arXiv:2304.03694v1 [physics.chem-ph] 5 Apr 2023

基于等变贝叶斯神经网络的高精度不确定性感知原子间力建模

Tim Rensmeyer, Benjamin Craig, Denis Kramer, Oliver Niggemann

摘要

尽管贝叶斯神经网络为不确定性建模、主动学习和融入先验物理知识提供了一个颇具前景的框架，但在原子间力建模领域，针对它们的应用仍然寥寥无几。将其应用于学习原子间作用力所面临的主要挑战之一，是缺乏适用于后验密度的蒙特卡罗马尔可夫链采样算法，因为对于许多最先进的架构，通常使用的算法无法在实际可接受的时间内收敛。为应对这一挑战，本文提出了一种新的蒙特卡罗马尔可夫链采样算法，可以规避现有采样方法所面临的问题。此外，我们基于 NequIP 架构提出了一种新的随机神经网络模型，并证明当其与我们新颖的采样算法相结合时，能够获得具有最先进精度的预测结果，同时还能很好地量化不确定性。

1 引言

尽管作为化学基础的量子力学定律早在近一个世纪前就已被发现，但将其应用于许多化学现象的从头算预测，至今仍是一项艰巨的任务[1, 2]。这在很大程度上是因为数值求解这些方程所涉及的计算复杂度极高，而且需要考量的相空间极其庞大[1, 2]。

分子动力学（MD）研究分子系统的时间演化，尤其受到这一挑战的影响。尽管已经发展出密度泛函理论（密度泛函理论）等能够以极高精度计算原子间作用力的方法，但这些方法通常计算成本高昂，并且随着系统规模的增大而表现出较差的扩展性。因此，要以高精度对更大的系统进行建模，并覆盖足够长的时间范围，仍然非常困难。

为克服这些困难，利用神经网络等机器学习模型学习原子间力已成为一个备受关注的研究领域。[3]然而，从从头算 模拟生成训练数据的成本，阻碍了通常构建具有足够预测能力的神经网络所需的大规模训练集。神经网络设计领域近期的创新，已经通过将对称性和能量守恒以硬约束的形式纳入模型架构 [4, 5, 6, 7, 8, 9]，使得仅凭有限数据为给定化合物学习高度精确的原子间力场成为可能。然而，要使模型对整个化合物类别达到化学精度，仍然需要大量数据 [4, 5, 8, 9], ，这在很大程度上是由于可能的原子构型空间极其庞大。因此，创建具有预测能力且可迁移的模型仍然是一项具有挑战性的任务。

此外，现有的最先进模型通常缺乏对其预测结果的不确定性度量[3, 10], ，因此无法判断哪些预测是可靠的。这限制了主动学习策略的应用：利用不确定性，可以更高效地从庞大的构型空间中选择有用的训练数据 [11, 12, 13, 14, 15] ，只标注模型预测其误差可能较大的异常构型。

此外，对于已部署的模型，检测到的异常值可以在运行过程中通过 DFT 重新计算，以确保基于神经网络的 MD 轨迹的准确性；也可以用于事后舍弃可能存在物理不准确性的轨迹。所有这些因素都阻碍了模型向更通用的方向扩展，而这类模型是预测和优化材料复杂（动态）性质所必需的；在这些应用中，需要筛选大量多样的原子构型，同时确保整体准确性高且具有良好的不确定性度量。后者可用于决定哪些预测样本应通过从头算 计算或实验进行进一步研究，哪些样本可能是假阳性。不确定性度量还有助于确保对稀有事件进行高精度建模；这类事件往往主导更长时间尺度的动力学，却在 MD 轨迹中严重缺乏代表性 [14]。与接近基态的构型相比，接近过渡态的构型被采样到的概率呈指数级降低。

构型。但这些状态通常具有特殊的电子构型和成键特性，而这些并不出现在低能构型中。因此，不确定性度量可以实时识别需要额外学习的构型，从而将主动学习策略偏向于相关的“新化学”，例如过渡态。

此外，许多现有的最先进模型目前都通过一种颇为复杂的方式在较小的数据集上进行优化，以提升性能；在这一过程中，通过较大的学习率和较小的批量大小将人为噪声引入优化过程 [4, 5, 6, 7, 9]。尽管这种方法在经验上表现良好，但对于此类优化的有效性尚无公认的理论基础，因此，除了反复试验之外，也没有指导如何选择适当的批量大小和学习率。针对这些挑战，我们在此探究将贝叶斯神经网络（BNN）与最先进的神经网络相结合、用于建模原子间力的可行性，因为贝叶斯神经网络具有稳健的不确定性度量，并且能够通过蒙特卡罗马尔可夫链（MCMC）方法，以一种内在且有原则的方式向优化过程中添加噪声。此外，它们还为纳入并基于数据更新先验信念提供了良好的框架，例如近似的物理引导的原子间力场；未来可以利用这些力场进一步提升数据效率和泛化能力。

然而，我们发现，目前现成的马尔可夫链蒙特卡罗方法不适用于处理现代模型中不同参数组所呈现的巨大梯度尺度差异。这导致参数空间的遍历速度慢得不切实际，以至于无法在合理的时间内实现收敛，这也解释了它们目前为何鲜少得到应用。因此，针对基于神经网络的原子间力模型，需要一种新的马尔可夫链蒙特卡罗算法，以实现高质量的贝叶斯不确定性量化。

本文的贡献如下：

•我们基于 NequIP 架构开发了一种新的随机神经网络模型， [7] 用于具备不确定性感知能力的原子间力建模，并实现了最先进的精度。

• 我们提出了一种新颖的贝叶斯马尔可夫链蒙特卡罗算法，无需依赖经过精细调优的批量大小和学习率，即可在实际可行的时间内生成精度高且不确定性估计质量高的模型。

•我们表明，当前基于较高学习率和较小批量大小的优化方法，在一定程度上已经模拟了贝叶斯马尔可夫链蒙特卡罗方法。

• 我们讨论了所提出的神经网络模型和采样算法的不足之处，以及可能的进一步改进和扩展。

本文常用的符号汇总于表 1 中。

表 1：常用符号说明

Notation	Meaning
Bold case symbol	Vector-valued quantities
$N(\boldsymbol{\mu}, \Sigma)$	Normal distribution
$\boldsymbol{\mathcal{N}}(\boldsymbol{\mu}, \Sigma)$	Random variable ~ $N(\boldsymbol{\mu}, \Sigma)$
E	Potential energy
D	The dataset
I	The identity matrix
$\boldsymbol{\theta}$	Neural network parameters
$\boldsymbol{x}$	Independent variable
$\boldsymbol{y}$	Dependent variable
$\boldsymbol{y}\|\boldsymbol{x}$	$\boldsymbol{y}$ conditioned on $\boldsymbol{x}$
$\boldsymbol{r}$	Atomic coordinates
z	Nuclear charges
$\boldsymbol{ab}$	Element-wise product of $\boldsymbol{a}$ and $\boldsymbol{b}$

1.1 相关工作

迄今为止，贝叶斯神经网络在这一领域的应用案例寥寥无几，但文献中仍有一些应用实例[15, 16, 17, 18]。然而，在 [16, 17] 中，研究人员使用了蒙特卡罗丢弃法进行贝叶斯推断。在某些情况下，这种方法可以被解释为贝叶斯变分推断 [19], ，但即使对于基于变分推断的方法，它对真实后验分布的近似也较差 [19] ，其不确定性量化效果通常远逊于马尔可夫链蒙特卡罗方法 [20]。尽管 [15], 采用了准确得多的贝叶斯算法，但其应用范围非常有限，仅涉及双原子系统。此外，上述所有案例都局限于非常小的全连接神经网络，而我们采用的是最先进的基于图神经网络的架构。我们在文献中只找到一项尝试，旨在将基于图神经网络的架构与通过模拟随机过程实现的采样方案相结合。这项工作由 Thaler 等人近期发表的一篇预印本完成 [18]。然而，他们在尝试对贝叶斯后验分布进行采样时未能取得有竞争力的结果，不得不转而对人为锐化的分布进行采样。即便如此，他们仍发现，要可靠地量化不确定性，必须从多个不同的随机过程中生成样本；而我们只需从单条马尔可夫链中生成样本。

2 基础模型

2.1 形式化设定

基于神经网络的原子间力模型旨在将原子构型 $\boldsymbol{x}$ ={($\boldsymbol{r}_1$, z_1).....，($\boldsymbol{r}_n$, z_n)}映射为作用于各个原子核的力 $\{\boldsymbol{F}_1, ..., \boldsymbol{F}_n\}$。其中， $\boldsymbol{r}_i$ 表示原子 i 的原子核位置，z_i 表示其核电荷。

