Global whole-rock geochemical database compilation

Matthew Gard[1], Derrick Hasterok[1,2], and Jacqueline A. Halpin[3]

[1]Department of Earth Sciences, University of Adelaide, North Terrace, SA, 5005, Australia
[2]Centre for Tectonics Research and Exploration (TRaX), University of Adelaide, North Terrace, SA, 5005, Australia
[3]Institute for Marine and Antarctic Studies (IMAS), University of Tasmania, Hobart, Tasmania, 7001, Australia

Correspondence: Matthew Gard (matthew.gard@adelaide.edu.au)

Abstract. Collation and dissemination of geochemical data are critical to promote rapid, creative and accurate research and place new results in an appropriate global context. To this end, we have compiled a global whole-rock geochemical database, sourced from various existing databases and supplemented with an extensive list of individual publications. Currently the database stands at 1,022,092 samples with varying amounts of associated sample data including major and trace element concentrations, isotopic ratios, and location information. Spatial and temporal distribution is heterogeneous, however temporal distributions are enhanced over some previous database compilations, particularly in ages older than ∼1000 Ma. Also included are a range of geochemical indices, various naming schema and physical property estimates computed on a major element normalized version of the geochemical data for quick reference. This compilation will be useful for geochemical studies requiring extensive data sets, in particular those wishing to investigate secular temporal trends. The addition of physical properties, estimated from sample chemistry, represent a unique contribution to otherwise similar geochemical databases. The data are published in .csv format for the purposes of simple distribution, but exists in a structure format acceptable for database management systems (e.g. SQL). One can either manipulate this data using conventional analysis tools such as MATLAB®, Microsoft® Excel, or R, or upload to a relational database management system for easy querying and management of the data as unique keys already exist. The data set will continue to grow and be improved, and we encourage readers to contact us or other database compilations within about any data that is yet to be included. The data files described in this paper are available at https://doi.org/10.5281/zenodo.2592822 (Gard et al., 2019a).

1 Introduction

Geochemical analyses in conjunction with other temporal, spatial, and physical property information have been vital sources of information for understanding the Earth and investigating both local, and global geodynamic histories (e.g. Keller and Schoene, 2018). Effective collection, collation and dissemination of this type of data is critical to promote rapid, creative and accurate research. Every year, the amount of data recorded globally increases, dispersed among many hundreds of individual publications. Since the 1960's and 70's, broad element suites have been promptly accumulated due to the commercial availability of methods

such as x-ray fluorescence (XRF) and inductively coupled plasma mass spectrometry (ICP-MS), and thus modern publications are swiftly expanding our cumulative global data records. However, due to the rate of new publications, in conjunction with significant partitioning between different journals, this data is not always easy to find and can be incredibly time consuming to collate. It is pertinent that this information be readily available for future studies as all benefit from taking advantage of the full suite of data available to produce more robust models and constrained analyses.

Geochemical compilations have been used in a range of studies such as examining crustal magma reservoirs (e.g. Carbotte et al., 2013), proposing changes in mantle dynamics (e.g. Iwamori and Nakamura, 2015), to look at regional and global tectonic histories (e.g. Keller and Schoene, 2018), and examine the connections between life and the solid Earth (e.g. Cox et al., 2018). Not only does this information have implications for the scientific community, but also for issues such as environmental management, land use, and mineral resources development.

In this paper we present a global whole-rock geochemical database compilation consisting of modified whole-rock subsets from existing database compilations, in conjunction with significant supplementation from individual publications not yet included in these other collections. Additionally, we have generated naming schema, various geochemical indices, and other physical property estimates including density, seismic velocity and heat production for a range of the data contained within.

2 Existing Initiatives

Many existing initiatives have worked to construct and maintain database compilations with great success, but often restrict themselves to certain tectonic environments or regimes, regions, or rock types. EarthChem (https://www.earthchem.org) is currently the most notable general use geochemical data repository. It consists of many federated databases such as NAVDAT, PetDB, GEOROC, SedDB, MetPetDB and the USGS National Geochemical Database, as well as other individually submitted publications. The constituent databases are mostly more specialized compilations, for example:

- The North American Volcanic and Intrusive Rock Database (NAVDAT) has existed since 2002 and is primarily aimed at geochemical and isotopic data from Mesozoic and younger igneous samples of western North America (Walker et al., 2006). (http://www.navdat.org/)

- The Petrological Database of the Ocean Floor (PetDB) is the premier geochemical compilation suite for the igneous and metamorphic hosted data from mid-ocean ridges, back-arc basins, sea-mounts, oceanic crust and ophiolites (https://www.earthchem.org/petdb).

- Geochemistry of Rocks of the Oceans and Continents (GEOROC) is a more holistic compilation effort of chemical, isotope, and other data for igneous samples, including whole-rock, glass, minerals and inclusion analyses and metadata (http://georoc.mpch-mainz.gwdg.de).

- SedDB focuses on sedimentary samples, primarily from marine sediment cores. It has been static since 2014, and includes information such as major and trace element concentrations, isotopic ratios, and organic and inorganic components. (http://www.earthchem.org/seddb).

2

- MetPetDB is a database for metamorphic petrology, in a similar vein to PetDB and SedDB. This database also hosts large swathes of images collected through various methods such as x-ray maps and photomicrographs, although this information is not utilized in this paper (http://metpetdb.com/).

5
- The USGS National Geochemical Database archives geochemical information and its associated metadata from USGS studies and made available online (https://www.usgs.gov/energy-and-minerals/mineral-resources-program/science/national-geochemical-database).

Many other government initiatives and national databases exist, with notable examples including PETROCH from the Ontario Geological Survey (Haus and Pauk, 2010), New Zealand's national rock database (Petlab) (Strong et al., 2016), Australia's national whole-rock geochemical database (OZCHEM) (Champion et al., 2016), the Finnish lithogeochemical rock

10 geochemistry database (RGDB) (Rasilainen et al., 2007), the Newfoundland and Labrador Geoscience Atlas (Newfoundland and Labrador Geological Survey, 2010), and the basement rock geochemical database of Japanese islands (DODAI) (Haraguchi et al., 2018).

While all of these are generally exceptional enterprises, we personally found that the variety of structures were cumbersome to reconcile or otherwise deficient in some respect for our own research. Some examples included databases being deficient

15 in aged data (1000 Ma+), or lacking many recent publications. Some issues in certain existing databases were also evident; we found many samples missing information available in the original individual publications. It was quite common for age resolutions to be significantly larger than the values quoted within the paper itself, on the order of hundreds of millions of years in some cases, or not included at all because they were not found in a table but within the text itself.

Thus, we sought to produce a database incorporating refined samples from previous databases, and supplementing signif-

20 icantly from other, often recent, publications. Computed properties, naming schemes, and various geochemical indices have also been calculated where the data permits. Smaller subsets of previous iterations of this database have already been utilized for studies of heat production and phosphorus content (Hasterok and Webb, 2017; Hasterok et al., 2018; Cox et al., 2018; Gard et al., 2019b; Hasterok et al., 2019b), and this publication represents the totality of geochemical information gathered. As an ongoing process we have corrected some errors or omissions from previous databases as we have come across them, but have

25 not undergone a systematic effort to quality check the prior compilations. We intend to continue updating the database both in additional entries and in further clean up when necessary.

3 Database aggregation and structure

While other database structures are incredibly efficient, some of the intricacies of the systems make it difficult to utilize the information contained within. For example, we had issues when seeking estimated or measured ages of rock samples. In order

30 to examine temporal variations of chemistry and physical properties, an accurate and precise age is required. Under some of the present data management schemes it may be difficult to recover the desired data. Crystallization ages for older samples are often determined by U-Pb or Pb-Pb measurements from a suite of zircons. For a given sample, the individual zircon dates

may be contained within the database, and stored under mineral analyses. However, a search for rock chemistry may only return an estimated age (often a geologic timescale division). To get the crystallization age one would have to also download the individual mineral analyses, conduct an analysis on a concordia diagram (or similar), determine whether each individual analysis was valid, and then associate the result with the bulk chemistry. This process can be tedious and may be intractable.

