计算机科学
COMPUTER SCIENCE

优化器对神经网络力场性能的影响与分析

李恩吉, 胡思宇, 谭光明, 贾伟乐

引用本文

李恩吉, 胡思宇, 谭光明, 贾伟乐. 优化器对神经网络力场性能的影响与分析[J]. 计算机科学, 2025, 52(5): 50-57.

LI Enji, HU Siyu, TAN Guangming, JIA Weile. Impact and Analysis of Optimizers on the Performance of Neural Network Force Fields [J]. Computer Science, 2025, 52(5): 50-57.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241100176

优化器对神经网络力场性能的影响与分析

李恩吉　胡思宇　谭光明　贾伟乐
中国科学院计算技术研究所处理器芯片全国重点实验室　北京 100190
中国科学院大学　北京 100190
(lienji23s@ict.ac.cn)

摘　要　分子动力学模拟是一种广泛应用于多个学科(如材料科学、计算化学等)的关键研究方法。近年来,随着计算能力的提升、神经网络模型的发展以及第一性原理数据的增加,神经网络力场模型已经展现出高精度的预测能力。目前存在多种神经网络力场模型的训练算法,而神经网络力场模型处于一个快速迭代的阶段,当前仍然缺乏神经网络力场模型及与之适配的优化器的指导建议。选取3种有代表性的神经网络力场模型和目前3种用于神经网络力场模型上的优化算法,在4个真实数据集上进行测试和评估,分析影响其收敛性的原因。设计实验对其进行全方位的评估,包括模型参数量对优化器的影响,神经网络宽度对收敛性的影响,以及模型训练时间与优化器的关联等。文中工作可以针对神经网络力场模型,给出优化器算法的建议。

关键词:分子动力学模拟;神经网络;力场训练;优化器
中图分类号　TP391

Impact and Analysis of Optimizers on the Performance of Neural Network Force Fields

LI Enji, HU Siyu, TAN Guangming and JIA Weile
State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
University of Chinese Academy of Sciences, Beijing 100190, China

Abstract　Molecular dynamics (MD) simulation is widely used in various fields, such as materials science and computational chemistry. In recent years, with the improvement in computational power, the development of neural network models, and the accumulation in first-principle data, neural network force field (NNFF) models have demonstrated high predictive accuracy. Currently, there are multiple training algorithms available for NNFF models, and these models are undergoing rapid iteration. However, there remains a lack of guidance on NNFF models and their compatible optimizers. This paper selects three representative NNFF models and the three most commonly used optimization algorithms for these models, testing and evaluating them on four real-world datasets to analyze factors affecting their convergence. We have designed numerous experiments for a comprehensive evaluation, including the impact of model parameter size on the optimizer, the influence of model depth and width on convergence, and the relationship between model training time and the optimizer. Our work provides recommendations for optimizer algorithms specific to NNFF models.

Keywords　Molecular dynamics simulations, Neural networks, Force field, Optimizer

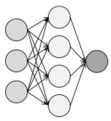
1　引言

分子动力学(Molecule Dynamic, MD)模拟,是一种广泛应用于各学科(如化学工程、生物医学、材料科学工程、物理等)的关键研究手段。其主要用于探究分子体系的结构特征,也可用于研究体系的热力学特性,还可研究分子体系的速度自相关函数,进而计算体系的均方位移、扩散系数等性质。分子动力学模型包括:基于物理模型的分子动力学和基于数学模型的分子动力学。其中,基于物理模型的分子动力学包括第一性原理分子动力学(Ab Initio MD, AIMD)和势函数法。基于数学模型的分子动力学主要指机器学习力场。3类分子动力学模拟的特点如表1所列。总体来说,分子动力学模拟朝着更高的效率、更高的精度、更大的体系、更长的演化时间发展。

到稿日期:2024-11-28　返修日期:2025-03-03
基金项目:中国科学院战略性先导科技专项资助(XDB0500102);国家自然科学基金(92270206, T2125013, 62372435, 62032023, 61972377, 61972380, T2293702);中国科学院青年基础研究项目(YSBR-005);国家创新人才博士后计划(BX20240383)
This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences(XDB0500102), National Natural Science Foundation of China(92270206, T2125013, 62372435, 62032023, 61972377, 61972380, T2293702), CAS Project for Young Scientists in Basic Research(YSBR-005) and China National Postdoctoral Program for Innovative Talents(BX20240383).
通信作者:贾伟乐(jiaweile@ict.ac.cn)

表 1 3 种模型的分子动力学模拟的特点

Table 1 Characteristics of molecular dynamics simulations of the three models

模型类型	图示	空间尺度/m	时间尺度/s	理论方法
量子力学模型		1×10^{-10}	1×10^{-15}	量子力学
经典力学模型		1×10^{-9}	1×10^{-9}	牛顿力学
神经网络分子动力学模型		1×10^{-9}	1×10^{-9}	机器学习

第一性原理方法是基于量子力学的基本理论，不依赖于任何经验数据来模拟微观尺度的物理现象的计算方法。其利用量子力学中的基本方程，即薛定谔方程，来计算电子的行为和原子间的相互作用。基于密度泛函理论[1]来求解 Kohn-Sham 方程[2]，是一种有代表性的第一性原理计算方法。该方法已被应用于很多电子结构计算软件中，例如 VASP[3]，Quantum Espresso[4]，PWmat[5]等。第一性原理计算涉及到对复杂系统中所有电子的量子态进行精确计算，具有较高的精度和较高的计算复杂度（一般复杂度为 $O(N^3)$），目前只能处理皮秒（ps）尺度的运动和最多数千原子量级的规模。时空尺度的约束，限制了第一性原理方法在更广泛实际场景下的应用。

势函数方法不再涉及量子力学问题的求解，简称经典势。大部分经典势的计算复杂度都是线性标度的，即 $O(N)$。势函数一般写作多个粒子坐标的解析函数，输出是体系的势能，常见的势函数法包括 Lennard-Jones[6]，Embedded Atom Method[7]，Reactive Force Field[8]，Reactive Empirical Bond Order[9]和 Tersoff[10]势等。势函数方法的精度取决于人们对目标体系的了解程度，对于不同的体系需要选择合适的原子间势及其参数（如键、角、二面角等）。目前势函数法的主要问题为精度问题。

