Data Science in Finance and Economics

DSFE, 6(2): 277–314.
DOI: 10.3934/DSFE.2026010
Received: 08 September 2025
Revised: 22 December 2025
Accepted: 13 March 2026
Published: 10 May 2026

https://www.aimspress.com/journal/dsfe

Research article

ESG-XAI: An explainable unsupervised feature selection pipeline for ESG datasets

Tristan Lim[1,*], Zhiyuan Wang[2], Yun Teng[1]

[1] SUSS Academy, Singapore University of Social Sciences, Singapore
[2] School of Business, Singapore University of Social Sciences, Singapore

* **Correspondence**: E-mail: tristanlimms@suss.edu.sg.

Abstract: Problem. Environmental, Social, and Governance (ESG) panels are difficult to model, as variables live on mixed scales, redundancy within silos is high, cross-silo alignment is weak, and labels suitable for supervision are scarce or disputed.

Method: We present a deployment-ready, end-to-end workflow that turns such disclosure matrices into train-only labels for evaluation, out-of-sample performance estimates, and audit-ready explanations that trace to ESG indicators. For the system, we separated three planes, Discovery, Prediction, and Explanation, with strict train/test boundaries. Discovery built a balanced multi-view latent (per-silo PCA-fusion) and used evidence accumulation with stability rule to choose the number of segments (minimizing Proportion of Ambiguous Clustering). When global discovery was weak, an evaluation-feasibility guard constructed label-agnostic, stratified base splits and deferred labeling to training folds only. Prediction adopted nested labeling with nearest-centroid assignment to test, closing common leakage paths. Explanation was fold-local; we selected a balanced set of features per silo on the training fold and computed SHAP only on held-out tests.

Results: On a 2024 ESG panel (57 firms; 61 indicators: E=15, S=25, G=21), 2D linear discriminants (PCA/NCA-LDA) attained balanced accuracy ≈ 0.917, with low dispersion under strict nested evaluation, outperforming higher-capacity baselines. Performance metrics quantified the out-of-sample consistency and generalization of discovered latent structure, remaining stable even when raw cluster assignments varied. Fold-local explanations with human-interpretable drivers indicated social variables as dominant contributors, followed by environmental and governance variables.

Contributions. A portable, auditable ESG analytics pipeline enables evaluation when discovery is weak, reduces model risk, and delivers transparent outputs suitable for ESG use cases in small-n settings.

Key Words: ESG analytics; consensus clustering; stability-based model selection; nested evaluation; explainable AI (XAI)

JEL Codes: C38, G34, M14, Q56

1. Introduction

ESG data are challenging to analyze, especially when searching for strong predictors within a high-dimensional feature set. Variables use very different scales (for example, tons of emissions versus counts of board members) (Blank, Sgambati & Truelson, 2016), many indicators move together within each silo (E, S, or G), and alignment across silos is often weak (Billio et al., 2024; Dormann et al., 2013). Panels may be small relative to the number of indicators, which increases the risk of overfitting (Cawley & Talbot, 2010). Moreover, reliable "ground-truth" labels (e.g., who is "good" or "poor" on ESG) are rarely agreed upon (Kotsantonis & Serafeim, 2019; Berg et al., 2022). Furthermore, decision-makers need outputs they can trust and explain (Demartini & Pagliei, 2023): Are there clear patterns in the data? Do those patterns hold up for unseen firms? Which original indicators drive the observed separation?

As a result, we need an analytics workflow that: (i) Processes and normalizes high-dimensional and heterogeneous E/S/G data; (ii) reduces redundancy while preserving E/S/G domain coverage, even after projection to a lower-dimensional feature space; (iii) learns leakage-safe predictive relationships to future outcomes (e.g., financial returns, ESG risk or ESG controversies); and (iv) produces explanations that map directly back to concrete, auditable ESG indicators for governance and rebalancing.

In this paper, we address these needs by proposing and validating a simple, end-to-end workflow that cleanly separates phases of discovery, prediction, and explanation. The workflow is designed to be practical to run, resistant to common sources of leakage, and easy to audit. The idea is straightforward. First, in a high-dimensional feature space, we search for structure without labels. We compress each silo to a small, balanced latent space and look for segments (Xu et al., 2013; Yu et al., 2025). Instead of guessing how many segments to use, we choose the smallest number that yields clear and repeatable assignments under random perturbations (Fred & Jain, 2005; Monti et al., 2003; Şenbabaoğlu et al., 2014); if assignments remain ambiguous, we do not force a segmentation. In those weak-signal cases, we keep evaluation feasible by creating label-agnostic base splits that reflect broad variation in the data and postpone any labeling to the training portion of each split. Second, for every split, all transformations and all label induction are fit on training data; test firms are assigned in that training space by nearest centroid, and performance is computed on held-out data. In the lower-dimensional feature space, we deliberately favor low-capacity models (two-dimensional embeddings followed by linear discriminants) so that strong results, if present, are attributable to real signal rather than model complexity (Jolliffe & Cadima, 2016; Goldberger et al., 2004). Splits that are uninformative (e.g., single-class tests) are rejected by design. Third, within each training fold, we select

the same number of features from each silo to ensure balance and capacity control, fit a small tree surrogate on those features (Lundberg et al., 2018), and compute test-only SHAP (SHapley Additive exPlanations) values (Lundberg & Lee, 2017). SHAP values quantify conditional predictive salience within the fitted model and are not intended to support causal interpretation. This produces explanations that are clean (no leakage), compact, and easy to audit.

We demonstrate the workflow on a 2024 panel of listed firms (57 firms; 61 ESG indicators: $E = 15, S = 25, G = 21$). This setting is intentionally demanding, comprising small n, redundancy-rich with many correlated indicators, and no recorded labels. Under the strict protocol, simple two-dimensional linear discriminants (PCA-LDA and NCA-LDA) achieve high balanced accuracy and AUC with low variability, outperforming higher-capacity baselines evaluated the same way (Qu & Pei, 2024). Performance remains stable across discovery draws even when raw cluster assignments vary, indicating that the evaluation guard and leak-free nesting are doing their job. In this panel, explanations are silo-balanced by construction and point to a clear ordering of contributions. Social indicators dominate, followed by environmental and governance indicators. This provides a direct link from the latent separation back to human-interpretable drivers. A critical finding validated under this rigorous protocol is that simple, highly interpretable models, specifically, two-dimensional linear discriminants (e.g., PCA followed by LDA), achieve high balanced accuracy and AUC with low dispersion, often outperforming more complex, higher-capacity models evaluated under the same leak-free conditions. This result has significant practical implications, suggesting that effective ESG screening can be achieved with lower computational cost and greater transparency.

Research objective and contributions. In this research, we introduce a system-level, deployment-ready analytics pipeline for ESG scoring that systematically addresses key challenges in unsupervised financial modeling: data leakage, auditability, and model risk. The core contribution is an integrated methodology that cleanly separates the workflow into three distinct phases, Discovery, Prediction, and Explanation, enforcing strict train-test boundaries at every stage. This design ensures that resulting metrics, scores, and explanations are out-of-sample, auditable, and portable.

The contributions, organized by pipeline phase, are as follows:

(1) **Discovery plane – feasibility guard for weak discovery via stratified splits.** In small-n ESG panels, global clustering often yields ambiguous or single-class outcomes. If we try to score models on such labels, test metrics can become meaningless (e.g., a "classifier" could be trivially perfect on a single class) (Cawley & Talbot, 2010). To address this issue, prior to any label induction, we create label-agnostic, stratified base splits. This ensures that the subsequent evaluation remains feasible even if the clustering procedure on a training fold collapses to a single class. We defer all labeling logic to the training data, and report metrics only for splits that remain bi-class after train-only labeling. This prevents silent failure modes (e.g., single-class test sets), and ensures that balanced accuracy and AUC are meaningful.

(2) **Discovery plane – data-driven cluster selection.** The number of clusters is determined not heuristically or a priori, but via a stability analysis of the consensus co-assignment matrix across runs (Meinshausen & Bühlmann, 2010). In other words, we run multiple clustering variants and compute a stability score (PAC) on the consensus co-assignment matrix (Fred & Jain, 2005; Monti et al., 2003; Şenbabaoğlu et al., 2014). Thereafter, the smallest K is selected whose assignments remain consistently clear under these perturbations. Further, a

minimum-mass constraint is applied to prune tiny, unstable clusters. This enables cluster count selection to be robust to perturbations, data-driven, and reproducible.

(3) **Prediction plane – validation of unsupervised labels.** Within each data split, all preprocessing, scaling, dimensionality reduction, and the final cluster label induction are fitted exclusively on the training fold. Test-set firms are then projected into this pre-defined feature space and assigned labels based on the nearest training-derived centroid. This nested approach closes common leakage paths, enabling comparison to supervised baselines (Cawley & Talbot, 2010).

(4) **Explanation Phase – attributions.** For each split, we implement a strict explanation protocol. A pre-set number of top features (e.g., five per E, S, and G silo) are selected from the training fold to control capacity and ensure coverage across ESG domains. A compact surrogate model (e.g., a tree model) is then trained on these selected features using only the training data (Lundberg et al., 2018). Finally, SHAP values are computed on the held-out test data (Lundberg & Lee, 2017). These values yield balanced, leak-free, and traceable explanations that link the latent cluster-based separation to concrete, interpretable indicators, making them suitable for scrutiny by risk committees and regulators.

Novelty statement. While the individual technical components (e.g., clustering stability and SHAP explanations) are established in the literature, the primary contribution of this work is their integration into a cohesive, end-to-end pipeline. This integration, combining feasibility-guarded discovery, stability-driven segment choice, nested leak-free evaluation, and fold-local explanations, is validated as a single, auditable workflow. It produces valid out-of-sample metrics and explanations, thereby addressing a gap in the practical application of unsupervised feature selection to ESG analytics.

The remainder of this paper is organized as follows: In Section 2, we review the background and relevant literature on ESG analytics and related methodological works. In Section 3, we detail the 3-plane methodology, including its formulation and algorithm, where applicable. In Section 4, we present the empirical results. In Sections 5 and 6, we discuss the implications and threats to validity of the findings, while in Section 7, we conclude with reflections on contributions and future research directions.

2. Background and literature review

2.1. Unsupervised discovery for ESG panels

When ground-truth labels are sparse or disputed in a high dimensional dataset, as in many ESG settings, analysts typically begin by uncovering coarse segments in a lower-dimensional space and only later test whether those segments have predictive value. Choosing a reliable segmentation and an appropriate number of segments is challenging in heterogeneous, high-dimensional, small-sample panels. Classical internal geometry indices (e.g., silhouette, gap statistic) and likelihood criteria (AIC/BIC) are widely used for model order selection, yet numerous studies have shown that these criteria can be problematic. There may exist mixed scales, redundancy, and covariance regularization, which causes the outputs to fluctuate across preprocessing choices, seeds, and subsamples (Rousseeuw, 1987; Tibshirani et al., 2001).

To mitigate such instability, researchers have developed consensus (ensemble) clustering, which aggregates multiple base partitions, varying algorithms, resolutions, and seeds into a co-association (consensus) matrix whose entries reflect the empirical frequency with which pairs of observations co-cluster across runs (Liu & Blair, 2022; Fred & Jain, 2005; Golalipour et al., 2021). Clustering a dissimilarity derived from this matrix operationalizes evidence accumulation, thereby favoring pairwise relations that persist under algorithmic and data perturbations, and reducing variance relative to any single method (Monti et al., 2003; Şenbabaoğlu et al., 2014). Consensus methods have also been extended to multi-view settings, where representations from different blocks (such as E, S, and G) are combined. This enables complementary signals to be captured while reducing noise and redundancy within each block. (Xu et al., 2013; Zhao et al., 2024).

Within this consensus space, stability-based model order selection is commonly advocated over purely geometric or parametric criteria. A recurring principle is to prefer resolutions whose pairwise co-assignments concentrate near "almost always together" or "almost never together" rather than in ambiguous mid-ranges (Tao et al., 2017; Zhao et al., 2024). Resampling-based diagnostics, such as the consensus cumulative distribution and related ambiguity measures computed on the consensus matrix, have been used to summarize how decisively the data support a given resolution, with selection rules that minimize ambiguity and prioritize repeatability over parametric fit. Researchers conducting ensemble studies corroborate that such stability criteria are model-agnostic, naturally pool over algorithms and perturbations, and are well suited to small-n regimes (Liu et al., 2023; Zhao et al., 2024). Practical guidance also emphasizes parsimony when several resolutions appear similarly stable and warns against pathologically small segments that compromise downstream evaluation (e.g., stratified validation) (Chen & Yang, 2021; Hao et al., 2023).

As ESG variables naturally exist in semantically distinct blocks with different scales and collinearity profiles, the multivariate analysis literature recommends multi-view (multiblock) representations prior to discovery. A robust pattern is 'compress-then-fuse', which includes standardizing within each block, applying a small per-block projection (e.g., PCA) to collapse near-duplicates, concatenating the block scores, and optionally applying a compact global projection. This strategy improves redundancy control, enforces balance by construction across views, and preserves a clean mapping back to the original blocks for interpretability (Westerhuis et al., 1998; Smilde et al., 2003; Jolliffe & Cadima, 2016). In multi-view learning, such designs are widely used to stabilize clustering and downstream modeling in the presence of correlated, unequally scaled blocks.

A separate but crucial strand concerns leak-aware evaluation. Decades of work document how data leakage, computing scaling parameters, dimensionality reductions, model order, or labels on the full dataset lead to overly optimistic performance estimates, especially in small samples (Guignard et al., 2024; Cawley & Talbot, 2010; Varma & Simon, 2006; Kaufman et al., 2012; Varoquaux, 2018). Best-practice reviews therefore recommend nested protocols in which any data-dependent statistic is fit within training folds, with held-out folds mapped into the training representation and assigned labels via simple, reproducible rules such as nearest centroid or nearest neighbor (Cover & Hart, 1967; Hastie et al., 2009). Although standard in supervised learning, the literature notes that leak-aware nesting is often under-specified for unsupervised-label workflows, where the representation and the labels are induced without supervision; this gap is highlighted in methodological audits and reproducibility critiques (Sasse et al., 2025; Kaufman et al., 2012; Varoquaux, 2018; Kapoor & Narayanan, 2023).

