Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Smart Process Manufacturing toward Carbon Neutrality—Review

化学中的机器学习——基础与应用

史云飞[a,#]，杨正新[a,#]，马思聪[b]，康沛林[a]，商 城[a]，胡培君[c,*]，刘智攀[a,b,*]

[a] *Collaborative Innovation Center of Chemistry for Energy Material, Shanghai Key Laboratory of Molecular Catalysis and Innovative Materials, Key Laboratory of Computational Physical Sciences of the Ministry of Education, Department of Chemistry, Fudan University, Shanghai 200433, China*
[b] *Key Laboratory of Synthetic and Self-Assembly Chemistry for Organic Functional Molecules, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China*
[c] *School of Chemistry and Chemical Engineering, Queen's University Belfast, Belfast BT9 5AG, Northern Ireland*

ARTICLE INFO

Article history:
Received 31 August 2022
Revised 19 January 2023
Accepted 6 April 2023
Available online 31 July 2023

关键词
机器学习
原子模拟
催化
逆合成分析
神经网络势函数

摘要

过去的十年间，机器学习（ML）在科学研究中得到了广泛应用。本文介绍了机器学习的基本组成部分，包括数据库、特征和算法，并着重介绍了机器学习在化学领域取得的一些重要成就。首先我们介绍了一些化学方面最流行的数据库，这些数据库来自实验或理论模拟，收录了有关小分子或固体材料的各种数据。其次简要介绍了部分重要的表示小分子和固体材料化学环境的二维和三维特征。本文对决策树和深度学习神经网络算法进行了综述，重点介绍了它们的框架和典型应用场景。随后，我们讨论了机器学习在化学中的三个重要应用领域：①逆合成分析，通过机器学习预测有机物的合成途径；②原子模拟，利用机器学习势函数加速势能面采样；③多相催化，使用机器学习辅助催化设计中从合成条件优化到反应机理探索的各个方面。最后我们对机器学习在化学中的应用前景进行了展望。

1. 引言

长期以来，发明具有类人智能并且可以自动完成复杂任务的机器一直是人类的梦想。这个梦想在过去的十年中从未如此成真，这十年见证了机器学习（ML）技术和人工智能（AI）机器在人类活动各个领域的快速应用。新的ML模型的开发——特别是深度学习方法[1]——以及数据存储能力的急剧提高是最近ML案例激增的关键。在现代科学研究中，ML除了在日常生活中的成就，如图像识别[2]和语音识别[3]，也引起了大量的关注；例如，用于预测蛋白质结构的AlphaFold算法已经证明了其在结构生物学中改变规则的能力[4–5]。本文将综述ML在化学研究中应用的最新进展，化学研究本身包含了大量的数据，与材料的复杂性和有机分子的丰富多样性有关。

化学家虽然接受过进行实验和收集数据的教育，但通常不太熟悉现代ML算法[6]。与20世纪90年代主要基于理论/经验规则[7]的计算机辅助化学研究不同，目前的ML应用依赖于承载所有基本信息的大数据集[8–9]。质量不佳的数据集可能会给ML应用带来不必要的困难，这些应用原则上应该是可行和直接的[10]。化学数据集的一个

* Corresponding author.
E-mail address: p.hu@qub.ac.uk (P. Hu), zpliu@fudan.edu.cn (Z.-P. Liu)..
# These authors contributed equally to this work.

英文原文：Engineering 2023, 27(8): 70–83
引用本文：Yun-Fei Shi, Zheng-Xin Yang, Sicong Ma, Pei-Lin Kang, Cheng Shang, P. Hu, Zhi-Pan Liu. Machine Learning for Chemistry: Basics and Applications. *Engineering*, https://doi.org/10.1016/j.eng.2023.04.013

常见问题是偏重于成功的实验数据。事实上，为了提供化学领域的一个平衡视角，不仅需要好的数据（例如，生产所需的产品），还需要坏的数据（如失败的实验）。此外，由于化学实验的复杂性，文献中记录的合成条件往往不完整，重要的变量被忽略。由于这些原因，与实验领域相比，ML应用在计算化学中更受欢迎，在计算化学中，可以通过量子力学（QM）计算可靠、一致地构建数据集。这些计算数据集可以用来直接对小分子和固体材料的物理化学性质进行基准测试，并用来开发先进的计算方法。因此，化学家有必要掌握基本的机器学习知识，这将使他们从数据记录到实践ML引导的实验中受益匪浅。

为此，本文将首先介绍流行的化学数据库，这些数据库为实践ML模型提供了基础。其次，介绍了一些广泛使用的分子二维（2D）和三维（3D）特征表示，它们将分子结构转化为ML模型可接受的输入。然后简要介绍了流行的ML算法，重点介绍了它们的基本理论框架和合适的应用场景。最后，更详细地描述了在ML领域取得重要进展的三个化学领域，包括有机化学中的逆合成、基于ML势函数的原子模拟和多相催化的ML。这些应用要么通过降低实验/模拟成本极大地加快了原来的研究速度，要么为合理解决复杂问题提供了新的途径。最后对未来的挑战进行了展望。

2. 数据

没有数据就没有人工智能（AI）。因此，数据的可用性是现代机器学习应用的先决条件，其中数据集的大小和质量都很重要。在化学领域，收集和整合数据有着悠久的传统，这些数据涵盖了从元素原子光谱到材料的宏观特性的广泛范围。化学数据科学的兴起促进了化学信息学的发展，这进一步大大推动了ML在化学中的应用。事实上，尽管从头开始构建一个大型数据集看起来令人生畏，但许多化学数据库在ML时代之前就已经可用了。表1列出了化学领域选定的流行数据库，其中许多数据库都有悠久的数据收集和整合历史。这些数据的来源包括开放专利和研究文章、针对特定属性的高通量实验，以及通常基于密度泛函理论（DFT）的QM计算。

2.1. 化学反应数据库

化学反应数据库对实验人员在合成路线的设计方面具有很高的价值，尤其是在有机化学中。在互联网出现之前，文献中的反应已经被化学文摘社（CAS）索引。这些数据现在可以从SciFinder上访问，其中包括来自期刊、专利、书籍和其他来源的化学和文献信息。然而，SciFinder和类似的商业数据库Reaxys无法批量导出大量的化合物和化学反应数据，这限制了深度ML所需的训练数据集的大小。出于这个原因，研究人员使用文本处理技术从美国专利商标局（USPTO）专利中提取反应信息[11]，这些专利是开源的，可以从互联网上下载。最近，开放反应数据库（Open Reaction Database, ORD）[12]建立了化学反应存储的数据格式模板，支持公共化学反应数据集的数据共享。应该提到的是，越来越多的计算机辅助合成领域的研究人员现在公开了他们的数据库——例如，使用NextMove软件[13]，该软件提供了识别化学品的开源文本挖掘工具——并共享他们的数据集，用于下载和在线查询。

2.2. 化学性质数据库

由于化学性质的种类繁多，因此存在许多不同类型的化学性质数据库。PubChem [14]是一个开放的化学数据库，主要关注物质的化学和物理性质、生物活性以及毒性。自1996年以来，美国国家标准与技术研究所（NIST）发布了Chemistry WebBook [15]，收集最初在手册和表格中发布的光谱和热力学数据；它还包括其他物理和化学的基础数据，如电离能量、溶解度、光谱、色谱和计算数据。这些数据集可以在该网站上批量下载。类似地，ChemSpider [16]是一个编译公开的网络数据库，提供分子的结构和属性。除了通用数据库外，还有一些关注特定属性的数据集，如ChemBL [17]和DrugBank [18]专注于药物的生物活性，Tox21数据集[19] 包含了化合物的毒性效应（包括通过高通量毒性分析获得的12 707种代表性化合物和12种不同的毒性效应），ESOL [20]提供有关小分子溶解度的实验数据（包括有机小分子的水溶性数据），FreeSolv [21]则提供小分子在水中的溶解度及计算得到的水合自由能的实验数据，以及Lipophilicity [22]提供的有机小分子的辛醇-水分配系数的实验数据。

2.3. 材料数据库

对于固体材料，剑桥结构数据库（CSD）[23]是最受认可的数据库之一；它从文献中收集有机晶体结构信息，包括X射线或中子衍射数据、结晶条件和构象测定的实验记录。无机晶体结构数据库（ICSD）[24]包含超过272 000个晶体结构，以及分子式、原子坐标、晶胞参数、空间群等信息，这些信息大部分通过实验确定。粉末衍射文件（PDF）[25]数据库提供了1 143 236种材料（2023版）的衍射和晶体学数据。PDF最初是单相X射线粉末衍射图

表1 ML中常用的流行化学数据库列表

Classification	Name	Content	URL
Chemical reaction databases	SciFinder	Information on chemical compounds, bibliographic data, and chemical reactions (commercial database)	https://scifinder.cas.org/
	Reaxys	Chemical reaction and bibliographic information (commercial database)	https://www.reaxys.com/
	USPTO	Chemical structure and reaction	https://www.repository.cam.ac.uk/handle/1810/244727
	ORD	Organic chemical reaction data	https://github.com/open-reaction-database
	NextMove	Chemical reaction data	https://www.nextmovesoftware.com/about.html
Chemical property databases	PubChem	Chemical and physical properties, biological activities, and toxicity of substances	https://pubchem.ncbi.nlm.nih.gov/
	NIST	Standard physicochemical properties of compounds	https://webbook.nist.gov/chemistry/
	ChemSpider	Structure and property of compounds	http://www.chemspider.com
	ChemBL	Drug-like properties of bioactive molecules	https://www.ebi.ac.uk/chembl/
	DrugBank	Properties of drug molecules	https://go.drugbank.com/releases/latest
	Tox21	Toxic effects of substances	https://ntp.niehs.nih.gov/whatwestudy/tox21/index.html
	ESOL	Water solubility of compounds	https://pubs.acs.org/doi/10.1021/ci034243x
	FreeSolv	Water solubility of small neutral molecules	https://github.com/MobleyLab/FreeSolv
	Lipophilicity	Lipid solubility of organic compounds	https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/cem.2718
Material databases	CSD	Organic and metal–organic crystal structures	https://www.ccdc.cam.ac.uk/
	ICSD	Inorganic and metal–organic crystal structures	https://icsd.products.fiz-karlsruhe.de/
	PDF	Diffraction data of inorganic and organic compounds	https://www.icdd.com/pdfsearch/
	MatWeb	The thermoplastic and thermoset of polymers, metals, and other engineering materials	https://matweb.com/
	Li-ion Battery Aging Datasets	Charge and discharge curves of lithium batteries	https://data.nasa.gov/dataset/Li-ion-Battery-Aging-Datasets/uj5r-zjdb
	HTEM	Experimental information of inorganic thin-film materials	https://htem.nrel.gov/
Computational chemistry database	GDB-17	Structures of organic molecules up to 17 atoms	https://www.gdb.unibe.ch/downloads/
	QM9	Quantum chemical properties of organic molecules	http://quantum-machine.org/datasets/
	ANI-1	Energy and force of non-equilibrium molecules	https://github.com/isayev/ANI1_dataset
	Materials Project	DFT relaxed material structures and their thermal, electronic, and elastic properties	https://materialsproject.org/
	OQMD	DFT relaxed material structures and their thermal, electronic, and elastic properties	https://oqmd.org/
	Aflowlib	DFT relaxed material structures and their thermal, electronic, and elastic properties	http://aflowlib.org/
	MD17/ISO-17	Energy and force of non-equilibrium molecules	http://quantum-machine.org/datasets/
	LASP	Global PES dataset of molecules/materials	http://www.lasphub.com
	OC20	Adsorption energy of molecules in catalysts	https://opencatalystproject.org/
	Atom3D	3D structure of molecules, RNA, and proteins	https://www.atom3d.ai/

URL: uniform resource locator; USPTO: United States Patent and Trademark Office; ORD: Open Reaction Database; NIST: National Institute of Standards and Technology; CSD: Cambridge Structural Database; ICSD: Inorganic Crystal Structure Database; PDF: Powder Diffraction File; HTEM: High-Throughput Experimental Materials; OQMD: Open Quantum Materials Database; OC20: Open Catalyst 2020; DFT: density functional theory; PES: potential energy surface.

