·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

基于机器学习的石英微量元素研究

王瑀[1]，邱昆峰[1,2*]，侯照亮[3]

1. 中国地质大学（北京）地球科学与资源学院，北京 100083；
2. 中国地质大学地质过程与矿产资源国家重点实验室，北京 100083；
3. 维也纳大学地质系，维也纳 1090

岩石中的石英微量元素记录了石英结晶过程的物理化学环境。通过微量元素研究石英原岩历史已久。传统方法基于石英微量元素分布图解，构建与恢复石英类型。经典图解包括区分矿床类型的石英 Ti-Al 二元图解（Rusk，2012），和判别岩浆岩类型的石英 Ti-Al-Ge 三元图解（Schrön 等，1988）。现阶段研究表明，传统经典图解无法对多种石英类型进行判别分类研究，进而无法准确的恢复石英生成环境。随着石英原位微区测试方法的成熟，高精度石英微量元素数据的逐渐丰富，为运用机器学习深入研究石英微量元素提供了理论与大数据基础。

在本研究中，我们汇编了全球 48 个典型矿床中的 5397 条石英微量元素数据，包含造山型矿床、斑岩型矿床、浅成低温热液矿床、与侵入岩有关的金成矿系统、卡林型矿床、矽卡岩型矿床以及不含矿的花岗岩、伟晶岩和云英岩共九种类型。通过穷举和聚类分析，本研究发现取对数的 Ti/Ge 和 P 作为两坐标轴的二维图解能较为有效地区分不同类型矿床中的石英。基于石英微量元素的典型性与代表性，本研究选取数据集中 Ti、Al、Li、Ge 和 Sr，运用支持向量机，构建了高维度石英微量元素分类器。为确保分类准确性，所用数据集随机分为训练样本（80%数据）和测试样本（20%数据）。通过训练样本得到的新石英微量元素分类器，经测试样本检验计算，石英生成环境预测准确度高达 86%。本研究建立新型石英微量元素分类器，可基于新的石英微量元素数据，准确预测石英类型。本研究同时证明，在地球化学研究中，若以微量元素作为训练特征，特征数量的训练与机器学习模型的准确度呈正相关。随着未来更多高质量的石英元素数据的公开，该分类方法仍有改进空间。本研究展示了运用石英微量元素与机器学习方法，建立相较于传统图解，更为高效准确的石英生长环境判别方法。我们的工作极大地提高了石英类型鉴别的准确度和可信度，为通过石英微量元素数据判别该石英所属母岩是否含矿，以及恢复其所属的矿床和岩石类型提供了理论基础。此分类器可通过此网站使用：https://quartz-classifier.herokuapp.com。

基金项目：国家自然科学基金（41702069、42072087）、高等学校学科创新引智计划（BP0719021）

第一作者简介：王瑀（1997–），博士研究生，矿物学、岩石学、矿床学专业. E-mail：yuwangcugb@qq.com

*通信作者简介：邱昆峰（1986–），教授，博士生导师，从事矿床学教学与科研工作. E-mail：kunfengqiu@qq.com

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

基于极性脂类机器学习分析的黄河—渤海湾土壤和沉积物有机质源汇及保存研究

陶柯宇[1*]，许云平[2]，王映辉[3]，王云涛[4]，何丁[1,4]

1. 浙江省地学大数据与地球深部资源重点实验室，浙江大学地球科学学院，杭州，310007；
2. 上海海洋大学海洋科学学院，上海，201306；
3. 南方科技大学环境科学与工程学院，深圳，518055；
4. 卫星海洋环境动力学国家重点实验室，国家海洋局第二海洋研究所，杭州，310007

有机碳通过河流由陆地向海洋的搬运和转化是全球短周期碳循环的一个重要组成部分。生物标志化合物（生标）研究在精确识别有机碳成因来源方面具有不可替代的优势，并对理解有机碳循环过程和影响因素至关重要。应用更多生源信息可相互佐证的生标分子通常意味着对有机碳来源更严格的约束，但巨大的分子信息量也给数据分析和解释带来了挑战。

本文应用随机森林分类模型（Random Forest classification model）对中国东部黄河中下游—渤海湾土壤以及沉积物中所鉴定的 123 个，包括 6 类（脂肪醇 Fatty alcohols，脂肪酸 Fatty acids，烷基 2-酮 Alkan-2-ones，甾类 Steroids，三萜类 Triterpenoids 和单烷基甘油醚 MAGEs）极性生标分子进行了分析。样品由河流至陆架海区采自三类生境，分别为河流土壤，河流及海相沉积物。基于随机森林模型对不同生境样本的有效区分，评估了极性生标分子的环境特异性，并以此切入确定了生物圈有机碳 4 种主要的成因来源，包括细菌贡献、藻类/浮游动物输入、陆源高等植物输入和人类活动输入。空间上有机碳来源的规律性分布为这一典型的“河流—海洋”系统提供了合理的生物圈有机碳源汇图景。此外，一类硫酸盐还原菌（SRB）示踪物 MAGEs（Pattern I MAGEs）作为随机森林模型中最重要的变量，被高效挖掘并有效地应用为底水氧含量指标。利用 Pattern I MAGEs 与其他极性脂类生标浓度的相关关系来评价影响不同生境下生物圈有机碳富集的主导因素。缺氧保存效应是是控制海洋原位有机碳在表层沉积物中富集的主要影响因素，计算结果显示在渤海表层沉积物中约 37% 的原位极性脂类在不同的氧化还原条件下发生了再矿化。

本文研究结果指示了有机碳埋藏过程中在经过水体中大量再矿化后，在表层沉积物中进一步的变化特征。显示了在这样一个典型的陆架海中，海洋自生有机碳在还原条件下的选择性保存，为了解沉积物中碳循环提供了有意义的见解。此外，Pattern I MAGEs 作为一种潜在的底水氧含量指标值得进一步研究，在海洋生物地球化学循环研究中具有广阔的应用前景。

第一作者简介：陶柯宇（1991–），技术职称：中级。研究方向：地质学，石油地质学，地球化学，有机地球化学，生物有机地球化学，地球科学大数据。电子邮箱：Taoky_gd@163.com

