DOI: https://doi.org/10.2138/am-2022-8899. http://www.minsocam.org/

Revision 2

A Machine Learning Approach to Discrimination of Igneous Rocks and Ore Deposits by Zircon Trace Elements

Word Count: 9214

Zi-Hao Wen[1,2,3], Lin Li[3,4], Christopher L. Kirkland[5], Sheng-Rong Li[3,6], Xiao-Jie Sun[7], Jia-Li Lei[8], Bo Xu[1,3] and Zeng-Qian Hou[9]

[1]*School of Gemology, China University of Geosciences, Beijing 100083, China*

[2]*Deutches GeoForschungsZentrum GFZ, Potsdam 14473, Germany*

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

[4]*Institute of Earth Science, China University of Geosciences, Beijing 100083, China*

[5]*Timescales of Mineral Systems Group, School of Earth and Planetary Sciences, Curtin University, Perth 6845, Australia*

[6]*School of Earth Science and Resources, China University of Geosciences, Beijing 100083, China*

[7]*China Telecom Co., LTD. Beijing Branch, Beijing 100010, China*

[8]*School of Earth Sciences, Zhejiang University, Hangzhou 310027, China*

[9]*Institute of Geology, Chinese Academy of Geological Sciences, Beijing 10037, China*

Corresponding author: Lin Li (clark.li@cugb.edu.cn)

ABSTRACT

The mineral zircon has a robust crystal structure, preserving a wealth of geological information through deep time. Traditionally, trace elements in magmatic and hydrothermal zircon have been employed to distinguish between different primary igneous or metallogenic growth fluids. However, classical approaches based on mineral geochemistry are not only time-consuming, but often ambiguous due to apparent compositional overlap for different growth environments. Here, we report a compilation of 11004 zircon trace element measurements from 280 published articles, 7173 from crystals in igneous rocks and 3831 from ore deposits. Geochemical variables include Hf, Th, U, Y, Ti, Nb, Ta, and the REEs. Igneous rock types include kimberlite, carbonatite, gabbro, basalt, andesite, diorite, granodiorite, dacite, granite, rhyolite and pegmatite. Ore types include porphyry Cu-Au-Mo, skarn-type polymetallic, intrusion-related Au, skarn-type Fe-Cu, and Nb-Ta deposits. We develop Decision Tree, XGBoost and Random Forest algorithms with this zircon geochemical information to predict lithology or deposit type. The F1-score indicates that the Random Forest algorithm has the best predictive performance for classification of both lithology and deposit type. The eight most important zircon elements from the igneous rock (Hf, Nb, Ta, Th, U, Eu, Ti, Lu,) and ore deposit (Y, Eu, Hf, U, Ce, Ti, Th, Lu) classification models, yielded reliable F1-scores of 0.919 and 0.891, respectively. We present a web page portal (http://60.205.170.161:8001/) for the classifier and employ it to a case study of Archean igneous rocks in Western Australia and ore deposits in Southwest China. The machine learning classifier successfully determines the known primary lithology of the samples, demonstrating significant promise as a classification tool, where host rock and ore deposit type is unknown.

Keywords: Zircon trace elements; igneous rocks classification; ore deposits classification; machine learning; Random Forests

INTRODUCTION

Zircon ($ZrSiO_4$) is common accessory mineral which grows in most silicate rocks and in many ore deposits. Zircon trace element chemistry reflects the partitioning of elements in the melt or fluid environment and the mineral during its crystallization (or later during recrystallization). Trace elements from a melt or other fluid can replace Zr, Si, or sit within interstitial spaces and become incorporated into the crystal during magmatic growth or during later metamorphism (Geisler et al. 2007; Hanchar et al. 2001; Hoskin and Schaltegger, 2003). Different trace elements within the zircon crystal record different information, for example the radioactive elements Th, U, and Pb can be used to calculate ages (Lee et al. 1997) but also retain crude relationships with magma fractionation state and bulk rock chemistry (e.g. Kirkland et al. (2015) on Th/U), Ti content is temperature dependent (Watson et al. 2006), Ce and Eu content is a key parameter related to magma oxygen fugacity (Trail et al. 2012), and Nb and Ta content reflects the degree of magmatic differentiation (Chen et al. 2021). Hf readily substitutes for Zr in the zircon structure meaning that the $^{176}Hf/^{177}Hf$ isotopic ratio, reflecting source Lu/Hf fractionation, is a powerful tool for crustal evolution studies (e.g. Belousova et al. 2010). A wide range of other geochemical parameters in zircon have been used to understand this mineral and hence a rocks crystallization and later alteration history (Bell et al. 2019; Claiborne et al. 2010; Olson et al. 2017; Zeng et al. 2017).

Studies on the classification of igneous rocks based on zircon compositions are abundant (Belousova et al. 2002; Breiter et al. 2014; Gudelius et al. 2020; Nardi et al. 2013). Utilizing a series of binary diagrams for zircon trace elements, Belousova et al. (2002) found that the content of specific elements varied between different igneous rock types. Belousova et al. (2002) used this information to construct a trace element Decision Tree to distinguish between potentially different igneous rocks precipitating zircon from their primary magma. Zircon composition has also been used as a pathfinder for mineralization (Lu et al. 2019), as there are differences in temperature, oxygen fugacity, water content, and magma fractionation state for barren and mineralized fluids which become encoded into zircons mineral chemistry. Porphyry-type Cu-Au-Mo deposits are commonly associated with intrusive bodies with high oxygen fugacity and water content (Lu et al. 2016). W-Sn deposits are associated with generally low oxygen fugacity (Yang et al. 2020). Nb-Ta deposits are often associated with highly evolved rocks (Yang et al. 2014). In the last 20 years, LA-ICPMS (Laser Ablation – Inductively Coupled Plasma Mass Spectrometry) has become a popular tool for both geochronology and geochemical analysis of zircon, allowing large datasets to be rapidly collected from relatively small sample volumes within individual zircon crystals (Jackson et al. 1992). As more zircon data are published, there is the potential to search for patterns within this "big data" and use the resulting information to address geological problems that may have lacked clear resolution with smaller datasets.

Machine learning is important in the context of "big data" and uses computational power to develop algorithms and statistical models to address a broad range of geological questions. With these algorithms and models, computer systems can process and analyze massive amounts of data in a short time, and make predictions or decisions on their own without explicit instructions (Mitchell 1997). Supervised learning is an important branch of machine learning, which predicts class labels by training a model. It requires input information to be labeled and divides it into a training dataset and a test dataset. The training dataset is used to teach the model and the test dataset serves to evaluate the performance of the constructed model (Hastie et al. 2009). Common supervised learning models include Decision Tree, Support Vector Machine (SVM), Random Forest, Extreme Gradient Boosting (XGBoost), and KNN. These models have already yielded some promising results for mineralogy. For example, models have been developed to predict the host rocks of quartz (Wang et al. 2021) and garnet (Schönig et al. 2021), tracing the possible provenance of detrital apatite in sedimentary rocks (O'Sullivan et al. 2020), and estimating the temperature and storage depth of clinopyroxene-bearing magma (Petrelli et al. 2020). For zircon grains, recently, Zou et al. (2022) successfully distinguished fertile and barren porphyries with the help of Random Forests and neural networks.

Distinct from Zou et al. (2022)'s study, here we aim to discriminate different magmatic rocks and different mineralizing fluids with zircon trace elements. Such classification model would have important use in provenance analysis of detrital zircon and ore prospecting. Specifically, with lithological context removed from a detrital zircon this tool may help refine provenance interpretations including lithology of the source (Hoskin and Ireland 2000), its potential geodynamic setting (Grimes et al. 2015), and expand the exploration search space for mineral systems (Lu et al. 2016). In this study, we prepared separate databases of zircon chemical composition for igneous rocks and ore deposits. We show that a Random Forests algorithm yields the best prediction for both igneous rock and ore deposit type. We also filtered the most significant elements from the compilation and developed a model using fewer variables which is able to achieve a similar classification effect. Compared with conventional methods, machine learning is both more efficient and reliable in classifying igneous rocks and ore deposits.

ZIRCON DATABASES AND CONVENTIONAL CLASSIFICATION METHODS

Zircon databases

We collected 11004 zircon trace element measurements from 280 published articles, with samples widely distributed over both space and time (**Fig. 1**). Part of the data is extracted from the online database https://data.goettingen-research-online.de. The elements in the database are Hf, U, Th, Y, Ti, Nb, Ta, and REE. Although zircon also contains P, Ca, Al, Fe, Sc, and Sr, the amount of data currently available for these elements is limited and thus is not yet suitable for inclusion in this form of analysis.

The igneous rock or ore deposit classification and primary publication is given in **Table 1**, and detailed zircon information can be found at https://github.com/ZihaoWen123/geology_class, including sample location, trace element contents and references. The Igneous Rocks Database includes nine different igneous rock types, with rock names extracted from the lithological descriptions within in the source publications. However, some of these samples have similar mineral assemblages. To improve classification efficiency, closely comparable mineral assemblages were integrated (**Table 1**). Ultimately, the Igneous Rock Database contains six discrete rock types: kimberlite, carbonatite, basic rocks (BR), intermediate rocks (IR), acid rocks (AR), and pegmatite. The Ore Deposit Database covers five discrete deposit types (**Table 1**): porphyry Cu-Au-Mo deposit, skarn-type polymetallic deposit, intrusion-related Au deposit, skarn-type Fe-Cu deposit and Nb-Ta deposit. Skarn-type polymetallic deposits in the database are mainly found in southern China and Southeast Asia, and are dominated by W, Sn, with minor Pb, Zn and Sb. Notably the above classification of igneous rocks and ore deposits is based on the description of field lithology and deposits in the published source articles.

Conventional classification methods

Before developing a machine learning method, we analyzed the zircon data from the igneous rocks (Belousova et al. 2002; Claiborne et al. 2010; Gagnevin et al. 2010; Gudelius et al. 2020) and ore deposits (Large et al. 2018; Lee et al. 2017; Lu et al. 2016) using more traditional two-dimensional classification methods (**Fig. 2**). REE depletion is regarded as an important feature of kimberlites (Hoskin and Ireland 2000). We found that not only REE (med. 299 ppm), but also Th (med. 33.5 ppm), U (med. 66.9 ppm) and Y (med. 248 ppm) are depleted in zircon crystals from kimberlites. In pegmatites Nb (med. 19.4 ppm), Ta (med. 8.32 ppm), REE (med. 1789ppm), U (med. 2123 ppm), Y (med. 2256 ppm) are all enriched (**Fig. 2a-c**). Niobium, Ta, and REE deposits are often associated with pegmatites (Van Lichtervelde et al. 2009; Seidler et al. 2005; Zhang et al. 2004), and some U deposits are found in areas which have significant pegmatite shows (Chen et al. 2019). Some Nb deposits are also spatially correlated with carbonatites (Melgarejo et al. 2012; Wu et al. 2021), as Nb (med. 52.5 ppm) is also enriched in carbonatites, while Ti (med. 3.03), REE (med. 519 ppm), U (med. 31.5 ppm), Y (med. 554 ppm) are generally deficient. **Figure 4d** shows the method proposed by (Grimes et al. 2007) for tracing zircon source area, which can constrain kimberlites but places few limits on the source of zircon from other rock types. The elemental contents of AR (Acid rock), IR (Intermediate rock), and BR (Basic rock) are not significantly enriched or depleted and all significantly overlap and cannot be uniquely identified via bivariate plots (**Fig. 2**).

