

人工智能助力当代化学研究

朱博阳^{*,a} 吴睿龙^a 于曦^{*,a,b}

(^a 天津大学化学系 天津 300072)

(^b 天津市分子光电科学重点实验室 天津 300072)

摘要 以机器学习为代表的人工智能在当代的科学研究中正在发挥越来越重要的作用。不同于传统的计算机程序,机器学习人工智能可以通过对大量数据的反复分析和自身模型的优化,即“学习”过程,从而在大量的数据中寻找客观事物的相互联系,形成具有更好预测和决策能力的新模型,做出合理的判断。化学研究的特点恰恰是机器学习人工智能的强项。化学研究经常要面对十分复杂的物质体系和实验过程,从而很难通过化学物理原理进行精准的分析 and 判断。人工智能可以挖掘化学实验中产生的海量实验数据的相关性,帮助化学家做出合理分析预测,大大加速化学研发过程。本文介绍了当代人工智能方法及用其解决化学问题基本原理,并通过具体案例展示了人工智能辅助解决不同化学研发问题的方法以及对应的机器学习算法。将人工智能运用在化学科学的尝试正处于蓬勃上升期,人工智能已经初步展示对化学研究的强大助力,希望本文能帮助更多的国内的化学工作者了解和运用这一有力的工具。

关键词 人工智能; 机器学习; 分子描述符; 深度学习

Artificial Intelligence for Contemporary Chemistry Research

Zhu, Boyang^{*,a} Wu, Ruilong^a Yu, Xi^{*,a,b}

(^a Department of Chemistry, Tianjin University, Tianjin 300072, China)

(^b Tianjin Key Laboratory of Molecular Optoelectronic Sciences, Tianjin 300072, China)

Abstract Artificial intelligence (AI), especially the machine learning, is playing an increasingly important role in contemporary scientific research. Unlike the traditional computer program, machine learning can analyze a large number of data repeatedly and optimize its own model, a process which is called a “learning process”. So that the AI can find the relationship underlying the experiments from a large number of data, form a new model with better prediction and decisionmaking ability, and make an optimized strategy. The characteristics of chemical research just hit the strengths of machine learning. Chemical research often faces very complex material system and experimental process, so it is difficult to accurately analyze and making judgment through physical chemistry principles. Artificial intelligence can mine the correlation of massive experimental data generated in chemical experiments, help chemists make reasonable analysis and prediction, and therefore greatly accelerate the process of chemical research. This review presents the modern artificial intelligence method and its basic principles on solving chemical problems, by representative examples with specific machine learning algorithm. The application of artificial intelligence in chemical science is in a period of vigorous rise. Artificial intelligence has initially shown a powerful assist to chemical research. We hope this review can help more domestic chemical workers understand and use this powerful tool.

Keywords artificial intelligence; machine learning; molecular descriptor; deep learning

1 引言

用机器代替人的工作,从而减轻人的劳动负担、提高工作效率是人类长久追寻的梦想。人类近代历史上每一次技术革命都以新型机器的出现和大量应用为标志。与通常的机器代替人的体力劳动不同,人工智能是让机器代替人类的部分脑力工作,是人类对机器能力的进一步拓展。

人工智能的最初设计可以追溯至自动化思想,即用机器来代替人发出重复、有规则的、按部就班的指令,指导工作一步一步按流程完成。这一思路最终产生了传统的冯·诺依曼计算机结构。这种结构中,指令在电脑程序中运行,完成某项特定的工作。例如在一个汽车生产车间中,程序控制机械的运行,完成从生产加工的各个环节,代替了人的规划、指导、控制的工作。然而这一模式中,工作程序是人们预先设定好的,执行过程是机

* E-mail: luciszhu@outlook.com; xi.yu@tju.edu.cn

Received July 12, 2020; published August 21, 2020.

Project supported by the National Natural Science Foundation of China (Nos. 21973069, 21773169, 21872103), National Key R&D Program (Nos. 2017YFA0204503, 2016YFB0401100), the PEIYANG Young Scholars Program of Tianjin University (No. 2018XRX-0007) and the College Student Innovation and Entrepreneurship Training Program of Tianjin University (No. 201910056451).

项目受国家自然科学基金(Nos. 21973069, 21773169, 21872103)、国家重点研究开发计划(Nos. 2017YFA0204503, 2016YFB0401100)、天津大学北洋青年学者计划(No. 2018XRX-0007)和天津大学大学生创新创业训练计划(No. 201910056451)资助。

械重复的, 而不存在程序自我“学习”“改进”的过程, 因而这种方式不是现代意义上的人工智能。

当代的人工智能强调机器的“学习”、“思考”能力, 即代替人做分析、模拟、预测、决策等脑力工作。在冯·诺依曼型的自动化过程中, 尽管机器会依据外界客观情况的变化做出不同的响应、发出不同的指令。然而, 其处理的外界信息必须是某种规范的数据, 输入的数据与执行操作间的联系由明确的人工编辑的程序设定。而当代人工智能是以大量的外界复杂多样的数据为驱动的计算机结构, 要求其可以对外界提供的大量复杂多样的数据进行分析总结, 自动提取事物特征, 寻找客观世界的规律, 从而形成应对某一问题的新的模型(程序), 这就是机器的“学习”能力。之后再依据这一模型程序做出相应的判断和操作。这就是目前人工智能的前沿领域, 即所谓“机器学习”的研究。从结构和运行过程上看(如图1所示)机器学习区别于传统自动化主要在于机器学习首先要通过对外界数据的不断的“学习”过程而改进自己, 产生应对某一问题独特的模型(程序), 而后再用这一模型或者程序来实现判断和操作。这一过程更接近真实人类学习思考过程。

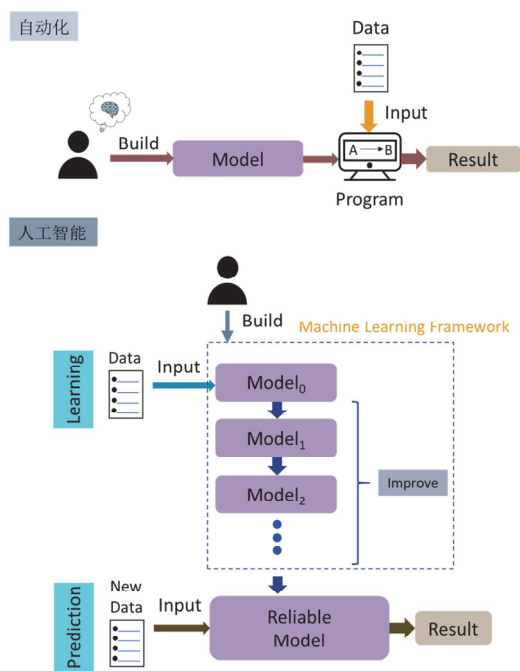


图1 人工智能思想的发展。人工智能与自动化最大的区别是程序的来源: 自动化程序是手动设置的固定模型, 对特定输入作出响应; 人工智能程序是一套模板, 可以从已有数据中挖掘规律, 并进行自我学习和改进, 自动生成可信的模型, 这种模型对预料之外的输入也能作出正确响应, 对全新情境有预测分析能力。

Figure 1 The development of artificial intelligence concept. The biggest difference between artificial intelligence and automation is the source of the program: the automation program is a fixed model manually set by human, which responds to specific inputs; the artificial intelligence program is a model that can mine laws from existing data and conduct self-learning to improve itself, generate reliable model automatically, this model can also correctly respond to unexpected inputs, and has the ability to predict and analyze new situations.

近些年来, 以机器学习为代表的人工智能领域的研究进入了高速发展时期, 重要的成果不断展现。最著名的例子莫过于基于深度学习算法开发的 AlphaGo。它能够自主学习围棋, 最终经过多次“学习”训练后成为人类无法匹敌的围棋智能^[1]。如今人工智能可以实现语言识别、图像识别, 辅助医生分析和诊断患者病情^[2], 选取肿瘤的特征基因^[3]等功能。机器学习的理论和技术日益成熟, 应用范围也越来越广, 正在深刻的改变着人类社会的工作方式, 化学领域也不例外。

近些年, 机器学习已经越来越多的被应用于化学研发领域, 机器学习可以在大量的化学信息中寻找发现数据间的联系和规律, 帮助化学工作者做出更合理的判断和决策, 加速化学研发过程。机器学习应用于化学领域已经取得很多引人注目的结果。如用人工智能学习寻找分子结构与其物理性质间的联系, 如偶极矩、折光指数、沸点等^[4]。机器学习可以分析数以万计的有机化学反应, 然后用做逆合成分析, 寻找产物的最佳合成路径; 利用机器学习从海量化合物中寻找具有特定药理性质的化合物, 等等。机器学习正在逐渐成为继计算化学和分子模拟以外另一种计算化学研究的新范式, 取得了快速的发展。

国内化学研究领域对现代人工智能认知度尚比较有限, 将机器学习引入化学研究的工作刚刚起步, 也取得了一定的成果。如马雷等^[5]利用神经网络等方法, 基于大量实验数据的探索, 找到了电极材料参数与超级电容器比电容性能的相关关系和相对重要性。孙宽等^[6]用人工智能筛选 PV 材料。钟苗等^[7]通过机器学习辅助开发高性能电催化材料, 实现高效率电催化还原二氧化碳制乙烯。本文旨在对当前以机器学习为代表的人工智能在化学研究领域的发展做一个简要浅显的介绍, 为国内广大化学工作者提供参考。我们首先概述了机器学习的基本原理和机器学习解决化学问题的基本思路。而后梳理总结了近些年机器学习在化学研究中的部分有代表性的应用。最后对这一领域做了展望。

2 人工智能在化学研究中的应用

2.1 机器学习的基本原理

人类学习过程可以粗略分为“收集”信息和“处理”信息两个步骤, 相应的当代人工智能研究也包含“知识处理”和“机器学习”两个方面。

知识处理研究机器的“感知”能力, 即机器如何正确、有效地提取数据的关键特征, 为学习过程提供优秀的“学习资料”。对于化学研究而言, 知识处理即是化学特性和信息符号化、数字化, 使其成为计算机可处理的对象。例如, 对于有机合成化学, 主要特征信息即是化合物的分子结构, 对于无机材料而言则是组成和晶体结构, 对于高分子材料, 除了分子结构外还应包括平均分子量、分布和聚集态等等。化学特征信息也会依据

所研究的问题的不同而变化。如研究有机溶剂的沸点,则分子量、分子偶极、氢键给受体结构、对称特征等就是必要的化学特征。

“机器学习”即是当代人工智能处理信息的主要内容。它区别于传统计算机的最显著的特点是其会通过输入信息的不断地分析,从多次“学习”中得到自我提升,从而具有更强的分析和预测、决策能力。按照解决问题性质的不同,机器学习可以分为两类。第一类被称为回归问题,用于寻找数据间的某种特定的联系,如分子组成与分子量的关系,分子结构与熔沸点的关系等化学中的结构-效能关系研究就是常见的回归问题。所采用的机器学习方法通常为监督学习,即在问题(条件数据)和答案(结果数据)间建立确定的联系。第二类问题为分类问题,如对大量的化学物质、化学反应按其特点进行分类。一般用无监督学习应对这类问题,通过对数据间的相关关系的分析找到相似性,从而进行归类。此外还有强化学习,它对于解决决策问题有很好的效果,特别是连续的决策,如解决对弈问题的 AlphaGo 应用的就是强化学习。在化学研发中的逆合成分析需要设计优化的合成路线,是一个连续多步骤的决策问题,因而需要采用强化学习。

通常来说,机器学习对数据的处理分为输入、训练和输出三个阶段,分别对应了人类学习过程中的经验积累、规律总结,解决实际问题的思维过程。在输入和训练阶段,数据代表机器的“学习资料”和“积累的经验”;而模型输出的数据,既是机器学习的结果,也是评价机器学习模型好坏的依据。以监督学习为例,机器学习辅助解决问题的过程如图 2 所示。首先准备足够的原始数据(包括所有正面反面的实验数据);选择合适的机器学习

习算法,搭建程序结构,并将原始数据输入到模型中,使模型能够针对一组输入给出相应的输出;通过多次训练并优化机器学习模型直到模型输出值符合预期;最后用未被“学习”过的数据验证模型对新情境的预测能力。可对新的情境做出正确响应的模型被认为是优秀的机器学习模型,这样的模型可以解决实际问题,实现特定应用。