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

A Refined Estimation of Li in Mica by a Machine Learning Method

Lu Wang, Cheng Su, Luo-Qi Wang, J ZhangZhou, Qun-Ke Xia, Qin-Yan Wang*

Key Laboratory of Geoscience Big Data and Deep Resource of Zhejiang Province, School of Earth Sciences, Zhejiang University, Hangzhou, 310027, China

Li-rich micas are crucial in the exploration for and exploitation of Li resources. The determination of Li in mica using classical bulk chemical methods or in-situ microanalytical techniques is expensive and time-consuming and has a high-quality requirement for micas and reference materials. Although simple linear and nonlinear empirical equations have been proposed, they are inconsistent with the complex physico-chemical mechanisms of Li incorporation and commonly lead to large errors. In this study, we introduce a refined method of multivariate polynomial regression using a machine learning algorithm to estimate Li from multiple major oxide abundances. The performance of our regression model is evaluated using the coefficient of determination (R^2) and the root-mean-square error (RMSE) of the independent test sets. The best-performed models show R^2 of 0.95 and a RMSE of 0.35 wt% for the test set of dataset 1 (all compiled data, n = 2124) and R^2 of 0.96 and a RMSE of 0.22 wt% for the test set of dataset 2 (only data obtained using in-situ techniques, n = 1386). Our results indicate that integration of electron probe microanalysis and multivariate polynomial regression (based on dataset 1) presents a robust and convenient approach to quantify Li in micas. The application of the proposed approach to micas from central Inner Mongolia, NE China, suggests that in addition to the Weilasituo ore bodies, the Jiabusi granite and greisen and the Shihuiyao metamorphic sediment formation have good potential for Li exploration. Our study also provides preliminary constraints on the genesis of Li deposits.

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

用机器学习方法研究长白山火山活动的动力学过程
Dynamic Evolution of Changbaishan Volcanism in Northeast China Illuminated by Machine Learning

Yong Zhao[1, 2], Yigang Zhang[3], Dongdong Ni[1, 2]

1. State Key Laboratory of Lunar and Planetary Sciences, Macau University of Science and Technology, Macau, PR China;
2. CNSA Macau Center for Space Exploration and Science, Macau, PR China;
3. Key Laboratory of Computational Geodynamics, College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, 100049 Beijing, China

中国东北地区广泛分布板内玄武岩，其中长白山玄武岩位于日本海沟以西约 1300 公里处，是中国东北部最大的岩浆活动中心。前人的研究结果普遍认为，长白山新生代玄武岩是地幔上升流形成的，然而，其主要的驱动来源和机制仍存在很大争议。

本研究利用 GEOROC 数据库中全球洋岛(IAB)和岛弧玄武岩(OIB) 的主微量元素，训练了可以用于预测俯冲流体影响程度的机器学习模型。随机森林和主成分分析方法提取了 IAB 和 OIB 的主要地球化学特征。如图 1a 所示，Nb, Ta, TiO_2, K_2O, Ba, Sr 在区分 IAB 和 OIB 是非常重要的元素。对于 IAB，K_2O、Ba 和 Sr 具有较大的正主成分载荷值，Nb、Ta 和 TiO2 具有较大的负主成分载荷值(图 1b)。OIB 具有与 IAB 相反的特点。K_2O、Ba 和 Sr 在俯冲过程中流体活动时非常活跃（fluid mobile）的元素，易于进入流体相；Nb、Ta 和 TiO_2 在俯冲过程中是不活跃元素，不易进入流体相（Pearce 和 Peate，1995；McCulloch 和 Perfit，1981）。显然，IAB 和 OIB 的主要特征是：IAB 由于俯冲流体的加入而富集活跃元素，OIB 由于俯冲板片释放流体后残留物质的加入而富集不活跃元素。因此，用 IAB 和 OIB 训练的机器学习模型可以告诉本文玄武岩样本是如何受到俯冲过程影响的，特别是这些样本是否有俯冲流体组分的加入，强弱是多少。主成分分析得到了与随机森林相似的结果（图 1c）。

将长白山玄武岩送入到训练好的随机森林和深度神经网络模型中进行预测，结果显示（图 2a）：1、随机森林与深度神经网络预测结果基本一致，这说明不同的机器学习方法从获取了岩石数据中类似的信息，给出了相似的预测结果。2、长白山玄武岩具有很高的 SFE 指数，说明很大程度上受到了滞留板片流体作用的影响。3、从 5 Ma 开始，流体作用逐渐减弱，1 Ma 之后又有所增强。结合图 2b、c，可以看

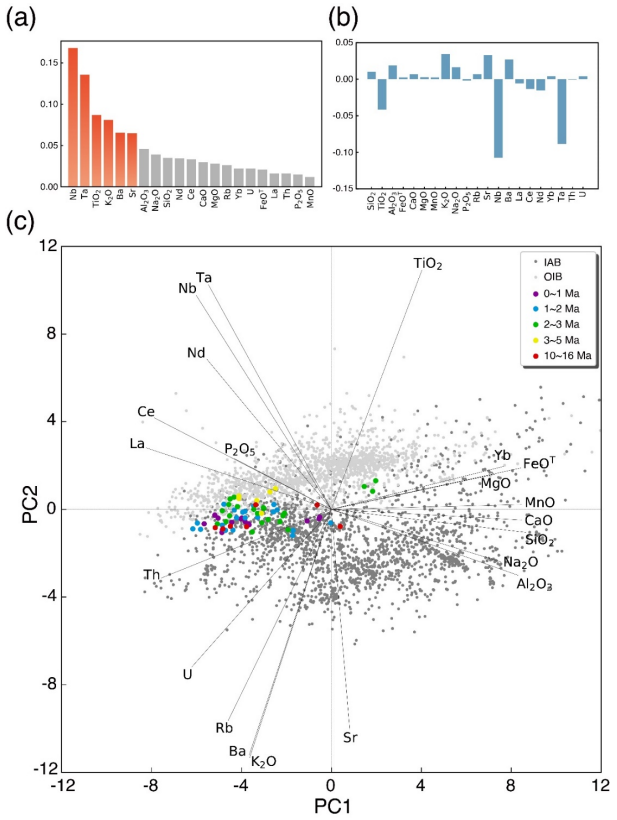

图 1 (a) 随机森林提取的重要特征。(b)洋岛玄武岩在随机森林分类模型中各元素的重要性 (c) 长白山玄武岩、洋岛和岛弧玄武岩主成分分析成分双标图。黑色线表示每个元素的主成分特征值，在主成分坐标轴上的投影长度表示对相应主成分的影响大小。

