



Anchored Targeted Projection Pursuit for Hypothesis-Driven Quality Control of Drill-Cuttings Geochemical Data

Yupeng Li^{1,*}, Saeed S. Shahrani² and Zhaoxian Yuan³

¹ Beijing Research Center, Upstream Advanced Research Center, Saudi Aramco, Beijing 102200, China

² Geological Operations Department, EXPEC, Saudi Aramco, Dhahran, Saudi Arabia

³ Hebei Key Laboratory of Strategic Critical Mineral Resources, Hebei GEO University, Shijiazhuang 050031, China

Abstract

Wellsite X-ray fluorescence (XRF) analysis of drill cuttings provides multi-element profiles for chemostratigraphy, well correlation and reservoir characterization, but the data are exposed to artefacts of the drilling workflow, such as drilling-fluid residue, carry-over between samples and detector drift. Quality control (QC) usually relies on principal component analysis (PCA), which ranks samples by variance rather than geological meaning, or on supervised classifiers, which need labels that do not exist at the QC stage. In practice, analysts often suspect a problem and can point to a few samples that show it. We introduce anchored targeted projection pursuit (anchored TPP), which builds its projection target from these anchors alone and leaves all other samples unconstrained. The fit reduces to a weighted orthogonal Procrustes problem with a closed-form solution and no tuning parameters. It remains

well posed when variables outnumber samples, and its loadings name the elemental contrast behind each flag. We test it on a synthetic XRF benchmark and on a semi-synthetic experiment that plants a potassium chloride (KCl) brine mud signature into real cuttings data from an unconventional hydrocarbon well. Anchored TPP detects contamination that PCA and single-element thresholds miss, and log-ratio pre-treatment more than doubles its recovery. Positive-unlabelled learning recovers marginally more (0.55 of 12 hidden samples; 95% confidence interval 0.07–1.03), but anchored TPP is interpretable, tuning-free and fast enough for interactive use. Because all contamination labels are planted, field validation against laboratory-confirmed intervals remains necessary.

Keywords: projection pursuit, drill cuttings, data quality control, chemostratigraphy, reservoir characterization, X-ray fluorescence, drilling-fluid contamination, compositional data, human-in-the-loop analysis.



Submitted: 14 July 2026

Revised: 05 October 2026

Accepted: 06 October 2026

Published: 14 October 2026

Vol. 3, No. 1, 2027.

10.62762/JGEE.2026.403029

*Corresponding author:

✉ Yupeng Li

yupeng.li@aramcoasia.com

Citation

Li, Y., Shahrani, S. S., & Yuan, Z. (2026). Anchored Targeted Projection Pursuit for Hypothesis-Driven Quality Control of Drill-Cuttings Geochemical Data. *Journal of Geo-Energy and Environment*, 3(1), 48–71.



© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Abbreviation	Definition
AR(1)	First-order autoregressive process
CI	Confidence interval
clr	Centred log-ratio
ilr	Isometric log-ratio
LDA	Linear discriminant analysis
LOF	Local outlier factor
MCD	Minimum covariance determinant
ML	Machine learning
PCA	Principal component analysis
PLS-DA	Partial least squares discriminant analysis
PPMT	Projection pursuit multivariate transform
PU	Positive-unlabelled
QC	Quality control
SVM	Support vector machine
<i>t</i> -SNE	<i>t</i> -distributed stochastic neighbour embedding
TPP	Targeted projection pursuit
UMAP	Uniform manifold approximation and projection
XRF	X-ray fluorescence

1 Introduction

The increasing adoption of machine learning (ML) and automated analytical tools in subsurface characterization, from geo-energy reservoir studies to mineral exploration, has raised the importance of geochemical data quality control [1–5]. In conventional deterministic geological workflows, isolated corrupted samples typically affect only a small number of interpretive outputs and are readily identified through manual inspection or basic univariate screening. In data-driven predictive modelling, by contrast, anomalous or contaminated samples distort high-dimensional feature spaces and bias model weights, degrading downstream tasks such as chemostratigraphic correlation, chemofacies and reservoir-quality classification, ore-body delineation and mineral prospectivity modelling [1, 6]. Data quality control (QC) has therefore moved from a peripheral post-processing step to a prerequisite for geochemical modelling and geological interpretation, and the quality-assurance chain that produces geochemical data is now itself an object of systematic study [7].

High-resolution X-ray fluorescence (XRF) core and cuttings scanning has become a standard technique in petroleum-well chemostratigraphy and

mineral exploration, providing dense, multi-element geochemical datasets with 20–30 elemental channels and fine spatial sampling resolution [8–11]. These multivariate measurements support detailed lithological mapping, alteration identification, and stratigraphic correlation [12]. The same instrumentation and the same sample-handling chain serve mineral exploration drill core, mineral prospectivity and chemofacies studies, and petroleum-well cuttings alike; portable XRF is routinely applied to core and cuttings from the same borehole and the two are treated as interchangeable inputs to a single down-hole profile [13].

What varies between these settings is the commodity being sought, not the structure of the measurement. Many XRF datasets contain closed or approximately compositional subsets—normative mineralogy normalized to 100%, or major-element oxides whose total is close to fixed—alongside variables reported on an open scale, such as trace elements in mg kg^{-1} ; closure should therefore be assessed separately for each analytical block rather than assumed for the dataset as a whole (Section 3.4). The artefacts that corrupt these data—drilling-fluid residue, carry-over between samples, surface coatings, detector drift—arise from the drilling and scanning workflow rather than from the deposit type; sample preparation, surface condition and matrix effects are documented sources of bias in portable XRF measurement irrespective of the material analysed [14]. Therefore, QC for XRF scan data is common ground across exploration geochemistry, and a method developed for one of these settings applies to the rest. Dedicated and effective QC workflows for XRF geochemical datasets nevertheless remain limited.

The problem is sharpest in geo-energy drilling. Horizontal and unconventional wells are now routinely analysed with wellsite XRF on cuttings, and the resulting profiles support chemostratigraphic correlation, lateral placement and facies interpretation [9, 15]; they support shale-play chemostratigraphic correlation and facies discrimination, including identification of clay-rich, carbonate-rich and organic-matter-rich intervals that inform completion and lateral-placement decisions [16]. Cuttings are also the sample type most directly exposed to the drilling system. They are carried to surface in the drilling fluid, and where hole cleaning is difficult, as in long horizontal sections, they can accumulate in the annulus, mixing material from different depths. The residues of drilling fluid retained on cuttings

surfaces have been shown to introduce systematic geochemical artefacts that compromise subsequent elemental analysis [17]. They carry residues of water-based drilling fluids, commonly KCl or NaCl brines, and fluids returned from shale gas wells are characteristically saline and rich in Na and Cl [18]. The drilling fluid is itself an object of wellsite measurement: mud logging provides continuous monitoring of gas shows, drilling parameters and cuttings descriptions that together characterize the formation being drilled [19]. A cuttings profile contaminated in this way can create a spurious chemostratigraphic boundary, or bias a clay- or carbonate-volume estimate that feeds the reservoir model, so QC of cuttings geochemistry is a direct input to subsurface interpretation rather than a clerical step. Section 6.2 returns to these applications.

Existing QC strategies fall into three established categories, none of which is sufficient on its own: univariate screening, unsupervised multivariate analysis, and supervised machine learning.

Univariate range checks assess elemental concentrations independently against empirical or statistical thresholds. While simple and widely adopted, univariate methods cannot detect geochemically implausible multi-element combinations where all individual elemental values fall within acceptable ranges. This limitation has long been recognized in exploration geochemistry and motivated the shift toward multivariate QC frameworks [20, 21].

Unsupervised multivariate methods integrate full multi-element information to identify data structure. They include principal component analysis (PCA), *k*-means clustering, *t*-distributed stochastic neighbour embedding (*t*-SNE), and uniform manifold approximation and projection (UMAP), and are widely combined with clustering and fractal filters to delineate geochemical anomalies [12, 21–23]. Nevertheless, such techniques optimize purely statistical objective functions, such as global variance or local geometric similarity, without explicit consideration of geological context or analytical artefact signatures. Drilling mud contamination, sample preparation biases, and instrumental drift typically constitute low-variance anomalies superimposed upon dominant high-variance lithological and alteration trends. Subtle QC artefacts are therefore frequently relegated to minor components and remain undetected in standard unsupervised analysis [20].

Supervised multivariate and machine learning methods, such as linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), and supervised classifiers, are powerful for targeted geochemical discrimination and alteration mapping [1]. Their usefulness for QC is limited, however, by their reliance on fully labelled training datasets. Complete labels are not available at the QC stage: if analysts could confidently label every sample as clean or contaminated, formal QC screening would be unnecessary. Partial labelling further introduces circular reasoning, as treating unexamined samples as implicitly clean masks unidentified anomalies and biases model performance.

Labelled samples are scarce throughout exploration geochemistry, and this is a long-recognized obstacle. It has motivated semi-supervised formulations that learn from a small labelled subset alongside a large unlabelled remainder [24], and few-shot frameworks that build a trainable set from a handful of known occurrences [25]. The difficulty is not only that positives are few: genuine negatives are unavailable, and arbitrarily designated negative samples have been shown to change the resulting predictions materially [26].

This study addresses the middle ground those methods leave open. In routine drill-core and cuttings work an analyst is seldom without information, even when labels are absent: there is usually a tentative but physically grounded suspicion—drilling-fluid residue, an unexpected alteration signature—and a few samples confidently interpreted as showing it. These anchor samples are sparse guidance, not a labelling scheme, and this is the ordinary situation rather than a special case. Neither variance-driven unsupervised methods nor label-dependent supervised tools are designed for it.

Targeted projection pursuit (TPP) [27] offers a way into this gap. Unlike conventional projection pursuit, which optimizes generic statistical indices, TPP finds linear projections that best approximate user-specified target configurations, so that domain knowledge guides dimensionality reduction. Although TPP has been applied to gene-expression data, and broader projection pursuit methods are widely applied in remote sensing and hyperspectral mineral mapping [28, 29], interactive knowledge-guided TPP workflows have not been adapted or formalized for geochemical QC or mineral exploration applications.

This work makes three contributions. First, we

formalize *anchored targeted projection pursuit* (*anchored TPP*), a knowledge-guided multivariate QC method that incorporates domain expertise exclusively through analyst-defined anchor samples, without requiring formal class labels. By framing the anchored target as a weighted orthogonal Procrustes alignment, we obtain a classical closed-form solution that provides deterministic, tuning-free model fitting. Second, we analytically characterize the estimator, proving that the anchored cross-product matrix depends on the data solely through the anchor rows. This structural property explains the method's parameter inertness, its $O(mp)$ fitting efficiency for interactive re-use, its stability under high dimensions ($p \gg n$), and its close alignment with the anchor centroid. Third, we evaluate anchored TPP against unsupervised PCA, one-class [30], and positive-unlabelled (PU) [31, 32] learning baselines in a unified benchmark. This controlled comparison shows where the method works and where it does not, including its sensitivity to down-hole spatial autocorrelation during anchor selection.

We evaluate detection performance, robustness to anchor choice, and the effect of compositional log-ratio pre-treatment on two paired benchmarks. The fully synthetic benchmark simulates subtle contamination hidden beneath geological confounding; the semi-synthetic one plants the same signal in the real covariance structure of published drill-cuttings data from the lateral of an unconventional hydrocarbon well [15]. We chose that dataset because it is openly published and carries exactly the down-hole collinearity that defeats univariate screening. Such a combination is rare among wellsite geochemical datasets, most of which are proprietary. The method works on the standardized compositional matrix and is indifferent to the commodity behind the survey, so the KCl-mud scenario here stands in for any multi-element artefact an analyst can exemplify with a few samples.

We acknowledge that all evaluations in this study rely on planted synthetic or semi-synthetic ground truth rather than independently verified field contamination. Reliable quantitative performance assessment requires known ground-truth labels, which are not available for unprocessed field datasets [33, 34]. Although our semi-synthetic design preserves realistic inter-element geochemical relationships, semi-synthetic results cannot fully replicate field operational conditions. We therefore emphasize that field validation with laboratory-confirmed contamination datasets remains

essential for operational deployment and constitutes a clear direction for future methodological refinement and application.

2 Background: Projection Pursuit and Anomaly Detection

Anchored TPP is assembled from two well-established components—the projection pursuit family and the orthogonal Procrustes solution—joined by one new element, the anchored target. This section sets out what is inherited from each and where it falls short for quality control (QC); Section 3 then develops the method itself. Results already available in the literature are recalled with citation rather than re-derived; only the properties specific to the anchored construction are established from scratch.

