Dear Editor,

We would like to thank the Editor for giving us the opportunity to respond to the reviewer comments, and the reviewers for their constructive feedback. Concerns and points highlighted by the reviewers are addressed in the revised manuscript. The majority of the edit is incorporated in the addition of an age dating method field, and further discussion about computed properties in Section 4.2. Minor rewrites to tidy sentence structure were also scattered throughout the manuscript.

Kind regards,

Matthew Gard, Derrick Hasterok, Jacqueline A. Halpin

Reviewer 1: Kent Condie

"Excellent contribution to the earth science community, should be widely used. A couple of suggestions to improve the paper: 1) p4, line 10. Indicate isotopic dating method (U/Pb zircon, Sm/Nd garnet, Rb/Sr whole rock, 40Ar/39Ar mica etc) 2) Add uncertainties to isotopic ages 3) Do you include detrital zircon ages? If so they should be reported separately from igneous zircon ages."

Dating method has been parsed out as its own field now (age_method) where available (Figure 1 and Table 3). If a method is supplied, the range of age dates (age_sd, or age_max and age_min range) can be taken to the uncertainty in isotopic age. We do not include detrital zircon ages.

"Data availability Investigators should be able to download specific parts of the geochemical database, by sorting before download (such as mafic igneous rocks, detrital sediments, tonalities > 2.5 Ga etc.)"

While we understand the Reviewer's desire to have pre-sorted datasets, we philosophically disagree on this point. It is partly this pre-sorting that has led us to design the database in the manner we have. Another reason opted for a single database is we found it annoying to have to download pre-existing databases through web forms in parts because of download limitations. Our dataset is large, but not overly so and smaller than many global datasets. First, pre-sorting requires that several files will be needed to be maintained during subsequent updates. Where does one end with pre-sorting? There are a number of geologically interesting datasets that one might consider e.g. TTGs, komatiites, kimberlites, plume-related magmatism, etc. However, as our understanding of Earth processes grows, our definitions of some of these interesting rocks has evolved. Not to mention the debate surrounding some of these definitions at present. We prefer the database to remain agnostic in this regard. One last point here is that we have developed a set of codes (in Matlab) to parse, filter, plot and run basic computations on the database that is available through github (repository address is included in the text). Not all the codes are available yet because they have not been fully documented, but a complete set of codes is forthcoming. In the future, we may also provide similar codes in Python and/or R.

Reviewer 2 - Anonymous Reviewer

"Gard and coauthors present a curated database of (primarily) whole-rock elemental and isotopic analyses. This work fills a useful niche between domain-specific manual compilations and large but less-curated online repositories such as EarthChem. I particularly applaud the authors for ensuring that it is relatively easy to download the full dataset and bibliography in open formats – in this case, csv and bib. There is one minor issue here though that I would request the authors consider addressing: right now, it would not be easy for a user without significant database experience to figure out, e.g., which age corresponds with which sample metadata, or elemental composition, or so on, for any of the ten individual tables provided. The simplest way to address this would be to provide a flat csv of the entire dataset. While this would weigh in at perhaps 1 GB, it would be sparse and highly amenable to compression (e.g. gzip, for a standard and open option). A similar dataset I have worked with compresses to a relatively manageable 160 MB when treated in such a manner."

A complete merged csv is now included and compressed, and does indeed total around 180 MB. Included in the github codes link is also a script to merge desired subtables into a singular table.

"Finally (though I suspect this may have already been done) since it is not immediately clear from the text, I would echo the request from Prof. Condie's review, to indicate the isotopic dating method and (critically!) uncertainty for samples with newly-attributed ages."

As addressed in Condie's response, we now have a column age_method which contains this information.

Reviewer 3 - Juan Carlos Afonso

"This compilation and associated database make up a fantastic contribution to the geo- science community. It should be useful for a wide range of researchers. The manuscript is well-written and easy to follow. The csv files are clear and easy to download and manipulate. Besides the comments from the other reviewers, the only shortcoming I found at the moment is the lack of interrogation and/or manipulation tools. The authors clearly state that they are creating such tools in matlab, which is terrific, but I would have loved to have at least some basic interrogation codes with this publication! Maybe something the authors can work on for a final version?"

A number of codes are now available on github (link within text) for managing, filtering and working with the data.

"Another comment is about the computed properties (Vp, RHP, density). The authors refer to other works in the text for the methods, which is fine, and then include some equations in Table 3. Can the authors say anything about the uncertainties associated with these estimates? or even better, provide any sort of validation of the predictions against real measurements? I guess that at least some of the samples that made it into the database/s have been characterized well enough to include measured density and perhaps ultrasonic measurements of Vp (?). Such a validation would be great for us readers/users. Perhaps this has been done in the cited works, and if so, all good. I'd then just mention it in the manuscript and give a brief summary to help the reader. Overall, a really nice contribution. Well done and thanks for your efforts!"

The computed properties sections was a little sparse on details of uncertainty and discussion of validity in the original manuscript. These details are available in the cited works as you suggested, but are now also with a brief description in this manuscript in Section 4.2.

