

D.Hasterok, M.Garda, C.M.B.Bishop, D.Kelsey

Chemical identification of metamorphic protoliths using machine learning methods

Computers and Geosciences, 2019; 132:56-68

© 2019 Elsevier Ltd. All rights reserved.

This manuscript version is made available under the CC-BY-NC-ND 4.0 license

<http://creativecommons.org/licenses/by-nc-nd/4.0/>

Final publication at: <http://dx.doi.org/10.1016/j.cageo.2019.07.004>

PERMISSIONS

<https://www.elsevier.com/about/policies/sharing>

Accepted Manuscript

Authors can share their [accepted manuscript](#):

24 Month Embargo

After the embargo period

- via non-commercial hosting platforms such as their institutional repository
- via commercial sites with which Elsevier has an agreement

In all cases [accepted manuscripts](#) should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license – this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our [hosting policy](#)
- not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article

17 November 2021

Chemical identification of metamorphic protoliths using machine learning methods

D. Hasterok^{a,b,*}, M. Gard^a, C.M.B. Bishop^{a,c}, D. Kelsey^{a,d}

^a*Department of Earth Sciences, University of Adelaide, North Terrace, SA, 5005, Australia*

^b*Mawson Centre for Geoscience, University of Adelaide, North Terrace, SA, 5005, Australia*

^c*now at Research School of Earth Sciences, Australian National University, Building 142 Mills Rd, Acton, ACT, 2601, Australia*

^d*now at Geological Survey of Western Australia, WA, Australia*

Abstract

The fundamental origins of metamorphic rocks as sedimentary or igneous are integral to the proper interpretation of a terrane's tectonic and geodynamic evolution. In some cases, the protolith class cannot be determined from field relationships, texture, and/or compositional layering. In this study, we utilize machine learning to predict a metamorphic protolith from its major element chemistry so that accurate interpretation of the geology may proceed when the origin is uncertain or to improve confidence in field predictions. We survey the efficacy of several machine learning techniques to predict the protolith class (igneous or sedimentary) for whole rock geochemical analyses using 9 major oxides. The data are drawn from a global geochemical database with >533 000 geochemical analyses. In addition to metamorphic samples, igneous and sedimentary analyses are used to supplement the dataset based on their similar chemical distributions to their metamorphic counterparts. We train the classifiers on most of the data, retaining ~10% for post-training validation. We find that the RUSBoost algorithm performs best overall, achieving a true-positive rate of >95% and >85% for igneous- and sedimentary-derived samples, respectively. Even the traditionally-difficult-to-differentiate metasedimentary and metaigneous rocks of granitic-granodioritic composition were consistently identified with a >75% success rate (92% for granite; 85% for granodiorite; 88% for wacke; 76% for arkose). The least correctly identified rock types were iron-rich shale (58%) and quartzolitic rocks (6%). These trained classifiers are able to classify metamorphic protoliths better than common discrimination methods, allowing for the appropriate interpretation of the chemical, physical, and

*Corresponding author

Email addresses: dhasterok@gmail.com (D. Hasterok, matthew.gard@adelaide.edu.au (M. Gard), david.kelsey@adelaide.edu.au (D. Kelsey)

¹DH conducted classification tests, added data to the database and wrote the manuscript. MG compiled the geochemical database. CB conducted classification tests. DK provided guidance on existing chemical discrimination methods. All authors helped with manuscript preparation and revision.

²)

tectonic contextual history of a rock. The preferred classifier is available as a MATLAB function that can be applied to a spreadsheet of geochemical analyses, returning a predicted class and estimated confidence score. We anticipate this classifier's use as a cheap tool to aid geoscientists in accurate protolith prediction and to increase the size of global geochemical datasets where protolith information is ambiguous or not retained.

Keywords: data processing, machine learning, protolith discrimination, igneous geochemistry, sedimentary geochemistry

1. Introduction

Accurately identifying a protolith is crucial to unravelling the geologic evolution of terranes, allowing one to understand past tectonic and geodynamic environments. Differentiation between igneous and sedimentary protoliths is often determined based on field relationships, mineral grain habits, and/or evidence of inherited relict structures such as bedding (Bucher and Grapes, 2011). Another common method of protolith discrimination, particularly for felsic gneisses, is examination of the zircon date spectrum. Sedimentary protoliths typically have more complex date spectra due to the integration of multiple sources of differing dates. Igneous protoliths tend to have uni- or bimodal zircon date spectra indicating the timing of crystallization and a record of ensuing metamorphic event. However, recrystallization, severe deformation, and/or partial melting can mask the diagnostic indicators of an original protolith. It is also common for geochemical databases to exclude sufficient descriptions of geological samples that readily indicate the protolith. For instance, it is common for many rocks to be identified simply as gneiss or schist within geochemical databases (Hasterok et al., 2018), a textural description that is ambiguous with regard to the protolith type.

In some cases, the protolith class can be reasonably inferred through the use of chemical scatter plots. Several chemical classification methods have been devised (e.g., Moine and De La Roche, 1968; Irvine and Baragar, 1971; Tarney, 1977; Roser and Korsch, 1988), but overlapping chemical ranges between igneous and sedimentary fields add uncertainty that results in misclassification (e.g., Lindsey, 1999). At present, no single method developed to identify a metamorphic protolith is optimized for the global geochemical dataset (Gard et al., 2019). The use of multiple discrimination diagrams can reduce uncertainty where the chemical ranges overlap, but the potential combinations of such diagrams are so numerous that it is difficult to

24 produce an optimal scheme by trial and error.

25 Machine learning methods are well-suited to developing predictive models from multidimen-
26 sional datasets and have been used effectively in geologic settings. Carranza and Laborte (2015)
27 used Random Forests to investigate epithermal gold deposits in the Philippines. In addition
28 to Random Forests, Rodriguez-Galiano et al. (2015) used neural networks, regression trees and
29 support vector machines to identify areas of mineral prospectivity. Machine learning has also
30 been used to predict lithology from soil geochemistry or chemically altered samples (Kirkwood
31 et al., 2016; Hood et al., 2018). Cracknell and Reading (2014) used a number of these methods
32 to investigate their potential to develop geologic maps based on remotely sensed geophysical
33 data.

34 Here we evaluate the accuracy of several machine learning methods for predicting a meta-
35 morphic protolith as either igneous or sedimentary on the basis of major element composition.
36 Because we wish to identify a method that has broad applicability, we focus only on major el-
37 ements as part of this study. We utilize a large whole-rock global geochemical dataset to train
38 and validate these classification methods. We also explore the preprocessing steps, including
39 log-ratio transforms and principal component analysis, to yield the best predictive capability.
40 Finally, we detail the success of the model in distinguishing protoliths among a variety of rock
41 types.

42 **2. Existing Chemical Discriminants**

43 Several methods to predict a protolith class from geochemistry have been employed but a
44 comprehensive comparison of methods has not been made. The discrimination method em-
45 ployed by studies is chosen to suit the chemistry of the protoliths. As only methods which
46 perform well are highlighted for publication, it is unknown what additional tests were at-
47 tempted but disregarded because of poor performance. Consequently, there is little guidance
48 in the literature for a best performing set of chemical discrimination tools that accurately
49 identifies protoliths in a majority of cases. Below we highlight a few of the more general dis-
50 crimination methods. In this study, we refer to a protolith as broad term that describes igneous
51 or sedimentary samples, either metamorphosed or unmetamorphosed. In Figure 1 we demon-
52 strate three discrimination methods, displaying the igneous and sedimentary distributions in
53 two-dimensional histograms.

Several studies employ TiO_2 as a metamorphic protolith discriminator in combination with other elements (Misra, 1971; Tarney, 1977; Winchester et al., 1980; Werner, 1987). For instance, Tarney (1977) suggests igneous and sedimentary rocks can be discriminated by a single division in SiO_2 – TiO_2 space. Tarney’s calibration works poorly on the global dataset (Figure 1a and b), resulting in 52% igneous and 55% sedimentary true positive rates. From a global geochemical dataset (Section 3), we suggest that samples with > 2.5 wt.% TiO_2 are more likely igneous, which identifies 13% of igneous samples as probably igneous and 2% of sedimentary samples as probably igneous. However, this simple test does not allow prediction of the protolith for igneous samples < 2.5 wt.% TiO_2 or any samples as sedimentary. Taking cells within a 2-dimensional histogram that contain a single class will identify 8.8% of igneous samples and 1% of sedimentary samples (Figure 1a and b). Hence, the range of both classes overlap such that it is very difficult to clearly identify either class definitively.

Ternary systems are a popular way of examining three rather than two dimensions, potentially separating samples better than a simple Cartesian plot. Commonly used ternary systems A–C–F, A–CN–K, and MgO – CaO – FeO_T are used to suggest a protolith for a suite of rocks (Misra, 1971; Winkler, 1979; Best, 1982; Ehlers and Blatt, 1982). For example, rocks with negative A values on a A–C–F diagram are typically igneous (Figure 1c and d). However, sedimentary samples do extend into the negative field.

More complex combinations and/or ratios of elements are also used to predict protoliths such as Niggli indices (Winkler, 1979) and discriminant function analysis (Roser and Korsch, 1988). The igneous field, as identified by Simonen (1953), performs considerably better than the TiO_2 discriminate above, identifying 60% true positive igneous and 88% true positive sedimentary. (Figure 1e and f). Given the distribution of protolith chemistries within the global dataset, Simonen’s igneous field may not be optimal. While the region occupied by igneous rocks is more concentrated, it overlaps most of the sedimentary class range as only 12% of cells contain one class. As a result, any growth in the size of the igneous field to increase the accuracy of igneous identification will reduce the sedimentary accuracy because of the overlapping chemical ranges.

3. Geochemical Dataset

3.1. Global geochemical database

We use a global dataset of whole rock major element data from a combination of online databases, government reports, and ~ 2000 journal papers. A full description of the database and sources is given by Gard et al. (2019), which is updated from Hasterok and Webb (2017) and Hasterok et al. (2018). The data extracted for use in this study as described below, can be found archived at Zenodo.org (Hasterok et al., 2019). The full database contains over one million samples and is derived from EarthChem.org linked databases, governmental reports and data releases, and the academic literature (Gard et al., 2019). About 17% of the samples within the database are sedimentary and $\sim 9\%$ are metamorphic. Nearly half the metamorphic samples include sufficient descriptions that igneous and sedimentary protoliths may be identified.

In order to ensure consistent treatment of the data we normalize 9 major elements (SiO_2 , TiO_2 , Al_2O_3 , FeO_T , MgO , CaO , Na_2O , K_2O , P_2O_5) to 100%, creating an Aitchison simplex geometry (Aitchison, 1986). Only data which contain all the required major elements above detection limits are used. The remaining dataset for analysis contains 533 360 samples, with 497 401 igneous and 35 959 sedimentary samples. Below detection limit (BDL) values can be used to improve classifier accuracies in some cases (Templ et al., 2016). However, we choose not to include BDL data because the detection limits vary by orders of magnitude with respect to K_2O and P_2O_5 depending on the study and method of analysis. Many studies simply report BDL, but do not report the detection limit, thus limiting the utility of these for classification. Excluding BDL values could potentially bias the results, but the compositional spaces occupied by such samples are likely filled by others given the size of the remaining dataset.