2.1 Projection Pursuit and Its Geoscience Lineage

Projection pursuit [35, 36] seeks low-dimensional linear projections of high-dimensional data that maximize an index of “interestingness”, most commonly a measure of departure from normality on the grounds that Gaussian projections are uninformative. The approach is well established in the geosciences through the projection pursuit multivariate transform (PPMT) [37], which projects, Gaussianizes and back-transforms multivariate geological data to satisfy the multi-Gaussian assumptions of geostatistical simulation. PPMT is widely used for multivariate transformation and, because variables that resist Gaussianization are often anomalous, has incidental diagnostic value.

The distinction between PPMT and the present work is the source of the objective. PPMT targets a fixed distributional property (multi-Gaussianity) determined by the requirements of downstream simulation. Anchored TPP targets an arbitrary configuration supplied by the analyst, and is therefore applicable to hypotheses that have no distributional expression—“these four samples were touched by the same mud system” is not a statement about skewness or kurtosis. It is this fixed distributional limitation that motivates the move to a user-specified target.

2.2 Targeted Projection Pursuit, and Circular Labelled Targets

Faith et al. [27] turned projection pursuit around: instead of letting the algorithm decide what is interesting, the user places the points where they should fall in the projected view, and the algorithm finds the projection that comes closest. Two variants

were proposed, one based on Procrustes analysis yielding orthogonal projections, and one based on a single-layer perceptron yielding non-orthogonal projections. We adopt the Procrustes variant, because the orthogonality constraint preserves the interpretation of the projection as a rotation of the original variable axes and hence keeps the loading vector physically meaningful.

In the original TPP formulation the target is built from known class labels, which suits the visualization problem it was designed for. Carried straight over to QC, that construction becomes circular. The objection deserves a direct answer. If every sample's known class is used to build the target, then recovering those classes afterwards establishes nothing beyond the sufficiency of the labels: it is discriminant analysis with additional steps. Nor would that be novel. Finding linear combinations of elements that separate known groups is long-standing practice in litho-geochemistry, and discriminant analysis, PLS-DA and supervised classifiers are standard tools for chemofacies work [1, 12].

The anchored construction developed below avoids the objection by a specific route rather than by assertion. Only $m \ll n$ analyst-chosen anchors enter the fit; those anchors are selected by eye and may be wrong; the remaining samples—more than 95% of the dataset in every experiment reported here—are left unconstrained; and class labels are used only after the fact, to score the result. The test that matters is accordingly not whether known anomalies are recovered, but whether a handful of uncertain anchors surfaces samples that were never flagged. Sections 4.1 and 5.6 are designed around that question.

2.3 Unsupervised Anomaly Detection and the Missing Hypothesis

Robust multivariate statistics provide an established alternative, notably the minimum covariance determinant estimator [38] and robust Mahalanobis distances, which flag samples departing from an elliptical consensus. Machine-learning detectors such as isolation forests [39–41] and the local outlier factor [42] relax the ellipticity assumption and scale to large datasets, and have been extended in exploration geochemistry to the PU setting, where a handful of known occurrences is combined with a large unlabelled remainder to score high-dimensional anomalies [43].

All are unsupervised or weakly supervised in

the same sense: they identify samples that are unusual, but do not indicate *in what respect*, and cannot distinguish genuine but unusual geology from corrupted measurement. That distinction requires domain input, and is the specific function that anchored TPP performs. For compositional data, Filzmoser and Hron [20] demonstrate that outlier detection must account for closure; we return to this point in Sections 3.4 and 6.4.

3 Anchored Targeted Projection Pursuit

Section 3.1 states the workflow as the analyst experiences it; Sections 3.2 and 3.3 formalize it and establish the properties that follow from the anchored construction; and Sections 3.4–3.6 cover compositional pre-treatment, the reading of the loading vector, and practice.

3.1 The Anchored Workflow

Anchored TPP hands the definition of “interesting” back to the analyst, and asks for it in the crudest form that still carries information: the analyst drags a small number of samples believed to share a common cause towards one side of a two-dimensional scatter plot. Nothing further is asserted—no threshold, no class definition, no claim about any other sample. The remaining rows sit at the origin of the target with a small weight, which regularizes the least-squares objective. As Section 3.3 shows, neither that weight nor those rows influence the projection actually fitted. The workflow is a cycle rather than a single pass: the analyst marks a few anchors, fits the projection, reads the ranked list and the loading vector, then revises the anchors and refits. Figure 1 shows these steps, with the dashed arrow closing the loop.

Two features of this loop set it apart from one-pass QC. First, a mistaken hypothesis shows itself inside the loop rather than only at the end. If the anchors do not in fact share a common multi-element cause, no projection separates them from the background, the loading vector carries no coherent geochemical meaning, and the ranked list shifts erratically as anchors are added or removed. That instability is diagnostic information rather than noise, and we report it as such in Sections 5.3 and 5.6.4. Second, the loop is cheap enough to be exploratory. Because the fit depends on the data only through the anchor rows (Section 3.3), adding an anchor and refitting costs $O(mp)$ operations, which is what makes drag-and-refit practical on a laptop at the data sizes considered here.

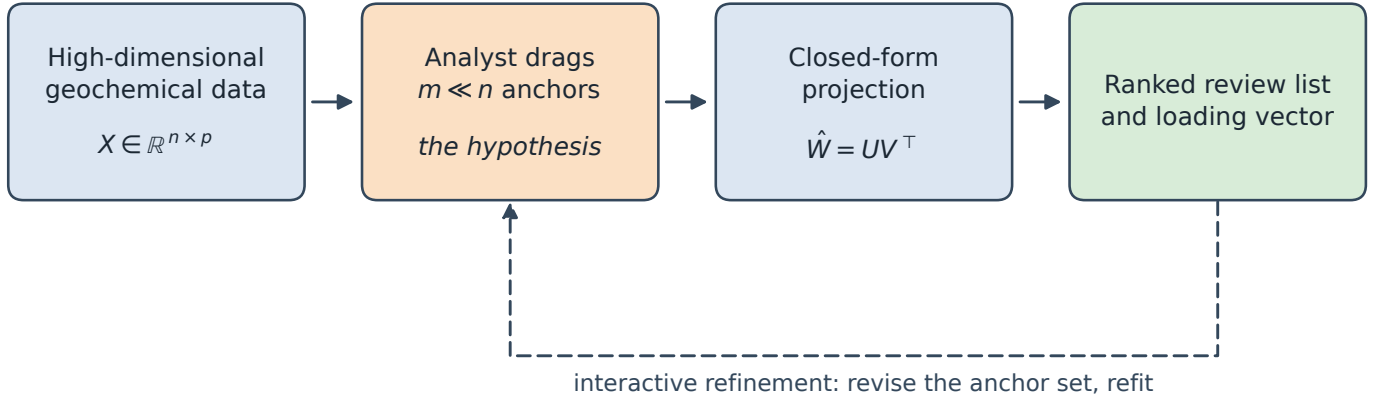


Figure 1. The anchored targeted projection pursuit workflow. The analyst displaces a small anchor set—the only hypothesis supplied, with no class labels and no thresholds—and the closed-form fit returns the projection $\hat{W} = UV^T$, a ranked review list and a loading vector. The dashed arrow closes the loop: revise the anchor set and refit.

3.2 Formulation

Let $X \in \mathbb{R}^{n \times p}$ denote the standardized data matrix for n samples and p variables, and let $x_i \in \mathbb{R}^p$ be the column vector holding the i -th row of X , so that $X^T = [x_1, \dots, x_n]$. The analyst selects a small anchor set $\mathcal{A} \subset \{1, \dots, n\}$, $|\mathcal{A}| = m \ll n$, and assigns each anchor a desired position in the two-dimensional view, collected in the target matrix $T \in \mathbb{R}^{n \times 2}$.

Anchors are displaced a common distance $\rho > 0$ along the first target axis and given small random offset (jitter) ε_i along the second; all remaining rows sit at the origin. Writing $\Omega = \text{diag}(\omega_1, \dots, \omega_n)$,

$$t_i \equiv (T_{i,1}, T_{i,2})^T = \begin{cases} (\rho, \varepsilon_i)^T & i \in \mathcal{A}, \\ (0, 0)^T & \text{otherwise,} \end{cases} \quad (1)$$

$$\omega_i = \begin{cases} 1 & i \in \mathcal{A}, \\ \omega_{\text{bg}} & \text{otherwise,} \end{cases} \quad (2)$$

with $t_i \in \mathbb{R}^2$ the i -th row of T written as a column vector, $\omega_{\text{bg}} \ll 1$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon \ll \rho$. The offset carries no geological meaning. It only keeps T at rank two: without it every anchor would sit at the same point, and the second axis of the projection would be arbitrary. The displacement ρ pulls the anchors away from the background along the first axis, which is the axis used for ranking.

The background weight ω_{bg} acts as a pure regularization term rather than a tunable substantive parameter. We adopt fixed default values $\omega_{\text{bg}} = 0.05$ and $\rho = 4$ throughout all experiments; Section 3.3

shows that neither value influences the fitted projection, so no empirical tuning is required.

We seek the projection $W \in \mathbb{R}^{p \times 2}$, constrained to have orthonormal columns, that brings the projected data into closest weighted agreement with the analyst-defined target configuration. The natural least-squares objective function is formulated as:

$$\hat{W} = \arg \min_{W^T W = I} \|\Omega^{1/2} (XW - T)\|_F^2. \quad (3)$$

Expanding the squared Frobenius norm yields three distinct terms governing the optimization:

$$\begin{aligned} \|\Omega^{1/2} (XW - T)\|_F^2 &= \underbrace{\text{tr}(W^T X^T \Omega X W)}_{\text{dispersion term}} \\ &\quad - 2 \text{tr}(W^T X^T \Omega T) + \text{tr}(T^T \Omega T). \end{aligned} \quad (4)$$

The third term is independent of the projection matrix W and can be excluded from optimization. While the dispersion term varies for rectangular low-rank projections and prevents closed-form minimization of the full objective, retaining only the cross-term that directly couples the projection to the user-defined target yields a solvable, well-posed optimization problem. We therefore adopt the alignment-focused objective as the operative definition for anchored TPP:

$$\hat{W} = \arg \max_{W^T W = I} \text{tr}(W^T X^T \Omega T), \quad (5)$$

which is the weighted orthogonal Procrustes problem in its standard form. Its solution is the orthonormal polar factor of the weighted cross-product matrix between the standardized data and the target configuration [44]:

$$M = X^\top \Omega T, \quad M = U \Sigma V^\top, \quad \hat{W} = UV^\top. \quad (6)$$

Projected sample coordinates are computed as $Y = X\hat{W}$. The solution is exact for rectangular orthonormal projections and requires no initialization, iteration or convergence control. Most importantly, no class label information enters the formulation or the fit at any stage.

3.3 Structural Properties of the Anchored Solution

Two properties of Equation (6) are classical and are recalled here rather than proved. First, the maximizer of $\text{tr}(W^\top M)$ over $W^\top W = I$ is the orthonormal polar factor $UV^\top$ of M [44]; the result follows from Von Neumann's trace inequality and holds for rectangular W . Second, that polar factor is invariant under $M \mapsto cM$ for any $c > 0$, since positive rescaling changes the singular values of M but not its singular vectors [45]. Neither result is specific to the present method.

What *is* specific to the anchored construction is the structure of M itself, and the practical behaviour of the estimator follows almost entirely from it.

Proposition 1 (Anchor reduction). *Let T and Ω be given by Equations (1) and (2). Then the weighted cross-product matrix satisfies*

$$M = X^\top \Omega T = \sum_{i \in \mathcal{A}} x_i t_i^\top,$$

so that M depends neither on ω_{bg} nor on the background rows $\{x_i : i \notin \mathcal{A}\}$.

Proof. Written row-wise, $M = \sum_{i=1}^n \omega_i x_i t_i^\top$, a sum of $p \times 2$ outer products. For $i \notin \mathcal{A}$ Equation (1) sets $t_i = 0$, so those terms vanish whatever value $\omega_i = \omega_{\text{bg}}$ takes; for $i \in \mathcal{A}$, Equation (2) gives $\omega_i = 1$. $\square$

Three consequences follow—inertness of the nominal hyper-parameters, locality of the fit, and proximity of the primary axis to the anchor centroid—and together they account for most of the method's operational character.

