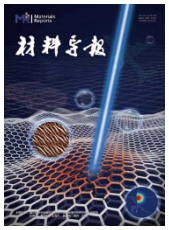

材料导报
Materials Reports
ISSN 1005-023X,CN 50-1078/TB

《材料导报》网络首发论文

题目：机器学习在材料信息学中的应用综述
作者：牛程程，李少波，胡建军，但雅波，曹卓，李想
网络首发日期：2020-08-28
引用格式：牛程程，李少波，胡建军，但雅波，曹卓，李想．机器学习在材料信息学中的应用综述．材料导报.
https://kns.cnki.net/kcms/detail/50.1078.TB.20200828.1449.004.html

网络首发时间：2020-08-28 16:02:45
网络首发地址：https://kns.cnki.net/kcms/detail/50.1078.TB.20200828.1449.004.html

机器学习在材料信息学中的应用综述

牛程程[1]，李少波[1,2]，胡建军[2, 3,*]，但雅波[2]，曹 卓[2]，李 想[2]

（1 贵州大学现代制造技术教育部重点实验室，贵州 贵阳 550025；

2 贵州大学机械工程学院，贵州 贵阳 550025；

3 美国南卡罗来纳大学计算机科学与工程系，南卡罗莱纳，哥伦比亚 29208 美国）

面对巨大的材料设计空间，基于理论研究、实验分析以及计算仿真的传统方法已经跟不上高性能新材料的发展需求。近年来材料数据库与机器学习的结合带动了材料信息学的进步，推动了材料科学的发展。当前，运用数据驱动的机器学习算法，建立材料性能预测模型然后应用于材料筛选与新材料开发的研究得到越来越多的应用。利用机器学习框架搭建材料研究设计平台对材料大数据资源进行分析与预测成为了开发新型材料的重要手段。

将机器学习运用于材料科学面临的一系列困难，包括根据预测对象确定材料特征的计算或自动抽取、不同精度的实验与计算数据的获取与预处理；选取或者开发合适的机器学习预测模型和训练算法；估计预测效果与预测性能的可靠性；处理材料机器学习问题所独有的小数据、异构数据、非平衡数据等特性。目前研究的焦点是针对不同的材料性能，收集相关的数据集、基于物理原理构造特征表示来训练机器学习模型并将机器学习的最新技术用于材料信息学。

现阶段机器学习已经被应用于光伏、热电、半导体、有机材料等几乎所有材料设计领域。通过采用机器学习算法训练材料性能的预测模型并用于筛选现有材料数据库或者搜索新的材料，大大加快了新材料发现的过程。目前，国内外科学家借助统计推理与机器学习算法开展了一系列的研究，开发适合预测不同材料属性的多种材料表征方法，应用了包括深度学习、贝叶斯网络等最新机器学习与人工智能方法，在多类功能材料设计领域取得了突破性的成果。

本文主要介绍了机器学习方法在材料性能预测中的相关研究与应用，包括目前最常用的材料数据库资源、多种适用的机器学习算法及应用实例，以及机器学习在材料性能预测中遇到的常见问题。最后对国内外的材料信息学发展现状进行了概括并对未来的发展进行了展望。

关键词：材料信息学 材料科学 材料性能 机器学习 大数据

通信作者：胡建军，jianjunh@cse.sc.edu

中图分类号：TP181; TB302 **文献标识码：**A

Application of Machine Learning in Material Informatics: A survey

NIU Chengcheng[1], LI Shaobo[1,2], HU Jianjun[2,3], DAN Yabo[2], CAO Zhuo[2], LI Xiang[2]

1 *Key Laboratory of Advanced Manufacturing Technology (Ministry of Education), Guizhou University, Guiyang 550025, China;*

基金项目：国家自然科学基金 (51741101)
This work was supported by the National Science Foundation of China (51741101).

2 School of Mechanical Engineering, Guizhou University, Guiyang, 550025, China;

3 Department of Computer Science and Engineering, University of South Carolina, SC, 29208, USA

In the face of huge material design space, the traditional methods based on theoretical research, experimental analysis and computational simulation cannot keep up with the development of new materials with high performance. In recent years, the combination of material database and machine learning has led to the progress of material informatics and the development of material science. At present, the material performance prediction model is established by the application of data-driven machine learning algorithm, and then applied to the research of material screening and new material development research has been more and more applications. Using the machine learning framework to build the material research and design platform to analyze and predict the material big data resources has become an important means to develop new materials.

Machine learning is applied to a series of difficulties faced by materials science, including the calculation or automatic extraction of material characteristics according to the predicted objects, the acquisition and preprocessing of experimental and computational data with different precision, the selection or development of appropriate machine learning prediction models and training algorithms, estimating the reliability of prediction effect and predictive performance, and deal with material machine learning problems with the unique characteristics of small data, heterogeneous data and unbalanced data. At present, the focus of research is to collect relevant data sets, construct feature representations based on physical principles to train machine learning models and apply the latest techniques of machine learning to material informatics for different material properties.

Machine learning has been used in photovoltaic, thermoelectric, semiconductor, organic materials and other materials design fields. The process of new material discovery is greatly accelerated by using machine learning algorithm to train the prediction model of material performance and to screen existing material database or search new material. At present, scientists at home and abroad have carried out a series of research with the help of statistical reasoning and machine learning algorithms, developed a variety of material characterization methods suitable for predicting the properties of different materials, and applied the latest machine learning and artificial intelligence methods, including deep learning, Bayesian network, etc. Breakthrough achievements have been made in the field of multi-functional material design.

This paper mainly introduces the related research and application of machine learning methods in material performance prediction, including the most commonly used material database resources, various applicable machine learning algorithms and application examples, and the common problems encountered by machine learning in material performance prediction. Finally, the development status of material informatics at home and abroad is summarized and the future development is prospected.

Keywords: materials informatics, material science, material properties, machine learning, big data

Correspondence: Jianjun Hu, jianjunh@cse.sc.edu

0 引言

竞争日益激烈的制造业形势和经济的快速发展向材料科学家和工程师提出了挑战：如何使用已有经验缩短新材料从发现到应用的研发周期，以满足工业领域对更高性能材料的需求，从而在关键领域取得竞争优势。近几年来，随着材料数据的积累，数据挖掘[1, 2]与机器学习[3]在材料学研究设计平台搭建和基于大数据的材料分析与预测方面得到了越来越多的应用[4]。在新材料发现方面，机器学习算法已经被用于多种关键材料的研究并取得了令人瞩目的成效，如新能源材料(2017) [5]、软材料(2018) [6]、聚合物电介质[7]、钙钛矿材料[7-12]、压电材料[13]、催化剂[14, 15]、感光材料[12]等。其中，日本国家材料科学研究所的Takahashi[16]等人使用高通量第一性原理(Density Function Theory, DFT)[17, 18]计算得到了15 000个ABC2(C1,C2)D型钙钛矿材料的带隙值，随后利用机器学习方法训练得到了材料的带隙值预测模型并对一些钙钛矿材料进行筛选，发现了诸多新的高性能钙钛矿材料。上述研究正是材料信息学在材料学研究中的优势所在。材料信息学[19]是应用信息学方法解决材料学问题的一门学科，本文将着重综述机器学习方法在材料信息学方面的应用，特别是其在传统材料学研究方法之上所发挥的重要作用。

