数据+人工智能是材料基因工程的核心

汪洪[1,2],项晓东[3],张澜庭[1,2]

1. 上海交通大学材料基因组联合研究中心,上海 200240
2. 上海交通大学材料科学与工程学院,上海 200240
3. 南方科技大学材料科学与工程系,深圳 518055

摘要　材料基因工程的工作模式,可大致总结为实验驱动、计算驱动和数据驱动3种。以“数据+人工智能”为标志的数据驱动模式围绕数据产生与数据处理展开,代表了材料基因工程的核心理念与发展方向。材料研究由“试错法”向科学第四范式的根本转变,将更快、更准、更省地获得成分-结构-工艺-性能间的关系。在数据密集型科学时代,快速获取大量材料数据的能力成为关键,而基于高通量实验与高通量计算的“数据工厂”是满足材料基因工程数据需求的重要平台。

关键词　材料基因工程;数据驱动;高通量实验;高通量计算

2012—2014年,本文部分作者有幸参与了中国工程院关于中国版材料基因组计划咨询报告的撰写工作,之后以咨询报告为基础在《科技导报》发表了题为《材料基因组——材料研发新模式》的文章[1],对材料基因组的理念进行了归纳总结。在之后的3年中,国内外对材料基因工程的认识与理解在不断加深。本文试图进一步探讨有关材料基因工程工作模式、材料数据的内涵、人工智能在材料基因工程中的核心作用、材料基因工程与第四科学范式间的关系等方面问题,以期引发材料科学界对材料基因工程的更多思考与重视。

材料基因组计划(Materials Genome Initiative,MGI)的出现反映了全球对加速材料从发现到应用进程的需要。进入21世纪以来,科技革命作为产业革命先导的趋势日见明朗,依赖于科学直觉与试错的传统材料研究方法已无法适应时代的发展。2011年6月,时任美国总统奥巴马宣布启动材料基因组计划,意在改革传统材料研究的封闭型工作方式,培育开放、协作的新型“大科学”研发模式,从而实现将材料从发现到应用的速度至少提高1倍,成本减半[2-3]的目标。具体措施包括:(1) 发展高通量材料模拟计算工具和方法,加快材料筛选和设计,减少耗时费力的“试错实验”;(2) 发展和推广高通量材料实验技术及装备,快速、准确地获取材料计算所需大量的关键数据,对候选材料进行筛选和验证;(3) 发展和完善材料数据库/信息学工具,有效管理与利用材料从发现到应用全过程的数据链。

材料基因组(materials genome)这个名词的出现是受到人类基因组计划的启发[4]。在生命世界中,人们发现脱氧核糖核酸是组成蛋白质的基本单元,其排列及缺陷结构被称为生物基因组,它决定了生物体的功能及疾病。因此,生物基因组的信息可以用于预测从而

收稿日期:2018-05-15;修回日期:2018-06-19

基金项目:国家重点研发计划项目(2017YFB0701900);上海市科学技术委员会研发平台专项(16DZ2260602)

作者简介:汪洪,教授,研究方向为材料基因工程,电子信箱:hongwang2@sjtu.edu.cn

引用格式:汪洪, 项晓东, 张澜庭. 数据+人工智能是材料基因工程的核心[J]. 科技导报, 2018, 36(14): 15-21; doi: 10.3981/j.issn.1000-7857.2018.14.003

改变生物体的性状和功能。人类基因组计划实施20多年来,人类对生物基因基础数据的采集技术以及掌握程度发生了翻天覆地的变化,获得全套人体基因图谱的时间和成本已由2001年的数周和上亿美元降至2015年的2 h和1000美元左右,根据对生物基因基础数据的认识进而改良物种、治疗疾病已开始成为现实[5]。

与此类比,材料基本单元(原子、分子、功能团等)的排序及缺陷结构决定了材料的性质或功能,或可称之为"材料基因组数据"。人们希望通过掌握材料基因组信息来实现对材料的按需设计。由于上述排序及缺陷结构取决于材料的热力学合成参数与加工工艺,材料体系的成分-组织-工艺-性能间的关联关系构成了材料设计的基础。关于材料基因的定义,国内外虽多有讨论,但众说纷纭,迄今仍无统一标准,通常仅作为设计预测型材料研发模式的代称。材料基因工程意味着通过交叉融合高通量模拟计算、高通量实验和人工智能数据挖掘技术,使掌握成分-组织-工艺-性能间关联规律的速度更快、效率更高、成本更少。