It is recommended that log-ratio rescaling to a Euclidean geometry from an Aitchison simplex will improve machine learning performance. Two transformations are commonly employed, the centered log-ratio (clr) and isometric log-ratio (ilr) transformations (Egozcue et al., 2003). The clr transformation is of equivalent dimensionality to the original simplex whereas the ilr transformation reduces the dimensionality by one component. Because compositional data sum is normalized to 1 (sum to 100%), there is one less degree of freedom than the number of compositional variables. The ilr transformation removes this redundancy to create a set of compositional vectors that form an independent basis (Egozcue et al., 2003).

3.2. Protolith chemistry

3.2.1. Metamorphosed versus unmetamorphosed

To discriminate metaigneous and metasedimentary samples on the basis of chemistry an observable difference must exist. However, the set of metamorphic samples in the global database is comparatively small relative to the number of igneous and sedimentary samples. Before supplementing the metamorphic dataset, it must be shown that the chemistry of protoliths are negligibly changed by metamorphic processes (Figure 2). The chemical variability of nine major oxides bear similarities between the metamorphic samples and those not indicated as metamorphic. There are no clear trends in chemistry between igneous and metaigneous or sedimentary and metasedimentary that suggests a significant chemical alteration process in response to metamorphism. Hence the majority of metamorphic systems may be chemically closed or at least isochemical (i.e., no change in major element cations) when partial melting has not resulted in extraction of melt.

There are some peaks observed in the igneous and metamorphic rocks that are not visible or as prominent in the metamorphic data (Figure 2). For example, igneous carbonatites with high CaO, low Al_2O_3 and low SiO_2 are not visible among the metaigneous samples. There is a similar concentration of sedimentary and metasedimentary samples coincident with this high CaO igneous peak associated with marbles. It is possible that many metamorphic carbonatites are indistinguishable from marbles and are misclassified in the database (Le Bas et al., 2002). Among the igneous and metaigneous data there is a set of bimodal peaks centered at approximately 50 and 75 wt.% SiO_2 (Figure 2). These are from sampling bias associated with mafic volcanics and felsic plutonics as noted by (Hasterok and Webb, 2017).

Metamorphic samples have lower variance, but this does not prevent the datasets from being meaningfully combined (Figure 2). First, many samples labeled as igneous or sedimentary in the global database experienced some degree of metamorphism—especially true for Precambrian samples where few unmetamorphosed rocks exist. The decision to report metamorphism is generally related to the questions probed by a particular study (Hasterok et al., 2018), e.g., metamorphic descriptions are often excluded from studies of igneous and sedimentary petrogenesis. In many studies, metamorphic facies and textures are described in the text, but the tables may only provide a protolith’s igneous or sedimentary name. Some databases (e.g., GEOROC),

record data from tables but not the text losing this information. We make this claim based on comparisons between descriptions of samples from papers and samples contained within the global dataset. Therefore, the chemistry of igneous and sedimentary samples in the global database are not independent of their metamorphic counterparts.

Second, we assume that the narrower chemistry of metamorphic samples within the database results from biased sampling. It is possible that a majority of metamorphic rocks in the global database are selected for their mineralogy that is useful for assessing metamorphic conditions as opposed to characterizing the natural variability in chemistry (M. Hand, pers. comm.). We see this phenomenon in rocks described as marbles, which are mostly pure calcite in outcrop but are more likely marls based on the compositions contained within the global database. Studies of igneous and sedimentary petrogenesis may not be so discriminating in their selection of compositions and therefore display a larger range. Hence, we assume the igneous and sedimentary samples can be combined with the metaigneous and metasedimentary samples to develop a protolith classification scheme.

3.2.2. Basic analysis of chemical differences

The chemical ranges of igneous and sedimentary protoliths largely overlap making it difficult when only 2 to 3 compositional parameters are used to determine the protolith from bulk chemistry alone. But there are some differences in chemistry between the sets, possibly allowing for a relatively definitive determination of some samples (Figure 2). For example, sedimentary rocks rarely have $\text{TiO}_2 > 2.5 \text{ wt.}\%$ or $\text{MgO} > 30 \text{ wt.}\%$. This difference allows us to identify the likely origins of these chemical characteristics as igneous (Figure 2). However, the overlaps outside these chemical ranges are considerable making it difficult to predict the protolith for most of the data from these simple Harker diagrams alone.

The Mahalanobis distance, D^2 , provides another metric for differences between multivariate data relative to a centroid (Maesschalck et al., 2000). We compute the Mahalanobis distance for the ilr transformed data (Figure 3), although the Aitchison and clr transformed data yield similar results. The distribution of D^2 for the igneous and metaigneous data are very similar as are the sedimentary and metasedimentary data (Figure 3a and b). In each case, the distances rapidly decrease in frequency from the centroid, with a small fraction $<10\%$ of samples, extending beyond D^2 of 20. However, the cumulative distribution of D^2 for the meta+sedimentary

172 data extend significantly farther beyond the meta+igneous centroid suggestive of chemical dif-
173 ferences between the two subsets (Figure 3c). For example, a D^2 of 20 from the meta+igneous
174 centroid contains nearly 50% of the meta+sedimentary data while capturing $\sim 90\%$ of the
175 meta+igneous data.

176 Principal component analysis (PCA) tends to highlight similarities rather than differences
177 (Abdi and Williams, 2010), but it can also be used to prefilter outlying data prior to applying
178 machine learning techniques. PCA of the Aitchison compositions is dominated by the largest
179 chemical components but may not reflect the underlying chemical processes that lead to the
180 variations in rock chemistry (Figure 4a). Both clr and ilr transformed data return relatively
181 similar principle component diagrams. Although the ilr transform reduces the dimensionality,
182 it is apparent that many of the ilr basis vector components are similar to the clr results (Fig-
183 ure 4b and c), indicating that the elements associated with the greatest variance are largely
184 independent despite the restrictions on dimensionality.

185 While the ilr transform eliminates the dependent dimension, the clr transformed data are
186 more straightforward to interpret. The clr transformed PCA result indicates the greatest vari-
187 ance in igneous rocks ($\sim 60\%$, Figure 5) is explained predominantly by K_2O and MgO and the
188 largest coefficient for the second principle component is P_2O_5 . The first two principal compo-
189 nents account for $\sim 80\%$ of the total variance. In contrast, the first two principal components
190 of sedimentary rocks account for ~ 44 and 17% of the total variance. CaO has the largest
191 coefficient for the first principal component and Na_2O is the largest for the second.

192 Distinguishing between igneous and sedimentary-derived protoliths for the majority of sam-
193 ples is not possible from PCA analysis. The sedimentary and igneous scores display considerable
194 overlap for the majority of their respective distributions, but there are significant differences in
195 the data density where one class may be more likely (Figure 6). Machine learning techniques
196 can exploit these variations to produce an optimal classification method.

197 4. Methods

198 4.1. Machine Learning Techniques

199 There are a number of machine learning classification schemes that have been developed
200 (Kotsiantis, 2007). In this study, we focus on using several common approaches that are

201 included in the MATLAB[®] Classification App (MATLAB[®], 2018): discriminant analysis,
 202 logistic-regression analysis, support vector machines (SVM), nearest neighbor classifiers (KNN),
 203 and decision trees (for reviews of the various methods see Kotsiantis, 2007; Crisci et al., 2012;
 204 Praveena and Jaiganesh, 2017). In this study, we test the effectiveness of each of these methods
 205 to develop an accurate protolith classifier. Below we discuss a few selected models that are
 206 singled out for additional study due to their performance. We are using MATLAB version
 207 9.4.0.813654 (R2018a) to perform our analysis, which gives us the ability to test these methods
 208 under a simple common framework. However, one is not limited to MATLAB as these methods
 209 are also available within the R and Python programming languages.

210 For all of the training methods, we use 5-fold cross-validation to select hyperparameters.
 211 The k -fold cross-validation, $k = 5$ in our study, is an option for MATLAB machine learning
 212 algorithms that randomly splits the data into k subsets. Each group as a holdout to score the
 213 performance of a model trained on the remaining subsets. This process is repeated for each
 214 subset and the scores of each test are summarized. This cross-validation processes reduces
 215 overfitting in the classifier model. We performed some early trials with larger k values (7 and
 216 10), but we found a negligible change in performance while significantly increasing the training
 217 time, so a k of 5 was deemed acceptable.

218 Although we tested a number of machine learning algorithms mentioned above, only a select
 219 few, KNN and ensemble decision trees, were chosen for more in-depth analysis based on their
 220 performance. In the interests of space, we limit our summary to these methods in greater depth
 221 below.

222 4.1.1. *K-Nearest Neighbor*

223 The KNN classifier is perhaps the simplest of the methods tested. The KNN algorithms
 224 produce a classifier by collecting a subset of data near a point within the compositional space.
 225 The score for each class is determined by the number of data for said class near the investigation
 226 point. The winning class is assigned by the highest score. The KNN methods may be improved
 227 by changing the number of points included in the subset and by weighting the samples contri-
 228 bution to the score by some distance metric (e.g., inverse square, Gaussian). MATLAB includes
 229 an ensemble option for the KNN method which allows for a set of models to be produced by
 230 using a subspace, with randomly selected combinations of a reduced set of predictor variables.

231 4.1.2. *Decision trees*

232 Decision trees produce a sequence of binary tests (branches) that split the dataset until a
233 branch terminates in a leaf that contains a single class (Breiman et al., 1984). The number
234 of branches is limited to reduce the likelihood of over-fitting, at which point the leaves are
235 determined by majority vote from the distribution of classes it contains. The branch tests
236 are randomly generated and chosen based on the test that results in the best discrimination
237 between the leaves and nucleating the next branch. A single decision tree is produced from the
238 process which results in an optimal tree that discriminates the most classes properly.

239 4.1.3. *Ensemble Trees*

240 While the single decision tree methods search for an optimal tree, it may not be optimal for
241 certain compositional subsets. For example, a tree that works well for silicate-dominated rocks
242 may not perform well on carbonates. In this case, an ensemble of individually less accurate
243 trees are combined to produce an overall more accurate result (Breiman, 1996). There are
244 several methods for developing ensemble decision trees. We focus on testing three algorithms:
245 Bagged, AdaBoost, and RUSBoost trees. There is no one algorithm that performs best for all
246 applications, so it is important to test multiple ensemble methods.

247 Bagged trees is a bootstrapping method that develops several trees using several random
248 subsets of the data (Breiman, 1996). The collection of random trees is then used to produce
249 a set of predicted classes for each sample. The final predicted class is then determined by a
250 simple majority vote for each sample.

251 AdaBoost generates data weights following production of each classification tree and pro-
252 duces a final classifier based on a weighted average of the individual classifiers. MATLAB uses
253 the AdaBoost algorithm by Freund and Shapire (1996).

254 The RUSBoost algorithm is a modified boosting method, similar to AdaBoost, that includes
255 random sampling of the training dataset, similar to bagging (Seiffert et al., 2010). The advan-
256 tage of the RUSBoost algorithm is improved performance when the training dataset is highly
257 skewed towards a single class, which is beneficial as the global dataset contains >90% igneous
258 samples.

