Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Deep Matter & Energy—Review

数据驱动型研究方法在矿物学领域里的新发现——矿物数据资源、数据分析和可视化的最新研究进展

Robert M. Hazen [a,*], Robert T. Downs [b], Ahmed Eleish [c], Peter Fox [c], Olivier C. Gagné [a], Joshua J. Golden [b], Edward S. Grew [d], Daniel R. Hummer [e], Grethe Hystad [f], Sergey V. Krivovichev [g], Congrui Li [c], Chao Liu [a], Xiaogang Ma [h], Shaunna M. Morrison [a], Feifei Pan [c], Alexander J. Pires [b], Anirudh Prabhu [c], Jolyon Ralph [i], Simone E. Runyon [a,j], Hao Zhong [c]

[a] *Geophysical Laboratory, Carnegie Institution for Science, Washington, DC 20015, USA*
[b] *Department of Geosciences, The University of Arizona, Tucson, AZ 85721-0077, USA*
[c] *Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*
[d] *School of Earth and Climate Sciences, University of Maine, Orono, ME 04469, USA*
[e] *Department of Geology, Southern Illinois University, Carbondale, IL 62901, USA*
[f] *Mathematics, Statistics, and Computer Science, Purdue University Northwest, Hammond, IN 46323-2094, USA*
[g] *Kola Science Centre of the Russian Academy of Sciences, Apatity, Murmansk Region 184209, Russia*
[h] *Department of Computer Science, University of Idaho, Moscow, ID 83844-1010, USA*
[i] *Mindat.org, Mitcham CR4 4FD, UK*
[j] *Department of Geology and Geophysics, University of Wyoming, Laramie, WY 82071-2000, USA*

ARTICLE INFO

Article history:
Received 15 November 2018
Revised 18 February 2019
Accepted 13 March 2019
Available online 2 May 2019

关键词
矿物演化
矿物生态学
Skyline 图
网络分析
聚类分析
Chord 图
Klee 图

摘要

随着矿物种类多样性、矿物（时空）分布特征和矿物性质等领域海量数据的快速增长，矿物学迎来了数据驱动型研究发现的新纪元。当前，最全面的国际性矿物数据库是 IMA 数据库和 mindat.org 数据库，其中，IMA 数据库包含了超过 5300 种被国际矿物学协会（International Mineralogical Association，IMA）批准认可的矿物及其属性信息。此外，mindat.org 数据库包含了超过 100 万种矿物种类及其产地信息，这些矿物来自于世界各地，有登记在册的产地来源就超过了 30 万个。采用各种现代化分析方法对这些海量地学数据进行分析解读和可视化处理，进一步增进了对地球圈和生物圈协同演化过程的理解认识，这些分析方法包括 chord 图、cluster 图、Klee 图、skyline 图，以及各式各样的网络分析方法。新型数据驱动型分析策略包括矿物演化分析、矿物生态学分析和矿物网络分析，这些分析策略能够系统性地综合考虑矿物的时空分布特征及其多样性。这些分析策略正在增进对矿物共生现象的深入认识，并且首次推动了对“地球上存在但尚未被发现和记录在册矿物”的预测。

1. 引言

一直以来，地球矿物资源的发现、记录和开发都是地球科学的核心追求。在很长一段时期内，新矿藏和矿物种类的发现仅依赖于偶然发现和经验性指导。古谚“Gold is where you find it”（黄金就在你找到它的地方）

* Corresponding author.
E-mail address: rhazen@ciw.edu (R.M. Hazen).

英文原文: Engineering 2019, 5(3): 397–405
引用本文: Robert M. Hazen, Robert T. Downs, Ahmed Eleish, Peter Fox, Olivier C. Gagné, Joshua J. Golden, Edward S. Grew, Daniel R. Hummer, Grethe Hystad, Sergey V. Krivovichev, Congrui Li, Chao Liu, Xiaogang Ma, Shaunna M. Morrison, Feifei Pan, Alexander J. Pires, Anirudh Prabhu, Jolyon Ralph, Simone E. Runyon, Hao Zhong. Data-Driven Discovery in Mineralogy: Recent Advances in Data Resources, Analysis, and Visualization. *Engineering*, https://doi.org/10.1016/j.eng.2019.03.006

阐述了绝大多数自然矿产资源的发现过程，然而当前数据驱动型矿物发现这一新策略正在逐渐颠覆这句古谚。本文中，我们回顾了不断增长的海量矿物数据资源库的本征属性，并介绍了一些适用于开展矿物时空分布特征及其多样性研究的数据分析方法与可视化技术。

最近的研究可以总结归纳为三大类。① 矿物演化。矿物演化主要研究在“超过”45亿年的地质时期近地表矿物的演化过程，这方面研究主要揭示了地球圈和生物圈的协同演化过程，以及因地球的化学分异而与日俱增的矿物种类多样性和复杂性[1–27]。② 矿物生态学。矿物生态学作为一种补充诉求，主要研究地球矿物的种类多样性和空间分布特征，包括研究地表层稀有矿物的独特分布[28–39]。③ 矿物网络分析。矿物网络分析为分析和可视化矿物在时空中的复杂分布及其性质提供了有力的手段[40]。

总而言之，这些方法将有可能改变我们对地球和其他类地星球矿物演化规律的认识。

2. 矿物数据资源库

数据驱动型新发现依赖于全面可靠的矿物数据库，包括矿物种类、性质及其时空分布特征。IMA数据库[†]是国际矿物学协会（International Mineralogical Association，IMA）官方认可的矿物数据库，包含了所有经过国际矿物学协会批准认可的矿物类型，该数据库由亚利桑那大学的地球科学学院进行更新维护[41]。RRUFF数据库除了包含超过5400种矿物种类信息外，还包含了晶体结构、化学成分、拉曼谱以及其他物理性质等重要信息。此外，开展矿物演化研究所需要的矿物年龄、矿物产地和矿物背景信息也都可以在RRUFF数据库的子库——矿物演化数据库（Mineral Evolution Database）[‡]中找到，在这个快速增长的开放数据库里可以获取超过18.5万种不同矿物的产地/年龄信息。

最大的全球矿物分布信息数据库是mindat.org[††]，这是一项由Jolyon Ralph和哈德森矿物研究所（Hudson Institute of Mineralogy）领导的国际性、开放性、大众化研究项目。Mindat数据库已经记录了超过110万条矿物种类/产地信息，这些登记在册的矿物来自全球各地约30万个不同地点。这些数据对于矿物分布和种类多样性以及矿物产地间联系等问题的研究分析与可视化处理都至关重要。

IMA数据库和mindat.org数据库的核心数据资源与其他数据库资源相结合将能够进一步凸显其重要性。比如，一个著名例子是由跨学科地球数据联盟（Interdisciplinary Earth Data Alliance, IEDA[‡‡]）组织的岩石学（petrological）和地球化学（geochemical）资源数据库，包括EarthChem[†††]子数据库（如文献[42]）。

开发与完善这些重要数据库过程中遇到的一个持续性挑战是大量的“dark data”（暗数据），换句话说，就是这些涉及矿物化学成分、产地和其他信息的重要数据只能够通过已经发表过的文章复印稿、专有的公司文件（特别是矿产资源企业）或者是个人的研究记录。数据驱动主导的矿物学新发现要实现其最大化效益的一个先决条件是：数据共享文化在地球科学界深入人心并广为接纳，而且研究人员能够将“FAIR方法”（findable, accessible, interoperable, and reusable，即可查找、可访问、可互操作、可重用）践行于科研实践中，并形成一种数据共享与数据重用的文化氛围[43]。