目前，对于新材料的研发主要采用研究者对材料的直觉判断和大量的“试错法”实验来进行，其效率较低且难以有效发现大量可能的新材料组合。另外，基于第一性原理(DFT)与分子动力学仿真计算方法对材料进行的研究也已经较多，但目前此类方法在快速计算时其准确性有限，若需获取高精度的计算结果其计算量惊人，难以高效的用于材料的筛选[20]与新材料的发现。然而，使用实验方法和第一性原理(DFT)计算得到的材料数据在不断的积累，且大多已作为材料数据库进行存储，这些材料数据虽然没有直接地被用于新材料的发现，其中仍存在着一定的材料信息与规律，若将其和机器学习等信息学方法有效结合，则可实现对基于材料性能的快速新材料预测和筛选，从而能够有效的发现大量候选新材料，在缩小目标范围后配以实验方法和第一性原理计算方法进行验证，则可加快新材料的研发过程。这一过程中，机器学习方法可成为重要的技术手段。

近几年，国内在该领域也开展了一系列研究，主要包括上海材料基因工程研究院、中国科学院宁波材料基因研究院、北京大学深圳研究生院新材料学院、清华大学深圳研究生院、北京计算科学研究中心、南方科技大学、电子科技大学等科研单位与院校。其中，上海材料基因组工程研究院在张统一[21]院士的带领下，将新材料的发现与应用作为重点，在建设材料基因数据库、集成计算与软件开发、高通量材料制备与表征、服役与失效机理等方面进行了大量工作并取得了一定的进展；国际高通量“组合材料芯片”技术的发明人、宁波国际材料基因工程研究院项晓东[22-24]研究员（现为南方科技大学教授）进行了高通量组合材料实验及原位实时高通量组合材料实验技术研究；中国科学院物理研究所陈立泉[22, 25-27]院士所领导的团队对自动化高通量计算方法进行了研究并开发了相关软件平台且已获得著作权，在超级计算机上将快速“键价和”法与高精度第一性原理分子动力学方法相结合，从众多材料数据库中收集得到了 20 万余条无机晶体数据，并对其中含锂材料的电子结构、三维离子导电通道和离子迁移活化能等进行计算，最终建立起了包含电解质与电极材料的数据库，并通过数据挖掘技术对新的固体电解质材料进行了一定的筛选。但从总体来看，我国与美国等率先开展材料研究新模式的国家仍然存在着较大的差距，因此，使用机器学习方法与众多材料数据相结合以加速新材料的设计与发现的研究方法需要引起更多国内材料学研究者的注意。

最近，统计推理和机器学习算法逐渐广泛的应用于新材料的发现与设计[28]中，其旨在建立一个以计算模拟和理论预测优先、实验验证在后的新型材料学研究方法体系，从而取代现有的以经验和实验为主的传统材料研发模式。当前材料信息学中的关键问题主要集中在材料的表征方法[29]以及基于机器学习与深度学习的材料性能预测模型的建立与优化[7]等方面。本文主要对材料信息学中无机化合物材料性能预测方法、相关材料数据库和常用的机器学习模型[3]等进行了综述。

1 无机化合物材料性能预测方法

材料学研究中的重要问题就是如何从无穷多种可能的元素组合中找到满足特定性能要求的材料。限于材料结构与功能关系的复杂性以及计算资源、实验成本等的局限性，如今，仅通过传统的理论研究、实验研究及计算仿真方法对新材料的发现已经跟不上众多新技术对高性能新材料的需求，高性能新材料的研发速度成为了众多其它领域技术发展的瓶颈，然而，通过数据驱动[1]的方法使用机器学习建立起材料性能预测模型用于未知性能材料的筛选，可有效提高高性能新材料的研发速度。

由于机器学习算法的发展及其在众多领域取得的巨大成功，信息学方法也开始逐渐应用于材料学研究中。很多材料性能因难以运用传统方法进行测量或计算，数据驱动下的材料性能预测模型开始发挥愈加重要的作用，这些方法基于已有的材料数据使用机器学习算法对材料性能进行预测并具有较快的预测速度和较高的精确度。材料性能的预测问题主要包括材料数据的获取、材料的表征方法、基于机器学习方法的预测模型的选择等内容[2]。如使用神经网络算法利用数字指纹（又称材料描述符，Descriptor）[30]的材料表征方法下的材料数据对材料的导电率进行预测，通过对数据的学习建立起了材料描述符与材料性能间的映射关系。另外，将机器学习用于双钙钛矿材料的带隙预测[7]中，可避免第一性原理计算时所需的巨大计算量。Deml等人[17]通过第一性原理计算得到了材料的总能量并生成焓的线性模型最终发现了影响这些性能的主要因素。上海大学材料基因组研究院使用基于材料局部结构特性的“中心-环境(center-environment, CE)”模型的数据挖掘方法与高通量第一性原理密度泛函方法（GGA-PBE 泛函）计算结合，对镍基单晶高温合金掺杂元素的置换能和几何结构进行了预测，并用于其它多元合金掺杂元素的能量稳定性和几何结构问题的研究中。常用于预测的材料性能以及所使用的材料描述符分别如表 1 和表 2 所示。

表 1　材料性能

Table 1　Material properties

性能类别	性能指标	相关文献
电子性能	形成能　Formation energy	[31]
	压电性　piezoelectric properties	[13]
	带隙　Band gap / E_g	[32]
	绝对能　Absolute energy	[10]
	自由能　Free energy	
	缺陷能　Defect energetics	
	费米能　Fermi energy	[7]
	电容率　Dielectric constant / ε_e	
力学性能	体积模量　Bulk moduli / *M (N/m^2)*	[33]

	剪切模量　Shear moduli / *G(N/m^2)*	
	弹性模量　Elastic moduli　/ *E(N/m^2)*	
	泊松比　Poisson ratio / μ	
物理化学性能	密度　Density / ρ *(Kg/m^3)*	[7]
	最近邻距离　Nearest neighbor distance / *Nnd*	
	生成焓　Formation enthalpy /*(KJ/mol)*	[34, 35]
	熔点　Melting temperature / T_m	[36]
	热导率　Thermal conductivities / λ	[37]
	催化活性　Catalytic activities	[38, 39]
	耐辐射性　Radiation resistance	[40]
	超导材料临界温度 Critical temperature	[41, 42]

表2　常用材料描述符

Table 2　Common material descriptors

分类	预测特征	相关文献
总水平分类	总水平的描述符与绝缘子带隙(band gap)的预测有关 总水平的描述符可以对原子尺寸、晶体结构、电负性、及两金属元素氧化态进行预测	[7]
分子片段/模式分类	利用第一性原理(DFT)预测机械性能	[43, 44]
	通过库伦矩阵(Coulomb matrix)来预测晶体结构形式	[45, 46]
	通过高通量计算进行数据管理及数据挖掘方法进行分析和预测	[14, 47]
原子级分类	采用对称函数(Symmetric function)的指纹映射到神经网络	[48, 49]
	根据高斯近似势(GAP)框架预测相邻原子密度谱	[50]
	根据 SOAP(smooth overlap of atomic positions)以适应基于量子力学数据的原子间相互作用	[51]

2 相关材料数据库

材料信息学的一个核心内容就是材料数据库的建立。材料数据是材料信息技术和材料基因工程开展的基础，但材料数据较难、较慢的获取是材料学领域一直以来存在的问题，因此，通过高通量计算、高通量实验和材料数据库等技术加快材料数据积累和新材料的研发成为了新的研发模式，也是材料基因组计划的核心内容之一。