2.2. Interpretable ESG explanations

In high-stakes settings such as ESG screening, stewardship, and index construction, auditors and regulators require clear, indicator-level traceability and credible out-of-sample evidence. The interpretability literature emphasizes that decision contexts with policy or compliance exposure should prioritize models or explanation protocols that support faithful, stable reasoning about drivers (Doshi-Velez & Kim, 2017; Rudin, 2019). For tabular problems, post-hoc explanations are common, but they are only defensible if (i) the feature set is reproducible under perturbations and (ii) the attribution computation does not contaminate evaluation (Adebayo et al., 2018).

A central challenge in small-sample, high-dimensional panels is that naïve single-run feature selection is unstable (subsample perturbations produce different lists). Stability selection provides a statistically principled remedy by repeating selection under resampling and retaining variables with consistently high selection frequency, thereby improving reproducibility and controlling false discoveries (Hofner et al., 2015; Meinshausen & Bühlmann, 2010). Beyond unstructured sparsity, group-aware penalties (e.g., group lasso and sparse-group lasso) encourage block-level parsimony and help prevent over-representation of correlated feature families. This is important when heterogeneous variables exist in domain silos such as E/S/G (Simon et al., 2013; Lim & Hastie, 2015). Complementary approaches like knockoffs provide false discovery guarantees for feature importance under broad conditions and have been adapted to complex tabular settings (Barber & Candès, 2015; Candès et al., 2018). These strands establish that shortlist construction can be made sparse and replicable, and that group structure can be used to preserve domain balance.

Among post-hoc methods, SHAP has become prominent because it yields additive, instance-level attributions with desirable axiomatic properties (Lundberg & Lee, 2017). For tree ensembles, TreeSHAP provides efficient, theoretically consistent attributions and supports aggregation from local effects to feature- and group-level importance. These are capabilities that align with indicator- and silo-level reporting in ESG (Lundberg et al., 2020). Surveys and frameworks for tabular explainability reinforce these advantages while cautioning that explanations must be contextualized with uncertainty and stability checks. Methodological work also highlights limitations, where local additive explanations can be sensitive to data distribution shifts, correlated features, and model misspecification, so practitioners could pair SHAP with robustness diagnostics and domain structure (Covert, et al., 2021).

A recurring failure mode in applied ML is data leakage, where statistics computed on the full dataset (e.g., scaling, feature screening, or even explanation models) leak information from test folds into training decisions, inflating reported performance and interpretability claims (Guignard et al., 2024; Cawley & Talbot, 2010; Varma & Simon, 2006; Kaufman et al., 2012; Varoquaux, 2018). This risk extends to the interpretability layer. Explanations computed after refitting on train-test or after using test data to tune feature selection create optimistic narratives that do not reproduce (Kaufman et al., 2012; Varoquaux, 2018). Best-practice analyses therefore recommend nesting interpretability inside cross-validation; that is, perform any selection or explainer fitting within training folds, compute attributions only on held-out data, and running sensitivity checks to detect explanation issues (Adebayo et al., 2018).

2.3. ESG analytics constraints

The empirical properties of ESG disclosure pose well-documented challenges for machine learning. Heterogeneous measurement scales, e.g., emissions volumes, resource intensities, board counts, and composition ratios, distort distance geometry and can bias scale-sensitive methods when normalization and variance control are weak (Blank et al., 2016). In addition, strong within-silo collinearity is common. Environmental indicators frequently track the same underlying intensity factors; social metrics cluster around workforce benefits and compensation; and governance variables co-move through ownership concentration and board structure. High redundancy inflates effective dimensionality and complicates feature selection and inference, a pattern mirrored across applied domains. By contrast, cross-silo alignment is typically weak. Environmental, social, and governance blocks encode partially orthogonal constructs. Studies comparing ESG blocks and composite scores report modest cross-block correlations and divergent explanatory content, implying that pooled representations risk dominance by the largest or noisiest block and may obscure complementary structure (Billio et al., 2024; Dormann et al., 2013).

Many ESG datasets are small-n, high-p ($n \leq p$), where the number of firms is comparable to, or smaller than, the number of indicators (Wang et al., 2013; Klaaßen et al., 2024). This regime magnifies variance and makes overfitting easy, especially when models are complex or selection steps are not strictly separated from evaluation (Cawley & Talbot, 2010). In addition, label scarcity and disagreement are pervasive: Agency ratings and controversy labels diverge substantially, limiting reliable supervised targets and motivating unsupervised discovery as a first step (Kotsantonis & Serafeim, 2019; Berg et al., 2022).

These statistical and institutional realities coincide with tightening governance expectations for auditability and explainability in sustainability analytics. Disclosure and assurance frameworks (e.g., emerging reporting standards) emphasize traceable modeling, reproducible metrics, and stakeholder-interpretable drivers (Demartini & Pagliei, 2023). Moreover, in ML, transparent and testable rationale for predictions is repeatedly recommended to manage model risk and enable oversight (Rudin, 2019; Varoquaux, 2018).

Relative to prior art (Lim, 2024), we integrate a single, auditable pipeline designed for small-n ESG. This comprises a feasibility guard that preserves evaluability when discovery is weak; stability-based selection of segment count with minimum-mass protection; nested, leak-free labeling/testing for unsupervised targets; and fold-local, test-only explanations using balanced per-silo shortlists. To our best knowledge, this particular combination, optimized for small-n ESG data with explicit E/S/G constraints, has not been systematically articulated in the literature.

3. Methodology

3.1. Process Overview

We design a workflow that is organized into three planes: (i) *discovery,* (ii) *prediction*, and (iii) *explanation*. The central premise is that ESG data may be small-n, high-p, heterogeneous across silos (E, S, G), and only weakly labeled at best. Consequently, any credible pipeline must (i) extract structure

without using labels, (ii) evaluate predictive discrimination strictly on held-out data with labels induced *inside* training folds only, and (iii) attribute model decisions with explanations computed on held-out tests. Our methodology instantiates these requirements in a linear process with explicit feasibility guards and fallbacks, so that analysis proceeds even when the data regime is unfavorable (e.g., extremely imbalanced or near-degenerate clusterings), but not at the expense of evaluation hygiene.

Plane I: *Discovery*. The workflow begins with data hygiene and pruning. ESG indicators vary widely in scale, redundancy, and measurement quality. We remove quasi-constant columns that cannot support discrimination and apply rank-based, within-silo de-duplication to eliminate near-duplicate variables while retaining the more informative representative. The result is a cleaned feature matrix $X \in \mathbb{R}^{n \times p}$ and an updated mapping of variables to the E, S, and G silos. This step reduces variance in downstream estimates without imposing strong modeling assumptions or feature selection that could collapse in small-n settings.

To prevent any single silo from dominating, we embed the data into a multi-view consensus space. Each silo is standardized and compressed with a small principal component analysis (PCA). The resulting low-dimensional views are concatenated, re-standardized, and projected via a final PCA to a compact latent space $Z \in \mathbb{R}^{n \times K}$. This two-stage construction balances contributions from E, S, and G. Crucially, this latent space is learned *without labels* and is used only for unsupervised discovery and train-fold labeling in later steps.

We then perform evidence-accumulation consensus clustering in Z. Rather than trusting a single algorithm or a single choice of the number of clusters K, we run a portfolio of base clusterers (e.g., k-means, Gaussian mixture models, agglomerative, and spectral) across a grid of candidate cluster counts and over many randomized subsamples of the rows. For each run, we record whether two observations co-occur in the same cluster; averaging over runs yields a co-association matrix whose entries quantify the *stability* of pairwise assignments. We select K automatically by minimizing the Proportion of Ambiguous Clustering (PAC) computed on the averaged co-association: The chosen K with the smallest density in the mid-range of $[0,1]$. Final labels are obtained by clustering the dissimilarity $1 - \bar{C}$. A light-touch "minimum class size" repair merges only the smallest, clearly non-viable classes into their nearest neighbors, preserving feasibility without masking genuine structure.

Plane II: *Prediction*. As global unsupervised labels can legitimately collapse in small-n ESG data (e.g., one class after stability filtering), we introduce an evaluation safeguard that activates when label collapse prevents stratified cross-validation. The guard constructs label-free, quantile-stratified train/test splits using an unsupervised score (first principal component or PC1 of standardized inputs), with a sampler that ensures the *test* portion of a split will contain at least two classes after train-only labeling. This mechanism preserves the integrity of downstream evaluation without inventing labels on the test set or over-fitting the splitting rule to labels.

Predictive performance is assessed via nested, leak-free evaluation. For each split independently, we refit the discovery transforms on the *training* fold only, induce train-only labels by the same consensus procedure (with conservative fallbacks if needed), and then assign labels to test observations by nearest centroid in the train-fit latent space. Only after labels are fixed do we train the candidate dimensionality-reduction-plus-classifier pipelines (e.g., PCA-LDA, NCA-LDA) on the training data and score on held-out tests using accuracy, balanced accuracy, macro-F1, and AUC. No statistic from the test partition participates in discovery, label induction, feature screening, or model selection. This

separation of concerns, discovery on train, assignment to test via a fixed rule, evaluation on held-out, constitutes the core leakage defense.

Plane III: *Explanation*. Finally, we provide fold-local interpretability under the same hygiene. Within each training fold, we run a group-sparse stability selection procedure that yields exactly five variables per silo (5 E, 5 S, 5 G). The selector first screens by mutual information, stabilizes via bootstrap resampling with group-wise regularization (group lasso) when groups are defined, and breaks ties/redundancy using a small mRMR ensemble. The resulting top 15 are used to train a compact, transparent model on the *training* fold; SHAP values are computed only on the held-out test fold and aggregated across splits to obtain robust, leakage-free explanations at the feature and silo levels. As selection is fold-local and silo-balanced, the explanations are comparable across splits and resilient small-samples.

The proposed design thus offers three properties that are essential for ESG inference: stability (via evidence accumulation and Auto-K by PAC), validity (via nested, label-inducing evaluation with nearest-centroid assignment and explicit feasibility guards), and interpretability (via fold-local, group-balanced stability selection and test-only SHAP). In the following subsections, we formalize each component, specify the algorithms and hyperparameters, and provide complexity and robustness considerations.

The high-level process flowchart i**s** described in Figure 1. In subsections 3.2 to 3.7, we formalize each block with algorithmic detail and hyperparameter settings, corresponding to Blocks I-VI in Figure 1. Block A performs data hygiene and redundancy pruning within each ESG **si**lo to remove quasi-constants and near-duplicate indicators. Block B constructs a balanced multi-view latent space via per-silo dimensionality reduction and fusion. Block C applies stability-based evidence accumulation to determine whether meaningful unsupervised structure exists and to select K when feasible. When global discovery is degenerate, Block D introduces an evaluation-feasibility guard that generates label-agnostic, stratified base splits. Block E induces labels strictly within training folds and evaluates predictive performance under a fully nested, leakage-aware protocol. Finally, Block F produces fold-local, test-only explanations via capacity-controlled surrogates and SHAP, preserving auditability and reproducibility.

Notation and conventions. We denote by n, the number of entities, and by p, the number of variables. The raw feature matrix is $X \in \mathbb{R}^{n\times p}$,$X[:,j]$ in column j. ESG silos are $g\epsilon\{\mathcal{E},\mathcal{S},\mathcal{G}\}$. For a silo $g\epsilon\mathcal{G}$, let $I_g \subseteq \{1,\dots,p\}$ index its variables and write $X_g := X[:,I_g]$. Upper-case letters denote matrices, lower-case italics scalars, and sets appear in roman/calligraphic braces. Absolute value is $|\cdot|$. All computations in pruning operate on the currently surviving columns; ties (e.g., $m(a*) = m(b*)$) are broken deterministically (variance, then lexicographic index).

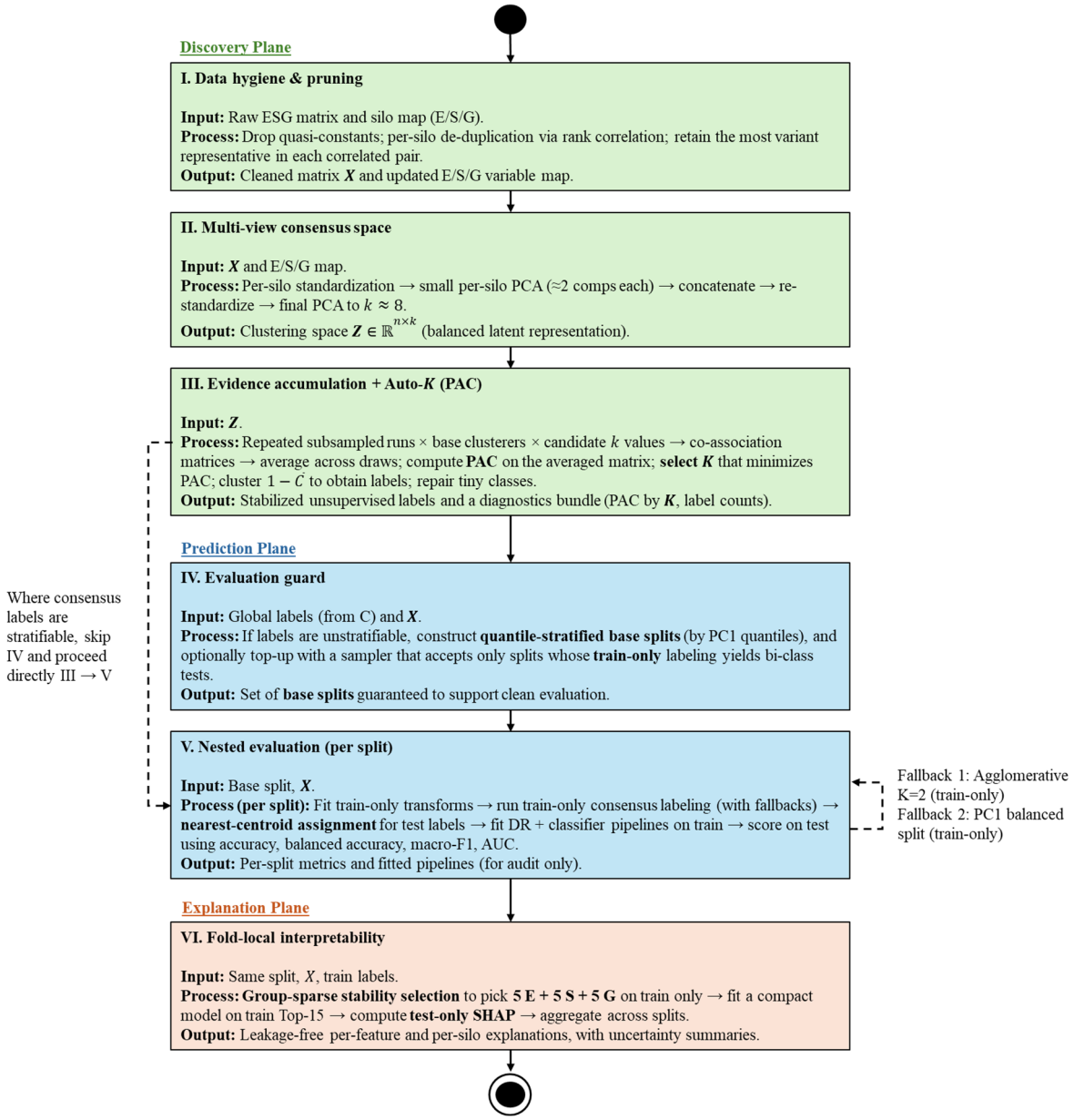

Figure 1. A leakage-averse three-plane workflow for ESG analytics.