5   Had the estimated crystallization age been attributed to the sample directly as often reported in the original study, much of this process could be short cut. Instead, our database attributes these estimated crystallization ages directly to the whole rock sample entry, which allows us to include estimated ages for the same unit or formation more readily. As a result the database presented here allows for a higher density of temporal sampling than other compilations.

     The database is provided in two formats; the first as a compressed single spreadsheet for people unfamiliar with database

10   management systems, and the second as a mixed flat file and relational database structure. Codd (1970) was the first to propose a relational model for database management. A relational structure organises data into multiple tables, with a unique key identifying each row of the sub-tables. These unique keys are used to link to other sub-tables. The main advantages of a relational database over a flat file format are that data is uniquely stored just once, eliminating data duplication, as well as performance increases due to greater memory efficiency and easy filtering and rapid queries.

15   Rather than utilize an entirely relational database format, we have adopted some flat file formats for the sub tables as to reduce the number of total tables to an amount more manageable for someone unfamiliar with SQL database structure. This format raises storage memory due to data duplication in certain fields (e.g. repetition of certain string contents across multiple samples, such as rock name). However, we believe this is a reasonable trade off for an easier to utilize structure for distribution, and makes using this data for someone unfamiliar with SQL simpler. Ideally we would host a purely relational database

20   structure online and be accessed via queries similar to the EarthChem Portal, but this is yet to be done.

     PostgreSQL was utilized as the relational database management system (RDBMS) to update and administer the database. PostgreSQL contains many built in features and useful addons including the geospatial database extender PostGIS which we utilize, has a large open source community and runs on all major operating systems.

     Python in conjunction with a PostgreSQL database adapter Psycopg are used to import new data efficiently. Data is copied

25   into a .csv template directly from publications to reduce any chance of transcribing errors, and dynamically uploaded to a temporary table in PostgreSQL. From here, the desired columns are automatically partitioned up and added to the database in their respective sub-tables. We iterate through a folder of new publications in this way, and are able to add data rapidly as a result.

     The database consists of 10 tables: trace elements, major elements, isotope ratios, sample information, rock group/origin/facies

30   triplets, age information, reference information, methods, country, and computed properties. The inter-connectivity of these tables is depicted in Figure 1, with tables linked via their respective id keys. A description of each of these tables is included in Table 1, and column names that require further details as well as computed property methods are detailed in Table 3. Individual subtables have been output as csv files for use. We suggest inserting these into a RDBMS for efficient queries and extraction of desired data. However, we have exported these in csv format in case people not familiar with database systems

35   wish to work with them in other programs such as Microsoft®Excel, MATLAB® or R. While technically inefficient, the largest

sub-table currently stands at only 280 MB uncompressed, which we believe to be an acceptable size for data manipulation. The compressed merged spreadsheet is only 130 MB.

Many samples include multiple geochemical analyses. These can vary from separate trace and major measurements with no overlap, to duplicate element analyses using different methods. In the case of some subsets of this data we have chosen to merge these multiple analyses into a singular entry in the database. This methodology has both benefits and drawbacks. While it reduces the difficulty in selecting individual samples analyses, it means that lower resolution geochemical methods are sometimes averaged with higher precision ones. In the future we hope to prioritise these higher precision methods where applicable (e.g. ICP-MS for many trace elements over XRF). Using a singular entry is simpler for many interdisciplinary scientists who don't wish to be slowed down by the complexity of managing duplicate samples and split analyses. We have generally kept track of this with the method field; where merging has occurred and both methods are known, we have concatenated the method in most cases.

4 Data statistics

4.1 Raw data

The largest existing database contributions to this database are listed in Table 2. Individual publication supplementation includes both new additions we have found in the literature, as well as cleaned up and modified entries from existing databases. The subsets of existing databases do not represent the entire collections for many of these programs as we have done pre-filtering to remove non-whole rock data or encountered issues with accessing the entire data set using online web forms.

Figure 2 denotes histograms of the various major, trace and isotope analyses within the database. The majority of isotope data was recently sourced from the GEOROC database. Unsurprisingly, major element analyses in general dwarf the number of trace element measurements recorded.

Despite the heterogeneous nature of geochemical sampling, there is still reasonable spatial coverage around the world. However, there are a noticeable dominance of samples sourced from North America, and additionally Canada, Australia, and New Zealand (Figure 3). The United States tops of the list with 352,761 samples, including those from their non-contiguous states. The African continent suffers the most from lack of data with regards to the rest of the globe (Figure 3).

Age distributions unsurprisingly show a significant dominance towards very recent samples (<50 Ma), due largely to the oceanic subset (Figure 4b). Age here is indicated as being an assumed crystallization age. Excluding major time-period associated ages (e.g. Paleoproterozoic age range of 2,500–1,600 Ma as the max and min age of a sample), there are 355,467 samples with estimated crystallisation age values. Of these, 282,147 have age uncertainty estimates and observing the cumulative distribution function of these values indicates that $\sim 99\%$ of the age uncertainties fall below ~ 150 Ma (Figure 4a).

Rock group and rock origin are described in Table 3. There is a clear dominance towards igneous samples, making up 72.37% of the data with known rock group information (Figure 5). About 99% of these igneous samples have a distinction noted as volcanic or plutonic in the rock origin field, with just over two thirds of these being volcanic. Sedimentary samples are the next most common rock group, however the vast majority of these have no classification in rock origin, and we aim to improve this

in future updates. Finally metamorphic rocks have $\sim 44\%$ of the samples with rock origin classifications. Meta-sedimentary origin is slightly more common than meta-igneous, however meta-igneous includes two further subdivisions of meta-volcanic and meta-plutonic where known.

4.2 Naming schema - rock_type

5   Nomenclature varies significantly within geology and unsurprisingly rock names within the database differ wildly as a result. Different properties such as texture, mineralogical assemblages, grain sizes, thermodynamic histories, and chemistry make up the majority of the basis for the various naming conventions utilized throughout, interspersed with author assumptions and/or inaccuracies. Thus, we sought a robust and consistent chemical classification scheme to assign rock names to the various samples of the database. This chemical basis classification scheme is stored in the computed table, within the rock_type field.

10   Differing naming work flows are applied to (meta-)igneous, and (meta-)sedimentary samples. For igneous, meta-igneous, and unknown protolith origin metamorphic samples, we use a total alkali-silica (TAS) schema (Middlemost, 1994) modified to include additional fields for further classification of high-Mg volcanics (Le Bas and Streckeisen, 1991). See Figure 6c and d for a partial visual description of the process. Furthermore we classify igneous rocks as carbonatites when the CO2 concentration exceeds 20 wt.%. These entries are assigned either the plutonic or volcanic equivalent rock names depending if the sample is

15   known to be of plutonic or volcanic origin.

For sedimentary and meta-sedimentary rocks, we first separate out carbonates and soils using ternary plot divisions of SiO_2, $Al_2O_3 + Fe_2O_3$, and $CaO + MgO$ (Mason, 1952; Turekian, 1969). Additionally, we further partition clasic sediments using the SedClass™classification method from Herron (1988). Quartzites are identified separately where SiO_2 exceeds 0.9 in the ternary system. See Hasterok et al. (2018) for further discussion.

20   A break down of the classification distributions are included in Figure 6a and b. Sub-alkalic basalt/gabbro is a significantly large contribution to the volcanic samples, due to the extent of samples of oceanic nature.

4.3 Computed properties

In numerical models, rock types are often assigned physical property estimates that have been derived from limited data sets. We compute a number of properties and naming schema for a significant subset of the database, a new addition over many previous

25   database compilations. This includes heat production, density and p-wave velocity estimates, as well as various geochemical indices and descriptors such as modified TAS, QAPF and SIA classifications. A full list of referenced methods and computed columns are given in Table 3.

Where computed values require major element concentrations, these properties and values have been calculated based on an LOI free major element normalised version of the database i.e. major element totals are normalised to 100, while preserving the

30   relative proportions of each individual elements contribution to the total. This normalisation occurs only on samples with major element totals between 85 and 120 wt.%. Totals lying outside this range are ignored, and properties requiring these values are not computed. The exact value of normalisation for each sample is recorded in the computed table, within the norm_factor field. Figure 7a, b and c denote some property estimates calculated from the normalised analyses.

