arXiv:2305.01754v1 [cs.LG] 2 May 2023

神经网络势中的单模型不确定性量化并不能始终优于模型集成

Aik Rui Tan[1], ShingoUrata[2], Samuel Goldman[3], JohannesC. B.Dietschreit[1], RafaelG´omez-Bombarelli[1*]麻省理工学院材料科学与工程系。[2]AGC株式会社创新技术实验室。[3]麻省理工学院计算与系统生物学系。

*通讯作者。电子邮箱：rafagb@mit.edu；

摘要

神经网络（NN）往往对其预测赋予很高的置信度，即使面对远离分布的数据点也是如此，这使不确定性量化（UQ）成为一项挑战。当神经网络用于对材料体系中的原子间势进行建模时，这一问题会导致产生干扰模拟的非物理结构，或造成无法反映真实物理规律的有偏统计与动力学。可微分的不确定性量化技术能够发现新的有效数据，并推动用于构建鲁棒势的主动学习循环。然而，原子模拟中存在多种不确定性量化技术，包括新近开发的方法，但目前没有明确的指南说明哪些方法最有效，或最适用于特定情形。在本研究中，我们考察了多种通过主动学习提升神经网络原子间势（NNIP）鲁棒性的不确定性量化方案。具体而言，我们将当前常用的基于集成的方法与采用单个确定性神经网络的策略进行比较：均值-方差估计、深度证据回归和高斯混合模型。我们探索了三个数据集，涵盖从域内插值学习到更具外推性的域外泛化挑战：rMD17、氨反转和块体二氧化硅玻璃。我们通过多个关联模型误差与不确定性的指标来衡量性能。实验结果表明，在各种指标上，没有一种方法始终优于其他方法。集成在泛化能力和神经网络原子间势的鲁棒性方面仍然表现更好；均值方差估计仅在域内插值中有效，而高斯混合模型在域外表现更好；并且

尽管证据回归具有一定的潜力，但在所有案例中，它都不是更优的替代方案。更广泛而言，在神经网络原子间势（NNIPs）的不确定性量化方面，目前成本效益较高的单一确定性模型尚无法始终达到或超越集成方法的表现。

关键词： 不确定性量化、神经网络原子间势、对抗采样、单一确定性神经网络、神经网络集成

1 引言

在过去十年中，神经网络（NN）越来越多地被用于研究复杂材料体系。除了对定量构效关系/定量构效关系（QSPR/QSAR）进行建模外，神经网络还被广泛用于材料原子模拟中的原子间势建模[1–5]。当用于预测材料体系的势能面（PESes）时，神经网络原子间势（NNIPs）能够达到从头算方法的精度，同时速度提高几个数量级，从而支持更大时间尺度和长度尺度的模拟[6, 7]。此外，神经网络原子间势展现出足够的灵活性，能够学习不同的化学组成和结构[8–10]。因此，研究人员开发了多种神经网络原子间势架构[11–16] ，并将其应用于反应过程[17, 18], 蛋白质设计[19, 20], 固体[21, 22],固液界面[23], 粗粒化[24, 25],等领域。然而，神经网络原子间势仍容易在外推区域产生较差的预测结果。这通常会导致原子模拟过程中出现非物理结构或不准确的热力学预测，从而损害研究的科学有效性[26, 27]。

为最大限度提升神经网络原子间势的鲁棒性并避免分布偏移，训练数据应能代表模拟将遍历的同一系综。然而，由于高质量的从头算计算成本高昂，通过穷尽探索化学空间来获取新的数据点并不可行。因此，量化模型不确定性和主动学习是训练鲁棒神经网络原子间势的关键。良好的不确定性量化（UQ）方法应同时估计由测量噪声引起的不确定性（即偶然不确定性），以及由模型误差导致的预测不确定性（即认知不确定性）[28]。认知不确定性源于数据稀缺、模型架构局限和/或模型参数优化不佳。由于通过从头算计算得到的训练数据不存在偶然误差，神经网络原子间势主要受到认知不确定性的影响，这是因为它们的插值能力很强，而外推能力较弱。因此，使用良好的UQ指标高效获取新的训练数据，对于促进主动学习并减轻认知不确定性至关重要[29, 30]。在众多UQ方案中，贝叶斯神经网络（BNN）将权重和输出视为概率分布，提供了一种内在的UQ方案[31, 32]。然而，BNN存在扩展性问题，用作块体材料体系的原子间势时，计算成本高得令人难以承受。另一种常用于UQ的架构是神经网络集成[33]。神经网络集成

提供了一种直接的不确定性估计方法，因为神经网络权重初始化、训练乃至超参数选择都具有随机性。尽管神经网络集成简单易用且能够达到较高精度，但对于大型系统，训练和推理由多个独立神经网络组成的集成所需的计算成本可能非常高。其他不确定性估计方法试图通过单个确定性神经网络估计不确定性，以解决这些瓶颈。由于单个神经网络模型的训练和推理计算成本较低，这类策略尤其值得关注。此外，只需对数据进行一次评估，就能估计单个确定性神经网络的预测不确定性，无需通过随机采样来近似底层不确定性函数。采用单个确定性神经网络的不确定性量化策略，大体可分为前置方法和后置方法[34]。在后置方案中，神经网络经过训练后，根据数据集在特征空间中的表示[35]，潜在空间[36]，或训练所得神经网络的梯度矩阵[37，38]来估计预测器不确定性。Janet 等人提出了一种后置方法[36]，用于非神经网络原子间势应用，并通过潜在特征中的欧几里得范数距离来近似不确定性。在此基础上，近期的一项进展是在神经网络的潜在空间上拟合高斯混合模型[39]。另一方面，在前置方法中，会对输入数据施加先验分布[40，42] 或对神经网络的训练目标施加约束[43]。均值-方差估计方法是一种回归先验方案，也是常用的前置方案：它对输入数据施加高斯先验分布，并训练神经网络预测均值和方差（方差也被视为不确定性）[40]。近期提出的另一种前置方法是深度证据回归，它在输入数据之上施加高阶先验分布[41]。

据我们所知，目前尚无任何不确定性量化方法在神经网络原子间势（NNIP）领域显著优于其他所有方法。事实上，针对这一任务，不确定性量化方法之间似乎缺乏比较。与物理性质预测（定量构效关系/定量构性关系，QSAR/QSPR）不同，神经网络原子间势除了拟合势能面外，还必须依赖力匹配梯度，这使得神经网络原子间势应用中的不确定性量化更具挑战性，也更为独特[44，45]。在本研究中，我们探讨采用单个确定性神经网络的不确定性量化方法，能否在降低计算成本的同时，持续地在生成高质量不确定性排序方面优于神经网络集成。具体而言，我们采用均值-方差估计（MVE）、深度证据回归和高斯混合模型（GMM）。我们使用多项指标，在三个不同数据集上评估这些方法的不确定性排序表现，并分析通过主动学习提升分子动力学模拟稳定性的效果。总体而言，我们发现，基于集成的方法在训练数据域之外的不确定性排序方面表现稳定，并能提供最稳健的神经网络原子间势。研究表明，MVE主要擅长识别训练域内误差较大的数据点。深度证据回归对认知不确定性的预测准确度较低，而高斯混合模型比深度证据回归和MVE更准确、计算开销也更小。

2 方法

给定一个包含一组元组（x, y）的数据集 $\mathcal{D}$，神经网络接收输入 x，并学习一个假设函数来预测相应的输出 $\hat{}y$。对于神经网络原子间势（NNIP）， x 包含材料体系中各原子的原子序数$\mathbf{Z} \in \mathbb{Z}_+^n$ 和核坐标$\mathbf{r} \in \mathbb{R}^{n\times 3}$，其中原子数为$n$；而 $\hat{}y$ 表示其势能值 $\hat{}\hat{}$，即 $E \in \mathbb{R}$，和/或作用于原子i的能量守恒力$\mathbf{F}_i$。能量守恒力$\mathbf{F}_i$ 是通过对预测势能关于原子坐标 r_i求负梯度计算得到的。

$$\hat{\mathbf{F}}_i = -\frac{\partial \hat{E}}{\partial r_i} \tag{1}$$

采用可极化原子相互作用神经网络（PaiNN）这一等变消息传递神经网络来学习原子间势[12], ，因为它兼顾了速度与精度。架构和训练的详细信息见补充信息第 I 节。

2.1 不确定性量化（UQ）

比较了四种不确定性量化（UQ）方法：基于集成的不确定性[33], 均值-方差估计（MVE）[40], 深度证据回归[41],和高斯混合模型（GMM）[39]。由于力是能量的导数，因此在过拟合或处于训练域之外时，力通常表现出更大的变化性，因此比能量方差更能描述认知不确定性。因此，对所有模型的不确定性估计均按照 σ_F [29]进行评估。

基于集成的模型是深度学习中最常见、最具成本效益且最可靠的不确定性量化方法之一。由于神经网络训练中存在多种随机性来源，从初始化到随机优化或超参数选择，基于集成的模型可以得到预测输出值的分布，该模型由 M 个不同的神经网络[29, 33, 34, 46]组成。对于输入系统 x，每个神经网络 m 都会预测总能量 $E_m(x)$和原子力$\mathbf{F}_m(x)$（使用公式1计算）。随后，将集成模型的预测值定义为各个神经网络预测值的算术平均值

$$\mu_E(x) = \frac{1}{M}\sum_m^M \hat{E}_m(x), \tag{2}$$

$$\boldsymbol{\mu}_F(x) = \frac{1}{M}\sum_m^M \hat{\mathbf{F}}_m(x) \tag{3}$$

Lakshminarayanan 等人提出，可以将集成中各个神经网络的预测方差用作不确定性估计，因为预测值

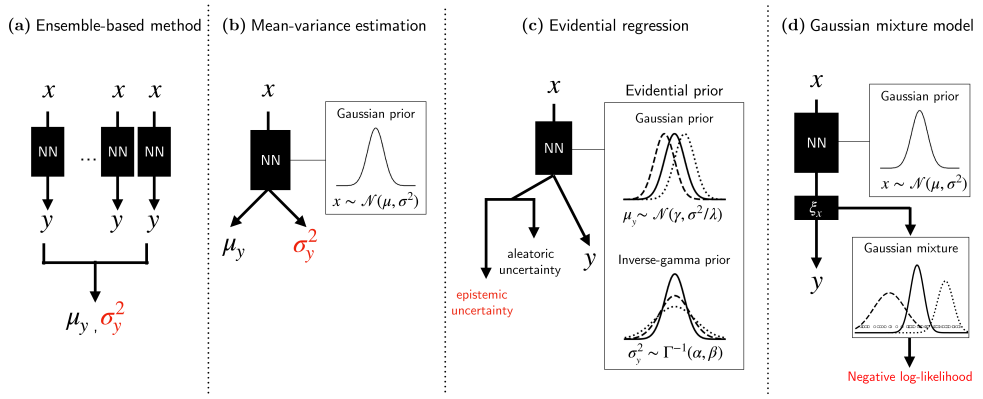

图 1 所用不确定性量化（UQ）方法示意图。 x 表示神经网络（NN）的输入，而 y 表示预测属性。对于 NNIP， x通常表示输入结构的位置和原子序数，而 y 表示相应输入结构的能量和/或力。此外，红色文字表示用作不确定性估计的变量。 **(a)** 多个经集成训练的神经网络对同一结构的目标属性进行预测。随后可计算该属性的均值和方差；属性变化越大，表示不确定性越高[33, 46]。 **(b)** 在均值-方差估计（MVE）方法中，对输入数据应用高斯先验分布，并使神经网络预测描述该高斯分布的均值和方差。方差参数越大，表示不确定性越高[40, 46]。 **(c)** 在深度证据回归方法中，对输入数据应用证据先验分布，并由神经网络预测目标属性以及描述偶然不确定性和认识不确定性的参数[41, 42]。 **(d)** 在高斯混合模型（GMM）方法中，假设输入数据来自多个高斯分布。根据结构的学习得到的特征向量拟合高斯混合模型后，计算其负对数似然（NLL）函数，ξ_x 。NLL 值越高，表示不确定性越高[39]。

神经网络在训练数据域之外[33]的不确定性可能会更高。在神经网络原子间势（NNIPs）的应用中，不确定性估计通过预测能量和力的方差来提供，这些方差可以通过以下方式计算：

$$\sigma_E^2(x) = \frac{1}{M-1}\sum_m^M (\hat{E}_m(x) - \mu_E(x))^2 \tag{4}$$

$$\sigma_F^2(x) = \frac{1}{M-1}\sum_m^M \left[\frac{1}{3n}\sum_i^n (\hat{\mathbf{F}}_{i,m}(x) - \boldsymbol{\mu}_{F_i}(x))^2\right] \tag{5}$$

均值-方差估计（MVE）。作为一种函数逼近器，标准神经网络（NN）本身并不具备评估其预测可信度的能力。将训练数据集视为具有方差的高斯随机变量后，可以构建神经网络，使其预测参数化高斯分布 $\mathcal{N}(\mu, \sigma^2)$[40] 的均值和方差对。换言之，神经网络的输出是一个由均值和方差参数化的分布，而不是一个固定点。

这种方法称为均值-方差估计（MVE），神经网络（NN）通过最大似然估计为每个数据点预测均值和方差。

对神经网络的最后一层进行修改，使其同时预测每个数据点的均值 μ_{E_i}（随后计算梯度以获得 μ_{F_i}）和方差 σ_F^2，并应用 softplus 激活函数，以确保方差严格为正。随后将预测方差作为不确定性度量。与集成方法类似，不确定性估计仅由原子力的方差来描述。

深度证据回归。尽管均值方差估计已广泛用于不确定性量化，但用作适当不确定性替代指标的预测方差，无法提供认知不确定性的估计。这是因为均值方差估计所依据的假设是，每个数据点的预测方差仅描述数据上的高斯分布，而不描述神经网络参数上的分布。因此，均值方差估计无法严格反映训练数据区域之外的分布[47]。为解决这一局限，人们开发了一种新兴的证据回归算法，通过对学习到的神经网络表征的似然函数进行概率分布参数化，直接处理认知不确定性的表征[41]。

与 MVE 直接对数据集设置先验分布不同，证据回归是在数据集的似然函数上设置先验分布。假设数据集来自均值和方差未知的独立同分布（i.i.d.）高斯分布（$\mathcal{N}(\cdot)$）（μ, σ^2）。这里假设数据由原子力表示，$\mathbf{F}_i$。在此假设下，未知均值可描述为高斯分布，而方差则服从逆伽马分布（$\Gamma(\cdot)$）[41]。

$$\begin{aligned}(\mathbf{F}_1, \mathbf{F}_2, ..., \mathbf{F}_N) &\sim \mathcal{N}(\mu_F, \sigma_F^2) \\ \mu_F \sim \mathcal{N}(\gamma_F, \sigma^2\nu^{-1}), &\quad \sigma_F^2 \sim \Gamma^{-1}(\alpha, \beta)\end{aligned} \tag{6}$$

其中，γ、ν、α 和 β 由神经网络预测，而 $\gamma \in \mathbb{R}$、$\nu > 0$、$\alpha > 1$ 和 $\beta > 0$。因此，这个高阶证据分布可表示为均值和方差上的分布：

$$p(\mu_F, \sigma_F^2|\gamma_F, \nu, \alpha, \beta) = \frac{\beta^\alpha\sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi\sigma_F^2}}\left(\frac{1}{\sigma_F^2}\right)^{\alpha+1}\exp\left(-\frac{2\beta + \nu(\gamma_F - \mu_F)^2}{2\sigma_F^2}\right) \tag{7}$$

详细推导请参见参考文献 [41, 47]。与均值方差估计类似，我们修改神经网络的最后一层，使其预测 γ_E（γ_F 随后根据公式中所示的导数计算1），ν，α 和 β。对这 3 个参数应用 softplus 激活函数 ν，α 和 β，以确保它们严格为正（$+1$ 对于 α，则取 $\mathbb{E}[\sigma_F^2]$。最后，分别按如下方式计算偶然不确定性、$[\mu_F]$ 和认知不确定性的 Var

$$\mathbb{E}[\sigma_F^2] = \frac{\beta}{\alpha - 1} \tag{8}$$

$$\mathrm{Var}[\mu_F] = \frac{\beta}{\nu(\alpha - 1)}. \tag{9}$$

高斯混合模型（GMM）。GMM 对数据集的分布进行建模，将其表示为由以下各项加权混合而成： K 个维度为 D[48]的多元高斯分布。混合中的每个D维高斯分布 k 均具有一个均值向量 $\mu_k \in \mathbb{R}^{D\times 1}$ 和一个协方差矩阵 $\Sigma_k \in \mathbb{R}^{D\times D}$。数据分布可描述如下：

$$p(x|\pi_k, \mu_k, \Sigma_k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k), \tag{10}$$

其中， π_k 为混合权重，并满足约束条件 $\sum_{k=1}^{K} \pi_k = 1$。首先，从训练好的 NN 中紧邻逐原子能量预测 [39]之前的层提取每个训练集数据点 $\xi_{\text{train}} \in \mathbb{R}^{D\times 1}$ 的潜在特征。使用期望最大化（EM）算法，在 ξ_{train} 上构建高斯混合模型（GMM）；其协方差矩阵为满秩矩阵，均值向量则通过 K 均值聚类预先确定。为确定所需的高斯分布数量，我们研究了贝叶斯信息准则（BIC）与轮廓系数之间的权衡（见图 S4 和 S5）。利用拟合好的 GMM，我们可以通过获取测试集数据的潜在特征 $\xi_{\text{test}} \in \mathbb{R}^{D\times 1}$ ，并计算测试集数据的负对数似然函数 NLL($\xi_{\text{test}}|\xi_{\text{train}}$)，来评估其不确定性。对于所有模型，潜在特征的维度 D 均设为 128。较高的 NLL($\xi_{\text{test}}|\xi_{\text{train}}$) 表明数据与 GMM 中高斯分布的均值向量相距较远，因此对应较高的不确定性。

$$\text{NLL}(\xi_{\text{test}}|\xi_{\text{train}}) = -\log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\xi_{\text{test}}|\mu_k, \Sigma_k)\right) \tag{11}$$

2.2 数据集

修订版 MD17 (rMD17)。rMD17 数据集包含十种小型有机分子长时间分子动力学轨迹中的快照[49, 50],，其能量和力是在电子结构理论的 PBE/def2-SVP 级别下计算得到的。每种分子有 5 个划分，每个划分包含 1000 个训练数据点和 1000 个测试数据点。该数据集取自 https://doi.org/10.6084/m9. figshare.12672038.v3。由于测试集是从与训练数据相同的轨迹中随机抽取的，因此 rMD17 更能代表域内插值型不确定性量化挑战。

