Efficiency Parameterization with Neural Networks

*C. Badiali, F. A. Di Bello, G. Frattari, E. Gross, V. Ippolito, M. Kado, **N. Kakati**, J. Shlomi*

(nilotpal.kakati@cern.ch)

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

SAPIENZA
UNIVERSITÀ DI ROMA

INFN

AI and Physics Conference
28 April, 2022

Computing and Software for Big Science (2021) 5:14
https://doi.org/10.1007/s41781-021-00059-x

ORIGINAL ARTICLE

Efficiency Parameterization with Neural Networks

Francesco Armando Di Bello[1,2] · Jonathan Shlomi[3] · Chiara Badiali[1,2] · Guglielmo Frattari[1,2] · Eilam Gross[3] · Valerio Ippolito[2] · Marumi Kado[1,2,4]

Received: 17 May 2020 / Accepted: 28 April 2021 / Published online: 28 May 2021
© The Author(s) 2021

Abstract
Multidimensional efficiency maps are commonly used in high-energy physics experiments to mitigate the limitations in the generation of large samples of simulated events. Binned efficiency maps are however strongly limited by statistics. We propose a neural network approach to learn ratios of local densities to estimate in an optimal fashion efficiencies as a function of a set of parameters. Graph neural network techniques are used to account for the high dimensional correlations between different physics objects in the event. We show in a specific toy model how this method is applicable to produce accurate multidimensional efficiency maps for heavy-flavor tagging classifiers in HEP experiments, including for processes on which it was not trained.

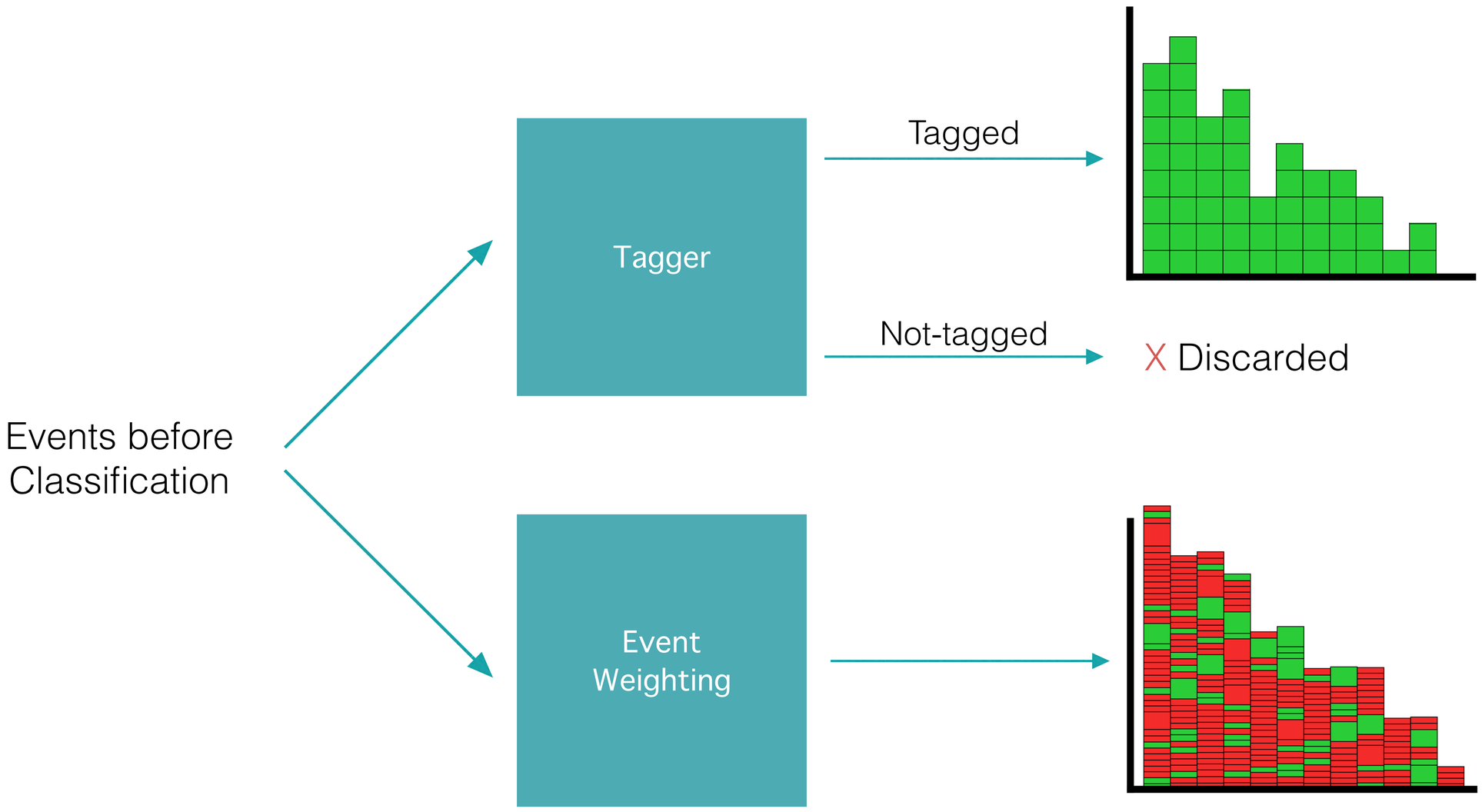
https://arxiv.org/abs/2004.02665v2

Introduction

- We often care about events occurring in a very restricted regions of phase space, but the MC simulations are quite inclusive
- Example: B-jet tagging
  - We have events with jets
  - Requirement: Events with two b-jets
- Trivial solution: Apply a b-Jet tagger

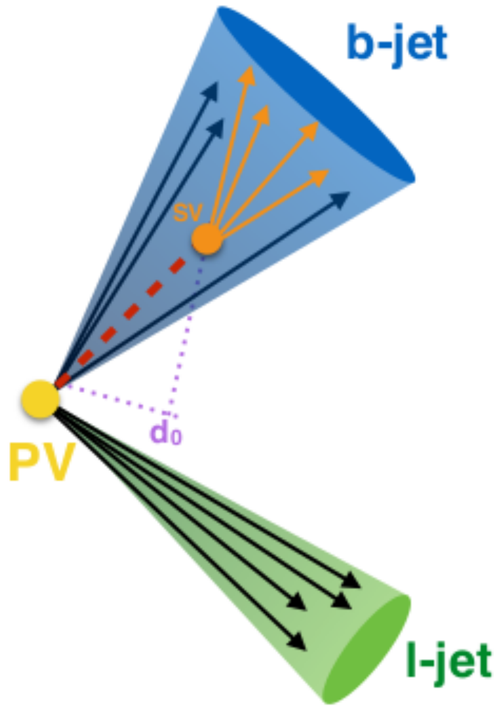
- Issue: Low statistics when rejection rate is high
- Alternate solution: Event Weighting Technique

Event Weighting



Efficiency Parameterization

Jet (X) → Tagger f(x) → Tagged/not tagged



- X: ~30 variables (mostly track vars)
- Distribution of 'X' depends on another set of variables 'θ' (jet pt, eta, phi etc.)
- Goal : To know the efficiency, $\epsilon(\theta)$ of the classifier

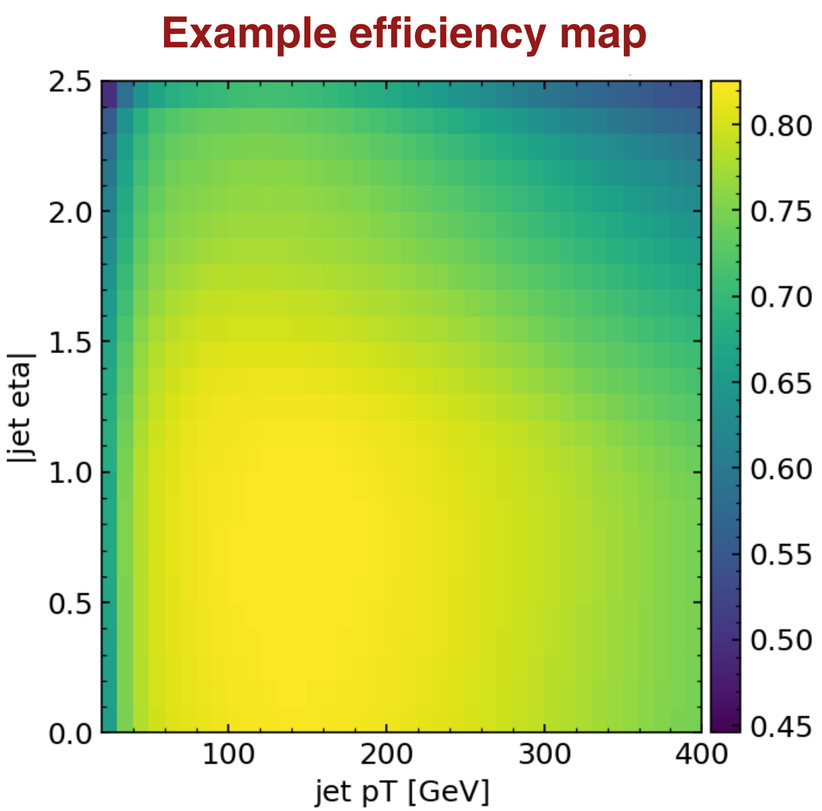
$$\epsilon(\theta) = \frac{N(tagged\,|\,\theta)}{N(\theta)}$$

What constitutes θ ?