出 5-1 Ma 期间流体作用的减弱伴随着$(^{87}Sr/^{86}Sr)_0$ 和$\varepsilon_{Nd(t)}$的微小变化，而 1 Ma 之后，流体作用的增强对应着$(^{87}Sr/^{86}Sr)_0$ 的下降和 $\varepsilon_{Nd(t)}$的升高。

本研究认为，由于太平洋板片后撤，在大约 5 Ma 时，使得长白山下面滞留板片中出现撕裂（图 3b），板片下方含水的大洋软流圈物质在 1 Ma 时从板片裂隙处上涌，使得长白山玄武岩富含流体，并且相对亏

损（图 3c）。

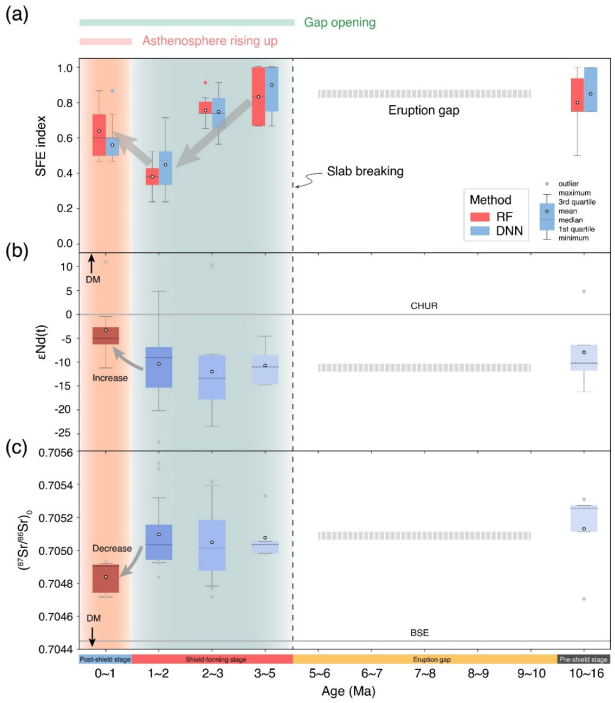

图 2 (a) 长白山玄武岩的随机森林和深度神经网络的预测结果。(b、c) 长白山玄武岩 Sr、Nd 同位素值随时间的变化情况。

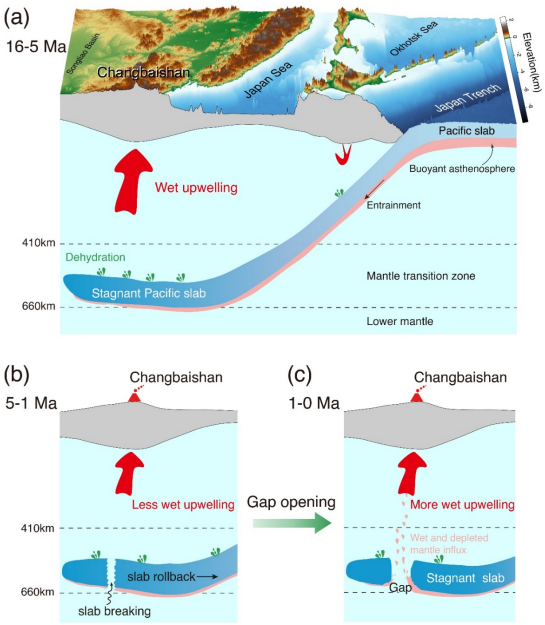

图 3. 长白山火山形成的动力学模型示意图。(a) 16-5 Ma。太平洋板片携带大洋软流圈穿过地幔滞留在长白山下方的地幔过渡带中。滞留板片脱水诱发的上升流形成了长白山玄武岩。(b) 5-1 Ma。滞留板片局部撕裂，致使板片局部缺失，从而使得流体作用减弱。(c) 1Ma 至今。滞留板片下方软流圈物质通过裂隙上升，使得长白山玄武岩更加富流体，相对亏损。

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

蛇纹石多型的机器学习分类器及其对俯冲带元素迁移的意义

黄放*

CSIRO Mineral Resources, 26 Dick Perry Ave, Kensington, WA 6151, Australia

蛇纹石分为利蛇纹石、纤蛇纹石和叶蛇纹石三种多型，广泛分布于地幔，洋壳以及俯冲带内。已有研究表明，不同蛇纹石多型形成于不同化学环境中，表明不同蛇纹石多型各具有独特的化学特征。因此，对蛇纹石多型的鉴定可反演蛇纹石形成的化学环境，明确蛇纹石多型相互转变过程中元素的迁移机制。在传统方法中，鉴定蛇纹石多型需结合其结构信息，尚无从化学成分角度出发提出的分类标准。近年来，机器学习方法的应用为矿物化学成分的分类提供新思路和新手段。在本研究中，我们采用了 XGBoost 算法对包含 1603 条蛇纹石主量成分数据集进行了处理，总计训练四个分类器，包括三种蛇纹石多型、利蛇纹石和叶蛇纹石、纤蛇纹石和叶蛇纹石，以及利蛇纹石和纤蛇纹石的化学成分分类模型。训练结果显示这些模型的准确度分别为 82.5%、87.7%、92.8%和 75.4%，通过计算特征重要性，发现叶蛇纹石具有高 SiO_2 含量的化学特征。利蛇纹石和纤蛇纹石可由 Cr_2O_3、TiO_2 和 NiO 含量进行区分。此外，结合三分类的混淆矩阵和地球化学数据投图的结果显示，约有 50%的纤蛇纹石被错误的分类，这表明纤蛇纹石可转变为利蛇纹石和叶蛇纹石。结合特征重要性结果，在俯冲带浅层部位，当纤蛇纹石转变为利蛇纹石时，环境中 Cr、Ti 和 Ni 迁入利蛇纹石；在俯冲带深部，当纤蛇纹石向叶蛇纹石转变时，Si 元素富集于叶蛇纹石内。以上结果说明，XGBoost 模型可有效应用于基于化学成分对矿物多型进行识别和分类；同时，基于 XGBoost 算法对特征重要性的计算，可探究解释矿物转变过程中元素的迁移机制。

