超越扩散模型与流模型的多模态科学学习

Leonardo Ferreira Guilhoto应用数学与计算科学研究生项目，宾夕法尼亚大学，美国宾夕法尼亚州费城
guilhoto@sas.upenn.edu

Akshat Kaushal计算机与信息科学系，宾夕法尼亚大学，美国宾夕法尼亚州费城
akaush@seas.upenn.edu

Paris Perdikaris机械工程与应用力学系，宾夕法尼亚大学，美国宾夕法尼亚州费城 pgp@seas.upenn.edu

arXiv:2602.00960v1 [cs.LG] 1 Feb 2026

摘要

科学机器学习（SciML）日益需要能够刻画由病态逆问题、多稳态和混沌动力学所产生的多模态条件不确定性的模型。尽管近期工作倾向于采用扩散模型和基于流的方法等表达能力很强的隐式生成模型，但这些方法往往数据需求大、计算成本高，且与科学问题中常见的结构化解空间并不契合。我们表明，混合密度网络（MDNs）为科学机器学习中的多模态不确定性量化提供了一种有原则、却在很大程度上被忽视的替代方案。作为显式的参数化密度估计器，混合密度网络施加了针对低维多模态物理的归纳偏置，从而能够在不同解分支之间直接进行概率质量的全局分配。这一结构带来了很强的数据效率，使其能够在科学数据稀缺的情形下可靠地恢复分离的模态。我们通过一个对比显式与隐式分布网络的统一概率框架将这些洞见形式化，并实证表明，在一系列逆问题、多稳态和混沌科学回归任务中，混合密度网络在泛化、可解释性和样本效率方面均表现更优。

关键词 不确定性量化 · 科学机器学习 · 人工智能 for Science · 可信人工智能 · 多模态分布

1 引言

深度学习融入自然科学，通常称为科学机器学习（SciML）或科学人工智能，已使科学建模从纯粹的确定性代理模型转向概率化表述。在经典科学计算中，不确定性通常被视为干扰噪声，并假设其以加性高斯分布的形式围绕某个单一的底层真值波动。对于适定的正问题，这种单峰假设是合理的，但在许多前沿应用中却不再成立，包括混沌流体动力学 [1] 和逆问题 [2, 3]。在此类情形下，输入 x 与输出 y 之间的关系往往不可逆或不适定。相反，支配物理规律可能允许多个符合物理规律的结果，这意味着一组给定的控制参数或初始条件

x 会在输出 y上诱导出多模态条件分布 $p(y|x)$。因此，可靠的科学发现模型必须表征的不只是单一的点预测，还包括条件分布的结构，其中包含彼此分离且具有物理意义的多个模态。

尽管存在这些要求，科学机器学习中的不确定性量化（UQ）仍以单峰建模假设为主。大多数回归模型通过最小化均方误差（MSE）进行训练，该目标严格恢复条件期望 $\mathbb{E}[y|x]$。当真实条件分布为多峰时，这一平均值可能落在概率密度较低的区域。因此，所得预测可能在物理上无效。更具表达能力的方法，如贝叶斯神经网络（BNNs），原则上能够表示复杂的后验分布，但在实践中，由于平均场变分推断的局限性以及标准集成聚合的集中效应，往往退化为单峰近似 [4, 5]。

针对单峰不确定性量化的局限性，机器学习领域的近期工作倾向于采用高度表达性的隐式生成模型，包括变分自编码器 [6], 去噪扩散概率模型（DDPMs） [7]以及条件流匹配（CFM） [8]。这些方法通过学习借助迭代去噪或连续时间动力学将简单基准测度变换为复杂目标分布，在高维生成任务中取得了显著成功。然而，将其应用于科学回归问题时，会引入主流生成建模中通常不会遇到的挑战。近期的理论分析 [9, 10] 表明，扩散与流模型难以准确学习模态分离良好的分布，需要指数级数量的样本才能正确捕获不相连解分支的相对权重。此外，推断的高计算成本——每次预测往往涉及数百次模型评估——限制了它们在实时或资源受限的科学场景中的适用性。这些局限性在科学机器学习典型的低数据量场景中被进一步放大，其中训练集比主流生成建模小若干个数量级。在此类约束下，扩散与流模型往往急剧退化，表现为模态覆盖不佳和训练不稳定。

鉴于这些观察，本文重新审视混合密度网络（MDN） [11] ，将其视为科学机器学习中多模态不确定性量化的一种有理论依据的替代方案。尽管混合密度网络在机器学习文献中早已是一类成熟模型，但它们在现代科学机器学习流程中的应用并不广泛，且此前尚未系统考察其在多模态、低数据量物理状态下的表现。基于近期理论，我们认为，混合密度网络显式的混合结构带来了一种天然适用于不连通解集的归纳偏置，使其能够表示多模态条件密度，同时避免连续输运模型引入的拓扑畸变。我们通过逆问题、多稳态动力系统和混沌时间序列预测实验验证了这些观点。结果表明，与扩散模型和基于流的方法相比，混合密度网络能够可靠地恢复模态，并以大幅更低的数据和计算需求实现具有竞争力的密度估计性能。

本文的主要贡献如下：

- 我们为科学机器学习提出一个统一概率框架，在同一形式体系中涵盖点估计模型、显式密度估计器和隐式生成方法。
- 我们基于统计学习理论，对混合密度网络与非参数模型在低到中等维度问题上进行了系统比较，表明参数混合模型在科学机器学习的典型情形中能够实现更可靠的模态恢复和更高的数据效率。
- 我们强调可解释性 在科学建模中的作用，展示 MDN 混合成分和权重如何具有直接的物理解释，从而能够与相图、稳定性边界和状态分类建立联系，而这些联系在扩散模型或基于流的模型中并不容易获得。
- 我们在具有代表性的科学机器学习基准上开展数值实验，表明混合密度网络在低数据情形下运行时，其条件密度估计性能可与现代基于扩散模型的方法持平或更优。
- 我们发布了 JaxMix，这是一个基于 JAX 的 Python 软件包，供计算科学家探索混合密度网络及其他多模态建模方法。它可在 https://github.com/ PredictiveIntelligenceLab/JaxMix 公开获取。

2 科学问题中的多模态不确定性

科学机器学习日益应用于这样的情境：确定性预测既不可能，也没有意义。许多物理系统具有噪声、观测不完整、混沌或规格不足等特征，因此同一个输入可能对应多个有效结果。然而，大多数深度学习方法仍以点估计为基础，通过最小化均值

平方误差（均方误差，MSE），隐含地将可变性视为噪声，而非真实的物理歧义。当真实的条件分布呈多峰时，这一范式产生的预测可能不符合物理规律或具有误导性。

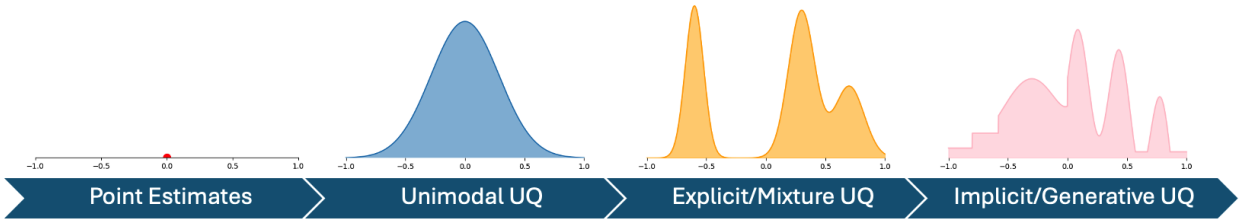

图1：不确定性量化复杂程度逐渐增加的示意图。点估计不提供任何不确定性信息（左），而任意生成模型虽然功能强大，却需要大量数据和计算资源（右）。许多科学问题都可以用更简单的多模态混合分布来近似。

图1展示了各种不确定性量化方法所处的不同层次。一端是完全忽略不确定性的点估计，另一端则是能够表示任意复杂分布、但需要大型数据集和大量计算资源的非参数生成模型。许多科学问题处于中间状态：输出维度较低或中等，且不确定性具有结构化的多模态特征。对于这类问题，显式的基于混合的模型能够兼顾表达能力、统计效率和可解释性。

不确定性量化中的一个核心区别是认知不确定性与偶然不确定性。认知不确定性源于数据有限或模型设定错误，原则上可以通过收集更多观测值来降低。偶然不确定性则无法消除：它反映了固有的随机性、尚未解决的自由度，或模型中未包含的潜变量。在许多科学问题中，多模态是偶然不确定性的一种表现。即使拥有无限的数据和完美的模型，条件分布 $p(y|x)$ 仍可能是多模态的。主要针对认知不确定性的方法，例如深度集成和贝叶斯神经网络，通常无法捕捉偶然不确定性。本文关注的是对偶然不确定性本身所形成的解空间结构进行建模。

2.1 单峰预测模型的局限性

大多数回归不确定性量化方法都隐含地假设预测分布是单峰的，通常为高斯分布。这一假设在若干科学场景中存在严重局限。

高斯过程虽然在理论上严谨，但采用高斯先验和似然函数，导致后验预测呈单峰，并将概率质量分散到不同结果之间 [12]。同样，贝叶斯神经网络（贝叶斯神经网络）在实践中往往会将输出上的多模态分布压缩为宽广的单峰密度，即使权重上的分布是多模态的 [4]。深度集成能够提升鲁棒性并改善分布外检测，但若使用均方误差目标进行训练，预测仍会集中在条件均值附近，从而捕捉认知不确定性，而非数据生成过程的结构 [5]。

病态逆问题。 逆映射通常是一对多的。单个观测可能对应多个物理上有效的构型，例如非线性逆回归问题中的情况。单峰预测器会返回这些解的凸平均值，产生的估计可能不对应任何可实现的物理状态。

混沌动力系统。在具有正李雅普诺夫指数的系统中，例如低维混沌流，长期逐点预测从根本上是不可能的。有意义的预报必须表示轨迹上的分布，而不是单一路径。

多稳态系统。许多物理系统表现出由吸引域分隔的多个稳定吸引子。噪声或微小扰动可以决定到达哪个吸引子，从而使条件结果分布呈多峰，即使底层动力学是确定性的。

3 科学学习的概率框架

为统一讨论科学机器学习中的建模选择，并阐明混合密度网络相较于现有方法的定位，我们引入一个概率框架，按照模型表示条件不确定性的方式对其进行分类。该框架刻意保持精简，但足以说明不同方法之间的差异。

3.1 从点预测到条件分布

设 $\mathcal{D} = \{(x_i,\ y_i)\}_{i=1}^{N}$ 为一个由输入–输出对构成的数据集，其中 $x \in \mathcal{X}$ 为控制参数、初始条件或观测值，$y \in \mathcal{Y}$ 表示一个关注量。

主流的 深度学习回归范式是点预测：一种形式如下的确定性映射

$$f_\theta : \mathcal{X} \to \mathcal{Y}, \tag{1}$$

通过最小化均方误差等逐点损失进行训练。这种方法隐含地假设不确定性为单峰，且可由条件均值很好地概括。对于适定确定性问题，这一假设是合适的；但当不确定性反映的是真实的物理歧义而非测量噪声时，这一假设便不再成立。

概率科学机器学习模型则旨在学习完整条件分布 $p(y|x)$。我们将任何此类模型称为分布网络。分布网络的主要区别在于如何表示这一条件分布，以及训练过程中如何分配概率质量。

3.2 显式与隐式分布网络

D分布网络大致可分为两类：显式和隐式模型 s.

显式分布网络。显式模型直接对条件概率密度函数进行参数化，并通过最大似然进行优化。形式上，它们输出一个可处理密度 $p_\theta(y|x)$，并最小化负对数似然 (NLL)，

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(y_i|x_i). \tag{2}$$

显式模型的决定性特征是，概率质量分配是目标函数的显式组成部分。模态权重、尾部行为或低概率区域中的误差都会受到直接惩罚。因此，当下游决策依赖于校准良好的似然而非仅依赖点预测或样本时，这些模型天然适用。

隐式分布 网络。隐式模型通过采样过程定义条件分布，

$$y = G_\theta(x, \varepsilon), \qquad \varepsilon \sim p(\varepsilon), \tag{3}$$

而无需密度的闭式表达式。该类包括生成对抗网络、条件变分自编码器 [13], 扩散模型 [7],以及条件流匹配 [8]。

隐式模型具有很强的表达能力，能够在高维设定中逼近复杂分布。然而，由于密度并未被直接优化，训练依赖于对抗损失或分数匹配等代理目标。这些目标优先保证样本的局部几何保真度，而非对全局概率质量分配的直接控制。

3.3 混合密度网络作为结构化条件模型

在显式分布网络这一类中，混合密度网络（混合密度网络）将条件密度参数化为高斯分布的有限混合：

$$p_\theta(y|x) = \sum_{k=1}^{K} \alpha_k(x;\theta)\, \mathcal{N}\left(y|\mu_k(x;\theta), \Sigma_k(x;\theta)\right), \tag{4}$$

其中，混合权重 α_k（x）构成一个单纯形（即 $\sum \alpha_k = 1$ 且 $\alpha_k \geq 0$），并与分量参数 μ_k 和 Σ_k联合学习。分布参数 $\{\alpha_k, \mu_k, \Sigma_k\}$ 是神经网络骨干的输出。

这一表述编码了一项简洁但强大的结构假设：不确定性可以分解为少量彼此 distinct 的结果，而每个结果都可以用高斯分布进行良好近似。重要的是，该表示将概率质量所在的位置（分量均值和协方差）与每个分量所承载的质量大小（混合权重）区分开来。

因此，混合密度网络将显式似然优化的校准优势与适用于不确定性分解为离散且可解释分支的问题的结构先验相结合。它们与现代生成模型的关键区别在于归纳偏置：混合密度网络将不确定性分解为离散分量，而隐式模型则无法提供这种保证

。对于许多具有结构化不确定性、中等维度和有限数据的科学问题而言，这在表达能力、效率与可解释性之间实现了强有力的权衡，从而使混合密度网络成为不确定性感知科学机器学习的一个有原则的基础。附录 B 提供了更多见解，并为计算科学家更完整地介绍了混合密度网络的实现，其中包括如何选择混合分量数量 K 的讨论。

4 多模态科学学习中的归纳偏置

科学机器学习问题通常处于这样一种情形：输出维度较低至中等，具有结构化的多模态不确定性，且数据有限。这种情形——输出维度为 $\mathbb{R}^d$，具有 $d \leq 20$、结构化的多模态不确定性，训练集规模为 $N < 10,000$——涵盖了广泛的实际科学问题，包括降阶建模、参数估计，以及动力系统中的状态推断。在这类情况下，性能更多取决于归纳偏置，而非原始表达能力。本节将论证，混合密度网络与现代扩散模型或基于流的模型体现了根本不同的归纳假设，因此在数据效率、拓扑保真度和可解释性方面呈现出系统性差异。

4.1 显式与隐式分布网络的样本效率

第 3 节介绍的显式与隐式之分会直接影响统计性质。混合密度网络是参数化的估计器：它们假设真实的条件密度属于某个有限混合分布族，或可由该分布族良好近似。满足这一假设时，统计学习理论的经典结果保证最大似然 [14]具有 $O(n^{-1/2})$ 阶的参数收敛速率。近期研究进一步细化了针对高斯混合的这些保证，表明要达到误差 $\epsilon > 0$，样本复杂度随输出维度 d 和分量数量 K呈多项式增长，其阶为 $\tilde{O}(Kd^2/\epsilon^2)$ [15, 16]。这种情形与科学机器学习尤其相关，因为其输出往往是低维量，例如位置或降阶摘要。

隐式生成模型在一种根本不同的机制下运行。扩散与流模型作为非参数 密度估计器，学习环境空间上的连续输运映射或得分场。在缺乏对目标分布的强结构假设时，非参数估计已知会受到形式为 $O(n^{-s/2(d+s)})$ 的较慢收敛速率的影响，其中 s 反映光滑性假设 [17]。近期研究从理论上并通过实验证明，扩散模型与流匹配等基于得分的方法的样本复杂度通常为 $\tilde{O}(n^{-1/2(d+5)})$ [18]。

这一区别在数据有限的场景中变得至关重要。由于隐式模型通过局部更新进行学习，它们需要足够的样本，不仅要填充高密度区域，还要填充连接这些区域的路径。相比之下，显式的基于似然的模型可以通过其参数在全局范围内重新分配概率质量，并且对分布的衰减速率具有内置假设。在混合密度网络中，来自稀有众数的单个观测会在相应的混合权重上产生直接的梯度信号，从而立即修正估计的混合比例。因此，在混合假设合理的前提下，混合密度网络能够以少得多的样本恢复多模态结构。
这一对比凸显出，混合密度网络与扩散模型的相对优势源于归纳偏置，而非表达能力。显式混合模型编码了这样一种假设：不确定性具有结构性，且可分解为少数几个组成部分。对于许多科学反问题和动力系统而言，这一假设反映了物理结构，并能带来更高的统计效率。

4.2 隐式分布网络中的似然伪影

当目标分布的支撑集不连通或仅弱连通时，基于分数和基于流的模型会面临一个核心难题。近期对分数匹配的分析表明，样本复杂度与目标分布的等周常数成反比 [9]。直观而言，当各个模态被低密度区域隔开时，它们之间的分数信号会很弱，或难以定义。因此，扩散模型和基于流的模型可能准确学到每个模态的整体形状，却无法估计其相对权重，从而系统性地高估低概率区域。