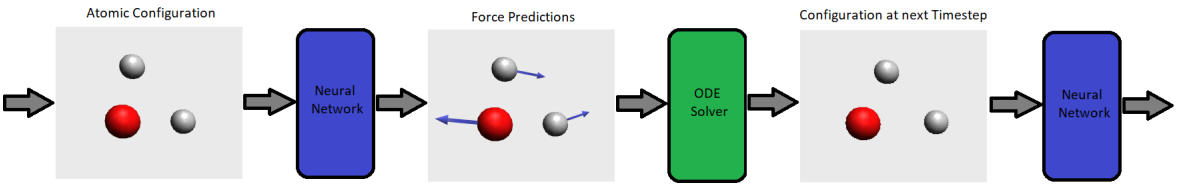


图 1：分子动力学工作流程示意图，其中使用神经网络预测力。

随后，通过常微分方程（ODE）求解器，根据牛顿运动方程利用这些力来模拟原子核的运动（图 1）。

这类数据的一个重要特点是其无结构性，也就是说，原子的枚举顺序是任意的。因此，合适的神经网络模型应对数据重新排序保持等变性。

另一个重要约束是，这些力具有保守性，因此可以由单个势能面梯度的负值导出

$$\boldsymbol{F}_i = -\nabla_{\boldsymbol{r}_i} E((\boldsymbol{r}_1, z_1), ..., (\boldsymbol{r}_n, z_n)).$$

势能面本身在原子坐标的任何保距变换下保持不变。最初，人们通过仅将原子间距离 r_{ij} 和核电荷作为输入，将这一性质纳入神经网络模型；这些输入本身具有上述不变性。尽管原则上可以从原子间距离集合中提取方向信息，但在实践中，显式纳入方向信息可以显著提高数据效率和准确性 [10]，并已成为许多最先进神经网络模型的关键特征 [4, 5, 6, 8, 7, 9]。

2.2 NequIP模型

目前用于原子间力建模的最强大神经网络模型之一是 NequIP 模型 [7](图 2)。该模型考虑了所有先前的因素，由模型主体和投影层组成。模型主体将原子构型 $\{(\boldsymbol{r}_1, z_1) ...,(\boldsymbol{r}_n, z_n)\}$ 映射为一组在保持距离的变换下不变的高维潜在特征向量 $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_n\}$。随后，投影层将 $\{(\boldsymbol{v}_1, z_1) ..., (\boldsymbol{v}_n, z_n)\}$ 映射为（虚拟）原子能量 $\{E_1, ..., E_n\}$。

势能 E 随后被计算为原子能量之和。最后，作用于原子核上的力被计算为势能相对于核坐标的负梯度。最后一步可以利用现代深度学习库的自动微分功能直接实现。

3 贝叶斯模型

3.1 贝叶斯神经网络

经典神经网络优化的目标是找到一组单一的参数，即权重和偏置，使其能够很好地对底层数据分布进行建模，而训练数据正是从该分布中采样得到的。相比之下，在贝叶斯方法中，神经网络的参数以概率方式建模。为使符号说明保持简洁，我们使用 $\boldsymbol{\theta}$ 来表示一个包含完整神经网络参数集的向量。对于BNN，假设我们对什么样的参数集构成良好参数集已有一些先验知识，并将其表示为先验密度 $p(\boldsymbol{\theta})$的形式。随后，通过训练数据$D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1) ...,(\boldsymbol{x}_l, \boldsymbol{y}_l)\}$ 并调用贝叶斯规则来计算后验密度，从而进一步细化该先验密度：

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)} = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

现在可以通过以下方式对新的数据点（$\boldsymbol{x}, \boldsymbol{y}$）进行预测：

$$p(\boldsymbol{y}|\boldsymbol{x}, D) = \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\theta$$

$$= \mathbb{E}_{p(\boldsymbol{\theta}|D)} \left[p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\right].$$

对于BNN 而言，该积分几乎总是无法进行解析求解。然而，通过从 $p(\boldsymbol{\theta}|D)$ 中生成样本 $\boldsymbol{\theta}_1, ...\boldsymbol{\theta}_k$，根据大数定律，可以将其估计为：

$$p(\boldsymbol{y}|\boldsymbol{x}, D) \approx \frac{1}{k} \sum_{i=1}^{k} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_i).$$

3.2 随机模型

为了在原子间作用力建模中实现最先进的精度，并同时获得良好的预测不确定性度量，我们需要一种新的随机神经网络模型，本节将对此进行介绍。为构建数据的随机模型，我们对数据的条件独立性作出如下假设：

$$p(D) = p(\boldsymbol{y}_1, ..., \boldsymbol{y}_l|\boldsymbol{x}_1, ..., \boldsymbol{x}_l)p(\boldsymbol{x}_1, ..., \boldsymbol{x}_l)$$

$$= p(\boldsymbol{x}_1, ..., \boldsymbol{x}_l)\Pi_{i=1}^{l} p(\boldsymbol{y}_i|\boldsymbol{x}_i).$$

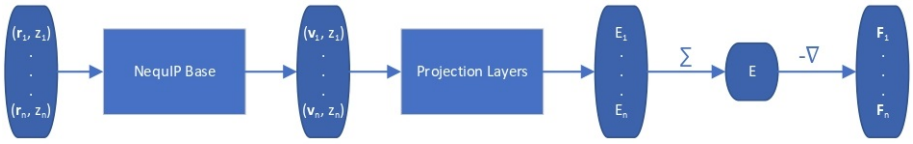


图 2：用于根据原子序数 z_i 和坐标 $\boldsymbol{r}_i$ 预测势能 E 和原子力 $\boldsymbol{F}_i$ 的 NequIP 模型计算图。

$p(\boldsymbol{y}_i|\boldsymbol{x}_i)$ 将由神经网络推断为 $p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta})$，而 $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l)$ 则取决于数据生成过程。

更具体地说，我们对条件概率密度进行建模

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) = p(E, \boldsymbol{F}_1, ..., \boldsymbol{F}_n|(\boldsymbol{r}_1, z_1), ..., (\boldsymbol{r}_n, z_n), \boldsymbol{\theta})$$

as

$$\begin{aligned}\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta} &\sim N(\mu_E(\boldsymbol{\theta}, \boldsymbol{x}), \sigma_E^2(\boldsymbol{\theta}, \boldsymbol{x})) \\ &\times \Pi_{i=1}^n N(\boldsymbol{\mu}_{\boldsymbol{F}i}(\boldsymbol{\theta}, \boldsymbol{x}), \sigma_{\boldsymbol{F}_i}^2(\boldsymbol{\theta}, \boldsymbol{x})I),\end{aligned}$$

其中 $\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}$ 表示在给定 $\boldsymbol{x}$ 和 $\boldsymbol{\theta}$ 的条件下的 $\boldsymbol{y}$，$\boldsymbol{x} = \{(\boldsymbol{r}_1, z_1) ....., (\boldsymbol{r}_n, z_n)\}$；均值 $\mu_E(\boldsymbol{\theta}, \boldsymbol{x})$ 和 $\boldsymbol{\mu}_{\boldsymbol{F}i}(\boldsymbol{\theta}, \boldsymbol{x})$ 是 NequIP 模型的常规点预测；标准差 $\sigma_E(\boldsymbol{\theta}, \boldsymbol{x})$ 和 $\sigma_{\boldsymbol{F}_i}(\boldsymbol{\theta}, \boldsymbol{x})$ 则由两个独立的多层感知器（MLP）根据 NequIP 基础模型的输出 $\boldsymbol{v}_1, ... \boldsymbol{v}_n$ 预测得到；I 表示单位矩阵（图 3）。请注意，我们为所有三个力分量计算一个不变的标准差，以确保预测密度在原子坐标的任意保距变换下保持等变。

我们在此将势能 E 纳入模型，作为一个额外的因变量，因为它在确定基态构型等应用中可能有用，而且无论如何，在密度泛函理论（DFT）中，势能都是力计算的副产品。最后，我们采用简单的高斯均值场先验 $\boldsymbol{\theta} \sim N(\boldsymbol{0}, I)$。

3.3 从后验分布中采样

通常通过模拟一个在分布上收敛至后验分布的马尔可夫链，来生成高质量的样本权重。尽管已有多种针对神经网络应用的马尔可夫链方法，但我们发现它们并不适用于这一场景。我们遇到的主要困难是，神经网络不同层对应的不同参数集，其梯度尺度相差数个数量级。因此，所有参数共用一个步长是不可能的，因为这会导致梯度较小的参数集在优化过程中被冻结。现代神经网络优化器通过为每个参数设置自适应步长来规避这一问题[21]，但对于马尔可夫链而言，这种做法只有在非常有限的程度上才可行

