20

Deep Carbon through Deep Time

Data-Driven Insights

ROBERT M. HAZEN, YANA BROMBERG, ROBERT T. DOWNS,
AHMED ELEISH, PAUL G. FALKOWSKI, PETER FOX,
DONATO GIOVANNELLI, DANIEL R. HUMMER, GRETHE HYSTAD,
JOSHUA J. GOLDEN, ANDREW H. KNOLL, CONGRUI LI, CHAO LIU,
ELI K. MOORE, SHAUNNA M. MORRISON, A.D. MUSCENTE,
ANIRUDH PRABHU, JOLYON RALPH, MICHELLE Y. RUCKER,
SIMONE E. RUNYON, LISA A. WARDEN, AND HAO ZHONG

20.1 Introduction: Data and the Deep Carbon Observatory

For most of the history of science, data-driven discovery has been difficult and time-consuming: a lifetime of meticulous data collection and thoughtful synthesis was required to recognize previously hidden, higher-dimensional trends in multivariate data. Recognition of processes such as biological evolution by natural selection (1,2), continental evolution by plate tectonics (3,4), atmospheric and ocean oxygenation by photosynthesis (5,6), and climate change (7,8) required decades of integrated data synthesis preceding the discovery and acceptance of critical Earth phenomena. However, we stand at the precipice of a unique opportunity: to dramatically accelerate scientific discovery by coupling hard-won data resources with advanced analytical and visualization techniques (9,10). Today, Earth and life sciences are generating a multitude of data resources in numerous subdisciplines. Integration and synthesis of these diverse data resources will lead to an abductive, data-driven approach to investigating Earth's mineralogical and geochemical history, as well as the coevolution of the geosphere and biosphere (11–13).

In this chapter, we examine applications of data science in deep carbon research through three "use cases." The first example focuses on geochemical and mineralogical anomalies from a period in Earth history (~1.3 to 0.9 Ga) when the supercontinent Rodinia was being assembled from previously scattered continental blocks. The second case study examines the diversity and distribution of minerals, notably carbon-bearing minerals, through deep time from the contexts of mineral evolution, mineral ecology, and mineral network analysis. The third and most speculative use case considers ways to analyze and visualize data that relate microbial protein expression to growth environments – complex interconnections that may shed light on Earth's coevolving microbial ecosystems and near-surface geochemical environments. In each example, discoveries related to Earth's deep-time evolution have resulted from the analysis and visualization of large data resources fostered by the Deep Carbon Observatory (DCO).

20.2 Use Case #1: Global Signatures of Supercontinent Assembly

Large and growing geochemical and mineralogical data resources facilitate global surveys of trends in crustal evolution through deep time. Over the past 3 billion years, Earth has undergone five periods of supercontinent assembly, during which most continents converged and concentrated into one more or less contiguous landmass. Each of these assembly episodes was followed by intervals of supercontinent stability, rifting, and dispersal (14–16).

In spite of some shared geochemical, mineralogical, and tectonic characteristics, each of these five supercontinent episodes is distinct in detail. The Mesoproterozoic Rodinian supercontinent, in particular, displays several unique mineralogical and geochemical characteristics that point to a unique outcome of collisional events between ~1.3 and 0.9 billion years ago (16–22). Rodinia represents an important transitional period for Earth's carbon cycle in terms of both geochemical and biological evolution. In this section, we examine rapidly growing data resources in mineralogy and geochemistry that shed light on the unique character of this interval of Earth's history.

20.2.1 Mineralogical Evidence

Evidence for five cycles of supercontinent assembly, stability, and dispersal are strikingly preserved in the age distributions of high-temperature minerals (including many igneous, metamorphic, and hydrothermal species), which may be preferentially formed and/or preserved during continental suturing. The most notable mineralogical proxy is detrital zircon grains (23–29). As with other supercontinents (Figure 20.1a), the assembly of Rodinia saw a significant peak in the production and/or detrital preservation of zircon, with a global maximum at ~1.1 to 1.0 Ga (23,29).

Important mineralogical insights into supercontinent cycles are provided by minerals other than zircon (Figure 20.1) (16,30,31), and these mineral species can be explored through deep time thanks to the creation and rapid expansion of the Mineral Evolution Database (MED; rruff.info/evolution), an important contribution of DCO mineralogists. The MED incorporates more than 195,000 mineral/locality/age data, mostly for minerals from well-constrained magmatic, metamorphic, or hydrothermal events (data as of June 10, 2019). Liu et al. (16,22) employed the MED to explore and document age distributions of minerals and found that minerals containing niobium and yttrium (Figure 20.1b and c) exhibit similar trends to those of zircon; these minerals display maxima slightly later than zircon, at ~1.1 to 0.95 Ga. By contrast, minerals of most other elements, including Ni, Co, Au, S, Hg, Li, and C (Figure 20.1d to j), record significant pulses of mineralization during the assembly of Kenorland, Nuna, Gondwana (Pannotia), and Pangaea, but notably indicate decreased mineralization during Rodinian assembly (30–35). From these observations, we conclude that the currently expressed patterns of mineralization associated with the Rodinian assembly are unique relative to those of the other aforementioned supercontinents.
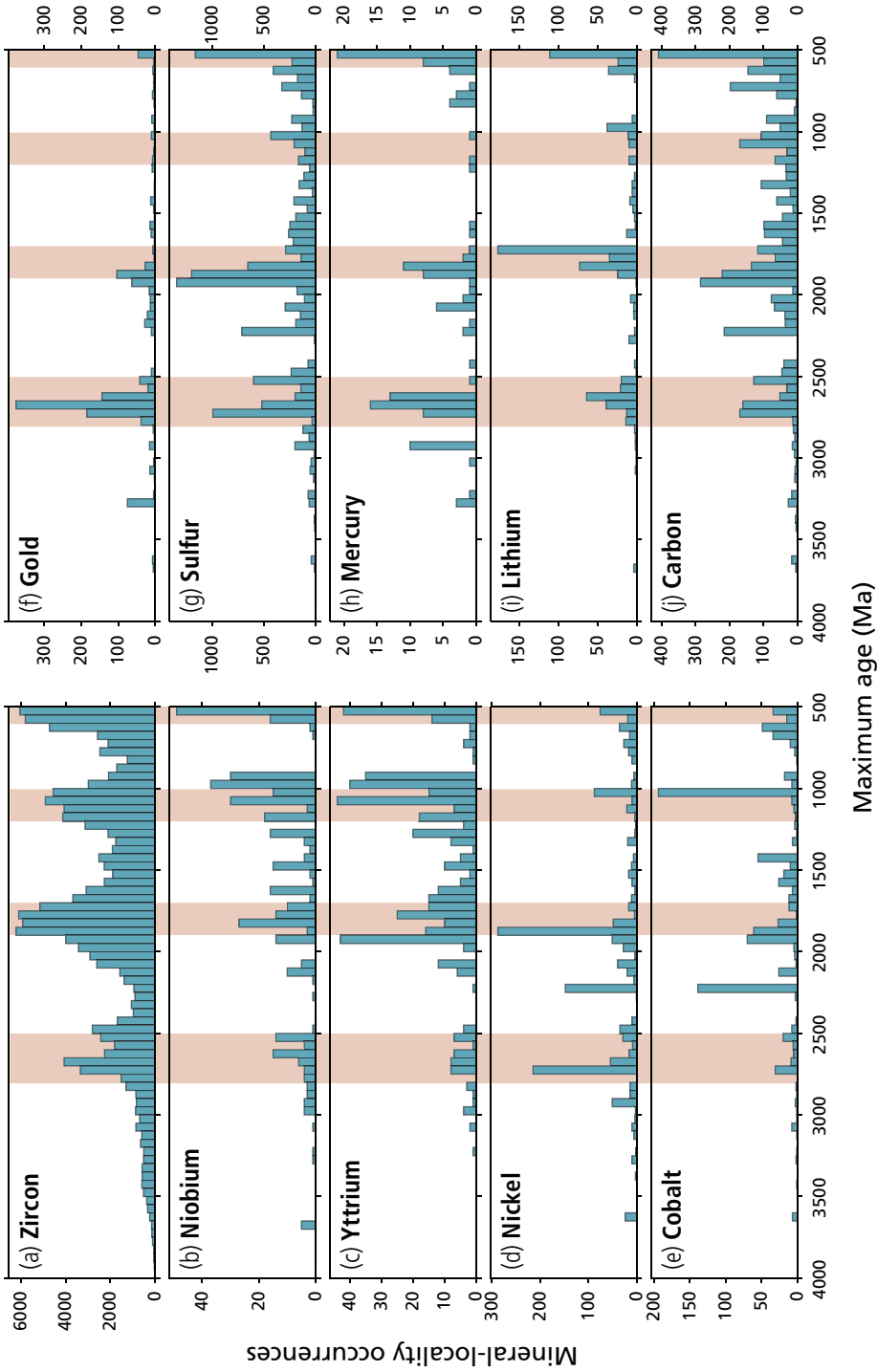
Figure 20.1 Mineral frequency of occurrence through deep time for: (a) detrital zircon (data from 29), (b) minerals with essential niobium, (c) yttrium, (d) nickel, (e) cobalt, (f) gold, (g) sulfur, (h) mercury, (i) lithium, and (j) carbon (data from earthchem.org). Bar graphs display 50-million-year bins. Vertical bars indicate periods of supercontinent assembly.

622

20.2.2 Trace Element Distributions

Temporal changes in the global averages of trace elements in igneous rocks complement and amplify mineral evolution data. Liu et al. (16,22) compiled trace element data for globally distributed igneous rocks from the EarthChem database (earthchem.org) and the United States Geological Survey (USGS) Mineral Resources Online Spatial database (mrdata.usgs.gov). They compiled age/concentration data for 129,161 samples with reported Zr analyses, 105,045 with Nb analyses, 121,373 with Y analyses, 77,835 with Co analyses, and 82,611 with Ni analyses from igneous rocks, all of which are associated with SiO_2 content (wt.%) and modern geographic coordinates (Figure 20.2).

The period of Rodinian assembly from 1.3 to 0.9 Ga saw significantly greater niobium, yttrium, and zirconium concentrations in igneous rocks than at any other time during the last 3 billion years (Figure 20.2). Furthermore, these trace element maxima apply to both mafic and felsic igneous rocks. By contrast, Liu et al.'s (16,22) survey found that average nickel and cobalt concentrations in igneous rocks display no significant enrichments or depletions during this interval (Figure 20.2).

20.2.3 Why Is Rodinian Assembly Unique?

Rodinia has long been recognized as distinct from other supercontinents. In addition to the mineralogical and geochemical anomalies noted above (Figure 20.1), the time from 1.3 to 0.9 Ga is marked by enhanced anorogenic magmatism, as well as a relative minimum extent of continental margins and collisional belts (21,36–40). Liu et al.'s (16,22) observation of significant maxima in the Nb, Y, and Zr composition of Rodinian igneous rocks (Figure 20.2) amplifies evidence that Rodinian assembly was unique, while pointing to possible reasons for these differences.

The enrichments of Nb, Y, and Zr, coupled with the greater relative abundances of minerals of these three elements, point to a distinctive tectonic setting for Rodinia. Rodinian assembly was dominated by "non-arc" magmatism, in contrast to other intervals of supercontinent assembly when collision-related mineralization and island arc magmatism were of greater relative significance (41–46). In particular, these tectonic conditions at 1.3 to 0.9 Ga led to enhanced production of NYF-type (i.e. Nb-, Y-, and F-enriched) pegmatites, with associated increases in the occurrence and diversity of Nb-, Y-, and Zr-bearing minerals (46–49). This mineralization may have been associated with a warmer mantle and/or a thickened continental crust during Rodinian assembly (50,51) – characteristics that may reduce scavenging of high-field-strength elements by interaction with the depleted mantle during arc magmatism (52,53).