We would like to thank all reviewers for their constructive suggestions which aided in improving the manuscript and database, and their positive comments towards our work.

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. ~~Dissemination and collation~~ 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 ~~assembled~~ compiled a global whole-rock geochemical database, ~~with other associated sample information and properties,~~ sourced from various existing databases and supplemented with ~~numerous individual publicationsand corrections~~an extensive list of individual publications. Currently the database stands at 1,~~023,490~~ 022,092 samples with varying amounts of associated ~~information~~ sample data including major and trace element concentrations, isotopic ratios, and location ~~data. The distribution both spatially and temporally is quite~~ information. Spatial and temporal distribution is heterogeneous, however temporal distributions are enhanced over some previous database compilations, particularly in ~~terms of~~ ages older than ~1000 Ma. Also included are a ~~wide range of computed~~ range of geochemical indices, various naming schema and physical property estimates ~~and naming schema~~ 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~~datasets, in particular those wishing to investigate secular temporal trends. The addition of physical properties, estimated ~~by~~ from sample chemistry, ~~represents~~ represent a unique contribution to otherwise similar geochemical databases. The data is 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. ~~This~~ The data set will continue to grow and be improved, and we encourage readers to contact us or other database compilations ~~contained~~ 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.~~2592823~~ .2592822 (**?**).

1 Introduction

Geochemical analyses in conjunction with other temporal, spatial, and physical property ~~data from rock samples~~ information have been vital sources of information for understanding the Earth and investigating both local, and global geodynamic histories (e.g. **?**). 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~~dramatically, often~~ , dispersed among many hundreds of individual publications. Since the 1960~~'~~'s and 70~~'~~'s, broad element suites have been ~~rapidly~~ 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 ~~have swiftly expanded~~ are swiftly expanding our cumulative global data records. However, due to the rate of new publications, in conjunction with significant partitioning ~~of this data~~ between different journals, ~~countries, authors etc.~~ 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. **?**), proposing changes in mantle dynamics (e.g. **?**), to look at regional and global tectonic histories (e.g. **?**), and examine the connections between life and the solid Earth (e.g. **?**). Not only does this information have implications for the scientific community, but also for issues such as environmental management, land use, and ~~minerals~~ 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 ~~data~~ 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 ~~'go-to'~~ general use geochemical data repository~~for general use~~. It consists of many federated databases such as PetDB, NAVDAT, SedDB, the USGS National Geochemical Database and GEOROC, 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 (**?**). (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).

- 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 ~~used in this study~~ utilised in this paper (http://metpetdb.com/).

- 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 (**?**), New Zealand's national rock database (Petlab) (**?**), Australia's national whole-rock geochemical database (OZCHEM) (**?**), the Finnish lithogeochemical rock geochemistry database (RGDB) (**?**), the Newfoundland and Labrador Geoscience Atlas (**?**), and the basement rock geochemical database of Japanese islands (DODAI) (**?**).

While all of these are generally exceptional enterprises, we personally ~~ran into issueswithin our own research~~met with issues. Some examples included databases being deficient in aged data (1000 Ma+), or lacking many recent publications. ~~Issues with existing data within these~~ Some issues in certain existing databases was 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 ~~seek~~ sought to produce a ~~global whole-rock geochemical database incorporating~~ database incorporating refined samples from previous databases~~as necessary~~, and supplementing significantly from other, often recent, publications. Computed properties, naming schemes, and various geochemical indices ~~have~~ has also been calculated where the data permits. Smaller subsets of previous ~~versions~~ of this database have already been utilised for studies of heat production and phosphorus content (**?????**), 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 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 utilise the information contained within. We ~~personally~~ had issues when seeking estimated or measured ages of rock samples. For studies which examine temporal variations of chemistry or physical properties an accurate and precise age is required. Under some of the present data management schemes for ~~online~~ databases it may be difficult to recover the desired data. Crystallization

dates 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 ages may be contained within the database, and stored under mineral analyses. However, a search for rock chemistry may return an age (often a geologic timescale division), but not a precise date. To get the date one would have to also download the individual mineral analyses, conduct an age 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. Had the estimated crystallization date been attributed to the sample directly as often reported in the original study, much of this process could be short cut. ~~Our database seeks to do just that, by attributing an estimated crystallization age~~ Instead, our database attributes these estimated crystallization ages to the sample as provided in the original reference~~at the point of data entry. This also~~ . This allows us to include estimated dates for the same unit or formation. As a result the database presented here allows for a higher density of temporal sampling than other compilations.

~~We have chosen a~~ The database is provided in two formats; the first as a compressed single spreadsheet for people unfamiliar with database management systems, and the second as a mixed flat file and relational database structure~~for simpler distribution~~.

? 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.

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

~~We utilise PostgreSQL~~ PostgreSQL was utilised as the relational database management system (RDMS) to update and administer the database. PostgreSQL ~~containins~~ contains many built in features and useful addons including the geospatial database extender PostGIS which we utilise, 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 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 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 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 200 Mb, which we believe to be an acceptable size for data manipulation.

