



计算化学中的机器学习

申林*, 贾璐阳, 汤典东, 张战运, 彭鼎

北京师范大学化学学院, 理论及计算光化学教育部重点实验室, 北京 100875

*通讯作者, E-mail: lshen@bnu.edu.cn

收稿日期: 2022-01-30; 接受日期: 2022-04-02; 网络版发表日期: 2022-04-13

国家自然科学基金(编号: 22193041, 21903005)资助项目

摘要 机器学习带来的技术革命波及甚广, 计算化学也从中受益. 正确地将机器学习引入计算化学, 有望突破现有方法在计算精度和计算效率上的瓶颈. 本文针对计算化学中的机器学习方法及其应用, 从势能面计算、分子动力学模拟和材料基因组工程中的分子设计三个方面进行了概述.

关键词 势能面, 分子动力学, 材料基因组, 机器学习

1 引言

计算化学主要以电子结构理论和分子模拟为工具来计算和预测各类化学反应的静态特征和动态行为, 为实验化学提供了强有力的理论支撑和科学指导. 随着计算数据的积累和计算资源的丰富, 机器学习在计算化学领域也取得了一系列突破, 得到人们的广泛关注. 机器学习是一门从已有的数据中学习规律, 再利用获得的规律对未知数据进行预测的科学. 对于计算化学而言, 机器学习能够将真实化学问题中不能用现有理论模型精确解释的部分完全数值化, 对从抽象物理规律到复杂化学过程之间的潜在关系进行推断, 从而显著地拓展了计算化学的应用范围, 有望解决现有计算方法精度与效率无法兼顾的难题. 本文主要针对计算化学的三个核心问题: 如何准确高效地构建反应体系的势能面, 如何通过微观尺度的模拟预测体系的宏观性质, 如何根据结构-性能关系实现对分子特别是功能材料的理论设计, 对计算化学中的机器学习方

法及其应用进行评述.

2 机器学习在势能面计算中的应用

构建准确的势能面是计算化学的基础. 计算势能面的方法包括两类: 量子力学方法和分子力学方法. 前者以电子结构的多体波函数理论或密度泛函理论为基础, 原则上不需要经验参数, 计算精度高, 但是计算量大, 对于复杂环境中的化学体系或者长时间尺度的化学问题, 即使当前最先进的计算机也难以处理. 后者以分子力场模型为基础, 假设分子势能由原子核构型唯一确定, 利用以分子结构为自变量的势能函数及其经验参数来拟合分子的真实势能. 分子力场的计算速度快得多, 但精度受限于势能函数的形式和分子各部分对能量贡献的可转移性, 也不能处理建立力场时没有“学习”到的分子类型. 针对上述问题, 人们设计了众多以神经网络(neural network, NN)为代表的机器学习方案, 极大地提升了现有势能面计算方法的精度

引用格式: Shen L, Jia L, Tang D, Zhang ZY, Peng D. Machine learning in computational chemistry. *Sci Sin Chim*, 2022, 52: 858–868, doi: 10.1360/SSC-2022-0022

和效率, 实现了多种复杂化学体系的理论计算.

计算势能面是一个回归问题. 神经网络相当于用一个多参数的非线性函数来拟合势能面, 和传统分子力场并无本质区别. 对于包含 N 个原子的分子, 将原子 i 的坐标记作 x_i , 分子势能 V 可以表示成:

$$V = \sum_j w_j z_j + b_0 = \sum_j w_j f\left(\sum_i w_{ij} x_i + b_j\right) + b_0 \quad (1)$$

其中 x_i 、 z_j 和 V 所占的节点分别构成神经网络的输入层(input layer)、隐藏层(hidden layer)和输出层(output layer), w 和 b 是神经网络的参数, 分别称作权重和阈值, $f(x)$ 是一个非线性的激活函数, 可选用双曲正切(tanh)函数、sigmoid函数或线性修正单元(rectified linear unit, ReLU)函数(图1).

我们首先利用量子化学方法计算势能面上多个构型的势能 V^{QM} 作为训练集, 再对损失函数最小化, 即可求得神经网络所有参数的值, 即:

$$\min_{w,b} \left\{ \sum_k (V_k - V_k^{\text{QM}})^2 + \lambda \|w, b\| \right\} \quad (2)$$

其中最后一项是用于避免过拟合的正则项, k 遍历所有训练样本. 此外, 如果量子化学方法已经得到原子 i 上的受力 $\mathbf{F}_i^{\text{QM}}$, 还可考虑将该项引入损失函数, 通常可以进一步提高神经网络的拟合精度, 即:

$$\min_{w,b} \left\{ \sum_k (V_k - V_k^{\text{QM}})^2 + \frac{\lambda_1}{3N} \sum_k \sum_{i=1}^N (\mathbf{F}_{i,k} - \mathbf{F}_{i,k}^{\text{QM}})^2 + \lambda_2 \|w, b\| \right\} \quad (3)$$

早在1995年, Blank等^[1]就利用神经网络建立了CO

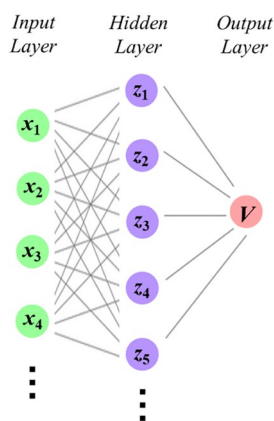


图1 神经网络的结构示意图 (网络版彩图)
Figure 1 Structure of neural network (color online).

在Ni表面以及H₂在Si表面的势能面模型. 张东辉等^[2]基于神经网络构造了一系列多原子体系的精确反应势能面, 将其成功地用于高精度的量子反应动力学模拟; 张东辉、傅碧娜等^[3]还设计了一种基于基本不变量(fundamental invariant, FI)的神经网络, 在满足置换不变性的前提下极大地减少了输入节点的数目, 提高了多原子反应体系全维势能面的计算速度. 但是, 上述神经网络只适用于个案(case-by-case). 以水溶液体系为例: 一旦水分子数目增加, 神经网络输入层的节点数也随之增加, 必须重新拟合参数; 因此, 这类神经网络不具备可转移性, 尚不能称为“力场”. 2007年, Behler和Parrinello^[4]提出了高维神经网络(high-dimensional neural network, HDNN), 将每种原子对能量的贡献用一个子网络表示(图2), 使高维势能面的构建成为可能. 仍以水溶液体系为例: 模型由多个子网络组成, 氧原子和氢原子各自对应一个拟合好的子网络; 每增加一个水分子, 就添加一个氧原子和两个氢原子的子网络, 并将新增原子的坐标信息作为输入变量, 不再重新拟合子网络的参数. 此处输入的“原子坐标信息”描述了新增原子与邻近原子之间的相互作用, HDNN称其为“广义坐标”. 将新增原子记作 i , 直观的做法是用 δ 函数表示其余原子 j 的相对位置, 即:

$$G_i(\mathbf{r}) = \sum_{j \neq i} \delta(\mathbf{r} - \mathbf{r}_j) \quad (4)$$

