Check for updates

Received: 31 May 2021 | Accepted: 5 June 2021

DOI: 10.1111/gbi.12462

PERSPECTIVE

gebiology WILEY

The Sedimentary Geochemistry and Paleoenvironments Project

1 | INTRODUCTION

Geobiology explores how Earth's system has changed over the course of geologic history and how living organisms on this planet are impacted by or are indeed causing these changes. For decades, geologists, paleontologists, and geochemists have generated data to investigate these topics. Foundational efforts in sedimentary geochemistry utilized spreadsheets for data storage and analysis, suitable for several thousand samples, but not practical or scalable for larger, more complex datasets. As results have accumulated, researchers have increasingly gravitated toward larger compilations and statistical tools. New data frameworks have become necessary to handle larger sample sets and encourage more sophisticated or even standardized statistical analyses.

In this paper, we describe the Sedimentary Geochemistry and Paleoenvironments Project (SGP; Figure 1), which is an open, community-oriented, database-driven research consortium. The goals of SGP are to (1) create a relational database tailored to the needs of the deep-time (millions to billions of years) sedimentary geochemical research community, including assembling and curating published and associated unpublished data; (2) create a website where data can be retrieved in a flexible way; and (3) build a collaborative consortium where researchers are incentivized to contribute data by giving them priority access and the opportunity to work on exciting questions in group papers. Finally, and more idealistically, the goal was to establish a culture of modern data management and data analysis in sedimentary geochemistry. Relative to many other fields, the main emphasis in our field has been on *instrument measurement* of sedimentary geochemical data rather than *data analysis* (compared with fields like ecology, for instance, where the post-experiment ANOVA (analysis of variance) is customary). Thus, the longer-term goal was to build a collaborative environment where geobiologists and geologists can work and learn together to assess changes in geochemical signatures through Earth history.

With respect to the data product, SGP is focused on assembling a well-vetted and comprehensive dataset that is tractable to multivariate statistical analyses accounting for multiple geological and methodological biases. Phase 1 of the project, which focused on the Neoproterozoic and Paleozoic, has been completed. Future phases will capture a broader range of geologic time, data types, and geography. The database contains tens of thousands of unpublished data points provided by consortium members, as well as detailed metadata that go beyond what is contained in papers. In many cases, these represent measurements that are tangential to a given published study but still of high utility to database studies; these allow the community to address questions that would be impossible to answer solely with the published data. For instance, in order to use a proxy such as Mo/TOC (total organic carbon) ratios in mudrocks deposited under a euxinic water column, the full suite of trace metal, iron speciation, and total organic carbon data is needed. Likewise, geospatial information is required to account for sampling biases, and many statistical learning approaches cannot accept, or have difficulty with, incomplete geological predictor variables. Ultimately, it is this complete data matrix that will allow for SGP's most insightful analyses.

This paper serves as an introduction to SGP, the process by which our data products are created, a description of the Phase 1 data product and a citable reference for that product, a description of the SGP website and API (Application Programming Interface) for open access, and a statement of our future goals.

2 | WHY SGP?

In recent years, there has been a welcome trend in the broader geochemical community toward increased data accessibility, documentation of sample context, and sample curation, albeit with challenges still ahead (Brantley et al., 2020; Cutcher-Gershenfeld et al., 2016; Planavsky et al., 2020). First, progress has been made through journals and organizations adopting stringent data archiving rules and promoting adherence to FAIR principles—findability, accessibility, interoperability, and reusability ("FAIR Play in Geoscience Data," 2019; Wilkinson et al., 2016). Second, several databases now house geochemical data at different scales and with different focuses (Brantley et al., 2020; Gard et al., 2019; He et al., 2019; Lehnert et al., 2000). Among the largest and most active are projects such as EarthChem (earthchem.org), the Geobiodiversity Database (geobiodiversity.com), Pangaea (https://www.pangaea.de), and the StabisoDB (https://cnidaria.nat.uni-erlangen.de/stabisodb/). The SGP database was built with the data structures and standards of these

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2021 The Authors. *Geobiology* published by John Wiley & Sons Ltd.

FIGURE 1 The Sedimentary Geochemistry and Paleoenvironments Project (SGP) is an open, collaborative consortium focused on understanding how the Earth has changed through time through analyses of large sedimentary geochemical datasets

other projects in mind, in keeping with FAIR principles and with the hope that data can be easily shared in the future. Consistent with the stance taken by other organizations in the community (Hanson, 2016), we also strongly encourage all members to register their samples for an International Geo Sample Number (IGSN; i.e., globally unique alphanumeric sample identifiers), which can be obtained from the System for Earth Sample Registration (www.geosamples.org). However, SGP is a domain-specific project that differs from other databases in the way the data are collected, the nature of the data collected, and the tailored way in which they are presented to our research community.

Specifically, SGP is focused on addressing how geochemical proxy records change through deep time. Central to these goals are the following:

1. Compilation of a large quantity (i.e., millions of records) of sedimentary geochemical data spanning deep time.
2. Appropriate age models (with uncertainty), especially for Proterozoic/Archean samples.
3. Information on interpreted depositional environment and specific rock type.
4. Information necessary to gauge whether samples are likely to preserve primary, environmental geochemical signals.
5. Detailed methodological information on how the data were generated.
6. An ability to download the data of interest flexibly and easily.

Although some other databases contain sedimentary geochemical data, the vast majority of deep-time data is not available from any single source, and samples are not readily associated with critical contextual data—such as age constraints and environmental data—necessary for the types of proxy-through-time and/or environmental studies typically conducted in historical geobiology. When the SGP was founded in 2015, we believed that a "team science" philosophy would be the most effective way to move beyond spreadsheets to the type and abundance of data required. The research consortium framework we have implemented is modeled after mature consortia in human statistical genetics, such as the Psychiatric Genomics Consortium (PGC). In the PGC, researchers have aggregated data to make statistically robust observations and landmark findings not possible with the data generated by any single research group alone (Duncan et al., 2017; Schizophrenia Working group of the Psychiatric Genomics Consortium, 2014; Wray et al., 2018). Similar to biomedical research consortia, we hope that the intellectual and collaborative environment fostered by SGP will ultimately be as important as our data products or specific insights in research papers.

The first priority for Phase 1 of SGP was to assemble or generate multi-proxy sedimentary geochemical data (carbon and sulfur abundances and isotopes, iron speciation, major and trace metal abundances, and trace metal isotopes, primarily from fine-grained siliciclastic rocks) from multiple regions worldwide for every Paleozoic Epoch and equivalent ~25 Myr Neoproterozoic time slice. In addition to data compilation, this has involved an effort by SGP members to generate new geochemical data from "background" intervals in the Paleozoic (i.e., not associated with events such as mass extinctions or significant climatic shifts). The first phase of data collection came to an end in 2019. At that point, a copy of the database was vetted by SGP team members and then archived—the first data "freeze" (following the best-practices approach used in medical consortia). Working groups were formed (with working group leadership established through an open call to SGP team members), and data were made available to Working group analysts via the website and through tailored queries. The first working group papers have recently been published (LeRoy et al., 2021; Lipp et al., 2021; Mehra et al., 2021), and more are in progress. Meanwhile, data collection continues, and the Phase 2 goal is to include more Mesozoic–Cenozoic and pre-Neoproterozoic time intervals and to expand the geochemical record to more diverse lithologies and grain-specific phases. The Phase 2 data freeze is currently anticipated for 2023, followed by data vetting and analyses toward group papers.