The relative enrichment of Nb, Y, and Zr contrasts with the behavior of many other elements during the period of Rodinian assembly. The minerals of most elements are notably lacking during the 1.3 to 0.9 Ga interval, as manifest in the relatively few ore deposits associated with the time of Rodinian assembly (30,31,36,38). However, the trace element concentrations of Co, Ni, and many other elements in igneous rocks do not show
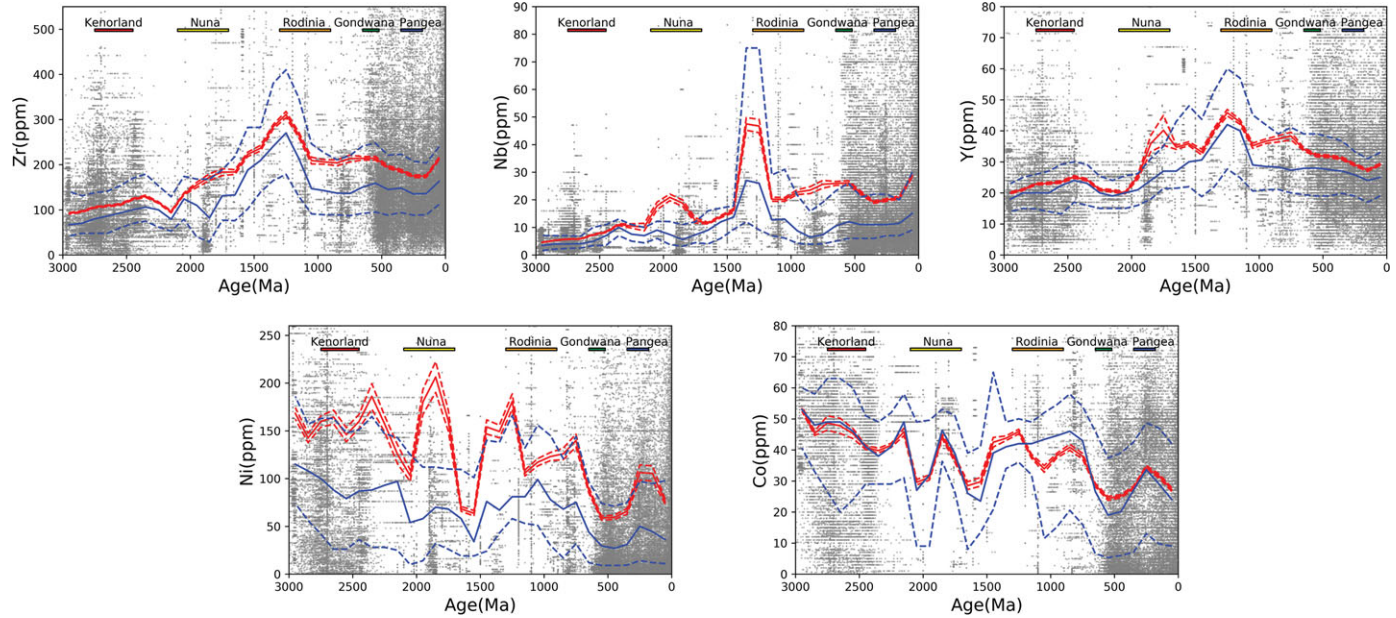
Figure 20.2 Trace element concentrations of Zr, Nb, Y, Ni, and Co in global igneous rocks through the last 3.0 Ga. Maximum values for Zr, Nb, and Y occur during and immediately before Rodinian assembly, in contrast to Ni and Co. Gray-filled circles are data resampled from earthchem.org with bootstrap resampling. Moving averages and medians of samples within ± 100 Ma bin sizes are calculated for each 100 Ma. Red solid lines are averages; red dashed lines are 95% confidence intervals of the moving average; blue solid lines are medians; blue dashed lines indicate the lower (25%) and upper (75%) quantiles

(after 16).

corresponding depletions compared to other supercontinent episodes (Figure 20.2) (16). Given this consistency in metal concentrations, reduced Rodinian ore deposition seems unlikely. Rather, the lack of Mesoproterozoic ore deposits may be a consequence of enhanced erosion of near-surface deposits that formed preferentially near active margins. This style of erosion was perhaps more characteristic of Rodinia than other supercontinents for two reasons. First, pre-collisional erosion of Rodinia may have been more aggressive than with other supercontinents, because the accretion of Rodinia is thought to have been both prolonged and "extrovert," with assembly by two-sided subduction (54–56). Such a tectonic context would have caused the loss of most volcanic-hosted massive sulfide deposits, which require rapid accretion of continental margins for preservation (38). Furthermore, the major orogens associated with Rodinian assembly experienced cycles of collisional distension that must have led to enhanced deep erosion. These processes are reflected in the high regional metamorphic grade of many surviving rocks associated with two major Rodinian sutures: the Grenville and Sveconorwegian orogens (20,57–60). Thus, for example, the absence of Rodinian-age gold deposits likely reflects removal of the shallower loci of mineralization, whereas the enhanced production of Grenvillian fluvial sediments led to the abundance of detrital zircon crystals of that age (61–63) Consequently, the observed distribution and diversity of minerals during the period of Rodinian assembly reflects a unique combination of mineralization events and preservational biases.

20.2.4 Implications for the Carbon Cycle

Tectonic events such as supercontinent assembly and dispersal have direct effects on carbon cycling at Earth's surface (64–66; Chapter 11, this volume). How might the distinct aggregation and breakup of Rodinia have influenced the carbon cycle and, related to this, redox conditions and life?

In principle, uplift and erosion associated with supercontinent assembly might have affected both atmospheric pCO_2 and nutrient fluxes into the oceans. Denudation rates of modern active margins (e.g. New Zealand, Taiwan) were reported to be highest on continents/islands – orders of magnitude higher than mountain belts (e.g. Alps, Himalaya) and shields away from the coast (67). The Rodinian supercontinent was proposed to be formed via closure of Pacific-type oceans (62,68), with abundant active but rare passive continental margins (38). On geologic timescales, continental erosion/weathering is the major sink for atmospheric CO_2 (69), and the high erosion/weathering rate of Rodinian active margins could have sequestered CO_2 more rapidly, paving the way for Neoproterozoic global glaciations (36). The fact that global ice ages postdate Rodinian assembly by more than 200 million years indicates that while Rodinian CO_2 drawdown might have contributed to later Proterozoic climate change, other factors must be considered as well.

Enhanced weathering and erosion had the potential to increase P fluxes into the oceans, thus promoting primary production. For example, the later Mesozoic and

Cenozoic uplift of major mountain belts appears to have impacted primary production, driving ecosystem-wide biological changes in the oceans (70). In addition, enhanced formation of rapidly subsiding sedimentary basins during the Rodinian breakup might have increased rates of organic carbon burial, thereby contributing to Neoproterozoic oxygenation (71).

We have several geochemical tools for exploring secular variations in carbon cycling, most notably the carbon isotopic record of carbonate and organic carbon (72). In addition, a variety of proxies permit inferences about changing redox conditions in the oceans and atmosphere (73), and fossils record the course of early evolution (74,75). Interestingly, supercontinental events correlate only weakly with the carbon isotopic, paleo-redox, and fossil records. Rodinian assembly correlates with a moderate increase in the secular variation of carbon isotopes, following a long interval of near-invariant values (76), whereas a much larger amplitude of C-isotopic variations is associated with the Rodinian breakup and its aftermath (77). Proxies for redox conditions show little change in association with either Rodinian assembly or breakup, perhaps because limited P availability (78) muted Earth system responses to these tectonic events. Global changes in oxygen levels and biological complexity occur only near the end of the Proterozoic Era, in association with a state change in P availability linked by some to climate rather than directly to tectonics (79).

Thus, at our present state of knowledge, the momentous tectonic events of Rodinian assembly and dispersal seem to have exerted only a limited influence on the surficial carbon cycle, with dispersal correlating more closely with enhanced organic carbon burial, perhaps minor oxygen enrichment, and protistan diversification (75) than with supercontinent assembly.

20.3 Use Case #2: Carbon Mineral Evolution, Mineral Ecology, and Mineral Network Analysis

Data-driven exploration is built on open-access data resources and the application of advanced analytical and visualization techniques. Databases, such as that of the RRUFF Project (rruff.info), which includes information on all approved mineral species, and that of mindat.org, which documents species found at more than 300,000 localities with greater than 1,000,000 mineral/locality data, provide opportunities to explore mineral data with new analytical tools. The effects of preservational and/or sampling bias in these data are poorly understood and are the subject of further investigation. The DCO has seized this opportunity by facilitating significant advances in the accumulation, analysis, and visualization of mineral data – notably information housed in the MED related to the more than 400 approved carbon-bearing mineral species (80–82). As such, carbon minerals constitute an important test case for new approaches to mineralogy, while providing unique insights into the evolving roles of carbon through deep time (Figure 20.3).
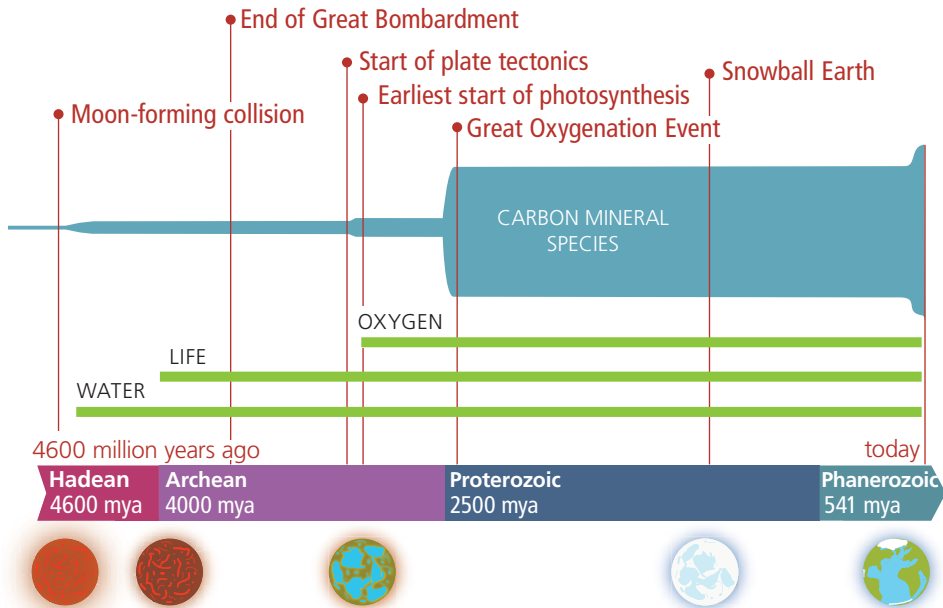
Figure 20.3 Carbon mineral evolution timeline over 4.5 billion years. Carbon played a key role throughout this evolutionary path, with an explosion in carbon mineral diversity in the Proterozoic and Phanerozoic.

20.3.1 Carbon Mineral Evolution

Mineral evolution is the study of the changing diversity and distribution of minerals through deep time – the consequence of varied physical, chemical, and, in the case of Earth, biological processes (11,83–85). Hazen et al. (80) surveyed carbon mineral evolution from a qualitative viewpoint, tracing changes in the nature and extent of carbon-bearing minerals through ten stages of Earth's evolution. From the most primitive Stage 1, characterized by chondrite meteorites, which contain several carbide minerals and allotropes of carbon, to the thriving terrestrial biosphere of Stage 10, with more than 400 approved carbon mineral species, Earth's 4.567-billion-year history saw significant increases in the diversity and complexity of C-bearing phases. The number of crystalline forms of C-bearing compounds has seen a dramatic rise with the creative contributions of chemists in the "Anthropocene Epoch" – an explosion of new mineral-like forms that some observers have dubbed "Stage 11" of Earth's mineral evolution (86,87).

