材料导报 MATERIALS REPORTS
综述
2021,Vol.35,No.15 www.mater-rep.com

机器学习技术在材料科学领域中的应用进展

米晓希,汤爱涛✉,朱雨晨,康 靓,潘复生

重庆大学材料科学与工程学院,重庆 400044

材料是国民经济的基础,新材料的发现是推动现代科学发展与技术革新的源动力之一,传统的实验“试错型”研究方法具有成本高、周期长和存在偶然性等特点,难以满足现代材料的研究需求。近些年,随着人工智能和数据驱动技术的飞速发展,机器学习作为其主要分支和重要工具,受到的关注日益增加,并在各学科领域展现出巨大的应用潜力。将机器学习技术与材料科学研究相结合,从大量实验与计算模拟产生的数据中挖掘信息,具有精度高、效率高等优势,给新材料的研发和材料基础理论的研究提供了新的契机。

机器学习技术结合了计算机科学、概率论、统计学、数据库理论以及工程学等知识,计算速度快、泛化能力强,能有效地处理一些难以运用传统实验及模拟计算方法解决的体系和问题。近 10 年,机器学习在材料科学研究中的应用呈现出爆炸式的增长,尤其在新材料的合成设计、性能预测、材料微观结构深入表征以及改进材料计算模拟方法几个方面,均有着出色的表现。当然,作为一项数据驱动技术,如何获取大量实验数据并将其构建为行之有效的数据集仍是现阶段机器学习技术在材料科学领域应用的热点和难点。

本文概述了机器学习技术的基本原理、主要工作流程和常用算法,简述了机器学习技术在材料科学领域中的研究重心及应用进展,分析了机器学习在材料学研究中尚存在的问题,并对未来此领域的发展热点进行了展望。

关键词 机器学习 性能预测 结构表征 计算模拟

中图分类号:TP181 文献标识码:A

Research Progress of Machine Learning in Material Science

MI Xiaoxi, TANG Aitao✉, ZHU Yuchen, KANG Jing, PAN Fusheng

College of Materials Science and Engineering, Chongqing University, Chongqing 400044, China

Materials are the foundation of the national economy, the discovery of new materials gives impetus to the development of modern science and technological innovation. The traditional “trial and error” experimental methods are no longer applicable to the research of modern materials owing to the disadvantages of high cost, long period and great contingency. In recent years, with the rapid development of artificial intelligence and data-driven approach, as a main branch and an important tool of them, machine learning is receiving increasing attention and showing tremendous potential. The integration of machine learning into material science research can greatly improve the precision and efficiency, and provide new opportunities for the research and development of new materials and the study of the basic theory.

Machine learning technology combines knowledge of computer science, probability theory, statistics, database theory and engineering. It shows a faster computing speed and good generalization ability, and can effectively deal with some systems and problems difficult to tackle by traditional experiments and numerical simulation. In the past decade, the applications of machine learning in material science research have shown explosive growth, especially in the synthesis and design of new materials, the property prediction, characterization of the microstructure, and the improvement of material calculation and simulation methods. Machine learning will be indispensable in the development of material science and engineering in the future. At present, how to obtain a large number of experimental data and build effective data set is still a hot spot and difficulty in the application of machine learning in the field of material science.

This paper outlines the basic principles, workflows and common algorithms of machine learning, briefly describes the research focus and application progress of machine learning technology in the field of materials science, and analyzes the existing problems of machine learning in materials science research. Meanwhile, some hot spots of the material field in the future are pointed out.

Key words machine learning, performance prediction, microstructure characterization, calculation and simulation

0 引言

自 20 世纪 90 年代以来,随着计算机技术与网络的飞速发展,如何利用计算机实现大规模的智能化已经成为一个备受关注的课题,大数据与人工智能的结合被称为“科学的第四范式”或者“第四次工业革命”,这表明当代人工智能的出现和发展有可能极大地改变和提高计算机在科学研究及工程应用中的作用[1-4]。机器学习作为人工智能发展最为迅速的一个分支,近些年受到人们的广泛关注,目前机器学习已被广泛应用于自然语言处理[3-4]、图像处理[5]、医学诊断[6]、生物医药[7-8]、智能机器人[9]、固态物理学[10]、证券市场分析[11]等众多领域,并取得了巨大的成功。

2011 年,“材料基因组计划(MGI)”被提出,标志着以数据驱动技术为核心的信息学战略开始在材料科学中形成[12]。机器学习技术相较于传统的“试错型”研究方法,具有成本低、效率高、周期短、尺度广等特点,渐渐成为材料基因工程必不可少的研究手段,在材料领域内引起了极大的重视。一方面,材料学领域在信息时代产生了海量数据,并建立了较大规模的数据库[13],由于机器学习核心统计算法对大数据具有良好的处理和泛化能力,能够从已有的实验数据中挖掘出新的信息,探寻各类参数间复杂的隐含关系,建立精准的预测模型,充分发挥实验数据的作用。另一方面,在计算材料领域,目前常用的计算模拟方法(如第一性原理计算、分子动力学、有限元模拟等)不但需要耗费大量的时间和资源,还存在诸多限制,而运用机器学习不但可以大大节省计算时间,扩大计算体系的空间和时间尺度,还可以通过数据拟合优化

基金项目:国家重点研发计划项目(2016YFB0301100);重庆市自然科学基金(cstc2017jcyjBX0040);国家自然科学基金(51531002)

This work was financially supported by the National Key Research and Development Program of China (2016YFB0301100), Natural Science Foundation of Chongqing (cstc2017jcyjBX0040), National Natural Science Foundation of China (51531002).

✉ tat@cqu.edu.cn

DOI:10.11896/cldb.20060168

原子势函数,使已有的计算模拟方法更加快速、精确。此外,随着机器学习技术在计算机视觉以及图像识别等领域的不断发展,将机器学习用于材料宏观组织分析和微观结构表征,甚至晶体结构建模和预测都已经成为可能,且将会是未来研究材料结构的有效手段。

将机器学习技术与材料科学相结合,充分发挥数据驱动技术的优势,给材料科学研究提供了新的手段、新的方向。目前越来越多的研究者开始将目光投向这种新型的研究方式,机器学习在材料科学领域的应用数量也以惊人的速度增长。

1 机器学习的基本原理与常用算法

机器学习是一门多学科交叉专业,涵盖计算机科学、概率论、统计学、近似理论和复杂算法等知识,它的本质是基于大量的数据和一定的算法规则,使计算机可以自主模拟人类的学习过程,并能够通过不断的数据"学习"提高性能并做出智能决策的行为。在传统的计算方法中,计算机只是一个计算工具,按照人类专家提供的程序运算。在机器学习中,只要有足够的数据和相应的规则算法,计算机就有能力在不需要人工输入的情况下对已知或未知的情景做出判断及预测,学习数据背后的规则。简而言之,机器学习就是研究如何让机器像人类一样"思考与学习"[3],这与机器按照人类专家提供的程序"工作"有本质的区别。通常,人类的学习过程要经历知识积累、总结规律,最终才能达到灵活运用的阶段。类似地,机器学习也分为输入、学习、输出三个阶段,机器学习的基本工作流程如图1所示。