1 MGI的3种工作模式

在新型材料研发模式下,大致可以总结出材料基因工程的3种工作模式。

1)模式1:以实验驱动的模式,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。受集成电路芯片与基因芯片启发,在一块基底上,通过精妙设计,以任意元素为基本单元,组合集成并且快速表征成千上万种成分、结构、物相等。随着实验通量的大幅提高带来研究效率质的转变,使通常需数年完成的三元相图(结构及物理特性)可在几天内完成,实现材料搜索的"多-快-好-省"[6-7]。在化学反应合成方面,Merck公司、Pfizer公司相继开发了自动化高通量反应筛选平台[8-9],能够与直接测定反应产物对靶标蛋白的亲和性,并进行排序[10]。

2)模式2:以计算驱动的模式,或称理性设计指导下的高效筛选。首先基于计算模拟,预测有希望的候选材料,缩小实验范围,再进行实验验证。Ceder研究组[11]在美国Materials Project高通量计算平台上,通过大规模自动计算流程并按照一定的判据对电池的电极材料、固态电解质材料进行筛选,例如从130000候选者中筛选出200多种潜在的碱性电池电极材料,再进行实验研究。Allison等[12]在福特汽车公司设计了一套针对铝合金的虚拟铸造系统,根据产品设计和选材,利用有限元、相场模拟等方法,对发动机缸体的制造流程从材料制备、器件制造和加工工艺等进行了全方位的设计、模拟和实验研究,实现了工艺过程和组织性能可设计可控制的效果。Olson等[13]结合计算相图热力学方法(CALPHAD)指导合金设计,在飞机起落架高强钢Ferrium M54的开发中,成功地从Ferrium S53的8.5年周期缩短为5年。

3)模式3:以数据驱动的模式,或称为材料信息学模式。基于大量数据,采用机器学习找出特征性参量,进行数据挖掘(人工智能+数据),预测出候选材料。所谓机器学习,一个比较严谨的定义是指计算机代码能从经验E中学习完成以表现P为考量的任务T,且它在完成任务T方面的表现(由P考量)能随着经验E而改进[14-15]。近年来,利用人工智能进行材料研究的成果开始大量涌现。Raccuglia等[16]采用机器学习中决策树方法从之前"不成功"的实验数据中学习规律,用于成功预测新的金属有机氧化物材料。对比有经验的化学家人工判断,机器预测结果成功率以89%:78%胜出,充分展示了机器学习方法的强大能力。Xue等[17-19]利用贝叶斯线性回归等多种回归模型加速了形状记忆合金、压电材料等的开发;Ren等[20]报道了通过高通量实验结果与机器学习模型间的迭代加速发现金属玻璃的工作。2018年3月,Waller等[21]发表了人工智能技术进行药物自动化开发的工作,利用深度神经网络+蒙特卡洛树的方式实现了化学反应逆向合成路线设计。

从模式3中得到的一个重要启示是:只有不好的"结果",没有不好的数据。结果好坏取决于人为判断,是主观的;而数据永远是对自然规律的反映(如果排除了操作失误的因素),是客观的。在实际工作中,存在大量被认为与应用目标不符的"失败"数据,通常被遗忘在数据本上多年,客观上是对社会资源的巨大浪费。如能挖掘它们潜在的价值,相信会得到更多的启示。

另一个重要的启示是:人工智能方法擅长在纷繁的数据中发现、建立背后的关联。材料是由极大数量原子构成的,描述材料的重要参量不仅有成分、结构,还包括缺陷等,十分复杂。材料性能通常是多个物理机制耦合的结果,很少只受单一因素影响。因此,仅仅

建立起与某一个参量相关的简单模型,很难描述。利用人工智能方法可以同时研究多参量耦合的效果,增加理解问题的维度。人工智能方法的引入对于理解与发现各种材料参数与性能间的关联极有帮助,Mueller等[22]和Liu等[15]分别综述了机器学习方法在材料学领域中的应用现状。