The development of the MED (88), which tabulates 17,455 ages for C-bearing mineral/locality data (data as of May 21, 2018), facilitates a more quantitative examination of carbon mineral evolution. A detailed investigation of these minerals, including their paragenetic modes, associated species, geochemical contexts, tectonic settings, and other parameters, is beyond the scope of this chapter. However, an overview of the temporal
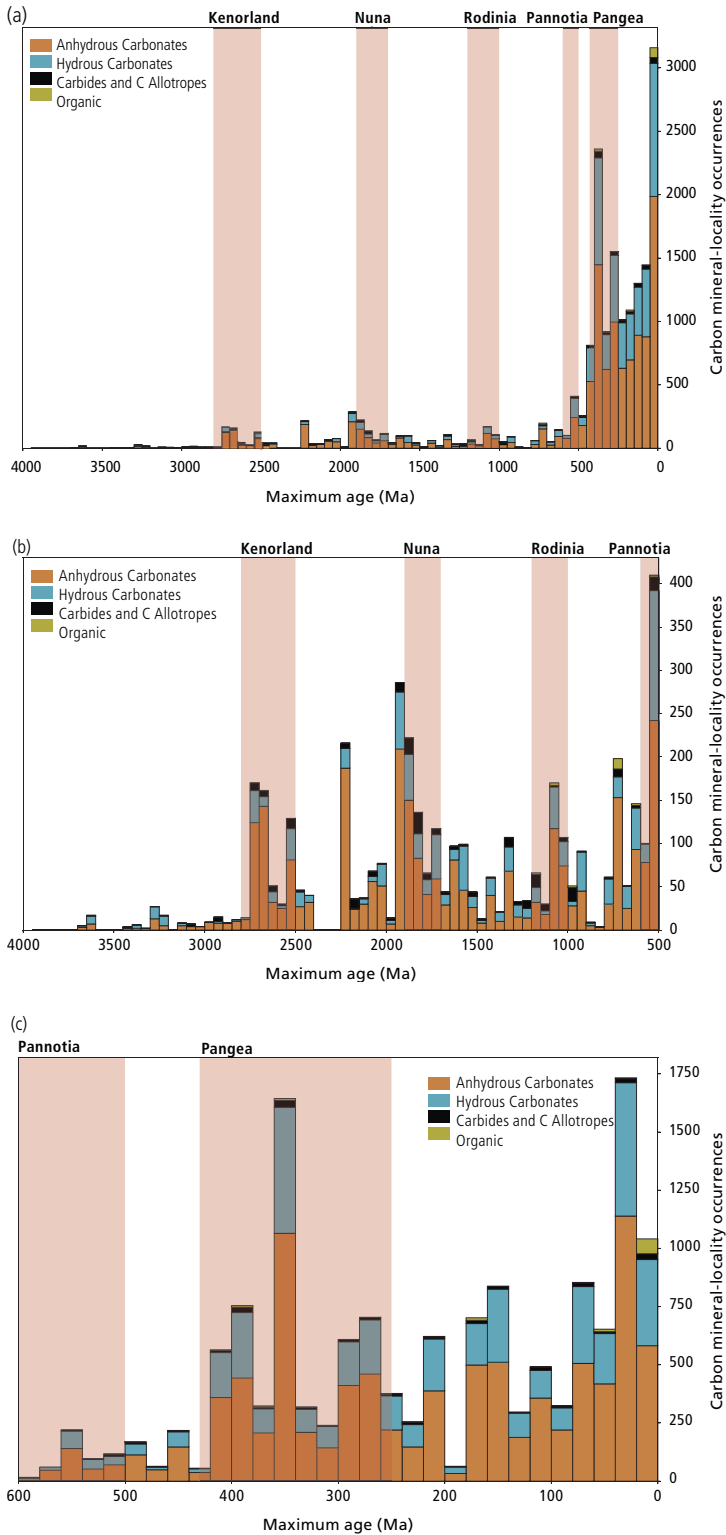
Figure 20.4 Temporal distribution of carbon minerals. (a) The past 4 billion years with 50-million-year bins. (b) Precambrian occurrences (4.0 to 0.5 Ga) with 50-million-year bins. (c) Neoproterozoic

distributions of C-bearing minerals reveals important physical, chemical, and biological processes that influence carbon mineralization. Figure 20.4 illustrates these newly expanded MED carbon mineral data.

The temporal distribution of carbon minerals reveals significant trends. As with most other groups of minerals, C-bearing species display striking episodicity, with pulses of mineralization as well as time intervals with few recorded carbon minerals. For example, significant maxima in preserved carbonate minerals are recorded at 2.75 to 2.70 Ga and at 2.55 to 2.50 Ga, with each interval having more than 150 points of reported carbon mineral/locality/age data. Those two 50-million-year intervals frame the assembly of Kenorland, the earliest well-documented supercontinent. By contrast, the 200-million-year interval from 2.45 to 2.25 Ga, a period of presumed Kenorland stability and generally low mineralization, has fewer than 20 total reported carbon mineral occurrences. As noted in Section 20.2, such a sharp contrast in numbers of mineral occurrences likely reflects a combination of episodic mineralization and preservational biases.

A similar contrast is observed for Nuna, the next widely recognized supercontinent episode in Earth's history. Approximately 800 mineral/locality/age data are recorded for the 250-million-year period of presumed Nuna assembly from 1.95 to 1.70 Ga. By contrast, the 250-million-year interval of Nuna breakup from 1.60 to 1.35 Ga is represented by fewer than 250 reports of C-bearing minerals.

Though less dramatic, the assembly of Rodinia is also reflected in the carbon mineral record. Approximately 400 mineral/locality/age data are recorded for the assembly period from 1.1 to 0.9 Ga, as opposed to fewer than 20 data points from the subsequent 100-million-year interval from 0.9 to 0.8 Ga. As suggested in Section 20.2, the relatively modest mineral inventory from Rodinian assembly likely reflects significant erosional loss of near-surface (i.e. more carbonate-rich) deposits compared to Kenorland and Nuna.

Approximately 80% of reported carbon mineral occurrences in the MED are from the Phanerozoic Eon, which spans the last 540 million years when carbonate biomineralization became an important mode of near-surface carbon mineralization. The greater number of data from the Phanerozoic Eon allows a more detailed examination of carbon mineral evolution during the past 500 million years. Figure 20.4c underscores the nonuniform distribution of documented carbon mineralization during the past 600 million years. Of note is that almost 1700 mineral/locality/age data are recorded from the 20-million-year interval from 360 to 340 Ma, a time of the supercontinent Pangaea's assembly, and thus a plausible time of enhanced mineralization and preservation.

Figure 20.4 (*cont.*) and Phanerozoic occurrences (760 to 0 Ma) with 20-million-year bins. Anhydrous carbonates (orange, lowest segment), hydrous carbonates (blue, next lowest segment), other (i.e. diamond and carbides, black, next lowest segment), and organic minerals (green; topmost segment). Graphs are based on 17,455 mineral/locality/age data tabulated in the MED (rruff.info/ima; as of February 15, 2018). Note that this tabulation is based on mineral specimens collected from specific localities and does not include sedimentary carbonate formations.

An important concurrent event was the expansion of late Paleozoic ice sheets in Gondwana, a scenario linked to enhanced burial of organic matter associated with the evolution of trees and diversification of seed plants, stem group ferns, and lycopods. This interval was also notable for the 359 Ma Devonian–Mississippian boundary, which marks the last pulse of elevated extinctions that occurred through much of the Devonian Period. A notable degree of ecological reorganization also occurred in marine environments, including the complete turnover of rugose corals, a once-abundant order of corals that are now extinct, at the family level. It is not obvious how these paleobiological developments might have led to enhanced mineralization, although it is possible that at least some of the observed paleobiological events might reflect responses to tectonic events and their environmental consequences, as recorded by carbon mineral occurrences.

By contrast, the interval from 200 to 180 Ma is represented by fewer than 15 C-bearing mineral/locality/age data points worldwide. This 20-million-year period occurred at the beginning of Pangaea's breakup and the opening of the modern Atlantic Ocean, a time characterized by tectonic conditions that might be associated with reduced carbon mineralization or deposition and enhanced erosional loss. The beginning of this interval corresponds to the end-Triassic mass extinction associated with massive volcanism, whereas a minor extinction event at 182 Ma is also associated with a large igneous province (89). However, neither of these short intervals of species loss have obvious connections to the mineral record.

Note that the distribution of mineral occurrences during the Precambrian at 50-million-year intervals (Figure 20.4b) is not unlike the peak distributions of the Phanerozoic Eon at 20-million-year intervals (Figure 20.4c). An unresolved question in mineral evolution research is the extent to which the temporal distribution of mineral groups, including C-bearing species, is fractal; in other words, does the same pattern of mineral distribution repeat at finer and finer temporal scales? This question can only be answered by gathering many more mineral/locality/age data with the highest possible time resolution. We are currently limited to the 195,000 mineral/locality/age data compiled in the MED, but there are likely many more data yet to be extracted from the existing literature, as well as many rock and mineral samples that have yet to be analyzed. For instance, rock-forming minerals are particularly underrepresented in the MED simply due to sampling bias.

20.3.2 Carbon Mineral Ecology

Mineral ecology is the study of mineral diversity–distribution relationships of minerals at the global scale –an effort that depends on large and growing data resources on mineral species and their localities on Earth's crust. Hazen et al. (81) applied a large number of rare events (LNRE) formalism (90–93) to model the distribution of 403 approved mineral species of carbon. Using 82,922 mineral species/locality data tabulated in mindat.org (as of January 1, 2015), they demonstrated that all C-bearing minerals as well as several compositional subsets containing C conform to LNRE distributions.

Number of minerals found at m localities

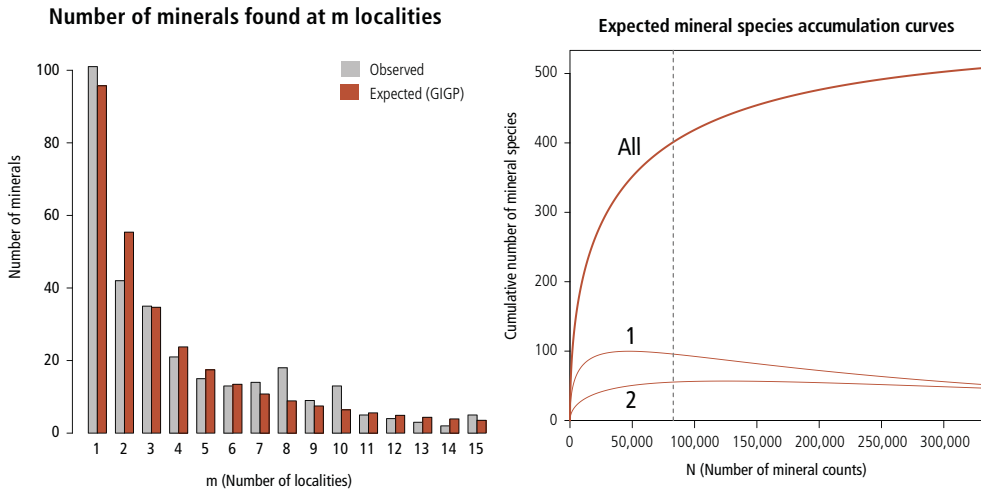**Expected mineral species accumulation curves**

Figure 20.5 (a) Frequency spectrum analysis of 403 C-bearing minerals, with 82,922 individual mineral-locality data (from mindat.org as of January 2015), employing a generalized inverse Gauss–Poisson (GIGP) function to model the number of mineral species for minerals found at between 1 and 15 localities (90). (b) This model facilitates the prediction of the mineral species accumulation curve (upper curve, "All"), plotting the number of expected C mineral species (y-axis) as additional mineral species-locality data (x-axis) are discovered. The vertical dashed line indicates data recorded as of January 2015 in mindat.org. The model also predicts the varying numbers of mineral species known from exactly one locality (curve 1) or from exactly two localities (curve 2). Note that the model predicts that the number of C-bearing mineral species known from only one locality is now decreasing, whereas the number from two localities is now increasing, though it will eventually decrease. We predict that the number of minerals known from two localities will surpass those known from one locality when the number of species-locality data exceeds ~400,000.
Reproduced from Hazen et al. (81) with permission.

The LNRE model is particularly useful because it can be used to determine an "accumulation curve" – a formalism that enables estimations of the probability that the next carbon mineral/locality discovery will represent a new species (Figure 20.5). Figure 20.5a displays the frequency spectrum analysis for 403 C-bearing mineral species based on 82,922 individual mineral-locality data (from mindat.org as of January 2015). We found that 101 minerals – more than 25% of known C-bearing species – have been identified from only one locality worldwide. Another 42 species have been found at exactly two localities. Based on this information, we employed a Generalized Inverse Gauss–Poisson function to model the number of mineral species for minerals found at between 1 and 14 localities (90).

This LNRE model facilitated the prediction of the mineral species accumulation curve (Figure 20.5b). In Figure 20.5b, the upper curve (labeled "All") plots the expected number of approved C mineral species (y-axis) as additional mineral species/locality data (x-axis) are discovered. The vertical dashed line indicates data recorded as of January 2015 in mindat.org. The model also predicts the varying numbers of mineral species known from exactly one

locality (curve "1") or from exactly two localities (curve "2"). Note that the model predicts that the number of C-bearing mineral species known from only one locality is now decreasing, whereas the number from two localities is now increasing, though it too will eventually decrease. We predict that the number of minerals known from exactly two localities will surpass those from one locality when the number of species-locality data exceeds ~400,000.

Employing this model, Hazen et al. (81) predicted that at least 548 carbon mineral species occur in Earth's crust today –a result that suggests at least 145 C-bearing minerals exist but have yet to be discovered. Additional hints regarding the nature of these "missing" carbon minerals are gleaned by analyzing compositional subsets of common additional elements in C-bearing minerals, including oxygen, hydrogen, calcium, and sodium. Accordingly, Hazen et al. (81) predicted that 129 missing carbon minerals contain oxygen (primarily carbonates) and 118 species contain hydrogen (mostly hydrous carbonates). In addition, more than 50 of the missing species contain calcium, while more than 60 contain sodium. Additional studies of the distributions of known minerals according to their distinctive sizes, colors, crystal forms, and physical properties (93) suggest that many of the missing carbon minerals may have been overlooked because they are colorless, poorly crystalized, water soluble, and/or occur in minute grains. Similarly, these same factors are likely why nearly 35% of Na minerals have yet to be discovered and, conversely, why fewer than 20% of Cu, Mg, Ni, S, Te, U, and V minerals are still unknown (93). This powerful data-driven approach has allowed the systematic prediction and discovery of large numbers of previously unknown mineral species for the first time.