随着开放获取型矿物数据库的不断丰富与完善，综合运用各种数据分析方法和可视化处理的前沿技术开展矿物学分析研究也日趋成熟[44,45]。本文主要回顾了与矿物演化、矿物生态学和矿物网络分析相关的一些数据分析和可视化处理方法。

3. 矿物演化

矿物演化主要研究地球或者是其他类地星球近地表面矿物在漫长地质时期的演化过程[5,19]。我们对过去45亿年漫长时期里地球矿物演化过程的认识逐渐深入，对太阳系中其他类地星球表面矿物演化的认识也日益增长[46, 47]。这些认识表明行星的矿物演化包含了一系列连续的演变阶段，每一阶段都是矿物共生条件下新物理、新化学和（地球）新生物模式共同作用的结果。

地球矿物演化数据库罗列了超过18.5万种矿物的产地/年龄信息，虽然这还远未能够记录下所有可获取的矿物种类/年龄信息，但也足以揭示地球矿物演化规律。

[†] https://rruff.info/ima/.
[‡] https://rruff.info/evolution.
[††] https://www.mindat.org.
[‡‡] https://www.iedadata.org/.
[†††] https://www.EarthChem.org.

其中，最为显而易见的三大演化趋势（规律）总结如下。

矿物在时间分布上的第一个趋势是其具有显著的“阶段性”，这反映了过去30亿年来的超大陆旋回运动[8, 12]。我们发现地球保留了在大陆板块集聚运动过程中爆发式产生新型矿物的证据。这些事件在时间上对应着被认为的5次分离的地块汇聚成一个超大陆的事件（图1，[39]）。大陆的汇聚及其伴随着的造山事件不仅引起了大范围的矿化事件，而且这些矿化事件也极有可能就保留在造山运动形成的山脉核心岩层中。对这些趋势的详细研究分析还提供了更多的额外细节信息，比如大约形成于13亿~9亿年前（1.3~0.9 Ga）的罗迪尼亚超大陆（Rodinia）具有独特的构造和地球化学环境[27]。

第二大重要趋势是：在地质演化时期里观察到了几次过渡金属平均氧化态（化学价）的大幅度增长[20, 48]。比如，在过去的5亿多年里，含锰矿物质的氧化态出现了系统性增加，而在地球早期也有几次波动（图2）。类似趋势也出现在所有对氧化还原反应敏感的第四周期过渡金属（图3†）以及铀[6]和徕[20]等体系中。

第三大矿物演化趋势是：随着地质时间的推移，矿

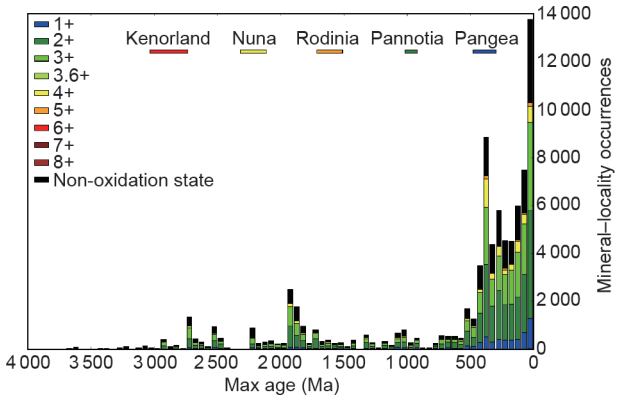


图1. 地质时期里，地球矿物质在五大超大陆（Kenorland, Nuna, Rodinia, Pannotia以及Pangea）的属地分布特性（以含第四周期过渡金属元素平均氧化态的最高价态为统计依据）。我们记录的地球矿物演化特性反映了地质时期里地球矿化事件的周期性与超大陆的形成周期有关。在图中有大约6万组包含第四周期过渡金属元素的地球矿物数据，每组数据都包含有矿物属地和矿物年龄信息。从图中可以清楚地看出脉冲式的矿化事件与五大超大陆（Kenorland, Nuna, Rodinia, Pannotia以及Pangea）的形成相关联。我们注意到，大约形成于1.3~0.9 Ga的Rodinian板块矿化事件则没有其他超大陆的矿化事件那么明显，其主要原因是Rodinian板块的独特板块构造背景[39]。1+~8+表示不同的氧化态。

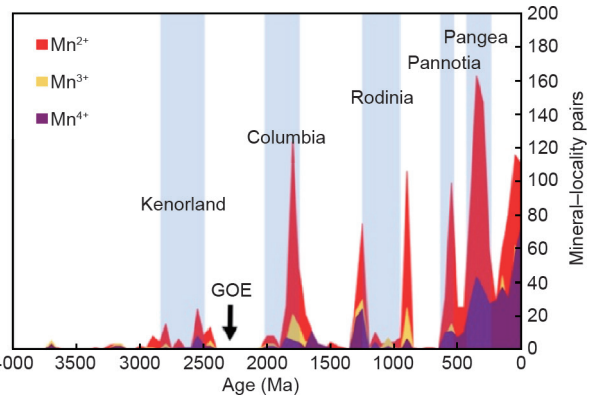


图2. 地质时期里，地球近地表面矿物平均氧化态演化图。其中，锰氧化物的三个不同价态（如Mn^{2+}, Mn^{3+}, Mn^{4+}）矿物比例的变化反映了含氧光合作用的演化结果。锰氧化物平均氧化态在过去5亿年里的增长尤为显著。GOE：大氧化事件。

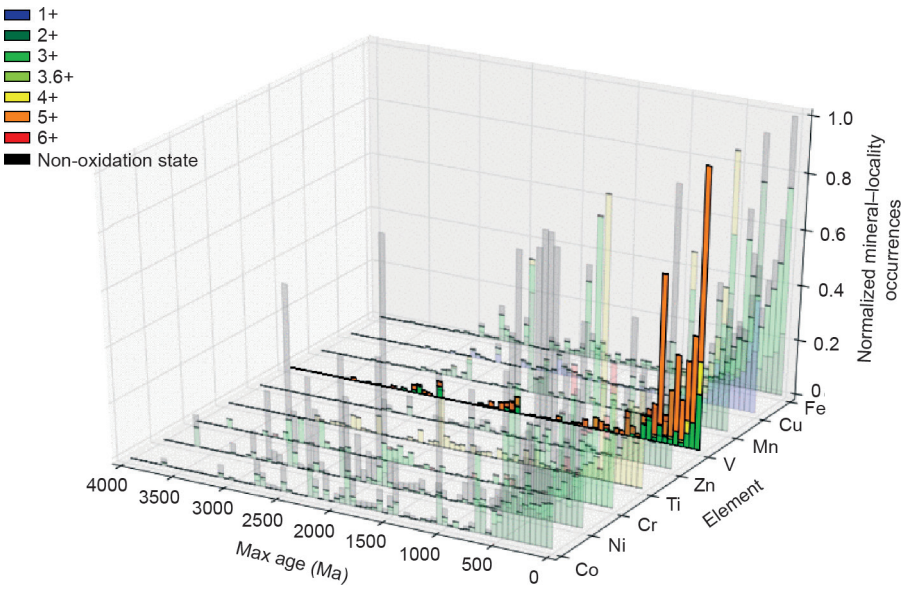


图3. 按照不同地质时期和所含的不同元素类型，进行归一化的矿物产地信息分布图。其中，含第四周期过渡金属元素矿物的“skyline图”反映了与超大陆形成周期和地球大气成分变化相关的系统趋势。

† See https://dtdi.carnegiescience.edu for an animated version.