Relevant to deposit formation, oxygen fugacity and water content are known to be related to the transport and deposition of metals (Wyborn et al., 1994). Some studies have found that Eu and Ce anomalies in zircon are controlled by magma temperature and also the crystallization of other minerals such as titanite, plagioclase, and hornblende, in addition to oxygen fugacity (Nathwani et al. 2021; Loader et al. 2022). Nonetheless, exploration approaches using Eu/Eu* (Dilles et al. 2015) and Ce* (Loader et al. 2017) have proved useful in distinguishing fertile from barren porphyry systems (Shen et al. 2015; Shu et al. 2019; Pizarro et al. 2020) (**Fig. 2e**). Recent studies have found that the water content of zircon crystals can be measured directly to estimate the amount of water within the primary magma (Xia et al. 2019). Another geochemical signature in zircon, with relevance for ores, is that water-rich magmas promote hornblende crystallization that suppresses plagioclase crystallization, resulting in Eu enrichment and Y deficiency in zircon. Lu et al. (2016) proposed that Eu/Eu*/Y×10000 and Ce/Nd/Y of zircon is positively correlated with magma water content (**Fig. 2f**). We find that skarn-type polymetallic deposits are associated with low oxygen fugacity and water content environments, while porphyry-type deposits, intrusion-related Au deposits, and skarn-type Fe-Cu deposits are associated with high oxygen fugacity and water content (**Fig. e, f**). Garnet is widespread in skarn rocks, which have a greater preference for HREE (Lee et al.

2017; Rubatto 2002). This chemical affinity may be responsible for the HREE deficit and low Yb/Gd ratios in zircons from skarn-type polymetallic deposits and skarn-type Fe-Cu deposits (**Fig. 2g**). In addition, zircons in Nb-Ta deposits unsurprisingly have high Nb and Ta contents (**Fig. 2h**).

DATA PRE-PROCESSING FOR MACHINE LEANING METHODS

Data pre-processing and model building was completed in Python on the scikit-learn platform (Pedregosa et al. 2011).

Addressing missing values – Imputation

In data analysis, data integrity is very important to obtain accurate and reliable results. Therefore, filling in missing values with appropriate estimates (imputation) is an essential step in data pre-processing. There are some missing compositional values in the data set, either because the elemental content was below the detection limit of the LA-ICPMS, because the analyst simply did not collect that element, or there was some other analytical limitation imposed on the acquisition.

For the first missing data case, we elected to remove elements with very low contents, such as La and Pr. Lanthanum and Pr contents are often below the detection limit and measurement of these elements are susceptible to reflecting the content of mineral inclusions within the zircon grains, rather than the zircon itself. These two elements were also avoided by other researchers, for example when calculating Ce^{3+} content from rare earth elements of zircon and estimating oxygen

DOI: https://doi.org/10.2138/am-2022-8899. http://www.minsocam.org/

fugacity (Zhong et al. 2019). In addition, we do not consider elements with >20% missing values in the dataset. This is because estimating a large number of missing values brings a heightened degree of uncertainty and could cause the model to poorly reflect the true data distribution. Niobium and Ta in the ore deposits database suffers from a large number of missing measurements.

For the second and third case of missing data (elements not measured for whatever reason), we are able to use the "knn-classification" and "iterative" vacancy filling methods as there is sufficient information to estimate the missing parameter in the dataset (Emmanuel et al. 2021). The term "knn-classification" uses the known characteristics of the data points to determine the nearest K samples to the missing data according to Euclidean distance (**Eqs. 1,2**), and then fills the missing values by averaging the results of these K samples (Emmanuel et al. 2021).

$$d_{xy} = \sqrt{(\text{weight} \times \text{squared distance from present coordinates})} \quad (1)$$

$$\text{weight} = (\text{Total number of coordinates})/(\text{Number of present coordinates}) \quad (2)$$

The alternative "iterative" method involves defining a model that predicts each missing element as a function of all other elements and repeating this process of estimating feature values multiple times (Emmanuel et al. 2021). Initially the procedure assumes that the missing data has a mean value. The concentration is then re-estimated based on the pattern within the entire dataset. The imputed values are used to update the missing values in the original data set. This repetition allows refined estimates for other features and can be used as the input in subsequent iterations of predicting missing values.

Data standardization

Data standardization unifies the units of measure and magnitudes of different features (in our case, elements), eliminating the effects of order-of-magnitude differences and making the data more comparable. We compared different data standardization strategies, including "Min-Max", "Log" and "Z-score".

"Min-Max" scales the original data in the range [0,1], i.e., to map the data to the specified interval by linear transformation of the original data (**Eq 3**).

$$x = (xi-\min(xi))/(\max(xi)-\min(xi)) \qquad 3$$

"Log" (log transformation) standardizes the data by taking the logarithm of the data (**Eq 4**).

$$x = \log(x_i+1) \qquad 4$$

The "Z-score" transforms the data into a data structure with mean of 0 and standard deviation equal to 1 (**Eq 5**). Where μ is the mean of the original data and σ is the standard deviation of the original data.

$$x = (x_i-\mu)/\sigma \qquad 5$$

Class imbalance

In the igneous rock database, the AR lithology has the most data with 2594 samples. The pegmatite lithology has the least data in the database with 218 samples (**Table 1**). In the case of the ore deposits database, the porphyry type Cu-Mo-Au deposit is the most numerous with 2122 samples and the Nb-Ta deposit is the least numerous with 81 samples (**Table 1**). This imbalance in the number of samples in different classes (lithology or deposit types) could cause the model to be more inclined to predict specific classes with more data and thus perform worse on classes with less data, resulting in biased model output (Japkowicz and Stephen 2002).

To address the apparent class imbalance a synthetic minority over-sampling technique (SMOTE) can be used (Chawla et al. 2002). This method first calculates the distance of each data point, in a minority class, from the adjacent K data. Then a number of data points are randomly selected from the K nearest neighbors to generate a new synthetic data point. This new synthetic data point is added to the original minority class dataset, increasing its number.

MACHINE LEARNING METHODS

Data is divided into training and testing sets with the training set : testing set ratio set at 9:1. The training set was used to develop the model and for parameter tuning. The test set was used to evaluate the performance of the model (Hastie et al. 2009). We developed Decision Tree (Myles et al. 2004), XGBoost (Chen and Guestrin 2016) and Random Forest (Tin Kam Ho 1995; Breiman 2001) to fit the compiled data. These methods are all tree-based algorithms, which are non-parametric and work regardless of the distribution/collinearity of the input data. Other methods, such as SVM, Artificial Neural Network, and Logistic Regression, can be limited compared to tree-based algorithms on geochemical data due to a

DOI: https://doi.org/10.2138/am-2022-8899. http://www.minsocam.org/

constant sum effect (Rollinson 1992).

Decision Tree

A Decision Tree model is often regarded as "weak classifier" and the basis for building integrated algorithms such as XGBoost and Random Forest. A Decision Tree is built by constructing a tree model that outputs the possible outcomes and probabilities under different conditions. Specifically, it selects the best feature from all the features as the root node and repeats this process for the selected features until a Decision Tree is generated (Myles et al. 2004). In the tree model, the Gini coefficient (**Eq 6**) is used for feature selection (Breiman 2001).

$$\mathrm{Gini}(t) = 1-\sum_{i=0}^{c-1} p(i|t)^2 \qquad 6$$

XGBoost

In the XGBoost algorithm the basic principle is to iteratively add Decision Trees to a model, with each tree attempting to correct the errors of the previous tree. During training, the model starts with a single Decision Tree and calculates the error (or loss) of the predictions on the training data. The algorithm then adds another Decision Tree to the model, but this time aims to correct the errors within the first Decision Tree. The combined output of both Decision Trees are then used to calculate a new error estimate, and the process repeats with additional Decision Trees added until the error is minimized. The predictions from each tree are combined by adding them together to produce the final output (Chen and He 2015; Chen and Guestrin 2016).

Random Forest

In a Random Forest model the algorithm builds a forest of Decision Trees, where each tree is constructed using a random subset of the data and features (**Fig. 3**). The trees are trained independently and are not correlated with each other. When making predictions, each Decision Tree in the forest is used to classify a given input, and the final prediction is made by averaging or taking the majority vote of the predictions from all the trees (**Eq. 7**) (Breiman, 2001). The algorithm can provide insight into the importance of each feature in the data during training by tracking the reduction in misclassification caused by each feature in each tree.

Equation 7 is the majority voting expression (Breiman 2001), H(x) denotes the combined classification model, h_i is the individual Decision Tree classification model Y denotes the output variable and I(·) is the indicative function.

$$H(x) = \arg\max \sum_{i=1}^{k} I(h_i(x)=Y) \qquad 7$$

Parameter tuning and cross validation

We adopt a Bayesian optimization algorithm to automatically adjust the parameters of the model (Snoek et al. 2012).

Five-fold cross-validation was employed to verify the reliability of the classification model (Hastie et al. 2009) (**Fig. 3**). This computational operation divides the data into five equal parts and takes one part at a time for validation with the remainder of the data set used for training the model. This calculation was repeated five times and the average computed.

RESULTS AND DISCUSSION

Traditional classification methods and its limitation

Despite our efforts to use our knowledge of geology to distinguish between the different rocks and deposits, there are still many overlapping areas in **Figure 2**. We take the Y-U plot for igneous rocks and the Eu/Eu*-Ce/Nd plot for ore deposits as examples, to calculate the accuracy of a conventional classification approach (**Fig. 4**). To avoid altered samples and select the most representative chemistry of a rock, the highest and lowest 5% of elemental concentration data were not considered. Figures 4b and 4d show examples of BR and porphyry-type Cu-Au-Mo deposits, respectively. First, we count the number of data points within overlapping intervals and also calculate the overlap rate on the X- and Y-axes. We then subtract the product of the two overlap rates from 1, which is the accuracy of identifying an igneous rock or ore deposit. BR reveals a complete overlap with an identification rate of 0 (**Fig. 4b**). The porphyry type Cu-Au-Mo deposit has 151 data distinguishable on a Ce/Nd plot, giving an identification rate of only 9% (**Fig. 4d**).

In **Figure 4a** even the most accurate classification rate, that for pegmatite, is only 64%, followed by carbonatite and kimberlite that are very similar with rates of 19% and 13% prediction, respectively. AR, IR, and BR are completely undistinguished. A similar result is evident in **Figure 4c**, which completely fails to discriminate between Nb-Ta deposits and skarn-type Fe-Cu deposits. The highest classification accuracy is for intrusion-related Au deposits, at 52%. Skarn-type polymetallic deposits and porphyry-type Cu-Au-Mo deposits are similar with only 11% and 9% prediction, respectively. In summary, traditional methods have generally poor performance in identifying different igneous rocks or ore deposits. It may be feasible to improve the identification of some rocks and deposits by making additional two-dimensional geochemical plots. However, such strategy will be both time consuming and may still be unable to uniquely distinguish between overlapping fields on discrimination plots and thus may lead to erroneous classifications.

Machine learning model construction

Before the selection of a machine learning algorithm, a lot of data pre-processing work is required, including treatment of missing values, data standardization, and addressing class imbalance. These steps aim to improve the accuracy, stability, and computational efficiency of the model. We ran the model on the compositional database with Decision Tree, XGBoost, and Random Forest algorithms and list the results in **Table 2**. Precision, recall, and F1-scores provide evaluation criteria for the classification models (see detailed description in Nathwani

DOI: https://doi.org/10.2138/am-2022-8899. http://www.minsocam.org/

et al. (2022)). The F1-score is the summed average of precision and recall, and is thus a useful summary of the function of the model. We observe that for igneous rocks and ore deposits, the best results are obtained by using the "knn-classification" method of filling in missing values, the "z-score" method for data standardization, and with "SMOTE" for class balance.

Improperly filling in the missing values would introduce new noisy data, increasing the uncertainty of the model, leading to biased results (Pearson 2006). In our models, "knn-classification" performs better than the "iterative" imputation method (**Table 2**). A possible reason for this observation is that the KNN algorithm is a similarity-based algorithm, and as the same sample group of data has a high similarity, so the "knn-classification" works better. A disadvantage of the "iterative" method is that it is computationally intensive. For data standardization, both classification models perform best with the "z-score". This may be because the “z-score” method can better preserve the information between features, avoid the influence of outliers, and does not change the shape of the original data. For class imbalance, “SMOTE” effectively increased the number of minority samples and improved their identification.