(i) **The nominal hyper-parameters are inert.** The background weight ω_{bg} is absent from M and therefore from $\hat{W}$: it regularizes the least-squares objective of Equation (3) but is structurally inoperative in the alignment solution actually used. The pull magnitude enters only through the scale of M , so by the recalled invariance any common rescaling $(\rho, \sigma_\varepsilon) \mapsto (c\rho, c\sigma_\varepsilon)$ leaves $\hat{W}$ unchanged; only the ratio ρ/σ_ε has any effect. That effect falls mainly on the secondary visualization axis. It does not vanish on the primary axis, because the polar factor couples the two columns of M , but it becomes negligible when $\sigma_\varepsilon \ll \rho$, which is the regime used throughout. Numerically, varying ω_{bg} over four orders of magnitude (10^{-4} to 1), and rescaling ρ together with σ_ε over three (10^{-1} to 10^2), changes the fitted projection by less than 10^{-15} , which is machine precision. The method has no hyper-parameters to tune, and reporting a tuned value for either would be misleading.

(ii) **The fit is local to the anchors.** Because M is a sum of m rank-one terms, fitting costs $O(mp)$ operations via the singular value decomposition of a $p \times 2$ matrix, which is what makes the interactive loop of Figure 1 viable. No covariance matrix is formed or inverted, so the estimator remains well posed when $p \gg n$, a regime in which conventional robust multivariate estimators fail without heavy regularization. A minimum of two anchors is required for a rank-two target and hence a two-dimensional view; a single anchor yields a valid primary ranking axis with a degenerate second axis. Note that locality applies to the *fit* only: the projected scores $Y = X\hat{W}$, and hence the ranking, depend on every row of X .

(iii) **The primary axis is close to the anchor centroid.** Across all benchmark replicates the first column of $\hat{W}$ aligns with the normalized centroid of the standardized anchors to $\cos \theta = 0.9998 \pm 0.0002$. In effect, the primary axis is a normalized anchor-centroid direction. This explains both its anchor economy and its inherited weaknesses—sensitivity to outlying anchors, and an inability to represent multi-modal target populations (Section 6.4). We state this plainly because it bounds the claim: the contribution of anchored TPP is not a novel direction estimator, but the anchored target construction, the orthonormal projection framework built on it, and the interactive, parameter-free workflow that the closed form makes possible.

The substitution of Equation (5) for Equation (3) has a cost. Direct minimization of the full least-squares

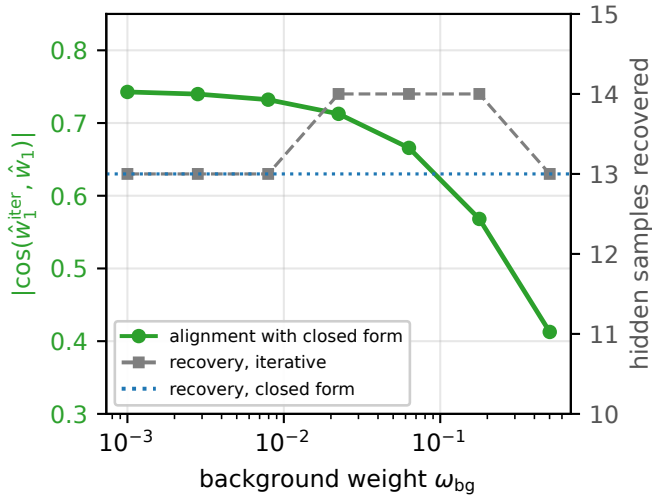


Figure 2. Cost of the objective substitution. Alignment between the leading axis of the iterative least-squares solution and that of the closed form (left axis, green), and hidden samples recovered by each (right axis, grey and dotted blue), as the background weight ω_{bg} is varied. The iterative axis drifts steadily away from the closed form while recovery stays flat, so detection alone does not reveal the divergence. $m = 20$ anchors; a single synthetic realization, which recovers 13 of 15 rather than the 11 of Table 2 because the random offset on the secondary target axis is drawn with a different seed.

objective by projected gradient descent on the Stiefel manifold attains lower residual error, but reintroduces explicit dependence on ω_{bg} and yields no consistent detection gain. Figure 2 shows what the dependence costs: as ω_{bg} rises from 10^{-3} to 0.5, the leading iterative axis drifts away from the closed-form axis, their alignment falling from 0.74 to 0.41, while recovery stays between 13 and 14 of 15 throughout. The iterative solution therefore finds a materially different projection that scores about the same, and detection alone does not reveal the difference. We retain the closed-form alignment solution for its stability, speed and freedom from tuning, while noting that it optimizes a related but not identical criterion.

3.4 Compositional Pre-Treatment

XRF data are often compositional, at least in part, and where a constant-sum constraint applies it makes raw covariances partly artefactual. The log-ratio transformation before multivariate analysis is long-established practice in exploration geochemistry [20, 46, 47]. We follow it, and record here only the choices specific to this workflow, because they turn out to matter more than the choice of detection algorithm (Section 5.6.2). We apply the centred log-ratio (clr) transform,

$$\text{clr}(z) = \left(\log \frac{z_1}{g(z)}, \dots, \log \frac{z_D}{g(z)} \right), \quad (7)$$

$$g(z) = \left(\prod_{j=1}^D z_j \right)^{1/D},$$

where $g(z)$ is the sample-wise geometric mean, to each closed subcomposition separately: a dataset containing more than one closed system is transformed one system at a time, since pooling them would form ratios across a boundary that carries no physical meaning. Closure is a property of an analytical block, not of the dataset, and is assessed block by block. A block is treated as compositional when its parts are normalized to a fixed total, as normative mineralogy is, or when its parts are major constituents whose total is approximately fixed, as major-element oxides are; an approximately closed block is analysed as a subcomposition, whose information lies in the ratios among its parts, which log-ratios preserve whether or not the parts sum exactly to a constant. A block is treated as open when its values are far from any sum constraint, as trace elements reported in mg kg^{-1} usually are. Variables reported on an open scale are log-transformed instead, to reduce skew. Zeros, which the logarithm cannot take, are handled by multiplicative replacement, which preserves the ratios among the remaining parts [48]. Standardization follows.

3.5 Interpretation of the Loading Vector

Because $\hat{W}$ has orthonormal columns, the projection rotates the standardized variable axes rather than stretching or shearing them, so each entry of a loading vector can be read directly as a weight on a measured variable. A black-box detector, which returns only an anomaly score, offers nothing comparable.

Each column of $\hat{W}$ is a unit vector in the original variable coordinate system. Entries in the primary column quantify the relative contribution of each elemental variable to the targeted contrast, and both magnitude and sign carry diagnostic meaning: positive loadings indicate variables enriched in the anchor group, while negative loadings mark variables whose elevation suppresses alignment with it. This signed structure is what allows co-occurring geological and anthropogenic processes with overlapping elemental signatures to be separated, and Section 5 gives a worked instance in which a small negative loading does the discriminating work.

One consequence of the clr transform deserves emphasis, because it changes how every loading vector reported on transformed data must be read. The entries of $\hat{W}$ are then weights on log-ratios relative to the geometric mean of the (sub)composition, not on concentrations. The same definition also forces the clr components to sum to zero, so loadings are not independent: adding mass to one part necessarily depresses the log-ratios of the others, and a scatter of small negative loadings across unrelated variables is an artefact of closure rather than a set of findings. Section 5.6 gives a concrete instance: the expected strong negative loading on alumina does not materialize, for exactly this reason. The caution applies only where clr has been used. Loadings from the fully synthetic benchmark, which is generated in open rather than compositional space, and loadings on open-scale variables may be read directly [49].

3.6 Algorithm and Interactive Use

The full anchored TPP pipeline, including pre-processing, projection fitting and anomaly ranking, is summarized in Algorithm 1. It supports both standalone batch analysis and the iterative refinement loop of Figure 1.

Algorithm 1: Anchored targeted projection pursuit.

Input: Raw data $Z \in \mathbb{R}^{n \times p}$; anchor set $\mathcal{A}$; pull magnitude ρ ; background weight ω_{bg}

Output: Projection $\hat{W}$; projected coordinates Y ; ranked review list

Apply compositional pre-treatment (Equation (7)) to closed subcompositions;

Standardize column-wise to obtain X ;

Initialize $T \leftarrow 0_{n \times 2}$, $\omega \leftarrow \omega_{bg} \mathbf{1}_n$;

foreach $i \in \mathcal{A}$ **do**

$T_{i,1} \leftarrow \rho$; $T_{i,2} \leftarrow \varepsilon_i$ (small random offset);
 $\omega_i \leftarrow 1$;

end

$M \leftarrow X^\top \text{diag}(\omega) T$;

$(U, \Sigma, V^\top) \leftarrow \text{svd}(M)$;

$\hat{W} \leftarrow UV^\top$; $Y \leftarrow X\hat{W}$;

Rank all $i \notin \mathcal{A}$ by $Y_{i,1}$ in descending order;

return top- k as the review list;

In practical exploratory QC the analyst inspects the projected distribution, revises the anchor set by adding newly surfaced candidates or removing points that prove inconsistent, and refits immediately. This adaptive, knowledge-driven screening is the practical

distinction from static one-pass QC methods, whether unsupervised or supervised.

4 Benchmark Design and Evaluation Protocol

Two benchmarks are used, and they are complementary rather than alternative. The fully synthetic benchmark (Section 4.1) buys exact ground truth at the cost of a simplified covariance structure; the semi-synthetic benchmark (Section 4.2) recovers realistic collinearity by planting a signal in real drill-cuttings data, at the cost of a contamination signature that is still artificial. Both are scored under the single review-list protocol of Section 4.3, so results from the two are directly comparable.

4.1 Fully Synthetic Benchmark

The synthetic benchmark is built so that the difficulties which defeat routine QC are present by construction: a contamination signature subtle enough to escape visual inspection, a geogenic confounder that reproduces part of it, and lithological variance large enough to bury both.

4.1.1 Data Generation

We simulate a dataset consisting of $n = 800$ samples and $p = 25$ elemental variables, mimicking high-resolution XRF core scanning data acquired from fine-grained clastic sedimentary successions. Bulk compositional variability is governed by three independent latent geological factors sampled from standard normal distributions, corresponding to dominant sedimentary and diagenetic processes:

- **Detrital sediment input:** controls concentrations of lithogenic elements including Si, Ti, Al, K, Zr, Rb, and Nb;
- **Carbonate and biogenic productivity:** governs Ca, Sr, Mg, and Mn abundances, and is partially anti-correlated with detrital input to simulate sediment dilution effects;
- **Redox environmental conditions:** modulates redox-sensitive elements including Fe, V, Mo, U, Ni, Cu, Zn, and Cr.

Additional natural background variability is incorporated for Ba and S, with Ba coupled to detrital signals and S linked to redox fluctuations, ensuring neither element acts as an exclusive tracer of the artificially planted contamination signature. Independent Gaussian noise is superimposed on all elemental channels to simulate instrumental

measurement uncertainty and residual geological heterogeneity.

4.1.2 Planted Populations and the Confounder

Three distinct subpopulations with anthropogenic and geogenic signatures are artificially embedded within the background dataset (Table 1) to simulate realistic QC challenges. Drilling-mud contamination, representative of common KCl-based mud systems with barite weighting agents, jointly elevates Ba, K, Cl, and Sr concentrations across affected samples. This contaminated population is further divided into two subgroups: a *visible contamination group* with full-amplitude elemental anomalies corresponding to visually identifiable core staining, and a *hidden contamination group* with identical multi-element signatures scaled to 60% amplitude, representing subtle contamination undetectable via visual core inspection. The low-amplitude hidden subgroup serves as the primary target for quantitative detection evaluation.

A geogenic confounding population consisting of natural barite nodules is additionally introduced, characterized by elevated Ba and S concentrations without anomalous enrichment in Cl or K. This confounder is explicitly designed to defeat simplistic single-element threshold strategies, ensuring benchmark evaluation strictly tests multi-element signature discrimination rather than trivial univariate screening.

4.2 Semi-Synthetic Data from Real Drill Cuttings

The fully synthetic benchmark provides exact ground-truth control but relies on artificially simplified covariance structures constructed from orthogonal latent factors. Real geochemical datasets exhibit far more complex collinearity, where elemental variations are simultaneously modulated by mineral composition, diagenetic alteration, and grain-size sorting. Methods performing well on idealized synthetic data may fail to generalize to realistic geological collinearity. Meanwhile, purely field-based evaluation is inherently limited by the absence of verifiable ground truth: visually indiscernible anomalous samples cannot be definitively labelled for quantitative assessment.

To bridge this gap, we adopt a semi-synthetic evaluation framework that preserves authentic real-data covariance structures while embedding precisely controlled artificial contamination signals, enabling quantitative evaluation with realistic

geological confounding; the contamination labels, however, remain planted.