氨。氨训练数据集由 78 个几何构型组成，其能量和力采用 ORCA 中实现的 BP86-D3/def2-SVP 理论水平计算得出。这些几何构型通过沿初始分子构象的正规模式向量进行 Hessian 位移生成；初始分子构象使用 RDKit 和 MMFF94 力场生成[29]。该数据集取自 https://doi.org/10.24435/materialscloud:2w-6h。测试集包含 200 个几何构型，能量范围为 0 至 100 千卡/摩尔（每个 1 千卡/摩尔区间内有 2 个几何构型），并采用相同的理论水平计算得出。这些几何构型是在本研究的试运行期间通过对抗采样生成的，与本研究分析的模型和数据集完全独立。由于测试集由高能几何构型组成，而训练集包含基态附近的低能结构，

氨数据集是域外外推不确定性量化挑战的一个基础示例。

二氧化硅玻璃。二氧化硅玻璃 (SiO_2) 训练数据集包含 1590 个结构，这些结构取样自多条较长的分子动力学轨迹；这些轨迹采用文献 [51] 所述的力匹配势，在不同条件下进行模拟，包括不同速率下的温度平衡、冷却、单轴拉伸、单轴压缩、双轴拉伸、双轴压缩和双轴剪切。每个二氧化硅玻璃结构包含 699 个原子（233 个 Si 原子和 466 个 O 原子）。总计从分子动力学轨迹中取样了 1590 个二氧化硅结构，构成训练集；另通过对抗采样选取了 402 个结构（这些样本是在本研究的初步测试中获得的，与本文所报告的结果分开采集），以构成测试集。使用 Vienna Ab-initio Simulation Package（维也纳从头算模拟软件包，VASP）对这些结构进行了密度泛函理论计算[53–56]。补充信息第 II.A 节讨论了分子动力学模拟[57, 58],、密度泛函理论计算以及二氧化硅结构的对抗采样详情。由于二氧化硅数据集具有较高的构型复杂性——包括高能断裂构型——但化学复杂性较低，因此它代表了不确定性量化在域外外推区域泛化能力的一项进步。

2.3 评估指标

斯皮尔曼秩相关系数。由于不确定性量化应衡量模型对所提供数据的未知程度，我们预期真实误差较大时，预测不确定性也较高。换言之，对于良好的不确定性估计器，我们预期预测不确定性 U 与真实误差 ϵ 之间存在单调关系。由于我们特别关注在主动学习中利用不确定性量化来改进神经网络原子间势，使其在误差最大的区域得到提升，因此我们使用斯皮尔曼秩相关系数来评估不确定性估计器的可靠性，从而量化预测不确定性的秩变量 R_U 与真实误差的秩变量 R_ϵ 之间的相关性。相关系数可定义为

$$\rho(U,\epsilon) = \frac{\mathrm{cov}(R_U, R_\epsilon)}{\sigma_{R_U}\sigma_{R_\epsilon}} \tag{12}$$

其中，cov(R_U， R_ϵ) 是秩变量的协方差，而 σ_{R_U} 和 σ_{R_ϵ} 是秩变量的标准差。对于完美的单调函数，预期斯皮尔曼相关系数为 $+1$ 或 -1，而相关系数为 0 则表示两个变量的秩之间不存在关联。然而，由于该指标仅比较变量的秩，因此即使估计器在误差较大时始终预测较低的不确定性，仍有可能获得较高的斯皮尔曼相关系数。因此，结合另一项指标进行评估，有助于更全面地了解其表现。

受试者工作特征曲线下面积（ROC-AUC）。由于预期真实误差较高的点具有较高的不确定性估计，因此建立了一项将不确定性和误差划分为高值和低值的标准[39, 60]。更准确地说，指定一个误差阈值 ϵ_c，使得真实误差高于该阈值的数据 马会有较高的真实误差。

高于误差阈值的 $\epsilon > \epsilon_c$ 被归类为“高误差”点，而真实误差较低的数据 $\epsilon \leq \epsilon_c$ 被归类为“低误差”点。类似地，将指定的不确定性阈值应用于不确定性 U_c。只有当真实误差和估计的不确定性都高于各自对应的阈值时，该结构才被视为真正例（TP）(($\epsilon > \epsilon_c$) $\wedge$ ($U > U_c$))。假正例（FP）点是指真实误差低于其阈值，而估计的不确定性高于阈值的结构 (($\epsilon \leq \epsilon_c$) $\wedge$($U > U_c$))。相反，当 ($\epsilon \leq \epsilon_c$) $\wedge$ ($U \leq U_c$) 时，该点为真负例（TN）；当 ($\epsilon > \epsilon_c$) $\wedge$ ($U \leq U_c$) 时，该点为假负例（FN）。

然而，阈值的设定需要针对每种不确定性方法进行手动优化，尤其是因为本研究中比较的各种方法所对应的不确定性的幅度和分布存在很大差异。与以往研究中评估真正率（TPR）和预测阳性值（PPV）的方法不同[39, 60]，我们计算受试者工作特征（ROC）曲线下面积（AUC），以便在不受判别阈值影响的情况下评估不确定性方法的质量[43]。需要注意的是，为了绘制ROC曲线，仍然必须指定“目标”得分的阈值，在本例中，该阈值即为真实误差的阈值，ϵ_c。对于所有模型，我们将 ϵ_c 设定为误差分布的第20百分位数，因为我们发现，改变误差百分位数阈值并不会显著影响ROC-AUC得分（见表S2）。我们还发现，使用依赖于系统的绝对截断值与使用百分位数截断值所得到的趋势相同。方法越准确，ROC-AUC得分就越高。ROC-AUC得分为0.5表示该方法无法区分高不确定性样本和低不确定性样本。得分低于0.5则意味着该方法在反转不确定性预测方面表现更好，即对高误差点预测出较低的不确定性。

失校准面积。评估不确定性估计质量的另一种方法，是量化模型的校准误差，即衡量校准曲线准确性的数值指标。校准曲线展示了数据点落在不同范围内的预测频率与观测频率之间的关系。Tran 等人 [61]提出，通过比较每个数据点的预期误差比例（即其误差落在均值的 z 个标准差范围内的比例）与根据不确定性估计构造的观测高斯随机变量（以 $U(x)$ 作为方差）来评估校准效果。换言之，如果不确定性估计质量良好，我们预计使用不确定性估计[46, 59, 61, 62]构造的高斯分布中，约有 68% 的误差落在均值的一个标准差范围内。根据这条校准曲线，我们可以计算曲线与完美校准线之间的面积，以提供一个校准情况的单一数值概括。这个量称为失校准面积。完美的不确定性量化方法会使观测比例与预期比例完全一致，因此失校准面积为 0。

校准负对数似然函数 (cNLL)。对于观测到的（真实）误差，可以假设误差服从均值为 0、方差由预测不确定性给定的正态分布，并据此计算负对数似然作为评估指标。然而，鉴于某些方法不会直接产生等同于方差的不确定性估计，可以对不确定性估计进行校准，使其更接近方差[36, 46]。具体而言， the

“方差”可以近似为 $\hat{\sigma}^2(x) = aU(x) + b$，其中标量可以是 a，而 b 可以估计为使验证集中的误差的负对数似然函数最小化的值，$\mathcal{D}_{val}$，

$$a_*, b_* = \underset{a,b}{\operatorname{argmin}} \frac{1}{2} \sum_{x,\mathbf{F}\in\mathcal{D}_{val}} \ln(2\pi) + \ln(aU(x)+b) + \frac{(\hat{\mathbf{F}}-\mathbf{F})^2}{(aU(x)+b)} . \tag{13}$$

利用校准后的标量 a 和 b，随后即可在测试集 $\mathcal{D}_{test}$ 上计算 cNLL 指标，结果为

$$\mathrm{cNLL} = \frac{1}{2} \sum_{x,\mathbf{F}\in\mathcal{D}_{test}} \ln(2\pi) + \ln(a_*U(x)+b_*) + \frac{(\hat{\mathbf{F}}-\mathbf{F})^2}{(a_*U(x)+b_*)} . \tag{14}$$

2.4 对抗采样

除了使用斯皮尔曼秩相关系数和 ROC-AUC 等指标评估不确定性估计器的可靠性之外，我们还希望评估：使用根据每种方法估计的不确定性采样的新结构，是否能够以不同方式提高神经网络原子间势的鲁棒性。为此，我们采用了 Schwalbe-Koda 等人提出的对抗采样方法。[29] 该方法用于采样使估计不确定性最大化的新构象，从而将这些对抗样本纳入神经网络的重新训练中，形成主动学习（active learning）循环。这将测试不确定性量化方法的可靠性、平滑性和实用性，因为对抗采样要求不确定性量化方法系统地识别模型最不确定的代表性不足区域。实际上，主动学习后神经网络原子间势鲁棒性的提升，是衡量不确定性估计方法的更好指标，因为理想情况下，良好的不确定性估计器能够获取新的训练样本，使模型的泛化能力比随机采样得到更大的提升。

在对抗采样策略中，扰动 δ 会施加到从初始训练数据 x 中随机选取的结构上。随后，通过梯度上升迭代更新该扰动，以最大化扰动结构的不确定性 $U(x_\delta)$，定义如下

$$\max_{\delta} p(x_\delta)U(x_\delta). \tag{15}$$

更确切地说，$U(x_\delta)$ 在基于集成的方法（方程5）和 MVE 中取为 $\sigma_F^2(x_\delta)$；在证据回归中，$\mathrm{Var}[\mu_F(x_\delta)]$ 取为（方程9），而 GMM 中的 $\mathrm{NLL}(x_\delta|x_{\mathrm{train}})$（方程11）。$p(x_\delta)$ 是根据模型已获得数据的分布计算出的扰动结构的热力学概率

$$p(x_\delta) = \frac{\exp\left(-\frac{\hat{E}(x_\delta)}{k_\mathrm{B}T}\right)}{\sum_{x\in\mathrm{train}} \exp\left(-\frac{E(x)}{k_\mathrm{B}T}\right)} \tag{16}$$

其中， k_{B} 是玻尔兹曼常数， T 是绝对温度，分子中的 $\hat{E}(x_{\delta})$ 是扰动结构的预测能量，而分母中的 $E(x)$ 表示初始训练数据集内结构的真实能量。关于对抗采样所用的参数，请参见补充信息第三节。

2.5 分子动力学（MD）模拟。

基于神经网络的分子动力学模拟是在 NVT系综中使用 Nos´e-Hoover 恒温器进行的。对于氨，在四种不同温度（300 K、500 K、750 K 和 1000 K）下，针对每种不确定性量化方法，均以 0.5 fs 的时间步长运行了 100 条时长为 5 ps 的分子动力学模拟。所有模拟的初始构型均从包含 78 个几何构型的初始训练数据集中随机选取。如果原子间距小于 0.75 $^{\circ}$A 或大于 2.25 $^{\circ}$A，或者预测能量低于基态能量，则认为轨迹不稳定；在本研究中，基态能量的参考值为 0 千卡/摩尔。对于二氧化硅模拟，在所有代中，针对每种不确定性量化方法，均以 0.25 fs 的时间步长，在 2500开尔文下运行了 10 条时长为 1 ns 的轨迹。所有模拟的初始结构均从初始训练数据集中 2500 开尔文平衡模拟采样得到的结构中随机选取。如果动能变为 0、空值或大于 10,000 千卡/摩尔（这是由于假极小值处预测势能失稳所致[27]），或者任意原子间的距离小于 1.0 $^{\circ}$A，则认为轨迹不稳定。由于模拟往往会在发生键拉伸之前就失败，因此未对原子间距设置上限。所有模拟均使用 Python 语言中的原子模拟环境（ASE）库进行[63]。

3 结果与讨论

3.1 修订版MD17数据集（rMD17）

首先，我们在 rMD17 数据集上进行了不确定性量化（UQ），该数据集可轻松从网上获取。我们评估了数据集中所有分子的预测不确定性是否能相对于真实误差正确排序（图 2a）。一个显而易见的观察结果是，基于集成的方法和均值-方差估计（MVE）方法的不确定性—误差分布呈正相关，这正是良好 UQ 方案所期望的。同样，高斯混合模型（GMM）的预测不确定性相对于真实误差呈现出较为线性但略有偏斜的分布。另一方面，证据回归呈现出双层分布，与能量没有良好的相关性。此外，值得注意的是，证据回归预测的不确定性跨越了八个数量级，这是由于训练过程中对 ν、α和 β 参数进行优化，使其趋近于 0。就各方法使用的计算资源而言（图 2b），包含 5 个独立神经网络（NN）的基于集成的方法，其训练时间和 GPU 内存使用量约为其他单个确定性神经网络方法的 5 倍。

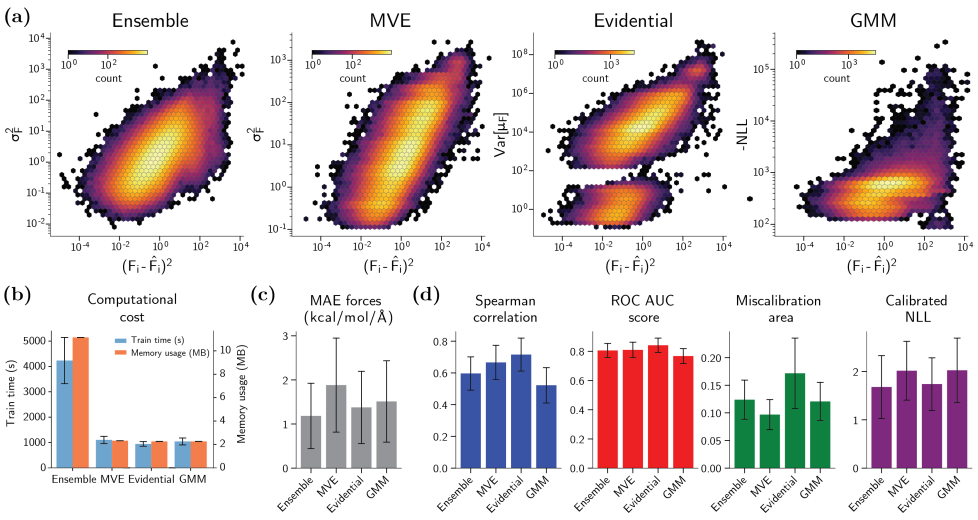

图 2 (a) 六边形分箱图展示了各不确定性量化方法在 rMD17 数据集的所有五折测试集中，所有分子原子力的（预测）不确定性与平方误差。请注意，由于 GMM 方法的不确定性（-NLL）包含负值，因此绘图时将所有不确定性均按最小值进行缩放，以得到正值，从而绘制对数刻度图。**(b)** 训练神经网络所用的总墙钟时间，以及神经网络产生的 GPU 内存使用量。**(c)** 条形图展示了使用各自对应损失函数训练的各不确定性量化模型在测试集上的预测误差。每个条形表示 rMD17 数据集所有五折测试集中所有分子各项指标的总体平均值。预测误差以原子力的平均绝对误差（MAE）表示，单位为千卡/摩尔/˚A。**(d)** 条形图展示了各不确定性量化方法的评估指标，包括斯皮尔曼秩相关系数、ROC AUC 分数、失校准面积和校准负对数似然（NLL）分数。对于**(c)** 和**(d)**，各分子指标的细分情况请参见图 S2。

由于 UQ 常用于神经网络原子间势（NNIP）以提高原子模拟的鲁棒性，因此首先评估各模型的预测准确度至关重要。从图 2c 可以看出，基于集成的方法在推理时的总体误差低于其他方法，而其他方法都只使用单个神经网络。鉴于所有 UQ 方法在相同数据划分中使用相同的模型架构、输入表示和训练数据，我们可以假设所有模型具有相同的模型不确定性（也称为模型偏差），即与模型形式不足和数据覆盖范围不足相关的认知不确定性部分。这表明模型准确度的差异源于参数优化不够理想，而这主要是由训练过程中使用的损失函数不同造成的。研究表明，集成学习能够消除参数优化歧义所导致的参数不确定性（也称为模型方差）[28, 34, 64]，从而提升泛化能力并降低总体误差[29, 65, 66]。另一方面，MVE 的平均测试集误差最高，这可能是由于其负对数似然（NLL）损失函数更难优化，文献 [67]已有相关报道。由于 NNIP 的鲁棒性很大程度上取决于模型的准确度，因此在选择要纳入其中的 UQ 方法时，总体误差是一个重要的考量标准。

主动学习工作流程。在不确定性估计的质量方面（图 2d），没有哪一种方法始终优于其他方法。相比之下，基于集成的方法在所有评估指标上的表现都处于良好范围内。有趣的是，尽管 MVE 的预测准确度较低，但其不确定性排序表现良好。相比之下，证据回归除失校准面积外，在所有评估指标上的表现均优于其他方法，这可能是由于图 2a 中的双峰分布所致。最后，GMM 在所有指标上的表现都最差，不过仍处于其他方法的误差条范围内。我们还发现，GMM 的表现与潜在特征向量的维度和高斯分布的数量无关（见图 S4）。

3.2 氨

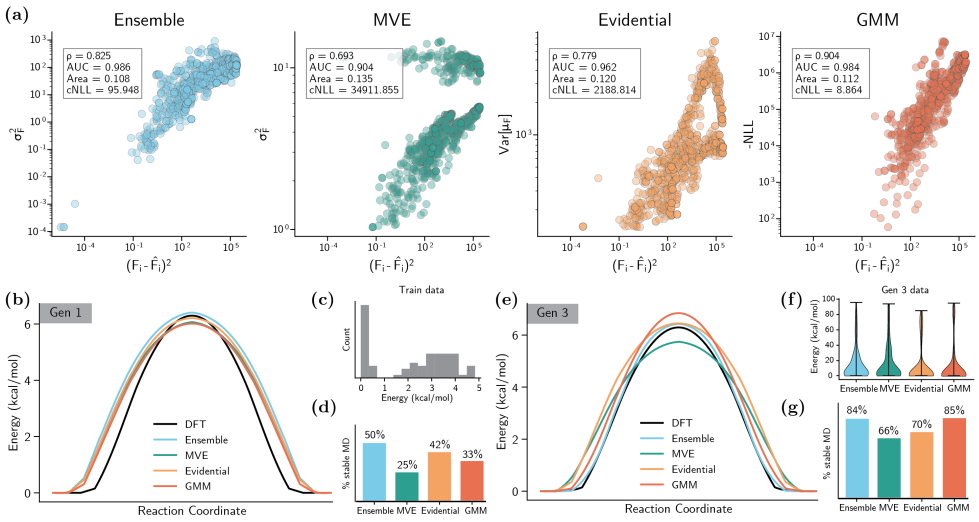

图 3 (a) 展示了仅使用初始训练数据集训练的各个所研究不确定性量化方法，在测试数据集中预测不确定性与原子力平方误差之间关系的图（见下方子图 **(c)** ）。每个子图右上角的方框显示了各方法的评估指标，即斯皮尔曼秩相关系数（ρ）、ROC AUC 分数（AUC）、失校准面积（Area）和校准负对数似然（cNLL）。**(b)** 使用密度泛函理论（DFT）和第 1 代中所研究的不确定性量化方法，通过弹性带方法（NEB）计算得到的氮反转能垒。各方法中的所有神经网络均使用相同的初始训练数据进行训练。**(c)** 展示初始训练数据中几何构型能量分布的直方图。**(d)** 使用第 1 代不确定性量化方法中的神经网络作为力场生成的稳定分子动力学轨迹所占的比例。**(e)** 使用密度泛函理论（DFT）和第 3 代不确定性量化方法，通过弹性带方法（NEB）计算得到的氮反转能垒。这些神经网络在初始训练数据的基础上，使用各自的不确定性量化方法生成的新对抗样本进行训练。**(f)** 经过两轮对抗采样后，训练集中能量的分布。**(g)**使用第 3 代不确定性量化方法中的神经网络作为力场生成的稳定分子动力学轨迹所占的比例。