样的集合；然而，近年来，它也部分包括来自CSD、ICSD、NIST等的原子坐标项。MatWeb数据库涵盖了广泛的工程材料，如热塑性和热固性聚合物、金属材料和陶瓷材料，记录了物理性能（如吸水率、比重）、力学性能（如弹性模量）、热力学性能（如熔点）和电学性能（如偶极矩、电阻）。其他更具体的数据库包括来自美国国家航

空航天局（NASA）艾姆斯预测中心的锂（Li）离子电池材料的锂离子电池老化数据集[26]和用于无机薄膜材料的高通量实验材料（HTEM）数据集[27]。前者收集电池材料的充电、放电和电化学阻抗谱等操作曲线，而后者包括有关薄膜材料的合成条件、化学成分、晶体结构和特性等信息。

2.4. 计算化学数据库

为了便于第一性原理的计算，计算化学数据库正成为当今化学数据的主要来源。计算数据的明显优势包括其较高的精度、自洽性和良好的重现性（即使是对于在实验中难以合成的化合物）。GDB-17数据库[28]在文献中经常被用于ML应用，因为它包含1664亿个有机分子，其中，碳（C）、氮（N）、氧（O）、硫（S）和卤素原子多达17个。这些分子通过应变拓扑结构和稳定性标准进行列举和过滤，并使用简化分子输入线性输入系统（SMILES）[29]名称进行索引，以通过分子组成和连接进行区分。QM9数据集[30]是量子化学性质的基准数据集；它由来自GDB-17数据库的平衡有机化合物组成，其含有多达9个来自C、N、O和氟（F）的“重”原子[30]。它还提供了类似的谐波频率、偶极矩、极化率、能量、焓和自由能，以及能量最小值，这些都是在DFT B3LYP/6-31G（2Df, p）水平上计算出来的。与小分子数据库并行的，还有许多材料数据集，包括材料项目（Materials Project）[31]、开放量子材料数据库（OQMD）[32]和Aflowlib数据库[33–34]，他们提供了基于网络的对DFT优化[主要是Perdew-Burke-Ernzerhof (PBE)函数]结构和数以百万计的已知或预测材料的计算属性的开放访问。这些项目通常伴随着Python包，如用于材料项目的pymatgen [35]、用于OQMD的QMpy [32]和用于Aflowlib的AFLOW [33]，它们提供了高吞吐量DFT计算框架来扩展数据集，以及用于分析数据的后处理工具。

为了扩大化学空间，人们已经做出了大量的努力来创建非平衡态结构数据集，例如，通过使用分子动力学（MD）模拟。ANI-1数据集[36]就是一个例子，其中包含2000万个非平衡分子。该数据集是由57 000种不同的分子构型组成，包括化学元素C、氢（H）、N和O。MD17[37]和ISO-17数据集[38]是量子化学性质基准的其他示例；它们包含非平衡分子，这些分子是通过对不同构象分子进行有限温度MD模拟得到的。此外，LASP软件[39]为通过随机表面行走（SSW）全局势能面（PES）探测获得的分子和材料提供了PES数据集，并包含反应构型和高能结构。这些数据集已被用来构建ML势函数（见下文）。除了一般的分子数据集，还有特定应用数据集，如Open Catalyst 2020 (OC20)数据集[40]，其中包含各种各样的表面上饱和或不饱和分子片段的872 000个吸附状态，以及Atom3D数据库[41]，其具有生物分子的三维结构，包括分子、RNA和蛋白质。

3. 特点

数据和特征决定了ML模型的上限。从源数据中进行预处理得到的特征，通常也被称为表示或描述符，它们是ML模型的输入。重要特征的选择（称为特征工程）过去是ML模型的训练中最耗时和劳动密集型的工作。虽然深度学习技术可以允许ML模型学习如何提取特征，但它们通常需要相对较大的训练数据集和模型参数空间；因此，它们有更高的计算成本，且最终创建的ML模型可解释性较差。在化学中，不同ML模型的输入特征可能是不同的[42–44]，但分子/晶体结构的表示是特征工程的一般任务。由于关于这个主题的优秀评论文章[45–46]已经发表，因此我们只简要介绍了一些与第4节和第5节中提到的应用相关的文章。

分子描述符基本上有两类，即二维特征和三维特征。二维特征关注分子中的键合模式，而忽略了空间构象。这些特征是由分子图（以原子为节点、键为边）或邻接矩阵（即键矩阵）推导出来的。例如，SMILES使用人类可读的字符串（例如，“CCO”代表乙醇）描述饱和分子，国际纯粹与应用化学联合会（IUPAC）国际化学标识符（InChI）[47]使用严格唯一但人类可读性较差的字符串来表示化合物。除了字符串之外，分子的拓扑结构也可以被抽象为浮点数的向量。利用Morgan算法开发的扩展连接性指纹（ECFP）[48]，迭代地搜索分子中的子结构，并将其编码为哈希值。

三维特征是由原子坐标编码的，由于缺乏排列、平移和旋转的不变性，原子坐标很难直接作为ML模型的输入[49]。研究者们设计了一些精简的方法来保持排列、平移和旋转的不变性，并灵敏地区分不同的三维结构。这些方法通常基于原子间距离和原子间角度导出的数值函数，如最小埋藏体积百分比[50]、原子中心对称函数（ACSF）[51]、Steinhardt型有序参数[52]和功率型结构描述符（PTSD）[53–54]。其他方法是基于原子密度相似的函数，包括但不限于平均空间占用率（ASO）[55]、原子位置平滑重叠（SOAP）[56]和基于高斯型轨道的密度向量[57]。

4. ML模型

在特征将数据编码为机器可读的输入之后，ML模型将输入转换为输出，即所预测的属性。ML模型并不从理论中推导出物理定律，而是在容易获取的变量与关心的属性之间建立数值联系，而这些属性往往过于复杂，无法用理论直接求解。一般来说，ML算法根据对数据集的学习方式可以分为三个主要类别：用于拟合标记数据的监督学习、用于对未标记数据进行分类的无监督学习以及利用奖励机制来指导数据学习的强化学习。其中，监督学习是科学研究中应用最广泛的方法，因为它对特定目标具有更好的数值预测能力。尽管在ML中有许多方法和类别，但在实践中实现ML并不难，这要归功于许多公开可用的软件包，如scikit-learn [58]、PyTorch [59]和TensorFlow [60]。下面，我们将介绍在监督学习中常用的算法，特别是那些过去十年开发的涉及（深度）神经网络（NN）的算法。读者应该参考更深入的ML书籍以了解数学细节。

4.1. 决策树

决策树可被可视化为一系列相关选择的可能结果的映射，如图1（a）所示，末尾节点代表结果[图1（a）中的A、B、C类]，分支中的节点代表选择（属性；如图1（a）中的x [2]）。为了训练决策树，数据集按选定的属性递归分割，最大限度地对子组进行分类，以获得相同的结果[61]。该算法具有可解释性、超参数少、计算成本低、适用于相对较小的数据集（如200个样本）等优点，被广泛用于分类和预测。然而，随着数据的微小变化，预测可能会有显著的变化。

为了增强模型的鲁棒性，研究者们开发了随机森林（RF）[62]，它可以独立地训练多棵决策树，并收集所有的结果，通过投票或平均来做出最终的预测。每棵树都是在不同子数据集上进行训练的，各个子数据集从源数据中随机采样得到。这也被称为引导聚集方法（bootstrap aggregating or bagging）。通过决策树的集合，RF模型的鲁棒性得到增强，从而获得了更好的预测能力。这些模型更适合于预测离散目标值；因此，典型的应用是通过将合成条件与所需产物[64–65]的选择性相关联来优化实验变量[63]。

4.2. 前馈神经网络

前馈神经网络（FFNN），也被称为多层感知机（MLP）[66]，由多个全连接的神经元层（即节点）组成，它们同时执行线性和非线性操作。如图1（b）所示，从输入$\boldsymbol{x}$到输出$\boldsymbol{y}$，每个全连接层执行线性运算，如等式（1）所示，其中，权重$\boldsymbol{W}_{m\times n}$和偏差$\boldsymbol{b}_{m\times 1}$分别为可训练参数，$m$和$n$分别为输出和输入的维度。

$$\boldsymbol{y}_{m\times 1}=\boldsymbol{W}_{m\times n}\boldsymbol{x}_{n\times 1}+\boldsymbol{b}_{m\times 1} \quad (1)$$

随后可以对每个节点上接收到的数据执行非线性转换，即激活。有许多可使用的激活函数，如双曲切线、sigmoid函数和修正线性单位（ReLU）函数。FFNN的训练是通过最小化预测值和真实值之间的误差（称为代价函数）来实现的，如等式（2）所示。

$$\boldsymbol{W}^*,\boldsymbol{b}^*=\underset{\boldsymbol{W},\boldsymbol{b}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i-\mathrm{FFNN}\left(\boldsymbol{W},\boldsymbol{b},\boldsymbol{x}_i\right)\right\|^2 \quad (2)$$

式中，$\boldsymbol{y}_i$和$\boldsymbol{x}_i$为训练集中第i个样本的标签和特征。可以利用多种基于梯度的优化方法，如随机梯度下降[67]、Adam优化[68]和L-BFGS [69]，来寻找FFNN中的最优参数。随着中间层（隐藏层）数量的增加，拟合参数越来越多，因此模型原则上具有更高的拟合能力[1]。在FFNN中，由于梯度消失问题，隐藏层的数量通常最多为三个，这表现为训练中的改进速度缓慢。不过，借助残差连接[70]（即跳跃连接），这个问题可以得到缓解，尽管一个大型网络的拟合需要大量计算。