We use the published drill-cuttings XRF dataset of Hall [15], consisting of 269 down-hole samples acquired from the lateral segment of an unconventional hydrocarbon well at a nominal sampling interval of 9.14 m (30 ft), with gaps where samples are absent. Each sample comprises nine normative components—eight mineral phases plus organic matter—forming a closed compositional system summing to 100%, eleven major-element oxides together with chlorine, and zirconium trace-element concentrations in mg kg^{-1} (ppm). The mineral block is closed by construction. The oxide block is treated as approximately closed and analysed as a subcomposition, and zirconium as an open-scale variable. Following the compositional pre-treatment protocol in Section 3.4, mineralogical and oxide subsets are independently centred log-ratio (clr) transformed, while zirconium abundances are log-transformed to mitigate skewed distributions, yielding a final set of 22 predictive features.

A realistic KCl-brine drilling-mud contamination signature is artificially implanted within the real background data via proportional elevation of K_2O , Na_2O , and Cl concentrations. Mass addition is implemented in raw untransformed concentration space prior to compositional closure and log-ratio transformation, mimicking the physical process of exogenous mud residue accumulation on drill cuttings. We designate 25 fully contaminated samples as the *visible anchor group*, simulating anomalies identifiable by routine visual inspection of the cuttings, and 12 low-amplitude contaminated samples, implanted at 55% of the visible-group amplitude, as the *hidden target group* for blind detection evaluation. Two scale parameters are therefore in play and should not be confused: the 55% ratio fixes the hidden group relative to the visible group, while the overall implant magnitude is a separate multiplier, held at a reference value of 1.3 except in the strength sweep of Figure 8, where it is varied from 0.6 to 2.0.

Unlike the fully synthetic benchmark requiring manually constructed confounders, this real-data background contains intrinsic geological confounding mechanisms. In the original dataset, potassium is predominantly clay-hosted: K_2O exhibits strong positive correlation with Al_2O_3 ($r = 0.94$) and clay mineral components including illite, smectite, and mica ($r = 0.45$). Naturally clay-rich

Table 1. Planted subpopulations in the synthetic benchmark dataset ($n = 800$).

Population	Elevated Elements	n	Known to Analyst?
Visible mud contamination	Ba, K, Cl, Sr (full amplitude)	40	Yes
Hidden mud contamination	Ba, K, Cl, Sr (60% amplitude)	15	No—detection target
Natural barite nodules	Ba, S	35	No—geological confounder
Unaffected background	—	710	—

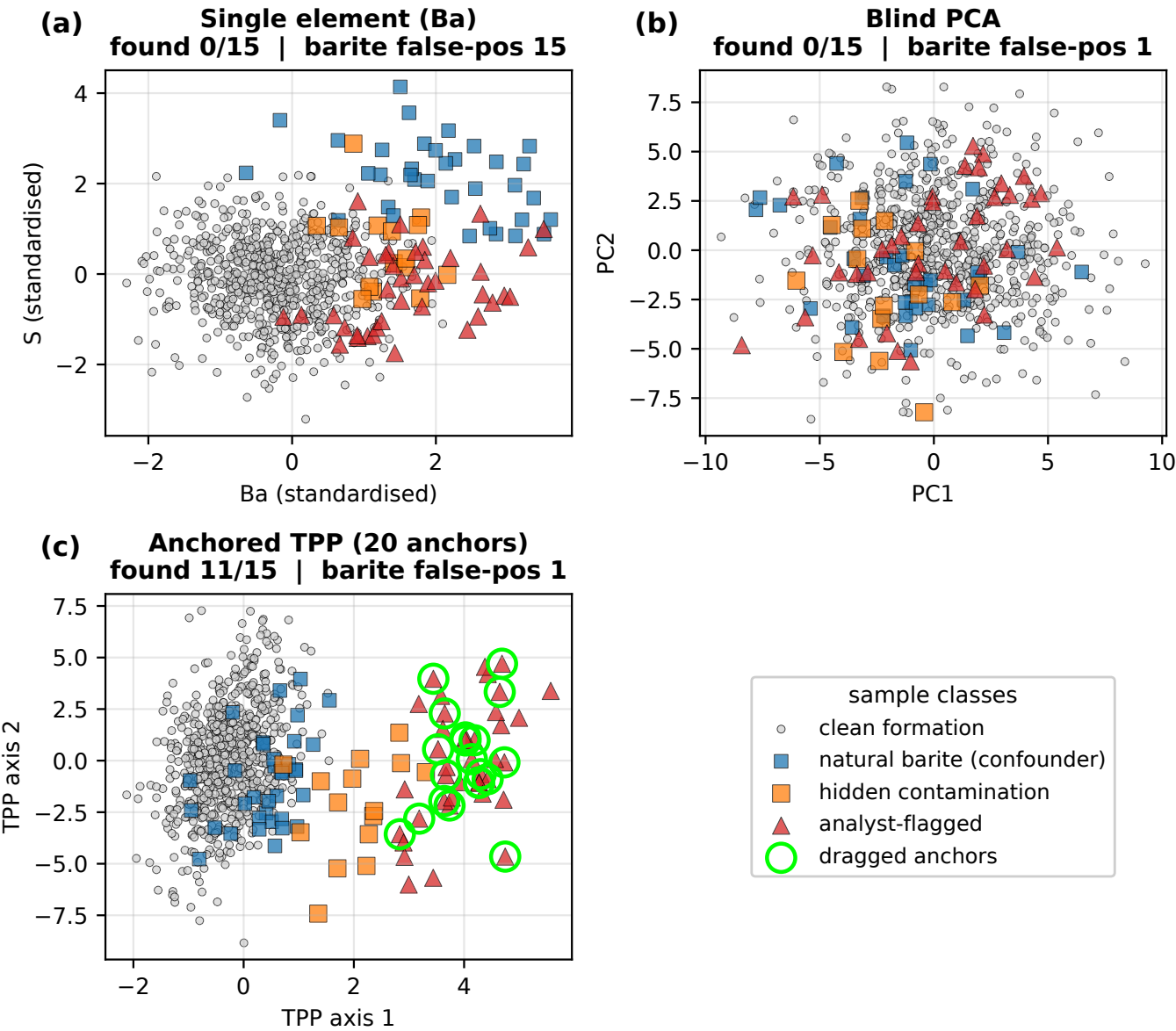


Figure 3. Detection performance of hidden contamination across three comparative methods. (a) Single-element view: bivariate scatter plot of standardized Ba versus S. Natural barite nodules overlap substantially with mud-contaminated samples in Ba abundance, precluding effective discrimination via univariate Ba criteria. (b) Blind PCA projection; subtle contamination signals are overwhelmed by dominant lithological variance and cannot be resolved in the first two principal components. (c) Anchored TPP projection; green circles denote anchor samples. Hidden contaminated samples (orange) are coherently displaced alongside anchors, while natural barite confounders are spatially separated on the opposite side of the projection axis. The fourth cell carries the shared legend.

lithologies therefore exhibit endogenous potassium enrichment unrelated to drilling contamination. This intrinsic collinearity undermines univariate potassium

screening, which tends to confuse natural lithological variability with anthropogenic mud contamination. Reliable discrimination requires identification of the

multi-element co-enrichment of K, Na, and Cl in the absence of clay-associated aluminium enrichment—a compound criterion inaccessible via univariate or pairwise analysis.

4.3 Evaluation Protocol

A practically oriented evaluation protocol is adopted to replicate operational geochemical QC workflows. All visually confirmed contaminated samples (visible group) are excluded from the candidate screening pool, as re-identifying known anomalies does not constitute valid discovery. Remaining samples are ranked according to the primary projection coordinate generated by anchored TPP, and the top $k = 15$ samples are retained as the inspection review list. The value of k is matched to the number of hidden contaminated samples, so that every method is scored on a list of the same length.

Method performance is quantified by two core metrics: the number of true hidden contamination samples successfully recovered in the review list, and the number of natural barite confounder samples falsely included as false positives. Two widely applied baseline methods are evaluated under the same protocol: univariate ranking based on standardized Ba concentrations, and unsupervised PCA ranking using the first principal component.

Uncertainty is reported as follows. Where a quantity is averaged over realizations, $\pm$ denotes the standard error of the mean unless stated otherwise, and 95% confidence intervals (CIs) are t intervals over the same realizations. When two methods, or two anchor strategies, are scored on the same realization, they are compared by the mean paired difference with its 95% CI and the paired standardized effect size d_z , the mean difference divided by the standard deviation of the differences; the Wilcoxon signed-rank test supplies the accompanying p -value. Each realization re-draws the implanted signal, the anchor subset and the random target offset within a single synthetic generator or a single real background, so realizations are not independent geological datasets, and the p -values are conditional on the benchmark design (Section 6.3).

5 Results

5.1 Detection Performance on the Fully Synthetic Benchmark

Detection performance across all methods is visualized in Figure 3 and quantitatively summarized in Table 2, where anchored TPP is implemented using $m = 20$

anchor samples randomly selected from the visually confirmed contamination group.

Table 2. Fully synthetic benchmark: detection performance under the standardized review-list protocol ($k = 15$).

Method	Hidden Samples Recovered (of 15)	Barite False Positives
Single-element Ba criterion	0	15
Blind PCA (PC1)	0	1
Anchored TPP ($m = 20$)	11	1

Both baseline methods fail to recover hidden contamination signals under the operational protocol. The univariate Ba criterion fails completely: all 15 samples on its list are natural barite nodules. Blind PCA also yields no true detection, as subtle anthropogenic contamination variance is negligible relative to dominant lithological and environmental variability captured by principal components. In contrast, anchored TPP recovers 11 of the 15 hidden contaminated samples with a single false positive (Table 2 reports one anchor draw; the averages below are over thirty draws). Detection does not require a large anchor set, although it does improve steadily as anchors are added. Averaging over thirty independent anchor draws from the same synthetic dataset, a single anchor already recovers 3.90 ± 0.59 of the 15 hidden samples, five recover 6.03 ± 0.68 , ten recover 8.30 ± 0.50 , and twenty recover 9.67 ± 0.45 ; the curve continues to climb to 12.00 at $m = 40$ rather than saturating. Both baselines remain at zero throughout. Useful performance is therefore available from a handful of anchors, but anchor count is a genuine lever rather than an inert parameter, and we quantify it with replicate variability on real data in Section 5.6.3.

5.2 Recovered Loadings

The loading vector corresponding to the primary anchored TPP projection axis (Figure 4) quantitatively reconstructs the planted contamination signature, with dominant positive weights: Cl +0.86, Ba +0.41, K +0.22, and Sr +0.12. All remaining elemental loadings have magnitudes below 0.08, with a slight negative loading for S (−0.09).

Two points follow from the loading structure. First, anchored TPP recovers the complete multi-element contamination signature without explicit prior

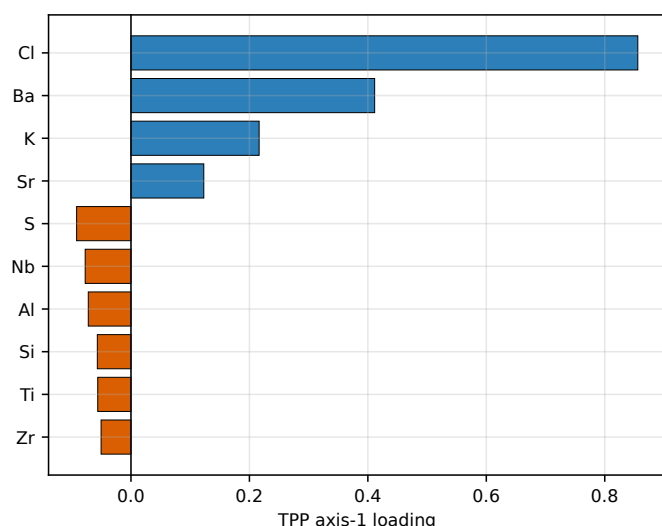


Figure 4. Elemental loadings of the primary anchored-TPP projection axis. Positive loading coefficients (blue) reproduce the multi-element drilling-mud contamination signature. The weak negative sulfur loading (orange) suppresses interference from natural barite confounders. All loadings correspond to standardized variables; for clr-transformed compositional data, interpretations follow the log-ratio contrast conventions defined in Section 3.5.

specification of target elemental associations. Second, the negative S loading does the discriminating work: natural barite nodules contain coupled Ba and S enrichment, whereas the simulated drilling-mud contamination elevates Ba without concurrent S anomalies. By down-weighting S abundance, the projection actively suppresses geogenic confounder signals. This compound discriminative criterion—elevated Cl, Ba, K, and Sr in the absence of S enrichment—cannot be derived from univariate thresholds or pairwise elemental cross-plots, and is learned entirely from sparse anchor constraints.