Figure 1 summarizes a six-block pipeline that separates unsupervised discovery (I–III) from supervised evaluation (V) and fold-local interpretability (VI), with an evaluation guard (IV) that activates only when consensus labels are not stratifiable. Plane I (Discovery) begins with data hygiene and pruning of the raw matrix in Block I. In Block II, a balanced multi-view latent space is formed by per-silo standardization, small per-silo PCA, concatenation, re-standardization, and a final PCA (k≈8). Block III performs evidence accumulation across subsamples, base clusterers, and candidate k values to build co-association matrices; K is chosen by minimizing PAC on the averaged co-association, followed by minimum-class-size repair. Moving to Plane II (Prediction), a dashed path III→V indicates the default bypass of the guard when labels are stratifiable; the solid III→IV→V path indicates activation of the quantile-stratified base-split guard (and optional top-up) when they are not. Within each split (V), all transforms and labeling are fit on train only; test points receive labels via

nearest centroids in the train-fit space; no test statistic influences discovery or selection. At Plane III (Explanation), Block VI conducts fold-local group-sparse stability selection to pick exactly 5 E + 5 S + 5 G features for training, trains a compact model, and computes test-only SHAP; explanations are aggregated across splits. Dashed callouts indicate conservative fallbacks (Agglomerative K=2, then PC1 split) used only when train-only consensus degenerates.

3.2. Discovery Plane – Data hygiene and pruning

The goal of Block A is to define a stable measurement space before any modeling takes place. ESG indicators are often heterogeneous in scale, subject to reporting plateaus (producing quasi-constants), and highly collinear within silos (E, S, G). Retaining degenerate or near-duplicate signals inflates variance downstream (most visibly in co-association matrices) and weakens interpretability. We therefore apply two label-free filters, namely quasi-constant removal and within-silo de-duplication, that are invariant to monotone transformations and robust to outliers. Block A is executed once to define the fixed feature universe; all learned transforms used for prediction (e.g., scalers, PCA/NCA) are later re-fit inside training folds, preserving evaluation hygiene. The algorithm is described in Algorithm 1.

Local notation. For each feature X_j, let $nunique(X_j)$ denote the number of distinct values, and let $\rho_s(a, b)$ denote the Spearman rank correlation between features a and b.

Quasi-constant removal. For each variable X_j, we compute the number of distinct values across the n entities. The screening threshold for near-constants is $\mathcal{u} \in \mathbb{N}$. Variables with fewer than $\mathcal{u}$ unique values are discarded (we implement $\mathcal{u} = 3$). The minimum class size $\mathcal{u}$ enforces evaluability under stratified testing by preventing degenerate or single-instance classes in small-n settings. This threshold serves to reduce near-zero variance issues (Kuhn & Johnson, 2013) that cannot support meaningful discrimination or stable scaling while preserving discretized signals (e.g., trinary scores) that exhibit variation.

Within-silo de-duplication. To reduce redundancy while scale invariant, within each silo $g\epsilon\{\mathcal{E}, \mathcal{S}, \mathcal{G}\}$, we compute the absolute Spearman correlation $|\rho_s(a, b)|$ within each silo separately. For any pair (a, b) with $|\rho_s(a, b)| \geq \tau$, we retain the more "invariant" member of the pair using a robust spread score by comparing their absolute deviation magnitudes (implemented as the mean absolute deviation from the mean) and dropping the lower-spread variable.

The redundancy threshold $\tau \in [0,1]$ is applied to absolute Spearman correlation $|\rho_s(a, b)|$ between variables a and b. For any current silo, the most strongly correlated pair is

$$(a*, b*) = \underset{a \neq b \in I_g}{\operatorname{argmax}} |\rho_s(a, b)|$$

To obtain the more variable member of a correlated pair, we use a robust spread score $m(\cdot)$, instantiated as the mean absolute deviation (MAD):

$$m\,(X[:, j]) = mean(|\,X[:, j] - mean\,(X[:, j])\,|)$$

This favors the feature that exhibits greater variability after controlling for monotone re-scalings. Rank correlation is chosen because it is invariant to monotone transforms and less sensitive to outliers

than Pearson correlation. When pruning a correlated pair, we drop $argmin\{m(a*), m(b*)\}$. The correlation-based redundancy pruning procedure iterates over pairs until no correlations exceed the threshold, yielding a pruned, non-redundant set per silo (Guyon & Elisseeff, 2003; Dormann et al., 2013). Ties are broken deterministically (by column order) to ensure reproducibility. The cleaned matrix is X_{clean}, with a pruned silo map $g' \epsilon \{\mathcal{E}', \mathcal{S}', \mathcal{G}'\}$.

Algorithm 1: Data Hygiene and Pruning

Input: feature matrix $X \in \mathbb{R}^{n \times p}$; silo map $g \epsilon \{\mathcal{E}, \mathcal{S}, \mathcal{G}\}$; parameters: $\mathcal{u}$ (min unique values, default $\mathcal{u} \leftarrow 3$), τ (corr. threshold, default $\tau \leftarrow 0.92$)
Output: X_{clean}, pruned silo map $g' \epsilon \{\mathcal{E}', \mathcal{S}', \mathcal{G}'\}$

drop columns with $nunique < \mathcal{u}$ *# near-zero variance screening*
for each silo $g \epsilon \{\mathcal{E}, \mathcal{S}, \mathcal{G}\}$:
    compute $|\rho_s(a, b)|$(Spearman) for all pairs $a, b \epsilon g$
    while $max_{a \neq b} |\rho_s(a, b)| \geq \tau$:
        let $(a*, b*)$ be the most correlated pair
        compute absolute-deviation score $m(\cdot)$ for $a*, b*$
        drop $argmin\{m(a*), m(b*)\}$ *# correlation-based redundancy pruning*
return X_{clean} and pruned silo map g'.

ESG indicators within the same silo (e.g., multiple governance turnover ratios) frequently encode the same latent factor under different scaling. Removing near-duplicates decreases estimation variance by preventing a single latent signal from being over-counted. This also improves the stability and readability of downstream explanations. As both filters are unsupervised and involve no fitted models, they do not create opportunities for label leakage. Furthermore, we implement a fixed, high-correlation pruning threshold ($|\rho_s| \geq 0.92$), consistent with standard de-duplication practice in high-dimensional settings. Results are not sensitive within this near-duplicate regime, and the threshold is chosen a priori, trading a small amount of redundancy for substantial reductions in variance and computational load, without risking the removal of distinct, weak signals that may be important at a later stage.

Within-silo rank correlations require $O(p_g^2)$ pairwise computations per silo of size p_g, which is tractable for typical ESG panels. Absolute-deviation calculations are linear in n. The use of rank correlation makes the decision rule insensitive to heavy-tailed scales; the absolute-deviation tie-breaker avoids the brittleness of variance under outliers. Finally, because pruning is performed per silo, contributions from E, S, and G remain balanced as the pipeline proceeds to the multi-view embedding in Block B.

3.3. Discovery Plane – Multi-View consensus space

Given the small-n, high-p structure of our data, dimensionality reduction is required to mitigate overfitting and stabilize estimation (Cawley & Talbot, 2010; Jolliffe & Cadima, 2016). ESG variables are naturally partitioned into three silos (E, S, and G) with markedly different dimensionalities and scales. If we embed all variables at once, the largest or most variable silo can dominate, making any

downstream discovery sensitive to silo imbalance rather than to genuine common structure. In the multi-view learning approach (Smilde et al., 2003; Westerhuis et al., 1998; Xu et al., 2013; Yu et al., 2025; Lock et al., 2013), Block B constructs a balanced, label-free latent space that (i) respects the silo partition, (ii) mitigates collinearity within and across silos, and (iii) yields a compact representation well-conditioned for consensus clustering.

Local notation. Let $X \in \mathbb{R}^{n\times p}$ be the cleaned matrix from Block A and let $g\epsilon\{\mathcal{E},\mathcal{S},\mathcal{G}\}$ index a silo with column set I_g.

Procedure.

1. **Per-silo standardization.** For each silo g, transform the submatrix $X_g = X[:, I_g]$ to zero mean and unit variance column-wise:

$$Z_g = Std(X_g)$$

where $Std(\cdot)$ refers to the column-wise z-scaling (mean 0, variance 1) computed on the *training fold* when used inside CV. PCA is scale-dependent and requires standardization when variables differ in units (Jolliffe & Cadima, 2016).

2. **Small per-silo PCA.** Fit a PCA on Z_g and retain

$$d_g = min(d_{per-view}, \quad rank(Z_g), |I_g|)$$

components for silo g, where $d_{per-view} \in \mathbb{N}$ is the small per-silo PCA target (we use $d_{per-view} \approx 2$ by default).

Denote the score matrix by

$$H_g = X_g P_g,$$

where P_g contains the first d_g loading vectors of the silo-specific PCA, and $H_g \in \mathbb{R}^{n\times d_g}$ is the silo-level score.

3. **View fusion.** Concatenate the three silo scores horizontally to combine the scores,

$$H = [H_E H_S H_G] \in \mathbb{R}^{n\times d_\Sigma}, d_\Sigma = \sum_{g\in\{E,S,G\}} d_g,$$

and re-standardize columns of H to equalize scale across silos:

$$\widetilde{H} = Std(H)$$

4. **Global PCA.** Apply a second PCA to $\widetilde{H}$ and keep

$$k = min(k_{max}\ ,\ d_\Sigma,\ n-1)$$

components (we apply a one-digit $k \approx 8$, while respecting the sample cap $n-1$). $k_{max} \in \mathbb{N}$ is the cap for the global dimensionality (e.g., $k_{max} \in [6,10]$).

The resulting consensus space is

$$Z = \widetilde{H} Q_k$$

where $Q_k \in \mathbb{R}^{d_\Sigma \times k}$ is the first k loading vectors of PCA on $\widetilde{H}$, and $Z \in \mathbb{R}^{n \times k}$ is the multi-view consensus input for Block C and for train-fold embeddings in Block E (evidence-accumulation consensus clustering).

For edge cases, if a silo is empty after pruning, we form H from the remaining silos. If all silos would be empty (pathological), we fall back to $Std(X)$) followed by a single PCA with the same k cap.

The two-stage multi-view learning small-per-silo PCA to global PCA follows sequential multiblock approaches (Smilde, Bro, & Geladi, 2003) and the shared vs. individual variation framework (Lock et al., 2013). The first stage compresses each silo into a fixed, low-rank summary so that a high-dimensional silo (e.g., S) cannot overwhelm a low-dimensional silo (e.g., G). The second stage then captures cross-silo covariation in a compact basis while the intermediate re-standardization ensures comparable variance for each view. All steps are unsupervised and fitted without labels, which preserves the separation between discovery and evaluation required by our leakage-averse design.

The following hyperparameter guidance are considered for robustness:

- $d_{per-view}$ may be small (1–4) to avoid over-parameterizing in small-n regimes; we set $d_{per-view} = 2$.
- k is capped by $n - 1$ and d_Σ (Bishop, 2006; Jolliffe & Cadima, 2016). In practice, $k \in [6,10]$ suffices for stable clustering while filtering measurement noise.
- Standardization at both stages neutralizes arbitrary units and prevents heteroscedasticity from steering the embedding.

Through this balanced per-silo approach, Block B produces a latent space that is not dominated by any single silo and is therefore more faithful to multi-criteria ESG structure. This consensus space Z is the common input for the evidence-accumulation procedure in Block C and for the train-fold embeddings used in nested evaluation (Block E), while remaining label-agnostic.

3.4. Discovery Plane – Evidence Accumulation + Auto-k (PAC)

Next, latent ESG regimes are estimated with an evidence-accumulating consensus procedure that pools signals across clustering methods, resolutions, subsamples, and feature bags. The goal is to produce labels that are stable and usable for downstream evaluation (i.e., admit stratified train/test splits without leakage). The algorithm is described in Algorithm 2.

Local notation. Let $Z \in \mathbb{R}^{n \times k}$ denote the latent representation from Block B. For a given clustering method $m \in \mathcal{M}$ and resolution $K \in \mathcal{K}$, each run produces labels $\ell \in \{1, \dots, K\}^n$. The ensemble updates a co-association matrix $C \in [0,1]^{n \times n}$, with averaged entries C_{ij} representing the empirical probability that items i and j are co-clustered across perturbations.

Consensus construction. For each draw $t = 1, \dots, T$, we (i) randomly subsample $\lfloor \phi n \rfloor$ rows and (ii) randomly bag features (a fixed fraction with a small absolute floor to avoid degeneracy). Within each draw, we run an ensemble of base clusterers

$$\mathcal{M} = \{\text{kmeans}, \text{GMM}, \text{agglomerative}, \text{spectral}\},$$

at resolutions $K \in \mathcal{K}$ (we consider a small grid, $\mathcal{K} = \{2,3,4\}$), producing hard labels $\ell \in \{1, \dots, K\}^n$ on the active subsample.

Each run updates a co-association matrix $C \in [0,1]^{n\times n}$ by

$$C_{ij} \leftarrow C_{ij} + \mathbf{1}[\ell_i = \ell_j],$$

restricted to the indices in that subsample; pairs outside the subsample are left unchanged. After normalizing by the number of runs, averaging over draws and methods yields an ensemble co-association $\bar{C} \in [0,1]^{n\times n}$ whose (i,j) entry estimates the probability that items i and j co-cluster across perturbations of data, features, and algorithms.

