Learning to Classify from Impure Samples with High-Dimensional Data

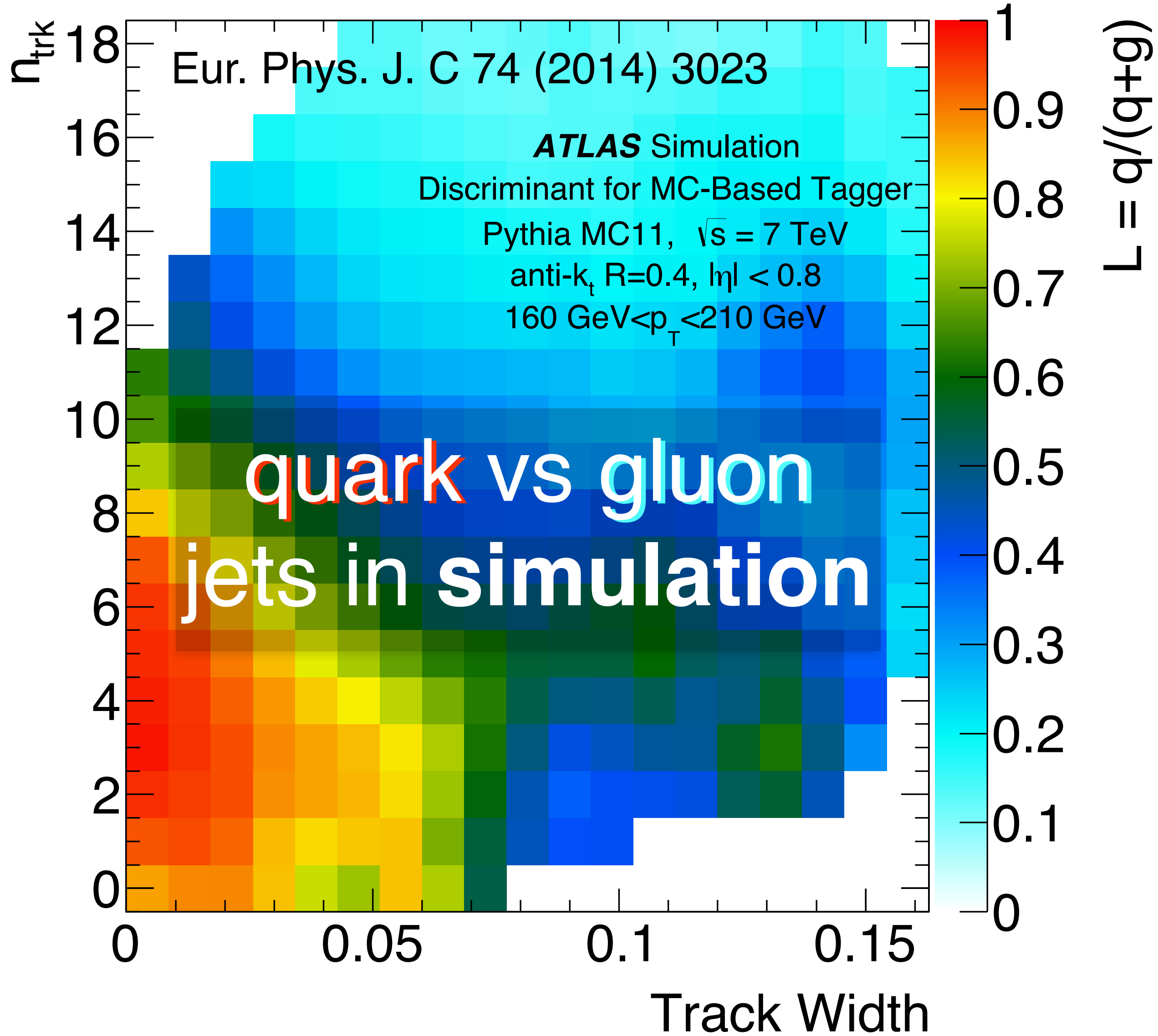
based on Phys. Rev. D 98, 011502(R) [published yesterday!]

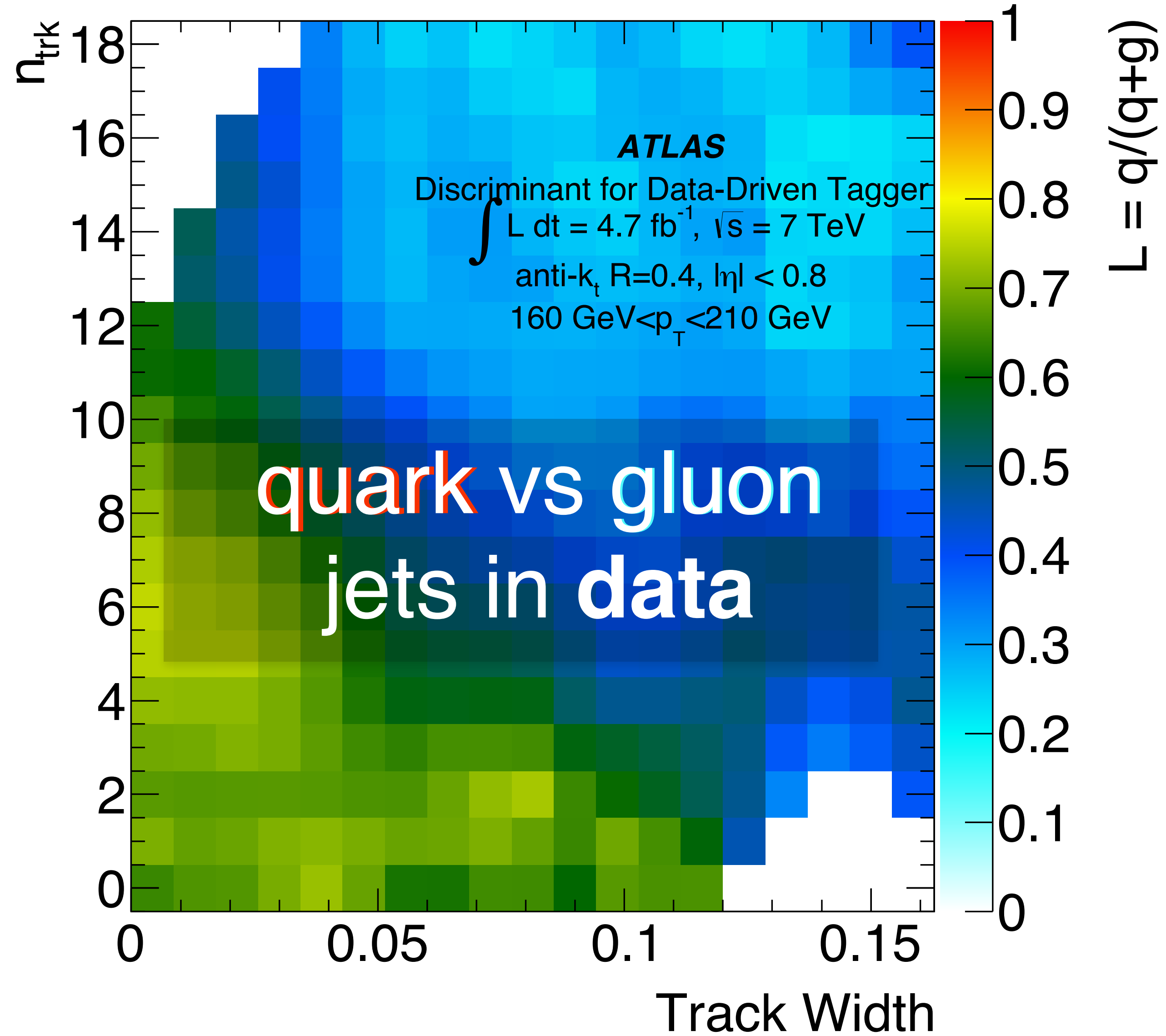
Patrick Komiske, Eric Metodiev,
Benjamin Nachman, Matt Schwartz

…building on work also in collaboration with Lucio Dery, Francesco Rubbo, Ariel Schwartzman, Jesse Thaler