物结构和化学成分逐渐趋于复杂（图4，[5, 11, 26]）。以数据信息为基础开展的复杂性数值估计研究推动了对数据库中矿物化学成分和结构复杂性间相关性的定量研究，该数据库包含了4962个矿物化学成分数据集和3989个矿物晶体结构数据集[23, 26]。该研究分析给出了一个总体趋势：矿物结构复杂性随其化学成分复杂性的增大而增大。此外，对不同地质时期[5, 15]的矿物群进行平均化学成分和结构复杂性分析时发现：两者都在矿物演化的过程中逐渐增大，也就是说地质时期晚期出现的矿物在化学成分和晶体结构方面都会比地质时期早期出现的矿物更加复杂。与生物进化过程类似[49]，矿物化学成分和结构复杂性的增加也呈现出整体被动跟随模式。随着地质时间的推移，逐渐形成越来越复杂的矿物，然而早期的简单矿物依然存在（[35]）。所观察到的相关性结果表明：在一级近似中，地球的化学分异是促进地球演化过程中矿物复杂性增加的一个主要动力。进一步研究发现：局部特定稀有元素的富集以及新地球化学环境的产生增加了矿物演化过程中的复杂性和种类多样性。

4. 矿物生态学

矿物生态学主要研究矿物空间分布及其多样性，与生态系统主要研究记录生物种群分布特性类似。地球矿物分布遵循大量的罕见事件（large number of rare event，LNRE）分布特征。该分布特征具有普遍性，也同样适用于生态系统中的生物种群分布，以及书籍中的单词分布[29, 31, 37]。在各自情形下，少数目（数量少）的生物种群和单词频繁出现（概率高），然而其余的大量（数量大）生物种群或者是单词则很罕见（概率低）。

对数据库中矿物/产地信息的分析增进了我们对矿物分布特征的详细认识。这些数据构成了“累积曲线”（accumulation curve），并能够据此大致估计“缺失”矿物的数量。“缺失”矿物是指那些在地球上存在但还未被发现与记录在册的矿物[28, 32]。例如，在一项针对400多种含碳矿物的详细研究中，Hazen等[33]预测说还有大约145种含碳矿物有待进一步发现（图5）。更进一步，他们列举了几百种潜在的“缺失”矿物类型，并指出大部分都以含水碳酸盐的形式存在，同时还特别强调可能被忽略了的含钙和钠的化学相，因为它们相对不显眼，通常为白色或灰色，并且结晶性不好[32]。这项工作[33]进一步激发了由Deep Carbon Observatory[†]组织所支持的一项国际性研究项目——Carbon Mineral Challenge[‡]，该项目旨在尽可能多地发现缺失的含碳矿物。截至2018年7月5日，至少有13种新型含碳矿物质进一步被发现与记录在册，并得到了IMA组织的认可。

5. 矿物共存和网络分析

矿物学研究中最重要的挑战之一是从矿物数据库的数百种矿物种类间寻找矿物共存组合信息，从而进一步理解矿物多样性和时空分布特征。与日俱增的数据库

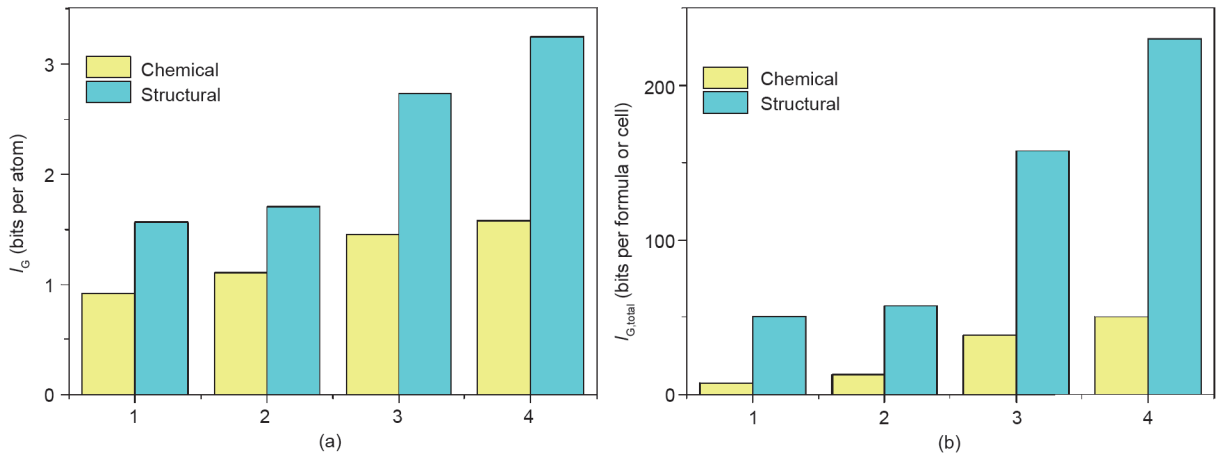


图4. 不同地质时期，矿物平均化学成分和结构复杂度信息演化图。1为12种“ur型矿物”[5]；2为60种球粒陨石(chondritic meteorites)矿物[5]；3为420种冥古宙时代(Hadean epoch)矿物[11]；4为后冥古宙时代(post-Hadean era)的所有矿物质，由数据库中的4962种含不同化学成分的矿物和3989种不同晶体结构的矿物计算而得[26]。（a）每个原子的Shannon信息（I_G）；（b）每个结构单元或者化学式单元的Shannon信息（$I_{G,total}$）。

[†] https://deepcarbon.net/.
[‡] https://mineralchallenge.net/.