2011 年 6 月美国启动了“先进制造业伙伴关系”(Advanced Manufacturing Partnership, AMP)计划，该计划呼吁政府与高校和企业之间加强合作，而“材料基因组计划”[52](Materials Genome Initiative, MGI)正是 AMP 计划中的重要组成部分，材料基因组计划的相关研究开始在全球多个国家蓬勃开展。在中国工程院和科学院的积极推动与广大材料领域研究者的努力下，2015 年科技部设立了“材料基因工程关键技术与支撑平台”（简称“材料基因工程”[53]）重点专项。材料基因工程项目将材料的模拟计算、实验表征和数据库技术作为重点，使用高通量-多尺度集成计算、高通量组合材料实验和数据挖掘技术，以实现低耗高效的材料研发过程。中国科学院金属研究所开发了材料数据库查询系统(http://www.materials.csdb.cn/SDB)，其中包含了大量高温合金、钛合金和腐蚀材料的相关材料数据。另外，由国家先进材料网络和信息中心建立的材料数据服务平台——材料信息网(http://www.materials.gov.cn/)，包含了一些特殊的材料数据库，如材料企业和产品数据库等。

随着材料基因组计划的实施，为了能够快速推进材料学发展，以材料数据共享为理念的数据库不断涌现，大大降低了材

料大数据技术中对材料数据的获取难度。这些数据库包括无机材料晶体结构数据库(ICSD)、剑桥晶体结构数据库(CSD)、第一性原理计算材料数据的 Materials Project[54]以及具有超过一百万种不同材料和一亿多种材料性能性能数据的 Aflow[55, 56]数据库等。表 3 列出了材料信息学中常用的材料数据库的名称和其包含的材料数据种类。这些数据库中，有的可以直接进行材料数据下载，有的数据库提供 REST API[57]接口以供批量下载。例如，Materials Project 数据库作为材料研究领域的谷歌，其所包含的材料数据可以通过 API[58]下载获取，该数据库通过计算机建模和机器学习技术，可以方便的建立材料性能预测模型以用于材料筛选。使用这种预测模型可使得材料计算速度得到极大提升，为加快新材料的研发和设计提供了有力支撑。从事于应用计算材料设计和储能材料研究的美国工程院院士 CEDER 等利用数据挖掘、高通量计算等技术加速了传统材料下的工业发展，并已经在新材料设计领域取得一些成功案例[59-62]。

表 3 材料数据库

Table 3 Material database

数据库名称	材料数据种类	网址
无机材料晶体结构库 (Inorganic Crystal Structure Database, ICSD)	该库提供除了金属和合金以外、不含 C–H 键的所有无机化合物晶体结构信息，包括化学名和化学式、矿物名和相名称、晶胞参数、空间群、原子坐标、热参数、位置占位度、R 因子及有关文献等各种信息。	http://www.fiz-karlsruhe.de/ icsd.html
剑桥晶体结构数据库 (Cambridge Structural Database, CSD)	数据库含有 875 000 多个有机及金属有机化合物的 X 射线和中子射线衍射的分析数据。它只负责收集并提供具有 C-H 键的所有晶体结构，包括有机化合物、金属有机化合物、配位化合物的晶体结构数据。	http://www.ccdc.cam.uk/
金属和合金晶体数据库 (Metals and Alloys Crystallographic Database, CRYSTMET)	CrystMet 数据库包含金属、合金和金属间化合物的晶体学信息。里面收集了 1913 年以来金属单质、金属化合物和固溶体的晶体数据，包括金属元素与硼、硫、硅、锗等元素的化合物。	http://crystalworks.ca/
开放晶体结构数据库 (Crystallography Open Database, COD)	COD 是储存晶体学数据、原子坐标参数以及详细的化学内容和参考文献的数据库。它对所收集的大量分子结构数据进行了全面、广泛的整理、核对和质量评价，因此它所提供的数据要比原始文献更为准确。可以方便地检索、筛选和进行系统的分析，还可对数据进行加工并绘成各种规格的图形。	http://www.crystallography.net/cod/
沸石结构数据库 (the Database of Zeolite Structures, DZS)	DZS 提供了所有沸石骨架类型材料的结构信息,包括每个框架式的说明和图纸、晶体学数据和代表性材料模拟粉末衍射图案、建筑模型的详细说明、无序沸石结构的的描述等。	http://www.iza-structure.org/databases/
Paulingfile 数据库 (Paulingfile Database)	一个集相图、晶体结构和物理性质的无机化合物数据库。	http://paulingfile.com
材料计划 (Materials Project)	该库专门用于搜索查找各种材料的性质，以较高的标准衡量是否将计算机预测的材料纳入数据库。例	http://materialsproject.org

	如：锂电池相关(约 15 000 个结构)；沸石、金属有机骨架 MOF(约 13 万种)。	
材料云项目 (Materials Cloud Project)	该库主要以石墨烯等二维材料为主，初步预测产生 1 500 种可能的二维结构。	http://www.materialscloud.org
NIMS 材料数据库 (NIMS Materials Database)	世界上最大的聚合物、陶瓷、合金、超导材料、复合材料和扩散材料数据库之一。	https://www.nist.gov/
AFLOWlib 数据库 (Automatic-FLOW for Materials Discovery, Aflowlib)	AFLOWlib 是目前最大的材料数据库，主要是金属合金，该库拥有超过一百万的不同材料和一亿左右的性能属性。	http://www.aflowlib.org
开放量子材料数据库 (Open Quantum Materials Database, OQMD)	该库主要以钙钛矿数据居多，用户可以下载整个数据库而不仅仅是单个搜索结果。	http://oqmd.org

