Comparative Machine Learning-Based Prediction of Gold Enrichment in a Sulphide-Hosted Orogenic System Using Multielement Geochemistry

Gilbert Yaw Bimpong *, Justina Senam Lotsu , Kwaku Boakye

Posted Date: 21 May 2026

doi: 10.20944/preprints202605.1411.v1

Keywords: machine learning; SHAP; geochemical modelling; sulphide-hosted systems; orogenic gold deposits

 doi:10.20944/preprints202605.1411.v1

Article

Comparative Machine Learning-Based Prediction of Gold Enrichment in a Sulphide-Hosted Orogenic System Using Multielement Geochemistry

Gilbert Yaw Bimpong [1,*], Justina Senam Lotsu [2] and Kwaku Boakye [3]

[1] Mining and Mineral Engineering Department, University of Alaska Fairbanks, AK, USA
[2] Mining and Explosives Engineering Department, Missouri University of Science and Technology, MO, USA
[3] Mining Engineering Department, Heidelberg Materials, Fluortown, PA, USA
* Correspondence: gybimpong@alaska.edu; Tel.: +1 (907) 405-2700

Abstract

Accurate prediction of gold enrichment is critical for mineral exploration and resource evaluation, particularly in data-limited environments where only geochemical information is available. This study evaluates machine learning (ML) models against linear baselines for predicting relative gold enrichment defined as $\ln(Au/G_{10})$, where G_{10} represents the geometric mean of ten predictor elements (S, Fe, Al, Si, Mn, Sr, Ni, Cu, K, Ti). A total of 53,126 samples from a sulphide-hosted orogenic gold system were subjected to compositional data analysis (CoDA) preprocessing, including multiplicative replacement of below-detection values, closure to a constant sum, centred log-ratio (CLR) transformation of predictor variables, and robust outlier filtering using the Minimum Covariance Determinant (MCD) method. After screening, 41,626 samples were retained for modelling. Comparative modelling included linear baselines—Ordinary Least Squares (OLS), Ridge, Lasso, and Huber—and non-linear ML algorithms: Random Forest (RF), Support Vector Regression (SVR), k-Nearest Neighbours (kNN), and Multi-Layer Perceptron (MLP). Under the non-circular formulation, nonlinear models consistently outperformed linear baselines. Random Forest achieved the strongest validation performance ($R^2 \approx 0.51$), followed closely by MLP and SVR, while linear models remained substantially weaker ($R^2 \approx 0.31$). SHapley Additive exPlanations (SHAP) applied to the Random Forest model identified sulphur (S) and iron (Fe) as the most influential predictors, consistent with sulphide-controlled gold mineralization processes in orogenic systems. The model predicts relative gold enrichment based solely on multielement geochemistry, providing a robust and interpretable proxy for mineralization intensity in data-constrained environments. This study demonstrates that machine learning models, when combined with CoDA-correct preprocessing and a non-circular target formulation, can provide geologically meaningful and methodologically robust predictions of gold enrichment. The workflow offers a transparent and reproducible framework for early-stage exploration targeting using multielement geochemistry.

Keywords: machine learning; SHAP; geochemical modelling; sulphide-hosted systems; orogenic gold deposits

1. Introduction

Orogenic gold deposits (OGDs) are among the most economically significant gold systems globally, particularly within Precambrian terranes such as the West African, Yilgarn, and Superior cratons [1–3]. These structurally controlled, epigenetic systems are typically associated with compressional tectonics in metamorphic belts and are characterized by quartz-carbonate veins containing gold in both visible and invisible forms, often within sulfide minerals such as arsenopyrite, pyrite, and pyrrhotite [4–6]. Pathfinder elements such as As, Sb, Fe, Ni, and Co are strongly associated with gold, enabling more effective geochemical exploration [7,8]. Mineral zoning

 doi:10.20944/preprints202605.1411.v1

in sulphide minerals has also been employed to differentiate mineralized from barren zones, reflecting fluid evolution during ore formation [9,10]. Recent advances have highlighted the importance of integrating multielement geochemistry with machine learning (ML) to better delineate mineralized zones, especially in complex and covered terrains [11–13].

Over the last two decades, the mineral exploration landscape has been transformed by the convergence of improved geoscientific data acquisition, computational capabilities, and ML tools [14,15]. Traditional exploration techniques, although foundational, are increasingly challenged by limitations of cost, time, and subjectivity, especially in geologically complex areas where indicators are obscured [16,17]. Sulphide-hosted OGDs in greenstone belts and shear zones typify this complexity, underscoring the need for predictive modelling in regions like the West African Craton [18,19].

The widespread availability of multielement geochemical datasets from soil and rock sampling campaigns provides a rich resource for ML applications. These algorithms are particularly adept at modelling nonlinear and high-dimensional relationships, offering great promise in mineral prospectivity analysis [20,21]. Emerging analytical technologies such as LA-ICP-MS, portable XRF, and high-resolution elemental mapping have further expanded data access [22,23]. Techniques like Random Forest (RF), Support Vector Regression (SVR), and Convolutional Neural Networks have been applied to anomaly detection and predictive modelling tasks, enabling higher confidence in target delineation [24–26]. Geochemical indices and elemental ratios also continue to support anomaly classification and mineral potential mapping [27,28].

ML has enabled a shift from deterministic to probabilistic modelling, improving flexibility and accuracy in mineral systems analysis. Ensemble learning strategies, including bagging and boosting, are now widely used and recognized for superior predictive power in geoscientific contexts [29,30]. Nonetheless, many studies emphasize advanced algorithms without benchmarking against elementary baselines such as linear regression models, limiting the ability to evaluate true ML gains under consistent preprocessing and geological settings [31,32].

Geochemical datasets pose unique challenges as they are inherently compositional, constrained by closure to a constant sum, and thus require log-ratio transformations to avoid spurious correlations [33]. Without compositional data analysis (CoDA), interpretations may lack subcompositional coherence and risk misleading geological conclusions [34]. Multicollinearity, redundancy, and sparsity in geochemical data further complicate modelling tasks [35,36].

While integrating geophysical and structural datasets into predictive models can improve accuracy, these datasets are often unavailable in early-stage exploration, making it essential to assess the predictive utility of geochemical data on its own [37,38]. The demand for interpretable machine learning solutions has increased the adoption of explainable artificial intelligence (XAI), with SHapley Additive exPlanations (SHAP) emerging as a leading method. SHAP enhances model transparency by assigning importance scores to each feature, thereby linking model outcomes to geological processes [39–43].

In mineral exploration, SHAP has helped validate pathfinder elements including Al_2O_3, MgO, Sr, S, Fe, and As, particularly in sulphide-rich systems [35,39]. It has also been employed in unsupervised learning workflows and anomaly detection, enabling the delineation of lithological boundaries and mineralized zones [44,45]. Although computationally intensive, SHAP represents a significant breakthrough in the fusion of ML with geological understanding [46,47].

