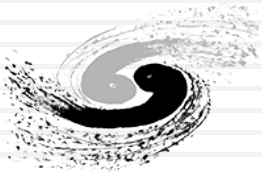


Institute of High Energy Physics
Chinese Academy of Sciences

使用生成型模型进行探测器快速准确模拟

王锦

实验物理中心CMS组

机器学习技术在CMS实验中的应用

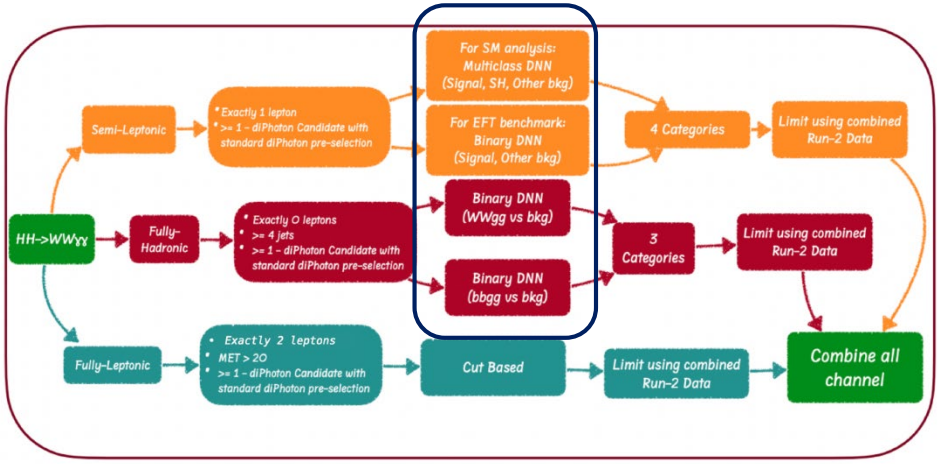
- 欧洲核子中心大型强子对撞机CMS实验中机器学习技术被深入地探索与应用
  - 实验运行监控、硬件触发及高级触发、数据清洗与降噪
    - 在线数据异常检测
    - 快速硬件触发及模型压缩技术
    - 高级触发以进行特殊物理事例筛选，数据降噪技术
    - 典型的ML方法：自动编码器、深度聚类等
  - 事例分类，物理对象重建、物理分析优化
    - 区分信号、排除本底
    - 重建和识别各类粒子，如喷注/tau粒子/电子/光子/μ子
    - 高级对象重建/标记，如希格斯粒子、顶夸克、W等的重建和标记
    - 无似然技术用于有效场理论（EFT）研究，用ML来降低系统误差的影响
    - 典型ML方法：决策树群（BDTs），深度神经网络（DNN），卷积神经网络（CNN），递归神经网络（RNN）、图神经网络（GNN）及新的改进算法（如particleNet）等
  - 事例模拟与探测器模拟
    - 快速、准确地模拟粒子碰撞和探测器响应
    - 探测器几何重建，优化设计
    - 典型ML方法：生成对抗网络（GANs），变分自编码器（VAEs），流模型（Normalizing Flow），扩散模型（Diffusion models）等
- 机器学习技术的研究与应用有效提高了高能实验数据处理和分析的效率，降低了人力与计算资源消耗，提升了物理测量精度与新物理发现的可能性