Motivation

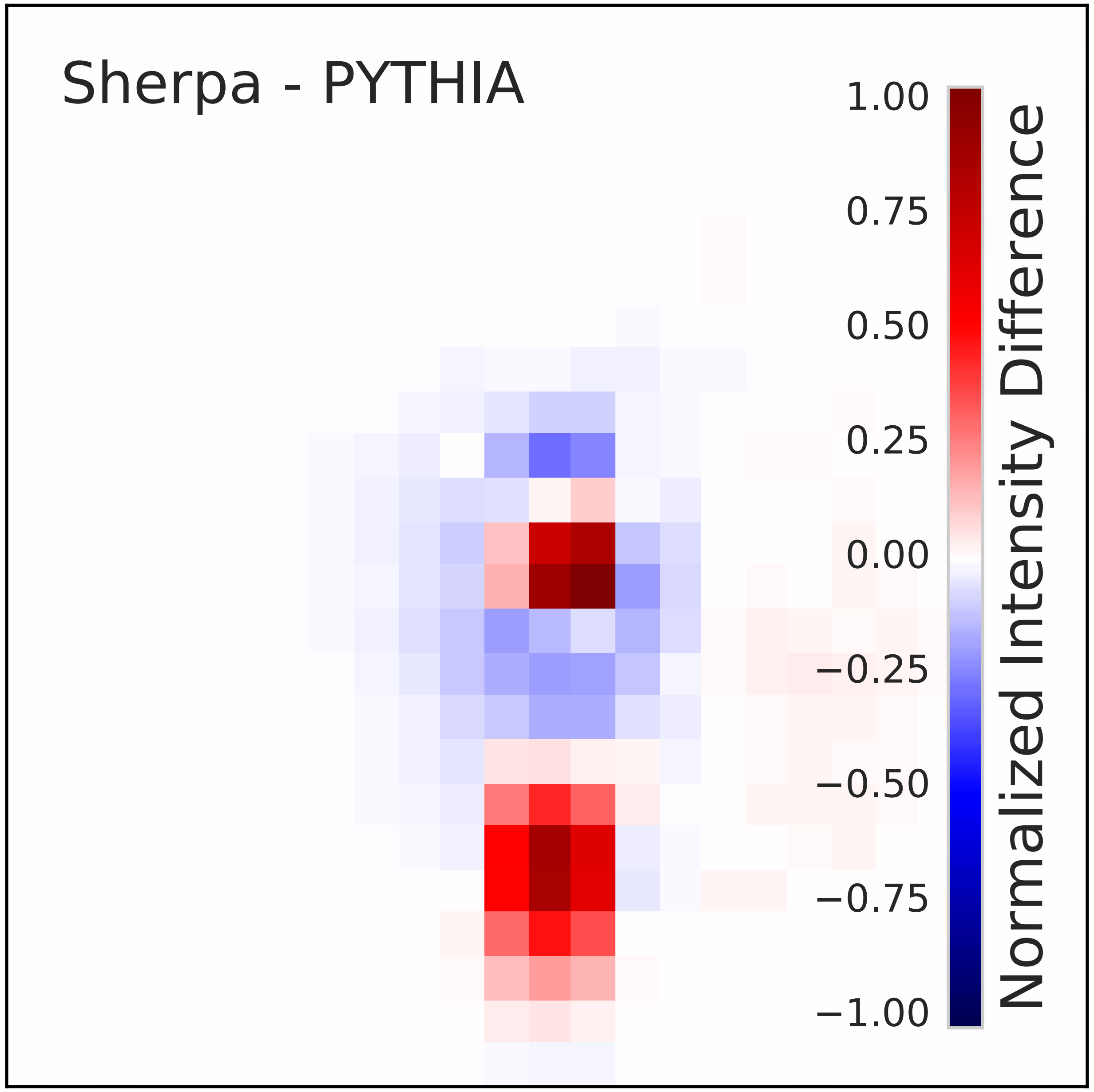
Usual paradigm: train in simulation, test on data.





If data and simulation differ, this is **sub-optimal**!

Motivation, continued



J. Barnard, E. Dawe, M. Dolan, N. Rajcic, Phys. Rev. D 95 (2017) 014018

Especially important for **deep learning** using subtle features → hard to model!

W boson radiation pattern - same physics, different simulators!

Solution 1: Use class proportions

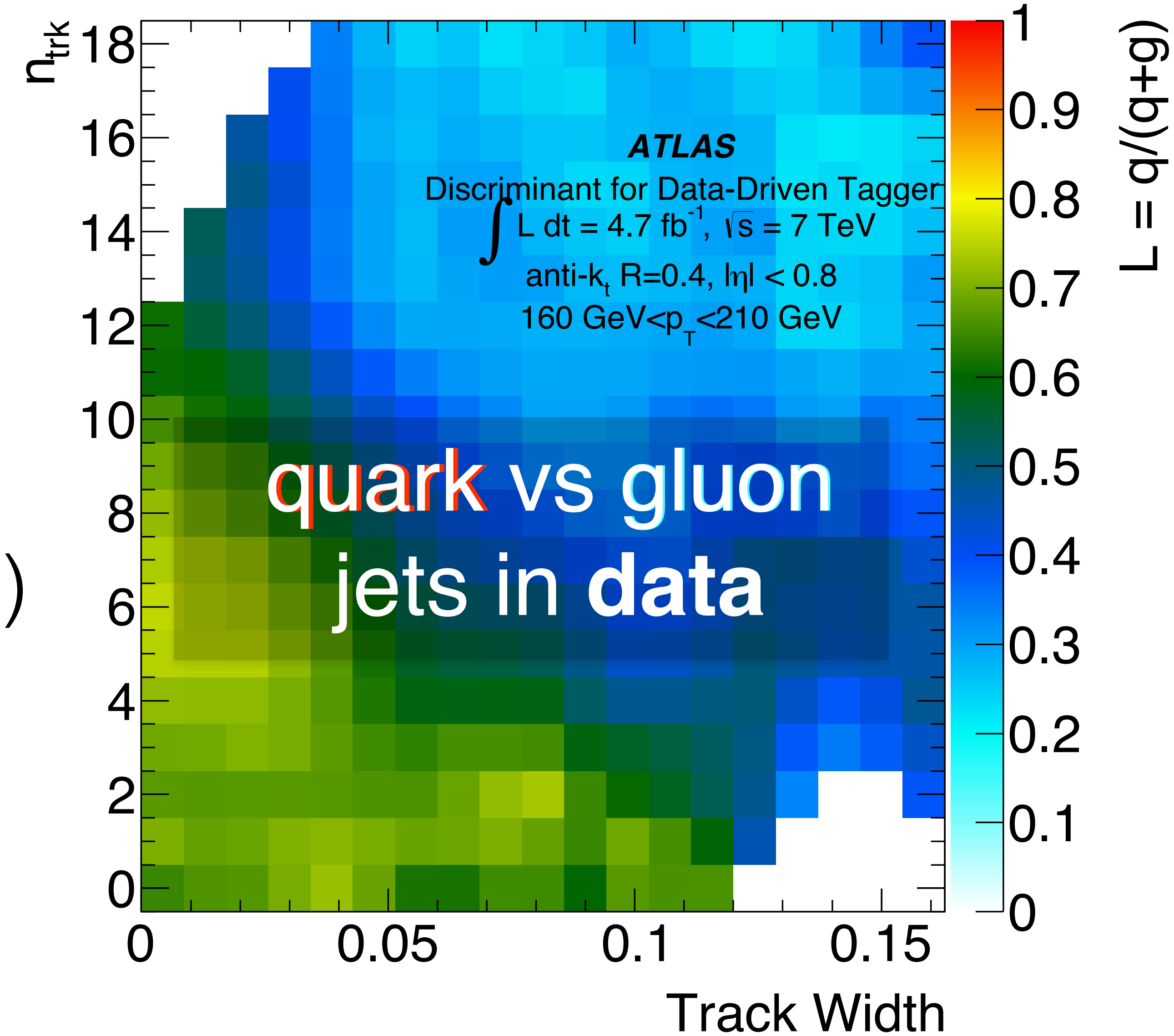
How did we make this plot?

dijets = f_q x Q + (1-f_q) x G

Z+jets = g_q x Q + (1-g_q) x G

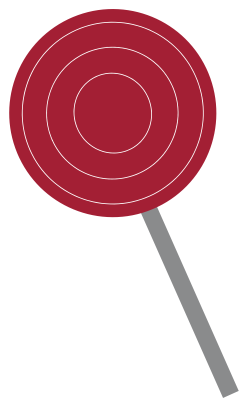
two equations, two unknowns (Q, G)

We often know f, g
(from ME + PDF) much better than
full radiation pattern inside jets.



This doesn’t work well when you have more than 2 observables because the templates become sparse.