This study responds to the need for robust, interpretable, and geology-informed predictive models in mineral exploration by: (1) benchmarking linear baselines—Ordinary Least Squares (OLS), Ridge, Lasso, and Huber—against non-linear ML algorithms (RF, SVR, kNN, and MLP) within a CoDA-correct preprocessing workflow; (2) applying SHAP to interpret RF outputs and identify key geochemical predictors; and (3) ensuring reproducibility by publishing full code, hyperparameter grids, and a synthetic proxy dataset. By tackling these objectives, this research contributes to best practices in model selection, preprocessing, and interpretation in ML-based geochemical targeting.

 doi:10.20944/preprints202605.1411.v1

2. Materials and Methods

This study employed a structured compositional data analysis (CoDA) framework following Aitchison [48], in which the response variable was defined as $y = \ln(Au/G_{10})$, where G_{10} is the geometric mean of the predictor elements. This formulation represents gold enrichment relative to the multielement geochemical background and ensures that the predictor variables are independent of Au, thereby eliminating circularity and enabling valid application to unseen samples [49,50]. The computational workflow was implemented in Python 3.10 within a Jupyter Notebook environment. The pipeline incorporated systematic preprocessing of geochemical assays using a CoDA-correct approach (multiplicative replacement of below-detection values, closure to constant sum, and centered log-ratio transformation with robust outlier screening), ensuring statistically valid treatment of compositional data.

Modelling included both elementary baselines (Ordinary Least Squares, Ridge, Lasso, Huber) and advanced non-linear algorithms (Random Forest, Support Vector Regression, k-Nearest Neighbors, and Multi-Layer Perceptron) to benchmark ML performance gains. A rigorous experimental design with stratified splits and five-fold cross-validation was adopted to guarantee robust evaluation. Performance metrics (R^2, RMSE, MAE) were compared across baselines and non-linear models, while SHAP analysis provided geological interpretability by linking predictive features to sulphide-associated mineralization.

2.1. Study Area and Data Description

The dataset comprises 53,126 multielement geochemical samples collected from a sulphide-hosted orogenic gold system within a Paleoproterozoic Birimian greenstone belt. The geological setting is characterized by structurally controlled gold mineralization associated with sulphide phases, particularly pyrite and arsenopyrite, consistent with typical orogenic gold deposit models.

Each sample contains concentrations of Au and a suite of major and trace elements, including S, Fe, Al, Si, Mn, Sr, Ni, Cu, K, and Ti. These elements were selected based on their geological relevance to mineralization processes, association with sulphide phases, and data completeness across the dataset. The final dataset provides a robust multivariate geochemical representation suitable for machine learning-based predictive modelling. A summary of the dataset is shown in Table 1.

Table 1. Raw compositional summaries of geochemical data before CoDA preprocessing.

Element	Min (ppm)	Max (ppm)	Mean (ppm)	Std Dev (ppm)	25th (ppm)	Median (ppm)	75th (ppm)
Au	0.01	2,240	0.82	11.80	0.01	0.02	0.09
Al	1,300	139,000	70,103	13,786	63,000	71,000	79,000
Cu	5	5,790	49.95	51.22	30	40	50
Fe	15	170,000	41,909	12,345	35,700	42,600	48,700
K	300	46,100	18,856	5,283	16,000	18,900	21,900
Mn	15	62,500	818.46	1,482.43	460	560	670
Ni	10	2,530	36.23	24.86	20	40	50
S	100	440,000	3,518	19,330	500	1,200	3,900
Si	1,150	550,000	238,281	42,271	216,000	233,000	253,000
Sr	2.5	1,100	257.01	80.74	213	246	287
Ti	50	22,000	3,580	1,105	3,300	3,700	4,000

2.1.1. Sample Preparation and Analytical Methods

All drillhole samples were prepared using industry-standard laboratory protocols. Sample preparation included crushing to <2 mm followed by pulverization to approximately 75 µm to ensure analytical homogeneity. Gold (Au) was analysed using fire assay with atomic absorption spectrometry (AAS) finish. Major and trace elements (S, Fe, Al, Si, Mn, Sr, Ni, Cu, K, Ti) were analysed

 doi:10.20944/preprints202605.1411.v1

using X-ray Fluorescence (XRF) following multi-acid digestion. All analyses were conducted at a single accredited commercial laboratory using consistent analytical methods and detection limits throughout the dataset.

2.1.2. Rationale for Element Selection

Element selection was guided by geochemical relevance, data completeness, and analytical consistency. The selected elements (S, Fe, Al, Si, Mn, Sr, Ni, Cu, K, Ti) are commonly associated with sulphide-hosted orogenic gold systems and exhibited low censoring rates (<25%) across the dataset. Sulphur and iron were specifically retained as key proxies for sulphide abundance, given their strong mineralogical association with gold-bearing phases such as pyrite and arsenopyrite. Although arsenic (As) is a well-known pathfinder element in orogenic gold systems, it was excluded due to high censoring rates and inconsistent detection limits within the available dataset. The retained element suite captures the dominant sulphide-controlled geochemical signal while ensuring CoDA robustness.

2.2. Data Preprocessing

Geochemical data are inherently compositional, meaning that only relative relationships between components carry meaningful information. To ensure statistically valid analysis and avoid spurious correlations, preprocessing was conducted within a CoDA framework.

2.2.1. Treatment of Below-Detection-Limit Values

Below-detection-limit (BDL) and zero values among the predictor elements were treated using a multiplicative replacement strategy. A small constant ($\delta = 1\times10^{-6}$ of the row total) was distributed proportionally across components, ensuring strictly positive values while preserving the relative structure of the data.

2.2.2. Closure and CLR Transformation

To account for the compositional constraint, predictor-element vectors (S, Fe, Al, Si, Mn, Sr, Ni, Cu, K, Ti) were closed to a constant sum and transformed using the centred log-ratio (CLR) transformation as used by Aitchison [48]: $CLR(x_i) = \ln(x_i/g(x))$, where $g(x)$ is the geometric mean of the predictor-element vector. Gold (Au) was excluded from this transformation to avoid circularity in the predictive framework.

2.2.3. Outlier Detection

Outliers were identified in the transformed predictor space using the Minimum Covariance Determinant (MCD) estimator. Mahalanobis distances were computed and compared against a chi-square threshold at $\alpha = 0.999$. Samples exceeding this threshold were removed, resulting in a filtered dataset of 41,626 samples. This procedure ensures robust modelling by removing multivariate anomalies without biasing high-grade mineralization trends.

2.3. Target Variable Definition

To eliminate circularity, the response variable was defined independently of the predictor transformation as $y = \ln(Au/G_{10})$, where G_{10} is computed from the predictor elements only. This formulation ensures that Au is not embedded within the predictor space and allows the model to be applied to unseen samples using only measured predictor-element concentrations. The formulation represents gold enrichment relative to the multielement geochemical background and provides a methodologically sound and non-circular predictive framework.