Many samples include multiple 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 ~~online~~in the literature, as well as ~~clean up of entries previously entered into~~ cleaned up and modified entries previously from existing databases. The subsets of ~~previous data sets~~ 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~~ or encountered issues with accessing the entire ~~data sets online~~ 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. ~~Major~~ Unsurprisingly major element analyses in general dwarf the ~~amount~~ number of trace element measurements recorded~~in terms of consistency which is unsurprising~~.

~~The samples are distributed reasonably well around the globe~~Despite the heterogeneous nature of geochemical sampling, there is still reasonable spatial coverage around the world. However, there ~~is~~ are a noticeable dominance of samples sourced from North America, and ~~in a smaller way from~~ additionally Canada, Australia, and New Zealand (Figure 3). The United States tops of the list with ~~335,266~~ 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 ~~here is indicated as being an assumed crystallization age. Age~~ distributions unsurprisingly show a significant dominance ~~for~~ 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 ~~361,815~~ 355,467 samples with mean crystallisation age values. Of these, 282,~~375~~ 147 have age uncertainty estimates and observing the cumulative distribution function of these values indicates that $\sim 99\%$ of the age uncertainties fall below $\sim$150 Ma (Figure 4a).

Rock group and rock origin are described in Table 3. There is a clear dominance towards igneous samples, making up ~~73.80~~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 $\sim$ ~~43~~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.

4.1.1 ~~Naming schema - rock_type~~

type]~~Computed propertiesWe compute a number of properties and naming schema for a significant subset of the database, a new addition over many previous 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 relative proportions of each individual elements contribution to the total. This normalisation occurs only on samples with major element totals between 85 and 120 . 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~~Naming schema - rock_~~factor field.~~

4.1.1 ~~Naming schema - rock_type~~

type 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 utilised 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.

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 (**?**) modified to include additional fields for further classification of high-Mg volcanics (**?**). 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 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 $CaO + MgO$ (**??**). Additionally, we further partition clasic sediments using the SedClass™classification method from **?**. Quartzites are identified separately where SiO_2 exceeds 0.9 in the ternary system. See **?** for further discussion.

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, unsurprisingly due to the extent of samples of oceanic nature.

4.1.1 ~~Computed physical properties~~

4.2 Computed properties

~~Physical properties for rock types used in numerical modelsare often based on averages based on limited samples from individual publications. This database provides an opportunity for in-depth analysis of physical properties of rock types with specified chemistry that can be used to improve geodynamic models.~~ In numerical models, rock types are often assigned physical property estimates that have been derived from limited datasets. We compute a number of properties and naming schema for a significant subset of the database, a new addition over many previous 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 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. ~~Estimates of density and p-wave velocities are based on the methods contained within **?**, albeit with a larger data set provided here. By utilising the density estimate we can also compute heat production estimates by employing the relationship from **?**. The multi-modal nature of the density and p-wave velocity estimates are driven largely by the dichotomy between mafic and felsic samples within the database. Heat production estimates however are resolved by a smoother distribution in log-space~~

4.2.1 Density estimates

Density is an important input for a wide range of models but only a small fraction of samples often have measured density values associated with them. Contained within the database are a number of publications hosting density observations (e.g. **???**) . Following the method of **?**, we produce a set of simple oxide-based linear regression density models.

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

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

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

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 **?** for further discussion of the model fits.

Density estimates peak at ~2680 and ~2946 kg m^{-3} due to mafic and felsic sample medians respectively ~~, and~~

4.2.2 Seismic velocity - Vp

We utilise the empirical model of ? for estimating anhydrous p-wave seismic velocity. Their model was calibrated on ~ 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. Utilising their 3 oxide model, estimated uncertainty (1σ) is $\sim \pm 0.13 \mathrm{kms}^{-1}$. P-wave velocity estimates depict maximums at ~~6.183 and 7.135~~$\sim$6.2 and $\sim$7.1 $\mathrm{m\,s}^{-1}$ (Figure 7c). For further details or discussion, refer to ? and ?.

$$Vp = 6.9 - 0.011 \times \mathrm{SiO_2} + 0.037 \times \mathrm{MgO} + 0.045 \times \mathrm{CaO},$$

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

4.2.3 Heat Production and heat production mass

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

$$A(\mu W m^{-3}) = \rho \times (9.67 \times \mathrm{U} + 2.56 \times \mathrm{Th} + 2.89 \times \mathrm{K_2O}) \times 10^{-5},$$

with concentrations of U, Th in ppm, K_2O in weight-percent and ρ in $\mathrm{kg\ m}^{-3}$. Heat production has a median value of ~~1.009~~1.0 $\mu\mathrm{W\,m}^{-3}$, with first and third quartiles (25th and 75th percentiles) of ~~0.3886 and 2.199~~0.39 and 2.2 $\mu\mathrm{W\,m}^{-3}$ 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 utilising this database additionally cite the original authors and data sources.

6 ~~Bibliography information file~~

~~Where DOI exists or where we have manually cleaned up individual publications, we have attached a bibtex file of the entries, containing further information over the reference .csv file. We hope to expand this .bib file as we continue to clean up the reference lists and make adjustments to other compilations. Many do not have this information however as we have inherited many database reference lists, and for those which don't, the information required to find the sample are included in reference table to the best of our ability.~~

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 this 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.

