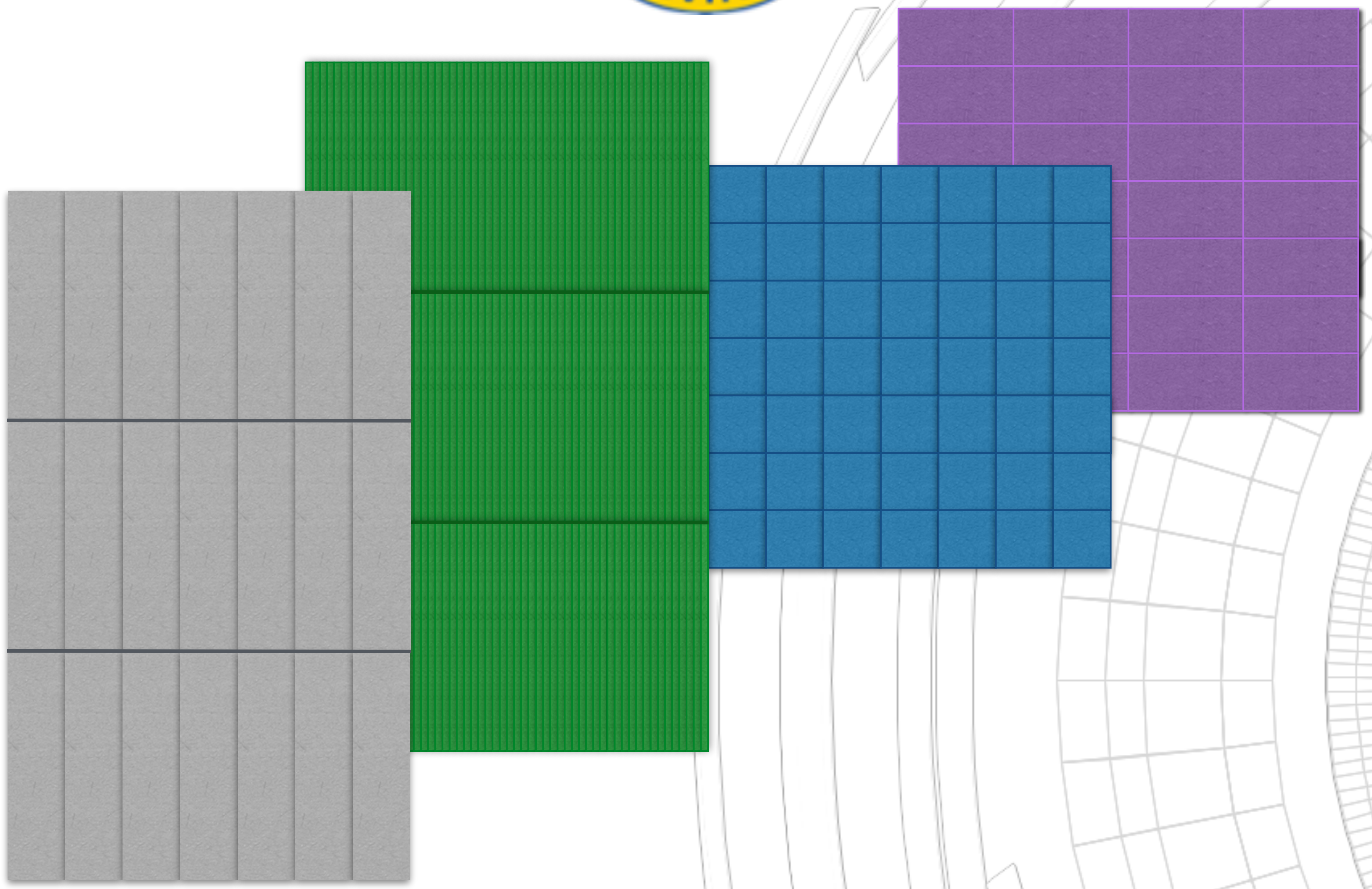
Uncertainties in the era of ML
(For Particle and Astrophysics)

Aishik Ghosh

Aspen Centre for Physics

28 Mar 2023

BERKELEY LAB



Uncertainties, the bedrock of experimental science

Uncertainties, the bedrock of experimental science

mH = 125.25 ± 0.17 GeV

Uncertainties, the bedrock of experimental science

$$m_H = 125.25 \pm 0.17 \text{ GeV}$$

Uncertainties, the bedrock of experimental science

mH = 125.25 ± 0.17 GeV

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

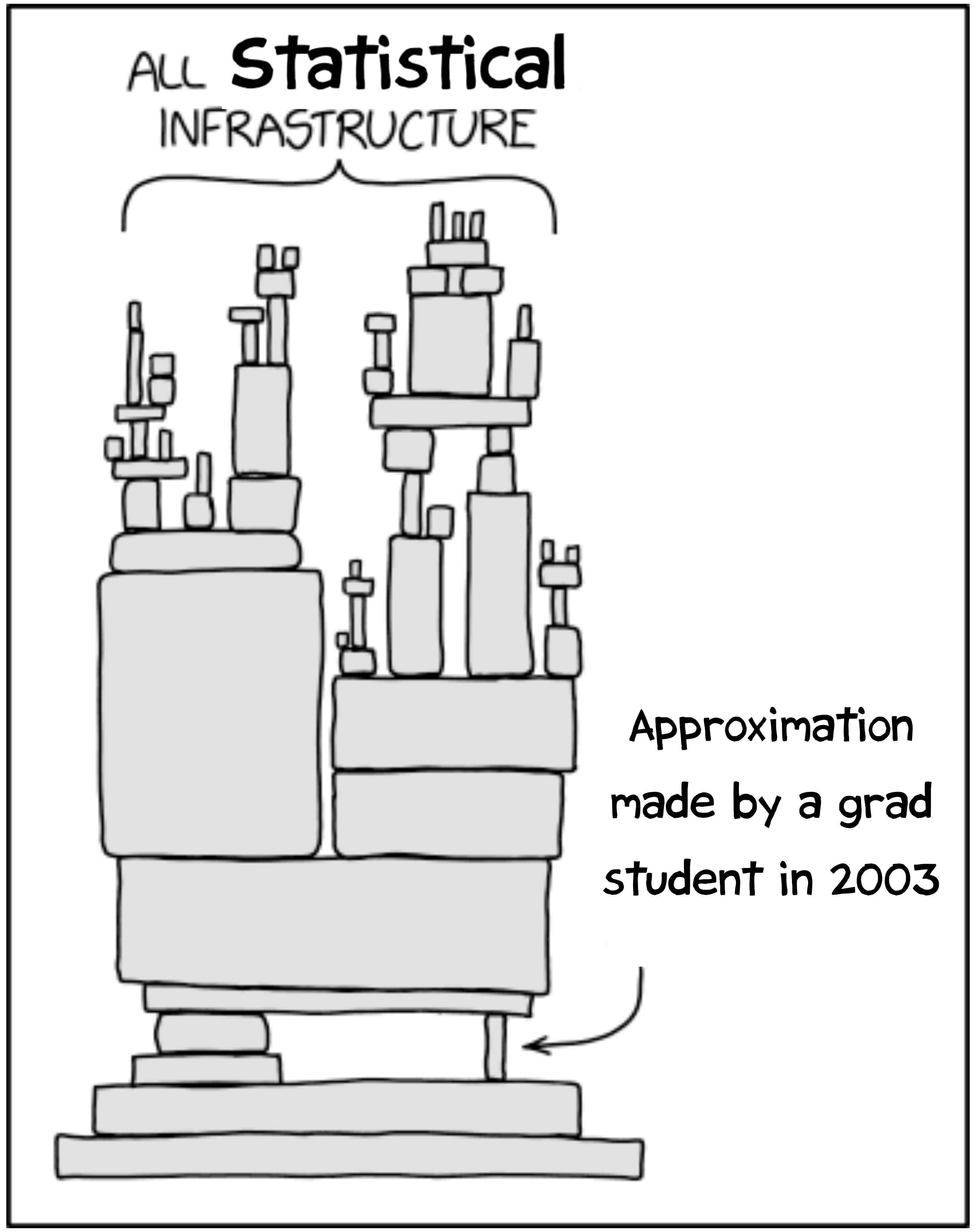
Uncertainties, the bedrock of experimental science

$$mH = 125.25 \pm 0.17 \text{ GeV}$$

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

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

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?

Outline: A predictable evolution over ten years

Fear: Will ML exacerbate uncertainties in a way human-designed strategies naturally avoid ?

Solution: Find ML equivalents of uncertainty mitigation tricks we implicitly use in classical methods. Understand good and bad ways to use ML

Opportunity: ML *for* uncertainty – Realising that ML unlocks completely new methods to tackle uncertainties in a way classical methods couldn't

Revolution: Novel ML uncertainty quantification & mitigation methods have wider applications, also back-ported to traditional algorithms

Outline: A predictable evolution over ten years

Fear: Will ML exacerbate uncertainties in a way human-designed strategies naturally avoid ?

Solution: Find ML equivalents of uncertainty mitigation tricks we implicitly use in classical methods. Understand good and bad ways to use ML

Opportunity: ML *for* uncertainty – Realising that ML unlocks completely new methods to tackle uncertainties in a way classical methods couldn't

We are here

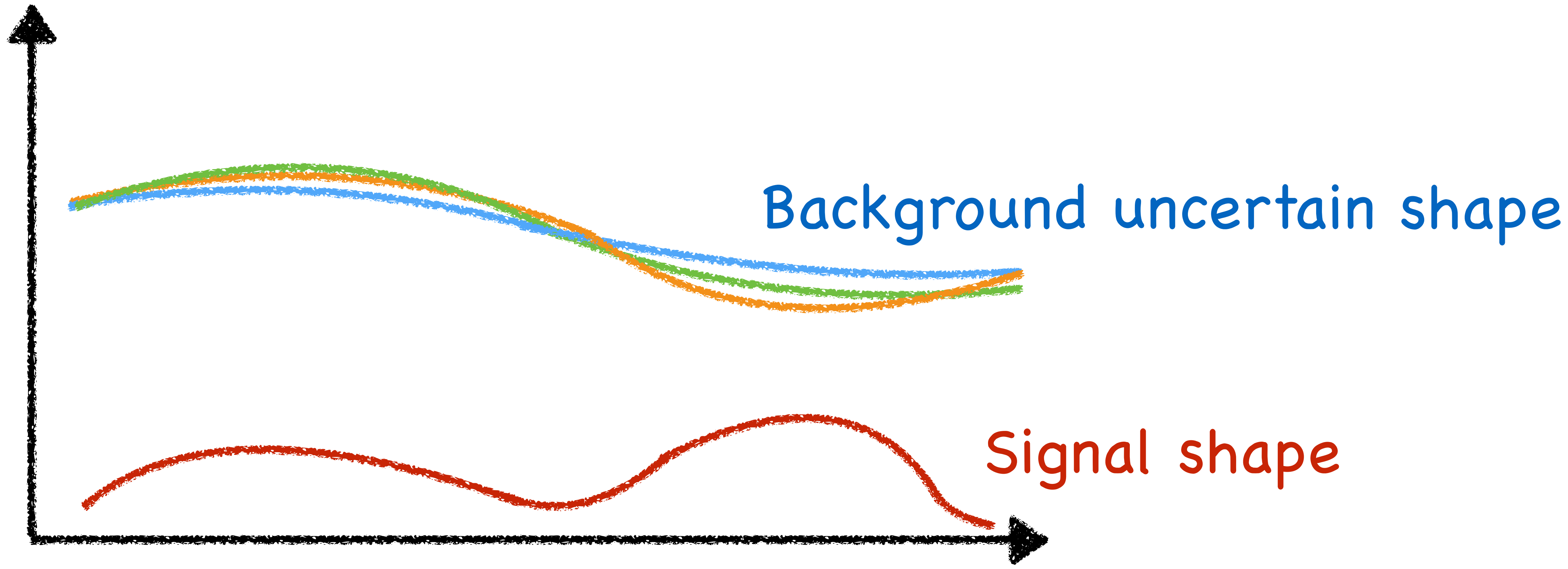
Revolution: Novel ML uncertainty quantification & mitigation methods have wider applications, also back-ported to traditional algorithms

Observable Sensitive to Nuisance Parameters

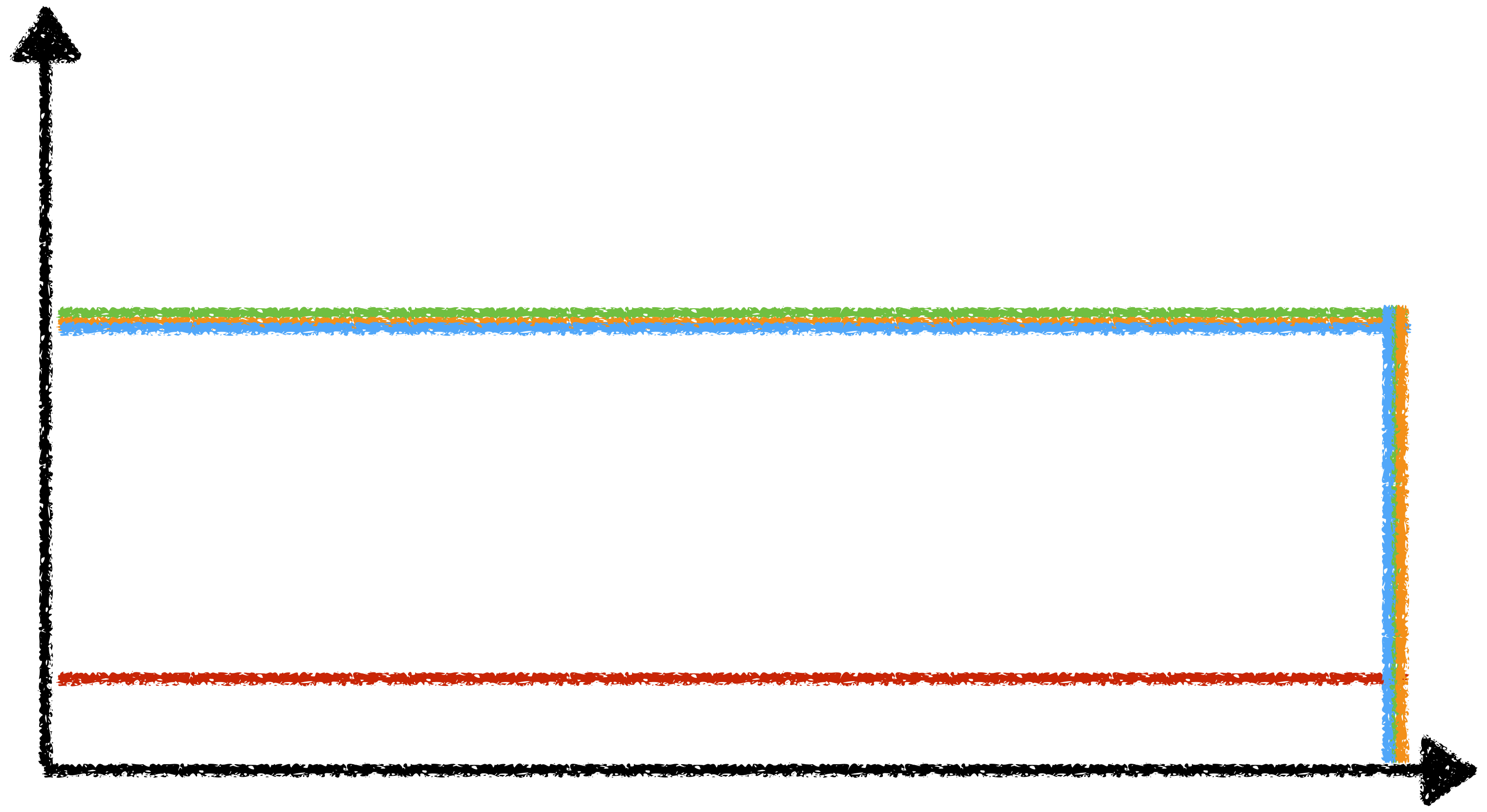
Traditionally, we reduce impact of NP by sacrificing something:

- 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

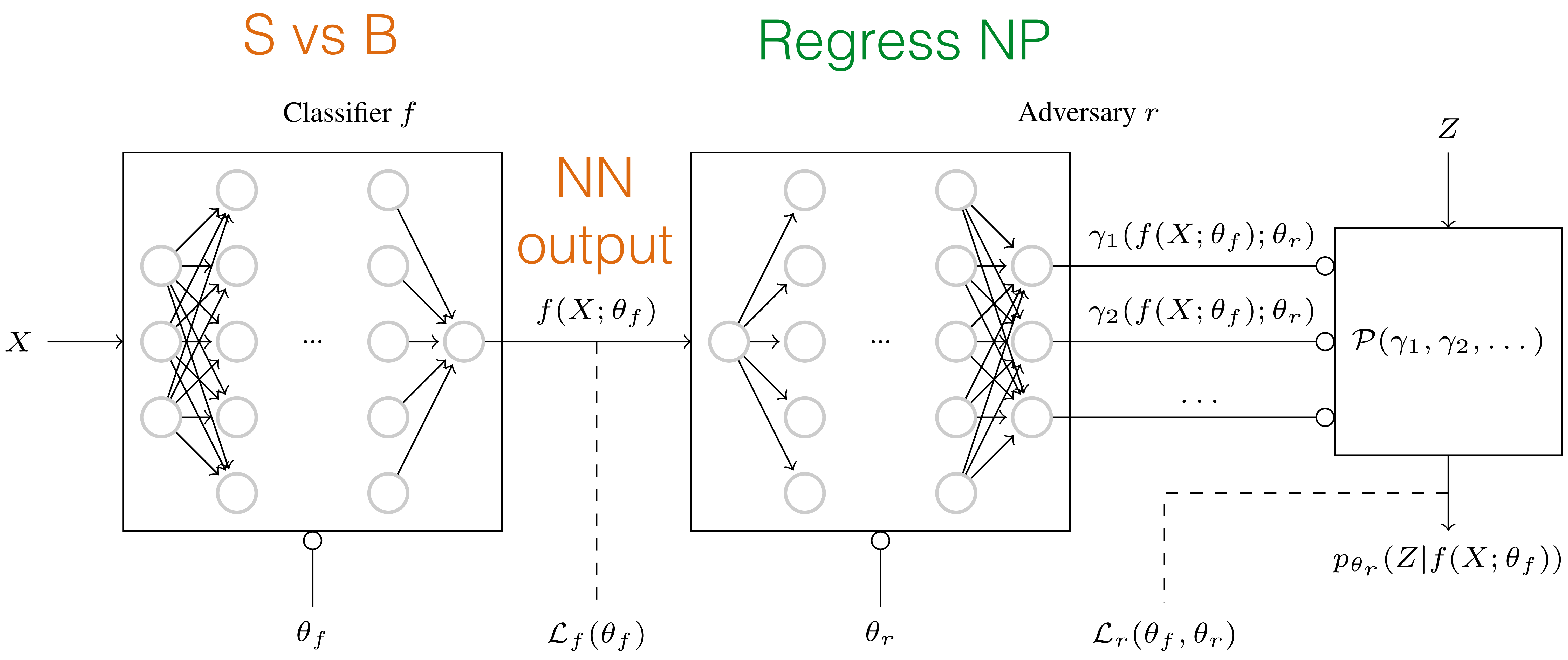


ML equivalent problem: Domain Adaptation

arXiv:1505.07818

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

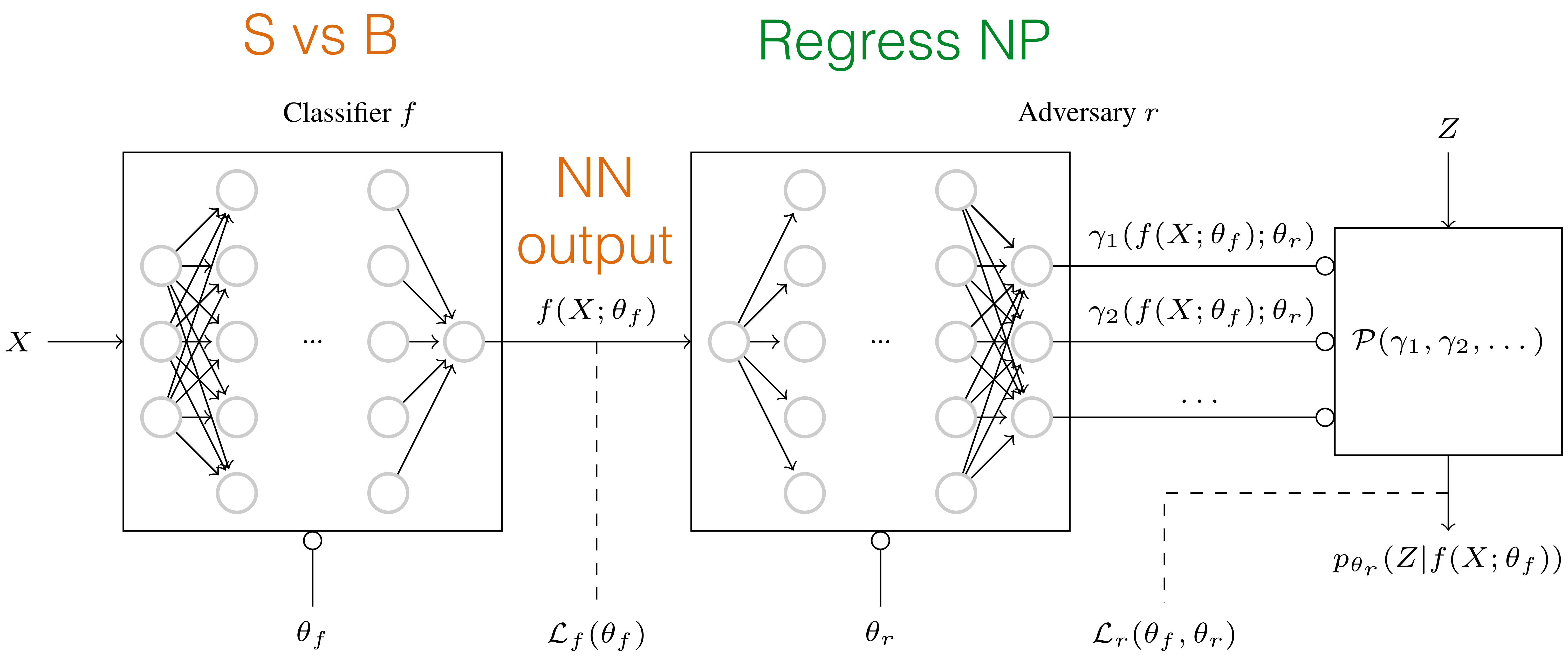
Adversarial decorrelation



Learning to Pivot, Louppe et al.