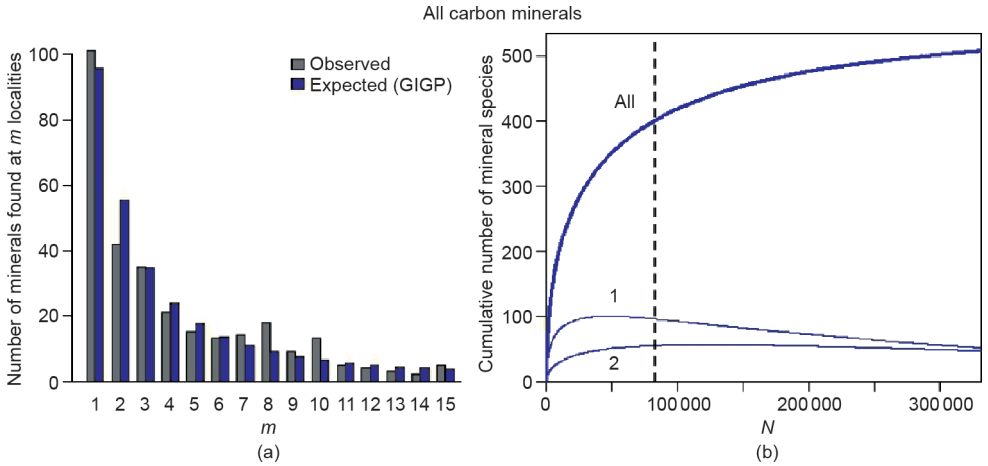


图5. 矿物产地分布图。(a) 含碳矿物的频谱分布图表明大部分矿物还是非常稀少。横坐标记录了发现含碳矿物的确切属地数量（*m*），纵坐标则记录了这些分布地点的矿物种类数量；其中，灰色柱状图是实验观测值，而蓝色柱状图则是模型值。在2016年发现的403种登记在册的含碳矿物，有超过100种是来自一个相同属地，而另外40种则有报道称来自两个不同属地。(b)“大量的罕见事件分布律”有助于“累积曲线”的加速计算（上面的蓝色曲线）。此处显示的是观察到的矿物/属地数据数量（*N*，*X*轴）与不同矿物种类的估计数量（*Y*轴）的关系图。将该曲线外推到右侧表明还有145种含碳矿物有待进一步被发现与描述[33]。垂直虚线则表示截至2016年发现的矿物/属地数量（82 922种）和已知的含碳矿物种类数量（403种）。曲线1和2则表示在一个或者两个不同属地出现的不同矿物种类演化信息。我们注意到，这两条曲线（曲线1和2）都会有一个极大值；尽管登记在册的矿物/属地数据数量越来越多，然而现在只有一个属地的已知矿物种类正在下降。

（mindat.org）资源，以及日新月异的数据分析与可视化处理方法的交汇融合给我们研究分析这类复杂多维系统带来前所未有的颠覆性能力。

5.1. Chord 图

矿物共存分析的第一步是构建一个数据对象，其中，每个矿物种类作为一个单独的作用域。在成对矿物共生矩阵的简单情形下，每个矩阵元素代表两种矿物质共同出现的次数。这些数据可以采用各种不同技术来进行展示与呈现。其中，chord图的构成方法是：先将一组相关的矿物种类依次排列成圆上的一组圆弧，再用曲线将共存矿物所代表的圆弧相连接（图6）。Chord图在基因研究中已经得到了广泛应用。更重要的是，chord图也是矿物学研究中的一种非常实用的分析方法，在chord图中能够通过单个视觉图表来说明大量成对出现的矿物种类。我们将矿物出现次数、矿物出现地点和其他共存物种信息等数据作为嵌入式元数据加入到该数据对象中，并在交互式显示中研究chord图。

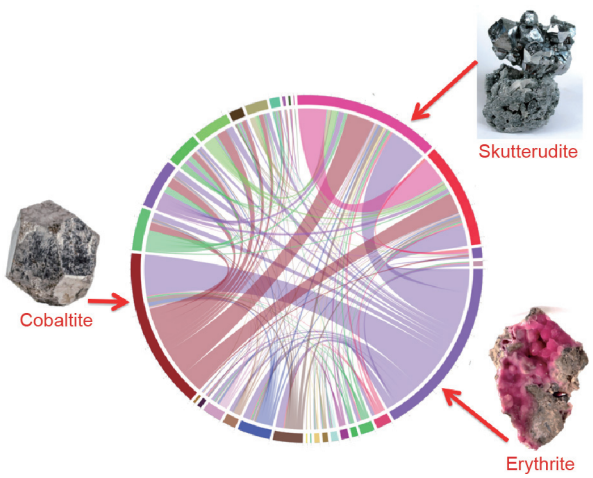


图6. 43种最常见的含钴矿物Chord图描述了共存共生的矿物对。该描述表明了次生矿物赤藓矿物[erythrite, $Co_3(AsO_4)_2{\cdot}8H_2O$]是最丰富的钴矿物，并且它经常与钴酸盐（cobaltite, CoAsS）和方钴矿（skutterudite, $CoAs_{3-x}$）这两种最常见的主要钴矿石矿物共存共生。

5.2. Klee 图

Klee图（有时候又称作“heat maps”，图7）表示物体对（比如矿物质或基本化学元素）共存的概率，因此Klee图也被认为是图6所示chord图的互补可视化工具。Klee图有助于快速分析理解共存矿物或者元素对，但是实际应用中往往需要同时理解两个以上对象间的关联性。因此，Ma等[50]采用交互式三维Klee图研究矿物间的共存元素对（图8）。尽管Klee图能够快速揭示数千种矿物对的共存趋势，但是迄今为止仍未在矿物共存关系研究领域得以普及。

5.3. 网络分析

网络分析是矿物学研究里的一个特别实用的工具，主要用于研究揭示大量矿物种类间的复杂相互关系[40]。众所周知，由于能够便捷地揭示网络成员间的联系，网络图被广泛应用于社交网络[51–54]、技术网络[55–58]，以及生物系统研究[59–62]。每个网络由顶点（节点）和连接顶点的边（链接）组成。节点之间的距

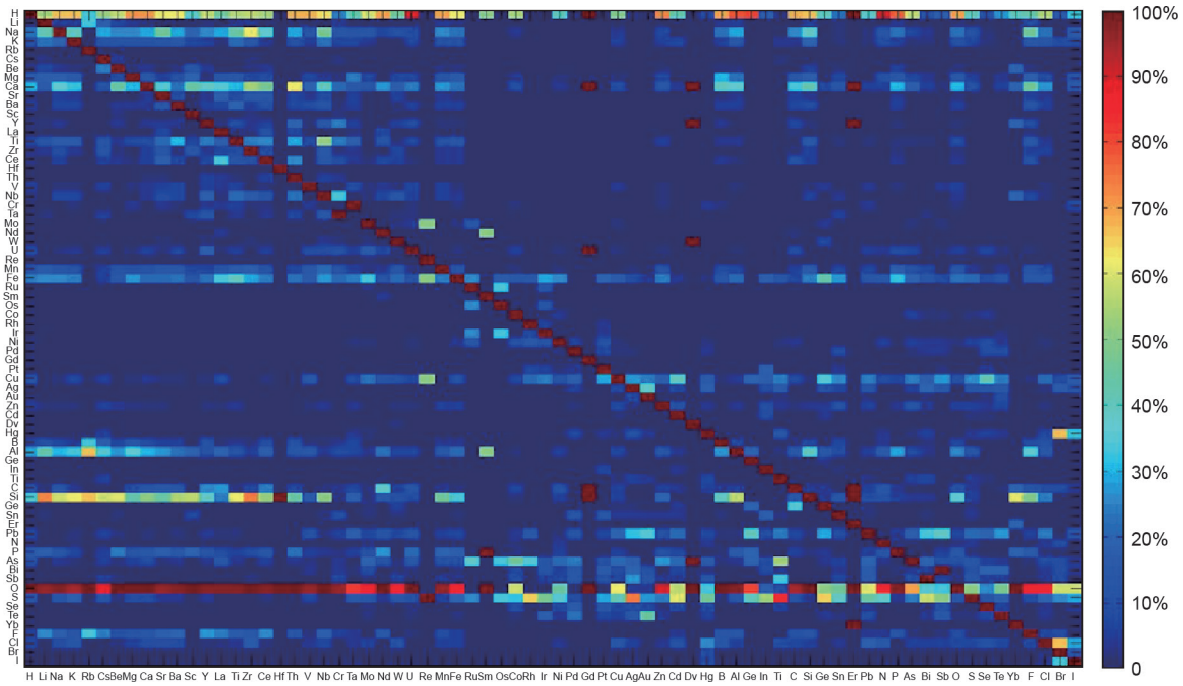


图7. Klee图揭示了矿物、元素，或者是其他物质的共存概率。图中展示了72 × 72的矿物矩阵中共存的化学元素，其中每个矩阵元素代表既包含元素*X*又包含元素*Y*的矿物比例。该矩阵是不对称矩阵，例如，所有含有铍的矿物质也都含有氧元素，但是只有一小部分含氧矿物质也同时含有铍元素。