机器学习力场的基本概念最早由 Behler 和 Parrinello 提出，他们采用全连接神经网络作为其机器学习力场模型。机器学习力场的一个重要假设是系统的总能量被表达成各个原子的势能之和，同时每一个原子的势能都是同一神经网络力场模型的输出。早期的机器学习力场模型，采用了一系列对称函数得到原子描述符，并将其作为神经网络的输入。描述符一般是原子坐标和类型的函数，且该函数需满足空间的平移、旋转、置换不变性或等变性。描述符用于捕捉原子的局部化学环境。机器学习力场是一个原子局部化学环境到原子势能的映射。主流的机器学习力场模型包括：基于余弦基组的神经网络[11]、SNAP[12]、SIMPLE-NN[13]、HDNNP[14]、DeePMD[15]、NEP[16]、BPNN[17]、ANI[18]、PAINN[19]、NequIP[20]、NewtonNet[21]等。随着人工智能模型的发展和大量高精度第一性原理数据的计算，近年来提出的机器学习力场方法可以在接近第一性原理方法的精度的同时，保持与势函数方法相似的低计算代价。例如，基于机器学习力场的分子动力学可以将第一性原理精度分子动力学的规模扩展到上亿原子[22]。

如图 1 所示，近年来学术界对机器学习力场的研究成果呈指数增长的趋势（数据于 2024 年 3 月 26 日提取于 Web of Science）。现如今，机器学习力场模型处于蓬勃发展的阶段。精度和收敛速度是评价机器学习力场模型性能的关键指标，快速训练得到一个高精度的机器学习力场模型极为重要。神经网络的训练过程本质上是权重更新的过程，不同的权重更新算法（优化器）的收敛性不同。运用于神经网络力场中的优化算法具有多样性，如典型的梯度下降法，其权重的更新沿着负梯度方向以一定的步长移动；再如拟牛顿法、基于扩展卡尔曼滤波理论的权重更新算法以及遗传算法等。本文在多种真实数据集上对不同的神经网络力场模型及其适用的优化器算法做了测试和分析，并给出一般性的结论。

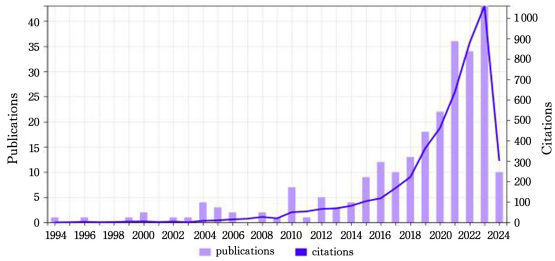


图 1 机器学习力场已发表的论文数和引用量

Fig. 1 Number of published papers and citations in the field of machine learning

本文第 2 章介绍有代表性的神经网络力场模型和优化器算法；第 3 章展示不同的优化器在不同的神经网络力场模型中的表现；第 4 章根据测试结果对神经网络结构（包括网络结构、网络深度和宽度等）与优化器的影响进行分析，并探究其原因；最后总结全文并展望未来。

2 背景

2.1 神经网络模型

本文选取基于高斯基组的神经网络力场模型、深度势能模型（Deep Potential Model）、遗传网络力场模型（Neuroevolution Potentials Model）进行分析。这 3 种神经网络分别于 2007 年、2018 年和 2019 年被提出。第一种网络代表了神经网络力场模型最早期的网络，手动设计解析式作为描述符，采用多层的全连接神经网络。第二种网络在多种实际应用中取得了较高的拟合精度，无需手动设计解析表达式，参数量朝着逐渐增加的趋势发展。第三种网络结合模拟的精度和效率，回归到早期的手动设计特征的模式，并朝着比第一种模式下更多的参数量方向发展，以追求极致的模拟效率。

2.1.1 基于高斯基组的神经网络力场模型

首先构造高斯基组。高斯基组分为两体高斯基组和三体高斯基组，其构造方式如下：

$$G_i=\sum_{j\neq i}\mathrm{e}^{-\eta(R_{ij}-R_s)^2}f_c(R_{ij}) \tag{1}$$

其中，η 和 R_s 是用户定义的参数。

$$G_i=2^{1-\zeta}\sum_{j,k\neq i}(1+\lambda\cos\theta_{ijk})^{\zeta}\mathrm{e}^{-\eta(R_{ij}^2+R_{ij}^2+R_{ij}^2)}f_c(R_{ij})f_c(R_{ij})f_c(R_{ij}) \tag{2}$$

其中：

$$\cos\theta_{ijk}=\frac{R_{ij}\cdot R_{jk}}{|R_{ij}\|R_{jk}|} \tag{3}$$

通常，两体高斯基组和三体高斯基组会同时产生描述符作为神经网络的输入。如图 2 所示，每一个原子根据给定截断半径下的邻居原子，基于用户给定的超参数，即可算出两体

和三体高斯基组的向量表示。将一个原子的高斯基组向量表示输入全连接神经网络，即可计算得到该原子的能量。遍历体系中的各个原子，即可得到体系的总能量。

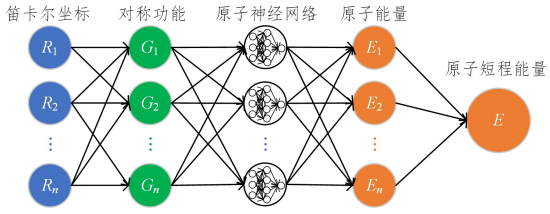


图 2 基于高斯基组的神经网络力场模型示意图

Fig. 2 T Schematic diagram of a neural network force field model based on Gaussky groups

2.1.2 深度势能模型

深度势能模型通常包括一个嵌入网络(Embedding Network)和一个拟合网络(Fitting Network)，如图 3 所示。嵌入网络将原子的局部环境映射到一个特征空间，再通过切比雪夫多项式进行维度的扩充，得到满足平移、旋转和置换不变性的描述符，而拟合网络则将这些描述符映射到原子能量上。

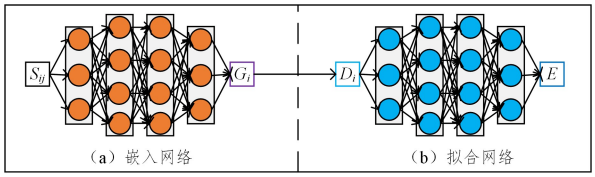


图 3 深度势能模型的网络结构示意图

Fig. 3 Schematic diagram of the network structure of the deep potential model

首先生成中心原子的局部环境矩阵$\widetilde{\boldsymbol{R}}_i$：中心原子的局部环境矩阵$\widetilde{\boldsymbol{R}}_i$由$\boldsymbol{R}_i$变换得到，$\boldsymbol{R}_i$是一个$N_m\times 4$的矩阵，其中$N_m$是所有原子的邻居列表的最大长度，在实际应用中通常指定为 100。根据$\boldsymbol{R}_i$生成$\widetilde{\boldsymbol{R}}_i$的过程为：针对中心原子 i 的每一个邻居原子 j，执行式(4)的变换：

$$\begin{aligned}&(x_{ji},y_{ji}z_{ji})\to(s(r_{ji}),\widehat{x_{ji}},\widehat{y_{ji}},\widehat{z_{ji}})\\&\widehat{x_{ji}}=\frac{s(r_{ji})x_{ji}}{r_{ji}}\\&\widehat{y_{ji}}=\frac{s(r_{ji})y_{ji}}{r_{ji}}\\&\widehat{z_{ji}}=\frac{s(r_{ji})z_{ji}}{r_{ji}}\end{aligned}\tag{4}$$

