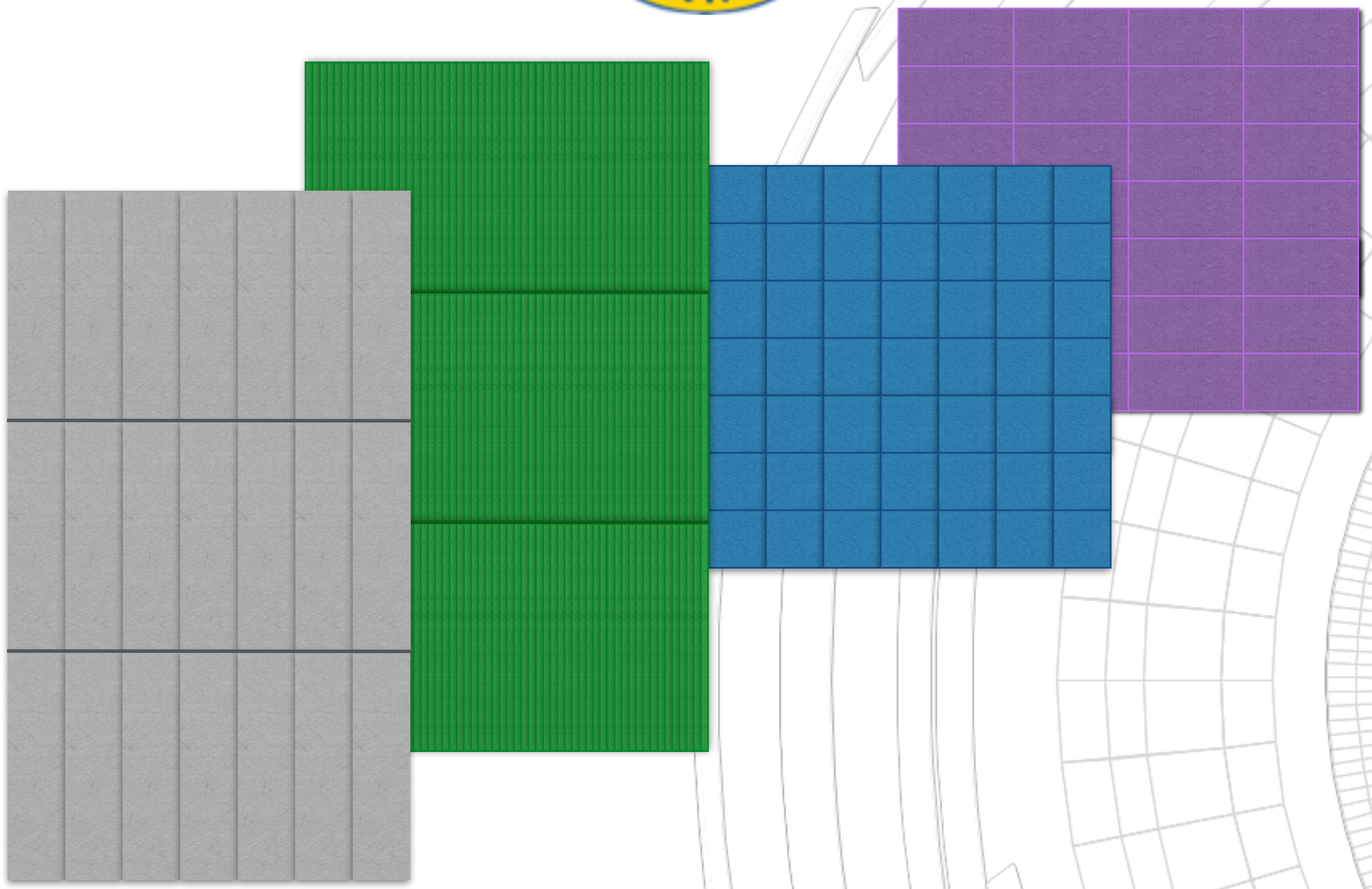
Uncertainties and ML

Aishik Ghosh

Les Houches

20 June 2023



Uncertainties, the bedrock of experimental science

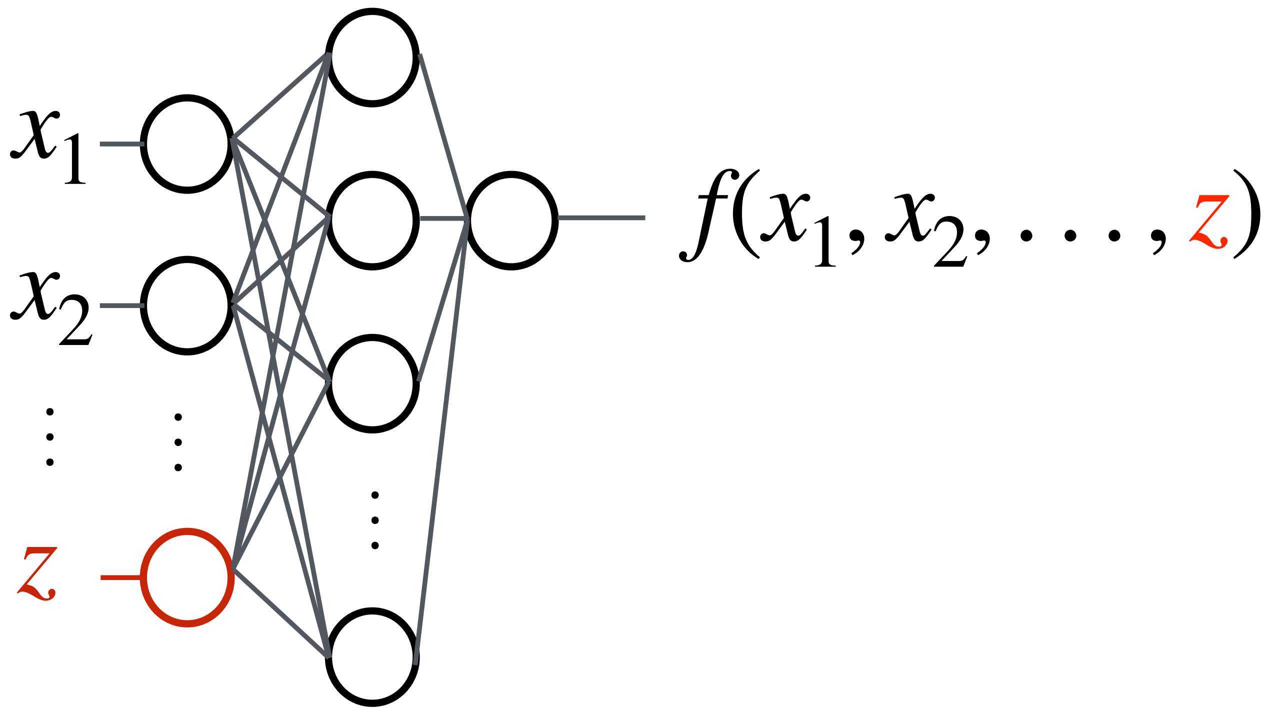
mH = 125.25 ± 0.17 GeV

{statistical, detector systematic, theory systematic, epistemic, ….}

How sure am I ? How can I reduce my uncertainty ?

Three pesky uncertainties for inference

Experimental Systematics



Propagate and profile

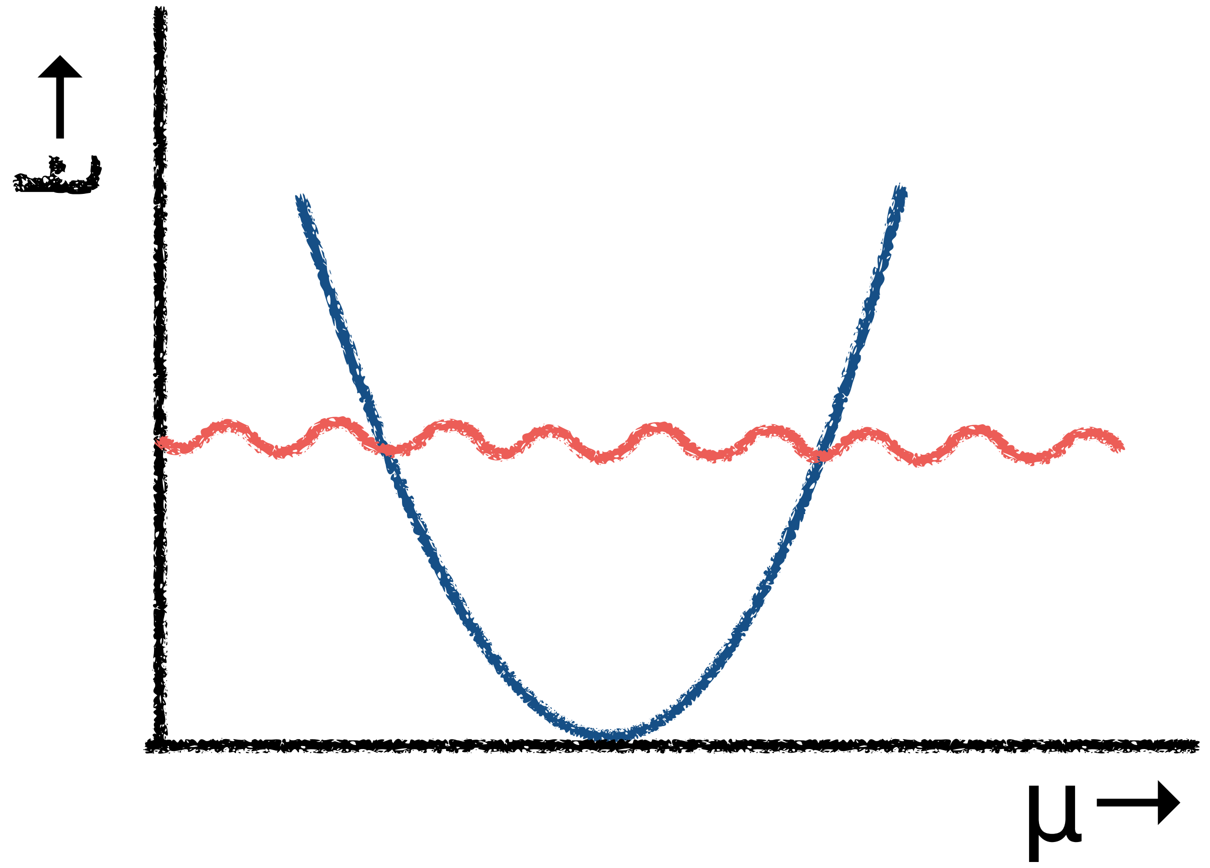
[Ghosh et al.](#)

Theory Systematics

???

[Ghosh and Nachman](#)

Epistemic Uncertainties



Calibrate by histogramming observables
/
Neyman Construction with test statistic

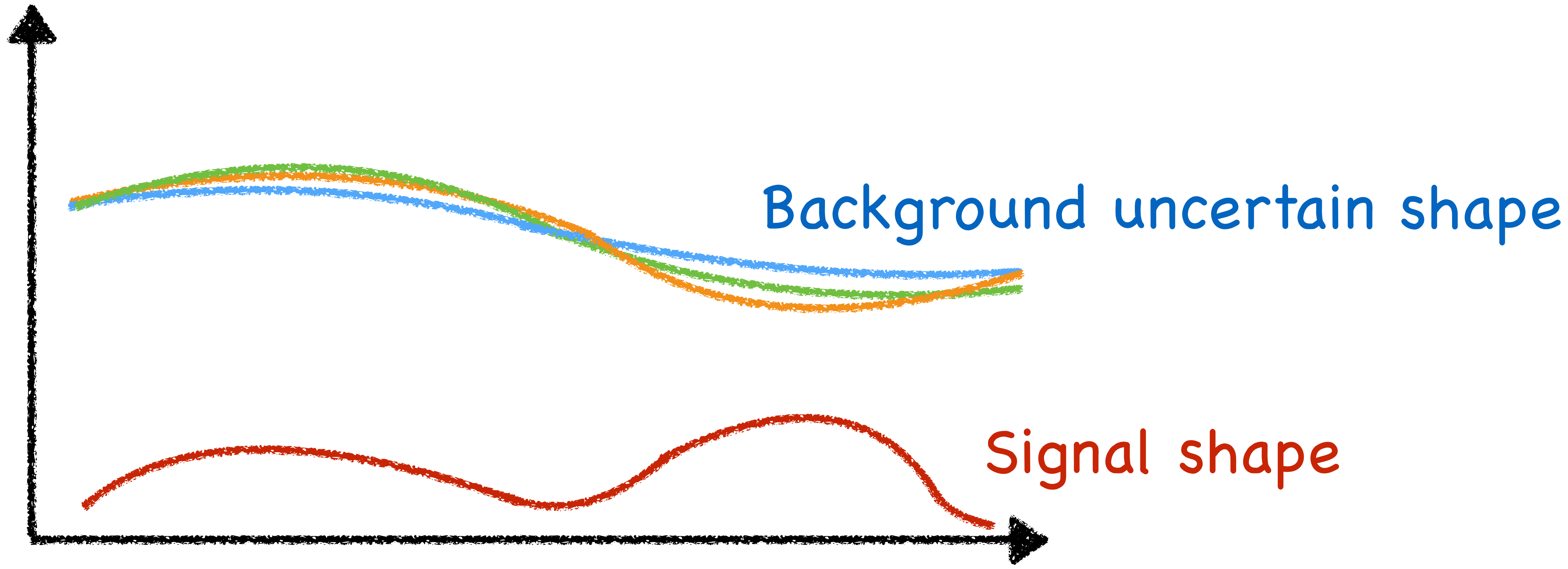
[Cranmer et al.](#)

Observable Sensitive to Nuisance Parameters

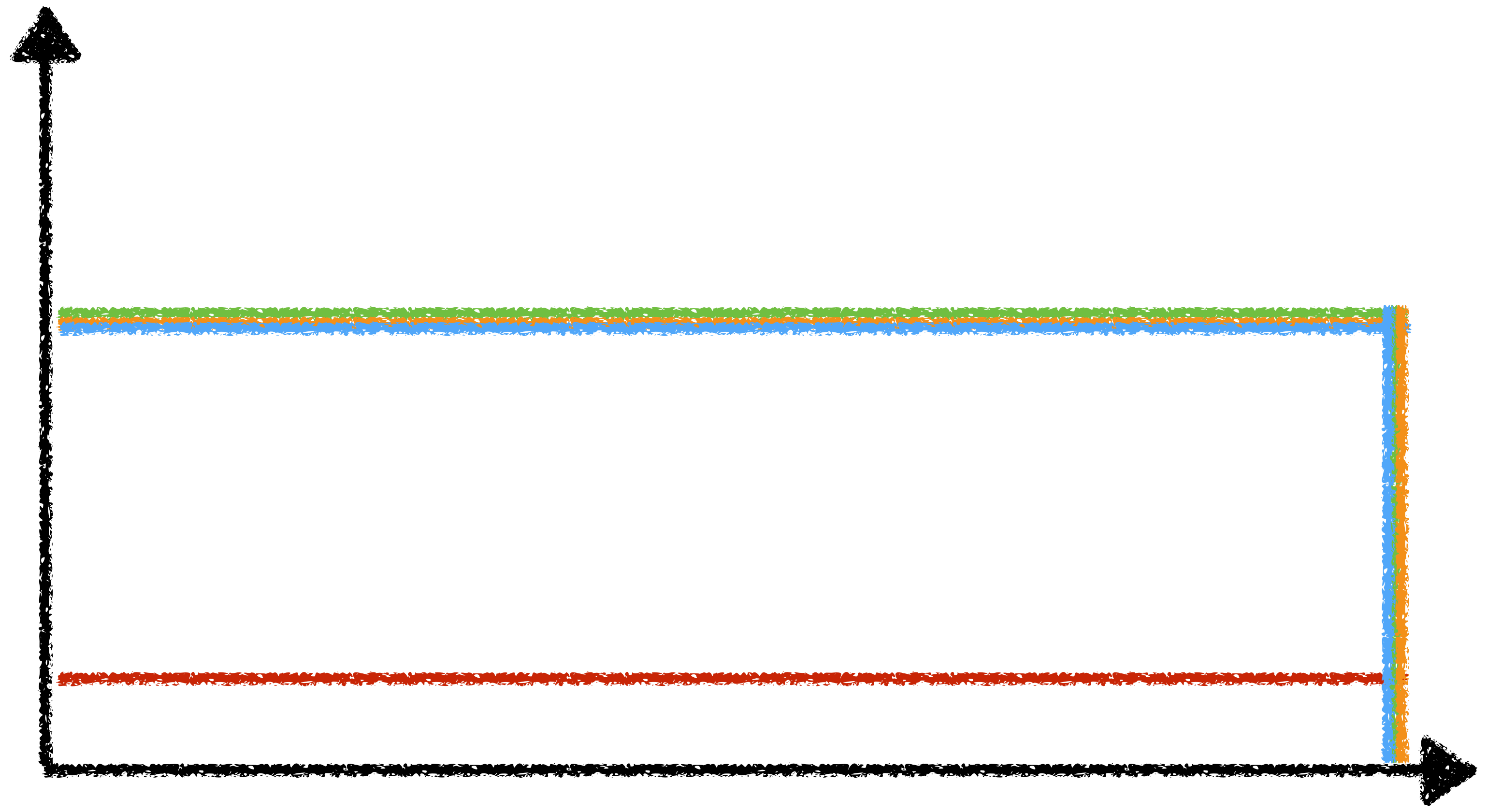
Traditionally, we reduce impact of NP by sacrificing something:

- Don't use observable
- Don't use phase space which is badly modelled by simulation
- Reduce sensitivity some other way

Infinite bin analysis, very sensitive to shape uncertainty



Single bin analysis, insensitive to shape uncertainty



ML equivalent problem: Domain Adaptation

arXiv:1505.07818

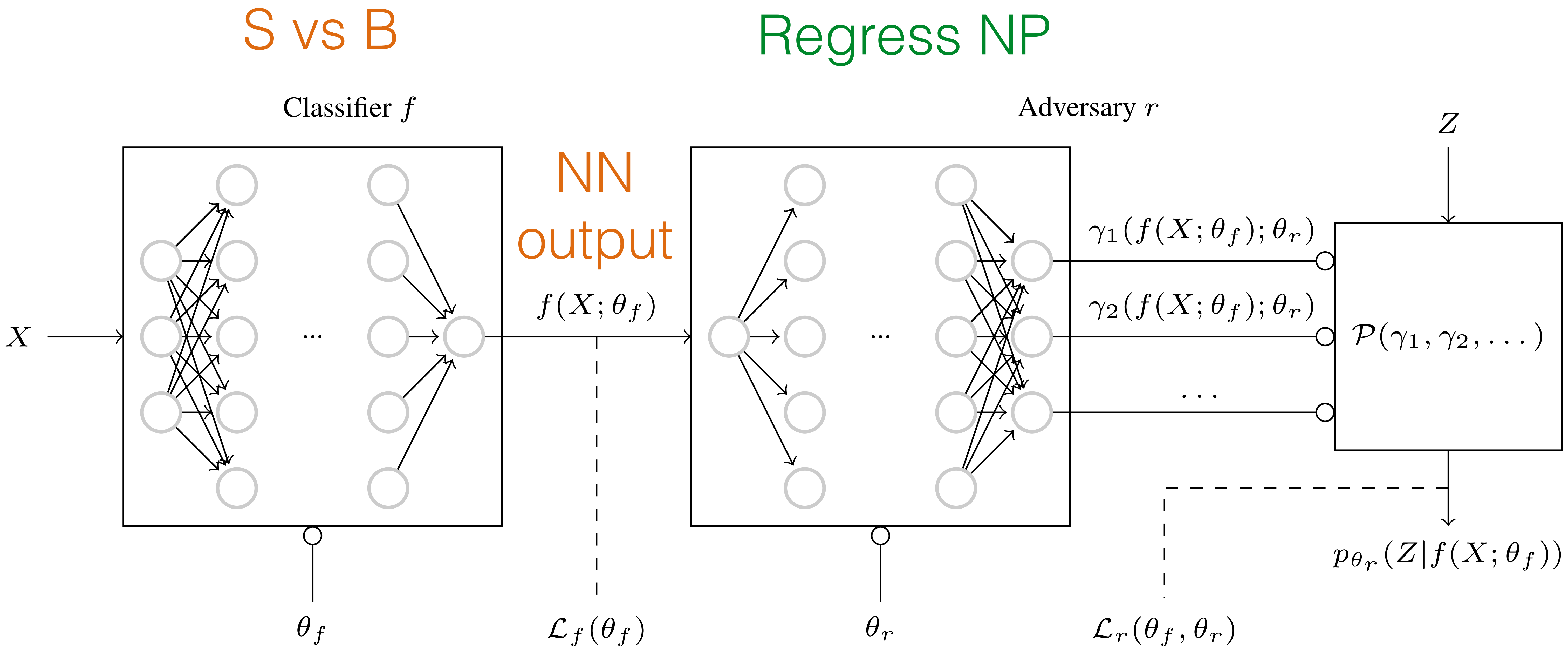
MNIST

MNIST-M

Learning to Pivot, Louppe et al.

Similar ideas: Blance et al., Stevens et al., Wunsch at al., Estrade at al. Kasieczka at al.

Adversarial decorrelation



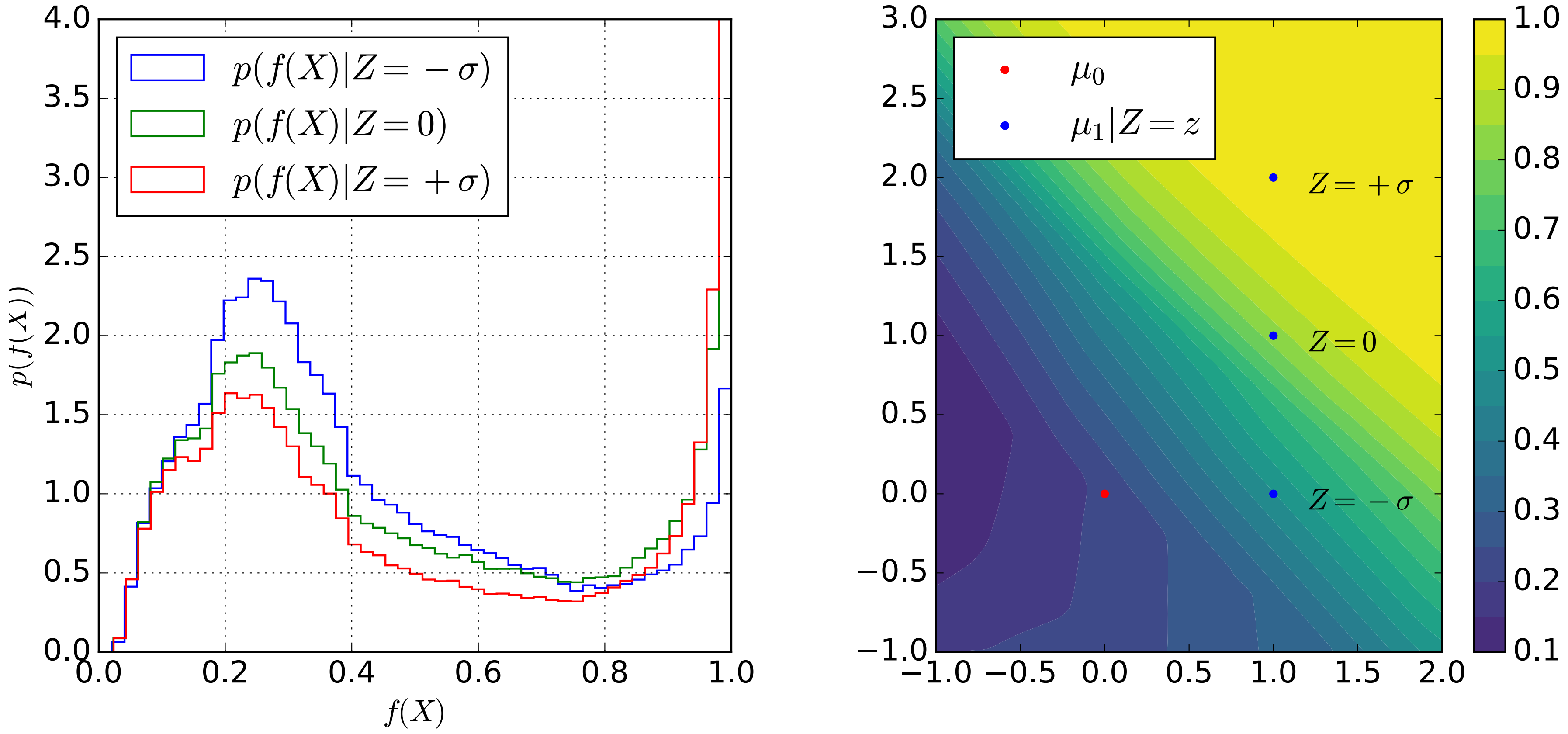
To fool the adversary, classifier output should be decorrelated to Z

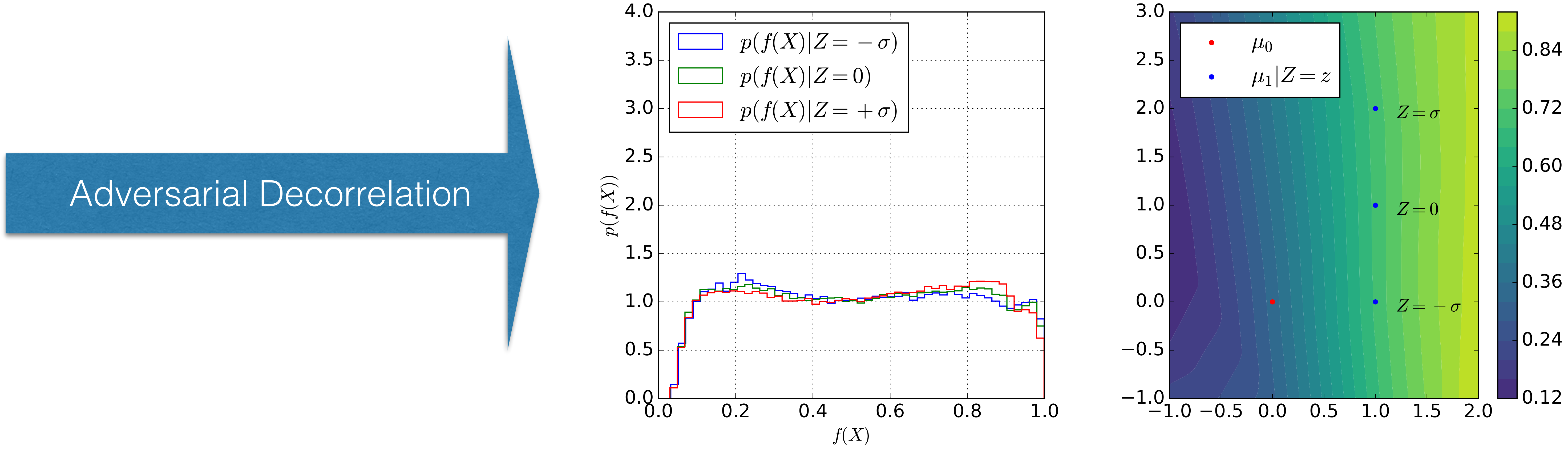
Learning to Pivot, Louppe et al.

$$L_{Classifier} = L_{Classification} - \lambda \cdot L_{Adversary}$$

Similar ideas: Blance et al., Stevens et al., Wunsch at al., Estrade at al. Kasieczka at al.

