Contents lists available at ScienceDirect

Engineering

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

Views & Comments

数据驱动的材料创新基础设施

汪洪 [a]，项晓东 [b]，张澜庭 [a]

[a] *Materials Genome Initiative Center & School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[b] *Department of Materials Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China*

近年来，材料基因组已经成为材料科学领域的一个热门话题。“材料基因组”（materials genome）一词的出现，很大程度上受到成功的人类基因组计划的启发。传统上，新材料和新工艺的发现和开发依赖于科学直觉和漫长的试错过程。多年来，材料科学家渴望找到某种类似于生物基因的材料基本构造单元，其排序及缺陷结构或可决定材料的性质或功能。通过了解这些构件，他们希望能够按需设计材料，从而加速材料的发现和开发，并降低成本。自2011年美国启动“材料基因组计划”[1,2]以来，其他主要经济体如欧盟[3,4]、日本[5]和中国都在国家层面设立了类似的科学计划。然而关于什么是“材料基因组”，一直众说纷纭，难下定论。近期取得的共识是其仅作为设计预测材料研发模式的代称[6]。材料基因工程（materials genome engineering, MGE）意味着通过交叉融合高通量计算、高通量实验和材料信息学技术，速度更快、效率更高、成本更少地掌握成分-组织-工艺-性能间的关联关系——这些恰恰构成了材料设计的基础。

材料基因工程的工作模式可大致可分为实验驱动、计算驱动和数据驱动[7]三种。实验驱动模式基于高通量合成与表征实验，直接快速优化与筛选材料。这种模式的典型代表是高通量组合材料芯片技术[8]。计算驱动模式基于计算模拟，预测有希望的候选材料，再进行实验验证[9]，大大缩小实验范围。数据驱动模式基于大量数据，借助材料信息学方法建立模型，即利用人工智能（AI）方法，如机器学习，解析多参数间复杂的关联关系，预测出候选材料[10]。从人类认识自然的过程来看，数千年来，科学探索跨越了实验观测、理论推演、计算仿真几个阶段。今天，利用前所未有的计算能力和大规模的数据收集能力，现代科学正在进入“第四范式”[11]，即密集数据+人工智能。材料基因工程的数据驱动模式正是“第四范式”的体现。

应该看到，实验和计算驱动模式的实质是基于事实的判断或基于物理规律的推演，并未从根本上改变材料科学的既有思维模式与工作套路。与之形成对照的是，数据驱动模式使用人工智能来揭示隐藏在海量数据背后的关联关系，为现有的常规研究增加了新的维度和视角，因此这个“工具箱里的新成员”必将带来新的能力，它的运用可能产生颠覆性的效果。尽管如此，我们还须指出，数据驱动模式绝不意味着对实验驱动和计算驱动模式的简单取代。相反，它应看作是对传统认知范式的有力补充和延伸，适时发挥各自的作用。此外，还应将领域知识引入机器学习机制，为基于人工智能的模型提供指导并提高其有效性。

充足的材料数据是全面实施数据驱动模式的基本前提。尽管世界各地的数据库已经收集的数据数以亿计，但在材料问题的多样性和复杂性面前，这不过是沧海一粟。据简单估计[7]，任取四个元素可组成200万个四元体系，按数据密度为1%成分计算，共应有上万亿个多维数据点。事实上，材料数据的高度匮乏，严重限制了数据驱动模式在材料领域全面展开。在数据驱动模式下，快速产生大量材料数据的能力变得至关重要。

英文原文: Engineering 2020, 6(6): 609–611
引用本文: Hong Wang, X.-D. Xiang, Lanting Zhang. On the Data-Driven Materials Innovation Infrastructure. *Engineering*, https://doi.org/10.1016/j.eng.2020.04.004

在许多方面，现有的材料研究基础设施是为满足当前的需要而设计和开发的。作为全新的材料科学研究套路，材料基因工程需要发展与之相适应的基础设施，从而保证新型技术体系得到有效实施。新型材料创新基础设施应以数据为中心，聚焦于数据的产生与利用。数据平台具有数据收集、存储、处理、交换、共享和网络协作的综合能力[12]，包括基于AI方法的建模软件工具库与符合材料基因工程理念的数据库；高通量实验与高通量计算平台技术恰好为快速获取大量数据提供了有效途径，同时也满足实验驱动与计算驱动模式的技术需求。这样，材料基因工程的三个技术要素实现了内在的协同，形成了缺一不可的深度融合关系。

