arXiv:2611.29363v1 [astro-ph.HE] 22 Nov 2026

学习器还是摘要构成了极限？审计 KASCADE 和 GRAPES-3 空气簇射数组的机器学习能量重建

Mugdha Raje Ashiya[1] 和 Samarjeet Malik[1*]

[1]焦特布尔印度理工学院计算机科学与工程系，印度焦特布尔。

*通讯作者。电子邮箱：samarjeetmalik.gs5@gmail.com；

现地址：阿育王大学计算机科学系，印度哈里亚纳邦索尼帕特

摘要

宇宙线阵列的机器学习能量估计器通常使用少数拟合的簇射参数进行训练，而非台站信号。当其分辨率不再提升时，标准诊断无法判断是更好的模型会有所帮助，还是摘要所含信息已经耗尽。我们采用五步审计，将训练好的模型与一个与学习器无关的估计值进行比较，以确定摘要所允许的最低误差（即其平方损失贝叶斯风险，或下限）；检验哪些输入能增加信息，并标记仅由拟合计算得出的输入。我们将此方法应用于采用三种强子相互作用模型（QGSJET-II-04、EPOS-LHC 和 SIBYLL-2.3d）的公开 KASCADE 模拟、13,848,509 个 KASCADE 实测事件，以及 GRAPES-3 模拟事件。在已知的 10^{16}电子伏特以下能量数量级范围内，基于九个重建特征的梯度提升在所有估计器设置下都与估计下限相差不超过 2.2%，因此，增加模型容量预计也无法带来超过约 2% 的提升。该下限是一个估计值；在区间覆盖不足的半合成基准测试中，其准确度约为 1%，并非真实贝叶斯风险的上界。将重建粒子数替换为真实粒子数后，在三个模型中，10^{15}电子伏特以下的误差方差减少 70% 至 72%，而在 10^{16} 至 10^{17}电子伏特之间减少 24% 至 26%；然而，模拟台站信号仅弥补了其中约十分之一的差距。在 GRAPES-3 上，使用拟合结果生成的图像作为输入的神经网络，表现并不优于拟合本身。在实测数据中，固定电子数时的平均缪子数落在每个模型的质子至铁核区间内；无标签探测也发现，μ子通道的使用方式与模拟一致；但这些都无法确定能量标度。带有能量标签的结果均来自模拟。

关键词： 广延大气簇射、能量重建、机器学习、贝叶斯风险、KASCADE、强子相互作用模型

1 引言

问题。 地面空气簇射阵列在数百个点位对簇射的粒子前沿进行采样。标准重建拟合侧向分布，在大多数数组中使用 Nishimura– Kamata–Greisen（NKG）函数 [1, 2], 来拟合台站信号，并保留少量参数：电子数、簇射年龄和核心位置。结合到达方向和一个缪子观测量，这些参数构成了摘要，该摘要会被归档、分发，并用作学习型估计器的初级能量和质量[3–5]的输入。

基于该摘要的学习型估计器无法恢复拟合过程舍弃的信息。当其分辨率不再提升时，可能有两个原因：估计器的容量、数据或调优不足，或者摘要已无更多信息可供利用。从模型角度看，两者并无区别：训练收敛，验证表现正常，学习曲线趋于平坦。通常的应对方法——增大模型，或根据存档量重建类图像输入——只在第一种情况下有效。基于台站信号进行重建可以优于摘要 [6–9], , 但这需要那些信号，而对于大多数存档数据来说，这些信号已无法获取。

我们的方法。我们考察学习型估计器的表现与摘要所能达到的水平有多接近，以及摘要损失了多少信息，并分析这些情况如何随能量和强子相互作用模型而变化。为此，我们采用五步审计（第 2 节）：

1. 在不训练学习器的情况下，估计摘要的平方损失贝叶斯风险，即其下限；2. 逐项添加特征，构建嵌套集合阶梯；3. 检查每项输入是测量所得，还是根据摘要计算得出，因为计算所得的输入无法降低贝叶斯风险；4. 探查模型输出是否在函数上依赖于摘要，这不需要标签；5. 重新测量任何沿用的数据筛选。

在模拟提供真实粒子数时，用真实粒子数替换重建粒子数的预言机，可作为参考，衡量重建过程造成的信息损失。这些组成部分都不是新的。本文的贡献在于将它们结合成一套流程，并为每个估计步骤设置对照，并将其应用于两个数组。与特征消融不同，特征消融只表明哪些输入有助于给定的学习器，而审计则加入了不依赖该学习器的参考。只有步骤 3 中的保证是精确的；下限和探测结果都是估计值，每次运行都设有答案已知的对照（图 1）。

数据。KASCADE 实验的公开模拟数据是主要测试平台 [10] ，由 KASCADE 宇宙线数据中心（KCDC）[11, 12], 发布，涵盖三种强子相互作用模型：QGSJET-II-04 [13], EPOS-LHC [14] 和 SIBYLL-2.3d [15]。除重建的摘要外，KCDC 还发布了真实的电子数和缪子数，以及每个探测器站的信号，因此可以将摘要与其未涵盖的信息以及预言机进行比较。第二个测试平台是 GRAPES-3 模拟事件 [16],„ 其中只导出了摘要；因此，所有可构建的输入都是摘要本身或其函数。

主要结果。表1总结了研究发现 gs.

•**提升模型容量并非改善摘要的主要途径**。在低于 10_{16} 电子伏特的能区内，对于所有估计器设置和三种模型，基于九个重建特征进行梯度提升的结果与估计下限相差不超过 2.2%（第 4 节）。

• **重建成本随能量增加而降低**。真实粒子数在 10_{15}电子伏特以下可消除 70–72% 的误差方差，在 10_{16} 至 10_{17}电子伏特之间可消除 24–26%，且在各模型中结果一致。

• **台站信号只能恢复其中一小部分**。我们尝试的编码器弥补了与预言机之间约十分之一的差距（第 4.6 节）。

•**渲染图像没有增加任何信息**。在 GRAPES-3 上，使用拟合结果渲染出的图像作为输入的神经网络，表现并不优于拟合本身（第 6 节）。

•**实测数据并不与模拟相矛盾**。无标签检验与测量的 KASCADE 事件上的模拟结果一致（第 5 节）；它们不检验能量标度。

表1 主要结果一览。按十年区间给出的数值以已知真实能量十年区间为条件（第4.5节将取消这一条件），能量标记结果为模拟结果

问题	结果	Sec.
学习器是否是 10^{16}电子伏特以下的极限?	提升法的结果在**2.2%**以内估计下限	4
固定的 KCDC公式 比较结果如何?	校准后，高于下限**6–32%** 下限估计值	4.2
什么 重建大小 代价?	**70–72%** 的误差方差 低于 10^{15}电子伏特；**24–26%** 位于 10^{16}–10^{17} eV	4.4
站点在多大程度上 神经网络能恢复多少?	**7.8–11.2%** 的预言机差距	4.6
渲染后的 在 GRAPES-3 上?	没有新信息：卷积神经网络 0.1227，基于拟合数据进行提升法 0.1230	6

2 方法

2.1 下限、估计下限与分辨率

设 X 为一个事件的台站信号，$T = t(X)$ 为导出的摘要，$y = \log_{10}(E/\mathrm{GeV})$ 为目标。对于平方损失，从 T 预测 y 的贝叶斯风险为

$$R^\star(T) = \inf_f \mathbb{E}\big[(y - f(T))^2\big] = \mathbb{E}\big[\mathrm{Var}(y \mid T)\big],$$

由 $f(T) = \mathbb{E}[y \mid T]$ 达到。我们将 $\sigma^\star(T) = \sqrt{R^\star(T)}$ 称为 T 的下限 。它是总体的一个属性，无法直接观测。估计下限 $\hat{\sigma}_{\mathrm{B}}(T)$ 是在伽马检验假设下对其作出的与学习器无关的估计（它不训练预测器，但取决于度量、缩放和邻居范围），且误差可能为正也可能为负；它不是下界，估计器的得分可能低于它。已训练估计器的分辨率 σ 是其在留出事件上的残差标准差。选择平方损失，是因为其贝叶斯预测器是条件期望，且风险可以精确分解（附录A）。以下所有论述均针对能量的条件均值；即使某个通道不改变条件均值，仍可能与成分或校准不确定性有关。

估计值。伽马检验 [17, 18] esti- mates $R^\star(T)$，无需拟合回归器。对于每个查询事件 i let $N_k(i)$ be its k-th nearest neighbourinstandardised feature spaceand

$$\gamma(k) = \frac{1}{2n}\sum_{i=1}^{n}\big(y_i - y_{N_k(i)}\big)^2,$$

$$\delta(k) = \frac{1}{n}\sum_{i=1}^{n}\big\|T_i - T_{N_k(i)}\big\|^2,$$

其中 n 是查询（测试）事件的数量。对于导数有界的回归函数$\gamma(k) \to R^\star(T)$，当 $\delta(k) \to 0$ 时，以及对 δ在 $k = 1,\dots,10$ 个估计值上的 γ进行线性拟合所得的截距。测试事件就是查询，近邻仅从其他簇射中抽取，而区间则对簇射进行重采样。在每次使用估计器之前，都会在真实特征矩阵上运行该估计器，并使用一个噪声已知的合成目标（由标准化的电子数、μ子尺寸和年龄的固定平滑组合 $0.9z_1 + 0.35z_2 + 0.08\tanh 2z_4$ 构成，缩放至目标离散程度的 85%，再加上标准差为 0.10 或 0.15 的高斯噪声）；它必须将噪声水平恢复到 10% 以内，否则分析将停止。该对照可防止估计器在特征几何结构上失效；其精度要求低于所需精度，而第 4 节的灵敏度研究和基准测试将处理后者。在具有已知贝叶斯风险的基准测试（附录C）中，该估计值在最多六个维度时的误差为 0.5%，在八个维度时为 1.9%，但在 32 维测量值上会失效。因此，本文仅将其用于最多包含九个特征的摘要，从不用于台站信号。

2.2 两个命题

命题1（确定性渲染无法降低贝叶斯风险） *Let y be square-integrable and g a deterministic measurable map. Then* $\inf_h \mathbb{E}[(y - h(T))^2] \geq \inf_f \mathbb{E}[(y - f(T))^2]$, *with equality if andonly if* $\mathbb{E}[y \mid g(T)] = \mathbb{E}[y \mid T]$ *almost surely.*

根据拟合参数计算得到的探测器密度图像就是这样一种 $g(T)$。无论编码器、架构或训练预算如何，都无法改变这一界限；只有单独测量的通道才能改变它。该表述涉及总体风险，而不是模型学习的难易程度。

命题 2（测量通道能够增加什么）
If $y \in [a,b]$ almost surely and M is any further observable, then

$$R^\star(T) - R^\star(T,M) = \mathbb{E}\big[(\mathbb{E}[y \mid T,M] - \mathbb{E}[y \mid T])^2\big] \le \tfrac{(b-a)^2}{2}\, I(y; M \mid T).$$

两者都可由塔式性质推出；不等式使用了 Pinsker 不等式（附录A）。第二个结论的逆命题不成立：通道可以携带关于 y 的信息，却不改变条件均值，因此平方损失下的零结果只涉及条件均值。第 4 节的阶梯逐级估计的正是这个等式；该界仅作定性用途，用来说明不含条件信息的通道无助于改善结果。

2.3 审计

图 1 展示了各个步骤。

1. **下限**。将具有不同归纳偏差的估计器与估计下限进行比较：就该目标和损失而言，如果估计结果所能说明的范围内，两者在下限处一致，则可认为受表征限制；如果一致时远高于下限，或两者不一致，则可认为受学习器限制。

2. **阶梯**。 从拟合结果出发，逐次向外构建摘要，每次加入一个量，并在相同事件上用配对区间评估每一级。如果某一级降低了某个学习器的误差，就说明该级对这个学习器有帮助；只有当新误差低于前一级的估计下限时，才能据此认为贝叶斯风险有所降低。

3. **呈现检查**。 对于任何呈现得比摘要更丰富的输入，都应确定其是测量得到的，还是根据摘要计算得到的；如果是计算得到的，则适用命题1。

4. **退化探测**。 使用以下内容探测训练好的模型输出 m

$$R^2_{\mathrm{deg}}(T) = 1 - \frac{\mathbb{E}[(m - \mathbb{E}[m \mid T])^2]}{\mathrm{Var}(m)},$$

用梯度提升法估计的平方相关比 [19],（在模拟数据中按簇射分组进行五折折外估计），仅使用 T 时，以及使用 T 和其他实测通道 M时的结果。使用 M 的模型在 T 上的得分较低，而在（T，M）上的得分接近 1；类噪声变化会使两者的得分都降低。其尺度由构造已知的对照项（添加噪声、与M 的函数混合）确定，而不是由数值 1 确定：由于 $\mathbb{E}[m \mid T]$是由容量有限的回归器近似的，探测衡量的是该回归器能够解释的部分，而 T 的函数得分略低于 1。我们将 M的探测增量定义为 R^2_{deg}（T，M）$- R^2_{\mathrm{deg}}$（T）。探测只需要预测结果和所列输入，因此适用于实测数据。

5. **筛选**。在应用继承的筛选的区间内重新测量它：分别在有筛选和无筛选的情况下进行评分，在该筛选条件内重新训练，并报告该筛选旨在保护的量，同时区分三个估计目标：操作变化，即评分总体随截断值变化；在固定总体上于该截断范围内训练的影响；以及根据其下限估计得到的保留事件和拒绝事件的难度。

预言机。 在模拟提供真实粒子数的情况下，用真实粒子数替代重建的粒子数：使用预言机特征训练的梯度提升模型在 σ 上的变化，就是观测和重建在总尺寸方面损失的部分。这是一个实用的参考，而不是表示台站信号能够恢复内容的严格上限，因为其空间模式携带了总尺寸之外的信息。

标题与探索性比较。审计围绕以下比较展开：每个样本中的梯度提升法与下限估计、相对于预言机的差距，以及预先记录的渲染结果（第 6 节）。阶梯梯级、台站表示、编码器变体和选择变体均属探索性分析：我们为它们给出配对簇射区间，但不进行多重性校正，并将其视为描述性结果。

估计器。对于摘要，我们使用岭回归、随机森林、直方图梯度提升（600 次迭代，学习率为 0.06）和多层感知机（两个隐藏层，每层 64 个单元，提前停止），设置均预先固定，且未针对测试事件进行调优。对于台站信号，我们使用小型卷积网络、残差网络、台站Transformer和图网络（附录E）。全文中，lg 表示 $\log_{10}$，分辨率 σ 是