4.3.1 Density estimates

Density is an important input for a wide range of models but only a small fraction of samples have measured density values associated with them. Contained within the database are a number of publications hosting density observations (e.g. Haus and Pauk, 2010; Barette et al., 2016; Slagstad, 2008). Following the method of Hasterok et al. (2018), we produce a set of simple oxide-based linear regression density models.

$$\rho_{\text{Low-Mg}} = 2506.22 + 204.82 \times \text{Fe}^* + 791.72 \times \text{Maficity} - 4.56 \times \text{MALI}, \qquad \text{Misfit} = 97 \text{ kg m}^{-3}$$

$$\rho_{\text{High-Mg}} = 3159.18 - 10.40 \times \text{MgO} + 1.36 \times \text{CaO}, \qquad \text{Misfit} = 149 \text{ kg m}^{-3}$$

$$\rho_{\text{Carb.}} = 3268.04 - 6.23 \times \text{SiO}_2 - 6.37 \times \text{CaO} - 2.88 \times \text{MgO}, \qquad \text{Misfit} = 147 \text{ kg m}^{-3}$$

where Fe^* is iron number, MALI is modified alkali-lime index, oxides are in weight percent and ρ is density in kg m^{-3}. Low-Mg, High-Mg and Carb. (carbonated rocks) refer to the specific models for different rock groups. See Hasterok et al. (2018) for further discussion of the model fits. Density estimates peak at $\sim$2680 and $\sim$2946 kg m^{-3} due to mafic and felsic sample medians respectively

4.3.2 Seismic velocity

We utilize the empirical model of Behn and Kelemen (2003) for estimating anhydrous p-wave seismic velocity. Their model was calibrated on $\sim$ 18,000 igneous rocks and validated against 139 high quality laboratory measurements. However this model does have limitations, as it was calibrated to anhydrous compositions only. utilizing their 3 oxide model, estimated uncertainty (1σ) is $\sim \pm 0.13\text{kms}^{-1}$. P-wave velocity estimates depict maximums at $\sim$6.2 and $\sim$7.1 m s^{-1} (Figure 7c). For further details or discussion, refer to Behn and Kelemen (2003) and Hasterok and Webb (2017).

$$Vp = 6.9 - 0.011 \times \text{SiO}_2 + 0.037 \times \text{MgO} + 0.045 \times \text{CaO},$$

where oxides are in weight percent and Vp is in ms^{-1}.

4.3.3 Heat production

Heat production is computed by employing the relationship from Rybach (1988). Heat production estimates are resolved by a smoother distribution in log-space than the dichotomous nature of the density and Vp estimates.

$$A(\mu W m^{-3}) = \rho \times (9.67 \times \text{U} + 2.56 \times \text{Th} + 2.89 \times \text{K}_2\text{O}) \times 10^{-5},$$

with concentrations of U, Th in ppm, K_2O in weight-percent and ρ in kg m^{-3}. Heat production has a median value of $\sim$1.0 $\mu\text{W m}^{-3}$, with first and third quartiles (25th and 75th percentiles) of 0.39 and 2.2 $\mu\text{W m}^{-3}$ respectively.

5 Improvements and future developments

5.1 Bibliographic information

Due to a high variety of sources and database formats, merging bibliographic information proved difficult. For individual publications and adjustments made manually, we have collated bibliographic information in higher detail. We hope to expand

this .bib file as we continue to clean up the reference lists and make adjustments to other compilations. For other inherited bibliographic information from external databases, the exact format can vary. These details are contained within the reference .csv and linked to each sample through the ref_id as seen in Figure 1.

5.2 Ownership and accuracy

Although every effort is made to ensure accuracy, there are undoubtedly some errors, either inherited or introduced. We make no claims to the accuracy of database entries or reference information. It is up to the user to validate subsets for their own analyses, and ideally contact the original authors, previous database compilation sources, or ourselves to correct errors where they exist. We make no claim on ownership of this data; when utilizing this database additionally cite the original authors and data sources.

6 Future Work

We have published portions of the database in the course of prior studies and will continue to expand this data set for our own research purposes. Small individual corrections have occurred incrementally with every version, and unfortunately we did not keep records of these improvements. Going forward, we plan to include a record of these corrections and forward them to the other database compilations as needed. We hope to work with existing compilation authors in the future to assist with new additions as well. This version of the database may be of use for these database initiatives to supplement their own records.

Utilizing this database we have worked on methods for predicting protoliths of metamorphic rocks (Hasterok et al., 2019a). As over 57% of the samples lack that information (Figure 5) this methodology may be included in future database versions. We are also making progress on a geologic provinces map that captures tectonic terranes. An associated set of software that can be used in MATLAB® to explore the database, including many of the individual methods cited above for the computed properties is also available on github at https://github.com/dhasterok/global_geochemistry.

7 Data availability

The BIB file and CSV tables of this data set are available on Zenodo: https://doi.org/10.5281/zenodo.2592822 (Gard et al., 2019a)

Author contributions. M. Gard and D. Hasterok worked on the processing codes and computed property estimates, as well as collation of data sources. M. Gard organised the database structure and framework codes, and prepared the manuscript with contributions from all co-authors. J. A. Halpin collated the Antarctic geochemical set.

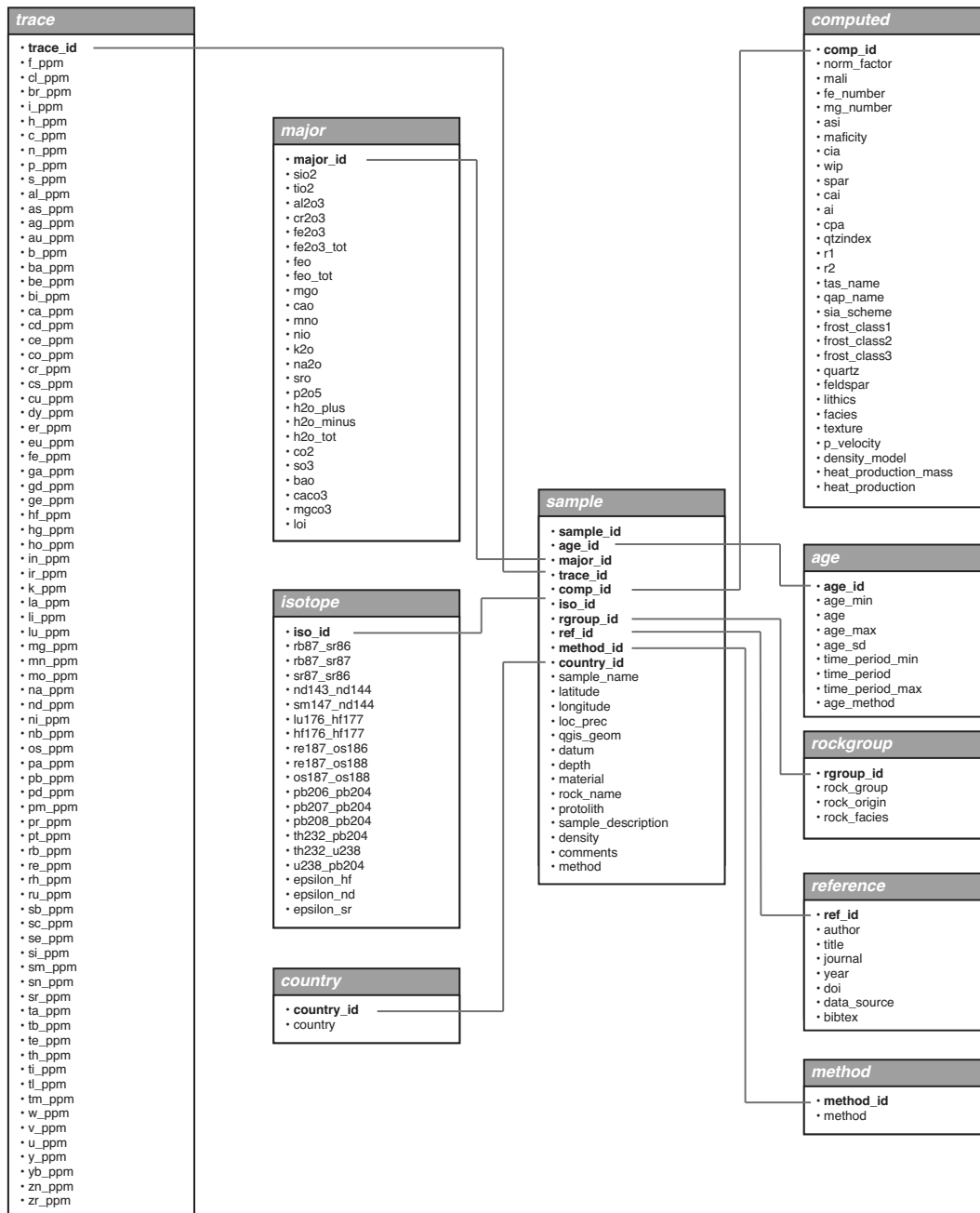*Competing interests.* The authors declare that they have no conflict of interest.