其中，$s(r_{ji})$为平滑的距离的倒数函数，$f_c(r)$函数为输入数据做平滑处理操作：

$$s(r_{ji})=\frac{f_c(r_{ji})}{r_{ji}}\tag{5}$$

$$f_c(r_{ji})=\begin{cases}1, & r_{ji}\leqslant R_{c2}\\ \dfrac{1}{2}\cos\left(\pi\dfrac{r_{ij}-R_{c2}}{R_{ij}-R_{c2}}\right)+\dfrac{1}{2}, & R_{c2}\leqslant r_{ji}\leqslant R_c\\ 0, & r_{ji}\geqslant R_c\end{cases}\tag{6}$$

其中，R_{c2}是一个平滑的截断参数，它允许在由 R_{c2}到R_c之间的区域平滑地将 $s(r_{ji})$ 减小到零。

接下来生成中心原子的嵌入矩阵$\boldsymbol{G}_i$。由 $s(r_{ji})$生成$\boldsymbol{G}_i$由嵌入网络(Embedding Net)完成，记为G_i，其表达式为：

$$G^{\alpha_j,\alpha_i}(s(r_{ji}))\tag{7}$$

其中，α_i和α_j表示原子 i 和原子 j 的化学元素类型，即由中心原子和其任一邻居所组成的原子对的每一种元素类型组合对应一组 G 的参数。在化学元素类型确定的情况下，神经网络 G 简写为：

$$G(S_{ji})\tag{8}$$

神经网络 G 的输入是标量$s(r_{ji})$，输出是 $M1$ 维向量。通过将$\widetilde{\boldsymbol{R}}_i$的第一列即$s(r_{ji})$列依次输入网络 $\boldsymbol{G}$，即可得到 $N\times M1$ 维的嵌入矩阵$\boldsymbol{G}_i$。

然后，生成中心原子的局部环境描述符$\boldsymbol{D}_i$。在得到嵌入矩阵$\boldsymbol{G}_i$后，首先取$\boldsymbol{G}_i$的前 $M2$ 列，通常 $M2\leqslant M1$，将其称为$\boldsymbol{G}^{i2}$，原$\boldsymbol{G}_i$矩阵称为$\boldsymbol{G}^{i1}$，然后按照式(9)生成$\boldsymbol{D}_i$：

$$\boldsymbol{D}_i=(\boldsymbol{G}^{i1})^{\mathrm{T}}\widetilde{R}_i(\widetilde{R}_i)^{\mathrm{T}}\boldsymbol{G}^{i2}\tag{9}$$

其中，$\boldsymbol{D}_i$是 $M1\times M2$ 维的矩阵。式(9)保证了$\boldsymbol{D}_i$对原子体系构型的平移、旋转以及交换不变性。将矩阵$\boldsymbol{D}_i$拉平为向量后输入拟合网络，即可拟合得到中心原子的势能E_i，如图 3(b)所示。

2.1.3 NEP 模型

NEP 采用单个隐藏层的前馈神经网络得到能量，输入的描述符向量的长度记为N_{neu}。NEP 的网络示意图如图 4 所示。NEP 可用数学符号表示为：

$$U_i=\sum_{\mu=1}^{Neu}\boldsymbol{\omega}_{\mu}^{(1)}\tanh\left(\sum_{v=1}^{N_{\mathrm{des}}}\boldsymbol{\omega}_{\mu v}^{(0)}q_v^i-\boldsymbol{b}_{\mu}^{(0)}\right)-\boldsymbol{b}^{(1)}\tag{10}$$

其中，$\tanh(a)$是隐藏层中的激活函数；$\boldsymbol{\omega}^{(0)}$是从输入层(描述符向量)到隐藏层的连接权重矩阵；$\boldsymbol{\omega}^{(1)}$是从隐藏层到输出节点的连接权重向量；输出节点得到能量U_i；$\boldsymbol{b}^{(0)}$是隐藏层中的偏置向量；$\boldsymbol{b}^{(1)}$是节点U_i的偏置。

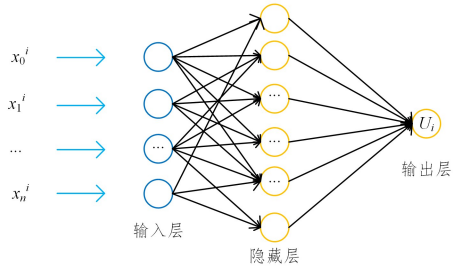


图 4 NEP 网络结构

Fig. 4 NEP network structure

NEP 的描述符包括径向描述符和角度描述符[16]。q_n^i为角度描述符，$g_n(r_{ij})$为径向描述符。

$$q_n^i=\sum_{j\neq i}g_n(r_{ij})\ \text{ with }0\leqslant n\leqslant n_{\max}^R\tag{11}$$

$$g_n(r_{ij})=\sum_{\mu=1}^{N_{\mathrm{bas}}^R}c_{nk}^{ij}f_k(r_{ij})\tag{12}$$

$$f_k(r_{ij})=\frac{1}{2}[T_k(2(r_{ij}/r_c^R-1)^{2-1})+1]f_c(r_{ij})\tag{13}$$

$$f_c(r_{ij})=\begin{cases}\dfrac{1}{2}\left[1+\cos\left(\pi\dfrac{r_{ij}}{r_c^R}\right)\right], & r_{ij}\leqslant r_c^R\\ 0, & r_{ij}>r_c^R\end{cases}\tag{14}$$

其中，求和操作作用于原子 i，在给定截止距离 r 内的所有邻居；$n_{\max}^R$和 n_{bas}^R为用户自定义的超参数；$T_k(x)$是 k 阶切比雪夫多项式；$f_c(r_{ij})$为截止函数；r_c^R是径向描述符分量的截断距离；系数c_{nk}^{ij}取决于 n 和 k，也取决于原子 i 和 j 的类型。