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

Adversarial decorrelation



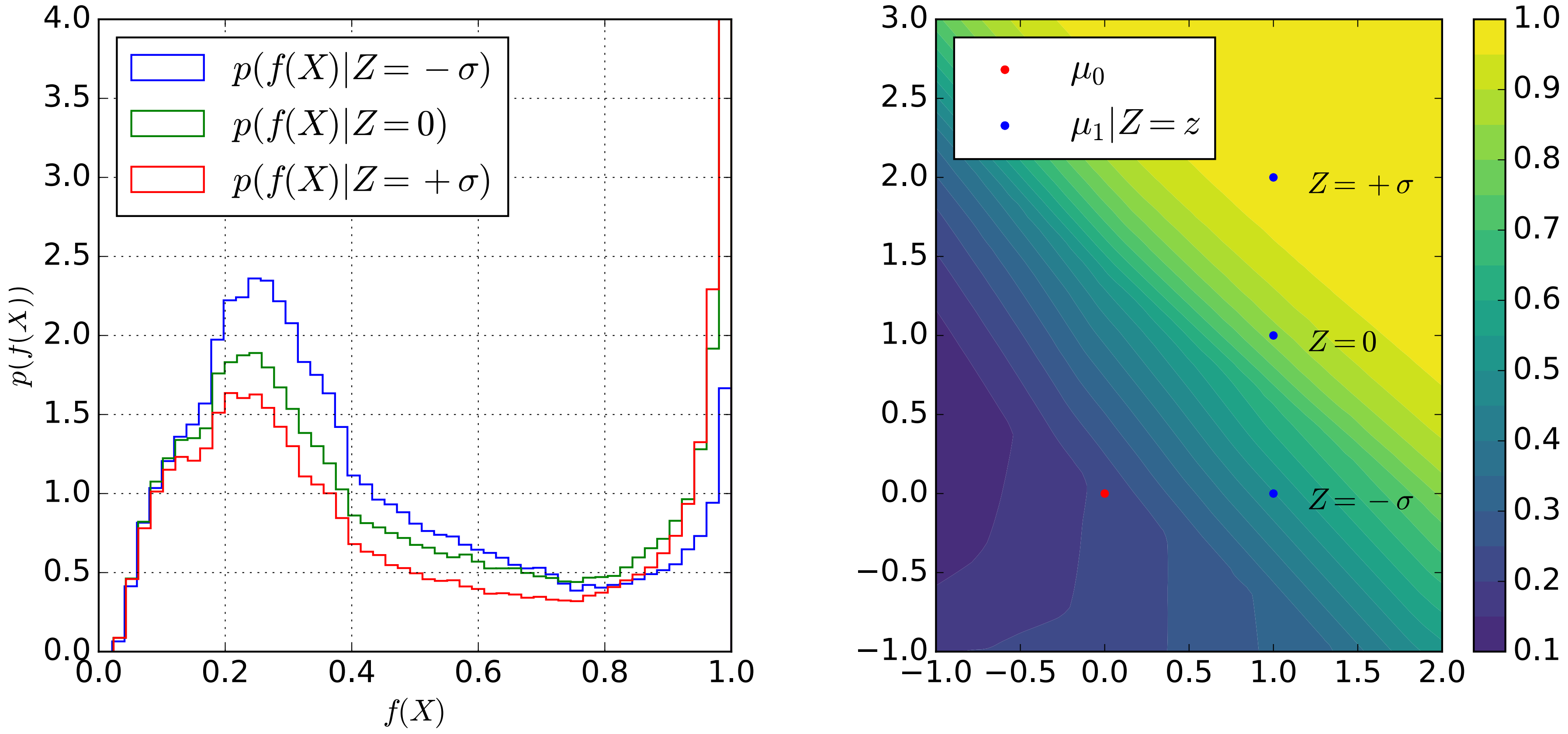
To fool the adversary, classifier output should be decorrelated to Z

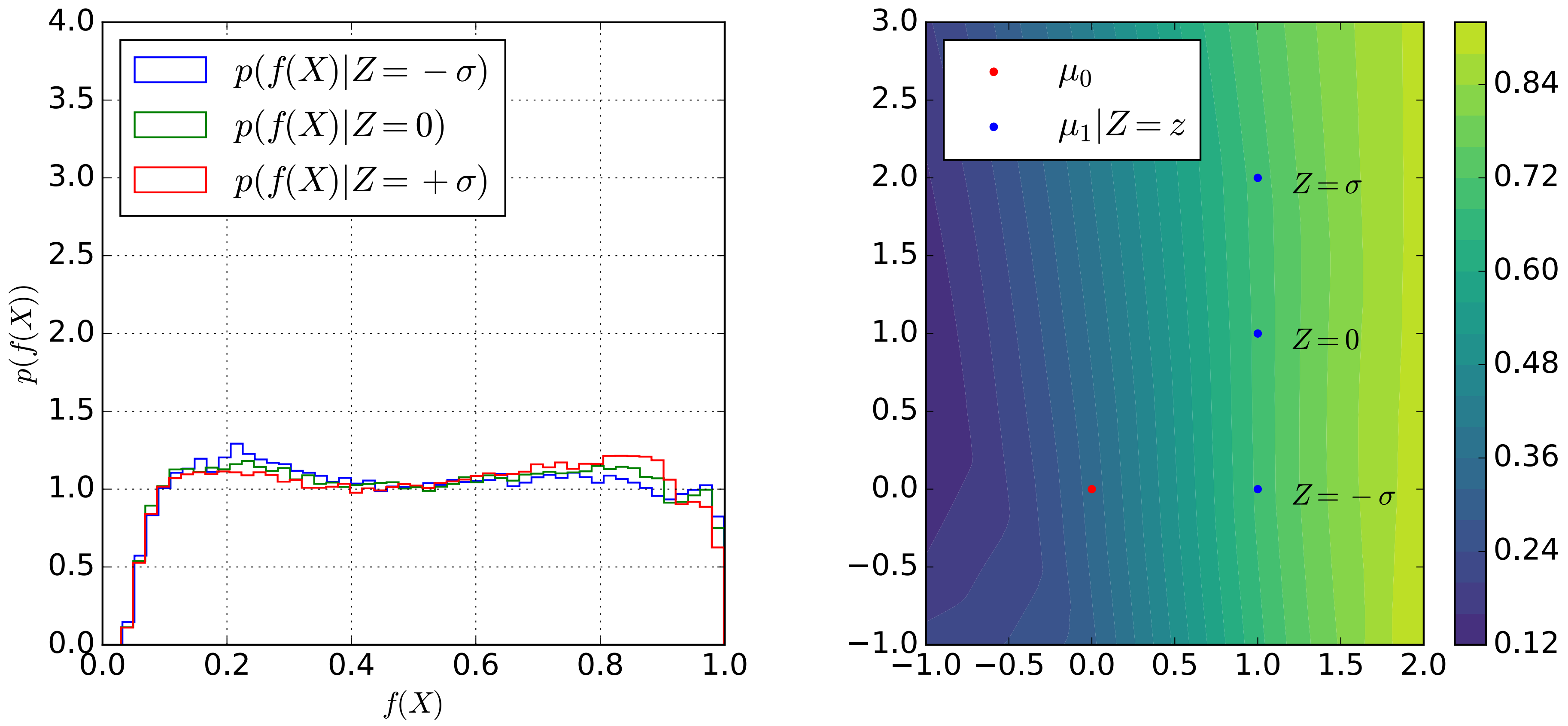
Learning to Pivot, Louppe et al.

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

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

ML-Decorrelation Methods



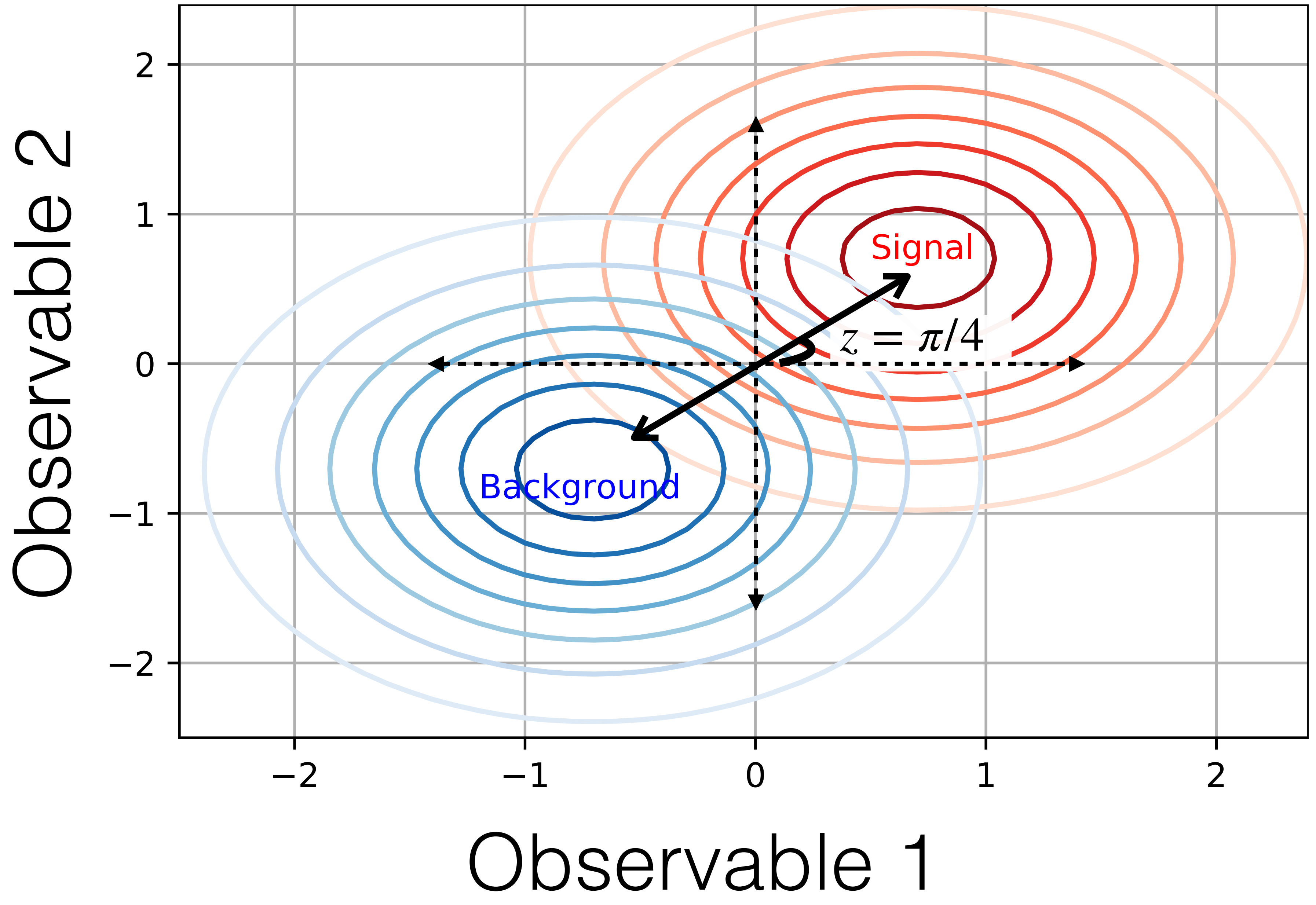


Classifier output for various values of Z

Learning to Pivot, Louppe et al.

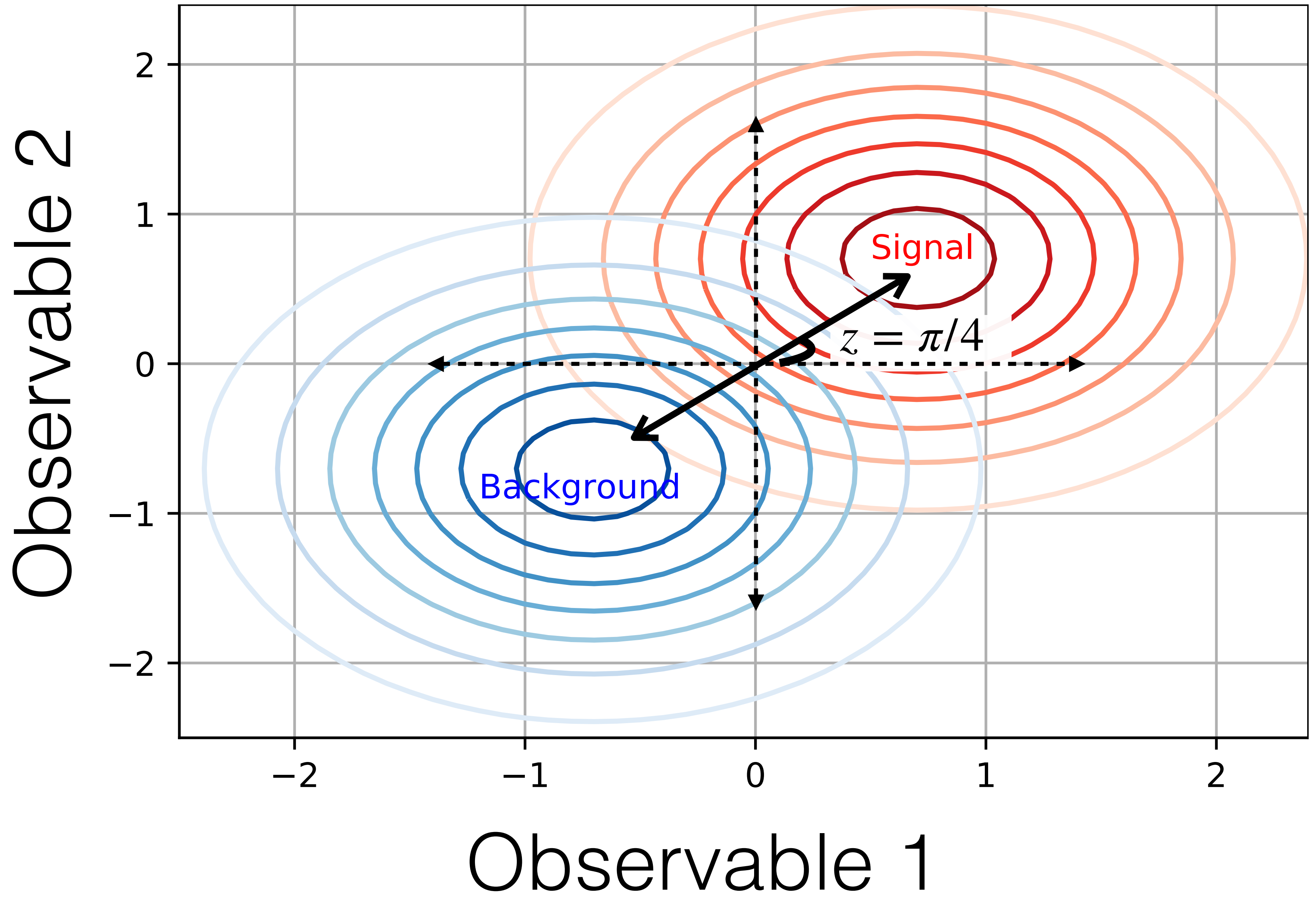
What if we could do better ?

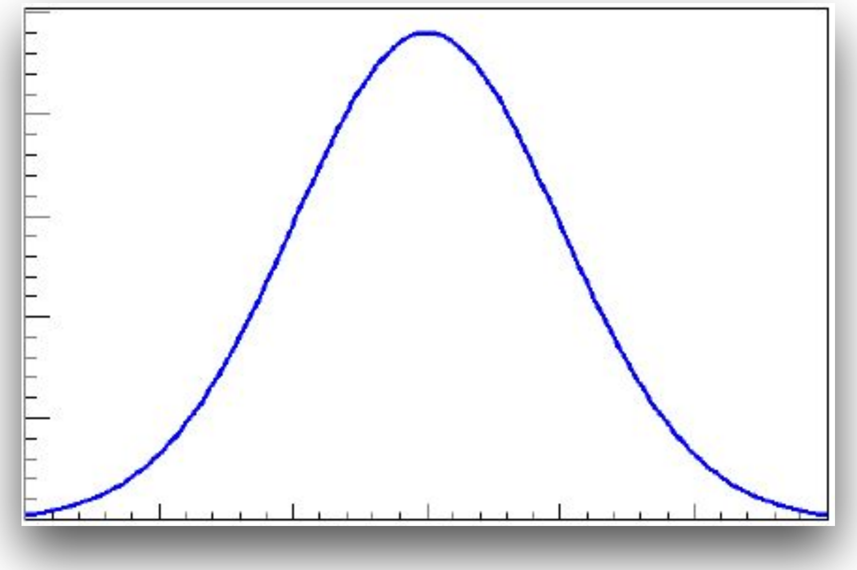
What if we could do better ?



z = Nuisance Parameter

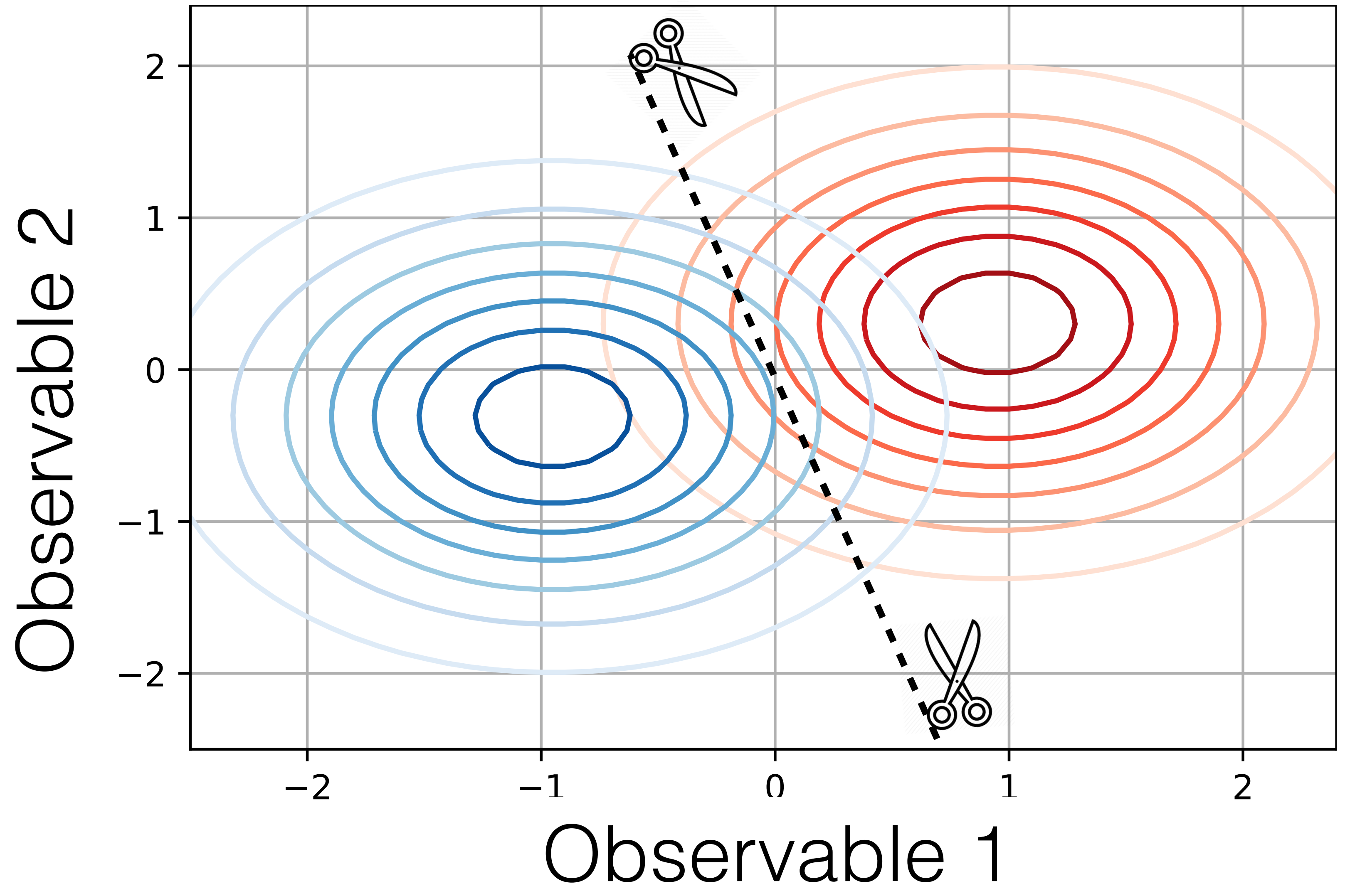
What if we could do better ?