用高斯函数替代 δ 函数并引入截断函数 $f_{\text{cut}}(R_{ij})$, 就构造出径向对称函数(radial symmetry function):

$$G_i^{\text{rad}} = \sum_{j \neq i} e^{-\eta(R_{ij} - r_s)^2} f_{\text{cut}}(R_{ij}) \quad (5)$$

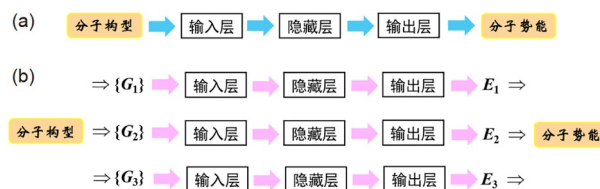


图2 传统神经网络(a)与高维神经网络(b). 以计算H₂O的势能为例: G_i ($i = 1, 2, 3$) 是原子 i 的广义坐标, 一个原子对应一个子网络, 同一元素的原子对应的子网络参数相同, E_i 是原子 i 的子网络的输出值, 分子势能是所有子网络输出的 E_i 之和 (网络版彩图)

Figure 2 Conventional neural network (a) and high-dimensional neural network (b). Take the potential energy calculation on a water system as an example. The generalized coordinates of the atom i enter the subnet yielding the contribution E_i to the total potential energy of the system. The subnets corresponding to the same element share parameters (color online).

其中 $R_{ij}=|r_i-r_j|$, r_s 是超参(hyperparameter). 该函数只包含两体相互作用, 描述多体作用需要进一步构造角向对称函数(angular symmetry function):

$$G_i^{\text{ang}} = \sum_{j,k \neq i} \left(1 + \lambda \cos \theta_{ijk}\right)^\zeta e^{-\eta(R_{ij}^2 + R_{jk}^2 + R_{ki}^2)} f_{\text{cut}}(R_{ij}) f_{\text{cut}}(R_{jk}) f_{\text{cut}}(R_{ki}) \quad (6)$$

其中 η , ζ 和 λ 都是超参. 在HDNN中, 对每个原子都可以构造相应的径向和角向对称函数, 并输入该原子的子网络. 在实际应用中往往根据体系的复杂度预先设定不同的超参, 从而构造出多个对称函数, 增加神经网络的输入节点数目, 提高拟合精度. 经过十余年的发展, HDNN已成为一套成熟的机器学习力场, 实现了若干复杂化学体系的高效理论计算和长时间动力学模拟, 解决了从小分子到大分子、从生物到材料等多个领域的实际问题^[5,6].

除了神经网络, 基于核函数的核岭回归(kernel ridge regression, KRR)模型也经常使用^[7]. 直观地看, 分子构型越相似, 在势能面上的位置就越接近, 势能也应当越相近. 以高斯核函数为例, 构型为 $\mathbf{R}$ 的分子势能可表示为

$$V(\mathbf{R}) = \sum_k \alpha_k \exp\left(-\frac{\|\mathbf{R}-\mathbf{R}_k\|_2^2}{2\sigma^2}\right) \quad (7)$$

其中 $\mathbf{R}_k$ 是训练集样本 k 的构型, 在实际应用中可将分子的平衡构型作为参考进行归一化处理^[8] (relative to equilibrium structure, RE), σ 是高斯核函数的宽度. α_k 可表示成矩阵形式的闭合解, 即:

$$\boldsymbol{\alpha}=(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{V}^{\text{QM}} \quad (8)$$

其中 λ 是正则项因子. 对称函数也可直接作为核函数方法的输入变量(即取代式7中的 $\mathbf{R}$ 和 $\mathbf{R}_k$). 一般而言, 神经网络的预测能力更强, 但需要反复调节和拟合大量的网络参数, 离不开充足的训练样本; 核函数方法的解释性更强, 适用于小样本的情形. 最近, Pinheiro等^[9]对各种机器学习力场进行了测试, 建议当损失函数只有能量误差项时(式2), 将50000作为样本充足与否的判断阈值; 如果引入了梯度误差项(式3), 则将1000作为阈值.

不同类型的分子表示法对机器学习的表现也有较大影响. 以原子坐标为自变量的任意函数, 只要满足一

定的约束条件(平移不变性、旋转不变性、置换不变性等), 原则上都可输入到机器学习模型, 预测分子的势能. 除了对称函数, von Lilienfeld等^[10]提出的库仑矩阵 (Coulomb matrix)和Bartók等^[11]提出的平滑重叠原子位置 (smooth overlap of atomic position, SOAP)是近年来应用比较广泛的两种分子构型表达方式. 前者将 N 个原子的三维坐标转换成一个 $N \times N$ 的矩阵, 矩阵元 M_{ij} 为:

$$M_{ij} = \begin{cases} 0.5Z_i^{2.4} & i=j \\ \frac{Z_i Z_j}{|\mathbf{r}_i - \mathbf{r}_j|} & i \neq j \end{cases} \quad (9)$$

其中 Z_i 是原子 i 的核电荷数. 后者将原子 i 与其余原子的相对位置表示为径向项和角向项的乘积, 即:

$$p_{nl}^i = \sum_{m=-l}^l \left| \sum_j e^{-\alpha(R_{ij}-R_n)^2} Y_{lm}(\theta_{ij}, \varphi_{ij}) Z_j f_{\text{cut}}(R_{ij}) \right|^2 \quad (10)$$

其中 n , l , α 和 R_n 都是根据体系特点预先设定的超参, Y_{lm} 是球谐函数.

何种分子表示法在机器学习中表现最优, 应当从原子坐标中提取什么样的特征, 是一个有待解决的重要问题. 深度学习的兴起提供了新的思路. 当前习惯上将包含三个或更多个隐藏层的神经网络称为深度神经网络(deep neural network, DNN). 与传统神经网络相比, 深度神经网络能够自动实现分层特征提取来代替人工设计. 例如对于图片分类问题, 深度学习可以从像素中提取出线条、边角等单元用于识别, 因此只需将图片像素作为输入值. 类似地, 在计算势能面时, 只需将分子的键长、键角或原子间距离等直观的结构特征作为输入变量, 就可以逐层建立起分子构型与势能的关联(图3). 除了深度神经网络, 人们还设计了卷积神经网络和循环神经网络等结构更复杂的深度学习模型, 也给势能面的构建带来了启发. 例如, Schütt等将原子间相互作用通过滤波卷积或张量递归的形式分别嵌入神经网络框架, 发展了高精度的SchNet^[12]和深度张量神经网络^[13](deep tensor neural-network, DTNN); 鄂维南等^[14,15]实现了基于深度学习势能面的分子动力学, 并于2020年首次完成了上亿原子的第一性原理精度的分子动力学模拟.