ML-Decorrelation Methods



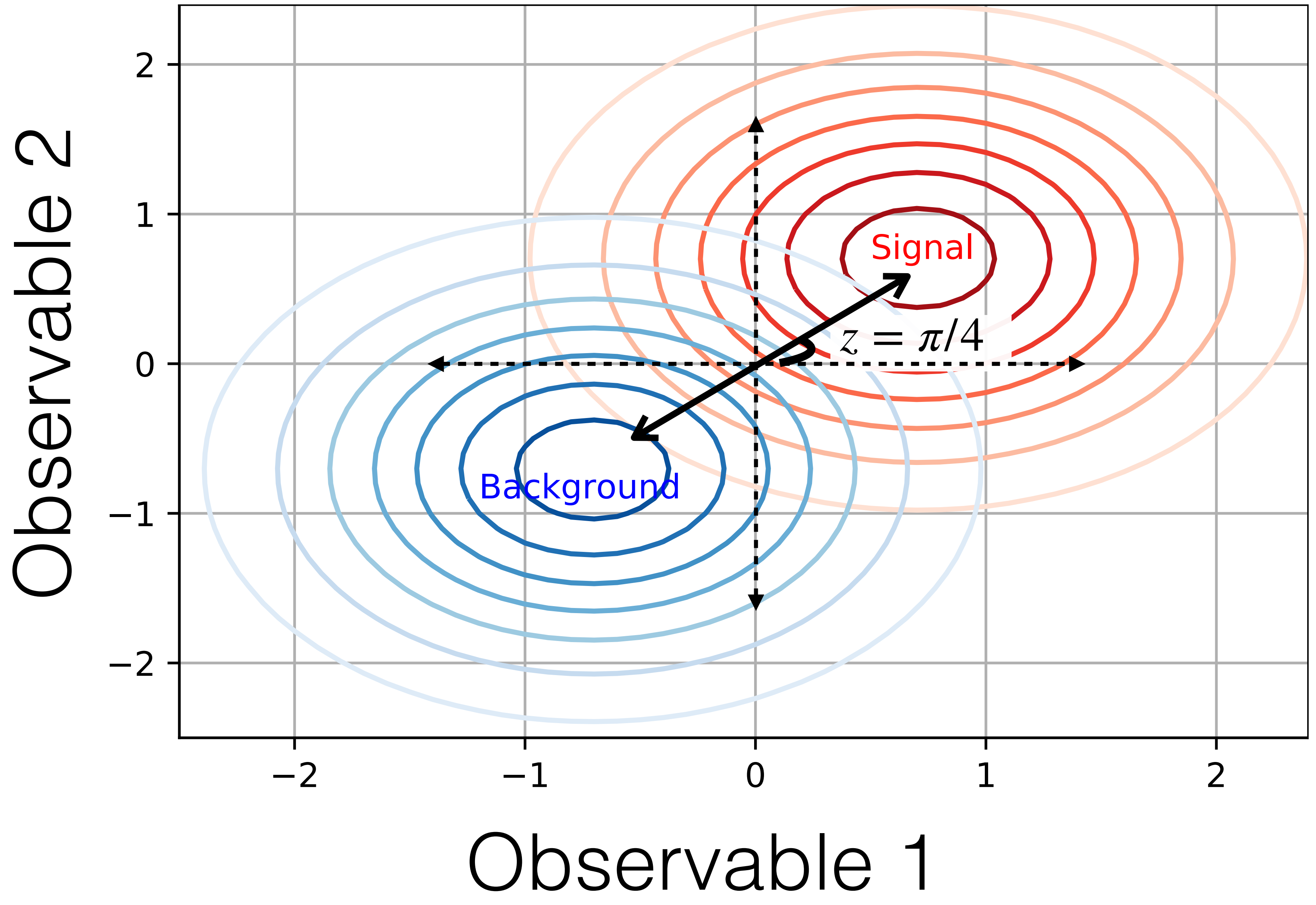


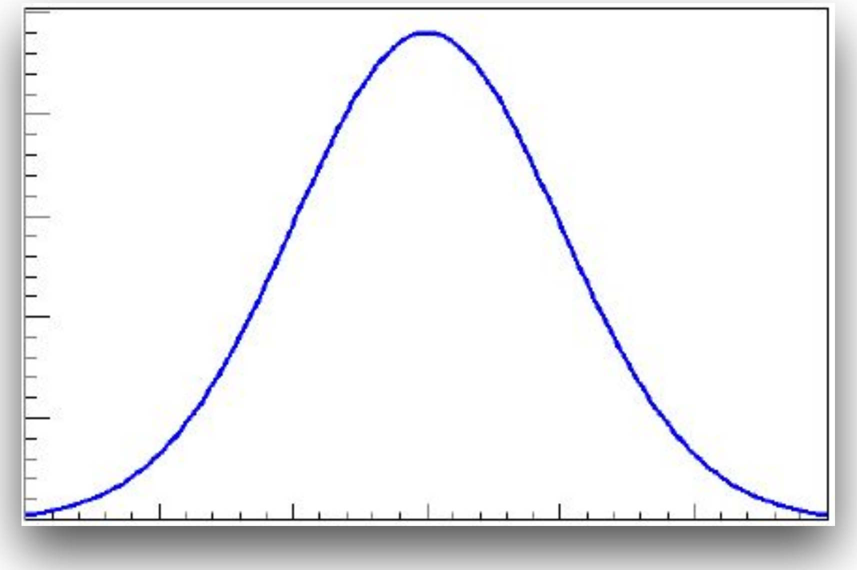
Classifier output for various values of Z

Learning to Pivot, Louppe et al.

Sacrifice separation power for robustness to NPs

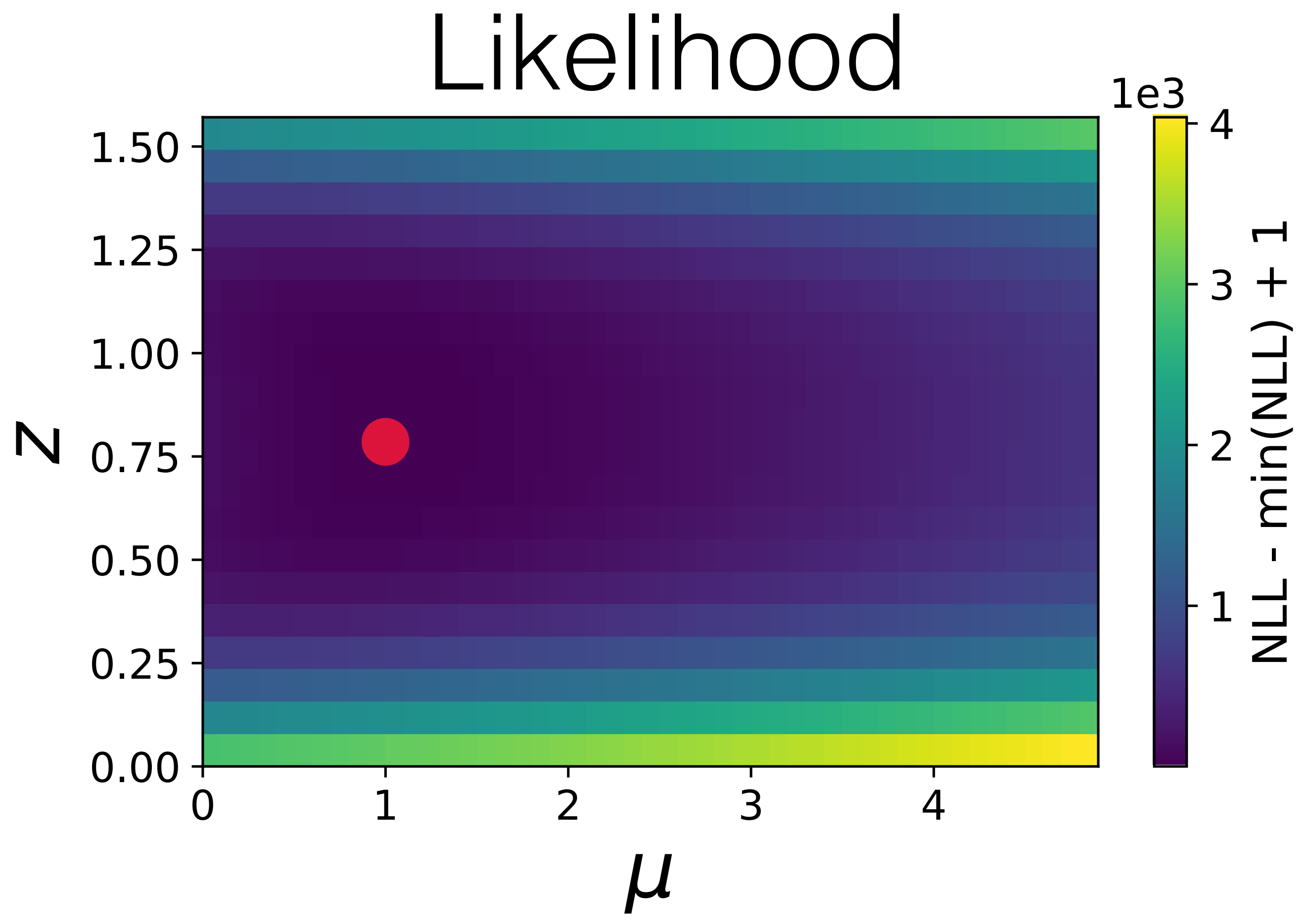
What if we could do better ?

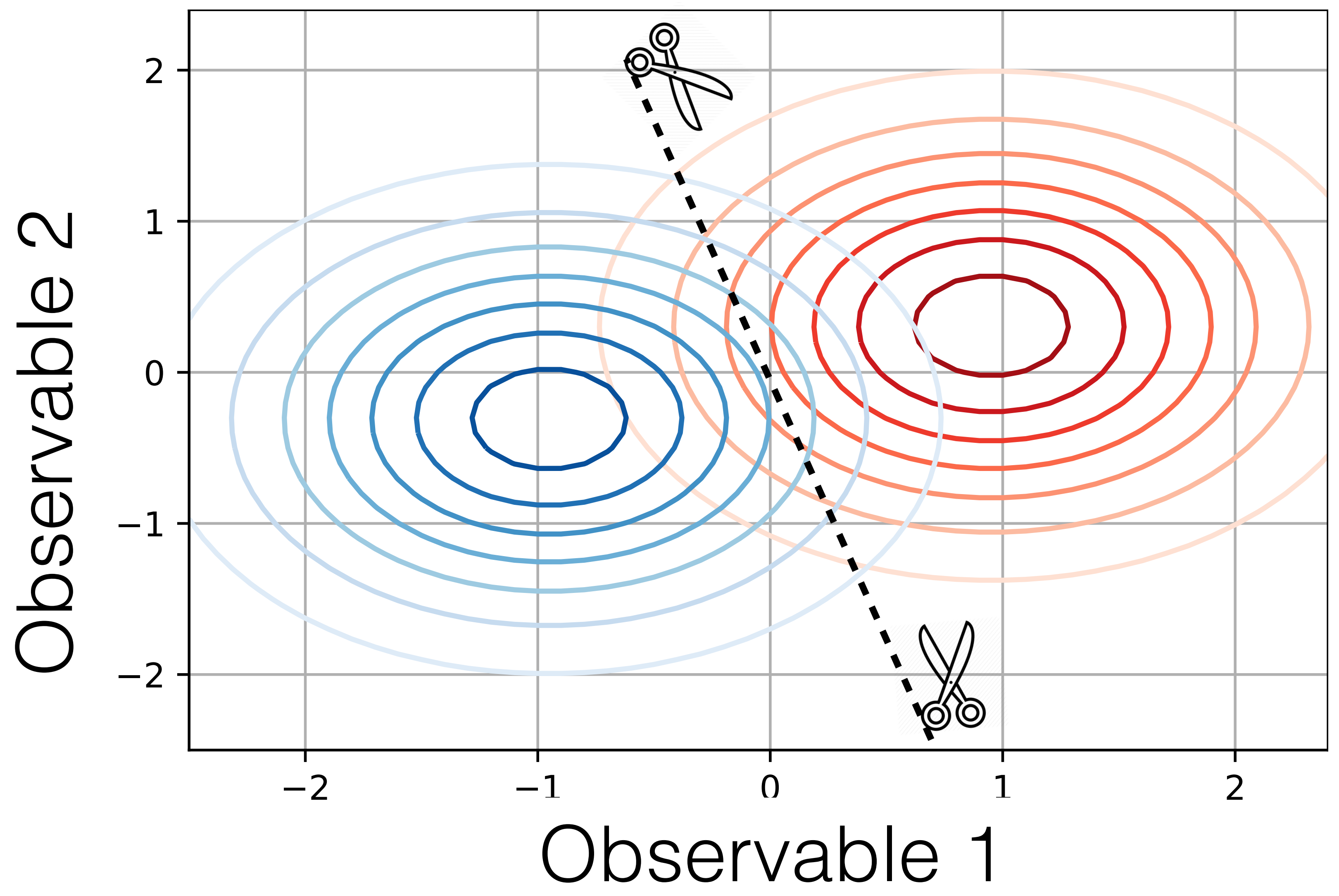


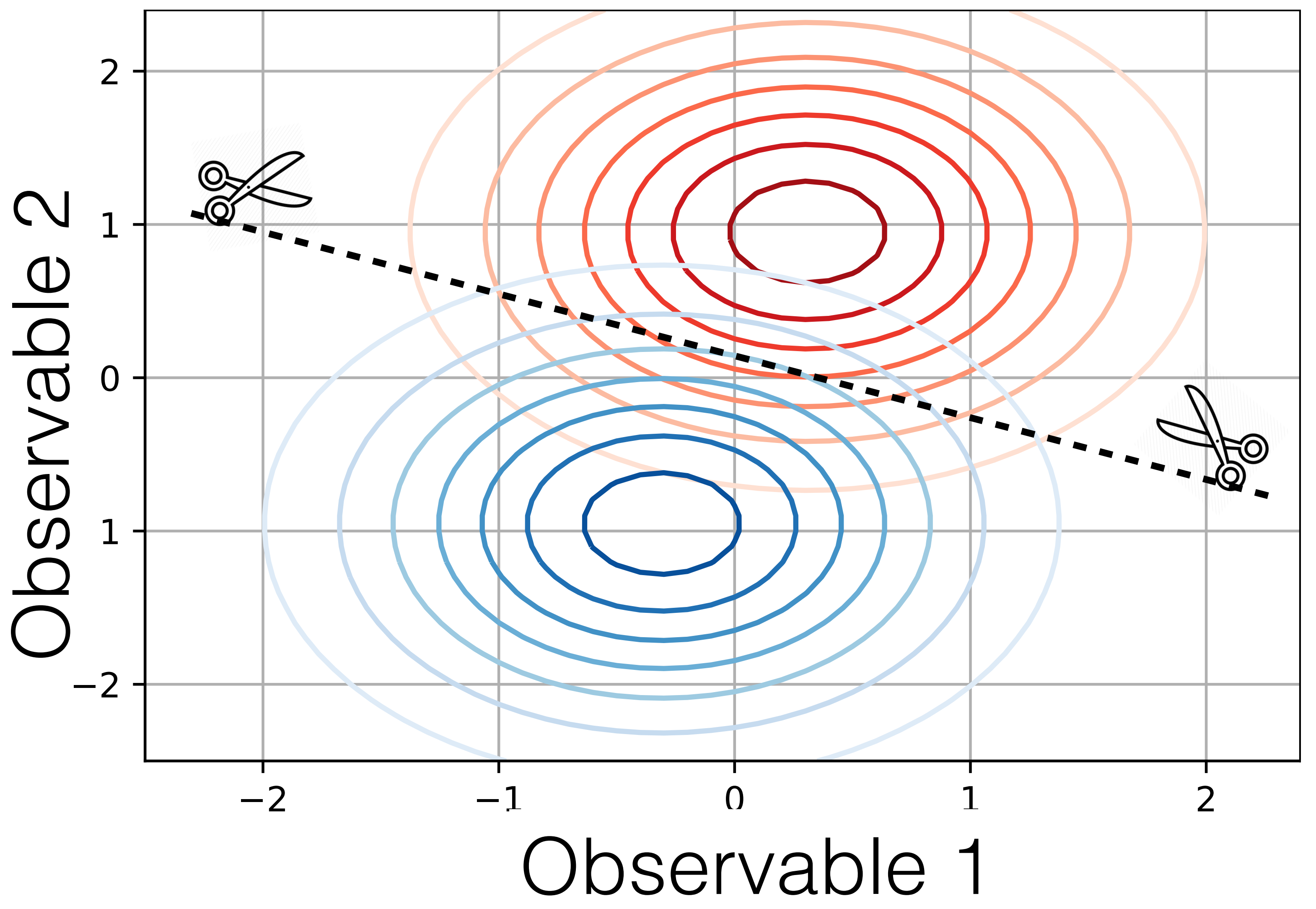


z = Nuisance Parameter

Prior

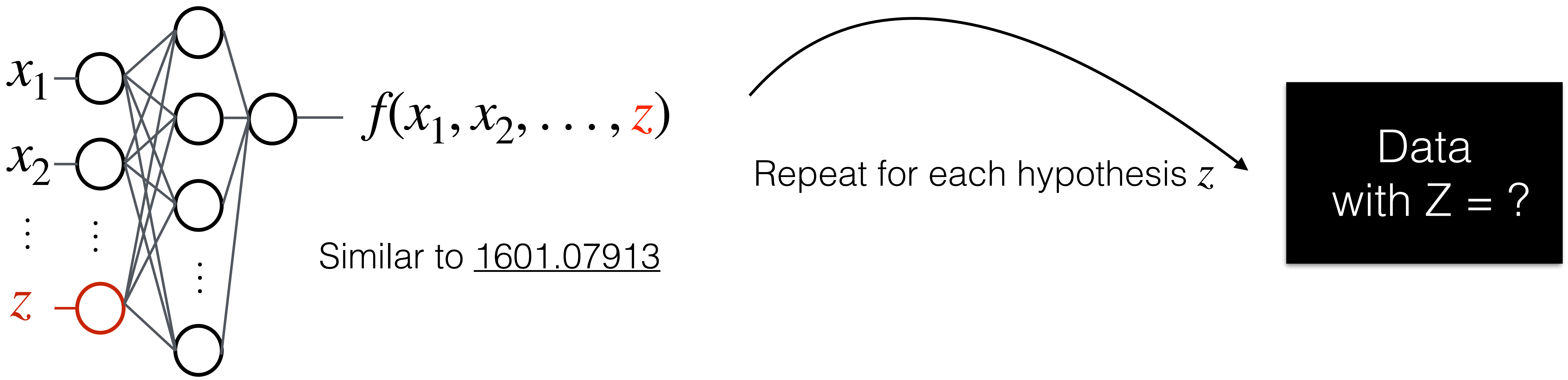






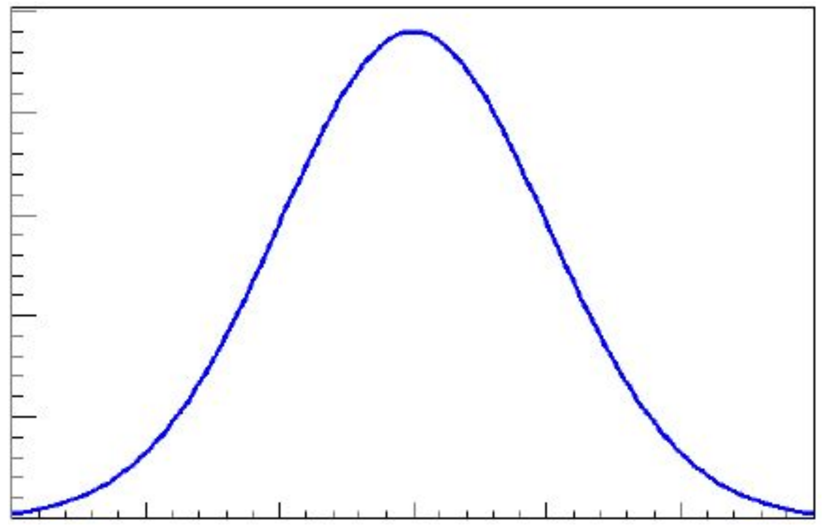
Opposite of decorrelation: Uncertainty-aware learning

- Propagate uncertainties through the classifier in an “uncertainty aware” way



- Intuition: Allow the analysis technique to vary with Z
  You always get the best classifier for each value of Z

- Profile Z + incorporate prior

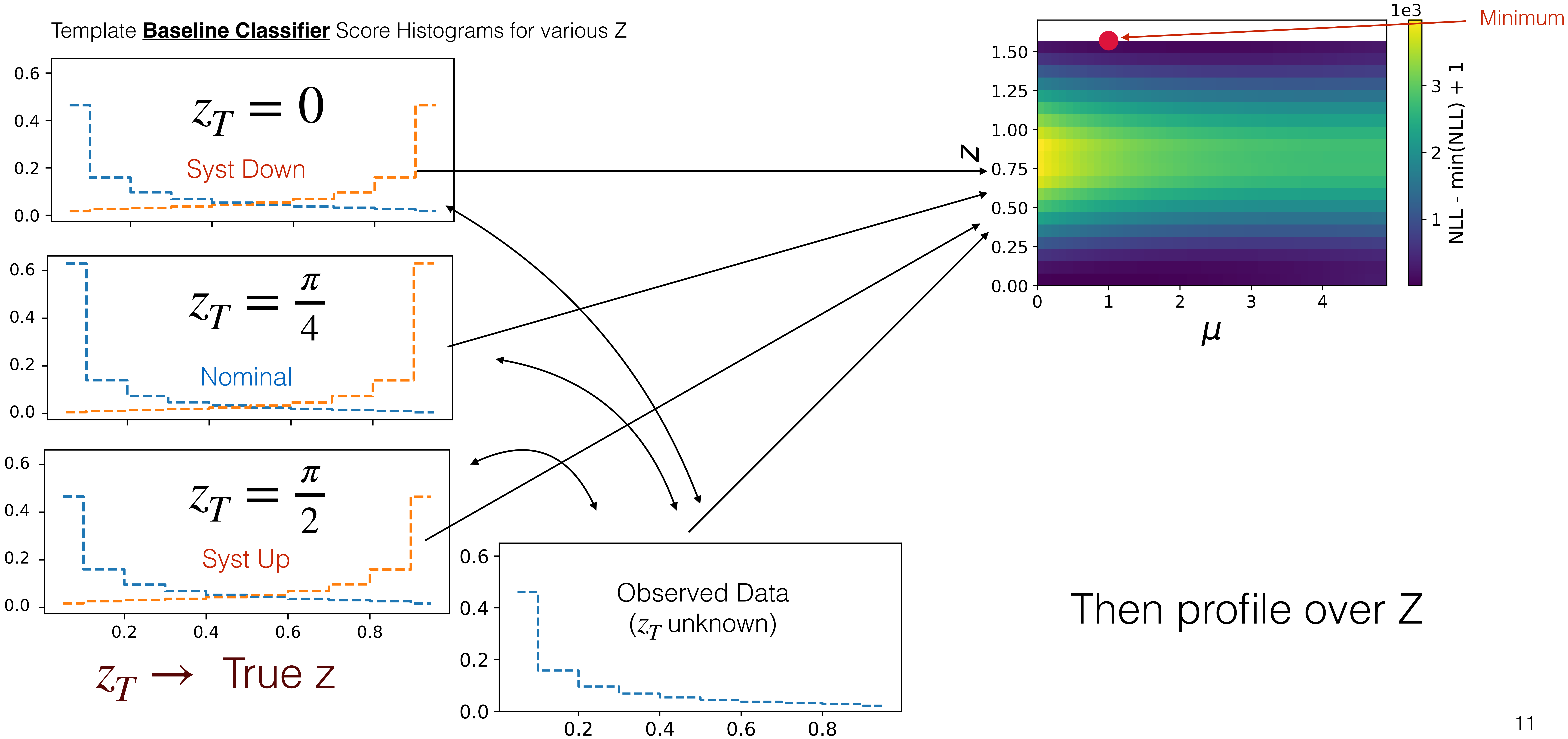


Use a more general function

Instead of building an observable for assumed NPs $O(x_i) := O(x_i, \nu_0)$, build a general one $O(x_i, \nu)$

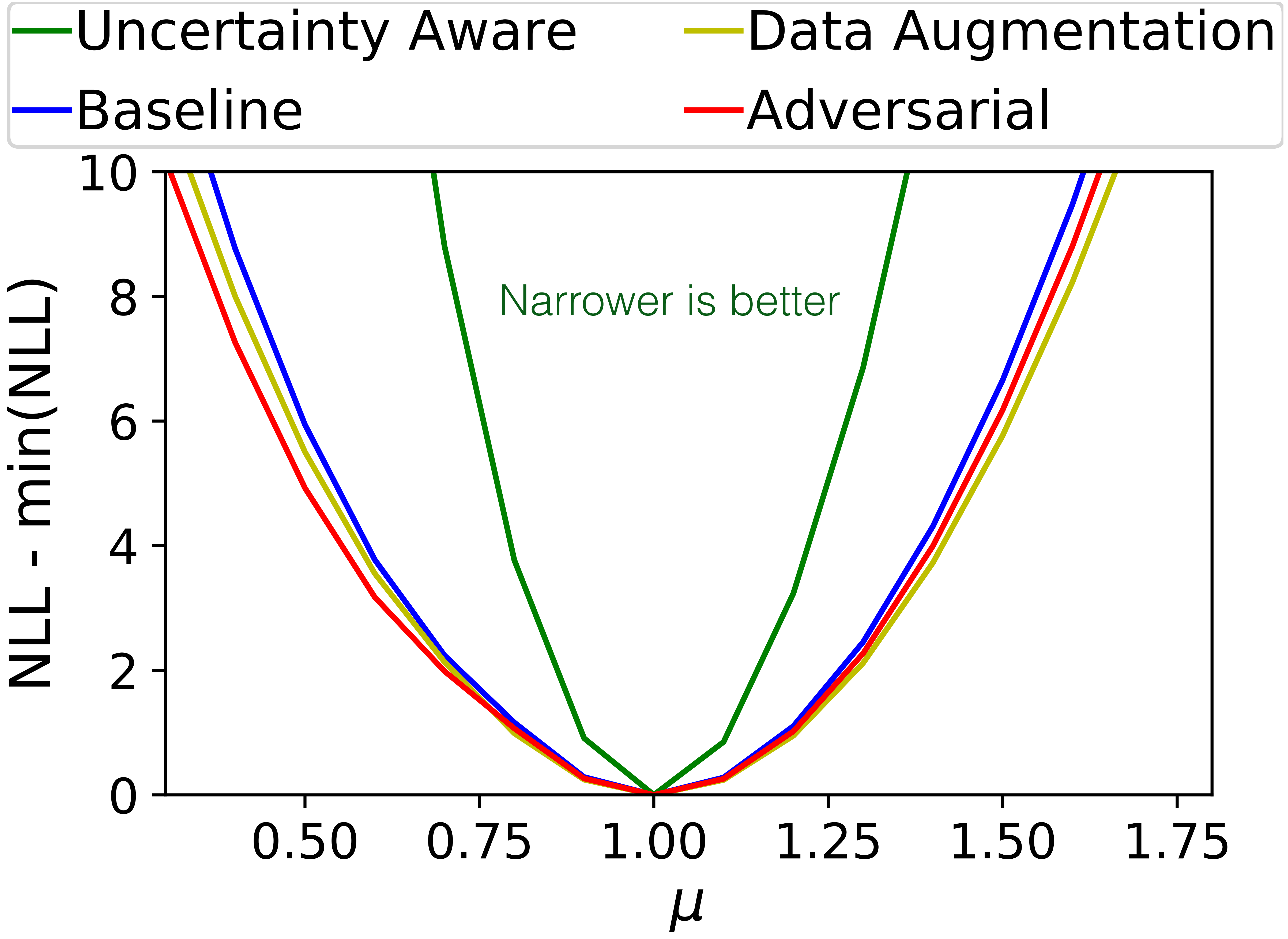
Promote NPs to PIO and scan over all possibilities of μ, ν

Scan the 2D Likelihood space in Z ($:= \nu$) vs μ



Then profile over Z

More sensitivity !