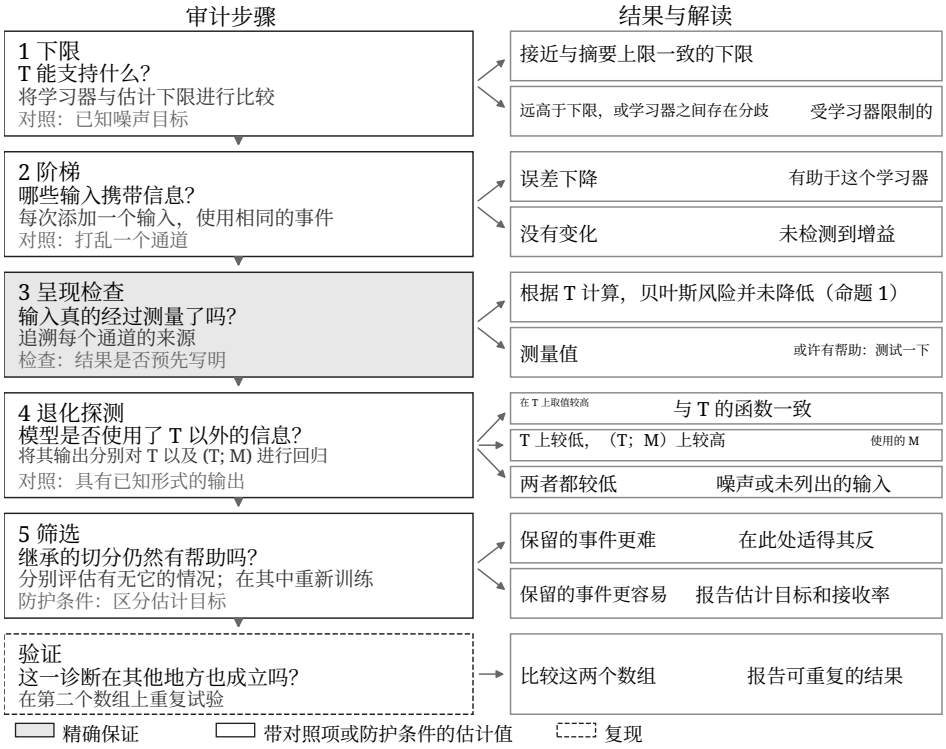


图 1 审计：五个诊断步骤，以及在使用不同探测器的第二个数组上进行的验证，用于区分表征受限平台期与受学习器限制的平台期。T 是摘要，M 是另一个测量通道。步骤 3 基于一项保证（命题1）。步骤 1、2 和 4 是带对照项的估计值，步骤 5 是带防护条件的估计值，而验证则是一次复现

残差的标准差；若均方根误差有所不同，则两者均予以报告。写作 $\log_{10}(E/\mathrm{GeV}) \in [5,6]$ 的能量范围等同于 10^{14}–10^{15} 电子伏特。区间为针对簇射计算的 95% 自助法区间；配对区间对比较双方使用相同的簇射进行重采样，且不包含不同训练运行之间的变化；在相关之处，这种变化将另行报告。

人工智能助手的使用。使用了人工智能助手（Claude，Anthropic）来编写部分分析代码，并起草和编辑部分文本。作者检查了代码及其输出，文中的每个数字均由分析流水线根据结果文件生成，并通过脚本与结果文件核对；作者对本文内容承担全部责任。

3 数据

***KASCADE*模拟。**KCDC 发布了 KASCADE 数组的模拟数据（由 252 个站点构成，按 $200 \times 200\,\mathrm{m}^2$ 网格排列，并配有独立的电磁和 μ子探测器 [10]），使用 CORSIKA[20]，FLUKA（低于 200 GeV）以及三种高能模型之一 [12]：QGSJET-II-04、EPOS-LHC 和 SIBYLL-2.3d 生成。每组数据包含五种初级粒子（质子、氦、碳、硅、铁），按 E^{-2} 谱在 10^{14} 至 10^{18}eV 范围内生成；每种初级粒子对应数十万个簇射，每个簇射均以三个核心位置为投射点。发布的事件均通过了 KCDC 的重建和质量筛选。每个事件包含重建的 $\log_{10} N_{\mathrm{e}}$、$\log_{10} N_{\mu}$、年龄 s、核心位置和天顶角 θ，以及真实能量、初级粒子、粒子数、各站沉积量和到达时间。九个特征分别为四个拟合参数（$\log_{10} N_{\mathrm{e}}$，s，X_c，Y_c）、θ 和 $\cos\theta$、核心距离、$\log_{10} N_{\mu}$ 以及比值 $\log_{10} N_{\mu} - \log_{10} N_{\mathrm{e}}$；其中三个特征是其他特征的函数。图 2 展示了两个事件，图 3 展示了平均侧向分布。

抽样与数据划分。发布数据集文件按初级粒子和能量排序，因此，若干个连续区块并非随机样本；

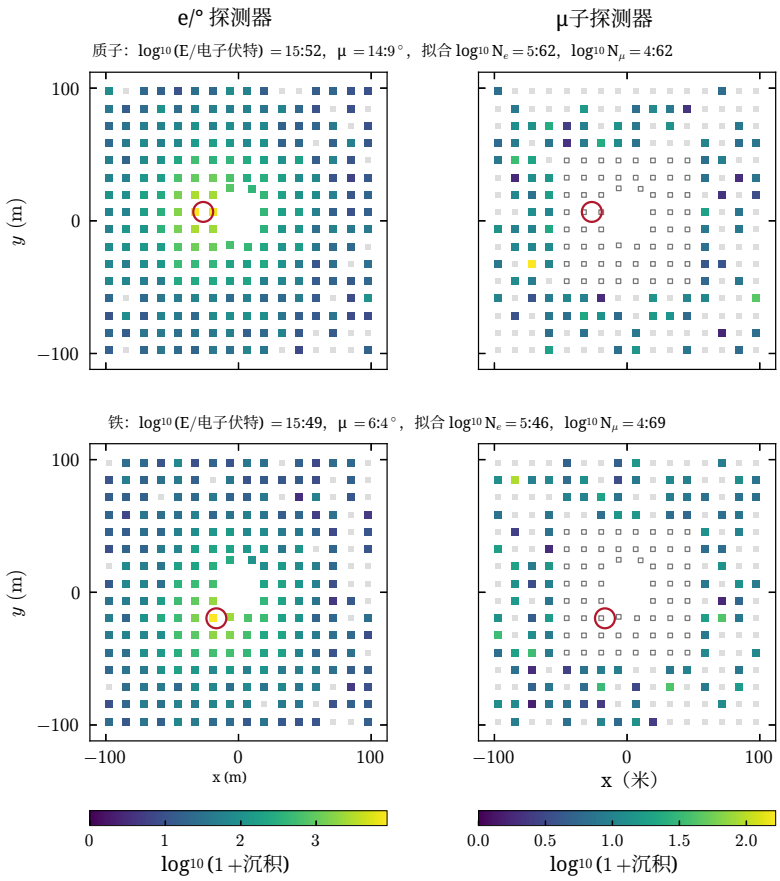


图 2 两个能量几乎相同的 KASCADE 模拟簇射（QGSJET-II-04）：第一个是质子簇射（上方，天顶角 14.9°），第二个是铁原子核簇射（下方， 6.4°）。文件中真实能量与 $\log_{10}(E/\text{电子伏特}) = 15.5$ 相差不超过 0.03、天顶角低于15° 且核心位于 30 米范围内的事例：252 个站点的电磁沉积和 μ子沉积，单位采用 KCDC 发布数据集中的单位，并标出重建核心（圆圈）。灰色方块：无信号；μ子面板中的空心方块：60 个内部站点，这些站点没有 μ子探测器。拟合保留规模、年龄和核心；摘要舍弃的正是台站分布

我们发现，这类区块会使第 4.4 节的预言机值变化超过其区间。按能量分层的分析会均匀随机抽取事件（每个十年最多 300，000 个）；站点级分析则限定在 $\log_{10}(E/\text{GeV}) \in [5, 6]$，，读取 2，400 个由各 200 个事件组成的等间距区块。这些区块的初级粒子比例与完整文件相差不超过 0.0003，中位能量在 $\log_{10}(E/\text{GeV})$ 上相差不超过 0.0023。对于 $\log_{10}(E/\text{GeV}) \in [6, 7]$，从每个发布数据集在该窗口内随机抽取最多 400，000 个事件。由于每个簇射会被模拟三次，其副本共享能量、初级粒子和方向，并且所有划分都以簇射为单位：75% 的簇射用于训练，25% 用于测试。所有评分均按事件取平均，因此，每个通过发布数据集筛选条件的副本都会使相应簇射计数一次，这就是对发布数据集进行分析时所见到的总体；下限估计取仅来自其他簇射的近邻。

KASCADE 测量数据。KCDC 还发布了采用相同重建方法得到的数组测量事例；第 5 节将使用这些数据。

GRAPES-3。第二个测试平台是 GRAPES-3 合作组针对 Ooty 的 GRAPES-3 数组模拟的 171，776 个 CORSIKA（QGSJET-II-04）/GEANT4 事件，涵盖质子、氦、氮、铝和铁，且均处于同一个簇射尺寸区间，$\log_{10} N_e \in [4.8, 5.2]$，位于 $\theta \leq 24.6°$。该数组约有 400 个间距为 8 米的闪烁体，以及一台读出为 16 个模块的 560 米^2 μ子望远镜 [16, 21, 22]。导出的数据只有摘要和 16 个模块计数。这些事件来自 47，394 个簇射（平均每个簇射有 3。6 个副本）；结果基于按簇射进行的五折交叉拟合给出（附录B）。

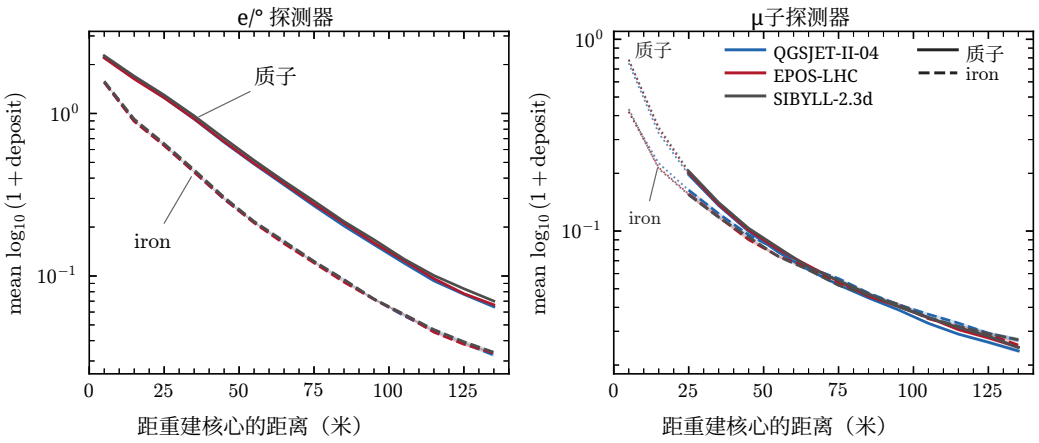


图3 质子（实线）和铁（虚线）在真实能量为 $10^{14.4}$–$10^{14.6}$ 电子伏特、天顶角小于 30° 且核心距阵列中心 60 m 以内时，三种相互作用模型下各站点的 $\log_{10}(1+\text{沉积量})$ 均值与距重建核心距离的关系。均值包含无信号的站点。μ子均值仅采用配备 μ子探测器的 192 个站点；点状线段标示事件占比不到一半的区间，因为内侧站点没有这些区间的数据。区间表示按事件计算的标准误差，并将簇射副本视为独立样本；区间宽度最多约为 1.7 个线宽。点状线段上未绘出区间，此处误差可达约 3 个线宽。在所有距离上，铁簇射的电磁信号都较少。质子在核心附近产生的 μ子信号较多；距离超过约 60 m 后，两者相近；在 QGSJET-II-04 模型下，铁在这一量上最多比质子高 14%。

4 KASCADE：不同能量和相互作用模型下的摘要

4.1 估计下限处的平台区

关键结果。在 10^{16}电子伏特以下，基于九个特征的提升法在所有三个模型上的表现与估计下限相差不超过 2。2%。

表2和图4按真实能量十倍区间给出了审计结果，模型均在各自区间内训练并评分；第4.5节则在不具备这一信息的情况下重复了比较。在从 10^{14} 到 10^{17} 电子伏特的三个十倍区间中，基于九个特征的梯度提升与估计下限的差距，在最低区间为 -0.2%至 $+0.5$%，在下一区间为 -0.3%至+0.6%，在第三个区间为 -0.9%至 $+1.8$%；对于此样本，九个区间均包含零。这些区间仅反映簇射抽样的影响；它们忽略了估计器的偏差，而且如下文的基准测试所示，区间过窄。多层感知器在较低的两个区间内与梯度提升相差不超过 1.7%，在第三个区间则高出 2.1 至 5.2% ；该区间的训练事件较少。岭回归在每个区间的表现都远远不及其他方法。已知噪声对照在每个区间和每种模型下均通过。最高能量区间中，每种模型有 3,682 至 7,872 个事件，对于学习器或下限估计而言都太少，因而仅报告结果而不作解读。

下限估计在多大程度上值得信赖。估计值有两个变异来源（图5）：簇射抽样，以及估计器设置的选择（邻居范围、特征缩放、度量、坐标）。附录D中的表 D5 给出了每个模型在每个十年区间（对发布数据集进行第二次随机抽取，抽取结果与表2的样本有重叠，且在 10^{16} 至 10^{17} 电子伏特之间包含相同的事件）以及站点级样本中的这两项变异。在12种情形中，有9种情形下，设置对估计值的影响超过其簇射区间，因此设置是主要影响因素；仅曼哈顿度量就能使其变动多达 3.2%。在参考设置下，梯度提升在所有情形中的结果都与估计值相差不超过1.8%，且其区间在12种情形中有2种不包含零，低端最多仅为0.1%。在所有设置下，低于 10^{16} 电子伏特时，梯度提升与估计值相差不超过2.2%，有时高于、有时低于估计值；在 10^{16} 至 10^{17} 电子伏特之间，则与估计值相差 -8.1% 至 $+3.9$%，此处事件较少，使得离散范围更大；其中较宽的一端几乎完全由秩缩放造成，不采用秩缩放时，范围为 -2.4 % 至 $+3.8$%。使用六个无冗余坐标而非全部九个坐标，最多使估计值变动1.3%，因此这三个派生特征不会扭曲邻居几何结构。

表2 按真实初级能量十年区间对三种强子相互作用模型进行的审计。事件从每个发布数据集中均匀随机抽样（每个十年区间最多300，000个），按 75/25 CORSIKA簇射划分。 σ 是梯度提升在九个重建特征上所得残差（以lg(E/GeV)表示）的标准差（模型在各十年区间内训练）；下限估计采用伽马检验，以测试事件作为查询，并从其他簇射中选取近邻；超额是梯度提升相对于下限估计的结果，并附带簇射自助法区间，该区间仅反映簇射抽样，且覆盖不足（表 D6）。预言机将重建的电子和μ子尺寸替换为其真实值；最后一列是以预言值为参照的误差方差差距（$\sigma^2 - \sigma^2_{\text{oracle}}$）/$\sigma^2$。最高能量十年区间受数据限制，因此不作解读。粗体表示预言机差距，即本表针对所解读十年区间的主要结果

lg(E/电子伏特)	模型	事件	提升法	下限估计	超额	区间	预言机	预言机差距
14–15	QGSJET-II-04	300，000	0.1481	0.1479	+0.1%	$[-0.6, +0.8]$	0.0807	**70%**
	EPOS-LHC	300，000	0.1485	0.1478	+0.5%	$[-0.2, +1.3]$	0.0783	**72%**
	SIBYLL-2.3d	300，000	0.1490	0.1493	−0.2%	$[-0.9, +0.6]$	0.0805	**71%**
15–16	QGSJET-II-04	300,000	0.0869	0.0864	+0.6%	$[-0.3, +1.4]$	0.0643	**45%**
	EPOS-LHC	300，000	0.0850	0.0851	−0.1%	$[-0.9, +0.9]$	0.0607	**49%**
	SIBYLL-2.3d	300，000	0.0872	0.0875	−0.3%	$[-1.1, +0.5]$	0.0639	**46%**
16–17	QGSJET-II-04	78，023	0.0614	0.0603	+1.8%	$[-0.3, +4.3]$	0.0536	**24%**
	EPOS-LHC	36，493	0.0596	0.0586	+1.7%	$[-1.6, +5.5]$	0.0521	**24%**
	SIBYLL-2.3d	36，940	0.0623	0.0628	−0.9%	$[-3.7, +1.7]$	0.0535	**26%**
17–18	QGSJET-II-04	7，872	0.0564	0.0481	+17.2%	$[+5.6, +33.7]$	0.0513	17%
	EPOS-LHC	3，682	0.0604	0.0482	+25.3%	$[+8.4, +54.5]$	0.0575	9%
	SIBYLL-2.3d	3,702	0.0583	0.0493	+18.1%	$[-0.8, +45.9]$	0.0539	14%