For machine learning algorithms, Random Forest performs the best for both databases no matter what data preprocessing method was used (**Table 2**). It is conceivable that the Decision Tree algorithm does not perform well because Random Forest and XGBoost are integrated algorithms and they are better at

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-2022-8899. http://www.minsocam.org/

handling data with a high level of dimensions (i.e. a large number of attributes within the dataset). The lower F1-score of XGBoost than Random Forest may be due to its tendency to overfit the data. Random Forest randomly selects some features in the training of each Decision Tree, avoiding possible overfitting caused by too many features.

Bayesian optimization is employed to parameterize the best igneous rocks and ore deposits models. It improves the predictive performance and accuracy of the model, reduces the risk of overfitting or underfitting, and improves the generalization ability of the model (Snoek et al. 2012). In **Table 3**, we list the parameter combinations (Detailed parameter tunning results in GitHub). The F1-scores of both igneous rocks and ore deposits classification models are significantly improved with the optimization, with scores of 0.963 and 0.961, respectively. The results of the five-fold cross-validation show that for Random Forest (**Table 4**) the precision of the classification models for igneous rocks and ore deposits has mean values of 0.947 and 0.897 respectively, suggesting that the classification models are both stable and reliable.

A confusion matrix was used to measure the performance of the classification model. We can see from **Figure 5a** that kimberlite has the highest value (0.959), followed by AR (0.938), IR (0.891), BR (0.882), carbonatite (0.87) and pegmatite (0.75). Some pegmatites are mistaken for AR (0.125), which may be due to the fact that they underwent a longer chemical evolution sharing ultimate compositional affinity to AR. For the Ore Deposits Database (**Fig. 5b**), porphyry-type Cu-Au-Mo deposits (0.945) and intrusion-related Au deposits (0.95) have a better precision, followed by Nb-Ta deposits (0.909) and skarn-type polymetallic deposits (0.841), with skarn-type Fe-Cu deposits being the lowest (0.712). Skarn-type Fe-Cu deposits can be mistaken for porphyry-type Cu-Au-Mo deposits (0.076) and intrusion-related Au deposits (0.076) and polymetallic silica deposits (0.136). The lower scores may be because both skarn-type Fe-Cu deposits and skarn-type polymetallic deposits are spatially associated with the same geological environment. However, skarn-type Fe-Cu deposits prefer an oxidized and H_2O-rich environment, as does porphyry-type Cu-Au-Mo deposits, and intrusion-related Au deposits (Sun et al. 2019).

Feature importance and model simplification

Feature importance highlights how relevant a feature (e.g. trace elements in a zircon) is to the classification (e.g. the type of igneous rock or ore deposit). Permutation Feature Importance (PFI) is a method for assessing the importance of features (Altmann et al. 2010). It evaluates the influence of the feature on the model by randomly replacing the value of a feature (Altmann et al. 2010).

For the igneous rocks and ore deposits classification models, 19 and 17 (Nb, Ta missing values >20% were not included in the model) elements were taken into account, respectively. In **Figures 6** we present the importance scores of the features for the Igneous Rocks and Ore Deposits Databases, respectively. In the igneous classification model Hf (0.123) is considered to be the most important, followed by Nb (0.120), Ta (0.089), Th (0.086) etc. and Sm (0) is considered to be the least important. In the deposit classification model, Y (0.119) is the most important, followed by Eu (0.097) ,Hf (0.067), U (0.067) etc. There are also some elements that are negative values, and they are usually considered to have a negative impact on the model, with Gd (-0.006) having the biggest negative impact.

To explore the relationship between the number of elements and the model scores, we first selected the top two most important elements and then added elements in descending order (**Figure 7**). The F1-score of the rock classification model increases from 0.612 to 0.902, while the deposit classification model increased from 0.478 to 0.851 until the 8th element was added. This is very close to the scores obtained with all elements in the Igneous Rock Database (0.914) and Ore Deposit Database (0.868). Therefore, we consider it acceptable to use the most important eight elements for the igneous rock and ore deposit classification models, respectively. Such approach aids in the decision of what trace elements to analyze in zircon when the goal is classifying the igneous rock source or ore deposits host, saving analytical time and costs, but arguably most importantly allowing element count times to be optimized to those most powerful elements for classification. From an algorithmic standpoint, using fewer elements in the final model will reduce its susceptibility to overfitting the training set (i.e. increases the signal to noise ratio). We additionally performed Bayesian optimal tuning for the simplified model (**Table**

3), which yielded F1-scores for igneous rock and deposit classification models of 0.919 and 0.891, respectively.

Both the simplified igneous rock and ore deposit classification models contain Hf, Th, U, Eu, Ti, and Lu. The igneous rock model also contains Nb and Ta, whereas the ore deposit model also contains Y and Ce. From a geological perspective, the contents of Hf, Y, U, Th, Nb, Ta, and Lu are known to correlate with degree of magmatic evolution. Fluorine is typically abundant in evolved magmas and zircon crystals generated in such fluids (Breiter et al. 2006). Zr/Hf (Claiborne et al. 2006), Hf/Y (Gagnevin et al. 2010), Th/U (Claiborne et al. 2006; Gagnevin et al. 2010; Kirkland et al. 2015) and Nb/Ta (Gudelius et al. 2020) ratios also evaluate the degree of magma fractionation. Cerium and Eu have variable valences and thus estimate magma oxygen fugacity (Ballard et al. 2002; Loader et al. 2017; Zhong et al. 2019). Europium and Y in zircon may also reflect water content in the magma (Triantafyllou et al., 2022). Oxygen fugacity and water content track the migration and potential enrichment of metals in the crust (Dilles et al. 2015; Lu et al. 2016). Hence, the most important elements selected by the PFI algorithm appear geologically significant with established relationships to both magmatic evolution and ore deposit formation.

The classification models for igneous rocks and ore deposits, discussed above, based on the eight most important elements, is provided via a web page front end http://60.205.170.161:8001/. Users can select the most appropriate model for

DOI: https://doi.org/10.2138/am-2022-8899. http://www.minsocam.org/

classification and upload their zircon compositional data. The model outputs the counts per classification (also expressed as a percentage of the total number of samples). A detailed results spreadsheet can be downloaded which appends the classification onto the input file.

Case study of igneous rocks and ore deposits classification model

Igneous rocks in Yilgarn Craton, western Australia

In order to explore the performance of the machine learning model we apply it to a case study on magmatic zircon crystals from the Archean Yilgarn Craton of Western Australia. The Yilgarn Craton has an exposed area of about 65×10^4 km^2 and is well endowed with a range of different mineral systems (Cassidy et al. 2006) (**Figure 8a**). We consider a compilation of zircon geochemical data collected by LA-ICPMS which is paired with whole rock geochemistry (Lu et al., 2019). This dataset has been used to evaluate the zircon trace element content of barren granitic rocks to that paragenetically associated with mineralization. Zircon grains were filtered for U-Pb isotopic discordance as a means to exclude those that would have seen secondary alteration effects. The whole rock dataset has been filtered to include only samples with loss on ignition values <63 wt% and Al_2O_3 <20 wt%. This filtering aims to exclude samples that are strongly altered or are plagioclase cumulates. Some samples were also excluded due to the effects of metamorphism. Whole-rock geochemical and zircon trace element data for 30 rocks in the Yilgarn

Craton (https://github.com/ZihaoWen123/geology_class) reflects primary compositions and is available to test the classification methods (see Lu et al. 2019). First, we classified these rocks using traditional methods based on whole-rock geochemistry, 12 are "granodiorite" and 18 are "granite" fields according to the TAS diagram (Le Maitre et al 2002), indicating that they are mainly intermediate-acid to acid rocks. We used the zircon trace elements from these 30 samples in the classification model and list the results with the whole-rock geochemical classification results in **Table 5** for comparison. The classification model indicates the rock type predicted by each zircon trace element analysis and can be expressed as the proportion of each rock type classified within any sample, as shown in the pie chart in **Figure 8b**. It is clear that the zircon based IR classification is dominant in the whole rock defined "granodiorite" field, and the AR classification is elevated in those defined by whole rock as "granite" (**Fig. 8b**). As with the classification results of the whole-rock geochemical measurements, the lithology classification model based on zircon trace elements correctly predicts that these igneous rocks are mainly intermediate-acid in composition. An obvious application of this approach would be to detrital zircon grains that are not in association with their primary magmatic source rock. The zircon classification model would enable a prediction on the most likely source lithology.

Ore deposits in Sanjiang region, southwest China

The Sanjiang metallogenic belt, located in southwestern China (**Fig. 9a**), is one of the most important polymetallic belts in China which includes several porphyry copper-gold and polymetallic skarn deposits (Hou et al. 2007; Xu et al. 2021). We compiled zircon compositional data (find data on https://github.com/ZihaoWen123/geology_class) from the Yangla skarn-type polymetallic deposit, the Pulang, and the Beiya porphyry-type Cu-Au deposits (**Fig. 9b**). Zircon compositions were used to determine the deposit type following the deposit classification model discussed above. The Yangla polymetallic skarn deposit, formed in the Triassic-Early Jurassic (Wang et al. 2022). It was traditionally considered as a copper deposit, but a high-grade tungsten ore in this deposit was recently identified (Yang et al. 2023). Wang et al. (2022) studied a quartz diorite from this deposit. In **Figure 9c**, the deposit classification model gives predictions for three zircon populations from this quartz diorite. Skarn-type polymetallic deposits are the dominant classification, consistent with the known situation. In the same area, Pulang and Beiya are two super large porphyry-type Cu and Au deposits formed in the Early Jurassic and Eocene, respectively (**Fig. 9b**) (Meng et al. 2018). Zircon compositional data from Meng et al. (2018) was used in classification. Three zircon populations of the Pulang deposit and five of the Beiya deposit yielded classifications dominated by porphyry-type Cu-Au-Mo deposits (**Fig. 9c**). It is notable that porphyry Cu-Au-Mo deposits and skarn-type polymetallic deposits always ranked within the top two for number of classifications. In summary, the zircon composition based ore deposit classification model seems to offer a useful indication of the potential mineralization type within an area.

CONCLUSIONS

Here we show that traditional methods of classifying magmatic rocks and deposits using zircon trace elements is inefficient at best and at worst can lead to misclassification. Random Forest models are an efficient multi-dimensional computation algorithm, although such classification results are difficult to show in the form of a flow chart. Many Decision Trees are computed independently, which can save computation time. Even if we use only the most important eight elements to predict igneous rock and ore deposit types, this limited compositional information still enables good classification. A case study of igneous rocks in the Yilgarn Craton and ore deposits in the Sanjiang region demonstrates that the zircon classifier has its own unique advantages in terms of ease of use and accuracy. It offers significant potential for tracing the origin of detrital zircon grains and enhancing exploration search space by indicating metallogenic fluids.

IMPLICATIONS

Zircon is a stable mineral that can preserve primary geological information and previous studies have confirmed that trace elements in this mineral are effective for tracing the origin of both igneous rocks and ore deposits. With large compilations of trace element data in zircon machine learning offers an attractive proposition to classifying igneous rocks and ore deposits source based on grain chemistry. Here we collect 7173 zircon chemical data from 11 different igneous rock types and 3831 analyses of 5 deposit types, worldwide. Based on this computational approach we identify the 8 most important zircon trace elements that influence zircon classification in igneous rocks and ore deposits. We then build classification models for both igneous rocks and ore deposits and validate their reliability. In addition, a web page portal (http://60.205.170.161:8001/) has been developed for the two (igneous / deposit) classification models. The approach is applied to a case study of zircon from known rock types in 30 igneous plutons from Western Australia. Classification models of igneous rocks and ore deposits using zircon chemical data will be clearly useful in tracing the provenance of detrital zircon grains and in reducing exploration risk by increasing the deposit halo in detrital zircon sampling surveys.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (Grant no. 42202084, 4222304, 91962101, 41603063, 41872038 and IGCP-662) and DDE Big Science Program. This is the 16th contribution of Bo Xu for National Mineral Rock and Fossil Specimens Resource Center. We thank Dr. Ci-Jian Yang of the Deutsches GeoForschungsZentrum, Dr. Zhangzhou J. of Zhejiang University, and Dr. Liu Haiming of the Chinese Academy of Geological Science for constructive comments. We also thank Chetan L. Nathwani and an anonymous reviewer for comments that improved this work.

