中国科学: 物理学 力学 天文学 2020 年 第 50 卷 第 6 期: 067002
SCIENTIA SINICA Physica, Mechanica & Astronomica physcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
评 述 非晶合金材料科学前沿论坛专题

机器学习在非晶材料中的应用

吴佳琦[1], 孙奕韬[2], 汪卫华[2], 李茂枝[1*]

1. 中国人民大学物理系, 光电功能材料与微加工北京市重点实验室, 北京 100872;
2. 中国科学院物理研究所, 北京 100190
*联系人, E-mail: maozhili@ruc.edu.cn

收稿日期: 2019-10-08; 接受日期: 2019-11-26; 网络出版日期: 2020-02-25
国家自然科学基金(编号: 51631003, 51801230)和国家重点基础研究发展计划(编号: 2015CB856800)资助项目

摘要 作为新兴非晶材料的金属玻璃由于其优异的力学、物理以及化学性能而被广泛研究. 玻璃形成能力一直是制约着非晶材料发展的重要问题, 为了设计出具有良好玻璃形成能力的非晶材料, 对非晶材料的玻璃形成能力已经有大量的研究. 研究表明单一的影响因素不足以全面解释非晶材料的玻璃形成能力, 即玻璃形成能力是由多种因素共同影响的. 另一方面, 由于非晶材料具有复杂且无序的结构, 传统的方法难以全面、清晰地理解非晶材料的结构与本质. 机器学习这一新的研究范式为解决非晶材料领域的关键瓶颈问题提供了新的途径和契机. 本文首先简单介绍了一些机器学习算法, 如支持向量机、人工神经网络和K均值聚类. 随后介绍了机器学习在非晶材料中的应用, 包括非晶结构分类、非晶结构-性能关联和非晶宏观性质的预测, 并提出了基于机器学习方法在未来非晶研究中的应用前景, 包括非晶数据库的建立、高通量计算方法的发展和机器学习势函数的发展.

关键词 机器学习, 非晶材料, 玻璃形成能力, 结构-性能关联

PACS: 02.70.-c, 61.43.Fs, 64.70.Pf , 61.25.Mv

1 引言

机器学习是一门计算机科学的子学科, 是计算机科学和统计学的交叉, 其核心是人工智能和数据科学. 机器学习的目的是赋予计算机自动学习的能力[1,2]. 机器学习拥有长远的历史, 最早可以追溯到17世纪的最小二乘法和马尔可夫链. 20世纪50年代, 通过赋予机器逻辑推理能力使机器获得智能, 称为“推理期”. 70年代, 通过将人类的知识赋予机器使其获得智能, 称为“知识期”. 随着计算机技术和互联网的飞速发展, 机器学习也获得了前所未有的发展. 近年来, 机器学习受到了人们的广泛关注, 并且在许多领域中展现出超人的能力, 使人们看到了它的巨大潜力, 如机器学习在围棋领域取得的举世瞩目的成功[3]、汽车的自动驾驶功能[4]、图像分类[5]等. 如今, 机器学习已经广泛应用于图像和语音识别[6]、智能网络搜索[7]、欺诈检测[8]、垃圾邮件过滤率[9]、信用评级[10]等领域, 给人们的生产、生活提供了极大的便利. 由于具有强大的科研潜力, 机器学习早已被广泛应用在科学研究中. 例如, 基于图像识别对肿瘤类型与生长速率进行分

引用格式: 吴佳琦, 孙奕韬, 汪卫华, 等. 机器学习在非晶材料中的应用. 中国科学: 物理学 力学 天文学, 2020, 50: 067002
Wu J Q, Sun Y T, Wang W H, et al. Application of machine learning approach in disordered materials (in Chinese). Sci Sin-Phys Mech Astron, 2020, 50: 067002, doi: 10.1360/SSPMA-2019-0345

类[11], 基于血清细胞预测帕金森病的发展[12], 基于大数据建立糖尿病预测模型[13]等医学领域的研究, 为疑难杂症的治疗提供了很大的帮助. 在地震学研究方面, 利用机器学习, 研究人员研究了南极洲微震和海冰之间的关系[14], 建立了地震引起的土壤液化预测模型[15]等, 使地震的预测更为精确. 此外, 机器学习还被用来研究量子体系. 例如, 利用人工神经网络识别量子相变[16]; 利用深度神经网络预测和开发量子拓扑材料[17], 将机器学习与量子计算相结合进行电子结构计算[18], 并且发展了量子机器学习算法以解决量子计算中编码问题[19]等方面的研究. 由此可见, 机器学习方法已经成为继理论研究、实验、计算模拟之后全新的思路和研究范式.

对于材料科学领域而言, 新材料的设计研发过程是非常艰难的, 更多地依赖于研究人员的个人经验, 经历无数次实验, 才能在机缘巧合之下取得成果, 还需要反向检测这种新材料的各种性质. 随着材料科学、物理学、化学等学科, 以及计算机科学的发展, 利用高性能计算机进行材料设计、开发变得可行. 在材料计算领域, 对于微观原子或电子尺度, 最常用的是第一性原理计算(基于密度泛函理论)[20,21]、分子动力学模拟[22,23]以及蒙特卡罗方法. 计算材料学的发展极大地推动了新材料的设计与研发, 缩短了材料研发的周期, 大大降低了新材料的研发成本[24,25]. 这些计算方法也存在一些缺陷, 例如, 无法模拟大量原子的体系及其长时间演化行为, 计算可靠性也不够理想. 虽然, 随着人们对原子间相互作用势认识的加深, 以及计算机计算能力的提升, 这两个典型的问题都在不断改善, 但是问题依然存在. 目前的理论计算和模拟结果与实际应用之间仍然存在一定的差距. 因此, 发展指导新材料设计、开发的新方法是非常必要的.

材料科学领域至今已经发展了很多年, 积累了大量的实验数据, 为以大数据分析为主的机器学习方法在材料科学领域的应用提供了契机, 奠定了基础. 近年来, 机器学习已经被广泛应用于新材料的设计和研发, 并发表了一系列高水平论文, 例如电催化剂的筛选[26]、辅助设计有机发光二极管[27]、晶界结构的探索[28]、无机材料性能的预测[29]等. 2016年5月*Nature*刊登的封面文章显示, 机器学习能够充分挖掘隐藏在大量废弃的(“失败的”)实验数据背后有价值的信息, 帮助研究人员更加高效地预测新材料的构成[30]. 文章中提出“从失败中学习”, 即利用“失败”的实验数据预测并指导新材料的设计开发. 在材料科学领域, 能找到的文献里所报道的数据都是成功实验的数据, 但是绝大部分的实验数据都是“失败”的实验数据, 而这些“失败”实验的数据则“沉睡”在每个课题组的记录本上. Alexander J. Norquist团队[30]利用实验室未成功的水热反应的数据训练机器学习模型, 并用得到的模型来预测新的反应, 所得的模型能够成功预测新的有机-无机材料的合成条件, 合成成功率达89%. 当输入数据被模型处理后, 计算机将会给出实验改进建议. 将改进后的实验所得到的数据再反馈给计算机, 进一步改进实验. 这项研究表明, “失败”的实验数据所包含的信息对预测反应成功和失败的边界条件有着重要的价值. 事实上, 传统的材料设计研发, 即根据材料科学家经验的“试错”法, 也体现着“失败”的实验数据的重要性, 但是, 更多的是经验性的, 且传统的研究范式也很难系统地利用这些数据进行材料的设计和研究. 这个成果备受材料科学家关注的原因在于它改变了材料科学家对“失败”的实验数据的认识, 而对于机器学习而言, 实验中所得到的数据都能够加以利用, 建立机器学习模型, 这为材料科学的研究和发展开辟了新的天地.

2 机器学习的分类及部分算法简介

机器学习通过建立模型的方法从而实现对已有数据的学习并且做出预测. 机器学习算法一般可分为三大类: 监督式学习、非监督式学习以及强化学习.