高能所CMS机器学习应用与研究

- 新的堆积事例排除算法 – Fabio Iemmi
  - 不依赖事例真实标记，可以广泛应用的噪声排除机器学习技术

arXiv:2211.02029

- 事例分类、粒子重建、高级对象重建



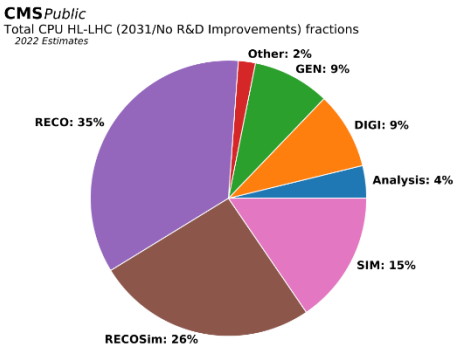
HH→ $WW\gamma\gamma$ 分析流程图

CMS-PAS-HIG-21-014

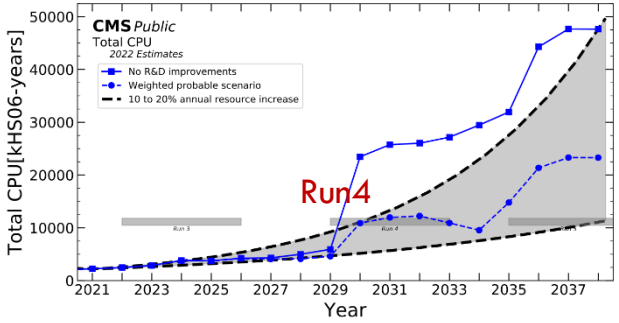
  - 深度神经网络（DNN），多类别机器学习区分技术，EFT参数化机器学习训练等
  - Higgs jet tagger, permutational DNN W-jet tagger
  - 基于自注意力机制的high p_T b-jet tagging
- CMS电磁量能器模拟优化
  - 使用生成型模型进行探测器快速准确模拟

目前探测器模拟中的问题

- CMS实验模拟计算需求大
  - 模拟CMS探测器响应和数据处理需要大量的计算资源。
  - 未来HL-LHC探测器升级进一步提升了模拟复杂程度与数据需要
  - 事例堆叠（Pileup）的提高使CPU需求超线性增长
  - 如果没有进一步的优化，将在Run4超过计算资源预算。
  - ATLAS, LHCb等实验面临同样问题
- 蒙卡模拟与数据有一定区别
  - 无法准确描述探测器随着时间和辐射的变化
  - 一些过程的理论描述不精确或是非常难产生
    - 稀有过程，未知过程等
  - 是一些测量主要的误差来源之一
- 使用生成模型机器学习技术进行探测器快速准确模拟是主要的模拟优化研究方向之一



CMS-NOTE-2022-008



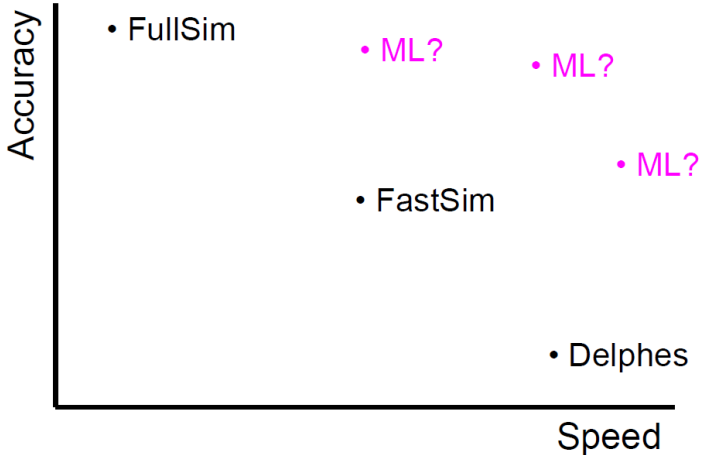
CMS希格斯质量测量误差来源

Source	Contribution (GeV)
Electron energy scale and resolution corrections	0.10
Residual p_T dependence of the photon energy scale	0.11
Modelling of the material budget	0.03
Nonuniformity of the light collection	0.11
Total systematic uncertainty	0.18
Statistical uncertainty	0.18
Total uncertainty	0.26

Phys. Lett. B 805 (2020) 135425

CMS模拟与机器学习模拟

- CMS实验全模拟采用GEANT4模拟，是基于物理过程的详细模拟
  - 理论模型来描述模拟探测器几何、材料、粒子传输和各种粒子相互作用等。
  - 优点: 模型的预测基于基础物理原理，可解释性强，准确性通常很高。
  - 缺点: 模拟计算成本很高, 对模型依赖性高。
- CMS实验的快速模拟（Fast Simulation）使用近似参数化模型来代替一些耗时的模拟过程
  - 优点：显著提高模拟的效率
  - 缺点：模拟精确度差
- 机器学习模拟技术主要生成模型的机器学习方法
  - 是基于数据驱动的方法，从大量的训练数据学习预测新的事件
  - 优点:
    - 机器学习模型被训练后可以快速准确生成大量的模拟数据
    - 可以学习到复杂的模式和关联，即使物理模型中难以描述。
  - 缺点
    - 机器学习模型受限于训练数据的质量和数量
    - 机器学习模型通常是一个“黑箱”，缺乏模型背后的物理原理与可解释性。