REFERENCE

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010) Permutation importance: A corrected feature importance measure. Bioinformatics, 26, 1340–1347.

Bell, E.A., Boehnke, P., Barboni, M., and Harrison, T.M. (2019) Tracking chemical alteration in magmatic zircon using rare earth element abundances. Chemical Geology, 510, 56–71.

Belousova, E.A., Griffin, W.L., O'Reilly, S.Y., and Fisher, N.I. (2002) Igneous zircon: Trace element composition as an indicator of source rock type. Contributions to Mineralogy and Petrology, 143, 602–622.

Belousova, E.A., Kostitsyn, Y.A., Griffin, W.L., Begg, G.C., O'Reilly, S.Y., and Pearson, N.J. (2010) The growth of the continental crust: Constraints from zircon Hf-isotope data. Lithos, 119, 457–466.

Breiter, K., Lamarão, C.N., Borges, R.M.K., and Dall'Agnol, R. (2014) Chemical characteristics of zircon from A-type granites and comparison to zircon of S-type granites. Lithos, 192–195, 208–225.

Cassidy, K.F., Champion, D.C., Krapez, B., Barley, M.E., Brown, S.J.A., Blewett, R.S., Groenewald, P.B., and Tyler, I.M. (2006) A revised geological framework for the Yilgarn Craton, Western Australia, 8pp p. Western Australia Geological Survey.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002) Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Chen, T., and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. In In

Kdd '16: Proceedings of the 22nd ACM SIGKDD Inter- national Conference on Knowledge Discovery and data mining pp. 785–794. Association for Computing Machinery.

Chen, T., and He, T. (2015) xgboost: eXtreme Gradient Boosting. R package version 04-2.

Chen, W., Zhang, G., Ruan, M., Wang, S., and Xiong, X. (2021) Genesis of Intermediate and Silicic Arc Magmas Constrained by Nb/Ta Fractionation. Journal of Geophysical Research: Solid Earth, 126, 1–15.

Chen, Y., Hu, R., Bi, X., and Luo, J. (2019) Genesis of the Guangshigou pegmatite-type uranium deposit in the North Qinling Orogenic Belt, China. Ore Geology Reviews, 115, 103165.

Claiborne, L.L., Miller, C.F., and Wooden, J.L. (2010) Trace element composition of igneous zircon: A thermal and compositional record of the accumulation and evolution of a large silicic batholith, Spirit Mountain, Nevada. Contributions to Mineralogy and Petrology, 160, 511–531.

Dilles, J.H., Kent, A.J.R., Wooden, J.L., Tosdal, R.M., Koleszar, A., Lee, R.G., and Farmer, L.P. (2015) Zircon compositional evidence for sulfur-degassing from ore-forming arc magmas. Economic Geology, 110, 241–251.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., and Tabona, O. (2021) A survey on missing data in machine learning. Journal of Big Data, 8.

Gagnevin, D., Daly, J.S., and Kronz, A. (2010) Zircon texture and chemical composition as a guide to magmatic processes and mixing in a granitic environment and coeval volcanic system. Contributions to Mineralogy and Petrology, 159, 579–596.

Geisler, T., Schaltegger, U., and Tomaschek, F. (2007) Re-equilibration of zircon in aqueous fluids and melts. Elements, 3, 43–50.

Grimes, C.B., John, B.E., Kelemen, P.B., Mazdab, F.K., Wooden, J.L., Cheadle, M.J., Hanghøj, K., and Schwartz, J.J. (2007) Trace element chemistry of zircons from oceanic crust: A method for distinguishing detrital zircon provenance. Geology, 35, 643–646.

Grimes, C.B., Wooden, J.L., Cheadle, M.J., and John, B.E. (2015) "Fingerprinting" tectono-magmatic provenance using trace elements in igneous zircon. Contributions to Mineralogy and Petrology, 170, 1–26.

Gudelius, D., Zeh, A., and Wilson, A.H. (2020) Zircon formation in mafic and felsic rocks of the Bushveld Complex, South Africa: Constraints from composition, zoning, Th/U ratios, morphology, and modelling. Chemical Geology, 546, 119647.

Hanchar, J.M., Finch, R.J., Hoskin, P.W.O., Watson, E.B., Cherniak, D.J., and Mariano, A.N. (2001) Rare earth elements in synthetic zircon: Part 1. Synthesis, and rare earth element and phosphorus doping. American Mineralogist, 86, 667–680.

Hastie, T., Tibshirani, R., and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining,Inference,and Prediction, 762 p. Springer.

Hoskin, P.W.O., and Ireland, T.R. (2000) Rare earth element chemistry of zircon and its use as a provenance indicator. Geology, 28, 627–630.

Hoskin, P.W.O., and Schaltegger, U. (2003) The composition of zircon and igneous and metamorphic petrogenesis. Reviews in Mineralogy and Geochemistry, 53, 27–62.

Hou, Z., Zaw, K., Pan, G., Mo, X., Xu, Q., Hu, Y., and Li, X. (2007) Sanjiang Tethyan metallogenesis in S.W. China: Tectonic setting, metallogenic epochs and deposit types. Ore Geology Reviews, 31, 48–87.

Jackson, S.E., Longerich, H.P., Dunning, G.R., and Fryer, B.J. (1992) The application of laser-ablation microprobe - inductively coupled plasma-mass spectrometry (LAM-ICP-MS) to in situ trace-element determinations in minerals. Canadian Mineralogist, 30, 1049–1064.

Japkowicz, N., and Stephen, S. (2002) The class imbalance problem: A systematic study. Intelligent Data Analysis, 6, 429–449.

Kirkland, C.L., Smithies, R.H., Taylor, R.J.M., Evans, N., and McDonald, B. (2015) Zircon Th/U ratios in magmatic environs. Lithos, 212–215, 397–414.

Large, S.J.E., Von Quadt, A., Wotzlaw, J.F., Guillong, M., and Heinrich, C.A. (2018) Magma evolution leading to porphyry Au-Cu mineralization at the Ok Tedi deposit, Papua New Guinea: Trace element geochemistry and high-precision geochronology of igneous zircon. Economic Geology, 113, 39–61.

Le Maitre, R.W. (2002). Igneous Rocks: a Classification and Glossary of Terms: Recommendations of the International Union of Geological Sciences Subcommission on the Systematics of Igneous Rocks. Cambridge University Press, Cambridge. 236 pp.

Lee, J.K.W., Williams, I.S., and Ellis, D.J. (1997) Pb, U and Th diffusion in natural zircon. Nature, 390, 159–162.

Lee, R.G., Dilles, J.H., Tosdal, R.M., Wooden, J.L., and Mazdab, F.K. (2017) Magmatic evolution of granodiorite intrusions at the El Salvador porphyry copper deposit, Chile, based on trace element composition and U/Pb age of zircons. Economic Geology, 112, 245–273.

Loader, M.A., Wilkinson, J.J., and Armstrong, R.N. (2017) The effect of titanite crystallisation on Eu and Ce anomalies in zircon and its implications for the assessment of porphyry Cu deposit fertility. Earth and Planetary Science Letters, 472, 107–119.

Loader, M.A., Nathwani, C.L., Wilkinson, J.J., and Armstrong, R.N. (2022) Controls on the magnitude of Ce anomalies in zircon. Geochimica et Cosmochimica Acta, 328, 242–257.

Lu, Y., Loucks, R.R., Fiorentini, M., Mccuaig, T.C., Evans, N.J., Yang, Z., Hou, Z., Kirkland, C.L., Parra-avila, L.A., and Kobussen, A. (2016) Zircon Compositions as a Pathfinder for Porphyry Cu ± Mo ± Au Deposits. Economic

Geology Special Publication, 19, 329–347.

Lu, Y., Evans, N., and Campbell McCuaig, T. (2019) Zircon fingerprinting of magmatic-hydrothermal systems in the Archean Yilgarn Craton, 1–22 p.

Melgarejo, J.C., Costanzo, A., Bambi, A.C.J.M., Gonçalves, A.O., and Neto, A.B. (2012) Subsolidus processes as a key factor on the distribution of Nb species in plutonic carbonatites: The Tchivira case, Angola. Lithos, 152, 187–201.

Meng, X., Mao, J., Zhang, C., Zhang, D., and Liu, H. (2018) Melt recharge, f O2-T conditions, and metal fertility of felsic magmas: zircon trace element chemistry of Cu-Au porphyries in the Sanjiang orogenic belt, southwest China. Mineralium Deposita, 53, 649–663.

Mitchell, T.M. (1997) Machine Learning, 264 p. McGraw-Hill.

Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., and Brown, S.D. (2004) An introduction to decision tree modeling. Journal of Chemometrics, 18, 275–285.

Nardi, L.V.S., Formoso, M.L.L., Müller, I.F., Fontana, E., Jarvis, K., and Lamarão, C. (2013) Zircon/rock partition coefficients of REEs, Y, Th, U, Nb, and Ta in granitic rocks: Uses for provenance and mineral exploration purposes. Chemical Geology, 335, 1–7.

Nathwani, C.L., Simmons, A.T., Large, S.J.E., Wilkinson, J.J., Buret, Y., and Ihlenfeld, C. (2021) From long-lived batholith construction to giant porphyry copper deposit formation: petrological and zircon chemical evolution of the Quellaveco District, Southern Peru. Contributions to Mineralogy and Petrology, 176.

Nathwani, C.L., Wilkinson, J.J., Fry, G., Armstrong, R.N., Smith, D.J., and Ihlenfeld, C. (2022) Machine learning for geochemical exploration: classifying metallogenic fertility in arc magmas and insights into porphyry copper deposit formation. Mineralium Deposita, 57, 1143–1166.

O'Sullivan, G., Chew, D., Kenny, G., Henrichs, I., and Mulligan, D. (2020) The trace element composition of apatite and its application to detrital provenance studies. Earth-Science Reviews, 201, 103044.

Olson, N.H., Dilles, J.H., Kent, A.J.R., and Lang, J.R. (2017) Geochemistry of the Cretaceous Kaskanak Batholith and genesis of the Pebble porphyry Cu-Au-Mo deposit, Southwest Alaska. American Mineralogist, 102, 1597–1621.

Pearson, R.K. (2006) The problem of disguised missing data. ACM SIGKDD Explorations Newsletter, 8, 83–92.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and others (2011) Scikit-learn: Machine learning in Python. Journalof Machine Learning Research, 12, 2825–2830.

Petrelli, M. (2021) Introduction to Python in Earth Science Data Analysis, 229 p. Springer Textbooks in Earth Sciences, Geography and Environment. Springer International Publishing.

Petrelli, M., Caricchi, L., and Perugini, D. (2020) Machine Learning Thermo-Barometry: Application to Clinopyroxene-Bearing Magmas. Journal of Geophysical Research: Solid Earth, 125.

Pizarro, H., Campos, E., Bouzari, F., Rousse, S., Bissig, T., Gregoire, M., and Riquelme, R. (2020) Porphyry indicator zircons (PIZs): Application to exploration of porphyry copper deposits. Ore Geology Reviews, 126, 103771.

Rollinson, H.R. (1992) Another look at the constant sum problem in geochemistry. Mineralogical Magazine, 56, 469–475.

Rubatto, D. (2002) Zircon trace element geochemistry: Partitioning with garnet and the link between U-Pb ages and metamorphism. Chemical Geology, 184, 123–138.

Schönig, J., von Eynatten, H., Tolosana-Delgado, R., and Meinhold, G. (2021) Garnet major-element composition as an indicator of host-rock type: a machine learning approach using the Random Forest classifier. Contributions to Mineralogy and Petrology, 176, 1–21.