这一现象与拓扑密切相关。基于流的模型被构造为光滑的可逆映射，因此会保持连通性等拓扑不变量 [19, 20]。当基分布连通而目标分布不连通时，连续传输无法在不引入人为概率“桥接区域”或高度刚性的向量场的情况下分离模态。这些伪影在许多科学应用中对应于物理上不合理的状态。图3通过持续高估位于高似然模态之间的罕见事件的似然，在实践中展示了这一点。扩展分析见附录D和E。

混合密度网络从构造上避免了这一问题。由于它们优化显式混合分布的负对数似然，概率质量可以通过混合权重在全局范围内重新分配，而无需穿越低密度区域。此外，各个高斯分量具有一个强烈但合理的假设，即概率分布呈指数衰减，这有助于刻画低似然区域。

4.3 混合密度网络的可解释性

混合密度网络具有显式的参数化结构，由此带来的另一项重要优势是可解释性。通过将条件分布建模为有限混合，混合密度网络可将不确定性分解为少数几个不同的组成部分。在科学场景中，这些由 $\mu_k(x)$ 和 $\sigma_k(x)$ 表示的组成部分，往往对应有意义的解分支；相应的混合权重 $\alpha_k(x)$ 则刻画了它们随输入 x 变化的相对似然。具体而言，标量场 $\alpha_k(x)$ 充当学习得到的序参量：$\alpha_k(x) \approx 1$ 的区域对应确定性结果，而主导组成部分之间的转换则标示出相界或分界曲线。这种结构可以直接从训练后的模型中获取，无需辅助损失、事后探查或蒙特卡罗分析。因此，混合密度网络能够清晰展现不确定性的组织方式，而不仅仅是样本的生成方式。相比之下，即使隐式生成模型能够生成高质量样本，也无法以可直接使用的形式揭示概率结构，因此科学解释仍依赖于额外的采样和分析。

这一现象的一个例子见 5.2 节所述的引力吸引子示例，其中混合密度网络学会为三个可能的吸引子各分配一个混合分量。此外，每个吸引子对应的混合权重 $\alpha_k(x)$对输入的依赖关系，展现了不同情境下可解释的结构，如图4所示。

5 实验

我们通过一系列多样化的实验，对关于样本复杂度、拓扑鲁棒性和可解释性的主张进行实证验证。我们在四个具有内在多模态性的问题上比较混合密度网络、条件流匹配和均方误差。我们选择条件流匹配作为隐式生成方法的代表，因为它体现了当前基于传输的条件生成最佳实践，并且与扩散模型共享核心的连续传输机制。完整的实验细节见附录 F。

5.1 正弦逆问题

我们从一个受 [11] 启发的正弦逆问题基准开始。考虑由 $y = f(x) + \epsilon$ 定义的正向过程，其中 $f\colon \mathbb{R} \to \mathbb{R}$ 由 $f(x) = \frac{x}{2} + 0.7\sin(5x)$ 给出，且 $\epsilon \sim N(0,\ \sigma^2)$ 满足 $\sigma = 0.2$。给定由该过程生成的数据，我们的目标是根据含噪观测推断条件密度 $p(x|y)$，而不是求得单一的逆解。

尽管这是一个低维问题，却体现了科学反问题的一个典型特征：前向映射适定且具有单峰性，但其逆映射本质上是一对多的。因此，使用均方误差训练的确定性回归会将解集压缩为单一的条件期望，产生与任何可实现解都不对应的估计值。图 12（附录）展示了这种失效模式：模型预测了逆映射中互不相容分支之间的平均值。

图2展示了该逆问题的几何结构，其中解流形描绘出一条折叠的非单射曲线。在定义域的中心区域，单个观测 y 在 f下有多个互不相连的原像，从而形成多峰条件密度。由于这种条件结构可以用平滑分量的有限混合很好地近似，混合密度网络提供了一种自然的参数化归纳偏置。与这一观点一致，混合密度网络在低数据情形下的误差显著更低（见图3），并且随着样本数量增加，估计精度迅速提升。

相比之下，条件流匹配通过连续传输学习密度，而不显式利用结构；如图3上图所示，要达到相当的拟合效果，它需要显著更多的数据。随着训练集扩大到 $N = 10,\ 000$，两种方法都会收敛，但条件流匹配仍持续倾向于在不连通的解分支之间引入过度平滑的概率桥接区域。这种伪影并非条件流匹配所独有；任何连续归一化流都会面临类似的拓扑约束。图3中的条件负对数似然显示了这一现象：条件流匹配会将质量分散到低概率区域，高估罕见事件，并人为降低负

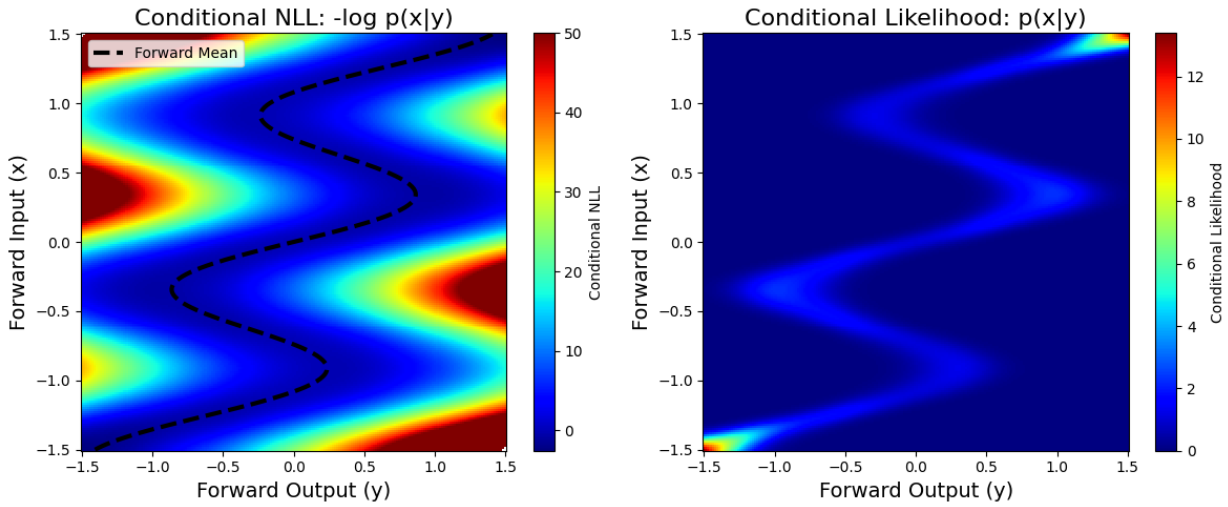

图2：在反正弦问题中给定 y 时 x 的真实条件密度。（左）条件负对数似然。（右）条件似然。

对数似然。这些结果从实证上支持了第4节所述混合密度网络的参数效率和拓扑对齐特性，凸显了其在具有不连通解集的逆问题中的优势。

5.2 部分观测的引力吸引子

在第4节中，我们论证了可解释性在科学机器学习中的重要性，即能够从模型的内部表征中恢复有意义的宏观结构。在这一视角下，MDN 的混合权重 $\alpha_k(x)$ 充当学习到的序参量，直接从数据中编码相结构和吸引域归属。我们在一个具有多个稳定吸引子的简单耗散动力系统上验证了这一主张。

我们模拟了一个粒子在二维平面中的运动：该粒子在最多三个固定天体的引力作用下运动，这些天体位于单位圆上。系统包含阻力项，因此具有耗散性：所有轨迹最终都会收敛到某个吸引子。初始位置从半径为 0.2 的圆盘中均匀采样，并加入小幅各向同性观测噪声。学习任务是给定初始位置，预测带噪声的离散化轨迹。

为单独考察不同不确定性来源的影响，我们针对三种实验设置分别训练了一个混合密度网络。在所有情形下，每个预测的混合均值 $\mu_k(x)$ 都会自发学会编码通向某个特定吸引子的轨迹。在此过程中，模型区分了粒子最终位置的各种彼此独立、离散的可能结果。

在我们考察的第一种情形中，每条轨迹都在一个随机选定的单一吸引子的影响下演化。混合密度网络正确学会了与输入无关的混合权重，这表明最终结果与初始位置 x 脱钩。在第二种情形中，所有吸引子均处于激活状态，但初始位置受到噪声扰动。此时，学得的混合权重揭示了原点附近的不确定性较高，因为此处吸引域重叠；而在更远处，归属则愈发明确，因为结果也变得更加可预测。在第三种设置中，所有吸引子均处于激活状态，且不存在输入噪声。即使在这种确定性情形下，混合密度网络也会将状态空间划分为不同的吸引域，实际上表现得如同一个混合专家模型 [21, 22]，能够学会系统的底层动力学。

图4展示了所有三个场景中的主导混合成分及其对应权重。相比之下，在同一数据上训练的条件流匹配模型无法还原吸引盆结构或决策边界。在所有三个问题中，混合密度网络的测试负对数似然均优于条件流匹配（见表5）。这一比较进一步印证了本文的核心观点：混合密度网络的参数化混合结构兼具效率与语义可解释性，能够提供隐式生成模型难以直接揭示的洞见。

5.3 常微分方程鞍结分岔

动力系统中的分岔代表临界跃迁，即系数的微小变化会大幅改变系统行为。在这些系数隐藏的欠定系统中，必须进行准确的不确定性量化，才能确定可能的轨迹 [23, 24]。

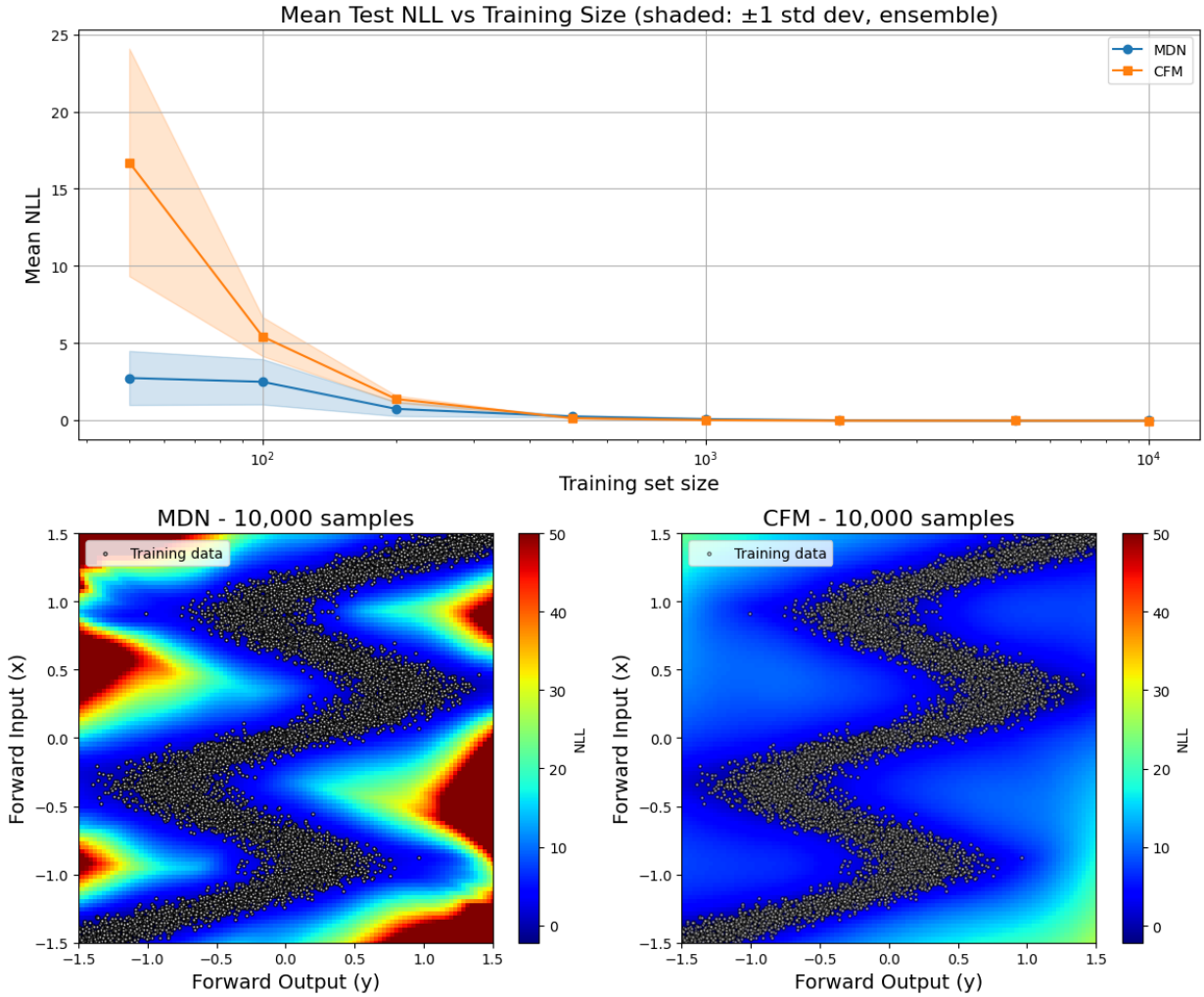

图3：混合密度网络和条件流匹配在正弦逆问题上的比较。上：测试负对数似然随训练集规模的变化，带状区域表示基于12个随机种子的1个标准差。下：平均负对数似然曲面。

鞍结分岔是动力系统中最基本的分岔之一。我们考虑标准型（附录 F 中有进一步说明），其中 x 是状态变量，$r \in \mathbb{R}$ 是分岔参数，

$$\frac{dx}{dt} = r + x^2. \tag{5}$$

系统的行为高度依赖于 r。对于 $r < 0$，存在两个不动点，分别位于 $x^* = \pm\sqrt{-r}$；其中，点 $-\sqrt{-r}$稳定（吸引的），点 $+\sqrt{-r}$ 不稳定（排斥的）。对于 $r = 0$，存在一个半稳定不动点，位于 $x^* = 0$（鞍结分岔点）。对于 $r > 0$，不存在不动点，所有轨迹都会发散至无穷远。

对于这个问题，我们将分岔参数 r 视为潜变量：模型仅接收初始条件x_0 作为输入。这就构成了一个欠定问题，模型必须学会针对每个可能的 r预测一条不同的轨迹，从而产生高度多模态的分布。

图5展示了混合密度网络与条件流匹配之间的样本效率比较。两种方法都成功捕捉到了该问题的多模态特性，能够正确识别不同的轨迹。然而，混合密度网络在低数据情形下展现出明显优势：在使用 $N = 50$ 个训练样本时，混合密度网络的测试负对数似然为 85.6 ± 6.5，而条件流匹配为691.2 ± 146.5。这一数量级的优势在不同超参数配置和随机种子下均持续存在，证实这反映的是样本复杂度差异，而非优化失败。该差距具有重要的实际意义：在许多实验设置中，即使收集50条标注轨迹也可能成本高昂。随着数据量增加（$N = 10,000$），两种方法的性能趋于相近。

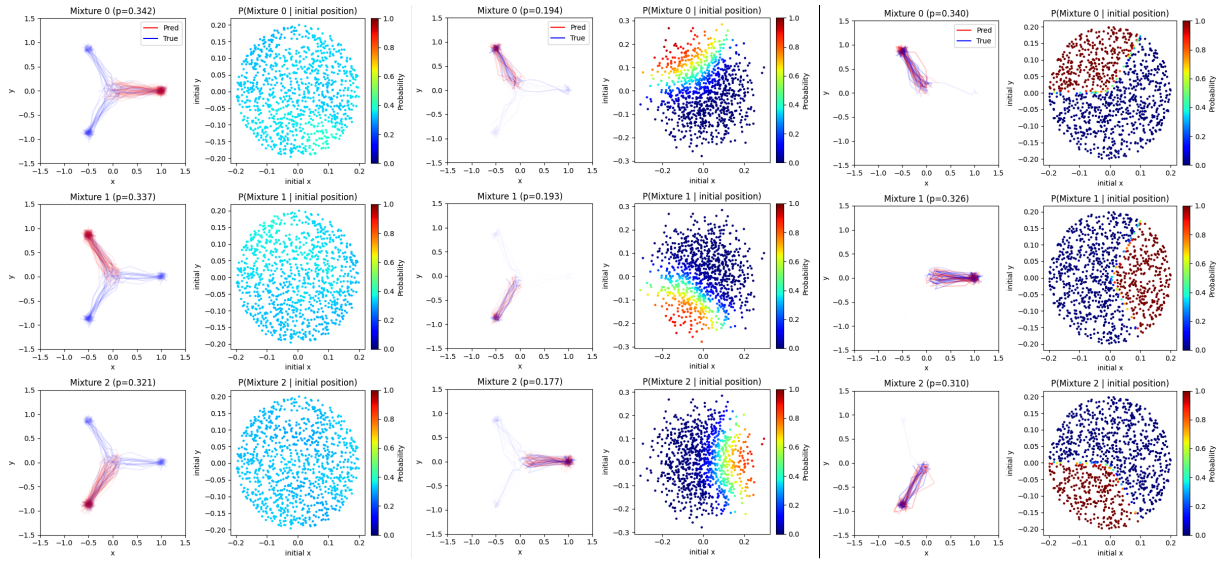

图4：引力吸引子示例中三种情形下按边缘概率排名前三的混合成分。左列显示混合成分样本，右列显示给定输入时该混合成分的权重。样本不透明度按混合概率加权。（左）案例1。（中）案例2。（右）案例3。