材料基因工程数据除了体量大外，还应保证数据具有高度完整性、系统性、一致性和多参量综合性。在理想条件下，这些数据应产生于一个集中建立或虚拟链接的平台，或可称之为“数据工厂”（图1），它们能够像工业生产线一样以标准化的方式批量地生产数据。实验“数据工厂”可以是基于大型科学设施（如同步加速器光源、中子源等）的大规模系统性的高通量综合制备与表征平台，或集成原位制备和多参数表征手段为一体的实验设施。计算“数据工厂”可以是一个拥有各种高通量计算软件和硬件的平台，能够通过批量计算生成大量的综合材料数据。“数据工厂”将给数据生成带来一系列重大变化。第一，为了更广泛的长远的目标，全面的材料数据将被大规模地有意识地产生，而不再局限于作为分散的具有特定目的的实验或计算的副产物；第二，“数据工厂”将数据的产生由个体活动转变为有组织的社会活动；第三，这种有组织的努力将把数据的社会属性从私有财产转变为公共资源。因此，数据的质量、一致性和全面性将得到提高，数据共享将变得更加简单，社会总成本也将降低。这种新型的数据产生形式是材料科学的革命性变化。

目前，国内外已开发了一系列基于高通量计算平台或计算“数据工厂”的数据库，如Materials Project [13]、Automatic Flow for Materials Discovery (AFLOW) [14]、Open Quantum Materials Database (OQMD) [15]、Novel *Materials* Discovery (NOMAD) [16]和MatCloud [17]。High-Throughput Experimental Materials *Database* (HTEM-DB) [18]是由美国国家可再生能源实验室（NREL）开发的利用高通量薄膜技术合成无机材料的开放实验数据库。它已初步具有实验“数据工厂”的特征。由中国国家重点研发计划支持的计算与实验“数据工厂”目前正在建设中。

材料基因工程的另一项重要任务是改革材料界多年来形成的封闭型工作方式，培育开放、协作的新型“大科学”研发模式。为了突破长期以来研究数据私有性的局限，让数据为全体研究者共享，Mons与他的合作者[19,20]共同提出了数据可发现、可访问、可交互、可重复使用的FAIR（findable, accessible, interoperable, reusable）数据原则。建立与之适应的数据标准是确保数据符合FAIR原则的一个重要方面。为此，最近发布的中国试验与材料标准（CSTM）《材料基因工程数据通则》[21]是对数据内容进行标准化的首次尝试（尽管具体数据格式标准仍有待建立）。这里，数据分为样品、源数据（未经处理的数据）与衍生数据（经分析处理得到的数据）三类，以每次操作（样品制备/表征/计算/数据处理）为条目单位，分别赋予独立资源标识（如DOI或符合GB/T 32843—2016的标识）。样品可以是实验产生的实物，也可以是经计算产生的虚拟物。同理，源数据可以是直接测量的结果，也可以在给定条件下通过计算/模拟生成。每个数据条目都应尽可能完整地收集与操作相关的元数据。将样品单独列为一类数据是之

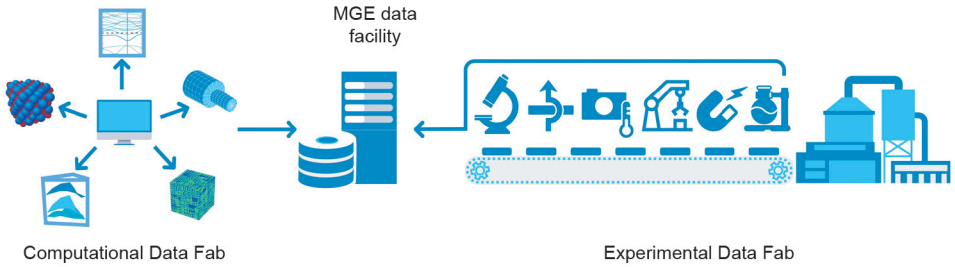

图1.“数据工厂”——像一条工业生产线一样以标准化方式批量生产数据的专用设施——概念图示。如图中右侧所示，实验数据工厂包含系统的、高通量的合成和表征设置，产生的多参量数据集包括机械、电气、光学、热学、磁学和声学特征及性能等。理想情况下，所有性能测量都是在同一样品上完成，最好同时进行，甚至是实时原位表征。如图中左侧所示，计算数据工厂可以是一个拥有各种高通量计算软硬件的计算中心，通过密度泛函理论、分子动力学、CALPHAD方法、相场模拟、有限元分析等多种方法，批量计算生成从原子尺度到宏观尺度的大量综合数据。数据工厂可以在同一地点集中建立，也可以是一组虚拟链接站点组成的平台。

