Natural Resources Research (© 2025)
https://doi.org/10.1007/s11053-025-10568-w

Original Paper

Uncertainty-Aware Deep Neural Network Training for Imbalanced Geochemical Data Distributions

Ali Dashti[1,3] **Michael Trumpp,**[1] **Lars H. Ystroem,**[1] **Valentin Goldberg,**[1,2] **Nancy Seimetz,**[1] **and Fabian Nitschke**[1]

Received 16 December 2024; accepted 25 October 2025

The growing interest in raw material extraction, particularly in trace elements, highlights the need for innovative geochemical modeling techniques to predict element concentrations accurately. This paper explores the predictive capabilities of a deep neural network (DNN) in estimating the concentrations of 20 trace elements based on 11 major elements and pH values. Using data from the BrineMine project, we applied DNNs to a challenging dataset characterized by a small sample size and imbalanced distributions. In total, 1000 independent DNN models were generated to address prediction accuracy and uncertainty instead of relying on a single model. Two preprocessing methods, including synthetic minority oversampling technique for regression with Gaussian noise (SMOGN) statistical transformation, were applied to improve the accuracy and decrease uncertainty further. Despite issues such as low initial correlations between input features and target variables, imbalanced data distributions, and extremely low concentrations, the DNN models provided reliable and robust results, except for Cu and V. For 13 trace elements, the DNN models achieved acceptable reliability with $R^2 > 0.8$. Analyzing the weight distribution of the DNN revealed that input features with high cross-correlation are prone to sharing the same information. While input features such as Fe, pH, and Mg are highly correlated to several target variables, accumulated local effects (ALE) scores indicate that Li has the highest influence, as it is the only input feature with a high correlation coefficient to some of the target variables.

KEY WORDS: Deep neural network, Trace element, Uncertainty, BrineMine, Data distribution, ALE values.

INTRODUCTION

Artificial intelligence (AI) and its subsets, such as machine learning (ML), artificial neural networks (ANNs), and deep neural networks (DNNs), have been extensively utilized across a broad range of applications (Brunton & Kutz, 2022; Rabczuk & Bathe, 2023). ML is typically defined as computer programs that learn from data (Jo 2021) and are expected to become an industry standard, replacing traditional analytical and numerical solutions. As one of the most promising branches of ML, ANNs consist of networks of neuron-like units that are optimized by minimizing the loss between the true and predicted values (Kanwisher et al., 2023). ANNs generally comprise three layers: input, hidden, and output. An ANN paradigm can have different architectures resulting in variations such as recurrent

Nancy Seimetz: Formerly at Institute of Applied Geosciences, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany.

[1]Institute of Applied Geosciences, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany.
[2]BWG Geochemische Beratung GmbH, Seestraße 7 A, 17033 Neubrandenburg, Germany.
[3]To whom correspondence should be addressed; e-mail: Ali.dashti@kit.edu

Published online: 08 November 2025

neural networks (RNNs) and convolutional neural networks (CNNs) (LeCun et al., 2015). Each variant is designed to address specific problems: for example, CNNs are widely used in computer vision tasks (Zhao et al., 2024b), while RNN is a state-of-the-art technology dealing with sequential data such as time series, letters constructing words, and sentences (Salehinejad et al., 2018). The number of hidden layers distinguishes two types of models: shallow neural networks with a single hidden layer and deep neural networks (DNNs) with multiple hidden layers. The latter architecture allows the network to understand more complex relations (Kufel et al., 2023). The multiple hidden layers of DNN enable the model to comprehend the hierarchical feature representations of data and yield balanced weight matrices (LeCun et al., 2015). Although DNNs impose higher computational costs, recent advances in graphics processing units (GPUs) have made it feasible to operate computational tasks efficiently (Wang et al., 2019). In particular, DNNs rely heavily on simple dot products, which can be parallelized effectively.

The accelerated expansion of ML in a wide variety of applications, including the geothermal sector, has been remarkable (Okoroafor et al., 2022). Several studies applied ML methods for mineral exploration and resource estimation from geochemical data (Oh and Lee 2010; Dornan et al., 2020; Enkhsaikhan et al., 2021; Liu et al., 2022; Zhang et al., 2023; Ngombe et al., 2024). Most research utilized shallow NNs, or methods such as K-means clustering, regression trees, and so on, for the existing geochemical data from ore deposits/mines (Puzyrev et al., 2023). Recently, ML algorithms have been deployed for re-purposing the legacy multi-element geochemical data to gain advanced knowledge, guide exploration, and detect anomalies (Zhang et al., 2022; Bourdeau et al., 2023). For example, Zhang et al. (2021) used the residual maps to characterize possible geochemical anomalies. The development of data-driven ML models for predicting geospatial information, e.g., mineral mapping, can be prone to different types of uncertainties (Zhang et al., 2024b), noteworthy to be mentioned in such kind of studies. The application of DNNs to predict the concentrations of multiple elements in geothermal brines remains largely unexplored (Drumm et al., 2023). Ystroem et al. (2023) and Gurgenc et al. (2024) demonstrated the effectiveness of ANNs in correlating aqueous element concentrations with subsurface temperatures across different geothermal settings. A key challenge in developing robust DNN models is the acquisition of sufficiently high-quality hydrogeochemical data from geothermal brines, due to the extremely challenging sampling techniques required (Arnórsson et al., 2006) and the site-specific geochemical settings.

Geothermal brine has a highly sensitive thermodynamic equilibrium (Hawkins & Tester, 2018) and requires delicate sampling to ensure the reliability of field campaigns. Changes in system parameters, such as temperature, pressure, and pH, can lead to significant responses in the chemical system, e.g., precipitation, dissolution, or degassing, which in turn alter the chemical composition (Bourcier et al., 2005; Cosmo et al., 2022). A range of technical challenges contributes to the scarcity of geochemical data from geothermal brines, as highlighted by Arnórsson et al. (2006). Firstly, access is often limited, with sampling only feasible through fluid circulation in boreholes. Representative downhole sampling techniques, such as positive displacement downhole samples, are costly and can only be conducted directly after the completion of the wells or during workovers while the pump is dismantled. During regular geothermal operations, fluid can interact with the extraction equipment before even reaching the sampler (Gunnlaugsson et al., 2014; Nitschke et al., 2017; Wanner et al., 2017). Secondly, labor and equipment costs associated with sampling campaigns can be prohibitively high. Thirdly, due to the site-specific (and even depth-specific) nature of concentrations, the analytical data are rarely comparable, making it difficult to obtain a representative geochemical dataset. Lastly, a range of interrelated factors—such as phase, pH, Eh, flow rate, temperature, and pressure—govern concentration ranges, further complicating system interpretation.

Data shortage and lack of precision can be mentioned as the main outcomes of this highly complex acquisition process (Zhang et al., 2024a). In some cases, even a dataset may exist; however, the unfavorable highly skewed (imbalanced) multimodal distributions for some elements hinder any further investigation. The size and representativeness of a dataset play a crucial role in the performance and reliability of machine learning models (Drumm et al., 2023; Al-Fakih et al., 2024; Zhao et al., 2024a; Mohammed et al., 2025). When trained on small or unbalanced datasets, ML models often exhibit higher predictive uncertainty due to limited exposure to data variability. This is particularly

critical in scientific and engineering applications where data collection is costly or limited (Zhang & Ling, 2018). Small datasets can amplify epistemic uncertainty, making it difficult for the model to learn the underlying data distribution accurately (Abdar et al., 2021; Xu et al., 2023). Small dataset sizes not only contribute to both bias and variance components of epistemic uncertainty because they cause some ambiguity in the optimal model parameters due to sparsity in some regions but also hinder the model convergence to a meaningful minimum generally (Heid et al., 2023).

While data insufficiency and imbalanced distributions are widely acknowledged and addressed in classification tasks (López et al., 2013; Wu et al., 2022; Talaei-Khoei & Motiwalla, 2023), they are less frequently discussed in the context of regression tasks. Classification problems are well-established due to their broad applicability in areas such as image recognition, spam detection, and disease diagnosis, which have led to significant advancements in this field. Various data augmentation methods, particularly in computer vision and natural language processing, have been developed to enhance dataset diversity (Mumuni & Mumuni, 2022). Synthetic data generation, noise addition, and bootstrapping are routine data augmentation methods (Batista et al., 2004; Li et al., 2018). One promising approach for handling data imbalances in classification problems is the synthetic minority over-sampling technique (SMOTE), which generates new samples from the minority class (Nitesh 2002; Blagus & Lusa, 2013). Branco et al. (2017) adapted SMOTE for regression tasks by incorporating Gaussian noise, resulting in a method known as SMOGN (synthetic minority over-sampling technique for regression with Gaussian noise). This technique can be effectively applied to geochemical data, which are often sparse and imbalanced. Common data augmentation methods are not well-suited for small datasets, as they primarily rely on duplicating existing data or adding noise rather than introducing genuinely new information. This can lead to an artificial inflation of model accuracy, as the ML models may learn to recognize duplicated patterns instead of capturing meaningful, generalizable relationships within the data.

In this study, we aimed to evaluate the performance of DNNs on a small dataset with imbalanced target distributions. Specifically, the concentrations of 20 trace elements serve as the target variables for the multi-output regression, while the concentrations of 11 major elements, along with the pH value, make the input feature set. To address the model-related uncertainties, rather than relying on a single DNN model, we trained and tested 1000 DNN models. We generated 1000 predictions for each target variable, providing a range of possible outcomes. Such an approach can project the existing uncertainties merely generated by the model due to sampling sparsity. Hence, the probability of the outcomes (concentration of elements) can be rendered rather than single deterministic results. The hyperparameters of the DNN models were optimized to create the most efficient and reliable configuration for a base model. To improve predictive accuracy and reduce uncertainty of the DNNs, we applied two preprocessing strategies: SMOGN resampling and statistical transformations (including Yeo-Johnson, Box–Cox, and square root). Our primary contribution is to establish multiple independent DNN models capable of predicting multiple trace element concentrations in geothermal fluids. The numerical insights gained from the improved performance of the DNN models can later serve as a guideline to optimize the sampling campaigns by focusing on the most significant input features. Finally, we quantified and visualized the impact of each input feature within the DNN models to help demystify inner workings, i.e., transforming the traditional ‘‘black box’’ nature of NNs into a more transparent ‘‘glass box’’.

DATA AND METHOD

This study dealt with a classical multi-dimensional regression problem for geochemical data. Therefore, data-driven DNNs were employed, as the nature of the problem dictates these types of models. Advanced physics-informed NN (PINN) techniques and their constantly developing sub-branches are promising tools that can conserve the laws of physics and leverage the speed-up of the NNs (Raissi et al., 2019; McClenny & Braga-Neto, 2023). PINNs solve forward regression problems, but there is no straightforward analytical/numerical relation between the input features and target variables in our case. The only evident constraint is that the concentration of an element is always positive, which can be enforced using appropriate activation functions (Sharma et al., 2017).

Dataset

The dataset used in this study comprised 109 measurements collected as part of the BrineMine project (Goldberg et al., 2023). The geothermal fluid was sampled at the Insheim geothermal power plant, located in the Upper Rhine Graben (URG), targeting a reservoir approximately 3600 m deep with a temperature of 165 °C. The sampled brine may originate from different parts of the reservoir, but its exact source cannot be accurately determined due to the nature of brine production from geothermal boreholes. Therefore, we assumed a single location for all samples, disregarding the impact of geospatial variability. The dataset contained monitoring data from the production well and experimental data from precipitation processes at the cold site of the power plant after heat extraction. To account for natural variations in the hydrochemical system, several samples from the production well were compared over one month, and no influence beyond the measurement error was found. For the artificially controlled experimental setup, the operational parameters (pH and residence time) were varied under continuously monitored flow-through conditions. This source diversity reinforces the already existing challenges with geochemical datasets. The technical details of the sampling are thoroughly described by Goldberg et al. (2023). Here, we only focus on the data quality for setting up the DNN model.

The constituents of a hydrochemical system are typically categorized into major and trace elements based on their concentrations. In geochemistry, major elements represent ions that form the dominant rock-forming minerals in a reservoir (Kaasalainen et al., 2015). Major elements—such as Na^+, K^+, Ca^{2+}, and Mg^{2+} as principal cations, and Cl^- and SO_4^{2-} as major anions—are routinely included in standard fluid analyses and are strongly linked to temperature-dependent equilibrium between the geothermal fluid and reservoir mineral assemblages (Ellis & Mahon, 1964; Giggenbach 1988). Fe^{2+}/Fe^{3+} and Al^{3+} are similarly important for the formation of ferromagnesian and alumino-silicate minerals. Silica concentrations reflect equilibrium with quartz or chalcedony (Fournier 1966), while pH, representing H^+ activity, influences ionic strength and mineral solubility (Davies 1938). These thermodynamic and geochemical controls indicate that the fluid composition carries a strong imprint of reservoir conditions and mineral–fluid interactions. The 20 target variables in this study represent trace elements, which often substitute for major elements in mineral lattices due to similarities in ionic radius and charge or occur as structural defects. According to the International Union of Pure and Applied Chemistry, trace elements are defined as those present at concentrations below 100 ppm (Bulska & Ruszczyńska, 2017). Their presence is indirectly governed by the same thermodynamic equilibria and mineral assemblages that determine major element concentrations. Because trace element analysis is more costly, technically demanding, and sparsely sampled, models that can infer their concentrations from more routinely available major element data offer a valuable tool for geochemical monitoring and exploration. Therefore, the regression strategy employed here, using 11 major elements and pH to predict 20 trace elements, was grounded in well-established geochemical principles and reflects domain-specific reasoning about elemental co-occurrence and solubility control in geothermal systems.

One of the primary challenges in our dataset was the significant variation in element concentrations. The large discrepancies in the absolute ranges of the parameters of the dataset are presented in Figure 1, while the most important characteristics are summarized in Table 1. In total, 10 orders of magnitude differences exist in the dataset. For some elements like Fe or Mg several orders of magnitude were measured in the sampling campaign, whereas the bars of elements like Na or Y look like a dot in Figure 1. The varying experimental setup, e.g., change in the pH, was an important driving force behind the wide ranges of distributions for Fe and Mg. Without a proper scaling method, parameters with more diverse distributions will completely dominate the behavior of the DNN model. The data were directly split into the input features and the target variables. The input features consisted of the concentrations of 11 major elements and pH of the fluid, while the concentrations of 20 trace elements served as the target variables. Regarding the concentration variation, Cl^- had a maximum concentration of 71,973.6 ppm due to the high salinity of the geothermal fluids in the URG (Pauwels et al., 1993; Stober & Bucher, 2015; Sanjuan et al., 2016; Drüppel et al., 2020), whereas the maximum concentration of elements like V or Ga was as low as 0.002 ppm.