作者简介：黄放（1989–），研究员，研究方向：地球化学数据驱动型研究. E-mail: f.huang@csiro.au

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

制约地幔交代作用的全球分布：来自针对地幔单斜辉石化学成分的机器学习研究

秦奔[1]，黄放[2]，黄士春[3]，陈云枫[1]，张舟[1*]

1. 浙江大学地球科学学院，杭州 310027;
2. CSIRO Mineral Resources, Kensington, WA 6151, Australia;
3. Department of Geoscience, University of Nevada Las Vegas

地幔交代作用影响着岩石圈的化学分异、克拉通稳定性和地幔的地球物理特性，是了解地球内部成分演化和动力学过程的一个重要考虑因素。因此，评估全球尺度的地幔交代作用对于理解地幔的不均一性具有重要意义。由于只有少数样品保留了地幔显性交代的矿物，前人提出了不同的地球化学元素比值作为指标来指示交代作用。尽管前人提出的地球化学指标对于局部地区的样品是否经历了地幔交代作用具有较好的判断，但是将前人提出的元素比值应用于全球样品，我们发现通过元素比值判断交代作用的结果和显性交代矿物的证据并不自洽。此外，不同化学指标判断的结果也不一致。因此，我们试图从数据科学角度出发，训练出和显性交代矿物证据自洽，并在化学成分上可区分交代作用的机器学习模型。为此，我们从 GEOROC 数据库中下载并清洗了原始数据，得到单斜辉石主量元素的有效数据约 22000 条，微量元素数据约 3000 条。我们以岩相学出现显性交代矿物和微量元素轻稀土元素含量单调递增分别作为发生交代作用的正标签和负标签，获得了监督学习的训练样本（主量元素数据 2089 条，微量元素数据 872 条）。通过 Xgboost 模型训练这些监督学习的样本，分类准确率可以达到 95 %，并且模型显示所有元素对于其判断均有一定的重要性。随后，利用无监督学习 K 均值模型证明有标签的数据具有全部数据的分簇特征，从而保障训练的模型可以预测无标签的数据（主量元素数据 19516 条，微量元素数据 2094 条）。将训练获得的机器学习模型应用于全球样品进行预测，结果显示大多数区域的单斜辉石均受到了交代作用影响，而个例结果则呈现交代概率分布的不均一性。此外，研究发现交代作用和地震观测数据尚无明显相关性。

第一作者简介：秦奔（1998–），博士研究生，研究方向：地球化学数据驱动型研究. E-mail：charlesbenq@zju.edu.cn
*通信作者简介：张舟（1986–），百人计划研究员，研究方向：地球化学数据驱动型研究. E-mail：zhangzhou333@zju.edu.cn

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

不同克拉通中的金刚石有何异同？—来自地球化学数据统计和机器学习的研究

雷佳莉[1]，黄放[2]，张舟[1*]

1. 浙江大学地球科学学院，杭州 310027；
2. CSIRO Mineral Resources, Kensington, WA 6151, Australia

金刚石及其包裹体是认识岩石圈演化和深部碳循环的重要载体。目前研究认为，金刚石有多种成因，可以在不同的流体/熔体碳过饱和的条件下形成（Stachel et al.，2008；Shirey et al.，2013）。同时，不同克拉通的金刚石及其包裹体成分可能存在统计学上的差异，而宝石市场上批量交易的金刚石来自多个未知产地的混合（Shor et al.，2010）。因此，研究不同克拉通的金刚石及其包裹体地球化学特征的异同，不仅可以为探求不同克拉通金刚石的形成机制提供制约，也可以为通过地球化学分析有效判别金刚石产地提供依据。在本研究中，我们搜集和整理了前人发表的 172 篇天然金刚石文献，整理并清洗有效数据约 10000 条，包括金刚石的碳同位素数据、金刚石的氮元素含量及其同位素比值、金刚石中包裹体的种类以及包裹体的化学成分。在前人已发表的数据中，Kaapvaal、Slave 和 Siberia 三个克拉通中的金刚石数据量占据前三位。统计分析这三个克拉通中金刚石的氮含量（ppmw）、$\delta^{15}N$ 和 $\delta^{13}C$ 发现：Slave 克拉通中金刚石氮含量变化区间较大（0~2000 ppmw），Kaapvaal 和 Siberia 克拉通中金刚石氮含量在 0~1000 ppmw 之间；P 型（橄榄岩型）金刚石氮含量多集中于 300 ppmw 以下，E 型（榴辉岩型）金刚石氮含量多在 300~800 ppmw 之间。金刚石 $\delta^{15}N$ 值显示 Siberia 和 Slave 克拉通中金刚石主要在-5‰至-3‰间。 Kaapvaal 和 Siberia 两个克拉通中的金刚石 $\delta^{13}C$ 值相似，更接近地幔平均值-5‰。使用 XGBoost 和 LightGBM 的机器学习手段，依据包裹体的主微量元素对克拉通来源进行预测，结果显示两种模型的准确度都达到了 0.94。因此，通过包裹体种类、金刚石碳同位素组成、金刚石氮元素含量等参数综合判断不同克拉通的金刚石及其包裹体地球化学成分的异同，可以设计出不同克拉通间金刚石相似度指数，用于区分金刚石产自的克拉通。

第一作者简介：雷佳莉（1994–），博士研究生，研究方向：地球化学数据驱动型研究. E-mail: leijiali@zju.edu.cn
*通信作者简介：张舟（1986–），百人计划研究员，研究方向：地球化学数据驱动型研究. E-mail: zhangzhou333@zju.edu.cn

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

QGrain：开源易用的沉积物粒度综合分析软件

刘宇明[1,2*]，刘星星[1]，孙有斌[1]