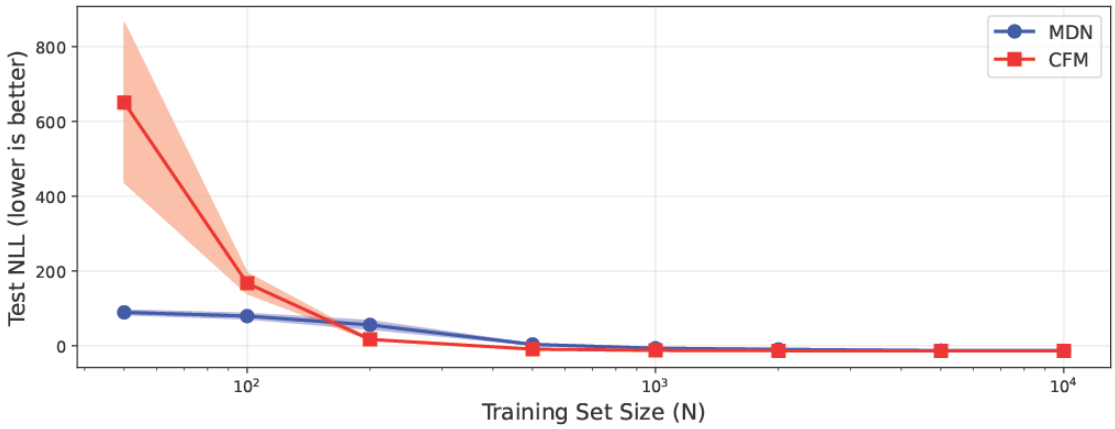

图5：数据效　　率比较，分岔系统中。阴影区域表示12个集成的 ±1 标准差　　e
成员　·

在该任务中，采用均方误差训练的标准网络会因预测条件均值而失败；条件均值是发散轨迹与收敛轨迹的平均值，而真实轨迹则呈现出明显的多模态行为。这种“平均化”行为是点估计模型的根本局限：当存在多个彼此不连通的可能结果时，均值预测可能落在概率密度较低的区域，从而产生物理上不合理的预测。相比之下，混合密度网络和条件流匹配能够正确捕捉这两种模态，生成跟踪发散轨迹和收敛轨迹的样本。

5.4 洛伦兹系统

最后，我们评估不同的不确定性感知模型捕捉洛伦兹63系统 [25] 混沌动力学的能力。在这种状态下，对初始位置的无穷小扰动会随时间呈指数增长，因此逐点长期预测在根本上是不适定的。加入观测噪声会加剧这一现象，实际上即使在较短的时间范围内，也必然会导致预测发散。因此，要获得有意义的预测，就必须对未来轨迹的分布进行建模，而不是仅预测一条确定性路径。

我们训练自回归模型来预测状态增量 $\Delta x_t := x_{t+\Delta t} - x_t$，并通过滚动预测评估其长期行为。我们比较了使用均方误差训练的标准神经网络、前馈混合密度网络（混合密度网络），

以及循环混合密度网络（循环神经网络混合密度网络） [26, 27]。正如预期，观测噪声会导致使用均方误差训练的模型无法再现吸引子的定性结构，并生成因对不相容的局部动力学取平均而向外漂移的轨迹。前馈混合密度网络能够捕捉局部变异性，但在长时间范围内会累积噪声。相比之下，循环神经网络混合密度网络始终能够生成保留系统真实几何和拓扑特征的轨迹。

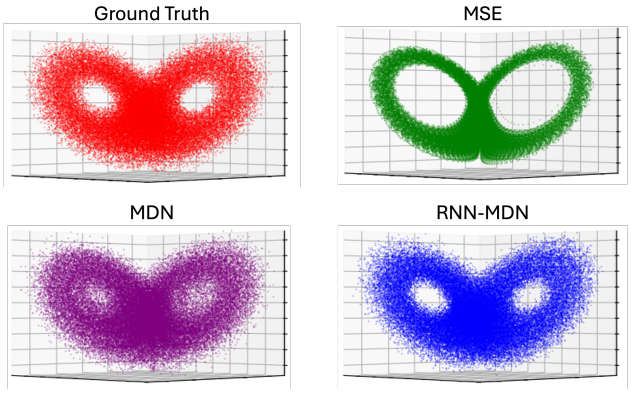

图6：洛伦兹系统中的滚动预测轨迹。只有循环神经网络混合密度网络能够恢复真实吸引子的几何与拓扑。

这些差异在图6中显而易见：只有循环神经网络混合密度网络能够在不产生人为失真的情况下，恢复吸引子特有的双叶结构。从定量角度看，这一表现反映在表1中预测轨迹与真实轨迹之间的最大均值差异（MMD）上，其中循环神经网络混合密度网络的差异最小。值得注意的是，循环神经网络混合密度网络使用的参数（58K）少于均方误差基线模型（67K）。此外，在架构匹配的情况下，前馈混合密度网络的最大均值差异已比均方误差低27倍，这单独体现了混合输出头的贡献。循环神经网络混合密度网络带来的进一步提升表明，显式的多模态输出能够自然地与循环结构相结合。循环神经网络混合密度网络还能够将空间划分为可解释的组成部分，这一点可见图18（附录）。

	MSE	MDN	循环神经网络混合密度网络
最大均值差异 (↓)	2.251e-2	8.104e-4	**3.813e-4**

表1：不同方法对洛伦兹系统真实值数据的点云与自回归预测结果之间的最大均值差异（MMD）。数值越低越好。

这些结果自然意味着，混合密度网络可以与针对具体问题设计的骨干架构有效结合，使动力学模型中基于定义域知识的归纳偏置与显式的多模态不确定性表示相结合，同时不牺牲稳定性或可解释性。

6 结论

总结。 本文重新审视了混合密度网络，认为它是科学机器学习中建模多模态条件不确定性的一种有原则的方法。尽管扩散模型和流匹配等隐式生成模型具有很强的表达能力，但它们往往样本复杂度高，并存在拓扑偏差，因而在科学场景中产生不符合物理规律的伪影。相比之下，混合密度网络的参数化归纳偏置能够在科学机器学习中常见的数据有限情形下，实现统计效率较高的密度估计。其混合结构产生的可解释组件与物理状态和相变自然对应。

局限性。 我们的实验比较以条件流匹配作为代表性的隐式生成模型；包含归一化流、条件VAE以及概率集成模型的全面比较将进一步阐明其中的权衡，但超出了本工作的范围。选择合适的混合分量数量会引入一种建模选择，可能需要验证或领域知识，而固定的参数化分量对于极高维或弱结构化不确定性可能具有限制性。虽然混合密度网络非常适合中等维度的科学输出，但将显式混合模型扩展到大规模、非结构化空间仍是一个开放挑战。这些考量将混合密度网络定位为并非隐式生成模型的通用替代，而是一种互补工具，其优势与许多科学问题的结构高度匹配。

未来方向。 有前景的方向包括将MDN 概率框架与特定架构解耦，并将其与更具表达能力或与领域对齐的骨干网络相结合。RNN-MDN在序列建模中的既有成功表明，显式混合输出能够自然地与结构化网络集成，同时保持可解释性与校准不确定性 [27, 26]。在科学人工智能中，将这一范式扩展到算子学习，为面向偏微分方程的不确定性感知神经代理模型提供了一条路径 [28], ，其中多模态解算子自然地源于分岔与参数歧义。最后，将混合密度网络与Kolmogorov-Arnold网络 [29] 或变体架构 [30], 相结合，可能为低维问题中的可解释性开辟新的机会。

致谢

我们谨此感谢美国能源部先进科学计算研究计划（资助编号 DE-SC0024563）和美国国家科学基金会（NSF）Soft AE 研究培训计划（NRT）（NSF 资助编号 2152205）提供的支持。我们还要感谢促成这项研究的软件开发者，包括 JAX [31], Flax [32] Matplotlib [33]和 NumPy [34]。

代码与数据可用性

我们将整个代码库作为 JaxMix 软件包公开发布，可从 https://github.com/PredictiveIntelligenceLab/JaxMix 获取。此代码库中的 examples 文件夹包含本文所用的所有实验，包括图表的生成。我们希望其他计算科学家也能发现此代码库有用，并利用它在各自的问题上开展多模态建模实验。

影响声明

我们的贡献推动了深度学习和不确定性量化的发展。这有望影响广泛的下游应用。虽然我们预计这项工作不会带来具体的负面影响，但与任何强大的预测工具一样，它也存在被滥用的可能性。我们鼓励研究界在敏感定义域中应用这些技术时，考虑其伦理影响和潜在的两用场景，并完全避免将其应用于武器及其他军事技术。

作者贡献

L.F.G. 和 P.P. 构思了方法并对实验进行了概念设计。L.F.G. 和 A.K. 开展了实验并分析了结果。P.P. 提供了资金并负责监督本研究。所有作者均参与了稿件的撰写和审阅。

参考文献

[1] Jonas Mikhaeil、Zahra Monfared 和 Daniel Durstewitz。论使用循环神经网络学习混沌动力学的难度。神经信息处理系统进展，35：11297–11312，2022。[2] Ming Li、Xue-Song Yan 和 Ming-Zhao Zhang。基于神经网络的地震反演综述。地球科学信息学，16(4)：2991–3021，2023。[3] Yang Song、Liyue Shen、Lei Xing 和 Stefano Ermon。使用基于分数的生成模型解决医学成像中的逆问题，2022。[4] Pavel Izmailov、Sharad Vikram、Matthew D Hoffman 和 Andrew Gordon Gordon Wilson。贝叶斯神经网络后验分布究竟是什么样的？载于国际机器学习会议，第4629–4640页。PMLR，2021。[5] Andrew Gordon Wilson 和 Pavel Izmailov。深度集成作为近似贝叶斯推断。https: //cims.nyu.edu/~andrewgw/deepensembles/，2021。[6] Diederik P Kingma 和 Max Welling。自编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*，2013。[7] Jonathan Ho、Ajay Jain 和 Pieter Abbeel。去噪扩散概率模型。神经信息处理系统进展，33：6840–6851，2020。

[8] Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matt Le。用于生成建模的流匹配。arXiv 预印本 arXiv:2210.02747，2022。[9] Frederic Koehler、Alexander Heckett 和 Andrej Risteski。分数匹配的统计效率：等周性视角。载于国际学习表征会议（ICLR），2023。[10] Giulio Biroli、Tony Bonnaire、Valentin De Bortoli 和 Marc Mézard。扩散模型的动力学状态。《自然·通讯》，15(1)：9957，2024。[11] Christopher M Bishop。混合密度网络。神经计算研究组报告，1994。[12] Carl E. Rasmussen 和 Christopher K. I. Williams。《机器学习中的高斯过程》。MIT出版社，2006。[13] Kihyuk Sohn、Honglak Lee 和 Xinchen Yan。使用深度条件生成模型学习结构化输出表征。神经信息处理系统进展，28，2015。[14] Aad W Van der Vaart。《渐近统计》，第3卷。剑桥大学出版社，2000。[15] Hassan Ashtiani、Shai Ben-David、Nicholas Harvey、Christopher Liaw、Abbas Mehrabian 和 Yaniv Plan。通过样本压缩方案学习高斯混合模型的近乎紧致样本复杂度界。载于神经信息处理系统进展（NeurIPS），2018。[16] Ananda Theertha Suresh、Alon Orlitsky、Jayadev Acharya 和 Ashkan Jafarpour。球面高斯混合模型的近乎最优样本估计量。载于神经信息处理系统进展（NIPS），2014。[17] Larry Wasserman。《非参数统计大全》。Springer Science & Business Media，2006。[18] Minshuo Chen、Kaixuan Huang、Tuo Zhao 和 Mengdi Wang。低维数据上扩散模型的得分近似、估计与分布恢复。载于国际机器学习会议，第4672–4712页。PMLR，2023。[19] Kyriakos Flouris 和 Ender Konukoglu。用于流形学习的规范化归一化流。载于神经信息处理系统进展（NeurIPS），2023。[20] Rob Cornish、Anthony Caterini、George Deligiannidis 和 Arnaud Doucet。利用连续索引归一化流放宽双射约束。载于 Hal Daumé III 和 Aarti Singh（编），第37届国际机器学习会议论文集，机器学习研究论文集第119卷，第2133–2143页。PMLR，2020年7月13–18日。[21] Michael I Jordan 和 Robert A Jacobs。分层专家混合模型与EM算法。神经计算，6(2)：181–214，1994。[22] Yanqi Zhou、Tao Lei、Hanxiao Liu、Nan Du、Yanping Huang、Vincent Zhao、Andrew M Dai、Quoc V Le、James Laudon 等。采用专家选择路由的混合专家模型。神经信息处理系统进展，35：7103–7114，2022。[23] Fleur Hendriks、Ondˇrej Rokoš、Martin Došká˘r、Marc G. D. Geers 和 Vlado Menkovski。针对对称性破缺分岔问题的等变流匹配，2025。[24] Yorgos M Psarellis、Themistoklis P Sapsis 和 Ioannis G Kevrekidis。用于寻找分岔的主动搜索。《混沌：非线性科学跨学科期刊》，35(5)，2025。[25] Ed ward N. Lorenz。确定性非周期流。《大气科学杂志》，20，1963。[26] David Ha 和 Jürgen Schmidhuber。循环世界模型有助于策略演化。神经信息处理系统进展，31，2018。[27] David Ha 和 Douglas Eck。草图绘画的神经表征。arXiv 预印本 arXiv:1704.03477，2017。[28] Akshay Thakur 和 Souvik Chakraborty。Md-NOMAD：用于模拟随机微分方程及不确定性传播的混合密度非线性流形解码器。arXiv 预印本 arXiv:2404.15731，2024。[29] Ziming Liu、Yixuan Wang、Sachin Vaidya、Fabian Ruehle、James Halverson、Marin Solja˘ci´c、Thomas Y. Hou 和 Max Tegmark。KAN：Kolmogorov-Arnold 网络，2024。[30] Leonardo Ferreira Guilhoto 和 Paris Perdikaris。Kolmogorov叠加定理的深度学习替代方案。载于第十三届国际学习表征会议（ICLR），2025。[31] James Bradbury、Roy Frostig、Peter Hawkins、Matthew James Johnson、Chris Leary、Dougal Maclaurin、George Necula、Adam Paszke、Jake VanderPlas、Skye Wanderman-Milne 和 Qiao Zhang。JAX：Python+NumPy 程序的可组合变换，2018。[32] Jonathan Heek、Anselm Levskaya、Avital Oliver、Marvin Ritter、Bertrand Rondepierre、Andreas Steiner 和 Marc van Zee。Flax：适用于 JAX 的神经网络库及生态系统，2023。