3 | DATABASE

SGP utilizes a relational database implemented with the PostgreSQL database management system. A full database diagram and documentation are available at https://github.com/ufarrell/sgp_phase1, and a simplified diagram is shown in Figure 2. The design was inspired by several existing data models in the geological and natural history museum communities. Tables for analytical geochemistry are from the British Geological Survey (BGS) geochemistry data model (Watson et al., 2014), with minor modifications. Tables for geological, geographical, and sample details are based on established museum collection management databases (Specify 6 https://www.specifysoftware.org/ and Arctos https://arctosdb.org/) in addition to the Observations Data Model 2 (ODM2, Horsburgh et al., 2016; Hsu et al., 2017), an information model for Earth observations.

The SGP database is centered on the sample table (Figure 2). Samples are generally characterized by an individual rock sample and all resulting analyzed powders. The three key sections of the database linked to samples are (1) analytical results and associated methods, (2) geographical context, and (3) geological context. Dictionary tables (standardized lists of terms, also known as "controlled

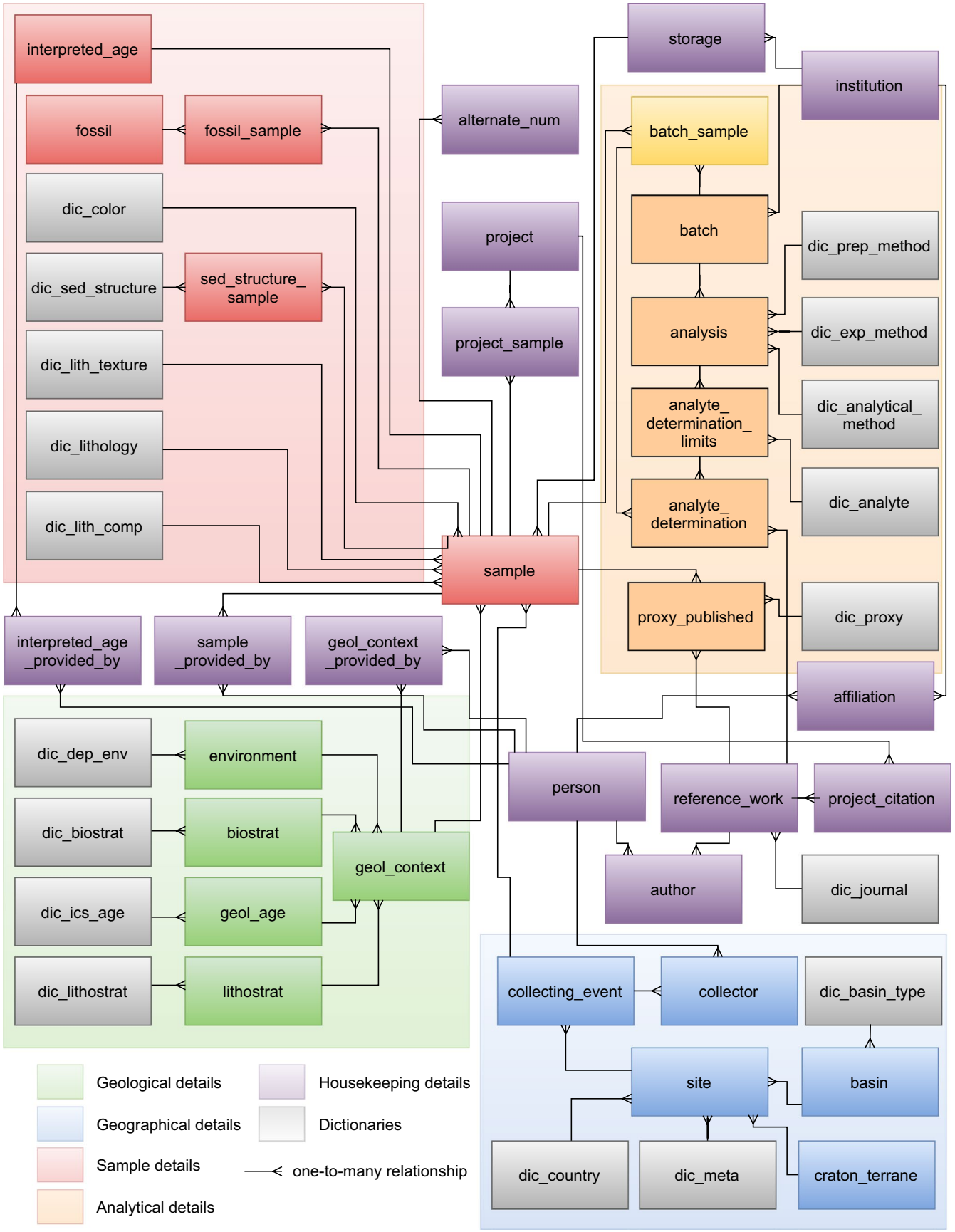

FIGURE 2 Simplified schema showing tables and table relationships in the SGP database (https://ufarrell.github.io/sgp_phase1/ for a detailed description). Tables are grouped according to the kind of information they store. Analytical tables (orange) are from the British Geological Survey model (Watson et al., 2014). Geographical, geological (green), and sample (red) tables are primarily based on natural history museum databases. "Housekeeping" tables (purple) record information such as how samples are grouped into projects, where they are stored, and who has contributed contextual information

vocabularies") are based on existing community vocabularies where possible (e.g., from EarthChem, ODM2, Macrostrat, U.S. Geological Survey (USGS), and BGS). However, in many cases, these vocabularies required additions, such as the inclusion of specific sedimentary geochemical experimental methods (e.g., sequential iron extraction techniques; Poulton & Canfield, 2005).

The BGS data model for analytical methods and geochemical results has been adopted almost without modification. We store analytical data in their submitted or published format and do not standardize the results to any given unit. An analytical result may be empty (NULL) only if it is below or above detection limits, and those values are also stored if they are available. If the results are published, they are linked directly to a reference work on an individual basis so that a fine-level distinction can be made between published and related unpublished data from the same samples. Any geostandards that are analyzed alongside samples in a study are also recorded.

In the SGP, we make every effort not to include the same result twice. However, replicates may legitimately be added if the same sample has undergone analysis for the same analyte more than once (this could include anything from true replicate analyses using the same methods in the same laboratory to analyses of the same sample by different research groups using different methods). We do not currently assign new sample identifiers to sub-samples. A parent–child relationship may be added in Phase 2 when the focus will expand to include carbonate data.

4 | DATA COLLECTION

The SGP welcomes contributions from any interested researchers. Specifically, contributing data automatically makes a researcher part of the SGP Collaborative Team, rather than one needing to "join" SGP to contribute data. In the first consortium-building stage, potential collaborators were targeted if their work was particularly relevant to the Phase 1 goals, and additional researchers were recruited via SGP representation at multiple conferences. SGP collaborators are involved in providing details about their samples and providing published data tables and unpublished data from their own archives. In addition, some data have been collected from relevant published studies where the authors are not directly involved. In such cases, contextual information was coded by SGP team members using information provided in the paper.

