



# Interpretability-Enhanced Mineral Prospectivity Models: A Synergistic Approach Using Large Language Models, Knowledge Graphs, and Machine Learning

Zhen-Jie Zhang<sup>1</sup> · Zi-Xing Yang<sup>1</sup> · Fu-Yuan Jian<sup>1</sup> · Si-Yuan Ban<sup>1</sup> ·  
Yi-Ming Wang<sup>1</sup>

Received: 11 November 2024 / Accepted: 17 August 2025  
© International Association for Mathematical Geosciences 2025

## Abstract

Developing interpretable predictive models for mineral prospectivity remains a persistent challenge in geoscience, as traditional data-driven methods often lack transparency, limiting geological understanding and validation of predictions. This study introduces an approach that integrates large language models (LLMs) with text data to construct a knowledge graph (KG), which is subsequently mined to identify and select predictive variables related to specific ore deposit types. Based on the extracted knowledge, an interpretable prediction framework (KG-SA-GC) is developed by combining S-A fractal filtering (SA) with the gcForest machine learning algorithm (GC). The framework is applied to the Nanling region of South China as a case study, successfully identifying key geochemical predictors associated with rare metal mineralization. The results demonstrate the potential of utilizing LLMs and KGs to build interpretable, accurate, and scalable models for mineral prospectivity, highlighting the broader applicability of artificial intelligence (AI)-driven methodologies in advancing mineral exploration across diverse geological settings.

**Keywords** Mineral prospectivity mapping · Large language models · Knowledge graph · Interpretable mineral predictive models · Machine learning

---

✉ Zhen-Jie Zhang  
zjzhang@cugb.edu.cn

<sup>1</sup> Frontiers Science Center for Deep-Time Digital Earth, State Key Lab of Geological Processes and Mineral Resources, School of Earth Sciences and Resources, China University of Geosciences, Beijing 100083, China

## 1 Introduction

Mineral deposits are fundamental to human civilization, serving as the cornerstone of social and economic development. As easily accessible surface deposits are progressively depleted, the discovery of new mineral resources has become an increasingly urgent challenge (Richards 2016). Mineral prospectivity mapping (MPM) has emerged as a vital tool to address this challenge by delineating areas with high mineral potential based on geological, geochemical, and geophysical evidence (Singer and Mosier 1981; Agterberg 1989; Bonham-Carter 1994; Harris et al. 2001; Carranza et al. 2008; Hronsky and Kreuzer 2019; Zuo 2020; Yousefi et al. 2021). MPM methodologies can generally be categorized into two primary types: (1) data-driven approaches, which statistically or algorithmically model spatial associations between known mineral deposits and predictor variables (e.g., Brown et al. 2000; Porwal et al. 2003; Zuo and Carranza 2011; Harris et al. 2015; Zhang et al. 2016; Chen et al. 2020; Li et al. 2020a; Mao et al. 2020; Wang et al. 2020b; Harris et al. 2022), and (2) knowledge-driven approaches, which utilize expert judgment to interpret and integrate geological information (Bonham-Carter 1994; Harris et al. 2015). The effectiveness of each approach depends on factors such as data availability, data quality, and the depth of prior geological knowledge (Montsion et al. 2019).

Driven by advances in big data and computational technologies, data-driven models have gained increasing attention. However, despite algorithmic innovations and access to large datasets, MPM applications remain largely confined to academic and governmental geoscience sectors, with limited adoption by the broader mineral exploration industry (Hronsky and Kreuzer 2019). One major barrier is the so-called black-box nature of many machine learning (ML) models, which complicates the interpretation of model outputs by geologists and engineers and undermines confidence in their practical deployment (Porwal et al. 2015; Hronsky and Kreuzer 2019; Zuo 2020). Addressing this interpretability gap has become a critical research frontier. Recent efforts have explored the integration of expert geological knowledge into ML workflows, thereby enhancing model transparency and trustworthiness (Zuo et al. 2023). Notably, knowledge graphs (KGs), structured semantic representations of geological knowledge, have emerged as a promising avenue for embedding domain expertise into predictive modeling processes (e.g., Yan et al. 2023; Wang et al. 2024b).

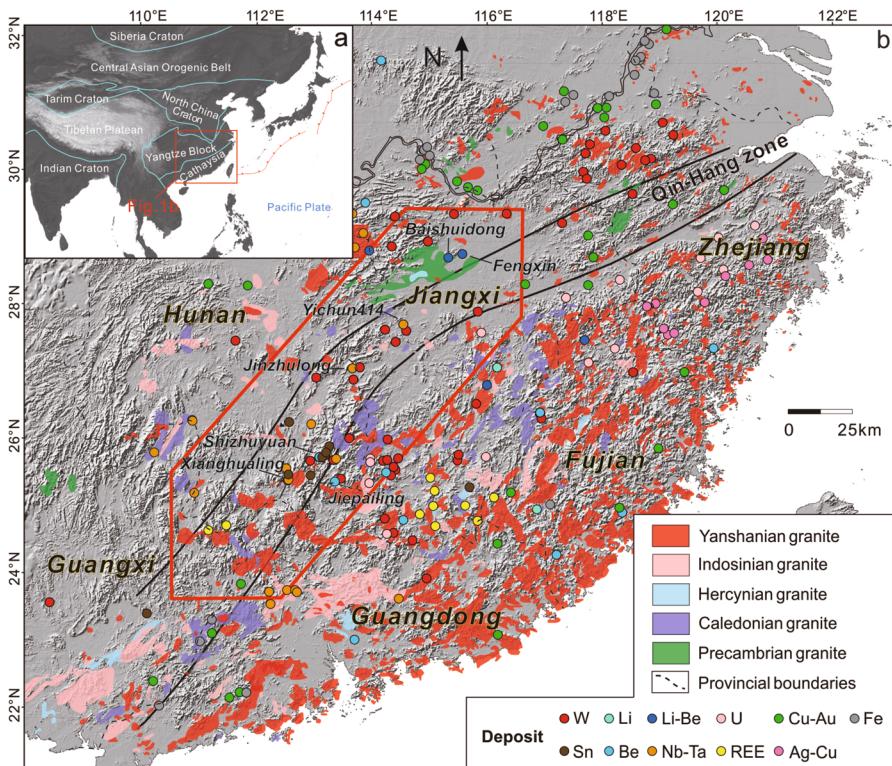
Despite these advances, significant challenges persist. Existing KG construction methods often rely heavily on manual encoding by domain experts, which is time-consuming, labor-intensive, and difficult to scale across diverse mineral systems. Moreover, integrating such KGs into ML frameworks in a way that systematically improves both interpretability and predictive performance remains nontrivial.

To address these limitations, this study proposes a novel interpretability-enhanced MPM framework that synergistically integrates large language models (LLMs), KGs, and ML techniques. First, LLMs are employed to automatically extract and organize expert knowledge from geological literature, thereby constructing a domain-specific KG tailored to the target mineralization system. This KG explicitly captures key geological features and their interrelationships, serving as a transparent knowledge base for guiding feature selection. Second, data mining techniques are applied to the KG to prioritize predictive features based on their inferred geological significance. Finally,

these KG-informed features are mapped onto geospatial datasets and used to train ML models for mineral prospectivity prediction. The proposed methodology is demonstrated through its application to the MPM of rare metal deposits in the Nanling region of South China, highlighting its potential to enhance both the interpretability and predictive accuracy of MPM workflows.

## 2 Geological Background

The Nanling region of South China is a major repository of rare and nonferrous metal resources, particularly rich in W, Sn, Sb, Bi, Pb, Zn, Li, Be, Nb, Ta, ion-adsorption rare earth elements (REEs), and U (Fig. 1). These mineral deposits predominantly formed during the Yanshanian period and are closely associated with the evolution of Yanshanian granitoids (Mao et al. 2011). Consequently, the Nanling region represents a key area for investigating continental mineralization systems and the petrogenesis of Yanshanian granitoids. Extensive studies have addressed the tectonic dynamics and metallogenic characteristics of critical mineral deposits within the region (e.g., Yuan et al. 2019; Su et al. 2020; Wang et al. 2020a; Yin et al. 2022).



**Fig. 1** Geological map of the Nanling region (modified after Zhang et al. 2023b)

Situated in central-southern South China, the Nanling region is bounded by the Pacific Plate to the east and the Indian Plate to the west, lying at the intersection of the circum-Pacific and Tethyan tectonic domains (Fig. 1a). Its geotectonic evolution has been influenced by multiple orogenic events, including the Jinning, Caledonian, Indosinian-Hercynian, Yanshanian, and Himalayan orogenies. A critical transition during the Mesozoic, from the Tethyan to the Paleo-Pacific tectonic regimes, significantly influenced both magmatism and mineralization. This transition resulted in the formation of NE-trending fault systems during the Yanshanian period, superimposed on NW- and EW-trending faults inherited from the Indosinian period (Wang et al. 2011; Xu 2023). Among these, NE-trending faults served as major conduits for magmatic activity and ore formation, controlling the emplacement of granitic intrusions and the development of W-Sn and other rare metal polymetallic deposits across the Nanling region and South China.

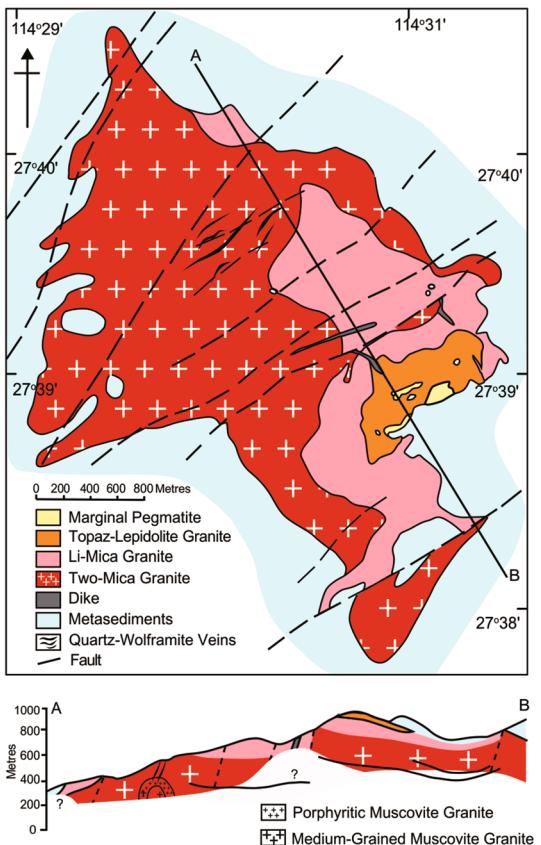
The Nanling region has experienced widespread magmatic activity, with numerous exposed and concealed granitic plutons. Deposits of Li, Be, Nb, and Ta are commonly spatially associated and often co-occur. Multiple tectonic events contributed to the genesis of Li-Be-Nb-Ta-bearing granites, most of which formed during the Yanshanian period (Wang et al. 2020a). These granites typically developed in relatively stable compressional or post-orogenic settings and are characterized by peraluminous to slightly metaluminous composition, high-potassium calc-alkaline affinities, and pronounced magmatic differentiation (Yin et al. 2022).

Various types of Li-Be-Nb-Ta deposits occur in the Nanling region, including granite-type, greisen-type, skarn-type, pegmatite-type, and quartz vein-type deposits (Zhao et al. 2021). These deposits formed through multiple episodes of tectonic—magmatic reworking and material cycling within mature continental crust and ancient basement terrains (Yu et al. 2023). A representative example is the Yichun super-large Nb-Ta-Li deposit (also known as the 414 deposit), hosted within anticlinal structures of fold belts (Fig. 2). Such deposits are mainly distributed in late-stage plutons of composite intrusions, altered granites, and pegmatite veins, often located near the upper margins of younger intrusive bodies (Fig. 2) (Lin et al. 1995; Yin et al. 2022). The migration and enrichment of ore-forming materials were primarily driven by crustal maturity, the high degree of granite evolution, and metasomatism by volatile-rich fluids (Zhao et al. 2021). Major Li-bearing minerals include spodumene, lepidolite, and ferro-lepidolite, whereas niobium and tantalum are commonly hosted in manganocolumbite, columbite, and tantalite, and beryllium predominantly in beryl, bertrandite, and helvite, particularly within greisen- and granite-type deposits.

### 3 Data and Methods

To enhance clarity and reproducibility, the overall workflow of this study is summarized in Fig. 3. The methodological framework integrates literature-based KG construction, S-A fractal filtering of geochemical anomalies, and interpretable ML using multiple algorithms. This flowchart illustrates the step-by-step process, from data

**Fig. 2** Geological map and cross section of the Yashan complex and Yichun deposit (after Zhang et al. 2023b)

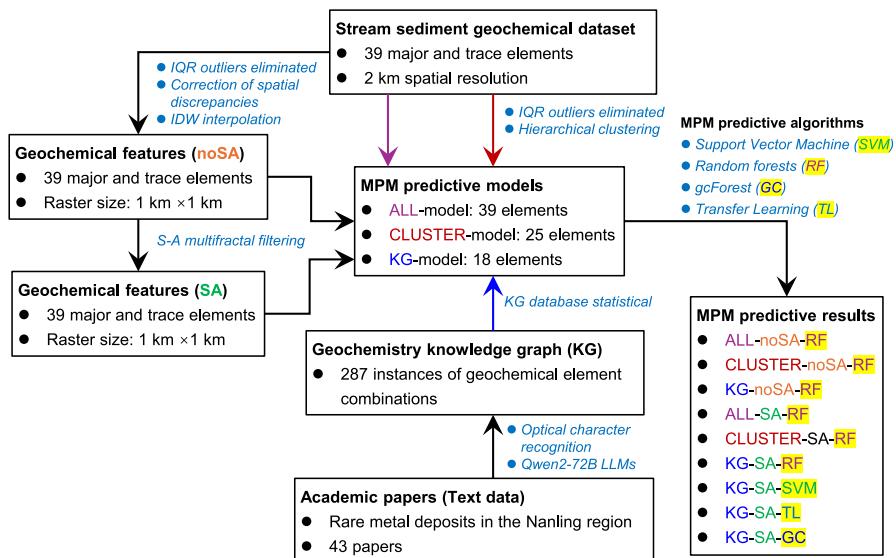


acquisition and knowledge extraction to model construction and mineral prospectivity prediction, thereby bridging the gap between methodological design and practical application.