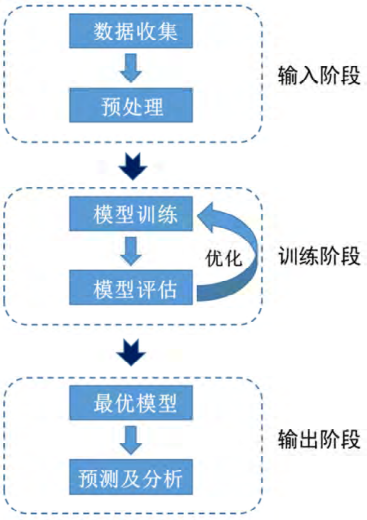

图1 机器学习的基本工作流程

Fig.1 Basic workflow of machine learning

数据的输入阶段,包括数据的收集及预处理,机器学习的核心是数据,收集充足的数据并建立有效的数据集是数据挖掘的前提,数据应尽可能完整且分布均匀,原始数据可以是文本、数值甚至音像,但数据呈现的形式往往会影响模型学习。对于相同的原始数据,机器学习算法使用一种格式可能比使用另一种更有效,输入数据的表示形式越合适,算法将其映射到输出数据的精度就越高,将原始数据转换成更适合的算法形式的过程被称为特征化或特征工程。

模型的学习阶段,是指通过一定的算法来对数据进行识别分析或探寻数据间的隐含关系,此阶段通常包括算法选择、模型结构参数优化、训练及测试等过程。不同算法依据不同的数学原理,也对应不同的模型结构参数,算法与数据的契合程度决定了学习模型的准确度,为了获得最优模型,可以通过增加有效训练数据、优化模型结构和参数等方式。

最后的输出阶段就是利用优化好的模型对未知的数据做出预测或者分析,机器学习适用范围非常广,实际应用效果取决于模型的精度,通俗地说,就是是否已经通过学习大量相似的老问题而总结出非常可靠的经验规律来解决一个新的问题。

机器学习系统通常被分为监督学习、无监督学习以及强化学习三大类。在监督学习问题中,一个程序的学习内容通常要包含标记的输入和输出,并从新的输入预测新的输出,通俗地说,即程序从提供了"正确答案"的例子中学习。相反,在无监督学习中,一个程序不会从标记数据中学习,它尝试在数据集中发现模式,即程序自己去寻找规则并预测答案。强化学习靠近监督学习一端,区别是强化学习程序不会从标记的输出中学习,而是从决策中接收反馈。例如,当机器程序对一个项目完成度较高时进行奖励,完成度较低时给予惩罚,从而提高此项目的完成效果,但是强化学习并不能指导程序如何完成项目。根据学习任务的不同,机器学习可以分为四类,如表1所示,分类和回归是两种最常用的监督学习任务,聚类和降维是两种最常用的无监督学习任务。

机器学习模型中的具体计算规则被称为算法。对于相同数据集,即使采用同类不同种的算法计算,结果也会有差别,因此选择契合数据的算法是机器学习过程中至关重要的一环。机器学习的算法有很多种,一些较常见的算法有:朴素贝叶斯(NB)、K近邻(KNN)、支持向量机(SVM)、决策树(DT)、随机森林(RF)、最大期望(EM)算法、人工神经网络(ANN)以及深度学习(DL)等,在这里对以上算法进行简单介绍。

表1 机器学习任务及主要算法功能

Table 1 Main tasks and functional algorithms of machine learning

学习任务	分类	回归	聚类	降维
主要功能	对已有数据进行分类,将给定样本放入相应类别	用函数拟合已知数据,从而预测未知样本	不经过训练,将样本划分为若干类别	减少要考虑的随机样本的数量
类型	监督学习	监督学习	无监督学习	无监督学习
常用算法	K近邻; 决策树; 支持向量机; 朴素贝叶斯	线性回归; Logistic回归; 人工神经网络; 深度学习	K均值聚类; AP聚类; 层次聚类; DBSCAN	主成分分析(PCA); 判别分析(LDA)

(1)朴素贝叶斯[14]:基于贝叶斯概率理论的一种常用的分类算法,通过计算不同独立特征的条件概率来进行类别划分。该算法在数据较少的情况下依然有效且可以处理多类别问题,但是对数据的输入方式较为敏感。

(2)K近邻算法[15]:KNN算法是最简单的机器学习算法之一,通过计算空间中样本与训练数据之间的距离,再以K个"最近邻"点中大多数点的类别决定样本类别的算法。K近邻算法精度高,无数据输入假定,但是当数据量增加时,空间计算复杂程度也相应提升。

(3)支持向量机[16]:SVM模型是一种二分类模型,其学习策略是通过构造一个边距最大的超平面将多维空间分为两个区域来进行分类。此算法的核心是依靠"核函数"将低维数据提升至更高维度的空间来寻找到具有最大间隔的超平面。

(4)决策树[17]:决策树基于树形结构,结构简单、效率较高,是一种十分常用的分类方法。决策树以流程图的方式将一个类标签分配给一个实例,从包含训练集中所有数据的根节点开始,根据一个属性的值分成两个子节点(子集)。选择属性和相应的决策边界,使用其他属性从两个子节点继续分离,直到一个节点中的所有实例都属于同一个类,结束节点通常被称为叶节点。其核心思想是递归选择最优特征进行分类。

(5)随机森林[18]:随机森林是一种重要的基于Bagging的集成学习算法,可以用于处理分类与回归问题。此算法通过构建多棵决策树提高信息增益,从而降低噪音数据带来的干扰,随机森林的输出由决策树输出的众数决定。

(6)最大期望[19]:EM算法是一种启发式的迭代算法,可实现用样本对含有隐变量的模型的参数做极大似然估计。已知的概率模型内部存在隐含的变量,导致不能直接用极大似然法来估计参数,EM算法就是通过迭代逼近的方式用实际的值代入求解模型内部参数。

(7)人工神经网络[20]:人工神经网络通过模仿人类大脑的神经元神经网络处理进行数据信息处理,分为输入层、隐藏层以及输出层三部分,输入层单元接受外部的信号与数据,隐藏层处在输入和输出单元之间,主要用于调整神经元间的连接权值及单元间的连接强度,最终将处理结果传递到输出层,输出层单元被激活函数激活后实现系统处理结果的输出。

(8)深度学习[21]:深度学习可以简单地理解为具有深层网络结构的人工神经网络,与人工神经网络相比,其网络结构更加复杂,计算量也更大。深度学习网络在处理图像方面有极大的优势,常用的深度学习模型包括卷积神经网络(CNN)及深度置信网络(DBN)等。

2 机器学习加速材料科学研究的进程

材料科学研究近些年发展十分迅猛,无论是依靠实验手段还是计算模拟,均产生了大量数据信息,依靠机器学习算法可以从数据中挖掘有效信息,尤其是对于计算产生的大规模、高维度的数据集,可以有效识别大数据集中的特征模式,提取数据集中的隐含规律和相关性[22-24]。此外,利用机器学习技术还可以实现特征提取、图像识别等[25]。基于机器学习技术的空间能量场和力场拟合还可以辅助优化现有的材料理论势函数,促进计算模拟技术的进步,甚至与新兴起的高通量计算也有很好的融合性,从而充分地发挥其高效性和良好的泛化能力。现阶段,机器学习技术在材料学研究中的应用主要体现在以下四个方面。

2.1 指导新材料的合成与开发

新材料的合成往往伴随着大量的数据和冗杂的参数,在材料化学合成路线中,每一个步骤可能发生的转变数量从几十到几千不等,由此需要考虑极端庞杂的系统和大量潜在的解决方案组合,在这些组合中,往往还存在许多相互竞争的参数(如时间、成本、纯度、毒性等),因此传统实验方式非常不适合当今形势下新材料的合成与开发。有机化学家们最早认识到计算机技术在化学合成领域的巨大潜力,在50多年前,Corey[26]开发的"有机物模拟合成程序(OCSS)"就已经依靠计算机对材料自动化学合成进行了尝试。近些年大量研究成果显示,在指定条件和合成规则确定的情况下,利用机器学习技术,计算机完全可以取代人类专家,甚至比人类更加高效。