Utilising this database we ~~are working~~ have worked on methods for predicting protoliths of metamorphic rocks, as over 57% of the samples lack that information (Figure 5) and may be included in the future (**?**). We are also making progress on a geologic provinces map that captures tectonic terranes. ~~These projects are being done separately, but may be utilised in conjunction with this database in future updates.~~ 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 subtable ~~will be released some time in the future~~is also available on github at https://github.com/dhasterok/global_geochemistry.

7 Data availability

The BIB file and CSV tables of this dataset are available on Zenodo: https://doi.org/10.5281/zenodo.2592822 (**?**)

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.

Acknowledgements. We thank Bärbel Sarbas for supplying the GEOROC database in its entirety. We would also like to thank the following individuals for providing data sets and/or personal compilations: D. Champion (GA) D. Claeson (SGU), T. Slagstad (NGU), Lorella Francalanci (UNIFI), Yuri Martynov (FEGI-RAS), Takeshi Hanyu (JAMSTEC), J. Clemens (SUN), H. Furness (UIB), A. Burton-Johnson (BAS) and M. Elburg (UJ). Peter Johnson provided a collection of papers with data for the Arabian-Nubian Shield. M. Gard is supported by Australian Government Research Training Program Scholarship. This research was supported partially by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP180104074). J. A. Halpin was supported under Australian Research Council's Special Research Initiative for Antarctic Gateway Partnership SR140300001. The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council.

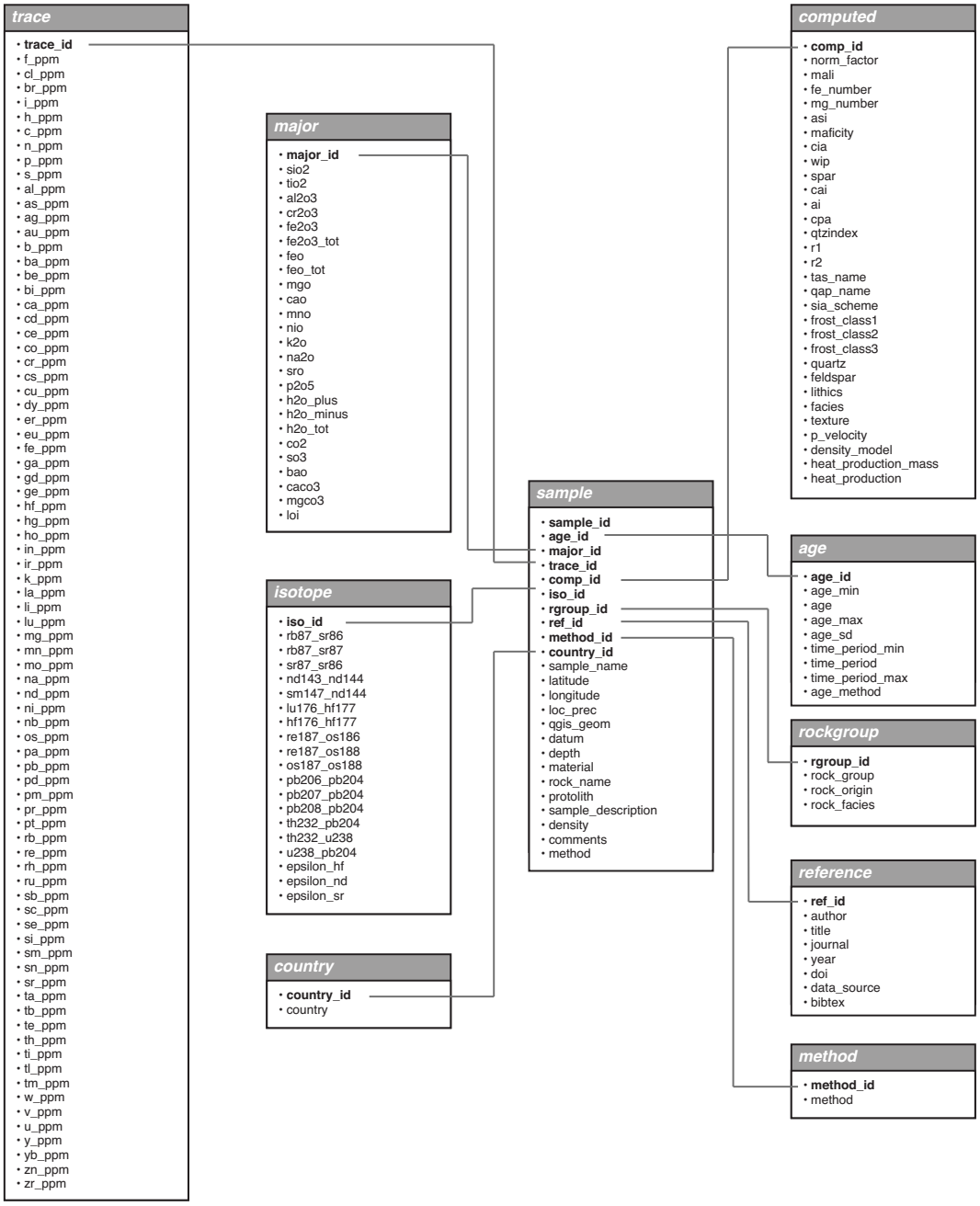

Figure 1. Database relational structure. Sub-tables are linked through foreign id keys. Ambiguous field names are described in detail in the supplemental material.

