Applied Geochemistry 140 (2022) 105273

Contents lists available at ScienceDirect

Applied Geochemistry

journal homepage: www.elsevier.com/locate/apgeochem

A review of machine learning in geochemistry and cosmochemistry: Method improvements and applications

Yuyang He [a,b,*,1], You Zhou [c,d,1], Tao Wen [e], Shuang Zhang [f], Fang Huang [g], Xinyu Zou [h], Xiaogang Ma [i], Yueqin Zhu [j]

[a] Key Laboratory of Earth and Planetary Physics, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, 100029, China
[b] State Key Laboratory of High Temperature Gas Dynamics, Institute of Mechanics, Chinese Academy of Sciences, Beijing, 100190, China
[c] International Research Center for Planetary Science, College of Earth Sciences, Chengdu University of Technology, Chengdu, 61005, China
[d] CAS Center for Excellence in Comparative Planetology, Hefei, 230026, China
[e] Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY, 13244, USA
[f] Department of Oceanography, Texas A&M University, College Station, TX, 77843, USA
[g] CSIRO Mineral Resources, Kensington, WA, 6151, Australia
[h] Key Laboratory of Mineral Resources, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, 100029, China
[i] Computer Science Department, University of Idaho, 875 Perimeter Drive, MS 1010, Moscow, ID, 83844-1010, USA
[j] National Institute of Natural Hazards, Ministry of Emergency Management of the People's Republic of China, Beijing, 100085, China

ARTICLE INFO

Editorial handling by Dr. Zimeng Wang

Keywords:
LIBS
XAFS
Mapping
Water/soil prediction
Molecular machine learning
Reactive-transport modeling

ABSTRACT

The development of analytical and computational techniques and growing scientific funds collectively contribute to the rapid accumulation of geoscience data. The massive amount of existing data, the increasing complexity, and the rapid acquisition rates require novel approaches to efficiently discover scientific stories embedded in the data related to geochemistry and cosmochemistry. Machine learning methods can discover and describe the hidden patterns in intricate geochemical and cosmochemical big data. In recent years, considerable efforts have been devoted to the applications of machine learning methods in geochemistry and cosmochemistry. Here, we review the main applications including rock and sediment identification, digital mapping, water and soil quality prediction, and deep space exploration. Research method improvements, such as spectroscopy interpretation, numerical modeling, and molecular machine learning, are also discussed. Based on the up-to-date machine learning/deep learning techniques, we foresee the vast opportunities of implementing artificial intelligence and developing databases in geochemistry and cosmochemistry studies, as well as communicating geochemists/cosmochemists and data scientists.

1. Introduction

Earth system is characterized by its complexity on scales from 10^{-10} m (size of atoms) to 10^{12} m (heliopause) on space, and from 10^{-10} s (equilibration time of fast chemical reactions) to 10^{17} s (age of Earth) in time. Each component of the system interacts with others and constitutes a constantly involving system. To describe the system, a great number of spatial-distributed and time-sequential variables are involved. As a part of Earth science, geochemistry has an inherent feature of complexity. With the development of space exploration, cosmochemistry has also shown its high-dimensionality in data.

In recent years, our abilities of collecting, storing, transferring, managing, and processing data are improved drastically. Lunar and Mars exploration is currently in full swing. It is expected that more data will be accessible, such as the samples brought back by the Chang'E−5 project and the data collected by the Mars Rover ZhuRong. While the data amount is exploding, the data structure becomes more and more intricate. The geochemical feature of an object has multiple dimensions, such as its compounds, elements, and isotope compositions. Information contained in the multi-dimensional spaces are largely undiscovered, due to limitation on ability of processing massive amount of multi-dimensional data. Effective and efficient data analysis techniques are

* Corresponding author. Key Laboratory of Earth and Planetary Physics, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, 100029, China.
E-mail addresses: yhe@mail.iggcas.ac.cn (Y. He), zhouyou06@cdut.cn (Y. Zhou), twen08@syr.edu (T. Wen), shuang-zhang@tamu.edu (S. Zhang), f.huang@csiro.au (F. Huang), zouxinyu@mail.iggcas.ac.cn (X. Zou), max@uidaho.edu (X. Ma), yueqinzhu@ninhm.ac.cn (Y. Zhu).
[1] Equal contribution.

https://doi.org/10.1016/j.apgeochem.2022.105273
Received 30 November 2021; Received in revised form 17 March 2022; Accepted 17 March 2022
Available online 23 March 2022

more needed now to improve the induction and deduction processes.

Big data and artificial intelligence are flourishing, which brings optimistic expectations for the future Earth and Space sciences (Bergen et al., 2019; Reichstein et al., 2019). Machine learning (ML) methods are powerful tools for finding and describing structural patterns in data, which help to extract information contained in data and assist in making predictions or decisions (Jordan and Mitchell, 2015). Each step of the scientific approach, i.e., the iteration of scientific problem determining, phenomenon observing, hypothesis formulating, prediction testing, and testable models or theories abstracting, is submerged by the flood of data (Mjolsness and DeCoste, 2001). The advantages in data processing and model optimization for specific tasks have already been demonstrated, such as rock and sediments identification (e.g., Petrelli et al., 2017), mineral prospectivity (e.g., Gregory et al., 2019), soil mapping (e.g., Hengl et al., 2017), and water/soil quality prediction (e.g., Wen et al., 2021). The applications in cosmochemical researches, such as Lunar surface geochemical mapping (e.g., Wang and Niu, 2012), decoding Laser-induced breakdown spectroscopy (LIBS) data on the Curiosity Mars Science Laboratory Rover (e.g., Boucher et al., 2015), and determining chemical contents of Apollo Lunar glasses by X-ray absorption fine structure (XAFS, e.g., Lanzirotti et al., 2018), have also been explored. In addition, ML methods have also been used to advance and improve research techniques, such as analysis techniques (e.g., Chen et al., 2020; Sutton et al., 2020), numerical modeling (e.g., Lee, 2020; Prasianakis et al., 2020), and theoretical calculation (e.g., Han et al., 2018; Wu et al., 2018; Pfau et al., 2020).

Liberating human from repeated works by computers can free great minds from tedious laboring, and can allow us devoting into creative works. Thus, the collaboration between geochemists/cosmochemists and data scientists is necessary. In this work, we review the recent research improvements in data acquiring, data processing, and research tool developing in geochemical and cosmochemical studies (Fig. 1). In section 2, we start from a brief history of ML and its main tasks to give the reader an impression about ML and its ability. Then, in section 3, we present the research methods improved by ML. In section 4, we review the most popular applications of ML on geochemistry and cosmochemistry, including lithology classification, geochemical mapping, soil/water quality predictions, and look forward the successful experiences that can be used on cosmochemistry in the near future. The main goal of this review is to communicate geochemists/cosmochemists with data scientists. Therefore, in section 5, we discuss the problems and developments of specialized databases and propose several feasible approaches that can strengthen geochemists/cosmochemists with data science knowledges.

2. Brief History of Machine Learning and its main tasks

Arthur Samuel coined the term “Machine Learning” for a checker program in 1959 (Samuel, 1959). A ML model, which is data-driven, can learn from data and improve its accuracy to the extent that is not explicitly programmed. After decades of developments, it has gradually evolved into different learning technologies, such as connectionism, symbolism and statistical learnings.

2.1. Symbolism

In the 1960s–1970s, the symbolic learning algorithms that are based on logical representation, were flourishing. Decision tree (DT) algorithm is one of the remarkable algorithms for symbolic learning. A DT model is a flowchart-like structure, in which each internal node represents a “test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (Michalski and Baskin, 1983). Its structure highly depends on the training dataset. Introducing any new data can lead to completely different tree structures. Thus, a DT model can hardly be transferred to another or an expanded dataset. In order to reduce the high variance of the DT algorithm, random forests (RF) algorithm was developed (Breiman, 2001). The RF algorithm adapted DT and bootstrap aggregating algorithms, in which the decision is made by the ensemble result of all the randomly generated DTs.

2.2. Statistical learning

The symbolic learning became less popular in the latter half of the 1990s, since its limited ability in handling problems with massive

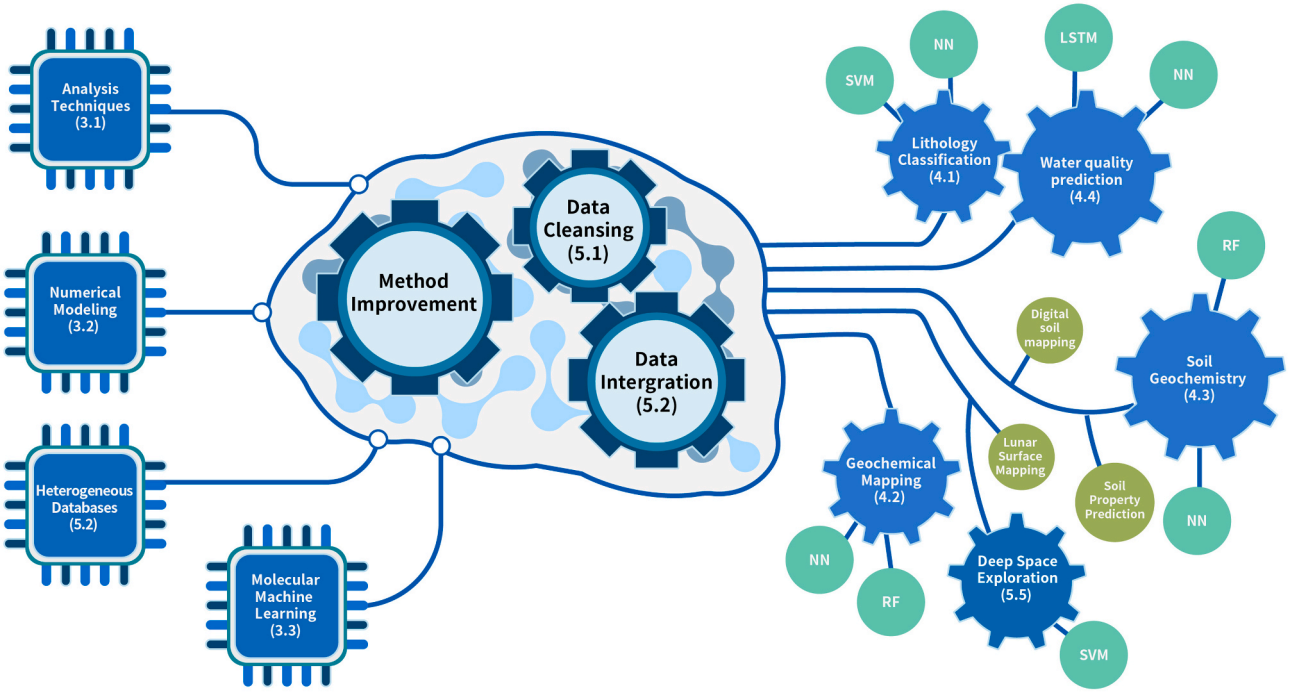

Fig. 1. Data acquiring, data processing, and research tool developing for geochemistry and cosmochemistry with machine learning methods. Numbers in text denotes the corresponding sections in this paper.

volume of data. The support vector machine (SVM) algorithm, which is a representative algorithm of the statistical learning, showed its superior performance on automatic text categorization (Cortes and Vapnik, 1995). It led the trend of statistical learning here after. A SVM model constructs the maximum margin hyperplanes to separate the vectors in a training dataset, which is suitable for solving classification problems with a small number of samples. However, for large-size data, it is prone to overfitting and requires special treatments to solve multi-classification problems.

2.3. Connectionism

2.3.1. Neural network

The representative algorithm for connectionism, perceptron, i.e., a single layer neural network (NN), was brought out in the 1950s (Rosenblatt, 1958). The core idea is to imitate human neurons, which receive information and transmits it to all adjacent neurons after processing. At that time, single layer NN models worked poorly even on some simple logic functions, such as the exclusive-or function (Minsky and Papert, 2017). The NN algorithm gained widespread recognition again in 1980s, when the Hopfield NN (Hopfield et al., 1983) and the back propagation (BP) algorithm (McClelland and Rnmelhart, 1986; Rnmelhart and McClelland, 1986) emerged. A BP neural network (BPNN) model has a hierarchical structure, in which every neuron in a layer is fully connected with all the neurons in the next layer (Kohonen, 1988). The learning process of a BPNN model is to adjust the weights between neurons and the bias of each neuron, which makes it to possess excellent nonlinear fitting abilities. The NN models are suitable for identification, classification, and prediction tasks.

2.3.2. Deep learning

Deep learning (DL) was proposed in 2006 based on the NN algorithms. Compared to traditional NN algorithms, deep neural network (DNN) algorithms have more (deeper) layers, which can map more complex nonlinear relationships hidden in data. In computer vision domain, a typical example is that a convolution neural network (CNN) model, the AlexNet, executed image classification tasks with extraordinary performance (Krizhevsky et al., 2012). The CNN algorithm has been successfully applied to object detection, facial detection/recognition, image segmentation, and edge detection (Chua, 1998). On sequential analysis problems, the recurrent neural network (RNN) algorithm improved the accuracy of voice recognition and natural language processing tasks to practical standards (Sutskever et al., 2014). Nevertheless, the major obstacles of DNN models are the lack of data, high training costs, and troublesome external parameter adjusting process.