Although normalizing the data partially addresses the large discrepancies in the absolute ran-

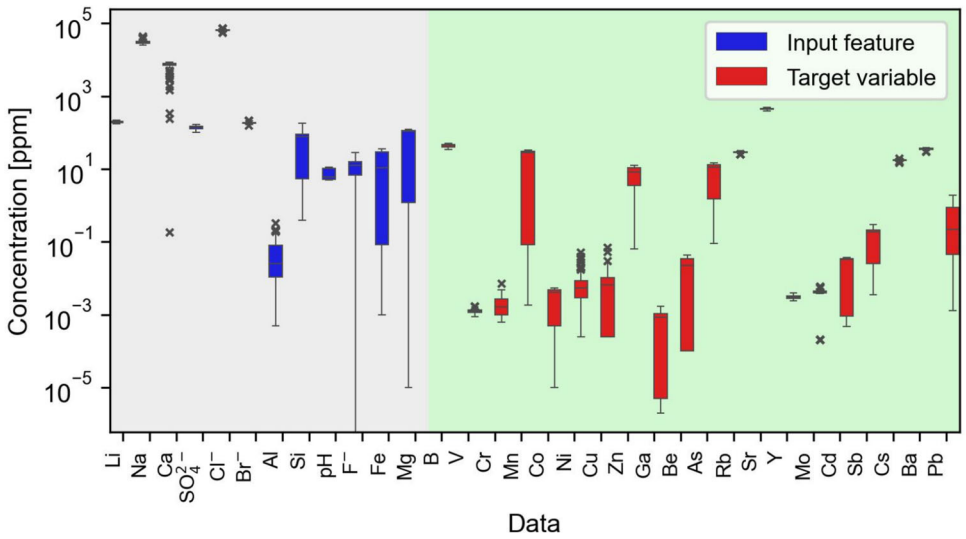

Figure 1. Measured concentration variation of 31 investigated ions and pH. Input features are in blue, target variables in red. The 20 target variables and 11 input features are also separated by different background colors, i.e., gray and green, respectively. The boxes represent the 25–75% interval. The *y*-axis is on a logarithmic scale to have a better view of the magnitude differences in the measured concentrations.

ges, the substantial variation still reflected the dataset's heterogeneity. Another concern was the limited size of the dataset. Neural network algorithms typically assume that a dataset is sufficiently large to represent an entire population, which is not the case here. Small datasets are likely to exhibit imbalanced distributions. The difference between the mean and median values in Table 1 suggests this imbalance. Figure 2 presents a violin plot of the normalized data. As shown in this figure, many elements have highly skewed distributions, and in some cases, such as Mo and Al, a pronounced bimodal distribution was observed due to their sensitivity to the experimental setup. This imbalance is likely a consequence of both the small dataset size and the inherent complexity of the data. Figure 3 illustrates the relationships between three input features (Ca, Na, and pH) and three target variables (V, Cu, and Mo). Including all 12 input features and 20 target variables would render the figure unreadable. The diagonal of the pair plot shows the distribution of each parameter, highlighting the imbalanced distributions, weak initial correlations, and large discrepancies in the absolute ranges of the parameters.

Low initial correlations between input features and some of the target variables can also degrade the precision of a NN model. Figure 4 is a heat map presenting the initial R^2 score measuring the correlation between each input feature and the target variable. Vanadium has extremely low correlation coefficients with input features. Copper can be nominated as the next one with a maximum R^2 score of 0.26 with Al. Besides the low initial correlation, the distribution of the Cu is also highly skewed with possibly some outliers. The high correlation between the pH and some of the target variables also pronounces the impact of the experimental condition.

NN Development

In this study, four ML algorithms among several existing ones were tested to map 12 input features into 20 target variables. Random forest regression (RFR) (Breiman, 2001), gradient boosting regression (GBR) (Friedman, 2001), support vector regression (SVR) (Cortes & Vapnik, 1995), and DNN were chosen as state-of-the-art AI methods closely related to geochemical data predictions (Rodriguez-Galiano et al., 2015; Okoroafor et al., 2022). An extensive preliminary study revealed the superiority of the DNN models for capturing the complex relation between the 12 input features and 20 target variables. Table 2 contains the summation of medians and standard deviations of the R^2 score as a measure of the accuracy of the predictions. As each algorithm was run 1000 times, the summations

Table 1. Summary statistics including mean, median, standard deviation (Std), minimum (Min), and maximum (Max) of the available dataset (numbers are rounded to avoid extra digits)

Variable	Category	Mean	Median	Std	Min	Max
Li	Input feature	196.4	197.0	10.1	172.1	216.6
Na		30592.8	29955.9	2972.6	25174.9	42292.7
Ca		6899.7	7434.5	1795.17	0.2	8458.4
SO_4^{2-}		138.3	142.3	16.0	100.2	166.9
Cl^-		64314.2	63962.1	2853.5	54316.6	71973.6
Br^-		184.0	181.8	11.4	151.2	212.3
Al		0.06	0.02	0.07	0.0005	0.32
Si		66.3	78.1	56.7	0.4	180.5
pH		7.3	6.0	2.3	5.0	11.3
F^-		10.5	12.53	6.0	0	28.56
Fe		15.1	10.86	14.5	0.001	36.1
Mg		75.3	107.7	49.5	0.00001	121.1
B	Target variable	43.4	44.2	3.7	33.8	49.4
V		0.001	0.001	0.0002	0.0009	0.002
Cr		0.002	0.002	0.001	0.0006	0.007
Mn		17.8	27.9	14.3	0.002	32.9
Co		0.003	0.004	0.002	0.00001	0.005
Ni		0.008	0.005	0.009	0.0002	0.05
Cu		0.008	0.006	0.01	0.0002	0.07
Zn		6.9	8.1	3.8	0.06	12.4
Ga		0.0006	0.0008	0.0005	0.0	0.002
Be		0.02	0.02	0.02	0.0001	0.04
As		8.5	11.0	5.3	0.09	14.5
Rb		28.6	28.8	1.5	24.9	31.8
Sr		442.8	442.7	21.6	386.2	496.5
Y		0.003	0.0030	0.0003	0.0024	0.004
Mo		0.0037	0.0042	0.0016	0.0002	0.0058
Cd		0.02	0.033	0.01	0.0005	0.04
Sb		0.13	0.18	0.09	0.003	0.29
Cs		17.53	17.57	0.68	14.84	19.3
Ba		35.0	35.1	1.8	28.8	38.6
Pb		0.5	0.2	0.4	0.001	1.9

The concentration of all elements is measured in ppm.

of the medians and standard deviations are presented in this table. The ML models predicted 20 target variables; therefore, the summation of the median score should be less than 20.0. The DNN models, with the highest summation of the median of R^2 score, showed a slightly higher accuracy, compared to other three methods. The standard deviation of the 1000 predictions also indicated the more stable behavior of the DNN models.

PyTorch library (Paszke et al., 2019; Ketkar & Moolayil, 2021) of Python was chosen due to its efficiency in DNN applications, especially for academia and research-focused projects. To address the uncertainty in predictions, 1000 independent DNN models were trained, with each model initialized randomly utilizing different random train–test splits. Additional models were tested to ensure robustness; however, the distribution of R^2 scores across target variables remained consistent, indicating that the 1000 models sufficiently encapsulated the variability in performance. The random initialization of the model and train–test split allowed us to capture the possible model-related, i.e., a subset of epistemic uncertainties (Hüllermeier & Waegeman, 2021). Our approach varied both the train-test splits and model initialization, which can affect model training and generalization. Given the small and imbalanced nature of our geochemical dataset, this study focused on how variability in data sampling influences model predictions, rather than on measurement noise (i.e., aleatoric uncertainty). Among various statistical measures of uncertainty, we chose the quartile coefficient of dispersion (QCD) because it offers a robust and interpretable way to quantify variability in the R^2 scores of DNNs. Unlike standard deviation or coefficient of variation, QCD is

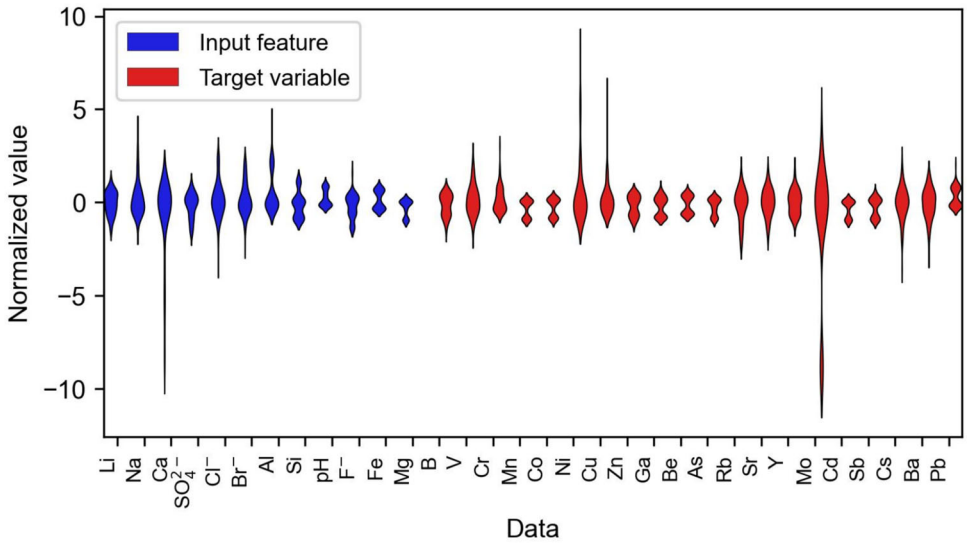

Figure 2. Violin plot showing the normalized, using the robust scaler, distribution of the input features (blue) and target variables (red). Despite being normalized using the robust scaler, most of the input features and target variables still have highly skewed distributions.

not affected by outliers or skewed distributions, which are common in small and imbalanced geochemical datasets. Additionally, being a relative measure, QCD allows meaningful comparison of prediction stability across different trace elements. QCD is defined as:

$$\mathrm{QCD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \tag{1}$$

where Q_1 and Q_3 are the first and third quartiles, respectively. QCD captures the spread of the middle 50% of the data relative to its central tendency (Rayat, 2018).

We first created a base case of 1000 DNN models with optimized hyperparameters (see Table 3 and Appendix). The optimal model architecture and hyperparameters were determined after a systematic analysis. Grid search was used as a hyperparameter tuning technique that explores a predefined range of input arguments by performing an exhaustive search over all possible combinations (Liashchynskyi and Liashchynskyi 2019). This method aims to identify the model configuration that yields the lowest error, i.e., the highest performance score. Among the tested architectures, six are presented in Figure 12. Based on accuracy and calculation time, a DNN with two hidden layers was chosen in this study. Given the dataset's wide range, spanning 10 orders of magnitude, it was necessary to standardize the range of the features (Sola & Sevilla, 1997). Among the tested scaling techniques (min–max, standard, and robust), the robust scaler was chosen for its ability to handle outliers effectively. For activation functions, the rectified linear unit (ReLU) was selected due to its simplicity and efficiency in preventing negative predictions for the target variables, such as concentration values (Ramachandran et al., 2017). Although we tested the hyperbolic tangent (tanh) activation function (Lau & Lim, 2018) and Leaky ReLU (J. Xu et al., 2020), they added computational complexity without improving the performance (Fig. 13). A learning rate of 0.1 was chosen, as it provided a better balance between accuracy and convergence compared to learning rates of 0.01 and 0.001. Adaptive moment estimation (Adam) (Kingma, 2014), stochastic gradient descent (SGD) (Bottou, 2012), and Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) (Liu & Nocedal, 1989) algorithms were compared to select an efficient optimizer (Fig. 14); Adam and SGD outperformed L-BFGS in terms of accuracy and computational cost. The accuracy of Adam and SGD remained almost the same, but Adam achieved this with 3 × lower computational cost. Regarding the loss criterion, mean absolute error (MAE), mean squared error (MSE), and log–cos*h* loss functions (Saleh & Saleh, 2022) were tested (Fig. 15). With MAE, small errors are as important as the big ones, making the gradient independent of error size. MSE heavily penalizes large errors by