4.2. Preparation of training and post-training validation datasets

We test several approaches that prepare the dataset for classification to identify which procedure results in the best metamorphic protolith classifier. The approaches follow three separate choices: (1) 10% reserve or equal sampling; (2) prefiltering using PCA; and (3) using the Aitchison data or transforming using clr or ilr. As a result, we test 9 separate input datasets into the classification algorithms. Note the PCA analysis is not performed on the equal sampling datasets.

For the first choice, we reserve a portion of the dataset for testing the trained classifier to independently validate its accuracy. We use two methods to select a training and post-training validation dataset. Hereafter, the post-training validation data are simply referred to as the validation data. In one case, we select an equal number of igneous and sedimentary data for training by randomly selecting 10% from the total number of sedimentary samples for the validation dataset. The remaining 90% are used to train the classifier. We then select an equal number of igneous samples for the training dataset and reserve the remainder for validation. The second method of sample selection is made by randomly selecting 10% of the total dataset for validation and using the remaining 90% for the training dataset. In the latter case, the percentages selected for training vary somewhat between the igneous and sedimentary datasets but are roughly proportional to the total dataset.

For the three datasets created with PCA prefiltering, we select 95% of the data with the lowest Hotelling's t^2 -statistic. In each of the three cases we use the ilr transformed data before computing the PCA and t^2 results. This choice ensures consistent treatment of the data, although the choice of transform has little effect on the samples excluded by PCA filtering. The PCA filtering is only applied to the larger training datasets.

5. Results

To identify an accurate classification method, we conduct 332 tests from each of the 9 training datasets. A full list of methods, parameters, and performance is provided in the Supplemental Material.

5.1. PCA filtering and log-ratio transforms

A few gross observations are apparent from the classification results with different preprocessing methods. First, PCA prefiltering results in less accurate classification in most tests (Figure 7a). Because our dataset is quite large, PCA filtering may not help because the tails of our distributions are well-sampled. PCA filtering on the dataset cuts these tails off, thereby restricting the compositional range and increasing misclassification in the tail regions of compositional space and resulting in less accurate classifiers. For example, notice the contraction in range that occurs with PCA filtering (i.e., change from the 99 to 95 percentile contours in Figure 6). Second, we find that transforming the dataset does systematically improve the classifier performance (Figure 7b). The differences in true-positive rate are typically within 2% and varies whether Aitchison or log-ratio transforms are best. Therefore, we see no particular advantage to transforming the data in order to predict a protolith class.

5.2. Comparison of classification methods

When the subsets of igneous and sedimentary samples are equal, most of the classification methods perform well, correctly identifying >80% of sample protoliths. Accuracy of sedimentary protoliths identification is typically 10% lower than for igneous protoliths (Supplementary Table). There are two methods that consistently have true-positive rates >90% regardless of the data geometry: weighted KNN and the bagged trees ensemble. The Gaussian SVM and KNN ensemble also achieve >90% true-positive rates for the Aitchison geometry. These methods typically have a <4% difference between the accurate identification of igneous and sedimentary protoliths.

Many of these methods perform well when each class size is roughly equal but perform poorly when a single class dominates the dataset (Figure 8). Because of our disparity in subset size, a classifier that identifies all samples as igneous obtains an overall true positive rate of 93.2% and is therefore, a relatively meaningless measure of performance. We prefer to examine the true-positive rate of classification of the igneous and sedimentary classes separately or as the average of each class’s accuracies, which we refer to as the normalized true-positive rate,

$$\text{normalized true-positive rate} = \frac{1}{2} \left(\frac{TP_{\text{ig.}}}{T_{\text{ig.}}} + \frac{TP_{\text{sed.}}}{T_{\text{sed.}}} \right), \quad (1)$$

where TP is the number of true positives and T is the number of true values in the igneous (ig.) and sedimentary (sed.) classes, respectively.

Generally, the true-positive rate of igneous protolith identification is $>98\%$ whereas sedimentary protolith identification is lower by typically $>30\%$ (Figure 8). One significant exception is the RUSBoost ensemble method, which produces more equitable performance between classes when a large disparity in class sizes exists. While RUSBoost has the lowest igneous true-positive rate in the initial test, it has the largest normalized true-positive rate of all the methods (Figure 8). The SVM, discriminant and logistic regression methods perform very poorly and will no longer be considered. Single decision trees also perform poorly among sedimentary protolith identification. The KNN and ensemble methods perform relatively well, achieving igneous true-positive rates $>98\%$ and sedimentary true-positive rates $>50\%$.

5.3. Refined ensemble classifiers

The ensemble methods are improved by increasing the number of learners, and for decision trees, branches as the number of branches (20) and learners (30) are relatively low in the initial tests. All further tests are conducted using a starting dataset with an Aitchison geometry without PCA filtering and utilizing 90% of the original dataset with 10% held for independent validation.

5.3.1. Ensemble KNN method

We test a few additional ensemble KNN classifiers, changing the number of subspace dimensions or number of learners. An increase in learners results in a negligible improvement in accuracy (Supplementary Table). An increase in the number of subspace dimensions does improve the true-positive rate of sedimentary protoliths from 39.3% with 3 subspace dimensions to 68.7% with 7 subspace dimensions. The igneous protolith accuracy does not change much since the true-positive rate is $>98\%$ for all ensemble KNN classifiers. The number of subspace dimensions is limited by the number of data dimensions. None of the ensemble KNN unweighted classifiers perform as well as the single weighted KNN method. While the KNN method performed well on the training dataset, the true-positive rate was significantly lower on the validation dataset. Therefore, we do not consider the method as reliable as the methods described below for protolith determination.

5.3.2. Ensemble decision trees

Bagging results are mildly better ($\sim 1\%$) than AdaBoost for igneous protoliths. Bagging is the poorest method among the ensemble classifiers for correctly sorting sedimentary protoliths. The RUSBoost model performs best overall ($>86\%$ normalized true-positive rate), but worst among ensemble decision tree methods for igneous protolith classification—note that worst still identifies $\sim 95\%$ true positive igneous samples.

All of the ensemble decision trees improve with the number of branches. AdaBoost and Bagging are relatively unaffected by additional learners whereas RUSBoost experiences a drop in normalized and sedimentary true-positive rates with additional learners and an increase in igneous true-positive rate (Figure 9). The RUSBoost classifiers improve to about 1000 branches.

We choose our preferred RUSBoost model to have 30 learners and 1000 branches because it represents the parameters for which the method performance plateaus and results in the highest normalized true-positive rate of the methods tested. The results of the RUSBoost classifier are nearly identical on both the training and validation dataset, both in gross performance (Table 1 and Figure 10) and the performance on individual rock types (Figure 11). We generated two random testing and validation datasets for the suite of RUSBoost classifiers and find the performance to be very similar for both testing and validation datasets (Figure 9). Each identifies $\sim 95\%$ of true igneous and $>85\%$ of true sedimentary protoliths correctly.

6. Discussion

6.1. RUSBoost performance

Since the RUSBoost classifier performs better than typical discrimination methods (Section 2), we suggest machine learning provides an advantage over conventional methods.

6.1.1. Sample scores

Up to this point, we have examined the performance of classifiers using a set of metrics that provide little insight into the reliability of individual predictions. A classifier can also provide a score for each individual sample that indicates the certainty in the predicted class. MATLAB returns a RUSBoost score, $f(\xi)$, for a new sample, ξ , determined by

$$f(\xi) = \sum_{t=1}^T \alpha_t h_t(x), \quad (2)$$

where

$$\alpha_t = \frac{1}{2} \log \frac{1 - \varepsilon_t}{\varepsilon_t} \quad (3)$$

and the weighted classification error, ε_t is given by

$$\varepsilon_t = \sum_{i=1}^N d_i^{(t)} I(y_i \neq h_t(x_i)). \quad (4)$$

The values x_i indicate a point within the training dataset, y_i , is the true class, h_t is the prediction hypothesis of index t , I is the indicator function, and $d_i^{(t)}$ is the weight of observation i at step t . The theoretical scores can range from $(-\infty, \infty)$ with one score for each unique class. Since we only have two classes, the scores mirror in value.

To simplify the interpretation, we use a single score that is normalized such that $[-1, 0]$ indicates the assigned class is sedimentary and $(0, 1]$ indicates the assigned class is igneous (Figure 10). For both true classes, the majority of samples have relatively high scores $> |0.5|$, indicating a high confidence in most predictions. The scores for each class have long tails that extend into the misclassified values, some of which predict the incorrect class with high confidence. As result it is difficult to remove all the misclassified samples by placing a threshold on the scores and one is likely to remove more correct than incorrect class determinations by doing so.

Scores on both the training and validation datasets yield similar distributions (Figure 10), providing further confidence that the classifier will work on an independent dataset with unknown protolith classes.

6.1.2. Performance by rock type

The performance of the preferred classifier on the dataset as a whole is different than the performance on individual rock types (Table 2). Therefore, it is necessary to evaluate the performance as a function of rock types in order to properly assess the confidence in the classifier for a specific field site.

Figure 11 and Table 2 identify the true rock types for the misclassified samples. To determine rock types, we employ several common chemical classification systems for igneous (Middlemost, 1994; Le Bas and Streckeisen, 1991) and sedimentary rocks (Mason, 1952; Turekian, 1969; Herron, 1988). These systems are slightly modified to provide additional divisions among

383 the largest compositional fields (Hasterok et al., 2018). The foidolite field is divided into
384 an ultramafic, mafic and intermediate field to allow finer compositional resolution. Mantle
385 peridotite and pyroxenite fields are added to capture samples with <42 wt.% SiO_2 and MgO
386 >18 wt.% (Reverdatto et al., 2008). Because we often do not know if metaigneous samples are
387 metavolcanic or metaplutonic we choose to group the compositions regardless of the igneous
388 emplacement process.

389 We use two metrics to assess the performance by rock type: the raw number of misclassified
390 samples and the percentage misclassified for a given rock type (e.g., percentage of granites
391 misclassified as sedimentary to all granites). The former metric provides an indication of the
392 types of rocks that will make up the bulk of the misclassified samples for large chemically diverse
393 datasets. Although a score can be determined from the classifier, the latter metric provides an
394 indication of how likely a set of samples of a given rock type may be misclassified.

395 In general, the percentages of individual rock types misidentified as igneous or sedimentary
396 are relatively low ($<5\%$). The performance of the RUSBoost classifier is very similar on both
397 the training and validation datasets (Figure 11), an indicator of the reliability of the classifier.

398 Igneous protoliths of granitic and granodioritic composition are the most commonly misiden-
399 tified samples as sedimentary absolute number and percentage of the individual rock types (Ta-
400 ble 2 and Figure 11a and c). The classifier scores for granitic and granodioritic igneous rocks
401 generally have high confidence (Table 2 and Figure 12a and b). Because these compositions are
402 in great abundance, the absolute number of misclassifications appear larger than when viewed
403 as a percentage of the number of these specific rock types. The percentage of samples misiden-
404 tified using the RUSBoost algorithm with respect to their overall abundance is $<8\%$ of granitic
405 and $<15\%$ of granodioritic rocks. Quartzolitic rocks are the poorest classified igneous rocks
406 with $\sim 94\%$ misclassified. This result is unsurprising since quartzolitic samples contain nearly
407 pure quartz, often occurring as vein quartz with little difference in major element chemistry to
408 quartzite. The classifier also has difficulty with carbonatite and silicocarbonatites, which can
409 be difficult to distinguish from marbles (i.e., limestone and dolomite) (Figure 12; Le Bas et al.,
410 2002).