The dominant Cl loading reflects the strong contrast of chlorine in the synthetic contamination signature. In practical XRF analysis, chlorine measurements often carry elevated instrumental uncertainty, so on real data the weights may shift toward Ba and K; this has not been tested here.

5.3 Robustness to Anchor Misspecification

In practical workflows, visually selected anchor samples will contain some misclassification errors. To evaluate method robustness, we progressively replace valid contamination anchors with unaffected background samples, with performance averaged over twelve independent random realizations (Figure 5).

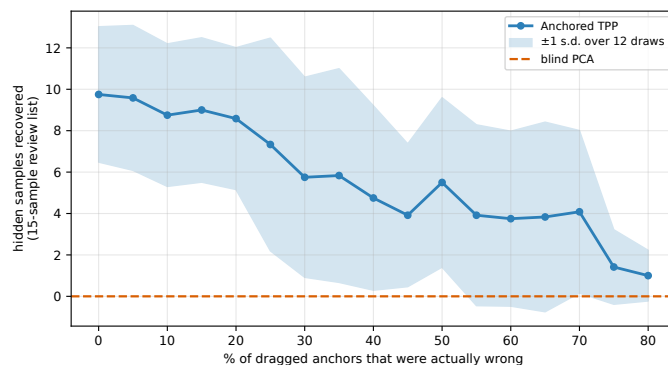


Figure 5. Performance degradation under progressive anchor misspecification. The shaded interval denotes ± 1 standard deviation across twelve independent realizations.

Detection performance degrades gradually and predictably with increasing anchor corruption. Mean recovery decreases from 9.8 (error-free anchors) to 7.3 at 25% misspecification, 5.5 at 50% misspecification, and 1.4 at 75% misspecification. No catastrophic failure occurs under moderate anchor error, and anchored TPP retains a performance advantage over blind PCA, which recovers none, across the entire range tested: mean recovery is still 1.0 of 15 when 80% of anchors are wrong. This is the expected behaviour: a projection guided by anchors should respond to the quality of those anchors.

For operational application, we propose a practical stability diagnostic: repeated projection fitting using disjoint anchor subsets. Loading vector instability across replicates indicates unreliable anchor selection, which should be explicitly reported alongside detection results to quantify uncertainty.

5.4 Spatial Autocorrelation and Anchor Placement

The baseline synthetic benchmark assumes independent sample distributions, which deviates from real down-hole geochemical profiles characterized by strong spatial autocorrelation. Lithological units and contamination intervals persist over contiguous depth intervals, reducing the effective number of independent observations—a well-known consequence of spatial autocorrelation on sample variance estimates [50]—and introducing multivariate structure that complicates anchor selection [51]. To assess spatial effects, we regenerate the benchmark dataset with all latent geological factors following a first-order autoregressive [AR(1)] process, and embed planted subpopulations within contiguous depth intervals.

Two anchor-selection strategies are compared at

fixed $m = 20$: dispersed anchors sampled across multiple discontinuous contamination intervals, and clustered anchors extracted from a single contiguous contaminated segment. Results are averaged over eighty realizations, with the standard error of each strategy shown in Figure 6 and the paired difference between strategies reported with its 95% CI.

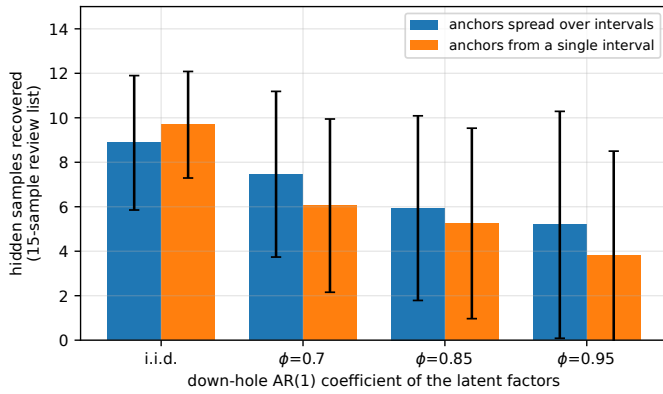


Figure 6. Effects of down-hole spatial autocorrelation and anchor placement strategy on detection recovery. Latent geological factors follow AR(1) processes with varying autocorrelation coefficients ϕ ; contamination populations occupy contiguous depth intervals. Error bars represent ± 1 standard error of the mean across eighty realizations; standard errors range from 0.27 to 0.57 across configurations.

Two conclusions follow from the spatial simulations. First, spatial autocorrelation substantially degrades detection performance by reducing effective sample size. For dispersed anchor placement, mean recovery declines monotonically from 8.88 ± 0.34 (independent samples) to 7.46 ± 0.42 ($\phi = 0.7$), 5.94 ± 0.46 ($\phi = 0.85$), and 5.19 ± 0.57 ($\phi = 0.95$), a reduction of roughly 40% under strong autocorrelation. Realization-to-realization variability increases concurrently, as autocorrelated samples fail to capture full population heterogeneity.

Second, anchor placement exerts a measurable but conditional effect. Both strategies are applied to the same generated core in each realization, so we compare them by the paired difference, dispersed minus clustered. With independent samples, clustered selection is slightly better (-0.81 samples, 95% CI $[-1.49, -0.14]$, $d_z = -0.27$). Once the data are autocorrelated, dispersed selection is better, by $+1.41$ samples at $\phi = 0.7$ ($[0.42, 2.41]$, $d_z = 0.32$) and $+1.38$ at $\phi = 0.95$ ($[0.55, 2.20]$, $d_z = 0.37$), while at $\phi = 0.85$ the difference ($+0.69$, $[-0.29, 1.67]$) is not resolved. The penalty for clustered anchors therefore appears once samples are autocorrelated, is of the

order of one hidden sample in fifteen, and does not grow systematically with ϕ . The effect sizes are small: anchor placement is a secondary lever compared with the loss of recovery caused by autocorrelation itself.

Operationally, analysts are recommended to select anchors from multiple spatially separated intervals wherever possible to maximize representativeness. Where only a single contaminated interval is available, detection results should be treated as optimistic, and anchor stability diagnostics should be prioritized to quantify uncertainty.

5.5 A Limitation of the Benchmark Itself

To ensure transparent evaluation, we explicitly acknowledge a design limitation of the synthetic benchmark that constrains the generality of its comparative conclusions. The benchmark successfully defeats Ba-based single-element screening via natural barite confounders, but it contains an unintended methodological shortcut: chlorine exhibits negligible natural geological variability and functions as an almost pure contamination tracer. In this synthetic configuration, a simple chlorine-based univariate ranking outperforms anchored TPP. While the benchmark validates that fixed *a priori* criteria such as Ba thresholds fail, it does not rule out optimally selected single-element tracers.

Because this artefact distorts the relative ranking of methods that exploit the same sparse positive information, we refrain from drawing comparative conclusions against anchor-aware semi-supervised learners, including PU logistic regression and the one-class support vector machine (SVM) [30], from this benchmark. The definitive comparison with such methods is deferred to the semi-synthetic experiment (Section 5.6), where authentic geochemical collinearity removes tracer shortcuts and provides a neutral basis for evaluating the trade-offs between detection recall, interpretability, and computational efficiency.

We therefore restrict the synthetic conclusions to one statement: in this fully controlled setting, anchored TPP outperforms blind unsupervised methods and fixed *a priori* single-element rules.

5.6 Semi-Synthetic Evaluation on a Real Geochemical Background

This section reports results on the semi-synthetic benchmark of Section 4.2. It carries the authentic inter-element collinearity that the fully synthetic benchmark lacks, and admits no single-element

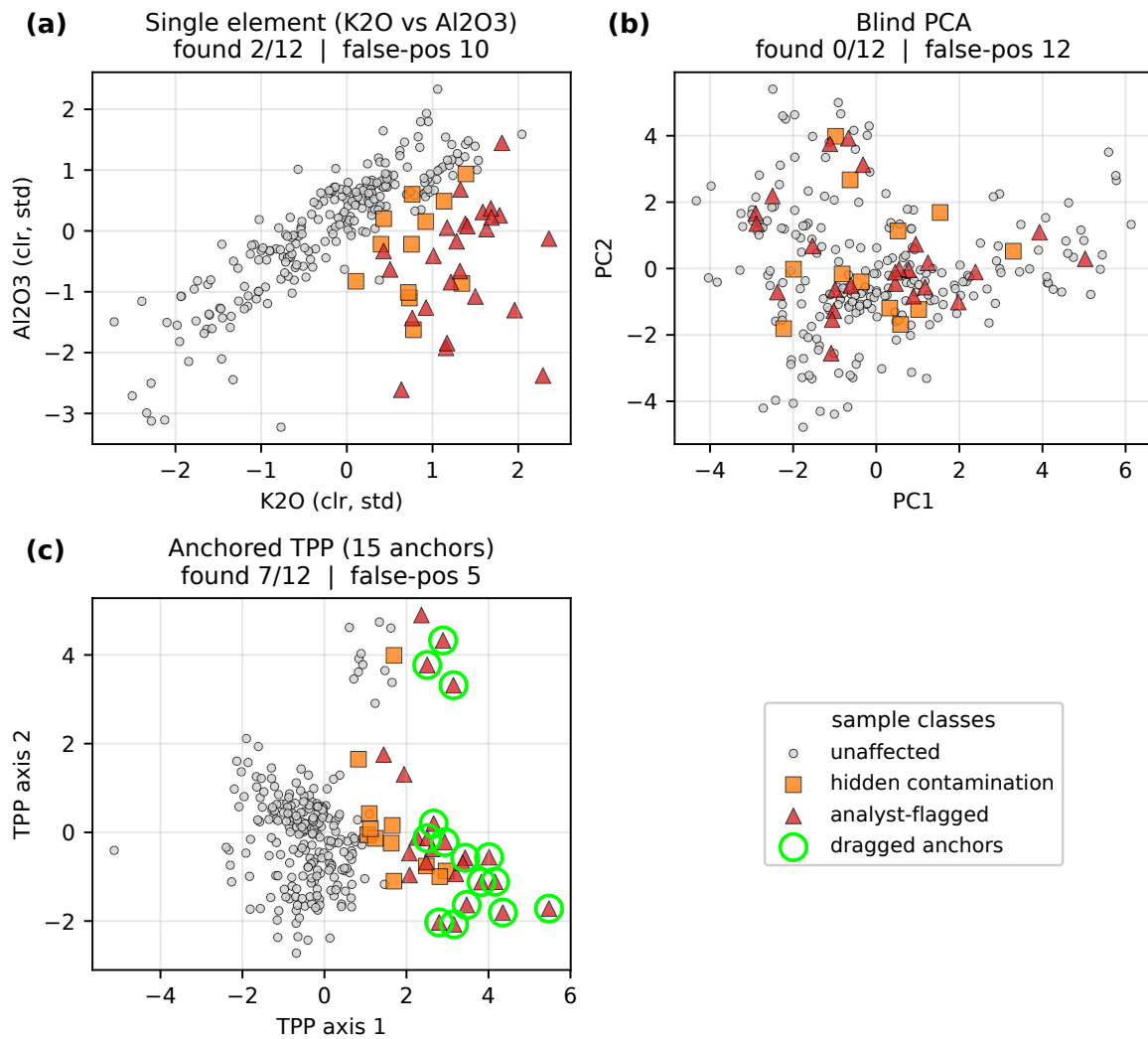


Figure 7. Detection of planted drilling-mud contamination against a real geochemical background. (a) Single-element view: clr-transformed K_2O - Al_2O_3 bivariate plane. The dominant linear trend reflects natural clay-hosted potassium variability; contaminated samples deviate from this lithological trend but overlap with geogenic potassium-rich samples in univariate potassium space. (b) Blind PCA projection fails to resolve subtle contamination signals amid dominant lithological variance. (c) Anchored TPP projection with 15 anchor samples (green circles); hidden contaminated samples are coherently separated from the natural geological background alongside anchor constraints. The fourth cell carries the shared legend.

shortcut, so it provides the primary, though still semi-synthetic, evidence for the method.

5.6.1 Detection Performance

We adopt the standard review-list evaluation protocol established in Section 4.3. All visually confirmed visible contamination samples are excluded from candidate screening to avoid trivial re-detection. Remaining samples are ranked by the primary anchored TPP projection coordinate, with the top 12 samples retained as the inspection short-list, matching the total number of hidden contaminated targets. A random ranking of the 244 candidates (269 samples less the 25 visible anchors) would place on average $12 \times 12/244 = 0.59$ hidden samples in the list; this is

the chance level.