Method 1: Learn from Proportions

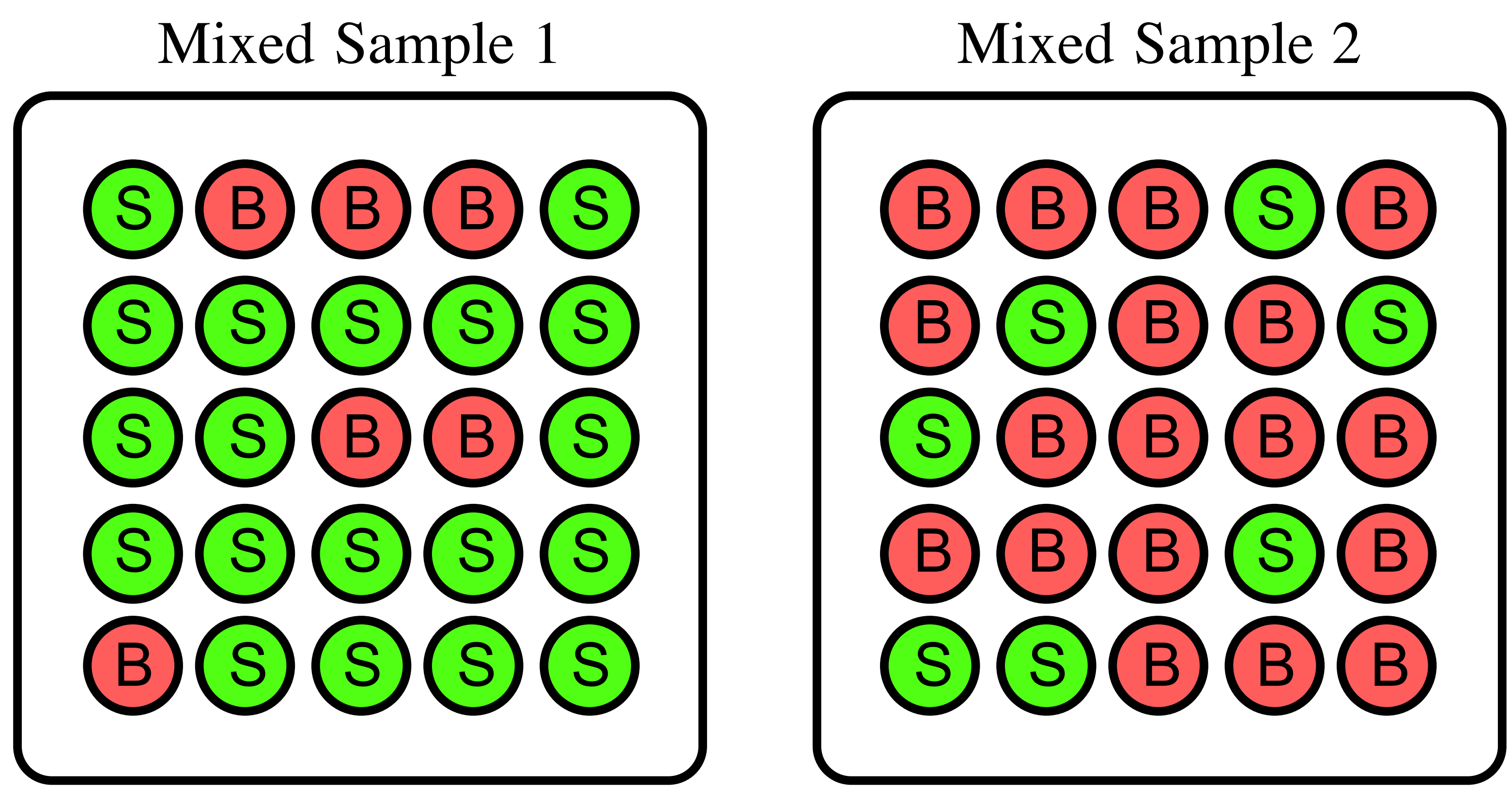


LoLiProp

Learning from Label Proportions

loss fcn. labels

$$f_{\text{full}} = \text{argmin}_{f':\mathbb{R}^n \to \{0,1\}} \sum_{i=1}^{N} \ell\left(f'(x_i) - t_i\right)$$



Solution: Train using class proportions. Work "on average"

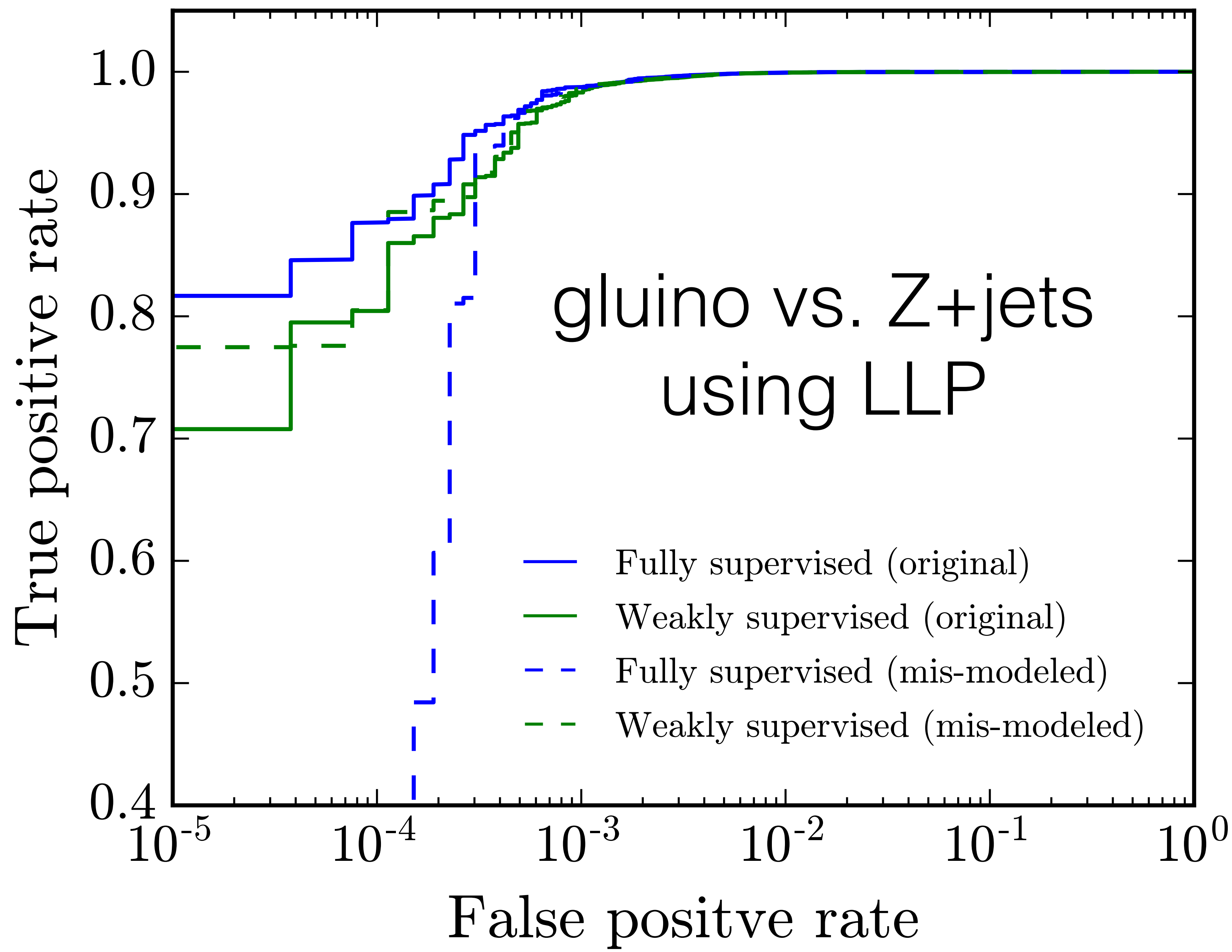
$$f_{\text{weak}} = \text{argmin}_{f':\mathbb{R}^n \to [0,1]} \ell\left(\sum_{i=1}^{N} \frac{f'(x_i)}{N} - y\right)$$

proportions

L. Dery, **BPN**, F. Rubbo, A. Schwartzman, JHEP 05 (2017) 145

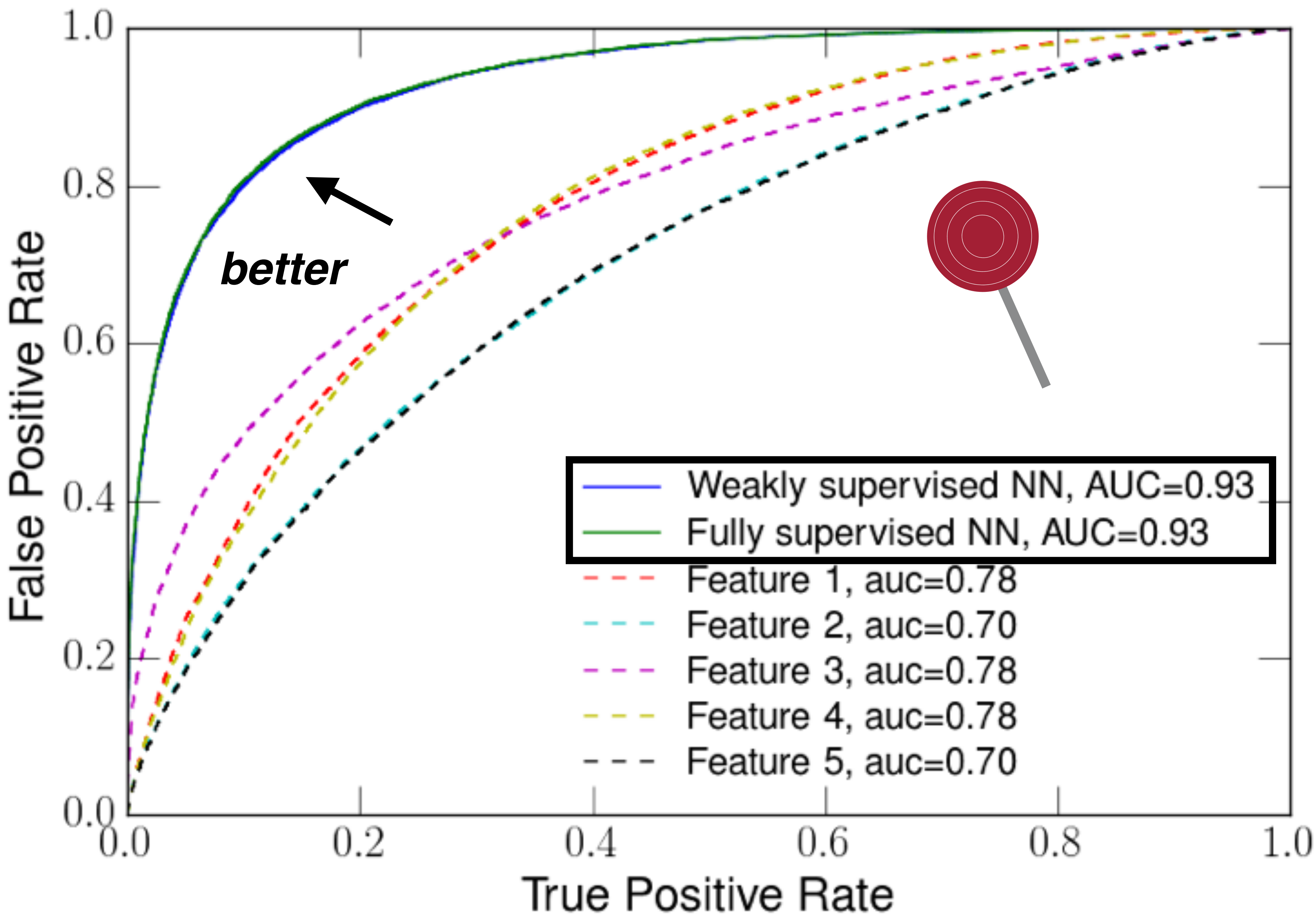
N.B. Don't need 100% fraction accuracy

Even though the proportions are required as input, if they are slightly wrong, you can end up with the correct classifier.



