中国科学: 化学 2018 年 第 48 卷 第 3 期: 243 ~ 255
SCIENTIA SINICA Chimica chemcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

评 述

材料基因组学的发展现状、研究思路与建议

李云琦[1*], 刘伦洋[1], 陈文多[1], 安立佳[2*]

1. 中国科学院长春应用化学研究所, 中国科学院合成橡胶重点实验室, 长春 130022
2. 中国科学院长春应用化学研究所, 高分子化学与物理国家重点实验室, 长春 130022
*通讯作者, E-mail: yunqi@ciac.ac.cn; ljan@ciac.ac.cn

收稿日期: 2017-11-02; 接受日期: 2017-12-06; 网络版发表日期: 2018-02-11
国家自然科学基金(编号: 21374117, 21774128)和中国科学院百人计划资助项目

摘要 回顾了材料基因组学的发展历程, 分析了材料基因组学在能源、气体分离、合金、催化和高分子等新材料开发中的成功应用案例、典型流程和理论基础等共性特征, 整理出了材料基因组学研究思路的要点和主要方法, 并对当前机遇下大力发展材料基因组学提出了建议.

关键词 材料基因组学, 新材料, 计算机辅助材料设计, 高通量筛选, 组成-工艺-结构-性能关系

1 引言

材料基因组学, 又称材料基因工程或材料信息学, 是材料与信息科学交叉的一门新学科. 为了改变像Edison寻找灯丝材料遍历方法的费时费力的局限性, 在丰富的科学数据和模型的驱动下, 利用计算机辅助设计加速材料创新的时代已经来临. 实现的主要途径包括: (1) 通过量子化学计算与统计力学方法发展; (2) 通过定量结构-性能关系(quantitative structure-property relationship, QSPR)的建立, 提高人们的知识储备和设计、制备新材料的智慧. 材料基因组学的早期应用(1995年前)主要包含组合理论和伴随发展的高通量筛选方法[1], 实现高通量实验向高通量知识发现的转化[2], 侧重于结构-性能关系的阐述. 近年来, 涵括了组成-工艺-结构-性能关系的定量化发展, 推动材料基因组学应用于新材料、新器件的开发更进一个层次[3]. 在此期间, 随着人类基因工程和生物信息学在世界范围的飞速发展, 材料基因组学在2006年前后逐步成熟[4], 并成为Obama执政时期重要的科学促进政绩之一. 材料基因组学要为认识材料性能的本质和促进新材料研发提供信息, 将材料基因组学与大数据联系起来, 通过数据在理论和机器学习的时空间外延, 实现材料的高效发现和组成工艺最优化, 大幅降低新材料开发成本, 是当前新材料发现的重要途径之一. 材料基因组学的核心是数据驱动创新(data-accelerated research paradigm), 类比于生物基因工程, 材料基因组学以材料的组成工艺参数为基因子, 参数与性能联系作为基因序列, 利用一级(组成/序列)、二级(分子结构、基团)、三级(相互作用、相区/折叠)和四级(结构形貌界面/聚集)多层次结构信息, 并结合工艺参数, 预测材料和器件的性质, 明确材料结构响应规律的本质.

材料基因组学有其独特的研究模式, 通用的是DIKW (Data→Information→Knowledge→Wisdom)层级模型[5], 表述为研究人员通过机器学习技术的挖掘

引用格式: Li Y, Liu L, Chen W, An L. Materials genome: research progress, challenges and outlook. *Sci Sin Chim*, 2018, 48: 243–255, doi: 10.1360/N032017-00182

© 2018 《中国科学》杂志社

数据, 获得有价值的信息, 在专家经验和理论指导下转化为可靠的知识并能够辅助决策的智慧, 而目前的重点是明确或预测给定体系的组成-工艺-结构-性能关系[6]. DIKW模型包括: 数据的采集、整理、分析和再传播. 利用的手段, 除了基于特定数据库的计算机建模、模拟和机器学习外, 还涵盖计算模拟与实验数据采集同步协同分析, 利于开展进一步的机器学习和大数据背景下的非显著参考体系分布建立. 在理论、高通量计算模拟与实验的协同框架下, 通过结构-性能定量的建立和验证, 获得知识和智慧, 一方面可以深化并超越所研究材料的物理本质, 另一方面可以进行预测, 获得新的预期. 高通量的计算机辅助材料设计强烈依赖于量子力学-统计热力学的进展, 以及数据库的构建和智能数据分析挖掘. 近年来的一些综述, 对材料基因组学的成功应用案例模式进行了深入分析[7,8]. 材料基因组学中的研究内容包含组成(ϕ: 含器件中材料最终驻留组成和加工制备过程试剂组成)、工艺(P)、结构(S: 含电子、原子、基团、聚集体、相区多尺度粒子分布以及场中密度分布; 分为本体结构和表界面结构; 也可分为平衡态结构、非平衡态结构和稳态结构下的物质和能量涨落、迁移和转化)、性质(E: 含材料和器件的结构对环境条件的连续或突变响应关系)以及性能(Y: 器件中外观或宏观结构在应用中对多个特定环境变化因素的协同响应关系), 其关系示意于图1(a). 结构是理论和计算模拟的研究核心, 宏观参数都可以通过微观的分布函数统计出来. 经典理论和模拟主要是统计热力学, 典型如Boltzmann分布, 而基于数据挖掘的机器学习却主要依据条件概率分布, 即Bayes分布. 在统计热力学下, 任意宏观量可写作$Y=\Sigma y_i\omega_i/\Sigma\omega_i$, 其中$y_i$和$\omega_i$是微观分量及其权重(简并度), 而在Bayes原理指导下, 任意性能指标总可以写为Y(S|ϕ|P; S|E), 其中"|"为条件概率符号. 该表达式包含材料基础研究的两个重要方面, 即特定组成工艺下的结构形成机制, 以及材料的结构-性质关系. 在Bayes原理, 即Y(A|B)=Y(A)Y(B|A)/Y(B)指导下, 通过概率置换得到新的分布, 进一步在统计热力学框架下可得到任意宏观量和性质性能参数. 在材料基因组学中, 围绕结构分布变化主要有如下几种模式(图1(b)): A→B, 同一分布中的某显著态; A→C, 同一分布下某显著态通过连续相变的富集; A→D, 最可几在同一分布下的某显著态富集, 局部分布发生突变; A→E, 整体发生相变, 微观和宏观结构存在显著差异. 由于材料基因组学可以在合理的计算消耗下, 实现大范围多分布的探索, 已经成为理论计算模拟不可或缺的重要发展方向之一. 在近十几年的研究中, 取得了令人瞩目的成果. 下面就材料基因组学在能源材料、气体分离材料、合金材料、催化材料、高分子材料和其他面向特定应用的材料等领域的一些典型应用进行介绍.