钙钛矿的合成研究就是一个典型的例子,早期钙钛矿的研究是通过手动收集上百种ABX_3型化合物的实验数据[27-28]并人为判断其是否能够形成钙钛矿,这种方式存在数据筛选不全面、主观性强等问题,极容易出现错误。将机器学习技术引入钙钛矿研究中,庞杂的化学合成问题可以得到有效解决。Xu等[29-30]采用八面体共顶连接形成的拓扑网络作为钙钛矿结构的判断标准,基于"Materials Project"数据库,筛选出590种ABX_3和538种$A_2B'B''X_6$化学式的化合物。他们利用机器学习方法,以原子序数、离子半径、电负性、八面体因子以及容忍因子等作为描述符,判断上述结构是否具有钙钛矿的结构属性,精度达到90%。Raccuglia等[31]以实验结果为基础建立了化合反应数据库,结合支持向量机及决策树算法,对3 955组完整的水热合成反应数据进行训练和测试,预测了钒亚硒酸盐的结晶情况,其预测精度为89%,甚至比具有丰富经验的化学家人工筛选还要精准。值得注意的是,他们使用的化合反应数据中既包含成功的实验数据,也有失败的数据。Jatin等[32]利用机器学习模型实现了聚(2-恶唑啉)浊点范围的精确预测,在四个重复单元和一系列分子质量的设计空间中,采用梯度增强决策树在24~90 ℃实现了4 ℃均方根误差的精度(比线性和多项式回归高三倍)。通过粒子群优化算法进行设计,在37~80 ℃的四个目标浊点下预测并合成了17种具有约束设计的有最低预测方差的新聚合物。Pillong等[33]应用卷积神经网络提取晶体结构特征,建立了一种能够预测给定分子结晶倾向的双参数机器学习模型,这个模型收集了超过20 000种晶体和非晶化合物参数的训练集,这个机器学习模型不但可以用于新材料合成设计,还可用于评估产品结晶倾向,其准确度超过80%。

此外,受益于机器学习的前沿算法发展,采用强化学习,如生成对抗网络(GAN)可以有效设计和开发新材料分子结构。如图2所示[3],在一个生成式对抗网络中,同时训练两个模型,即生成式模型和判别式模型,生成式模型用于捕获

数据的分布,而判别式模型则用来估计样本来自训练集而不是生成器的概率。生成器的训练过程是使鉴别器出错的概率最大化,基于目标增强技术,生成反求网络的模型能够从零开始生成新的有机分子。通过类似于心理学经典条件反射的奖赏机制,这些模型可以被训练生成包含特定化学特征和物理反应的多种分子。使用强化学习,新生成的化学结构可以偏向于那些具有期望的物理和生物特性的结构(从头设计)。

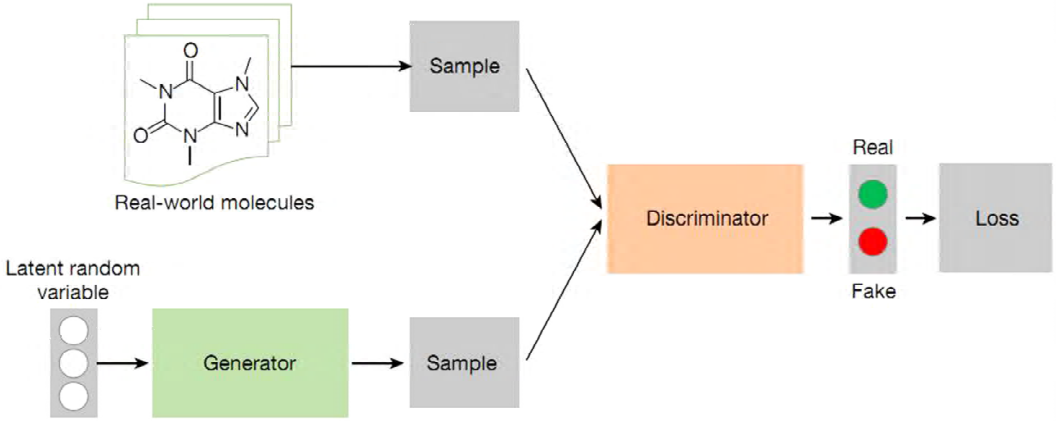

图2 分子发现的生成对抗网络(GAN)方法[3]

Fig.2 The generative adversarial network (GAN) approachof molecular discovery[3]

采用机器学习进行新材料的合成设计,渐渐成为新材料合成及设计的新选择,也给深入研究新化合物的合成准则带来了可能,从大量的实验数据出发,通过训练探究规律、积累经验,机器学习可以像一位经验丰富的材料学家一样,筛选分析,做出判断,给出预测,甚至做得比人类更好,大大节省了时间与实验损耗。随着算法的丰富与复合框架的更新,新材料的合成设计与开发会以难以想象的速度发展。

2.2 材料属性预测

目前,材料的设计及研发主要还是依靠传统的"试错型"实验方法来制备和测试样品,周期长、效率低、成本高且存在一定的偶然性,将机器学习应用在材料的设计与研发中,从大量实验数据(包括成分、宏观性能、组织结构、能量特征参数以及工艺参数等)中提取出各参数间的隐含关系,寻找属性影响因素及变化规律,获得预测模型,从而指导新材料的设计。

在具有特定功能的材料(如锂离子电极、半导体、超导、磁性材料等)的研发过程中,找到结构与性质之间的对应关系十分重要,在很多情况下,材料的结构与性质之间并非简单的线性关系,在现有的理论知识无法解释结构与性质之间的关系时,利用机器学习进行训练学习,预测材料的结构和性质,挖掘其中的对应关系也是一种可行的方法[23]。Atta-rian等[35]应用包含线性回归、人工神经网络、支持向量机、K近邻及随机森林五种机器学习算法对硅酸盐正极材料进行晶体结构上的分类,并探究了阴极材料晶体系统与不同物理性能间的相关性。类似地,Fujimur[36]采用支持向量机回归模型来预测全固态电池中锂离子导体材料的离子电导率。人工神经网络是最常用的预测模型回归算法,Jie等[37]采用人工神经网络模型进行材料带隙结构预测,从数据库约4 000种化合物中确定了34种带隙特殊的化合物。Gu等[38]利用卷积神经网络模型预测了二维复合材料的力学性能,对上百种软质和硬质材料数据进行组合,预测了超过百万种目标材料的比强度和比刚度,对材料性能的"好"与"坏"进行了判断,并研究了不同参数对训练过程的影响。结果表明,采用机器学习,无论训练数据密度低、批量小、训练循环次数少,还是数据量庞大、结构复杂、训练次数多,均可以实现高精度的预测,预测精度可达95%。Li等[39]提出了一种被称为原子表卷积神经网络的机器学习方法,该方法可以在训练中不断学习合适的特征来预测化合物的形成能、带隙和超导转变温度T_c(机器学习模型和预测结果如图3所示),该模型的精度超过了标准DFT计算的结果。通过数据增强的方法,这种模型不仅能够准确预测超导体的超导转变温度,还能够区分超导体和非超导体。利用这种模型,他们从数据库中筛选出20种可能具有高超导转变温度的材料。

