This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2022-8613. http://www.minsocam.org/

Revision 3—Submitted January 21, 2023

What is Mineral Informatics?

Anirudh Prabhu[1*], Shaunna M. Morrison[1], Peter Fox[2], Xiaogang Ma[3], Michael L. Wong[1], Jason R. Williams[1], Kenneth N. McGuinness[4], Sergey V. Krivovichev[5], Kerstin Lehnert[6], Jolyon Ralph[7], Barbara Lafuente[8], Robert T. Downs[9], Michael J. Walter[1], Robert M. Hazen[1]

[1]Earth and Planets Laboratory, Carnegie Institution for Science, 5241 Broad Branch Rd NW, Washington, DC 20015.

[2]Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th St, Troy, NY 12180.

[3]Department of Computer Science, University of Idaho, 875 Perimeter Dr, Moscow, ID 83844.

[4]Department of Biochemistry and Molecular Biology, Rutgers University, 57 US Highway 1. New Brunswick, NJ 08901-8554.

[5]Kola Science Centre of the Russian Academy of Sciences, Leninskiy Prospekt, 14, Moscow, Russia, 119991.

[6]Lamont-Doherty Earth Observatory, Columbia University, 61 Rte 9W, Palisades, NY 10964.

[7]Mindat.org, 1113 Cambridge Hill Lane, Keswick, VA 22947-2749.

[8]Carl Sagan Center, SETI Institute, 189 Bernardo Ave, Suite 200, Mountain View, CA 94043.

[9]Department of Geosciences, University of Arizona, 1040 E 4th St, Tucson, AZ 85721.

*** Email:** aprabhu@carnegiescience.edu

ORCID ID: 0000-0002-9921-6084

ABSTRACT

Minerals are information-rich materials that offer researchers a glimpse into the evolution of planetary bodies. Thus, it is important to extract, analyze, and interpret this abundance of information in order to improve our understanding of the planetary bodies in our solar system and the role our planet's geosphere played in the origin and evolution of life. Over the past several decades, data-driven efforts in mineralogy have seen a gradual increase. The development and application of data science and analytics methods to mineralogy, while extremely promising, has also been somewhat *ad-hoc* in nature. In order to systematize and synthesize the direction of these efforts, we introduce the concept of "Mineral Informatics," which is the next frontier for researchers working with mineral data. In this paper, we present our vision for Mineral Informatics and the X-Informatics underpinnings that led to its conception, as well as the needs, challenges, opportunities, and future directions of the field. The intention of this paper is not to create a new specific field or a sub-field as a separate silo, but to document the needs of researchers studying minerals in various contexts and fields of study, to demonstrate how the systemization and enhanced access to mineralogical data will increase cross- and interdisciplinary studies, and how data science and informatics methods are a key next step in integrative mineralogical studies.

Keywords: Minerals; X-informatics; Data; Data science; Information; Scientific discovery.

INTRODUCTION

The potential for data-driven methods to make novel, unintuitive, and groundbreaking discoveries in Earth and planetary science research will only grow as the volume and variety of data increases with time. Mineralogy, in particular, is ripe for the application of data-driven methods. Minerals form as a result of their unique chemical and physical conditions and, in the process, retain information regarding their formation that offers an opportunity to study the complex geologic and biologic past of planetary bodies (Prabhu et al. 2021b).

Mineralogy has been the subject of scientific curiosity and study for millennia (Agricola and Bandy 1955; Needham and Wang 1995; Bandy and Bandy 2004). In addition to their roles as captivating specimens for collection and study, minerals and their ores are essential in the survival and industrialization of humankind (Coates 1985; Murray 1995). This interest and utility has led to the characterization and systemization of mineralogy and mineral occurrence on Earth and other planetary bodies (Dana 1895; Bragg and Bragg 1913; Strunz and Tennyson 1941; Lehnert et al. 2000; Lafuente et al. 2015; Hazen and Morrison 2020). As a result of this rich history of scientific investigation, vast amounts of information are available on the occurrence and attributes of minerals. These data provide a robust platform for the analysis of more complex, multidimensional, and larger mineralogical systems; the integration of heterogeneous data types, linking to data from other fields of science; and predictive, data-driven scientific exploration - all of which leads to the answering of complex, multidisciplinary questions. The potential of data-driven mineralogical research has been exemplified by important scientific advances in the last decade. Recent discoveries have demonstrated periodicity of mineral formation and diversification associated with supercontinent assemble (Bradley 2011; Voice et al. 2011; Hazen et al. 2014; Nance et al. 2014), an association of mineral redox state to the oxidation of Earth's atmosphere

(Liu et al. 2021; Hummer et al. 2022; Large et al. 2022), and that much of Earth's mineral inventory is the direct or indirect result of interactions with water and/or biology (Hazen and Morrison 2020, 2022), as well as the prediction of the number of as-yet undiscovered mineral species (Hazen et al. 2015; Hystad et al. 2015, 2019), the chemical composition of minerals on Mars (Morrison et al. 2018c, 2018a, 2018b), and the location of undiscovered mineral deposits (Prabhu et al. 2019; Morrison et al. 2023). Mineralogy is rapidly entering the data-driven era, tackling previously unanswerable questions, while demonstrating the need and opportunity for a symbiotic relationship between mineralogy and the fields of data science and informatics.

Data-driven efforts in mineralogy have been gradually increasing in the past decades and there are some promising studies that have helped researchers uncover patterns hidden in the data—patterns that have led to scientific discoveries (Morrison et al. 2017, 2020; Gregory et al. 2019; Hazen et al. 2019; Prabhu et al. 2019; Hazen and Morrison 2020, 2022; Zhao et al. 2020; Boujibar et al. 2021; Hystad et al. 2021). While still nascent, application of data science and data analytics methods in mineralogy shows a promising trajectory, though the development of these methods and advances in the past have been somewhat ad-hoc in nature. However, development of mineral informatics can be guided in a more deliberate and systematic way by taking into account the underpinnings from information theory and data science advances, as exemplified by collaborations in other fields, including biology, medicine, chemistry, and astronomy. We believe this is the start of a new era in mineralogy, where utilizing data-driven methods to answer mineralogical (and broader scientific) questions takes center stage.

In this paper, we take a high-level look at our vision for "Mineral Informatics", the underpinnings that led to its conception, as well as the needs, challenges, and opportunities for this emerging field. We also discuss implications such advances will have on the field of mineralogy.

Informatics is the study of the structure, algorithms, behavior, and interactions of natural and artificial systems that store, process, access and communicate information (Fox 2011). The term informatics has often been used in conjunction with the name of a domain/discipline, for example, Bioinformatics, Geoinformatics, Astroinformatics, and Cheminformatics. In the past, researchers with expertise in a specific domain worked on processing and engineering information systems designed for that domain only. But in the last decade, informatics has gained a much wider visibility across a range of disciplines (Prabhu 2018). This wider visibility is in large part due to successful efforts at systematizing the core (i.e., discipline neutral) aspects of informatics, for example, use-cases, human-centered design, iterative approaches, and information models (Fox 2020). The core methods of informatics are used as a foundation to explore raw data and extract information from the data that lead to scientific discoveries. As the volume and complexity of the data increase, so does the need for utilizing the solid foundations provided by informatics methods and combining them with needs of the specific domain to pursue data-driven scientific discoveries.

Mineral informatics is a nascent approach compared to fields like Bioinformatics, Medical Informatics, and Geoinformatics that have been pursued for decades (Collen 1986; Fox et al. 2006; Sinha et al. 2010; Gauthier et al. 2019). The intention of this paper is not to create a new specific field or a sub-field as a separate silo, but to think of and document the needs of researchers studying minerals in various contexts and how data science and informatics methods are a key next step in mineralogical studies. We also need to learn from the successes and failures of more mature domains that have applied the informatics approach. Lastly, a very important factor to keep in mind is the truly interdisciplinary and important questions that can be explored by studying minerals. So, while the term “mineral informatics” may seem to be creating a new subclass of

geoinformatics, we assert that we are instead tying together various disciplines that use minerals as a key part of the pursuit for answers to big science questions.

A METHODOLOGY FOR MINERAL INFORMATICS EXPLORATIONS

In this paper, we present a general methodology for mineral informatics (see Figure 1). This methodology, adapted from Fox and McGuinness' Semantic Web Methodology (Beaulieu et al. 2017), includes all the steps typically followed in a data-driven scientific exploration. This approach was created for mineral informatics but, as is the case with many data science and informatics approaches, is transferable and applicable to other domains.

Most informatics explorations start in one of two ways: 1) Scientists have a research question they want to answer, or 2) scientists have data ready to be explored. In the second case, we perform preliminary data exploration, which helps generate new hypotheses and research questions based on interesting trends and anomalies in the data.

Once a specific research question has been selected for scientific exploration, we start by dividing the large problem into smaller more tractable parts. Next, we iteratively develop use cases for every one of these parts. A "use case" is a documented collection of possible sequences of actions and interactions between a system and its users in pursuit of a particular goal. Identification and development of use cases helps to define the needs (e.g., data, personnel, infrastructure) for this data-driven approach. The next steps in the methodology includes creating an (or assigning roles to an already established) interdisciplinary team to conduct the data-driven research.

Next, we inventory the preliminary dataset and/or existing mineral data resources to determine if they are what is necessary for the desired exploration. In some cases, we need to collect, compile, and extract data from other repositories or sources, including scientific literature, websites, digital

PDFs, and experimental results. We then create an information model to better understand and mediate data from heterogeneous sources and data types, which provides a holistic picture of the relationships among the various data sources, types, and attributes. The information model allows us to extract the datasets and data attributes most relevant to answering the desired research question. Note that this step differs from the statistical and machine learning approaches used for feature selection.

We then begin applying data analytics methods (i.e., data visualization as well as descriptive, predictive, and prescriptive analysis) to identify and explore patterns and anomalies seen in the data. A team of domain and data scientists iteratively examine the results of the analytics methods and use their respective expertise to (1) provide interpretations and/or insight, and/or (2) recommend changes to the analysis. The data analysis and scientific interpretation are usually done over multiple iterations with small modifications to the approach, algorithms, and/or code to explore different aspects of the data.

If scientists come to an agreement that parts of the analysis would be widely used in the larger community, then they can choose to generalize and adapt their work into a system, technology, or infrastructure. This development can include creation of tools, code snippets, reusable workflows, R packages, Python libraries, and other resources. Irrespective of whether there is a decision to create a general tool, technology, or package, we recommended using rapid prototyping coding practices (Gordon and Bieman 1995) for data science and informatics activities.

After obtaining the desired results from our data analysis, it is important to disseminate and effectively communicate the research products generated by mineral informatics explorations. Research products can include datasets, code, scientific literature, and executable workflows. Establishing best practices for disseminating research products is an ongoing effort, especially in

the geoscience community. Datasets can be published as part of a data paper or they can be assigned their own DOIs by data repositories such as Zenodo, Dryad, Figshare, or Dataverse (Assante et al. 2016). Existing mineral data repositories, including the EarthChem Library (ECL), Astromat, and the Open Data Repository (ODR) also provide DOIs for datasets deposited by researchers. Additionally, some journals host data associated with their publications. Similar to releasing data used in scientific exploration, code can be maintained and released in many ways, including Github (with a persistent identifier pointing to the repository), figshare, or Zenodo. Saving executable code for an experiment in an interactive environment like Jupyter or R notebooks adds to the reproducibility of the code and of the scientific workflow in general (Prabhu and Fox 2021). Dissemination of scientific advances through scientific publications has been practiced for more than 300 years (Fyfe et al. 2015). In addition to journal publications, conference proceedings, preprint servers (such as arXiv, ESSOAr, and EarthArXiv), and even press releases associated with publications have considerably improved the landscape of disseminating research products.

The final stage of our informatics methodology follows the sharing of the research products. If researchers follow FAIR (Findable Accessible Interoperable and Reusable) and Open Science practices (Wilkinson et al. 2016; Stall et al. 2019; Ramachandran et al. 2021) not only for the dissemination of their scientific results, but also during the use case development, information modeling, and analysis stages, then it becomes easier to evolve, improve, redesign, or adapt their work. Ongoing research and recommendations on designing FAIR and Open scientific workflows will help improve the methodology of data-driven exploration (Sandve et al. 2013; Kluyver et al. 2016; Prabhu and Fox 2021).