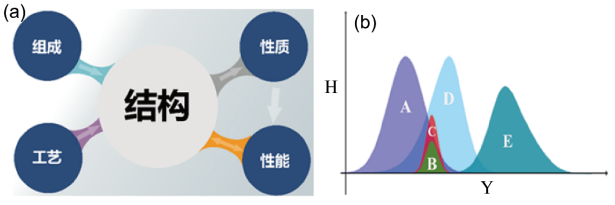


图 1 材料研究中的组成-工艺-结构-性质-性能关系(a), 以及经典统计热力学与Bayesian分布相结合的典型结构分布改变模式(b) (网络版彩图)

在能源材料研究中, Harvard大学洁净能源项目(Harvard Clean Energy Project, CEP)为制备出高性能的有机光伏材料, 通过神经网络学习了有限的实验数据, 并进行了拓展搜索和高通量筛选, 开发出了精度可与量子化学计算相媲美的大数据手段. 通过遍历两类分子组合, 26种分子片段共350万种分子组合, 利用条件概率指引需求目标分子, 有效地缩小了化学空间搜索, 加速了多来源多种材料的筛选, 降低了材料设计中所需的计算资源、合成与表征[9~11]. 通过模式识别和机器学习, 优化CEP项目中光伏高分子给体材料, 对最高已占轨道(HOMO)、最低未占轨道(LUMO)和能量转换效率(PCE)建立了可靠模型, 发现含Benzothiadiazole和Thienopyrrole基团的高分子具有高的光电转换效率, 并得到了实验验证[12]. 除了光电转换材料, 利用材料基因组学方法还发现了高性能的电极材料. 例如, Cheng等[13]通过发展高通量的量子化学计算, 发掘出可替代锂离子电池的新型电池材料, 通过LUMO、HOMO、氧化还原势、溶解能和结构稳定性的计算预测, 从1400余种有机材料中遴选出了200余种候选材料, 构建了氧化还原电位与LUMO和HOMO的线性相关模型, 进而确定了新型电池的电极材料. Ryder等[14]利用原子探针与密度泛函理论(DFT)计算黑磷材料, 通过最优化涂层插嵌材料的稳定性和荷电性, 开发出了具有优异性能的表面修饰光反应器件. Sevov等[15]利用DFT能带计算, 结合结构和电解能谱, 分析

了吡啶及不同衍生物阳极液, 通过预测模型的拓展, 找到了单电子液流电池的优化组成. 在材料基因组学辅助能源材料的研究中, 高通量的DFT计算是材料筛选和设计的关键, 功能基团能带的近程和长程耦合(如聚集诱导发光[16])是材料性能的重要调节因素. 目前, 能源材料中的物质输运和能量转换效率的上限仍被不断刷新, 材料基因组学的重要性逐步由计算机辅助设计向计算机主导的材料设计转化.

在气体分离材料的开发中, 金属骨架(MOF)和碳骨架(COF)材料是研究的前沿. 对于甲烷气体分离储存材料, Snurr等[17]整理了102个已知MOF构成的金属中心和有机连接模块, 将这些模块遍历组合, 分析了孔径分布、比表面积和甲烷储存能力, 找出了300余种MOF组合, 其性能超过目前已合成出的材料, 进一步通过预测的结构性能关系, 发现甲基化能有效提升MOF材料性能, 并获得了实验验证. 进一步, 他们利用DFT模拟了不同组成基团的气体吸附行为, 通过65万多种虚拟MOF材料基因组学研究, 建立了小规模数据无法观察到的结构-性能关系. 在此指导下, 制备出了高性能的CO_2、H_2和CH_4等的捕获分离材料[18~20]. 此外, 他们还对应用于气体储存、分离、传感、药物筛选、光捕获和催化等特定的MOF材料进行了研究, 利用Monte Carlo模拟对分子的组合进行了遍历, 对MOF材料的比表面积、孔径和气体吸附能力, 建立了实验与计算模拟相结合的可靠统计学预测模型[21]. Haranczyk等[22]利用巨正则系综Monte Carlo模拟, 计算分析了沸石及其配体的孔穴结构, 筛选出CO_2捕获的高性能材料. 进一步在优化电子结构基础上, 利用巨正则Monte Carlo模拟计算筛选了18000种多孔网络高分子材料的甲烷吸收和脱附性质, 明晰了结构-性能关系, 以及孔尺寸是吸附性能的主要决定因素, 指导实验制备出了性能优异的CH_4吸附储存高分子材料, 实现了A→D分布类型的拓展[23]. 小分子分离材料设计的关键是合理权衡选择性与渗透性, 寻求处于分离上限(upper boundary)的物质[24]以及特殊的结构和形貌, 材料基因组学正是瞄准了这一要点, 围绕结构和性质的显著性分布指导新材料开发, 不断刷新这类材料的综合性能上限.

合金材料的材料基因组学研究将硬度和韧性这两个很难统一的特性, 在同一材料上得以实现, 让新材料“无懈可击”. 材料基因组学可用于合金材料的定性分类. 例如, 通过对80余种二元合金原子轨道半径与光谱性质的耦合和DFT计算, 更新验证了3种传统经验认为的脆性合金实为柔性合金[25]. 材料基因组学也可用于合金材料的组成相图的准确预测, Xiong等[26]通过组成-工艺-结构关系, 明确体系特定结构的形成机制, 利用X射线衍射图案的特征分类并开展了图像像素化和Delauney马赛克化分析, 针对三元合金的形貌, 建立了高通量结构的快速分类方法, 实现了对相图的可靠预测. 也可以利用材料基因组学方法, 通过大数据挖掘, 分析元素的丰度产量、化合物的电阻等与热电材料性能关系的关联, 明确热电转换材料的物理本质及其与组成元素的决定性作用, 制备出高性能热电材料并替代了稀有金属的使用[27]. 此外, 利用材料的结构和电子分布特征的统计分布, 分析了超导体的临界温度, 建立了大数据辅助材料设计的分析、可视化、QSPR预测模型和筛选, 可以实现A→E模式的转换, 得到全新的材料组合. 对二元、三元和多元超导材料的性能关系与能带联系进行了深入剖析, 根据材料组成与超导临界点的相似度和关联度, 明晰了材料结构与性能联系的物理本质[28].

在催化材料的研发中, 非贵重金属替代一直处于研究前沿. 科研人员通过荧光强度与电动化学势的关联, 构建了二元和三元氧化物催化剂, 通过预测模型的建立, 用计算筛选出了CuNiMn氧化物并得到实验验证, 其具有与Pt/C催化剂相近的催化活性[29]. Zhan等[30]利用组合化学, 结合电镜, 研究了TiO_2基底上不同组成、工艺的量子点材料对光吸附和催化反应效率的影响, 通过理论与实验结合的高通量模型的建立和验证, 筛选出了界面优化且光电转换效率高达5.33%的两组分量子点材料. Le和Winkler[31]对基因遗传算法(GA)用于开发气体分子转化催化剂进行了综述, 表明GA能够指导筛选出性能优异的多元催化材料. 同时, 利用机器学习辅助合成路线设计, 预测主要产物, 可以帮助选择合适的催化剂, 提高特定目标产物的成功合成概率[32]. 可以看到, 在亲核与亲电反应、异构反应、光催化反应以及增强反应物的吸附与生成物的脱附、新型高效率催化材料的设计和制备中, 材料基因组学手段逐步得到广泛应用.