Signal Strength

Narrower ⇒ Smaller [statistical + systematic] uncertainty on measurement

Practical for LHC analysis: Parameterise your main nuisance parameter but no need to train on all 100 NPs

Subsequently applied to astrophysics problems

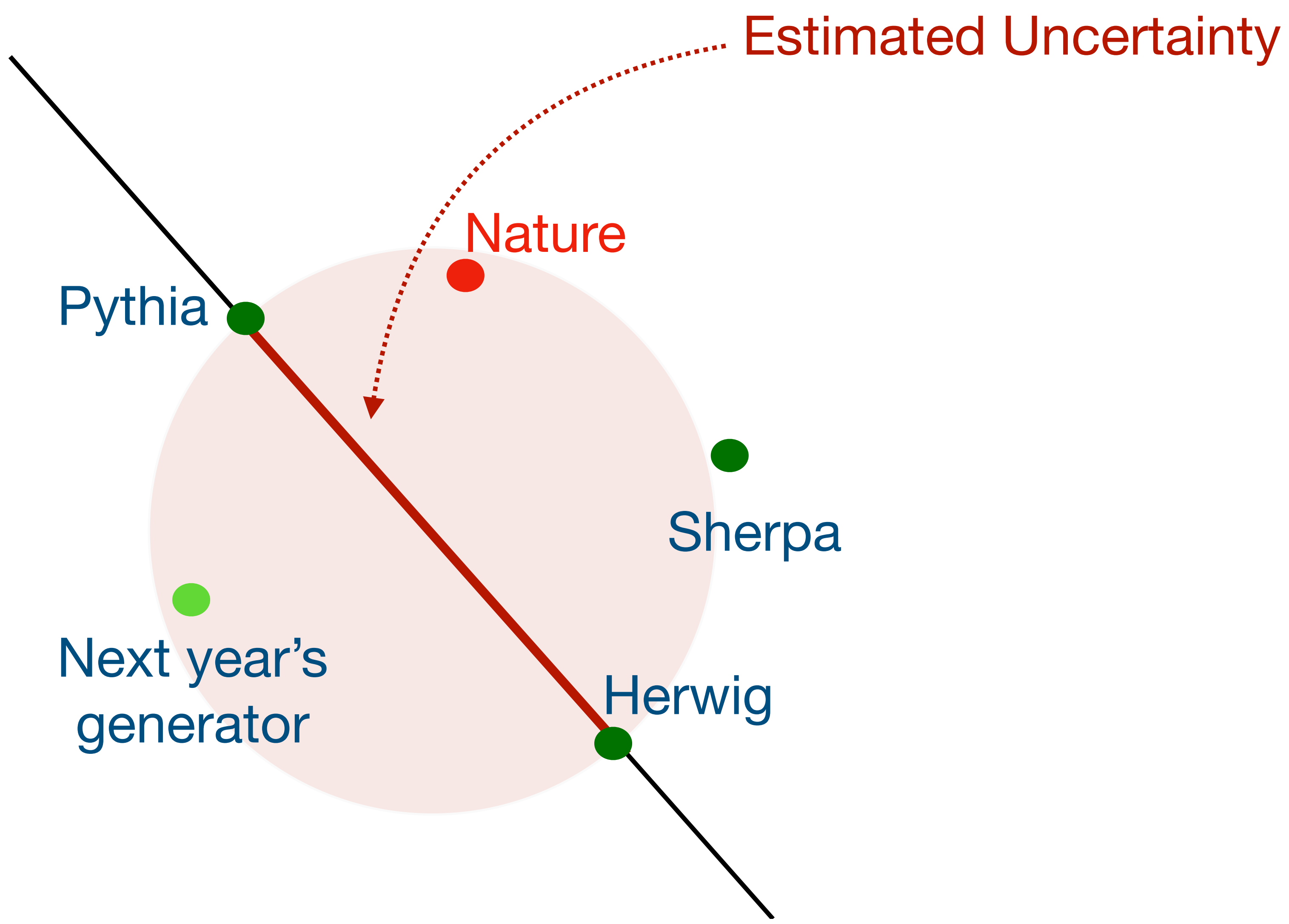
Can we do similar things for theory uncertainties ?

Not at the moment..

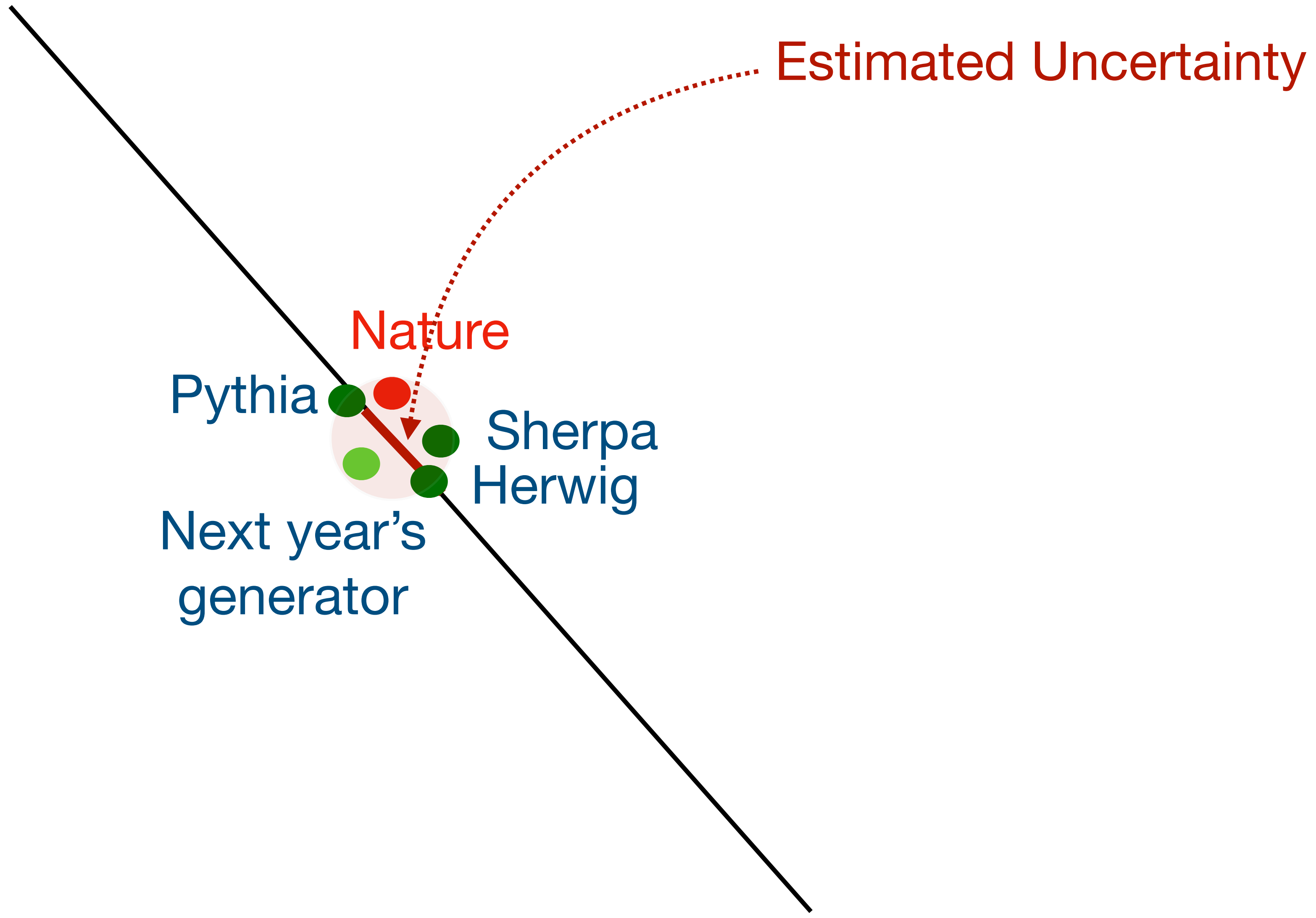
ML-decorrelating theory uncertainties

EPJC:s10052.022.10012.w: Aishik Ghosh, Benjamin Nachman

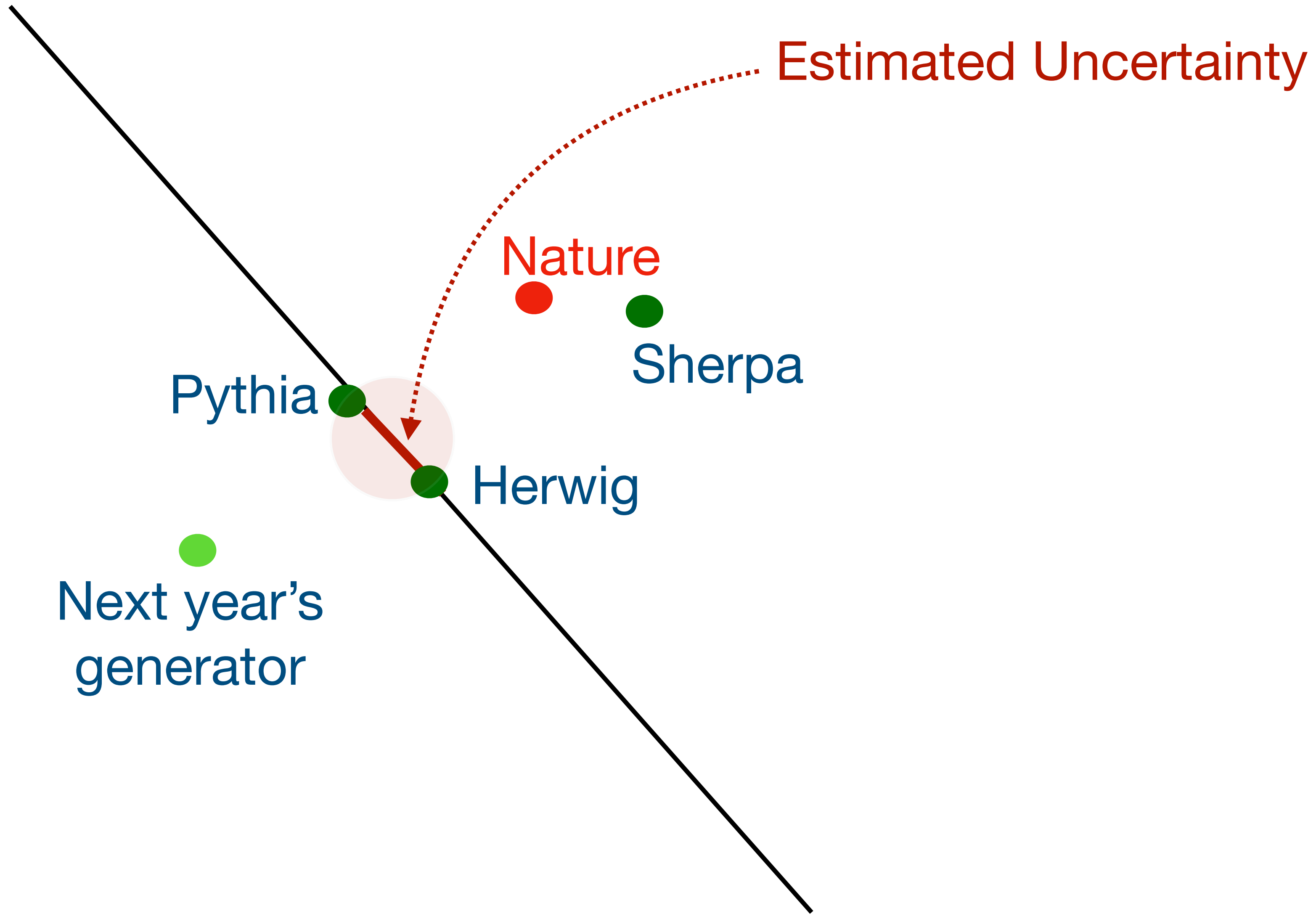
Default



What you want with decorrelation



What you get with decorrelation



Instruction to ML: "Please shrink Pythia vs Herwig difference"

Model will learn to fool you !

ML methods don't often generalise the way you would hope

Case Study 2: Uncertainties from varying unphysical scales at LO

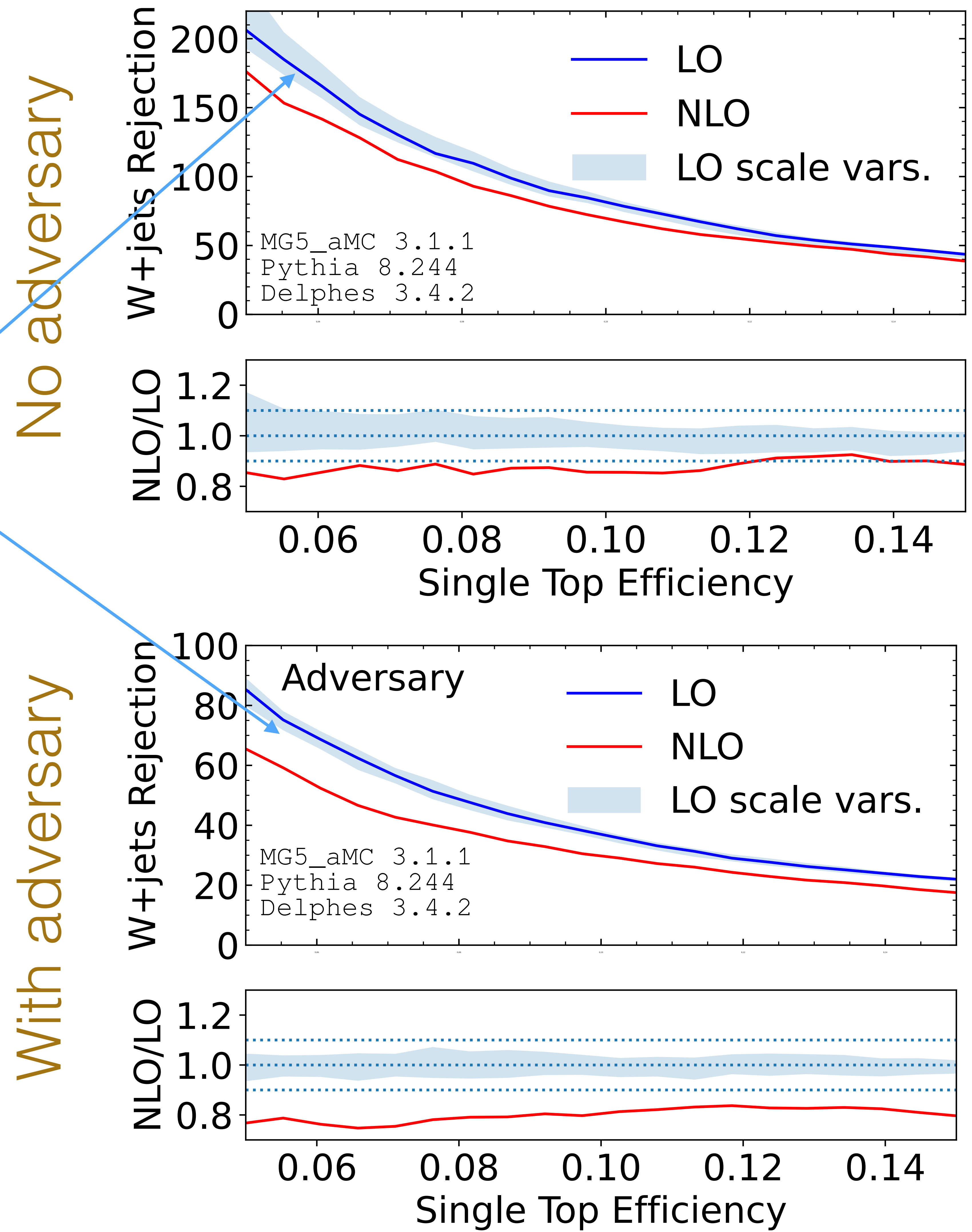
Adversary successfully **sacrifices separation power** in order to reduce difference in performance between scale variations

Cross-check with NLO reveals **uncertainty severely underestimated** by decorrelation approach

In an typical LHC analysis, a cross-check with higher-order usually unavailable

Decorrelation:
Only the error bars shrink, not the actual distance to NLO

ROC curve (higher is better)



As an experimentalist, I want to understand theory uncertainties better

If left to our own devices, here's how we'd go…

Questions

Up: $\mu_{+} = 2\ \mu_0$

Down: $\mu_{-} = \frac{1}{2}\ \mu_0$

- How accurate are these scale uncertainties ?
- Is 1/2 to 2 a good range ?

Study pull distribution

$$t_{scale} = \frac{\sigma_{NLO} - \sigma_{LO}}{\Delta\sigma_{LO\ scale}}$$

Madgraph paper

The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations

J. Alwall[a], R. Frederix[b], S. Frixione[b], V. Hirschi[c], F. Maltoni[d], O. Mattelaer[d], H.-S. Shao[e], T. Stelzer[f], P. Torrielli[g], M. Zaro[hi]

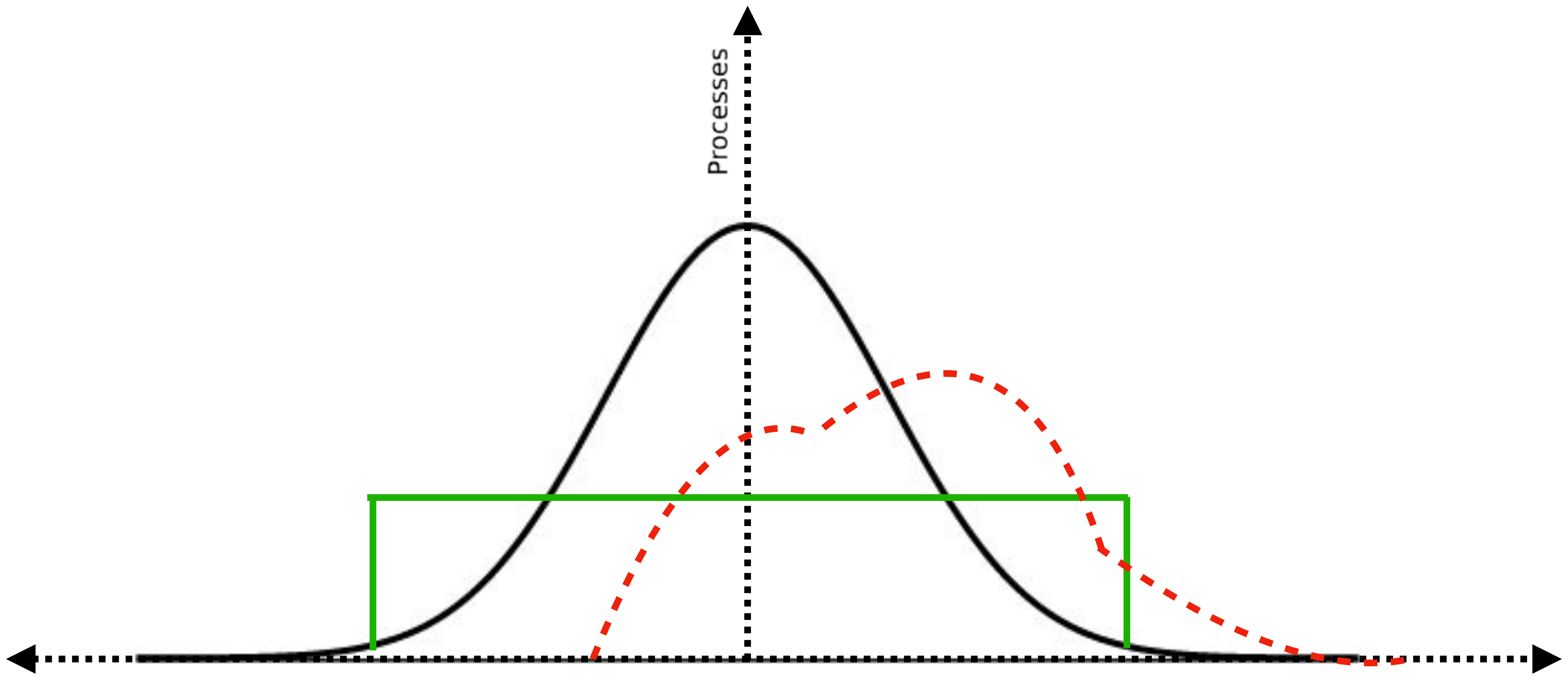
	Process	Syntax	Cross section (pb) LO 13 TeV		NLO 13 TeV	
	Vector boson +jets					
a.1	$pp \to W^{\pm}$	`p p > wpm`	$1.375 \pm 0.002 \cdot 10^5$	$^{+15.4\%}_{-16.6\%}$ $^{+2.0\%}_{-1.6\%}$	$1.773 \pm 0.007 \cdot 10^5$	$^{+5.2\%}_{-9.4\%}$ $^{+1.9\%}_{-1.6\%}$
a.2	$pp \to W^{\pm}j$	`p p > wpm j`	$2.045 \pm 0.001 \cdot 10^4$	$^{+19.7\%}_{-17.2\%}$ $^{+1.4\%}_{-1.1\%}$	$2.843 \pm 0.010 \cdot 10^4$	$^{+5.9\%}_{-8.0\%}$ $^{+1.3\%}_{-1.1\%}$
a.3	$pp \to W^{\pm}jj$	`p p > wpm j j`	$6.805 \pm 0.015 \cdot 10^3$	$^{+24.5\%}_{-18.6\%}$ $^{+0.8\%}_{-0.7\%}$	$7.786 \pm 0.030 \cdot 10^3$	$^{+2.4\%}_{-6.0\%}$ $^{+0.9\%}_{-0.8\%}$
a.4	$pp \to W^{\pm}jjj$	`p p > wpm j j j`	$1.821 \pm 0.002 \cdot 10^3$	$^{+41.0\%}_{-27.1\%}$ $^{+0.5\%}_{-0.5\%}$	$2.005 \pm 0.008 \cdot 10^3$	$^{+0.9\%}_{-6.7\%}$ $^{+0.6\%}_{-0.5\%}$
a.5	$pp \to Z$	`p p > z`	$4.248 \pm 0.005 \cdot 10^4$	$^{+14.6\%}_{-15.8\%}$ $^{+2.0\%}_{-1.6\%}$	$5.410 \pm 0.022 \cdot 10^4$	$^{+4.6\%}_{-8.6\%}$ $^{+1.9\%}_{-1.5\%}$
a.6	$pp \to Zj$	`p p > z j`	$7.209 \pm 0.005 \cdot 10^3$	$^{+19.3\%}_{-17.0\%}$ $^{+1.2\%}_{-1.0\%}$	$9.742 \pm 0.035 \cdot 10^3$	$^{+5.8\%}_{-7.8\%}$ $^{+1.2\%}_{-1.0\%}$
a.7	$pp \to Zjj$	`p p > z j j`	$2.348 \pm 0.006 \cdot 10^3$	$^{+24.3\%}_{-18.5\%}$ $^{+0.6\%}_{-0.6\%}$	$2.665 \pm 0.010 \cdot 10^3$	$^{+2.5\%}_{-6.0\%}$ $^{+0.7\%}_{-0.7\%}$
a.8	$pp \to Zjjj$	`p p > z j j j`	$6.314 \pm 0.008 \cdot 10^2$	$^{+40.8\%}_{-27.0\%}$ $^{+0.5\%}_{-0.5\%}$	$6.996 \pm 0.028 \cdot 10^2$	$^{+1.1\%}_{-6.8\%}$ $^{+0.5\%}_{-0.5\%}$
a.9	$pp \to \gamma j$	`p p > a j`	$1.964 \pm 0.001 \cdot 10^4$	$^{+31.2\%}_{-26.0\%}$ $^{+1.7\%}_{-1.8\%}$	$5.218 \pm 0.025 \cdot 10^4$	$^{+24.5\%}_{-21.4\%}$ $^{+1.4\%}_{-1.6\%}$
a.10	$pp \to \gamma jj$	`p p > a j j`	$7.815 \pm 0.008 \cdot 10^3$	$^{+32.8\%}_{-24.2\%}$ $^{+0.9\%}_{-1.2\%}$	$1.004 \pm 0.004 \cdot 10^4$	$^{+5.9\%}_{-10.9\%}$ $^{+0.8\%}_{-1.2\%}$

+127 more pp processes from 1405.0301!