SGP collaborators are asked to fill in a template with contextual information as completely as possible, but with an emphasis on key fields such as modern latitude and longitude, stratigraphic unit name, depositional environment, and lithology. A particularly important field is interpreted age, which is a numerical estimate for the age of each sample in millions of years (Ma). Whenever possible, the original authors, who are most familiar with the samples and stratigraphic sections, are asked to provide the interpreted age. They can use whatever method with which they feel most comfortable; for example, ages may be estimated based on assumed sedimentation rates and/or linear interpolation, or groups of samples can be assigned one age based on proximity to any available time markers. A brief justification is required for each age provided, which may be used in the future to refine ages further. Maximum and minimum age estimates can also be stored, and indeed, are critical for the type of re-weighted bootstrap analyses employed by many SGP working groups (Mehra et al., 2021).

A subset of samples from two USGS databases has been integrated into the SGP database. The first of the databases used is the National Geochemical Database: Rock (USGS NGDB, U.S. Geological Survey, 2008), comprising data from USGS projects from the 1960s to1990s, largely from North America. The second is the Global Geochemical Database for Critical Metals in Black Shales project (USGS CMIBS, Granitto et al., 2017), which includes predominantly Phanerozoic shale data from all continents. Data from both USGS databases lack much of the contextual information available for samples directly coded by the SGP team members (most specifically basin type, metamorphic/maturity grade, depositional environment, and detailed age justification) and there are a higher proportion of analytes with less detailed geochemical methodology. Nevertheless, they represent large numbers of samples (74% of samples in Phase 1 are from USGS sources) with age, lithology, and geographic information that can be utilized for many types of analysis.

In the case of USGS NGDB, only sedimentary samples were incorporated into SGP, and in the case of USGS CMIBS, we did not include samples with lithologies indicative of ore or studies where the authors were primarily concerned with mineral deposits or studying the effects of metamorphism on shales. An attempt was made to match USGS fields to SGP fields, with some data cleaning needed in order to extract important information such as up-to-date stratigraphic names. Samples can easily be traced back to the original USGS databases using their original identifiers.

The USGS NGDB data were enhanced by adding interpreted ages. Samples were matched, using a combination of stratigraphy and location, to the continuous-time age model in Macrostrat (Peters et al., 2018). Specifically, the minimum and maximum age estimates from the Macrostrat model were entered, and the interpreted age was entered as the average of these values. Only samples with matched interpreted ages were included from USGS NGDB. The USGS CMIBS samples were associated with Macrostrat continuous-time age models where possible and given age information by SGP team members where not. However, a proportion (36%) remain without ages, and filling those in is a key goal for Phase 2.

These three sources of data (direct entry by SGP team members (26% of samples), the CMIBS compilation (16% of samples), and the USGS NGDB (58% of samples)) provide a robust base platform for statistical analyses of aggregated sedimentary geochemical data through Earth history. Moving forward, we will continue direct entry from SGP team members, and work toward incorporating

geochemical data compiled by additional geological surveys (for instance, incorporation of the OZCHEM whole-rock database from Geoscience Australia is currently in progress).

5 | DATA DESCRIPTION PHASE 1

Phase 1 of data collection ended in August 2019. A static version of the database was archived and made available to collaborators through the website (sgp-search.io) and via tailored queries. Time was allowed for vetting, and any errors discovered were corrected before the final freeze in February 2020. The Phase 1 data freeze includes 82,578 samples, with 2,701,236 analytical results, and was made public through our search website in December 2020. This paper should be cited in the future use of Phase 1 data downloads. More complete information on the Phase 1 data product can be found on the SGP wiki (https://github.com/ufarrell/sgp_phase1/wiki), including summaries by age, lithology, and geochemical methodology, as well as the specifics of how USGS databases were incorporated into the SGP structure.

6 | SGP

The SGP-contributed dataset includes 20,811 samples with 518,291 results. Approximately two thirds of the data (64%) come from 160 published sources (https://github.com/ufarrell/sgp_phase1/wiki/SGP-data-references). The remaining 36% are from unpublished sources, including new and legacy data. The samples come from 942 individual sites from 46 countries (Figure 3). Consistent with the Phase 1 goals, 84% of samples were from the Neoproterozoic–Paleozoic (Figure 4). Sixty-four percent of samples are fine-grained siliciclastic rocks (shale, mudstone, or siltstone), as are the majority of uncoded lithologies (Figure 5).

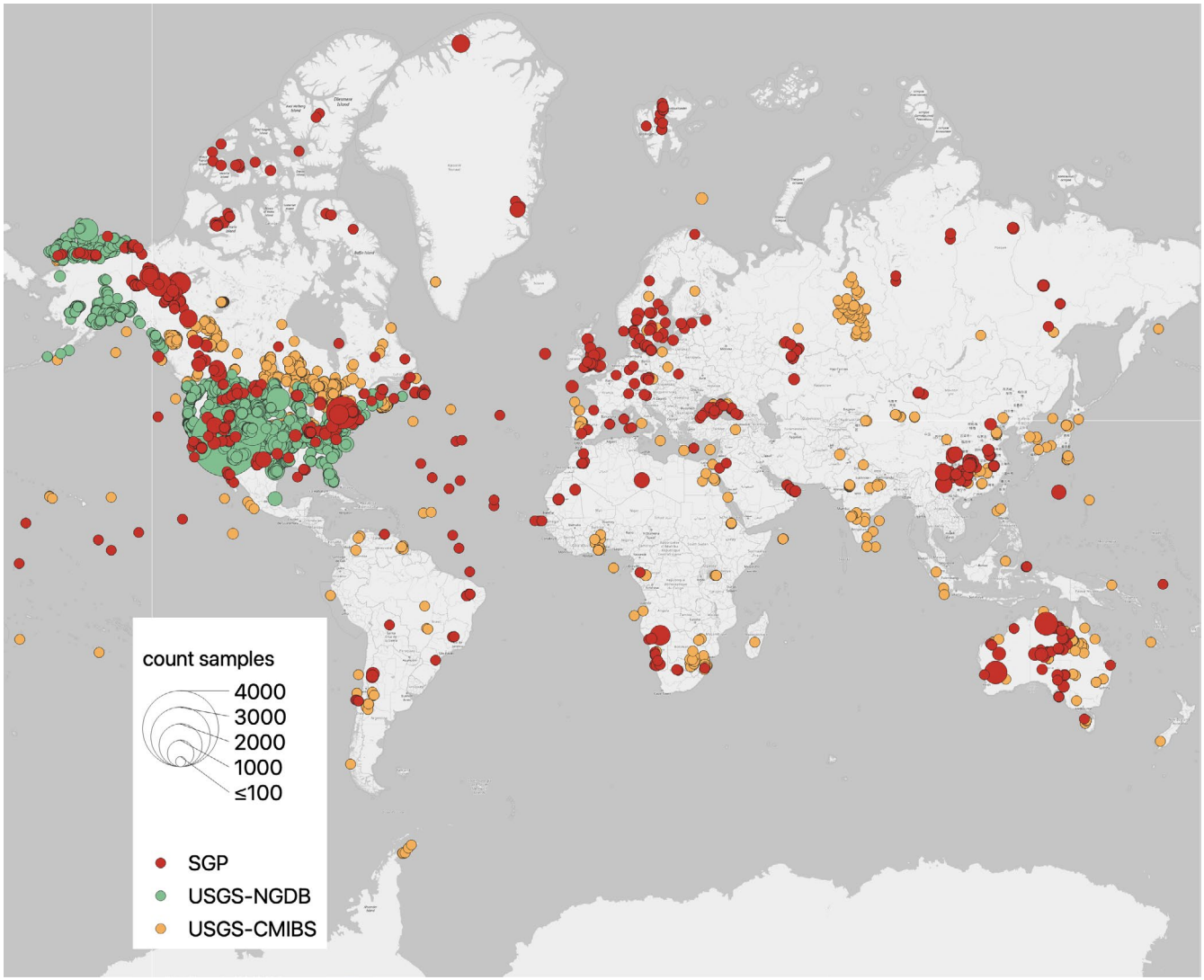

FIGURE 3 Geographic distribution of samples in the Phase 1 dataset, separated by our three main data sources (SGP direct entry, USGS CMIBS, and USGS NGDB)