[22] 而不改变其收敛到的平稳分布。此外，这通常需要计算高阶导数 [22] ，计算成本很高。为应对这些挑战，我们在此提出一种基于 Chen 等人提出的随机梯度哈密顿蒙特卡洛（SGHMC）算法 [23]、用于对后验分布进行采样的新方法。该算法的基本形式由以下马尔可夫链给出

$$\begin{aligned}\Delta\boldsymbol{w}_t &= -\left(\nabla_{\boldsymbol{\theta}_t} u(\boldsymbol{\theta}_t) + \boldsymbol{M}^{-1}\boldsymbol{C}\boldsymbol{w}_{t-1}\right)\Delta t + \sqrt{2\boldsymbol{C}\Delta t}\boldsymbol{\mathcal{N}}_t(\boldsymbol{0}, I), \\ \Delta\boldsymbol{\theta}_t &= \boldsymbol{M}^{-1}\boldsymbol{w}_t\Delta t.\end{aligned}$$

其中

$$u(\boldsymbol{\theta}) := -\ln p(\boldsymbol{\theta}) - \sum_{i=1}^{l} \ln p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}),$$

$\boldsymbol{M}^{-1}$ 和 $\boldsymbol{C}$ 是包含严格正值的向量，$\Delta t << 1$ 是步长，$\boldsymbol{\mathcal{N}}_t(\boldsymbol{0}, I)$表示一个具有多元标准正态分布的随机变量。为使符号说明简洁，我们在此约定，向量上的所有运算（乘法、求逆等）均按元素进行。辅助变量 $\boldsymbol{w}_t$ 与 $\boldsymbol{\theta}_t$具有相同的维度。

通过注意到这些方程也描述了在摩擦作用下，坐标为 $\boldsymbol{\theta}$、势能为 $u(\boldsymbol{\theta})$、动量为 $\boldsymbol{w}_t$ 、质量为 $\boldsymbol{M}$ 的粒子的扩散，可以获得对随机梯度哈密顿蒙特卡洛的直观理解。在这个类比中，向量 $\boldsymbol{C}$ 表示摩擦系数。从机器学习的角度来看，$\boldsymbol{w}_t$ 表示一种动量项，类似于许多现代神经网络优化器 [21] 中使用的动量项。事实上，Chen 等人采用了一些替换，从而得到带动量的随机梯度下降的贝叶斯对应形式[23], ；但我们将采用略有不同的替换，以便自然地引入自适应步长：$\alpha = \Delta t\boldsymbol{M}^{-1}\boldsymbol{C}$, $\gamma = (\Delta t)^2|D|\alpha^{-1}$ 和 $\boldsymbol{v}_t = \gamma^{-1}\Delta t\boldsymbol{w}_t$，由此得到：

$$\begin{aligned}\Delta\boldsymbol{v}_t &= -\alpha\frac{1}{|D|}\nabla_{\boldsymbol{\theta}_t} u(\boldsymbol{\theta}_t) - \alpha\boldsymbol{v}_{t-1} + \alpha\sqrt{\frac{2\boldsymbol{M}}{|D|\gamma}}\boldsymbol{\mathcal{N}}_t(\boldsymbol{0}, I), \\ \Delta\boldsymbol{\theta}_t &= \gamma\boldsymbol{M}^{-1}\boldsymbol{v}_t\end{aligned}$$

其中，$|D|$ 是数据集 D的大小。

这在噪声项 $\boldsymbol{\mathcal{N}}_t(\boldsymbol{0}, I)$ 的范围内，等同于 Adam 优化器的更新 [24]。

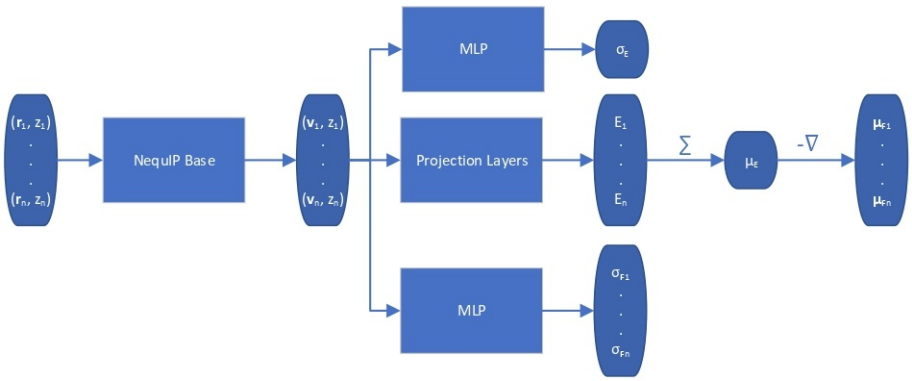


图 3：用于根据原子序数 z_i 和坐标 $\boldsymbol{r}_i$ 预测势能 E 和原子力 $\boldsymbol{F}_i$ 的均值 μ_E，$\boldsymbol{\mu}_{\boldsymbol{F}_i}$ 及标准差 σ_E，$\sigma_{\boldsymbol{F}_i}$ 的随机模型计算图。

遗憾的是，标准 Adam 优化器中使用的质量项 $\boldsymbol{M}$ 即使在训练后期也可能变化很大，而且这种变化取决于前几个时间步的 $\boldsymbol{\theta}_t$ 值。因此，产生的过程 $(\boldsymbol{\theta}_t,\ \boldsymbol{v}_t)$ 甚至不再是马尔可夫链，而贝叶斯后验分布很可能也不再是 $\boldsymbol{\theta}_t$ 收敛到的分布 [22]。对于许多神经网络架构，只需设置 $\boldsymbol{M} = I$ 即可实现及时收敛到后验分布。然而，模型中不同参数组的梯度尺度差异极大，使得这种方法不可行。为此，我们现在为随机梯度哈密顿蒙特卡洛算法引入一种新的自适应步长方法，该方法无需计算高阶导数，仍能收敛到后验分布。为实现这一点，我们在优化的第一阶段将 $\boldsymbol{M}$ 设为 AMSGrad 算法 [25] 的分母：

$$\boldsymbol{a}_t = (1-\beta)\frac{1}{|D|^2}\left(\nabla_{\boldsymbol{\theta}_t} u(\boldsymbol{\theta}_t)\right)\left(\nabla_{\boldsymbol{\theta}_t} u(\boldsymbol{\theta}_t)\right) + \beta \boldsymbol{a}_{t-1},$$

$$\boldsymbol{D}_t = \max(\boldsymbol{D}_{t-1}, \boldsymbol{a}_t),$$

$$\boldsymbol{M}_t = \sqrt{\frac{\boldsymbol{D}_t}{1-\beta^t}} + \epsilon_{stability}.$$

这里，$\max(\cdot,\cdot)$ 是逐元素最大值，$\epsilon_{stability}$ 是稳定性常数，β 是用于计算运行平均值 $\boldsymbol{a}_t$ 的超参数，通常设置在 0.99和 0.9999 之间。

由于这一质量项仍然依赖于时间和路径，$\boldsymbol{\theta}_t$ 通常无法在此阶段精确收敛到后验分布。然而，与大多数自适应步长方法不同，$\boldsymbol{M}_t$ 并非基于梯度平方的滑动平均，而是基于这种滑动平均的最大值，因此在优化后期通常变化很小。因此，在优化的这一阶段，该过程的分布已经相当接近贝叶斯后验分布。此外，由于 $\boldsymbol{M}_t$ 一段时间后已几乎保持不变，我们可以在一定步数后将其完全固定，而不会造成不稳定

，此时该过程 $\boldsymbol{\theta}_t$ 便成为一个常规的随机梯度哈密顿蒙特卡洛过程；已知当步长 [23]足够小时，该过程会收敛到贝叶斯后验分布。

由于深度学习数据集通常非常庞大，精确评估 $\nabla_{\boldsymbol{\theta}} u(\boldsymbol{\theta})$通常非常耗时。实际实现则在一个较小的、随机采样的子集 $D \subset D$上估计$\hat{}\nabla_{\boldsymbol{\theta}} u(\boldsymbol{\theta})$，通过

$$\nabla_{\boldsymbol{\theta}} u(\boldsymbol{\theta}) = \boldsymbol{\epsilon}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \hat{u}(\boldsymbol{\theta})$$

$$:= \boldsymbol{\epsilon}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \left(-\ln p(\boldsymbol{\theta}) - \frac{|D|}{|\hat{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\hat{D}} \ln p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) \right).$$

所得算法总结于算法 1中。这里，$\boldsymbol{\epsilon}(\boldsymbol{\theta})$ 是对 $\nabla_{\boldsymbol{\theta}} u(\boldsymbol{\theta})$ 的估计误差，其中 $\mathbb{E}\left[\boldsymbol{\epsilon}(\boldsymbol{\theta})\right] = \mathbf{0}$。

只要 $\frac{|\boldsymbol{\epsilon}(\boldsymbol{\theta}|}{|D|} << \sqrt{\frac{2\boldsymbol{M}_t}{\gamma|D|}}$)很明显，这一附加-