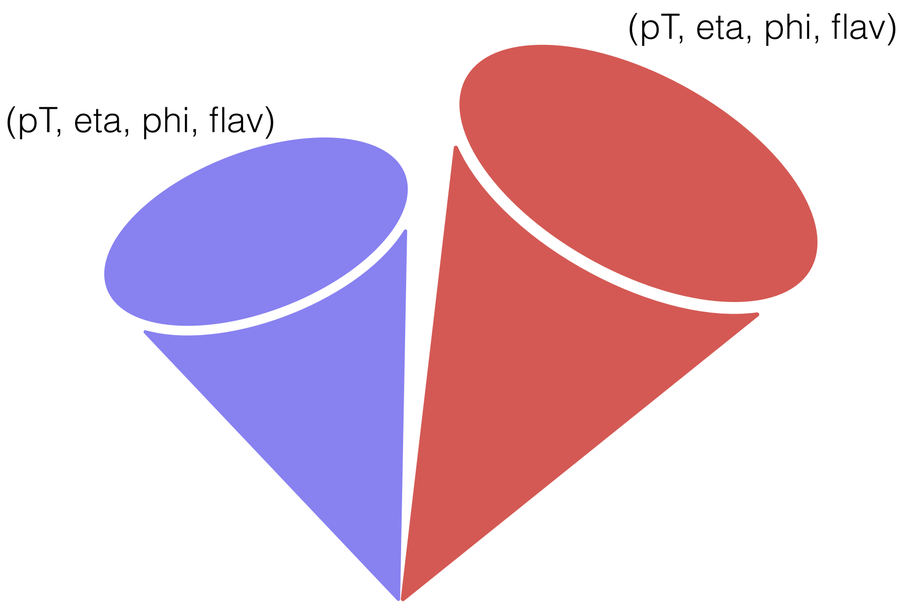
- Typically, 'θ' is known only partially
- For b-jet tagging in ATLAS

  p_T and η are the most dominant components of θ.

- Common Practice: binned efficiency maps (p_T, η)
- Fails to capture the complete picture



Limitations of the binned maps



- Issues -
  - θ is known partially
  - Dimension of θ is large
  - Dimension of θ is not constant (influence of neighboring jet)
- Histogram based maps cannot capture the full dependencies of the classifier efficiency

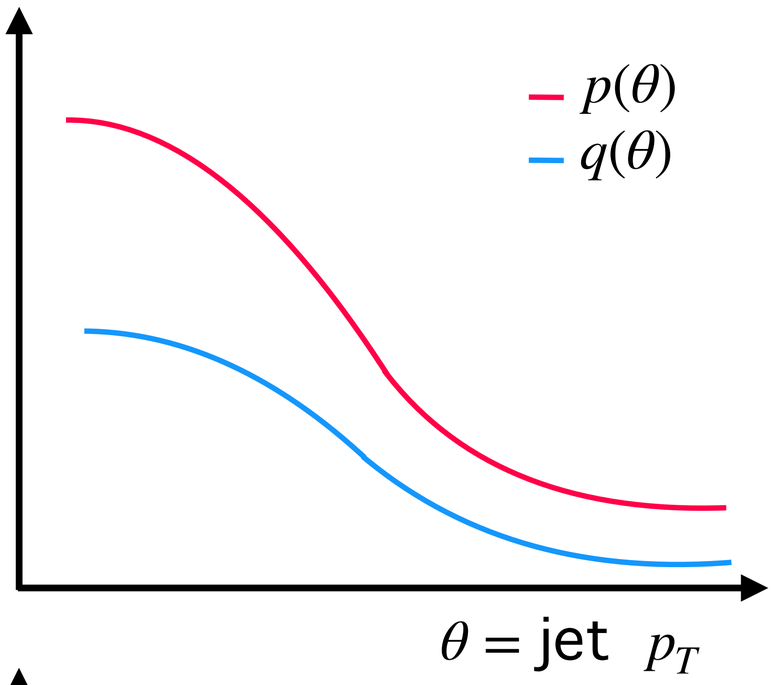
- The paper proposes an NN based approach to address these issues

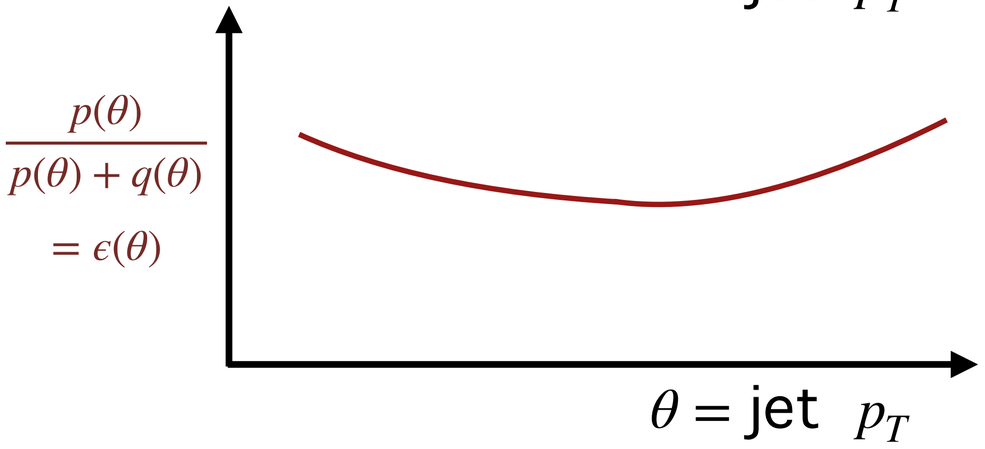
The NN approach (background)

- Density Ratio Estimation

- Two distributions - $p(\theta)$ and $q(\theta)$
  - $p(\theta)$ - Distribution of tagged jets
  - $q(\theta)$ - Distribution of non-tagged jets

- If we train a binary classifier $g(\theta)$, it converges to -
  - $g(\theta) \approx \dfrac{p(\theta)}{p(\theta) + q(\theta)} = \epsilon(\theta)$ = efficiency





The NN approach

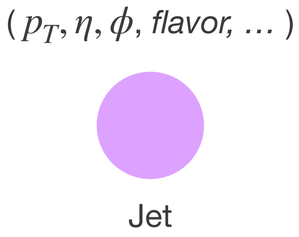
- " θ is not fully known "

- We want the network to -
    - Infer θ during training
    - Consider the jet-jet dependency
    - Work with variable number of jets in an event
    - Be permutation invariant wrt the jets