### 3.1 Data

#### 3.1.1 Text Data

In contrast to the well-established metallogenic models for deposits such as porphyry copper (e.g., Richards 2003), rare metal deposits in the Nanling region lack unified and effective metallogenic and prospecting models. The complexity and diversity of elemental associations and mineralization styles in this region pose challenges to traditional statistical analyses of geological data (including geochemical, geophysical, and related data), which often fail to capture commonalities among different mineralization types. Nevertheless, valuable insights beyond conventional statistical methods can be obtained from textual data within the academic literature.



**Fig. 3** Workflow for building an interpretability-enhanced model for mineral prospectivity mapping

To enhance metallogenetic predictions, this study directly determines the degree of involvement of each element in the mineralization processes across various deposits based on the published literature, supplemented by statistical analysis. Key information regarding elemental associations relevant to metallogenetic prediction was extracted using the open-source LLM Qwen2-72B, with customized prompts designed to improve the quality and relevance of data extraction. The extracted information was subsequently organized and visualized in the form of a KG.

Academic publications provide an accurate reflection of evolving research trends and theoretical achievements. A total of 43 academic papers on rare metal deposits in the Nanling region, published between 2012 and 2024, were selected for analysis. These papers cover eight typical deposits or mineralization areas (Table 1 and S1). To avoid the overrepresentation of any single deposit, priority was given to the most recent publications. In cases where recent studies were limited, earlier key papers were selectively incorporated to ensure balanced representation and to prevent bias in the statistical outcomes. All selected papers were processed using optical character recognition (OCR), followed by manual verification to ensure accurate interpretation by the LLM.

### 3.1.2 Geochemical Data

The stream sediment geochemical datasets used for prospectivity mapping were collected at a scale of 1:200,000 under the framework of the Chinese National Geochemical Mapping (CNGM) project (Xie et al. 1997). Each geochemical datum represents a composite sample comprising four stream sediment subsamples collected

**Table 1** Deposits utilized for information extraction in this study

No.	Sub-area	Deposits
1	Yichun, Jiangxi Province	Yichun 414, Ta-Li
2	Dajishan, Jiangxi Province	Dajishan, W-polymetal
3	Xianghualing, Hunan Province	Xianghualing, Sn-polymetal
4	Jianfengling, Southern Hunan	Jianfengling, Ta-Nb-W-Sn
5	Shizhuyuan, Southern Hunan	Shizhuyuan, W-Sn
6	Jiepailing, Hunan Province	Jiepailing, Sn-polymetal
7	Liangyuan, Guangdong Province	Liangyuan, Nb-Ta-Rb-W
8	Limu, Guangxi Province	Limu, Sn-polymetal

\*Detailed references used for KG can be found in Appendix Table S1

within 1 km<sup>2</sup> grid cells, ensuring local representativeness. Sampling points are uniformly distributed with an approximate spatial resolution of 2 km across the study area. The analytical procedures followed two main protocols: 39 major and trace elements were determined using either X-ray fluorescence (XRF) or inductively coupled plasma–atomic emission spectrometry (ICP-AES), depending on the element and project phase. Quality control measures included duplicate analyses for approximately 5% of the samples, with analytical bias controlled within 5%. These procedures ensure both spatial consistency and analytical reliability.

For the North American region, the geochemical data utilized in the transfer learning model training consisted of stream sediment geochemistry at an approximate mapping scale of 1:200,000. These datasets included 20 elements (Al, Ba, Ca, Co, Cr, Cu, Fe, La, Mg, Mn, Na, Ni, Pb, Sc, Sr, Th, Ti, V, Y, Zn), totaling 56,207 samples. Data were primarily obtained from (1) the United States Geological Survey (USGS 2016), (2) the British Columbia Geological Survey (Han and Rukhlov 2020), and (3) the Yukon Geological Survey (Geomatics Yukon 2021). Sampling points were generally spaced at intervals of ~2 km, with each geochemical datum representing a composite of multiple subsamples to ensure local representativeness. Analytical procedures followed agency-specific protocols, typically involving XRF or ICP-AES. All datasets are publicly available via their respective geological survey repositories, and detailed spatial coverage is provided in Dong and Zhang (2024).

To ensure data quality and consistency, the following preprocessing steps were applied:

1. Duplicate and outlier removal: Duplicate values and zero values for individual elements were eliminated. Outliers were identified and removed using the interquartile range (IQR) method, where the IQR was calculated by subtracting the first quartile (Q1) from the third quartile (Q3). Values outside the range of 1.5 times the IQR below Q1 or above Q3 were classified as outliers.
2. Correction of spatial discrepancies: The study area encompasses 40 spatially adjacent geochemical survey maps, resulting in occasional “step-like” discrepancies in element concentrations between neighboring survey images. To address these

discontinuities, a standard normal transformation method, a linear leveling technique based on data standardization, was employed. The transformation formula is expressed as

$$x_{it} = \overline{U_i} \times S_0 + x_0 = \frac{x_i - \bar{x}_i}{S_i} \times S_0 + \bar{x}_0,$$

where  $x_{it}$  represents the transformed sample content,  $x_i$  denotes the original content,  $\bar{x}_i$  and  $S_i$  are the mean and standard deviation of the original sample data, respectively, and  $\bar{x}_0$  and  $S_0$  refer to the corresponding parameters of the standard sample.

A visual inspection of the leveled maps was performed to confirm that spatial consistency and continuity between adjacent survey sheets were effectively maintained after normalization.

3. Interpolation to raster format: After leveling, each element dataset was interpolated into raster images using the inverse distance weighting (IDW) interpolation method at a spatial resolution of 1 km<sup>2</sup>. These high-resolution geochemical maps provide a critical foundation for detailed prospectivity analysis.

### 3.1.3 Label Data

The positive samples used for model training were derived from the National Mineral Deposit Database of China (<http://ngac.org.cn>), supplemented by coordinates extracted from published literature and exploration reports. All coordinates were carefully verified and corrected using high-resolution satellite imagery available in Google Earth to ensure spatial accuracy. A total of 47 mineralized sites were selected, each exhibiting evidence of mineralization in at least one of the following rare metals: Li, Be, Nb, or Ta.

A total of 94 negative samples were selected to represent non-mineralized locations. Specifically, regions within 2 km of major faults and igneous rock bodies—areas considered to have higher mineralization potential—were excluded from the candidate pool. Random sampling was then applied to the remaining area to generate spatially independent negative samples, ensuring that the classification task was both geologically meaningful and spatially unbiased.

## 3.2 Methods

### 3.2.1 Qwen LLM

The development of LLMs marks a significant leap in natural language processing (NLP), transforming methodologies applied to various NLP tasks over recent years (Raiaan et al. 2024; Yang et al. 2024b). Although a standardized definition for LLMs is lacking, the term generally refers to models with vast parameter scales—typically measured in billions—that exhibit capabilities comparable to human linguistic proficiency and intelligence (Raiaan et al. 2024). Prominent examples include OpenAI's

GPT series, Alibaba Cloud's Qwen series, Anthropic's Claude series, Tsinghua University's ChatGLM, and Tencent's Hunyuan.

LLMs are trained on expansive datasets, often encompassing terabytes of data, enabling them to learn the structures of human language while amassing a broad base of general knowledge. Through instruction tuning, LLMs acquire the ability to comprehend and execute complex tasks based on human commands. Their capacity to process and interpret academic literature with near-human understanding makes them particularly valuable in domains such as geological research, where they assist in repetitive and labor-intensive tasks traditionally reliant on expert intervention.

Among open-source LLMs, the Qwen series, developed by Alibaba Cloud, has demonstrated outstanding performance in both Chinese and English language tasks (Bai et al. 2023). With parameter sizes ranging from 1.5 billion to 110 billion, the first generation of the Qwen model was released in 2023, followed by the launch of the Qwen2 series in 2024. Built on the Transformer architecture, Qwen2 was pretrained on a vast corpus of web content and literary works, amounting to approximately 3 trillion tokens (Yang et al. 2024a).

To qualitatively evaluate the geological knowledge embedded in these models, several LLMs were prompted with the same domain-specific question: "Could you define the Nanling granite-type lithium deposit in the most professional way possible?" (Fig. 4). Qwen2-72B responded with a detailed explanation encompassing the tectonic background (Late Mesozoic to Early Cenozoic granites in the Nanling region), magmatic evolution (deep crustal melting and emplacement of peralkaline granites),



You are a geology expert. Could you define the Nanling granite-type lithium deposit in the most professional way possible?

**Qwen:** The Nanling granite-type lithium deposit refers to lithium ore bodies located in the Nanling Mountain region of southern China, associated with specific types of granite intrusions. These deposits mainly formed during the late Mesozoic to early Cenozoic, under regional tectonic settings. They resulted from a series of geological processes such as deep crustal material ascent, melting, and emplacement. Specifically, Nanling granite-type lithium deposits are usually related to alkali-rich peralkaline granites (e.g., rare metal peralkaline granites), which contain high concentrations of alkali metals (potassium, sodium), rare earth elements, and other trace elements, including lithium. The lithium mineralization typically occurs near the contact zones between these granite bodies and their surrounding rocks, with various forms of mineralization, such as lithium-bearing minerals like lepidolite and spodumene.



**ERNIE Bot:** The Nanling granite-type lithium deposit refers to lithium ore bodies formed by lithium-rich granitic magma in the Nanling orogenic belt and surrounding regions in southern China. These lithium deposits are primarily hosted within granite bodies, especially those that have undergone significant greisenization alteration. Lithium is mainly present in minerals like lepidolite (e.g., fine-grained and microcrystalline lepidolite) and spodumene, sometimes accompanied by other rare metals like rubidium, cesium, and beryllium.



**ChatGLM:** Nanling granite-type lithium deposits are formed by lithium-rich granitic magma that has undergone a high degree of fractional crystallization under specific geological conditions. During this process, rare metals like lithium concentrate in the magma chamber and eventually form economically valuable lithium ore bodies in specific parts of the granite. These ore bodies are typically found in granites rich in lithium-bearing minerals such as lepidolite.



**Fig. 4** Qwen, ERNIE Bot, and ChatGLM's answer to Nanling granite-type lithium deposits. Translated from Chinese, because the professional comprehensiveness of the English answers of the three models is weaker than that of the Chinese answers

and mineral associations (e.g., lithium-bearing minerals such as lepidolite and spodumene). ERNIE Bot produced a technically correct and concise response, accurately referencing greisenization alteration and relevant trace metals (e.g., rubidium, cesium, beryllium). In contrast, ChatGLM provided a more generalized summary, emphasizing fractional crystallization without specifying geological context or mineralogical details.

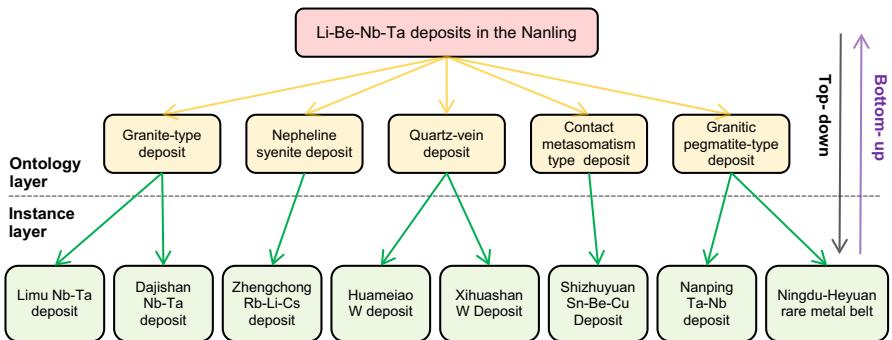
This comparison demonstrates that Qwen2 not only exhibits a higher degree of geoscientific fluency, but also delivers more comprehensive, domain-relevant outputs. These differences likely stem from the inclusion of geoscientific literature and textbooks in its pretraining corpus. Notably, the Qwen2 model used in this study was the raw base version, without fine-tuning or external tool integration. Despite this, it performed competently across multiple geology-related tasks. Furthermore, Qwen2 supports model fine-tuning and plug-in integration (Yang et al. 2024a), offering potential for future domain-specific adaptation in geoscientific applications.

### 3.2.2 KGs

The concept of KG was first introduced by Google in 2012 to enhance search engine optimization and has since become a pivotal tool for representing, managing, and analyzing knowledge across diverse disciplines (Ma 2022). Fundamentally, a KG is a graph database that stores relationships between entities, typically represented as triples consisting of a head entity, a tail entity, and the relationship linking them (Zhou et al. 2021a; Ma 2022). These triples mirror the subject-object-relationship structure in natural language, allowing objective facts to be systematically represented. KGs have become widely used in fields such as medicine, evident in the Unified Medical Language System (UMLS) (Bodenreider 2004), and in general knowledge management through projects like DBpedia (Lehmann et al. 2015). In the geosciences, KGs have been applied to systems such as porphyry copper deposit databases (Zhou et al. 2021b; Wang et al. 2024a).

In the era of exponential data growth, geoscientists face challenges in managing and extracting meaningful insights from vast datasets. KGs offer an effective solution, organizing geological knowledge into two primary layers: the ontology (or schema) layer and the instance layer (Fig. 5) (Ma 2022). The ontology layer defines abstract geological concepts, specifies their relationships, and outlines the attributes of corresponding instances. Ontology design is a crucial stage that often determines the ultimate capabilities and applications of the KG. The instance layer populates these ontologies with specific examples (instances). For instance, the Dajishan Nb-Ta deposit would be categorized as an instance under the ontology for granite-type rare metal deposits (Fig. 5).