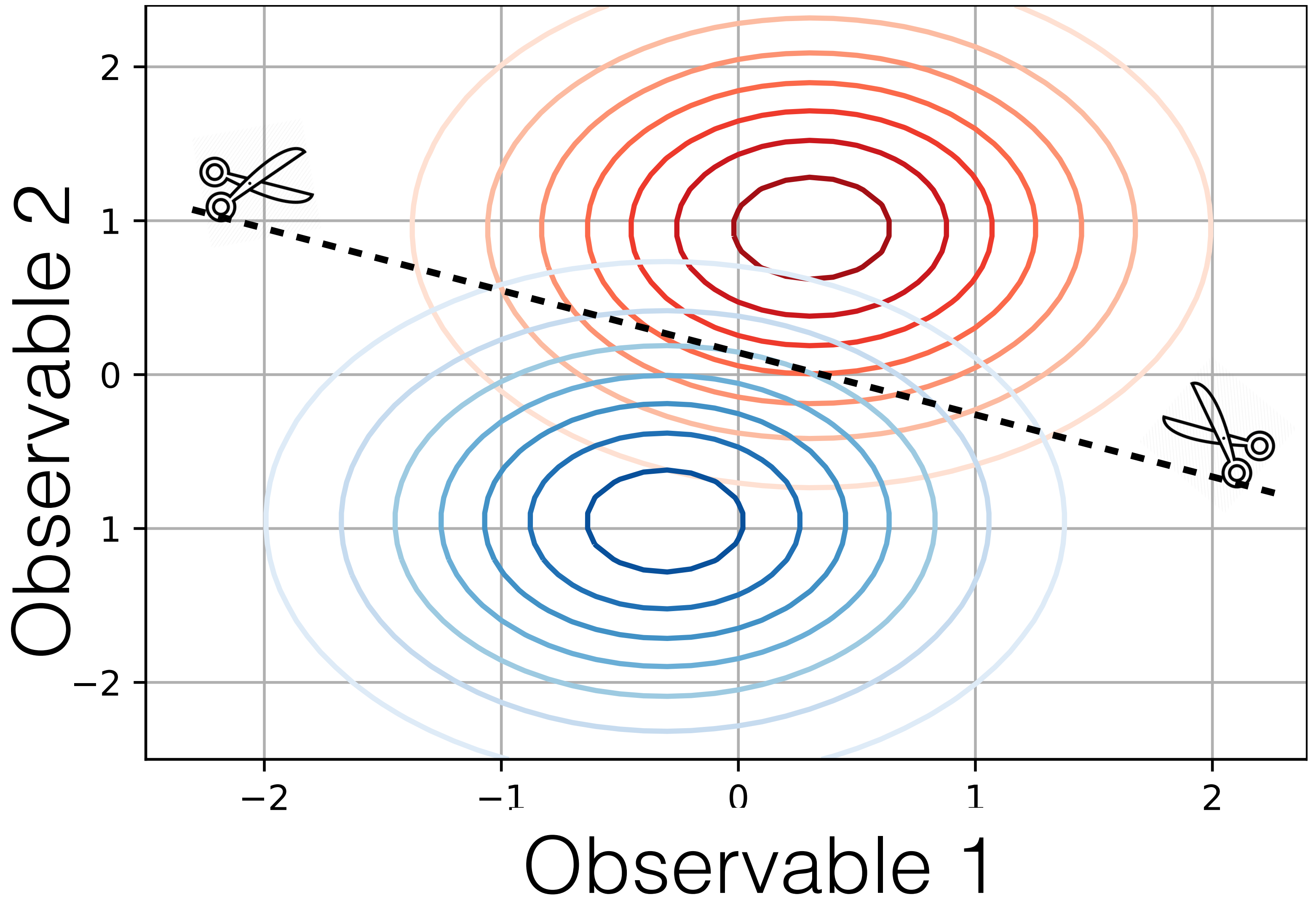




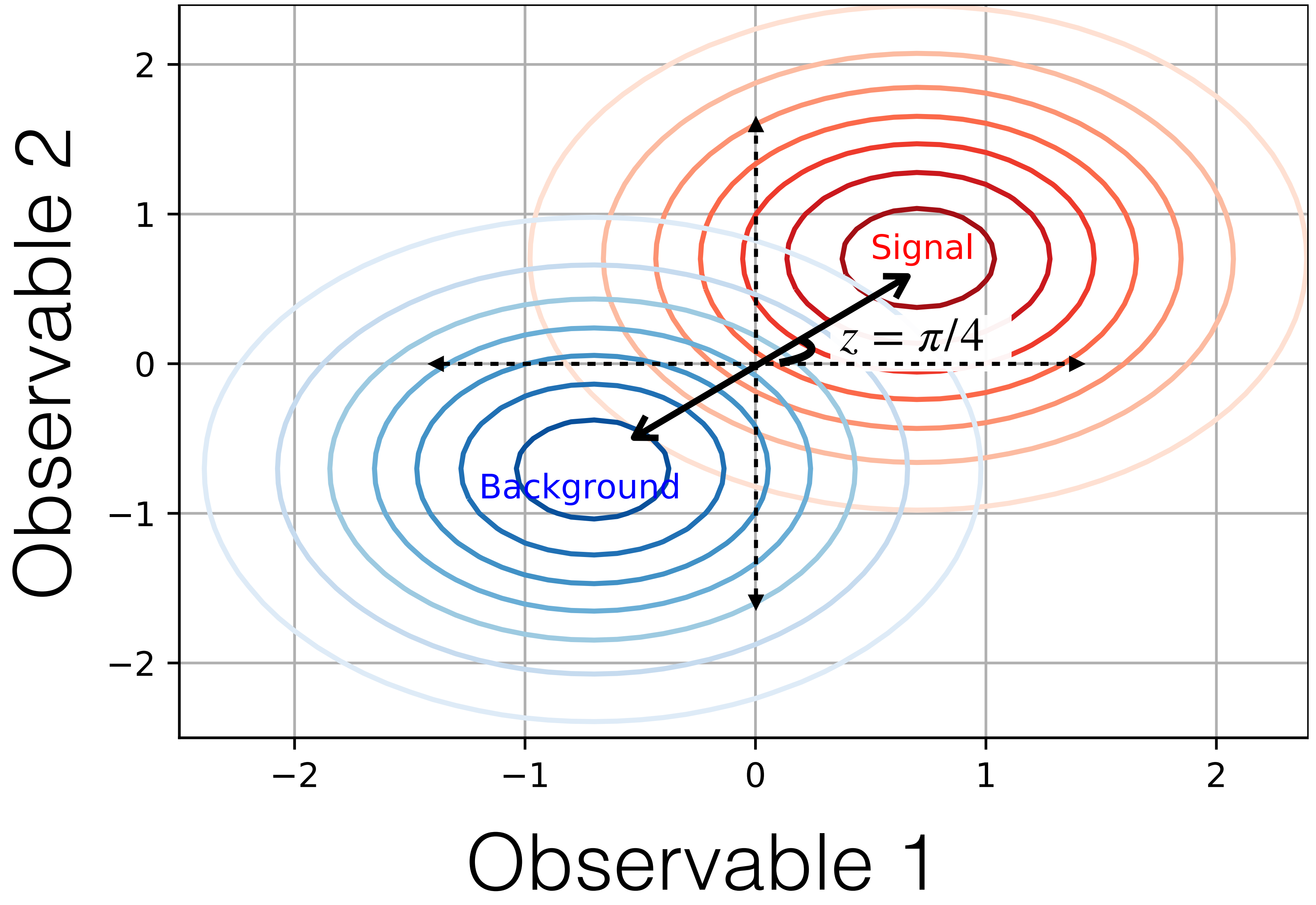
z = Nuisance Parameter

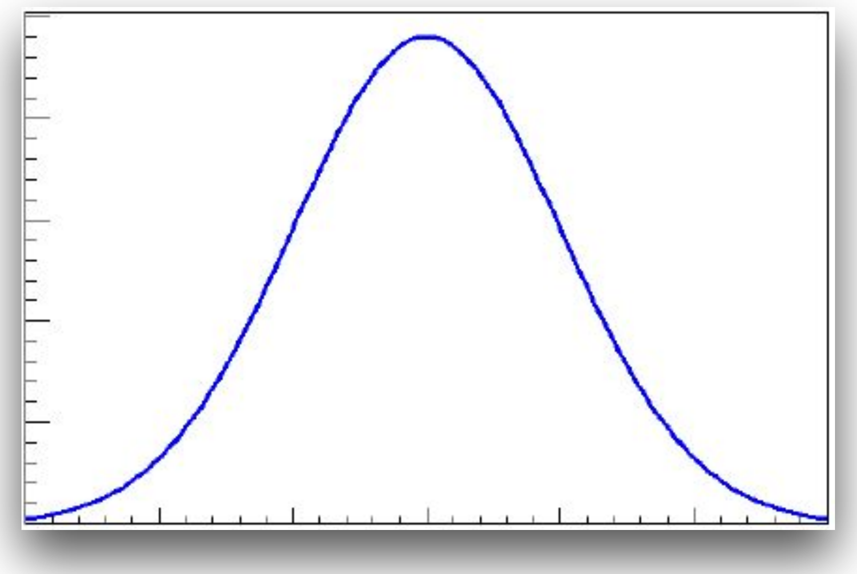
Prior





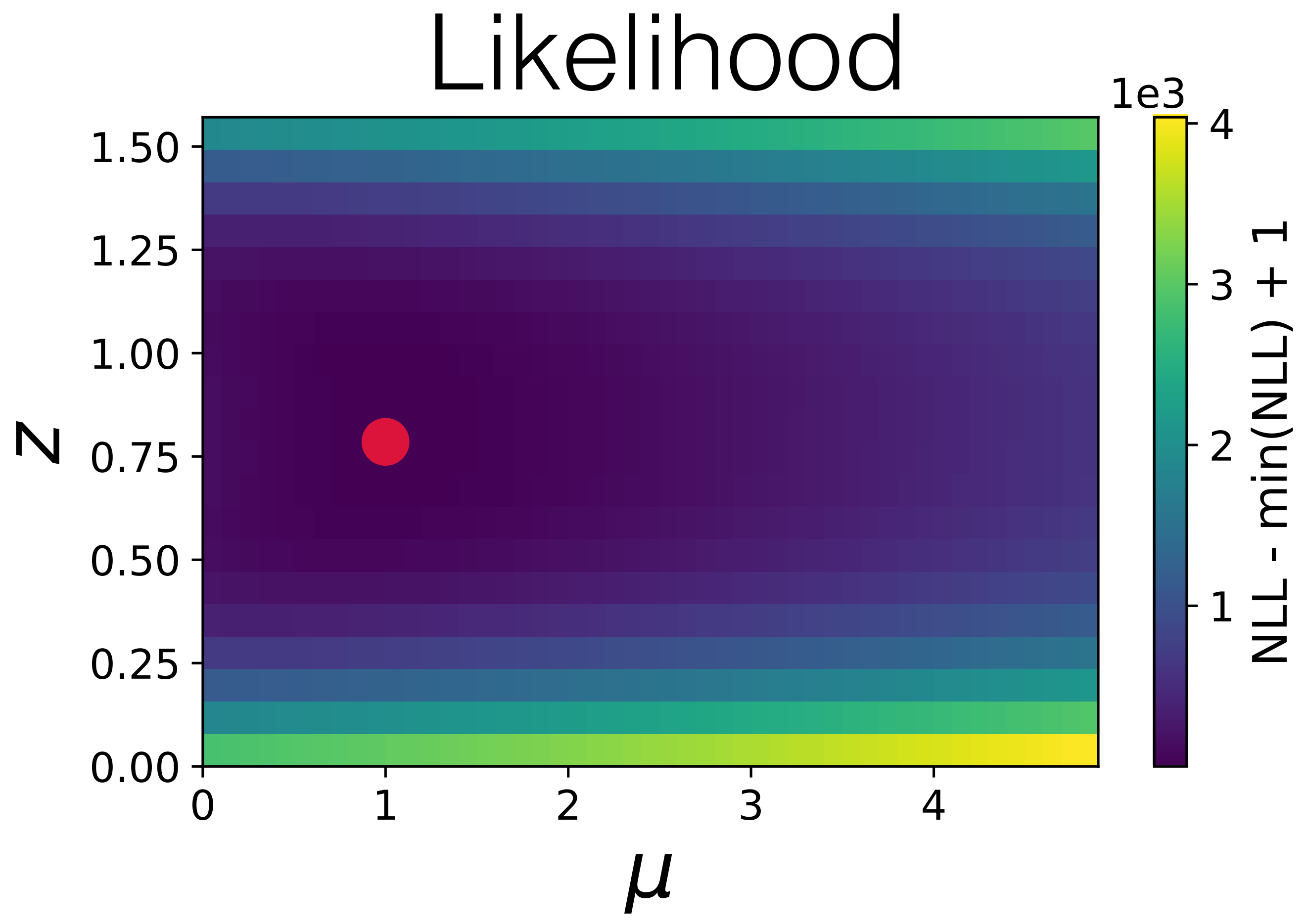
What if we could do better ?

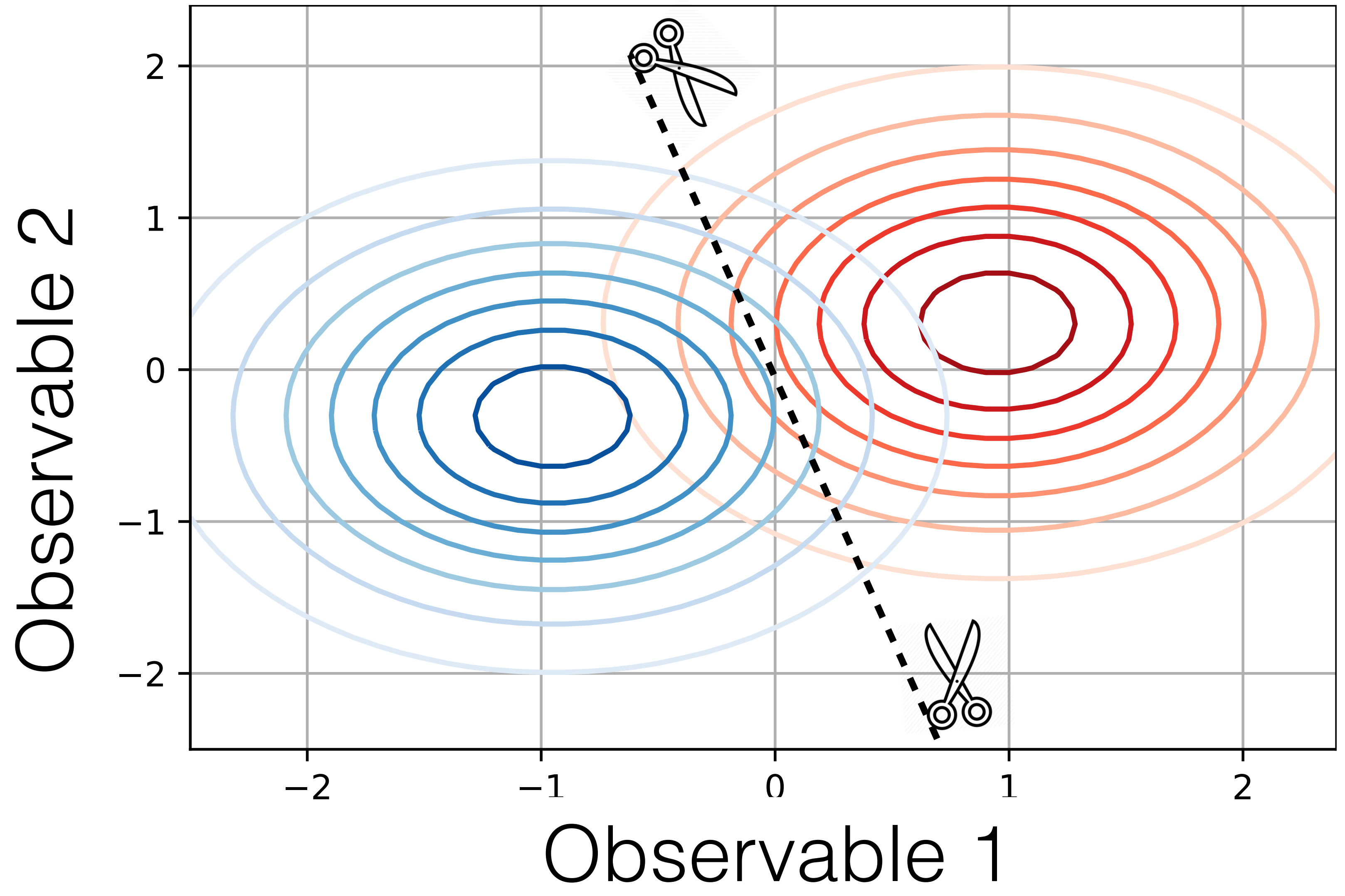


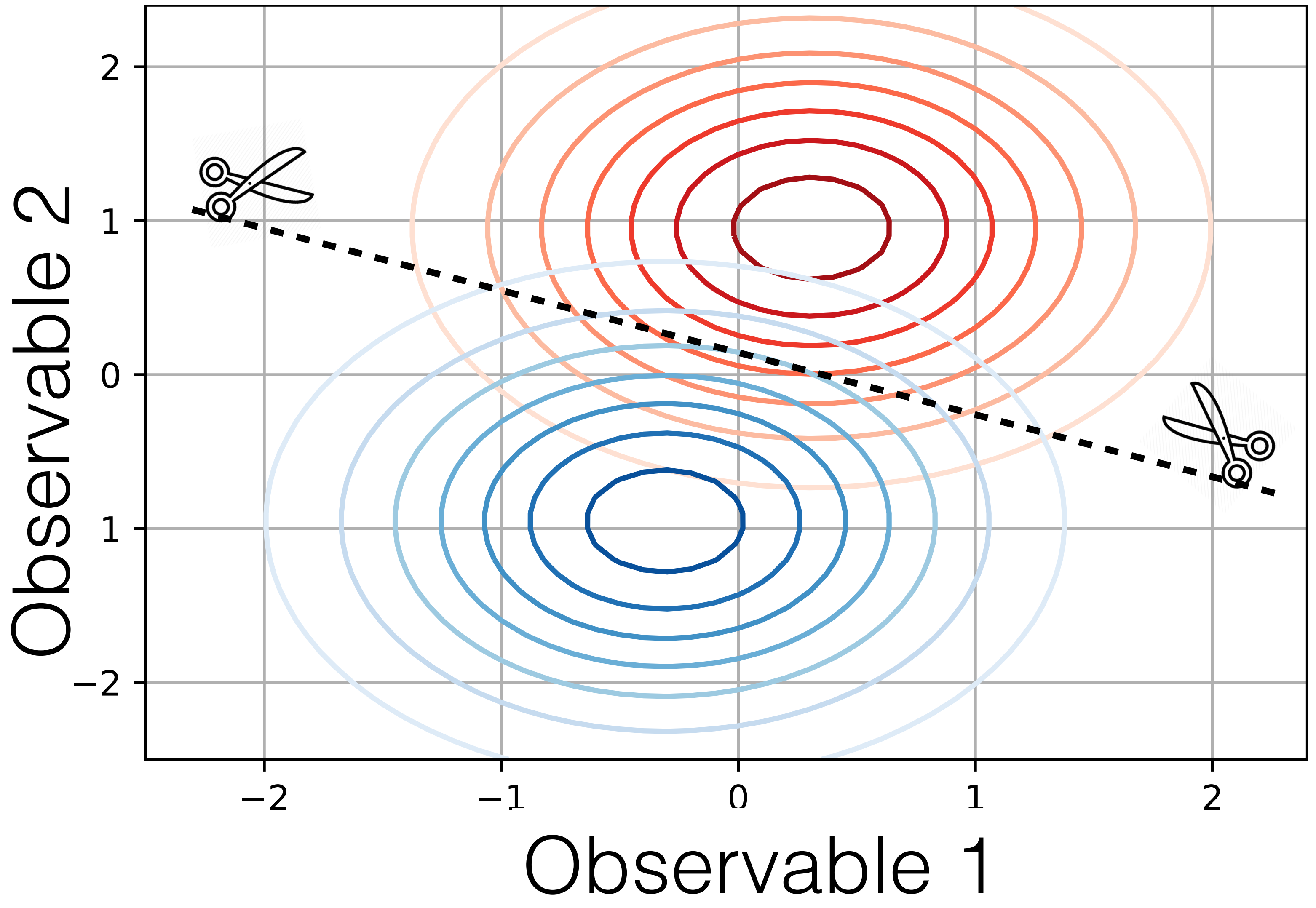


z = Nuisance Parameter

Prior

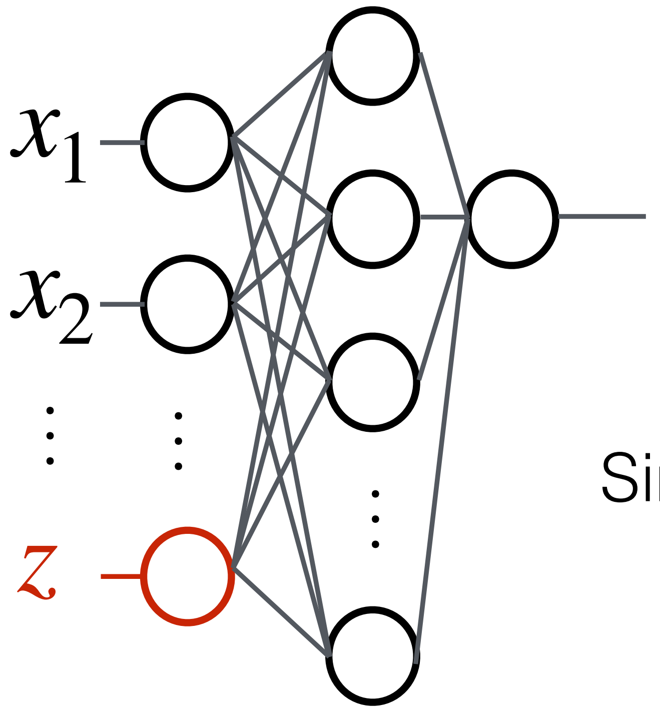






Opposite of decorrelation: Uncertainty-aware learning

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



$f(x_1, x_2, \ldots, z)$

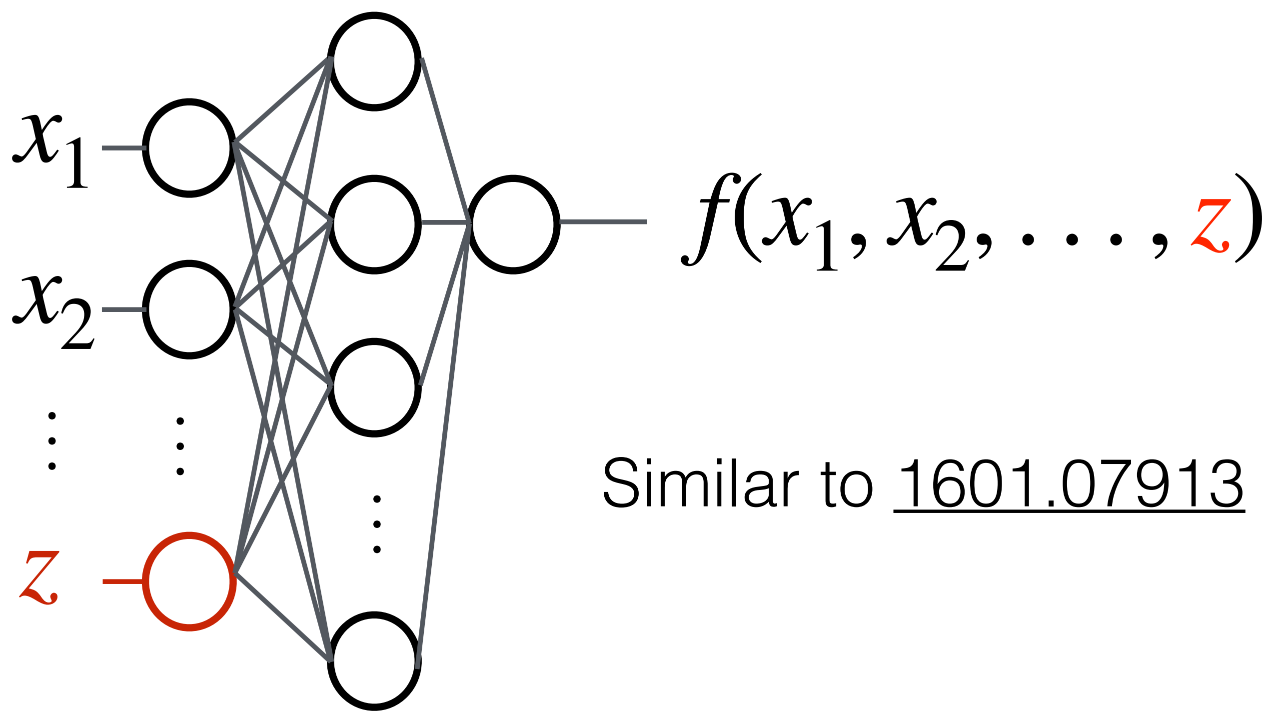
Similar to 1601.07913

Repeat for each hypothesis z

Data
with Z = ?

Opposite of decorrelation: Uncertainty-aware learning

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



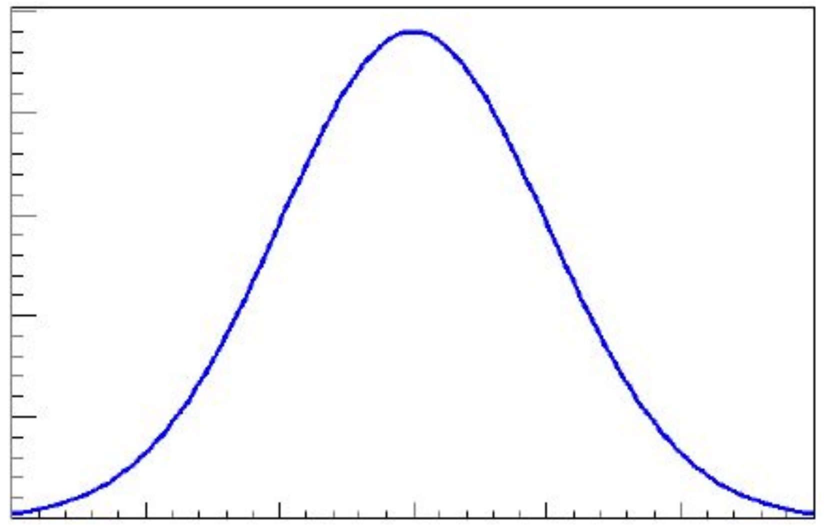
Similar to 1601.07913

Repeat for each hypothesis z

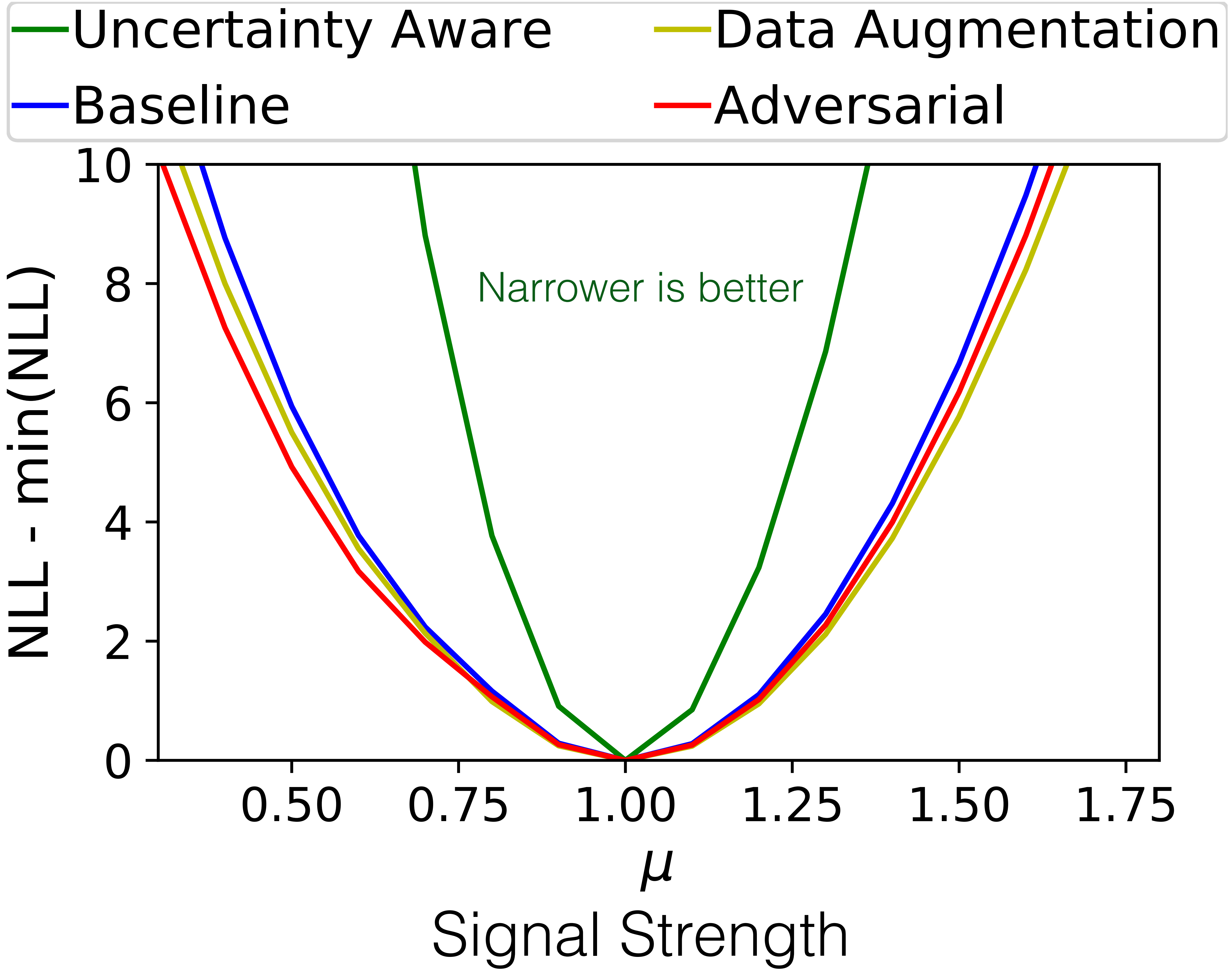
Data
with Z = ?

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

- Profile Z + incorporate prior



Better final measurements!