离以及链路的长度是由两个节点的关联程度决定；最短间距代表最强链接。顶点和边的大小、形状和颜色也可以表示系统的其他属性。

共存矿物网络则提供了生动的网络图应用实例[†]。如图9[40]所示，每个节点代表一种矿物；这些节点大小代表每种矿物产地的相对数量，而节点颜色则代表其化学成分、晶体结构、共生或其他信息。这些高度交互式的视觉图像是由多维空间到二维或三维空间的投影，并揭示了从每个矿物节点到所有其他共存矿物节点的连通性。一般而言，对于含有*N*种矿物的连通网络，渲染是来自*N*−1维度空间的投影。在许多情况下，即便投影可能来自更高维度，其三维渲染图也能够提供非常重要的额外信息。

矿物（种类）网络图不仅反映了矿物的局域特性（比如给定矿物的共存矿物种类信息），而且也揭示了在单个数据中不容易发现的整体趋势（例如，以化学或共生模式进行的矿物聚类、矿物网络间的相互关联程度，以及其他隐藏的化学成分和时间演化信息）。网络统计分析的一个明显优势是能够通过网络度量表征网络的整体和局域统计特性[63, 64]。网络度量参数包括密度、中心性和直径，这些参数有助于进行相关网络比较，如表征含不同化学元素的矿物种类或者是给定元素的矿物时间序列网络[40]。

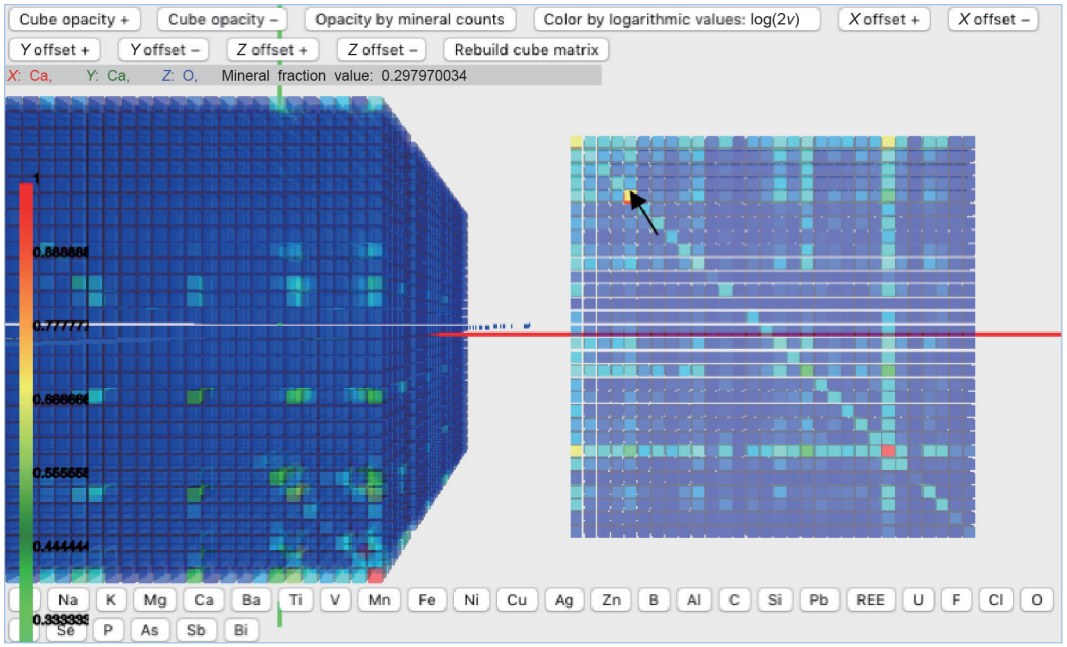
5.4. 网络二分图

矿物学研究所用的网络图有多种表达（rendering）方式，其中最重要的一种方式是二分图[65]，二分图能够清晰地区分两种明显不同的节点类型，如矿物种类信息或者产地信息（图10）。与其他自然或人工系统所不同的是，矿物二分网络有一个显著特征：其局部节点以U形（或三维“花瓶形状”）分布，少数常见矿物分布在U形（或“花瓶”）里面，而大量稀有矿物则呈点缀型分布在周围（图10）。这种分布是LNRE分布特征的直观表示，简明地揭示了矿物分布规律：分布广泛且非常常见的矿物种类数量极少，而其余大部分（种类数目多）矿物都是分布稀少（数量少）的稀有矿物。

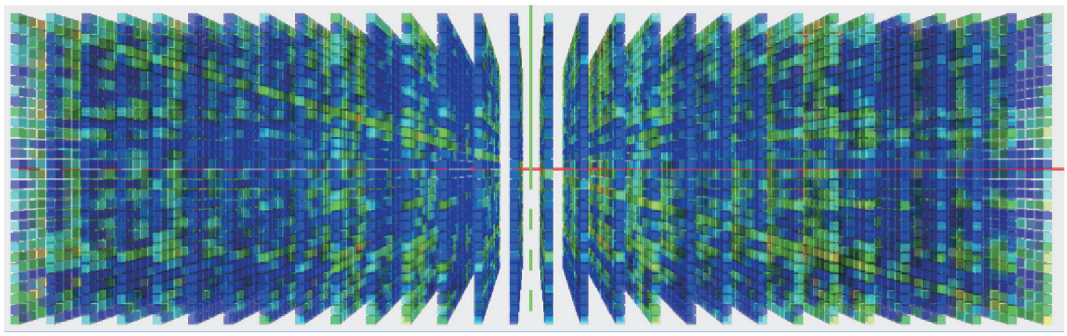
6. 展望

矿物学领域的数据驱动型研究仍处于起步阶段。当前开源获取的矿物数据规模至少还要再扩大10倍，并尝试努力恢复那些即将消失的暗数据。未来，新型数据分

[†] See https://dtdi.carnegiescience.edu for interactive examples.