KG construction typically follows two primary strategies: the top-down and the bottom-up approaches (Ma 2022). The top-down approach involves designing the ontology layer first, followed by extracting instances based on the predefined schema (Fig. 5). In contrast, the bottom-up approach begins with the extraction of instances, from which the ontological structure is subsequently inferred and constructed. Each methodology presents unique advantages depending on the research objectives and data characteristics.



**Fig. 5** Construction example of ontology layer and instance layer of geological KG

### 3.2.3 Hierarchical Clustering

Hierarchical clustering groups data points based on a hierarchical structure by progressively merging or splitting clusters at each level (Murtagh and Contreras 2012). This method creates a nested sequence of clusters, enabling analysis at different levels of granularity. Two primary types of hierarchical clustering exist: agglomerative and divisive (Fig. 6a). Agglomerative clustering, also referred to as bottom-up clustering, begins with each data point as an individual cluster and iteratively merges the closest pairs of clusters until a single cluster encompassing all data points is formed. In contrast, divisive clustering, or top-down clustering, starts with the entire dataset as one cluster and recursively splits it into smaller clusters until each data point constitutes its own cluster.

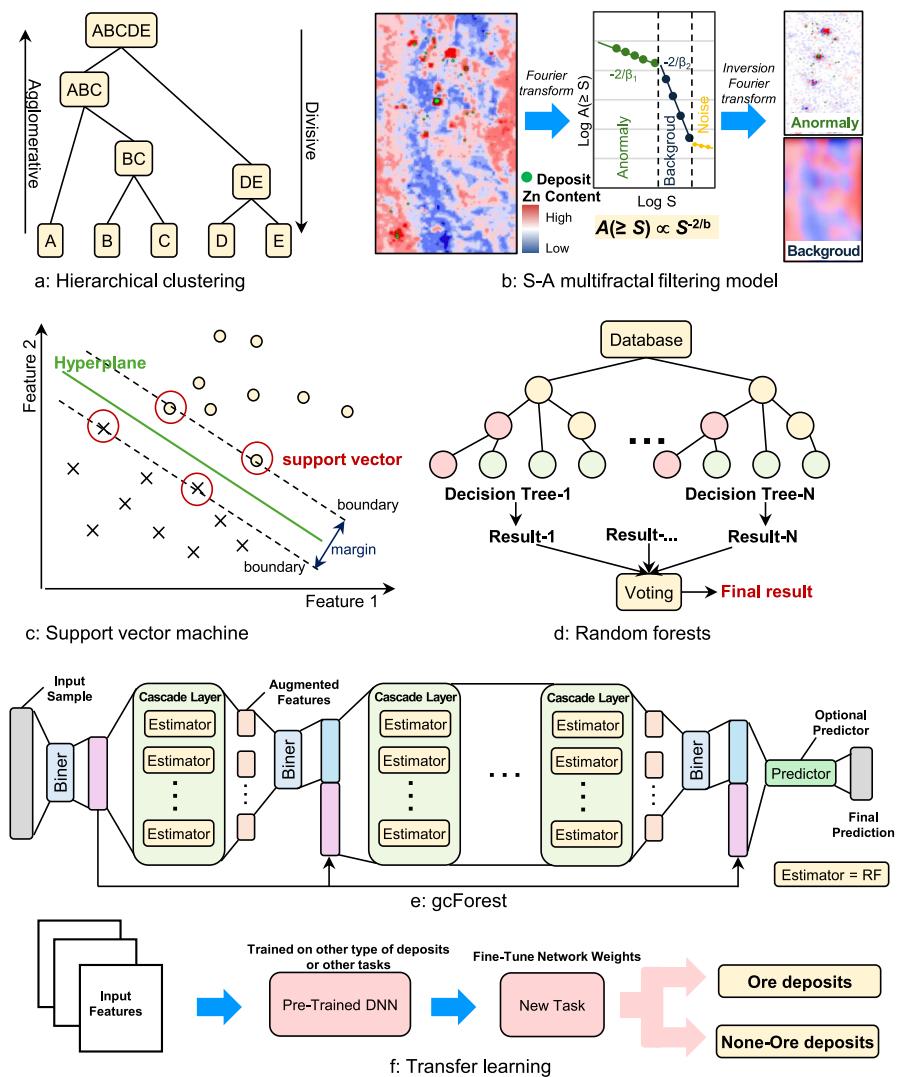
### 3.2.4 S-A Multifractal Filtering Model

The spectrum-area (S-A) multifractal filtering technique (Cheng et al. 2000) employs the Fourier transform to convert geochemical distribution maps from the spatial domain into the frequency domain, where image attributes are represented by energy spectral density. The energy spectral density distribution, exhibiting varying generalized self-similarity features, is decomposed through the application of different filters. This procedure enables the separation and preservation of the energy spectral components corresponding to anomalous fields and background fields. Subsequently, the inverse Fourier transform is applied to revert the data from the frequency domain back to the spatial domain, facilitating the differentiation between anomalous fields and background fields within the geochemical distribution map (Fig. 6b).

In this approach, self-similarity features in the frequency domain differ between anomalies and background, and the relationship can be described by the expression

$$A(\geq S) \propto S^{-\beta},$$

where  $S$  represents the spectral density,  $A(\geq S)$  denotes the area in the spectral density space where the spectral density exceeds  $S$ , and  $\beta$  is the fractal model's exponent.



**Fig. 6** Data processing and machine learning methods used in this study. **a** Hierarchical clustering; **b** S-A multifractal filtering model; **c** SVM; **d** random forests; **e** gcForest; **f** transfer learning

In the application of the S-A method for separating geochemical anomalies, carefully defined filters enable the extraction of anomaly and background features with varying intensities, characteristics, and spatial scales.

### 3.2.5 Machine Learning

#### (1) Support vector machine

The support vector machine (SVM) is a widely used supervised learning method, extensively applied to both classification and regression tasks. By utilizing data vectors with known labels, SVM constructs a linear hyperplane to separate different classes (Hearst et al. 1998). For nonlinear problems, SVM transforms the data into a higher-dimensional space to identify the optimal separating hyperplane (Fig. 6c). Once trained, the SVM can classify new, unlabeled data based on their features. Kernel functions play a crucial role in this process by mapping input features into a high-dimensional space, making the problem linearly separable. Commonly used kernel functions include the polynomial kernel, the radial basis function (RBF) kernel, and the sigmoid kernel (Zuo and Carranza 2011). Among these, the RBF kernel is particularly favored due to its efficient performance and minimal parameter requirements. In the present study, the RBF kernel was selected to leverage these advantages.

#### (2) Random forests

Random forests (RF) (Breiman 2001) are a widely adopted ensemble learning technique for MPM (e.g., Carranza and Laborde 2015; Zhang et al. 2016) consisting of multiple decision trees (DTs) applied to both classification and regression. Each DT is constructed from randomly sampled data vectors, with all trees generated independently from the same distribution (Fig. 6d). The RF model employs bootstrap aggregation (bagging) to create diverse training sets, typically using approximately two-thirds of the available samples for training and reserving the remaining one-third for validation. As the number of trees increases, the generalization error tends to stabilize, thus preventing overfitting. The Gini impurity index is utilized to determine the optimal splitting attribute at each node, facilitating the construction of complete decision trees. The final prediction of the RF model is obtained by averaging the predictions from all individual trees.

#### (3) gcForest

The gcForest (GC) technique is an optimization of the deep forest framework, which operates using a multilayer structure composed of non-differentiable decision trees (Zhou and Feng 2019). The version of deep forest applied in this study represents an enhanced form of GC, maintaining a cascade structure while achieving representation learning through forests (Fig. 6e). To better capture representative features from the data, deep forest introduces a multi-granularity scanning mechanism, further optimized by incorporating a binner layer. This layer reduces the number of splitting candidates when constructing decision trees, thereby enhancing memory efficiency and reducing computational costs associated with multi-granularity scanning. Compared to traditional deep learning methods, deep forest requires fewer hyperparameters and can automatically adjust its model complexity based on the dataset characteristics, making it a highly adaptable and efficient alternative for various learning tasks (Dong and Zhang 2024).

#### (4) Transfer learning

The framework of transfer learning involves a source domain, a target domain, and their respective task models (Weiss et al. 2016). The objective is to leverage knowledge and experience acquired from the source domain to enhance learning performance in the target domain, thereby minimizing the required number of samples, training time, and computational resources for the target task. The core principle of this approach is the transfer of knowledge and experience between tasks to improve the learning effectiveness of the target domain (Fig. 6f).

Specifically, a feature-based deep neural network (DNN) transfer learning strategy was applied using a fine-tuning approach. The source domain consisted of feature datasets derived from porphyry copper systems in the North American Cordillera, where mineralization is genetically linked to magmatic-hydrothermal processes associated with Pacific plate subduction—similar to the tectonic setting of rare metal deposits in South China. The target domain was the South China metallagenic belt, where model performance was to be optimized.

The pretraining phase involved training a DNN model on the North American dataset. The model architecture consisted of a six-layer fully connected neural network with layer sizes of 32–128–256–64–32–1. Each hidden layer used the rectified linear unit (ReLU) activation function and included dropout layers ( $p = 0.1$ ) to reduce overfitting. The output layer employed a sigmoid activation function for binary classification. During the fine-tuning phase, the pretrained model parameters were used as initialization for the target task, and all layers were updated using the South China dataset. Input features were standardized using StandardScaler.

The model was trained using the Adam optimizer (learning rate = 0.001) and binary cross-entropy loss. A ReduceLROnPlateau scheduler was applied (factor = 0.1, patience = 12) to adapt the learning rate during training. Early stopping (patience = 50) was used to prevent overfitting. The batch size was 64, and the maximum number of training epochs was set to 200. The data were split into training and validation sets at a ratio of 8:2. This transfer learning strategy enabled efficient reuse of mineralization-related knowledge from the source domain and enhanced the model's predictive capability in geologically similar but data-scarce target regions.

## (5) Model evaluation metrics

To quantitatively assess the classification performance of the ML models, four widely used evaluation metrics were employed: accuracy, precision, recall, and the F1 score. These metrics provide complementary perspectives on model effectiveness, particularly under conditions of class imbalance, which are common in MPM tasks.

Accuracy refers to the proportion of correctly classified instances among all predictions, offering a general indication of overall performance. Precision measures the proportion of true positive predictions among all instances predicted as positive, thereby reflecting the reliability of positive predictions. Recall (also known as sensitivity) quantifies the proportion of true positive instances correctly identified by the model, highlighting its ability to detect mineralized samples. The F1 score is the harmonic mean of precision and recall, providing a balanced evaluation metric that is particularly informative when class distributions are skewed. For each model, an 80/20 split was applied at the sample level, ensuring that the training and testing datasets

were disjoint in both space and class label. To reduce overfitting and improve robustness, the experiments were carried out 10 times with different random seeds, and the average performance was reported.

#### (6) Parameter tuning

To enhance model performance and ensure fair comparisons among different ML approaches, parameter optimization was conducted using a grid search strategy. This method systematically evaluates combinations of key parameters to identify those that yield optimal classification results on the validation set. For each model, the search space was defined based on prior knowledge and empirical ranges commonly used in mineral prospectivity studies.

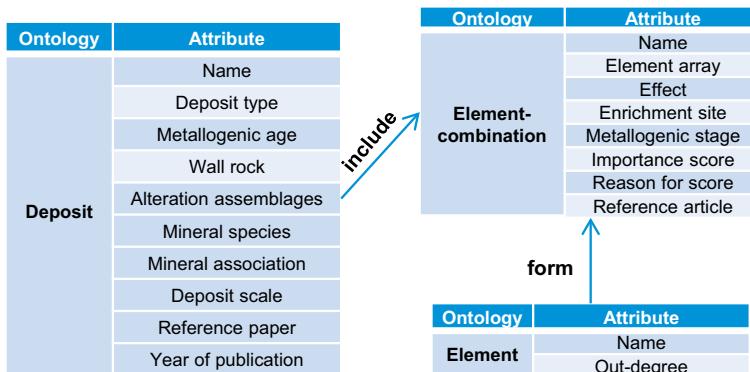
In the case of RF, the primary parameters tuned include the number of estimators (`n_estimators`), maximum tree depth (`max_depth`), and the minimum number of samples required to split an internal node (`min_samples_split`) or to be at a leaf node (`min_samples_leaf`). For the SVM, parameters including the regularization coefficient (`C`), kernel type, and kernel-specific settings (e.g., `gamma`) were optimized. In transfer learning (TL) models, the structure of the fully connected layers was predefined but adjusted based on input dimensionality and learning efficiency. The `gcForest` model was optimized by tuning the number and depth of trees in each cascade layer.

## 4 Geochemistry KG for Nanling Rare Metal Deposits

### 4.1 Ontology Design

As an effective tool for knowledge management, a KG enables the organization and storage of diverse information within a graph database. When combined with appropriate information extraction tools and advanced graph algorithms, the Geochemistry KG provides a novel framework for creating structured, domain-specific representations among geochemical elements. A top-down approach was employed to construct the Geochemistry KG for the Nanling rare metal deposits. This approach begins with the definition of the graph's ontology structure (Fig. 7), proceeds with the adjustments to the prompt terms according to this structure, and concludes with the extraction of the specified information.

To ensure the accuracy of the LLM's extraction results, minimize ambiguities arising from multiple extractions, and streamline subsequent alignment tasks, a concise and targeted schema was developed (Fig. 7). The model output was constrained using a structured JSON schema, which explicitly defined the expected fields and formats for each ontology type. This structured output format helps reduce the risk of hallucinations and ensures consistency across extractions. Within this schema, the ontology "Element-combination" represents a self-summarization generated by the model, derived from the discussion sections of the literature and grounded in the factual content of the original text. The LLM summarizes the characteristics of these element combinations and their contributions to the mineralization process. All outputs were stored in the KG as attributes of the ontology Element-combination, ensuring traceability and interpretability for subsequent analyses.