2.4. Basic tasks of machine learning

The basic tasks of ML are classification, regression and clustering. Classification is to segment and separate data based on the given rules, which is used in supervised training. New data item can be mapped to a certain category according to the classifier. The commonly used algorithms for classification tasks include SVM, DT, Logistic Regression (LR), and K-nearest neighbor (KNN). Regression refers to a statistical analysis method to determine the quantitative relationship between two or more variables. Typical regression analysis algorithms include linear regression, LR, and least absolute shrinkage and selection operator (Lasso). Clustering is similar to classification, which separates data into similar groups, but it is used in unsupervised training that does not give requirements in advance. K-Means and hierarchical clustering are representative clustering algorithms. ML algorithms are used in geochemical and cosmochemical researches in tasks like recognition, recommendation, dimensionality reduction, and predictive analytics.

3. Improvements of research methods

Geochemical research methods, such as spectroscopy, numerical modeling, and theoretical calculations have made great progresses by incorporating ML technologies.

3.1. Analysis techniques

ML methods have been used to improve the effectiveness and efficiency of analysis techniques. In this section, we focus on the improvements of LIBS and XAFS spectroscopies, and briefly introduce the applications in micro x-ray fluorescence (μXRF) and micro x-ray diffraction (μXRD), as well as electron probe microanalyzer (EPMA).

3.1.1. Laser-induced breakdown spectroscopy

LIBS uses simple atomic emission by highly energetic laser pulse for *in situ* qualitative or semi-quantitative geochemical analysis (e.g., Boucher et al., 2015). Natural samples are mixtures, which result in complex LIBS spectra. To extract quantitative and qualitative information from LIBS spectra and improve analysis speed and accuracy, multiple ML techniques have been introduced (Chen et al., 2020). By integrating SVM or RF models with the chemical properties obtained from LIBS, the identification and discrimination of ten iron ore grades were completed (Sheng et al., 2015). A binary search algorithm (BSA) model with Calibration-free LIBS detected the CaO/SiO_2 mass ratios of iron ores (Wang et al., 2016). An independent component analysis-wavelet NN model was built for coal ash classification (Zhang et al., 2017). Kernel-based Extreme learning machine (ELM) models predicted carbon and sulfur contents, as well as calorific values in coal (Yan et al., 2017; 2019a; 2019b). Combining the variable reduction algorithm (Wootton, Sergent, Phan-Tan-Luu, V-WSP) and a wrapper method particle swarm optimization (PSO), a hybrid feature selection model V-WSP-PSO improved the accuracy of LIBS analysis for coal calorific value determination (Yan et al., 2019a). A deep belief network model with principal component analysis (PCA) improved the Pd detection in tobacco planted soil (Zhao et al., 2019a). A BPNN model determined trace element concentration in soil from generalized LIBS spectra (Sun et al., 2019). A graph theory model with PCA was constructed for classification of sedimentary and igneous rocks by LIBS and nanoparticle-enhanced LIBS (El-Saeid et al., 2019). A partial least square (PLS) regression model detected rare earth elements in natural geological samples (Bhatt et al., 2017). PLS, SVM, and RF models determined Zn, Cd, and Pb in seafood *Tegillarca granosa* (Ji et al., 2017). A locally linear embedding for regression model outperformed PLS and traditional locally linear embedding models on predicting the abundance of major elements in dust, rocks, and soils under Mars conditions, which was used to calibrate the LIBS is installed on the Curiosity Mars Science Laboratory Rover (Boucher et al., 2015).

3.1.2. X-ray absorption fine structure spectroscopy

XAFS spectroscopy uses the average electronic and molecular energy levels associated with a specific element to determine its valence state, coordination number, thereby to identity the nearest neighbors and bond lengths (e.g., McCanta et al., 2017). Recently, ML methods have been used to eliminate external factors that complicate the XAFS analysis and to aid data interpretation (Sutton et al., 2020). For instance, the PLS analysis of the entire X-ray absorption near edge structure (XANES) spectral region yielded accurate predictions of Fe^{3+} in garnets (Dyar et al., 2012) and iron valence state in amphiboles (Dyar et al., 2016a). In the predictions of iron redox state in silicate glasses, a Lasso model provided significantly better results than a PLS model (Dyar et al., 2016b). Using the Lasso model trained by Dyar et al. (2016b), McCanta et al. (2017) determined Fe^{3+} contents of Lunar glass beads collected during Apollo missions 11, 14, 15, and 17. PLS and Lasso models also predicted the magmatic oxygen fugacity of equilibration in basaltic glasses, which avoided the need for an external measure of the V valence

(Lanzirotti et al., 2018).

3.1.3. Other analysis techniques

Similar applications were done for other analysis techniques. For instance, calibrated by Alternative least square multivariate curve resolution, NN, SVM, and kernel ridge, the iron oxidation state in mid-ocean ridge basalt glasses determined by Raman spectroscopy agrees with the Fe K-edge XANES and wet-chemistry results (Le Losq et al., 2019). A NN model was trained to extract relationships between the element abundances from μXRF and mineral identity from μXRD. Based on this model, a Synchrotron-based Machine learning Approach for RasTer (SMART) mapper was developed that can identify minerals using only μXRF data (Kim et al., 2021). Integrating EPMA data with a multivariate polynomial regression (MPR) model provides an efficient and effective method for microanalysis of lithium in mica (Wang et al., 2022). Compare with other models, RF and extra-trees models reduced the forecast error by 30–40% in the determination of the iron oxidation state in clinopyroxene (cpx) by Mössbauer spectroscopy (Huang et al., 2022). ML techniques can greatly improve our ability of analyzing geochemical features of objects in laboratory and in field.

3.2. Numerical modeling

Numerical modeling conducts virtual experiments on different scales, which are important tools for geochemical and cosmochemical studies. There exists a common problem in different numerical modeling techniques, which is the great demands of computational resources. ML methods can reduce their computational complexities from atomic scale (e.g., Wu et al., 2018), to catchment scale (e.g., El Tabach et al., 2007), and to the scale of giant impacts (e.g., Cambioni et al., 2019; Zhou et al., 2021).

3.2.1. Chemical reaction kinetics

Chemical reaction kinetics simulations can predict products and can calculate reaction kinetic parameters of geochemical reactions under given initial conditions. It was used to estimate organic matter characters and to evaluate source and reservoir rocks in oil industry (Lee, 2020), and to predict propane isotopomer pyrolysis products for calibrating position-specific isotope analysis results (Goldman et al., 2019). Decomposition of organic matters involves an extensive number of reactions, where a series of nonlinear partial differential equations need to be solved. NN models accelerated chemical reaction kinetics simulation, and enabled a simple and efficient characterization of chemical reactions (Lee, 2020; Li et al., 2020a).

3.2.2. Reactive-transport modeling

The physics-based models (PBMs), including reactive-transport models (RTMs), simulate and reconstruct the movements and reactions of chemicals in systems like soil and groundwater (Li et al., 2017; Goddéris et al., 2019; Maher and Navarre-Sitchler, 2019). The quantitative prediction usually requires solving a series of coupled nonlinear partial differential equations, such as a coupled transient convection-advection-diffusion-reaction equation that involves spatial/temporal fields as well as variable fractional orders (He and Bao, 2019). The integration of gauge/well datasets (e.g., river discharge, element concentration, shallow water table depth) with various predictors (e.g., temperature, land topography) in a ML model could bypass the limitations of PBMs by identifying parameters that infer multi-complex spatiotemporal relations between various predictors and gauge/well observations.

Eutrophication modeling (EM) is considered one of the most difficult simulations of water-driven nutrient transport as well as the interactions between nutrients and biological communities (Shen and Kuo, 1998). It considers complexity in system interaction and nonlinear kinetics involved in biological dynamics. A NN-embedded genetic algorithm was used for inverse modeling of dissolved oxygen and chlorophyll-a from EM (Zou et al., 2007). Using the basic soil properties, a NN model determined transport velocity, dispersivity, and retardation factors of solutes in waters infiltrating soils (Mojid et al., 2019). In comparison to a complete RTM, emulator models including Gaussian processes, polynomial chaos expansion, and DNN models completed the tasks of direct emulation, global sensitivity analysis, uncertainty propagation, and calibration (Laloy and Jacques, 2019). Compared with an exact RTM simulation using the law of mass action, a NN model performance was four orders of magnitude faster with negligible reduction in accuracy in a model of a macroscopic system (Prasianakis et al., 2020).

3.2.3. Plate tectonics dynamics

Plate tectonics dynamics simulations model mantle convection to understand the driving mechanism for plate movements. Based on 300 mantle convection models, an SVM model determined the magnitude of the spin transition-induced density anomalies in mantle flows (Shahnas et al., 2018). Such model can be extended to predict more mantle properties, such as viscosity, elastic parameters, and the nature of thermal and chemical anomalies.

3.3. Molecular machine learning

Theoretical calculations, including molecular dynamics (MD), quantum mechanics (QM), and molecular mechanics, reveal geochemical reaction behaviors on atomic scale. The calculations are limited by the molecular size. For instance, the QM computational time of a molecule with n atoms increases approximately proportional to n^3 (Stern and Wolfsberg, 1966) or n^4 (Rustad, 2009). The tradeoff between efficiency and accuracy is a long-lasting challenge. Simplification, such as cluster model (He and Liu, 2015; Gao et al., 2018; Zhang et al., 2020; Li et al., 2021a), cutoff (He et al., 2020, 2021), and ONIOM (Chung et al., 2015) methods are used to reduce the computational resources. In recent years, machine-learning architectures are used to predict properties of molecules, which is called molecular machine learning (Wu et al., 2018).

3.3.1. Molecular dynamics

The MD simulations predict the time-dependent atomic processes of a system. The classical MD is based on Newton's equation of motion with simple analytical potential functions (Tsuchiyama et al., 1994). It does not require massive computational resources, but its accuracy is insufficient. The *ab initio* molecular dynamics (AIMD) describes atomic interactions using first-principles electronic structure methods such as density functional theory (DFT, Luo et al., 2020a). The AIMD uses a path integration algorithm to obtain an accurate potential energy surface (PES), which leads to significantly higher accuracy with considerably low computational efficiency. To address the dilemma of accuracy versus efficiency in AIMD simulations, Behler and Parrinello (2007) designed a special multilayer BPNN model to fit the PES in DFT, which improved the computing speed of a 64-atom system by 5 magnitudes. Schütt et al. (2017) designed a deep tensor neural network (DTNN) to fit the PES of a system. However, the BPNN and DTNN models have their own shortcomings. The BPNN model needs to introduce manual operations to maintain the input to meet local symmetry, and the DTNN model is incapable of handling large organic molecules (Han et al., 2018).

The Deep Potential (DP) method was established to address the inadequacies (Han et al., 2018). The DP performs in the speed close to the classical MD, and its accuracy reached the AIMD level. On basis of the DP, an open-source software DeePMD-kit was developed, which can directly complete NN models training and parameter setting by calling the TensorFlow (Wang et al., 2018). With interfaces in the commonly used MD software LAMMPS and i-PI, the DeePMD-kit becomes a user-friendly, accessible tool. Using the DeePMD-kit, Jia et al. (2020) calculated a system with 403 million water molecules and 113 million copper atoms in the accuracy of the AIMD level. Before that, the largest

system calculated has 1 million silicon atoms (Hasegawa et al., 2011). The DeePMD-kit calculated a system with 13 times more atoms at a speed of 5000 times faster.

Other molecular machine learning, such as TensorMol (Yao et al., 2017) and TorchMD (Doerr et al., 2021), also show strong potential to compete with or even outperform conventional MD methods in biomolecular simulations.

3.3.2. Quantum mechanics

The QM calculation describes a system by solving its Schrödinger equation. In practice, obtaining the wave functions of a target system requires many approximations due to the limitation of computational resources. NN models have been trained to calculate electronic properties (Montavon et al., 2013) and molecular atomization energies (Rupp et al., 2012; Hansen et al., 2013), as well as to improve the accuracy of the Møller-Plesset perturbation theory (McGibbon et al., 2017). Pfau et al. (2020) of the famous DEEPMIND team built the Fermionic neural network (FermiNet) for wave-function optimization. Although the FermiNet cannot converge the result for a large system, the electron density and optimized geometry calculation performances are comparable or better than all-electron CBS CCSD(T) method. The FermiNet also predicted the dissociation energy curves of the nitrogen molecule and hydrogen chain, which reaches a higher accuracy than other *ab initio* QM methods.

The application of ML on accelerating theoretical calculations can benefit geochemical studies, which need further explorations.

4. Applications in geochemistry and cosmochemistry

Under the guidance of experts' knowledge, ML methods have been used to in many geochemical fields with great success in a relatively short time. Due to the limitation of data, ML are mostly used in several fields. In the following sections, we present these state-of-the-art applications in lithology classification, geochemical mapping, soil/water quality predictions, and deep space explorations.

4.1. Lithology classification

4.1.1. Trace element indiscrimination diagram and basalt identification

The identification of lithologies, source of materials, special geological events, and mineralization zones from geochemical features is the basis of Earth and Space sciences. In the early stage, the dominant trend was to reduce the dimensionality of data and used only the representative geochemical features to serve research purposes. Taking basalt classification as an example, traditional geochemical methods compare different samples manually (e.g., mantle-normalized trace element diagrams, White, 1985), or reduce the variables into two to three key representatives (e.g., trace element discrimination diagrams, Debon and Le Fort, 1983; Batchelor and Bowden, 1985). Such methods, which are still widely used, appear to be applicable when the data amount is in handful sizes. However, with the increasing amount of data, some representative features fall short in drawing the whole picture. For instance, basalt trace element discrimination diagrams tend to fail when the datasets became tall, and they turned into "*trace element indiscrimination diagrams*" since different basalts largely overlapped on the diagrams (Li et al., 2015). Using a set of 13 rare-earth elements in zircon, Li et al. (2020c) exhaustively explored 4095 binary diagrams and 12485 ternary diagrams and found that none of them is able to distinguish mineralization and barren granites. Such discrepancies brought out challenges in finding rigorous and unbiased ways of handling the constantly increasing amount of data.