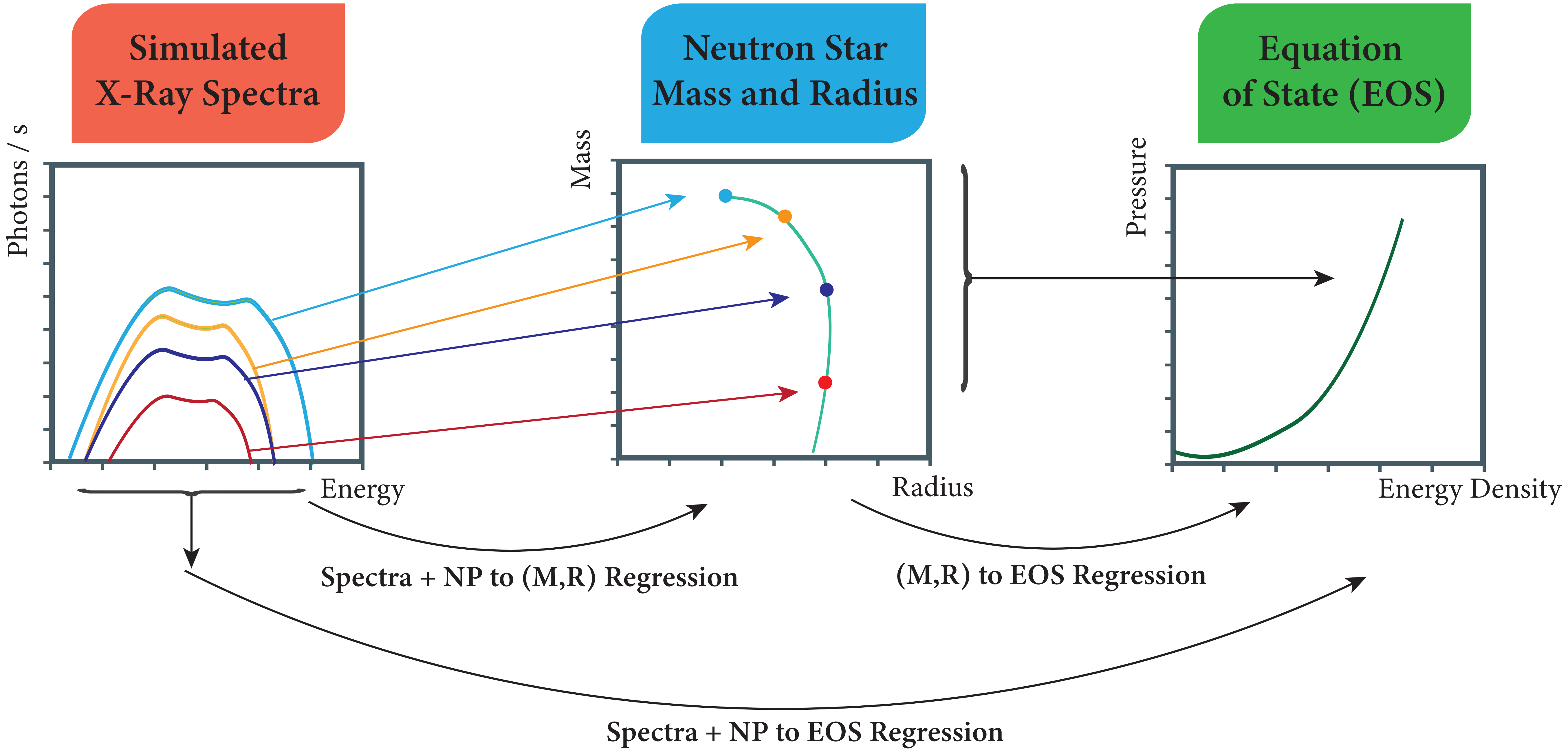


Narrower ⇒ Smaller [statistical + systematic] uncertainty on measurement

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

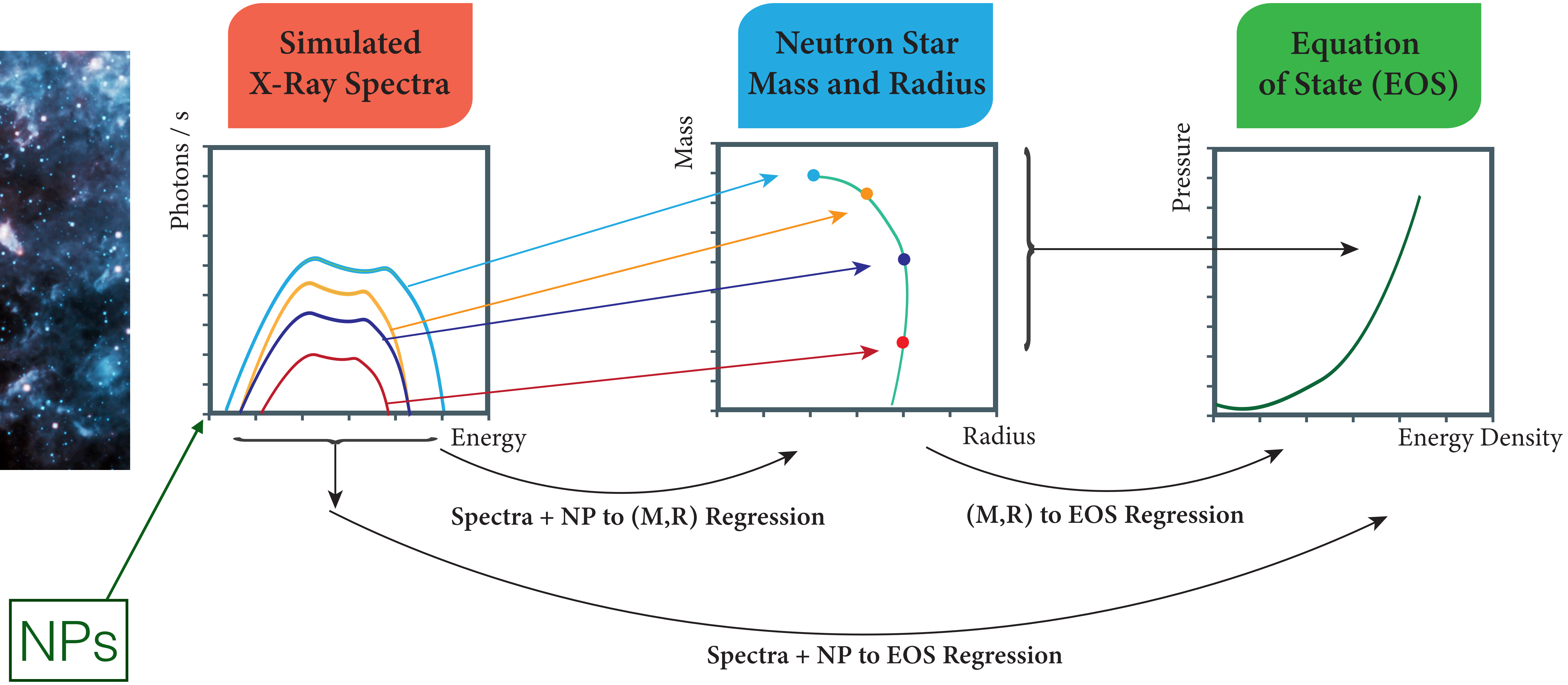
An application in astrophysics

Application in Astrophysics: Full propagation of uncertainties



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

Application in Astrophysics: Full propagation of uncertainties

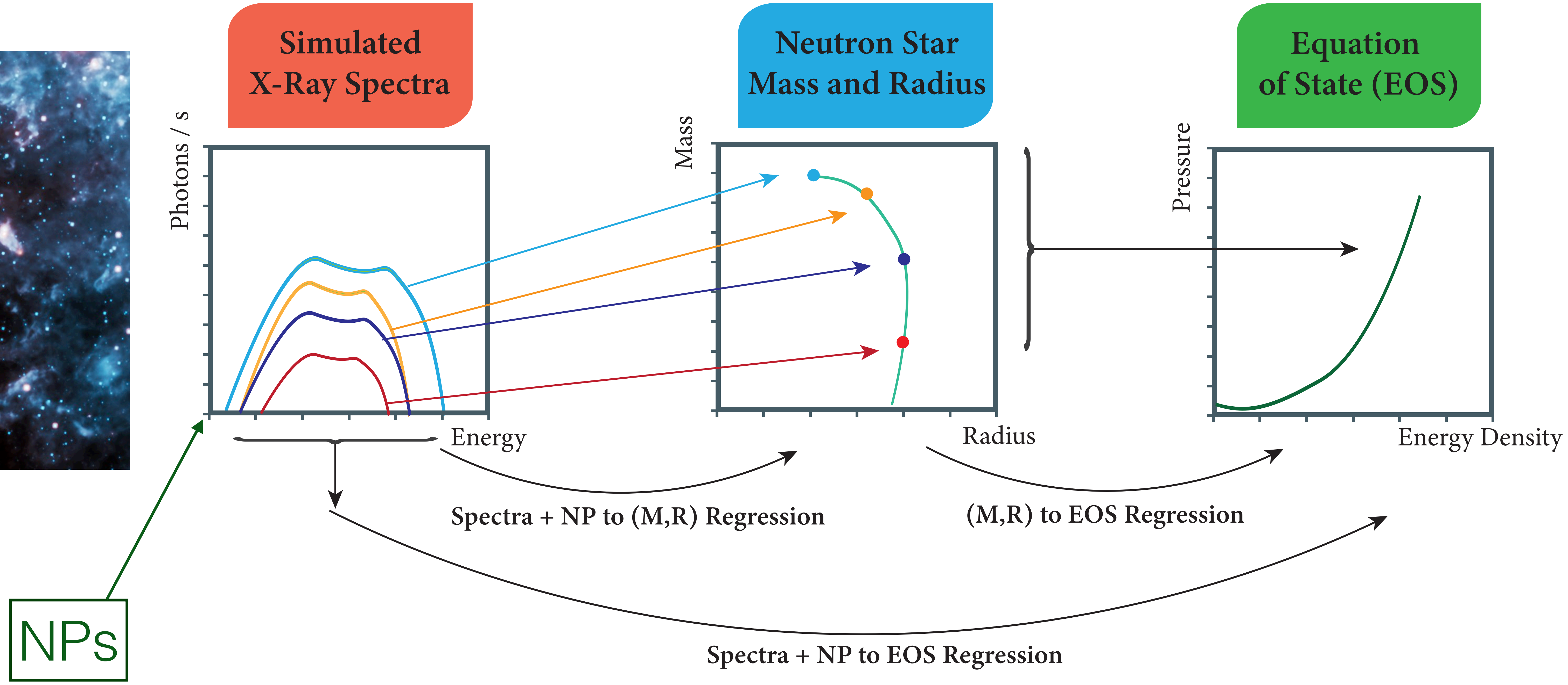


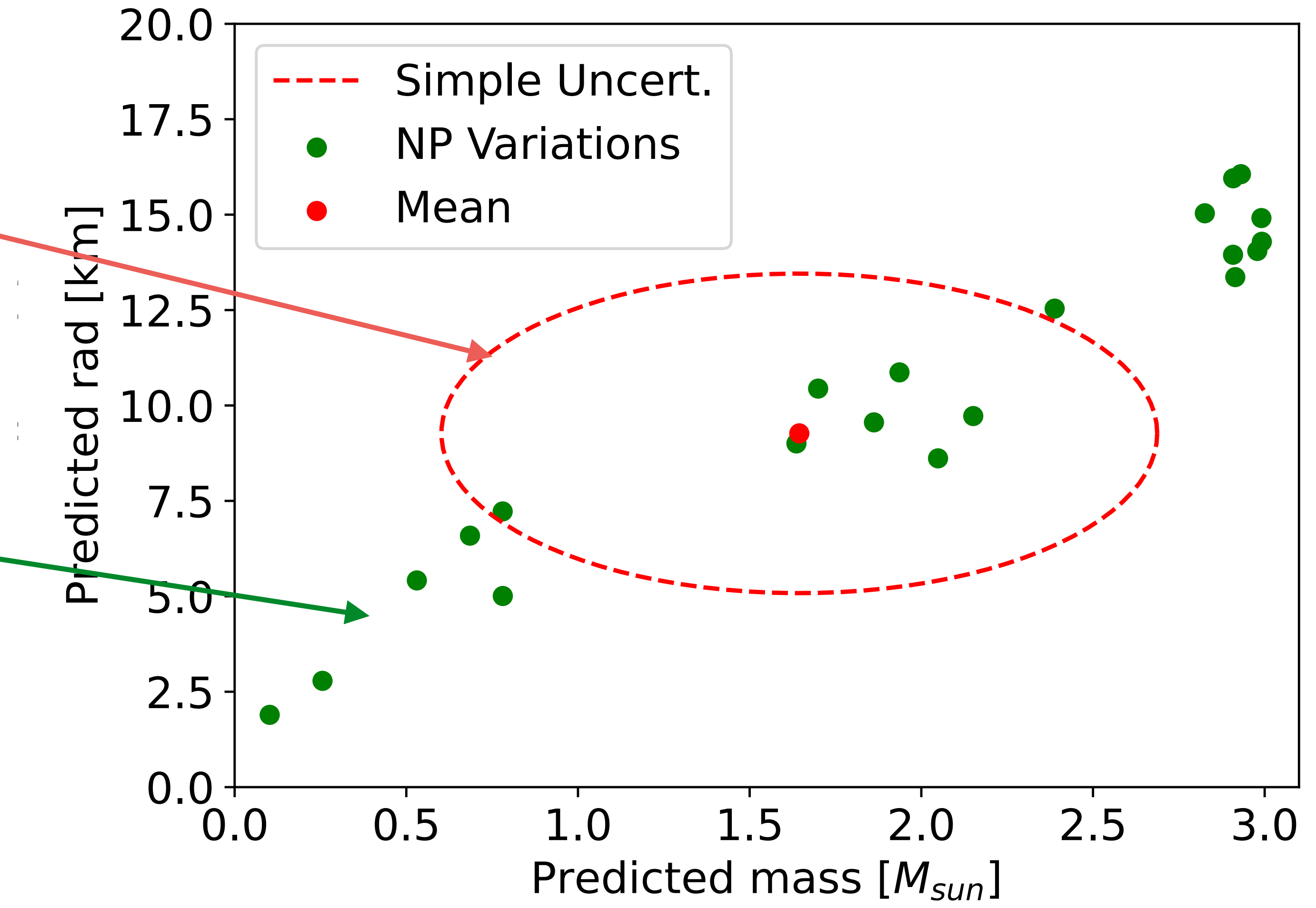
JCAP.020P.0922: Delaney Farrell, Pierre Baldi, Jordan Ott, **Aishik Ghosh**, Andrew W. Steiner, Atharva Kavitkar, Lee Lindblom, Daniel Whiteson, Fridolin Weber

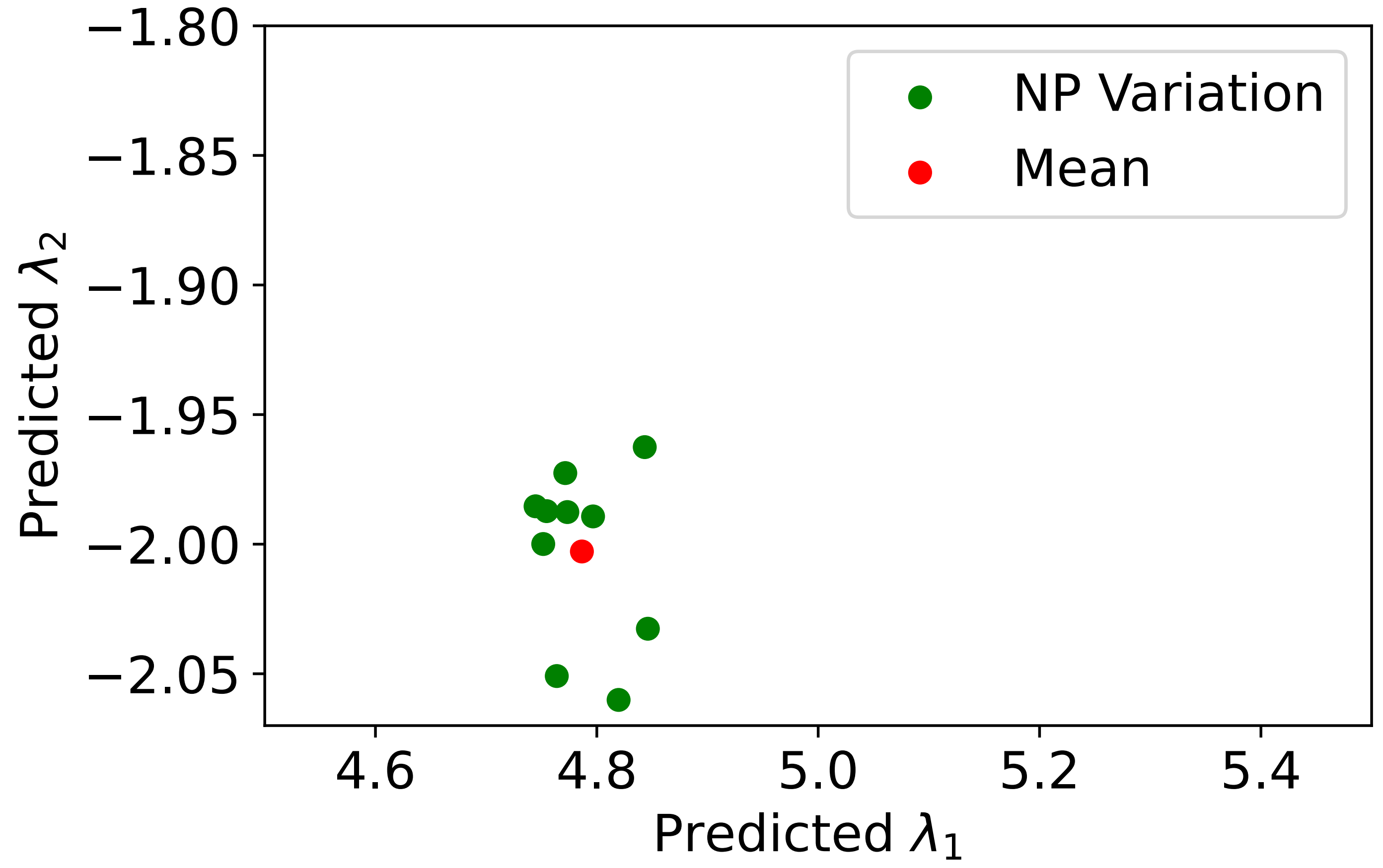
Application in Astrophysics: Full propagation of uncertainties

SOTA made a single point estimate + assumed uncorrelated Gaussian uncertainties

Real uncertainties look quite different

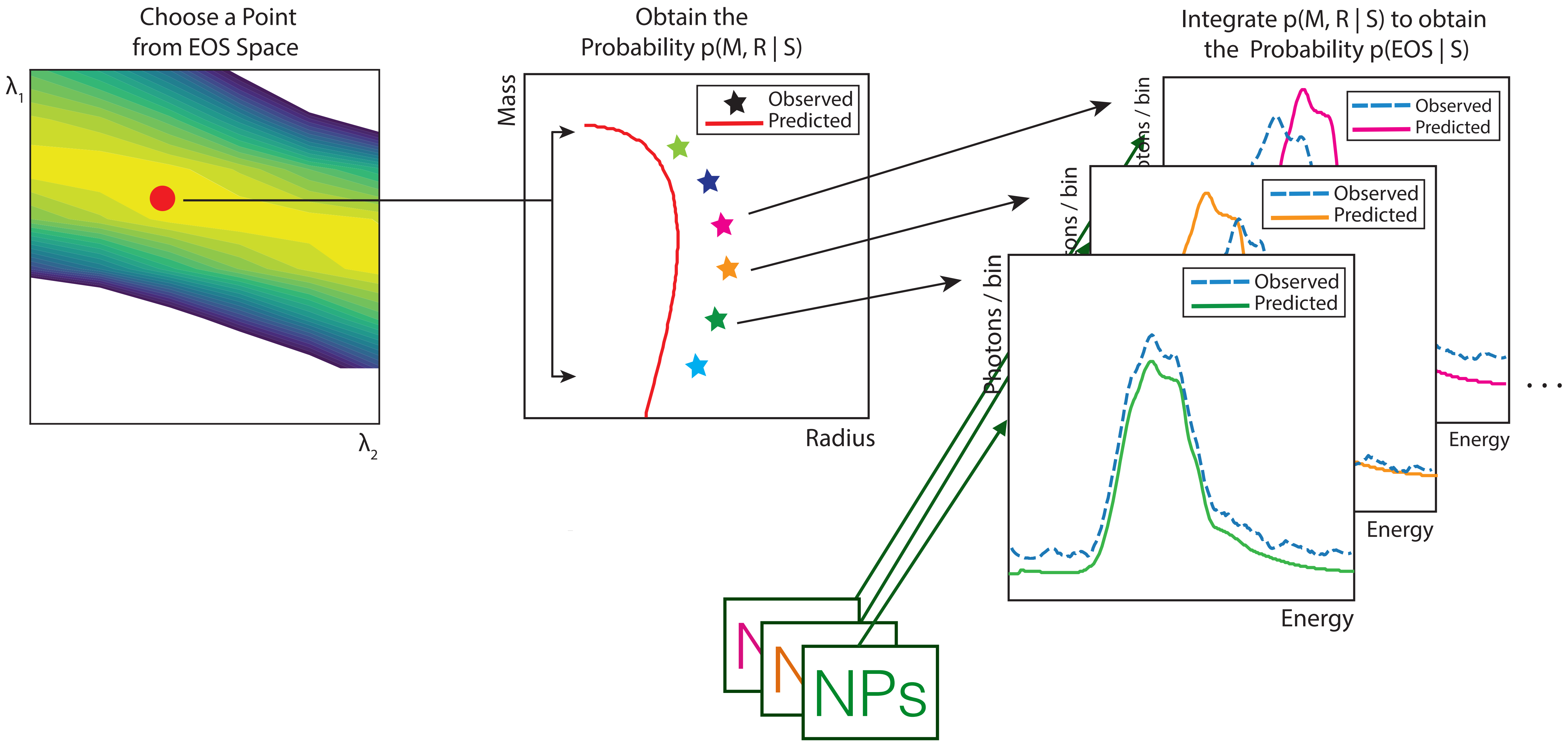






JCAP.020P.0922: Delaney Farrell, Pierre Baldi, Jordan Ott, **Aishik Ghosh**, Andrew W. Steiner, Atharva Kavitkar, Lee Lindblom, Daniel Whiteson, Fridolin Weber

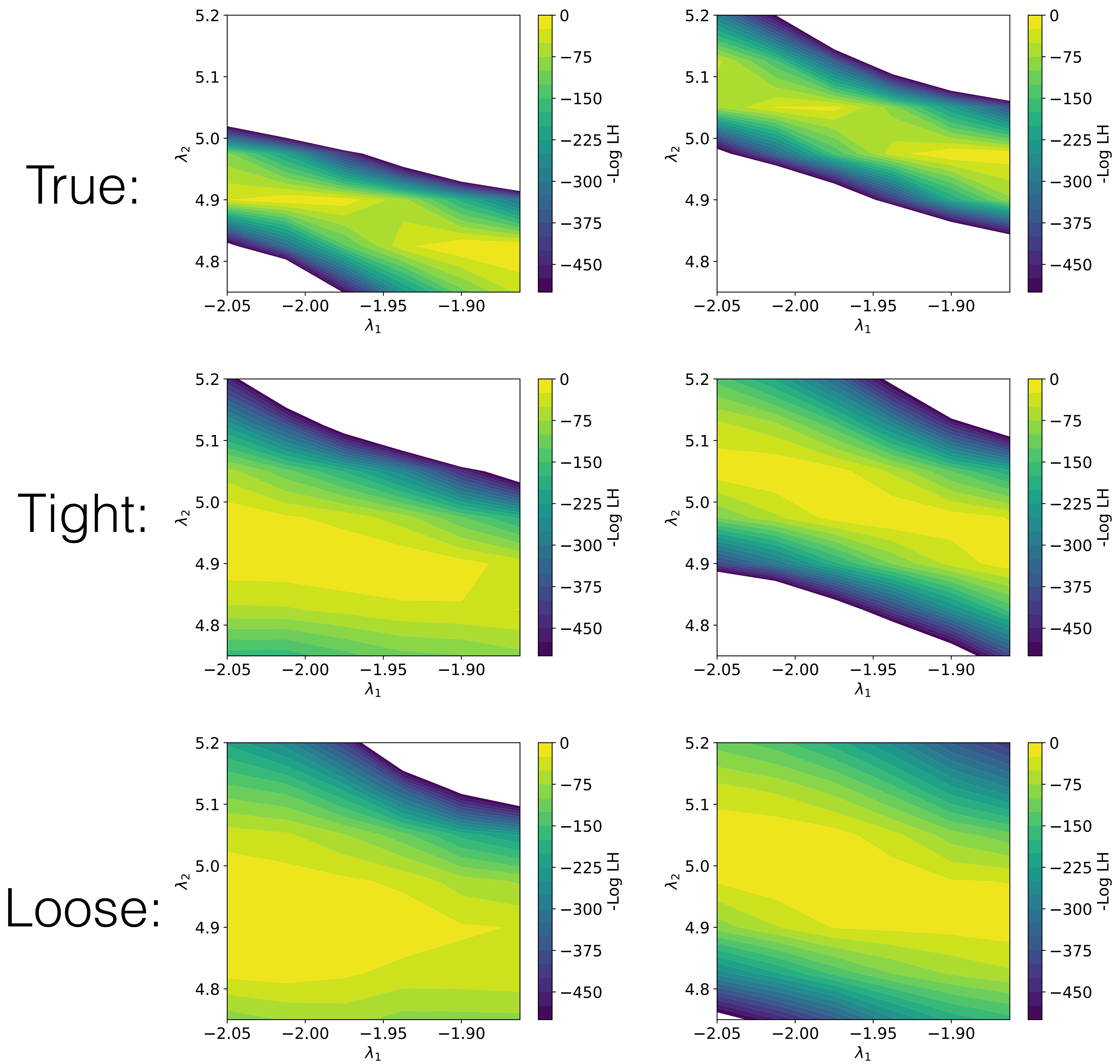
Learn forward process to access the likelihood