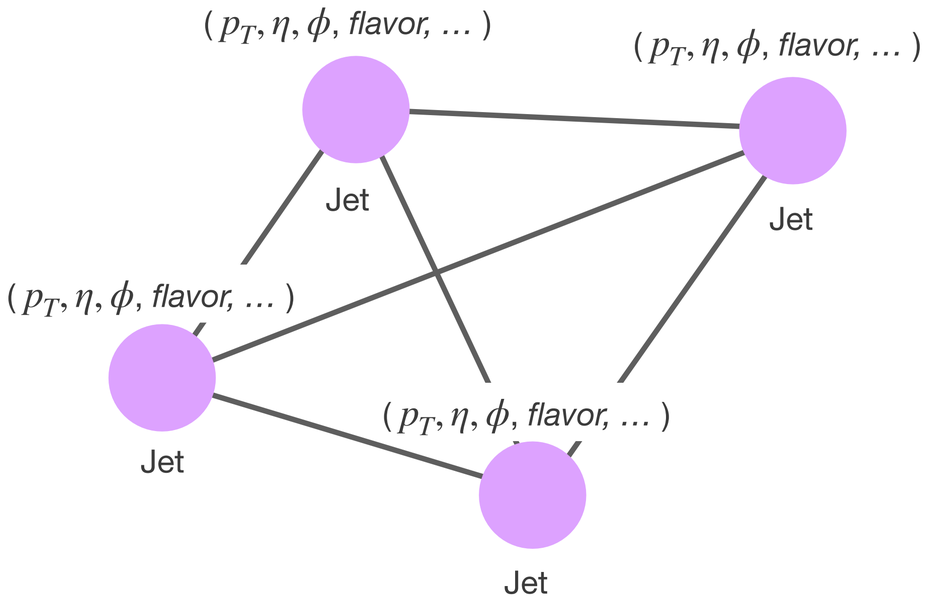
-> Graph Neural Network (GNN)

Event Representation

A node

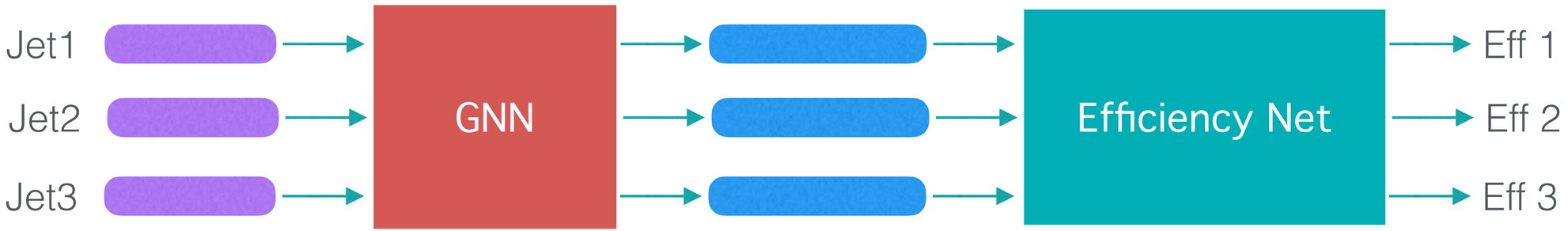


The graph



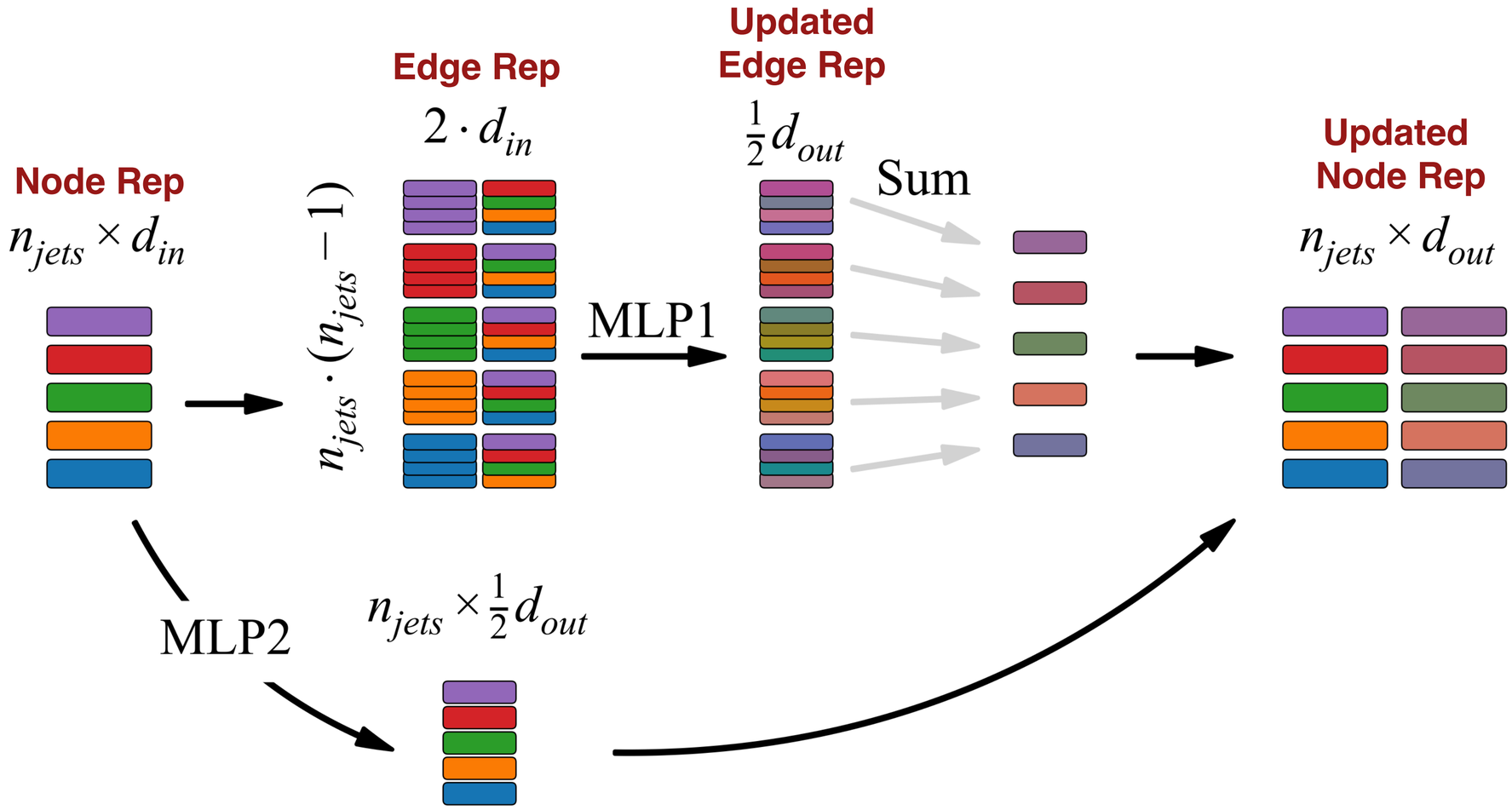
- Set of node features = main ingredients of θ.
- The network will try to construct full θ, from these inputs (hidden representation)

The NN architecture



- Construct a high dimensional representation of the jets, which is also aware of the neighborhood (θ)
- Pass the jets through a binary classifier → get the efficiencies
- The GNN consists of multiple GNN blocks (MPNNs)

A GNN block



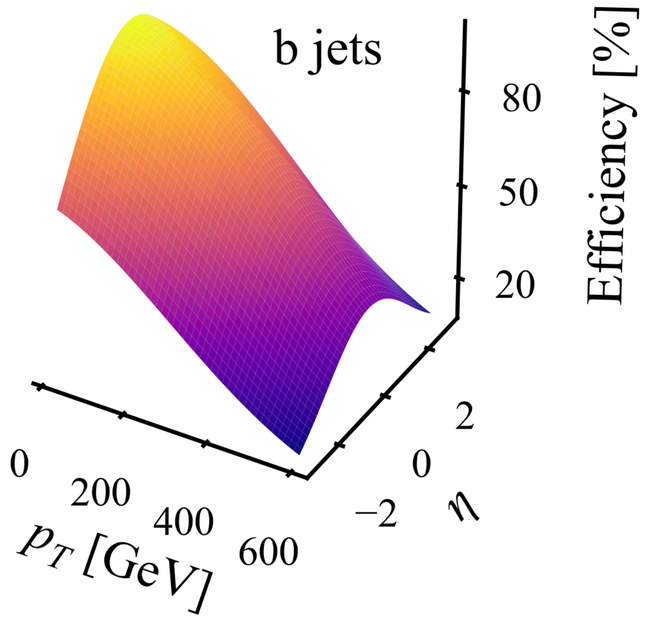
- Efficiency of a jet is conditioned on the near-by jets