Auto-K via the PAC criterion. Rather than fix K a priori, we select it adaptively by minimizing the PAC:

$$PAC(\bar{C};\ell,h) = \frac{1}{n(n-1)}\sum_{i\neq j}\mathbf{1}\{\ell < \bar{C}_{ij} < h\},$$

with a conservative ambiguity band $(\ell,h) = (0.1, 0.9)$. Intuitively, PAC penalizes co-association values that are neither near 0 (consistently apart) or 1 (consistently together). K that minimizes PAC is the resolution with the fewest ambiguous pairs. We compute $\bar{C}_K$ for each $K \in \mathcal{K}$ and set

$$K^* = \operatorname*{argmin}_{K\in\mathcal{K}} PAC(\bar{C}_K).$$

Labels and size repair. Given $\bar{C} = \bar{C}_{K^*}$, we define a dissimilarity $D = 1 - \bar{C}$ and apply average-linkage agglomerative clustering to obtain K^* groups. Finally, to ensure feasibility for stratified validation and to avoid pathologies in small samples, we apply a minimum-class-size repair: While any class has size $< m_{min}$, we merge the smallest class into the nearest-centroid neighbor in the same Z space, stopping once all classes satisfy the threshold or a single non-trivial split would be violated. This guarantees labels that are stable (by construction) and usable (each class is represented).

Algorithm 2: Auto-K Consensus with PAC

Input: latent space Z, candidate K grid $\mathcal{K}$, methods $\mathcal{M}$, number of draws T, repetitions per draw R, subsample fraction ϕ, PAC band (ℓ,h), minimum class size m_{min}
Output: consensus labels y, averaged co-association $\bar{C}$, chosen K^*

1 for each draw $t = 1, \ldots, T$, and for each $K \in \mathcal{K}$:
  a. initialize $C^{(t,K)} \leftarrow 0$.
  b. repeat R times: subsample $\lfloor \phi n \rfloor$ rows, bag features, run each $m \in \mathcal{M}$ at $k = K$; update $C^{(t,K)}$.
  c. normalize $C^{(t,K)}$ by the number of runs.
2 for each K: compute $\bar{C}_K = \frac{1}{T}\sum_t C^{(t,K)}$; evaluate $PAC(\bar{C};\ell,h)$.
3 select $K^* = \operatorname*{argmin}_{K\in\mathcal{K}} PAC(\bar{C}_K)$; set $\bar{C} \leftarrow \bar{C}_{K^*}$.
4 cluster $D = 1 - \bar{C}$ into K^* groups via average linkage → labels y.
5 apply minimum-class-size repair until all classes $\geq m_{min}$ or only one non-trivial split remains.
6 return $y, \bar{C}, K^*$.

We include the implementation details and robustness guards for this process:

- To prevent singularities at small n or near-collinearity, GMM is run with progressive covariance/regularization fallbacks and an automatic clamp on K to the number of distinct patterns observed (after rounding), guaranteeing a well-posed fit.
- Spectral clustering uses a nearest-neighbors graph with $n_b \approx \sqrt{n}$ neighbors (clamped within a small range), which is empirically stable in low-sample regimes (Lucińska & Wierzchoń, 2012).
- Clamp K to the number of distinct rows to avoid "distinct clusters < k" failures, and use multiple starts as available.
- Feature bagging retains a minimum number of variables (e.g., ≥ 12) to avoid degenerate runs; $\phi \in (0.7, 0.8)$ balances perturbation and geometric fidelity.

This block seeks a balance between parsimony (small K), stability (low ambiguity under perturbations), and feasibility (classes large enough for unbiased validation). PAC provides an interpretable scalar that declines when co-assignments concentrate near 0 or 1 and rises when assignments are equivocal. Minimizing PAC therefore selects the resolution that is most decisive under perturbations. The subsequent size repair is conservative. It alters labels only as needed to meet m_{min} and preserves at least one non-trivial split, ensuring that downstream stratified procedures are well-defined.

The procedure yields (i) low-ambiguity $\bar{C}$ with clear block structure, (ii) valid stratified splits without resorting to ad-hoc relabeling, and (iii) stable labels across draws, as quantified by standard agreement metrics (e.g., ARI/NMI) reported elsewhere in the results. By aggregating co-association averaging, Auto-K selection via PAC, and minimal size repair, Block C delivers data-driven, stable, and CV-feasible labels. These properties are useful for the subsequent leak-free evaluation and for fold-local interpretability.

3.5. Prediction Plane – Evaluation guard

Unsupervised discovery in small-n ESG settings can yield trivial solutions, such as a single dominant cluster or partitions that collapse under stability constraints. While these outcomes are valid from a data-driven perspective, they present a methodological problem for evaluation: Cross-validation procedures require multiple classes to construct stratified splits and estimate discriminatory performance. Fabricating labels in test folds would compromise validity, while discarding degenerate cases altogether risks biasing results toward overly "favorable" configurations. Block D therefore introduces an evaluation-feasibility guard, designed to preserve test informativeness without leaking information across folds. The algorithm is described in Algorithm 3.

Local notation. Let $X \in \mathbb{R}^{n \times p}$ denote the cleaned ESG matrix from Block A, and let $Z \in \mathbb{R}^{n \times k}$ denote the consensus latent space from Block B. Denote by π the set of target base splits to be generated. Let $\mathrm{PC1}(X)$ represent the first principal component of standardized X. For a quantile parameter $q \in \mathbb{N}$, define B_q as the partition of $\mathrm{PC1}(X)$ into q equal-frequency bins.

Quantile-stratified base splits. The default guard procedure constructs train/test partitions that are independent of any discovered labels but nonetheless ensure class separability ex post.

1. Standardize X to zero mean and unit variance column-wise.
2. Project to the first principal component (PC1).
3. Bin entities into q quantiles along PC1, yielding B_q.
4. Sample train/test folds stratified by these bins.

This procedure ensures that train and test sets contain representation across the dominant unsupervised axis of variation, thereby increasing the likelihood that subsequent label induction (Block E) yields bi-class structures in both partitions. Importantly, PC1 is used solely as a label-agnostic axis of dominant variance and is computed without reference to any induced labels; the stratification neither encodes class information nor optimizes performance, as all label induction, model fitting, and evaluation remain strictly confined to training folds under the nested protocol, with test data used only for out-of-sample assignment and scoring.

Top-up sampler. If the number of feasible splits (i.e., those with at least two classes in train and test folds) falls below the target $|\pi|$,

1. Draw a candidate train/test split.
2. Within the train partition, induce labels via Block E (train-only consensus).
3. Assign test labels by nearest centroid in the train latent space.
4. Retain the split only if both train and test contain at least two classes.

This procedure preserves test informativeness while respecting the leakage constraint, such that no statistic from the test set influences label induction or model training.

Algorithm 3: Evaluation Guard for Feasible Splits

Input: cleaned matrix X, consensus latent Z, quantile parameter q, target number of splits $|\pi|$
Output: feasible split set π

1 Standardize X; compute PC1(X).
2 Bin into q quantiles; sample base train/test splits stratified by bins.
3 Evaluate feasibility: discard splits where induced train/test labels (from Block E) collapse to one class.
4 If $|\pi|$ < target:
  a. sample candidate split.
  b. induce train-only labels via Block E.
  c. assign test labels by nearest centroid.
  d. retain split if both partitions are bi-class.
5 Return final set π.

The evaluation guard is a useful hinge between unsupervised discovery (Block C) and leak-free prediction (Block E). The evaluation guard provides three safeguards. First, by stratifying on PC1 rather than on discovered labels, it prevents data leakage while aligning splits with the strongest unsupervised axis of variation. Second, the quantile mechanism ensures that small-n imbalances do not result in empty test folds. Third, the top-up sampler provides a fallback that maintains the target number of splits without artificially inflating class structure. This design enables evaluation to proceed even under degenerate discovery outcomes, without introducing bias from label engineering or test leakage.

3.6. Prediction Plane – Nested evaluation (Per Split)

The central challenge in evaluating predictive models on small-n, weakly labeled ESG datasets lies in preventing leakage, such that no statistic from the test partition should influence label induction,

feature selection, or model fitting. Block E implements a nested evaluation protocol, in which discovery and labeling are refit inside each training fold, test items are assigned via a fixed mapping rule, and performance is assessed only on held-out data. This ensures that results reflect genuine generalization rather than inadvertent reuse of test information. The algorithm is described in Algorithm 4.

Local notation. Let $(\mathcal{I}_{train}, \mathcal{I}_{test})$ denote a train/test split from the guard procedure. For the training subset, let $X_{train} \in \mathbb{R}^{n_{tr} \times p}$ be the raw feature matrix and $X_{test} \in \mathbb{R}^{n_{te} \times p}$ the corresponding test features. The goal is to induce labels y_{train}, y_{test} in a leakage-free manner and evaluate pipelines $\pi \in \Pi$, where Π is the set of dimensionality-reduction (DR) and classifier combinations.

Procedure.

1. **Train-fit representation.** Standardize X_{train} column-wise, optionally adding small jitter for numerical stability, and fit a PCA capped at ≤ 10 components. This yields the train latent space Z_{train}. Apply the same scaler and PCA mapping to obtain Z_{test} for the held-out data.
2. **Train-only labeling.** Induce labels on Z_{train} using the consensus procedure earlier described, but restrict it to the training fold and a compact grid of candidate K. If consensus clustering collapses to a single class, we apply conservative fallbacks: (i) $K = 2$ agglomerative clustering on Z_{train}, or (ii) a balanced PC1 split, which guarantees two groups. These fallbacks maintain feasibility while avoiding fabricated test labels.
3. **Test assignment.** Compute class centroids in Z_{train}. Each test point in Z_{test} is assigned to the nearest centroid, ensuring that no test information influences the induction of cluster structure. This nearest-centroid rule provides a reproducible mapping from train-induced labels to the held-out test set.
4. **Modeling and scoring.** With labels fixed, we train candidate pipelines $\pi \in \Pi$ on (X_{train}, y_{train}) and evaluate on (X_{test}, y_{test}). Pipelines include combinations of dimensionality reduction (i.e., PCA, NCA, and LDA) with classifiers (i.e., LDA, RBF-SVM, and bagging), optionally augmented by graph-based features. Metrics reported are accuracy, balanced accuracy, macro-F1, and AUC (where applicable).

Algorithm 4: Leak-Free Nested Labeling and Evaluation

Input: feature matrix X, split set $\mathcal{S}$, pipeline family Π.
Output: per-pipeline held-out metrics.
For each $(\mathcal{I}_{train}, \mathcal{I}_{test}) \in \mathcal{S}$:
1 Fit scaler + PCA on X_{train}; obtain Z_{train}, Z_{test}.
2 Apply consensus on Z_{train} to get y_{train}; if degenerate, apply fallbacks.
3 Compute centroids in Z_{train}; assign y_{test} by nearest centroid in Z_{test}.
4 For each $\pi \in \Pi$: train on (X_{train}, y_{train}); score on (X_{test}, y_{test}).

The nested evaluation avoids the pitfall where test samples inadvertently influence labeling or feature extraction. The nearest-centroid assignment rule ensures that test labels are determined solely by train-induced structure, while the fallback mechanisms guarantee that evaluation remains feasible even when consensus clustering degenerates. Reporting metrics (accuracy, balanced accuracy, macro-

F1, AUC) captures complementary aspects of out-of-sample discriminatory consistency under class imbalance. As labels are induced exclusively within training folds, reported metrics should be interpreted as measures of out-of-sample stability and consistency of the discovered structure, rather than as predictive performance against an exogenous or ground-truth target. This protocol yields performance estimates that are robust (resistant to small-sample issues), valid (free of leakage), and interpretable (aligned with train-only induced structure).

3.7. Explanation Plane – Fold-Local Interpretability

A key requirement of the proposed workflow is that interpretability be derived without contaminating evaluation. To this end, Block F implements fold-local feature selection and explanation, ensuring that (i) features are selected solely within training folds, (ii) explanatory signals are computed only on held-out test partitions, and (iii) outputs are aggregated across folds for robust, leakage-free insights. This design prevents common pitfalls in high-dimensional ESG data, where overfitting and feature leakage can otherwise lead to spurious explanatory narratives (Cawley & Talbot, 2010). The goal is to obtain a Top-15 feature set (5 per silo – E, S and G), fit a compact predictive model on training data, and compute SHAP values exclusively on held-out test points. The algorithm is described in Algorithm 5.

Algorithm 5: Group-Sparse Stability Top-15 (Per-Fold)

Input: (X_{train}, y_{train}), silo map g, repetitions R
Output: Top-15 feature set (5 per silo).

1 For $r = 1 \ldots R$ subsample rows; fit group lasso (or per-silo L1); update per-feature selection counts.
2 For each silo g:
  a. rank features by selection frequency; lock top-m anchors above threshold.
  b. from remaining pool, run ensemble mRMR to fill to 5.
3 Concatenate silo lists to obtain 15 features.

Local notation. Let (X_{train}, y_{train}) denote the feature–label pairs in a training fold with silo map $g \epsilon \{\mathcal{E}, \mathcal{S}, \mathcal{G}\}$. For a given silo g, let I_g index its features. Let R denote the number of resampling repetitions, and m the minimum number of "anchor" features to be locked in via stability selection.

Procedure.

1. **Group-sparse stability selection (train only).** Within each silo, rows are repeatedly subsampled (≈70%) across R resampling iterations. On each subsample, we fit either (i) a group lasso with silo-level groupings (when available) or (ii) per-silo L1-penalized logistic classifiers with a lightweight inner cross-validation for penalty tuning. Selection frequencies are computed per feature, and Wilson confidence intervals are used to identify stable anchors. For each silo, the top- m features above a conservative stability threshold (e.g., $p \geq 0.10$) are locked as anchors. This guarantees that highly reproducible features are preserved across folds, reducing instability in explanatory outcomes.

