Imperial College London

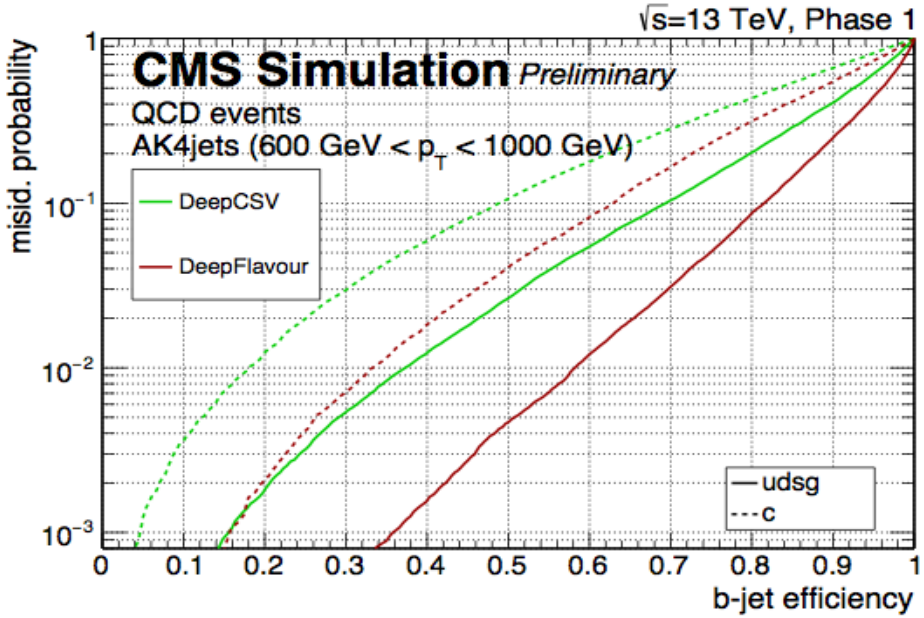
Data Science Institute

Data simulation difference mitigation studies in b-tagging

Jan Kiesele[3], Arabella Martelli[2], Markus Stoye[12], Mauro Verzetti[3]

[1]Imperial College London, [2]DSI, [3]CERN

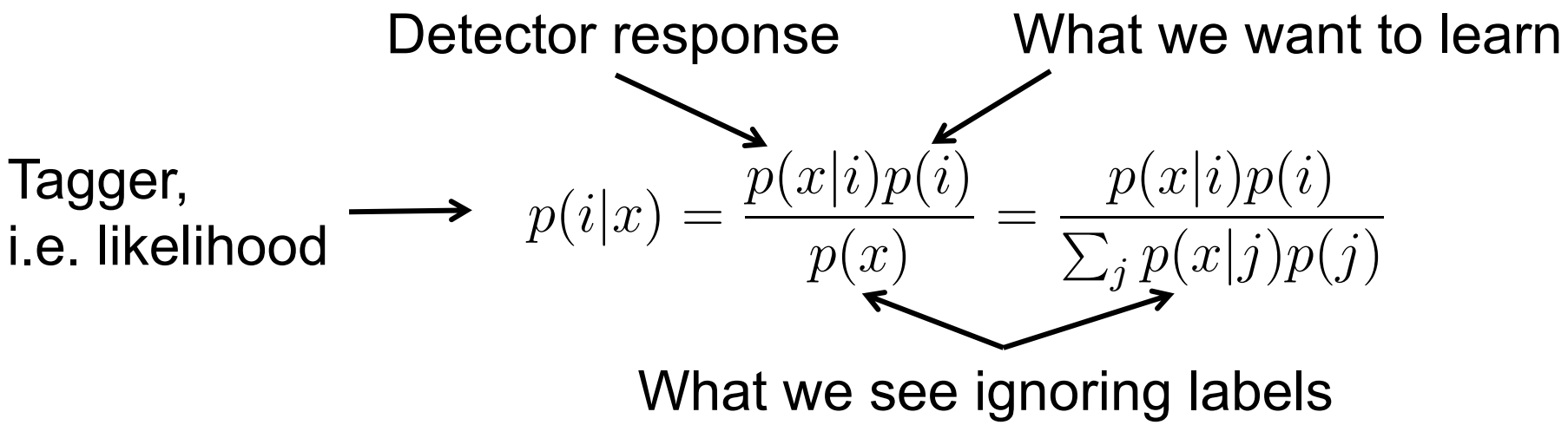
Motivation



DNN achieved big improvements (see Mauro's talk on Monday)