The Geochemistry of Rocks of the Oceans and Continents (GEOROC, Sarbas and Nohl, 2009) and the Petrological Database of the Ocean Floor (PetDB, Lehnert et al., 2000) are mature basalt geochemical databases, which are desirable well-structured training datasets. Using the databases, classification tree models were used to distinguish basalts from different tectonic environments (Vermeesch, 2006). Similar applications are made with SVM (Petrelli and Perugini, 2016; Petrelli et al., 2017; Ueki et al., 2018), RF (Ueki et al., 2018; Zhao et al., 2019b), k–means and fuzzy c–means (Yoshida et al., 2018), swarm optimized neural fuzzy inference system (Ren et al., 2019), and DNN (Zhao et al., 2019b) models. Besides using geochemical data as the training data directly, one application represented geochemical data by 2-dimensional gray images and used a CNN model to identify the tectonic settings (Ge et al., 2021).

4.1.2. Lithology and mineralogy classification

Besides basalt identifications, other lithology classifications were also explored for specific tasks. A RF model was more accurate than a SVM model for lithological mapping of serpentinite, talus and terrace deposits, argillites, conglomerates, basic lavas, limestone, and shales (Othman and Gloaguen, 2017). RF and gradient tree boosting models showed similar accuracy for formation lithology identification, which exceeds the performance of Naïve Bayes, SVM, and NN models. Nevertheless, all the models have difficulties in distinguishing sandstone (Xie et al., 2018). A RF model was used to identify intrusive lithologies in volcanic terrains (Kuhn et al., 2020). Combined a DNN model with thermodynamic equations, 'i-Melt' was built to predict 18 properties of melts and glasses in the K_2O-Na_2O-Al_2O_3-SiO_2 system. It has been used to explore the effect of the K/(K + Na) ratio on the properties of alkali aluminosilicate melts (Le Losq et al., 2021). Compared with a supervised PCA model using ten major-element geochemical data, an unsupervised self-organized map model obtained better lithological mapping results (Wu et al., 2021). SVM and Gaussian mixture models classified highly fractionated boninite series glasses from trace elements (Valetich et al., 2021).

Machine learning have also been used in the mineralogy classification. Based on petrography and mineral chemistry features, a PLS model identified magnetite in magmatic, hydrothermal, and metamorphic volcanogenic massive sulfide deposits (Makvandi et al., 2016). Huang et al. (2019) used trace element compositions to classify igneous and hydrothermal magnetite from porphyry deposits by a PLS model. A RF model distinguished ore deposit type and barren sedimentary pyrite from pyrite trace element data (Gregory et al., 2019). A SVM model slightly outperformed a NN model in discriminating the genesis of pyrites sampled from sedimentary rock, orogenic and volcanic hosted massive sulfide deposits, which revealed the multi-stage ore-forming history (Zhong et al., 2021). A linear discrimination analysis model outperformed traditional discrimination diagrams in discriminating ore-bearing and fertile granites based on trace element compositions of zircon (Li et al., 2020c). Iron oxide-copper-gold and iron oxide-apatite deposits were distinguished using magnetite composition by a RF model (Hong et al., 2021). A SVM model characterized the crystallization environments of quartz using trace elements (Wang et al., 2021b). PCA and SVM models traced the source rocks of detrital apatite (O'Sullivan et al., 2020).

In the classification tasks, the SVM classifiers usually yielded outstanding performance (Abedi et al., 2012; Kuwatani et al., 2014; Gonbadi et al., 2015; Heung et al., 2016; Petrelli and Perugini, 2016; Othman and Gloaguen, 2017; Petrelli et al., 2017; Mohammadi and Hezarkhani, 2018; Xie et al., 2018; Ueki et al., 2018; Hao et al., 2019; Sun et al., 2019; Lin et al., 2020; Flores et al., 2021; Zhong et al., 2021).

4.2. Geochemical mapping

The book *Geochemical method of prospecting for ore deposits* (Sergeev and Sokoloff, 1941) marked the dawn of geochemistry exploration and mapping. Since then, multivariate statistical methods, such as cluster analysis and factor analysis, have been developed to determine geochemical anomalies, element combinations, and to predict mineral deposits or soil contaminations (Cheng, 2007; 2012; Zuo and Cheng, 2008). Nowadays, researchers are exploring feasible ML methods on digital mapping problems.

4.2.1. Random forest

The RF models result in notable higher precision than other models, even the NN models, in many cases. Thus, it has become one of the most popular models in geochemical mapping, such as Cu (Rodriguez-Galiano et al., 2014, 2015), Cu-Au (Keykhay-Hosseinpoor et al., 2020), Ag–Pb–Zn (Wang et al., 2020), and terrestrial $^{87}Sr/^{86}Sr$ isoscapes (Bataille et al., 2018). Incorporating more than 9000 river sediments and 400 rock samples, tectonic unit, geological background, and geomorphic landscape, the geochemical anomalies identified by a RF model were in good agreements with the known Cu deposits (Tian et al., 2019). Using fuzzy-transformed input variables, three new highly prospective tungsten mineral targets were identified by a RF model (Yeomans et al., 2020).

4.2.2. Neural networks

The NN algorithm have been gradually implemented in geochemical explorations. The potential of hydrothermal gold and silver deposits predicted by a NN model had the accuracy higher than 70% (Oh and Lee, 2010). Polymetallic prospectivity mapping by an ELM model was 26 times faster than a LR model (Chen and Wu, 2017). Integrating spatial characteristics of shape, overlap, and zoning of multivariate geochemical anomalies and haloes, a CNN model was constructed for polymetallic deposits prospectivity mapping (Li et al., 2020b). Using labeled training data obtain from geochemical maps by the pixel-pair feature method, a CNN model identified geochemical anomalies that were consistent with the known mineral deposits (Zhang et al., 2021). Integrating information extracted from a deep variational autoencoder network, the obtained geochemical anomaly maps were in good agreement with known Fe polymetallic deposits (Luo et al., 2020b).

4.2.3. Other algorithms

ML methods were also used to identify possible mine sites (Abedi et al., 2012; Mohammadi and Hezarkhani, 2018; Dornan et al., 2020; Roshanravan, 2020), geochemical anomalies (Gonbadi et al., 2015; Esmaeiloghli and Tabatabaei, 2020), ore-forming stages of mineral deposits (Zhong et al., 2021). Combining PCA and local singularity analysis (LSA), the hybrid model performed better than a traditional factor ratio model in felsic intrusion mapping (Xiong and Zuo, 2016). An isolation forest model outperformed Boltzmann machine and LR models on both accuracy and efficiency in predicting Fe deposits (Chen and Wu, 2019). For both Sn and W prospectivity mapping, a DT model achieved the best accuracy (Iglesias et al., 2020). A SVM model showed the best performance in Cu-Au mineralization target prediction (Ghezelbash et al., 2021). Compared to other models, an Adaptive boosting (AdaBoost) model not only yielded the best results, but also simplified data preprocessing and hyperparameter tuning in fitting geochemical logging curves (Blanes de Oliveira and de Carvalho Carneiro, 2021). The eXtreme Gradient Boosting and RF models performed well in carbonate-hosted Zn-Pb mineral prospectivity mapping (Parsa, 2021).

4.3. Digital soil mapping and soil property predictions

Detailed global soil information is needed for climate change, sustainable environment development, and agricultural productivity problems (Hengl et al., 2017). Based on comprehensive soil databases, ML algorithms have been successfully applied to digital soil mapping and soil property predictions.

4.3.1. Neural networks

The NN algorithm is very suitable for predicting the distribution of soil pollutants and soil properties. A radial basis function neural network (RBFNN) model predicted acid sulfate in soil (Beucher et al., 2013). Compared with a RF model, a CNN-based “hybrid scale” model for digital soil mapping returns the most accurate results (Behrens et al., 2018). A CNN model predicted soil properties from regional spectral data with higher accuracy than the most commonly used models for soil spectroscopy (Padarian et al., 2019). The application in estimating soil organic matter content showed that NN models with different complexities resulted in the same accuracy (Fernandes et al., 2019). A NN model combined with geostatistics technology improved the prediction accuracy for spatial heavy metal content in topsoil (Sergeev et al., 2019). The determination and consistency correlation coefficients of a CNN model could greatly improve the prediction accuracy for total organic carbon at different soil depths (Wadoux et al., 2019). A DNN model had the best accuracy on soil organic carbon prediction (Emadi et al., 2020).

4.3.2. Other algorithms

The RF algorithm also showed good performance in soil studies. It worked the best in predicting soil categories (Brungard et al., 2015), soil carbon concentration (Keskin et al., 2019; Mahmoudzadeh et al., 2020), and ion concentration and geochemical gradients (Diaz et al., 2021). A RF model effectively extracted the relationship between soil parent material and terrain, and the soil maps generated were highly consistent with soil surveys (Heung et al., 2014). Increasing detail, accuracy, interpretability, and uncertainty awareness for high-resolution regional soil mapping were provided by a quantile RF model (Kirkwood et al., 2016). The soil redox interface depth obtained from an RF model only reconstructed half of the real condition, and the local uncertainty strongly depended on the variance scaling method (Koch et al., 2019).

Besides, universal kriging, sequential Gaussian simulation, quantile return to the Forest, Bayesian network, PLS, alternative regression equations and rule-based regression tree, and spectrum-based learner models were also applied to the predictions of soil contamination, texture content, organic carbon, calcium carbonate, pH, and electrical conductivity etc. (Albuquerque et al., 2017; Mikkonen et al., 2018; Boente et al., 2019, 2020; Duarte-Guardia et al., 2019; Szatmári and Pásztor, 2019; Tziolas et al., 2019). In the prediction of soil water retention, a KNN models not only obtained similar accuracy with NN models (Nemes et al., 2006b; Coopersmith et al., 2014), but also is insensitive to different datasets, data densities, and input attribute weights (Nemes et al., 2006a). A KNN model was also used to estimate near-surface soil moisture from deeper *in situ* records (Coopersmith et al., 2016).

4.4. Water quality prediction and forecasting

Given recent advances in high resolution environmental/geological variables, growing computational resources and algorithms processing geographic data efficiently, the implementation of the ML approach in modeling hydrological features is advancing at a rapid speed. Predictor variables considered in these data-driven hydrological ML models can be very diverse, including ecological variables (e.g., land cover), climatic variables (e.g., temperature), geomorphic variables (e.g., slope), lithological variables (surface rock type), hydrological variables (e.g., precipitation), soil properties (e.g., soil density and composition), and other watershed-related properties. Take tree-based algorithms as examples. They have been used to successfully assess the driving forces for the variations of hydrogeochemical parameters and to predict their distribution in both groundwater (Lek et al., 1999; Nolan et al., 2015; Tesoriero et al., 2017; Stackelberg et al., 2021), surface water (Lintern et al., 2018a, 2018b; Peterson et al., 2019; Shen et al., 2020).

4.4.1. Neural networks

The NN algorithms are widely used in water quality prediction and forecasting (Maier et al., 2010). In 1999, a NN model already predicted stream nitrogen concentrations using watershed parameters (Lek et al., 1999). By analyzing time-series hydrochemical properties, a NN model investigated the driving forces of the sulfate (Lischeid, 2001). A NN model forecasted river salinity 14 days in advance (Bowden et al., 2002). Compared to a BPNN model, a RBFNN model required less training efforts and yield more robust results on modeling yearly nitrate concentrations in rivers (Suen and Eheart, 2003). A BPNN model assessed flash

floods and their attendant water quality parameters (Sahoo et al., 2006). A three-layer cascade correlation artificial neural network model successfully estimated missing monthly values of water quality (Diamantopoulou et al., 2007). A DNN model predicted reservoir temperatures accurately based on typical hydrogeochemical parameters (Tut Haklidir et al., 2020).

Nevertheless, NN models were not always the best candidate for hydrogeochemical problems. In the prediction of copper concentrations in acid mine drainage, a SVM model with polynomial kernel showed the highest accuracy compared to other models, including to a NN model (Betrie et al., 2013). However, another study using the same dataset showed that a NN model exceeded a SVM model in predicting Cu and Zn concentrations in acid mine drainage (Betrie et al., 2014). A BPNN model performed worse in predicting total dissolved solids of an urban aquifer, while a hybrid principal component regression model not only minimized multicollinearity of the input parameters but also yielded accurate and precise results (Pan et al., 2019). Compared to RBFNN, Multilayer Perceptron (MLP), and Gene Expression Programming models, a least square support vector machine with Firefly Algorithm (FFA) model was demonstrated to be the best method in estimating carbon dioxide solubility at high pressure and temperature conditions (Hemmati-Sarapardeh et al., 2020).

More recently, the long short-term memory (LSTM) model, one specialized RNN model, gained popularity in predicting water chemistry using time-series data. For example, Zhi et al. (2021) construct a LSTM model using dissolved oxygen data plus 56 meteorological and watershed attributes from 236 minimally human-disturbed U.S. watersheds. The model successfully predicted the dissolved oxygen content in ungauged watersheds. Using LSTM models to make predictions for time-series data is particularly effective, which is due partly to that LSTM take in account the temporal context on the both short- and long-term scales.