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

Nuisance Priors:

EOS parameter likelihoods:



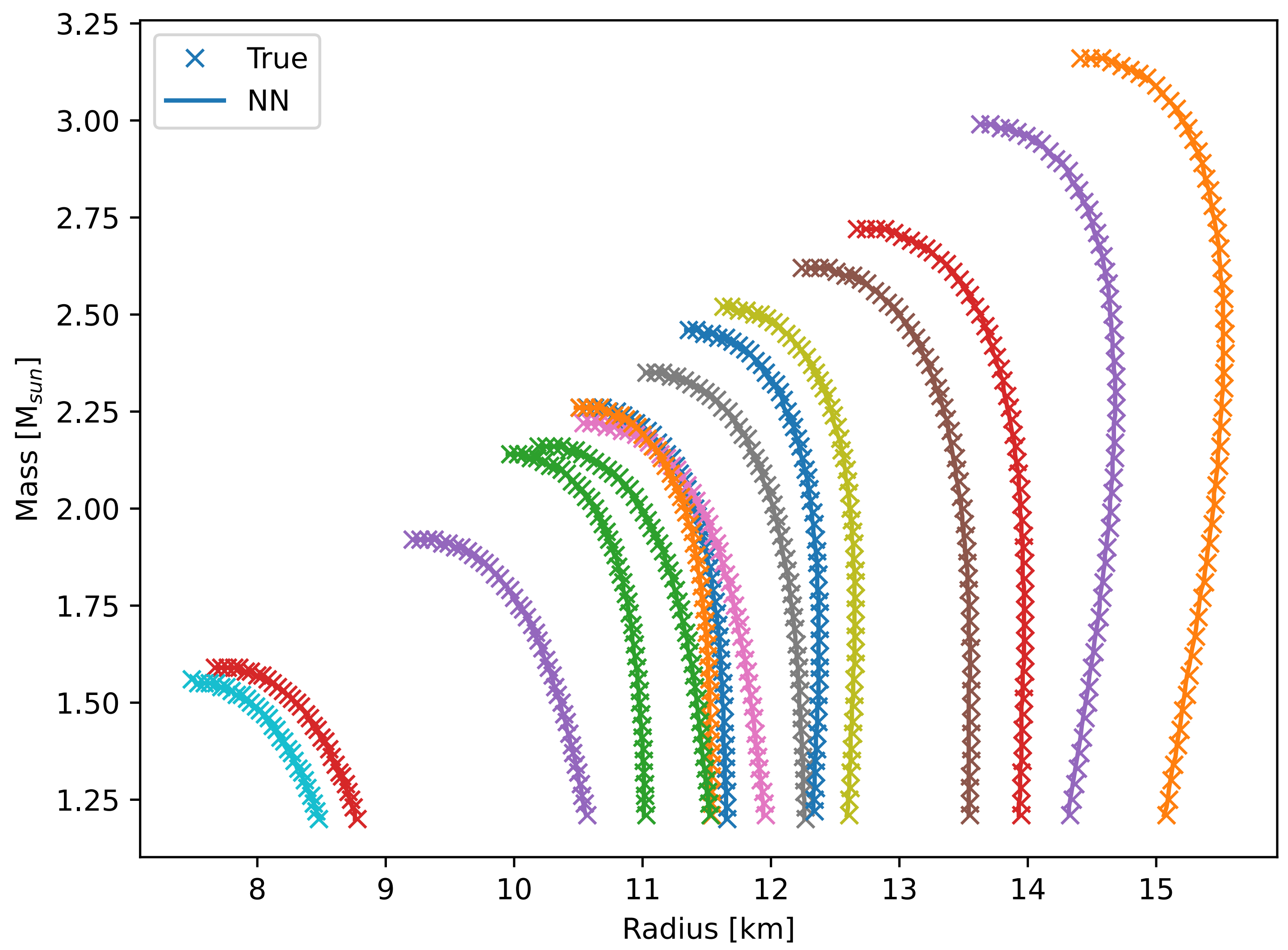
Paper in preparation

Forward process step-by-step

Intermediate steps remain interpretable physical quantities

Forward process step-by-step

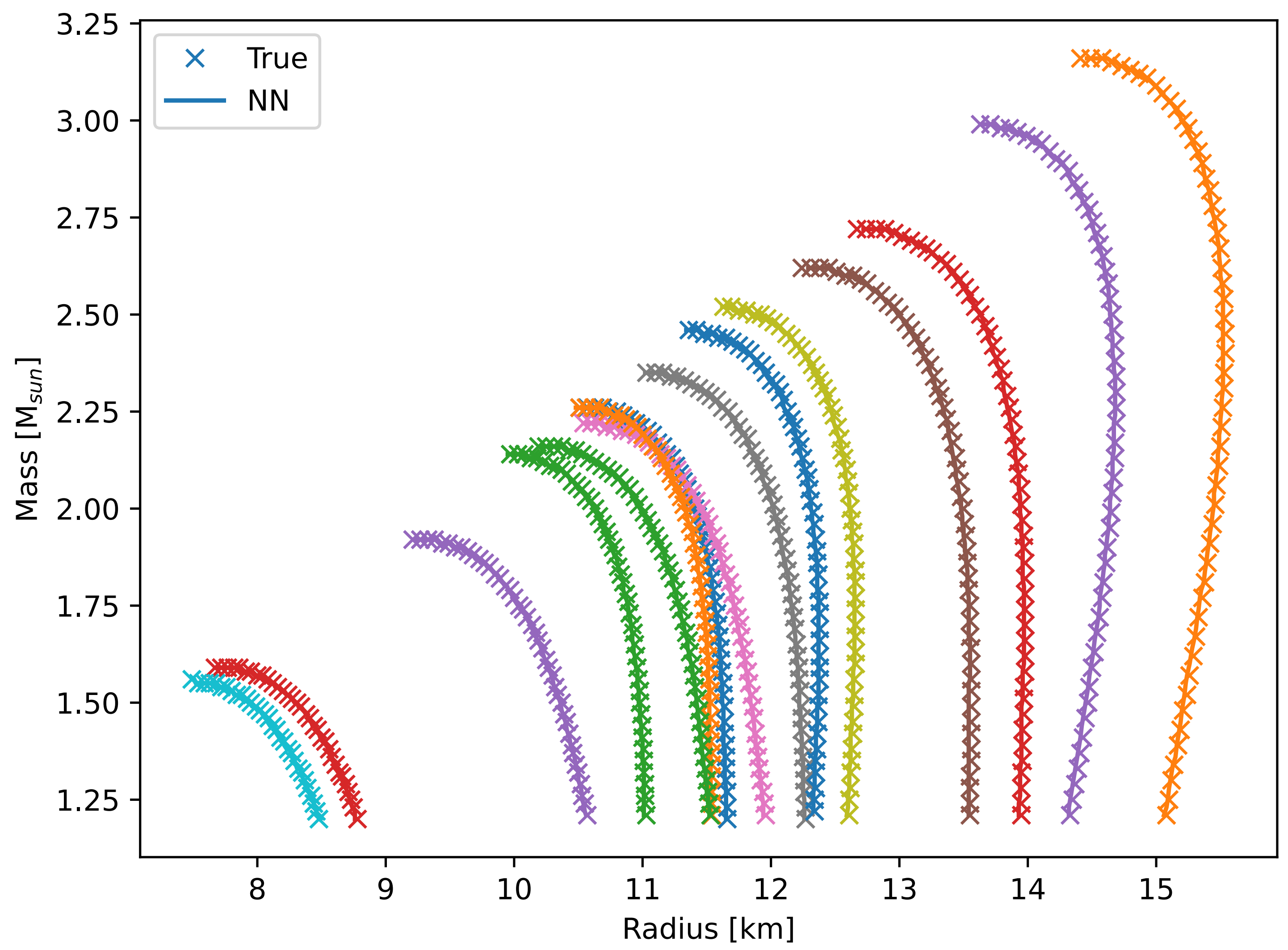
Intermediate steps remain interpretable physical quantities

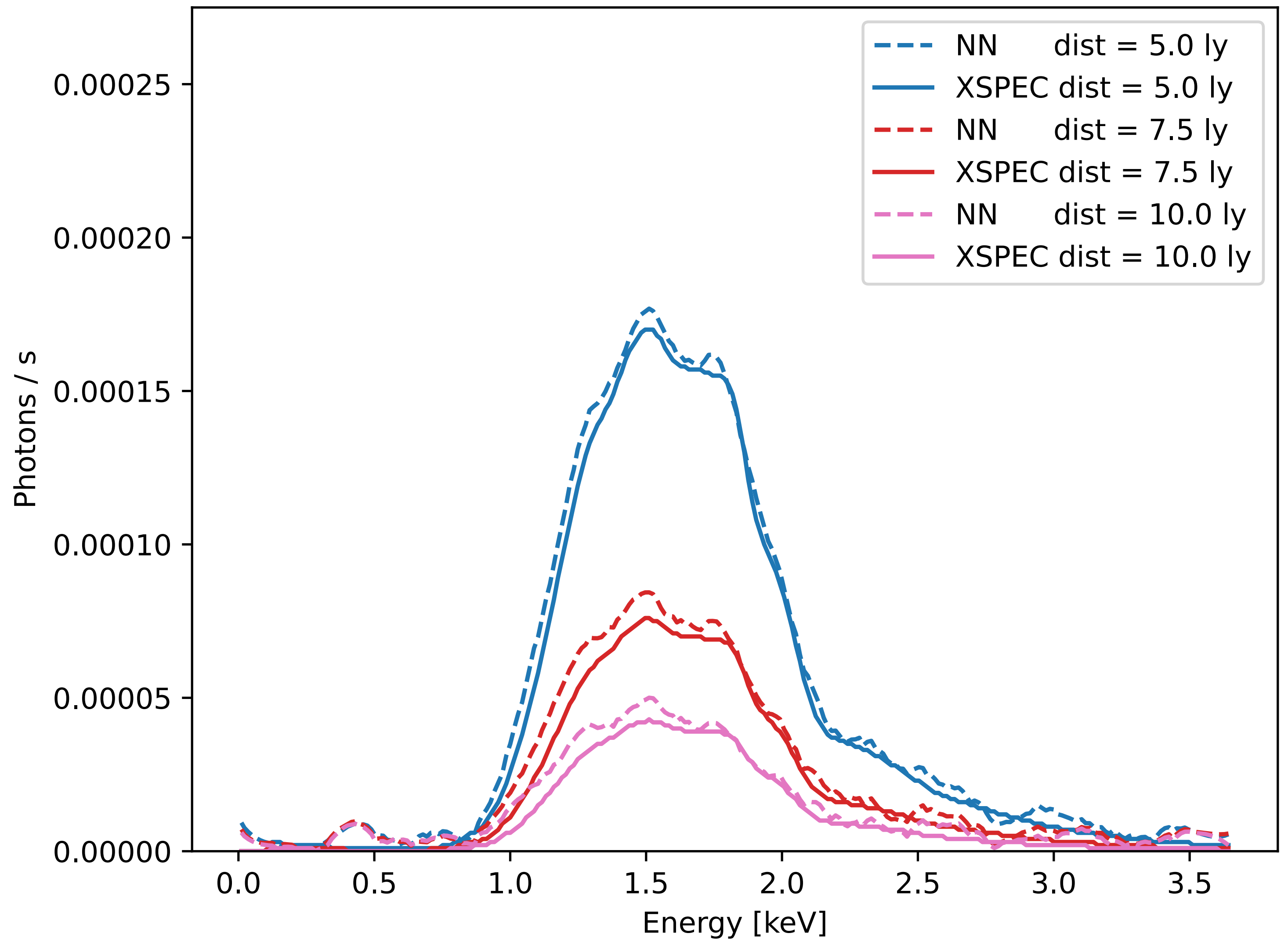


Learn EOS to M-R

Forward process step-by-step

Intermediate steps remain interpretable physical quantities





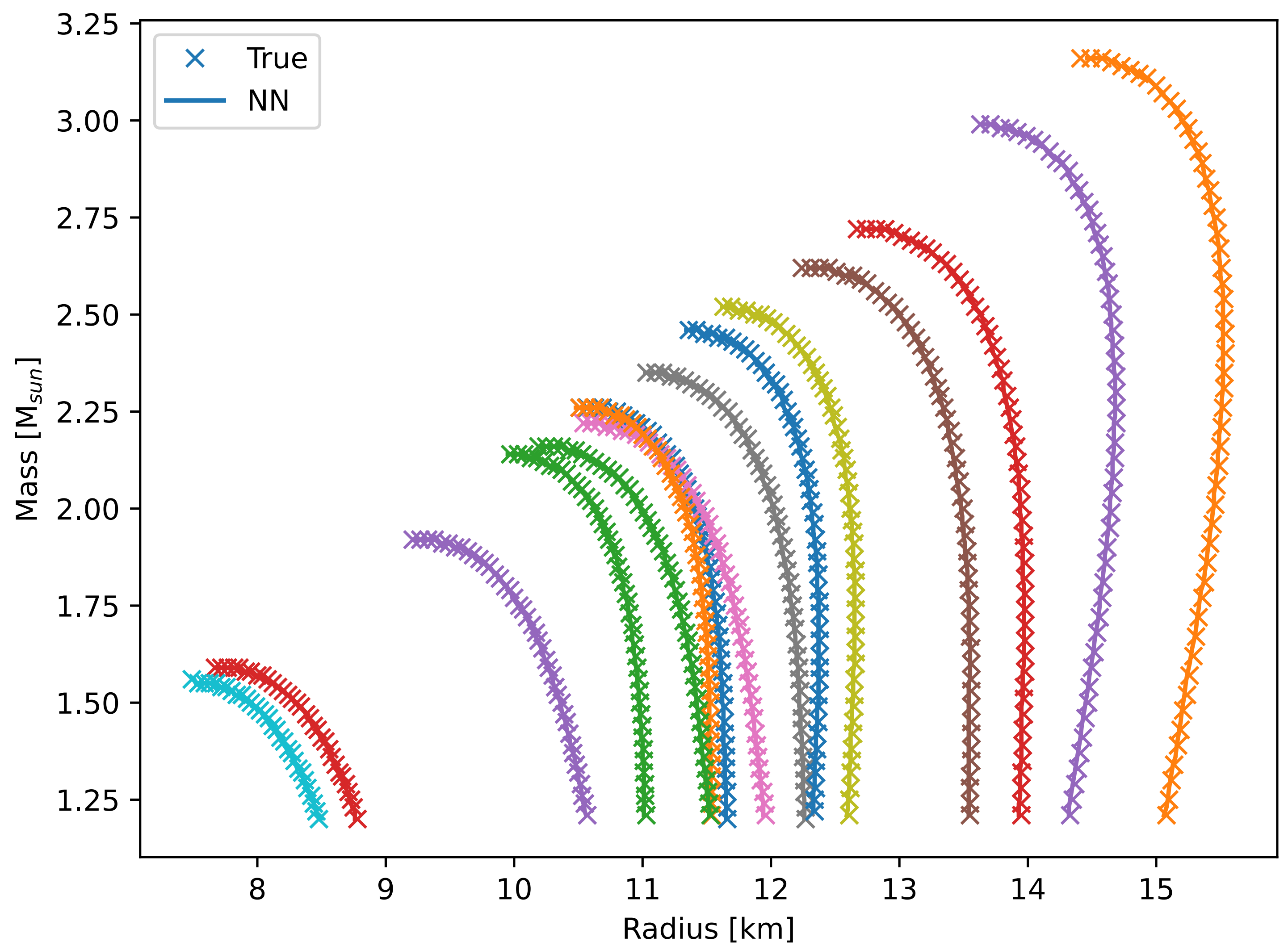
Learn EOS to M-R

Learn {M,R,NPs} to Spectrum

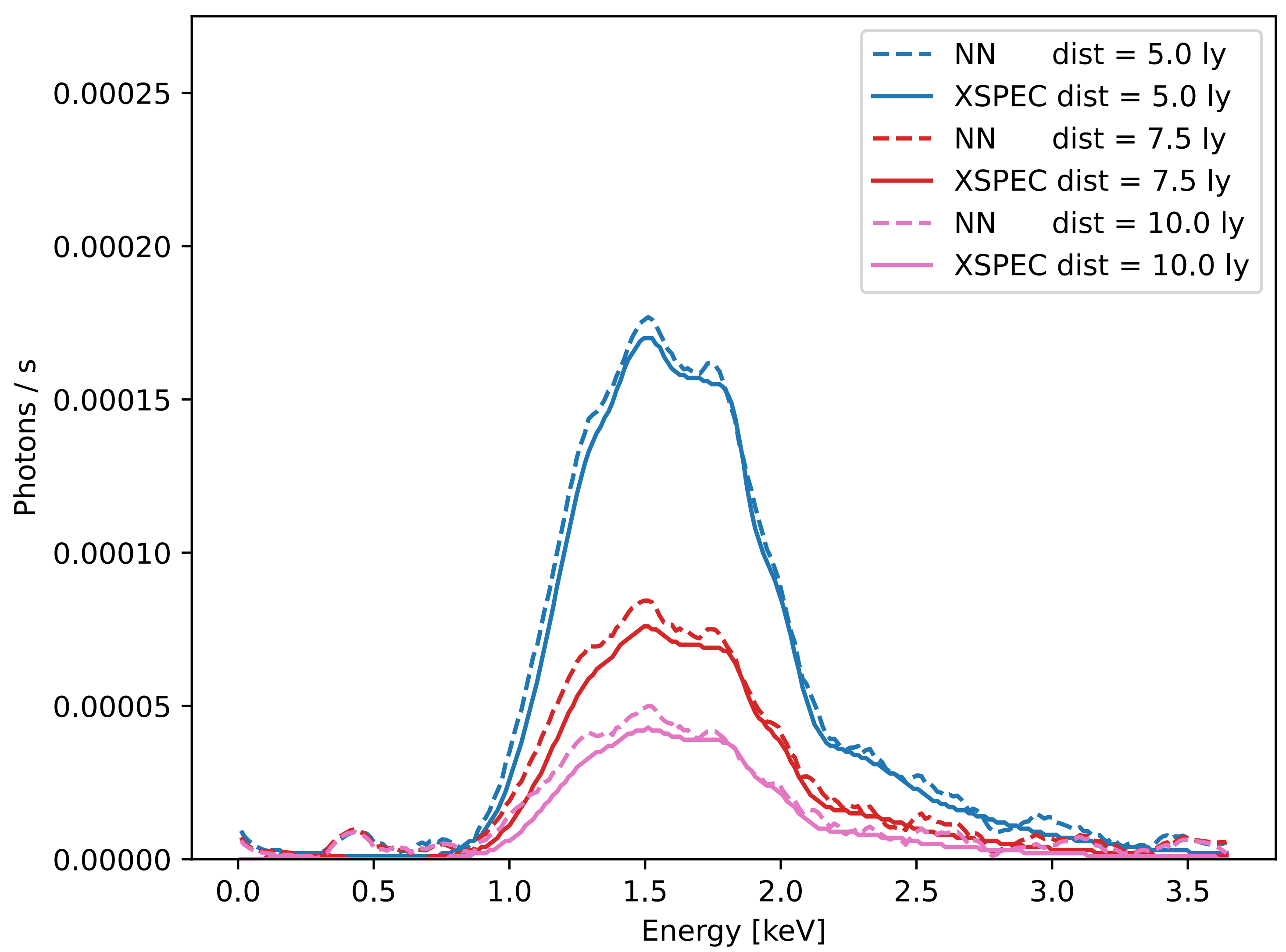
Paper in preparation

Forward process step-by-step

Intermediate steps remain interpretable physical quantities



Learn EOS to M-R



Learn {M,R,NPs} to Spectrum

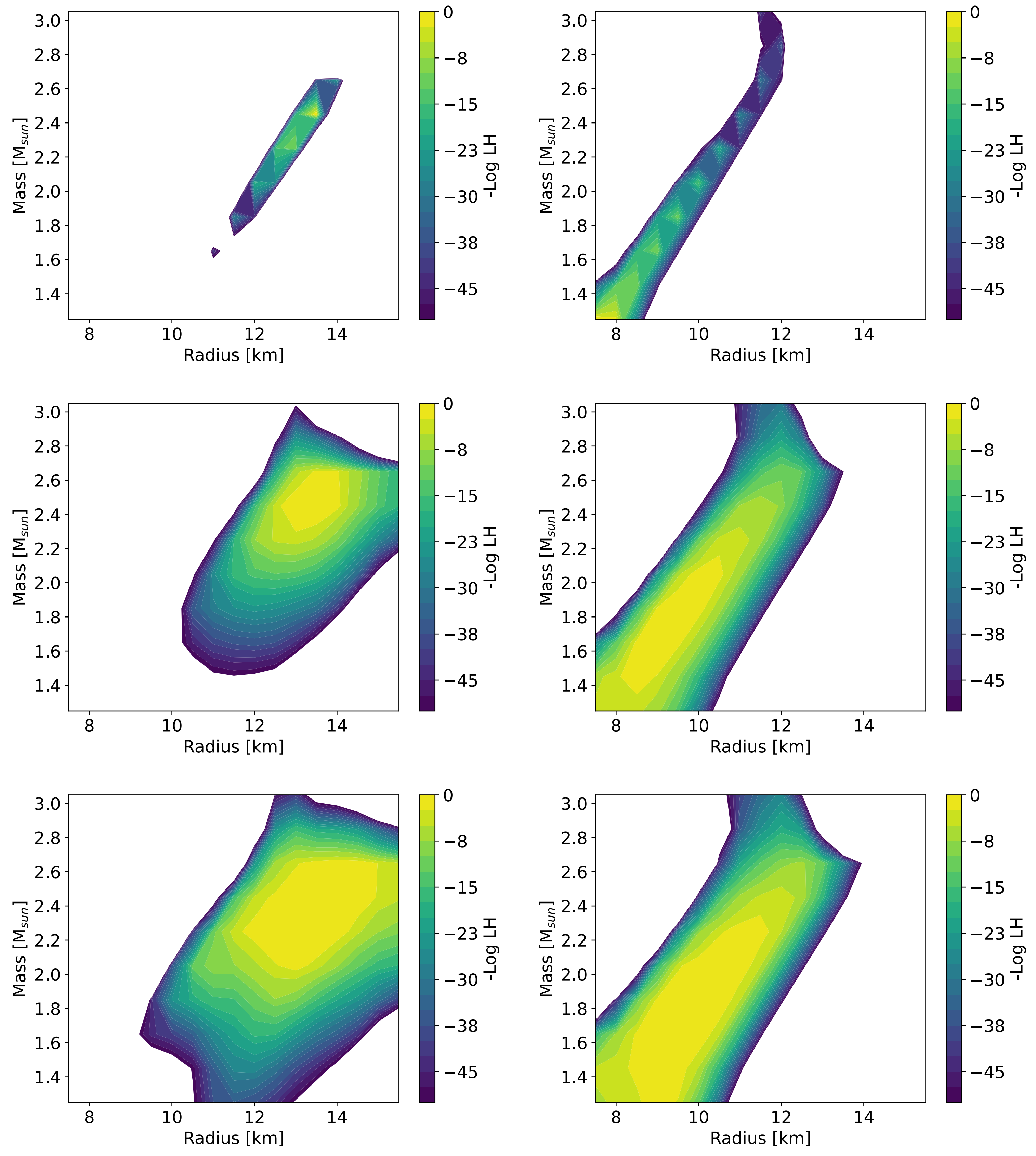
Nuisance Priors:

M-R likelihoods:

True:

Tight:

Loose:



Back to particle physics

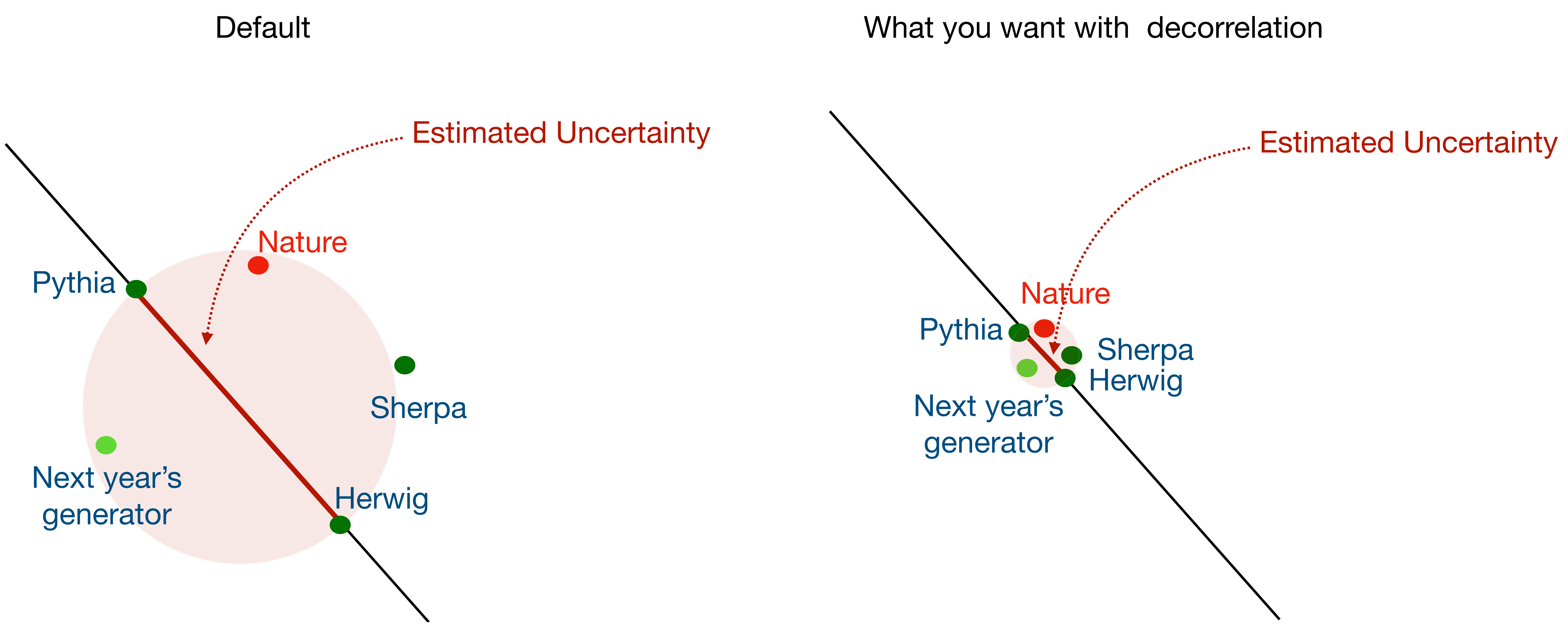
Back to particle physics

We also learnt what not to do …

ML-decorrelating theory uncertainties

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

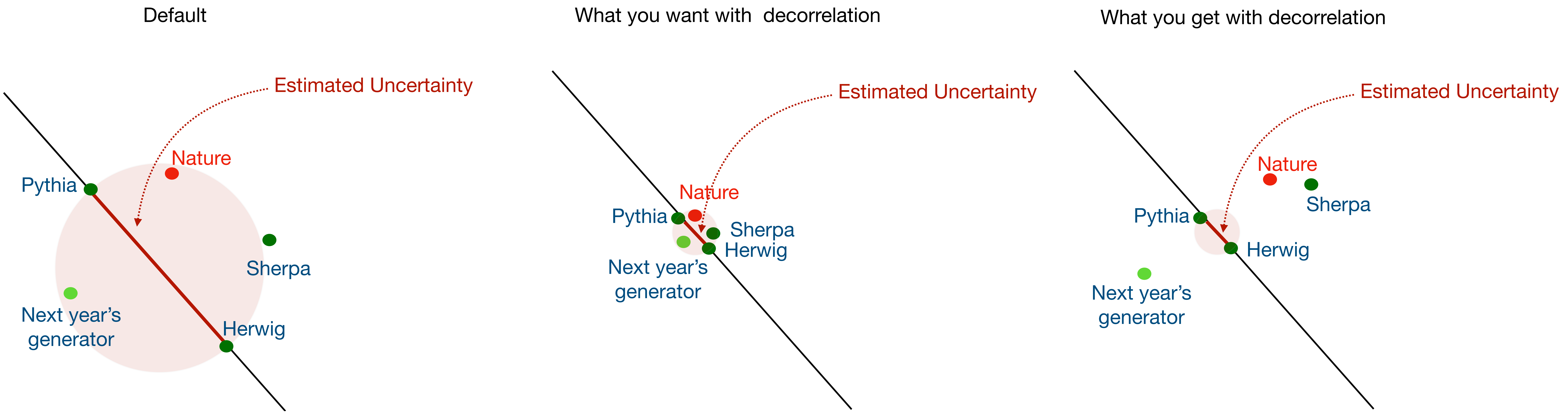
Default — What you want with decorrelation



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

ML-decorrelating theory uncertainties

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

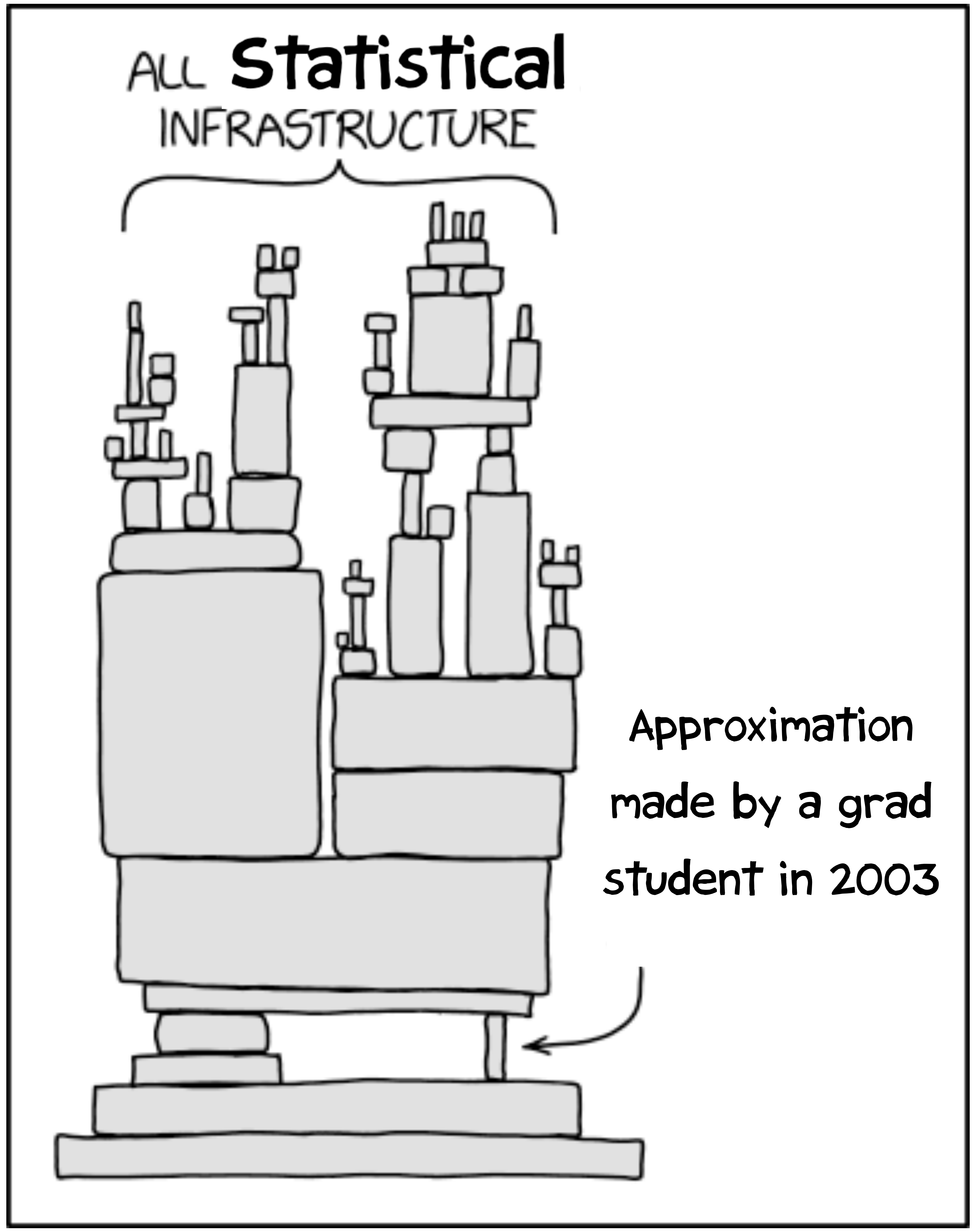


Instruction to ML: "Please shrink Pythia vs Herwig difference"

Model will learn to fool you !

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

Theory uncertainties



From Daniel Whiteson

It’s dangerous to use ML methods to mitigate theory uncertainties

But we continue to treat Δ_{theory} and Δ_{exp} on same footing in statistical fits

What even is their statistical behaviour?

Scale Uncertainties

Uncertainty of cross-section from truncating QFT series

Sensitivity to scale variation quantifies ‘uncertainty’

Scale Uncertainties

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

$$\mu_0 = \frac{H_T}{2} = \frac{1}{2} \sum_{final\ state} \sqrt{m^2 + p_T^2}$$

Uncertainty of cross-section from truncating QFT series

Sensitivity to scale variation quantifies ‘uncertainty’

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

Questions

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

Study pull distribution

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

Questions

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

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)			
Vector boson +jets			LO 13 TeV		NLO 13 TeV	
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)

Study pull distribution

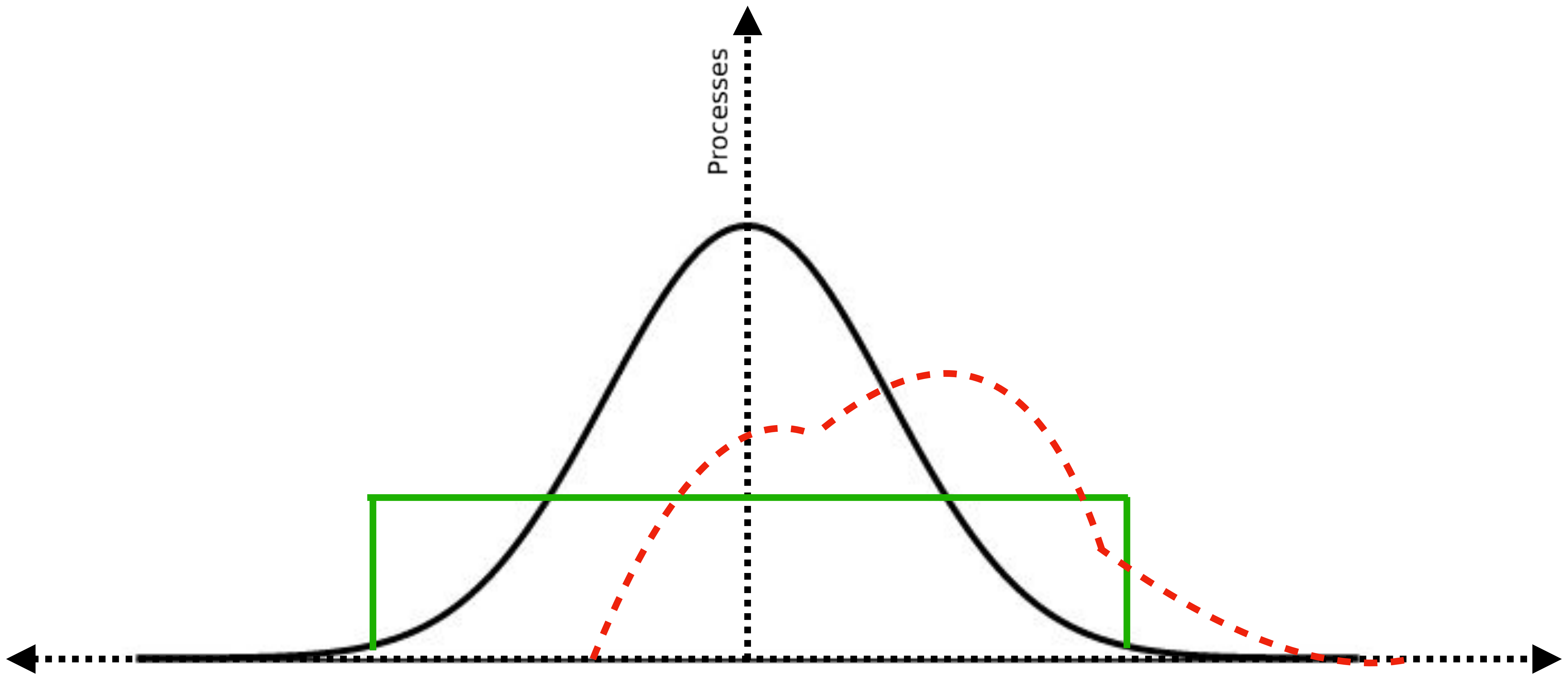
$$t_{scale} = \frac{\sigma_{NLO} - \sigma_{LO}}{\Delta\sigma_{LO\ scale}}$$

Alwall et al.

Plot the pulls

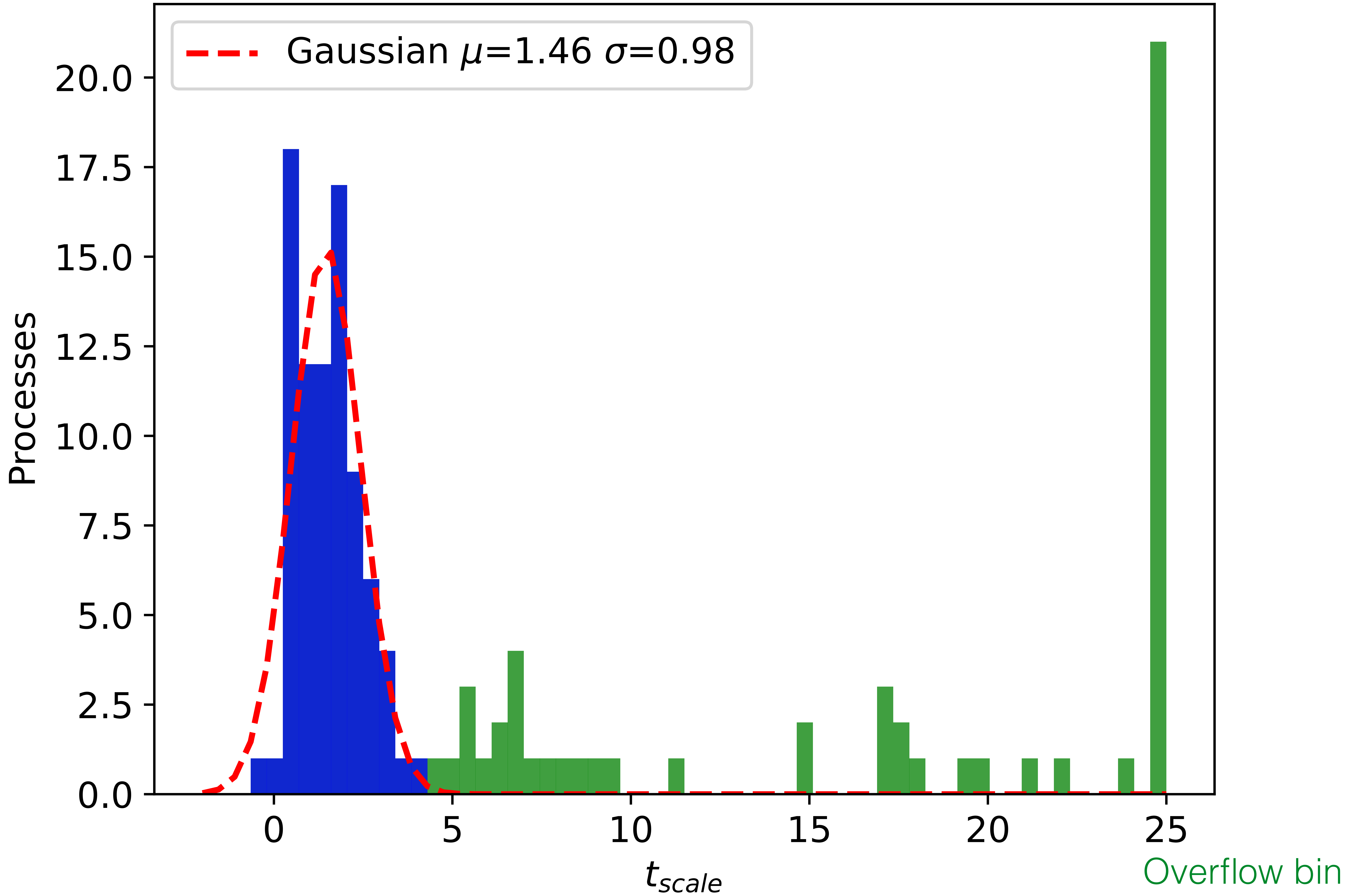
$$t_{scale} = \frac{\sigma_{NLO} - \sigma_{LO}}{\Delta\sigma_{LO\ scale}}$$

Which of these distributions do you expect?

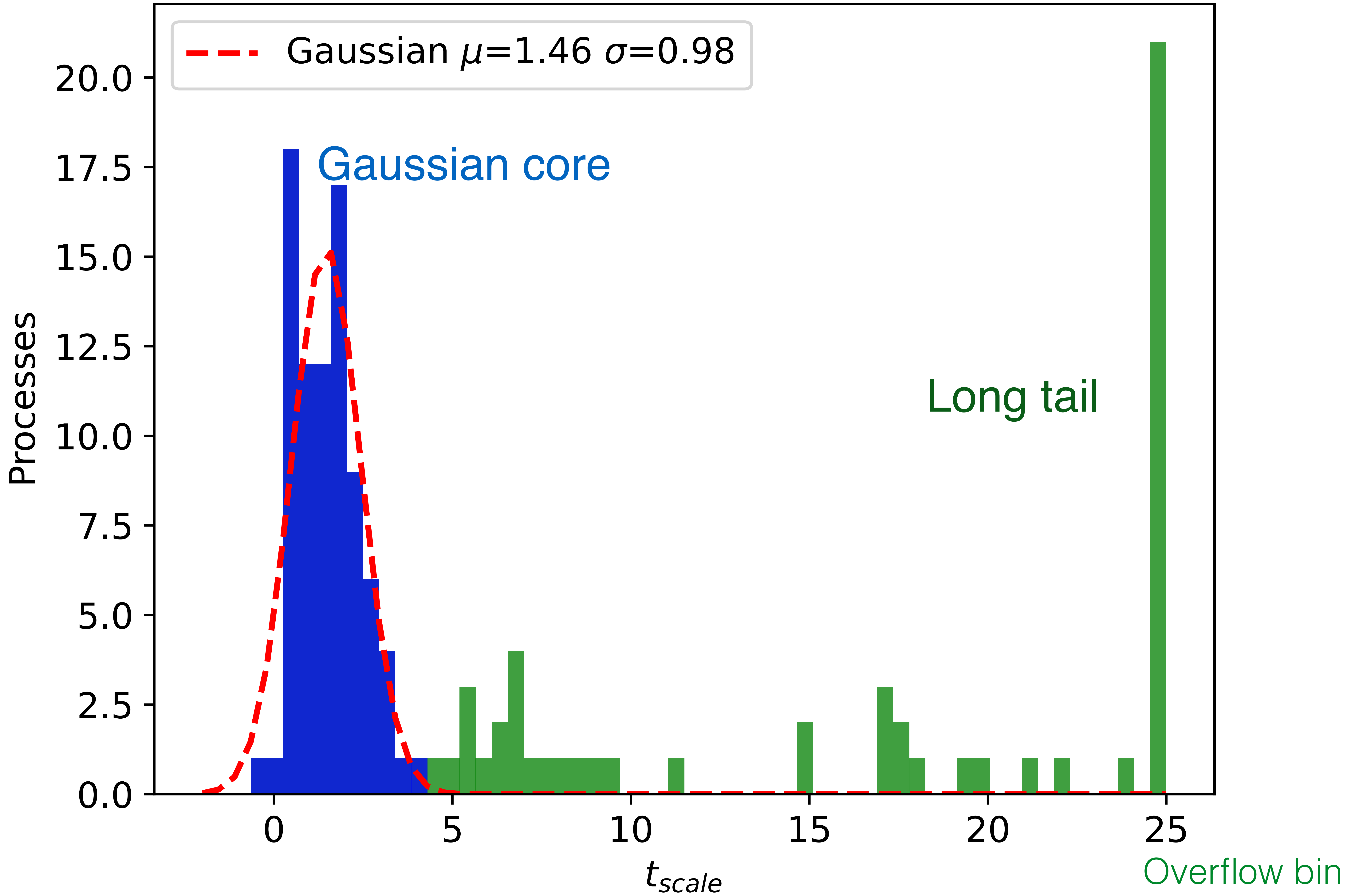


$$t_{scale} = \frac{\sigma_{NLO} - \sigma_{LO}}{\Delta\sigma_{LO\ scale}}$$

Pull distribution



Pull distribution



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

QCD processes follow (an expected) pattern

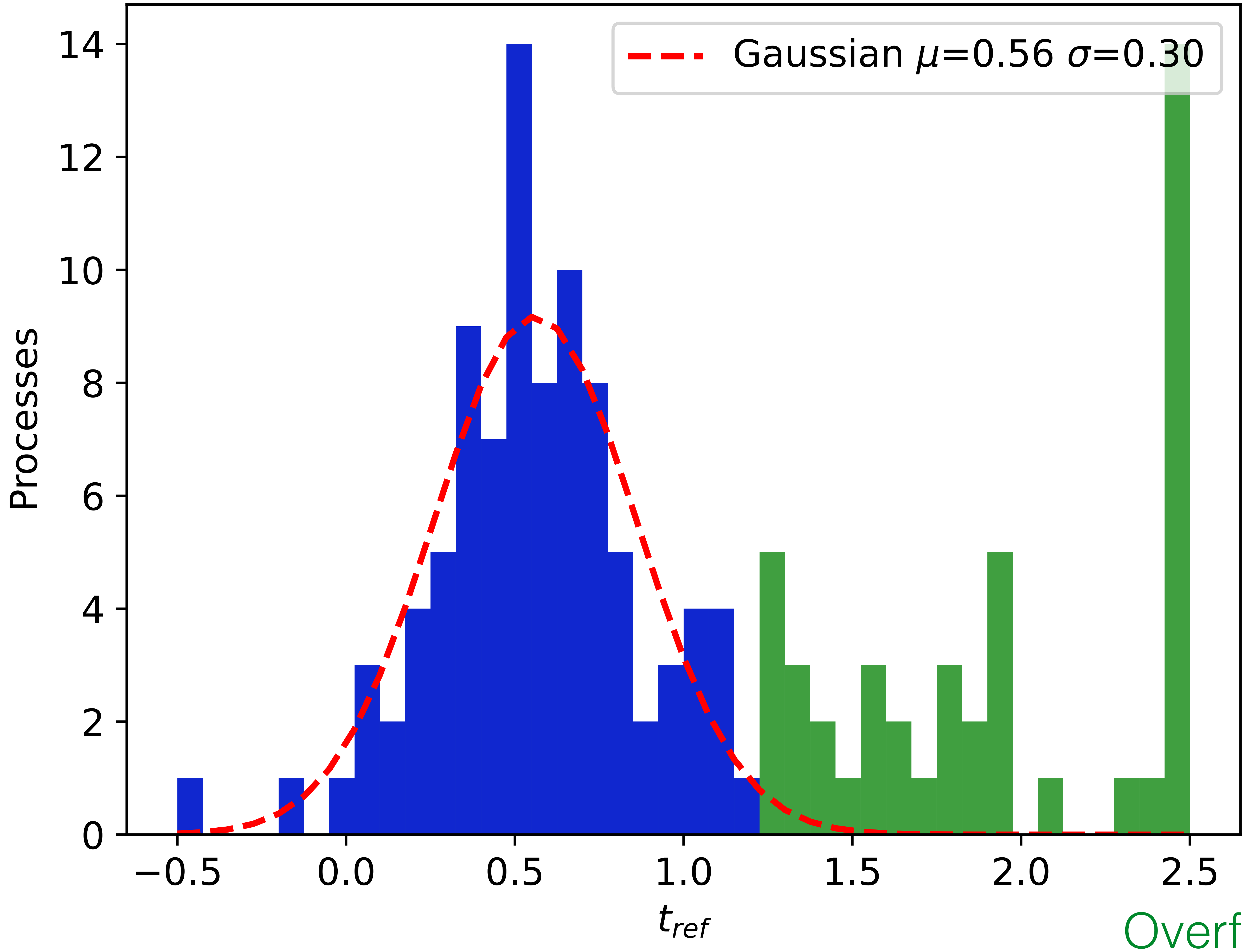
Process	$\frac{\Delta\sigma}{\sigma_0}$		n	$\frac{\Delta\sigma}{n\,\sigma_0}$	
p p > j j	$+2.49\times10^{-1}$	-1.88×10^{-1}	2	$+1.24\times10^{-1}$	-9.40×10^{-2}
p p > b b	$+2.52\times10^{-1}$	-1.89×10^{-1}	2	$+1.26\times10^{-1}$	-9.45×10^{-2}
p p > t t	$+2.90\times10^{-1}$	-2.11×10^{-1}	2	$+1.45\times10^{-1}$	-1.06×10^{-1}
p p > j j j	$+4.38\times10^{-1}$	-2.84×10^{-1}	3	$+1.46\times10^{-1}$	-9.47×10^{-2}
p p > b b j	$+4.41\times10^{-1}$	-2.85×10^{-1}	3	$+1.47\times10^{-1}$	-9.50×10^{-2}
p p > t t j	$+4.51\times10^{-1}$	-2.90×10^{-1}	3	$+1.50\times10^{-1}$	-9.67×10^{-2}
p p > b b j j	$+6.18\times10^{-1}$	-3.56×10^{-1}	4	$+1.54\times10^{-1}$	-8.90×10^{-2}
p p > b b b b	$+6.17\times10^{-1}$	-3.56×10^{-1}	4	$+1.54\times10^{-1}$	-8.90×10^{-2}
p p > t t j j	$+6.14\times10^{-1}$	-3.56×10^{-1}	4	$+1.53\times10^{-1}$	-8.90×10^{-2}
p p > t t t t	$+6.38\times10^{-1}$	-3.65×10^{-1}	4	$+1.60\times10^{-1}$	-9.12×10^{-2}
p p > t t b b	$+6.21\times10^{-1}$	-3.57×10^{-1}	4	$+1.55\times10^{-1}$	-8.93×10^{-2}
average				$+1.47\times10^{-1}$	-9.34×10^{-2}

Table 1: Scale dependence for LHC processes with only QCD particles in the final state. For each process, we report the relative scale uncertainty, the number of final state particles, and the per-particle relative scale uncertainty.