2.4. Exploratory Compositional Analysis and Predictor Selection

 doi:10.20944/preprints202605.1411.v1

Before model development, exploratory compositional data analysis was conducted to evaluate relationships among predictor elements and justify their inclusion. A CLR-based correlation matrix was examined to assess inter-element associations and potential redundancy within the predictor set as shown in Figure 1. Additionally, geological relevance was considered, particularly the role of sulphur and iron as key indicators of sulphide mineralization and gold deposition. This approach preserves the compositional structure of the data and avoids spurious correlations associated with raw concentrations.

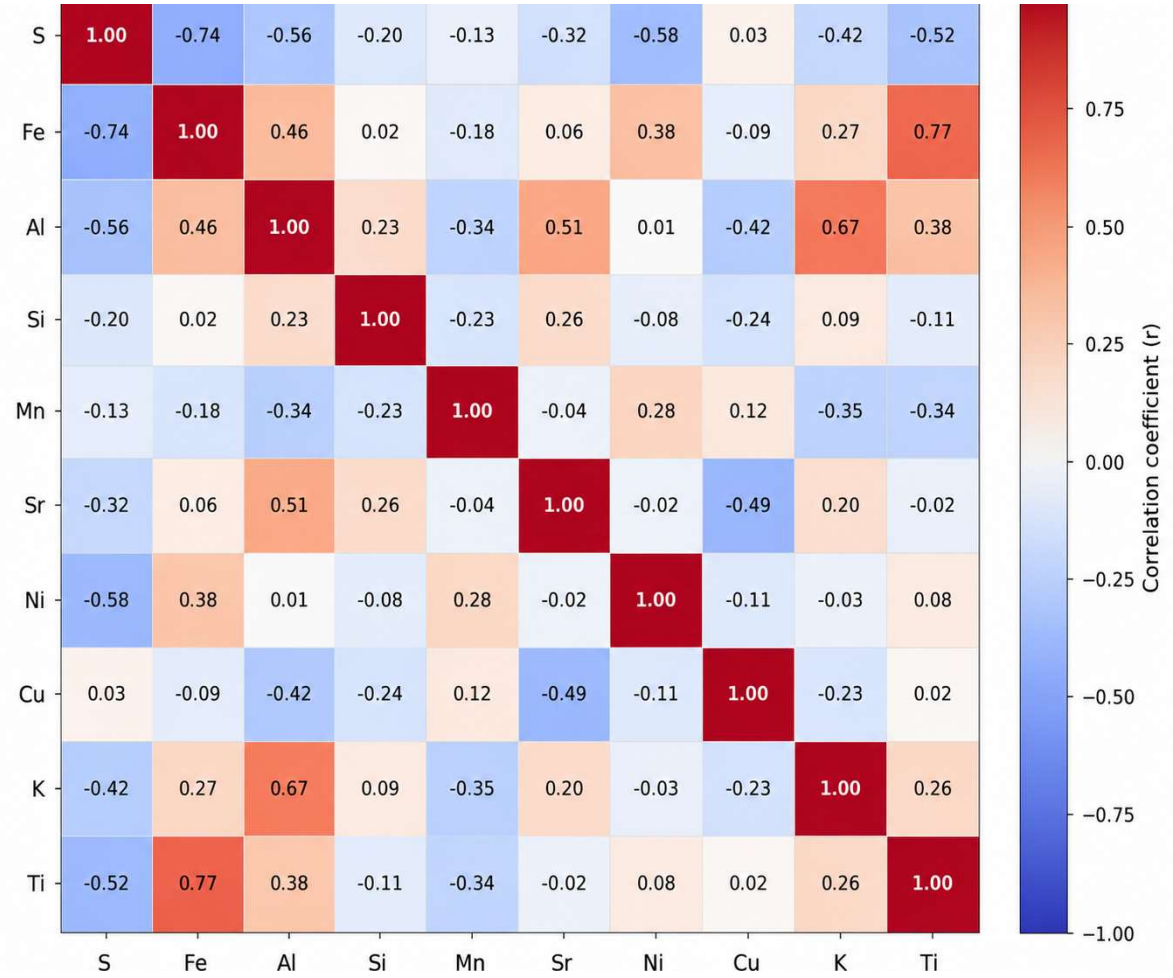

Figure 1. CLR-based correlation matrix of predictor elements for gold enrichment modelling.

The CLR-based correlation matrix (Figure 1) reveals distinct geochemical associations among predictor elements. Strong relationships between Fe and Ti, as well as Al and K, reflect lithological controls, while associations involving sulphur (S) highlight sulphide-related mineralization processes. The absence of extreme redundancy among variables indicates that the selected predictor set retains meaningful compositional information suitable for machine learning modelling.

To further investigate compositional structure and validate element selection, principal component analysis (PCA) was performed on CLR-transformed predictor variables. The resulting biplot in Figure 2 provides a visual representation of inter-element relationships and their contribution to variance within the dataset.

 doi:10.20944/preprints202605.1411.v1

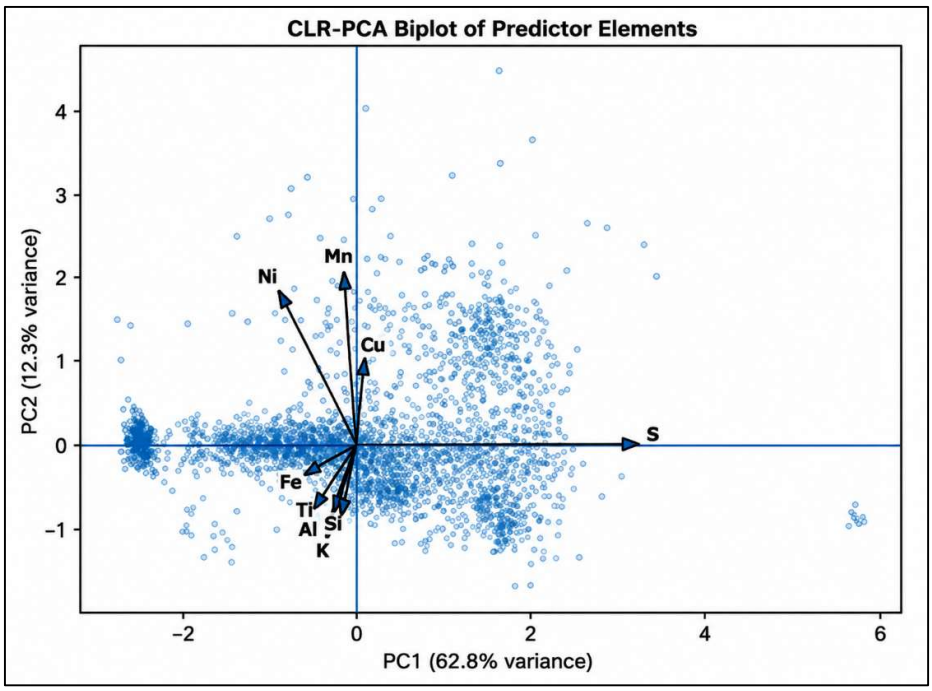

Figure 2. CLR-PCA biplot of predictor elements.