411 It is unsurprising the sedimentary determination is less accurate as some sedimentary rocks
412 are basically disaggregated igneous rocks with little chemical alteration (e.g., volcanoclastics

413 and arkose). The greatest number of misclassified true sedimentary protoliths are among iron-
414 rich shale, arkose, shale and wacke (Table 2 and Figure 11b and d). These sedimentary com-
415 positions are chemically similar to intermediate to felsic igneous rocks, explaining why granitic
416 and granodioritic compositions are the most likely to be incorrectly predicted.

417 Figure 12c to f shows the scores for common sedimentary rock types. Aside from iron-rich
418 shales, the remaining rock types have significant peaks in scores at values <-0.5 , indicating
419 high confidence in a sedimentary class for most of the samples. However, the pattern of scores
420 also have long tails that are relatively constant between -0.5 and 1 . Iron-rich shale, is the worst
421 classified sedimentary rock, also has a slight increase in scores above 0.5 suggesting a high
422 confidence in the prediction as an igneous protolith. As a result, removing samples with low
423 $(0,0.5)$ igneous scores will not be able to filter out the majority of these misclassified samples.
424 Iron-rich shales are a very common sedimentary protolith; representing $\sim 12\%$ of sedimentary
425 samples in the global database. However, because of the dominance of igneous samples in the
426 database, filtering misclassified sedimentary rocks from the igneous predicted classes may not
427 result in a significant bias in chemical analyses. Furthermore, while the misclassification rate of
428 iron-rich shales is relatively high ($>32\%$; Figure 11b and d), the identification of a metamorphic
429 protolith within a large database will likely be correct in most cases.

430 How well the classifier generally performs on individual units with unknown class is still
431 uncertain. What we do not yet know from this analysis is whether the errors are due to
432 wholesale misidentification of individual shale units, or whether a collection of samples within
433 individual shale units each have a 20 to 30% probability of being misclassified. If the latter
434 is true, then viewing the predicted protoliths as a collection of samples for the same unit will
435 increase confidence in the accuracy of the prediction.

436 6.2. *Extrinsic uncertainties*

437 Beyond classification-based uncertainties that arise from these models, there are extrinsic
438 uncertainties that are not addressed by this study. Beyond the large bias towards igneous
439 rocks, the database is not necessarily representative of the proportions of specific rock types
440 within Earth. The proportions of various rock types change between the continents and oceans,
441 vertically within the crust, and from one terrane to another. Sedimentary rocks are more
442 common in the shallow Earth whereas igneous and metaigneous rocks are more common with

depth (Wilkinson et al., 2009). Rifts are filled with sediments whereas arcs are constructed from volcanics and plutonics. Among igneous rocks, more felsic compositions are typically concentrated in the upper continental crust with more mafic concentrations in the lower oceanic crust (Rudnick and Gao, 2003). How these variations affect the reliability of the classifiers is beyond the scope of this study, but present interesting avenues of future work.

7. Conclusions

The first step to interpreting the tectonic and geodynamic history of a terrane requires the basic identification of protoliths as igneous or sedimentary, which can be obscured or destroyed by metamorphic processes. Existing methods of chemical discrimination of protoliths are generally poor and/or are not optimized for the observed global distribution of chemical compositions. Using a recent global geochemical compilation of whole-rock chemical analyses, we demonstrate utility of machine learning methods for estimating a metamorphic protolith as igneous or sedimentary based on major oxide composition. We combine unmetamorphosed and metamorphosed samples in this analysis based on the geochemical similarity between igneous protoliths and their metamorphosed counterparts and similar patterns among sedimentary and metasedimentary samples. The method is simple to implement and provides more accurate estimates of protolith discrimination than common discrimination methods.

Machine learning improves the ability of protoliths to be discriminated by their major element composition. We find that it is possible to accurately determine a sample’s protolith using ensemble decision tree classification schemes, specifically RUSboost (95% of true igneous and 85% of true sedimentary). Our preferred classifier contains 30 learners and 1000 branches. A classification function is constructed that can be used to classify unknown samples. The classifier performs similarly well on a training and validation dataset. The true-positive rate varies for individual rock types, performing best for mafic igneous rocks and quartz-rich sedimentary samples. The classifier performs poorest—though a majority are still classified correctly (>75% correct)—among intermediate to felsic sedimentary rocks (i.e., iron-rich shales) because they are very similar chemically to felsic igneous rocks.

The performance of the RUSBoost method is better than conventional chemical discrimination diagrams. Therefore, we recommend using the protolith classifier in cases where the protolith is unknown. While there is no substitute for field relationships, textural indicators

473 and zircon date spectra, the classifier can be used as a cheap and independent tool to improve
474 confidence in observational-based predictions. Additionally, samples with ambiguous or miss-
475 ing protolith origin in global geochemical databases (e.g., Earthchem.org or Gard et al. (2019))
476 can be now be estimated to increase the potential size of datasets used to study specific rock
477 types or environments. We provide a MATLAB function that can be applied to a spreadsheet
478 of geochemical analyses, returning a predicted class and estimated confidence score.

479 8. Acknowledgements

480 We thank A. Reading, R. Holder and two anonymous reviewers, whose comments greatly im-
481 proved this work. M. Gard is supported by Australian Government Research Training Program
482 Scholarship. D. Hasterok and D. Kelsey are supported by the Australian Government through
483 the ARC Discovery Projects funding scheme (projects DP180104074 and DP160101006) and C.
484 Bishop was supported by the ARC Linkage Project funding scheme (project LP160100578). The
485 views expressed herein are those of the authors and are not necessarily those of the Australian
486 Government or Australian Research Council. This paper is Mawson Centre for Geoscience
487 contribution XXXX.

488 9. Computer Code Availability

489 The preferred classifiers, codes to run them, and templates for preparing geochemical data
490 can be found at <https://dx.doi.org/10.5281/zenodo.2586461> and http://github.com/dhasterok/global_geochemistry/tree/master/protolith/. MATLAB is required to run
491 the scripts, load the training and validation datasets, and use the classifier.
492

493 References

- 494 Abdi, H., Williams, L.J., 2010. Principal component analysis. *Wiley Interdisciplinary Reviews:*
495 *Computational Statistics* 2, 433–459. doi:10.1002/wics.101.
- 496 Aitchison, J., 1986. The statistical analysis of compositional data: Monographs on statistics
497 and applied probability. Chapman & Hall Ltd.
- 498 Best, M., 1982. Igneous and Metamorphic Petrology. W.H. Freeman, New York.
- 499 Breiman, L., 1996. Bagging predictors. *Machine Learning* 24, 123–140. doi:10.1023/A:
500 1018054314350.
- 501 Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees.
502 Chapman & Hall.
- 503 Bucher, K., Grapes, R., 2011. Petrogenesis of Metamorphic Rocks. Springer Berlin Heidelberg.
504 doi:10.1007/978-3-540-74169-5.
- 505 Carranza, E.J.M., Laborte, A.G., 2015. Random forest predictive modeling of mineral prospec-
506 tivity with small number of prospects and data with missing values in Abra (Philippines).
507 *Computers & Geosciences* 74, 60–70. doi:10.1016/j.cageo.2014.10.004.
- 508 Cracknell, M.J., Reading, A.M., 2014. Geological mapping using remote sensing data: A com-
509 parison of five machine learning algorithms, their response to variations in the spatial distri-
510 bution of training data and the use of explicit spatial information. *Computers & Geosciences*
511 63, 22–33. doi:10.1016/j.cageo.2013.10.008.
- 512 Crisci, C., Ghattas, B., Perera, G., 2012. A review of supervised machine learning algorithms
513 and their applications to ecological data. *Ecological Modelling* 240, 113–122. doi:10.1016/
514 j.ecolmodel.2012.03.001.
- 515 Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric
516 logratio transformations for compositional data analysis. *Mathematical Geology* 35, 279–
517 300. doi:10.1023/a:1023818214614.
- 518 Ehlers, E., Blatt, H., 1982. Petrology: igneous, sedimentary, and metamorphic. W.H. Freeman,
519 San Francisco.

520 Freund, Y., Shapire, R., 1996. Experiments with a new boosting algorithm, in: Machine
 521 Learning: Proceedings of the Thirteenth International Conference, pp. 148–156.

522 Gard, M., Hasterok, D., Halpin, J., 2019. Global whole-rock geochemical database compilation.
 523 Earth System Science Data Discussions , 1–23doi:10.5194/essd-2019-50.

524 Hasterok, D., Gard, M., Bishop, C., Kelsey, D., 2019. Geochemical data for protolith classifi-
 525 cation testing. doi:10.5281/zenodo.2586461.

526 Hasterok, D., Gard, M., Webb, J., 2018. On the radiogenic heat production of metamorphic,
 527 igneous, and sedimentary rocks. Geoscience Frontiers 9, 1777–1794. doi:10.1016/j.gsf.
 528 2017.10.012.

529 Hasterok, D., Webb, J., 2017. On the radiogenic heat production of igneous rocks. Geoscience
 530 Frontiers 8, 919–940. doi:10.1016/j.gsf.2017.03.006.

531 Herron, M.M., 1988. Geochemical classification of terrigenous sands and shales from
 532 core or log data. SEPM Journal of Sedimentary Research 58. doi:10.1306/
 533 212f8e77-2b24-11d7-8648000102c1865d.

534 Hood, S.B., Cracknell, M.J., Gazley, M.F., 2018. Linking protolith rocks to altered equiva-
 535 lents by combining unsupervised and supervised machine learning. Journal of Geochemical
 536 Exploration 186, 270–280. doi:10.1016/j.gexplo.2018.01.002.

537 Irvine, T.N., Baragar, W.R.A., 1971. A guide to the chemical classification of the common
 538 volcanic rocks. Canadian Journal of Earth Sciences 8, 523–548. doi:10.1139/e71-055.

539 Kirkwood, C., Cave, M., Beamish, D., Grebby, S., Ferreira, A., 2016. A machine learning
 540 approach to geochemical mapping. Journal of Geochemical Exploration 167, 49–61. doi:10.
 541 1016/j.gexplo.2016.05.003.

542 Kotsiantis, S., 2007. Supervised machine learning: A review of classification techniques. Infor-
 543 matica 31, 249–268.

544 Le Bas, M., Streckeisen, A., 1991. The IUGS systematics of igneous rocks. J. Geol. Soc.,
 545 London 148, 825–833.

546 Le Bas, M., Subbarao, K., Walsh, J., 2002. Metacarbonatite or marble? — the case of the
547 carbonate, pyroxenite, calcite–apatite rock complex at Borra, Eastern Ghats, India. *Journal*
548 *of Asian Earth Sciences* 20, 127–140. doi:10.1016/s1367-9120(01)00030-x.