对金属材料而言,影响其性能最为关键的限制因素就是合金成分及组织结构。近期,高熵合金(HEA)和复杂成分合金(CCA)因具有卓越的力学和物理性能而引起人们极大的研究兴趣。尽管已有许多有用的HEA或CCA的报道,但用来指导合金筛选的相设计规则尚不明确。通过收集已有的实验数据,利用机器学习进行相结构设计的针对性强、准确度高,Yang等[40]基于人工神经网络等三种不同的算法并采用601组多元合金数据集训练了模型(见图4),从ML建模中提取出敏感性矩阵,由此可以定量评估如何调整设计参数以形成特定的相,如固溶体、金属间化合物或非晶相等。基于该模型,他们定量评估了文献中已有的高熵合金相结构的设计规则,并探索提出了一组全新的设计参数,该研究表明基于机器学习的技术有望成为发展高熵或多组元合金设计的新工具。

传统材料设计是很难以性能指标作为参数进行精准设计的,但是由于监督机器学习的训练数据具有输入与输出一一映射的关系,这就给材料性能预测以按需设计提供了可能。Liu[41]采用BP神经网络构建了AZ31镁合金力学性能预测模型,以取样方向、退火温度、退火时间工艺参数为输入,屈服强度、抗拉强度和延伸率作为输出,采用参数全排列的方式对模型进行优化,得到了精度较高的能够预测AZ31

镁合金经不同退火工艺处理后的力学性能预测模型。Wang[43]以实验数据为训练和测试集构建了“成分—性能”及“性能—成分”两个人工神经网络预测模型，提出了一个包含机器学习建模、成分设计和性能预测三个功能的机器学习设

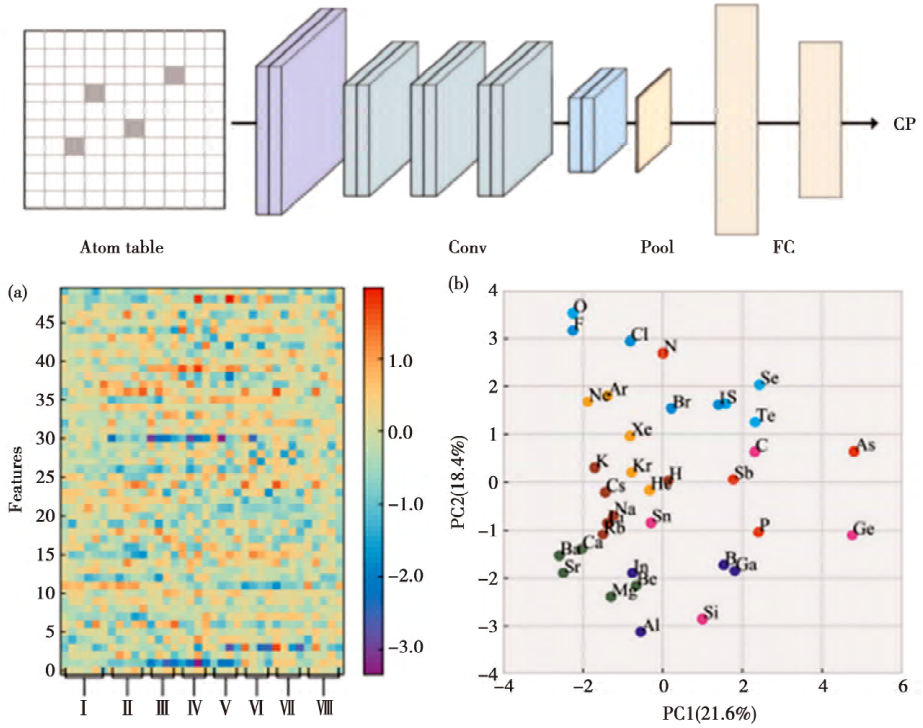

图 3 原子表卷积神经网络预测 T_c 的机器学习模型[39]（电子版为彩图）

Fig.3 Schematic diagram of the ATCNN model for T_c prediction[39]

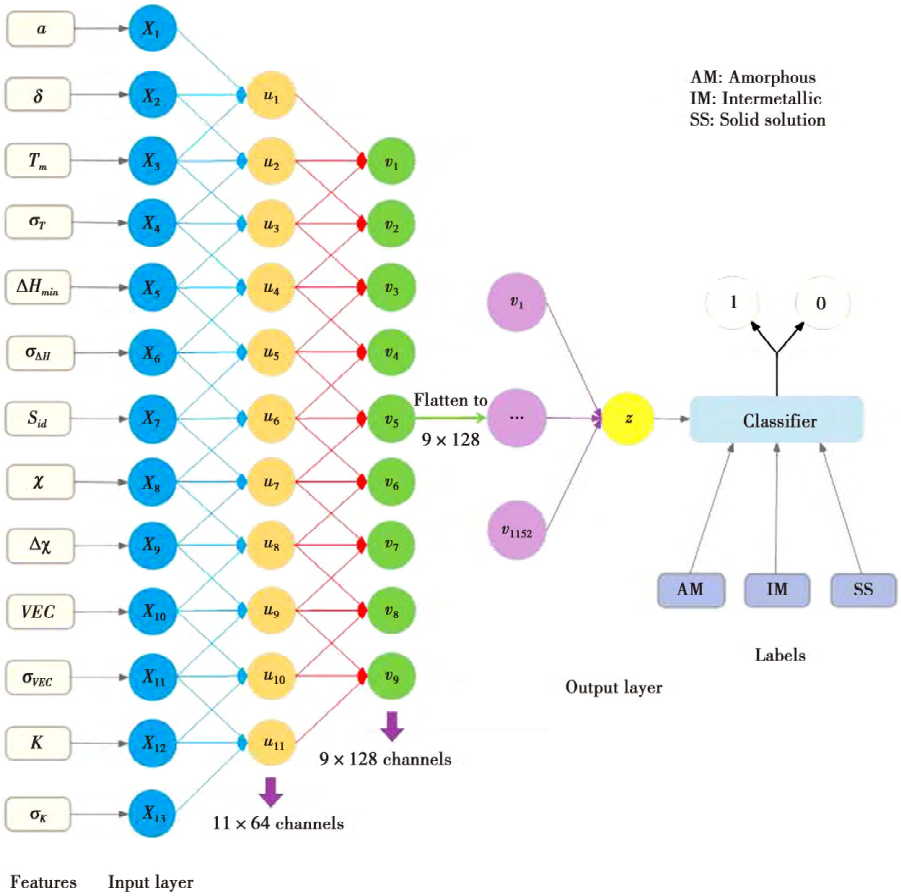

图 4 高熵合金设计人工神经网络模型结构[40]

Fig.4 The structure of artificial neural network model of the HEA[40]

计系统,其在高性能铜合金的快速成分设计上有很好的效率。以目标极限抗拉强度为600~950 MPa、导电性为50.0%的铜为标准,在机器学习设计系统的指导下(MLDS),设计了不同成分的铜基合金,且设计合金的性能预测值与文献记录的十分相近。Tagade[43]则提出了一种"新型条件采样属性驱动的材料设计的结构学习(SLAMDUNCS)"的深度学习逆向预测框架,将贝叶斯推论与深度学习网络相结合,成功设计出具有所需性能的有机分子,该框架可以很容易地推广于无机材料的设计,如半导体、压电和热电材料。