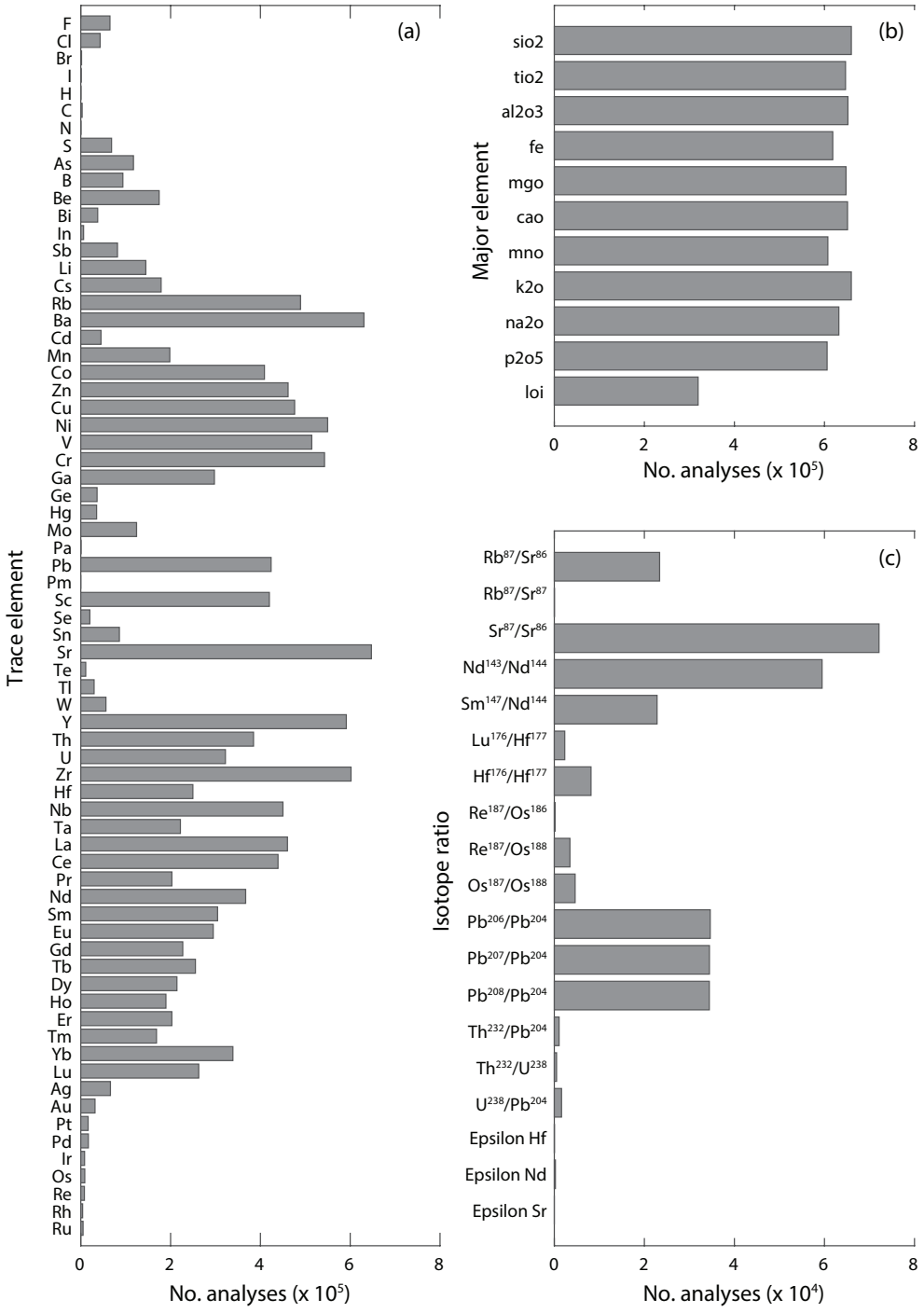

Figure 2. Histogram of analyses. a) Trace elements b) Major oxides. Fe denotes any one or more entries for feo, feo total, fe2o3, or fe2o3 total. c) Isotope ratios and episilon values