549 Li, X.P., Wang, X., Chen, S., Storey, C., Kong, F.M., Schertl, H.P., 2018. Petrology and zircon
550 U–Pb dating of meta-calcsilicate from the Jiaobei terrane in the Jiao-Liao-Ji Belt of the
551 North China craton. *Precambrian Research* 313, 221–241. doi:10.1016/j.precamres.2018.
552 04.018.

553 Lindsey, D.A., 1999. An evaluation of alternative chemical classifications of sandstones. Open-
554 file Report 99-346. Reston, VA. URL: <http://pubs.er.usgs.gov/publication/ofr99346>.

555 Maesschalck, R.D., Jouan-Rimbaud, D., Massart, D., 2000. The mahalanobis distance. *Chemo-*
556 *metrics and Intelligent Laboratory Systems* 50, 1–18. doi:10.1016/s0169-7439(99)00047-7.

557 Mason, B., 1952. *Principles of Geochemistry*. J Wiley & Sons.

558 MATLAB[®], 2018. MATLAB Documentation: Classification Learner. r2018b ed. Mathworks.
559 URL: <https://au.mathworks.com/help/stats/classificationlearner-app.html>.

560 Middlemost, E., 1994. Naming materials in the magma/igneous rock system. *Earth Sci. Rev.*
561 37, 215–224. doi:10.106/0012-8252(94)90029-9.

562 Misra, S., 1971. Chemical distinction of high-grade ortho- and para-metabasites. *Norsk Geol-*
563 *ogisk Tidsskrift* 51, 311–316.

564 Moine, B., De La Roche, H., 1968. Nouvelle approche du problème de l'origine des amphibolites
565 à partir de leur composition chimique. *CR Acad. Sci. Paris D* 267, 284–287.

566 Praveena, M., Jaiganesh, V., 2017. A literature review on supervised machine learning algo-
567 rithms and boosting process. *International Journal of Computer Applications* 169, 32–35.
568 doi:10.5120/ijca2017914816.

569 Reverdatto, V., Selyatitskiy, A., Carswell, D., 2008. Geochemical distinctions between ‘crustal’
570 and mantle-derived peridotites/pyroxenites in high/ultrahigh pressure metamorphic com-
571 plexes. *Russian Geology and Geophysics* 49, 73–90. doi:10.1016/j.rgg.2008.01.002.

- 572 Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine
573 learning predictive models for mineral prospectivity: An evaluation of neural networks, ran-
574 dom forest, regression trees and support vector machines. *Ore Geology Reviews* 71, 804–818.
575 doi:10.1016/j.oregeorev.2015.01.001.
- 576 Roser, B., Korsch, R., 1988. Provenance signatures of sandstone-mudstone suites determined
577 using discriminant function analysis of major-element data. *Chemical Geology* 67, 119–139.
578 doi:10.1016/0009-2541(88)90010-1.
- 579 Rudnick, R., Gao, S., 2003. Composition of the continental crust, in: Rudnick, R. (Ed.),
580 Treatise on Geochemistry: The Crust. Elsevier. volume 3. chapter 1, pp. 1–64. doi:10.1016/
581 B978-0-08-095975-7.00301-6.
- 582 Seiffert, C., Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A., 2010. RUSBoost: A hybrid ap-
583 proach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics*
584 - Part A: Systems and Humans 40, 185–197. doi:10.1109/tsmca.2009.2029559.
- 585 Simonen, A., 1953. Stratigraphy and sedimentary of the svecofenidic, early archean supracrustal
586 rocks in south-western finland. *Bull. Commun. Geol. Finland* 160.
- 587 Tarney, J., 1977. Petrology, mineralogy, and geochemistry of the falkland plateau basement
588 rocks, site 330, deep sea drilling project, in: Initial Reports of the Deep Sea Drilling Project,
589 36. U.S. Government Printing Office. doi:10.2973/dsdp.proc.36.123.1977.
- 590 Templ, M., Hron, K., Filzmoser, P., Gardlo, A., 2016. Imputation of rounded zeros for high-
591 dimensional compositional data. *Chemometrics and Intelligent Laboratory Systems* 155,
592 183–190. doi:10.1016/j.chemolab.2016.04.011.
- 593 Turekian, K., 1969. The oceans, streams and atmosphere, in: Handbook of geochemistry.
594 Springer-Verlag Berlin, Heidelberg, New York. volume 1, pp. 297–323.
- 595 Werner, C., 1987. Saxonian granulites: a contribution to the geochemical diagnosis of original
596 rocks in high-metamorphic complexes. *Gerlands Beitrage zur Geophysik* 96, 271–290.
- 597 Wilkinson, B.H., McElroy, B.J., Kesler, S.E., Peters, S.E., Rothman, E.D., 2009. Global

598 geologic maps are tectonic speedometers—rates of rock cycling from area-age frequencies.
599 Geological Society of America Bulletin 121, 760–779. doi:10.1130/b26457.1.

600 Winchester, J.A., Park, R.G., Holland, J.G., 1980. The geochemistry of lewisian semipelitic
601 schists from the gairloch district, wester ross. Scottish Journal of Geology 16, 165–179.
602 doi:10.1144/sjg16020165.

603 Winkler, H., 1979. Petrogenesis of Metamorphic Rocks. 5th ed., Springer-Verlag.

Table 1: RUSBoost classifier overall performance using 30 trees and 1000 branches using the untransformed and unfiltered dataset.

		predicted protolith			
		igneous		sedimentary	
	true	N	%	N	%
<i>training dataset</i>					
igneous	447669	428440	95.7	19229	4.3
sedimentary	32355	3258	10.1	29097	89.9
<i>validation dataset</i>					
igneous	49732	47475	95.5	2257	4.5
sedimentary	3604	530	14.7	3074	85.3

Table 2: RUSBoost classifier performance for individual rock types using 30 trees and 1000 branches using the untransformed and unfiltered dataset.

rock type ^a	training dataset			validation dataset		
	true positives	false negatives	% FN	true positives	false negatives	% FN
<i>true igneous samples</i>						
quartzolite	36	563	94	5	66	93
granite	62077	5037	7.5	6704	573	7.9
granodiorite	31549	5538	14.9	3553	627	15
diorite	33692	1826	5.1	3659	206	5.3
gabbroic diorite	56262	610	1.1	6209	76	1.2
subalkalic gabbro	99156	306	0.3	11090	35	0.3
peridotgabbro	2626	67	2.5	313	4	1.3
crustal peridotite	621	26	4	62	2	3.1
syenite	7883	222	2.7	921	35	3.7
quartz monzonite	14400	822	5.4	1568	103	6.2
monzonite	15443	1010	6.1	1736	113	6.1
monzodiorite	18304	385	2.1	2020	39	1.9
monzogabbro	14227	111	0.8	1546	17	1.1
alkalic gabbro	27075	144	0.5	3077	19	0.6
foid syenite	3376	57	1.7	350	5	1.4
foid monzosyenite	1802	73	3.9	214	10	4.5
foid monzodiorite	2729	54	1.9	320	11	3.3
foid gabbro	12975	143	1.1	1498	19	1.3
ultra-high alkali igneous	280	7	2.4	36	4	10
foidolite	3377	91	2.6	399	12	2.9
sanukitoid	1931	84	4.2	190	8	4
picrite/alkali picrite	3021	45	1.5	313	6	1.9
komatiite/meimechite	3697	36	1	381	8	2.1
mantle peridotite/pyroxenite	2627	2	0.1	303	3	1
carbonatite	665	433	39.4	81	69	46
silicocarbonatite	924	284	23.5	87	36	29
<i>true sedimentary samples</i>						
quartzite	3087	57	1.8	328	10	3
quartz arenite	147	0	0	24	0	0
litharenite	1300	10	0.8	156	2	1.3
sublitharenite	171	0	0	16	0	0
arkose	1953	455	18.9	222	72	24.5
subarkose	261	0	0	26	0	0
wacke	6136	478	7.2	639	84	11.6
shale	7129	617	8	754	96	11.3
iron-rich shale	3167	1469	31.7	304	219	41.9
iron-rich sand	1464	16	1.1	148	4	2.6
laterite/bauxite	308	1	0.3	37	2	5.1
limestone	2539	3	0.1	276	7	2.5
dolomite	1433	151	9.5	144	34	19.1

Only plutonic names for igneous rocks. See Figure 11 for volcanic equivalents.

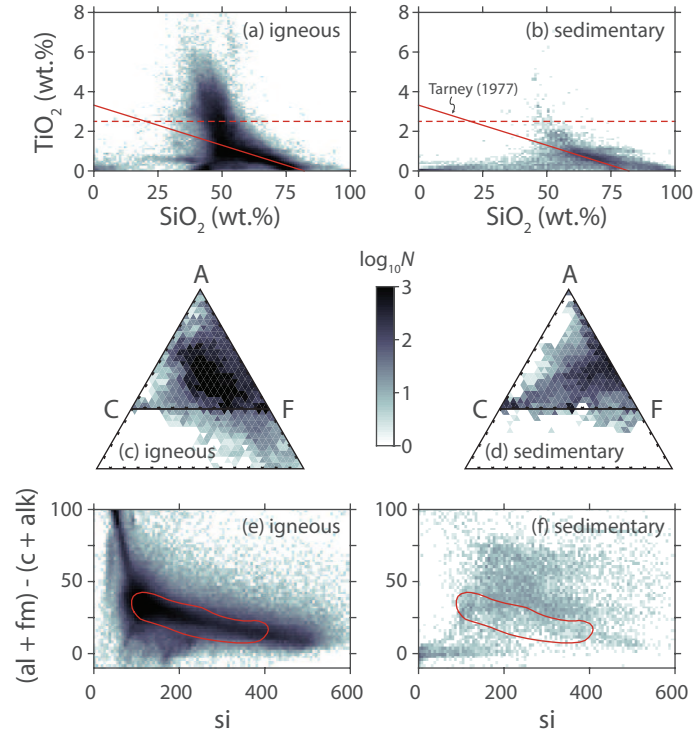


Figure 1: Examples of geochemical plots previously employed to identify igneous and sedimentary protoliths. Although extremes of igneous protoliths are distinguishable from sedimentary protoliths, the vast majority of igneous and sedimentary rock compositions overlap in discrimination diagrams designed to differentiate the two. The various shades of grey represent the number of data contained within each bin and share the same range located in the center of the figure. (a,b) TiO_2 content. The solid line represents the predicted division between sedimentary (above) and igneous (below) protoliths (Tarney, 1977). The dashed line indicates 2.5 wt.% TiO_2 with likely igneous field above the line. (c,d) ACF ternary system, where A is $\text{Al}_2\text{O}_3 - \text{Na}_2\text{O} - \text{K}_2\text{O}$, C is $\text{CaO} - 10/3\text{P}_2\text{O}_5$, and F is $\text{FeO} + \text{MgO} - \text{TiO}_2$, all expressed as molar quantities. (e,f) The parameters are determined from the Niggli indices ($\text{al} = 100\text{Al}_2\text{O}_3/n$; $\text{c} = 100\text{CaO}/n$; $\text{alk} = 100(\text{Na}_2\text{O} + \text{K}_2\text{O})/n$; $\text{fm} = 100(\text{FeO} + \text{MgO})/n$; and $\text{si} = 100\text{SiO}_2/n$; where $n = \text{Al}_2\text{O}_3 + \text{CaO} + \text{Na}_2\text{O} + \text{K}_2\text{O} + \text{FeO} + \text{MgO}$). The field encompasses the predicted igneous field (Simonen, 1953), digitized from Li et al. (2018). The data are the global geochemical database used throughout this study (Section 3).