4.3. 卷积神经网络

卷积神经网络（CNN）是在FFNN的基础上开发的一种深度学习方法，它将多个卷积层和池化层添加到FFNN中，如图1（c）所示。CNN最初被用于图像识别，取得了巨大的成功，因此在学习网格数据方面特别强大[2]。以单通道（灰度）图像为例[图1（c）]，一个卷积层聚焦于图像内部的一个预定义大小的小窗口（如3×3像素）。通过在权重矩阵（称为滤波器）与小窗口输入数据（3×3矩阵）之间进行卷积（实际上是互相关），并在整个图像上滑动小窗口，从局部窗口中提取的图像特征被平铺成一个二维地图。在实践中，往往在CNN中应用多个滤波器来捕获不同的特征并生成多个二维地图。在卷积层之后，池化层进一步以预先定义的模式扫描2D地图（如3×3窗口），并计算该区域的平均值或最大值，用于聚合和粗化特征。在CNN中，拟合参数不仅包括FFNN中使用的参数，还包括卷积层中滤波器的权重。

CNN可以用于处理二维数据的化学问题，如用红外摄像机进行气体泄漏检测[71]；它也是AlphaFold1 [4]中的基本单元。在实践中，一维（1D）数据，如来自化学传感器的信号，也可以作为输入，并使用一维CNN应用于化学工程的故障检测和诊断[72–75]。

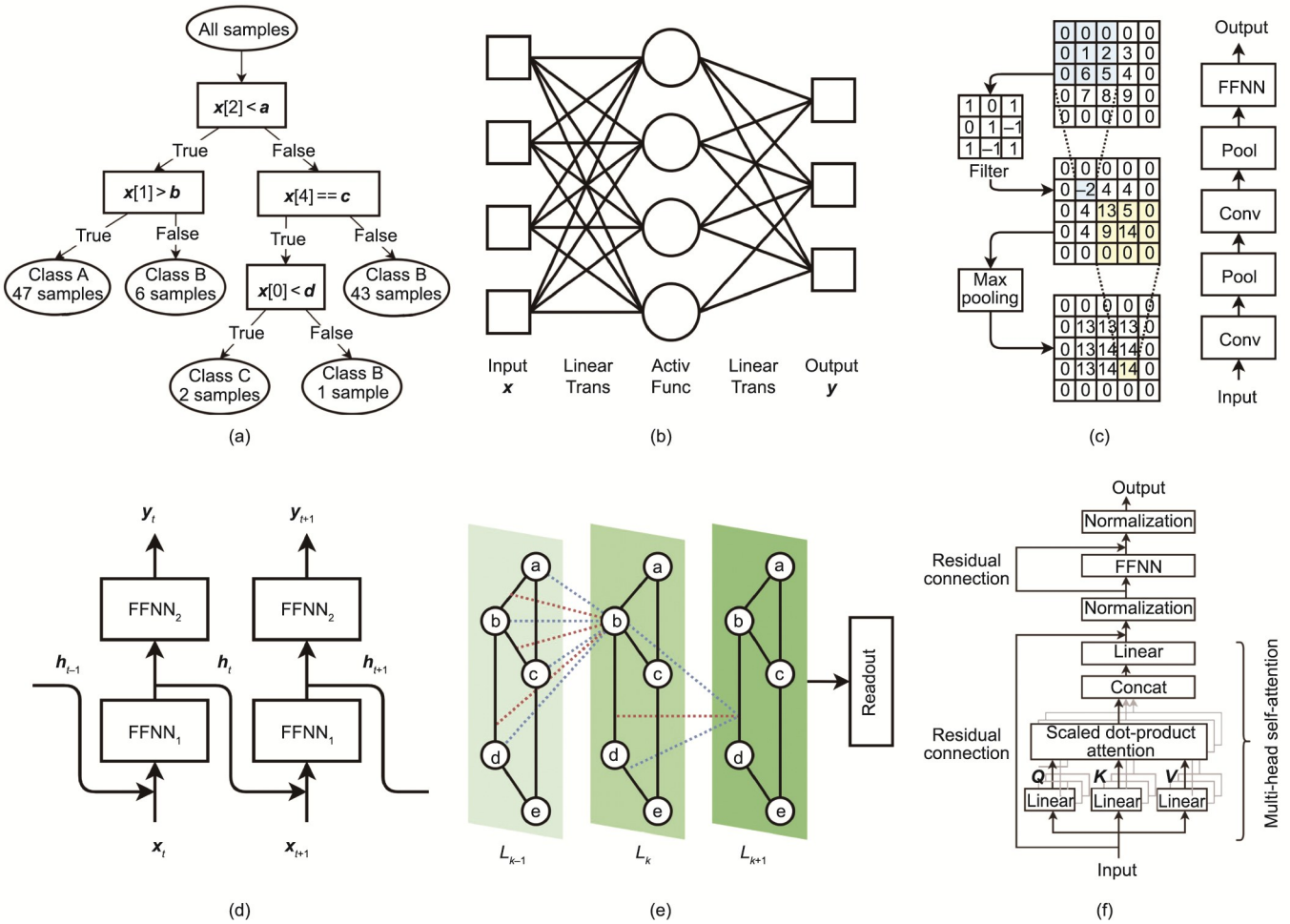


图1. 六种流行的机器学习模型。（a）决策树；（b）前馈神经网络（Trans：转换；Activ Func：激活函数）；（c）卷积神经网络（Conv：卷积；Pool：池化）；（d）循环神经网络；（e）图神经网络；（f）Transformer神经网络。

4.4. 循环神经网络

循环神经网络（RNN）是另一类人工神经网络，它允许某些节点的输出作为附加的输入重新输入到相同的节点，如图1（d）所示。这使得RNN适用于处理序列数据[76]，如语音识别[3]。对于t时刻的序列数据，$\boldsymbol{x}_t$和$\boldsymbol{y}_t$分别为输入和输出。从$\boldsymbol{x}_t$到$\boldsymbol{y}_t$，一个简单的RNN模型可以表示如下：

$$\boldsymbol{h}_t = \phi\left(\boldsymbol{W}_{h\times h}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{h\times n_x}\boldsymbol{x}_t + \boldsymbol{b}_{h\times 1}\right) \tag{3}$$

$$\boldsymbol{y}_t = \phi\left(\boldsymbol{W}_{n_y\times h}\boldsymbol{h}_t + \boldsymbol{b}_{h\times 1}\right) \tag{4}$$

式中，$\boldsymbol{h}_t$为t时刻的隐藏变量；$\boldsymbol{W}_{h\times h}$、$\boldsymbol{W}_{h\times n_x}$、$\boldsymbol{W}_{h\times n_y}$为可训练的权重矩阵；$h$、$n_x$、$n_y$分别为隐藏变量、输入和输出的维数。显然，$\boldsymbol{W}_{h\times h}\boldsymbol{h}_{t-1}$是前一个时间$t-1$时刻的附加项，它会影响$t$时刻的输出。如果没有附加项，RNN会退化为标准FFNN。RNN特别适合于学习类似序列的数据，比如一串化学名称。通过使用反应物的SMILES名称作为输入，RNN已经被用来预测有机反应的产物[77]（见第5.1节）。

4.5. 图神经网络

图神经网络（GNN）是一类深度学习方法，它可以通过图中节点之间的成对消息传递来处理图数据；它通常也被称为消息传递神经网络（MPNN）[78–79]。一个GNN通常会堆叠多个消息传递层，如图1（e）所示；因此图中的一个节点可以与几个相邻节点之外的其他节点进行通信。在每个MPNN层L_k中，节点N_k^{b}（即第k层中的节点b）根据来自前一层L_{k-1}的信息进行更新，包括节点本身（N_{k-1}^{b}）、其第一个相邻节点（N_{k-1}^{a}、N_{k-1}^{c}和N_{k-1}^{d}）和它连接到的边（E_{k-1}^{ab}、E_{k-1}^{bc}和E_{k-1}^{bd}）。边缘表示法也可以用类似的方法进行更新。MPNN中的更新策略可以非常自由地设计，例如，使用邻域表示的和，然后进行非线性激活。在消息传递层之后，利用读出函数（如FFNN）来获得基于最后一个消息传递层的输出。

化学家对GNN特别感兴趣，因为可以自然地用图表示分子。GNN作为一类前沿但略欠成熟的方法，已成功应用于预测分子[78]和晶体[80]的性质。也有一些尝试用

GNN拟合材料的PES [38,81]（详见第5.2节）。

4.6. Transformer

Transformer是一种新的深度学习模型，最初被设计用于处理序列数据（如自然语言处理）[82]，并显示出了取代RNN模型的巨大潜力。Transformer的关键特点是多头自注意力机制，它允许一次性处理整个输入序列。如图1（f）所示，一个Transformer层可以表示为等式（5）。

$$\mathrm{Atten}\left(\boldsymbol{Q}_{d_k\times d_m},\boldsymbol{K}_{d_k\times d_m},\boldsymbol{V}_{d_k\times d_m}\right)=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (5)$$

$$\mathrm{softmax}\left(\boldsymbol{z}\right)_i=\frac{\exp\left(z_i\right)}{\sum_{i=1}^{K}\exp\left(z_i\right)}\ \text{for}\ i=1,\cdots,K \quad (6)$$

该方程计算查询词向量$\boldsymbol{Q}$和关键词向量$\boldsymbol{K}$的内积，并将其发送到等式（6）中定义的softmax函数，获得一组对值向量$\boldsymbol{V}$的权重。这里d_k和d_m分别是关键词向量和模型的维数。这三个矩阵$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$是由线性变换的同一输入生成的（因此，这种方法称为自注意），其中，线性变换权值$\boldsymbol{W}_{\boldsymbol{Q}}$、$\boldsymbol{W}_{\boldsymbol{K}}$和$\boldsymbol{W}_{\boldsymbol{V}}$是学习参数。通过使用具有不同权值的并行的多个注意力单元集合，即所谓的多头注意力，该模型可以同时关注不同位置的特征信息。多头自注意力层的输出可利用FFNN进行进一步的处理。由于Transformer模型可以很深，有很多层，因此利用残差连接[70]来避免梯度消失；这直接将某一层（如FFNN）的输入与其输出相加，并将其和作为下一层的输入。凭借多头注意力带来的强大的特征提取能力，Transformer模型已被证明对序列文本数据[83–84]和网格图像数据[85]都是成功的，从而统一了ML的两个重要应用领域。

得益于其强大的ML框架，Transformer近年来已经有了许多重要应用。例如，AlphaFold2利用Transformer的一种变体，即所谓的Evoformer [5]，来取代AlphaFold1 [4]中残差连接的CNN。Graphormer [86]是一种改进的用于图数据的Transformer，在Open Catalyst Challenge 2021中，它在从非弛豫结构预测弛豫能量方面显示出较高的准确性，优于经典的MPNN。Schwaller等[87]使用一个Transformer来学习有机反应的产物与反应物之间的原子映射关系，而不需要监督或人工标记，从而确定反应规则。

5. 应用

在接下来的章节中，我们将提供ML的一些重要应用来说明这些ML技术如何被用于解决化学问题，包括有机化学中的逆合成、计算化学中的ML势函数，以及物理化学中的多相催化。表2总结了一些相关文献[38,56–57,63,88–106]，列出了ML任务、输入数据、特征、ML模型和预测目标的信息。

