This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

Revision 2

Word Count: 7995

Machine Learning Applied to Apatite Compositions for Determining Mineralization Potential

Yu-yu Zheng[1], Bo Xu[1,2,3], David R. Lentz[4], Xiao-yan Yu[1], Zeng-qian Hou[5] and Tao Wang[5]

[1] School of Gemology, China University of Geosciences Beijing, 29 Xueyuan Road, Beijing 100083, China.

[2] State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Beijing 100083, China.

[3] Frontiers Science Center for Deep-time Digital Earth, China University of Geosciences (Beijing), Beijing 100083, China.

[4] Department of Earth Sciences, University of New Brunswick, P.O. Box 4400, Fredericton, NB E3B 5A3, Canada.

[5] Beijing SHRIMP Center, Institute of Geology, Chinese Academy of Geological Sciences, Beijing 100037, China.

Corresponding author: Bo Xu (xubo@outlook.com.cn)

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

ABSTRACT

Apatite major and trace element chemistry is a widely used tracer of mineralization, as it sensitively records the characteristics of the magmatic-hydrothermal system at the time of its crystallization. Previous studies have proposed useful indicators and binary discrimination diagrams to distinguish between apatites from mineralized and unmineralized rocks; however, their efficiency has been found to be somewhat limited in other systems and larger scale datasets. This work applied a machine learning (ML) method to classify the chemical compositions of apatites from both fertile and barren rocks, aiming to help determine the mineralization potential of unknown system. Approximately 13,328 apatite compositional analyses were compiled and labeled from 241 locations in 27 countries worldwide, and three apatite geochemical datasets were established for XGBoost ML model training. The classification results suggest that the developed models (accuracy: 0.851–0.992; F1 score: 0.839–0.993) are much more accurate and efficient than conventional methods (accuracy: 0.242–0.553). Feature importance analysis of the models demonstrates that Cl, F, S, V, Sr/Y, V/Y, Eu*, $(La/Yb)_N$, and La/Sm are important variables in apatite that discriminate fertile and barren host rocks and indicates that V/Y and Cl/F ratios and the S content, in particular, are crucial parameters to discriminating metal enrichment and mineralization potential. This study suggests that ML is a robust tool for processing high-dimensional geochemical data and presents a novel approach that can be applied to mineral exploration.

Keywords: Apatite; Major and Trace Element; Machine Learning; Mineralization Potential; XGBoost

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

INTRODUCTION

Apatite ($Ca_5[PO_4]_3[F,Cl,OH]$) is a ubiquitous accessory mineral in most igneous and metamorphic rocks and derived clastic sediments and is relatively resistant to weathering (O'Sullivan et al., 2020). It is considered to be an ideal indicator mineral, given its chemical composition sensitivity to the crystallization environment (Bruand et al., 2017; Mao et al., 2016). Trace elements and volatile chemistry and isotopic signature of apatites can characterize diverse crystallization environments, including magmatic systems (Cao et al., 2022; Gao et al., 2020; Li et al., 2021; Long et al., 2023; Palma et al., 2019; Qu et al., 2021; Tang et al., 2021; Xu et al., 2023; Zhang et al., 2021), low-grade metamorphic systems (Bea and Montero, 1999; El Korh et al., 2009; Henrichs et al., 2018; Nutman, 2007), and sedimentary environments (Joosu et al., 2016). Accordingly, the trace element chemistry of apatite is widely used to characterize the lithology of source rocks (Belousova et al., 2002), including tracing detrital provenance (Bruand et al., 2017; Dill, 1994; O'Sullivan et al., 2018; O'Sullivan et al., 2020), and used to constrain petrogenetic process (Chu et al., 2009; La Cruz et al., 2020; Sun et al., 2022; Tollari et al., 2008; Zafar et al., 2019), especially for revealing the origin and evolution of magma (Gao et al., 2020; Meng et al., 2021; O'Reilly and Griffin, 2000) . Moreover, the major and trace element chemistry of apatite is applied to mineral exploration (Belousova et al., 2002; Cao et al., 2012; Mao et al., 2016; Sha and Chappell, 1999; Xu et al., 2015). A series of indicators, including Sr/Y, Mn, Eu/Eu*, Th/U, La/Sm, and $(Ce/Yb)_N$ (Belousova et al., 2002), and several binary classification diagrams, such as Sr vs. F (Mn, Y, $(La/Yb)_N$, Eu/Eu*), F/Cl vs. F (Azadbakht et al., 2018; Cao et al., 2012; Zhong et al., 2018), Cl vs. Eu/Eu* (Mao et al., 2016), V/Y vs. REE+Y, Cl vs. SO_3, and $^{87}Sr/^{86}Sr$ vs. Cl/F, are commonly used to diagnose the metallogenic fertility of magmatic rocks. (Xu et al., 2021). Unfortunately, as interest in apatite has recently increased and numerous major

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

and trace element data have been reported (Adlakha et al., 2018; Bruand et al., 2019; Cao et al., 2022; Chakhmouradian et al., 2017; Chen and Zhang, 2018; Chen et al., 2019; Gao et al., 2020; Glorie et al., 2019; Henrichs et al., 2018; Hoshino et al., 2017; La Cruz et al., 2020; Li et al., 2021; Liu et al., 2021; Long et al., 2023; Lupulescu et al., 2017; Meng et al., 2021; Mercer et al., 2020; Palma et al., 2019; Qu et al., 2021; Sun et al., 2022; Tang et al., 2021; Xie et al., 2018; Xu et al., 2023; Yang et al., 2018; Zafar et al., 2019; Zhang et al., 2021), it is challenging to validate these individual indicators and binary discrimination techniques due to a large area overlap of compositional spots, suggesting that those traditional low-dimensional classifiers that seemed to work well in specific systems might be invalid in other systems or datasets of larger scales. Consequently, a novel data processing method that can handle high-dimensional compositional data is imperative for identifying robust indices to aid in exporing various systems for new mineral resources.

The field of machine learning (ML) encompasses the use of computational algorithms to discern patterns within datasets, which can subsequently be applied to make predictions. ML offers a robust toolkit for decoding latent information within high-dimensional data. Over the past few years, there has been an explosion of interest in the applications of ML to solid Earth geoscience (Li et al., 2023). ML has been widely applied in earthquake phase detection and seismicity classification (Cianetti et al., 2021; Linville, 2022), geophysical data processing and image interpretation (Xiao et al., 2021), geophysical inversion (Cai et al., 2022), and multi-physical and multidisciplinary information integration. Given the complexity and diversity of geochemistry data, ML-based classification methods have emerged as a promising approach that outperforms conventional methods, especially in large-scale geological processes, such as in predicting mantle metasomatism worldwide (Qin et al., 2022), revealing source compositions of intraplate

basaltic rocks (Guo et al., 2021), identifying primary water concentrations in mantle pyroxene (Chen et al., 2021), determining the quartz-forming environments (Wang et al., 2021), and classifying the source rocks of detrital zircons (Zhong et al., 2023a, 2023b). In the field of mineral exploration, two studies tried to apply ML to characterize magma fertility based on zircon compositional data, aiming to identify porphyry copper mineralization potential (Zhou et al., 2022; Zou et al., 2022). Tan et al. (2023) employed partial least squares discriminant analysis (PLS-DA) to the apatite trace element dataset (4,298 data) to distinguish between apatites from different types of deposits and rocks. Their plots could not directly discriminate ore magmatic and hydrothermal apatites, but showed a great potential in classifying barren and ore magmatic apatites from granitoid-related deposits and highlighted the role of V, Eu, and Sr for classification.

Here, three global datasets of the major and/or trace element chemistry of apatites were compiled from both mineralized and unmineralized rock samples, and a series of XGBoost models were trained to determine the mineralization potential. The classification results compared with traditional binary diagrams demonstrated an improvement in accuracy and efficiency in discriminating whether apatite is derived from a fertile rock suite or a barren suite. In addition, the feature importance analysis suggested that V/Y and Cl/F ratios and the S content are crucial to metal enrichment and mineralization.

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

DATA COMPILATION AND LABELING

Data Compilation

All of the apatite compositional data used for modeling were collected and compiled from 241 locations in 27 countries worldwide (Figure 1) from preexisting literature. Each location included multiple samples and analyses. This raw dataset (Table S1) contains 13,382 rows of compositional data, including spot analyses and mean values for those references in which spot analyses were not given. Figure 2 shows an overview of the elements and geochemical parameters contained in this dataset. All the data and related sources can be found at https://github.com/YuyuJo/Supplementary-file-for-AM-9115R.

Data Labeling

Analyses of apatites collected from rock samples with obvious mineralization that formed in association with mineral deposits were labeled as "Mineralized." The analyses of apatites in barren rocks were labeled as "Unmineralized." Descriptions regarding the deposit and rock samples (including location, mineralization, and alteration information) can be easily found in the literatures when collecting apatite data. Based on these criteria, 9,104 and 4,278 analyses were labeled as "Mineralized" and "Unmineralized," respectively. The deposit type of each data was also identified based on the classification in Mao et al. (2016). For data with "Mineralized" labels, their deposit types included porphyry (no. = 2,251), skarn (5,075), orogenic Au (875), carbonatite deposits (207), iron oxide Cu–Au (IOCG, 80), Kiruna type (IOA, 579), orogenic Ni–Cu ± platinum group element (28), and epithermal Au–Ag (9).

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

Sub-Dataset Construction

The raw dataset was divided into three subsets to further differentiate the role of major and trace elements. The analyses of samples containing CaO, P_2O_5, SO_3, Cl, and F were selected as "Major" dataset, and the analyses of samples including trace elements were selected as "Trace" dataset. The analyses with both major and trace elements were set into the "Major and Trace" dataset.

To preprocess the collected data, the initial step involved handling the missing values, whereby any element with missing values >60% of the entire column was excluded. After this filtering, the "Major" dataset comprised 5,618 analyses (Table S2). The features therein included CaO, P_2O_5, SO_3, F, Cl, FeO, MnO, Na_2O, SiO_2, and Cl/F. The "Trace" dataset (Table S3) contained 9,979 data and included V, Mn, Rb, Sr, Y, Zr, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, and Lu. Additionally, certain geochemical parameters, which are considered significant for mineralization and magma evolution, were computed and added to the "Trace" dataset. These parameters included LREE, HREE, Sr/Y, V/Y, Ce/Nd, Eu*, Ce^*_N, Eu_N/Eu^*_N, Ce/Ce*, Eu/Eu*/Y, REE+Y, $(La/Yb)_N$, and La/Sm. The "Major and Trace" dataset (Table S4) included 2,448 analyses and 43 features (CaO, P_2O_5, SO_3, F, Cl, FeO, Cl/F, SiO_2, Na_2O, MgO, Rb, Sr, Y, Zr, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Th, U, Sr/Y, V/Y, Ce/Nd, Eu*, Ce^*_N, Eu_N/Eu^*_N, Ce/Ce*, Eu/Eu*/Y, REE+Y, $(La/Yb)_N$, La/Sm, LREE, HREE).