The CLR-PCA biplot (Figure 2) explains approximately 75% of the total variance, with PC1 accounting for 62.8% and PC2 for 12.3%. Sulphur (S) shows a strong loading along PC1, indicating its dominant influence on compositional variability and its association with sulphide mineralization. The distribution of variables confirms the presence of meaningful compositional patterns and supports the validity of the selected predictor elements for modelling.

2.5. Machine Learning Models

A comparative modelling framework was implemented to evaluate the performance of machine learning algorithms against elementary baselines.

2.5.1. Linear Baseline Models

Linear regression models were used as baseline comparators, including: Ordinary Least Squares (OLS), Ridge regression, Lasso regression, and Huber regression. These models provide a reference for assessing whether nonlinear approaches offer improved predictive capability.

2.5.2. Nonlinear Machine Learning Models

The following nonlinear models were evaluated: Random Forest (RF), Support Vector Regression (SVR), k-Nearest Neighbours (kNN), and Multi-Layer Perceptron (MLP). These models were selected due to their ability to capture nonlinear relationships within complex geochemical systems.

2.6. Model Training and Evaluation

The dataset was randomly partitioned into training (70%), testing (15%), and validation (15%) subsets. Model performance was evaluated using the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). All models were trained to predict the revised target variable $\ln(Au/G_{10})$. Performance metrics were computed for training, testing, and validation sets, with emphasis placed on validation results to assess generalization performance.

 doi:10.20944/preprints202605.1411.v1

2.7. Reconstruction of Gold Concentration

For interpretational completeness, the relationship between the predicted target variable and gold concentration can be expressed as: $Au = \exp(y) \times G_{10}$, where G_{10} is calculated directly from the predictor-element concentrations. However, model evaluation and interpretation in this study are performed exclusively in the transformed target space $\ln(Au/G_{10})$, ensuring methodological consistency and avoiding reliance on derived quantities for validation.

2.8. Model Interpretability Using SHAP

Model interpretability was assessed using SHapley Additive exPlanations (SHAP), applied to the best-performing model (Random Forest). SHAP values quantify the contribution of each predictor variable to the predicted output, allowing identification of key geochemical drivers of gold enrichment. In this study, SHAP analysis was conducted on the revised target variable $\ln(Au/G_{10})$, ensuring that feature importance reflects valid predictive relationships rather than artefacts of compositional circularity [39].

3. Results

3.1. Model Performance in Revised Target Space

All models were retrained using the revised non-circular target variable $\ln(Au/G_{10})$, where G_{10} is the geometric mean of the predictor elements. Under this corrected formulation, nonlinear machine learning models consistently outperformed linear baselines. Random Forest (RF) achieved the best validation performance with $R^2 = 0.505$, RMSE = 1.264, and MAE = 1.001, followed closely by the Multi-Layer Perceptron (MLP) and Support Vector Regression (SVR). In contrast, linear models—including Ordinary Least Squares (OLS), Ridge, Lasso, and Huber regression—showed substantially lower performance, with validation R^2 values around 0.31, as shown in Table 2.

Table 2. Performance comparison of machine learning models in the revised target space $\ln(Au/G_{10})$. Validation metrics are used for model comparison. ΔR^2 indicates improvement relative to the OLS baseline.

Model	Validation R^2	Validation RMSE	Validation MAE	ΔR^2 vs OLS
RF	0.505	1.264	1.001	+0.198
MLP	0.505	1.264	0.999	+0.198
SVR	0.479	1.297	0.994	+0.172
kNN	0.476	1.300	1.028	+0.169
OLS	0.307	1.495	1.196	0.000
Ridge	0.307	1.495	1.196	0.000
Lasso	0.307	1.495	1.196	0.000
Huber	0.306	1.497	1.191	−0.002

The comparative performance of the models is illustrated in Figure 3, which summarizes validation R^2 across all models.

 doi:10.20944/preprints202605.1411.v1

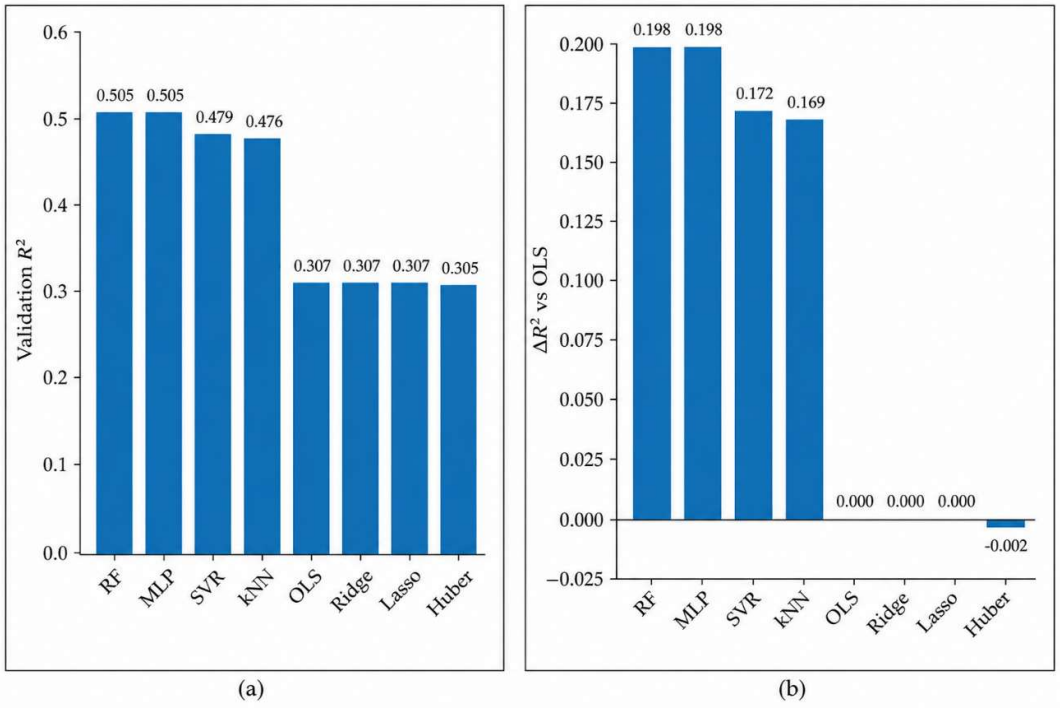

Figure 3. Combined model-performance summary for the corrected workflow predicting ln(Au/G_{10}). (a) Validation R^2 across candidate models. (b) Improvement in validation R^2 relative to OLS, highlighting the advantage of nonlinear learners after circularity was removed.