人工智能大语言模型与模拟生成型模型

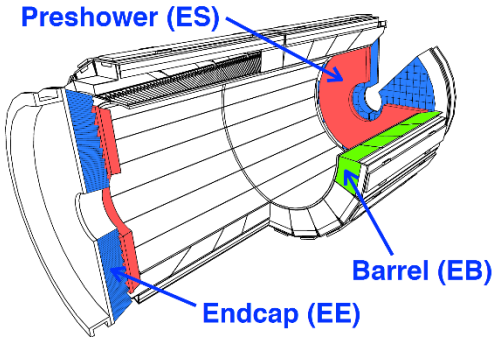
- 人工智能大语言模型（e.g. chatGPT） - 理解和生成复杂的自然语言文本
  - 需要大量高质量数据，获取、清洗和标记这些数据可能需要大量人力和时间成本。
  - 大模型的参数数量庞大，优化过程困难耗时，模型部署和优化迭代周期长
  - 大模型需要大量的计算资源进行训练，成本很高

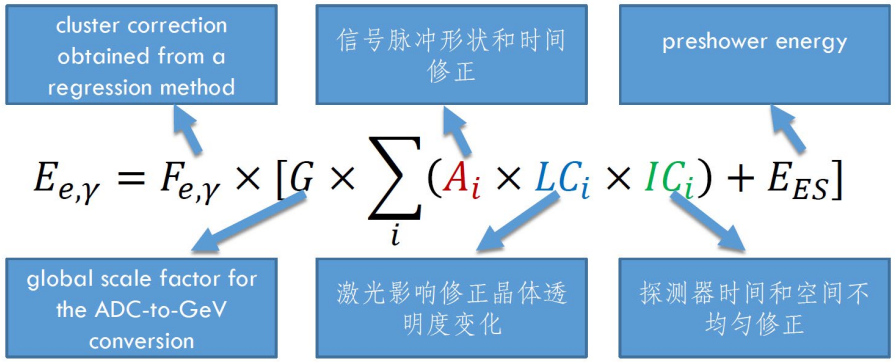
- 模拟生成型模型 – 精确生成预测现实物理过程，另一类拥有广阔前景的ML技术
  - 高能物理实验数据是用于训练生成型模型的优质资源，可以高效地优化和研究机器学习技术

量级大	结构化	复杂性
高能物理实验产生的数据量极大，LHC每秒钟都会产生大约一千万个碰撞事件，每个事件可能包含几十到几百个粒子。	高能物理数据高度结构化，构建模型中可以引入物理规律，对机器学习模型进行更好地约束和优化	高能实验数据通常涉及非常多不同的变量和粒子事件，是训练和测试高级生成模型的理想资源

CMS电磁量能器模拟

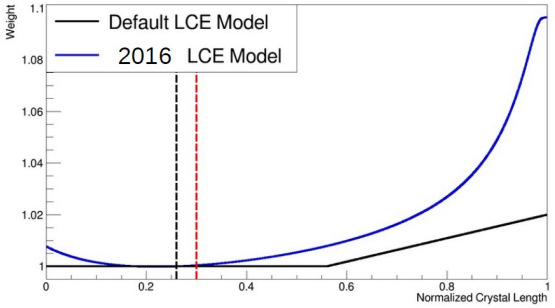
- CMS电磁量能器性能组（ECAL DPG）与CMS机器学习团队开展新的用于探测器模拟优化的机器学习技术研究
  - ECAL DPG convener: 王锦，Thomas Reis
  - 降低未来升级后模拟的计算资源需求
  - 提高模拟与真实数据的符合度，降低精确测量系统误差
    - 使模拟更好地反应探测器状态