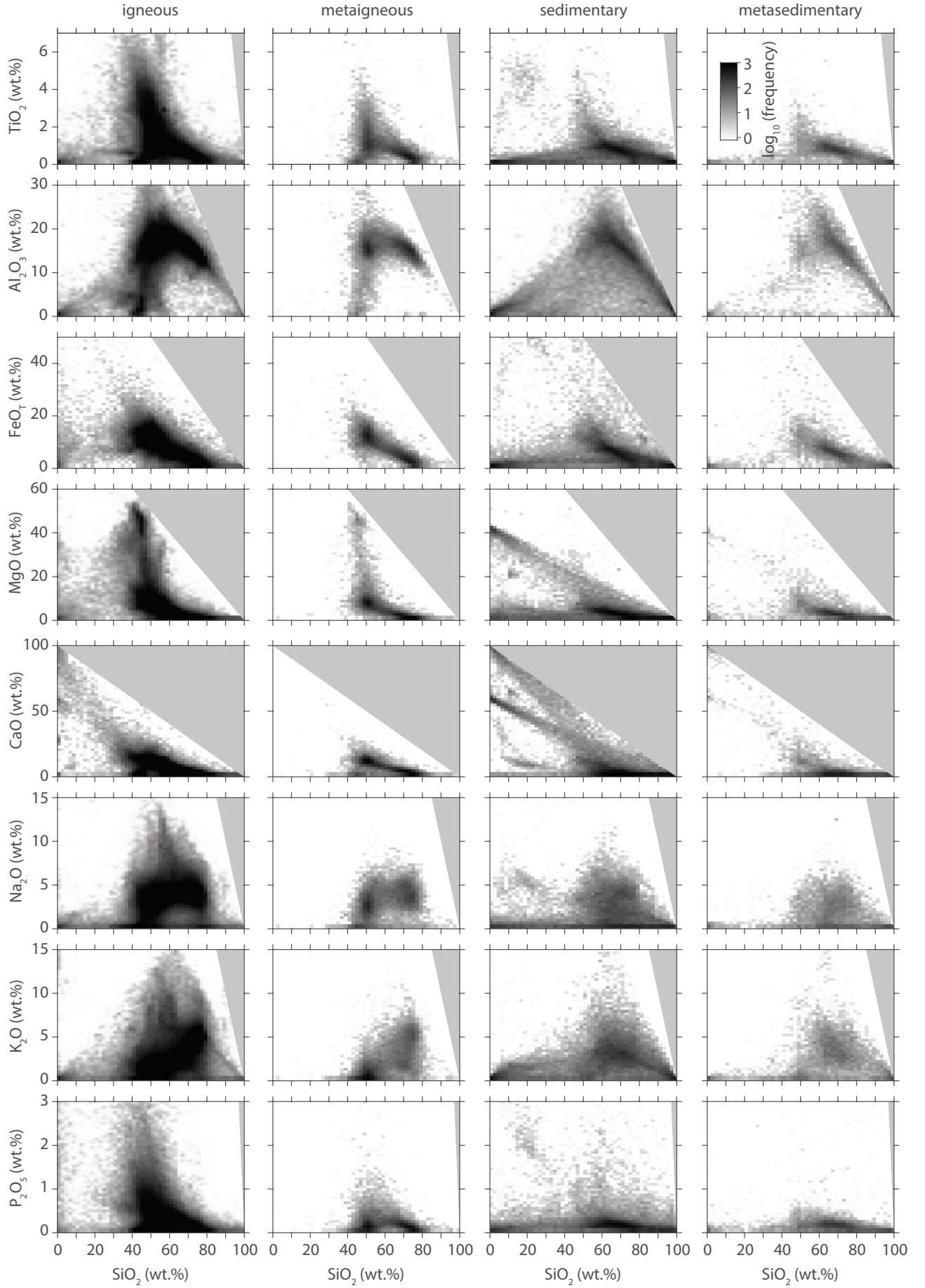


Figure 2: Harker diagrams illustrating variations in major oxides with respect to SiO_2 for igneous and sedimentary protoliths normalized on a volatile free basis to 100%. Because the igneous and sedimentary protoliths are similar to their metamorphosed counterparts, they may be used to increase the size of the dataset used to develop a machine-learning based classification scheme. The grey regions are regions of no data.

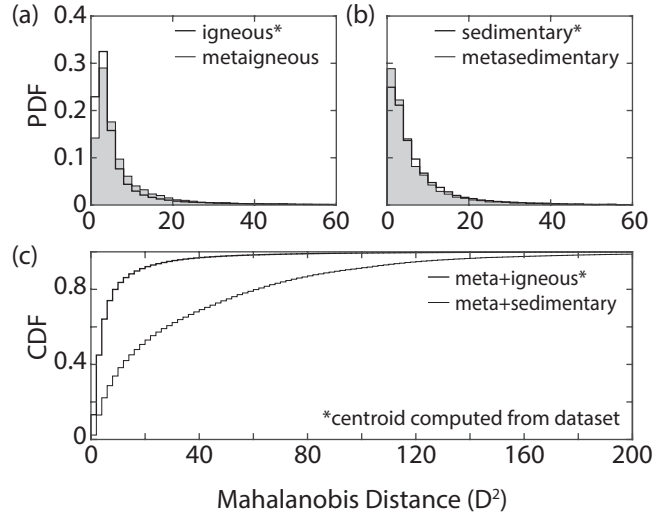


Figure 3: To measure the difference between igneous and sedimentary protolith compositions with many compositional parameters we use the Mahalanobis distance. Comparison of Mahalanobis distances for (a) igneous and metaigneous, (b) sedimentary and metasedimentary and (c) igneous + metaigneous and sedimentary + metasedimentary samples computed with reference to the (a) igneous centroid, (b) sedimentary centroid, and igneous + metaigneous centroid. PDF, probability density function and CDF, cumulative density function.

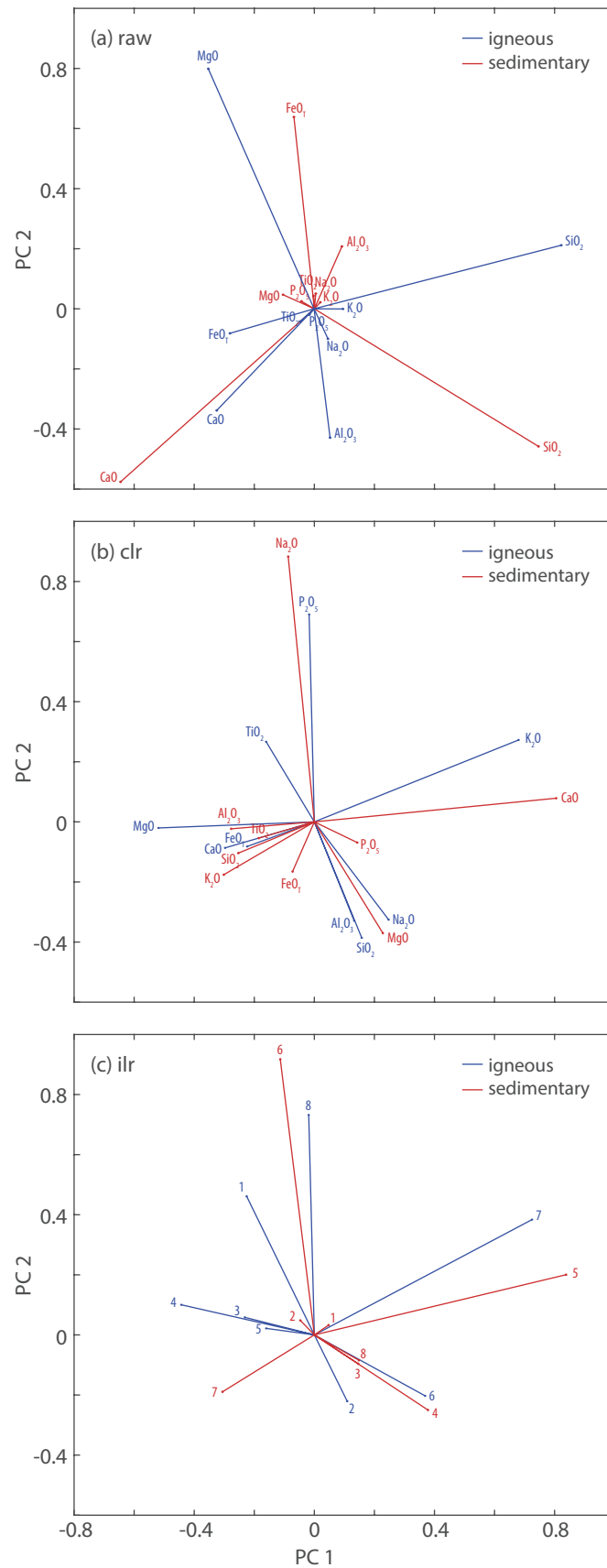


Figure 4: Principle component analysis is used to identify the source of dominant compositional variations within the geochemical dataset. Three methods are used to examine these as there are potential issues with simply using raw data due to the compositional dominance of SiO_2 and a redundant compositional dimension. Variations in chemistry with respect to the first two principal components for igneous + metigneous (blue) and sedimentary + metasedimentary (red) samples. Principle components are computed for (a) raw data, and the (b) centered log-ratio (clr) and (c) isometric log-ratio (ilr) transformed data. The ilr vectors are labeled with numbers rather than major oxides because the transformation removes the redundancy created by N compositional dimensions that sum to a fixed value. As a result, the scores (now $N - 1$) no longer represent individual compositional dimensions, instead each score becomes a variable combination of all the major oxides.