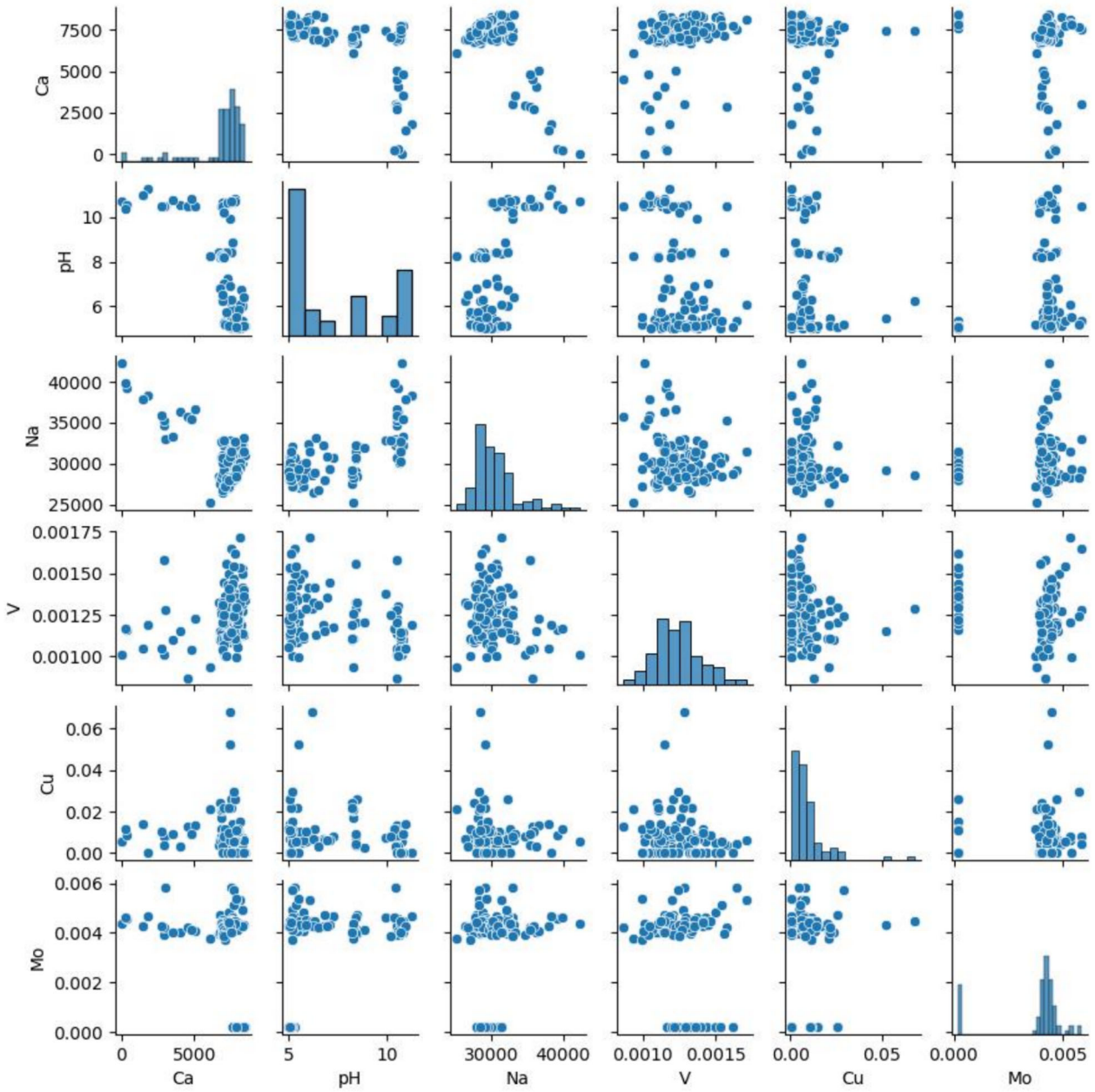

Figure 3. Cross-relation of three input features (Ca, pH, and Na) and three target variables (V, Cu, and Mo) in their original (measured) ranges. The unit of concentration for all elements is ppm.

squaring the value that can make the caveat of sensitivity to the outliers. Our small dataset was very sensitive regarding outliers, which is also proved by the better performance of MAE over MSE. The log–cos*h* loss function is a mixture of MAE and MSE:

$$L_{\log-\cosh} = \frac{1}{N}\sum_{i=1}^{N} \log(\cosh(f(x_i) - y_i) \qquad (2)$$

where N is sample size, $f(x_i)$ denotes the predicted value and y_i is the true value. Then, for small errors, $L_{\log-\cosh}$ is approximately $\frac{x^2}{2}$, i.e., quadratic. For large errors, it behaves linearly: $L_{\log-\cos}h = |x| - \log(2)$. The high computation cost is the main drawback of the log–cos*h* loss function. In this study, the results of the 1000 DNN models did not improve by

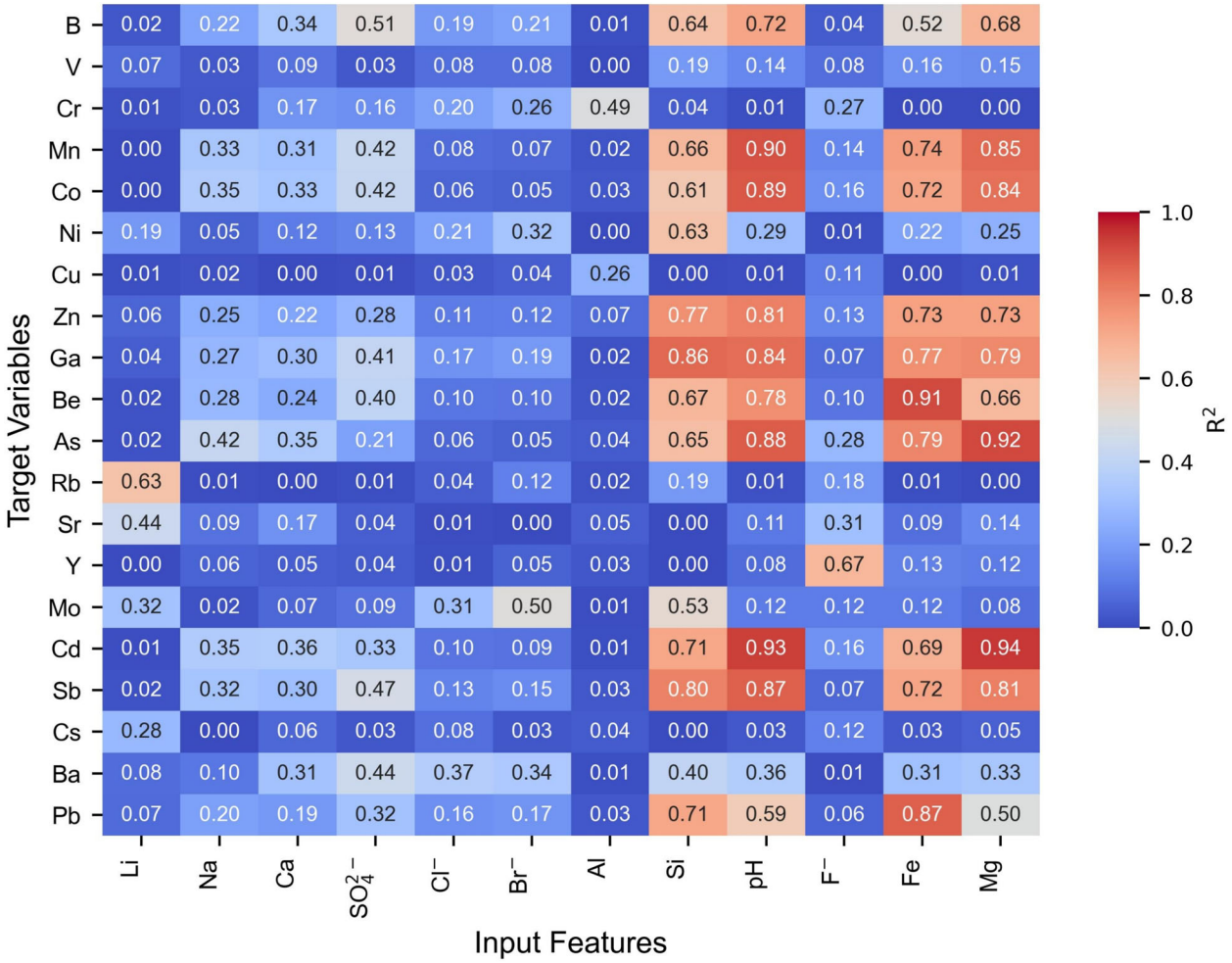

Figure 4. Heat map representing the correlation coefficients (R^2) between input features and target variables. The value of R^2 between each pair is written in the related cell.

Table 2. Summation of the medians and standard deviations of the four tested ML algorithms

ML algorithm	$\sum R^2$ median	$\sum R^2$ Std
DNN	15.78	1.50
RFR	15.49	1.54
GBR	15.33	1.67
SVR	15.55	1.52

switching from MAE to the log–cos*h* loss function; hence, the former option was chosen due to its lower computational cost.

Additional advanced techniques were employed to further customize the DNN models. To reduce the computational cost of training of the 1000 DNN models, early stopping was implemented. In addition to minimizing runtime, early stopping helps prevent overfitting (Yao et al., 2007). This technique monitors changes in the validation loss. The optimal fraction of the training–validation split can be determined based on the number of data points and model parameters (Amari et al., 1995; Hsieh 2009), thus:

$$f_{\text{opt}} \approx \frac{1}{\sqrt{2\mathrm{N}_p}}, \quad \text{for large } N_p \tag{3}$$

where f_{opt} is the optimum fraction and N_p represents the number of model parameters.

As a general rule, overfitting is less of a concern when ample data are available, and using a large

Table 3. Architecture and hyperparameters of the finalized base DNN model

Layers	Neurons	Scaler	Activation	Learning rate	Optimizer	Test size	Loss	Patience
Input	12	Robust scaler	ReLU	0.1	Adam	0.2	MAE	100
Hidden 1	32							
Hidden 2	16							
Output	20							

validation set may unnecessarily reduce training efficiency. In our case, however, the dataset was limited, making overfitting a potential issue. To mitigate this, we employed early stopping and allocated 20% of the training data for validation. The method proposed by Amari et al. (1995) was not applicable here, as it relies on the assumption of a large dataset. The base DNN models were trained for 2000 epochs while early stopping broke the training loop if the validation loss did not improve for 100 consecutive epochs, i.e., patience equals 100. L2 regularization was added to the Adam optimizer with the weight decay parameter (G. Zhang et al., 2018). Then, Adam can apply a penalty proportional to the square of the magnitude of the weights. This helps prevent overfitting by discouraging the weights from becoming too large, i.e., overfitting. Despite normalizing the dataset using the robust scaler, the batch normalization technique was also used to normalize the inputs to each layer (Santurkar et al., 2018). In our DNN model, batch normalization was used after the linear transformation and before the ReLU activation function.

To effectively address the imbalanced distribution of the dataset, SMOGN was applied as an extra preprocessing step. The Python smogn library, developed by Branco et al. (2017) (available at https://github.com/nickkunz/smogn), was used to implement the SMOGN algorithm. The method is thoroughly explained by its developers, and only a brief overview is provided here. SMOGN resamples the original dataset to improve representation in the undersampled regions of the population. It redistributes data by randomly undersampling the majority class and over-sampling the minority class to enhance diversity. When generating new data for over-sampling, two strategies are employed: interpolation between two minority samples or the addition of Gaussian noise to existing minority samples. SMOGN first distinguishes minority samples from the majority. Then, for each minority sample, it calculates the distance between that sample and the others. Based on this distance, SMOGN selects the appropriate over-sampling technique. If the sample is close to others within the minority group, new samples are generated through interpolation. Otherwise, the sample is far from the others and a Gaussian noise is added instead of interpolation with unrelated samples.

To further improve the data distribution, an additional step of redistribution was applied using various transformation techniques. Specifically, the square root, Box–Cox (Sakia, 1992), and Yeo-Johnson (Weisberg, 2001) transformations were employed on selected input features and target variables. The square root method simply takes the square root of the values, which can result in less variance and skewness. The Box–Cox technique is particularly useful for normalizing distributions and stabilizing variance. Machine learning algorithms often assume that the variance of residuals is constant regardless of the values of the independent variables, an assumption known as homoscedasticity (Wang & Zhou, 2007). However, the fan-shaped relationship between Na and V (Fig. 3) suggests heteroscedasticity in our data. This can be addressed by Box–Cox, although it only functions for positive values, based on the following formulation:

$$y(\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & if\ \lambda \neq 0 \\ \log(x), & if\ \lambda = 0 \end{cases} \tag{4}$$

where x is original data, $y(\lambda)$ is transformed data, and the parameter λ determines the nature of the transformation and is automatically estimated by the SciPy library (Virtanen et al., 2020) in a way to transform the data into the most normal distribution. In the default mode, SciPy uses maximum likelihood estimation (MLE) method to estimate λ. Yeo-Johnson transformation is very similar to the Box–Cox with slight differences. Yeo-Johnson can be used to transform negative values and is very efficient in dealing with skewed and heteroscedastic datasets. The SciPy documentation provides a comprehensive overview of the mathematical back-

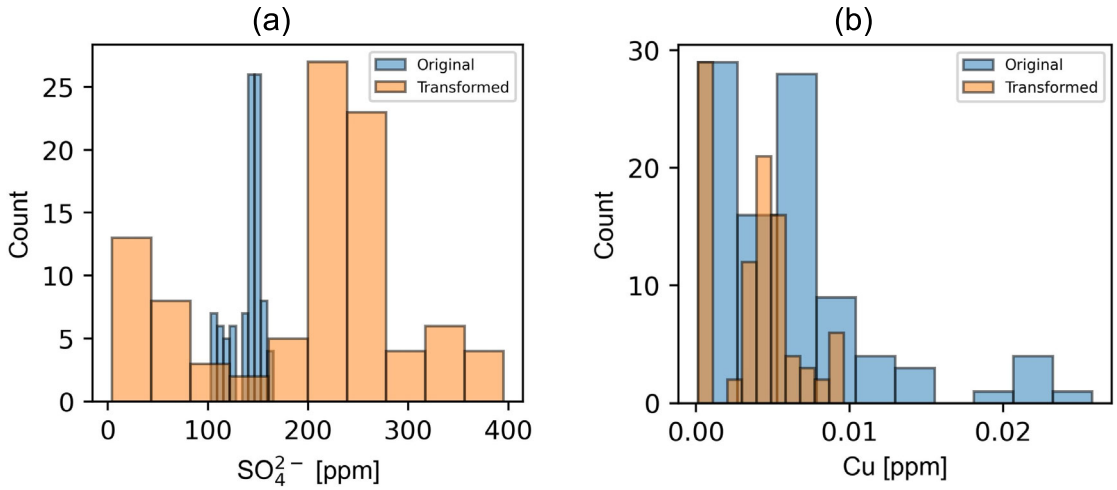

Figure 5. Effect of Yeo-Johnson transform on the distribution of (**a**) SO_4^{2-} and (**b**) Cu. The transformation resulted in more uniform distributions and a better data balance for both cases.

ground of the Yeo-Johnson transformation. Figure 5 shows two examples of the Yeo-Johnson transformation. In both cases, the distribution became more uniform after the transformation. This effect on the skewness was especially noticeable with Cu, where a significant shift was visible before (heavily right skewed) and after the transformation.

In this study, all preprocessing steps were conducted after the data were split into training, validation, and test subsets, thereby avoiding any information leakage. Specifically, the SMOGN resampling technique was applied only to the training set, without any access to or influence from the validation or test data. The test set remained untouched and was used exclusively for final evaluation to ensure unbiased performance assessment. Similarly, all data transformations, including feature scaling (e.g., RobustScaler), were fit on the training data only and then applied to the validation and test sets.

.