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2022-8613. http://www.minsocam.org/

It is important to evaluate the outcomes at almost every stage of the informatics methodology. The evaluation method or metric used at each stage will be significantly different, but it is important to stop at the end of every stage and assess not only the progress made, but also lessons learned for future iterations in the same exploration or the beginning of a different exploration. For example, a data collection/resource may be evaluated based on a set of quality criteria (e.g.,(Prabhu et al. 2021b)), but results from the data analysis may need to use quantitative metrics to evaluate results from a descriptive, prescriptive, or predictive model (e.g., (Statnikov et al. 2008; Hossin and Sulaiman 2015; Tomašev and Radovanović 2016; Zhou et al. 2021)). Established evaluation methods exist for each stage of the informatics methodology, and we recommend following those established best practices and standards set by the scientific community. Issues found during evaluation will need to be documented in the use case and thus improve the data-driven exploration during the next iteration or redesign of the approach.

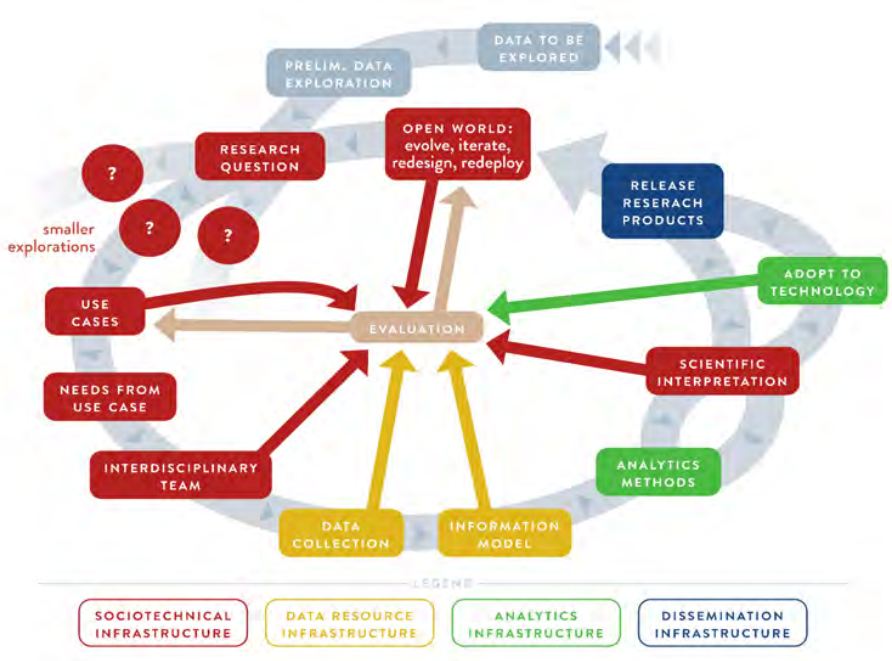

Figure 1. The mineral informatics methodology adapted from Semantic Web methodology by Fox & McGuinness' Semantic Web Methodology (Beaulieu et al. 2017).

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2022-8613. http://www.minsocam.org/

CHALLENGES AND OPPORTUNITIES IN MINERAL INFORMATICS

Mineral informatics methods not only systematize the mineral data landscape, but also provide a path to answering longstanding interdisciplinary scientific questions. Figure 2 gives an example of the domains influenced by the research questions being broached with mineral informatics methods. In the following section we outline some significant scientific questions that can be addressed with mineral informatics.

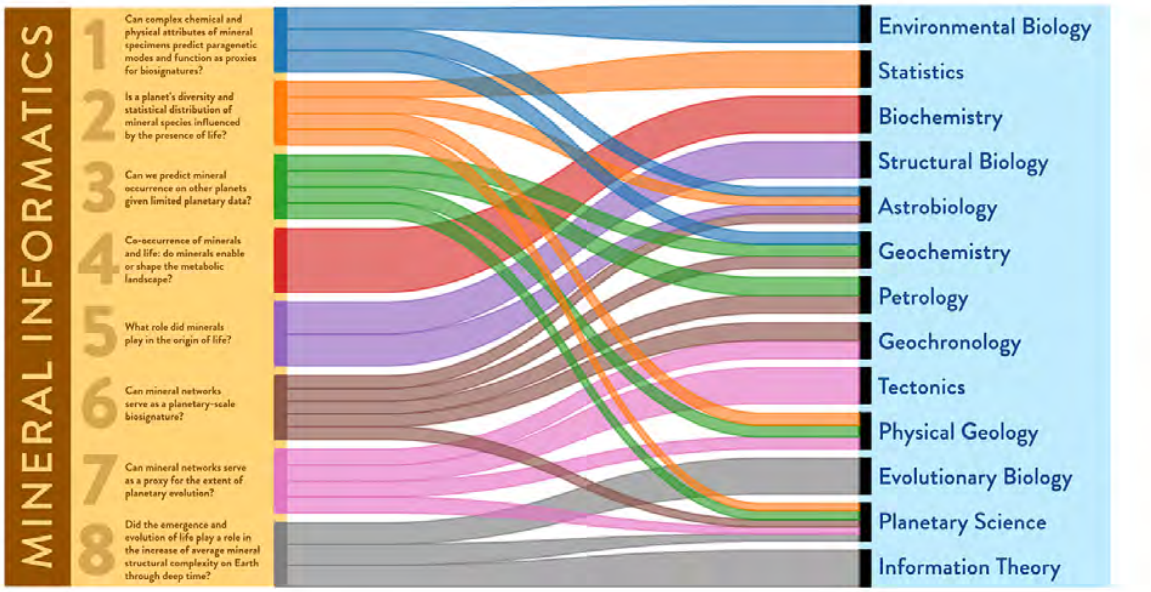

Figure 2: Interdisciplinary research questions related where mineral informatics play a key role.

Can complex chemical and physical attributes of mineral specimens reveal their paragenetic modes and function as proxies for biosignatures? Minerals record the physical, chemical, and, in some cases, biological conditions of their paragenetic modes (i.e., formational and alteration environments). This information is stored in the myriad attributes of mineral specimens, including major, minor, and trace elements; stable isotopes and their ratios; solid and fluid inclusions;

texture, twinning, exsolution, and other structural characteristics; grain size and shape; and much more. Therefore, conditions of mineralization, including whether or not there was biological input, can be characterized with cluster analysis performed on the various properties of mineral samples (Gregory et al. 2019). Furthermore, robust classification schemes can be developed from the clustering models that will enable prediction not only of the geologic environment of formation but also of any biogenic origins (Hazen 2019). Therefore, this work will deconvolve our understanding of the minerals that formed in environments influenced by life from those that formed under strictly abiotic conditions.

Is a planet's diversity and statistical distribution of mineral species influenced by the presence of life? Life creates unique niches of chemical disequilibrium for minerals to exploit. These processes likely drove a significant fraction of the mineral diversity we see on Earth today, influencing the spatial and temporal patterns of mineral distribution (Hazen 2018; Morrison et al. 2020; Hazen and Morrison 2022). These trends on Earth and other planetary bodies can be modeled, compared, and used to develop statistical biosignatures and abiosignatures that are reflected in the diversity and distribution of mineral species across a planetary body (Hystad et al. 2019) and provide models for planetary-scale mineralogical biosignatures of inhabited worlds.

Can we predict mineral occurrences on other planets given limited planetary data? From orbital infrared spectroscopy, we have obtained global or near-global datasets of the mineralogy of other terrestrial worlds, including Mars, Mercury, Vesta, and Ceres (Murchie et al. 2009; De Sanctis et al. 2012; Ehlmann and Edwards 2014; Namur and Charlier 2017; Prettyman et al. 2019). Informatics methods, such as association analysis, can be used to predict the existence of minerals

that cannot be detected from space. By understanding mineral affinities for assemblages, localities, and geochemical parameters, we may be able to use a sparse mineralogical dataset to anticipate future discoveries (Prabhu et al. 2019), but first a robust small/sparse-data framework must be developed. Enhancing predictive capabilities will help to prioritize landing sites for future landers and rovers with broad science goals that relate to mineralogy, like understanding planetary history or searching for signs of life. Such predictions would be strategically important because interplanetary missions cost hundreds of millions to billions of dollars and take years to decades to develop, build, and launch.

We also have geochemical indicators of the mineralogy of the ice-covered ocean world Enceladus from plume flybys and E-ring analyses performed by the Cassini spacecraft (Postberg et al. 2008; Waite et al. 2017; Glein and Waite 2020). Mineral informatics methods can help predict the mineral composition of ice-covered ocean worlds, whose mineralogy is planetologically and perhaps astrobiologically relevant but cannot be accessed directly in the near future.

Co-occurrence of minerals and life: do minerals enable or shape the metabolic landscape? Minerals play a key role in biological redox transformations. Many microorganisms (e.g., of the genus *Geobacter*) are able to use metals in their environment to power their metabolisms (Childers et al. 2002). Several studies have suggested deep similarities between minerals and metalloenzymes (Nitschke et al. 2013; Zhao et al. 2020; McGuinness et al. 2022). Thus, minerals may play an important role in shaping the metabolic landscape of ecosystems by providing electron donors/acceptors or raw materials (Novikov and Copley 2013) that organisms assimilate to create metalloenzymes. If minerals and their structures are found to be critical in shaping which metabolisms occur/do not occur in certain environments, these mimneralogical data may allow for

the prediction of metabolisms in terrestrial and extraterrestrial environments for which we have mineralogical data.

What role did minerals play in the origin of life? Several studies have posited that minerals played critical roles at the emergence of life on Earth, whether by catalyzing critical biomolecular reactions, templating the formation of biopolymers, influencing the homochirality of organic molecules, or performing redox transformations and carbon fixation (Hazen and Sholl 2003; Hazen 2005; Hazen and Sverjensky 2010; Nitschke et al. 2013; Russell 2018). Others have even suggested that clays and other minerals with layered structures may have been the first self-replicating entities (Cairns-Smith and Hartman 1986; Cairns-Smith 1990; Greenwell and Coveney 2006; Brack 2013), though these hypotheses have not been confirmed experimentally (Bullard et al. 2007; Krivovichev et al. 2011). Mineral informatics, combined with phylogenetics, geology, and laboratory experiments, could be informative for deducing the likely role(s) that minerals played at the origin of life in Earth's deep past. If certain minerals are found to be uniquely critical to the emergence of life on Earth, then this discovery would have profound implications for the emergence of life on other planetary bodies where those minerals may or may not occur. The origin of life from a non-living substance involves considerable jump in the informational (static) complexity of the underlying molecular structures, which should be taken into account in any possible scenario of molecular evolution/revolution that led to the appearance of self-replicating living entities. The sudden rise in structural complexity corresponds to the drop in configurational entropy (Krivovichev 2016). Can the (local) entropic changes associated with the origin of life be measured quantitatively and understood using mineral informatics data?

Can mineral networks serve as a planetary-scale biosignature? Roughly half of all known minerals are mediated by biology and 34% are exclusively biotic (Hazen et al. 2021, 2022; Morrison et al. 2021; Hazen and Morrison 2022). Many of these minerals are formed when life opens up a new compositional space for the planet, such as the Great Oxidation Event (Hazen et al. 2008; Sverjensky and Lee 2010). However, some of this biogenic chemical space may be abiotically accessed on other worlds. Abundant atmospheric O_2, for instance, may be abiotically generated by various star–planet interactions (Meadows et al. 2018) and references therein. Earth and planetary mineral network analysis may reveal whether mineral networks of environmental, biological, geochemical, and mineralogical attributes can distinguish living from nonliving worlds.

Can mineral networks serve as a proxy for the extent of planetary evolution? Mineralogical evolution occurs when processes create new pressure–temperature–compositional regimes where solids can form (Hazen et al. 2008, 2021; Hazen and Morrison 2020; Cleland et al. 2021). Each stage of mineral evolution expands the network of mineralogy through the introduction of new minerals, localities, and paragenetic modes. The network of martian mineralogy, therefore, is thought to be a subset of the network of Earth's mineralogy, due to the halting or slowing of mineral-generating geological processes on Mars. One can consider Mars and Earth to be two points along a spectrum of terrestrial worlds whose geological (and biological) activities have differed in temporal extent. A hypothetical world where plate tectonics was sustained for ~1 Gyr but then ceased should have a mineral network that surpasses Mars's mineral diversity, but is still a subset of Earth's. In this way, mineral informatics helps us interpret the extent of a planet's mineralogical network as a record of ancient and extinct processes, revealing a planet's geological history.

When considering exoplanetary systems where element ratios (e.g., C:O or Mg:Si) differ greatly from those of our own solar system, this linear spectrum on which Mars and Earth lie becomes a multidimensional phase space (Unterborn et al. 2016; Hinkel and Unterborn 2018; Unterborn and Panero 2019; Putirka et al. 2021). Understanding mineral networks from an informatics point of view may help to predict how planetary mineralogy might evolve in vastly different geochemical contexts.