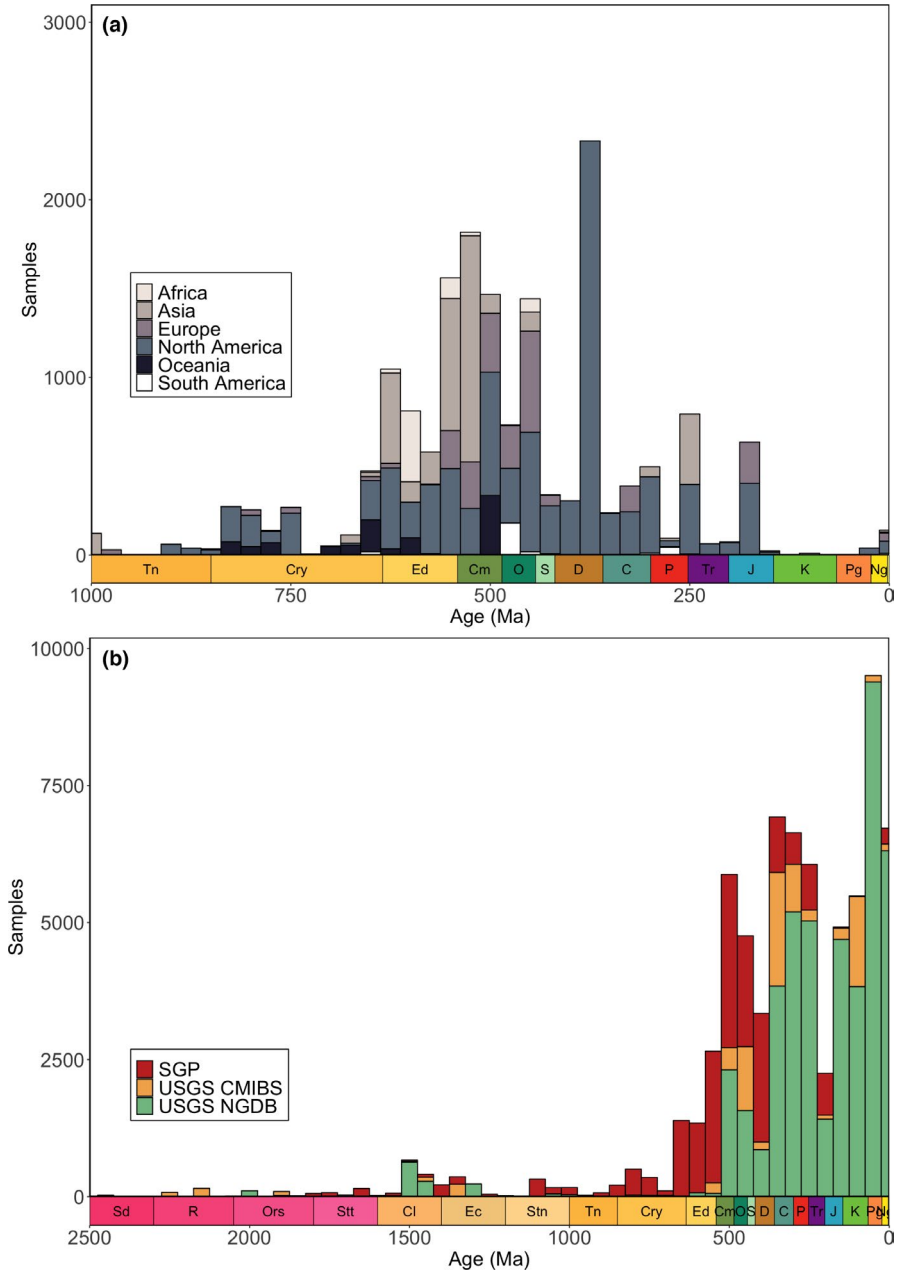

FIGURE 4 Distribution by age and continent for SGP direct entry data (a). Distribution by age for SGP, USGS CMIBS, and USGS NGDB data (a small number of samples (489) with ages >2500 Ma are not included in the figure) (b)

7 | USGS NGDB

The data from USGS NGDB that are incorporated into the SGP database include 48,234 samples with 1,769,696 results. Nearly all (99%) of the samples are from the United States. Nineteen percent are sandstone, 13% are shale, and 29% do not have a specific lithology (although lithological details may be available in verbatim fields; Figure 5). Contextual details, including depositional environment and low-grade metamorphic bin, are mostly not available for these samples, and methodological information is sparse. In general, the USGS NGDB samples skew younger than the SGP samples: 39% are from the Paleozoic, 25% from the Mesozoic, and 33% from the Cenozoic (~3% of samples are from the Proterozoic/Archean).

The USGS database provides excellent coverage of the United States, but given the remit of the organization, with strong focus on economic deposits (petroleum-producing units, phosphatic units, and sedimentary mineral deposits), the sampling may not be representative of the entire country. This is distinct from the bias present in geochemical data produced by academic researchers, which are often focused on mass extinction intervals, Earth system perturbations, and other stratigraphic boundaries.

8 | USGS CMIBS

The data incorporated from USGS CMIBS into the SGP database include 12,797 samples with 409,188 results. The samples are from 45 countries, with 40% from Canada, 27% from the United States, and 13% from Australia. The majority of samples are fine-grained siliciclastic sediments (69% shale, mudstone, siltstone, or