监督式学习需要给每组训练数据一个明确的标签, 包括最小二乘法、支持向量机、人工神经网络、决策树等算法. 监督式学习常用来处理分类问题和回归问题. 其中最常用的监督式学习算法是支持向量机(Support Vector Machine, SVM)算法以及人工神经网络算法. 对于非监督式学习, 训练数据不需要给定标签, 训练模型是为了找出数据样本之间存在的内在关系. 非监督式学习包括K均值聚类算法、均值漂移算法、马尔可夫随机场、层次聚类算法等. 非监督式学习常用于关联规则的学习和聚类. 与监督式学习不同, 强化学习的输入数据直接反馈到模型, 模型将会立刻作出调整. 强化学习常用于动态系统和机器人的控制. 本文将着重介绍支持向量机算法, 并简要介绍人工神经网络(Artificial Neutral Network, ANN)算法和K均值

聚类算法.

2.1 支持向量机

支持向量机算法是一种对数据进行二分类的算法[31]. 在训练时需要输入的数据是已知其类别的数据样本, 并根据这些数据建立模型, 最后根据所建立起来的模型用以区分未知分类的数据样本, 从而达到预测类别的目的. 支持向量机算法的本质是运用一个分类超平面将高维空间中的数据样本进行分割, 并且使得正负样本之间的间隔最大化.

根据训练数据的不同, 支持向量机可分为三类: (1) 如果训练数据线性可分, 那么就可以通过硬间隔最大化, 得到线性可分支持向量机; (2) 如果训练数据近似线性可分, 那么就可以通过软间隔最大化, 得到线性支持向量机; (3) 如果训练数据线性不可分, 那么可以将训练数据映射到高维空间, 使映射后的训练数据在高维空间中线性可分, 即将低维空间中的非线性问题转化为高维空间中的线性问题, 得到非线性支持向量机. 由于训练过程中不关心单个数据样本的情况, 只关心高维空间中数据样本两两之间的距离, 所以没必要将低维空间中的数据样本一个个地映射到高维空间中, 只需要通过一个核函数将低维空间中的距离映射到高维空间中. 由于核函数的选择没有通用的标准, 所以想要找到一个合适的核函数比较困难.

图1为线性支持向量机示意图, 其中符号“●”和“▲”代表两类不同的数据样本, SVM超平面通过软间隔最大化将两类数据样本近似划分. 也就是说, 软间隔支持向量机允许有一定的数据样本被误分类. 除去误分类的数据外, 距离SVM超平面最近的称为支持向量. 两支持向量之间垂直于超平面的距离称为边界宽度. 需要注意的是, 在支持向量机寻找分类超平面的训练过程中, 起作用的只有支持向量.

当支持向量机被用来处理回归问题时, 称为支持向量回归(Support Vector Regression, SVR). 与支持向量机处理分类问题不同, 支持向量回归是为了寻找一个线性回归方程来拟合所有的数据样本, 它所构造的超平面不是使间隔最大化, 而是使数据样本离超平面的总方差最小.

2.2 人工神经网络

人工神经网络(ANN)是20世纪80年代以来人工智能领域兴起的研究热点[32,33]. 它从信息处理角度对人脑的神经元网络进行抽象, 建立某种简单模型, 按不同的连接方式组成不同的网络. 如图2所示, 人工神经网络是一种运算模型, 它是由大量的节点(或称神经元)之间相互连接构成, 每个节点代表一种特定的输出函数, 称为激励函数(Activation Function). 每两个连接的节点之间存在着一个权值, 表示该连接的权重. 节点一般分为三类: 位于输入层(Input Layer)的输入节点、位于输出层(Output Layer)的输出节点以及位于隐藏层(Hidden Layer)的隐藏节点. 图2(a)展示的是只有一个隐藏层的简单人工神经网络; 图2(b)展示的是存在多个隐藏层的深度人工神经网络. 人工神经网络的训练过程中, 根据最终输出结果与预期结果的差别, 选择以下三种方式减小误差: (1) 调整每个突触中的权值; (2) 神经网络结构的修改, 包括隐藏节点层数的增减、层内节点数目的增减以及突触的连接与删除; (3) 非线性映射的函数的选择与更改.

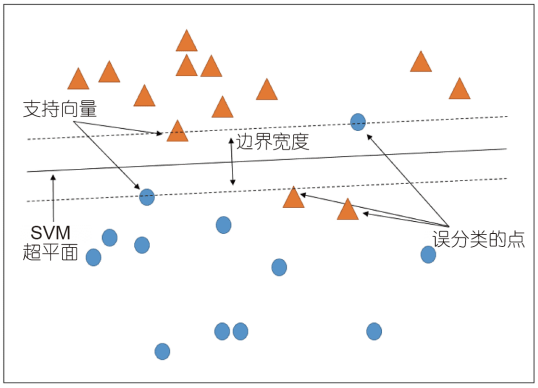

图 1 (网络版彩图)线性支持向量机示意图. 利用一个超平面划分两类样本, 并使间隔最大化. 其中符号“●”和“▲”代表两类不同的数据样本, 黑色实线代表SVM分类超平面, 除去误分类的数据外, 距离SVM超平面最近的称为支持向量, 而两支持向量之间垂直于超平面的距离称为边界宽度

Figure 1 (Color online) The diagram of linear support vector machine (LSVM) uses a hyperplane to divide two kinds of samples and maximize the interval, where the symbols “●” and “▲” represent two different types of data samples, respectively. The black solid line represents the SVM classification hyperplane. Except the misclassified data, the nearest to the SVM hyperplane is called the support vector, and the distance between the two support vectors perpendicular to the hyperplane is called the margin width.

2.3 *K*均值聚类

*K*均值聚类算法是一种常见的聚类算法, 该算法对没有标签的数据样本进行训练, 然后将数据样本聚

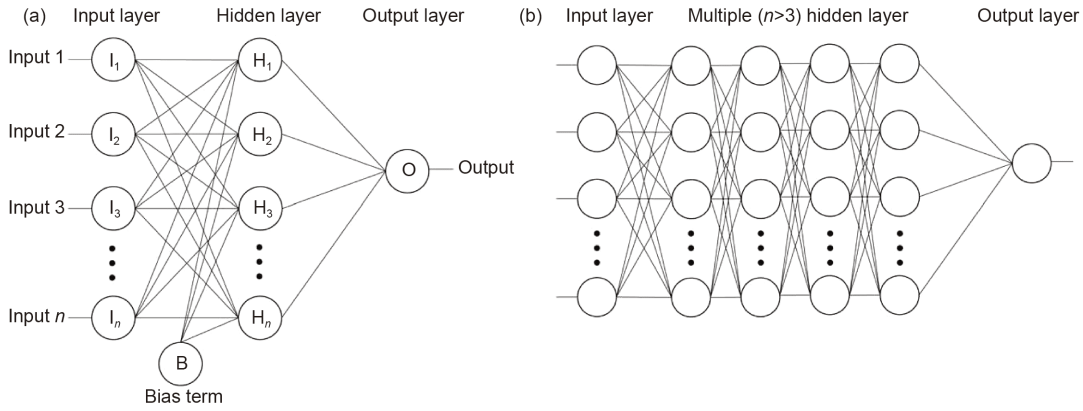

图 2 人工神经网络原理图, 包含输入层、输出层以及隐藏层. (a) 包含一个隐藏层的简单人工神经网络; (b) 包含多个隐藏层的深度人工神经网络[32]

Figure 2 Schematic diagrams of artificial neural network, including input layer, output layer and hidden layer. (a) A simple artificial neural network with a hidden layer; (b) depth artificial neural network containing multiple hidden layers (reprinted with permission from ref. [32]).

类成不同的类别[34]. K均值聚类是一种迭代算法, 该方法的过程如下:

(1) 选择K个随机点, 作为聚类中心;

(2) 针对数据样本中的每个数据, 计算其与K个聚类中心的距离, 将其与距离最近的聚类中心关联起来, 在此基础上, 将属于同一个聚类中心的样本聚为一类(也称为簇);

(3) 计算每一个簇所有样本的均值, 并将聚类中心移动到平均值的位置. 然后, 对步骤(2)和(3)进行迭代, 直到聚类中心不发生移动.

