使用自动编码器对 BESIII EMC 异常检测

李明润，刘春秀，季晓斌

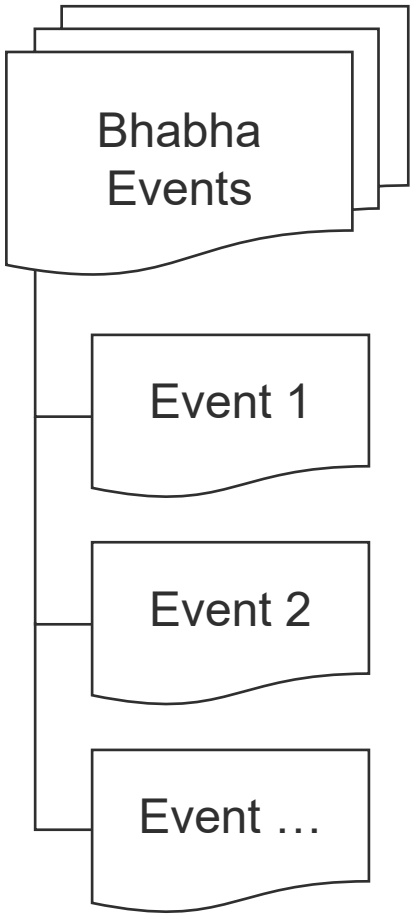
2024-10-17

1. 简介

- 高精度的物理分析需要高质量的实验数据，更高效的探测器异常检测能够有效提高取数效率和数据质量。
- 近年来，机器学习发展迅猛，使用机器学习方法进行异常检测有望显著提高BESIII的探测器异常检测能力。
- 我们开发了一个使用自动编码器对BESIII EMC进行异常检测的方法，并且发现了传统方法不易发现的探测器异常。

2. 数据处理

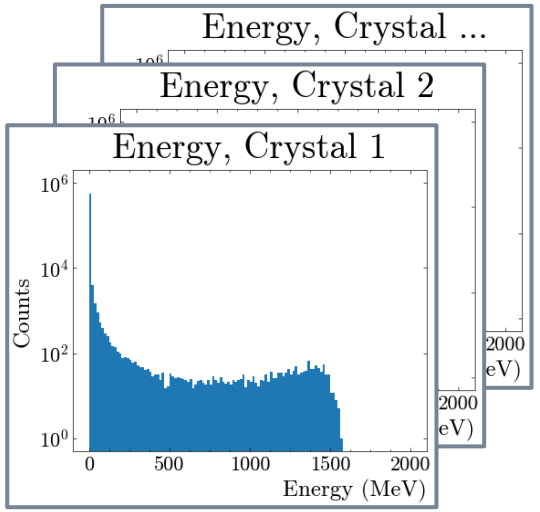
量能器晶体数量：桶部5280+端盖960=总共6240块



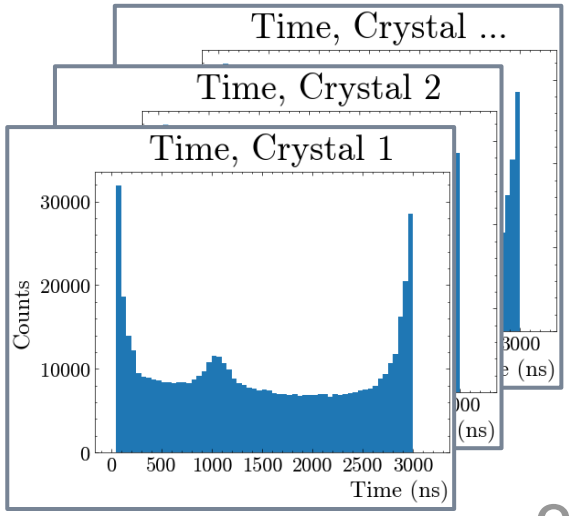
EMC 击中格式

Crystal ID	Energy	Time
ID 1	Energy 1	Time 1
ID 3	Energy 2	Time 2
ID 2	Energy 3	Time 3
ID 1	Energy 4	Time 4
ID 2	Energy 5	Time 5
…	…	…