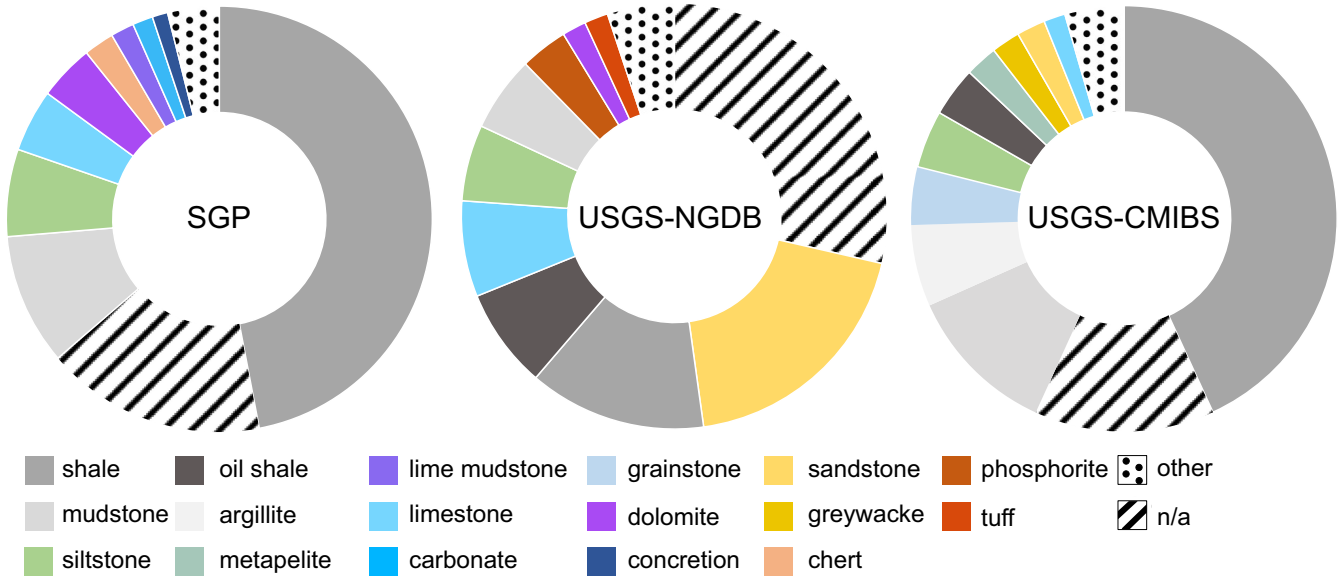

FIGURE 5 Representation of lithologies in the Phase 1 dataset. Note that most unclassified samples from SGP direct entry and USGS CMIBS will be fine-grained clastic rocks (e.g., shale), whereas USGS NGDB unclassified samples are more heterogeneous

argillite; Figure 5). Sixty percent of samples with interpreted ages are Paleozoic, 24% are Mesozoic, 2% are Cenozoic, and 15% are Proterozoic/Archean.

As was the case for USGS NGDB, contextual details, including depositional environment and low-grade metamorphic bin, are often missing for these samples. However, more detailed geochemical methodological information is available. Each sample in CMIBS has a "best value" result per analyte, selected from multiple values that were originally available (Granitto et al., 2017). The choice of "best value" was made using a rubric which included consideration of the sample weight, the sample "decomposition" (e.g., full vs. partial acid digestion), the instruments used in the analysis, and the detection limits (Granitto et al., 2013).

9 | DATA PRESENTATION AND ACCESS

The SGP search website (sgp-search.io) utilizes an intuitive user interface to query the Phase 1 database via an API. The two main search types are "samples" and "analyses," with "nhhxrf" simply being a "samples" search that excludes any handheld XRF (X-ray fluorescence) data. This methodological distinction is made because while handheld XRF data can be accurate for some elements (e.g., Ca and Fe), it is highly inaccurate for many others (e.g., S, Ni) (Rowe et al., 2012). Handheld XRF data represent 1% of the total results and 4% of SGP-contributed data; although this is a small percentage now, we anticipate continued growth given the popularity and utility of handheld XRFs. A "samples" search will list an individual sample on each row, with geological context information and geochemical analytes taking up the columns. Data are converted to one standard unit, and oxides are converted to elements (e.g., Al_2O_3 to Al), and values are averaged if more than one analysis was made per sample. Note, this search may average values produced using different analytical methods, although the number of samples in the database with multiple analytical values for a specific analyte is relatively small. Further, any analyses below or above detection limit are removed, as these cannot be averaged. This has implications for queries involving very low abundance elements (e.g., Ag in sedimentary rocks), as only results above detection limits, and thus higher values, will be included. We anticipate that this search will produce the optimal data output for most end-users interested in Earth history: a file with age, geological context, and geochemical data for each sample.

If users are looking to delve deeper into the data and understand the analyses and procedures that were executed to obtain each sample's geochemical data, then the "analyses" search is useful because it lists every analysis recorded in the database in a separate row. The "analyses" search also allows users to show data relating to the laboratory where the sample was analyzed, the person who made the measurement, geochemical methodology, etc. At the current time, aside from the ability to exclude handheld XRF data, the "samples" and "nhhxrf" search types will not report information about, or have the ability to filter by, geochemical methodology. Users who are interested in methodological details or who would like to export a data file beyond the size limit (10 Mb) should contact the SGP Leadership Team regarding a custom SQL query.

Once the user has selected a search type, samples can be filtered based on both geological context and geochemical attributes. Note that for many samples some aspects of geological contextual information are incomplete. Thus, for example, a search filtering for samples deposited in a rift basin will only return samples positively described as such and not necessarily all samples in the database deposited in rift basins. Given that samples will have non-overlapping missing data, too many filters may result in a smaller-than-expected dataset.

Search results will appear in a "preview" window that can be used to check the output. Each sample also has an information icon associated with it; clicking this icon will bring up a lightbox with detailed sample information. Finally, the user may request to show reference information for their search. For "analyses" searches (where every analysis is shown as an individual row), this will return the specific literature citation for that individual analytic result. For other search types, this will return, for every sample, a concatenated list of all references whose geochemical data contributed to that specific search.

When the user is satisfied with their search, they can then download a.csv file of the data and export a map showing the location and age of samples in their search.

The SGP website uses an API to interact with the database, and users can make a copy of the API call using the API icon next to their search results. However, users can also bypass the user interface entirely and access data via a direct API call. This comprises three parts:

- type: Selects the search type (samples, analyses or nhhxrf)
- filters: Contains a list of search options that are logically ANDed in the results
- show: Contains search options that determine which columns will appear in the results

Thus, an example API call would be

{"type":"samples","filters":{"country":["Argentina","Brazil","Chile","Bolivia","Colombia","Venezuela"],"toc":[2,100]},"-show":["toc","fe","height_meters","section_name","country","interpreted_age"]}.

This API call is making a "samples" type search for samples that originate from Argentina, Brazil, Chile, Bolivia, Colombia, or Venezuela and have 2%–100% total organic carbon (TOC) content. In other words, searching for organic-rich samples from South America. In addition, the API call is asking for a results output table with columns that show TOC (wt%), Fe (wt%), section or core name, collection height in meters, each sample's country, and the age in millions of years. Full documentation and a tutorial video are available on the website.

10 | FUTURE GOALS AND DIRECTIONS

The overarching goal of SGP was to provide intellectual and geoinformatic resources for the Earth Science community to advance our understanding of environmental changes on Earth through time. A better understanding of Earth's history requires sufficient data density, but equally importantly it means training a new generation of researchers with the data science and statistical skills to make meaningful conclusions from large sedimentary geochemical datasets. Much of the focus in SGP Phase 1 was in initiating the consortium and increasing the data product to the point where it was useful for analyses by the community. We now aim to increasingly move toward developing a community-initiated set of best practices for data management, a culture of publishing metadata, and a shared intellectual framework for analyzing such datasets. Over the course of Phase 2, we plan to continue holding annual meetings at Goldschmidt while also beginning regular video calls to share progress and ideas for data analysis. We will also develop accessible "Proxy Primer" videos to help the geobiological community understand the strengths and weaknesses of different proxies.

Beyond these broad community and educational goals, we have the following more concrete goals during SGP Phase 2:

- Expand the geological and geographic scope of samples in our database. Most samples with complete context information (SGP direct entry), and indeed most samples, are Neoproterozoic–Paleozoic in age and from North America (Figure 4). Younger and older samples, and worldwide sampling, are necessary for accurate analyses through the full swath of Earth history.
- Expand the carbonate geochemical record. Our database structure is appropriate for carbonate data (and indeed, >8000 carbonate samples are already in the database). However, this goal will require community discussion regarding how best to incorporate methodologies and phase-specific analyses.
- Continue correcting errors in previously entered data. Although we have been as careful as possible during data entry, mistakes are inevitable in a dataset of this size. Paleobiological analyses and basic statistical logic suggest that such mistakes (random error) will not affect results as long as they are not biased (systematic error) (Sepkoski, 1993). Nonetheless, we would like to present the most accurate results, and we welcome users to notify us of true errors (rather than geologic disagreement) that are found during their database searches.
- Continue developing the SGP search website and API to best serve the sedimentary geochemistry and Earth history communities.
- Expand the community and user group. Anyone who is interested in contributing to the project is welcome, and helping the community grow our data resource is the only requirement to join the SGP Collaborative Team. Details, including contact information and sample submission templates, are available at https://sgp.stanford.edu/. We want SGP to be a hub for deep-time sedimentary geochemical research, and researchers from diverse backgrounds, early-career researchers, and researchers working or studying outside Europe and North America (where the bulk of SGP members reside) are especially invited to become involved.