高分子材料的材料基因组学应用相对较少, 这与体系的计算复杂度高、材料性质严重依赖于加工工艺且工艺参数难以统一有关. Schubert等[33]研究了甲基丙

烯酸聚合物结构单元上的3种不同取代基团对控制输运材料的极性和pH响应性及其对β-萘酚橙的释放动力学, 通过材料基因组学寻求最佳吞服载体, 使药物分子在胃和小肠中有最佳的释放率. Sharma等[34]提出了含高通量DFT、力场模拟和微扰杂化能带理论, 结合样品制备、红外光谱、X射线衍射和介电谱等多层次建模研究基础上的材料选择途径, 找到了几种令人振奋的全新有机高分子介电材料(A→E模式). 利用材料基因组学构建和应用定量的结构性能关系, 围绕组成的原子、分子和聚集体信息与其分散、界面能量及响应关系, 结合粒子分散性与界面能的预测模型, 在纳米二氧化硅与聚苯乙烯(PS)、聚甲基丙烯酸甲酯(PMMA)、聚乙基丙烯酸甲酯(PEMA)和聚2-乙烯基吡啶(P2VP)等复合物中, 从微观组成结构, 到介观形貌和宏观性质得到了验证, 预测了近平衡态下多种复合物的性质[35]. von Lilienfeld等[36]利用大数据拓展了量子化学计算对大分子的限制(特别是A→D模式的拓展), 在较低的计算消耗下, 对分子的能量进行了准确预测, 其准确度与通过冗繁计算的DFT相当, 有力地推动了量子化学计算在大数据挖掘中的应用. 本课题组[37]也通过对高分子膜材料的数据挖掘和统计学分析, 阐明了质子交换膜的组成-工艺-结构-性质关系, 为开发高性能的磺酸基质子交换膜提供了切实指导.

在其他面向特定应用领域, 为了适应材料基因组学对高通量计算的需求, 机器学习与量子计算结合, 建立准确可靠的力场参数, 拓展第一性原理计算用于更长时间、更大空间的模拟方法得到了持续关注和发展[38,39]. Oliynyk等[40]使用机器学习方法, 训练了随机森林模型, 通过对具有AB_2C通式的复合物进行筛选, 预测出12种具有热电和自旋特性的、新的Heusler物质并得到实验证实. Wang等[41]通过多元线性回归(MLR)和支持向量机(SVM)分析, 基于61种化学物质的实验最小点火能量(MIE)数据和分子模拟, 构建了QSPR模型, 可以帮助理解分子结构对碳氢燃料燃烧特性的影响. Kuenemann等[42]从胺的分子结构出发, 开发了基于随机森林和神经网络的机器学习模型, 用于预测胺的二氧化碳吸收性能, 被用来虚拟筛选优化新的胺分子, 用于二氧化碳捕获和利用. Liu等[43]通过材料基因组学方法, 优化组成材料的结构熵, 开发了一系列性能优异的热电转换材料. Ren等[44]通过硅基嵌段共聚物直接自组装结构的透射电镜和X射线散射的背刻蚀高通量筛选, 为制备特定图案的光刻材料提供了全面快速的参考方案.

利用材料基因组学取得的代表性成果不限于上述列举, 为了帮助大家明晰它的特色, 实现有效利用并取得更好的研究结果, 下面我们将其研究思路及建议介绍如下.

2 研究思路

材料基因组学的核心是: 通过结构建立组成、工艺与性质的定量联系, 该定量联系可以是线性或非线性的、显式或隐式表达的. 在可靠数据集整理的基础上, 性能指标预测模型的成功主要依赖于两方面: 一是描述符集合的建立; 二是多参数定量关系函数(预测模型)的建立. 材料基因组学方法包含4个方面: 数据集建立、描述符集合建立、预测模型建立和高效能模型的可靠性检验与应用[8](图2). 利用该思路研究高性能MOF材料的一个具体实例流程见图3[45]. 下文针对这4个方面展开说明.

2.1 数据集整理

2.1.1 数据集的收集整理

一般地, 在材料基因组学研究项目中, 数据整理占据整个过程60%~80%的时间, 数据的完备性(comprehensive)和代表性(representative)直接决定最终预测模型的可靠性. 由于数据源存在系统误差、噪声、不确

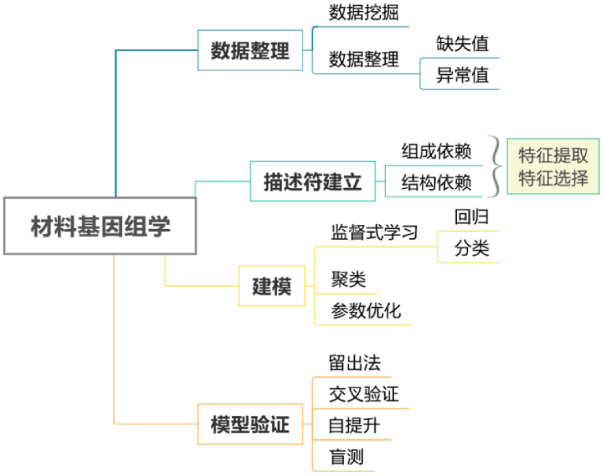


图 2 材料基因组学的主要研究内容和思路(网络版彩图)

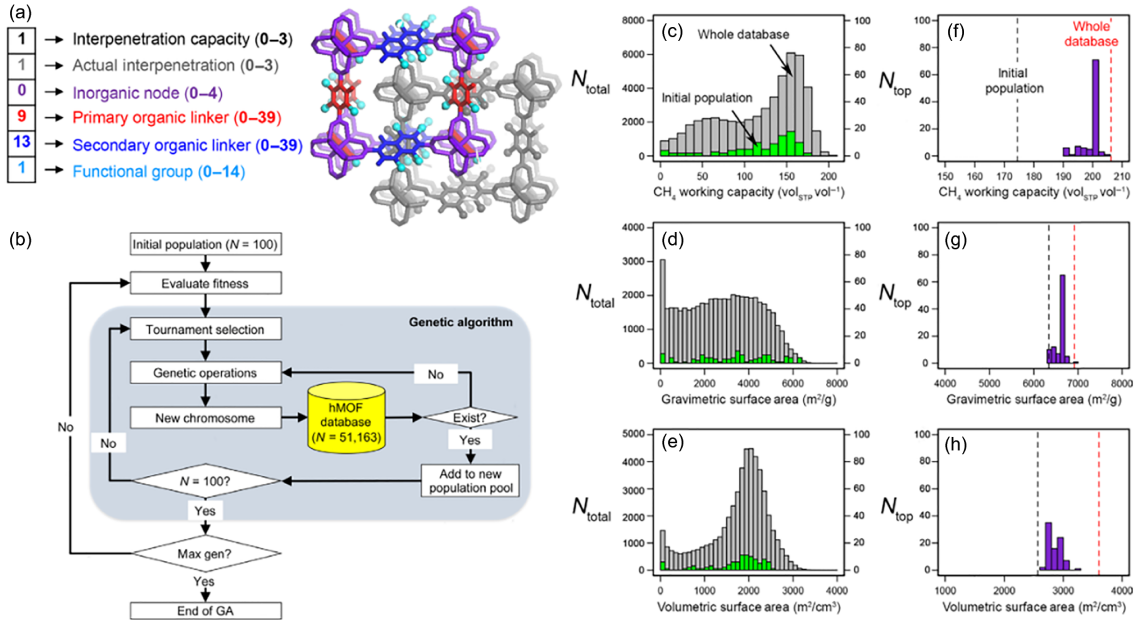


图 3　利用材料基因组学方法中的基因遗传算法开发用于捕获二氧化碳的MOF材料研究思路[45]. (a) 一个特定MOF结构的标准描述; (b) 基因遗传算法流程图; (c~h) MOF材料特征性质参数在优化前后的分布(网络版彩图)

定性和多分散性等特点，需要用数据挖掘的智慧来分析处理, 建立可靠的数据驱动创新模型[46].