6240张能量直方图

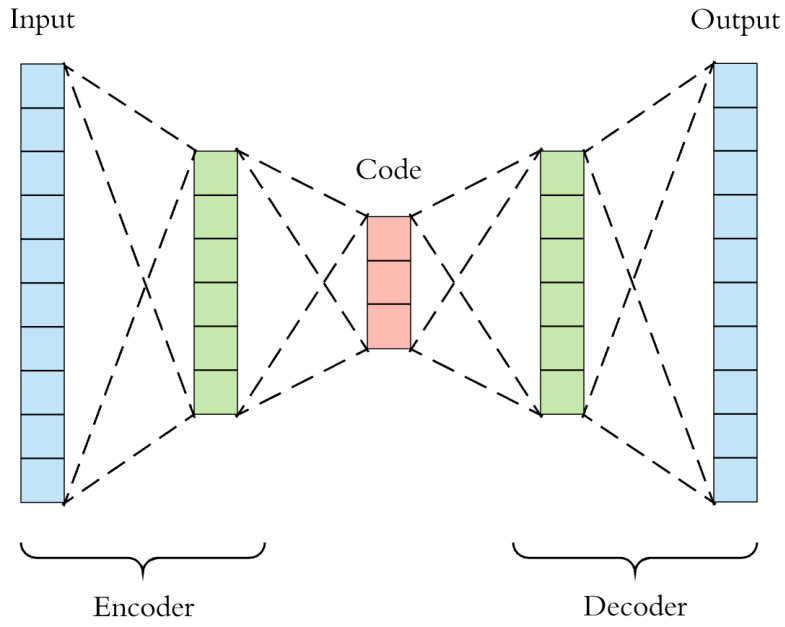


6240张时间直方图



3. 使用自动编码器进行异常检测

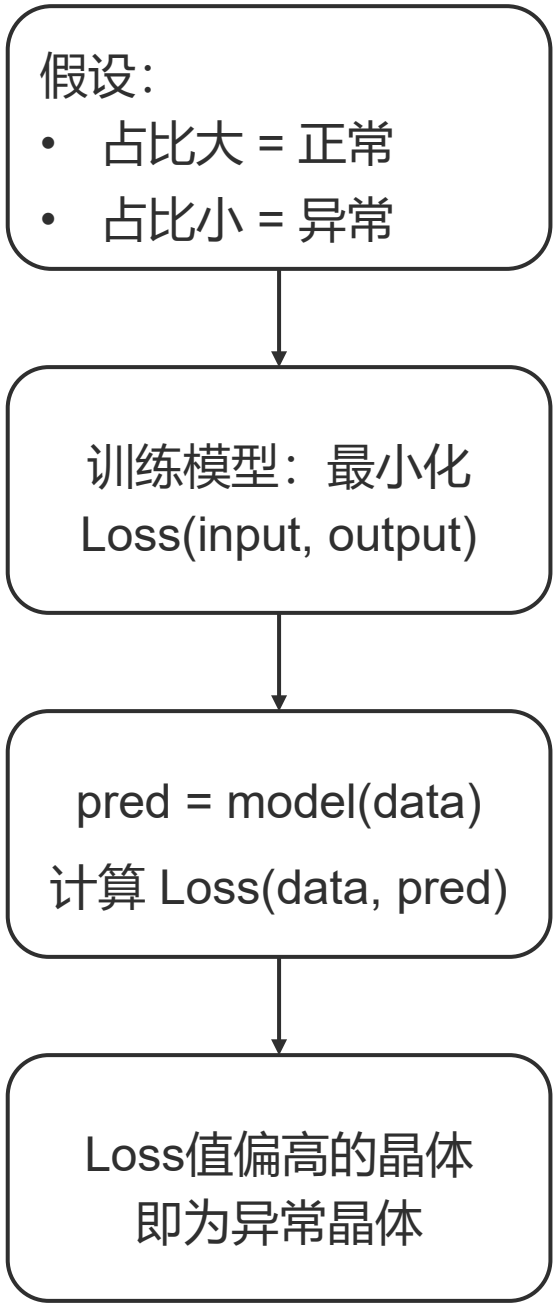
- **自动编码器（AE）：** 特殊结构的神经网络
  - 输入维度与输出维度相同；
  - 中间层维度更低。
- **训练：**要求模型的输出与输入保持一致（最小化输出与输入的损失函数）
- **损失函数：**Jensen-Shennon Distance
- **异常检测原理：**
  - 自动编码器会找到适合数据集中**占比大**的那一类数据的编/解码（降/升维）规则；
  - 编解码规则不适合占比小的数据，会导致模型对其的输出与输入差异较大（Loss值偏大）。