(a)



(b)

图8. 交互式三维Klee图有助于加速发现矿物或者是矿物元素间的三重共存态。本例来自文献[50]，详细记录了矿物中化学元素三重共存态出现的概率。（a）其中，立方体形状的描述渲染比较难解释，但是立方体的任何平面切片都可以独立查看；（b）又或者采用“爆炸型”的视角描述该三维立方Klee图，从而有助于用户观察研究立方体内部的信息。其中，红线表示3维图像的中心线，该箭头指向众多“热点”信息中的一个。在钙-钙-氧的情形中，矿物质中的元素组合比基于地壳丰度的预测更为常见。REE:稀土元素（rare earth element）。

析方法与可视化技术，以及针对矿物学特性而定制的分析研究策略将得到进一步研究与综合运用。此外，随着来自火星、月球和其他星球的数据逐步收集，这些方法也将运用于其他类地星球矿物研究分析。

其中一个关键需求是将各种矿物数据库和时间深度（deep-time）数据集相融合，并与其他数据片段进行关联。当前，正在努力将矿物数据库与其他时间深度数据库（如地球化学、古生物学和蛋白质数据库）相关联，以便更全面地了解地球圈和生物圈[†]的协同演化过程。这些研究将有可能揭示一直在随时间变化的地球近地表矿物与地球化学环境是如何影响生命体的生物化学反应机制，以及生命体又如何反过来创造新的矿物种类和地球化学圈。

6.1. 关联度分析

也许最激动人心的应用前景是采用关联度分析方法有针对性地发现新矿物，包括有经济价值的新矿产资源。最近，Jolyon Ralph（个人通讯作者，2018年5月）的研究预测提供了一种新的尝试。采用成对矿物相关性预测方法，Ralph指出那些不常见的矿物钼铅矿（wulfenite，$PbMoO_4$）应该存在于新墨西哥州库克斯峰（Cookes Peak）的一个铅-锌-银开采矿区，该矿区迄今

[†] For example, https://dtdi.carnegiescience.edu.

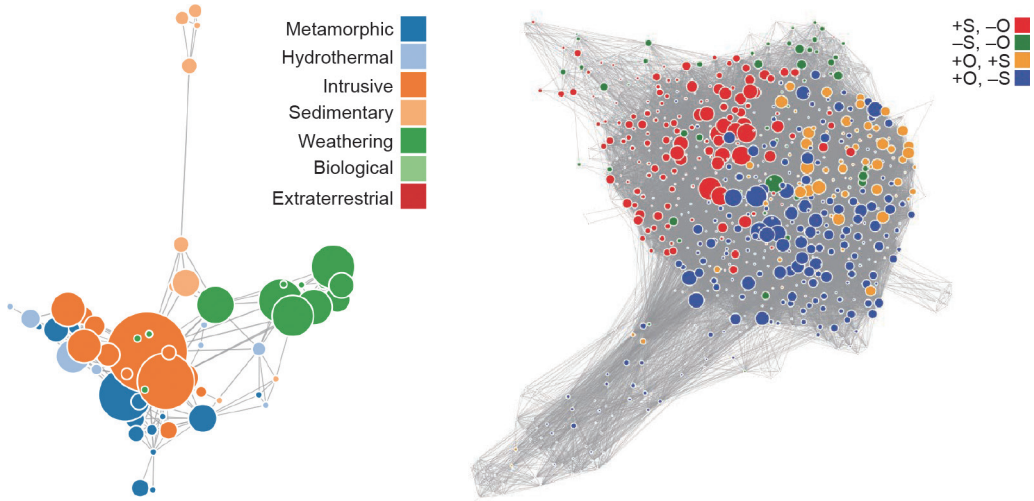


图9. 矿物种类网络图。（a）58种含铬矿物网络图，网络节点尺寸大小依据矿物出现频率设计，而节点颜色则根据形成方式设计（如插图）。该低密度网络揭示出了强烈的共生化和集群化趋势。（b）664种含铜矿物网络图，与图（a）类似，该图的节点尺寸大小也依据矿物出现频率设计，而节点颜色则根据含S/O或者不含S/O的富集方式设计（如插图，文献[40]）。

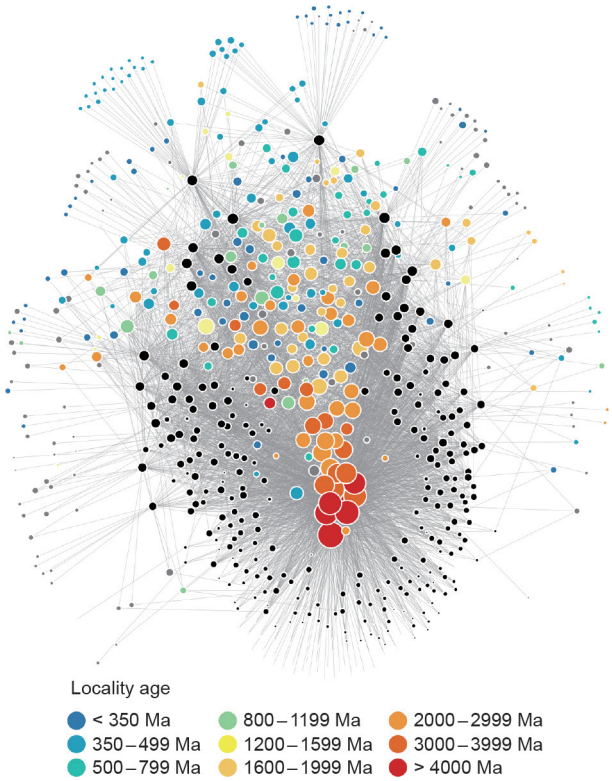


图10. 由403种含碳矿物构成的二分网络图。彩色圆代表含碳矿物种类，其中，圆的大小代表矿物出现的相对概率，而颜色则代表最早发现这些矿物的年代（如插图）。黑色圆表示矿物属地，其大小表示在这些区域发现不同含碳矿物的相对数量。该网络图反映了含碳矿物多样性和时空分布方面的重要信息。特别是，黑色圆呈现出U形分布——大量稀有含碳矿物分布在外围，而少数非常常见的含碳矿物则分布在U形内部。该图像也是图4中描述的LNRE的另外一种形象表述。我们注意到，大多数的常见矿物发现年代较为久远，而大多数的稀有矿物则发现年代较近。

报道过的矿物种类超过了75种，但是仍然没有关于钼铅矿的报道。在Ralph提出预言之后，当地矿产收藏家随后在该矿区进行了详细检查，发现了这种美丽但以前被忽视了的矿物质。

关联度分析（当用于推荐采购产品的时候又称作购物篮分析）采取了类似购物篮分析的策略，但是包含了更多维度的正面和负面共存信息[66–68]。对矿物进行关联度分析的初步试验是将搜索算法从成对共存数据信息扩展到包含矿物特征、种类等数据信息的更广泛应用组合场景。在不久的将来，我们希望能够通过查询mindat.org数据库来编制“缺失”矿物清单及其在特定区域的出现概率，这将成为未来矿物预测发展的一种可操作方法。

我们项目研究的目标之一是通过对更广泛的数据资源采用关联度分析来搜索新的矿物或其他自然资源，这些数据资源包括多个维度的数据信息，涉及矿物矿化事件、矿物化学成分信息（包括微量元素和同位素数据）及其物理性质，以及在那些矿化事件中涉及的物理、化学和生物环境背景信息。我们预期该专家推荐系统将会在下一代自然资源勘探中发挥关键作用。

6.2. 晶体化学系统学

最近，Gagné和Hawthorne[69–72]以及Gagné[73]等研究者在数据驱动型矿物学研究方面的前期研究给我们提供了氧化物、含氧盐和氮化物晶体有关化学键的基础性统计知识。这些核心知识也能够拓展到硫化物和硫酸盐矿物体系，并结合晶体结构的理想键价理论[74]、外加离子成键条件的理论预测，最终将能够更加精确地预测“缺失矿物”化学成分的最概然成分。这些价键数据

信息的引入将进一步推动获取高质量价键参数（详见参考文献[75]），这些参数在矿物演化研究中非常实用。比如在研究不断变化的近地表环境时，有了这些价键参数将能够更好地推断对氧化还原敏感的过渡金属氧化态。