这里需要注意的是, K均值聚类的K个聚类中心是几何中心, 并不一定是某一个数据样本.

3 机器学习在非晶体系中的应用

由于非晶材料结构复杂无序, 用传统的方法很难建立起结构-性能关联, 从而更好地理解非晶材料的本质并指导新型高性能非晶材料的设计和研发. 过去几十年, 非晶合金材料前沿领域不断有新的重要进展和突破, 包括新材料和新应用, 在能源、信息、国防、航空航天等高新技术领域发挥着越来越重要的作用. 同时, 非晶合金材料的研发与应用也面临着如何提高合金玻璃形成能力、室温塑性等关键瓶颈问题. 提高合金的玻璃形成能力是非晶合金材料制备中需要解决的首要问题, 不仅涉及制备工艺, 更与合金液体的基本物理问题密切相关, 直接决定着非晶合金材料的力学性能和稳定性等. 该瓶颈问题的突破将极大推动非晶合金材料的基础研究和应用开发, 产生巨大的经济和社会效益. 此外, 对非晶合金材料而言, 相关研究已经经历了半个多世纪, 积累了大量成功和"失败"的实验数据. 机器学习这一新的研究范式为解决非晶合金材料领域的关键瓶颈问题提供了新的途径和契机, 将有力推动非晶合金材料的基础研究和应用开发. 在现有的研究当中, 机器学习在非晶体系中主要有三方面的应用: (1) 从结构间的相似性角度出发, 利用机器学习分析非晶材料的结构特征; (2) 利用机器学习分析模拟或实验得到的结构、动力学、力学性能等数据, 建立起结构与性能之间的关联; (3) 利用机器学习分析现有的实验数据得到模型, 并根据模型给出理论预测, 指导新材料的设计与开发. 本节将从以上三个方面简要介绍机器学习方法在非晶中的研究.

3.1 非晶材料结构分析

理解和预测合金的玻璃形成能力是一个复杂的、多方面的问题, 而这种复杂性的一部分原因在于对玻璃和液体结构的认识不充分, 特别是对由中心原子及其最近邻原子组成的短程序的认识和表征. 在中程序的范围之内, 由于短程序单元连接时存在的旋转自由度而导致长程有序的缺失. 因此, 想要定量地描述金属玻璃体系的短程序结构非常困难.

现有最常用的描述金属玻璃体系短程序的方法是Voronoi空间分割法[35]. Voronoi分析方法通过中心原子及其最近邻原子之间的中垂面将空间切割, 并将中垂面包络的包含中心原子的多面体称为Voronoi多面体.

通过多面体的不同形状面的数目定义所谓的Voronoi指数: $\langle n_3, n_4, n_5, n_6\rangle n_i$($i$=3, 4, 5, 6)代表多面体含有的$i$边形的面的个数. 这里, 相同Voronoi指数的多面体认为是相似的, 但是由于缺乏衡量相似度的量化标准而存在着一定的缺陷.

Maldonis等人[36]从短程序单元之间的相似性出发, 利用基于密度的空间聚类(HDBSCAN)的机器学习聚类算法将短程序单元分成30类. 对比两个短程序单元之间的几何相似度需要解决短程序单元的旋转问题. 为此, 他们发展了Point-Pattern Matching (PPM)方法[37]. PPM方法的工作原理是, 使用一种近似的刚性图匹配技术将两组三维点尽可能地排列成一个方向和位置的相似点, 以处理这两个结构之间的轻度无序. 在对齐后, 便可以使用任意几何度量来比较结构的相似性. 使用PPM方法可以量化$Cu_{50}Zr_{45}Al_5$体系中相同配位数的短程序单元之间的相似性, 并生成相似度矩阵. 根据相同配位数的短程序单元之间的相似性, 通过基于密度的空间聚类(HDBSCAN)的机器学习聚类算法, 而不是人为选取一定的截断, 将相同配位数的短程序单元分为若干类, 所有的短程序单元共分为30类. 可以认为, 金属玻璃体系是由这30类短程序单元相互连接而成的无序体系. 这种方法为在金属玻璃中建立起清晰的结构-性能关联奠定了基础.

3.2 非晶材料的结构-性能关联

固体在足够高的应力下将会发生流动, 在足够高的温度下将会熔化. 对于晶体而言, 流动和熔化优先发生于结构缺陷——位错处. 也就是说位错控制着晶体的流动以及熔化. 但是对于非晶等无序体系而言, 用传统的方法从结构上找到流动缺陷是非常困难的. 虽然, 自由体积、键取向序等物理结构量与流动缺陷有关联[38], 但是不能从结构方面先验地识别流动缺陷.

Liu课题组[39]使用支持向量机从结构上区分了容易发生重排的原子与不容易发生重排的原子, 并以此识别容易发生重排的原子, 即识别了非晶中的流动缺陷. 他们引入了表征原子近邻径向密度特征的结构量 $G_Y^X(i;\mu)=\sum_j \mathrm{e}^{-(R_{ij}-\mu)^2/L^2}$ 和表征键取向信息的结构量 $\Psi_{YZ}^X(i;\xi,\lambda,\zeta)=\sum_j\sum_k \mathrm{e}^{-\left(R_{ij}^2+R_{ik}^2+R_{jk}^2\right)/\xi^2}\left(1+\lambda\cos\theta_{ijk}\right)^\zeta$ 来表征非晶中原子的结构信息, 这些物理量已经被用来表示从量子力学计算得到的复杂材料的势能面[40]. 其中, R_{ij}表示原子之间的距离, θ_{ijk}表示i, j和k角, L, μ, ξ, λ和ζ都是常数, X, Y和Z是识别体系中不同原子的标签, 分别对应于$i\leftrightarrow X$, $j\leftrightarrow Y$和$k\leftrightarrow Z$. 通过改变常数μ, ξ, λ, ζ的值, 最终建立起了一个维度M=160的结构空间. 虽然对于单一的$G_Y^X(i;\mu)$和$\Psi_{YZ}^X(i;\xi,\lambda,\zeta)$而言, 给出的结构信息非常有限, 但是对于给定一组这样的数值, 情况就大大不同, 即可以完整地描述局域结构. Liu课题组通过非仿射形变量 $D_{\min}^2\equiv\min_{\wedge}\left\{\frac{1}{z}\sum_j\left[R_{ij}(t+\Delta t)-\wedge R_{ij}(t)\right]^2\right\}$ 来表征原子在时间间隔Δt内的形变量, 其中, i表示中心原子, j表示近邻原子, z为中心原子的近邻原子数, $R_{ij}(t)=\mathbf{r}_i(t)-\mathbf{r}_j(t)$, $\wedge$为局域变形梯度张量. 如果$D_{\min}^2(i)$大于某一截断值$D_{\min,0}^2$, 则认为原子i在Δt内发生了重排, 标记为r_i=1; 反之, 则认为原子未发生重排, 标记为r_i=0. 这里, Liu课题组选取截断值$D_{\min,0}^2$, 使得$D_{\min}^2(i)$值为前0.15%的原子发生结构重排. 在此基础上, 使用支持向量机算法对发生重排的原子和没发生重排的原子进行分类, 并构造超平面. 需要注意的是, 虽然在构造超平面的过程中通过动力学量给定标签, 但是这个超平面仅仅表征的是结构属性, 它表征发生重排的原子与未发生重排的原子在结构上有什么区别. 一旦超平面被确定下来, 那么就可以从结构上区分原子是否容易发生重排, 即是否是所谓的流动缺陷.

上述机器学习方法是否能有效地识别原子是否发生结构重排呢? Liu课题组随后将上述方法应用到三个体系当中用以验证, 即二维颗粒柱的压缩实验、二维以及三维Lennard-Jones玻璃的剪切模拟. 如图3所示, 对于三个不同的体系及其学习模型都显示, 被识别为“软”的原子比例随着$D_{\min}^2$增大而增大. 这表明被SVM识别为“软”的原子更有可能参与塑性变形. 同时也说明了这些被识别出来的原子并不是将来会参与重排的特定原子, 而是有可能会发生重排的原子群. 但是, 由于在剪切或热力学过程中的涨落导致重排事件具有随机性, 因此对倾向于发生结构重排的原子群的识别更有意义.