2.3 辅助材料微观结构表征

材料的微观结构直接影响到其宏观性能,微观结构的表征和调控是发展材料理论基础及新材料设计开发的关键手段,传统的表征从早期的金相组织,到借助扫描电镜和透射电镜进行晶体结构和原子缺陷分析,以及最新的原位分析、环境电镜等更加精细的观测方法,这些研究方式除了依赖先进的实验设备外,还需要研究人员具有丰富的标定经验,通过一些成熟的商用标定软件进行定性、定量处理,费时费力,也有很大概率出现错误。近些年,机器学习为研究材料微观结构提供了一种崭新的途径,例如,用玻尔兹曼机模拟MoS_2化学沉淀相的单分子层结构[44],利用卷积神经网络建立"微结构—大性能"间的映射关系[45],利用无监督概率学习方法来识别并去除测量谱中的背景,从而实现了XRD谱和拉曼光谱的自动识别[46]。此外,随着机器学习尤其是深度学习技术的飞速发展,图像识别技术也得到了空前的进步[21]。将图像识别、分类算法、降维处理、数据增强等机器学习方法应用于材料的结构表征和分析中,对材料理论基础的认识和新材料的研发有着极大的推动作用。

二维晶体由于具有低维特性以及与体材截然不同的性能,在工程、物理、化学、药学和生物学等诸多领域引起了学者越来越多的兴趣。机械剥离是一种十分重要的制备二维晶体的方法,但是在衬底上机械剥离出的晶体通常含有较厚的薄片,手动筛找和清除这些薄片费时费力,限制了原子厚度的二维晶体和范德华异质结的高通量制备。Saito等[47]提出了一种通过深度神经网络自动划分和识别二维晶体厚度的通用技术,构建由卷积U-Net组成的框架,该神经网络只使用少量(24和30张)实际晶体图像的数据进行训练就可成功区分单层和双层MoS_2、石墨烯,成功率可达70%~80%,该研究表明很大一部分的实验室手工工作将可能被基于AI的系统所取代,为探索基于AI的大规模快速制造2D材料和范德华异质结的新方法开辟了新的途径。

利用深度学习技术还可以实现三维样品微观结构的表征,Subramanian等[48]结合无监督机器学习技术、拓扑分类和图像处理方法建立了一种微结构分析方案,可以自动识别并分析三维数据样品中的微结构(见图5)。该技术不需要先验地描述目标系统的微结构,对无序不敏感(例如由线缺陷和平面缺陷引起的多晶中扩展缺陷),定量地证明了该技术能获得无偏微结构信息(例如3D多晶样品中晶粒的精确定量及其尺寸分布),表征了3D聚合物样品中的孔隙和孔隙度以及3D复杂流体中胶束尺寸分布等特征。他们还采用金属、聚合物和复杂流体等各种模拟数据及实验表征数据进行了比较研究。这种方法计算效率高,可以快速识别、跟踪和量化那些影响材料性能的复杂微结构特征。

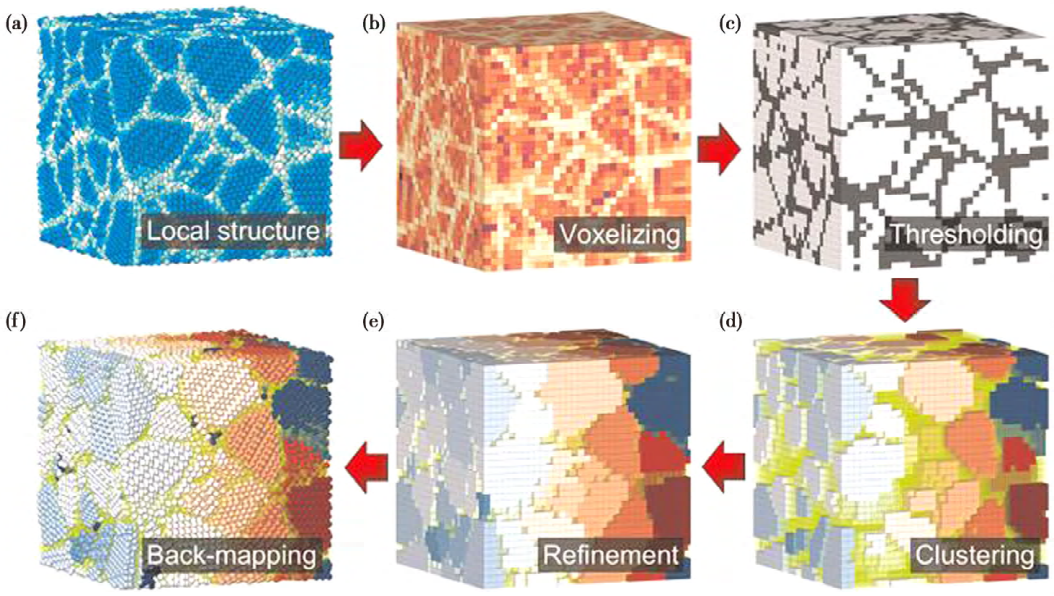

图5 三维样品微观结构表征[48]

Fig.5 Microscopic characterization of three-dimensional sample structure[48]

缺陷的识别及分析也是金属材料结构研究中非常重要的一部分,借助于机器学习的图像识别及分类技术,可以快速识别缺陷并对缺陷进行表征,甚至对金属材料的强韧化机制研究也有一定的促进作用。Andrew等[49]为了探索AZ31合金中拉伸孪晶形核的驱动力,利用机器学习框架对电子背散射衍射(EBSD)扫描得到的数据进行挖掘,提取了导致孪晶的物理特性之间的相关性。他们选择了j48决策树模型来捕捉微观结构与孪晶形核的关系,研究表明:AZ31合金孪晶的成核主要受晶粒尺寸、基面施密特因子和体位错密度的共同控制,而晶界长度、基面施密特因子、晶界面到RD面的夹

角和晶界错位对孪晶的生长影响最大。在此基础上，Tong[50]采用了决策树、梯度提升树、支持向量机、朴素贝叶斯、人工神经网络五种机器学习算法，进一步探究了具有不同织构的AZ31镁合金孪晶形核成长及组织演化机理，模型准确率达到87%。与传统实验方法相比，机器学习技术可以更容易地用于研究微观结构和材料属性之间的响应关系。在工业生产中，对产品宏观缺陷的控制至关重要，为此，Gao[51]开展了一系列基于深度学习的轧制镁薄板缺陷分类算法研究——基于Tensor Flow研究镁薄板缺陷分类的卷积神经网络及其算法，用约50 000张镁薄板边裂纹图像进行训练，将两个经典的卷积神经网络模型Le Net-5和Alex Net结合，并在此基础上进行了不同程度的改进，在卷积神经网络模型的基础上加入了迁移学习的策略。实验测试结果表明，基于迁移学习的卷积神经网络对镁薄板缺陷的分类准确率达到96.0%，将此研究成果应用于轧制镁合金生产线，可以大大减少人为误判和资源损耗，并且具有很高的准确度。

随着人们对材料性能的要求不断提高，材料基础理论研究也愈发重要，想要从根本上提升材料的性能，就必须从材料微观结构上进行探究，机器学习技术对推动材料的微观结构表征有巨大的潜力。

2.4 加速计算模拟技术发展

材料科学与计算机学科的交叉应用推动了材料科学的第一次计算革命，传统的计算模拟手段包括密度泛函理论(DFT)、蒙特卡洛法(MC)、分子动力学(MD)和有限元分析(FEA)等，通过理论计算及模拟技术，研究人员能够更有效地探索更加微观的相位和成分空间，借助抽象的能量概念处理更加复杂的体系，减少浪费。但是，每种计算模拟手段的适用条件是有限的，例如密度泛函理论模拟规模小，且在高温、高压、强磁场环境中的模拟结果误差较大；分子动力学模拟则依赖较为精确的势函数；有限元模拟需要准确的边界条件设定，且精确度浮动较大等。将机器学习方法与计算模拟手段相结合，发挥机器学习处理大数据及可外延的能力，是计算材料领域的一种新的研究思路。