机器学习在势能面计算中已取得相当的进展, 为计算化学提供了新的理论工具, 并呈现出两方面的发

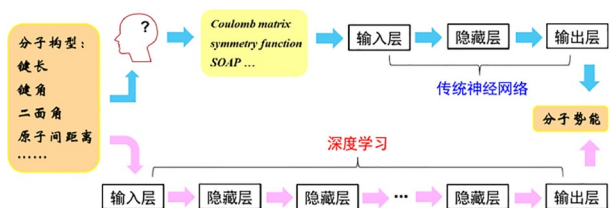


图 3 传统神经网络和深度神经网络的对比. 以计算给定核构型的分子势能为例: 传统神经网络需要事先将分子构型转换成恰当的函数作为输入值; 深度学习可以将直观的结构作为输入值, 再通过多个隐藏层提取出与势能更相关的特征 (网络版彩图)

Figure 3 Comparison between conventional neural network (top) and deep neural network (bottom). Take potential energy calculations as an example. Conventional NN must be fed with some well-designed functions of atomic coordinates, while deep NN starts with the raw input of atomic coordinates and transforms the representation into a higher and more abstract level through several hidden layers (color online).

展趋势. 一方面, 机器学习力场逐渐拓展到更复杂的化学体系, 开始深入研究更真实的化学问题. 例如, 刘智攀、商城等^[16]结合神经网络模型和随机表面行走方法发展了大规模原子模拟软件包LASP, 通过对表界面化学反应过渡态结构的搜索和表征, 预测了多种催化剂的表面活性位; 鄂维南、张林峰等^[17]开发了基于深度学习的开源软件DeePMD-kit, 被全球多个团队应用于小分子、大分子、水、半导体、表面材料和高熵合金等体系的理论研究; 兰峥岗、崔刚龙等^[18,19]分别利用核函数和神经网络, 率先实现了激发态分子的非绝热动力学模拟; Westermayr等^[20–22]系统地发展了基于深度学习的非绝热动力学计算方案, 成功地将光化学反应的模拟时间提升到纳秒量级. 另一方面, 机器学习与计算化学正在发生深度的融合, 帮助现有理论方法完成“最后一公里”. 2015年, von Lilienfeld等^[23]提出了 Δ -machine learning的策略, 利用机器学习预测高精度与低精度方法的计算结果之差, 避免了昂贵耗时的高精度计算, 低精度方法则起到了提升机器学习的数值稳定性和泛化能力的作用. 又如, 杨伟涛等^[24,25]将HDNN与多尺度量子力学/分子力学(QM/MM)模型结合起来用于分子动力学模拟和反应自由能面计算, 同时达到了从头算方法的精度和半经验方法的效率; Miller等^[26]将Hartree–Fock计算获得的电子结构信息输入到机器学习模型, 不仅实现了后Hartree–Fock水平电子相关能的预测, 而且在原则上解决了机器学习力场的可移植性问题; Small等^[27]设计了基于电子布居的

神经网络(Becke population neural network, BpopNN), 将势能分为短程项和长程项, 前者由HDNN求解, 后者由修正的库仑模型求解, 而库仑模型所需的原子电荷来自自洽场迭代得到的电子布居; 马晶等^[28]将样本的对称函数(式5、式6)作为辅助学习目标, 提出了一套可迁移的多任务学习方案, 成功预测了分子偶极矩、前线轨道能量、Gibbs自由能等12种量子化学性质.

3 机器学习在分子模拟中的应用

在前面列举的工作中, 机器学习能够绕过昂贵的量子化学计算, 实现势能面的高效估算. 但是对于复杂体系, 既无法遍历高维势能面上的所有区域, 也很难从计算数据中直接得出一个能解释实验的宏观规律. 分子动力学模拟利用“以时间换空间”的思想, 在微观与宏观之间搭建起了一座桥梁, 使得研究者可以通过有效抽样和统计平均来计算复杂体系的宏观性质. 当前分子模拟还存在诸多局限, 例如计算结果依赖于使用者的初始猜测和操作经验、抽样速度明显慢于绝大多数化学过程对应的时间尺度、计算数据冗余度高等. 机器学习可以从动力学轨迹的大量数据中提炼出有用的知识, 进而帮助人们更好地估计与动力学模拟相关的反应路径、反应坐标、分布函数或力场参数等, 为设计更加高效可靠的分子模拟方案提供了新的思路.

反应路径能够给出重要的热力学和动力学信息, 进而完整揭示反应机理. 当前各种优化算法大都只能给出离初始猜测最近的最小能量路径, 好的初始猜测能起到事半功倍的效果. 另一方面, 对于给定的化学反应, 通常很容易通过分子动力学得到反应物和产物在其能量极小区域的大量构型. Parrinello等^[29]利用线性判别分析(linear discriminant analysis, LDA)算法, 将这些构型投影到一条直线上, 属于反应物的投影点之间、以及属于产物的投影点之间尽可能接近, 而反应物和产物的投影点之间尽可能远离; 该直线即投影的方向就是反应路径的初始猜测(图4). 这一简单算法能够显著加速反应路径的优化, 已成功地推广到复杂化学反应的路径搜索和增强抽样^[30,31].

分子动力学抽样速度慢, 主要原因在于克服较高势垒需要花费很长的计算时间. 如果预先知道准确的反应坐标, 就可以利用伞形取样等增强抽样方法来加

速热力学性质的计算, 拓展动力学模拟的时间尺度. 但是, 如何确定反应坐标是一个具有挑战性的问题, 对于蛋白质折叠等复杂过程更是难以下手. Ferguson等^[32,33]在自编码-解码(autoencoder-decoder)的框架下设计了一种神经网络, 将高维分子构型同时作为输入和输出, 而居中的一个隐藏层即瓶颈层(bottleneck)只保留了很少的节点数目(图5). 高维输入信息首先被编

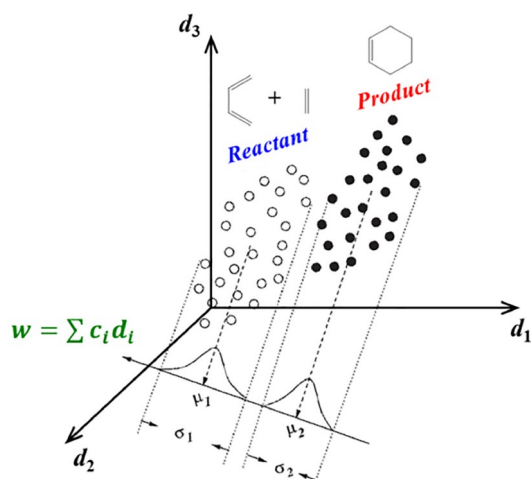


图4 利用线性判别分析方法搜索初始反应路径. 以Diels-Alder反应为例: 反应物和产物在 N 维空间的数据点被投影到直线 w 上, w 的方向即是反应路径(原文献中 $N=6$, 图中简化为 $N=3$) (网络版彩图)

Figure 4 Search for reaction path as an initial guess with linear discriminant analysis. Take the Diels-Alder reaction as an example. The N -dimensional data points related to reactants and products are projected into a line (denoted as w) that indicates the reaction path (color online).