Did the emergence and evolution of life play a role in the increase of average mineral structural complexity on Earth through deep time? It has been shown that complexity of Earth's mineral kingdom increased gradually during planetary evolution (Krivovichev et al. 2018), but it is unclear whether this trend is related to the contemporaneous increase in complexity in the course of biological evolution. The average structural complexity of minerals on the abiotic Moon, for example, does not follow the same trend of increasing complexity through time. Minerals are relatively less complex than biological organisms, both in terms of their static (Krivovichev 2013, 2015) and functional (Hazen et al. 2007) complexities. However, since life and the mineral kingdom co-evolved, the character of the evolution of mineral complexity on Earth (Krivovichev et al. 2018) may have been influenced by biological activity, and is thereby a potential bio-signature.

SUCCESSFUL USE CASES IN MINERAL INFORMATICS

Strategies for future advances in mineral informatics are informed by previous efforts—"use cases" that have applied data science analytics and visualization to tackle key mineralogical problems. In the following section we review five of these recent and ongoing studies.

The evolution of mineralizing environments, as characterized by their myriad, complex attributes: Mineralization, and associated formational environments, vary significantly across Earth and neighboring planetary bodies, as well as throughout the different historical stages of planetary evolution. These stages and environmental parameters dictate the types of mineralization that occur and, likewise, leave their mark in the complex chemical and physical attributes of the resulting mineral specimens. Understanding the changing characteristics of mineralizing environments spatially and temporally across our planetary systems requires the examination of huge volumes of mineralogical information. The beginning steps of this work included a survey of all formational environments of ~5700 known mineral species, resulting in a compiled dataset ripe for exploration (Hazen and Morrison 2022; Hazen et al. 2022). Initial exploration has led to the discovery that (1) more than 80% of all mineral species formed through processes that involved water; (2) 50% of minerals formed through processes directly or indirectly related to biology, with 34% of minerals forming exclusively through biotic processes; (3) 42% of minerals contain one or more rare elements (e.g., REE, PGE, As, Mo, Sn), elements which all together represent only 0.01% of crustal atoms; and (4) most minerals have only one (59%) or two (24%) modes of formation, with a few notable exceptions, including pyrite with the most modes of formation at 21 (Hazen and Morrison 2022).

An additional component of this work involves analyzing those myriad attributes of mineral specimens via cluster analysis to relate their complex characteristics to their modes of formation, thereby determining the natural kind clustering of these mineral systems. There are many such projects underway, including those examining the formation of pyrite (Gregory et al. 2019; Zhang et al. 2019), garnet minerals (Chiama et al. 2020, 2022a, 2022b in preparation), spinel oxide phases

(Hindrichs et al. 2022), and presolar moissanite (SiC) (Boujibar et al. 2021; Hystad et al. 2021). Boujibar et al. (2021) performed cluster analysis on a range of isotopic data from presolar SiC grains in order to examine and compare the origins of these materials. This study made several exciting discoveries - while the clustering model agreed with previously defined grain types and origins in several aspects, there were notable and important deviations, including: (1) a division of one previously defined grain type into three distinct kinds based on the varying metallicity of the parent star; (2) the arbitrary nature of certain prior divisions in systems that in fact are continuous rather than discrete; (3) the observation that asymptotic giant branch (AGB) stars with narrow ranges of mass and metallicity tend to have enhanced production of SiC; and (4) enrichments in ^{15}N and ^{26}Al that are not explained by existing AGB models.

Next steps: This exploration of mineralizing environments and their characteristics not only provides an opportunity to integrate data from heterogeneous sources and types (e.g., X-ray diffraction, electron microprobe analysis, inductively coupled plasma mass spectrometry), but also to link data from different fields of science to better understand mineral paragenesis. Handling heterogeneous data is a challenge (Reichman et al. 2011; Wang 2017) and many researchers have been actively working on using heterogeneous data for their analysis by creating methods, approaches, and pipelines to seamlessly clean, integrate, process, and analyze data (Wiederhold 1999; Beneventano and Bergamaschi 2004; Wang 2017; Zhang et al. 2018; Nazábal et al. 2020). Additionally, the exploration conducted by Boujibar et al. (2021) provided another use case to test machine learning methods on sparse data sets, thereby aiding in the eventual development of a sparse data framework.

Mineral association analysis: Prediction of the locations of as yet undiscovered mineral deposits has long been a point of great scientific and economic interest. Mineralization and mineral co-occurrence across the varied geologic terrains of Earth and other planetary bodies has a level of complexity that makes prediction of mineral locations, or even the mineral inventory at a locality of interest, difficult. However, recent advances in the mineral locality data resources (e.g., mindat.org and the Mineral Evolution Database) have provided an opportunity to begin tackling this tough problem with machine learning. Association analysis can be used to create a recommender system (Burke et al. 2011; Shah et al. 2017) that generates association rules based on known co-occurrences and these rules can be queried to determine the likelihood of currently unknown co-occurrences. In the case of minerals, we can query our mineral association rules to predict: (1) previously unknown locations of a mineral species, (2) previously unknown locations of mineral assemblages, including those that represent analog environments for study, and (3) the mineral inventory at a locality of scientific interest. The mindat.org team have conducted preliminary explorations using pairwise associations to predict the occurrence of certain minerals on Earth.

Next Steps: Mineral association analysis provides a powerful approach to new types of data problems. We need to modify the association analysis algorithms to better handle larger mineral occurrence datasets. For example, our models can currently handle only 2,473 minerals occurring in 87,306 localities (Prabhu et al. 2019; Morrison et al. 2023), but there are at present more than 5800 mineral species in the International Mineralogical Association's (IMA) list of approved mineral species (https://rruff.info/ima/, accessed 17 January 2023), which occur in more than 375,000 localities (https://www.mindat.org/stats.php, accessed 20 December 2022). In order to increase the scalability of the association analysis algorithm, we plan to introduce threshold checks

and additional parameters during the association rule generation process, so that the number of rules generated is controlled. In addition to improving the scalability of association analysis methods, we also need to work on the dimensionality and reducing the minimum support of our method. For example, our method currently develops rules containing 4 minerals at a time, but there are localities with more than 50 coexisting minerals. Therefore, an important next step in our research is to increase the dimensionality of the association analysis method to handle more complex mineral assemblages. We plan to reduce the number of rules in a rule base by better identifying redundant rules or similar rules, thus leaving more disk space for higher dimensional rules. We also need to adapt our methods to enable inclusion of rarer mineral species that are known to occur in 17 or fewer localities (Prabhu et al. 2019). We plan to include rarer mineral species by weighting the mineral occurrence by other factors including tonnage, its paragenetic mode diversity, and criticality of the mineral. Lastly, we are currently developing a new approach to evaluate association rule mining methods (Prabhu et al. 2021a).

Martian crystal chemistry: The scientific payload onboard the NASA Mars Science Laboratory (MSL) rover, *Curiosity*, is the one of the most advanced instrument suites ever landed on another planet. Part of this payload is the CheMin X-ray diffraction (XRD) instrument, which is used to characterize the mineralogy of rock and soil samples. CheMin is capable of identifying mineral phases present in samples, as well as their abundances and, for phases with an abundance $\geq$ 1 to 3 wt %, their unit-cell parameters. While there are instruments that analyze the bulk composition of martian samples, there is no instrument that directly measures the chemical composition of these mineral phases. However, in compiling data resources on mineral unit-cell parameters and compositions measured on Earth, the CheMin XRD patterns and resulting mineralogical data are

used to predict the composition of the mineral phases observed on the martian surface (Morrison et al. 2018c, 2018a).

These initial studies, as with many investigations predating it, used unit-cell parameters to predict mineral composition in chemically limited systems, generally 2- or 3-element systems such as Fe-Mg olivine or Mg-Fe-Ca pyroxene (Morrison et al. 2018c, 2018a). This limitation was due to the complexity of the compositional and structural parameter space when four or more elements are considered together. One way to develop a model that accounts for the complexity associated with multi-component systems and predicts the chemical composition of crystalline phases based on their crystallographic parameters is by using Label Distribution Learning (LDL) (Geng et al. 2013, 2014; Geng 2016). LDL is a machine learning algorithm originally created for facial recognition applications. When the approach was adapted for application to crystallographic and chemical parameters, it resulted in a model that accurately predicted the multi-component chemical compositions (up to 12 elements, in some mineral systems) of samples based solely on their unit-cell parameters (Morrison et al. 2018b). This crystal-chemical method has expanded the capability of XRD on spacecraft to that of a powerful chemical analysis tool, such as an electron microprobe, and has dramatically deepened our understanding of the geologic history of Mars.

Next steps: This exploration was the initial inspiration that motivated us to create a framework for small and sparse data. In addition to our work developing a framework for small and sparse data, we will also need to develop methods to evaluate the accuracy of predictions made by our data models. This evaluation will attempt to address sources of uncertainty and how that affects our predictions. The LDL evaluation method being developed will address uncertainty of measurement (instrument errors), uncertainty from sampling (various sampling strategies to train predictive models), and most interestingly, scope compliance (Kläs 2018) of the LDL method.

Machine learning majorite barometer: Diamond-hosted majoritic garnet inclusions provide important insights in processes that occur in Earth's deep mantle. Majoritic garnets provide the most accurate estimates for diamond formation pressures because laboratory experiments have shown that garnet chemistry varies as a function of pressure (Akaogi and Akimoto 1977; Irifune 1987; Collerson et al. 2010; Wijbrans et al. 2016; Beyer and Frost 2017; Thomson et al. 2021). Thomson et al. (2021) show that none of the available barometers in the literature reliably reproduces the pressures of experimentally synthesized majoritic garnet over the entire pressure-temperature-composition space investigated. Hence, they developed a barometer by using machine learning algorithms (specifically random forest regression) and experimental training data. This machine learning approach, tested with various cross-validation methods, produces a barometer with a much-improved fit to the experimental data, especially at the highest pressures and at extremes of composition space, and thus provides more reliable estimates of formation pressures of diamond-hosted majoritic inclusions. Applying the machine learning barometer to the global database of diamond-hosted inclusions reveals that their formation occurs over specific depth intervals that can be related to melting and decarbonation of subducted oceanic crust.

Next Steps: While the machine learning approach improved the fit to the available experimental data, it also revealed regions in pressure, temperature, and most critically, composition space where the experimental data set is sparse. Because many of the mineral inclusions have compositions lying near or within sparse data regions, uncertainty remains as to whether the barometer is accurately capturing their pressure (and depth) of origin. Experiments can now be targeted to these specific P-T-X regimes for an even more improved barometer. Machine learning methods also can be used to predict the compositional variables that correlate most strongly with

changes in pressure, leading to an improved crystal chemical and thermodynamic understanding of pressure-sensitive substitutions in garnet. These methods can also be applied to other mineral thermometers and barometers where large experimental datasets are fitted to extract thermodynamic solution parameters.

Comparison of mineral and protein metal clusters: Understanding the evolutionary stages of biology on a geological timescale is hampered by the propensity of organic matter to degrade within thousands of years without leaving physical fossil records. To understand how life evolved over the course of billions of years, proxy data are required.

At least five observations suggest that minerals can act as a source of proxy data from which to infer how biology evolved: (1) biology and geology are intimately connected, for instance, cellular organisms excrete minerals as metabolic end products (hazenite; (Yang et al. 2011); greigite; (Gorlas et al. 2018)) and cellular organisms transmit electrons to and from minerals (Shi et al. 2016); (2) cellular organisms and minerals use transition metals (Fe, Mn, Co, Mo, Cu, V, W, Ni) to perform electron transfer reactions; (3) mineral surfaces are hypothesized and shown to be capable of prebiotic reactions similar to those that extant proteins perform (Wächtershäuser 1988; Novikov and Copley 2013); (4) minerals are similar to the rings of a tree in that they provide information (e.g., temperature, humidity, etc.) about the environment of formation; and (5) metal cluster structures of extant proteins were observed to be so similar to the structure of bulk mineral metal clusters as to be considered vestiges of minerals that were co-opted and assimilated into biological systems (Russell and Hall 1997; Nitschke et al. 2013; Zhao et al. 2020).