Algorithm 1 The Stochastic Optimization Algorithm

for $t = 1$ **to** T **do**
  Sample minibatch $\hat{D}$
  Evaluate $\nabla_{\boldsymbol{\theta}_t} \hat{u}(\boldsymbol{\theta}_t)$
  $= \nabla_{\boldsymbol{\theta}_t} \left(-\ln p(\boldsymbol{\theta}_t) - \frac{|D|}{|\hat{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\hat{D}} \ln p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_t) \right)$
  if $t \le t_{max}$ **then**
    $\boldsymbol{a}_t = (1-\beta)\frac{1}{|D|^2}\left(\nabla_{\boldsymbol{\theta}_t} \hat{u}(\boldsymbol{\theta}_t)\right)\left(\nabla_{\boldsymbol{\theta}_t} \hat{u}(\boldsymbol{\theta}_t)\right) + \beta \boldsymbol{a}_{t-1}$
    $\boldsymbol{D}_t = \max(\boldsymbol{D}_{t-1}, \boldsymbol{a}_t)$
    $\boldsymbol{M}_t = \sqrt{\frac{\boldsymbol{D}_t}{1-\beta^t}} + \epsilon_{stability}$
  else
    $\boldsymbol{M}_t = \boldsymbol{M}_{t_{max}}$
  end if
  $\Delta\boldsymbol{v}_t = -\frac{\alpha}{|D|}\nabla_{\boldsymbol{\theta}_t} \hat{u}(\boldsymbol{\theta}_t) - \alpha\boldsymbol{v}_{t-1} + \alpha\sqrt{\frac{2\boldsymbol{M}_t}{|D|\gamma}}\boldsymbol{\mathcal{N}}_t(0, I)$
  $\Delta\boldsymbol{\theta}_t = \gamma\boldsymbol{M}_t^{-1}\boldsymbol{v}_t$
end for

测量噪声不会对动态产生很大影响。
因为高斯噪声项将占主导地位。
这总可以通过选择合适的

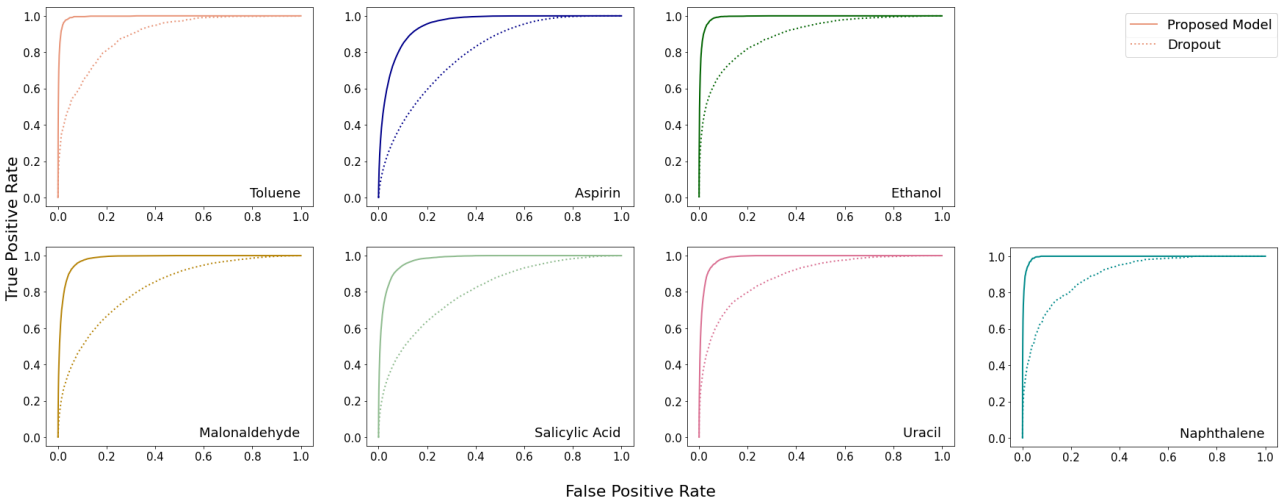


图4：使用k=8蒙特卡罗样本，在RMD17 数据集上对预测误差至少为 1 kcal/(mol· ˚A) 的力分量进行基于不确定性的检测所得的受试者工作特征曲线。

批量大小和步长 γ 来实现。
然而，对于较大的学习率和较小的批量大小，额外的噪声会变得显著。
事实上，即使我们不向这一过程中添加高斯噪声，随后仅使用由此得到的AMS-Grad 优化，常用的大步长和小批量大小所引入的噪声也会使优化过程变成一种随机过程，在一定程度上模仿所提出的算法，只是带有一个噪声项 $|\boldsymbol{w}_{(\boldsymbol{\theta}|}$ $_{|D|}$)，而该噪声项没有任何理论基础。

4 实证评估

4.1 基准测试

为了评估模型的准确性和预测不确定性，我们使用了两个不同的数据集。为了分析模型在不同蒙特卡洛样本数量下以及领域偏移情况下的表现，我们使用了一个由 PEDOT聚合物组成的数据集。PEDOT聚合物是一类导电聚合物，凭借其化学稳定性和可调节的性质 [26] ，已广泛用于传感器 [27], 、超级电容器 [28], 、电池电极 [29], 、生物电子学、太阳能电池、电致变色显示器、电化学晶体管 [30], 和自旋电子学 [31]等领域。该数据集包含长度为 8、12 和 16 的聚合物，其中仅使用较短的链进行训练（详情见附录 B）。训练集总共仅包含 100 个构型，其中长度为 8 的聚合物有 50 个，长度为 12 的聚合物有 50 个。同样，我们构建了一个规模较小、包含 30 个样本的验证集。

使用的第二个数据集是 MD17数据集 [32] ，由若干种小型有机分子的长分子动力学轨迹组成，这些轨迹通过密度泛函理论计算得到。由于该数据集存在一些数值误差，因此推出了一个修订版本RMD17最近以更高精度完成了计算[33] ，我们将利用它来评估模型性能。由于随机失活是用于原子间力建模的最常见贝叶斯方法，我们将在该数据集上将模型性能与所提出随机模型的基于随机失活的版本进行比较（详情请参见附录 A.4）。

对于每种化合物，我们在训练过程中随机抽取了 1000 个构型，其余构型用于测试。在抽取的 1000 个构型中，30 个留作小型验证集，其余构成实际的训练集。我们提出的模型和基于随机失活的模型使用了相同的训练集、验证集和测试集样本。

4.2 评估指标

为衡量预测准确性，我们评估力的平均绝对误差（MAE-F）和能量的平均绝对误差（MAE-E），以及均方根误差（RMSE-F、RMSE-E）。在评估中，我们使用估计后验密度下的期望值作为点预测。即 $\boldsymbol{F}_{pred}(\boldsymbol{x}) = \frac{1}{k}\sum_{i=1}^{k}\boldsymbol{\mu}_{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{\theta}_i)$ 和 $E_{pred}(\boldsymbol{x}) = \frac{1}{k}\sum_{i=1}^{k}\mu_E(\boldsymbol{x},\boldsymbol{\theta}_i)$。用于评估和比较预测不确定性的一个指标是平均对数似然（MLL）。由于在最常见的应用中，不确定性度量的主要任务是异常值检测，我们还基于力分量预测分布的方差，评估检测误差大于 1 kcal/(mol· ˚A) 的力分量的 ROC AUC 分数。由于 PEDOT 数据集的规模相对较小，我们仅在规模大得多的 RMD17 数据集上评估异常值检测。