1. 中国科学院地球环境研究所；
2. 中国科学院大学

摘要：在过去的一百年里，沉积物的粒度分析方法被广泛地研究，许多有效的方法已经被发表（Blott and Pye，2001；Folk，1966；Hateren et al.，2018；Vandenberghe，2013；Weltje and Prins，2003）。在最近的研究中，对沉积物粒度的分解是一个热点。单样本分解（曲线拟合）是一种经典的分解方法，已经被广泛地运用于多种沉积物的粒度分析中（Qin et al.，2005；Sun et al.，2011；Chen et al.，2013；Liu et al.，2016；Xiao et al.，2015）。但是另一种分解方法，即端元分析，却认为它不具有普遍意义（Weltje and Prins，2007）。端元分析已经逐渐成为粒度分解的主流方法，并衍生出了许多的变种算法。最近的一些研究已经注意到，端元分析也存在一定的局限性，而且不同的端元分析算法可能会给出不一样的结果（van Hateren et al.，2018）。然而，还没有研究对这两种方法的异同及其应用范围进行详细的讨论。本研究基于这两种方法的数学模型和一系列的实验，结合数值优化和信息论，对他们的特性进行了详细的讨论。之后，我们客观总结了两种方法的适用范围。尽管单样本分解方法的拟合难度更大，但是我们认为该方法具有更大的潜力。我们同样强调综合分析的重要性，一些传统方法（主成分和聚类分析）同样可以发挥重要的作用，让我们对粒度数据的认识更为全面。最后，我们介绍一个开源易用的粒度综合分析软件，QGrain。QGrain 不仅集成了许多传统的分析工具，而且提供了新的端元分析和单样品分解算法。与其他端元分析算法相比，新方法具有杰出的表现。此外，我们对单样本分解方法做了大量的扩展以提高其稳定性，使其可以与端元分析算法进行竞争或合作。

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

太古代陆壳火成岩地球化学数据库建设

刘纯韬，张舟*

浙江大学地球科学学院，杭州 310000

随着火成岩地球化学数据的不断积累，对其开展数据挖掘成为揭示大陆地壳演化的新方向。前人通过整合主流地球化学数据平台的数据（如：GEOROC，EarthChem 等），建立了包含约 70000 件火成岩样品的地球化学数据集。依据该数据集，Keller 和 Schoene（2012）计算得到 38 亿年以来陆壳镁铁质火成岩的平均化学成分（如：K_2O，Na_2O，Ni 和 MgO）和元素比值（如：La/Yb）演化的图谱。研究发现，这些化学成分和比值均在 25 亿年左右发生突变。结合地幔温度和熔融比例变化的关系，该研究从火成岩地球化学大数据的视角提供了地球逐渐冷却的证据（Keller 和 Schoene, 2012）。但我们注意到，在太古代时期，镁铁质火成岩不同的化学成分（如：MgO，Cr 和 Ni 等）在 25 至 38 亿年之间均呈现出近似正态分布的特点。鉴于该作者在计算过程中假设火成岩样品的年龄误差呈正态分布，我们推测特定的年龄误差导致了这种变化。而这表明该地球化学数据集中年龄数据的评估不够准确，亟待修正。

高质量的地球化学数据是进行大数据研究的前提条件，本研究旨在完善陆壳火成岩地球化学数据库。我们系统检查了 EarthChem 数据库中太古代火成岩样品的年龄数据（约 12000 件样品），发现其中约 65%样品的平均年龄为 31.75 亿年，其年龄误差为±6.75 亿年，年龄区间覆盖整个太古代。因此，该年龄数据与岩石的真实年龄可能存在较大偏差。这可能直接导致了计算得到的地球化学成分在 25 至 38 亿年之间呈现出正态分布特征，并影响我们对早期地球陆壳演化过程的理解。本研究依据发表数据的原始文献，系统地检查了太古代火成岩样品的年龄和地球化学数据，并就输入数据的年龄错误进行了修正。基于对已完成检查样品的地球化学数据（约 3600 件样品），经过 10000 次蒙特卡洛抽样计算得到了太古代镁铁质火成岩的平均成分。修正年龄后显示的化学成分演化图谱（MgO，Cr 和 Ni 等）与前人结果不同，并没有呈现出正态分布的演化特征。

第一作者简介：刘纯韬（1994–），博士研究生，研究方向：地球化学数据驱动型研究. E-mail：chuntliu@zju.edu.cn

*通信作者简介：张舟（1986–），百人计划研究员，研究方向：地球化学数据驱动型研究. E-mail: zhangzhou333@zju.edu.cn

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

Machine Learning for Source Identification of Dust on the Chinese Loess Plateau

Xin Lin[1*], Hong Chang[2], Kaibo Wang[2]

1. School of Earth Sciences and Resources, Chang'an University，Xi'an 710054;
2. Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an, 710061

Central northern China is occupied by a unique landscape called the Loess Plateau which is formed by accumulation of extensive eolian dust (640,000 km^2 in area and 105 km^3 in volume). However, the provenance of voluminous eolian dust on the Chinese Loess Plateau (CLP) is still highly debated. We apply machine learning methods of support vector machine and convolutional neural network to train models using element compositions of surface sediments from eight potential source regions, accordingly, to determine the dust sources and contributions by classifying the last glacial loess and present interglacial sediments on the CLP. The trained models succeed in differentiating major secondary sources and quantitatively estimating the contributions of both primary and secondary sources at least during the last glacial-interglacial cycle. Our results indicate that the contributions of major secondary sources are dominated (approximately 50%) by the recycled Yellow River sediments from the Hetao Graben, followed by materials from inland basins in the northwest of China (>15% for eastern Tibetan Plateau, approximately 15% for Qaidam Basin, and >5% for Tarim Basin, Junggar Basin, and Alxa Plateau). Two primary sources, that is, northern Tibetan Plateau and Central Asian Orogenic Belt, thus contribute about 80% and 20%, respectively, to the dust deposition. The understanding that a constant dust source despite changing climate conditions agrees with those derived from Sr-Nd isotopes and U-Pb age spectra. Our observations demonstrate that big geochemical data sets coupled with machine learning technology are fully capable of tracing sources.