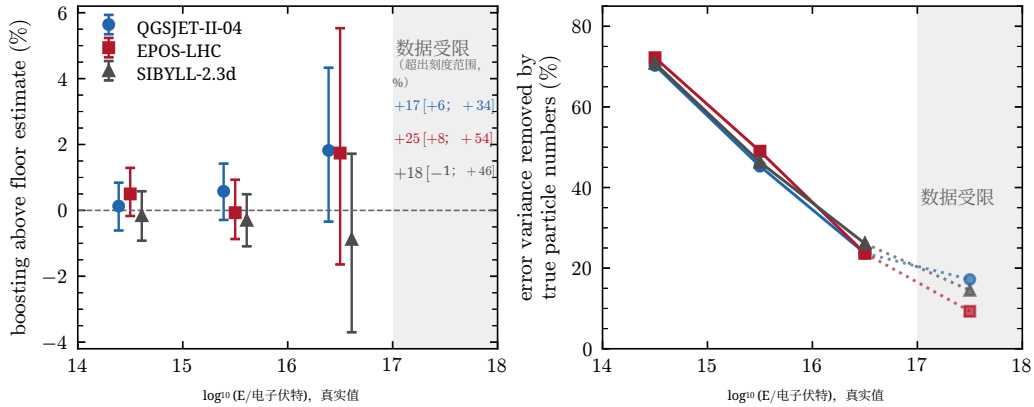


图 4 按真实能量十倍区间划分的 KASCADE 模拟，涵盖三种强子相互作用模型。左：在九个重建特征的 Gamma 检验下限估计之上，使用梯度提升法；簇射自助法区间仅反映簇射抽样，且覆盖不足；最高能量十倍区间（灰色）受数据量限制，超出此图的刻度，因此列出了其数值和区间。右：以预言值为参照的误差方差差距，即用真实粒子数替换重建粒子数后，摘要的误差方差减少的比例。此处未显示区间；虚线段延伸至数据量受限的十倍区间

直接的最近邻值 $\sqrt{\gamma(1)}$，即不采用 δ 中的外推时，比估计值高1%至17%；稀疏十年区间中的差异最大，这符合预期，因为该估计值包含近邻之间回归函数的变化。

***KASCADE*几何中的准确度。** 设置敏感性说明估计值变化了多少，而非其与真实值相差多远。为了在重要的几何结构中衡量后者，我们保留每个样本的真实九个特征、样本量及其簇射副本，并将目标替换为 $g(X) + s(X)\xi$：g 和 s 根据真实目标进行学习（梯度提升法用于条件均值和残差平方的情况也是如此，因此回归曲面和异方差性都具有现实性），而 ξ 服从标准正态分布，每个 CORSIKA 簇射仅抽取一次。这样就知道了真实下限 $\sqrt{\langle s^2 \rangle}$，并可在 20 次噪声重复试验中，以它为基准对每种设置进行评分（附录D中的表 D6 ）。低于 10^{16} 电子伏特时，参考估计值偏高 $+0.1$ 至 $+0.6$%；除秩缩放和直接1近邻值外，所有设置的估计值与真实值的差异都在 0。0 至 $+0.6$ %以内；在稀疏十倍区间 10^{16}–10^{17}电子伏特内，参考估计值偏高$+1.7$ 至 $+2.4$%，秩缩放则最高偏高 $+8$ %，这正是设置间取值范围较宽的原因。

参考估计值的簇射区间在 60% 至 95% 的重复试验中覆盖了真实下限，低于其名义上的 95%，这是因为它忽略了这一小幅向上偏差。根据这种方式测得的偏差校正下限估计后，在每个低于 10^{16} 电子伏特的样本中，提升法的估计值比校正后的估计值高 0。0 至 $+1.4$%。因此，有数据支持的结论是：在这种几何结构中，梯度提升与一个准确度经过测量的估计值相近——在所有设置下，低于 10^{16}电子伏特时，梯度提升与估计下限相差约 2%；而该估计值本身在这一区间内的误差小于 1%。这并不能界定它与真实目标的真实贝叶斯风险之间的差距；基准测试所模拟的回归曲面并不等同于真实目标的回归曲面。

不确定性带来的变化。有三项结论在所有设置下都成立：测量得到的偏差，以及区间的覆盖不足。在低于10^{16}eV时，梯度提升在所有设置下都处于估计值的2.2%以内；一旦根据基准测试中测得的偏差对下限估计进行校正，其结果最多比下限估计高 $+1.4$%。具有不同归纳偏置的非线性估计器彼此一致（在$\log_{10}(E/\mathrm{GeV}) \in [5,6]$ 内相差不超过0.0013 in σ），而岭回归则明显偏低；因此，增加九个特征的模型容量，预计不会使其在inσ 处再提高约百分之二以上。另有两项结论并不成立。无法确定提升法位于真实下限之上还是之下，也无法确定其在该范围内偏离多少。在稀疏十倍区间 10^{16}–10^{17}eV内，即使不进行秩缩放，也没有任何比较结果能达到优于设置范围的分辨率；该范围为 -2.4 至 $+3.8$%，而估计值最高偏高$+2.4$%。本文为下限估计给出的簇射区间仅反映簇射抽样的影响，因此应被视为其不确定性的下限。

4.2 标准重建与下限的对比

关键结果。 校准后，固定的 KCDC公式比下限估计高出 6 至 32%，且这一差值随能量增加而扩大。

KCDC 会随每个事件一并发布合作组自己的能量估计值 E_{rec}，这是电子数、μ子尺寸和天顶角的固定函数。表3和图6将其与梯度提升法及下限估计值在同一测试事件上进行比较。未经处理时，其 $\lg E$ 中的平均偏差为 $+0.010$ 至 $+0.055$ 。校准后（仅使用 E_{rec} 进行梯度提升，从而消除偏差和非线性），在所有三个模型上，其分辨率在低于 10^{15} 电子伏特时比下限估计值高 6% 至 7%，在 10^{15} 至 10^{16} 电子伏特之间高 15% 至 16%，在 10^{16} 至 10^{17} 电子伏特之间高 28% 至 32%。与此处的每个模型一样，校准是在十年能量区间内拟合的（第4.5节比较了不按十年区间拟合的情况）。由于 E_{rec} 本身是摘要的函数，它不可能优于下限（命题 1）；这一差距随能量增大，体现了固定的三项公式未能利用九特征摘要所能提供的信息，而使用相同摘要的标准学习器则能弥补这一差距。

按初级粒子分类的偏差。对于能量标度，各初级粒子之间的平均偏差跨度与分辨率同样重要。在混合样本上以平方误差进行训练，会使平均误差最小化，并允许在不同初级粒子之间权衡偏差，因此较低的 σ 未必意味着成分依赖性较弱。在 10^{15}电子伏特以下，梯度提升法的偏差跨度（$\lg E$中为0.075至0.085）大于校准后的 KCDC 估计值（0.057至0.070）；在 10^{15} 至 10^{16}电子伏特之间，其偏差跨度较小（0.011至0.017，而后者为0.023至0.036）；在更高能量下，校准后的 KCDC 估计值在三个模型中有两个模型的偏差跨度更小。偏差跨度只是粗略的汇总指标——提升法的偏差并不随质量单调变化（图6，中间），因为在混合样本上进行平方误差训练会将每种初级粒子都拉向混合样本，从而可能使最大和最小偏差相近，却不代表估计值不依赖质量。在每种初级粒子内部，提升法在初级粒子、十年能区和模型的全部45种组合中都比另一种方法更精确；因此，当其偏差跨度更大时，它是在以分辨率换取成分依赖性，而不是两方面都更差。使用真实粒子数的预言机在 10^{15}电子伏特以下的偏差跨度与提升法相当（0.069至0.081），而在更高能量下则比两者都大（在 10^{15} 至 10^{16}电子伏特之间为0.049至0.060），因此成分依赖性源自簇射：在能量固定时，簇射规模取决于初级粒子质量，而非重建过程。若分析需要与成分无关的能量标度，就必须校正质量，如同 KASCADE 的展开所做的那样[23]；较低的平方误差并不能实现这一点。

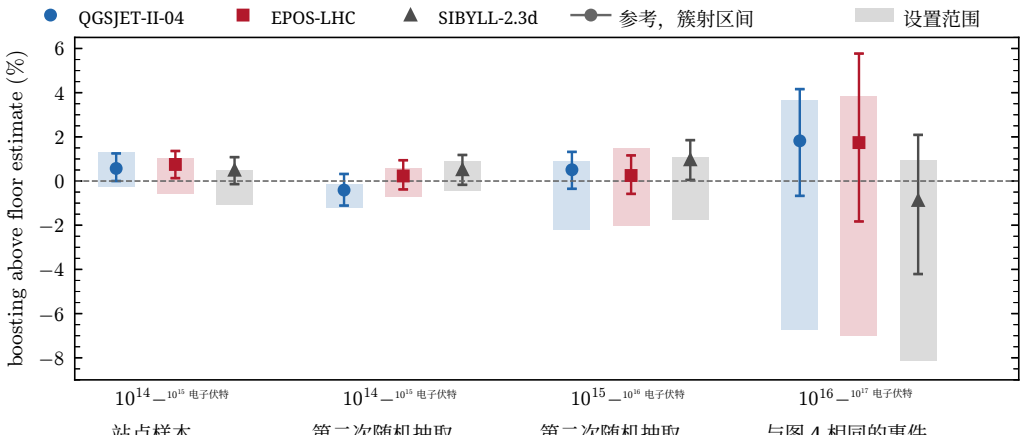


图5 每个模型和样本中高于下限估计的梯度提升：点和误差线表示带有联合簇射区间的参考设置，阴影条表示所有估计器设置的范围（表 D5）。站点样本是包含约 410，000 个事件的站点级样本。标记为第二次随机抽取的组与图4的样本重叠； 10^{16}–10^{17}电子伏特样本中的事件相同。灵敏度运行的站点样本下限与站点分析中的下限在各模型间相差 0。0002 至 0。0004（图7的 QGSJET-II-04 样本相差 0。0002）。在 10^{16} 电子伏特以下，除九种情况中的一种外，设置范围均包含零；在 10^{16}–10^{17} 电子伏特范围内，离散程度更大。此处以及图4中的区间分别来自独立的联合自助法运行（分别进行 500 次和 300 次重复试验），因此对于相同事件，其结果会略有不同。簇射区间在基准测试中覆盖不足（表 D6），因此是不确定性的下限

表3 KASCADE自身的能量估计值 E_{rec} （$\log_{10} N_e$、$\log_{10} N_\mu$ 和天顶角的固定函数）与基于九个特征和下限估计的梯度提升法在同一批测试事件上的表现比较。原始值：$\lg E_{\mathrm{rec}}$ 的均方根误差及其平均偏差；校准后：仅基于 E_{rec} 的提升法，可消除偏差和非线性（残差的标准差）。最后三列分别给出五种初级粒子之间平均偏差的跨度（最大值减最小值）：提升法、校准后的 KCDC 估计值，以及使用真实尺寸的预言机。粗体表示提升法和校准后的 KCDC 估计值中分辨率较高者，以及两者中偏差跨度较小者，无论哪种方法更占优

				KCDC估计值			不同初级粒子之间的偏差跨度		
lg(E/电子伏特)	模型	提升法	下限估计	原始（偏差）	校准后	高于地板	提升法	KCDC	预言机
14–15	QGSJET-II-04	**0.1481**	0.1479	0。1942（+0.027）	0.1585	+7%	0.075	**0.057**	0.081
	EPOS-LHC	**0.1485**	0.1478	0.1956 (+0.025)	0.1588	+7%	0.083	**0.066**	0.069
	SIBYLL-2.3d	**0.1490**	0.1493	0。1972 (+0.034)	0.1586	+6%	0.085	**0.070**	0.073
15–16	QGSJET-II-04	**0.0869**	0.0864	0。1216 (+0.010)	0.0994	+15%	**0.015**	0.023	0.060
	EPOS-LHC	**0.0850**	0.0851	0。1229 (+0.022)	0.0984	+16%	**0.017**	0.036	0.049
	SIBYLL-2.3d	**0.0872**	0.0875	0。1264 (+0.028)	0.1002	+15%	**0.011**	0.035	0.054
16–17	QGSJET-II-04	**0.0614**	0.0603	0.0946 (+0.025)	0.0790	+31%	0.030	**0.009**	0.048
	EPOS-LHC	**0.0596**	0.0586	0。1036 (+0.050)	0.0772	+32%	**0.021**	0.024	0.039
	SIBYLL-2.3d	**0.0623**	0.0628	0.1089 (+0.055)	0.0801	+28%	0.024	**0.018**	0.042

4.3 信息从何处进入

关键结果。 天顶角和 μ子尺寸承载着阶梯中的信息。

在每个十年和每个模型中，天顶角和μ子尺寸构成阶梯（图7；附录D中的表 D3）。在 $\log_{10}(E/\mathrm{GeV}) \in [5, 6]$中，它们在三个模型上的降幅分别为 0.0250 至 0.0265，以及 0.0259 至 0.0264。核心距离和缪子比是已有量的函数，因此根据命题1，它们无法降低贝叶斯风险；它们对学习器的 σ 影响最多为 0.0001，而在这种变化可分辨的情况下（即三个模型中的其中两个模型上的缪子比），这种变化带来的是学习器的收益，而不是信息的收益；学习器发现，使用缪子比比构造缪子比更容易，而不是信息本身更容易使用。