5.1. 逆合成

合成规划，也被称为逆合成，是化学的核心，它回答了如何从现有材料中合成所需的化合物的问题。长期以来，这项任务在很大程度上依赖于经验丰富的化学家的知识；因此，计算机辅助合成规划（CASP）——早在20世纪60年代由Corey等[107–108]提出——一直是化学领域的热门话题。从那时起，许多成功的CASP程序被开发出来，如LHASA [109]、SECS [110]、Chematica [111]、IBM RXN [112]、3N蒙特卡罗树搜索（MCTS）[88]和AI-ynthFinder [113]（表2）。由于有机反应丰富，且这类数据库也相对容易访问，多年来，特别是在过去十年的ML技术的帮助下，逆合成得到了积极发展[88,111–117]。

反应预测和逆合成是CASP中的两个关键模块。反应预测是逆合成的基础，关注于单步反应，目的是在一定的反应条件下建立反应物与生成物之间的一一对应关系。预测必须选择正确的反应规则（即模板），这取决于分子结构和反应条件。因此，反应预测可以分为两类：基于模板的方法和无模板的方法[89–92,118]。前者需要一个先验的模板库，它既可以由专家使用化学信息学进行编码[108–109]，也可以通过最近流行的原子映射算法从反应数据库中提取[93]。无模板的方法通常侧重于预测分子中的反应中心，从而确定最适合连接（断开）的键。

在基于模板的方法中，通常会从一种反应物产生太多可能的产物，从而产生过多的候选反应而导致过载。2016年，Wei等[94]尝试使用ML来预测模板的适用性。利用以指纹为基础的神经网络算法，他们在仅给出反应物和试剂作为输入的情况下，预测了卤代烷和烯烃的16个基本反应中最有可能的反应类型。最后的反应是通过对反应物进行SMARTS转换而产生的。他们的模型在测试反应中达到了85%的准确率，在选定的教科书问题中达到了80%的准确率。后来，Segler和Waller [93]将该方法应用于来自Reaxys的一个更复杂的实验数据集。如图2（a）所示[93]，每个反应物指纹在8720个算法提取的模板库中产生了一个概率分布，准确率达到78%。需要注意的是，基于模板的方法在CASP中相对成熟，主要关注的问题包括预测的相关性和模板库的范围。在ML模型的训练中，通常须排除稀有模板。

近年来出现的无模板方法，有可能打破基于模板的方

表2 ML在逆合成、ML势函数和多相催化中的应用总结

Application	Task	Input data	Feature	Model	Prediction target	Refs.
Retrosynthesis	Template-based reaction prediction	Reactant molecule	ECFP	FFNN	The most probable reaction type	[93–94]
	Template-free reaction prediction	Product molecule, reaction type	SMILES	RNN	SMILES of reactant	[89]
	Template-free reaction prediction	Reactant molecule	SMILES	RNN	SMILES of product	[90]
	Template-free reaction prediction	Reactant molecule	SMILES	Transformer	SMILES of product	[91]
	Template-free reaction prediction	Reactant molecule	Molecule graph	GNN	Reaction center and product	[92]
	Retrosynthesis	Product molecule	ECFP	FFNN	SCScore	[95]
	Retrosynthesis	Product molecule	ECFP	MCTS	Retrosynthetic route	[88]
ML potential	ML potential	Atomic coordinates	SOAP	Gaussian process regression	DFT energy	[56]
	ML potential	Atomic coordinates	ACSF/PTSD	FFNN	DFT energy	[99]
	ML potential	Atomic coordinates	Interatomic distance	CNN	DFT energy	[96–97]
	ML potential	Atomic coordinates	Interatomic distance	GNN	DFT energy	[38,98]
	ML potential	Atomic coordinates	Gaussian-type-orbital based atomic density vector	FFNN	DFT energy	[57]
	ML potential	Atomic coordinates	ACSF	FFNN	DFT energy by atomic charge	[100]
Heterogeneous catalysis	Optimizing catalysts	Experimental data	Experiment condition	FFNN, RF	Product yield, selectivity	[101–102]
	Optimizing catalysts	Literature experimental data	Experiment condition, the characteristic results	RF	Product yield, selectivity	[63]
	Optimizing catalysts	Robot-produced experimental data	Experiment condition	Bayesian	Catalyst activity	[103]
	Predicting reactivity	Atomic coordination environments	Coordination number, element type	RF	Adsorption energy	[104]
	Predicting reactivity	Element information	Elementally, the atomic radius, number of valence electrons	RF	d–p band center	[105]
	Research reaction mechanism	Atomic coordinates	PTSD	FFNN	DFT energy	[106]

MCTS: Monte-Carlo tree search; SCScore: synthetic complexity score.

法在质量和完整性方面的局限性。基于RNN的seq2seq模型是最具代表性的无模板ML模型[89–91,118]。在seq2seq模型中，反应预测被认为是反应物和产物之间的SMILES字符串[29]的机器翻译问题，而前体或产物输出的SMILES代码通过图形转换模块来生成真实的化学结构，如图2（b）[89]所示。值得一提的是，seq2seq模型只输出SMILES序列，因此由于存在对SMILES表示语法的误解，模型输出的SMILES序列有时无法转换为合理的化学结构式。2017年，Liu等[89]在美国USPTO数据集的50 000个实验反应示例上训练了seq2seq模型，在测试数据集上获得了37.4%的top-1准确率和70.7%的top-50准确率。最近，Schwaller等[91]用一个Transformer取代了seq2seq模型中的RNN，并在一个公共基准数据集上实现了90.4%的top-1精度（93.7%的top-2精度）。类似地，GNN可以用于无模板预测[92,119]。Jin等[92]的一项研究使用了一种MPNN的Weisfeiler-Lehman网络（WLN），在USPTO-15K数据集上实现了76%的top-1精度，在USPTO数据集上实现了79%的top-1精度。

逆合成更为复杂，因为其目的是提供一个全局最优的合成途径，而不像连接最佳单步反应或选择最短路径那么简单。传统上，CASP项目（如LHASA和SECS）会建议一些候选的合成方案，最终的选择是由经验丰富的化学家做出的[107,109]。更进一步，Coley等[95]提出了合成复杂性评分（SCScore）作为逆合成中分子排序的度量标准。如图2（c）[95]所示，他们构建了一个FFNN模型，根据ECFP [48]计算SCScore，并对来自Reaxys数据库的超过

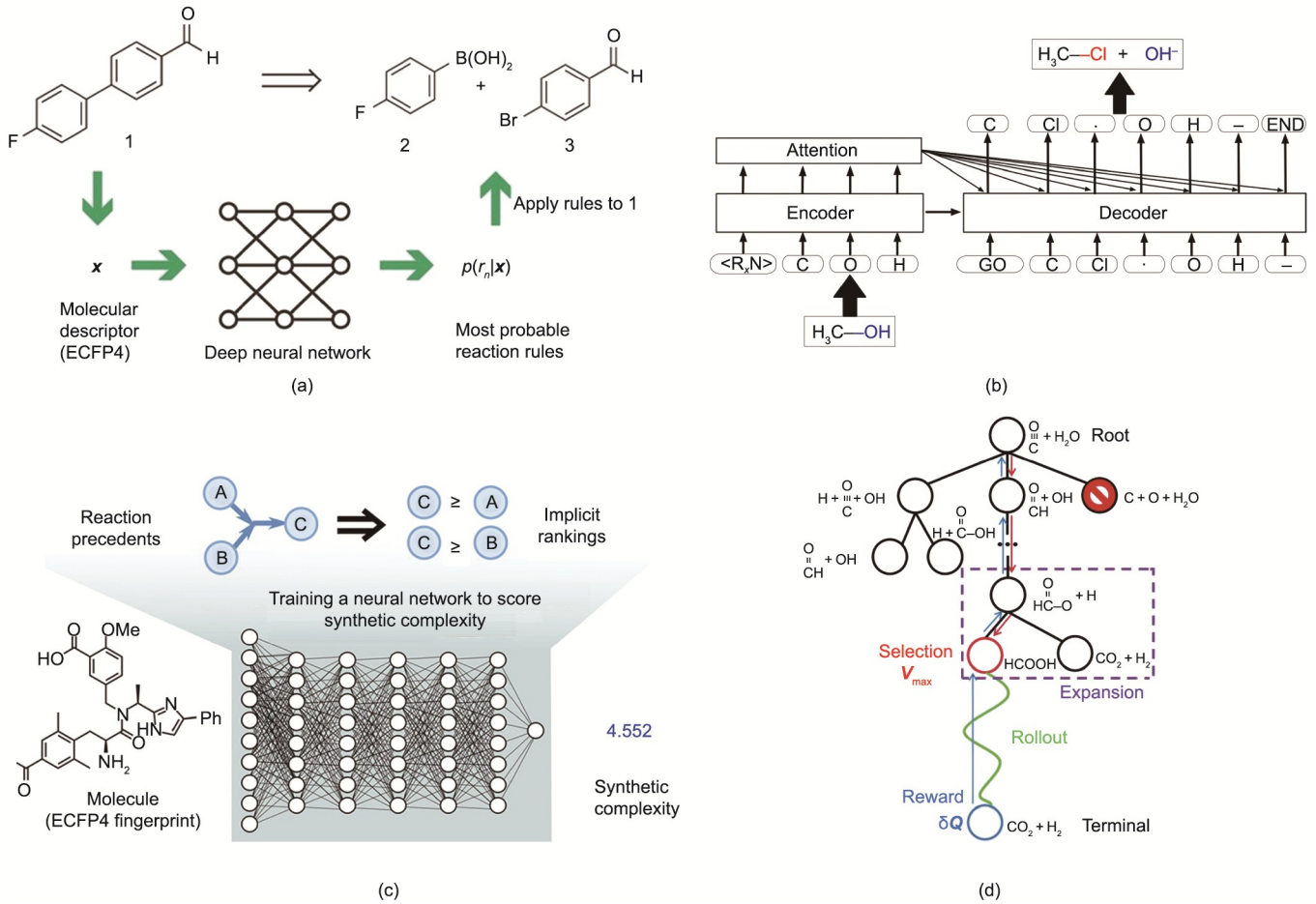


图2. （a）基于模板的反应预测的神经符号方法示意图，该方法通过反应物的ECFP4描述符来预测可能的反应规则[93]；（b）用于无模板反应预测的Seq2seq模型结构，该模型将反应物的SMILES名称转化为产物[89]；（c）指导逆合成的SCScore模型示意图[95]；（d）MCTS算法的说明图，算法由四个步骤组成：选择、扩展、仿真和奖励[120]。

1200万个反应进行了训练。平均而言，对于出版物中的化学反应，其产物的合成复杂程度应高于相应的反应物。基于这一前提，他们在训练中使用了一种铰链损失函数，以鼓励反应物和产物之间的SCScore分离。在该方案下，高价值合成路径的SCScore应呈单调增加。

Segler等[88]没有使用SCScore来评估合成路线，而是开发了一种基于MCTS的方法（图2（d）[120]）来生长有前景的不对称的子合成树，其中利用一种范围内滤波器网络来预测一个反应是否实际可行。滤波网络以产物和反应指纹为输入，并作为分类器，过滤出MCTS扩展阶段的荒谬的化学反应。通过结合其他两个预测反应模式的神经网络模型（即策略模型），研究人员认为，在对不同分子的9种路线的双盲A/B测试中，计算机生成的反应路线在平均水平上与文献报道的路线一样好（根据45位有机化学家的判断，MCTS的优选率为57%，文献的优选率为43%）。尽管取得了这些成功，但天然产物的合成仍然是一个挑战。除了对复杂分子的训练数据稀少外，在大多数模型中通常缺少对映异构体的定量产率，而这对于正确评估合成路线很重要。