Seidler, J.K., Lentz, D.R., Hall, D.C., and Susak, N. (2005) Zircon-rich Ta-Nb-REE mineralization in the McKeel Lake pegmatite-aplite system, Welsford intrusion, southwestern New Brunswick. Exploration and Mining Geology, 14, 79–84.

Shen, P., Hattori, K., Pan, H., Jackson, S., and Seitmuratova, E. (2015) Oxidation condition and metal fertility of granitic magmas: Zircon trace-element data from porphyry cu deposits in the central asian orogenic belt. Economic Geology, 110, 1861–1878.

Shu, Q., Chang, Z., Lai, Y., Hu, X., Wu, H., Zhang, Y., Wang, P., Zhai, D., and Zhang, C. (2019) Zircon trace elements and magma fertility: insights from porphyry (-skarn) Mo deposits in NE China. Mineralium Deposita, 645–656.

Snoek, J., Larochelle, H., and Adams, R.P. (2012) Practical Bayesian Optimization of Machine Learning Algorithms. Advances in neural information processing systems.

Tin Kam Ho (1995) Random Decision Forests Tin Kam Ho Perceptron training. Proceedings of 3rd International Conference on Document Analysis and Recognition, 278–282.

Trail, D., Bruce Watson, E., and Tailby, N.D. (2012) Ce and Eu anomalies in zircon as proxies for the oxidation state of magmas. Geochimica et Cosmochimica Acta, 97, 70–87.

Triantafyllou, A., Ducea, M.N., Jepson, G., Hernández-Montenegro, J.D., Bisch, A., and Ganne, J. (2023) Europium anomalies in detrital zircons record major transitions in Earth geodynamics at 2.5 Ga and 0.9 Ga. Geology, 51, 141–145.

Van Lichtervelde, M., Melcher, F., and Wirth, R. (2009) Magmatic vs. hydrothermal origins for zircon associated with Tantalum mineralization in the Tanco pegmatite, Manitoba, Canada. American Mineralogist, 94, 439–450.

Wang, X., Li, B., Tang, G., Lei, Z., and Chang, H. (2022) Petrogenesis and Metallogenesis of Granitoids in the Yangla Cu-W Polymetallic Deposit, Southwest China: Evidence from Zircon Trace Elements and Hf Isotope.

DOI: https://doi.org/10.2138/am-2022-8899. http://www.minsocam.org/

Minerals, 12.

Wang, Y., Qiu, K.F., Müller, A., Hou, Z.L., Zhu, Z.H., and Yu, H.C. (2021) Machine Learning Prediction of Quartz Forming-Environments. Journal of Geophysical Research: Solid Earth, 126, 1–11.

Watson, E.B., Wark, D.A., and Thomas, J.B. (2006) Crystallization thermometers for zircon and rutile. Contributions to Mineralogy and Petrology, 151, 413–433.

Wu, B., Hu, Y.Q., Bonnetti, C., Xu, C., Wang, R.C., Zhang, Z.S., Li, Z.Y., and Yin, R. (2021) Hydrothermal alteration of pyrochlore group minerals from the Miaoya carbonatite complex, central China and its implications for Nb mineralization. Ore Geology Reviews, 132, 104059.

Wyborn, L.A.I., Heinrich, C.A., and Jaques, A.L. (1994) Australian Proterozoic mineral systems: Essential ingredients and mappable criteria: Proceedings of the Australasian Institute of Mining and Metallurgy. In The AusIMM Annual Conference. Darwin.

Xia, X.P., Cui, Z.X., Li, W., Zhang, W.F., Yang, Q., Hui, H., and Lai, C.K. (2019) Zircon water content: Reference material development and simultaneous measurement of oxygen isotopes by SIMS. Journal of Analytical Atomic Spectrometry, 34, 1088–1097.

Xu, B., Hou, Z.Q., Griffin, W.L., Lu, Y., Belousova, E., Xu, J.F., and O'Reilly, S.Y. (2021) Recycled volatiles determine fertility of porphyry deposits in collisional settings. American Mineralogist, 106, 656–661.

Yang, W.B., Niu, H.C., Shan, Q., Sun, W.D., Zhang, H., Li, N.B., Jiang, Y.H., and Yu, X.Y. (2014) Geochemistry of magmatic and hydrothermal zircon from the highly evolved Baerzhe alkaline granite: Implications for Zr-REE-Nb mineralization. Mineralium Deposita, 49, 451–470.

Yang, J.H., Zhou, M.F., Hu, R.Z., Zhong, H., Williams-Jones, A.E., Liu, L., Zhang, X.C., Fu, Y.Z., and Mao, W. (2020) Granite-Related Tin Metallogenic Events and Key Controlling Factors in Peninsular Malaysia, Southeast Asia: New

DOI: https://doi.org/10.2138/am-2022-8899. http://www.minsocam.org/

Insights from Cassiterite U-Pb Dating and Zircon Geochemistry. Economic Geology, 115, 581–601.

Yang, X., Wang, X., Li, B., Liu, Y., Liu, C., Liu, F., Huang, Z., and Hu, P. (2023) Discovery of tungsten orebody in the depth of the Yangla large Cu deposit, northwest Yunnan. Acta Mineralogical Sinica, 43, 93-96 (In Chinese with English abstract).

Zeng, L.J., Niu, H.C., Bao, Z.W., and Yang, W.B. (2017) Chemical lattice expansion of natural zircon during the magmatic-hydrothermal evolution of A-type granite. American Mineralogist, 102, 655–665.

Zhang, A.C., Wang, R.C., Hu, H., Zhang, H., Zhu, J.C., and Chen, X.M. (2004) Chemical evolution of Nb-Ta oxides and zircon from the Koktokay No. 3 granitic pegmatite, Altai, northwestern China. Mineralogical Magazine, 68, 739–756.

Zhong, S., Seltmann, R., Qu, H., and Song, Y. (2019) Characterization of the zircon Ce anomaly for estimation of oxidation state of magmas: a revised Ce/Ce* method. Mineralogy and Petrology, 113, 755–763.

Zhu, J.J., Hu, R., Richards, J.P., Bi, X., and Zhong, H. (2015) Genesis and magmatic-hydrothermal evolution of the Yangla Skarn Cu Deposit, Southwest China. Economic Geology, 110, 631–652.

Zou, S., Chen, X., Brzozowski, M.J., Leng, C., and Xu, D. (2022) Application of Machine Learning to Characterizing Magma Fertility in Porphyry Cu Deposits. Journal of Geophysical Research: Solid Earth, 127, 1–17.

FIGURE CAPTIONS

FIGURE 1. World map with sample positions labelled.

FIGURE 2. Zircon trace element scatter plot. (a-d) zircons collected from igneous rocks; (a) Y vs. U; (b) Th vs. U; (c) Nb vs. U; (d) U/Yb vs. Y; (e-h) zircons collected from ore deposits; (e) Eu/Eu* vs. Ce/Nd; (f) Eu/Eu* vs. Yb/Gd; (g) Eu/Eu*/Y×10000 vs. Ce/Nd/Y; (h) Nb vs. Ta.

FIGURE 3. Cartoon of the workflow. For data pre-processing, we perform missing value processing, data normalization and data balancing for magmatic rocks and deposits database. The purple boxes denote the optimal method. For the machine learning model, Random Forest works best, and the cartoon image is put here for easy understanding. Parameter tuning and 5-fold cross-validation were also done for the model.

FIGURE 4. Interval graphs for calculating the accuracy of conventional classification. (a) Y vs. U plot of igneous rocks; (b) The case of BR; (c) Eu/Eu* vs. Ce/Nd plot of ore deposits; (d) The case of porphyry-type Cu-Au-Mo deposit. Accuracy = $1-P(X)\times P(Y)$, P(X) and P(Y) means the ratio of overlapping data on the X- and Y-axis.

DOI: https://doi.org/10.2138/am-2022-8899. http://www.minsocam.org/

FIGURE 5. Confusion matrix plot for the testing set using Random Forest. (a) Data from Igneous Rocks Database; (b) Data from Ore Deposit Database. The data in the table represents the precision of prediction (**Eq. 7**). Each column in the matrix represents the predicted category, while each row represents the true category of the data. The sum of the scores in each column is 1.

FIGURE 6. Ranking of feature importance using Random Forest. (a) Based on Igneous Rocks Database; (b) Based on Ore Deposit Database.

FIGURE 7. Trend curve of F1-score with increasing number of features using Random Forest. Element symbols are listed in descending order of feature importance in the table. Colored element symbols indicate that they are decisive for classification and can be used to simplify the model.

FIGURE 8. Case study of igneous rocks in Yilgarn Craton, western Australia. (a) Geological map of Yilgarn Craton with sampling points; (b) The predicted results of zircon compositions on rock samples, the division of "granite" and "granodiorite" is based on the TAS rock classification proposed by Le Maitre (2002).

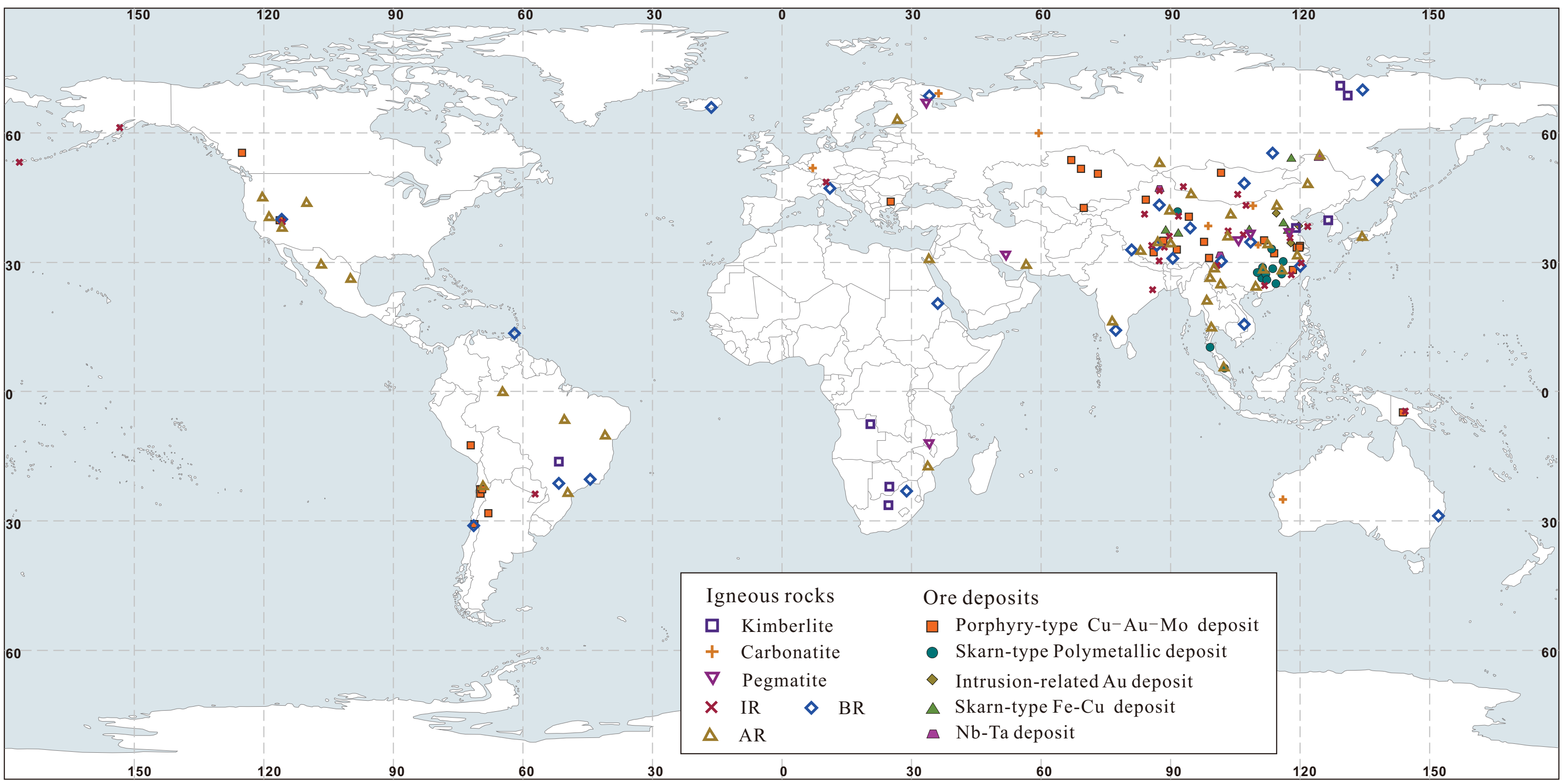
Figure 9. Case study of ore deposits in Sanjiang region, southwest China. (a) Geological map showing the location of Sanjiang region (Zhu et al. 2015); (b)

Tectonic framework of the Sanjiang region in southwest China showing the major terranes, suture zones, arc volcanic belts, and locations of the Yangla polymetallic skarn deposit, Pulang Cu porphyry deposit and Beiya Au-Cu porphyry deposit (Zhu et al. 2015); (c) Pie chart of the classification results of ore deposits based on zircon populations. Zircon samples 45-R1, 3250-41Lb1, 3250-41Lb1 were selected from quartz diorite at the Yangla deposit (Wang et al. 2022); sample PL01 and PL 02 were selected from a quartz diorite porphyry and sample PL03 was selected from a quartz monzonite porphyry at the Pulang deposit (Meng et al. 2018); sample BY01 and BY04 were selected from a quartz monzonite porphyry and BY02, BY03, and BY05 were selected from quartz syenite porphyry at the Beiya deposit (Meng et al. 2018).