Echoing this final point, we reiterate that the SGP is a community-oriented research consortium, and we welcome suggestions on how to best move toward our shared goals.

KEYWORDS
consortium, database, Earth history, geochemistry, website

ACKNOWLEDGMENTS

We thank Sufian Lattouf for developing the initial version of the SGP website, and Kai Lenz, Kassie Sharp, Aaron Cole, Clare Swan, Lyna Kim, and John Freshwaters for computational assistance. We thank Erin Saupe, Itay Halevy, Jordon Hemingway, Minming Cui, Maya Gomes, Matthew Granitto, Alf Lenz, Charles Henderson, Chengsheng Jin, Clint Scott, David Champion, Jinghai Yang, Joe Shaffer, Kathy Doyle, Lei Xiang, Liam Bhajan, Patrick Sack, Paul Hoffman, Paulo Linarde Dantas Mascena, Will Thompson-Butler, and Yu Liu for their contributions to SGP. We thank Patrick Sullivan and Laramie Duncan for discussions regarding the PGC and research consortium organization. We thank the donors of The American Chemical Society Petroleum Research Fund for partial support of SGP website development (61017-ND2). EAS is funded by National Science Foundation grant (NSF) EAR-1922966. BGS authors (JE, PW) publish with permission of the Executive Director of the British Geological Survey, UKRI. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. The authors declare no conflicts of interest.

Úna C. Farrell[1]
Rifaat Samawi[2]
Savitha Anjanappa[3]
Roman Klykov[3]
Oyeleye O. Adeboye[4]
Heda Agic[5]
Anne-Sofie C. Ahm[6]
Thomas H. Boag[7]
Fred Bowyer[8]
Jochen J. Brocks[9]
Tessa N. Brunoir[10]
Donald E. Canfield[11]
Xiaoyan Chen[12]
Meng Cheng[13]
Matthew O. Clarkson[14]
Devon B. Cole[15]
David R. Cordie[16]
Peter W. Crockford[17]
Huan Cui[18,19]
Tais W. Dahl[20]
Lucas D. Mouro[21]
Keith Dewing[22]
Stephen Q. Dornbos[23]
Nadja Drabon[24]
Julie A. Dumoulin[25]
Joseph F. Emmings[26]
Cecilia R. Endriga[2]
Tiffani A. Fraser[27]
Robert R. Gaines[28]
Richard M. Gaschnig[29]
Timothy M. Gibson[7]
Geoffrey J. Gilleaudeau[30]
Benjamin C. Gill[31]
Karin Goldberg[32]
Romain Guilbaud[33]
Galen P. Halverson[34]
Emma U. Hammarlund[35]
Kalev G. Hantsoo[36]
Miles A. Henderson[37]
Malcolm S.W. Hodgskiss[38]
Tristan J. Horner[39]
Jon M. Husson[40]
Benjamin Johnson[41]
Pavel Kabanov[22]
C. Brenhin Keller[42]
Julien Kimmig[43]
Michael A. Kipp[44]
Andrew H. Knoll[45]
Timmu Kreitsmann[46]
Marcus Kunzmann[47]
Florian Kurzweil[48]
Matthew A. LeRoy[31]
Chao Li[13]
Alex G. Lipp[49]
David K. Loydell[50]
Xinze Lu[51]
Francis A. Macdonald[5]
Joseph M. Magnall[52]
Kaarel Mänd[53]
Akshay Mehra[42]
Michael J. Melchin[54]
Austin J. Miller[51]
N. Tanner Mills[55]
Chiza N. Mwinde[56]
Brennan O'Connell[57]
Lawrence M. Och[58]
Frantz Ossa Ossa[59]
Anais Pagès[60]
Kärt Paiste[61]
Camille A. Partin[62]
Shanan E. Peters[63]
Peter Petrov[64]
Tiffany L. Playter[65]
Stephanie Plaza-Torres[66]
Susannah M. Porter[5]
Simon W. Poulton[8]
Sara B. Pruss[67]
Sylvain Richoz[68]
Samantha R. Ritzer[2]
Alan D. Rooney[7]
Swapan K. Sahoo[69]
Shane D. Schoepfer[70]
Judith A. Sclafani[2]
Yanan Shen[12]

Oliver Shorttle[38]
Sarah P. Slotznick[42]
Emily F. Smith[36]
Sam Spinks[47]
Richard G. Stockey[2]
Justin V. Strauss[42]
Eva E. Stüeken[71]
Sabrina Tecklenburg[2]
Danielle Thomson[72]
Nicholas J. Tosca[73]
Gabriel J. Uhlein[74]
Maoli N. Vizcaíno[2]
Huajian Wang[75]
Tristan White[7]
Philip R. Wilby[26]
Christina R. Woltz[5]
Rachel A. Wood[76]
Lei Xiang[77]
Inessa A. Yurchenko[78]
Tianran Zhang[42]
Noah J. Planavsky[7]
Kimberly V. Lau[79]
David T. Johnston[24]
Erik A. Sperling[2]