Liu课题组运用支持向量机算法又研究了过冷液体的结构(软度)与动力学间的关联. 研究结果表明, 体

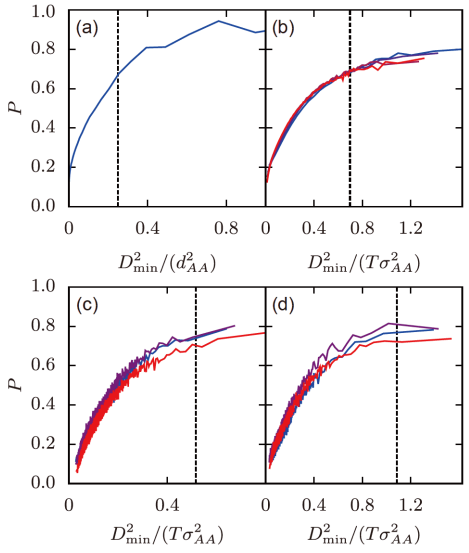

图 3 (网络版彩图)对于给定$D^2_{\min}$的原子被识别为“软”原子的概率[39]. (a) 二维柱的结果, $D^2_{\min}$用d^2_{AA}进行约化, d_{AA}为大颗粒的直径; (b) 二维L-J体系在T=0.1, 0.2, 0.3, 0.4下的分类结果; (c), (d)分别为3维L-J体系在T=0.4, 0.5, 0.6下A, B原子的分类结果; (b)–(d), $D^2_{\min}$均用σ^2_{AA}进行约化, σ_{AA}为L-J势中两个A原子之间的平衡距离

Figure 3 (Color online) Probability that a particle of a given $D^2_{\min}$ value is soft. (a) The result of the two-dimensional pillar system, $D^2_{\min}$ is reduced by d^2_{AA}, where d_{AA} refers to the large grain diameter; (b) the classification results of the two-dimensional L-J system at T = 0.1, 0.2, 0.3, and 0.4; (c), (d) results for species A and B, respectively, for the 3-dimensional L-J system at T=0.4, 0.5 and 0.6; (b)–(d), $D^2_{\min}$ is reduced by σ^2_{AA}, where σ_{AA} is the equilibrium distance between the two A atoms in the L-J potential (reprinted with permission from ref. [39]).

系动力学的变慢与软度的演化具有对应关系[41]. 在玻璃转变过程中, 结构没有明显的变化. 虽然, 过冷液体的动力学特征可以由假设具有均匀局域结构的平均场理论定性地描述, 然而研究发现, 实际体系的结构和动力学之间的关联很微弱[38,42,43]. 那么, 对于三维体系而言, 结构对动力学是否重要? 与寻找流动缺陷的方法类似, 需要定义弛豫过程的结构重排. 对于力学加载而言, 可以用非仿射形变量$D^2_{\min}$来表征原子是否发生结构重排; 而对于弛豫过程而言, 利用函数$p_{\text{hop}}(t)=\sqrt{\left\langle\left(\mathbf{r}_i-\langle\mathbf{r}_i\rangle_B\right)^2\right\rangle_A\left\langle\left(\mathbf{r}_i-\langle\mathbf{r}_i\rangle_A\right)^2\right\rangle_B}$ [44,45], 从动力学的角度识别原子是否发生重排. 其中, $\mathbf{r}_i$表示i原子的位置矢量, $\langle\cdots\rangle_A$, $\langle\cdots\rangle_B$分别表示在时间间隔A=[t–t_R/2, t]和B=[t, t+t_R/2]内的平均, t_R表征原子重排的时间尺度. 如果原子发生结构重排, p_{hop}将很大. 当超平面确定后, 可以定义软度S_i为原子在高维空间中距离超平面的最短距离. 当S_i>0时, 原子位于超平面的“软”侧; 当S_i<0时, 原子位于超平面的“硬”侧. 研究发现, 原子的软度具有很强的空间关联性; 并且整个体系中只有不到一半原子的S_i>0; 但是对于经历了结构重排的原子而言, 有90%原子的S_i>0. 因此, 对于所选择的结构函数集而言, 软度能很好地预测原子是否发生结构重排. 值得注意的是, S_i的数值越大则原子的动力学越快. 尽管一些比较容易计算的结构参量, 如局域势能和配位数, 与软度和动力学都有很强的关联, 但是它们只能预测60%–65%的结构重排. 而用支持向量机所定义为“软”的原子可以预测90%左右的结构重排, 极大地提高了模型的预测能力.

3.3 预测非晶材料宏观性质

提高合金的玻璃形成能力是非晶合金材料制备中需要解决的首要问题, 不仅涉及制备工艺, 更与合金液体的基本物理问题密切相关, 直接决定着非晶合金材料的力学性能和稳定性等. 如何设计并开发具有良好玻璃形成能力的合金, 一直是一个具有重要产业价值的基础科学问题. 合金的玻璃形成能力通常被定义为临界冷却速率, 当冷却速率高于临界冷却速率时, 液体将会避免晶化并发生玻璃转变形成玻璃. 为了设计和开发出具有良好玻璃形成能力的合金, 研究人员提出了各种方法或参量来预测合金的玻璃形成能力. 例如, 与玻璃转变温度T_g相关的T_{rg}, T_{gl}, $T_x/(T_g-T_l)$等, 以及几何堆积[46–48]、混合焓[49]、关联长度[50]等物理量. 但是, 这些参量都不能全面地解释合金的玻璃形成能力. 这是由于在玻璃形成的过程中, 许多变量都起到了关键的作用, 因此用单一的参量无法全面地描述玻璃形成能力. 受到经验性判据的准确性和通用性的限制, 新型高性能非晶合金材料的研发进程非常缓慢. 如何提高材料设计的效率, 寻找具有更优性能的材料, 是一个非常具有挑战性的问题.

在大数据的时代背景下, 机器学习这种先进的数据分析方法能够综合分析各种影响合金玻璃形成能力的实验和理论数据, 建立较全面的理论模型, 为解决非晶合金材料领域的关键瓶颈问题提供了新的途径和契

机. 最近, Sun等人[51]利用支持向量机的机器学习算法, 对二元合金的玻璃形成能力进行了系统分析, 建立了合金成分与性能之间的关联, 并对可能具有良好玻璃形成能力的新材料进行了预测. 他们利用支持向量机算法, 对所构建的多维空间内的带有标签的数据进行分割, 从而建立输入参量与输出参量之间的关联. 如图4所示, 整个建模过程分为四个阶段, 即数据库的建立、SVM模型训练、模型评估和模型预测. 需要注意的是, 由于训练数据集不够大, 传统的将数据样本切割成小子集的交叉验证方法就不适用了. 通过调整具有径向基函数的SVM的两个参数C与γ, 可以得到不同的SVM模型, 并将模型应用于目标组和整体组, 分别得到玻璃形成能力好的样品的概率$P_{目标}$和$P_{整体}$, 并定义参量$E=P^2_{目标}/P_{整体}$, 用以评估模型的预测能力. 其中, “目标组”包含339个已经被报道可以通过甩带方式形成金属玻璃的成分, “整体组”包括所有能得到输入数据的二元合金成分, 共1131对. 最好的SVM模型必须同时满足$P_{目标}>0.3$和E最大.

通过选取原子质量($aw1$, $aw2$)、混合焓(ΔH)、原子半径($r1$, $r2$)、元素单质的液化温度($T_{liq}1$, $T_{liq}2$)、虚拟液化温度(T_{fic})、液化温差(ΔT_{liq})、元素组分($c1$, $c2$)中的一个或多个参量组合对模型进行重复训练, 可以探讨合金的不同性质对其玻璃形成能力的影响. 其中, $T_{fic}=T_{liq}1\cdot c1+T_{liq}2\cdot c2$, $\Delta T_{liq}=(T_{fic}-T_{liq})/T_{fic}$. 如图5所示, 上述各参量与玻璃形成能力之间有着不同程度的关联, E越大表示关联越显著. 其中液化温差ΔT_{liq}与合金的玻璃形成能力之间的关联最为显著, 而且组合参量ΔT_{liq}与T_{fic}作为输入参量, 可以得到具有最佳预测效率的SVM模型.