在这个示例中，我们通过研究氨中的氮反转过程来测试不确定性量化方法。在第一代中，所有神经网络都使用相同的初始训练数据进行训练，该数据仅包含 78 个几何构型，能量范围为 0 至 5 千卡/摩尔（图 3c）。在能量介于 0 至 100 千卡/摩尔的 200 个构象上进行测试（每个 1 千卡/摩尔区间包含 2 个构象）时，可以清楚地看到，基于集成的方法和 GMM 模型所预测的不确定性与真实误差之间呈现良好的单调关系（图3a）。通过查看集成方法和 GMM 的评估指标也可以证实这一点：这两种方法在所有指标上的表现都优于另外两种方法。值得注意的是，GMM 在该数据集上的表现与 rMD17 数据集上的结果截然不同；在后者中，MVE 和证据回归的表现明显优于 GMM（图2d）。最可能的原因是训练集和测试集的分布不同。具体而言，rMD17 数据集中训练集和测试集中的构象能量范围有所重叠（图 S7）；而对于氨，训练集仅包含低能量构象（0–5 千卡/摩尔），测试集则包含低能量和高能量范围内的构象（0–100 千卡/摩尔）。这表明，基于集成的方法和 GMM 方法更擅长预测训练数据域之外的认识不确定性。证据回归在根据真实误差对不确定性进行排序方面也表现良好，但其分布再次呈现出奇特的三角形趋势。相比之下，MVE 在所有评估指标上的表现均不如其他模型，且产生了双层分布。将 MVE 在该数据集上的表现与其在 rMD17 数据集上的表现进行比较，可以推断，MVE 可能不擅长预测训练域之外数据的不确定性。

在预测氮反转能垒时，所有方法似乎都低估或高估了整体真实能量曲线（图 3b）。这一观察是合理的，因为所有模型在初始（第一代）训练数据中只见过低能量构象（图 3c）。当使用不同模型作为分子动力学模拟的原子间势时，神经网络集成产生的稳定分子动力学轨迹比例最高，为 50%；这些轨迹中没有原子解离或碰撞（图 3 d）。证据回归产生了 42% 的稳定轨迹，其次是高斯混合模型（33%），最后是均值方差估计（25%）。

经过两代对抗采样后，每种不确定性量化方法都向神经网络原子间势的训练中添加了 40 个数据点。需要注意的是，每种不确定性量化方法分别独立生成了新的数据点，因此各方法生成的数据点并不相同。图3f 展示了各方法的新训练数据的能量分布，其中包括含有 78 个几何构型的初始训练数据。仔细观察能量分布可以发现，尽管基于集成的方法能够将一些构象推向更高能量，但大多数新数据点仍集中在 0-30 千卡/摩尔的低能区域。均值方差估计方法也呈现类似情况，但其能量在 20-60 千卡/摩尔范围内的新构象数量多于基于集成的方法。相比之下，证据回归采样得到的能量在 40-85 千卡/摩尔范围内以及低于 20 千卡/摩尔的新几何构型更多。这可能是由于图 3a 所示的发散趋势：在对抗采样过程中最大化不确定性，会以分层的方式将几何构型推向低能和高能区域。

对于 GMM 方法采样得到的几何构型，也能观察到类似现象，但新的几何构型更多集中在 70-90 千卡/摩尔范围内。在尝试预测氨中的氮反转势垒时，只有基于集成的方法在两轮对抗采样后有所改进，并给出了准确的预测（图3e）。然而，经过两轮对抗采样后，基于集成的模型、均值方差估计和证据回归所生成的稳定分子动力学轨迹比例提高了约 30-40 个百分点（图 3g）。另一方面，高斯混合模型提高了 52 个百分点，这表明，与其他方法相比，不确定性估计的质量提升使神经网络原子间势在训练域之外的鲁棒性更强。

3.3 二氧化硅玻璃

最后，我们在二氧化硅玻璃上测试了不确定性量化方法；原子间势拟合在这一领域由来已久[51, 68–71]。这是因为二氧化硅玻璃由大型块体结构组成，对这些块体系统进行量子化学计算的计算成本非常高。神经网络日益广泛地用于玻璃态二氧化硅的原子间势拟合[72, 73], ，但遗憾的是，在训练数据域之外仍不可靠。因此，需要高效的采样方法来生成少量额外数据，以改善高误差区域。这些方法在很大程度上依赖不确定性量化方法的质量，而这些方法还应当轻量且快速。我们在相同的初始训练数据上训练了单个神经网络和一个包含 4 个网络的神经网络集成，并在测试数据集上评估了相应不确定性量化方法给出的误差和不确定性。初始训练集和测试集的分布如图 4a 所示。测试集包含能量和力都非常高的结构，这些结构未用于任何神经网络的训练或验证，因此可以确保评估得到的预测不确定性属于训练数据域之外的认识不确定性。图 4b 展示了不确定性量化方法的真实误差和不确定性评估结果。结果表明，所有方法都能较好地按照真实误差对不确定性进行排序，即使评估指标对哪个模型最佳给出了不一致的结果。可以说，考虑到评估指标、预测不确定性的数量级以及不确定性与误差的分布都不太理想，证据回归在这个数据集上的不确定性排序表现似乎最差。

随后，我们使用每种不确定性量化方法进行对抗采样，获得新的数据点，并使用密度泛函理论对其进行评估，然后将其纳入后续各代神经网络的再训练中。请注意，从第 2 代开始，每种不确定性量化方法所用神经网络的训练数据各不相同，因为新数据点是使用不同的不确定性量化方法采样得到的（图 4a）。在第 2 代中，各方法的对抗样本能量分布相似，但证据回归除外。有趣的是，证据回归并未将对抗样本引导至非常高的能量，而是引导至低于 0 千卡/摩尔（参考能量）的能量。参考能量选为初始训练数据中的最低能量；这些数据仅由从 300 K 或更高温度下的分子动力学模拟中提取的结构组成（参见图 S1）。而证据回归采样得到的结构能量低于训练数据中包含的能量。基于集成的方法，另一方面

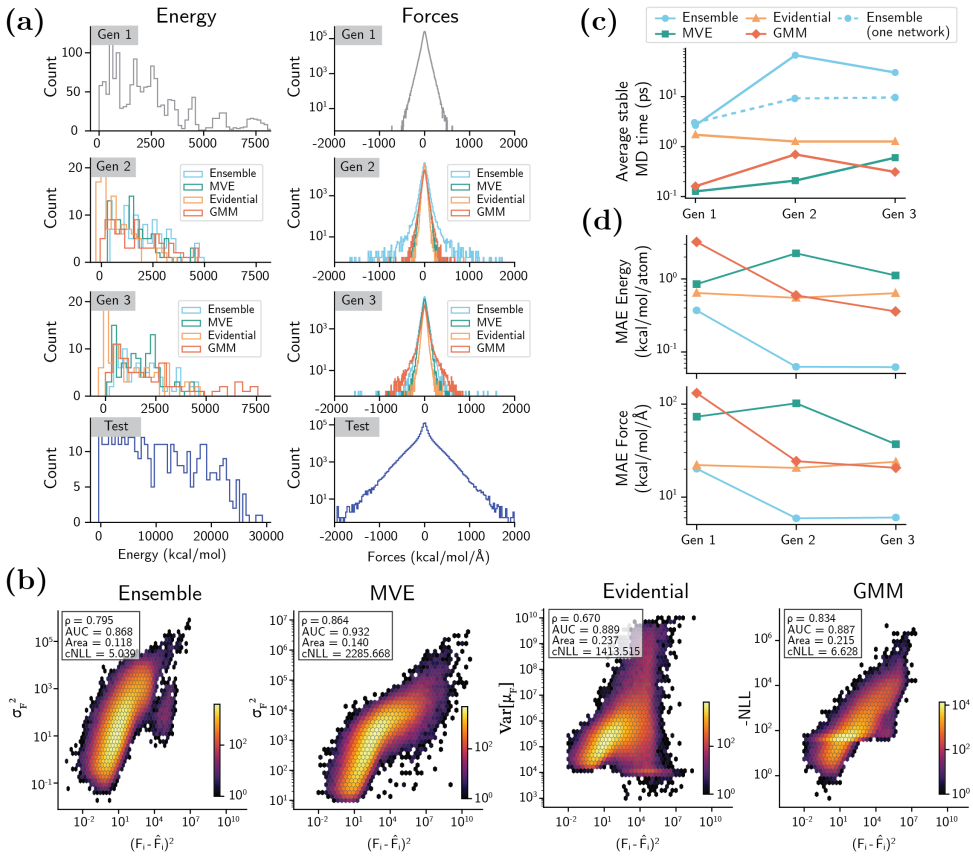

图 4 (a) 训练集和测试集中能量与力的分布。第一代（第一行）中，所有模型均使用相同的初始训练集进行训练。在后续各代中，神经网络模型使用各自对应的不确定性量化方法，同时基于初始训练数据和新采样的数据点进行训练。最后，在测试集上评估所有模型；与训练集相比，测试集中的能量和力的幅值明显更大。注意，测试集能量分布（最后一行第一列）的横轴尺度与其他能量分布不同。**(b)** 六边形分箱图，显示仅使用初始训练数据训练的神经网络所预测的不确定性与原子力平方误差之间的关系。**(c)** 分子动力学轨迹中未发生原子解离或碰撞的平均时长。该平均值基于 10 条轨迹计算。图例中的“集成（单个网络）”表示仅使用集成中的单个神经网络进行预测所得到的分子动力学轨迹。**(d)** 神经网络模型在测试集上预测能量和力的平均绝对误差。

手动采样能量范围在 0-5000 千卡/摩尔内、但原子力较高的结构。虽然这可以说是为对抗采样构建的约束（方程 15），但该约束的效果在基于集成的方法中最为明显，这些新增数据点显著提升了神经网络原子间势的鲁棒性。在第三代中，高斯混合模型采样得到的新几何构型，其能量和原子力均远高于其他方法。

每一代的神经网络训练完成后，使用这些神经网络进行 10 次分子动力学模拟。模拟保持稳定的时长记录于图 S15 中，平均值绘制在图 4c 中。图中显示，第 1 代的模拟很早就出现了非物理行为（训练仅使用了初始数据）。这可能是因为神经网络原子间势预测出了虚假的极小值，导致模拟过程中产生非物理结构[27]。此外，尽管训练数据和模型架构相同，稳定的分子动力学轨迹持续时间仍有差异。这与氨数据集上的结果一致：神经网络原子间势的鲁棒性直接受到预测准确性的影响，而预测准确性以能量和原子力的平均绝对误差表示（图 4d 和图 S14）。在第 2 代中，随着预测准确性的提高，基于集成的方法所获得的稳定分子动力学模拟时长大幅增加，这可能是因为纳入了新的、力值极高的对抗数据。高斯混合模型也呈现类似趋势，但改进幅度较小。然而，在第 3 代中，尽管纳入了更多数据，且图 4d 所示的测试集预测误差有所降低，基于集成的方法和 GMM 方法的改进似乎都出现了逆转。这表明，增加数据并不总能提升神经网络的鲁棒性，而新数据的质量更为关键[74]。均值方差估计表现出持续改进，但所有分子动力学模拟的时长都未能超过 1 ps。相比之下，随着通过对抗采样获得的结构不断增加，证据回归并未表现出改善（图 S16）。由于不同不确定性量度的梯度可能不同，我们将对抗采样中所有方法的玻尔兹曼温度和学习率设为相同值，同时增加迭代次数，以尽可能增大不确定性。对于该数据集可以看到，基于集成的方法不仅能产生稳定性更高的模拟，还能采样得到质量更高的不确定数据，而这些数据的质量无法通过评估指标量化。

3.4 一般性说明

总体而言，这些方法似乎在预测准确度、不确定性量化的可信度和计算成本之间存在权衡。对于基于集成的方法，其预测准确度和不确定性量化表现表明，这类方法比其他方法更可靠。然而，训练模型所需的时间和内存会随着集成中使用的神经网络数量增加而增长，对于超大数据集而言，可能会变得难以承受。另一方面，均值方差估计在根据真实误差对不确定性进行排序方面表现相对较好，且计算成本较低。不过，与其他方法相比，即使采用相同的训练参数和条件，这种方法所达到的预测准确度也较低。此外，训练过程中通过损失函数（Eq. S3）优化的预测不确定性，本质上只描述偶然不确定性。对于通过从头算方法生成的训练数据，数据集中的噪声相对较少，严格来说并不存在有意义的偶然不确定性，这意味着均值方差估计在某种程度上取代了认知不确定性。在证据回归的情况下，该方法在不确定性估计方面表现出色，但其预测准确度略低于

基于集成的方法。尽管如此，值得注意的是，证据回归方法在训练时往往会出现数值不稳定性，因为损失函数（Eq. S6）的优化会驱使参数（ν、 $\alpha-1$和 β）趋近于零。至于高斯混合模型，该模型易于实现、计算成本低，并且可以与任何神经网络配合使用，而不会降低已达到的预测准确度。不过，这种方法的缺点是需要为每个数据集手动配置高斯分布的数量（Figures S3 and S4），并且依赖高斯混合模型拟合的准确度。

4 结论

不同数据集和评估指标下，各种不确定性量化（UQ）方法的表现差异很大。本研究中，基于集成的不确定性量化方法在所有数据集和指标上都始终表现出色。此外，由于平均效应，集成预测通常具有较低的预测误差，但代价也更高。不过，其他研究表明，基于集成的方法只能用于处理源自参数不确定性（模型方差）的认知不确定性，而且在识别训练域之外的数据点方面可靠性较低[28]。与这些发现相反，我们的研究表明，在进行对抗采样时，基于集成的方法能够识别训练域之外的数据点。

另一方面，均值-方差估计（MVE）在不同数据集上的表现不一致。我们推断，MVE不太擅长系统地预测代表性不足区域的高不确定性，但更善于识别训练域内对应高误差的区域。研究表明，深度证据回归不确定性量化表现尚可，但不及基于集成的方法，而且在参数优化过程中往往数值不稳定。高斯混合模型（GMM）方法在不确定性量化方面表现良好，尤其是对于域外帧；它也更为通用，因为训练完成后可应用于任何模型架构，但在插值区间内表现不够理想，而MVE在该区间似乎更胜一筹。

然而，对于神经网络原子间势（NNIPs）的应用而言，泛化能力在确保原子模拟的鲁棒性方面发挥着非常重要的作用。一个看似能够很好地估计不确定性、预测准确度也不错，但无法从头开始或通过主动学习实现稳定生产模拟的模型，帮助有限。因此，尚待解决的挑战是找到一种能够准确排序不确定性、同时具备良好实用扩展性和较高预测准确度，并能产生稳定原子模拟的模型。神经网络原子间势中的不确定性量化和鲁棒性可能面临独特挑战，因为力是一个导数函数，在训练时可由能量轻易获得，但本文考察的所有方法都未能妥善处理力方差与能量方差之间的关系。

虽然成本低于基于集成的方法，但单个确定性神经网络的泛化能力较弱，因此表现未能超过基于集成的方法，尤其是在主动学习循环中，因为集成学习会降低

参数不确定性。对于均值方差估计（MVE）而言，这一点尤为明显，因为已有研究表明，联合优化均值和方差估计器会降低预测的稳定性[67]。当单一神经网络架构的泛化能力高于神经网络集成时，对单个神经网络采用深度证据回归或高斯混合模型等不确定性量化方法便有其合理性。

更广泛地说，不存在适用于所有情况的解决方案、更稳健的基于集成的方法成本高昂，以及其他方面颇具前景的新型证据方法表现令人失望，这些都表明，神经网络原子间势中的不确定性量化仍是科学人工智能方法开发面临的一项持续挑战。

5 致谢

本研究获得了 AGC公司的资助。

6 声明

- **利益冲突** 作者声明不存在利益冲突。
- **数据可用性** 本研究生成的分子动力学轨迹已存入 Materials Cloud Archive，登录代码为https://doi.org/10.24435/materialscloud:55-sd[52]。
- **代码可用性**本文用于生成结果的代码可在以下地址获取：
https://github.com/learningmatter-mit/UQ singleNN.git 采用 MIT许可证。
- **作者贡献** R.G.-B 和 A.R.T. 构思了该项目。A.R.T. 设计并开展了实验，进行了数据分析并撰写了手稿。S.U. 对初始二氧化硅训练数据集进行了分子动力学模拟和密度泛函理论计算。S.G. 和 J.C.B.D. 提供了有益的建议和讨论。J.C.B.D. 参与了内容组织和手稿审阅。R.G.-B. 负责监督研究、落实经费，并参与撰写手稿。