[1]*Department of Geology, Trinity College Dublin, Dublin, Ireland*
[2]*Department of Geological Sciences, Stanford University, Stanford, California, USA*
[3]*Aionis, Los Gatos, California, USA*
[4]*Boone Pickens School of Geology, Oklahoma State University, Stillwater, Oklahoma, USA*
[5]*Department of Earth Science, University of California, Santa Barbara, Santa Barbara, California, USA*
[6]*Department of Geosciences, Princeton University, Princeton, New Jersey, USA*
[7]*Department of Earth and Planetary Sciences, Yale University, New Haven, Connecticut, USA*
[8]*School of Earth and Environment, University of Leeds, Leeds, UK*
[9]*Research School of Earth Sciences, Australian National University, Canberra, ACT, Australia*
[10]*Department of Earth and Planetary Sciences, University of California, Davis, Davis, California, USA*
[11]*Nordic Center for Earth Evolution (NordCEE), University of Southern Denmark, Odense, Denmark*
[12]*School of Earth and Space Science, University of Science and Technology of China, Hefei, China*
[13]*State Key Laboratory of Biogeology and Environmental Geology, China University of Geosciences, Wuhan, China*
[14]*Department of Earth Sciences, Institute of Geochemistry and Petrology, ETH Zurich, Zurich, Switzerland*
[15]*School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA*
[16]*Division of Physical, Computational, and Mathematical Sciences, Edgewood College, Madison, Wisconsin, USA*
[17]*Earth and Planetary Science, Weizmann Institute of Science, Rehovot, Israel*
[18]*Equipe Géomicrobiologie, Institut de Physique du Globe de Paris (IPGP), Université de Paris, Paris, France*
[19]*Stable Isotope Laboratory, Department of Earth Sciences, University of Toronto, Toronto, Ontario, Canada*
[20]*GLOBE Institute, University of Copenhagen, Copenhagen, Denmark*
[21]*Instituto de Geociências, University of São Paulo, São Paulo, SP, Brazil*
[22]*Natural Resources Canada, Geological Survey of Canada, Calgary, Alberta, Canada*
[23]*Department of Geosciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA*
[24]*Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA*
[25]*U.S. Geological Survey, Alaska Science Center, Anchorage, Alaska, USA*
[26]*British Geological Survey, Keyworth, UK*
[27]*Yukon Geological Survey, Government of Yukon, Whitehorse, Yukon, Canada*
[28]*Department of Geology, Pomona College, Claremont, California, USA*
[29]*Department of Environmental Earth and Atmospheric Sciences, University of Massachusetts Lowell, Lowell, Massachusetts, USA*
[30]*Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, Virginia, USA*
[31]*Department of Geosciences, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA*
[32]*Department of Geology, Kansas State University, Manhattan, Kansas, USA*
[33]*Géosciences Environnement Toulouse, Université de Toulouse, CNRS, Toulouse, France*
[34]*Department of Earth and Planetary Sciences/Geotop, McGill University, Montreal, QC, Canada*
[35]*Department of Laboratory Medicine, Lund University, Lund, Sweden*
[36]*Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, Maryland, USA*
[37]*Department of Geosciences, University of Texas Permian Basin, Odessa, Texas, USA*
[38]*Department of Earth Sciences, University of Cambridge, Cambridge, UK*
[39]*Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, USA*
[40]*Department of Earth and Ocean Sciences, University of Victoria, Victoria, British Columbia, Canada*
[41]*Department of Geological and Atmospheric Sciences, Iowa State University, Ames, USA*

[42]Department of Earth Sciences, Dartmouth College, Hanover, New Hampshire, USA
[43]Earth and Mineral Sciences Museum & Art Gallery, Pennsylvania State University, University Park, Pennsylvania, USA
[44]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California, USA
[45]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, USA
[46]Department of Physics and Earth Sciences, Jacobs University Bremen, Bremen, Germany
[47]Australian Resources Research Centre, CSIRO Mineral Resources, Kensington, Western Australia, Australia
[48]Department of Geology and Mineralogy, University of Cologne, Cologne, Germany
[49]Department of Earth Sciences and Engineering, Imperial College London, London, UK
[50]School of the Environment, Geography and Geosciences, University of Portsmouth, Portsmouth, UK
[51]Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, Ontario, Canada
[52]GFZ German Research Centre for Geosciences, Potsdam, Germany
[53]Department of Geology, University of Tartu, Tartu, Estonia
[54]Department of Earth Sciences, St. Francis Xavier University, Antigonish, Nova Scotia, Canada
[55]Department of Geology and Geophysics, Texas A&M University, College Station, Texas, USA
[56]Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois, USA
[57]School of Earth Sciences, University of Melbourne, Melbourne, Victoria, Australia
[58]Dr. von Moos AG, Zurich, Switzerland
[59]Department of Geosciences, University of Tuebingen, Tuebingen, Germany
[60]Department of Water and Environmental Regulation, Government of Western Australia, Joondalup, Western Australia, Australia
[61]Faculty of Science and Technology, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia
[62]Department of Geological Sciences, University of Saskatchewan, Saskatoon, Canada
[63]Department of Geoscience, University of Wisconsin-Madison, Madison, Wisconsin, USA
[64]Geological Institute, Russian Academy of Sciences, Moscow, Russia
[65]Alberta Geological Survey, Edmonton, Alberta, Canada
[66]Geological Sciences, University of Colorado Boulder, Boulder, Colorado, USA
[67]Department of Geosciences, Smith College, Northampton, Massachusetts, USA
[68]Department of Geology, Lund University, Lund, Sweden
[69]Equinor, Houston, Texas, USA
[70]Geoscience and Natural Resources, Western Carolina University, Cullowhee, North Carolina, USA
[71]School of Earth and Environmental Sciences, University of St Andrews, St Andrews, UK
[72]Shell Canada, Calgary, Alberta, Canada
[73]Department of Earth Sciences, University of Oxford, Oxford, UK
[74]Department of Geology, Federal University of Minas Gerais, Belo Horizonte, Brazil
[75]Research Institute of Petroleum Exploration and Development, China National Petroleum Corporation, Beijing, China
[76]School of Geosciences, University of Edinburgh, Edinburgh, UK
[77]State Key Laboratory of Palaeobiology and Stratigraphy, Nanjing Institute of Geology and Palaeontology and Center for Excellence in Life and Paleoenvironment, Nanjing, China
[78]Chevron Technical Center, Houston, Texas, USA
[79]Department of Geosciences, Earth and Environmental Systems Institute, Pennsylvania State University, University Park, Pennsylvania, USA

Correspondence
Erik Sperling, Stanford University, Department of Geological Sciences, 450 Jane Stanford Way Stanford, CA 94305 USA.
Email: esper@stanford.edu

ORCID

Úna C. Farrell https://orcid.org/0000-0002-3489-8512
Oyeleye O. Adeboye https://orcid.org/0000-0002-3262-0664
Jochen J. Brocks https://orcid.org/0000-0002-8430-8744
Donald E. Canfield https://orcid.org/0000-0001-7602-8366
Devon B. Cole https://orcid.org/0000-0002-5669-3817
Timothy M. Gibson https://orcid.org/0000-0003-0836-4565
Benjamin Johnson https://orcid.org/0000-0001-6925-3223
Matthew A. LeRoy https://orcid.org/0000-0002-3572-9841
Chao Li https://orcid.org/0000-0001-9861-661X
Kaarel Mänd https://orcid.org/0000-0003-1575-3710
N. Tanner Mills https://orcid.org/0000-0001-8578-3232
Brennan O'Connell https://orcid.org/0000-0002-5652-1222
Lawrence M. Och https://orcid.org/0000-0001-6207-5348
Susannah M. Porter https://orcid.org/0000-0002-4707-9428
Sara B. Pruss https://orcid.org/0000-0003-1751-2697
Alan D. Rooney https://orcid.org/0000-0002-5023-2606
David T. Johnston https://orcid.org/0000-0002-2487-1084
Erik A. Sperling https://orcid.org/0000-0001-9590-371X

REFERENCES

Brantley, S. L., Tao, W., Agarwal, D., Catalano, J., Schroeder, P. A., Lehnert, K. A., Varadharajan, C., Pett-Ridge, J., Engle, M., Castronova, A. M., Hooper, R., Ma, X., Jin, L., McHenry, K., Aronson, E., Shaughnessy, A. R., Derry, L., Richardson, J., Bales, J., & Pierce, E. (2020). A vision for the future low-temperature geochemical data-scape. *EarthArXiv*, https://doi.org/10.31223/X5ZP5W.

Cutcher-Gershenfeld, J., Baker, K. S., Berente, N., Carter, D. R., DeChurch, L. A., Flint, C. C., Gershenfeld, G., Haberman, M., King, J.

L., Kirkpatrick, C., Knight, E., Lawrence, B., Lewis, S., Lenhardt, W. C., Lopez, P., Mayernik, M. S., McElroy, C., Mittleman, B., Nichol, V., … Zaslavsky, I. (2016). Build it, but will they come? A Geoscience cyberinfrastructure baseline analysis. *Data Science Journal*, *15*, 8. https://doi.org/10.5334/dsj-2016-008.