最后，为了评估所提出的模型是否得到了适当的校准，我们使用期望校准误差（

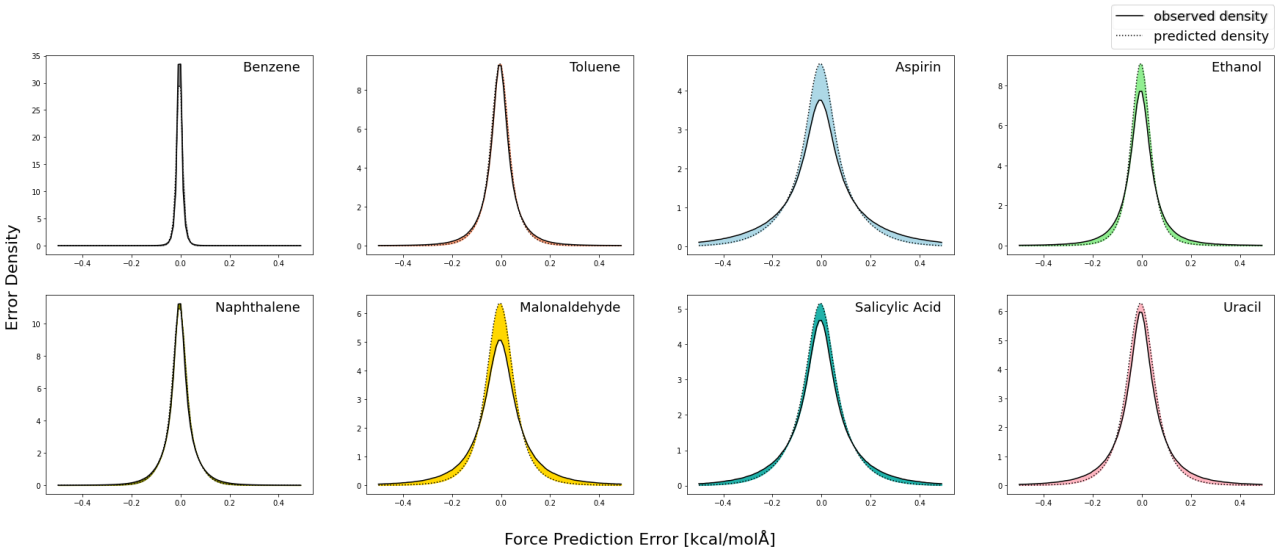


图 5：RMD17 数据集上，使用 k=8 个蒙特卡罗样本时，力分量误差密度的观测值和预测值，单位为 kcal/(mol· ˚A)。

差（ECE） [34] ，并对力分量的预测误差密度与观测误差密度进行目视检查（图 4 和图 5）（详情见附录 A.5）。计算 ECE 时，将预测结果 y_i 划分为m 个等宽区间。随后，ECE 的计算方式如下：

$$ECE = \sum_{i=1}^{m} f_i \left| f_i - e_i \right|$$

其中， f_i 是落入分箱 i 的观测样本比例， e_i 是样本落入分箱 i的预测概率。由于 ECE 高度依赖于分箱宽度，我们将 ECE 除以分箱宽度δ ，得到归一化期望校准误差（NECE）：

$$NECE = \sum_{i=1}^{m} \frac{f_i}{\delta} \left| f_i - e_i \right|$$

这是对.....的估计

$$\mathbb{E}_{\rho(y)} \left[\left| \rho(y) - \rho_{predicted}(y) \right| \right].$$

4.3 结果

PEDOT 数据集上的结果

如表 2 所示，该模型在三个测试集上的力预测均具有较高精度，且精度会随链长增加而降低。从一个蒙特卡罗样本增加到八个蒙特卡罗样本时，精度略有提升。与力预测相比，该模型在能量预测这一辅助任务上的误差要大得多。不过这并不令人意外，因为 PEDOT 分子相当大，能量尺度也相应较大。此外，PEDOT 训练集包含的数量要少得多。

能量标签多于力标签。此外，从表 3 可以看出，单个蒙特卡罗样本无法为力提供良好的不确定性估计，而使用八个蒙特卡罗样本时，这种情况得到了显著改善。对于能量预测的不确定性量化，则未观察到类似趋势。

尽管训练集中不包含长度为 16 的聚合物链，但我们发现，即使在使用 k=8 个蒙特卡罗样本的情况下，模型仍能在力预测中达到较高的准确度，并对不确定性进行良好的量化（表 3）；而使用单个蒙特卡罗样本时，不确定性量化效果则差得多。预测和实际力分量的完整直方图见附录 C。即使在 k=8 的情况下，我们也观察到一定程度的过度自信（图 6）。

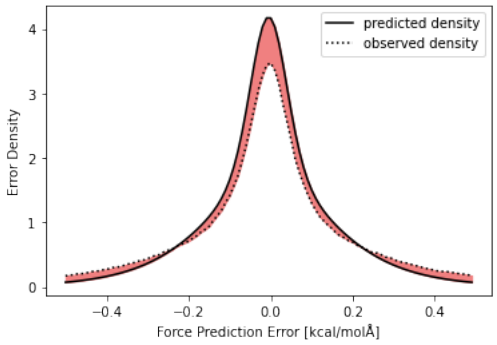


图 6：长度为 16 的 PEDOT 数据集上，使用 k=8 个蒙特卡罗样本得到的力分量的观测误差密度和预测误差密度，单位为 kcal/(mol· ˚A)。

蒙特卡罗样本数量的影响如表 3 所示，校准效果和

表 2：PEDOT 数据集上的结果。力的单位为 kcal/(mol· ˚A)，能量的单位为 kcal/mol。

Chain Length		Proposed Model (k=1)				Proposed Model (k=8)			
		MAE	RMSE	NECE	MLL	MAE	RMSE	NECE	MLL
16	Energy	15.86	23.22	0.009	-4.57	14.31	22.87	0.006	-4.33
	Forces	0.222	0.385	0.649	-1.79	0.210	0.371	0.313	-0.20
12	Energy	5.40	6.79	0.023	-3.70	5.11	6.24	0.022	-3.68
	Forces	0.184	0.276	0.652	-0.87	0.169	0.254	0.317	0.14
8	Energy	4.15	5.27	0.044	-3.59	5.37	6.54	0.045	-3.65
	Forces	0.151	0.221	0.619	-0.284	0.139	0.203	0.215	0.44

力的 MLL 均随着样本量的增加而显著提升。由于这些样本是从单条马尔可夫链中生成的，这表明所提出的采样算法能够高效遍历参数空间。此外，随着样本量增加，MLL 的提升似乎呈指数级减弱。对于能量预测这一辅助任务，则未发现类似趋势。总体而言，我们认为，$k = 8$ 个蒙特卡罗样本是在不确定性量化质量与计算复杂度之间取得良好平衡的选择。

表 3：蒙特卡洛样本数量对长度为 16 的 PEDOT 数据集的影响。力的单位为 kcal/(mol· ˚A)，能量的单位为 kcal/mol。

Number of Samples k		Proposed Model			
		MAE	RMSE	NECE	MLL
1	Energy	15.86	23.22	0.009	-4.57
	Forces	0.222	0.385	0.649	-1.79
2	Energy	15.16	23.00	0.008	-4.44
	Forces	0.219	0.381	0.473	-0.87
4	Energy	14.36	22.88	0.007	-4.34
	Forces	0.216	0.378	0.376	-0.44
8	Energy	14.31	22.87	0.006	-4.33
	Forces	0.210	0.371	0.313	-0.20
16	Energy	14.56	22.86	0.007	-4.38
	Forces	0.201	0.356	0.254	0.00

RMD17 数据集上的结果

两个模型均使用八个蒙特卡罗样本时，所提出的模型在许多较简单的 RMD17 数据集上与 Dropout 模型的准确度相当，但在阿司匹林和丙二醛数据集上显著优于后者（表 4）。在这些数据集上，所达到的准确度与原始 NequIP 模型非常一致[7]（单个蒙特卡罗样本以及 NequIP 模型在该数据集上的结果见附录 D）。此外，就力和能量而言，该模型的平均对数似然始终优于 Dropout 模型。唯一的例外是尿嘧啶能量的平均对数似然，两种模型的结果相同。

在异常值检测任务中，各模型的性能差异非常大；所提出的模型始终取得更接近最优值 1 的 ROC AUC 分数（表 5）。其优

表 4：使用 $k = 8$蒙特卡罗样本在 RMD17 数据集上得到的平均结果。力的单位为 kcal/(mol· ˚A)，能量的单位为 kcal/mol。

Molecule		Dropout			Proposed Model		
		MAE	RMSE	MLL	MAE	RMSE	MLL
Aspirin	Energy	0.072	0.096	0.88	0.045	0.067	1.41
	Forces	0.215	0.340	-1.85	0.144	0.229	0.22
Ethanol	Energy	0.012	0.021	2.61	0.010	0.017	2.99
	Forces	0.078	0.147	0.74	0.064	0.115	1.03
Uracil	Energy	0.012	0.020	2.63	0.013	0.020	2.63
	Forces	0.082	0.141	0.71	0.085	0.138	0.83
Malonaldehyde	Energy	0.021	0.038	1.82	0.016	0.027	2.51
	Forces	0.145	0.259	-0.18	0.099	0.170	0.49
Salicylic acid	Energy	0.022	0.042	1.87	0.022	0.038	2.13
	Forces	0.108	0.199	0.14	0.109	0.182	0.63
Naphthalene	Energy	0.009	0.013	2.55	0.009	0.012	2.74
	Forces	0.044	0.069	1.19	0.043	0.069	1.63
Touluene	Energy	0.011	0.015	2.63	0.010	0.014	2.79
	Forces	0.052	0.084	1.11	0.051	0.082	1.43
Benzene	Energy	0.005	0.006	3.20	0.004	0.005	3.53
	Forces	0.020	0.032	1.50	0.010	0.016	3.06

表5：使用 k=8 个蒙特卡洛样本，基于不确定性检测预测误差至少为 1 kcal/(mol· ˚A) 的力分量，在 RMD17 数据集上的 ROC AUC 分数。