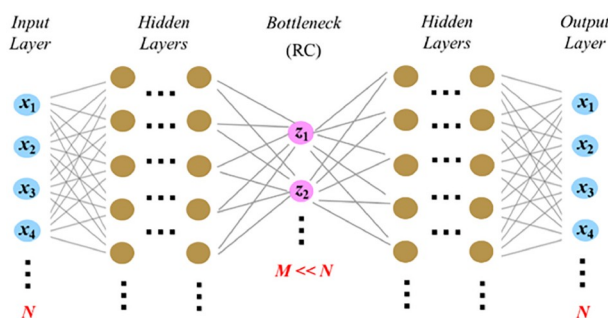


图5 利用自编码-解码神经网络寻找反应坐标. 以蛋白质的折叠过程为例: x 代表主链氨基酸残基间的 N 个二面角, z 代表神经网络找到的 M 个反应坐标 (网络版彩图)

Figure 5 Search for reaction coordinates with autoencoder-decoder neural network. Here, the N -dimensional space of backbone dihedral angles of proteins (denoted as x) is reduced to the M -dimensional subspace of reaction coordinates (denoted as z) to describe protein-folding process (color online).

码到瓶颈层, 随后被解码到输出层; 如果输入层和输出层的高维信息相同, 说明瓶颈层的低维信息足以描述不同样本在高维空间的相关性, 因而可以作为反应坐标来表示分子构型的变化. 自编码-解码神经网络模拟了人类认知复杂事物的过程, 从大量数据中提炼出“面面俱到”的低维向量用于增强抽样, 是一项有发展潜力的技术. 例如, Ensing等^[34]用遗传算法代替了神经网络的自编码结构, 求得反应坐标更容易被化学家理解; Belkacemi等^[35]基于迭代流程改进了增强抽样的权重因子校正方案, 有助于加速收敛.

体系在热平衡状态下服从玻尔兹曼分布. 对于不同的化学体系, 由于其势能面不同, 研究者往往需要花费很多时间对各种增强抽样方案进行反复调试, 才能选择出合适的算法和参数, 实现该体系重要微观状态的有效抽样. 另一方面, 既然现实世界中的玻尔兹曼分布难以处理, 如果能够在理想状态下的分布中采样, 再根据两个分布函数之间的变换关系来校正样本的权重因子, 原则上就可以直接得到符合玻尔兹曼分布的大量样本. 随着机器学习的发展, 人们不断提出新的神经网络模型, 为实现上述设想提供了强有力的工具. Noé等^[36]聚焦损失函数的表达形式和物理意义, 训练了可逆的玻尔兹曼生成器(Boltzmann generator)来描述玻尔兹曼分布和正态分布之间的可逆变换; 只要在变换后的正态分布中一次性采样, 就可以将抽取的样本映射回真实的玻尔兹曼分布, 计算正则系综下的自由能. 由此可见, 机器学习不仅能够优化已有的增强抽样方法, 还可以在不同的分布状态之间架起桥梁. 最近, 张斌等^[37]利用自回归模型(neural autoregressive model)设置自由能计算的参考态, Wirnsberger等^[38]利用标准化流(normalizing flows)构造自由能微扰的中间态, 为模拟和计算热力学性质提供了更多方案.

分子力场的不确定性是当前分子模拟的另一个局限. 目前优化和改进力场的策略包括“自下而上”(bottom-up)和“自上而下”(top-down). 以高精度的量子化学计算结果作为参考值训练得到的机器学习力场使用的是“自下而上”的策略. 但是对于生物大分子等复杂体系, 不采用大量的实验信息即“自上而下”的策略, 很难对力场进行有效的优化. “自上而下”通常需要通过反复的分子动力学不断调节力场参数, 并基于各组参数下的动力学轨迹计算出特定的热力学性质, 直至

重现实验结果. 整个流程非常繁琐耗时, 对于参数数目庞大的机器学习力场而言更加困难. 最近, Thaler等^[39]从样本权重校正因子与热力学性质之间的定量关系入手, 将“自上而下”的策略引入了基于神经网络的机器学习力场. 具体流程如下: 首先在当前力场 V^0 下进行动力学模拟, 并从轨迹中抽取若干样本; 随后通过最小化损失函数拟合出神经网络的所有参数, 即:

$$\min_{w,b} \left[\frac{\sum_i A_i e^{-\beta(V_i - V_i^0)}}{\sum_i e^{-\beta(V_i - V_i^0)}} - A^{\text{exp}} \right]^2 \quad (11)$$

其中 i 和 j 遍历抽取的样本, A 是需要重现的热力学性质, 实验值为 A^{exp} , 计算值通过对所有样本相对应的物理量做加权平均求得, V 是待求的机器学习力场的势能函数; 最后用拟合得到的 V 代替上式的 V^0 , 重复最小化过程(式11)直至有效样本数量不足, 再在替换后的力场 V^0 下重新进行动力学模拟, 迭代重复整个流程直至结果收敛. 这一方法拓展了机器学习力场在分子模拟中的应用范围, 有助于准确预测体系的宏观性质.

分子动力学方程的数值积分需计算每一步构型的势能对原子坐标的导数(即梯度), 计算量大, 耗时最多. 如果能够将这些构型看作一个时间序列, 将“过去”时刻若干个分子构型输入机器学习模型, 利用数值方法预测出“未来”时刻的构型, 就可以绕开梯度计算, 提高分子动力学的抽样效率. 传统神经网络难以处理时间序列; 循环神经网络虽然允许输入信息持续产生作用, 但只能展开到非常有限的时间, 不满足分子模拟的要求. 近年提出的长短期记忆(long-short term memory, LSTM)网络通过引入长期状态向量(cell)和多个门限控制器, 实现了对时间序列样本的“长期记忆”. 最近, Koumoutsakos等^[40]将LSTM与自编码-解码网络相结合, 首先利用自编码网络将原子坐标映射到低维空间, 然后由LSTM预测低维向量的动力学演化过程, 最后通过解码网络得到演化若干步后的原子坐标, 计算效率提升了三个数量级. 多个团队还将LSTM应用于非绝热过程中电子自由度的量子动力学演化^[41-43], 体现出这一技术在计算化学领域的普适性.

