

基于机器学习势函数的原子模拟软件的开发及应用

商 城, 康沛林, 刘智攀

(复旦大学化学系, 上海 200433)

摘 要: 近年来, 基于机器学习的大规模原子模拟技术的发展为许多化学分支的发展带来了巨大的希望。这些模拟具有高速度和高精度的特点。本文概述了基于机器学习势函数的原子模拟的 3 个关键方面的最新进展, 即机器学习模型和结构描述子的开发, 全局势能面训练集的生成, 以及基于主动学习的势函数自动训练。研究表明, 最近设计的指数型结构描述子和前馈神经网络模型非常适合生成高度复杂的全局势能面。通过神经网络势函数在材料和反应模拟中的 2 个最新应用来说明基于机器学习势函数的原子模拟如何有助于发现新材料和反应。

关键词: 机器学习; 势函数; 全局势能面; 原子模拟软件

中图分类号: O6 文献标志码: A 文章编号: 0454-5648(2023)02-0476-12

网络出版时间: 2022-12-27



Development and Application of Atomic Simulation Software Based on Machine Learning Potentials

SHANG Cheng, KANG Peilin, LIU Zhipan

(Department of Chemistry, Fudan University, Shanghai 200433, China)

Abstract: Recent development of large-scale atomic simulation techniques based on machine learning has brought a great promise in chemistry. These simulations are featured by both high speed and high accuracy. This review outlined recent development on three key aspects of atomic simulation based on machine learning potential, *i.e.*, machine learning models and structure descriptors, generation of global potential energy surface training sets, and automatic training of potential functions based on active learning. It is indicated that the designed structure descriptor and feedforward neural network model are suitable for generating a highly complex global potential energy surface. In addition, the applications of LASP software in material and reaction simulations were also selected to illustrate how ML-based atomic simulation could assist the discovery of novel materials and reactions.

Keywords: machine learning; potential; global potential energy surface; atomic simulation software

1 机器学习辅助原子模拟在化学与材料领域中的机遇与挑战

针对材料和化学反应的模拟, 一直是化学研究的前沿问题。其中最具挑战的问题是如何在复杂度化学环境中处理结构中的化学键演化, 比如催化材料表界面发生的化学反应, 以及不同材料在界面处

形成的异质结构等。在过去很长的一段时间里, 对这些问题的研究只能通过基于量子力学^[1-3]的分子动力学模拟来实现。通过手动建立的原子模型, 在超级计算机上求解 Schrodinger 方程, 获得模型所对应的精确能量以及能量梯度, 并根据 Newton 力学对结构随时间进行演化。虽然当代科技的发展已经使得计算机的计算能力获得了飞跃, 然而量子

收稿日期: 2022-10-01。 修订日期: 2022-10-13。

基金项目: 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)。

第一作者: 商 城(1985—), 男, 博士, 教授。

Received date: 2022-10-01. Revised date: 2022-10-13.

First author: SHANG Cheng (1985—), male, Ph.D., Professor.

E-mail: cshang@fudan.edu.cn

力学计算受到方法本身等的限制, 对所研究的体系依然具有时间和空间尺度上的限制, 通常仅能对亚纳米(约 100 个原子)在皮秒级别进行模拟^[4]。这与化学合成通常所需的秒级甚至小时级的反应时间, 以及合成材料一般能达到的几十纳米至微米级的尺寸依然具有数量级的差别。因此, 完全依赖量子力学计算的原子模拟很难实现接近真实反应条件的原子模拟, 而如何在计算误差可控的前提下提高原子模拟的体系规模, 并快速对构象空间进行广泛采样的也成为理论化学家一直以来追求的目标。

随着深度学习方法和 GPU 硬件的发展, 人工智能技术在最近几年得到了飞速的发展^[5], 并逐渐应用于各个研究领域, 其中就包括针对化学反应与材料结构的理论模拟。现代机器学习技术用于原子模拟最早可以追溯到 20 世纪 90 年代, 针对具有有限自由度的小分子体系的势能面构建^[6-7], 其中机器学习技术主要是基于神经网络模型, 被用于拟合基于量子力学计算的势能面数据集, 形成机器学习势函数, 用于快速势能面搜索。与传统的经验分子立场相比, 机器学习势函数具有更为复杂的数学函数形式和更高维度的参数空间, 并因此实现了传统力场难以达到的能够恰当描述化学过程的计算精度。然而, 将这一方法推广至大规模原子模拟的过程却并不是一帆风顺, 其最主要的制约因素是难以获得高质量的量子力学势能面数据集^[8-27]。

与量子力学计算中通过求解 Schrodinger 方程获得高精度的势能面不同, 基于机器学习势函数依赖于通过对高质量的势能面数据集进行学习才有可能得到较高的高精度。作为第 1 步的数据集产生, 其实是整个流程当中最耗费计算资源的一步, 解决这一问题的关键则是设计高效的势能面采样方法来生成具有代表性和紧凑性的势能面数据集。高效势能面搜索方法因此在基于机器学习的原子模拟方法中扮演着至关重要的作用, 它一方面直接决定了模拟过程在有限的时间内能够探索的构象空间的范围, 以及发现关键结构及过程的能力, 另一方面, 作为产生机器学习数据集结构的重要手段, 也决定了机器学习势函数的可靠性和使用场景的广泛性。

一旦建立了数据集, 就可以利用势能面数据集的定量信息(结构坐标、总能量和原子力), 来训练机器学习模型, 这一过程属于有监督机器学习的过程。训练样本的大小和质量决定了有监督学习在产生机器学习潜能方面的效率。一般来说, 机器学习单纯学习训练集中的信息, 但对训练数据的质

量一无所知, 因此可以称之为被动学习。传统的被动监督学习往往需要人工设置采样参数, 分析训练数据的质量, 添加所需的新数据。这样生成的训练集可能包含大量冗余信息, 比如构型相似的结构。这不仅浪费了量子力学计算, 而且由于数据集分布不均, 容易导致过拟合。为了克服被动监督学习这一主要缺陷, 目前的机器学习势函数主要通过主动学习产生^[28-30]。主动学习方法整合了通过迭代的方式逐步改进机器学习模型。它从有限数量的标记训练样本开始, 然后迭代执行机器学习模型训练, 运用阶段性的机器学习势函数进行原子模拟, 产生新的数据, 并根据一些预定义的数据选择规则追加标记训练数据集, 直到满足停止条件。

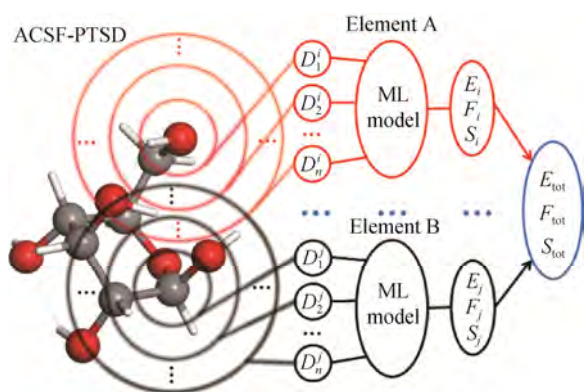
本文从机器学习模型、原子模拟产生训练集、势函数主动学习方法等 3 个方面对机器学习原子模拟的进展进行了简单概述, 并着重介绍了本小组开发的基于神经网络势函数的大规模原子模拟软件(LASP)^[27]及其主要功能随机势能面行走全局优化神经网络势(SSW-NN)方法^[20-21]。该软件实现了势能面数据的生成、神经网络势函数的训练和基于势函数的原子模拟。运用 LASP 软件, 通过自动的全局搜索方法进行全局势能面采样, 建立了大量的全局神经网络(G-NN)势^[20], 涵盖了元素周期表中的大量元素并广泛应用于包括非均相催化, 揭示有机反应网络等大规模原子模拟中。

2 机器学习模型和结构描述子

为了保证良好的机器学习潜力, 机器学习模型的设计及其输入(结构描述符)的选择必须遵守 2 条基本规则: 1) 体系大小不变性。机器学习模型需要保持灵活性, 以准确描述从分子体系到固体等不同大小的体系。结构的超胞(原子数的增加)不应改变每个原子的能量; 2) 结构描述符的连续性。为了获得具有高数值的原子力, 机器学习模型的输入应该是连续的, 并且可以相对于原子坐标解析求导。受这 2 个规则的限制, 机器学习模型采用与体系大小无关的局部结构信息作为输入; 它们的结构描述符保持了对体系的平移、旋转和置换不变性, 以生成连续可导的势能面。

上述 2 个基本规则自 20 世纪 60 年代以来发展起来的经验力场方法就已经满足^[31], 研究者们将总能量分解为两体(键)、三体(角)和四体(二面角)项等。然而这些内部坐标通常为短程描述符, 无法描述晶体材料。另一方面这些函数形式也不够复杂, 无法