3.2. Prediction Accuracy in the Revised Target Space

The relationship between observed and predicted values for the RF model is shown in Figure 4. The predictions exhibit a clear positive relationship with the observed values and are distributed close to the 1:1 line, indicating good agreement between model outputs and measured gold enrichment. Although some dispersion is present, particularly at higher enrichment values, the model captures the overall structure of the data effectively.

 doi:10.20944/preprints202605.1411.v1

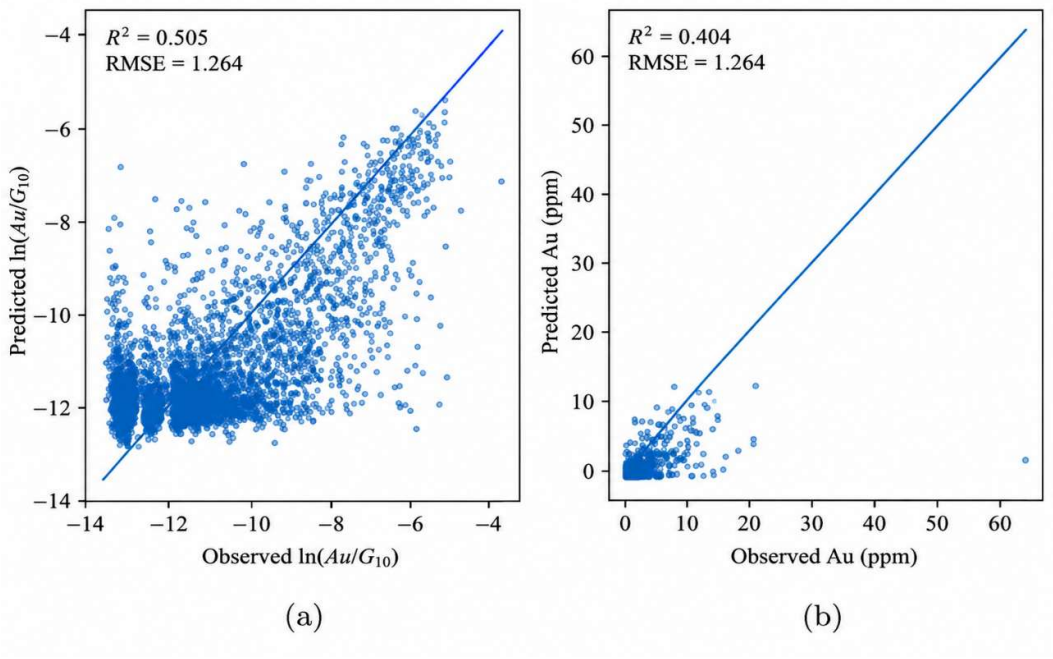

Figure 4. Prediction diagnostics for the best-performing Random Forest model under the corrected framework on the validation dataset. (a) Observed versus predicted ln(Au/G10). (b) Observed versus predicted Au (ppm) after back-transformation to the original concentration scale.

3.3. Residual Behaviour

Residual diagnostics in Figure 5 indicate that errors are approximately symmetrically distributed around zero, with no clear systematic bias across the prediction range. However, an increase in residual spread is observed at higher predicted values, suggesting reduced precision for extreme enrichment conditions.

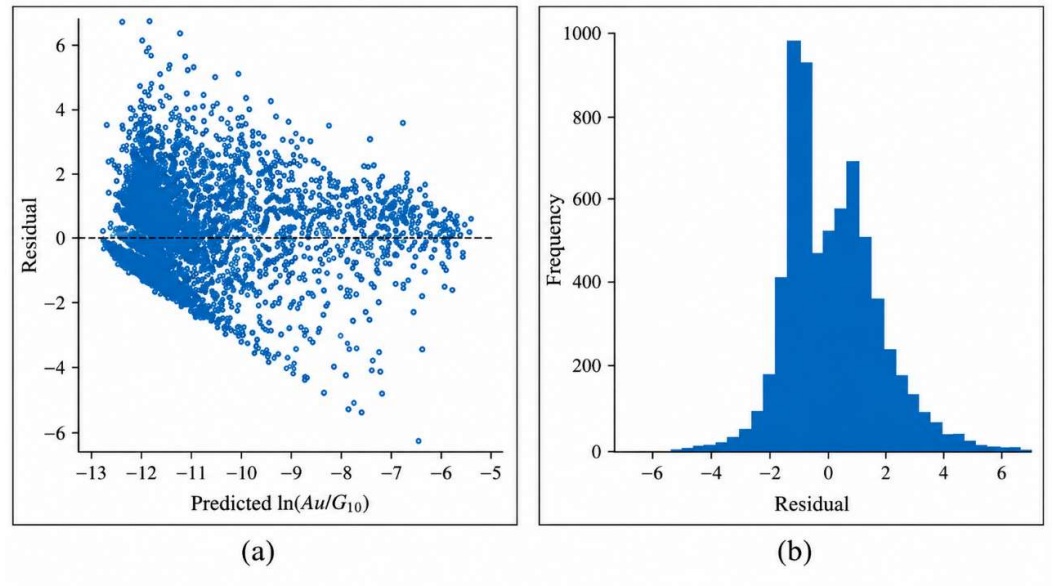

Figure 5. Residual diagnostics for the Random Forest model in the corrected target space. (a) Residuals versus predicted ln(Au/G10). (b) Distribution of residuals on the validation set.

 doi:10.20944/preprints202605.1411.v1

3.4. Model Interpretation Using SHAP

SHapley Additive exPlanations (SHAP) were used to evaluate the contribution of predictor variables to model predictions. The SHAP summary plot in Figure 6 shows that sulphur (S) and iron (Fe) are the most influential predictors, with higher values of these elements generally associated with increased predicted gold enrichment. Secondary contributions are observed from elements such as Cu, Ni, and Mn. The global importance ranking in Figure 6a confirms the dominant role of S and Fe, with other elements contributing to a lesser extent.