4.4.2. Other algorithms

Other methods, such as KNN, DT, boosted regression trees, gradient boosting machine, RF, and SVM were also used to study concentrations of nitrate (Nolan et al., 2018; Ransom et al., 2018; Uddameri et al., 2020), dissolved oxygen (Coopersmith et al., 2011; Bertone et al., 2015; Erickson et al., 2021), arsenic (Meliker et al., 2008; Podgorski et al., 2020), chlorophyll-a (Yajima and Derot, 2018), methane (Wen et al., 2021), and Uranium (Lopez et al., 2021), acid rock drainage chemistry (Betrie et al., 2014; Flores et al., 2021), water salinity (Tran et al., 2021), pH value (Astray et al., 2021; Stackelberg et al., 2021), as well as oxygen isotope composition and water temperature (Astray et al., 2021).

Another important application in hydrogeochemistry is to identify the sources of geochemical constituents in water. Several inverse models were developed to tackle this mixing problem (Gaillardet et al., 1999; Jacobson and Blum, 2003). However, these inverse models often require some prior knowledge of the chemical composition of endmembers, which we commonly have little information. Unlike inverse models, ML models do not require such prior knowledges. A non-negative matrix factorization model deconvolved the sources of sulfate content in surface water without incorporating source information (Shaughnessy et al., 2021).

4.5. Deep space explorations

ML methods have been used to improve the performance for well-defined science problems and hypotheses. Although the cosmochemical data are comparatively limited, the development of observation technology and implementation of space exploration projects will bring us new data in the very near future. The successful applications of ML methods in geochemical researches offer valuable insights into cosmochemical researches.

An accurate geochemical map can serve as a valuable assistant for Lunar exploration projects, such as mineral and rock identification, resources exploration, and landing spot selection. Element distributions on Lunar surface are also fundamental for understanding its evolution and collision history. The element distribution on Earth's surface can be investigated by detailed geochemical surveys. However, the direct Lunar surface mapping is still beyond our ability. Two types of approaches are used to explore element distributions on Lunar surface. The most accurate way is to analyze Lunar samples. However, Lunar samples with known location information are limited to the Apollo project and the Chang'E project. The samples are sparse and only represent limited locations. Spectroscopy, such as gamma ray and neutron spectroscopy (e.g., Prettyman et al., 2006), X-ray spectroscopy (e.g., Swinyard et al., 2009), and optical spectroscopy (e.g., Wu, 2012), can provide widespread Lunar maps but require further corrections. Traditional studies based on linear relationships to derive oxide contents from optical images have limited accuracies (Xia et al., 2019).

In recent years, ML methods are used to depict complex nonlinear relationships between spectral characteristics and geochemical components. Based on the Clementine ultraviolet–visible spectroscopy maps and Lunar Soil Characterization Consortium data, Lunar surface TiO_2 abundance map was constructed by a NN model, which reproduced the distribution of mare basalts (Korokhin et al., 2008). Based on the whole 32 channels of Interference imaging spectrometer (IIM) images of Chang'E−1 and 36 Apollo and Luna station samples, a DT-SVM model predicted TiO_2 abundance of Lunar surface soil (Wang and Niu, 2012). Since there were only 4 high-Ti samples accessible, the DT-SVM model performed better in regions containing very low TiO_2 contents, but underestimated the high-Ti units. To improve the DT-SVM model, the authors removed 13 abnormal bands and used an SVM-SVM model that requires less training data (Wang and Zhu, 2013). Based on the IIM images and the analytical results of 39 Lunar samples, a NN model constructed the distribution maps of Mg/(Mg + Fe) ratio, SiO_2, Al_2O_3, CaO, FeO, MgO, and TiO_2 abundances on Lunar surface (Xia et al., 2019). Such attempts can be expanded to other spectroscopies with one limitation, which is the shortage of Lunar sample analysis data.

It is expected that more interesting applications will emerge. For instance, the Thermal Emission Imaging System (THEMIS) instrument on board the Mars Odyssey spacecraft monitors the Martian surface mineral distributions, atmospheric dust, and surface temperatures (e.g., Viviano and Moersch, 2013), and the LIBS on Curiosity and ZhuRong, can provide direct geochemical information on Mars. Combines with THEMIS and LIBS data, ML algorithms should be able to provide detailed Mars surface geochemical maps.

In addition, determining the surface age of a planet is the first step of studying its evolution history. Due to the limitation of extraterrestrial samples, the surface age of a planet is determined mainly by crater statistics, i.e., the counting of the number of craters. Traditionally, the craters were performed by visual inspection of images by researchers. Image recognition is one of the strengths of DL techniques. Using a CNN model, researchers achieved accurate and automatic crater determination on Lunar surface, which improves the efficiency of crater-dating (Silburt et al., 2019; Ali-Dib et al., 2020). Nevertheless, the crater chronology method is empirical, which needs further verification and calibration by isotope dating of extraterrestrial samples with known geological context (e.g., Che et al., 2021; Li et al., 2021b; Yue et al., 2022). ML algorithms can bridge crater-dating and isotope dating methods, which could generate a more accurate crater chronology.

4.6. Other applications

Lithology/mineral classification, geochemical mapping, and water/soil quality prediction are the most intensive users of ML techniques. Besides, other attempts were also done, such as the identifications of tsunami deposits (Kuwatani et al., 2014), geochemical distribution patterns of light rare earth elements (Zaremotlagh and Hezarkhani, 2017), heavy minerals in river sands (Hao et al., 2019), dust sources of Chinese loess plateau (Lin et al., 2020), controlling factors of lacustrine

shale lithofacies (Liu et al., 2020), natural gas origins (Snodgrass and Milkov, 2020), and heavy metal pollution sources (Khorshidi et al., 2021).

Recently, ML methods were used to determine the geochemical features of minerals. Based on ordinary least squares linear regression plus ML-based stochastic gradient boosting, extremely randomized trees (ERTs), RF, KNN, and decision trees models, Petrelli et al. (2020) presented a machine learning thermo-barometry to estimate pre-eruptive pressures and temperatures and storage depths using cpx-melt pairs and cpx-only chemistry. A SVM model distinguished the cpx phenocrysts that do or do not undergo hydrogen diffusion, in which the initial water contents of basalt magmas were estimated from the cpx without H diffusion (Chen et al., 2021).

5. Discussions and prospective

Data mining from geochemical data can be the key and the engine to understand the evolution of the Earth system and to guide our space exploration of extraterrestrial bodies. Modern ML schemes take advantage of big data to extract new information and knowledge, which enables remarkable engineering improvements and scientific discoveries. However, unlike ML applications in biology, physics, and solid Earth (e.g., Bergen et al., 2019; Kates-Harbeck et al., 2019; Jumper et al., 2021), the application of ML in geochemistry and cosmochemistry is still in an initial stage. The main hindrances are: 1) the absence of large geochemical and cosmochemical databases that can handle heterogeneous, time-sequential, high-complexity, multi-scale, and multi-dimensional data; and 2) new algorithms that suitable for small datasets. Designing algorithms requires strong computer science and mathematics skills, which are not the strengths of geochemists and cosmochemists. However, geochemists and cosmochemists' domain knowledge and fundamental understanding of the geoscientific problems can be really helpful in constructing specialized databases. Therefore, in this section, we focus on discussing database problems. In addition, data science is moving very quickly while the adoption of data science techniques in geoscience is relatively slow partly due to the lack of data science expertise in geoscientists. Educating graduate students with geochemical and cosmochemical fields with data science knowledge will further promote interdisciplinary researches benefiting both Earth and Space sciences, as well as data science communities. Thus, we further prospect the educating next-generation geochemists and cosmochemists, which will help address challenges of the algorithms mentioned before.

5.1. Problems in current databases

Current applications of ML based on current geochemical databases have some shortages. First, most datasets are acquired by individual research groups or institutions. The major challenges for such geochemical datasets include biased sampling, low resolution, and disconnected data sources. The ML models trained by regional datasets have restricted applications, which is hard to be generalized. Data cleaning and preprocessing can result in better outcomes (Grunsky and de Caritat, 2019; Grunsky and Arne, 2020; Daviran et al., 2021). Based on geochemical anomalies recognized from a generalized additive model, a Bayesian framework model constructed three-dimensional models of deep alteration zones for mineral resources (Chen et al., 2020). Combined with two swarm intelligence optimization algorithms, i.e., bat algorithm and FFA, which optimized the initial hyperparameters setting, the accuracy of MLP, AdaBoost, and one-class SVM models were greatly improved (Lin et al., 2021).

Second, the insufficient amount of geochemical and cosmochemical data hindered further applications. For example, the number of samples required for geochemistry exploration is generally in the magnitude up to tens of thousands, which is usually the largest in geochemistry field. However, this amount of data is difficult to drive most of the exiting DNN models that designed for larger data sizes.

Third, the consistency of raw data is insufficient in many cases. Geochemical and cosmochemical data are affected by the sample collection process, analysis process, and even different sampling seasons. Therefore, in the same database, the original data of different regions may have deviations, which potentially affect the results.

Last but not the least, current databases are mostly designed for special research topics. Databases in limited disciplines tend to restrict correlation efficacy for broader purposes. The development of open-source, high-quality, comprehensive databases that are labeled by experts' knowledge would facilitate the extensive exploration of ML methods with potentials on knowledge discovery to the extent of changing the research paradigm.

5.2. Developments of geochemical databases

Several Earth big data grand projects are pushing forward the integration of multi-source data to help generate more relevant, richer, and complete information. Given the large variety of scientific questions being addressed by the geochemistry community, a large variety of databases and data repositories (e.g., structured versus unstructured) are demanded by the geochemistry community (Brantley et al., 2021). The projects aim to promote accessing, sharing, and using geochemical data to support study, analysis and decision making.

For instance, the EarthChem established in 2005 manages open-access digital geochemical resources including the PetDB (Lehnert et al., 2000), the North American Volcanic and Intrusive Rock Database (Walker et al., 2006), the GEOROC (Sarbas and Nohl, 2009), the metamorphic petrology database (MetPetDB, Spear et al., 2009), the marine and terrestrial sediment geochemical data, (SedDB, Johansson et al., 2012), the GANSEKI (Tomiyama et al., 2013), and the USGS National Geochemical Database. The EarthChem participates in the Interdisciplinary Earth Data Alliance (IEDA) that was funded by the US National Science Foundation (NSF) in 2010. As a primary community data collection, it provides data services for data from the ocean, Earth, and polar sciences that supports interdisciplinary research and data integration. One of IDEA's goals is to bridge the gap between scientists and data. The NSF Directorate for Geosciences and the Division of Advanced Cyber infrastructure initiated the EarthCube project in 2011. The EarthCube aims to transform geoscience research by improving access, sharing, visualization, and analysis of geosciences data. The IEDA is an active member of EarthCube's Council of Data Facilities.

Currently, Chinese scientists are leading the development of Earth science databases. The Chinese Academy of Sciences Strategic Priority Research Program supported the Big Earth Data Science Engineering (CASEarth) project that focuses on modern Earth's surface data, the establishment of discrete global grid systems, and geospatial information processing and visualization platforms (Guo et al., 2020). The Deep-time Digital Earth (DDE) project, which is one of the International Union of Geological Sciences big science programs, is working on standardize and digitize geological data in deep-time (Normile, 2019; Wang et al., 2021a). Based on the comprehensive databases that will be established, the DDE aims to use data sciences and information techniques to understand the Earth system evolution.

All the above-mentioned datasets have great potentials to be the success factors of applying ML techniques in geochemical researches. It must be noted that ML methods are powerful pattern recognition methods, which are still statistics techniques. The correlations in data cannot be conclusive evidence for certain problems. Currently, some research studies focus on testing the most accurate and efficient multiple ML methods for specific tasks. The ML models, especially the DNN models that contain multiple non-linear hidden layers, can learn very complicated relationships between their inputs and outputs. Increasing of fitting parameters, sampling noises, overtraining, and insufficient training data can result in overfitting in the training dataset. Many of these complicated relationships may not exist in real data (Srivastava

et al., 2014). Biased data, mixing training and testing data, overfitting, and improper validation will lead to unreliable results (Schaffer, 1993). Therefore, ML should not be applied naively to complex geoscience problems, and one urgent question we are facing is to distinguish useful and meaningful information from the mist of correlations, and to understand the fundamental principles under the superficial characteristics of the phenomenon.

5.3. Preparing next-generation geochemists and cosmochemists

To efficiently and effectively self-educate or educate graduate students, one or several of the following approaches might be worth considering.

First, taking advantage of massive open online courses (MOOCs) education resources (e.g., Coursera and edX) that are accessible for free. These online courses are particularly useful if geochemistry graduate students need to learn basics in data science as many are designed for students from all backgrounds and majors. In addition, more and more online data science courses developed by geoscience educators become available for instructors to adapt or for students to learn in a self-paced way. For example, Wen et al. (2020) developed a data science course hosted on the platform of HydroLearn funded by the NSF to educate geoscience students about the basics and advanced knowledge of data science using genuine research data and peer-reviewed geoscience research works.

Second, attending a boot-camp style training workshop. Geoscience conferences like AGU and Goldschmidt started offering data science workshops. These workshops often enable participants to gain first-hand experiences in using data science knowledge for addressing geoscience questions in a fast-paced manner. These workshops and the MOOC resources can be used as the first bite of data science for geochemistry graduate students.

Third, developing new data science curricula specifically designed for geochemistry graduate students that are built on existing geoscience courses. Instructors can start with incorporating data science modules into their courses by focusing on one or a few selected topics. The feedbacks from students on these small modules can help shape the scope and structure of the new geo-data science course to be developed.