模拟化学反应过程。Behler 和 Parrinello 等提出了原子能量框架的高维神经网络体系结构(HDNN)^[8]以解决实际应用中体系大小变化的需求,其中体系总能量被表示为单原子能量的加和。而单原子能量由原子周围的化学环境决定,通过将一组结构描述符 D_i 表示的局部化学环境信息与总能量相关联,可以使用前馈神经网络对原子能量进行训练。每个 D_i 都是一个原子中心对称函数(ACSF),它是一系列两体径向和三体角向函数的求和形式(图 1)。此后,研究者们提出了许多机器学习的模型,如高斯过程回归(GPR)^[13–14, 32–33]和核方法脊回归(KRR)^[25],支持向量机(SVM)^[34–35]和谱邻域分析势(SNAP)^[36]等。这些方法中基于核函数和基于神经网络的势函数应用较多。相比较而言,基于核函数的模型所需的参数量远小于神经网络模型,也因此更容易训练。然而由于核函数方法中,运算代价随训练集呈 n^3 增长(n 为训练集大小),因此核函数模型严重受限训练集大小。总体而言,神经网络模型仍是目前势函数中使用最为广泛的模型,而这些神经网络模型几乎无一例外地采用了 HDNN 的基本原理。相比之下,局部化学环境有许多可替代的表示方法,如 Behler 和 Parrinello 提出的高斯型原子中心对称函数(ACSF)^[8–9]、在 Deep Potential 中使用的基于原子间两体距离的卷积神经网络架构^[37–38],基于图网络的 MPNN^[39]自学习描述符^[40–41],基于高斯型轨道密度向量结构描述符^[42]等。对于这些模型与描述符之间的比较,可以参考最近的一些综述性文献^[28, 43–46]。



Left: the vector D_i is built from a series of element-discriminated functions based on atomic coordinates; Right: ML models utilize the atomic vector D_i as the input and decompose the system properties including energy, force and stress (E_{tot} , F_{tot} , and S_{tot}) as the sum of atomic contribution.

图 1 机器学习模型的示意图^[47]

Fig. 1 An illustration of ACSF-PTSD based machine learning model^[47]

刘智攀等提出了指数函数型结构描述符(PTSD)作为结构描述子式(1)~式(10)^[21],目前采用的 PTSD 函数类似于 ACSF,也采用加和形式并包含两体和三体函数。不同之处在于使用幂指数函数代替了 Gaussian 函数,此外,增加了四体项、球谐函数和内侧截断函数来增强结构识别能力。式(4)和式(7)中的幂函数和球谐函数的组合模拟了类 H 原子波函数,这为耦合原子的径向和角度信息提供了一种方便的方法。式(8)中的四体项可以描述二面角,对于描述有机分子构型和反应至关重要。式(9)和式(10)则屏蔽了近距离原子的信息在加和项中的干扰,帮助更好地突出描述长距离原子信息。

$$f_c(r_{ij}) = \begin{cases} 0.5 \times \tanh^3 \left(1 - \frac{r_{ij}}{r_c} \right), & \text{for } r_{ij} \leq r_c \\ 0 & \text{for } r_{ij} > r_c \end{cases} \quad (1)$$

$$R^n(r_{ij}) = r_{ij}^n \cdot f_c(r_{ij}) \quad (2)$$

$$S_i^1 = \sum_{j \neq i} R^n(r_{ij}) \quad (3)$$

$$S_i^2 = \left[\sum_{m=-L}^L \left| \sum_{j \neq i} R^n(r_{ij}) Y_{Lm}(r_{ij}) \right|^2 \right]^{\frac{1}{2}} \quad (4)$$

$$S_i^3 = 2^{1-\zeta} \sum_{j,k \neq i} (1 + \lambda \cos \theta_{ijk})^\zeta \cdot R^n(r_{ij}) \cdot R^m(r_{ik}) \cdot R^p(r_{jk}) \quad (5)$$

$$S_i^4 = 2^{1-\zeta} \sum_{j,k \neq i} (1 + \lambda \cos \theta_{ijk})^\zeta \cdot R^n(r_{ij}) \cdot R^m(r_{ik}) \quad (6)$$

$$S_i^5 = \left\{ \sum_{m=-L}^L \left| \sum_{j,k \neq i} R^n(r_{ij}) \cdot R^m(r_{ik}) \cdot R^p(r_{jk}) \cdot [Y_{Lm}(r_{ij}) + Y_{Lm}(r_{ik})] \right|^2 \right\}^{\frac{1}{2}} \quad (7)$$

$$S_i^6 = 2^{1-\zeta} \sum_{j,k,l \neq i} [(1 + \lambda \cos \delta_{ijkl})^\zeta \cdot R^n(r_{ij}) R^m(r_{ik}) R^p(r_{il})] \quad (8)$$

$$S_i^7 = \sum_{j \neq i} \{ R^n(r_{ij}) \cdot [1 + \tanh \alpha \cdot (r_{ij} - r_m)] \} \quad (9)$$

$$S_i^8 = \left[\sum_{m=-L}^L \left| \sum_{j \neq i} R^n(r_{ij}) Y_{Lm}(r_{ij}) \cdot [1 + \tanh[\alpha \cdot (r_{ij} - r_m)]] \right|^2 \right]^{\frac{1}{2}} \quad (10)$$

式中: r_{ij} 是原子 i 和 j 之间的核间距; θ_{ijk} 是以 i 原子为中心,由邻居原子 j 和 k 定义的夹角(i 、 j 和 k 是原子序号),其余参数: r_c 、 n 、 L 、 ζ 、 λ 、 m 、 p 、 r_m , 均为结构描述符相关的用户设定参数; PTSD 的主要成分是截断函数 f_c [公式(1),用以确保函数在截断半径 r_c 处衰减为零],幂指数径向函数,三角函

数, 球谐函数和双曲正切函数; S_1 和 S_2 是两体函数; S_3 、 S_4 和 S_5 是三体函数; S_6 是四体函数, S_7 和 S_8 则是对应 S_1 和 S_2 添加了内侧屏蔽。

代表性的 PTSD 径向函数特点如图 2 所示, S_1 中只有 1 个参数 n 的幂指数函数与截断函数结合后, 可以产生灵活的峰值和形状的径向分布, 从而有效地提取特定径向窗口中的结构信息。类似地三体函数中不同幂指数(n, m, p)的组合可以方便地耦合不同的径向分布。图 2 显示了 1 组典型的 S_1 到 S_6 函数, 用于描述具有不同配位情况的原子, 依次为两配位(直线型)、三配位(三角形)、四配位(正方

形)、五配位(五边形)、六配位(六边形)、四配位(四面体)、八配位(立方体)、六配位(八面体)。在这些结构中, 所有相邻原子与中心原子的距离都是相同的(2 Å), 所有 PTSD 中的幂指数取 $n=2$ 。如图 2 所示, 具有不同角量子数($S_2, l=2, 4, 6$)的球谐函数虽然是两体函数, 但可以有效提取原子周围环境中的角向信息。图 2 显示, $S_3 \sim S_6$ 三体项和四体项可以进一步在描述原子周围环境时产生不同的值, 尤其是对于高配位数环境(大于 3), 因此是对两体函数的有效补充, 正是 PTSD 值的巨大变化实现了敏感区分不同结构的能力。