5

References

Barette, F., Poppe, S., Smets, B., Benbakkar, M., and Kervyn, M.: Spatial variation of volcanic rock geochemistry in the Virunga Volcanic Province: Statistical analysis of an integrated database, Journal of African Earth Sciences, https://doi.org/10.1016/j.jafrearsci.2016.09.018, 2016.

5   Barette, F., Poppe, S., Smets, B., Benbakkar, M., and Kervyn, M.: Spatial variation of volcanic rock geochemistry in the Virunga Volcanic Province: Statistical analysis of an integrated database, Journal of African Earth Sciences, 134, 888–903, https://doi.org/10.1016/j.jafrearsci.2016.09.018, http://www.sciencedirect.com/science/article/pii/S1464343X16303119, 2017.

Behn, M. and Kelemen, P. B.: Relationship between seismic P-wave velocity and the composition of anhydrous igneous and meta-igneous rocks, Geochem. Geophys. Geosyst., 4, 1041, https://doi.org/10.1029/2002GC000393, 2003.

10 Buggle, B., Glaser, B., Hambach, U., Gerasimenko, N., and Marković, S.: An evaluation of geochemical weathering indices in loess–paleosol studies, Quaternary International, 240, 12–21, https://doi.org/10.1016/j.quaint.2010.07.019, http://www.sciencedirect.com/science/article/pii/S1040618210002892, 2011.

Carbotte, S. M., Marjanović, M., Carton, H., Mutter, J. C., Canales, J. P., Nedimović, M. R., Han, S., and Perfit, M. R.: Fine-scale segmentation of the crustal magma reservoir beneath the East Pacific Rise, Nature Geoscience, 6, 866–870, https://doi.org/10.1038/ngeo1933,

15 https://doi.org/10.1038/ngeo1933, 2013.

Champion, D., Budd, A., Hazell, M., and Sedgmen, A.: OZCHEM National Whole Rock Geochemistry Dataset, Tech. rep., Geoscience Australia, https://ecat.ga.gov.au/geonetwork/srv/eng/catalog.search?node=srv#/metadata/65464[AccesssedJune2016], 2016.

Codd, E. F.: A Relational Model of Data for Large Shared Data Banks, Communications of the ACM, 13, 377–387, https://doi.org/10.1145/362384.362685, http://doi.acm.org/10.1145/362384.362685, 1970.

20 Cox, G. M., Lyons, T. W., Mitchell, R. N., Hasterok, D., and Gard, M.: Linking the rise of atmospheric oxygen to growth in the continental phosphorus inventory, Earth and Planetary Science Letters, 489, 28–36, https://doi.org/https://doi.org/10.1016/j.epsl.2018.02.016, http://www.sciencedirect.com/science/article/pii/S0012821X18300785, 2018.

Debon, F. and Le Fort, P.: A chemical–mineralogical classification of common plutonic rocks and associations, Transactions of the Royal Society of Edinburgh: Earth Sciences, 73, 135–149, https://doi.org/10.1017/S0263593300010117, 1983.

25 Frost, B. R., Barnes, C. G., Collins, W. J., Arculus, R. J., Ellis, D. J., and Frost, C. D.: A Geochemical Classification for Granitic Rocks, Journal of Petrology, 42, 2033–2048, https://doi.org/10.1093/petrology/42.11.2033, https://dx.doi.org/10.1093/petrology/42.11.2033, 2001.

Gard, M., Hasterok, D., and Halpin, J.: Global whole-rock geochemical database compilation (data files), https://doi.org/10.5281/zenodo.2592823, 2019a.

Gard, M., Hasterok, D., Hand, M., and Cox, G.: Variations in continental heat production from 4 Ga to the present - Evidence from geo-

30 chemical data, Lithos, 342-343, 391–406, https://doi.org/10.1016/j.lithos.2019.05.034, 2019b.

Geological Survey of Greenland: Ujarassiorit 1989–2011, Tech. rep., http://www.ujarassiorit.gl. Downloaded June 2016., 2011.

Haraguchi, S., Ueki, K., Yoshida, K., Kuwatani, T., Mohamed, M., Horiuchi, S., and Iwamori, H.: Geochemical database of Japanese islands for basement rocks: compilation of domestic article, Journal of Geological Society of Japan, 124, 1049–1054, https://doi.org/10.5575/geosoc.2018.0027, 2018.

35 Hasterok, D. and Webb, J.: On the radiogenic heat production of igneous rocks, Geoscience Frontiers, 8, 919–940, https://doi.org/10.1016/j.gsf.2017.03.006, http://www.sciencedirect.com/science/article/pii/S1674987117300579, 2017.

Hasterok, D., Gard, M., and Webb, J.: On the radiogenic heat production of metamorphic, igneous, and sedimentary rocks, Geoscience Frontiers, 9, 1777–1794, https://doi.org/10.1016/j.gsf.2017.10.012, http://www.sciencedirect.com/science/article/pii/S1674987117301937, 2018.

Hasterok, D., Gard, M., Bishop, C., and Kelsey", D.: "Chemical identification of metamorphic protoliths using machine learning methods", "Computers and Geosciences", 132, 56–68, https://doi.org/10.1016/j.cageo.2019.07.004, 2019a.

Hasterok, D., Gard, M., Cox, G., and Hand, M.: A 4 Ga record of granitic heat production - Implications for geodynamic evolution and crustal composition of the early Earth, Precambrian Research, 331, https://doi.org/10.1016/j.precamres.2019.105375, 2019b.

Haus, M. and Pauk, T.: Data from the PETROCH lithogeochemical database, Miscellaneous release—data 250, Ontario Geol. Surv., 2010.

Herron, M. M.: Geochemical Classification of Terrigenous Sands and Shales from Core or Log Data, SEPM Journal of Sedimentary Research, 58, https://doi.org/10.1306/212f8e77-2b24-11d7-8648000102c1865d, 1988.

Iwamori, H. and Nakamura, H.: Isotopic heterogeneity of oceanic, arc and continental basalts and its implications for mantle dynamics, Gondwana Research, 27, 1131–1152, https://doi.org/10.1016/j.gr.2014.09.003, http://www.sciencedirect.com/science/article/pii/S1342937X14002743, 2015.

Keller, B. and Schoene, B.: Plate tectonics and continental basaltic geochemistry throughout Earth history, Earth and Planetary Science Letters, 481, 290–304, https://doi.org/10.1016/j.epsl.2017.10.031, http://www.sciencedirect.com/science/article/pii/S0012821X1730599X, 2018.

la Roche, H. D., Leterrier, J., Grandclaude, P., and Marchal, M.: A classification of volcanic and plutonic rocks using R1R2-diagram and major-element analyses – Its relationships with current nomenclature, Chemical Geology, 29, 183–210, https://doi.org/10.1016/0009-2541(80)90020-0, http://www.sciencedirect.com/science/article/pii/0009254180900200, 1980.

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

Lett, R. and Ronning, C.: BC rock geochemical database - British Columbia Geological Survey Geofile 2005-14, Tech. rep., British Columbia Ministry of Energy and Mines, 2005.