(Not a random sampling)

[Alwall et al.](#)

Which of these distributions do you expect?

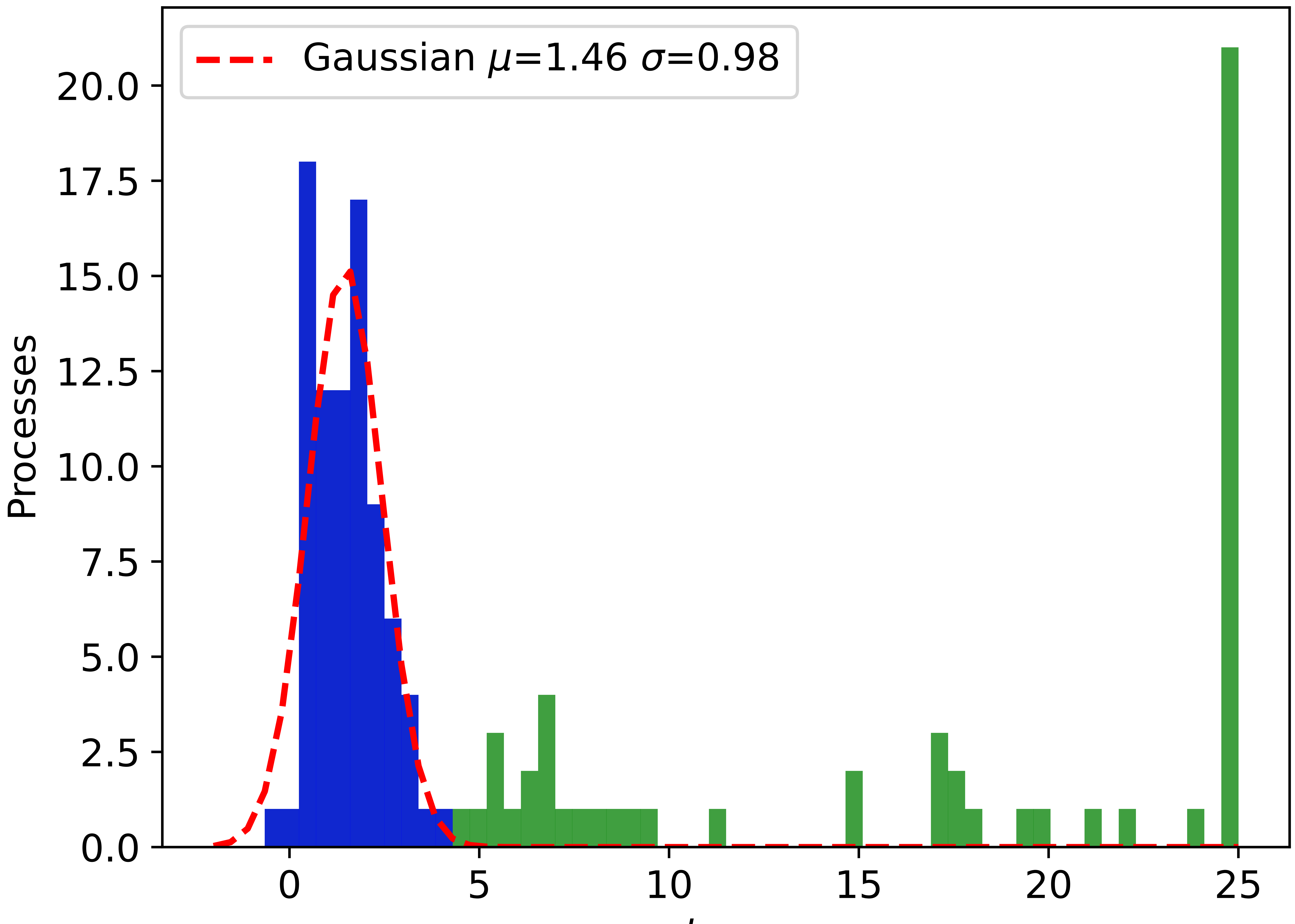


$$t_{scale} = \frac{\sigma_{NLO} - \sigma_{LO}}{\Delta\sigma_{LO\ scale}}$$

Statistical patterns of scale variation uncertainties at LO

Study pull distribution

$$t_{scale} = \frac{\sigma_{NLO} - \sigma_{LO}}{\Delta\sigma_{LO\ scale}}$$



A desire to have a more meaningful NPs

Up: $\mu_+ = 2\ \mu_0$

Down: $\mu_- = \frac{1}{2}\ \mu_0$

Experiments interpolate between up / down variations and fit NPs

Could we have a more physically motivated description of uncertainties ? [Eg. Suggestion at Les Houches 2019]

Then we could meaningfully think of propagating / constraining them…, better account for correlations when combining measurements

A Possible Solution.

$$\sigma = c_0 + \alpha_s(\mu)[c_1 + \alpha_s(\mu)\, c_2 + \cdots]$$

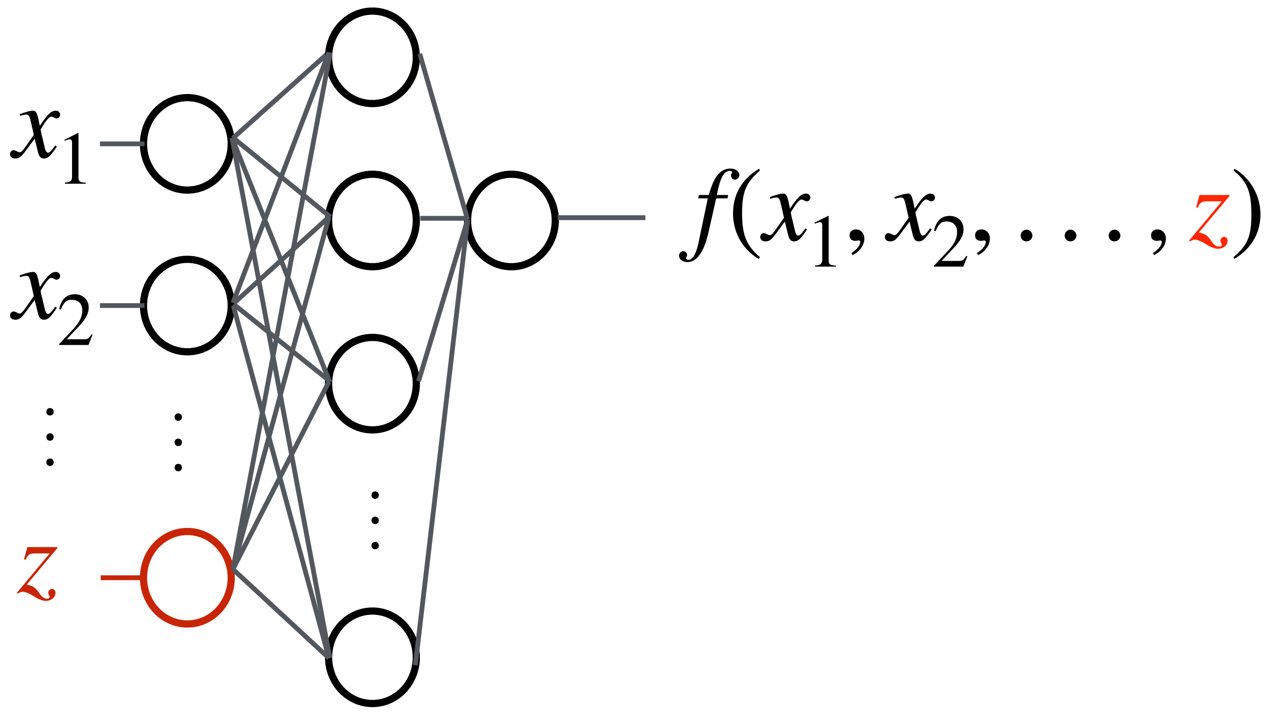
Identify the actual source of uncertainty

- The unknown higher-order corrections: $\alpha_s(\mu)\, c_2 + \cdots$

Parametrize and vary the unknown

- We often know quite a lot about the general structure of c_2
  - μ dependence, color structure, partonic channels, kinematic structure, ...
- Suitably parametrize the missing pieces
  - Simplest case: c_2 is just a number
  - More generally, have to parametrize an unknown function
- Common/independent pieces between different predictions determine the correlations between them

Frank Tackmann (DESY) Theory Uncertainties from Nuisance Parameters. 2019-06-14 7 / 17



Conclusion

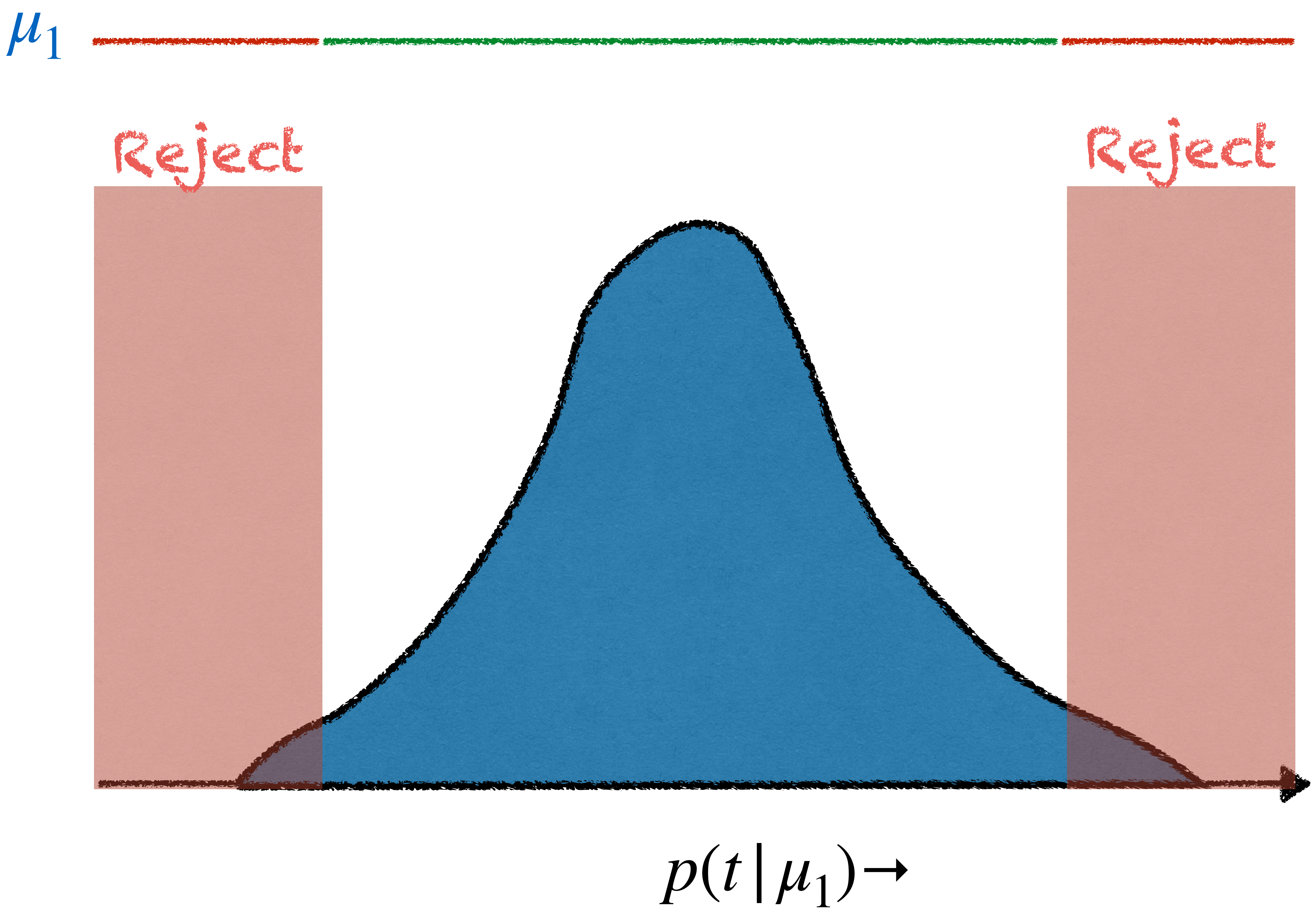
- ML more sensitive to simulation artefacts → building better uncertainty propagation tools
- If we have meaningful theory NPs, we could do more: constrain these terms, better quantify impact on measurements
- Opens the door to ML as interpretability tools to understand constrains

Neyman Construction

Hypothesis tests using arbitrary test statistic

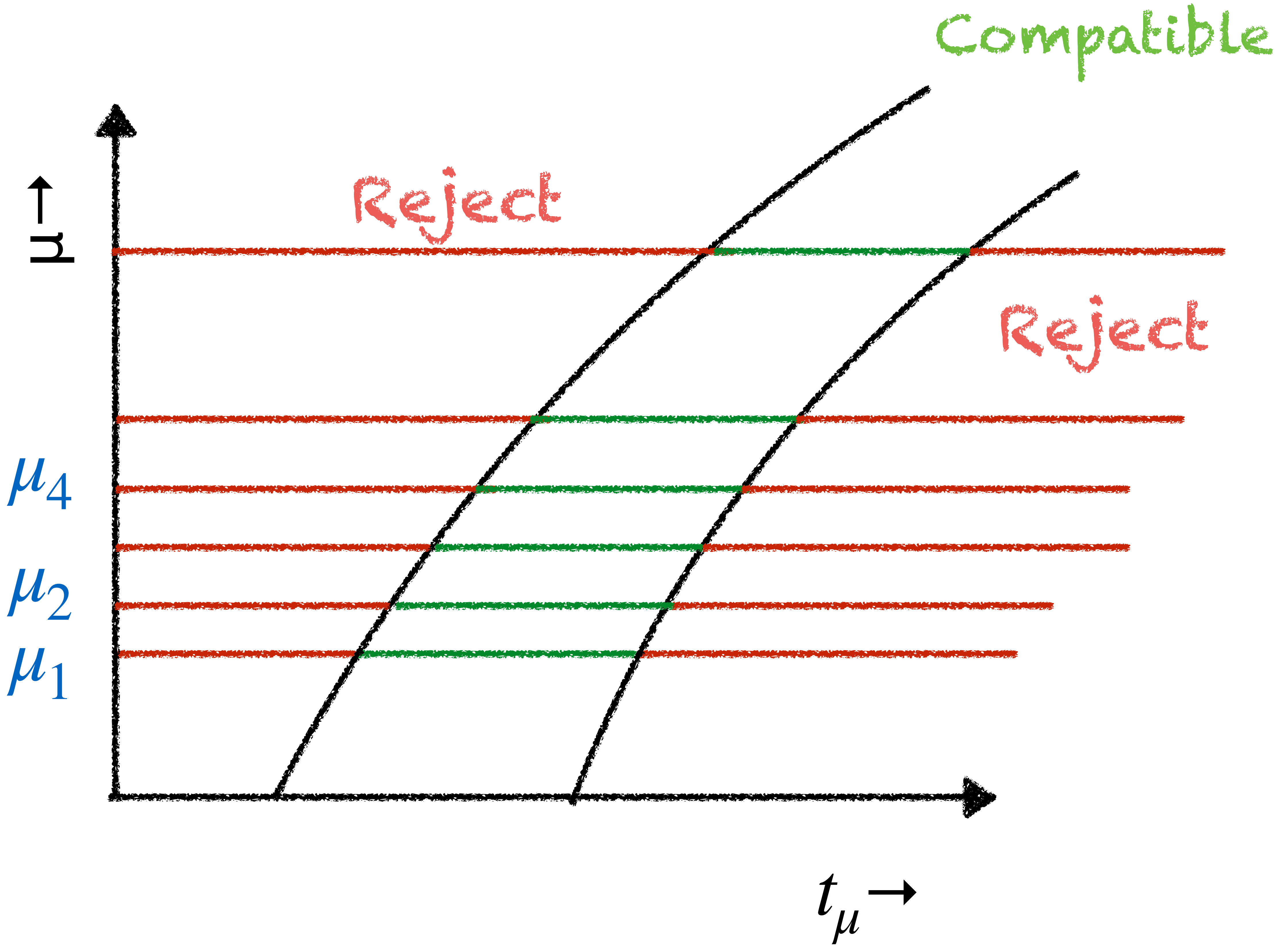
$$H_0 : \mu = \mu_1$$

$$P(t \in \omega \,|\, H_0) = \alpha$$



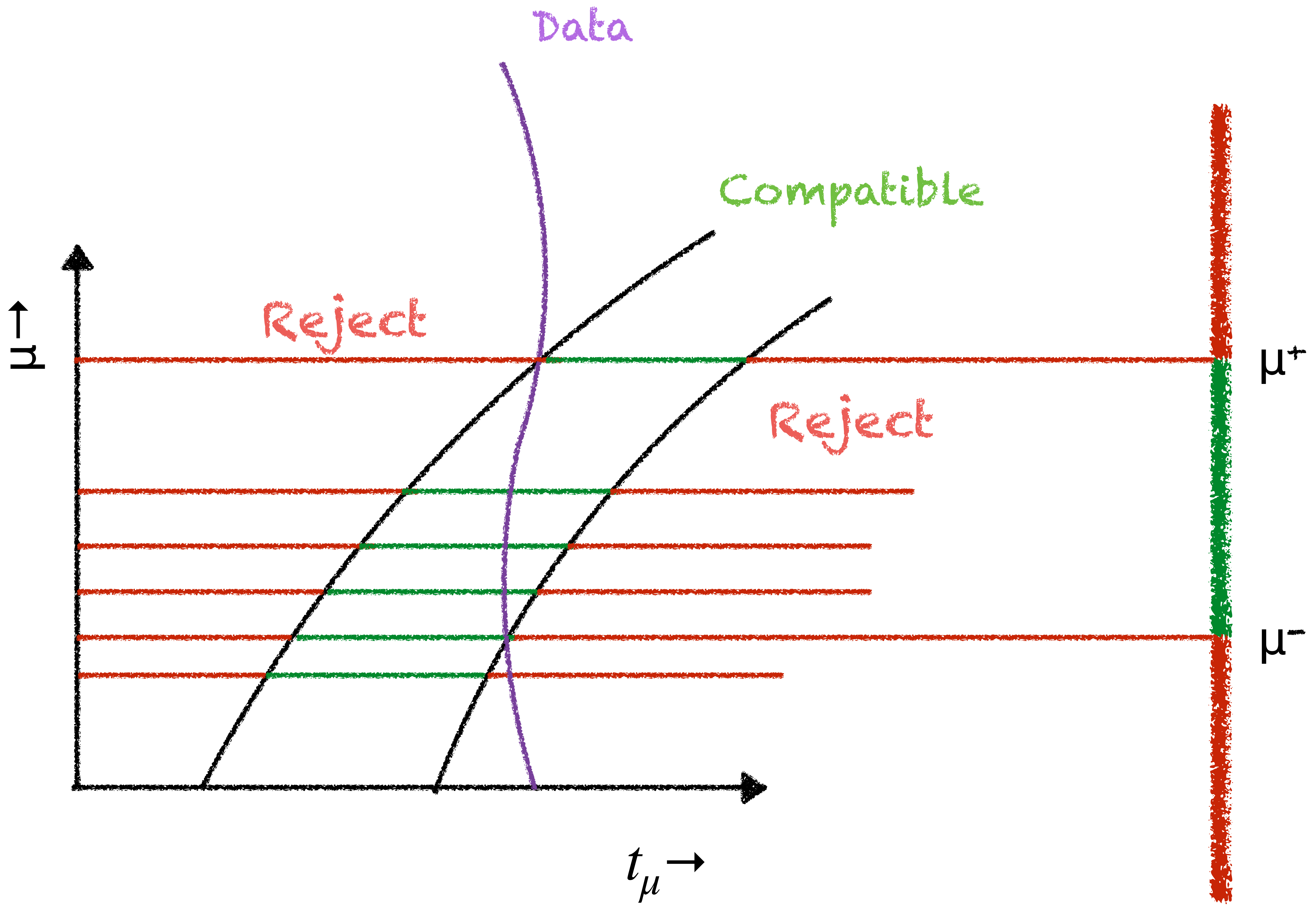
We can find the correct cuts by throwing toys

Neyman Construction



Notice t_μ can be different for each μ

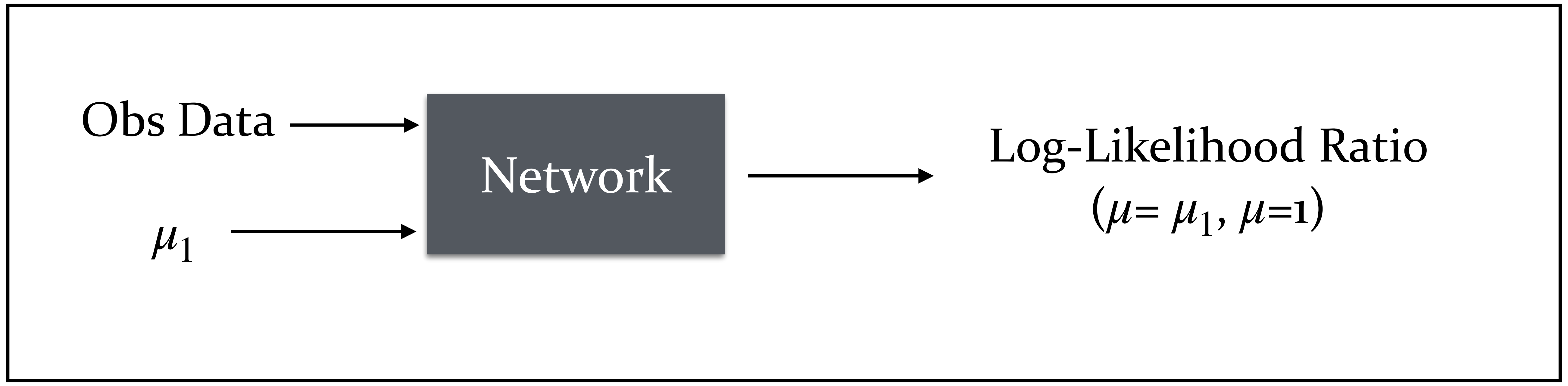
Neyman Construction



Constructing the test statistic with neural networks

Brehmer et al

Bypass the need for histograms & likelihood model based on Poisson distributions

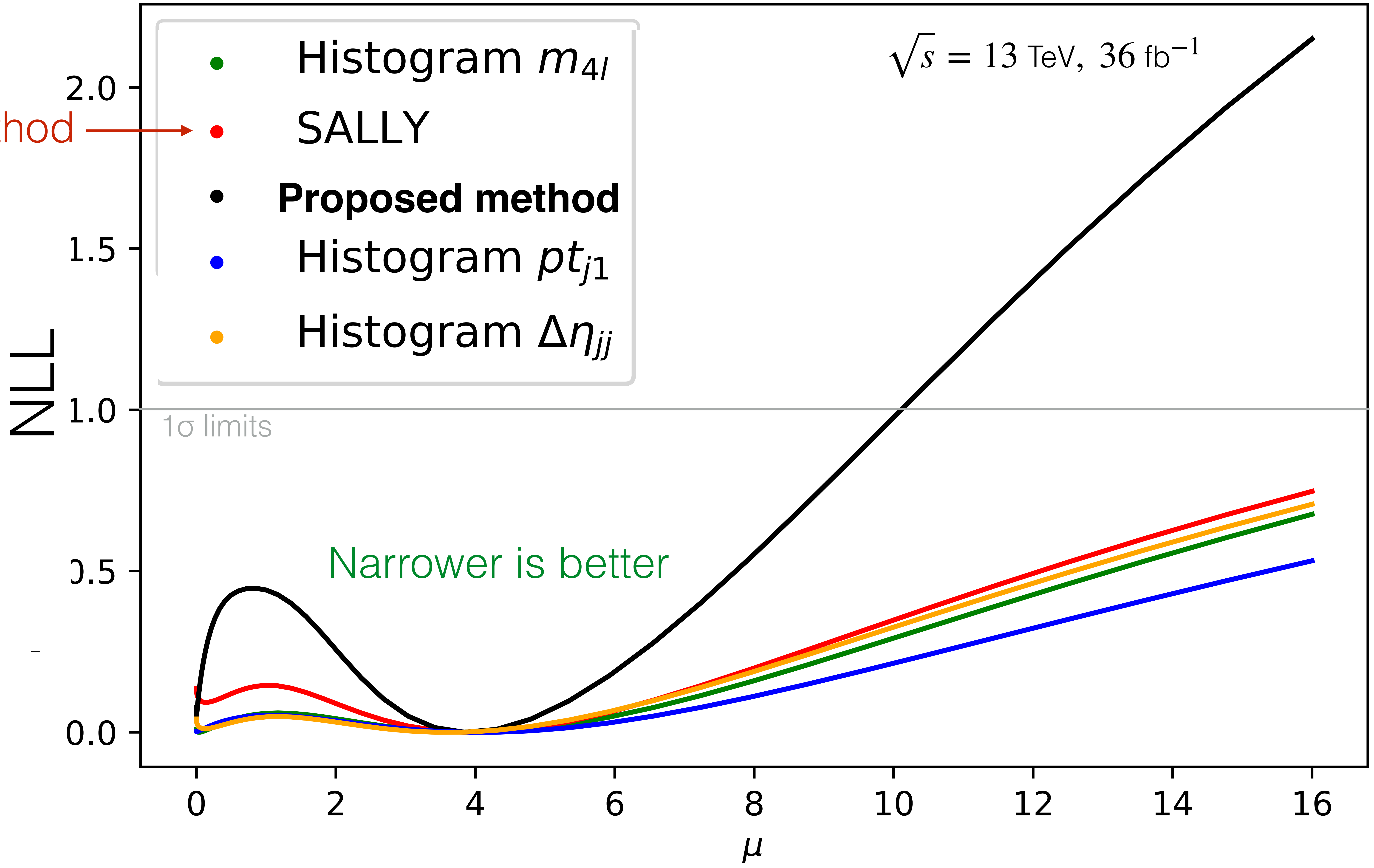


Even if the LR is only approximate, Neyman Construction treats it as "just another test statistic" and finds you the correct confidence intervals