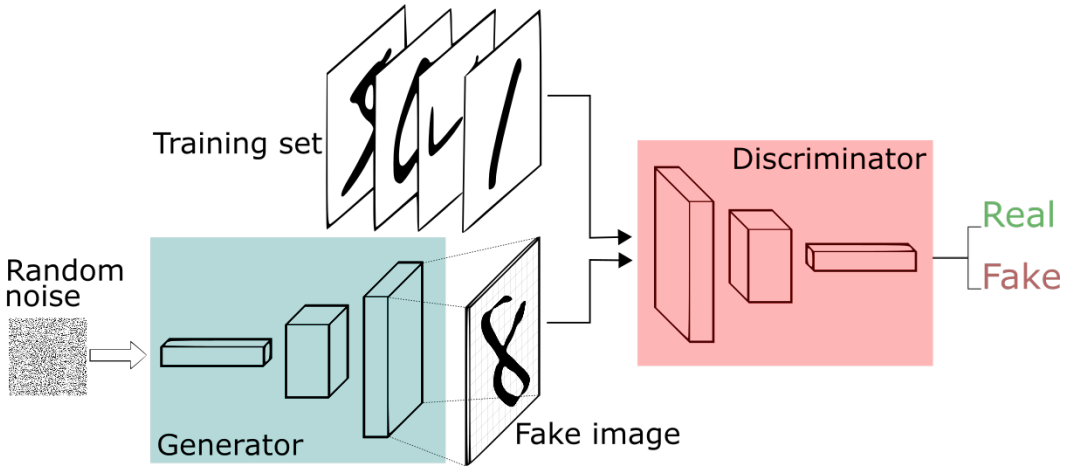
电子光子能量重建

- 模拟一些辐射造成的探测器响应变化
  - 晶体光线收集效率随晶体深度，辐射而变化
  - 对电子与光子能量重建的影响不相同
  - 目前蒙卡模拟无法准确模拟这些区别
    - 目前希格斯质量测量的主要误差来源之一



主流的模拟生成型模型

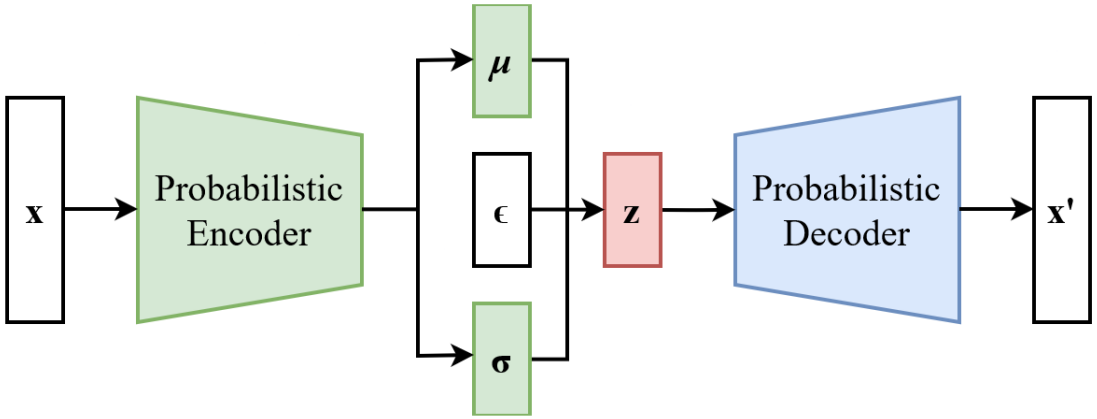
- 对抗生成网络（Generative Adversarial Networks，GANs）：GANs是一种深度学习方法，由两个神经网络组成——一个生成器和一个判别器，两者相互竞争。生成器的目标是生成假的数据，使其尽可能地接近真实数据；判别器的目标是区分生成的数据和真实的数据。