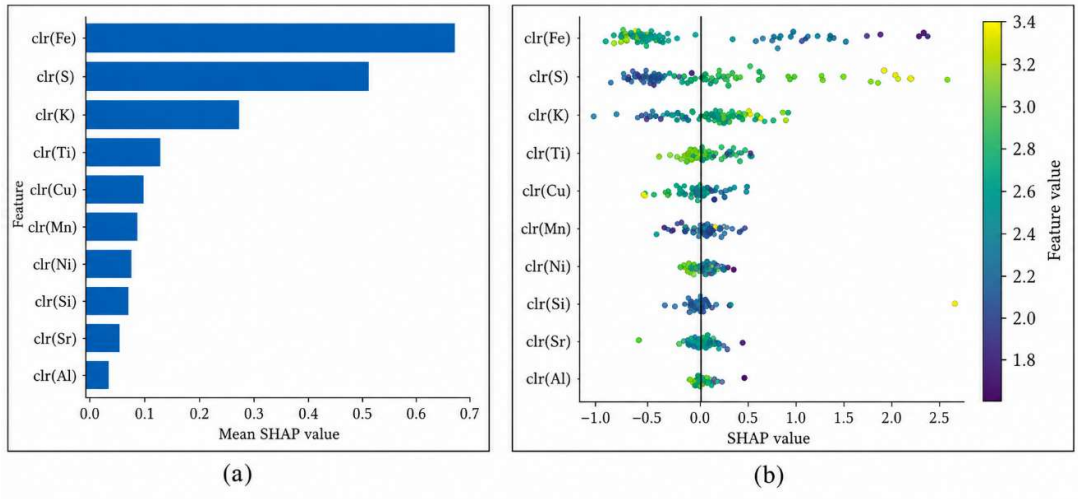

Figure 6. SHAP-based interpretation of a compact Random Forest model trained on the corrected target $\ln(Au/G_{10})$. (a) Global feature-importance ranking based on mean absolute SHAP value. (b) SHAP summary plot showing the direction and magnitude of feature contributions across validation samples.

4. Discussion

The results demonstrate that multielement geochemical data alone can provide meaningful predictive capability for gold enrichment within a sulphide-hosted orogenic system. The best-performing model, Random Forest, achieved a validation R^2 of approximately 0.51, indicating that the selected predictor elements capture key geochemical signals associated with gold mineralization. While this level of performance may be considered moderate, it is consistent with the inherent complexity of geological systems, where mineralization is controlled by multiple interacting factors including lithology, structure, hydrothermal fluid flow, and alteration processes.

The consistent outperformance of nonlinear models relative to linear baselines highlights the importance of capturing nonlinear relationships within geochemical datasets. Linear models assume additive and proportional relationships between predictor variables and the response, which are often insufficient to represent the complexity of geochemical processes. In contrast, nonlinear models such as Random Forest, MLP, and SVR are capable of modelling interactions and non-additive effects among elements, allowing them to better represent the underlying geochemical system. The observed improvement in validation performance, particularly the approximately 0.20 increase in R^2 of Random Forest over OLS, demonstrates the advantage of nonlinear learning approaches for mineral exploration applications.

The application of SHAP provides valuable insight into the geochemical controls on model predictions and has been widely applied in geosciences for interpretable modelling and uncertainty-aware prediction [42,43]. The dominance of sulphur (S) and iron (Fe) as primary predictors is consistent with the geological setting of sulphide-hosted orogenic gold systems, where gold is commonly associated with sulphide minerals such as pyrite and arsenopyrite. Elevated concentrations of S and Fe likely indicate zones of increased sulphide abundance, which are often

 doi:10.20944/preprints202605.1411.v1

correlated with gold mineralization. Secondary contributions from elements such as Cu, Ni, and Mn may reflect additional geochemical processes, including trace metal substitution within sulphide minerals, hydrothermal alteration, and variations in host-rock composition.

The proposed workflow has practical implications for mineral exploration, particularly in early-stage or data-limited environments. By relying solely on multielement geochemical data, the approach provides a cost-effective method for identifying zones of potential gold enrichment without requiring extensive geological, structural, or geophysical data. The interpretability provided by SHAP supports decision-making by linking model predictions to geologically meaningful variables.

Despite its strengths, the model exhibits several limitations. The increase in prediction dispersion at higher enrichment levels suggests reduced precision in extreme conditions, which may be influenced by data imbalance, analytical uncertainty, or localized geological variability. Furthermore, the use of random data partitioning does not explicitly account for spatial dependencies within the dataset. In practice, spatial autocorrelation can influence model performance, and future studies should consider spatially aware validation strategies, such as block cross-validation, to better reflect real-world prediction scenarios. Additionally, the current model is limited to geochemical variables and does not incorporate structural, lithological, or geophysical information, which are known to influence mineralization processes.

Future work should focus on integrating multielement geochemical data with spatial and geological information to enhance predictive accuracy and robustness. The application of spatial machine learning methods, including geostatistical learning and spatial cross-validation, may provide improved generalization in exploration contexts. Expanding the framework to other deposit types and geological settings would also provide insight into its broader applicability.

5. Conclusions

This study presents a machine learning-based framework for predicting relative gold enrichment in a sulphide-hosted orogenic system using multielement geochemical data. By integrating compositional data analysis (CoDA) with a non-circular target formulation defined as $\ln(Au/G_{10})$, where G_{10} represents the geometric mean of selected predictor elements, the methodology ensures statistical validity and eliminates dependency between predictor variables and the response.

The comparative evaluation of models demonstrated that nonlinear machine learning algorithms consistently outperform linear baselines in capturing the complexity of geochemical relationships. Among the models tested, Random Forest achieved the highest validation performance, confirming its suitability for modelling multivariate geochemical systems characterized by nonlinear interactions and element associations.

Interpretability analysis using SHAP revealed that sulphur (S) and iron (Fe) are the most influential predictors of gold enrichment, consistent with the well-established association of gold with sulphide minerals such as pyrite and arsenopyrite in orogenic systems. Secondary contributions from elements such as Cu, Ni, and Mn further highlight the role of additional geochemical processes, including trace metal substitution, hydrothermal alteration, and host-rock variability.

It is important to note that the compositional relationships and model outcomes are specific to the selected element subset and geological context. Future work should evaluate the robustness of model performance across alternative element subsets and compositional transformations (e.g., isometric log-ratio transformations) to further assess generalizability. Overall, this study demonstrates that machine learning models, when combined with appropriate compositional data preprocessing and a rigorously defined target variable, can provide reliable and geologically interpretable predictions of gold enrichment.

Author Contributions: Conceptualization, G.Y.B.; methodology, G.Y.B an JSL.; software, G.Y.B.; formal analysis, G.Y.B., J.S.L., and K.B.; investigation, G.Y.B., J.S.L., and K.B.; data curation, G.Y.B.; writing—original draft preparation, G.Y.B.; writing—review and editing, G.Y.B., J.S.L., and K.B.; visualization, G.Y.B.; supervision, K.B. All authors have read and agreed to the published version of the manuscript.

 doi:10.20944/preprints202605.1411.v1

Funding: This research received no external funding.

Data Availability Statement: All code, hyperparameter grids, and a synthetic proxy dataset supporting this study are openly available on Zenodo (Version 1.0): https://doi.org/10.5281/zenodo.19412270. The original assay data cannot be shared due to confidentiality agreements, but the proxy dataset reproduces the dimensionality, compositional structure, and Au–S/Fe associations, enabling full replication of preprocessing, model training, and evaluation.