4.4 观测和重建损失了什么

关键结果。 与重建规模相关的误差方差占比，从最低十年至第三个十年间由 70–72% 降至 24–26%。

将重建的电子和μ子规模替换为其真实值，并用这些真实值训练同一个学习器，即可得到表2中的预言机列。我们将摘要误差

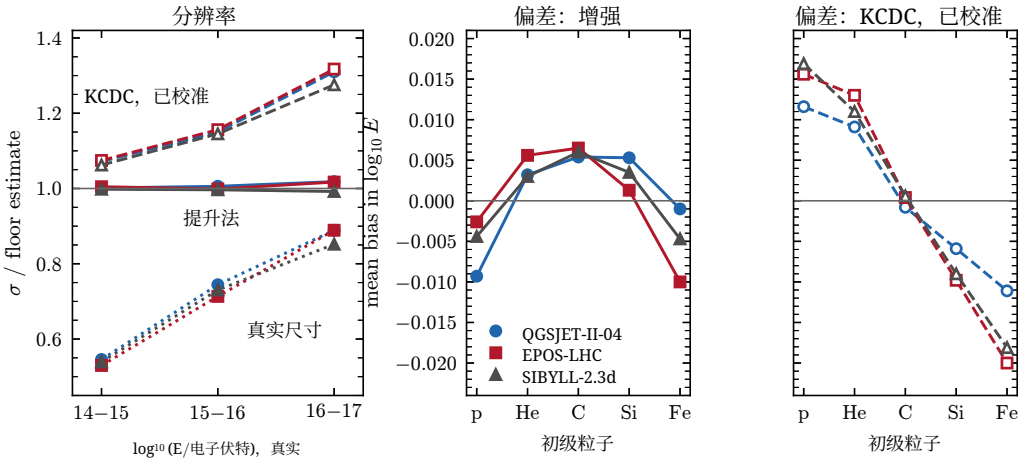


图 6 左：在三种相互作用模型下，九个特征的梯度提升法（实线）、校准后的 KCDC 估计值（虚线，空心标记）以及以真实尺寸为输入的预言机（点线）相对于下限估计的分辨率。中间和右侧：在 10^{15} 至 10^{16} 电子伏特之间，提升法和校准后的 KCDC 估计值的每个初级粒子 $\log_{10} E$ 平均偏差。图中未显示偏差的统计区间

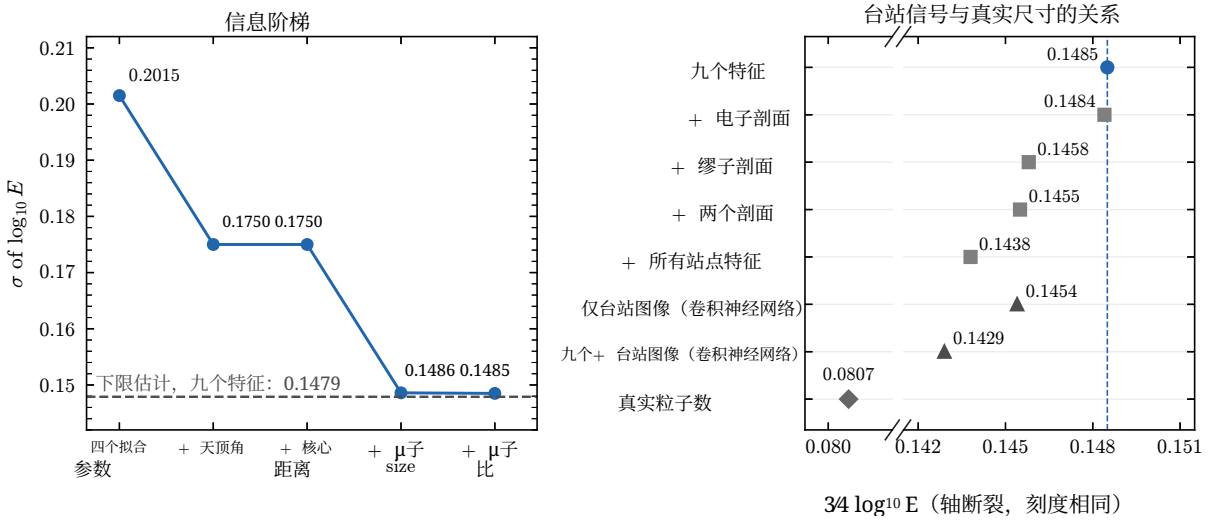


图 7 KASCADE 站点级样本（QGSJET-II-04, 10^{14}–10^{15} 电子伏特）。左：信息阶梯，从拟合结果开始，每次向外添加一个量（“+天顶角”添加 θ 和 $\cos\theta$），并与九个特征的下限估计对比。右：九个特征的分辨率（圆圈、虚线）；加入人工构建的台站表示后（方块）；三种随机种子的卷积神经网络集成在模拟台站图像上的结果，单独使用时以及作为摘要的校正时（三角形）；以及以真实尺寸为输入的预言机（菱形）。横轴采用刻度恒定的断轴。没有任何探测器能够测量真实尺寸；真实尺寸是用于参考重建总量代价的基准，并非完整台站信号估计器所能达到性能的上限

方差中由此消除的部分称为以预言值为参照的误差方差差距，$(\sigma^2 - \sigma^2_{\text{oracle}})/\sigma^2$。它比较的是两个经过训练的估计器，因此衡量的是相较于真实总量，重建总量给这个学习器带来的代价，而不是探测器丢失的所有信息。在低于 10^{15}电子伏特时，该差距为70%至72%；在 10^{15} 至 10^{16}电子伏特之间为45%至49%；在 10^{16} 至 10^{17}电子伏特之间为24%至26%；三种相互作用模型的结果相差不超过4个百分点

在每个十年能区内（图4，右）；数据受限的最高能区不作解释。在低能量下，摘要误差大多与测得的粒子数相关；在较高能量下，误差大多仍与真实尺寸相关，这表明簇射间涨落和未知的初级粒子质量也会产生影响，而无论如何更准确地测量总量，都无法消除这些影响。没有探测器能测量真实粒子数，而

台站信号还包含总量之外的空间和时间信息，因此预言机是一个实用的参考，而非台站信号可能达到的恢复能力上限。

4.5 不知道十年区间时

关键结果。 不知道十年区间时，平台区和损失的能量依赖性依然存在，但绝对分辨率更大。

到目前为止，分析都是在已知的真实能量十倍区间内训练和评分，而实验并不知道这一信息；上文所有逐能区分辨率和下限估计均以 $\lg(E/\mathrm{eV}) \in [a, a+1]$ 为条件，而这种条件限制会减小十倍区间边缘附近的误差。表4取消了这一条件：模型在整个发布数据集上只训练一次，再按真实能量十倍区间评分。在生成的 E^{-2} 能谱上训练时，提升法因高能簇射数量不足而表现不佳：在 10^{16} 到 10^{17}eV之间，其误差为 0.1194 至 0.1261，差于 KCDC 的原始公式（0.0939 至 0.1105）。按 $\lg E$ 中的平坦分布加权后，在 10^{15} 到 10^{16}eV之间，其误差比逐能区数值高 23% 至 25%，在 10^{16} 到 10^{17}eV之间则高 12% 至 15%；最低能区还包含模拟的下边界10^{14} eV。

不考虑十年信息，仍有三个结论成立。在整个范围内，基于生成谱训练的提升法比同一总体的九个特征下限估计高 0.7 至 0.8%（0.1449 至 0.1465；超出部分的联合簇射区间为 0.4 至 1.1%）：其数值可测地高于估计值，但不到百分之一，因此平台区并非窗口造成的假象。随能量增加，重建所损失的误差方差占比也以相同方式下降：在三个有数据的十年区间内分别为 73% 至 75%、50% 至 53% 和 26% 至 28%；已知十年信息时则分别为 70% 至 72%、45% 至 49% 和 24% 至 26%。此外，在 10^{15} 至 10^{17}电子伏特之间，KCDC 的原始估计在均方根误差方面仍比基于平坦分布训练的提升法高 12% 至 58%。因此，按十年划分的下限以及第 4.2 节中与下限的比较，都是针对已知十年信息的结论；实验从摘要中可预期的绝对分辨率，则是平坦列中的数值。

4.6 台站信号能恢复什么

关键结果。 台站信号优于汇总值，但我们尝试的编码器仅弥合了约十分之一的预言机差距。

表5和图8将由模拟台站级可观测量构成的表示加入九个特征中，针对 $\log_{10}(E/\mathrm{GeV}) \in [5,6]$ ，以及在其下方区块中的 $[6,7]$。在 $[5,6]$，中，电磁信号和μ子信号的径向剖面、击中计数、信号总量和到达时间离散度与九个特征一起输入梯度提升模型，在 QGSJET-II-04、EPOS-LHC 和 SIBYLL-2.3d 上分别弥合了相对于预言机的方差差距的 8.8%、7.7% 和 8.0%。其中几乎全部增益都来自缪子剖面：其形状在轻、重初级粒子之间存在差异（图3），而缪子总规模无法捕捉这种差异；在 σ 中，电磁剖面弥合的差距分别为 0.2%、0.2% 和 0.2%。

台站图像。 针对台站图像（16×16 沉积和到达时间网格）训练的卷积网络，作为对九个特征上梯度提升结果的校正，将 σ中的差距缩小了10.5%、9.3%和9.6%，比手工构建的剖面分别好0.0009、0.0008和0.0009。因此，摘要在预测能量方面并不充分；对于每种模型，测量信号恢复了重建所损失信息的大约十分之一。

能量高一个数量级。 对于$\log_{10}(E/\mathrm{GeV}) \in [6,7]$（表5，下半部分），网络缩小了较小差距的7.8%至9.8%，因此这一比例并不强烈依赖能量；而手工构建的特征缩小的差距较少（4.4%至5.1%）：缪子剖面在此处贡献较小（缩小 σ中差距的1.2%至2.3%），电磁剖面的贡献则稍大（0.6%至1.0%）。

其他编码器。更大且构造方式不同的编码器带来的变化很小（附录E中的表 E7，QGSJET-II-04）。一个规模约为卷积神经网络三十倍的残差网络，以及一个每个站点对应一个标记的 Transformer，都使组合结果维持原状。在这些站点上构建的图网络与其他模型不同，它获得了每个站点相对于重建核心的位置，是仅使用图像时表现最好的编码器（两个随机种子的平均值为 0.1432，优于使用九个特征并包含所有手工构造的台站特征进行的提升法）；作为摘要的校正项时，它在 QGSJET-II-04 上弥合了方差差距的 11.2%，在 EPOS-LHC 上弥合了 10.3%，比卷积神经网络多约一个百分点。因此，四个编码器系列对于测量信号所增加的信息量给出了大致相差两个百分点以内的一致结论；这是关于这些编码器的经验性表述，而非一个上界。

表4 在不知道真实能量十年区间的情况下进行审计。模型在整个发布数据集（10^{14}–10^{18}电子伏特，最多 2×10^6 个事件，按簇射划分）上训练一次，并按真实能量十年区间进行评分（lg E的均方根误差），与在每个十年区间内训练的提升法进行比较（表2）。“Generated”：在模拟的 E^{-2} 谱上训练；“flat”：对训练进行加权，使lg E呈平坦分布。KCDC：其估计值分别通过仅使用 E_{rec} 进行提升法校准（平坦权重）以及使用原始值获得。最后两列：归因于重建而损失的误差方差占比（使用真实尺寸的预言机），分别对应已知十年区间和采用平坦全范围训练的情况。粗体表示在不知道十年区间时可用的四个估计器中误差最低者（仅解释各十年区间）

		提升法，九个特征			KCDC估计值		归因于重建的损失	
lg(E/电子伏特)	模型	每十年	生成的	flat	校准后	raw	每十年	flat
14–15	QGSJET-II-04	0.1481	**0.1498**	0.1626	0.1768	0.1945	70%	73%
	EPOS-LHC	0.1485	**0.1499**	0.1628	0.1762	0.1958	72%	75%
	SIBYLL-2.3d	0.1490	**0.1511**	0.1641	0.1770	0.1981	71%	73%
15–16	QGSJET-II-04	0.0869	0.1218	**0.1086**	0.1234	0.1215	45%	50%
	EPOS-LHC	0.0850	0.1172	**0.1044**	0.1203	0.1217	49%	53%
	SIBYLL-2.3d	0.0872	0.1224	**0.1088**	0.1245	0.1269	46%	52%
16–17	QGSJET-II-04	0.0614	0.1261	**0.0696**	0.0871	0.0939	24%	26%
	EPOS-LHC	0.0596	0.1194	**0.0688**	0.0842	0.1028	24%	26%
	SIBYLL-2.3d	0.0623	0.1221	**0.0699**	0.0910	0.1105	26%	28%
17–18	QGSJET-II-04	0.0564	0.3182	0.0653	0.0784	0.0826	17%	18%
	EPOS-LHC	0.0604	0.3197	0.0629	0.0740	0.1063	9%	19%
	SIBYLL-2.3d	0.0583	0.3164	0.0643	0.0747	0.0983	14%	30%

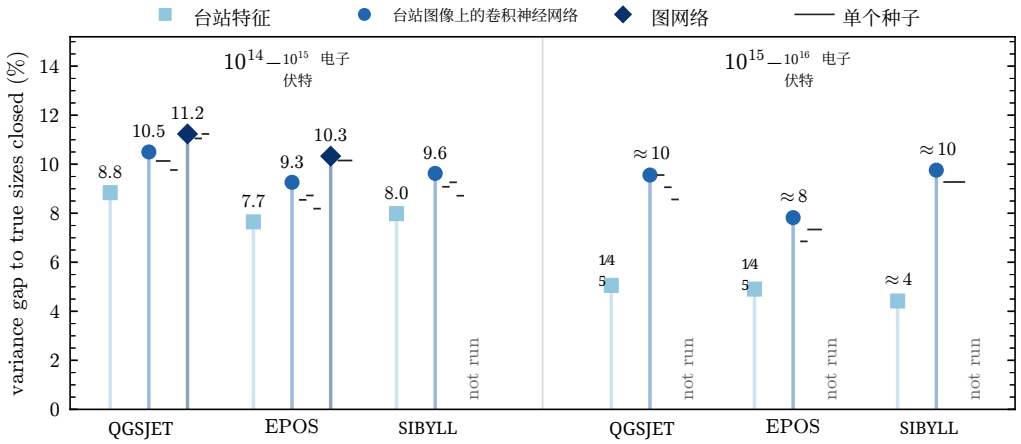


图8 按表示、相互作用模型和能量窗口划分，并分别与摘要结合后，模拟台站信号弥合九特征摘要与预言机之间方差差距的比例。网络数据点表示随机种子集成（卷积神经网络使用三个种子，图网络使用两个种子），短横线表示单个种子，其中一些数据点彼此重合。上方能量窗口的比例以整数显示，因为这些比例使用 σ 四舍五入到四位小数的数值，因此该窗口中的精度约为半个百分点，下方窗口中则约为0.2个百分点。表5以一位小数列出了相同的比例。比例本身没有可用的区间；表5给出了卷积神经网络相较于 σ 台站特征的增益所对应的配对簇射区间。“未运行”表示未执行的图网络运行

核心位置起作用吗？只有图网络获得了相对于重建核心的位置，因此我们在相同预算下，改用台站绝对位置、去除核心位置，重新训练了该网络：仅使用图像时，其结果为 0.1431（对比 0.1432）；作为校正项时为 0.1423（对比 0.1425），弥合了方差差距的 11.6%。绝对坐标对照表明，图网络的优势并不能由其能够访问重建核心来解释；这与局部消息传递所带来的益处相一致