Pheno study to recover sensitivity lost due to quantum interference

hal-02971995v3: **Aishik Ghosh**, David Rousseau

“Traditional ML” method

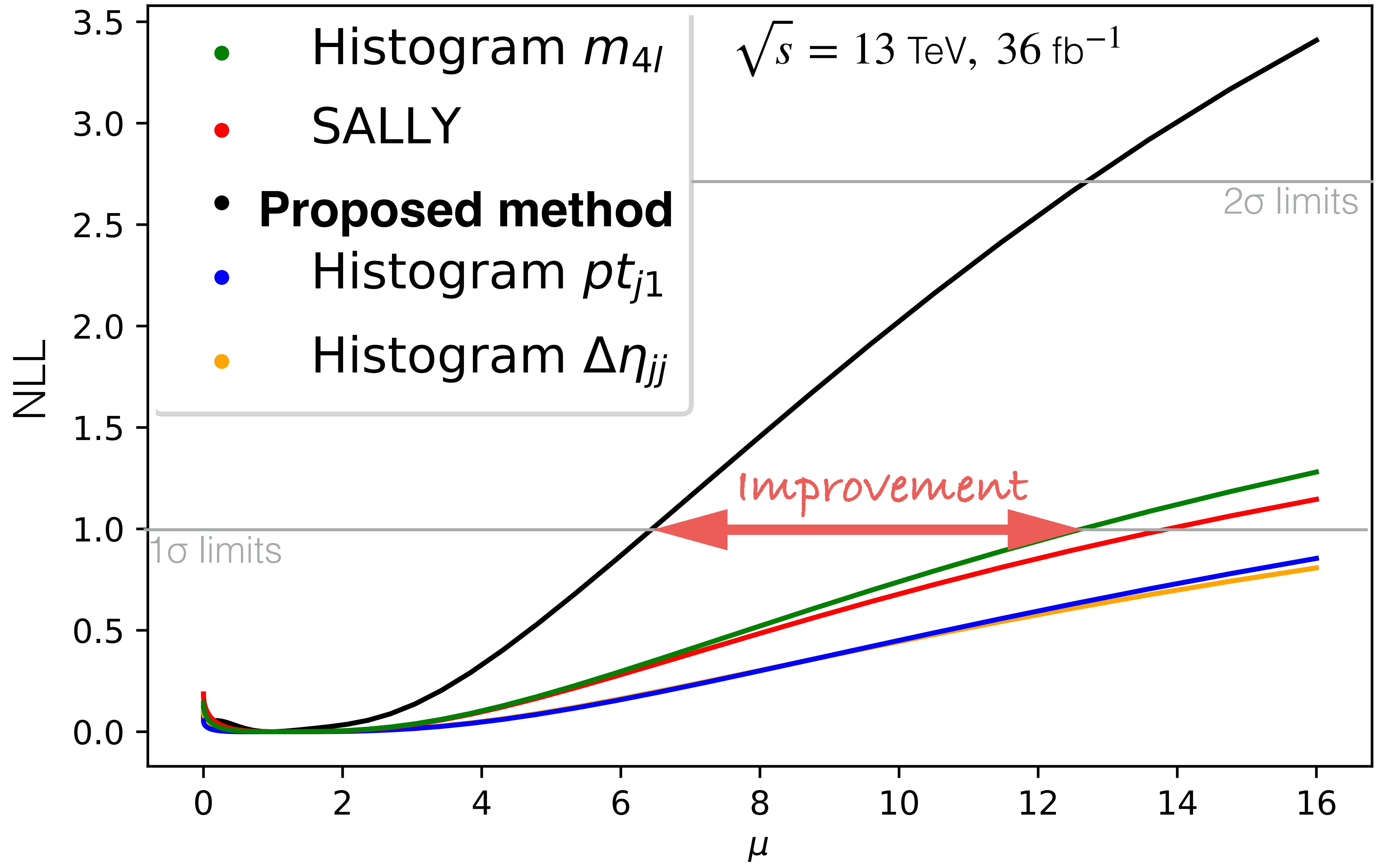


(e) $\mu = 4$, without rate

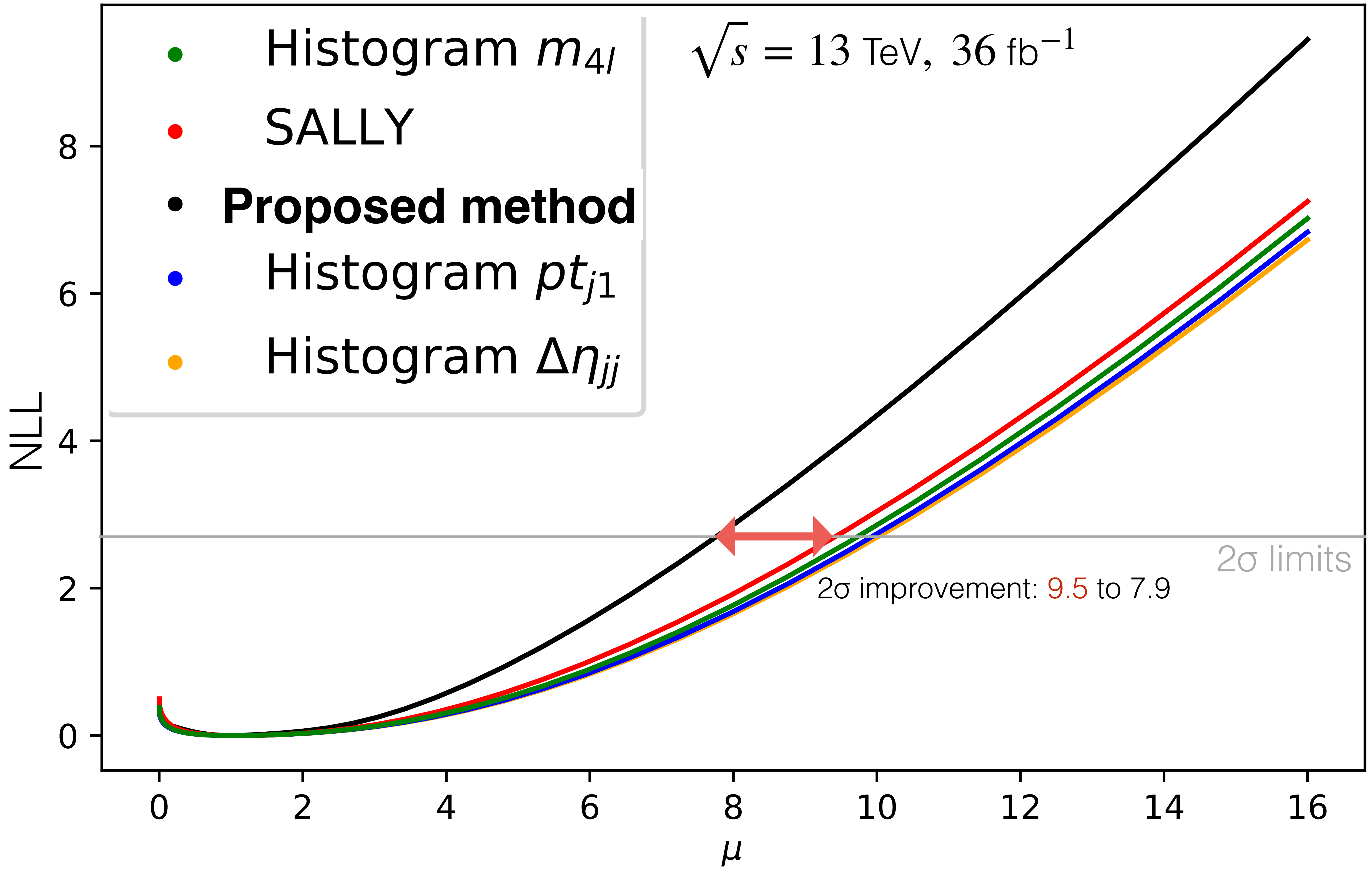
Pheno study: Madgraph+Pythia+Delphes
VBF samples

Expected sensitivity at μ=1

hal-02971995v3: **Aishik Ghosh**, David Rousseau



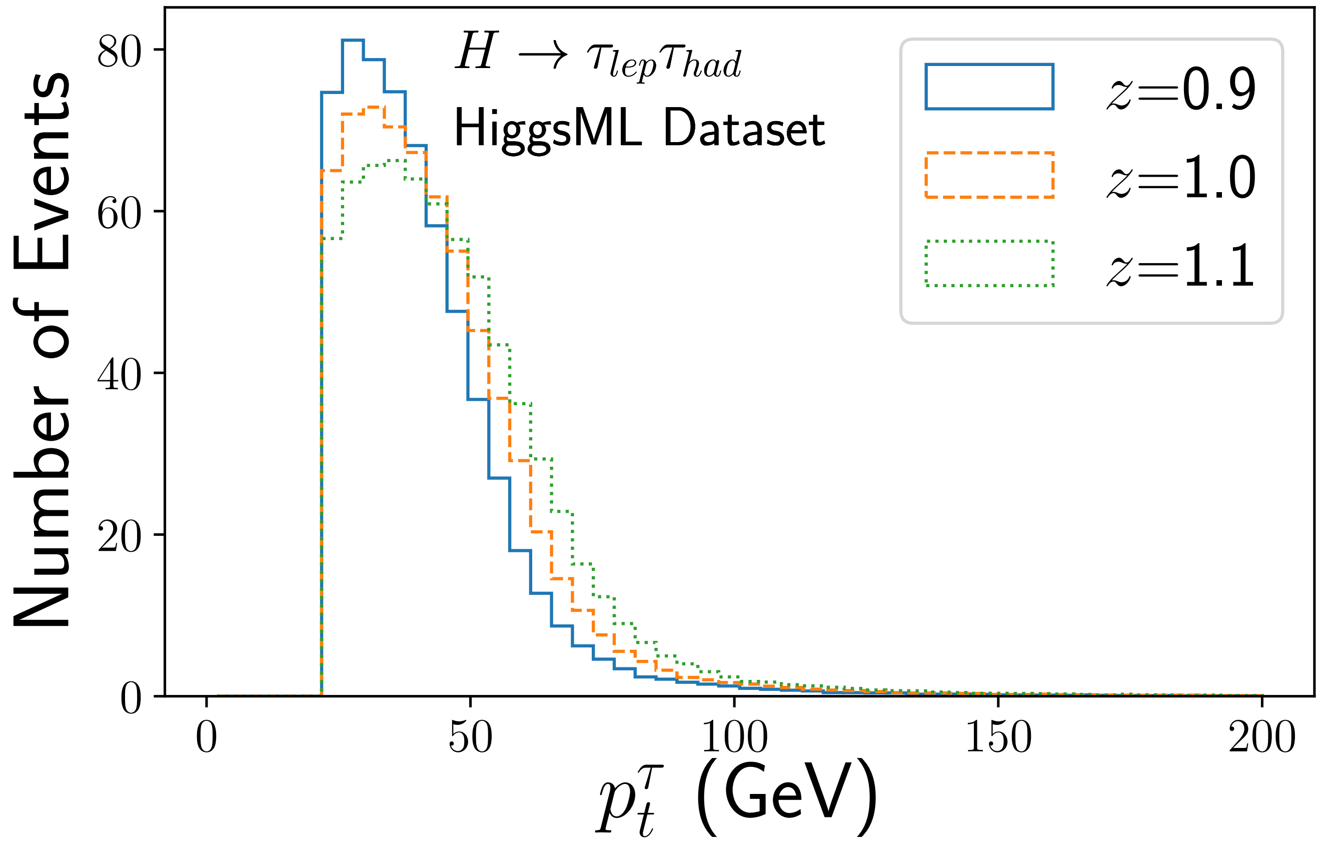
(a) SM, without rate

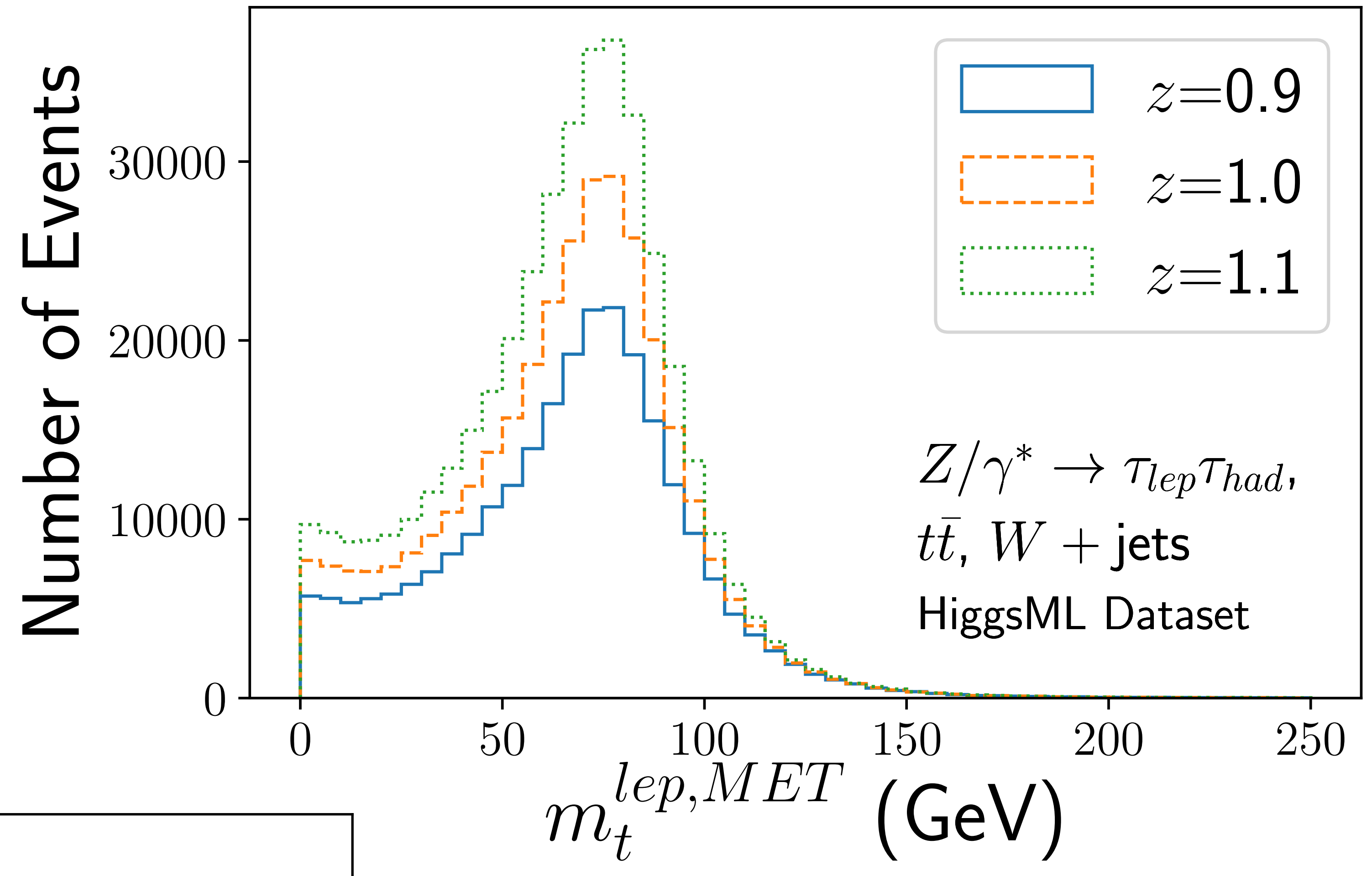


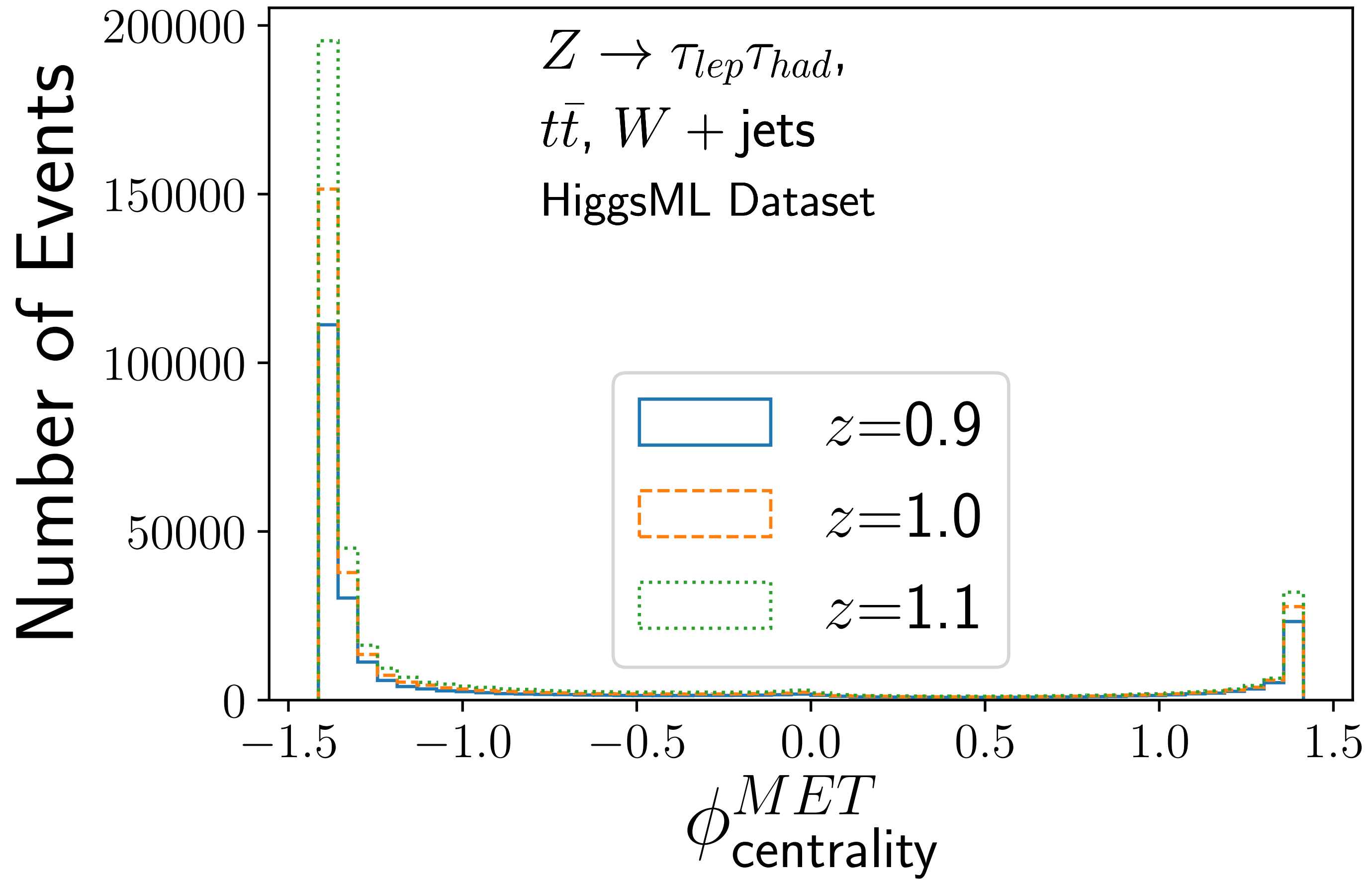
(b) SM with rate

Physics Data: HiggsML + Tau Energy Scale (TES) Uncertainty

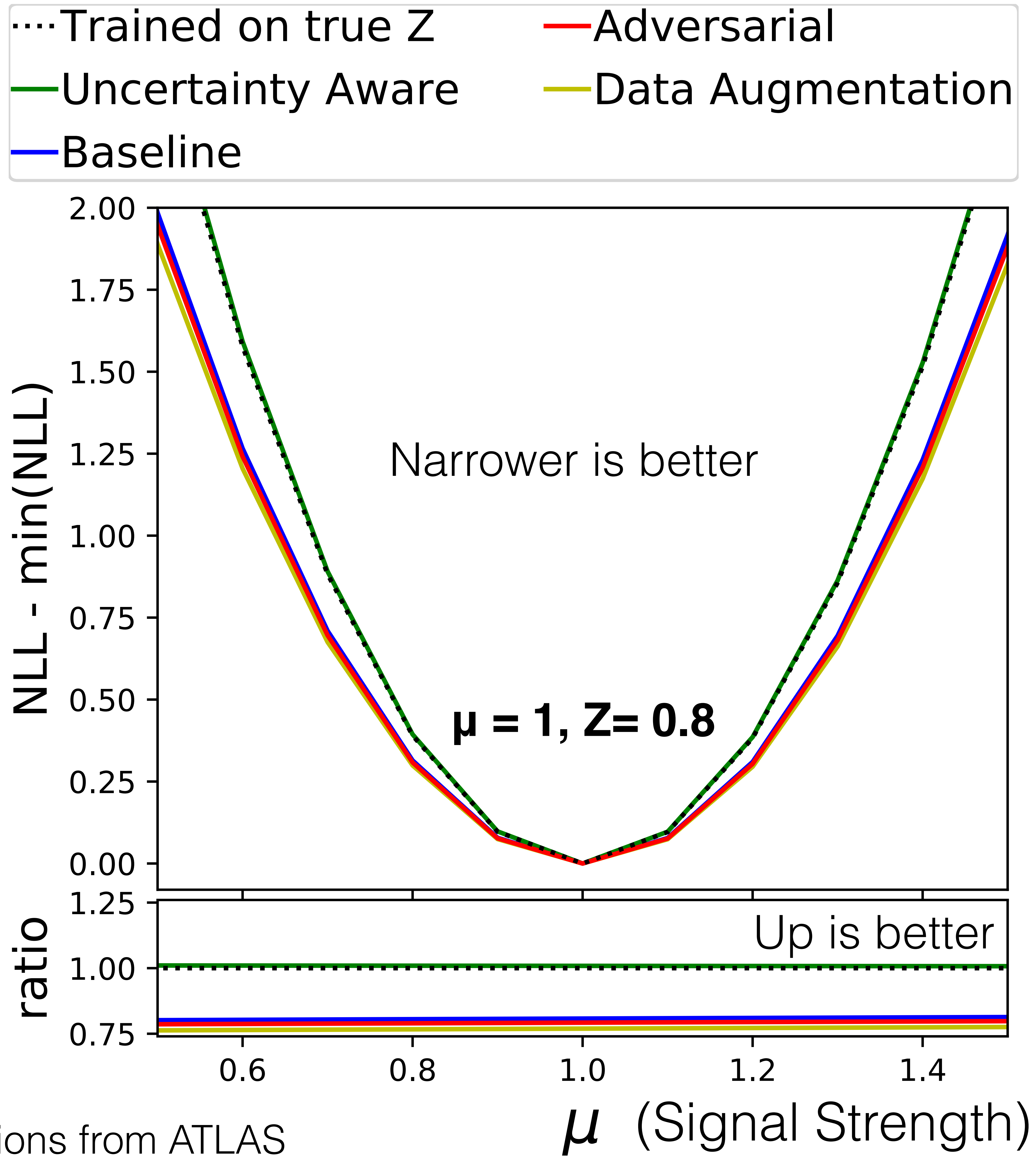
Parameter of Interest is Higgs signal strength μ, and TES is the nuisance parameter Z