2.2 优化器

优化器负责神经网络权重的更新。权重更新时，已知第 t 次迭代的权重，根据当前批次样本可以经过神经网络的前向

计算得到损失值，再根据损失值进行反向传播，得到权重更新的增量。结合第 t 次的权重，即可获得第 $t+1$ 次迭代的权重值，如图 5 所示。

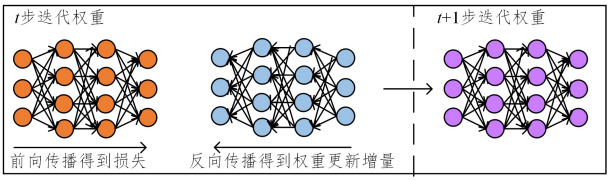


图 5 权重更新示意图

Fig. 5 Schematic diagram of weight updates

接下来介绍 3 种神经网络力场模型中用到的优化器：Adam 优化器[23]、LKF 优化器[24-25]、遗传算法[26-27]。

2.2.1 Adam 优化器

Adam 优化器是基于梯度下降的一阶优化算法，目前在多种任务和模型上都表现出良好的性能，包括图像分类、自然语言处理和其他复杂的机器学习任务。在神经网络力场模型中，Adam 优化器也是应用最为广泛的一种优化器。其使用偏差校正项来调整一阶矩和二阶矩的估计，有助于加速收敛并减少震荡。

2.2.2 KF 优化器

卡尔曼滤波(Kalman Filter，KF)是鲁道夫·卡尔曼在 1960 年提出的一种数据处理算法。其目标是通过估计器最优来估计线性过程的状态，基于具有高斯噪声的状态测量模型，旨在最小化预测状态的均方误差，以噪声测量作为输入。由于卡尔曼滤波具有快速收敛和有效的噪声过滤特性，因此被广泛应用于自主系统、导航和交互式计算机图形学等领域。

然而，KF 的公式仅限于线性优化问题。对于非线性系统，通过对非线性的测量模型和过程模型进行泰勒展开，可以将其线性化，从而推广为扩展卡尔曼滤波(Extended Kalman Filter，EKF)。神经网络是典型的非线性系统，当状态值映射到神经网络的权重时，全局卡尔曼滤波(Global Extended Kalman Filter，GEKF)可用于神经网络权重的更新。

重组卡尔曼滤波(Reorganized Layer-wised Extended Kalman Filter，LKF)优化器是在全局卡尔曼滤波优化器的基础上，通过一种特殊的解耦合方法加以改进。给定一个块大小阈值 N_b，当神经网络层的参数量小于该阈值时，聚合多个小的神经网络层；当神经网络的参数量大于该阈值时，将该层按照阈值进行拆分。

2.2.3 遗传算法

遗传算法(Genetic Algorithm)作为神经网络力场模型的训练算法，其工作流程如下。1)参数初始化。权重的个数记为 N_{par}，权重的随机初始化服从均值向量为 m 和标准差向量为 s 的分布。2)进行 N_{gen} 次迭代。首先根据当前 m 和 s 创建一簇解，记作：$z_k(1\leqslant k\leqslant N_{pop})$。

$$z_k \leftarrow m+s\odot r_k \tag{15}$$

其中，$\odot$ 表示逐元素乘，N_{pop} 是这一簇解的数目，r_k 是一组服从均值为 0、方差为 1 的高斯分布的随机数。在生成不同的解 z_k 时，r_k 的取值不同；采用损失函数 $L(z_k)$ 对每一个解 z_k 进行评估，并根据损失函数从小到大进行排序；按照如下方式更新自然梯度：

$$\nabla_m J \leftarrow \sum_{k=1}^{N_{pop}} u_k r_k \tag{16}$$

$$\nabla_s J \leftarrow \sum_{k=1}^{N_{pop}} u_k (r_k \odot r_k - 1) \tag{17}$$

其中，u_k 用于将当前解朝着损失更低的方向演变。更新 m 和 s：

$$m \leftarrow m+\eta_m (s\odot \nabla_m J) \tag{18}$$

$$s \leftarrow s\odot \exp\left(\frac{\eta_s}{2}\nabla_s J\right) \tag{19}$$

其中，η_m 和 η_s 可以视为 m 和 s 的学习率。

3 实验与结果分析

数据集：神经网络力场的核心是利用高精度的第一性原理计算数据并利用深度学习网络拟合高维的势函数。经前期调研，本文中的数据均使用 PWmat 软件[28]生成。首先在 The Materials Project 下载 Ag，Cu，C，H2O 等分子的结构文件，其可视化展示如图 6 所示。通过 PWmat 生成 MOVEMENT 文件，我们对这 4 种数据集分别生成了 2 015，1 646，4 000，4 000 个样本。将 MOVEMENT 文件中前 80% 的数据作为训练集，后 20% 的数据作为测试集。然后，根据不同神经网络力场模型的需求生成相应的描述符，并进行训练。

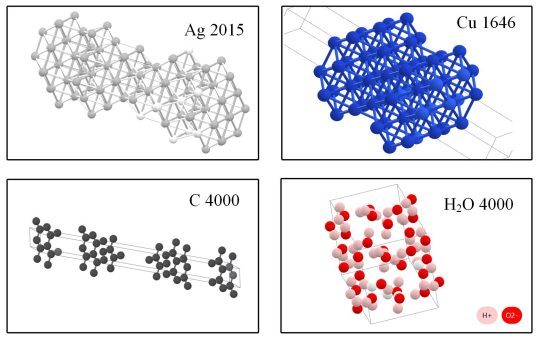


图 6 体系可视化

Fig. 6 System visualization

实验环境：使用了配备 80GB 显存的 A100 GPU 测试。PWMLFF 代码[1)](版本 ID：14b264a)在实现过程中调用了 MKL(Math Kernel Library)2022.0.2 版本的数学函数库，还调用了 numpy，pandas，scikit-learn-intelex 等包。代码的编译工作由 GCC 8.0 完成。基于高斯基组的神经网络力场模型[29]、深度势能模型、NEP 均基于深度学习框架 Pytorch 2.2.0。

参数配置：基于高斯基组的神经网络力场模型采取 3 层全连接神经网络，每一层的神经元个数为 15，激活函数为 tanh，默认的优化器为 Adam。深度势能模型包含嵌入网络和拟合网络，文中设计 3 组不同大小的网络参数，其维度如表 2 所列，其默认的优化器为 Adam。

表 2 不同规模的深度势能模型的参数

Table 2 Parameters of depth potential energy models at different scales

scales	嵌入网络	拟合网络
small	[25,25,25]	[50,50,50,1]
middle	[25,50,100]	[120,120,120,1]
large	[25,50,120]	[240,240,240,1]

1) https://github.com/LonxunQuantum/PWMLFF

NEP 模型为单隐藏层的神经网络力场模型，隐藏层的神经元个数默认值为 100，其默认的优化器为遗传算法，其中平均值矢量 m 的各个分量可在 $-1/2$ 到 $1/2$ 之间均匀地取随机值，方差矢量 s 的每一个分量都取常数值 0.1。