APPENDIX AND WEB PORTAL

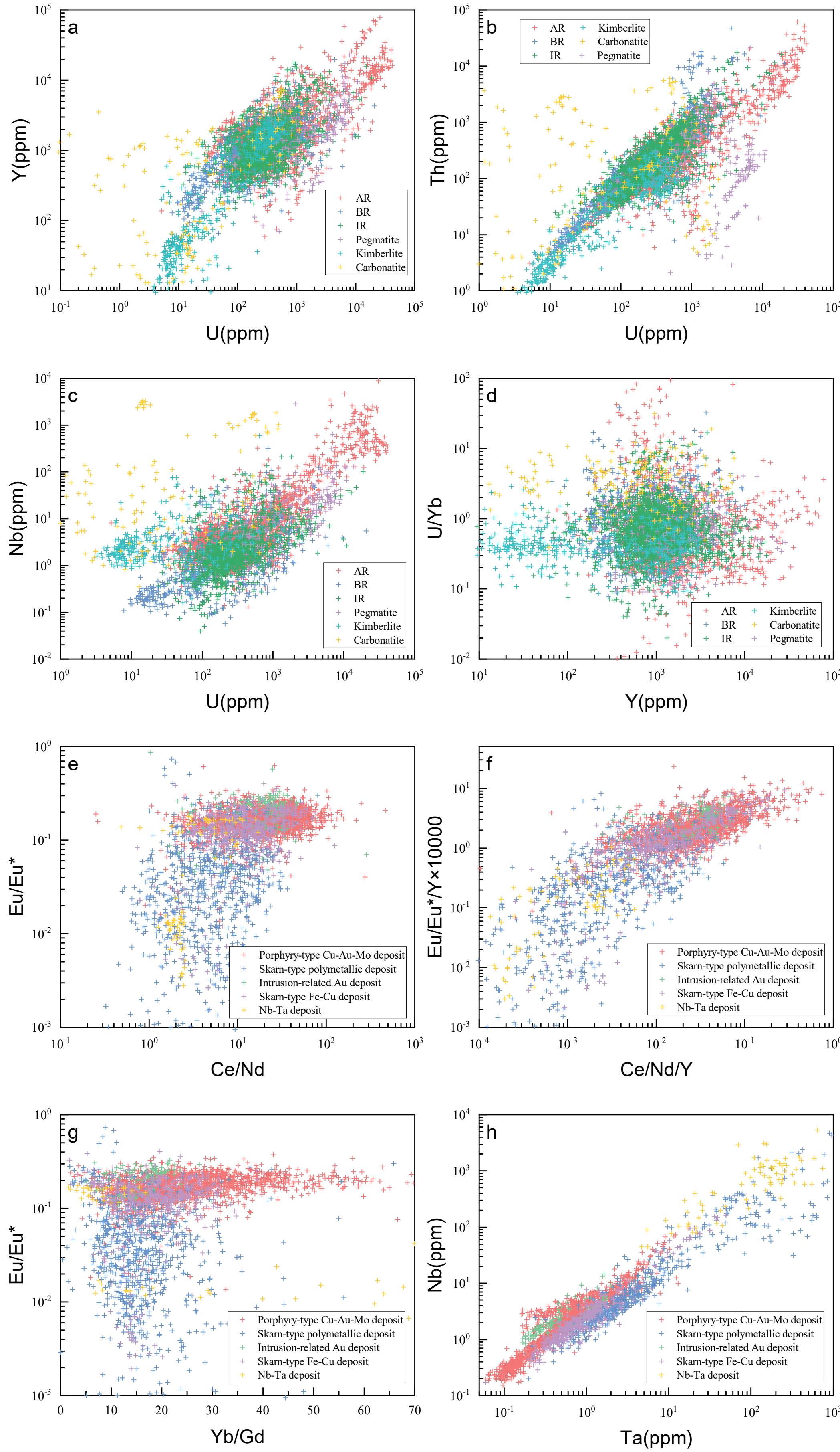
To train models and validate the applicability of models in machine learning, this study collated a large amount of data. These data are peer-reviewed and published and the program code for the zircon classification model is made publicly available at (https://github.com/ZihaoWen123/geology_class). Furthermore, to aid users we have developed a website front end for the zircon classification model which should facilitate ease of use (http://60.205.170.161:8001/).

Figure 1


150
120
90
60
30
0
30
60
90
120
150
60
30
0
30
60
Igneous rocks
Kimberlite
Carbonatite
Pegmatite
IR
BR
AR
Ore deposits
Porphyry-type Cu−Au−Mo deposit
Skarn-type Polymetallic deposit
Intrusion-related Au deposit
Skarn-type Fe-Cu deposit
Nb-Ta deposit

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-2022-8899. http://www.minsocam.org/

Figure 2



DOI: https://doi.org/10.2138/am-2022-8899. http://www.minsocam.org/

Figure 3

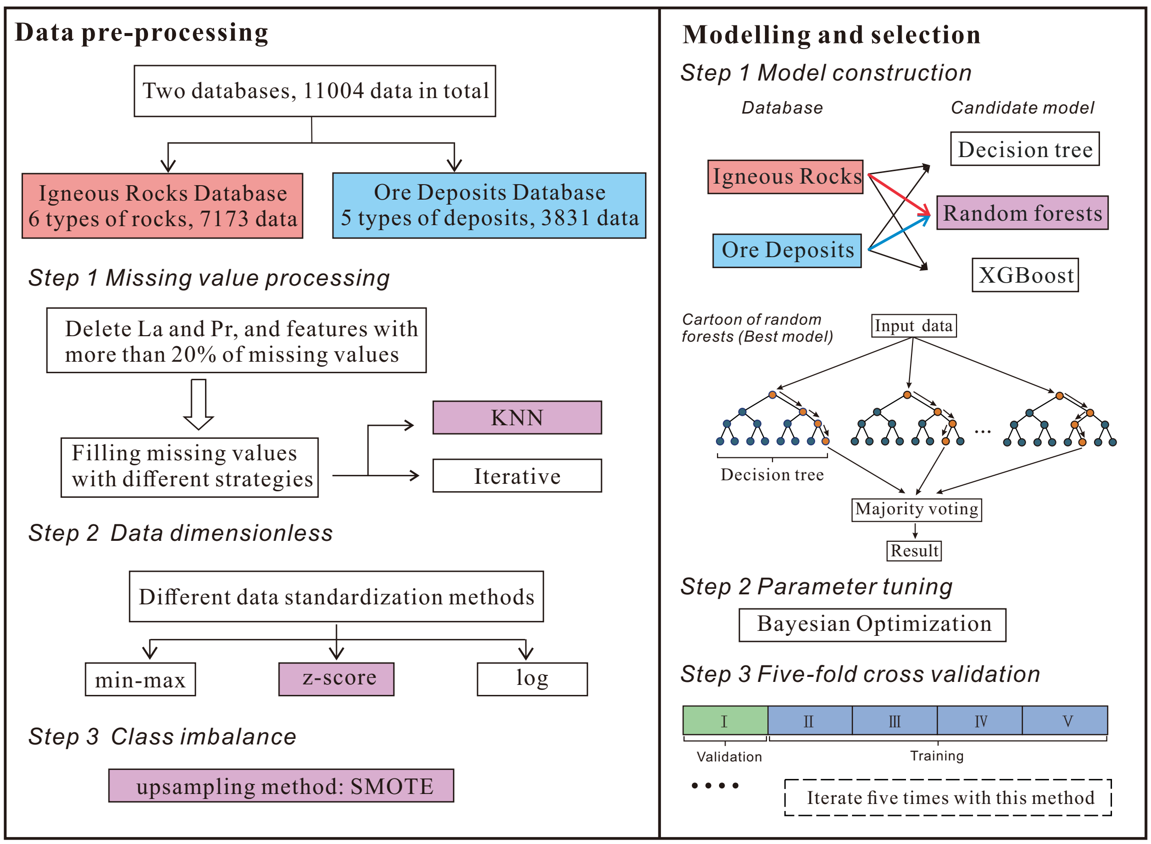
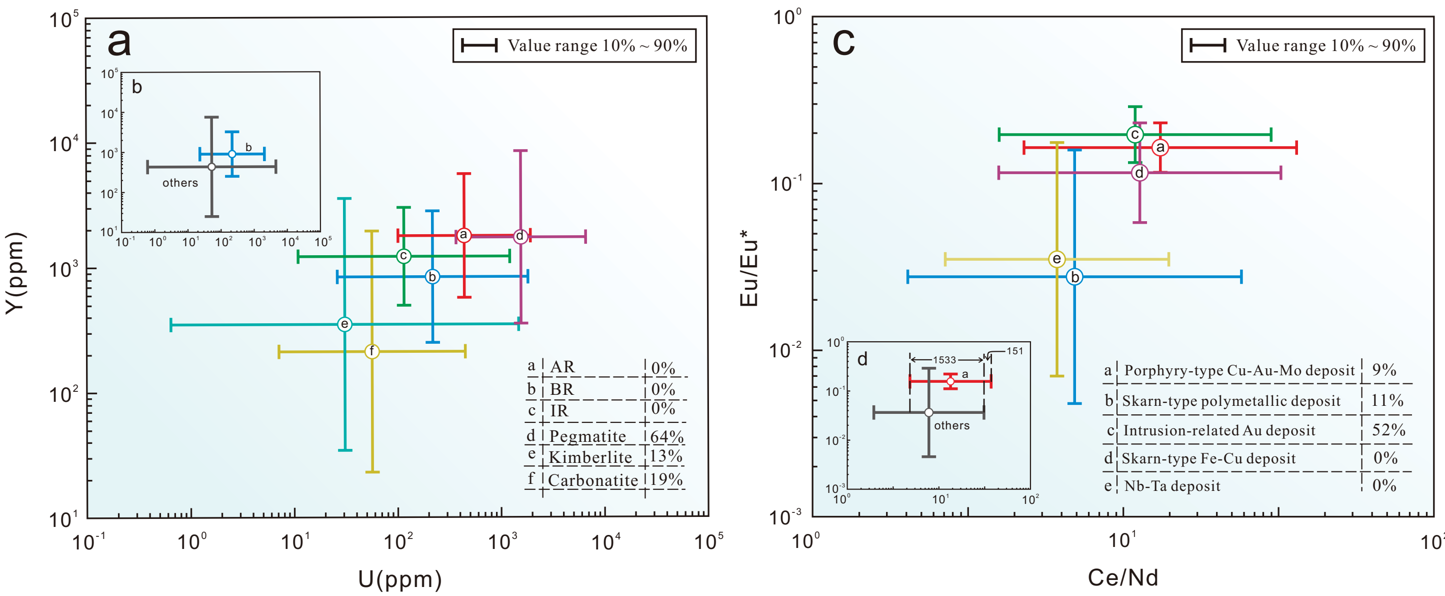