站点，不过其较长的训练时间以及归纳偏置方面的其他差异也不能排除。编码器之间的差异与随机种子间离散度处于同一量级（表 E7）；真正明显偏离零的是每个网络相较于摘要和人工构造的站点特征所取得的提升，表5给出了相应的配对簇射区间。表5中报告的卷积神经网络结果是三个随机种子的集成，而表 E7中报告的是随机种子的均值。我们无法将未能恢复的剩余部分分摊到数组对

表5 在三种强子相互作用模型下，模拟台站级可观测量带来的提升（对于 lg(E/GeV) $\in [5,6]$2，每个发布数据集从中均匀抽取 400，000 个事件，划分为每块 200 个事件；筛选后约有 410，000 个事件；对于 $[6,7]$ 最多从该区间随机抽取 400，000 个事件；按簇射划分）。对九个重建特征使用提升法和下限估计值；使用真实尺寸的预言机；对九个特征加上所有站点特征（径向电子/γ 和缪子剖面、击中计数、总量及时间分布宽度）使用提升法；以及将卷积网络应用于模拟台站图像，并训练其校正提升法的结果（三个随机种子的集成）。最后几列给出站点特征和网络分别弥合的相对于预言机的方差差距比例，以及网络相对于站点特征的变化，并附有配对簇射自助法区间。所有数值均为残差的标准差。粗体表示探测器可测输入中的最低值；其相较于站点特征的提升，其配对区间不包含零。比例使用 σ 四位小数取整，因此在较高区间内可能存在最多约半个百分点的取整误差

lg(E/GeV)	模型	测试	提升法	估计值	下限 预言机	+站点 特征	+图像 网络特征网络	差距已缩小		网络 −特征 （配对区间）
5–6	QGSJET	102，483	0.1485	0.1479	0.0807	0.1438	**0.1429**	8.8%	10.5%	−0.0009 [−0.0011, −0.0007]
	EPOS	102,629	0.1484	0.1477	0.0772	0.1442	**0.1433**	7.7%	9.3%	−0.0008 [−0.0010, −0.0006]
	SIBYLL	103，001	0.1495	0.1491	0.0805	0.1452	**0.1443**	8.0%	9.6%	−0.0009 [−0.0011, −0.0007]
6–7	QGSJET	99，767	0.0868	0.0862	0.0642	0.0858	**0.0849**	5.1%	9.6%	−0.0009 [−0.0010, −0.0008]
	EPOS	89，809	0.0853	0.0844	0.0618	0.0843	**0.0837**	4.9%	7.8%	−0.0006 [−0.0007, −0.0005]
	SIBYLL	92，905	0.0872	0.0865	0.0638	0.0863	**0.0852**	4.4%	9.8%	−0.0011 [−0.0013, −0.0010]

簇射的稀疏采样上——任何站点信号函数都无法弥补这一点——以及更优的信号模型能够提取的信息上。

5 测量的 KASCADE 数据

关键结果。 测量事例的平均缪子数位于每个模型的质子—铁区间内，且 μ子通道的使用方式与模拟中相同；两项检验均无法修正能量标度。

没有真实能量，就无法对测量事例执行步骤 1、2 和 5，也无法运行预言机。但可以进行两项检查。首先，在重建能量的各个十年区间内，将数据中 l og$_{10}\,N_{\rm e}$、log$_{10}\,N_{\mu}$、年龄和天顶角的分布与各相互作用模型进行比较；比较前，将模拟数据从其 E^{-2} 谱重加权为膝区谱（在 $10^{15.6}$电子伏特以下指数为 -2.7，在 [23]以上为 -3.1 ），分别针对纯质子、纯铁，以及由所有已发布事件构成的五种成分混合样本（经过筛选和膝区加权后，每种初级粒子占 11% 至 30%），并将固定 log$_{10}\,N_{\rm e}$ 时的平均 log$_{10}\,N_{\mu}$ 置于各模型的质子至铁标度上。其次，将在每种模拟数据上训练的估计器应用于数据并进行探测（步骤 4）：如果估计器对数据中 μ子通道的依赖方式与其在模拟中相同，那么探测值就会一致。我们首先在从一半 EPOS-LHC 簇射中抽取的伪数据上运行这两项检查；这些伪数据具有已知的谱和成分：在每个十年区间的 μ 子数分布中，伪数据都最接近 EPOS-LHC ；其平均 log$_{10}\,N_{\mu}$ 在每个有数据的 log$_{10}\,N_{\rm e}$区间内，都落在 EPOS-LHC 的质子至铁标度上，且与混合样本的取值相差不超过 0.02 ；伪数据上 μ子通道的探测增量为 0.280，而留出的 EPOS-LHC 模拟数据上为 0.282（三种模型中的最大差异为 0.017）。

数据。 测量数据集包含 389，893，448 个事件，其中 $14 \le$ log$_{10}$($E_{\rm rec}$/电子伏特) ≤ 18；我们随机选取其中的 5.1%。在最低的重建能量处（$14 \le$ log$_{10}$($E_{\rm rec}$/电子伏特) < 14.2），测量事件中只有 0.15% 具有 log$_{10}\,N_{\rm e}$ <3.5，而重加权模拟中的纯质子比例为 38% 至 40%，更重的初级粒子比例则更高。没有任何成分能产生如此少的小簇射，因此测量数据集带有模拟数据集无法再现的簇射大小选择效应，这与触发器相符：数组只有在 log $_{10}\,N_{\rm e} > 4$ [24] 时才达到探测效率达到100%。因此，我们对数据和模拟统一应用 log$_{10}\,N_{\rm e}$ ≥4.0，即全效率阈值，最终保留 13，848，509 个测量事例。于是，共同筛选条件为 $14 \le$lg($E_{\rm rec}$/电子伏特) < 18、天顶角低于 $40°$、簇射核心位于距阵列中心 91 米以内、log$_{10}\,N_{\rm e} \ge 4.0$ 以及有限的重建量；下方在固定电子数下的比较还增加了天顶角低于 $30°$ 的条件。由于该筛选条件针对电子数，因此在最低的重建能量附近，它对重簇射的剔除多于对轻簇射的剔除，而这一点在数据和模拟中相同。上文的闭合检验已使用相同的筛选条件重新进行。

数据与模拟对比。 在每个十年区间以及每个模型中，数据的平均 log$_{10}\,N_{\mu}$ 都介于纯质子和纯铁之间（表6，图9）。在 10^{14}–10^{15},10^{15}–10^{16} 和 10^{16}–10^{17} 电子伏特的区间内，数据的 μ子数

表6 应用共同筛选条件（包括 $\log_{10} N_{\mathrm{e}} \geq 4.0$）后，按重建能量每十年统计的 KASCADE 测量事件与模拟结果。经重新加权以符合膝区谱后，数据以及纯质子、五成分混合和纯铁各模型的平均 $\log_{10} N_{\mu}$；以及数据的 $\log_{10} N_{\mu}$ 分布与各模型之间的 Wasserstein-1 距离。数据为发布数据集的随机 5.1%

			平均 $\log_{10} N_{\mu}$				与数据的距离		
lg(E_{rec}/电子伏特)	事件	模型	data	质子	混合物	iron	质子	混合物	iron
14–15	12,308,663	QGSJET-II-04	3.486	3.456	3.526	3.623	0.031	0.040	0.137
		EPOS-LHC	3.486	3.459	3.521	3.611	0.028	0.034	0.124
		SIBYLL-2.3d	3.486	3.455	3.522	3.612	0.031	0.035	0.126
15–16	1，522，324	QGSJET-II-04	4.156	4.126	4.195	4.248	0.030	0.039	0.092
		EPOS-LHC	4.156	4.135	4.197	4.247	0.021	0.041	0.091
		SIBYLL-2.3d	4.156	4.126	4.193	4.249	0.030	0.037	0.093
16–17	17，325	QGSJET-II-04	5.122	5.037	5.129	5.214	0.085	0.019	0.092
		EPOS-LHC	5.122	5.057	5.140	5.224	0.065	0.022	0.102
		SIBYLL-2.3d	5.122	5.038	5.129	5.213	0.084	0.017	0.091

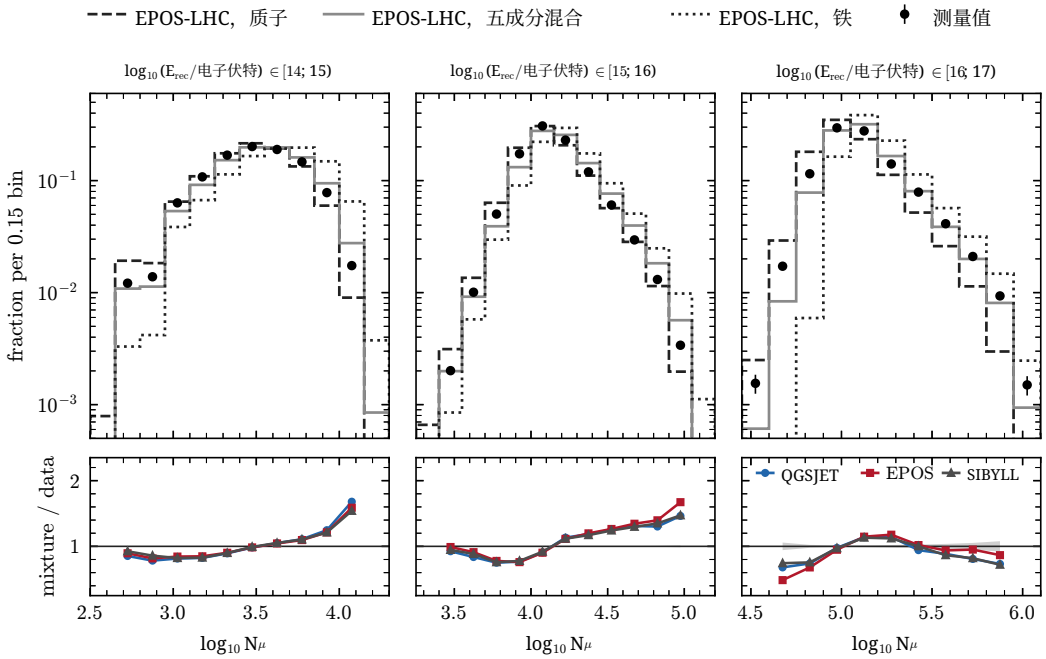


图 9 重建能量分为三个十年区间时，在满足共同筛选条件（天顶角小于 40°）后，测量的 KASCADE 事件（黑色，泊松误差）以及 EPOS-LHC 模拟中质子（虚线）、五成分混合（灰色）和铁（点线）的 $\log_{10} N_{\mu}$ 分布；模拟结果已重新加权至膝区谱。由于该分布在更小区间宽度下具有精细结构，因此区间宽度为 0.15。下方各图：对于三个相互作用模型，在数据占比超过 2×10^{-3} 的区间内，混合结果除以数据；灰色带表示数据的泊松误差，在前两幅图中窄得难以看见，且未显示模拟的统计误差

分布与质子、质子和混合成分的分布最为接近，且在全部三个模型中都是如此。与预期相同，由于真实成分并非这种混合成分，这些距离比闭合测试中的大一个数量级。对于其他量，数据与该混合成分之间的 Wasserstein-1 距离在 $\log_{10} N_{\mathrm{e}}$ 上为 0。015 至 0。065，在年龄上为 0。007 至 0。093，在天顶角上为 0.1° 至 1.2°，涵盖所有十年和模型（每十年及各成分对应的数值见结果文件）。在固定电子数下（图 10），当 $\log_{10} N_{\mathrm{e}}$ 不超过 5.5 时，数据位于从质子到铁范围的 0.04 到 0.32 之间，而混合成分对应的范围为 0.42 到 0.55；当 $\log_{10} N_{\mathrm{e}} \approx 5.25$ 以上时，该数值随电子数增加，在最高区间（197 个事件）达到 0.42 到 0.54。每个区间的数据都落在所有模型的范围内。这是对审计所依赖的模拟的一致性检验，

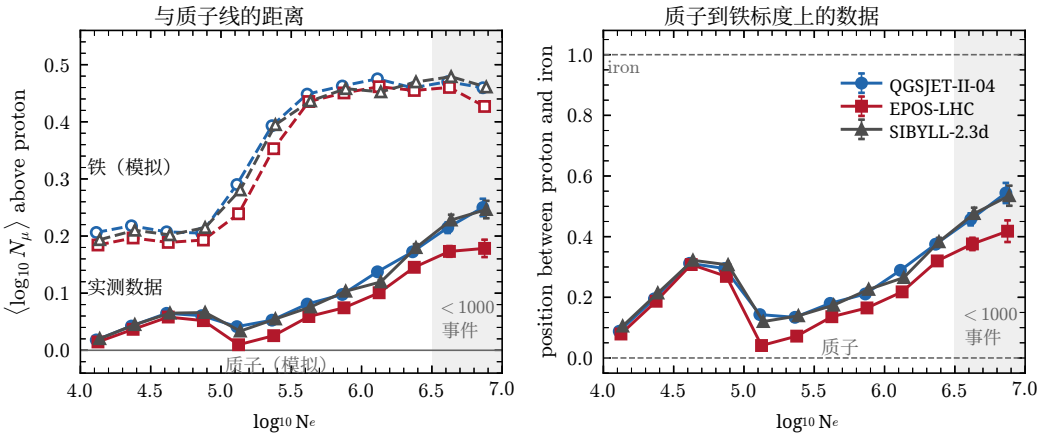


图 10 测量得到的 KASCADE 事件（天顶角小于 30°）。左：数据（实线）和铁（虚线）的平均 $\log_{10} N_\mu$，相对于各模型的质子值。右：数据在各模型质子至铁标度上的位置（0 表示质子，1 表示铁）。误差棒仅表示数据的标准误差。在 $\log_{10} N_e \approx 5$ 与 5.75 之间，质子至铁的区间在 $\log_{10} N_\mu$ 中从约 0.2 扩大到约 0.45，数据相对于质子值的偏移也随之改变；我们尚未确定原因，因此不应将 $\log_{10} N_e \approx 5.1$–5.4 附近的低谷解读为成分差异。灰色：事件数少于 1,000 的分箱

而非成分测量；膝区以下的轻成分，以及由此推测出的膝区以上成分增加，在定性上与 KASCADE 的反卷积结果 [23] 一致。模拟所得的质子到铁区间本身在 $\log_{10} N_e \approx 5$ 至 5.75 之间变宽，原因尚未查明，因此不应将该电子数附近尺度上的结构解读为成分。

对实测数据的探测。 梯度提升模型分别在每种模拟上训练（应用共同筛选条件后，在整个能量范围内，不加权），训练输入为 $\phi = (\log_{10} N_e, s, X_c, Y_c)$，并结合 $M = (\log_{10} N_\mu, \theta)$；随后在其对实测事件（五个随机折）以及重采样为膝区谱的留出模拟簇射所产生的输出上，计算 M 的探测增量。对于固定的训练函数而言，一致意味着数据和模拟经过该函数后呈现相同的 (ϕ, M) 联合结构；这并非对能量标度的检验。该估计器在实测事件上的探测增量为 0.276 至 0.283，在留出模拟事件上为 0.282 至 0.287，差值至多为 0.008，处于闭合检验的尺度范围内。模拟训练的估计器在实测簇射和模拟簇射上都使用 μ子通道，因此，审计在模拟中发现的结构，至少不会与数据相矛盾；这是在没有标签的情况下唯一可检验的一方面。

