A Protolith Reconstruction Model (PRM) for Metabasalt: Quantitative Protolith and Mass Transfer Estimation Based on Machine-learning Approach

Satoshi Matsuno
Tohoku University
Masaoki Uno (uno@geo.kankyo.tohoku.ac.jp)
Tohoku University
Atsushi Okamoto
Tohoku University
Noriyoshi Tsuchiya
Tohoku University

Research Article

Keywords: fluid–rock interaction, Earth, including metamorphism, metasomatism, PRMs

Posted Date: May 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-558656/v1

A protolith reconstruction model (PRM) for metabasalt: Quantitative protolith and mass transfer estimation based on machine-learning approach

Satoshi Matsuno[1], Masaoki Uno[1]*, Atsushi Okamoto[1], and Noriyoshi Tsuchiya[1]

[1]*Graduate School of Environmental Studies, Tohoku University, 6-6-20 Aza-Aoba, Aramaki, Aobaku, Sendai 980-8579, Japan*

*e-mail: uno@geo.kankyo.tohoku.ac.jp

Submitted to the journal *Scientific Reports*

Submitted 25th May 2021

Abstract

Mass transfer in rocks provides a direct record of fluid–rock interaction within the Earth, including metamorphism, metasomatism, and hydrothermal alteration. However, mass transfer analyses are usually limited to local reaction zones where the protoliths are evident in outcrops (1–100 m in scale), from which regional mass transfer can be only loosely constrained due to uncertainty in protolith compositions. In this study, we developed protolith reconstruction models (PRMs) for metabasalt based on a machine learning approach. We constructed PRMs through learning multi-element correlations among basalt compositional datasets, including mid-ocean ridge, ocean island, and island arc basalts. The PRMs were designed to estimate trace-element compositions from inputs of 2–9 selected trace elements, and basalt trace-element compositions (e.g., Rb, Ba, U, K, Pb, Sr, and rare earth elements) were estimated from only four inputs with a reproducibility of ~0.1 $\log_{10}$ units (i.e., ±25%). Using Th, Nb, Zr, and Ti, which are typically immobile during metamorphism, as input elements, the PRM was verified by applying it to seafloor altered basalt with known protoliths. When suitable immobile elements are incorporated, a PRM can yield unbiased and accurate mass transfer analysis of any metabasalt with unknown protolith.

Background

Chemical alteration of rocks, or mass transfer, provides direct evidence for fluid–rock interactions within the Earth and represents various geochemical processes such as seafloor alteration, subduction zone metamorphism, geothermal fluid activity, and fault zone processes. Mass transfer analyses for subduction-related metamorphism reveal element transport via dehydration reactions in subducting slabs[1–3] and element cycling in subduction zones[4,5] that are chemically linked to arc basalt[6,7]. The distribution of mass transfer in a regional metamorphic belt can reveal the spatial distributions of fluid flow in the crust and mantle[8–10]. Mass transfer analyses in mineral-filled veins and fault zones are related to the fluid flux[11], duration of fluid infiltration[2,12,13], and/or degree of fault heating[14,15]. Hydrothermal alteration of Archean seafloor basalt is known to be linked to the chemistry of seawater[16,17]. Therefore, analyses of mass transfer in chemically altered rocks are essential to better understand fluid-related processes within the Earth and the evolution of surface environments.

Mass transfer analyses are generally achieved by comparing the compositions of protolith with those of metamorphosed/altered rocks. Mass transfer at the outcrop scale (<100 m) can be estimated by comparing compositions of altered rocks with those of adjacent unaltered host rocks[18–20]. For larger scales (i.e., comparisons of rocks from various metamorphic belts), mass transfer can be qualitatively evaluated by comparing chemical differences between metamorphosed rocks (i.e., metabasalt) and their likely protoliths[4,5] (i.e., mid-ocean ridge basalt or MORB). Such mass transfer analyses, which compare protolith and metamorphosed/altered rocks, have contributed to understanding material recycling in subduction zones[4,5].

In many cases, information about the exact protoliths of metamorphosed/altered

rocks is unavailable, except for cases where the protoliths are evident in outcrops. As the compositional variations of likely protoliths (i.e., basalt and sediment) are generally large[21–24], it is difficult to quantitatively evaluate the amount of mass transfer for each sample. Recent analyses of regional metamorphic belts have also revealed that protoliths of metamorphic rocks differ in their depositional ages and/or tectonic settings among different units or grades of rock[5,25,2627,28], suggesting that it is unrealistic to assume a uniform protolith composition in regional metamorphic belts or alteration zones. Therefore, to better understand quantitative mass transfer, estimation of protolith compositions from individual samples is required.

Natural observations and experiments have revealed the tendency of mass transfer to differ according to the elements, pressure, temperature, and/or fluid chemistry involved. High-field-strength elements (HFSEs) generally show little mass transfer in seafloor alteration[29–31] and low solubility in typical pressure–temperature (P–T) conditions of metamorphism[32–34], and are generally considered as “immobile” [4,5,29–31,35]. Other elements, such as large-ion lithophile elements (LILEs; e.g., Rb, Ba, and Sr), have large mass transfer in seafloor alteration and high solubility in metamorphic fluids[31–34]. The mobility of these elements during metamorphism has been confirmed by various mass transfer analyses of mineral veins and alteration zones[1,3,36]. Compilations of mass transfer in various metamorphic conditions and environments have suggested that the mobility of HFSEs decreases roughly in the order of rare earth elements (REEs) > U > Nb > Ti > Th ~ Zr for high-pressure subduction zone environments[35]. These elements are widely considered as immobile elements and are therefore used for discriminating the tectonic settings of metabasalt. The general success of discrimination diagrams[37,38] suggests that these immobile elements retain protolith information, meaning that it should

be possible to use these elements to reconstruct protolith compositions from metamorphosed/altered rocks.

Advances in data science have provided tools for extracting information from large numbers of multidimensional data. In particular, machine learning is an effective way of recognizing complex patterns in images and extracting information from multidimensional table data. The recent increase in data held in geochemical compositional databases (e.g., PetDB and GeoRock) has allowed machine learning to become established in its application to research problems in geochemistry. For example, machine learning has been successfully applied to discriminate tectonic settings of basalt from chemical componsitions[24], and classify metamorphic protoliths from major elements[39]. However, previous applications of machine-learning to geochemical data have been limited mainly to classification problems and have not dealt with regression problems (i.e., predicting quantitative chemistry). For quantitative mass transfer/protolith analyses, a new quantitative and predictive machine-learning scheme for geochemical data is needed.

In this study, we develop protolith reconstruction models (PRMs) that estimate protolith compositions of metabasalt using machine learning. First, using a basalt whole-rock compositional dataset, we develop empirical models that learn multi-elemental correlations among the dataset and then estimate trace-element compositions of basalt based on the contents of a few (two to nine) elements. The numbers and combinations of input elements are optimized to precisely predict the output contents. Results show that basalt trace-element contents (i.e., Rb, Ba, U, K, Pb, Sr, and REEs) can be estimated from only four element contents (i.e., Th, Nb, Zr, and Ti). Finally, we apply the selected four-element PRM to seafloor altered basalt and metabasalt, and demonstrate the validity of

the model and examples of mass transfer analyses for metamorphic rocks.

Model description

PRMs were developed through machine learning of a compositional dataset of basalt. The empirical models were designed to estimate contents of a particular trace element as an output from between two and nine HFSE contents as inputs (Fig. 1). Basalt compositional data were taken from the PetDB database (https://search.earthchem.org/). Trace-element contents of 16 elements (Rb, Ba, U, K, La, Ce, Pb, Sr, Nd, Y, Yb, Lu, Zr, Th, Ti, and Nb) from a total of 8253 basalt samples were compiled, including data for mid-ocean ridge basalt (MORB), ocean island basalt (OIB), and island arc basalt (IAB) as potential protoliths of metabasalt. The distribution of compositional data for these basalts varies depending on the elements, with Th and Ba having relatively high correlations, Zr and Y having variable correlations depending on the type of basalt, and Nb and Sr having low correlations (Fig. 1a–c). These data distributions suggest non-linear and multidimensional relationships among the contents of these 16 elements.

PRMs were constructed with the machine learning algorithm of the gradient boosting decision tree (GBDT), with separate training/test data being used to evaluate the model (Fig. 1d). The GBDT is one of several decision tree algorithms that are capable of fitting complex datasets (i.e., non-linear structural data) and which perform with high accuracy[40]. The models were evaluated by the root mean squared error (RMSE) in log space between the estimated output and the measured data:

$$E_i = \frac{1}{N} \sum_j^{sample} \sqrt{\left(\log_{10} y_{i,j}^{estimated} - \log_{10} y_{i,j}^{test}\right)^2} \qquad (1)$$

where E_i is the RMSE for element i, N is the number of samples, $y_{i,j}^{\mathrm{estimated}}$ is the estimated content of element i in sample j, and $y_{i,j}^{\mathrm{test}}$ is the measured content of element i in sample

j.

The elements used as input and output elements were determined from the degree of mass transfer reported in previous studies; LILEs are mobile during fluid activity in subduction zones, contact metamorphism, and seafloor alteration, whereas HFSEs are immobile during fluid activity[29–31,33–35]. The order of mobility of HFSEs is REEs > U > Nb > Ti > Th ~ Zr, as determined from observations of natural metamorphic rocks and experiments on metamorphic conditions[35]. For this reason, combinations of 2–9 elements from Zr, Th, Ti, Nb, La, Ce, Nd, Yb, and Lu were selected as input elements, and Rb, Ba, U, K, La, Ce, Pb, Sr, Nd, Y, Yb, Lu, Zr, Th, Ti, and Nb were selected as output elements. Elements used as output elements were excluded from input elements. For each model (combinations of particular input elements and an output element), basalt compositional data were chosen to ensure that there were no missing values for input and output elements in the utilized dataset (typically 3000–5000 samples).