[33] J. D. Hunter。《Matplotlib：一个二维图形环境》。《计算机在科学与工程中的应用》，9(3):90–95，2007年。[34] Charles R. Harris、K. Jarrod Millman、Stéfan J. van der Walt、Ralf Gommers、Pauli Virtanen、David Cournapeau、Eric Wieser、Julian Taylor、Sebastian Berg、Nathaniel J. Smith、Robert Kern、Matti Picus、Stephan Hoyer、Marten H. van Kerkwijk、Matthew Brett、Allan Haldane、Jaime Fernández del Río、Mark Wiebe、Pearu Peterson、Pierre Gérard-Marchant、Kevin Sheppard、Tyler Reddy、Warren Weckesser、Hameer Abbasi、Christoph Gohlke 和 Travis E. Oliphant。使用 NumPy 进行数组编程。《自然》，585(7825):357–362，2020年9月。[35] Yihui He 和 Jianren Wang。用于概率目标检测的深度混合密度网络。载于：2020年IEEE/RSJ智能机器人与系统国际会议（IROS），第10550–10555页。IEEE，2020年。[36] Ali Varamesh 和 Tinne Tuytelaars。用于目标检测和人体姿态估计的混合密集回归。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第13086–13095页，2020年。[37] Jaeseok Choi、Yeji Song、Yerim Kim、Jaeyoung Yoo 和 Nojun Kwak。MD3D：基于混合密度的点云三维目标检测。《IEEE Access》，10:104011–104022，2022年。[38] Hexuan Li、Tarik Kanuric 和 Arno Eichberger。基于混合密度网络的虚拟仿真汽车雷达建模。《IEEE传感器期刊》，23(11):11117–11124，2022年。[39] Rohit Unni、Kan Yao 和 Yuebing Zheng。用于层状光子结构逆向设计的深度卷积混合密度网络。《ACS光子学》，7(10):2703–2712，2020年。[40] Christopher N Davis、T Deirdre Hollingsworth、Quentin Caudron 和 Michael A Irvine。混合密度网络在复杂流行病学个体模型仿真中的应用。《PLoS计算生物学》，16(3):e1006869，2020年。[41] Fabio Tosi、Yiyi Liao、Carolin Schmitt 和 Andreas Geiger。SMD-Nets：立体混合密度网络。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第8942–8952页，2021年6月。[42] Jacob H. Seidman、Georgios Kissas、Paris Perdikaris 和 George J. Pappas。NOMAD：用于算子学习的非线性流形解码器，2022年。[43] Djork-Arné Clevert、Thomas Unterthiner 和 Sepp Hochreiter。利用指数线性单元（ELU）实现快速、准确的深度网络学习，2016年。[44] David Ha。教程：使用 JAX 的混合密度网络。https://github.com/hardmaru/mdn_jax_tutorial/blob/master/mixture_density_networks_jax.ipynb，2020年。[45] Hanzhong Guo、Cheng Lu、Fan Bao、Tianyu Pang、Shuicheng Yan、Chao Du 和 Chongxuan Li。扩散模型的高斯混合求解器。载于：神经信息处理系统进展（NeurIPS），2023年。[46] Alexis V. Kudryashov、Alexey L. Rukosuev、Alexander N. Nikitin、Ilya V. Galaktionov 和 Julia V. Sheldakova。实时 1.5 kHz 自适应光学系统，用于校正大气湍流。Opt. Express，28(25):37546–37552，2020年12月。[47] Ian Osband、Zheng Wen、Mohammad Asghari、Morteza Ibrahimi、Xiyuan Lu 和 Benjamin Van Roy。认知神经网络。CoRR，abs/2107.08924，2021年。[48] Xilu Wang、Yaochu Jin、Sebastian Schmitt 和 Markus Olhofer。贝叶斯优化的最新进展，2022年。[49] Burr Settles。主动学习文献综述。计算机科学技术报告1648，威斯康星大学麦迪逊分校，2009年。[50] Leonardo Ferreira Guilhoto 和 Paris Perdikaris。利用 NEON（神经认知算子网络）在函数空间中进行复合贝叶斯优化。《科学报告》，14(1):29199，2024年。[51] Ethan Pickering、Stephen Guth、George Em Karniadakis 和 Themistoklis P. Sapsis。通过神经算子中的主动学习发现并预测极端事件。《自然·计算科学》，2(12):823–833，2022年12月。[52] Samuel Kim、Peter Y Lu、Charlotte Loh、Jamie Smith、Jasper Snoek 和 Marin Solja˘ci´c。利用深度学习对具有高维结构的科学问题进行贝叶斯优化。《机器学习研究汇刊》，2022年。[53] DeepMind、Igor Babuschkin、Kate Baumli、Alison Bell、Surya Bhupatiraju、Jake Bruce、Peter Buchlovsky、David Budden、Trevor Cai、Aidan Clark、Ivo Danihelka、Antoine Dedieu、Claudio Fantacci、Jonathan Godwin、Chris Jones、Ross Hemsley、Tom Hennigan、Matteo Hessel、Shaobo Hou、Steven Kapturowski、Thomas Keck、Iurii Kemaev、Michael King、Markus Kunesch、Lena Martens、Hamza Merzic、Vladimir Mikulik、Tamara Norman、George Papamakarios、John Quan、Roman Ring、Francisco Ruiz、Alvaro Sanchez、Laurent Sartran、Rosalia Schneider、Eren Sezener、Stephen Spencer、Srivatsan Srinivasan、Miloš Stanojevi´c、Wojciech Stokowiec、Luyu Wang、Guangyao Zhou 和 Fabio Viola。DeepMind JAX生态系统，2020年。

[54] Dan Hendrycks 和 Kevin Gimpel。高斯误差线性单元（GeLU），2023年。[55] Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。*arXiv* 预印本 *arXiv:1711.05101*，2017年。[56] Andrew Brock、Soham De、Samuel L. Smith 和 Karen Simonyan。无需归一化的高性能大规模图像识别。*CoRR*，abs/2102.06171，2021年。[57] Fleur Hendriks、Vlado Menkovski、Martin Doškář、Marc GD Geers、Kevin Verbeek 和 Ondˇrej Rokoš。基于壁纸群的力学超材料：包含力学响应的数据集。*Scientific Data*，2025年。[58] Patrick Kidger。论神经微分方程。博士论文，牛津大学，2021年。[59] John R Dormand 和 Peter J Prince。嵌入式龙格-库塔公式族。*Journal ofcomputational andapplied mathematics*，6(1):19–26，1980年。[60] Junyoung Chung、Caglar Gulcehre、KyungHyun Cho 和 Yoshua Bengio。门控循环神经网络在序列建模中的实证评估。*arXiv* 预印本 *arXiv:1412.3555*，2014年。[61] Jacob H Seidman、Georgios Kissas、George J Pappas 和 Paris Perdikaris。变分自编码神经算子。载于：第*40*届国际机器学习会议论文集，第30491–30522页，2023年。

数学符号表示

表 2 总结了本文使用的符号和符号表示。

表 2：本文使用的符号和符号表示总结。

符号	含义
$\boldsymbol{x}$	一个向量 $\boldsymbol{x} = (x_1, \ldots\ldots, x_d)$ 在 $\mathbb{R}^d$
d	通常用于表示 向量的维度
$\mathcal{D}$	一个由输入/输出对组成的数据集 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ 输入/输出对
N	通常用于表示大小 数据集的
θ	神经网络的参数
p_θ	一个概率分布，其 可学习参数 θ
$N(\mu, \Sigma)$	一个均值为.....的正态分布 μ 和协方差 Σ
$\mathcal{N}(y; \mu, \Sigma)$	正态分布的概率密度函数 均值为 μ 且协方差为 Σ 在点 y处求值
K	混合成分的数量 在混合密度网络中
α_k	第 k个混合的权重 混合密度网络的元素
$\mathcal{L}$	用于训练神经网络模型的 损失函数
$\boldsymbol{I}_d$	大小为 d的单位矩阵
$\mathbb{E}$	期望运算数
对数和指数$_k$	“对数和指数”运算 对某个索引 k

B 面向科学机器学习的混合密度网络

近年来，混合密度网络已应用于深度学习的多个领域，包括图像分割和目标检测任务 [35, 36, 37], ，以及时间序列数据相关任务，例如生成逼真的手写文本 [27]。在 [38],中，混合密度网络用于对汽车雷达数据建模，以促进自动驾驶；在 [39], 中，混合密度网络用于光子结构的逆向设计；在 [40]中，混合密度网络用于流行病学建模。尽管有这些应用案例，混合密度网络在自然科学中用于物理系统建模的情况仍然很少。
在阐明了混合密度网络在低维、多模态科学任务中的理论优势后，我们将介绍与计算科学家相关的具体实现细节。

B.1 架构

我们采用标准的混合密度网络架构，其中骨干网络可以是适用于输入数据类型的任意神经网络（例如，用于向量输入的多层感知机、用于图像/场的卷积神经网络 [35, 36, 37, 41], 用于时间序列的循环神经网络/长短期记忆网络 [27, 26], 或用于算子学习的 NOMAD [28, 42]）。关键的修改在于输出头：对于一个目标值 $y \in \mathbb{R}^d$ 和由 K 个高斯分布组成的混合，网络输出 $K(1+d+d)$ 个参数（假设为对角协方差）或 $K(1+d+d(d+1)/2)$（完整协方差）。尽管预测完整协方差矩阵能够提供更强的表达能力，但在实践中，大多数混合密度网络实现（包括本文所使用的实现）都采用对角协方差。

网络输出三个向量：对数几率 $z_\alpha \in \mathbb{R}^K$、均值 $z_\mu \in \mathbb{R}^d$ 和尺度 $z_\sigma \in \mathbb{R}^d$，随后将它们分别（可能经过变换）转换为关注量的权重、均值和方差。

$$\text{Logits } z_\alpha \mapsto \alpha_k = \frac{\exp(z_{\alpha,k})}{\sum_j \exp(z_{\alpha,j})} \quad (\texttt{Softmax}), \tag{6}$$

$$\text{Means } z_\mu \mapsto \mu_k = z_{\mu,k} \quad (\text{Identity}), \tag{7}$$

$$\text{Scales } z_\sigma \mapsto \sigma_k = \log\left(1 + \exp(z_{\sigma,k})\right) + \epsilon \quad (\texttt{Softplus}). \tag{8}$$

使用 Softplus 激活函数并加上一个小的艾普西隆 $\epsilon > 0$ 来设置尺度参数 σ，优于使用指数函数，因为这样既能严格保证参数为正，又能避免数值爆炸（e^x 的梯度可能不稳定）。另一种方法是使用 $z_\sigma \mapsto 1 + ELU(z_\sigma) + \epsilon$，其中 ELU 是指数线性单元 [43]。

B.2 混合密度网络损失的稳定数值计算

给定目标值 $y \in \mathbb{R}^d$，混合密度网络所建模的条件似然为

$$p(y \mid x) = \sum_{k=1}^{K} \alpha_k(x)\, \mathcal{N}\big(y \mid \mu_k(x), \sigma_k^2(x)\big)\,, \tag{9}$$

训练通过最小化负对数似然（NLL）进行。

直接计算该表达式在数值上不稳定，因为其中涉及指数运算和可能很小的方差。因此，我们使用 LogSumExp 技巧计算负对数似然。定义

$$F_k(x,y) = z_{\alpha,k}(x) - \texttt{LogSumExp}_j(z_{\alpha,j}(x)) - \log\left(\mathcal{N}(y|\mu_k(x), \Sigma_k(x)\right)$$

单个观测 $(x,\ y)$ 的损失可简洁地写为

$$\mathcal{L}(x,y) = -\texttt{LogSumExp}_k\left(F_k(x,y)\right). \tag{10}$$

此公式改编自 [44], ，可避免对混合权重进行显式归一化，并防止在计算低概率分量时发生数值下溢。在实践中，基于对数和指数的实现对于稳定的混合密度网络训练至关重要，尤其是在混合权重极不均衡或分量方差较小时。

对数和指数运算可通过以下方式稳定计算：

$$\texttt{LogSumExp}_k\left(a_k\right) := \max_k(a_k) + \log\left(\sum_k \exp\left(a_k - \max(a_k)\right)\right).$$

B.3 解读潜在结构

混合密度网络的一项显著优势是混合权重 $\alpha_k(x)$ 具有可解释性。在科学情境中，这些权重实际上能够学习到类似于系统相图的内容。

标量场 $\alpha_k(x)$ 充当序参量。 $\alpha_k(x) \approx 1$ 的区域对应于稳定相。权重从一个分量转移到另一个分量的过渡区域（例如 $0 \ll \alpha_1 \approx \alpha_2 \ll 1$）可确定分界曲线或相变。通过计算混合权重的熵 $H(\alpha(x))$，可以自动检测相共存区域。

在隐式生成模型中，无法直接访问模型的内部“信念状态”，只能通过代价高昂的蒙特卡罗采样或常微分方程迹估计来获取概率密度。

B.4 推理延迟与实时控制

最后，在实时控制等应用中，推断速度至关重要。MDN 的推断复杂度为 $O(1)$，因为神经网络只需进行一次前向传播，就能得到条件分布的解析参数（$\alpha,\ \mu,\ \Sigma$）。另一方面，DDPM 和 CFM 需要进行 $O(T)$ 次前向传播，并求解常微分方程/随机微分方程，通常需要$T = 20$ 至 $T = 100$ 次神经网络的连续评估。即使采用近期的“蒸馏”求解器， [45],扩散模型仍然明显慢于 MDN。例如，在以 1.5 kHz 运行的低延迟控制回路中， [46],MDN 往往是唯一可行的多模态概率深度学习方案。

B.5 混合成分数量的消融实验

选择合适的骨干架构后，MDN 引入了一个需要设定的新超参数：混合分量的数量 K。实验中我们发现，采取谨慎做法、添加多于严格必要数量的混合成分，通常弊大于利的情况很少，因为与骨干网络相比，增加的计算成本微不足道，而且 MDN 通常会学会不激活不必要的混合成分。由于实践中很难事先知道描述一个系统需要多少个分量，因此采取谨慎做法可能更有利。

在图7中，我们绘制了主文中所述反正弦问题的测试负对数似然，针对使用大小为 $N = 1000$ 的数据集、包含$K = 1, 2, \ldots, 16$ 个混合成分的混合密度网络。如图所示，当成分数量足以描述系统中的不确定性后（本例中约为 5 个），继续增加K带来的收益会逐渐减少，但不会导致性能退化。

使用过多成分的一个潜在缺点是，多个混合成分可能表示相同的众数，从而造成歧义并影响可解释性。例如，在图8中可以看到，混合成分 2 与混合成分 4 和 6 重叠。在我们希望将定义域整齐地划分为“影响区域”的情形下，这可能会带来挑战。然而，更多时候，不必要的成分只会被赋予零权重或非常小的权重，如图8中的混合成分 8 所示。此外，在对混合模型进行过度参数化时，还可以通过舍弃权重低于较小阈值的成分（例如 0.01 ），在推理时进一步抑制伪影，避免从概率极低的混合分布中产生伪样本 [1]。有关此消融实验的完整说明（包括具体的混合成分）可见于补充材料中的 mixture_component_ablation.ipynb 文件。

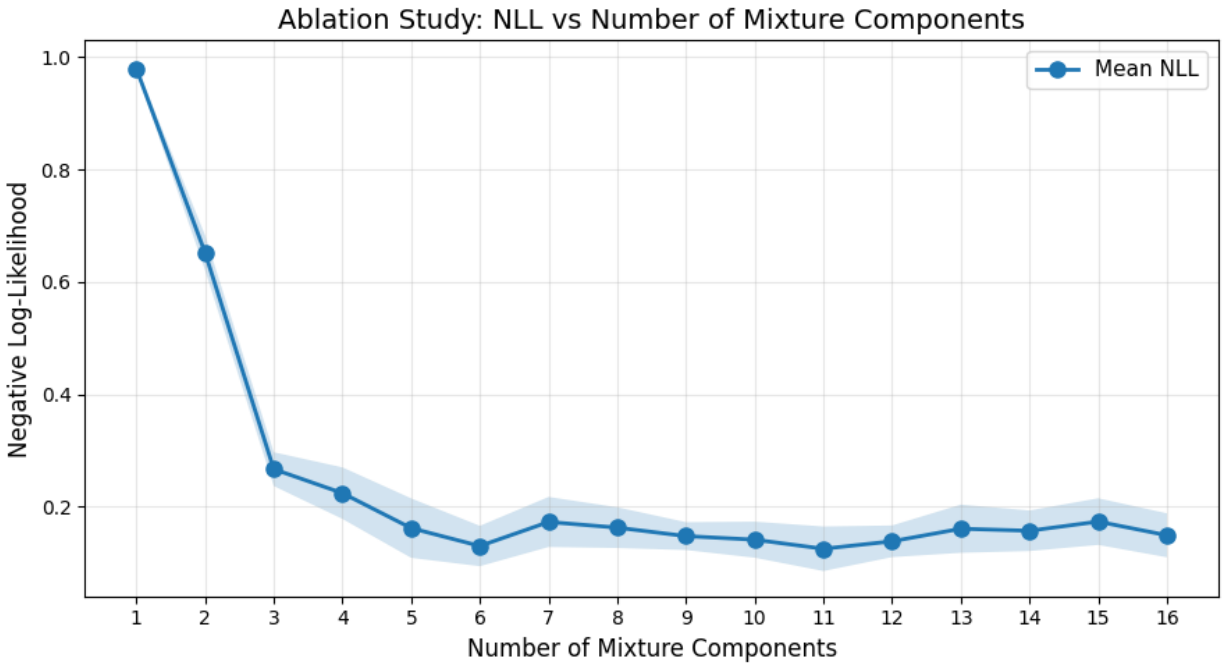

图7：反正弦问题中不同混合密度网络混合成分数量对应的测试负对数似然。 ± 1 标准差带根据 12 种不同的网络初始化计算得出。

C 科学机器学习中的偶然不确定性与认知不确定性

科学机器学习中的不确定性通常分解为偶然不确定性和认知不确定性两部分。偶然不确定性源于系统或测量过程的内在变异性，即使拥有无限数据也依然存在。典型来源包括传感器噪声、随机强迫，或真实存在的多值物理响应，本文正文中讨论的就是此类响应。相比之下，认知不确定性反映了对

[1] 此功能已在 通过 restrict_rare_event 使用我们的 JaxMix 软件包 _速率选项。

由于数据有限以及模型设定错误而产生的底层映射误差，原则上可以通过增加观测值或改进建模来降低。本文的贡献重点在于对偶然不确定性进行建模。
混合密度网络（MDN）主要针对偶然不确定性，而认知不确定性量化则可通过集成模型、贝叶斯神经网络等互补技术，或 EpiNets [47] 等轻量级方法来实现。在实践中，区分这两种不确定性来源对于科学决策至关重要，因为它们对实验设计、模型改进和物理解释有着不同的影响。

认知不确定性在自适应数据获取策略中也发挥着核心作用，例如贝叶斯优化 [48]和主动学习 [49], 。这些策略旨在通过有针对性的采样，高效降低模型不确定性。在科学机器学习中，认知不确定性能够揭示现有数据尚未充分探索的输入空间区域，从而指引我们找到最值得开展新模拟或实验的地方。通过明确区分偶然不确定性和认知不确定性，模型可以避免对本质上具有随机性的区域过度采样，而将数据获取重点放在那些新增数据能够显著提升预测准确度的区域。这一区分在成本高昂的科学场景中尤为重要，因为要高效地实现科学发现，必须审慎权衡探索与利用；近年来已有多篇论文致力于推动这一领域的发展 [50, 51, 52]。