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

METHODS

ML Algorithms

XGBoost is an ML system based on gradient tree boosting, which was originally proposed by Friedman (2001). It has gained widespread recognition in numerous ML and data mining challenges due to its ability to solve real-world-scale problems using minimal resources (Chen and Guestrin, 2016; Python et al., 2021). XGBoost is a distributed gradient boosting library that has been optimized for high efficiency and flexibility. Its flexibility is exemplified by its ability to handle sparse data with multiple possible causes, including missing values and frequent zeros. In addition, its parallel and distributed computing capabilities facilitate faster learning, enabling quicker model exploration. The highly scalable end-to-end tree boosting system allows for efficient scaling to larger datasets with minimal cluster resources (Chen and Guestrin, 2016). Moreover, the tree structure of XGBoost enables the identification of important features and enhances the interpretability of results (Azodi et al., 2020; Qin et al., 2022), which is beneficial in elucidating the relationship between apatite composition and mineralization and exploring the geochemical implications.

XGBoost is an ML algorithm that operates under a gradient boosting framework. Its training methodology is additive, with each new tree added to fit the residuals of the prior predictions. The results of all the trees are summed up to obtain the final predictions. Given a dataset $D = \{(xi, yi)\}$ ($|D| = n, xi \in Rm, yi \in R$) with n examples and m features, the output of a tree ensemble model that uses K additive functions is predicted as a sum of k times scores:

$$\hat{y}_i = \varphi(x_i) = \sum_{k=1}^{K} f_k(X_i), \quad f_k \in \mathcal{F}, \tag{1}$$

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

where $\mathcal{F} = \{f(x) = w_{q(x)}\}$ $(q: \mathbb{R}^m \to T,\ w \in \mathbb{R}^T)$ represents the space of regression trees, the function q denotes the structure of each tree that maps an example to the corresponding leaf index, T is the number of leaves in the tree, each f_k corresponds to an independent tree structure q and leaf weights w, and w_i represents the score on the i-th leaf (Chen and Guestrin, 2016).

The following regularized objective is constructed and minimized to evaluate the quality of a tree structure q:

$$\mathcal{L}(\varphi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (2)$$

where $\Omega(f) = \gamma \mathrm{T} + \frac{1}{2}\lambda\|w\|^2$. The regularization term penalizes the complexity of the model and helps to smooth the final learned weights to avoid overfitting. The parameter γ controls the degree of regularization, while λ controls the strength of the penalty. In this equation, l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i.

Model Construction Processes

A four-step modeling process was employed to construct a classification model that best fitted the apatite compositional data (Figure 3).

Data Preprocess and Splitting. All three sub-datasets were used to train the classification models. Taking the "Trace" dataset as an example, the elemental data were used as input without any transformation. The inputed dataset was first split into "Features" and "Class" subsets, which

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

were uniformed as 0 (Unmineralized) and 1 (Mineralized) using the "LabelEncoder" function. Maintaining the original proportion of each class, both subsets were randomly split into training (80%) and test (20%) sets.

XGBoost Modeling. To avoid overfitting, a fivefold cross-validation (Kohavi, 1995) was employed to train the model. The training set was divided into five folds of equal sizes, where four subsets were used to train the ML model, and the left-out fold was used for validation and classification evaluation. This process was repeated five times, with each validation fold being different, and the output score represented the mean value of all five predictions.

Model Hyperparameter Tuning. A fivefold cross-validation approach was utilized in conjunction with a grid search strategy to optimize the XGBoost model. This strategy exhaustively generated candidates from a grid of parameter values and selected the candidate with the highest output scores, as evaluated by a predefined metric. Specifically, the goal of the grid search procedure was to identify the optimal combination of hyperparameters (eta, gamma, max depth, and alpha) and to generate 3,600 candidates from which the optimal model was selected.

Model Validation and Evaluation. Predictions were obtained by applying the test set to the above optimal XGBoost model. To clearly observe the classification results, the predictions were generally displayed as a confusion matrix (Stehman, 1997), of which the rows represent the true number of each class (from labeled dataset) and the columns display the predicted number of each class. The commonly used classification metrics for evaluating the model performance can be calculated based on the confusion matrix. Here, the accuracy and the F1 score were used as the evaluation indicators of the model. For the convenience of description, true "Mineralized," which was also predicted as "Mineralized," is abbreviated as "MM"; true "Mineralized," which

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

205 was falsely predicted as "Unmineralized," is abbreviated as "MU"; true "Unmineralized," which
206 was also predicted as "Unmineralized," is abbreviated as "UU"; and true "Unmineralized,"
207 which was falsely predicted as "Mineralized," is abbreviated as "UM."

208 Accuracy is a metric that measures the number of correctly predicted cases relative to the total
209 number of samples used. It is calculated as the ratio of the number of correct predictions to the
210 sum of all the utilized samples, which can be expressed as

$$Accuracy = \frac{MM + UU}{MM + MU + UU + UM} \quad (3)$$

211 The F1 score is a measure of classification accuracy that combines precision and recall.
212 Specifically, it is the harmonic mean of precision and recall and is expressed as

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

213 Precision is a measure of the accuracy of predictions, and it indicates the probability that a
214 sample is truly positive among all samples predicted to be positive. Taking class "Mineralized"
215 as an example, precision can be calculated as follows:

$$Precision = \frac{MM}{MM + UM} \quad (5)$$

216 Recall is a measure of how well the classifier identifies the actual positive cases, and it indicates
217 the probability that a sample predicted to be positive is actually positive. Taking class
218 "Mineralized" as an example, recall can be calculated as follows

$$Recall = \frac{MM}{MM + MU} \quad (6)$$

219

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

RESULTS

Classification Results and Feature Importance

A total of 14 XGBoost models were trained using three apatite compositional datasets based on different feature selection. Five models were trained using the "Major and Trace" dataset, and the number of selected features was 43, 35, 22, 12, and 6, respectively (Table 1). Two models were trained using the "Major" dataset. All the ten major elements were applied to train model M-1, while four selected elements to model M-2. The "Trace" dataset was thought to be important to identify mineralization and was thus employed to train seven XGBoost models. The related feature numbers were sequentially set as 33, 28, 21, 14, 7, 3, and 2. The classification results of these XGBoost models are displayed as confusion matrices. Figure 4 shows the results of the four representative models. Table 1 presents the F1 scores and accuracies of 14 models, which were calculated based on the classification confusion matrices.

The relative importance of all the features used in each model was obtained from the XGBoost algorithm to determine the elements of apatite that are highly relevant to mineralization (Table 1). Among all the models for the "Trace" dataset, V, Sr, Y, Eu, Ce, and Rb appeared most frequently in the top ten feature relative importance. Vanadium was particularly important in the rank. Of all five models in which V was selected, the relative importance of V was highest in four and ranked second in the remaining one. Some geochemical parameters also contributed to the rank, including Sr/Y, V/Y, Eu*, $(La/Yb)_N$, and La/Sm. In the two models for the "Major" dataset, the content of SO_3 showed the highest relative importance. However, the proportion of each feature was quite uniform. The features that played a key role in the models for the "Major

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

241 and Trace" dataset were somehow familiar to those in the models for the "Trace" dataset. In
242 addition, Cl, F, and Cl/F were also noteworthy.

243

244 **Feature Selection and ML Model Performance Evaluation**

245 The variance of classification metrics presented in Table 1 has been shown to be related to the
246 input dataset type and the number of selected features in the model. The highest F1 scores of the
247 models are 0.9939 and 0.9933 obtained by models T-1 and M-T-1, respectively, while that of
248 models trained by the "Major" dataset is 0.9259 obtained by model M-1. The accuracies of T-1
249 and M-T-1 are 0.9900 and 0.9918, respectively, while that of M-1 is 0.9386. This indicates that
250 the models trained using the "Trace" and "Major and Trace" datasets achieved better
251 performance than those using the "Major" dataset.

252 The classification results in Table 1 also suggest a positive correlation between feature number
253 and model performance. As displayed in Figure 5, the XGBoost models achieved higher scores
254 when they are trained on more elements and geochemical parameters, which was also observed
255 in other research (Qin et al., 2022). For instance, the accuracy and F1 score increased from
256 0.9146 and 0.8507 for model T-7 (no. of features = 2) to 0.9682 and 0.9474 for model T-5 (no.
257 of features = 7) and 0.9939 and 0.9900 for model T-3 (no. of features = 33).

258 Furthermore, feature selection also showed a salient effect on the model performance, as
259 evidenced by models M-T-4 and T-3. The features selected to train model M-T-4 were almost
260 the same as the crucial elements and geochemical parameters summarized in Section 4.1, leading
261 to the result that although the model was trained only on 12 features, its score is slightly higher

than that of M-T-2 and M-T-3 that were trained on 35 and 22 features, respectively. This effect was even more pronounced in models trained using "Trace" datasets. Model T-4 was trained on a feature dataset consisting entirely of 14 rare-earth elements, but scored worse than model T-5 (no. of features = 7). This indicates that REEs might not play a positive role in training the XGBoost model for discriminating between fertile and barren apatites in general. As shown in Figure 5, the performance of model T-3 trained on 21 features without REEs was better than that of T-2 trained with 28 features with REEs.

On the whole, 10 of the 12 models could correctly classify more than 90% samples of the test set (accuracy > 0.9), indicating the excellent performance of the models in this study for distinguishing between "Mineralized" and "Unmineralized" apatites. Among all 14 models, model M-T-1 achieved the highest scores for both training and test sets. In the results of this model, all samples in the training set were correctly classified (accuracy = 1), and more than 99% samples of the test set were correctly classified (accuracy = 0.9918). Elemental data obtained in practice might not be enough as those used for model M-T-1; however, model M-T-4 can achieve a similar performance with an accuracy and an F1 score of 0.9878 and 0.900, respectively, when only using 9 elements (12 features). This demonstrates that the classification models in this study can work in a variety of situations. However, the XGBoost model performance sharply declined when the number of selected features decreased to 2 (Figure 5). Given the overall classification results, it was clear that the XGBoost models in this study can achieve excellent performance when proper feature selection was performed and can be applied in various scenarios.