第一作者简介：林鑫（1987–），副教授，研究方向：应用地球化学、地学大数据与机器学习. E-mail：xinlin@chd.edu.cn

*通信作者简介：林鑫，长安大学地球科学与资源学院副教授，加拿大 Laurentian University 访问学者，长期从事应用地球化学领域的研究，担任 Journal of Geochemical Exploration 副主编。目前以第一或通讯作者发表论文 14 篇，代表性成果发表于 Geophysical Research Letters, Journal of Geochemical Exploration, Applied Geochemistry, Environmental Science & Policy 和《地球科学进展》等国内外知名地学期刊。

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

基于机器学习模型估算单斜辉石三价铁含量

黄伟桦，张舟*

浙江大学地球科学学院，杭州 310027

单斜辉石中三价铁的含量对于地幔氧逸度的估算，地幔含水性的评估以及石榴子石-单斜辉石地质温度计的准确应用有重要作用。准确获取单斜辉石中三价铁含量存在测试难度大、成本高、需要样品量较大等问题。然而，基于电价平衡计算单斜辉石中三价铁含量存在较大的误差。因此，亟需一种方便，准确且低成本的方式获取单斜辉石中的三价铁。为此，我们收集了 407 组已知主量元素和 $Fe^{3+}/\Sigma Fe$ 的单斜辉石数据。其中，$Fe^{3+}/\Sigma Fe$ 由穆斯堡尔谱测试获得。我们以单斜辉石的主量元素为输入，$Fe^{3+}/\Sigma Fe$ 为输出，采用了 7 种机器学习的方法（线性回归，多项式回归，决策树，人工神经网络，集成学习神经网络，随机森林和极端随机树）训练了用主量元素估算 $Fe^{3+}/\Sigma Fe$ 的模型。相较于传统基于电价平衡的算法，机器模型有着更高的准确度（从±1 提升为±0.2）和更低的均方根误差（从 0.32 降低为 0.05），是一种便捷可靠的计算单斜辉石中三价铁含量的方法。将模型应用于石榴子石-单斜辉石地质温度计，我们发现忽视单斜辉石中的三价铁或者使用电价平衡法计算三价铁含量均会导致结果产生较大误差（±250℃）。与之相对的是，基于机器学习模型计算的单斜辉石三价铁含量可以将温度误差降低到 50℃以内，显著改善了该地质温度计的应用。此外，我们将模型应用于玄武岩中经历了氢扩散和未经历氢扩散的单斜辉石，发现经历氢扩散的单斜辉石相对于未经历氢扩散的单斜辉石有着较高的 $Fe^{3+}/\Sigma Fe$。该结果支持了单斜辉石中氢扩散机制为 Fe^{2+} 的氧化这一观点。

基金项目：中央高校基本科研业务费专项资金

第一作者简介：黄伟桦（1996–），博士研究生，研究方向：地球化学数据驱动型研究. E-mail：21938008@zju.edu.cn

*通信作者简介：张舟（1986–），百人计划研究员，研究方向：地球化学数据驱动型研究. E-mail：zhangzhou333@zju.edu.cn

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

可解释机器学习揭示地球深部过程机理：单斜辉石中的氢扩散

李岸洲*，陈欢，吴森森，夏群科，杜震洪*

浙江大学地球科学学院，杭州 310027

地球深部的水影响了矿物和岩石的物理化学性质，对地幔演化与全球水循环具有重要影响。幔源岩浆是认识地球内部的水含量的一个重要窗口。其中，从冷却的玄武岩中结晶的单斜辉石是计算其地幔源区中水含量的关键矿物。通过单斜辉石计算地幔源区的水含量需要考虑的重要因素是单斜辉石中的氢元素是否发生了扩散。传统方法来判断氢扩散仅考虑了个别元素之间的相关关系，然而高温高压的实验与理论已经表明存在多种单斜辉石中氢的扩散机制。近年来，已经有研究运用支持向量机（Support Vector Machines）机器学习的方法综合主量元素判断单斜辉石斑晶中的氢是否扩散。但是，支持向量机模型的解释性较差，无法为影响氢元素扩散的元素提供更多的指示意义。

本研究在支持向量机的研究基础上，通过极致梯度提升（XGBoost）的机器学习方法实现了精准建模。在此基础上，运用树型模型本身的强可解释性以及可解释机器学习模型，按照样本、特征、整体的顺序定量揭示了各元素对氢扩散的贡献程度。从数据与机理的角度总结出 Na 对单斜辉石中的氢元素扩散具有较强控制，并定量描述了各元素在不同水含量下对扩散过程的影响程度。

第一作者和通信作者简介：李岸洲（1998–），硕士研究生，研究方向：地球科学大数据. E-mail: anzhouli@zju.edu.cn
通信作者简介：杜震洪（1981–），教授，研究方向：大数据与地球-海洋系统. E-mail: duzhenhong@zju.edu.cn

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

深度卷积神经网络在地质图像识别中的应用及线上部署

刘小康，宋海军*，楚道亮，侯杰，佘振兵

生物地质与环境地质国家重点实验室，中国地质大学（武汉），武汉 430074

传统地质学领域的诸多研究多基于实体材料开展，期间会产生大量图像数据，并基于此开展进一步研究。以化石的鉴定为例，其系统分类学在古生物学研究中起到了举足轻重的地位，被广泛应用于生物演化、古生态重建、地层学厘定等诸多方面。传统的系统分类学研究往往需要大量的先验知识，并在特定领域有长时间的耕耘，而不同学者的知识储备和经验存在差异，这为鉴定结果增添了不确定性。这个过程也同样存在于岩石、岩相分析等地质图像识别领域。因此，如何使用人工智能手段解决诸多繁琐、人员密集型的常规识别工作成为当下的热点问题之一。

我们选择对最常见的地质图像进行自动识别，现已实现三个模块，包括微相模块、岩石模块、大化石模块。微相模块是对碳酸盐岩微相分析中的生物、非生物颗粒进行识别。从 1133 篇文献和本团队材料中搜集了 22 种最常见颗粒组成的 3 万余张照片（其中三分之二以上来自发表资料），用来实现四个经典的深度卷积神经网络训练和预测，最终使用迁移学习的方法在 Inception ResNet v2 网络中获得了平均 95%的鉴定准确率。对于岩石模块，我们通过对博物馆馆藏标本、室内教学实习标本和野外经典地质路线中的野外露头岩石进行拍照，共收集 48 种常见岩石 2.2 万余张图片。最终实现了 84%的识别准确率。而对于古生物化石识别，我们使用网络爬虫技术从开放网络中搜集了一个包含 41.5 万张图片的化石图像数据集（Fossil Image Dataset，FID），通过人工数据清洗方式得到了包括无脊椎动物、脊椎动物、植物、微体化石和遗迹化石在内的 50 个类别，最终在 Inception ResNet v2 网络中获得了平均 90%的鉴定准确率。当下硬件条件已不再是制约深度学习研究发展的主要障碍，而如何搜集整理更大量的数据和优化模型算法是推进其在地质领域应用中所需要突破的环节。