- However, many physics analysis sensitivity limited by systematics uncertainties that are caused by a data and simulation difference
- Try ML methods to mitigate the differences

Tagging uncertainties, label $\mathbf{i}$ and features $\mathbf{x}$



We only have $\mathrm{p(x)}$ in data and simulation

Decrease data and simulation differences

Typically we derive a correction δ as we can observe

$$p(x)^{sim} \sim p(x)^{data}$$

a) $\delta(x)$ same for data and simulation:
   - Reduces impact of disagreeing features

b) $\delta(x)^{sim}$ for simulation only:
   - Corrects label integrated features to data

c) $\delta(x,i)^{sim}$ for simulation only conditional on i
   - Has the potential to lead to full data simulation agreement

Show some work in progress on a) and c) type

Delphes details for b-tagging study

A CMS-Delphes sample with ~ 500k ttbar & QCD jets

Inputs:

- 2D and 3D impact parameter significance
- Relative transverse momentum
- Transverse momentum correlation
- Parallel component of transverse momentum

Smeared

- Increase track impact parameter resolution 2 in Delphes in “data”

➢ The most important features are smeared by the same amount

Scaled

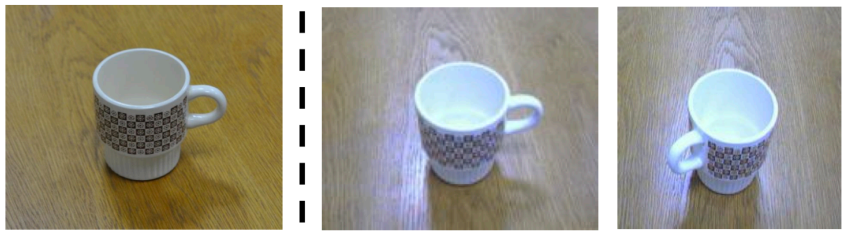
- Scaled by factor 1.2 3D impact parameters of b-jets in “simulation”

➢ Half of the most important features are changed

Domain adaptation: simulation and data

Source domain (sim) Target domain (data)

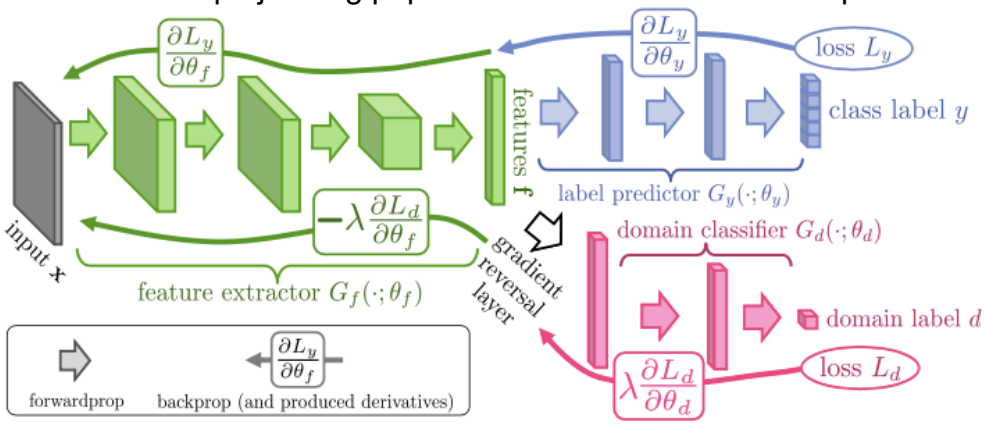
Good samples with **labels** for training a classifier