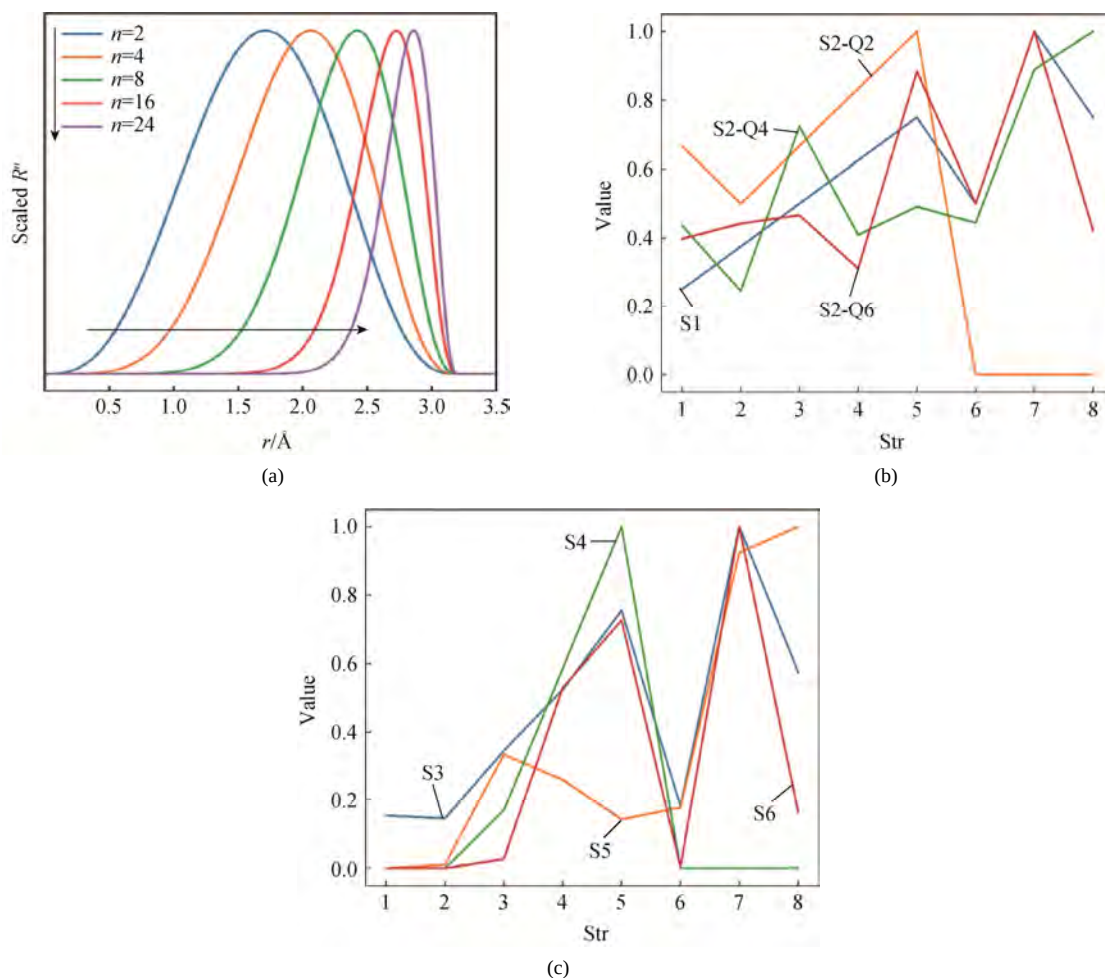


图 2 (a) PTSD 径向部分幂指数函数示意图。(b)和(c)在不同配位环境中, 对应于 x 轴上标记的不同结构的中心原子的两体 PTSD(S_1 – S_2)的和三体及四体 PTSDs(S_3 – S_6)函数值示意图。结构 1: 直线; 2: 三角形; 3: 正方形; 4: 五边形; 5: 六边形; 6: 四面体; 7: 立方体; 8: 八面体^[47]

Fig. 2 (a) Illustration of the radial part of PTSDs. (b) and (c) PTSDs of different chemical environment. Str1: line; 2: triangle; 3: square; 4: pentagon; 5: hexagon; 6: tetragonal; 7: cubic; 8: octahedron^[47]

包括 PTSD 在内的结构描述符虽然形式不同, 但本质上均为了更好的实现提取结构特征以区分结构的目的, 而结构特征与能量之间的关系则完全交给了机器学习模型。模型的数学形式与势能没有明确的物理意义, 也就是所谓的“黑盒”模型。这类

模型所具有的一个普遍问题是为了实现对所学习的目标性质的精确描述, 需要大量的参数或复杂的结构描述符架构。为了克服这一问题, 理论学家首先将静电能作为突破口, 通过显示的学习原子电荷, 将长程静电相互作用作为能量的一部分与网络

的输出共同作为体系的总能量^[48]。Behler 等最近提出了 4G-HDNNP 方法^[49](4G:第 4 代),用于研究共轭长链有机分子和非中性金属离子团簇^[49],该方法可以通过一种特殊的电荷平衡方案包括非局域静电相互作用。

3 全局势能面训练集的生成

获取代表性数据集的一种简单方法是利用现有数据库。例如,对于有机分子来说,GDB-X^[50]和 QM-X^[51–52]数据集(X:非氢原子的原子数量)已经经过精心设计,包含了大量已知的满足八隅律规则的分子。Smith 等利用正态抽样(NMS)方法置换数据库分子,获得了 1 720 万个结构的巨大数据集^[18]。然而这些结构均局限于靠近势能面底部的稳定结构区域,不包括化学反应涉及到的过渡态等结构信息。

在迄今应用的各种方法中,最受欢迎的仍然是基于分子动力学(MD)的模拟方法,以模拟退火方法为代表^[8, 10, 23–24]。模拟退火方法通过重复加热和冷却循环来探索势能面,并已在人们的工作中广泛使用^[53–55]。增强采样分子动力学如 Metadynamics^[56–57]和 iMD-VR^[58]方法则可以作为产生化学反应相关结构数据的补充方法。Smith 等开发的主动学习方法也用于从 MD 轨迹中选择不同的数据^[59–60]。然而,用于数据生成的 MD 采样存在“短视”问题,因为在低温下,MD 模拟采样到化学反应发生的概率,随反应所具有的能垒呈指数级降低,而且在高温下的模拟轨迹则偏好获得构型熵较高的结构,对稳定结构的采样不足。因此,由此生成的势能面数据通常是非常冗余的,局限在输入结构附近的局部势能面区域。这将不可避免地导致由此获得的机器学习势函数在预测未知材料和反应方面的不足。另一方面,其他的全局优化方法,如盆地跳跃方法^[61],遗传算法^[25]和粒子群优化方法(PSO)等^[26],也在近年来被用于材料的动态机器学习模型训练。这些方法通过忽略最小值之间的过渡区域对势能面进行变换,从而实现快速的全局最小值搜索。他们通常从结构弛豫轨迹获得势能面数据,因此很可能错过过渡区域的关键反应通道。

刘智攀等提出利用随机势能面行走方法(SSW)的全局优化轨迹生成势能面数据集,结果表明该方法能够有效运用在材料结构搜索和化学反应机理预测等问题上^[20–21, 63–69]。SSW 方法最初设计为针对非周期性体系进行全局优化和路径搜索,如分子和团簇结构,之后进一步扩展到可以应用于周期性

体系的结构搜索^[70–72]。与其他具有剧烈结构扰动的全局优化方法相比,SSW 方法利用数值二阶梯度信息(化学键振动模式)高效搜索势能面的低能量区域。同时在这一过程中,SSW 的轨迹包含了过渡态区域的结构特征,这使得发现未知的化学反应成为可能。

在 SSW 方法中,刘智攀等结合了 Metadynamics 方法^[73]和 Metropolis Monte Carlo (MC)方法^[74]的主要特点。Metadynamics 方法是一种用于克服势能面上最小值之间的高势垒的标准技术,如式(11)及图 3 所示所示。其中添加高度为 w_n 的连续偏置势(Gaussian 函数, v_n , $n=1$ 到 H)以修改势能面(V_{mod}),将结构构型从能量极小值点移动到势能面上的高能位置[式(1)中的 R_t^H]。这一过程可分为 3 个步骤,即:1) 优化反应坐标、2) 势能面爬升、3) 局部优化。优化反应坐标可以获得结构的一种随机运动模式($3 \times N$ 维向量, N 为原子数),即式(11)中的 N_t^n ,并根据这一模式添加偏置势函数 v_n 。势能面爬升即在当前的 v_{mod} 势能面上向前运动 d_s ,这也是 SSW 与 Metadynamics 方法最大的不同。SSW 方法是沿着反应坐标单向运动的,而 Metadynamics 方法则是往复运动。局部优化则是在当前的 v_{mod} 势能面上进行局部的结构优化,并最终得到收敛的结构,即式(11)中的 R_t^n ,同时进入下一次迭代($n=n+1$),直至 $n=H$ 。MC 方法则是势能面采样中根据 Boltzmann 分布选

$$V_{\text{mod}} = V_{\text{real}} + \sum_{n=1}^H v_n = V_{\text{real}} + \sum_{n=1}^H w_n \times \exp \left[-\frac{((R_t - R_t^n) \cdot N_t^n)^2}{2 \times d_s^2} \right] \quad (11)$$

$$P = \begin{cases} \exp \left[\frac{E_{\text{new}} - E_{\text{old}}}{RT} \right], & \text{when } E_{\text{new}} > E_{\text{old}} \\ 1, & \text{otherwise} \end{cases} \quad (12)$$

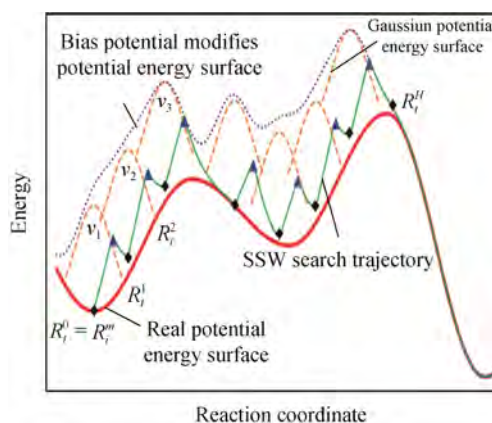


图 3 SSW 方法在一维势能面上的示意图^[73]