当前,随着硬件技术和软件平台的发展,云计算使数据存储和访问成为一种廉价商品。当真正能获取大量数据的时候,如何从中提取出有效信息则成为关键。人脑推理活动的本质是建立因素间的关联性,每个人解决问题的能力各不相同,取决于知识的丰富程度和推理能力的高低。以统计、拟合算法为基础的人工智能方法,利用计算机长于重复运算的特点,可以突破人脑所能关联因素的数量限制,从而在高维参量空间中构建关联关系。神经元网络具有学习高层次抽象特征的能力。利用深层神经元网络(深度学习),已能够在两幅照片间构建像素级的关联关系,使图像判读的精准度达到甚至超过了人脑的水平。

2　科学探索的4种范式

数千年来,从人类认识自然的过程来看,科学探索跨越了实验观测、理论推演、计算仿真的阶段,正进入"人工智能分析密集型数据"的"第四范式"。从远古开始,人类对自然的认识是从亲身经验(也就是实验观测)开始的。17世纪前后,当实验观测积累达到一定程度,从现象中可以归纳总结出理论规律,人们开始使用数学方程这种简明的语言来描述具有共性的现象及其规律,并由此通过假设推演出结论(理论推演)。最具代表性的理论如牛顿定律、热力学三大定律、电磁波麦克斯韦方程、狭义及广义相对论、规范量子场论等。然而,现实中许多问题的数学模型过于复杂,受限于求解能力,无法获得解析解。于是,出现了数学方程的数值近似解。

自1946年电子计算机问世以来,特别是1980年以来,计算机的计算能力出现了爆炸式增长,模拟仿真技术也随之快速发展。如今,根据已知关系模拟结果做出预估的方法,已经逐渐成为科学与技术领域通行的做法。

随着数据量的迅速增长,科学探索正在进入数据密集型的第四范式。正如已故微软公司著名科学家、图灵奖获得者吉姆·格雷(Jim Gray)在《The fourth paradigm》[23](《第四范式》)一书中所描绘的:"今天在科学的很多领域里,科学家们已不再直接透过望远镜观察,……新的模式是由仪器采集或模拟产生数据,经过软件处理,将产生的信息或知识存储在计算机里。"应该看到,"第四范式"中的数据处理计算与"第三范式"中的模拟仿真计算有着截然不同的意义。模拟仿真计算是基于由已知物理规律决定的因果律的认识进行的推演,而数据密集型范式则是基于算法对数据进行分析,从而建立起多维参数间的复杂关联关系。科学范式改变的基础在于当今数字时代强大的数据产生能力和处理能力,同时它也为分析解决复杂体系科学问题提供了新的途径。

材料基因工程的3种工作模式与科学探索的4个范式是密切关联的。实验驱动在认识过程上属典型的"第一范式",其加速效果的实质是以量取胜,类似于快速穷举法;计算驱动是地地道道的"第三范式",根据现有理论的模拟仿真计算,再进行少量的实验验证。这个过程避免了大量试错实验的进行,取得降本增效的结果。但不可否认,二者均是在传统思维下基于事实的判断或基于物理规律的推演,并未从根本上改变原有思维模式与工作套路。

数据驱动与前2种模式形成鲜明对比,它以大量数据为前提,运用机器学习、数据挖掘技术,更快、更准、更省地建立起成分-结构-工艺-性能间的关联关系。数据驱动模式是科学"第四范式"在材料科学中的具体体现,它秉承了完全不同的思维逻辑,为材料科学引入了真正的革命性元素,也代表了认识的更高境界,它的全面应用必将产生颠覆性的效果。

鉴于认识范式的差别,材料基因工程数据驱动模式的研发路径与传统研发路径有着较为根本性的不同,远大于与实验模式和计算模式间的差别。受限于人脑对信息的处理能力,传统思维是以单一目标为导向,在实验设计中尽量降低变量维度(基本上每次只变化一个参数),按照理论与经验人为地确定探索方向。当结果符合目标方向,将沿同一方向继续尝试;如果与目标渐行渐远,便进行调整。经过大量试错,最终得到一条沿目标方向曲折前行、不断渐近的轨迹。形成鲜明对照的是,数据驱动模式基于对大量数据进行分析,这些数据可能来自于现有数据库,或高通量表征,也可能通过高通量计算得到。它们覆盖较广阔的参数空

间,其中既包含了传统意义上与目标一致的“好”数据,也包含与目标不一致的“不好”数据,因此分布不再局限于起点至目标连线周边,所得到的规律也将更具有普适性。简单来说,传统路径是以目标为导向,追求直接效果;而数据驱动模式的路径更加注重全局,通过对完整、系统的数据的分析,找出背后隐含的关系。显然,数据驱动模式对问题的认识更加深刻,更加全面。