D 拓扑约束：微分同胚与断裂

流匹配和基于常微分方程的扩散模型是以**微分同胚**的形式构建的——即具有可微性和可逆性的映射。微分拓扑的一个基本定理指出，微分同胚会保持拓扑不变量，包括连通分支的数量 [19, 20]。

$$\text{Topology}(\mathcal{Z}) \cong \text{Topology}(\mathcal{Y}), \quad \text{under diffeomorphism } \phi. \tag{11}$$

当基分布 $p(z)$ 是单个连通的高斯分布，而目标物理分布 $p(y|x)$ 由两个或更多不连通的峰构成时，这就造成了一个深刻的问题。连续流无法将单个高斯团块撕裂成两个独立的团块，除非在向量场中引入无限大的梯度或奇点。为了弥补这一点，流模型总会生成连接各个模态的细小“细丝”或“桥接区域”，使其具有非零概率。这些桥接区域可能对应物理上不允许的状态（例如，预测粒子存在于势垒内部）。要近似这种“撕裂”，学习到的向量场必须具有极高的利普希茨常数，导致刚性常微分方程在积分时速度缓慢且不稳定。实验部分的图3展示了这一现象：条件流匹配难以在概率景观中学习不连通模态。

另一方面，混合密度网络在输出空间中不受拓扑结构限制。混合分布通过对离散核求和构建而成。当控制参数 x 跨越决策边界时，门控网络（用于预测分量逻辑值）可以通过调节权重 α，平滑地（$K = 1 \to K = 2$）将活动分量从一个快速切换为两个。这种准离散的“切换”能力与本文所述问题的离散本质完美契合。

E 分数匹配中的概率等周问题

扩散模型在多模态情境下一个关键且常被忽视的理论局限，是**概率等周问题**。我们用这一术语指代学习具有不连通模态的分布时所面临的困难。这一问题已在 [9], 中得到广泛研究，该研究对分数匹配的统计效率进行了严格分析。他们的工作证明，分数匹配的样本复杂度与目标分布的等周常数成反比。

考虑一个包含两个模态 A 和 B 的分布，它们被一个概率密度可忽略不计的区域（“深能井”）隔开。如果分数函数 $\nabla \log p(y)$ 是训练所用的信号，那么在 A 和 B之间的空白区域中，密度接近零，分数也趋于平坦或无法定义。因此，要学习 A 与 B 的相对质量（例如， A 的可能性为 90%， B 的可能性为 10%），扩散过程必须在噪声注入阶段让样本在这些模态之间混合。如果模态相距很远，或者数据集较小，扩散轨迹就很少能有效穿越低密度屏障，因而无法将全局混合比例传递给分数网络。因此，基于分数的模型可以较好地学习模态的形状，但在估计其相对权重时会出现指数级的失效。这也适用于峰之间的低概率区域；正如上一节进一步讨论的那样，DDPM 和 CFM 等方法会高估这些区域的似然。
相比之下，混合密度网络直接最小化负对数似然： $\mathcal{L} = -\sum \log(\sum \alpha_k \mathcal{N})$。这个目标函数是全局性的。如果数据点出现在众数 B中，当 α_B 较小时，损失函数会对模型施加很大的惩罚，从而立即直接地更新混合系数。混合密度网络能够更有效地通过权重 α_k“传送”概率质量，无需穿越低密度区域。

F 实验细节

F.1 软件

我们的代码使用 JAX [31]实现，并借助 Flax [32] 和 Optax [53] 库来定义和训练神经网络。若论文被接收，我们计划在 GitHub 上发布完整代码库。在此期间，我们已将代码作为 ZIP 文件附在本次投稿的补充材料中。

F.2 硬件

本文中的每项实验均使用单个 NVIDIA A6000 GPU 完成。

F.3 在流匹配中计算负对数似然

为了计算 CFM 的条件负对数似然，我们不得不对流常微分方程进行反向积分，同时沿着这条路径计算流的散度，具体方法如原始 CFM 论文的附录 C 所述 [8]。这一过程计算开销很大，可能需要数百次神经网络评估。相比之下，混合密度网络只需一次前向传播即可给出完整的概率分布。

要计算训练好的 CFM 模型的条件负对数似然，我们通过从 $t = 1$ 处的数据分布出发，沿时间反向求解概率流常微分方程，直至 $t = 0$ 处的先验噪声分布，从而计算数据样本 x 的对数概率。该过程使用瞬时变量替换公式，需要沿样本的轨迹对向量场的散度进行积分。最终的对数似然由结果噪声样本 x_0 在先验分布 p_0 下的对数概率、估计得到的累积散度，以及用于校正任何数据缩放或预处理变换的对数行列式项共同确定。

不同于原论文使用 Hutchinson 迹估计器近似散度，我们直接计算散度。与使用近似方法相比，这样计算更加准确（且具有确定性）；鉴于本文聚焦于低维问题，这种做法也是可行的。

F.4 正弦逆问题

系统描述。如正文所述，我们使用了由 $y = f(x) + \epsilon$ 定义的正向过程，其中 $f\colon\ \mathbb{R} \to \mathbb{R}$ 由 $f(x) = \frac{x}{2} + 0.7\sin(5x)$ 给出，且 $\epsilon \sim N(0,\ \sigma^2)$ 满足 $\sigma = 0.2$。随后，我们希望计算在给定某个观测 y 的情况下，x 的条件概率分布 $p(x|y)$。

真实逆似然。尽管逆条件似然 $p(x|y)$ 没有闭式解（据我们所知），但可以使用贝叶斯定理进行数值计算：

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{12}$$

$$= \frac{\mathcal{N}(y; f(x), 0.2^2) \cdot \frac{1}{3}}{\int_{-1.5}^{1.5} p(y|x')p(x')dx'} \tag{13}$$

$$= \frac{\mathcal{N}(y; f(x), 0.2^2)}{\int_{-1.5}^{1.5} \mathcal{N}(y; f(x'), 0.2^2)dx'}. \tag{14}$$

我们使用上述公式生成了图2中的似然和负对数似然（NLL）图。

数据生成。 我们生成了大小各异的训练数据集 $N \in \{50, 100, 200, 500, 1000, 2000, 5000, 10000\}$。输入从区间 $[-1.5, 1.5]$ 中均匀采样。所有情况下，我们都使用了一个包含 1000 个样本的独立测试数据集。

模型架构。 我们比较了三种模型架构：混合密度网络（MDN）、条件流匹配（CFM）以及采用均方误差（MSE）训练的标准基线模型。所有方法都使用相同的骨干架构，即由 5 个宽度为 128、采用 GeLU 激活函数的全连接层组成 [54]。MDN 模型共使用 $K = 8$ 个混合分量。为稳健地估计负对数似然（NLL）并量化初始化带来的方差，我们针对每个架构与数据集的组合训练了一个包含 12 个模型的集成。

训练。 所有方法均使用 AdamW [55] 训练，共进行 30,000 次迭代。MDN 模型采用如下学习率调度：先进行 100 步线性预热，达到峰值学习率 $5e-3$，随后每 1,000 步按 0.9 的比例指数衰减。为使需要较低学习率的 CFM 获得更好的结果，我们采用 $1,000$ 步线性预热，达到峰值学习率 $1e-3$，随后采用与 MDN 相同的指数衰减调度。

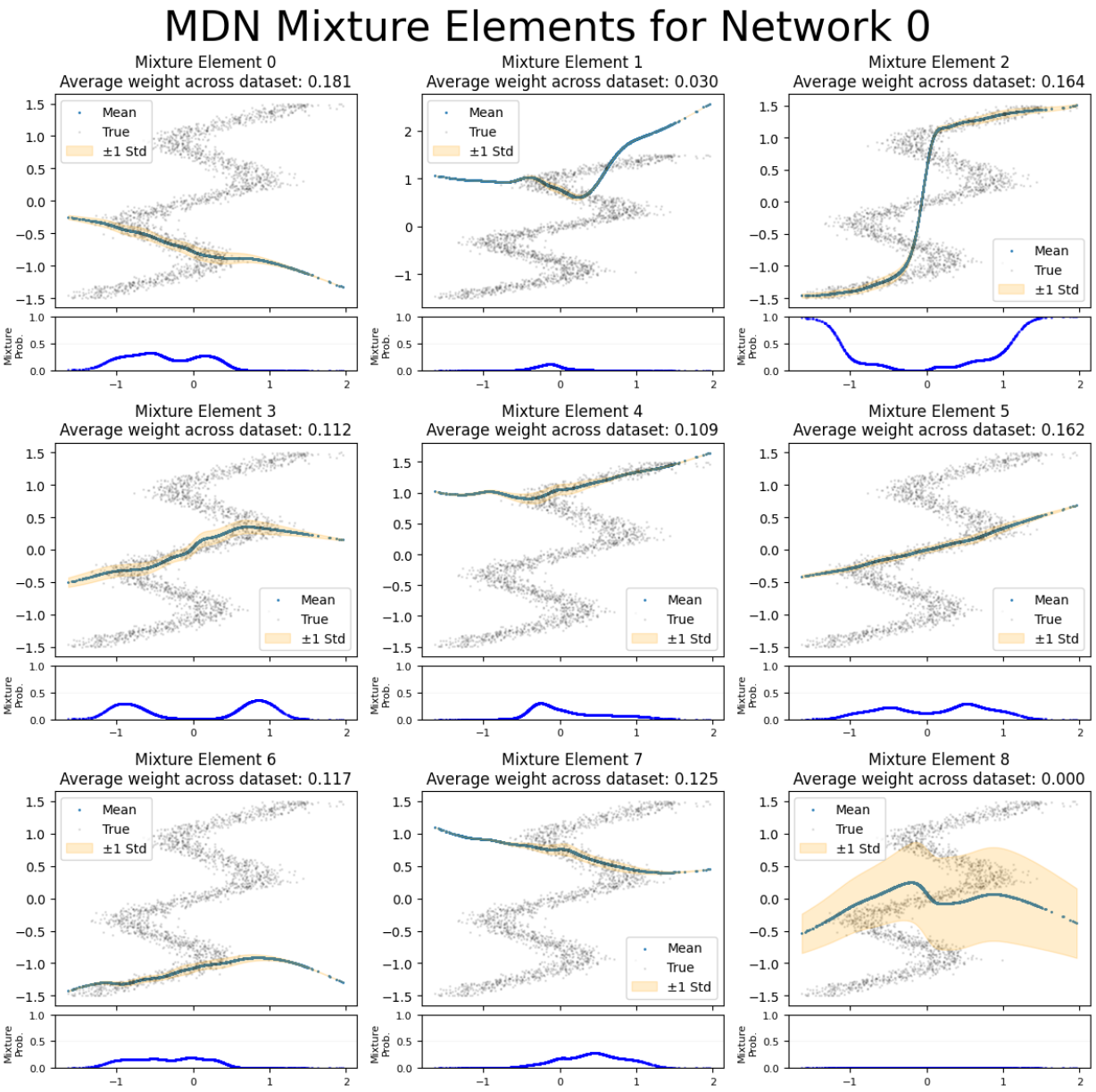

图8：反正弦基准测试中学习到的混合分量示例，包含 $K=9$ 个混合成分。每个子图显示每个混合的均值和预测标准差带，下方较小的图则显示给定输入（x轴）时的条件混合概率。

结果。 如正文所述，混合密度网络在低数据情形下显著优于条件流匹配。即使在高数据设置下，条件流匹配仍会高估罕见事件的似然。在图9中，我们用来自图3的结果加以扩展，补充了NLL曲面图以及 $-log(p(x|y=0.5))$的切片，从而更容易看出条件流匹配的过度平滑倾向。为完整起见，我们还将图3和图9顶部图中绘制的数值报告于表3。

N （样本）	混合密度网络负对数似然（↓）	条件流匹配负对数似然（↓）
50	2.727 ± 1.753	16.709 ± 7.379
100	2.478 ± 1.469	5.409 ± 1.265
200	0.723 ± 0.462	1.360 ± 0.227
500	0.235 ± 0.074	0.124 ± 0.010
1000	0.059 ± 0.068	-0.002 ± 0.003
2000	-0.017 ± 0.006	-0.036 ± 0.002
5000	-0.042 ± 0.003	-0.040 ± 0.002
10000	-0.040 ± 0.003	-0.049 ± 0.001

表3：图9上方图中展示的正弦逆问题里，混合密度网络与条件流匹配的数据效率对比。

网络规模消融实验。作为网络规模消融实验，我们使用仅含3层、每层宽度为64的较小骨干网络，重复了图3所示的数据效率实验。如图10所示，在这种情形下，我们的发现与较大网络的结果一致：在低数据情形下，混合密度网络的表现优于条件流匹配。对于这种网络配置，即使使用更大的数据集（$N = 10,\ 000$），混合密度网络在负对数似然方面仍然优于条件流匹配。此外，第4节所述的“桥状”拓扑伪影也同样出现，条件流匹配仍然会高估低概率区域的似然。为求完整，我们还在表4中报告了图10上方图中绘制的数值。

N （样本）	混合密度网络负对数似然（↓）	条件流匹配负对数似然（↓）
50	1.548 ± 0.445	3.006 ± 1.311
100	0.487 ± 0.192	0.499 ± 0.052
200	0.244 ± 0.106	0.246 ± 0.032
500	0.071 ± 0.011	0.057 ± 0.009
1000	0.010 ± 0.037	0.037 ± 0.015
2000	-0.016 ± 0.002	0.038 ± 0.021
5000	-0.048 ± 0.006	-0.003 ± 0.003
10000	-0.043 ± 0.003	0.003 ± 0.007

表 4：使用较小的骨干网络，在图 10 顶部图中展示的反正弦问题上，对混合密度网络和条件流匹配的数据效率进行比较。

点估计的失效。 在所有数据区间内，均方误差基线的表现都极其糟糕。正如预期，它学习的是条件期望 $\mathbb{E}[x|y]$，对应于条件样本的算术均值，如图 12 所示。该均值通常位于概率密度接近于零的区域。这凸显了逆问题中概率建模的必要性。

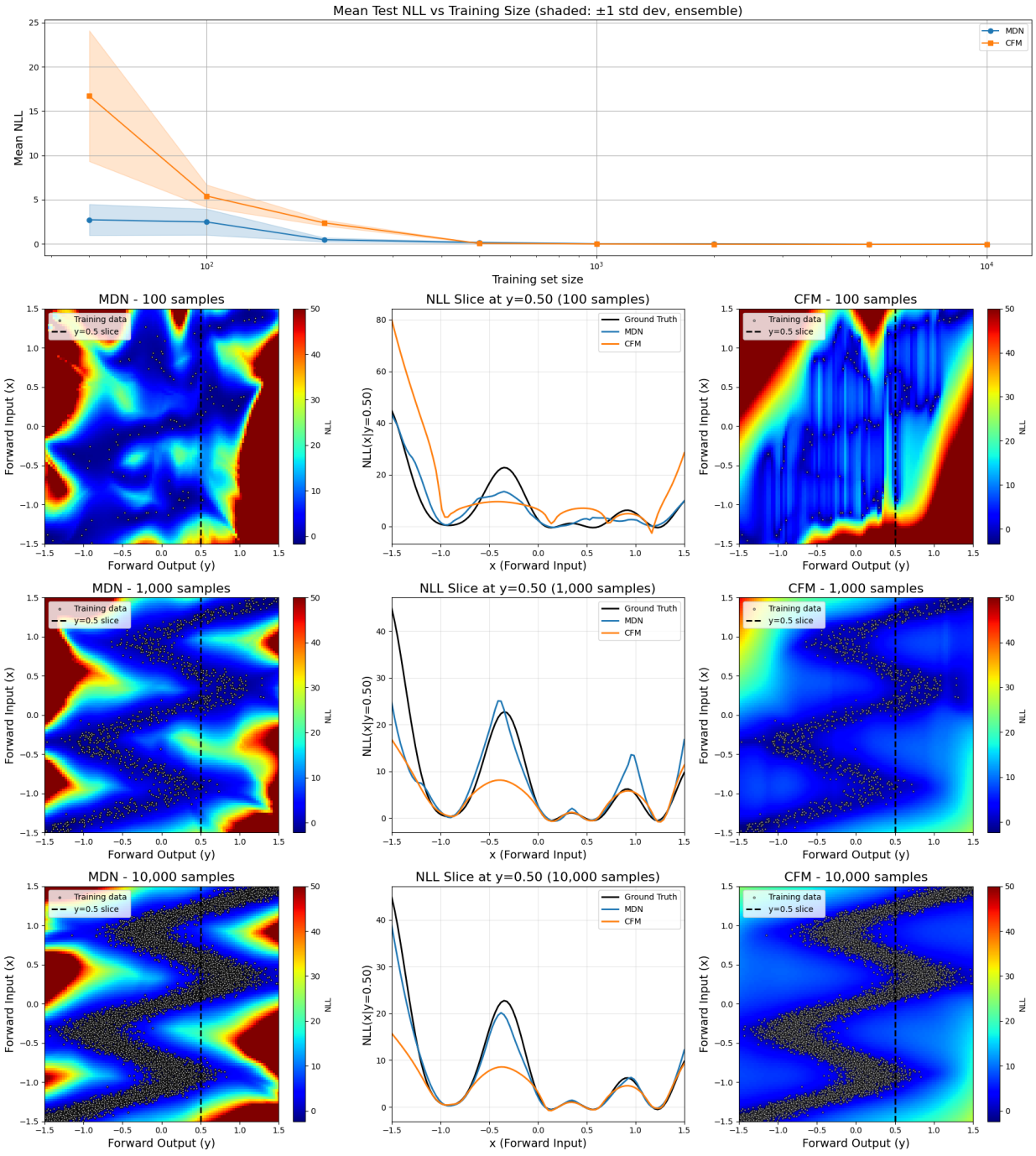

图9：混合密度网络与条件流匹配在正弦逆问题中的比较。上方图表比较了两种方法在不同训练集规模下的测试负对数似然（越低越好）。标准差带对应于12个不同随机初始种子的结果。其余图表展示了部分实验中的负对数似然曲面（对随机种子取平均）。