数据挖掘常被转换为表格形式，数据储存为以行和列分布的电子表格，或者以每行为特征矢量的矩阵，每行对应于所收集样本的实例，每列代表一个特征(变量: 预测因子). 当然，也存在一些特殊情形(如长字符串型)，但大多数机器学习算法需要结构化的数据. 对于表格或矩阵数据，我们需要注意到各行可能包括各种类型(如字符型、数字型和逻辑型)的数据，各列包括相同类型的数据，但是，实际上数据可能以各种其他“凌乱”的方式(如缺失值、通配值)存储. 例如，对于数据收集阶段, 假设某个量不可测, 且无法找到值域区间，那么，数据表将包含一个或多个单元格中的缺失值，这样，可能会使模型构建和后续使用模型做预测时更为困难. 此时，在某些情况下(如数据记录等重复任务中)，人工参与数据收集就容易犯错误，可能导致一些数据出现异常. 因此, 在数据结构化的基础上，需要明确这些异常值可能导致预测模型的不确定性.

材料研究领域数据存储的多分散性来源于数据测试流程有不同标准，分散的文献、专利和专著的侧重点也不一致. 虽然材料研究者长期以来在生产与材料的加工-结构-性能关系相关方面积累了相当丰富的知识与数据, 但是, 这些信息是以各种杂乱方式生成与存储和分析与传播的, 如材料计算者使用不同的软件, 材料实验者使用不同的表征方法和标准. 这些直接导致了研究者在重新利用这些数据时面临着巨大的困难，使得材料数据的结构化和标准化变得更具挑战性. 因此，为方便机器学习，建立可靠的材料基因组学预测模型，我们必须对数据进行必要的“清洗”和处理. 这个过程可以分为3个步骤，即数据选择、数据预处理和数据转换.

第一步: 数据选择. 我们收集数据时总是强烈希望获取所有可用的数据，使数据集合尽量靠近完备性和代表性. 具体的数据要结合实际面临的问题来挑选，并且在获取数据时可以对这些数据进行一些假设，以便以后进行统计检验.

第二步: 数据预处理. 在数据使用之前，需要将数据加工处理成方便机器学习建模的数据格式. 该步包含3个常见的预处理方案: (1) 格式化. 被选择数据的存储格式可能不适用于建模，如数据可能存储于网页、图片或者其他非传统型的数据库中，这就需要将其格式转化为易于建模的表格型数据格式. (2) 清洗. 清洗数据就是移除或者修正缺失数据、异常值等. 在数据集中，可能存在包含缺失值的样本或者某个样本

的某个属性值过于突出异常, 这时可以考虑将其移除. 另外, 在一些属性中, 可能存在对结构突变的奇点值, 这些属性可以通过合理的连续化处理或删除. (3) 采样. 当数据点在某些搜索空间过于冗余, 造成运算代价过大时, 可以采用归类集成, 抽提出代表性样本, 以利于深层次建模.

第三步: 数据转换, 即特征工程. 这是依据所选用算法和模型的具体应用, 通过特征定标(scaling)、特征分解(decomposition)和特征合并(aggregation), 提高材料基因组学模型的可靠性和可应用性. 特征定标常包含量纲归一化和值域归一化; 特征分解常需剔除基准值, 提高变化依赖值的响应性和单调性; 特征集成则刚好相反, 通过特征值的正交组合, 构建新的特征值, 使其在性能指标上具有更大的投影分量.

2.1.2 材料学物性数据库

在材料研究领域, 已经出现了一批较为成熟且实用的物性数据库, 可以分为: 实验数据库和计算模拟数据库[47]. 实验数据部分来自晶体结构数据库, 并分散在多个数据库或材料物性手册中. 这些不同的数据源经历版本更新迭代, 精心整理和验证, 已经成为材料数据集的重要可靠资源. 在这些数据库基础上, 进行数据挖掘是有可能的, 但由于数据库的限制(如访问权限、程序抓取数据的限制和数据自身不完整性等), 数据挖掘仍然面临重重阻碍. 就完整性而言, 许多材料的性质(如形成能、带隙和弹性张量等)仅仅是已知晶体结构中极小的一部分. 化合物性质通常与物质组成和结构紧密关联, 但是, 对这些材料的结构和组成(如晶体结构、微结构和掺杂水平等)却缺乏详细描述, 对于开发可靠的预测模型非常具有挑战性. 最后, 在数据访问方面, 大多数数据库只能通过设计用于“单次查询”的机制进行访问, 而不能通过程序批量访问数据, 这大大降低了大规模数据挖掘效率. 在材料的计算模拟数据库方面, 通常经过实验数据库的晶体结构信息验证, 使用系统化的高通量计算技术(最典型的是基于DFT的方法)生成材料数据的能力已经为数据挖掘构建了不少高质量的数据库. 一些数据库已初具规模, 如Materials Project至今已包含6万多种化合物的物性数据, AFLOWlib拥有超过60条材料性质. 常见材料物性数据库及相关信息列于表1. 不幸的是, 目前还没有综合的搜索引擎或类似的工具来进行跨数据库的检索.

2.2 描述符集合的建立

为了使不同方面的描述符最终能够收敛地表达性质参数, 描述符被分为多个子集, 常见的有: (1) 本征参数(constitutional) (分子量及其分布、原子数、功能基团数和芳环数等); (2) 拓扑几何结构(topological & geometric) (主惯量、分子/单体体积、排除体积、Weiner指数、Randic指数、Kier & Hall分子连接性指数和Shannon信息熵等); (3) 表面和电荷(surface &electrostatic) (溶剂可接触面积、本体电荷、内聚能密度、部分电荷分布、等电点、极性和表面电荷密度等); (4) 量子化学参数(quantum chemical) (偶极矩、σ和π键序列、HOMO与LUMO能带和前线轨道(FMO) (5) 热力学参数(thermodynamic) (玻璃化温度、熔融结晶温度和热膨胀指数等); (6) 溶度参数(solvation) (色散力、氢键给受体和极性指数等).

这些描述符可以基于二维(2D)化学结构(组成依赖)或三维(3D)结构进行计算得出, 基于3D结构的计算往往涉及电子和原子结构的优化, 需要使用量子化学计算、全原子或粗粒化的分子力场, 利用场论、分子动力学或Monte Carlo模拟实现. 在优化的3D结构基础上, 对给定的体系, 可以计算出上千种描述符. 依据描述符所处维度的不同, 可以划分为不同的类别(表2).

由于这些描述符并不是严格正交的, 为了加快建立定量关系的统计学模型和尽可能简化关联描述符, 需要对描述符集合利用统计学手段进行整理. 例如, 为了减小不同描述符数值在量级上差异引起的误差, 常对训练集的单一描述符(X_i)进行约化处理, 具体是:

$$X'_i = \frac{X_i - \langle X_i \rangle}{\mathrm{SD}(X_i)} \tag{1}$$

其中< >代表平均值, SD()代表方差函数.