Figure 4


a
Value range 10% ~ 90%
b
others
Y(ppm)
U(ppm)
a AR 0%
b BR 0%
c IR 0%
d Pegmatite 64%
e Kimberlite 13%
f Carbonatite 19%
c
Value range 10% ~ 90%
d
1533
151
others
Eu/Eu*
Ce/Nd
a Porphyry-type Cu-Au-Mo deposit 9%
b Skarn-type polymetallic deposit 11%
c Intrusion-related Au deposit 52%
d Skarn-type Fe-Cu deposit 0%
e Nb-Ta deposit 0%

DOI: https://doi.org/10.2138/am-2022-8899. http://www.minsocam.org/

Figure 5

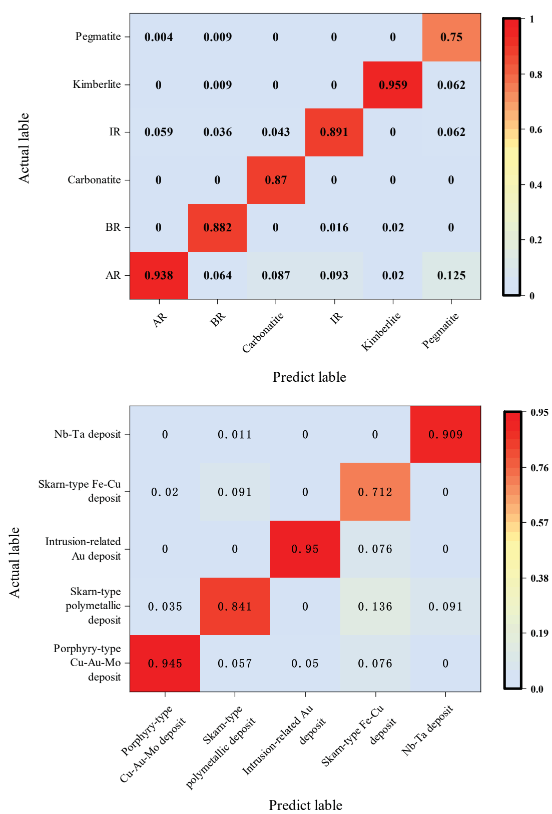


Figure 6

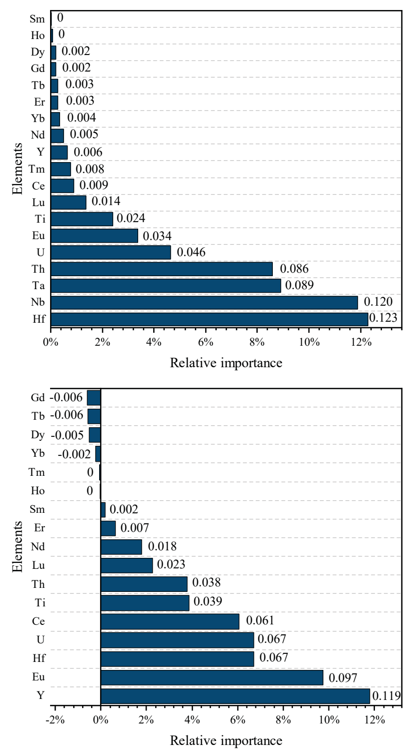

Sm
Ho
Dy
Gd
Tb
Er
Yb
Nd
Y
Tm
Ce
Lu
Ti
Eu
U
Th
Ta
Nb
Hf
0
0
0.002
0.002
0.003
0.003
0.004
0.005
0.006
0.008
0.009
0.014
0.024
0.034
0.046
0.086
0.089
0.120
0.123
Elements
0%
2%
4%
6%
8%
10%
12%
Relative importance
Gd
Tb
Dy
Yb
Tm
Ho
Sm
Er
Nd
Lu
Th
Ti
Ce
U
Hf
Eu
Y
-0.006
-0.006
-0.005
-0.002
0
0
0.002
0.007
0.018
0.023
0.038
0.039
0.061
0.067
0.067
0.097
0.119
Elements
-2%
0%
2%
4%
6%
8%
10%
12%
Relative importance

Figure 7

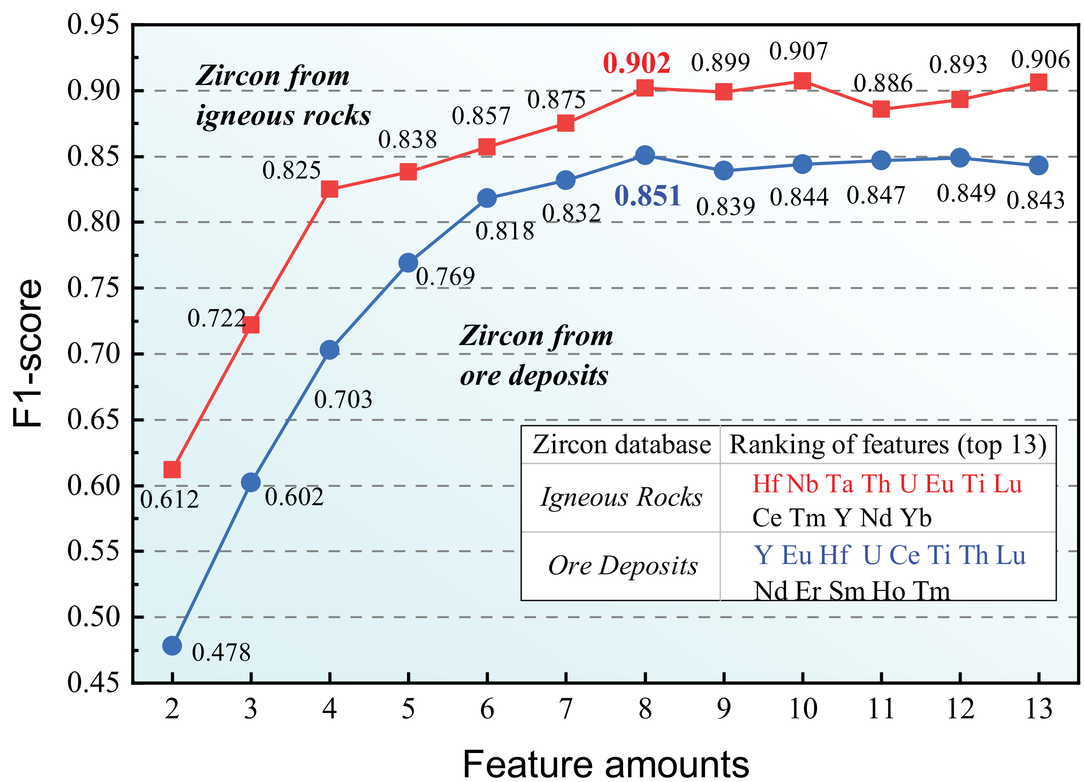
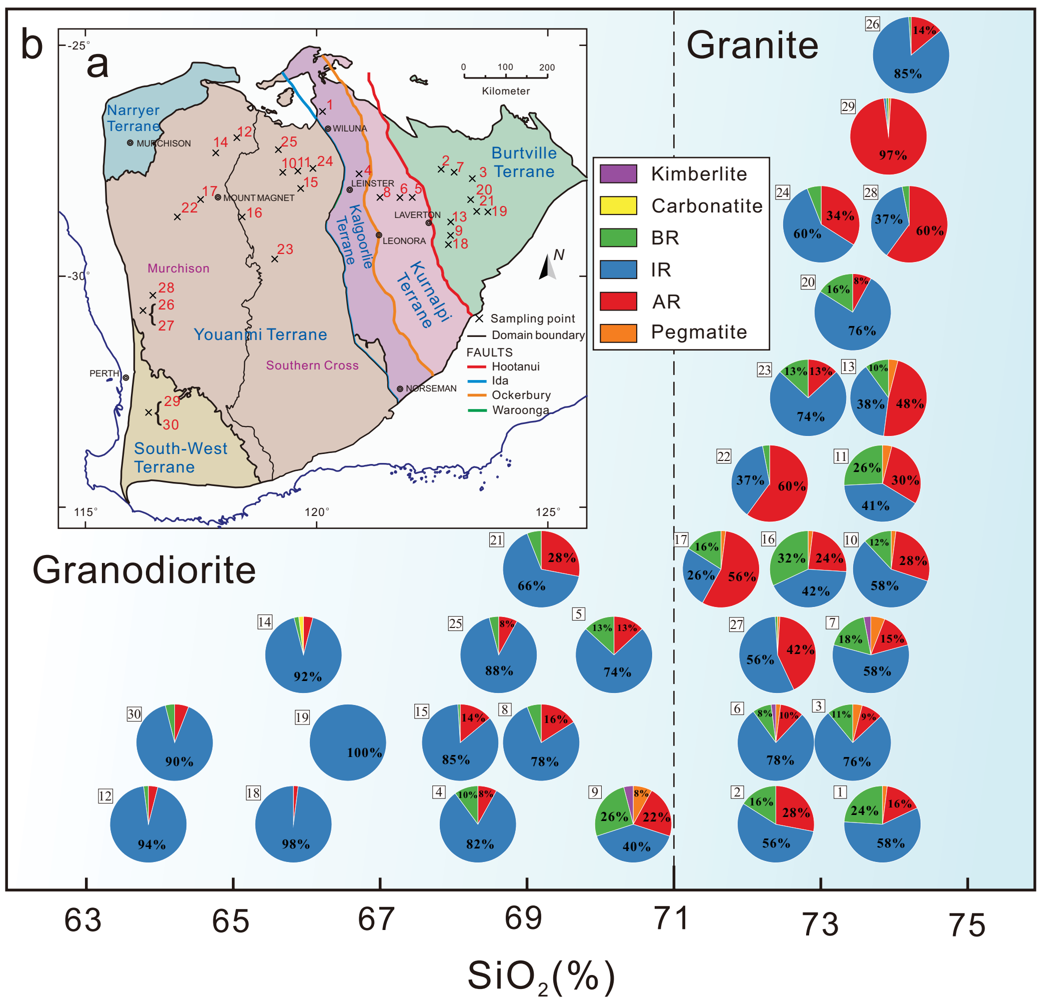

F1-score
Feature amounts
Zircon from igneous rocks
Zircon from ore deposits
0.612
0.722
0.825
0.838
0.857
0.875
0.902
0.899
0.907
0.886
0.893
0.906
0.478
0.602
0.703
0.769
0.818
0.832
0.851
0.839
0.844
0.847
0.849
0.843
Zircon database
Ranking of features (top 13)
Igneous Rocks
Hf Nb Ta Th U Eu Ti Lu Ce Tm Y Nd Yb
Ore Deposits
Y Eu Hf U Ce Ti Th Lu Nd Er Sm Ho Tm

Figure 8


b
a
Granite
Granodiorite
Narryer Terrane
Burtville Terrane
Youanmi Terrane
Kalgoorlie Terrane
Kurnalpi Terrane
South-West Terrane
Murchison
Southern Cross
MURCHISON
MOUNT MAGNET
WILUNA
LEINSTER
LAVERTON
LEONORA
NORSEMAN
PERTH
Kilometer
0 100 200
Sampling point
Domain boundary
FAULTS
Hootanui
Ida
Ockerbury
Waroonga
Kimberlite
Carbonatite
BR
IR
AR
Pegmatite
63
65
67
69
71
73
75
SiO2(%)

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-2022-8899. http://www.minsocam.org/