Mamani, M., Worner, G., and Sempere, T.: Geochemical variations in igneous rocks of the Central Andean orocline (13 S to 18 S): Tracing crustal thickening and magma generation through time and space, Geological Society of America Bulletin, 122, 162–182, https://doi.org/10.1130/B26538.1, 2010.

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

Middlemost, E. A.: Naming materials in the magma/igneous rock system, Earth Sci. Rev., 37, 215–224, https://doi.org/10.106/0012-8252(94)90029-9, http://www.sciencedirect.com/science/article/pii/0012825294900299, 1994.

Nesbitt, H. W. and Young, G. M.: Formation and Diagenesis of Weathering Profiles, The Journal of Geology, 97, 129–147, http://www.jstor.org/stable/30065535, 1989.

Newfoundland and Labrador Geological Survey: Newfoundland and Labrador GeoScience Atlas OnLine, Tech. rep., http://geoatlas.gov.nl.ca[AccesssedJune2016], 2010.

Parker, A.: An Index of Weathering for Silicate Rocks, Geological Magazine, 107, 501–504, https://doi.org/10.1017/S0016756800058581, 1970.

Rasilainen, K., Lahtinen, R., and Bornhorst, T.: The Rock Geochemical Database of Finland Manual. (online), Report of Investigation 164, Geol. Surv. Finland, 2007.

Rybach, L.: Determination of heat production rate, in: Terrestrial Handbook of Heat-Flow Density Determination, edited by Hänel, R., Rybach, L., and Stegena, I., chap. 4.2, pp. 125–142, Kluwer Academic Publishers, Dordrecht, 1988.

Slagstad, T.: Radiogenic heat production of Archean to Permian geological provinces in Norway, Norwegian Journal of Geology, 88, 149–166, 2008.

Strong, D., Turnbull, R., Haubrock, S., and Mortimer, N.: Petlab: New Zealand's national rock catalogue and geoanalytical database, New Zealand J. Geol. Geophys., 53, 475–481, https://doi.org/10.1080/00288306.2016.1157086, 2016.

Turekian, K.: The oceans, streams and atmosphere, in: Handbook of geochemistry, vol. 1, pp. 297–323, Springer-Verlag Berlin, Heidelberg, New York, 1969.

Walker, J. D., Bowers, T. D., Black, R. A., Glazner, A. F., Lang Farmer, G., and Carlson, R. W.: A geochemical database for western North American volcanic and intrusive rocks (NAVDAT), in: Geoinformatics: Data to Knowledge, Geological Society of America, https://doi.org/10.1130/2006.2397(05), 2006.

trace
- **trace_id**
- f_ppm
- cl_ppm
- br_ppm
- i_ppm
- h_ppm
- c_ppm
- n_ppm
- p_ppm
- s_ppm
- al_ppm
- as_ppm
- ag_ppm
- au_ppm
- b_ppm
- ba_ppm
- be_ppm
- bi_ppm
- ca_ppm
- cd_ppm
- ce_ppm
- co_ppm
- cr_ppm
- cs_ppm
- cu_ppm
- dy_ppm
- er_ppm
- eu_ppm
- fe_ppm
- ga_ppm
- gd_ppm
- ge_ppm
- hf_ppm
- hg_ppm
- ho_ppm
- in_ppm
- ir_ppm
- k_ppm
- la_ppm
- li_ppm
- lu_ppm
- mg_ppm
- mn_ppm
- mo_ppm
- na_ppm
- nd_ppm
- ni_ppm
- nb_ppm
- os_ppm
- pa_ppm
- pb_ppm
- pd_ppm
- pm_ppm
- pr_ppm
- pt_ppm
- rb_ppm
- re_ppm
- rh_ppm
- ru_ppm
- sb_ppm
- sc_ppm
- se_ppm
- si_ppm
- sm_ppm
- sn_ppm
- sr_ppm
- ta_ppm
- tb_ppm
- te_ppm
- th_ppm
- ti_ppm
- tl_ppm
- tm_ppm
- w_ppm
- v_ppm
- u_ppm
- y_ppm
- yb_ppm
- zn_ppm
- zr_ppm

major
- **major_id**
- sio2
- tio2
- al2o3
- cr2o3
- fe2o3
- fe2o3_tot
- feo
- feo_tot
- mgo
- cao
- mno
- nio
- k2o
- na2o
- sro
- p2o5
- h2o_plus
- h2o_minus
- h2o_tot
- co2
- so3
- bao
- caco3
- mgco3
- loi

isotope
- **iso_id**
- rb87_sr86
- rb87_sr87
- sr87_sr86
- nd143_nd144
- sm147_nd144
- lu176_hf177
- hf176_hf177
- re187_os186
- re187_os188
- os187_os188
- pb206_pb204
- pb207_pb204
- pb208_pb204
- th232_pb204
- th232_u238
- u238_pb204
- epsilon_hf
- epsilon_nd
- epsilon_sr

country
- **country_id**
- country

sample
- **sample_id**
- age_id
- major_id
- trace_id
- comp_id
- iso_id
- rgroup_id
- ref_id
- method_id
- country_id
- sample_name
- latitude
- longitude
- loc_prec
- qgis_geom
- datum
- depth
- material
- rock_name
- protolith
- sample_description
- density
- comments
- method

computed
- **comp_id**
- norm_factor
- mali
- fe_number
- mg_number
- asi
- maficity
- cia
- wip
- spar
- cai
- ai
- cpa
- qtzindex
- r1
- r2
- tas_name
- qap_name
- sia_scheme
- frost_class1
- frost_class2
- frost_class3
- quartz
- feldspar
- lithics
- facies
- texture
- p_velocity
- density_model
- heat_production_mass
- heat_production

age
- **age_id**
- age_min
- age
- age_max
- age_sd
- time_period_min
- time_period
- time_period_max
- age_method

rockgroup
- **rgroup_id**
- rock_group
- rock_origin
- rock_facies

reference
- **ref_id**
- author
- title
- journal
- year
- doi
- data_source
- bibtex