2.3 预测模型建立

由于描述符的数量可能远多于训练集中的样本数, 因此, 需要对描述符集合利用统计学手段进行整理, 包含训练集中描述符对性能参数的回归、排序、主成分分析(PCA)和工作特征曲线(ROC)分析等. 依据关联度排序, 排除部分信息含量较低的描述符. 在降维后的数据矩阵中, 计算关联矩阵, 排除部分高度冗余的描述符矢量. 此时, 描述符集合可用于统计预测模型QSPR的建立中. 在QSPR预测模型建立中, 选择合适

表 1　常见材料物性数据库及相关信息列表

数据库名称	网址	类别
AFLOWLIB	http://www.aflowlib.org	计算
AIST Research Information Databases	http://www.aist.go.jp/aist_e/list/database/riodb	材料表征
American Mineralogist Crystal Structure Database	http://rruff.geo.arizona.edu/AMS/amcsd.phP	矿物质
ASM Alloy Database	https://www.asminternational.org/materials-resources/online-databases	合金材料
CALPHAD	http://www.opencalphad.com/index.html	热力学
Cambridge Crystallographic Data Centre	http://www.ccdc.cam.ac.uk	晶体学
SUNCAT	http://suncat.stanford.edu/theory/it-facilities	催化剂
ChemSpider	http://www.chemspider.com	化学结构
CINDAS Alloys Database	http://www.cindasdata.com/products/hpad	合金材料
Citrination	http://www.citrination.com	材料数据
CRC Handbook	http://www.hbcpnetbase.com	材料数据
CrystWorks	https://cds.dl.ac.uk	晶体数据
Crystallography Open Database	http://www.crystallography.net	晶体数据
Granta CES Selector	http://www.grantadesign.com/products/ces	材料数据
Handbook of Optical Constants of Solids	N/A	书籍
Harvard Clean Energy Project	http: //cepdb.molecularspace.org	计算
Inorganic Crystal Structure Database	http: //cds.dl.ac.uk/cds/datasets/crys/icsd/llicsd.html	晶体学
International Glass Database System	http://www.newglass.jp/interglad_n/gaiyo/info_e.html	玻璃
Knovel	http://app.knovel.com/web/browse.v	材料工程
Matbase	http://www.matbase.com	材料数据
Materials Project	http://www.materialsproject.org	计算
MatNavi (NIMS)	http://mits.nims.go.jp/index_en.html	材料数据
MatWeb	http://www.matweb.com	材料数据
Mindat	http://www.mindat.org	矿物质
NanoHUB	http://www.nanohub.org	纳米材料
Nanomaterials Registry	http://www.nanomaterialregistry.org	纳米材料
NIST Materials Data Repository	https://materialsdata.nist.gov	材料数据
NIST Standard Reference Data	http://www.nist.gov/srd/dblistpcdatabases.cfm	标准数据
Open KIM	http://www.openkim.org	材料模拟
Open Quantum Materials Database	http://www.oqmd.org	计算
Pauling File	http://www.paulingfile.com	材料数据
Pearson's Handbook: Crystallographic Data	N/A	书籍
Powder Diffraction File (PDF)	http://www.icdd.com/products/index.htm	晶体学
Reaxys	http://www.elsevier.com/solutions/reaxys	化学数据
Scifinder/ChemAbstracts	http://scifinder.cas.org	化学数据
Springer Materials	http://materials.springer.com	材料数据
Metallurgical Thermochemistry	N/A	书籍
TE Design Lab	http://www.tedesignlab.org	热力学
Total Materia	http://www.totalmateria.com	材料数据
UCSB-MRL Thermoelectric Database	http://www.mrl.ucsb.edu:8080/datamine/thermoelectric.jsp	热电材料

表 2 常见描述符维度分类

维度	描述	例子
零维	原子数, 键数, 分子量	C和H原子数量, 杂原子/非氢原子数量, 分子量
一维	分子片段数	氢键给体原子数量, 氢键受体原子数量, 分子片段极性表面积
二维	拓扑指数	Zagreb指数, Wiener指数, Balaban指数, 连接指数chi(χ), kappa(κ)形状指数
三维	几何描述符	回转半径, E-State拓扑参数, 3D Wiener指数, 3D Balaban指数, 径向分布函数

的机器学习算法也是非常关键的一步，它会极大影响模型预测的准确性和可靠性. 每个算法都有其适用范围，目前还没有一个算法适合解决所有问题. 特别是，概率估计算法主要用于新材料的发现，回归、集成分类则通常用于材料宏观和微观性质的预测. 另外，机器学习还常常搭配智能优化算法(如基因遗传算法GA、粒子群优化算法SAA等)进行模型的参数优化.

2.3.1 特征选择

特征选择就是从特征空间中选择相关的特征(描述符和预测因子)子集用于建模, 其被广泛使用主要是为了: 避免过拟合, 提升模型性能; 加速模型训练, 提升模型性价比; 简化预测模型, 使其更具可解释性. 使用特征选择技术的核心前提是: 数据集中包含冗余或者不相关(信息含量很低)的特征，移除这些特征并不会过多损失数据信息. 冗余和不相关是两个完全不同的概念，冗余是指多个特征在数据集中强相关时，除一个特征外，其他特征都是冗余的，而不相关是指特定描述符对预测因子几乎正交.

特征选择是一种组合搜索技术，用于提取新的特征子集并进行评估. 最简单的算法是遍历测试，找到最小化错误率的特征子集. 这是一种贪婪搜索算法，当特征空间较大时，会直接导致消耗大量的计算资源. 特征选择与机器学习存在紧密联系，根据特征选择中子集评价标准和后续学习算法的结合方式，可分为过滤式(filter)、封装式(wrapper)和嵌入式(embedded) 3种.

(1) 过滤式特征选择: 特征的评价标准从数据集本身获得，与特定的学习算法无关，具有较好的通用性. 通常选择与类别相关度大的特征或者特征子集，因为相关度较大的特征或者特征子集可以在分类器上获得较高的准确率. 过滤式特征选择的评价标准分为4种，即距离、信息、关联性和一致性.

(2) 封装式特征选择: 利用学习算法的性能来评价特征, 对于一个待评价的特征子集, 封装式方法需要训练一个分类器，根据分类器的打分对该特征子集进行评价. 封装式方法中用以评价特征的学习算法包括神经网络、Bayes分类器、近邻法和支持向量机等.

(3) 嵌入式特征选择: 特征选择算法本身作为组成部分嵌入到学习算法. 最典型的就是决策树算法，如ID3、C4.5和CART算法等，在树增长过程的每个递归步都必须选择一个特征，将样本集划分成较小的子集, 选择特征的依据通常是划分后子节点的纯度，划分后子节点越纯，则说明划分效果越好，可见决策树生成的过程也就是特征选择的过程.

2.3.2 建模

发现性能优异的新材料是材料基因组学的最终目的，其中机器学习能够处理发散分布(如A→E模式)和高维相关下的快速收敛. 目前，实验或者计算机筛选新材料主要是对分子进行元素替换和结构变换. 然而，材料组成搜索空间、结构搜索空间或两者都倾向于受到限制[33]. 两种筛选方法都可能需要大量的计算或实验，并且通常会导致在“穷举搜索”中以不正确的方向进行前进，这会消耗相当多的时间和资源. 结合这一事实和机器学习的优势，进一步发展材料基因组学，开发将机器学习与计算机模拟结合在一起的自适应性方法, 用于“虚拟”筛选新材料, 以期探索到某些材料性质存在显著分布(4种类型如图1(b)所示)，为发现新的和更好的材料提供指导.