6 GRAPES-3：摘要即全部内容

关键结果。 基于拟合结果渲染的图像训练的神经网络并不比拟合本身更好；只有测量通道能提供帮助。

GRAPES-3 展示了摘要即全部内容的情形。具有不同归纳偏置的估计器结果一致，而交叉拟合梯度提升法（0.1021）在所有估计器设置下均比九个特征的估计下限高 0.2% 至 1.8%（参考设置为 0.1018，簇射级区间$[0.1006, 0.1030]$）。阶梯呈现出 KASCADE 的形态：天顶角和 μ子总计数降低 σ（分别降低 0.0042 和 0.0164），而望远镜中 16 个 μ子模块计数的分布模式，相较于它们的总计数并未提供可分辨的新信息（配对区间 $[-0.00003, +0.00005]$；在每个事件内置换这些计数也没有影响，$[-0.00006, +0.00002]$）。

拟合结果的渲染并不是一次测量。图11展示了该结构。

此前针对这一样本的一个网络，采用 3.2 百万参数的卷积与 Transformer 阶段混合架构，以“粒子密度、到达时间和 μ子多重数”图像作为输入，达到了与岭回归相当的 $\sigma = 0.1063$ 水平。导出文件不包含每个闪烁体的密度或时间，因此，由这些文件构建的密度图像

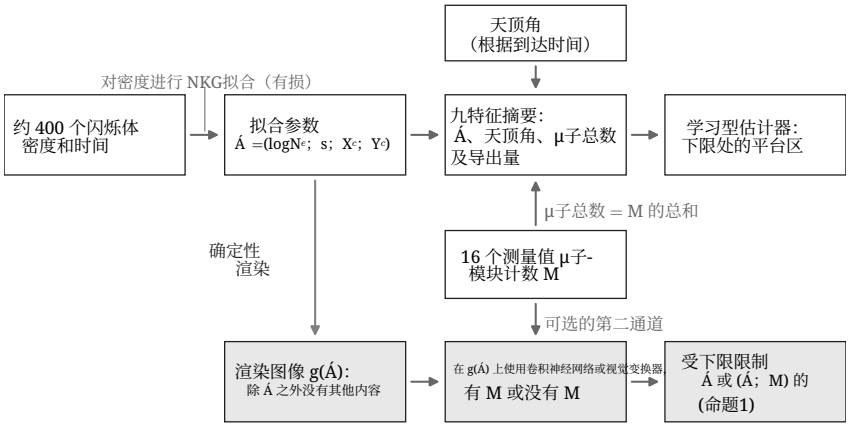


图11 GRAPES-3 的情况。标准重建将约 400 个闪烁体信号替换为四个拟合参数 ϕ；九特征摘要由 ϕ、天顶角和 μ子总数构成。由 ϕ 渲染出的图像是 ϕ的确定性函数，因此其平方损失贝叶斯风险不可能更低（命题1）；只有 ϕ以外的输入（例如 16 个 μ子模块计数 M 或天顶角）才能增加信息

表7 GRAPES-3 渲染实验（60,000 个训练事件）。簇射交叉拟合：每个事件都由一个未见过其簇射副本的模型进行预测（初级粒子）；随机划分：存档协议，会泄漏簇射副本。C′ 和 C′′ 行是之后添加的表格对照组；在运行 A 至 D 行之前，预测列就已记录。四参数下限估计为 0.1219。在交叉拟合条件下，对渲染图像使用卷积神经网络的结果比对四个参数使用梯度提升的结果低 0.0003（配对区间 $[0.0001, 0.0005]$）：这是使用相同信息时不同学习器之间的差异，命题1允许这种情况

	输入和模型	预测值	随机划分	簇射交叉拟合
A	四个拟合参数 ϕ，梯度提升	0.1234	0.1230	0.1230
B	基于由 ϕ 渲染的图像的卷积神经网络	0.1234	0.1226	0.1227
C	基于渲染图像的卷积神经网络 + 16 测量值 μ子计数	0.1063	0.1063	0.1064
C′	ϕ + 16 μ子计数，梯度提升	—	0.1071	—
C′′	作为碳′ +天顶角	—	0.1029	—
D	九个特征，梯度提升	0.1019	0.1022	0.1024

只能是 NKG 函数在探测器位置处的取值，这是拟合的确定性函数（图 12）。

我们重建了该输入，并提前记录了结果（表 7）。命题1指出，渲染图像无法将总体贝叶斯风险降至生成它的四个参数所对应的风险以下；但它并未说明两个特定的有限样本学习器能多好地利用这些信息。根据按簇射进行的无泄漏交叉拟合，实测结果显示，渲染图像上的卷积神经网络达到 0.1227，而四个参数上的梯度提升达到 0.1230；两者都在四参数下限估计 0.1219 的 1% 以内。卷积神经网络低 0.0003 （区间 $[0.0001, 0.0005]$），这反映的是使用相同信息的不同学习器之间的差异，而非新增信息。在随机划分协议的 5 个随机种子中，情况也相同（范围 0.1226 至 0.1228）；视觉Transformer在图像上的表现与卷积神经网络相当；而对像素上的线性模型和全连接模型而言，由于它们还原渲染过程的能力较弱，表现更差。将测量的μ子计数作为第二通道加入后，结果为 0.1064；与九个特征（0.1024）之间剩余的差距来自到达方向，而图像中不包含这一信息。

在 KASCADE 上，台站图像是探测器可观测量，而不是拟合结果的渲染；同类网络的表现优于摘要（第 4.6 节）；而在图像经过渲染的 GRAPES-3 上，情况则并非如此。这是对单个尺寸切片上命题1的佐证性说明，而不是对整个审计的独立验证。在重建之前，对早期网络归档的预测运行的退化探测已经指出了这一点：其输出得分

0.860 作用于拟合参数，0.993 作用于 μ子计数和天顶角，这正是这些输入的函数特征（图 13）。

继承而来的筛选。GRAPES-3 分析通常会接受簇射年龄为 $0.90 \leq s \leq 1.15$ 的事件；这一切割旨在用于全范围重建。在这个狭窄的规模区间内，它保留了 49.1% 的事件。三种比较分别回答不同的问题，我们将它们区分开来（全程采用簇射交叉拟合）。从实际操作来看，应用该切割的分析在保留的事件上得到 $\sigma = 0.1115$，而不应用切割时为 0.1021；各初级粒子的平均偏差离散程度为 0.112，而不应用切割时为 0.060；但这些对应的是不同的总体。对于固定的已接受事件总体，与在全部事件上训练相比，仅在切割范围内训练会使σ 变化 -0.00004 （区间 $[-0.00025, +0.00014]$），这是必然的：切割是摘要的函数，因此将训练限制在该范围内，不可能降低保留事件上的贝叶斯风险。切割的作用是筛选出较难处理的一半事件。九个特征的下限估计在已接受事件上为 0.1106（$[0.1084, 0.1121]$），在被拒绝的事件上为 0.0912（$[0.0898, 0.0927]$），而梯度提升达到 0.0923。因此，切割并不会损害重建；它舍弃了摘要显示最容易重建的事件，这对于希望在该区间获得最佳分辨率的分析而言适得其反。在 KASCADE 上，类似的中心年龄筛选也使其自身事件上的分辨率在重新训练后保持不变（变化至多为 0.1%），但所保留总体的 σ 相差 -10% 至 $+1\%$，且差异方向取决于规模区间。因此，审计会明确筛选依据哪个估计目标进行评判，并在实际应用该筛选的范围内重新测量该目标。

7 讨论

对于类 ***KASCADE*** 阵列。在重建的摘要上，标准学习器已经达到估计下限；在能量范围内以及三种相互作用模型下，其结果与该估计值的精度（约百分之二）一致。至于它们高于还是低于真实下限，以及在这一误差范围内具体相差多少，目前无法确定；在相同的九个特征基础上增加模型容量，预计不会使能量性能提升超过该估计值的精度。这些结论是有条件的，适用于一个已知的十年区间。实际使用中的估计器并不知道所处的十年区间：在表4的全范围训练中，其分辨率在 10^{15} 至 10^{17}电子伏特之间高出12%至25%；而在生成谱上进行训练、且不对高能部分加权时，在该区间内会失效。摘要在低能量处损失最大，因为此处规模的测量最不准确；在我们尝试的编码器下，模拟的台站级可观测量只能弥补其中很小一部分损失。更好的台站信号模型能否恢复更多性能仍是未知数；预言机展示了可提升空间有多大，但并未说明其中有多少能够实现。

相互作用模型。 这三个模型在μ子含量和簇射发展方面存在差异，其绝对分辨率也略有不同（表2）。但结果的结构并无不同：在估计下限处的平台区（在其精度范围内）、阶梯、损失的能量依赖性，以及台站信号带来的小幅恢复，在三个模型中都成立。因此，审计的结论并非某一个模型的假象，而基于这些估计器开展的任何成分分析都继承了通常的模型依赖性。

超越这些数组。命题1适用于拟合参数经过分布、并据此重建类图像输入的任何场景；成本低廉的保障措施是逐通道记录哪些输入是测量值、哪些是计算得到的。下限估计、阶梯和探针迁移属于程序，其设置必须重新确定；没有任何数值可以直接迁移。审计成本很低：在笔记本电脑CPU上，对一个测试平台进行下限估计、阶梯和探针迁移只需几分钟，而唯一需要进行的网络训练，就是对正在接受审计的输入进行的训练。

局限性。下限是一个估计值，而非界限；在我们的半合成测试中，其簇射区间覆盖真实下限的频率低于名义水平。此处测得的准确度针对的是一个模拟真实目标的目标，而非真实目标本身：低于 10^{16}电子伏特时，其设置使与提升法的比较结果最多变化 2。2%（仅曼哈顿度量最多变化 3。2%）；在 10^{16}至10^{17}电子伏特之间，变化幅度为 -8.1 至 $+3.9$%；在基准测试中，八个维度上的误差最多为 1。9%。在数据受限的最高能区，它甚至低于预言机（表2），而摘要的任何真实下限都不可能做到这一点，因此在事件稀少时会产生向下偏差。结论针对的是能量的平方损失预测

；在此不增加信息的通道，对成分研究或许有用。站点级分析涵盖两个十年（$\log_{10}(E/\mathrm{GeV}) \in [5,7]$）；高于$10^{17}$ 电子伏特时， 200×200米×^{2} 的KASCADE数组收集到的事件过少，而KASCADE-Grande在KCDC中的数据和模拟则是进一步研究的途径。预言机使用的是没有任何探测器能够测量的粒子数。这两个测试平台都是用CORSIKA模拟的地面数组，因此第二个数组在探测器方面独立，但在领域方面并不独立；其模拟还与主要的KASCADE数据集共用了QGSJET-II-04。合成基准测试是唯一在该数据集之外进行的检验。GRAPES-3样本只涵盖一个尺寸区间；其存档网络的代码和权重已丢失，因此我们对该网络输入的描述只是推测，但重建结果和探查结果为此提供了支持。

8 结论

当能量估计器基于拟合得到的簇射摘要进行训练时，摘要会限制它的能力，而模型侧诊断无法揭示这一极限。我们通过审计将两者区分开来。

在采用三种强子相互作用模型的 KASCADE模拟中，基于九特征摘要的学习型估计器在 10^{14}至 10^{16}电子伏特范围内，与估计下限的差距约在百分之二以内；在达到 10^{17}电子伏特之前，差距也处于估计值较宽的不确定范围内。我们在半合成目标上测量了该估计值的偏差和区间覆盖不足，但无法在真实目标上测量这些指标。

观测与重建造成的误差方差损失占比，在这一范围内从约 72% 降至约 26%，且对每种模型都呈现相同趋势。在已知的十年能量区间内，合作组的固定公式估计值比估计下限高出 6% 至 32%，且能量越高，高出的幅度越大。在整个能量范围内训练时，标准学习器仍至少能将其改进 12%。即使不知道所属的十年能量区间，这些结果仍然成立，不过绝对分辨率会更大。模拟的台站级可观测量在所研究的两个十年能量区间内，都能弥补约十分之一的差距；对于 GRAPES-3，由拟合结果生成的输入并不比拟合本身更好。

除了对 KASCADE 测量数据进行的无标签检验外，所有结果都基于 CORSIKA 模拟，因此，在不同探测器和相互作用模型上的复现，并不构成超出该模拟领域的验证在扩大模型或重新设计其输入之前，应先估计这种表征所允许的范围，并记录哪些输入是测量值。

致谢。 感谢 GRAPES-3合作组允许我们使用本研究所用的模拟事件，也感谢 KASCADE 宇宙线数据中心公开发布模拟和数据。

声明

经费。 本研究未获得任何资助机构的专项经费。

利益冲突。 作者声明不存在与本文内容相关的利益冲突。

数据可用性。 KASCADE模拟（QGSJET-II-04、EPOS-LHC 和 SIBYLL-2.3d 数据集，KASCADE 组分）及实测数据可通过 KASCADE 宇宙线数据中心公开获取（https://kcdc.iap.kit.edu） [11, 12]；各站点的位置列于其模拟手册中。经 GRAPES-3合作组许可，两个 GRAPES-3模拟文件（171，776 个事件）及存档的预测表已发布于 https://doi.org/10.5281/zenodo.23096836 [25]。

代码可用性。 分析代码、从数据重新生成每个数字、表格和图的流水线，以及检查正文中每个数字是否与结果文件一致的脚本，均位于 https://github.com/SamarjeetMalik/summary-representation-limits；本稿件所使用的版本标记为`epjc-submission-v5`。README 说明了如何在不使用数据的情况下，根据存储的结果文件重新生成每个表格和数字宏。

作者贡献。 两位作者共同构思并设计了这项研究。Mugdha Raje Ashiya 完成了 GRAPES-3 分析。Samarjeet Malik 完成了 KASCADE 分析，编写了分析软件并起草了稿件。两位作者都对稿件进行了批判性修订，并阅读和批准了最终版本。

附录A 证明

对于子-σ-代数 $\mathcal{G}$ ，平方损失贝叶斯预测器为 $\mathbb{E}[y \mid \mathcal{G}]$；对于嵌套的 $\mathcal{G} \subseteq \mathcal{H}$ ，塔式性质给出

$$\begin{aligned}\mathbb{E}\big[(y-\mathbb{E}[y \mid \mathcal{G}])^2\big] &= \mathbb{E}\big[(y-\mathbb{E}[y \mid \mathcal{H}])^2\big] \\ &\quad + \mathbb{E}\big[(\mathbb{E}[y \mid \mathcal{H}]-\mathbb{E}[y \mid \mathcal{G}])^2\big],\end{aligned} \tag{A1}$$

因为在给定 $\mathcal{H}$ 的条件下，交叉项为零。

命题*1.* $\sigma(g(T)) \subseteq \sigma(T)$; 将 $\mathcal{G}=\sigma(g(T))$ 和 $\mathcal{H}=\sigma(T)$ 代入 (A1)。下确界由条件期望取得，因此它们的差就是第二项，该项非负，并且当且仅当 $\mathbb{E}[y \mid g(T)]=\mathbb{E}[y \mid T]$ 几乎处处成立时为零。