Acknowledgments: The authors acknowledge the use of Grammarly for grammatical and spelling corrections during manuscript preparation. The authors reviewed and edited the content and take full responsibility for the content of the published article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
RF	Random Forest
SVR	Support Vector Regression
MLP	Multi-Layer Perceptron
kNN	k-Nearest Neighbours
OLS	Ordinary Least Squares
SHAP	SHapley Additive exPlanations
CoDA	Compositional Data Analysis
CLR	Centered Log-Ratio
MCD	Minimum Covariance Determinant
PCA	Principal Component Analysis
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
XAI	Explainable Artificial Intelligence
AAS	Atomic Absorption Spectrometry
XRF	X-ray Fluorescence
LA-ICP-MS	Laser Ablation Inductively Coupled Plasma Mass Spectrometry
OGD	Orogenic Gold Deposit
BDL	Below Detection Limit

References

1. Groves, D.I.; Santosh, M.; Zhang, L. A scale-integrated exploration model for orogenic gold deposits based on a mineral system approach. Geoscience Frontiers 2020, 11, 719–738. https://doi.org/10.1016/j.gsf.2019.12.007
2. Perret, J.; Jessell, M.W.; Masurel, Q.; et al. Review of Paleoproterozoic tectonics in the southern West African Craton: Insights from multi-disciplinary data integration. Precambrian Research 2025, 422, 107707. https://doi.org/10.1016/j.precamres.2025.107707
3. Djagre, L.; Ali, K.; Kra, L.K.; Koffi, B.G. Geological controls on gold mineralization of the Nyangboué prospect in the southern part of the Boundiali-Syama belt, northwest Ivory Coast. Scientific African 2025, 27, e02584. https://doi.org/10.1016/j.sciaf.2025.e02584
4. Beaudin, D.; Partin, C.A.; Ansdell, K.; Yang, P. Comparative lithology and alteration mineral chemistry of host rocks at the Seabee Gold Operation. Ore Geology Reviews 2024, 166, 105950. https://doi.org/10.1016/j.oregeorev.2024.105950
5. Naumov, E.; Kalinin, Y.; Palyanova, G.; et al. Combined study of Au-bearing arsenopyrite of orogenic gold deposits (NE Asia). Geoscience Frontiers 2025, 16, 101953. https://doi.org/10.1016/j.gsf.2024.101953
6. Rusk, B. Cathodoluminescent Textures and Trace Elements in Hydrothermal Quartz. In Springer Geology 2012, pp. 307–329. https://doi.org/10.1007/978-3-642-22161-3_14

 doi:10.20944/preprints202605.1411.v1

7. Adama, A.; Eric, B.E.; Bertrant, B.S.; et al. Geochemical dataset of laterites soils in Koubou gold district (Zone A) East Cameroon. Data in Brief 2024, 57, 111039. https://doi.org/10.1016/j.dib.2024.111039
8. Campos, L.M.; Toledo, C.L.B.; Silva, A.M.; et al. The hydrothermal footprint of the Crixás deposit. Ore Geology Reviews 2022, 146, 104925. https://doi.org/10.1016/j.oregeorev.2022.104925
9. Chen, B.; Zuo, Y.; Zheng, L.; et al. Relationship between silicification and gold mineralization. Ore Geology Reviews 2025, 176, 106394. https://doi.org/10.1016/j.oregeorev.2024.106394
10. Li, J.; Yang, Z.M.; Wang, C.W.; et al. Metallogeny of the Xiaotongjiapuzi gold deposit. Ore Geology Reviews 2023, 157, 105455. https://doi.org/10.1016/j.oregeorev.2023.105455
11. Ge, Y.Z.; Zhang, Z.J.; Zhou, Y.Z.; et al. Explainable machine learning reveals apatite fertility and porphyry copper mineralization. Ore Geology Reviews 2025, 183, 106679. https://doi.org/10.1016/j.oregeorev.2025.106679
12. Boadi, B.; Raju, P.S.V.; Wemegah, D.D. Lode-gold prospectivity mapping in the Ahafo gold district. Ore Geology Reviews 2022, 148, 105059. https://doi.org/10.1016/j.oregeorev.2022.105059
13. Behera, R.C.; Singh, S.; Srivastava, S.; et al. Trace elemental systematics of auriferous sulfides in dolerites. Ore Geology Reviews 2025, 180, 106569. https://doi.org/10.1016/j.oregeorev.2025.106569
14. Davies, R.S.; Trott, M.; Georgi, J.; Farrar, A. AI and ML to enhance critical mineral deposit discovery. Geosystems and Geoenvironment 2025, 4, 100361. https://doi.org/10.1016/j.geogeo.2025.100361
15. Ahmed, A.A.; Sayed, S.; Abdoulhalik, A.; et al. Applications of machine learning to water resources management. Journal of Cleaner Production 2024, 441, 140715. https://doi.org/10.1016/j.jclepro.2024.140715
16. Hansen, T.F.; Erharter, G.H.; Liu, Z.; Torresen, J. ML approaches for rock mass classification. Applied Computing and Geosciences 2024, 24, 100199. https://doi.org/10.1016/j.acags.2024.100199
17. Mantilla-Dulcey, A.; Goyes-Peñafiel, P.; Báez-Rodríguez, R.; Khurama, S. Porphyry-type mineral prospectivity mapping. Gondwana Research 2024, 136, 236–250. https://doi.org/10.1016/j.gr.2024.09.004
18. Zou, X.; et al. Ore fluid pathways at the giant Lannigou Carlin-type gold deposit. Ore Geology Reviews 2025, 179, 106523. https://doi.org/10.1016/j.oregeorev.2025.106523
19. Sumail, T.; Thébaud, N.; Masurel, Q.; et al. Temporal constraints on gold mineralisation at the Jundee deposit. Precambrian Research 2024, 410, 107479. https://doi.org/10.1016/j.precamres.2024.107479
20. Liu, J.; Bao, X.; Kou, S.; et al. LA-ICP-MS U-Pb geochronology of monazite in the Xinjiazui gold deposit. Ore Geology Reviews 2023, 161, 105626. https://doi.org/10.1016/j.oregeorev.2023.105626
21. Mahboob, M.A.; Celik, T.; Genc, B. Predictive modelling of mineral prospectivity using ML. Remote Sensing Applications: Society and Environment 2024, 36, 101316. https://doi.org/10.1016/j.rsase.2024.101316
22. Xue, X.F.; Feng, Y.C.; Tamer, M.T.; et al. Comparison of gold precipitation processes between disseminated and quartz vein ores of orogenic gold deposits: insights from the Linglong gold field, Jiaodong Peninsula, China. Ore Geology Reviews 2025, 183, 106639. https://doi.org/10.1016/j.oregeorev.2025.106639
23. Liang, Y.; Xue, W.; Li, L.; et al. Multi-stage evolution of a gold mineralization from southern China: Implications for the ore-forming processes. Ore Geology Reviews 2025, 181, 106618. https://doi.org/10.1016/j.oregeorev.2025.106618
24. Chehreh Chelgani, S.; Nasiri, H.; Alidokht, M. Interpretable modeling of metallurgical responses for an industrial coal column flotation circuit by XGBoost and SHAP. International Journal of Mining Science and Technology 2021, 31, 1135–1144. https://doi.org/10.1016/j.ijmst.2021.10.006
25. Antonini, A.S.; Tanzola, J.; Asiain, L.; et al. Machine learning model interpretability using SHAP values: Application to Igneous Rock Classification task. Applied Computing and Geosciences 2024, 23, 100178. https://doi.org/10.1016/j.acags.2024.100178
26. Zhang, S.; Chen, C.; Xu, J.; et al. Deterministic modelling for driving factors of mineralization in Shanggong gold deposit (China). Ore and Energy Resource Geology 2024, 17, 100062. https://doi.org/10.1016/j.oreoa.2024.100062
27. Raič, S.; Molnár, F.; O'Brien, H.; et al. Building geochemical vectors with trace element compositions of sulfides in orogenic gold mineral systems in northern Finland. Journal of Geochemical Exploration 2023, 251, 107252. https://doi.org/10.1016/j.gexplo.2023.107252
28. Zhao, H.; Wang, Q.; Groves, D.I.; et al. Genesis of orogenic gold systems in the Daduhe belt. Ore Geology Reviews 2022, 145, 104861. https://doi.org/10.1016/j.oregeorev.2022.104861