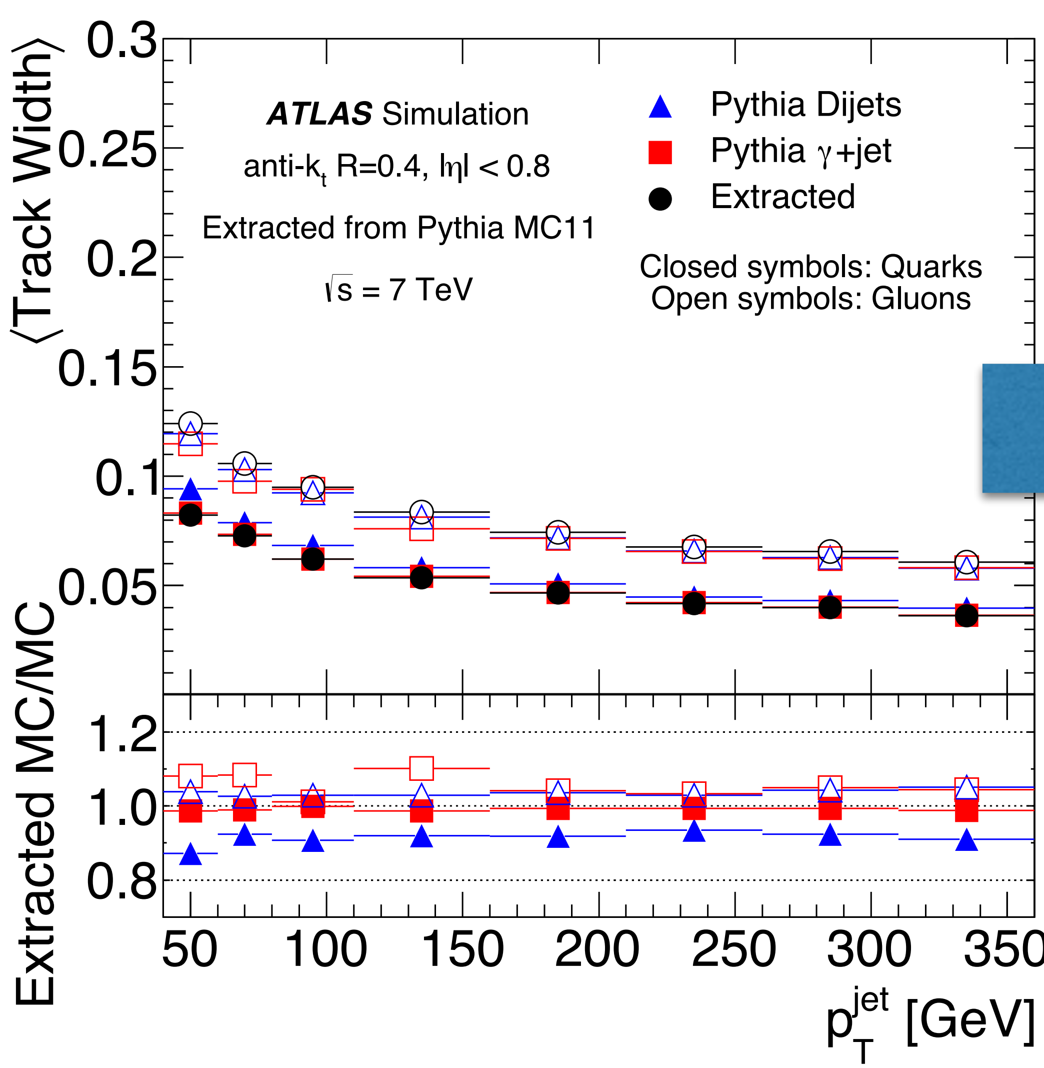
T. Cohen, M. Freytsis, B. Ostdiek,https://arxiv.org/abs/1706.09451