Model dependence on input elements

We firstly selected the numbers and combinations of input elements to estimate basalt composition. Machine-learning models were constructed for each combination of input and output elements (e.g., input: Th, Nb, and Zr; output: Rb). Therefore, the number of possible combinations of the input elements is $2^9 - 1 = 511$. As each machine-learning model was developed for each output element independently, 5872 machine learning models were developed in total.

Examples of estimated compositions for a specific basalt sample are shown for different sets of input elements in Figure 2a, b, and c. The reproducibility of the estimation is dependent largely on input elements. For example, in the case of the input elements

being Yb and Lu, the reproducibility (i.e., the difference between the actual and estimated compositions) for each element is large (Fig. 2a; i.e., >1 in $\log_{10}$ units); in contrast, when the input elements are Th and Ti, or Nd, Ti, Yb and Lu, the reproducibility for each element is greatly improved and is <0.2 in $\log_{10}$ units (Fig. 2b and c). Consequently, this dependence of reproducibility on input elements indicates that the numbers and combinations of elements affect the estimation of composition.

Effects of input elements were evaluated by taking averages of RMSE scores. Figure 2d shows the average RMSE scores of all output elements for each combination of input elements (511 cases; best model score of Zr, Th, Nb, La, and Yb = 0.089; worst model score of Lu = 0.32). The top 12% of models all include Th, and 18% of models include Nb. Figure 2e shows average RMSE scores for all of the models classed by the number of input elements. For the case where the number of input elements is more than four, the averaged RMSE scores converge around 0.11 (0.115 for four input elements and 0.110 for five input elements). The effect of each input element was evaluated by taking the average of all of the models containing a particular element as inputs (Fig. 2f). Models using Th and Nb as inputs have slightly lower average scores than the other models. These results indicate that the trace-element composition of basalt can be suitably estimated from only four (or five) input elements. The RMSE score of all output elements does not change substantially with the combinations of input elements.

The top 37% of models fall within the range of RMSE $\leq$ 0.11. The three best models consist of five input elements: Th, Nb, La, Zr, and Yb (RMSE = 0.089); Th, Nb, La, Nd, and Lu (RMSE = 0.091); and Th, Nb, La, Zr, and Lu (RMSE = 0.092). Among the models with four input elements, the best combinations are Th, Nb, La, and Y (RMSE = 0.092), Th, Nb, La, and Lu (RMSE = 0.092), and Th, Nb, Ce, Yb (RMSE = 0.093). The

top 37% of models (189 combinations of input elements) have almost identical RMSE values (0.09–0.11), or reproducibilities of ±0.09–0.11 in $\log_{10}$ units, or ±24%–28%.

The performance of a particular output element improves in cases where input elements have similar incompatibility to that of the target element. For example, the RMSE of Ce is improved when the input combinations include La and Nd (Supplementary Fig. 1). The dependence of RMSE on input elements indicates that input elements with closer compatibility to that of the output element contain more identifying information on protolith composition. For example, the RMSE of Ce gradually improves when the input elements have closer compatibility with Ce[41]. As a whole, to improve the overall estimation, it is necessary to choose input elements that have a wide range of incompatibility when combined.

The constructed models for estimating basalt composition can be used to reconstruct the protolith composition of metamorphosed or altered basaltic rocks. Assuming that the contents of immobile elements in metabasalt are identical to those of its protolith, these contents can be assigned as input elemental contents of PRMs (Fig. 1d). Then, the amounts of transfer of the other elements (mobile elements during metamorphism or alteration) can be obtained by comparing their observed and predicted contents. It is noted that elements that are immobile during alteration or metamorphism may vary from case to case[35]; as such, users can choose PRMs with other input combinations by selecting the appropriate four to five immobile elements for the geochemical system of interest.

PRM reproducibility using the example of models incorporating Th, Nb, Zr, and Ti

In the following application to metabasalt, the combination of the four elements

of Th, Nb, Zr, and Ti was chosen as the input of the PRM, as these elements are the most immobile elements from both natural observations and experiments[32–35] and have a wide variety of compatibility[41]. The PRM was constructed by using ~3000 basalt samples (i.e., data containing all of the input elements and an output element) and can estimate protolith compositions with an RMSE of ~0.1 (i.e., ±25%; Fig. 2d).

We applied the PRM using Th, Nb, Zr, and Ti as input elements to the test data of the compositional dataset for basalt. The estimated contents show largely linear relationships with the raw (measured) contents in log–log space, with a slope of 1.0 (Fig. 3). These results show that the PRM closely reproduces individual elements through a wide range of their contents. Scatter plots of La, Ce, Nd, Y, Yb, and Lu show relatively small deviations from the 1:1 line and show almost no dependence on tectonic setting. In comparison, distributions of Rb, Ba, U, K, Pb, and Sr have relatively large dispersions. In particular, dispersions of Rb and K are apparent for low contents of elements. These results also affect the distribution of reproducibility of each element (Fig. S2). The reproducibility of Rb, U, K, Pb, and Sr differs according to tectonic setting, with the other elements showing no or only slight dependence on tectonic setting. The distributions of reproducibility for MORB have a wider range than those for OIB and IAB for Rb, and U, whereas those for IAB are slightly wider than those for MORB and/or OIB for Ce, Sr, and Nd.

One explanation for the dependence of element content on tectonic setting is the analytical detection limit. In particular, the raw data for K have identical values for samples with low contents ($\leq 10^3$ ppm), and the reproducibility of such data are large, probably because of the detection limit of K in X-ray fluorescence (XRF) analyses and/or the resolution of the original dataset (i.e., ~0.1 wt.%). An alternative explanation is

seafloor alteration, for which Rb, Ba, U, K, Pb, and Sr are mobile[30,31,42]. Some samples of MORB and IAB might have already undergone mass transfer by hydrothermal alteration because parts of these were collected from the ocean seafloor, with the sample data being correspondingly affected. It is likely that some of the "fresh" basalts in the training data have been affected by the detection limit and/or seafloor alteration, contributing to enlargement of the reproducibility of models; the estimation of Rb, Ba, U, K, and Pb can be potentially changed by removing such alteration-affected data.

Examples of PRM estimation for each tectonic setting are presented in Figure 4. These estimations were derived by models incorporating only Th, Nb, Zr, and Ti. The varied compositional patterns of different tectonic settings can be reasonably estimated from these four input elements only, within a reproducibility of ±25%.

Application to seafloor altered basalt: validation of the PRM

To validate the PRM-reconstructed compositions, we applied the four-element PRM incorporating Th, Nb, Zr, and Ti to seafloor altered basalt, whose protolith composition has been already estimated from fresh volcanic glass[31]. The reconstructed protolith compositions were compared with the volcanic glass compositions identified as protolith[31].

Altered-sample compositions were derived from Ocean Drilling Program (ODP) Site 801[31] (http://www-odp.tamu.edu/). ODP Site 801 is located in 170 Ma crust to the east of Mariana Island in the Pacific plate. Altered minerals are commonly composed of saponite and calcite. PRM was applied to samples 801-MORB-110-222_ALL and 801C Super, which are characterized by enrichment in Rb, U, K, and Li.

The PRM was used to reconstruct protolith compositions from altered basalt.

The PRM-based primitive-mantle-normalized protolith compositions have smooth patterns, and elements with higher compatibility have higher contents[41] (Fig. 5a and c). These PRM-based compositions are within the range of protolith compositions estimated from fresh glass. Protolith composition can be accurately reconstructed from seafloor basalt.

Furthermore, mass mobility (i.e., the ratio of element content in the altered sample to that in the protolith) was calculated from the estimated protolith composition and altered sample composition (Fig. 5b and d). Compared with previous estimates of mobility[31], results from the PRM show an accurate estimation of mass mobility, ensuring the accurate reconstruction of protolith composition from altered or mass-transferred samples within the uncertainty of the estimation (±0.1 in $\log_{10}$ units or ±25%). The PRM can therefore reconstruct protolith composition for metabasalt.

Application to metabasalt: An example of metamorphic mass transfer analysis using a PRM

We also applied the PRM incorporating Th, Nb, Zr, and Ti as inputs to an eclogite sample (Z139-6) obtained from central Zambia within the Zambezi belt, part of the Pan-African orogenic system between the Conga and Kalahari cratons[5]. Peak metamorphic conditions have been estimated as 2.6–2.8 GPa and 630–690 °C[43]. The sample is porphyroblastic eclogite composed of omphacite, garnet, kyanite, and quartz that has replaced plagioclase. The sample shows no evidence of prograde blueschist- or amphibolite-facies metamorphism but displays evidence of direct eclogitization from gabbroic assemblages. Reaction textures and chemical analyses have revealed that this sample records prograde eclogitization and mass transfer influenced by fluid derived from

the serpentinized lithospheric mantle of a subducting slab[5]. On the basis of comparisons with empirically determined likely protolith composition, the fluid is inferred to have been strongly undersaturated in light REEs (LREEs) and LILEs[5]. We applied the PRM to sample Z139-6, which is characterized by depletion in Rb, Ba, La, Ce, Sr, and Nd.

The reconstructed primitive-mantle-normalized protolith composition shows that elements with higher compatibility have higher contents (Fig. 5e). Compared with its protolith, the eclogite is depleted in LREEs (La, Ce, and Nd) and LILEs (Rb, Ba, and Sr), with LREEs and Sr decreased by about 95%, and Rb and Ba decreased by 60% and 50%, respectively (Fig. 5f). In addition, U, and heavy REEs (HREEs) do not show mass transfer. This chemical pattern of protolith composition and element mobility is consistent with the empirically estimated protolith composition and mass transfer[5]. These results suggest that the PRM can accurately reconstruct protolith compositions from metamorphic rock geochemistry.

Implications of PRM-based estimates of mass transfer

The mass transfer estimated using a PRM is an integral value between fresh basalt and an altered sample. In the case where an analyzed sample has undergone regional metamorphism, this value includes the mass transfer that occurred during seafloor alteration, prograde metamorphism, and/or retrograde metamorphism. By utilizing multi-elemental mass transfer data as well as petrological indexes such as reaction extent, these complex mass transfers can be assigned to each process; a comparison of PRM-based mass transfer with the degree of alteration or retrogression can reveal element transport at a particular stage of alteration or retrogression.