One of the key limitations of data-driven NN models is their lack of interpretability (Baker et al., 2019; Willcox et al., 2021; Degen et al., 2023). In our DNN model, direct human interpretation of the relationships between the layers was simply impossible. For instance, there were 384 weight and 32 bias values between the input layer and the first hidden layer. Overall, our DNN model contained 1380 learnable parameters. To address this interpretability issue, the accumulated local effects (ALE) method was used as a model-agnostic approach for global explanations of the DNN models (Apley & Zhu, 2020). ALE scores address a proper functional decomposition of the model and, as such, they are not sensitive to the possible interactions among variables (Molnar et al., 2020). ALE calculates how much the prediction changes, on average, when a slight change happens in an input feature. The ALE method works by dividing the range of each input feature into small, non-overlapping intervals. For a given feature, ALE calculates the local effect by measuring how the model's prediction changes when the feature value moves from one interval to the next, while keeping other features fixed. These local differences are then averaged over all data points within each interval, and the effects are accumulated across the feature range to construct a function that represents the marginal effect of the feature on the model's output, which is demonstrated mathematically as:

$$\mathrm{ALE}_j(z) = z_{\min} z \int E_{X-j|x_j=t}\left[\frac{\partial f(t, X-j)}{\partial t}\right] dt \quad (5)$$

where $f(\cdot)$ is the prediction function, and $X - j$ represents all features except x_j. This expression captures the accumulated average effect of feature x_j on the model output as its value increases from the minimum observed value z_{min} to a specified point z. The variable t is used as a dummy integration vari-

able to avoid notational ambiguity, since x_j is also used to denote a fixed value at which the ALE is evaluated. The term $\frac{\partial f(t,X-j)}{\partial t}$ represents the local sensitivity of the prediction with respect to small changes in x_j, while the expectation $E_{X-j|x_j=t}$ ensures that this sensitivity is averaged over the conditional distribution of the other features, thereby accounting for their dependence on x_j.

One key advantage of ALE is that it provides a reliable decomposition of feature effects, making it particularly effective for interpreting complex DNNs in the presence of correlated or interacting features, similar to our case with 12 input features and 20 target variables. The PyALE library developed in Python (https://github.com/DanaJomar/PyALE) was used here to quantify the impact of the input features. All the aforementioned techniques and methods were developed in different Jupyter notebooks and can be accessed via the Zenodo and GitHub (see Code and Data Availability section) of the first author.

RESULTS

Figure 6 presents the distribution of the R^2 scores (measuring the correlation between real and predicted target variables) for 1000 DNN models. As the figure shows, the accuracy of the prediction increased in a stepwise manner with different color codes. The highest performance was achieved when SMOGN was applied first, followed by data transformation. This combination led to the most notable improvement in the DNN models' predictive accuracy compared to the base case, as revealed through a systematic evaluation of different preprocessing strategies. While the order of SMOGN and data transformation can be reversed, doing so generally results in lower model accuracy, indicating that the sequence of preprocessing steps plays a crucial role. Figure 16 compares four distinct preprocessing strategies, highlighting their relative impact on model performance. The base DNN model shows an acceptable level of accuracy for the majority of the trace elements. The predictions of the base model were the worst for V, Cu, and Cs.

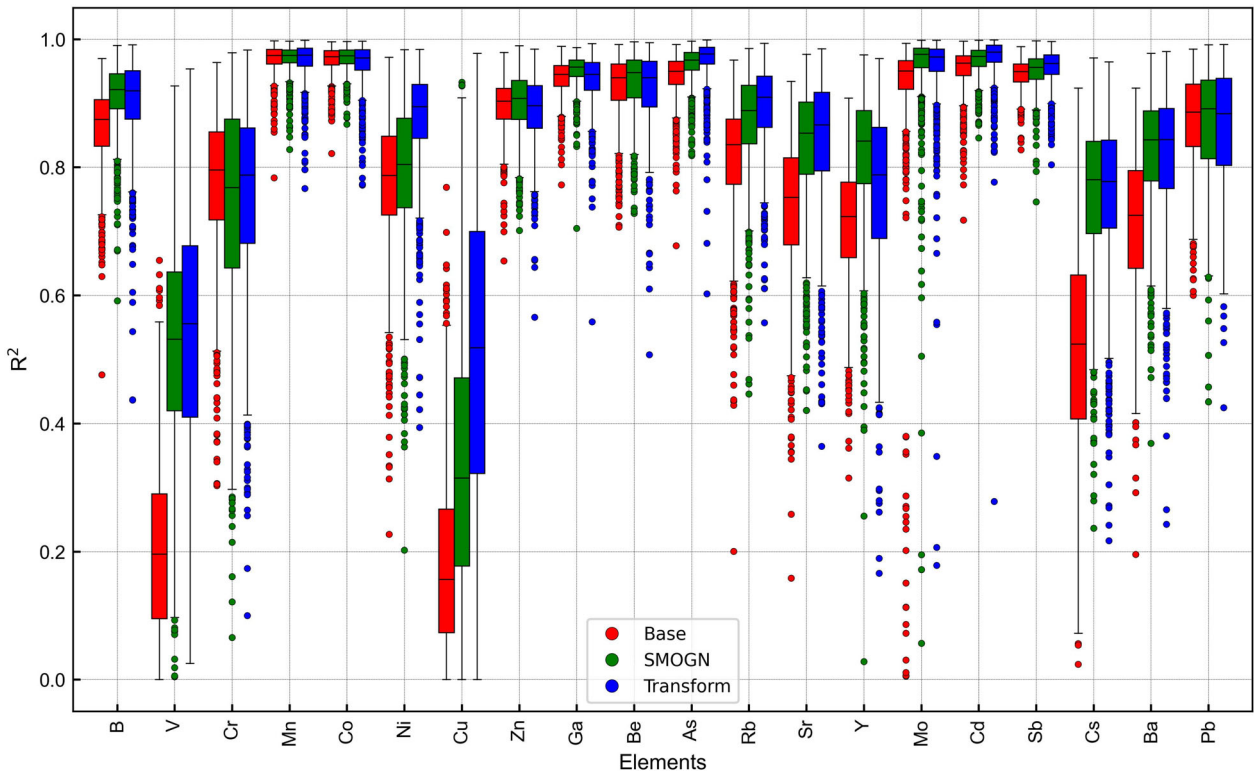

Figure 6. Distribution of R^2 scores for 1000 different neural networks. For each target variable, three color-coded boxes and whiskers show the changes in accuracy from the base DNN model (blue) to the models modified with the SMOGN resampling technique (green) and data transformation methods (red).

For many elements, a broad range of accuracies was observed, primarily due to the initial distribution and limited size of the dataset. One major factor contributing to this variability was the train–test split, which can lead to inconsistent accuracy across models. Highly accurate predictions for V, Cu, and Cs often result from ''lucky'' splits, where the test set contains data points similar to those in the training set. Additionally, random weight initialization and differing optimization paths contribute to the variability in results. Although *k*-fold cross-validation is typically used to ensure that each data point is tested representatively (Fushiki, 2011) our small dataset limited its ability to reduce the variability in model precision.

In our study, early stopping not only prevented overfitting but also significantly reduced computation time. In terms of computational cost, early stopping lowered the runtime by one order of magnitude, reducing it from 75 to 8 min for the 1000 DNN models on a Lenovo laptop equipped with Intel Core i7-10510U CPU (1.8–2.3 GHz, 4 cores), 16 GB of RAM, running a 64-bit operating system. No GPU acceleration or specialized hardware accelerators were used. The learning curves of the 1000 DNN models of the base case are plotted in Figure 7. The training loss (blue curves) was calculated during the training phase, while testing loss measured the model accuracy for predicting the data not seen during training. As Figure 7 indicates, the model was well-trained and generalized effectively to unseen data. The variability observed in the loss trends across the 1000 models can be attributed to the random initialization of the models and the variation in train-test splits. Although the learning curves could be smoothed by lowering the learning rate, this would not enhance the model's final precision. Moreover, a lower learning rate would significantly increase computation time; for instance, using a learning rate of 0.001 instead of 0.01 resulted in four times longer training durations due to the additional iterations required to optimize the model's weights.

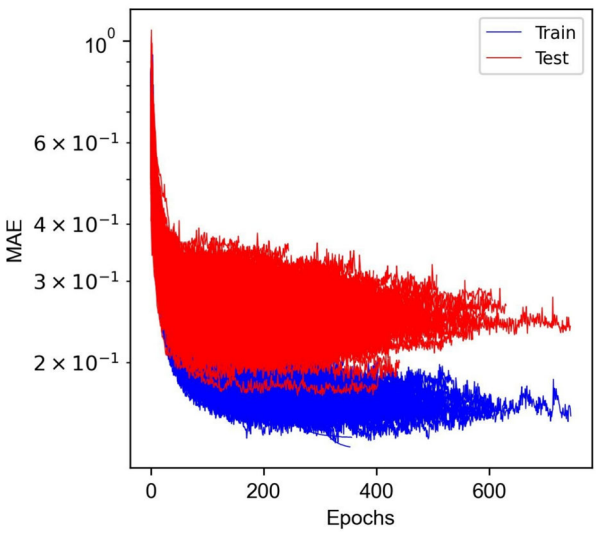

Figure 7. Learning curves of training (blue lines) and testing (red lines) of the 1000 DNN models of the base case. An early stopping with patience of 100 epochs decided when to break the training loop.

There was a significant increase in accuracy when transitioning from the base case to the preprocessed data with SMOGN, particularly for V and Cu, where precision nearly doubled. Accuracy improved across all 18 remaining trace elements as well (Fig. 6). In addition to applying SMOGN, modifying the distribution of certain input features and target variables also increased the DNN model's accuracy. We compared the model results by applying different data transformation techniques and choosing the most optimal distribution for each variable to deliver the highest prediction accuracy. For example, adjusting the distribution of Cu using the Yeo-Johnson method significantly enhanced the prediction accuracy. Likewise, transforming the distribution of pH as an input feature slightly improved the prediction accuracy for most of the target variables. The applied transformation methods can be summarized as follows:

- Yeo-Johnson: Fe, SO_4^{2-}, Pb, B, Y, Cs, V, Cu• Box-Cox: pH, Ni• Square root: Rb

Figure 8 shows the impact of the Yeo-Johnson transformation on the Fe (as an input feature) for the prediction of Be. While the true relationship between the 12-dimensional input features and the target variables was far more complex, this figure provides a simplified visualization of the cross-correlation between two elements. Although the residuals between the two cases (with and without transformation) remained the same, their distributions differed significantly. Without transformation, the residuals exhibited a fan-shaped pattern, indicating heteroscedasticity. Applying a simple Yeo-Johnson transformation to Fe altered its relationship with Be, effectively addressing this issue.

.

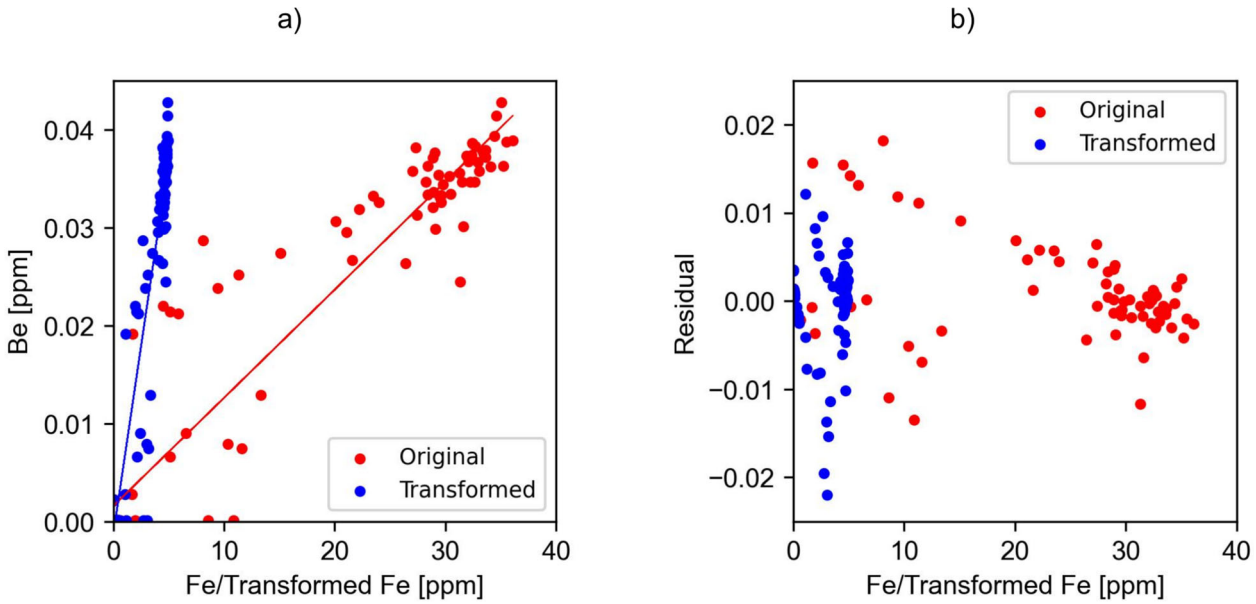

Figure 8. Comparison of relationship of Fe and transformed Fe with Be. (**a**) Red dots (labeled as Original) represent scatter plot between Fe and Be. The continuous red line is the regression line between Fe and Be. The relation of the transformed Fe with Be is shown by the blue dots and line. (**b**) Residuals of the regression line and real values from subplot are visualized here. The regression between the original Fe and Be makes a fan-shaped distribution for the red dots, while for the transformed Fe (blue dots), the distribution of the residual is more homogeneous.

Figure 9 visualizes the extracted standard deviation and median of 1000 calculated R^2 scores of the base, SMOGN, and transformed cases. The SMOGN and manual transformation successfully increased the accuracy, i.e., the median R^2 scores for all elements. For some elements like Ni, Sr, Mo, and Cs, the SMOGN and manual transformation not only increased the accuracy but also decreased the uncertainty, i.e., standard deviation, of the DNN models. For V and Cu the increase in the median was significant, with slightly increased standard deviations.

The performance of the DNN model improved significantly with the application of SMOGN and manual transformations. However, these enhancements came at the cost of increased model complexity. This was evident in the higher computational cost and the increased number of required epochs for training each model. The base model required approximately 230 epochs (Fig. 7) to map the input features to the target variables, while the SMOGN and transformed cases required about 350 epochs. In terms of computational time, the SMOGN and transformed cases took about 12 min to train the 1000 DNN models, compared to 8 minutes for the base case. While the base model struggled to find meaningful relationships between the input features and some target variables, the more complex SMOGN and transformed models took longer but ultimately established more robust relationships.