6. Conclusions

This is the best of time, that we have unprecedented capability of acquiring and processing vast amount of data. This is the worst of time, that we are overwhelmed with data, and our ability of information and knowledge extraction might not catch up with the high data volume and the complexity of data structures. Earth and Space sciences are experiencing the positive hype triggered by big data and ML. The efficient data collecting, producing, and processing, as well as model iteration by ML can be amplifiers of scientific approaches, which is expected to change the research methods in many realms of Earth and Space sciences fundamentally. At present, ML can greatly improve sampling tools, speed up calculation methods, and be used for classification, regression, and clustering tasks. Possible hypotheses may be suggested from data patterns that can guide the research directions, which could be the new growing point of Earth and Space sciences.

Author contributions

Y. He and Y. Zhou prepared the original draft. All authors provided critical feedback and helped shape the manuscript. Y. He is responsible for *Improvements of research methods* and *Deep space explorations*. Y. Zhou is responsible for *Brief History of Machine Learning and its main tasks*. T. Wen is responsible for *Water quality prediction and forecasting* and *Preparing next-generation geochemists and cosmochemists*. S. Zhang is responsible for *Digital soil mapping and soil property predictions*. F. Huang is responsible for *Geochemical mapping*. X. Zou is responsible for *Lithology classification*. X. Ma and Y. Zhu are responsible for *Problems in current databases* and *Development of geochemical databases*.

Funding

This work was supported by the National Science Foundation of China (NSFC) project [42150202, 4217030170], China Postdoctoral Science Foundation [2019M660811], and the pre-research project on Civil Aerospace Technologies of China National Space Administration [D020203] to Y. He, the NSFC projects [41973063, 42011530431] to Y. Zhou, the Earth Science Information Partners Lab Grant [05088] to T. Wen, the NSFC project [42003021] to X. Zou, the U.S. National Science Foundation (NSF) project [2126315] to X. Ma, and the NSFC projects [41872253] to Y. Zhu.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The project is benefited from valuable discussions with Xianhua Li, Yun Liu, Jianfeng Gao, Chao Ma, and Zhiyong Xiao.

References

Abedi, M., Norouzi, G.-H., Bahroudi, A., 2012. Support vector machine for multi-classification of mineral prospectivity areas. Comput. Geosci. 46, 272–283.
Ali-Dib, M., Menou, K., Jackson, A.P., Zhu, C., Hammond, N., 2020. Automated crater shape retrieval using weakly-supervised deep learning. Icarus 345, 113749.
Albuquerque, M.T.D., Gerassis, S., Sierra, C., Taboada, J., Martin, J.E., Antunes, I.M.H.R., Gallego, J.R., 2017. Developing a new Bayesian Risk Index for risk evaluation of soil contamination. Sci. Total Environ. 603, 167–177.
Astray, G., Soto, B., Barreiro, E., Gálvez, J.F., Mejuto, J.C., 2021. Machine learning applied to the oxygen-18 isotopic composition, salinity and temperature/potential temperature in the Mediterranean Sea. Mathematics 9, 1–15.
Bataille, C.P., von Holstein, I.C., Laffoon, J.E., Willmes, M., Liu, X.-M., Davies, G.R., 2018. A bioavailable strontium isoscape for Western Europe: a machine learning approach. PLoS One 13, e0197386.
Batchelor, R.A., Bowden, P., 1985. Petrogenetic interpretation of granitoid rock series using multicationic parameters. Chem. Geol. 48 (1–4), 43–55.
Behler, J., Parrinello, M., 2007. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys. Rev. Lett. 98, 146401.
Behrens, T., Schmidt, K., MacMillan, R.A., Rossel, R.A.V., 2018. Multi-scale digital soil mapping with deep learning. Sci. Rep. 8.
Bergen, K.J., Johnson, P.A., de Hoop, M.V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid Earth geoscience. Science 363, eaau0323.
Bertone, E., Stewart, R.A., Zhang, H., Veal, C., 2015. Data-driven recursive input–output multivariate statistical forecasting model: case of DO concentration prediction in Advancetown Lake, Australia. J. Hydroinf. 17, 817–833.
Betrie, G.D., Sadiq, R., Morin, K.A., Tesfamariam, S., 2014. Uncertainty quantification and integration of machine learning techniques for predicting acid rock drainage chemistry: a probability bounds approach. Sci. Total Environ. 490, 182–190.
Betrie, G.D., Tesfamariam, S., Morin, K.A., Sadiq, R., 2013. Predicting copper concentrations in acid mine drainage: a comparative analysis of five machine learning techniques. Environ. Monit. Assess. 185, 4171–4182.
Beucher, A., Österholm, P., Martinkauppi, A., Edén, P., Fröjdö, S., 2013. Artificial neural network for acid sulfate soil mapping: application to the Sirppujoki River catchment area, south-western Finland. J. Geochem. Explor. 125, 46–55.
Bhatt, C.R., Jain, J.C., Goueguel, C.L., McIntyre, D.L., Singh, J.P., 2017. Determination of rare earth elements in geological samples using laser-induced breakdown spectroscopy (LIBS). Appl. Spectrosc. 72, 114–121.
Blanes de Oliveira, L.A., de Carvalho Carneiro, C., 2021. Synthetic geochemical well logs generation using ensemble machine learning techniques for the Brazilian pre-salt reservoirs. J. Petrol. Sci. Eng. 196, 108080.
Boente, C., Albuquerque, M.T.D., Gerassis, S., Rodriguez-Valdes, E., Gallego, J.R., 2019. A coupled multivariate statistics, geostatistical and machine-learning approach to address soil pollution in a prototypical Hg-mining site in a natural reserve. Chemosphere 218, 767–777.
Boente, C., Gerassis, S., Albuquerque, M.T.D., Taboada, J., Gallego, J.R., 2020. Local versus regional soil screening levels to identify potentially polluted areas. Math. Geosci. 52, 381–396.
Boucher, T., Carey, C.J., Dyar, M.D., Mahadevan, S., Clegg, S., Wiens, R., 2015. Manifold preprocessing for laser-induced breakdown spectroscopy under Mars conditions. J. Chemometr. 29, 484–491.

Bowden, G.J., Maier, H.R., Dandy, G.C., 2002. Optimal division of data for neural network models in water resources applications. Water Resour. Res. 38, 1–11.
Brantley, S.L., Wen, T., Agarwal, D., Catalano, J.G., Schroeder, P.A., Lehnert, K., Varadharajan, C., Pett-Ridge, J., Engle, M., Castronova, A.M., Hooper, R.P., Ma, X., Jin, L., McHenry, K., Aronson, E., Shaughnessy, A.R., Derry, L.A., Richardson, J., Bales, J., Pierce, E.M., 2021. The future low-temperature geochemical data-scape as envisioned by the U.S. geochemical community. Comput. Geosci. 104933 https://doi.org/10.1016/j.cageo.2021.104933.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards Jr., T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239, 68–83.
Cambioni, S., Asphaug, E., Emsenhuber, A., Gabriel, T.S.J., Furfaro, R., Schwartz, S.R., 2019. Realistic on-the-fly outcomes of planetary collisions: machine learning applied to simulations of giant impacts. Astrophys. J. 875, 40–54.
Che, X., Nemchin, A., Liu, D., Long, T., Wang, C., Norman, M.D., Joy, K.H., Tartese, R., Head, J., Jolliff, B., Snape, J.F., Neal, C.R., Whitehouse, M.J., Crow, C., Benedix, G., Jourdan, F., Yang, Z., Yang, C., Liu, J., Xie, S., Bao, Z., Fan, R., Li, D., Li, Z., Webb, S. G., 2021. Age and composition of young basalts on the Moon, measured from samples returned by Chang'e-5. Science 374, 887–890.
Chen, H., Su, C., Tang, Y.-Q., Li, A.-Z., Wu, S.-S., Xia, Q.-K., ZhangZhou, J., 2021. Machine learning for identification of primary water concentrations in mantle pyroxene. Geophys. Res. Lett. 48, e2021GL095191.
Chen, T., Zhang, T., Li, H., 2020. Applications of laser-induced breakdown spectroscopy (LIBS) combined with machine learning in geochemical and environmental resources exploration. Trends Anal. Chem. 133, 116113.
Chen, Y., Wu, W., 2017. Mapping mineral prospectivity using an extreme learning machine regression. Ore Geol. Rev. 80, 200–213.
Chen, Y., Wu, W., 2019. Isolation forest as an alternative data-driven mineral prospectivity mapping method with a higher data-processing efficiency. Nat. Resourc. Res. 28, 31–46.
Cheng, Q., 2007. Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geol. Rev. 32 (1–2), 314–324.
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.
Chua, L.O., 1998. CNN: A Paradigm for Complexity, vol. 31. World Scientific.
Chung, L.W., Sameera, W.M.C., Ramozzi, R., Page, A.J., Hatanaka, M., Petrova, G.P., Harris, T.V., Li, X., Ke, Z.F., Liu, F.Y., Li, H.B., Ding, L.N., Morokuma, K., 2015. The ONIOM method and its applications. Chem. Rev. 115, 5678–5796.
Coopersmith, E.J., Cosh, M.H., Bell, J.E., Boyles, R., 2016. Using machine learning to produce near surface soil moisture estimates from deeper in situ records at US Climate Reference Network (USCRN) locations: analysis and applications to AMSR-E satellite validation. Adv. Water Resour. 98, 122–131.
Coopersmith, E.J., Minsker, B., Montagna, P., 2011. Understanding and forecasting hypoxia using machine learning algorithms. J. Hydroinf. 13, 64–80.
Coopersmith, E.J., Minsker, B.S., Wenzel, C.E., Gilmore, B.J., 2014. Machine learning assessments of soil drying for agricultural planning. Comput. Electron. Agric. 104, 93–104.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.
Daviran, M., Maghsoudi, A., Ghezelbash, R., Pradhan, B., 2021. A new strategy for spatial predictive mapping of mineral prospectivity: automated hyperparameter tuning of random forest approach. Comput. Geosci. 148, 104688.
Diamantopoulou, M.J., Antonopoulos, V.Z., Papamichail, D.M., 2007. Cascade correlation artificial neural networks for estimating missing monthly values of water quality parameters in rivers. Water Resour. Manag. 21, 649–662.
Diaz, M.A., Gardner, C.B., Welch, S.A., Jackson, W.A., Adams, B.J., Wall, D.H., Hogg, I. D., Fierer, N., Lyons, W.B., 2021. Geochemical zones and environmental gradients for soils from the central Transantarctic Mountains, Antarctica. Biogeosciences 18, 1629–1644.
Doerr, S., Majewski, M., Pérez, A., Krämer, A., Clementi, C., Noe, F., Giorgino, T., De Fabritiis, G., 2021. TorchMD: a deep learning framework for molecular simulations. J. Chem. Theor. Comput. 17, 2355–2363.
Dornan, T., O'Sullivan, G., O'Riain, N., Stueeken, E., Goodhue, R., 2020. The application of machine learning methods to aggregate geochemistry predicts quarry source location: an example from Ireland. Comput. Geosci. 140, 104495.
Debon, F., Le Fort, P., 1983. A chemical–mineralogical classification of common plutonic rocks and associations. Earth Environ. Sci. Trans. Roy. Soc. Edinb. 73 (3), 135–149.
Duarte-Guardia, S., Peri, P.L., Amelung, W., Sheil, D., Laffan, S.W., Borchard, N., Bird, M. I., Dieleman, W., Pepper, D.A., Zutta, B., Jobbagy, E., Silva, L.C.R., Bonser, S.P., Berhongaray, G., Piñeiro, G., Martinez, M.-J., Cowie, A.L., Ladd, B., 2019. Better estimates of soil carbon from geographical data: a revised global approach. Mitig. Adapt. Strategies Glob. Change 24, 355–372.
Dyar, M.D., Breves, E.A., Emerson, E., Bell, S.W., Nelms, M., Ozanne, M.V., Peel, S.E., Carmosino, M.L., Tucker, J.M., Gunter, M.E., Delaney, J.S., Lanzirotti, A., Woodland, A.B., 2012. Accurate determination of ferric iron in garnets by bulk Mössbauer spectroscopy and synchrotron micro-XANES. Am. Mineral. 97, 1726–1740.
Dyar, M.D., Breves, E.A., Gunter, M.E., Lanzirotti, A., Tucker, J.M., Carey, C.J., Peel, S.E., Brown, E.B., Oberti, R., Lerotic, M., Delaney, J.S., 2016a. Use of multivariate analysis for synchrotron micro-XANES analysis of iron valence state in amphiboles. Am. Mineral. 101, 1171–1189.
Dyar, M.D., McCanta, M., Breves, E., Carey, C.J., Lanzirotti, A., 2016b. Accurate predictions of iron redox state in silicate glasses: a multivariate approach using X-ray absorption spectroscopy. Am. Mineral. 101, 744–747.
El Tabach, E., Lancelot, L., Shahrour, I., Najjar, Y., 2007. Use of artificial neural network simulation metamodelling to assess groundwater contamination in a road project. Math. Comput. Model. 45, 766–776.
El-Saeid, R.H., Abdel-Salam, Z., Pagnotta, S., Palleschi, V., Harith, M.A., 2019. Classification of sedimentary and igneous rocks by laser induced breakdown spectroscopy and nanoparticle-enhanced laser induced breakdown spectroscopy combined with principal component analysis and graph theory. Spectrochim. Acta B Atom Spectrosc. 158, 105622.
Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., Scholten, T., 2020. Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. Rem. Sens. 12, 2234.
Erickson, M.L., Elliott, S.M., Brown, C.J., Stackelberg, P.E., Ransom, K.M., Reddy, J.E., 2021. Machine learning predicted redox conditions in the Glacial aquifer system, Northern Continental United States. Water Resour. Res. 57, e2020WR028207.
Esmaeiloghli, S., Tabatabaei, S.H., 2020. Comparative analysis of geochemical data processing methods for allocation of anomalies and background. Geochem. Int. 58, 472–485.
Fernandes, M.M.H., Coelho, A.P., Fernandes, C., Silva, M.F.d., Dela Marta, C.C., 2019. Estimation of soil organic matter content by modeling with artificial neural networks. Geoderma 350, 46–51.
Flores, H., Lorenz, S., Jackisch, R., Tusa, L., Contreras, I.C., Zimmermann, R., Gloaguen, R., 2021. UAS-based hyperspectral environmental monitoring of acid mine drainage affected waters. Minerals 11.
Gaillardet, J., Dupré, B., Louvat, P., Allègre, C.J., 1999. Global silicate weathering and CO2 consumption rates deduced from the chemistry of large rivers. Chem. Geol. 159, 3–30.
Gao, C., Cao, X., Liu, Q., Yang, Y., Zhang, S., He, Y., Tang, M., Liu, Y., 2018. Theoretical calculation of equilibrium Mg isotope fractionations between minerals and aqueous solutions. Chem. Geol. 488, 62–75.
Ge, C., Huo, J., Gu, H.-O., Wang, F., Sun, H., Li, X., Li, W., Yuan, F., 2021. Tectonic discrimination and application based on convolution neural network and incomplete big data. J. Geochem. Explor. 220, 106662.
Gregory, D.D., Cracknell, M.J., Large, R.R., McGoldrick, P., Kuhn, S., Maslennikov, V.V., Baker, M.J., Fox, N., Belousov, I., Figueroa, M.C., Steadman, J.A., Fabris, A.J., Lyons, T.W., 2019. Distinguishing ore deposit type and barren sedimentary pyrite using laser ablation-inductively coupled plasma-mass spectrometry trace element data and statistical analysis of large data sets. Econ. Geol. 114, 771–786.
Ghezelbash, R., Maghsoudi, A., Bigdeli, A., Carranza, E.J.M., 2021. Regional-scale mineral prospectivity mapping: support vector machines and an improved data-driven multi-criteria decision-making technique. Nat. Resourc. Res. 30, 1977–2005.
Goddéris, Y., Schott, J., Brantley, S.L., 2019. Reactive transport models of weathering. Elements 15, 103–106.
Goldman, M.J., Vandewiele, N.M., Ono, S., Green, W.H., 2019. Computer-generated isotope model achieves experimental accuracy of filiation for position-specific isotope analysis. Chem. Geol. 514, 1–9.
Gonbadi, A.M., Tabatabaei, S.H., Carranza, E.J.M., 2015. Supervised geochemical anomaly detection by pattern recognition. J. Geochem. Explor. 157, 81–91.
Grunsky, E.C, Arne, D., 2020. Mineral-resource prediction Using advanced data analytics and machine learning of the QUEST-South Stream-Sediment Geochemical Data, Southwestern British Columbia, Canada. Geochem-Explor. Env. A. 21 (1) geochem2020-2054.
Grunsky, E.C., de Caritat, P., 2019. State-of-the-art analysis of geochemical data for mineral exploration. Geochem-Explor. Env. A. 20 (2), 217–232.
Guo, H., Goodchild, M.F., Annoni, A., 2020. Manual of Digital Earth. Springer, Singapore.
Han, J., Zhang, L., Car, R., 2018. Deep potential: a general representation of a many-body potential energy surface. Commun. Comput. Phys. 3, 629–639.
Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., von Lilienfeld, O. A., Tkatchenko, A., Müller, K.-R., 2013. Assessment and validation of machine learning methods for predicting molecular atomization energies. J. Chem. Theor. Comput. 9, 3404–3419.
Hao, H., Guo, R., Gu, Q., Hu, X., 2019. Machine learning application to automatically classify heavy minerals in river sand by using SEM/EDS data. Miner. Eng. 143, 105899.
Hasegawa, Y., Iwata, J.-I., Tsuji, M., Takahashi, D., Oshiyama, A., Minami, K., Boku, T., Shoji, F., Uno, A., Kurokawa, M., Inoue, H., Miyoshi, I., Yokokawa, M., 2011. First-principles calculations of electron states of a silicon nanowire with 100,000 atoms on the K computer. In: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. Association for Computing Machinery, New York, NY, USA, pp. 1–11.
He, H., Liu, Y., 2015. Silicon isotope fractionation during the precipitation of quartz and the adsorption of H4SiO4 (aq) on Fe (III)-oxyhydroxide surfaces. Chin. J. Geochem. 34, 459–468.
He, Y., Bao, H., 2019. Predicting high-dimensional isotope relationships from diagnostic fractionation factors in systems with diffusional mass transfer. ACS Earth Space Chem. 3, 120–128.
He, Y., Bao, H., Liu, Y., 2020. Predicting equilibrium intramolecular isotope distribution within a large organic molecule by the cutoff calculation. Geochem. Cosmochim. Acta 269, 292–302.
He, Y., Zhang, Y., Zhang, S., Liu, Y., 2021. Predicting nitrogen and oxygen kinetic isotope effects of nitrate reduction by periplasmic dissimilatory nitrate reductase. Geochem. Cosmochim. Acta 293, 224–239.
Hemmati-Sarapardeh, A., Amar, M.N., Soltanian, M.R., Dai, Z., Zhang, X., 2020. Modeling CO2 solubility in water at high pressure and temperature conditions. Energy Fuels 34, 4761–4776.