Model Structure

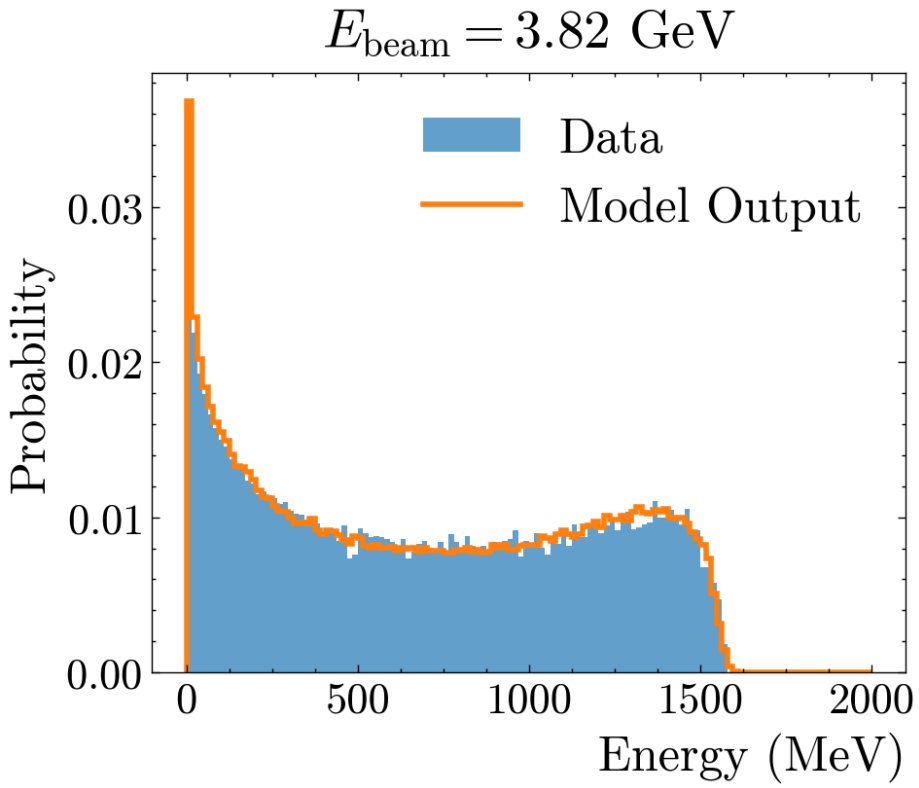
3. 使用自动编码器进行异常检测

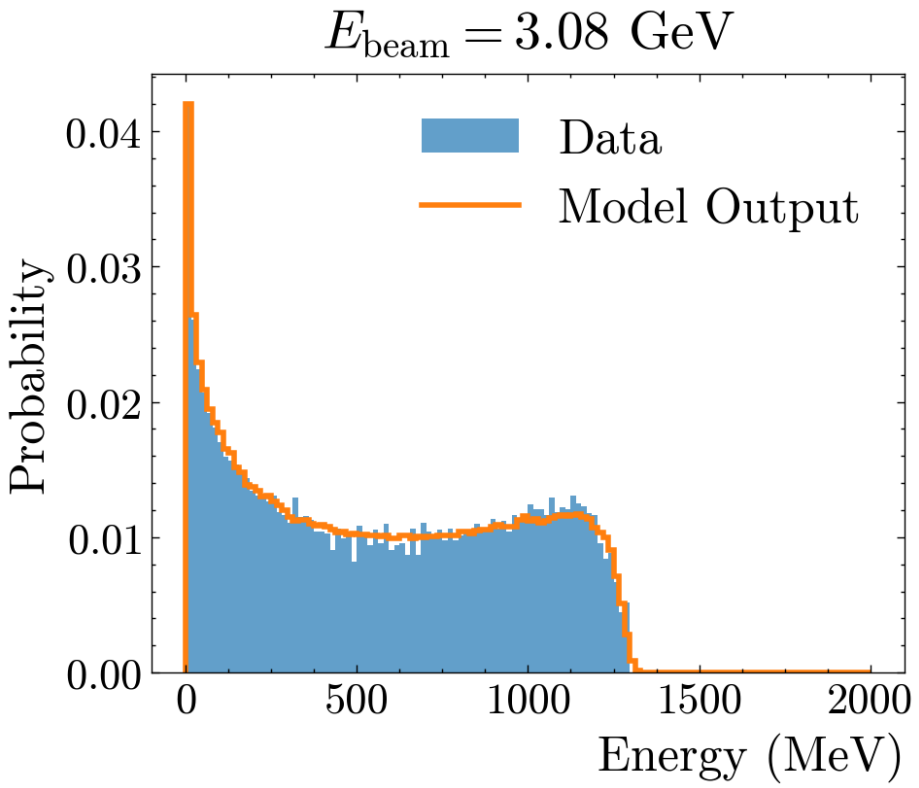
- **自动编码器（AE）：** 特殊结构的神经网络
  - 输入维度与输出维度相同；
  - 中间层维度更低。
- **训练：**要求模型的输出与输入保持一致（最小化输出与输入的损失函数）
- **损失函数：**Jensen-Shennon Distance
- **异常检测原理：**
  - 自动编码器会找到适合数据集中**占比大**的那一类数据的编/解码（降/升维）规则；
  - 编解码规则不适合占比小的数据，会导致模型对其的输出与输入差异较大（Loss值偏大）。