从上述工作可以看出, 机器学习不仅可以提高分子模拟中单次抽样的计算效率, 还能够直接对各种抽样策略进行优化, 改进抽样的稳定性和收敛性. 分子模拟为机器学习的应用提供了更大的平台. 除了有监

督学习, 值得关注的机器学习算法还包括: 其一, 增强抽样往往需要对算法的各个参数进行反复调试和迭代, 才能得到收敛的结果. 强化学习可以通过奖励机制自动调整抽样策略, 从而加速迭代过程^[44]. 其二, 随着计算机硬件速度的不断提高, 百万甚至上亿原子量级的化学体系已经进入了分子模拟的视野. 无监督学习能够根据化学家的兴趣, 从如此海量的数据中挖掘出关联特征和统计规律, 正逐渐成为分子模拟的又一强大工具^[45].

4 机器学习在材料基因组工程中的应用

美国政府于2011年启动了材料基因组计划, 旨在建立以理论预测在先、实验验证在后的新的研究范式, 从而取代现有的以经验为主的材料研发手段, 降低新材料的研发周期和成本. 中国自己的材料基因组工程也正在快速崛起^[46,47]. 材料基因组工程包括实验、计算和数据三大平台. 机器学习能够有效整合实验、计算的结果以及数据库的信息, 进而破译材料的结构-性能关系, 实现三大平台的融会贯通, 为材料基因组工程提供创新源泉. 可以说, 当前新材料研发的关键问题就在于如何将计算化学和机器学习正确地结合起来, 通过多层筛选(图6)实现大数据驱动下的分子设计.

数据平台是材料基因组工程的基础. 高效的信息学工具能够帮助人们探索广阔的化学空间, 进而突破现有材料基因的限制, 得到新颖的结果. 我国科学工作者已意识到数据积累对材料设计的支撑作用, 开展了大量工作. 例如, 合肥机数量子科技公司建成了包含9000万分子、90万晶体数据的大规模材料数据库; 烟台京师材料基因组工程研究院在国内率先开发了高端显示材料数据平台; 张兵兵等^[48]报道了基于第一性原理的非线性光学材料高通量筛选管线系统, 并从数据库中发现了若干性能优异的新材料. 上述工作为进一步开展机器学习奠定了基础. 另一方面, 新材料的设计是一个从无到有的过程, 因此还需要从已知化合物的数据库出发, 生成更多的、尚未研究过的全新分子结构. 最近, Kulik等^[49]在设计氧化还原液流电池(redox flow battery)时, 报道了一套分层组装方案: 首先选择8种五元或六元杂环, 对环上的原子和基团进行突变, 生成38种杂环; 随后将环上的氧原子或氮原子定义为配位原子, 将与配位原子相邻的碳原子定义为连

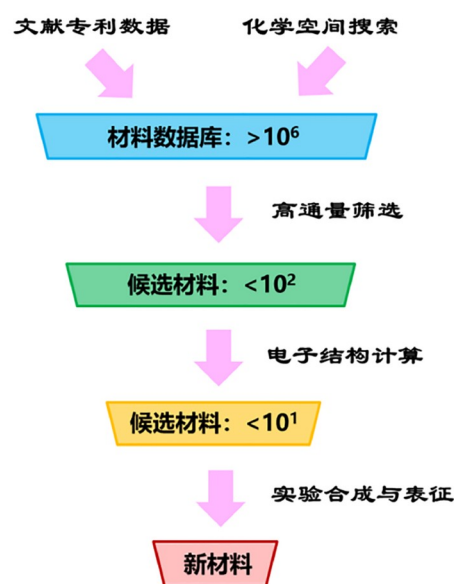


图6 新材料的发现和设计流程。机器学习在“化学空间搜索”(例如ACSESS方法)和“高通量筛选”(例如SISSO方法)的环节起关键作用(网络版彩图)

Figure 6 The workflow of new-material discovery. Many powerful tools based on machine learning have been developed for the chemical space exploration (e.g., ACSESS) and high-throughput screening (e.g., SISSO) (color online).

接点,生成741种双齿配体;再选择897种官能团对双齿配体进行官能团化,得到70万种配体,最终将280万种金属配合物(Cr, Mn, Fe, Co)用于高通量筛选。该方案的设计思路与遍历法类似,虽然不易遗漏性能优异的未知材料,但后续筛选的工作量大。杨伟涛和Beratan等^[50]基于遗传算法提出了化学空间随机搜索策略(algorithm for chemical space exploration with stochastic search, ACSESS),将分子中原子、基团和化学键的增减变化定义为“突变”,将分子切断、片段互换和重新连接的过程定义为“杂交”,并根据分子设计的需求提出若干判据用于“选择”以实现“进化”,最终获得最大差异分子集。与遍历法相比,ACSESS能够更高效地搜索与特定性能相关的化学空间;而如何根据材料设计的需求定义合适的“进化”方向,是这一策略在实际应用中能否成功的关键。

高精度计算是材料基因组工程的重点。例如,发光材料通常需要多参考态电子结构计算才能得到半定量的理论预测结果;这种水平的高精度计算离不开庞大的高性能计算资源和研究者的丰富经验,无法遍历材料数据库中数以万计的候选分子。机器学习能够构造

出与材料性能直接相关的描述符(descriptor),并利用这套描述符从海量分子中快速筛选出数十个候选材料,再对这些分子开展高精度计算,从而大大减少了工作量。以二元金属/非金属材料 A_mB_n 的分类为例:根据已有的化学知识,可将第一电离能、电子亲和能、Pauling电负性、原子共价半径、原子间距、原子体积与单位晶胞体积之比、配位数、价电子数和元素组分等特性作为原始描述符。然而,如果将这些描述符直接作为自变量,并不足以准确地区分金属和非金属材料。欧阳润海等^[51]基于压缩感知理论提出了一种寻找最佳描述符的方法(sure independence screening and sparsifying operators, SISSO)。该方法首先对原始描述符进行各种数学运算以生成新的描述符,然后将样本数据点之间的重叠度进行最小化以提升分类的准确性,最终找到两个最佳描述符:

$$D_1 = \frac{x_A}{\sum V_{\text{atom}}/V_{\text{cell}}} \cdot \frac{\text{IE}_B/\chi_B}{\chi_A} \quad (12)$$

$$D_2 = \chi_A^2 \left\| 1 - 2x_A \right\| - x_A^2 \chi_B / \chi_A$$

D_1 和 D_2 包含元素组分 x 、电离能(IE)、Pauling电负性 χ 以及原子体积 V_{atom} 与单位晶胞体积 V_{cell} 之比等四个原始描述符,对二元金属/非金属材料分类的准确度超过99%。更重要的是,所有原始描述符都可以直接从分子结构式或化学手册中获取,不需要高精度计算。由此可见,机器学习有望建立起一个高通量筛选在先、高精度计算在后的研发策略,为材料基因组工程提供新一代“加速器”。

从上述工作可以看出,机器学习在材料设计的两个关键环节即“化学空间搜索”和“高通量筛选”中分别起到了数据引导和加速的作用(图6)。当然,机器学习在材料基因组工程中的应用尚处于起步阶段,还不能仅由机器学习得到性能优良的新材料,需要理论计算和实验的通力合作。2016年,Aspuru-Guzik等^[52]利用神经网络和电子结构计算从160万个化合物中筛选出近2500种候选分子,并根据预期性能、合成难度和新颖性制备了4种有机发光二极管器件,外量子效率最高可达到22%。该工作的关键在于机器学习与高通量计算的融合:首先通过大量测试,在含时密度泛函理论的水平上建立了热活化延迟荧光速率的计算方案;随后利用该方案获得40000个化合物的速率构建初始数

数据集,在此基础上训练神经网络用于快速预测其余数百万个分子的热活化延迟荧光速率;再对预测性能较好的样本开展理论计算,并将结果加入训练集,更新神经网络.这套流程实际共执行数十万次电子结构计算,即使机器学习已经节省一个数量级的计算量,仍需要1300万个CPU核时,可见高精度第一性原理计算的重要性.