Hengl, T., de Jesus, J.M., Heuvelink, G.B., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., 2017. SoilGrids250m: global gridded soil information based on machine learning. PLoS One 12, e0169748.
Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma 214, 141–154.
Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77.
Hopfield, J.J., Feinstein, D., Palmer, R., 1983. 'Unlearning' has a stabilizing effect in collective memories. Nature 304, 158.
Hong, S., Zuo, R., Huang, X., Xiong, Y., 2021. Distinguishing IOCG and IOA deposits via random forest algorithm based on magnetite composition. J. Geochem. Explor. 230, 106859.
Huang, W.-H., Lyu, Y., Du, M.-H., He, C., Gao, S.-d., Xu, R.-j., Xia, Q., ZhangZhou, J., 2022. Estimating Ferric Iron Content in Clinopyroxene Using Machine Learning Models. American Mineralogist (in press).
Huang, X.-W., Sappin, A.-A., Boutroy, É., Beaudoin, G., Makvandi, S., 2019. Trace element composition of igneous and hydrothermal magnetite from porphyry deposits: relationship to deposit subtypes and magmatic affinity. Econ. Geol. 114, 917–952.
Iglesias, C., Antunes, I.M.H.R., Albuquerque, M.T.D., Martínez, J., Taboada, J., 2020. Predicting ore content throughout a machine learning procedure – an Sn-W enrichment case study. J. Geochem. Explor. 208, 106405.
Jacobson, A.D., Blum, J.D., 2003. Relationship between mechanical erosion and atmospheric CO2 consumption in the New Zealand Southern Alps. Geology 31, 865–868.
Ji, G., Ye, P., Shi, Y., Yuan, L., Chen, X., Yuan, M., Zhu, D., Chen, X., Hu, X., Jiang, J., 2017. Laser-induced breakdown spectroscopy for rapid discrimination of heavy-metal-contaminated seafood Tegillarca granosa. Sensors 17.
Jia, W., Wang, H., Chen, M., Lu, D., Lin, L., Car, R., Weinan, E., Zhang, L., 2020. Pushing the Limit of Molecular Dynamics with Ab Initio Accuracy to 100 Million Atoms with Machine Learning, SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1–14.
Johansson, A., Lehnert, K., Hsu, L., 2012. Status Report on the SedDB Sediment Geochemistry Database. March.
Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349 (6245), 255–260.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589.
Kates-Harbeck, J., Svyatkovskiy, A., Tang, W., 2019. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. Nature 568, 526–531.
Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40–58.
Keykhay-Hosseinpoor, M., Kohsary, A.-H., Hossein-Morshedy, A., Porwal, A., 2020. A machine learning-based approach to exploration targeting of porphyry Cu-Au deposits in the Dehsalm district, eastern Iran. Ore Geol. Rev. 116, 103234.
Khorshidi, N., Parsa, M., Lentz, D., Sobhanverdi, J., 2021. Identification of heavy metal pollutionsources and its associated risk assessment in an industrial town using theK-means clustering technique. Appl. Geochem. 135.
Kim, J.J., Ling, F.T., Plattenberger, D.A., Clarens, A.F., Lanzirotti, A., Newville, M., Peters, C.A., 2021. SMART mineral mapping: synchrotron-based machine learning approach for 2D characterization with coupled micro XRF-XRD. Comput. Geosci. 156, 104898.
Kirkwood, C., Cave, M., Beamish, D., Grebby, S., Ferreira, A., 2016. A machine learning approach to geochemical mapping. J. Geochem. Explor. 167, 49–61.
Koch, J., Stisen, S., Refsgaard, J.C., Ernstsen, V., Jakobsen, P.R., Højberg, A.L., 2019. Modeling depth of the redox interface at high resolution at National scale using random forest and residual Gaussian simulation. Water Resour. Res. 55, 1451–1469.
Kohonen, T., 1988. An introduction to neural computing. Neural Network. 1, 3–16.
Korokhin, V.V., Kaydash, V.G., Shkuratov, Y.G., Stankevich, D.G., Mall, U., 2008. Prognosis of TiO2 abundance in lunar soil using a non-linear analysis of Clementine and LSCC data. Planet. Space Sci. 56, 1063–1078.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.
Kuhn, S., Cracknell, M.J., Reading, A.M., Sykora, S., 2020. Identification of intrusive lithologies in volcanic terrains in British Columbia by machine learning using random forests: the value of using a soft classifier. Geophysics 85, B249–B258.
Kuwatani, T., Nagata, K., Okada, M., Watanabe, T., Ogawa, Y., Komai, T., Tsuchiya, N., 2014. Machine-learning techniques for geochemical discrimination of 2011 Tohoku tsunami deposits. Sci. Rep. 4, 7077.
Lanzirotti, A., Darby Dyar, M., Sutton, S., Newville, M., Head, E., Carey, C.J., McCanta, M., Lee, L., King, P.L., Jones, J., 2018. Accurate predictions of microscale oxygen barometry in basaltic glasses using V K-edge X-ray absorption spectroscopy: a multivariate approach. Am. Mineral. 103, 1282–1297.
Laloy, E., Jacques, D., 2019. Emulation of CPU-demanding reactive transport models: a comparison of Gaussian processes, polynomial chaos expansion, and deep neural networks. Comput. Geosci. 23, 1193–1215.
Le Losq, C., Valentine, A.P., Mysen, B.O., Neuville, D.R., 2021. Structure and properties of alkali aluminosilicate glasses and melts: insights from deep learning. Geochem. Cosmochim. Acta 314, 27–54.
Le Losq, C., Berry, A.J., Kendrick, M.A., Neuville, D.R., O'Neill, H.S.C., 2019. Determination of the oxidation state of iron in Mid-Ocean Ridge basalt glasses by Raman spectroscopy. Am. Mineral. 104, 1032–1042.
Lee, K.J., 2020. Characterization of kerogen content and activation energy of decomposition using machine learning technologies in combination with numerical simulations of formation heating. J. Petrol. Sci. Eng. 188, 106860.
Lehnert, K., Su, Y., Langmuir, C.H., Sarbas, B., Nohl, U., 2000. A global geochemical database structure for rocks. Geochem. Geophys. Geosys. 1, 1–14.
Lek, S., Guiresse, M., Giraudel, J.L., 1999. Predicting stream nitrogen concentration from watershed features using neural networks. Water Res. 33, 3469–3478.
Li, B., Lee, Y., Yao, W., Lu, Y., Fan, X., 2020a. Development and application of ANN model for property prediction of supercritical kerosene. Comput. Fluids 209, 104665.
Li, C., Arndt, N.T., Tang, Q., Ripley, E.M., 2015. Trace element indiscrimination diagrams. Lithos 232, 76–83.
Li, L., He, Y., Zhang, Z., Liu, Y., 2021a. Nitrogen isotope fractionations among gaseous and aqueous NH4+, NH3, N2, and metal-ammine complexes: theoretical calculations and applications. Geochem. Cosmochim. Acta 295, 80–97.
Li, L., Maher, K., Navarre-Sitchler, A., Druhan, J., Meile, C., Lawrence, C., Moore, J., Perdrial, J., Sullivan, P., Thompson, A., 2017. Expanding the role of reactive transport models in critical zone processes. Earth Sci. Rev. 165, 280–301.
Li, Q.-L., Zhou, Q., Liu, Y., Xiao, Z., Lin, Y., Li, J.-H., Ma, H.-X., Tang, G.-Q., Guo, S., Tang, X., Yuan, J.-Y., Li, J., Wu, F.-Y., Ouyang, Z., Li, C., Li, X.-H., 2021b. Two billion-year-old volcanism on the Moon from Chang'E-5 basalts. Nature 600, 54–58.
Li, S., Chen, J., Xiang, J., 2020b. Applications of deep convolutional neural networks in prospecting prediction based on two-dimensional geological big data. Neural Comput. Appl. 32, 2037–2053.
Li, X., Zhou, Y., Wang, J., Ye, M., Geng, T., 2020c. Contrasting granite metallogeny through the zircon REE composition: perspective from data mining. Appl. Geochem. 122, 104758.
Lin, X., Chang, H., Wang, K., Zhang, G., Meng, G., 2020. Machine learning for source identification of dust on the Chinese loess plateau. Geophys. Res. Lett. 47, e2020GL088950.
Lin, N., Chen, Y., Liu, H., Liu, H., 2021. A comparative study of machine learning models with hyperparameter optimization algorithm for mapping mineral prospectivity. Minerals 11, 159.
Lintern, A., Webb, J.A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., Leahy, P., Wilson, P., Western, A.W., 2018a. Key factors influencing differences in stream water quality across space. WIREs Water 5, e1260.
Lintern, A., Webb, J.A., Ryu, D., Liu, S., Waters, D., Leahy, P., Bende-Michl, U., Western, A.W., 2018b. What are the key catchment characteristics affecting spatial differences in riverine water quality? Water Resour. Res. 54, 7252–7272.
Lischeid, G., 2001. Investigating short-term dynamics and long-term trends of SO4 in the runoff of a forested catchment using artificial neural networks. J. Hydrol. 243, 31–42.
Liu, Y., Huang, C., Zhou, Y., Lu, Y., Ma, Q., 2020. The controlling factors of lacustrine shale lithofacies in the Upper Yangtze Platform (South China. using artificial neural networks. Mar. Petrol. Geol. 118, 104350.
Lopez, A.M., Wells, A., Fendorf, S., 2021. Soil and aquifer properties combine as predictors of groundwater Uranium concentrations within the central valley, California. Environ. Sci. Technol. 55, 352–361.
Luo, H., Karki, B.B., Ghosh, D.B., Bao, H., 2020a. First-principles computation of diffusional Mg isotope fractionation in silicate melts. Geochem. Cosmochim. Acta 290, 27–40.
Luo, Z., Xiong, Y., Zuo, R., 2020b. Recognition of geochemical anomalies using a deep variational autoencoder network. Appl. Geochem. 122, 104710.
Maher, K., Navarre-Sitchler, A., 2019. Reactive transport processes that drive chemical weathering: from making space for water to dismantling Continents. Rev. Mineral. Geochem. 85, 349–380.
Mahmoudzadeh, H., Matinfar, H.R., Taghizadeh-Mehrjardi, R., Kerry, R., 2020. Spatial prediction of soil organic carbon using machine learning techniques in western Iran. Geoderma Reg. 21, e00260.
Makvandi, S., Ghasemzadeh-Barvarz, M., Beaudoin, G., Grunsky, E.C., McClenaghan, M. B., Duchesne, C., Boutroy, E., 2016. Partial least squares-discriminant analysis of trace element compositions of magnetite from various VMS deposit subtypes: application to mineral exploration. Ore Geol. Rev. 78, 388–408.
Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environ. Model. Software 25, 891–909.
McCanta, M.C., Dyar, M.D., Rutherford, M.J., Lanzirotti, A., Sutton, S.R., Thomson, B.J., 2017. In situ measurement of ferric iron in lunar glass beads using Fe-XAS. Icarus 285, 95–102.
McClelland, J., Rnmelhart, D., 1986. Parallel distributed processing: explorations in the microstructure of cognition. In: Psychological and Biological Models, vol. 2. MIT Press, Cambridge, MA.
McGibbon, R.T., Taube, A.G., Donchev, A.G., Siva, K., Hernández, F., Hargus, C., Law, K.-H., Klepeis, J.L., Shaw, D.E., 2017. Improving the accuracy of Møller-Plesset perturbation theory with neural networks. J. Chem. Phys. 147, 161725.
Meliker, J.R., AvRuskin, G.A., Slotnick, M.J., Goovaerts, P., Schottenfeld, D., Jacquez, G. M., Nriagu, J.O., 2008. Validity of spatial models of arsenic concentrations in private well water. Environ. Res. 106, 42–50.
Michalski, R.S., Baskin, A.B., 1983. Integrating Multiple Knowledge Representations and Learning Capabilities in an Expert System: the ADVISE System. IJCAI.
Mikkonen, H.G., van de Graaff, R., Clarke, B.O., Dasika, R., Wallis, C.J., Reichman, S.M., 2018. Geochemical indices and regression tree models for estimation of ambient