Fig. 3 Illustration of SSW method on a one-dimensional potential energy surface^[73]

择状态的常用方法, 如式(12)所示, 当新找到的结构能量更低时将新结构作为下一步 SSW 的起始结构, 否则根据式(12)计算指数函数与随机数进行比较来判断, 式(12)中 E_{new} 、 E_{old} 分别为 SSW 发现的新结构和起始结构的势能, T 为 MC 方法的温度参数, P 为接收概率。

在图 4 中, 对含有 48 个原子的 TiO_2 晶体进行 MD 模拟和变晶胞 SSW 搜索^[20], 并从中看出 MD 和 SSW 方法在势能面采样中的区别。MD 模拟在等温等压(NPT)系综下进行了 0.5 ns 的模拟, 模拟条件为 1 500 K 和 1 个标准大气压。SSW 和 MD 模拟轨迹具有相似的计算代价(5×10^5), 从 2 个方法轨迹分别随机收集了 5 000 个结构, 并行成图 4a 和图 4b 所示的二维势能面等高线图, 其中能量是根据

距离加权的 Steinhardt 结构序参量(OP_2)^[75-76]绘制的。可以看到, SSW 的搜索轨迹访问了势能面的大片区域, 覆盖了几十个不同的晶体相, 而 MD 即使在 1 500 K 也只探索了初始结构周围的局部势能面, 即金红石相。此外, 还利用主成分分析(PCA)对所有数据进行了分析。主成分分析是一种根据数据集的特征对数据进行分类的技术。基于 TiO_2 G-NN 势(可从 LASP 网站访问)中使用的 1 组 Ti 元素的原子结构描述符, PCA 识别出 3 个最易区分的特征, 即 PC1 到 PC3。可以发现, 从 MD 轨迹得到的数据集所包含的化学环境相对局限在较小的范围内。而作为比较, SSW 方法所获得的数据更加多样化, 它涵盖了大部分 MD 数据, 并扩展到属于许多其他固相的原子环境。

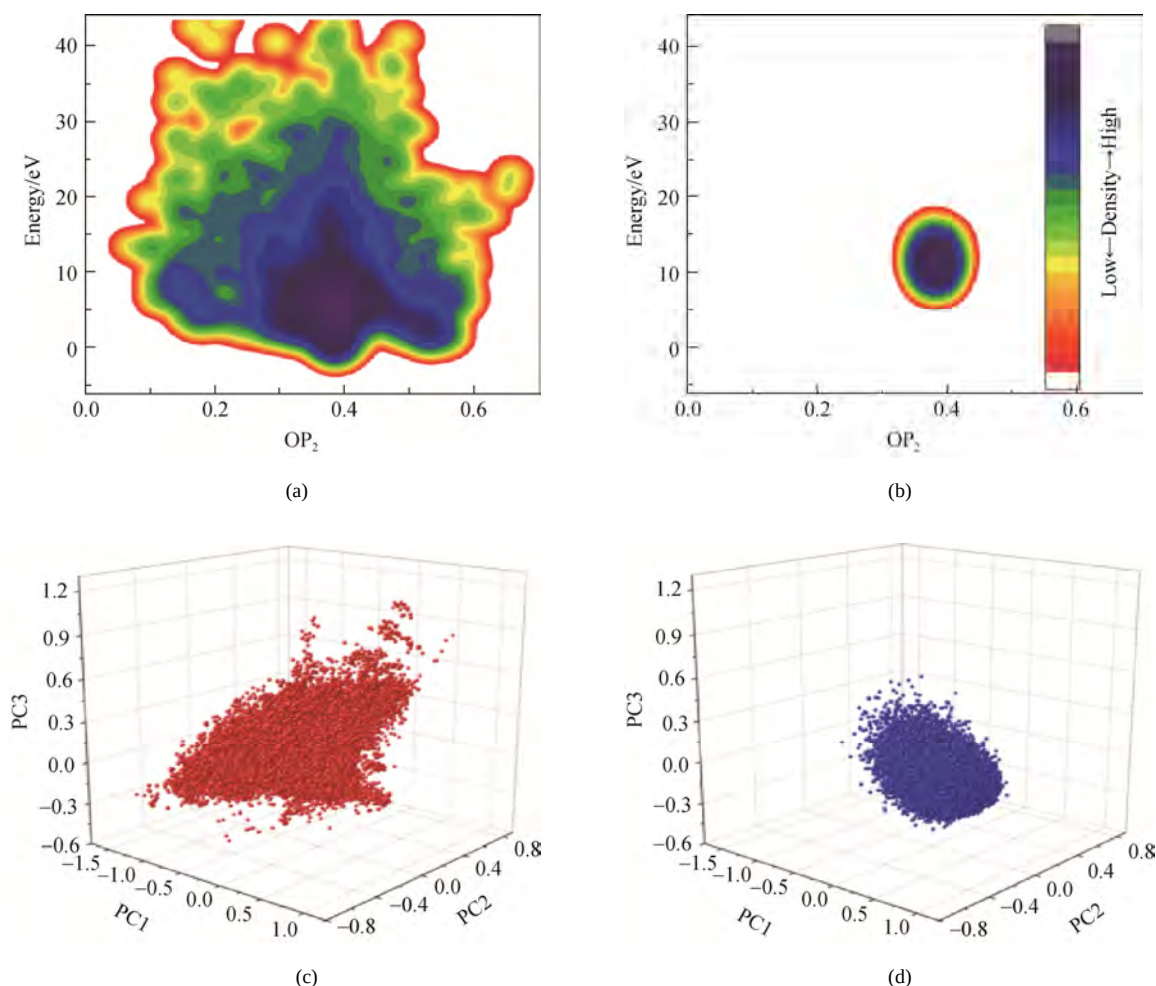


图 4 SSW 方法与 MD 方法产生的数据集的比较. 图(a)和(b)是 SSW 方法和 MD 方法对 TiO_2 体系模拟产生的轨迹在势能面上的投影图。X 轴是距离加权的 Steinhardt 结构描述符(OP_2), y 轴是体系不同结构的能量, 以全局最低点作为能量零点。图 (c)和(d)是将数据集投影在通过主成分分析得到的 Ti 原子对于 G-NN 势函数最有区分度的三个最主要的结构描述符^[47]

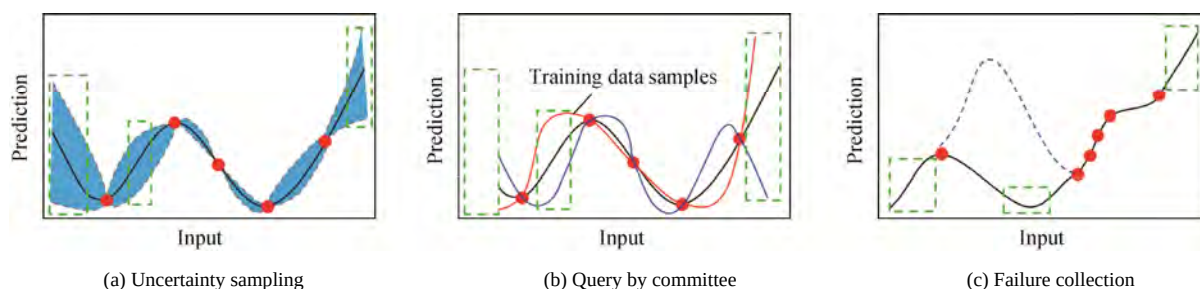
Fig. 4 Comparison of data generated by SSW (a, c) and MD (b, d). (a) and (b) are projected SSW and MD trajectories for TiO_2 system. The x-axis is a distance-weighted Steinhardt order parameter (OP_2), the y-axis is the total energy of the system corresponding to the GM. (c) and (d) are the three most distinct features of Ti atoms in SSW and MD trajectories from principle component analysis (PCA)^[47]

4 主动学习方式自动训练全局神经网络势函数

主动学习的根本目的是为了取代人工产生并筛选数据集,实现高效的势函数生成。主动学习可以遵循 2 种不同的策略 1 基于数据池或基于数据流^[77–79]。基于数据池针对已有的数据集,每次迭代时,分析这些数据的不确定性和多样性,通过数据点相互之间的比较,将不确定性最大或多样性最大的最有价值的数据点进行标记(比如进行一次量子力学能量计算),添加到训练数据集中。基于数据流的策略是动态生成新的数据,并根据预先设定的规则,选取数据点进行标记。这一策略显然非常适合基于机器学习势函数的原子模拟。事实上,2 种策略在一定条件下也可以通用,比如基于流的策略中,如果不是实时做出判断,而是积累一定数据,形成一个临时数据集,就也可以转化为基于数据池的策略。