Physics Data: HiggsML + Tau Energy Scale (TES) Uncertainty

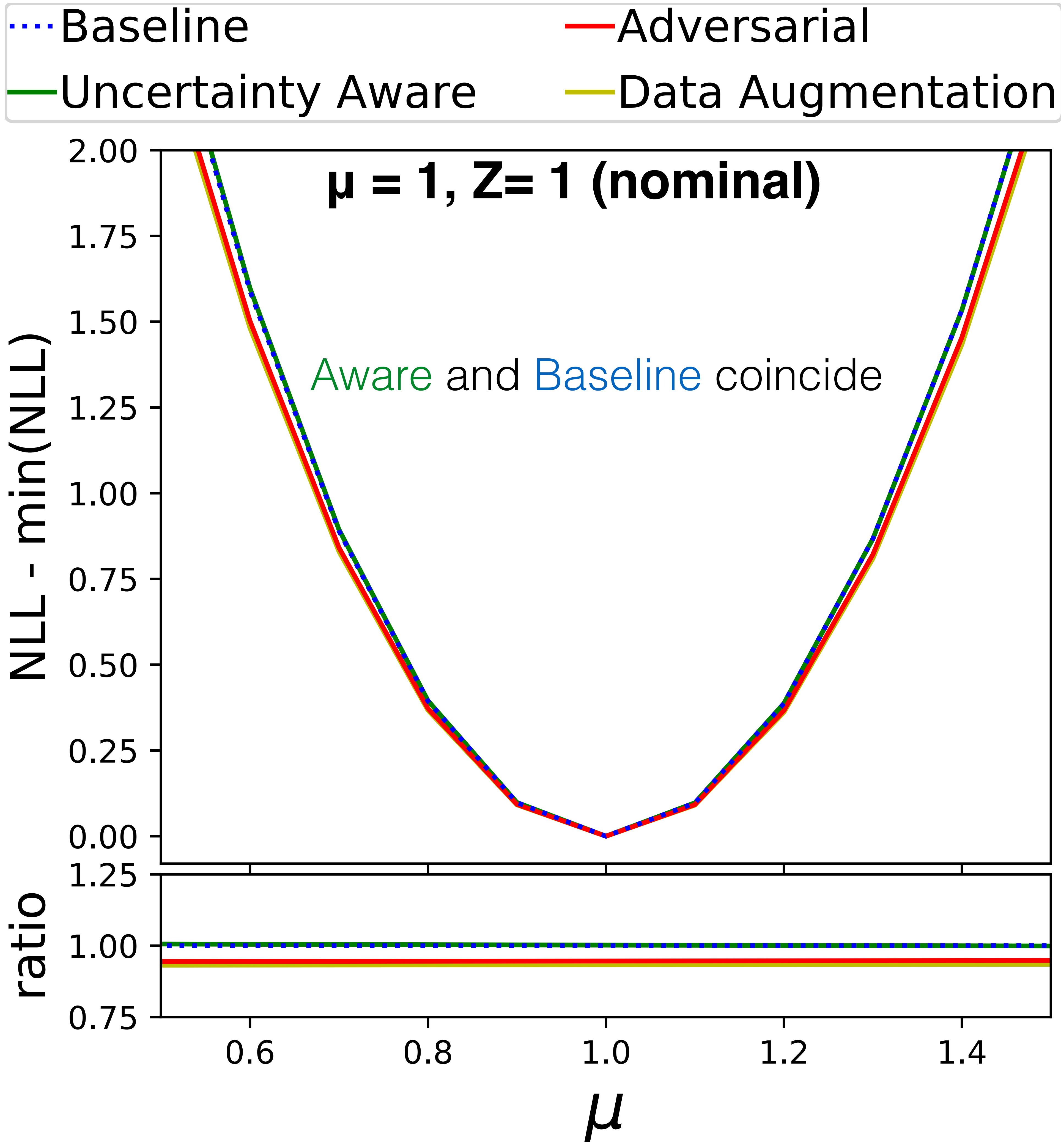


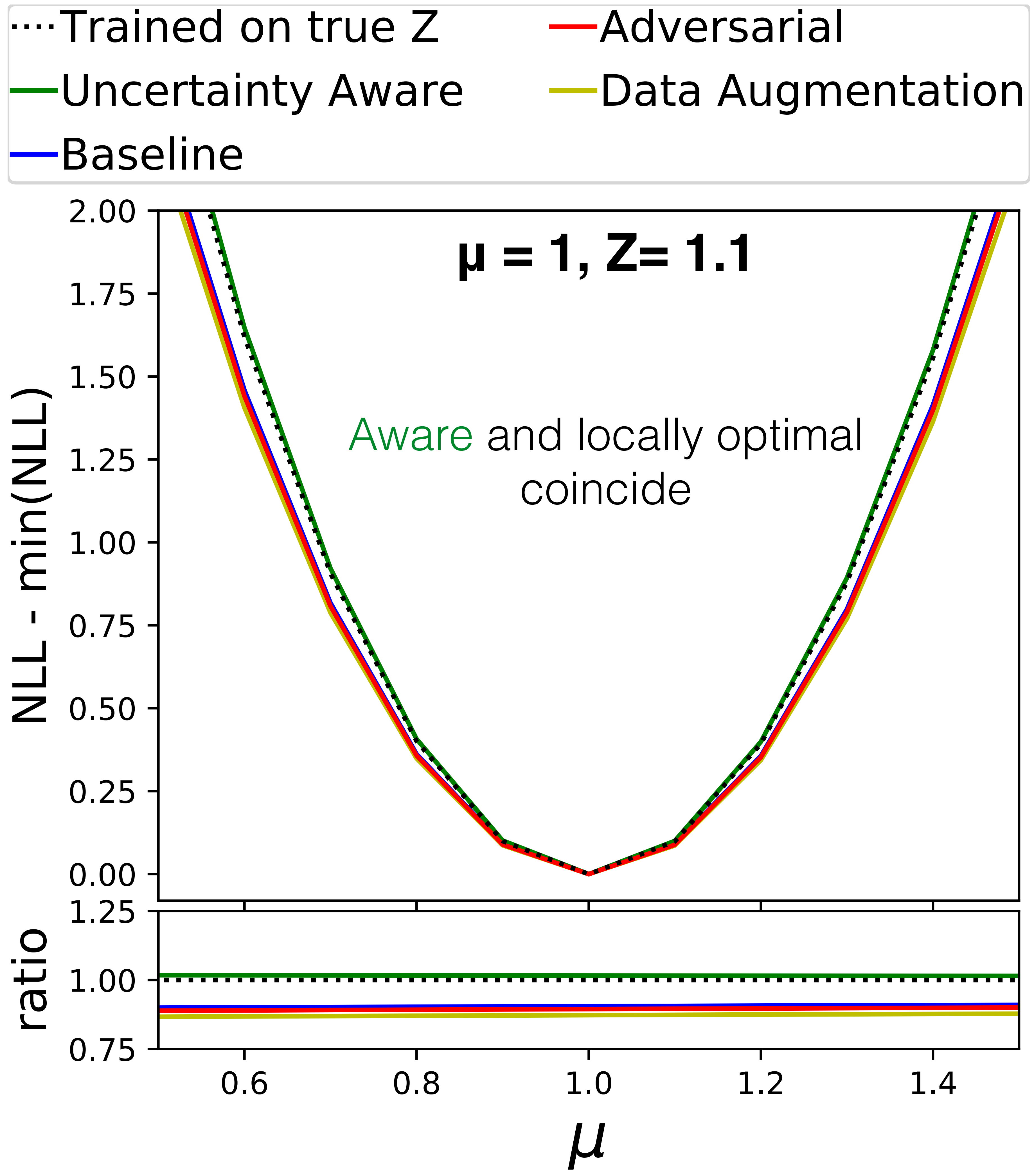
Uncertainty-Aware coincides with classifier trained on true Z

⇒ Can't get much better than that!

Simulations from ATLAS

Test performance for “observed” data at nominal and above nominal Z

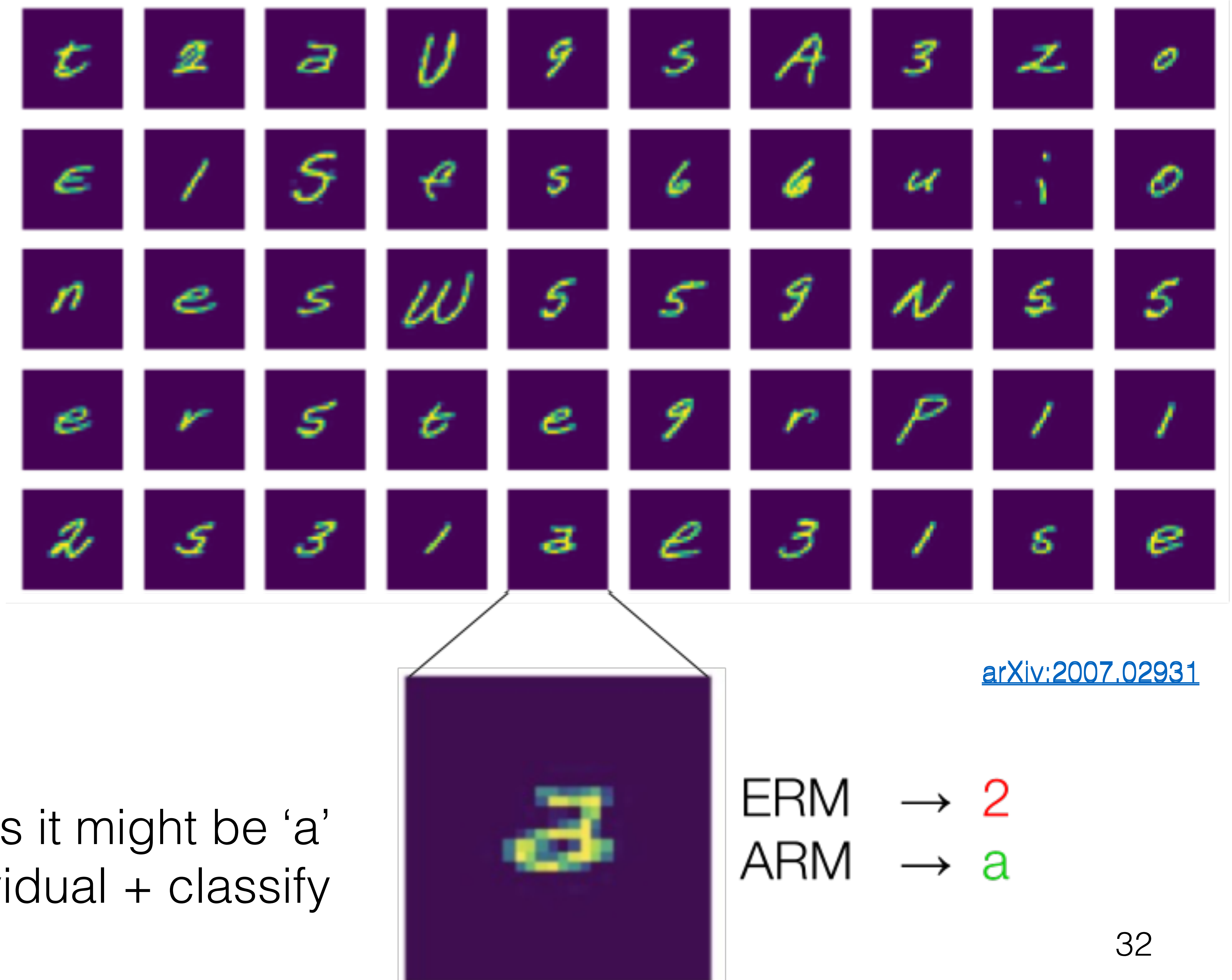




In every case the Aware Classifier is as good as the optimal one, no other technique matches its performance everywhere

Idea fascinating also to ML researchers !

- ML researchers assume i.i.d
- This technique exploits correlations between samples – a different paradigm
- Interesting applications outside of physics



arXiv:2007.02931

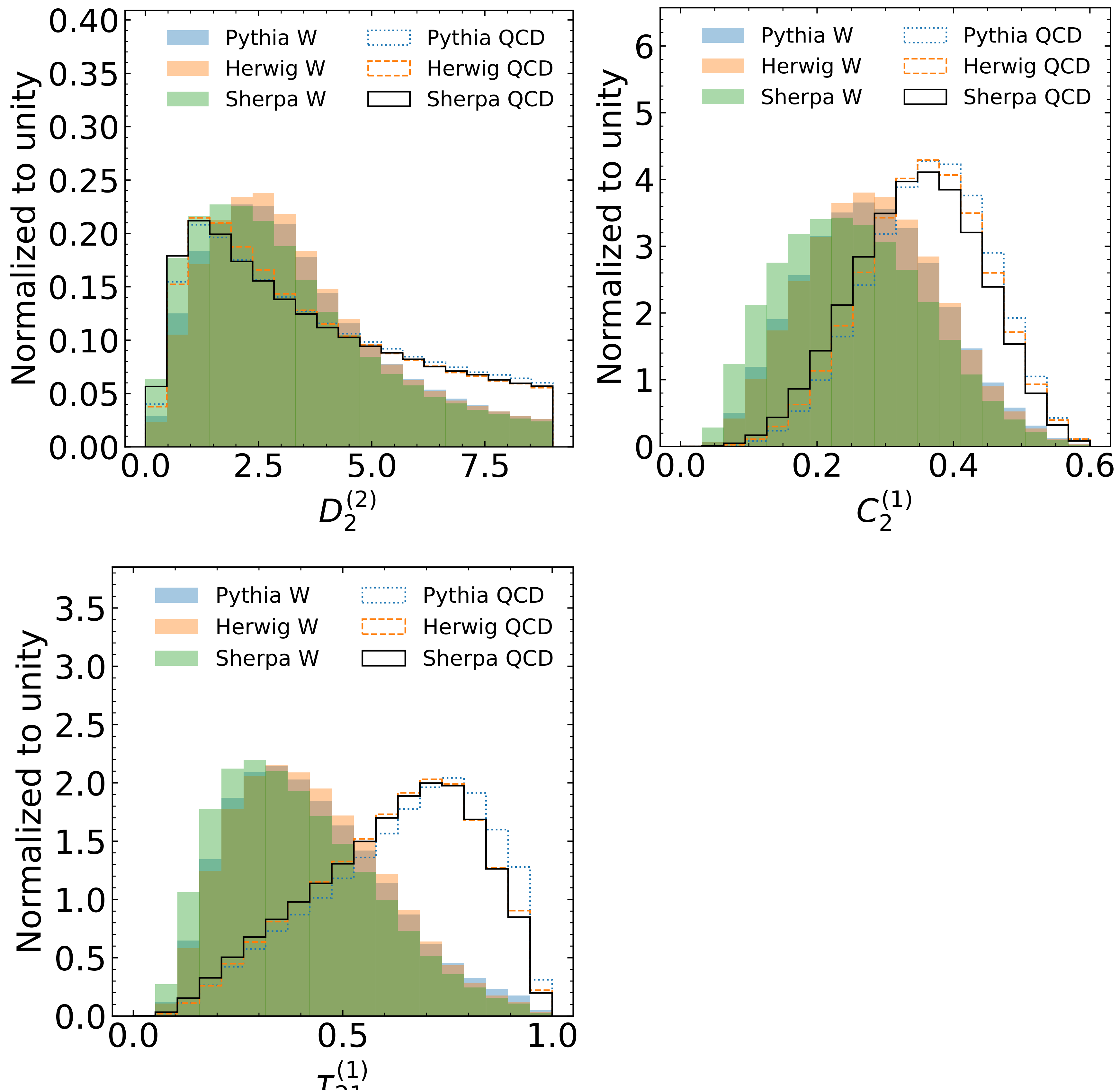
For my handwriting this is ‘2’, for yours it might be ‘a’
ARM: Adapt to the individual + classify

Case Study 1: Two-point uncertainty (fragmentation modelling)

Goal: W jets vs QCD jets

Decorrelation: Reduce difference in performance on `Herwig` vs `Pythia`

Cross-check: Test uncertainty estimate from {`Herwig` vs `Pythia`} using `Sherpa`

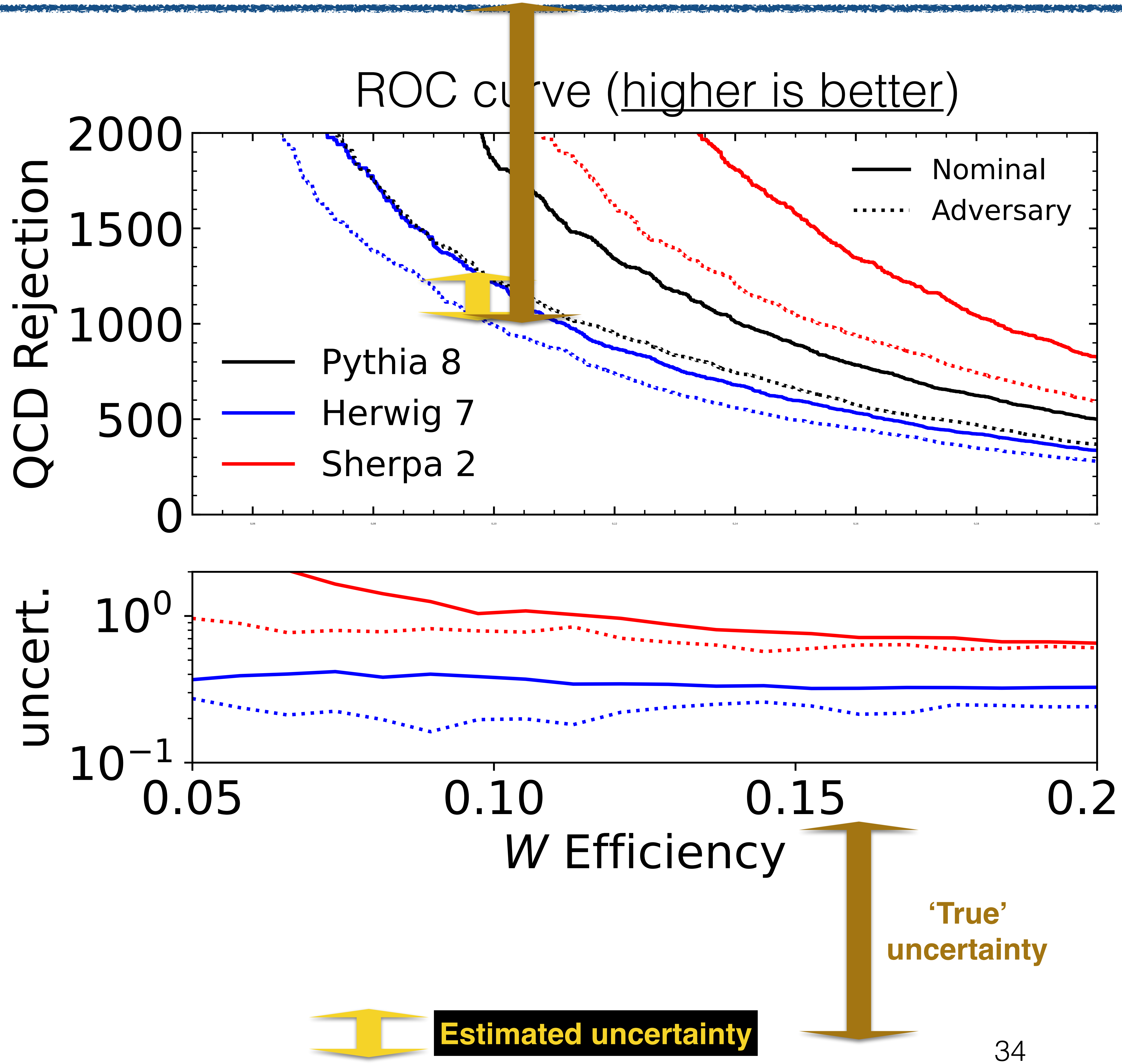


Case Study 1: Two-point uncertainty - Result

Adversary successfully sacrifices separation power in order to reduce difference in performance between Herwig and Pythia

Cross-check with Sherpa reveals uncertainty severely underestimated by usual Herwig vs Pythia comparison

In an typical LHC analysis, a cross-check with third generator rarely performed, similar to prior work suggesting decorrelation for theory uncertainties



Case Study 2: Higher-order corrections

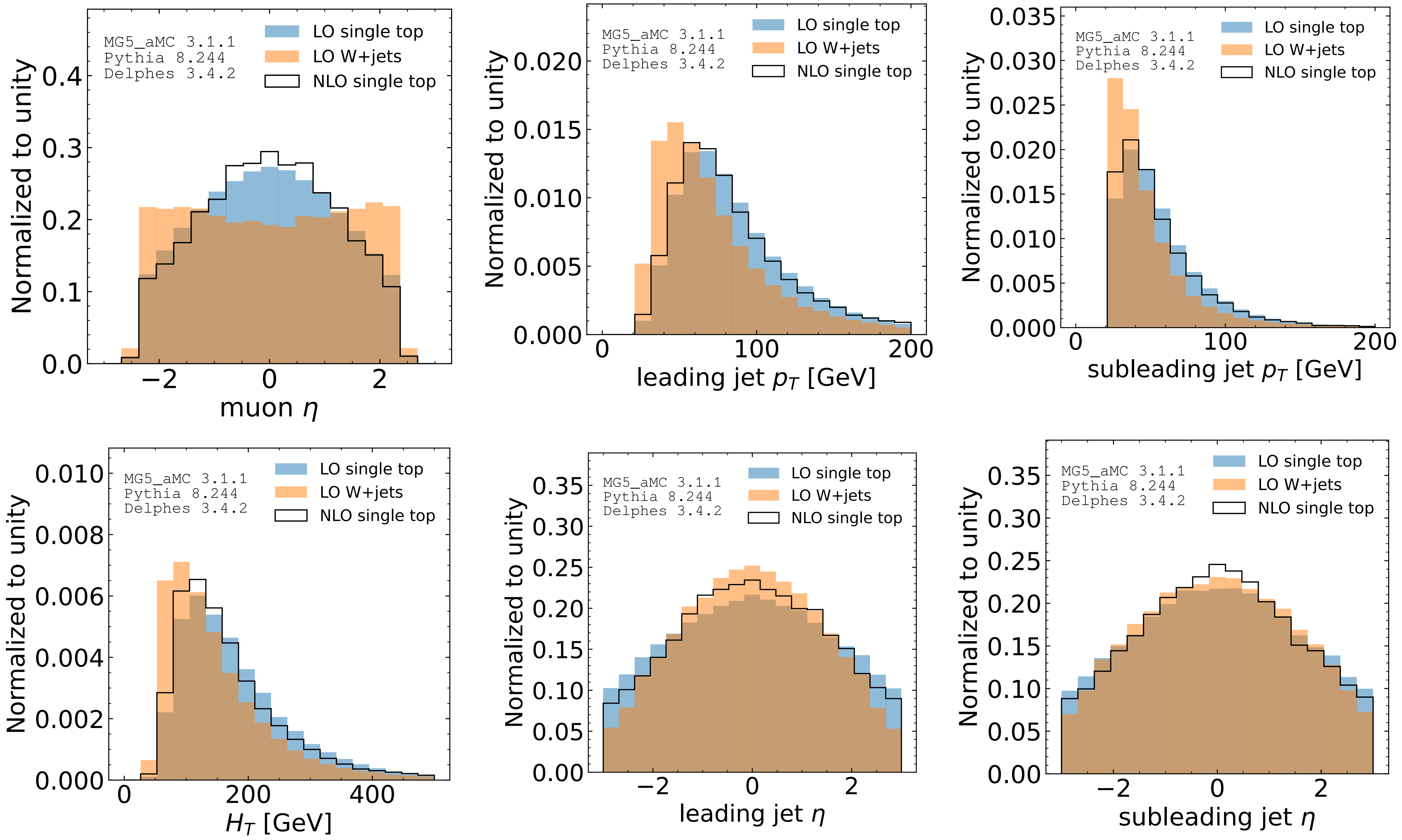
- We can’t calculate QFT to infinite order
- Artefact of truncation of series: Varying certain unphysical scales changes predictions
- Uncertainty quantification: Vary scales (renormalization scale, factorisation scale) between 1/2 to 2 in MC, see change in prediction

Scale uncertainty – Problem Setup

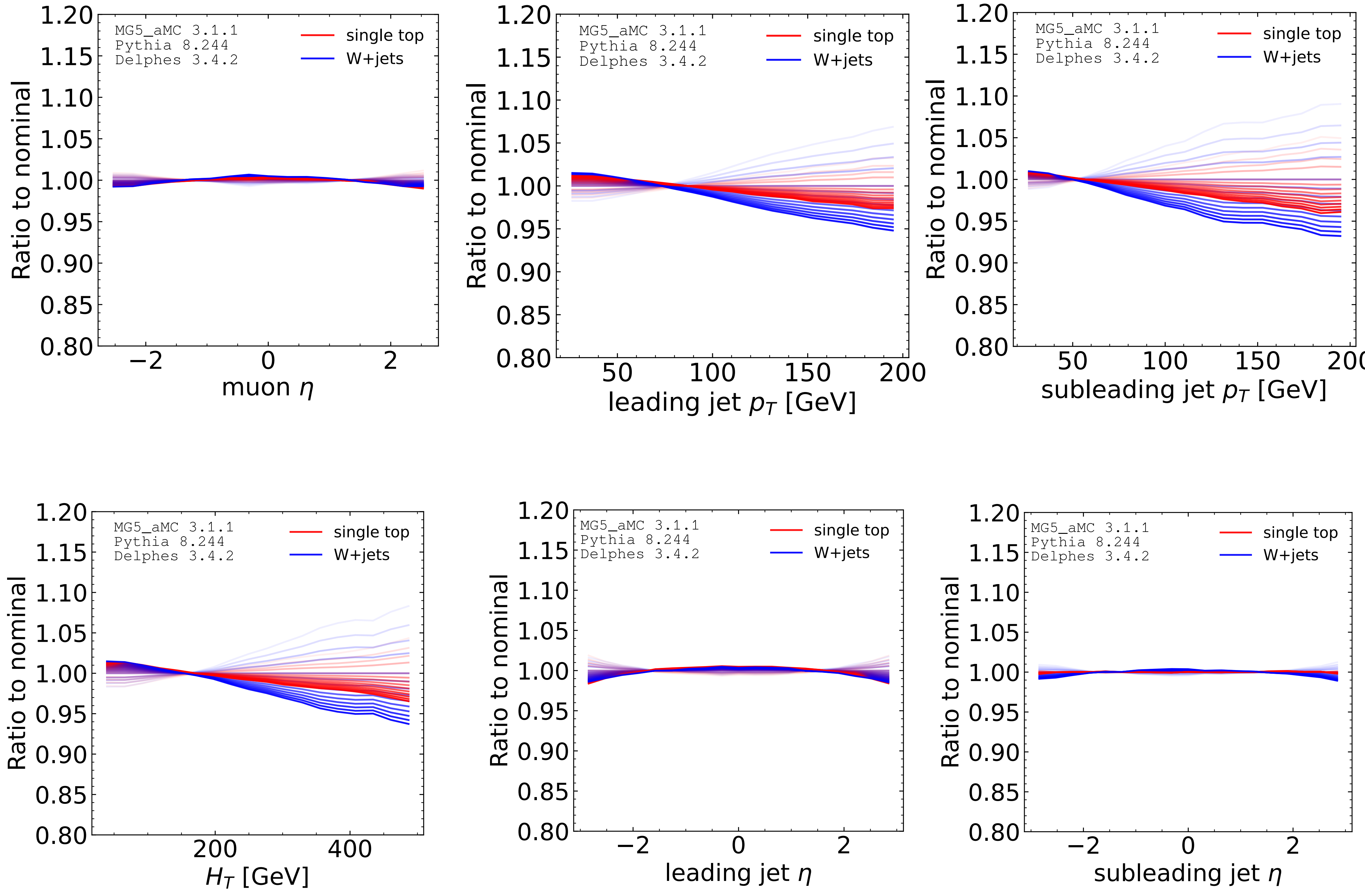
Goal: Single top vs W+Jets

Decorrelation: Reduce difference in performance on scale variations at LO

Cross-check: Test uncertainty estimate from {scale variations at LO} using NLO



NLO vs LO



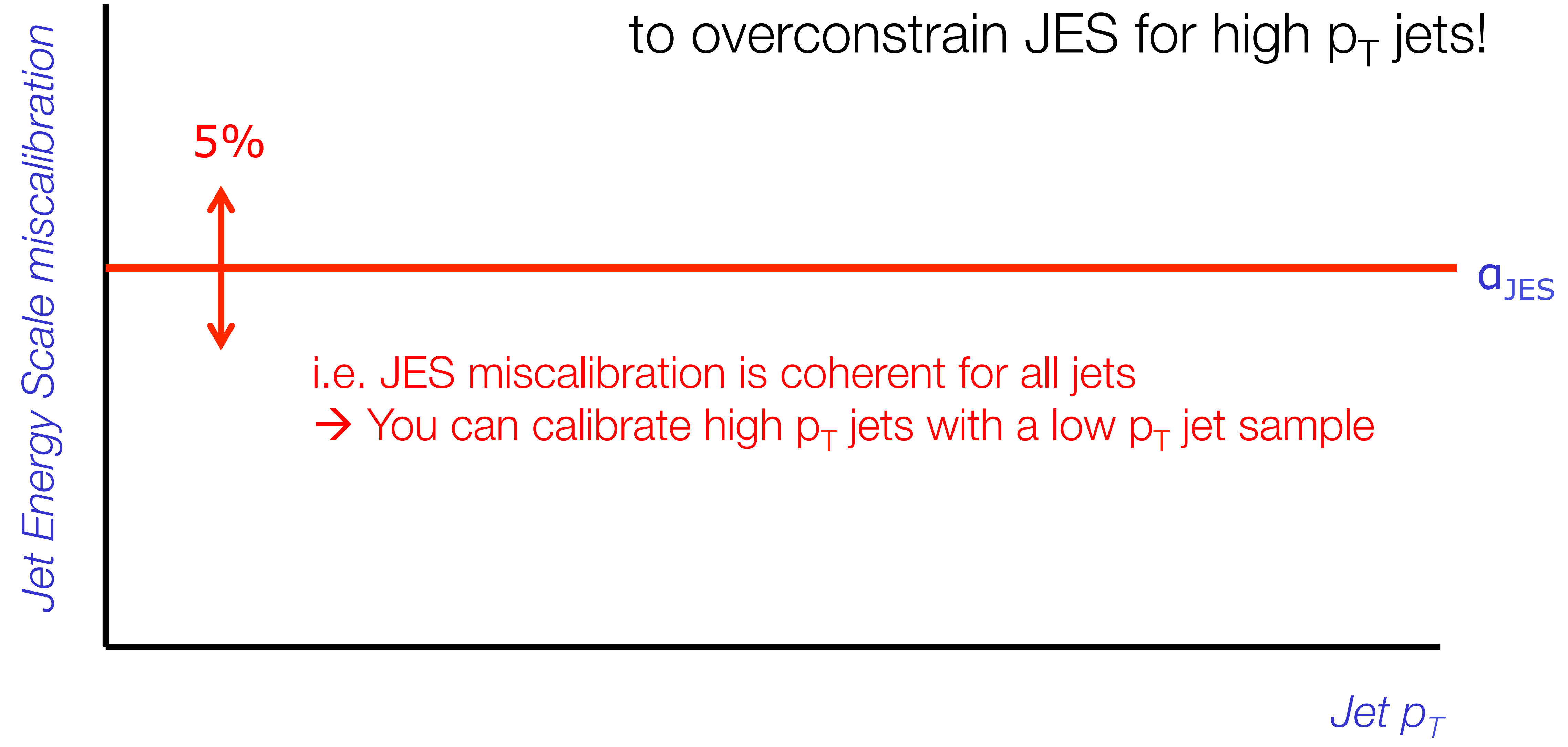
Factorisation scale variations going from 1/2 to 2