- 优点：性能优秀，不依赖数据标记
- 缺点：训练困难，模式崩溃问题（忽视部分信息），模型结构不易控制和变化

主流的模拟生成型模型

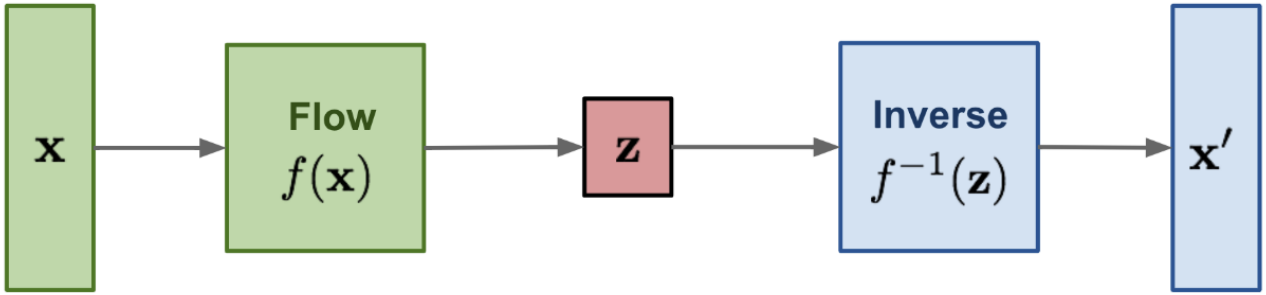
- 变分自编码器（Variational Autoencoders，VAEs）：VAEs是一种生成模型，其将输入数据编码为潜在表示，学习概率分布，然后从这个表示中解码生成数据。VAEs特别适合处理具有连续潜在变量的数据，可以用于模拟复杂的物理过程。



- 优点：训练稳定，隐空间结构化，可控性强
- 缺点：生成的数据质量低，需要假设后验分布，复杂数据处理能力弱

主流的模拟生成型模型

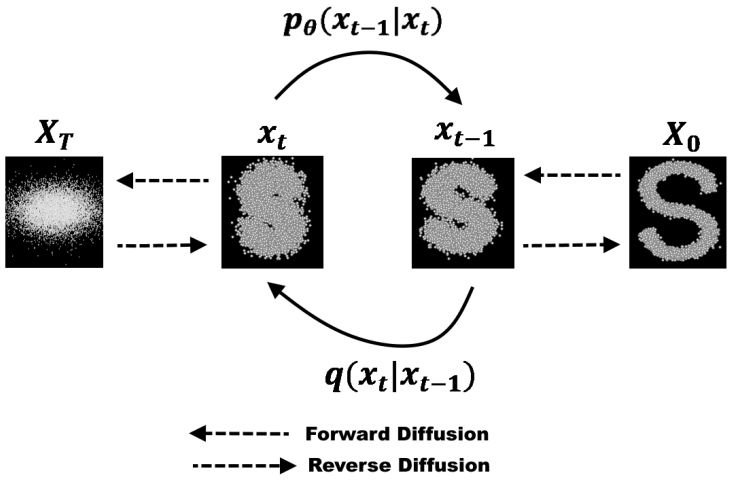
- 流模型（Flow Models）：流模型，通过构造一个可逆的变换（或者说是流），将简单的分布变换到复杂的数据分布，可以同时完成精确的生成、密度估计和采样。