**Fig. 7** Ontology design of KG, which includes Deposit, Element-combination, and Element, and corresponding attributes

The ontology “Deposit” is synthesized by the LLM to facilitate a comprehensive review of the literature, requiring the simultaneous extraction of its predefined properties. The ontology “Element” represents a singular entity linked to all instances of geochemical element combinations containing the particular element, thus enabling streamlined statistical analysis and further exploration of the roles of geochemical elements within various ore deposits.

#### 4.2 Triple Extraction by LLMs

Traditional approaches to KG extraction for mineral deposits have primarily relied on deep learning models. For instance, the BERT-BiLSTM [Bidirectional Long Short-Term Memory]-CRF [conditional random field] model has been applied for the joint extraction of rock-related named entities and their relationships (Chen et al. 2022), while models such as convolutional networks (CNN), attention-based BiLSTM (Att-BiLSTM), and Transformers have been employed to identify relationships between gold mining entities (Zhang et al. 2023a). These deep learning models benefit from well-established theoretical frameworks and training methodologies. However, a significant limitation of such approaches is their high demand for human resources. Unlike the field of computer science, geology lacks standardized training datasets such as CoNLL03 (Sang et al. 2003) and CoNLL04 (Roth and Yih 2004), as well as specialized evaluation datasets. Furthermore, the development of these models requires substantial contributions from geological experts for data collection and annotation, alongside computer specialists for model training. These factors significantly extend the timeline for constructing KGs related to mineral deposits, often requiring several months.

Additionally, models trained for specific deposit-related tasks are often not transferable to other tasks, necessitating the collection of new data and retraining. To address these challenges, the use of LLMs offers a more flexible and efficient framework for

information extraction in mineral-type studies. LLMs can significantly reduce development and deployment time, and possess capabilities that traditional deep learning models lack, such as synthesizing and interpreting geochemical data directly from academic literature to provide more nuanced semantic information for KG construction.

Since the release of GPT-3 by OpenAI (Floridi and Chiriatti 2020), LLMs have rapidly advanced to become prominent tools for NLP subtasks. They have demonstrated effectiveness in tasks such as named entity recognition (NER) (Tolegen et al. 2024) and relation extraction (RE) (Li et al. 2024). Unlike traditional deep learning models, LLMs exhibit a form of general intelligence due to their large parameter counts and fine-tuning via human instructions. With carefully designed prompts, LLMs are capable of handling complex tasks including NER, RE, and event extraction (EE). Fine-tuning with task-specific datasets has enabled LLMs to achieve or exceed state-of-the-art performance in these tasks (e.g., Cao et al. 2023; Dagdelen et al. 2024). Current capabilities of LLMs are therefore sufficient to support information extraction tasks related to mineral deposit.

It is important to note that information extraction represents only a subset of the overall linguistic capabilities of LLMs. These models are also capable of understanding instructions, analyzing literature, summarizing content, and answering complex questions—skills that can partially substitute for geological experts when interpreting academic texts. In this study, LLMs are leveraged to synthesize geochemical element data from the discussion sections of papers, enriching the KG with refined and structured information.

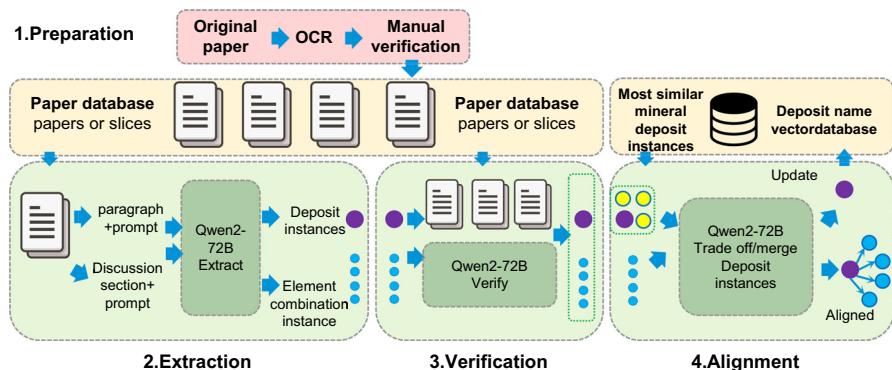
The effectiveness of LLM-based information extraction depends primarily on two factors: the underlying model performance and prompt engineering. The model's performance is influenced by factors such as fine-tuning and parameter count. However, fine-tuning LLMs requires high-quality datasets, which are costly to develop, particularly in specialized domains like geology where established datasets are scarce. Therefore, models with large parameter counts (over 7 billion) were selected, relying on their emergent capabilities to perform information extraction tasks effectively (Schaeffer et al. 2024).

An initial evaluation was conducted to assess the information extraction capabilities of six Qwen-series LLMs hosted on the Alibaba Cloud Bailian platform. Three geological experts evaluated the models by answering five conceptual questions related to rare metal deposits in the Nanling region. Each expert rated the responses on a scale of 1 to 5, and the average scores were compiled (Table 2). Based on model parameters, performance, cost, and geological expertise, the open-source Qwen2-72B-instruct model was selected as the optimal choice for information extraction tasks.

The prompt serves as the directive provided to the LLM, which generates an output in accordance with the instructions. The final extraction process (Fig. 8) incorporates several stages, including extraction, retrieval-augmented generation (RAG) (Lewis et al. 2020), verification, and alignment. This multistage approach ensures the accurate and targeted extraction of required information while minimizing the inclusion of irrelevant content.

**Table 2** Results of the Qwen-series LLM preliminary tests and associated API costs

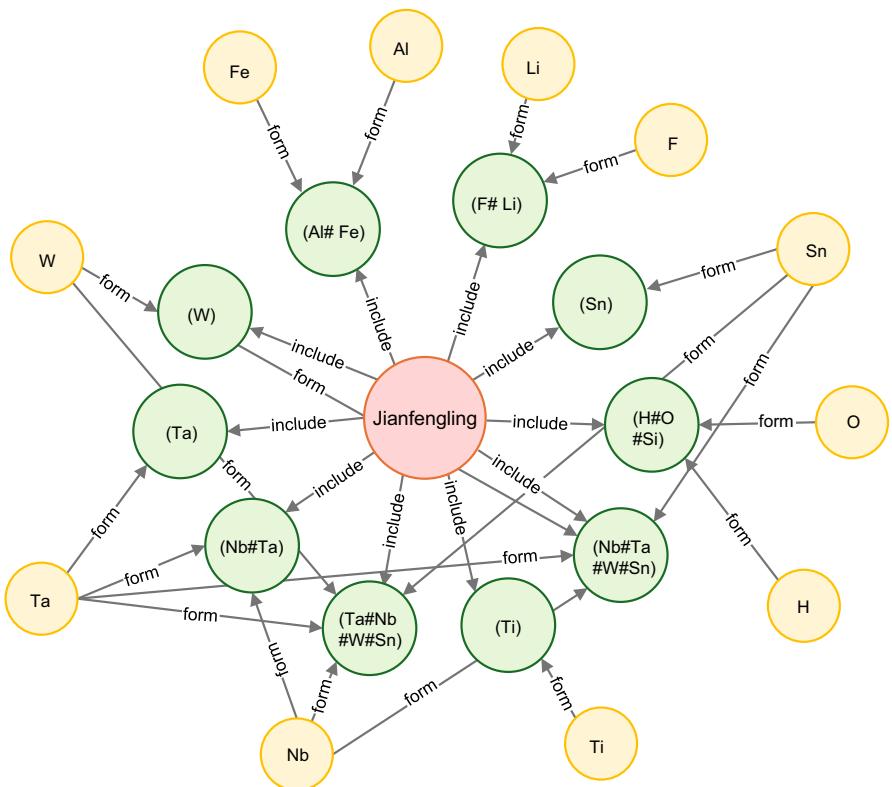
	Model name	Parameter count (in billions)	Output cost (RMB yuan/1000 tokens)	Average score
1	Qwen2-72b-instruct	72	0.012	4.200
2	Qwen2-57b-a14b-instruct	57	0.007	4.200
3	Qwen2-7b-instruct	7	0.002	3.533
4	Qwen-turbo	~ 14	0.0006	4.133
5	Qwen-max	~ 110	0.06	4.466
6	Qwen-long	7	0.002	3.933

**Fig. 8** Mineral deposit and element combination information extraction process

### 4.3 Geochemistry KG

Utilizing the predefined schema and appropriate prompt words, the Qwen2-72B model was applied to extract a KG from a collection of 43 papers. Following judgment and alignment pruning conducted by the LLM, a total of 287 instances of geochemical element combinations were extracted, representing 60 distinct elements (Table S2). All extracted data were imported into the Neo4j KG tool for visualization. An example featuring the Jianfengling deposit, along with its corresponding element combinations and elements, is presented in Fig. 9.

However, due to the phenomenon of LLM hallucinations (Verspoor 2024), inaccuracies may arise, potentially complicating subsequent data analyses. Although a validation section was designed to ensure accuracy, additional manual verification was performed, as the LLM might yield conclusions that deviate from the literature based on its internal knowledge. All mineral deposit examples and half of the extracted element combinations were randomly selected for manual comparison with the original literature. For the deposit names, their consistency with the names in the original texts was verified, with concise answers being considered acceptable. For the element combinations, it was assessed whether the text discussed them holistically and



**Fig. 9** Example for knowledge graph: Jianfengling deposit and its corresponding element combination and element

whether any elements were omitted. Additionally, regarding the effects of the element combinations, ambiguity relative to the original text was evaluated, and the comprehensiveness of the summaries was assessed. The statistical results are summarized in Table 3, indicating a high level of accuracy across the three evaluated items, thus supporting further analyses.

## 5 MPM Predictive Models

In this study, stream sediment geochemical data comprising 39 elements from the Nanling region were utilized. Using this comprehensive dataset, an ALL-model for MPM was developed, incorporating all 39 elements into the predictive model (Fig. 10). However, not all spatial distribution characteristics of these elements accurately represent the mineralization signatures of rare metals. The inclusion of all features indiscriminately introduces noise and reduces model performance (Zekri et al. 2019). Therefore, feature selection is essential in MPM to enhance the accuracy, efficiency, and interpretability of predictive algorithms (e.g., Sun et al. 2024).

**Table 3** The results of manual examination of deposit: name, element-combination: name, and element-combination: effect

Error type	Deposit: name		Element-combination: name		Element-combination: effect	
	Number	Percentage	Number	Percentage	Number	Percentage
Serious discrepancy	0	0%	2	1.40%	1	0.70%
Partial deviation	0	0%	13	9.15%	9	6.33%
Basically consistent	43	0%	127	89.43%	132	92.95%
Sum	43	100%	142	100%	142	100%

\*All deposit names were manually checked, and half of the element combination names and element combination effects were randomly checked

MPM predictive models	Feature selection method	Elements
<b>ALL (39 individual elements)</b>	No feature selection	Ag, Al, As, Au, Ba, Be, Bi, B, Ca, Cd, Co, Cr, Cu, Fe, F, Hg, K, La, Li, Mg, Mn, Mo, Na, Nb, Ni, Pb, P, Sb, Si, Sn, Sr, Th, Ti, U, V, W, Y, Zn, Zr
<b>CLUSTER (25 individual elements)</b>	Data-driven feature selection	Bi, Fe, Zr, Al, Sb, Ag, Ca, Mo, B, Ti, V, Nb, U, Li, Ba, Na, As, P, K, La, Hg, Cu, Sn, Be, Si
<b>KG (18 individual elements)</b>	Knowledge + Data-driven feature selection	Nb, Ta, Re, Sn, Li, W, Si, Al, F, Q, Be, Fe, Zr, Na, Eu, K, Cs, Hf, Nd, Sm, Mo, Ge

Ag – Element in CLUSTER-model, but not in KG-model

Al – Element in all three models

Mg – Element in KG-model, but not in CLUSTER-model

Ta – Element not in ALL-model

Au – Element not in KG- and CLUSTER-models

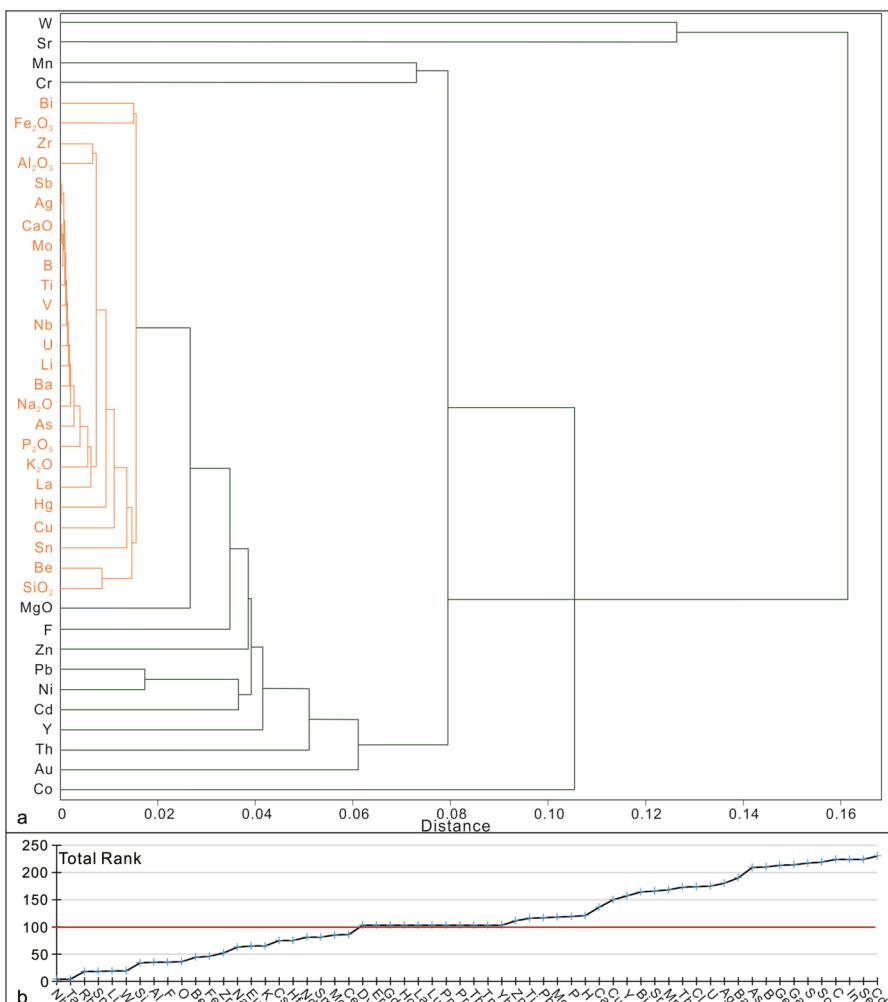
(stream sediment geochemical data)

**Fig. 10** MPM predictive models

By carefully selecting relevant and informative features, noise and irrelevant data can be reduced, thereby improving model precision and generalization capabilities. Feature selection simplifies the model, reducing the risk of overfitting, and optimizing computational resources, thereby accelerating both training and inference. Moreover, this approach promotes a more transparent decision-making process, which is crucial for stakeholders in the mineral exploration industry. It mitigates the “curse of dimensionality,” ensuring that models remain robust and adaptable across diverse geological

contexts. Ultimately, this leads to more reliable and actionable mineral prospectivity insights.