Duncan, L., Yilmaz, Z., Gaspar, H., Walters, R., Goldstein, J., Anttila, V., Bulik-Sullivan, B., Ripke, S., Thornton, L., Hinney, A., Daly, M., Sullivan, P. F., Zeggini, E., Breen, G., Bulik, C. M., Duncan, L., Yilmaz, Z., Gaspar, H., … Bulik, C. M. (2017). Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa. *American Journal of Psychiatry*, *174*, 850–858. https://doi.org/10.1176/appi.ajp.2017.16121402.

FAIR (2019). FAIR Play in geoscience data. *Nature Geoscience*, *12*, 961. https://doi.org/10.1038/s41561-019-0506-4.

Gard, M., Hasterok, D., & Halpin, J. A. (2019). Global whole-rock geochemical database compilation. *Earth System Science Data*, *11*, 1553–1566. https://doi.org/10.5194/essd-11-1553-2019.

U.S. Geological Survey. (2008). *Geochemistry of rock samples from the National Geochemical Database*. U.S. Geological Survey. https://mrdata.usgs.gov/ngdb/rock

Granitto, M., Giles, S. A., & Kelley, K. D. (2017). *Global Geochemical Database for Critical Metals in Black Shales*. U.S. Geological Survey data release. https://doi.org/10.5066/F71G0K7X.

Granitto, M., Schmidt, J. M., Shew, N. B., Gamble, B. M., & Labay, K. A. (2013). *Alaska Geochemical Database, Version 2.0 (AGDB2)—Including "best value" data compilations for rock, sediment, soil, mineral, and concentrate sample media*. U.S. Geological Survey Data Series 759, 20 p. pamphlet and database, 1 DVD. https://pubs.usgs.gov/ds/759/

Hanson, B. (2016). AGU opens its journals to author identifiers. *Eos*, *97*. https://doi.org/10.1029/2016EO043183.

He, Y., Bai, Y., Tian, D., Yao, L., Fan, R., & Chen, P. (2019). A review of geoanalytical databases. *Acta Geochimica*, *38*, 718–733. https://doi.org/10.1007/s11631-019-00323-3.

Horsburgh, J. S., Aufdenkampe, A. K., Mayorga, E., Lehnert, K. A., Hsu, L., Song, L., Jones, A. S., Damiano, S. G., Tarboton, D. G., Valentine, D., Zaslavsky, I., & Whitenack, T. (2016). Observations Data Model 2: A community information model for spatially discrete Earth observations. *Environmental Modelling & Software*, *79*, 55–74. https://doi.org/10.1016/j.envsoft.2016.01.010.

Hsu, L., Mayorga, E., Horsburgh, J. S., Carter, M. R., Lehnert, K. A., & Brantley, S. L. (2017). Enhancing interoperability and capabilities of Earth science data using the Observations Data Model 2 (ODM2). *Data Science Journal*, *16*, 4. https://doi.org/10.5334/dsj-2017-004.

Lehnert, K., Su, Y., Langmuir, C. H., Sarbas, B., & Nohl, U. (2000). A global geochemical database structure for rocks: geochemical database structure. *Geochemistry, Geophysics, Geosystems*, *1*. https://doi.org/10.1029/1999GC000026.

LeRoy, M. A., Gill, B. C., Sperling, E. A., McKenzie, N. R., & Park, T.-Y.- S. (2021). Variable redox conditions as an evolutionary driver? A multibasin comparison of redox in the middle and later Cambrian oceans (Drumian-Paibian). *Palaeogeography, Palaeoclimatology, Palaeoecology*, *566*, 110209. https://doi.org/10.1016/j.palaeo.2020.110209

Lipp, A. G., Shorttle, O., Sperling, E. A., Brocks, J. J., Cole, D. B., Crockford, P. W., Del Mouro, L., Dewing, K., Dornbos, S. Q., Emmings, J. F., Farrell, U. C., Jarrett, A., Johnson, B. W., Kabanov, P., Keller, C. B., Kunzmann, M., Miller, A. J., Mills, N. T., O'Connell, B., … Yang, J. (2021). The composition and weathering of the continents over geologic time. *Geochemical Perspectives Letters*, *7*, 21–26. https://doi.org/10.7185/geochemlet.2109.

Mehra, A., Keller, C. B., Zhang, T., Tosca, N. J., McLennan, S. M., Sperling, E. A., Farrell, U. C., Brocks, J., Canfield, D., Cole, D., Crockford, P., Cui, H., Dahl, T. W., Dewing, K., Emmings, J., Gaines, R. R., Gibson, T., Gilleaudeau, G. J., Guilbaud, R., … Strauss, J. V. (2021). Curation and analysis of global sedimentary geochemical data to inform Earth History. *GSA Today*, *31*, 4–10. https://doi.org/10.1130/GSATG484A.1.

Peters, S. E., Husson, J. M., & Czaplewski, J. (2018). Macrostrat: A platform for geological data integration and deep-time earth crust research. *Geochemistry, Geophysics, Geosystems*, *19*, 1393–1409. https://doi.org/10.1029/2018GC007467.

Planavsky, N., Hood, A., Tarhan, L., Shen, S., & Johnson, K. (2020). Store and share ancient rocks. *Nature*, *581*, 137–139. https://doi.org/10.1038/d41586-020-01366-w.

Poulton, S. W., & Canfield, D. E. (2005). Development of a sequential extraction procedure for iron: Implications for iron partitioning in continentally derived particulates. *Chemical Geology*, *214*, 209–221. https://doi.org/10.1016/j.chemgeo.2004.09.003.

Rowe, R., Hughes, N., & Robinson, K. (2012). The quantification and application of handheld energy-dispersive x-ray fluorescence (ED-XRF) in mudrock chemostratigraphy and geochemistry. *Chemical Geology*, *324–325*, 122–131. https://doi.org/10.1016/j.chemgeo.2011.12.023.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*, 421–427. https://doi.org/10.1038/nature13595.

Sepkoski, J. J. (1993). Ten years in the library: New data confirm paleontological patterns. *Paleobiology*, *19*, 43–51. https://doi.org/10.1017/S0094837300012306.

Watson, C., Baker, G., & Nayembil, M. (2014). *Open Geoscience Data Models: End of project report [Publication - Report]*. British Geological Survey. http://nora.nerc.ac.uk/id/eprint/508791/

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Agerbo, E., Air, T. M., Andlauer, T. M. F., Bacanu, S.-A., Bækvad-Hansen, M., Beekman, A. F. T., Bigdeli, T. B., Binder, E. B., Blackwood, D. R. H., Bryois, J., Buttenschøn, H. N., Bybjerg-Grauholm, J., Cai, N., Castelao, E., … the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, *50*, 668–681. https://doi.org/10.1038/s41588-018-0090-3.

How to cite this article: Farrell, Ú. C., Samawi, R., Anjanappa, S., Klykov, R., Adeboye, O. O., Agic, H., Ahm, A.-S. C., Boag, T. H., Bowyer, F., Brocks, J. J., Brunoir, T. N., Canfield, D. E., Chen, X., Cheng, M., Clarkson, M. O., Cole, D. B., Cordie, D. R., Crockford, P. W., Cui, H., … Sperling, E. A. (2021). The Sedimentary Geochemistry and Paleoenvironments Project. *Geobiology*, 19, 545–556. https://doi.org/10.1111/gbi.12462