3.1 同一优化器在多种网络上的表现

对于高斯基组的神经网络力场模型(Gua-NN)、深度势能模型(DP)和基于径向函数和角度函数的 NEP 力场模型，分别采用 Adam 优化器、LKF 优化器和遗传算法进行训练。NEP 中网络参数更新的默认算法为遗传算法，但遗传算法中有较多的超参数待指定，直接复用 NEP 中的超参数设置，Gua-NN 和 DP 在这 4 个数据集上无法收敛。不同的网络结构采用遗传算法时，还需对参数继续微调。因此，暂时评估 Adam 和 LKF 在 3 个网络上的收敛效果，其在测试集上的逐原子能量(eV)和原子受力(eV/Å)的收敛误差(RMSE)分别参考表 3 和表 4。表 3 和表 4 中采用 Adam 优化器的收敛结果是在各个模型上经过 300 个 epoch 训练后评估的，采用 LKF 优化器的收敛结果是各个模型经过 30 个 epoch 训练完评估的。这是由于 Adam 优化器是一种梯度下降算法，随着学习率的降低，收敛逐渐变慢，300 个 epoch 是这些体系下的经验参数。LKF 是一种拟牛顿法，每一次迭代的计算开销比 Adam 大，但收敛得比 Adam 快，根据经验，一般 30 个 epoch 即可收敛。因此，表 3 和表 4 在各个模型上，采用 LKF 优化器，本文测试的 4 个数据集均训练 30 个 epoch。

表 3 能量(E/atom，eV)在多个力场模型下的均方根误差

Table 3 Root mean square error of energy(E/atom, eV) under multiple force field models

体系	Adam			LKF		
	GuaNN↓	DP↓	NEP↓	GuaNN↓	DP↓	NEP↓
Ag	0.000 22	0.000 44	**0.000 21**	0.000 31	0.000 18	**0.000 13**
Cu	**0.001 36**	0.003 47	0.005 64	0.020 17	0.000 31	**0.000 27**
C	**0.000 68**	0.004 75	0.001 80	0.000 63	0.000 16	**0.000 11**
H_2O	**0.000 77**	0.002 64	0.004 81	0.004 51	0.000 25	**0.000 11**

表 4 原子受力(F，eV/Angstrom)在多个力场模型下的均方根误差

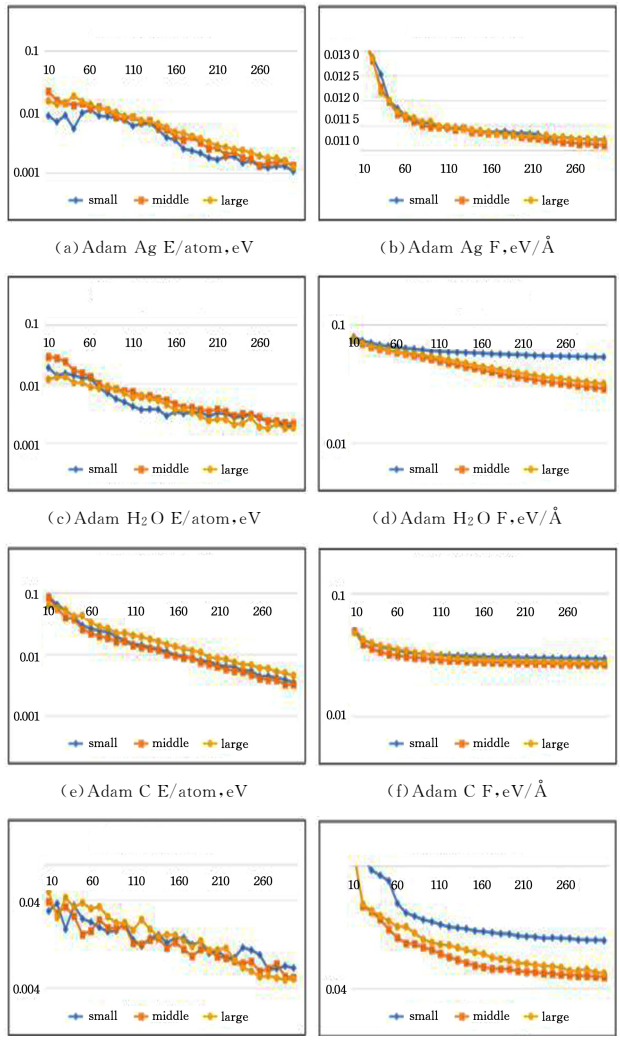
Table 4 Root mean square error of the atomic force(F, eV/Angstrom) under multiple force field models

体系	Adam			LKF		
	GuaNN↓	DP↓	NEP↓	GuaNN↓	DP↓	NEP↓
Ag	0.015 58	0.011 24	**0.011 00**	0.022 34	0.011 40	**0.00953**
Cu	0.082 38	0.046 73	**0.048 42**	0.879 34	0.047 78	**0.045 56**
C	0.075 78	0.029 39	**0.024 78**	0.112 44	0.027 89	**0.023 27**
H_2O	0.079 34	0.054 77	**0.037 81**	0.389 98	0.054 73	**0.028 13**

从测试集的收敛结果来看，基于高斯基组的神经网络力场模型采用 Adam 优化器时，能量的收敛比采用 LKF 优化器更低，使用 Adam 模型比使用 LKF 优化器精度高数倍到数十倍。对于 DP 和 NEP 模型，使用 LKF 优化器比 Adam 优化器在能量的收敛上更好，在力的预测上，两者相当。对于多个模型的比较，NEP 在这 4 个模型上预测结果与第一性原理的 DFT 结果更接近。根据分析，Gua-NN 和 NEP 在模型上均属于自定义描述符的网络，通过指定的高斯基组、径向基组和角度基组得到神经网络力场模型的输入。关于模型层数及每一层的神经元上的设置，NEP 单层宽度大，默认值为 100，Gua-NN 采用三层神经网络，各层的神经元的个数为[15，15，1]。实验结果符合“通用逼近定理”的观点[30](即单层足够宽的神经网络可以近似任何连续函数)，意味着 NEP 采用更宽的网络能带来更强的表达能力。其次，与第一性原理的 DFT 最为接近的是 DP 的预测结果。NEP 和 Gua-NN 在训练时需要用户自行指定基组的数量，基组数及其他超参数是影响 NEP 和 Gua-NN 训练的因素之一。而 DP 无需用户指定物理参数，在训练时，是一种纯数据驱动的神经网络力场模型。DP 相比 Gua-NN 和 NEP 更为稳定，模型的预测结果接近 DFT 结果，与 NEP 的预测结果相当。同时，DP 在 Adam 和 LKF 上的表现也较为稳定，基于 LKF 的 DP 比基于 Adam 的 DP 的收敛精度高 0.002 11～0.011 16。总体看来，DP 是一种较为稳健的模型。