对于机器学习的过程,算法是其核心,即人工智能学习和处理信息的方式。合适的算法可以增强模型的表达能力与应用价值^[8]。研究者需要根据所研究问题的不同,选择具有相应功能的机器学习算法对数据进行训练。常见的机器学习算法有支持向量机^[9]、决策树^[10]、神经网络^[11]、蒙特卡洛树搜索^[12]等。后文将结合案例对这些算法进行具体介绍。

2.2 人工智能解决化学问题的基本方法

2.2.1 化学特征的提取

将机器学习应用于化学研究的第一步是让机器能够接受化学的特征和信息。化学体系中的信息非常多样而繁杂,将这些信息进行某种形式的筛选或者编码,形成化学信息的某种表达方式,这即是“分子描述符”^[13,14]。例如,对于一个有机分子而言,它化学组成就可以看作一种分子描述符,它的标准命名法也是一种描述符。选择合适的分子描述符,需要依据待研究的特性及其与化学结构间的关系来考虑。例如,如果待研究的性质为材料的电学性质,那么可能就要考虑采用描述分子极化率等方面的性质的描述符;而如果研究分子的药理学活性,则要求描述符能够包含分子的三维结构和手性性质^[15-18]。对于一些结构性性质相对简单的情况,分子描述符可以简单直接。如图 3(a)所示,研究分子的液晶性质时,由于理论上已经比较明确液晶相的出现与分子的刚性和柔性相关,因而分子描述符就包含了四个数据:分子柔性部分的长度、刚性部分的长度、分子总长度、以及分子量^[19]。而当结构性性质比较复杂的时候则需要更复杂的描述符。如图 3(b)所示为研究巴比妥酸衍生物作为尿素酶抑制剂的时候所采用的 16 种不同的描述符^[19]。因为已经知道蛋白质与抑制剂的相互作用会与分子间相互作用力以及分子的形状有很大的关系,因而描述符包含了分子的氢键给受体原子数、输水原子数、偶极矩、形状因子等的信息。

依据具体应用场景和机器学习算法的不同,人们已经开发出了几千种不同的分子描述符。如图 4 所示^[20,21],分子描述符可以分成不同的类别。如根据数据来源可以分为实验描述符和理论描述符^[22];根据分子的结构维数可以分为零至三维。零维分子描述符即分子化学式,不涵盖任何分子结构信息。一维分子描述符描述分子的组成序列,如官能团,结构片段,分子结构信息含量低。二维分子描述符考虑原子如何连接,三维分子描述符又称几何描述符,来自空间中分子的坐标^[23]。

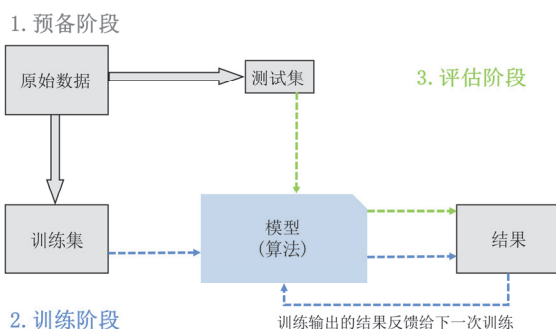


图 2 机器学习过程。(1)预备阶段:将原始数据分割成训练集和测试集,训练集作为模型的“学习资料”,测试集验证模型对新情境的响应能力(2)训练阶段:搭建训练集—模型—结果的训练结构;调整算法的优化参数,多次迭代直到模型输出结果符合预期值(3)测试阶段:使用测试集验证模型的表达能力。

Figure 2 The machine learning process. (1) Preprocessing: Divide the raw data into a training set and a test set, the training set serves as the model's "learning material", and the test set verifies the model's ability to respond to new situations. (2) Training: Build the structure of training set to model to test set. Adjust the optimization parameters of the algorithm inside the model and iterate multiple times until the model output meets the expected value. (3) Evaluation: Use the test set to evaluate the expressive ability of the model.

根据描述符的数据类型可以分为布尔值^[24](如是否是手性)、实数(如分子量)、向量(如偶极矩)、张量(如极化率)等等。除了这种分子性质的直接描述外,描述符还可以是通过对分子特征(如分子结构、性质、片段或子结构)进行某种形式的编码形成对分子具有特定描述符的分子指纹,称为定性描述符^[25];亦可以通过将分子的结构通过场或者图论的方法抽象成描述符,这称为定量描述符。

a.

Sample class	Three-dimensional (3D) structure for several organic compounds	L_{rig} (Å)	L_{flex} (Å)	S	M (a.m.u.)
1		10.74	8.08	0.146	326.4
2		11.78	13.95	0.110	363.5
3		11.88	12.76	0.120	346.5
4		12.04	12.76	0.119	345.5
5		11.52	21.56	0.089	484.7
6		17.89	5	0.126	420.5

b.

Code	Description
a_acc	Number of hydrogen bond acceptor atoms
a_don	Number of hydrogen bond donor atoms
a_hyd	Number of hydrophobic atoms.
vs_a_acc	Approximation to the sum of VDW surface areas (Å ²) of pure hydrogen bond acceptors
Q_PC+PEOE_PC+	Total positive partial charge
Q_PC-PEOE_PC-	Total negative partial charge
i3D	Dipole moment
1K	Kier first shape index
2K	Kier second shape index
3K	Kier third shape index
Weight	Molecular weight (including implicit hydrogens) in atomic mass units with atomic weights
SlogP	Log of the octanol/water partition coefficient.
TPSA	Polar surface area (Å ²)
a_nN	Number of nitrogen atoms

图3 (a)分子 3D 结构和分子性质描述符。(b)利用 MOE 计算分子描述符^[4,19]。

Figure 3 (a) Values of 3D molecular descriptors. (b) Molecular descriptors calculated with MOE^[4,19]. Reprinted with permission from ref. [19]. Copyright [MDPI, Basel, Switzerland].

分子描述符的生成在简单情况下可以由化学工作者手动提取,如上面研究液晶性质的例子,直接用分子的理论长度。也可以通过专业软件,如巴比妥酸的例子,分子描述符是由 MOE 软件 Molecular Operating Environment 计算得到的。目前已有很多分子描述符软件和数据库可供使用,如 DRAGON^[26], GRID^[27], MOLE db^[28]等(图 5)。

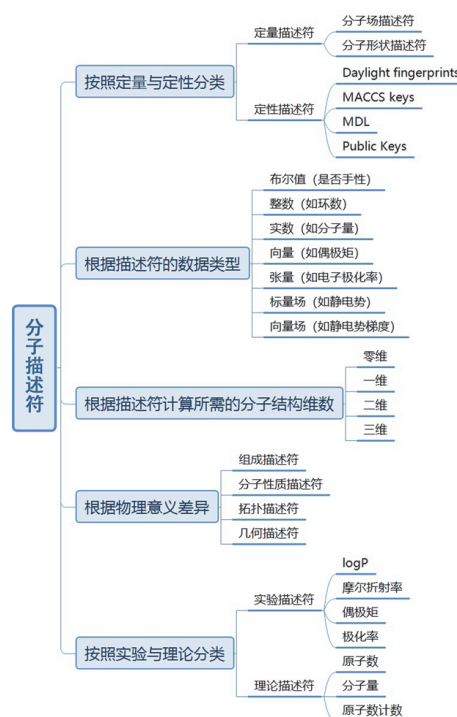


图4 分子描述符的粗略分类。其中,定性描述符又被称为分子指纹,即将分子结构、性质、片段或子结构信息用某种编码来表示^[21]。

Figure 4 Rough classification of molecular descriptors. The qualitative descriptor is also called molecular fingerprint, by using a certain code to represent the molecular structure, property, fragment or substructure information^[21].

软件 Software	描述符数量 Descriptor-Num	软件 Software	描述符数量 Descriptor-Num
Dragon	4885	MOE	>300
MODEL	3778	Joelib	>40
MolGen	707	ADAPT	>260
Cerius 2	73	SYBYL	>100
Sarchitect	1084	PowerMV	>1000
PreADMET	>2500	CODESSA	>1500
MOLE-db	1124	Pipeline Pilot	>390
ADMET Predictor	297	SARNavigator	>70

b. MOLE db - Molecular Descriptors Data Base

The MOLE db - Molecular Descriptors Data Base is a free on-line database comprised of 1124 molecular descriptors calculated for 2,497,23 molecules. At the present moment, several queries have been made on the database. This data base is intended as a research and teaching tool: please, read here general details, references, credits, limitations, warranty and conditions. Otherwise, click here for help on how to query the data base.

Developed by Milano Chemometrics and QSAR Research Group :: Info and details on Molecular Descriptors Data Base

图5 (a)常用分子描述符计算软件及其可计算的描述符数量。(b) MOLE-db 界面^[28]。

Figure 5 (a) Commonly used molecular descriptor calculation software and the number of descriptors that can be calculated. (b) MOLE-db interface^[28].

2.2.2 机器学习辅助化学研发过程

“Chem is try”描绘了化学研究的典型依赖实验和试错过程的特点。经典的实验科学研究流程一般可概括为：设计实验、进行实验、收集数据、分析数据、改进实验、得到结论的过程。化学和材料的开发研究也是如此(如图 6 所示)，而化学的实验过程更加复杂，条件繁多，数据量大。这主要是因为化学研究对象是分子原子组成的多尺度的复杂体系。尽管现代化学理论的发展使得宏观化学反应和微观物质结构的变化的规律已经有理可循，现代表征工具的发展也使得对化学过程的检测和控制更精细。然而化学体系的复杂和多样性使得物理化学规律仍然只是在原则上有效，而常常无法做到精准。另一方面，借助于现代计算机高速计算能力，运用量子力学的原理，从原子出发计算化学体系的性质(即第一性原理计算的方法)，或者利用分子运动的规律，模拟分子体系的演变(即分子动力学模拟)的计算化学方法在近些年也取得飞速的发展，理论对于实验的指导能力也不断提高。然而同样是由于体系的复杂性使得理论计算的计算量极大，所能应对的体系尺度和复杂程度依然有限。因此大量的化学的研究只能在物理化学原则所划定的很大的空间内依靠经验和试错来进行。造成了化学研发的高成本和高风险。分析现有的化学实验数据，做出正确的新实验的决策是减少试错、加速开发的有效方法。这恰恰是机器学习擅长的战场^[29]。机器学习在这一过程中可以用于数据分析，总结数据中反映出来的规律，寻找因果间的关系，提出更精准实验设计和策略，减少试错过程，以节约实验成本，加速研究进度。

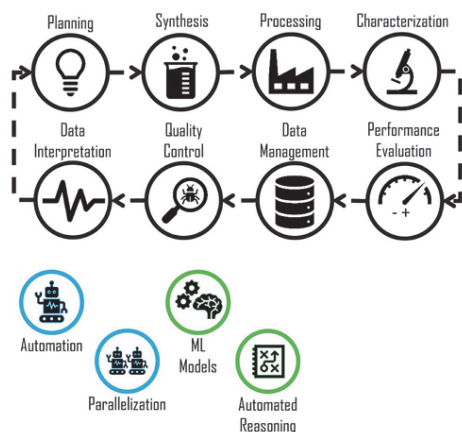


图 6 机器学习辅助化学研发^[29]

Figure 6 Machine learning assists chemical research^[29]. Reprinted with permission from ref. [29]. Copyright [Chemical Science].