→Tilman Plehn's 'reference process' method

$$\frac{\Delta\sigma_{\rm ref}}{\sigma_0} = n \times \left\langle \frac{\Delta\sigma}{n\sigma_0} \right\rangle_{\rm QCD} .$$

Make correction in UQ for EW processes

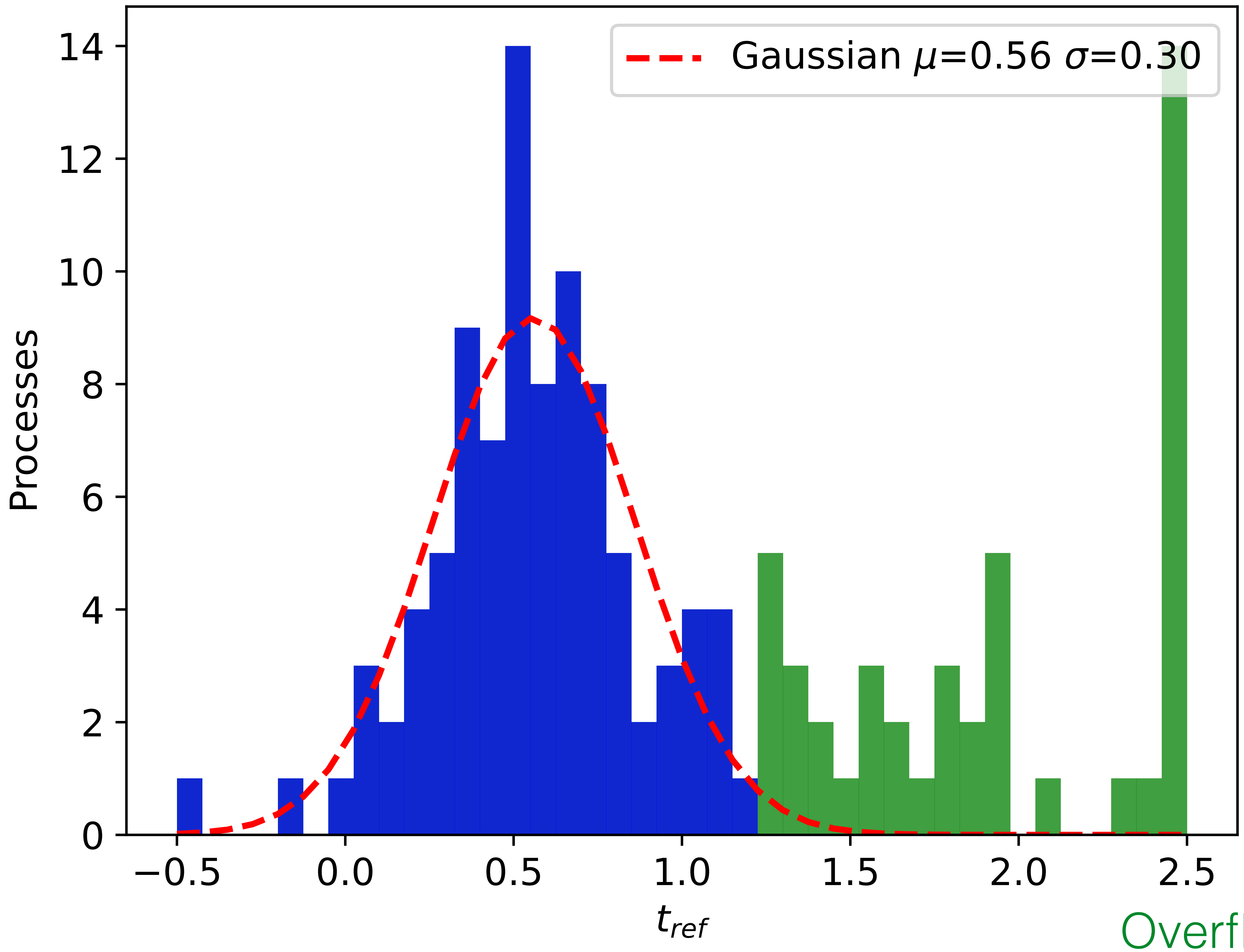


Overflow bin

Much reduced tails

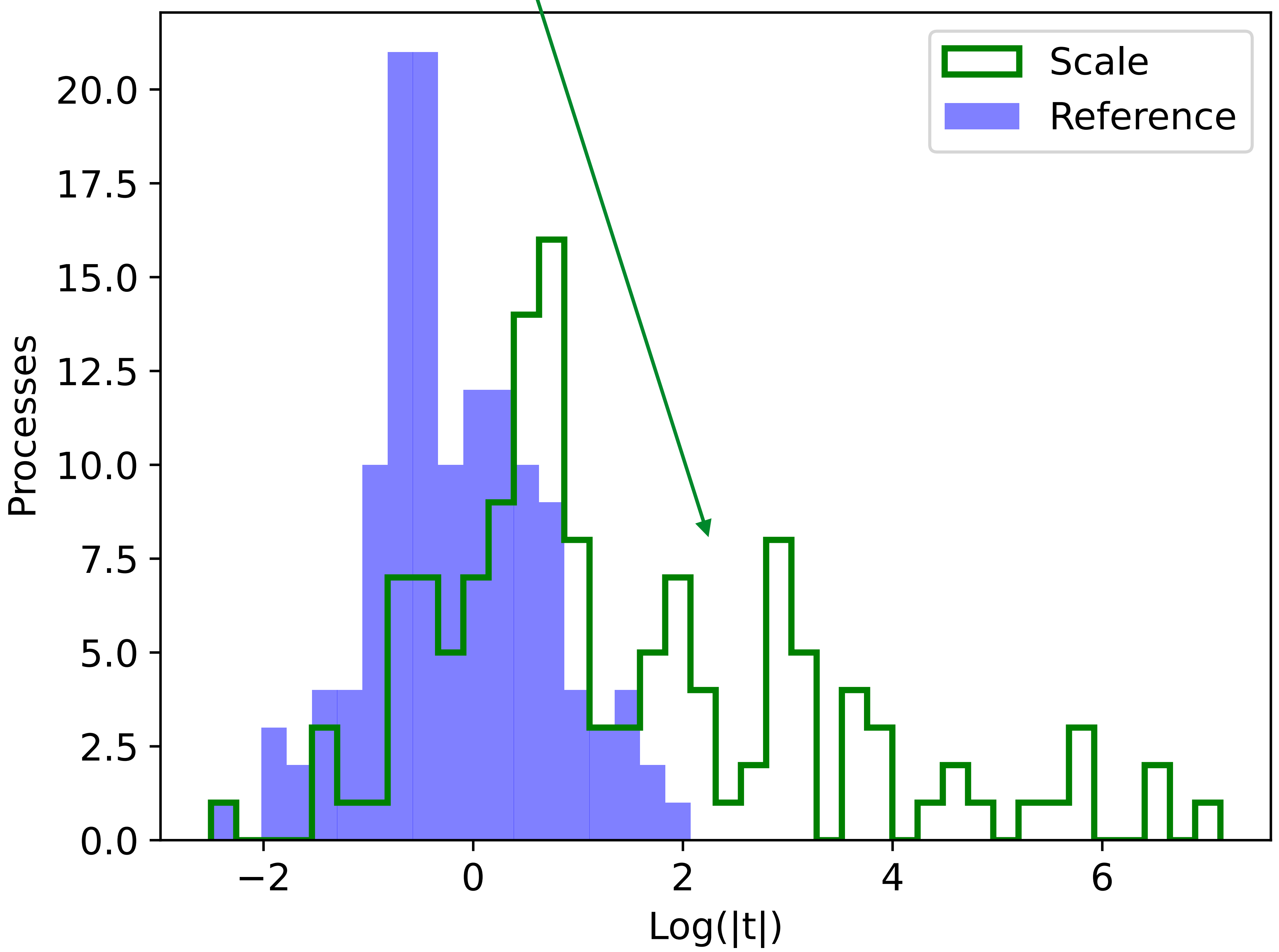
Tilman Plehn's 'reference process' method

Make correction in UQ for EW processes



Overflow bin

Much reduced tails



Tilman Plehn's 'reference process' method

Leaves us wanting more …

- Would be even more interesting to repeat study for NLO → NNLO, differential distributions
- Can we use ML to automatically find patterns of failure ?
- Why did we find a Gaussian-ish core ?
- Impact: A new method for cross-checks within experiment collaborations

Snowmass Whitepaper: Recommendations for the future

- Common language for uncertainty between ML and Physics communities
- Funding to test ML UQ methods for physics
- Create benchmark datasets for uncertainty tests
- Develop and study interpretability methods

arXiv:2208.03284: Thomas Y. Chen, Biprateep Dey, **Aishik Ghosh**, Michael Kagan, Brian Nord, Nesar Ramachandra

Snowmass Whitepaper: Recommendations for the future

arXiv > hep-ex > arXiv:2208.03284

High Energy Physics – Experiment

[Submitted on 5 Aug 2022 (v1), last revised 6 Sep 2022 (this version, v3)]

Interpretable Uncertainty Quantification in AI for HEP

Thomas Y. Chen, Biprateep Dey, Aishik Ghosh, Michael Kagan, Brian Nord, Nesar Ramachandra

Snowmass 2021: Summary of past work and future roadmap

- Common language for uncertainty between ML and Physics communities
- Funding to test ML UQ methods for physics
- Create benchmark datasets for uncertainty tests
- Develop and study interpretability methods

arXiv:2208.03284: Thomas Y. Chen, Biprateep Dey, **Aishik Ghosh**, Michael Kagan, Brian Nord, Nesar Ramachandra

Conclusion

- ML more sensitive to simulation artefacts → building better uncertainty quantification tools
- ML lets us better propagate experimental uncertainties and build analyses optimised for all possibilities: HEP, Astro
- Solutions have wider use cases
  - Tractable likelihoods
  - Optimise true objective with differentiable programming [Inferno, NEOS]
  - Uncertainty quantification of ML-simulators? [Performance metrics, Bayesian networks]
  - Learn physics from machine: Mapping ML into a human-readable space [CNN to EFPs]

And more cool solutions to come !

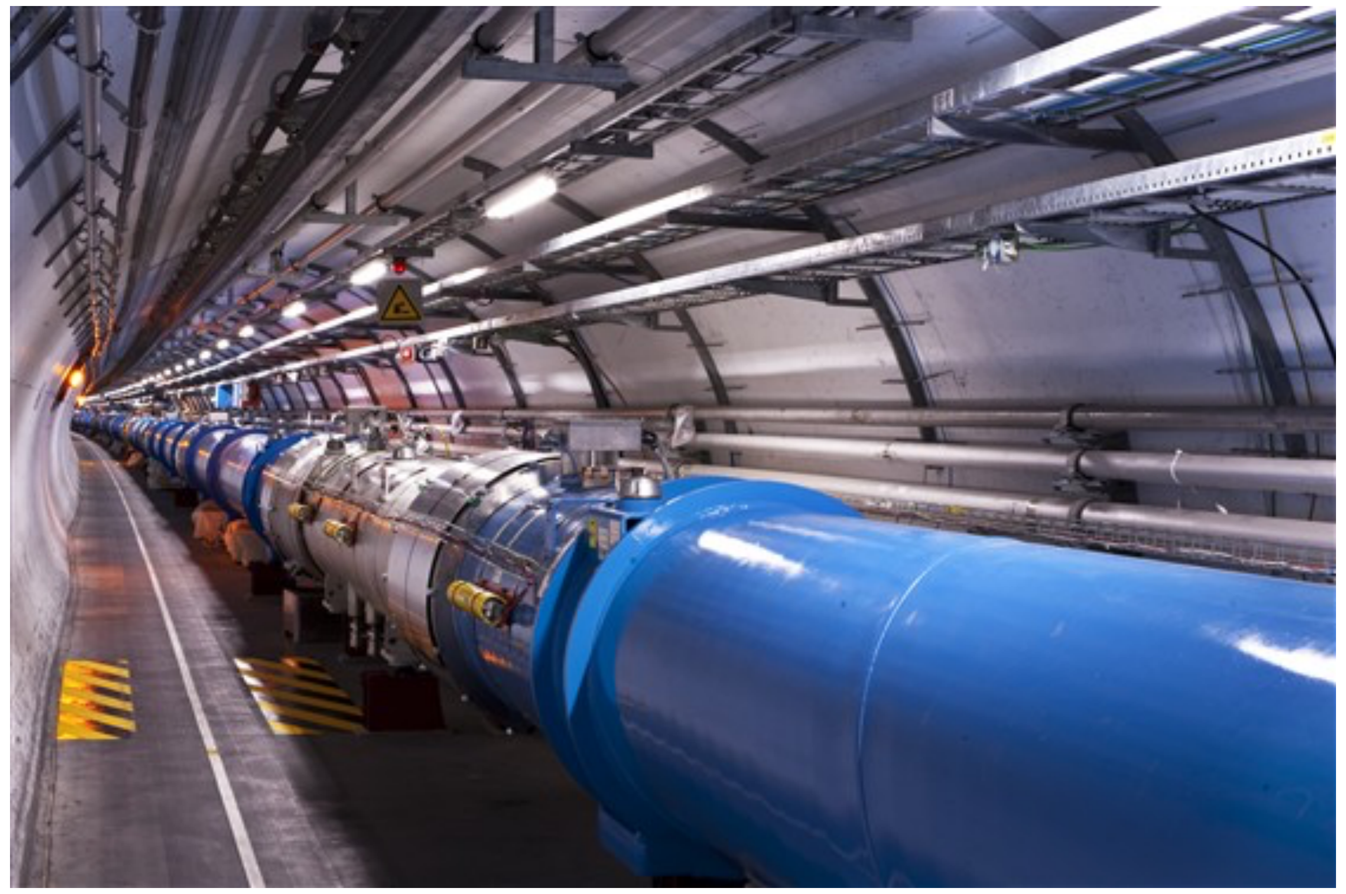
Thank you!

Known unknowns

Simulation using Standard Model of particle physics

Train ML models on simulation, apply on data

Unlabelled data from LHC



Detector state in sim: Z=1

Detector state in data: Z = ?

Systematic differences lead to systematic uncertainties 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

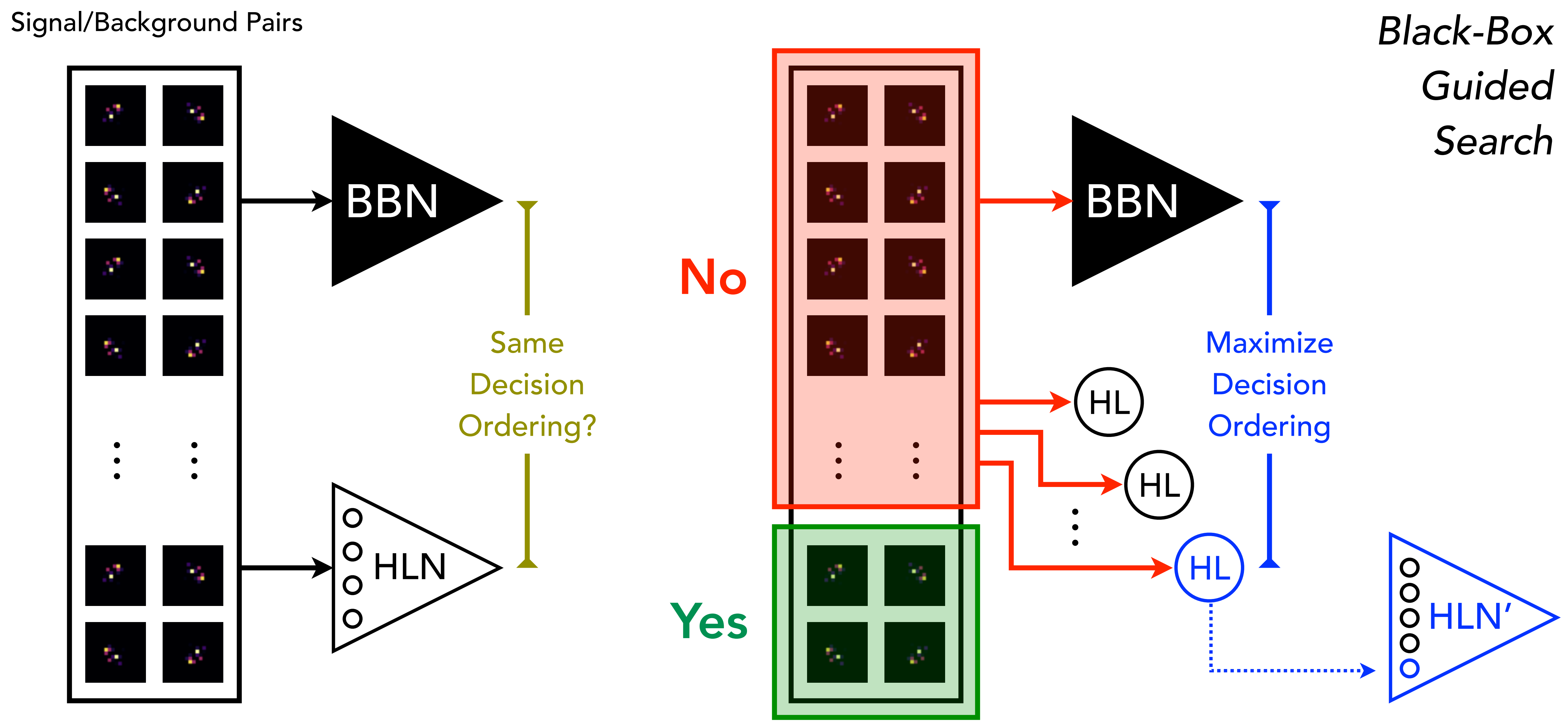
Tilman Plehn's 'reference process' method

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

Mapping machine-learned physics into a human-readable space



Rank	EFP	κ	β	Chrom #	ADO[EFP, CNN]$_{X_6}$	AUC[EFP]	ADO[6HL + EFP, CNN]$_{X_{\rm all}}$	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

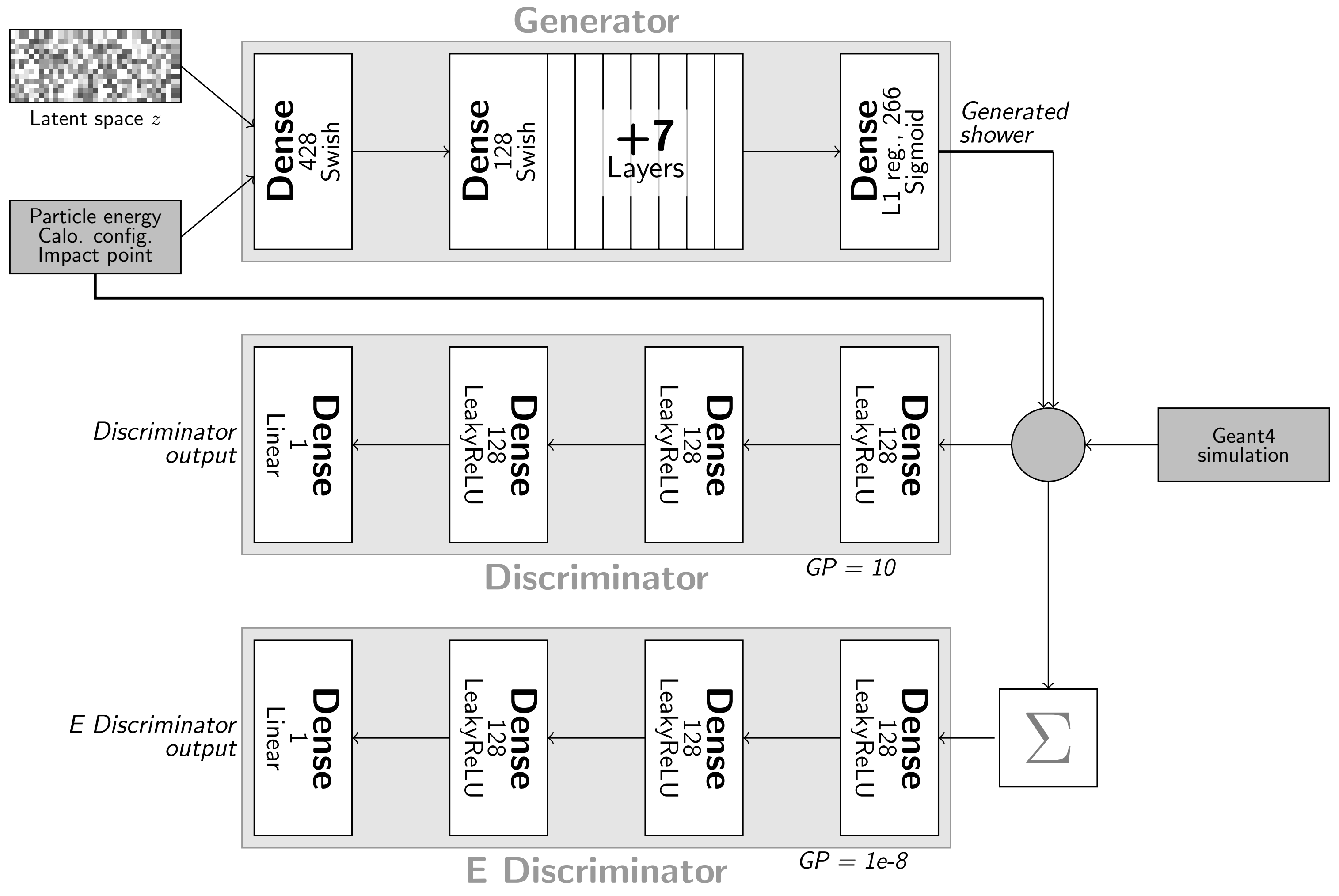
Faucett et al.