5.2. ML势函数

ML在化学中的另一个重要应用与复杂系统的原子模拟有关，其中，ML势函数[121]可替代计算代价高昂的QM计算来评估PES。因为ML势函数是在QM计算的数据集上训练的，所以ML势函数计算可以达到与QM相当的精度，但速度要快几个数量级。因此，ML势函数方法显著地将第一性原子模拟的范围扩展到具有数千个原子的多元素体系，而传统上可能只能通过经验力场来模拟这些系统，尽管经验力场的可用性非常有限，仅限于具有相对简单PES的体系。自从1995年出现第一个神经网络势函数[122]以来，已经提出了许多不同类型的ML模型，其中两类ML结构（表2）——神经网络势函数[81,123–124]和基于核的势函数[125–127]——是最受欢迎的。虽然基于核的势函数，如高斯近似势（GAP）[128–129]及其使用原子位置平滑重叠核（SOAP-GAP）[56]的升级版，其超

参数比神经网络势函数少得多，但它们的计算速度受到训练集大小的限制。因此，在特别大的训练集使用基于核的势函数在本质上是困难的，而且它们更适合于单元素系统，如碳和硅[128–133]。下面，我们将重点关注神经网络势函数，它正在成为ML势函数计算的主流。

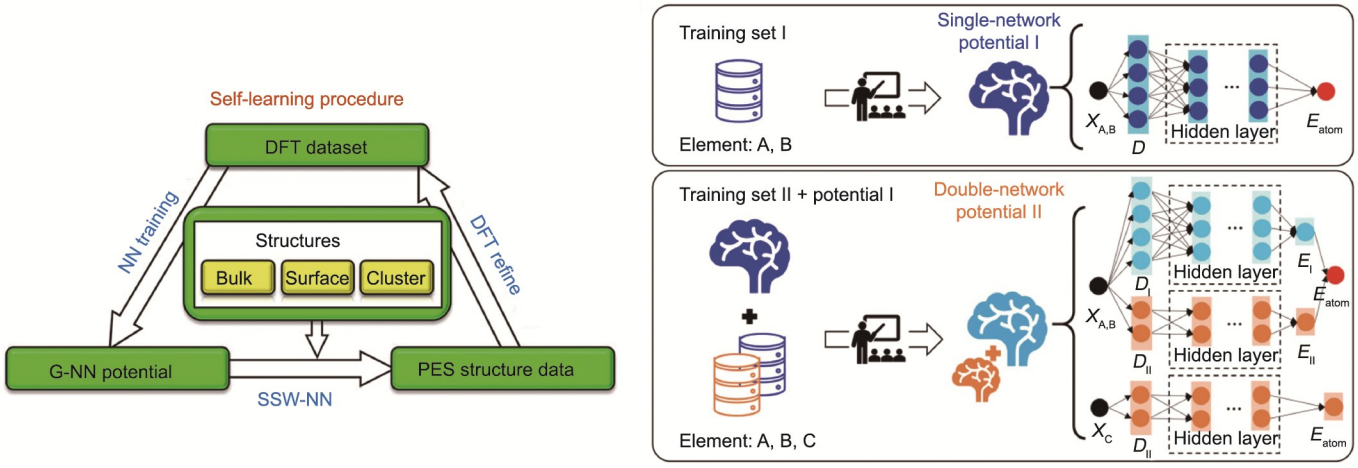
尽管在分子系统中有许多早期的应用，但复杂系统的神经网络势函数始于Behler和Parrinello [123]在2007年提出的高维神经网络（HDNN）框架。通过假设结构的总能量为单个原子的总能量，HDNN建立了一个FFNN，将原子的局部化学环境与原子能联系起来。Behler和Parrinello进一步发明了一组对原子平移、旋转和置换不变的AC-SF，作为神经网络输入层的结构描述符。HDNN框架的一个主要优点是它满足了总能量的扩展性，从而能够平等对待数据集中原子数和化学成分可变的不同结构构型。

此后，HDNN架构得到了积极的研究和改进，尤其是在结构描述符方面。例如，可以使用CNN架构提取局部原子环境，如在Deep Potential [96–97]中实现的那样，利用原子中心成对距离作为网格数据。类似地，GNN的MPNN [78]也可以用于从成对的原子距离中提取描述符，这已经在用于分子的DTNN [38]和用于周期性固体的SchNet [98]中实现。由Zhang等[57]提出的嵌入式原子神经网络势函数，利用一个高斯型的基于轨道的密度向量作为神经网络的输入，这已被证明具有与其他ML模型一样好的精度。

Liu课题组[39,134]提出的全局神经网络（G-NN）势函数（图3）实现了用于预测反应体系的自动数据生成程序，并改进了结构描述符和网络结构。G-NN势是根据从SSW全局PES搜索收集的全局PES数据集进行迭代训练的[135–136]。为了更好地拟合全局PES数据，他们开发了一套新的结构描述符，即PTSD [53–54]，可以更好地描述原子的局部化学环境。通过重用数据集和预先训练的神经网络势函数，实现了一种多网络架构，以快速生成多元素G-NN势函数。SSW-NN方法[图3（a）] [134]现在已在LASP软件中实现[39,99]，并已被应用于解决许多复杂的PES问题，如催化剂结构的确定和反应网络的预测[137–141]。

为了提供一个G-NN势的例子，我们参考了第一个Ti–O–H G-NN势，它被构造用于描述在H_2下处理的非晶TiO_2结构的PES [142]。G-NN势采用大量的PTSD，每个元素包含201个描述符，包括77个二体、108个三体和16个四体描述符，网络包含两个隐藏层（201-50-50-1网络），相当于总共约38 000个网络参数。对于包含140 000个结构的大型TiO_xH_y全局数据集，最终能量和力的均方根误差（RMSE）分别约为每个原子9.8 meV和0.22 eV·A^{-1}。利用这种Ti–O–H G-NN势，Ma等[142]解决了加氢过程中非晶TiO_2的形成机理，并发现了TiH氢化物介导的制氢途径。

上述ML模型中使用的局部化学环境描述符通常不足以捕获长程相互作用，如分子中的电荷转移。Ghasemi等[100]提出了一种可能的解决方案。他们使用了电荷平衡神经网络技术（CENT），使用相同的HDNN架构结构来学习显式的原子电荷，然后利用这些电荷来计算长程静电相互作用。Ko等[143]最近提出了第四代HDNN势（4G-HDNNP）方法，用于研究共轭长链有机分子及非中性金属和离子团簇[143]。该方法可以通过特殊的电荷平衡方案将非局域静电相互作用包括在内。



(a) (b)

图3. （a）G-NN势的SSW-NN自学习过程的方案。G-NN通过SSW采样、DFT细化和神经网络训练的循环进行迭代改进[134]。（b）在LASP中实现的双网络框架方案。通过对元素A和元素B进行训练的势I，它被重新用作势函数II中子网的起点，其数据集包含元素A、B和C [39]。*X*：每个原子的笛卡尔坐标；E_{atom}：每个原子的原子能；*D*：NN中使用的PTSD。

5.3. ML用于多相催化

由于催化剂结构的复杂性和催化剂在工业中的重要意义，多相催化一直是新技术的主要试验场。早期的ML应用可追溯到20世纪90年代[144–145]，通常是在现象学水平上，使用简单的ML模型学习实验数据，以此优化催化剂的合成和反应条件[101–102]。这些ML应用似乎受限于实验数据集的可用性，并且缺乏基本的理解，很可能忽略了实验中潜藏的重要变量，导致ML模型失败。随着深度学习和ML方法的出现，出现了许多令人兴奋的应用场景，如ML辅助文献分析[65, 146–148]和AI机器人[103]（表2）。

ML辅助文献分析利用自然语言处理模型的数据挖掘能力，从文献中提取实验数据。深入的数据分析将有助于揭示不同实验中的关键配方。例如，Suvarna等[63]从文献中收集了Cu基、Pd基、In_2O_3基和ZnO/ZrO_2基催化剂上CO_2加氢制甲醇相关的1425个实验数据点。如图4 [63]所示，他们建立了RF模型（R^2>0.85），将甲醇时空收率与实验操作条件相关的12个描述符相关联，从中确定了排名最高的因素（如空速、压力和金属含量）。随后进行实验验证，结果显示较小的RMSE（0.11 $g_{MeOH}\cdot h^{-1}\cdot g_{cat}^{-1}$）和高的$R^2$值（0.81），说明了ML模型的有效性。

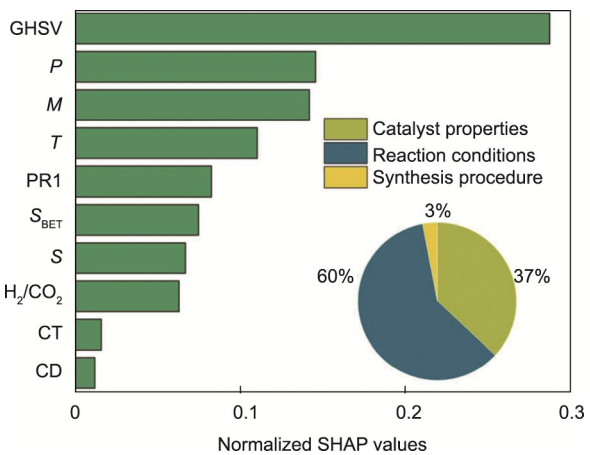


图4. CO_2加氢制甲醇的特征重要性分析。SHAP：Shapley加性解释；GHSV：气时空速；*P*：压力；*M*：金属含量；*T*：温度；PR1：促进剂1的含量；S_{BET}：催化剂表面积；*S*：载体含量；CT：煅烧温度；CD：煅烧时间[63]。

化学机器人被认为是化学的未来，因为它们能自动高效地进行实验，同时最大限度地保持实验之间的数据一致性[103,149–150]。例如，Burger等[103]开发了一种移动机器人，以搜寻改进分解水制氢的光催化剂。在8天的时间里，机器人在批量贝叶斯搜索算法的指导下（根据之前的实验优先选择有益的成分），在10个变量的实验空间中进行了688次实验。他们成功地开发了一种由新配方合成的催化剂，配方包括P10（5 mg）、氢氧化钠（6 mg）、L-半胱氨酸（200 mg）、$Na_2Si_2O_5$（7.5 mg）和水（5 mL），其活性是原配方催化剂的6倍。

从理论的角度来看，ML模型也可以用来学习低计算成本的物理量，如分子的吸附能和电子能带结构，这些性质对于催化很重要[151–152]。Tran和Ulissi [104]使用了一个基于RF的模型，基于一个包含31种不同元素的合金的数据库，将结构指纹与合金上的CO和H吸附能联系起来。最后，从54种体相合金中得到131个用于CO_2还原的候选表面，并从102个体相合金中得到258个用于析氢的候选表面。实验进一步证明了具有接近最佳CO结合能的CuAl合金是一种对CO_2还原选择性很好的催化剂[153]。Sun等[105]最近发现尖晶石氧化物的析氧反应（OER）活性本质上是由四面体和八面体位点之间的共价竞争决定的，这可以使用金属d带和氧p带之间的距离进行量化，表示为D_M。因此，他们开发了一个RF模型来预测D_M，并通过实验证实了预测的$[Mn]_T[Al_{0.5}Mn_{1.5}]_OO_4$混合氧化物具有较高的OER活性，在25 $\mu A\cdot cm_{ox}^{-2}$处有240 mV（vs RHE）的过电位。