2. **Ensemble mRMR.** From the remaining stable pool in each silo, the set is expanded to exactly 5 features per silo using an ensemble of minimum-redundancy maximum-relevance (mRMR) runs. Here, mutual information with the induced labels is maximized while redundancy is penalized via absolute correlation $|\rho|$. Multiple seeds are used to prevent dependence on initialization.
3. **Held-out SHAP attribution.** Random forest (RF) is trained on X_{train} restricted to the Top-15 features. Explanations are then computed on X_{test} using SHAP values. By construction, no test statistic influences feature selection, model training, or parameter tuning. For aggregation, we compute the mean absolute SHAP value per feature across splits, and summarize at the silo level (E/S/G) to reveal contribution patterns.

This fold-local interpretability protocol enforces three safeguards:

- Stability through repeated resampling and group-sparse penalties, mitigating the volatility of small-n regimes.
- Balance via the 5-per-silo rule, ensuring that E, S, and G contribute comparably rather than being dominated by a single silo.
- Validity by restricting SHAP attribution to held-out tests, thus preserving the separation of training and evaluation.

The resulting explanations are robust to small samples and comparable across folds, making them suitable for cross-study generalization. By design, Block F yields fine-grained feature-level insights (via SHAP) and silo-level interpretability (via balanced Top-15 selection), bridging the gap between predictive modeling and ESG domain interpretability.

4. Data and findings

4.1. Dataset Description

The empirical analysis is conducted on a 2024 ESG disclosure dataset of Chinese listed firms, compiled through systematic integration of publicly reported ESG indicators. The dataset is structured at the firm-year level, covering 57 unique firms with 61 mapped ESG variables across the three canonical silos: E: 15; S: 25; and G: 21. The silo allocation follows established ESG taxonomies, with E variables emphasizing emissions and resource use, S variables focused on workforce and welfare indicators, and G variables capturing board composition, independence, and turnover dynamics (Table 1).

Table 1. Dataset composition by silo.

Silo	No. of Variables	% of Total	Key Themes
E	15	24.6%	Emissions, energy use, water, resource efficiency
S	25	41.0%	Workforce, insurance, welfare, R&D intensity
G	21	34.4%	Board composition, independence, turnover
Total	61	100%	

Note: The dataset is balanced across silos, with S variables representing the largest share.

The dataset exhibits several defining properties. First, E indicators show very high dispersion (average MAD ≈ 1.05M), reflecting the wide reporting range of emissions and energy use. Second, S indicators display moderate variability (average MAD ≈ 4.5K), driven by workforce and insurance heterogeneity. Third, G indicators are comparatively discrete (median number of unique values = 16), reflecting bounded institutional attributes such as board counts and independence ratios. All silos achieve full non-missing coverage (1.0).

For redundancy and multicollinearity, we compute the pairwise Spearman correlation matrix across all 61 ESG variables. Figure 2 presents the correlation heatmap. Several robust observations emerge:

- Strong intra-silo collinearity. Bright diagonal blocks ($|\rho| > 0.9$) highlight redundancy within silos, such as R&D expenditure ratios in the S silo and director turnover indicators in the G silo. This validates the correlation-based redundancy pruning rule ($\tau = 0.92$) in Block A.
- Weaker cross-silo association. Cross-block intensities are lower and more diffuse, suggesting that E, S, and G silos encode complementary information. This supports the multi-view design of Block B, where per-silo PCA compresses local structure before fusion.
- Silo-specific heterogeneity. E variables exhibit balanced correlations across emissions, energy, and water metrics; S variables show localized redundancy clusters around welfare and R&D; and G variables present concentrated blocks in ownership and board composition.

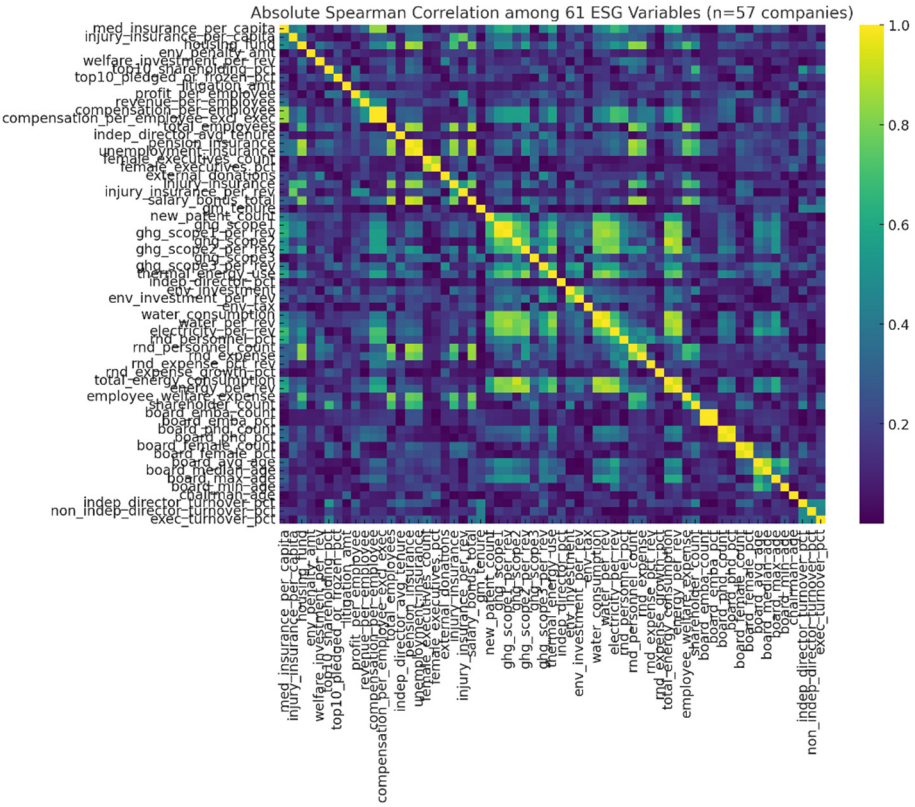

Figure 2. Spearman correlation heatmap of ESG variables (n = 61).

Figure 2 indicates strong within-silo, weak cross-silo collinearity, and high silo correlation heterogeneity.

The dataset's properties motivate the pipeline design:

- Small-n, high-p regime. With 57 firms and 61 variables, naive modeling risks overfitting, supporting the use of dimensionality reduction and consensus discovery.
- Heterogeneous completeness. The presence of nearly complete anchors alongside sparse measures justifies the dual-stage evaluation guard and fold-local feature selection.
- Redundancy within silos. Intra-silo correlations highlight the need for within-silo de-duplication and balanced per-silo PCA, avoiding dominance by any one silo.
- Cross-silo complementarity. The weaker inter-silo correlations justify a multi-view latent space rather than collapsing all variables into a single pool. This supports interpretability and fairness across E, S, and G.

4.2. Discovery plane (Blocks B-C) and evaluation-feasibility guard (Block D)

The discovery stage implements multi-view dimensionality reduction (Block B) followed by evidence-accumulation consensus clustering with automatic K selection (Block C). PAC criterion is used to evaluate candidate resolutions, $K \in \{2,3,4\}$. Across all K, PAC values remain uniformly high, and the procedure selects the smallest $K = 2$. However, even so, the consensus labels collapse to a single dominant class ($n = 57$), indicating weak separability of ESG indicators in a small-n, high-p regime.

To safeguard against degenerate clustering, the evaluation-feasibility guard (Block D) is triggered. This guard replaces global consensus labels with quantile-stratified base splits constructed along the first principal component (PC1) of the standardized feature space. By stratifying entities into PC1 quantile bins, the procedure distributes samples more evenly across splits while maintaining strict separation between training and test folds. Out of 20 candidate cross-validation splits generated by this guard, 6 splits (30%) remain bi-class feasible after train-only label induction; 11 fail because the induced train labels are too small/imbalanced, and 3 fail because the test is single-class after assignment. This process ensures that subsequent evaluation proceeds only under conditions where stratification is meaningful, preserving the statistical informativeness of the test folds and the methodological integrity of the nested design.

4.3. Nested evaluation results (Block E)

Nested evaluation is conducted within each guard-generated split under strict leakage-free conditions: (i) Training folds are standardized and reduced in dimensions (jitter of 10^{-9} is added to break ties in standardized features and has no measurable effect on eigenvalues, loadings, or downstream performance; PCA cap $\leq$ 10 comps); (ii) labels are induced only on training data using the consensus clustering procedure or fallback schemes (with fallbacks: $K = 2$ average-linkage; else PC1 split in the train latent); (iii) test labels are assigned by nearest centroids in the latent training space; and (iv) classifiers are fit solely on training data and scored on held-out test sets.

Primary findings (6 validated test folds). Low-dimensional linear discriminants perform best under strict nesting. DR-classifier pipelines achieve consistently high discrimination. The strongest performance is observed for LDA applied to low-dimensional embeddings. A 1-NN baseline in standardized space is included for reference.

- *NCA-LDA (2D):* Accuracy = 0.9889 ± 0.0272, Balanced Accuracy = 0.9167, AUC = 0.9643.
- *PCA-LDA (2D):* Accuracy = 0.9889 ± 0.0272, Balanced Accuracy = 0.9167, AUC = 0.9762.
- *MultiView-LDA (2D):* Accuracy = 0.9778 ± 0.0344, Balanced Accuracy = 0.8333, AUC = 0.9405.
- Reference *1-NN* baseline: Accuracy = 0.933 ± 0.038.

By contrast, more complex variants (e.g., bagging ensembles, SVM classifiers, or graph-augmented embeddings) deliver comparable raw accuracy but consistently underperform in balanced accuracy (as low as 0.50–0.66). These results demonstrate that simple low-dimensional linear embeddings generalize more reliably than more elaborate non-linear designs when evaluated under strict leakage-free nesting.

Table 2. Nested evaluation results.

Design	MeanAcc	StdAcc	MeanBAcc	MeanF1m	MeanAUC
NCA-LDA	**0.988889**	2.72E-02	**0.916667**	**0.913793**	0.964286
PCA-LDA	**0.988889**	2.72E-02	**0.916667**	**0.913793**	0.97619
MultiView-LDA	0.977778	3.44E-02	0.833333	0.827586	0.940476
MultiView-Bagging	0.955556	3.44E-02	0.666667	0.655172	0.690476
MultiView-SVM	0.944444	2.72E-02	0.583333	0.568966	0.833333
NCA-Graph-LDA	0.944444	2.72E-02	0.583333	0.568966	0.97619
PCA-Graph-LDA	0.944444	2.72E-02	0.583333	0.568966	0.964286
NCA-Bagging	0.933333	4.22E-02	0.577381	0.565887	**0.988095**
NCA-Graph-SVM	0.933333	1.22E-16	0.500000	0.482759	0.904762
NCA-SVM	0.933333	1.22E-16	0.500000	0.482759	0.904762
PCA-Graph-SVM	0.933333	1.22E-16	0.500000	0.482759	0.940476
PCA-SVM	0.933333	1.22E-16	0.500000	0.482759	0.940476
NCA-Graph-Bagging	0.922222	2.72E-02	0.494048	0.47968	0.97619
PCA-Bagging	0.922222	**6.55E-02**	0.571429	0.56258	0.869048
PCA-Graph-Bagging	0.922222	2.72E-02	0.494048	0.47968	0.839286
LDA-Bagging	0.911111	1.56E-01	0.720238	0.72342	0.720238
LDA-Graph-Bagging	0.911111	1.56E-01	0.720238	0.72342	0.750000
LDA-LDA	0.911111	1.56E-01	0.720238	0.72342	0.714286
LDA-Graph-LDA	0.888889	1.44E-01	0.553571	0.551006	0.553571
LDA-SVM	0.888889	1.44E-01	0.553571	0.551006	0.857143
LDA-Graph-SVM	0.877778	1.36E-01	0.470238	0.464799	0.833333

Note: The table reports leakage-free nested cross-validation results across 20 base splits, of which 6 are bi-class feasible. Performance is summarized by mean accuracy, standard deviation of accuracy, balanced accuracy, macro-F1, and AUC. The results demonstrate that low-dimensional embeddings with linear discriminants dominate more complex ensemble and kernel-based methods.

4.4. Stability across discovery draws (Blocks C–E)

A central concern in unsupervised learning pipelines is the sensitivity of clustering outcomes to random initialization, subsampling, or feature bagging. To assess robustness, the discovery process (Blocks C-E) is repeated under multiple random seeds, with consensus clustering, dimensionality reduction, and nested evaluation re-executed in each draw. To test sensitivity to discovery randomness, we repeat the discovery-to-evaluation pipeline under multiple seeds: re-sampling methods/ks, feature-bagging (≥ 12 or 70% of features), row subsampling ($\phi = 0.8$), and refitting the train-only

transforms/labels inside each split. We summarize label stability via ARI/NMI across draws and predictive stability via between-draw/within-draw dispersion of held-out accuracy.

Discovery variability *vs.* predictive reproducibility. As expected in small-n, high-p regimes, raw consensus partitions exhibit variability across draws, reflected in relatively low stability indices (Adjusted Rand Index, ARI ≈ 0.09 ± 0.18; Normalized Mutual Information, NMI ≈ 0.06 ± 0.08). Such dispersion is consistent with the literature on high-dimensional clustering, where weakly separated structures often yield unstable partitions. Crucially, however, downstream predictive evaluation remains stable despite this discovery-level variability.

We empirically test two combinations of relatively top performing DR-classifier pipelines: *Top 15: PCA/NCA-LDA* (fold-local top 15 only, group-balanced selection of 5 E + 5 S + 5 G) and *Base: PCA/NCA-LDA* (full-feature set) (Table 3). Both dominate accuracy and balanced accuracy under strict nesting and remain stable across seeds. Performance metrics aggregated across draws reveal narrow between-draw standard deviations in accuracy (≤ 0.034), with within-draw SD absorbed (≤ 0.059) effectively by the evaluation-feasibility guard and nested design.

Table 3. Stability across discovery draws under multiple seeds.

Design	Mean Accuracy	Between-draw SD	Avg within-draw SD
Top 15: PCA-LDA	**0.943**	**0.020**	**0.045**
Base: NCA-LDA	**0.933**	**0.019**	**0.056**
Top 15: NCA-LDA	0.922	**0.020**	0.058
Base: PCA-LDA	0.900	0.034	0.059

Table reports mean accuracy across seeds, with between-draw and average within-draw standard deviations.

Despite low ARI/NMI at the clustering stage, the guard and nested framework yields reproducible predictive performance. The highlighted pipelines bracket our design choices (fold-local interpretability vs. full-feature supervised DR) and remain stable across seeds, supporting credible inference under stringent leakage controls.