A PRM can reconstruct protolith compositions from samples in which mass

transfer has occurred and for which the protoliths are unknown. For example, in previous studies, quantitative analyses of mass transfer during metamorphism and metasomatism have usually been limited to a scale of <10 m, where protolith homogeneity can be assumed[1,3]. Provided that the distribution of data is retained within training data (i.e., mafic rocks with either a MORB, OIB, or IAB origin), the mass transfer can be estimated by a PRM for individual samples independently, and thus their spatial variation provides important information for constraining the regional-scale (i.e., >1 km) mass transfer, even if the protolith compositions are heterogeneous. The PRMs utilized in this study allow analysis of various types of sample that have undergone mass transfer (e.g., seafloor altered basalt or contact metamorphic rock) with incorporation of appropriate immobile elements. Immobile elements used for PRM inputs can be selected from 511 combinations of 9 elements according to petrological and geochemical observations.

In cases where protoliths are unknown, conventional mass transfer analyses have relied on the experience and intuition of the trained geochemist, including empirical fitting or assuming a suitably representative basalt as the protolith, such as MORB or OIB. "Anomalies" on normalized multi-elemental variation diagrams (i.e., spidergrams) are considered to show mobile elements. In contrast, the data-driven approach of the present study is applicable to investigating heterogeneities of protolith compositions and provides a less biased and more accurate estimation of metamorphic mass transfer for independent samples. Such a data-driven method is suitable for quantitative mass transfer analysis, especially in cases where protoliths are unknown and/or when there is a need to analyze mass transfer from a compiled dataset with samples from various tectonic origins.

Conclusion

In this study, we developed protolith reconstruction models (PRMs) for metabasalt, using machine-learning with large numbers of basalt compositional data. The best PRMs can estimate trace-element contents of basalt with an error of around ±0.1 in $\log_{10}$ units or ±25% incorporating only four or five input element contents. Using immobile elements as input elements, a four-element PRM was used to estimate protolith compositions of metabasalt. Application to seafloor altered basalt and eclogite verified the accuracy of protolith reconstruction within reasonable uncertainty of the estimation (0.1 in $\log_{10}$ units or 25%). This machine-learning-based method enabled an analysis of mass transfer of metabasalt with unknown protolith and can be applied to regional metamorphic belts or alteration zones where the protolith is heterogeneous.

Method

PRMs were constructed using the machine learning algorithm of the gradient boosting decision tree, specifically, the LightGBM algorithm. To improve empirical model reproducibility, hyperparameters of LightGBM were automatically tuned through Bayesian optimization by using a partial training dataset. Partial training datasets for hyperparameter tuning were prepared by K-fold cross validation, which enabled us to use all training data for constructing the PRMs. Details of the machine-learning calibrations for PRMs are described below.

Gradient boosting decision tree (LightGBM)

Gradient boosting decision tree (GBDT) is a supervised machine-learning method from which prediction models can be constructed from multidimensional data and used to solve classification and regression problems[44]. In the field of geochemistry, this machine-learning method has been applied to extract information, discriminate classes, and predict values; for example, to discriminate and extract characteristics from a volcanic rock dataset of eight different tectonic settings[24], classify metamorphic protolith(s) from the major element contents of a rock[39], and complement geochemical mapping for improvement of accuracy and interpretation [45].

Both random forests and GBDT have been proposed as explainable models with high accuracy. GBDT is an ensemble method that combines multiple decision trees to build a powerful model. In the GBDT method, decision trees are built one after another in such a way that the next decision tree corrects the errors of the previous one[40]. The development of GDBT has allowed various algorithms such as Xgboost[46] and Catboost[47] to be proposed, of which LightGBM is an algorithm with fast calculation time and high

accuracy[48]. For this reason, LightGBM was used as the machine-learning algorithm and for constructing models to predict element contents.

Tuning hyperparameters

LightGBM is a decision-tree-based nonparametric model. A nonparametric model has higher degrees of freedom compared with a linear model because of the fewer assumptions needed regarding the training data. However, the flexibility of a decision tree model makes it easier to overfit the training data. To solve this overfitting problem, each model has hyperparameters to restrict the degrees of freedom. Given that appropriate values can be assigned depending on the structure and number of dimensions of datasets, the hyperparameters need to be selected.

To choose appropriate hyperparameters, we used Bayesian optimization to tune them automatically for the dataset. Bayesian optimization is a method that uses the framework of Bayesian probability to select the next parameter to be explored based on the history of previously calculated parameters[49]. In this study, Optuna was used as the optimization software[50], with a part of the data being used as the validation for hyperparameter tuning by Bayesian optimization. The number of hyperparameter searches was set to 50. The tuned hyperparameters were as follows:

num_leaves: the maximum number of leaves in one tree;

max_depth: limit the depth for the tree model. This can be used to deal with over-fitting; and

min_data_in_leaf: the minimum number of data in one leaf.

These three parameters are specified in the official LightGBM documentation as the first to be tuned. The other parameters are set with default values.

Model construction

K-fold cross-validation

Data with no missing values in the input and output elements were extracted from the basalt compositional dataset and divided into training or test data. One-fifth of the data were used as test data to evaluate the accuracy of model, and the remaining data were used as training data to construct machine-learning models.

K-fold cross-validation is a way of evaluating the effects of tuning hyperparameters and to prevent a reduction in the number of available data (Fig. S3). The training data are randomly split into K distinct subsets. K − 1 subsets are assigned for training the model, and the other subset is used for evaluating the hyperparameters (i.e., validation data). By changing the subsets used for training and validation, the model is evaluated K times (i.e., K folds)[40]. The average RMSE obtained from all folds is used for hyperparameter tuning by Bayesian optimization. In this study, we constructed a 4-fold cross validation. The reproducibility of the model was evaluated by using the test data (which are independent from the training and validation data used for hyperparameter tuning).

Preprocessing of each set of compositional data and Bayesian optimization

To improve the estimation error, input variables are transformed to ratios and products, along with dimensional compression, with a search for the best data representation (i.e., feature engineering). Feature engineering is a common technique for constructing machine-learning models[40]. In this study, data were transformed as ratios and products of contents between two arbitrary elements, and scores of Principal

Component Analysis (PCA) and Independent Component Analysis (ICA) were calculated from log-transformed datasets for the training data of each fold. The validation and test data of each fold were also transformed using the same procedures, and their PCA and ICA scores were calculated by projecting the validation/test data onto the PCA/ICA eigenvectors of the training data. All of the measured content data, products, ratios, and PCA/ICA scores were used as preprocessed data for training and estimation of the machine learning models.

Preprocessed training data were used to construct machine-learning models, and we applied the models to preprocessed validation data to evaluate the reproducibility by RMSE (Fig. S3). Based on the averages of the obtained RMSE, Bayesian optimization software (Optuna) performed to tune the model's hyperparameters. We repeated model construction and evaluation 50 times to find the appropriate hyperparameters for each set of compositional data.

References

1. Taetz, S., John, T., Bröcker, M. & Spandler, C. Fluid–rock interaction and evolution of a high-pressure/low-temperature vein system in eclogite from New Caledonia: insights into intraslab fluid flow processes. *Contrib. to Mineral. Petrol.* **171**, (2016).
2. Taetz, S., John, T., Bröcker, M., Spandler, C. & Stracke, A. Fast intraslab fluid-flow events linked to pulses of high pore fluid pressure at the subducted plate interface. *Earth Planet. Sci. Lett.* **482**, 33–43 (2018).
3. Beinlich, A., Klemd, R., John, T. & Gao, J. Trace-element mobilization during Ca-metasomatism along a major fluid conduit: Eclogitization of blueschist as a consequence of fluid-rock interaction. *Geochim. Cosmochim. Acta* **74**, 1892–1922 (2010).
4. Bebout, G. E. Metamorphic chemical geodynamics of subduction zones. *Earth Planet. Sci. Lett.* **260**, 373–393 (2007).
5. John, T., Scherer, E. E., Haase, K. & Schenk, V. Trace element fractionation during fluid-induced eclogitization in a subducting slab: Trace element and Lu-Hf-Sm-Nd isotope systematics. *Earth Planet. Sci. Lett.* **227**, 441–456 (2004).
6. Pearce, J. A., Stern, R. J., Bloomer, S. H. & Fryer, P. Geochemical mapping of the Mariana arc-basin system: Implications for the nature and distribution of subduction components. *Geochemistry, Geophys. Geosystems* **6**, (2005).
7. C. J. Hawkesworth;, Gallagher;, K., Hergt;, J. M. & McDermott;, F. P. Mantle and Slab Contributions in ARC Magmas. *Earth* 175–204 (1993).
8. Ferry, J. M. Regional metamorphism of the waits river formation, Eastern Vermont: Delineation of a new type of giant metamorphic hydrothermal system. *J. Petrol.* **33**, 45–94 (1992).
9. Ague, J. J. Mass transfer during Barrovian metamorphism of pelites, south-central Connecticut. I: evidence for changes in composition and volume. *Am. J. Sci.* **294**, 989–1057 (1994).
10. Masters, R. L. & Ague, J. J. Regional-scale fluid flow and element mobility in Barrow's metamorphic zones, Stonehaven, Scotland. *Contrib. to Mineral. Petrol.* **150**, 1–18 (2005).
11. Penniston-Dorland, S. C. & Ferry, J. M. Development of spatial variations in reaction progress during regional metamorphism of micaceous carbonate rocks, northern New England. *Am. J. Sci.* **306**, 475–524 (2006).
12. John, T. *et al.* Volcanic arcs fed by rapid pulsed fluid flow through subducting slabs. *Nat. Geosci.* **5**, 489–492 (2012).
13. Mindaleva, D. *et al.* Rapid fluid infiltration and permeability enhancement during

middle–lower crustal fracturing: Evidence from amphibolite–granulite-facies fluid–rock reaction zones, Sør Rondane Mountains, East Antarctica. *Lithos* **372–373**, 105521 (2020).