3 “数据+人工智能”是材料基因工程的核心

数据驱动模式代表了材料基因工程核心的理念和最先进的方法。互联网时代令数据传播、分享的门槛大大降低,移动终端设备的普及令数据的产生发生了爆炸式的增长,计算机硬件计算能力的提升又令大数据的计算分析成为可能,从而催生了科学第四范式。随着第四范式的诞生,所能解决问题的复杂度有进一步提升,在这样的循环中推进了科学技术的发展。可以看到,在人工智能的时代,数据是最核心的资源,也是实践材料科学第四范式的必要基础。

当前数据分析在不同科学领域中的应用状况,与这些领域中数据量是有着重要关系的。例如,天文学和粒子物理方面每年产生的数据超过1 PB[24],主要由大型科学装置产生。美国的大型综合巡天望远镜(LSST)每晚的观测数据量是15 TB[25]。中国郭守敬望远镜(LAMOST)截至2016年12月已经发布了768万条光谱,成为世界上获取光谱数目最多的望远镜[26]。在生命科学领域中,数据则主要来自高通量实验。根据维基百科报道[27],美国国立卫生研究院(NIH)的生物基因序列库GenBank迄今已收录了超过2亿条基因序列,并正以大约每18个月翻一番的速度增长;深圳华大基因研究院每月仅原始测序相关的数据量就达到60 TB以上。与此同时,随着计算模拟能力的不断提高,高通量计算也成为大量数据的重要来源之一[23-24]。

数据是材料基因组工程的要素之一,各国都十分重视材料数据库的建设。美国国家标准技术院(NIST)Materials Data Facility收集的数据量已达到12.5 TB[28];美国的Materials Project、OQMD和AFLOW等高通量计算平台收录了超过280万种化合物数据;瑞士的Pauling File数据库[29],收录了4.6万余条相图数据、32万条晶体结构数据、12.5万余条物理性能数据,是世界上最大的无机化合物数据库;英国的Granta Design公司提供的材料天地(Material Universe)和工艺天地(Process Universe)数据库收录了3900种材料、240种工艺的数据[30];日本物质·材料研究机构(NIMS)建设的MatNavi数据库是关于高分子、陶瓷、合金、超导材料、复合材料和扩散的世界上最大的数据库之一[31]。据估计,中国公开的材料数据库中也收录了数百万条材料数据[32-33]。然而,由于材料的多样性与复杂性,已获得的材料数据只是沧海一粟,还远不能满足数据科学的要求。例如从元素周期表60个元素中任取3个元素组成三元体系,可组成近10万个三元体系,按照数据密度为1%进行估算,每个三元体系5000个数据点(多维热力学及物理性能参数),共应有5亿个多维数据点;任取4个元素可组成200万个四元体系,每个体系50万个多维参数数据点,共应有10000亿个多维数据点。因此要使材料科学全面进入科学探索的第四范式,必须首先解决材料数据匮乏这一全球性瓶颈问题。

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,荷兰莱顿大学的Barend Mons等提出了数据可发现、可访问、可交互、可重复使用的FAIR(findable,accessible,interoperable,reusable)数据原则[34]。其中,数据可重复使用在材料基因工程中非常重要。传统材料数据库一般收集由源数据处理而得到的分析结果(如各种材料性能参数等),而源数据通常分散在实验者手中,不被收录,且源数据格式多样,不便为其他人再次利用。再有,这些数据往往以特定应用为目标,包含的材料属性相对有限,缺乏综合性。这样,数据可关联的参数就比较有限。这与传统材料研究方式与数据产生方式有极大关系。同时,符合材料基因工程思想的材料数据模型标准和存储架构尚未建立,因此现有的材料数据库大多不能满足材料基因工程的需要。

作为在科学第四范式下的全新的材料科学研究套路,材料基因工程需要发展和建立新的技术体系及与之相适应的基础设施。材料数据基础设施建设应包括数据存储库、数据工具和e-合作平台3个核心组成部分[35]。针对中国当前的实际情况,一方面,需要建立以人工智能工具为基础的数据平台,同时构建起符合材料基因工程理念的数据库,或将已有数据库按照材料