These newly applied data analytic methods have led to DCO's Carbon Mineral Challenge (mineralchallenge.net), which enlists professional mineralogists and amateur mineral collectors around the world in the search for new species. More than 30 new carbon minerals – roughly 20% of the predicted total missing inventory – have been reported since January of 2016. Two of those species, abellaite ($NaPb_2(CO_3)_2(OH)$) and parasite-(La) ($CaLa_2(CO_3)_3F_2$), were predicted as possible new carbon minerals by Hazen et al. (81). Other new carbon species were not predicted. Of note is the organic mineral tinnunculite ($C_5H_4N_4O_3 \cdot 2H_2O$), which crystallizes when the excrement of the kestrel, *Falco tinnunculus*, bakes in the hot gases of a burning coal fire. Though tinnunculite was not anticipated by our analysis, we did predict that several new organic minerals would be included in the list of new finds.

Mineral ecology and data-driven approaches to predicting and discovering new mineral species (as well as valuable mineral resources identified using similar statistical approaches) are in their infancy. In addition to further studies of carbon mineral ecology on Earth, efforts concentrating outward, focusing on other planetary bodies, will be necessary. Some work has begun, including hypothesizing the mineral diversity of Saturn's moon, Titan (94,95). Maynard-Caseley et al. (94) propose a rich, diverse population of carbon minerals, specifically organic molecular minerals, on Titan's frozen surface. The applications of such data-driven methods as cluster analysis, network analysis, and affinity analysis to mineral systems are poised to revolutionize the way we think about the

diversity and distribution of minerals on Earth and other worlds by providing a more complete, multivariate understanding of these systems.

20.3.3 Carbon Mineral Network Analysis

Advances in data-driven discovery rely on application of creative analytical and visualization methods to complex multi-dimensional systems. Mineral network analysis (82) is a particularly powerful approach to understanding complex relationships among mineral species, their localities, paragenetic modes, and varied physical and chemical properties.

Figure 20.6a displays a force-directed network graph in which colored circles (nodes) indicate C-bearing mineral species, while lines between circles (edges) denote coexisting pairs of minerals. The sizes of nodes indicate the relative abundances of the minerals, while colors represent major C-bearing mineral groups. In this force-directed graph, each edge has an optimal length like a spring; edges are stretched or compressed to achieve a "lowest energy" state for the entire network. Similarly, Figure 20.6b shows a bipartite network of 403 C-bearing mineral species from approximately 300 mineralized regions on Earth. These graphs are interactive; each node can be clicked and dragged to more closely examine the number and nature of edges (see dtdi.carnegiescience.edu for interactive renderings).

An important characteristic of network visualizations is that they can be analyzed with numerous metrics, each of which quantifies aspects of the local and global distributions of nodes and links (96–98). For example, the carbon network (Figure 20.6a) has density $D = 0.24$ (i.e. 24% of all possible edges are present) – a value that is intermediate between those of copper minerals ($D = 0.12$) and igneous minerals ($D = 0.64$) (82). The network diameter, which measures the maximum degree of separation between any two network nodes, is $d = 4$, while the network affinity is $a = 0.55$.

One of the surprising findings related to networks of minerals is that they may embed information not coded into the network layout. For example, a slight chemical trend is visible in Figure 20.6a, with nearly all of the anhydrous carbonates not containing transition elements, lanthanides, and/or actinides (orange nodes) plotting on the left side of the network and the majority of the organics and hydrous carbonates containing transition elements, lanthanides, and/or actinides (green and purple nodes, respectively) plotting on the right. In Figure 20.6b, a few trends regarding the diversity and distribution of minerals in space and time are evident. First, the "U-shaped" distribution of black locality nodes, with a few very common carbon minerals "inside" and many more rare carbon minerals "outside," is a visual representation of the LNRE distribution illustrated in Figure 20.5. Second, there is an embedded timeline, with the oldest minerals in the center of the locality "U" radiating outward as the mineral species' age of first occurrence becomes younger.

Mineral network analysis, a direct outgrowth of interactions among diverse members of the DCO community, is in its infancy. We anticipate that open-access data resources, as well as freely available analytical and visualization software, will lead to a transformation in the ways that we study complex mineral systems on Earth and other worlds.
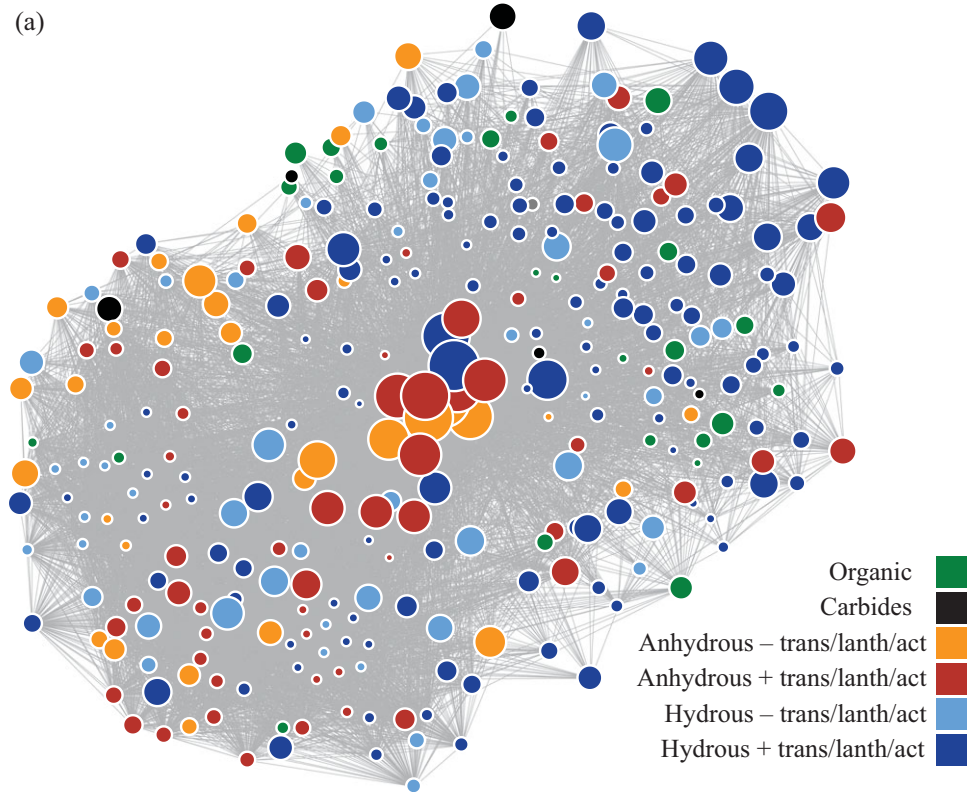
(a)

Figure 20.6 (a) Force-directed, unipartite network graphs of 403 C-bearing mineral species. Nodes represent C-bearing mineral species, while lines between nodes denote coexisting pairs of minerals. Node diameters indicate the relative abundances of the minerals, while colors represent compositional groups (dark blue = hydrous carbonates with transition elements, lanthanides, and/or actinides; light blue = hydrous carbonates without transition elements, lanthanides, and/or actinides; red = anhydrous carbonates with transition elements, lanthanides, and/or actinides; orange = anhydrous carbonates without transition elements, lanthanides, and/or actinides; black = carbon allotropes and carbides; green = organic minerals). (b) Force-directed, bipartite network of 403 C-bearing mineral species and their localities on Earth (see also http://dtdi.carnegiescience.edu/node/ 4557 for an interactive version). Colored nodes represent carbon mineral species, with node size corresponding to the frequency of occurrence and color corresponding to the age of earliest known occurrence of each mineral species. Black nodes represent regional localities, with diameter corresponding to the relative numbers of distinct C-bearing mineral species found at each locality. The network rendering reveals important information regarding the diversity and distribution of carbon minerals through space and time. In particular, the "U-shaped" distribution of black locality nodes, with a few very common carbon minerals "inside" and many more rare carbon minerals "outside," is an alternative visual representation of the LNRE distribution illustrated in Figure 20.5.
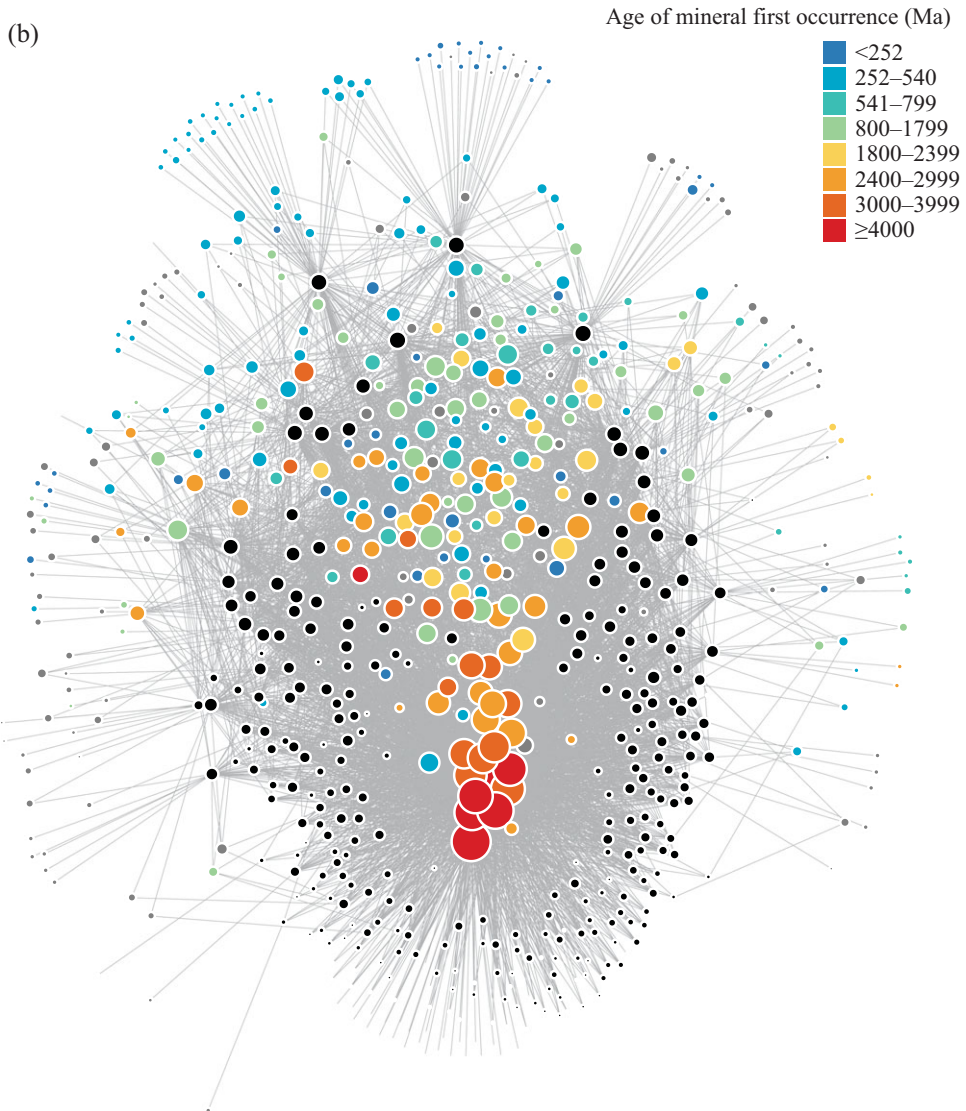
(b)

Age of mineral first occurrence (Ma)

■	<252
■	252–540
■	541–799
■	800–1799
■	1800–2399
■	2400–2999
■	3000–3999
■	≥4000

Figure 20.6 (*cont.*)

20.4 Use Case #3: Enzyme Evolution and the Environmental Control of Protein Expression

Microbes in Earth's crust have played key roles in the carbon cycle throughout space and time (99; Chapters 17 and 18, this volume). In order to better understand "whole-Earth carbon," we must examine the relationships among: (1) the physical and chemical characteristics of varied microbial environments (Chapters 16 and 19, this volume); (2) the metabolic strategies adopted by microbial consortia in these environments (Chapter 17,

this volume); and (3) the consequent variation of microbial gene molecular function and expression (Chapter 18, this volume). The exploration of the complex interconnections among the physical, chemical, and biological aspects of microbial ecosystems represents an as yet unrealized opportunity for understanding the coevolving geosphere and biosphere.

A fundamental stumbling block in documenting the role of microbes in Earth's carbon cycle through deep time is the lack of relevant data on the nature and expression of proteins in ancient microbial ecosystems. In spite of the occasional preservation of Precambrian microfossils, scant biomolecular traces survive in ancient rock formations (100–102). Therefore, an understanding of the biochemical evolution of microbes might seem beyond our reach.