从图6中能够看出最佳SVM预测效率模型的可靠性. 不同的颜色代表输入不同参量ΔT_{liq}与T_{fic}后, 最佳预测模型预测的玻璃形成能力P_{GFA}的分布, 红色区域表示具有最佳玻璃形成能力的区域, 其中玻璃形成能力P_{GFA}定义为数据样本距离超平面的最短距离. 符号“×”标记了已知的具有良好玻璃形成能力的二元合金的分布. 尽管还有一部分玻璃形成能力好的二元合金落在红色区域之外, 但是就整体而言这两者之间具有很好的一致性. 使用这个模型, 可以对未知的合金成分进行预测, 这样由已有的实验数据分析指导设计实验, 可以极大地缩短非晶合金材料的研发周期. 该工作作为使用新的工具对经典问题进行分析的一种尝试, 得到了初步成果. 这表明, 机器学习的方法在非晶材料设计与研发领域具有重要的应用前景. 采用更全面、完善的数据库, 运用更深入的人工智能算法, 机器学习方法能够为非晶等领域研究人员提供更精准的信息, 进一步加速新材料的研发过程.

改善玻璃的机械性能对于解决能源、通信和基础设施方面的重大挑战至关重要[52]. 发现性能增强的新材料一直是一项艰巨的任务, 对于非晶材料而言更是提出了一些独特的挑战. 只要降温速率足够快, 周期表中所有的元素都可以制备成非晶样品. 与晶体不同, 由于非晶结构无序, 没有固定的晶体相, 因此在制备过

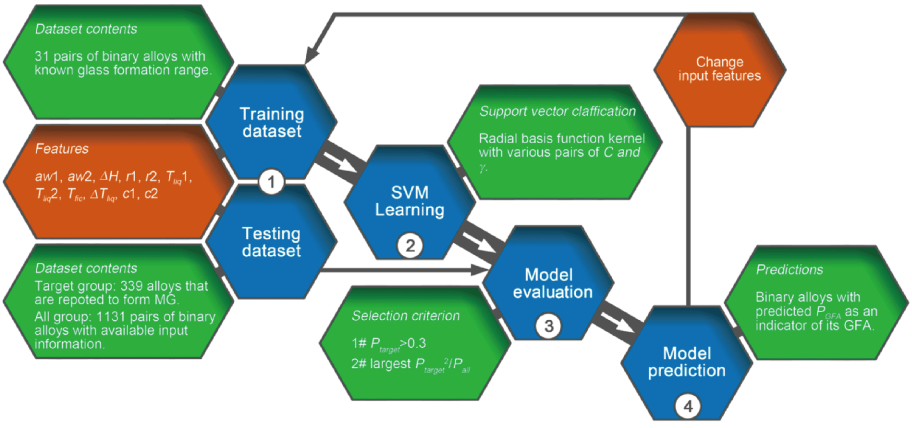

图 4 (网络版彩图)支持向量机算法预测玻璃形成能力的基本过程, 包括数据库的建立、SVM模型训练、模型评估和模型的预测[51]

Figure 4 (Color online) The basic process of predicting glass formation ability by support vector machine algorithm, includes database establishment, SVM model training, model evaluation and model prediction (reprinted with permission from ref. [51]).

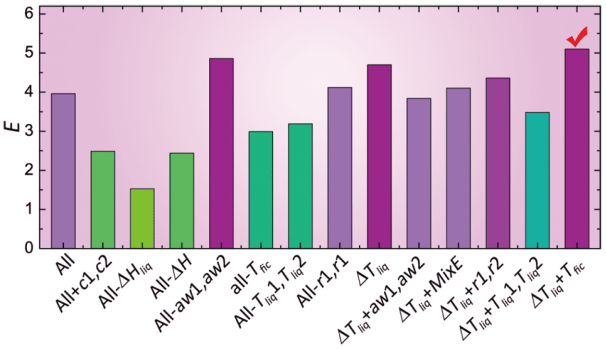

图 5 (网络版彩图)根据不同输入参量训练所得SVM模型评估, E表示SVM模型的预测效率, E越大, 模型效果越好[51]

Figure 5 (Color online) According to the evaluation of SVM model trained by different input parameters, E represents the prediction efficiency of SVM model. The greater the E, the better the effect of the model (reprinted with permission from ref. [51]).

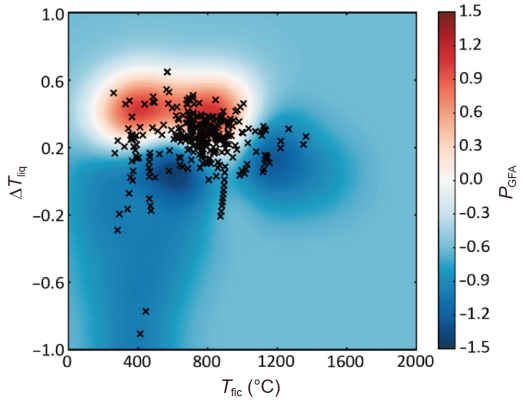

图 6 (网络版彩图)最佳SVM模型预测玻璃形成能力P_{GFA}[51]. 从蓝到红, 玻璃形成能力依次变好. P_{GFA}表示具有良好的玻璃形成能力

Figure 6 (Color online) The best SVM model predicts glass formation ability P_{GFA}. From blue to red, glass formation ability becomes better in turn. P_{GFA}>0 indicates good glass formation ability (reprinted with permission from ref. [51]).

程中不需要特定的化学组分. 正是因为非晶制备过程中的特点, 为开发增强性能的新材料提供了无限的可能. 也正是由于这些原因, 非晶材料的成分调控具有很大的不确定性. 如果存在一个预测模型, 那么新材料的开发将变得方便很多. 在理想情况下, 物理模型可以提供最大的预测. 对于玻璃而言, 其杨氏模量的预测最有效的模型是Makishima-Mackenzie (MM)模型[53,54], 但是它无法很好地识别任何非线性的依赖关系. 而机器学习方法则与基于物理的模型方法不同, 基于机器学习的模型是通过对数据库的"学习"而生成的. 由于机器学习的"学习"过程需要大量完整、一致、准确的数据, 但是在数据库没有建立起来的情况下获取并处理这些数据是非常困难的.

为此, Yang等人[55]结合高通量计算模拟和人工神经网络的方法, 预测了硅酸盐玻璃的杨氏模量与其组分的关系. 他们首先通过高通量计算模拟的方法模拟了231个CaO-Al_2O_3-SiO_2玻璃体系, 并计算了它们的杨氏模量. 使用高通量分子动力学模拟可以高效、系统、全面地探索所有组分. 图7(a)展示了高通量分子动力学模拟的杨氏模量E随着不同组分的变化. 总的来说, 可以看出, 杨氏模量分别随着Al_2O_3和CaO含量的增加而增加. 但是, 杨氏模量对组分的依赖性是非线性的. 现在最常用的预测杨氏模量的方法即MM模型预测杨氏模量随着组分的变化. 如前文所述, MM模型只能预测线性关系, 而对于非线性的依赖关系则没有那么好的预测. 并且, MM模型预测的杨氏模量与分子动力学模拟所得到的杨氏模量在数值上有所差异. 因此, MM模型只能推断某些组分的粗略的变化趋势,