method
- **method_id**
- method

Figure 1. Database relational structure. Sub-tables are linked through foreign id keys. Ambiguous field names are described in detail in the supplemental material.
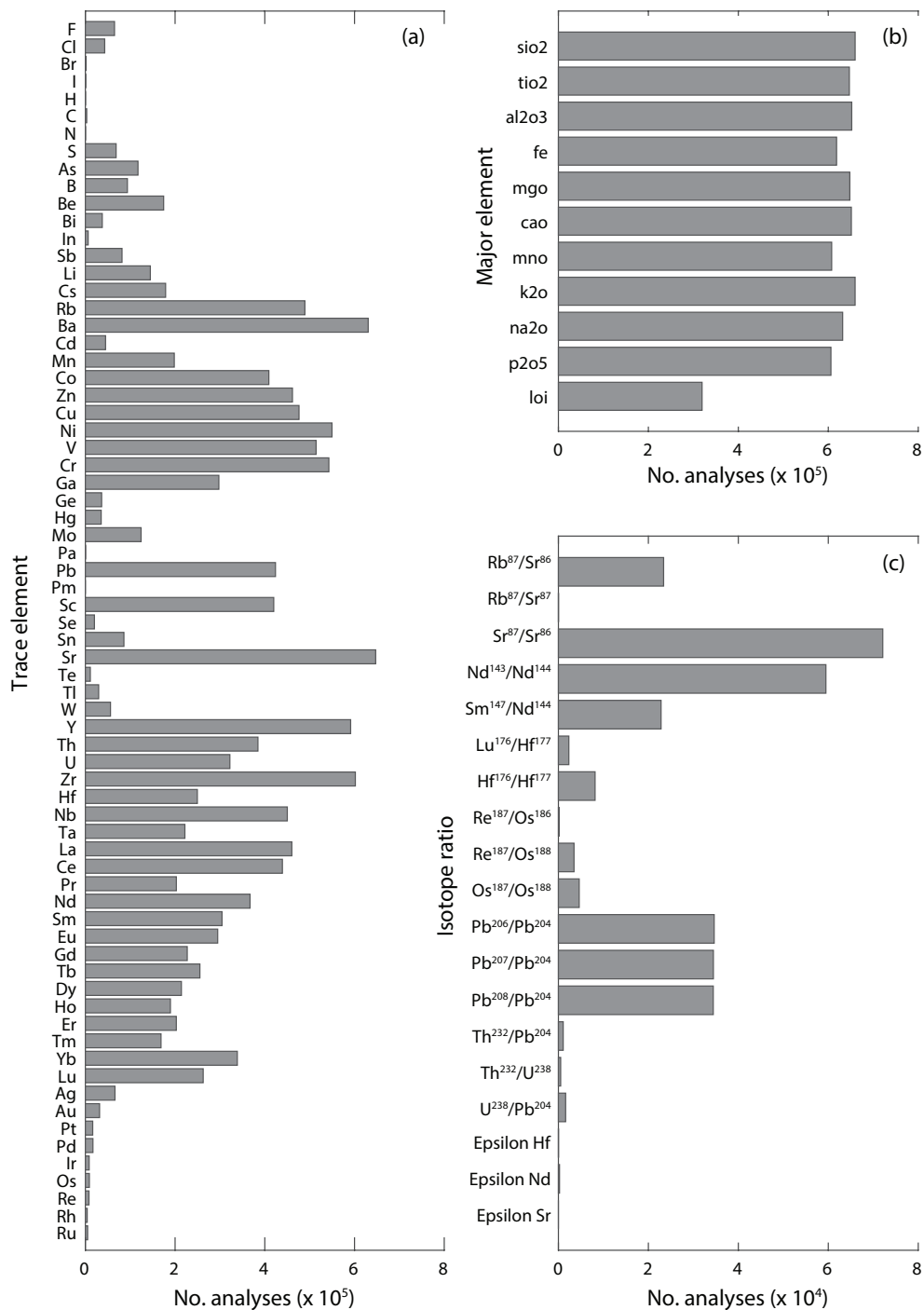
Figure 2. Histogram of analyses. a) Trace elements b) Major oxides. Fe denotes any one or more entries for feo, feo total, fe2o3, or fe2o3 total. c) Isotope ratios and episilon values
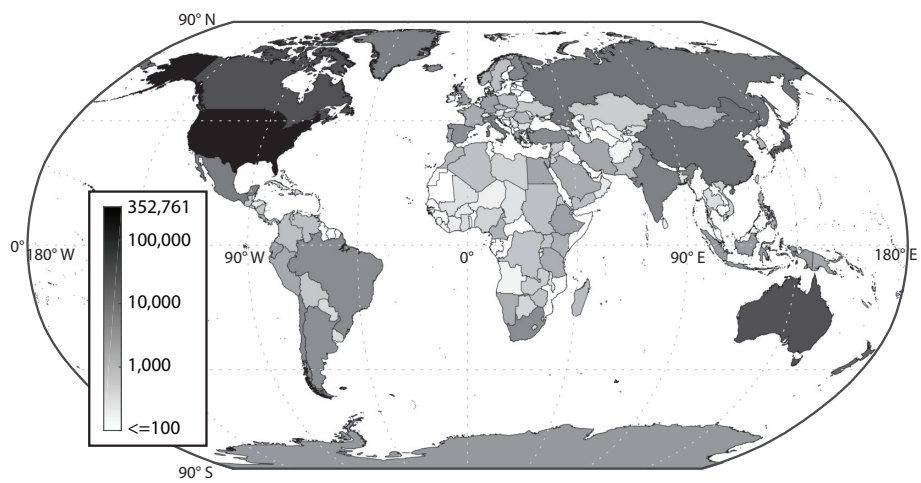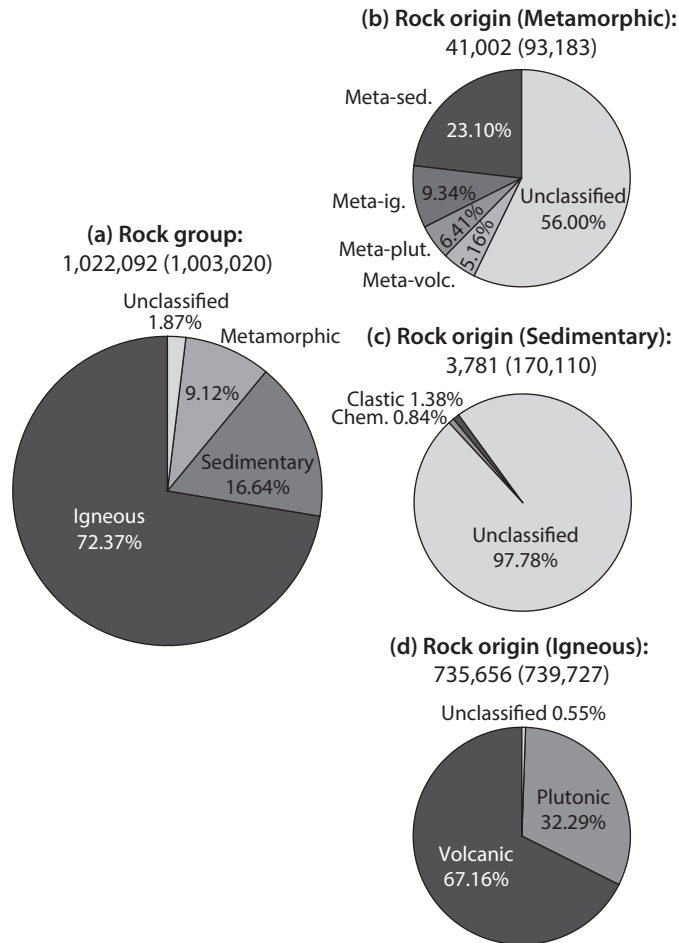
Figure 3. Spatial distribution of geochemical samples. Countries are shaded based on the amount of data points within the polygons.
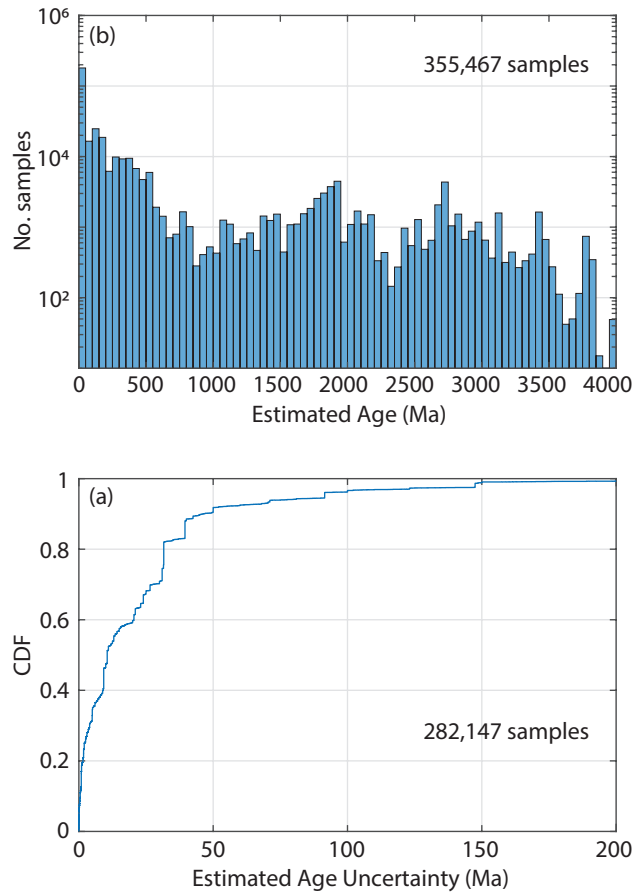
Figure 4. Temporal distribution of geochemical samples. a) Histogram of mean ages in 50 Ma intervals b) Empirical CDF of age uncertainty (major time-period associated ages removed)
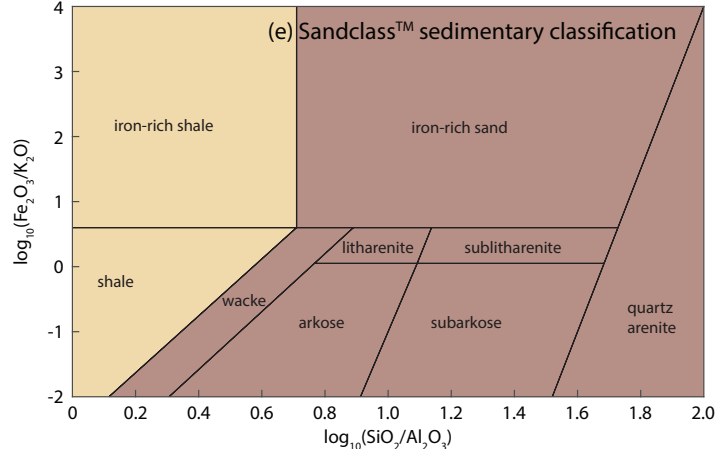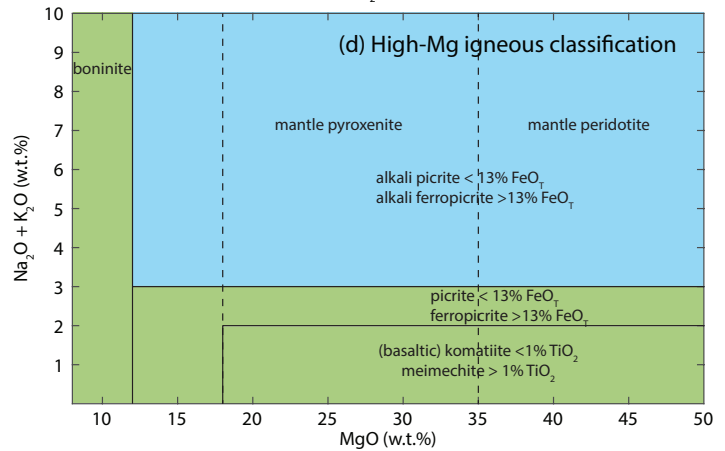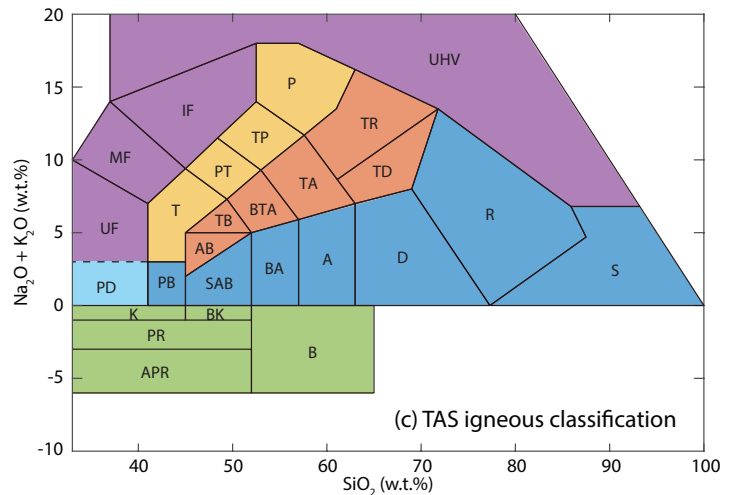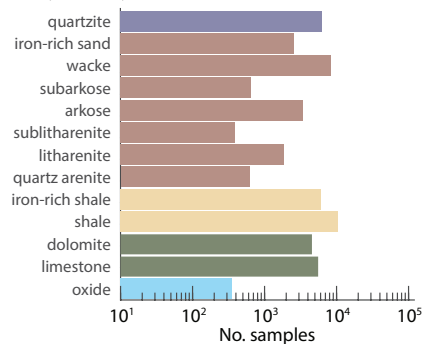