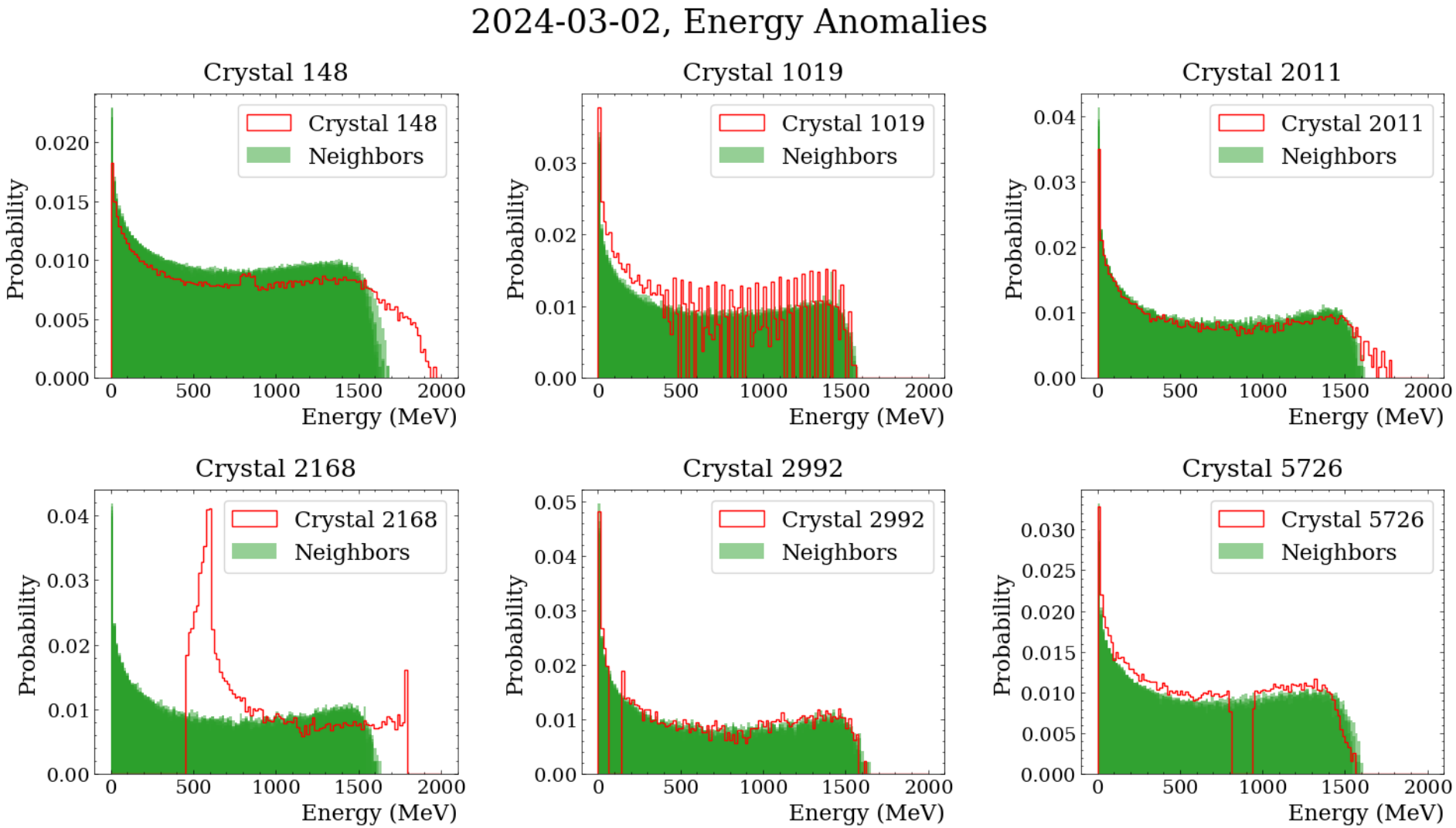
3. 使用自动编码器进行异常检测

优势：能够自适应外部条件（质心系能量、束流本底等）的变化



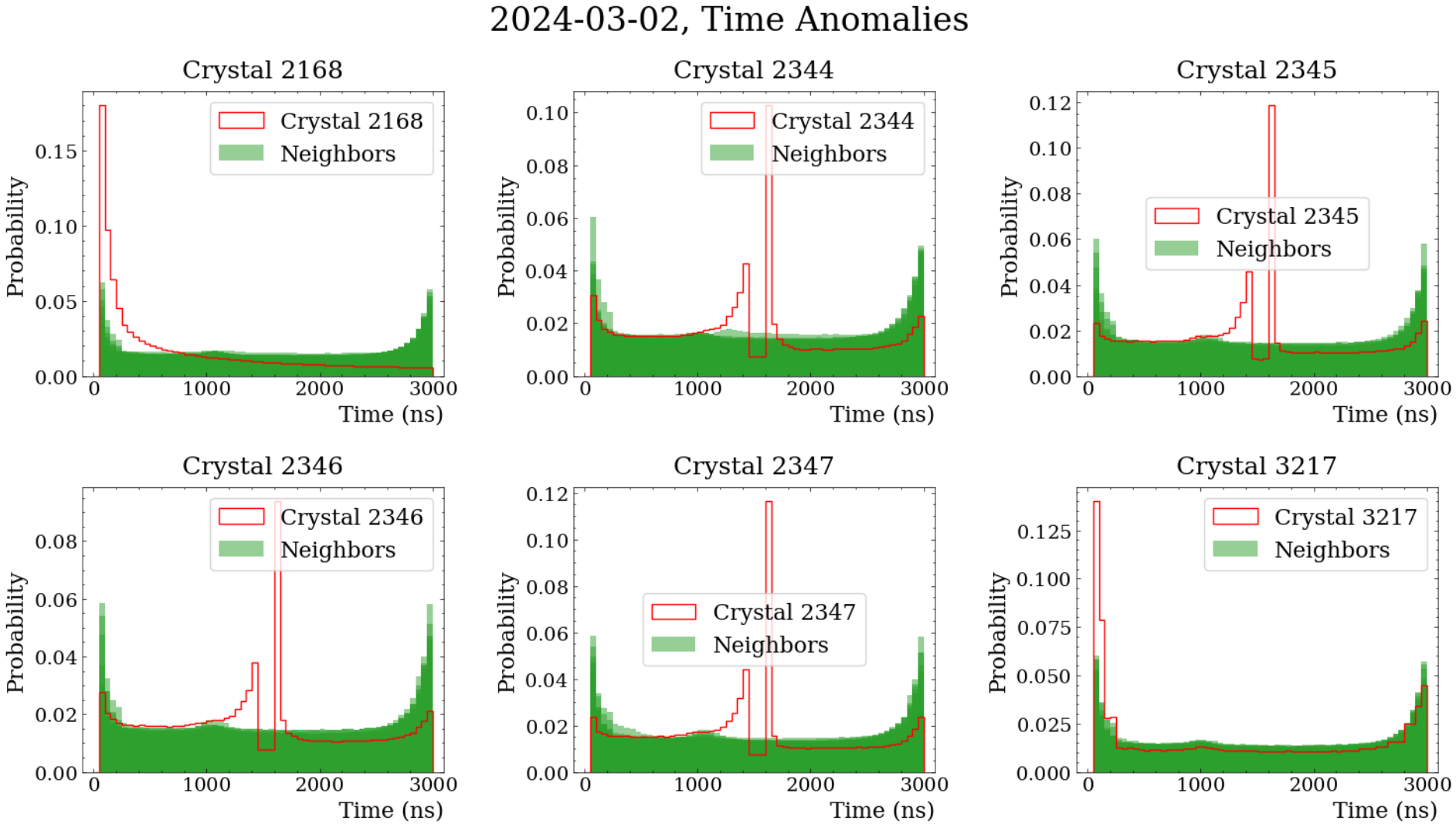


4. 对 EMC 进行异常检测



“Neighbors”: 异常晶体周围的正常晶体

4. 对 EMC 进行异常检测



“Neighbors”: 异常晶体周围的正常晶体

总结

- 开发了一个高效的基于机器学习的离线EMC异常检测方法；
- EMC的事件响应信息现在也能被检测；