参考文献

[1]Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A.：《分子与材料科学中的机器学习》。Nature 559(7715)，547–555（2018）https://doi.org/10.1038/s41586-018-0337-2[2]Schwalbe-Koda, D., G´omez-Bombarelli, R.：《用于自动化化学设计的生成模型》。载于：《物理学讲义》第968卷，第445–467页。Springer，Cham（2020）。https://doi.org/10.1007/978-3-030-40245-7 21 . http://link.springer.com/10.1007/978-3-030-40245-7 21[3] Kocer, E., Ko, T.W., Behler, J.：《神经网络势：方法简明概述》。《物理化学年度评论》73(1)，163–186（2022）https: //doi.org/10.1146/annurev-physchem-082720-034254[4] Behler, J.：《高维神经网络势的四代发展》。《化学评论》121(16)，10037–10072（2021）https://doi.org/10.1021/acs. chemrev.0c00868[5]Yang, K., Swanson, K., Jin, W., Coley, C.W., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T.S., Jensen, K.F., Barzilay, R.：《分析用于性质预测的学习到的分子表示》。《化学信息与建模杂志》59(8)，3370–3388（2019）https://doi.org/10.1021/acs.jcim.9b00237[6] Mueller, T., Hernandez, A., Wang, C.：《原子间势模型的机器学习》。《化学物理学杂志》152(5)，050902（2020）https://doi.org/ 10.1063/1.5126336[7] Botu, V., Batra, R., Chapman, J., Ramprasad, R.：《机器学习力场：构建、验证与展望》。《物理化学C》121(1)，511–522（2017）https://doi.org/10.1021/acs.jpcc.6b10908[8] Behler, J., Parrinello, M.：《高维势能面的广义神经网络表示》。《物理评论快报》98(14)，146401（2007）https://doi.org/10.1103/PhysRevLett.98.146401[9] Onat, B., Cubuk, E.D., Malone, B.D., Kaxiras, E.：《植入式神经网络势：在锂硅合金中的应用》。《物理评论B》97(9)，1–9（2018）https: //doi.org/10.1103/PhysRevB.97.094106[10]Gokcan, H., Isayev, O.：《利用神经网络学习分子势》。《WIREs计算分子科学》12(2)，1–22（2022）https://doi.org/10.1002/wcms.1564[11] Sch¨utt, K.T., Sauceda, H.E., Kindermans, P.-J.J., Tkatchenko, A., M¨uller, K.- R.R.：《SchNet——一种面向分子与材料的深度学习架构》。《杂志》

[12]Sch¨utt, K.T., Unke, O.T., Gastegger, M.：用于预测张量性质和分子光谱的等变消息传递。arXiv 预印本 arXiv:2102.03150（2021年）[13] Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J.P., Kornbluth, M., Molinari, N., Smidt, T.E., Kozinsky, B.：用于数据高效且精准的原子间势的 E(3)等变图神经网络。《自然·通讯》13(1), 2453（2022年）https://doi.org/10.1038/s41467-022-29939-5[14] Zubatyuk, R., Smith, J.S., Nebgen, B.T., Tretiak, S., Isayev,O.：教会神经网络从分子中附着和脱离电子。《自然·通讯》12(1), 1–11（2021年）https://doi.org/10.1038/s41467-021-24904-0[15] Unke, O.T., Meuwly, M.：PhysNet：用于预测能量、力、偶极矩和部分电荷的神经网络。《化学理论与计算杂志》15(6), 3678–3693（2019年）https://doi.org/10.1021/acs.jctc.9b00181[16] Klicpera, J., Groß, J., G¨unnemann, S.：分子图的方向性消息传递。载于：ICLR，第 1–13 页（2020年）。http://arxiv.org/abs/2003.03123[17] Gastegger, M., Marquetand, P.：用于有机反应的高维神经网络势及改进的训练算法。《化学理论与计算杂志》11(5), 2187–2198（2015年）https://doi.org/10.1021/acs. jctc.5b00211[18] Ang, S.J., Wang, W., Schwalbe-Koda, D., Axelrod, S., G´omez-Bombarelli, R.：主动学习加速反应性势能面上的从头算分子动力学。《Chem》7(3), 1–32（2021年）https://doi.org/10.1016/j.chempr.2020.12.009[19] Wang, Z., Han, Y., Li, J., He, X.：结合碎片化方法与碎片的神经网络势能面，实现蛋白质能量的精确计算。《物理化学杂志》2020, 3035（2020年）https://doi.org/10.1021/ acs.jpcb.0c01370[20] Wang, W., Axelrod, S., G´omez-Bombarelli, R.：用于控制和学习的可微分子模拟。arXiv 预印本 arXiv: 2003.00868（2月）（2020年）[21] Marchand, D., Jain, A., Glensk, A., Curtin, W.A.：金属冶金中的机器学习 I：用于铝-铜的神经网络势。《物理评论：材料》4(10), 1–21（2020年）https://doi.org/10.1103/PhysRevMaterials.4.103601[22] Jakse, N., Sandberg, J., Granz, L.F., Saliou, A., Jarry, P., Devijver, E., Voigt- mann, T., Horbach, J., Meyer, A.：用于铝的机器学习原子间势：在凝固现象中的应用。《物理学杂志》：

[23] Natarajan, S.K., Behler, J.：固液界面的神经网络分子动力学模拟：低指数铜表面上的水。《物理化学化学物理》18(41), 28704–28725 (2016年) https://doi.org/10.1039/C6CP05711J[24] Morawietz, T., Singraber, A., Dellago, C., Behler, J.：范德华相互作用如何决定水的独特性质。《美国国家科学院院刊》113(30), 8368–8373 (2016年) https://doi.org/10.1073/pnas.1602375113[25]Ruza, J., Wang, W., Schwalbe-Koda, D., Axelrod, S., Harris, W.H., Gomez-Bombarelli, R., G´omez-Bombarelli, R.：利用双图卷积神经网络对离子液体进行温度可迁移粗粒化。《化学物理学杂志》153(16), 164501 (2020年) https://doi.org/10.1063/5.0022431[26] Fu, X., Wu, Z., Wang, W., Xie, T., Keten, S., Gomez-Bombarelli, R., Jaakkola, T.：仅有力还不够：通过分子模拟对机器学习力场进行基准测试与批判性评估 (2022年) https://doi.org/10.48550/arxiv.2210.07237[27] Morrow, J.D., Gardner, J.L.A., Deringer, V.L.：如何验证机器学习原子间势。arXiv预印本 arXiv: 2211.12484, 1–14 (2022年)[28] Heid,E., Mcgill, C.J., Vermeire,F.H., Green,W.H.：化学领域机器学习中的不确定性表征。《ChemRxiv》预印本。ChemrXiv:10.26434 (2023年) https://doi.org/10.26434/chemrxiv-2023-00vcg[29] Schwalbe-Koda, D., Tan, A.R., G´omez-Bombarelli, R.：利用基于不确定性的对抗攻击对分子几何构型进行可微采样。《自然通讯》12(1), 1–12 (2021年) https://doi.org/10.1038/s41467-021-25342-8[30] Shuaibi, M., Sivakumar, S., Chen, R.Q., Ulissi, Z.W.：利用简单的基于物理的先验分布，实现机器学习势的稳健离线主动学习。《机器学习：科学与技术》2(2), 025007 (2021年) https://doi.org/10.1088/2632-2153/abcc44[31]Vandermause, J., Torrisi, S.B., Batzner, S., Xie, Y., Sun, L., Kolpak, A.M., Kozinsky, B.：针对原子尺度稀有事件的可解释贝叶斯力场在线主动学习。《npj计算材料》6(1), 20 (2020年) https: // doi.org/10.1038/s41524-020-0283-z[32] Jinnouchi, R., Lahnsteiner, J., Karsai, F., Kresse, G., Bokdam, M.：通过基于贝叶斯推断的在线训练机器学习力场模拟杂化钙钛矿的相变。《物理评论快报》122(22), 225701 (2019年) https://doi.org/10.1103/PhysRevLett.122.225701

[33] Lakshminarayanan, B., Pritzel, A., Blundell, C.：使用深度集成进行简单且可扩展的预测不确定性估计。载于：Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R.（编）《神经信息处理系统进展》，第30卷，第6402–6413页。Curran Associates, Inc.（2017）。https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf[34] Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X.：深度神经网络中的不确定性综述。arXiv 预印本 arXiv: 2107.03342，第1–41页（2021年）[35] Nandy, A., Duan, C., Janet, J.P., Gugler, S., Kulik, H.J.：过渡金属化学中机器学习加速发现的策略与软件。《工业与工程化学研究》57(42)，13973–13986（2018）https: //doi.org/10.1021/acs.iecr.8b04015[36] Janet, J.P., Duan, C., Yang, T., Nandy, A., Kulik, H.J.：一种定量不确定性度量控制着神经网络驱动的化学发现中的误差。《化学科学》10(34)，7913–7922（2019）https://doi.org/10.1039/c9sc02298h[37] Oberdiek, P., Rottmann, M., Gottschalk, H.：基于梯度信息的深度神经网络分类不确定性。《计算机科学讲义》（包括《人工智能讲义》和《生物信息学讲义》子系列）11081 LNAI，113–125（2018）https://doi.org/10.1007/ 978-3-319-99978-4 9[38] Lee, J., Alregib, G.：梯度作为神经网络不确定性的度量。《国际图像处理会议论文集》，ICIP 2020-Octob，2416–2420（2020）https://doi.org/10.1109/ICIP40778.2020.9190679[39] Zhu, A., Batzner, S., Musaelian, A., Kozinsky, B.：深度学习原子间势中的快速不确定性估计。arXiv 预印本 arXiv: 2211.09866，第1–38页（2022）[40]Nix, D.A., Weigend, A.S.：估计目标概率分布的均值和方差。《IEEE神经网络国际会议——会议论文集》第1卷，55–60（1994）https://doi.org/10.1109/icnn.1994. 374138[41] Amini, A., Schwarting, W., Soleimany, A., Rus, D.：深度证据回归。《神经信息处理系统进展》2020-Decem（NeurIPS），第1–20页（2020）[42] Soleimany, A.P., Amini, A., Goldman, S., Rus, D., Bhatia, S.N., Coley, C.W.：用于引导分子性质预测与发现的证据深度学习。《ACS中央科学》7(8)，1356–1367（2021）https://doi.org/10.1021/acscentsci.

[43] Amersfoort, J., Smith, L., Teh, Y.W., Gal, Y.：使用单个深度确定性神经网络进行不确定性估计。第37届国际机器学习大会（ICML 2020），第F16814部分，9632–9642（2020）[44] Behler, J.：展望：用于原子模拟的机器学习势。《化学物理学杂志》145(17)，170901（2016）https://doi.org/10.1063/1. 4966192[45] Kulichenko, M., Smith, J.S., Nebgen, B., Li, Y.W., Fedik, N., Boldyrev, A.I., Lubbers, N., Barros, K., Tretiak, S.：用于材料与化学动力学的神经网络兴起。《物理化学快报》12(26)，6227–6243（2021）https://doi.org/10.1021/acs.jpclett.1c01357[46] Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R., Coley, C.W.：利用神经网络进行分子性质预测的不确定性量化。arXiv预印本 arXiv:2005.10036 60(8)，3770–3780（2020）https://doi.org/10.1021/acs. jcim.0c00502[47] Gurevich, P., Stuke, H.：梯度共轭先验与多层神经网络。《人工智能》278（2020）https://doi.org/10.1016/j.artint.2019. 103184[48]Reynolds, D.：高斯混合模型，载于《生物识别百科全书》，第659–663页。Springer，波士顿，马萨诸塞州（2009）。https://doi.org/10.1007/ 978-0-387-73003-5{ } 。https://doi.org/10.1007/978-0-387-73003-5 196[49] Christensen, A.S., Lilienfeld, O.：梯度在分子能量与力机器学习中的作用。《机器学习：科学与技术》1(4)（2020）https://doi.org/10.1088/2632-2153/abba6f[50] Chmiela, S., Tkatchenko, A., Sauceda, H.E., Poltavsky, I., Sch¨utt, K.T., M¨uller, K.-R.：准确的能量守恒分子力场的机器学习。《科学进展》3(5)，1603015（2017）https://doi.org/10.1126/sciadv.1603015[51] Urata, S., Nakamura, N., Aiba, K., Tada, T., Hosono, H.：氟如何使二氧化硅玻璃的密度涨落最小化：采用机器学习辅助力匹配势的分子动力学研究。《材料与设计》197，109210（2021）https://doi.org/10.1016/j.matdes.2020.109210[52] Tan, A.R., Urata, S., Goldman, S., G´omez-Bombarelli, R.：神经网络势中的单模型不确定性量化并不总是优于模型集成。Materials Cloud Archive（2023）https://doi.org/10.24435/materialscloud:55-sd[53] Kresse, G., Furthm¨uller, J.：采用平面波基组进行从头算总能量计算的效率，适用于

金属和半导体。《计算材料科学》6(1)，15–50（1996）https://doi.org/10.1016/0927-0256(96)00008-0[54] Kresse, G., Furthm¨uller, J.：采用平面波基组进行从头算总能量计算的高效迭代方案。《物理评论B》54(16)，11169–11186（1996）https://doi.org/10.1103/PhysRevB.54.11169[55]Kresse, G., Joubert, D.：从超软赝势到投影缀加波方法。《物理评论B》59(3)，1758–1775（1999）https: //doi.org/10.1103/PhysRevB.59.1758[56] Perdew, J.P., Burke, K., Ernzerhof, M.：让广义梯度近似变得简单。《物理评论快报》77(18)，3865–3868（1996）https://doi.org/ 10.1103/PhysRevLett.77.3865[57] Plimpton, S.：短程分子动力学的快速并行算法。《计算物理学杂志》117(1)，1–19（1995）https://doi.org/10.1006/JCPH. 1995.1039[58] Evans, D.J.：库埃特流非线性热力学的计算机“实验”。《化学物理学杂志》78(6)，3297–3302（1983）https://doi.org/10.1063/1.445195[59]Varivoda, D., Dong, R., Omee, S.S., Hu, J.：带不确定性量化的材料性质预测：基准研究。arXiv预印本 arXiv: 2211.02235（Ml）（2022）[60] Kahle, L., Zipoli, F.：神经网络势集成的不确定性估计质量。《物理评论E》105(1)，1–11（2022）https://doi.org/10.1103/ PhysRevE.105.015311[61] Tran, K., Neiswanger, W., Yoon, J., Zhang, Q., Xing, E., Ulissi, Z.W.：比较材料性质预测不确定性量化方法的方法。《机器学习：科学与技术》1(2)，025006（2020）https://doi.org/10.1088/ 2632-2153/ab7e1a[62] Hu, Y., Musielewicz, J., Ulissi, Z.W., Medford, A.J.：利用保形预测对机器学习原子间势进行稳健且可扩展的不确定性估计。《机器学习：科学与技术》3(4)，045028（2022）https: //doi.org/10.1088/2632-2153/aca7b1[63] Hjorth Larsen,A.,JørgenMortensen, J., Blomqvist, J., Castelli, I.E., Christensen, R., Du-lak, M., Friis, J., Groves, M.N., Hammer, B., Hargus, C., Hermes, E.D., Jennings, P.C., Bjerre Jensen, P., Kermode, J., Kitchin, J.R., Leonhard Kolsbjerg, E., Kubal, J., Kaasbjerg, K., Lysgaard, S., Bergmann Maronsson, J., Maxson, T., Olsen, T., Pastewka, L., Peterson, A., Rostgaard, C., Schiøtz, J., Sch¨utt, O., Strange, M., Thygesen, K.S., Vegge, T., Vilhelmsen, L., Walter, M., Zeng,

Z., Jacobsen, K.W.：《原子模拟环境：用于处理原子的 Python 库》。《凝聚态物质物理学》29(27)，273002（2017）https://doi.org/10.1088/1361-648X/aa680e[64] Gabriel, J.J., Paulson, N.H., Duong, T.C., Tavazza, F., Becker, C.A., Chaudhuri, S., Stan, M.：《原子尺度金属建模中的不确定性量化及其对介观和连续体建模的影响：综述》。《Jom》73(1)，149–163（2021）https://doi.org/10.1007/s11837-020-04436-6[65]Tumer, K., Ghosh, J.：《集成分类器中的误差相关性与误差降低》。《连接科学》8(3-4)，385–404（1996）https://doi.org/10.1080/ 095400996116839[66] Mendes-Moreira, J.,Soares, C., Jorge, A.M., De Sousa, J.F.：《回归的集成方法：综述》。《ACM 计算调查》45(1)（2012）https://doi.org/ 10.1145/2379776.2379786[67] Seitzer, M., Tavakoli, A., Antic, D., Martius, G.：《概率神经网络中异方差不确定性估计的陷阱》，1–24（2022）[68] Beest, B.W.H.v., Kramer, G.J.J., Van Beest, B. W.H.H., Kramer, G.J.J., Van San- ten, R.A.：《基于从头算计算的二氧化硅和铝磷酸盐力场》。《物理评论快报》64(16)，1955–1958（1990）https://doi.org/ 10.1103/PhysRevLett.64.1955[69] Tangney, P., Scandolo, S.：《用于二氧化硅的从头算参数化原子间力场》。《化学物理学杂志》117(19)，8898–8904（2002）https://doi.org/ 10.1063/1.1513312[70] Tsuneyuki, S., Tsukada, M., Aoki, H., Matsui, Y.：《应用于分子动力学的二氧化硅第一性原理原子间势》。《物理评论快报》61(7)，869–872（1988）https://doi.org/10.1103/PhysRevLett.61.869[71] Pedone, A., Bertani, M., Brugnoli, L., Pallini, A.：《氧化物玻璃的原子间势：过去、现在与未来》。《非晶固体杂志：X》15(August)，100115（2022）https://doi.org/10.1016/j.nocx.2022.100115[72] Urata, S., Nakamura, N., Tada, T., Hosono, H.：《减少二氧化硅玻璃瑞利散射的 Al2O3 与氟共掺杂效应的分子动力学研究》。《美国陶瓷学会杂志》104(10)，5001–5015（2021）https: // doi.org/10.1111/jace.17774[73] Balyakin,I.A.,Rempel,S.V., Ryltsev, R.E., Rempel, A.A.：《液态二氧化硅的深度机器学习原子间势》。《物理评论E》102(5)，52125（2020）https://doi.org/10.1103/PhysRevE.102.052125

[74] Koh, P.W., Liang, P.：通过影响函数理解黑箱预测。第34届国际机器学习大会（ICML 2017）4，2976–2987（2017）

《单模型不确定性量化在神经网络势中的表现并不总是优于模型集成》的补充信息

Aik Rui Tan，[1] Shingo Urata，[2] Samuel Goldman，[3] Johannes C. B. Dietschreit，[1] 以及 Rafael G´omez-Bombarelli[1,a)]1) 麻省理工学院材料科学与工程系2) *AGC*株式会社创新技术实验室3) 麻省理工学院计算与系统生物学（日期：2023年5月2日）

a)电子邮件：rafagb@mit.edu

I. 神经网络原子间势（NNIPS）

我们使用可极化原子相互作用神经网络（PaiNN）——一种等变消息传递神经网络——来学习分子和材料的原子间势。所有系统采用相同的神经网络原子间势架构：使用3个卷积层，每层的隐藏维度均为128、20个可学习高斯函数、5.0 $^{\circ}$A的截断距离以及平移软加激活函数。随后，将128维特征向量传入两个大小为64的全连接层，并将其求和以得到最终输出值。对于特定系统，所有模型均使用Adam优化器，在不同的数据集划分上进行训练，初始学习率为 10^{-4}。如果验证集上的误差连续30个训练轮次未改善，则使用调度器将学习率降低为原来的二分之一。