前面已经提到，机器学习解决化学问题集中在分类和寻找因果关系(回归)两大方面，而在此基础上还有进一步的决策问题，分别对应了机器学习中的非监督学习，监督学习和强化学习。以寻找因果关系，即实验条件与实验结果间的关系问题为例。应用人工智能过程中，输入人工智能程序的训练样本是实验条件以及人们

关心的可能会对实验结果产生影响的因素(即输入)，以及实验结果(即输出)。人工智能通过分析，找到条件与结果(输入与输出)间的关系，从而帮助设计新的实验和做出新的决策。在构建机器学习算法模型时，需要建立从输入到输出的关系。例如在研究介电材料结构对超级电容器性质的影响时，输入是电容材料的结构参数，如比表面积、杂原子、纵横比，纤维素含量，输出是超级电容器的电容；在寻找反应条件与反应产物间的关系时，输入是反应物与反应条件，输出是产物。通过大量输入输出数据不断对机器学习模型反复训练，直到模型对输入数据能够产生正确输出数据，这时模型训练就完成了，就得到了输入和输出数据间的相关关系，也就是实验条件和实验结果间的关系，完成了数据分析的工作。接下来，向人工智能模型输入未知条件(未经过实验验证的条件)，就可以得到预测的结果。这就是预测过程。

可以看出，相比于精细复杂的理论计算，人工智能只在因果之间建立经验性的联系，而通常不涉及深层次的理论。只要所选取的训练数据和预测数据在相似的条件里，机器学习就可以快速准确的做出判断。而正因为这样的经验判断特性，其预测速度比理论计算快得多。特别是面对复杂的大量数据的体系，理论可能太复杂以至于无法建模或者计算非常缓慢的情况下，机器学习可以以很低的时间和计算成本得到相当准确的结果。因此正在逐渐成为化学研究中继实验、理论计算、计算机模拟以外的另一研究范式。利用机器学习算法可以高效的实现分子构效关系分析，分子设计和实验规划(图 7)^[30]。

此外传统研究方法难以靠理论或者实验遍历全部化学空间，而机器学习辅助解决化学问题可以触及经验主义的盲区，对寻找目标化合物工作有着难以估量的价值。例如 Isayev 等利用机器学习算法通过对少量无机材料的物理和化学性能进行学习，从而对数据库中全部空间近 3 万种无机材料进行预测。对其中绝缘体材料的 7 种热机械性能，带隙能(E_{BG})，体积模量(B_{VRH})，剪切模量(G_{VRH})，德拜温度(θ_D)，定压热容(C_P)，定容热容(C_V)以及热扩散系数(α_v)进行回归。机器学习模型所预测的热机械性能与实验和计算的数据基本吻合^[31]。同时，机器学习模型在自我优化和对其它数据进行学习时，就相当于进行了多组平行实验与对照试验，在实验过程中也能有效为化学家减负。

机器学习运用于化学研究，特别是在近几年中，已经取得了很多非常优秀的成果。对于机器学习在化学领域的研究热度分布，Cova 和 Pais^[32]做了统计。图 8 显示了 2008~2019 年(6 月 30 日)期间，以化学领域(如有机，无机，分析，物理和生物化学等)和解决问题的种类作为两个分类标准，分析机器学习在化学领域和化学问题中研究的分布状况。如图 8 所示，红色为机器学习研究较多的领域，而黄色为较少的领域。我们可以发现

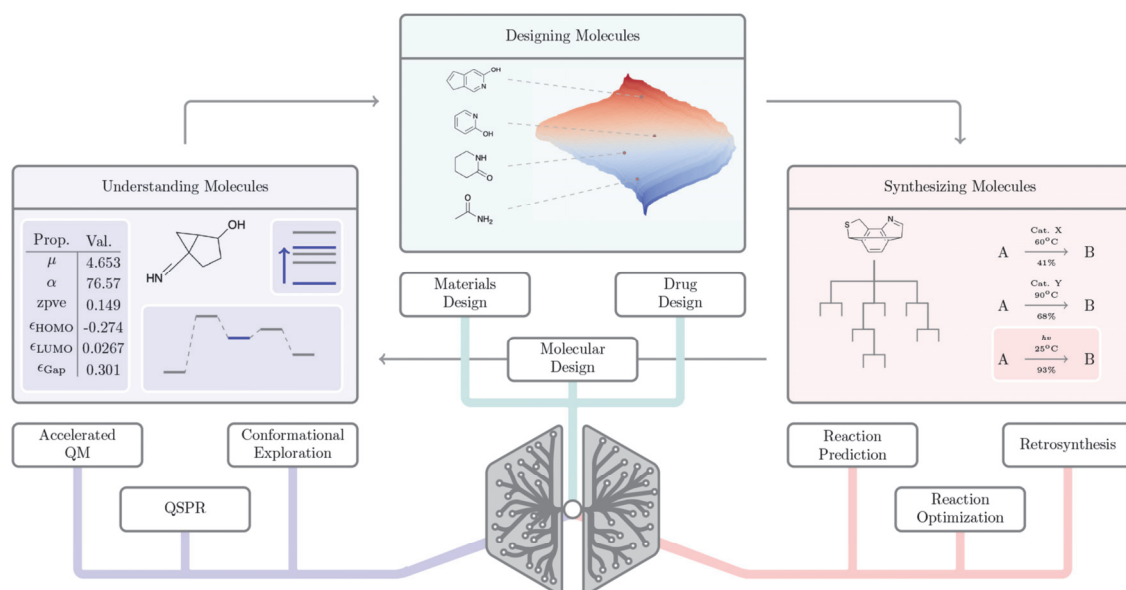


图7 机器学习对化学工作流程的影响. 分子的表征, 设计, 合成的说明性示例分别在有颜色的对话框中显示, 每个对话框表示一个封闭工作过程^[30].

Figure 7 Influence of machine learning on the chemical workflow. Illustrative examples of molecular characterization, design, and synthesis are displayed in colored dialog boxes, each of which represents a closed working process^[30]. Reprinted with permission from ref. [30]. Copyright © 2019, American Chemical Society.

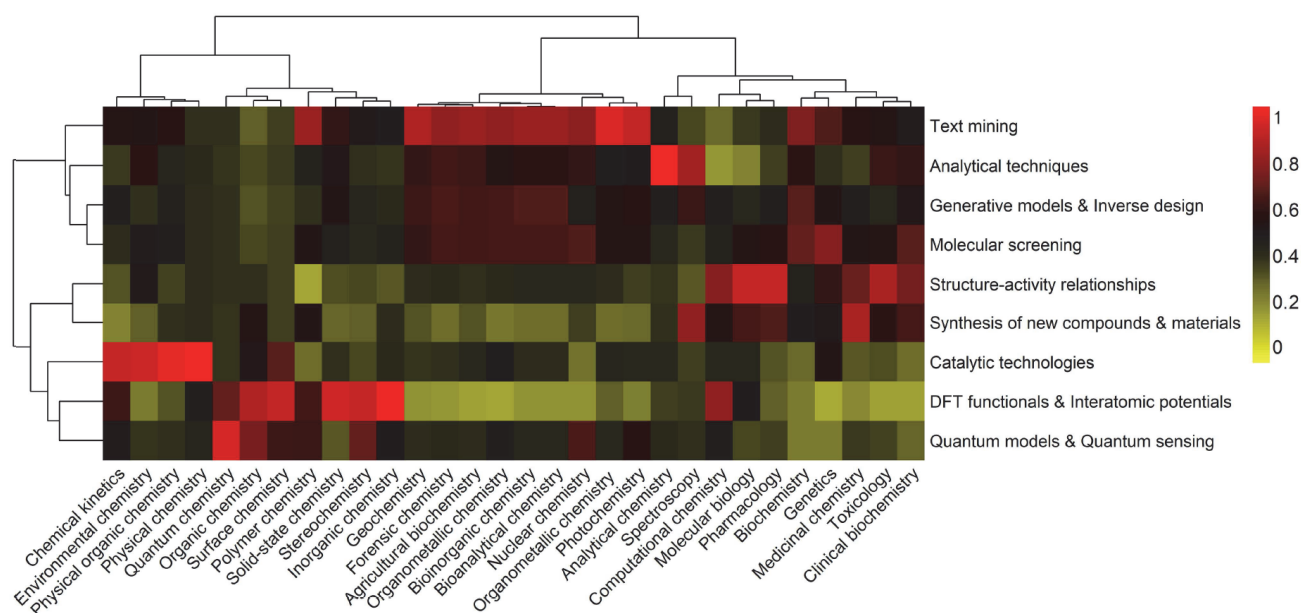


图8 用聚类热图表现 ML 在化学研究中的贡献. 数据表示为出版物总数最多的部分, 包括文章, 评论和书籍, 包含特定的同时出现的关键词, 并遵循标准的标准化程序. 在化学子字段和应用类型上都进行了具有欧氏距离和 Ward 链接的层次聚类. 使用黄色到红色的配色方案进行着色, 最高和最低相对贡献分别对应于 1(红色)和 0(黄色)值^[32].

Figure 8 Using clustering heat map to show the contribution of ML in chemistry research. Data are expressed as fractions of the highest number of publications, including articles, reviews and books, containing specific co-occurring keywords, and following a standard normalization procedure. Hierarchical clustering with Euclidean distances and Ward linkage was performed on both Chemistry sub-fields and type of application. Co-occurrences are colored using a yellow-to-red color scheme. Highest and lowest relative contributions correspond to 1 (red) and 0 (yellow) values, respectively^[32]. Reprinted with permission from ref. [32]. Copyright © 2019 Cova and Pais.

机器学习的热点主要还是以解决问题的种类分布为主, 在图中表现为横向集中分布. 结合化学学科分类, 可以发现主要有三个热点区域, 分别是(i)文本挖掘, 分析技术, 生成方式和逆向设计以及分子筛选(ii)构效关系和

合成新化合物和新材料(iii)催化技术, DFT 理论和原子间势能.

我们接下来通过部分典型的例子进一步具体介绍人工智能在化学研究中的应用方法. 由于篇幅的限制,

这些例子仅仅只是管中窥豹, 远远不能涵盖这一领域的发展.

2.3 人工智能解决化学问题的例子

2.3.1 寻找结构与性能之间的关系

(1) 影响超级电容器电容的因素分析

能量存储是当今社会的热点问题, 超级电容器由于其本身的高功率密度、短充/放电时间和优秀的循环稳定性, 是非常有竞争力的能量存储装置. 活性炭有着比表面积大, 化学性质稳定, 导电性能高和结构可调整的特点, 是超级电容器研究的重点对象.

影响超级电容器性能的因素有很多, 例如: 活性炭高比表面积提供了丰富的接触面, 有利于吸收电解质离子从而增加比电容; 活性炭中的微孔有利于提升双层电容器的性能, 中孔则提供离子传输通道; 杂原子, 纵横比, 纤维素含量等因素同样会对其产生影响. 问题涉及因素多、化学模型复杂是超级电容器研究的主要难点.

马雷等^[5]用酸浸法从念珠藻中合成了可被制作成超级电容器的活性炭材料, 应用神经网络和随机森林两种机器学习方法, 预测了活性炭材料的各个特征参数与超级电容器比电容之间的关系, 分析比较了活性炭材料各个特征参数之间的相对重要性. 他们首先通过多种分析测试方法, 例如使用热重分析仪检测碳化期间的重量变化; 扫描电子显微镜和透射电子显微镜表征样品的表面形貌和微观结构; X 射线衍射仪测定结晶结构; 利

用 196 °C 下的氮吸附-脱附过程表征孔的结构; 多点 Brunauer-Emmett-Teller 法测定样品的特定表面积; 环形/狭缝密度泛函理论估计孔径分布等方式, 将总计十个活性炭特征参数, 即所谓描述符, 纳入考量. 将这些特征输入到人工神经网络的算法中去, 将整体数据的 90% 作为训练集, 其余 10% 作为测试集. 训练子集用于训练模型, 最终在结构和超级电容器电容之间建立关系(图 9).

该实验中主要使用的神经网络(NN)是一种目前应用最广泛的主流机器学习算法, 常用于监督学习, 在化学中常用的就是结构效能关系的学习. 它是一种“黑箱模型”, 特点是面对一些复杂或难以用清晰的数学模型表达的问题, 使用者不需要知道其内部规律即可完成对数据的分析预测^[33]. 这一模型来源于生物的神经元结构, 于 1943 年由心理学家 McCulloch 和数学家 Pitts^[34]提出. 模型中数据的输入可以类比为神经元的树突, 数据的计算过程比作细胞核, 输出过程比作轴突(图 10). 通过对输入的数据进行加权求和, 得到一个输出值, 将得到的输出值与预期值进行对比, 不断调整每条线上的权值, 整个模型得到了训练和学习. 这一方法主要依靠多层多个节点中权值的调整, 从而可以综合多个因素对于某一个性质的影响, 非常适合于化学中复杂体系的关系的学习判断, 是化学研究中应用最广泛的机器学习算法之一. 不过在使用神经网络方法时, 需要注意神经网络存在参数较多、学习时间长, 有可能陷入局部最小值等缺点.