Figure 5. Rock group partitioning. a) Pie chart depicting distribution of samples containing a rock group, b) c) and d) denote the rock origin distributions of the rock group fields where rock origin is listed.
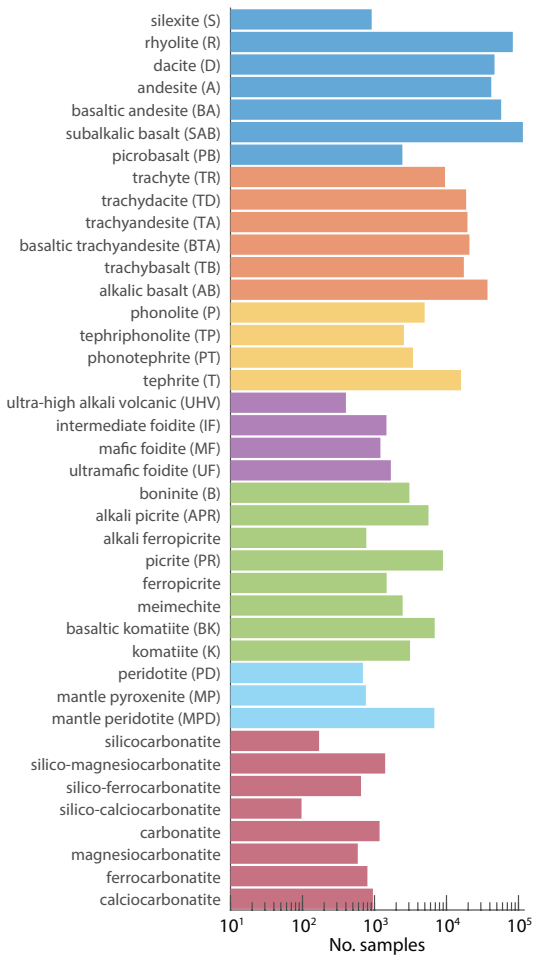
Figure 6. Rock type classification information. a) Igneous and metaigneous sample histograms of assigned rock names b) Sedimentary and metasedimentary sample histograms of assigned rock names c) TAS igneous classification (Middlemost, 1994) d) High-Mg igneous classification. See Le Bas and Streckeisen (1991) for further information on classification methods. e) Sedimentary classification, after Herron (1988) (Sandclass™)
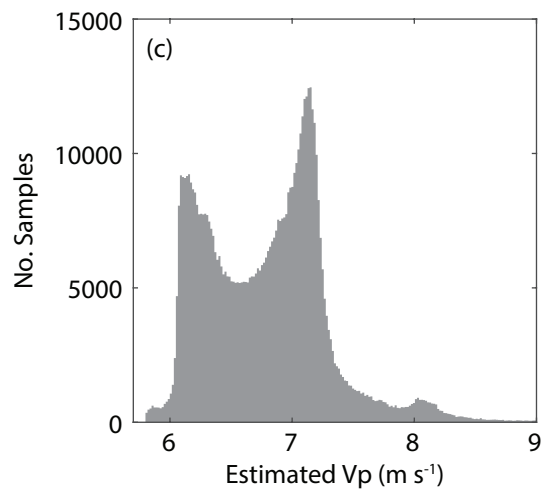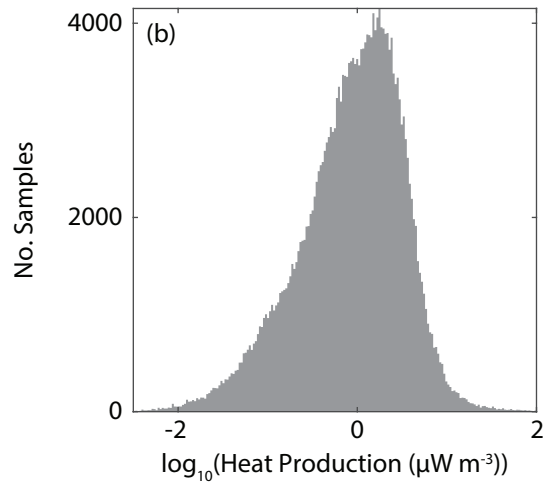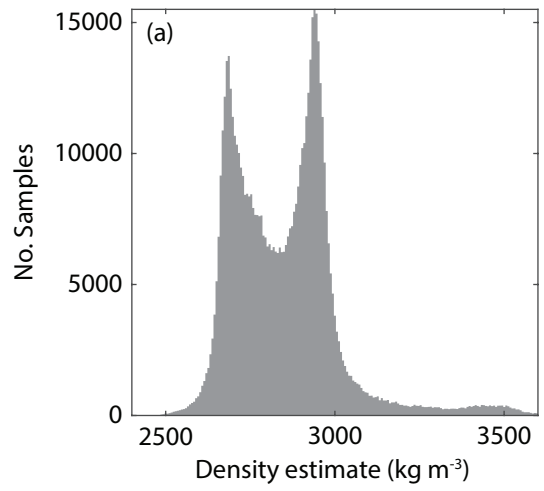
Figure 7. Example computed physical property estimate distributions. a) Density b) Heat Production c) P-wave velocity