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

DISCUSSION

Limitation of Conventional Apatite Fertility Indicators

Previous research has suggested that magmas characterized by high water content, high Sr/Y ratio, and high oxidation state play a vital role in the genesis of porphyry Cu deposits (Lu et al., 2015; Richards, 2011; 2015). Recent investigations have indicated that chlorine and sulfur are crucial components of ore-forming fluids due to their ability to form complexes with ore metals, including Cu, Au, Pb, Zn, Fe, and Mo (Hsu et al., 2019; Xu et al., 2021; 2022). These geochemical signatures of magma could be inherited by apatite crystallized from such fertile magmas. Accordingly, various apatite fertility indicators, such as Sr/Y, $(Ce/Yb)_N$, Cl/F, V/Y, and $(Ce/Pb)_N$, have been proposed to distinguish between fertile and barren suites (Belousova et al., 2002; Mao et al., 2016; Xu et al., 2021). In this study, the performance of several traditional apatite fertility indicators was evaluated using the raw dataset (Figure 6). However, the efficacy of these indicators in predicting mineralization at a global scale was found to be limited, despite their effectiveness in specific metallogenic systems, such as porphyry deposits; for instance, Xu et al. (2021) proposed three indicators in apatite that worked effectively in differentiating fertile and barren porphyries. However, it only showed the best accuracy of 0.553 (Figure 6a) when applied to the dataset in this study. More precisely, the Cl/F ratio-based classification (Figure 6a) yielded a true-positive rate (TPR) of 0.421 for fertile apatite and a true-negative rate (TNR) of 0.580 for barren apatite. The biplot of V vs. Y (Figure 6b) has an accuracy, TPR, and TNR of 0.261, 0.866, and 0.026, respectively, indicating its ability to identify fertile apatite but not barren apatite. Comparative scores are 0.423, 0.007, and 0.919 for Cl/F vs. $(Ce/Pb)_N$ biplot (Figure 6c), 0.242, 0.181, and 0.640 for V/Y vs. $(Ce/Pb)_N$ biplot (Figure 6d), and 0.299, 0.172,

and 0.750 for V/Y vs. Cl/F biplot (Figure 6e), indicating that the $(Ce/Pb)_N$ ratio might perform better in determining barren apatite.

The traditional discrimination diagrams exhibit low accuracies (from 0.242 to 0.553) on a global scale dataset and could result in inconclusive findings and imprecise mineralization targets when applied to mineral exploration. As an increasing amount of geochemical data pertaining to apatite becomes publicly accessible, the limitations of individual geochemical indicators are becoming progressively conspicuous. One major limitation is the lack of transferability of a geochemical indicator that accurately identifies fertile rocks in one system to another. Additionally, traditional methods that combine limited indicators could not comprehensively introduce the underlying pattern of multiple elements and efficiently assess the metallogenic fertility.

Accordingly, an ML model that can process high-dimensional geochemical data is considered to be a robust mineral exploration tool. The XGBoost models in this study are apparently more accurate and efficient than traditional elemental biplots with accuracies ranging from 0.8507 to 0.9918, suggesting a higher success rate during prospecting and exploration. In addition, ML can simultaneously integrate all apatite trace element features and directly capture the relationships between geochemical data and mineralization. The advantage of this approach is the applicability of results to any geological environment, while the disadvantage is that it required a systematic and comprehensive apatite geochemical dataset from worldwide occurrences. With the growth in the volume of geochemical data on apatite from various deposit types, ML models trained on such datasets are likely to become more sophisticated and accurate. This is because ML

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

algorithms excel at identifying complex patterns and relationships in large datasets, which can lead to more precise discrimination of mineralization. As a result, hopes are high for the robustness of this ML approach in mineral exploration. With the continuous expansion of geochemical databases and the ongoing refinement of ML algorithms, further improvements in the performance of these models are expected.

Model Application and Limitation

To further substantiate the reliability of our model and elucidate its potential applications, a set of unlabelled apatite compositional data reported by Xu et al. (2021) was employed as a validation dataset (Table S5). These apatites were extracted from rock samples collected in 12 distinct localities spanning Iran and western China (Tibet and Yunnan) which include both barren localities (Liuhe, Nanmuqie, Renduoxiang, Songgui, Wolong) and porphyry deposits (Beiya, Chongmuda, Jiama, Machangqing, Masjed Daghi, Qulong, Zhunuo). In an effort to strike a balance between feature quantity and model performance, we utilized the cost-effective model M-T-5 to predict the fertility of the host rocks of these apatites and the corresponding mineralization potential in the respective regions. To render these data points visually interpretable, Principal Component Analysis (Smith, 2002) was employed to reduce their dimensionality to two dimensions. Figure 7a illustrates that their distribution primarily aligns with the clusters of the "Major & Trace" subset, signifying the comprehensive spectrum covered by our established database.

After applying model M-T-5 on the validation dataset and then organizing the resulting probability values and prediction results (Figure 7b-f and Table 2), the robus performance of our model was reconfirmed. As evident from Figure 7b and Table 2, apatites originating from fertile

rock occurrences in all seven instances were accurately identified, with five of them exhibiting a 100% likelihood of mineralization, while the remaining two displayed probabilities exceeding 70%. Additionally, four barren rock samples were predicted with high probabilities, whereas the prediction results for apatite from Renduoxiang displayed suboptimal performance.

These cases emphasize the optimistic perspective for utilizing ML models and the compositional data of apatite to predict the mineralization potential in this region. However, it is crucial to note that our current database exhibits a degree of data imbalance, with a predominant proportion of apatite data from porphyry and skarn deposits. This imbalance may render our models more sensitive to these two systems. Hence, it is imperative to conduct a thorough examination of data distribution before deploying the models, and data clusters that substantially deviate from our database's distribution should be used judiciously. This highlights the need for larger-scale and more diverse datasets, which relies on contributions of more geological researchers.

Geochemical Explanation for ML Model

The feature importance ranks of 14 XGBoost models reveal that several indicators are highly relevant with mineralization based on our dataset, including Cl, F, S, V, Sr/Y, V/Y, Eu*, $(La/Yb)_N$, and La/Sm. This is in consistent with the conclusions of previous studies (Lu et al., 2015; Richards, 2011; 2015; Xu et al., 2021; 2022). Several observations of higher Cl and S contents in fertile than barren apatites (Chelle-Michou and Chiaradia, 2017; Xu et al., 2021; Zhu et al., 2018) and fluid inclusion studies (Sillitoe, 2010) highlighted the significance of chlorine and sulfur in supporting the transport and deposition of ore metals at magmatic hydrothermal systems (Duan et al., 2021; Wang et al., 2021a; Zheng et al., 2021). These two elements form ligands with ore metals, such as Cu, Au, Pb, Zn, Fe, and Mo, allowing their transport to the site

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

of ore deposition and are involved in causing hydrothermal alteration. Additionally, the V/Y ratio was proved to be high in apatite crystallized from ore-forming magma (Xu et al., 2021). The presence of elevated V contents in the host magma indicates high levels of dissolved H_2O in the melt, which was also recognized as a crucial factor for mineralization (Chiaradia, 2014; Lu et al., 2015). The mechanism therein is that amphibole has a more wide-ranging crystallization sequence than titanomagnetite in high H_2O content melt environment, leading to the retention of more V in the residual melt and the efficient extraction of Y from the melt into amphibole. Therefore, apatite crystallized from such magmas exhibited small negative Eu anomalies and high V/Y, reflecting early amphibole fractionation and suppression of plagioclase crystallization in hydrous melts (Davidson et al., 2007). Considering the feature importance analysis of this study and the suggestions of previous studies, Cl- and S-enriched hydrous magma seem to be crucial factors in metal enrichment and mineralization.

IMPLICATIONS

The XGBoost models developed in this study exhibit strong discriminatory power in distinguishing apatite samples from mineralized fertile suites and those from barren suites, with high accuracy and efficiency. This indicates that ML, when integrated with conventional geological and geochemical techniques, can offer a cost-effective and efficient approach to evaluate mineralization. Furthermore, this methodology holds potential for identifying fertile and barren magmas in other systems by other minerals, including quartz, zircon, and titanite. The versatility of ML models trained on different target variables can extend beyond solid Earth science to other fields.

In conclusion, this study demonstrates the efficacy of ML methods in capturing the intricate relationship between 43-dimensional apatite geochemical data and mineralization. The findings underscore the significant feasibility of ML in analyzing and processing high-dimensional data in solid Earth sciences, which could help elucidate underlying geological events.

However, the application of ML methods also has potential pitfalls. Firstly, ML algorithms do not include an inferential component such as an adequate assessment of uncertainty (Frenzel, 2023). Secondly, care of the breadth and representativeness of the data should be taken. Within different deposits of same type, factors such as the alteration degree of host rock and the mineral assemblage of apatite can significantly influence its composition. Furthermore, variations in the composition of rock samples from different locations in a deposit are noteworthy. Hence, the volume and representativeness of the data are of paramount importance. A small number of samples cannot adequately represent the characteristics of the entire deposit. Instead, a relatively large sample size is likely necessary to reasonably encompass the extent of the observed variability. In our database, there exists an imbalance in the quantity of "Mineralized" apatite data originating from different types of deposit. For instance, data from porphyry and skarn deposits are more abundant, while data from other deposit types are less represented. Consequently, this disparity may result in our model performing more effectively when applied to these two deposit types. Thirdly, interpretational pitfalls must be acknowledged. Owing to the influence of data imbalance and volume, it remains to be verified whether parameters identified by our models as highly sensitive to mineralization are equally effective across all ore systems.

ACKNOWLEDGMENTS

This research was funded by National Natural Science Foundation of China (42222304, 42073038), and the "Deep-time Digital Earth" Science and Technology Leading Talents Team Funds for the Central Universities for the Frontiers Science Center for Deep-time Digital Earth, China University of Geosciences (Beijing) (Fundamental Research Funds for the Central Universities; grant number: 2652023001), Young Talent Support Project of CAST, the Fundamental Research Funds for the Central Universities (Grant no. 265QZ2021012) and International Geoscience Programme (IGCP-741, IGCP-662). This is the XXth contribution of B.X. for National Mineral Rock and Fossil Specimens Resource Center. The authors would like to express their sincere gratitude to Jiali Lei for her valuable contributions in enhancing the model's performance. Furthermore, we extend our appreciation to Jiaxing Yu and Wenxuan Wu for their assistance in data collection. Their efforts significantly contributed to the successful completion of this study. Finally, we heartfeltly thank the Laboratory of the Jewelry College (China University of Geosciences, Beijing) for their assistance.