14. Ishikawa, T. *et al.* Coseismic fluid-rock interactions at high temperatures in the Chelungpu fault. *Nat. Geosci.* **1**, 679–683 (2008).
15. Tanikawa, W. *et al.* Transport properties and dynamic processes in a fault zone from samples recovered from TCDP Hole B of the Taiwan Chelungpu Fault Drilling Project. *Geochemistry, Geophys. Geosystems* **10**, 1–18 (2009).
16. Kitajima, K., Maruyama, S., Utsunomiya, S. & Liou, J. G. Seafloor hydrothermal alteration at an Archaean mid-ocean ridge. *J. Metamorph. Geol.* **19**, 583–599 (2001).
17. Shibuya, T. *et al.* Decrease of seawater CO2 concentration in the Late Archean: An implication from 2.6Ga seafloor hydrothermal alteration. *Precambrian Res.* **236**, 59–64 (2013).
18. Grant, J. A. The isocon diagram-a simple solution to Gresens' equation for metasomatic alteration. *Econ. Geol.* **81**, 1976–1982 (1986).
19. Grant, J. A. Isocon analysis: A brief review of the method and applications. *Phys. Chem. Earth* **30**, 997–1004 (2005).
20. Kuwatani, T. *et al.* Sparse isocon analysis: A data-driven approach for material transfer estimation. *Chem. Geol.* **532**, 119345 (2020).
21. Moss, B. E., Haskin, L. A., Dymek, R. F. & Shaw, D. M. Redetermination and reevaluation of compositional variations in metamorphosed sediments of the Littleton Formation, New Hampshire. *Am. J. Sci.* **295**, 988–1019 (1995).
22. Moss, B. E., Haskin, L. A. & Dymek, R. F. Compositional variations in metamorphosed sediments of the Littleton Formation, New Hampshire, and the Carrabassett Formation, Maine, at sub-hand specimen, outcrop, and regional scales. *Am. J. Sci.* **296**, 473–505 (1996).
23. Plank, T. & Langmuir, C. H. The chemical composition of subducting sediment and its consequences for the crust and mantle. *Chem. Geol.* **145**, 325–394 (1998).
24. Ueki, K., Hino, H. & Kuwatani, T. Geochemical discrimination and characteristics of magmatic tectonic settings: A machine-learning-based approach. *Geochemistry, Geophys. Geosystems* **19**, 1327–1347 (2018).
25. Uno, M. *et al.* Elemental transport upon hydration of basic schists during regional metamorphism: Geochemical evidence from the Sanbagawa metamorphic belt, Japan. *Geochem. J.* **48**, 29–49 (2014).
26. Spandler, C., Hermann, J., Arculus, R. & Mavrogenes, J. Geochemical heterogeneity and element mobility in deeply subducted oceanic crust; insights from high-pressure mafic rocks from New Caledonia. *Chem. Geol.* **206**, 21–42

(2004).

27. Aoki, K. *et al.* U-Pb zircon dating of the Sanbagawa metamorphic rocks in the Besshi-Asemi-gawa region, central Shikoku, Japan, and tectono-stratigraphic consequences. *J. Geol. Soc. Japan* **125**, 183–194 (2019).
28. Cluzel, D., Aitchison, J. C. & Picard, C. Tectonic accretion and underplating mafic terranes in the late Eocene intraoceanic fore-arc of New Caledonia (Southwest Pacific): Geodynamic implications. *Tectonophysics* **340**, 23–59 (2001).
29. Alt, J. C. *et al.* Hydrothermal Alteration of a Section of Upper Oceanic Crust in the Eastern Equatorial Pacific: A Synthesis of Results from Site 504 (DSDP Legs 69, 70, and 83, and ODP Legs 111, 137, 140, and 148). *Proc. Ocean Drill. Program, 148 Sci. Results* **148**, (1996).
30. Staudigel, H., Plank, T., White, B. & Schmincke, H. U. Geochemical fluxes during seafloor alteration of the basaltic upper oceanic crust: DSDP sites 417 and 418. in *Geophysical Monograph Series* vol. 96 19–38 (1996).
31. Kelley, K. A., Plank, T., Ludden, J. & Staudigel, H. Composition of altered oceanic crust at ODP Sites 801 and 1149. *Geochemistry, Geophys. Geosystems* **4**, (2003).
32. Kessel, R., Schmidt, M. W., Ulmer, P. & Pettke, T. Trace element signature of subduction-zone fluids, melts and supercritical liquids at 120-180 km depth. *Nature* **437**, 724–727 (2005).
33. Tsay, A., Zajacz, Z. & Sanchez-valle, C. Efficient mobilization and fractionation of rare-earth elements by aqueous fluids upon slab dehydration. *Earth Planet. Sci. Lett.* **398**, 101–112 (2014).
34. Tsay, A., Zajacz, Z., Ulmer, P. & Sanchez-Valle, C. Mobility of major and trace elements in the eclogite-fluid system and element fluxes upon slab dehydration. *Geochim. Cosmochim. Acta* **198**, 70–91 (2017).
35. Ague, J. J. Element mobility during regional metamorphism in crustal and subduction zone environments with a focus on the rare earth elements (REE). *Am. Mineral.* **102**, 1796–1821 (2017).
36. Ague, J. J. Extreme channelization of fluid and the problem of element mobility during Barrovian metamorphism. *Am. Mineral.* **96**, 333–352 (2011).
37. Pearce, J. A. & Cann, J. R. Tectonic setting of basic volcanic rocks determined using trace element analyses. *Earth Planet. Sci. Lett.* **19**, 290–300 (1973).
38. Hollocher, K., Robinson, P., Walsh, E. & Roberts, D. Geochemistry of amphibolite-facies volcanics and gabbros of the støren nappe in extensions west and southwest of Trondheim, Western Gneiss Region, Norway: A key to correlations and paleotectonic settings. *Am. J. Sci.* **312**, 357–416 (2012).

39. Hasterok, D., Gard, M., Bishop, C. M. B. & Kelsey, D. Chemical identification of metamorphic protoliths using machine learning methods. *Comput. Geosci.* **132**, 56–68 (2019).
40. Géron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow_ Concepts, Tools, and Techniques to Build Intelligent Systems*. (O'Reilly Media, Inc., 2011).
41. Sun, S. S. & McDonough, W. F. Chemical and isotopic systematics of oceanic basalts: Implications for mantle composition and processes. *Geol. Soc. Spec. Publ.* **42**, 313–345 (1989).
42. Alt, J. C. & Teagle, D. A. H. The uptake of carbon during alteration of ocean crust. *Geochim. Cosmochim. Acta* **63**, 1527–1535 (1999).
43. John, T. & Schenk, V. Partial eclogitisation of gabbroic rocks in a late Precambrian subduction zone (Zambia): Prograde metamorphism triggered by fluid infiltration. *Contrib. to Mineral. Petrol.* **146**, 174–191 (2003).
44. Loh, W. Y. Classification and regression trees. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1**, 14–23 (2011).
45. Kirkwood, C., Cave, M., Beamish, D., Grebby, S. & Ferreira, A. A machine learning approach to geochemical mapping. *J. Geochemical Explor.* **167**, 49–61 (2016).
46. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* **13-17-Augu**, 785–794 (2016).
47. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018-Decem**, 6638–6648 (2018).
48. Ke, G. *et al.* LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017-Decem**, 3147–3155 (2017).
49. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst.* **4**, 2951–2959 (2012).
50. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. *arXiv* 2623–2631 (2019).
51. Daisuke, K., Ryuji, S., Keisuke, H. & Yuji, H. *Data analysis techniques to win Kaggle*. doi:4297108437.

Acknowledgments

This study was financially supported in part by JSPS KAKENHI grants 18K13628 awarded to M.U. S.M., M.U., and A.O. were partly funded by Joint Usage/Research Center programs (ERI JURP) 2015-B-04, 2018-B-01, and 2021-B-01 of the Earthquake Research Institute, University of Tokyo, Japan. We thank the members of the ERI JURP for constructive discussions.

Author contributions

S.M. designed and coded machine-learning algorithms. M.U. designed the research strategy. A.O. and N.T. critically discussed the research strategy and outcomes. All of the authors discussed the results and commented on the manuscript. All authors read and approved the final manuscript.

Competing of interests

The authors declare that they have no competing interests.

Additional Information

Corresponding author

Masaoki UNO

Graduate School of Environmental Studies, Tohoku University, Sendai 980-8579, Japan