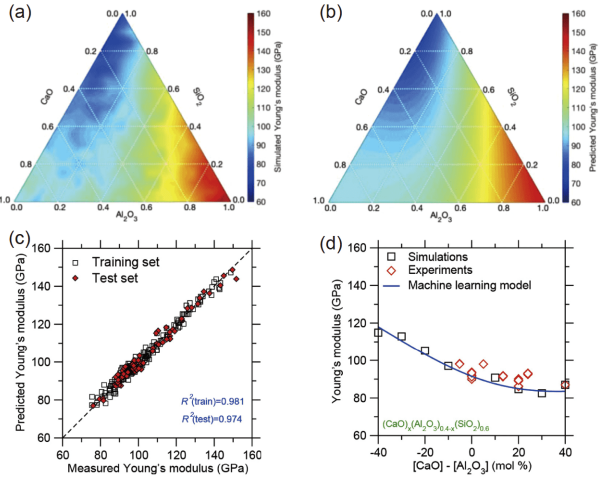

图 7 (网络版彩图)三角图表示杨氏模量E随着组分的变化. (a) 高通量分子动力学模拟所得的结果; (b) 人工神经网络(ANN)预测的结果; (c) ANN模型预测的结果与高通量计算模拟的结果之间的比较; (d) 分子动力学模拟数值、实验数值和ANN模型预测数值之间的对比[56]

Figure 7 (Color online) The trigonometric diagram shows the change of Young's modulus E with the composition. (a) Results from high-throughput molecular dynamics simulations; (b) results from artificial neural network (ANN) predictions; (c) comparison between the results predicted by the ANN model and the results of high-throughput computational simulations; (d) comparison of molecular dynamics simulation values, experimental values, and ANN model prediction values (reprinted with permission from ref. [56]).

不能准确预测硅酸盐玻璃的杨氏模量.

Yang等人[55]为了找到一个合适的机器学习算法预测硅酸盐玻璃的杨氏模量, 分别应用多项式回归(Polynomial Regression, PR)、LASSO、随机森林(Random Forest, RF)和人工神经网络(ANN)的机器学习算法训练模型并预测数值. 对比发现, PR, LASSO, RF以及ANN都可以拟合成分与杨氏模量之间的非线性关系, 并且能够对杨氏模量值进行实际预测. 这些机器学习算法在预测杨氏模量方面比MM模型更准确. 评估各个机器学习算法在预测杨氏模量上的表现后发现, 人工神经网络对杨氏模量的预测最为准确. 图7(b)展示了人工神经网络训练出的模型所给出的杨氏模量的预测值. 可以看到, 人工神经网络很好地拟合了杨氏模量随着组分的变化关系. 从图7(c)可知, 无论是训练数据集还是测试数据集, 都展示了人工神经网络得到的预测数据和高通量分子动力学模拟得到的数据有很好的一一对应关系. 图7(d)展示了人工神经网络模型预测数值与高通量模拟得到的数值具有很好的对应关系, 同时与实验得到的数值也存在很好的对应关系.

4 总结与展望

虽然机器学习在非晶材料领域已经有许多应用, 如结构的分类、结构-动力学关联、宏观性能预测等, 但是仍旧处于一个刚刚起步的探索阶段. 因此, 将机器学习应用到非晶领域中还存在许多困难之处. 这些困难主要源于机器学习所需数据的缺乏以及模拟与实验之间的差异.

非晶材料设计研发与人们对玻璃形成能力的认识密切相关. 现有的研究从热力学、动力学和结构等角度综合理解非晶合金的形成, 提出了一些经验准则和参数来评估合金的玻璃形成能力. 由于大多数参数是与特征温度(如T_g, T_x, T_l等)密切相关, 因此只能在实际制备出玻璃态样品后进行测量, 进而评估其玻璃形成能力. 利用这些参数无法在制备样品之前对玻璃形成能力进行预测, 缺乏对非晶材料设计开发的指导作用. 新材料的开发有很多种驱动模式, 如基于样品制备与表征的手段属于实验驱动模式, 相图计算属于计算驱动模式, 机器学习方法则属于数据驱动模式[57]. 非晶材料的玻璃形成能力由多个因素共同影响, 并且需要归纳分析大量数据, 因此应用传统的方法进行分析存在巨大的困难, 但这正是机器学习方法的优势所在. 这给用机器学习方法解决玻璃形成能力问题提供了可能性. 但是, 作为数据驱动模式的机器学习方法, 将它运用到非晶材料领域的过程中会存在着一些挑战.

4.1 非晶数据库的建立

将机器学习应用在非晶材料领域最大的挑战就是数据库的建立. 虽然有着长期的实验累积, 但是要把能够收集到的原始数据转化为可以进行机器学习的数据库, 需要很大的工作量. 而且, 文献中所报道的数据只选取了部分成功的实验数据. 而对于机器学习方法, 完整的实验数据能够更有效地训练模型、提高模型精度. 虽然, 已经有一些非晶数据库可以使用[58], 但是不同课题组生成的实验数据之间的不一致对机器学习方法的应用带来了很多限制. 只要拥有可获得、完整、一致、精确以及数量多等特征的数据, 机器学习方法就可以发挥其优势, 从已有的数据中发掘出重要的信息, 对实验、新材料的设计研发进行针对性的指导. 因此, 建立完整、规范的非晶材料数据库是非常有必要的.

那么需要建立什么样的数据库? 在非晶材料领域中, 最典型、最受关注的问题就是非晶材料的玻璃形成能力. 如前文所述, 已经有研究表明, 玻璃形成能力与T_{rg}, K_{g1}, $T_x/(T_g-T_l)$等参量相关, 这些参量都与玻璃转变温度T_g相关, 以及与合金组分、原子半径差异、几何堆积、混合焓、关联长度等物理量相关. 因此, 我们需要包含上述数据的数据库, 用以研究非晶体系的玻璃形成能力. 除此之外, Fe基非晶材料由于拥有优异的软磁性能而受到广泛关注. Fe基非晶材料的组分、含量等变化对软磁性能具有很大的影响, 例如掺杂少量Cu等其他元素时, 可以提高Fe基非晶的软磁性能, 但是它们的物理机制还不清楚. 因此, 想要开发尽可能少元素组分的Fe基非晶材料还存在一定的困难. 这就需要建立起具有Fe基非晶的组分、含量、原子半径差以及软磁性能等物理量的数据库, 用以辅助开发具有更高性能的Fe基非晶.

在当前数据库还没建立或者说还没完全建立的情况下, 是否就没办法应用机器学习来处理非晶材料的典型问题呢? 事实上, 在数据量不够的情况下也有办法应用机器学习方法. 一方面, 目前已经发展了一些

处理小样本数据的机器学习方法，如朴素贝叶斯算法、决策树以及其他线性模型. 通常，机器学习算法越简单，处理小样本数据的效果越好. 因为小样本数据需要低复杂度的模型，以避免过度拟合的情况出现. 另一方面，将所有的数据利用某些方法进行预处理后，就可以提高模型预测的精度. 有研究表明, 通过分析数据集与机器学习预测能力后发现，小样本数据不是直接影响模型的精度，而是以模型的自由度作为中介，从而导致了精度和自由度之间存在关联现象[59]. 然而，通过增加额外的实验数据而提高精度，需要增加大量的实验成本，并且成本与精度的提升并不成正比. 因此，通过某些方法提高模型的精度而不需要更高的模型自由度是材料特性建模中实践机器学习的关键挑战. 在机器学习方法应用到材料领域之前, 已经出现了许多预测性质的方法. 尽管这些方法预测的结果不尽如人意，但是如果将这些预测结果应用到机器学习方法中就可以提高小数据样本机器学习模型的精度. 这些在预测二元半导体带隙、晶格导热系数以及沸石的弹性模量三个研究中得到了验证. 如何将这种方法应用在非晶材料领域中还需要进一步研究.