此外，当前深度学习在地质学的应用大多集中在线下研究，鲜有研究将训练成熟的人工智能识别模型提供到开放平台以形成端到端服务。为此，我们建立了智能化石识别平台（网页版：www.ai-fossil.com；微信小程序：ai 化石），我们已完成对上述三个模块的线上部署，可供同行研究和教学使用，同时搭建的数据库模块可以实现化石图片的上传和下载，以作后续研究。

基金项目：国家自然科学基金项目（41821001）；中国科学院战略性先导研究项目（XDB26000000）

第一作者简介：刘小康（1994–），博士研究生，从事地学大数据与深度学习研究. E-mail：xkliu@cug.edu.cn

*通信作者简介：宋海军（1983–），教授，博士，从事地学大数据与地球生物学研究. E-mail：haijunsong@cug.edu.cn

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

通过地球化学指标计算地壳厚度：基于全球岛弧岩浆和青藏高原火成岩的机器学习研究

栾志康，刘佳*，张舟

浙江大学地球科学学院，杭州 310027

岛弧岩浆的全岩地球化学指标与地壳厚度存在一定的相关性。前人研究通过统计分析手段提出了 Sr/Y 和 La/Yb 等地球化学指标与地壳厚度之间的经验公式。这些经验公式被应用在地质历史时期地壳厚度演变，以及造山带中地壳厚度演变的研究之中。但是，利用地球化学指标传统经验公式计算地壳厚度存在着较大误差。此外，不同的地球化学指标计算得出结果之间也存在着不自洽，在造山带地壳厚度的计算中误差尤为明显。相较于传统经验公式，机器学习的手段具有结合多组指标并拟合非线性关系的特性，可以更有效率地利用地球化学元素特征反演地壳厚度。

通过收集前人已经发表的全球岛弧的地球化学数据，我们利用监督学习算法，对全球岛弧地区全岩主量和微量数据进行机器学习模型训练。研究发现，机器学习模型可以显著提高利用地球化学指标计算得到地壳厚度的精度。随后，我们将训练获得的模型应用于青藏高原的火成岩地球化学数据，反演了青藏高原地壳厚度变化的历史。

第一作者简介：栾志康（1998–），博士研究生，研究方向：地幔地球化学. E-mail：zinkluan@zju.edu.cn

*通信作者简介：刘佳（1985–），研究员，研究方向：地幔地球化学. E-mail：liujia85@zju.edu.cn

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

通过机器学习算法进行磷灰石含矿性判别

郑育宇，许博*

中国地质大学（北京）地质过程与矿产资源国家重点实验室，北京 100083

磷灰石作为岩浆岩常见矿物，其微量元素特征可以记录生长环境，并且可用于含矿性和矿床类型的判别。传统判别手段一般使用二元或三元的二维投图以及部分特征元素的含量。这些方法由于样本数量以及图像维度的局限表现出较差的精确度和普适性。本研究利用机器学习的方法分析磷灰石的微量元素数据，以判别矿床的含矿性。本次研究一共使用了来自 55 个矿化矿床的 674 个磷灰石微量元素数据以及 2 个未矿化矿床的 248 个磷灰石数据进行机器学习，数据维度包括 46 种元素含量以及 10 个常用地球化学参数。本研究主要使用 XGBoost 机器学习算法，并使用网格搜索寻找最优参数。研究结果表明机器学习方法在磷灰石含矿性判别上表现出极高的精度和召回率，分类模型中可能对含矿性判别结果影响较强元素为：Na、(Th/U)、Pb、Ba、V、Mn、La、F、Th 和(Eu/Eu*)。本工作完善并扩展了机器学习在矿物中的应用，识别出磷灰石微量元素对矿床的研究具有相对可靠的应用前景，但具体工作还需要继续挖掘。

第一作者简介：郑育宇（1996–），博士研究生，研究方向：成因矿物学. E-mail：zhengyuyu2019@163.com
*通信作者简介：许博（1988–），副教授，研究方向：矿物学、岩石学、矿床学. E-mail：bo.xu@cugb.edu.cn

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

基于大数据的磷灰石微量元素研究

周统[1]，邱昆峰[1,2*]，于皓丞[1]

1. 中国地质大学（北京）地球科学与资源学院，北京 100083；
2. 中国地质大学地质过程与矿产资源国家重点实验室，北京 100083

磷灰石广泛分布于火成岩、沉积岩和变质岩中，其晶格容纳丰富的微量元素，能够记录形成的物理化学环境，因此，碎屑磷灰石是指示沉积物来源，约束构造演化的理想矿物。磷灰石的微量元素判别图解是进行物源判别的常用手段，经典判别图解包括 Sr-Y、Sr-Mn、Y-（Eu/Eu*）和（Ce/Yb）$_{cn}$-REE 图解。近年来，微区测试技术日益成熟，地球化学数据大量积累，传统分析方法逐渐无法有效利用这些数据所携带的信息，上述图解已经无法准确评估磷灰石母岩类型。建立能准确判别碎屑磷灰石物源的新型判别图解故而迫切。数据的丰富和大数据技术的发展，为以大数据为依托的分析方法应用到磷灰石物源判别研究中奠定了数据基础。

本次研究将大数据技术与地球化学数据相结合，共收集整理了全球 1925 个代表性磷灰石微量元素数据，对富碱性火成岩、超镁铁质岩石、镁铁质火成岩、长英质花岗岩、中-低级变质岩、高级变质岩六种母岩类型中磷灰石微量元素数据进行穷举端元处理，共获得 7140 个磷灰石物源判别图解端元组合，在轮廓系数限定下，进一步有效筛选并提取出能判别磷灰石物源类型的最优图解端元。基于此，我们构建了 Eu/Y-Ce 磷灰石判别新图解。Eu/Y-Ce 图解的数据分布显示，磷灰石形成时的氧化还原状态以及微量元素的不同配分行为可能会影响碎屑磷灰石物源判别效果。相较于之前的磷灰石判别图解，Eu/Y-Ce 图解涵盖了更全面的物源类型，可以更准确地进行物源判别。

本研究对磷灰石微量元素的数据挖掘工作，是将大数据技术运用在地球科学研究中初步探索。随着未来磷灰石地球化学数据的更加丰富，结合更多算法，高纬度元素判别图解的建立值得进一步探索。