命题 *2.* 等式由 (A1) 取 $\mathcal{G}=\sigma(T)$、 $\mathcal{H}=\sigma(T,M)$ 得出。固定 (T,M)，令 P 和 Q 分别为在给定 (T,M) 和给定 T 的条件下 y 的条件分布。令 $c=(a+b)/2$,

$$\begin{aligned}|\mathbb{E}_P y-\mathbb{E}_Q y| &= \Big|\int (y-c)\,\mathrm{d}(P-Q)\Big| \\ &\le (b-a)\,\mathrm{TV}(P,Q) \\ &\le (b-a)\sqrt{\mathrm{KL}(P\|Q)/2}\end{aligned}$$

根据 Pinsker 不等式， [26],, 其中 $\mathrm{TV}(P,Q)=\sup_A |P(A)-Q(A)| = \frac{1}{2}\|P-Q\|_1$，KL 的单位为纳特；对其平方并在 (T,M) 上取平均，其中 $\mathbb{E}[\mathrm{KL}(P\|Q)] = I(y;\,M \mid T)$，即可得出该界。

附录B GRAPES-3协议

每个CORSIKA簇射都会投射到多个核心位置，因此随机划分事件会导致测试簇射的副本泄漏到训练集中。删除训练集中那些簇射出现在测试集中的事件可以消除泄漏，但会使训练样本偏向投射次数较少的簇射。因此，我们按簇射进行交叉拟合：将簇射分配到五折，并用其余四折训练的模型预测每一折中的事件。随机划分下的泄漏最多为 σ 中的 0.0009，且结论不取决于所用方案。表B1列出了九个特征上的估计器。

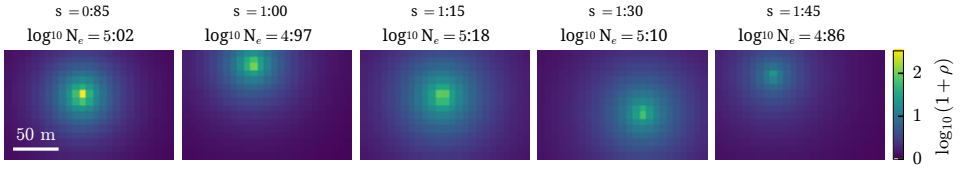


图12 根据NKG拟合生成的图像，采用矩形 25×16 图像网格，间距为8米，展示了簇射年龄逐渐增加的五个事件 s；这些事件的规模（已标注）和核心位置也各不相同。 ρ 为 m^{-2} 中的NKG密度。这些图像看起来像探测器图像，但除四个拟合参数外，不包含任何探测器信号

附录C 已知真值基准测试

测量值 $X \in \mathbb{R}^{32}$ 服从协方差为 $\Sigma = Q\Lambda Q^{\top}$ （Q 随机正交矩阵，$\Lambda_{jj}=j^{-1/2}$）的高斯分布，摘要是一个固定的有损线性映射 $T=AX\in\mathbb{R}^6$ （ A 的元素取自 $N(0,1/32$；嵌套子集是其前 k 个坐标），而 $y = h(w^{\top}X) + \tau(0.5+V)\xi$，其中 $h(u) = u + 0.5\sin 2u$, w归一化，使 $\mathrm{Var}(w^{\top}X) = 1$, $\tau = 0.30$, $\xi \sim N(0,1)$；另有一个测量通道 $M = m^{\top}X + 0.30\,\eta$，其 $m = w + 0.8\,\zeta/\sqrt{32}$ (η, ζ为标准正态分布)，包含关于 $w^{\top}X$的信息；以及一个噪声尺度

表B1 九特征摘要上的估计器和下限估计，均在相同的 34,356 个 GRAPES-3 评估事件上进行评分。 σ 是$\hat{y}-y$ 的标准差，其中 $y=$ 为$_{10}\log E/$（GeV）。对于估计器，两列分别对应 B 节中的两种训练方案：随机补集（五个随机种子的均值），以及按簇射进行的五折交叉拟合，在这种拟合中，没有任何模型会看到其评分所针对的簇射的副本。下方区块给出了 Gamma 检验对相同特征的平方损失贝叶斯风险平方根的估计值$\hat{\sigma}_{\mathrm{B}}$，并附有 95%置信区间；这些方法不训练回归器，所得结果是估计值而非界限。除最后一行外，区间均通过对簇射进行重采样获得。

	随机划分	簇射交叉拟合
Learned estimators, nine features		
岭回归（10 个系数）	0.1066	0.1067
随机森林	0.1034	0.1036
梯度提升	0.1019	0.1021
多层感知机 64×64	0.1022	0.1031
XGBoost	0.1024	—
k-最近邻（$k=50$）	0.1041	—
References		
尺寸–年龄校准（2 个系数）	0.1284	—
已存档的卷积神经网络/视觉变换器，渲染后的输入（佐证性）	0.1063	—
Estimated floor $\hat{\sigma}_{\mathrm{B}}$, nine features (Gamma test)	估计值	区间
以评估事件作为查询，近邻来自所有其他簇射	0.1018	$[0.1006, 0.1030]$
仅评估事件	0.1007	$[0.0995, 0.1018]$
评估查询，六个非冗余坐标	0.1015	$[0.1003, 0.1027]$
全部 171,776 个事件（事件级区间）	0.1003	$[0.0998, 0.1009]$

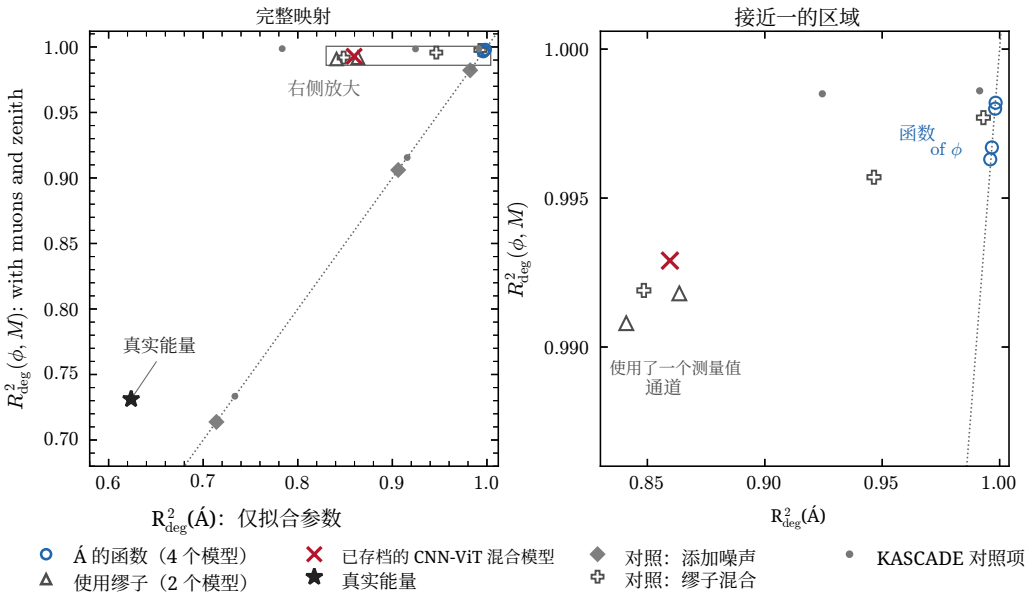


图 13 退化探测。除小灰点外，每个点均为 GRAPES-3 的输出或参考值（“存档”指较早的 CNN-ViT 混合模型所保存的预测），仅基于拟合参数（横轴）以及结合 μ子计数和天顶角（纵轴）进行探测；右图放大了 1 附近的区域。 ϕ 的函数位于右上方，但少量混合成分也可能落在其旁边（10% 的 μ子混合仅凭 ϕ 得分为 0.993），因此高分是必要条件，但并非充分条件；添加噪声会使输出沿对角线向下移动；来自 ϕ 之外的信息会使其在顶部向左移动。真实能量并非这些输入的函数，因此在两个轴上的得分都低于 1。灰点表示 KASCADE 上相应的对照项，其 μ子混合也使用 θ；位于 (0.78,1.00) 的孤立点是其 50% 混合。这里， M 指 GRAPES-3 上的 16 个 μ子计数及天顶角，而 KASCADE 上则为 $(\log_{10} N_\mu, \theta)$

通道 $V \sim U(0、1)$。共有 170，000 个样本，最后 34，000 个留作测试集。CNN 行的渲染图像是在 16×25 格点上的一个高斯斑点，其振幅、宽度和中心是 T 的四个标准化坐标的平滑函数，另两个坐标则决定一个线性梯度；其中不含任何测量值。图像的维度远超估计器的适用范围，因此省略了 Gamma 检验行；根据命题 1，其真实下限与 T 相同。每个贝叶斯风险都可化为高斯–埃尔米特积分。Gamma 检验估计值与摘要及其嵌套子集（一至六个维度）的真实值相差不到 0。5%；加入 M 后低估 0。9%，在八个维度中加入 V 后低估 1。9%，而对 32 维测量值则高估 25%（表 C2，图 14）。对 T 使用梯度提升，结果比真实下限高 0。5%。对于测量值，多层感知机达到 0。318，而下限为 0。312；梯度提升则停在 0。449：这是一个受学习器限制的情形。审计正是为了区分这种情况，而在此处，只有已知真实值才能解决这一问题。六维摘要在维度上与 KASCADE 的六个非冗余特征相符，因此该基准测试验证的是估计器在相关低维范围内的表现，而非其普遍适用性；它在实际 KASCADE几何结构中的准确度另行测量（表 D6）。

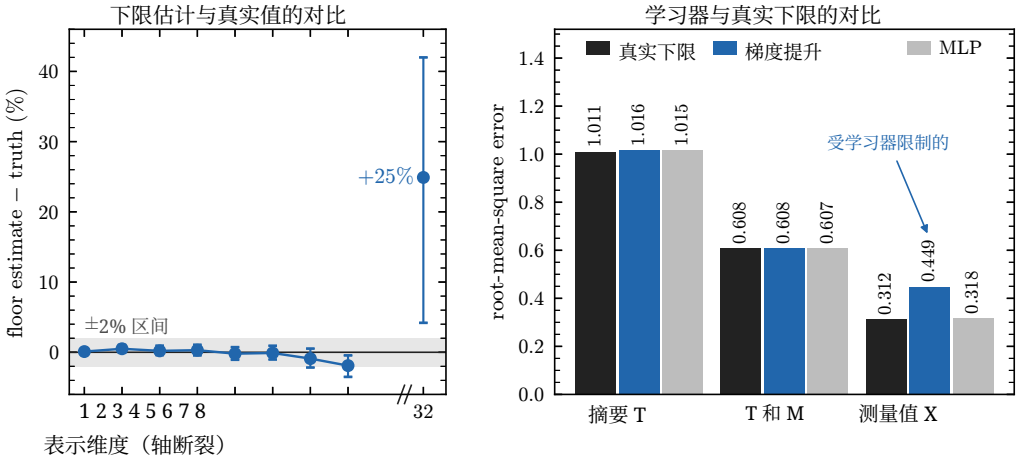


图 14 已知真值基准测试。左：随着表示从摘要的一个坐标扩展到再增加两个通道（八个维度）的摘要，再到原始测量值（32 个维度，轴断开），Gamma 检验下限估计相对于真实下限的误差及其存储的区间。 ±2% 区间仅作视觉参考；八个维度时的区间低于零，而 32 个维度时的区间高于零（+4% 至 +42%）。右：梯度提升和多层感知机相对于真实下限的表现；对于原始测量值，提升法受学习器限制。数值来自有限测试样本，因此学习器输出的数值可能略低于真实下限

表 C2 在贝叶斯风险已知的基准测试上进行的审计。对于每种表示：真实下限 $\sqrt{R^{\star}}$、Gamma 检验估计值及其相对误差，以及最佳学习型估计器（梯度提升或多层感知机）的均方根误差。最后一行是基于 T 计算所得图像的卷积神经网络；其下限为 T 的下限。

表示	真实下限	$\hat{\sigma}_{\mathrm{B}}$	误差	最佳学习结果
摘要 T （六个数字）	1.011	1.010	−0.1%	1.015
T 和测量通道 M	0.608	0.603	−0.9%	0.607
T, M 和噪声尺度通道 V	0.608	0.596	−1.9%	0.608
测量值 X （32 个数字）	0.312	0.390	+24.9%	0.318
由 T 渲染的图像，卷积神经网络	1.011	—	—	1.029

附录D 阶梯和底限敏感性

表 D3 按十年给出了阶梯；其最后一列使用六个非冗余特征（拟合、天顶角和 μ子尺寸），因此在事件数量较少时，结果可能与表2中使用九个特征的提升法列略有不同。

表 D4 将梯度提升与下限估计进行比较，表 D5 则给出了每种设置下的估计值。表 D6 衡量了每种设置在相同样本中、具有已知下限的半合成目标上的准确度。

表 D3 按真实能量十倍区间划分的阶梯： σ 对四个拟合参数进行提升法，并加入天顶角和 μ子尺寸

lg(*E*/电子伏特)	模型	fit	+ 天顶角	+ μ子尺寸
14–15	QGSJET-II-04	0.2017	0.1753	0.1481
	EPOS-LHC	0.2010	0.1743	0.1486
	SIBYLL-2.3d	0.2012	0.1755	0.1491
15–16	QGSJET-II-04	0.1579	0.1149	0.0870
	EPOS-LHC	0.1553	0.1121	0.0851
	SIBYLL-2.3d	0.1553	0.1144	0.0873
16–17	QGSJET-II-04	0.1213	0.0808	0.0615
	EPOS-LHC	0.1155	0.0755	0.0600
	SIBYLL-2.3d	0.1187	0.0794	0.0626
17–18	QGSJET-II-04	0.1022	0.0720	0.0573
	EPOS-LHC	0.1063	0.0698	0.0622
	SIBYLL-2.3d	0.1058	0.0720	0.0578

表 D4 KASCADE模拟中梯度提升相对于下限估计的结果，针对站点级样本及各真实能量十倍区间：其分辨率、参考下限估计、相对于该估计值的超额百分比（附联合簇射区间）、表 D5 中所有估计器设置下超额的范围，以及在同一样本中以具有已知下限的半合成目标测得的参考估计偏差（表 D6）。偏差为正表示估计值偏高，因此梯度提升相对于真实下限的超额比相对于该估计值的超额大约高出相应数值。簇射区间仅反映抽样，应视为不确定性的下限