3 材料信息学中的机器学习算法与模型

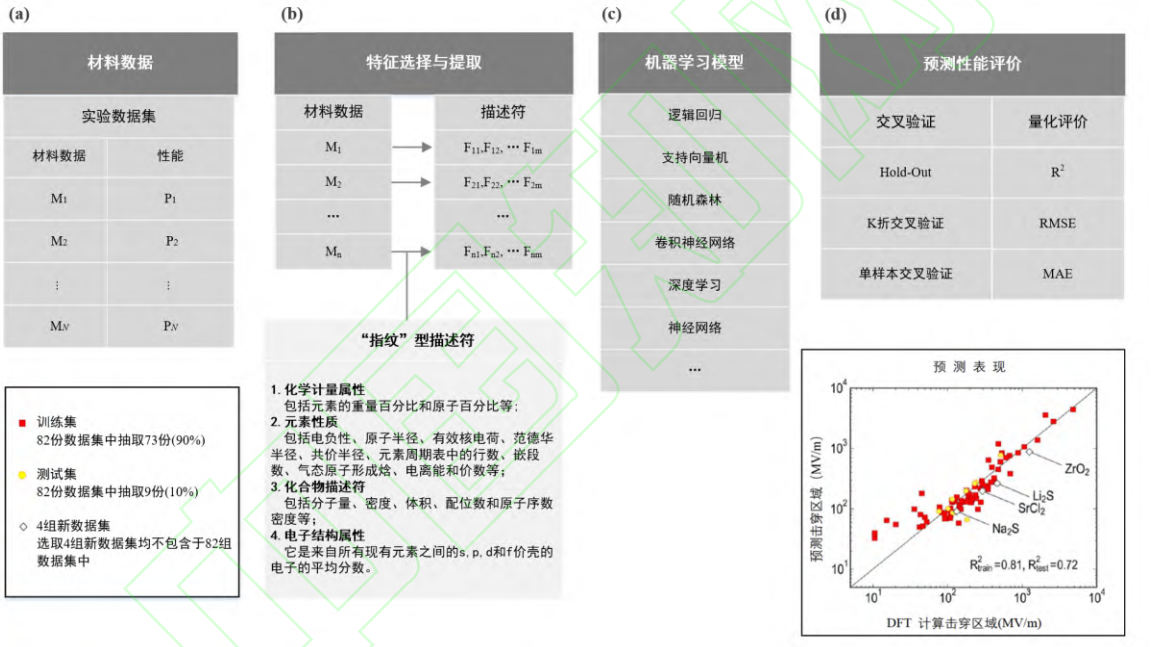

图 1　以建立用于预测绝缘子固有电击穿场现象模型[63]为例来解析材料科学中应用机器学习的关键要素。(a) 收集材料数据并进行编码，将实验数据分为训练集、验证集、测试集；(b) 通过人工定义或自动学习提取材料描述符（指纹），N 和 M 分别是训练样本的数量和描述符的数量，化学描述符（“指纹”）主要具有四大类[64]；(c) 选择机器学习模型；(d) 最终使用交叉验证的方法将模型预测性能与数据集的标准值比较，通过量化评价指标进行预测性能评价。

Figure 1　Taking the establishment of a model[63] for predicting the phenomenon of intrinsic electrical breakdown of insulators as an example to analyze the key elements of applied machine learning in material science. (a) Collect material data and encode it into training set, validation set and test set;(b) Extract material descriptors (fingerprints) by manual definition or automatic learning; N and M are the number of training samples and descriptors, respectively; chemical descriptors (fingerprints) have four main categories[64]; (c) Select machine learning model; (d) Finally use the cross-validation method to compare the predictive performance of the model with the standard value of the dataset, and evaluate the predictive performance through quantitative evaluation indicators.

材料信息学中常见的机器学习相关问题大致可分为有监督学习和无监督学习两种问题，根据预测值是连续值或离散值可分为回归和分类任务，另外常用的机器学习方法还包括特征选择和降维等。由于材料性能值多为连续值，所以回归问题占有主要地位，常用的机器学习模型框架如表 4 所示，其中的相关文献介绍了一些主流算法及其在材料信息学中的应用。

回归分析是在自变量和因变量之间建立出回归方程以描述其相关关系且具有一定的预测功能。回归分析在机器学习问题中有着广泛的应用，如传统线性回归中的拟合分析、Logistic 回归[65]用于分类问题的实现。根据自变量的个数不同，可分为一元回归分析和多元回归分析；根据采用的拟合函数的不同，又可分为线性回归和非线性回归。常用的回归方法还包括 LASSO 回归(Least Absolute Shrinkage and Selection Operator, LASSO) [32, 66, 67]和核岭回归(Kernel Ridge Regression, KRR) [46, 68]。

最近，深度学习已逐渐代替传统的机器学习并在众多领域中有着优异的表现。在学习模型中，深度神经网络（DNN）[69]、卷积神经网络（CNN）[70]、循环或递归神经网络（RNN）[71]等学习模型是较常用的深度学习算法。深度神经网络的灵活性使模型理论上可以从数据最原始的表示中不断学习更高阶的特征。比如，在计算机视觉领域，应用于图像识别的卷积神经网络可以学习如何检测网络中的中间层数据的边缘，最后在终端层中进行检测。卷积神经网络最近已成功地应用于大型图像分类任务，并在其他方面取得了许多成功。在材料信息学任务中，该方法已经应用于预测晶体体系、消光基和空间群的 X 射线衍射(XRD)图谱的分类问题中[72]。最近，卷积神经网络还被成功应用于钢铁珠光体、奥氏体、马氏体等材料图像的识别任务中[73]。例如，由 Ahmet Cecen[74]等人采用三维卷积神经网络来学习材料微观结构的显著特征，使用随机梯度下降(SGD)方法来训练网络并对可靠的结构-性能关系进行探索，以提高所提取的降阶模型的准确性，使得能够获得高对比度复合材料的有效弹性性能；Ward[64]等人使用决策树方法对多维数据进行分析以预测结晶材料的带隙能量和无机材料的 GFA。人工神经网络还被用来探索晶体结构-性能关系以及捕捉非线性吸附底物的相互作用关系[75]；深度学习算法还被用于预测化学式为 AB 型晶体结构和性能关系[7]。另外，在一部分任务中将深度神经网络与传统回归方法相结合以得到更高精度的预测结果，如使用神经网络从材料结构中抽取高层特征然后利用传统的回归模型根据高层特征对材料性能进行预测。

表 4 材料信息学中常用的机器学习模型

Table 4 Machine learning models commonly used in materials informatics

机器学习模型	特色或应用案例	相关文献
支持向量机 (Support Vector Machine, SVM)	无需依赖整个数据，能够处理非线性特征的相互作用，可以提高泛化能力、解决高维问题，但若样本过多时，由于数据敏感度的缺失而导致效率不是很高。	[76, 77]
线性回归 (Linear Regression, LR)	适用于简单的回归问题，用梯度下降法对最小二乘法形式的误差函数进行优化，计算简单训练速度快，但不能拟合非线性数据。	[78, 79]
逻辑回归 (Logistic Regression, LR)	进行分类问题时，计算速度快， 可结合正则化模型来解决问题。	[65, 80]
K 最邻近算法 (K-Nearest Neighbor, KNN)	即可做回归问题也可做分类问题，可用于非线性的分类问题，但当样本不平衡时，可能会导致结果不准确。	[81, 82]
决策树	适合处理含有缺失属性的样本，短时间内能够对大数据源做出较	[64, 83]

(Decision Tree, DT)	好的结果，但容易发生过拟合，忽略数据间的相关性。	
随机森林 (Random Forests, RF)	基于 Bagging 的集成学习方法，可以用来做分类、回归等问题，因为随机性的引入，使得有较好的抗噪声能力且不易过拟合，训练速度快，有极高的准确率，但训练时需较大的空间和时间。	[84, 85]
深度神经网络 (Deep Neural Networks, DNN)	适合非线性函数拟合特征抽取。	[69, 86]
卷积神经网络 (Convolutional Neural Network, CNN)	特别适合于图像模式识别、计算机视觉等领域。应用于描述基于 X 射线衍射的空间群预测。	[70, 74]
循环或递归神经网络 (Recusive or Recurrent Neural Network, RNN)	深度学习主要框架之一，主要是将上一次迭代的输出作为当前迭代的输入，从而实现循环。可用于建立结构和材料性质之间的函数关系。	[71, 87]
（经典）人工神经网络 (Artificial Neural Network，ANN)	用于探索大量的晶体结构和可视化的结构-性能关系。应用于预测小分子力场的电子性质。	[72, 75]
自动编码器 (Auto- Encoder, AE)	主要是通过逐层的无监督学习先将输入数据进行表征的压缩，然后通过网络进行有监督的学习。	[88, 89]
核岭回归 (kernel ridge regression, KRR)	经常用于材料性能的回归分析预测。如带隙和合成焓（能）预测	[46, 68]
压缩感知 (Compressive Sensing, CS)	基于压缩感知的特征选择方法。被用于建立一个定量预测二元化合物半导体晶体结构的物理模型。	[90, 91]

4 机器学习方法在材料信息学中的常见应用问题

机器学习与材料科学的碰撞，为新材料的发现带来了新的机遇，由于材料数据的特殊性以及材料性能预测的复杂性，在应用机器学习进行材料学研究时存在着一些需要注意的关键问题，如训练集与测试集的划分与验证评估方法的选择、过拟合与欠拟合问题的处理、提高模型训练速度、不平衡数据集的处理、材料表征方法的选择（材料描述符的选择）等问题，下面就这些常见问题进行阐述。