 doi:10.20944/preprints202605.1411.v1

29. Abraham, E.; Usman, A.; Amano, I. Machine learning-based classification of geological structures from magnetic anomaly data. Machine Learning with Applications 2025, 20, 100678. https://doi.org/10.1016/j.mlwa.2025.100678
30. Zhou, S.; Cheng, Z.; Wang, J.; et al. Uncover implicit associations among geochemical elements using machine learning. Ore Geology Reviews 2025, 179, 106506. https://doi.org/10.1016/j.oregeorev.2025.106506
31. Cui, Q.Y.; Li, J.; Cai, W.Y.; et al. Episodic fluid pulses in the Baiyun gold deposit, Liaodong Peninsula, Eastern China. Ore Geology Reviews 2024, 174, 106313. https://doi.org/10.1016/j.oregeorev.2024.106313
32. Ma, G.; Zhao, X.; Mao, Q.; et al. Gold and antimony metallogenic relationship of the Awanda Au (Sb) deposit. Ore Geology Reviews 2025, 176, 106384. https://doi.org/10.1016/j.oregeorev.2024.106384
33. Buccianti, A.; Grunsky, E.C. Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes? Journal of Geochemical Exploration 2014, 141, 1–5. https://doi.org/10.1016/j.gexplo.2014.03.022
34. Filzmoser, P.; Hron, K.; Templ, M. Applied Compositional Data Analysis. With Worked Examples in R. Springer International Publishing 2018. https://doi.org/10.1007/978-3-319-96422-5
35. Yu, P.Y.; Li, C.; Fu, J.N.; et al. LA-ICP-MS/MS Rb-Sr sericite geochronology in orogenic gold deposits. Ore Geology Reviews 2025, 180, 106543. https://doi.org/10.1016/j.oregeorev.2025.106543
36. Nassabeh, M.; You, Z.; Keshavarz, A.; Iglauer, S. Sub-surface geospatial intelligence in carbon storage using ML. Energy 2024, 305, 132086. https://doi.org/10.1016/j.energy.2024.132086
37. Negrello Bergami, G.; de Souza Filho, C.R.; Haddad-Martim, M.P.; Carranza, E.J.M. The multifractal nature of world-class orogenic gold systems in greenstone belts. Ore Geology Reviews 2024, 165, 105909. https://doi.org/10.1016/j.oregeorev.2024.105909
38. Morgan, H.; Elgendy, A.; Said, A.; et al. Enhanced lithological mapping using explainable AI and remote sensing. Computers & Geosciences 2024, 193, 105738. https://doi.org/10.1016/j.cageo.2024.105738
39. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 2017, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874
40. Zhu, D.; Wang, J.; Kuwatani, T.; Tsuchiya, N. ML applications for magmatic-hydrothermal systems: Quartz trace-element insights. Applied Geochemistry 2025, 189, 106431. https://doi.org/10.1016/j.apgeochem.2025.106431
41. Zhu, C.; Liu, Y.; Wang, D.; et al. Exploration of highly stable and efficient lead-free halide perovskite solar cells by ML. Cell Reports Physical Science 2024, 5, 102321. https://doi.org/10.1016/j.xcrp.2024.102321
42. Wang, H.; Wu, Y.; Zhang, Y.; et al. Uncertainty and Explainable Analysis of Machine Learning Model for Reconstruction of Sonic Slowness Logs. Artificial Intelligence in Geosciences 2023, 4, 182–198. https://doi.org/10.1016/j.aiig.2023.11.002
43. Chen, M.; Wang, H. Explainable machine learning model for prediction of ground motion parameters with uncertainty quantification. Chinese Journal of Geophysics 2022, 65, 3386–3404. https://doi.org/10.6038/cjg2022P0428
44. Sharapatov, A.; Saduov, A.; Assirbek, N.; et al. Prediction of rare and anomalous minerals using anomaly detection and ML. Applied Computing and Geosciences 2025, 26, 100250. https://doi.org/10.1016/j.acags.2025.100250
45. Sun, B.; Cui, W.; Liu, G.; et al. A hybrid strategy of AutoML and SHAP for explainable concrete strength prediction. Case Studies in Construction Materials 2023, 19, e02405. https://doi.org/10.1016/j.cscm.2023.e02405
46. Fang, X.; Gu, F.H.; Tang, J.X.; et al. Mesozoic orogenic gold metallogenesis in Tibet. Ore Geology Reviews 2024, 170, 106135. https://doi.org/10.1016/j.oregeorev.2024.106135
47. Dai, Q.Y.; Zhang, L.M.; Zhang, K.; et al. Integrated optimization of reservoir production using ML. Petroleum Science 2025. https://doi.org/10.1016/j.petsci.2025.06.001
48. Aitchison, J. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological) 1982, 44, 139–177. https://www.jstor.org/stable/2345821
49. Pawlowsky-Glahn, V.; Egozcue, J.J.; Tolosana-Delgado, R. Modeling and Analysis of Compositional Data. Wiley 2015. https://doi.org/10.1002/9781119003144

 doi:10.20944/preprints202605.1411.v1

50. Quinn, T.P.; Erb, I.; Gloor, G.; et al. A field guide for the compositional analysis of any-omics data. GigaScience 2019, 8, giz107. https://doi.org/10.1093/gigascience/giz107