当前各类功能材料的数据集往往存在着样本量不足、类别不平衡、物理性能信息缺失等局限^[53],而材料数据的积累和优化是一项长期性的工作,因此特别需要新的机器学习方法以缩短研发周期.近年来,王金兰等^[54,55]针对这些问题发展了一系列机器学习策略,在钙钛矿材料、二维铁磁性材料等领域取得了显著进展.例如在寻找有机无机杂化钙钛矿材料时将回归模型与特征提取相结合,探讨了材料带隙和容忍因子(tolerance factor)、八面体因子(octahedral factor)、一价阳离子极化率、二价阳离子电负性等特征之间的统计相关性,不仅显著提高了机器学习的筛选效率,还揭示了材料结构与性能之间的内在规律,最终从5000余种候选化合物中成功筛选出两种综合性能优良的无铅太阳能电池材料.又如在寻找二维铁磁材料时基于梯度提升算法、边际抽样以及第一性原理计算设计了主动学习(active learning)的框架,并集成原子磁性、晶体场、配位环境和近邻金属原子等四类描述符,实现了对磁性材料化学空间的高效搜索,最终获得700余种高居里温度的二维铁磁材料.由此可见,与上千万核时的高通量计算相比,如何设计具有物理意义的描述符,如何从预测模型中提取具有解释能力的特征,如何实现自适应的机器学习流程,是材料研发中从小样本到大数据的关键.最近,人们还将迁移学习、多保真数据(multi-fidelity data)和生成对抗网络等技术应用于材料的性能预测与优化^[56-58],为设计新材料提供了新途径.需要注意的是,当已有样本量和特征维度均较少时, k -近邻、高斯过程回归等简单算法的表现有可能反超更复杂的集成学习和神经网络模型^[59].如何选择合适的机器学习模型,离不开研究者对数据的深入分析和对材料结构-性能关系的深刻理解.

最近,我们发展了一个适用于无机非线性光学材料的机器学习模型^[60].人们在寻找和制备非线性光学晶体的实验过程中,提出了阴离子基团理论和电荷转移理论等经验规则,初步揭示了影响晶体光学性能的

几何因素和物理因素.设计一种新的无机非线性光学材料,其核心策略是引入与倍频效应相关的阴、阳离子基团,并将这些结构基元进行优化组合和定向组装.当前已经发现了数百种具有实用意义的非线性光学晶体,但主要局限于硼酸盐和磷酸盐体系.我们首先借鉴阴离子基团理论,根据“化学键-倍频基因-晶体”的三级材料结构,设计了多层描述符(图7).前两层描述符根据晶体的化学组分即可唯一定义.底层描述符包括晶体中原子的范德华半径、Pauling电负性、其稳定单质的密度、熔点、沸点等,都可直接从化学手册中获取其数值;中间层次的描述符包括相应的酸根离子(对应晶体中的阴离子基团)和金属氧化物(对应阳离子基团)的偶极矩、极化率、超极化率、前线轨道能隙等,只需要对这些小分子在气相中进行电子结构计算即可求得,在高通量筛选中非常方便.第三层次的描述符包括晶体空间群、晶胞参数、等价原子对称性(Wyckoff position)等信息,在训练样本充足且已解出候选材料晶体结构的情况下,可提升机器学习的预测精度.随后,我们基于前两个层次的描述符,利用随机森林和支持向量机构建了近似分类器,用于判断给定晶体在倍频系数和双折射率两方面的表现,并对2534种非线性光学性能未报道过的晶体进行虚拟扫描,利用SISSO得到了最佳描述符及其分布规律(图8).最后,我们对机器学习筛选出的30种候选材料开展高精度的第一性原理计算,获得6种有前景的新型非线性光学晶体:Cs(Al₄Be₄B₁₂O₂₈)、Pb(Ga₂O₄)、BiCa₂(AsO₄)O₂、

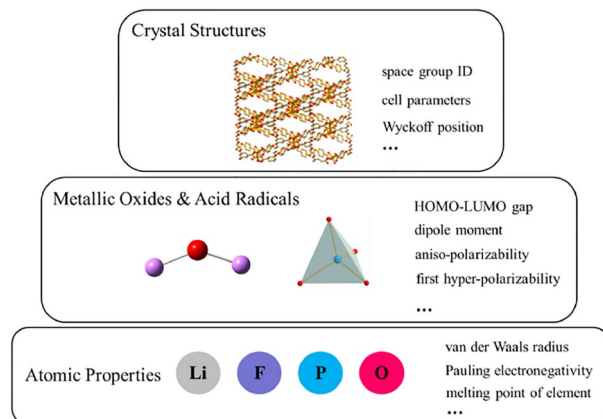


图7 适用于无机非线性光学材料的多层描述符(网络版彩图)

Figure 7 Multilevel descriptors for inorganic nonlinear optical crystals (color online).

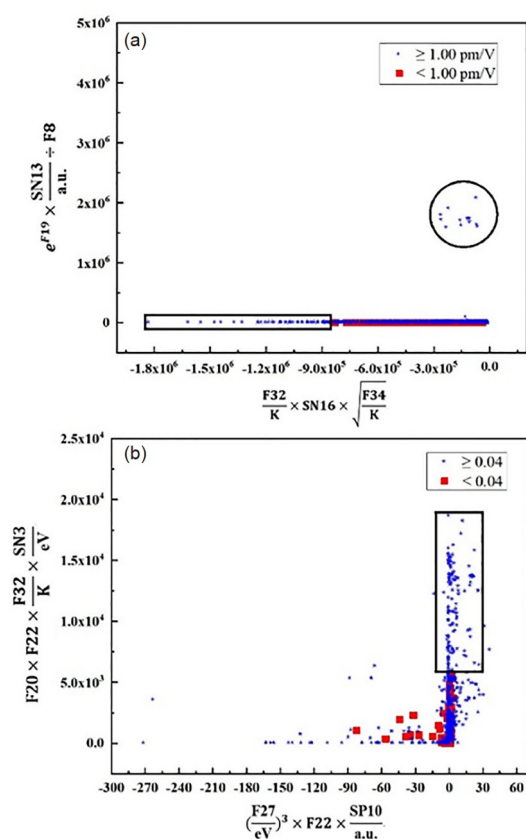


图8 非线性光学晶体的倍频系数(a)和双折射率(b)对应的最佳描述符。在蓝色样本点集中的区域(黑框内部)进行搜索可以更高效地筛选出性能优良的晶体(网络版彩图)

Figure 8 Distribution of crystals with d_{ij} (a) and Δn (b) on optimal descriptors. The highlighted regions with high-density positive samples (in blue) suggest a larger possibility to discover NLO crystals (color online).