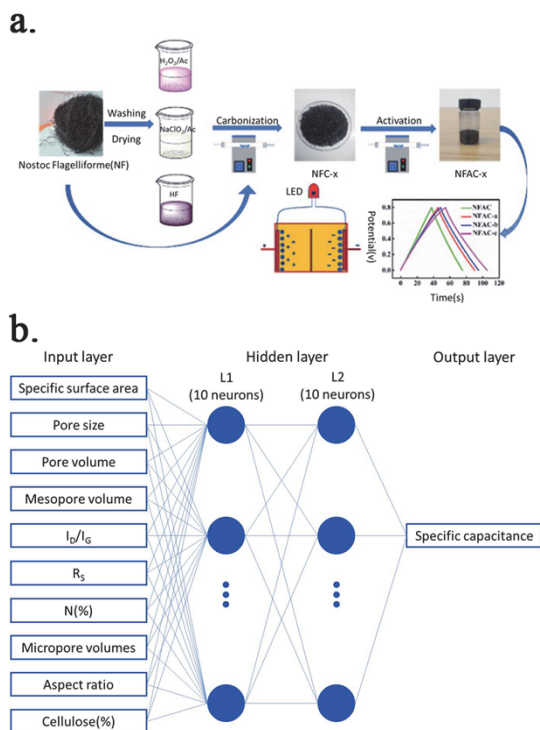


图9 (a) NFAC-x 合成方式. (b) 超级电容器电容分析的人工神经网络模型^[5]

Figure 9 (a) NFAC-x synthesis method. (b) Artificial neural network model for supercapacitor capacitance analysis^[5]. Reprinted with permission from ref. [5]. Copyright [RSC Adv.]

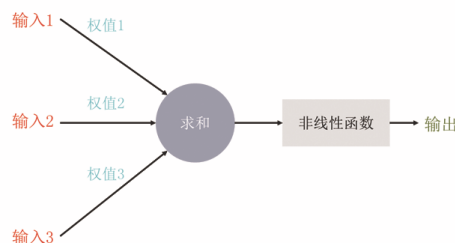


图10 M-P 类神经元运算模型

Figure 10 M-P neural model

为进一步评估电容材料中各参数对于电容性质的相对重要性和贡献, 他们进一步使用了基于决策树的随机森林算法^[35,36], 得到了各输入参数对结果影响的占比, 发现其中影响最大的比表面积相对重要性达到了 30.1%. 这里所用的随机森林是一种灵活的、便于使用的机器学习算法, 经常用来进行分类和回归任务. 随机森林算法可以对变量的重要性进行度量评估^[37], 由多个决策树集成得到, 每个决策树都对输入的数据产生分类效果, 随机森林则综合了所有分类的投票结果, 并将投票次数最多的结果作为最优输出(图 11). 随机森林可以处理特征值很多的高维数据, 是一种十分有效的降维方法. 在部分数据缺失的情况下, 随机森林仍能保持较

高的精确性. 而随机森林的缺点在于在面对噪音较大的问题时可能会存在过拟合现象.

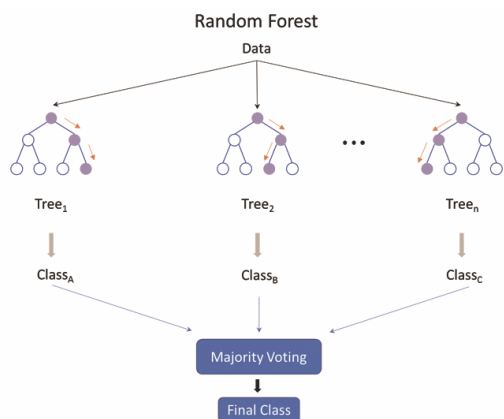


图 11 随机森林模型

Figure 11 Random Forest Model

本案例利用机器学习模型精确地预测了生物基碳材料的电性能, 分析了各参数对比电容的贡献, 为未来储能材料的合成提供了指导. 作者也详细解释了利用人工神经网络和随机森林进行数据预测的方法过程, 包括算法结构的设置, 输入数据类型与数据集的选取, 优化过程中各优化参数的调整, 揭示了采用人工智能机器学习辅助新材料开发的一般流程.

(2) 机器学习辅助高性能有机光伏材料分子设计

有机光伏(OPV)电池利用有机光电材料将太阳能转化为电能. 功率转换效率(PCE)是人们最关心的性能. 近些年来 OPV 的研究迅速发展, 功率转换效率(PCE)已超过 17%. 传统研究方式包括光伏材料的设计和合成, 材料的光电性能表征以及光伏电池的组装和优化. 传统方法依赖于实验装置和对合成过程的精细控制, 需要大量的资源投入和较长的研究周期. 因此, OPV 的开发效率仍然有待提高. 例如, 自 1973 年首次报告以来, 仅在光伏电池中合成并测试了不到 2000 个 OPV 供体分子^[38]. 同时, 由于分析方法的落后, 这些数据的潜在价值没有得到充分利用, 对寻找新型高性能 OPV 材料时提供的指导意义有限.

孙宽等^[6]利用机器学习方法辅助设计高性能有机光伏材料. 在解决信息输入问题时, 孙宽等首先建立了一个数据库, 其中包含文献中报道的 1700 多种供体材料. 如果某种供体材料已被多次报告, 选择其中最高的 PCE 值作为原始数据. 以 PCE 为 3% 对供体材料的好坏进行定义: PCE 在 0 到 2.99% 范围内的分子被认为是“低性能”(图 12a 中以绿色表示), 而 PCE 高于 3.00% (在图 12a 中以紫色表示). 作为“高性能” OPV 分子. 为了获得对 PCE 影响最大的输入类型, 对不同类型的分子表征方式分别进行测试, 包括图像, ASCII 字符串, 两种类型的描述符和七种类型的分子指纹. 发现长度超过 1000 位的分子指纹可以获得较高的预测精度, 准确度为

81.76%, 且分子指纹长度与预测精度呈正相关.

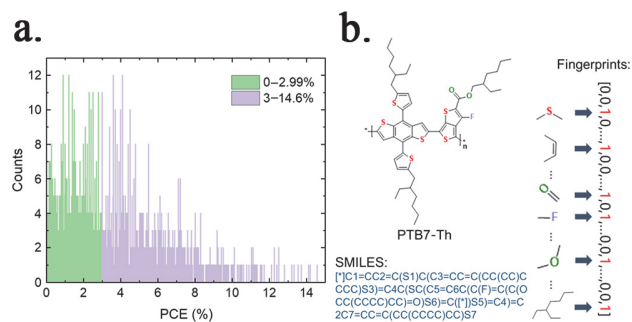


图 12 OPV 供体材料数据库的信息. (a) 数据库中 1719 个分子 PCE 值分布. (b) 分子表达示意图^[6]

Figure 12 Information about our database of OPV donor materials. (a) Distribution of PCE values of the 1719 molecules in our database. (b) Schematics of expressions of a molecule, including image, simplified molecular-input line-entry system (SMILES), and fingerprints.^[6] Reprinted with permission from ref. [6]. Copyright © 2019 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science.

在数据处理阶段, 分别将约 90% (1549 个分子) 和 10% (170 个分子) 的数据分为独立的训练集和测试集, 这两个子集彼此独立. 采用五种类型的有监督机器学习算法, 并对比它们的预测准确度(图 13a). 初期深度学习模型的预测精度令人不够满意, 这是由于深度学习模型需要大量的数据作为训练集. 当数据库中的分子数量扩大到 50,000 个时, 深度学习模型的准确性最终达到了 91.02%^[39].

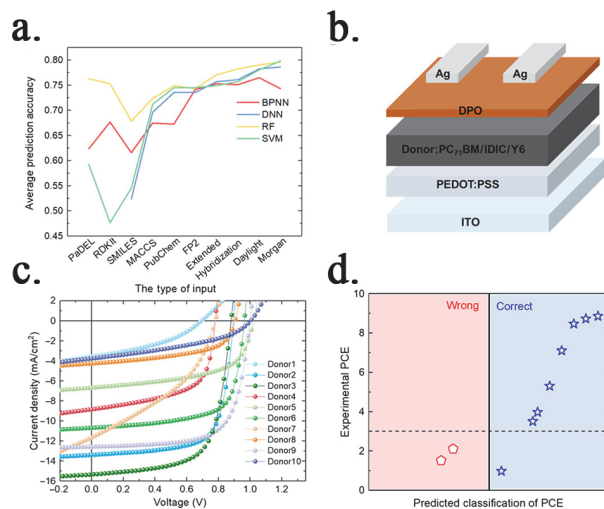


图 13 通过实验验证机器学习模型. (a) 四种不同模型分析结果. (b) OPV 器件架构示意图. (c) 供体材料太阳能电池的 $J-V$ 曲线. (d) 使用 Daylight 分子指纹和 RF 算法的预测结果与实验数据的对比^[6].

Figure 13 Verification of ML models with experiment. (a) Comparison of the results from four different models. (b) Schematic diagram of the cell architecture used in this study. (c) $J-V$ curve of the solar cell with the active layer using the predicted donor material. (d) Prediction results versus experimental data for the predicted donor materials with the RF algorithm and Daylight fingerprints^[6]. Reprinted with permission from ref. [6]. Copyright © 2019 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science.

最后他们设计了 10 种新型供体材料, 其中九种是

从未被发现的供体. 这些器件在 AM1.5G 光照环境中进行了测试, OPV 器件的电流密度-电压(J - V)曲线如图 13c 中所示. 通过实验验证, 得到了模型预测与实验结果间良好的一致性(图 13d).

孙宽等^[6]为 OPV 研究提供了全新的研究方式, 在寻找具有高 PCE 特征的 OPV 材料的过程中, 在实验合成和测定光伏特性之前就可以用机器学习模型建立它们之间的关系, 找出具有目标特性的潜在化合物, 从而加快 OPV 材料的研发速度.

2.3.2 机器学习辅助逆合成分析

逆合成分析是有机化学研究的重要内容. 化学家需要依据产物的结构, 结合已知的有机反应, 将产物不断分解成多个片段, 最终从简单的原料出发构造出全部合成步骤. 逆合成分析除了专业知识, 还需要良好的直觉和创造性, 需要经过多年大量艰苦的训练才能不断提高分析能力. 由于逆合成分析原理上“分步决策”与棋类

游戏类似, Segler 等^[40]受到了 AlphaGo 的启发, 试图用人工智能方法解决逆合成分析问题.

在进行实际操作时, 逆合成分析比下围棋更加复杂, 原因有三: (1)化学家们对于好的合成路径的定义各不相同; (2)有时在合成中需要暂时增加保护基团或引导基团, 反而增加了化学复杂性, 会对人工智能产生障碍; (3)合成路线的价值取决于前体的可用性, 因此每条合成路径都要走到最后才能判断出它的价值如何.

为了解决这些问题, Segler 等^[40]将蒙特卡罗树搜索(MCTS)与三种不同人工神经网络(NN)进行了集成, 构建 3N-MCTS 的训练模型(图 14). 第一种神经网络负责筛选有价值的前体, 第二种神经网络判断反应的可行性, 第三种神经网络寻找其它合成路径. 三种神经网络对迄今为止全部近 1200 万种有机化学基本反应进行了训练. 最后, 蒙特卡罗树搜索负责判断反应是否确实能生成目标分子.