前其他数据中都没有的独特做法，其最大优点是使样品本身和数据一样成为符合FAIR原则的公共资源，便于被发现、共享和重复使用。

综上所述，材料基因工程的数据驱动模式提出了一种新的材料创新范式，它与当前的思维和行为方式有着根本的不同，需要相应的全新基础设施来支撑。必要的基础设施包括一个以数据为中心的集成平台，整合了数据设施、高通量实验和高通量计算模块，以全面覆盖数据生产、存储、分析、共享和协同能力。这样的平台将产生和利用大量符合FAIR原则的数据，以支撑数据驱动模式的发展，同时也服务于实验驱动和计算驱动模式的实践。

致谢

本文作者感谢国家重点研发计划项目（2018YFB0703600）的经费支持。

References

[1] Holdren JP. Materials genome initiative for global competitiveness [Internet]. Washington, DC: National Science and Technology Council; 2011 Jun [cited 2018 Mar 31]. Available from: https://www.mgi.gov/sites/ default/files/documents/materials_genome_initiative-final.pdf.
[2] MGI strategic plan [Internet]. Washington, DC: National Science and Technology Council; [cited 2018 Mar 31]. Available from: https://www. mgi.gov/sites/default/files/documents/mgi_strategic_plan_-_dec_2014.pdf.
[3] Accelerated metallurgy—the accelerated discovery of alloy formulations using combinatorial principles [Internet]. Luxembourg: The Community Research and Development Information Service; c1994–2019 [updated 2019 Aug 2; cited 2019 Mar 5]. Available from: https://cordis.europa.eu/project/id/263206.
[4] Jarvis D, Raabe D, Singer R, van Houtte P, Vahlas C, Alford N, editors. Metallurgy Europe: a renaissance programme for 2012–2022. Strasbourg: EFS; 2012.
[5] JST support program for starting up innovationhub, Information Integration, new scientific approach for materials research [Internet]. Tsukuba: National Institute for Materials Science; c2020 [cited 2019 Mar 5]. Available from: https://www.nims.go.jp/eng/research/MII-I/index.html.
[6] Wang H, Xiang Y, Xiang X, Chen L. Materials genome enables research and development revolution. Sci Technol Rev 2015;33(10):13–9.
[7] Wang H, Xiang X, Zhang L. Data + AI is the core of materials genomic engineering. Sci Technol Rev 2018;36(14):15–21.
[8] Xiang X, Sun X, Briceño G, Lou Y, Wang K, Chang H, et al. A combinatorial approach to materials discovery. Science 1995;268(5218):1738–40.
[9] Ceder G, Persson K. The stuff of dreams. Sci Am 2013;309(6):36–40.
[10] Raccuglia P, Elbert KC, Adler PD, Falk C, Wenny MB, Mollo A, et al. Machinelearning- assisted materials discovery using failed experiments. Nature 2016;533(7601):73–6.
[11] Hey T, Tansley S, Tolle KM, editors. The fourth paradigm: data-intensive scientific discovery. Redmond: Microsoft Research Press; 2009.
[12] Ward C, Brinson LC, Galli G, Kalidindi SR, MehtaA, Meredig B, et al. Building a materials data infrastructure: opening new pathways to discovery and innovation in science and engineering [Internet]. Pittsburgh: The Minerals, Metals & Materials Society; [cited 2019 Mar 5]. Available from: http://www.tms.org/Publications/Studies/Materials_Data_ Infrastructure/Materials_Data_Infrastructure.aspx?hkey=d228f86c-e269-49a2- a638-395285b760e4.
[13] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. The materials project: a materials genome approach to accelerating materials innovation. APL Mater 2013;1(1):011002.
[14] Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, Taylor RH, et al. AFLOW: an automatic framework for high-throughput materials discovery. Comput Mater Sci 2012;58:218–26.
[15] Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, et al. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. Comput Mater 2015;1(1):15010.
[16] Draxl C, Scheffler M. NOMAD: the FAIR concept for big data-driven materials science. MRS Bull 2018;43(9):676–82.
[17] MatCloud [Internet]. Beijing: Materials Genetic Engineering Information Technology Application Laboratory; c2016 [cited 2019 Mar 5]. Available from: http://matcloud.cnic.cn/static/view/about.html.
[18] Zakutayev A, Wunder N, Schwarting M, Perkins JD, White R, Munch K, et al. High throughput experimental materials database (2018) [Internet]. Denver: National Renewable Energy Laboratory; [cited 2019 Mar 5]. Available from: https://htem.nrel.gov/#/about/content.
[19] 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(1):160018.
[20] Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European open science cloud. Inf Serv Use 2017;37(1):49–56.
[21] China Standards of Testing and Materials (CSTM). T/CSTM 00120–2019: general rule for materials genome engineering data. Chinese standard. Beijing: Metallurgical Industry Press; 2019. Chinese.