对于 rMD17 数据集，每一折/划分中，所有模型均使用 800 个数据点（分子构象）进行训练，并使用 200 个数据点进行验证，训练 500 轮，批量大小为 64。随后，使用 1000 个数据点对模型进行测试。对每种分子和不确定性量化方法均进行了五折交叉验证，各折训练集和测试集的索引由 Christensen等人提供。对于氨数据集，第一代模型使用 62 个几何构型进行训练，并使用 16 个几何构型进行验证，以评估模型在数据点较少时进行自助采样的能力。测试集包含 200 个几何构型，其能量范围为 0 至 100 千卡/摩尔。模型训练了 500 轮，批量大小为 64。对于二氧化硅数据集，模型使用包含 1590 个几何构型的数据集进行训练（训练集与验证集的比例为 80:20），并使用外部测试集中的 402 个结构进行测试；这些结构的能量范围为 0 至 30,000 千卡/摩尔（更多详情见数据集部分）。模型训练了 1000 轮，批量大小为 3。

A. 基于集成的方法

rMD17 和氨数据集使用 5 个独立的神经网络，而二氧化硅玻璃数据集使用 3 个独立的神经网络。在集成模型的神经网络训练过程中，损失 $\mathcal{L}$ 随后计算为一个批量大小为 N 的批次中预测值与目标值的平均绝对误差

$$\mathcal{L} = \frac{1}{M}\frac{1}{N}\sum_{m}^{M}\left[\sum_{i}^{N}\left(w_E|E_{i,m} - \hat{E}_{i,m}| + w_F|\mathbf{F}_{i,m} - \hat{\mathbf{F}}_{i,m}|\right)\right] \tag{1}$$

其中， w_E 和 w_F 分别是训练期间能量与力之间权衡的权重。在所有模型中，这些权重分别设为 0.1 和 1.0。

B. 均值方差估计（MVE）

MVE 方法中的神经网络采用正态分布的负对数似然（NLL）损失函数进行训练，而不是常规的平均绝对误差（MAE）。在我们的工作中，NLL 损失函数仅用于力的预测，而能量输出则采用常规的平均绝对误差（MAE）。这提高了神经网络的预测准确度。与集成方法类似，不确定性估计仅通过原子力的方差来描述。

$$\mathrm{MAE}_E = \frac{1}{N}\sum_i^N |E_i - \mu_{E_i}| \tag{2}$$

$$\mathrm{NLL}_F = \frac{1}{2N}\sum_i^N \log(2\pi\sigma_{F_i}^2) + \frac{(\mathbf{F}_i - \mu_{F_i})^2}{2\sigma_{F_i}^2} \tag{3}$$

$$\mathcal{L} = w_E\mathrm{MAE}_E + w_F\mathrm{NLL}_F \tag{4}$$

其中， w_E 和 w_F 分别设为 0.1 和 1.0。

C. 深度证据回归

深度证据神经网络的训练所使用的损失函数计算如下

$$\mathrm{MAE}_E = \frac{1}{N}\sum_i^N |E_i - \gamma_{E_i}| \tag{5}$$

$$\mathrm{NLL}_F = \frac{1}{N}\sum_i^N \frac{1}{2}\log\left(\frac{\pi}{\nu}\right) - \alpha\log\left[2\beta(1+\nu)\right] \tag{6}$$

$$+\left(\alpha+\frac{1}{2}\right)\log\Big\{(\mathbf{F}_i - \gamma_{F_i})^2\nu + \left[2\beta(1+\nu)\right]\Big\} + \log\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})}\right) \tag{7}$$

$$+\lambda|\mathbf{F}_i - \gamma_{F_i}|\cdot(2\nu+\alpha) \tag{8}$$

$$\mathcal{L} = w_E\mathrm{MAE}_E + w_F\mathrm{NLL}_F \tag{9}$$

其中， λ设为 0.2，用于控制在错误预测时减少证据（即降低 ν 和 α ）与最大化模型拟合度之间的权衡。同样， w_E 和w_F 分别设为 0.1 和 1.0。

D. 高斯混合模型（GMM）

在 GMM 方法中，使用常规的平均绝对误差（MAE）函数训练神经网络（NN），如方程 1 所示，其中 $M = 1$、 $w_E = 0.1$ 和 $w_F = 1.0$

II. 数据集

A. 二氧化硅

二氧化硅玻璃（SiO_2）训练数据集由 1590 个结构组成，这些结构从多条较长的分子动力学轨迹中采样而来；轨迹在不同条件下进行，包括温度平衡、冷却、单轴拉伸、单轴压缩、双轴拉伸、双轴压缩以及不同速率下的双轴剪切，所用力匹配势见 Urata 等人所述[1]。每个非晶二氧化硅结构包含 699 个原子（233 个 Si 原子和 466 个 O 原子）。在冷却条件的模拟中，5 个二氧化硅结构以 1 K/ps 的速率从 2300 K 冷却至 300 K，每隔 25 ps 保存一次帧。温度平衡模拟中，非晶二氧化硅结构分别在 300 K、500 K、1000 K、1500 K、2000开尔文、2500开尔文、3000开尔文和 3500开尔文下平衡 11 ns，并从模拟的最后 10 ns 中每隔 100 ps 保存一次帧。请注意，冷却模拟与温度平衡模拟彼此独立。在单轴变形条件下，我们进行了两种不同的模拟，分别为单轴拉伸和单轴压缩。具体而言，二氧化硅体系受到恒定应变率 $\pm 10^{-4}$ ps^{-1} （$+$ 表示拉伸， $-$ 表示压缩），在一个方向上达到 ± 0.30 的最终应变幅值，并每隔 100 ps 采样一次帧，对应的应变变化为 ± 0.1。在双轴拉伸和压缩情况下（分别进行模拟），体系在 x 方向上受到恒定应变率 ± 0.0001 ps^{-1} ，直至最终应变幅值达到 ± 0.30。在 y 方向上，应变率为 $\pm 2.5 \times 10^{-5}$ ps^{-1},

[1]Urata, S., Nakamura, N., Aiba, K., Tada, T., Hosono, H. **材料设计**, 197,109210. https://doi.org/10.1016/j.matdes.2020.109210 (2021年)。

$\pm 5.0 \times 10^{-5}$ ps^{-1}、 $\pm 7.5 \times 10^{-5}$ ps^{-1} 和 $\pm 10^{-4}$ ps^{-1}，最终应变幅值分别达到 ± 0.075、± 0.015、± 0.0225 和 ± 0.3。在双轴剪切模拟中，体系在 x 方向上以 $\pm 10^{-4}$ ps^{-1} 的应变率施加拉伸应力，直至应变达到最大值 -0.3。在 y 方向上，以 2.5×10^{-5} ps^{-1}、5.0×10^{-5} ps^{-1}、7.5×10^{-5} ps^{-1} 和 10^{-4} ps^{-1} 的应变率施加压缩应力，直至应变分别达到最大值 0.075、0.015、0.0225 和 0.3（模拟的汇总详情见表 S1，数据分布见图 S1）。所有帧均每隔 100 ps 采样一次。所有分子动力学模拟均使用 LAMMPS 软件包，并通过 Nos´e-Hoover 恒温器和恒压器控制温度和压力。此外，本研究在试运行期间通过对抗采样创建了一个独立测试集。选取了 402 个能量介于 0 至 30,000 千卡/摩尔（43 千卡/摩尔/原子）之间的结构组成测试集。根据能量对结构分箱并去重后，从每个分箱中选出原子力最大的 2 个结构。训练集和测试集中没有重复结构。

总计从分子动力学轨迹中采样了 1590 个二氧化硅结构，以构成训练集；另从对抗采样中选取了 402 个结构，以构成测试集。使用投影缀加平面波（PAW）方法，通过 Vienna Ab-initio Simulation Package（VASP）对这些结构进行了密度泛函理论计算。采用广义梯度近似（GGA）下的 Perdew-Burke-Ernzerhof（PBE）泛函作为交换-相关泛函。平面波的动能截断值设为 600 eV。自洽场（SCF）循环中的能量收敛阈值设为 10^{-5} eV。

模拟 Type	Name	温度 (K)	加热 / 冷却-速率 (K/ps)	应变率 in x di-方向 (10^{-4}/皮秒)	应变率 in y di-方向 (10^{-4}/皮秒)	Num 帧
温度300 K		300	-	-	-	100
平衡化 500 K		500	-	-	-	100
	1000 K	1000	-	-	-	100
	1500 K	1500	-	-	-	100
	2000开尔文	2000	-	-	-	100
	2500开尔文	2500	-	-	-	99
	3000开尔文	3000	-	-	-	98
	3500开尔文	3500	-	-	-	92
冷却	冷却	-	1.00	-	-	410
单轴	单轴压缩 300		-	-1.00	-	30
变形 单轴拉伸		300	-	1.00	-	30
双轴	双轴 m1-m0.25 300		-	-1.00	-0.25	30
双轴压缩 m1-m0.5		300	-	-1.00	-0.50	30
	双轴 m1-m0.75 300		-	-1.00	-0.75	30
	双轴 m1-m1	300	-	-1.00	-1.00	30
双轴	双轴 1-0.25	300	-	1.00	0.25	23
拉伸	双轴 1-0.5	300	-	1.00	0.50	26
	双轴 1-0.75	300	-	1.00	0.75	23
	双轴 1-1	300	-	1.00	1.00	22
双轴	双轴 1-m0.25	300	-	1.00	-0.25	28
剪切	双轴 1-m0.5	300	-	1.00	-0.50	29
	双轴 1-m0.75	300	-	1.00	-0.75	30
	双轴 1-m1	300	-	1.00	-1.00	30
						1590

表 S1. 二氧化硅初始训练数据集细分表。

III. 对抗采样

在对抗采样策略中，对从初始训练数据中随机选取的结构施加扰动 δ 。随后使用梯度上升技术迭代更新该扰动，以最大化扰动结构的不确定性 $U(x_\delta)$，其定义见主论文中的公式 15。具体而言，在归一化温度 kT 为 0.7 千卡/摩尔/原子时，对扰动更新进行了 80 次迭代优化。氨和二氧化硅所使用的学习率分别设为 5×10^{-5} 和 3×10^{-5} 。

氨在每一代（即主动学习循环）中采样 20 个对抗构象，而二氧化硅采样 100 个对抗构象。最后，我们将氨的新对抗样本截断距离设为 100 千卡/摩尔，以避免几何构型发生极端畸变。对于二氧化硅，第 2 代的截断距离设为 5000 千卡/摩尔（7.2 千卡/摩尔/原子），并在第 3 代提高至 10000 千卡/摩尔（14.3 千卡/摩尔/原子）。

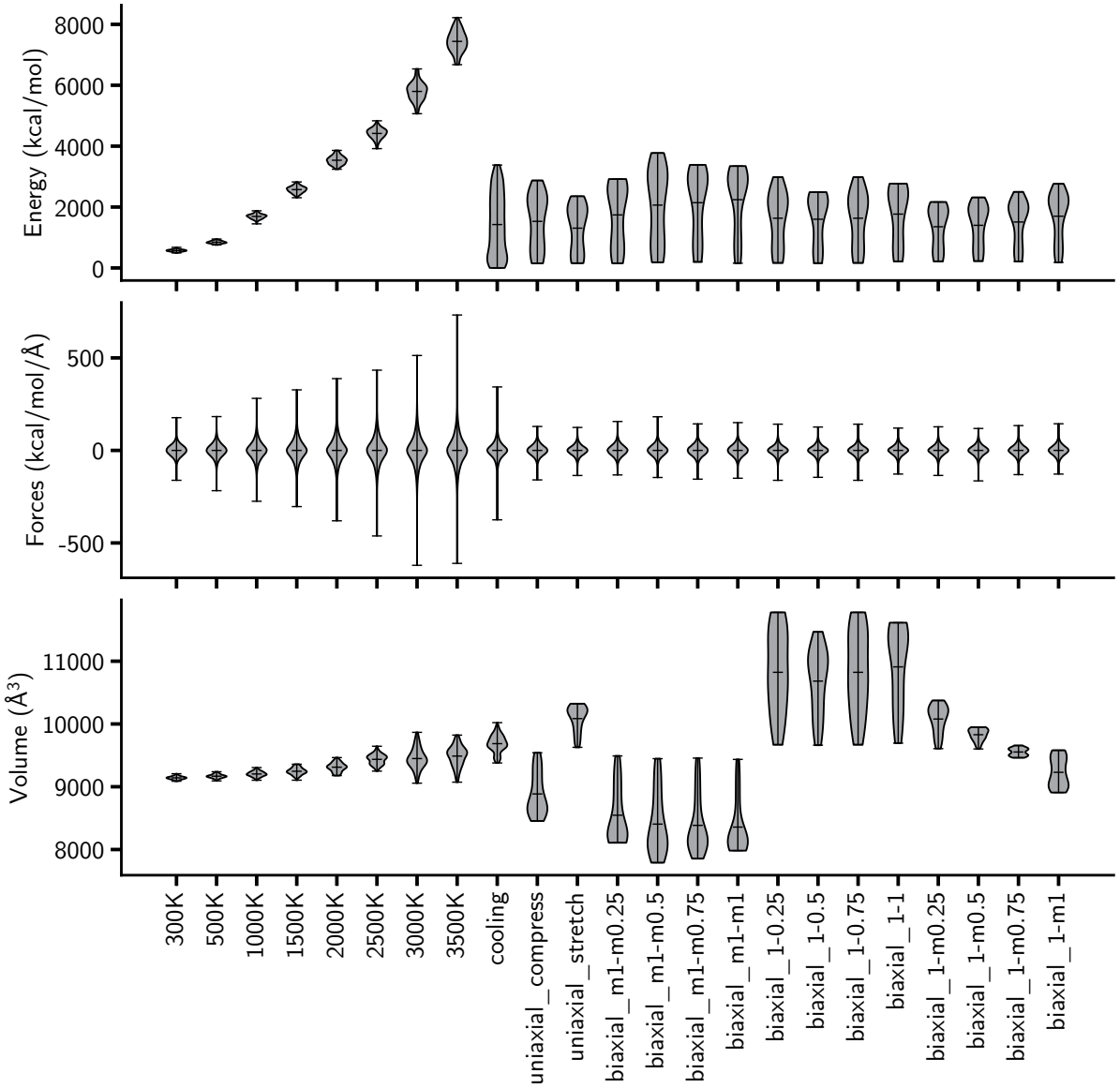

图 S1. 小提琴图展示了初始二氧化硅数据集中能量、力和晶胞体积分布。该数据集包含在各种条件下进行的分子动力学（MD）轨迹中提取帧的密度泛函理论（DFT）计算结果（见表 S1）。所有二氧化硅系统均包含 699 个原子。在双轴拉伸情形中，下划线（“ ”）后的第一个和第二个数字分别表示 x 方向和 y 方向上大小为 $1 \times 10^{-4}\mathrm{ps}^{-1}$ 的应变率。“m”表示负值，即施加了压缩应变。

IV. 附加结果

A. rMD17 数据集

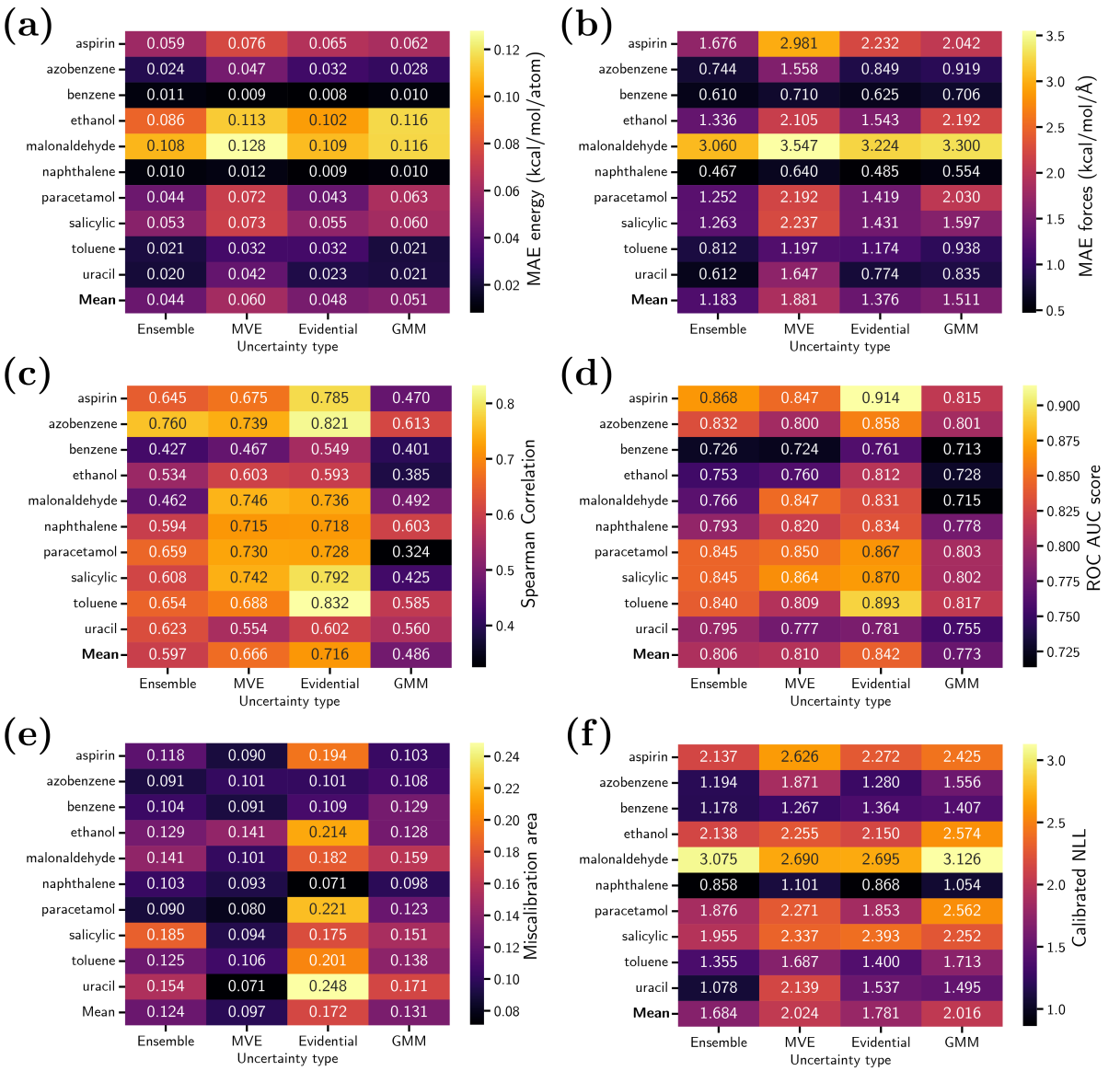

图 S2. 各种不确定性估计方法在 rMD17 数据集测试集上的预测准确率与不确定性估计评估指标热图。每个单元格表示各分子所有划分的平均值。**(a)** 能量的平均绝对误差，单位为千卡/摩尔/原子。**(b)** 原子力的平均绝对误差，单位为千卡/摩尔/Å。**(c)** 斯皮尔曼秩相关系数。**(d)** 受试者工作特征曲线下面积（ROC AUC 分数）。**(e)** 失校准面积。**(f)** 校准负对数似然得分。