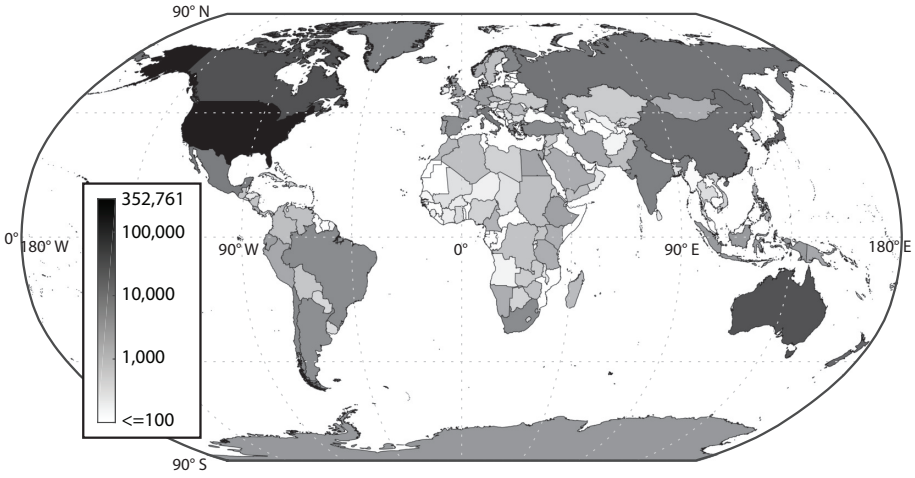

Figure 3. Spatial distribution of geochemical samples. Countries are shaded based on the amount of data points within the polygons.

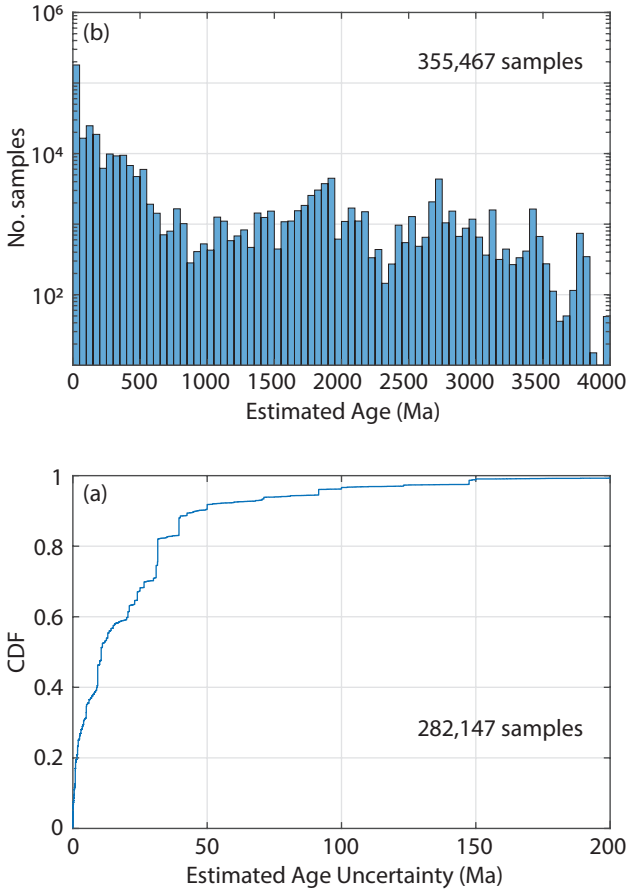

Figure 4. Temporal distribution of geochemical samples. a) Histogram of mean ages in 50 Ma intervals b) Empirical CDF of age uncertainty (major time-period associated ages removed)

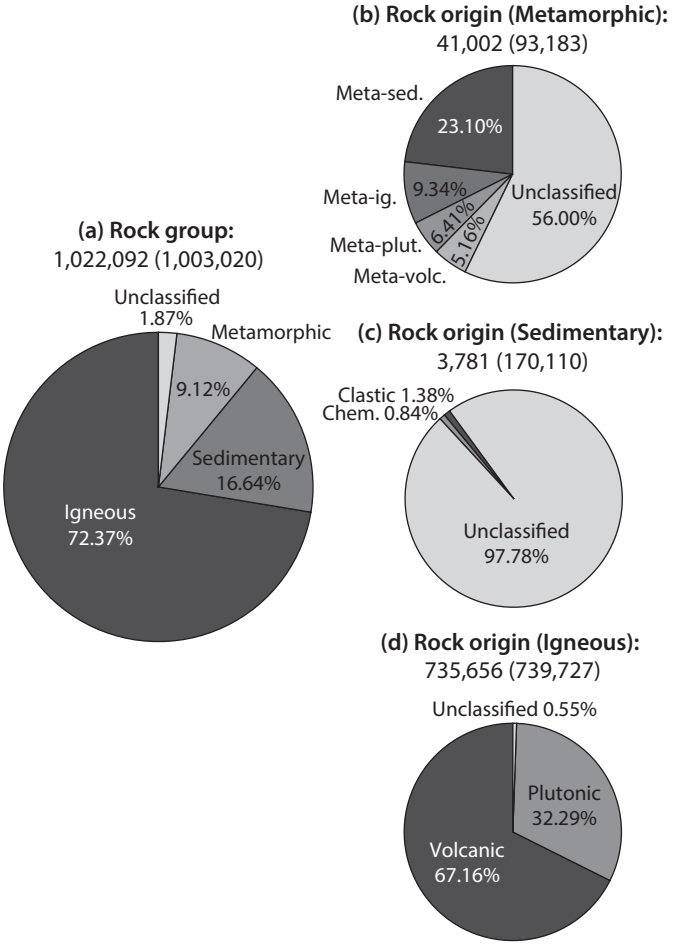

Figure 5. Rock group partitioning. a) Pie chart depicting distribution of samples containing a rock group, b) c) and d) denote the rock origin distributions of the rock group fields where rock origin is listed.