A promising strategy to understand aspects of the coevolution of geochemical and biochemical systems is based on the analysis of the large and growing data resources describing microbial ecosystems. Extant microbial communities span a wide range of physical and chemical environmental conditions (e.g. high and low pH, temperature extremes, high salinity and pressure, low consumable resource availability, and low water activity), some of which likely mimic a range of ancient conditions extending back to the dawn of life (103). While extant microorganisms living in these ecosystems are modern organisms that coevolved with our planet and adapted to its changing conditions, they still harbor ancestral metabolic traits. Consequently, today's microorganisms contain both inherited traits as well as recently acquired ones.

Considering that ancient protein structures and functions are at least to some extent conserved in modern organisms, then modern analogs of presumed ancient environments may resemble life's earliest enzymatic systems. For instance, extant strict anaerobes that inhabit anoxic, geothermal environments must have inherited the metabolic machinery necessary to conserve energy using redox couples abundant in geothermally influenced environments (e.g. hydrogen and sulfur) and to fix carbon dioxide of magmatic origin (103). These same organisms also must have acquired the ability to cope with reactive oxygen species in order to adapt as atmospheric oxygen levels on Earth increased over the last 700 million years. However, being unable to accurately differentiate new adaptations of older functions from truly new innovations complicates the process of reconstructing the emergence and evolution of metabolisms. The integration of large data sets obtained from the study of extant microorganisms and their protein structures, coupled with detailed environmental, geochemical, and mineralogical information, may allow us to better understand the emergence and evolution of microbial metabolism. In particular, it may provide new insights into how the geosphere and biosphere have coevolved, ultimately resulting in the complex network of metabolic reactions we see today (104,105).

Here, we propose strategies for applying methods of data analysis and visualization in order to answer questions about microbial ecology, protein evolution, and their relationship to carbon mineralization through deep time.

20.4.1 Network Analysis of Protein Structures: Geo–Bio Interactions on Evolutionary Scales

Methods of network analysis are well suited to the exploration of the evolution of and relationship between protein structure and function (106–108). The combination of geo-chemically identifiable timescales with biologically determined timelines permits glimpses into the history of life on Earth. For example, Bromberg and colleagues have employed similarity networks to analyze relationships among the structures of nearly 4700 oxidor-eductases from varied microbial and multicellular organisms. Since electron transfer reactions are necessary to fulfill the energy requirements of all life-forms, the ability to carry out redox reactions must have been among the first functions acquired by early life. Understanding the evolution of biological redox machinery can thus shine light on the history of life and on its interactions with Earth's environment.

Ideally, the evolution of redox abilities could be traced through the analysis of the relevant enzyme sequences. However, the origins of biological redox, which likely corres-pond to the origins of life, as well as the dramatic environmental changes that have since taken place (e.g. the Great Oxidation Event and the "fold explosion" of protein structures), are ancient. This fact makes the exploration of the mutations in sequence space that led to the current biological "state of the art" nearly impossible (109,110). Protein three-dimensional structures, on the other hand, retain evolutionary evidence for significantly longer stretches of time. Note that the process of the divergent evolution of folded structures implies that existing folds emerged from prior ones. However, functionally similar folds may also arise independently via convergent evolution. Using network analysis, augmented by metrics of sequence similarity in structural alignments, it is possible to trace distant relations between redox proteins to estimate whether they have common ancestors or whether they developed independently.

Bromberg et al. have created a method, *sahle* (structure-annotated homology, ligand-extended), for evaluating the reliability of *structural* similarity of transition metal binding sites in proteins, defined as spheres of 15-Å radius from the active metal-containing site (111). A *sahle* score, ranged 0–100, gives weight to an edge between two spheres/nodes in the resulting network (Figure 20.7). The color of the nodes indicates the primary metal at the active site of a given sphere in a protein. Interestingly, network connectivity illustrates that the biological use of metals may be traceable through evolutionary time; in other words, the earliest proteins preferentially incorporated Fe, with later proteins using Mn and then Cu –the same sequence seen in the network graph – although metal information was not explicitly encoded in the network topology. This network reinforces previous findings from geochemistry (112) and biochemistry (104,105) that suggest that Fe proteins are ancient, whereas Cu-bearing proteins evolved later, possibly related to the presence and bioavailability of Fe and Cu in Earth's oceans through deep time (113).

An important finding of these and other network applications is that graphs of evolving systems (i.e. fossil taxa or mineral species) inevitably embed a time axis (Figure 20.8). This discovery points to possible data-driven approaches to gaining insights into the evolution
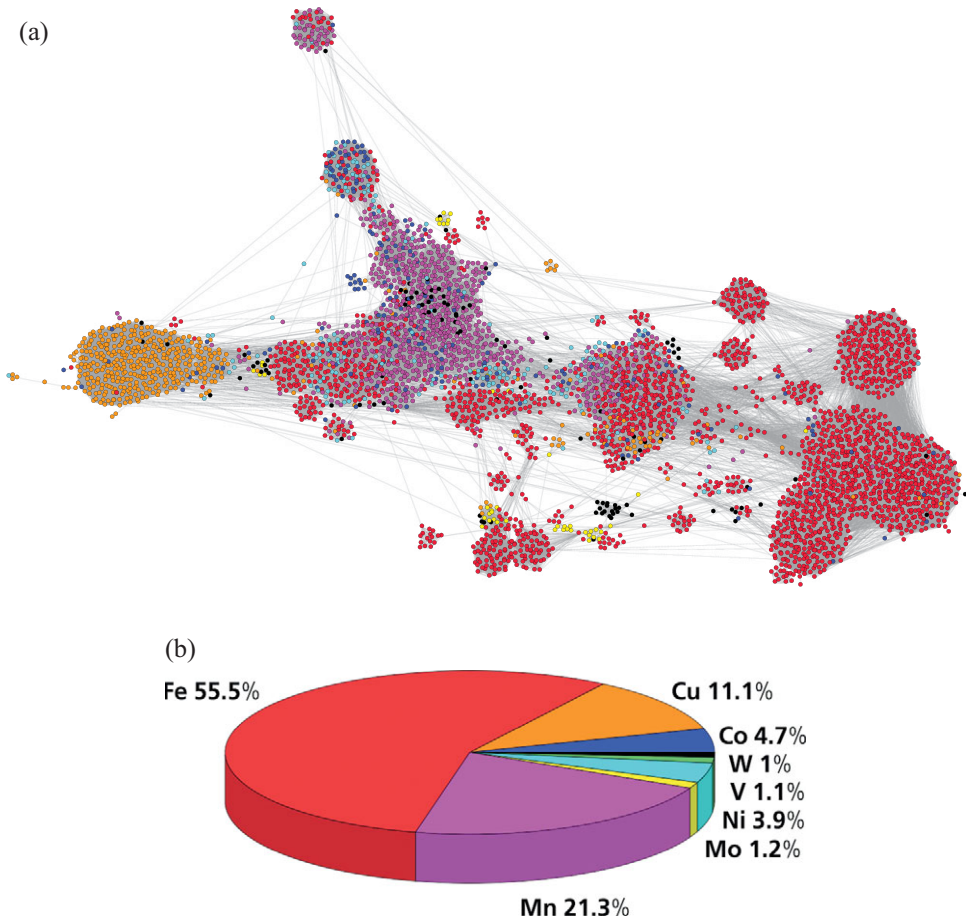
(a)

(b)

Figure 20.7 (a) Similarity network diagram of relationships among protein structures. The 4686 circular nodes represent oxidoreductases for which the three-dimensional structure is known. The linking and therefore distribution of nodes relates to similarities in protein fold structure in a 15-Å radius from the active metal-containing site. Nodes are colored according to the principal metal cation at the active site. Network connectivity illustrates that functional similarity of spheres may be traceable through evolutionary time (i.e. although metal information was not used in the building of this network, the time-related sequence Fe to Cu is embedded). This network indicates that Fe proteins are ancient, whereas Cu proteins evolved later. The pie chart (b) shows the relative abundance of metals in the graph.

of specific protein groups. For example, clustering of spheres in the network provides a means for reducing experimental bias in favor of generating a more naturally representative set of nodes and edges, which can be further used to build evolutionary trees of redox reactions on global timescales. These approaches can also inform synthetic biology, directing possible experimental mutagenesis efforts for designing and evaluating evolutionary intermediates that no longer exist in nature.

(a) **Fossil network of co-occuring marine animal orders**
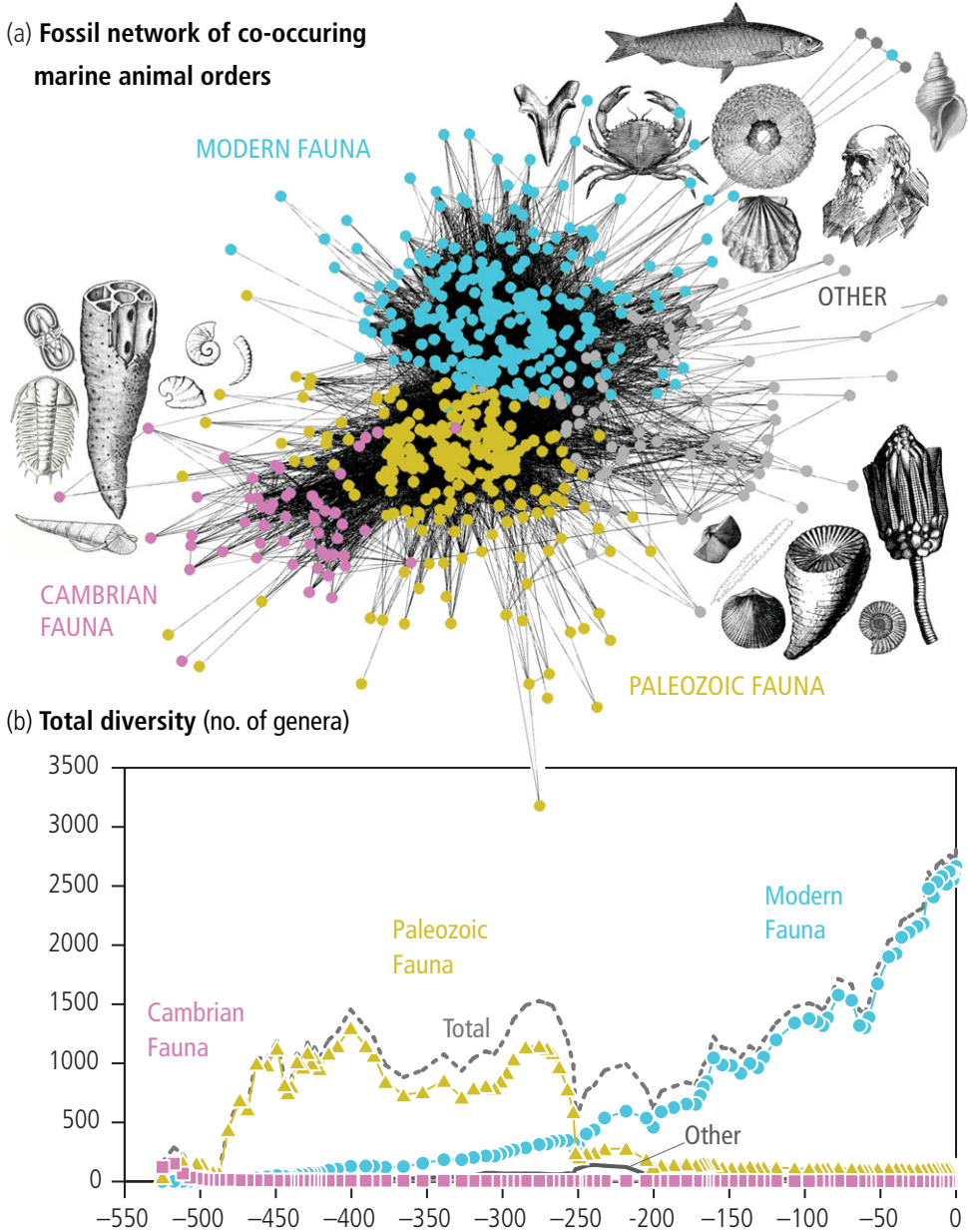

(b) **Total diversity** (no. of genera)

Figure 20.8 Networks that illustrate structural or coexistence relationships among individual members of an evolving system (i.e. mineral species, fossil taxa, or protein structures) inevitably embed a time axis, even though no age information is used in the generation of the graphs. (a) Phanerozoic fossil animals: nodes represent family-level taxa, while lines indicate coexisting fauna. The network was partitioned using the Louvain (multilayer) algorithm for community detection (138), resulting in the discovery of five modules, or "evolutionary paleocommunities." An embedded time axis is visible from the Cambrian to modern fauna and each partition represents a major extinction event. (b) Plot of diversity (total number of genera) versus time for each of the modules in (a).