REFERENCES

Adlakha, E., Hanley, J., Falck, H., and Boucher, B. (2018) The origin of mineralizing hydrothermal fluids recorded in apatite chemistry at the Cantung W–Cu skarn deposit, NWT, Canada. European Journal of Mineralogy, 30(6), 1095–1113.

Azadbakht, Z., Lentz, D.R., and McFarlane, C.R.M. (2018) Apatite Chemical Compositions from Acadian-Related Granitoids of New Brunswick, Canada: Implications for Petrogenesis and Metallogenesis. Minerals, 8(12), 598–628.

Azodi, C.B., Tang, J.L., and Shiu, S.H. (2020) Opening the Black Box: Interpretable Machine Learning for Geneticists. Trends in Genetics, 36(6), 442–455.

Bea, F., and Montero, P. (1999) Behavior of accessory phases and redistribution of Zr, REE, Y, Th, and U during metamorphism and partial melting of metapelites in the lower crust: An example from the Kinzigite Formation of Ivrea-Verbano, NW Italy. Geochimica et Cosmochimica Acta, 63(7/8), 1133–1153.

Belousova, E.A., Griffin, W.L., O'Reilly, S.Y., and Fisher, N.I. (2002) Apatite as an indicator mineral for mineral exploration: trace-element compositions and their relationship to host rock type. Journal for Geochemical Exploration, 76(1), 45–69.

Bruand, E., Fowler, M., Storey, C., and Darling, J. (2017) Apatite trace element and isotope applications to petrogenesis and provenance. American Mineralogist, 102(1), 75–84.

Bruand, E., Storey, C., Fowler, M., and Heilimo, E. (2019) Oxygen isotopes in titanite and apatite, and their potential for crustal evolution research. Geochimica et Cosmochimica Acta, 255, 144–162.

Cai, A., Qiu, H., and Niu, F. (2022) Semi-Supervised Surface Wave Tomography With Wasserstein Cycle-Consistent GAN: Method and Application to Southern California Plate Boundary Region. Journal of Geophysical Research: Solid Earth, 127(3).

Cao, J., Yang, X., Yang, S., Zhong, C., and Wang, Y. (2022) Records of apatite for multiple injections of magmas in adakitic plutons: A case study of Mesozoic plutons in the Shatanjiao region of the Tongling ore cluster, south China. Journal of Asian Earth Sciences, 242.

455 Cao, M., Li, G., Qin, K., Seitmuratova, E.Y., and Liu, Y. (2012) Major and Trace Element
456 Characteristics of Apatites in Granitoids from Central Kazakhstan: Implications for
457 Petrogenesis and Mineralization. Resource Geology, 62(1), 63–83.
458 Chakhmouradian, A.R., Reguir, E.P., Zaitsev, A.N., Couëslan, C., Xu, C., Kynický, J., Mumin,
459 A.H., and Yang, P. (2017) Apatite in carbonatitic rocks: Compositional variation, zoning,
460 element partitioning and petrogenetic significance. Lithos, 274–275, 188–213.
461 Chelle-Michou, C., and Chiaradia, M. (2017) Amphibole and apatite insights into the evolution
462 and mass balance of Cl and S in magmas associated with porphyry copper deposits.
463 Contributions to Mineralogy and Petrology, 172(11), 105.
464 Chen, H., Su, C., Tang, Y.Q., Li, A.Z., Wu, S.S., Xia, Q.K., and ZhangZhou, J. (2021) Machine
465 Learning for Identification of Primary Water Concentrations in Mantle Pyroxene.
466 Geophysical Research Letters, 48(18).
467 Chen, L., and Zhang, Y. (2018) In situ major-, trace-elements and Sr-Nd isotopic compositions
468 of apatite from the Luming porphyry Mo deposit, NE China: Constraints on the
469 petrogenetic-metallogenic features. Ore Geology Reviews, 94, 93–103.
470 Chen, T., and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. Proceedings of
471 the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data
472 Mining, p. 785–794. Association for Computing Machinery, San Francisco, California,
473 USA.
474 Chen, X., Liang, H., Zhang, J., Huang, W., Ren, L., and Zou, Y. (2020) Geochemical
475 characteristics and oxidation states of the Xietongmen ore-bearing porphyries: Implication
476 for the genetic types of the Xietongmen No. I and No. II deposits, southern Tibet.
477 Geological Journal, 55(6), 4691–4712.

478 Chiaradia, M. (2014) Copper enrichment in arc magmas controlled by overriding plate thickness.
479 Nature Geoscience, 7(1), 43–46.
480 Chu, M.-F., Wang, K.-L., Griffin, W.L., Chung, S.-L., O'Reilly, S.Y., Pearson, N.J., and Iizuka,
481 Y. (2009) Apatite Composition: Tracing Petrogenetic Processes in Transhimalayan
482 Granitoids. Journal of Petrology, 50(10), 1829–1855.
483 Cianetti, S., Bruni, R., Gaviano, S., Keir, D., Piccinini, D., Saccorotti, G., and Giunchi, C. (2021)
484 Comparison of Deep Learning Techniques for the Investigation of a Seismic Sequence: An
485 Application to the 2019, Mw 4.5 Mugello (Italy) Earthquake. Journal of Geophysical
486 Research: Solid Earth, 126(12).
487 Davidson, J., Turner, S., Handley, H., Macpherson, C., and Dosseto, A. (2007) Amphibole
488 "sponge" in arc crust? Geology, 35(9), 787–790.
489 Dill, H.G. (1994) Can REE patterns and U-Th variations be used as a tool to determine the origin
490 of apatite in clastic rocks? Sedimentary Geology, 92(3–4), 175–196.
491 Duan, D.-F., Jiang, S.-Y., Tang, Y.-J., Wu, Y., Zhou, B., and Zhu, J. (2021) Chlorine and sulfur
492 evolution in magmatic rocks: A record from amphibole and apatite in the Tonglvshan Cu-Fe
493 (Au) skarn deposit in Hubei Province, south China. Ore Geology Reviews, 137.
494 El Korh, A., Schmidt, S.T., Ulianov, A., and Potel, S. (2009) Trace Element Partitioning in HP–
495 LT Metamorphic Assemblages during Subduction-related Metamorphism, Ile de Groix,
496 France: a Detailed LA-ICPMS Study. Journal of Petrology, 50(6), 1107–1148.
497 Frenzel, M. (2023) Making sense of mineral trace-element data – How to avoid common pitfalls
498 in statistical analysis and interpretation. Ore Geology Reviews, 159.
499 Friedman, J.H. (2001) Greedy function approximation: A gradient boosting machine. The Annals
500 of Statistics, 29(5), 1189–1232.

501 Gao, X., Yang, L., Wang, C., He, W., Bao, X., and Zhang, S. (2020) Halogens and trace
502 elements of apatite from Late Mesozoic and Cenozoic porphyry Cu-Mo-Au deposits in SE
503 Tibet, China: Constraints on magmatic fertility and granitoid petrogenesis. Journal of Asian
504 Earth Sciences, 203.
505 Glorie, S., Jepson, G., Konopelko, D., Mirkamalov, R., Meeuws, F., Gilbert, S., Gillespie, J.,
506 Collins, A., Xiao, W., and Dewaele, S. (2019) Thermochronological and geochemical
507 footprints of post-orogenic fluid alteration recorded in apatite: implications for
508 mineralisation in the Uzbek Tian Shan. Gondwana Research, 71, 1–15.
509 Guo, P., Yang, T., Xu, W.L., and Chen, B. (2021) Machine Learning Reveals Source
510 Compositions of Intraplate Basaltic Rocks. Geochemistry, Geophysics, Geosystems, 22(9).
511 Henrichs, I.A., O'Sullivan, G., Chew, D.M., Mark, C., Babechuk, M.G., McKenna, C., and Emo,
512 R. (2018) The trace element and U-Pb systematics of metamorphic apatite. Chemical
513 Geology, 483, 218–238.
514 Hoshino, M., Watanabe, Y., and Kon, Y. (2017) Implication of Apatite and Anhydrite for
515 Formation of an Iron-Oxide-Apatite(IOA) Rare Earth Element Prospect, Benjamin River,
516 Canada. Resource Geology, 67(4), 361–383.
517 Hsu, Y.-J., Zajacz, Z., Ulmer, P., and Heinrich, C.A. (2019) Chlorine partitioning between
518 granitic melt and H_2O-CO_2-NaCl fluids in the Earth's upper crust and implications for
519 magmatic-hydrothermal ore genesis. Geochimica et Cosmochimica Acta, 261, 171–190.
520 Joosu, L., Lepland, A., Kreitsmann, T., Üpraus, K., Roberts, N.M.W., Paiste, P., Martin, A.P.,
521 and Kirsimäe, K. (2016) Petrography and the REE-composition of apatite in the
522 Paleoproterozoic Pilgujärvi Sedimentary Formation, Pechenga Greenstone Belt, Russia.
523 Geochimica et Cosmochimica Acta, 186, 135–153.

524 Kohavi, R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and
525 Model Selection. Proceedings of the 14th International Joint Conference on Artificial
526 Intelligence (IJCAI-95), p. 1137–1145. International joint conference on Artificial
527 intelligence, Montreal.
528 La Cruz, N.L., Ovalle, J.T., Simon, A.C., Konecke, B.A., Barra, F., Reich, M., Leisen, M., and
529 Childress, T.M. (2020) The Geochemistry of Magnetite and Apatite from the El Laco Iron
530 Oxide-Apatite Deposit, Chile: Implications for Ore Genesis. Economic Geology, 115(7),
531 1461–1491.
532 Li, Q., Sun, X., Lu, Y., Wang, F., and Hao, J. (2021) Apatite and zircon compositions for
533 Miocene mineralizing and barren intrusions in the Gangdese porphyry copper belt of
534 southern Tibet: Implication for ore control. Ore Geology Reviews, 139, 104474.
535 Li, Y.E., O'Malley, D., Beroza, G., Curtis, A., and Johnson, P. (2023) Machine Learning
536 Developments and Applications in Solid-Earth Geosciences: Fad or Future? Journal of
537 Geophysical Research: Solid Earth, 128.
538 Linville, L.M. (2022) Event-Based Training in Label-Limited Regimes. Journal of Geophysical
539 Research: Solid Earth, 127.
540 Liu, L., Zhang, W., Jin, C., and Chen, W.T. (2021) Mineralogical and geochemical constraints
541 on the origin of the variable REE enrichments in the Kangdian IOCG province, SW China.
542 Ore Geology Reviews, 138.
543 Long, X.-Y., Tang, J., Xu, W.-L., Sun, C.-Y., Luan, J.-P., Xiong, S., and Zhang, X.-M. (2023)
544 Trace element and Nd isotope analyses of apatite in granitoids and metamorphosed
545 granitoids from the eastern Central Asian Orogenic Belt: Implications for petrogenesis and
546 post-magmatic alteration. Geoscience Frontiers, 14(2), 101517.