Case Study

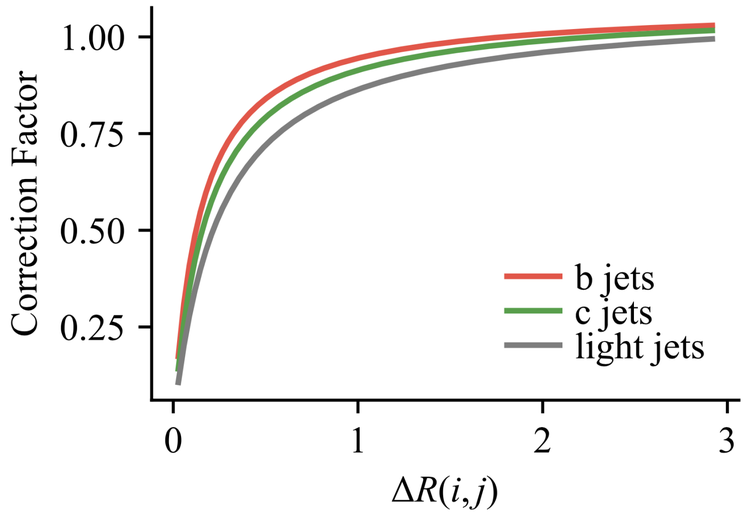
- Toy dataset - emulating W/Z+jets sample
- True efficiency of each jet is calculated as-

$$\epsilon_{jet_i} = \epsilon_{f_i}(p_T, \eta) . \prod \hat{\epsilon}_{ij} \left(\Delta R_{ij}, f_j\right)$$

Correction factor to emulate proximity effect



- B-tagging emulation
  - Generate a random number, s_i in [0,1]
  - If $s_i < \epsilon_i$, jet$_i$ is tagged



- In reality we don't know this true efficiency and the NN is supposed to learn this

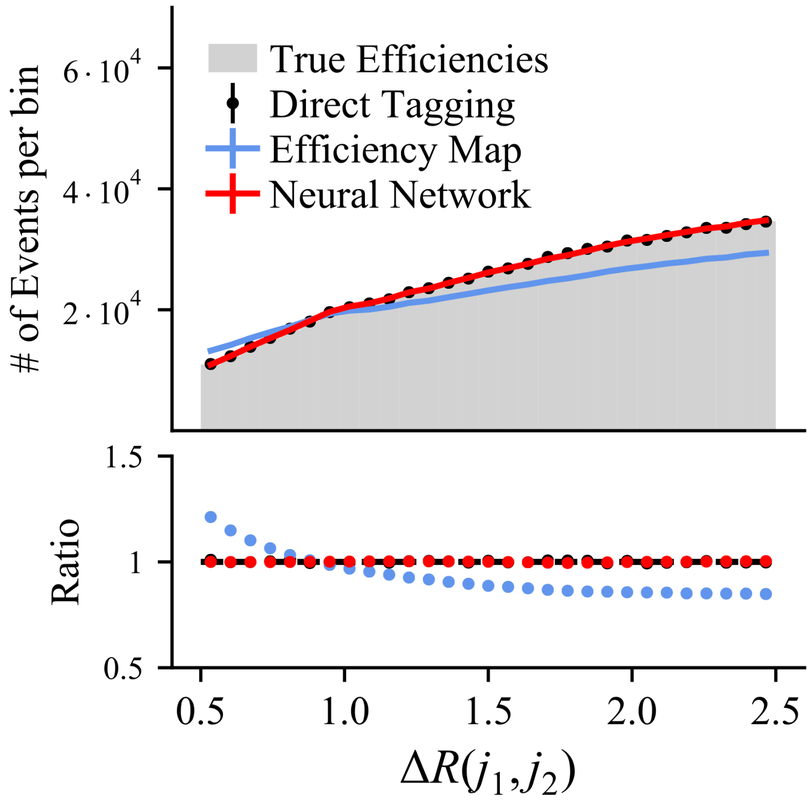
Training

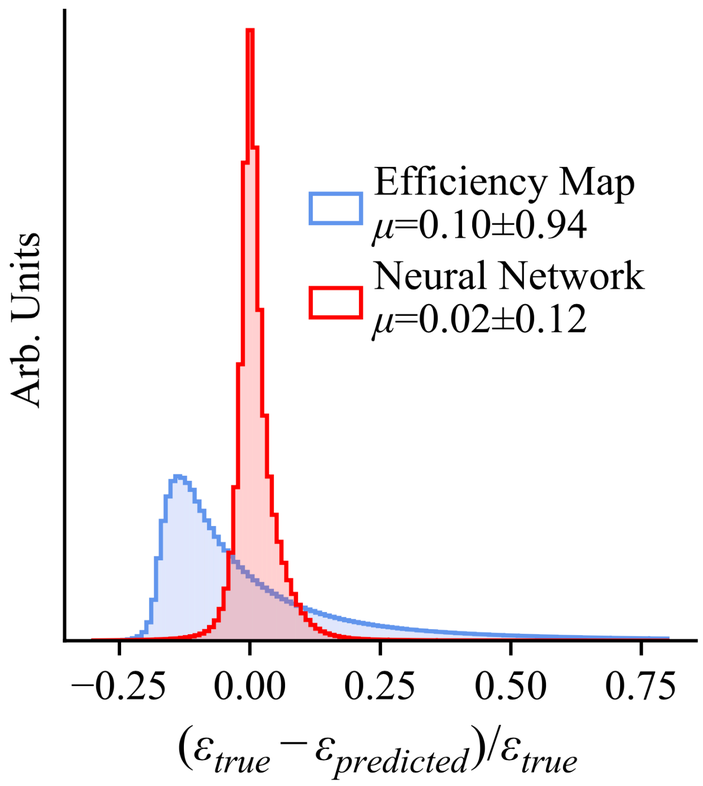
- Targets = 1/0 (jet is tagged/not tagger)
- BCE loss
- Prediction converges to (Density ratio estimation)

$$\text{pred}(\theta)_i = \frac{p_{tag}(\theta)_i}{p_{tag}(\theta)_i + p_{non-tag}(\theta)_i} = \text{efficiency}(\theta)_i \quad \text{, for jet } i$$

Result I

- Event is tagged if the leading jet is b-tagged; event weight = efficiency of leading jet