为了实现高效的主动学习训练,从数据集中选择数据的具体策略就显得尤为关键。目前常用的甄别代

表性数据的方法主要有 3 种:不确定性抽样(Uncertainty sampling)、委员会查询(Query-by-committee)以及失败样本收集(Failure collection)^[59, 77–80]。前 2 种广泛运用于各种机器学习模型中,但是第 3 种方法在产生神经网络势函数中应用的最为广泛。不确定性选择主要针对数据池策略,通过计算所有数据的不确定性,选出不确定性最大的数据点,进行标记(图 5a)。此类方法一般需要对数据点之间两两进行相似性计算,因此适用于训练集规模有限的模型,或高斯过程回归这种原本就需要计算数据点相似性的机器学习模型^[79, 81–82]。委员会查询策略需要建立一个不同机器学习模型的集合,每个模型具有不同的架构或参数,基于整个训练数据集的不同部分进行训练。所有未标记的数据将由委员会的所有模型进行评估并选择偏差最大^[83–85],或落在一定范围内的数据进行标记^[86–87](图 5b)。这一策略的效果取决于“委员会”中机器学习模型的数量,由于每个模型都需要单独训练,因此决定了委员会的数目不会太大,一般局限在 10 个以内。



Red points—Training data samples; Blue, black and red lines—Trained ML Potentials; Green frame—Samples could be selected as labeling candidates; Blue area in Fig. 5a—Uncertainty value of input; Blue dotted line in Fig. 5c—Actual PES.

图 5 3 种不同的主动学习查询策略示意图^[88]

Fig. 5 Schematic illustration of three different AL query strategy in generating ML potentials^[88]

失败样本收集是目前基于神经网络模型最为常用的策略,也是在 LASP 软件^[89]自动寻路势函数过程中使用的策略。这一策略运用阶段性训练的势函数进行模拟,并通过一些结构和能量相关的规则,收集导致模拟失败的关键结构进行标记。对失败事件的判断可以发生在模拟的不同阶段,如描述符生成、模型评估和轨迹分析。当描述符的计算值显著超过当前训练集定义的值范围时,定义为结构生成描述符失败(图 5c)。由于大多数 ML 模型的外推能力比插值能力差得多,因此这种结构的能量预测误差往往远远高于训练精度。另一种类型的失败事件可以通过 PES 的连续性来判断。能量的骤增或骤减,一般意味着结构不合理,在这些结构附近进行分子

动力学模拟可能导致恒温器无法达到平衡,从而破坏能量守恒定律,或者 Hessian 矩阵产生数值过大的本征值。

5 基于神经网络势函数的大规模原子模拟软件的开发及应用

5.1 LASP 软件的开发

为了实现更为便捷的机器学习辅助原子模拟,本课题组开发了“基于神经网络势函数的大尺度原子模拟软件”(LASP)^[47, 89]。软件的主要功能如图 6 所示,包括势能面计算、结构模拟和神经网络势能训练。在势能面计算部分, LASP 提供了神经网络势和标准的数据接口,链接常用势能面计算程序。

结构模拟模块包括基于方法部分中介绍的 SSW 方法的全局结构搜索和路径采样, 可用于结构预测和反应机理研究。以及标准的分子动力学模块。LASP 软件支持运用主动学习方法进行神经网络势函数的自动训练, 首先通过低精度的第一性原理计算, 或根据其他相似体系通过元素变换快速获得初步结构数据集, 并进行高精度第一性原理单点能计算, 得到初始训练集。根据当前训练集进行训练, 获得阶段神经网络势函数。运用阶段势函数进行 SSW 势能面搜索, 并根据失败样本收集策略对 SSW 搜索轨迹中的代表性结构进行收集。对这些结构进行高精度第一性原理单点能计算并扩充训练集。迭代训练-SSW 模拟-结构收集-扩充训练集这一过程, 直到训练集不再增长且势函数精度达到预期为止。

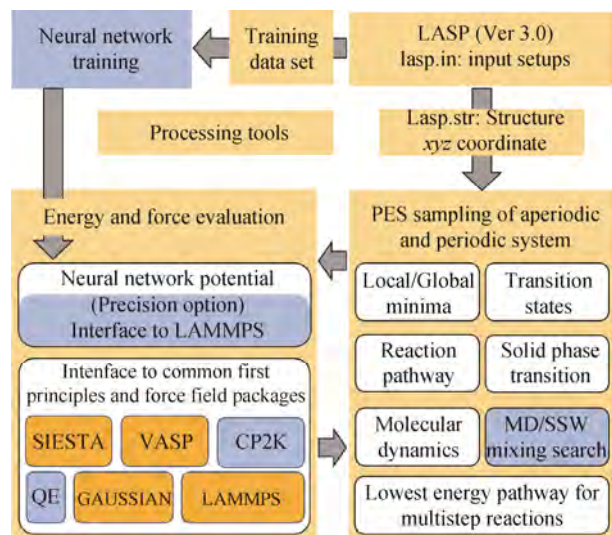


图 6 LASP 软件基本功能示意图^[88]

Fig. 6 Illustration of the main functions of LASP software package^[88]

5.2 通过结构模拟预测材料相图

由于机器学习模型中结构描述符和数值函数的局部表示特点, 基于机器学习势的原子模拟原则上与系统大小呈线性标度, 即使对于中等大小的系统, 其速度也比密度泛函理论快 4 个数量级以上。因此, 机器学习势函数潜在应用的一个主要领域是结构预测, 例如寻找新的晶相, 以及最近机器学习势也被用来加快反应空间探索以寻找可能的反应机理。近年来, 在材料科学、团簇科学和生物科学中有着大量的应用, 下面, 将分别介绍 LASP 软件在结构预测和反应网络探索中的应用。

工业条件下铁基催化剂上原位形成的碳化铁 (FeC_x) 是费托合成 ($\text{CO} + \text{H}_2 \rightarrow \text{长链烃类}$) 的关键活性

组分^[90-92]。尽管该体系已历经 1 个世纪的研究, 由于 FeC_x 结构与 CO 加氢机理非常复杂, 催化活性中心仍存在很大争议。通过使用 SSW-NN 方法, 探索了百万个在费托合成条件下 FeC_x 体相与表面的候选结构, 构建了 FeC_x 的体相热力学凸图, 涵盖不同的 Fe:C 的比率, 找到了具有低表面能的表面, 并评估了 CO 分子和 H 原子在这些表面的吸附能力, 最后确定了 CO 活化的最低能量反应路径。从热力学凸图中我们发现 Fe_5C_2 、 Fe_7C_3 与 Fe_2C 是在费托合成产烯烃条件下稳定的体相, 其中 Fe_7C_3 与 Fe_2C 有许多能量近似简并的晶体相; 在费托合成条件下, FeC_x 表面是动态的, 涉及重构与成分的变化。在上述体相中, 较低表面能的表面通常采用相同的 Fe_2C 表面模式, 其中只有 $\chi\text{-Fe}_5\text{C}_2(510)$ 、 $\chi\text{-Fe}_5\text{C}_2(111)$ 与 $\eta\text{-Fe}_2\text{C}(111)$ 3 个较稳定表面能暴露对 H 原子吸附为放热的 Fe 位点, 其表面 Fe:C 比分别为 2、1.75、2; CO 可以在加氢动态产生的表面碳空位上直接解离, 比如在 $\chi\text{-Fe}_5\text{C}_2(510)$ 上, CO 活化能垒为 1.1 eV^[93]。

5.3 通过结构模拟预测复杂团簇稳定结构

负载型贵金属团簇中一类十分重要的催化剂。早在 20 世纪七八十年代, 化学家就发现当金、铂等贵金属以纳米尺寸分散在氧化物表面时, 可以产生块状材料所不具有的优异催化活性^[94-95]。然而由于材料尺寸太小, 实验上难以对催化中心进行表征, 另一方面, 对理论模拟来说, 由于有载体的存在, 理论模型通常包含上百甚至上千个原子, 传统上基于量子力学的计算难以准确进行结构预测。