3.2 网络参数量对优化器的影响

DP 包含嵌入网络和拟合网络，其嵌入网络和拟合网络的各层神经元数量可由用户指定。我们规定了 3 组不同的参数规模，分别为 small size DP，middle size DP，large size DP，其嵌入网络和拟合网络的配置如表 2 所列。为了查看参数规模对优化器的影响，挑选 Ag 和 H_2O 数据集，查看不同参数规模的深度势能模型。采用 Adam 优化器和 LKF 优化器时的学习曲线如图 7 和图 8 所示。



(a) Adam Ag E/atom, eV (b) Adam Ag F, eV/Å

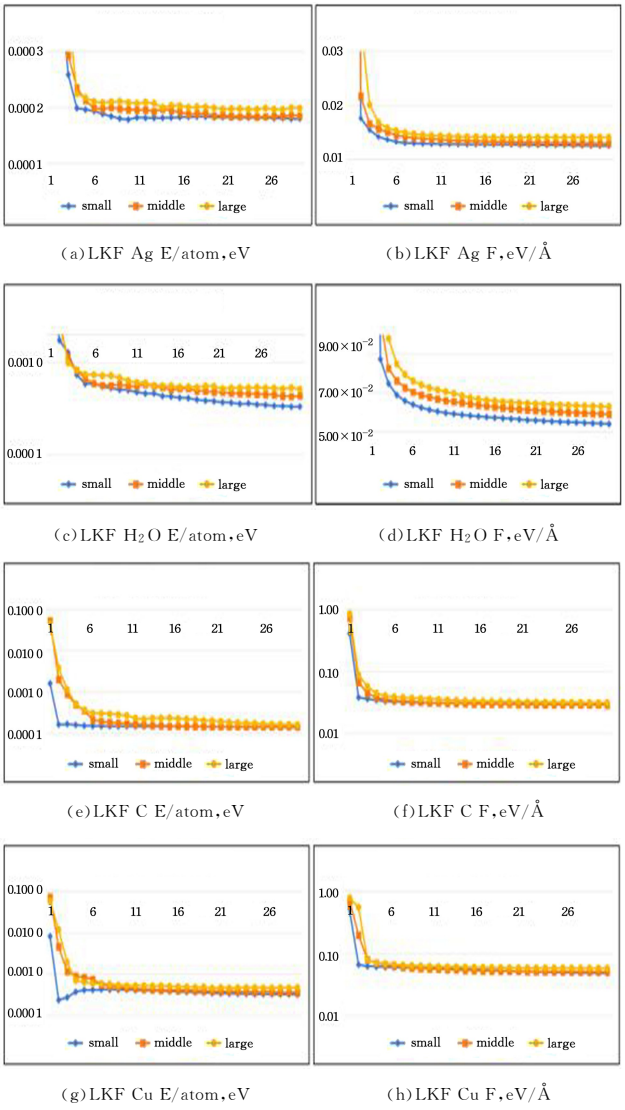
(c) Adam H_2O E/atom, eV (d) Adam H_2O F, eV/Å

(e) Adam C E/atom, eV (f) Adam C F, eV/Å

(g) Adam Cu E/atom, eV (h) Adam Cu F, eV/Å

图 7 不同参数规模下 Adam 优化器的学习曲线

Fig. 7 Learning curve of the Adam optimizer at different parameter scales



(a)LKF Ag E/atom，eV (b)LKF Ag F，eV/Å

(c)LKF H_2O E/atom，eV (d)LKF H_2O F，eV/Å

(e)LKF C E/atom，eV (f)LKF C F，eV/Å

(g)LKF Cu E/atom，eV (h)LKF Cu F，eV/Å

图 8 不同参数规模下 LKF 优化器的学习曲线

Fig. 8 Learning curve of the LKF optimizer at different parameter scales

训练样本数量和网络参数量有着复杂的关联，在训练神经网络力场模型时，需要根据给定的问题规模给出合理的网络参数设置。不难发现，在采用 Adam 优化器对深度势能模型进行训练时，small size 在 H_2O 和 Cu 体系关于原子受力的预测比 middle size 和 large size DP 差。这是由于参数量小的网络通常模型复杂度较低，表现在实际训练任务中，即拟合能力不够。在 4 个训练数据集上，middle size 和 large size DP 具有极为接近的收敛效果。

采用 LKF 进行训练时，通过 4 个体系在训练数据集上的收敛曲线可以发现，采用 small size DP 时，RMSE 的下降比 middle size 和 large size DP 快。这是由于 LKF 是一种近似算法，对 KF 中的 P 矩阵按照聚合毗连的小的权重，切分参数量大的权重矩阵，进行解耦合。当给定分块阈值时，small size，middle size，large size DP 关于 P 矩阵的切分方式如图 9 所示。随着参数量变大，需要维护的权重与权重的误差协方差矩阵的相关性会更多，而采用固定阈值进行切分时，网络参数大时遗失掉的相关性建模会增多，对应图 9 中非对角线块的空白部位。当参数量较大，在给定分块阈值的切分下，出现大量的不被更新的相关性时，LKF 在该参数设置下的收敛性就难以保证。因此，在综合考虑计算资源和误差协方差矩阵相关性的情况下，无需将阈值设得太大。

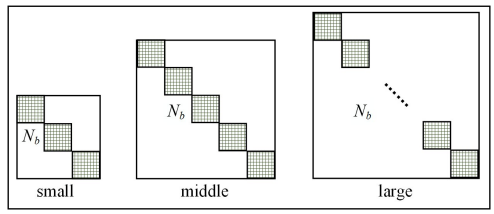


图 9 误差协方差矩阵示意图

Fig. 9 Schematic diagram of the error covariance matrix

同时我们发现，采用 Adam 进行深度势能模型的训练时，单样本的网络前向计算和梯度更新过程对于计算资源的消耗无法占满 A100 加速卡。以 Ag 体系为例，采用 Adam 优化器时，每个 epoch 的训练时间均在 13 s 左右。而在采用 LKF 进行训练时，前面已分析参数过大带来的收敛性问题。从计算效率的角度看，适用 LKF 优化器也不应设置过大的参数量。这是由于参数量过大时，计算和更新的对角协方差子矩阵的数目大幅增加，而代码的实现采用串行的方式执行。因此，参数量过大时，会给神经网络的训练带来更大的存储占用和计算开销。在 Ag 体系上，随着网络参数量的增加，单个 epoch 的训练时间也会增加，如表 5 所列。