background concentrations of copper, chromium, nickel and zinc in soil. Chemosphere 210, 193–203.
Minsky, M., Papert, S.A., 2017. Perceptrons: an Introduction to Computational Geometry. MIT press, Cambridge, MA.
Mjolsness, E., DeCoste, D., 2001. Machine learning for science: state of the art and future prospects. Science 293, 2051–2055.
Mohammadi, N.M., Hezarkhani, A., 2018. Application of support vector machine for the separation of mineralised zones in the Takht-e-Gonbad porphyry deposit, SE Iran. J. Afr. Earth Sci. 143, 301–308.
Mojid, M.A., Hossain, A.B.M.Z., Ashraf, M.A., 2019. Artificial neural network model to predict transport parameters of reactive solutes from basic soil properties. Environ. Pollut. 255.
Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., Müller, K.-R., Anatole von Lilienfeld, O., 2013. Machine learning of molecular electronic properties in chemical compound space. New J. Phys. 15, 095003.
Nemes, A., Rawls, W., Pachepsky, Y.A., Van Genuchten, M.T., 2006a. Sensitivity analysis of the nonparametric nearest neighbor technique to estimate soil water retention. Vadose Zone J. 5, 1222–1235.
Nemes, A., Rawls, W.J., Pachepsky, Y.A., 2006b. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. Soil Sci. Soc. Am. J. 70, 327–336.
Nolan, B.T., Fienen, M.N., Lorenz, D.L., 2015. A statistical learning framework for groundwater nitrate models of the Central Valley, California, USA. J. Hydrol. 531, 902–911.
Nolan, B.T., Green, C.T., Juckem, P.F., Liao, L., Reddy, J.E., 2018. Metamodeling and mapping of nitrate flux in the unsaturated zone and groundwater, Wisconsin, USA. J. Hydrol. 559, 428–441.
Normile, D., 2019. Earth scientists plan a 'geological Google'. Science 363, 917.
O'Sullivan, G., Chew, D., Kenny, G., Henrichs, I., Mulligan, D., 2020. The trace element composition of apatite and its application to detrital provenance studies. Earth Sci. Rev. 201, 103044.
Oh, H.-J., Lee, S., 2010. Application of artificial neural network for gold–silver deposits potential mapping: a case study of Korea. Nat. Resourc. Res. 19, 103–124.
Othman, A.A., Gloaguen, R., 2017. Integration of spectral, spatial and morphometric data into lithological mapping: a comparison of different Machine Learning Algorithms in the Kurdistan Region, NE Iraq. J. Asian Earth Sci. 146, 90–102.
Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning to predict soil properties from regional spectral data. Geoderma Reg. 16, e00198.
Pan, C., Ng, K.T.W., Fallah, B., Richter, A., 2019. Evaluation of the bias and precision of regression techniques and machine learning approaches in total dissolved solids modeling of an urban aquifer. Environ. Sci. Pollut. Res. 26, 1821–1833.
Parsa, M., 2021. A data augmentation approach to XGboost-based mineral potential mapping: an example of carbonate-hosted ZnPb mineral systems of Western Iran. J. Geochem. Explor. 228, 106811.
Peterson, K.T., Sagan, V., Sidike, P., Hasenmueller, E.A., Sloan, J.J., Knouft, J.H., 2019. Machine learning-based ensemble prediction of water-quality variables using feature-level and decision-level fusion with proximal remote sensing. Photogramm. Eng. Rem. Sens. 85, 269–280.
Petrelli, M., Bizzarri, R., Morgavi, D., Baldanza, A., Perugini, D., 2017. Combining machine learning techniques, microanalyses and large geochemical datasets for tephrochronological studies in complex volcanic areas: new age constraints for the Pleistocene magmatism of central Italy. Quat. Geochronol. 40, 33–44.
Petrelli, M., Caricchi, L., Perugini, D., 2020. Machine learning thermo-barometry: application to clinopyroxene-bearing magmas. J. Geophys. Res. Solid Earth 125, e2020JB020130.
Petrelli, M., Perugini, D., 2016. Solving petrological problems through machine learning: the study case of tectonic discrimination using geochemical and isotopic data. Contrib. Mineral. Petrol. 171, 81.
Pfau, D., Spencer, J.S., Matthews, A.G., Foulkes, W.M.C., 2020. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. Phys. Rev. Res. 2 (3), 033429.
Podgorski, J., Wu, R., Chakravorty, B., Polya, D.A., 2020. Groundwater arsenic distribution in India by machine learning geospatial modeling. Int. J. Environ. Res. Publ. Health 17, 7119.
Prasianakis, N.I., Haller, R., Mahrous, M., Poonoosamy, J., Pfingsten, W., Churakov, S.V., 2020. Neural network based process coupling and parameter upscaling in reactive transport simulations. Geochem. Cosmochim. Acta 291, 126–143.
Prettyman, T.H., Hagerty, J.J., Elphic, R.C., Feldman, W.C., Lawrence, D.J., McKinney, G.W., Vaniman, D.T., 2006. Elemental composition of the lunar surface: analysis of gamma ray spectroscopy data from Lunar Prospector. J. Geophys. Res.: Planets 111, E12007.
Ransom, K.M., Bell, A.M., Barber, Q.E., Kourakos, G., Harter, T., 2018. A Bayesian approach to infer nitrogen loading rates from crop and land-use types surrounding private wells in the Central Valley, California. Hydrol. Earth Syst. Sci. 22, 2739–2758.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204.
Ren, Q., Li, M., Han, S., Zhang, Y., Zhang, Q., Shi, J., 2019. Basalt tectonic discrimination using combined machine learning approach. Minerals 9, 376.
Rnmelhart, D., McClelland, J., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations. MIT Press, Cambridge, MA.
Rodriguez-Galiano, V., Chica-Olmo, M., Chica-Rivas, M., 2014. Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain. Int. J. Geogr. Inf. Sci. 28, 1336–1354.
Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818.
Roshanravan, B., 2020. Translating a mineral systems model into continuous and data-driven targeting models: an example from the Dolatabad chromite district, southeastern Iran. J. Geochem. Explor. 215, 106556.
Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. 65 (6), 386.
Rupp, M., Tkatchenko, A., Müller, K.-R., von Lilienfeld, O.A., 2012. Fast and accurate modeling of molecular atomization energies with machine learning. Phys. Rev. Lett. 108, 058301.
Rustad, J.R., 2009. Ab initio calculation of the carbon isotope signatures of amino acids. Org. Geochem. 40, 720–723.
Sahoo, G.B., Ray, C., De Carlo, E.H., 2006. Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii. J. Hydrol. 327, 525–538.
Samuel, A.L., 1959. Some studies in machine learning using the game of checkers. IBM J. Res. Dev. 3 (3), 210–229.
Sarbas, B., Nohl, U., 2009. The GEOROC database - a decade of "online geochemistry. Geochem. Cosmochim. Acta Suppl. 73, A1158.
Schaffer, C., 1993. Overfitting avoidance as bias. Mach. Learn. 10 (2), 153–178.
Schütt, K.T., Arbabzadah, F., Chmiela, S., Müller, K.R., Tkatchenko, A., 2017. Quantum-chemical insights from deep tensor neural networks. Nat. Commun. 8, 13890.
Sergeev, A., Buevich, A., Baglaeva, E., Shichkin, A., 2019. Combining spatial autocorrelation with machine learning increases prediction accuracy of soil heavy metals. Catena 174, 425–435.
Sergeev, E.A., Sokoloff, V.P., 1941. Geochemical Method of Prospecting for Ore Deposits (No. 48-21-B). USGPO.
Shahnas, M., Yuen, D., Pysklywec, R., 2018. Inverse problems in Geodynamics using machine learning algorithms. J. Geophys. Res. Solid Earth 123, 296–310.
Shaughnessy, A.R., Gu, X., Wen, T., Brantley, S.L., 2021. Machine learning deciphers CO2 sequestration and subsurface flowpaths from stream chemistry. Hydrol. Earth Syst. Sci. 25, 3397–3409.
Shen, J., Kuo, A.Y., 1998. Application of inverse method to calibrate estuarine eutrophication model. J. Environ. Eng. 124, 409–418.
Shen, L.Q., Amatulli, G., Sethi, T., Raymond, P., Domisch, S., 2020. Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. Sci. Data 7, 161.
Sheng, L., Zhang, T., Niu, G., Wang, K., Tang, H., Duan, Y., Li, H., 2015. Classification of iron ores by laser-induced breakdown spectroscopy (LIBS) combined with random forest (RF). J. Anal. At. Spectrom. 30, 453–458.
Silburt, A., Ali-Dib, M., Zhu, C., Jackson, A., Valencia, D., Kissin, Y., Tamayo, D., Menou, K., 2019. Lunar crater identification via deep learning. Icarus 317, 27–38.
Snodgrass, J.E., Milkov, A.V., 2020. Web-based machine learning tool that determines the origin of natural gases. Comput. Geosci. 145, 104595.
Spear, F.S., Hallett, B., Pyle, J.M., Adalı, S., Szymanski, B.K., Waters, A., Linder, Z., Pearce, S.O., Fyffe, M., Goldfarb, D., Glickenhouse, N., Buletti, H., 2009. MetPetDB: a database for metamorphic geochemistry. Geochem. Geophys. Geosys. 10.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.
Stackelberg, P.E., Belitz, K., Brown, C.J., Erickson, M.L., Elliott, S.M., Kauffman, L.J., Ransom, K.M., Reddy, J.E., 2021. Machine learning predictions of pH in the Glacial aquifer system, Northern USA. Groundwater 59, 352–368.
Stern, M.J., Wolfsberg, M., 1966. Simplified procedure for the theoretical calculation of isotope effects involving large molecules. J. Chem. Phys. 45, 4105–4124.
Suen, J.-P., Eheart, J.W., 2003. Evaluation of neural networks for modeling nitrate concentrations in rivers. J. Water Resour. Plann. Manag. 129, 505–510.
Sun, C., Tian, Y., Gao, L., Niu, Y., Zhang, T., Li, H., Zhang, Y., Yue, Z., Delepine-Gilon, N., Yu, J., 2019. Machine learning allows calibration models to predict trace element concentration in soils with generalized LIBS spectra. Sci. Rep. 9, 11363.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. Adv. NIPS 27, 1–8.
Sutton, S.R., Lanzirotti, A., Newville, M., Dyar, M.D., Delaney, J., 2020. Oxybarometry and valence quantification based on microscale X-ray absorption fine structure (XAFS) spectroscopy of multivalent elements. Chem. Geol. 531, 119305.
Swinyard, B.M., Joy, K.H., Kellett, B.J., Crawford, I.A., Grande, M., Howe, C.J., Fernandes, V.A., Gasnault, O., Lawrence, D.J., Russell, S.S., Wieczorek, M.A., Foing, B.H., 2009. X-ray fluorescence observations of the moon by SMART-1/D-CIXS and the first detection of Ti Kα from the lunar surface. Planet. Space Sci. 57, 744–750.
Szatmári, G., Pásztor, L., 2019. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma 337, 1329–1340.
Tesoriero, A.J., Gronberg, J.A., Juckem, P.F., Miller, M.P., Austin, B.P., 2017. Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. Water Resour. Res. 53, 7316–7331.
Tian, M., Wang, X., Nie, L., Liu, H., Wang, W., Yan, T., 2019. Spatial distributions and the identification of ore-related anomalies of Cu across the boundary area of China and Mongolia. J. Geochem. Explor. 197, 37–47.
Tomiyama, T., Ichiyama, Y., Horikawa, H., Sato, Y., Soma, S., Hanafusa, Y., 2013. GANSEKI: JAMSTEC Deep Seafloor Rock Sample Database Emerging to the New Phase. AGU Fall Meeting Abstracts, 1626.
Tran, D.A., Tsujimura, M., Ha, N.T., Nguyen, V.T., Binh, D.V., Dang, T.D., Doan, Q.-V., Bui, D.T., Anh Ngoc, T., Phu, L.V., Thuc, P.T.B., Pham, T.D., 2021. Evaluating the predictive power of different machine learning algorithms for groundwater salinity

prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam. Ecol. Indicat. 127, 107790.
Tsuchiyama, A., Kawamura, K., Nakao, T., Uyeda, C., 1994. Isotopic effects on diffusion in MgO melt simulated by the molecular dynamics (MD) method and implications for isotopic mass fractionation in magmatic systems. Geochem. Cosmochim. Acta 58, 3013–3021.
Tut Haklidir, F.S., Haklidir, M., 2020. Prediction of reservoir temperatures using hydrogeochemical data, western Anatolia geothermal systems (Turkey): a machine learning approach. Nat. Resourc. Res. 29, 2333–2346.
Tziolas, N., Tsakiridis, N., Ben-Dor, E., Theocharis, J., Zalidis, G., 2019. A memory-based learning approach utilizing combined spectral sources and geographical proximity for improved VIS-NIR-SWIR soil properties estimation. Geoderma 340, 11–24.
Uddameri, V., Silva, A.L., Singaraju, S., Mohammadi, G., Hernandez, E.A., 2020. Tree-based modeling methods to predict nitrate exceedances in the Ogallala aquifer in Texas. Water 12, 1023.
Ueki, K., Hino, H., Kuwatani, T., 2018. Geochemical discrimination and characteristics of magmatic tectonic settings: a machine-learning-based approach. Geochem. Geophys. Geosys. 19, 1327–1347.
Valetich, M.J., Le Losq, C., Arculus, R.J., Umino, S., Mavrogenes, J., 2021. Compositions and classification of fractionated boninite series melts from the Izu–Bonin–Mariana arc: a machine learning approach. J. Petrol. 62, egab013.
Vermeesch, P., 2006. Tectonic discrimination of basalts with classification trees. Geochem. Cosmochim. Acta 70, 1839–1848.
Viviano, C.E., Moersch, J.E., 2013. Using THEMIS data to resolve the discrepancy between CRISM/OMEGA and TES modeled phyllosilicate abundance in Mawrth Vallis. Icarus 226, 497–509.
Wadoux, A.M.J.C., Padarian, J., Minasny, B., 2019. Multi-source data integration for soil mapping using deep learning. Soil 5, 107–119.
Walker, J.D., Bowers, T.D., Black, R.A., Glazner, A.F., Farmer, G.L., Carlson, R.W., Sinha, A.K., 2006. A geochemical database for western North American volcanic and intrusive rocks (NAVDAT). Spec. Pap. Geol. Soc. Am. 397, 61.
Wang, C., Hazen, M.R., Cheng, Q., Stephenson, H.M., Zhou, C., Fox, P., Shen, S.-Z., Oberhänsli, R., Hou, Z., Ma, X., Feng, Z., Fan, J., Ma, C., Hu, X., Luo, B., Wang, J., Schiffries, C.M., 2021a. The deep-time digital earth program: data-driven discovery in geosciences. Natl. Sci. Rev. 8 (9), nwab027.
Wang, L., Su, C., Wang, L.-Q., ZhangZhou, J., Xia, Q.-K., Wang, Q.-Y., 2022. A Refined Estimation of Li in Mica by a Machine Learning Method. American Mineralogist (in press).
Wang, X., Niu, R., 2012. Lunar titanium abundance characterization using Chang'E-1 IIM data. Sci. China Phys. Mech. Astron. 55, 170–178.
Wang, Y., Qiu, K.-F., Müller, A., Hou, Z.-L., Zhu, Z.-H., Yu, H.-C., 2021b. Machine learning prediction of quartz forming-environments. J. Geophys. Res. Solid Earth 126, e2021JB021925.
Wang, Z., Yan, C., Dong, J., Zhang, T., Wei, J., Li, H., 2016. Acidity analysis of iron ore based on calibration-free laser-induced breakdown spectroscopy (CF-LIBS) combined with a binary search algorithm (BSA). RSC Adv. 6, 76813–76823.
Wang, H., Zhang, L., Han, J., E, W., 2018. DeePMD-kit: a deep learning package for many-body potential energy representation and molecular dynamics. Comput. Phys. Commun. 228, 178–184.
Wang, J., Zhou, Y., Xiao, F., 2020. Identification of multi-element geochemical anomalies using unsupervised machine learning algorithms: a case study from Ag–Pb–Zn deposits in north-western Zhejiang, China. Appl. Geochem. 120, 104679.
Wang, X., Zhu, P., 2013. Refinement of lunar TiO2 analysis with multispectral features of Chang'E-1 IIM data. Astrophys. Space Sci. 343, 33–44.
Wen, T., Bandaragoda, C., Harris, L., 2020. Data Science in Earth and Environmental Sciences (v1.0). Zenodo. https://doi.org/10.5281/zenodo.6399063.
Wen, T., Liu, M., Woda, J., Zheng, G., Brantley, S.L., 2021. Detecting anomalous methane in groundwater within hydrocarbon production areas across the United States. Water Res. 200, 117236.
Wu, G., Chen, G., Cheng, Q., Zhang, Z., Yang, J., 2021. Unsupervised machine learning for lithological mapping using geochemical data in covered areas of Jining, China. Nat. Resourc. Res. 30, 1053–1068.
Wu, Y., 2012. Major elements and Mg# of the moon: results from Chang'E-1 interference imaging spectrometer (IIM) data. Geochem. Cosmochim. Acta 93, 214–234.
Wu, Z., Ramsundar, B., Feinberg, Evan N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K., Pande, V., 2018. MoleculeNet: a benchmark for molecular machine learning. Chem. Sci. 9, 513–530.
White, W.M., 1985. Sources of oceanic basalts: radiogenic isotopic evidence. Geology 13 (2), 115–118.
Xia, W., Wang, X., Zhao, S., Jin, H., Chen, X., Yang, M., Wu, X., Hu, C., Zhang, Y., Shi, Y., 2019. New maps of lunar surface chemistry. Icarus 321, 200–215.
Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X., Tu, M., 2018. Evaluation of machine learning methods for formation lithology identification: a comparison of tuning processes and model performances. J. Petrol. Sci. Eng. 160, 182–193.
Xiong, Y., Zuo, R., 2016. A comparative study of two modes for mapping felsic intrusions using geoinformatics. Appl. Geochem. 75, 277–283.
Yajima, H., Derot, J., 2018. Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. J. Hydroinf. 20, 206–220.
Yao, K., Herr, J.E., Parkhill, J., 2017. The many-body expansion combined with neural networks. J. Chem. Phys. 146, 014106.
Yan, C., Liang, J., Zhao, M., Zhang, X., Zhang, T., Li, H., 2019a. A novel hybrid feature selection strategy in quantitative analysis of laser-induced breakdown spectroscopy. Anal. Chim. Acta 1080, 35–42.
Yan, C., Qi, J., Ma, J., Tang, H., Zhang, T., Li, H., 2017. Determination of carbon and sulfur content in coal by laser induced breakdown spectroscopy combined with kernel-based extreme learning machine. Chemometr. Intell. Lab. Syst. 167, 226–231.
Yan, C., Zhang, T., Sun, Y., Tang, H., Li, H., 2019b. A hybrid variable selection method based on wavelet transform and mean impact value for calorific value determination of coal using laser-induced breakdown spectroscopy and kernel extreme learning machine. Spectrochim. Acta B Atom Spectrosc. 154, 75–81.
Yeomans, C.M., Shail, R.K., Grebby, S., Nykänen, V., Middleton, M., Lusty, P.A.J., 2020. A machine learning approach to tungsten prospectivity modelling using knowledge-driven feature extraction and model confidence. Geosci. Front. 11, 2067–2081.
Yoshida, K., Kuwatani, T., Yasumoto, A., Haraguchi, S., Ueki, K., Iwamori, H., 2018. GEOFCM: a new method for statistical classification of geochemical data using spatial contextual information. J. Mineral. Petrol. Sci. 113, 159–169.
Yue, Z., Di, K., Wan, W., Liu, Z., Gou, S., Liu, B., Peng, M., Wang, Y., Jia, M., Liu, J., Ouyang, Z., 2022. Updated lunar cratering chronology model with the radiometric age of Chang'e-5 samples. Nat. Astron. 1–5.
Zaremotlagh, S., Hezarkhani, A., 2017. The use of decision tree induction and artificial neural networks for recognizing the geochemical distribution patterns of LREE in the Choghart deposit, Central Iran. J. Afr. Earth Sci. 128, 37–46.
Zhang, S., Liu, Q., Tang, M., Liu, Y., 2020. Molecular-Level mechanism of phosphoric acid digestion of carbonates and recalibration of the 13C–18O clumped isotope thermometer. ACS Earth Space Chem. 4, 420–433.
Zhang, T., Yan, C., Qi, J., Tang, H., Li, H., 2017. Classification and discrimination of coal ash by laser-induced breakdown spectroscopy (LIBS) coupled with advanced chemometric methods. J. Anal. At. Spectrom. 32, 1960–1965.
Zhang, C., Zuo, R., Xiong, Y., 2021. Detection of the multivariate geochemical anomalies associated with mineralization using a deep convolutional neural network and a pixel-pair feature method. Appl. Geochem. 130, 104994.
Zhao, Y., Lamine Guindo, M., Xu, X., Sun, M., Peng, J., Liu, F., He, Y., 2019a. Deep learning associated with laser-induced breakdown spectroscopy (LIBS) for the prediction of lead in soil. Appl. Spectrosc. 73, 565–573.
Zhao, Y., Zhang, Y., Geng, M., Jiang, J., Zou, X., 2019b. Involvement of slab-derived fluid in the generation of cenozoic basalts in Northeast China inferred from machine learning. Geophys. Res. Lett. 46, 5234–5242.
Zhi, W., Feng, D., Tsai, W.P., Sterle, G., Harpold, A., Shen, C., Li, L., 2021. From hydrometeorology to river water quality: can a deep learning model predict dissolved oxygen at the Continental scale? Environ. Sci. Technol. 55, 2357–2368.
Zhong, R., Deng, Y., Li, W., Danyushevsky, L.V., Cracknell, M.J., Belousov, I., Chen, Y., Li, L., 2021. Revealing the multi-stage ore-forming history of a mineral deposit using pyrite geochemistry and machine learning-based data interpretation. Ore Geol. Rev. 133, 104079.
Zhou, You, Liu, Yun, Christian, Reinhardt, Deng, Hongping, 2021. The core-merging giant impact in Earth's accretion history and itsimplications. Acta Geochim. 1–15.
Zou, R., Lung, W.S., Wu, J., 2007. An adaptive neural network embedded genetic algorithm approach for inverse water quality modeling. Water Resour. Res. 43, W08427.
Zuo, R., Cheng, Q., 2008. Mapping singularities—a technique to identify potential Cu mineral deposits using sediment geochemical data, an example for Tibet, west China. Mineral. Mag. 72 (1), 531–534.