Figure 7 presents a representative single-realization projection result, while Table 3 and Figure 9(a) summarize performance across sixty realizations with randomized contamination placement and anchor selection.

Unsupervised blind PCA achieves marginal improvement over random chance, consistent with the view that subtle anthropogenic contamination signals are overwhelmed by dominant natural lithological variability in real datasets and cannot be resolved via variance-maximizing decomposition. The conventional *a priori* potassium single-element

Table 3. Semi-synthetic benchmark: detection performance of competing methods for 12 hidden contaminated samples embedded within a real geochemical background.

Method	Uses Anchors	Recovered (of 12)	95% CI
Random ranking (theoretical expectation)	–	0.59	–
Blind PCA (PC1 unsupervised)	no	0.77 ± 0.09	0.59-0.95
<i>A priori</i> single-element K_2O criterion	no	1.97 ± 0.14	1.69-2.25
Anchored TPP (raw weight-percent data)	yes	3.08 ± 0.24	2.60-3.56
Anchor-selected single-element criterion	yes	3.93 ± 0.18	3.57-4.29
One-class SVM (anchor-trained)	yes	4.95 ± 0.20	4.55-5.35
Anchored TPP (clr-transformed compositional data)	yes	7.03 ± 0.26	6.51-7.55
PU logistic regression (anchor-positive learning)	yes	7.58 ± 0.15 ¹	7.28-7.88

¹ Reported values denote mean recoveries $\pm$ standard errors across sixty realizations at fixed contamination strength (1.3) and 15 anchors; the 95% CI is the t interval over the same realizations, which are repeated implantations in one real background rather than independent datasets (Section 6.3). Bold marks the method proposed in this work; it does not denote the highest score, which is attained by PU logistic regression.

criterion yields only modest gains, limited by clay-hosted potassium. In contrast, anchored TPP recovers several times more true positives than the unsupervised baselines at every contamination strength tested: roughly threefold at the weakest planted signal, rising at the strongest to more than tenfold against blind PCA and about threefold against the *a priori* K_2O criterion (Figure 8).

The semi-synthetic results change the ranking of methods seen in the fully synthetic experiment. Paired over the sixty realizations, anchored TPP recovers more hidden samples than anchor-selected univariate screening (mean difference 3.10 samples, 95% CI [2.51, 3.69], $d_z = 1.37$) and one-class SVM (2.08, [1.49, 2.67], $d_z = 0.92$), with Wilcoxon $p < 10^{-4}$ in both cases. PU logistic regression recovers marginally more than anchored TPP: the difference is -0.55 samples of 12 ($[-1.03, -0.07]$, $d_z = -0.29$, $p = 0.035$), a small effect whose interval lies within about one sample, compared with the large gaps observed in idealized synthetic data. These p -values are conditional on the benchmark design (Section 6.3).

This reversal relative to univariate methods is the result that matters for practice. The fully synthetic benchmark contained an artificial pure-element tracer shortcut enabling univariate rule superiority, whereas real geochemical backgrounds feature universal elemental collinearity and genuine geological variability across all channels. Under realistic confounding, a compound multi-element criterion is what separates contamination from lithology, and no single channel substitutes for it.

Absolute recovery rates are lower than in fully synthetic tests, as expected. Authentic lithological

collinearity reduces unique signal variance attributable to contamination, giving a more realistic, though still optimistic, indication of field detection performance than idealized synthetic benchmarks (Section 6.4).

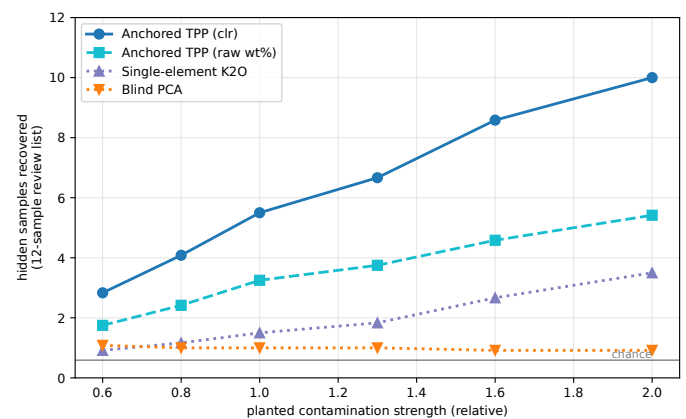


Figure 8. Detection recovery as a function of planted contamination signal strength. Performance curves represent mean values across sixty independent realizations. Both raw and clr-corrected anchored TPP variants consistently outperform all unsupervised baselines across the full strength range, while blind PCA remains near chance-level recovery.

5.6.2 Effect of Compositional Pre-Treatment

Comparing raw and clr-transformed anchored TPP isolates the effect of compositional pre-processing. At the reference contamination strength of 1.3, clr transformation improves mean recovery from 3.08 ± 0.24 to 7.03 ± 0.26 , a 2.3-fold gain. The gain from this single pre-processing step exceeds the spread between the competing anchor-aware algorithms in Table 3, which makes it the most consequential choice in the workflow.

The physical mechanism driving performance

improvement originates from compositional closure artefacts. Exogenous mud mass addition elevates target elemental concentrations while proportionally depressing all other components via constant-sum closure in raw concentration space, diffusing contamination signatures across the entire compositional spectrum. The centred log-ratio transformation removes much of this closure-induced covariation, because ratios are unaffected by the common rescaling, so the contamination signal stays concentrated in the elements actually enriched. We are not aware of a prior quantitative estimate of this effect for projection-pursuit-based geochemical QC.

5.6.3 Anchor Economy

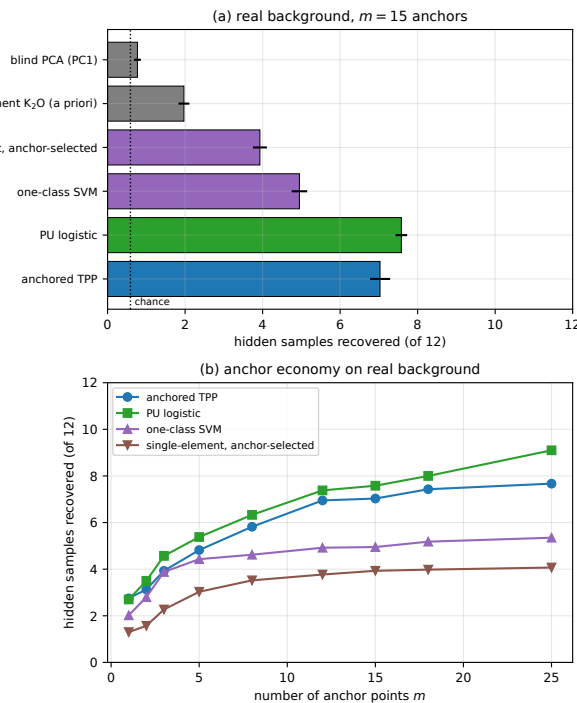


Figure 9. Comprehensive method benchmarking on a real geochemical background. (a) Comparative recovery at fixed $m = 15$ anchors and contamination strength 1.3; error bars denote standard errors over sixty realizations; the horizontal dotted line marks chance-level recovery. Grey bars indicate anchor-free unsupervised methods. (b) Recovery as a function of the number of anchors, for every anchor-aware method under an identical protocol; curves are means over sixty realizations.

Anchor efficiency on real geochemical data follows the same pattern, and is reported here alongside the competing anchor-aware methods so that the comparison is made on one axis rather than two (Figure 9(b)). A single anchor yields a mean recovery of 2.75 hidden anomalies; five anchors achieve 4.82; recovery then rises more slowly, reaching 6.95 at twelve anchors, 7.43 at eighteen and 7.67 at twenty-five. The

curve begins to flatten beyond about fifteen anchors, whereas the fully synthetic curve was still climbing at forty, which is what the greater background collinearity of real data would predict.

The cross-algorithm hierarchy is stable from two anchors upward: anchored TPP consistently outperforms one-class SVM and anchor-selected univariate screening while remaining marginally subordinate to PU logistic regression. At a single anchor the two are indistinguishable (2.75 versus 2.70, $p = 0.73$). The PU advantage widens moderately with increasing anchor counts, from 0.37 samples at $m = 2$ (95% CI [0.03, 0.71]) to 1.43 at $m = 25$ ([0.97, 1.89]), which suggests that the discriminative PU model makes slightly better use of additional anchors.

The operational point stands regardless of the ordering: about ten confidently identified samples suffice, which is what a core inspection session already yields.

5.6.4 Anchor Placement on Real Down-Hole Structure

Spatial autocorrelation analysis in synthetic experiments (Section 5.4) recommended dispersed multi-interval anchor selection to mitigate spatial bias. The real down-hole background allows this guideline to be tested under an authentic geological correlation structure, with the contamination itself still planted. Clr-transformed elemental features exhibit a median lag-one down-hole autocorrelation coefficient of 0.61, closely matching the moderate autocorrelation regime ($\phi = 0.7$) where synthetic spatial effects were statistically resolvable.

Contamination signals are replanted as continuous ten-sample depth intervals to replicate realistic mud-invasion geometry, which affects contiguous borehole segments rather than discrete random samples. Two anchor-selection strategies are compared: dispersed anchors sampled across multiple discontinuous contaminated intervals, and clustered anchors extracted from a single contiguous contaminated segment. Averaged over one hundred independent realizations, dispersed anchor placement yields a mean recovery of 6.04 ± 0.30 against 5.02 ± 0.28 for clustered placement; the paired difference is 1.02 samples (95% CI [0.56, 1.48], $d_z = 0.44$, Wilcoxon $p < 10^{-4}$).

The spatial sampling guideline is therefore supported in the semi-synthetic experiment on a real geochemical background, where the placement effect is more consistent than in the fully synthetic experiments. Exclusive anchor selection from a single visually

Table 4. Methodological positioning of anchored TPP relative to competing multivariate and anomaly detection approaches.

Method	Requires Labels	Uses Domain Hypothesis	Interpretable Output	Well Posed for $p \gg n$	Automatable
PCA, k -means	No	No	Partial	Yes	Yes
t -SNE, UMAP	No	No	No	Yes	Yes
LDA	Yes (complete)	Yes	Yes	No	Yes
PLS-DA	Yes (complete)	Yes	Yes	Yes	Yes
Isolation forest, LOF	No	No	No	Yes	Yes
Robust Mahalanobis, MCD	No	No	Partial	No	Yes
One-class SVM	Positives only	Partial	No	Yes	Yes
PU learning	Positives only	Partial	Partial	Partial	Yes
Anchored TPP	Positives only	Yes	Yes	Yes	Partial

stained depth interval reduces detection performance by approximately one true positive sample per twelve targets. Operationally, anchor stability diagnostics should employ depth-blocked resampling rather than simple random subsetting to account for down-hole spatial structure. On this real dataset, anchored TPP loading vectors exhibit high reproducibility across disjoint anchor subsets (mean absolute cosine similarity = 0.92), with 80% overlap in the top-15 inspection lists.

5.6.5 Behaviour with Respect to the Natural Confounder

Quantitative false-positive analysis confirms the distinct confounding failure modes of univariate versus multi-element screening strategies. Over twenty realizations, false positives generated by the potassium single-element criterion correspond to samples at the 0.89 ± 0.02 quantile of aluminium abundance, demonstrating near-exclusive misclassification of naturally clay-rich lithologies as contaminated. In contrast, anchored TPP false positives occur at the 0.69 ± 0.24 aluminium quantile and are less than half as numerous (4.9 versus 10.1 per realization; Mann-Whitney U , $p = 0.03$, conditional on the benchmark design, Section 6.3).

The confounding-suppression mechanism differs from that of the synthetic benchmark. Unlike the explicit negative sulfur loading that suppressed natural barite signals in synthetic data, anchored TPP does not rely on strong negative clay-associated Al_2O_3 loadings (mean loading = -0.17 ± 0.07 , fifth-largest magnitude among non-planted variables). Pervasive negative loadings across all unplanted components arise from compositional closure mass-balance effects rather than targeted confounder suppression.

Instead, discrimination efficacy originates from strict multi-element joint-signature requirements: valid contamination must exhibit simultaneous elevation of Cl, K, and Na, a combination that clay mineralogy alone does not produce. The three planted contamination elements consistently dominate positive loading weights across replicates (Cl: $+0.48 \pm 0.03$; K_2O : $+0.41 \pm 0.04$; Na_2O : $+0.34 \pm 0.05$), yielding stable, interpretable, and geologically consistent multi-element anomaly criteria.