TEL: +81-22-795-6336

FAX: +81-22-795-6336

E-mail: uno@geo.kankyo.tohoku.ac.jp

Figures

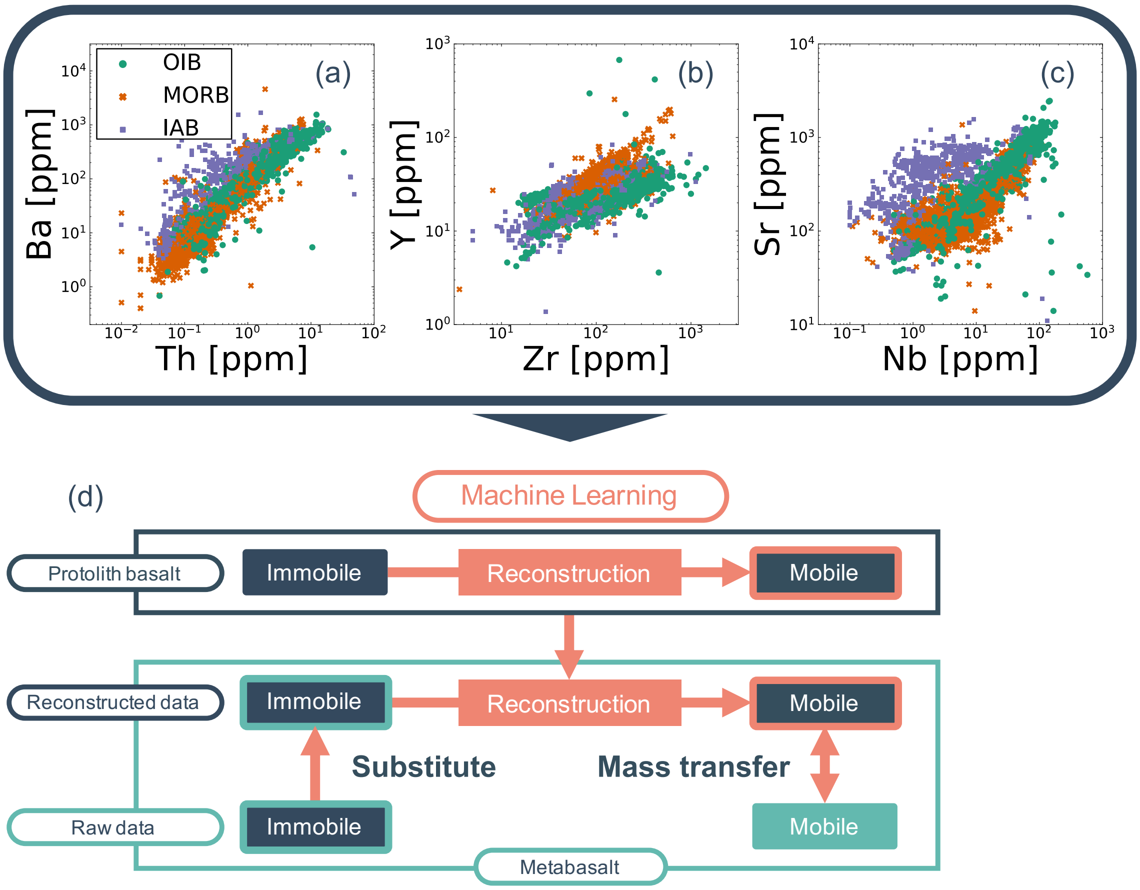

Figure 1. Distribution of the compositional dataset used in this study (compiled from the geochemical database at https://search.earthchem.org/). (**a**) Th and Ba, (**b**) Zr and Y, and (**c**) Nb and Sr. (**d**) Schematic overview of protolith reconstruction models (PRMs). Empirical models were calibrated by the protolith (basalt) compositional dataset and applied to metabasalt compositions. Assuming that the contents of immobile elements in metabasalt are identical to those in protolith basalt, these contents can be assigned as inputs and used to obtain protolith compositions.

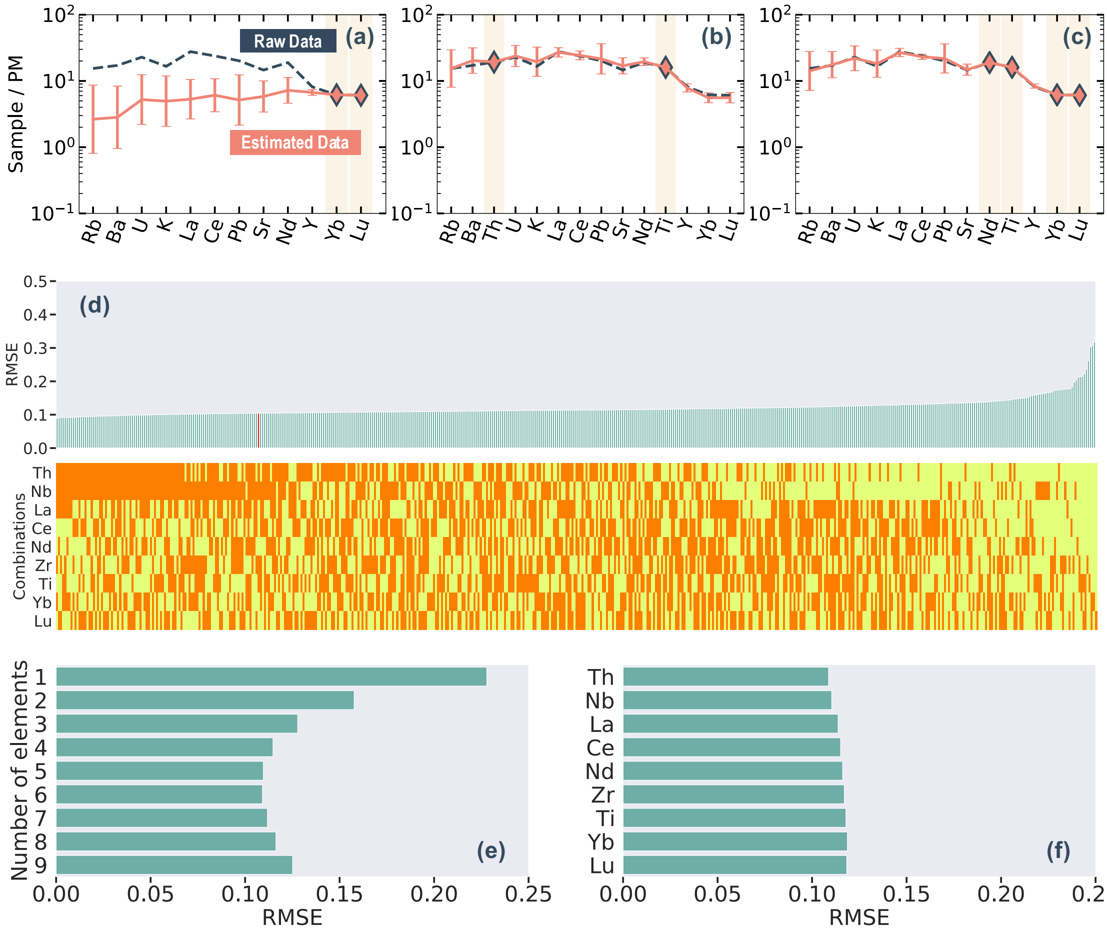

Figure 2. (**a–c**) Estimated primitive-mantle-normalized contents of basalt. Pink diamonds indicate the input contents. Predicted data were obtained from the input contents of (**a**) Yb and Lu; (**b**) Th and Ti; and (**c**) Nd, Ti, Yb, and Lu. Raw basalt compositional data are shown as a dashed dark-blue line, and estimated basalt compositional data are shown as a pink line. Compositions of the primitive mantle are from Sun and McDonough (1989). (**d**) Average RMSE scores of all output elements for each combination of input elements (511 cases), and combinations of input elements for each model shown in below. In the upper plot, the red line indicates the input combination of Th, Nb, Zr, and Ti. In the lower plot, the orange elements are used in combinations, and yellow elements are not used. (**e**) Average RMSE scores for all of the models using a particular number of input elements. (**f**) Average of all of the models containing a particular element as an input.

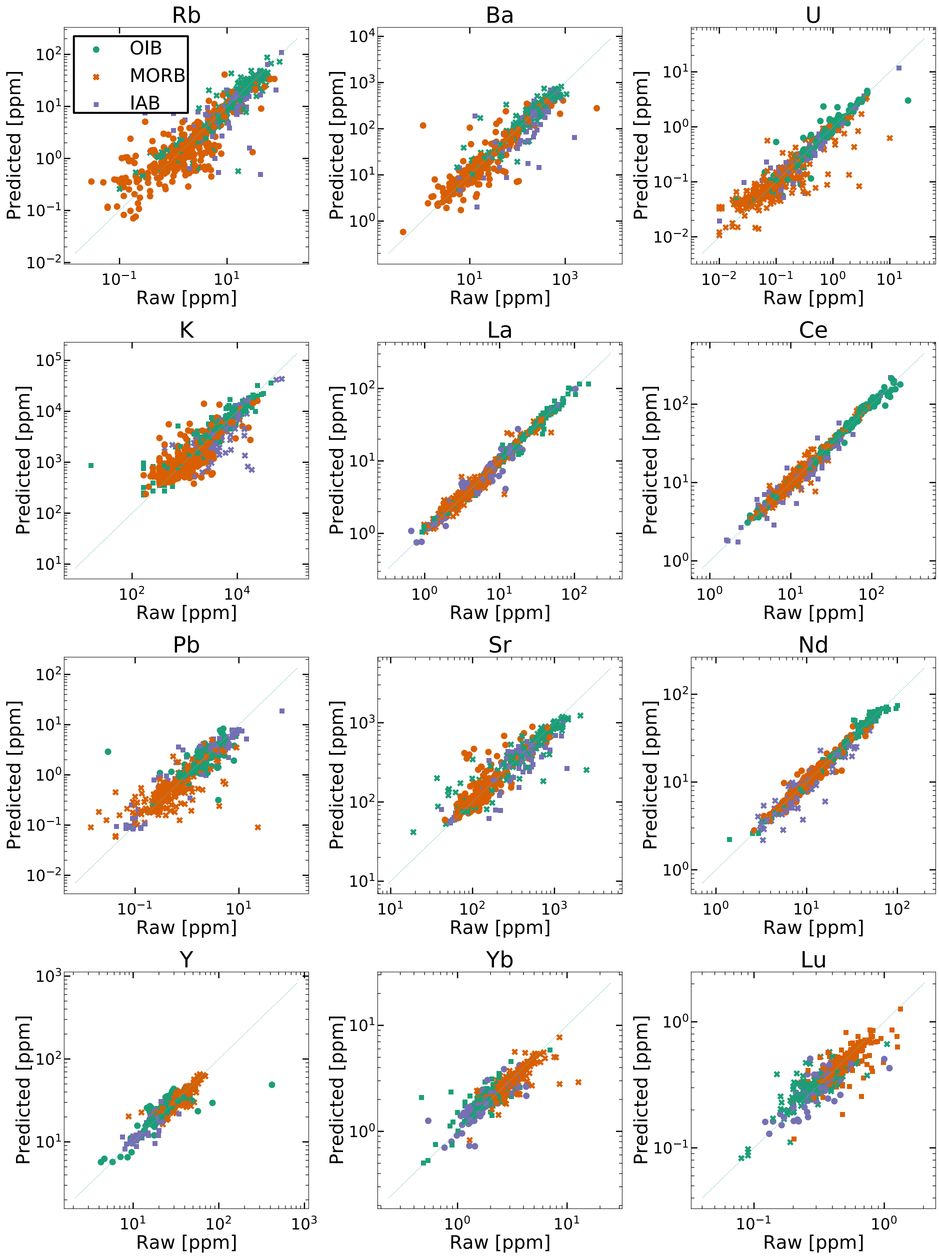

Figure 3. Scatter plots of predicted contents versus raw (measured) contents with the final PRM using Th, Nb, Zr, and Ti as input elements. The PRM was applied to test data of the basalt dataset, which covers three different tectonic settings (mid-ocean ridge basalt, ocean island basalt, and island arc basalt).