基金项目：国家自然科学基金（41702069、42072087）、高等学校学科创新引智计划（BP0719021）

第一作者简介：周统（1996–），硕士研究生，矿物学、岩石学、矿床学专业. E-mail：zhoutong_1996@163.com

*通信作者简介：邱昆峰（1986–），教授，博士生导师，从事矿床学教学与科研工作. E-mail：kunfengqiu@qq.com

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

机器学习方法对岩浆-热液系统高钛磁铁矿的成因判别

曾丽平[1*]，胡斌[1]，赵新福[1]

1. 中国地质大学（武汉）资源学院，地质过程与矿产资源国家重点实验室，武汉 430074

磁铁矿因其反尖晶石结构可以容纳多种微量元素从而记录其成因信息，据此前人制作了多个判别图解用以判别不同矿床的成因类型。但磁铁矿复杂的形成过程或遭受叠加改造作用（如溶解-再沉淀作用或重结晶作用）可能会改变磁铁矿中微量元素的原始组成，导致无法正确提取关键信息，使得现有磁铁矿微量元素成分判别图解在实际使用时存在很多问题。例如，磁铁矿的 Ti 含量通常被当作判断岩浆磁铁矿的最重要的依据。但近年来在一些高温热液矿床（如斑岩铜矿、IOCG 矿床）中也发现了具有钛铁矿出溶结构的高钛磁铁矿（Ti >1 wt.%），这些高钛磁铁矿与热液蚀变矿物紧密共生，表明其为热液成因，使得这些已有图解对高钛磁铁矿的成因判别基本失效。这一问题在在玢岩铁矿（又称铁氧化物-磷灰石矿床或 IOA 型铁矿）的研究尤为突出，该矿床长期存在岩浆成因和热液交代成因两种截然不同的模型。因此，确定磁铁矿中的微量元素之间的关系及成因联系，对于正确提取关键元素，优化已知的判别图解，解决玢岩型铁矿床的成因问题均具有重要意义。

通过大数据统计分析和机器学习技术，本研究选取了来自全球高温成矿系统的 876 个、共计 59 种元素的原生高钛磁铁矿的 LA-ICP-MS 微量数据；对数据集进行了两种无监督方法（主成分分析 PCA 和 t 分布随机邻域嵌入 t-SNE）的机器学习分析；建立了三个不同元素组合的机器学习模型来识别不同成因的磁铁矿。对已建立的模型通过支持向量机分类器（SVM）以及接受者操作特性曲线（ROC）进行进一步的评估其分类效能，证明模型能够描述并区分不同成因高钛磁铁矿的微量元素组成特征。我们的模型表明，Mg、Mn、Al、Ti、V、Co 等元素可以明确区分不同成因的高钛磁铁矿。

在上述研究基础上，我们进一步的将降维后结果转换成一般形式的元素组合，并提出新的 lg(Al)+lg(Ti)+lg(V) vs lg(Mn)/[lg(Co)+lg(Mg)]判别图解，用以区分岩浆和热液成因的高钛磁铁矿。新的判别图解可以避免原有判别图解带来的问题，在指示磁铁矿成因类型等方面提供更加丰富和准确地解释。我们的研究结果还表明，IOA 矿床中发育的高钛磁铁矿成分无论从含量或者微量元素之间的相关性上和高温岩浆热液成矿系统（包括 IOCG 和斑岩型矿床）中的高钛磁铁矿具有十分相似的特征，而与典型的岩浆成因磁铁矿间却存在不小的差异。因此，基于高钛磁铁矿微量元素组成，从机器学习分析角度说明玢岩铁矿床为热液成因。

基金项目：矿床学（2019-2021），国家自然科学基金优秀青年项目（批准号：41822203)

第一作者简介：曾丽平（1990–），助理研究员，研究方向：玢岩铁矿床成因及矿物学研究. E-mail：liping_zeng@cug.edu.cn

*通信作者简介：曾丽平（1990–），助理研究员，研究方向：玢岩铁矿床成因及矿物学研究. E-mail：liping_zeng@cug.edu.cn

·专题 28：矿物岩石地球化学基础科学问题的数据驱动型研究进展·

Geochemistry π: 自动调参的机器学习 Python 工具

何灿[1,2]，孙建昊[3]，赵健铭[4]，吕洋[1]，王盛鑫[5]，赵文钰[1]，李岸洲[1]，张舟[1*]

1. 浙江大学地球科学学院；
2. 新加坡国立大学计算机系；
3. 中国地质大学（武汉）地球科学学院；
4. 吉林大学地球探测科学与技术研究学院；
5. 兰州大学地质科学与矿产资源学院

近年来，机器学习在地球化学领域的应用为相关科学问题带来了新认识。但是，地球化学科研人员需要花费大量时间精力实践机器学习方法。为了降低地球化学家应用数据挖掘方法研究的门槛，我们开发了 Geochemistry π，一款基于 Python 的高度自动化机器学习框架（Framework）。通过一键安装和运行 Geochemistry π，地球化学科研人员只需提供 Excel 表格数据，再从终端界面选择提供的选项，就可获得经过自动调参后的机器学习模型及其相关图表，为进一步分析提供依据。

Geochemistry π 以 Scikit-learn 库为基础，建立了定制化的自动化数据挖掘流程，为数据预处理环节提供统计分析方法，能够训练监督学习和无监督学习两类算法模型。本工具通过构建数据传输和算法功能应用相分离的层级式流水线架构，保证了框架的可扩展性和可移植性。同时，利用 FLAML 库的 CFO 和 BlendSearch 两种超参数搜索方法来实现 AutoML 模块，并结合 Ray 计算框架来加速模型参数优化过程。该工具操作简单，能有效省时省力，并采用 FLAML 和 Ray 框架相结合的方式，保障了模型的最优性能。总的来说，Geochemistry π 是地球化学领域研究人员快速挖掘表格数据潜在价值的高效易用工具。

第一作者简介：何灿（2000–），男，硕士研究生，研究方向：地球科学大数据. E-mail：sanyhew1097618435@163.com
*通信作者简介：张舟（1986–）男，百人计划研究员，研究方向：地球化学数据驱动型研究 E-mail: zhangzhou333@zju.edu.cn