基因工程需要进行改造,更重要的是系统、快速地充实大量新数据。为此,快速获取大量材料数据的能力成为关键,而高通量实验与高通量计算技术恰恰为快速获取大量数据提供了有效途径,可以作为数据的重要来源。于是,材料基因工程的3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。因此除数据平台外,材料基因工程基础设施还必须包括高通量实验平台和高通量计算平台。

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,这些数据可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"。实验"数据工厂"可以是基于大科学平台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,流水线般标准化地批量产生数据。计算"数据工厂"可以是各种高通量计算软件及硬件平台,通过批量计算产生大量系统的综合的材料数据。利用数据标识码技术[36],结合高通量实验(或高通量计算)数据格式标准,就可以从实验线站上导出记录样品信息、实验条件和实验源数据(或计算条件和计算源数据)的具有唯一标识的、符合FAIR原则的数据,供社会使用。数据工厂将数据产生由个体活动变为社会活动,数据由个体所有变为了社会资源,提高了共享程度,节约了社会成本,这种新型的数据产生形式必将引发材料科学的革命性变化。

迄今国际上尚未建成以标准化流水线般产生实验数据的实验平台。当前提出的数据库框架仅着眼于将各家产生的数据集中收集处理。如此收集到材料数据具有多源、分散、关联关系复杂的特点,不方便使用。例如美国密西根大学的Materials Commons[37]和NIST资助建立的Materials Data Facility[28,38]等数据平台则突出其数据收集的功能,将格式问题留给用户自行处理。美国材料数据公司Citrine Informatics公司建立了以物理信息文件(PIF)为标准数据模式的Citrination平台[39],试图在普适性、灵活性和结构化之间找到平衡,使数据的存储与使用过程尽可能简单。

与之相比,将材料基础数据在统一的公益性平台上集中产生,可以极大地简化由各家格式不统一带来的麻烦。与其他国家相比,中国有可能建立集中的、系统的、为社会提供基础数据的"数据工厂"。这也为中国在材料领域带来机遇。

4　数据驱动模式是未来材料科学的趋势

与科学"第四范式"相对应,材料基因组工程以前所未有的大量数据为基础,将人工智能与高通量实验数据采集和高通量计算深度融合,更快、更准地获得成分-结构-工艺-性能间的关系,从而实现对先进材料及工艺进行设计预测。因此,以数据为基础是材料基因工程方法与传统方法的根本不同点。高通量是数据时代的需求,数据采集技术是技术革命要素,而数据分析技术则是思维模式的变革,带来更加深刻、更加久远的变化。可以预见,材料科学的未来将构筑于数据与人工智能的基础之上。

人工智能在材料中的应用正在成为大数据经济的下一个战场。事实上,2017年12月,国际领先的人工智能企业DeepMind(AlphaGo和AlphaGo Zero的开发者)的联合创立人Hassabis表示已将下一个挑战目标放在了材料科学问题上[40]。2018年4月19日,美国Citrine Informatics公司宣布腾讯和奥地利私募股权公司B&C工业控股联合向他们投资8百万美元用于发展材料人工智能,则是这个趋势的最新明证[41]。

5　结论

在新型材料研发模式下,可以大致总结出材料基因工程的3种工作模式,即实验驱动、计算驱动和数据驱动。以"数据+人工智能"为标志的数据驱动模式围绕数据产生与数据处理展开,代表了材料基因工程的核心理念与发展方向。实现材料研发由"试错法"向"数据+人工智能"科学"第四范式"的根本转变,将更快、更准、更省地获得成分-结构-工艺-性能间的关系。目前材料数据的数量还远不能满足数据驱动模式的要求,因此,建设快速获取大量材料数据的能力是关键,基于高通量实验与高通量计算技术的"数据工厂"是满足材料基因工程数据需求的重要平台。在此框架下,材料基因工程的3个技术要素缺一不可,实现了完美的协同。当前,人工智能在材料中的应用正在成为大数据经济的下一个战场。未来的材料科学将构筑于数据与人工智能的基础之上。应该抓住材料基因组计划历史契机,抢占技术创新高地和发展先机,实现材料领域的弯道超车,摆脱中国战略性关键材料受制于人的窘境。

参考文献(References)

[1] 汪洪, 向勇, 项晓东, 等. 材料基因组——材料研发新模式[J]. 科技导报, 2015, 33(10): 13-19.
Wang Hong, Xiang Yong, Xiang Xiaodong, et al. Materials genome enables research and development revolution[J]. Science & Technology Review, 2015, 33(10): 13-19.