4.2 高通量计算的发展

为了缩短材料研发到应用的周期，高通量制备实验方法应运而生. 高通量制备，就是在短时间内制备出大量样品，然后从大量样品中筛选出具有优异性能的目标材料[60]. 与高通量实验制备相对应的计算方法是高通量计算模拟，二者在理念上并无差异. 高通量计算模拟也是希望一次性计算大量体系，利用计算的方法，快速并且无盲区地对目标体系进行全面的结构、性能分析，从而得出具有最佳性能的材料. 近年来，高通量计算已经被广泛地运用到材料科学领域中. Andersson等人[61]采用高通量的密度泛函研究了60多种合金表面对氧的结合能力及其对甲烷化反应的催化活性，设计出了效率更高的Ni-Fe合金催化剂. 基于第一性原理，Curtarolo等人[62]设计了高通量计算程序AFLOW，获得了15万种合金的热力学参数和13000种化合物的电子结构. Yang等人[63]采用高通量的计算方法成功挑选出了28种拓扑绝缘体. 然而, 关于非晶材料的高通量计算模拟，至今只有个别工作已经发表[55]. 也就是说，非晶合金的高通量计算依然是相对空白，这一方面说明非晶材料的高通量计算模拟拥有很大的发展前景，另一方面说明开展非晶材料的高通量计算模拟非常迫切.

4.3 势函数的拟合

非晶合金的优异物性很大程度上来源于其无序的原子排列结构，深入理解非晶态材料结构与性能的关联，是解决非晶材料领域重要问题的关键. 从微观原子排列结构出发理解材料结构与性能的关联是所有材料科学领域面临的共性问题，金属玻璃的无序结构使得这个问题更具挑战. 长期以来，金属玻璃领域围绕结构与性能关联性的研究进展缓慢，结构上的无序和能量上的亚稳极大地限制了实验上对非晶合金材料的制备、表征，建立材料宏观性能与原子排列结构跨尺度关联性更是难上加难. 尽管现代结构分析技术发展迅速，但是由于非晶材料缺乏周期性重复的结构单元，通过二维信息还原三维结构的方法无法在非晶材料中适用，而真正对金属玻璃三维原子排列重构的技术对时间和计算的消耗巨大，还没能生产出具有统计意义的结果. 分子动力学模拟对解决上述问题具有一定的指导意义.

分子动力学模拟是基于粒子之间的相互作用求解经典牛顿方程得到体系的演化过程，如何描述粒子之间的相互作用是分子动力学模拟的关键问题. 相互作用的选取与模拟结果的可靠性和真实性相关. 针对各种不同的模型体系，人们开发出了各种各样的相互作用势函数来描述粒子之间的相互作用，目前比较精确的描述方法是第一性原理的方法. 第一性原理计算基于密度泛函理论，通过计算电子密度分布来描述体系物理、化学性质. 尽管该方法是目前最精确的计算方法, 但是由于计算量大, 目前只能计算较小的体系, 模拟时长也因此受到限制. 对于更大的体系或更长时间的模拟计算，必须牺牲一定的精度. 早期发展了一些简单的势函数如Lennard-Jones势，但是这些势函数无法描述真实的金属体系. 在保证一定精度的前提下，人们又发展了嵌入原子势，它是用以描述金属体系相互作用的势函数. 用经典的分子动力学的方法可以模拟超过10000个原子的体系，模拟时长超过1 ns. 嵌入原子势的方法，是通过一个成对的原子相互作用项和一个多体局域密度项来描述体系的能量. 这种方法又发展出不同的势函数形式，用以描述不同的体系，但其本质都是相同的. 如何去平衡模拟时长、模拟体系

大小以及模拟精度这三者是非晶计算领域中的一个重要问题. 另一方面, 有很多合金体系, 尤其是多元合金体系的嵌入原子势函数还未被开发.

为解决这些问题, 用机器学习拟合势函数应运而生. 机器学习方法往往能在很短的时间内给出第一性原理级别精度的能量、力等合金性质的预测. 近些年来, 机器学习经验势已经得到了广泛的关注并取得了显著的进展. 现阶段势函数拟合研究最主要的是基于神经网络的势函数拟合[40,64–66]. 但是非晶合金所需要的势函数还很少有人拟合, 因此发展势函数以模拟更大的时间尺度、空间尺度以及模拟精度是非常必要的. 同时也需要克服人工神经网络的计算速度比嵌入势的计算时间要大两个数量级的问题.

2007年, 图灵奖获得者Jim Gray[67]提出了科学研究的第四范式——数据密集型科学发现. 他认为, 人类科学研究活动已经历过三种不同范式的演变过程: 原始社会的“实验科学范式”、以模型和归纳为特征的“理论科学范式”、以模拟仿真为特征的“计算科学范式”. 目前正在从“计算科学范式”转向“数据密集型科学发现范式”. 第四范式, 即“数据密集型科学发现范式”的主要特点是科学研究人员只需要从大数据中查找和挖掘所需要的信息和知识, 无须直接面对所研究的物理对象. 这为科学研究注入变革性元素, 为非晶乃至整个材料领域的研究提供了新思路和新契机.

参考文献

1 Lopez de Mantaras R, Armengol E. Machine learning from examples: Inductive and lazy methods. Data Knowledge Eng, 1998, 25: 99–123
2 Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects. Science, 2015, 349: 255–260
3 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484–489
4 Kim J, Park C. End-to-end ego lane estimation based on sequential transfer learning for self-driving cars. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Honolulu: IEEE, 2017. 30–38
5 He K M, Zhang X Y, Ren S Q, et al. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 1026–1034
6 Liu S S, Tian Y T. Facial expression recognition method based on gabor wavelet features and fractional power polynomial kernel PCA. Lect Notes Comput Sc, 2010, 6064: 144–151
7 Pazzani M, Billsus D. Learning and revising user profiles: The identification of interesting web sites. Machine Learning, 1997, 27: 313–331
8 Chan P K, Stolfo S J. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. New York: AAAI Press, 1998. 164–168
9 Guzella T S, Caminhas W M. A review of machine learning approaches to Spam filtering. Expert Syst Appl, 2009, 36: 10206–10222
10 Huang C L, Chen M C, Wang C J. Credit scoring with a data mining approach based on support vector machines. Expert Syst Appl, 2007, 33: 847–856
11 Tang T T, Zawaski J A, Francis K N, et al. Image-based classification of tumor type and growth rate using machine learning: A preclinical study. Sci Rep, 2019, 9: 12529
12 Rastegar D A, Ho N, Halliday G M, et al. Parkinson’s progression prediction using machine learning and serum cytokines. npj Parkinson’s Disease, 2019, 5: 14
13 Makino M, Yoshimoto R, Ono M, et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Sci Rep, 2019, 9: 11862
14 Cannata A, Cannavò F, Moschella S, et al. Exploring the link between microseism and sea ice in Antarctica by using machine learning. Sci Rep, 2019, 9: 13050
15 Alobaidi M H, Meguid M A, Chebana F. Predicting seismic-induced liquefaction through ensemble learning frameworks. Sci Rep, 2019, 9: 11786
16 Rem B S, Käming N, Tarnowski M, et al. Identifying quantum phase transitions using artificial neural networks on experimental data. Nat Phys, 2019, 15: 917–920
17 Ming Y, Lin C T, Bartlett S D, et al. Quantum topology identification with deep neural networks and quantum walks. npj Comput Mater, 2019, 5: 88

18 Xia R, Kais S. Quantum machine learning for electronic structure calculations. Nat Commun, 2018, 9: 4195

19 Alvarez-Rodriguez U, Lamata L, Escandell-Montero P, et al. Supervised quantum learning without measurements. Sci Rep, 2017, 7: 13645

20 Hohenberg P, Kohn W. Inhomogeneous electron gas. Phys Rev, 1964, 136: B864–B871

21 Kohn W, Sham L J. Self-consistent equations including exchange and correlation effects. Phys Rev, 1965, 140: A1133–A1138

22 Alder B J, Wainwright T E. Studies in molecular dynamics. I. General method. J Chem Phys, 1959, 31: 459–466

23 Rahman A. Correlations in the motion of atoms in liquid argon. Phys Rev, 1964, 136: A405–A411

24 Olson G B. Designing a new material world. Science, 2000, 288: 993–998

25 Oganov A R, Glass C W. Crystal structure prediction using *ab initio* evolutionary techniques: Principles and applications. J Chem Phys, 2006, 124: 244704

26 Tran K, Ulissi Z W. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. Nat Catal, 2018, 1: 696–703

27 Gómez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel T D, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nat Mater, 2016, 15: 1120–1127