digital SLR camera low-cost camera, flash

User samples to apply the training, **no labels** available

http://jmlr.org/papers/volume17/15-239/15-239.pdf

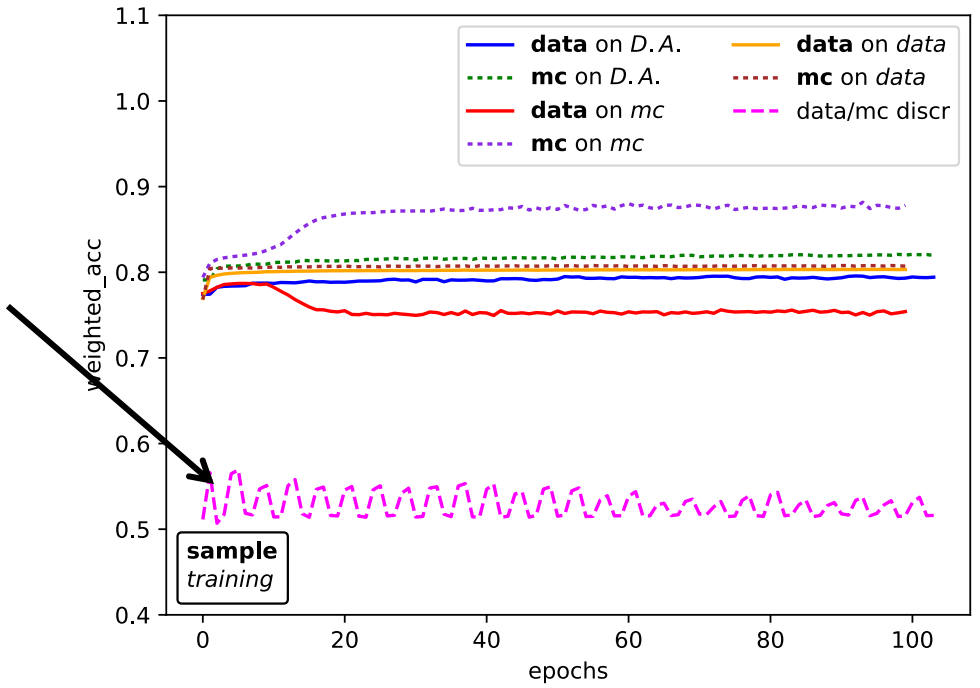


- Built intermediate features x for which data and simulation are hard to distinguish (type a)
- Use adversarial loss to ensure data and simulation similarity

Accuracy for gradient reversal

Smeared and scaled

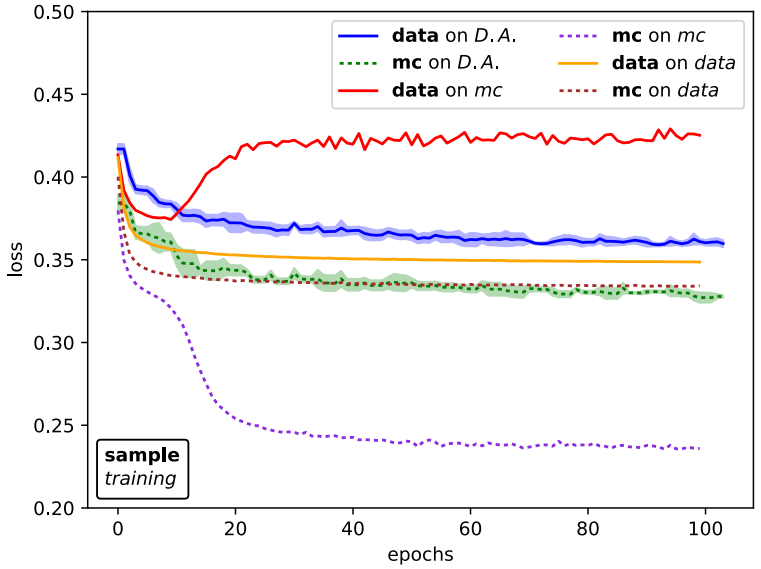
Alternate between training only the data simulation classifier and the full NN