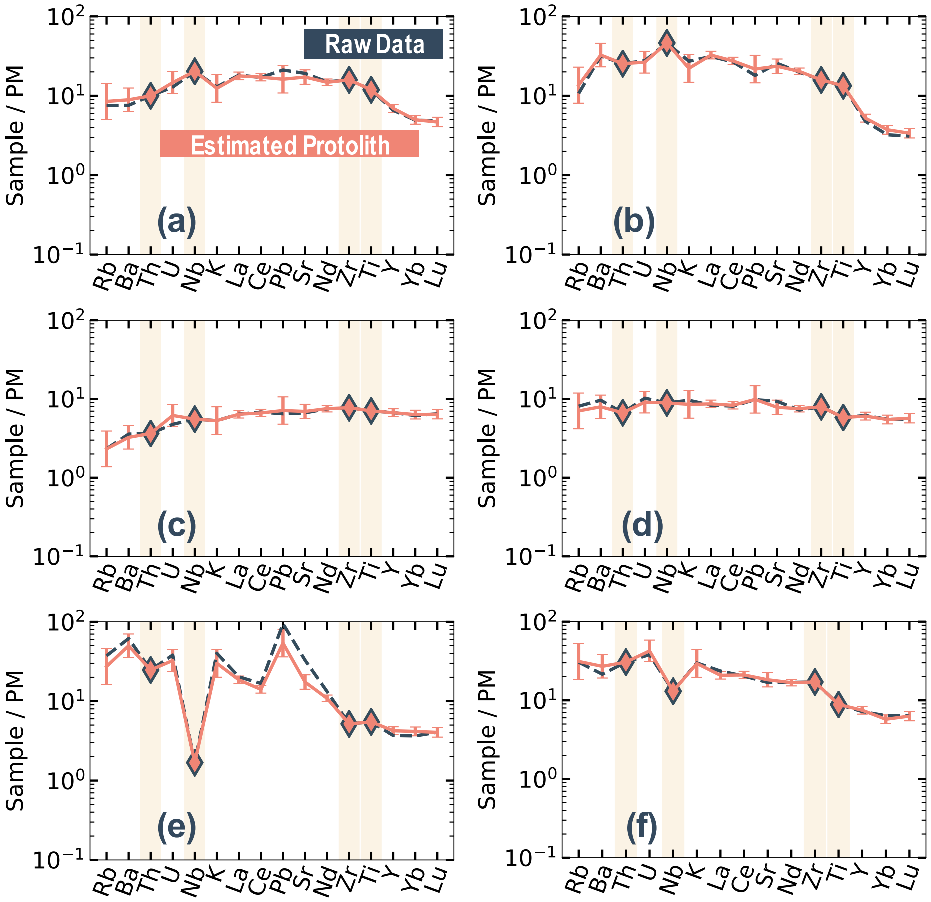

Figure 4. Primitive-mantle-normalized contents of estimated basalt using the four-element PRM with Th, Nb, Zr, and Ti as input elements. Samples for each plot are examples from (**a, b**) OIB, (**c, d**) MORB, and (**e, f**) IAB. Diamonds indicate input data. Raw basalt compositional data are shown as a dashed dark-blue line, and estimated basalt compositional data are shown as a pink line. Compositions of the primitive mantle are from Sun and McDonough (1989).

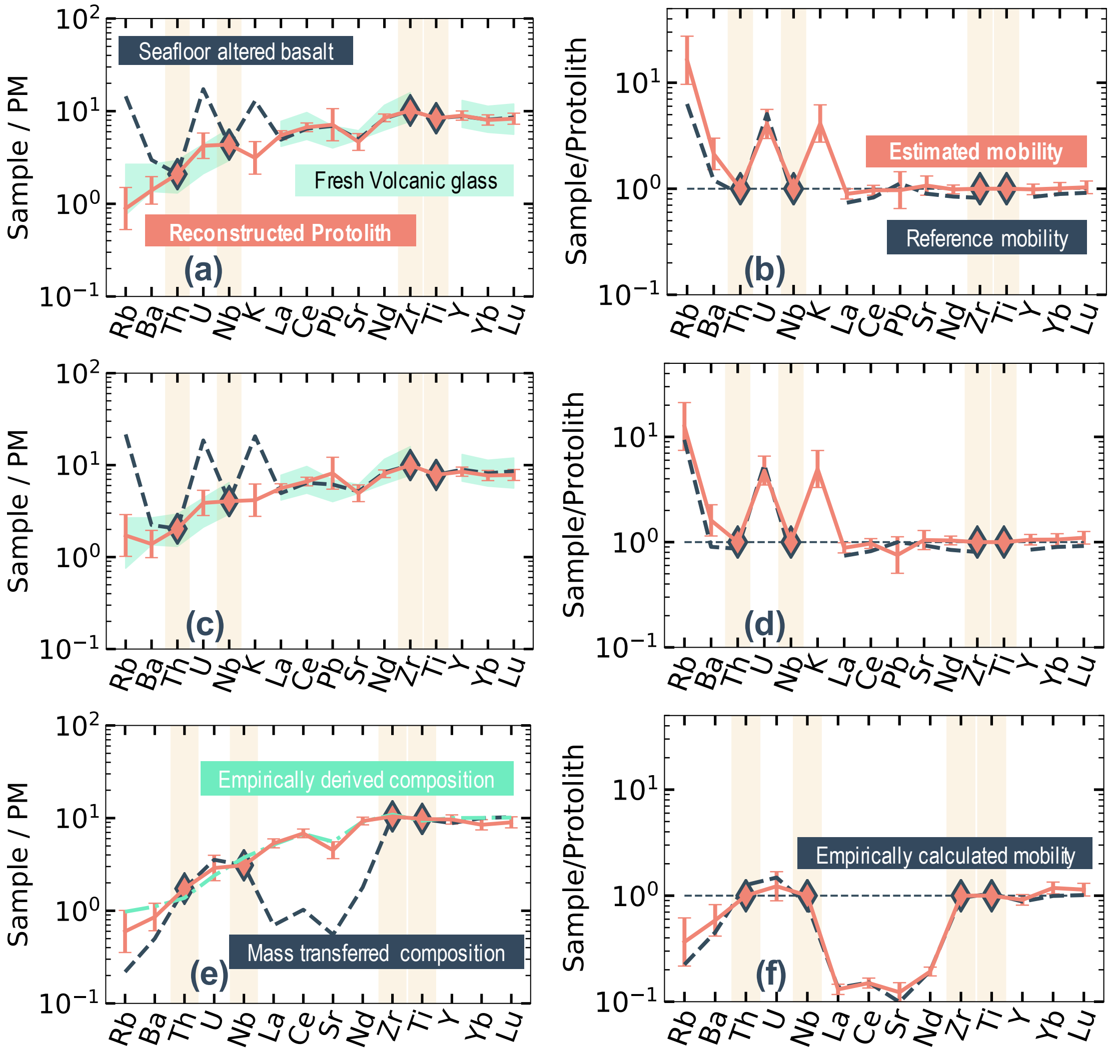

Figure 5. Results of the selected four-element PRM applied to seafloor altered basalt and metabasalt, and calculated mass mobility. Samples for each plot are (**a, b**) 801-MORB-110-222_ALL[31], (**c, d**) 801_SUPER, and (**e, f**) Z139-6[5]. (**a, c**) Primitive-mantle-normalized contents of estimated protolith basalt using the PRM. Diamonds indicate input data (Th, Nb, Zr, and Ti). Seafloor altered and metamorphic rock contents are shown as a dashed dark-blue line, and estimated protolith basalt contents are shown as a pink line. The range in protolith contents derived from fresh glass is shown as a sky-blue region. (**b, d**) Calculated mass mobility using fresh glass composition (dashed dark-blue line)

and estimated protolith (pink line). (**e**) Primitive-mantle-normalized contents of estimated protolith basalt. Protolith compositions empirically derived[5] are shown as a sky-blue line. (**f**) Calculated mass mobility using empirically derived composition (dark-blue line) and estimated protolith (pink line).