材料性质(如硬度、熔点、离子电导率、玻璃化转变温度、分子雾化能和晶格常数等)均可以在宏观或微观层面描述. 也存在一些材料性质, 其在多个尺度上体现为不同组分间的协同作用，如膜材料的分离和传质性质、高分子材料的相转变导致的机械和热与光电性质、多相催化性质等. 材料性质参数的获取可通

过计算模拟和实验测量, 在一系列的标准模型和定义域限制下, 要统一众多的材料本征参数与特定结构的单一或协同参数间的关系, 这些关系可能是线性或非线性的, 有时甚至不存在显式表达, 模型的构建无疑是极具挑战的. 因此, 通过计算机模拟, 探索在有限空间去完全捕捉材料属性与其影响因素之间的关系就显得捉襟见肘. 此外, 用于表征材料性质的实验通常发生在材料研发的后期阶段, 如果结果不尽如人意, 投入的大量时间和实验资源将被浪费. 实际上, 在多数情况下, 即使耗费大量的计算或实验资源, 也难以完全达到预期. 此时, 迫切需要开发能够以较低时间和计算成本, 正确预测材料性质的高性能预测模型. 机器学习算法可以帮助我们从已有数据中学习模式和发掘新知识, 用以辅助新材料的设计和开发. 常见的用于材料发现和材料性质预测的机器学习算法包括: 部分线性回归(PLS)[48]、多元线性回归(MLR)[49~51]、随机森林(RF)[40,52]、支持向量机(SVM)[52]、神经网络(NN)[42,53]和基因遗传算法(GA)[45]等.

2.4 预测模型可靠性检验及应用

当前已知的材料基因组学预测模型几乎都着力于QSPR的建立, 经典软件有CODESSA[54], 近年来开源的R-Project结合Python、Java的开发非常活跃(详情可参阅Pirhadi等[55]的总结). 可靠的预测模型除了在训练集有良好表现外, 还需对未知数据集(含标准测试集和盲测集)提供准确预测. 一般情况下, 可以根据测试集预测模型的泛化误差来对可靠性进行评估. 目前有3种广为应用的方法用于模型评估, 即留出法(hold out)、交叉验证法(cross validation)和自提升法(boot strapping).

(1) 留出法: 将数据集D划分成训练集S和测试集T, S与T互补且正交. 用S训练出预测模型, 用T来评估预测误差, 作为对泛化误差的近似估计.

(2) 交叉验证法: 将数据集D划分为k个大小相似的互斥子集, 即$D=D_1\cup D_2\cup\cdots\cup D_k$, 各子集相互正交且尽可能保持数据分布的一致性. 训练时, 每次用$k-1$个子集的并集作为训练集, 余下的一个子集作为测试集; 如此, 可获得k组训练集和测试集, 从而进行k次训练和测试, 最终返回k次测试结果的均值. k值决定了交叉验证法评估结果的稳定性和保真性. 因此, 也称为k折交叉验证或k倍交叉验证.

(3) 在留出法和交叉验证法中, 训练集S的样本数小于数据集D, 因样本规模不同会导致所训练的模型及评估结果偏差. 据此, 发展出了基于自助采样获取训练集和测试集的自提升法. 给定包含m个样本的数据集D, 每次随机从D中选一个样本, 放入D', 然后, 将该样本放回初始数据集D中, 使得该样本在下次采样时仍有可能被采到; 这个过程重复执行m次后, 就得到了包含m个样本的数据集D', 规模和D一样, 不同的是D'中部分样本可能重复, 也有部分样本可能不出现. 一个样本在m次自助采样中都没有被采到的概率是$(1-1/m)^m$, 取极限得到0.368. 自助采样后, 将样本规模和数据集D一样的采样数据集D'作为训练集, 将$T=D-D'$作为测试集(不在D'中的样本作为测试集). 如此, 实际评估的模型(基于D')与期望评估的模型(基于D)使用了同样的样本规模(m个样本), 同时又有大概36.8%的样本作为测试集T用于测试, 产生的测试结果, 称为"包外"(out-of-Bag)估计, 其对模型的可靠性有直接关联.

在训练集的评估基础上, 严格的验证还包括在标准集测试(benchmark test)和未知数据集测试(盲测, blind test). 盲测往往与材料的研发制备紧密结合, 是应用材料基因组学预测模型非常重要的一环. 目前, 材料基因组学的预测模型主要有分类和回归两种, 在数据划分的基础上, 这两种模型的评估指标是不一样的.

2.4.1 分类模型评估指标

分类模型的性能评估多采用混淆矩阵. 混淆矩阵展示了分类模型对样本进行预测所得到的预测准确与不准确的数量. 它是一个$N\times N$矩阵, N为类别数. 模型性能通常可以从这个矩阵中的数据来评估. 表3给出了常见分类模型评估指标的计算表达方式.

在分类模型中, 通过连续移动true/false和positive/negative的阈值, 可以得到工作特征曲线(ROC), 其线下面积(AUC)也是评估分类模型的重要指标. 根据预测模型的用途, 在模型建立过程中可以最大化灵敏度、特异度、准确度和AUC等, 同样地, 在模型验证中, 这些指标具有较高的值(一般要大于0.6)说明模型具有较可靠的预测能力.

2.4.2 回归模型评估指标

回归模型的性能评估指标主要有均方误差(RMSE)、相对平方误差(RSE)和平均绝对误差(MAE)等. 均方误差是衡量回归模型预测率常用的量, 定

表 3 分类模型评估矩阵及其相关指标[a)]

混淆矩阵		目标			
		阳性	阴性		
模型	阳性	TP	FP	阳性预测值	TP/(TP+FP)
	阴性	FN	TN	阴性预测值	TN/(FN+TN)
		灵敏度= TP/(TP+FN)	特异度= TN/(FP+TN)	准确度=(TP+TN)/(TP+FP+FN+TN)	

a) 表中的符号意义如下: 阳性(positive, P), 阴性(negative, N), 真阳性(true positive, TP), 真阴性(true negative, TN), 伪阳性(false positive, FP), 伪阴性(false negative, FN), 灵敏度(sensitivity), 特异度(specificity, SPC)和准确度(accuracy, ACC).

义为:

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}} \tag{2}$$

其中y_i为实际值, $\hat{y}_i$为预测值. 而相对平方误差描述预测值在实验值分布中的相对偏差, 定义为:

$$\mathrm{RSE}=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(\bar{y}-y_i)^2} \tag{3}$$

其中$\bar{y}$为实验值的平均值. 与均方误差RMSE类似, 平均绝对误差的定义为:

$$\mathrm{MAE}=\frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{n} \tag{4}$$

这几个指标的值越小, 说明预测值与实验值越接近. 这些评估指标的使用依赖于预测值的分布情况, 在建模过程中, 需要根据具体分布, 选择合适的指标进行最小化. 在模型评估中, 特别是当标准集测试或盲测中发现指标远低于训练集时, 需要将训练集和测试集的预测值做统计分析, 确保预测模型在恰当的值域, 并且测试集中特征量的分布与训练集中的分布具有一定的相关性, 在实际预测中依然可靠.