Access to large mineral and protein structure databases allows the potential to understand how mineral and protein metal clusters are connected. Connecting the mineral world with biology will

allow a deeper understanding of how geology and biology co-evolved. Directly quantifying metal cluster similarity between minerals and proteins is a challenge due to comparing the finite protein cluster to a periodic lattice of a mineral. Solutions using graph-based methods have been proposed (Zhao et al. 2020; McGuinness et al. 2022). Each solution compared subgraphs of mineral and protein metal clusters, however without including metal coordination, and mineral dimensionality (2D-layer vs 3D lattice) metal clusters were quantified as being highly similar (Zhao et al. 2020). Subsequent studies, building off the quantitative pioneering work of Zhao et al. (2020), included these chemically important characteristics and found FeS minerals and protein were significantly less similar (McGuinness et al. 2022) than previously proposed (Russell and Hall 1997; Nitschke et al. 2013) Even though McGuinness et al. 2022 show that FeS mineral lattices and protein metal clusters are not structurally similar, this method has not been applied to other metal types such as Ni or Cu. Applying the method developed by McGuinness et al. 2022 to additional metal types may help understand the extent to which proteins and minerals co-evolved as cellular metabolism and minerals became more complex (Moore et al. 2017; Krivovichev et al. 2018).

Next Steps: An additional step towards a potentially clearer understanding of how minerals and proteins are related is to compare mineral surface and protein metal cluster structures. Mineral surfaces expose the chemically active components that may have catalyzed biologically relevant products under hydrothermal conditions on early Earth (Novikov and Copley 2013). Comparing the surface properties of minerals to the chemical properties of protein metal clusters might elucidate the extent to which minerals acted as primitive enzymes at the dawn of life. Did biology co-opt the chemical configuration of the chemically active surface of minerals to reproduce the reactions that were possible abiotically? Or did biology incorporate and reconfigure metal building blocks (e.g., Fe_2S_2) to meet growing cellular needs? Answering these questions is challenging

because mineral surfaces are complex, i.e., they are subject to structural relaxation, chemically active, display complexly irregular surface topologies, and are affected by many solution conditions (pH, salinity, temperature, etc.) Alternatively, there also exists the possibility that protein metal clusters do not bear any significant resemblance to minerals (neither surface, nor lattice structure), suggesting an alternative pathway and relationship between mineralogy and biology in which biology acts independently, only relying on minerals for the feedstock (i.e., metals) to nucleate the information-rich systems that remain far from equilibrium.

MINERAL INFORMATION SYSTEMS

Table 1 is a non-exhaustive list of open access mineral data resources that are among the most widely used in the community. Note that many other useful and important mineral data resources are not yet available as open resources.

The global research community of mineralogy has made impressive progress on information models for database construction and data sharing in the past decades. From the point of view of data management, a good information model should be correct, complete, and consistent. An effective way for information modeling in real-world practice is to follow or adapt existing community agreements or standards on mineralogy, such as those on the physical, chemical, and biological characteristics of minerals. For instance, the Database of Mineral Properties (https://rruff.info/ima/) maintained by the International Mineralogical Association (IMA) keeps an up-to-date list of mineral species. The main components in the information model include mineral name, chemistry, mineral groups, origins, paragenetic mode, IMA status, relevant references, and links to external sources such as mindat.org, Google Images, and Wikipedia.

As open data and data-driven studies are increasingly accepted in the geoscience community, many databases in the field of mineralogy also help in increasing the visibility of their information model and building machine interfaces for data query, access, and download. For instance, the RRUFF database (https://rruff.info, accessed 21 January 2023) has integrated records of Raman spectra, X-ray diffraction, and chemistry data for minerals. The user interface enables data query through mineral name and chemistry includes/excludes. Interested users can also contact the database manager for batch data download and sharing. Mindat.org (https://www.mindat.org, accessed 21 January 2023) is another widely used database in the field of mineralogy. Its construction and maintenance follow a crowd-sourcing style. Besides the physical and chemical attributes of mineral species, a unique attribute on mindat.org is a comprehensive list of the localities where that mineral species has been found. In the past years, many research activities have benefited from the open data shared by mindat.org. As each of those open databases has its own focus and information model, scientists in large-scale research activities often need to collect data from multiple sources. Recently, researchers in geoinformatics and data science also discussed the need for a more comprehensive mineral information model to document the extensive facets of mineral data, such as the Global Earth Mineral Inventory (GEMI) proposed by (Prabhu et al. 2021b). Complementing these efforts are initiatives using semantic technologies to build knowledge graphs for mineral species, as a preparation to explore new ways for annotating and discovering mineral data shared on the Internet (Brodaric and Richard 2020).

The FAIR (findable, accessible, interoperable, and accessible) data principles (Wilkinson et al. 2016) are now widely accepted in geoscience. Information models are an important part of FAIR data. More community efforts, such as through IMA, the Mineralogical Society of America

(MSA), and the Geoinformation Committee of the International Union of Geological Sciences (IUGS-CGI), are needed to promote the quality and usefulness of the model outputs.

INFORMATICS INNOVATIONS NEEDED FOR MINERALOGY

The previous sections of this paper (and many other informatics papers focusing on various domains) have clearly emphasized the value that informatics methods provide to their respective domains (Collen 1986; Lord et al. 2004; Goble and Stevens 2008; Gauthier et al. 2019; Heberling et al. 2021). However, a point often missed or overlooked in scientific literature discussions is that innovations in data science and informatics are usually driven by diverse datasets available in various domains and the needs of the use-cases utilizing those datasets. In this section we discuss some of the interesting data science challenges we have observed while working with mineral data to try to answer some of the unanswered questions in geoscience.

In the following section we summarize four examples of mineral data challenges that provide interesting and unique problems that limit the usability of existing machine learning methods meant to extract meaningful information from data.

Small and sparse data framework: It has been widely publicized that we live in the "Age of Big Data" (Borgman et al. 2008; Lohr 2012; Wise and Shaffer 2015; Yu 2016; Wachter 2019), and understandably there has been a lot of research done into scaling-up algorithms, methods, software, and hardware needed to enable the exploration and use of very large datasets to gain valuable information. This focus has led to the creation and constant improvement of "big data frameworks," which provide a roadmap on how to work with large datasets. However, mineralogy, along with many other fields in Earth and planetary sciences, provide a plethora of small and sparse datasets that do not fall into the realm of big data. These datasets therefore require the application

of methodologies that lie outside the focus of traditional big data researchers. The next major hurdle for mineral informatics (and geoinformatics in general) is to work towards creating a framework for small and sparse data.

For example, mineral data collected by the CheMin X-ray diffractometer onboard the Mars Science Laboratory (Morrison et al. 2018c; Rampe et al. 2018) have few data points, having analyzed ~40 samples, each with around a dozen mineral species (as of January 2022). The CheMin team used small (on the order of dozens to a few hundred data points) datasets of mineral composition and associated unit-cell parameters to build models capable of predicting the basic chemical composition of major mineral phases observed on Mars, based solely on their unit-cell parameters (Morrison et al. 2018b, 2018c). However, the team wished to push their chemical prediction further, to predict complex, multi-element mineral compositions for the martian crystallographic data. In order to do so, (Morrison et al. 2018b) assembled datasets of laboratory-analyzed complex, multi-element mineral compositions and unit-cell parameters, which contained only a few hundred data points for each of the major mineral groups identified by CheMin. (Morrison et al. 2018b) used the small data Label Distribution Learning approach to predict complex chemical compositions (up to 12 elements, in some mineral systems) of mineral samples collected by the CheMin instrument based on the unit-cell parameters of these samples. Significantly more work can be done here to increase the accuracy and performance of these models and such complex datasets with small sample sizes provide an interesting and rare challenge to data scientists.

Mineral geochemistry often contains information related to the geologic, chemical, and/or biological processes and materials that went into their formation and any subsequent weathering and alteration. However, geochemical data are inherently sparse due to chemical variability in

geologic deposits and materials, different elemental affinities amongst different mineral species, and analytical bias introduced by research aims or instrument limitations. The resulting frequency of "missing values" makes many geochemical datasets unsuitable for use with existing algorithms designed for complete or near-complete datasets. A prime example of the sparseness of geochemical data is the garnet dataset compiled by (Chiama et al. 2020, 2022b in preparation), which contains over 95,000 geochemical analyses of garnet group mineral samples collected from a variety of sources, ranging from large repositories (EarthChem, RRUFF, MetPetDB) to individual peer-reviewed literature. Even a compiled and curated dataset such as this is considered sparse, largely due to the chemical variability amongst the various garnet mineral species, resulting in missing values in the chemical compositions of these samples (Chiama et al. 2022a). For example, of the 95,000 analyses compiled, only 5 major elements (Mg, Fe, Ca, Al, and Si) are present and/or reported in most samples, while other elements, including Mn, Cr, and Ti, are much less common throughout the dataset. An additional contribution to this sparseness is that studies may not analyze for all elements in a sample (e.g., limited to elements of interest, difficulty measuring light elements), resulting in missing values for which it is not known whether that element is present. Thus, while analyzing these data (using descriptive, prescriptive, or predictive methods) we need to take into account these missing values and their effect on the results. Sparse data is not a problem new or unique to mineral data (Greenland et al. 2000, 2016; Sweeting et al. 2004; Rogers et al. 2018), but, as is the theme for the rest of this paper, we must learn from the successes and failures of other domains in addressing sparse data (Katz 1987; Shepperd and Cartwright 2001; Uzuner 2009; Derczynski et al. 2013).

Other examples of small and sparse data challenges can be encountered in efforts to understand other planets and moons including Venus and Titan through their mineralogy and geochemistry.

Frigid Titan's exotic mineralogy, with water ice as a principal rock-forming mineral, oceans of liquid hydrocarbons, and varied postulated organic minerals, is mostly understood through laboratory analogs (Fegley et al. 1992; Bullock and Grinspoon 1996; Hashimoto and Abe 2005; Treiman and Bullock 2012; Gilmore et al. 2017; Hazen 2018; Maynard-Casely et al. 2018; Zolotov 2018; Cable et al. 2021).

Small and sparse datasets are a common occurrence in Earth and planetary science. Despite the limitations of the available information, the answers to key scientific questions are tied to these datasets. Therefore, an effort to create a framework to handle small and/or sparse data will be highly beneficial to scientific research in Earth and planetary science. Many researchers are working on "High-Dimensional, Small Sample Size" (HDSSS) or "High-Dimensional, Low Sample Size" (HDLSS) and its use in data analytics (Hall et al. 2005; Golugula et al. 2011; Yata and Aoshima 2012; Liu et al. 2017; Shen et al. 2017). However, this area of research has received much less attention compared to its big data counterpart, and hence has lacked the synthesis and generalization that comes with the popularity and maturity of well-established fields. The aforementioned examples, clearly demonstrate how such a framework would open paths for exploring very important scientific questions within and beyond mineralogy.

Data discovery: An increasing trend of data science in recent years is doing research with open data shared by others (Fox and Hendler 2014). Several recent scientific advances in mineral informatics also reflect that trend (Hazen et al. 2019). From the point of view of data users, an ideal situation is that they can efficiently find data portals on the Internet, datasets on the portals, or subsets of the data. In comparison, from the point of view of data providers and data managers, they need to organize the data with shared community standards, detailed metadata, and persistent and stable facilities to increase the reusability. As illustrated in the FAIR data principles for open

data (Wilkinson et al. 2016), the first two key points to consider are the findability and accessibility of data. Correspondingly, three key technical items arise here. The first item is the metadata schema for describing the datasets. While there are many common-purpose metadata schemas, such as the Dublin Core, for describing datasets, for domain-specific data such as those in mineralogy there can also be specific metadata elements. The second item is the identifier for the datasets. Similar to the Digital Object Identifier (DOI) for publications, datasets shared on the Internet should also have specific identifiers to enable persistent and stable discoverability. The third item with respect to findability and accessibility is the protocol for retrieving metadata through the identifier of datasets. Community efforts such as DataCite (Brase 2009) have made solid progress toward that goal. Nevertheless, the wide implementation of those best practices for open data in geosciences, including mineralogy, still need more time. It is also important to remember that appropriate scientific credit must be given at every stage of informatics methodology, from the acquisition of data to data analytics, and finally the dissemination of the research products produced by the data analysis.