运用 LASP 软件, 针对经典的 Au-TiO₂ 体系展开了结构模拟, 针对不同金原子数, 不以及同的金-氧比结构的势能面搜索。针对金红石的(110)面, 探究了 3 种表面情况: 化学计量比表面、缺陷位表面和富氧表面, 并绘制了二维的 $\text{Au}_n\text{O}_m\text{-TiO}_2$ 的势能面图。发现氧化的 TiO₂ 表面可以稳定金的纳米团簇, 避免烧结, 尤其是 Au_{24}O_4 团簇非常稳定。该结构是 C_{2v} 对称的圆顶结构, 在 Au 层和 TiO₂ 表面之间有一个大的空隙。由于其单层的性质, Au 原子仍然处于低配位状态(平均配位数为 4.7)。对于 $\text{Au}_{24}\text{O}_4\text{-TiO}_2$ 存在大量低的未占据 Au 的 5d 态, 表明它们具有存储多余电子的能力。去除表面的晶格 O 后, 可以看到费米能级的上移确实跨越了 2 个系统中连续的 Au 的 5d 态。显然, 这是电子从氧化物表面转移到金上的结果, 导致电子在 Au 的 5d 态上积累。根据这一模型研究了 CO 氧化的反应机理, 并将理论预测的活性与实验中报告的高 CO 氧化活

性进行比较,发现本课题组提出的模型催化剂和不同的实验之间的速率是比较一致的。同时也进行了长时间的MD模拟以直接评估在CO和O₂存在下负载的Au团簇的热稳定性。发现拱形的Au₂₄O₄结构十分稳定,金原子仅在相邻区域之间局部移动,而Au₂₄在800 K时会发生剧烈的结构变形,最终从圆顶转变为紧密堆积的棒状结构,这表明在没有O₂气氛的情况下Au团簇在TiO₂表面上有烧结趋势^[88]。

5.4 通过反应模拟搜索反应路径和机理

长久以来,有机反应一直出于化学的核心地位。近年来,在基于现有有机合成实验数据库的合成路线逆向设计中,机器学习技术取得了令人鼓舞的进展^[96–99]。然而,这些基于数据的方法缺乏对化学反应机理的深入洞察,无法取代原子模拟。为了解决葡萄糖热解反应网络的难题,本课题组最近应用SSW-NN方法生成了一个通用的有机反应数据集。考虑到全局路径可以看作是简单局部模式的组合,SSW-NN数据集生成模拟侧重于收集基础的局部反应模式,最终数据集具有94 854个结构,包含几乎所有C–H–O–N四元素的键合模式。

得益于神经网络势函数的低成本势能面评估优势,从D-葡萄糖出发,深入探索超过120万个局部极小值结构,并收集了超过15万个反应对。对其中的重复反应进行去重并只记录最低的能垒通道后,最终反应数据库包含4 459个独特的分子、6 407个不同的反应和3 488种反应模式。本课题组仔细分析了反应数据库中的路径,确定生成了5-羟甲基糠醛(HMF)的路径,这是实验中观察到的一种重要且有价值的产物^[100–102]。在最低能垒路径中,β-D-吡喃葡萄糖经历开环、异构化、互变异构、脱水和成环生成HMF。总的来说,决速步是烯醇和酮互变异构反应(C→D),能垒为1.91 eV(相对于β-D-吡喃葡萄糖的最低能量构型),比之前的文献中的路径能垒低0.19 eV(2.10 eV,决速步为β消除脱水)。烯醇和酮互变异构反应在总能垒上使这一机理收益,因为其在随后的脱水反应中打开了逆Michael加成反应脱水路线,避免了直接的高能垒β消除脱水^[69]。

5.5 结合反应模拟与微观反应动力学自动预测表面反应机理

与小分子气相反应相比,表面反应具有更为复杂的化学环境,包括丰富的吸附位种类、多样的的中间体构象,也因此具有更为复杂的反应网络。以Cu–ZnO–Al₂O₃催化甲醇合成为例,这是一个已经工业化近60年的重要传统工业过程,但关于催化

的活性位点和甲醇中的碳源仍有争议。最近本课题组基于LASP软件的SSW方法和神经网络势函数,提出了一种微观动力学引导的机器学习反应路径搜索方法(MMLPS)。MMLPS方法采用了分治法思想,利用大规模机器学习原子模拟,结合图论最短路径、微观动力学求解器,对CO₂加氢和CO加氢反应在CuZn多个表面,包括Cu(111)、Cu(211)和Zn合金化的Cu(211),上的可能反应路径,进行了多达数百万结构的采样,从中确认了动力学最低的反应通道。研究发现,在反应条件下金属Zn在Cu(211)台阶位上覆盖度至多到0.22 ML,且Zn–Zn二聚位点不稳定。CO₂和CO加氢仅发生在(211)阶梯表面的台阶边缘,低覆盖的Zn(0.11 ML)对反应动力学影响不大,但较高覆盖度(0.22 ML)的Zn使催化剂中毒。微观动力学模拟显示,CO₂是甲醇合成的主要碳源,主要原因是CO加氢的决速步中间体[CHO]非常不稳定,动力学模拟结果与前人的同位素标记实验的结果一致。虽然Zn在金属阶梯位(活性位点)的直接催化效果并不显著,在工业反应条件下,Cu(111)金属表面可以生长出热力学稳定的[–Zn–OH–Zn–]的链状结构(阳离子Zn),表明了ZnO载体能够部分还原并覆盖金属Cu的主要晶面,这说明ZnO的使用可以有效分散Cu纳米颗粒防止高温失活,同时混合气中CO存在能有效还原Cu表面阳离子Zn来暴露用于甲醇合成的金属位点,而留下的金属Zn促进了CuZn合金形成,并重构出丰富的阶梯位活性位点^[103]。

6 结束语

现有的机器学习势函数已经极大地扩展了原子模拟的时间和空间尺度。本文概述了对于基于神经网络势函数进行原子模拟至关重要的3个方面:1)神经网络势函数模型;2)原子模拟获取高质量训练集;3)基于主动学习的势函数训练。通过将这种方法加一整合成为LASP软件包,在很大程度上简化了神经网络势函数的使用,并加速推广了这些新方法和新技术。运用LASP软件及SSW-NN方法的理论模拟工作证明,基于机器学习势函数的原子模拟可以广泛应用于对材料的结构预测和对化学反应机理的自动探索。未来,通过进一步优化机器学习模型和模拟方法,相信将进一步拓展机器学习势函数的应用场景,加速理论模拟研究周期,进一步扩展原子模拟时间和空间尺度,像接近真实实验条件下的理论模拟这一长远目标迈进。

参考文献:

- [1] HOHENBERG P, KOHN W J P r. Inhomogeneous electron gas[J]. Phys Rev, 1964, 136(3B): B864–871.
- [2] CAR R, PARRINELLO M. Unified approach for molecular dynamics and density-functional theory[J]. Phys Rev Lett, 1985, 55(22): 2471–2474.
- [3] ZHAO Y, TRUHLAR D G. Density Functionals with Broad Applicability in Chemistry[J]. Account Chem Res, 2008, 41(2): 157–167.
- [4] ROUSSEA A, SELLONI A. Theoretical insights into the surface physics and chemistry of redox-active oxides[J]. Nat Rev Mater, 2020, 5(6): 460–475.
- [5] JORDAN M I, MITCHELL T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255–260.
- [6] BLANK T B, BROWN S D, CALHOUN A W, et al. Neural-Network Models of Potential-Energy Surfaces[J]. J Chem Phys, 1995, 103(10): 4129–4137.
- [7] LIU T, FU B, ZHANG D H. Six-dimensional quantum dynamics study for the dissociative adsorption of HCl on Au(111) surface[J]. J Chem Phys, 2013, 139(18): 184705.
- [8] BEHLER J, PARRINELLO M. Generalized neural-network representation of high-dimensional potential-energy surfaces[J]. Phys Rev Lett, 2007, 98(14): 146401.
- [9] BEHLER J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials[J]. J Chem Phys, 2011, 134(7): 074106.
- [10] SOSSO G C, MICELI G, CARAVATI S, et al. Neural network interatomic potential for the phase change material GeTe[J]. Phys Rev B, 2012, 85(17): 174103.
- [11] RUPP M, TKATCHENKO A, MULLER K R, et al. Fast and accurate modeling of molecular atomization energies with machine learning[J]. Phys Rev Lett, 2012, 108(5): 058301.
- [12] BARTÓK A P, KONDOR R, CSÁNYI G. On representing chemical environments[J]. Phys Rev B, 2013, 87(18): 184115.
- [13] KERMOD L R, DE VITA A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces[J]. Phys Rev Lett, 2015, 114(9): 096405.
- [14] BARTÓK A P, CSÁNYI G. Gaussian approximation potentials: A brief tutorial introduction[J]. Int J Quantum Chem, 2015, 115(16): 1051–1057.
- [15] BEHLER J. Perspective: Machine learning potentials for atomistic simulations[J]. J Chem Phys, 2016, 145(17): 170901.
- [16] KOLB B, ZHAO B, LI J, et al. Permutation invariant potential energy surfaces for polyatomic reactions using atomistic neural networks[J]. J Chem Phys, 2016, 144(22): 224103.
- [17] DE S, BARTOK A P, CSANYI G, et al. Comparing molecules and solids across structural and alchemical space[J]. Phys Chem Chem Phys, 2016, 18(20): 13754–13769.
- [18] SMITH J S, ISAYEV O, ROITBERG A E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost[J]. Chem Sci, 2017, 8(4): 3192–3203.
- [19] SCHUTT K T, ARBABZADAH F, CHMIELA S, et al. Quantum-chemical insights from deep tensor neural networks[J]. Nat Commun, 2017, 8: 13890.
- [20] HUANG S D, SHANG C, ZHANG X-J, et al. Material discovery by combining stochastic surface walking global optimization with a neural network[J]. Chem Sci, 2017, 8(9): 6327–6337.
- [21] HUANG S D, SHANG C, KANG P-L, et al. Atomic structure of boron resolved using machine learning and global sampling[J]. Chem Sci, 2018, 9(46): 8644–8655.
- [22] JÄGER M O J, MOROOKA E V, FEDERICI CANOVA F, et al. Machine learning hydrogen adsorption on nanoclusters through structural descriptors[J]. NPJ Comput Mater, 2018, 4(1): 37.
- [23] SCHUTT K T, SAUCEDA H E, KINDERMANS P J, et al. SchNet-A deep learning architecture for molecules and materials[J]. J Chem Phys, 2018, 148(24): 241722.
- [24] ZHANG L, HAN J, WANG H, et al. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics[J]. Phys Rev Lett, 2018, 120(14): 143001.
- [25] JACOBSEN T L, JORGENSEN M S, HAMMER B. On-the-fly machine learning of atomic potential in density functional theory structure optimization[J]. Phys Rev Lett, 2018, 120(2): 026102.
- [26] TONG Q, XUE L, LV J, et al. Accelerating CALYPSO structure prediction by data-driven learning of a potential energy surface[J]. Faraday Discuss, 2018, 211, 31–43.
- [27] HUANG S D, SHANG C, KANG P L, et al. LASP: Fast global potential energy surface exploration[J]. WIREs Comput Mol Sci, 2019, 9(6): e1415.
- [28] FRIEDERICH, HäSE F, PROPPE J, et al. Machine-learned potentials for next-generation matter simulations, Nature Materials, 2021, 20(6), 750–761.
- [29] JINNOUCHI R, MIWA K, KARSAI F, et al. On-the-fly active learning of interatomic potentials for large-scale atomistic simulations[J]. J Phys Chem Lett, 2020, 11(17): 6946–6955.
- [30] COHN D A, GHAMRANI Z, JORDAN M I. Active learning with statistical models[J]. J. Artif Int Res, 1996, 4(1): 129–145.
- [31] TRUHLAR D G, GORDON M S. From force fields to dynamics: Classical and quantal paths[J]. Science, 1990, 249(4968): 491–498.
- [32] CARO M A, DERINGER V L, KOSKINEN J, et al. Growth mechanism and origin of high sp^3 content in tetrahedral amorphous carbon[J]. Phys Res Lett, 2018, 120(16): 166101.
- [33] MOCANU F C, KONSTANTINOU K, LEE T H, et al. Modeling the phase-change memory material, $\text{Ge}_2\text{Sb}_2\text{Te}_5$, with a machine-learned interatomic potential[J]. J Phys Chem B, 2018, 122(38): 8998–9006.
- [34] BALABIN R M, LOMAKINA E I. Support vector machine regression (LS-SVM)—an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data? [J]. Phys Chem Chem Phys, 2011, 13(24): 11710–11718.
- [35] ELAHE A, HOOMAN A, DAVOOD I. Experimental investigation and development of a SVM model for hydrogenation reaction of carbon monoxide in presence of Co–Mo/ Al_2O_3 catalyst[J]. Chem Eng J, 2015, 276: 213–221.
- [36] THOMPSON A P, SWILER L P, TROTT C R, et al. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials[J]. J Comput Phys, 2015, 285: 316–330.
- [37] ZHANG L F, HAN J Q, WANG H, et al. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics[J]. Phys Rev Lett, 2018, 120(14): 143001.
- [38] HAN J, ZHANG L, CAR R, et al. Deep potential: A general representation of a many-body potential energy surface[J]. Commun Comput Phys, 2018, 23(3): 629–639.
- [39] SANCHEZ-LENGELING B, REIF E, PEARCE A, et al. A gentle introduction to graph neural networks[J]. Distill, 2021, 6: e33.
- [40] SCHÜTT K T, ARBABZADAH F, CHMIELA S, et al. Quantum-

- chemical insights from deep tensor neural networks[J]. *Nat Commun*, 2017, 8: 13890.
- [41] SCHUTT K, KINDERMANS P, SAUCEDA H, et al. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions//*Advances in Neural Information Processing Systems*[M]. Curran Associates, Inc., 2017; Vol. 30.
- [42] ZHANG Y, HU C, JIANG B. Embedded atom neural network potentials: Efficient and accurate machine learning with a physically inspired representation[J]. *J Phys Chem Lett*, 2019, 10(17): 4962–4967.
- [43] LIU Y, GUO B, ZOU X, et al. Machine learning assisted materials design and discovery for rechargeable batteries[J]. *Energy Storage Mater*, 2020, 31: 434–450.
- [44] LIU Y, ZOU X, YANG Z, et al. Machine Learning Embedded with Materials Domain Knowledge[J]. *J Chin Ceram Soc*, 2022, 50(3): 863–876.
- [45] KOCER E, KO T W, BEHLER J. Neural network potentials: A concise overview of methods[J]. *Annual Rev Phys Chem*, 2022, 73(1): 163–186.
- [46] UNKE O T, CHMIELA S, SAUCEDA H E, et al. Machine learning force fields[J]. *Chem Rev*, 2021, 121(16): 10142–10186.
- [47] KANG P L, SHANG C, LIU Z P. Large-scale atomic simulation via machine learning potentials constructed by global potential energy surface exploration[J]. *Account Chem Res*, 2020, 53(10): 2119–2129.
- [48] GHASEMI S A, HOFSTETTER A, SAHA S, et al. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network[J]. *Phys Rev B Condensed Matter Mater Phys*, 2015, 92(4): 045131.
- [49] KO T W, FINKLER J A, GOEDECKER S, et al. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer[J]. *Nat Commun*, 2021, 12 (1): 398.
- [50] FINK T, BRUGGESSER H, REYMOND J L. Virtual exploration of the small-molecule chemical universe below 160 Daltons[J]. *Angew Chem Int Ed Engl*, 2005, 44(10): 1504–1508.
- [51] RUDDIGKEIT L, VAN DEURSEN R, BLUM L C, et al. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17[J]. *J Chem Inform Model*, 2012, 52(11): 2864–2875.
- [52] RAMAKRISHNAN R, DRAL P O, RUPP M, et al. Quantum chemistry structures and properties of 134 kilo molecules[J]. *Sci Data*, 2014, 1: 140022.
- [53] KIRKPATRICK S, GELATT C D, VECCHI M P. Optimization by simulated annealing[J]. *Science*, 1983, 220(4598): 671–680.
- [54] PANNETIER J, BASSAS-ALSINA J, RODRIGUEZ-CARVAJAL J, et al. Prediction of crystal structures from crystal chemistry rules by simulated annealing[J]. *Nature*, 1990, 346(6282): 343–345.
- [55] SCHÖN J C, JANSEN M. First step towards planning of syntheses in solid-state chemistry: Determination of promising structure candidates by global optimization[J]. *Angew Chem Int Ed Engl*, 1996, 35(12): 1286–1304.
- [56] GASTEGGER M, MARQUETAND P. High-dimensional neural network potentials for organic reactions and an improved training algorithm[J]. *J Chem Theory Comput*, 2015, 11(5): 2187–2198.
- [57] HERR J E, YAO K, MCINTYRE R, et al. Metadynamics for training neural network model chemistries: A competitive assessment[J]. *J Chem Phys*, 2018, 148(24): 241710.
- [58] AMABILINO S, BRATHOLM L A, BENNIE S J, et al. Training neural nets to learn reactive potential energy surfaces using interactive quantum chemistry in virtual reality[J]. *J Phys Chem A*, 2019, 123(20): 4486–4499.
- [59] SMIT J S, NEBGEN B, LUBBERS N, et al. Less is more: Sampling chemical space with active learning[J]. *J Chem Phys*, 2018, 148(24): 241733.
- [60] SMITH J S, ZUBATYUK R, NEBGEN B, et al. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules[J]. *Sci Data*, 2020, 7(1): 1–10.
- [61] WALES D J, DOYE J P K. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 Atoms[J]. *J Phys Chem A*, 1997, 101(28): 5111–5116.
- [62] ARTRITH N, URBAN A, CEDER G. Constructing first-principles phase diagrams of amorphous Li_xSi using machine-learning-assisted sampling with an evolutionary algorithm[J]. *J Chem Phys*, 2018, 148(24): 241711.
- [63] GUAN S H, SHANG C, HUANG S-D, et al. Two-stage solid-phase transition of cubic ice to hexagonal ice: Structural origin and kinetics[J]. *J Phys Chem C*, 2018, 122(50): 29009–29016.
- [64] KONG F C, LI Y F, SHANG C, et al. Stability and phase transition of cobalt oxide phases by machine learning global potential energy surface[J]. *J Phys Chem C*, 2019, 123(28): 17539–17547.
- [65] HUANG S D, SHANG C, LIU Z P. Ultrasmall Au clusters supported on pristine and defected CeO_2 : Structure and stability[J]. *J Chem Phys*, 2019, 151(17): 174702.
- [66] LI Y F, LIU Z P. Active Site revealed for water oxidation on electrochemically induced $\delta\text{-MnO}_2$: Role of spinel-to-layer phase transition[J]. *J Am Chem Soc*, 2018, 140(5): 1783–1792.
- [67] GUAN S H, ZHANG K X, SHANG C, et al. Stability and anion diffusion kinetics of yttria-stabilized zirconia resolved from machine learning global potential energy surface exploration[J]. *J Chem Phys*, 2020, 152(9): 094703.
- [68] MA S, HUANG S D, LIU Z P. Dynamic coordination of cations and catalytic selectivity on zinc–chromium oxide alloys during syngas conversion[J]. *Nat Catal*, 2019, 2(8): 671–677.
- [69] KANG P L, SHANG C, LIU Z P. Glucose to 5-hydroxymethylfurfural: origin of site-selectivity resolved by machine learning based reaction sampling[J]. *J Am Chem Soc*, 2019, 141(51): 20525–20536.
- [70] SHANG C, LIU Z P. Stochastic surface walking method for structure prediction and pathway searching[J]. *J Chem Theory Comput*, 2013, 9(3): 1838–1845.
- [71] ZHANG X J, SHANG C, LIU Z P. From atoms to fullerene: Stochastic surface walking solution for automated structure prediction of complex material[J]. *J Chem Theory Comput*, 2013, 9(7): 3252–3260.
- [72] SHANG C, ZHANG X J, LIU Z P. Stochastic surface walking method for crystal structure and phase transition pathway prediction[J]. *Phys Chem Chem Phys*, 2014, 16(33): 17845–17856.
- [73] LAIO A, PARRINELLO M. Escaping free-energy minima[J]. *Proceed National Academy Sci United States of America*, 2002, 99(20): 12562–12566.
- [74] METROPOLIS N, ROSENBLUTH A W, ROSENBLUTH M N, et al. Equation of state calculations by fast computing machines[J]. *J Chem Phys*, 1953, 21(6): 1087–1092.
- [75] ZHAN, SHAN C, LIU Z P. Pressure-induced silica quartz amorphization studied by iterative stochastic surface walking reaction sampling[J]. *Phys Chem Chem Phys*, 2017, 19(6): 4725–4733.