Molecule	Dropout	Proposed Model
Aspirin	0.800	0.952
Ethanol	0.897	0.993
Uracil	0.888	0.986
Malonaldehyde	0.824	0.983
Salicylic acid	0.809	0.977
Napthalene	0.902	0.995
Toluene	0.894	0.995

zene 数据集未被纳入此次比较，因为对于所提出的模型，不存在达到所需规模的力预测误差实例。图4给出了完整的受试者工作特征曲线图。如图所示，我们提出的模型在检测异常值方面的可靠性高得多。

最后，正如表6 和图5 所明显显示的，在使用8个蒙特卡罗样本的情况下，所得模型并未得到准确校准，并且倾向于过度自信。

表 6：使用 k=8个蒙特卡罗样本在 RMD17 数据集上的校准误差

Molecule		NECE
Aspirin	Energy	0.388
	Forces	0.389
Ethanol	Energy	6.068
	Forces	0.758
Uracil	Energy	3.814
	Forces	0.370
Malonaldehyde	Energy	3.376
	Forces	0.631
Salicylic acid	Energy	1.269
	Forces	0.296
Naphthalene	Energy	12.314
	Forces	0.309
Touluene	Energy	8.166
	Forces	0.273
Benzene	Energy	23.006
	Forces	3.439

5 结论与展望

结果表明，所提出的贝叶斯神经网络模型达到了最先进的精度，同时也实现了良好的不确定性量化。所提出的采样算法似乎能在合理的时间内收敛到后验分布，并在此后持续快速地遍历参数空间。最重要的是，该模型能够在从同一马尔可夫链中采样参数的同时，可靠地检测异常值。这为用于学习原子间作用力的贝叶斯主动学习流程开辟了新的可能性。此外，它还可能通过贝叶斯先验分布，以有原则的方式纳入现有的、由物理规律指导的近似原子间力场。这有望进一步提高模型的数据效率和泛化能力。

尽管所提出的模型已经取得了良好结果，但仍有一些方面可以改进。
尽管使用所提出的算法可以实现后验分布的收敛，但在实践中我们发现，与 [7] 中所述的非贝叶斯优化过程相比，达到收敛仍需要大约两倍的 GPU 小时数。不过，尽管该算法对于展示所提出采样器的收敛特性很有帮助，但只需在优化初始阶段减少注入的梯度噪声和批量大小，就很可能实现更快的收敛。尽管如此，即使不采取这些措施，在大多数应用中，以从头算 精度为训练集生成原子间作用力，仍很可能成为计算瓶颈。

此外，通过适当纳入原子之间以及力分量之间的协方差，随机模型仍有可能得到进一步改进。然而，这两种协方差都难以纳入。对于力分量之间的协方差，主要挑战在于保持预-

预测的密度。对于原子间协方差的纳入，对于较大的分子而言，稠密协方差矩阵很快就会变得不切实际，因为对数似然的评估会成为计算瓶颈。尽管由于原子间作用力主要具有局部性质，稀疏协方差矩阵可能已经足够，但仍需要一种对稀疏协方差矩阵进行参数化的方法；这并非易事，而且我们在现有文献中也未能找到相关方法。

最后，我们发现，预测的不确定性并不总能得到适当校准，并且往往过度自信。这可能是从同一条马尔可夫链中采样造成的，因为模型之间始终无法实现完全独立的采样。这会导致采样模型之间的预测方差降低，从而提高置信度。在这种情况下，使用两条或更多条并行链来减少样本间的依赖，并更高效地探索后验分布，或者在验证集上重新校准不确定性，可能会有所帮助。

6 致谢

本文由 dtec.bw——德国联邦国防军数字化与技术研究中心资助，我们对此谨致谢意 [项目 CoupleIT!]

参考文献

[1] P.W. Atkins和R.S. Friedman。《分子量子力学》。牛津大学出版社，2011年。[2] M. Gastegger 和P. Marquetand。使用神经网络势进行分子动力学。在《机器学习与量子物理的交汇》中，第 233–252 页。施普林格，2020年。[3] Emir Kocer、Tsz WaiKo和J¨orgBehler。神经网络势：方法简明综述。《物理化学年度评论》，73(1):163–186，2022年。PMID:34982580。[4]Johannes Klicpera、FlorianBecker和 StephanG¨unnemann。Gemnet：用于分子的通用方向图神经网络。在 A. Beygelzimer、Y. Dauphin、P. Liang 和 J. Wortman Vaughan 编，神经信息处理系统进展，2021年。[5] Oliver T. Unke、Stefan Chmiela、Michael Gastegger、KristofT. Sch¨utt、Huziel E. Sauceda 和 Klaus- RobertM¨uller。Spookynet：学习具有电子自由度和非局部效应的力场。《自然·通讯》，12，2021年。

[6]Kristof T. Sch¨utt、Oliver T. Unke 和 Michael Gastegger。用于预-

张量性质和分子光谱的预测。*CoRR*, abs/2102.03150，2021年。

[7] Simon Batzner、Albert Musaelian、Lixin Sun、Mario Geiger、Jonathan P. Mailoa、Mordechai Kornbluth、Nicola Molinari、TessE. Smidt和BorisKozinsky。用于数据高效且准确的原子间势的E(3)等变图神经网络。《自然·通讯》，13，2022年。

[8] M. Haghighatlari, J. Li,X. Guan, O. Zhang, A. Das,C. J. Stein, F. Heidar-Zadeh, M. Liu, M. Head- Gordon, L. Bertels, H. Hao, I. Leven, andT. Head-Gordon.Newtonnet：一种用于深度学习原子间势和力的牛顿式消息传递网络。《数字发现》，1(3)：333–343，2022年。

[9]Zhuoran Qiao、Anders S. Christensen、Matthew Welborn、Frederick R. Manby、AnimaAnandkumar和ThomasF.Miller。利用电子相互作用为几何深度学习提供信息，以加速量子化学研究。《美国国家科学院院刊》，119(31):e2205221119，2022年。

[10] P. AU Reiser、M. Neubert、A. Eberhard、L. Tor- resi、C. Zhou、C. Shao、H. Metni、C. van Hoe- sel、H. Schopmans、T. Sommer 和 P. Friederich。用于材料科学和化学的图神经网络。《通讯·材料》，3(93)，2022年。

[11] Evgeny V. Podryabinkin 和 Alexander V. Shapeev。线性参数化原子间势的主动学习。计算材料科学，140：171–180，2017年。

[12]Evgeny V. Podryabinkin、Evgeny V. Tikhonov、Alexander V. Shapeev 和 Artem R. Oganov。利用主动学习和机器学习原子间势加速晶体结构预测。《物理评论*B*》，99：064114，2019年。

[13]Konstantin Gubaev、Evgeny V. Podryabinkin、Gus L.W. Hart和 Alexander V.Shapeev。利用主动学习原子间势，加速新合金的高通量搜索。计算材料科学，156：148–156，2019。

[14] Ryosuke Jinnouchi、Kazutoshi Miwa、Ferenc Kar- sai、Georg Kresse 和 Ryoji Asahi。面向大规模原子模拟的原子间势在线主动学习。《物理化学快报》，11:6946–6955，2020年。

[15]A. Shapeev、K. Gubaev、E. Tsymbalov 和 E. Podryabinkin。主动学习与不确定性估计。载于机器学习与量子物理的交汇，第 309–329 页。施普林格，2020年。

[16]Mingjian Wen 和 Ellad B. Tadmor。《利用随机失活神经网络势对分子模拟中的不确定性进行量化》。《npj 计算数学》，6:124，2020年。[17] Sung-Ho Lee、Valerio Olevano 和 Benoit Skl´enard。《一种适用于 gesbte 的可泛化、具有不确定性感知能力的神经网络势：蒙特卡罗丢弃法》。《固态电子学》，第 108508 页，2022年。[18] G. Thaler、S. and Doehner 和 J. Zavadlav。《神经网络势的可扩展贝叶斯不确定性量化：前景与陷阱》。《arxiv》，2022年。

[19] YarinGal和Zoubin Ghahramani。将随机失活作为贝叶斯近似：在深度学习中表示模型不确定性。载于第 *33* 届国际机器学习国际会议论文集——第*48*卷，ICML'16，第 1050–1059 页。JMLR.org，2016。

[20] J. Yao、W. Pan、S. Ghosh 和 F. Doshi-Velez。《贝叶斯神经网络推断中不确定性量化的质量》。发表于国际机器学习大会论文集：深度学习中的不确定性与鲁棒性研讨会（*ICML*），2019 年。

[21] Sebastian Ruder。《梯度下降优化算法概述》。*ArXiv*，abs/1609.04747，2016 年。

[22] Yi-An Ma、Tianqi Chen 和 Emily Fox。《随机梯度 MCMC 完整指南》。载于 C. Cortes、N. Lawrence、D. Lee、M. Sugiyama 和 R. Garnett 编，神经信息处理系统进展，第 28 卷。Curran Associates, Inc.，2015 年。