目前应用最广泛的材料计算模拟方法是密度泛函理论，将密度泛函理论与机器学习结合起来的方法有很多，最直接的方法就是利用密度泛函理论的计算结果作为机器学习的训练数据，采用监督学习进行预测。例如，Tanaka等[52]利用第一性原理计算得到的339组晶体结构数据，结合五种机器学习算法，设计了锂离子电池阴极材料；Terentyev等[53]基于第一性原理计算的多个配置下的能量数构建数据集，以原子的迁移能为神经网络输出，原子的局部环境配置为神经网络输入，构建了用于计算不同体系结构下空位的迁移能，其精度与NEB计算得到的相当；Lu等[54]从212种已经报道的无铅有机-无机杂化钙钛矿(HOIPs)第一性原理计算的带隙值中训练ML模型，开发了一种靶向驱动法，该方法可用于发现稳定的HOIPs。他们的实验结果表明，在理论计算数据与实验数据的一致性有保证的情况下，机器学习可以同时利用实验数据和理论计算数据，更加高效地进行模拟计算。分子动

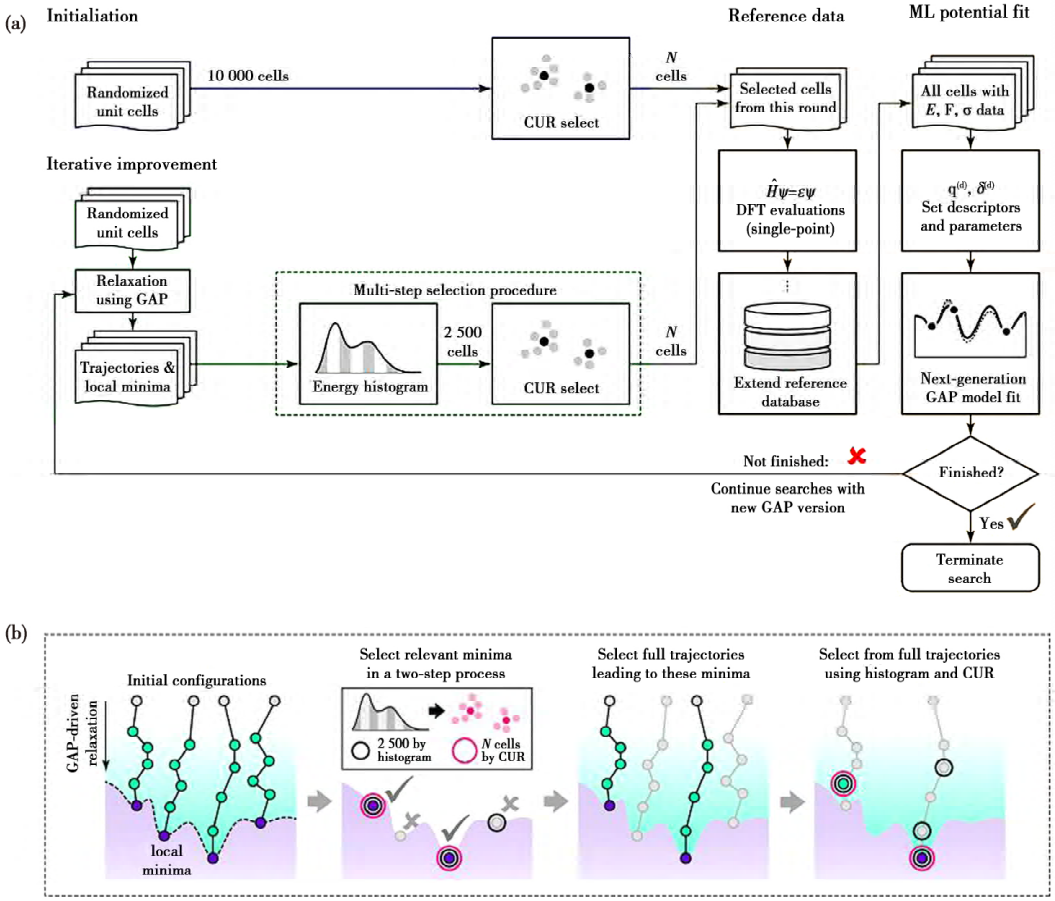

图6 机器学习拟合势函数模型[55]

Fig.6 Potential function model by ML[55]

力学模拟对具有较大的结构体系以及极端环境的模拟非常有效,然而,分子动力学计算依赖精确的势能信息,即势函数,势函数的开发工作量大、难度高,严重限制了分子动力学的发展和应用。运用机器学习的方法可以通过大量原始数据拟合空间力场,更加精确地描述原子间的位能关系,优化计算模拟方法。目前,已经有很多科研工作者利用机器学习手段拟合空间力场势能信息,完善势函数。Deringer[54]研究发现,ML原子间势可以以一种高度自动化的方式构建、探索和拟合不同的势能面,因此他们构建了一个势能面自学习的机器学习模型,通过建立区域原子力场"指纹",对原子力分量的每个原子环境进行"指纹"识别,然后利用优化聚类算法,拟合"指纹"和力之间的非线性关系,预测更加大范围的力场信息,实验结果与通过嵌入式原子势(EAM)函数的计算结果吻合较好,机器学习拟合结构流程如图6所示。

Ding等[56]将卷积神经网络与分子动力学模拟相结合,用DFT实验数据建立特征参量,设计了一个从头算分子动力学(AIMD)模型,成功拟合出锆合金马氏体相变过程的原子间势,弥补了锆合金马氏体相变势函数的空白,并将其应用于分子动力学计算。Cooper等[57]提出一种基于原子间作用力信息来高效训练高精度神经网络经验势函数的方法,他们将水和过渡族氧化物间的原子力信息转换为近似能量,由此构建机器学习人工神经网络经验势,该方法简化了神经网络经验势的构造过程,有望推广应用于任意类型的材料中。这些研究工作表明,机器学习技术拟合势函数十分有效,也为将来在材料发现中更常规地应用ML势函数铺平了道路。

3 问题及展望

机器学习在材料科学工程领域的初步应用中因处理复杂体系的灵活性、准确性以及良好的泛化能力,带来了与传统实验和计算模拟手段完全不同的研究视角,展现出巨大的潜力,然而,机器学习作为材料科学领域一种新的研究手段,其发展仍存在一些问题及限制因素。

首先,限制机器学习技术在材料学研究及应用的最主要因素就是缺乏有效数据集,机器学习作为一种数据驱动技术,对数据的依赖性比较强,但材料科学领域中的数据具有获取成本高、过于集中或分散、缺乏统一处理标准等特点,获取一个数据量大、分布均匀、特征参量完整且匹配的数据集往往是非常困难的。一方面,由于材料学领域的学者们往往从事某一方向的研究,实验数据集中,又或者研究方向相同但是实验条件差别较大,数据集容量小,容易出现过拟合现象;另一方面,将不同实验条件下得到的数据进行统一,或者将计算结果与实验测量值进行匹配也极其困难,这是因为计算模拟数据无法完全模拟特定的实验条件,而实验数据中又往往缺失与结构和能量相关的特征参数。建立有效数据集最重要的一点就是要求所有数据具有统一标准,即数据集的实验体系和实验条件相同、特征变量维数相同、输入与输出相对应等。但是在目前的材料科学研究中,绝大部分数据集都是通过收集实验数据而获得,收集和处理满足研究需求的实验数据往往要花费大量的时间,由此带来的问题就是其前端建立数据集的低效性掩盖了后端数据处理的高效性,使得现阶段机器学习技术在材料领域的发展受到一定的限制。