另一方面，ML原子模拟可以提供关于催化剂结构和反应机理的原子级认识，这有利于催化剂的合理设计。例如，Shi等[106]提出了一种微观动力学引导的ML路径搜索方法（MMLPS），该方法可以在G-NN势下自动探索反应网络并确定低能反应路径。MMLPS的每个分支从不同的分子和表面覆盖度开始，独立地对反应PES的不同部分进行取样。通过合并所有分支的反应来建立反应对数据集，从中可以确定出最低势垒的反应路径。图5（a）[106]利用MMLPS采样得到的14 958个反应对，绘制了铜表面上CO和CO_2加氢的完整二维反应图。在所有表面上，CO_2通过甲酸根路径加氢（CO_2–$HCOO^*$–$HCOOH^*$–H_2COOH^*–$HCHO^*$–CH_3O^*–CH_3OH^*–CH_3），CO通过甲酰基路径加氢（CO–CO^*–CHO^*–$HCHO^*$–CH_3O^*–CH_3OH^*–CH_3OH），如图5（b）和（c）[106]中的自由能分布所示。CO_2加氢的总能垒在Cu(211)表面上仅为1.40 eV，而CO的能垒为1.45 eV，说明CO_2是甲醇产品中的主要碳源。进一步的微观动力学模拟表明，锌合金化对反应速率没有显著影响，甚至使反应失活。

6. 前瞻

本文综述了最近化学领域ML应用的关键要素，从流行的数据库到常见特征、现代ML模型和标准应用场景。随着近期ML应用的成功，我们必须认识到ML在化学中

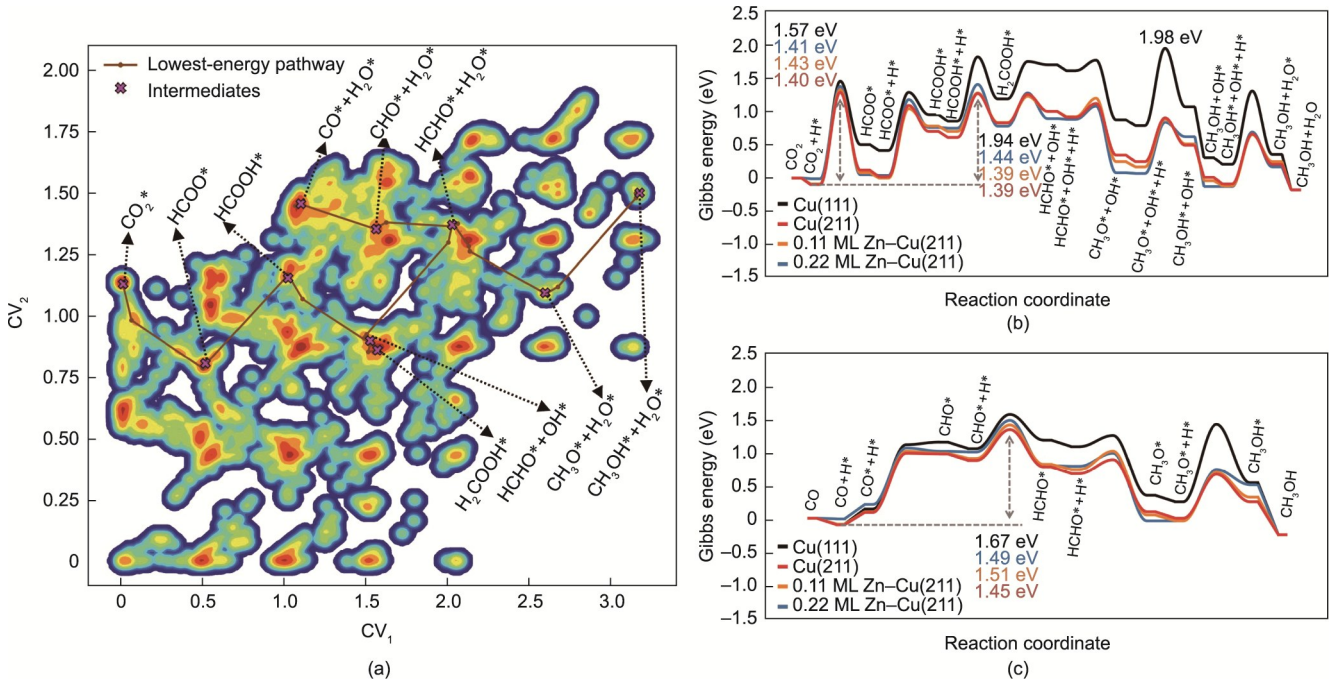


图5. （a）Cu(211)上MMLPS获得的14 958个反应对的等高线图。颜色表示该状态在反应对数据库中的出现频率。所有的结构都通过两个集体变量（CV_1和CV_2）被投影到图上。最低能量路径的关键中间体用棕色线突出显示。CO_2（b）以及CO加氢反应（c）在铜（211）和锌合金化的铜（211）表面上的吉布斯自由能轮廓。这里的“ML”代表单分子层结构[106]。

的使用面临许多挑战。例如，一个主要的障碍是缺乏高质量的数据，特别是涉及实验的数据。即使有了高通量的实验技术和实验机器人，在化学中仍有许多领域必须由人类来产生实验数据。此外，化学家往往不熟悉最先进的ML方法和相关的计算机科学技术，这导致了难以为目标应用设计合适的特征。如何针对不同的化学问题自动提取特征仍然是一个挑战。最后，大多数基于FFNN的ML研究的解释性很差，因此很难迁移到新的化学问题上。

随着计算设施的快速更新和新的ML算法的发展，可以期待更多令人兴奋的ML应用的到来，化学研究的未来肯定会在ML时代被重塑。虽然未来很难预测，特别是在这样一个快速发展的领域，但毫无疑问，ML模型的发展将带来更好的可访问性、通用性、准确性和智能性，从而获得更高的生产力。ML模型与互联网的结合是在全世界共享ML预测的一个好方法。Yoshikawa等[154]做出了有趣的贡献，他们在推特上建立了一个逆合成分析机器人，如果将目标分子的SMILES作为输入，该机器人可以自动回复逆合成结果。该机器人利用AIZynthFinder [113]包进行逆合成分析。

由于元素种类众多和材料的复杂性高，ML模型在化学中的可迁移性是一个常见的问题。一个预测通常局限于应用的数据库，而往往这只是广阔化学空间中的一个局部数据集。预测的精度在数据集之外迅速下降。这个问题可以通过新技术的出现来解决，这些新技术可以执行更有效的数据收集，如学习SSW全局PES数据的G-NN势函数，或者可以通过具有更多参数的ML模型拟合学习更复杂的体系。事实上，各种各样的ML竞赛都是由数据科学家举办的，比如Kaggle [98]，这催生了许多优秀的算法。在这方面，关于化学问题的公开ML竞赛仍然是有限的[40]，需要更多的努力来促进该领域的年轻人才的成长。

对于更智能的ML应用，端到端学习是一个很有前途的方向，因为它从原始输入而不是从手动设计的描述符生成最终输出。AlphaFold 2 [5]是一个典型的端到端学习框架，它将蛋白质的一维结构作为输入进行处理，最终输出蛋白质的三维结构。该框架为实验生物学家使用ML模型提供了很大的便利。类似地，在多相催化中，Kang等[120]最近展示了一种用于解析反应途径的端到端AI模型，展示了结合多个ML模型进行一次尝试预测来解决复杂问题的光明前景。这些先进的ML模型也应有助于构建更智能的实验机器人，以进行高通量实验[103,149–150]。

致谢

这项工作得到了国家重点研发计划（2018YFA0208600）、国家自然科学基金（12188101、22033003、91945301、91745201、92145302、22122301和92061112）、腾讯科学探