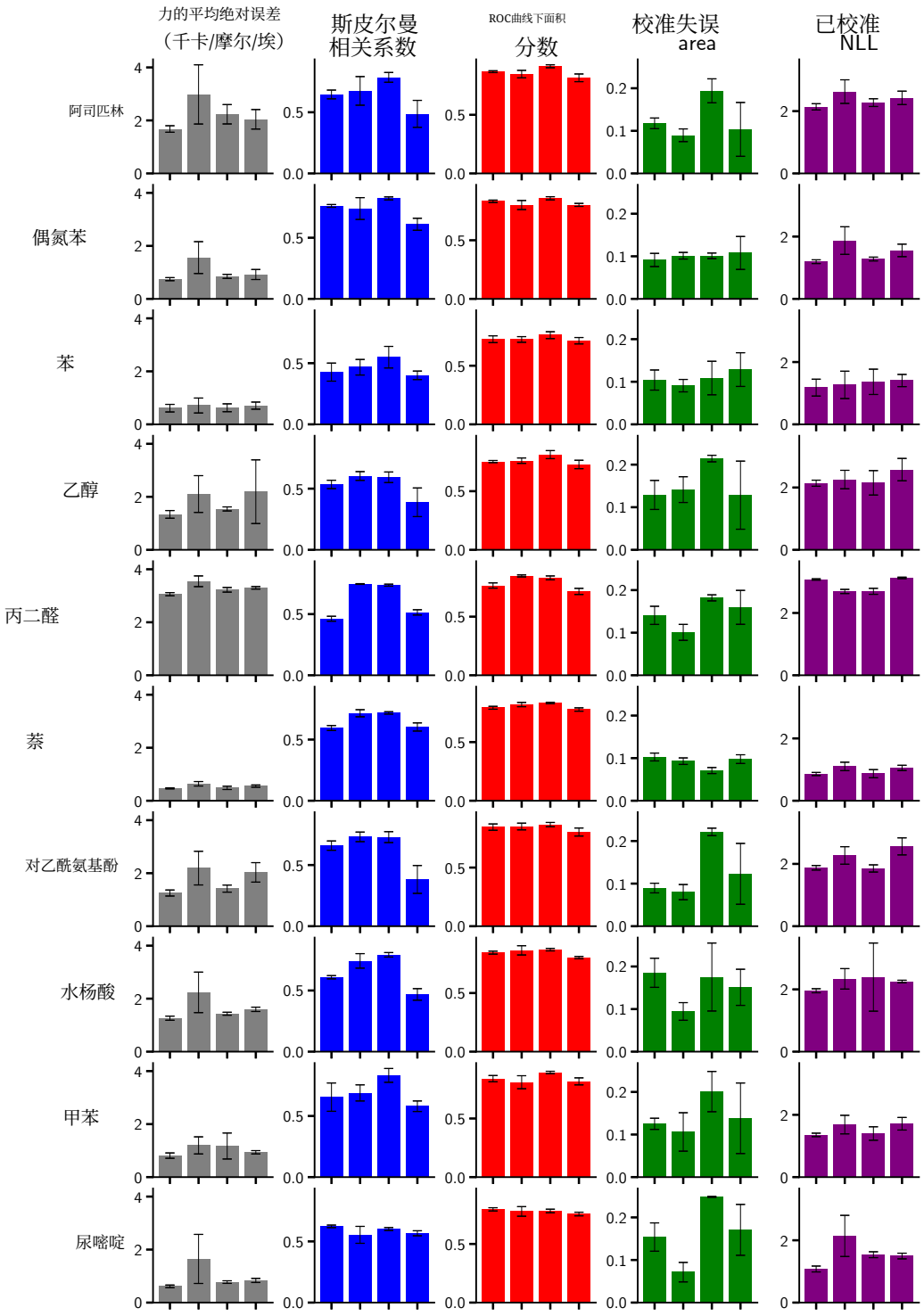

图 S3. 条形图便于比较所考察的各类不确定性估计方法在 rMD17 数据集中所有分子上的预测准确度（原子力的平均绝对误差）和不确定性估计评估指标。

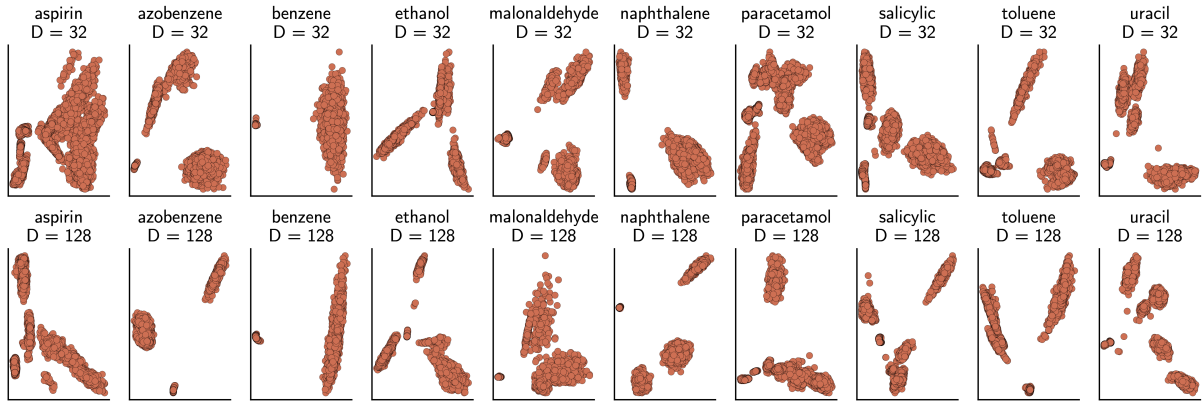

图 S4. 对 rMD17 数据集中每种分子所用的神经网络（NN）的潜在嵌入进行主成分分析（PCA），并将其用于高斯混合模型（GMM）（针对一次划分的训练集）。此图展示了如何合理设置拟合 GMM 的簇/组件数量。

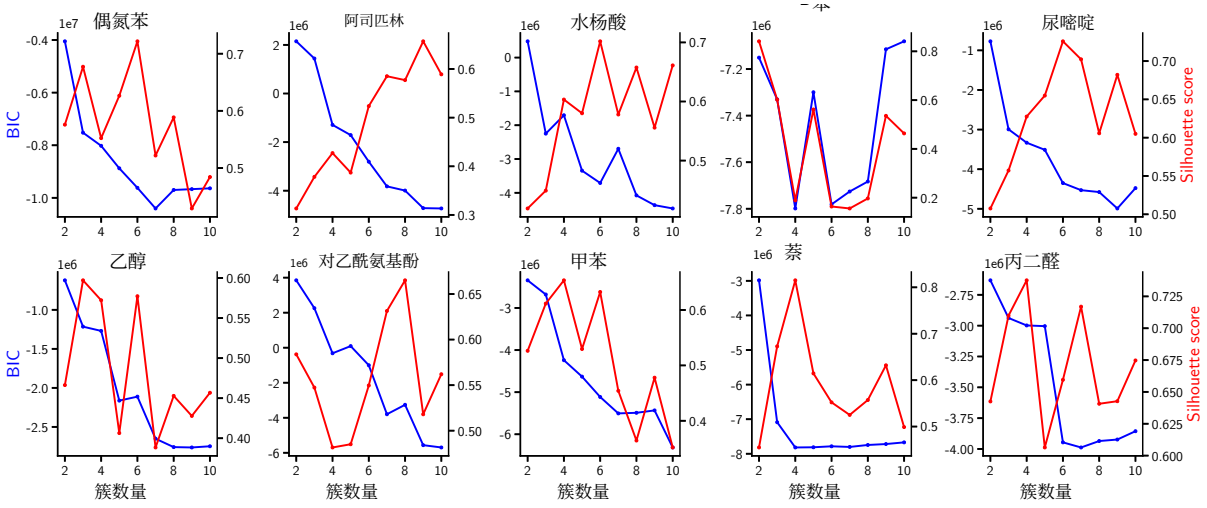

图 S5. 展示贝叶斯信息准则（BIC）和轮廓系数随用于拟合神经网络（NN）潜在嵌入（用于 GMM）的簇/组件数量变化的曲线。随后可通过考察 BIC 与轮廓系数之间的权衡，确定用于最佳拟合 GMM 的簇数量。理想情况下，拟合良好的模型应具有较低的 BIC 和较高（接近 1）的轮廓系数。

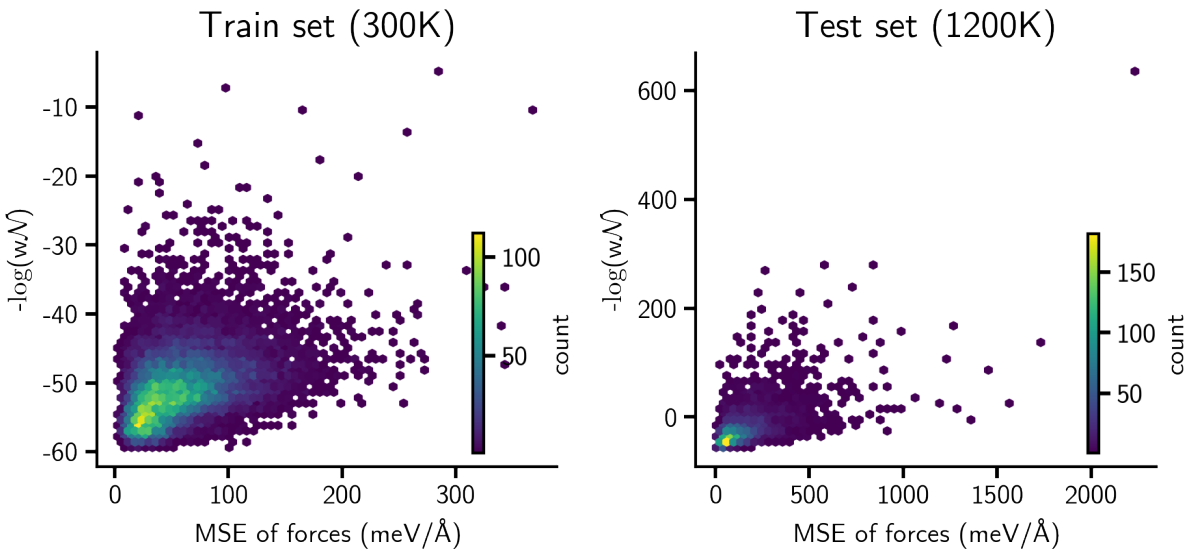

图 S6. 复现原论文中提出的用于不确定性估计方案的 3BPA 分子结果：高斯混合模型[1]

[2] A. Zhu、S. Batzner、A. Musaelian 和 B. Kozinsky，《深度学习原子间势中的快速不确定性估计》，arXiv:2211.09866，1（2022）。

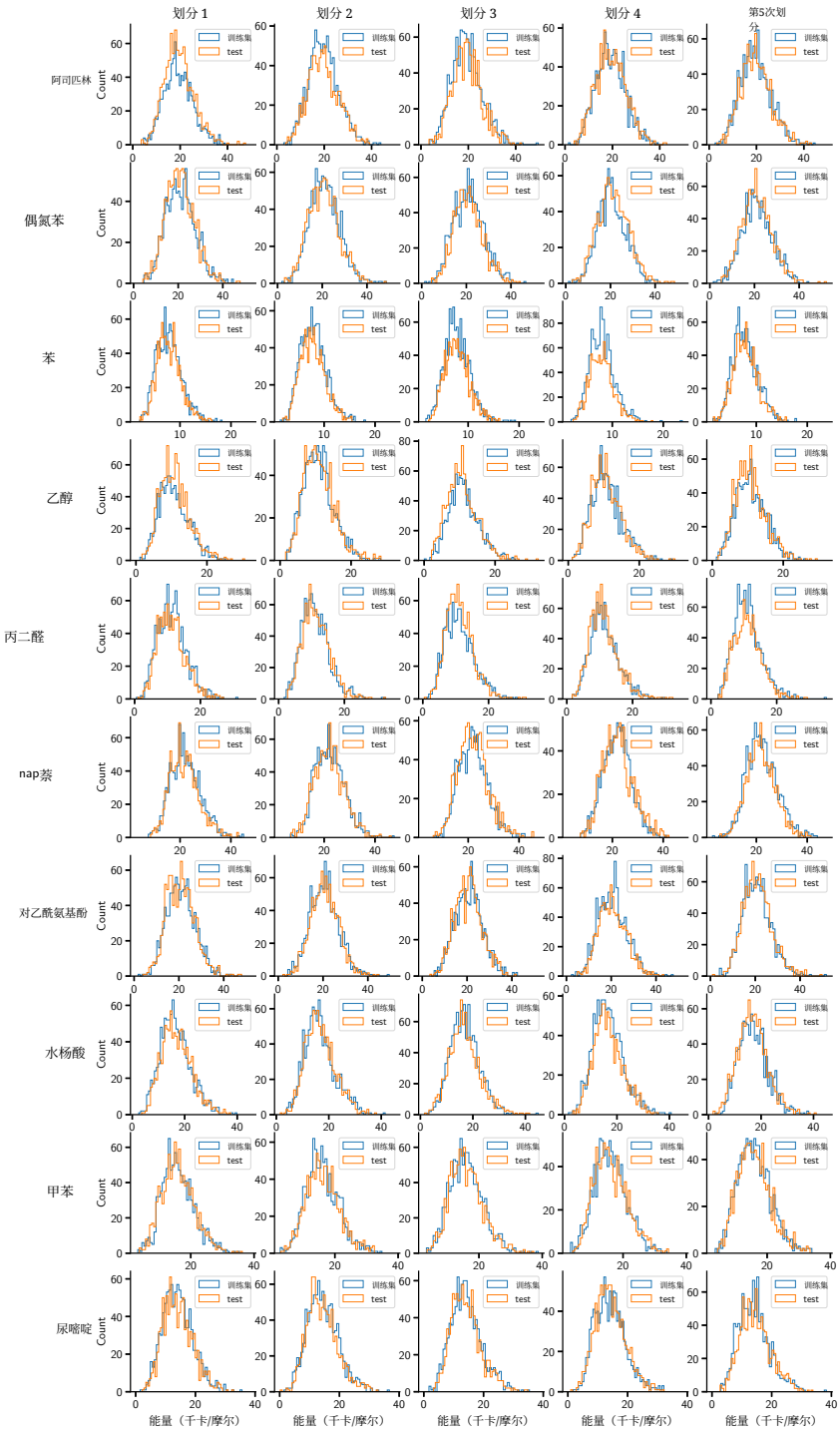

图 S7. rMD17 数据集中所有分子的所有划分（横向）的训练集和测试集能量分布（纵向）。

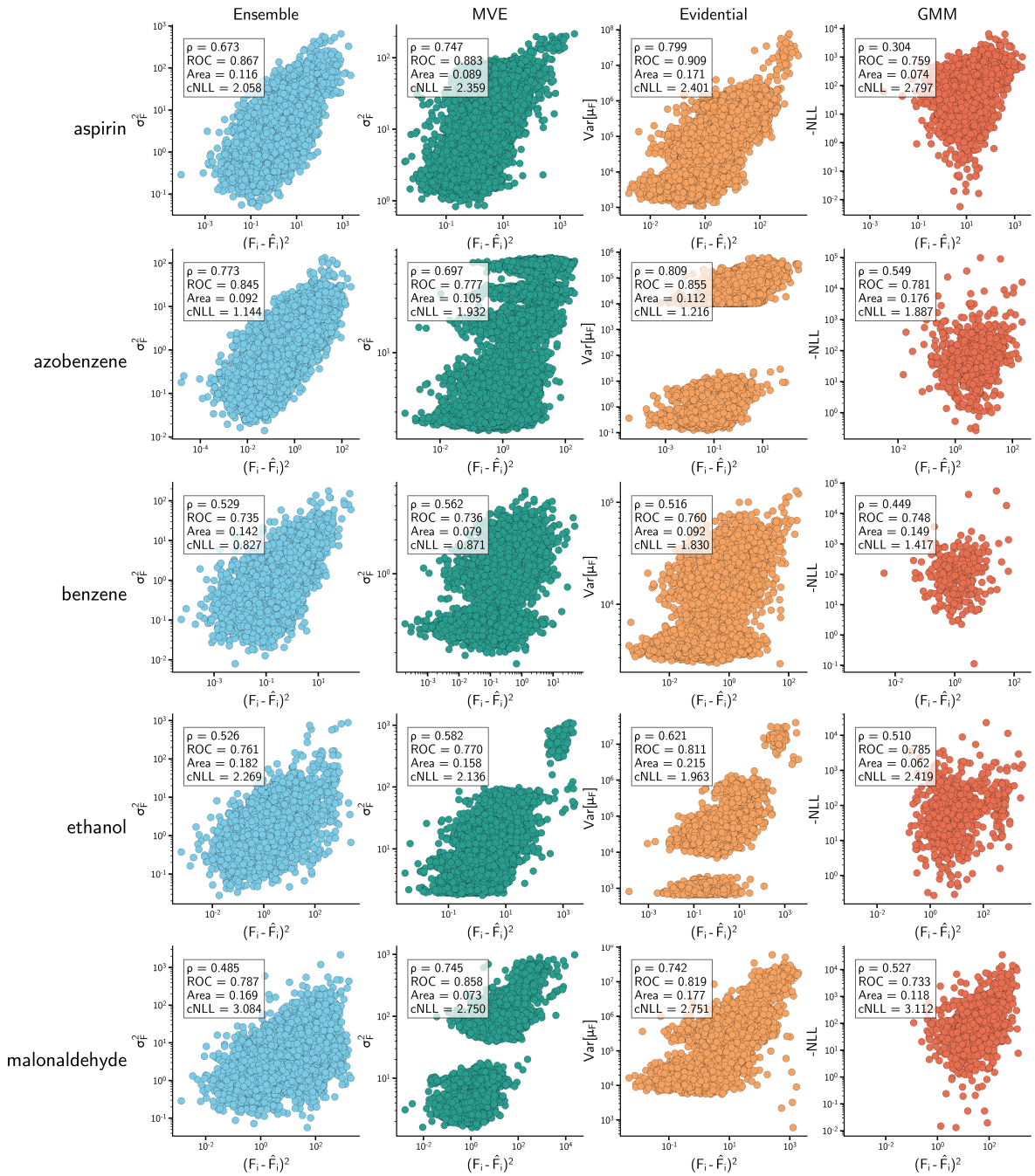

图 S8. 对于所考察的每种不确定性估计方法，展示 rMD17 数据集中阿司匹林、偶氮苯、苯、乙醇和丙二醛分子的原子力（预测）不确定性与平方误差的散点图。请注意，由于 GMM 方法的不确定性（-NLL）包含负值，因此将所有不确定性均按最小值进行缩放，使其成为正值，以便在对数刻度图中绘制。