To refine the feature selection process, hierarchical clustering was applied to explore the relationships between geochemical elements, allowing the creation of MPM predictive models based on element groupings. The clustering analysis identified that the elements Bi, Fe, Zr, Al, Sb, Ag, Ca, Mo, B, Ti, V, Nb, U, Li, Ba, Na, As, P, K, La, Hg, Cu, Sn, Be, and Si formed a distinct category (Fig. 11a). These elements are closely associated with rare metals, leading to the development of a CLUSTER-model for MPM, derived from the hierarchical clustering of stream sediment geochemical data (Fig. 10).



**Fig. 11** Feature selection for MPM predictive models. **a** Hierarchical clustering result for 39 elements; **b** total ranking of individual element in KG highlighting the importance of elements within Li–Be–Nb–Ta deposits

While data-driven feature selection offers significant advantages in improving model performance and efficiency, it is not without its challenges. One primary concern is the risk of overfitting, where selected features may correlate well with the target variable in the training data but fail to generalize to new geological contexts. The quality and balance of the training data are critical; poor or biased data can compromise feature selection and, consequently, model reliability. Additionally, the interpretability of data-driven methods can be problematic, as they often prioritize statistical significance over domain-specific knowledge, which may result in the selection of features that are statistically relevant but lack geological meaning (Zuo et al. 2021). Moreover, data-driven methods may fail to capture important feature interactions or nonlinear relationships.

To overcome these limitations, KG-driven feature selection was also explored, aiming to enhance both model interpretability and performance. The statistical methodology employed KG queries to evaluate the degree of correlation between ore-forming elements and individual elements. In the case of lithium deposits, for each deposit entity containing lithium, all combinations of elements were enumerated and their occurrence accounted for. Each pathway connecting a lithium-bearing deposit entity to an element represents a single occurrence, enabling the ranking of all associated elements by their frequency. This approach results in a ranking table highlighting the importance of elements within lithium deposits. Statistical results (Fig. 11b) for lithium, beryllium, niobium, and tantalum are shown in Table 4, with the final ranking reflecting the overall importance of elements across these deposit types. Due to the absence of certain geochemical elements (Ta, Rb, Eu, Cs, Hf, Nd, Sm, and Ce) in the stream sediment geochemical dataset, and given the similar geochemical characteristics of Nb to Ta, Li to Rb and Cs, and La to Eu, Nd, Sm, and Ce, the KG combinations of these elements include Nb and Li. Consequently, La is utilized as an additional surrogate for the four rare earth elements in the KG-model. This process facilitated the development of a KG-model for MPM based on statistical insights from the KG database (Fig. 10), offering a robust and interpretable approach to MPM.

## 6 Discussion

### 6.1 Comparison Experiments

#### 6.1.1 Weak Information Extraction Versus Raw Data

Geochemical measurements of stream sediments typically reflect a combination of weathering products from multiple geological bodies, representing various geological processes. Consequently, element concentrations in these sediments comprise both mineralization-related enrichment and background values from non-mineralized geological formations. Distinguishing genuine geochemical anomalies linked to mineralization from background noise is a critical practice in MPM, particularly for concealed ore deposits in covered regions.

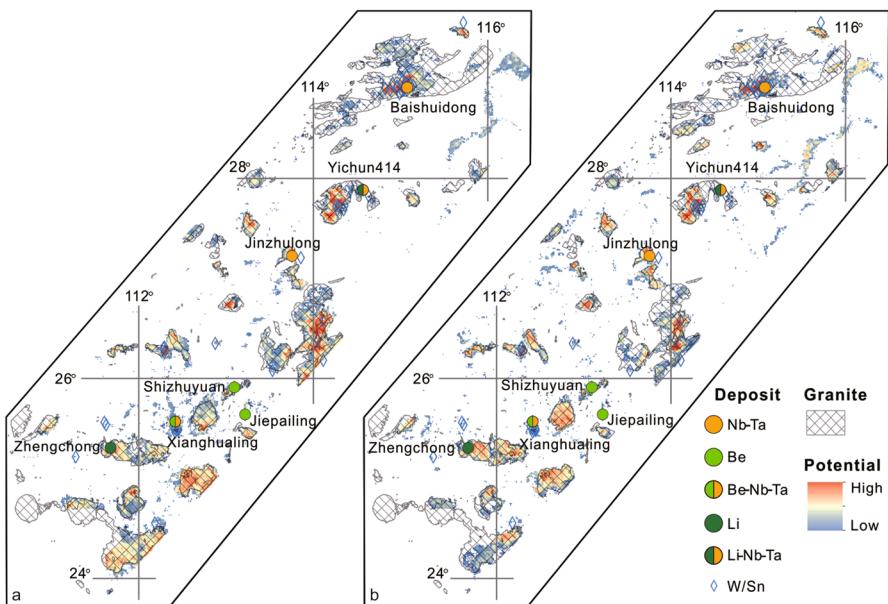
**Table 4** The importance of geochemical elements in Nanling rare metal deposit calculated from Geochemistry KG

Importance ranking*	Element	Importance ranking score					Total
		Lithium deposits	Beryllium deposits	Niobium deposits	Tantalum deposits		
1	Nb	1	1	1	1		4
2	Ta	1	1	1	1		4
3	Rb	5	5	4	4		18
4	Sn	4	6	4	4		18
5	Li	3	4	6	6		19
6	W	6	7	3	3		19
7	Si	10	8	8	8		34
8	Al	9	10	8	8		35
9	F	7	14	7	7		35
10	O	7	13	8	8		36
11	Be	13	3	14	14		44
12	Fe	10	12	12	12		46
13	Zr	12	18	11	11		52
14	Na	14	11	19	19		63
15	Eu	19	22	12	12		65
16	K	16	9	20	20		65
17	Cs	15	18	21	21		75
18	Hf	16	23	18	18		75
19	Nd	25	26	15	15		81
20	Sm	25	26	15	15		81
21	Mo	29	14	21	21		85
22	Ce	25	31	15	15		86

\*Only the first 22 elements are listed

The Nanling region, characterized by extensive forest and crop cover, serves as a typical example of such covered areas, where the geochemical signatures of ore-forming geological bodies are often weak. Accurately identifying these weak and complex anomalies is essential for effective exploration in this region. Cheng (2012) highlighted that traditional approaches, which rely solely on anomaly magnitudes, face limitations when applied in covered terrains due to the intricate nature of the anomaly patterns in such settings.

Recent studies have shown that geochemical distributions exhibit fractal and multifractal characteristics, which can be leveraged to distinguish mineralization-related processes from background geological processes. The S-A multifractal filtering model (Cheng et al. 2000), which assumes that the spatial distribution of geochemical elements follows distinct fractal patterns for mineralized and non-mineralized processes, has emerged as an advanced method for isolating weak geochemical anomalies. This



**Fig. 12** Comparative predictive mapping: absence versus presence of S-A fractal filtering methods. **a** ALL-noSA-RF; **b** ALL-SA-RF. Only predictions with high value ( $> 0.5$ ) are displayed

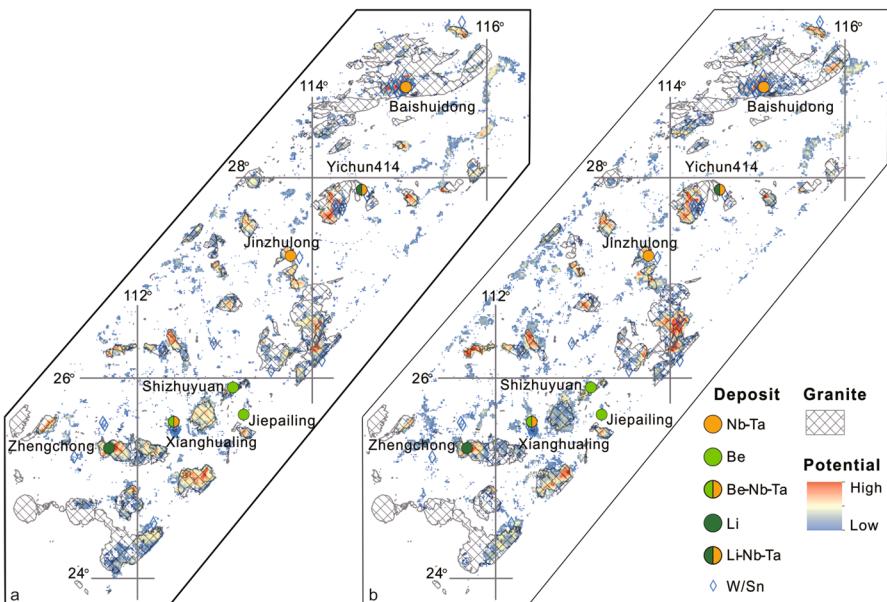
model has proven effective in characterizing singular mineralization events and predicting the locations of concealed ore deposits (Zuo et al. 2013; Zhang et al. 2016; Ghezelbash et al. 2019; Pourgholam et al. 2021).

In this study, a comparative experiment was conducted to evaluate the impact of applying S-A fractal filtering to geochemical elements in MPM. To ensure comparability, consistency in the prediction method, training set, and test set was maintained across all experiments. Optimal parameter searches were employed to configure the models appropriately. The final optimal parameters for each model are summarized in Table S3.

The results demonstrate that applying S-A fractal filtering significantly refined the delineation of high-potential areas for mineralization (Fig. 12). The filtered geochemical data resulted in a reduction in the size of these high-potential zones, effectively concentrating the focus on areas with higher mineralization potential. This outcome highlights the method's capacity to reduce exploration risk and investment by narrowing target zones, leading to improved success rates in subsequent detailed exploration efforts.

### 6.1.2 KG Versus Non-KG Methods

Three MPM predictive models—All-, CLUSTER-, and KG-model—were developed to assess their effectiveness in identifying mineralization potential. To validate model performance, a comparative experiment was conducted using geochemical elements that had been filtered through S-A fractal filtering. The MPM analysis was performed



**Fig. 13** Comparative predictive mapping: application of CLUSTER feature selection methods versus KG feature selection methods. **a** CLUSTER-SA-RF; **b** KG-SA-RF

using the RF method, ensuring consistency between the training and testing sets, and optimal parameter searches were employed to configure the models appropriately.

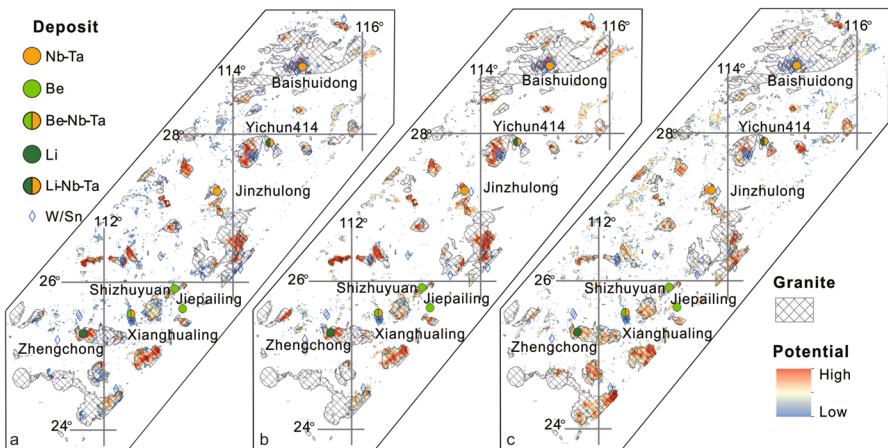
The experimental results (Figs. 12b, 13) indicated that the KG-model, which incorporates a KG with selected variables, led to a more precise delineation of high-potential mineralization areas. The area of these zones was reduced, thereby highlighting regions with true mineralization potential. This reduction in target zones reduces exploration risk and cost, ultimately enhancing the efficiency of mineral exploration efforts.

Moreover, the KG-model's predictions revealed that most of the high-potential mineralization areas are situated along the upper edges of known granite bodies. This spatial distribution aligns well with the characteristics of rare metal deposits in granite-hosted mineralization systems, which typically form during the final stages of magma differentiation. This consistency reinforces the relevance of the KG-model for predicting rare metal mineralization associated with granitic intrusions.