[2] National Science and Technology Council. Materials genome initiative for global competitiveness[R/OL]. [2018-03-31]. https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf.

[3] National Science and Technology Council. MGI strategic plan [R/OL]. [2018-03-31]. https://www.mgi.gov/sites/default/files/documents/mgi_strategic_plan_-_dec_2014.pdf.

[4] 刘梓葵. 关于材料基因组的基本观点及展望[J]. 科学通报, 2013, 58(35): 3618-3622.
Liu Zikui. Perspective on materials genome[J]. Chinese Science Bulletin, 2014, 58(35): 3618-3622.

[5] Green E D, Watson J D, Collins F S. Human Genome Project: Twenty-five years of big biology[J]. Nature, 2015, 526(7571): 29-31.

[6] Xiang X D, Sun X, Briceño G, et al. A combinatorial approach to materials discovery[J]. Science, 1995, 268(5218): 1738-1740.

[7] Xing H, Zhao B B, Wang Y J, et al. Rapid construction of Fe-Co-Ni composition-phase map by combinatorial materials chip approach[J]. ACS Combinatorial Science, 2018, 20: 127-131.

[8] Buitrago S A, Regalado E L, Pereira T, et al. Organic chemistry, Nanomole-scale high-throughput chemistry for the synthesis of complex molecules[J]. Science, 2015, 347(6217): 49-53.

[9] Perera D, Tucker J W, Brahmbhatt S, et al. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow[J]. Science, 2018, 359(6374): 429-434.

[10] Gesmundo N J, Sauvagnat B, Curran P J, et al. Cernak, nanoscale synthesis and affinity ranking[J]. Nature, 2018, 557 (7704): 228-232.

[11] Ceder G, Persson K. The stuff of dreams[J]. Scientific American, 2013, 6: 36-40.

[12] Allison J, Li M, Wolverton C, Su X M. Virtual aluminum castings: An industrial application of ICME[J]. JOM, 58(11): 28-35.

[13] Olson G B, Kuehmann C J. Materials genomics: From CALPHAD to flight[J]. Scripta Materialia, 2014, 70(1): 25-30.

[14] Mitchell T. Machine learning and data mining[J]. Communications of the ACM, 1999, 42(11): 30-36.

[15] Liu Y, Zhao T L, Ju W W, et al. Materials discovery and design using machine learining[J]. Journal of Materiomics, 2017, 3(3): 159-177.

[16] Raccuglia P, Elbert K C, Adler P D F, et al. Machine-learning-assisted materials discovery using failed experiments[J]. Nature, 2013, 533(7601): 73-76.

[17] Xue D, Balachandran P V, Hogden J, et al. Accelerated search for materials with targeted properties by adaptive design[J]. Nature Communications, 2016, 7: 11241.

[18] Xue D, Balachandran P V, Yuan R, et al. Accelerated search for $BaTiO_3$-based piezoelectrics with vertical morphotropic phase boundary using Bayesian learning[J]. Proceedings of the National Academy of Sciences, 2016, 113(47): 13301-13306.

[19] Xue D, Yuan R, Zhou Y, et al. An informatics approach to transformation temperatures of NiTi-based shape memory alloys[J]. Acta Materialia, 2017, 125: 532-541.

[20] Ren F, Ward L, Williams T, et al. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments[J]. Science Advances, 2018, 4: eaaq1566.

[21] Segler M H S, Preuss M, Waller M P. Planning chemical syntheses with deep neural networks and symbolic AI[J]. Nature, 2018, 555(7698): 604.

[22] Mueller T, Kusne A G, Ramprasad R. Machine learning in materials science: Recent progress and emerging aplications [M]//ParrillKenny A L, Lipkowitz B. Reviews in Computational Chemistry. Hoboken: John Wiley & Sons, Inc, 2016: 186-273.

[23] Hey T, Tansley S, Tolle K. 第四范式: 数据密集型科学发现[M].潘教峰, 张晓林, 译. 北京: 科学出版社, 2012.
Hey T, Tansley S, Tolle K. The fourth paradigm: Data-intensive scientific discovery microsoft research[M]. Pan Jiaofeng, Zhang Xiaolin, trans. Beijing: Science Press, 2012.