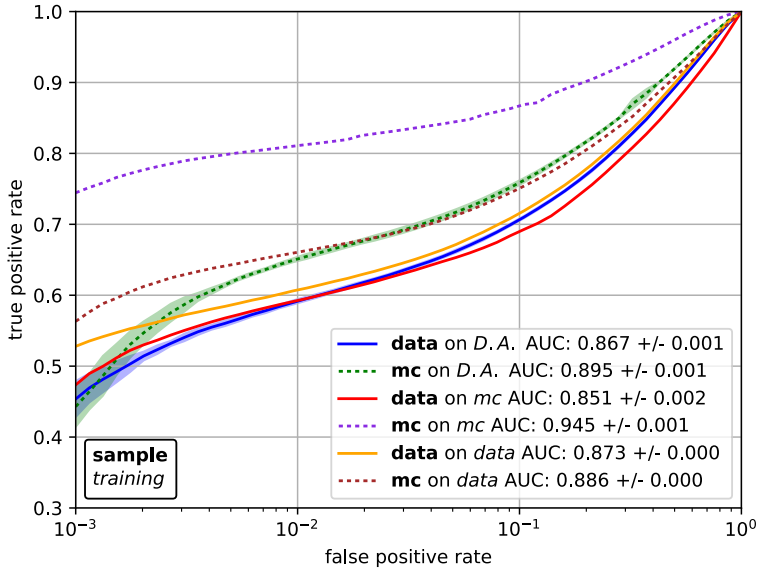


The accuracy for data improves a lot by using “data” labels

Performance for gradient reversal

Smeared and scaled

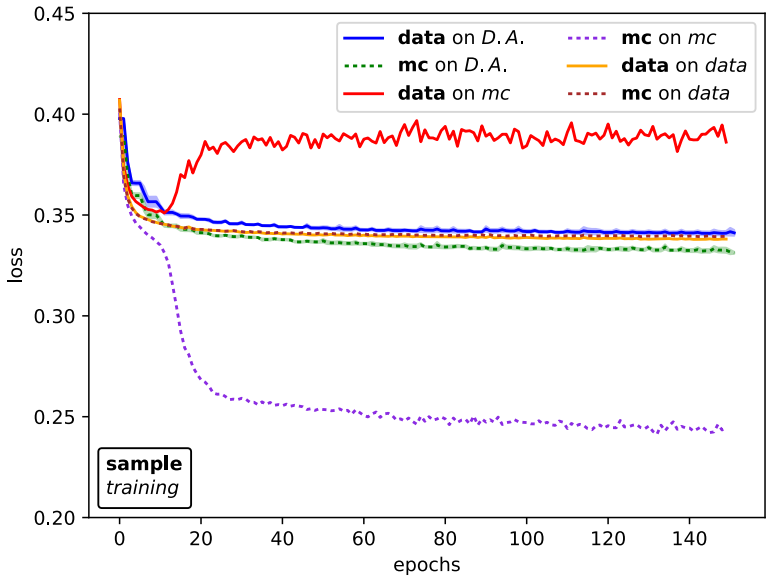


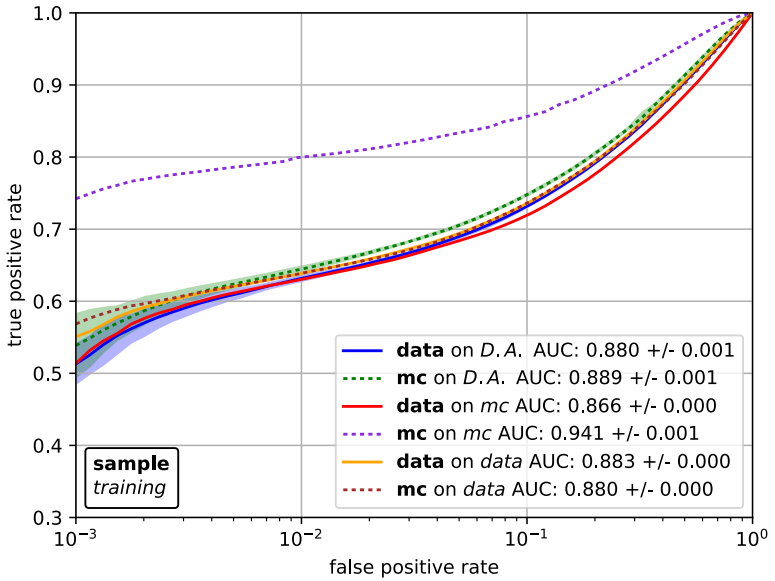


- ”Data” loss becomes better and closer to simulation
- Performance between data and simulation much closer
- Data performance better