28 Zhu Q, Samanta A, Li B, et al. Predicting phase behavior of grain boundaries with evolutionary search and machine learning. Nat Commun, 2018, 9: 467

29 Isayev O, Oses C, Toher C, et al. Universal fragment descriptors for predicting properties of inorganic crystals. Nat Commun, 2017, 8: 15679

30 Raccuglia P, Elbert K C, Adler P D F, et al. Machine-learning-assisted materials discovery using failed experiments. Nature, 2016, 533: 73–76

31 Mountrakis G, Im J, Ogole C. Support vector machines in remote sensing: A review. ISPRS J Photogrammetry Remote Sens, 2011, 66: 247–259

32 Goh G B, Hodas N O, Vishnu A. Deep learning for computational chemistry. J Comput Chem, 2017, 38: 1291–1307

33 Zhang G P. Neural networks for classification: A survey. IEEE Trans Syst Man Cybern C, 2000, 30: 451–462

34 Ralambondrainy H. A conceptual version of the K-means algorithm. Pattern Recognition Lett, 1995, 16: 1147–1157

35 Voronoi G. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Deuxième mémoire. Recherches sur les parallélloèdres primitifs. J für die reine und angewandte Mathematik, 1908, 134: 198–287

36 Maldonis J J, Banadaki A D, Patala S, et al. Short-range order structure motifs learned from an atomistic model of a $Zr_{50}Cu_{45}Al_5$ metallic glass. Acta Mater, 2019, 175: 35–45

37 Banadaki A D, Maldonis J J, Voyles P M, et al. Point-pattern matching technique for local structural analysis in condensed matter. arXiv: 1811.06098

38 Manning M L, Liu A J. Vibrational modes identify soft spots in a sheared disordered packing. Phys Rev Lett, 2011, 107: 108302

39 Cubuk E D, Schoenholz S S, Rieser J M, et al. Identifying structural flow defects in disordered solids using machine-learning methods. Phys Rev Lett, 2015, 114: 108001

40 Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys Rev Lett, 2007, 98: 146401

41 Schoenholz S S, Cubuk E D, Sussman D M, et al. A structural approach to relaxation in glassy liquids. Nat Phys, 2016, 12: 469–471

42 Patrick Royall C, Williams S R, Ohtsuka T, et al. Direct observation of a local structural mechanism for dynamic arrest. Nat Mater, 2008, 7: 556–561

43 Jack R L, Dunleavy A J, Royall C P. Information-theoretic measurements of coupling between structure and dynamics in glass formers. Phys Rev Lett, 2014, 113: 095703

44 Candelier R, Widmer-Cooper A, Kummerfeld J K, et al. Spatiotemporal hierarchy of relaxation events, dynamical heterogeneities, and structural reorganization in a supercooled liquid. Phys Rev Lett, 2010, 105: 135702

45 Smessaert A, Rottler J. Distribution of local relaxation events in an aging three-dimensional glass: Spatiotemporal correlation and dynamical heterogeneity. Phys Rev E, 2013, 88: 022314

46 Laws K J, Miracle D B, Ferry M. A predictive structural model for bulk metallic glasses. Nat Commun, 2015, 6: 8123

47 Zhang K, Liu Y, Schroers J, et al. The glass-forming ability of model metal-metalloid alloys. J Chem Phys, 2015, 142: 104504

48 Zhang K, Dice B, Liu Y, et al. On the origin of multi-component bulk metallic glasses: Atomic size mismatches and de-mixing. J Chem Phys, 2015, 143: 054501

49 Ramakrishna Rao B, Gandhi A S, Vincent S, et al. Prediction of glass forming ability using thermodynamic parameters. Trans Ind Inst Met, 2012,

65: 559–563
50 Bian X, Guo J, Lv X, et al. Prediction of glass-forming ability of metallic liquids. Appl Phys Lett, 2007, 91: 221910
51 Sun Y T, Bai H Y, Li M Z, et al. Machine learning approach for prediction and understanding of glass-forming ability. J Phys Chem Lett, 2017, 8: 3434–3439
52 Wondraczek L, Mauro J C, Eckert J, et al. Towards ultrastrong glasses. Adv Mater, 2011, 23: 4578–4586
53 Makishima A, Mackenzie J D. Direct calculation of Young's modulus of glass. J Non-Crystalline Solids, 1973, 12: 35–45
54 Makishima A, Mackenzie J D. Calculation of bulk modulus, shear modulus and Poisson's ratio of glass. J Non-Crystalline Solids, 1975, 17: 147–157
55 Yang K, Xu X, Yang B, et al. Predicting the Young's modulus of silicate glasses using high-throughput molecular dynamics simulations and machine learning. Sci Rep, 2019, 9: 8739
56 Liu H, Fu Z, Yang K, et al. Machine learning for glass science and engineering: A review. J Non-Crystalline Solids-X, 2019, 4: 100036
57 Mauro J C. Decoding the glass genome. Curr Opin Solid State Mater Sci, 2018, 22: 58–64
58 Priven A I, Mazurin O V. Glass property databases: Their history, present state, and prospects for further development. Adv Mater Res, 2008, 39-40: 147–152
59 Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater, 2018, 4: 25
60 Li M X, Zhao S F, Lu Z, et al. High-temperature bulk metallic glasses developed by combinatorial methods. Nature, 2019, 569: 99–103
61 Andersson M, Bligaard T, Kustov A, et al. Toward computational screening in heterogeneous catalysis: Pareto-optimal methanation catalysts. J Catal, 2006, 239: 501–506
62 Curtarolo S, Setyawan W, Hart G L W, et al. AFLOW: An automatic framework for high-throughput materials discovery. Comput Mater Sci, 2012, 58: 218–226
63 Yang K, Setyawan W, Wang S, et al. A search model for topological insulators with high-throughput robustness descriptors. Nat Mater, 2012, 11: 614–619
64 Lorenz S, Groß A, Scheffler M. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. Chem Phys Lett, 2004, 395: 210–215
65 Cubuk E D, Malone B D, Onat B, et al. Representations in neural network based empirical potentials. J Chem Phys, 2017, 147: 024104
66 Handley C M, Popelier P L A. Potential energy surfaces fitted by artificial neural networks. J Phys Chem A, 2010, 114: 3371–3383
67 Hey T, Trnsley S, Tolle K. The Fourth Paradigm-Data-Intensive Scientific Discovery. Remond, Washington: Microsoft Research, 2009

Application of machine learning approach in disordered materials

WU JiaQi[1], SUN YiTao[2], WANG WeiHua[2] & LI MaoZhi[1*]

[1] *Department of Physics, Beijing Key Laboratory of Opto-electronic Functional Materials & Micro-nano Devices, Renmin University of China, Beijing 100872, China;*
[2] *Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China*

As a new amorphous material, metallic glass has been widely studied because of its excellent mechanical, physical and chemical properties. Glass-forming ability has always been an important problem restricting the development of amorphous materials. In order to design amorphous materials with good glass-forming ability, a lot of research has been done on the glass-forming ability of amorphous materials. It has been shown that a single factor cannot describe glass-forming ability of amorphous materials. Because of the complex and disordered structure of amorphous materials, it is difficult to understand the structure and nature of amorphous materials comprehensively and clearly by traditional methods. Machine learning method, as a new research paradigm, provides new opportunities for exploring these bottleneck scientific issues in disordered materials. In this paper, some machine learning algorithms, such as support vector machine, artificial neural network and *K*-means clustering, are briefly introduced. Moreover, we briefly review the application of machine learning approach in amorphous materials, including the classification of amorphous structure, the correlation between amorphous structure and properties, and the prediction of macroscopic properties of amorphous materials. We also discuss the future application of machine learning approach to disordered materials, including the development of comprehensive database, high throughput calculation method and the development of machine learning potential function.

machine learning, disordered materials, glass-forming ability, structure-property relationships

PACS: 02.70.-c, 61.43.Fs, 64.70.Pf , 61.25.Mv

doi: 10.1360/SSPMA-2019-0345