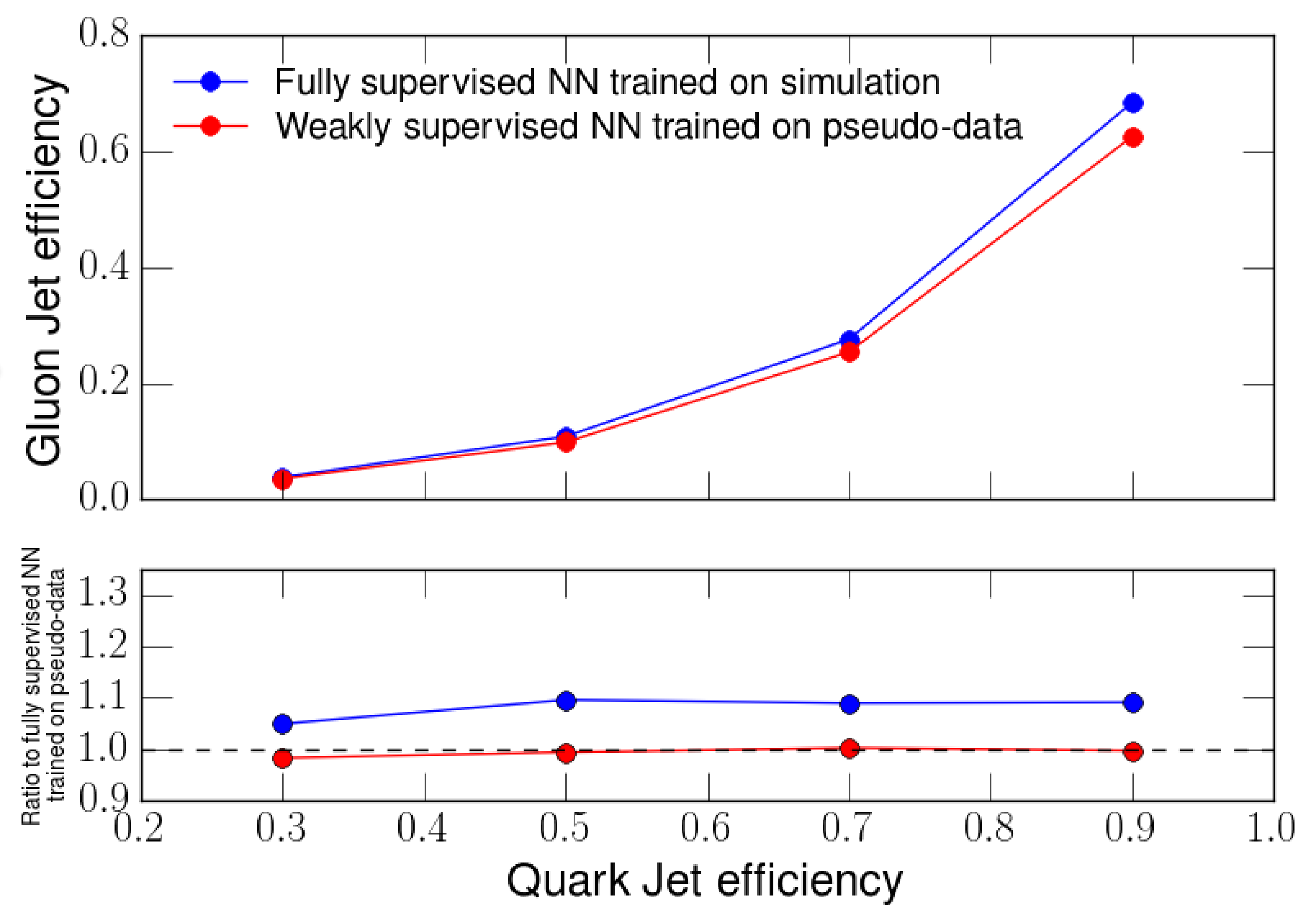
Works in low-dimensions



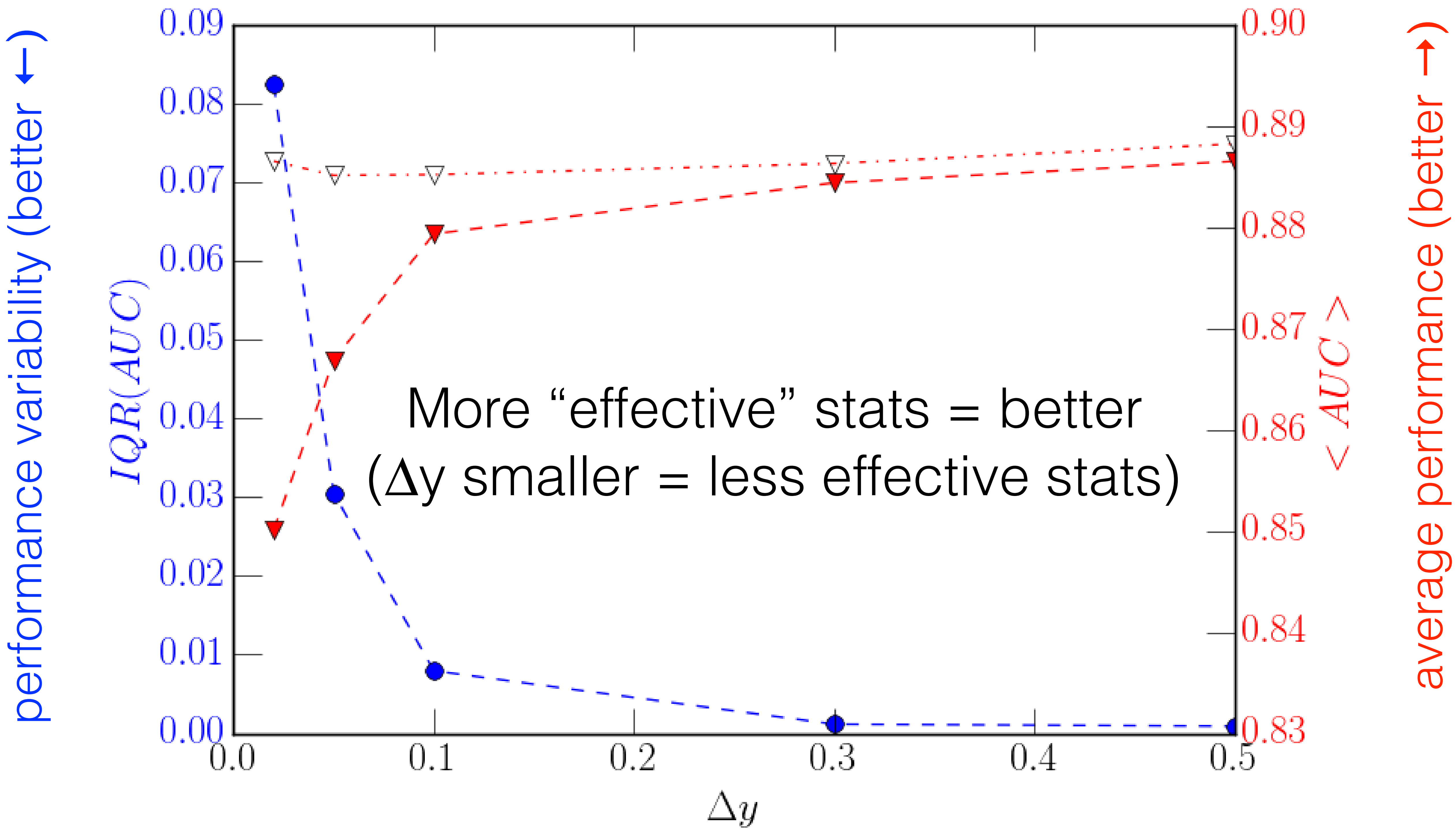
Works in low-dimensions … for q/g

Given the data/MC disagreement from the first slide, this is what you might expect in terms of the performance difference.





A note about training statistics



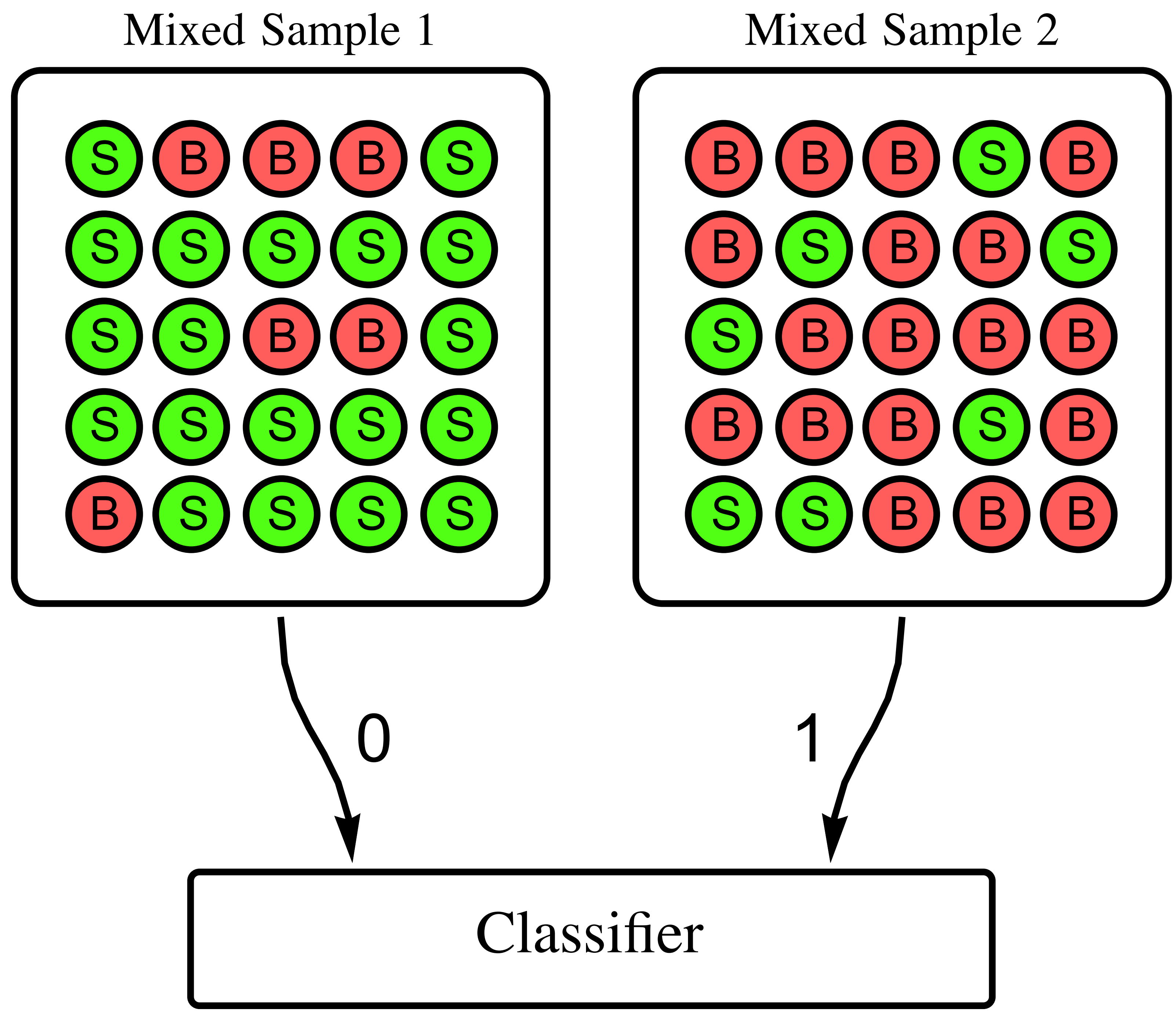
how different are the proportions for the two mixed samples