DISCUSSION

In this study, we established a logical framework for understanding the behavior of the model, drawing on factors such as the initial correlation between input features and target variables, the distribution of the dataset, and the absolute concentration values. Based on the performance of the base case with the lowest accuracy, the predictions of the DNN models can be categorized into three groups according to the median R^2 scores:

- Poor ($R^2 < 0.6$): V, Cu, Cs
- Moderate ($0.6 < R^2 < 0.8$): Ni, Sr, Y, Ba, Cr
- Good ($R^2 > 0.8$): B, Mn, Co, Zn, Ga, Be, As, Rb, Mo, Cd, Sb, Pb

Starting with the worst-predicted element, V, the primary issue behind the low accuracy of the predictions was the extremely low initial correlation with input features. The highest correlation coeffi-

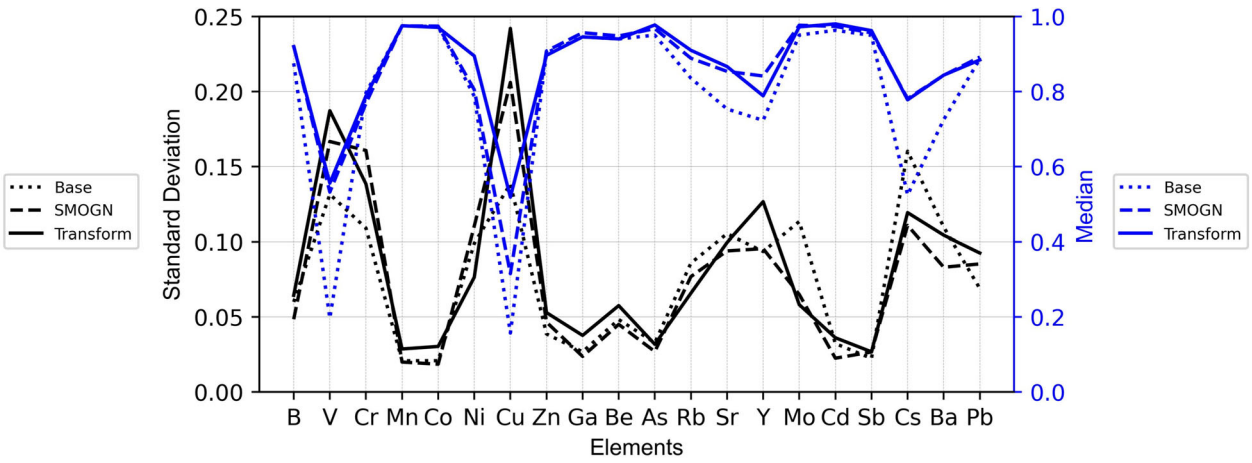

Figure 9. Standard deviation (black) and median (blue) of 1000 trained DNN models. The base, SMOGN, and transform cases are represented by dotted, dashed, and solid lines, respectively.

cient for V was only 0.19, with Si. The concentration of V typically ranges 5–80 ppm in granite (Nriagu and Pacyna 1988). In the URG, a range of 0.005–0.007 ppm concentration was measured for geothermal brines in the Soultz-sous-Forêts (Nitschke et al., 2014). However, in our dataset, V concentrations lay at the lower end, with values ranging 0.0009–0.002 ppm. For Cu, the maximum initial correlation was just 0.26, with Al. In addition to the low initial correlation, the distribution of Cu was highly skewed, a problem also seen with other elements such as Si and Al. In our dataset, the Cu concentration ranged 0.0002–0.07 ppm, whereas higher values (up to 0.37 ppm) were measured in the URG (Nitschke et al., 2014; Sanjuan et al., 2016). Furthermore, the concentration ranges of Al, which were the most correlated input feature to Cu, were also very low. Calculating the limit of detection (LOD), based on the proposed scheme of Schwarzbauer and Jovančićević (2020), revealed that eight elements, including Al and Cu, in our dataset fell below this threshold. Table 4 lists the elements and their respective counts of < LOD values. The redox sensitivity of Cu can also impact its concentration during sampling by forming copper sulfate/sulfide compounds. The formation of Cu-FeS solid solutions has been documented in other geothermal fields (Hardardottir et al., 2010), including URG (Nitschke et al., 2014; Goldberg 2024). The bimodal distribution of Fe highlights the complex geochemical interactions occurring in the geothermal fluids during the sampling. Cesium, another poorly predicted element, had a low initial correlation with the input features. While Cs showed initial correlation ranges similar to Cu, its most correlated feature, Li, was measured with greater accuracy and without any < LOD value. Consequently, the DNN models achieved better predictions for Cs than for Cu. This underscores that prediction accuracy is influenced by both the initial correlation coefficient and the distribution of the most correlated input feature(s).

Table 4. Number of < LODs for eight elements of the dataset.

Element	Al	Fe	Cu	Ni	Ga	Mo	Be	Co
No. < LODs	17	8	28	15	39	17	43	1

A comparison of the elements in the moderately predicted group highlights the significance of data distribution. Regarding the initial correlation, Ni and Y had the highest R^2 scores of 0.63 and 0.67 with Si and F^-, respectively. Meanwhile, Sr had the maximum R^2 score of 0.44 with Li. Despite this, the prediction accuracy for Ni, Y, and Sr was nearly the same. The reason is related to the multimodal distribution of Si and F^-, whereas Li exhibited a more balanced distribution. Similarly, for Ba the highest correlations were with Si and SO_4^{2-}, both of which had distributions that degrade the predictions.

To further investigate the uncertainty in DNN performance across different preprocessing strategies, the QCD was computed for the R^2 scores of

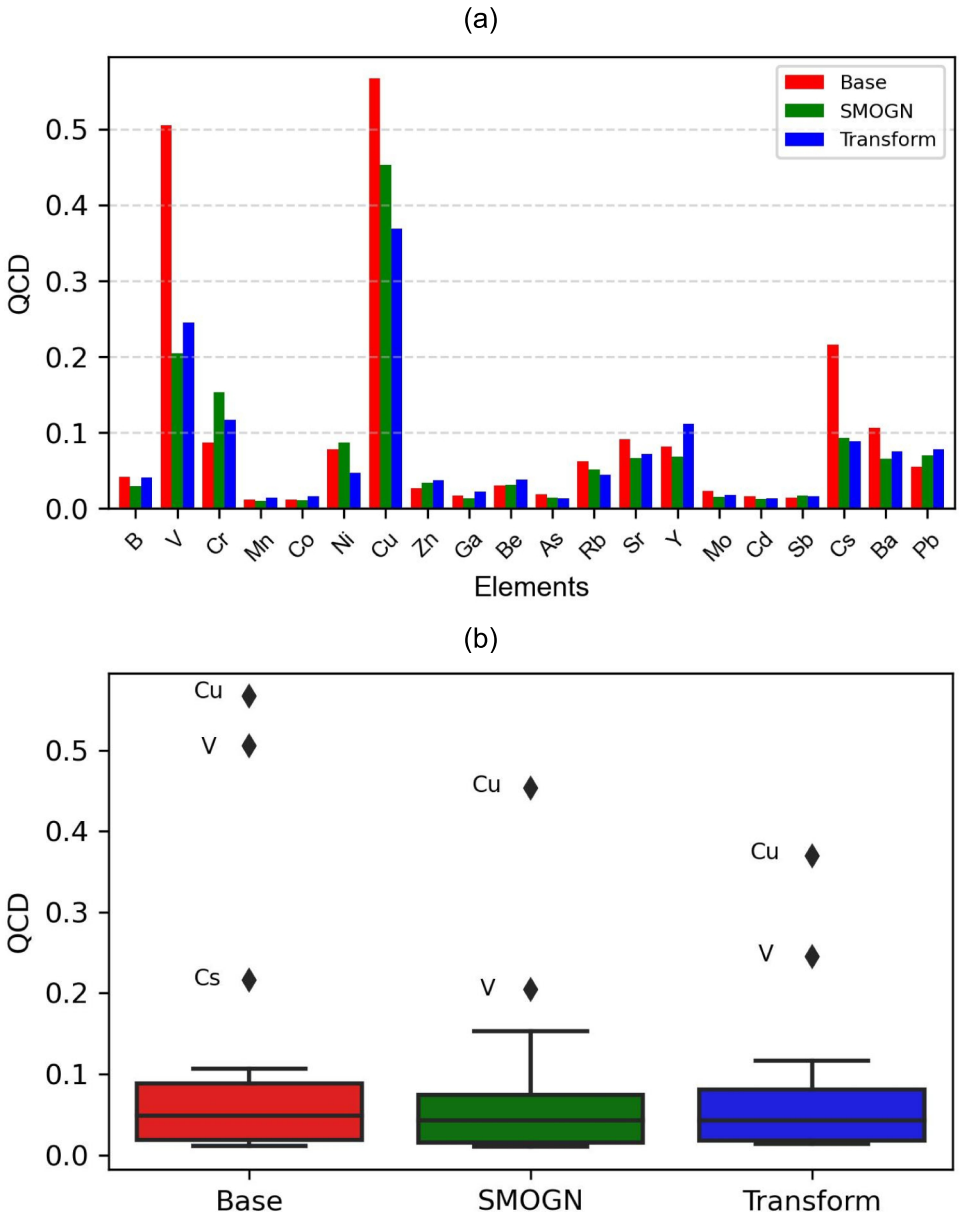

Figure 10. (**a**) QCD of 1000 R^2 scores for each target element across three preprocessing strategies: base case without any preprocessing (red), SMOGN (green), and transformed (blue). (**b**) Box plots of QCD values for all 20 target variables under three modeling approaches: Base (red), SMOGN (green), and Transform (blue). Diamonds represent statistical outliers, with labels indicating the corresponding elements.

each target variable (Fig. 10a). This figure provides a quantitative measure of the uncertainty by capturing the relative variability in model performance across 1000 predictions. A lower QCD value indicates greater consistency and reduced uncertainty in the model's R^2 outcomes, highlighting the stabilizing effect of SMOGN and data transformation methods. The results revealed that certain elements, such as V, Cu, and Cs, exhibited notably higher QCD values in the base model configuration, indicating greater instability in the neural network's predictive performance. Copper and V had relatively high uncertainty and were among the most problematic target variables, due to their very low concentration levels in the input data. Therefore, integrating external datasets with higher Cu and V concentrations could

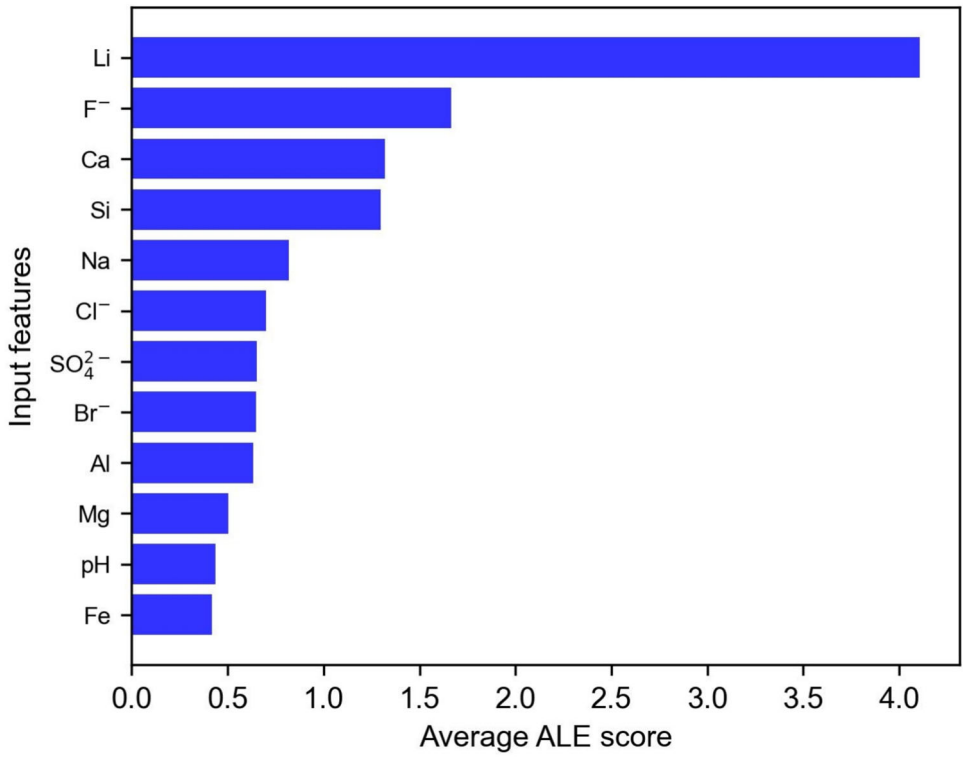

Figure 11. Average ALE scores of 12 input features.

improve prediction accuracy and reduce model uncertainty. The application of SMOGN and data transformations consistently reduced QCD values for most elements, suggesting that these preprocessing techniques not only improved average model accuracy but also contributed to a more stable and reliable prediction outcome.

Figure 10b presents box plots of the QCD values for the three modeling approaches across all target elements. Each box shows the median, interquartile range, and overall spread of QCD values for the respective method, while diamond markers indicate statistical outliers. These outliers, labeled with the corresponding element names, highlight cases where the QCD markedly deviates from the central distribution. For V, the QCD value increased when switching from the SMOGN to the Transform approach, indicating a wider spread in the model's performance across runs. However, this was accompanied by an upward shift in the distribution of R^2 values, suggesting that while the predictions exhibited greater variability, the overall accuracy improved. This reflects a trade-off between consistency and performance, where higher accuracy comes at the cost of increased variability. This analysis underscores the importance of addressing data imbalance and scale-related issues when modeling trace elements, particularly for those at the distributional extremes.

All the aforementioned interpretations were primarily based on single cross-correlations between the target variables and their most correlated input features. To more accurately capture the full dependency matrix of the DNN model with its 1380 tuned parameters, the average ALE score of the base model is visualized in Figure 11 as a summary plot for 1000 DNN models. The figure highlights the wide variation in the importance of the input features. As shown in Figure 11, Li had the highest impact on the performance of the models. The ALE scores showed discrepancies concerning the initial correlation coefficients, shown as a heat map in Figure 4. The reason can be linked to the shared information, the nonlinear influence of the input features, and possible interactions between the input features. For example, Li was the only input feature that had a high correlation coefficient with target variables such as Rb and Sr. Compared to Li, input

features like Mg, pH, and Fe had far higher correlation coefficients with several target variables. However, the shared information between these three input features decreased their importance. To evaluate the robustness of the ALE score, two new sets of DNN models, with 1000 runs, were tested: in the first set Fe was removed from the input features list, and in the second Li was removed. Figure 17 shows the accuracy of the three cases. As the figure shows, the performance of the DNN models remained almost the same after removing the Fe from the input features. The impact of the Li on the accuracy of at least five target variables, e.g., B, Rb, Sr, Cs, and Ba, is obvious.