547 Lu, Y.-J., Loucks, R.R., Fiorentini, M.L., Yang, Z.-M., and Hou, Z.-Q. (2015) Fluid flux melting
548 generated postcollisional high Sr/Y copper ore–forming water-rich magmas in Tibet.
549 Geology, 43(7), 583–586.
550 Lupulescu, M.V., Hughes, J.M., Chiarenzelli, J.R., and Bailey, D.G. (2017) Texture, Crystal
551 Structure, and Composition of Fluorapatites From Iron Oxide-Apatite (Ioa) Deposits,
552 Eastern Adirondack Mountains, New York. The Canadian Mineralogist, 55(3), 399–417.
553 Mao, M., Rukhlov, A.S., Rowins, S.M., Spence, J., and Coogan, L.A. (2016) Apatite Trace
554 Element Compositions: A Robust New Tool for Mineral Exploration. Economic Geology,
555 111(5), 1187–1222.
556 Meng, X., Kleinsasser, J.M., Richards, J.P., Tapster, S.R., Jugo, P.J., Simon, A.C., Kontak, D.J.,
557 Robb, L., Bybee, G.M., Marsh, J.H., and Stern, R.A. (2021) Oxidized sulfur-rich arc
558 magmas formed porphyry Cu deposits by 1.88 Ga. Nature Communication, 12(1), 2189.
559 Mercer, C.N., Watts, K.E., and Gross, J. (2020) Apatite trace element geochemistry and
560 cathodoluminescent textures—A comparison between regional magmatism and the Pea
561 Ridge IOAREE and Boss IOCG deposits, southeastern Missouri iron metallogenic province,
562 USA. Ore Geology Reviews, 116.
563 Nutman, A.P. (2007) Apatite recrystallisation during prograde metamorphism, Cooma, southeast
564 Australia: implications for using an apatite - graphite association as a biotracer in ancient
565 metasedimentary rocks. Australian Journal of Earth Sciences, 54(8), 1023–1032.
566 O'Reilly, S.Y., and Griffin, W.L. (2000) Apatite in the mantle: implications for metasomatic
567 processes and high heat production in Phanerozoic mantle. Lithosphere, 53(3), 217–232.

568 O'Sullivan, G.J., Chew, D.M., Morton, A.C., Mark, C., and Henrichs, I.A. (2018) An Integrated
569 Apatite Geochronology and Geochemistry Tool for Sedimentary Provenance Analysis.
570 Geochemistry, Geophysics, Geosystems, 19(4), 1309–1326.
571 O'Sullivan, G., Chew, D., Kenny, G., Henrichs, I., and Mulligan, D. (2020) The trace element
572 composition of apatite and its application to detrital provenance studies. Earth-Science
573 Reviews, 201.
574 Palma, G., Barra, F., Reich, M., Valencia, V., Simon, A.C., Vervoort, J., Leisen, M., and
575 Romero, R. (2019) Halogens, trace element concentrations, and Sr-Nd isotopes in apatite
576 from iron oxide-apatite (IOA) deposits in the Chilean iron belt: Evidence for magmatic and
577 hydrothermal stages of mineralization. Geochimica et Cosmochimica Acta, 246, 515–540.
578 Python, A., Bender, A., Nandi, A.K., Hancock, P.A., Arambepola, R., Brandsch, J., and Lucas,
579 T.C.D. (2021) Predicting non-state terrorism worldwide. 7(31), 1–13.
580 Qin, B., Huang, F., Huang, S., Python, A., Chen, Y., and ZhangZhou, J. (2022) Machine
581 Learning Investigation of Clinopyroxene Compositions to Evaluate and Predict Mantle
582 Metasomatism Worldwide. Journal of Geophysical Research: Solid Earth, 127(5).
583 Qu, P., Li, N.-B., Niu, H.-C., Shan, Q., Weng, Q., and Zhao, X.-C. (2021) Difference in the
584 nature of ore-forming magma between the Mesozoic porphyry Cu-Mo and Mo deposits in
585 NE China: Records from apatite and zircon geochemistry. Ore Geology Reviews, 135.
586 Richards, J.P. (2011) High Sr/Y arc magmas and porphyry Cu ± Mo ± Au deposits: just add
587 water. Economic Geology, 106(7), 1075–1081.
588 Richards, J. P. (2015) The oxidation state, and sulfur and Cu contents of arc magmas:
589 implications for metallogeny. Lithos, 233, 27–45.

590 Sha, L.K., and Chappell, B.W. (1999) Apatite chemical composition, determined by electron
591 microprobe and laser-ablation inductively coupled plasma mass spectrometry, as a probe
592 into granite petrogenesis. Geochimica et Cosmochimica Acta, 63(22), 3861–3881.
593 Sillitoe, R.H. (2010) Porphyry Copper Systems. Economic Geology, 105(1), 3–41.
594 Smith, L. I. (2002). A tutorial on Principal Components Analysis (Computer Science Technical
595 Report No. OUCS-2002-12). Available: http://hdl.handle.net/10523/7534. University of
596 Otago, Otago, New Zealand.
597 Stehman, S.V. (1997) Selecting and interpreting measures of thematic classification accuracy.
598 Remote Sensing of Environment, 62(1), 77–89.
599 Sun, C.-Y., Cawood, P.A., Xu, W.-L., Zhang, X.-M., Tang, J., Li, Y., Sun, Z.-X., and Xu, T.
600 (2022) In situ geochemical composition of apatite in granitoids from the eastern Central
601 Asian Orogenic Belt: A window into petrogenesis. Geochimica et Cosmochimica Acta, 317,
602 552–573.
603 Tan, H.M.R., Huang, X.-W., Meng, Y.-M., Xie, H., and Qi, L. (2023) Multivariate statistical
604 analysis of trace elements in apatite: Discrimination of apatite with different origins. Ore
605 Geology Reviews, 153.
606 Tang, P., Tang, J., Wang, Y., Lin, B., Leng, Q., Zhang, Q., He, L., Zhang, Z., Sun, M., Wu, C.,
607 Qi, J., Li, Y., and Dai, S. (2021) Genesis of the Lakang'e porphyry Mo (Cu) deposit, Tibet:
608 Constraints from geochemistry, geochronology, Sr-Nd-Pb-Hf isotopes, zircon and apatite.
609 Lithos, 380–381, 105834.
610 Tollari, N., Barnes, S., Cox, R., and Nabil, H. (2008) Trace element concentrations in apatites
611 from the Sept-Îles Intrusive Suite, Canada — Implications for the genesis of nelsonites.
612 Chemical Geology, 252(3–4), 180–190.

613 Wang, D.-Z., Zhu, J.-J., Bi, X.-W., Fu, S.-L., Lu, Z.-T., Wu, L.-R., and Hu, R. (2021a)
614 Increasing sulfur and chlorine contents in ore-forming magmas: The key to Pulang porphyry
615 Cu-Au formation, SW China. Ore Geology Reviews, 139.
616 Wang, Y., Qiu, K.F., Müller, A., Hou, Z.L., Zhu, Z.H., and Yu, H.C. (2021b) Machine Learning
617 Prediction of Quartz Forming-Environments. Journal of Geophysical Research: Solid Earth,
618 126(8).
619 Xiao, H., Zhang, F., Shen, Z., Wu, K., and Zhang, J. (2021) Classification of Weather
620 Phenomenon From Images by Using Deep Convolutional Neural Network. Earth and Space
621 Science, 8(5).
622 Xie, F., Tang, J., Chen, Y., and Lang, X. (2018) Apatite and zircon geochemistry of Jurassic
623 porphyries in the Xiongcun district, southern Gangdese porphyry copper belt: Implications
624 for petrogenesis and mineralization. Ore Geology Reviews, 96, 98–114.
625 Xu, B., Hou, Z.-Q., Griffin, W.L., Lu, Y., Belousova, E., Xu, J.-F., and O'Reilly, S.Y. (2021)
626 Recycled volatiles determine fertility of porphyry deposits in collisional settings. American
627 Mineralogist, 106(4), 656–661.
628 Xu, B., Hou, Z.-Q., Griffin, W.L., Yu, J.-X., Long, T., Zhao, Y., Wang, T., Fu, B., Belousova,
629 E., and O'Reilly, S.Y. (2022) Apatite halogens and Sr-O and zircon Hf-O isotopes:
630 Recycled volatiles in Jurassic porphyry ore systems in southern Tibet. Chemical Geology,
631 605.
632 Xu, L.-L., Bi, X.-W., Zhang, X.-C., Huang, M.-L., and Liu, G. (2023) Mantle contribution to the
633 generation of the giant Jinduicheng porphyry Mo deposit, Central China: New insights from
634 combined in-situ element and isotope compositions of zircon and apatite. Chemical
635 Geology, 616.

636 Xu, W., Fan, H., Hu, F., Santosh, M., Yang, K., and Lan, T. (2015) In situ chemical and Sr–Nd–
637 O isotopic compositions of apatite from the Tongshi intrusive complex in the southern part
638 of the North China Craton: Implications for petrogenesis and metallogeny. Journal of Asian
639 Earth Sciences, 105, 208–222.
640 Yang, J.-H., Kang, L.-F., Peng, J.-T., Zhong, H., Gao, J.-F., and Liu, L. (2018) In-situ elemental
641 and isotopic compositions of apatite and zircon from the Shuikoushan and Xihuashan
642 granitic plutons: Implication for Jurassic granitoid-related Cu-Pb-Zn and W mineralization
643 in the Nanling Range, South China. Ore Geology Reviews, 93, 382–403.
644 Zafar, T., Leng, C.-B., Zhang, X.-C., and Rehman, H.U. (2019) Geochemical attributes of
645 magmatic apatite in the Kukaazi granite from western Kunlun orogenic belt, NW China:
646 Implications for granite petrogenesis and Pb-Zn (-Cu-W) mineralization. Journal of
647 Geochemical Exploration, 204, 256–269.
648 Zhang, L., Hu, Y., Deng, J., Wang, J., Wang, K., Sui, Q., Chen, Y., Wu, K., and Sun, W. (2021)
649 Genesis and mineralization potential of the Late Cretaceous Chemen granodioritic intrusion
650 in the southern Gangdese magmatic belt, Tibet. Journal of Asian Earth Sciences, 217.
651 Zheng, X., Liu, Y., and Zhang, L. (2021) The role of sulfate-, alkali-, and halogen-rich fluids in
652 mobilization and mineralization of rare earth elements: Insights from bulk fluid
653 compositions in the Mianning–Dechang carbonatite-related REE belt, southwestern China.
654 Lithos, 386–387, 106008.
655 Zhong, S., Feng, C., Seltmann, R., Li, D., and Dai, Z. (2018) Geochemical contrasts between
656 Late Triassic ore-bearing and barren intrusions in the Weibao Cu–Pb–Zn deposit, East
657 Kunlun Mountains, NW China: constraints from accessory minerals (zircon and apatite).
658 Mineralium Deposita, 53(6), 855–870.