A very recent technical development regarding data discovery is the Dataset Search Engine released by Google (Brickley et al. 2019), which is able to index millions of datasets on thousands of data portals, including their identifiers or Web links. End users of the dataset search engine (https://datasetsearch.research.google.com, accessed 21 January 2023) have integrated access to thousands of data portals. When a dataset is found on the engine, users can go to its original data portal page through the identifier or Web link and then download. The Google Dataset Search Engine is built on the top of Schema.org, which is designed as a comprehensive metadata schema for annotating digital objects on the Web. The annotated objects, such as datasets, will then be indexed by the search engines. As its usage is expanding, Schema.org also provides space for

extending the metadata elements of certain objects. A potential here is to have specific metadata elements designed for datasets of mineralogy, and this should be based on community collaboration. In the past few years, the EarthCube community has leveraged a list of open geoscience data portals to develop the GeoCODES search engine (https://geocodes.earthcube.org, accessed 21 January 2023). It is also based on Schema.org but has made extensions specifically for the registration and discovery of geoscience data. Any future efforts on the findability and accessibility of open mineralogy data can significantly benefit from the technical structure and experience of GeoCODES. Community agreements and standards, such as those developed by IMA, MSA, and IUGS-CGI, as well as best practices in existing data portals, such as those in RRUFF and mindat.org, will also be helpful to enrich the metadata of open mineralogy data.

Data processing: Dozens of data repositories contain a wealth of mineralogical information from which large data resources can be extracted. Web-scraping algorithms allow for the retrieval and storage of large amounts of data from web sources (Glez-Peña et al. 2014; Zhao 2017). Scraping algorithms in scripting languages such as Python or R allow users to extract and compile large amounts of data from web sources or journal articles in minutes or seconds, but the structure (or lack thereof) of web pages can slow the production of new data resources. Open-access mineral databases tend to be very contributor friendly; thus, users can pick and choose which data to include for a particular entry. Recognizing the inconsistencies in the storage and representation of mineral attribute data within and across different mineralogical databases is essential when compiling large mineral datasets from open data sources.

Webpages associated with Webmineral and Mindat have hierarchical structures made up of Hypertext Markup Language (HTML), Extensible Markup Language (XML), or Cascading Style Sheet (CSS) elements that allow for the selection of nodes that can contain specific data a user is

interested in (Gunawan et al. 2019). The ubiquitous occurrence or rarity of a mineral, relative interest among the scientific community in a mineral, as well as the age of discovery cause significant differences in the amount of information available for a mineral, driving the differences in the structure of these webpages. Webpages and digital PDFs associated with the *Handbook of Mineralogy* (Anthony et al. 1990-2003) have very little structure, which places more importance on the use of keywords (e.g., space group or crystal system) or separators (e.g., each mineral attribute or property introduced may have a semicolon preceding the associated description) in the compilation of data. Nested conditional statements (i.e., if-else statements) are useful for compiling data from web databases that have variable or no structure, but this approach can be more time-consuming and prone to error. Some headers may be reused such as "beta (ß)" which is used as a descriptor of the refractive indices in biaxial minerals (Frazier et al. 1963; Gunter and Ribbe 1993) and it can also refer to the geometry of the unit-cell of dimensions (Grove and Hazen 1974; Nesse 2013).

Quantifying and correcting bias: Critical to all of these aspects of data resource development and use is an understanding of and, where possible, modeling of the biases that exist in each of these systems. For example, significant biases occur in mineral sampling based on the physical appearance of the phase (e.g., large, brightly colored, euhedral crystals), the economic value, the scientific interest, proximity to major universities or research centers, and analytical technology. Such biases can be corrected with models of each of these parameters (Hazen et al. 2015; Grew et al. 2017; Hystad et al. 2019). Natural preservational biases is more complex, as it involves geologic history and mineral properties (e.g., chemistry, solubility, hardness), but work is underway to begin unraveling the history of preservational biases in mineral systems on Earth and other planetary bodies (Liu et al. 2019).

INFORMATICS RESEARCH AS A SOCIO-TECHNICAL SYSTEM

Research in the field of informatics is heavily dependent on interactions between data scientists and domain scientists (e.g., mineralogists, planetary scientists) (Ma et al. 2017). Conducting and applying informatics research is very much a socio-technical system (Fischer and Herrmann 2011). It is as much about the researchers, their interactions, the hypotheses generated, and interpreting of results from visualizations or models as it is about the data, the algorithms, and the models. Collaborations in informatics include many iterations between data and domains scientists, starting from data explorations and the problem formulation to interpreting the results and documenting the scientific insights learned from the data.

We recommend starting an informatics exploration with an in-person or virtual "datathon" (Anslow et al. 2016; Fritz et al. 2020). During this datathon, which usually lasts a day or two, collaborators mainly focus on nine aspects: (1) Interactions and discussions among data scientists and domain scientists to frame their goals and expectations; (2) documenting the research questions to be explored; (3) collating the data resources required to explore the documented research questions; (4) exploring the methods needed to examine the data (both analytically and visually); (5) constructing a roadmap for dividing the research question and tasks into smaller, more tractable parts; (6) leveraging descriptive, prescriptive, and predictive methods to gain preliminary insights from the data; (7) forming short-term and long-term goals based on the preliminary results; (8) documenting the shortcomings of the methods explored and why these roadblocks hamper scientific exploration; and (9) documenting the innovation needs of both data science and domain science methods to overcome the previously documented hurdles.

Not all of these steps need to be done during the two-day datathon; steps 1 and 2 can be completed beforehand. The main goal of conducting a datathon is to expedite and streamline the initial data exploration to gain preliminary results that can be examined by the domain scientist, while also allowing the data scientist to explore and understand the intricacies of the data at hand. Additionally, all collaborators gain an understanding of the shortcomings, needs, and opportunities of their data and of the current methods to address the desired scientific questions. This inventory of needs and opportunities in both the data science and domain science can result in a datathon output of a list of projects and publications spurred by the creative and iterative processes of this closely collaborative effort.

After the initial datathon, each collaborator (or group of collaborators) has a plan of action for the projects and subtasks within the project they are leading. Subsequent communication and collaboration usually follows the preferred working model of the team. For example, weekly meetings between the group to discuss advances in the project, or email communications between the team for the same purpose. The steps taken after the datathon and methods to communicate and collaborate change depending on the work style and comfort levels of the collaborators. General recommendations for this step include "science of team science" best practices advocated by many communities (National Research Council 2015).

IMPLICATIONS: A VISION FOR THE FUTURE

Durable and information-rich, minerals are the only ancient relics that offer direct, solid glimpses of eons of planetary transformation (Hazen et al. 2022). It is important to extract the abundance of information contained in these mineral samples to improve our understanding of the evolution of our planet, our solar system, and the role our planet's evolving geosphere played in

the origin and proliferation of life. Key synergistic aspects of the ongoing paradigm shift in mineralogy includes systematic efforts to collect and curate mineralogical information in data resources that enable open and widespread dissemination, and the use of those data to make scientific discoveries.

As mentioned earlier in this paper, informatics methods have been followed, implemented, and improved upon in other fields over the past decades. The concept of "X-informatics" has also been around since its first conceptualization in 2007 (Gray and Szalay 2007; Hey 2009), and over the past decade there has been a steady decline in researchers conducting informatics research in the silos of their respective fields. When planning for a new paradigm like mineral informatics, it is important to learn from successes and failures of more mature fields of informatics (Lord et al. 2004; Goble and Stevens 2008; Heberling et al. 2021) and modify the methods developed by past researchers to apply them to comprehensively address our needs as a community.

Over the last decade, there have been some efforts at collating various data resources in the geosciences and providing these data to researchers with minimal barriers and maximum interoperability. These efforts include Onegeology (Jackson 2010), Onegeochemistry (Chamberlain et al. 2021; Wyborn et al. 2021), and Onestratigraphy (Wang et al. 2021). The Onegeochemistry initiative also includes plans to develop best practices for FAIR geochemical data, governance models to ensure participation and trust, and a business model to ensure long-term sustainability (https://www.earthchem.org/communities/onegeochemistry/; accessed 21 January 2023). Efforts to improve the access, usage, and impact of mineral data resources can learn from the successes and challenges faced by such global initiatives. Developing a set of best practices and recommendations for creating, linking, and releasing mineral data would improve

the mineral data landscape and make it easier for researchers to produce and use mineral data without too many barriers.

Just as increasing the findability, accessibility, interoperability, reusability, and other important aspects of mineral data management and stewardship, obtaining scientific insights from mineral data using data-driven methods are another key facet of mineral informatics. For this, too, we can look to and learn from the success and failures of other domains applying informatics methods to answer their research questions. We hope the research directions for informatics and other fields like mineralogy, planetary science, and other related fields using mineral data that have been documented in this paper act as an initial step towards the ultimate goal of systematizing data driven scientific exploration using mineral data.

Mineralogy is facing new opportunities and challenges with the increased interest in and applications of data-driven methods. We believe the next paradigm for the field of mineralogy is that of mineral informatics. Mineral informatics focuses on deciphering the patterns and trends hidden in mineralogical, geochemical, and related data and using these patterns to answer scientific questions, thus making important new discoveries. In this paper, we have shown how the study of minerals is essential to improving our understanding of the evolution of our planet, our solar system, and more. We present a broad methodology for the study and use of mineral informatics methods and document the needs of the field and important scientific questions that may be answered using mineral informatics. We reiterate the symbiotic relationship between data scientists and domain scientists (e.g., mineralogists, planetary scientists, biologists) to make continuous and sustainable scientific progress.

In summary, our vision for the next decade of mineralogical research is built upon the systematic and coordinated study of mineral data and of the data science methods used to gain scientific insights.

ACKNOWLEDGMENTS

We thank Editor Dr. Don Baker, Associate Editor Dr. Jennifer Kung, and the two anonymous reviewers for their thorough, thoughtful, and constructive reviews.

FUNDING

This publication is a contribution to the 4D Initiative and the Deep-time Digital Earth (DDE) program. Studies of mineral evolution and mineral ecology have been supported by the Alfred P. Sloan Foundation, the W. M. Keck Foundation, the John Templeton Foundation, NASA Astrobiology Institute (Cycle 8) ENIGMA: Evolution of Nanomachines In Geospheres and Microbial Ancestors (80NSSC18M0093), a private foundation, and the Carnegie Institution for Science. Any opinions, findings, or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration.

REFERENCES

Agricola, G., and Bandy, J.A. (1955) De Natura Fossilium. Geological Society of America.

Akaogi, M., and Akimoto, S.-I. (1977) Pyroxene-garnet solid-solution equilibria in the systems $Mg_4Si_4O_{12}$—$Mg_3Al_2Si_3O_{12}$ and $Fe_4Si_4O_{12}$—$Fe_3Al_2Si_3O_{12}$ at high pressures and temperatures. Physics of the Earth and Planetary Interiors, 15, 90–106.

Anslow, C., Brosz, J., Maurer, F., and Boyes, M. (2016) Datathons: an experience report of data hackathons for data science education. Proceedings of the ACM Technical Symposium on Computing Science Education, 47, 615–620.

Anthony, J.W., Bideaux, R.A., Bladh, K.W., and Nichols, M.C. (1990-2003) Handbook of Mineralogy, 6 volumes. Mineral Data Publishing.

Assante, M., Candela, L., Castelli, D., and Tani, A. (2016) Are scientific data repositories coping with research data publishing? Data Science Journal, 15, 1-24.

Bandy, M.C., and Bandy, J.A. (2004) De Natura Fossilium (Textbook of Mineralogy). Courier Corporation.

Beaulieu, S.E., Fox, P.A., Di Stefano, M., Maffei, A., West, P., Hare, J.A., and Fogarty, M. (2017) Toward cyberinfrastructure to facilitate collaboration and reproducibility for marine integrated ecosystem assessments. Earth Science Informatics, 10, 85-97.

Beneventano, D., and Bergamaschi, S. (2004) The MOMIS methodology for integrating heterogeneous data sources. In R. Jacquart, Eds., Building the Information Society, 156, p. 19–24. IFIP International Federation for Information Processing, Springer, Boston, MA.

Beyer, C., and Frost, D.J. (2017) The depth of sub-lithospheric diamond formation and the redistribution of carbon in the deep mantle. Earth and Planetary Science Letters, 461, 30–39.

Borgman, C.L., Abelson, H., Dirks, L., Johnson, R., Koedinger, K.R., Linn, M.C., Lynch, C.A., Oblinger, D.G., Pea, R.D., and Salen, K. (2008) Fostering learning in the networked world: The cyberlearning opportunity and challenge. A 21st century agenda for the National Science Foundation. Report of the NSF Task Force on Cyberlearning. No. nsf08204, 1-65.

Boujibar, A., Howell, S., Zhang, S., Hystad, G., Prabhu, A., Liu, N., Stephan, T., Narkar, S., Eleish, A., and Morrison, S.M. (2021) Cluster analysis of presolar silicon carbide grains: Evaluation of their classification and astrophysical implications. The Astrophysical Journal Letters, 907, L39.