表 5 不同规模 DP 模型单个 epoch 的训练时间

Table 5 Training time of a single epoch for a DP model of different scales

优化器	Ag(second/epoch)		
	small	middle	large
Adam	13	13	13
LKF	26	90	238

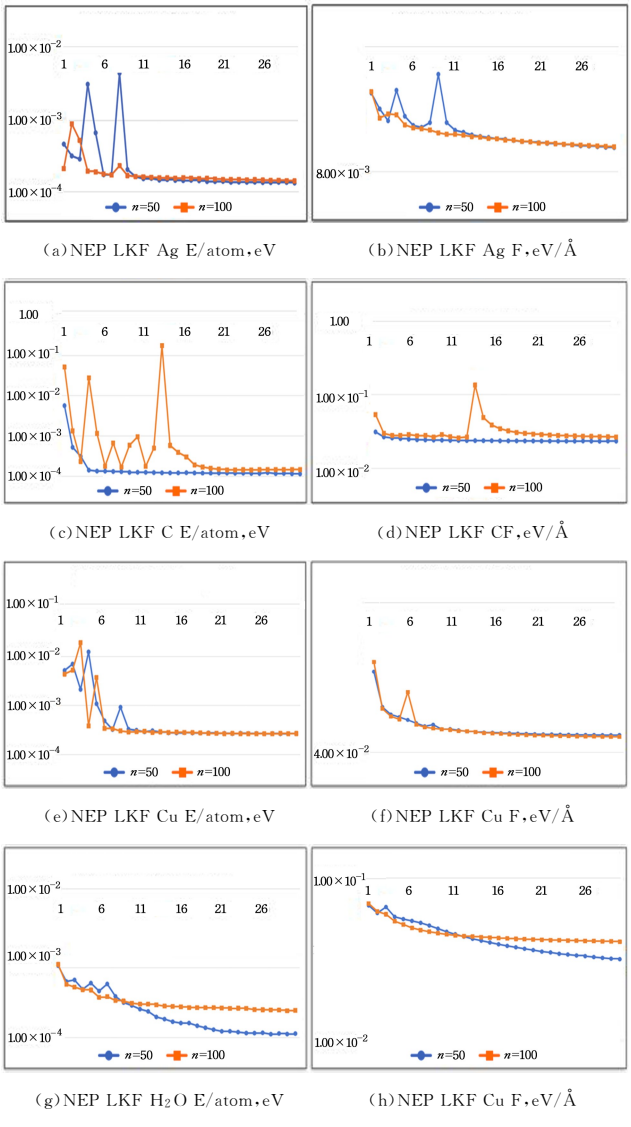
根据上文对 small size，middle size 和 large size 的深度势能模型在训练集上得到的结论，我们检验采用 Adam 和 LKF 优化器在 3 种不同规模的深度势能模型下在测试集上的收敛结果，如表 6 所列。可以得到同样的结论：Adam based DP 采用 middle size 和 large size 比采用 small size 的模型的均方根误差更小，LKF based DP 采用 small size 比采用 middle size 和 large size 的 DP 的均方根误差更小，其中更小的均方根误差用粗体表示。

表 6 不同规模的 DP 模型在测试集上的收敛结果

Table 6 Convergence results of DP models at different scales on the test set

		Adam		LKF	
		E/atom，eV	F，eV/ Å	E/atom，eV	F，eV/ Å
Ag	small	0.004 280	0.153 745	**0.000 190**	**0.011 304**
	middle	**0.001 677**	0.121 359	0.000 195	0.011 469
	large	0.004 161	**0.011 292**	0.000 209	0.011 986
Cu	small	0.002 644	0.054 772	**0.000 379**	**0.055 147**
	middle	**0.001 506**	**0.030 805**	0.000 574	0.059 497
	large	0.002 244	0.034 132	0.000 633	0.063 238
C	small	0.004 753	0.029 391	**0.000 154**	0.027 890
	middle	**0.000 979**	**0.026 703**	0.000 156	**0.027 338**
	large	0.002 066	0.027 695	0.000 184	0.029 955
H_2O	small	0.003 469	0.046 729	**0.000 310**	**0.047 784**
	middle	0.002 544	**0.041 263**	0.000 400	0.050 243
	large	0.001 609	0.041 750	0.000 433	0.057 461

另外，在 NEP 模型上对神经网络的宽度进行测试。NEP 默认的隐藏层的神经元个数为 100，另设置一组隐藏层神经元个数为 50 的对比实验。由于 NEP 为单层全连接网络，隐藏层为 50 或 100，在采用 LKF 优化器，分块阈值为 10240 进行训练时，均没有被切分开，即参数总量小于分块阈值，没有进行 P 矩阵的解耦合操作。NEP 在 LKF 上的表现只受到参数总量本身的影响，隐藏层为 100 个神经元的表达能力比为 50 个的强，收敛性更稳定，如图 10 所示。



(a) NEP LKF Ag E/atom, eV (b) NEP LKF Ag F, eV/Å

(c) NEP LKF C E/atom, eV (d) NEP LKF CF, eV/Å

(e) NEP LKF Cu E/atom, eV (f) NEP LKF Cu F, eV/Å

(g) NEP LKF H_2O E/atom, eV (h) NEP LKF Cu F, eV/Å

图 10 数据在 NEP 上的收敛曲线

Fig. 10 Convergence curve of the data on the NEP

4 总结与展望

优化器算法在神经网络的训练中起着关键作用。本文对已经用于神经网络力场模型中的 3 种不同的优化器（基于梯度下降的一阶优化器、基于拟牛顿法的卡尔曼滤波算法、遗传算法）进行测试和比较，将这 3 种优化器分别用于 3 种有代表性的力场模型上（基于高斯基组的力场模型、深度势能模型、NEP 模型），并在 4 个真实数据集上对模型进行测试，采用均方根误差来评估模型的精度，用模型的训练时间作为模型效率的评价指标。综合来看，Adam 优化器在收敛的过程中鲁棒性更高，目前是神经网络力场模型中使用范围最广泛的优化器。LKF 在一些参数量较小的模型，如 NEP（参数量为 2200）和深度势能模型（small size 参数量约为 27000）上，表现得比 Adam 的收敛效果好。遗传算法目前用于 NEP 模型上。我们发现 NEP 模型采用 LKF 进行权重参数的更新能达到与遗传算法更新 NEP 参数接近的收敛精度，而遗传算法在迁移到其他模型如深度势能模型和基于高斯基组的力场模型上时，无法较好地收敛（模型预测结果的均方根误差在 10^2 的量级）。同时，根据观察到的现象，推测了使用这些优化器进行训练时导致该现象的原因。

结束语 总体来看，优化器算法的性能和效率受到多种因素的影响：1）优化器在训练时本身的参数设置；2）模型的复杂度；3）数据的特性，包括数据涵盖的化学空间，例如是否包含多个相位的数据等；4）数据集的规模。优化器算法在选择和设计时，需要综合考虑以上各种因素，才能在实际的训练过程中表现出良好的收敛性。后续将继续进行更深入的研究：1）在模型变得更复杂，数据的自由度更大的情况下，优化器的选取情况；2）在推广 LKF 和遗传算法时，需要注意参数设置情况；3）尝试给出理论的证明，当给定模型和数据集时，指导优化器的选取。

参 考 文 献

[1] ENGEL E. Density functional theory[M]. Springer, 2011.