There was a sharp improvement in the model's accuracy after statistical resampling of the dataset using SMOGN. Our study confirmed that more homogenous and uniform data distributions can allow the DNNs to see the diversity and establish more robust relations, as also observed by Dashti et al. (2024). Sampling a full range of concentrations for all 31 major and trace elements at a single location was impossible due to the site-specific geochemical properties of geothermal fluids. Factors such as reservoir lithology, temperature, pressure, and regional flow impact water–rock interactions, which in turn influence the concentration of elements in the produced fluid (Gallup, 1998; Stober & Bucher, 2012). In our case, the fluid was highly enriched with Cl^- and Na, while severely depleted in V or Ga. Low concentration values posed significant challenges for the developed DNN in this study. Every measuring instrument has an intrinsic error range, which becomes more pronounced at lower concentration values. Impurities also contribute to errors in geochemical sampling campaigns. Therefore, the precision of the developed DNNs was tightly linked to the accuracy of the measured concentrations.

Applications and Limitations

Although geochemical data are typically sparse, the DNN models presented here were successfully trained and tested using a dataset with only 109 measurements. In this study, we aimed to reveal relationships between elements that may not be immediately apparent and to demonstrate the feasibility of making reliable predictions of trace elements in incomplete datasets, independent of the specific elements included in the model. This study was the development of 1000 DNN models rather than relying on a single model. By evaluating the accuracy, R^2, for each of these models, the study provides a comprehensive view of the performance variability across different DNN models. This approach allows for a quantitative representation of the uncertainty associated with the effects of sparse training data on a specific neural network architecture, offering deeper insights into how data scarcity affects model reliability. Such a methodology is particularly valuable in applications like geothermal brine analysis, where uncertainty quantification is critical for informed decision-making in resource evaluation and process optimization.

Many legacy geochemical datasets contain measurements for only major elements, often due to the absence of advanced measuring tools, with the samples no longer available. In cases where renewed interest arises in those locations, the DNN models could be used to predict the concentration of the desired trace elements.

The developed DNN models can significantly enhance raw mineral extraction from geothermal brines by enabling the prediction of valuable trace element concentrations using readily available major element data. By providing insights into possible ranges of trace elements, DNN models support optimized extraction processes tailored to specific brine compositions, maximizing yield and minimizing waste.

The development of DNN models could optimize sampling campaigns, as the DNN can predict the concentrations of 20 trace elements. Such models can be deployed to avoid repetitive samplings during the monitoring phase of the geothermal power plants. Therefore, a full set of analytically measured elements during the exploration phase can be used to avoid tedious samplings in the later (monitoring) phase, reducing both time and costs.

The improved performance of the DNN on preprocessed and transformed data further aids in designing informed sampling campaigns. While the size of the dataset remained constant, applied transformation techniques enhanced accuracy by creating more uniform data distributions. The Machine Learning for Enhancing Geothermal Energy Production (MALEG) (Nitschke, 2024) project will use insights from this study to optimize sampling campaigns at various geothermal power plants. The MALEG study and other geochemical sampling campaigns can also further benefit from the results achieved after quantifying the impact of each input

feature. The calculated ALE scores revealed that input features like Li have the biggest impact on the accuracy of the DNN models. Therefore, measuring more impactful elements should be prioritized to enable an intelligent sampling campaign. Such adjustments can significantly reduce the complexity of geochemical data acquisitions in terms of time and cost.

In terms of limitations, the findings of this study highlight the sensitivity of data-driven DNN models to the distribution of the input data. Explicit enforcement of the physics or chemistry laws can make the DNN models more robust. Our dataset lacks such constraints and relies on the mathematical robustness of the data-driven NN approaches. In our case, the highly biased distributions of the geochemical data made the training process more challenging and reduced the reliability of the DNN models. Extrapolation is another challenging aspect of such data-driven models. Our small dataset comes from one single location covering a limited part of the existing (wide range) geochemical concentrations. Therefore, it is impossible to apply 12 input features in a way to evaluate the reliability of the DNN models in terms of extrapolation.

CONCLUSIONS AND OUTLOOK

The lack of large, high-quality datasets has long been a central challenge for applying ML in fields (such as geosciences and particularly in geochemical studies) where data acquisition is costly, slow, and highly purpose-driven. Unlike domains where ML thrives on abundant data, geochemical datasets tend to be small and expensive to expand. This mismatch creates a systemic barrier to fully adopting data-hungry ML techniques. In this study, we addressed this challenge by demonstrating how deep learning methods can be deployed to extract meaningful patterns from limited and imbalanced geochemical datasets. Specifically, we explored the feasibility of predicting trace element concentrations based solely on major elements, re-purposing small geochemical datasets to support predictive modeling in a field where ''big data'' is often rare.

The small dataset, unfavorable data distributions, and extremely low concentration values posed challenges to DNN performance. To address the uncertainties, 1000 DNN models were used instead of a single model. Techniques such as early stopping and batch normalization improved the accuracy, with 12 out of 20 elements achieving median R^2 scores above 0.8. The worst predictions were for V, Cu, and Cs, which had R^2 scores below 0.6 due to the low initial correlations and imbalanced distributions.

The SMOGN resampling technique significantly enhanced the model's accuracy, particularly for challenging elements like V, where the median R^2 score improved from 0.19 to 0.55. Additional transformations, such as square root, Box–Cox, and Yeo-Johnson, further increased accuracy and reduced variability for trace elements such as Ni, Sr, Mo, and Cs. Therefore, small geochemical datasets with a limited number of measured elements can still be used to predict the concentration of trace elements. The inherent instability and uncertainty of data-driven models remain a limitation that can be addressed by multiple independent DNN models. Computed QCD for the R^2 scores of each target variable confirmed that applying SMOGN and data transformations reduced prediction uncertainty for most elements, highlighting the stabilizing effect of these preprocessing strategies.

The ALE score of the 12 input features revealed that Li was the most prominent input feature due to its unique correlation with target variables. Other input features had far higher correlation coefficients with several target variables; however, their impact was less than Li and they shared the same information. Expensive geochemical measurements and campaigns can be more informed by the development of such studies that quantify the impact of each input feature.

Future studies, as developed in the frame of the MALEG project, can consider integrating data from locations with different brines while ensuring standardized sampling protocols to avoid introducing artifacts and enhance the robustness of the DNN model. Combining small datasets from different geothermal brines could also help achieve more homogeneous distributions. DNN models can also be developed to capture the relationship between operational perturbations, such as modifying pH, temperature, and other factors, and the response of the geochemical system, such as changes in element concentrations. Relating the measured element concentrations from the more readily available

wellhead samples to the rarer downhole samples can also be a highly promising new application of DNN models in studies focused on the geochemistry of geothermal brines.

ACKNOWLEDGMENTS

This study is part of the subtopic ''Geoenergy'' in the program ''MTET—Materials and Technologies for the Energy Transition'' of the Helmholtz Association. The authors thank BMWK–PTJ project MALEG with the funding number 03EE4041B. Further, the BMBF (Federal Ministry of Education and Research) is thanked for funding the BrineMine project (Grant Number 033R190B) in the Client II framework. Finally, yet importantly, authors appreciate two anonymous reviewers who helped to improve the quality of the work with their invaluable comments.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY

Developed workflows for running the DNN models are fully documented and available on GitHub (https://github.com/Ali1990dashti/Ensemble-DNN, last access: 26 June 2025) and Zenodo (https://zenodo.org/records/15744274, last access: 26 June 2025) repositories.

DECLARATIONS

Conflict of Interest The authors declare that they have no conflict of interest.

APPENDIX

See Figures 12, 13, 14, 15, 16 and 17.

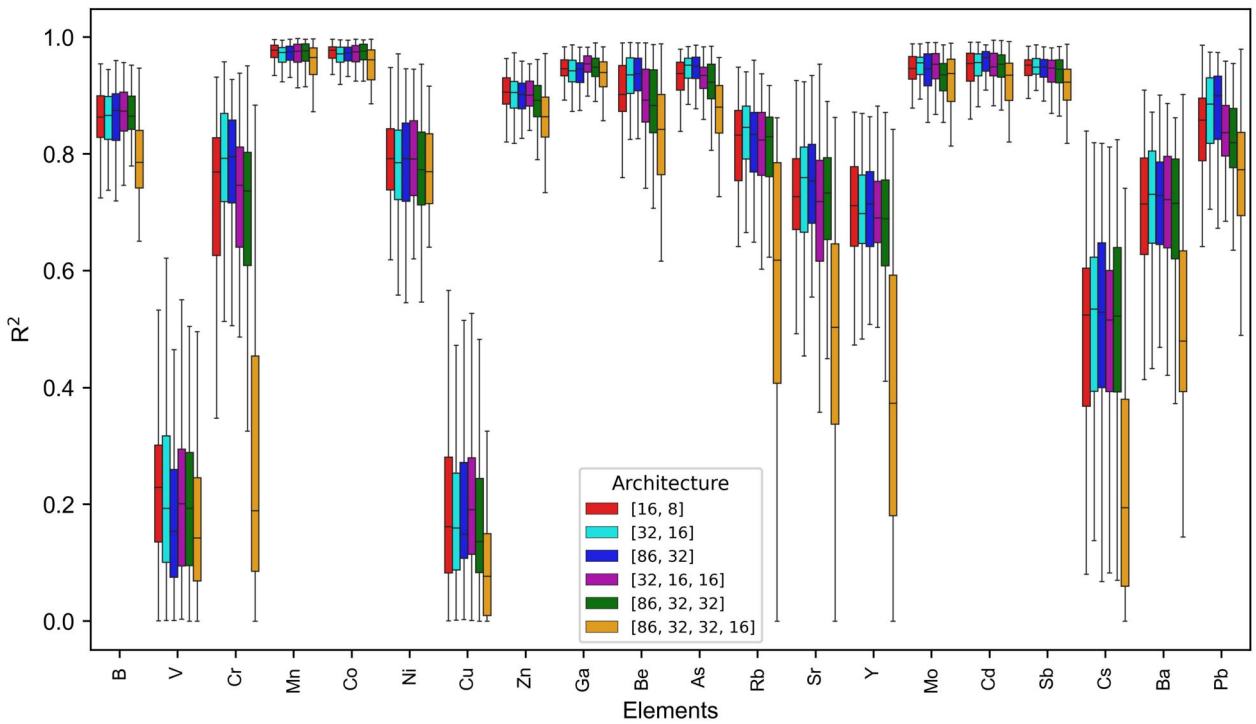

Figure 12. Accuracy of six tested architectures. Each color-coded box plot represents the R^2 score of 20 target variables. The NN model constructed by two hidden layers (with 32 and 16 neurons) shows the highest range of R^2 score and is chosen as the preferred architecture.

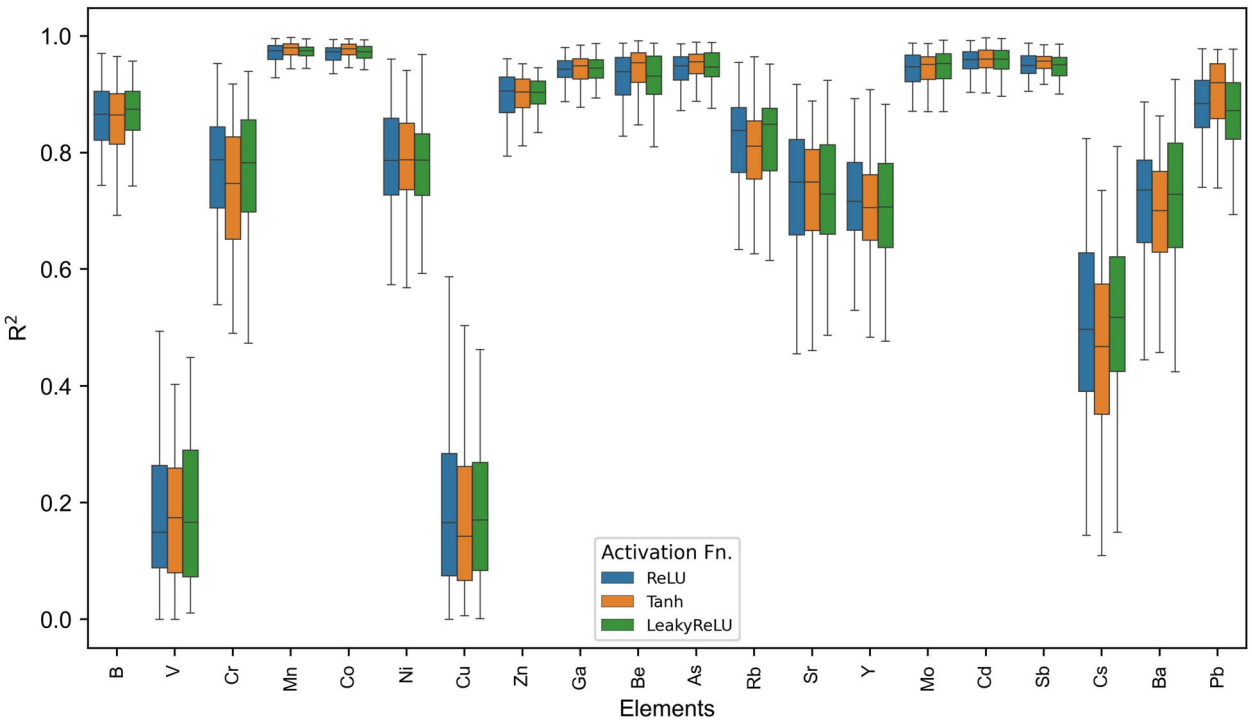

Figure 13. Accuracy of three tested activation functions. Each color-coded box plot represents the R^2 score of 20 target variables. The ReLU shows the highest range of R^2 score and is chosen as the preferred activation function.