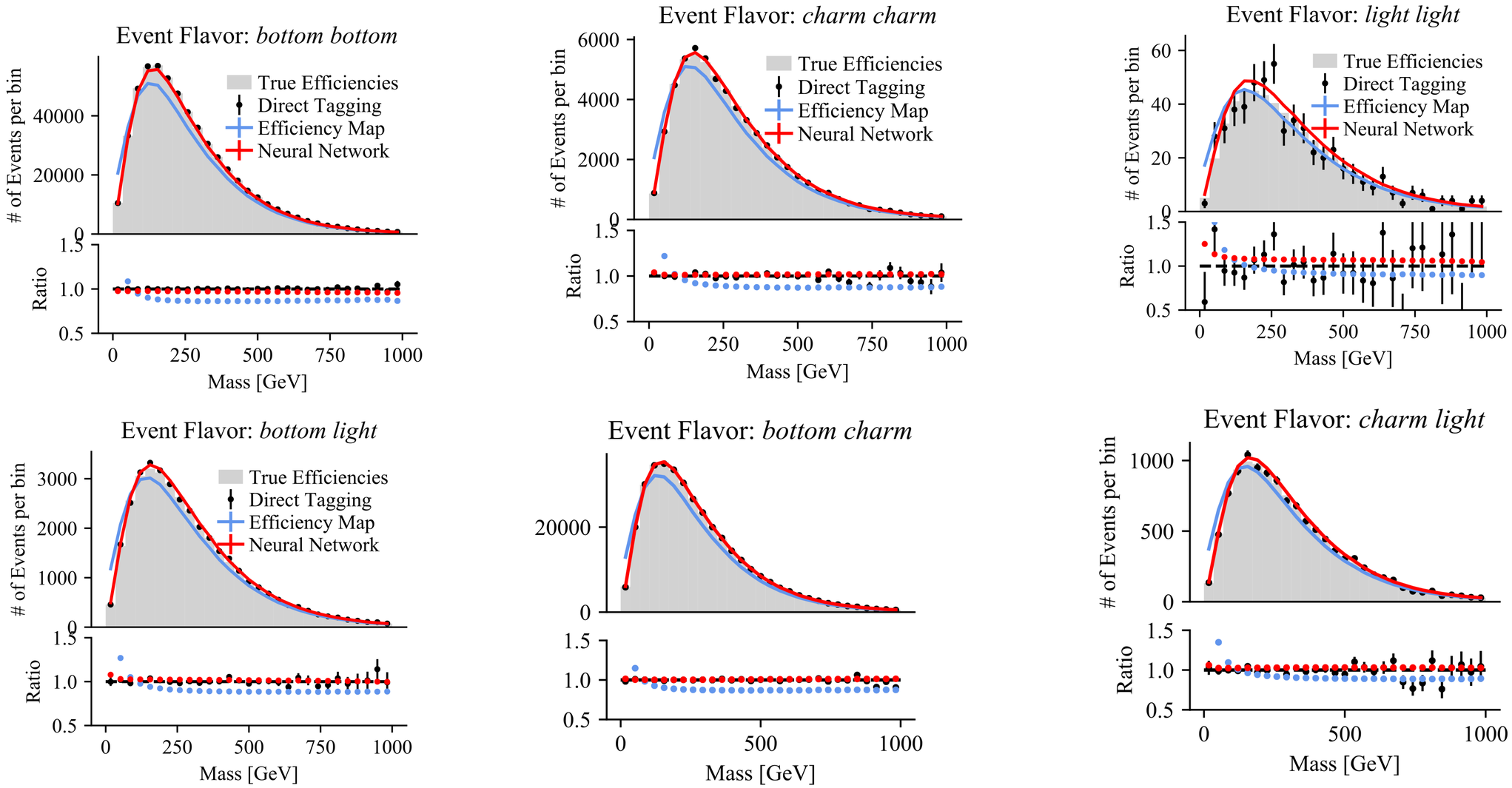




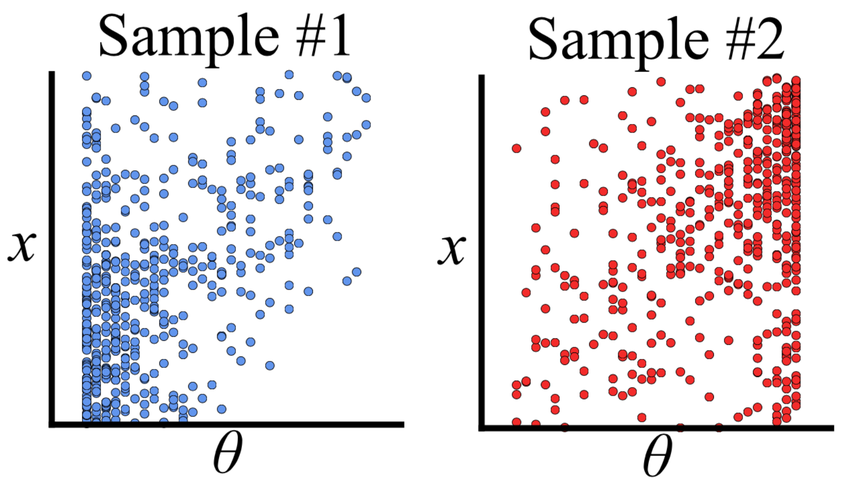
- We know that dR is also a part of θ, and the network learns it during training

Result II

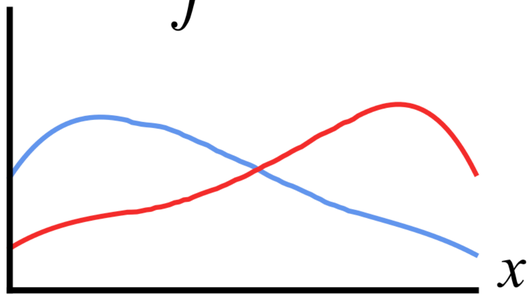
- Event is tagged if the leading and sub-leading jet are b-tagged. Event efficiency = $\epsilon_1 \epsilon_2$

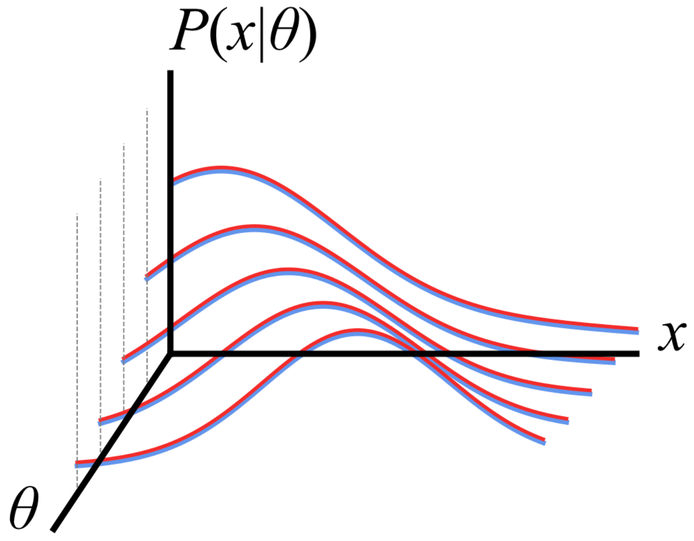


Generalization



$$P(x) = \int P(x|\theta) d\theta$$

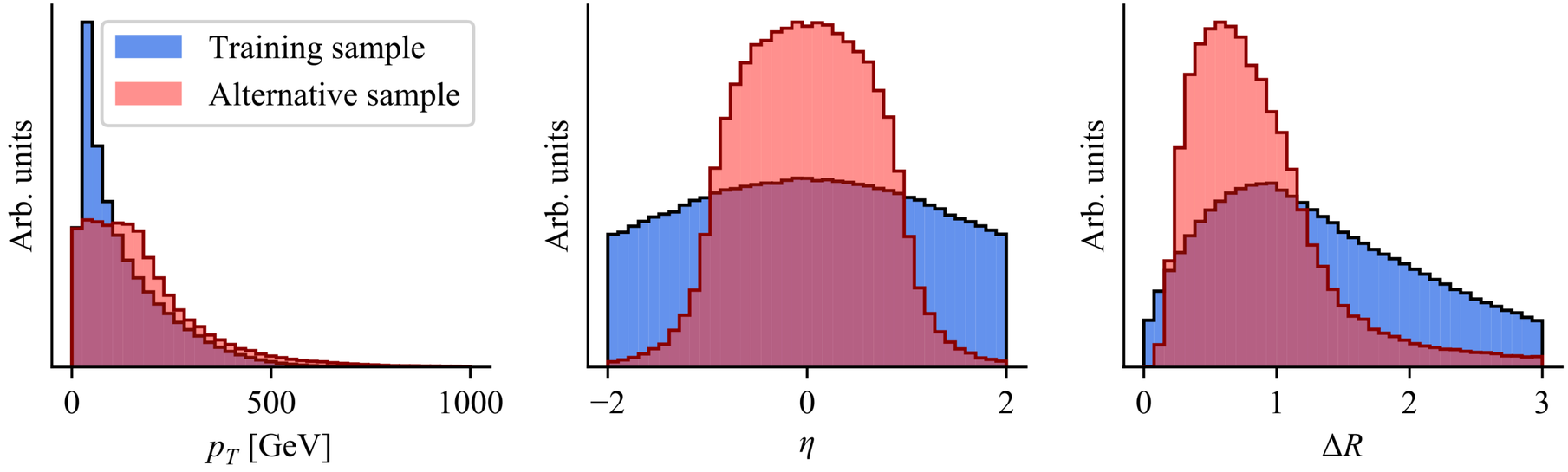




- $p(X)$ is different for both the samples
- But $p(X|\theta)$ is same for both
- Parametrizing the efficiency in theta with sample 1 should also work in sample 2