4.1 数据集划分方法与模型评价指标

训练与评价机器学习预测模型前对数据集的基本处理过程包括将训练数据集与测试数据集进行有效划分。通常为了评估一个模型的泛化能力，会将数据分成训练集和测试集两部分来训练和测试机器学习模型。目前常用的训练集和测试集的划分方法主要有三种，首先是留出法(Hold-out)，这种方法直接将数据集划分为两个互斥的训练集S和测试集T，在S上训练出模型后，用T来评估其误差，通过若干次随机划分、重复实验评估后取平均值作为留出法的评估结果。但是，训练集和测试集按照什么比例来划分仍是个问题。第二种划分方法是K折交叉验证法(K-fold Cross Validation)[92]，先将数据集D划分为K个数据分布一致的互斥子集，选择其中的一个子集为测试集，剩余的K-1个子集作为训练集。这样可以获得K组训练/测试集，从而可以将数据集进行K次训练和测试，最终将返回的K个测试结果的均值作为测试结果。显然，交叉验证法评估结果的稳定性和保真性在很大程度上取决于K的取值[30]，另外，本次训练中的训练集样本在下一次训练中

可能变成了测试集样本，但是对每一次训练而言，测试集样本和训练集样本是不会重叠的。此外，一个值得注意的问题是现有的机器学习模型的预测性能都因样本冗余和交叉验证方法中测试集数据过少而被过高估计 [93]，主要原因是数据集中包含较多掺杂产生或者高度类似的材料数据，这些相似样本在交叉验证时，很大概率会同时被分配到训练集和测试集，从而导致过高的预测准确度，而实际上的模型泛化能力却不及结果所示，并不能精确的预测出材料的性能，类似由高冗余度样本集导致过高估计的结果在高冗余度的OQMD数据集训练的模型上得到体现[94]，其预测性能严重高估失实。第三种方法则是自助法(Boot Strapping)。该方法采用重复抽样技术从原始样本中抽取一定数量来进行模型评估。这种方法用于数据集较小且难以有效划分训练集和测试集时效果较为突出。由于自助法得到的测试集改变了初始数据集的分布且引入了误差，因此在数据集比较大时应使用留出法和交叉验证法。我们发现机器学习模型在数据集上使用交叉验证法时普遍具有很好的效果，例如，针对121种纯二氧化硅沸石DFT数据集的交叉验证，对GBR(Gradient Boosting Regressor)模型进行3折交叉验证并重复100次，得到了最优的模型结果[95]，KRR模型通过5折交叉验证法对材料带隙进行预测得到了较好的结果[96]。

除此之外，通过利用一种数据分区策略，可以使机器学习模型的预测精度显著提高。通过将数据集分组并在每个子集上训练一个单独的模型，我们可以通过减少每个机器学习模型所需要捕获的材料信息的数量来提高预测的准确性。例如，稳定的金属化合物的物理效应可能与陶瓷不同，因此，在这种情况下，可以将数据划分为只包含金属元素的化合物和同时包括非金属元素的化合物两部分，试验表明，分区数据集可以显著提高预测属性的准确性，分区策略的选取可以根据实际问题的不同，通过自动探索的方法和无监督学习来确定不同材料集合的划分标准，通过对大量可能的划分标准进行尝试，可以最大限度地减少交叉验证时的误差。

4.2 过拟合与欠拟合问题

过拟合[97]是指机器学习模型在训练集数据上的误差较小但在训练集以外的数据样本上的误差较大，如果训练数据样本的数量较少或样本差异较小就易产生过拟合现象，同时，过拟合也可能在无参数非线性模型中发生，例如，决策树[83]就是一种无参数机器学习算法，其具有弹性且易受过拟合的影响，因此可根据“没有免费午餐定理”(No Free Lunch Theorem）[98] 在给定的数据集上进行相关的算法寻优解决过拟合的问题，常用的解决方法[99] 包括在目标函数中加入正则化项[100]、“Dropout”方法、扩充数据集[30]、“Early stopping”方法等。其中，L1、L2 正则化（机器学习中一种常用的技术，其主要目的是控制模型复杂度，减小过拟合）是通过修改损失函数来防止过拟合的，而 Dropout 则是通过在神经网络中断开部分神经元连接来实现的。由于小数据集而导致的过拟合现象，通常可通过数据增强的方法来防止过拟合现象的产生，“Early stopping”则是一种通过提前终止训练来防止过拟合的方法，即模型在训练数据集上将要出现过拟合问题之前停止迭代来防止过拟合，与之相反，由于机器学习算法对训练集数据拟合度不足而导致的欠拟合同样会

对训练模型的预测性能产生影响，但欠拟合问题在给定的模型评估指标下易被发现，因此通常不被讨论。

过拟合问题在使用机器学习模型对材料性能进行预测时较易产生，因为在大部分的材料性能预测时都面临严重的数据不足的问题。除了上述提到的处理过拟合问题的方法外，还可以通过融合多数据源数据（实验+仿真数据）、多尺度数据以及补充材料学分析等方法对材料数据进行补充以避免过拟合问题[101]。

4.3 不平衡数据集问题

在分类问题中，训练数据不平衡是指不同类别的样本数据数量相差很大，从而导致机器学习模型对不同类别样本的权重差异较大，难以对小类别样本作出正确的判断，从而影响模型预测效果。因此，数据不平衡是机器学习分类任务中的特别需要注意的典型问题，在训练机器学习模型之前，需要对类别样本不平衡的数据集进行相应的处理。目前为止，解决不平衡数据集问题的策略大致可分为两类[102]：一类是直接从训练集入手，通过改变训练集样本的分布，从而降低数据集的不平衡程度，即通过增加少样数据（即过采样）或减少大样数据（即欠采样）的方法对数据集进行补充或缩减以使得数据集保持均衡，并使用混淆矩阵(Confusion Matrix)、精确度(Precision)、召回率(Recall)、ROC 曲线[103]等评价指标对预测结果进行多角度的评价。另一类则是从算法入手，通过机器学习的算法来解决不平衡问题时的影响，正如 Lee 等[77]对 SVM 进行改进以解决汽车材料数据不平衡对分类结果的影响。此外，在材料性能预测问题中，如果样本集中大部分样本属于同一个材料体系而只有极少量的样本属于其他材料体系，那么也会造成数据不平衡问题，另外，机器学习在材料信息学中应用时的另一个常见问题是预测结果较差的材料数据通常都会被剔除，即材料数据选择时的偏向性也会导致一定的数据不平衡问题。

4.4 材料表征方法

材料性能预测模型的作用是使用给定材料的分子式或结构信息对其性能进行预测，机器学习模型的主要目的就是对表征后的材料数据和其性能之间的潜在关系进行学习，材料被表征后得到的数据又被称为材料描述符或材料特征，因此，材料描述符的设计需要能够尽可能多的保留下材料本身的信息。无论其是由计算获得还是实验测量产生，一个好的材料描述符应该比材料本身要更简单、更直接，如可以使用材料的原子半径、原子序数、元素组成或电负性作为材料描述符对化合物的结构进行预测，从而得到材料的唯一表示。虽然许多材料的元素组成和结构已知，但能够准确表征材料空间结构信息与其他方面信息关系的三维材料描述符很难设计，三维描述符的设计为解决电子分布和局部电位场量等三维固体材料系统的表示供了可能。Kajita 等[104]利用卷积神经网络在固态材料上开发出了一种简单通用的三维描述符，在 680 种氧化物上使用电子分布编码下的材料描述符数据进行材料性能预测并得到了较好的结果。