Figures

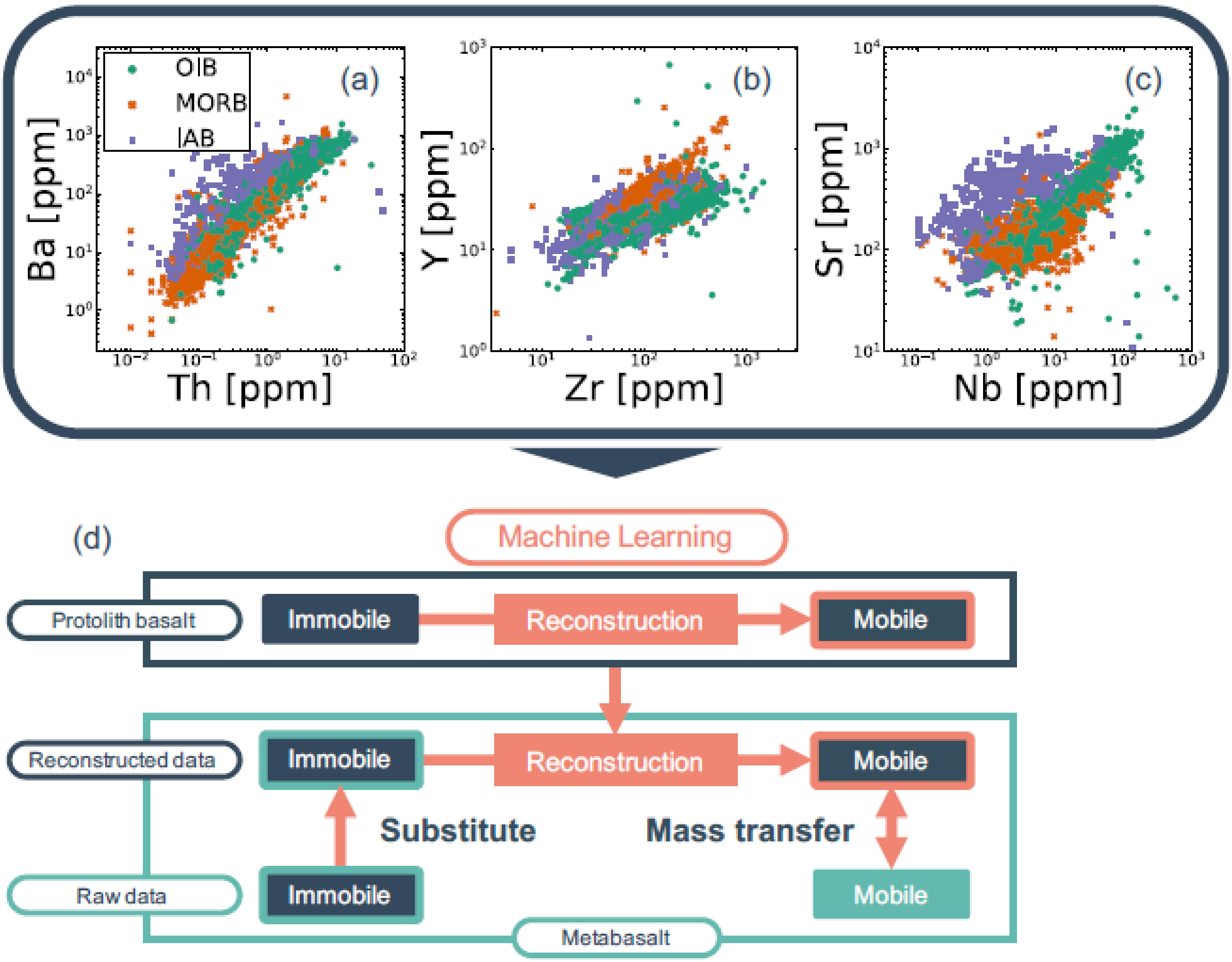

Figure 1

Distribution of the compositional dataset used in this study (compiled from the geochemical database at https://search.earthchem.org/). (a) Th and Ba, (b) Zr and Y, and (c) Nb and Sr. (d) Schematic overview of protolith reconstruction models (PRMs). Empirical models were calibrated by the protolith (basalt) compositional dataset and applied to metabasalt compositions. Assuming that the contents of immobile elements in metabasalt are identical to those in protolith basalt, these contents can be assigned as inputs and used to obtain protolith compositions.

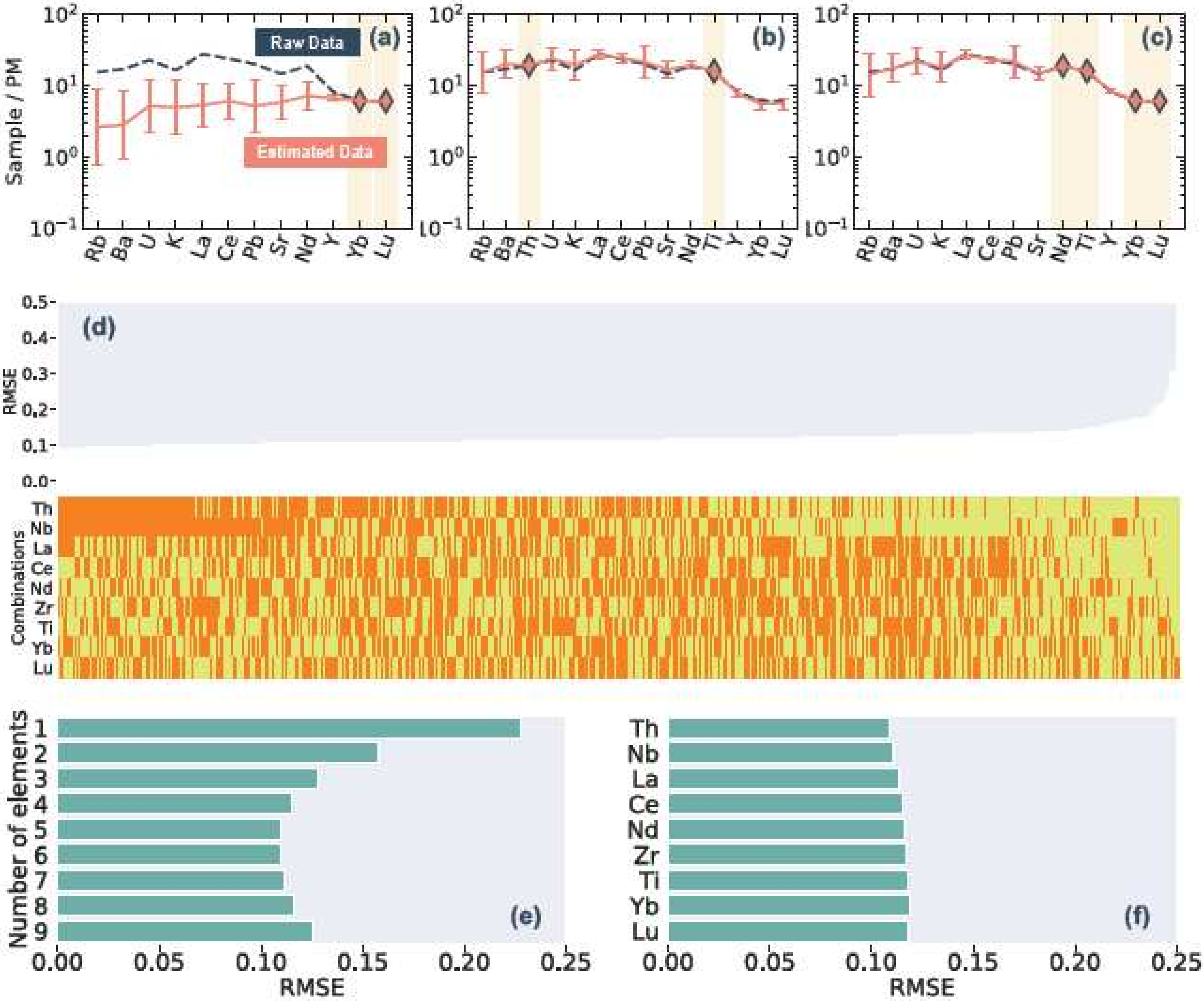

Figure 2

(a–c) Estimated primitive-mantle-normalized contents of basalt. Pink diamonds indicate the input contents. Predicted data were obtained from the input contents of (a) Yb and Lu; (b) Th and Ti; and (c) Nd, Ti, Yb, and Lu. Raw basalt compositional data are shown as a dashed dark-blue line, and estimated basalt compositional data are shown as a pink line. Compositions of the primitive mantle are from Sun and McDonough (1989). (d) Average RMSE scores of all output elements for each combination of input elements (511 cases), and combinations of input elements for each model shown in below. In the upper plot, the red line indicates the input combination of Th, Nb, Zr, and Ti. In the lower plot, the orange elements are used in combinations, and yellow elements are not used. (e) Average RMSE scores for all of the models using a particular number of input elements. (f) Average of all of the models containing a particular element as an input.

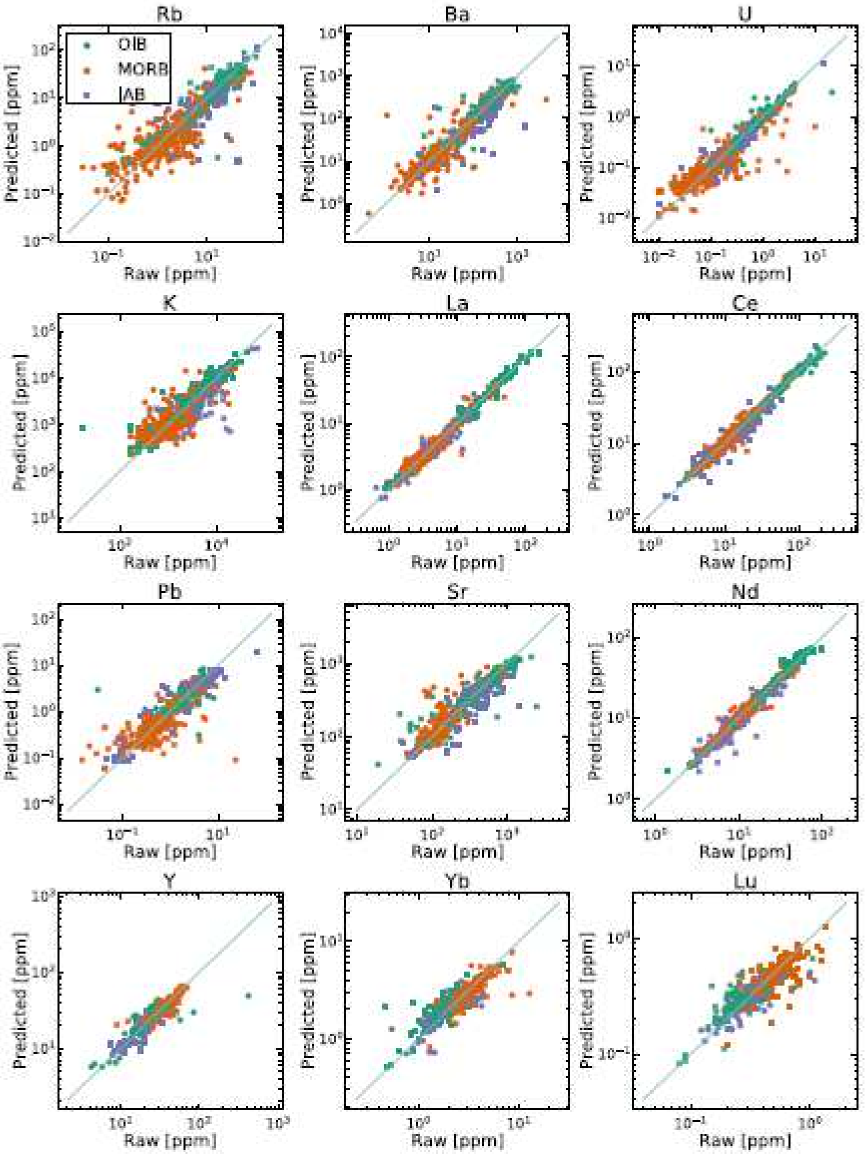

Figure 3

Scatter plots of predicted contents versus raw (measured) contents with the final PRM using Th, Nb, Zr, and Ti as input elements. The PRM was applied to test data of the basalt dataset, which covers three different tectonic settings (mid-ocean ridge basalt, ocean island basalt, and island arc basalt).