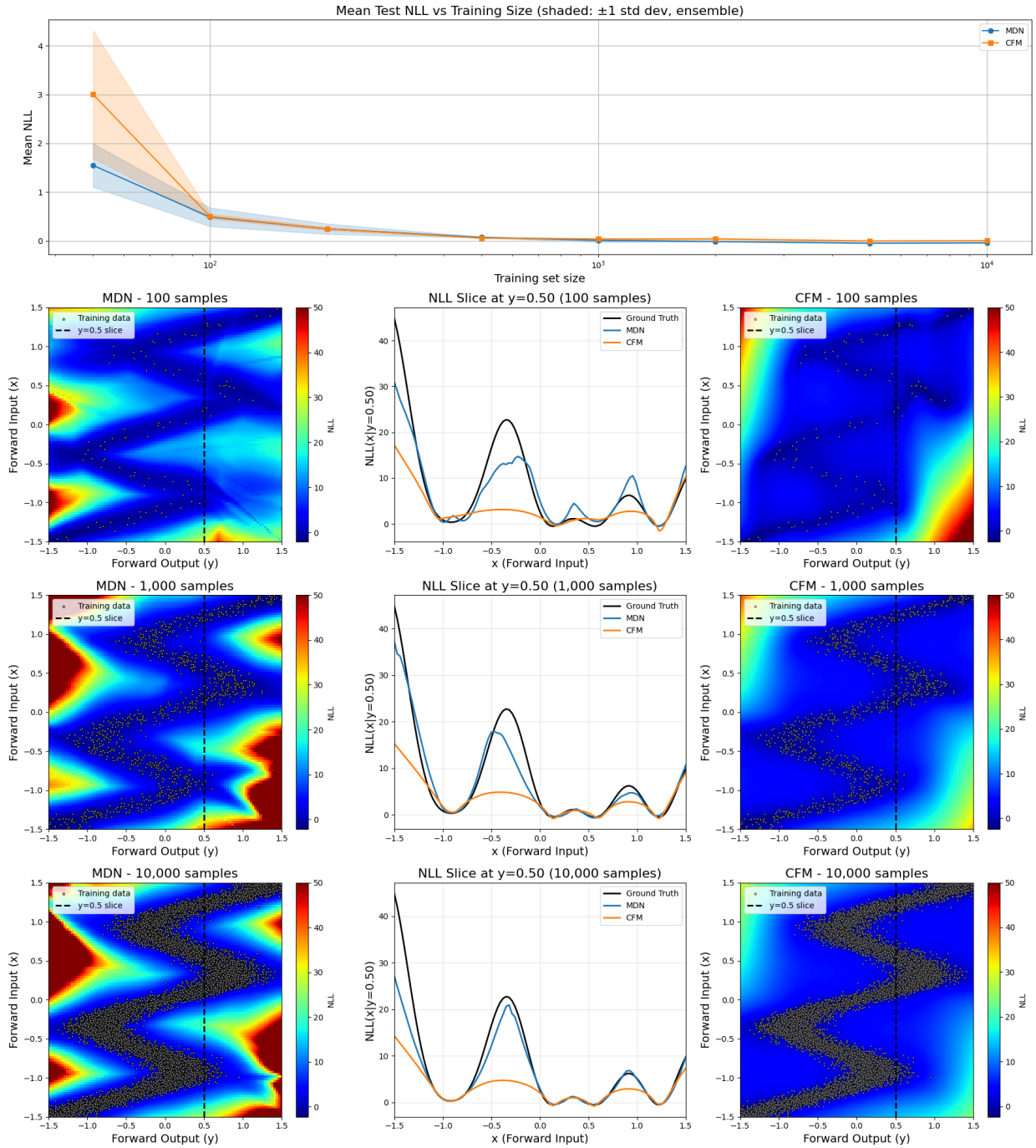

图10：混合密度网络与条件流匹配方法的数据效率比较，与图9类似，但使用了规模更小的骨干网络。

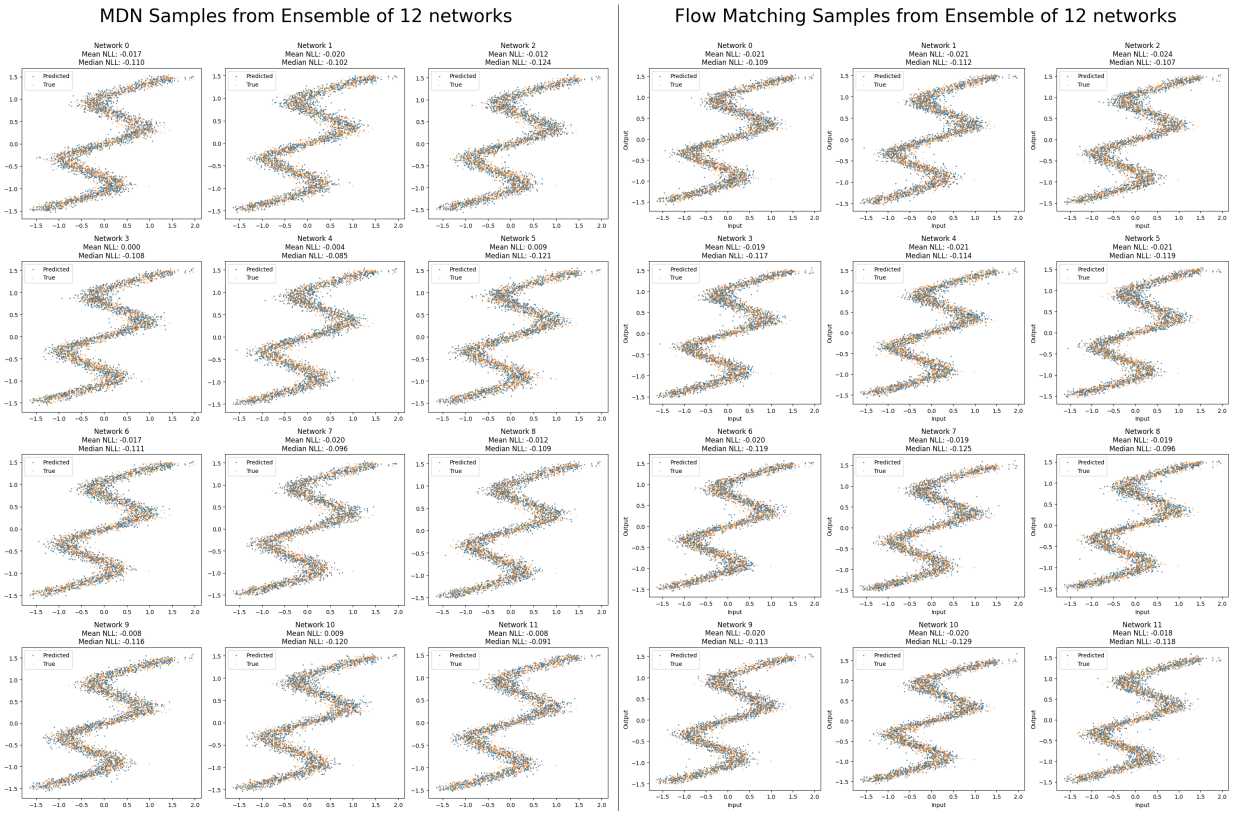

图 11：使用由12个网络组成的集成模型，在反正弦问题的条件后验分布中得到的代表性样本。（左）混合密度网络。（右）条件流匹配。

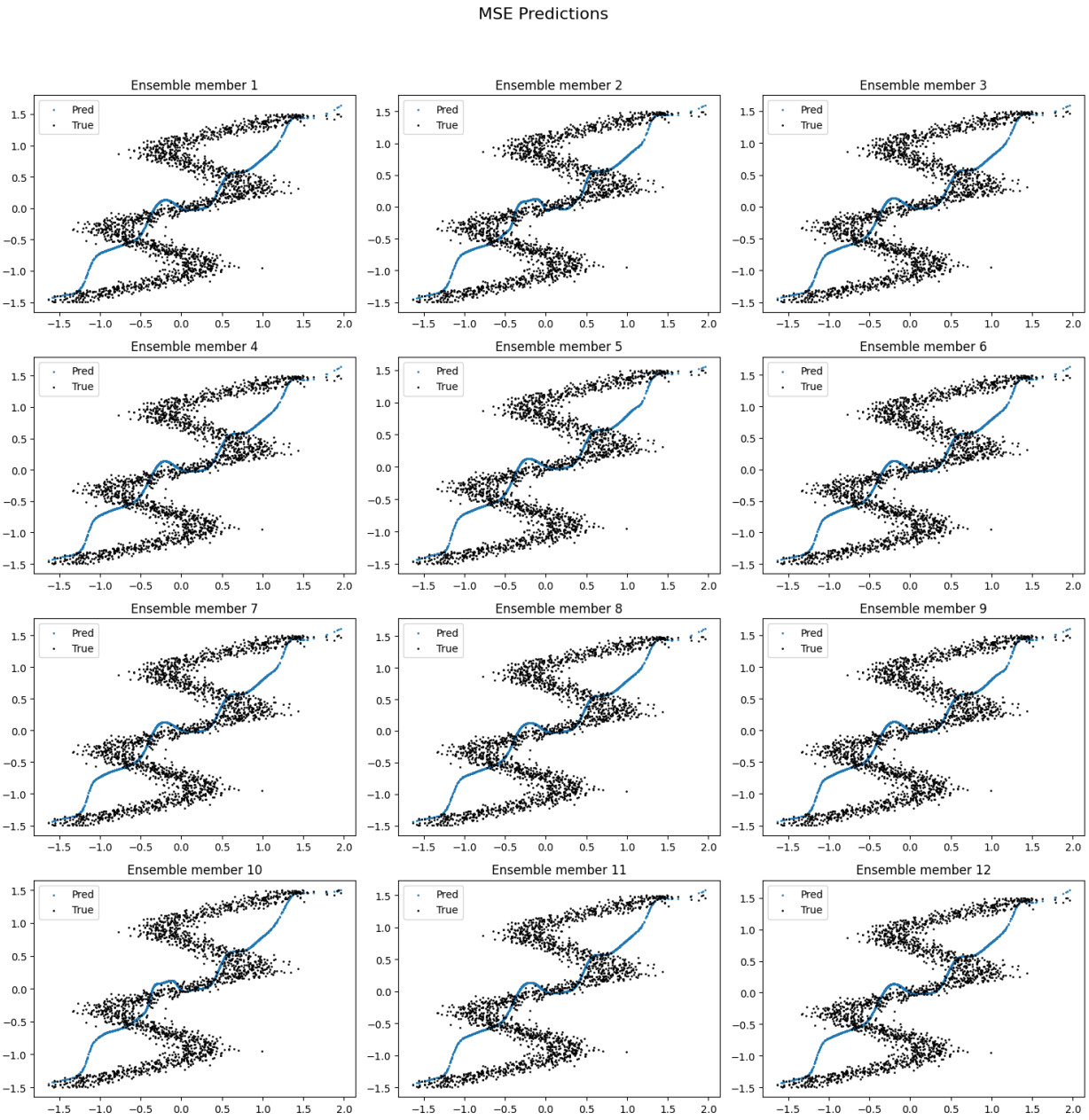

图 12：在反正弦问题上使用均方误差训练的网络集成的点预测。由于只预测均值，该网络未能捕捉数据集的真实结构。

F.5 引力吸引子

系统动力学。 我们模拟一个粒子在最多三个静止天体的引力作用下，在二维平面上的运动轨迹。这些天体位于固定位置 $\mathbf{c}_1$,$\mathbf{c}_2$,$\mathbf{c}_3$。位置向量 $\mathbf{r}$ 的运动方程如下：

$$\ddot{\mathbf{r}} = -\sum_{j=1}^{3} \frac{GM_j(\mathbf{r}-\mathbf{c}_j)}{||\mathbf{r}-\mathbf{c}_j||^2} - \gamma\dot{\mathbf{r}}^2, \tag{15}$$

其中，γ 表示阻尼系数（阻力），取值为 $\gamma = 10$。我们将引力常数设为 $G = 100$，并将每个天体的质量设为 $M_j = 1$。吸引子的位置在单位圆上均匀分布，粒子的初始位置则在以原点为圆心、半径为 0.2 的圆盘内均匀随机采样。由于存在阻尼项，系统是耗散的；每条轨迹最终都会失去能量，并渐近收敛到三个天体之一的位置，这些天体充当稳定吸引子。我们还添加了少量各向同性观测噪声，其服从均值为 0、标准差为 0.05 的正态分布。

所考虑情况的描述。 如正文所述，我们考虑该系统的三种变体。图 13 展示了每种情况下的示例轨迹。

1. **无输入噪声的部分可观测系统：** （吸引子位置有多种可能）：在此情景中，每条轨迹都会随机选取一个吸引子，并且系统演化时只使用这一个吸引子。这就构成了一个部分可观测性问题，我们需要不同的世界模型来涵盖每种可能情况。在这种情况下，可能的“世界”总数等于可能的吸引子数量，并且我们预计每个“活跃”混合都会被选中，且与初始位置无关。
2. **带输入噪声的一致系统**：在此情景中，控制系统是一致的（三个吸引子始终存在），但数据中的输入受到观测噪声的干扰。因此，靠近哪个吸引子牵引力最强这一决策边界的点可能并不可靠。因此，我们预计每个混合元素在远离该决策边界时会更加确定（更接近 1 或 0），而在原点附近则更加模糊。

3. **无输入噪声的一致系统：** （混合密度网络实际上充当专家混合模型）：在这种情况下，三个吸引子始终都存在，数据中的输入也都是干净的。在这种设置下，混合密度网络方法似乎没有太大优势，但正如我们所见，该方法依然有效，并且实际上充当了专家混合模型：对于给定的输入，我们以接近 1 的概率选取一个特定的混合，该混合专门处理平面上特定切片中的点。

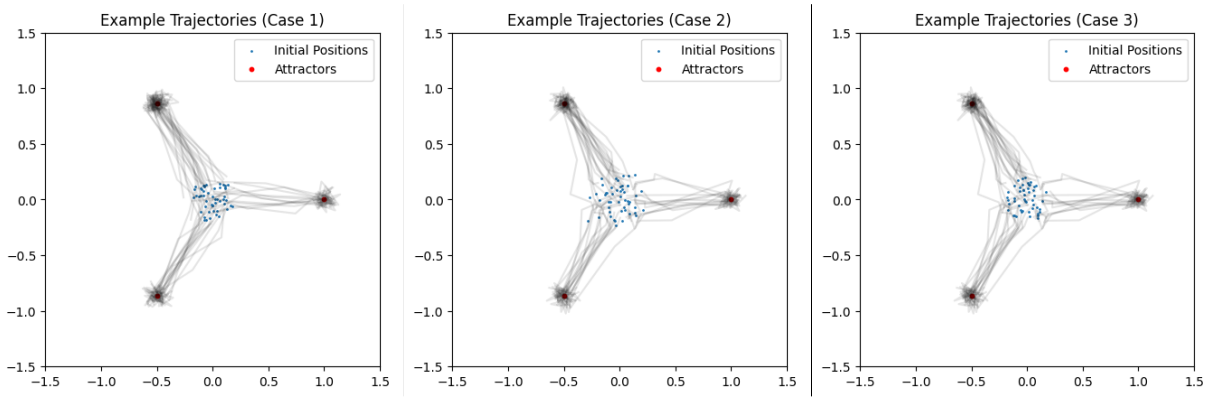

图 13：引力吸引子示例中所考虑的三种不同情形下的轨迹示例。尽管这些轨迹在定性上非常相似，但每个系统从根本上都与其他系统不同。然而，混合密度网络能够以人类可解释的方式学习轨迹对输入的依赖关系，从而揭示问题的结构，如图 4 所示。

数据生成。 粒子的真实轨迹采用简单的欧拉积分方案，从 $t = 0$ 计算至 $t = 1$，步长为 $dt = 0.0005$，从而得到2,000个时间步。在训练和推断中，我们将轨迹均匀下采样为仅11个时间步 $t \in \{0.0, 0.1, 0.2, \ldots, 1.0\}$。

模型架构。 混合密度网络和条件流匹配使用了相同的骨干架构，即四个维度为128的隐藏层和GeLU激活，其中混合密度网络使用10个混合分量。为控制网络初始化带来的变异性，我们针对每种方法训练了12 个网络的集成。