机器学习模型所使用的材料描述符需要围绕所要预测的目标材料性能进行设计，设计出好的材料描述符需要利用先验或专业知识，需考虑到与目标材料性能具有一定相关性的其它材料信息，然而，通常情况下材料描述符的设计需要通过尝试和试错来逐步优化的，该问题在机器学习领域称之为特征工程[105, 106]，

另外，特征选择也是机器学习中的重要方法之一，特征选择可将重要的材料信息保留，移除影响较弱的信息，从而能够提高预测精度并提高计算效率，常见的特征选择方法见文献[107, 108]。其中，Filter 方法借鉴了统计学的思想，对每个特征在预测中的影响进行评价，最优特征子集由其若干个优秀特征组成；相反，在 Wrapper 方法将机器学习算法嵌入到特征选择的过程中，通过测试的特征子集在此算法上的预测结果来决定其优劣。例如，Sendek 等[109] 通过使用多描述符来识别数据库中潜在的超离子结构，利用 Logistic 回归开发了离子电导率分类模型，从 12 831 种含锂离子固体材料中筛选得到 21 种可以用于锂电池的固体电解质的候选材料，从而加速对新型固体电解质的研究。统计学方法在理解和筛选材料方面的应用虽然很小但却越来越受欢迎，如对三元氧化物结构的预测假想[110]、预测热力学稳定性[111]以及筛选液体电解质[112, 113]等。

5 国内材料信息学发展现状

目前，国内使用机器学习方法对材料性能进行预测和发现新材料的研究还处于萌芽阶段，相关论文甚少，2018 年苏燕利用 RBF（径向基函数）神经网络建立了铸轧 7050 铝合金的力学性能预测模型[114]，2019 北京工业大学的年孙光民[115]推出了一种基于改进随机森林算法的铁磁材料硬度预测方法，然而，这些研究并没有针对新材料的发现展开。2019 年上海大学材料基因组工程研究院的徐永林等人[116]提出了一种基于高通量计算及机器学习的新材料带隙预测模型，可用于对指定类别的材料体系带隙进行快速而准确的预测，他们分析了 Lasso、SVR 和梯度提升树(Gradient Boosting Decision Tree, 以下简记为 GBDT)三种不同类型的机器学习模型的预测结果，并提出了一种稳定有效的新型集成学习方法。此外，上海大学材料学院的鲁文冲组[117]把支持向量机(SVM)、相关向量机(RVM)和随机森林(RF)等机器学习模型应用于钙钛矿材料的快速筛选中，中科大的江雪等人[118]将 SVM、神经网络等用于镍基单晶体超合金网格不适配性的预测，西北工业大学团队把神经网络机器学习方法用于合金设计。在我们的前期工作中，主要进行了将机器学习应用于高通量实验的相图生成与识别算法 [119-121]、[122]、基于遗传算法与第一性原理的掺杂材料的结构预测 [123, 124] ，基于卷积神经网络的材料性能预测研究[125]，以及超导材料临界温度预测等研究。

6 结论

本文针对机器学习在无机化合物性能预测方面的应用进行了综述，展示了机器学习在材料学领域广泛的适用性，通过使用已有的材料数据并训练有效的机器学习预测模型，可以一定程度避免成本昂贵的实验或者计算量巨大的计算仿真。由于材料固有的复杂性、材料多晶体的无定形性质以及多尺度几何结构的缺陷，使用传统计算模拟时往往达不到预期的效果。相比于传统计算或测量方法的繁杂与误差，利用现存可靠的材料数据集，机器学习算法在材料性能预测方面有着巨大的潜力。

自 2006 年以来，机器学习在新材料发现领域取得了突破性的进展，这是传统新材料研发模式的全新变革，对材料科学研究与新材料研发具有重要创新意义。机器学习采用先进的数据挖掘方法对材料数据库中的大量材料数据进行分析，缩短了材料研发周期、降低了试验过程中的错误率并减少了研发成本，极大地加快了新材料的研发进度。相比于国外，我国在材料基因组计划、基于数据驱动与机器学习的新材料发现方面还处于萌芽阶段，在开放式材料数据库的构建和材料信息学的应用方面还需要有突破性的进步。

参考文献

1 Ceder G, Morgan D, Fischer C, et al. *Mrs Bulletin*, 2006, 31, 981.

2 Saad Y, Gao D, Ngo T, et al. *Physical Review B*, 2012, 85,1092

3 Gaultois M W, Oliynyk A O, Mar A, et al. *Apl Materials*, 2016, 4(5), 053213.

4 Wang Z, Wang M, Yong Q L, et al. *Materials China*, 2017, 36(2), 132 (in Chinese).
王卓，王礞，雍歧龙，等. *中国材料进展*, 2017, 36(2), 132.

5 Phil D L , Jennifer W , Yoshua B , et al. *Nature*, 2017, 552(7683), 23.

6 Ferguson A L. *Journal of Physics-Condensed Matter*, 2018, 30(4), 043002.

7 Mannodi-Kanakkithodi A , Tran H , Ramprasad R. *Chemistry of Materials*, 2017, 29, 9001.

8 Pilania G , Mannodi-Kanakkithodi A , Uberuaga B P , et al. *Scientific Reports*, 2016, 6, 19375.

9 Kim C, Huan T D, Krishnan S, et al. *Scientific Data*, 2017, 4, 170057.

10 Legrain F, Carrete J, Roekeghem A V, et al. *Chemistry of Materials*, 2017, 29, 6220.

11 Kim C, Pilania G, Ramprasad R. *Journal of Physical Chemistry C*, 2016, 120, 14575.

12 Raccuglia P, Elbert K C, Adler P D, et al. *Nature*, 2016, 533, 73.

13 Xue D Z, Balachandran P V, Yuan R H, et al. *Proceedings of the National Academy of Sciences of the United States of America*, 2016, 113, 13301.

14 Wilson D R, Mishra B, Rui Y, et al. *Molecular Therapy*, 2017, 25, 21.

15 Pankajakshan P, Sanyal S, Noord O E D, et al. *Chemistry of Materials*, 2017, 29(10), 04229.

16 Seko A, Takahashi A, Tanaka I. *Physical Review B*, 2015, 92(5), 054113.

17 Deml A M, Hayre R O, Wolverton C, et al. *Physical Review B*, 2016, 93, 085142.

18 Lee J, Seko A, Shitara K, et al. *Physics*, 2015, 89, 611.

19 Agrawal A, Choudhary A. *Apl Materials*, 2016, 4, 1.

20 Greeley J, Jaramillo T F, BondeJ, et al. *Nature Materials*, 2006, 5, 909.

21 Sun S, Zhang T Y. In: The Chinese Congress of Theoretical and Applied Mechanics(CCTAM 2015). Shanghai, 2015, pp. 332 (in Chinese).
孙升，张统一. *中国力学大会-2015*, 2015, pp. 332.

22 Wang H, Xiang Y, Xiang X D, et al. *Science & Technology Review*, 2015, 33(10), 13 (in Chinese).
汪洪，向勇，项晓东，等. *科技导报*, 2015, 33(10), 13.