$\text{Li}_2\text{GeMo}_3\text{O}_8$ 、 $\text{Ba}_3\text{B}_6\text{F}_2\text{O}_{11}$ 、 $\text{Li}_2\text{SnMo}_3\text{O}_8$ (其中两种位于深紫外波段, 两种位于中远红外波段), 在半定量水平初步实现了数据驱动的非线性光学材料设计。

机器学习离不开足量的训练样本, 如果某种功能材料过于匮乏, 机器学习也无从谈起。另一方面, 当样本数量充足的时候, 研究者能够依赖长期积累的经验 and 化学直觉, 利用传统的试错法对现有材料直接加以改性, 也可得到综合性能更优的新材料。由于不同类型功能材料的样本数量差异很大, 机器学习在实践中是否必然优于传统范式, 还需进一步检验。最近, del Cueto等^[61]借助高斯混合函数对这一问题开展了深入研究。该工作首先基于已有样本建立机器学习模型用

于预测材料性能, 随后对材料的结构特征进行突变, 并考虑两种策略: 一是随机选择某个突变后的样本, 相当于试错; 二是利用机器学习预测突变样本的性能, 选择预测值最优者。计算结果表明, 当样本数量很少时, 或者当仅对性能优良的材料(前40%)进行突变时, 机器学习策略相对于随机选择并无优势; 如果放宽范围, 对性能前70%的材料进行突变, 机器学习出现了一个“收益窗口”, 即当样本数量在一定范围内(并非越多越好), 机器学习才有优势。因此, 只有深入了解各种功能材料的研发现状, 才能确保机器学习的有效性。把机器学习作为“黑箱”的尝试在材料设计中是难以成功的。

5 结论与展望

由于篇幅所限, 本文只列举了近年来计算化学和机器学习相结合的一小部分工作。机器学习在计算化学中的成功范例正在以惊人的速度增加。机器学习不仅可以节省理论模拟所需要的计算资源, 还能帮助研究者跳出前人经验的框架, 从越来越多的计算数据中挖掘全新的知识, 开发全新的材料。与此同时, 计算化学的发展也可以为机器学习提供更多高质量的训练样本, 并在机器学习中起着“路标”式的导向作用。

机器学习在计算化学领域的进一步突破有赖于数据“质”与“量”的不断提升。具体包括以下四点: 第一, 在计算化学中经常出现样本量很少的情形, 亟需在实践中发展并不断完善适用于小样本的机器学习方法。第二, 机器学习能够帮助计算化学更高效地探索未知的化学空间和构型空间, 但如何在这一过程中评估机器学习模型的可靠性, 实现“即时学习”^[62], 还需要着眼于机器学习的理论和算法, 开展深入研究。第三, 机器学习的能力上限取决于数据本身的质量, 而计算化学的数据集分布比较零散, 缺乏对误差来源和不确定性的评估, 一些急需的数据库尚未建立, 能够满足机器学习要求的数据体系标准还在摸索中。第四, 机器学习模型本身并无物理意义, 而化学体系必须服从物理学的基本规律; 这就需要通过设计有约束条件的机器学习算法^[63], 将这些规律嵌构到数值模型中, 机器学习才能真正成为可靠且实用的计算化学工具。

参考文献

- 1 Blank TB, Brown SD, Calhoun AW, Doren DJ. *J Chem Phys*, 1995, 103: 4129–4137
- 2 Chen J, Xu X, Xu X, Zhang DH. *J Chem Phys*, 2013, 138: 154301
- 3 Chen R, Shao K, Fu B, Zhang DH. *J Chem Phys*, 2020, 152: 204307
- 4 Behler J, Parrinello M. *Phys Rev Lett*, 2007, 98: 146401
- 5 Behler J. *Angew Chem Int Ed*, 2017, 56: 12828–12840
- 6 Behler J. *Chem Rev*, 2021, 121: 10037–10072
- 7 Rupp M. *Int J Quantum Chem*, 2015, 115: 1058–1073
- 8 Dral PO, Owens A, Yurchenko SN, Thiel W. *J Chem Phys*, 2017, 146: 244108
- 9 Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA. *Phys Rev Lett*, 2012, 108: 058301
- 10 Bartók AP, Kondor R, Csányi G. *Phys Rev B*, 2013, 87: 184115
- 11 Pinheiro M, Ge F, Ferré N, Dral PO, Barbatti M. *Chem Sci*, 2021, 12: 14396–14413
- 12 Schütt KT, Saucedo HE, Kindermans PJ, Tkatchenko A, Müller KR. *J Chem Phys*, 2018, 148: 241722
- 13 Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A. *Nat Commun*, 2017, 8: 13890
- 14 Zhang L, Han J, Wang H, Car R, E W. *Phys Rev Lett*, 2018, 120: 143001
- 15 Jia W, Wang H, Chen MH, Lu DH, Lin L, Car R, E W, Zhang LF. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. In *Proceedings of SC 2020: International Conference for High Performance Computing, Networking, Storage and Analysis*. Atlanta: IEEE Computer Society 2020. 20392450
- 16 Huang SD, Shang C, Kang PL, Zhang XJ, Liu ZP. *WIREs Comput Mol Sci*, 2019, 9: e1415
- 17 Wang H, Zhang L, Han J, E W. *Comput Phys Commun*, 2018, 228: 178–184
- 18 Hu D, Xie Y, Li X, Li L, Lan Z. *J Phys Chem Lett*, 2018, 9: 2725–2732
- 19 Chen WK, Liu XY, Fang WH, Dral PO, Cui G. *J Phys Chem Lett*, 2018, 9: 6702–6708
- 20 Westermayr J, Gastegger M, Menger MFSJ, Mai S, González L, Marquetand P. *Chem Sci*, 2019, 10: 8100–8107
- 21 Westermayr J, Gastegger M, Marquetand P. *J Phys Chem Lett*, 2020, 11: 3828–3834
- 22 Westermayr J, Marquetand P. *Chem Rev*, 2021, 121: 9873–9926
- 23 Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. *J Chem Theor Comput*, 2015, 11: 2087–2096
- 24 Shen L, Wu J, Yang W. *J Chem Theor Comput*, 2016, 12: 4934–4946
- 25 Shen L, Yang W. *J Chem Theor Comput*, 2018, 14: 1442–1455
- 26 Welborn M, Cheng L, Miller III TF. *J Chem Theor Comput*, 2018, 14: 4772–4779
- 27 Xie X, Persson KA, Small DW. *J Chem Theor Comput*, 2020, 16: 4256–4270
- 28 Liu Z, Lin L, Jia Q, Cheng Z, Jiang Y, Guo Y, Ma J. *J Chem Inf Model*, 2021, 61: 1066–1082
- 29 Mendels D, Piccini GM, Parrinello M. *J Phys Chem Lett*, 2018, 9: 2776–2781
- 30 Rizzi V, Mendels D, Sicilia E, Parrinello M. *J Chem Theor Comput*, 2019, 15: 4507–4515
- 31 Bonati L, Rizzi V, Parrinello M. *J Phys Chem Lett*, 2020, 11: 2998–3004
- 32 Chen W, Tan AR, Ferguson AL. *J Chem Phys*, 2018, 149: 072312
- 33 Chen W, Ferguson AL. *J Comput Chem*, 2018, 39: 2079–2102
- 34 Hooft F, Pérez de Alba Ortiz A, Ensing B. *J Chem Theor Comput*, 2021, 17: 2294–2306
- 35 Belkacemi Z, Gkeka P, Lelièvre T, Stoltz G. *J Chem Theor Comput*, 2022, 18: 59–78
- 36 Noé F, Olsson S, Köhler J, Wu H. *Science*, 2019, 365: aaw1147
- 37 Ding X, Zhang B. *J Phys Chem B*, 2020, 124: 10166–10172
- 38 Wirsberger P, Ballard AJ, Papamakarios G, Abercrombie S, Racanière S, Pritzel A, Jimenez Rezende D, Blundell C. *J Chem Phys*, 2020, 153: 144112
- 39 Thaler S, Zavadlav J. *Nat Commun*, 2021, 12: 6884
- 40 Vlachas PR, Zavadlav J, Praprotnik M, Koumoutsakos P. *J Chem Theor Comput*, 2022, 18: 538–549
- 41 Yang B, He B, Wan J, Kubal S, Zhao Y. *Chem Phys*, 2020, 528: 110509