3 建议

目前材料基因组学发展的主要障碍是材料领域研究的多样性, 缺乏数据储存标准, 以及缺乏对数据的分享激励机制. 就目前情况, 基于实验的材料数据库必须按照一定标准进行组织, 这些标准可以通过对实验结果取平均值或根据实验条件组织数据来完成; 基于计算模拟组织数据库相对容易, 因大量的材料属性目前已可以较为可靠地计算出来, 但基于第一性原理的计算仍然十分有限. 值得注意的是, 材料学家总是倾向于报道好的、正面的结果. 但是, 在材料基因组学中, 所谓的正面和负面结果具有同等重要性. 因此, 要鼓励材料研究者同时分享正面和负面的实验数据, 如构建公共数据分享平台, 发表文章中要求作者提供紧密相关的探索性数据等.

另一方面, 虽然数据库创建、数据库管理、统计学分析和机器学习是材料信息学的重要组成部分, 但数据收集分享的平台在材料信息学中同样不可或缺. 在一定的规范模式下, 实现全世界同行业科学家的数据共享和深入分析是材料基因组学数据源整合的理想目标. 我们可以利用互联网平台, 在使用一种主要的编程语言将数据收集、标准化、分享和传播过程中, 将其整合在一个平台中[56]. 目前, Dropbox、GitHub、HUBzero和nanoHUB等已出现活跃的大数据共享平台. 大科学装置(如德国DESY同步辐射中心)也建立了用户的数据和工具分享, 但国内类似平台还很缺乏. 在充足、可靠、全面的材料数据支持下, 通过DIKW过程, 将会极大地推进材料研究的颠覆性发展[5]. 同时也要积极开发自适应算法, 通过机器学习的自学习, 如AlphaGo Zero, 减少可靠预测性模型对先验性数据的依赖性.

通过材料基因组学大数据平台的建设, 最大程度地利用大数据辅助新材料研发, 是新材料绿色可持续发展的必然途径. 尽管已有越来越多的成功案例, 但还没有标准化(或模板化)的PSP (加工-结构-性质)连接协议, 包括物质、加工工艺的统一命名和数字化标准; 材料多种性质内在联系和统一, 以及基本属性、扩展属性和协同属性的明确; 数据存储交流的格式标准、统一符号和量纲等. 建立PSP链接标准化协议的主要障碍之一是: 大多数现存数据尚未被学术开源访问和

分析. 现代数据科学和信息学工具可以通过设计和实施分层文件格式转换器、电子协作平台、跟踪集成工作等来解决这一现状. 而采用材料基因组学的最大动力来自于在材料研究周期中材料研发成本和时间成本的节省. 因此, 材料科学家必须与理论、计算机和数据处理科学家紧密合作, 构建新的平台与新的成果数据分享方式. 具体需要便捷的数据收集和交流平台(包括界面和处理工具等), 数据分析结果的可视化呈现以及数据标准和数据使用协议等.

材料基因组计划是集成“官-产-学-研-用”多方力量加速材料研究的一个框架. 材料基因组项目的推进除了多学科研究者的积极参与外, 还需要发挥国家级科研项目和围绕大科学装置用户群的引导作用. 科技部于2016年启动了“材料基因工程关键技术与支撑平台”重点专项, 主要内容为构建高通量计算、高通量制备与表征和专用数据库3大示范平台; 研发多尺度集成化高通量计算方法与计算软件、高通量材料制备技术、高通量表征与服役行为评价技术, 以及面向材料基因工程的材料大数据技术4大关键技术. 该项目的启动是一个好的开端, 但仍需联合信息科学和材料理论研究人员, 发展面向特定应用和特定阶段材料开发的数据挖掘和机器学习算法和软件. 同时, 专业人才的培养, 材料基因组学的研究理念, 也需要通过大学教育、研究生培养和学术交流中进行推广.

4 结语

材料基因组学是一门新型的交叉学科, 其独特的研究模式以及对高质量数据、理论背景和材料前沿的综合要求, 促使理论和计算模拟工作者、信息科学和基于网络的大数据工作者与前沿的材料研究工作者必须深度合作. 虽然利用材料基因组学研究材料综合性能平衡的上限、从失败案例中学习新的合成路线、显著地提升针对特定应用的关键材料指标等领域已取得令人瞩目的成功, 但是, 正如最近关于高分子信息学(polymer informatics)的报道中指出的[57], 材料基因组学的发展有巨大的机遇, 也存在巨大的挑战. 目前, 材料基因组学也逐步渗透到更多的组成和工艺方案中, 能够极大地降低新材料研发成本, 提高研究效率. 对于理论和计算模拟工作者, 无论是粒子还是场为基础的结构体系, 拓展经典力场和模型, 开发高通量的快速算法, 并密切与面向应用的材料研究前沿结合, 将有力推进材料基因组学的成熟和完善. 对于材料的实验开发研究人员, 在特定的材料技术研究领域, 积极吸纳材料基因组学方法的知识和智慧, 加强与理论和信息研究的合作, 必能极大地提升新材料的研发效率, 降低开发成本. 通过材料基因组学的大力发展, 发挥其在先进材料和先进制造由辅助地位到不可或缺手段的积极作用.

参考文献

1 Katritzky AR, Lobanov VS, Karelson M. *Chem Soc Rev*, 1995, 24: 279–287
2 Rajan K. *Annu Rev Mater Res*, 2008, 38: 299–322
3 Kalidindi SR, De Graef M. *Annu Rev Mater Res*, 2015, 45: 171–193
4 Hill J, Mulholland G, Persson K, Seshadri R, Wolverton C, Meredig B. *MRS Bull*, 2016, 41: 399–409
5 Bird CL, Frey JG. *Chem Soc Rev*, 2013, 42: 6754–6776
6 Maier WF, Stöwe K, Sieg S. *Angew Chem Int Ed*, 2007, 46: 6016–6067
7 Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O. *Nat Mater*, 2013, 12: 191–201
8 Liu Y, Zhao T, Ju W, Shi S. *J Materiomics*, 2017, 3: 159–177
9 Hachmann J, Olivares-Amaya R, Jinich A, Appleton AL, Blood-Forsythe MA, Seress LR, Román-Salgado C, Trepte K, Atahan-Evrenk S, Er S, Shrestha S, Mondal R, Sokolov A, Bao Z, Aspuru-Guzik A. *Energ Environ Sci*, 2014, 7: 698–704
10 Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sanchez-Carrera RS, Gold-Parker A, Vogt L, Brockway AM, Aspuru-Guzik A. *J Phys Chem Lett*, 2011, 2: 2241–2251
11 Pyzer-Knapp EO, Li K, Aspuru-Guzik A. *Adv Funct Mater*, 2015, 25: 6495–6502
12 Olivares-Amaya R, Amador-Bedolla C, Hachmann J, Atahan-Evrenk S, Sánchez-Carrera RS, Vogt L, Aspuru-Guzik A. *Energ Environ Sci*, 2011, 4: 4849