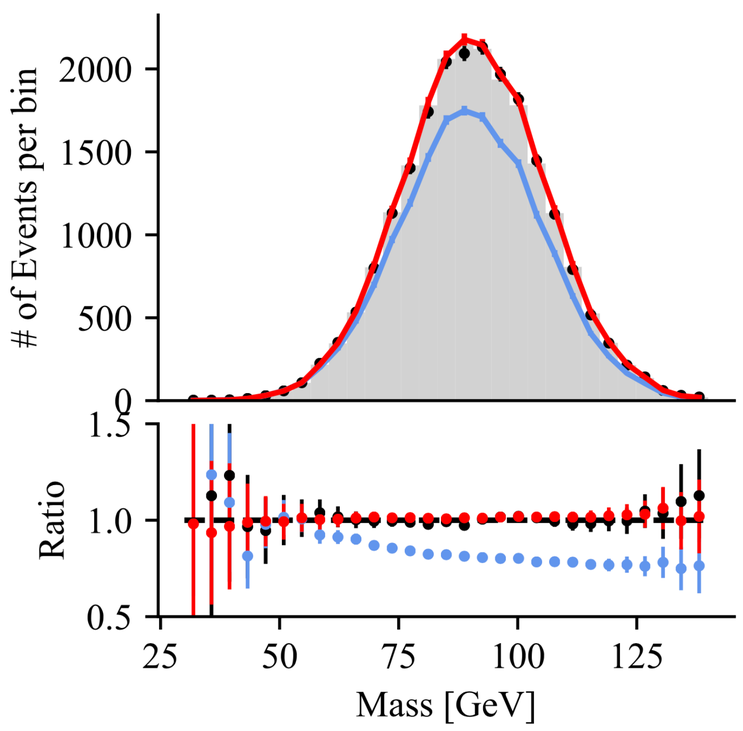
Alternative sample

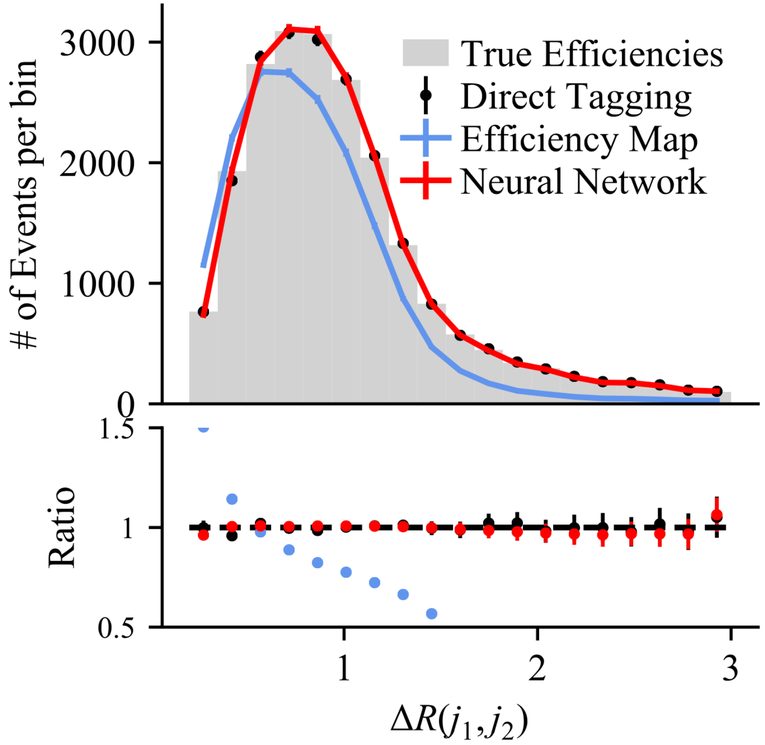
- Training sample - emulating W/Z + jets
- Alternative sample - emulating boosted scalar decay with exactly two jets
- The NN is NOT trained with the alternative sample



Result III - Generalisation

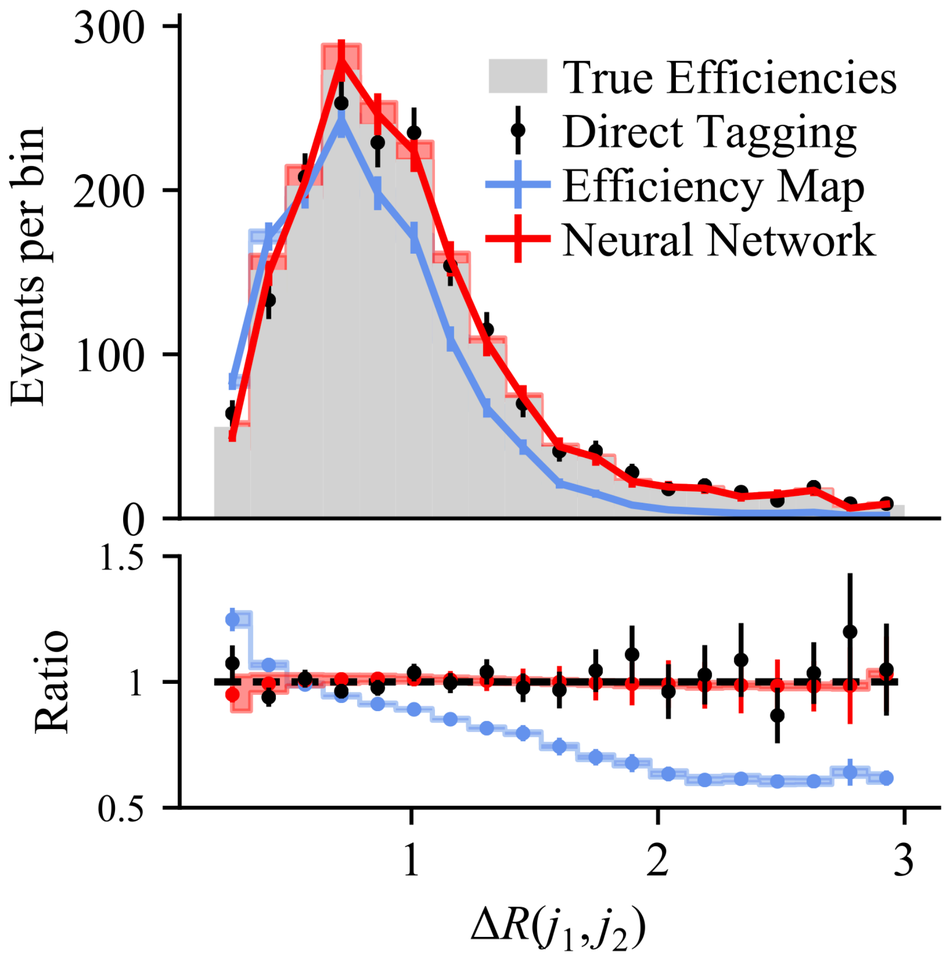
- Evaluating the NN in the alternative sample





Statistical uncertainty estimation

- Ensemble training to stabilize the model
- Bootstrapping for statistical uncertainty estimation
- Computationally expensive!
- In practice (from studies in ATLAS), quite small uncertainties, often can be neglected



Current work

- Results from ATLAS specific studies are very promising
- Significant gain in statistics while maintaining good closure with direct tagging
- Uoto 120x more events!!
- Helps in -
  - Signal vs Background discrimination
  - Alternative sample studies
  - Final fit

- *public plots coming soon…*

Summary

- We discussed an NN based approach for efficiency estimation in a multidimensional space
- Advantages -
  - Better Efficiency estimation, as it can account for much larger number of parameters than the binned maps
  - Automatically infers theta during training
  - Learns the jet-jet dependency
  - Generalize well on sample not used for training
- The approach can be used in other studies with a similar setup - eg. fake electron identification
- Results in ATLAS look very promising (public plots will be available soon)

Thank you for listening…

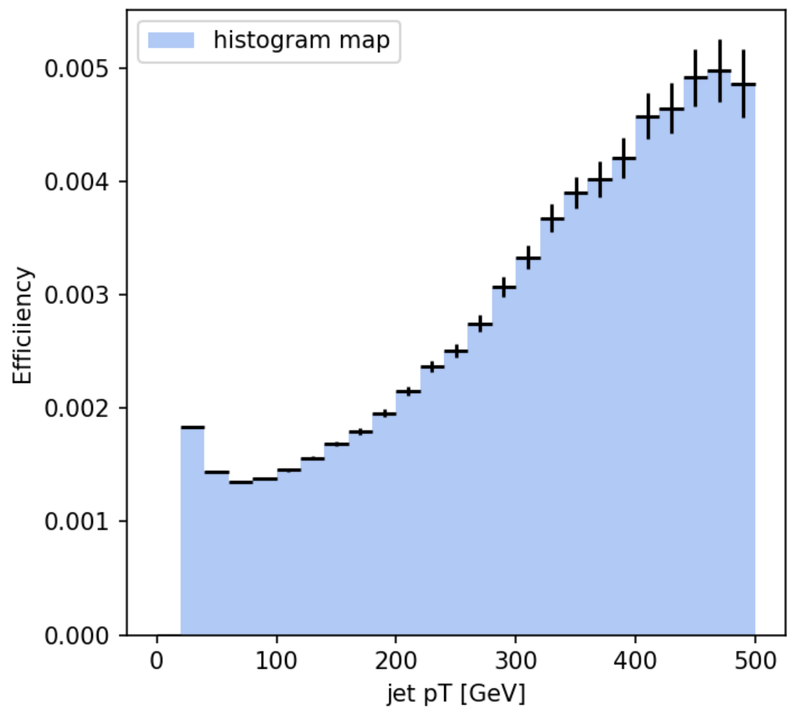
Backup

Uncertainty Estimation

- For histogram, we have limited statistics for each bin
- That helps us construct a Confidence Interval around the estimated efficiency
- Region with less data -> more uncertainty
- How to estimate the uncertainty of the NN estimator??