659 Zhong, S., Li, S., Liu, Y., Cawood, P.A., and Seltmann, R. (2023a) I-type and S-type granites in
660 the Earth's earliest continental crust. Communications Earth & Environment, 4(1), 61.
661 Zhong, S.H., Liu, Y., Li, S.Z., Bindeman, I.N., Cawood, P.A., Seltmann, R., Niu, J.H., Guo,
662 G.H., and Liu, J.Q. (2023b) A machine learning method for distinguishing detrital zircon
663 provenance. Contributions to Mineralogy and Petrology, 178(6).
664 Zhou, Y., Zhang, Z., Yang, J., Ge, Y., and Cheng, Q. (2022) Machine Learning and Singularity
665 Analysis Reveal Zircon Fertility and Magmatic Intensity: Implications for Porphyry Copper
666 Potential. Natural Resources Research, 31(6), 3061–3078.
667 Zhu, J.-J., Richards, J.P., Rees, C., Creaser, R., DuFrane, S.A., Locock, A., Petrus, J.A., and
668 Lang, J. (2018) Elevated Magmatic Sulfur and Chlorine Contents in Ore-Forming Magmas
669 at the Red Chris Porphyry Cu-Au Deposit, Northern British Columbia, Canada. Economic
670 Geology, 113(5), 1047–1075.
671 Zou, S., Chen, X., Brzozowski, M.J., Leng, C.B., and Xu, D. (2022) Application of Machine
672 Learning to Characterizing Magma Fertility in Porphyry Cu Deposits. Journal of
673 Geophysical Research: Solid Earth, 127(8).
674

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

675 **FIGURE CAPTIONS**

676 **FIGURE 1.** Representative locations of apatite samples of which analyses were collected in this
677 study. The green points indicate that the analyses were collected from apatites that formed in
678 mineralized rocks and were labeled as "Mineralized." The orange points indicate that the
679 analyses were collected from apatites in unmineralized rock samples and were labeled as
680 "Unmineralized."

681

682 **FIGURE 2.** Boxplots of major and trace elements and geochemical parameter data of apatite
683 samples worldwide, expressed in wt.% (a) and ppm (b). The box represents the interquartile
684 range (IQR), with the upper (75%) and lower (25%) quartiles demarcated. The outer whiskers
685 extended to 1.5 times the IQR. A horizontal line within the colored box represents the median
686 (50%). The black square symbols and circle symbols indicate the mean and outliers, respectively.

687

688 **FIGURE 3.** Workflow for the XGBoost modeling in this research. Step I: The labeled apatite
689 dataset ("Major and Trace," "Major," or "Trace" dataset) is read as input and preprocessed and
690 then randomly split into training set (80%) and test set (20%) by the holdout method. Step II:
691 The training set is applied to train the XGBoost model using the fivefold cross-validation
692 method. Step III: The optimal hyperparameters are determined by grid search techniques with the
693 fivefold cross-validation method. Step IV: The best model obtained by Step III is applied to the
694 test set. The classification results will be used to evaluate the model performance.

695

696 **FIGURE 4.** Confusion matrices (left) and feature importance ranks (right) of four representative
697 XGBoost models. The confusion matrices display the prediction results for each class.

698

699 **FIGURE 5.** Correlation between feature selection and XGBoost model performance. n = number
700 of selected features.

701

702 **FIGURE 6.** Scatterplots of elemental ratios of fertile ("Mineralized") and barren
703 ("Unmineralized") apatite in the raw dataset.

704

705 **FIGURE 7.** Distribution patterns of the validation dataset (a) and prediction results generated by
706 model M-T-5; (a) distribution patterns displayed after dimension reduction using PCA for the
707 "Major & Trace" dataset and validation dataset; (b) prediction results compilation for each
708 occurrence; (c-f) probabilities of mineralization for the Songgui, Liuhe, Machangqing, and
709 Jiama.

710

711

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

712 **APPENDIX**

713 The data sets and the code for the machine learning models developed in this study are available

714 at https://github.com/YuyuJo/Supplementary-file-for-AM-9115R .

715

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

716 **TABLE**

717 **TABLE 1. Summary of XGBoost Model Metrics and Feature Importance Ranks**

				Training Set		Test Set		
Model	Dataset	No. of Data	No. of Fea.[a]	F1[b]	Accuracy[c]	F1	Accuracy	Top 10 feature importance
M-T-1	"Major & Trace"	2448	43	1.0000	1.0000	0.9933	0.9918	Y, Rb, Cl, Eu, Ce/Ce*, Tm, V/Y, Zr, Sr/Y, Ce
M-T-2	"Major & Trace"	2448	35	1.0000	1.0000	0.9850	0.9816	Y, V/Y, Rb, Cl, FeO, REE+Y, Sr/Y, Sr, Tb, Cl/F
M-T-3	"Major & Trace"	2448	22	1.0000	1.0000	0.9884	0.9857	Rb, FeO, Y, Cl, V/Y, Eu, Sr, Eu_N/Eu^*_N, $(La/Yb)_N$, Na_2O
M-T-4	"Major & Trace"	2448	12	0.9996	0.9995	0.9900	0.9878	V/Y, Cl, Cl/F, Y, Sr/Y, F, $(La/Yb)_N$, Sr, Eu, La/Sm
M-T-5	"Major & Trace"	2448	6	0.9987	0.9985	0.9733	0.9673	V/Y, Cl/F, $(La/Yb)_N$, Sr/Y, Eu, SO_3
M-1	"Major"	5618	10	0.9921	0.9933	0.9259	0.9386	SO_3, Cl/F, Na_2O, MnO, F, Cl, SiO_2, FeO, P_2O_5, CaO
M-2	"Major"	5618	4	0.9308	0.9424	0.8391	0.8639	SO_3, F, Cl/F, Cl
T-1	"Trace"	9979	33	1.0000	1.0000	0.9939	0.9900	Eu*, V, Tb, Rb, Mn, Ce/Ce*, $(La/Yb)_N$, Sr/Y, Sr, Ce
T-2	"Trace"	9979	28	1.0000	1.0000	0.9849	0.9750	V, Sr/Y, Eu/Eu*/Y, Ce/Ce*, La/Sm, $(La/Yb)_N$, Tb, Pr, Dy, Sr
T-3	"Trace"	9979	21	0.9998	0.9996	0.9912	0.9855	V, Ce/Ce*, Eu*, Zr, Sr, Ce^*_N, V/Y, Rb, Mn, $(La/Yb)_N$
T-4	"Trace"	9979	14	0.9955	0.9926	0.9593	0.9319	Nd, Yb, La, Dy, Tb, Ce, Gd, Tm, Lu, Er
T-5	"Trace"	9979	7	0.9919	0.9866	0.9682	0.9474	V, Sr/Y, Sr, Y, Ce, V/Y, Eu
T-6	"Trace"	9979	3	0.9601	0.9345	0.9543	0.9243	V, Sr, Y
T-7	"Trace"	9979	2	0.9215	0.8626	0.9146	0.8507	Eu, Ce

718 [a]Number of elements and geochemical parameters used in the model. [b]Calculated by Equation 4.

719 [c]Calculated by Equation 3.

720

721 **TABLE 2. Prediction Results of Model M-T-5 Applied to Unlabeled Validation Set**

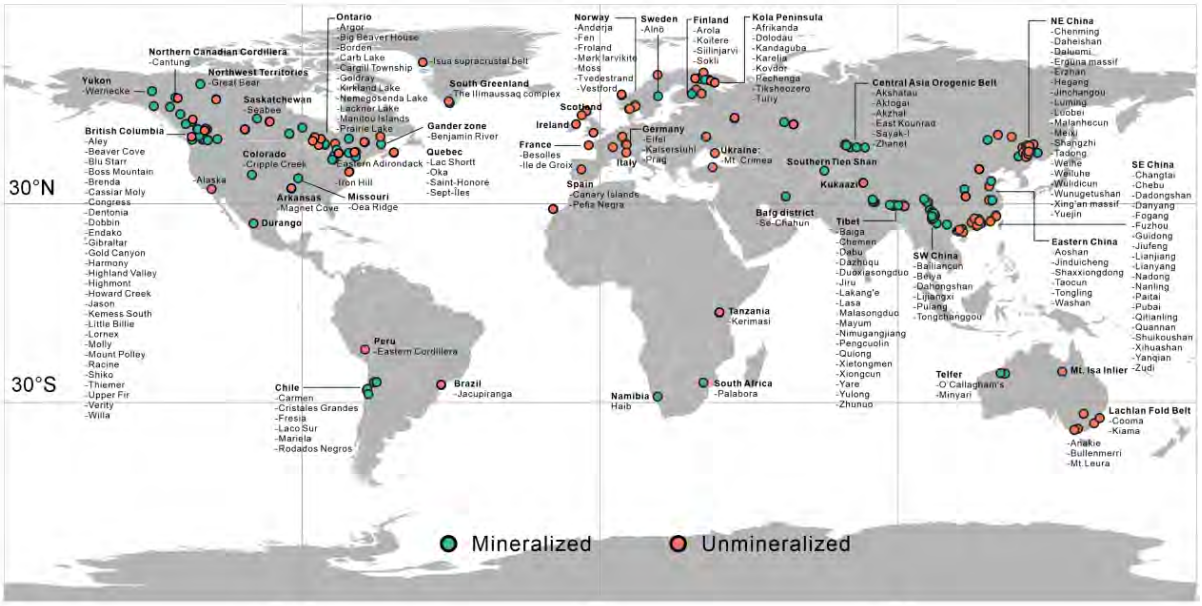
Occurrence	Sample source given by Xu et al. (2021)	Prediction results of apatite composition	
		Probability of barren rock	Probability of Mineralization
Beiya	Porphyry-skarn Cu-Au	30%	70%
Chongmuda	Porphyry-hydrothermal Cu-Mo	0	100%
Jiama	Porphyry-skarn Cu-Au	0	100%
Machangqing	Porphyry Cu-Au	25%	75%
Masjed Daghi	Porphyry Cu-Mo	0	100%
Qulong	Porphyry Cu-Mo	0	100%
Zhunuo	Porphyry Cu	0	100%
Renduoxiang	Barren porphyry	50%	50%
Liuhe	Barren granodiorite	74%	26%
Nanmuqie	Barren porphyry	89%	11%
Songgui	Barren porphyry	98%	2%
Wolong	Barren granodiorite	100%	0