6 Discussion

6.1 Positioning

Unsupervised machine-learning anomaly detectors are widely used for high-dimensional geochemical data screening, yet they produce purely statistical outlier scores without geological interpretation of the driving element associations [52, 53]. Blind PCA similarly prioritizes dominant global variance and may suppress subtle but geologically meaningful signals [54]. Interpretability has consequently become an explicit objective in geochemical machine learning, pursued mainly through post hoc attribution applied to an already-fitted black-box model [55]; anchored TPP instead obtains it *a priori*, because the fitted object is itself a signed vector in the original variable space. Anchored TPP is aimed at this gap: it is a hypothesis-guided, human-in-the-loop QC method that needs only a small set of anchor examples rather than a fully labelled training set.

Table 4 compares anchored TPP with established multivariate and anomaly-detection methods. Linear discriminative modelling for group separation is a well-established classical technique widely applied in

lithogeochemical facies analysis; no novelty is claimed for basic linear separation capability.

What distinguishes anchored TPP is the combination: hypothesis-guided analysis with sparse positive anchors, no negative labels, interpretability through variable-level loadings, stability when $p \gg n$, and suitability for interactive use. This combination matches routine geochemical QC practice, where analysts possess qualitative geological suspicions but no formal labelled datasets.

Anchored TPP uses sparse positive anchor samples and no negative labels, matching the information structure of one-class and PU learning. It thus avoids the dependence on arbitrarily designated negative samples that is known to affect predictions in this setting [26]. The distinction lies not in the input information but in the output: interpretable loadings, no tuning parameters, well-posedness when $p \gg n$, and fast interactive refitting.

Anchored TPP also complements automated black-box detectors within integrated workflows: unsupervised isolation forests and PU classifiers provide fast, scalable statistical screening, while anchored TPP delivers interpretable geological diagnosis to distinguish instrumental artefacts from genuine lithological anomalies. A two-stage pipeline—automated ranking, then anchored-TPP inspection of the ranked candidates—uses each where it is strongest.

6.2 Applications in Geo-Energy Workflows

Although the method is indifferent to the commodity behind the survey, its most immediate use is in geo-energy drilling, where cuttings geochemistry is acquired continuously and consumed quickly. We outline four settings. None has yet been tested against confirmed field contamination, so they are stated as potential applications rather than demonstrated ones.

Wellsite cuttings QC. Mud-logging and wellsite geochemistry teams usually know which intervals were drilled after a change of mud system, with a brine or lost-circulation material, or with poor hole cleaning, and they can often point to a few samples that look affected. Those samples are exactly the anchors the method needs. Because refitting costs $O(mp)$ operations, the anchored projection can be recomputed as each new cuttings sample arrives, and the loading vector states which elements drive a flag—K, Na and Cl for a KCl brine, Ba for barite weighting material—so that the flag can be checked against the mud report

before the data are released for interpretation.

Chemostratigraphy and well correlation. Chemostratigraphic zonation relies on element ratios and multivariate boundaries that a drilling-fluid artefact can mimic: an interval of KCl contamination raises K relative to Al and can be read as a change in clay mineralogy or provenance. Screening with anchored TPP before zonation, and removing or down-weighting flagged samples, reduces the risk of placing a correlation marker on an artefact, and the signed loading vector records why each sample was set aside.

Reservoir characterization. Whole-rock inorganic geochemistry of cuttings from shale resource plays enables discrimination of clay-rich, carbonate-bearing and organically enriched intervals, providing a continuous chemostratigraphic profile that directly informs lateral placement and completion design [16]. The confounder in the semi-synthetic experiment is clay-hosted potassium, so that experiment bears directly on this use: a multi-element contrast separates brine contamination from genuinely clay-rich intervals, whereas a potassium threshold would remove the clay-rich intervals the reservoir model most needs to represent.

Other subsurface geo-energy workflows. The same data type and the same drilling artefacts arise in wells drilled for geothermal energy, carbon dioxide storage and underground hydrogen storage, where cuttings and core geochemistry characterize reservoir and caprock mineralogy for reactivity and integrity assessments. Anchored TPP applies unchanged provided the analyst can exemplify the suspected artefact with a few samples. The elemental signature will differ, for instance where flowback brine from hydraulic fracturing rather than a KCl drilling mud is the contaminant source [18], but the workflow does not.

6.3 Statistical Scope of the Benchmark Results

The 20 to 100 realizations behind the reported intervals and p -values are repeated signal implantations, anchor draws and target offsets inside one synthetic generator or one real well. They measure how stable the algorithm's behaviour is under that benchmark design—how much recovery varies with where the contamination falls and which anchors happen to be picked—not how it would vary across wells, basins, lithologies, mud systems or XRF instruments. The p -values should therefore be read as conditional Monte

Carlo evidence: within this benchmark, a difference of the reported size is unlikely to be an artefact of the particular draws; they say nothing about its size in another geological setting.

Two consequences follow. First, with enough realizations any fixed difference becomes statistically detectable, so effect sizes and confidence intervals are the more informative quantities, and we report them alongside every test. In the semi-synthetic experiment the differences between anchored TPP and the anchor-free baselines are large ($d_z = 3.21$ against blind PCA and 2.16 against the *a priori* K_2O criterion, from the paired differences of 6.27 ± 0.25 and 5.07 ± 0.30 samples); those against one-class SVM and anchor-selected screening are moderate to large ($d_z = 0.92$ and 1.37); and the shortfall against PU learning is small (-0.55 samples, 95% CI $[-1.03, -0.07]$, $d_z = -0.29$), which is why we describe it as marginal rather than as a tie. Second, the between-setting variance that governs field performance—different backgrounds, contamination chemistry, sample spacing—is not sampled at all. Establishing that the ordering of methods holds across settings requires several independent wells or datasets, with the well rather than the realization as the unit of replication; we list this with laboratory-confirmed contamination as the principal requirement for field validation.

6.4 Limitations

The method and the evaluation have several limitations.

First, anchored TPP relies on strictly linear projections and is therefore unable to capture curved anomaly boundaries or complex variable interactions. While kernel and neural variants can recover non-linear structure, they do so by sacrificing the directly interpretable loading vectors that motivate the method in the first place.

Second, the method employs an objective substitution: the closed-form solution maximizes the alignment objective (Equation (5)) rather than minimizing the full weighted least-squares criterion (Equation (3)), which lacks an analytic solution for rectangular orthonormal projections. Detection performance remains comparable in practice, but the fitted projections do not: their leading axes align only to 0.74 at small ω_{bg} and to 0.41 at large (Figure 2), so the two criteria are not interchangeable.

Third, stability relies on an anchor-centroid

approximation, with the primary axis almost coinciding with the standardized anchor centroid ($\cos \theta = 0.9998 \pm 0.0002$). Consequently, the method inherits the well-known vulnerabilities of centroid-based estimators, including sensitivity to outlying anchors and an inability to model multi-modal target populations. The effects of robust or multi-group anchor targets remain untested.

Fourth, because projection quality depends on sample selection, analyst bias enters the workflow by design. Although performance degradation under anchor misspecification or spatially clustered selection is measurable (Sections 5.3 and 5.6.4), it cannot be eliminated entirely; this justifies including anchor-stability diagnostics within the standard pipeline output rather than delegating them to an appendix.

Fifth, a residual compositional dependence persists. Although clr pre-treatment removes closure artefacts that otherwise distort covariance structure [20, 46], clr components are linearly dependent by construction, meaning loadings cannot be interpreted as strictly independent contributions. While an isometric log-ratio basis would eliminate this constraint, it would do so at the expense of directly interpretable per-element coefficients.

Finally, the scope of empirical evidence is bounded by the synthetic nature of the evaluation datasets. All ground truth used here is planted: the synthetic benchmark is generated in open, non-compositional space—precluding tests of compositional correction—and contains an unintended near-pure tracer channel, which is why comparisons against anchor-aware baselines are excluded from the main evidence (Section 5.5). Furthermore, while the semi-synthetic experiment implants a fixed-composition signature into real background data to test authentic collinearity, it fails to capture the depth-dependent mixing and variable intensity typical of actual mud invasion. As a result, reported recovery rates represent upper bounds, the statistical tests are conditional on a single background (Section 6.3), and validation against laboratory-verified contamination intervals remains an important target for future work.

6.5 Further Work

Closing the gap between semi-synthetic simulation and operational field performance will primarily require field validation against independently laboratory-verified drilling contamination datasets.

Beyond empirical testing, several targeted methodological extensions offer promising avenues for development.

First, replacing the clr transformation with singularity-free isometric log-ratio (ilr) bases would eliminate linear dependence in the compositional input space, improving numerical stability [47, 56]. Second, generalizing from discrete anchor groups to continuous directional hypothesis constraints would enable the modelling of gradational stratigraphic transitions rather than restricting analysis to discrete anomalous populations. Third, developing robust, multi-group anchor targets would mitigate outlier sensitivity while allowing the framework to accommodate multi-modal anomalous distributions. Finally, explicitly integrating spatial coherence penalties would allow down-hole autocorrelation to be exploited as active supporting detection evidence rather than managed as a spatial nuisance.

7 Conclusions

This study presents anchored targeted projection pursuit (anchored TPP), a closed-form, tuning-free multivariate method for hypothesis-driven quality control of wellsite and core geochemical data when the analyst holds a geological suspicion but no class labels. A small set of analyst-selected anchors defines a target configuration; the resulting weighted orthogonal Procrustes alignment returns interpretable variable loadings rather than an opaque anomaly score, so that a flagged sample comes with a statement of the elemental contrast that flagged it.

The method's defining features originate from its closed-form Procrustes formulation. The alignment-focused objective structurally eliminates the background weight and target pull magnitude from the fitted projection, rendering the procedure intrinsically free of tunable hyper-parameters. Furthermore, the solution depends on the data solely through the anchor rows; this locality reduces refitting complexity to $O(mp)$ and preserves well-posedness in high-dimensional regimes where conventional covariance-based estimators are singular. By preserving variable-space orthonormality, the projection yields signed, unit-norm loading vectors that function as explicit geochemical contrasts, converting a flagged outlier into a diagnostically testable elemental signature rather than an opaque score.

In controlled synthetic and semi-synthetic

experiments, the latter on the real geochemical background of drill cuttings from a hydrocarbon well, anchored TPP consistently resolves subtle planted anomalies missed by unsupervised decompositions and univariate screening. Relative to other methods that exploit the same sparse positive information, anchored TPP sacrifices marginal recall compared to positive-unlabelled learning in exchange for direct interpretability, hyper-parameter independence, and interactive computational efficiency. It is not designed to maximize recall. Its role is diagnosis: when a sample is flagged, the signed loading vector tells the analyst which elements drove the flag. It is therefore a companion to automated screeners, not a replacement for them.

The current implementation is restricted to linear projections and, in this study, relies on planted ground truth; the results are therefore a proof of concept, and independent field validation against laboratory-confirmed contamination intervals, in several independent wells, remains an essential prerequisite for operational deployment in cuttings QC, chemostratigraphic and reservoir-characterization workflows. Future methodological development will target robust multi-group anchor aggregation, isometric log-ratio bases to eliminate compositional dependence, and spatially regularized formulations that explicitly encode down-hole autocorrelation as supporting evidence.

Data Availability Statement

The reference implementation of anchored TPP, the synthetic benchmark generator, and the scripts reproducing every figure and reported value in this paper are archived at Zenodo, DOI: [10.5281/zenodo.22668675](https://doi.org/10.5281/zenodo.22668675), and developed at <https://github.com/yupenglee2006-ui/anchored-tpp-reproducibility>, where the revision statistics script (`experiments/exp_revision_stats.py`) reproduces every confidence interval and effect size reported here (MIT licence for code; CC-BY-4.0 for figures and result files). The semi-synthetic experiments use the published cuttings XRF dataset of B. Hall ("Geochemical facies analysis using unsupervised machine learning," *CSEG Recorder*, pp. 26–30, 2018; [available here](#)), which is third-party and is not redistributed in the archive; it must be obtained from its original source. A synthetic stand-in generator is provided so that the pipeline can be exercised without it.

Funding

This work was supported without any funding.

Conflicts of Interest

Yupeng Li is affiliated with the Beijing Research Center, Upstream Advanced Research Center, Saudi Aramco, Beijing 102200, China, and Saeed S. Shahrani is affiliated with the Geological Operations Department, EXPEC, Saudi Aramco, Dhahran, Saudi Arabia. The authors declare that these affiliations did not influence the study design, data collection, analysis, interpretation, or decision to publish. The authors declare that no other competing interests exist.