  -> Bootstrapping

- We would also like to stabilize the model first



Uncertainty Estimation

Training uncertainty

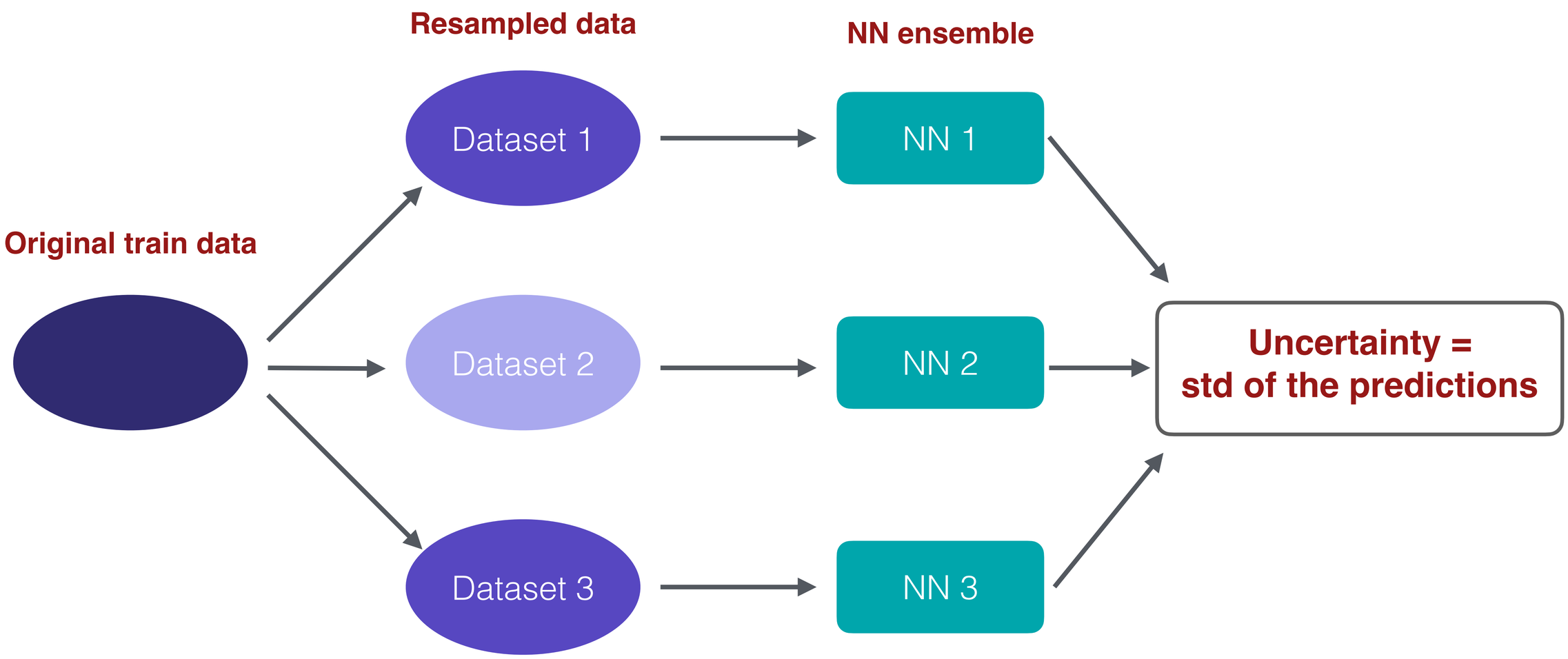
- Repeated training won't result in the same trained model
- Solution
  - Model = Ensemble of NN
  - Model prediction = ensemble prediction avg

Statistical uncertainty

- Uncertainty due to limited training data
- Solution
  - Bootstrapping

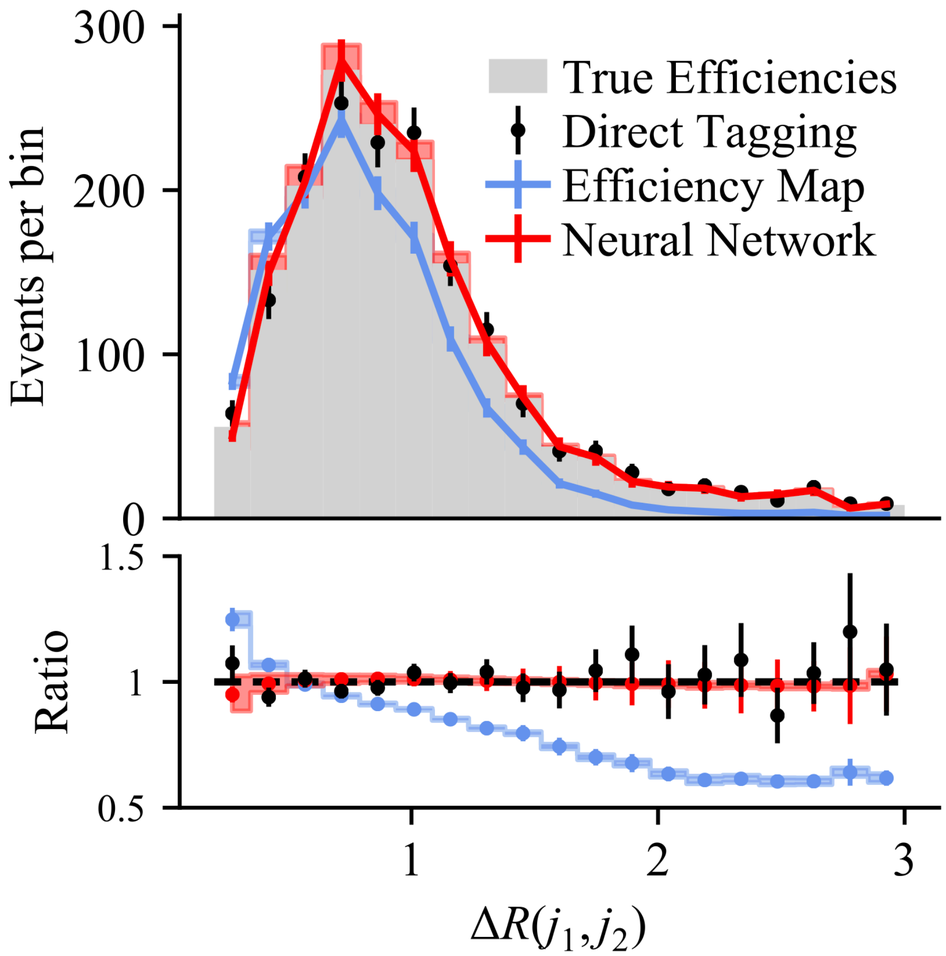
- Model Uncertainty cannot be decoupled from the stat uncertainty. So we try to reduce its impact during bootstrapping by using ensemble training