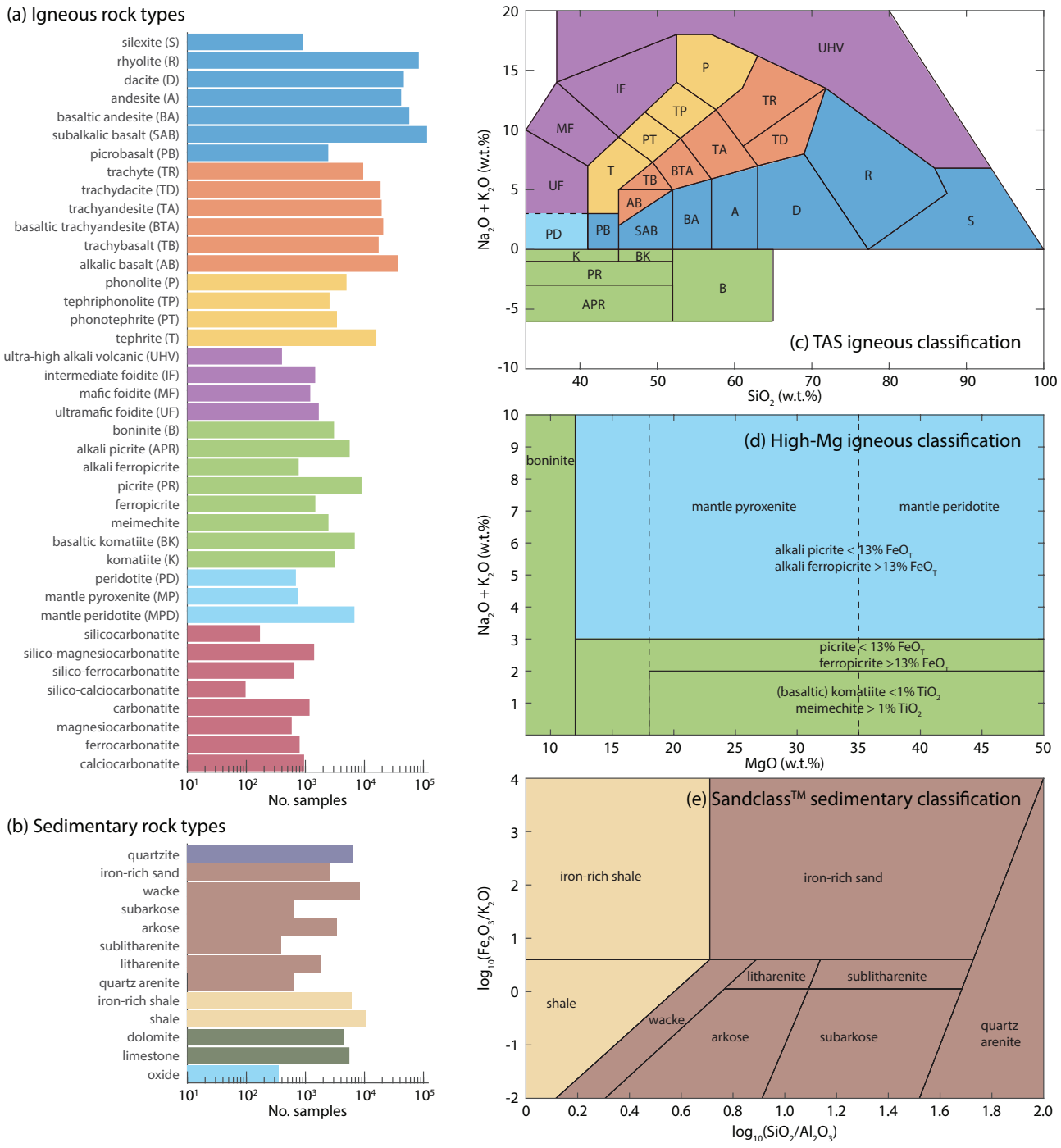

Figure 6. Rock type classification information. a) Igneous and metaigneous sample histograms of assigned rock names b) Sedimentary and metasedimentary sample histograms of assigned rock names c) TAS igneous classification (**?**) d) High-Mg igneous classification. See **?** for further information on classification methods. e) Sedimentary classification, after **?** (Sandclass™)