索奖基金和中央高校基本科研业务费（20720220011）的资助。

Compliance with ethics guidelines

Yun-Fei Shi, Zheng-Xin Yang, Sicong Ma, Pei-Lin Kang, Cheng Shang, P. Hu, and Zhi-Pan Liu declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[2] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84–90.
[3] Li X, Wu X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In: Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing; 2015 Apr 19–24; BrisbaneSouth, QLD, Australia. Piscataway: IEEE; 2015. p. 4520–4.
[4] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020; 577(7792):706–10.
[5] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021; 596(7873):583–9.
[6] Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens CV, Van Geem KM. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. Engineering 2021;7(9):1201–11.
[7] Venkatasubramanian V. The promise of artificial intelligence in chemical engineering: is it here, finally? AIChE J 2019;65(2):466–78.
[8] Zhou T, Song Z, Sundmacher K. Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. Engineering 2019;5(6):1017–26.
[9] Chen W, Iyer A, Bostanabad R. Data centric design: a new approach to design of microstructural material systems. Engineering 2022;10:89–98.
[10] Thebelt A, Wiebe J, Kronqvist J, Tsay C, Misener R. Maximizing information from chemical engineering data sets: applications to machine learning. Chem Eng Sci 2022;252:117469.
[11] Lowe DM. Extraction of chemical structures and reactions from the literature [dissertation]. Cambridge: University of Cambridge; 2012.
[12] Kearnes SM, Maser MR, Wleklinski M, Kast A, Doyle AG, Dreher SD, et al. The open reaction database. J Am Chem Soc 2021;143(45):18820–6.
[13] Akhondi SA, Klenner AG, Tyrchan C, Manchala AK, Boppana K, Lowe D, et al. Annotated chemical patent corpus: a gold standard for text mining. PLoS One 2014;9(9):e107477.
[14] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47(D1):D1102–9.
[15] Olver FW, Lozier DW, Boisvert RF, Clark CW, editors. NIST handbook of mathematical functions hardback and CD-ROM. Cambridge: Cambridge University Press; 2010.
[16] Ayers M. ChemSpider: the free chemical database. Ref Rev 2012;26(7):45–6.
[17] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40(D1):D1100–7.
[18] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 2008;36(D1):D901–6.
[19] Huang R, Xia M, Nguyen D-T, Zhao T, Sakamuru S, Zhao J, et al. Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs. Front Environ Sci 2016;3.
[20] Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci 2004;44(3):1000–5.
[21] Mobley DL, Guthrie JP. FreeSolv: a database of experimental and calculated hydration free energies, with input files. J Comput AIded Mol Des 2014;28(7): 711–20.
[22] Wang JB, Cao DS, Zhu MF, Yun YH, Xiao N, Liang YZ. *In silico* evaluation of logD7.4 and comparison with other prediction methods. J Chemometr 2015; 29(7):389–98.
[23] Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge Structural Database. Acta Cryst B 2016;72(Pt 2):171–9.
[24] Zagorac D, Müller H, Ruehl S, Zagorac J, Rehme S. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. J Appl Cryst 2019;52(Pt 5):918–25.
[25] Gates-Rector S, Blanton T. The Powder Diffraction File: a quality materials characterization database. Powder Diffr 2019;34(4):352–60.
[26] Lucu M, Martinez-Laserna E, Gandiaga I, Camblong H. A critical review on self-adaptive Li-ion battery ageing models. J Power Sources 2018;401:85–101.
[27] Zakutayev A, Wunder N, Schwarting M, Perkins JD, White R, Munch K, et al. An open experimental database for exploring inorganic materials. Sci Data 2018;5(1):180053.
[28] Ruddigkeit L, van Deursen R, Blum LC, Reymond JL. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model 2012;52(11):2864–75.
[29] SMILESWeininger D., a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 1988; 28(1):31–6.
[30] Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. Quantum chemistry structures and properties of 134 kilo molecules. Sci Data 2014;1(1):140022.
[31] JAIn A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. APL Mater 2013;1(1):011002.
[32] Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. npj Comput Mater 2015;1(1):15010.
[33] Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, Taylor RH, et al. AFLOW: an automatic framework for high-throughput materials discovery. Comput Mater Sci 2012;58:218–26.
[34] Calderon CE, Plata JJ, Toher C, Oses C, Levy O, Fornari M, et al. The AFLOW standard for high-throughput materials science calculations. Comput Mater Sci 2015;108:233–8.
[35] Ong SP, Richards WD, JAIn A, Hautier G, Kocher M, Cholia S, et al. Python Materials Genomics (pymatgen): a robust, open-source Python library for materials analysis. Comput Mater Sci 2013;68:314–9.
[36] Smith JS, Isayev O, Roitberg AE. ANI-1, a data set of 20 million calculated offequilibrium conformations for organic molecules. Sci Data 2017; 4(1): 170193.
[37] Bowman JM, Qu C, Conte R, Nandi A, Houston PL, Yu Qi. The MD17 datasets from the perspective of datasets for gas-phase “small” molecule potentials. J Chem Phys 2022;156(24):240901.
[38] Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A. Quantumchemical insights from deep tensor neural networks. Nat Commun 2017;8(1):13890.
[39] Kang P, Shang C, Liu Z. Recent implementations in LASP 3.0: global neural network potential with multiple elements and better long-range description. Chin. J Chem Phys 2021;34(5):583–90.
[40] Kolluru A, ShuAIbi M, Palizhati A, Shoghi N, Das A, Wood B, et al. Open challenges in developing generalizable large-scale machine-learning models for catalyst discovery. ACS Catal 2022;12(14):8572–81.
[41] Townshend RJL, Vögele M, Suriana P, Derry A, Powers A, Laloudakis Y, et al. ATOM3D: tasks on molecules in three dimensions. 2022. arXiv:2012.04035.
[42] Tolman CA. Steric effects of phosphorus ligands in organometallic chemistry and homogeneous catalysis. Chem Rev 1977;77(3):313–48.
[43] Al Hasan NM, Hou H, Sarkar S, Thienhaus S, Mehta A, Ludwig A, et al. Combinatorial synthesis and high-throughput characterization of microstructure and phase transformation in Ni-Ti-Cu-V quaternary thinfilm library. Engineering 2020;6(6):637–43.
[44] Plehiers PP, Symoens SH, Amghizar I, Marin GB, Stevens CV, Van Geem KM. Artificial intelligence in steam cracking modeling: a deep learning algorithm for detAIled effluent prediction. Engineering 2019;5(6):1027–40.
[45] Musil F, Grisafi A, Bartók AP, Ortner C, Csányi G, Ceriotti M. Physics-inspired structural representations for molecules and materials. Chem Rev 2021;121(16): 9759–815.
[46] Durand DJ, Fey N. Computational ligand descriptors for catalyst design. Chem

Rev 2019;119(11):6561–94.
[47] Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier. J Cheminform 2015;7(1):23.
[48] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50(5):742–54.
[49] Braams BJ, Bowman JM. permutationally invariant potential energy surfaces in high dimensionality. Int Rev Phys Chem 2009;28(4):577–606.
[50] Newman-Stonebraker SH, Smith SR, Borowski JE, Peters E, Gensch T, Johnson HC, et al. Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. Science 2021;374(6565):301–8.
[51] Behler J. Atom-centered symmetry functions for constructing highdimensional neural network potentials. J Chem Phys 2011;134(7):074106.
[52] Steinhardt PJ, Nelson DR, Ronchetti M. Bond-orientational order in liquids and glasses. Phys Rev B 1983;28(2):784–805.
[53] Huang SD, Shang C, Kang PL, Liu ZP. Atomic structure of boron resolved using machine learning and global sampling. Chem Sci 2018;9(46):8644–55.
[54] Huang SD, Shang C, Zhang XJ, Liu ZP. Material discovery by combining stochastic surface walking global optimization with a neural network. Chem Sci 2017;8(9):6327–37.
[55] Zahrt AF, Henle JJ, Rose BT, Wang Y, Darrow WT, Denmark SE. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. Science 2019;363(6424).
[56] Bartók AP, Kondor R, Csányi G. On representing chemical environments. Phys Rev B 2013;87(18):184115.
[57] Zhang Y, Hu C, Jiang B. Embedded atom neural network potentials: efficient and accurate machine learning with a physically inspired representation. J Phys Chem Lett 2019;10(17):4962–7.
[58] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-Learn: machine learning in Python. J Mach Learn Res 2011; 12(85): 2825–30.
[59] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada. Red Hook: Curran Associates Inc.; 2019. p. 8026–37.
[60] DevelopersTensorFlow. TensorFlow. Version 2.8.2 [software]. 2022 May 23 [cited 2022 Jun 8]. AvAIlable from: https://zenodo.org/record/6574269.
[61] Quinlan JR. Induction of decision trees. Mach Learn 1986;1(1):81–106.
[62] Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition; 1995 Aug 14–16; Montreal, QC, Canada. Piscataway: IEEE; 1995. p. 278–82.
[63] Suvarna M, Araújo TP, Pérez-Ramírez J. A generalized machine learning framework to predict the space-time yield of methanol from thermocatalytic CO_2 hydrogenation. Appl Catal B 2022;315:121530.
[64] Muraoka K, Sada Y, Miyazaki D, ChAIkittisilp W, Okubo T. Linking synthesis and structure descriptors from a large collection of synthetic records of zeolite materials. Nat Commun 2019;10(1):4459.
[65] Baysal M, Günay ME, Yıldırım R. Decision tree analysis of past publications on catalytic steam reforming to develop heuristics for high performance: a statistical review. Int J Hydrogen Energy 2017;42(1):243–54.
[66] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brAIn.Psychol Rev 1958;65(6):386–408.
[67] Bottou L. Large-scale machine learning with stochastic gradient descent. In: Lechevallier Y, Saporta G, editors. Proceedings of COMPSTAT' 2010; 2010 Aug 22–27; Paris, France. Heidelberg: Physica-Verlag HD; 2010. p. 177–86.
[68] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2017. arXiv: 1412.6980.
[69] Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Math Program 1989;45(1):503–28.
[70] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; VegasLas, NV, USA. Piscataway: IEEE; 2016. p. 770–8.
[71] Wang J, Tchapmi LP, Ravikumar AP, McGuire M, Bell CS, Zimmerle D, et al. Machine vision for natural gas methane emissions detection using an infrared camera. Appl Energy 2020;257:113998.
[72] Wang N, Li H, Wu F, Zhang R, Gao F. Fault diagnosis of complex chemical processes using feature fusion of a convolutional network. Ind Eng Chem Res 2021;60(5):2232–48.
[73] Wen L, Li X, Gao L, Zhang Y. A new convolutional neural network-based datadriven fault diagnosis method. IEEE Trans Ind Electron 2018;65(7):5990–8.
[74] Xing J, Xu J. An improved convolutional neural network for recognition of incipient faults. IEEE Sens J 2022;22(16):16314–22.
[75] Ge X, Wang B, Yang X, Pan Yu, Liu B, Liu B. Fault detection and diagnosis for reactive distillation based on convolutional neural network. Comput Chem Eng 2021;145:107172.
[76] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997; 9(8):1735–80.
[77] Bort W, Baskin II, Gimadiev T, Mukanov A, Nugmanov R, Sidorov P, et al. Discovery of novel chemical reactions by deep generative recurrent neural network. Sci Rep 2021;11(1).
[78] Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6 – 11; Sydney, NSW, Australia; 2017. p. 1263–72.
[79] Sanchez-Lengeling B, Reif E, Pearce A, Wiltschko AB. A gentle introduction to graph neural networks. Distill 2021;6(9):e33.
[80] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett 2018; 120(14):145301.
[81] Schütt KT, Sauceda HE, Kindermans P-J, Tkatchenko A, Müller K-R. SchNet-a deep learning architecture for molecules and materials. J Chem Phys 2018; 148(24):241722.
[82] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: von LuxburgU, GuyonI, BengioS, WallachH, FergusR, editors. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. Red Hook: Curran Associates, Inc.; 2017. p. 6000–10.
[83] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in neural information processing systems 33. Red Hook: Curran Associates, Inc.; 2020. p. 1877–901.
[84] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-trAIning of deep bidirectional transformers for language understanding. 2019. arXiv:1810.04805.
[85] Parmar N, Vaswani A, Uszkoreit J, KAIser L, Shazeer N, Ku A, et al. Image transformer. In: DyJ, KrauseA, editors. Proceedings of the 35th International Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. Red Hook: Curran Associates, Inc.; 2018. p. 4055–64.
[86] Ying C, CAI T, Luo S, Zheng S, Ke G, He D, et al. Do transformers really perform badly for graph representation? In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, editors. Advances in neural information processing systems 34. Red Hook: Curran Associates, Inc.; 2021. p. 28877–88.
[87] Schwaller P, Hoover B, Reymond J-L, Strobelt H, LAIno T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Sci Adv 2021;7(15).
[88] Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 2018;555(7698):604–10.
[89] Liu B, Ramsundar B, Kawthekar P, Shi J, Gomes J, Luu Nguyen Q, et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. ACS Cent Sci 2017;3(10):1103–13.
[90] Schwaller P, Gaudin T, Lányi D, Bekas C, LAIno T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chem Sci 2018;9(28):6091–8.
[91] Schwaller P, LAIno T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular Transformer: a model for uncertAInty-calibrated chemical reaction prediction. ACS Cent Sci 2019;5(9):1572–83.
[92] Jin W, Coley C, Barzilay R, Jaakkola T. Predicting organic reaction outcomes with Weisfeiler-Lehman network. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, editors. Advances in neural information processing systems 30. Red Hook: Curran Associates, Inc.; 2017. p. 2604–13.
[93] Segler MHS, Waller MP. Neural-symbolic machine learning for retrosynthesis and reaction prediction. Chemistry 2017;23(25):5966–71.
[94] Wei JN, Duvenaud D, Aspuru-Guzik A. Neural networks for the prediction of organic chemistry reactions. ACS Cent Sci 2016;2(10):725–32.
[95] Coley CW, Rogers L, Green WH, Jensen KF. SCScore: synthetic complexity learned from a reaction corpus. J Chem Inf Model 2018;58(2):252–61.
[96] Zhang L, Han J, Wang H, Car R, E W. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. Phys Rev Lett 2018; 120(14):143001.
[97] Han J, Zhang L, Car R, E W. Deep Potential: a general representation of a many-body potential energy surface. Commun Comput Phys 2018; 23(3): 629–39.
[98] Schütt K, Kindermans PJ, Sauceda Felix HE, Chmiela S, Tkatchenko A, Müller

KR. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, editors. Advances in neural information processing systems 30. Red Hook: Curran Associates, Inc.; 2017. p. 992–1002.