其次,尽管机器学习方法已在新材料的发现、成分及结构设计、性能预测等方面展现出巨大的潜力和优势,但是利用机器学习建立的模型是否具有实际的物理和化学意义还需进一步探究。机器学习的过程是一个进行数学运算的黑盒子,通过计算得到的数据间隐藏的关联性和规律是否能够真实反映材料本身的属性,还需要通过大量的事实验证,从而建立更精确的"描述符"来表示这种关系。因此,机器学习技术目前只能进行一定的探索性工作,给材料科学研究提供了新的思路及新的研究方法,但是无法替代传统的实验研究。

当然,机器学习技术带给材料科学领域的更多的是机遇,在未来的发展中,机器学习在材料科学中的应用热点可能将集中在以下几个方面:

(1)"大数据"材料。大数据是当前的一个热门话题,在各个领域都受到了广泛的关注[58-59],如何存储、管理和分析海量数据是材料科学研究和其他领域都亟待解决的难题。因此,研究大数据背景下的机器学习在材料科学中的应用是未来一个重要的研究方向。随着"材料基因组"概念的不断深化,准确全面地表征材料"成分—结构—性能"间的关系是研究和开发新材料的关键,只有真正掌握材料的本质特征,才能更好地应用,甚至实现真正的材料"按需设计"。目前已有的研究虽然还处于依靠有限的数据进行探索的阶段,但是也证明数据驱动型材料科学的研究是非常有效的。随着全世界对材料学数据库的不断重视、整合和完善,越来越多数据资源可供使用,如基于实验获得的无机晶体材料数据库(ICSD)[60]、剑桥结构数据库(CSD)[61]、基于理论模拟的量子材料数据库(OQMD)[62]、AFLOW[63]等都在不断丰富及开源化,可以极大地推动材料"大数据"进程。特别地,近些年广泛兴起的深度学习技术在处理大量数据方面表现良好,在图像处理,尤其是微观尺度的结构表征等领域取得了相当大的突破,建立适合的材料图像"描述符"数据库,以更好地探索材料组织结构。

(2)少数据机器学习。机器学习方法通常需要大量的数据才能有效地学习,但在现阶段的材料科学领域,实验数据稀疏,获取困难,速度缓慢,采用一些新型机器学习算法有助于提高有限数据下的学习效率和精确度。在数据有限的情况下,可以使用数据增广技术,如风格迁移、元学习等来产生相似但又不同的训练样本,增加数据集容量。此外,例如调整神经网络、模仿学习以及贝叶斯框架等新技术[6],可以在有限数据的情况下一次性解决学习问题,达到人类的水平,此类技术适用解决外延性问题,这在高分子材料和材料微观结构预测及模拟方面有着极大的潜力。

(3)材料机器学习与高通量计算相结合。高通量计算也是目前计算材料领域的热点之一,机器学习和高通量计算在本质上具有一定的相似性,都是从大量数据中提取有价值的信息,且具有可并行、可扩展的特点。但是,高通量计算更加偏向于计算,即按照设定好的规则完成指定的工作,可以进行数据的组合和筛选,具有较高的"自动化"程度,不具备外延性和泛化能力,相当于一台高效的计算机。而机器学习技

术则是模仿人类的思维模式，其算法本身带有决策性，有良好的外延性和泛化能力，偏向于“智能化”，相当于经验丰富且非常高效的科学家。将高通量技术与机器学习技术相结合，利用高通量技术参数标准化和体量大的优势，解决机器学习技术的前端问题，结合互补，扬长避短，有望进一步提升新材料的筛选和研发工作效率。

数据的本质是生产资料和资产，它不是社会生产的“副产物”，而是可被二次乃至多次加工的原料，从中可以探索更大的价值，机器学习的作用就是从已有的数据中挖掘有效信息，获取新的价值。数据驱动下的材料科学研究是一个充满无限可能的新方向，它颠覆了传统研究方式，给人们了解材料物质背后的潜在规律提供了新的途径，机器学习在材料科学与工程领域的应用还只是刚刚开始，未来将有无限可能。