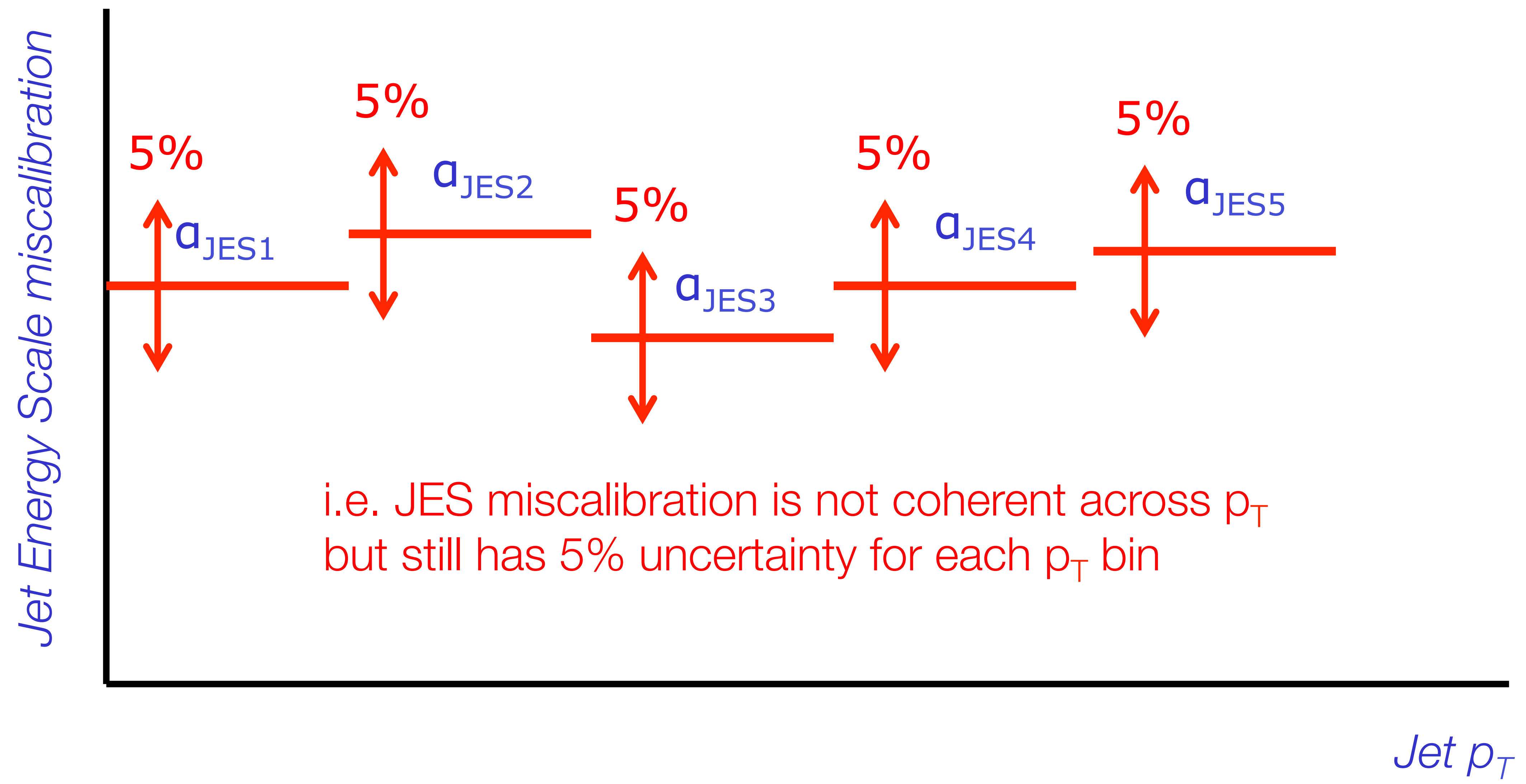
Overconstraining NP

From W. Verkerke:

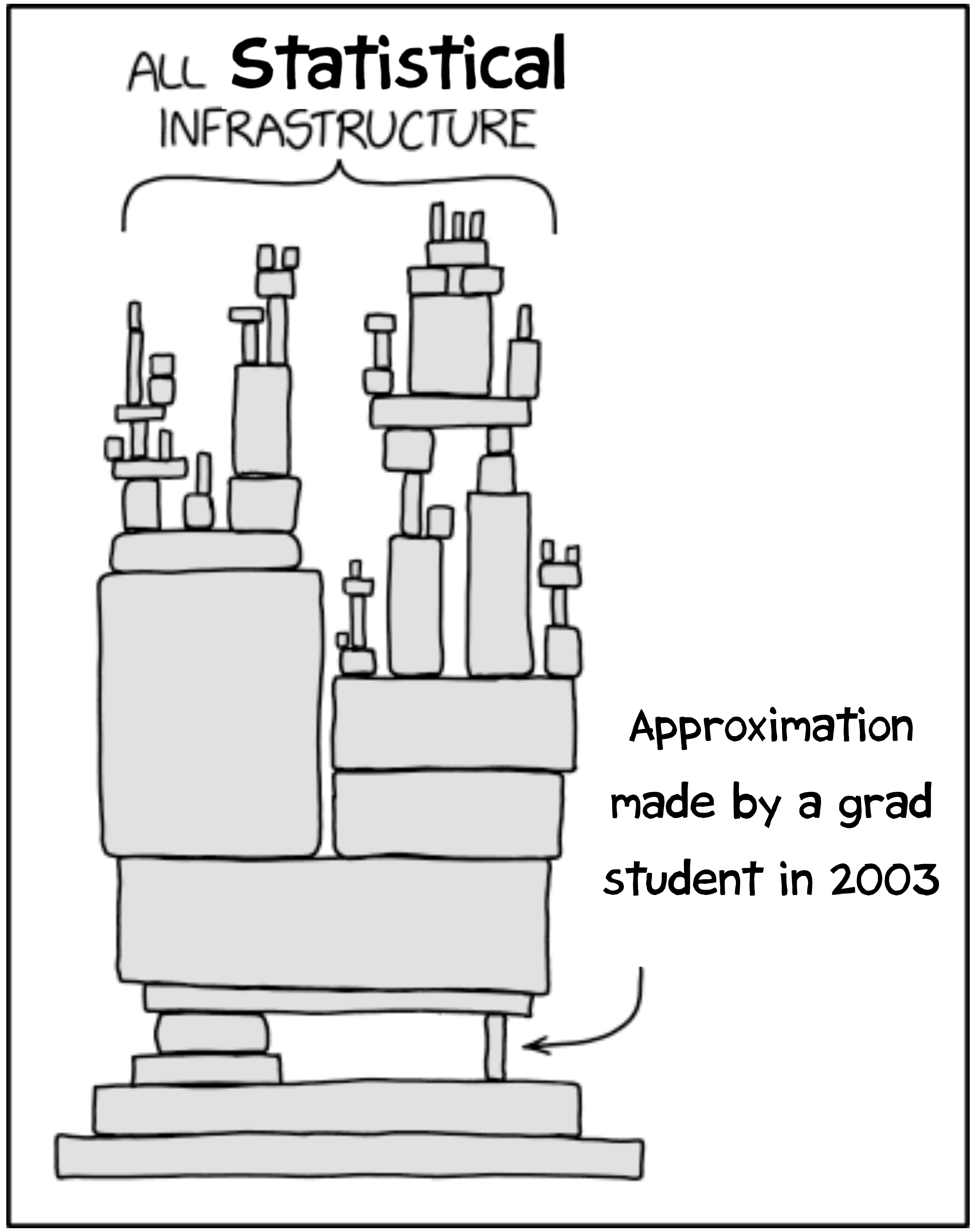
Our modelling of NPs might be over-simplified

- If you assume one NP – chances are that your physics Likelihood will exploit this oversimplified JES model to overconstrain JES for high p_T jets!





Nuisance Parameter Infrastructure



From Daniel Whiteson
Inspired by XKCD

Time to re-examine
some of the
underlying pieces

Are they up to the
task of the precision era?

What processes populate the tail ?

Process	n_{part}	$\Delta\sigma/\sigma_0$	$\frac{\sigma_{\text{NLO}}-\sigma_0}{\Delta\sigma}$
`p p > wpm`	1	1.54×10^{-1}	1.84
`p p > wpm j`	2	1.97×10^{-1}	1.96
`p p > wpm j j`	3	2.45×10^{-1}	0.59
`p p > wpm j j j`	4	4.10×10^{-1}	0.25
`p p > z`	1	1.46×10^{-1}	1.87
`p p > z j`	2	1.93×10^{-1}	1.82
`p p > z j j`	3	2.43×10^{-1}	0.56
`p p > z j j j`	4	4.08×10^{-1}	0.27
`p p > a j`	2	3.12×10^{-1}	5.33
`p p > a j j`	3	3.28×10^{-1}	0.85
`p p > w+ w- wpm`	3	1.00×10^{-3}	610.69
`p p > z w+ w-`	3	8.00×10^{-3}	92.39
`p p > z z wpm`	3	1.00×10^{-2}	85.00
`p p > z z z`	3	1.00×10^{-3}	302.75
`p p > a w+ w-`	3	1.90×10^{-2}	42.33
`p p > a a wpm`	3	4.40×10^{-2}	47.24
`p p > a z wpm`	3	1.00×10^{-3}	1244.49
`p p > a z z`	3	2.00×10^{-2}	17.24

https://arxiv.org/abs/2210.15167

Make correction in UQ for EW processes

Process	n_{part}	$\Delta\sigma/\sigma_0$	$\frac{\sigma_{\text{NLO}}-\sigma_0}{\Delta\sigma}$	$\Delta\sigma_{\text{ref}}/\sigma_0$	$\frac{\sigma_{\text{NLO}}-\sigma_0}{\Delta\sigma_{\text{ref}}}$
`p p > wpm`	1	1.54×10^{-1}	1.84	1.47×10^{-1}	1.92
`p p > wpm j`	2	1.97×10^{-1}	1.96	2.94×10^{-1}	1.31
`p p > wpm j j`	3	2.45×10^{-1}	0.59	4.41×10^{-1}	0.33
`p p > wpm j j j`	4	4.10×10^{-1}	0.25	5.88×10^{-1}	0.18
`p p > z`	1	1.46×10^{-1}	1.87	1.47×10^{-1}	1.86
`p p > z j`	2	1.93×10^{-1}	1.82	2.94×10^{-1}	1.19
`p p > z j j`	3	2.43×10^{-1}	0.56	4.41×10^{-1}	0.31
`p p > z j j j`	4	4.08×10^{-1}	0.27	5.88×10^{-1}	0.19
`p p > a j`	2	3.12×10^{-1}	5.33	2.94×10^{-1}	5.66
`p p > a j j`	3	3.28×10^{-1}	0.85	4.41×10^{-1}	0.63
`p p > w+ w- wpm`	3	1.00×10^{-3}	610.69	4.41×10^{-1}	1.39
`p p > z w+ w-`	3	8.00×10^{-3}	92.39	4.41×10^{-1}	1.68
`p p > z z wpm`	3	1.00×10^{-2}	85.00	4.41×10^{-1}	1.93
`p p > z z z`	3	1.00×10^{-3}	302.75	4.41×10^{-1}	0.69
`p p > a w+ w-`	3	1.90×10^{-2}	42.33	4.41×10^{-1}	1.82
`p p > a a wpm`	3	4.40×10^{-2}	47.24	4.41×10^{-1}	4.72
`p p > a z wpm`	3	1.00×10^{-3}	1244.49	4.41×10^{-1}	2.82
`p p > a z z`	3	2.00×10^{-2}	17.24	4.41×10^{-1}	0.78

Tilman Plehn's 'reference process' method

Surviving tails

Process	n_{part}	$\Delta\sigma/\sigma_0$	$\frac{\sigma_{\text{NLO}}-\sigma_0}{\Delta\sigma}$	$\Delta\sigma_{\text{ref}}/\sigma_0$	$\frac{\sigma_{\text{NLO}}-\sigma_0}{\Delta\sigma_{\text{ref}}}$
`p p > h`	1	3.48×10^{-1}	3.02	1.47×10^{-1}	7.15

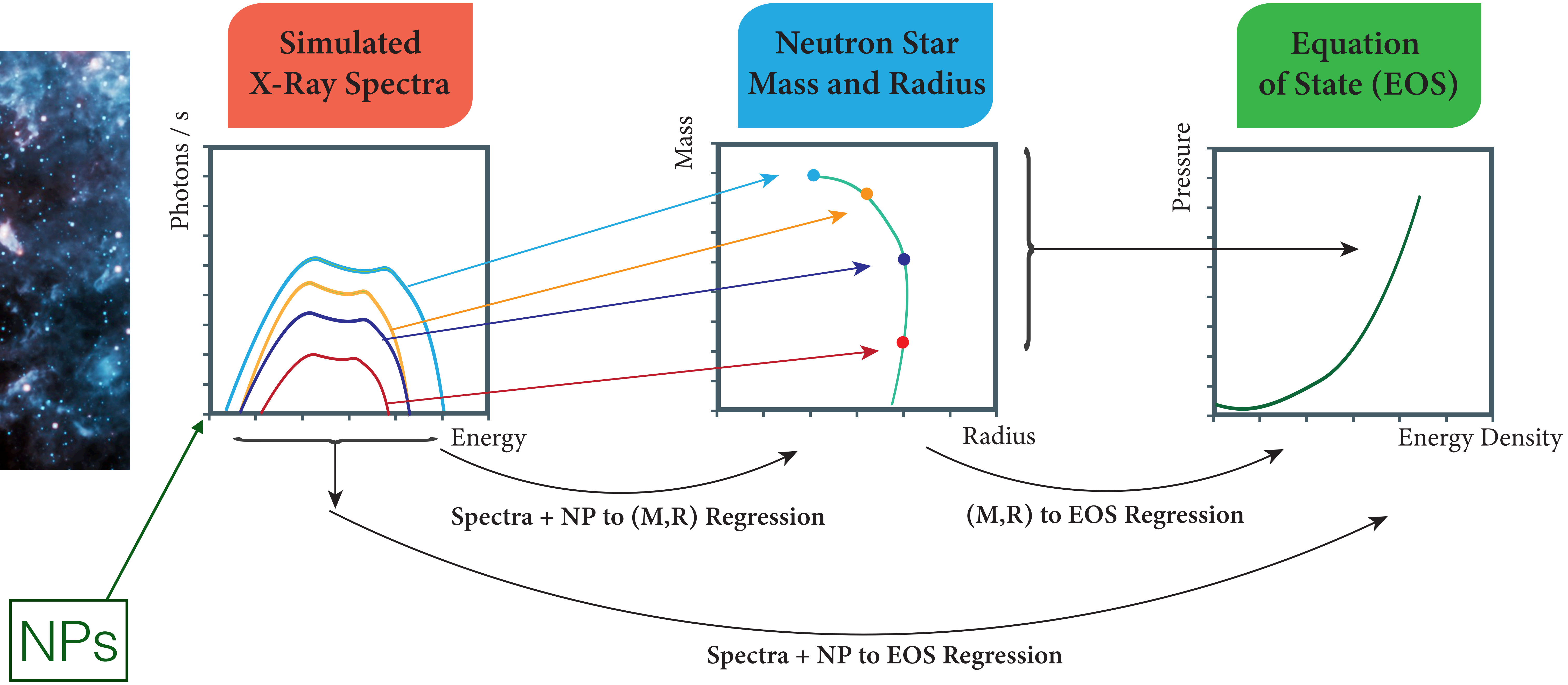
Large corrections loop-induced 2->1 process

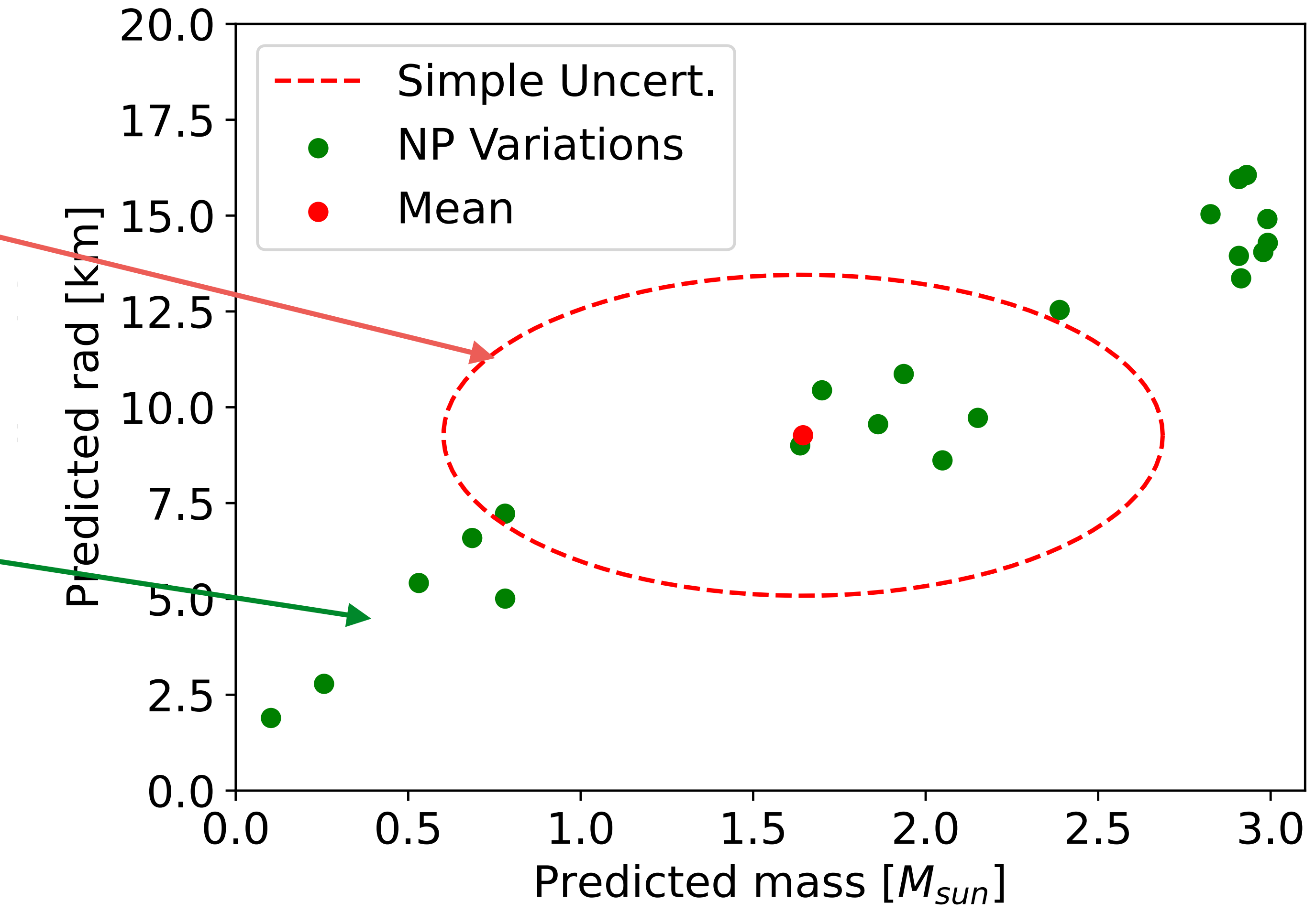
An application in astrophysics

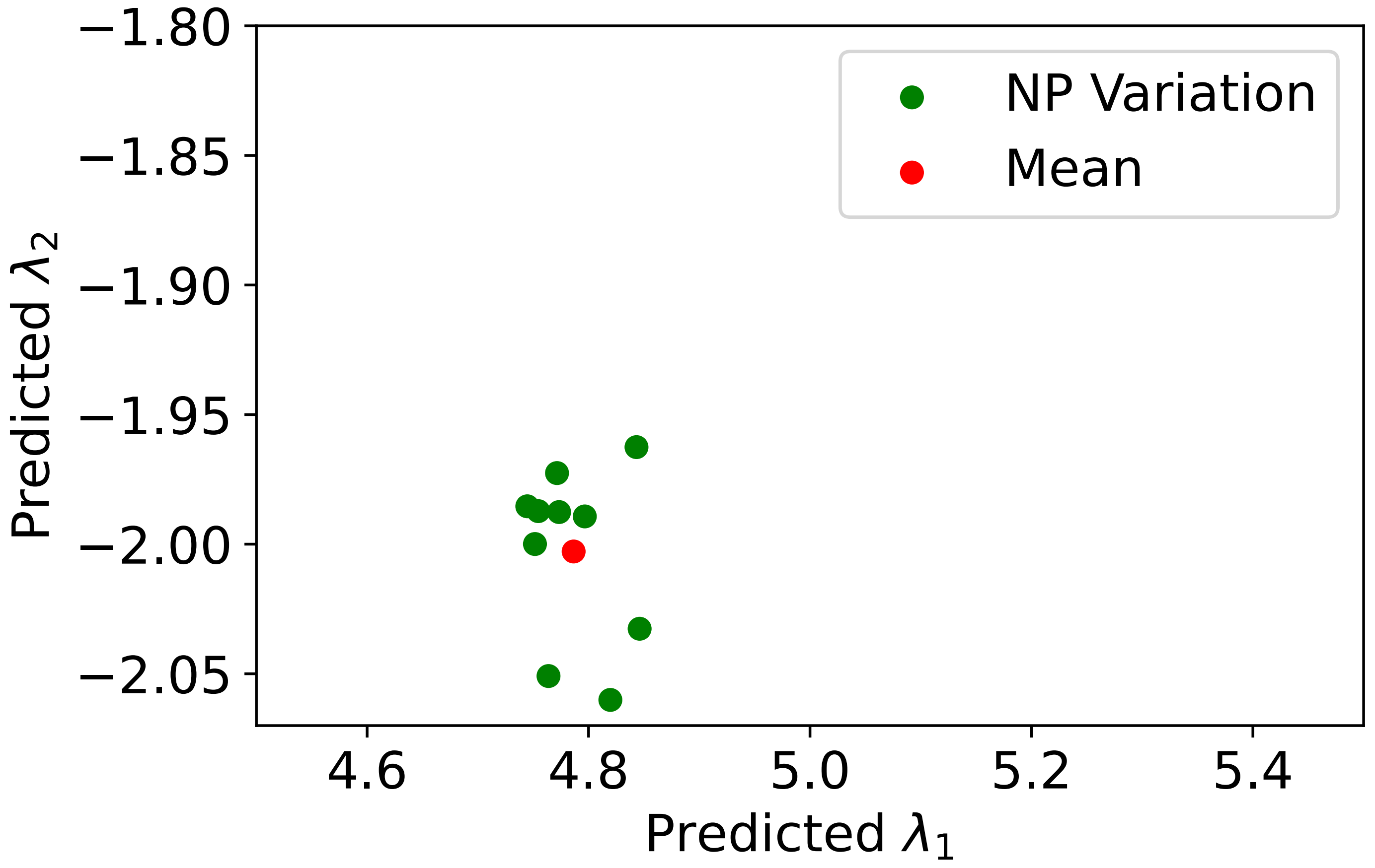
Application in Astrophysics: Full propagation of uncertainties