Bootstrapping



Comments about uncertainty estimation

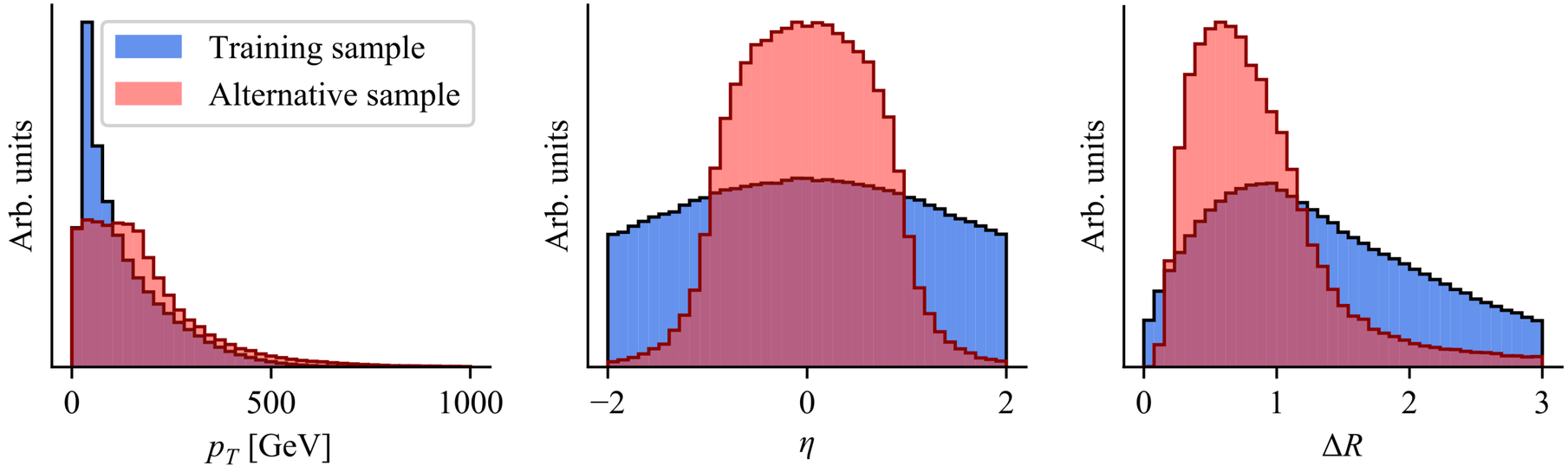
- Quite expensive!
- In practice (from studies in ATLAS), the trainings are quite stable
- Quite small, often can be neglected
- Nevertheless, this provides us a well-defined procedure to obtain them



Case Study - Dataset generation

- "Training sample" emulates W/Z+jets events
- number of jets in an event and the jet kinematics (p_T, η, ϕ) are sampled from a distribution
- The distribution of angular separation among jets (ΔR) also follows a predefined distribution

Let's ignore the alternative sample for now
We'll come to it later

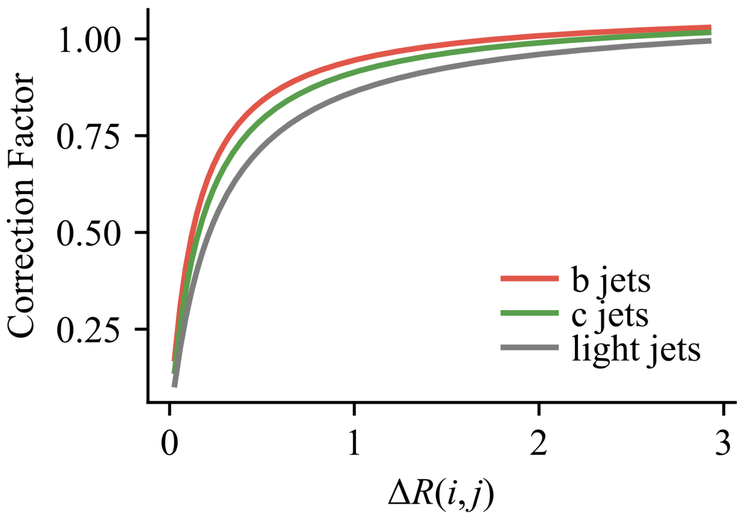


Case Study - Flavor tagging

- True efficiency each jet is calculated as-

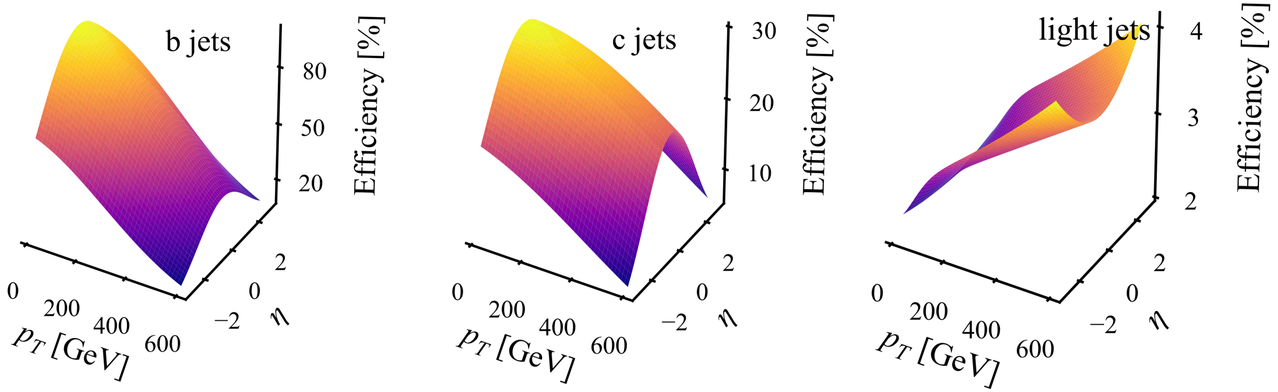
$$\epsilon_{jet_i} = \epsilon_{f_i}(p_T, \eta) . \prod \hat{\epsilon}_{ij} \left(\Delta R_{ij}, f_j\right)$$

Correction factor to emulate proximity effect



- B-tagging emulation
  - Generate a random number, s_i in [0,1]
  - If $s_i < \epsilon_i$, jet$_i$ is tagged

- In reality we don't know this true efficiency and the NN is supposed to learn this



Training

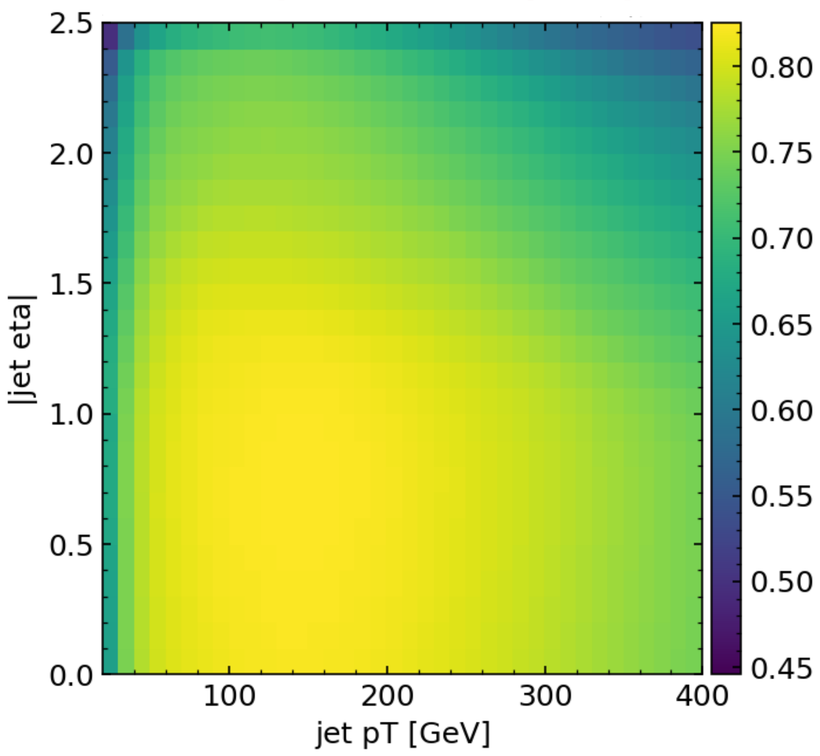
- Targets = 1/0 (jet is tagged/not tagger)
- BCE loss
- Prediction converges to (Density ratio estimation)

$$\text{pred}(\theta)_i = \frac{p_{tag}(\theta)_i}{p_{tag}(\theta)_i + p_{non-tag}(\theta)_i} = \text{efficiency}(\theta)_i \quad \text{, for jet } i$$

Binned maps (for comparison)

- Histogram based efficiency maps were constructed from the jets
- Binned in pT and η
- Goal - to mimic how we would address the issue with the “traditional” binned approach

Example efficiency map



*not from the actual study

Comments about theta

- Efficiency can be improved by including -
    - The pileup info
    - For b and c jets, the truth hadron info
    - Quark vs gluon light jets