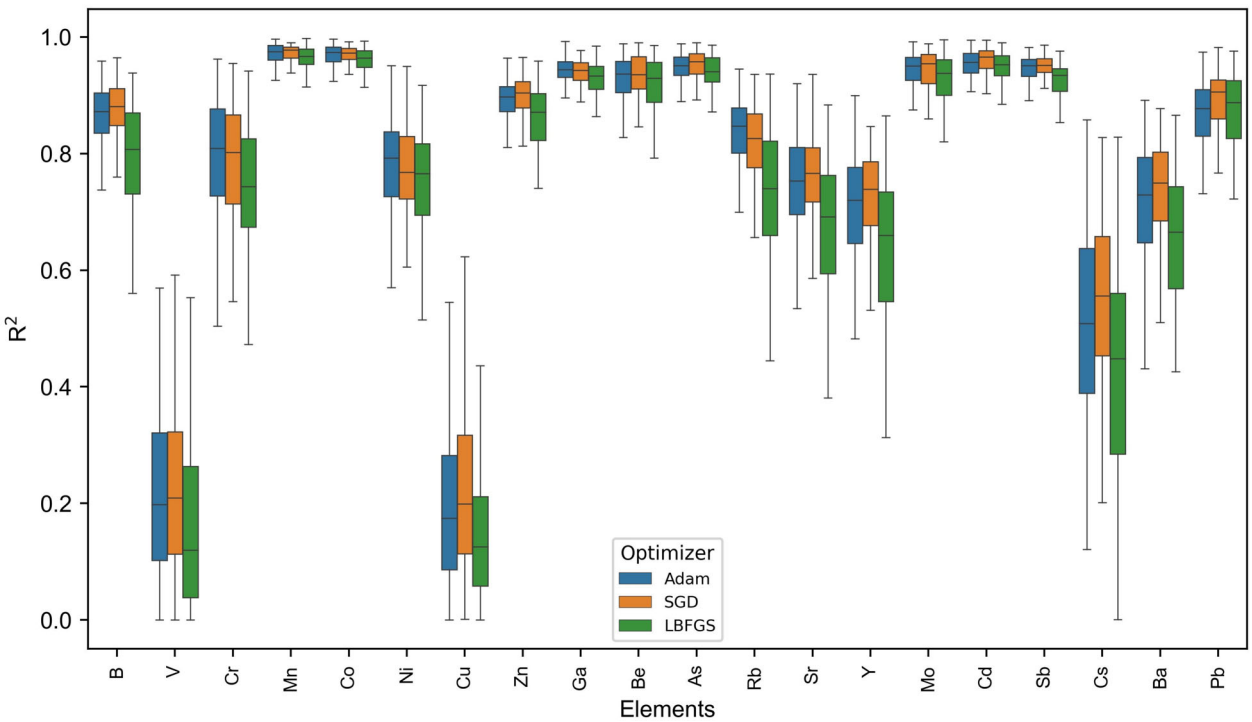

Figure 14. Accuracy of three tested optimizers. Each color-coded box plot represents the R^2 score of 20 target variables. The Adam and SGD algorithms show the highest range of R^2 score. Adam is chosen as the preferred optimizer due to its lower computational cost.

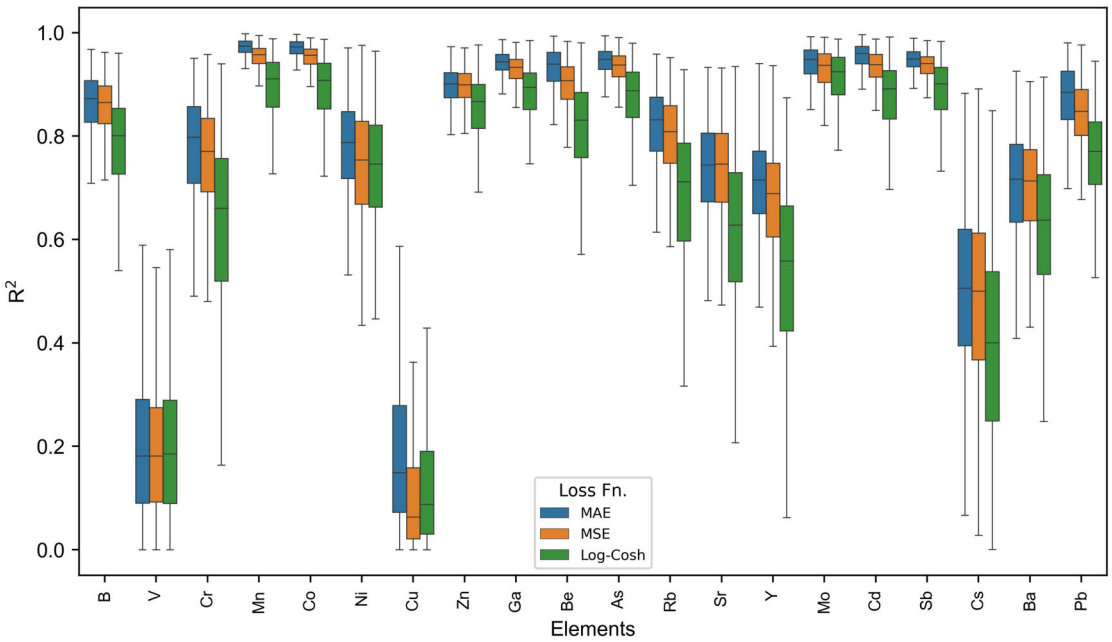

Figure 15. Accuracy of three tested loss functions. Each color-coded box plot represents the R^2 score of 20 target variables. The MAE shows the highest range of R^2 score and is chosen as the preferred loss function.

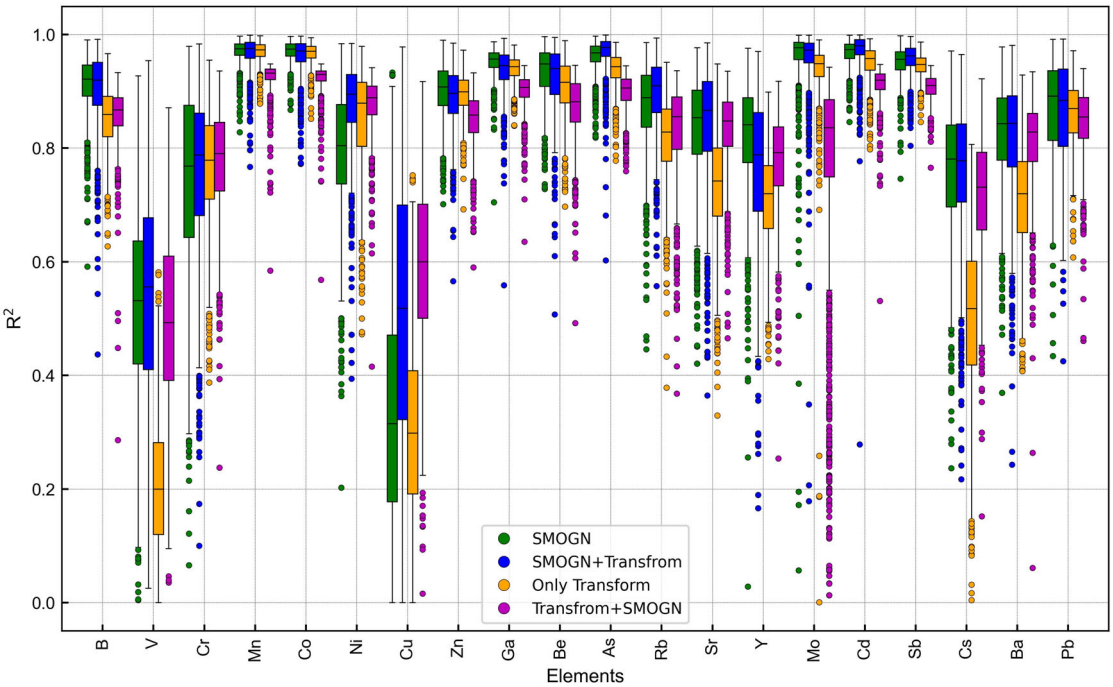

Figure 16. Distribution of R^2 scores for different modeling scenarios. Green boxes (SMOGN) and blue boxes (SMOGN followed by data transformation) are as introduced in Fig. 6. Orange boxes represent results from DNN models using data transformation techniques without SMOGN. Purple boxes correspond to DNN models where data transformation was applied first, followed by SMOGN as a preprocessing step.

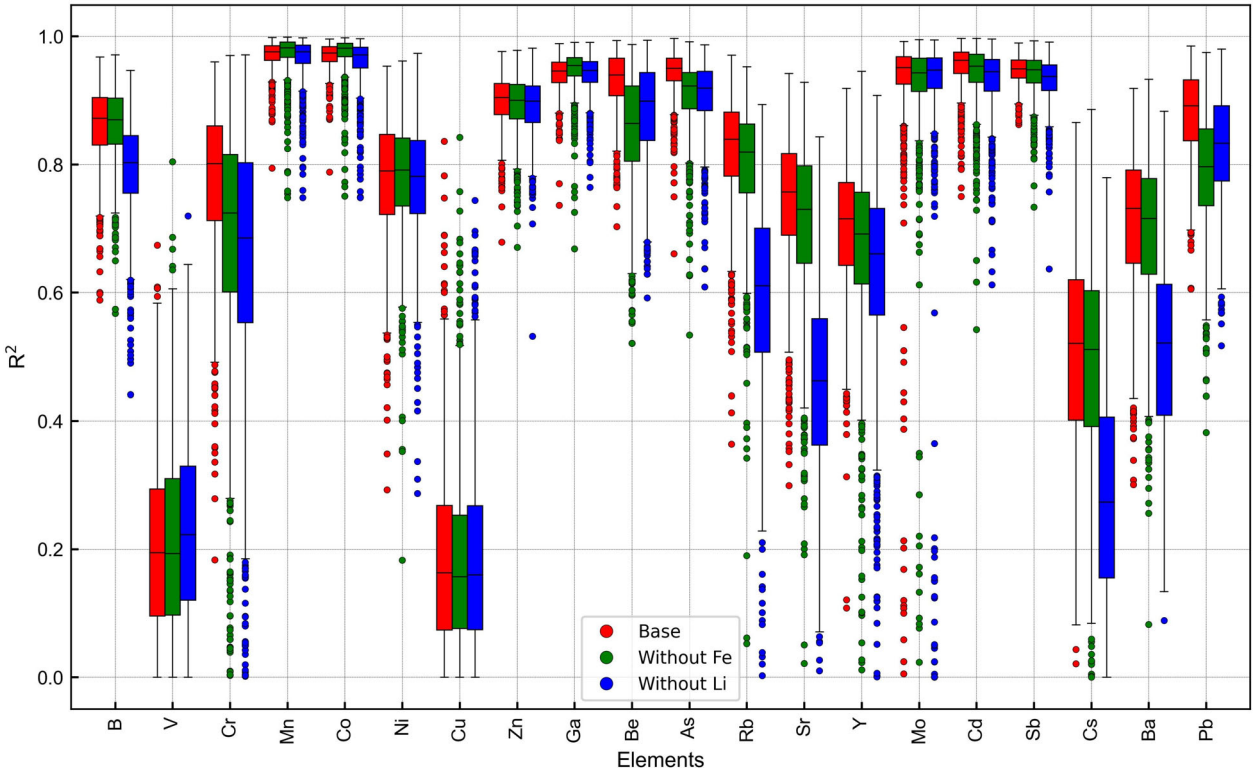

Figure 17. Comparison of accuracy of 1000 DNN models after removing either Fe or Li.

OPEN ACCESS

REFERENCES

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion, 76*, 243–297.

Al-Fakih, A., Abdulraheem, A., & Kaka, S. (2024). Application of machine learning and deep learning in geothermal resource development: Trends and perspectives. *Deep Underground Science and Engineering, 3*(3), 286–301.

Amari, S., Murata, N., Müller, K.-R., Finke, M., & Yang, H. (1995). Statistical theory of overtraining-Is cross-validation asymptotically effective? *Advances in Neural Information Processing Systems*, *8*.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology, 82*(4), 1059–1086.

Arnórsson, S., Bjarnason, J. Ö., Giroud, N., Gunnarsson, I., & Stefánsson, A. (2006). Sampling and analysis of geothermal fluids. *Geofluids, 6*(3), 203–216.

Baker, N., Alexander, F., Bremer, T., Hagberg, A., Kevrekidis, Y., Najm, H., et al. (2019). *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence* (No. None, 1478744) (p. None, 1478744). https://doi.org/10.2172/1478744.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing ma-