### 6.1.3 Different ML Methods

To further evaluate whether the choice of ML model impacts the results of MPM, a comparative experiment was conducted using several ML methods, including RF, SVM, GC, and transfer learning. In this experiment, the input variables were geochemical elements in the KG-model, all of which had undergone S-A multifractal filtering to enhance the true characteristics of mineralization anomalies.

The experimental results (Fig. 14) demonstrated consistent predictions across the different methods, indicating the stability of the prediction outcomes. Among the



**Fig. 14** Comparative predictive mapping for different ML methods. **a** KG-SA-SVM; **b** KG-SA-GC; **c** KG-SA-TL

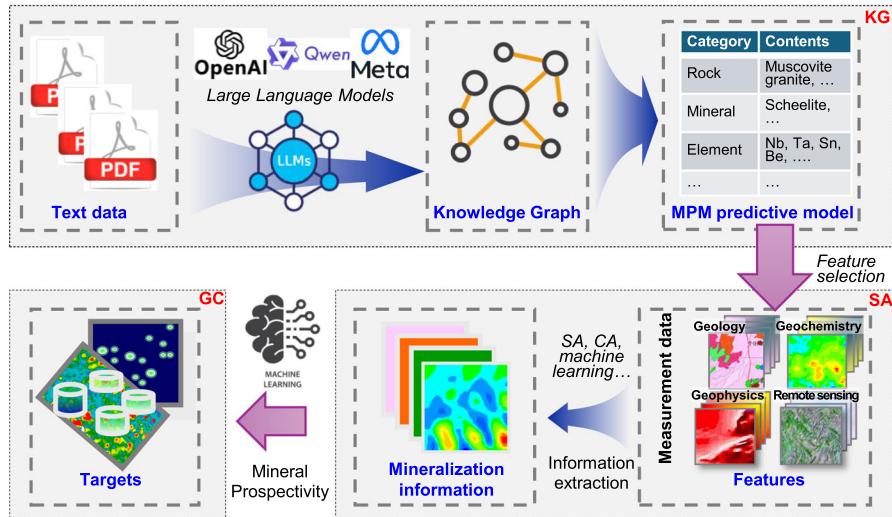
models, the GC ML method yielded the most accurate and reliable results, suggesting it is particularly well suited for MPM when applied to geochemically filtered data. These findings further confirm the effectiveness of using the KG-model to refine input data for ML models in mineral exploration.

#### 6.1.4 KG-SA-GC Framework for MPM

The experimental results consistently demonstrated that prediction models using the KG for feature selection outperformed other approaches in terms of accuracy, precision, recall, and F1 score (Fig. 15). These findings highlight the robustness and generalizability of KG-based models. Furthermore, the application of S-A fractal filtering for anomaly extraction significantly enhanced the prediction performance compared to models without anomaly filtering. However, the experiments also revealed

MPM predictive models	Weak information extraction method	Machine Learning Method	Accuracy	Precision	Recall	F1 Score
ALL	noSA	RF	0.9	1	0.7	0.82
CLUSTER			0.83	0.78	0.7	0.74
KG			0.833	0.778	0.7	0.737
ALL	SA	RF	0.83	0.86	0.6	0.71
CLUSTER			0.87	0.8	0.8	0.8
KG			0.867	0.875	0.7	0.777
KG	SA	SVM	0.833	0.857	0.6	0.706
KG		Transfer Learning	0.922	0.909	0.851	0.879
KG		GC	0.867	1	1	1

**Fig. 15** Model performance for different models used in this work



**Fig. 16** Workflow for the KG-SA-GC MPM framework

that without the integration of cluster analysis or KG-based feature selection, the benefits of S-A fractal filtering were less pronounced. This indicates that including all features indiscriminately, without proper selection, can negatively affect the performance of the prediction model. Overall, when KG-selected features were combined with S-A fractal filtering, the gcForest (GC) model exhibited stability across different ML methods, while maintaining superior predictive accuracy. Given its high performance and interpretability, the GC model is recommended for predicting rare metal deposits in the Nanling region.

Thus, the KG-SA-GC framework (Fig. 16) is proposed as a robust and accurate approach for predicting rare metal deposits. This framework employs LLMs to build a KG database, which serves as the foundation for an interpretability-enhanced MPM predictive model. The KG-model enables the selection of key features for ML predictions, which are further refined using S-A fractal filtering to isolate true mineralization anomalies. These filtered anomalies are then input into the GC ML model, producing reliable and precise predictions for mineral deposits.

## 6.2 Strong Interpretability of KG

Although data-driven hierarchical clustering methods can identify geochemical elements in stream sediments closely related to Li–Be–Nb–Ta mineralization, the relationships derived from these methods often lack sufficient explanatory power. Such approaches typically require geologists to apply strong inductive reasoning, consulting extensive literature and relying heavily on experience to interpret the interactions. By contrast, a KG constructed using LLMs allows for more direct tracing of dependencies between elements, enabling linkages to be made with relevant primary research

literature. This enhances both the interpretability and transparency of the geochemical associations identified.

The roles of key elements involved in ore formation were systematically assessed using LLMs. Elements such as Nb (1) [the order of importance of the elements are in parentheses], Ta (2), Sn (4), Li (5), W (6), and Be (11) emerged as the primary ore-forming constituents. Notably, Li, Nb, and Ta display a stronger association with Sn mineralization than with W mineralization. This discrepancy is attributed to the difference in partitioning between aqueous fluid and granitic melt for W and Sn, where wolframite deposition occurs earlier and preferentially to cassiterite during the magmatic-hydrothermal stage. Consequently, Li, Be, Nb, and Ta are more likely to be enriched in the later stages, in association with cassiterite (Michaud et al. 2020).

Elements including Si (7), Al (8), Na (14), and K (16) serve as indicators of magma differentiation. Alteration of potassium and sodium feldspar, linked to the crystallization of fractionated peralkaline magma rich in rare metals, plays a crucial role in the activation and enrichment of ore-forming elements, and thus is central to understanding the mechanisms of mineralization (Dostal 2016; Harlaux et al. 2017). The enrichment of alkali metals such as Rb (3), Li (5), and Cs (17) also reflects the high degree of differentiation that characterizes evolved granitic systems (Liu et al. 2022). Low Nb/Ta and Zr (13)/Hf (18) ratios serve as key indicators of highly differentiated granites, pointing to late-stage melt–fluid interactions during magma evolution (Linnen 1998; Dostal and Chatterjee 2000; Zaraisky et al. 2009), which are critical for rare metal mineralization processes.

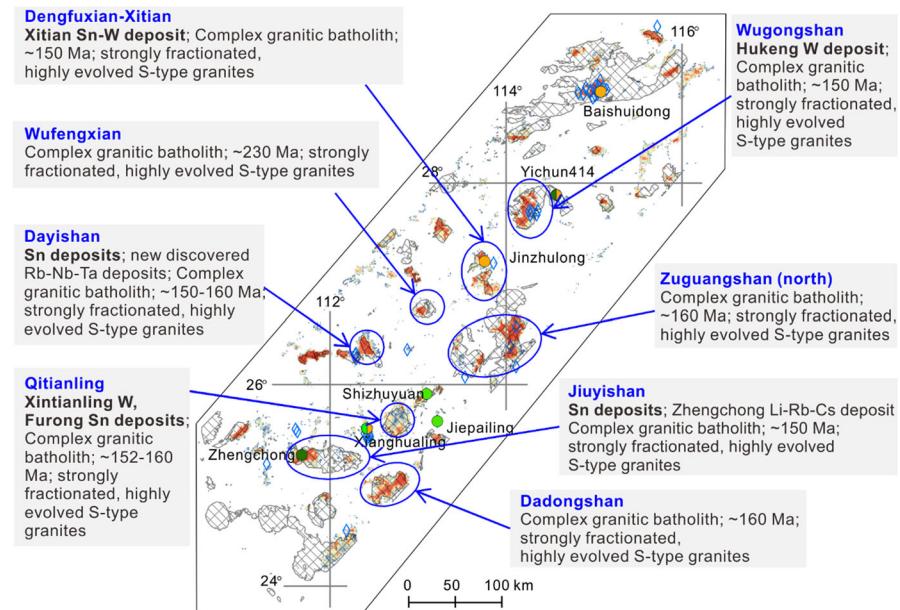
Elevated Fe (12) content in ore minerals, along with the presence of magnetite and hematite, indicates deep hydrothermal activity, offering insights into the thermodynamic conditions of the mineralization systems (Nadol et al. 2014). Element F (9), which reflects fluid and volatile compositions, plays a fundamental role in controlling the migration, enrichment, and deposition of rare metals (Sheard et al. 2012; Vasyukova and Williams-Jones 2019). Although REEs such as Eu (15), Nd (19), Sm (20), and Ce (22)], as well as Mo (21), are not the primary ore-forming elements, they contribute to the overall economic value of the deposits by serving as valuable associated resources (Hu et al. 2017).

### 6.3 Implications for Rare Metal Prospecting in Nanling

Using the KG-SA-GC framework, eight key prospective zones for rare metal mineralization were delineated (Fig. 17). All these zones are controlled by highly differentiated and evolved granites, which align well with previous studies. The areas of high mineralization favorability largely coincide with granitic bodies that have undergone significant differentiation and evolution, indicating potential mineralization along the margins of these intrusions.

#### 6.3.1 Dayishan Granitic Complex

The Dayishan granite body, an elongated structure aged between 144 and 160 Ma, is composed of highly differentiated A-type granites, characterized by consistently high



**Fig. 17** Eight key prospective zones for rare metal deposits obtained by the KG-SA-GC MPM framework

content of high-field-strength elements (HFSEs) with limited variation (Zhang et al. 2021, 2022). Predictive modeling results indicate a high potential for mineralization within the highly differentiated granite located in the northern Dayishan region and on the western side of the Tashan granite body (Fig. 17). While historical exploration has primarily targeted Sn deposits, recent geological surveys have emphasized the potential for Li–Rb–Nb–Ta–Be rare metals in these areas (Zhang et al. 2021; Zhao et al. 2022). These rare metals are mainly associated with altered granite, greisen, and pegmatite formations (Zhang et al. 2021).

### 6.3.2 Jiuyishan Granitic Batholith

The Jiuyishan granite complex consists of a sequence of composite plutons, including the Xuehuading, Pangxiemu, Jinjiling, Shaziling, and Xishan plutons, all dated to approximately 152 Ma (Li et al. 2020b). These plutons represent various evolutionary stages of a common parental magma, exhibiting a fractional crystallization trend progressing from the Shaziling to Xishan, Jinjiling, and finally the Pangxiemu plutons (Li et al. 2020b). High-value mineralization zones have been identified within the highly differentiated Pangxiemu pluton and the neighboring Jinjiling pluton, consistent with the predictive results (Fig. 17). This area hosts significant Li-polymetallic deposits, including the Zhengchong giant lithium deposit.

### 6.3.3 Qitianling Granitic Complex

The Qitianling complex, with a mineralization age of 150–163 Ma (Zhu et al. 2009; Zhao et al. 2012), consists of A-type granites derived from mixed crust-mantle magmas, exhibiting enriched Nd and Hf isotopic compositions (Zhao et al. 2012). The surrounding Carboniferous-Permian limestones host rare metal deposits associated with skarn, greisen, and altered rock types. Elevated concentrations of Li, Rb, Cs, and Be have been observed in the eastern Qitianling region, consistent with the high-value zones delineated by predictive modeling in the southeastern part of the intrusion (Fig. 17). These zones coincide with fine- to medium-grained biotite granites, where the giant Furong tin deposit is also located (Li et al. 2007).

### 6.3.4 Dadongshan Granitic Complex

The Dadongshan intrusive body, formed around 160 Ma, consists of highly differentiated and evolved I-type granites (Huang et al. 2008). The western part of Dadongshan complex shows significant potential for uranium-polymetallic mineralization. Although exploration in this region is limited due to its location within a national nature reserve, numerous W, Sn, and REE deposits have been identified around the intrusive body. The northern contact zone of the western Dadongshan body represents a high-value zone for mineralization (Fig. 17), with good potential for polymetallic deposits.

### 6.3.5 Wugongshan Granitic Complex

In the Wugongshan area, a series of composite intrusive bodies have been identified, with magmatic events spanning from over 400 Ma to around 150 Ma (Wang et al. 2001). Mineralization events, however, are dated to approximately 150 Ma (Liu et al. 2011). High-value mineralization zones are concentrated within both the western granites and the eastern Yashan granite (Fig. 17). The Yashan granite hosts the giant Yichun 414 Nb-Ta-Li deposit, which is characterized by altered granite-type mineralization (Lin et al. 1995; Yin et al. 2022). In the western granites, highly differentiated S-type granites contain smaller Nb-Ta deposits within pegmatites and granites in the Wankeng and Liujiashan area (Xu et al. 2022). Predictive modeling results indicate significant potential for further mineralization near these known Nb-Ta deposits.

### 6.3.6 Dengfuxian-Xitian Granitic Complexes

The Dengfuxian and Xitian granitic complexes contain both Triassic and Late Jurassic intrusive events, dated to approximately 230–220 Ma and 160–150 Ma, respectively, indicating a close temporal, spatial, and material relationship (Li et al. 2019). These complexes are associated with Mesozoic W-Sn mineralization events, with three distinct episodes of mineralization inferred in the Dengfuxian and Xitian ore fields: (1) Triassic (~220 Ma), (2) Jurassic (~150 Ma), and (3) Cretaceous (~80 Ma). The Dengfuxian intrusive body hosts the large Jinzhulong Nb-Ta deposit, characterized by

Nb–Ta–W–Sn mineralization related to altered granite (Mao et al. 2021). In contrast, the Xitian deposit shows less pronounced Li–Be–Nb–Ta mineralization. Predictive result suggests that both complexes exhibit high potential for rare metal mineralization, with Xitian considered a key area for Li–Be–Nb–Ta target exploration (Fig. 17).