- 新方法找到了一些隐蔽的EMC晶体异常。

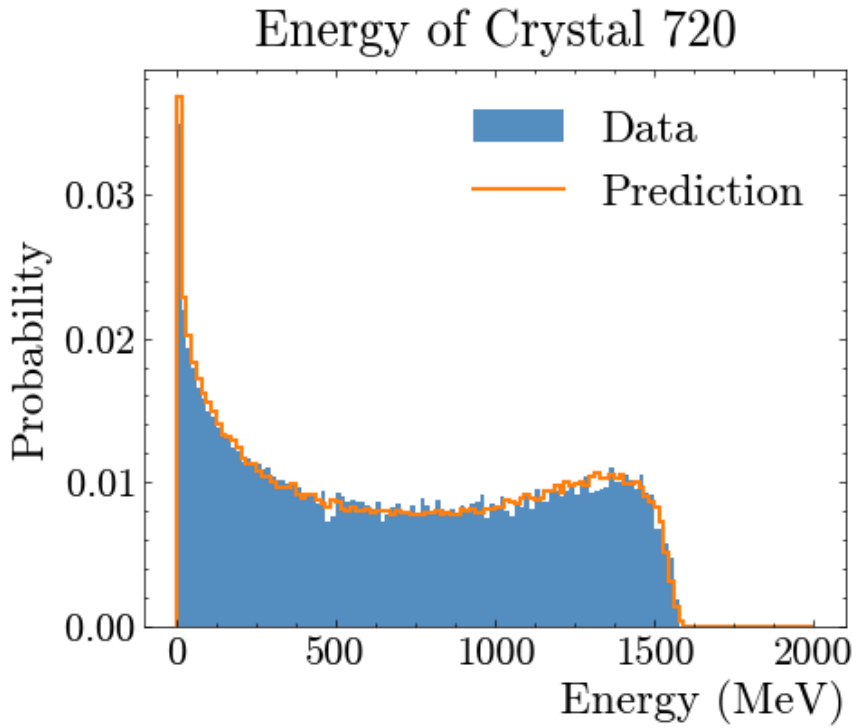
下一步

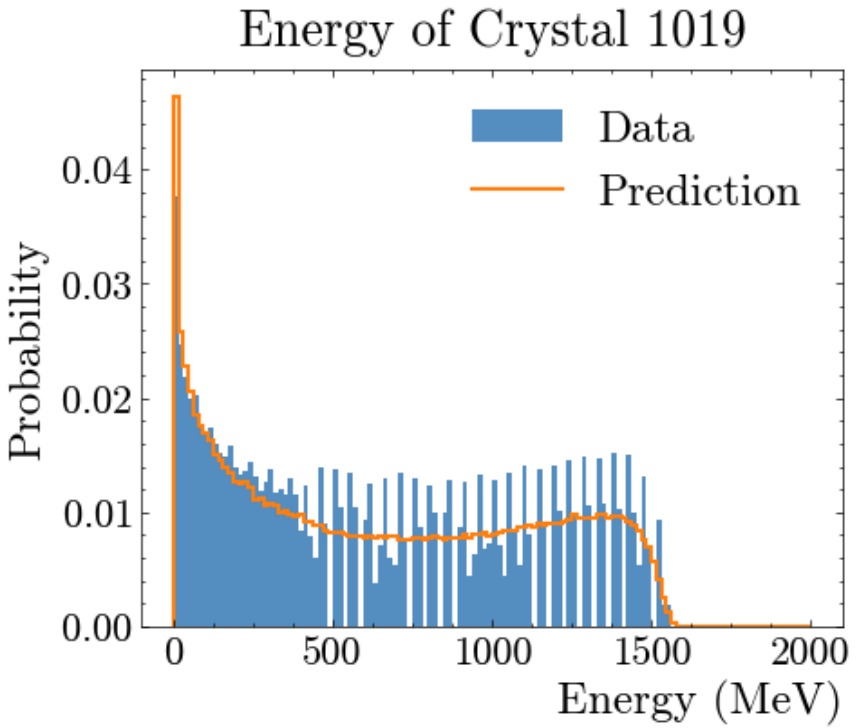
- 优化模型的超参数；
- 将这个方法自动化部署，在取数时进行常态化检测。

Back Up

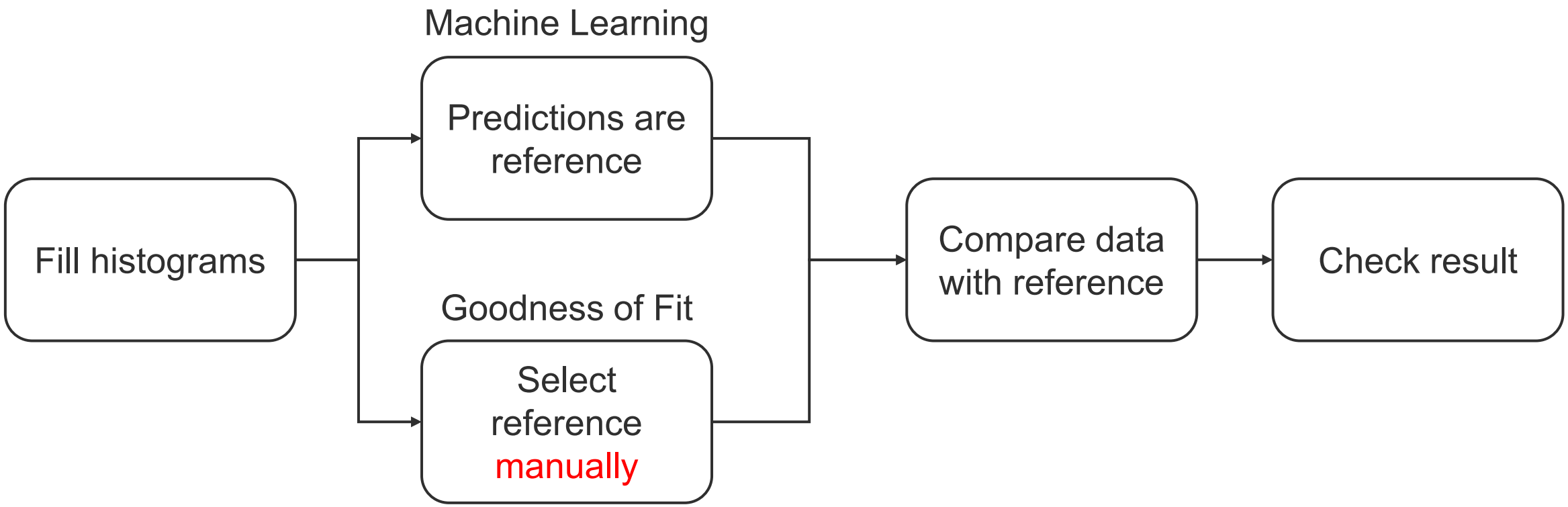
3. Detect Anomaly by Autoencoder

Model predictions can be regarded as reference