[23] Tianqi Chen、Emily B. Fox 和 Carlos Guestrin。《随机梯度哈密顿蒙特卡洛》。发表于*ICML*，第 1683–1691 页，2014 年。

[24]Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于 Yoshua Bengio 和 Yann LeCun 编，第三届国际学习表征会议（*ICLR 2015*）会议论文集，*2015* 年 *5* 月 *7* 日至 *9* 日，美国加利福尼亚州圣迭戈，2015 年。

[25] Sashank J. Reddi、SatyenKale和 Sanjiv Kumar。论 Adam 的收敛及其后续研究。发表于 国际学习表征会议，2018 年。

[26] Magatte N. Gueye、Alexandre Carella、Nicolas Mas- sonnet、Etienne Yvenou、Sophie Brenet、J´erˆome Faure-Vincent、St´ephanie Pouget、Fran˛cois Rieu- tord、Hanako Okuno、Anass Benayad、Renaud De- madrille 和 Jean-Pierre Simonato。PEDOT 薄膜中的结构与掺杂剂工程：显著提升电导率的实用工具。材料化学，28(10):3462–3468，2016年。

[27] Thanh-Hai Le、YukyungKim 和 HyeonseokYoon。导电聚合物的电学和电化学性质。聚合物，9(4)，2017年。

[28]Magatte N. Gueye、Alexandre Carella、J´erˆome Faure-Vincent、RenaudDemadrille 和 Jean-Pierre Simonato。对理解基于 PEDOT 的材料结构与传输特性的进展：一篇批判性综述。材料科学进展，108:100616，2020年。

[29]尼古拉斯·S·胡达克。氯铝酸盐掺杂导电聚合物作为可充电铝电池的正极。《物理化学杂志 *C*》，118(10):5203–5215，2014。

[30] Igor Zozoulenko、Amritpal Singh、Sandeep Kumar Singh、Viktor Gueskine、Xavier Crispin 和 Magnus Berggren。PEDOT 的极化子、双极化子与吸收光谱。*ACS* 应用聚合物材料，1(1):83–94，2019 年。

[31] Kazuya Ando、Shun Watanabe、Sebastian Mooser、Eiji Saitoh 和 Henning Sirringhaus。溶液加工的有机自旋–电荷转换器。*NatureMaterials*，12:622–627，2013。

[32] Stefan Chmiela、Alexandre Tkatchenko、Huziel E. Sauceda、Igor Poltavsky、Kristof T. Sch¨utt 和 Klaus-Robert M¨uller。机器学习准确且守恒能量的分子力场。科学进展，3(5)：e1603015，2017年。

[33] Anders S. Christensen 和 O. Anatole von Lilienfeld。梯度在分子能量和力的机器学习中的作用。机器学习：科学与技术，1，2020年。

[34] Mahdi Pakdaman Naeini、Gregory F. Cooper 和 Milos Hauskrecht。使用贝叶斯分箱获得校准良好的概率。载于 第二十九届 *AAAI* 人工智能大会论文集，AAAI'15，第 2901–2907 页。AAAI Press，2015年。

[35] Mario Geiger、Tess Smidt、Alby M.、Benjamin Kurt Miller、Wouter Boomsma、Bradley Dice、Kostiantyn Lapchevskyi、Maurice Weiler、Micha·l Tyszkiewicz、Martin Uhrin、Simon Batzner、Dylan Madisetti、Jes Frellsen、Nuri Jung、Sophia Sanborn、jkh、Mingjian Wen、Josh Rackers、MarcelRød和 Michael Bailey。e3nn/e3nn：2022年-12-12，2022年。

[36] Ruqi Zhang、Chunyuan Li、Jianyi Zhang、Changyou Chen 和 Andrew Gordon Wilson。用于贝叶斯深度学习的循环随机梯度 MCMC。载于国际学习表征会议，2020年。

[37] E. Apr`a, E. J. Bylaska, W. A. de Jong, N. Govind, K. Kowalski, T. P. Straatsma, M. Valiev, H. J. J. van Dam, Y. Alexeev, J. Anchell, V. Anisimov, F. W. Aquino, R. Atta-Fynn, J. Autschbach, N. P. Bau- man, J. C. Becca, D. E. Bernholdt, K. Bhaskaran- Nair, S. Bogatko, P. Borowski, J. Boschen, J. Brabec, A. Bruner, E. Cau¨et, Y. Chen, G. N. Chuev, C. J. Cramer, J. Daily, M. J. O. Deegan, T. H. Dunning, M. Dupuis, K. G. Dyall, G. I. Fann, S. A. Fischer, A. Fonari, H. Fr¨uchtl, L. Gagliardi, J. Garza, N. Gawande, S. Ghosh, K. Glaese- mann, A. W. G¨otz, J. Hammond, V. Helms, E. D. Hermes, K. Hirao, S. Hirata, M. Jacquelin, L. Jensen, B. G. Johnson, H. J´onsson, R. A. Kendall, M. Klemm, R. Kobayashi, V. Konkov, S. Krishnamoorthy, M. Krishnan, Z. Lin, R. D. Lins, R. J. Littlefield, A. J. Logsdail, K. Lopata, W. Ma, A. V. Marenich, J. Martin del Campo, D. Mejia-Rodriguez, J. E. Moore, J. M. Mullin, T. Nakajima, D. R. Nascimento, J. A. Nichols, P. J. Nichols, J. Nieplocha, A. Otero-de-la Roza, B. Palmer, A. Panyala, T. Pirojsirikul, B. Peng, R. Peverati, J. Pittner, L. Pollack, R. M. Richard, P. Sadayappan, G. C. Schatz, W. A. Shelton, D. W. Silverstein, D. M. A. Smith, T. A. Soares, D. Song, M. Swart, H. L. Taylor, G. S. Thomas, V. Tipparaju, D. G. Truhlar, K. Tsemekhman,´T. Van Voorhis, A. V´azquez-Mayagoitia, P. Verma, O. Villa, A. Vishnu, K. D. Vogiatzis, D. Wang, J. H. Weare, M. J. Williamson, T. L. Windus, K. Woli´nski, A. T. Wong, Q. Wu, C. Yang, Q. Yu, M. Zacharias, Z. Zhang, Y. Zhao, 和 R. J. Harri- son。Nwchem：过去、现在与未来。化学物理学杂志，152(18):184102，2020年。

[38] AxelD. Becke。《密度泛函热化学》。III. 精确交换的作用。《化学物理学杂志》，98(7):5648–5652，1993。

[39] John P. Perdew、MatthiasErnzerhof 和 Kieron Burke。将精确交换与密度泛函近似混合的理论依据。《化学物理学杂志》，105(22):9982–9985，1996。

[40] Vitaly A. Rassolov、John A. Pople、Mark A. Ratner 和 Theresa L.Windus。用于 K 至Zn原子的6-31G* 基组。《化学物理学杂志》，109(4)：1223–1229，1998。

[41] Stefan Grimme, AndreasHansen, Jan Gerit Bran- denburg,and Christoph Bannwarth。色散校正的平均场电子结构方法。《化学评论》，116(9):5105–5154，2016。PMID：27077966。

[42]Giovanni Bussi、Davide Donadio和 MicheleParrinello。通过速度重标度进行正则系综采样。《化学物理学杂志》, 126(1):014101, 2007。

A 实现细节

A.1 基础模型的详细信息

对于 NequIP 基础模型，我们使用了 64 的潜在维度，并纳入了截至 $l = 2$ 的所有具有偶宇称和奇宇称的不可约表示。

我们使用了一个径向网络，其径向截断为 4 Å，采用大小为 8 的可训练贝塞尔基，以及两个维度为 64、使用 SiLU 非线性的隐藏层。

偶特征和奇特征分别使用了门控 SiLU 非线性和基于 tanh 的非线性。模型采用了包含 5 个交互块的基础结构。

模 型基于 e3nn 库 [3，由官方 NequIP 软件包构建。 5].