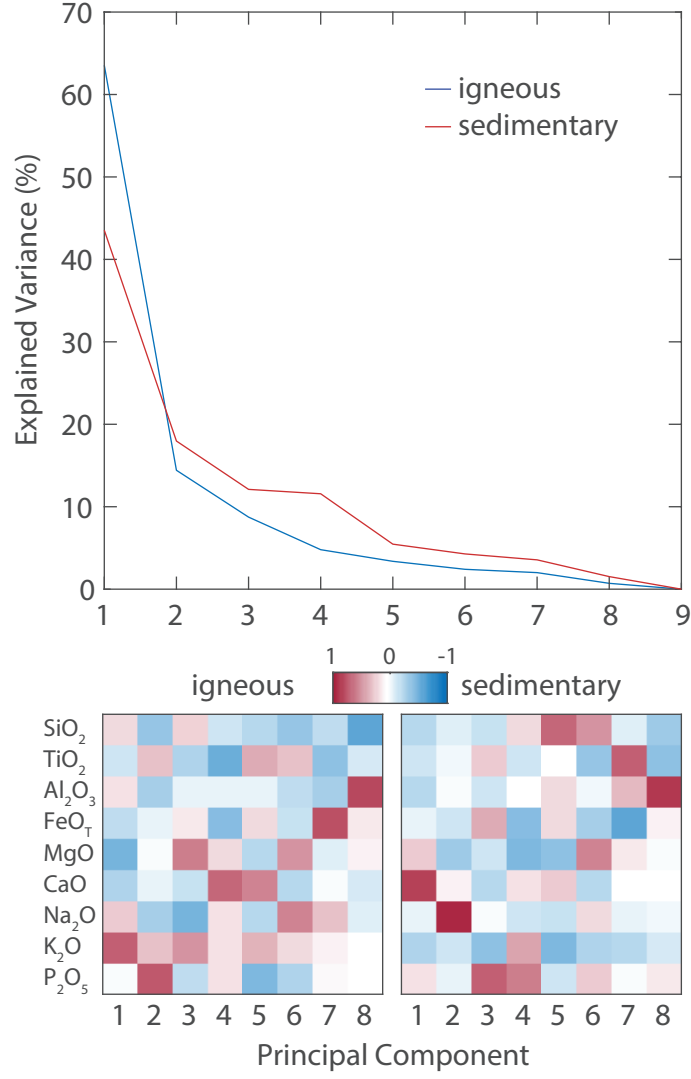


Figure 5: Principle component analysis on igneous and metamorphic training data scaled by a centered log-ratio transformation. Note the 9th principle component is excluded because it's eigenvalue is zero due to redundancy. (a) the variance reduction for each component. The principal component vectors for igneous + metaigneous (a) and sedimentary + metasedimentary (b) samples.

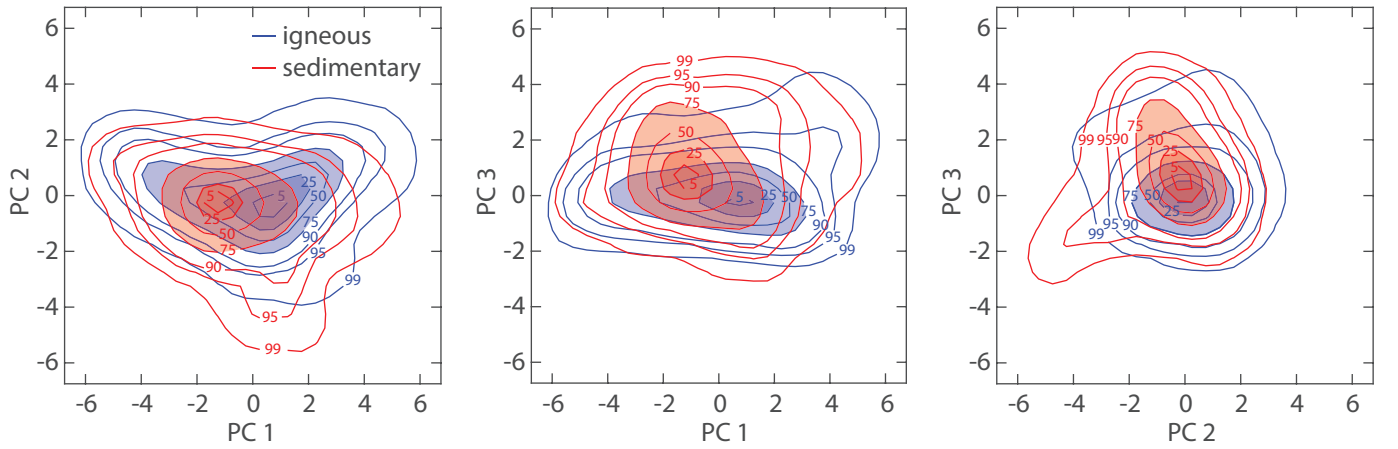


Figure 6: An examination of differences between the igneous and sedimentary protolith data using principle component analysis. Principle component scores for ilr transformed composition data with igneous (blue) and sedimentary (red) contours representing the shape of the PDF and labeled by the cumulative percentage contained within a given contour. The scores are computed with respect to the PCA only on the igneous subset. Shading emphasizes the contours which encircle the highest 25 and 75% of density data, respectively.

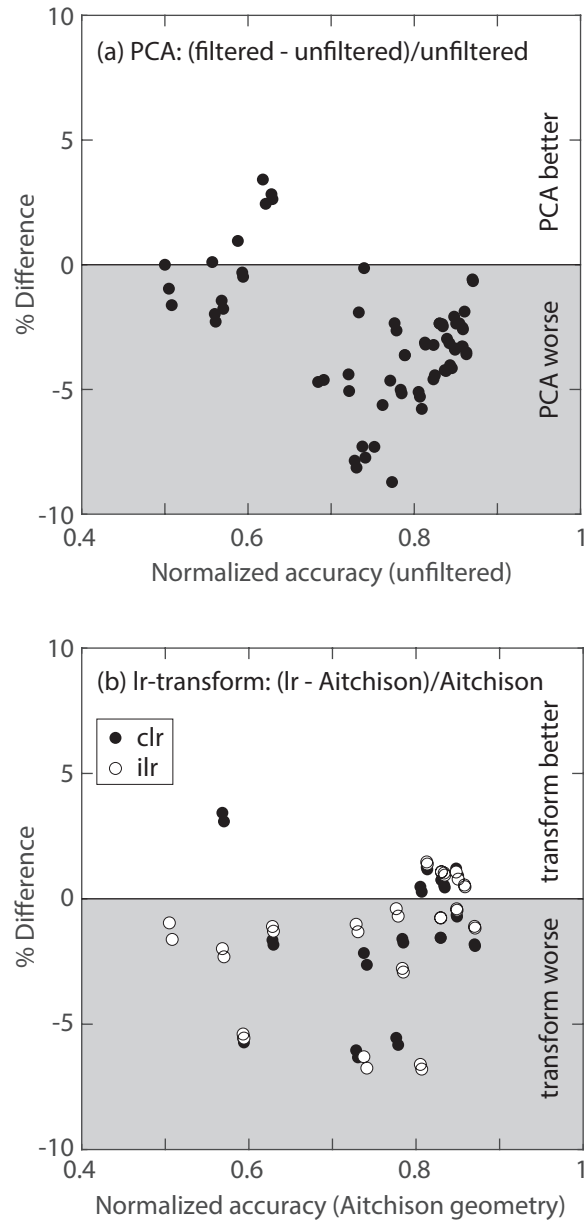


Figure 7: A test of log-ratio transforms of geochemical data on machine learning performance with respect to the untransformed normalized data. Comparison of models: (a) percent change in normalized true-positive rate of PCA filtered data trained classifier to unfiltered data trained classifier; and (b) percent change in normalized true-positive rate of clr or ilr transformed data trained classifier relative to untransformed (Aitchison) data trained classifier. A full list of models is provided in the supplementary material. Clusters occur as a result of changes to different parameter options for individual machine learning algorithms.

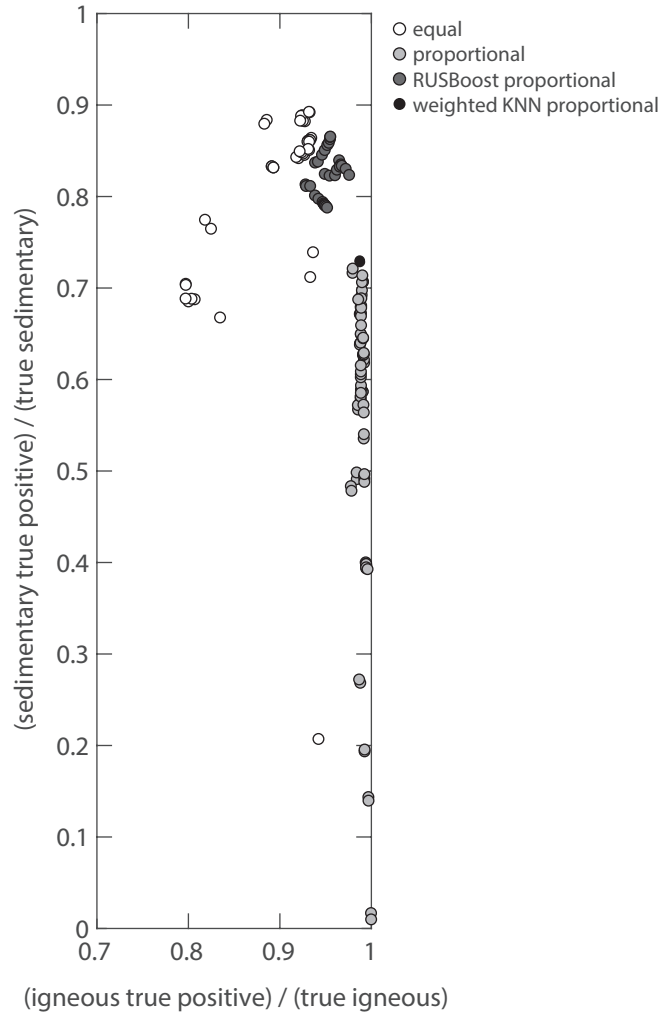


Figure 8: An assessment of classifier performance trained using a variety of machine learning algorithms. True positive igneous and sedimentary fractions for all classifiers trained in this study (Supplementary Table). Classifiers trained with equal set sizes (open circle), classifiers with 90% of the dataset randomly selected for training (light grey), RUSBoost classifiers with 90% of the dataset randomly selected for training (dark grey), and weighted KNN classifiers with 90% of the dataset randomly selected for training.

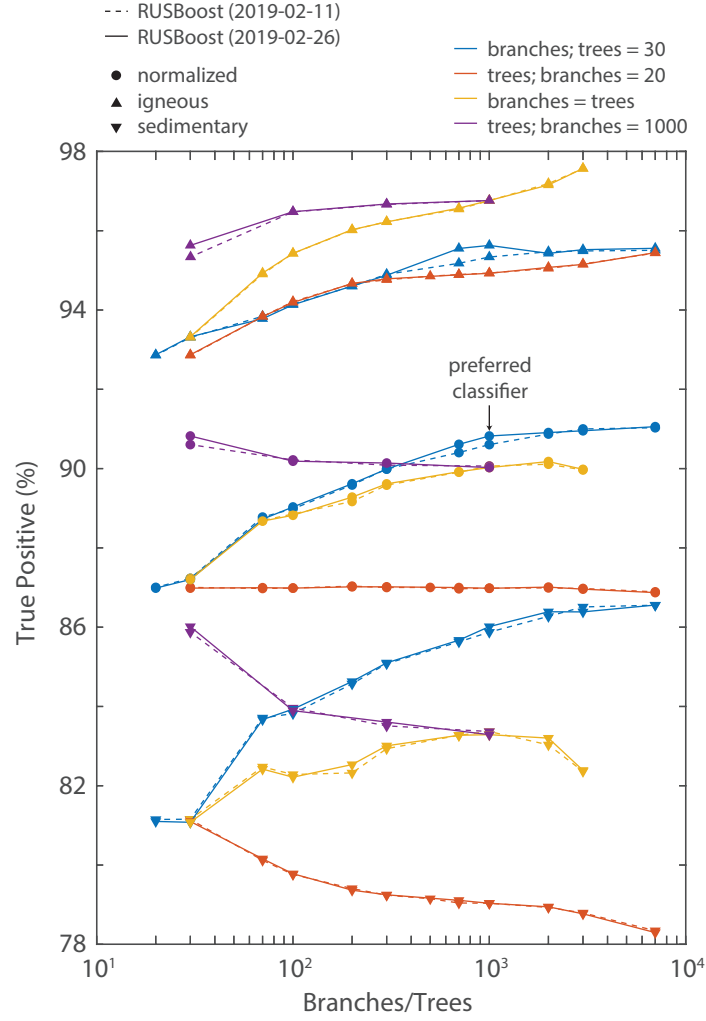


Figure 9: A suite of results for the RUSBoost classifier with differing combinations of branches and trees, used to identify the preferred classification parameters. Normalized (circles), igneous (up triangle), and sedimentary (down triangle) true-positive rate for a suite of RUSBoost classifiers. The results are shown for a constant number of trees (30, blue), a constant number of branches (20 orange, 1000 violet), and when the branches equal trees (yellow). The 90% of the global database with known protoliths are randomly selected to generate the training dataset. The test was run on two generated training datasets (2019-02-11 and 2019-02-26). The training dataset is untransformed and unfiltered.