- 优点：训练稳定，可以显示数据概率
- 缺点：需要复杂计算，保证函数可逆，函数形式有限制

主流的模拟生成型模型

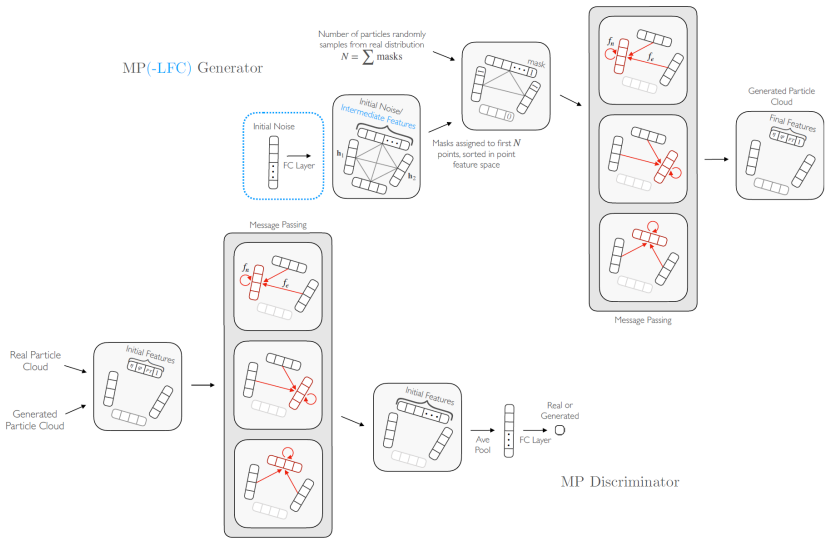
- 扩散模型（Diffusion Models），通过对原始数据加入噪声进行随机扩散，从数据分布逐渐“扩散”到一个已知的简单分布（如正态分布），然后再通过逆过程从简单分布生成新的数据。



- 优点：训练简单，显示概率，样本生成质量高
- 缺点：多步扩散慢，计算复杂，需要大量训练

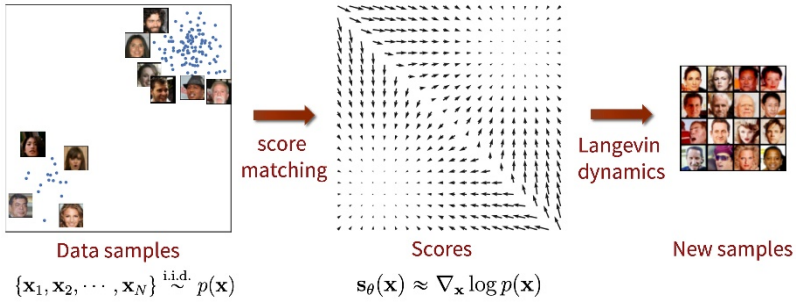
外面已有的前沿模拟生成型模型

- Message Pass GANs (MP-GAN)
  - 基于点云(point cloud)的生成器和判别器
  - Generic message-passing neural network (MPNN) framework
  - arXiv:2106.11535



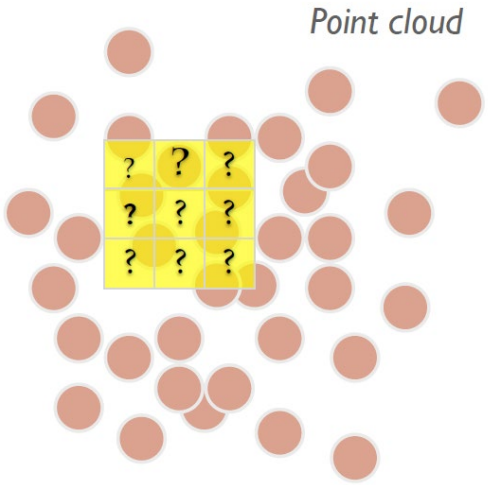
- EPiC-GAN — equivariant point cloud generative adversarial network
  - arXiv:2301:08128
  - Multiple EPiC layers with an interpretable global latent vector.

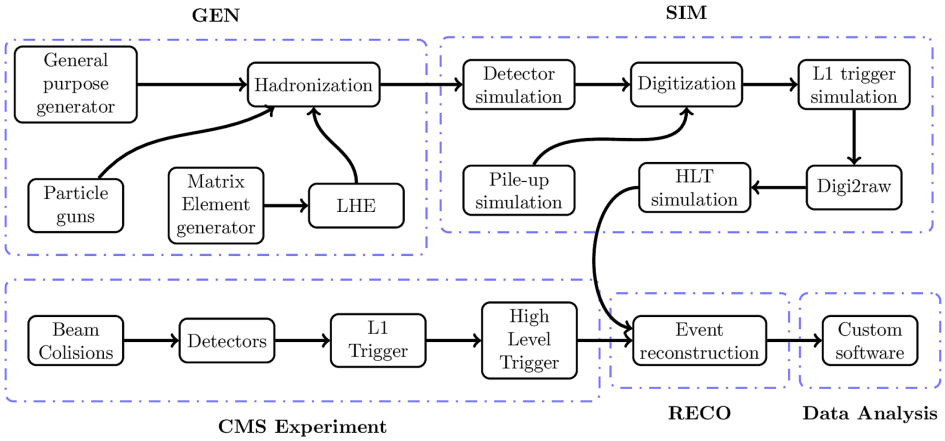
- Score based diffusion models
  - 用概率梯度取代扩散中使用的概率分布