### 6.3.7 Zhuguangshan Granitic Complex

The Zhuguangshan complex is divided into southern and northern segments, formed during the Caledonian, Indosian, and Yanshanian periods (Zhang et al. 2018). While the southern segment is well known for its high potential for granite-type uranium deposits (Zhang et al. 2017), predictive results highlight the northern segment (Fig. 17). Specifically, high-potential zones have been identified in the vicinity of the Linyang intrusive body, which formed around 230 Ma, and its northern periphery. This suggests that the northern segment of Zhuguangshan may hold significant potential for rare metal exploration.

### 6.3.8 Wufengxian Granitic Batholith

The Wufengxian granite, dated to 236 Ma, is composed of highly differentiated S-type granites (Wang et al. 2007). However, metallic mineralization in this body is underdeveloped and has received less attention than other intrusive bodies in the Nanling region. The predictive modeling identifies high-value mineralization zones corresponding to the most evolved biotite monzogranites, suggesting potential for Li–Be–Nb–Ta mineralization (Fig. 17). These findings indicate that further exploration in this area is warranted.

## 7 Conclusion

An innovative framework is proposed to address the challenges of MPM, focusing on the integration of LLMs with KG construction to improve interpretability and accuracy in predictive modeling. By leveraging academic literature, key relationships between geochemical elements were mined, and a comprehensive KG was constructed to capture the complex interactions influencing rare metal mineralization.

The KG-SA-GC framework constitutes a unified and interdependent workflow designed to enhance both the interpretability and accuracy of mineral prospectivity modeling. It systematically integrates KG-based feature selection, S-A fractal filtering for anomaly enhancement, and the gcForest algorithm for predictive modeling. Rather than a collection of independent tools, each component builds upon the outputs and logic of the previous step—domain-informed KG features guide the filtering process, which in turn shapes the structure of the predictive model. Applied to the Nanling region, this integrative framework effectively identified critical geochemical predictors, offering new insights into rare metal mineralization. Moreover, the use of domain knowledge throughout the pipeline strengthens geological interpretability and supports informed decision-making in exploration strategies.

However, certain limitations remain. In this study, the LLM was primarily employed for knowledge extraction and organization from the geological literature, rather than for direct interpretation of spatial geochemical patterns. Future work may explore the potential of LLMs to interpret geochemical anomalies directly, providing additional layers of explainability. The current model is based exclusively on stream sediment geochemical data, which, while offering wide spatial coverage, may not capture lithological or structural controls as effectively as other sampling media (e.g., rock chip or soil geochemistry). Moreover, input feature types and quality can vary considerably across different geological contexts and survey designs, which may influence the generalizability of ML-based MPM frameworks. Incorporating additional datasets such as geophysical, geological, mineral alteration, and isotopic information could enhance the robustness and applicability of the framework. Furthermore, the unified model developed for Li, Be, Nb, and Ta represents a simplification of their distinct mineralization processes. Refinement through element-specific predictive models is recommended to better account for their individual geological characteristics.

Despite these limitations, the results demonstrate the scalability and adaptability of the proposed approach. While applied to the Nanling region, the methodology is transferable to other geological settings, making it a versatile tool for mineral exploration. As global demand for rare metals continues to increase, this framework offers a valuable contribution to modern exploration strategies and lays the groundwork for further advancements in the field.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11004-025-10231-3>.

**Acknowledgements** This work was supported by the National Key R & D Program of China [No. 2023YFC2906402], the National Natural Science Foundation of China [Nos. 42430111, 42472358, and 42050103], and the Fundamental Research Funds for the Central Universities [No. 2652023001].

**Author Contributions** Conceptualization: Zhen-Jie Zhang; Data curation: Zi-Xing Yang, Fu-Yuan Jian, Si-Yuan Ban, & Yi-Ming Wang; Formal analysis: Zhen-Jie Zhang; Funding acquisition: Zhen-Jie Zhang; Investigation: Zhen-Jie Zhang; Methodology: Zhen-Jie Zhang; Project administration: Zhen-Jie Zhang; Resources: Zhen-Jie Zhang; Software: Zi-Xing Yang, Fu-Yuan Jian, Si-Yuan Ban, & Yi-Ming Wang; Supervision: Zhen-Jie Zhang; Writing—original draft: Zhen-Jie Zhang, Zi-Xing Yang, Fu-Yuan Jian, Si-Yuan Ban, & Yi-Ming Wang; Writing—review & editing: Zhen-Jie Zhang.

## Declarations

**Competing interests** The authors declare no competing interests.

## References

- Agterberg FP (1989) Computer programs for mineral exploration. *Science* 245:76–81
- Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, Fan Y, Ge W, Han Y, Huang F (2023) Qwen technical report. arXiv preprint [arXiv:2309.16609](https://arxiv.org/abs/2309.16609)
- Bodenreider O (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 32(suppl\_1):D267–D270
- Bonham-Carter G (1994) Geographic information systems for geoscientists: modelling with GIS vol 4. Computer Methods in the Geosciences Pergamon, vol 4

- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Brown WM, Gedeon T, Groves D, Barnes R (2000) Artificial neural networks: a new method for mineral prospectivity mapping. *Aust J Earth Sci* 47(4):757–770
- Cao M, Wang H, Liu X, Wu J, Zhao MLLM (2023) Collaboration PLM improves critical information extraction tasks in medical articles. China Health Information Processing Conference. Springer, pp 178–185
- Carranza EJM, Laborte AG (2015) Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: application of random forests algorithm. *Ore Geol Rev* 71:777–787
- Carranza EJM, Hale M, Faassen C (2008) Selection of coherent deposit-type locations and their application in data-driven mineral prospectivity mapping. *Ore Geol Rev* 33(3–4):536–558
- Chen Y, Wu W, Zhao Q (2020) A bat algorithm-based data-driven model for mineral prospectivity mapping. *Nat Resour Res* 29(1):247–265
- Chen Z, Yuan F, Li X, Zhang M (2022) Based on BERT-BiLSTM-CRF model the named entity and relation joint extraction of Chinese lithological description corpus. *Geol Rev* 68(2):742–750
- Cheng Q (2012) Singularity theory and methods for mapping geochemical anomalies caused by buried sources and for predicting undiscovered mineral deposits in covered areas. *J Geochem Explor* 122:55–70
- Cheng Q, Xu Y, Grunsky E (2000) Integrated spatial and spectrum method for geochemical anomaly separation. *Nat Resour Res* 9(1):43–52
- Dagdelen J, Dunn A, Lee S, Walker N, Rosen AS, Ceder G, Persson KA, Jain A (2024) Structured information extraction from scientific text with large language models. *Nat Commun* 15(1):1418
- Dong Y-L, Zhang Z-J (2024) Deep forest modeling: an interpretable deep learning method for mineral prospectivity mapping. *J Geophys Res Mach Learn Comput* 1(4):e2024JH000311
- Dostal J, Chatterjee AK (2000) Contrasting behaviour of Nb/Ta and Zr/Hf ratios in a peraluminous granitic pluton (Nova Scotia, Canada). *Chem Geol* 163(1):207–218
- Dostal J (2016) Rare metal deposits associated with alkaline/peralkaline igneous rocks. In: Verplanck PL, Hitzman MW (eds) Rare earth and critical elements in ore deposits, vol 18. Society of Economic Geologists
- Floridi L, Chiriatti M (2020) GPT-3: its nature, scope, limits, and consequences. *Mind Mach* 30:681–694
- Ghezelbasi R, Maghsoudi A, Carranza EJM (2019) Performance evaluation of RBF-and SVM-based machine learning algorithms for predictive mineral prospectivity modeling: integration of SA multi-fractal model and mineralization controls. *Earth Sci Inf* 12(3):277–293
- Han T, Rukhlov AS (2020) Update of the provincial Regional Geochemical Survey (RGS) database at the British Columbia Geological Survey. British Columbia Ministry of Energy, Mines and Petroleum Resources, British Columbia Geological Survey, GeoFile 2020-08, 3p
- Harlaux M, Mercadier J, Bonzi WME, Kremer V, Marignac C, Cuney M (2017) Geochemical signature of magmatic-hydrothermal fluids exsolved from the beauvoir rare-metal granite (Massif Central, France): insights from LA-ICPMS analysis of primary fluid inclusions. *Geofluids* 1:1925817
- Harris JR, Wilkinson L, Heather K, Fumerton S, Bernier MA, Ayer J, Dahn R (2001) Application of GIS processing techniques for producing mineral prospectivity maps—a case study: mesothermal Au in the Swayze Greenstone Belt, Ontario, Canada. *Nat Resour Res* 10(2):91–124
- Harris JR, Grunsky E, Behnia P, Corrigan D (2015) Data-and knowledge-driven mineral prospectivity maps for Canada's North. *Ore Geol Rev* 71:788–803
- Harris JR, Naghizadeh M, Behnia P, Mathieu L (2022) Data-driven gold potential maps for the Chibougamau area, Abitibi greenstone belt, Canada. *Ore Geol Rev* 150:105176
- Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B (1998) Support vector machines. *IEEE Intell Syst Appl* 13(4):18–28
- Hronsky JM, Kreuzer OP (2019) Applying spatial prospectivity mapping to exploration targeting: fundamental practical issues and suggested solutions for the future. *Ore Geol Rev* 107:647–653
- Hu X, Gong Y, Pi D, Zhang Z, Zeng G, Xiong S, Yao S (2017) Jurassic magmatism related Pb-Zn-W-Mo polymetallic mineralization in the central Nanling Range, South China: geochronologic, geochemical, and isotopic evidence from the Huangshaping deposit. *Ore Geol Rev* 91:877–895
- Huang HQ, Li XH, Li WX, Liu Y (2008) Age and origin of the Dadongshan granite from the Nanling range: SHRIMP U-Pb zircon age, geochemistry and Sr-Nd-Hf isotopes. *Geol J China Univ* 14(3):317–333 (in Chinese with English abstract)

- Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, Hellmann S, Morsey M, Van Kleef P, Auer S (2015) Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semant Web* 6(2):167–195
- Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-t, Rocktäschel T (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv Neural Inf Process Syst* 33:9459–9474
- Li Z, Hu R, Yang J, Peng J, Li X, Bi X (2007) He, Pb and S isotopic constraints on the relationship between the A-type Qitianling granite and the Furong tin deposit, Hunan Province, China. *Lithos* 97(1–2):161–173
- Li H, Sun HS, Algeo TJ, Wu JH, Cao JY, Wu QH (2019) Mesozoic multi-stage W-Sn polymetallic mineralization in the Nanling range, South China: an example from the Dengfuxian-Xitian ore field. *Geol J* 54(6):3755–3785
- Li H, Li X, Yuan F, Jowitt SM, Zhang M, Zhou J, Zhou T, Li X, Ge C, Wu B (2020a) Convolutional neural network and transfer learning based mineral prospectivity modeling for geochemical exploration of Au mineralization within the Guandian-Zhangbaling area, Anhui Province, China. *Appl Geochem* 122:104747
- Li J, Fu J, Ma C, Lu Y, Cheng S, Ma L, Qin Z, Sheng H (2020b) Petrogenesis and tectonic setting of the Shaziling pluton in Jiuyishan Area, Nanling: Evidence from Zircon U-Pb geochronology, petrogeochemistry, and Sr-Nd-Hf isotopes. *Earth Sci* 45(2):374–388 (in Chinese with English abstract)
- Li Y, Peng X, Li J, Zuo X, Peng S, Pei D, Tao C, Xu H, Hong N (2024) Relation extraction using large language models: a case study on acupuncture point locations. *J Am Med Inform Assoc*. <https://doi.org/10.1093/jamia/ocae233>
- Lin Y, Pollard PJ, Hu S, Taylor RG (1995) Geologic and geochemical characteristics of the Yichun Ta-Nb-Li deposit, Jiangxi Province, South China. *Econ Geol* 90(3):577–585
- Linnen RL (1998) The solubility of Nb-Ta-Zr-Hf-W in granitic melts with Li and Li + F; constraints for mineralization in rare metal granites and pegmatites. *Econ Geol* 93(7):1013–1025
- Liu J, Mao J, Ye H, Zhang W (2011) Geology, geochemistry and age of the Hukeng tungsten deposit, Southern China. *Ore Geol Rev* 43(1):50–61
- Liu X-H, Li B, Xu J-W, He B, Liao J, Peng H-W, Wang Y-H, Lai J-Q (2022) Monazite geochronology and geochemistry constraints on the formation of the giant Zhengchong Li-Rb-Cs deposit in South China. *Ore Geol Rev* 150:105147
- Ma X (2022) Knowledge graph construction and application in geosciences: a review. *Comput Geosci-UK* 161:105082
- Mao J, Pirajno F, Cook N (2011) Mesozoic metallogeny in East China and corresponding geodynamic settings—an introduction to the special issue. *Ore Geol Rev* 43(1):1–7
- Mao X, Zhang W, Liu Z, Ren J, Bayless RC, Deng H (2020) 3D mineral prospectivity modeling for the low-sulfidation epithermal gold deposit: a case study of the Axi gold deposit, western Tianshan, NW China. *Minerals* 10(3):233
- Mao Y, Shao Y, Xiong Y, Jiang S, Wen C (2021) Magmatic-hydrothermal metallogenic system in Nb-Ta-W-Sn-Pb-Zn Dengfuxian orefield, eastern Hunan: constraint from U-Pb geochronology of columbite-tantalite. *J Cent South Univ (Sci Technol)* 52(9):2959–2972 (in Chinese with English abstract)
- Michaud JA-S, Gumiaux C, Pichavant M, Gloaguen E, Marcoux E (2020) From magmatic to hydrothermal Sn-Li-(Nb-Ta-W) mineralization: the Argemela area (central Portugal). *Ore Geol Rev* 116:103215
- Montsion RM, Saumur BM, Acosta-Gongora P, Gadd MG, Tschirhart P, Tschirhart V (2019) Knowledge-driven mineral prospectivity modelling in areas with glacial overburden: porphyry Cu exploration in Quesnellia, British Columbia, Canada. *Appl Earth Sci* 128(4):181–196
- Murtagh F, Contreras P (2012) Algorithms for hierarchical clustering: an overview. *Wires Data Min Knowl Discov* 2(1):86–97
- Nadoll P, Angerer T, Mauk JL, French D, Walshe J (2014) The chemistry of hydrothermal magnetite: a review. *Ore Geol Rev* 61:1–32
- Porwal A, Carranza EJM, Hale M (2003) Artificial neural networks for mineral-potential mapping: a case study from Aravalli Province, Western India. *Nat Resour Res* 12(3):155–171
- Porwal A, Das RD, Chaudhary B, Gonzalez-Alvarez I, Kreuzer O (2015) Fuzzy inference systems for prospectivity modeling of mineral systems and a case-study for prospectivity mapping of surficial uranium in Yeelirrie Area, Western Australia. *Ore Geol Rev* 71:839–852
- Pourgholam MM, Afzal P, Yasrebi AB, Gholinejad M, Wetherelt A (2021) Detection of geochemical anomalies using a fractal-wavelet model in Ipak area, Central Iran. *J Geochem Explor* 220:106675