参考文献

1 Jordan M I, Mitchell T M. *Science*, 2015, 349, 255.
2 Butler K T, Davies D W, Cartwright H, et al. *Nature*, 2018, 559, 547.
3 Liu H, Xu C, Liang J. *Physics of Life Reviews*, 2017, 21, 233.
4 Cambria E, White B. *Computational Intelligence Magazine IEEE*, 2014, 9(2), 48.
5 Shin H C, Roth H R, Gao M, et al. *IEEE Transactions on Medical Imaging*, 2016, 35(5), 1285.
6 Criminisi A. *Medical Image Analysis*, 2016, 33, 91.
7 Yang K K, Wu Z, Arnold F H. *Nature Methods*, 2019, 16(8), 687.
8 Sean E, Ana C, Kimberley M, et al. *Nature Materials*, 2019, 18(5), 435.
9 Cully A, Clune J, Tarapore D, et al. *Nature*, 2014, 521(7553), 503.
10 Schmidt J, Marques M, Botti S, et al. *npj Computational Materials*, 2019, 5, 83.
11 Feng N, Wang H J, Li M. *Information Sciences*, 2014, 256, 57.
12 Ramprasad R, Batra R, Pilania G, et al. *npj Computational Materials*, 2017, 3(1), 54.
13 Jain A, Hautier G, Ong S P, et al. *Journal of Materials Research*, 2016, 31(8), 977.
14 Yu H K. *International Statistical Review*, 2001, 69(3), 385.
15 Shakhnarovich G, Darrell T, Indyk P. *Nearest-neighbor methods in learning and vision: theory and practice*, MIT Press, Boston, 2005.
16 Burges C J. *A tutorial on support vector machines for pattern recognition*, Kluwer Academic Publishers, Dordrecht, 1998.
17 Safavian S R, Landgrebe D. *IEEE Transactions on Systems Man & Cybernetics*, 2002, 21(3), 660.
18 Zhu J F. Design and application of greenhouse environmental regulation rules based on decision tree. Master's Thesis, Zhejiang University, China, 2017(in Chinese).
朱剑峰.基于决策树的温室环境调控规则设计及其应用研究.硕士学位论文，浙江大学，2017.
19 Dempster A P, Laird N M, Rubin D B. *Journal of the Royal Statistical Societ, Series B: Methodological*, 1977, 39(1), 1.
20 Goh G B, Hodas N O, Vishnu A. *Journal of Computational Chemistry*, 2017, 38(16), 1291.
21 Lecun Y, Bengio Y, Hinton G. *Nature*, 2015, 521(7553), 436.
22 Wu W, Sun Q. *Scientia sinica Physica, Mechanica & Astronomica*, 2018, 48(10), 58(in Chinese).
吴炜，孙强. *中国科学: 物理学力学天文学*，2018，48(10)，58.
23 Liu Y, Zhao T, Ju W, et al. *Journal of Materiomics*, 2017, 3(3), 159.
24 Lin Q Y. *Journal of Chongqing University of Technology (Natural Science)*, 2019, 33(10), 121(in Chinese).
林倩瑜. *重庆理工大学学报(自然科学)*，2019，33(10)，121.
25 Aritra C, Elizabeth K, Bülent Y, et al. *Computational Materials Science*, 2016, 123, 176.
26 Corey E J, Wipke W T. *Science*, 1969, 166(3902), 178.
27 Zhang H, Li N, Li K, et al. *Acta Crystallographica*, 2010, 63(6), 812.
28 Li C, Lu X, Ding W, et al. *Acta Crystallographica*, 2010, 64(6), 702.
29 Xu Q, Li Z, Liu M, et al. *Journal of Physical Chemistry Letters*, 2018, 9, 6948.
30 Sun Z T, Li Z Z, Cheng G J, et al. *Chinese Science Bulletin*, 2019(32), 3270(in Chinese).
孙中体，李珍珠，程观剑，等. *科学通报*，2019(32)，3270.
31 Raccuglia P, Elbert K C, Adler P D, et al. *Nature*, 2016, 533, 73.
32 Kumar J N, Li Q, Tang K Y, et al. *npj Computational Materials*, 2019, 5, 73.
33 Pillong M, Marx C, Piechon P, et al. *CrystEngComm*, 2017, 19, 3711.
34 Goodfellow I J, Pouget-Abadie J, Mirza M, et al. *Generative Adversarial Networks*, 2014, 3, 2672.
35 Attarian Shandiz M, Gauvin R. *Computational Materials Science*, 2016, 117, 270.
36 Fujimura K, Seko A, Koyama Y, et al. *Advanced Energy Materials*, 2013, 3(8), 980.
37 Jie J S, Hu Z X, Qian G Y, et al. *Science Bulletin*, 2019, 64(9), 612.
38 Gu G X, Chen C T, Buehler M J. *Extreme Mechanics Letters*, 2017, 18, 19.
39 Zeng S, Zhao Y, Li G, et al. *npj Computational Materials*, 2019, 5, 84.
40 Zhou Z, Zhou Y, He Q, et al. *npj Computational Materials*, 2019, 5, 128.
41 Liu B. Development and research of a magnesium alloy expert system. Ph. D. Thesis, Chongqing University, China, 2011(in Chinese).
刘彬.镁合金专家系统的开发研究.博士学位论文，重庆大学，2011.
42 Wang C, Fu H, Jiang L, et al. *npj Computational Materials*, 2019, 5, 87.
43 Tagade P M, Adiga S P, Pandian S, et al. *npj Computational Materials*, 2019, 5, 127.
44 Liu J, Mohan A, Kalia R K, et al. *Computational Materials Science*, 2020, 173, 15.
45 Pokuri B S, Ghosal S, Kokate A, et al. *npj Computational Materials*, 2019, 5, 95.
46 Ament S E, Stein H S, Guevarra D, et al. *npj Computational Materials*, 2019, 5, 77.
47 Saito Y, Shin K, Terayama K, et al. *npj Computational Materials*, 2019, 5, 124.
48 Chan H, Cherukara M, Loeffler T D. *npj Computational Materials*, 2020, 6, 1.
49 Orme A D, Chelladurai I, Rampton T M, et al. *Computational Materials Science*, 2016, 124, 353.
50 Tong Z, Wang L, Zhu G, et al. *Metallurgical and Materials Transactions A*, 2019, 35(9), 2719.
51 Gao Q. Study on defect detection of rolled sheet for magnesium alloy based on deep learning. Master's Thesis, University of Science and Technology Liaoning, China, 2018(in Chinese).
高青.基于深度学习的镁合金轧制薄板缺陷检测问题的研究.硕士学位论文，辽宁科技大学，2018.
52 Attarian Shandiz M, Gauvin R. *Computational Materials Science*, 2016, 117, 270.
53 Terentyev D, Bonny G, Castin N, et al. *Journal of Nuclear Materials*, 2011, 409(2), 167.
54 Shuaihua L, Qiong H. *Nature Communications*, 2018, 9, 1.
55 Bernstein N, Csanyi G, Deringer V L, et al. *npj Computational Materials*, 2019, 5(1), 99.
56 Hongxiang Z, Ghanshyam P, Xiangdong D, et al. *npj Computational Materials*, 2018, 4(1), 48.
57 Cooper A, Kstner J, Urban A, et al. *npj Computational Materials*, 2020, 6, 54.
58 Tian J, Zhu D J, Yang W H, et al. *Computer Science*, 2018, 45(11A), 58(in Chinese).
田娟，朱定局，杨文翰，等. *计算机科学*，2018，45(11A)，58.
59 Liu R, Zhang N. *Computer Science*, 2021, 48(7), 137(in Chinese).
刘荣，张宁. *计算机科学*，2021，48(7)，137.
60 Belsky A, Hellenbrandt M, Karen V L, et al. *Acta Crystallographica*, 2010, 58(3), 364.
61 Allen F H. *Acta Crystallographica Section B Structural Science*, 2002, 58, 380.
62 Kirklin S, Saal J E, Meredig B, et al. *npj Computational Materials*, 2015, 1, 15010.
63 Curtarolo S, Setyawan W, Hart G L W, et al. *Computational Materials Science*, 2012, 58, 218.
64 Shaikhina T, Khovanova N A. *Artificial Intelligence in Medicine*, 2017, 75, 51.

(责任编辑 杨 霞)

Xiaoxi Mi, a doctoral candidate, graduated from Chongqing University in 2018 with a master's degree. Now he is a doctoral candidate in Chongqing University, with the guidance of Professor Aitao Tang. The major research is the application of machine learning in magnesium alloys.

米晓希,博士研究生,2018 **年毕业于重庆大学获得硕士学位。现在为重庆大学博士研究生,在汤爱涛教授的指导下进行研究,主要从事基于机器学习的镁合金组织与性能的研究。**

Aitao Tang, Ph. D., professor, doctoral supervisor, key researchers of National Engineering research center for magnesium alloys. Focusing on magnesium alloy, aluminum alloy and composite materials. Mainly engaged in material database, material simulation and high performance materials research. In 1984, she graduated from the department of metallurgy of Chongqing University. In 2004, she received her Ph. D. degree at the School of Materials of Chongqing University. She served as the teaching staff of five undergraduate courses and one postgraduate course successively. She is the backbone teacher of the applied course of computer in material science and engineering. She has obtained a number of national authorized invention patents, published more than 60 important theses.

汤爱涛,博士,教授,博士研究生导师,国家镁合金材料工程技术研究中心骨干研究人员。以镁合金、铝合金和复合材料为重点,主要从事材料数据库、材料的计算模拟以及高性能材料的研究。1984 **年本科毕业于重庆大学冶金系,**2004 **年博士毕业于重庆大学材料学院,先后担任了五门本科课程和一门研究生课程的教学工作,是"计算机在材料科学与工程中的应用课程"的骨干教师。获得多项国家授权发明专利,发表重要论文** 60 **多篇。**

Fusheng Pan received his Ph. D. degree in Northwestern Polytechnical University. He is an academician of the Chinese academy of engineering, a professor of materials science at Chongqing University and a doctoral supervisor. In 1977, he joined the Jiusan Society, a member of the 11th CPPCC national committee, a member of the discipline evaluation group of the degree committee of the state council, and a member of the central committee of the Jiusan Society. He has served successively as lecturer, associate researcher, deputy director of the department, director of the institute, vice dean of the graduate school and director of the institute of light metals of Chongqing University. He has studied in Oxford University, Chiba University in Japan and national materials institute in Germany. He has published more than 180 papers and 7 books in important journals at home and abroad.

潘复生,中国工程院院士,西北工业大学工学博士,重庆大学材料学科教授,博士研究生导师。1977 **年参加工作,**1993 **年加入九三学社,第十一届全国政协委员,国务院学位委员会学科评议组成员,九三学社中央委员。历任重庆大学讲师、副研究员、系副主任、研究所所长、研究生院副院长、轻金属研究院院长等职。曾留学英国牛津大学、日本千叶大学和德国国家材料研究所。已在国内外重要刊物发表论文** 180 **余篇,出版著作** 7 **部(本)。**