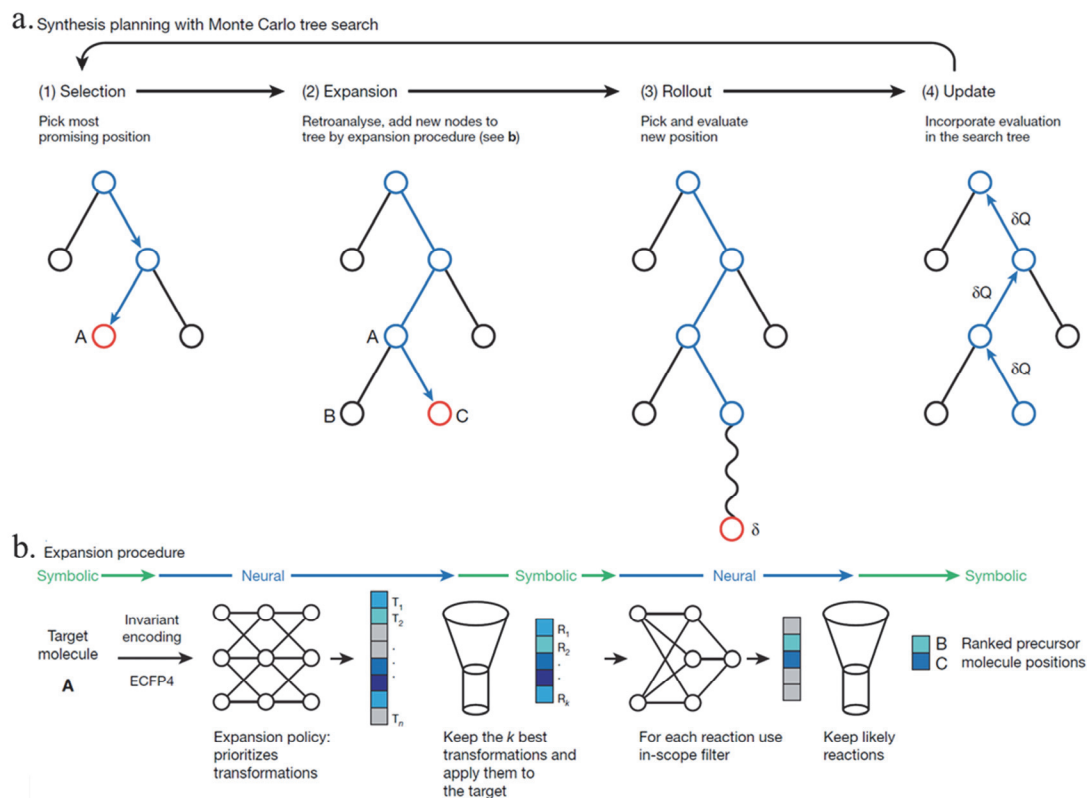


图 14 3N-MCTS 分析过程. (a) MCTS 搜索分为四个阶段: (1)选择阶段, 根据当前位置值选择最有可能的分析节点. (2)扩展阶段, 通过 b 中的扩展程序将新节点 B 和 C 添加到搜索树中. 然后选择最有希望的新位置. (3)采样阶段, 进行随机采样直到所有分子都被求解或超过特定深度. (4)更新阶段, 更新当前分支中的位置值并反馈给程序. (b) 扩展程序: 首先将要进行逆向分析分子 A 转换为指纹输入到分析网络中, 分析网络将返回所有可能的转换(T_1 到 T_n)的概率分布. 筛选出 k 个最有可能的转化(R_1 到 R_k)并作用于分子 A, 产生了制备 A 所必需的反应物, 从而完成了 R_1 到 R_k 的反应. 用过滤器对每个反应进行预测, 返回概率得分, 过滤掉不太可能发生的反应, 从而得出最有可能的前体^[40].

Figure 14 Synthesis planning with 3N-MCTS. (a) MCTS searches by iterating over four phases. In the selection phase (1), the most urgent node for analysis is chosen on the basis of the current position values. In phase (2) this node may be expanded by processing the molecules of the position A with the expansion procedure (b), which leads to new positions B and C, which are added to the tree. Then, the most promising new position is chosen, and a rollout phase (3) is performed by randomly sampling transformations from the rollout policy until all molecules are solved or a certain depth is exceeded. In the update phase (4), the position values are updated in the current branch to reflect the result of the rollout. (b) Expansion procedure. First, the molecule A to retroanalyse is converted to a fingerprint and fed into the policy network, which returns a probability distribution over all possible transformations (T_1 to T_n). Then, only the k most probable transformations are applied to molecule A. This yields the reactants necessary to make A, and thus complete reactions R_1 to R_k . For each reaction, the reaction prediction is performed using the in-scope filter, returning a probability score. Improbable reactions are then filtered out, which leads to the list of admissible actions and corresponding precursor positions B and C^[40].

3N-MCTS 的逆合成分析方法分为搜索、扩展、输出、训练四步。首先搜索目标分子的所有可能前体,筛选出最合理前体。第二步,将最合理的前体作为目标分子,重复上一步,找到该分子的最合理前体。接下来循环迭代,直到找到整条合成路径的起点,最后判断输出结果是否为合理解或最优解,对 3N-MCTS 进行训练。

训练后的 3N-MCTS 可以在 5 秒内找到超过 80% 分子的最优合成路径,以 5 秒为限,3N-MCTS 的解决效率是其它计算机方法的 30 倍以上。在双盲 AB 测试中,化学家认为由计算机生成的合成路径与现有的经过实验测试的合成路径一样好。

Segler 等^[40]开发了一个纯数据驱动的分析算法,不需要化学家的直觉和经验,就能帮助完成逆合成分析,设置简单、分析速度快、准确性高,是一种可以直接应用的成熟算法。然而美中不足的是,该算法无法解决天然产物的合成问题,这需要使用更强大、分析速度更慢的算法。此外 3N-MCTS 也无法区分化合物的立体化学性质,如互变异构、三维结构,仍需要进一步的改进。

这里使用的蒙特卡洛树搜索(MCTS)是一种以概率统计理论为指导的数值计算方法。其基本思想类似于穷举,是对于问题领域内所有的可能进行随机抽样,根据抽样结果会形成一个解的分布情况,随着抽样次数的不断增多,解的分布也越来越接近真实情况,得到局部最优解或近似解^[41]。程序目标是执行搜索步骤直到找到解决方案或达到最大深度,并通过人类游戏或自我游戏的方式进行训练。具体到解决实际问题,我们定义一个基本包括了所处理问题的全部可能性的域,然后随机生成一个值进行输入,并对这一输入进行计算与评估;重复进行这一过程就能得到我们所求问题的一个可接受结果(图 15)^[42]。作为一种通用的搜索技术,MCTS 经常用于解决具有大分支因子的顺序决策问题,如游戏 AI 或自动化定理证明。MCTS 求解过程不涉及问题本身领域的相关知识,是另一种应用较为广泛的“黑箱模型”。随着计算机计算性能的不断提高,这种近似于穷举法模型的弊端,即过高的算力要求与较慢的收敛速度也被克服,应用范围越来越广。

2.3.3 筛选目标化合物

基质金属蛋白酶-13(MMP-13)对预防和治疗骨关节炎(OA)有着非常显著的效果,通过抑制剂来阻断 MMP-13 的活性可能会对治疗 OA 疾病产生有利的影响,但是常见的宽谱抑制剂同样会抑制 MMP 家族的其它成员,特别是会导致肌与骨综合症的 MMP-1,因此寻找高效选择性抑制剂的工作显得尤为重要。

设计选择性 MMP 抑制剂相当困难,因为 MMP 各种类别的三维晶体结构有着很高的相似性,人工寻找潜在化合物费时费力,机器学习在预测结构不同的化合物的药效动力学、药代动力学和毒物学属性方面均显示除了非常好的效果^[43-46]。来自四川大学李秉轲等人使用分子描述符和机器学习方法预测和筛选 MMP-13 对 MMP-1 的选择性抑制剂^[47]。分别使用支持向量机和随机森林两种机器学习方法,将 189 个依赖于分子 3D 结构的分子描述符作为输入数据,建立机器学习模型并训练优化。选择预测能力最优的模型,对含有 458327 个分子的 ZINC 数据库中的“fragment-like”子集进行筛选,以获得潜在的 MMP-13 对 MMP-1 选择性抑制剂。并找到了 32 个对 MMP-13 选择性抑制剂和非抑制剂预测最相关的描述符。其中排名最高的描述符是 S(55)(S 原子类型电拓扑态之和)和 n_{sulph} (S 原子的数目)。除此之外,两种模型共同认可的描述符还有 $Q_{\text{N,Max}}$ (N 原子上最大的正电荷)、 $Q_{\text{O,Max}}$ (O 原子上最大的正电荷)、 $Q_{\text{H,Max}}$ (H 原子上最大的正电荷),由此可以认为电性参数对模型中选择抑制剂的活性起比较重要的作用。另外,疏水性相关描述符(Shpl(亲水区域)、S(22)($>\text{CH}$ -原子类型电拓扑态之和))与文献报道的 MMP-13 抑制剂活性与疏水性之间的关系相吻合^[48]。

2.3.4 开发新型材料

在开发新型材料时,潜在化合物的搜索空间巨大。随着化学复杂性的增加,可能的化合物组合有很多,从这个巨大的搜索空间中寻找符合要求的化合物的实验成本很高。以往的研究方式更依赖化学家的直觉,并时常伴随着错误。例如在寻找热滞后系数最低的记忆合金^[49]时,化学组成和微观结构自由度的相互作用产生

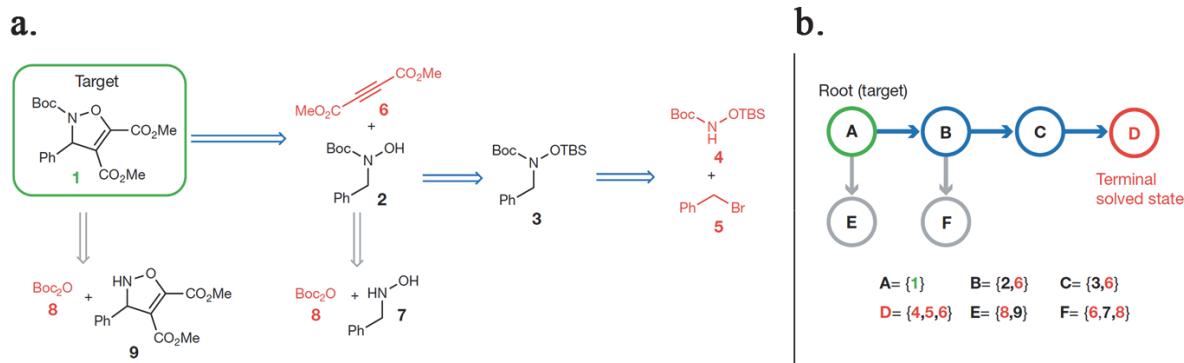


图 15 逆合成路径的化学表示与对应的搜索树结构^[40]

Figure 15 The chemical representation of the reverse synthetic pathway and the corresponding search tree structure^[40]

了巨大的复杂性. 在处理设计缺陷、固体溶液、非化学计量和多组分化合物的问题时, 研究者要面对成千上万种化合物, 对其一一合成显然是不现实的. 而一般的回归方法存在可用数据少, 所研究问题的模型大, 预测不准确, 对未开发的化学空间无法进行有效探索, 缺少实验反馈的缺点. 薛德祯等^[49]研究 $\text{Ni}_{50-x-y-z}\text{Ti}_{50}\text{Cu}_x\text{Fe}_y\text{Pd}_z$ 合金族群中热滞后系数最低的潜在化合物, 在范围 $50-x-y-z \geq 30\%$, $x \leq 20\%$, $y \leq 5\%$, $z \leq 20\%$ 内, 以 1% 为步长变化 x , y , z , 共得到近 80 万(797504)种化合物. 他们设计了一个利用机器学习方法与实验相结合进行研发的闭环路线. 选择合适间隔, 对一部分(22 种)化合物进行合成, 用它们的结构-性能数据作为训练数据集, 用其训练机器学习模型. 而后应用训练好的模型在可能的化合物空间中进行搜寻, 寻找合适的潜在材料, 并经过全局优化得到最好的新的实验方案. 再通过实验得到新的结果, 进一步优化机器学习模型. 不断循环这一过程, 材料和机器学习模型在这一循环中不断得到优化(图 16). 利用这一优化循环, 薛德祯等^[49]从 80 万种化合物中找到了 14 种热滞后系数最低的化合物. 展现了机器学习在新材料开发中的巨大潜力.

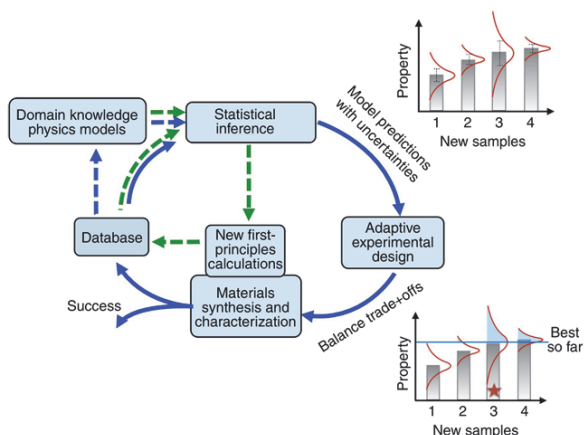


图 16 Dezhen Xue 等人设计的闭环研究路线. 用来自先前实验和物理模型的数据以及相关特征来描述材料. 将这部分信息输入机器学习框架进行预测, 使用全局优化方法优化机器学习模型. 根据模型预测结果进行新的合成表征实验, 具有改进模型和发现新材料的双重目的. 以上结果用来更新数据库, 更新后的数据库为下一次循环提供输入. 绿色箭头表示通过实验或理论计算方式的传统研究方法^[49].