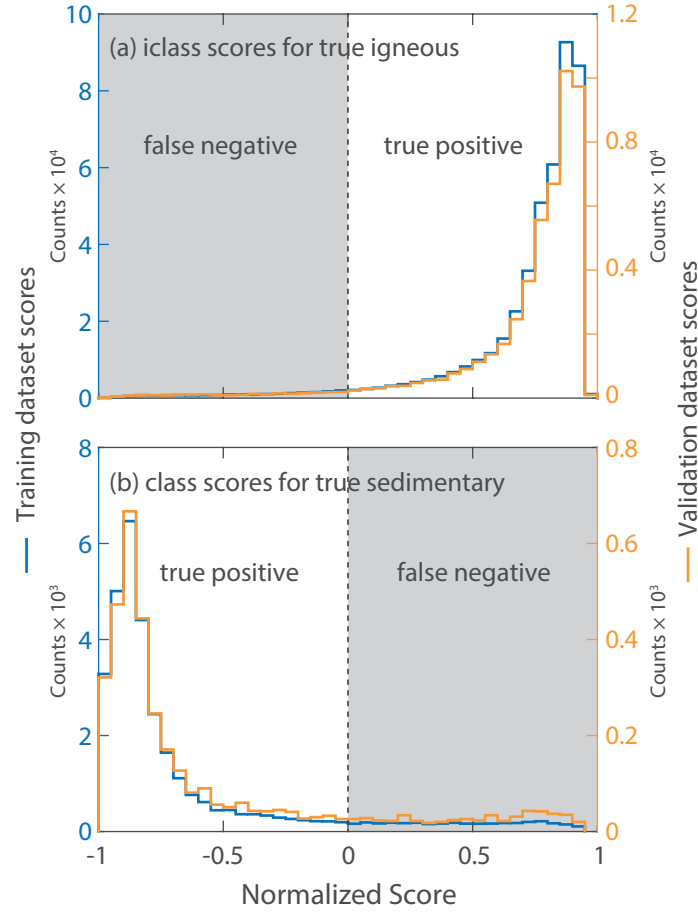


Figure 10: An assessment of the confidence in the predicted protolith class using the preferred classifier. Normalized scores for the preferred RUSBoost classifier (2019-02-26, 30 trees and a maximum of 1000 branches). Classifier scores for (a) true igneous (b) true sedimentary samples. The left axes correspond to the results for the training dataset (blue), whereas the right axes correspond to the results for the validation dataset (orange). The scores have been normalized such that the scores range from -1 to 1, with 0 demarcating the boundary between the assigned classes. Negative scores are predicted as sedimentary and positive values are predicted as igneous. The training dataset is the same as Figure 9 and the validation dataset represents the 10% of the data withheld from training.

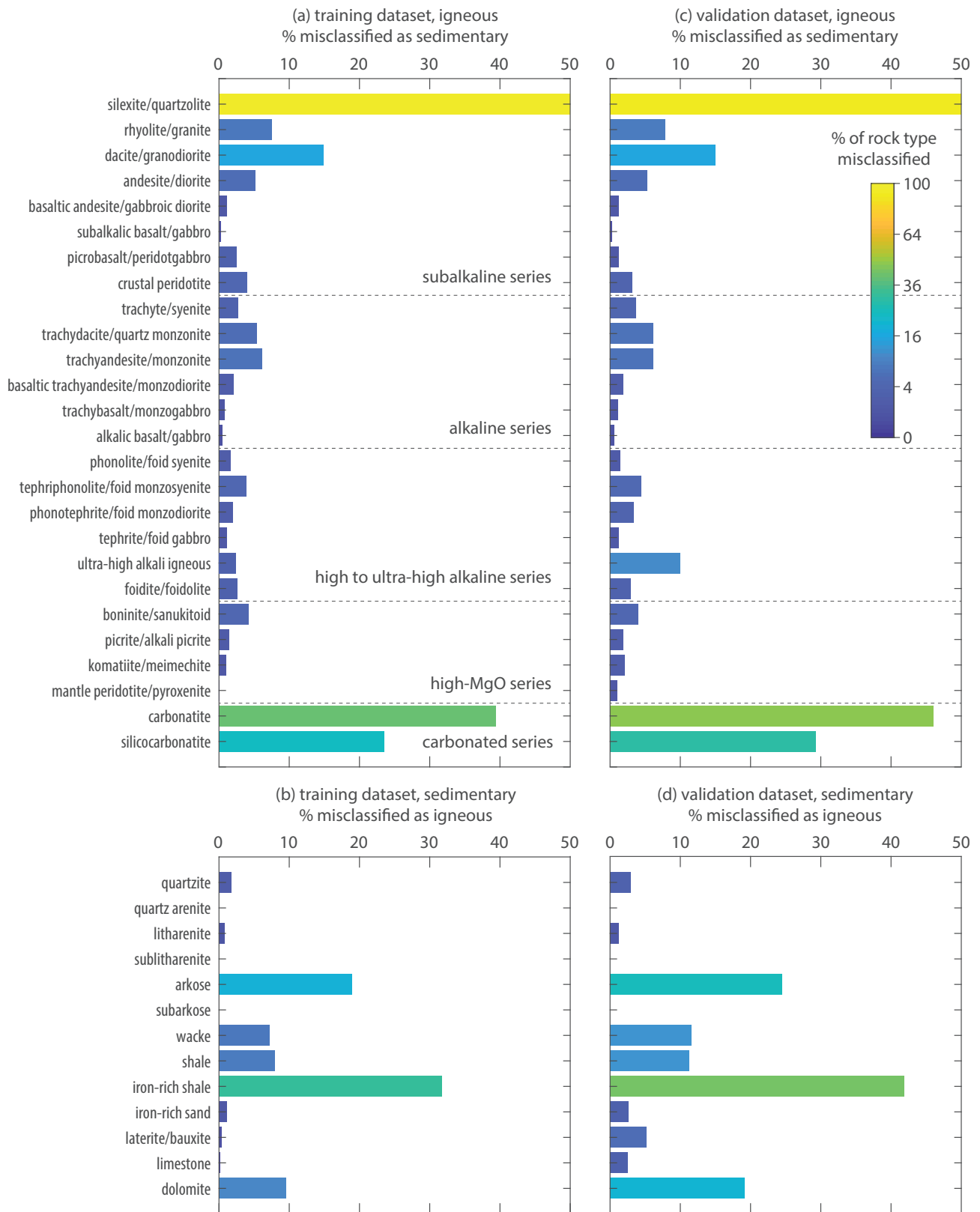


Figure 11: Misclassified protoliths by rock type for the preferred RUSBoost classifier within the training (a,b) and validation datasets (c,d), respectively. The bar lengths and color indicate the percentage of samples misclassified with respect to their total true rock type (a,c igneous and b,d sedimentary). The training dataset is the same as Figure 9 and the validation dataset includes 10% of the data withheld from training.

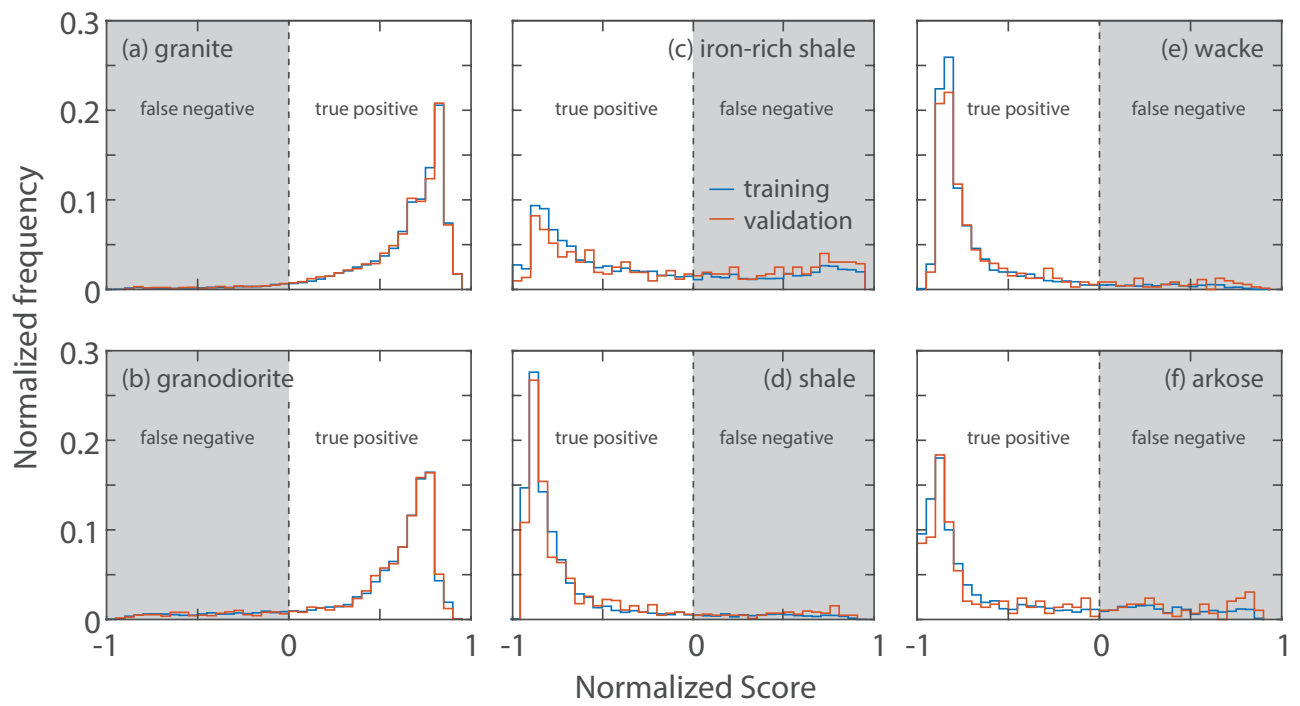


Figure 12: An examination of normalized scores for rock types that had higher rates of misidentification. Values <0 are classified as sedimentary and values >0 are classified as igneous.