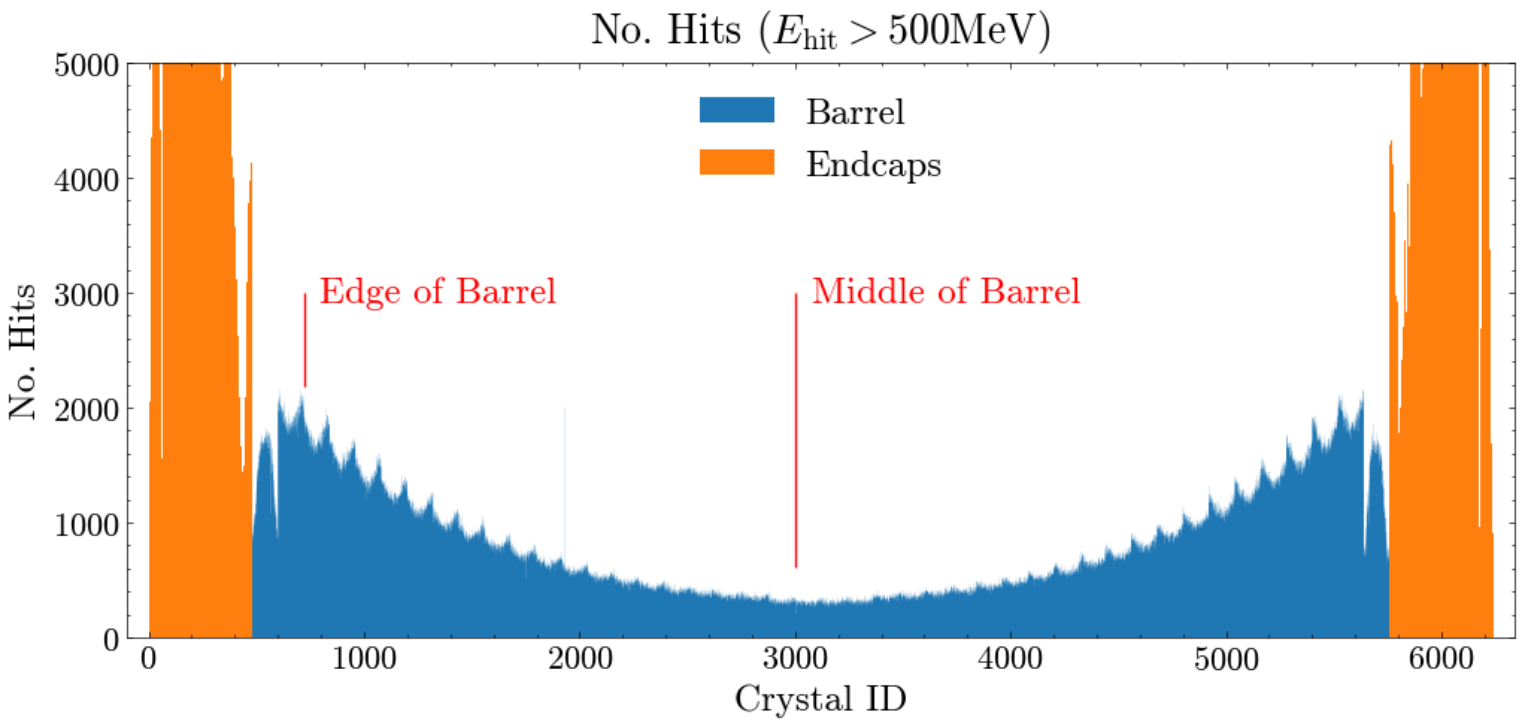
3. Detect Anomaly by Autoencoder

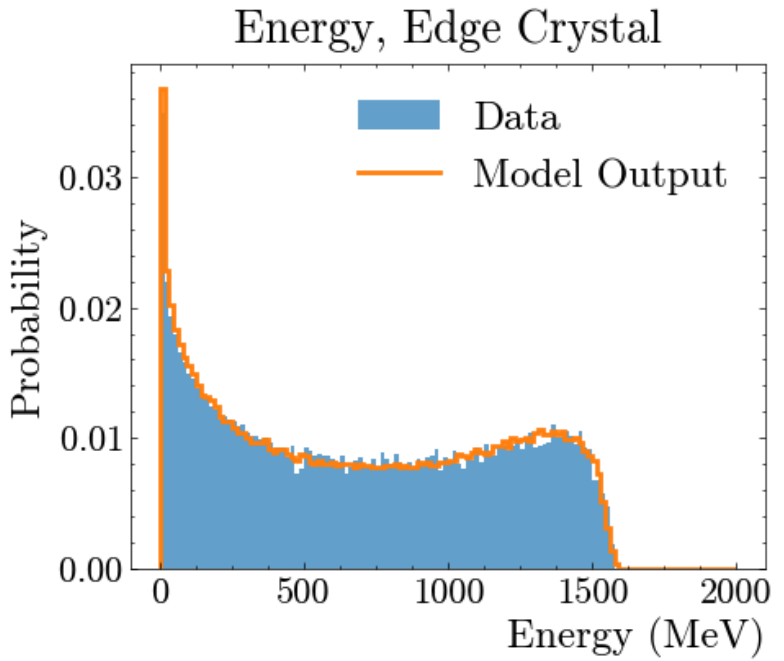


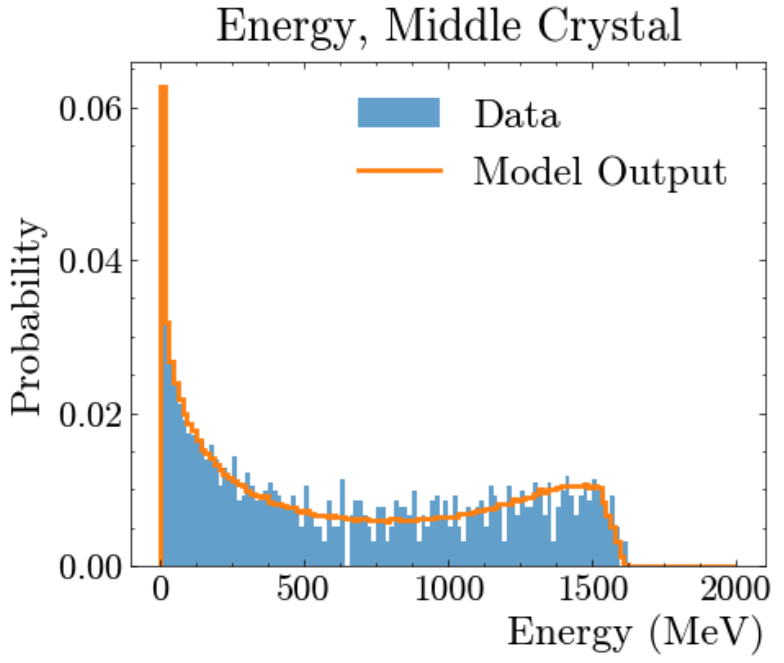
For Goodness of Fit method, we need to select new reference manually once outer condition (such as beam energy) changes.