AI Use Statement

The authors declare that DeepSeek-V4.1-Flash was used for language polishing during drafting, and Claude Opus 5.0 (Anthropic) was used for language editing and formatting during revision. The authors have carefully reviewed, revised, and verified the AI-assisted output and take full responsibility for the content of the manuscript.

Ethical Approval and Consent to Participate

Not applicable. This study did not involve human participants, human data or animals; it uses synthetic data and a published, third-party geochemical dataset of rock samples.

References

- [1] Kirkwood, C., Cave, M., Beamish, D., Grebby, S., & Ferreira, A. (2016). A machine learning approach to geochemical mapping. *Journal of Geochemical Exploration*, 167, 49-61. [CrossRef]
- [2] Caté, A., Perozzi, L., Gloaguen, E., & Blouin, M. (2017). Machine learning as a tool for geologists. *The Leading Edge*, 36(3), 215-219. [CrossRef]
- [3] Hall, B. (2016). Facies classification using machine learning. *The Leading Edge*, 35(10), 906-909. [CrossRef]
- [4] Lee, S., & Moon, I. (2026). Recent advances and future perspectives of AI-based mineral exploration: A review of machine learning, deep learning, and geologically informed approaches. *Minerals*, 16(6), 584. [CrossRef]
- [5] Howarth, R. J., Coles, B. J., & Ramsey, M. H. (2000). The potential of multivariate quality control as a diagnostic tool in geoanalysis. *Analyst*, 125(11), 2032-2037. [CrossRef]
- [6] Bourdeau, J. E., Zhang, S. E., Lawley, C. J., Parsa, M., Nwaila, G. T., & Ghorbani, Y. (2023). Predictive Geochemical Exploration: Inferential Generation of Modern Geochemical Data, Anomaly Detection and Application to Northern Manitoba. *Natural Resources Research*, 32(6), 2355-2386. [CrossRef]
- [7] Balaram, V., & Satyanarayanan, M. (2022). Data quality in geochemical elemental and isotopic analysis. *Minerals*, 12(8), 999. [CrossRef]
- [8] Rowe, H., Hughes, N., & Robinson, K. (2012). The quantification and application of handheld energy-dispersive x-ray fluorescence (ED-XRF) in mudrock chemostratigraphy and geochemistry. *Chemical geology*, 324, 122-131. [CrossRef]
- [9] Carr, R., Yarbrough, L., Lentz, N., Neset, K., Lucero, B., & Kirst, T. (2014). On-site XRF analysis of drill cuttings in the Williston Basin. In *Proceedings of the Unconventional Resources Technology Conference (URTeC 2014)*, Denver, Colorado, USA, Paper URTEC-1934308-MS. [CrossRef]
- [10] Ross, P. S., Bourke, A., & Fresia, B. (2014). Improving lithological discrimination in exploration drill-cores using portable X-ray fluorescence measurements: (1) testing three Olympus Innov-X analysers on unprepared cores. *Geochemistry: Exploration, Environment, Analysis*, 14(2), 171-185. [CrossRef]
- [11] Zhou, S., Wang, J., Wang, W., & Liao, S. (2023). Evaluation of portable X-ray fluorescence analysis and its applicability as a tool in geochemical exploration. *Minerals*, 13(2), 166. [CrossRef]
- [12] Grunsky, E. C., & Caritat, P. D. (2020). State-of-the-art analysis of geochemical data for mineral exploration. *Geochemistry: Exploration, Environment, Analysis*, 20(2), 217-232. [CrossRef]
- [13] Kowalska, S., Kubik, B., Skupio, R., & Wolański, K. (2020). Downhole lithological profile reconstruction based on chemical composition of core samples and drill cuttings measured with portable X-ray fluorescence spectrometer. *Minerals*, 10(12), 1101. [CrossRef]
- [14] Laperche, V., & Lemièrre, B. (2020). Possible pitfalls in the analysis of minerals and loose materials by portable XRF, and how to overcome them. *Minerals*, 11(1), 33. [CrossRef]
- [15] Hall, B. (2018). Geochemical facies analysis using unsupervised machine learning. *CSEG Recorder*, 26-30. https://cseg.ca/wp-content/uploads/2018-01-RECORDER-Geochemical_Facies_Analysis.pdf
- [16] Ratcliffe, K. T., Wright, A. M., & Schmidt, K. (2012). Application of inorganic whole-rock geochemistry to shale resource plays: An example from the Eagle Ford Shale Formation, Texas. *The Sedimentary Record*, 10(2), 4-9. [CrossRef]
- [17] Ratnayake, A. S., & Sampei, Y. (2019). Organic geochemical evaluation of contamination tracers in deepwater well rock cuttings from the Mannar Basin, Sri Lanka. *Journal of Petroleum Exploration and Production Technology*, 9(2), 989-996. [CrossRef]

- [18] Haluszczak, L. O., Rose, A. W., & Kump, L. R. (2013). Geochemical evaluation of flowback brine from Marcellus gas wells in Pennsylvania, USA. *Applied Geochemistry*, 28, 55-61. [CrossRef]
- [19] Ablard, P., Bell, C., Cook, D., Fornasier, I., Poyet, J.-P., Sharma, S., Fielding, K., Lawton, L., Haines, G., Herkommer, M. A., McCarthy, K., Radakovic, M., & Umar, L. (2012). The expanding role of mud logging. *Oilfield Review*, 24(1), 24-41. <https://www.scienceopen.com/document?vid=7e621021-d2df-4946-8182-Ocedca082a8b>
- [20] Filzmoser, P., & Hron, K. (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, 40(3), 233-248. [CrossRef]
- [21] Grunsky, E. C. (2010). The interpretation of geochemical survey data. *Geochemistry: Exploration, Environment, Analysis*, 10(1), 27-74. [CrossRef]
- [22] Hajihosseini, M., Maghsoudi, A., & Ghezelbash, R. (2024). Geochemical anomaly detection and pattern recognition: A combined study of the apriori Algorithm, principal component Analysis, and spectral clustering. *Minerals*, 14(12), 1202. [CrossRef]
- [23] Ren, X., Wang, G., & Mou, N. (2026). Identification of Geochemical Anomalies by Pattern Recognition: A Case Study of Wulonggou Area in Qinghai Province, China. *Minerals*, 16(4), 411. [CrossRef]
- [24] Wang, J., Zuo, R., & Xiong, Y. (2020). Mapping mineral prospectivity via semi-supervised random forest. *Natural Resources Research*, 29(1), 189-202. [CrossRef]
- [25] Zhou, K., Sun, T., Liu, Y., Feng, M., Tang, J., Mao, L., Pu, W., & Huang, J. (2023). Prospectivity mapping of tungsten mineralization in Southern Jiangxi Province using few-shot learning. *Minerals*, 13(5), 669. [CrossRef]
- [26] Zuo, R., & Wang, Z. (2020). Effects of random negative training samples on mineral prospectivity mapping. *Natural Resources Research*, 29(6), 3443-3455. [CrossRef]
- [27] Faith, J., Mintram, R., & Angelova, M. (2006). Targeted projection pursuit for visualizing gene expression data classifications. *Bioinformatics*, 22(21), 2667-2673. [CrossRef]
- [28] Jimenez, L. O., & Landgrebe, D. A. (1999). Hyperspectral data analysis and supervised feature reduction via projection pursuit. *IEEE Transactions on Geoscience and Remote Sensing*, 37(6), 2653-2667. [CrossRef]
- [29] Malpica, J. A., Rejas, J. G., & Alonso, M. C. (2008). A projection pursuit algorithm for anomaly detection in hyperspectral imagery. *Pattern Recognition*, 41(11), 3313-3327. [CrossRef]
- [30] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7), 1443-1471. [CrossRef]
- [31] Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 213-220). [CrossRef]
- [32] Bekker, J., & Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4), 719-760. [CrossRef]
- [33] Bourdeau, J. E., Zhang, S. E., Nwaila, G. T., & Ghorbani, Y. (2024). Data generation for exploration geochemistry: Past, present and future. *Applied Geochemistry*, 172, 106124. [CrossRef]
- [34] Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenkova, B., Schubert, E., ... & Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data mining and knowledge discovery*, 30(4), 891-927. [CrossRef]
- [35] Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, 100(9), 881-890. [CrossRef]
- [36] Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, 13(2), 435-475. [CrossRef]
- [37] Barnett, R. M., Manchuk, J. G., & Deutsch, C. V. (2014). Projection pursuit multivariate transform. *Mathematical Geosciences*, 46(2), 337-359. [CrossRef]
- [38] Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3), 212-223. [CrossRef]
- [39] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In *2008 eighth IEEE international conference on data mining* (pp. 413-422). IEEE. [CrossRef]
- [40] Chen, Y., & Wu, W. (2019). Isolation forest as an alternative data-driven mineral prospectivity mapping method with a higher data-processing efficiency. *Natural Resources Research*, 28(1), 31-46. [CrossRef]
- [41] Shahrestani, S., & Carranza, E. J. M. (2024). Effectiveness of LOF, iForest and OCSVM in detecting anomalies in stream sediment geochemical data. *Geochemistry: Exploration, Environment, Analysis*, 24(3), geochem2024-009. [CrossRef]
- [42] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93-104). [CrossRef]
- [43] Yang, Z., & Chen, Y. (2025). Anomaly detection-oriented positive-unlabeled metric learning for extracting high-dimensional geochemical anomalies linked to mineralization. *Natural Resources Research*, 34(3), 1219-1241. [CrossRef]
- [44] Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1), 1-10. [CrossRef]

- [45] Higham, N. J. (1986). Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4), 1160-1174. [CrossRef]
- [46] Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2), 139-160. [CrossRef]
- [47] Zhao, Z., Qiao, K., Liu, Y., Chen, J., & Li, C. (2022). Geochemical data mining by integrated multivariate component data analysis: The Heilongjiang Duobaoshan area (China) case study. *Minerals*, 12(8), 1035. [CrossRef]
- [48] Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3), 253-278. [CrossRef]
- [49] Filzmoser, P., Hron, K., & Reimann, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics: The Official Journal of the International Environmetrics Society*, 20(6), 621-632. [CrossRef]
- [50] Barnes, R. J. (1991). The variogram sill and the sample variance. *Mathematical geology*, 23(4), 673-678. [CrossRef]
- [51] Reimann, C., Filzmoser, P., & Garrett, R. G. (2002). Factor analysis applied to regional geochemical data: problems and possibilities. *Applied geochemistry*, 17(3), 185-206. [CrossRef]
- [52] Zuo, R. (2017). Machine learning of mineralization-related geochemical anomalies: A review of potential methods. *Natural Resources Research*, 26(4), 457-464. [CrossRef]
- [53] Puchhammer, P., Kalubowila, C., Braus, L., Pospiech, S., Sarala, P., & Filzmoser, P. (2024). A performance study of local outlier detection methods for mineral exploration with geochemical compositional data. *Journal of Geochemical Exploration*, 258, 107392. [CrossRef]
- [54] Liu, Y., Carranza, E. J. M., Zhou, K., & Xia, Q. (2019). Compositional Balance Analysis: An Elegant Method of Geochemical Pattern Recognition and Anomaly Mapping for Mineral Exploration. *Natural Resources Research*, 28(4), 1269-1283. [CrossRef]
- [55] Zhang, S., Carranza, E. J. M., Fu, C., Zhang, W., & Qin, X. (2024). Interpretable machine learning for geochemical anomaly delineation in the Yuanbo Nang District, Gansu Province, China. *Minerals*, 14(5), 500. [CrossRef]
- [56] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35(3), 279-300. [CrossRef]



Yupeng Li received the Ph.D. degree in mining and petroleum engineering from the University of Alberta, Edmonton, AB, Canada, in 2011. Since 2014, he has been a Geologic Specialist with the Aramco Beijing Research Center, Saudi Aramco. His research interests include quantitative geologic modeling, reservoir characterization, and artificial intelligence applications in geoscience. (Email: yupeng.li@aramcoasia.com)



Saeed S. Shahrani received the B.Sc. degree in geophysics from King Abdulaziz University, Jeddah, Saudi Arabia, in 2001, and the M.Sc. degree in geology from Bowling Green State University, Ohio, USA, in 2013. His professional and research interests include wellsite geology, biosteering, quantitative geoscience, geochemical data analysis, and applications of artificial intelligence and machine learning to geological interpretation and operational decision-making. (Email: saeed.shahrani.20@aramco.com)



Zhaoxian Yuan received the Ph.D. degree in mineral resource prospecting and exploration from China University of Geosciences, Wuhan, China, in 2014. Since 2016, he has been a researcher with Hebei GEO University. His research interests include applied geochemistry and quantitative prediction of mineral resources. (Email: yuanzhaoxian@hgu.edu.cn)