[24] Bell G, Hey T, Alex S. Beyond the data deluge[J]. Science, 2009, 323(5919): 1297-1298.

[25] The large synoptic survey telescope[EB/OL]. [2018-05-01]. https://www.lsst.org/lsst.

[26] 郭守敬望远镜(LAMOST,大天区面积多目标光纤光谱天文望远镜)[EB/OL]. [2013-04-26]. http://www.lamost.org/public/survey.
Guo Shoujing Telescope (LAMOST, large sky area multitarget fiber spectral astronomical telescope) [2013-04-26]. http://www.lamost.org/public/survey.

[27] GenBank[EB/OL]. [2018-03-31]. https://en.wikipedia.org/wiki/GenBank.

[28] Materials data facility[EB/OL]. [2018-03-31]. https://www.materialsdatafacility.org.

[29] Pauling file[EB/OL]. [2018-03-31]http://paulingfile.com.

[30] Granta design[EB/OL]. [2018-03-31] http://www.grantadesign.com.

[31] NIMS Materials Database (MatNavi) [EB/OL]. [2014-12-11]. http://mits.nims.go.jp/index_en.html.
[32] 国家材料环境腐蚀平台数据共享服务[EB/OL]. [2018-03-31]. http://www.ecorr.org/pingtai/.
[33] 国家材料科学数据共享网[EB/OL]. [2018-03-31] http://matsec.ustb.edu.cn/index.jsp.
[34] Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR guiding principles for scientific data management and stewardship[J]. Scientific Data, 2016(3): 167-172.
[35] Building a materials data infrastructure: Opening new pathways to discovery and innovation in science and engineering [EB/OL]. [2018-03-31]. http://www.tms.org/Publications/Studies/Materials_Data_Infrastructure/Materials_Data_Infrastructure.aspx?hkey=d228f86c-e269-49a2-a638-395285b760e4.
[36] 中华人民共和国国家质量监督检查检疫总局, 中国国家标准化管理委员会. GB/T 32843—2016 科技资源标识[S]. 北京：中国标准出版社, 2016.
General Administration of Quality Supervision, Inspection and Quaranitne of People's Republic of China, China National Standardization Management Committee. GB/T 32843—2016 Identification of scientific and technological resources[M]. Beijing: China Standard Press, 2016.
[37] Puchala B, Tarcea G, Marquis E A, et al. The materials commons: A collaboration platform and information repository for the global materials community[J]. JOM, 2016, 68(8): 2035-2044.
[38] Dima A, Bhaskarla S, Becker C, et al. Informatics infrastructure for the materials genome initiative[J]. JOM, 2016, 68(8): 2053-2064.
[39] Michel K, Meredig B. Beyond bulk single crystals: A data format for all materials structure-property-processing relationships[J]. MRS Bulletin, 2016, 41(8): 617-623.
[40] DeepMind wants to find the next miracle material—experts just don't know how they'll pull it off[EB/OL]. [2017-10-25]. https://qz.com/1110469/if-deepmind-is-going-to-find-the-next-miracle-material-experts-dont-know-how-theyll-pull-it-off.
[41] Citrine informatics Inc[EB/OL]. [2018-04-09]. https://citrine.io/news.

Data + AI: The core of materials genomic engineering

WANG Hong[1,2], XIANG Xiaodong[3], ZHANG Lanting[1,2]

1. Materials Genome Initiative Center, Shanghai Jiao Tong University, Shanghai 200240, China
2. School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
3. Department of Materials Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China

Abstract The working models of the Materials Genomic Engineering can be roughly classified into those of the experiment-driven, the computation-driven and the data-driven. The last kind of model is consistent with the fourth paradigm of scientific approach of a fundamental change from "trial and error" to "data-intensive". Such a paradigm shift allows one to acquire the composition-structure-process-performance relationship, as the basis for the rational design of materials, in a faster, cheaper and more accurate way. It represents the core concept and the future direction of the MGI. In this data-centric scientific era, the ability to quickly obtain a large amount of materials data becomes essential. Thus, the "data foundries"-the centralized materials data generation facilities based on high-throughput experiments and high-throughput computations are the key infrastructures for meeting the future data needs. It is contemplated that the data and the artificial intelligence will become the foundation for building the materials science of the future.

Keywords materials genome engineering; data-driven; high-throughput experiments; high-throughput computation

（责任编辑　刘志远）