4.5. Interpretability (Block F)

Block F provides fold-local, leakage-averse explanations that answer *which original ESG variables drive the discovered regimes*, without altering or contaminating the performance evaluation in Blocks D-E. The design mirrors our evaluation hygiene, where all selection and model fitting for explanation occur inside the training fold only, and explanations are computed only on the held-out test fold from the same split.

Protocol (per guard-validated split).

1. **Train-only variable shortlisting (balanced top 15).** On the training fold, we standardize features and select exactly 5 variables per silo (E/S/G) to ensure balanced coverage and to control capacity in small-n settings. Selection proceeds in two steps:
   (i) Group-sparse stability selection, where we repeatedly subsample ~70% of training rows and fit either a group-aware penalized model (group lasso when available) or silo-wise L1-logistic baselines, tallying per-feature selection frequencies and Wilson-score confidence intervals;

(ii) Ensemble mRMR fill, where we complete the set to 5 per silo using a small ensemble of mRMR runs that maximizes mutual information with the train-only labels while minimizing redundancy via $|\rho|$. The output is a top 15 list (5 E + 5 S + 5 G) for this fold.

2. **Train-only surrogate for attribution.** Still on the training fold, we fit a compact random-forest surrogate using only the fold's top 15 features. We use a tree-based surrogate because TreeSHAP provides fast, theoretically consistent additive attributions for tree ensembles, yielding stable, local explanations in tabular data. The surrogate is diagnostic only and does not replace or influence the PCA-LDA/NCA-LDA performance pipelines.
3. **Test-only SHAP.** We compute SHAP values on the held-out test fold using the train-fit surrogate. This ensures that explanations summarize out-of-sample behavior of the fold-local explanatory model, preserving the same leakage discipline as the main evaluation. We aggregate $\text{mean}|\text{SHAP}|$ by feature and by silo within fold, then summarize across folds.

Group-level contributions. Across validated splits, the group-level contribution ordering based on aggregated $\text{mean}|\text{SHAP}|$ is S > E > G (Table 4). This hierarchy is consistent across folds and robust to subsampling in the stability-selection stage.

Table 4. Group-level SHAP contributions.

Group	Total Absolute SHAP	Relative Weight	Rank
S	4.82E-10	0.446	1
E	4.09E-10	0.379	2
G	1.89E-10	0.175	3

The table reports absolute and normalized group-level SHAP contributions.

Feature-level insights. At the feature level:

- Social (dominant): Workforce and innovation intensity variables contribute the most influence. Drivers include *R&D expense growth rate*, *R&D intensity*, and *employee welfare/insurance* constructs (e.g., salary/bonus and medical insurance measures).
- Environmental (secondary): Energy intensity dominates (e.g., *electricity_per_rev*, *thermal_energy_use*), with policy/cost signals such as *environmental tax* offering consistent, directional support.
- Governance (modest): Ownership concentration and board structure carry most of the governance signal. Top 10 shareholding (%) and shareholder count dominate, with board PhD count as a secondary driver. Diversity and independence measures (e.g., independent director %, female %) are directionally stable but of smaller magnitude.

These findings are visualized in the test-only top 15 and per-silo Top 5 SHAP charts (Figure 3a, 3b, 3c, 3d), which jointly convey the magnitude and direction of effects.

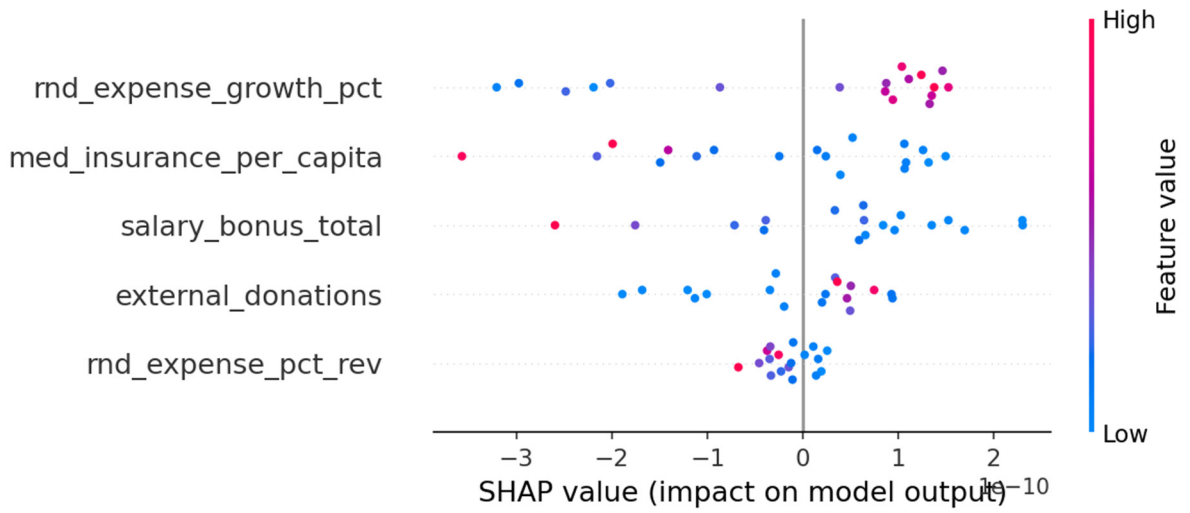

Figure 3a. Social (S) Top 5. R&D and workforce–benefit variables (e.g., rnd_expense_growth_pct, med_insurance_per_capita, and salary_bonus_total) exhibit the largest influence.

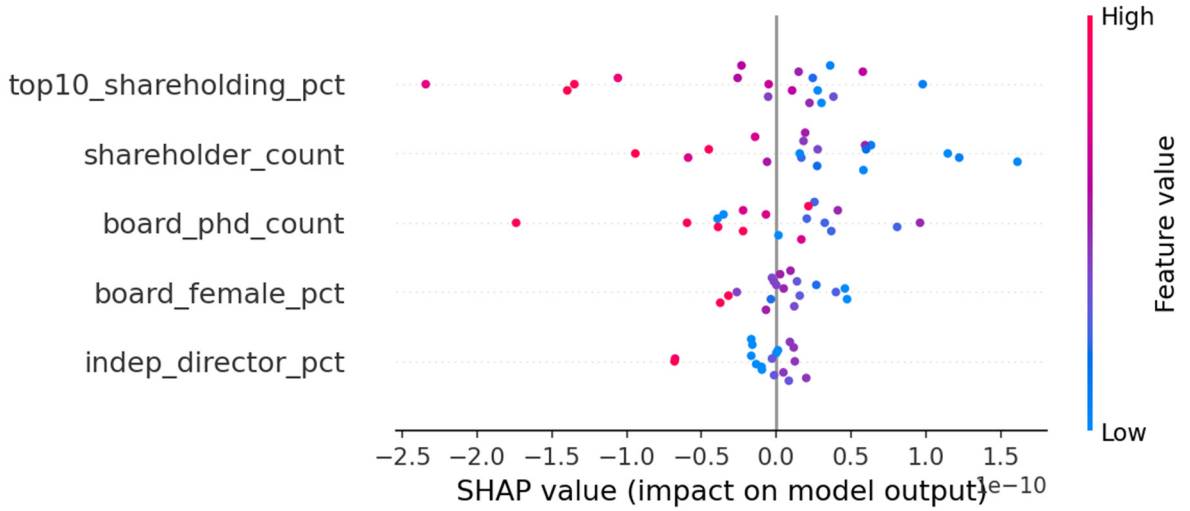

Figure 3b. Environmental (E) Top 5. Energy intensity indicators (notably electricity_per_rev and thermal_energy_use) dominate Environmental influence, with smaller but directionally consistent contributions from ghg_scope3_per_rev and env_tax.

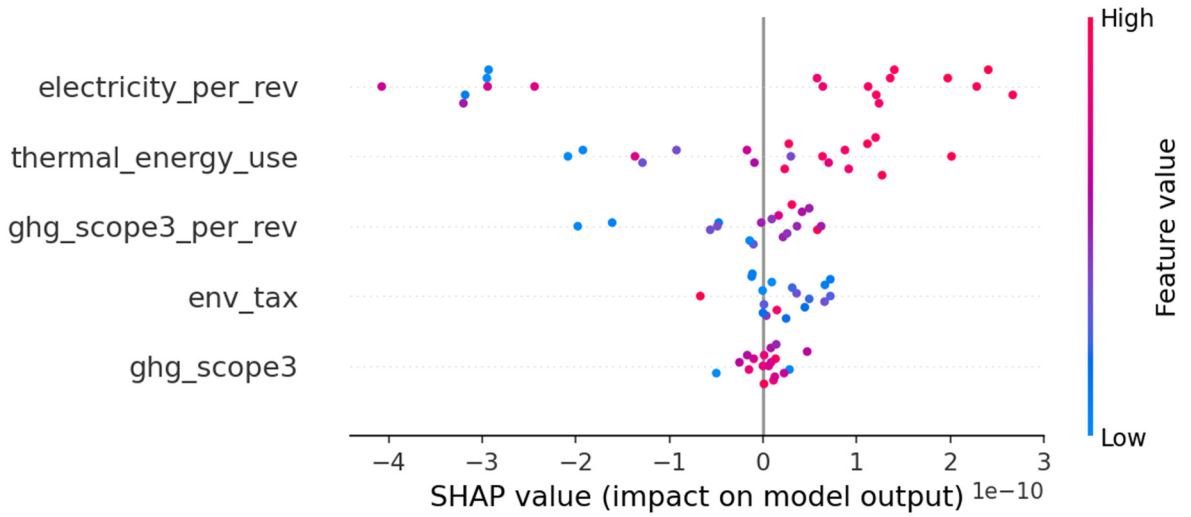

Figure 3c. Governance (G) Top 5. Ownership concentration and board structure as the principal Governance drivers (top10_shareholding_pct, shareholder_count, board_phd_count), while diversity proxies (board_female_pct, indep_director_pct) contribute smaller effects.

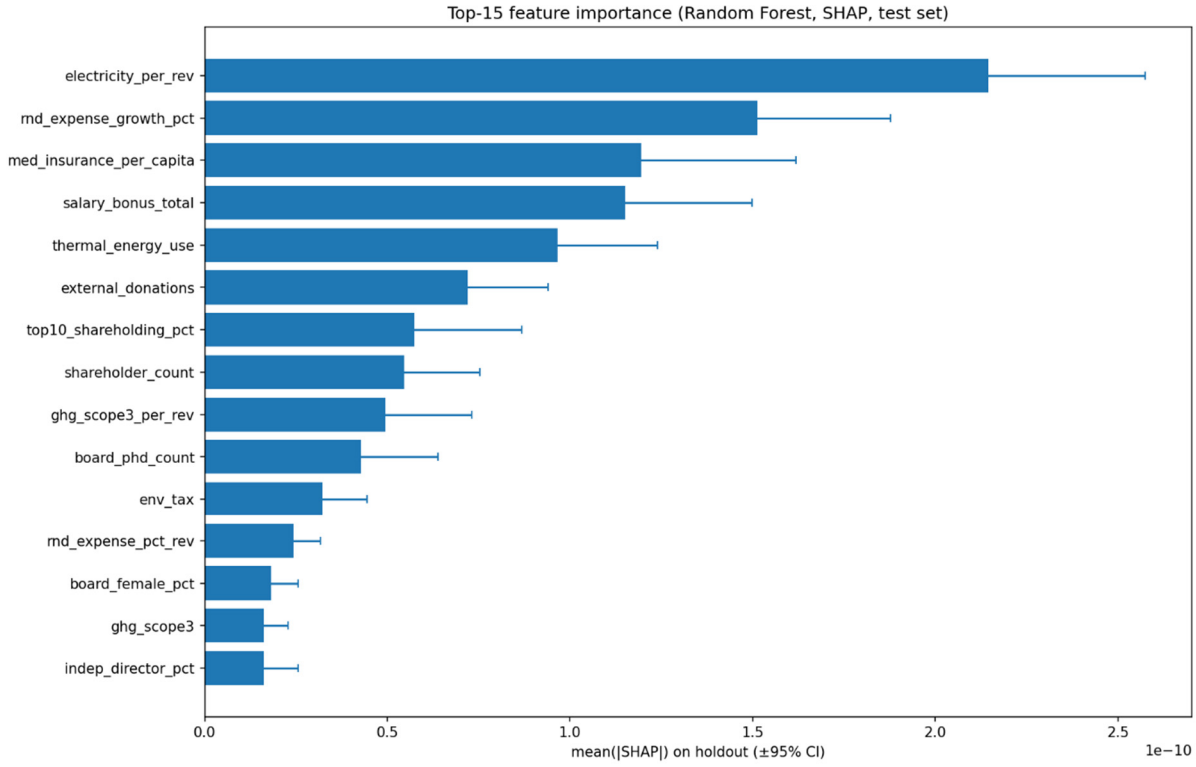

Figure 3d. Global Top 15 (5 E, 5 S, 5 G): Test-only mean|SHAP| with 95% Cis. Bars report the mean absolute SHAP aggregated over all test instances and splits, providing a direction-agnostic measure of influence; whiskers show ±95% normal-approximation CIs across splits. The ranking corroborates the beeswarm panels: electricity_per_rev (E) and Social investment variables (rnd_expense_growth_pct, med_insurance_per_capita, salary_bonus_total) are the most influential test-time predictors, followed by Governance concentration (top10_shareholding_pct, shareholder_count).

Validity and robustness checks.

(ii) Leakage control, where all selection and surrogate fitting are train-only. SHAP is computed only for the held-out test. No statistic from the test informs selection, model fitting, or thresholds.

(iii) Capacity control, where the top 15 (5 per silo) constraint and tree-count limits keep the surrogate's effective capacity aligned with small-n conditions, avoiding overfitting issues.

(iv) Stability selection uses repeated subsampling with Wilson intervals. Ensemble mRMR step explicitly penalizes redundancy, improving feature-set reproducibility and reducing dependence on any single indicator.

(v) Explanations are designed to be parallel to the evaluation pipeline (same folds, same labels, train/test segregation). They serve to interpret the discovered separation in the original ESG space, while the PCA-LDA/NCA-LDA results remain the sole basis for performance claims.