3. Detect Anomaly by Autoencoder

Machine learning method can also well handle position-depend difference:







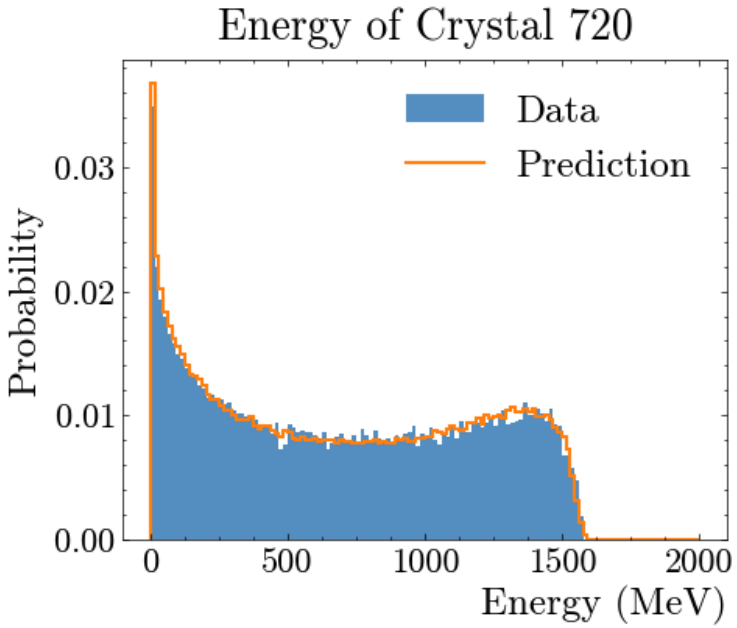
4. Anomaly Detection on EMC

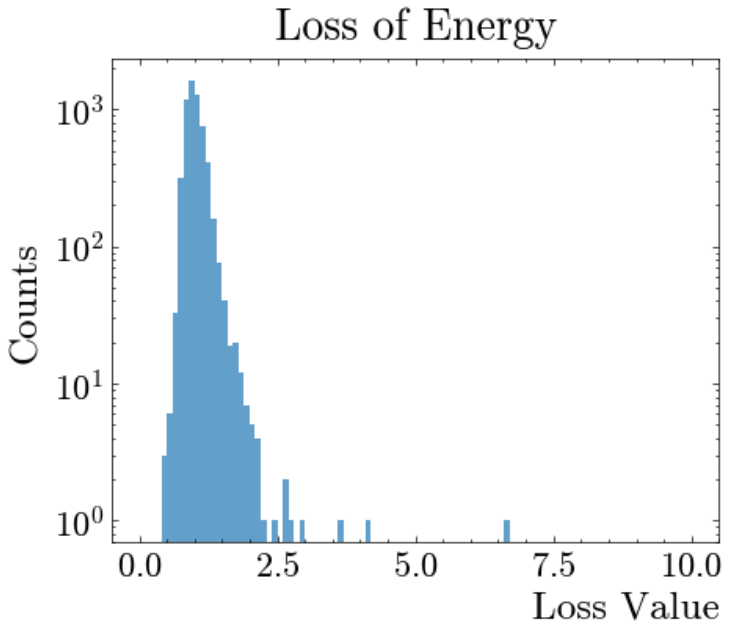
In practice on EMC:

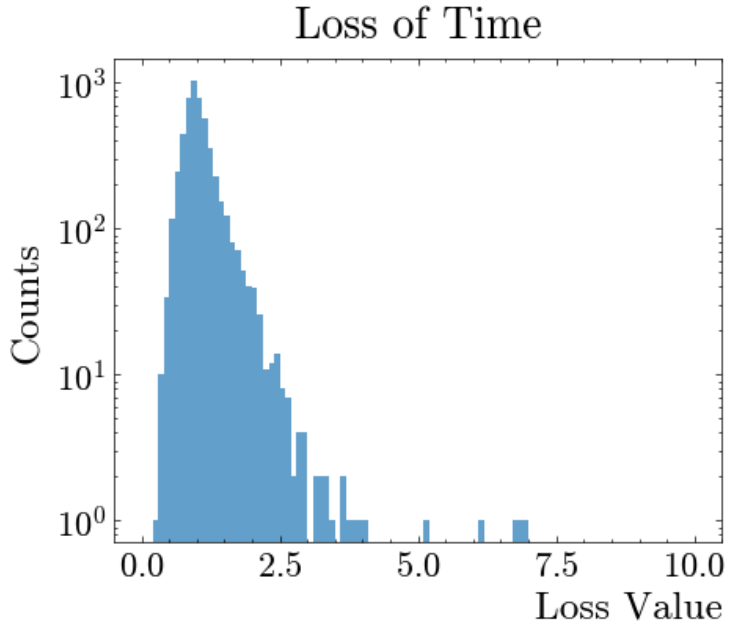
① Select Jensen-Shannon Distance as loss;

② Train model, obtain predictions, calculate losses for all crystals;

(Separately apply on energy and time histograms)