Figure 9

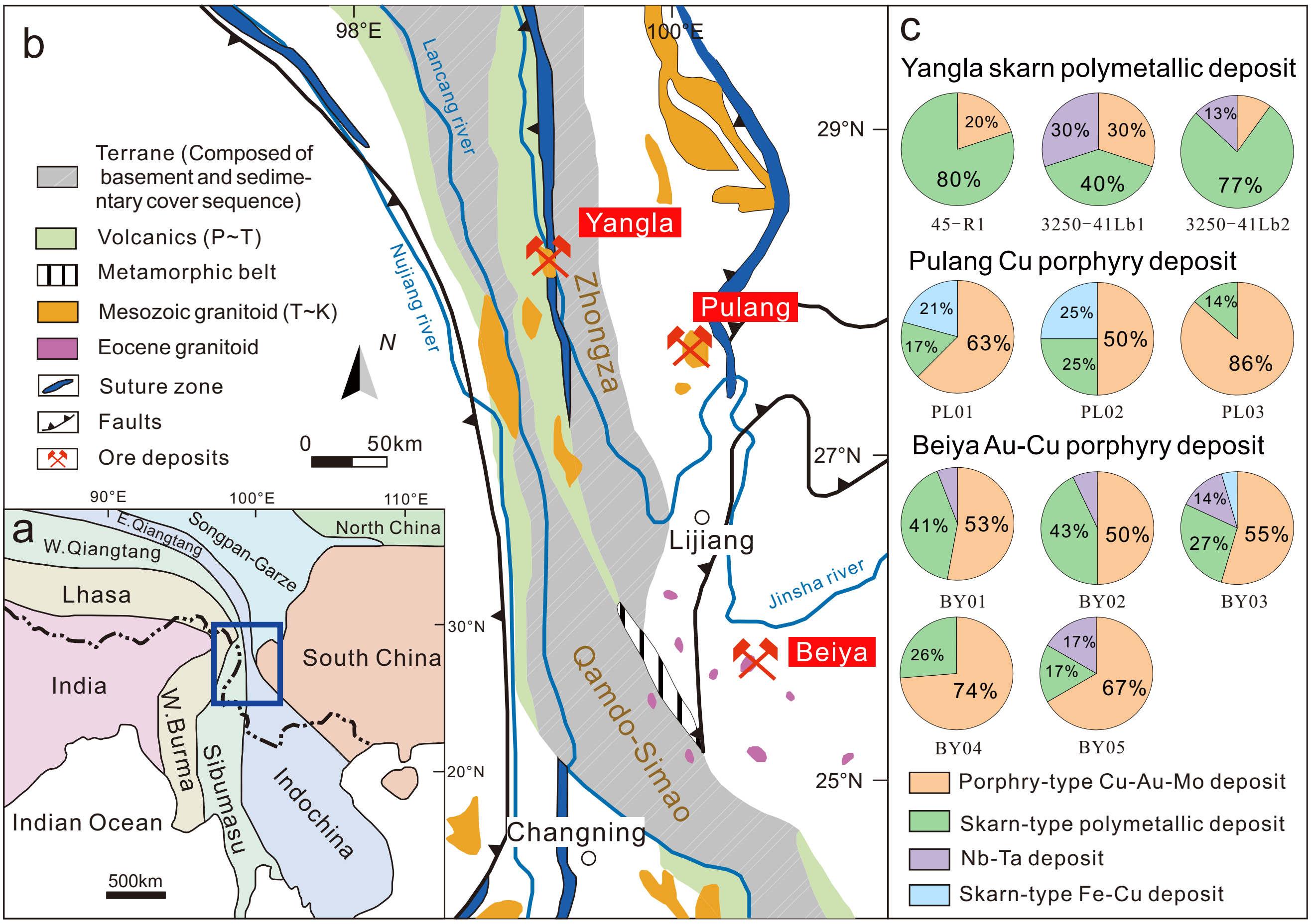


Table 1 Igneous rock and ore deposit type and data volume

Database types		Number of publications	Amount of data
Igneous rocks			
Kimberlite		10	549
Carbonatite		8	240
[a]BR	Gabbro Basalt	30	1058
[b]IR	Andesite Diorite	72	2514
[c]AR	Granite Rhyolite	66	2594
Pegmatite		8	218
Deposits			
Porphyry-type Cu-Mo-Au deposit		24	2122
Skarn-type Polymetallic deposit		13	896
Intrusion-related Au deposit		3	203
Skarn-type Fe-Cu deposit		8	529
Nb-Ta deposit		3	81

a. BR - basic rock, include gabbro and basalt;
b. IR - intermediate rock, include andesite and diorite;
c. AR – acid rock, include granite and rhyolite.

Table 2 Comparison of different data pre-processing strategies and machine learning algorit

Algorithm	Data pre-processing strategies			
	Missing values filling	Data standardization	Class imbalance	Accuracy
Igneous rocks classification model (19 features/elements)				
Decision Tree	knn-classification	z-score	smote	0.803
	knn-classification	log	smote	0.798
	knn-classification	minmax	smote	0.790
	iterative	z-score	smote	0.654
	iterative	minmax	smote	0.623
	iterative	log	smote	0.613
XGBoost	knn-classification	minmax	smote	0.870
	knn-classification	z-score	smote	0.868
	knn-classification	log	smote	0.854
	iterative	minmax	smote	0.682
	iterative	z-score	smote	0.689
	iterative	log	smote	0.742
Random Forest	knn-classification	minmax	smote	0.928
	knn-classification	log	smote	0.934
	knn-classification	z-score	smote	0.947
	iterative	z-score	smote	0.805
	iterative	log	smote	0.829
	iterative	minmax	smote	0.828
Ore deposits classification model (17 features/elements)				
Decision Tree	knn-classification	minmax	smote	0.749
	knn-classification	z-score	smote	0.742
	knn-classification	log	smote	0.725
	iterative	minmax	smote	0.533
	iterative	z-score	smote	0.575
	iterative	log	smote	0.591
XGBoost	knn-classification	minmax	smote	0.819
	knn-classification	log	smote	0.807
	knn-classification	z-score	smote	0.799
	iterative	minmax	smote	0.611
	iterative	z-score	smote	0.634
	iterative	log	smote	0.654
Random Forest	knn-classification	z-score	smote	0.856
	knn-classification	log	smote	0.838
	knn-classification	minmax	smote	0.807
	iterative	minmax	smote	0.722
	iterative	z-score	smote	0.704
	iterative	log	smote	0.733

thms

Performance	
F1-score	Recall
0.833	0.872
0.825	0.861
0.821	0.861
0.661	0.671
0.655	0.718
0.643	0.693
0.892	0.917
0.889	0.914
0.880	0.913
0.713	0.774
0.725	0.781
0.770	0.808
0.902	0.879
0.902	0.876
0.931	0.917
0.810	0.819
0.834	0.842
0.836	0.847
0.762	0.780
0.759	0.783
0.728	0.749
0.567	0.639
0.594	0.642
0.595	0.601
0.827	0.839
0.813	0.824
0.809	0.827
0.662	0.773
0.685	0.792
0.698	0.785
0.872	0.894
0.855	0.878
0.830	0.866
0.737	0.790
0.734	0.799
0.730	0.758

Table 3 Optimal parameter tuning results for Random Forests

Parameters	Igneous rocks classification model		Ore deposits classification model	
	19 features	8 features	17 features	8 features
max_depth	100	44	100	13
max_features	0.444	0.572	0.551	0.649
min_samples_leaf	1	1	1	1
min_samples_split	2	2	2	2
n_estimators	300	300	300	163
F1-score	0.963	0.914	0.890	0.877

Table 4 Results of Random Forest algorithm with five-fold cross validation

Database name	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold_avg
Igneous rocks	0.9369	0.9611	0.9501	0.9431	0.9427	0.9468
Ore deposits	0.9089	0.8988	0.9100	0.9187	0.8731	0.9019

Table 5 Case Study - Application of zircon classifier to igneous rocks in Yilgarn Craton

Rock number	Latitute and longitude	SiO_2 (%) content of rocks	results of whole-rock geochemitry	Predicti	
				Pegmatite	AR
1	-26.02°S, 120.32°E	73.7	granite	2%	16%
2	-27.86°S, 123.23°E	72.52	granite		28%
3	-27.76°S, 123.37°E	73.26	granite	4.3%	8.7%
4	-27.48°S, 121.02°E	68.46	granodiorite		8.3%
5	-27.89°S, 122.01°E	70.21	granodiorite		13%
6	-27.89°S, 121.88°E	72.43	granite	2%	10%
7	-27.35°S, 123.11°E	73.79	granite	5.9%	14.7%
8	-27.95°S, 121.37°E	69.18	granodiorite		16%
9	-28.77°S, 123.03°E	70.43	granodiorite	8%	22%
10	-27.53°S, 119.5°E	73.99	granite	2%	28%
11	-27.43°S, 119.6°E	73.74	granite	3.7%	29.6%
12	-26.75°S, 118.3°E	63.84	granodiorite		4%
13	-28.51°S, 123.02°E	73.8	granite	4%	48%
14	-27.41°S, 117.7°E	65.63	granodiorite		4%
15	-28.21°S, 119.86°E	67.87	granodiorite		14.1%
16	-28.44°S, 118.62°E	72.73	granite	2%	24%
17	-28.05°S, 117.73°E	71.6	granite	2%	56%
18	-29.02°S, 123.05°E	65.81	granodiorite		2%
19	-28.19°S, 123.67°E	66.54	granodiorite		
20	-27.99°S, 123.43°E	73.06	granite		8%
21	-28.19°S, 123.64°E	69.15	granodiorite		28%
22	-28.61°S, 116.85°E	72.3	granite		60.0%
23	-29.38°S, 119.17°E	72.21	granite		13.3%
24	-27.26°S, 119.96°E	72.95	granite		34%
25	-26.91°S, 119.27°E	68.45	granodiorite		8%
26	-31.03°S, 116.62°E	74.47	granite		14.1%
27	-31.03°S, 116.63°E	72.38	granite	1.40%	41.7%
28	-30.92°S, 116.65°E	74.02	granite		60.0%
29	-32.76°S, 116.38°E	73.89	granite	1.10%	96.7%
30	-32.76°S, 116.36°E	64.04	granodiorite		5.9%

Abbreviations: AR - acid rocks; IR - intermediate rocks; BR - basic rocks

on results of zircon composition				Data source
IR	BR	Carbonatite	Kimberlite	
58%	24%			Nelson DR (1998)
56%	16%			Wingate MTD, et al. (2011)
76.1%	10.9%			Wingate MTD, et al. (2011)
81.2%	10.4%			Nelson DR (1997)
73%	13.3%			Nelson DR (1997)
78%	8%		2%	Nelson DR (1997)
58.8%	17.6%		2.90%	Wingate MTD, et al. (2010)
78%	6%			Nelson DR (1997)
40%	26%		4%	Wingate MTD, et al. (2010)
58%	12%			Wingate MTD and Bodorkos S (2007)
40.7%	25.9%			Wingate MTD and Bodorkos S (2007)
94%	2%			Wingate MTD, et al. (2008)
38%	10%			Wingate MTD, et al. (2011)
92%	2%	2%		Wingate MTD, et al. (2011)
84.7%	1.20%			Wingate MTD, et al. (2012)
42%	32%			Wingate MTD, et al. (2015)
26%	16%			Wingate MTD, et al. (2014)
98%				Wingate MTD, et al. (2010)
100.0%				Wingate MTD, et al. (2011)
76%	16%			Wingate MTD, et al. (2009)
66%	6%			Wingate MTD, et al. (2011)
37%	2.90%			Wingate MTD, et al. (2015)
73.3%	13.30%			Nelson DR (2001)
60%	6%			Love GJ, et al. (2006)
88%	4%			Wingate MTD and Bodorkos S (2007)
84.7%	1.2%			Wingate MTD, et al. (2018)
55.6%	1.4%			Wingate MTD, et al. (2018)
37.1%	2.90%			Wingate MTD, et al. (2018)
1.1%	1.10%			McNaughton N. unpublished cited in Lu et al., (2019)
90.6%	3.5%			McNaughton N. unpublished cited in Lu et al., (2019)