Figure 16 Closed research route designed by Dezhen Xue et al. Describe materials with data from previous experiments and physical models and related features. This part of the information is input into the machine learning framework for prediction, and the global optimization method is used to optimize the machine learning model. Carrying out new synthetic characterization experiments based on the model prediction results has the dual purpose of improving the model and discovering new materials. The above results are used to update the database. The updated database provides input for the next cycle^[49].

2.3.5 寻找新反应

寻找新的化学反应是一件十分耗时费力的工作, 一个可行的方向是对反应性进行预测.

Granda 等^[50]将“化学智能”与“人工智能”相结合,

由专家训练机器学习模型, 控制实际反应系统, 探索化学反应空间, 寻找新的化学反应. 他们使用液体处理机器人进行化学反应, 通过实时传感器记录数据, 用 IR-NMR 进行自动反应分析. 将分析结果输入一个以支持向量机作为算法的模型中进行分类, 得到反应能否发生的结论, 再更新反应数据库, 优化机器学习模型(图 17).

为了研究这一自动化系统的学习能力, Granda 等^[50]做了进一步实验, 测试该模型在已有的 Suzuki-Miyaura 化学反应空间上的学习能力. 最终发现随机对 Suzuki-Miyaura 化学反应空间中 10% 的化学反应进行训练, 模型就能预测其余 90% 的反应. Jaroslaw 将传统自动化思想与人工智能相结合, 印证了机器学习方法的适用性和强大之处.

这一研究中所采用的支持向量机(SVM)是一种基于结构风险最小化(SRM)的分类机器学习方法^[51]. 其原理是将所给数据映射到一个高维空间中, 找到一个符合要求的超平面, 使其距离两组数据的最小距离最大, 从而对所给的数据起到分类的效果^[52](图 18). SVM 结构简单、性能优良, 模型泛化程度高, 有着坚实的理论支持, 善于对高维数据进行处理. 而其的缺点在于对数据完整性要求高, 对缺失数据敏感. 目前支持向量机已经广泛用于人脸识别, 语音识别, 手写体识别, 图像过滤处理等方面^[53]. 化学家常用 SVM 筛选目标化合物和目标反应.

决策树是一种树形结构算法, 用于解决多分类和决策问题. 当问题不再是简单的二分类或者二选一决策, 而是多分类或者多个选择问题时, 可以将其分解成多个独立的二分类问题, 在每个独立问题上应用某种分类方法, 如 SVM 方法, 这样决策树可看作多个 SVM 的集成^[54]. 根据问题本身重要程度的不同, 调整每一个子 SVM 的级别^[55,56], 可以对变量进行分类和解决简单的逻辑问题(图 19). 其优点在于易于理解和实现, 存在明确的层次结构, 对于给定的一个决策树模型, 有清晰的逻辑表达式. 决策树的缺点同样在于对缺失数据处理困难. 除此之外, 决策树还存在可能出现过拟合、不支持在线学习等缺点.

2.3.6 机器学习辅助理论计算化学

前面曾提到, 理论化学计算可以依据物理化学原理提供对化学体系的精确描述. 然而由于体系的复杂性导致理论化学计算的计算量异常巨大. 因而低成本的, 而又不失精度的计算就成为另一种应对复杂体系的选择. 机器学习通过已有的数据进行学习, 建立表观联系, 从而可以以极低的成本得到比较精确的预测结果. 用机器学习辅助理论化学计算因而成为机器学习深入化学研究的另一重要领域^[57,58].

将机器学习方法运用到理论化学计算中主要有两种方法. 第一种是利用比较可靠的理论化学计算算出化

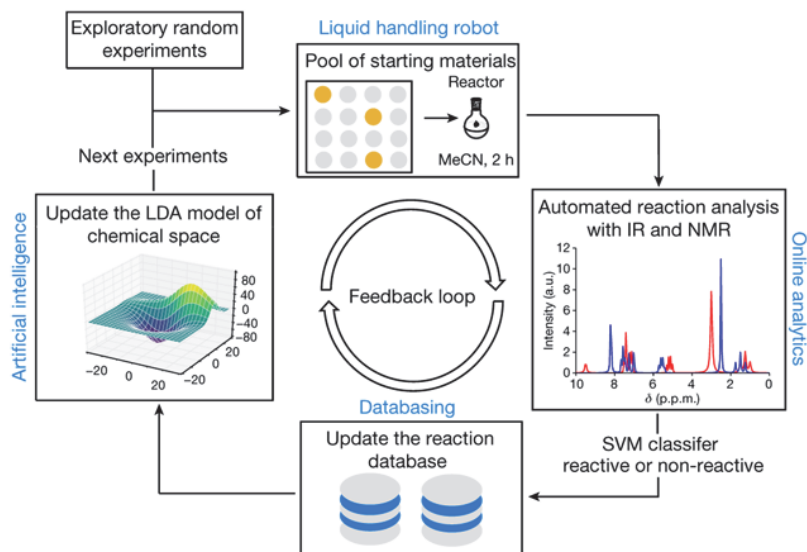


图 17 液体处理机器人与人工智能结合探索化学空间的实验过程. 液体处理机器人通过从原料库中选择反应物来执行反应. 在线分析用于将反应结果实时解释为反应性或非反应性, 反应数据库存储反应结果. 机器学习用于建立化学空间模型, 推荐下一个实验并控制机器人^[49].

Figure 17 Experimental process of combining liquid handling robot and artificial intelligence to explore chemical space. The liquid-handling robot performs reactions by choosing reactants from the pool of starting materials. Online analytics is used for real-time interpretation of reaction outcomes as reactive or non-reactive, and the reaction database stores reaction outcomes. Machine learning is used to build a model of the chemical space, recommend the next experiments and control the robot^[49].

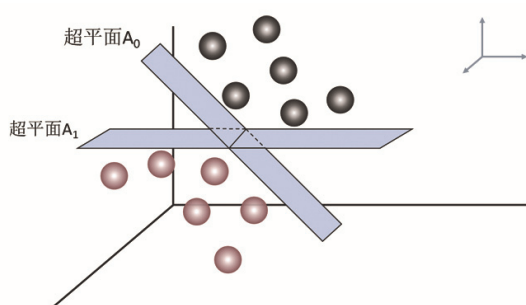


图 18 所给数据线性可分时 SVM 对数据分类示意图. 超平面 A_0 距离两组点的最小距离 $>$ 超平面 A_1 距离两组点的最小距离, 故超平面 P_0 是分类功能更好的超平面.

Figure 18 Schematic diagram of data classification by SVM when the given data is linearly separable. The minimum distance of hyperplane A_0 from two groups of points $>$ the minimum distance of hyperplane A_1 from two groups of points, so hyperplane P_0 is a hyperplane with better classification function.

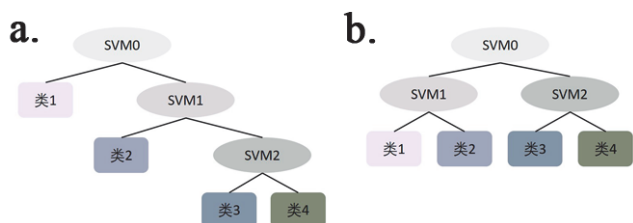


图 19 基于 SVM 的决策树结构

Figure 19 Structure of decision tree which based on SVM

合物的性质, 将化学结构和性质作为训练集输入机器学习模型. 提供足够多的训练数据, 机器学习模型就可以得到比较好的训练, 提供好的预测能力. 这一方法的优点是借助比较成熟的理论计算方法得到的数据通常可

靠性和稳定性比较好, 受实验其他因素的影响小. 通过高性能计算得到优质的少部分数据, 即可训练出好的模型, 预测大量的未知化学体系. 而机器学习模型预测所需的计算成本比理论计算小得多, 大大提高了预测效率. 一个典型的例子是江俊等^[59]利用神经网络, 模拟了肽键结构与光谱性质之间的构效关系.

紫外吸收被广泛用于表征蛋白质结构, 但是对蛋白结构的紫外光谱预测依赖于复杂的理论计算, 计算成本高、速度慢. 江俊等^[59]以 *N*-甲基乙酰胺(NMA)作为蛋白多肽的模型化合物. 首先利用分子动力学模拟, 在 200 K, 300 K 和 400 K 下产生出 70000 个 NMA 的构象. 而后利用含时密度泛函理论 TD-DFT 计算算出了这些构象的激发态性质, 包括跃迁能和跃迁偶极矩. 再以这 70000 个构象的 NMA 的基态几何参数和电荷信息作为描述符, 采用神经网络对相应的基态能、跃迁能和跃迁偶极矩进行学习. 经过这些数据的训练, 机器学习模型可以很好的预测在不同温度下不同构象分子的基态和激发态性质, 与 TD-DFT 计算结果相一致. NN 模型得到紫外光谱性质也与 TD-DFT 结果一致. 这一例子表明, 机器学习提供了一种经济高效的工具来模拟蛋白质的光学特性, 与量子化学计算相比, 神经网络模拟速度快了近 3000 倍(图 20).

另一机器学习运用于理论化学计算的方式是将机器学习嵌入理论计算的方法本身, 以代替理论计算中某些复杂、计算量大而耗时的步骤, 提高计算效率^[60,61]. 例如, Schütt 等^[58]提出了一个深度学习模型, 用于预测分子轨道的波函数, 再进一步从波函数计算出所有其它基态特性. 他们以组成分子的原子轨道为基组, 以成熟

以为图像中的不同对象分配重要性(可学习的权重和偏差), 并且还可以区分这些对象的深度学习算法. 卷积神经网络在化学上可用来将分子化学结构式直接作为机器学习模型的输入对象, 每个卷积层提取并保留分子信息, 再将这部分分子信息进行计算得到我们想要的结果^[76].

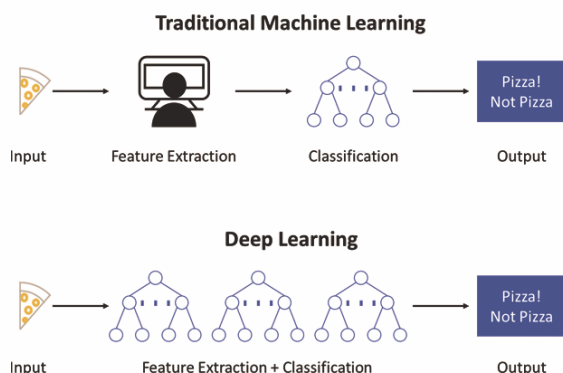


图 22 传统机器学习与深度学习的区别^[75]

Figure 22 The difference between traditional machine learning and deep learning^[75]

随着深度学习领域的发展成熟, 仅凭简单基础的原子层面信息就可以搭建优秀的 QSAR(定量分子构效关系)^[77]分析模型^[78]. 2012 年 Merck 公司举办的 QSAR 预测比赛中, 基于“深度神经网络(DNN)”的机器学习方法大幅(约 15%)超过了 Merck 公司内部的基线模型, 一举夺冠. 其重要特点就在于深度神经网络 DNN 算法不需要任何人为的特征构建, 完全基于数据来自动生成. 在其多层网络算法中, 每一层都会对分子结构进行不断抽象和特征提取, 从而不断提高表现能力^[79]. Hochreiter 等^[80]在 2016 年开发化合物毒性评估模型 DeepTox; Duvenaud 等^[81]将 SMILES(分子名称)直接与分子性质对映, 构建 QSAR 模型, 整个过程没有进行任何人工干预, 体现了高度智能化的思想. 深度学习有着优秀性能和广阔前景, 值得研究者进一步发掘探索.

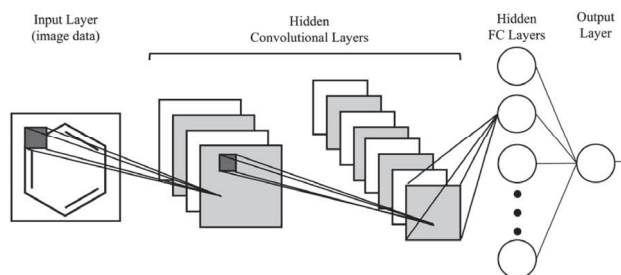


图 23 卷积神经网络(CNN)示意图. CNN 要求输入是图像数据形式. 每个卷积层提取并保留空间信息, 并学习表示形式, 然后将该表示形式通常传递到输出层之前的传统的完全连接的前馈神经网络^[76].

Figure 23 Schematic diagram of convolutional neural network (CNN). CNN requires the input to be in the form of image data. Each convolutional layer extracts and retains spatial information, and learns a representation, which is then passed to the traditional fully connected feedforward neural network before the output layer^[76].