23 Wang H Z, Wang H, Ding H, et al. *Science & Technology Review*, 2015, 33(10), 31 (in Chinese).
王海舟，汪洪，丁洪，等. *科技导报*, 2015, 33(10), 31.

24 Xiang X D, Wang H, Xiang X Y, et al. *Science & Technology Review*, 2015, 33(10), 64 (in Chinese).
项晓东，汪洪，向勇，等."组合材料芯片技术在新材料研发中的应用," *科技导报*, 2015, 33(10), 64.

25 Pan D, Qi X G, Liu L L, et al. *Journal of the Chinese Ceramic Society*, 2018, 46(4), 479 (in Chinese).
潘都，戚兴国，刘丽露，等. *硅酸盐学报*, 2018, 46(4), 479.

26 Xiao R J, Li Hong, Chen L Q. *Acta Physica Sinica*, 2018, 67(12), 291 (in Chinese).
肖睿娟，李泓，陈立泉. *物理学报*, 2018, 67(12), 291.

27 Chen L Q. *Battery Bimonthly*, 2002, 32(s1), 32 (in Chinese).
陈立泉, "锂离子电池正极材料的研究进展," *电池*, 2002, 32(s1), 32.

28 Lookman T, Balachandran P V, Xue D Z, et al. *Current Opinion in Solid State & Materials Science*, 2017, 21, 121.

29 Seko A, Togo A, Tanaka I. *Nanoinformatics*, Springer, Singapore, 2018.

30 Elton D C, Boukouvalas Z, Butrico M S, et al. *Scientific Reports*, 2018, 8, 9059.

31 Wang M Z, Wu J X, Lin L, et al. *Acs Nano*, 2016, 10, 10317.

32 Mannodi-Kanakkithodi A, Pilania G, Ramprasad R. *Computational Materials Science*, 2016, 125, 123.

33 Aryal S, SakidjaR, Barsoum M W, et al, *Physica Status Solidi B-Basic Solid State Physics*, 2014, 251, 1480.

34 Faber F A, Lindmaa A, Von Lilienfeld O A, et al. *Physical Review Letters*, 2016, 117(13), 135502.

35 Meredig B, Agrawal A, Kirklin S, *et al. Physical Review B*, 2014, 89(9), 094104.

36 Seko A, Maekawa T, Tsuda K, et al. *Physical Review B*, 2014, 89, 054303.

37 Seko A, Togo A, Hayashi H, et al. *Physical Review Letters*, 2015, 115(20), 205901.

38 Li Z, Ma X F, Xin H L. *Catalysis Today*, 2017, 280, 232.

39 Hong W T, Welsch R E, Shao-Horn Y. *Journal of Physical Chemistry C*, 2016, 120, 78.

40 Pilania G, Whittle K R, Jiang C, et al. *Chemistry of Materials*, 2017, 29, 2574.

41 Konno T, Kurokawa H, Nabeshima F, et al. *arXiv preprint arXiv*, 2018, 1812,01995.

42 Stanev V, Oses C, Kusne A G, et al. *npj Computational Materials*, 2018, 4, 29.

43 Mannodi-Kanakkithodi A, Pilania G, Huan T D, et al. *Scientific Reports*, 2016, 6, 20952.

44 Mueller T, Kusne A G, Ramprasad R. *Machine Learning in Materials Science: Recent Progress and Emerging Applications*, Reviews in Computational Chemistry, USA, 2016.

45 Chmiela S, Tkatchenko A, Sauceda H E, et al. *Science Advances*, 2016, 3(5), 1603015.

46 Rupp M, Tkatchenko A, Muller K R, et al. *Physical Review Letters*, 2012, 108(5), 58301.

47 Kirklin S, Meredig B, Wolverton C. *Advanced Energy Materials*, 2013, 3, 252.

48 Behler J, Martonak R, Donadio D, et al. *Physical Review Letters*, 2008, 100(18), 185501.

49 Behler J. *Journal of Physics Condensed Matter An Institute of Physics Journal*, 2014, 26, 183001.

50 Bartok A P, Payne M C, Kondor R, et al.*Physical Review Letters*, 2010, 104(13), 136403.

51 Bartok A P, Csanyi G. *International Journal of Quantum Chemistry*, 2016, 115(16), 1051.

52 Zhao J C. *Chinese Journal of Nature*, 2014, 36, 89.

53 Xiao J M, Wang H, Ouyang C Y. *Journal of Jiangxi Normal University(Natural Science Edition)*, 2015, 39(1), 73 (in Chinese).

肖建茂, 汪浩, 欧阳楚英. *江西师范大学学报(自然科学版)*, 2015, 39(1), 73.

54 Jain A, Ong S P, Hautier G, et al. *Apl Materials*, 2013, 1, 1049.

55 Curtarolo S, Setyawan W, Hart G L W, et al. *Computational Materials Science*, 2012, 58, 218.

56 Curtarolo S, Setyawan W, Wang S D, et al. *Computational Materials Science*, 2012, 58, 227.

57 Taylor R H, Rose F, Toher C, et al. *Computational Materials Science*, 2014, 93, 178.

58 Kalidindi S R, Graef M D. *Annual Review of Materials Research*, 2015, 45, 171.

59 Dompablo M E A Y D, Ven A V D, Ceder G. *Phys.rev.b*, 2002, 66, 340.

60 Jain A, Hautier G, Moore C J, et al. *Computational Materials Science*, 2011, 50, 2295.

61 Wu Y, Lazic P, Hautier G, et al. *Energy & Environmental Science*, 2012, 6, 157.

62 Yang L, Ceder G. *Phys.rev.b*, 2013, 88(22), 330.

63 Pilania G, Kim C, Ramprasad R. *Chemistry of Materials*, 2016, 28, 1304.

64 Ward L, Agrawal A, Choudhary A, et al. *Npj Computational Materials*, 2016, 2, 201628.

65 Fan R E, Chang K W, Hsieh C J, et al. *Journal of Machine Learning Research*, 2008, 9, 1871.

66 Seko A, Takahashi A, Tanaka I. *Physical Review B*, 2014, 90, 024101.

67 Gross S M, Tibshirani R. *Computational Statistics & Data Analysis*, 2016, 101, 226.

68 Schutt K T, Glawe H, Brockherde F, et al. *Physical Review B*, 2014, 89(20), 118.

69 Ravi D, Wong C, Deligianni F, et al. *Ieee Journal of Biomedical and Health Informatics*, 2017, 21, 4.

70 Park W B, Chung J, Jung J, et al. *Iucrj*, 2017, 4, 486.

71 Xu Y J, Pei J F, Lai L H. *Journal of Chemical Information and Modeling*, 2017, 57, 2672.

72 Fan M, Hu J, Cao R, et al. *Sci Rep*, 2017, 7(1), 18040.

73 Azimi S M, Britz D, Engstler M, et al. *Scientific Reports*, vol. 8, Feb 1 2018, 8(1), 2128.

74 Cecen A, Dai H, Yabansu Y C,et al. *Acta Materialia*, 2017, 146, 76.

75 Ma X, Li Z, Achenie L E, et al. *The journal of physical chemistry letters*, 2015, 6, 3528.

76 Chang C C, Lin C J. *ACM Transactions on Intelligent Systems and Technology*, 2011, 2, 1.

77 Lee K K, Harris C J, Gunn S R, et al. *Control Sensitivity SVM for Imbalanced Data A Case Study on Automotive Material*, Springer, Vienna, 2001.