训练。 混合密度网络和条件流匹配均使用AdamW训练30,000步，权重衰减为0.01，批量大小为64。学习率调度为：经500步线性预热达到峰值学习率5e-4，随后每1,000步以0.9的指数衰减率衰减，并在5e-4处趋于平稳。我们还额外使用了自适应梯度裁剪 [56] ，速率为0.01。

与隐式模型的比较。我们使用条件流匹配模型进行了相同的实验。如表 5 所示，在测试负对数似然方面，混合密度网络始终优于条件流匹配。虽然条件流匹配也能准确预测结果的边缘分布（生成的样本聚集在三个吸引子周围），但它没有提供直接可视化吸引盆结构的方法。这一结果印证了第 4 节中的观点：混合密度网络的参数化结构不仅是计算上的便利，更是一种语义特征，与多稳态物理系统的离散本质相契合；它能以“白盒”方式揭示隐式模型所掩盖的相拓扑。

	混合密度网络负对数似然（↓）	条件流匹配负对数似然（↓）
案例1	$\mathbf{-32.596 \pm 0.059}$	-30.583 ± 0.096
案例2	$\mathbf{-28.445 \pm 0.184}$	-27.504 ± 0.083
案例3	$\mathbf{-32.840 \pm 0.088}$	-30.718 ± 0.177

表5：测试 引力吸引子问题中所考察各案例下混合密度网络和条件流匹配的负对数似然。

F.6 常微分方程鞍结分岔

在动力系统中，当动力学系数发生微小变化却导致系统行为发生剧烈改变时，就会出现分岔。如果该系数未知，且在不同数据集成员之间可能有所不同（对于欠定系统而言），那么准确的不确定性量化对于确定关注量可能的轨迹至关重要。这一直是近期科学机器学习论文关注的重点，例如 [23, 57, 24]。我们展示了常微分方程中的一个此类系统示例，其中，多模态分布对于捕捉潜在的不确定性并给出有用的预测必不可少。

系统描述。鞍结分岔是动力系统中不动点产生和消失的一种基本机制。我们考虑如下常微分方程：

$$\frac{dx}{dt} = r + x^2, \tag{16}$$

其中， $r \in \mathbb{R}$ 是分岔参数。该系统的定性行为取决于 r的符号：

- **稳定区间（$r < 0$）：** 存在两个不动点，分别位于 $x^* = \pm\sqrt{-r}$。位于 $x^* = -\sqrt{-r}$ 的不动点是稳定的（吸引的），而 $x^* = +\sqrt{-r}$ 是不稳定的（排斥的）。从不稳定不动点下方（$x_0 < -\sqrt{-r}$）出发的轨迹会收敛到稳定平衡点，而从其上方（$x_0 > -\sqrt{-r}$）出发的轨迹则会发散至无穷远。
- **临界点（$r = 0$）：** 两个不动点在原点处碰撞并湮灭，发生鞍结分岔，此时存在一个半稳定不动点。
- **不稳定区间（$r > 0$）：** 不存在不动点，所有轨迹都会发散至无穷大。然而，分岔的影响仍以临界减速的形式残留：轨迹经过 $x \approx 0$附近的瓶颈区域时会减速。

数据生成。我们使用步长为 $\Delta t = 0.001$ 的前向欧拉格式对方程 (16) 进行数值积分，总时长为 $T = 5.0$，以此生成训练数据。随后从轨迹中抽取 20 个时间步。初始条件x_0 从 $[-2.0, 2.0]$中均匀采样，分岔参数 r 从 $[-1.5, 1.5]$中均匀分布的 5 个值组成的离散集合$r \in \{-1.50, -0.75, 0.00, 0.75, 1.50\}$ 中采样。向轨迹添加标准差为$\sigma = 0.1$ 的观测噪声。关键的是， r 被视为隐变量：模型仅接收 x_0 作为输入，必须预测完整的离散化轨迹 $x(t_1)$ $x(t_2) \ldots, x(T)$。此外，对于趋于无穷大的轨迹，我们将位置截断在 $x = 10$ ，以避免表示任意大的数值。

测试集包含 $N_{\text{test}} = 2000$ 个样本，并在所有实验中保持固定。

模型架构。 MDN 和 CFM 模型均采用由五层多层感知机组成的骨干架构，每层包含 256 个单元，并使用 GeLU 激活函数 [54]。为捕捉随机种子的变异性，两种方法均通过 12 个独立初始化的集成成员进行训练。MDN 使用 $K = 15$ 个混合分量。

训练。 两个模型均使用 AdamW 优化器训练 50,000 次迭代，批量大小为 128，权重衰减为 10^{-5} 。学习率采用如下调度：先经过 2,000 步线性预热，升至峰值 5×10^{-4}，随后每 2,000 步持续按指数衰减至原来的 0.9。我们还使用了自适应梯度裁剪[56] ，裁剪率为 0.1。

样本效率结果。 表6报告了不同训练集大小下的测试负对数似然（12 个集成成员的均值 $\pm$ 标准差）。在 50–100 个样本时，MDN 的负对数似然优于 CFM；而 CFM 需要约 200+ 个样本才能得出明显更好的估计。在 $N \geq 5000$ 时，两种方法的性能趋于相近。

成功的多模态预测：混合密度网络 vs. 条件流匹配。 图14和图15比较了混合密度网络和条件流匹配各自全部12个集成成员的预测结果。两种方法都成功捕捉到鞍结分岔的多模态特征，准确识别出不同的收敛（$r < 0$）和发散（$r > 0$）动力学状态。混合密度网络集成（图14）在所有成员中都呈现出一致的多模态行为，各众数彼此分离良好；条件流匹配集成（图15）也捕捉到了双峰结构。

失效模式：均方误差点估计。图16展示了使用均方误差损失训练的点估计网络的根本失效模式。所有12个集成成员预测的轨迹都聚集在条件均值附近，而该均值位于两种物理上可实现的结果之间。这种“平均”行为产生的预测

表6：鞍结分岔系统的样本效率比较。负对数似然值以12个集成成员的均值 $\pm$ 标准差报告。数值越低越好。

N（样本）	混合密度网络负对数似然（↓）	条件流匹配负对数似然（↓）
50	87.3 ± 7.2	665.9 ± 139.9
100	77.4 ± 3.5	146.0 ± 21.9
200	50.6 ± 13.1	16.1 ± 2.0
500	3.4 ± 1.4	-9.5 ± 0.2
1000	-5.5 ± 1.8	-12.8 ± 0.1
2000	-10.9 ± 1.3	-13.5 ± 0.1
5000	-13.4 ± 0.9	-13.9 ± 0.1
10000	-13.9 ± 0.3	-14.0 ± 0.1

既不对应收敛状态，也不对应发散状态，实际上预测出了一条在真实动力系统中永远不会观测到的轨迹。均方误差预测紧密聚集在均值附近，而非真实模态附近，这表明点估计方法不适用于具有多模态不确定性的问题。

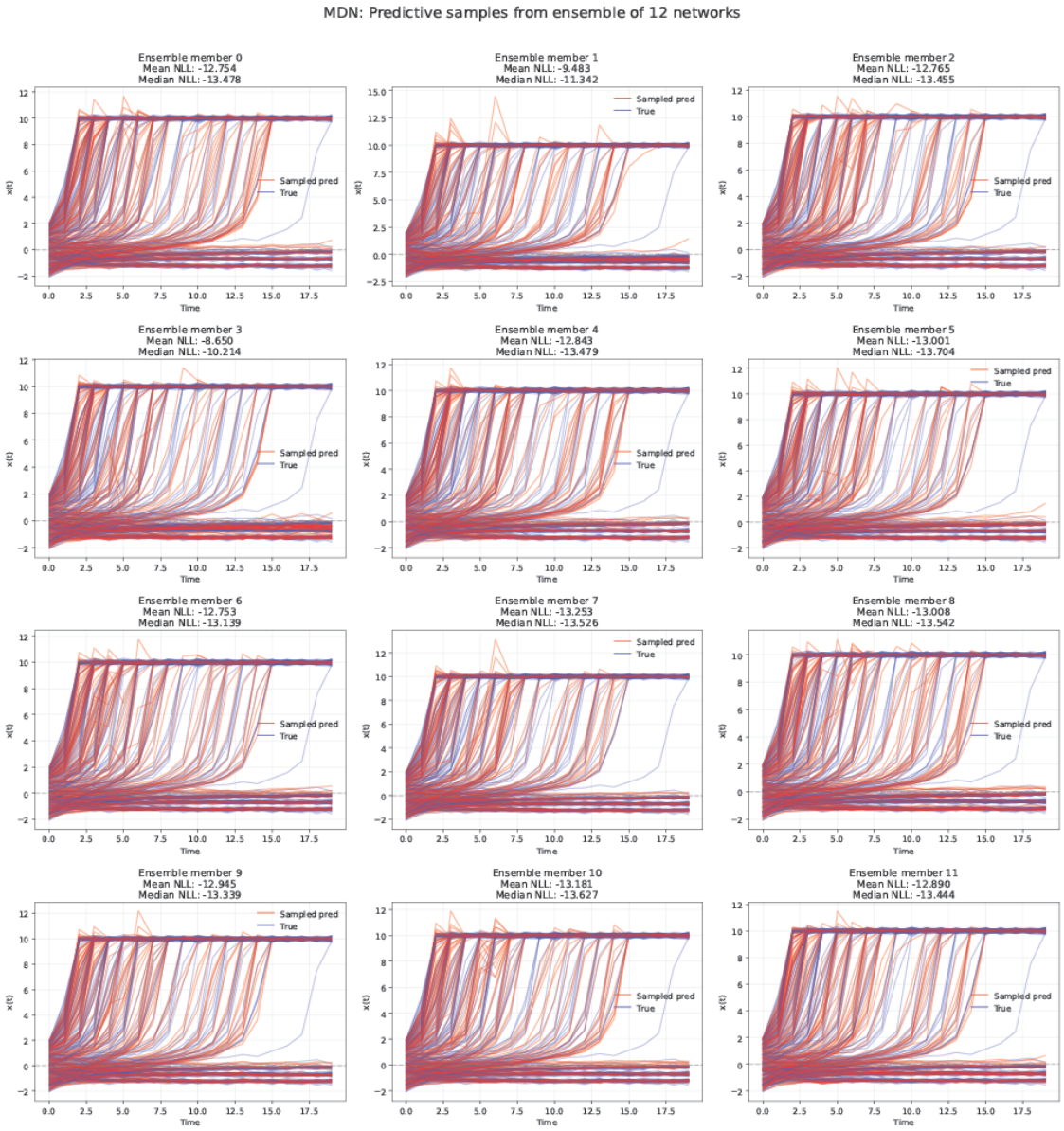

图14：混合密度网络在鞍结分岔系统上的集成预测。每个子图展示了同一组初始条件下，单个集成成员的预测分布样本。所有12个集成成员都呈现出一致的多模态结构，收敛轨迹与发散轨迹之间界限清晰。

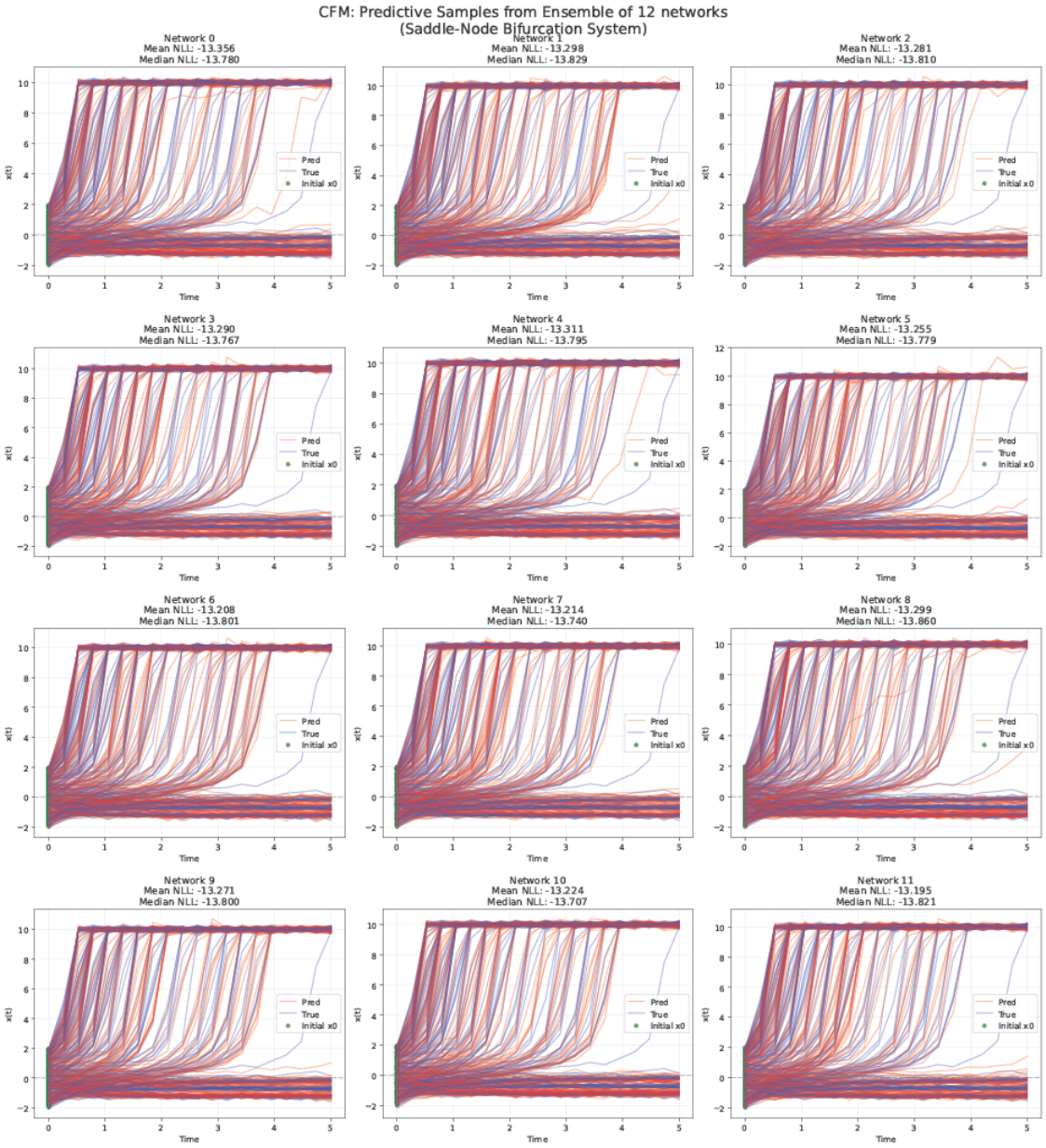

图1：鞍结分岔系统上的条件流匹配集成预测。每个子图显示同一组初始条件下，单个集成成员的预测分布样本。条件流匹配集成同样能够捕捉鞍结动力学中的双峰性。

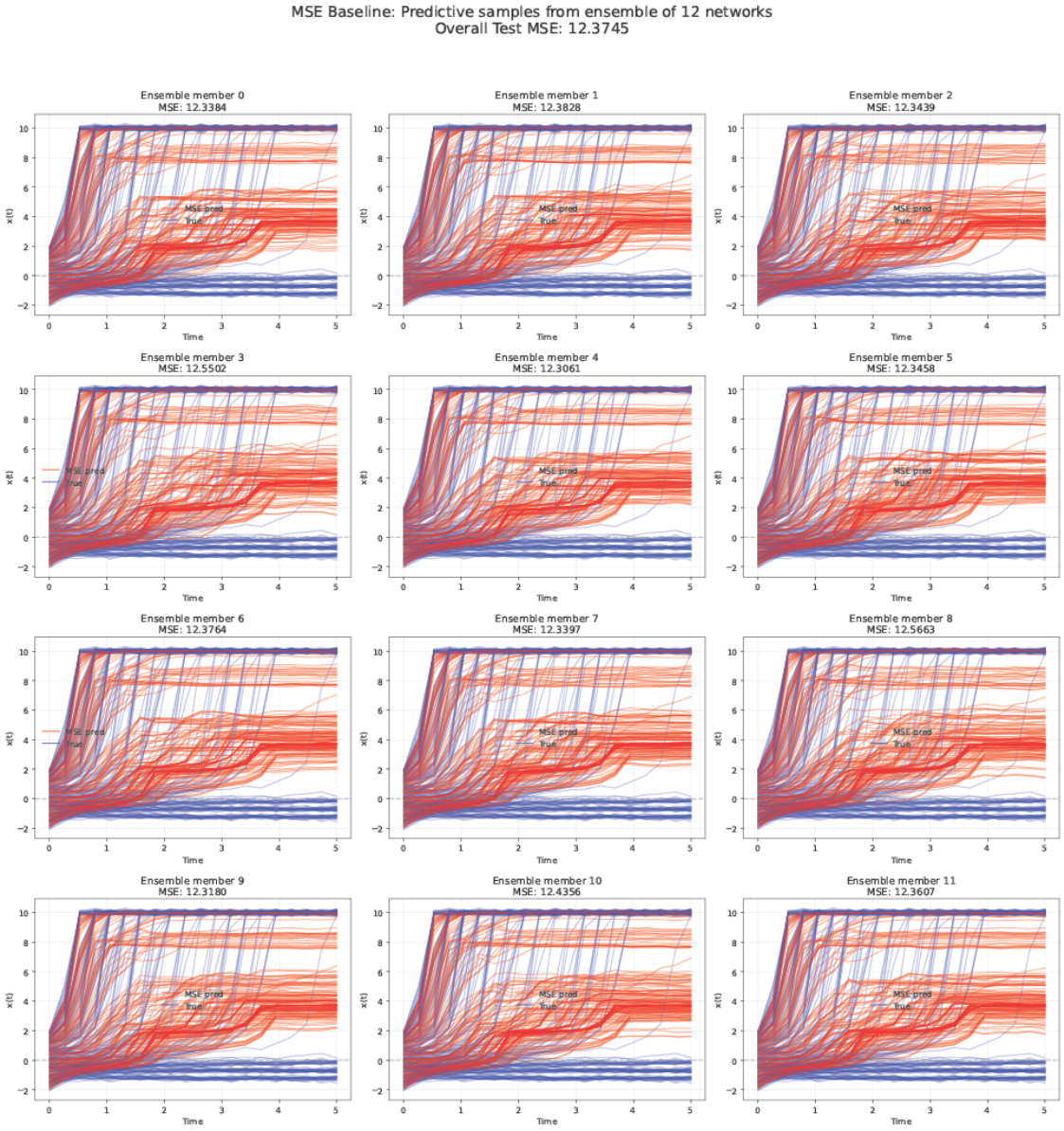

图16：在鞍结分岔系统上使用均方误差训练的集成模型的失效模式。每个子图展示一个集成成员的确定性预测。与混合密度网络和条件流匹配不同，所有集成成员的预测都趋于相似，集中在条件均值附近；该均值位于两个物理状态之间的低概率密度区域。预测轨迹对应于相互排斥的结果的平均值，并不代表任何物理可实现状态。