SOTA made a single point estimate + assumed uncorrelated Gaussian uncertainties

Real uncertainties look quite different

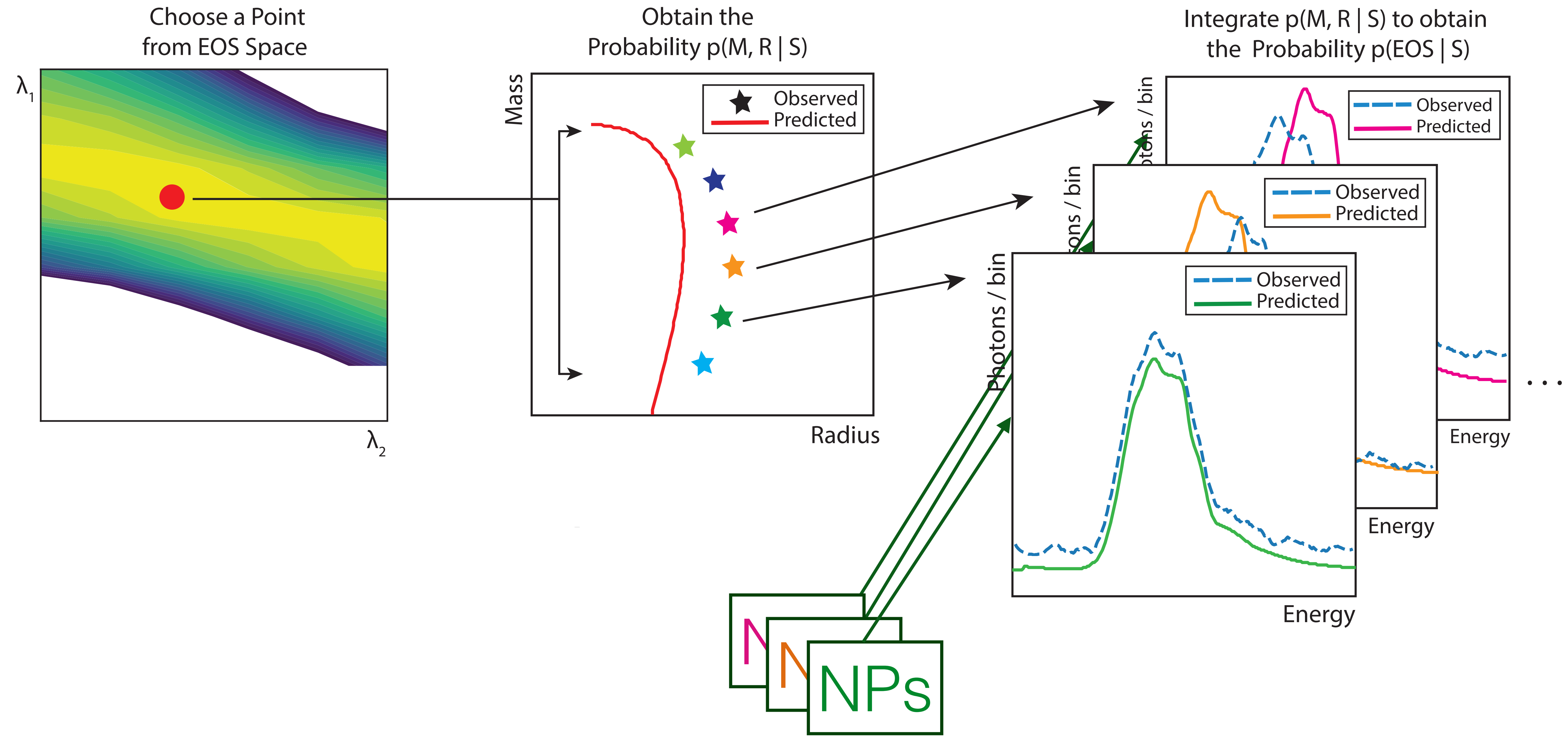






JCAP.020P.0922: Delaney Farrell, Pierre Baldi, Jordan Ott, **Aishik Ghosh**, Andrew W. Steiner, Atharva Kavitkar, Lee Lindblom, Daniel Whiteson, Fridolin Weber

Learn forward process to access the likelihood



Deploy with ONNX Runtime to compute likelihoods on-the-fly

ONNX RUNTIME

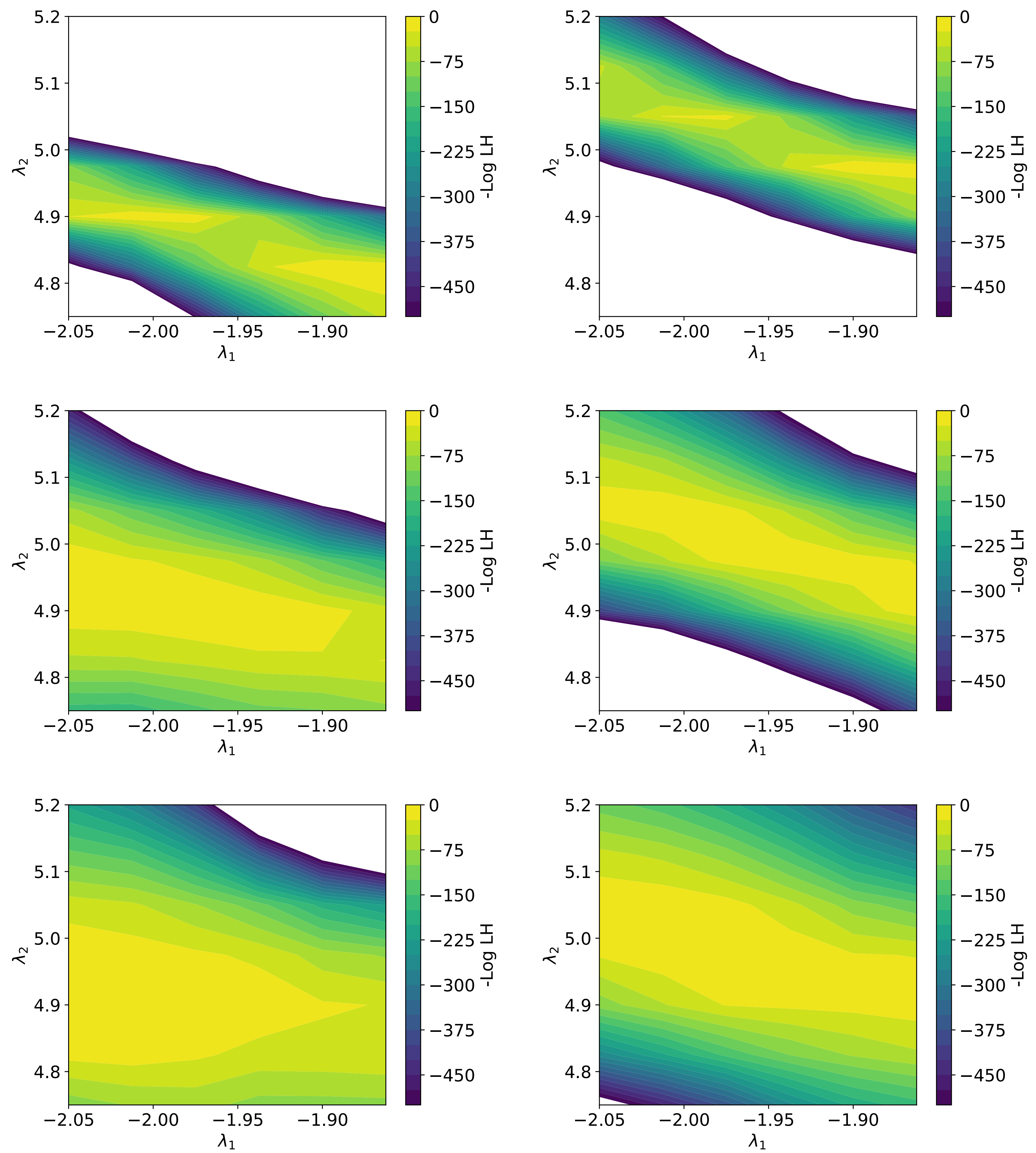
Nuisance Priors:

EOS parameter likelihoods:

True:

Tight:

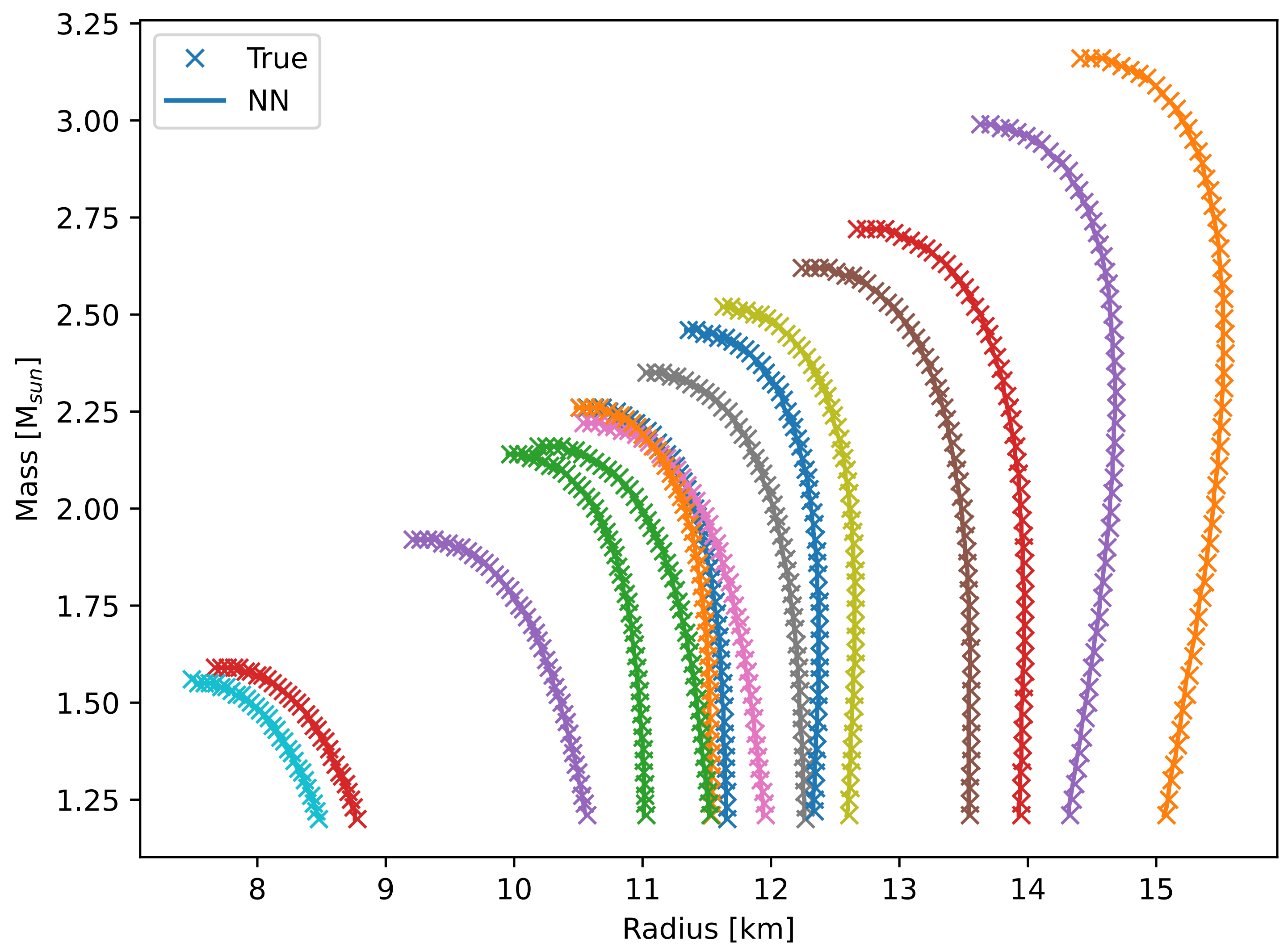
Loose:



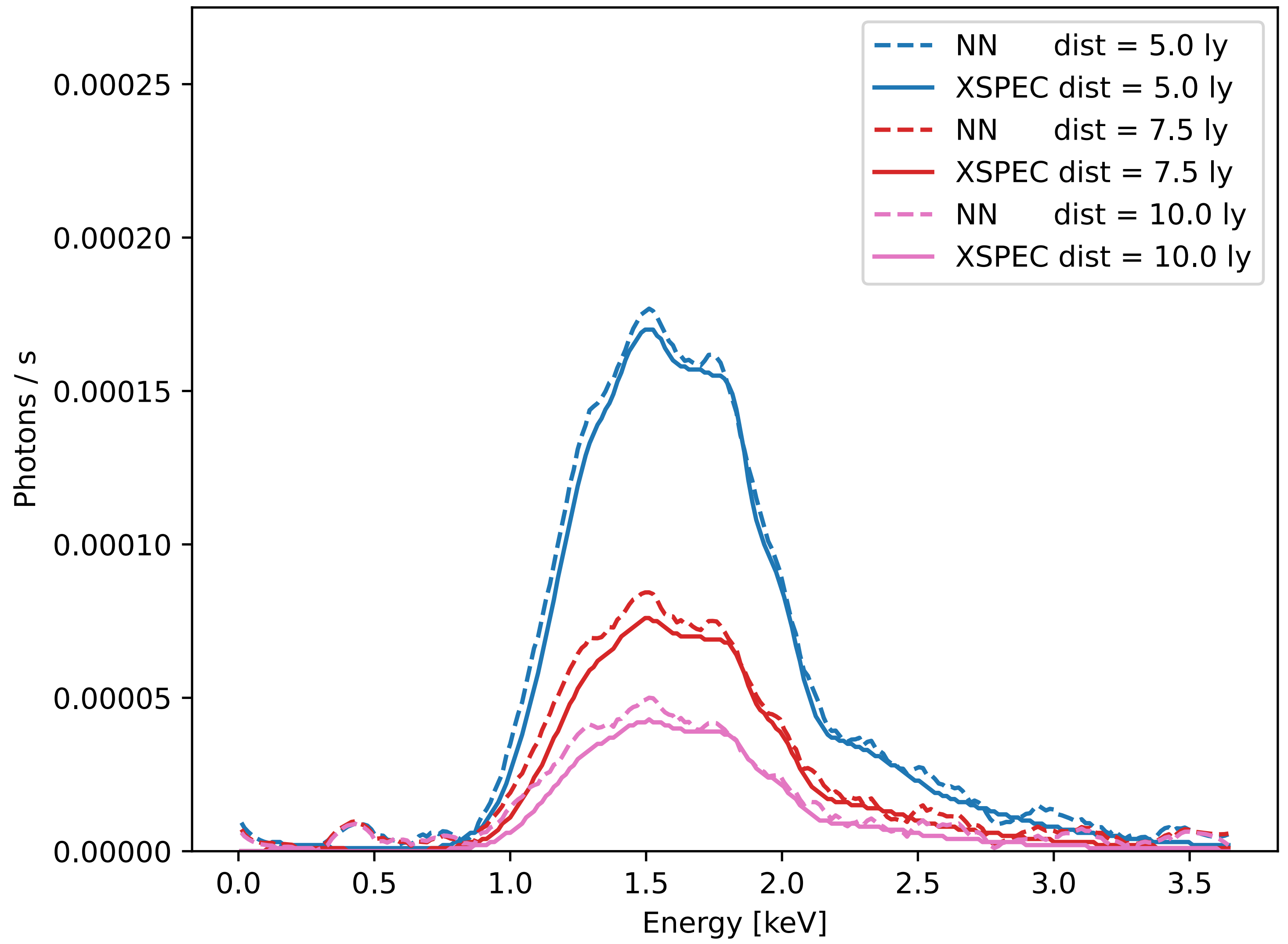
arXiv:2305.07442: Delaney Farrell, Pierre Baldi, Jordan Ott, **Aishik Ghosh**, Andrew W. Steiner, Atharva Kavitkar, Lee Lindblom, Daniel Whiteson, Fridolin Weber

Forward process step-by-step

Intermediate steps remain interpretable physical quantities



Learn EOS to M-R



Learn {M,R,NPs} to Spectrum

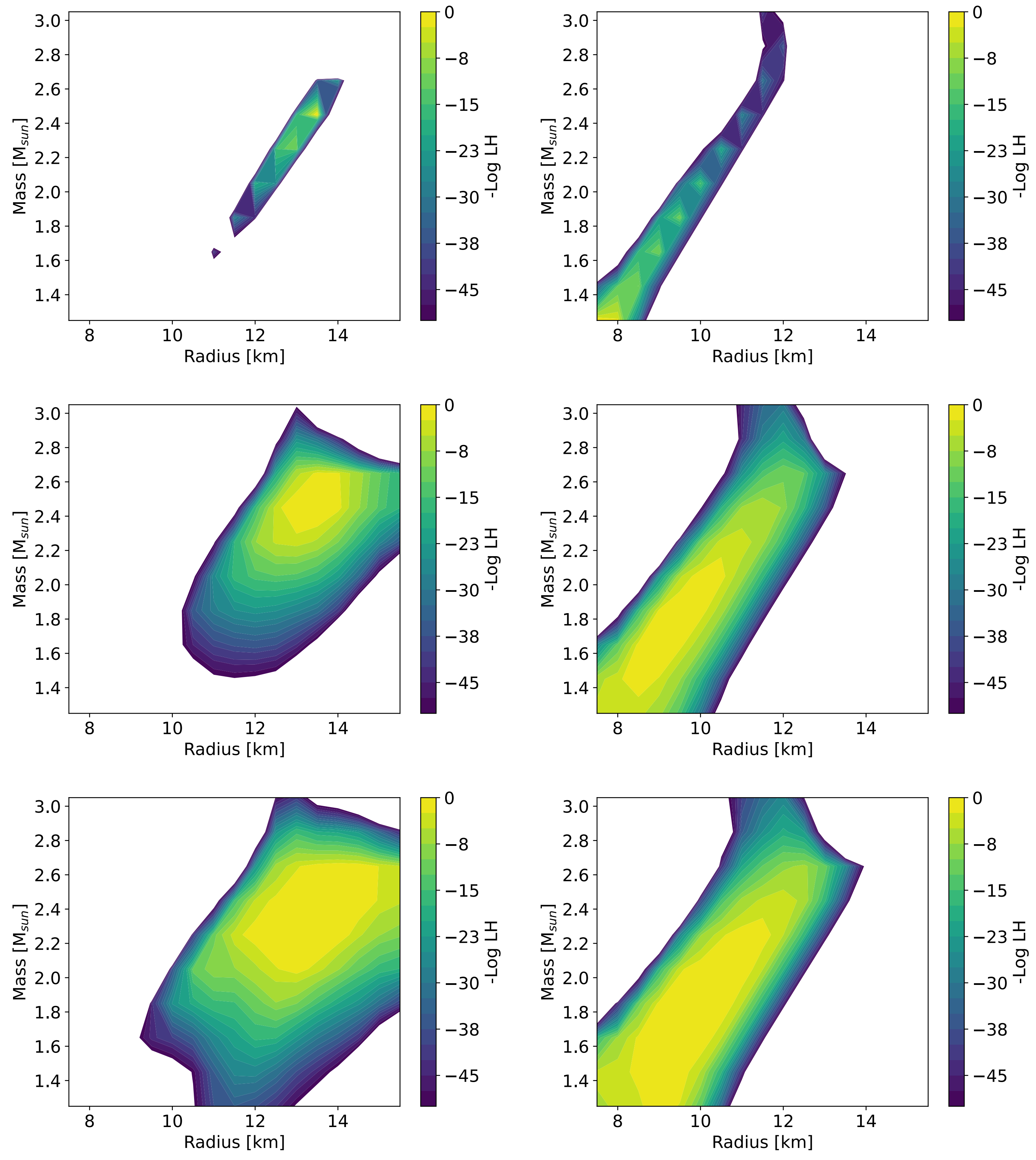
<u>Nuisance Priors:</u>

<u>M-R likelihoods:</u>

True:

Tight:

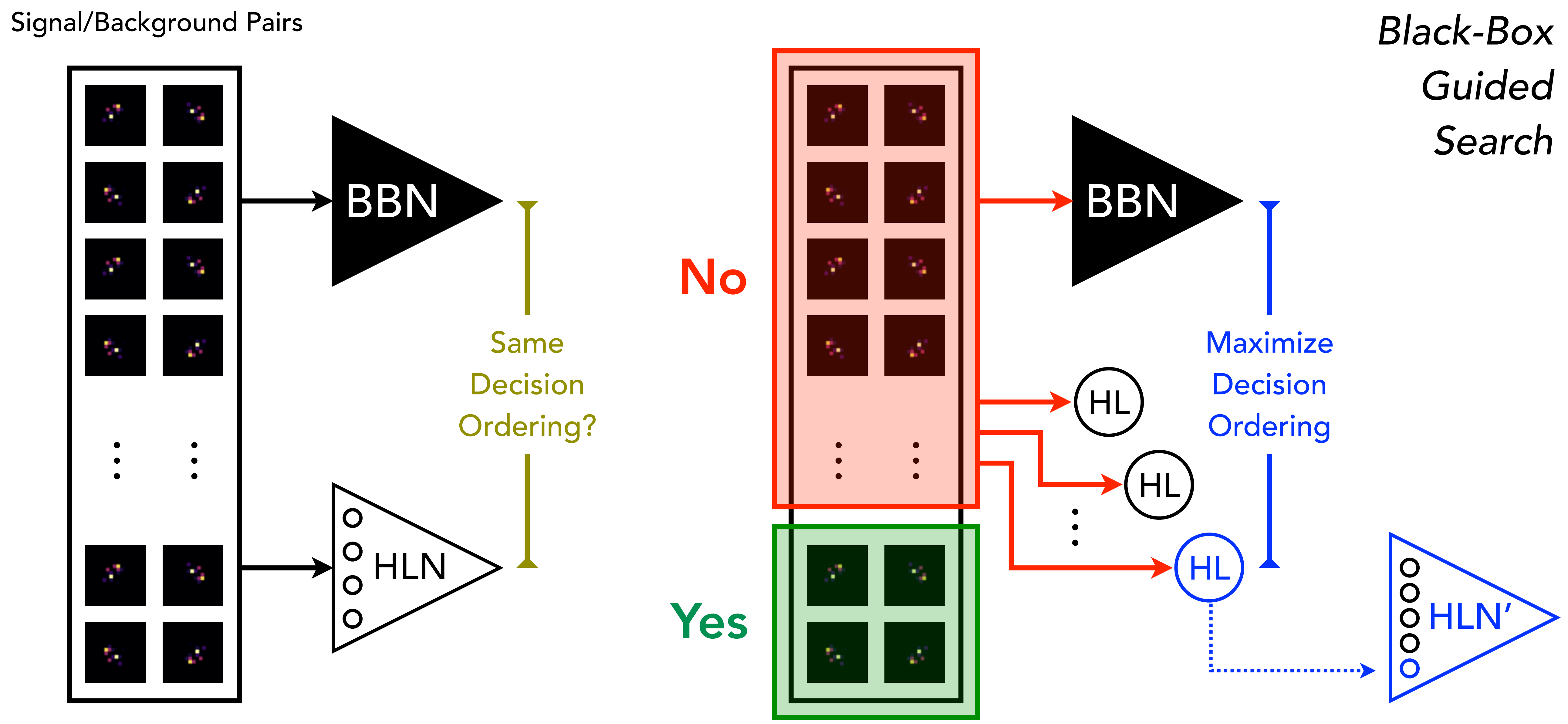
Loose:



[arXiv:2305.07442](https://arxiv.org/abs/2305.07442): Delaney Farrell, Pierre Baldi, Jordan Ott, **Aishik Ghosh**, Andrew W. Steiner, Atharva Kavitkar, Lee Lindblom, Daniel Whiteson, Fridolin Weber

New ML tools

Mapping machine-learned physics into a human-readable space



Rank	EFP	κ	β	Chrom #	$\mathrm{ADO}[\mathrm{EFP},\mathrm{CNN}]_{X_6}$	AUC[EFP]	$\mathrm{ADO}[\mathrm{6HL}+\mathrm{EFP},\mathrm{CNN}]_{X_{\mathrm{all}}}$	AUC[6HL + EFP]
1		2	$\frac{1}{2}$	3	0.6207	0.8031	0.9714	0.9528 ± 0.0003
2		2	$\frac{1}{2}$	3	0.6205	0.8203	0.9714	0.9524
3		0	–	1	0.6205	0.6737	0.9715	0.9525
4		2	$\frac{1}{2}$	3	0.6199	0.8301	0.9715	0.9527
5		2	$\frac{1}{2}$	3	0.6197	0.8290	0.9714	0.9527
6		2	$\frac{1}{2}$	3	0.6196	0.8251	0.9715	0.9522
7		0	$\frac{1}{2}$	2	0.6187	0.7511	0.9715	0.9526
8		2	$\frac{1}{2}$	3	0.6184	0.8257	0.9712	0.9527
9		2	$\frac{1}{2}$	3	0.6182	0.8090	0.9714	0.9527
10		2	$\frac{1}{2}$	3	0.6180	0.8314	0.9714	0.9526
60		0	1	2	0.6163	0.7194	0.9715	0.9525
341		-1	$\frac{1}{2}$	4	0.6142	0.6286	0.9714	0.9509
589		0	2	2	0.6109	0.7579	0.9714	0.9523
3106		-1	–	1	0.5891	0.5882	0.9714	0.9510
3519		$\frac{1}{2}$	$\frac{1}{2}$	2	0.5664	0.7698	0.9715	0.9524
3521		$\frac{1}{2}$	–	1	0.5663	0.7093	0.9714	0.9522
5531		1	2	1	0.5290	0.7454	0.9714	0.9507
5554		1	$\frac{1}{2}$	2	0.5279	0.8210	0.9713	0.9505
5610		2	–	1	0.5245	0.7117	0.9714	0.9507
5657		1	1	3	0.5224	0.8257	0.9712	0.9506
5793		1	1	2	0.5191	0.8640	0.9714	0.9505
6052		1	2	3	0.5153	0.8500	0.9716	0.9504
7438		1	2	2	0.5011	0.8835	0.9716	0.9506

[Faucett et al.]

Differentiable Programming: Optimise your final objective directly

Simpson et al.

Following Inferno [de Castro et al.]

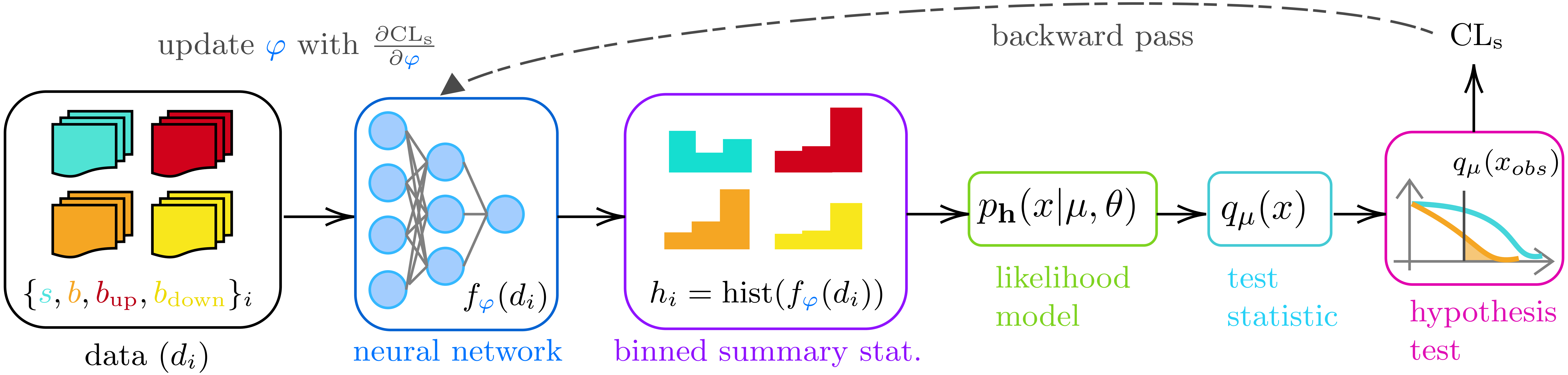


Figure 1. The pipeline for **`neos`**. The dashed line indicating the backward pass involves updating the weights φ of the neural network via gradient descent.