The inherent flexibility of network approaches allows for the incorporation of additional data, thus strengthening any inferences made. For example, as there are no protein fossils that can be used to establish dates of redox protein existence, one reliable piece of information that can be used for this purpose is transition metal availability, which would drive the selection of the molecular functionality necessary for life. By matching the currently existing microbiome molecular function (114) and metal cofactor annotations with mineralogical and geochemical data, it is possible to reveal the relationships between the presence and abundance of specific enzymatic functionalities and metal availabilities. Functional annotations can thus be mapped to metal availability and, further, to the corresponding evolutionary age. Additionally, using machine learning techniques to recognize patterns in molecular function to metal availability relationships, it is likely possible to pinpoint any discrepancies between expectations and existing annotations, suggesting areas for more extensive research. As a result, protein structure networks, in combination with geochemical evidence, could provide a glimpse into the emergence and evolution of life on our planet and an understanding of the principles that could govern life on other planets.

20.4.2 Network Analysis of Extant Microbial Ecosystems: Geo–Bio Interactions on Ecological Scales

Investigations of the relationships between individual microbial taxa, microbiomes, and environmental conditions are complicated by the large number of contributing physical, chemical, and biological parameters, culminating in a complexity that is not easily representable by two-dimensional graphical methods. It has been suggested that new analytical techniques will be necessary to explore the large data sets produced by high-throughput DNA sequencing to discover new connections between microbiomes and the environment (115). Quantitative gene content analysis of terrestrial and marine microbial communities has already revealed habitat-specific fingerprints that reflect known characteristics of the sampled environments (116). Metagenomic and amplicon sequencing of diverse environments and microbial communities are now paving the way toward outlining the global ecosystem network and the development of ecosystem-wide dynamic models (117,118).

Network analysis and machine learning can be used to investigate microbial communities from all types of ecosystems and are useful approaches for examining and determining patterns in large, complex data sets, and they provide predictive power in the absence of mechanistic models (115,119–121). Since microbes are notoriously difficult to culture, the primary source of information on their diversity and evolution comes from the environmental distribution of microbiome data (122,123). Metagenomics – the study of genetic material obtained directly from environmental samples – has opened the door to the incredible diversity of microbial communities in the biosphere. The large-scale analysis of metagenomes, in concert with a wide range of environmental characteristics and geological diversity, will allow for the identification of unknown geo–bio interactions in the near

future. This opportunity may lay the foundations for better understanding the geosphere and biosphere and their coevolution on this planet. As of the time of writing (January 2018), there were 6983 metagenomes available on the Department of Energy Joint Genome Institute public database (https://img.jgi.doe.gov), covering a variety of environments. Identifying relationships among physical and chemical parameters, such as temperature, pH, salinity, geochemistry, and the diversity in microbial communities, can reveal microbial responses to changing environmental conditions, and such information is critical to understanding microbial adaptations to different environments and their functions within those environments. Many studies have already shown the strong links between environmental conditions and microbial populations, a number of which did so with network analytical approaches (115,118,122,124–131). We suggest that the application of advanced analytical approaches to the microbial metagenomes and their corresponding environments, coupled with geochemical, geological, and mineralogical information, could transform the way we understand the role of microbial diversity in ecosystems.

Sharing and relating data sets between different disciplines, however, remains a great challenge. One way to deal with this challenge is through ensuring online availability of data. Currently, large amounts of sequenced data that represent a substantial portion of the total environmental diversity of Earth reside in online databases (e.g. MG-RAST, NCBI, JGI IMG, CAMERA). However, the quality of the associated metadata is generally low, with essential information like pH, temperature, salinity, redox state, and organic load often missing (132). Moreover, the links among sequence data, metadata, and any geochemical, geological, or other environmental data collected during the study are difficult or impossible to establish. Numerous attempts are being made by the scientific community to standardize the quality and type of metadata collected along with each sequenced sample in order to increase interoperative power. For example, efforts from the Genomic Standard Consortium (gensc.org) such as the Minimum Information about a Metagenomic Sequence (133), initiatives like the Earth Microbiome Project (earthmicrobiome.org), and the release of metadata-curated metagenomes (134) are pointing the metagenomics community in the right direction. Pioneering data sets of interdisciplinary, colocated data have been collected by the International Continental Drilling Programs (icdp-online.org), the International Ocean Discovery Program (iodp.org), and the DCO Integrated Field Site Initiatives (deepcarbon.net). These sampling programs will provide unprecedented environmental, geological, and geochemical metadata to analyze along with the associated metagenomes. Expansion of these efforts is crucial for advancing this important work in the future toward understanding geo–bio interactions on a global scale.

Our ability to generate predictive models of the relationships between -omic data and environmental data is further hindered by the varying data structures specific to the different fields of study (135,136). The poor resolution of our current understanding of the relationship between functional diversity and redundancy, biodiversity, ecosystem roles, and niche partitioning also presents challenges. A possible way to overcome this problem is by using predictive models that are not linked to specific hypothesis but take advantage of big data approaches that allow data-driven discoveries. Tools such as network

analysis and machine learning can identify hidden patterns in large-scale data and provide predictive power in the absence of mechanistic models (115,119–121). Similar techniques have been used in metagenomic modeling to predict microbial assemblages and their metabolic properties (e.g. 113–115,137), and they can be applied to the investigation of the interaction between the geosphere and biosphere.

Recently, we have attempted a preliminary exploration of large-scale patterns in the relationships among oxidoreductase metalloproteins and the mineral diversity present at the same location (Figure 20.9). Based on publicly available metagenomic data from
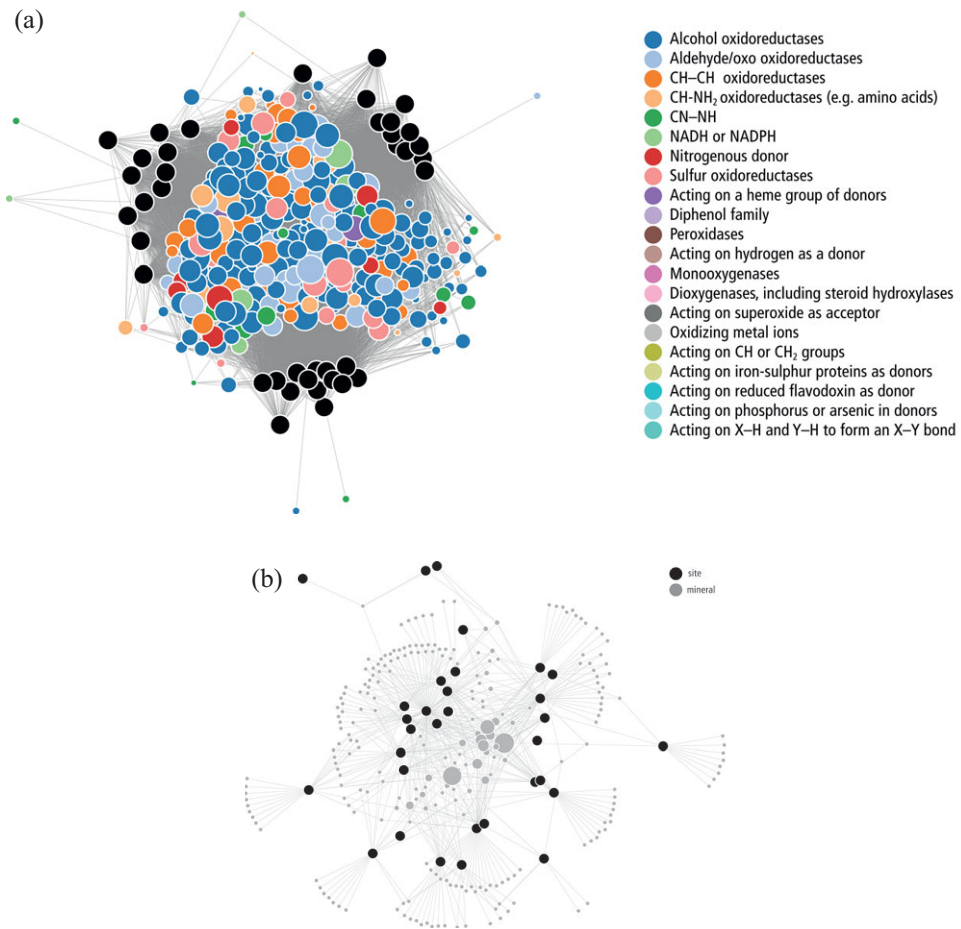

Figure 20.9 Bipartite networks of our preliminary analysis of geo–bio interactions based on 40 random metagenomes downloaded from MG-RAST and the mineral composition of the same site obtained from the Mindat database. (a) Bipartite network of the metalloprotein oxidoreductases (enzyme commission EC1 class) and the sites where they were found (in black). Enzyme nodes sized according to their counts and colored by their subclass. (b) Bipartite network of the mineral diversity at the same sites. Mineral nodes in gray, sized according to their mineral diversity; site nodes in black.

40 randomly selected microbial ecosystems (including samples from shallow-water and deep-sea hydrothermal vents, hot springs, permafrost, mines, soils, arctic soils, marine sediments, and salt marshes), our analysis reveals distinct patterns in the association between specific metalloprotein functions and the mineral settings where those functions are commonly abundant. In particular, geochemistry and redox conditions govern oxidoreductase gene diversity distribution in the observed environments. The microbial communities of certain locations had few or no distinctively expressed oxidoreductase proteins within the network, thus exhibiting overlap with other communities with similar environmental conditions. However, microbial communities from most locations expressed unique oxidoreductases that were not present in the communities of the other environments. This information is crucial to understanding niche partitioning among environmental taxa and may reveal key details regarding how environmental conditions and metal availability shape microbial community function.

We expected a great deal of overlap in gene expression between the microbial populations of many environments as we observed in our initial analysis. These functions will shed light on the expected and unexpected core functions of diverse communities. Additionally, numerous genes that are exclusively expressed in particular environments or under distinctive physical/chemical conditions will reveal geo–bio interactions that evolved in systems that are ancient Earth analogs to the modern day. We conclude that expanding data resources on microbial communities and ecosystems and better integration with geochemical, mineralogical, and geological databases will provide opportunities for documenting the effects of environmental parameters on gene distribution and functional diversity.

20.5 Conclusions: The Future of Data-Driven Discovery

Among the DCO's enduring legacies, and a tremendous opportunity for future advances, is the continued development and exploitation of data resources in the geosciences and biosciences. Our experiences over the decadal adventure of the DCO have convinced us that further advances in data-driven discovery will rest on three coequal pillars. The first ongoing demand is the creation and enhancement of comprehensive data resources, including those in geochemistry, petrology, mineralogy, paleobiology, paleotectonics, microbiology, proteomics, and other deep time aspects of carbon's global cycles in space and time.

Hand in hand with database enhancement, we require the development and adaptation of established and new methods for data analysis and visualization. Ongoing advances include new techniques to exploit geochemical data, novel LNRE formulations designed for specific applications to mineralogical and paleobiological systems, modified approaches to visualizing networks of varied geological and biological systems, and applications of affinity analysis to Earth systems.

Thirdly, data-driven discovery will advance through continued creative application of data resources and analytical methods targeted to answer complex problems related to

Earth's evolution through space and time. Our ambitions for the coming years include: estimating the erosional bias of the ancient rock record from differential mineral preservation through deep time; investigating the completeness of the fossil record with LNRE methods applied to the Paleobiology Database (paleobiodb.org); creating interactive networks of all known mineral species, fossil genera, and microbes and their environmental contexts; and applying affinity analysis to the discovery of new mineral and ore deposits.

The DCO has fostered the beginning of the era of data-driven discovery in carbon mineral science and has promoted the collection and assembly of a wide range of data resources. The DCO has employed existing analytical and visualization methods while developing new approaches and has raised and refined a suite of fundamental questions about Earth's carbon from crust to core – its forms, movements, quantities, and origins. Looking forward to the next decade of exploration, we predict that data-driven discovery will play an ever-greater role in our emerging understanding of carbon in Earth.

Acknowledgments

Questions for the Classroom

1 What are the "three pillars" of data-driven discovery and why are all three important?
2 What are some of the visualization methods that can enhance discovery and how many different parameters can be displayed simultaneously with each of these methods?
3 Why are time axes embedded in network graphs of evolving systems, even though no age information is used in the generation of these graphs?
4 What was "Rodinia" and what is the evidence for its unique signature in Earth's history?
5 What are some of the preservational biases likely affecting the rock record and how do these biases scale with time?
6 How many carbon mineral localities are in the MED today and how many of those localities are dated? Which locality has the most carbon mineral species?
7 What are the biases in sampling the carbon minerals listed in the text and what are additional biases not covered in the chapter?

8 What is an LNRE distribution and why is it a useful model for mineral distributions?

9 To what other systems could you apply an LNRE model and associated accumulation curve?

10 What factors might be important in describing a microbial ecosystem, such as a community of microbes living beneath the ocean floor?