2.5 使用机器学习方法时应注意的问题

尽管机器学习已经在化学研究中初步展示出强大的助力, 但是由于机器学习方法本身仍然存在的相当的局限, 其应用仍然要非常谨慎. 例如在 20 世纪 60 年代, 美国军方开发了一种能够从图片中识别坦克的机器学习模型, 但研究者发现, 程序只能识别晨光中或云朵下的坦克, 而类似于夜晚中的坦克无法被识别. 这里程序识别的实际上是“晨光”或“云朵”, 而非研究者期望的“坦克”^[82]. 在机器学习领域中, 这被称作假阳性^[83](False Positive)结果, 是指算法预测结果正确, 但没有挖掘出问题本身的核心驱动因素. 由于当代机器学习从数学原理上仍然是一种数理统计方法, 它只寻找数据之间的联系, 而不揭示原理, 因而不恰当的运用机器学习很容易导致错误的结果.

来自 Google 的 Riley^[84]提到了人们在应用机器学习时经常犯三类错误:

(1)对数据的不当处理. 如一味追求数据总量, 造成实验资源的浪费; 对训练集和测试集数据分割不当等.

(2)隐变量. 实验中必然会遇到不可控变量, 如仪器的老化, 人为因素带来的误差等等, 这些不可控变量是模型的干扰因素^[85]. 面对此类问题, 可以考虑使用多个机器学习模型相结合, 一个集中于解决问题本身, 其它则用来排除实验中的混杂因子, 如果后者的效果非常好, 则需要对实验数据进行标准化处理.

(3)定错训练目标. 研究者倾向于用一小部分函数来描述其面对的问题, 这可能无法捕捉到真正驱动问题的因素.

这些常见错误也会是将机器学习运用到化学研究中所要面临的. 不恰当的数据采集处理和筛选以及不恰当的算法的应用会导致机器学习无法发现化学体系的核心驱动因素, 造成错误判断. 例如, 机器学习需要大量的实验数据进行模型训练, 特别是面对复杂的体系时, 而只有输入和输出数据是稳定可靠的才能得到正确的相关关系. 然而当前很多化学实验的人为因素仍然很强, 因此对于化学实验数据和化学信息的筛选就是运用机器学习方法的一个重要挑战. 而对大量稳定标准实验的要求也催生了机器学习与传统自动化技术的结合, 将实验通过标准程序交给机器完成, 从而提供大量稳定可重复的数据. 如前面提到的 Granda 等^[50]开发的自动系统, 使用机器人自动进行化学反应, 对实验数据的分析预测也交给机器完成. 这也是未来化学研发的一个重要发展趋势.

此外机器学习所需要的大量数据也面临成本问题. 材料化学实验数据的获取耗时很长、成本高、数据库建立困难. 这些都影响了机器学习的应用. 应对数据匮乏的思路主要有两种, 一种是通过理论计算, 增加数据量. 我们前文已经提及, 利用已经发展比较成熟、对实验结果重复性比较好的理论化学计算方法, 结合高性能计算

机进行高通量计算得到大量学习数据。例如王金龙等^[86]基于第一性原理高通量计算数据库,搭建了设计二维铁磁功能材料的智能化平台,成功从近 4000 种二维材料中,快速筛选出了近百种未研究过的二维本征铁磁半导体、半金属和金属。很好的克服了二维磁性材料数据的匮乏,并提出了一种针对复杂材料性质的通用描述符。另一方法是发展具有可迁移性的机器学习算法。前文已经指出,机器学习方法不同于理论计算在于它只寻找数据间的表现联系,而不涉及其本质。因而通常机器学习只能在原有数据所覆盖的范围内学习,也只能在这一范围内进行预测和决策,而很难运用在其他情形中,这就是机器学习模型的可迁移问题。解决这一问题的一个例子是迁移学习。迁移学习^[87]依赖这样一个基本概念:即不同种类的性能(如物理性能、化学性能、电性能,热力学性能和机械性能)在物理上相互关联。对于某项待预测但缺少训练数据的性能,可以用其他数据充足的性能进行预训练,得到相应的多个神经网络模型。然后以目标性质为导向,以预训练得到的神经网络模型隐藏层中的相关参数为基础,进行再次训练。如此训练出的机器学习模型有一定程度的“举一反三”能力,即使在目标性质数据量非常少的情况下,仍然可以做出出色的预测。例如 Yamada 等^[88]构建了一个具有巨大潜力的神经迁移学习(neural transfer learning)框架,克服了材料数据量有限的问题。为了验证该迁移学习模型的多方面应用,Yamada 等分别用其预测了聚合物的热容量、有机聚合物的热传导性、无机晶体的热传导率,无机和有机材料间的迁移性,都得到了很好的结果。

最后需要指出的是化学家的知识和经验是不能被人工智能所代替的^[89],人工智能只是辅助化学家进行研究的一个工具,无论是在建立模型时需要的专业知识,还是分析其给出的结果来进行进一步的分析研究,都离不开化学家的参与。人工智能只是在数据分析整理中起到辅助作用,降低了数据分析的劳动强度。从发现新知识、新规律角度看,人工智能可以将大量的化学信息简化,降低其繁琐性,突出其中的重要因素,有助于化学家抓住关键因素进行规律的探索和知识的发现。从开发新功能新材料角度,人工智能有助于在海量化学品中筛选合适的化学品和实验条件。大大减少了繁琐的人工筛选过程。

3 总结与展望

人工智能应用于化学的研究在全世界,特别是发达国家,正处于迅猛发展时期。已经有越来越多的化学和材料学家开始运用人工智能工具解决化学问题,而众多数学、计算机、物理学家也在尝试开发新的机器学习算法应用于化学体系。人工智能可能带来的具有革命性的变化的力量已经吸引了国家和企业的重点关注和大量的投资。人工智能已经成为各国争夺未来化学研究的制

高点的重要组成部分。例如美国联邦政府支持的材料基因组计划(Materials Genome Initiative MGI)^[90]利用了机器学习人工智能加速新材料的开发^[91]。美国国家科学基金 NSF 化学部则于 2017 年启动了数据驱动的化学科学发现计划(Data-Driven Discovery Science in Chemistry (D3SC) initiative),支持了 30 个基于大数据、现代数据科学及机器学习方法的复杂化学体系的基础研究^[92]。欧盟的地平线(Horizon 2020)计划则支持了化学大数据的项目^[93],用于机器学习等方法挖掘化学研究中的数据。

我国在机器学习运用于化学方面的研究起步较晚,特别是化学信息学和现代数学和计算机技术运用于化学研究的传统相对较弱,与国外相比差距较大;不过近两年也取得了快速的发展,如我们前面所介绍的部分出色的成果。同时国家也开始增加投入支持人工智能在化学领域的研究。例如,从公开的国家自然科学基金委员会资助的项目情况看,2010 年到 2019 年基金委化学学部共资助了 10 个机器学习项目,其中 2019 年就有 7 个。这里包括通过机器学习模拟高精度分子;基于机器学习预测纯组分炼化性质;机器学习调控酶催化剂的立体选择性与多手性定向合成;机器学习设计重要能源转化过程中的催化剂;模拟研究多晶型预测模型;预测分子热力学和输运性质等。从交流情况看,从事相关研究的国内队伍也在不断壮大,交流不断增多。2018 年 6 月,华东师范大学-纽约大学计算化学联合研究中心(上海纽约大学)主办了以“分子科学中的机器学习方法及应用”为主题的国际暑期学校。介绍了机器学习方法技术基础及其在化学、生物和材料科学领域应用。2019 年 9 月厦门大学化学化工学院也举办了一次机器学习-计算化学的 Workshop,邀请了国内外多名学者介绍了对机器学习在分子动力学模拟和理论计算,晶体性质预测等方面的成果。

可以预见的是人工智能在化学领域的运用会越来越多、越来越广、越来越深入。有望成为继科学实验,理论计算,计算机模拟以外另一科学研究范式,为现代化学的发展注入全新的强大的力量。

作者简介



朱博阳. 主要研究方向: 人工智能, 程序后端开发。



吴睿龙. 2017 年考入天津大学, 现为理学院化学系大三学生. 目前在于曦老师指导下从事机器学习用于化学体系的大学生创新项目.



于曦. 天津大学理学院化学系、物理系研究员、教授. 主要研究方向为微观体系电荷的量子输运、分子电子学和微纳光电器件, 以及人工智能辅助的有机光电材料开发.