Performance for gradient reversal

Scaled only

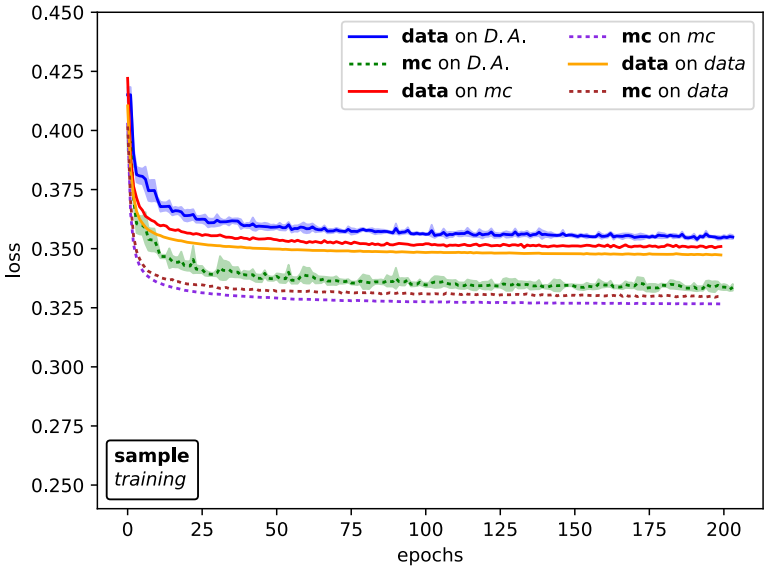


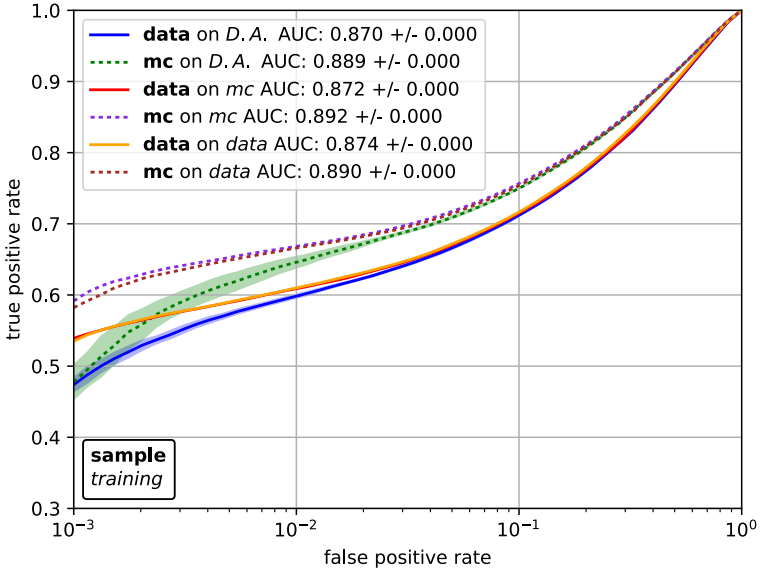


- Data performance almost reaches data on data training performance
- MC moves significantly towards data. Differences almost negligible