11 What is a metagenome and how is it sequenced? Why is shotgun metagenomics used instead of pure cultures?

References

1. Darwin C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray, 1859.

2. Beddall BG. Wallace, Darwin, and the theory of natural selection. *J Hist Biol*. 1968;1(2):261–323.

3. Muir Wood R. *The Dark Side of the Earth*. Illus. by M Woodhouse London: Allen & Unwin, 1985.

4. Hazen RM. *The Story of Earth*. New York: Viking, 2012.

5. Holland HD. *The Chemical Evolution of the Atmosphere and Oceans*. Princeton, NJ: Princeton University Press; 1984.

6. Canfield DE. *Oxygen: A Four Billion Year History*. Princeton, NJ: Princeton University Press, 2014.

7. Intergovernmental Panel on Climate Change. *Fifth Assessment Report (4 volumes)*. New York: Cambridge University Press, 2013.

8. Weart SR. *The Discovery of Global Warming*. Cambridge, MA: Harvard University Press, 2008.

9. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Mag*. 1996;17(3):37.

10. Bolukbasi B, Berente N, Cutcher-Gershenfeld J, Dechurch L, Flint C, Haberman M, et al. Open data: crediting a culture of cooperation. *Science*. 2013;342(6162):1041–1042.

11. Hazen RM, Bekker A, Bish DL, Bleeker W, Downs RT, Farquhar J, et al. Needs and opportunities in mineral evolution research. *Am Mineral*. 2011;96(7):953–963.

12. Keller CB, Schoene B. Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gyr ago. *Nature*. 2012;485(7399):490.

13. EarthCube. *A Community Roadmap for EarthCube Data: Discovery, Access, and Mining*. Washington, DC: National Science Foundation, 2012.

14. Spencer CJ, Hawkesworth C, Cawood PA, Dhuime B. Not all supercontinents are created equal: Gondwana–Rodinia case study. *Geology*. 2013;41(7):795–798.

15. Nance RD, Murphy JB, Santosh M. The supercontinent cycle: a retrospective essay. *Gondwana Res*. 2014;25:4–29.

16. Liu C, Knoll AH, Hazen RM. Geochemical and mineralogical evidence that Rodinian assembly was unique. *Nat Commun*. 2017;8(1):1950.

17. Bogdanova SV, Pisarevsky SA, Li ZX. Assembly and breakup of Rodinia (some results of IGCP project 440). *Stratigr Geol Correl*. 200917(3):259–274.

18. Huston DL, Pehrsson S, Eglington BM, Zaw K. The geology and metallogeny of volcanic-hosted massive sulfide deposits: variations through geologic time and with tectonic setting. *Econ Geol*. 2010;105(3):571–591.

19. Perssohn S, Eglington B, Huston D, Evans D. Why was Rodinia underendowed? Comparing the effects of paleogeography versus lithosphere thickness on secular ore

deposit preservation. *Mineralogical Magazine*, 2003. Available from: https://goldschmidtabstracts.info/abstracts/abstractView?id=2013006335.

20. Cawood PA, Hawkesworth CJ. Earth's middle age. *Geology*. 2014;42(6):503–506.

21. Bierlein FP, Groves DI, Cawood PA. Metallogeny of accretionary orogens – the connection between lithospheric processes and metal endowment. *Ore Geol Rev*. 2009;36(4):282–292.

22. Liu C, Runyon SE, Knoll AH, Hazen RT. The same and not the same: ore geology, mineralogy and geochemistry of Rodinia assembly. *Earth Sci Rev*., 2019;196:102860, DOI: 10.1016/j.earscirev.2019.05.004.

23. Kemp AIS, Hawkesworth CJ, Paterson BA, Kinny PD, Kemp T. Episodic growth of the Gondwana supercontinent from hafnium and oxygen isotopes in zircon. *Nature*. 2006;439(7076):580–583.

24. Campbell IH, Allen CM. Formation of supercontinents linked to increases in atmospheric oxygen. *Nat Geosci*. 2008;1(8):554–558.

25. Rino S, Kon Y, Sato W, Maruyama S, Santosh M, Zhao D. The Grenvillian and Pan-African orogens: world's largest orogenies through geologic time, and their implications on the origin of superplume. *Gondwana Res*. 2008;14(1–2):51–72.

26. Condie KC, Belousova E, Griffin WL, Sircombe KN. Granitoid events in space and time: constraints from igneous and detrital zircon age spectra. *Gondwana Res*. 2009;15:228–242.

27. Condie KC, Aster RC. Episodic zircon age spectra of orogenic granitoids: the supercontinent connection and continental growth. *Precambrian Res*. 2010;180 (3–4):227–236.

28. Hawkesworth CJ, Dhuime B, Pietranik AB, Cawood PA, Kemp AIS, Storey CD. The generation and evolution of the continental crust. *J Geol Soc London*. 2010;167 (2):229–248.

29. Voice PJ, Kowalewski M, Eriksson KA. Quantifying the timing and rate of crustal evolution: global compilation of radiometrically dated detrital zircon grains. *J Geol*. 2011;119(2):109–126.

30. Hazen RM, Golden J, Downs RT, Hystad G, Grew ES, Azzolini D, et al. Mercury (Hg) mineral evolution: a mineralogical record of supercontinent assembly, changing ocean geochemistry, and the emerging terrestrial biosphere. *Am Mineral*. 2012;97 (7):1013–1042.

31. Hazen RM, Liu X-M, Downs RT, Golden J, Pires AJ, Grew ES, et al. Mineral evolution: episodic metallogenesis, the supercontinent cycle, and the coevolving geosphere and biosphere. *Soc Econ Geol Spec Publ*. 2014;18(18):1–15.

32. Golden JJ, McMillan M, Downs RT, Hystad G, Goldstein I, Stein HJ, et al. Rhenium variations in molybdenite (MoS_2): evidence for progressive subsurface oxidation. *Earth Planet Sci Lett*. 2013;366:1–5.

33. Grew ES, Hazen RM. Beryllium mineral evolution. *Am Mineral*. 2014;99 (5–6):999–1021.

34. Grew ES, Krivovichev SV, Hazen RM, Hystad G. Evolution of structural complexity in boron minerals. *Can Mineral*. 2016;54(1):125–143.

35. Hummer D, Golden J, Hystad G, Downs R, Eleish A, Liu C, et al. The oxidation of Earth's crust: evidence from the evolution of manganese minerals. *Nat Geosci*., in revision.

36. Hoffman PF, Kaufman AJ, Halverson GP, Schrag DP. A neoproterozoic Snowball Earth. *Science*. 1998;281(5381):1342–1346.

37. Goldfarb RJ, Bradley D, Leach DL. Secular variation in economic geology. *Econ Geol*. 2010;105(3):459–465.

38. Bradley DC. Secular trends in the geologic record and the supercontinent cycle. *Earth Sci Rev*. 2011;108(1–2):16–33.

39. Van Kranendonk MJ, Kirkland CL. Orogenic climax of Earth: the 1.2–1.1 Ga Grenvillian superevent. *Geology*. 2013;41(7):735–738.

40. Dickinson WR. Impact of differential zircon fertility of granitoid basement rocks in North America on age populations of detrital zircons and implications for granite petrogenesis. *Earth Planet Sci Lett*. 2008;275(1–2):80–92.

41. Nicholson SW, Cannon WF, Schulz KJ. Metallogeny of the midcontinent rift system of North America. *Precambrian Res*. 1992;58(1–4):355–386.

42. Upton BGJ, Emeleus CH, Heaman LM, Goodenough KM, Finch AA. Magmatism of the mid-Proterozoic Gardar Province, South Greenland: chronology, petrogenesis and geological setting. *Lithos*. 2003;68(1–2):43–65.

43. Iriondo A, Premo WR, Martínez-Torres LM, Budahn JR, Atkinson WW, Siems DF, et al. Isotopic, geochemical, and temporal characterization of Proterozoic basement rocks in the Quitovac region, northwestern Sonora, Mexico: implications for the reconstruction of the southwestern margin of Laurentia. *Bull Geol Soc Am*. 2004;116 (1–2):154–170.

44. Greentree MR, Li ZX, Li XH, Wu H. Late Mesoproterozoic to earliest Neoproterozoic basin record of the Sibao orogenesis in western South China and relationship to the assembly of Rodinia. *Precambrian Res*. 2006;151(1–2):79–100.

45. McLelland J, Selleck B, Bickford M. Review of the Proterozoic evolution of the Grenville Province, its Adirondack outlier, and the Mesoproterozoic inliers of the Appalachians. In: Tollo R, Bartholomew M, Hibbard J, Karabinos P, eds. *From Rodinia to Pangea: The Lithotectonic Record of the Appalachian Region: Geological Society of America Memoir*. Boulder, CO: Geological Society of America, 2010, pp. 21–49.

46. Prol-Ledesma RM, Melgarejo JC, Martin RF. The El Muerto "NYF" granitic pegmatite, Oaxaca, Mexico, and its striking enrichment in allanite-(CE) and monazite-(Ce). *Can Mineral*. 2012;50(4):1055–1076.

47. Baadsgaard H, Chaplin C, Griffin WL. Geochronology of the Gloserheia pegmatite, Froland, southern Norway. *Nor Geol Tidsskr*. 1984;64(2):111–119.

48. Foord EE, Černý P, Jackson LL, Sherman DM, Eby RK. Mineralogical and geochemical evolution of micas from miarolitic pegmatites of the anorogenic pikes peak batholith, Colorado. *Mineral Petrol*. 1995;55(1–3):1–26.

49. McCauley A, Bradley DC. The global age distribution of granitic pegmatites. *Can Mineral*. 2014;52(2):183–190.

50. Dhuime B, Wuestefeld A, Hawkesworth CJ. Emergence of modern continental crust about 3 billion years ago. *Nat Geosci*. 2015;8(7):552–555.

51. Dijkstra AH, Dale CW, Oberthür T, Nowell GM, Graham Pearson D. Osmium isotope compositions of detrital Os-rich alloys from the Rhine River provide evidence for a global late Mesoproterozoic mantle depletion event. *Earth Planet Sci Lett*. 2016;452:115–122.

52. Kelemen PB, Johnson KTM, Kinzler RJ, Irving AJ. High-field-strength element depletions in arc basalts due to mantle–magma interaction. *Nature*. 1990;345 (6275):521–524.

53. Woodhead J, Eggins S, Gamble J. High field strength and transition element systematics in island arc and back-arc basin basalts: evidence for multi-phase melt extraction and a depleted mantle wedge. *Earth Planet Sci Lett*. 1993;114 (4):491–504.

54. Cawood PA, Strachan RA, Pisarevsky SA, Gladkochub DP, Murphy JB. Linking collisional and accretionary orogens during Rodinia assembly and breakup: implications for models of supercontinent cycles. *Earth Planet Sci Lett*. 2016;449:118–126.

55. Evans DAD, Li ZX, Murphy JB. Four-dimensional context of Earth's supercontinents. *Geol Soc London Spec Publ*. 2016;424(1):1–14.

56. Cawood PA, Pisarevsky SA. Laurentia–Baltica–Amazonia relations during Rodinia assembly. *Precambrian Res*. 2017;292:386–397.

57. Hoffman PF, Grotzinger JP. Orographic precipitation, erosional unloading, and tectonic style. *Geology*. 1993;21:195–198.

58. Bingen B, Stein HJ, Bogaerts M, Bolle O, Mansfeld J. Molybdenite Re–Os dating constrains gravitational collapse of the Sveconorwegian orogen, SW Scandinavia. *Lithos*. 2006;87(3–4):328–346.

59. Rivers T. Assembly and preservation of lower, mid, and upper orogenic crust in the Grenville Province – implications for the evolution of large hot long-duration orogens. *Precambrian Res*. 2008;167(3–4):237–259.

60. Möller C, Andersson J, Dyck B, Antal Lundin I. Exhumation of an eclogite terrane as a hot migmatitic nappe, Sveconorwegian orogen. *Lithos*. 2015;226:147–168.

61. Rainbird R, Cawood P, Gehrels G. The great Grenvillian sedimentation episode: record of supercontinent Rodinia's assembly. In: Cathy B, Antonio A, eds. *Tectonics of Sedimentary Basins: Recent Advances*. Chichester: John Wiley & Sons, Ltd., 2012, pp. 583–601.

62. Cawood PA, Hawkesworth CJ, Dhuime B. The continental record and the generation of continental crust. *Bull Geol Soc Am*. 2013;125(1–2):14–32.

63. Spencer CJ, Prave AR, Cawood PA, Roberts NMW. Detrital zircon geochronology of the Grenville/Llano foreland and basal Sauk Sequence in west Texas, USA. *Bull Geol Soc Am*. 2014;126(7–8):1117–1128.

64. Brune S, Williams SE, Müller RD. Potential links between continental rifting, CO_2 degassing and climate change through time. *Nat Geosci*. 2017;10(12):941–946.