Method 2: Learning without Proportions

CWoLa

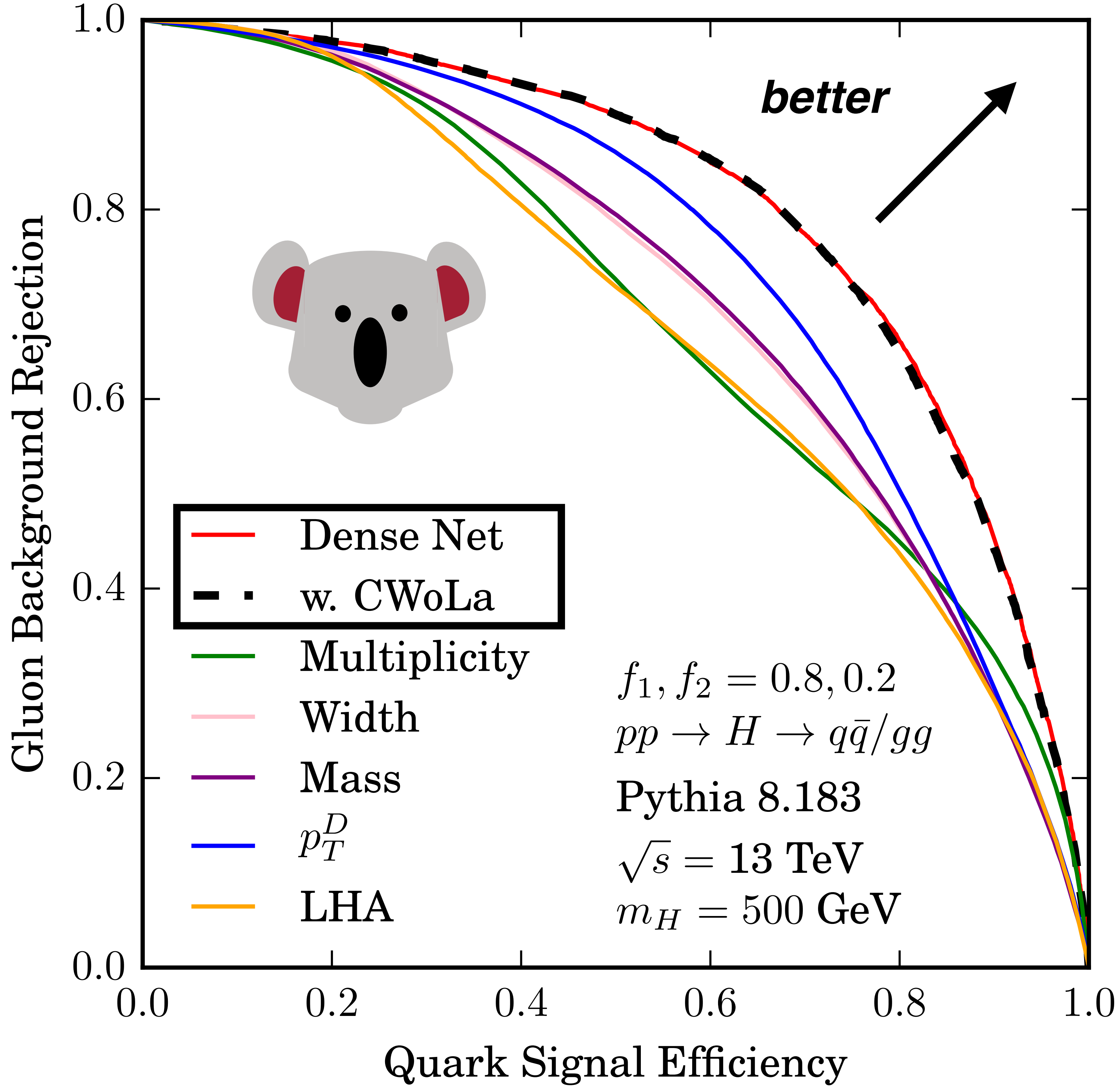
Classification Without Labels

Solution: Train **directly on data** using mixed samples

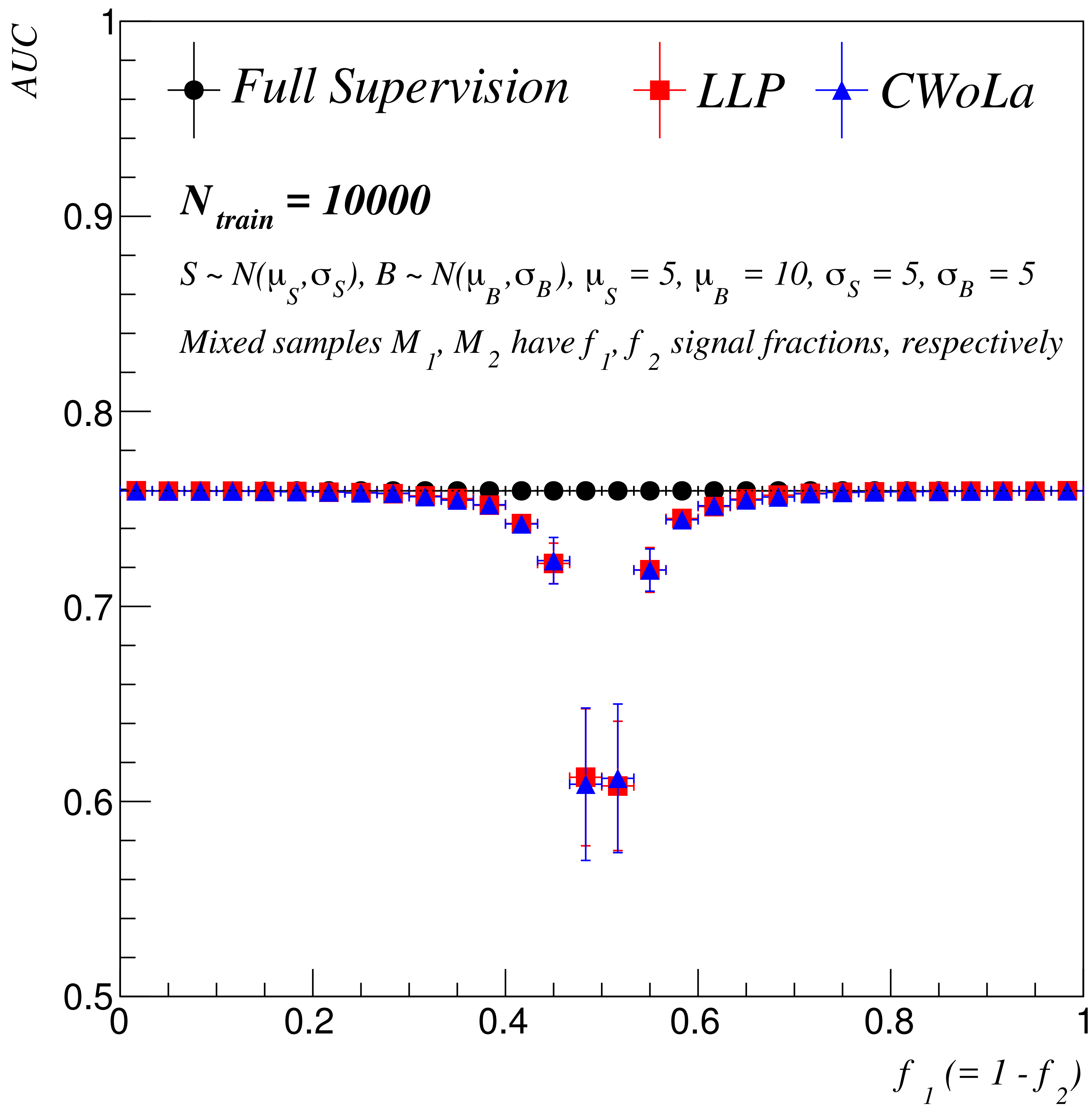


E. Metodiev, **BPN**, J. Thaler, JHEP 10 (2017) 51

Works in low-dimensions



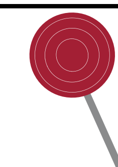
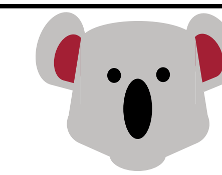
A note about training statistics



As with LLP, need sufficient effective statistics

Can't learn when the two proportions are the same.

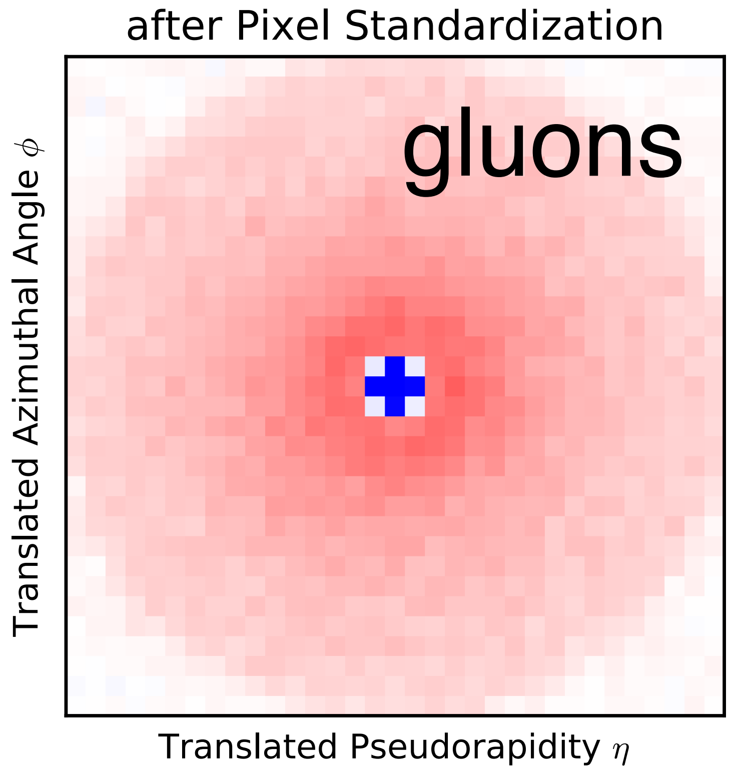
Methods Overview

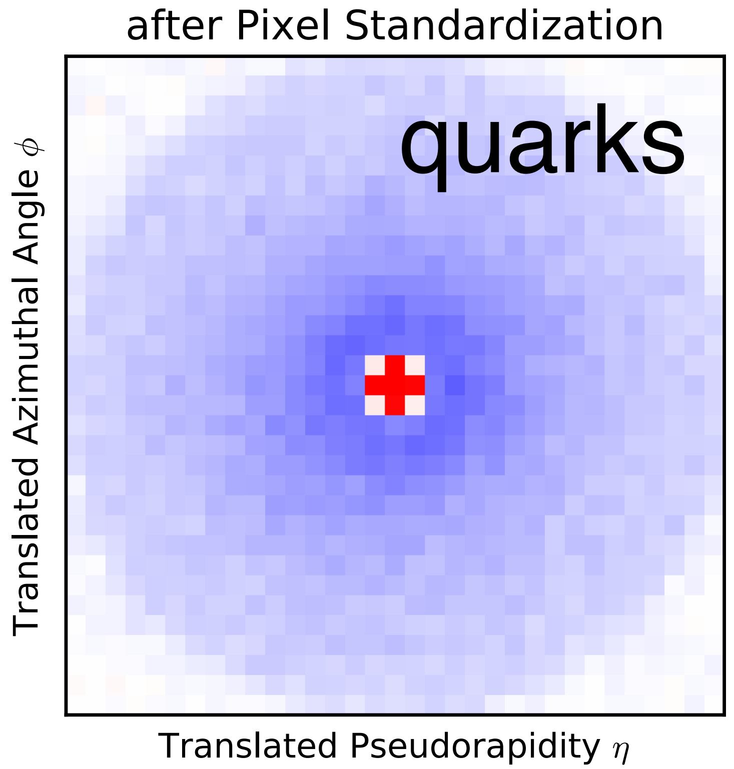
Property	 LLP	 CWoLa
Compatible with any trainable model	✓	✓
No training modifications needed	✗	✓
Training does not need fractions	✗	✓
Smooth limit to full supervision	✗	✓
Works for > 2 mixed samples	✓	?

Next step: what about high dim.?

There are many O(1)-dimensional ML problems for jets, but since the full radiation pattern is higher dimensional, need to go to bigger!