- 42 Lin K, Peng J, Gu FL, Lan Z. *J Phys Chem Lett*, 2021, 12: 10225–10234
- 43 Wu D, Hu Z, Li J, Sun X. *J Chem Phys*, 2021, 155: 224104
- 44 Shamsi Z, Cheng KJ, Shukla D. *J Phys Chem B*, 2018, 122: 8386–8395
- 45 Glielmo A, Husic BE, Rodriguez A, Clementi C, Noé F, Laio A. *Chem Rev*, 2021, 121: 9722–9758
- 46 de Pablo JJ, Jackson NE, Webb MA, Chen LQ, Moore JE, Morgan D, Jacobs R, Pollock T, Schlom DG, Toberer ES, Analytis J, Dabo I, DeLongchamp DM, Fiete GA, Grason GM, Hautier G, Mo Y, Rajan K, Reed EJ, Rodriguez E, Stevanovic V, Suntivich J, Thornton K, Zhao JC. *npj Comput Mater*, 2019, 5: 41
- 47 Li Y, Liu L, Chen W, An L. *Sci Sin-Chim*, 2018, 48: 243–255 (in Chinese) [李云琦, 刘伦洋, 陈文多, 安立佳. 中国科学:化学, 2018, 48: 243–255]
- 48 Yu J, Zhang B, Zhang X, Wang Y, Wu K, Lee MH. *ACS Appl Mater Interfaces*, 2020, 12: 45023–45035
- 49 Janet JP, Ramesh S, Duan C, Kulik HJ. *ACS Cent Sci*, 2020, 6: 513–524
- 50 Virshup AM, Contreras-García J, Wipf P, Yang W, Beratan DN. *J Am Chem Soc*, 2013, 135: 7296–7303
- 51 Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M, Ghiringhelli LM. *Phys Rev Mater*, 2018, 2: 083802
- 52 Gómez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel TD, Duvenaud D, Maclaurin D, Blood-Forsythe MA, Chae HS, Einzinger M, Ha DG, Wu T, Markopoulos G, Jeon S, Kang H, Miyazaki H, Numata M, Kim S, Huang W, Hong SI, Baldo M, Adams RP, Aspuru-Guzik A. *Nat Mater*, 2016, 15: 1120–1127
- 53 Zhou Q, Lu S, Wu Y, Wang J. *J Phys Chem Lett*, 2020, 11: 3920–3927
- 54 Lu S, Zhou Q, Ouyang Y, Guo Y, Li Q, Wang J. *Nat Commun*, 2018, 9: 3405
- 55 Lu S, Zhou Q, Guo Y, Wang J. *Chem*, 2022, 8: 769–783
- 56 Yamada H, Liu C, Wu S, Koyama Y, Ju S, Shiomi J, Morikawa J, Yoshida R. *ACS Cent Sci*, 2019, 5: 1717–1730
- 57 Chen C, Zuo Y, Ye W, Li X, Ong SP. *Nat Comput Sci*, 2021, 1: 46–53
- 58 Song Y, Siriwardane EMD, Zhao Y, Hu J. *ACS Appl Mater Interfaces*, 2021, 13: 53303–53313
- 59 Park C, Lee JW, Kim M, Lee BD, Singh SP, Park WB, Sohn KS. *Inorg Chem Front*, 2021, 8: 4610–4624
- 60 Zhang ZY, Liu X, Shen L, Chen L, Fang WH. *J Phys Chem C*, 2021, 125: 25175–25188
- 61 del Cueto M, Troisi A. *Phys Chem Chem Phys*, 2021, 23: 14156–14163
- 62 Li Z, Kermode JR, De Vita A. *Phys Rev Lett*, 2015, 114: 096405
- 63 Beucler T, Pritchard M, Rasp S, Ott J, Baldi P, Gentile P. *Phys Rev Lett*, 2021, 126: 098302

Machine learning in computational chemistry

Lin Shen^{*}, Luyang Jia, Diandong Tang, Zhan-Yun Zhang, Ding Peng

Key Laboratory of Theoretical and Computational Photochemistry of Ministry of Education, College of Chemistry, Beijing Normal University, Beijing 100875, China

Abstract: Machine learning (ML) is becoming a powerful tool in many domains in science. Its development on computational chemistry is still at an early stage but has dramatically and simultaneously improved the accuracy and efficiency of existing computational approaches in theoretical chemistry. Here, we summarized several typical and recent applications of machine learning in three aspects of computational chemistry: calculation on potential energy surface, molecular dynamics simulation, and new material discovery within the framework of the material genome initiative.

Keywords: potential energy surface, molecular dynamics, material genome, machine learning

doi: [10.1360/SSC-2022-0022](https://doi.org/10.1360/SSC-2022-0022)