78 FaridKianifard. *Technometrics*, 2004, 32, 352.

79 Montgomery D C, Peck E A. *Introduction to linear regression analysis*,Wiley, USA, 1982.

80 Friedman J, Hastie T, Tibshirani R. *Annals of Statistics*, 2000, 28, 337.

81 Weinberger K Q, Saul L K. *Journal of Machine Learning Research*, 2009, 10, 207.

82 Peterson L E. *Scholarpedia*, 2009, 4, 1883.

83 Quinlan J R. *Machine Learning*, 1986, 1, 81.

84 Breiman L. *Machine Learning*, 2001, 45, 5.

85 Cutler A, Cutler D R, Stevens J R. *Machine Learning*, 2004, 45, 157.

86 Yosinski J, Clune J, Bengio Y, et al.In: *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA ,2014.

87 Chow T W S, Yong F. *IEEE Transactions on Industrial Electronics*, 1998, 45, 151.

88 Xie T, Gong D L, Zhang W L, et al. *Superconductor Science & Technology*, 2017, 30(9), 095002.

89 Feng J, Zhou Z H. *ArXiv*, 2018, 1709, 09018.

90 Ghiringhelli L M, Vybiral J, Ahmetcik E, et al. *New Journal of Physics*, 2017, 19, 023017.

91 Baraniuk R G. *IEEE Signal Processing Mag*, 2007, 24(4), 118.

92 Moreno-Torres J G, Saez J A, Herrera F. *IEEE Trans Neural Netw Learn Syst*, 2012, 23, 1304.

93 Meredig B, Antono E, Church C, et al. *Molecular Systems Design & Engineering*, 2018, 3, 819.

94 Jha D, Ward L, Paul A, et al. *Scientific reports*, 2018, 8, 17493.

95 Evans J D, Coudert F. *Chemistry of Materials*, 2017, 29(18), 7833.

96 Zhang Y, Ling C. *npj Comput Mater*, 2018, 4, 25.

97 Hawkins D M. *Cheminform*, 2004, 35, 1.

98 Wolpert D H, Macready W G. *IEEE Trans on Evolutionary Computation*, 1997, 1, 67.

99 Chicco D. *Biodata Mining*, 2017, 10(1), 35.

100 Zhu H, Tsang E C C, Zhu J. *Soft Computing*, 2018, 22, 3477.

101 Karpatne A, Watkins W, Read J, et al. *ArXiv preprint arXiv*, 2017, 1710, 11431.

102 Lemaitre G, Nogueira F, Aridas C K. *Journal of Machine Learning Research*, 2017, 18, 1.

103 Hand D J. *Machine Learning*, 2009, 77, 103.

104 Kajita S, Ohba N, Jinnouchi R, et al. *Scientific Reports*, 2017, 7, 16991.

105 Guyon I, Elisseeff A. *Journal of Machine Learning Research*, 2003, 3, 1157.

106 Li G Z, Yang J, Liu G P, et al. In: 8th Pacific Rim International Conference on Artificial Intelligence (PRICAI-04), Auckland, New Zealand, 2004, pp. 292.

107 Cai J, Luo J, Wang S,et al. *Neurocomputing*, 2018, 300, 70.

108 Li J, Cheng K, Wang S, et al. *ACM Computing Surveys (CSUR)*, 2018, 50, 94.

109 Sendek A D, Yang Q, Cubuk E D, et al. *Energy & Environmental Science*, 2017, 10, 306.

110 Hautier G, Fischer C C, Jain A, et al.*Chemistry of Materials*, 2010, 22, 3762.

111 Meredig B, Agrawal B, Kirklin S, et al. *Phys.rev.b*, 2014, 89, 82.

112 Qu X, Jain A, Rajput N N, et al. *Computational Materials Science*, 2015, 103, 56.

113 Schütter C, Husch T, Viswanathan V, et al. *Journal of Power Source*, 2016, 326, 541.

114 Su Y, Liang W. *Hot Working Technology*, 2018, 47(21), 145 (in Chinese).
苏燕, 梁武. *热加工工艺*, 2018, 47(21), 145..

115 Sun G M, Liu H, He C F, *et al. Journal of Beijing University of Technology*, 2019, 45(2), 119 (in Chinese).
孙光民, 刘浩, 何存富, 等. *北京工业大学学报*, 2019, 45(2), 119.

116 Xu Y L, Wang X M, Li X, *et al. SCIENTIA SINICA Technologica*, 2019, 49(1), 44 (in Chinese).
徐永林, 王香蒙, 李鑫, 等. *中国科学:技术科学*, 2019, 49(1), 44.

117 Zhai X, Chen M, Lu W. *Computational Materials Science*, 2018, 151, 41.

118 Jiang X, Yin H Q, Zhang C, et al. *Computational Materials Science*, 2018, 143, 295.

119 Bunn J K, Han S, Zhang Y, et al. *Journal of Materials Research*, 2015, 30, 879.

120 Bunn J K, Hu J, Hattrick-Simpers J R. *JOM*, 2016, 68(8), 2116.

121 Xiong Z, He Y, Hattrick-Simpers J R, et al. *ACS Combinatorial Science*, 2017, 19(3), 137.

122 Li S, Xiong Z, Hu J. *Materials Science and Technology*, 2018, 34, 315.

123 Atilgan E, Hu J. In: Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation. Madrid, 2015, pp. 1349.

124 Atilgan E, Hu J. *Bulletin of Materials Science*, 2018, 41, 1.

125 Cao Z, Dan Y, Xiong Z, et al. *Crystals*, 2019, 9, 191.

Chengcheng Niu received the B.S. degree in mechanical engineering from ShanDong JiaoTong University in 2017. She is currently pursuing M.S. degree in mechanical engineering at the Key Laboratory of Advanced Manufacturing Technology (Ministry of Education), Guizhou University under the supervision of Prof. Jianjun Hu. Her research has focused on materials informatics.

牛程程，2017 年 6 月毕业于山东交通学院，获得工学学士学位。现为贵州大学现代制造技术教育部重点实验室硕士研究生，在胡建军教授的指导下进行研究。目前主要研究领域为材料信息学。

Jianjun Hu received the B.S. and M.S. degrees of Mechanical Engineering in 1995 and 1998 respectively from Wuhan University of Technology, China. He received the Ph.D. of Computer Science in 2004 from Michigan State University in the area of machine learning and evolutionary computation. He worked as postdoctoral fellow at Purdue University and University of Southern California from 2004 to 2007. He is currently associate professor at the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, United States and also visiting professor at School of Mechanical Engineering, Guizhou University. His research interests include machine learning, deep learning, data mining, evolutionary computation, bioinformatics, and material informatics. His research has been supported by NSF Career Award and NSF China. Dr. Hu is also the associate editors of Nature Scientific Report, PLOS ONE, and BMC Bioinformatics.

胡建军，贵州大学机械学院特聘教授、博士导师；美国南卡罗来纳大学计算机科学与工程系终身副教授。分别于 1995 年和 1998 年获得武汉理工大学机械工程学士与硕士学位。2004 年获得密歇根州立大学计算机科学专业博士学位。2004 年至 2007 年在普渡大学和南加州大学担任博士后研究员。主要研究方向：机器学习、材料信息学、生物信息学、智能制造。获得美国国家自然科学基金委 Career Award 以及国家自然科学基金项目《基于机器学习与图像处理算法的高通量组合材料实验相图生成与物相辨识方法研究》资助。已发表 SCI 论文 60 余篇。学术服务兼任《Nature Scientific Report》，《PLOS ONE》和《BMC Bioinformatics》的副编辑。