65. Lee CTA, Caves J, Jiang H, Cao W, Lenardic A, McKenzie NR, et al. Deep mantle roots and continental emergence: implications for whole-Earth elemental cycling, long-term climate, and the Cambrian explosion. *Int Geol Rev*. 2018;60(4):431–448.

66. Aulbach S, Creaser RA, Stachel T, Heaman LM, Chinn IL, Kong J. Diamond ages from Victor (Superior Craton): intra-mantle cycling of volatiles (C, N, S) during supercontinent reorganisation. *Earth Planet Sci Lett*. 2018;490:77–87.

67. von Blanckenburg F. The control mechanisms of erosion and weathering at basin scale from cosmogenic nuclides in river sediment. *Earth Planet Sci Lett [Internet]*. 2006;237(3–4):462–479.

68. Silver PG, Behn MD. Intermittent plate tectonics? *Science*. 2008;319(5859):85–88.

69. Berner RA. *The Phanerozoic Carbon Cycle: CO_2 and O_2*. Oxford: Oxford University Press, 2004.

70. Knoll AH, Follows MJ. A bottom-up perspective on ecosystem change in Mesozoic oceans. *Proc R Soc B Biol Sci*. 2016;283(1841):20161755.

71. Knoll AH. Biological and biogeochemical preludes to the Ediacaran radiation. In: Lipps JH, Signor PW, eds. *Origin and Early Evolution of the Metazoa*. Boston, MA: Springer, 1992, pp. 53–84.

72. Hayes JM, Waldbauer JR. The carbon cycle and associated redox processes through time. *Philos Trans R Soc B Biol Sci*. 2006;361:931–950.

73. Lyons TW, Reinhard CT, Planavsky NJ. The rise of oxygen in Earth's early ocean and atmosphere. *Nature*. 2014;506:307–315.

74. Knoll AH. Paleobiological perspectives on early eukaryotic evolution. *Cold Spring Harb Perspect Biol*. 2014;6(1):a016121.
75. Knoll AH. Paleobiological perspectives on early microbial evolution. *Cold Spring Harb Perspect Biol*. 2015;7(7):a018093.
76. Kah LC, Sherman AG, Narbonne GM, Knoll AH, Kaufman AJ. $\delta^{13}C$ stratigraphy of the Proterozoic Bylot Supergroup, Baffin Island, Canada: implications for regional lithostratigraphic correlations. *Can J Earth Sci*. 1999;36(3):313–332.
77. Halverson GP, Hoffman PF, Schrag DP, Maloof AC, Rice AHN. Toward a Neoproterozoic composite carbon-isotope record. *Bull Geol Soc Am*. 2005;117:1181–1207.
78. Reinhard CT, Planavsky NJ, Gill BC, Ozaki K, Robbins LJ, Lyons TW, et al. Evolution of the global phosphorus cycle. *Nature*. 2017;541(7637):386–389.
79. Laakso TA, Schrag DP. Regulation of atmospheric oxygen during the Proterozoic. *Earth Planet Sci Lett*. 2014;388:81–91.
80. Hazen RM, Downs RT, Kah L, Sverjensky D. Carbon mineral evolution. *Rev Mineral Geochem*. 2013;75(1):79–107.
81. Hazen RM, Hummer DR, Hystad G, Downs RT, Golden JJ. Carbon mineral ecology: predicting the undiscovered minerals of carbon. *Am Mineral*. 2016;101(4):889–906.
82. Morrison SM, Liu C, Eleish A, Prabhu A, Li C, Ralph J, et al. Network analyses of mineralogical systems. *Am Mineral*. 2017;102(1):1588–1596.
83. Hazen RM, Papineau D, Bleeker W, Downs RT, Ferry JM, Mccoy TJ, et al. Mineral evolution. *Am Mineral*. 2008;93(11–12):1693–1720.
84. Hazen RM, Ferry JM. Mineral evolution: mineralogy in the fourth dimension. *Elements*. 2010;6(1):9–12.
85. Zhabin AG. Is there evolution of mineral speciation on Earth? *Dokl Earth Sci Sect*. 1981;247:142–144.
86. Zalasiewicz J, Kryza R, Williams M. The mineral signature of the Anthropocene in its deep-time context. *Geol Soc London Spec Publ*. 2014;395(1):109–117.
87. Hazen RM, Grew ES, Origlieri MJ, Downs RT. On the mineralogy of the "Anthropocene Epoch." *Am Mineral*. 2017;102(3):595–611.
88. Golden JJ, Pires AJ, Hazen RM, Downs RT, Ralph J, Meyer M. Building the mineral evolution database: implications for future big data analysis. *Geol Soc Am Abstr Prog*. 2016;48:7.
89. Ernst RE. *Large Igneous Provinces*. Cambridge: Cambridge University Press, 2014.
90. Hystad G, Downs RT, Hazen RM. Mineral species frequency distribution conforms to a large number of rare events model: prediction of Earth's missing minerals. *Math Geosci*. 2015;47(6):647–661.
91. Hystad G, Downs RT, Grew ES, Hazen RM. Statistical analysis of mineral diversity and distribution: Earth's mineralogy is unique. *Earth Planet Sci Lett*. 2015;426:154–157.
92. Hazen RM, Grew ES, Downs RT, Golden J, Hystad G. Mineral ecology: chance and necessity in the mineral diversity of terrestrial planets. *Can Mineral*. 2015;53(2):295–324.
93. Hazen RM, Hystad G, Downs RT, Golden JJ, Pires AJ, Grew ES. Earth's "missing" minerals. *Am Mineral*. 2015;100(10):2344–2347.
94. Maynard-Casely HE, Cable ML, Malaska MJ, Vu TH, Choukroun M, Hodyss R. Prospects for mineralogy on Titan. *Am Mineral*. 2018;103(3):343–349.
95. Hazen RM. Titan mineralogy: a window on organic mineral evolution. *Am Mineral*. 2018;103(3):341–342.
96. Abraham A, Hassanien A-E, Snasel V. *Computational Social Network Analysis: Trends, Tools and Research Advances (Computer Communications and Networks)*. New York: Springer, 2010.

97. Scott J, Carrington PJ, eds. *The SAGE Handbook of Social Network Analysis*. Thousand Oaks, CA: SAGE Publications Ltd., 2011.

98. Newman M. *Networks: An Introduction*. Oxford: Oxford University Press, 2010.

99. Colwell FS, D'Hondt S. Nature and extent of the deep biosphere. *Rev Mineral Geochem*. 2013;75(1):547–574.

100. Tashiro T, Ishida A, Hori M, Igisu M, Koike M, Méjean P, et al. Early trace of life from 3.95 Ga sedimentary rocks in Labrador, Canada. *Nature*. 2017;549 (7673):516–518.

101. Dodd MS, Papineau D, Grenne T, Slack JF, Rittner M, Pirajno F, et al. Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature*. 2017;543 (7643):60–64.

102. Stüeken EE, Buick R, Anderson RE, Baross JA, Planavsky NJ, Lyons TW. Environmental niches and metabolic diversity in Neoarchean lakes. *Geobiology*. 2017;15 (6):767–783.

103. Giovannelli D, Sievert SM, Hügler M, Markert S, Becher D, Schweder T, et al. Insight into the evolution of microbial metabolism from the deep-branching bacterium, *Thermovibrio ammonificans*. *Elife*. 2017;6:e18990.

104. Jelen BI, Giovannelli D, Falkowski PG. The role of microbial electron transfer in the coevolution of the biosphere and geosphere. *Annu Rev Microbiol*. 2016;70(1):45–62.

105. Moore EK, Jelen BI, Giovannelli D, Raanan H, Falkowski PG. Metal availability and the expanding network of microbial metabolisms in the Archaean eon. *Nat Geosci*. 2017;10:629–636.

106. Greene LH. Protein structure networks. *Brief Funct Genomics*. 2012;11(6):469–478.

107. Zhu C, Delmont TO, Vogel TM, Bromberg Y. Functional basis of microorganism classification. *PLoS Comput Biol*. 2015;11(8):e1004472.

108. Zhu C, Mahlich Y, Miller M, Bromberg Y. Fusion DB: assessing microbial diversity and environmental preferences via functional similarity networks. *Nucleic Acids Res*. 2018;46(D1):D535–D541.

109. Harel A, Falkowski P, Bromberg Y. TrAnsFuSE refines the search for protein function: oxidoreductases. *Integr Biol*. 2012;4(7):765–777.

110. Harel A, Bromberg Y, Falkowski PG, Bhattacharya D. Evolutionary history of redox metal-binding domains across the tree of life. *Proc Natl Acad Sci*. 2014;111 (19):7042–7047.

111. Senn S, Nanda V, Falkowski P, Bromberg Y. Function-based assessment of structural similarity measurements using metal co-factor orientation. *Proteins*. 2014;82 (4):648–656.

112. Anbar AD, Knoll AH. Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *Science*. 2002;297:1137–1142.

113. Dupont S, Lundve B, Thorndyke M. Seawater carbonate chemistry and biological processes during experiments with a sea star *Crassaster papposus*. Supplement to: Dupont S et al. (2010): Near future ocean acidification increases growth rate of the lecithotrophic larvae and juveniles of the sea star *Crossaster papposus*. *Journal of Experimental Zoology Part B – Molecular and Developmental Evolution*. *Pangaea*. 2010;314B:382–389.

114. Zhu C, Miller M, Marpaka S, Vaysberg P, Rühlemann MC, Wu G, et al. Functional sequencing read annotation for high precision microbiome analysis. *Nucleic Acids Res*. 2018;46(4):e23.

115. Barberán A, Bates ST, Casamayor EO, Fierer N. Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J*. 2012;6(2):343–351.

116. Tringe SG, Von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, et al. Comparative metagenomics of microbial communities. *Science*. 2005;308 (5721):554–557.

117. Faust K, Raes J. Microbial interactions: from networks to models. *Nat Rev Microbiol*. 2012;10:538–550.

118. Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-González A, Eldridge DJ, Bardgett RD, et al. A global atlas of the dominant bacteria found in soil. *Science*. 2018;359(6373):320–325.

119. Weiss SM, Kapouleas I. An empirical comparison of pattern recognition, neural nets and machine learning classification methods. Shavlik JW, Dietterich TG, eds. *Readings in Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1990, pp. 177–183.

120. Hand DJ. Data mining: statistics and more? *Am Stat*. 1998;52(2):112–118.

121. Proulx SR, Promislow DEL, Phillips PC. Network thinking in ecology and evolution. *Trends Ecol Evol*. 2005;20:345–353.

122. Lozupone CA, Knight R. Global patterns in bacterial diversity. *Proc Natl Acad Sci*. 2007;104(27):11436–11440.

123. Solden L, Lloyd K, Wrighton K. The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Curr Opin Microbiol*. 2016;31:217–226.

124. Giovannelli D, D'Errico G, Fiorentino F, Fattorini D, Regoli F, Angeletti L, et al. Diversity and distribution of prokaryotes within a shallow-water pockmark field. *Front Microbiol*. 2016;7:941.

125. Fierer N, Jackson RB. The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci*. 2006;103(3):626–631.

126. Lauber CL, Hamady M, Knight R, Fierer N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol*. 2009;75(15):5111–5120.

127. Auguet JC, Barberan A, Casamayor EO. Global ecological patterns in uncultured Archaea. *ISME J*. 2010;4(2):182–190.

128. Barberán A, Casamayor EO. Global phylogenetic community structure and β-diversity patterns in surface bacterioplankton metacommunities. *Aquat Microb Ecol*. 2010;59(1):1–10.

129. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci*. 2011;108(Suppl. 1):4516–4522.

130. Hoppe B, Kahl T, Karasch P, Wubet T, Bauhus J, Buscot F, et al. Network analysis reveals ecological links between N-fixing bacteria and wood-decaying fungi. *PLoS One*. 2014;9(2):e88141.

131. Ruff SE, Biddle JF, Teske AP, Knittel K, Boetius A, Ramette A. Global dispersion and local diversification of the methane seep microbiome. *Proc Natl Acad Sci*. 2015;112(13):4015–4020.

132. Gilbert J. Metagenomics, metadata, and meta-analysis. In: Karen EN, ed. *Encyclopedia of Metagenomics*. Boston, MA: Springer US, 2015, pp. 439–442.

133. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, et al. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol*. 2008;26:541–547.

134. Pasolli E, Schiffer L, Manghi P, Renson A, Obenchain V, Truong DT, et al. Accessible, curated metagenomic data through ExperimentHub. *Nat Methods*. 2017;14:1023–1024.

135. Reed DC, Algar CK, Huber JA, Dick GJ. Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proc Natl Acad Sci.* 2014;111(5):1879–1884.
136. Schimel J. Microbial ecology: linking omics to biogeochemistry. *Nat Microbiol.* 2016;1:15028.
137. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech.* 2008;P10008.