7. 结语

数据驱动型矿物学研究是当前“开放数据行动”的一个方面，它将有助于推动科学新发现[76, 77]。该领域的研究总体进展将取决于综合数据资源库的建立、先进的数据分析方法和可视化技术应用开发，以及将这些先进技术应用于研究解决矿物学领域内突出问题等三大方面的协同进展。在某些情况下，这些数据科学方法将有助于推动矿物学研究朝着假说驱动型探索研究方向发展，从而增进我们对矿物多样性和分布特性的理解。换而言之，有助于揭开许多当前矿物学领域未解之谜的神秘面纱。更激动人心的是将多维分析方法应用于矿物学领域也将有助于促进全新的意想不到的新发现；换而言之，有助于探索前所未想、闻之未闻的新知识。

Acknowledgements

We are grateful to Ho-Kwang Mao and the organizers of this special issue for the opportunity to share our results. This publication is a contribution to the Deep Carbon Observatory. Studies of mineral evolution and mineral ecology are supported by grants from the Alfred P. Sloan Foundation (G-2016-7065), the W. M. Keck Foundation (grant entitled “Co-Evolution of the Geosphere and Biosphere”), the John Templeton Foundation (60645), the NASA Astrobiology Institute (1-NAI8_2-0007), a private foundation, and the Carnegie Institution for Science. Sergey V. Krivovichev acknowledges support from the Russian Science Foundation (19-17-00038).

Compliance with ethics guidelines

Robert M. Hazen, Robert T. Downs, Ahmed Eleish, Peter Fox, Olivier C. Gagné, Joshua J. Golden, Edward S. Grew, Daniel R. Hummer, Grethe Hystad, Sergey V. Krivovichev, Congrui Li, Chao Liu, Xiaogang Ma, Shaunna M. Morrison, Feifei Pan, Alexander J. Pires, Anirudh Prabhu, Jolyon Ralph, Simone E. Runyon, and Hao Zhong declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Gastil G. The distribution of mineral dates in time and space. Am J Sci 1960;258 (1):1–35.
[2] Nash JT, Granger HC, Adams SS. Geology and concepts of genesis of important types of uranium deposits. Econ Geol 1981:63–116.
[3] Zhabin AG. Is there evolution of mineral speciation on Earth? Dokl Earth Sci Sect 1981;247:142–4.
[4] Yushkin NP. Evolutionary ideas in modern mineralogy. Zap Vses Mineral Obshch 1982;116(4):432–42. Russian.
[5] Hazen RM, Papineau D, Bleeker W, Downs RT, Ferry J, McCoy T, et al. Mineral evolution. Am Mineral 2008;93(11–12):1693–720.
[6] Hazen RM, Ewing RJ, Sverjensky DA. Evolution of uranium and thorium minerals. Am Mineral 2009;94(10):1293–311.
[7] 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):953–63.
[8] Hazen RM, Golden JJ, Downs RT, Hysted 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):1013–42.
[9] Hazen RM, Papineau D. Mineralogical co-evolution of the geosphere and biosphere. In: Knoll AH, Canfield DE, Konhauser KO, editors. Fundamentals of geobiology. Oxford: Wiley-Blackwell; 2012. p. 333–50.
[10] Hazen RM, Jones AP, Kah L, Sverjensky DA. Carbon mineral evolution. In: Hazen RM, Jones AP, Baross J, editors. Carbon in Earth. Washington, DC: Mineralogical Society of America; 2013. p. 79–107.
[11] Hazen RM, Sverjensky DA, Azzolini D, Bish DL, Elmore S, Hinnov L, et al. Clay mineral evolution. Am Mineral 2013;98(11–12):2007–29.
[12] Hazen RM, Liu XM, Downs RT, Golden JJ, Pires AJ, Grew ES, et al. Mineral evolution: episodic metallogenesis, the supercontinent cycle, and the coevolving geosphere and biosphere. Soc Econ Geolog Special Pub 2014;18:1–15.
[13] Hazen RM, Grew ES, Origlieri M, Downs RT. On the mineralogy of the “Anthropocene Epoch”. Am Mineral 2017;102(3):595–611.
[14] Hazen RM. Evolution of minerals. Sci Am 2010;302(3):58–65.
[15] Hazen RM. Paleomineralogy of the Hadean Eon: a preliminary list. Am J Sci 2013;313(9):807–43.
[16] Hazen RM. Mineral evolution, the Great Oxidation Event, and the rise of colorful minerals. Mineralog Record 2015;46(805–816):34.
[17] Hazen RM. An evolutionary system of mineralogy: proposal for a classification based on natural kind clustering. Am Mineral. In press.
[18] Hazen RM, Eldredge N. Themes and variations in complex systems. Elements 2010;6(1):43–6.
[19] Hazen RM, Ferry JM. Mineral evolution: mineralogy in the fourth dimension. Elements 2010;6(1):9–12.
[20] Golden J, McMillan M, Downs RT, Hystad G, Stein HJ, Zimmerman A, et al. Rhenium variations in molybdenite (MoS2): evidence for progressive subsurface oxidation. Earth Planet Sci Lett 2013;366:1–5.
[21] Grew ES, Hazen RM. Evolution of the minerals of beryllium. Stein 2013:4–19.
[22] Grew ES, Hazen RM. Beryllium mineral evolution. Am Mineral 2014;99(5–6):999–1021.
[23] Krivovichev SV. Structural complexity of minerals: information storage and processing in the mineral world. Mineral Mag 2013;77(3):275–326.
[24] Krivovichev SV. Structural complexity of minerals and mineral parageneses: information and its evolution in the mineral world. In: Armbruster T, Danisi RM, editors. Highlights in mineralogical crystallography. Berlin/Boston: de Gruyter; 2015. p. 31–74.
[25] Grew ES, Dymek RF, De Hoog JCM, Harley SL, Boak JM, Hazen RM, et al. Boron isotopes in tourmaline from the 3.7–3.8 Ga Isua Belt, Greenland: sources for boron in Eoarchean continental crust and seawater. Geochim Cosmochim Acta 2015;163:156–77.
[26] Krivovichev SV, Krivovichev VG, Hazen RM. Structural and chemical complexity of minerals: correlations and time evolution. Eur J Mineral 2018;30(2):231–6.
[27] Liu C, Knoll AH, Hazen RM. Geochemical and mineralogical evidence that Rodinian assembly was unique. Nat Commun 2017;8(1):1950.
[28] 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):647–61.
[29] 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;426:154–7.