样本	模型	提升法	下限估计（参考）	高于下限估计的提升法（%） ref.	区间	所有设置	估计值的偏差（基准测试，%）
5–6（站点）	QGSJET	0.1485	0.1477	+0.6	[+0.0, +1.2]	[−0.2, +1.3]	+0.3
	EPOS	0.1484	0.1473	+0.8	[+0.1, +1.4]	[−0.6, +1.0]	+0.3
	SIBYLL	0.1495	0.1488	+0.5	[−0.1, +1.1]	[−1.1, +0.5]	+0.2
14–15	QGSJET	0.1476	0.1482	−0.4	[−1.1, +0.3]	[−1.2, −0.2]	+0.4
	EPOS	0.1483	0.1480	+0.2	[−0.4, +0.9]	[−0.7, +0.6]	+0.1
	SIBYLL	0.1500	0.1492	+0.5	[−0.2, +1.2]	[−0.4, +0.9]	+0.4
15–16	QGSJET	0.0866	0.0861	+0.5	[−0.3, +1.3]	[−2.2, +0.9]	+0.5
	EPOS	0.0853	0.0851	+0.2	[−0.6, +1.2]	[−2.0, +1.5]	+0.6
	SIBYLL	0.0875	0.0866	+1.0	[+0.1, +1.9]	[−1.7, +1.1]	+0.5
16–17	QGSJET	0.0614	0.0603	+1.8	[−0.7, +4.2]	[−6.7, +3.7]	+1.7
	EPOS	0.0596	0.0586	+1.7	[−1.8, +5.8]	[−7.0, +3.9]	+1.7
	SIBYLL	0.0623	0.0628	−0.9	[−4.2, +2.1]	[−8.1, +0.9]	+2.4

表 D5 KASCADE模拟中下限估计的敏感性，针对站点级样本（lg(E/GeV) ∈ [5, 6]）以及每个真实能量十倍区间（lg(E/电子伏特), 另取一个随机均匀样本，最多包含 300，000 个事件）。参考设置：将测试事件作为查询，从其他簇射中选取近邻，使用标准化的九个特征、欧几里得度量以及 $K = 10$；其区间通过对簇射重采样得到。接下来的各列每次改变一项设置：邻居范围 K、在 $K = 10$ 下的三种特征缩放方式（标准化、稳健缩放、秩）、曼哈顿度量，以及以六个非冗余坐标替代九个坐标；这些设置均通过了已知噪声对照。“直接1近邻”是 $\sqrt{\gamma}$(1)，未采用 δ 中的外推，因此会因近邻间回归函数的变化而产生偏高偏差

样本	模型	参考	簇射区间	$K = 5$–50	3 种缩放	曼哈顿	六配位	直接1近邻
5–6（站点）	QGSJET	0.1477	[0.1467, 0.1485]	0.1473–0.1481	0.1476–0.1477	0.1480	0.1478	0.1497
	EPOS	0.1473	[0.1465, 0.1482]	0。1469–0。1482	0.1473–0.1484	0.1480	0.1475	0.1493
	SIBYLL	0.1488	[0.1479, 0.1497]	0.1488–0.1494	0.1488–0.1505	0.1490	0.1493	0.1511
14–15	QGSJET	0.1482	[0.1472, 0.1494]	0.1478–0.1484	0.1478–0.1494	0.1485	0.1478	0.1499
	EPOS	0.1480	[0.1470, 0.1490]	0。1476–0。1484	0.1480–0.1489	0.1487	0.1487	0.1502
	SIBYLL	0.1492	[0.1481, 0.1504]	0.1488–0.1496	0.1492–0.1496	0.1491	0.1494	0.1511
15–16	QGSJET	0.0861	[0.0854, 0.0868]	0.0861–0.0865	0.0860–0.0876	0.0862	0.0860	0.0898
	EPOS	0.0851	[0.0843, 0.0858]	0.0847–0.0851	0.0850–0.0865	0.0848	0.0851	0.0885
	SIBYLL	0.0866	[0.0859, 0.0874]	0.0866–0.0867	0.0866–0.0887	0.0870	0.0870	0.0902
16–17	QGSJET	0.0603	[0.0588, 0.0616]	0.0601–0.0608	0.0603–0.0634	0.0606	0.0601	0.0679
	EPOS	0.0586	[0.0560, 0.0611]	0.0585–0.0606	0.0586–0.0623	0.0605	0.0592	0.0685
	SIBYLL	0.0628	[0.0608, 0.0650]	0.0622–0.0628	0.0617–0.0661	0.0632	0.0620	0.0713

表 D6 KASCADE几何中的下限估计值准确度。对于每个样本，保留真实的九个特征，并将目标替换为 $g(X) + s(X)\,\xi$；其中 g 和 s 根据真实目标进行学习（条件均值和残差平方采用梯度提升）， ξ 为标准正态分布，每个 CORSIKA簇射抽取一次；随后真实下限$\sqrt{\langle s^2 \rangle}$ 已知。表 D5 中每种设置下，20 次噪声重复试验所得下限估计值的平均偏差（百分比），以及参考设置的 95% 簇射区间覆盖率

样本	模型	真实地板参考值		下限估计值的偏差（%） $K = 5$–50	稳健	rank	曼哈顿	六配位	1-NN	覆盖率 区间的
5–6（站点）	QGSJET	0.1481	+0.3	+0.3至 +0.5	+0.2	+0.6	+0.5	+0.3	+1.6	85%
	EPOS	0.1479	+0.3	+0.3至 +0.5	+0.3	+0.6	+0.4	+0.2	+1.6	80%
	SIBYLL	0.1496	+0.2	+0.0至 +0.4	+0.1	+0.6	+0.4	+0.1	+1.4	90%
14–15	QGSJET	0.1486	+0.4	+0.4至 +0.5	+0.4	+0.8	+0.5	+0.3	+1.8	85%
	EPOS	0.1484	+0.1	−0.0至 +0.4	+0.0	+0.5	+0.3	+0.2	+1.6	95%
	SIBYLL	0.1495	+0.4	+0.2至 +0.4	+0.3	+0.8	+0.5	+0.4	+1.6	85%
15–16	QGSJET	0.0870	+0.5	+0.2至 +0.5	+0.4	+2.0	+0.6	+0.4	+4.4	95%
	EPOS	0.0849	+0.6	+0.6至 +0.6	+0.6	+2.1	+0.6	+0.3	+4.9	60%
	SIBYLL	0.0874	+0.5	+0.4至 +0.6	+0.4	+2.1	+0.6	+0.3	+4.5	90%
16–17	QGSJET	0.0617	+1.7	+1.4至 +2.1	+1.2	+6.8	+2.4	+2.4	+13.8	75%
	EPOS	0.0597	+1.7	+0.9至 +3.0	+1.5	+6.7	+2.2	+1.7	+17.6	90%
	SIBYLL	0.0627	+2.4	+2.4至 +4.1	+1.9	+7.5	+2.4	+2.5	+17.1	70%

附录E 站点图像编码器

小型卷积神经网络以 16×16 台站网格为输入，包含四个通道（电磁沉积、μ子沉积、到达时间、击中标志），并按照预先选定的固定训练轮次数进行训练。残差网络和台站Transformer采用相同的输入。图网络将252个站点视为节点，使用上述四种信号以及它们相对于重建核心的位置（Δx、 Δy 和距离），在每个站点与其八个最近邻之间传递消息四次，并通过注意力进行池化。三个较大的编码器最多训练40轮（图网络）或20轮，并在训练簇射的验证集上进行早停；图网络的每次训练中，最佳验证轮次都远早于最后一轮，而首次将训练限制为20轮时，模型在训练结束时仍在持续改进（两份训练日志均保存在分析记录中）。将这些模型用作基于九个特征的梯度提升校正项时，每个模型都初始化为输出零。

表 E7 展示了 QGSJET-II-04 在 $\log_{10}(E/\mathrm{GeV}) \in [5, 6]$ 条件下的结果，并给出了不同随机种子的均值和范围（表5给出了随机种子集成）。在校正模式下，网络学习的是基于九个特征的梯度提升法的残差；其在训练簇射上的预测采用折外方式计算。所有网络均使用 AdamW，权重衰减为 10^{-4}，批量大小为 512，并采用余弦学习率调度（初始学习率为 10^{-3}

卷积神经网络使用 3×10^{-4}，残差网络使用 5×10^{-4}，Transformer 和图网络使用 ）；卷积神经网络训练 20 个训练轮次，早停所用的验证集占训练簇射的十分之一。表 E7 的最后一行展示了同一图网络使用台站绝对位置而非相对于重建核心的位置（第 4 节.6）的结果。

表 E7 测量值 KASCADE 图像的编码器。“仅图像”表示单独使用编码器；“校正”表示将编码器作为零初始化校正项，用于基于九个特征的梯度提升（0.1485）。均值及范围按随机种子计算（第 4 节的卷积神经网络使用三个种子，其余模型使用两个）。卷积神经网络的结果报告为预先固定的最终训练轮次；较大的编码器则在训练簇射的验证集上采用提前停止。最后一列表示校正项弥补的、相对于蒙特卡洛真值预言器的方差差距比例。

编码器	size	仅图像	范围	校正	范围	差距闭合
小型卷积神经网络	0.2 M	0.1474	0.1469–0.1480	0.1432	0.1431–0.1433	10%
残差网络	6.3 M	0.1458	0.1458–0.1458	0.1435	0.1435–0.1435	9%
台站Transformer	1.2 M	0.1459	0.1459–0.1459	0.1431	0.1431–0.1431	10%
图网络	0.4 M	0.1432	0.1432–0.1433	0.1426	0。1425–0。1426	11%
图网络、绝对位置	0.4 M	0.1431	0.1431–0.1432	0.1424	0.1424–0.1425	11%

参考文献

[1] Kamata, K.，Nishimura, J.：电子簇射的横向结构函数和角结构函数。《理论物理学进展增刊》**6**，93–155（1958）https://doi.org/10.1143/PTPS.6.93

[2] Greisen， K.：宇宙线簇射。《核科学年度评论》**10**，63–108（1960）https://doi.org/10 .1146/an nurev.ns.10.120160.000431

[3] Kuznetsov, M.Y., Petrov, N.A., Plokhikh, I.A., Sotnikov, V.V.：基于 KASCADE 实验数据和机器学习的宇宙射线元素组成组能谱。《宇宙学与天体粒子物理学杂志》**2024**(05)，125（2024）https://doi.org/10.1088/1475-7516/2024/05/125

[4] Kuznetsov, M.Y.，*et al.*：利用 KASCADE 实验数据分析宇宙射线质量成分的机器学习方法。《仪器学报》**19**，01025（2024）https://doi.org/10.1088/ 1748-0221/19/01/P01025

[5] Zhang, H.，*et al.*：基于LHAASO-KM2A探测器的宇宙射线成分无关能量重建方法。《物理评论D》**106**，123028 (2022) https://doi.org/10.1103/PhysRevD.106. 123028

[6] Erdmann, M., Glombitza, J., Walz, D.：基于深度学习的宇宙射线诱发的广延空气簇射重建。《天体粒子物理学》**97**, 46–53 (2018) https://doi.org/10.1016/j.astropartphys.2017.10.006

[7] Aab, A.，*et al.*： $X_{\max}$ 利用皮埃尔·奥杰天文台的水切伦科夫探测器进行簇射最大值的深度学习重建。《仪器学报》**16**, 07019 (2021) https://doi.org/10. 1088/1748-0221/16/07/P07019

[8] Aartsen, M.G.，*et al.*：基于 IceTop 和 IceCube 三年数据研究拍电子伏特至艾电子伏特范围内的宇宙射线能谱与成分。《物理评论D》**100**, 082002 (2019) https://doi.org/10.1103/PhysRevD.100.082002

[9] Sarkar, S., Talwar, M., Mohanty, P.K.：GRAPES-3 实验的深度学习宇宙射线能量重建流水线。载于：第39届国际宇宙射线大会（ICRC2025）论文集。PoS，第 ICRC2025 卷，第 383 页（2025）。https://doi.org/10.22323/1.501.0383 。arXiv:2607.07265

[10] Antoni, T.，*et al.*：《KASCADE 宇宙射线实验》。《核仪器与方法物理研究A辑》**513**，490–510（2003）https://doi.org/10.1016/S0168-9002(03)02076-X

[11] Haungs, A., *et al.*：KASCADE 宇宙线数据中心 KCDC：开放获取天体粒子物理研究数据。《欧洲物理学杂志C》**78**，741（2018）https://doi.org/10.1140/epjc/ s10052-018-6221-2[12] Wochele, J., Kang, D., Wochele, D., Haungs, A.：《KCDC 模拟手册：KASCADE、GRANDE 和 CALORIMETER 模拟》，版本 V.02.3。卡尔斯鲁厄理工学院，https://kcdc.iap.kit.edu（2025）[13] Ostapchenko, S.：增强型庞美子方案中强子相互作用的蒙特卡罗处理：QGSJET-II 模型。《物理评论D》**83**，014018（2011）https://doi.org/10.1103/PhysRevD.83.014018[14] Pierog, T., Karpenko, I., Katzy, J.M., Yatsenko, E., Werner, K.：EPOS LHC：利用欧洲核子研究中心大型强子对撞机测量的数据检验集体强子化。《物理评论C》**92**，034906（2015）https: //doi.org/10.1103/PhysRevC.92.034906[15] Riehn, F., Engel, R., Fedynitch, A., Gaisser, T.K., Stanev, T.：强子相互作用模型 SIBYLL 2.3d 与广延大气簇射。《物理评论D》**102**，063002（2020）https://doi.org/10.1103/PhysRevD.102.063002[16] Gupta, S.K., *et al.*：GRAPES-3：用于研究膝区附近宇宙射线能谱结构的高密度空气簇射阵列。《核仪器与方法物理研究A辑》**540**，311–323（2005）https://doi.org/10.1016/j.nima.2004.11.025[17] Stef´ansson, A., Kon˘car, N., Jones, A.J.：关于伽马检验的简述。《神经计算与应用》**5**，131 133 1997 10 1007 01413858 – ( ) https://doi.org/ . /BF[18] Evans, D., Jones, A.J.：伽马检验的证明。《皇家学会会刊A辑》**458**(2027)，2759–2799（2002）https://doi.org/10.1098/rspa.2002.1010[19] R´enyi, A.：论依赖性度量。《匈牙利科学院数学学报》**10**，441–451（1959）https://doi.org/10.1007/BF02024507[20] Heck, D., Knapp, J., Capdevielle, J.N., Schatz, G., Thouw, T.：CORSIKA：用于模拟广延大气簇射的蒙特卡罗代码。技术报告 FZKA 6019，卡尔斯鲁厄研究中心（1998）[21] Hayashi, Y., *et al.*：用于超高能宇宙线天体物理学的大面积μ子径迹探测器：GRAPES-3 实验。《核仪器与方法物理研究A辑》**545**，643–657（2005）https://doi.org/10.1016/j.nima.2005.02.020[22] Varsi, F., *et al.*：用于 GRAPES-3 实验大面积μ子望远镜的 GEANT4 模拟框架。《仪器学报》**18**，03046（2023）https://doi.org/10.1088/1748-0221/18/03/P03046[23] Antoni, T., *et al.*：KASCADE 对宇宙射线元素组能谱的测量：结果与待解决问题。《天体粒子物理学》**24**，1–25（2005）https://doi.org/10.1016/j.astropartphys.2005.04.001[24] Antoni, T., *et al.*：KASCADE 测得的大尺度宇宙线各向异性。《天体物理学杂志》**604**，687–692（2004）https://doi.org/10.1086/382039[25] Malik, S., Ashiya, M.R.：GRAPES-3 模拟空气簇射摘要（$\log_{10} N_e$ 4.8–5.2）及存档预测表，版本 1.0。Zenodo，数据集（2026）。https://doi.org/10.5281/zenodo.23096836[26] Cover, T.M., Thomas, J.A.：《信息论基础》，第2版。Wiley，霍博肯，新泽西州（2006）