Brack, A. (2013) Clay minerals and the origin of life. Developments in Clay Science, 5, 507–521.

Bradley, D.C. (2011) Secular trends in the geologic record and the supercontinent cycle. Earth-Science Reviews, 108, 16–33.

Bragg, W.H., and Bragg, W.L. (1913) The reflection of X-rays by crystals. Proceedings of the Royal Society of London, A88, 428–438.

Brase, J. (2009) DataCite—A global registration agency for research data. International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 4, 257–261. IEEE.

Brickley, D., Burgess, M., and Noy, N. (2019) Google dataset search: Building a search engine for datasets in an open web ecosystem. The World Wide Web Conference, 2019, 1365–1375.

Brodaric, B., and Richard, S.M. (2020) The GeoScience Ontology. American Geophysical Union Fall Meeting, 2020, Abstract IN030-07.

Bullard, T., Freudenthal, J., Avagyan, S., and Kahr, B. (2007) Test of Cairns-Smith's 'crystals-as-genes' hypothesis. Faraday Discussions, 136, 231–245.

Bullock, M.A., and Grinspoon, D.H. (1996) The stability of climate on Venus. Journal of

Geophysical Research: Planets, 101, 7521–7529.

Burke, R., Felfernig, A., and Göker, M.H. (2011) Recommender systems: An overview. AI Magazine, 32, 13–18.

Cable, M.L., Runčevski, T., Maynard-Casely, H.E., Vu, T.H., and Hodyss, R. (2021) Titan in a test tube: Organic co-crystals and implications for Titan mineralogy. Accounts of Chemical Research, 54, 3050–3059.

Cairns-Smith, A., and Hartman, H. (1986) Clay Minerals and the Origin of Life. Cambridge University Press.

Cairns-Smith, A.G. (1990) Seven clues to the origin of life: a scientific detective story. Cambridge University Press.

Chamberlain, K.J., Lehnert, K.A., McIntosh, I.M., Morgan, D.J., and Wörner, G. (2021) Time to change the data culture in geochemistry. Nature Reviews Earth & Environment, 2, 737–739.

Chiama, K., Rutledge, R., Gabor, M., Lupini, I., Hazen, R.M., Zhang, S., and Boujibar, A. (2020) Garnet: A comprehensive, standardized, geochemical database incorporating locations and paragenesis. Geological Society of America Southeastern Section Annual Meeting, 69, 344505.

Chiama, K., Gabor, M., Lupini, I., Rutledge, R., Nord, J., Zhang, S., Boujibar, A., Bullock, E. S., Walter, M., Lehnert, K.A., and others (2022a) Garnet Dataset (ver. 1.0) (Online). Available: https://doi.org/10.48484/camh-xy98 (accessed January 21, 2023) Open Data Repository, Gray, ME.

——— (2022b) The secret life of garnets: A comprehensive, standardized dataset of garnet geochemical analyses integrating localities and petrogenesis. Earth System Science Data,

in prep.

Childers, S.E., Ciufo, S., and Lovley, D.R. (2002) *Geobacter metallireducens* accesses insoluble Fe(III) oxide by chemotaxis. Nature, 416, 767–769.

Cleland, C.E., Hazen, R.M., and Morrison, S.M. (2021) Historical natural kinds and mineralogy: Systematizing contingency in the context of necessity. Proceedings of the National Academy of Sciences, 118, e2015370118.

Coates, D.R. (1985) Mineral resources. In Geology and Society. Environmental Resource Management Series. p. 19–46. Springer, Boston, MA.

Collen, M.F. (1986) Origins of medical informatics. The Western Journal of Medicine, 145, 778–785.

Collerson, K.D., Williams, Q., Kamber, B.S., Omori, S., Arai, H., and Ohtani, E. (2010) Majoritic garnet: A new approach to pressure estimation of shock events in meteorites and the encapsulation of sub-lithospheric inclusions in diamond. Geochimica et Cosmochimica Acta, 74, 5939–5957.

Dana, J.D. (1895) The Geological Story Briefly Told. American Book Company.

De Sanctis, M.C., Ammannito, E., Capria, M.T., Tosi, F., Capaccioni, F., Zambon, F., Carraro, F., Fonte, S., Frigeri, A., Jaumann, R., and others (2012) Spectroscopic characterization of mineralogy and its diversity across Vesta. Science, 336, 697–700.

Derczynski, L., Ritter, A., Clark, S., and Bontcheva, K. (2013) Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. Proceedings of the International Conference on Recent Advances in Natural Language Processing, 2013, 198–206.

Ehlmann, B.L., and Edwards, C.S. (2014) Mineralogy of the martian surface. Annual Review of Earth and Planetary Sciences, 42, 291–315.

Fegley, B., Treiman, A.H., and Sharpton, V.L. (1992) Venus surface mineralogy - Observational and theoretical constraints. Proceedings of the Lunar and Planetary Science Conference, 22, p. 3-19, Lunar and Planetary Institute, Houston, TX.

Fischer, G., and Herrmann, T. (2011) Socio-technical systems: A meta-design perspective. International Journal of Sociotechnology and Knowledge Development, 3, 1–33.

Fox, P. (2011) The rise of informatics as a research domain WIRADA Science Symposium, 15, 125–131.

——— (2020) What is neo-informatics? American Geophysical Union, Fall Meeting, 2020, Abstract IN025-02.

Fox, P., and Hendler, J. (2014) The science of data science. Big Data, 2, 68–70.

Fox, P., Gundersen, L., Lehnert, K., McGuinness, D., Sinha, K., and Snyder, W. (2006) Toward broad community collaboration in geoinformatics. Eos, Transactions of the American Geophysical Union, 87, 513.

Frazier, A.W., Lehr, J.R., and Smith, J.P. (1963) The magnesium phosphates hannayite, schertelite and bobierrite. American Mineralogist, 48, 635–641.

Fritz, S., Milligan, I., Ruest, N., and Lin, J. (2020) Building community at distance: a datathon during COVID-19. Digital Library Perspectives, 36, 415–428.

Fyfe, A., McDougall-Waters, J., and Moxham, N. (2015) 350 years of scientific periodicals. Notes and Records: The Royal Society Journal of the History of Science, 69, 227–239.

Gauthier, J., Vincent, A.T., Charette, S.J., and Derome, N. (2019) A brief history of bioinformatics. Briefings in Bioinformatics, 20, 1981–1996.

Geng, X. (2016) Label distribution learning. IEEE Transactions on Knowledge and Data Engineering, 28, 1734–1748.

Geng, X., Yin, C., and Zhou, Z.-H. (2013) Facial age estimation by learning from label distributions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35, 2401–2412.

Geng, X., Wang, Q., and Xia, Y. (2014) Facial age estimation by adaptive label distribution learning. International Conference on Pattern Recognition, 22, 4465–4470.

Gilmore, M., Treiman, A., Helbert, J., and Smrekar, S. (2017) Venus surface composition constrained by observation and experiment. Space Science Reviews, 212, 1511–1540.

Glein, C.R., and Waite, J.H. (2020) The carbonate geochemistry of Enceladus' ocean. Geophysical Research Letters, 47, e2019GL085885.

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., and Fdez-Riverola, F. (2014) Web scraping technologies in an API world. Briefings in Bioinformatics, 15, 788–797.

Goble, C., and Stevens, R. (2008) State of the nation in data integration for bioinformatics. Journal of Biomedical Informatics, 41, 687–693.

Golugula, A., Lee, G., and Madabhushi, A. (2011) Evaluating feature selection strategies for high dimensional, small sample size datasets. Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 33, 949–952.

Gordon, V.S., and Bieman, J.M. (1995) Rapid prototyping: lessons learned. IEEE Software, 12, 85–95.

Gorlas, A., Jacquemot, P., Guigner, J.-M., Gill, S., Forterre, P., and Guyot, F. (2018) Greigite nanocrystals produced by hyperthermophilic archaea of Thermococcales order. PLOS ONE, 13, e0201549.

Gray, J., and Szalay, A. (2007) eScience - A transformed scientific method. In National Reseaech

Council-Computer Science and Telecommunication Board meeting, p. 146-156. National Research Council-CTSB Meeting, Mountain View, CA.

Greenland, S., Schwartzbaum, J.A., and Finkle, W.D. (2000) Problems due to small samples and sparse data in conditional logistic regression analysis. American Journal of Epidemiology, 151, 531–539.

Greenland, S., Mansournia, M.A., and Altman, D.G. (2016) Sparse data bias: a problem hiding in plain sight. BMJ, 352, i1981.

Greenwell, H.C., and Coveney, P.V. (2006) Layered double hydroxide minerals as possible prebiotic information storage and transfer compounds. Origins of Life and Evolution of Biospheres, 36, 13–37.

Gregory, D.D., Cracknell, M.J., Large, R.R., McGoldrick, P., Kuhn, S., Maslennikov, V.V., Baker, M.J., Fox, N., Belousov, I., Figueroa, M.C., and others (2019) Distinguishing ore deposit type and barren sedimentary pyrite using laser ablation-inductively coupled plasma-mass spectrometry trace element data and statistical analysis of large data sets. Economic Geology, 114, 771–786.

Grew, E.S., Hystad, G., Hazen, R.M., Krivovichev, S.V., and Gorelova, L.A. (2017) How many boron minerals occur in Earth's upper crust? American Mineralogist, 102, 1573–1587.

Grove, T.L., and Hazen, R.M. (1974) Alkali feldspar unit-cell parameters at liquid nitrogen temperature: Low temperature limits of the displacive transformation. American Mineralogist, 59, 1327–1329.

Gunawan, R., Rahmatulloh, A., Darmawan, I., and Firdaus, F. (2019) Comparison of web scraping techniques: Regular expression, HTML DOM and Xpath. Proceedings of the International Conference on Industrial Enterprise and System Engineering, 2018, 2, p. 283-287.

Yogyakarta, Central Java, Indonesia.

Gunter, M.E., and Ribbe, P.H. (1993) Natrolite group zeolites: Correlations of optical properties and crystal chemistry. Zeolites, 13, 435–440.

Hall, P., Marron, J.S., and Neeman, A. (2005) Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society, B67, 427–444.

Hashimoto, G.L., and Abe, Y. (2005) Climate control on Venus: Comparison of the carbonate and pyrite models. Planetary and Space Science, 53, 839–848.

Hazen, R.M. (2005) Genesis: Rocks, minerals, and the geochemical origin of life. Elements, 1, 135–137.

Hazen, R.M. (2018) Titan mineralogy: A window on organic mineral evolution. American Mineralogist, 103, 341–342.

——— (2019) An evolutionary system of mineralogy: Proposal for a classification of planetary materials based on natural kind clustering. American Mineralogist, 104, 810–816.

Hazen, R.M., and Morrison, S.M. (2020) An evolutionary system of mineralogy. Part I: Stellar mineralogy (>13 to 4.6 Ga). American Mineralogist, 105, 627–651.

——— (2022) On the paragenetic modes of minerals: A mineral evolution perspective. American Mineralogist, 107, 1262–1287.

Hazen, R.M., and Sholl, D.S. (2003) Chiral selection on inorganic crystalline surfaces. Nature Materials, 2, 367–374.

Hazen, R.M., and Sverjensky, D.A. (2010) Mineral surfaces, geochemical complexities, and the origins of life. Cold Spring Harbor Perspectives in Biology, 2, a002162–a002162.

Hazen, R.M., Griffin, P.L., Carothers, J.M., and Szostak, J.W. (2007) Functional information and the emergence of biocomplexity. Proceedings of the National Academy of Sciences, 104,

8574–8581.

Hazen, R.M., Papineau, D., Bleeker, W., Downs, R.T., Ferry, J.M., McCoy, T.J., Sverjensky, D.A., and Yang, H. (2008) Mineral evolution. American Mineralogist, 93, 1693–1720.

Hazen, R.M., Liu, X.-M., Downs, R.T., Golden, J., Pires, A.J., Grew, E.S., Hystad, G., Estrada, C., and Sverjensky, D.A. (2014) Mineral evolution: Episodic metallogenesis, the supercontinent cycle, and the coevolving geosphere and biosphere. In Building Exploration Capability for the 21st Century. Society of Economic Geologists Special Publication, 18, 1-15.

Hazen, R.M., Hystad, G., Downs, R.T., Golden, J.J., Pires, A.J., and Grew, E.S. (2015) Earth's "missing" minerals. American Mineralogist, 100, 2344–2347.