[30] Hystad G, Downs RT, Hazen RM, Golden JJ. Relative abundances for the mineral species on Earth: a statistical measure to characterize Earth-like planets based on Earth's mineralogy. Math Geosci 2017;49(2):179–94.
[31] 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):295–323.
[32] Hazen RM, Hystad G, Downs RT, Golden J, Pires A, Grew ES. Earth's "missing" minerals. Am Mineral 2015;100(10):2344–7.
[33] 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.
[34] Hazen RM, Hystad G, Golden JJ, Hummer DR, Liu C, Downs RT, et al. Cobalt mineral ecology. Am Mineral 2017;102(1):108–16.
[35] Grew ES, Krivovichev SV, Hazen RM, Hystad G. Evolution of structural complexity in boron minerals. Can Mineral 2016;54(1):125–43.
[36] Grew ES, Hystad G, Hazen RM, Krivovichev SV, Gorelova LA. How many boron minerals occur in Earth's upper crust? Am Mineral 2017;102(8): 1573–87.
[37] Hazen RM, Ausubel J. On the nature and significance of rarity in mineralogy. Am Mineral 2016;101(6):1245–51.
[38] Liu C, Hystad G, Golden JJ, Hummer DR, Downs RT, Morrison SM, et al. Chromium mineral ecology. Am Mineral 2017;102(3):612–9.
[39] Liu C, Eleish A, Hystad G, Golden JJ, Downs RT, Morrison SM, et al. Analysis and visualization of vanadium mineral diversity and distribution. Am Mineral 2018;103(7):1080–6.
[40] Morrison SM, Liu C, Eleish A, Prabhu A, Li C, Ralph J, et al. Network analysis of mineralogical systems. Am Mineral 2017;102(8):1588–96.
[41] Downs RT. The RRUFF project: an integrated study of the chemistry, crystallography, Raman and infrared spectroscopy of minerals. In: Proceedings of the 19th General Meeting of the International Mineralogical Association; 2006 July 23–28; Kobe, Japan; 2006.
[42] Lehnert KA, Walker D, Sarbas B. EarthChem: a geochemistry data network. Geochim Cosmochim Acta 2007;71:A559.
[43] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
[44] Fox P, Hendler J. Changing the equation on scientific data visualization. Science 2011;331(6018):705–8.
[45] Hazen RM. Data-driven abductive discovery in mineralogy. Am Mineral 2014;99(11–12):2165–70.
[46] Papike JJ, editor. Planetary materials. Chantilly: Mineralogical Society of America; 1998.
[47] Morrison SM, Downs RT, Blake DF, Vaniman DT, Ming DW, Rampe EB, et al. Crystal chemistry of martian minerals from Bradbury Landing through Naukluft Plateau, Gale crater, Mars. Am Mineral 2018;103(6):857–71.
[48] Liu XM, Kah LC, Knoll AH, Cui H, Kaufman AJ, Shahar A, et al. Tracing Earth's CO_2 evolution using Zn/Fe ratios in marine carbonate. Geochem Perspect Lett 2016;2:24–34.
[49] Carroll SB. Chance and necessity: the evolution of morphological complexity and diversity. Nature 2001;409(6823):1102–9.
[50] Ma X, Hummer D, Golden JJ, Fox PA, Hazen RM, Morrison SM, et al. Using visualized exploratory data analysis to facilitate collaboration and hypothesis generation in cross-disciplinary research. ISPRS Int J Geoinf 2017;6(11):368.
[51] Otte E, Rousseau R. Social network analysis: a powerful strategy, also for the information sciences. J Inf Sci 2002;28(6):441–53.
[52] Abraham A, Hassanien AE, Snasel V, editors. Computational social network analysis: trends, tools and research advances. New York: Springer; 2010.
[53] Pinheiro CAR. Social network analysis in telecommunications. Hoboken: Wiley; 2011.
[54] Kadushin C. Understanding social networks. New York: Oxford University Press; 2012.
[55] Hwang N, Houghtalen R. Fundamentals of hydraulic engineering systems. Upper Saddle River: Prentice Hall; 1996.
[56] Guimerà R, Mossa S, Turtschi A, Amaral LAN. The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. Proc Natl Acad Sci USA 2005;102(22):7794–9.
[57] Dong W, Pentland A. A network analysis of road traffic with vehicle tracking data. In: Proceedings of the American Association of Artificial Intelligence, Spring Symposium, Human Behavior Modeling; 2009 Mar 23–25; Palo Alto, CA, USA; 2009. p. 7–12.
[58] Pagani GA, Aiello M. The power grid as a complex network: a survey. Phys A 2013;392(11):2688–700.
[59] Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, et al. Network analysis of protein structures identifies functional residues. J Mol Biol 2004;344(4):1135–46.
[60] Banda-R K, Delgado-Salinas A, Dexter KG, Linares-Palomino R, Oliveira-Filho A, Prado D, et al. Plant diversity patterns in neotropical dry forests and their conservation implications. Science 2016;353(6306):1383–7.
[61] Corel E, Lopez P, Méheust R, Bapteste E. Network-thinking: graphs to analyze microbial complexity and evolution. Trends Microbiol 2016;24(3):224–37.
[62] Muscente AD, Prabhu A, Zhong H, Eleish A, Meyer MB, Fox P, et al. Quantifying ecological impacts of mass extinctions with network analysis of fossil communities. Proc Natl Acad Sci USA 2018;115(20):5217–22.
[63] Kolaczyk ED. Statistical analysis of network data. New York: Springer; 2009.
[64] Newman MEJ. Networks: an introduction. New York: Oxford University Press; 2013.
[65] Asratian AS, Denley TMJ, Häggkvist R. Bipartite graphs and their applications. New York: Cambridge University Press; 1998.
[66] Adomavicius G, Tuzhilin A. Context-aware recommender systems. In: Ricci F, Rokach L, Shapira B, Kantor PB, editors. Recommender systems handbook. Boston: Springer; 2011. p. 217–53.
[67] Ricci F, Rokach L, Shapira B. Introduction to recommender systems handbook. In: Ricci F, Rokach L, Shapira B, Kantor PB, editors. Recommender systems handbook. Boston: Springer; 2011. p. 1–35.
[68] Panniello U, Tuzhilin A, Gorgoglione M. Comparing context-aware recommender systems in terms of accuracy and diversity. User Model Useradapt Interact 2014;24(1–2):35–65.
[69] Gagné OC, Hawthorne FC. Bond-length distributions for ions bonded to oxygen: alkali and alkaline-earth metals. Acta Crystallogr B Struct Sci Cryst Eng Mater 2016;72(Pt 4):602–25.
[70] Gagné OC, Hawthorne FC. Bond-length distributions for ions bonded to oxygen: results for the non-metals and discussion of lone-pair stereoactivity and the polymerization of PO_4. Acta Crystallogr B 2018;74:79–96.
[71] Gagné OC, Hawthorne FC. Bond-length distributions for ions bonded to oxygen: metalloids and post-transition metals. Acta Crystallogr B 2018;74: 63–78.
[72] Gagné OC, Hawthorne FC. Bond-length distributions for ions bonded to oxygen: results for the transition metals and discussion of d^0 cations and the Jahn-Teller effect. Acta Cryst B 2018;74(Pt 1):79–96.
[73] Gagné OC. Bond-length distributions for ions bonded to oxygen: results for the lanthanides and actinides and discussion of the f-block contraction. Acta Crystallogr B 2018;74:49–62.
[74] Gagné OC, Mercier PHJ, Hawthorne FC. *A priori* bond-valence and bondlength calculations in rock-forming minerals. Acta Crystallogr B 2018;74: 470–82.
[75] Gagné OC, Hawthorne FC. Comprehensive derivation of bond-valence parameters for ion pairs involving oxygen. Acta Crystallogr B Struct Sci Cryst Eng Mater 2015;71(Pt 5):562–78.
[76] Schutt R, O'Neil C. Doing data science: straight talk from the frontline. New York: O'Reilly; 2013.
[77] Kitchin R. The data revolution: big data, open data, data infrastructures & their consequences. London: Sage; 2014.