References

- [1] Tang, Z. T.; Shao, K.; Zhao, D. B.; Zhu, Y. H. *Control Theory & Applications* **2017**, 034, 1529 (in Chinese). (唐振韬, 邵坤, 赵冬斌, 朱圆恒. 控制理论与应用, **2017**, 034, 1529.)
- [2] McKinney, S. M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafi, H.; Back, T.; Chesus, M.; Corrado, G. C.; Darzi, A.; Etemadi, M.; Garcia-Vicente F.; Gilbert, F. J.; Halling-Brown, M.; Hassabis, D.; Jansen, S.; Karthikesalingam, A.; Kelly, C. J.; King, D.; Ledsam, J.R.; Melnick, D.; Mostofi, H.; Peng, L.; Reicher, J. J.; Romera-Paredes, B.; Sridharan, R.; Suleyman, M.; Tse, D.; Young, K. C.; De, Fauw, J.; Shetty, S. *Nature* **2020**, 577, 7788.
- [3] Li, J. G.; Gao, Z. K. *Acta Biophysica Sinica* **2009**, 25, 51 (in Chinese). (李建更, 高志坤, 生物物理学报, **2009**, 25, 51.)
- [4] Leon, F.; Lisa, C.; Curteanu, S. *Mol. Cryst. Liq. Cryst.* **2010**, 518, 1542.
- [5] Wang, J.S.; Li, Z.; Yan, S.C.; Yue, X.; Ma, Y.Q.; Ma, L. *RSC Adv.* **2019**, 9, 14797.
- [6] Sun, W.B.; Zheng, Y.J.; Yang, K.; Zhang, Q.; Shan, Akeel A.; Wu, Z.; Sun, Y.Y.; Feng, L.; Chen, D.Y.; Lu, S.R.; Li, Y.; Sun, K. *Sci. Adv.* **2019**, 5, 4275.
- [7] Zhong, M.; Tran, K.; Min, Y. M.; Wang, C. H.; Wang, Z. Y.; Ding, C. T.; Luna, P.; Sedighian Rasouli, A.; Brodersen, P.; Sun, S.; Voznyy, O.; Tan, C. S.; Askerka, M.; Che, F. L.; Liu, M.; Seifitokaldani, A.; Pang, Y. J.; Lo, S. C.; Sargent, E. *Nature* **2020**, 581, 178.
- [8] Wu, W.; Sun, Q. *Scientia Sinica Physica, Mechanica & Astronomica* **2018**, 48, 54 (in Chinese). (吴炜, 孙强, 中国科学: 物理学 力学 天文学, **2018**, 48, 54.)
- [9] Saunders, C.; Stitson, M. O.; Weston, J.; Holloway, R.; Bottou, L.; Scholkopf, B. *Comput. Sci.* **2002**, 1, 1.
- [10] Safavian, S. R.; Landgrebe, D. *IEEE Trans. Syst., Man, Cybern.* **1991**, 21, 660.
- [11] Hagan, M. T.; Demuth, H. B.; Beale, M. H. *Neural Network Design*, China Machine Press, Beijing, **2002**.
- [12] Browne, C. B.; Powley, E.; Whitehouse, D.; Lucas, S. M.; Cowling, P. I. *IEEE Transactions on Computational Intelligence & Ai in Games*, **2012**, 4, 1.
- [13] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatic — Second, Revised and Enlarged Edition*, Volume I: Alphabetical Listing; Volume II: Appendices, Bibliography, **2009**.
- [14] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, WILEY-VCH, Weinheim, Germany, **2000**.
- [15] He, B.; Luo, Y.; Li, B. K.; Xue, Y.; Yu, L. T.; Qiu, X. L.; Yang, D. G. *Acta Physico-Chimica Sinica* **2015**, 09, 1795 (in Chinese). (何冰, 罗勇, 李秉轲, 薛英, 余洛汀, 邱小龙, 杨登贵, 物理化学学报, **2015**, 09, 1795.)
- [16] Wang, J. X.; Li, Y.; Yang, M.; Wang, Q. H.; Deng, G. W.; Yang, F.; Li, B. K. *Chemical Research & Application* **2019**, 031, 1313 (in Chinese). (王洁雪, 李瑶, 杨敏, 王琪慧, 邓国伟, 杨帆, 李秉轲. 化学研究与应用, **2019**, 031, 1313.)
- [17] Wang, L.; Mao, H. T.; Zhang, L.; Liu, L. L.; Du, J. *CIESC J.* **2019**, 70, 4722 (in Chinese). (王璐, 毛海涛, 张磊, 刘琳琳, 都健, 化工学报, **2019**, 70, 4722.)
- [18] Dai, Y.; Niu, L.; Zou, J.; Liu, D. Y.; Liu, H. *J. Cent. South Univ.* **2018**, 25, 1535.
- [19] Ul-Haq, Z.; Ashraf, S.; Al Majid, A.; Barakat, A. *Int. J. Mol. Sci.* **2016**, 17, 657.
- [20] Xu, Y. J.; Pei, J. F. *Big Data Research* **2017**, 003, 45 (in Chinese). (徐优俊, 裴剑锋, 大数据, **2017**, 003, 45.)
- [21] What is the molecular descriptor(in Chinese)? 什么是分子描述符? <https://zhuanlan.zhihu.com/p/113381716>.
- [22] Mauri, A.; Consonni, V.; Todeschini, R. *Molecular Descriptors*, Vol. 8, Eds.: Puzyn, T.; Leszczynski, J.; Cronin, M. T. D., Springer, New York, **2009**, p. 33.
- [23] Mauri, A.; Consonni, V.; Todeschini, R. *Molecular Descriptors*, Vol. 8, Eds.: Puzyn, T.; Leszczynski, J.; Cronin, M. T. D., Springer, New York, **2009**, p. 34.
- [24] Ren, W.; Kong, D. X. *Computers & Applied Chemistry*, **2009**, 11, 1455 (in Chinese). (任伟, 孔德信. 计算机与应用化学, **2009**, 11, 1455.)
- [25] Dickert, F. L.; Hayden, O. *Adv. Mater.* **2000**, 12, 311.
- [26] DRAGON <http://www.taletale.mi.it/>
- [27] GRID http://www.moldiscovery.com/soft_grid.php
- [28] MOLE db http://michem.disat.unimib.it/mole_db/
- [29] Stein, H. S.; Gregoire, J.M. *Chem. Sci.* **2019**, 10, 9640.
- [30] Mater, A. C.; Coote, M. L. *J. Chem. Inf. Model.* **2019**, 59, 2545.
- [31] Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. *Nat. Commun.* **2017**, 8, 15679.
- [32] Cova, Tânia F. G. G.; Pais, Alberto A. C. C. *Front. Chem.* **2019**, 7, 809.
- [33] Jordan, M. I.; Mitchell, T. M. *Science* **2015**, 349, 6245.
- [34] McCulloch, W. S.; Pitts, W. *Bull. Math. Biol.* **1943**, 52.
- [35] Gall, J.; Razavi, N.; Van Gool, L. *An Introduction to Random Forests for Multi-class Object Detection*, Springer-Verlag, Heidelberg, Germany, **2012**, pp. 243-263.
- [36] Lim, A.; Breiman, L.; Cutler, A. *Computer Science* **2014** (data package and software).
- [37] Ahneman, D. T.; Estrada, J. G.; Lin, S. S.; Dreher, S. D.; Doyle, A. G. *Science* **2018**, 360, 6385.
- [38] Ghosh, A. K.; Feng, T. *J. Appl. Phys.* **1973**, 44, 2781.
- [39] Sun, W.; Li, M.; Li, Y.; Wu, Z.; Sun, Y.; Lu, S.; Xiao, Z.; Zhao, B.; Sun, K. *Adv. Theor. Simul.* **2019**, 2, 1800116.
- [40] Segler, M. H. S.; Preuss, M.; Waller, M. P. *Nature* **2018**, 555, 7698.
- [41] Yu, Y. B. *M.S. Thesis*, Dalian Maritime University, Dalian, **2015** (in Chinese). (于永波, 硕士论文, 大连海事大学, 大连, **2015**.)
- [42] Fu, M. C. In *2016 Winter Simulation Conference*, Arlington Virginia, **2016**, pp. 659-670.
- [43] Xue, Y.; Li, H.; Ung, C. Y.; Yap, C. W.; Chen, Y. Z. *Chem. Res. Toxicol.* **2006**, 19, 1030.
- [44] Lü, W. J.; Chen, Y. L.; Ma, W. P.; Zhang, X. Y.; Luan, F.; Liu, M. C.; Chen, X. G.; Hu, Z. D. *Eur. J. Med. Chem.* **2008**, 43, 569.
- [45] Lü, W.; Xue, Y. *Acta Phys.-Chim. Sin.* **2010**, 26, 471.
- [46] Li, B. K.; Yong, C.; Yang, X. G.; Xue, Y.; Chen, Y. Z. *Comput. Biol. Med.* **2013**, 43, 395.
- [47] Li, B. K.; Cong Y.; Tian, Z. Y.; Xue, Y. *Acta Physico-Chimica Sinica* **2014**, 30, 171 (in Chinese). (李秉轲, 丛湧, 田之悦, 薛英, 物理化学学报, **2014**, 30, 171.)
- [48] Barta, T. E.; Becker, D. P.; Bedell, L. J.; Crescenzo, G. A. D.; McDonald, J. J.; Mehta, P.; Munie, G. E.; Villamil, C. I. *Bioorg. Med. Chem. Lett.* **2001**, 11, 2481.
- [49] Xue, D. Z.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D. Q.; Lookman, T. *Nat. Commun.* **2016**, 7, 11241.
- [50] Granda, J. M.; Donina, L.; Dragone, V.; Long, D. L.; Cronin, L. *Nature* **2018**, 559, 7714.
- [51] Ding, S. F.; Qi, B. J.; Tan, H. Y. *Journal of University of Electronic Science and Technology of China*, **2011**, 40, 1 (in Chinese). (丁世飞, 齐丙娟, 谭红艳, 电子科技大学学报, **2011**, 40, 1.)
- [52] Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.* **1998**, 2, 121.

- [53] Qi, H. N. *Computer Engineering* **2004**, *30*, 10 (in Chinese). (祁亭年, 计算机工程, **2004**, *30*, 6.)
- [54] Zhang, X. W.; Guo, L. *Firepower & Command Control* **2010**, *35*, 31 (in Chinese). (张先武, 郭雷, 火力与指挥控制, **2010**, *35*, 31.)
- [55] Wang, J. F.; Cao, Y. D. *Journal of Beijing Institute of Technology* **2001**, *21*, 225 (in Chinese). (王建芬, 曹元大, 北京理工大学学报, **2001**, *21*, 225.)
- [56] Zhang, Q. Y.; Jie, Y.; Li, K. *Journal of Computer Applications* **2008**, *28*, 3227 (in Chinese). (张秋余, 竭洋, 李凯, 计算机应用, **2008**, *28*, 3227.)
- [57] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. *Nature* **2018**, *559*, 547.
- [58] Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K. R.; Maurer, R. J. *Nat. Commun.* **2019**, *10*, 1.
- [59] Ye, S.; Hu, W.; Li, X.; Zhang, J. X.; Zhong, K.; Zhang, G. Z.; Luo, Y.; Mukamel, S.; Jiang, J. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 11612.
- [60] Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. *Phys. Rev. Lett.* **2018**, *120*, 036002.
- [61] Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K. Preprint at <https://arxiv.org/abs/1802.08219>, **2018**.
- [62] Ryczko, K.; Strubbe, D. A.; Tamblyn, I. *Phys. Rev. A* **2019**, *100*, 022512.
- [63] Behler, J.; Parrinello, M. *Phys. Rev. Lett.* **2018**, *98*, 146401.
- [64] Braams, B. J.; Bowman, J. M. *Int. Rev. Phys. Chem.* **2009**, *28*, 577.
- [65] Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. *Phys. Rev. Lett.* **2010**, *104*, 136403.
- [66] Smith, J. S.; Isayev, O.; Roitberg, A. E. *Chem. Sci.* **2017**, *8*, 3192.
- [67] Podryabinkin, E. V.; Shapeev, A. V. *Comput. Mater. Sci.* **2017**, *140*, 171.
- [68] Podryabinkin, E. V.; Tikhonov, E. V.; Shapeev, A. V.; Oganov, A. R. *Phys. Rev. B* **2019**, *99*, 064114.
- [69] Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K. R. *Sci. Adv.* **2018**, *3*, e1603015.
- [70] Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. *Nat. Commun.* **2018**, *9*, 3887.
- [71] Gastegger, M.; Behler, J.; Marquetand, P. *Chem. Sci.* **2018**, *8*, 6924.
- [72] Dral, P. O. *J. Phys. Chem. Lett.* **2020**, *11*, 2336.
- [73] Sun, Z. J.; Xue, L.; Xu, Y. M.; Wang, Z. *Application Research of Computers* **2012**, *029*, 2806 (in Chinese). (孙志军, 薛磊, 许阳明, 王正, 计算机应用研究, **2012**, *029*, 2806.)
- [74] Liu, J. W.; Liu, Y.; Luo, X. L. *Application Research of Computers* **2014**, *031*, 1921 (in Chinese). (刘建伟, 刘媛, 罗雄麟, 计算机应用研究, **2014**, *031*, 1921.)
- [75] The difference between machine learning and deep learning. (in Chinese). (机器学习和深度学习区别). https://mp.weixin.qq.com/s/h93LO6nlAVOUmG_vmJMGaQ.
- [76] Goh, G. B.; Hodas, N. O.; Vishnu, A. *J. Comput. Chem.* **2017**, *38*, 1291.
- [77] Sun, Y. Z. *M.S. Thesis*, China Medical University, Shengyang, **2009** (in Chinese). (孙也之, 硕士论文, 中国医科大学, 沈阳, **2009**.)
- [78] Lusci, A.; Pollastri, G.; Baldi, P. *J. Chem. Inf. Model.* **2013**, *53*, 1563.
- [79] Markoff, J. *New York Times*, **2012**, *10*, pp. 1-71.
- [80] Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. *DeepTox: Front Environ. Sci. Eng.* **2016**, *3*, 80.
- [81] Duvenaud, D.; Dougal, M.; Jorge, A. I.; Rafa, G. B.; Timothy, H.; Alán, A. G.; Ryan, P. A. In *Proceedings of Advances in Neural Information Processing Systems 28*, MIT Press, Montreal, **2015**, pp. 2215-2223.
- [82] Kanal, L. N.; Randall, N. C. *Proceedings of the 1964 19th ACM National Conference*, Association for Computing Machinery, New York, NY, USA, **1964**, pp. 42.501-42.5020.
- [83] Viola, J.; Snow, D.; Jones, M. J. In *Proceedings Ninth IEEE International Conference on Computer Vision*, Springer-Verlag, Nice, **2003**, pp. 734-741.
- [84] Riley, P. *Nature* **2019**, *572*, 27.
- [85] Baltz, E. A.; Trask, E.; Binderbauer, M.; Dikovsky, M.; Gota, H.; Mendoza, R.; Platt, J. C.; Riley, P. F. *Sci. Rep.* **2017**, *7*, 6425.
- [86] Lu, S.; Zhou, Q.; Guo, Y.; Zhang, Y.; Wu, Y.; Wang, J. *Adv. Mater.* **2020**, *32*, 2002658.
- [87] Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. In *International Conference on Neural Information Processing Systems*, MIT Press, Siem Reap, **2014**, p. 32.
- [88] Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. *ACS Cent. Sci.* **2019**, *5*, 1717.
- [89] Maryasin, B.; Marquetand, P.; Maulide, N. *Angew. Chem. Int. Ed.* **2018**, *57*, 6978.
- [90] Ward, Charles. **2012**. <https://www.mgi.gov/>
- [91] de Pablo, J. J.; Jackson, N. E.; Webb, M. A.; Chen, L. Q.; Moore, J. E.; Morgan, D.; Jacobs, R.; Pollock, T.; Schlom, D. G.; Toberer, E. S.; Analytis, J.; Dabo, I.; DeLongchamp, D. M.; Fiete, G. A.; Grason, G. M.; Hautier, G.; Mo, Y.; Rajan, K.; Reed, E. J.; Zhao, J. C. *npj Comput. Mater.* **2019**, *5*, 41.
- [92] <https://www.nsf.gov/pubs/2017/nsf17036/nsf17036.pdf>
- [93] <http://bigchem.eu/>

(Cheng, B.)