722

723

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

724 **FIGURE 1**



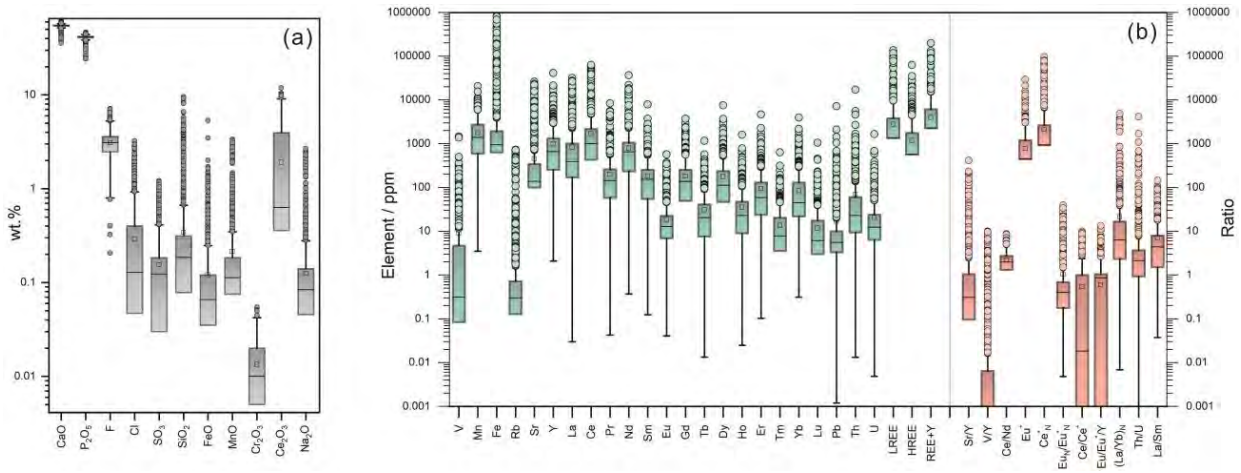
725

726

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

727

728 **FIGURE 2**

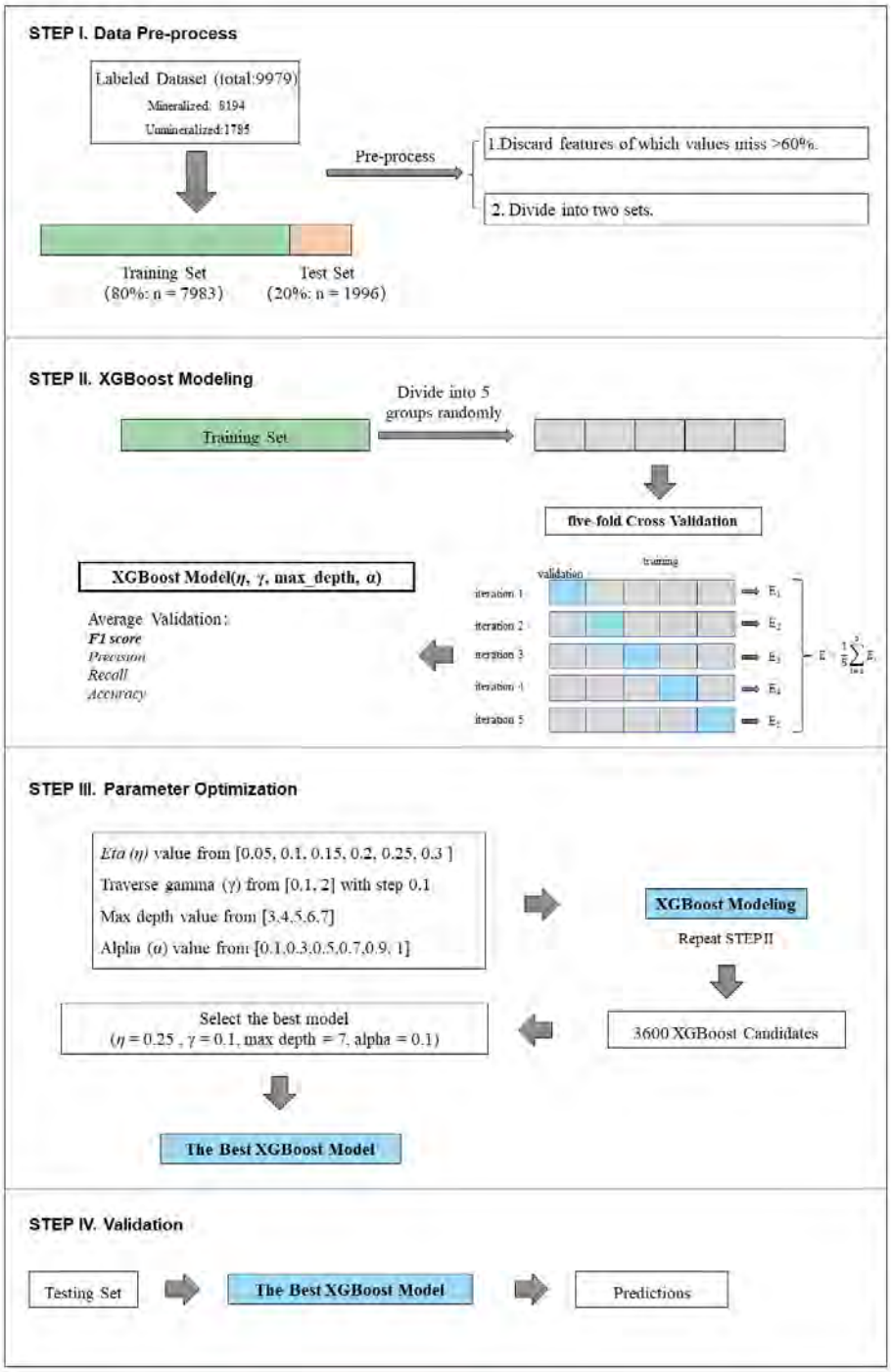


729

730

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

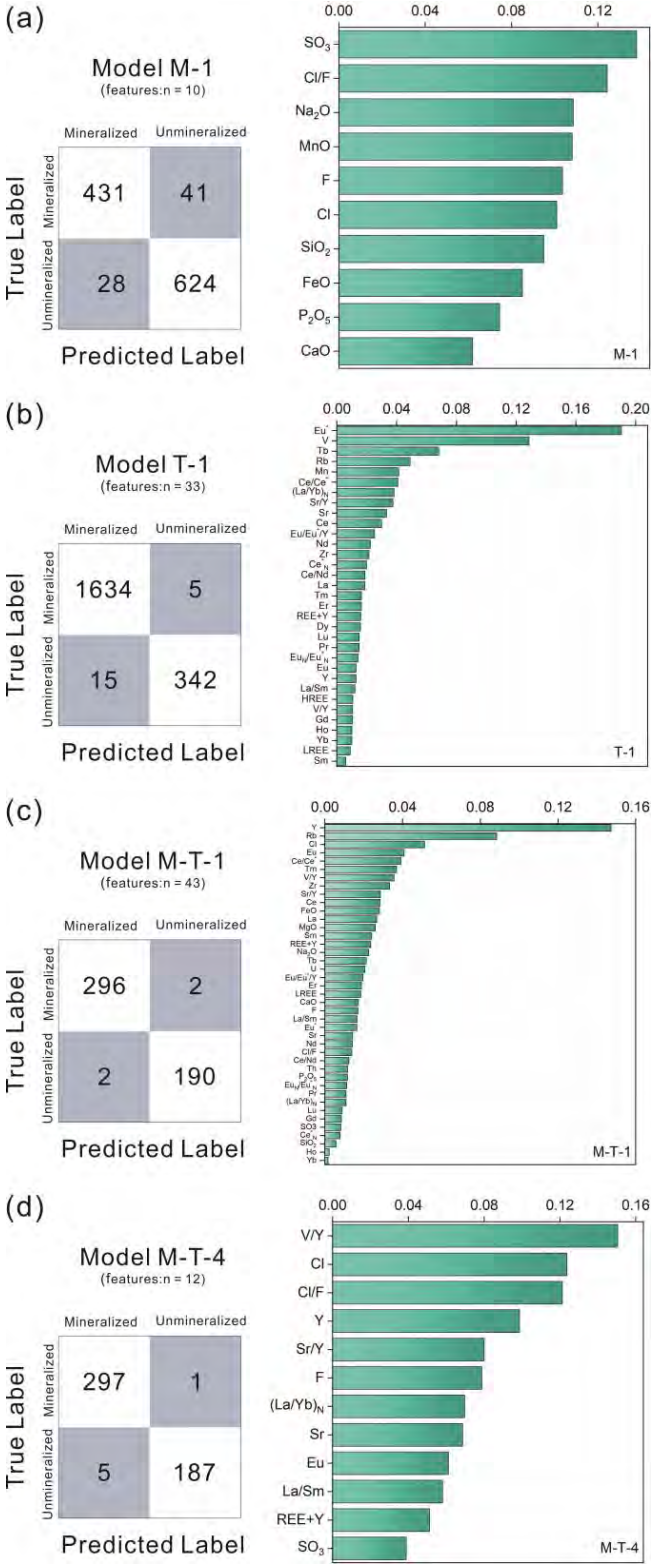
731 **FIGURE 3**



732

733

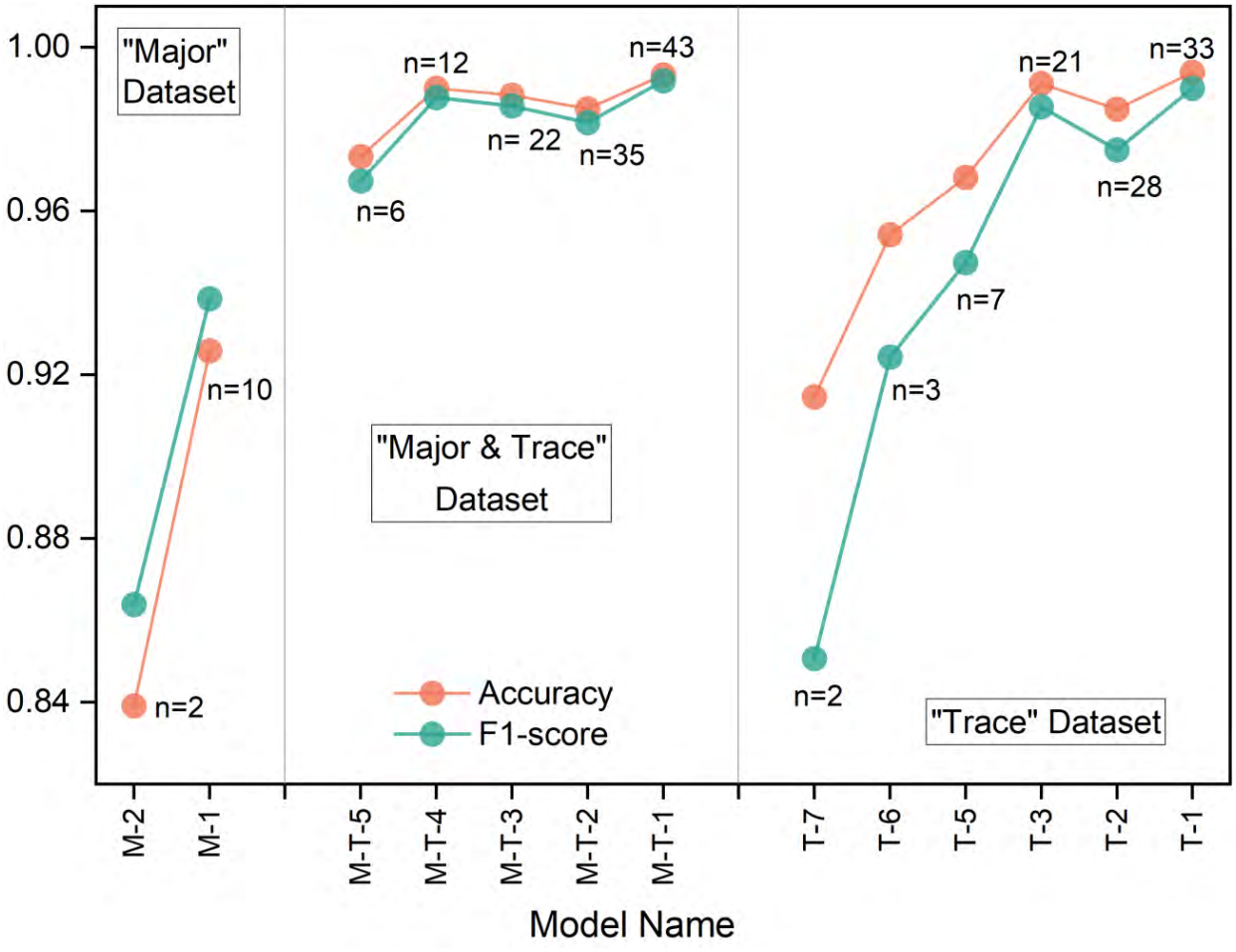