We'll use jet images as a testing ground, still focusing on quarks versus gluons.





P. Komiske, E. Metodiev, M. Schwartz, https://arxiv.org/pdf/1612.01551.pdf

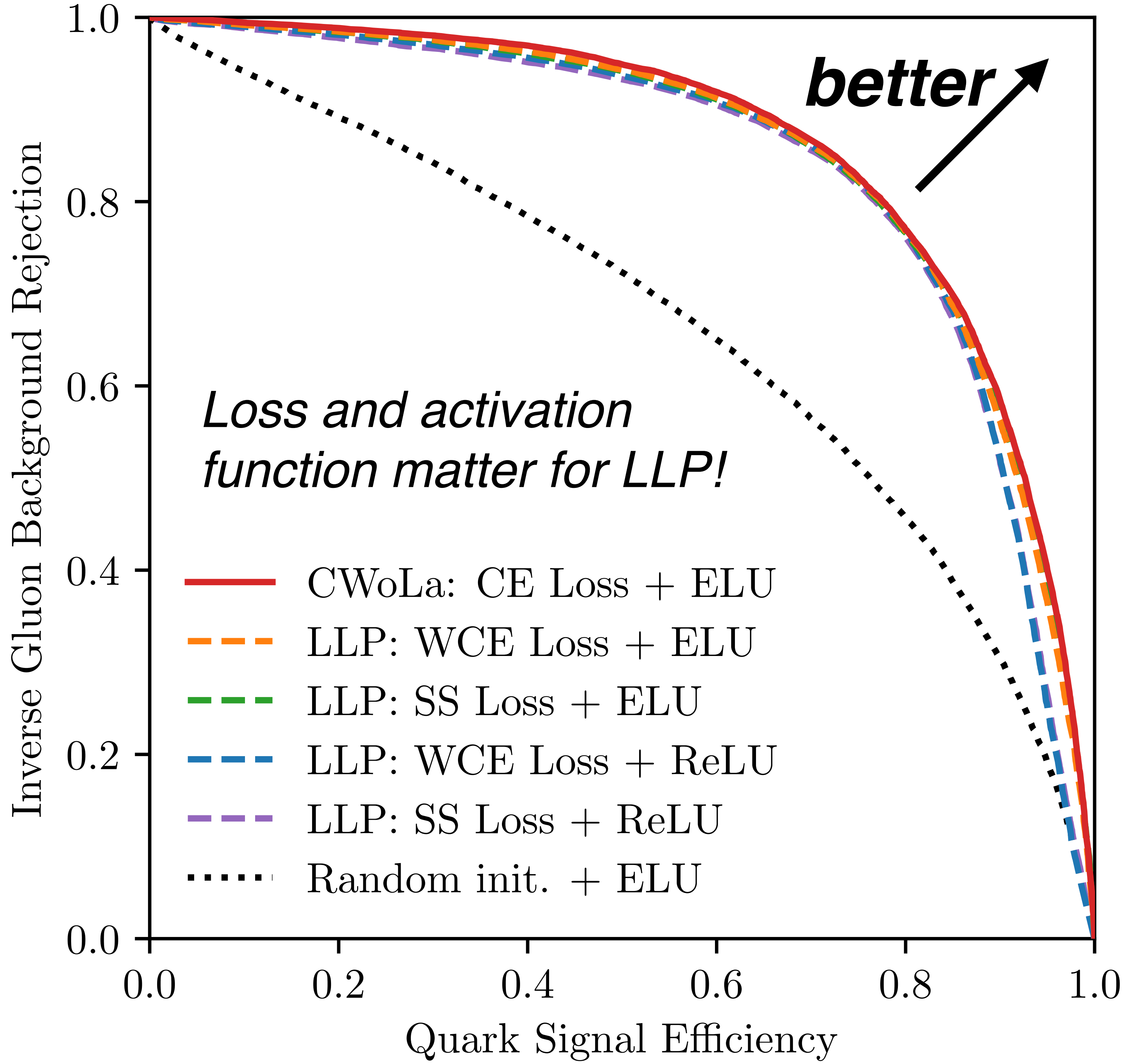
Some Technical Details

The CWoLa approach works out-of-the box - can use well-tested CNN architecture with usual cross-entropy loss.

On the other hand, LLP requires significant work on the technical implementation / optimization.

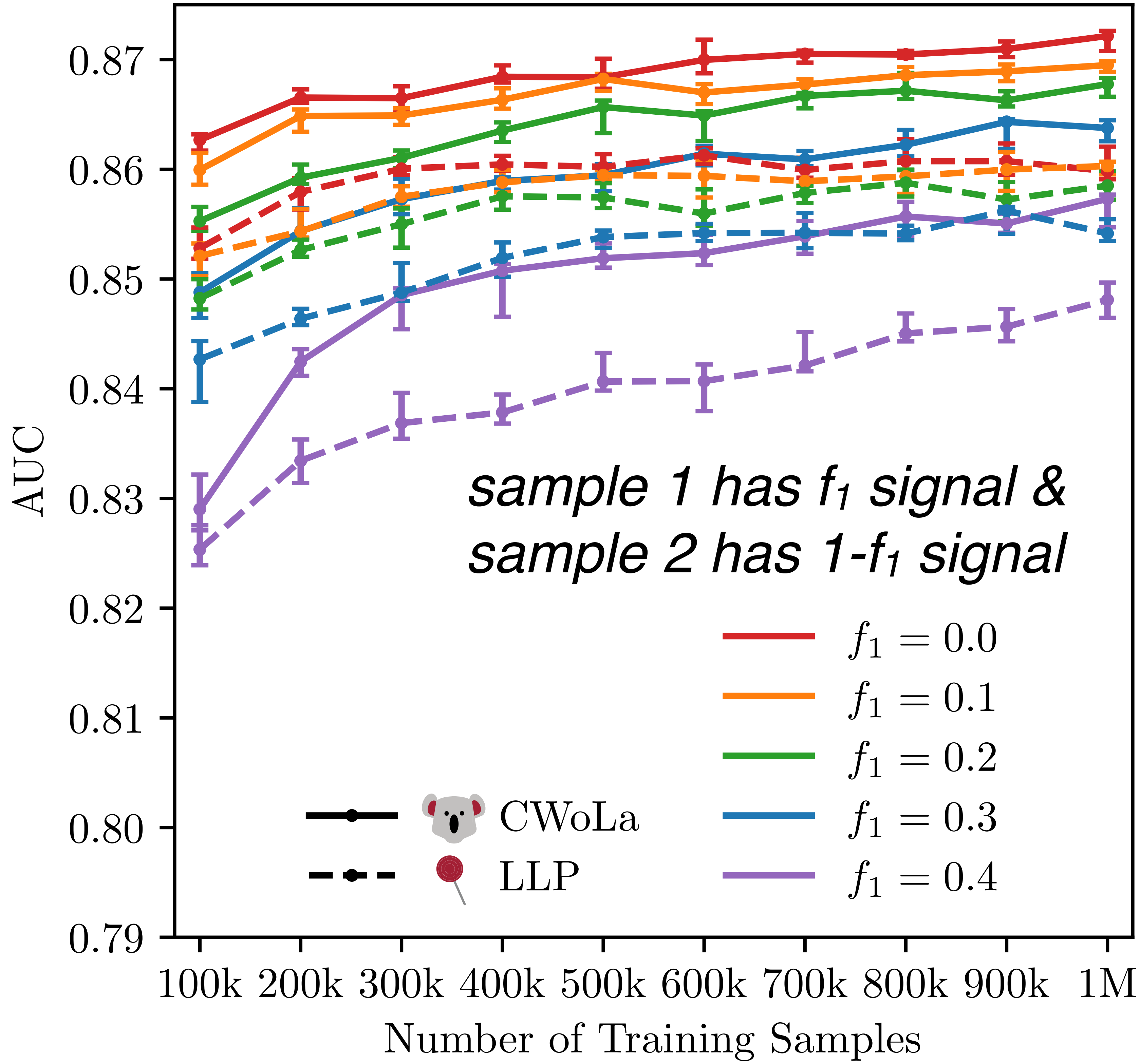
$$\ell_{\text{WMSE}} = \sum_a \left(f_a - \frac{1}{N} \sum_{i=1}^{N} h(\mathbf{x}_i) \right)^2 \qquad \ell_{\text{WCE}} = \sum_a \text{CE}\left(f_a, \frac{1}{N} \sum_{i=1}^{N} h(\mathbf{x}_i) \right)$$

Works in many-dimensions!