F.7 洛伦兹系统

系统描述。我们考虑洛伦兹-63 系统 [25]，其三维位置为（x， y， z），演化方程如下：

$$\dot{x} = \sigma(y - x), \tag{17}$$
$$\dot{y} = x(\rho - z) - y, \tag{18}$$
$$\dot{z} = xy - \beta z, \tag{19}$$

采用标准参数 $\sigma = 10$， $\beta = 8/3$， $\rho = 28$。该系统的特征是其主导李雅普诺夫指数为正 $\lambda_1 \approx 0.906$，这意味着初始状态中的任何无穷小误差都会随时间呈指数增长，如 $\|\delta(t)\| \approx \|\delta_0\| e^{\lambda_1 t}$ 所示。对于较长的预测时域（$t \gg 1/\lambda_1$），不可能进行逐点预测。因此，在机器学习语境下，对这个问题做出可信预测的形式不是单条轨迹，而是可能轨迹上的概率密度。

数据生成。用于训练模型的轨迹由 Diffrax 软件包生成 [58]，采用 Dormand–Prince (DOPRI) 5/4 求解器 [59]，时间步长为 $dt = 0.001$，时间范围从 $t = 0$ 到 $t = 10$，每条轨迹因此包含 10，000 个观测值。训练和推断时，我们对轨迹进行均匀下采样，使每条轨迹仅包含 500 个观测值，从而得到 $\Delta t = 0.02$ 的时间间隔。初始位置从方差为 1、均值为（0，0，24。5）的正态分布中随机采样 $\in \mathbb{R}^3$，以使粒子位于动力系统的轨迹流形附近，并尽量减少初始模拟步骤的预热过程。我们还添加少量各向同性观测噪声，该噪声服从均值为 0、标准差为 0.2 的正态分布。

模型架构。 MSE 和普通 MDN 模型使用相同的骨干架构：5 个隐藏层，每层宽度为 128，并采用 GeLU 激活函数；MDN 使用 9 个混合。MSE 和 MDN 的总参数量分别为 66,947 和 74,687。RNN-MDN 由单个 GRU 层 [60]组成，隐藏维度为 128，并包含 8 个混合分量，总参数量为 58,040。值得注意的是，RNN-MDN 以最少的参数（58K，而 MSE 为 67K）实现了更优的性能，这表明其优势源于概率输出结构，而非更大的模型容量。

训练。 所有三个模型均使用 AdamW 训练，权重衰减为 0.1，共迭代 30,000 次。学习率先经过 1,000 次迭代的线性预热，升至峰值学习率 $1e-3$，随后每 1,000 步按 0.9 的速率指数衰减。均方误差和混合密度网络模型使用 256 的批量大小进行训练，而循环神经网络混合密度网络使用批量大小为 5 的子轨迹，每条长度为 100。图 17 展示了四个具有代表性的训练子轨迹示例。

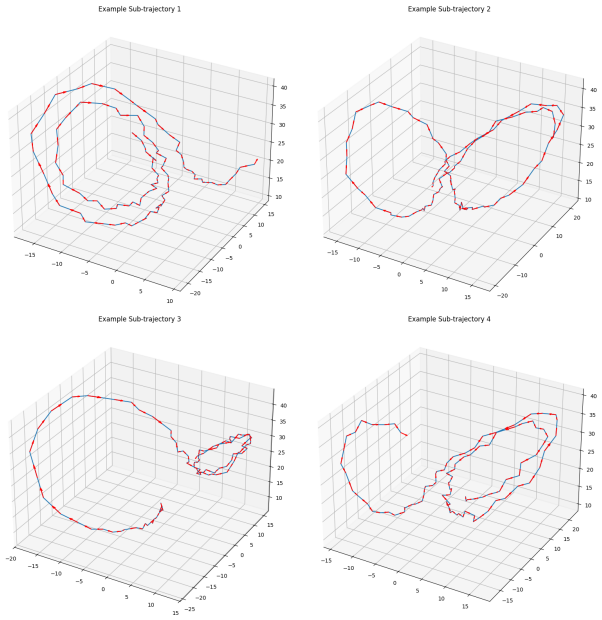

图17：用于训练洛伦兹系统 RNN-MDN 模型的子轨迹示例。

结果 我们训练了自回归模型来预测状态转移 $\Delta x_t := x_{t+\Delta t} - x_t$。随后，我们进行了滚动预测，以观察生成轨迹的长期行为。我们比较了使用均方误差训练的标准网络、普通混合密度网络以及循环混合密度网络（循环神经网络混合密度网络） [26, 27]。使用均方误差训练的模型无法在长期预测中捕捉轨迹的真实特性，而循环神经网络混合密度网络则成功捕捉了偶然不确定性。循环神经网络混合密度网络生成的轨迹始终保持在真实洛伦兹吸引子的流形上。

这种行为可以从图6所示的轨迹中直观地看出：MSE模型对吸引子上的可能位置变化取平均，因此生成的轨道过大；MDN模型则更能捕捉系统的真实变化，但在较长时间范围内会生成过于嘈杂的轨迹；最后，RNN-MDN模型是唯一能够成功捕捉解流形的几何与拓扑结构，并正确恢复每个吸引子附近两个“孔洞”的模型。我们还通过比较真实值与各方法预测结果之间的最大均值差异（MMD）进行定量分析，结果表明RNN-MDN的表现最佳，如表1及其图形表示图20所示。

可解释性 RNN-MDN能够以无监督的方式学习空间中有意义的划分，并学会将不同的混合分量分配给吸引子的两个“叶瓣”。图18展示了这一行为：该网络成功地学会了对空间进行有意义的分解，将问题中双叶结构的两个部分分开，同时识别出两种结构交汇处的模糊区域。

F.8 最大均值差异（MMD）计算

鉴于洛伦兹系统的混沌特性，均方误差（MSE）或条件似然等逐点误差估计，除了在极短时间范围内之外，无法提供有关预测轨迹真实性的有用信息。由于我们的系统还包含由观测噪声导致的固有随机性，这一问题更加突出。相反，我们计算真实值轨迹的统计特性，并将其与预测模型的统计特性进行比较。

在此类情形下，一种常见方法是计算不同方法预测的位置点云之间的经验最大均值差异（MMD）。对于两个点云 $X = (x_1, \dots x_{n_1})$ 和 $Y = (y_1, \dots y_{n_1})$，其中的点位于 $\mathbb{R}^d$ 中，具体做法是计算映射 $\phi\colon\ \mathbb{R}^d \to \mathcal{F}$ 从 $\mathbb{R}^d$ 映射到某个特征空间 $\mathcal{F}$。随后，两个数据集之间的最大均值差异被计算为特征均值之间的距离：

$$\left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x_i) - \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(y_i) \right\|_{\mathcal{F}}. \tag{20}$$

使用 核技巧，这可以利用适当的核 $k\colon\ \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ b 等价地计算 y

$$\begin{aligned} MMD_k(X,Y) = &\left(\frac{1}{n_1(n_1-1)} \sum_{i=1}^{n_1} \sum_{j \neq i} k(x_i, x_j) \right) \\ &- \left(\frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} k(x_i, y_j) \right) \\ &+ \left(\frac{1}{n_2(n_2-1)} \sum_{i=1}^{n_2} \sum_{j \neq i} k(y_i, y_j) \right). \end{aligned}$$

由于核的选择会影响该最大均值差异所捕捉的性质，因此通常会在一组核 [61] 上取最大值。对于这项工作，我们考虑径向基函数核族，并选择尺度参数 $\sigma \in [0.1, 50]$。每种方法的这一最大值对应于图20中的条形图，而对 sigma 值的完整依赖关系则见图19。

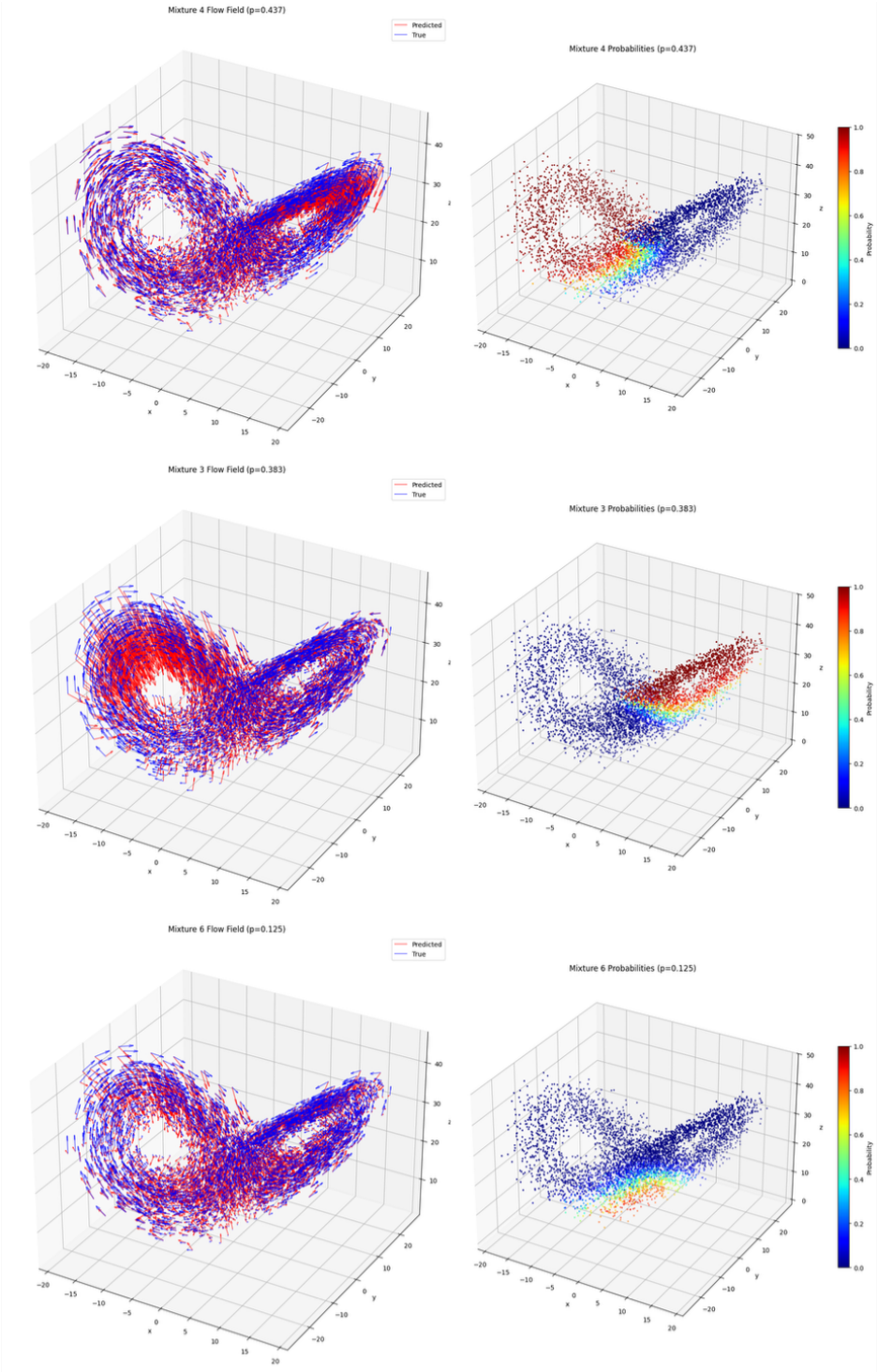

图18：洛伦兹问题中循环神经网络混合密度网络按边缘概率排序的前三个混合分量。左列显示预测位移场和真实位移场，右列显示基于输入的混合权重彩色图。

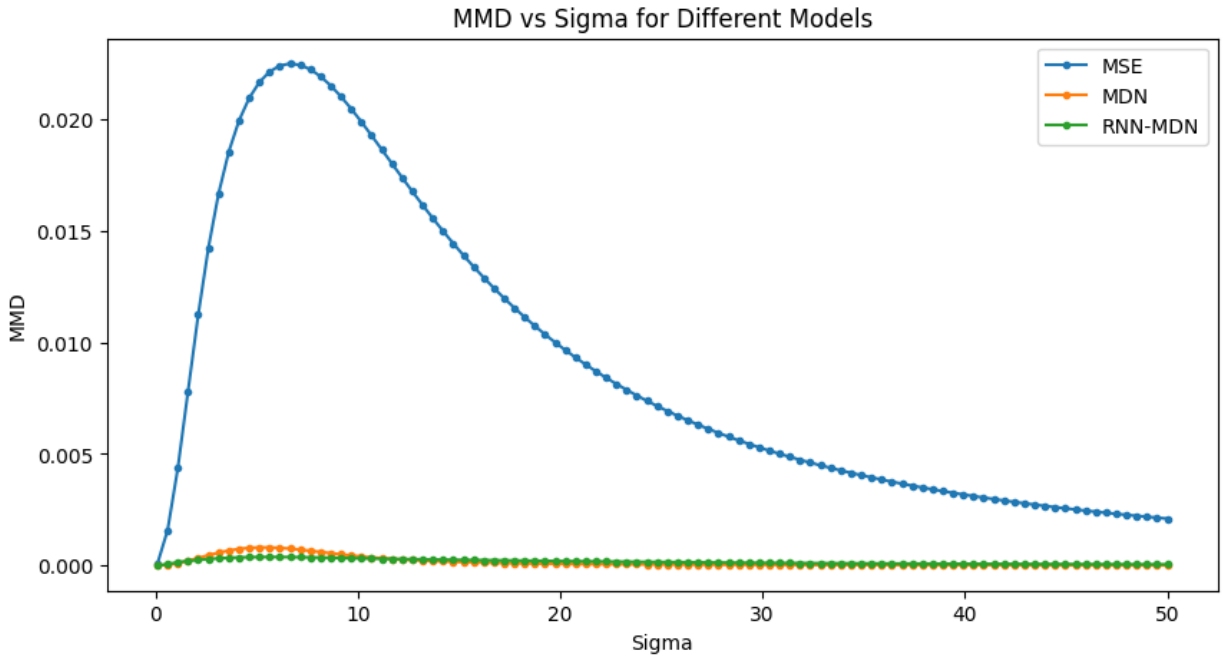

图19：MM　洛伦兹问题中 $\sigma \in [0.1, 50]$ 内不同方法在不同 sigma 值下的最大均值差异值　　.

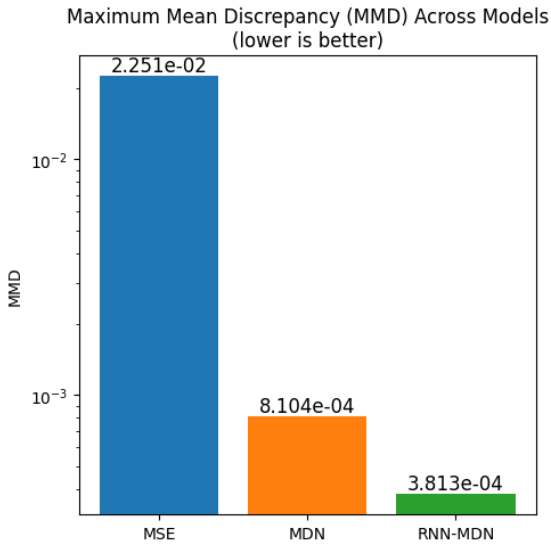

表1中报告的洛伦兹问题不同方法的最大最大均值差异值。