A.2 随机模型的详细信息

为了根据 NequIP 基础模型生成的不变特征计算力和能量的标准差，我们使用了两个独立的全连接前馈神经网络。这些神经网络的输入层维度为 64，包含两个隐藏层，维度分别为 32 和 16，输出层维度为 1。

隐藏层使用 SiLU 激活函数，输出层使用指数函数作为激活函数。

用于力标准差的神经网络将每个不变特征 $\boldsymbol{v}_i$ 直接映射到标准差 $\sigma_{\boldsymbol{F}_i}$。用于能量标准差的神经网络将每个不变特征 $\boldsymbol{v}_i$ 映射到标量潜变量 u_i，然后将标准差计算为 $\sigma_E = \frac{1}{n}\sum_{i=1}^{n} u_i$。

我们不对力标签和能量标签进行归一化，而是将力和能量的预测均值按在训练数据集上评估得到的力分量均方根进行重新缩放。我们不重新缩放标准差。

为了平滑预测分布，以便使用多个蒙特卡罗样本评估对数似然，我们根据预测分布的均值和方差拟合了一个正态分布，然后使用该正态分布评估对数似然。

A.3 采样算法的详细信息

对于所有实验，所有模型均从单个马尔可夫链中采样。我们始终将 α 设置为 0.1，将 β 设置为 0.999，并将 ϵ设置为 10^{-3}。

在 RMD17 数据集上，批量大小设为 32，学习率 γ 在前 10^6 训练步数内从 10^{-2}指数衰减至 10^{-5}。在第 10^6 步后，对第一个模型进行采样。随后，我们采用了与 [36] $\gamma_i = \frac{\gamma_0}{2}\left(\cos\left(\pi + \frac{i \cdot \pi}{K}\right) + 1\right)$ 中提出的方法非常相似的循环学习率调度，并使用 $\gamma_0 = 0.001$和周期长度 $K = 50000$ 从同一马尔可夫链中采样后续模型。每个周期后采样一个模型。在前 $9 \cdot 10^5$ 训练步数后，质量向量 $\boldsymbol{M}$ 保持不变。

PEDOT 数据集使用了相同的采样流程，但批量大小设为 10，初始学习率衰减和随后的采样发生在前 $5 \cdot 10^5$ 个训练步数内，并且质量向量 $\boldsymbol{M}$ 在前 10^5 个训练步数后保持不变。

A.4 Dropout 神经网络的详细信息

对于 Dropout 架构，在随机模型中，NequIP 基础模型之后添加了一个 Dropout 层。使用 AMSGrad 优化器对生成的模型进行优化，其批量大小和学习率计划与马尔可夫链采样算法初始收敛阶段所用的相同。目标函数采用平均对数似然。

我们使用了 1/64 的 Dropout 率，使得平均每个原子能量贡献中有一个激活值被随机失活。我们发现，必须使用这种相当低的 Dropout 率，因为更高的比率会显著降低所得模型的准确性，这与 [16] 中的发现一致。

A.5 校准误差和误差密度的评估 s

在校准误差和观测误差密度的评估中，我们采用分箱宽度 $\delta = 0.01$。为可视化测试数据集上误差分量（$\boldsymbol{F}_i - \boldsymbol{F}_{pred\,i}$）$i = 1$、..... n 的观测分布，我们将观测误差密度设为 $\rho_{pred}(x) = \frac{f_x}{\delta}$，其中 f_x 是落入区间 $[x, x+\delta$ 的误差所占比例。

B PEDOT 数据集

使用 NWChem 6.8 [37]，进行 PEDOT 的 从头算分子动力学，采用杂化密度泛函理论：B3LYP[38, 39] 6-31G* [40]，并在几何弛豫中采用 D3(BJ) 色散校正 [41]。先前的工作表明，弥散函数不会显著改变结果。关闭了阻尼、迭代子空间直接反演（DIIS）和能级位移，并使用二次收敛算法。结构弛豫使用 DRIVER 模块完成，该模块采用拟牛顿优化。使用其默认值：最大梯度为 0.00045，均方根梯度为 0.00030；最大笛卡尔步长为 0.00180，均方根笛卡尔梯度为 0.00120。分子动力学本身采用随机速度重标度（SVR）恒温器 [42]，，弛豫时间参数（τ）为 25 fs，时间步长为 0.5 fs。

训练集由长度为8 和长度为 12 的PEDOT在前 1500个时间步长中随机采样的 50 个构型组成。等价地，验证集由长度为 8 和长度为 12的 PEDOT 在第1501-1600个时间步长中随机采样的15个构型组成。第1601-2000 个时间步长分别作为相应长度的测试集。长度为16 的 PEDOT 的测试集包含全部 2000 个构型。对于每个 PEDOT 分子，我们将基态几何结构的势能定义为零。

C 力预测直方图

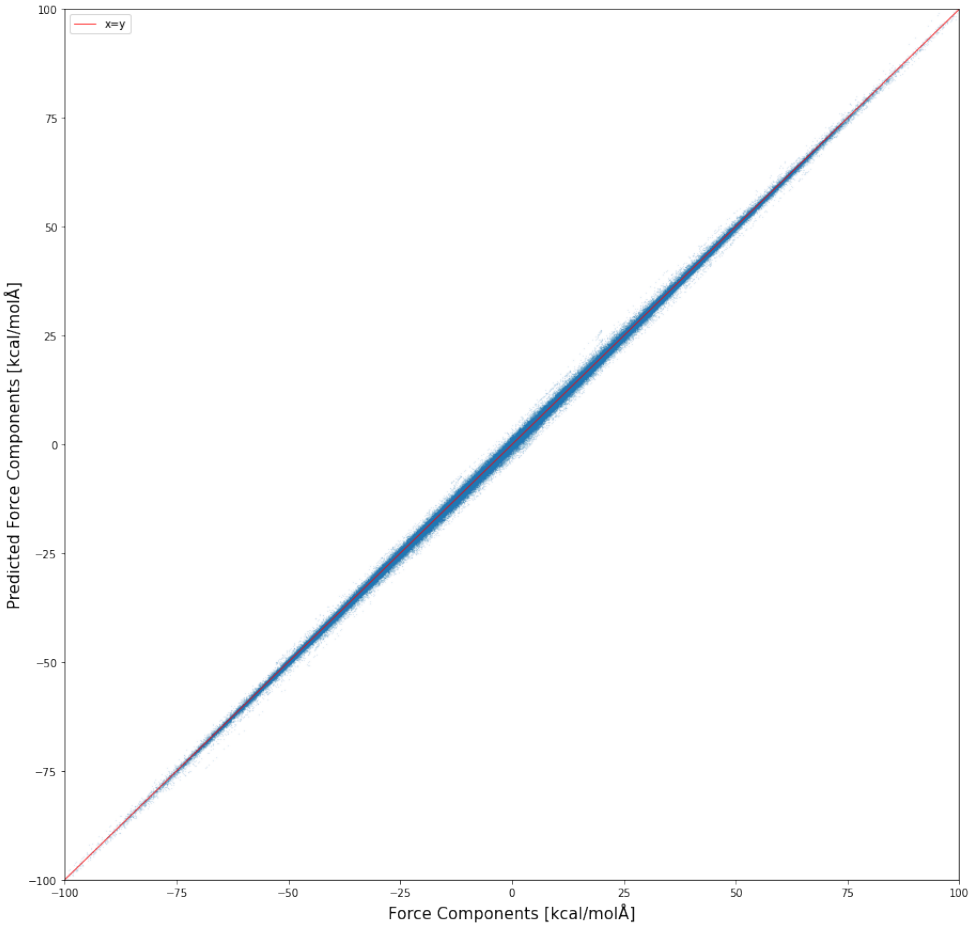


图 7：长度为16 的 PEDOT数据集上，以 kcal/(mol· ˚A)为单位的观测力分量和预测力分量，其中包含 k=8 蒙特卡罗样本。

D RMD17数据集上的其他结果

表 7：在 RMD17数据集上，将NequIP架构用作基础模型，以及随机模型的贝叶斯后验分布的单个蒙特卡罗样本的结果。两个模型均使用相同的训练、验证和测试数据集生成。

Molecule		Base Model		Proposed Model (k=1)			
		MAE	RMSE	MAE	RMSE	NECE	MLL
Aspirin	Energy	0.060	0.087	0.047	0.068	1.63	0.77
	Forces	0.197	0.308	0.154	0.244	1.03	-1.76
Ethanol	Energy	0.013	0.023	0.010	0.018	6.51	1.72
	Forces	0.078	0.139	0.067	0.117	2.69	-2.16
Uracil	Energy	0.010	0.017	0.014	0.021	4.21	2.07
	Forces	0.068	0.115	0.089	0.145	0.96	0.20
Malonaldehyde	Energy	0.020	0.036	0.016	0.029	4.35	1.04
	Forces	0.131	0.227	0.104	0.178	2.60	-8.68
Salicylic acid	Energy	0.020	0.046	0.025	0.042	4.33	0.95
	Forces	0.098	0.182	0.117	0.194	0.79	-0.30
Naphthalene	Energy	0.010	0.013	0.010	0.013	5.55	2.89
	Forces	0.035	0.055	0.046	0.074	1.25	1.31
Touluene	Energy	0.009	0.012	0.011	0.015	3.32	2.73
	Forces	0.044	0.070	0.055	0.087	1.35	0.92
Benzene	Energy	0.004	0.005	0.004	0.005	6.34	3.88
	Forces	0.008	0.012	0.010	0.016	0.04	3.06