734 **FIGURE 4**



735

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

736 **FIGURE 5**



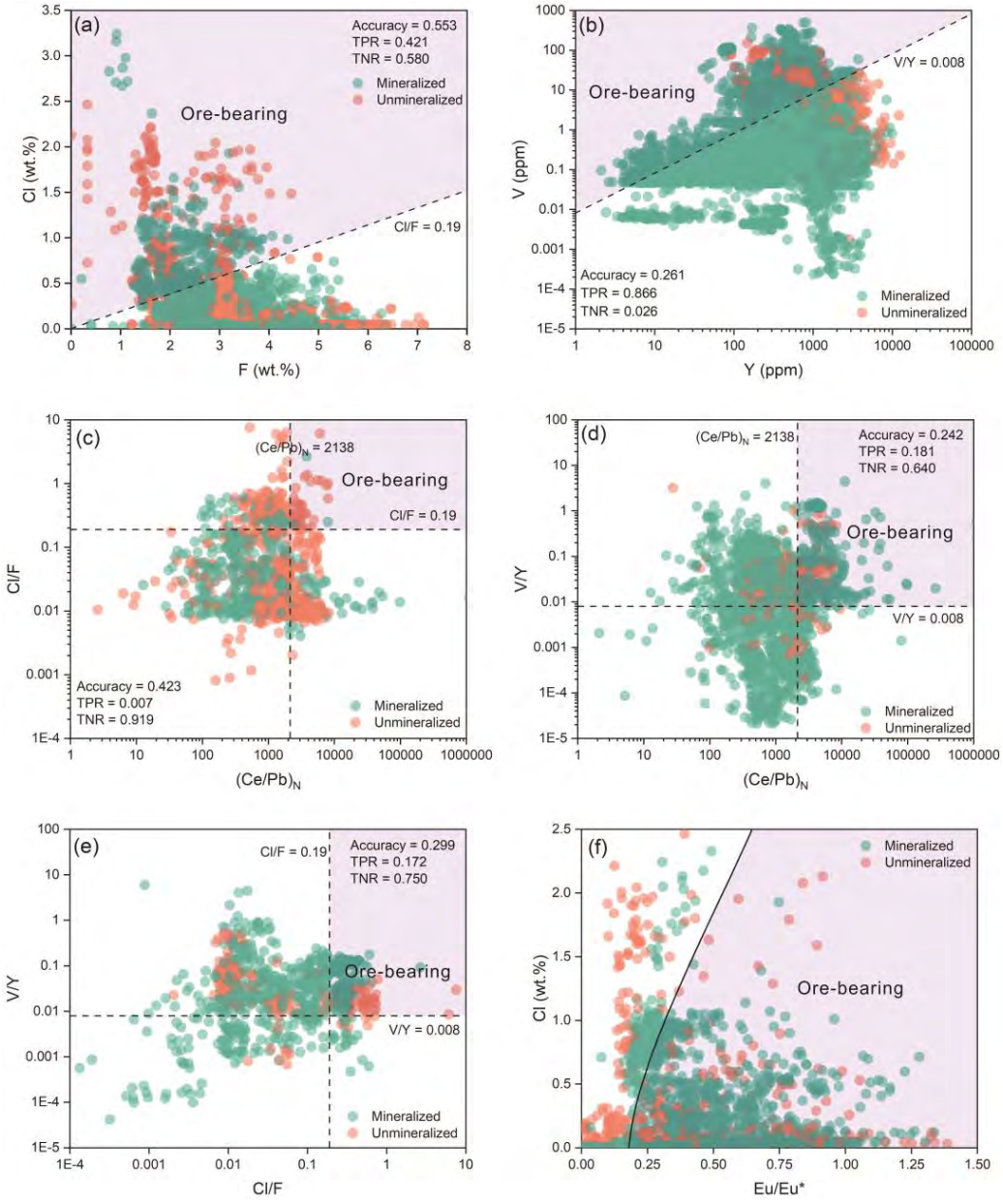
737

738

739

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

740 **FIGURE6**

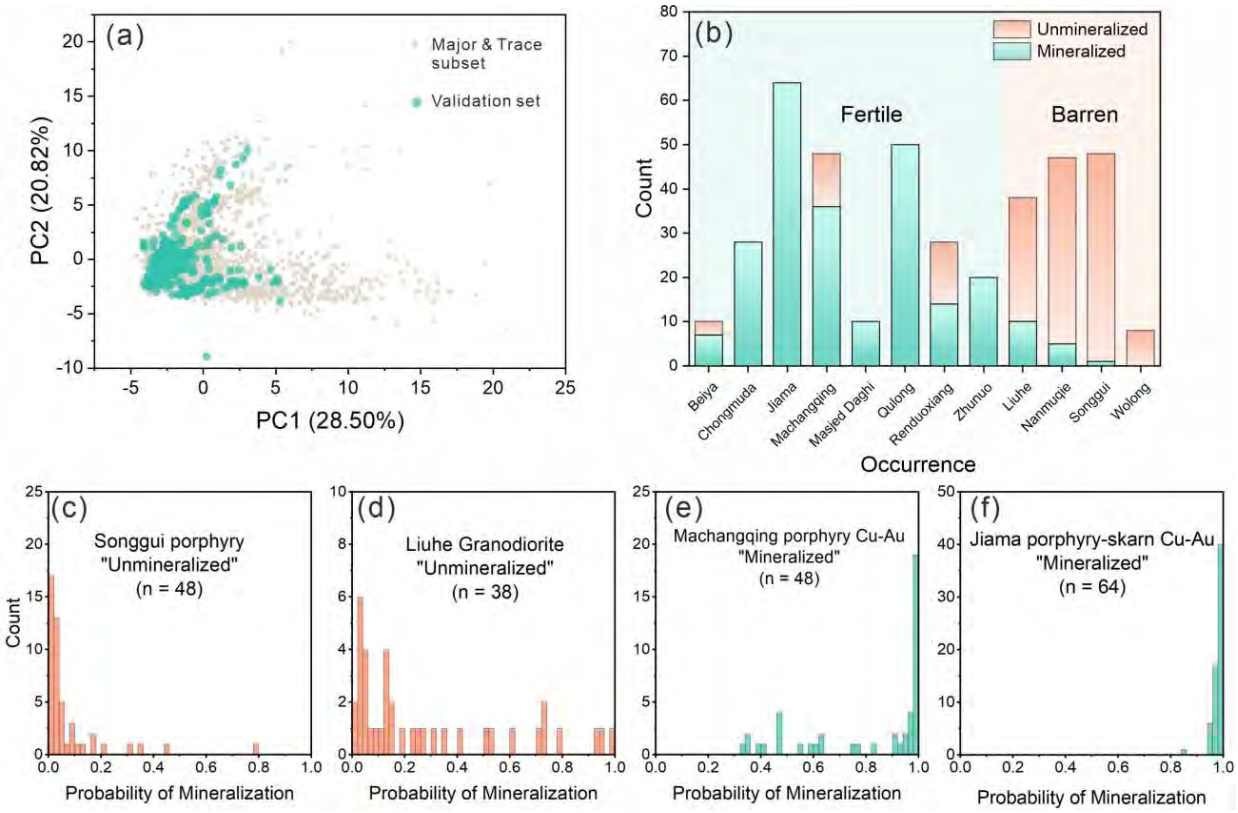


741

742

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2023-9115. http://www.minsocam.org/

743 **FIGURE 7**



744