The fold-local, test-only SHAP analysis demonstrates that the dominant separating signals, under strict nested evaluation, arise from human-capital and innovation (S), followed by energy intensity and environmental cost (E), with ownership/board structure (G) providing complementary but smaller influence. Because explanations are computed only on held-out data and features are selected within the fold, the narrative remains evaluation-clean, silo-balanced, and reproducible, supporting auditor-grade traceability from latent structure to concrete ESG indicators.

4.6. Summary findings

This study delivers a practical, auditable pipeline for working with small-n, high-p ESG panels: Guarded discovery to avoid degenerate labels, leak-free nested evaluation to obtain out-of-sample performance, and fold-local explanations to map latent separation back to concrete indicators. In summary:

(i) Discovery with guardrails preserves evaluability. When the unsupervised signal was weak, the protocol did not force unreliable labels. Instead, the evaluation-feasibility guard produced stratified base splits and performed train-only labeling within each split. This kept test folds informative while maintaining strict separation between discovery and evaluation.

(ii) Low-capacity, low-dimensional models are sufficient and reliable. Under nested, train-only labeling, 2-D embeddings with linear discriminants (NCA-LDA, PCA-LDA) delivered the highest balanced accuracy and AUC with tight fold-level dispersion, outperforming higher-capacity baselines whose balanced accuracy degraded under the same bi-class tests. This indicates that the discriminative structure, once properly de-noised and leak-controlled, is essentially linear in a compact latent space.

(iii) Explanations are clean, silo-balanced, and reproducible. Per split, we selected an exactly 5-per-silo feature set on the training fold only, fit a compact surrogate on the same training fold, and computed test-only SHAP on the hold-out. Aggregating across validated splits yielded an attribution hierarchy S > E > G, with influential variables that had clear semantics for index construction (e.g., *electricity_per_rev* in E, *rnd_expense_growth_pct* in S, *top10_shareholding_pct* in G). This ordering reflects the statistical salience of disclosed

indicators within the analyzed reporting panel and may serve as an interpretive signal to support ESG evaluation and assurance processes.

Summarized in Table 5, the approach produced a reproducible, interpretable, and leakage-averse pipeline suitable for ESG index design in constrained data regimes. It turned a noisy, redundancy-rich disclosure matrix into (a) defensible labels for evaluation, (b) robust, low-variance generalization estimates, and (c) transparent, silo-balanced drivers that can be traced end-to-end.

Table 5. Pipeline stages, principal results, and supporting evidence. This table summarizes the data set and the pipeline stages.

Pipeline Block	Method / Step	Key Result(s)	Section (§); Supporting Table(s) & Figure(s)
Blocks A-B: Dataset; Data Hygiene & Pruning; Multi-View Consensus Space	Data: 57 firms, 61 features (E=15, S=25, G=21). Pruning: drop quasi-constants and de-duplicate within silos. Two-stage multi-view learning to form Z.	Balanced coverage; strong within-silo redundancy and weak cross-silo association, which motivates pruning + multi-view; Z preserves complementary E/S/G structure for discovery.	• §3.2-3.3 (methodology) • Algorithm 1 • §4.1 (results) • Table 1 • Figure 2
Block C: Evidence accumulation + Auto-K (PAC)	PAC on co-association over $\mathcal{K} = \{2,3,4\}$; average-linkage with size repair.	Flat PAC across K, resulting in Auto- $K = 2$; labels collapsed after repair (weak global separation).	• §3.4 (methodology) • Algorithm 2 • §4.2 (results)
Block D: Evaluation-feasibility guard	Quantile-stratified PC1 binning; candidate splits; train-only labeling per split.	Guard restored evaluability: 20 base splits prepared, 6 validated (30%) bi-class tests.	• §3.5 (methodology) • Algorithm 3 • §4.2 (results)
Block E: Nested evaluation (primary)	Train-only labeling; nearest-centroid test assignment; empirical testing on DR-classifier pipelines.	Best performing DR-classifier pipeline {NCA-LDA, PCA-LDA}: Acc = 0.9889 ± 0.0272; BAcc = 0.9167; AUC = 0.964-0.976. Baseline 1-NN: Acc = 0.933 ± 0.038.	• §3.6 (methodology) • Algorithm 4 • §4.3 (results) • Table 2
Blocks C-E: Stability testing	Multi-seed discovery; re-evaluation per draw.	Reproducible performance: Between-draw SD ≤ 0.020; within-draw SD ≈ 0.045-0.059.	• §3.4-3.6 (methodology) • Algorithms 2-4 • §4.4 (results) • Table 3
Block F: Fold-local interpretability	Group-sparse stability selection + ensemble mRMR (exactly 5 E / 5 S / 5 G per fold); Train-fold surrogate RF; test-only SHAP.	Group ordering S > E > G; salient features include *electricity_per_rev, rnd_expense_growth_pct, med_insurance_per_capita.*	• §3.7 (methodology) • Algorithm 5 • §4.5 (results) • Table 4 • Figures 3a–d

5. Implications

5.1. Implications for System and application designers

The empirical results motivate a design principle that separates discovery, evaluation, and explanation into a 3-plane pipeline, which is useful for controlling leakage and variance in a small-n, high-p regime. In practice, evidence-accumulation with Auto-K establishes whether structure exists at all. When global separation is weak, the feasibility guard converts discovery noise into stratified,

auditable valid test folds. Subsequent train-only labeling and nearest-centroid assignment ensure that all downstream scores are computed in a space learned exclusively from training data. Finally, explanations are produced in strict parallel. Features are shortlisted within the training fold, while SHAP is evaluated only on held-out observations, so interpretability does not contaminate generalization measurement. This layering yields a system that is reproducible by construction and whose failure modes (e.g., single-class tests, unstable clustering) are intercepted before they can affect reported performance.

Across validated folds, 2D embeddings with linear discriminants (PCA-LDA and NCA-LDA) achieved the strongest balanced accuracy and AUC, while higher-capacity alternatives (RBF-SVM, bagging, graph-augmented variants) offered no systematic gains under bi-class tests and often exhibited degraded class balance. The inference is that, once pruning and standardization remove redundancy and scale artifacts, the discriminative signal is largely second-order and well captured by low-dimensional linear structure. For system builders, this translates into a default scoring stack that is computationally light, stable under resampling, and amenable to transparent monitoring; more complex learners are best relegated to sensitivity analysis rather than production scoring in similarly constrained settings.

5.2. Implications for ESG analysis, governance, and audit

The empirical results translate into actionable design principles for ESG analytics in settings characterized by sparse, noisy disclosures and stringent model-risk governance, with direct relevance to analysts, auditors, regulators, and governance committees. They include:

- The discovery to evaluation workflow converts a disclosure matrix into *defensible*, train-only pseudo-labels and low-variance generalization estimates. This enables ESG issuer segmentation and shortlist formation *prior* to the availability of any supervised target (ESG ratings, ESG controversies, excess returns). Because label induction and transformation are restricted to the training fold and test labels are assigned by train-space centroids, the resulting segments are not due to leakage, and can therefore be used as stable scaffolds for downstream supervised tasks or ESG analyst review.
- Group-level attribution hierarchy $S > E > G$ reflects the relative statistical salience of disclosed indicators within the analyzed reporting panel and can inform field surveys, vendor selection, and quality-control prioritization when resources are constrained. Further, as low-capacity 2D linear discriminants dominate balanced accuracy under bi-class validation, feature acquisition and compute can be trimmed without measurable loss. This is advantageous for quarterly rebalances where ESG disclosure lags and batch windows are short.
- Train-only labeling combined with test-only SHAP enables counterfactual "what-if" audits (e.g., perturb electricity intensity or R&D spend) and inspection of issuer movement in the latent space without retraining on test data. This scaffolding supports exploratory analyses for regulator and issuer dialogues on policy levers and transition pathways. Moreover, separation of discovery-prediction-explanation planes creates a transparent audit chain from latent separation to indicators. Fold-wise logs (train/test indices, train-fit transforms and centroids, top 15 selections, test-only SHAP) satisfy explainability of ESG reporting requirements.

5.3. Implications for research

Here are several implications for research:

- Empirically, 2-D embeddings coupled with linear discriminants (PCA-LDA, NCA-LDA) dominate balanced accuracy, with negligible advantage to higher-capacity kernels and ensembles. This pattern is consistent with a second-order view of separability after standardization: Most discriminative signal appears to be expressible through low-rank covariance structure once redundancy is pruned. Two research directions follow:
  i. Derive bounds for linear separability and generalization under correlation-pruned designs (e.g., $|\rho| \leq 0.92$) and dimensionality caps (e.g., $d = 2$), quantifying when linear decision rules saturate Bayes-optimal performance in small-n regimes.
  ii. Develop tests that decide, from train-fold statistics alone, whether nonlinear hypothesis classes can yield robust gains over LDA given the observed eigen-spectrum and class-conditional covariance geometry.
- Auto-K via PAC was flat and $K = 2$ collapsed after minimum-mass repair, indicating weak global separation in this panel. For larger universes or richer disclosure, multi-regime structure may emerge. Methodologically, extending auto-K with (a) stability-adjusted PAC, (b) explicit minimum-mass constraints, and (c) multiscale co-association (coarse-to-fine evidence accumulation) could enable principled selection of $K > 2$ without over-fragmentation. A related avenue is label-budgeted discovery, where the algorithm selects K subject to a constraint on the smallest testable class under stratified splits.
- The fold-local, test-only SHAP design yields leakage-clean *associations* between variables and predicted regimes. A natural next step is to fuse this with causal structure, e.g., invariant risk minimization (IRM) across folds/seeds, proximal causal inference with disclosure proxies, or SHAP constrained by graphical priors. The objective is to separate correlational salience from policy-relevant effects while preserving the strict train/test segregation that underwrites the validity of current explanations.
- The results are cross-sectional. Extending to rolling windows would enable the assessment of temporal stability under disclosure drift. Promising directions include discovery-time domain adaptation (train-fold alignment penalties across vintages) and distribution-shift detection tied to the PAC/feasibility diagnostics (e.g., triggering re-discovery only when co-association entropy or test-feasibility rates degrade).
- Finally, the train-only labels produced here can serve as scaffolds for semi-supervised or weak-supervised learning when partial outcomes (e.g., controversies and third-party scores) become available. Researchers may examine co-training schemes that treat consensus labels as priors, update them with sparse supervision, and prove that leakage control is preserved end-to-end, maintaining the methodological integrity that underpins these findings.

6. Limitations and threats to validity

6.1. Scope and design limitations

The feasibility guard retains only bi-class–feasible tests. Consequently, summary statistics are computed on 6 of 20 base splits. This prioritizes metric validity (balanced classes in test) over nominal coverage and prevents incorrect scores on single-class tests. While this may exclude edge cases produced by highly imbalanced discovery, the exclusion is ex ante and rule-based, not performance-driven, and is reported alongside results (feasibility is itself a KPI). In practice, this renders the claims conservative and interpretable.

In this panel, PAC is flat over $K \in \{2,3,4\}$, and minimum-mass repair yields an effective single dominant class pre-guard. Evaluation therefore proceeds via guard-generated base splits with train-only labeling per split. Richer universes (larger n, denser signal) may support $K > 2$. The same framework extends directly once PAC endorses higher K under minimum-mass constraints; critically, these performance estimates do not depend on assuming a particular global K.

ESG indicators vary in scale, discretization, and reporting incentives. Block-A pruning (e.g., $nunique$, $|\rho|$) reduces redundancy and quasi-constants; per-silo standardization and the multi-view latent prevent any silo from dominating. While such controls cannot eliminate all latent reporting bias, the train-only fitting and test-only scoring/explanation ensure that any residual heterogeneity cannot leak into the performance estimates, measuring true out-of-sample variability.

6.2. Threats to validity

By design, the reported classification metrics in this study assess the reproducibility and out-of-sample stability of unsupervised structure under strict evaluation hygiene, rather than predictive performance against an exogenous ground-truth target. This construct-validity framing reflects the absence of universally accepted ESG labels and clarifies that accuracy, balanced accuracy, and AUC are interpreted as indicators of structural generalization rather than outcome prediction.

The empirical contribution hierarchy ($S > E > G$) and the top 15 features are estimated on a single ESG panel and are therefore panel-specific. While empirical findings depend on the dataset, the proposed pipeline architecture is designed to be portable across ESG panels. Portability in this study refers to the design of the workflow, including the separation of discovery, prediction, and explanation planes, rather than to empirical generalization from a single panel. To support reproducibility, we release operational thresholds (e.g., $nunique$, $|\rho|$ for pruning), PAC bands for Auto-K, and split generators, enabling roll-forward re-discovery and replication. In deployment, portability is monitored via (i) feasibility rate (share of bi-class test folds) and (ii) selection stability (per-silo Jaccard forop 15). Threshold alarms trigger automatic re-pruning and re-discovery, ensuring adaptation without changing the evaluation hygiene.

Further, ESG policy shifts can alter scales, sparsity, and discretization, affecting separability. To counteract this threat, the pipeline re-induces labels inside each training fold and evaluates strictly out-of-sample, so temporal drift appears as changed train-fold geometry rather than leaked signal in test metrics. Operationally, we track (i) feasibility rate, (ii) balanced accuracy with confidence intervals,

and (iii) Top 15 stability as production KPIs. Sustained degradation triggers re-tuning of Block-A pruning and Block-B multi-view compression, followed by re-discovery. Drift therefore results in explicit re-estimation under the same leakage controls.

SHAP values quantify predictive salience, not causal effects. To counteract this threat, we interpret SHAP as associational and test-only under strict train/test segregation. For policy inference, we recommend leakage-clean counterfactual stress tests (e.g., perturb electricity intensity or R&D spend on held-out data and examine movements in the latent space and decision margins). Where appropriate, invariance penalties or causal regularizers can be layered without altering the guard/nesting scaffold, preserving evaluation hygiene while improving construct validity.

Future datasets may exhibit stronger nonlinearity, making 2-D LDA sub-optimal. To counteract this threat, higher-capacity learners (RBF-SVM, ensembles) are maintained as sensitivity models and are evaluated under the same nested protocol. Promotion to production requires dominance in balanced accuracy/AUC with comparable fold/seed dispersion and unchanged feasibility. Because discovery, evaluation, and explanation layers are model-agnostic, upgrading the scorer does not modify leakage controls, audit trail, or interpretability flow.