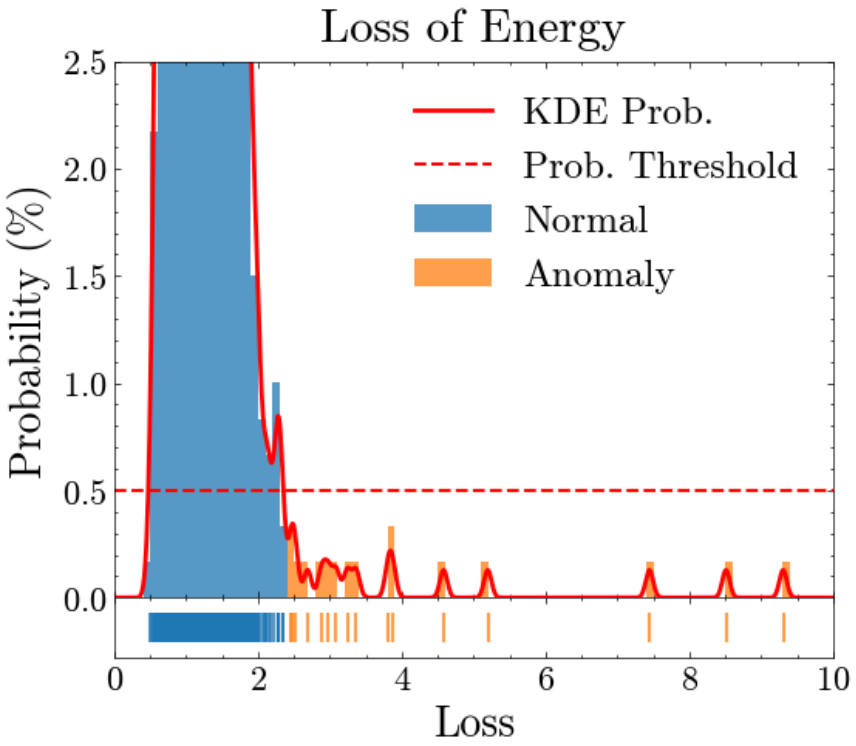


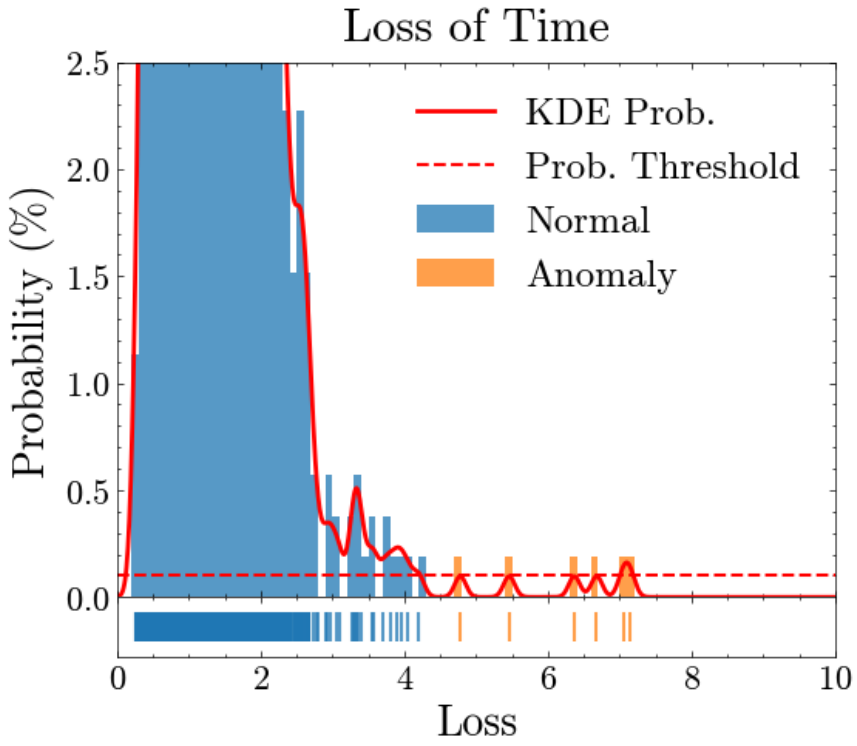


4. Anomaly Detection on EMC

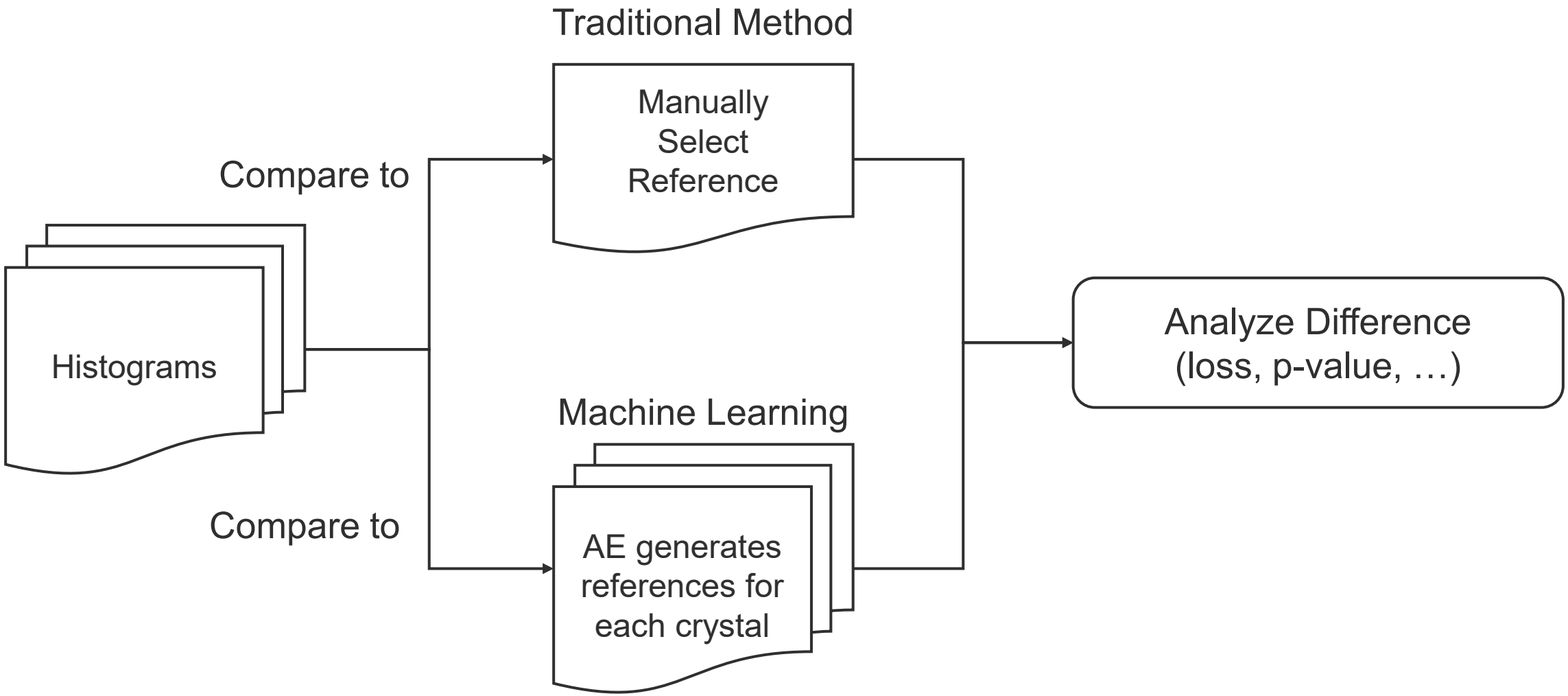
③ Estimate loss probability density by Kernel-Density Estimation (KDE);

④ Anomalies: (1) Energy: $\text{Prob.} < 0.5\%$ (2) Time: $\text{Prob.} < 0.1\%$





5. Advantages of Machine Learning



4. Strategy of Anomaly Detection

Loss Function: Jensen-Shannon Distance