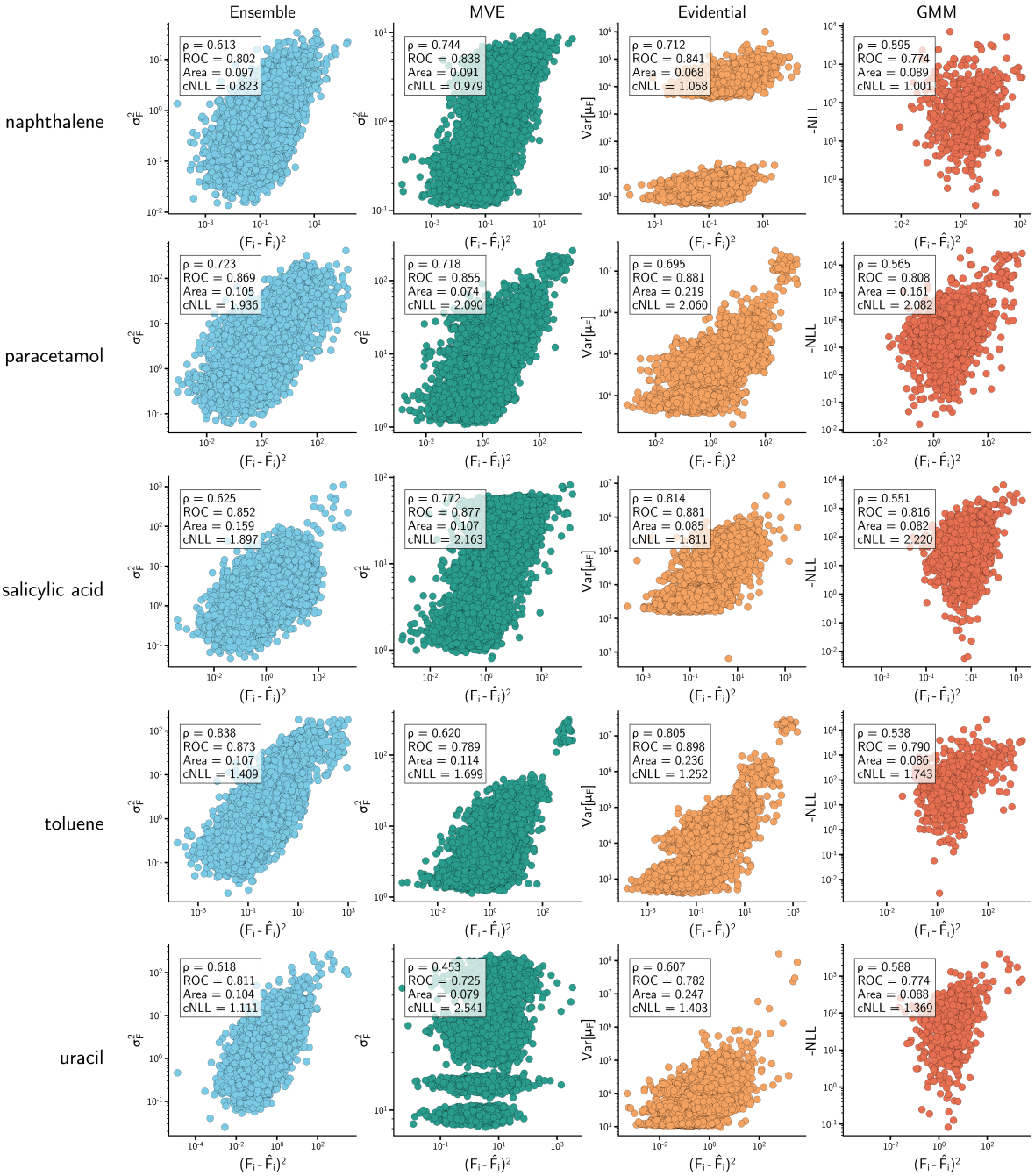

图 S9. 散点图展示了针对所考虑的每种不确定性估计方法，rMD17 数据集中萘、对乙酰氨基酚、水杨酸、甲苯和尿嘧啶分子的原子力（预测）不确定性与平方误差之间的关系。请注意，由于 GMM 方法的不确定性（-NLL）包含负值，因此将所有不确定性均按最小值进行缩放，以得到正值，从而绘制对数刻度图。

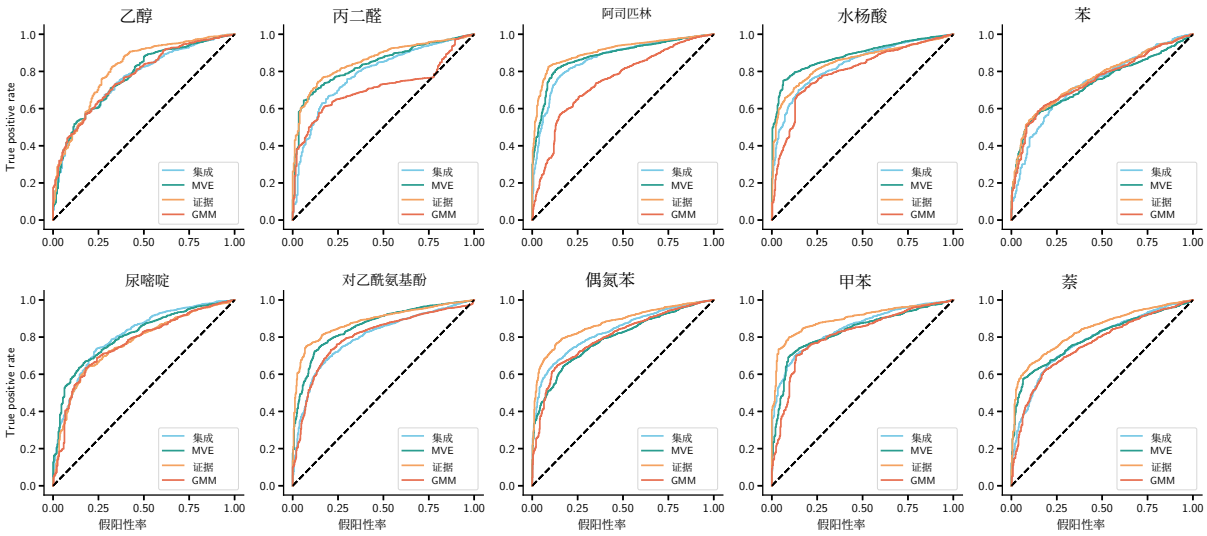

图 S10. rMD17 数据集中所有分子所采用的各不确定性估计方法的受试者工作特征（ROC）曲线。为“目标”（真实误差）指定的阈值为分布的第20百分位数。

	第20百分位数				第50百分位数			
	集成	均值方差估计	证据	高斯混合模型	集成	均值方差估计	证据	高斯混合模型
阿司匹林	0.853	0.840	0.921	0.783	0.815	0.834	0.887	0.733
偶氮苯	0.853	0.826	0.890	0.807	0.883	0.873	0.915	0.821
苯	0.700	0.716	0.764	0.702	0.702	0.733	0.777	0.698
乙醇	0.745	0.765	0.780	0.693	0.755	0.800	0.779	0.679
丙二醛	0.763	0.866	0.849	0.745	0.741	0.887	0.876	0.786
萘	0.786	0.840	0.843	0.781	0.788	0.868	0.867	0.800
对乙酰氨基酚	0.837	0.862	0.874	0.773	0.822	0.870	0.854	0.690
水杨酸	0.835	0.884	0.904	0.794	0.811	0.876	0.901	0.738
甲苯	0.842	0.821	0.924	0.821	0.808	0.839	0.912	0.784
尿嘧啶	0.793	0.771	0.787	0.758	0.805	0.779	0.802	0.779
Mean	0.801	0.819	0.854	0.766	0.793	0.836	0.857	0.751

表S2. 使用第20和第50百分位数确定真实误差（目标）分布中的二元截断值时，所考察的不确定性估计方法在 rMD17 数据集上的受试者工作特征曲线下面积（ROC曲线下面积）得分。

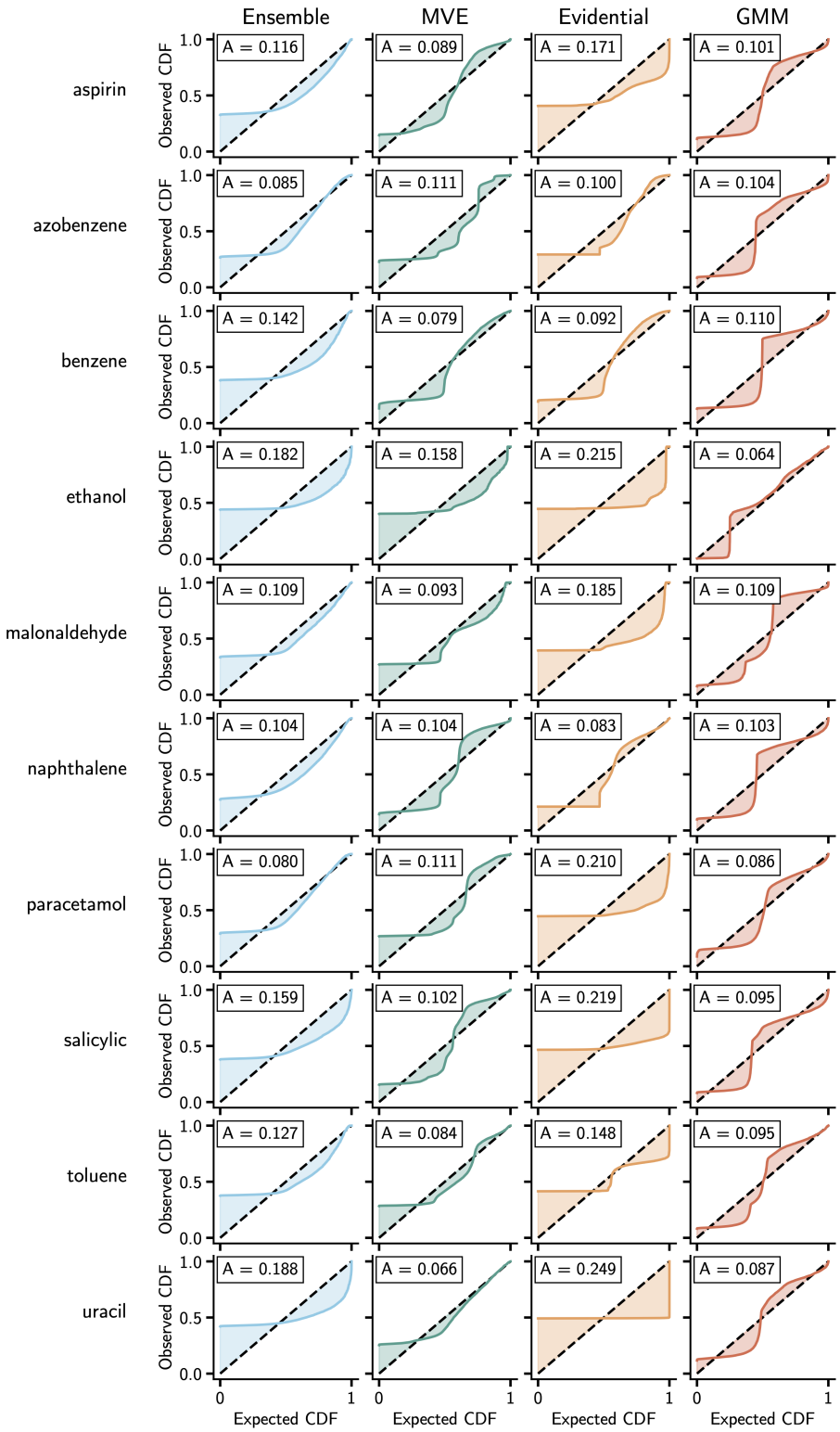

图 S11. rMD17 数据集中所有分子的各个所考察不确定性估计方法的校准曲线（跨分子）。每个子图左上角的方框标示失校准面积，即校准曲线与理想曲线（虚线）之间的阴影区域。