With $n = 57$ and $p = 61$, fine-grained structure may be under-powered. To counteract this threat, the design bounds variance via dimensionality reduction emphasizes balanced accuracy and AUC, and reports fold-level dispersion and between-seed variability. The feasibility guard rejects non-informative tests ex ante, and claims are framed with uncertainty summaries. As n grows, the same protocol yields tighter intervals, making improvements immediately auditable.

Last, mixed scales and discretization can potentially inflate redundancy. To counteract this threat, Block-A removes quasi-constants and de-duplicates within silos at $|\rho| \geq 0.92$; Block-B standardizes per silo and uses multi-view PCA to prevent dominance by any single silo. All transforms are fit on training data only, and tests are scored and explained out-of-sample.

7. Conclusions

In this study, we propose and evaluate an end-to-end 3-plane pipeline for ESG discovery, evaluation, and explanation designed for small-n, high-p disclosure matrices. The core contribution is an architecture that (i) prevents degenerate unsupervised outcomes from contaminating validation, (ii) delivers reliable out-of-sample estimates under train-only label induction, and (iii) provides auditable, silo-balanced attributions that map latent separation to ESG indicators.

On a 2024 ESG panel (57 firms, 61 variables), the discovery layer combined multi-view dimensionality reduction with evidence-accumulation consensus clustering and an Auto-K rule based on PAC. When global consensus was under-separated, the evaluation-feasibility guard replaced global labels with quantile-stratified base splits, and labeling was re-induced inside each training fold. This design preserved test informativeness while strictly isolating discovery from evaluation. Under this regime, compact 2-D embeddings with linear discriminants (PCA-LDA, NCA-LDA) achieved the strongest balanced accuracy and AUC with low dispersion, outperforming higher-capacity baselines whose balanced accuracy deteriorated under bi-class tests. Stability checks across discovery draws showed that predictive performance remained reproducible (tight between-draw standard deviations)

even when raw consensus partitions varied. This is a frequent characteristic of weakly separable, high-dimensional clustering.

The explanation layer was aligned with the same hygiene, such that per split, we learned exactly 5 features per silo (E/S/G) on the training fold only via group-sparse stability selection and ensemble mRMR. We then fit a compact, train-fold surrogate and computed test-only TreeSHAP. Aggregating across folds yielded an attribution ordering $S > E > G$: human-capital and innovation proxies dominated, followed by energy intensity, with ownership/board structure providing complementary but smaller signals.

Methodologically, the results support two key conclusions. First, in small-sample ESG panels, parsimonious linear models in low-dimensional spaces are sufficient and more stable than complex kernels/ensembles when labels are induced and evaluated under strict nesting. Second, guardrails (PAC-driven Auto-K, feasibility checks, rejection of single-class tests) form necessary architectural elements that prevent silent failure modes and enable valid out-of-sample estimation.

Limitations are explicit and mitigated. Reporting is confined to validated bi-class splits for metric interpretability; small-n uncertainty is quantified via fold/seed dispersion; and surrogates are used only for explanation, leaving performance claims grounded in the PCA/NCA-LDA stack. While the empirical ordering $S > E > G$ is panel-specific, the architecture generalizes to other sparse corporate disclosure regimes (e.g., supply-chain KPIs and sustainability-linked debt), where leakage risk is high and labels are scarce. The attribution ordering reflects the statistical salience of disclosed indicators within the analyzed panel and is intended as an interpretive signal rather than a direct assessment of sustainability quality or impact. In practice, such signals may be used to support exploratory analysis and review alongside established ESG evaluation and assurance processes.

In future studies, researchers should extend the framework along four axes: (i) Multi-regime structure ($K > 2$) using stability-adjusted PAC and minimum-mass constraints on larger universes; (ii) temporal robustness, with rolling re-discovery and drift diagnostics; (iii) causal augmentation, pairing test-only SHAP with counterfactual stress tests and invariance-based regularization; and (iv) benchmarks and open protocols to standardize pruning thresholds, PAC settings, and fold generators for reproducible cross-study comparisons.

This work demonstrates that doing less, carefully, can produce more reliable and explainable ESG analytics. The result is a reproducible, interpretable, and leakage-averse pathway from noisy ESG disclosures to deployable analytics pipeline design.

Author contributions

The authors declare to have contributed equally to the manuscript. All authors have read and approved the final manuscript.

Use of AI tools declaration

The authors declare that AI tools were used solely to improve the language and clarity of the manuscript. All scientific work, tables, and findings were created and validated exclusively by the authors.

Acknowledgment.

We thank Ms. Ann Guo from Jiran Think Tank for your input on initial data preprocessing, and Mr. Qian Zhong Chao from Qinglv for providing access to the dataset used in this study.

Declaration of funding.

No funding was received.

Conflict of interest.

The authors declare no conflict of interest

Availability of data and materials.

The data that support the findings of this study were obtained from QingLv (https://www.i-esg.com/) under a data-use agreement. Due to licensing and confidentiality restrictions, the underlying ESG dataset cannot be made publicly available. The manuscript provides full methodological details, parameter settings, and evaluation protocols to enable reproducibility using comparable ESG disclosure data.

References

Adebayo J, Gilmer J, Muelly M, et al. (2018) Sanity checks for saliency maps. *Adv Neur Inf Proces Syst* 31.

Barber RF, Candès EJ (2015) Controlling the false discovery rate via knockoffs. *Ann Stat* 43: 2055–2085. Available from: https://www.jstor.org/stable/43818570.

Berg F, Kölbel J F, Rigobon R. (2022) Aggregate confusion: The divergence of ESG ratings. *Rev Financ* 26: 1315–1344. https://doi.org/10.1093/rof/rfac033

Billio M, Fitzpatrick AC, Latino C, et al. (2024) *Unpacking the ESG ratings: Does one size fit all?* (No. 415). SAFE Working Paper. https://doi.org/10.2139/ssrn.4742445

Bishop CM (2006) *Pattern Recognition and Machine Learning*. Springer.

Blank H, Sgambati G, Truelson Z (2016) Best practices in ESG investing. *J Invest* 25: 103–112. https://doi.org/10.3905/joi.2016.25.2.103

Candes E, Fan Y, Janson L, et al. (2018) Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection. *J R Stat Soc B* 80: 551–577. https://doi.org/10.1111/rssb.12265

Cawley GC, Talbot NL (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 11: 2079–2107. Available from: https://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf.

Chen Y, Yang Y (2021) The one standard error rule for model selection: Does it work? *Stats* 4: 868–892. https://doi.org/10.3390/stats4040051

Cover T, Hart P (1967) Nearest neighbor pattern classification. *Ieee T Inform Theory* 13: 21–27. https://doi.org/10.1109/TIT.1967.1053964

Covert I, Lundberg S, Lee SI (2021) Explaining by removing: A unified framework for model explanation. *J Mach Learn Res* 22: 1–90. Available from: https://www.jmlr.org/papers/v22/20-1316.html.

Demartini P, Pagliei C (2023) Can we trust ESG ratings? Some insights based on a bibliometric analysis of ESG data quality and rating reliability. *Manag Control*, *Supplemento* 2: 161–187. Available from: https://www.torrossa.com/en/resources/an/5657265.

Dormann CF, Elith J, Bacher S, et al. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36: 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://doi.org/10.48550/arXiv.1702.08608

Fred AL, Jain AK (2005) Combining multiple clusterings using evidence accumulation. *Ieee T Pattern Anal Machi Intell* 27: 835–850. https://doi.org/10.1109/TPAMI.2005.113

Golalipour K, Akbari E, Hamidi SS, et al. (2021) From clustering to clustering ensemble selection: A review. *Eng Appl Artif Intell* 104: 104388. https://doi.org/10.1016/j.engappai.2021.104388

Goldberger J, Hinton GE, Roweis S, et al. (2004) Neighbourhood components analysis. *Adv Neural Inform Process Syst* 17. Available from: https://proceedings.neurips.cc/paper_files/paper/2004/file/42fe880812925e520249e808937738d2-Paper.pdf.

Guignard F, Ginsbourger D, Levy Häner L, et al. (2024) Some combinatorics of data leakage induced by clusters. *Stoch Env Res Risk Assess* 38: 2815–2828. https://doi.org/10.1007/s00477-024-02715-1

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *J Mach Learn Res* 1157–1182. https://www.jmlr.org/papers/v3/guyon03a.html

Hao Z, Lu Z, Li G, et al. (2023) Ensemble clustering with attentional representation. *Ieee T Knowl Data Eng* 36: 581–593. https://doi.org/10.1109/TKDE.2023.3292573

Hofner B, Boccuto L, Göker M (2015) Controlling false discoveries in high-dimensional situations: boosting with stability selection. *BMC Bioinformatics* 16: 144. https://doi.org/10.1186/s12859-015-0575-3

Jolliffe IT, Cadima J (2016) Principal component analysis: A review and recent developments. *Philos T R Soc A* 374: 20150202. https://doi.org/10.1098/rsta.2015.0202

Kaufman S, Rosset S, Perlich C, et al. (2012) Leakage in data mining: Formulation, detection, and avoidance. *Acm T Knowl Discov Data* 6: 1–21. https://doi.org/10.1145/2382577.2382579

Kapoor S, Cantrell EM, Peng K, et al. (2024) REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Sci Adv* 10: eadk3452. https://doi.org/10.1126/sciadv.adk3452

Kapoor S, Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4: 100804. https://doi.org/10.1016/j.patter.2023.100804

Klaaßen L, Lohmüller C, Steffen B (2024) Assessing corporate climate action: Corporate climate policies and company-level emission reductions. *PLOS Clim* 3: e0000458. https://doi.org/10.1371/journal.pclm.0000458

Kotsantonis S, Serafeim G (2019) Four things no one will tell you about ESG data. *J Appl Corp Financ* 31: 50–58. https://doi.org/10.1111/jacf.12346

Kuhn M, Johnson K (2013) Data pre-processing. In: *Applied Predictive Modeling,* 27-59. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-6849-3_3

Lim T (2024) Environmental, social, and governance (ESG) and artificial intelligence in finance: State-of-the-art and research takeaways. *Artif Intell Rev* 57: 76. https://doi.org/10.1007/s10462-024-10708-3

Lim M, Hastie T (2015) Learning interactions via hierarchical group-lasso regularization. *J Comput Graph Stat* 24: 627–654. https://doi.org/10.1080/10618600.2014.938812

Liu H, Chen J, Dy J, et al. (2023) Transforming complex problems into K-means solutions. *Ieee T Pattern Anal Mach Intell* 45: 9149–9168. https://doi.org/10.1109/TPAMI.2023.3237667

Liu T, Yu H, Blair RH (2022) Stability estimation for unsupervised clustering: A review. *Wires Comput Stat* 14: e1575. https://doi.org/10.1002/wics.1575

Lock EF, Hoadley KA, Marron JS, et al. (2013) Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat,* 7: 523–542. https://doi.org/10.1214/12-AOAS597

Lucińska M, Wierzchoń ST (2012) Spectral clustering based on k-nearest neighbor graph. In: *IFIP International Conference on Computer Information Systems and Industrial Management,* 254–265. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-33260-9_22

Lundberg SM, Erion G, Chen H, et al. (2020) From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2: 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg SM, Erion GG, Lee SI (2018) Consistent individualized feature attribution for tree ensembles. *arXiv preprint*. https://doi.org/10.48550/arXiv.1802.03888

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 30. Available from: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Meinshausen N, Bühlmann P (2010) Stability selection. *J R Stat Soc B* 72: 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

Monti S, Tamayo P, Mesirov J, et al. (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach learn* 52: 91–118. https://doi.org/10.1023/A:1023949509487

Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Ieee T Pattern Anal Mach Intell* 27: 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Qu L, Pei Y (2024) A comprehensive review on discriminant analysis for addressing challenges of class-level limitations, small sample size, and robustness. *Processes* 12: 1382. https://doi.org/10.3390/pr12071382

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 20: 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1: 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sasse L, Nicolaisen-Sobesky E, Dukart J, et al. (2025) Overview of leakage scenarios in supervised machine learning. *J Big Data* 12: 135. https://doi.org/10.1186/s40537-025-01193-8

Şenbabaoğlu Y, Michailidis G, Li JZ (2014) Critical limitations of consensus clustering in class discovery. *Sci Rep* 4: 6207. https://doi.org/10.1038/srep06207

Simon N, Friedman J, Hastie T, et al. (2013) A sparse-group lasso. *J Comput Graph Stat* 22: 231–245. https://doi.org/10.1080/10618600.2012.681250

Smilde AK, Bro R, Geladi P (2003) *Multi-Way Analysis: Applications in the Chemical Sciences*. Wiley.

Smilde AK, Westerhuis JA, de Jong S (2003) A framework for sequential multiblock component methods. *J Chemometr* 17: 323–337. https://doi.org/10.1002/cem.811

Tao Z, Liu H, Li S, et al. (2017) From ensemble clustering to multi-view clustering. In: C. Sierra (Ed.), *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17),* 2843–2849. AAAI Press. https://dl.acm.org/doi/10.5555/3172077.3172285

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc B* 63: 411–423. https://doi.org/10.1111/1467-9868.00293

Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7: 91. https://doi.org/10.1186/1471-2105-7-91

Varoquaux G (2018) Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage* 180: 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Wang C, Cao L, Miao B (2013) Optimal feature selection for sparse linear discriminant analysis and its applications in gene expression data. *Comput Stat Data Anal* 66: 140–149. https://doi.org/10.1016/j.csda.2013.04.003

Westerhuis JA, Kourti T, MacGregor JF (1998) Analysis of multiblock and hierarchical PCA and PLS models. *J Chemometr* 12: 301–321. https://doi.org/10.1002/(SICI)1099-128X(199809/10)12:5%3C301::AID-CEM515%3E3.0.CO;2-S

Xu C, Tao D, Xu C (2013) A survey on multi-view learning. *arXiv preprint*. https://doi.org/10.48550/arXiv.1304.5634

Yu Z, Dong Z, Yu C, et al. (2025) A review on multi-view learning. *Front Comput Sci* 19: 197334. https://doi.org/10.1007/s11704-024-40004-w

Zhao X, Niu X, Ma Y, et al. (2024) A multi-view ensemble clustering approach using joint entropy. *Expert Syst Appl* 255: 124683. https://doi.org/10.1016/j.eswa.2024.124683

Zhou S, Xu H, Zheng Z, et al. (2024) A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. *Acm Comput Surv* 57: 1–38. https://doi.org/10.1145/3689036

AIMS Press