P. Komiske, E. Metodiev, **BPN**, M. Schwartz, Phys. Rev. D 98, 011502(R), arXiv:1801:10158

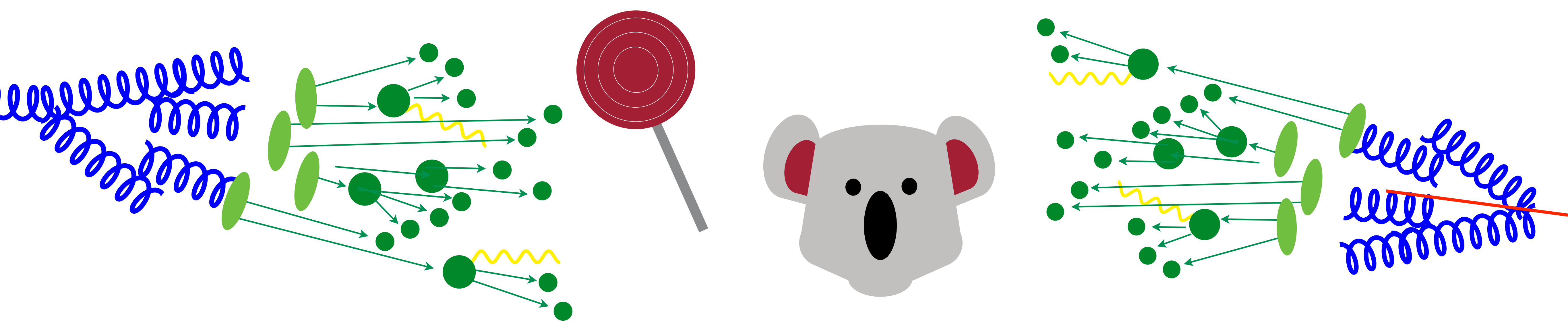
A note about training statistics



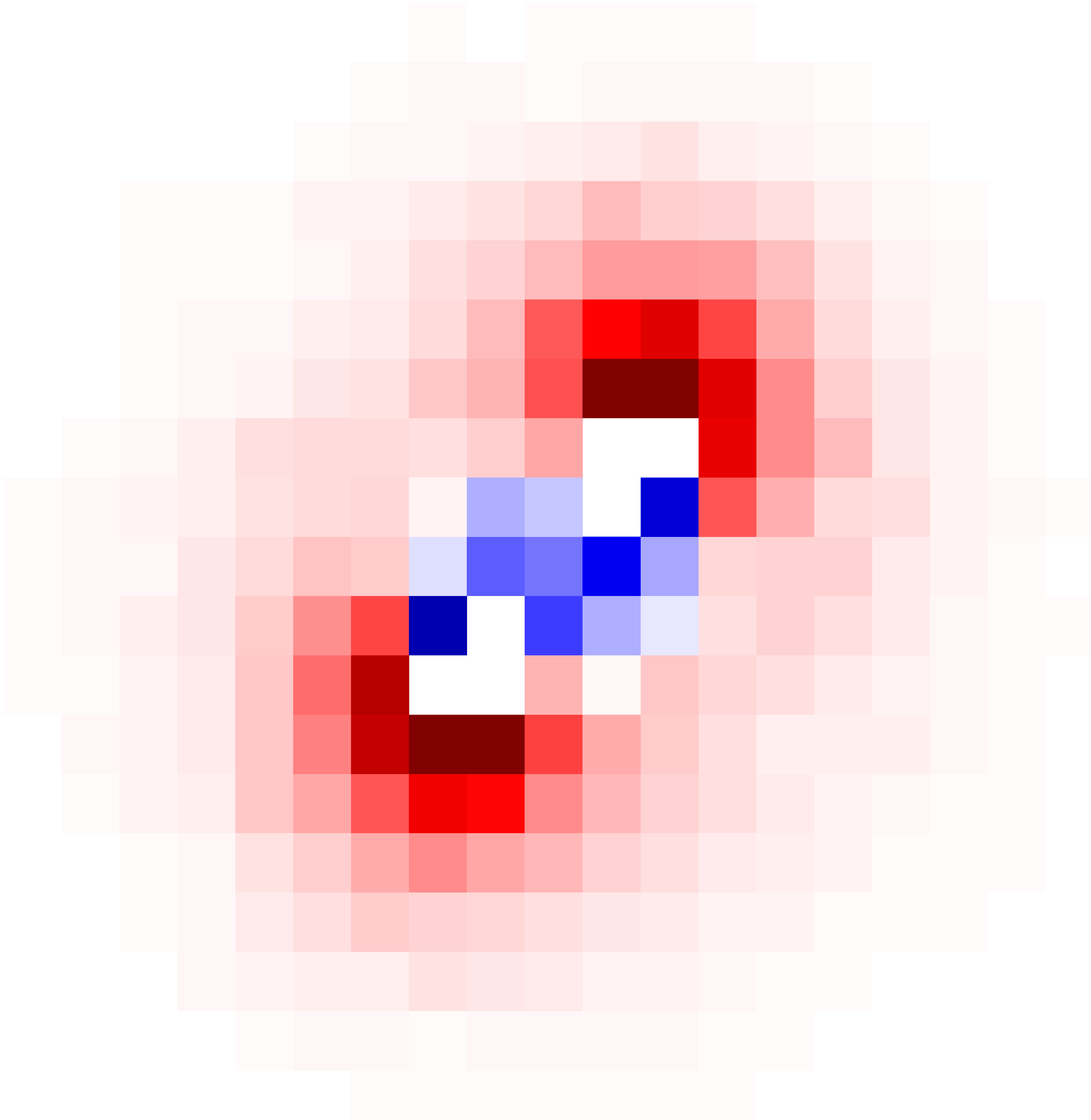
Conclusions and Outlook

Weak supervision is a new & exiting paradigm for training classifiers. We can learn directly from nature!

This is particularly important for jet physics, where there are concerns about mis-modeling. We have shown that the methods work even for high-dimensional data.



To see how else these ideas could be used, see Jack's CWoLa hunting talk and Eric's Jet Topics talk

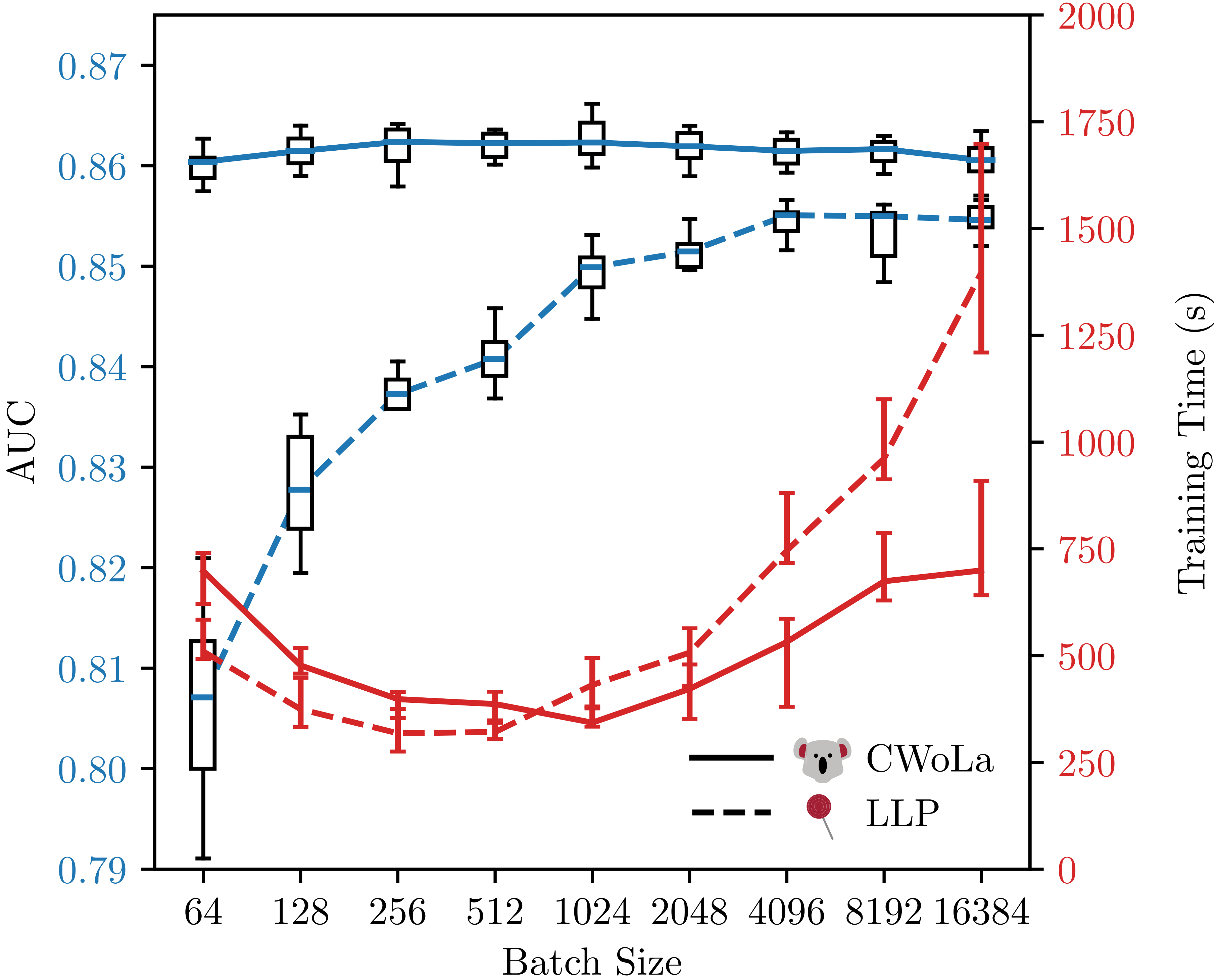


Fin.

Backup: Topology Dependence

Learning	**Sample**	**AUC**
CWoLa	Z+jet vs. dijets	0.8626 ± 0.0020
	Artificial $Z + q/g$	0.8621 ± 0.0019
LLP	Z+jet vs. dijets	0.8544 ± 0.0019
	Artificial $Z + q/g$	0.8549 ± 0.0018

Timing


AUC
0.87
0.86
0.85
0.84
0.83
0.82
0.81
0.80
0.79
2000
1750
1500
1250
1000
750
500
250
0
Training Time (s)
64
128
256
512
1024
2048
4096
8192
16384
Batch Size
CWoLa
LLP