B. 氨数据集

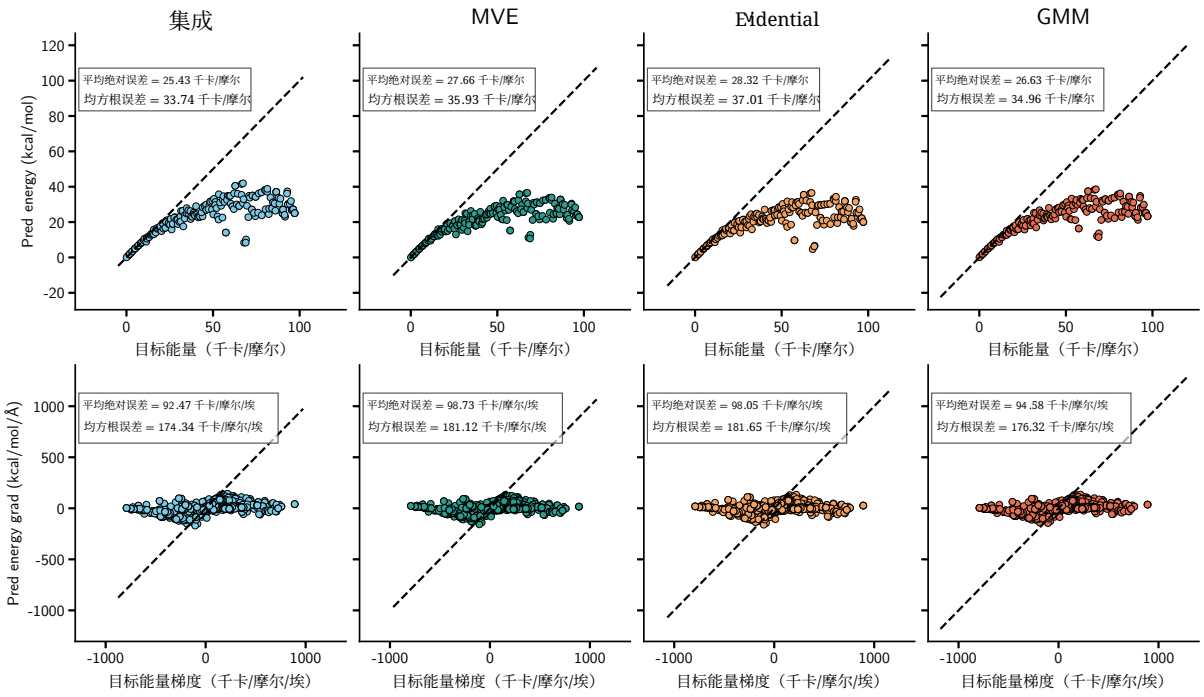

图 S12. 使用第 1 代训练的神经网络模型（仅使用 78 个几何构型进行训练）对氨测试集上的预测能量和力绘制的奇偶图。

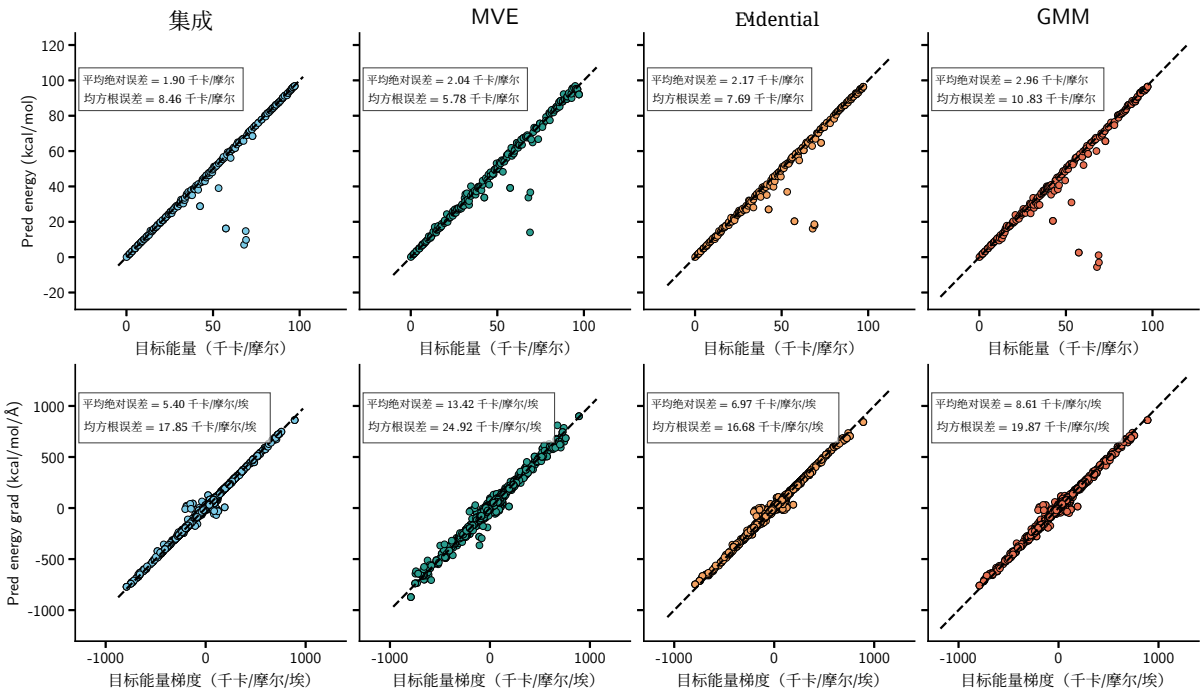

图 S13. 奇偶图显示了使用第 3 代训练的模型（基于 78 个初始几何构型和 40 个对抗样本训练）对氨测试集的预测能量和力。

C. 二氧化硅数据集

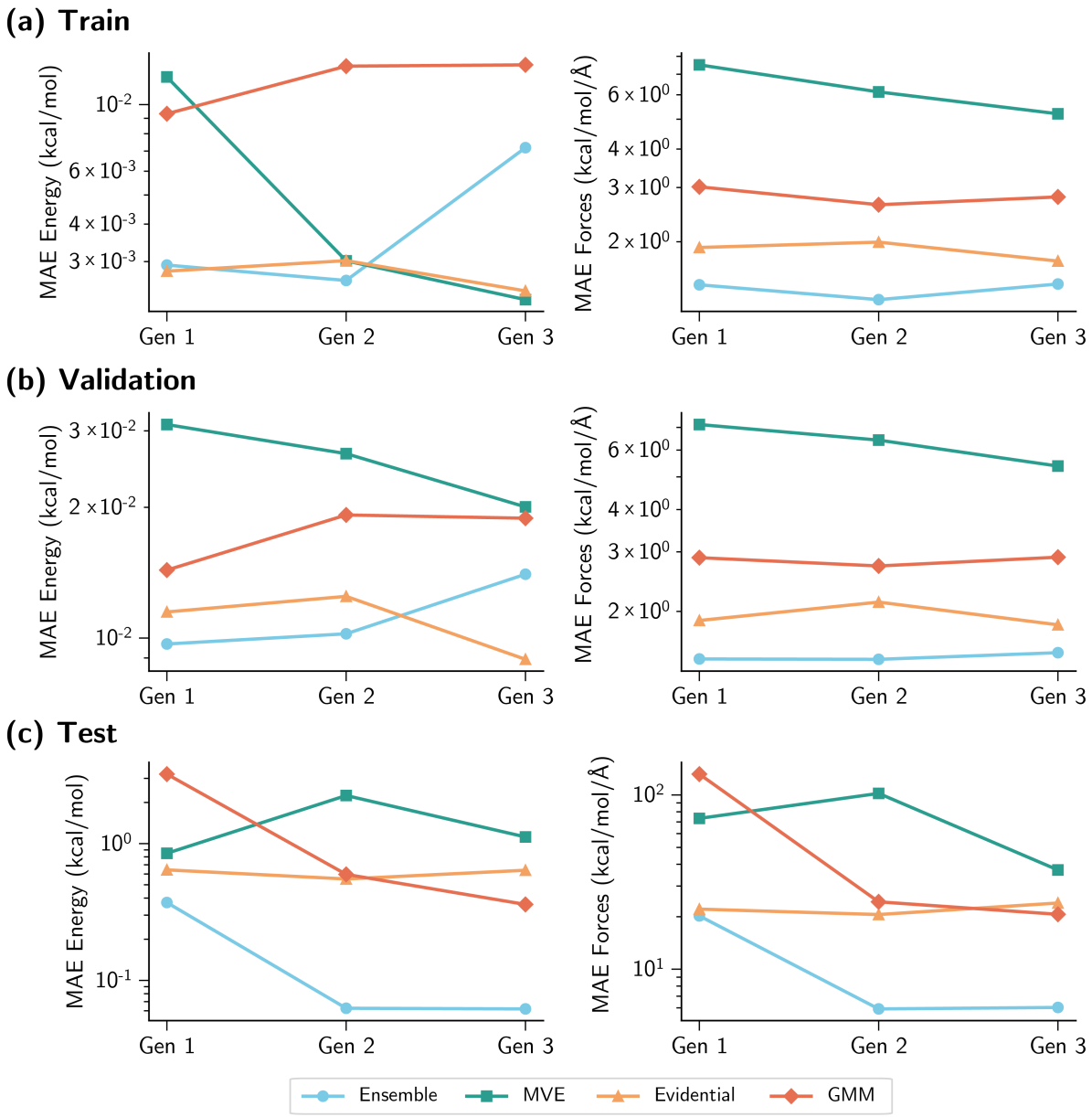

图 S14. 神经网络模型在不同代数下，相对于二氧化硅的（**a**）训练集、（**b**）验证集和（**c**）测试集的预测能量（第一列）和力（第二列）的平均绝对误差。

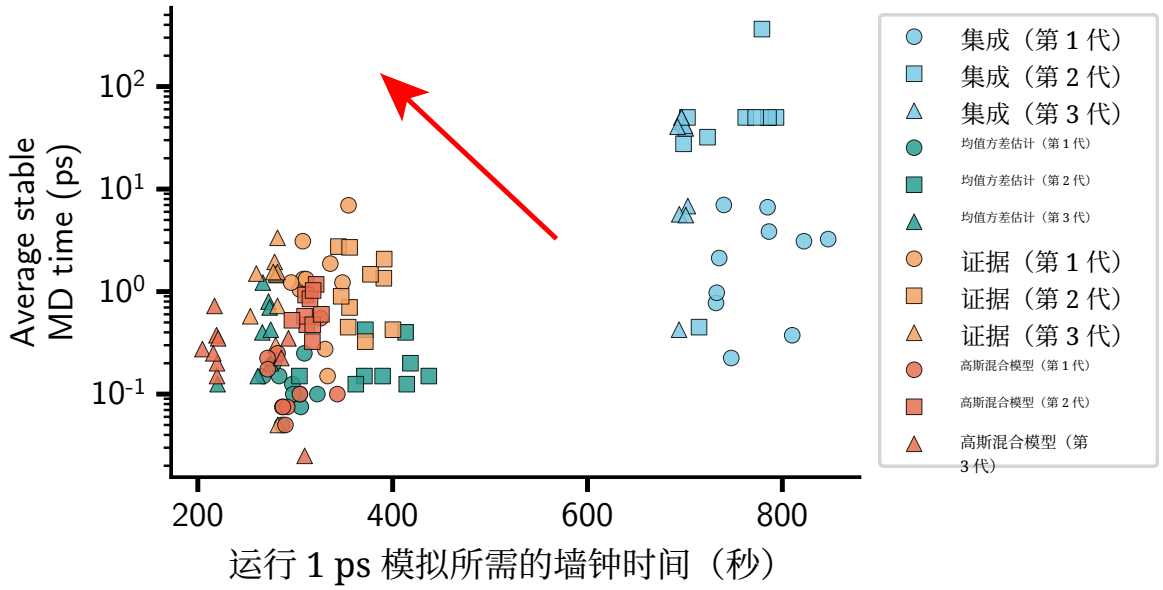

图 S15. 散点图展示了不同代中各不确定性估计方法所采用的神经网络模型，其稳定的分子动力学轨迹长度与运行 1 ps 模拟所需墙钟时间之间的关系。每个标记表示一条在 2500 K 下模拟的、时长为 1 ns 的 NVT 轨迹。

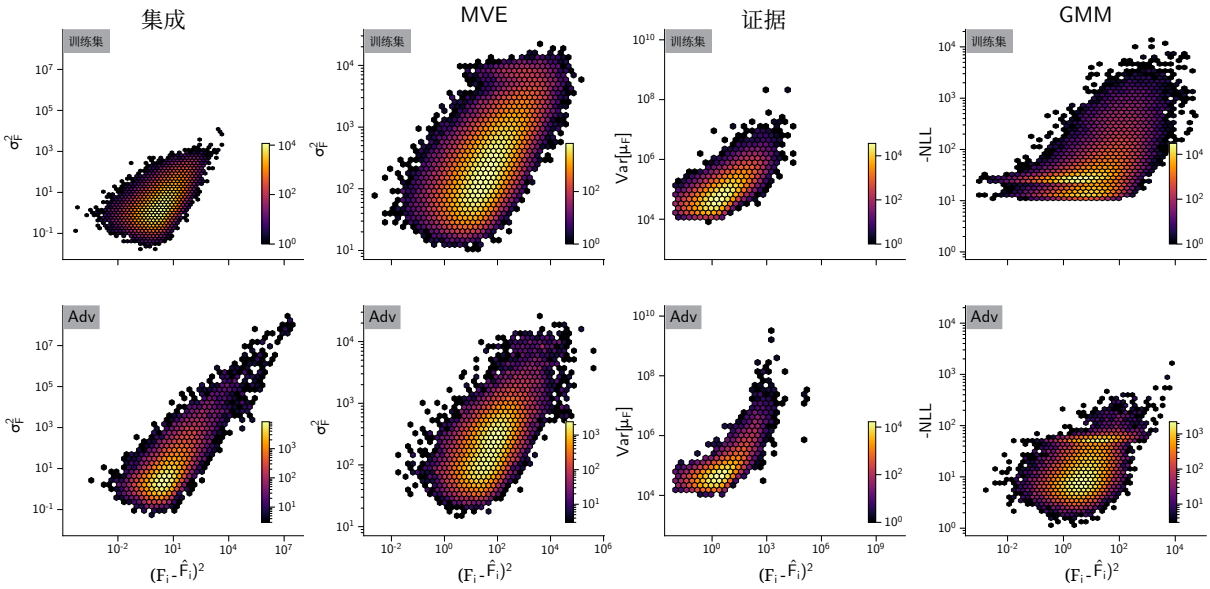

图 S16. 六边形分箱图展示了第 1 代中采用不确定性估计方法的神经网络模型的预测不确定性与真实误差之间的关系。上排展示训练集的分布，下排展示由相应不确定性估计方法生成的对抗样本的分布。

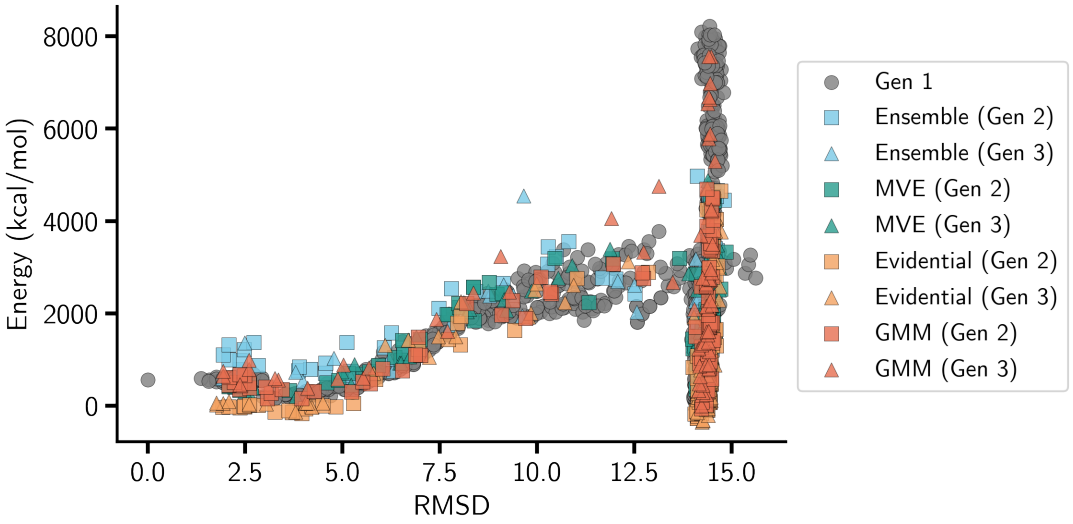

图 S17。训练数据中的二氧化硅结构以及各不确定性估计方法生成的新对抗样本的能量与均方根偏差（RMSD）关系图。所有二氧化硅结构的 RMSD 均相对于训练数据中通过 300 K 平衡模拟采样得到的一个二氧化硅结构计算。

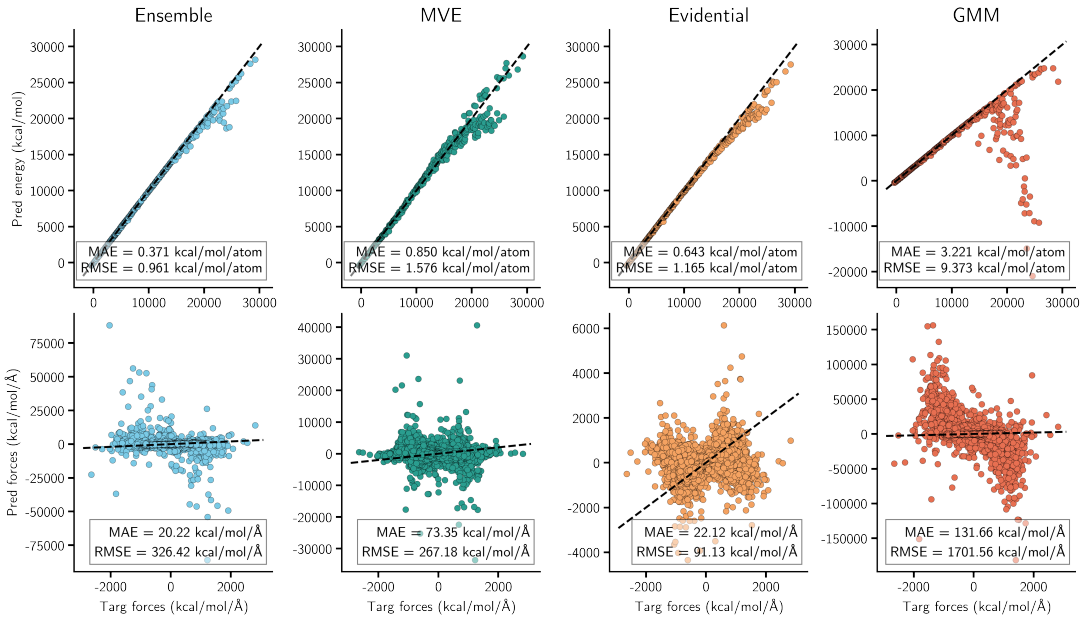

图 S18。使用第 1 代训练的模型（基于 1590 个初始训练数据训练）绘制的二氧化硅测试集预测能量和力的奇偶图。

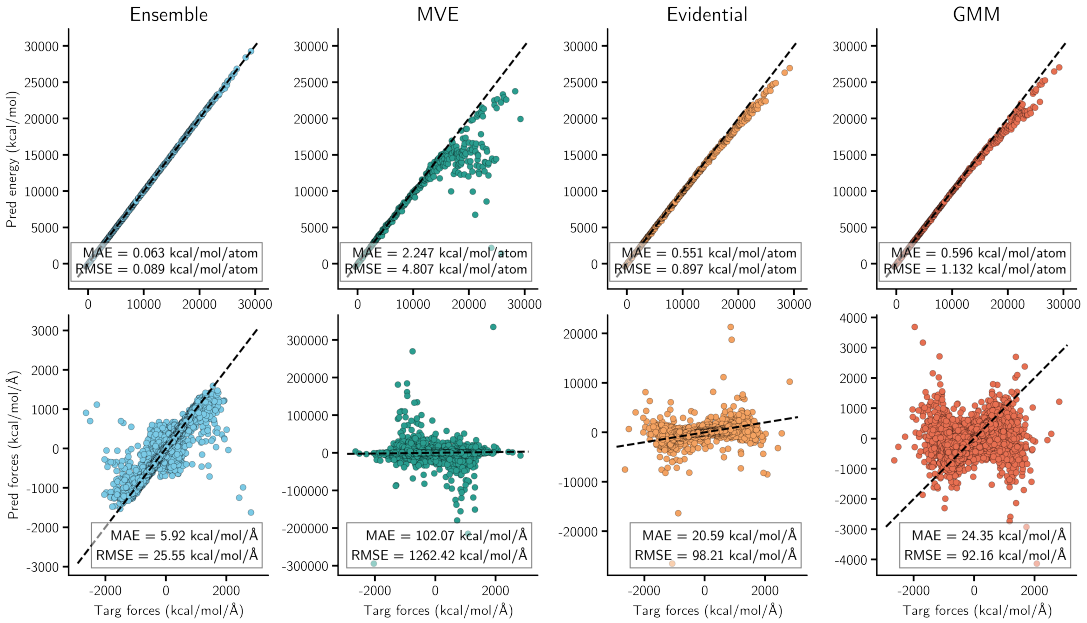

图 S19. 奇偶图显示了使用第 2 代训练的模型对二氧化硅测试集的预测能量和力（在 1590 个初始训练数据和 100 个对抗样本上训练）。

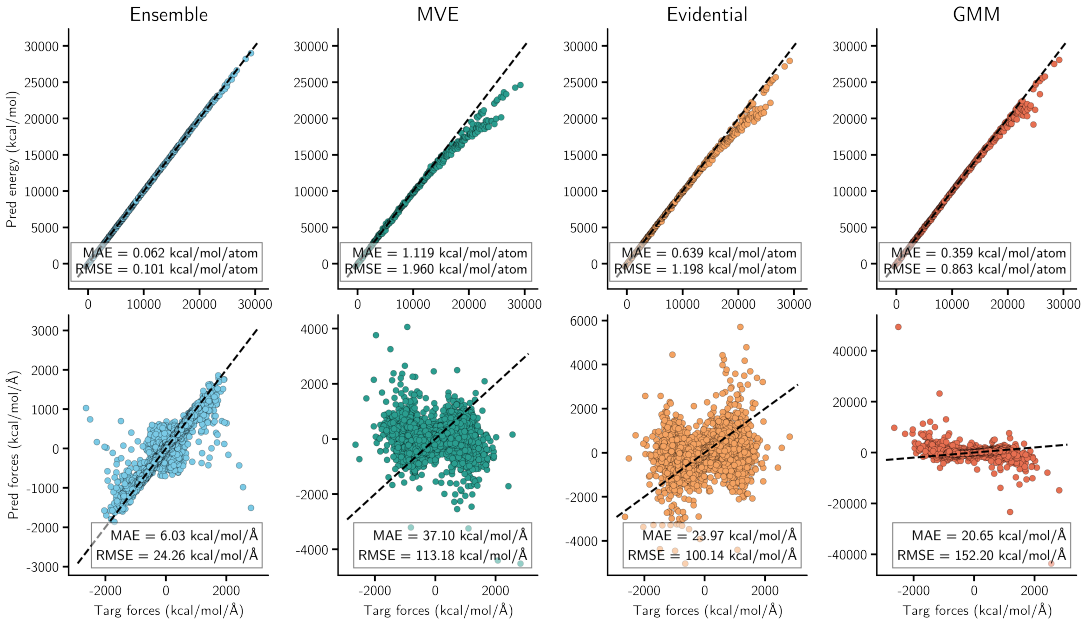

图 S20. 奇偶图显示了使用第 3 代训练的模型对二氧化硅测试集的预测能量和力（在 1590 个初始训练数据和 200 个对抗样本上训练）。