- [76] STEINHARDT P J, NELSON D R, RONCHETTI M. Bond-orientational order in liquids and glasses[J]. *Phys Rev B*, 1983, 28(2): 784–805.
- [77] SETTLES B. Active Learning Literature Survey, Computer Sciences Technical Report 1648[R]. 2009, University of Wisconsin Madison.
- [78] TIAN Y, LOOKMAN T, XUE D. Efficient sampling for decision making in materials discovery[J]. *Chin Phys B*, 2021, 30(5): 050705.
- [79] GUAN Y, YANG S, ZHANG D H. Construction of reactive potential energy surfaces with Gaussian process regression: active data selection[J]. *Mol Phys*, 2018, 116(7): 823–834.
- [80] PODRYABINKIN E V, SHAPEEV A V. Active learning of linearly parametrized interatomic potentials[J]. *Comput Mater Sci*, 2017, 140: 171–180.
- [81] PROPPE J, GUGLER S, REIHER M. Gaussian process-based refinement of dispersion corrections[J]. *J Chem Theory Comput*, 2019, 15(11): 6046–6060.
- [82] TANG Y H, DE JONG W A. Prediction of atomization energy using graph kernel and active learning[J]. *J Chem Phys*, 2019, 150(4): 044107.
- [83] IMBALZANO G, ZHUANG Y, KAPIL V, et al. Uncertainty estimation for molecular dynamics and sampling[J]. *J Chem Phys*, 2021, 154(7): 074102.
- [84] LIN Q, ZHANG Y, ZHAO B, et al. Automatically growing global reactive neural network potential energy surfaces: A trajectory-free active learning strategy[J]. *J Chem Phys*, 2020, 152(15): 154104.
- [85] LIN Q, ZHANG L, ZHANG Y, et al. Searching Configurations in Uncertainty Space: Active Learning of High-Dimensional Neural Network Reactive Potentials[J]. *J Chem Theory Comput*, 2021, 17(5): 2691–2701.
- [86] ZHANG Y, WANG H, CHEN W, et al. DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models[J]. *Comput Phys Commun*, 2020, 253: 107206.
- [87] ZHANG L, LIN D Y, WANG H, et al. Active learning of uniformly accurate interatomic potentials for materials simulation[J]. *Phys Rev Mater*, 2019, 3(2): 023804.
- [88] PENG Y, SHANG C, LIU Z P. The dome of gold nanolized for catalysis[J]. *Chem Sci*, 2021, 12(15): 5664–5671.
- [89] SHANG C, LIU Z P. Chapter 14-Constructing machine learning potentials with active learning, *Quantum Chemistry in the Age of Machine Learning*, Elsevier, 2023, 313–327.
- [90] HUANG S D, SHANG C, KANG P-L, et al. LASP: Fast global potential energy surface exploration[J]. *Comput Mol Sci*, 2019, 9: e1415.
- [91] ECKSTROM H C, ADCOCK W A. A New Iron Carbide in Hydrocarbon Synthesis Catalysts[J]. *J Am Chem Soc*, 1950, 72(2): 1042–1043.
- [92] HOFER L J E, COHN E M, PEEBLES W C. The Modifications of the Carbide, Fe₂C-Their Properties and Identification[J]. *J Am Chem Soc*, 1949, 71(1): 189–195.
- [93] LIU Q Y, SHANG C, LIU Z P. *In situ* active site for Fe-catalyzed fischer–tropsh synthesis: Recent progress and future challenges[J]. *J Phys Chem Lett*, 2022, 13(15): 3342–3352.
- [94] HARUTA M. Size- and support-dependency in the catalysis of gold[J]. *Catal Today*, 1997, 36(1): 153–166.
- [95] HARUTA M, TSUBOTA S, KOBAYASHI T, et al. Low-temperature oxidation of CO over gold supported on TiO₂, α -Fe₂O₃, and Co₃O₄[J]. *J Catal*, 1993, 144(1): 175–192.
- [96] SEGLER M H S, PREUSS M, WALLER M P. Planning chemical syntheses with deep neural networks and symbolic AI[J]. *Nature*, 2018, 555(7698): 604–610.
- [97] COLEY C W, GREEN W H, JENSEN K F. Machine Learning in Computer-Aided Synthesis Planning[J]. *Account Chem Res*, 2018, 51(5): 1281–1289.
- [98] COLEY C W, JIN W, ROGERS L, et al. A graph-convolutional neural network model for the prediction of chemical reactivity[J]. *Chem Sci*, 2019, 10(2): 370–377.
- [99] SCHWALLER P, LAINO T, GAUDIN T, et al. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction[J]. *ACS Cent Sci*, 2019, 5(9): 1572–1583.
- [100] FANG Y, LI J, CHEN Y, et al. Experiment and modeling study of glucose pyrolysis: Formation of 3-hydroxy- γ -butyrolactone and 3-(2h)-furanone[J]. *Energy Fuels*, 2018, 32(9): 9519–9529.
- [101] PATWARDHAN P R, SATRIO J A, BROWN R C, et al. Product distribution from fast pyrolysis of glucose-based carbohydrates[J]. *J Anal Appl Pyrolysis*, 2009, 86(2): 323–330.
- [102] MAYES H B, NOLTE M W, BECKHAM G T, et al. The alpha-bet(a) of glucose pyrolysis: Computational and experimental investigations of 5-hydroxymethylfurfural and levoglucosan formation reveal implications for cellulose pyrolysis[J]. *ACS Sustain Chem Eng*, 2014, 2(6): 1461–1473.
- [103] SHI Y F, KANG P L, SHANG C, et al. Methanol synthesis from CO₂/CO mixture on Cu–Zn catalysts from microkinetics-guided machine learning pathway search[J]. *J Am Chem Soc*, 2022, 144(29): 13401–13414.