- Raiaan MAK, Mukta MSH, Fatema K, Fahad NM, Sakib S, Mim MMJ, Ahmad J, Ali ME, Azam S (2024) A review on large language models: architectures, applications, taxonomies, open issues and challenges. *IEEE Access* 12:26839–26874
- Richards JP (2003) Tectono-magmatic precursors for porphyry Cu-(Mo-Au) deposit formation. *Econ Geol* 98(8):1515–1533
- Richards JP (2016) Clues to hidden copper deposits. *Nat Geosci* 9:195–196
- Roth D, Yih W-t (2004) A linear programming formulation for global inference in natural language tasks. Paper presented at the Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004
- Sang EF, De Meulder F (2003) Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003, Edmonton, Canada, 2003. pp 142–147
- Schaeffer R, Miranda B, Koyejo S (2024) Are emergent abilities of large language models a mirage? *Adv Neural Inf Process Syst* 36:55565–55581
- Sheard ER, Williams-Jones AE, Heiligmann M, Pederson C, Trueman DL (2012) Controls on the concentration of zirconium, niobium, and the rare earth elements in the Thor Lake rare metal deposit, Northwest Territories, Canada. *Econ Geol* 107(1):81–104
- Singer D, Mosier D (1981) A review of regional mineral resource assessment methods. *Econ Geol* 76(5):1006–1015
- Su H-M, Jiang S-Y, Cao M-Y, Luo P (2020) Rare-metal mineralization potential and petrogenesis of early Cretaceous I-type granitic rocks in the Liziwang volcanic basin of Jiangxi Province, South China: evidence from mineralogy, geochemistry, and geochronology. *Miner Deposita* 55(3):453–468
- Sun T, Feng M, Pu W, Liu Y, Chen F, Zhang H, Huang J, Mao L, Wang Z (2024) Fractal-based multi-criteria feature selection to enhance predictive capability of AI-driven mineral prospectivity mapping. *Fractal Fract* 8(4):224
- Tolegen, G, Toleu, A, Mussabayev, R (2024) Enhancing low-resource NER via knowledge transfer from LLM. In: Nguyen NT, Franczyk B, Ludwig A, Núñez M, Treur J, Vossen G, Kozierkiewicz A (eds) Computational collective intelligence. ICCC 2024. Lecture notes in computer science. Springer, Cham, pp 238–248
- USGS (2016) National geochemical database: sediment. U.S. Geological Survey. <https://mrdata.usgs.gov/ngdb/sediment>
- Vasyukova OV, Williams-Jones AE (2019) Closed system fluid-mineral-mediated trace element behaviour in peralkaline rare metal pegmatites: evidence from Strange Lake. *Chem Geol* 505:86–99
- Verspoor K (2024) 'Fighting fire with fire'—using LLMs to combat LLM hallucinations. *Nature* 630(8017):569–570
- Wang D, Shu L, Faure M, Sheng W (2001) Mesozoic magmatism and granitic dome in the Wugongshan Massif, Jiangxi province and their genetic relationship to the tectonic events in southeast China. *Tectonophysics* 339(3–4):259–277
- Wang Y, Fan W, Sun M, Liang X, Zhang Y, Peng T (2007) Geochronological, geochemical and geothermal constraints on petrogenesis of the Indosinian peraluminous granites in the South China Block: a case study in the Hunan Province. *Lithos* 96(3):475–502
- Wang F-Y, Ling M-X, Ding X, Hu Y-H, Zhou J-B, Yang X-Y, Liang H-Y, Fan W-M, Sun W (2011) Mesozoic large magmatic events and mineralization in SE China: oblique subduction of the Pacific plate. *Int Geol Rev* 53(5–6):704–726
- Wang D, Huang F, Wang Y, He H, Li X, Liu X, Sheng J, Liang T (2020a) Regional metallogeny of tungsten-tin-polymetallic deposits in Nanling region, South China. *Ore Geol Rev* 120:103305
- Wang J, Zuo Rg, Xiong Yh (2020b) Mapping mineral prospectivity via semi-supervised random forest. *Nat Resour Res* 29(1):189–202
- Wang C, Tan L, Li Y, Wang M, Ma X, Chen J (2024a) Ontology-driven relational data mapping for constructing a knowledge graph of porphyry copper deposits. *Earth Sci Inf* 17(3):2649–2660
- Wang C, Wang M, Wang B, Chen J, Ma X, Jiang S (2024b) Knowledge graph-infused quantitative mineral resource forecasting. *Earth Sci Front* 31(4):26–36 (in Chinese with English abstract)
- Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. *J Big Data* 3(1):3:9
- Xie X, Mu X, Ren T (1997) Geochemical mapping in China. *J Geochem Explor* 60(1):99–113
- Xu X (2023) Late Triassic to middle Jurassic tectonic evolution of the South China block: geodynamic transition from the paleo-Tethys to the paleo-Pacific regimes. *Earth Sci Rev* 241:104404

- Xu Z, Lou F, Wu Z, Zhang F, Wang D, Chen L, Zhang F (2022) The characteristics of rare metal mineralization, mineralization models, and exploration directions in the Wugong Mountain area of Jiangxi. Paper presented at the 2022 Advances in Geosciences in Jiangxi (in Chinese)
- Yan Q, Xue L, Li Y, Wang R, Wu B, Ding K, Wang J (2023) Mineral prospectivity mapping integrated with geological map knowledge graph and geochemical data: a case study of gold deposits at Raofeng area, Shaanxi Province. *Ore Geol Rev* 161:105651
- Yang J, Jin H, Tang R, Han X, Feng Q, Jiang H, Zhong S, Yin B, Hu X (2024b) Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. *ACM Trans Knowl Discov Data* 18(6):1–32
- Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, Li C, Li C, Liu D, Huang F (2024a) Qwen2 technical report. arXiv preprint [arXiv:2407.10671](https://arxiv.org/abs/2407.10671)
- Yin R, Huang X-L, Wang R-C, Sun X-M, Tang Y, Wang Y, Xu Y-G (2022) Rare-metal enrichment and Nb-Ta fractionation during magmatic-hydrothermal processes in rare-metal granites: evidence from zoned micas from the Yashan pluton, South China. *J Petrol* 63(10):egac093
- Yousefi M, Carranza EJM, Kreuzer OP, Nykänen V, Hronsky JMA, Mihalasky MJ (2021) Data analysis methods for prospectivity modelling as applied to mineral exploration targeting: state-of-the-art and outlook. *J Geochem Explor* 229:106839
- Yu J, Cai Y, Sun T, Jiang W, Zhang R, Griffin WL, Mao Z, Xia L (2023) Distribution and enrichment of rare metal elements in the basement rocks of South China: controls on rare-metal mineralization. *Ore Geol Rev* 163:105797
- Yuan S, Williams-Jones AE, Romer RL, Zhao P, Mao J (2019) Protolith-related thermal controls on the decoupling of Sn and W in Sn-W metallogenic provinces: insights from the Nanling region, China. *Econ Geol* 114(5):1005–1012
- Geomatics Yukon (2021) Regional Geochemical Surveys—RGS - All - 250k. Yukon Geological Survey—Government of Yukon. <https://yukon.maps.arcgis.com/home/item.html?id=5920059e9c1b48bfba26e70e988f6d93>.
- Zaraisky GP, Aksyuk AM, Devyatova VN, Udaratina OV, Chevychelov VY (2009) The Zr/Hf ratio as a fractionation indicator of rare-metal granites. *Petrology* 17(1):25–45
- Zekri H, Cohen DR, Mokhtari AR, Esmaeili A (2019) Geochemical prospectivity mapping through a feature extraction-selection classification scheme. *Nat Resour Res* 28(3):849–865
- Zhang Z, Zuo R, Xiong Y (2016) A comparative study of fuzzy weights of evidence and random forests for mapping mineral prospectivity for skarn-type Fe deposits in the southwestern Fujian metallogenic belt, China. *Sci China Earth Sci* 59(3):556–572
- Zhang L, Chen Z, Li S, Santosh M, Huang G, Tian Z (2017) Isotope geochronology, geochemistry, and mineral chemistry of the U-bearing and barren granites from the Zhuguangshan complex, South China: implications for petrogenesis and uranium mineralization. *Ore Geol Rev* 91:1040–1065
- Zhang L, Chen Z, Li X, Li S, Santosh M, Huang G (2018) Zircon U-Pb geochronology and geochemistry of granites in the Zhuguangshan complex, South China: Implications for uranium mineralization. *Lithos* 308–309:19–33
- Zhang Z, Ning Y, Lu Y, Cao J, Fu J, Zhao Z, Guo J, Ma L, Qin Z, Li J (2021) Geological characteristics and metallogenetic age of Tengshan'ao Sn deposit in Dayishan of South Hunan and its prospecting significance. *Solid Earth Sci* 6(1):37–49
- Zhang X, Liu W, Lentz DR, Wu Z, Wu Y, Zhang X, Yang S (2022) Tin enrichment in a highly fractionated A-type granite: origin and mineralization potential of the Dayishan granite batholith in the Shi-Hang magmatic zone, South China. *Ore Geol Rev* 140:104603
- Zhang C, Liu W, Zhang X (2023a) Knowledge graph construction method of gold mine based on ontology. *J Geo-Inf Sci* 25(7):1269–1281 (in Chinese with English abstract)
- Zhang Z-Y, Hou Z-Q, Lü Q-T, Zhang X-W, Pan X-F, Fan X-K, Zhang Y-Q, Wang C-G, Lü Y-J (2023b) Crustal architectural controls on critical metal ore systems in South China based on Hf isotopic mapping. *Geology* 51(8):738–742
- Zhao K-D, Jiang S-Y, Yang S-Y, Dai B-Z, Lu J-J (2012) Mineral chemistry, trace elements and Sr-Nd-Hf isotope geochemistry and petrogenesis of Cailing and Furong granites and mafic enclaves from the Qitianling batholith in the Shi-Hang zone, South China. *Gondwana Res* 22(1):310–324
- Zhao Z, Fu T-Y, Gan J-W, Liu C, Wang D-H, Sheng J-F, Li W-B, Wang P-A, Yu Z-F, Chen Y-C (2021) A synthesis of mineralization style and regional distribution and a proposed new metallogenic model of Mesozoic W-dominated polymetallic deposits in South China. *Ore Geol Rev* 133:104008

- Zhao Z, Yang X, Lu Y, Zhang Z, Chen S, Sun C, Hou Q, Wang Y, Li S (2022) Geochemistry and boron isotope compositions of tourmalines from the granite-greisen-quartz vein system in Dayishan pluton, Southern China: implications for potential mineralization. *Am Mineral* 107(3):495–508
- Zhou Z-H, Feng J (2019) Deep forest. *Natl Sci Rev* 6(1):74–86
- Zhou C, Wang H, Wang C, Hou Z, Zheng Z, Shen S, Cheng Q, Feng Z, Wang X, Lv H, Fan J, Hu X, Hou M, Zhu Y (2021a) Geoscience knowledge graph in the big data era. *Sci China Earth Sci* 64(7):1105–1114
- Zhou Y, Zhang Q, Huang Y, Yang W, Xiao F, Ji J, Han F, Tang L, Ouyang C, Shen W (2021b) Constructing knowledge graph for the porphyry copper deposit in the Qingzhou-Hangzhou Bay area: Insight into knowledge graph based mineral resource prediction and evaluation. *Earth Sci Front* 28(3):67–75 (in Chinese with English abstract)
- Zhu J, Wang R, Zhang P, Xie C, Zhang W, Zhao K, Xie L, Yang C, Che X, Yu A, Wang L (2009) Zircon U-Pb geochronological framework of Qitianling granite batholith, middle part of Nanling Range, South China. *Sci China Ser D Earth Sci* 52(9):1279–1294
- Zuo R (2020) Geodata science-based mineral prospectivity mapping: a review. *Nat Resour Res* 29(6):3415–3424
- Zuo R, Carranza EJM (2011) Support vector machine: a tool for mapping mineral prospectivity. *Comput Geosci-UK* 37(12):1967–1975
- Zuo R, Xia Q, Zhang D (2013) A comparison study of the C-A and S-A models with singularity analysis to identify geochemical anomalies in covered areas. *Appl Geochem* 33:165–172
- Zuo R, Kreuzer OP, Wang J, Xiong Y, Zhang Z, Wang Z (2021) Uncertainties in GIS-based mineral prospectivity mapping: key types, potential impacts and possible solutions. *Nat Resour Res* 30(5):3059–3079
- Zuo R, Xiong Y, Wang Z, Wang J, Kreuzer OP (2023) A new generation of artificial intelligence algorithms for mineral prospectivity mapping. *Nat Resour Res* 32(5):1859–1869

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.