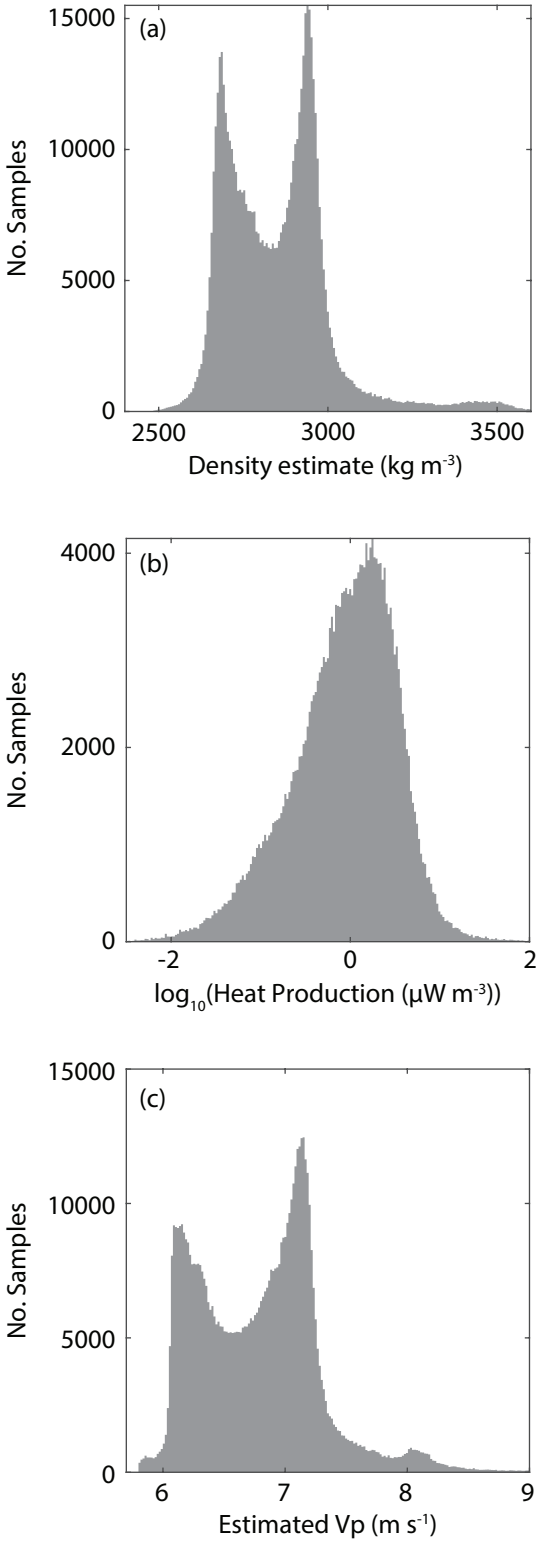

Figure 7. Example computed physical property estimate distributions. a) Density b) Heat Production c) P-wave velocity

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,~~620~~ 532
GEOROC (http://georoc.mpch-mainz.gwdg.de)	~~351,171~~ 349,037
OZCHEM (**?**)	~~64,462~~ 65,391
Petlab (**?**)	35,~~950~~ 499
Petroch (**?**)	27,388
Newfoundland and Labrador Geoscience Atlas (**?**)	10,073
The British Columbia Rock Geochemical Database (**?**)	8,990
Canadian Database of Geochemical Surveys Open File Reports	8,766
DODAI (**?**)	6,~~701~~ 588
Finnish Geochemical Database (**?**)	6,543
Ujarassiorit Mineral Hunt (**?**)	6,078
The Central Andes Geochemical GPS Database (**?**)	1,970
Geochemical database of the Virunga Volcanic Province (**?**)	908
Other sources (∼1,900 sources, misc. files, see reference csv and bib file)	~~113,870~~ 123,095
Total	1,~~023,490~~ 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 utilised 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 (**?**)
fe_number	Iron number (**?**)
mg_number	Magnesium number. Fe2+ estimated using $0.85 \times \mathrm{FeO}^T$.
asi	Alumina Saturation Index (**?**)
maficity	$n_{Fe} + n_{Mg} + n_{Ti}$
cia	Chemical index of alteration (**?**). Generally CaO* includes an additional correction for CO_2 in silicates, but CO_2 is not reported for a large fraction of the dataset so we do not include this term for consistency.
wip	Weathering Index of Parker (**?**)
spar	Modified from (**?**) to remove apatite
cai	Calcic-alkalic index (**?**)
ai	alkalic index (**?**)
cpa	Chemical proxy of alteration (**?**)
qtzindex	(**?**)
r1	R1R2 chemical variation diagram (**?**)
r2	R1R2 chemical variation diagram (**?**)

rock_type	compositionally based rock names, discussed in Section 4.1.1, following similar methods of **?**
sia_scheme	S-, I-, and A-type granite classification. For felsic compositions, A- and I-types are not properly discriminated with this method. (**?**)
frost_class1	Magnesian or Ferroan (**?**)
frost_class2	Calcic, calc-alkalic, alkali-calcic, alkalic(**?**)
frost_class3	Metaluminous, peraluminous, peralkaline (**?**)
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 (**??**)
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 (**??**)
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 (**??**)
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 **?**, and utilised in **?**. 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 $\mathrm{wt.}\%$.
density_model	We utilise the multiple density estimate methods as outlined by **?** 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 $\mathrm{wt.}\%$ (**?**)
heat_production	Heat production mass multiplied by the density estimate (in kg m^{-3}) (**?**)

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