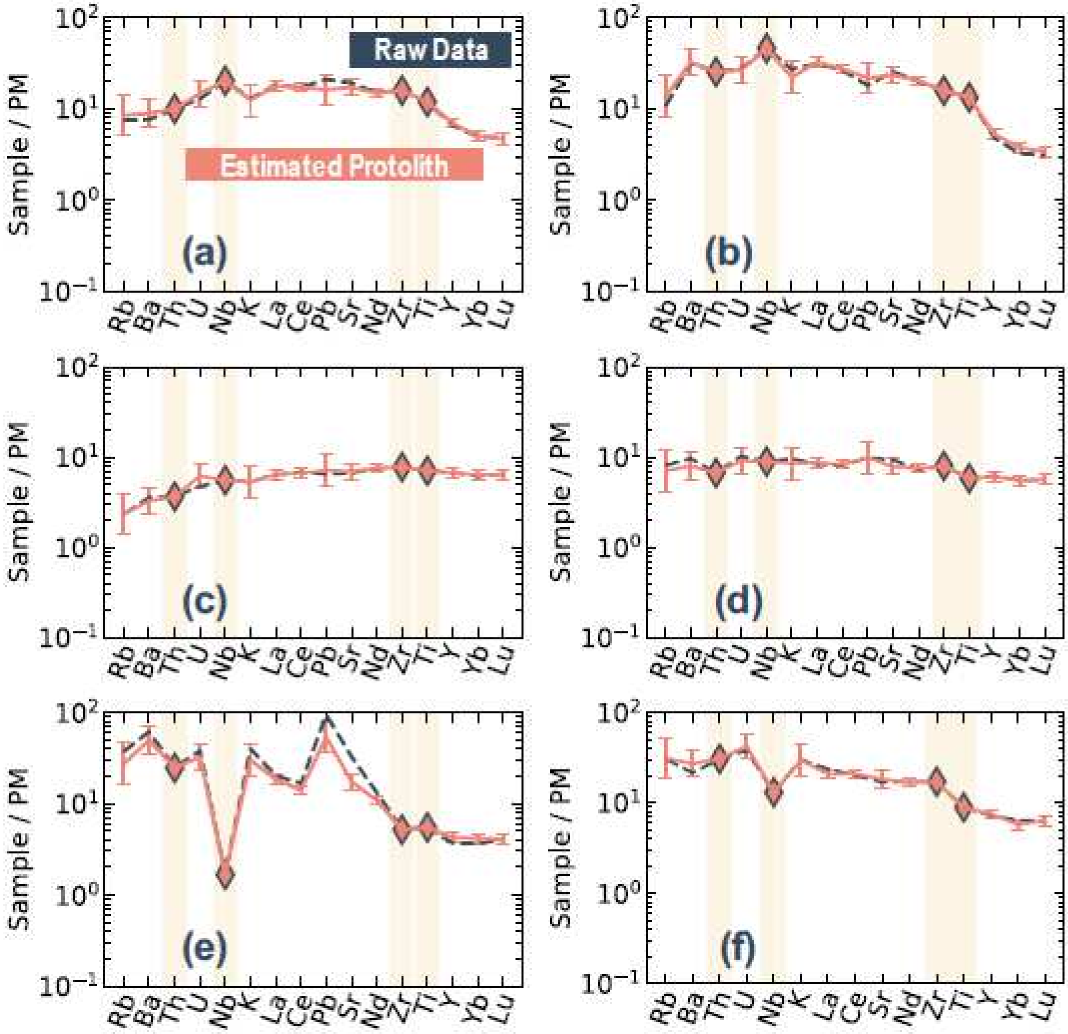

Figure 4

Primitive-mantle-normalized contents of estimated basalt using the four element PRM with Th, Nb, Zr, and Ti as input elements. Samples for each plot are examples from (a, b) OIB, (c, d) MORB, and (e, f) IAB. Diamonds indicate input data. Raw basalt compositional data are shown as a dashed dark-blue line, and estimated basalt compositional data are shown as a pink line. Compositions of the primitive mantle are from Sun and McDonough (1989).

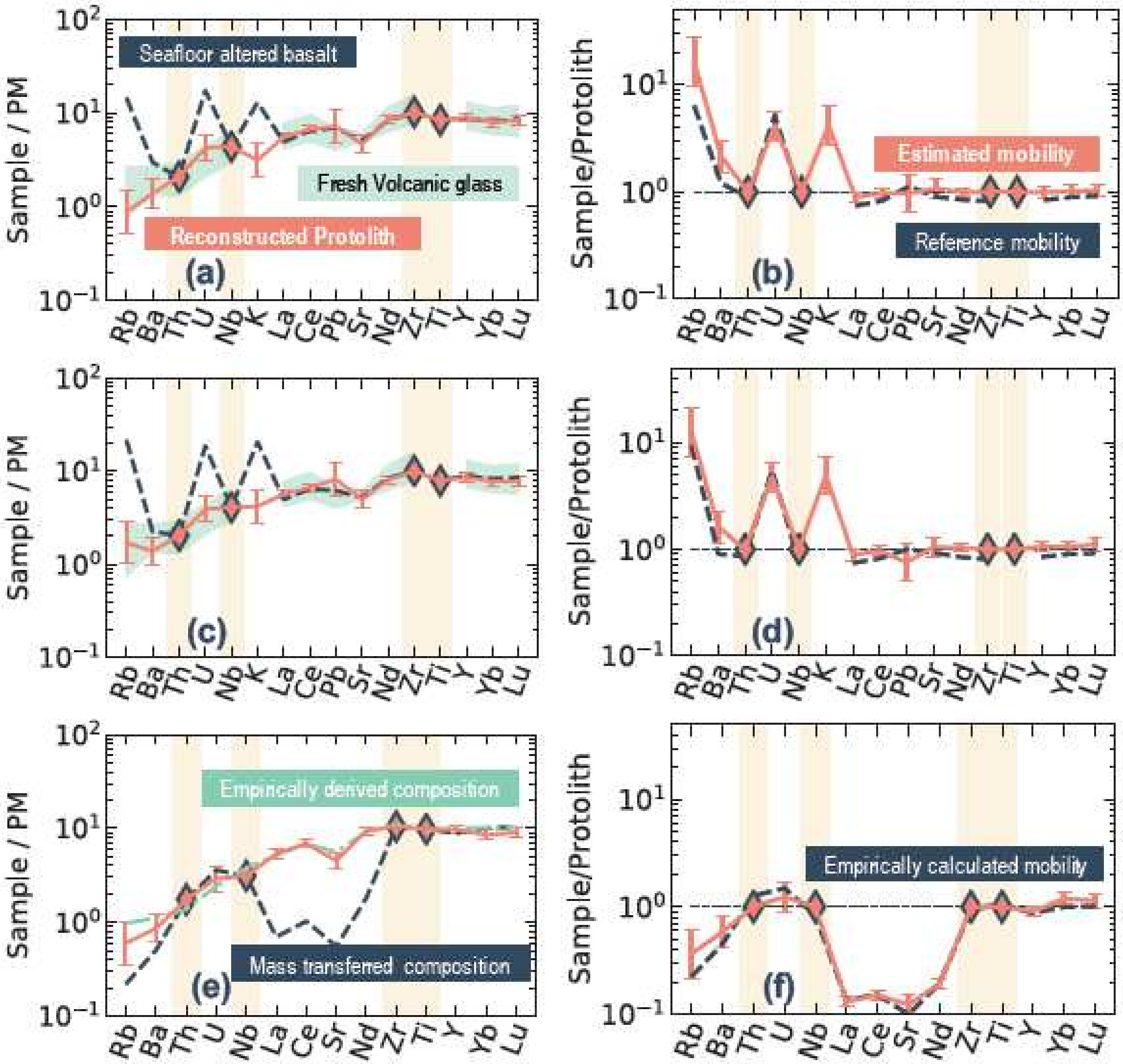

Figure 5

Results of the selected four-element PRM applied to seafloor altered basalt and metabasalt, and calculated mass mobility. Samples for each plot are (a, b) 801-MORB- 110-222_ALL31, (c, d) 801_SUPER, and (e, f) Z139-65. (a, c) Primitive-mantle normalized contents of estimated protolith basalt using the PRM. Diamonds indicateinput data (Th, Nb, Zr, and Ti). Seafloor altered and metamorphic rock contents are shown as a dashed dark-blue line, and estimated protolith basalt contents are shown as a pink line. The range in protolith contents derived from fresh glass is shown as a sky-blue region. (b, d) Calculated mass mobility using fresh glass composition (dashed dark-blue line) and estimated protolith (pink line). (e) Primitive-mantle-normalized contents of estimated protolith basalt. Protolith compositions empirically derived5 are shown as a sky-blue line. (f) Calculated mass mobility using empirically derived composition (dark-blue line) and estimated protolith (pink line).

Supplementary Files