13 Cheng L, Assary RS, Qu X, Jain A, Ong SP, Rajput NN, Persson K, Curtiss LA. *J Phys Chem Lett*, 2015, 6: 283–291
14 Ryder CR, Wood JD, Wells SA, Yang Y, Jariwala D, Marks TJ, Schatz GC, Hersam MC. *Nat Chem*, 2016, 8: 597–602
15 Sevov CS, Hickey DP, Cook ME, Robinson SG, Barnett S, Minteer SD, Sigman MS, Sanford MS. *J Am Chem Soc*, 2017, 139: 2924–2927
16 Hong Y, Lam JWY, Tang BZ. *Chem Soc Rev*, 2011, 40: 5361–5388
17 Wilmer CE, Leaf M, Lee CY, Farha OK, Hauser BG, Hupp JT, Snurr RQ. *Nat Chem*, 2012, 4: 83–89
18 Wilmer CE, Farha OK, Bae YS, Hupp JT, Snurr RQ. *Energ Environ Sci*, 2012, 5: 9849
19 Qiao Z, Peng C, Zhou J, Jiang J. *J Mater Chem A*, 2016, 4: 15904–15912
20 Simon CM, Kim J, Gomez-Gualdron DA, Camp JS, Chung YG, Martin RL, Mercado R, Deem MW, Gunter D, Haranczyk M, Sholl DS, Snurr RQ, Smit B. *Energ Environ Sci*, 2015, 8: 1190–1199
21 Colón YJ, Snurr RQ. *Chem Soc Rev*, 2014, 43: 5735–5749
22 Lin LC, Berger AH, Martin RL, Kim J, Swisher JA, Jariwala K, Rycroft CH, Bhown AS, Deem MW, Haranczyk M, Smit B. *Nat Mater*, 2012, 11: 633–641
23 Martin RL, Simon CM, Smit B, Haranczyk M. *J Am Chem Soc*, 2014, 136: 5006–5022
24 Park HB, Kamcev J, Robeson LM, Elimelech M, Freeman BD. *Science*, 2017, 356: eaab0530
25 Balachandran PV, Theiler J, Rondinelli JM, Lookman T. *Sci Rep*, 2015, 5: 13285
26 Xiong Z, He Y, Hattrick-Simpers JR, Hu J. *ACS Comb Sci*, 2017, 19: 137–144
27 Gaultois MW, Sparks TD, Borg CKH, Seshadri R, Bonificio WD, Clarke DR. *Chem Mater*, 2013, 25: 2911–2920
28 Isayev O, Fourches D, Muratov EN, Oses C, Rasch K, Tropsha A, Curtarolo S. *Chem Mater*, 2015, 27: 735–743
29 Dogan C, Stöwe K, Maier WF. *ACS Comb Sci*, 2015, 17: 164–175
30 Yuan D, Xiao L, Luo J, Luo Y, Meng Q, Mao BW, Zhan D. *ACS Appl Mater Interfaces*, 2016, 8: 18150–18156
31 Le TC, Winkler DA. *Chem Rev*, 2016, 116: 6107–6132
32 Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF. *ACS Cent Sci*, 2017, 3: 434–443
33 Krieg A, Arici E, Windhab N, Schattka JH, Schubert S, Schubert US. *ACS Comb Sci*, 2014, 16: 386–392
34 Sharma V, Wang C, Lorenzini RG, Ma R, Zhu Q, Sinkovits DW, Pilania G, Oganov AR, Kumar S, Sotzing GA, Boggs SA, Ramprasad R. *Nat Commun*, 2014, 5: 4845–4853
35 Breneman CM, Brinson LC, Schadler LS, Natarajan B, Krein M, Wu K, Morkowchuk L, Li Y, Deng H, Xu H. *Adv Funct Mater*, 2013, 23: 5746–5752
36 Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. *J Chem Theor Comput*, 2015, 11: 2087–2096
37 Liu L, Chen W, Li Y. *J Membr Sci*, 2016, 504: 1–9
38 Botu V, Batra R, Chapman J, Ramprasad R. *J Phys Chem C*, 2017, 121: 511–522
39 Brockherde F, Vogt L, Li L, Tuckerman ME, Burke K, Muller KR. *Nat Commun*, 2017, 8: 872
40 Oliynyk AO, Antono E, Sparks TD, Ghadbeigi L, Gaultois MW, Meredig B, Mar A. *Chem Mater*, 2016, 28: 7324–7331
41 Wang B, Zhou L, Xu K, Wang Q. *Ind Eng Chem Res*, 2017, 56: 47–51
42 Kuenemann MA, Fourches D. *Mol Inf*, 2017, 36: 1600143–1600157
43 Liu R, Chen H, Zhao K, Qin Y, Jiang B, Zhang T, Sha G, Shi X, Uher C, Zhang W, Chen L. *Adv Mater*, 2017, 29: 1702712
44 Ren J, Ocola LE, Divan R, Czaplewski DA, Segal-Peretz T, Xiong S, Kline RJ, Arges CG, Nealey PF. *Nanotechnology*, 2016, 27: 435303
45 Chung YG, Gomez-Gualdron DA, Li P, Leperi KT, Deria P, Zhang H, Vermeulen NA, Stoddart JF, You F, Hupp JT, Farha OK, Snurr RQ. *Sci Adv*, 2016, 2: e1600909
46 Rajan K. *Annu Rev Mater Res*, 2015, 45: 153–169
47 Jain A, Hautier G, Ong SP, Persson K. *J Mater Res*, 2016, 31: 977–994
48 Zang Q, Mansouri K, Williams AJ, Judson RS, Allen DG, Casey WM, Kleinstreuer NC. *J Chem Inf Model*, 2017, 57: 36–49
49 Hong WT, Welsch RE, Shao-Horn Y. *J Phys Chem C*, 2016, 120: 78–86
50 Kar S, Roy JK, Leszczynski J. *NPJ Comput Mater*, 2017, 3: 22
51 Leung SSF, Sindhikara D, Jacobson MP. *J Chem Inf Model*, 2016, 56: 924–929
52 Podlewska S, Czarnecki WM, Kafel R, Bojarski AJ. *J Chem Inf Model*, 2017, 57: 133–147
53 Thornton AW, Simon CM, Kim J, Kwon O, Deeg KS, Konstas K, Pas SJ, Hill MR, Winkler DA, Haranczyk M, Smit B. *Chem Mater*, 2017, 29:

2844–2854
54 Katritzky AR, Maran U, Lobanov VS, Karelson M. *J Chem Inf Comput Sci*, 2000, 40: 1–18
55 Pirhadi S, Sunseri J, Koes DR. *J Mol Graphics Model*, 2016, 69: 127–143
56 Takahashi K, Tanaka Y. *Dalton Trans*, 2016, 45: 10497–10499
57 Audus DJ, de Pablo JJ. *ACS Macro Lett*, 2017, 6: 1078–1082

Materials genome: research progress, challenges and outlook

Yunqi Li[1*], Lunyang Liu[1], Wenduo Chen[1], Lijia An[2*]

[1] *Key Laboratory of Synthetic Rubber, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, China;*
[2] *State Key Laboratory of Polymer Physics and Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences,Changchun 130022, China*
**Corresponding authors (email: yunqi@ciac.ac.cn; ljan@ciac.ac.cn)*

Abstract: We are facing a new era with plenty of scientific data, which stimulates the born of materials genome (MG). It tackles huge amount of data to informatics, and finally provides wisdom for the development of novel materials with superior properties. Representative successes in applying MG for variant purposes, including clean energy, gas separation, alloy, catalyst and functional polymer materials etc., have been summarized, and approaches and suggestions for further development of MG are presented in this review. Though there are still alternative challenges in the advancement of MG, it exhibits a powerful paradigm to satisfy the thirsty for advance materials beyond all doubts.

Keywords: material genome, advance materials, computer aided material design, high through-out screen, composition-process-structure-property relationship

doi: 10.1360/N032017-00182