Table 1. Brief table content information

Table name	Table description
sample	Lists all samples, where sample_id uniquely describes each row. Contains all foreign keys linking to the other tables. Other information such as coordinates, measured density and depth of sample, analysis method, as well as author prescribed sample descriptions, comments and rock names are also included.
major	Unique major analyses, linked via the key major_id to sample list. Includes major element oxides as well as volatile, carbonate and l.o.i. content where available.
trace	Unique trace element analyses analyses, linked via the key trace_id to sample list.
isotope	Unique isotopic ratio analyses, including epsilon values for Hf, Nd and Sr. Linked via the key iso_id to sample list.
computed	List of physical properties including heat production and density estimates, and classifications and indices based on schemes such as TAS (Total alkali-Silica) and ASI (aluminum-saturation index). Theses values are computed on a major element normalised (LOI free) version of the associated sample's trace and major compositions and may not match the raw values listed. We preserve the raw data in the database, and methods for normalisation and computed properties are included in the appendices if one wishes to recompute these computed properties and indices with different parameters. comp_id uniquely describes each row and is linked to the sample table.
reference	Includes information on the author of the original paper the data was sourced from, and/or reference to database or other previous compilation the data was sourced from e.g. EarthChem. ref_id links the reference table to the sample table.
rockgroup	Uniquely links triplets of rock group, rock origin and rock facies to sample table. For definitions of rock group, origin and facies see Table 3.
age	Uniquely links sets of age and time period information to sample table
country	Unique list of countries (ISO 3166 ALPHA-2 codes) as well as ocean
method	Lists unique method strings detailed in previous publications or databases

Table 2. Data sources

Data source	No. data
EarthChem family (excluding GEOROC) (https://www.earthchem.org/)	380,532
GEOROC (http://georoc.mpch-mainz.gwdg.de)	349,037
OZCHEM (Champion et al., 2016)	65,391
Petlab (Strong et al., 2016)	35,499
Petroch (Haus and Pauk, 2010)	27,388
Newfoundland and Labrador Geoscience Atlas (Newfoundland and Labrador Geological Survey, 2010)	10,073
The British Columbia Rock Geochemical Database (Lett and Ronning, 2005)	8,990
Canadian Database of Geochemical Surveys Open File Reports	8,766
DODAI (Haraguchi et al., 2018)	6,588
Finnish Geochemical Database (Rasilainen et al., 2007)	6,543
Ujarassiorit Mineral Hunt (Geological Survey of Greenland, 2011)	6,078
The Central Andes Geochemical GPS Database (Mamani et al., 2010)	1,970
Geochemical database of the Virunga Volcanic Province (Barette et al., 2017)	908
Other sources (∼1,900 sources, misc. files, see reference csv and bib file)	123,095
Total	1,022,092

Table 3. Potentially ambiguous column information

Column name	Description
sample_name	Author denoted title for the sample. Often non-unique e.g. numbered.
loc_prec	Location precision
qgis_geom	PostGIS ST_Geometry object based on the latitude and longitude of the sample.
material	Material/source of the sample e.g. Auger sample, core, drill chips, xenolith, vein
rock_name	Rock name designated by the original author
sample_description	Sample description mostly inherited from previous databases. Highly variable field.
density	Measured density
comments	Misc. comments, often additional information not included in the sample description field.
method	Method utilized to analyse chemistry and/or age. Variable due to inheritance from previous databases. Multiple methods may be listed, separated by semicolons.
norm_factor	Major element normalisation factor applied to the samples major element chemistry before computing properties
MALI	the modified alkali–lime index (Frost et al., 2001)
fe_number	Iron number (Frost et al., 2001)
mg_number	Magnesium number. Fe2+ estimated using $0.85 \times \mathrm{FeO}^T$.
asi	Alumina Saturation Index (Frost et al., 2001)
maficity	$n_{Fe} + n_{Mg} + n_{Ti}$
cia	Chemical index of alteration (Nesbitt and Young, 1989). Generally CaO* includes an additional correction for CO_2 in silicates, but CO_2 is not reported for a large fraction of the data set so we do not include this term for consistency.
wip	Weathering Index of Parker (Parker, 1970)
spar	Modified from (Debon and Le Fort, 1983) to remove apatite
cai	Calcic-alkalic index (Frost et al., 2001)
ai	alkalic index (Frost et al., 2001)
cpa	Chemical proxy of alteration (Buggle et al., 2011)
qtzindex	(Debon and Le Fort, 1983)
r1	R1R2 chemical variation diagram (la Roche et al., 1980)

r2	R1R2 chemical variation diagram (la Roche et al., 1980)
rock_type	compositionally based rock names, discussed in Section 4.2, following similar methods of Hasterok et al. (2018)
sia_scheme	S-, I-, and A-type granite classification. For felsic compositions, A- and I-types are not properly discriminated with this method. (Frost et al., 2001)
frost_class1	Magnesian or Ferroan (Frost et al., 2001)
frost_class2	Calcic, calc-alkalic, alkali-calcic, alkalic(Frost et al., 2001)
frost_class3	Metaluminous, peraluminous, peralkaline (Frost et al., 2001)
quartz	Estimate of quartz content from major element analyses. SiO_2/M_{SiO_2} where M_X is the molecular weight of the oxide X (Mason, 1952; Turekian, 1969)
feldspar	Estimate of feldspar/clay/Fe-Al oxide content from major element analyses. $Al_2O_3/M_{Al_2O_3} + Fe_2O_3(t)/M_{Fe_2O_3}$ where M_X is the molecular weight of the oxide X (Mason, 1952; Turekian, 1969)
lithics	Estimate of lithics (carbonate) content from major element analyses. $MgO/M_{MgO} + CaO/M_{CaO}$ where M_X is the molecular weight of the oxide X (Mason, 1952; Turekian, 1969)
facies	metamorphic facies information pulled from rock_name via partial string search
texture	metamorphic texture information pulled from rock_name via partial string search
p_velocity	To estimate seismic velocity we use an empirical model developed by Behn and Kelemen (2003), and utilized in Hasterok and Webb (2017). We use the compositional model $V_p(km/s) = 6.9 - 0.011C_{SiO_2} + 0.037C_{MgO} + 0.045C_{CaO}$ where the concentration of each oxide is in wt.%.
density_model	We utilize the multiple density estimate methods as outlined by Hasterok et al. (2018) for each compositional group, using multiple linear regression on the data set
heat_production_mass	Determined from the chemical composition with the relationship $HP_{mass} = 10^{-5}(9.67C_U + 2.56C_{Th} + 2.89K_2O)$ where C are the concentrations of the HPEs in ppm except K_2O in wt.% (Rybach, 1988)

heat_production	Heat production mass multiplied by the density estimate (in kg m^{-3}) (Rybach, 1988)
age_ or time_period_ min	Minimum crystallisation age estimate
age or time_period	Mean crystallisation age estimate
age_ or time_period_ max	Maximum crystallisation age estimate
age_sd	Age uncertainty
age_method	Method of age estimation, variable due to inheritance from previous databases
rock_group	The highest order rock type classifications: Igneous/metamorphic/sedimentary
rock_origin	Second order classifications of the rock groups - e.g. plutonic/volcanic, metaplutonic/metaigneous/metased, clastic/chemical
rock_facies	Third order classifications, mostly restricted to metamorphic rock facies e.g. granulite
data_source	Field reserved for existing database compilation e.g. if a sample is derived from EarthChem
bibtex	bibtex key corresponding to further reference information if it exists, contained in the attached bib file for easier citation