Performance for gradient reversal

Smeared only



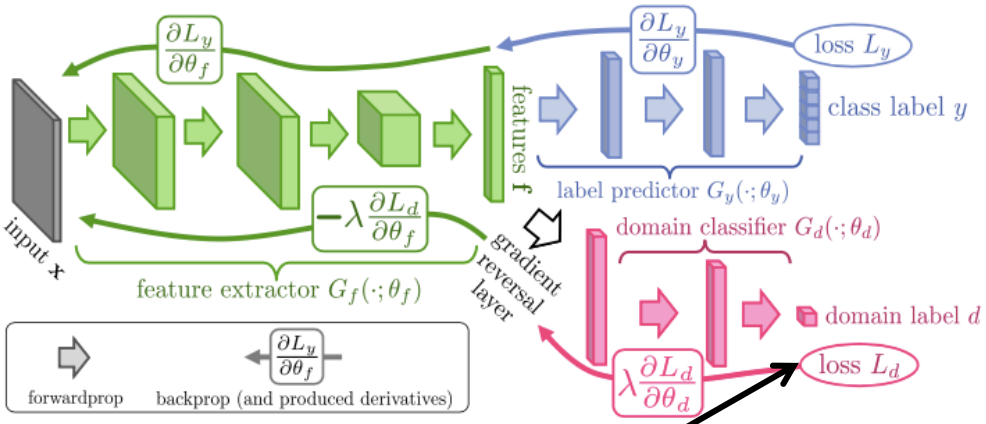


- Data only slightly improves by using “data” labels
- Performance between data and simulation equal at 0.1%
- Data performance simulation above 1% fake rate at worse at 0.1%

What if p(i) is uncertain as well

Reweight prior

$$p(x|i)w(i)p(i)$$



$$\text{weighted loss} = \frac{\sum w(i) loss}{\sum w(i)}$$

input i → NN → w^{sim}

This mitigates negative effect wrong simulated label fraction (or nuisance parameters)

AUC	no adv. noss	grad. rev. adv	grad. rev. adv. with label stift	grad. rev. adv. with label shift and w estimate
	0.91	0.94	0.89	0.94

Works in simple toy example, work in progress for b-tagging

What if we have more than on data sample?

Apply conditional corrections which keep data simulation similar (type c)

The data pdf is a linear combination of the conditional pdf for the label to be true or false

$$pdf_j^{data} = f_j pdf_{true}^{data} + (1 - f_j) pdf_{false}^{data} = pdf_{false}^{data} + f(pdf_{true}^{data} - pdf_{false}^{data})$$

If we have more than one sample with different f, we can determine the shape of the difference between the conditional pdfs

$$pdf_j^{data} - pdf_i^{data} = \beta(pdf_{true}^{data} - pdf_{false}^{data})$$

Problem: we do not precisely know *fs and thus not* β

Best fit regularization

We can reformulate the problem as determining γ_{true} and γ_{false}

$$pdf_{true}^{data} = pdf_{j}^{data}\gamma_{true}(pdf_{j}^{data} - pdf_{i}^{data})$$
$$pdf_{false}^{data} = pdf_{j}^{data}\gamma_{false}(pdf_{j}^{data} - pdf_{i}^{data})$$

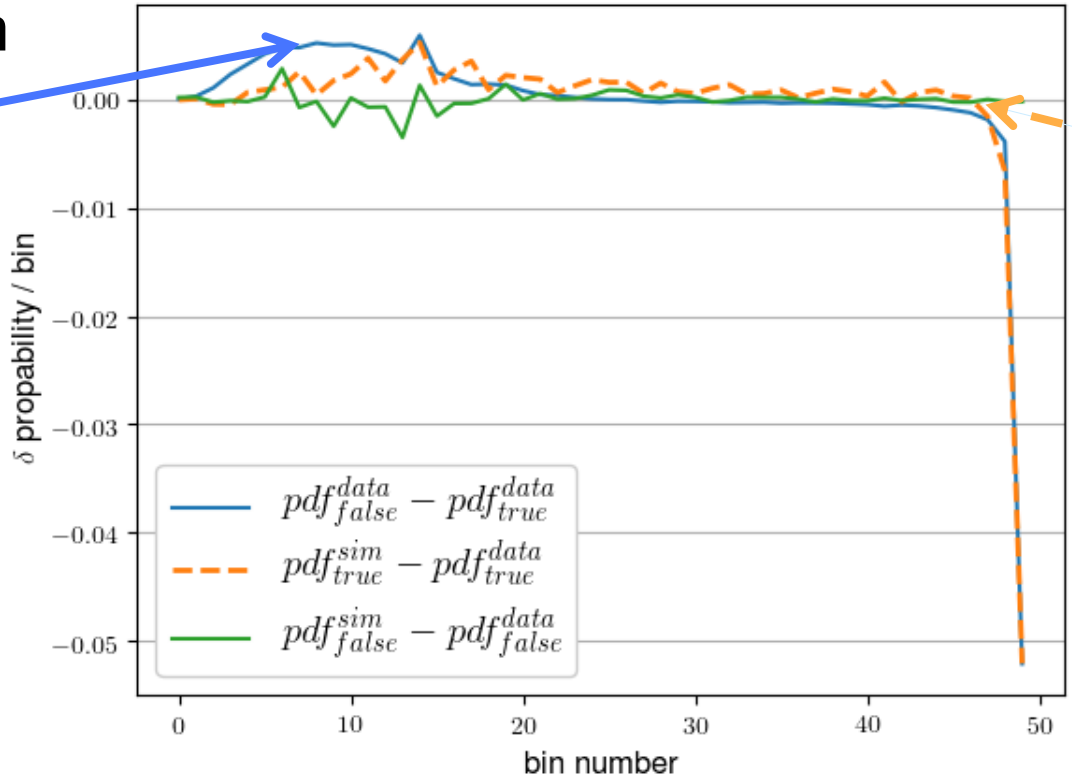
- We need a clever regularization that picks the right γs
- Assuming the simulation is a reasonable approximation we can choose:

$$\frac{(pdf_{true}^{data} - pdf_{true}^{sim})^2}{pdf_{true}^{sim}}$$

We use the solution that best fits (in a χ^2 sense) the simulation

Best fit regularization

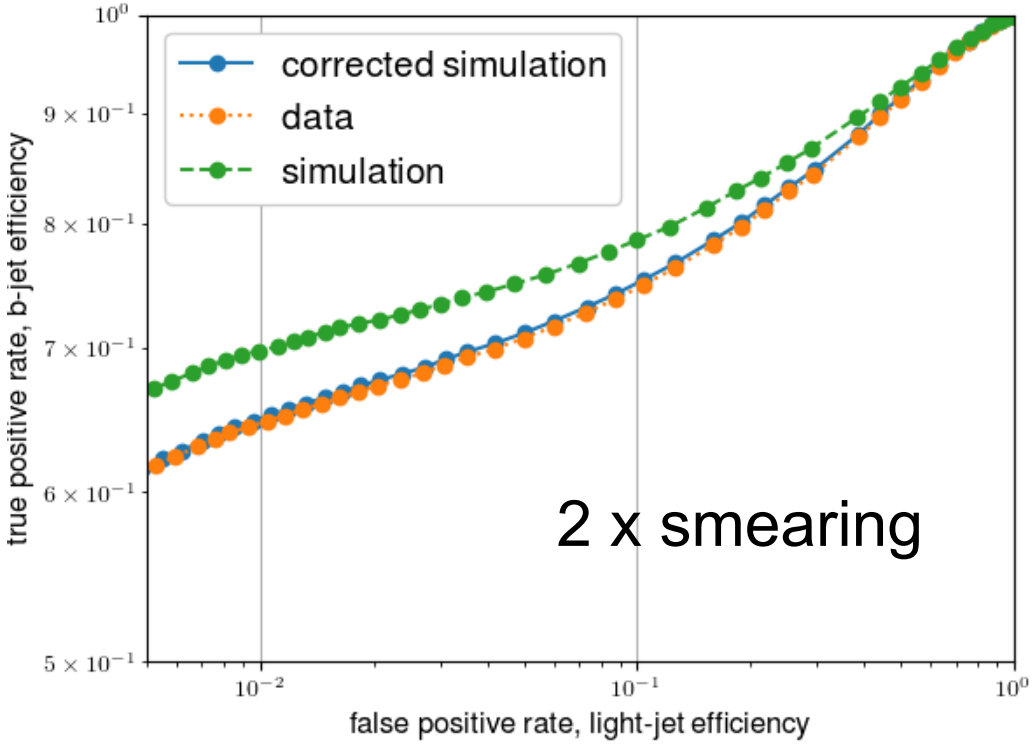
Peak from lights jets



Anti-correlation Due to bin migration

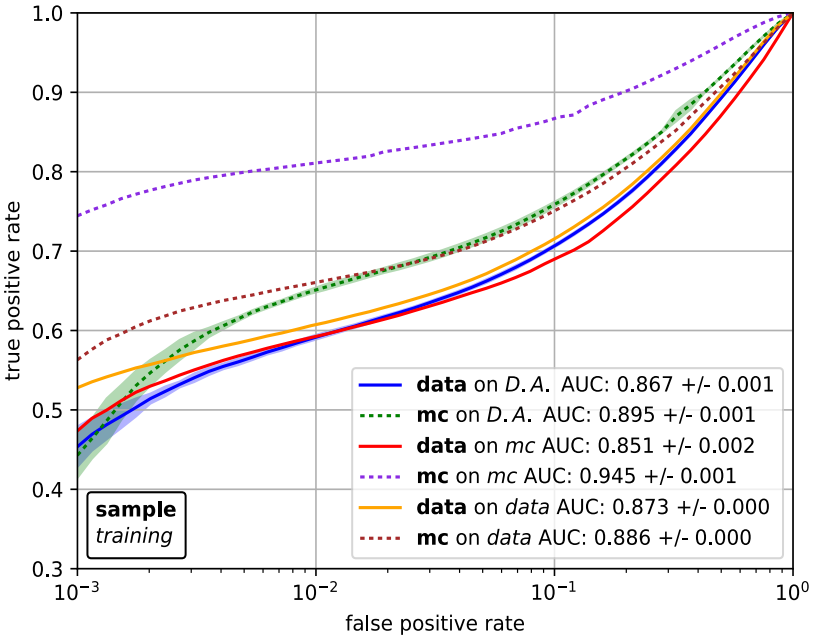
- True correction to simulation and the label shift
- High correlation between label shift and correction to b-jets

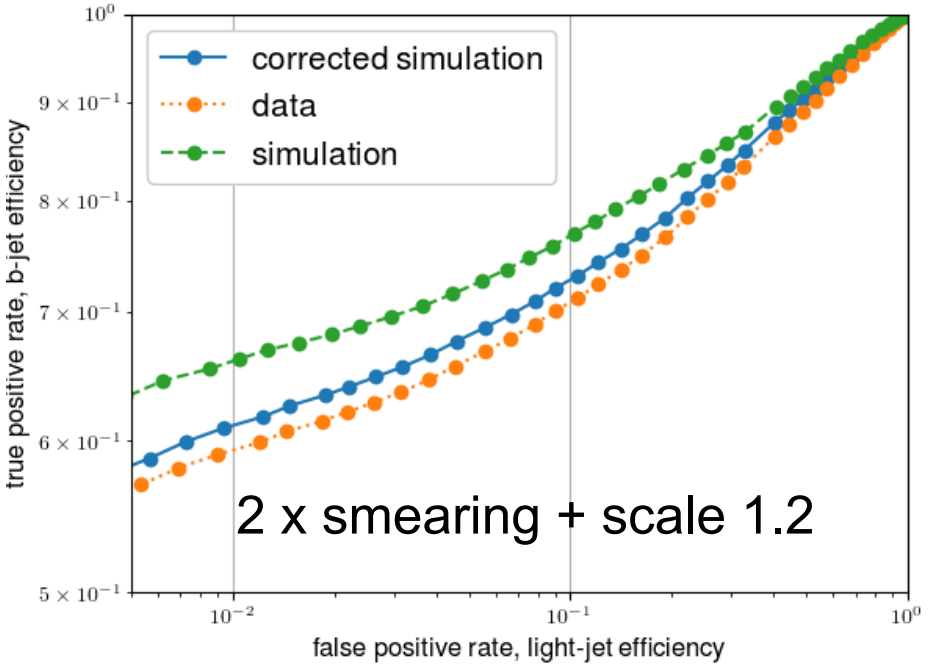
ROC with corrected simulation



Picking the correction according to best fit regularization yielded good correction.

ROC with corrected simulation





Picking the correction according to best fit regularization yielded reasonable correction.

Summary

We looked at a simple b-tagging example and tried tools to mitigate data simulation differences and improve

1. Using gradient reversal for DA some improvement was observed
2. Extended DA method to simultaneously estimate known and modeled nuisance parameters and fraction
3. Showed a regularization that picks the closest solution to simulation that matches data in case of several datasets

In this setting data and simulation disagreement reduced significantly and improved data performance.