[99] Huang SD, Shang C, Kang PL, Zhang XJ, Liu ZP. LASP: fast global potential energy surface exploration. WIREs Comput Mol Sci 2019;9(6):e1415.

[100] Ghasemi SA, Hofstetter A, Saha S, Goedecker S. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtAIned by a neural network. Phys Rev B 2015;92(4):045131.

[101] Kito S, Hattori T, Murakami Y. Estimation of catalytic performance by neural network-product distribution in oxidative dehydrogenation of ethylbenzene. Appl Catal A 1994;114(2):L173–8.

[102] Abdul Rahman MB, ChAIbakhsh N, Basri M, Salleh AB, Abdul Rahman RNZR. Application of artificial neural network for yield prediction of lipasecatalyzed synthesis of dioctyl adipate. Appl Biochem Biotechnol 2009; 158(3):722–35.

[103] Burger B, Maffettone PM, Gusev VV, AItchison CM, BAI Y, Wang X, et al. A mobile robotic chemist. Nature 2020;583(7815):237–41.

[104] Tran K, Ulissi ZW. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. Nat Catal 2018; 1(9): 696–703.

[105] Sun Y, Liao H, Wang J, Chen Bo, Sun S, Ong SJH, et al. Covalency competition dominates the water oxidation structure-activity relationship on spinel oxides. Nat Catal 2020;3(7):554–63.

[106] Shi YF, Kang PL, Shang C, Liu ZP. Methanol synthesis from CO_2/CO mixture on Cu-Zn catalysts from microkinetics-guided machine learning pathway search. J Am Chem Soc 2022;144(29):13401–14.

[107] Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses: pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. Science 1969;166(3902):178–92.

[108] Corey EJ, Cramer III RD, Howe WJ. Computer-assisted synthetic analysis for complex molecules. Methods and procedures for machine generation of synthetic intermediates. J Am Chem Soc 1972;94(2):440–59.

[109] Corey EJ, Long AK, Rubenstein SD. Computer-assisted analysis in organic synthesis. Science 1985;228(4698):408–18.

[110] Wipke WT, Ouchi GI, Krishnan S. Simulation and evaluation of chemical synthesis-SECS: an application of artificial intelligence techniques. Artif Intell 1978;11(1–2):173–93.

[111] Mikulak-Klucznik B, Gołębiowska P, Bayly AA, Popik O, Klucznik T, Szymkuć S, et al. Computational planning of the synthesis of complex natural products. Nature 2020;588(7836):83–8.

[112] Schwaller P, Petraglia R, Zullo V, NAIr VH, Haeuselmann RA, Pisoni R, et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. Chem Sci 2020;11(12):3316–25.

[113] Genheden S, Thakkar A, Chadimová V, Reymond JL, Engkvist O, Bjerrum E. AIZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. J Cheminform 2020;12(1):70.

[114] Coley CW, Green WH, Jensen KF. Machine learning in computer-AIded synthesis planning. Acc Chem Res 2018;51(5):1281–9.

[115] Wang Z, Zhang W, Liu B. Computational analysis of synthetic planning: past and future. Chin J Chem 2021;39(11):3127–43.

[116] Badowski T, Gajewska EP, Molga K, Grzybowski BA. Synergy between expert and machine-learning approaches allows for improved retrosynthetic planning. Angew Chem Int Ed Engl 2020;59(2):725–30.

[117] Jiang Y, Yu Y, Kong M, Mei Y, Yuan L, Huang Z, et al. Artificial intelligence for retrosynthesis prediction. Engineering, In press.

[118] Lin K, Xu Y, Pei J, LAI L. Automatic retrosynthetic route planning using template-free models. Chem Sci 2020;11(12):3355–64.

[119] Coley C, Jin W, Rogers L, Jamison TF, Jaakkola TS, Green WH, et al. A graphconvolutional neural network model for the prediction of chemical reactivity. Chem Sci 2019;10(2):370–7.

[120] Kang PL, Shi YF, Shang C, Liu ZP. Artificial intelligence pathway search to resolve catalytic glycerol hydrogenolysis selectivity. Chem Sci 2022; 13(27): 8148–60.

[121] Kocer E, Ko TW, Behler J. Neural network potentials: a concise overview of methods. Annu Rev Phys Chem 2022;73(1):163–86.

[122] Blank TB, Brown SD, Calhoun AW, Doren DJ. Neural network models of potential energy surfaces. J Chem Phys 1995;103(10):4129–37.

[123] Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys Rev Lett 2007;98(14):146401.

[124] Lorenz S, Groß A, Scheffler M. Representing high-dimensional potentialenergy surfaces for reactions at surfaces by neural networks. Chem Phys Lett 2004; 395(4–6):210–5.

[125] Bartók AP, Csányi G. Gaussian approximation potentials: a brief tutorial introduction. Int J Quantum Chem 2015;115(16):1051–7.

[126] Bartók AP, Payne MC, Kondor R, Csányi G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. Phys Rev Lett 2010;104(13):136403.

[127] Chmiela S, Sauceda HE, Poltavsky I, Müller KR, Tkatchenko A. sGDML: constructing accurate and data efficient molecular force fields using machine learning. Comput Phys Commun 2019;240:38–45.

[128] Szlachta WJ, Bartók AP, Csányi G. Accuracy and transferability of Gaussian approximation potential models for tungsten. Phys Rev B 2014;90(10):104108.

[129] Deringer VL, Csányi G. Machine learning based interatomic potential for amorphous carbon. Phys Rev B 2017;95(9):094203.

[130] Unruh D, Meidanshahi RV, Goodnick SM, Csányi G, Zimányi GT. Gaussian approximation potential for amorphous Si : H. Phys Rev Mater 2022; 6(6): 065603.

[131] Deringer VL, Caro MA, Csányi G. A general-purpose machine-learning force field for bulk and nanostructured phosphorus. Nat Commun 2020;11(1):5461.

[132] Bartók AP, Kermode J, Bernstein N, Csányi G. Machine learning a generalpurpose interatomic potential for silicon. Phys Rev X 2018;8(4):041048.

[133] Bernstein N, BhattarAI B, Csányi G, Drabold DA, Elliott SR, Deringer VL. Quantifying chemical structure and machine-learned atomic energies in amorphous and liquid silicon. Angew Chem Int Ed Engl 2019;131(21):7131–5.

[134] Ma S, Shang C, Liu ZP. Heterogeneous catalysis from structure to activity via SSW-NN method. J Chem Phys 2019;151(5):050901.

[135] Shang C, Zhang XJ, Liu ZP. Stochastic surface walking method for crystal structure and phase transition pathway prediction. Phys Chem Chem Phys 2014; 16(33):17845–56.

[136] Shang C, Liu ZP. Stochastic surface walking method for structure prediction and pathway searching. J Chem Theory Comput 2013;9(3):1838–45.

[137] Liu QY, Shang C, Liu ZP. *In situ* active site for Fe-catalyzed Fischer-Tropsch synthesis: recent progress and future challenges. J Phys Chem Lett 2022;13(15): 3342–52.

[138] Liu QY, Shang C, Liu ZP. In situ active site for CO activation in Fe-catalyzed Fischer-Tropsch synthesis from machine learning. J Am Chem Soc 2021; 143(29):11109–20.

[139] Li XT, Chen L, Shang C, Liu ZP. *In situ* surface structures of PdAg catalyst and their influence on acetylene semihydrogenation revealed by machine learning and experiment. J Am Chem Soc 2021;143(16):6281–92.

[140] Kang PL, Shang C, Liu ZP. Large-scale atomic simulation via machine learning potentials constructed by global potential energy surface exploration. Acc Chem Res 2020;53(10):2119–29.

[141] Kang PL, Shang C, Liu ZP. Glucose to 5-hydroxymethylfurfural: origin of siteselectivity resolved by machine learning based reaction sampling. J Am Chem Soc 2019;141(51):20525–36.

[142] Ma S, Huang SD, Fang YH, Liu ZP. TiH hydride formed on amorphous black titania: unprecedented active species for photocatalytic hydrogen evolution. ACS Catal 2018;8(10):9711–21.

[143] Ko TW, Finkler JA, Goedecker S, Behler J. A fourth-generation highdimensional neural network potential with accurate electrostatics including non-local charge transfer. Nat Commun 2021;12(1):398.

[144] Sasaki M, Hamada H, KintAIchi Y, Ito T. Application of a neural network to the analysis of catalytic reactions analysis of NO decomposition over Cu/ZSM-5 zeolite. Appl Catal A 1995;132(2):261–70.

[145] Mohammed ML, Patel D, Mbeleck R, Niyogi D, Sherrington DC, Saha B. Optimisation of alkene epoxidation catalysed by polymer supported Mo(VI) complexes and application of artificial neural network for the prediction of catalytic performances. Appl Catal A 2013;466:142–52.

[146] Günay ME, Yildirim R. Knowledge extraction from catalysis of the past: a case of selective CO oxidation over noble metal catalysts between 2000 and 2012. ChemCatChem 2013;5(6):1395–406.

[147] Günay ME, Yildirim R. Neural network analysis of selective CO oxidation over copper-based catalysts for knowledge extraction from published data in the literature. Ind Eng Chem Res 2011;50(22):12488–500.

[148] Omata K. Screening of new additives of active-carbon-supported heteropoly acid catalyst for Friedel-Crafts reaction by Gaussian process regression. Ind Eng Chem Res 2011;50(19):10948–54.

[149] Rohrbach S, Šiaučiulis M, Chisholm G, Pirvan P-A, Saleeb M, Mehr SHM, et al. Digitization and validation of a chemical synthesis literature database in the ChemPU. Science 2022;377(6602):172–80.

[150] Perera D, Tucker JW, Brahmbhatt S, Helal CJ, Chong A, Farrell W, et al. A platform for automated nanomole-scale reaction screening and micromolescale

synthesis in flow. Science 2018;359(6374):429–34.
[151] Ulissi ZW, Tang MT, Xiao J, Liu X, Torelli DA, Karamad M, et al. Machinelearning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO_2 reduction. ACS Catal 2017;7(10):6600–8.
[152] Liu X, Xiao J, Peng H, Hong X, Chan K, Nørskov JK. Understanding trends in electrochemical carbon dioxide reduction rates. Nat Commun 2017;8(1):15438.
[153] Zhong M, Tran K, Min Y, Wang C, Wang Z, Dinh C-T, et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. Nature 2020;581(7807):178–83.
[154] Yoshikawa N, Kubo R, Yamamoto KZ. Twitter integration of chemistry software tools. J Cheminform 2021;13(1):46.