[2] ZEPEDA-NÚÑEZ L, CHEN Y, ZHANG J, et al. Deep density: circumventing thekohn-sham equations via symmetry preserving neural networks[J]. arXiv: 1912.00775, 2019.

[3] HAFNER J. Ab-initio simulations of materials usingvasp: Densityfunctional theory and beyond[J]. Journal of Computational Chemistry, 2008, 29(13): 2044-2078.

[4] GIANNOZZI P, ANDREUSSI O, BRUMME T, et al. Advanced capabilities for materials modelling with quantum espresso[J]. Journal of Physics: Condensed Matter, 2017, 29(46): 465901.

[5] JIA W, FU J, CAO Z, et al. Fast plane wave density functional theory molecular dynamics calculations on multi-gpu machines [J]. Journal of Computational Physics, 2013, 251: 102-115.

[6] KOURA K, MATSUMOTO H. Variable soft sphere molecular model for inverse-power-law orlennard-jones potential[J]. Physics of Fluids A: Fluid Dynamics, 1991, 3(10): 2459-2465.

[7] FOILES S, BASKES M, DAW M S. Embedded-atom-method functions for thefcc metals cu, ag, au, ni, pd, pt, and their alloys [J]. Physical Review B, 1986, 33(12): 7983.

[8] SENFTLE T P, HONG S, ISLAM M M, et al. The reaxff reactive forcefield: development, applications and future directions [J]. Computational Materials, 2016, 2(1): 1-14.

[9] NI B, LEE K H, SINNOTT S B. A reactive empirical bond order (REBO) potential for hydrocarbon-oxygen interactions [J]. Journal of Physics: Condensed Matter, 2004, 16(41): 7261.

[10] NGUYEN T D. Gpu-accelerated tersoff potentials for massively parallel molecular dynamics simulations[J]. Computer Physics Communications, 2017, 212: 113-122.

[11] HUANG Y, KANG J, GODDARD III W A, et al. Density functional theory based neural network force fields from energy de-

compositions[J]. Physical Review B，2019，99(6)：064103.
[12] THOMPSON A，SWILER L，TROTT C，et al. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials[J]. Journal of Computational Physics，2015，285：316-330.
[13] LEE K，YOO D，JEONG W，et al. Simple-nn：An efficient package for training and executing neural-network interatomic potentials[J]. Computer Physics Communications，2019，242：95-103.
[14] BEHLER J. Representing potential energy surfaces by high-dimensional neural network potentials[J]. Journal of Physics：Condensed Matter，2014，26(18)：183001.
[15] WANG H，ZHANG L，HAN J，et al. Deepmd-kit：A deep learning package for many-body potential energy representation and molecular dynamics[J]. Computer Physics Communications，2018，228：178-184.
[16] FAN Z，WANG Y，YING P，et al. GPUMD：A package for constructing accurate machine-learned potentials and performing highly efficient atomistic simulations[J]. The Journal of Chemical Physics，2022，157(11)：114801.
[17] DAI H，MACBETH C. Effects of learning parameters on learning procedure and performance of abpnn[J]. Neural networks，1997，10(8)：1505-1521.
[18] SMITH J S，ISAYEV O，ROITBERG A E. Ani-1：an extensible neural network potential withdft accuracy at force field computational cost[J]. Chemical science，2017，8(4)：3192-3203.
[19] SCHÜTT K，UNKE O，GASTEGGER M. Equivariant message passing for the prediction of tensorial properties and molecularspectra[C]// International Conference on Machine Learning. PMLR，2021：9377-9388.
[20] BATZNER S，MUSAELIAN A，SUN L，et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials[J]. Nature Communications，2022，13(1)：2453.
[21] HAGHIGHATLARI M，LI J，GUAN X，et al. Newtonnet：a newtonian message passing network for deep learning of interatomic potentials and forces[J]. Digital Discovery，2022，1(3)：333-343.
[22] JIA W，WANG H，CHEN M，et al. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning[C]// SC20：International Conference for High Performance Computing，Networking，Storage and Analysis. IEEE，2020：1-14.
[23] KINGMA D，BA J. Adam：A method for stochastic optimization[J]. arXiv. 1412. 6980，2014.
[24] HU S，ZHANG W，SHA Q，et al. Rlekf：An optimizer for deep potential with ab initio accuracy[C]// Proceedings of the AAAI Conference on Artificial Intelligence：volume 37. 2023：7910-7918.
[25] HU S，ZHAO T，SHA Q，et al. Training onedeepmd model in minutes：a step towards online learning[C]// PPoPP '24：Proceedings of the 29th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming. New York，NY，USA：Association for Computing Machinery，2024：257-269.
[26] SCHAUL T，GLASMACHERS T，SCHMIDHUBER J. High dimensions and heavy tails for natural evolution strategies[C]// Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation(GECCO'11). New York，NY，USA：Association for Computing Machinery，2011：845-852.
[27] LAMBORA A，GUPTA K，CHOPRA K. Genetic algorithm- a literature review[C]// 2019 International Conference on Machine Learning，Big Data，Cloud and Parallel Computing(COMITCon). 2019：380-384.
[28] PIEPER A，KREUTZER M，ALVERMANN A，et al. High-performance implementation ofchebyshev filter diagonalization for interior eigenvalue computations[J]. Journal of Computational Physics，2016，325：226-243.
[29] ZAVERKIN V，HOLZMÜLLER D，STEINWART I，et al. Fast and sample-efficient interatomic neural network potentials for molecules and materials based on gaussian moments[J]. Journal of Chemical Theory and Computation，2021，17(10)：6658-6670.
[30] NISHIJIMA T. Universa l approximation theorem for neural networks[J]. arXiv：2102. 10993，2021.

LI Enji，born in 1994，master. His main research interests include machine learning and molecular dynamics simulations.

JIA Weile，born in 1985. Ph. D，researcher. His main research interests include AI4Science，HPC and AI.

（责任编辑：李亚辉）