Hazen, R.M., Downs, R.T., Eleish, A., Fox, P., Gagné, O.C., Golden, J.J., Grew, E.S., Hummer, D.R., Hystad, G., Krivovichev, S.V., and others (2019) Data-driven discovery in mineralogy: Recent advances in data resources, analysis, and visualization. Engineering, 5, 397–405.

Hazen, R.M., Morrison, S.M., and Prabhu, A. (2021) An evolutionary system of mineralogy. Part III: Primary chondrule mineralogy (4566 to 4561 Ma). American Mineralogist, 106, 325–350.

Hazen, R.M., Morrison, S.M., Krivovichev, S.V., and Downs, R.T. (2022) Lumping and splitting: Toward a classification of mineral natural kinds. American Mineralogist, 107, 1288–1301.

Heberling, J.M., Miller, J.T., Noesgaard, D., Weingart, S.B., and Schigel, D. (2021) Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences, 118, e2018093118.

Hey, A.J.G., Ed. (2009) The fourth paradigm: data-intensive scientific discovery, 251 p. Microsoft

Research.

Hindrichs, A.S., Eleazer, K., Lui, T., Williams, J., Nord, J., Gregory, D., Morrison, S., Hazen, R.M., and Ostroverkhova, A. (2022) Oxide spinel and data-driven discovery: A comprehensive mineralogical and geochemical data resource, incorporating composition, location, and paragenesis. Geological Society of America Southeastern Section Regional Meeting, 2022, Abstract 375662.

Hinkel, N.R., and Unterborn, C.T. (2018) The star–planet connection. I. Using stellar composition to observationally constrain planetary mineralogy for the 10 closest stars. The Astrophysical Journal, 853, 83.

Hossin, M., and Sulaiman, M.N. (2015) A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5, 1–11.

Hummer, D.R., Golden, J.J., Hystad, G., Downs, R.T., Eleish, A., Liu, C., Ralph, J., Morrison, S.M., Meyer, M.B., and Hazen, R.M. (2022) Evidence for the oxidation of Earth's crust from the evolution of manganese minerals. Nature Communications, 13, 960.

Hystad, G., Downs, R.T., and Hazen, R.M. (2015) Mineral species frequency distribution conforms to a large number of rare events model: Prediction of Earth's missing minerals. Mathematical Geosciences, 47, 647–661.

Hystad, G., Morrison, S.M., and Hazen, R.M. (2019) Statistical analysis of mineral evolution and mineral ecology: The current state and a vision for the future. Applied Computing and Geosciences, 1, 100005.

Hystad, G., Boujibar, A., Liu, N., Nittler, L.R., and Hazen, R.M. (2021) Evaluation of the classification of pre-solar silicon carbide grains using consensus clustering with resampling

methods: An assessment of the confidence of grain assignments. Monthly Notices of the Royal Astronomical Society, 510, 334–350.

Irifune, T. (1987) An experimental investigation of the pyroxene-garnet transformation in a pyrolite composition and its bearing on the constitution of the mantle. Physics of the Earth and Planetary Interiors, 45, 324–336.

Jackson, I. (2010) OneGeology: improving access to geoscience globally. Earthwise, 26, 14–15.

Katz, S. (1987) Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Transactions on Acoustics, Speech, and Signal Processing, 35, 400–401.

Kläs, M. (2018, November 28) Towards identifying and managing sources of uncertainty in AI and machine learning models - An overview. arXiv.

Kluyver, T., Ragan-Kelley, B., Perez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., and others (2016) Jupyter Notebooks – a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas, pp. 87–90.

Krivovichev, S.V. (2013) Structural complexity of minerals: information storage and processing in the mineral world. Mineralogical Magazine, 77, 275–326.

Krivovichev, S.V. (2015) Structural complexity of minerals and mineral parageneses: Information and its evolution in the mineral world. In T. Armbruster and R.M. Danisi, Eds., Highlights in Mineralogical Crystallography, pp. 31–74. De Gruyter.

——— (2016) Structural complexity and configurational entropy of crystals. Acta Crystallographica, B72, 274–276.

Krivovichev, S.V., Yakovenchuk, V.N., and Zhitova, E.S. (2011) Natural double layered

hydroxides: Structure, chemistry, and information storage capacity. In S.V. Krivovichev, Ed., Minerals as Advanced Materials, 2, 87–102. Springer.

Krivovichev, S.V., Krivovichev, V.G., and Hazen, R.M. (2018) Structural and chemical complexity of minerals: correlations and time evolution. European Journal of Mineralogy, 30, 231–236.

Lafuente, B., Downs, R.T., Yang, H., and Stone, N. (2015) 1. The power of databases: The RRUFF project. In T. Armbruster and R.M. Danisi, Eds., Highlights in Mineralogical Crystallography, pp. 1–30. De Gruyter.

Large, R.R., Hazen, R.M., Morrison, S.M., Gregory, D.D., Steadman, J.A., and Mukherjee, I. (2022) Evidence that the GOE was a prolonged event with a peak around 1900 Ma. Geosystems and Geoenvironment, 1, 100036.

Lehnert, K., Su, Y., Langmuir, C.H., Sarbas, B., and Nohl, U. (2000) A global geochemical database structure for rocks: Geochemical database structure. Geochemistry, Geophysics, Geosystems, 1, 1012.

Liu, B., Wei, Y., Zhang, Y., and Yang, Q. (2017) Deep neural networks for high dimension, low sample size data. Proceedings of the International Joint Conference on Artificial Intelligence, 26, 2287–2293.

Liu, C., Runyon, S.E., Knoll, A.H., and Hazen, R.M. (2019) The same and not the same: Ore geology, mineralogy and geochemistry of Rodinia assembly versus other supercontinents. Earth-Science Reviews, 196, 102860.

Liu, X.-M., Kah, L.C., Knoll, A.H., Cui, H., Wang, C., Bekker, A., and Hazen, R.M. (2021) A persistently low level of atmospheric oxygen in Earth's middle age. Nature Communications, 12, 351.

Lohr, S. (2012, February 11) The Age of Big Data. The New York Times.

Lord, P., Bechhofer, S., Wilkinson, M.D., Schiltz, G., Gessler, D., Hull, D., Goble, C., and Stein, L. (2004) Applying semantic web services to bioinformatics: Experiences gained, lessons learnt. In S.A. McIlraith, D. Plexousakis, and F. van Harmelen, Eds., The Semantic Web – ISWC 2004, 3298, 350–364. Springer.

Ma, X., Hummer, D., Golden, J., Fox, P., Hazen, R., Morrison, S., Downs, R., Madhikarmi, B., Wang, C., and Meyer, M. (2017) Using visual exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research. ISPRS International Journal of Geo-Information, 6, 368.

Maynard-Casely, H.E., Cable, M.L., Malaska, M.J., Vu, T.H., Choukroun, M., and Hodyss, R. (2018) Prospects for mineralogy on Titan. American Mineralogist, 103, 343–349.

McGuinness, K.N., Klau, G.W., Morrison, S.M., Moore, E.K., Seipp, J., Falkowski, P.G., and Nanda, V. (2022) Evaluating mineral lattices as evolutionary proxies for metalloprotein evolution. Origins of Life and Evolution of Biospheres, 52, 263–275.

Meadows, V.S., Reinhard, C.T., Arney, G.N., Parenteau, M.N., Schwieterman, E.W., Domagal-Goldman, S.D., Lincowski, A.P., Stapelfeldt, K.R., Rauer, H., DasSarma, S., and others (2018) Exoplanet biosignatures: Understanding oxygen as a biosignature in the context of its environment. Astrobiology, 18, 630–662.

Moore, E.K., Jelen, B.I., Giovannelli, D., Raanan, H., and Falkowski, P.G. (2017) Metal availability and the expanding network of microbial metabolisms in the Archaean eon. Nature Geoscience, 10, 629–636.

Morrison, S., Hazen, R.M., Prabhu, A., Williams, J., Eleish, A., and Fox, P. (2021) Mineral network analysis: Exploring geological, geochemical, and biological patterns in

mineralization via multidimensional analysis. Geological Society of America Annual Meeting, 2021, Abstract 370437.

Morrison, S.M., Liu, C., Eleish, A., Prabhu, A., Li, C., Ralph, J., Downs, R.T., Golden, J.J., Fox, P., Hummer, D.R., and others (2017) Network analysis of mineralogical systems. American Mineralogist, 102, 1588–1596.

Morrison, S.M., Downs, R.T., Blake, D.F., Vaniman, D.T., Ming, D.W., Hazen, R.M., Treiman, A.H., Achilles, C.N., Yen, A.S., Morris, R.V., and others (2018a) Crystal chemistry of martian minerals from Bradbury Landing through Naukluft Plateau, Gale crater, Mars. American Mineralogist, 103, 857–871.

Morrison, S.M., Pan, F., Gagné, O.C., Prabhu, A., Eleish, A., Fox, P.A., Downs, R.T., Bristow, T., Rampe, E.B., Blake, D.F., and others (2018b) Predicting multi-component mineral compositions in Gale Crater, Mars with label distribution learning. American Geophysical Union Fall Meeting, 2018, Abstract P21I-3438.

Morrison, S.M., Downs, R.T., Blake, D.F., Prabhu, A., Eleish, A., Vaniman, D.T., Ming, D.W., Rampe, E.B., Hazen, R.M., Achilles, C.N., and others (2018c) Relationships between unit-cell parameters and composition for rock-forming minerals on Earth, Mars, and other extraterrestrial bodies. American Mineralogist, 103, 848–856.

Morrison, S.M., Buongiorno, J., Downs, R.T., Eleish, A., Fox, P., Giovannelli, D., Golden, J.J., Hummer, D.R., Hystad, G., Kellogg, L.H., and others (2020) Exploring carbon mineral systems: Recent advances in C mineral evolution, mineral ecology, and network analysis. Frontiers in Earth Science, 8, 208.

Morrison, S.M., Prabhu, A., Eleish, A., Fox, P., Golden, J.J., Downs, R.T., Perry, S.N., Burns, P.C., Ralph, J., and Hazen, R.M. (2023) Machine learning approaches for predictive

mineralogy in Earth and planetary science: A study in mineral association analysis (in review). PNAS Nexus.

Murchie, S.L., Mustard, J.F., Ehlmann, B.L., Milliken, R.E., Bishop, J.L., McKeown, N.K., Noe Dobrea, E.Z., Seelos, F.P., Buczkowski, D.L., Wiseman, S.M., and others (2009) A synthesis of Martian aqueous mineralogy after 1 Mars year of observations from the Mars Reconnaissance Orbiter. Journal of Geophysical Research, 114, E00D06.

Murray, H.H. (1995) Industrial minerals—key to economic development. In R.L. Miller, G. Escalante, J.A. Reinemund, and M.J. Bergin, Eds., Energy and Mineral Potential of the Central American-Caribbean Region, 16, 335–337. Springer.

Namur, O., and Charlier, B. (2017) Silicate mineralogy at the surface of Mercury. Nature Geoscience, 10, 9–13.

Nance, R.D., Murphy, J.B., and Santosh, M. (2014) The supercontinent cycle: A retrospective essay. Gondwana Research, 25, 4–29.

National Research Council (2015) Enhancing the Effectiveness of Team Science. The National Academies Press.

Nazábal, A., Olmos, P.M., Ghahramani, Z., and Valera, I. (2020) Handling incomplete heterogeneous data using VAEs. Pattern Recognition, 107, 107501.

Needham, J., and Wang, L. (1995) Science and civilisation in China. Cambridge University Press.

Nesse, W.D. (2013) Introduction to optical mineralogy, 4th ed., 361 p. Oxford University Press.

Nitschke, W., McGlynn, S.E., Milner-White, E.J., and Russell, M.J. (2013) On the antiquity of metalloenzymes and their substrates in bioenergetics. Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1827, 871–881.

Novikov, Y., and Copley, S.D. (2013) Reactivity landscape of pyruvate under simulated

hydrothermal vent conditions. Proceedings of the National Academy of Sciences, 110, 13283–13288.

Postberg, F., Kempf, S., Hillier, J.K., Srama, R., Green, S.F., McBride, N., and Grün, E. (2008) The E-ring in the vicinity of Enceladus. Icarus, 193, 438–454.

Prabhu, A., and Fox, P. (2021) Reproducible workflow. In B.S. Daya Sagar, Q. Cheng, J. McKinley, and F. Agterberg, Eds., Encyclopedia of Mathematical Geosciences, pp. 1–5. Springer International Publishing.