chine learning training data. *SIGKDD Explorations Newsletter, 6*(1), 20–29.
Blagus, R., & Lusa, L. (2013). Smote for high-dimensional class-imbalanced data. *BMC Bioinformatics, 14*(1), 106.
Bottou, L. (2012). Stochastic gradient descent tricks. In G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade* (Vol. 7700, pp. 421–436). Springer. https://doi.org/10.1007/978-3-642-35289-8_25.
Bourcier, W. L., Lin, M., & Nix, G. (2005). *Recovery of minerals and metals from geothermal fluids*. Lawrence Livermore National Lab.(LLNL).
Bourdeau, J. E., Zhang, S. E., Lawley, C. J. M., Parsa, M., Nwaila, G. T., & Ghorbani, Y. (2023). Predictive geochemical exploration: Inferential generation of modern geochemical data, anomaly detection and application to Northern Manitoba. *Natural Resources Research, 32*(6), 2355–2386.
Branco, P., Torgo, L., & Ribeiro, R. P. (2017). SMOGN: a Pre-processing Approach for Imbalanced Regression. In P. B. Luís Torgo & N. Moniz (Eds.), *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications* (Vol. 74, pp. 36–50). PMLR. https://proceedings.mlr.press/v74/branco17a.html.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Brunton, S. L., & Kutz, J. N. (2022). *Data-driven science and engineering: machine learning, dynamical systems, and control* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781009089517.
Bulska, E., & Ruszczyńska, A. (2017). Analytical techniques for trace element determination. *Physical Sciences Reviews, 2*(5), Article 20178002.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018.
Cosmo, RdeP., Pereira, FdeAR., Soares, E. J., & Ferreira, E. G. (2022). Addressing the root cause of calcite precipitation that leads to energy loss in geothermal systems. *Geothermics, 98*, Article 102272.
Dashti, A., Stadelmann, T., & Kohl, T. (2024). Machine learning for robust structural uncertainty quantification in fractured reservoirs. *Geothermics, 120*, Article 103012.
Davies, C. W. (1938). The extent of dissociation of salts in water. Part VIII. An equation for the mean ionic activity coefficient of an electrolyte in water, and a revision of the dissociation constants of some sulphates. *Journal of the Chemical Society (Resumed)*, pp 2093–2098.
Degen, D., Caviedes Voullième, D., Buiter, S., Hendricks Franssen, H.-J., Vereecken, H., González-Nicolás, A., & Wellmann, F. (2023). Perspectives of physics-based machine learning strategies for geoscientific applications governed by partial differential equations. *Geoscientific Model Development, 16*(24), 7375–7409.
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. *Computers & Geosciences, 140*, Article 104495.
Drumm, E., Bolton, R., & Peter-Borie, M. (2023). Making the Most of Existing Data: Challenges and Opportunities for Geothermal Brine Resource Exploration. In *The Fourth EAGE Global Energy Transition Conference and Exhibition* (pp. 1–5). Presented at the Fourth EAGE Global Energy Transition Conference and Exhibition, Paris, France: European Association of Geoscientists & Engineers. https://doi.org/10.3997/2214-4609.202321033.
Drüppel, K., Stober, I., Grimmer, J. C., & Mertz-Kraus, R. (2020). Experimental alteration of granitic rocks: Implications for the evolution of geothermal brines in the Upper Rhine Graben, Germany. *Geothermics, 88*, Article 101903.
Ellis, A. J., & Mahon, W. A. J. (1964). Natural hydrothermal systems and experimental hot-water/rock interactions. *Geochimica Et Cosmochimica Acta, 28*(8), 1323–1357.
Enkhsaikhan, M., Holden, E.-J., Duuring, P., & Liu, W. (2021). Understanding ore-forming conditions using machine reading of text. *Ore Geology Reviews, 135*, Article 104200.
Fournier, R. O. (1966). Estimation of underground temperatures from the silica content of water from hot springs and wet-steam wells. *American Journal of Science, 264*, 685–697.
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232.
Fushiki, T. (2011). Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing, 21*(2), 137–146.
Gallup, D. L. (1998). Geochemistry of geothermal fluids and well scales, and potential for mineral recovery. *Ore Geology Reviews, 12*(4), 225–236.
Giggenbach, W. F. (1988). Geothermal solute equilibria. Derivation of Na-K-Mg-Ca geoindicators. *Geochimica et Cosmochimica Acta, 52*(12), 2749–2765.
Goldberg, V. (2024). The potential of direct mineral extraction from geothermal fluids. *Karlsruher Institut für Technologie (KIT)*. https://doi.org/10.5445/IR/1000171335.
Goldberg, V., Winter, D., Nitschke, F., Held, S., Groß, F., Pfeiffle, D., et al. (2023). Development of a continuous silica treatment strategy for metal extraction processes in operating geothermal plants. *Desalination, 564*, Article 116775.
Gunnlaugsson, E., Ármannsson, H., Thorhallsson, S., & Steingrímsson, B. (2014). *Problems in geothermal operation–scaling and corrosion* (pp. 1–18). United Nations University.
Gurgenc, E., Altay, O., & Altay, E. V. (2024). AOSMA-MLP: A novel method for hybrid metaheuristics artificial neural networks and a new approach for prediction of geothermal reservoir temperature. *Applied Sciences, 14*(8), 3534.
Hardardottir, V., Hannington, M., Hedenquist, J., Kjarsgaard, I., & Hoal, K. (2010). Cu-rich scales in the Reykjanes Geothermal System, Iceland. *Economic Geology, 105*(6), 1143–1155.
Hawkins, A. J., & Tester, J. W. (2018). Geothermal Systems. In W. M. White (Ed.), *Encyclopedia of Geochemistry* (pp. 592–597). Springer International Publishing. https://doi.org/10.1007/978-3-319-39312-4_106.
Heid, E., McGill, C. J., Vermeire, F. H., & Green, W. H. (2023). Characterizing uncertainty in machine learning for chemistry. *Journal of Chemical Information and Modeling, 63*(13), 4012–4029.
Hsieh, W. W. (2009). *Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels*. Cambridge University Press.
Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning, 110*(3), 457–506.
Jo, T. (2021). *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-65900-4.
Kaasalainen, H., Stefánsson, A., Giroud, N., & Arnórsson, S. (2015). The geochemistry of trace elements in geothermal fluids, Iceland. *Applied Geochemistry, 62*, 207–223.
Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences, 46*(3), 240–254.
Ketkar, N., & Moolayil, J. (2021). Introduction to PyTorch. In *Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch* (pp. 27–91). Apress. https://doi.org/10.1007/978-1-4842-5364-9_2.
Kingma, D. P. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., et al. (2023). What is machine learning, arti-

ficial neural networks and deep learning?—Examples of practical applications in medicine. *Diagnostics, 13*(15), Article 2582.
Lau, M. M., & Lim, K. H. (2018). Review of adaptive activation function in deep neural network. In *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)* (pp. 686–690). IEEE.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
Li, D.-C., Lin, W.-K., Chen, C.-C., Chen, H.-Y., & Lin, L.-S. (2018). Rebuilding sample distributions for small dataset learning. *Decision Support Systems, 105*, 66–76.
Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: A big comparison for NAS. arXiv. https://doi.org/10.48550/ARXIV.1912.06059.
Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming, 45*(1), 503–528.
Liu, L., Lu, J., Tao, C., Liao, S., Su, C., Huang, N., & Xu, X. (2022). Fuzzy forest machine learning predictive model for mineral prospectivity: A case study on Southwest Indian Ridge 48.7°E–50.5°E. *Natural Resources Research, 31*(1), 99–116.
López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113–141.
McClenny, L. D., & Braga-Neto, U. M. (2023). Self-adaptive physics-informed neural networks. *Journal of Computational Physics, 474*, Article 111722.
Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., et al. (2025). The effects of data quality on machine learning performance on tabular data. *Information Systems, 132*, Article 102549.
Molnar, C., Casalicchio, G., & Bischl, B. (2020). Quantifying model complexity via functional decomposition for better post-hoc interpretability. In P. Cellier & K. Driessens (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 1167, pp. 193–204). Springer International Publishing. https://doi.org/10.1007/978-3-030-43823-4_17.
Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array (San Diego, Calif.), 16*, Article 100258.
Ngombe, O. G. D., Walter, J., Chesnaux, R., & Molson, J. (2024). Application of hierarchical cluster analysis and principal component analysis to identify compositional trends of brine in crystalline basements and sedimentary basins. *Applied Geochemistry, 169*, Article 106030.
Nitesh, V. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal Of Artificial Intelligence Research, 16*(1), 321.
Nitschke, F., Scheiber, J., Kramar, U., & Neumann, T. (2014). Formation of alternating layered Ba-Sr-sulfate and Pb-sulfide scaling in the geothermal plant of Soultz-sous-Forêts. *Neues Jahrbuch Für Mineralogie - Abhandlungen, 191*(2), 145–156.
Nitschke, F., Held, S., Himmelsbach, T., & Kohl, T. (2017). THC simulation of halite scaling in deep geothermal single well production. *Geothermics, 65*, 234–243.
Nitschke, F. (2024). MALEG-Maschinelles Lernen zur Verbesserung der Effizienz Geothermischer Energienutzung. *GeoTHERM Abstracts Band*, *3*.
Nriagu, J. O., & Pacyna, J. M. (1988). Quantitative assessment of worldwide contamination of air, water and soils by trace metals. *Nature 333*, 134e139.
Oh, H.-J., & Lee, S. (2010). Application of artificial neural network for gold-silver deposits potential mapping: A case study of Korea. *Natural Resources Research, 19*(2), 103–124.
Okoroafor, E. R., Smith, C. M., Ochie, K. I., Nwosu, C. J., Gudmundsdottir, H., & (Jabs) Aljubran, M. (2022). Machine learning in subsurface geothermal energy: Two decades in review. *Geothermics*, *102*, 102401.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.
Pauwels, H., Fouillac, C., & Fouillac, A.-M. (1993). Chemistry and isotopes of deep geothermal saline fluids in the Upper Rhine Graben: Origin of compounds and water-rock interactions. *Geochimica et Cosmochimica Acta, 57*(12), 2737–2749.
Puzyrev, V., Zelic, M., & Duuring, P. (2023). Applying neural networks-based modelling to the prediction of mineralization: A case-study using the Western Australian Geochemistry (WACHEM) database. *Ore Geology Reviews, 152*, Article 105242.
Rabczuk, T., & Bathe, K.-J. (Eds.). (2023). *Machine Learning in Modeling and Simulation: Methods and Applications*. Springer International Publishing. https://doi.org/10.1007/978-3-031-36644-4.
Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics, 378*, 686–707.
Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv preprint* arXiv:1710.05941.
Rayat, C. S. (2018). Measures of Dispersion. In *Statistical Methods in Medical Research* (pp. 47–60). Springer Singapore. https://doi.org/10.1007/978-981-13-0827-7_7.
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 Geology Reviews, 71*, 804–818.
Sakia, R. M. (1992). The box-cox transformation technique: A review. *The Statistician, 41*(2), 169–178.
Saleh, R. A., & Saleh, A. K. (2022). Statistical properties of the log-cosh loss function used in machine learning. *arXiv preprint* arXiv:2208.04564.
Salehinejad, H., Sankar, S., Barfett, J., Colak, E., & Valaee, S. (2018). Recent Advances in Recurrent Neural Networks. *arXiv*. https://doi.org/10.48550/ARXIV.1801.01078.
Sanjuan, B., Millot, R., Innocent, Ch., Dezayes, Ch., Scheiber, J., & Brach, M. (2016). Major geochemical characteristics of geothermal brines from the Upper Rhine Graben granitic basement with constraints on temperature and circulation. *Chemical Geology, 428*, 27–47.
Santurkar, S., Tsipras, D., Ilyas, A., & Madry, A. (2018). How does batch normalization help optimization? *Advances in neural information processing systems*, *31*.
Schwarzbauer, J., & Jovančićević, B. (2020). Analytical Quality Control. In *Introduction to Analytical Methods in Organic Geochemistry* (pp. 129–134). Springer International Publishing. https://doi.org/10.1007/978-3-030-38592-7_6.
Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Sci, 6*(12), 310–316.
Sola, J., & Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science, 44*(3), 1464–1468.
Stober, I., & Bucher, K. (2012). *Geothermie*. Springer.
Stober, I., & Bucher, K. (2015). Hydraulic and hydrochemical properties of deep sedimentary reservoirs of the Upper Rhine Graben, Europe. *Geofluids, 15*(3), 464–482. https://doi.org/10.1111/gfl.12122.

Talaei-Khoei, A., & Motiwalla, L. (2023). A new method for improving prediction performance in neural networks with insufficient data. *Decision Analytics Journal, 6*, Article 100172.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272.

Wang, Y. E., Wei, G.-Y., & Brooks, D. (2019). Benchmarking TPU, GPU, and CPU Platforms for Deep Learning. *arXiv*. https://doi.org/10.48550/ARXIV.1907.10701.

Wang, L., & Zhou, X.-H. (2007). Assessing the adequacy of variance function in heteroscedastic regression models. *Biometrics, 63*(4), 1218–1225.

Wanner, C., Eichinger, F., Jahrfeld, T., & Diamond, L. W. (2017). Causes of abundant calcite scaling in geothermal wells in the Bavarian Molasse Basin, Southern Germany. *Geothermics, 70*, 324–338.

Weisberg, S. (2001). Yeo-Johnson power transformations. Department of Applied Statistics, University of Minnesota. Retrieved June, 1, 2003.

Willcox, K. E., Ghattas, O., & Heimbach, P. (2021). The imperative of physics-based modeling and inverse theory in computational science. *Nature Computational Science, 1*(3), 166–168.

Wu, Z., Xue, M., Hou, B., & Liu, W. (2022). Cross-domain decision making with parameter transfer based on value function. *Information Sciences, 610*, 777–799.

Xu, P., Ji, X., Li, M., & Lu, W. (2023). Small data machine learning in materials science. *Npj Computational Materials, 9*(1), Article 42.

Xu, J., Li, Z., Du, B., Zhang, M., & Liu, J. (2020). Reluplex made more practical: Leaky ReLU. In *2020 IEEE Symposium on Computers and Communications (ISCC)* (pp. 1–7). Presented at the 2020 IEEE Symposium on Computers and Communications (ISCC). IEEE. https://doi.org/10.1109/ISCC50000.2020.9219587.

Yao, Y., Rosasco, L., & Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation, 26*(2), 289–315.

Ystroem, L. H., Vollmer, M., Kohl, T., & Nitschke, F. (2023). AnnRG—An artificial neural network solute geothermometer. *Applied Computing and Geosciences, 20*, Article 100144.

Zhang, Y., & Ling, C. (2018). A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials, 4*(1), Article 25.

Zhang, S. E., Nwaila, G. T., Bourdeau, J. E., & Ashwal, L. D. (2021). Machine learning-based prediction of trace element concentrations using data from the Karoo large igneous province and its application in prospectivity mapping. *Artificial Intelligence in Geosciences, 2*, 60–75.

Zhang, S. E., Bourdeau, J. E., Nwaila, G. T., & Ghorbani, Y. (2022). Advanced geochemical exploration knowledge using machine learning: Prediction of unknown elemental concentrations and operational prioritization of re-analysis campaigns. *Artificial Intelligence in Geosciences, 3*, 86–100.

Zhang, P., Zhang, Z., Yang, J., & Cheng, Q. (2023). Machine learning prediction of ore deposit genetic type using magnetite geochemistry. *Natural Resources Research, 32*(1), 99–116.

Zhang, S. E., Bourdeau, J. E., Nwaila, G. T., Parsa, M., & Ghorbani, Y. (2024a). Denoising of geochemical data using deep learning-implications for regional surveys. *Natural Resources Research, 33*(2), 495–520.

Zhang, S. E., Lawley, C. J. M., Bourdeau, J. E., Nwaila, G. T., & Ghorbani, Y. (2024b). Workflow-induced uncertainty in data-driven mineral prospectivity mapping. *Natural Resources Research, 33*(3), 995–1023.

Zhang, G., Wang, C., Xu, B., & Grosse, R. (2018). Three mechanisms of weight decay regularization. *arXiv*. https://doi.org/10.48550/ARXIV.1810.12281.

Zhao, T., Wang, S., Ouyang, C., Chen, M., Liu, C., Zhang, J., et al. (2024a). Artificial intelligence for geoscience: Progress, challenges, and perspectives. *The Innovation, 5*(5), Article 100691.

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024b). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review, 57*(4), 99.