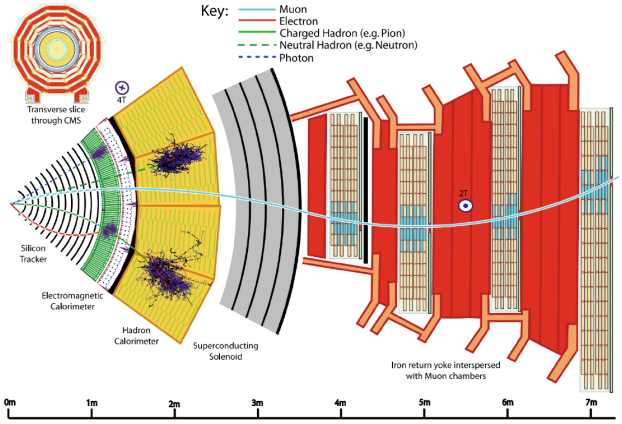


前沿生成模型的探索

- 数据表示
  - 向量、1D图像、分片2D图像、3D图形、点云
- 设计模型结构及演变方法、损失函数
- 不同模型的组合
- 平衡精确性与速度
  - 全部或部分替代全模拟
    - 如模拟粒子簇射、模拟从产生子到探测器沉积
  - 全部或部分替代快速模拟
    - 改进其中的近似模型
  - 从产生子到重建对象、重建事例的映射
  - 直接从噪声到最终事例的映射







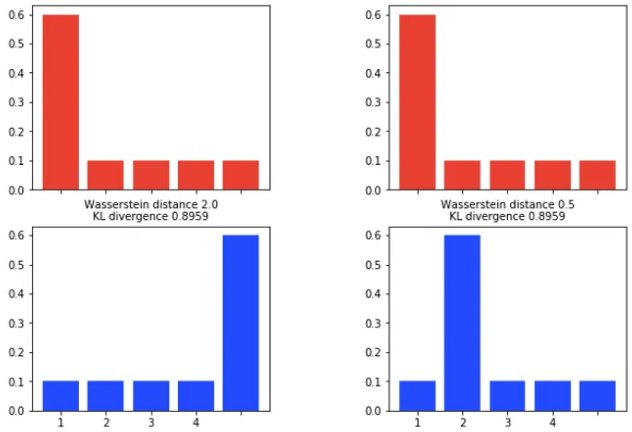
模型性能对比指标

- 训练时间，CPU/GPU、内存消耗

- 物理性能指标
  - 最优传输距离 - Wasserstein distance
    - 将一个分布移动并改变形状以覆盖另一个分布所需的最小工作量
    - 能够反应形状和位移，可以衡量复杂的变量
    - 优于KL divergence, KS, x2等指标

  - 基于机器学习技术的指标
    - Frechet distance （arXiv:2106.11535）

  - 训练区分器等



总结

- CMS高能物理实验中已经大规模应用机器学习技术
  - 高效处理数据，提升物理分析潜力，降低人力计算成本
  - 高能物理高质量数据非常有助于研发机器学习技术
- 探测器模拟优化需要研究使用机器学习技术
  - 降低探测器模拟计算资源需求
  - 提高模拟与真实数据的符合度，降低精确测量系统误差
- 研究应用及改进前沿生成型机器学习技术以快速、准确地探测器模拟
  - 基于CMS探测器电磁量能器的新研究
  - 致力于解决目前生成型模型的问题以及实现早期应用
    - MPGAN，EPIC-GAN，GVAE，Score based diffusion models等
- 模拟生成型模型预期有非常广阔的应用前景
  - 精确生成和预测现实物理过程
  - 相应技术也可用于CEPC，中微子，天文等