Prabhu, A., Morrison, S.M., Eleish, A., Narkar, S., Fox, P.A., Golden, J.J., Downs, R.T., Perry, S., Burns, P.C., Ralph, J., and others (2019) Predicting unknown mineral localities based on mineral associations. American Geophysical Union Fall Meeting, 2019, Abstract EP23D-2286.

Prabhu, A., Morrison, S., and Giovannelli, D. (2021a) A new way to evaluate association rule mining methods and its applicability to mineral association analysis. American Geophysical Union, 2021, Abstract IN45-08.

Prabhu, A., Morrison, S.M., Eleish, A., Zhong, H., Huang, F., Golden, J.J., Perry, S.N., Hummer, D.R., Ralph, J., Runyon, S.E., and others (2021b) Global earth mineral inventory: A data legacy. Geoscience Data Journal, 8, 74–89.

Prettyman, T.H., Yamashita, N., Ammannito, E., Ehlmann, B.L., McSween, H.Y., Mittlefehldt, D.W., Marchi, S., Schörghofer, N., Toplis, M.J., Li, J.-Y., and others (2019) Elemental composition and mineralogy of Vesta and Ceres: Distribution and origins of hydrogen-bearing species. Icarus, 318, 42–55.

Putirka, K.D., Dorn, C., Hinkel, N.R., and Unterborn, C.T. (2021) Compositional diversity of rocky exoplanets. Elements, 17, 235–240.

Ramachandran, R., Bugbee, K., and Murphy, K. (2021) From open data to open science. Earth and Space Science, 8.

Rampe, E.B., Lapotre, M.G.A., Bristow, T.F., Arvidson, R.E., Morris, R.V., Achilles, C.N., Weitz, C., Blake, D.F., Ming, D.W., Morrison, S.M., and others (2018) Sand mineralogy within the Bagnold Dunes, Gale Crater, as observed in situ and from orbit. Geophysical Research Letters, 45, 9488–9497.

Reichman, O.J., Jones, M.B., and Schildhauer, M.P. (2011) Challenges and opportunities of open data in ecology. Science, 331, 703–705.

Rogers, K., Thomson, B., Colwell, F., Eleish, A., Fontaine, K., Fox, P., Gaidos, E., Hoarfrost, A., Huang, F., Ladau, J., Magnabosco, C., Parsons, M.A., Prabhu, A., Ruff, E., and Twing, K.I. (2018) The Census of Deep Life: Metadata then and now. American Geophyscial Union, Fall Meeting, 2018, Abstract IN53C-0629.

Russell, M. (2018) Green rust: The simple organizing 'seed' of all life? Life, 8, 35.

Russell, M.J., and Hall, A.J. (1997) The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. Journal of the Geological Society, 154, 377–402.

Sandve, G.K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013) Ten simple rules for reproducible computational research. PLoS Computational Biology, 9, e1003285.

Shah, K., Salunke, A., Dongare, S., and Antala, K. (2017) Recommender systems: An overview of different approaches to recommendations. International Conference on Innovations in Information, Embedded and Communication Systems, 4, 1-4.

Shen, D., Shen, H., Zhu, H., and Marron, J.S. (2017) The statistics and mathematics of high dimension low sample size asymptotics. Statistica Sinica, 26, 1747-1770.

Shepperd, M., and Cartwright, M. (2001) Predicting with sparse data. IEEE Transactions on Software Engineering, 27, 987–998.

Shi, L., Dong, H., Reguera, G., Beyenal, H., Lu, A., Liu, J., Yu, H.-Q., and Fredrickson, J.K. (2016) Extracellular electron transfer mechanisms between microorganisms and minerals. Nature Reviews Microbiology, 14, 651–662.

Sinha, A.K., Malik, Z., Rezgui, A., Barnes, C.G., Lin, K., Heiken, G., Thomas, W.A., Gundersen, L.C., Raskin, R., Jackson, I., and others (2010) Geoinformatics: Transforming data to knowledge for geosciences. GSA Today, 20, 4–10.

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., and Wyborn, L. (2019) Make scientific data FAIR. Nature, 570, 27–29.

Statnikov, A., Wang, L., and Aliferis, C.F. (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics, 9, 319.

Strunz, H., and Tennyson, C. (1941) Mineralogische tabellen. Akademische Verlagsgesellschaft Becker & Erler.

Sverjensky, D.A., and Lee, N. (2010) The Great Oxidation Event and mineral diversification. Elements, 6, 31–36.

Sweeting, M.J., Sutton, A.J., and Lambert, P.C. (2004) What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Statistics in Medicine, 23, 1351–1375.

Thomson, A.R., Kohn, S.C., Prabhu, A., and Walter, M.J. (2021) Evaluating the formation pressure of diamond-hosted majoritic garnets: A machine learning majorite barometer. Journal of Geophysical Research: Solid Earth, 126, e2020JB02604.

Tomašev, N., and Radovanović, M. (2016) Clustering evaluation in high-dimensional data. In M.E. Celebi and K. Aydin, Eds., Unsupervised Learning Algorithms, pp. 71–107. Springer International Publishing.

Treiman, A.H., and Bullock, M.A. (2012) Mineral reaction buffering of Venus' atmosphere: A thermochemical constraint and implications for Venus-like planets. Icarus, 217, 534–541.

Unterborn, C.T., and Panero, W.R. (2019) The pressure and temperature limits of likely rocky exoplanets. Journal of Geophysical Research: Planets, 124, 1704–1716.

Unterborn, C.T., Dismukes, E.E., and Panero, W.R. (2016) Scaling the Earth: A sensitivity analysis of terrestrial exoplanetary interior models. The Astrophysical Journal, 819, 32.

Uzuner, O. (2009) Recognizing obesity and comorbidities in sparse data. Journal of the American Medical Informatics Association, 16, 561–570.

Voice, P.J., Kowalewski, M., and Eriksson, K.A. (2011) Quantifying the timing and rate of crustal evolution: Global compilation of radiometrically dated detrital zircon grains. The Journal of Geology, 119, 109–126.

Wachter, S. (2019) Data protection in the age of big data. Nature Electronics, 2, 6–7.

Wächtershäuser, G. (1988) Before enzymes and templates: theory of surface metabolism. Microbiological Reviews, 52, 452–484.

Waite, J.H., Glein, C.R., Perryman, R.S., Teolis, B.D., Magee, B.A., Miller, G., Grimes, J., Perry, M.E., Miller, K.E., Bouquet, A., and others (2017) Cassini finds molecular hydrogen in the Enceladus plume: Evidence for hydrothermal processes. Science, 356, 155–159.

Wang, C., Hazen, R.M., Cheng, Q., Stephenson, M.H., Zhou, C., Fox, P., Shen, S., Oberhänsli, R., Hou, Z., Ma, X., and others (2021) The deep-time digital Earth program: data-driven discovery in geosciences. National Science Review, 8, nwab027.

Wang, L. (2017) Heterogeneous data and big data analytics. Automatic Control and Information Sciences, 3, 8–15.

Wiederhold, G. (1999) Mediation to deal with heterogeneous data sources. In A. Včkovski, K.E. Brassel, and H.-J. Schek, Eds., Interoperating Geographic Information Systems, 1580, 1–16. Springer.

Wijbrans, C.H., Rohrbach, A., and Klemme, S. (2016) An experimental investigation of the stability of majoritic garnet in the Earth's mantle and an improved majorite geobarometer. Contributions to Mineralogy and Petrology, 171, 50.

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., and others (2016) The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018.

Wise, A.F., and Shaffer, D.W. (2015) Why theory matters more than ever in the age of big data. Journal of Learning Analytics, 2, 5–13.

Wyborn, L.A., Lehnert, K., and Klump, J.F. (2021) The future of X-informatics lies in collaborative convergence: An exemplar from the global OneGeochemistry initiative. American Geophysical Union Fall Meeting, 2021, Abstract IN13A-02.

Yang, H., Sun, H.J., and Downs, R.T. (2011) Hazenite, $KNaMg_2(PO_4)_2{\cdot}14H_2O$, a new biologically related phosphate mineral, from Mono Lake, California, U.S.A. American Mineralogist, 96, 675–681.

Yata, K., and Aoshima, M. (2012) Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. Journal of Multivariate Analysis, 105, 193–215.

Yu, S. (2016) Big privacy: Challenges and opportunities of privacy study in the age of big data.

IEEE Access, 4, 2751–2763.

Zhang, L., Xie, Y., Xidao, L., and Zhang, X. (2018) Multi-source heterogeneous data fusion. International Conference on Artificial Intelligence and Big Data, 2018, 47–51.

Zhang, S., Morrison, S.M., Prabhu, A., Ma, C., Huang, F., Gregory, D., Large, R.R., and Hazen, R. (2019) Natural clustering of pyrite with implications for its formational environment. American Geophysical Union Fall Meeting, 2019, Abstract EP23D-2284.

Zhao, B. (2017) Web scraping. In L.A. Schintler and C.L. McNeely, Eds., Encyclopedia of Big Data pp. 1–3. Springer International Publishing.

Zhao, D., Bartlett, S., and Yung, Y.L. (2020) Quantifying mineral-ligand structural similarities: Bridging the geological world of minerals with the biological world of enzymes. Life, 10, 338.

Zhou, J., Gandomi, A.H., Chen, F., and Holzinger, A. (2021) Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics, 10, 593.

Zolotov, M.Yu. (2018) Gas–solid interactions on Venus and other solar system bodies. Reviews in Mineralogy and Geochemistry, 84, 351–392.

This is the peer-reviewed, final accepted version for American Mineralogist, published by the Mineralogical Society of America.
The published version is subject to change. Cite as Authors (Year) Title. American Mineralogist, in press.
DOI: https://doi.org/10.2138/am-2022-8613. http://www.minsocam.org/

Name	URL
IMA list of approved minerals	https://rruff.info/ima/
RRUFF Project	https://rruff.info/
Mineral Evolution Database	https://rruff.info/Evolution/
American Mineralogist Crystal Structure Database	http://rruff.geo.arizona.edu/AMS/amcsd.php
Handbook of Mineralogy	
Mindat.org	https://www.mindat.org/
Mineral Properties Database	https://odr.io/MPD
Evolutionary System of Mineralogy Database	https://odr.io/esmd
CheMin Database	https://odr.io/chemin
Astromaterials Data System	https://www.astromat.org/
EarthChem	https://earthchem.org/
GEOROC	http://georoc.mpch-mainz.gwdg.de/georoc/

MetPetDB	https://tw.rpi.edu/project/MetPetDB
Planetary Data System	https://pds.nasa.gov/
Mineral RI	https://odr.io/mineralRI

Description

A searchable database of mineral species information, including chemical formula, unit-cell parameters, paragenetic modes, and links to other important mineralogical data resources

A mineral library and database of chemical, spectral, and diffraction data for mineral species (Lafuente et al. 2015).

A database of mineral locality and age information, with ~200,000 species/locality/age records extracted primarily from scientific literature and mindat.org. (Golden et al. 2016; Golden 2019).

A crystal structure database that includes every structure published in the *American Mineralogist* , *The Canadian Mineralogist* , *European Journal of Mineralogy* , and *Physics and Chemistry of Minerals* , as well as selected datasets from other journals.

A five volume set with each of the 4988 pages dedicated to a mineral species description, with information such as crystallographic and physical attributes, microprobe chemical analyses, paragenetic mode and locality information, and select references.

The world's largest open database of minerals, rocks, meteorites and the localities from which they were found.
A database of various mineral attributes including age, color, redox state, structural complexity, and method of discovery.
A database containing measured geochemical and physical characteristics of mineral samples, including major, minor, trace elements as well as isotopic ratios. (Chiama et al. 2022a)
A database containing the X-ray diffraction data from martian rock and soil samples analyzed by the CheMin instrument onboard the NASA Mars Science Laboratory.

A data infrastructure that stores, curates, and provides access to laboratory data acquired on samples curated in the NASA Johnson Space Center Astromaterials Collection, including the Apollo lunar samples and the Antarctic meteorite collection (Lehnert et al. 2019).

A data system providing open data services to the geochemical, petrological, mineralogical, and related communities, including data preservation, discovery, access, and visualization.

a global geochemical database containing published chemical and isotopic data as well as extensive metadata for rocks, minerals and melt/fluid inclusions.

A relational database and repository for global geochemical data on and images collected from metamorphic rocks from the earth's crust.

A long-term archive of digital data products returned from NASA's planetary missions, and from other kinds of flight and ground-based data acquisitions, including laboratory experiments.

A database containing the refractive indices minerals and synthetic compounds. (Shannon et al. 2017).