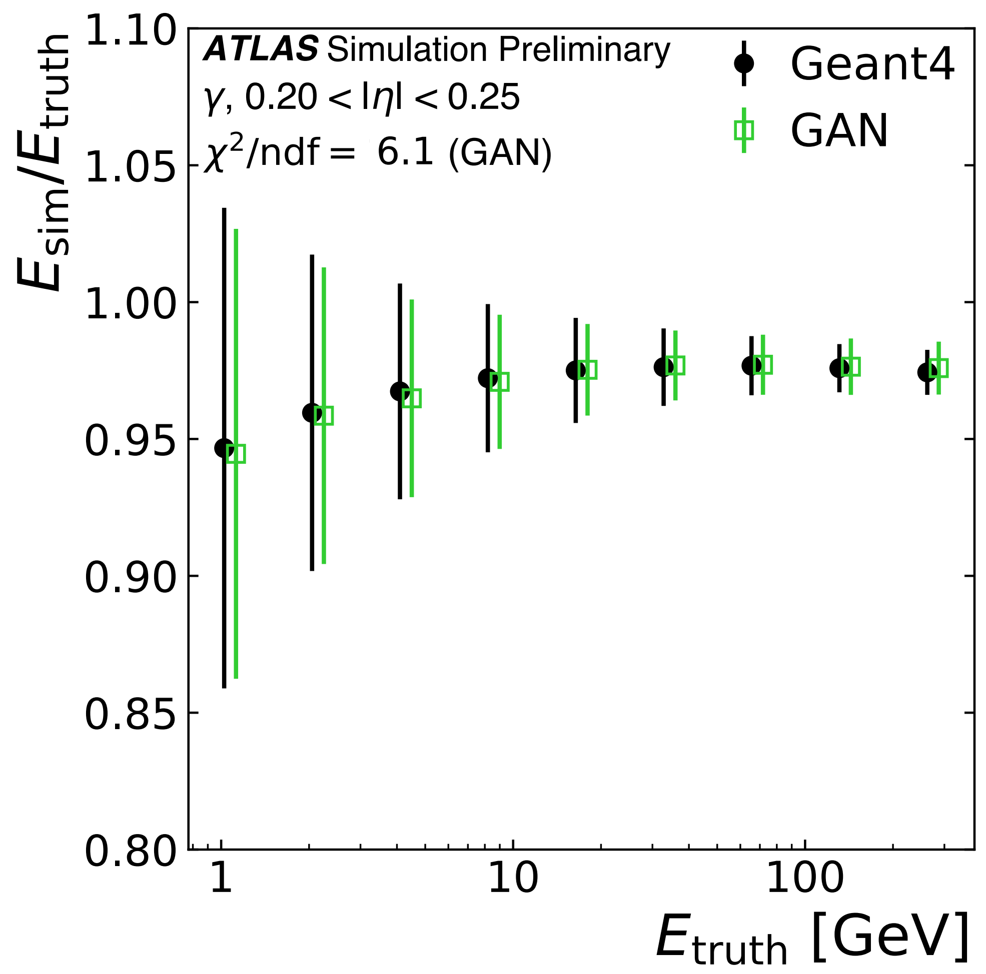
Performance Metrics for Generative Models

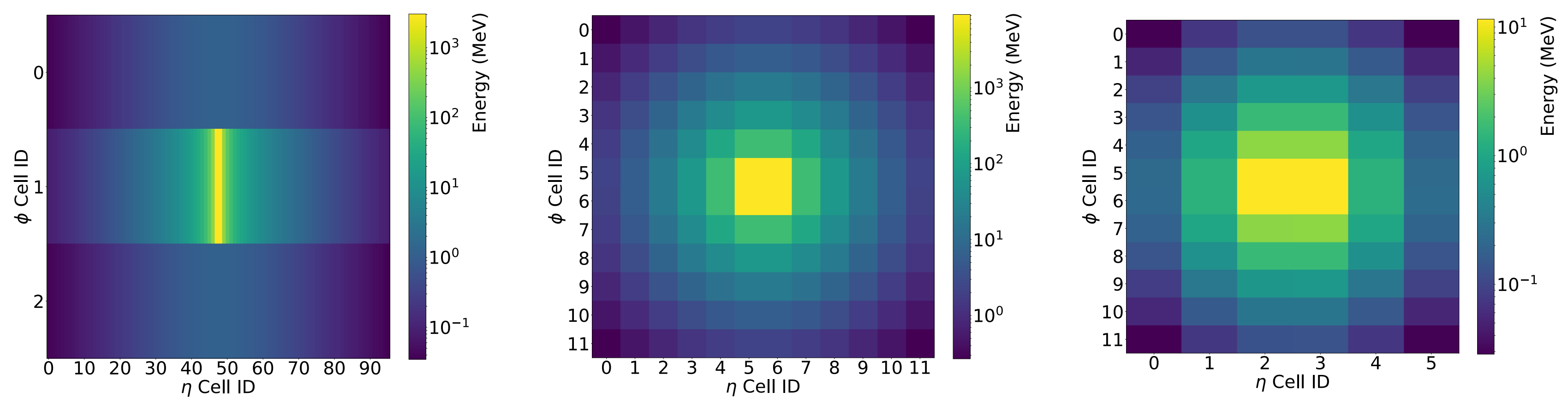
Generative Models for Simulation

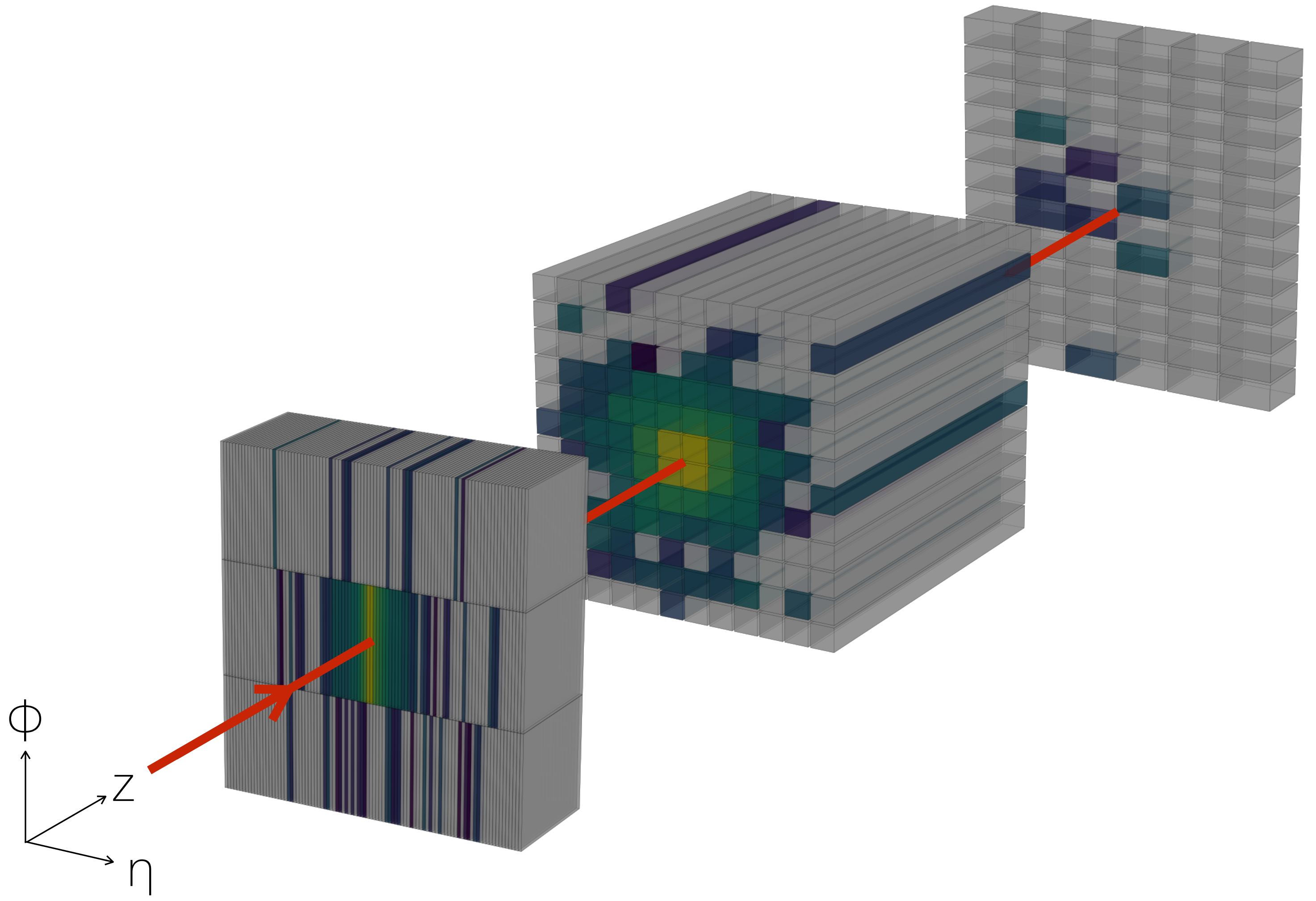
[Ghosh, ATLAS Collaboration, 2019](#)

[Paganini et al.](#)

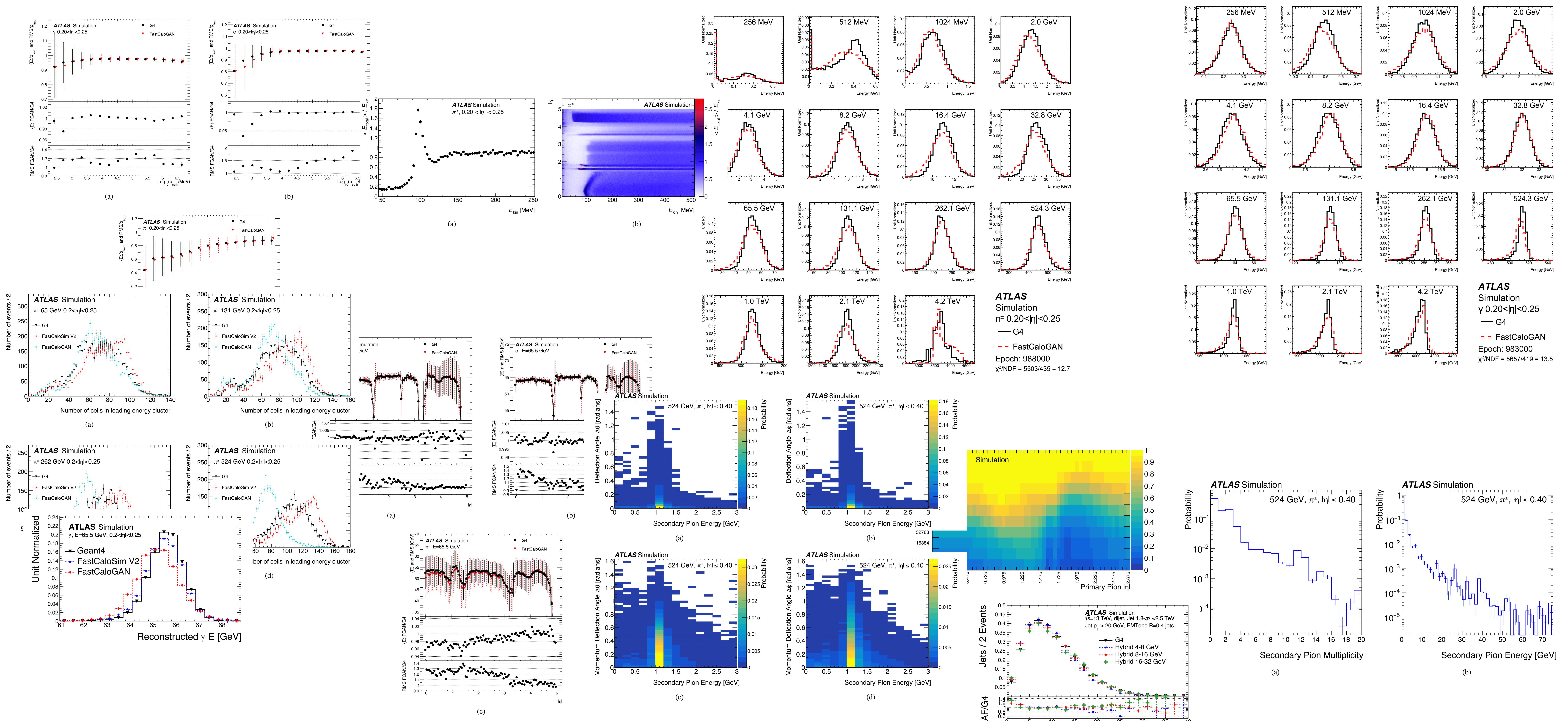








Evaluating Fast Calo Simulators



Can we automise the evaluation ?

Krause and Shih, 2021

5.4 Classifier metrics

In much of the GAN literature (see e.g. [8]), a common metric is to train classifiers to distinguish between different categories of data (e.g. e^+ vs. π^+), and to see if there is any difference in classifier performance when real data and generated data are interchanged. For example, one might train a classifier on e^+ vs. π^+ GEANT4 images, and compare this to a classifier trained on e^+ vs. π^+ GAN images. If the classifier trained on real images performs similarly to the classifier trained on generated images, then this is evidence that the generated images are approximating the real images well. One can repeat this test for different combinations of real and generated data.

The ultimate test of whether $p_{\text{generated}}(x) = p_{\text{data}}(x)$ would be a direct binary classifier between real and generated images of the *same* type. If the generated and true probability

Can we automise the evaluation ?

Krause and Shih, 2021

5.4 Classifier metrics

In much of the GAN literature (see e.g. [8]), a common metric is to train classifiers to distinguish between different categories of data (e.g. e^+ vs. π^+), and to see if there is any difference in classifier performance when real data and generated data are interchanged. For example, one might train a classifier on e^+ vs. π^+ GEANT4 images, and compare this to a classifier trained on e^+ vs. π^+ GAN images. If the classifier trained on real images performs similarly to the classifier trained on generated images, then this is evidence that the generated images are approximating the real images well. One can repeat this test for different combinations of real and generated data.

The ultimate test of whether $p_{\text{generated}}(x) = p_{\text{data}}(x)$ would be a direct binary classifier between real and generated images of the *same* type. If the generated and true probability

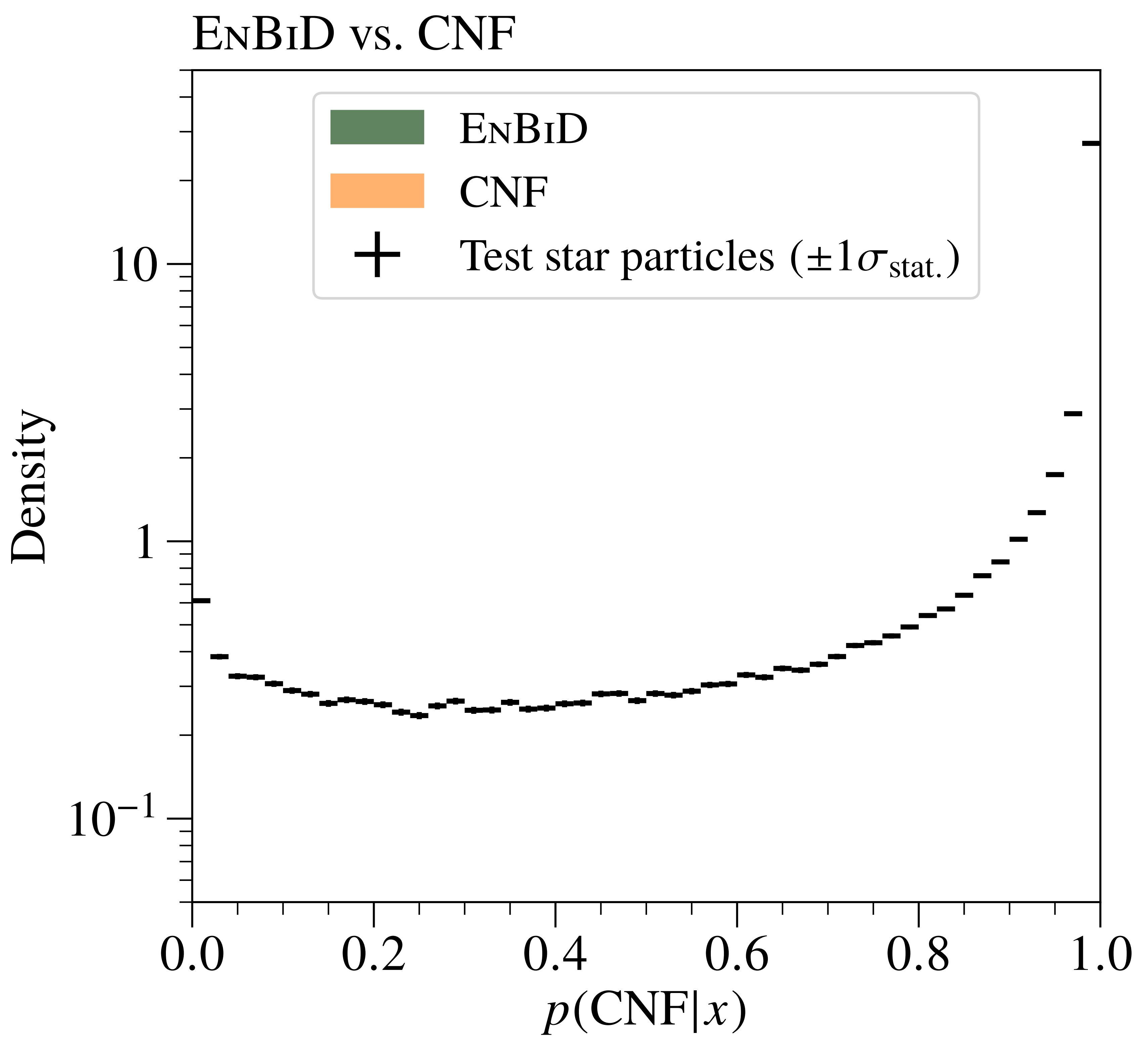
Classify Geant4 vs generated and use AUC as single metric

Another classifier test

[Lim et al, 2022]

Compare two generative models:

Classify generative model1 vs model2, check if test dataset agrees better with one or the other



A comparison of metrics

Kansal et al, 2022

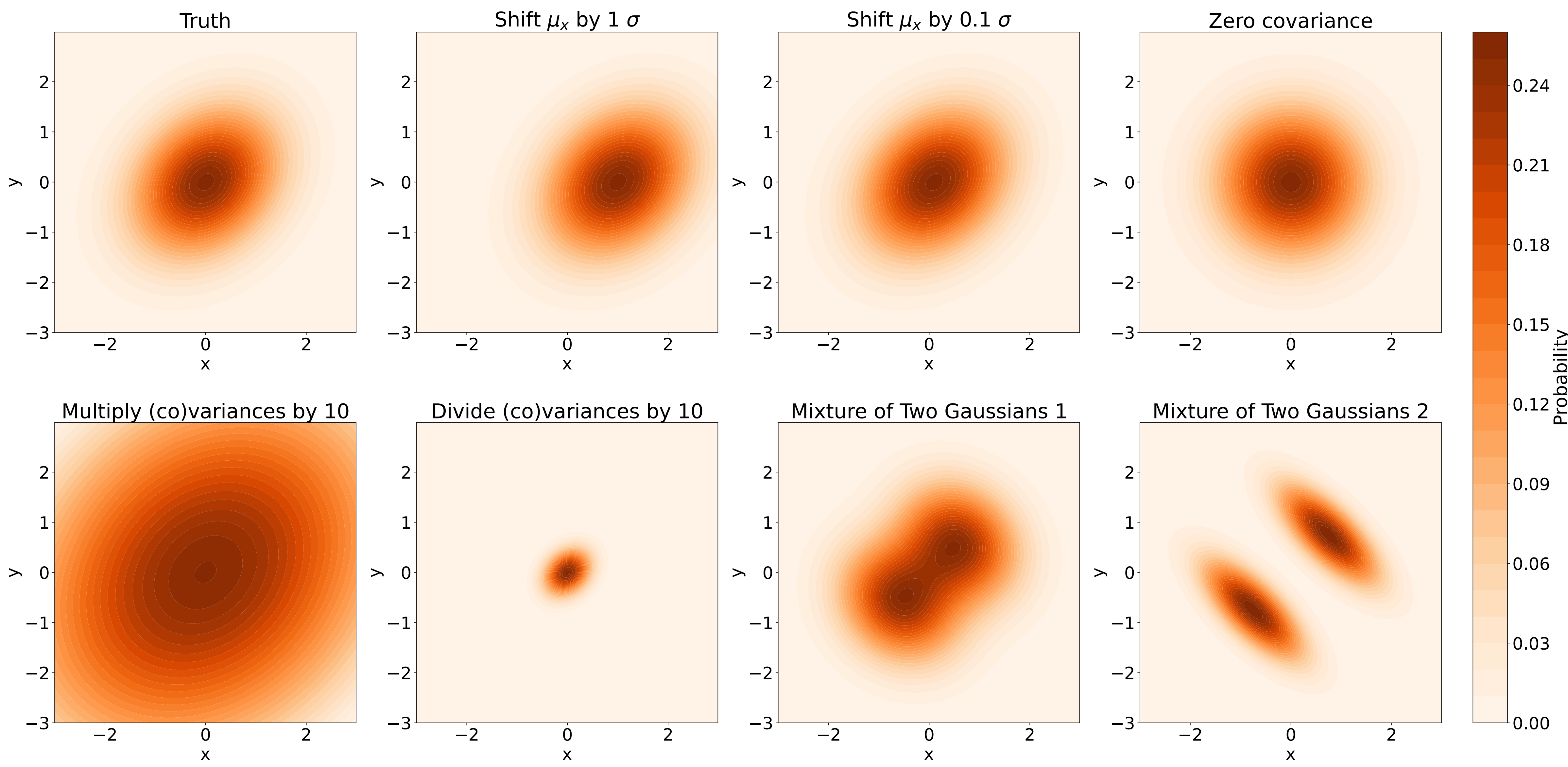
On the Evaluation of Generative Models in High Energy Physics

Raghav Kansal,* Anni Li, and Javier Duarte
University of California San Diego, La Jolla, CA 92093, USA

Nadezda Chernyavskaya, Maurizio Pierini
European Center for Nuclear Research (CERN), 1211 Geneva 23, Switzerland

Breno Orzari, Thiago Tomei
Universidade Estadual Paulista, São Paulo/SP, CEP 01049-010, Brazil

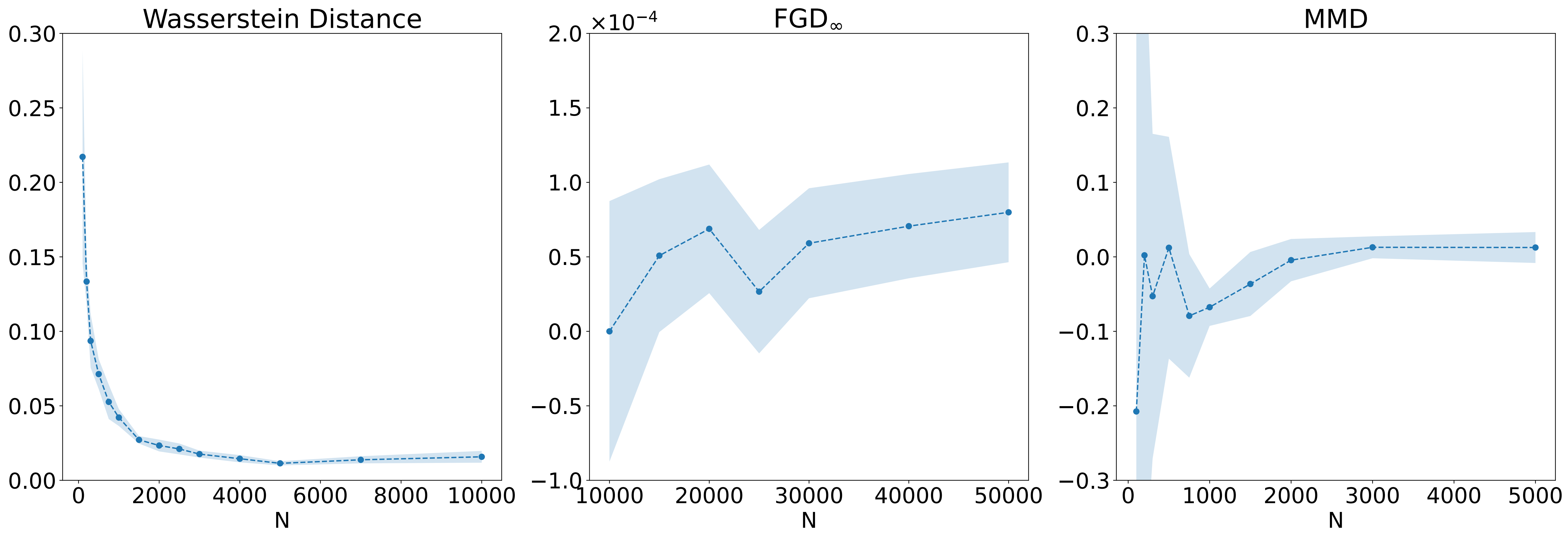
(Dated: November 21, 2022)



Detailed comparison on Gaussian toys where you have full control

Application on jet dataset with hand designed distortions

Gaussian Study



- FGD_∞, MMD unbiased
- W too expensive for large N

Metric	Truth	Shift μ_x by 1σ	Shift μ_x by 0.1σ	Zero covariance	Multiply (co)variances by 10	Divide (co)variances by 10	Mixture of Two Gaussians 1	Mixture of Two Gaussians 2
Wasserstein	0.016 ± 0.004	1.14 ± 0.02	0.043 ± 0.008	0.077 ± 0.006	9.8 ± 0.1	0.97 ± 0.01	$\mathbf{0.036 \pm 0.003}$	$\mathbf{0.191 \pm 0.005}$
FGD$_\infty$ $\times 10^3$	0.08 ± 0.03	$\mathbf{1011 \pm 1}$	$\mathbf{11.0 \pm 0.1}$	$\mathbf{32.3 \pm 0.2}$	9400 ± 8	$\mathbf{935.1 \pm 0.7}$	0.07 ± 0.03	0.03 ± 0.03
MMD	0.01 ± 0.02	16.4 ± 0.9	0.07 ± 0.04	0.40 ± 0.08	$\mathbf{19k \pm 1k}$	4.3 ± 0.1	0.06 ± 0.02	0.35 ± 0.03
Precision	0.972 ± 0.005	0.91 ± 0.01	0.976 ± 0.004	0.969 ± 0.006	0.34 ± 0.01	1.0 ± 0.0	0.975 ± 0.003	0.9976 ± 0.0007
Recall	0.997 ± 0.001	0.992 ± 0.003	0.997 ± 0.001	0.9976 ± 0.0006	0.998 ± 0.001	0.58 ± 0.02	0.996 ± 0.001	0.9970 ± 0.0009
Density	3.23 ± 0.06	2.48 ± 0.08	3.19 ± 0.07	3.1 ± 0.1	0.60 ± 0.02	5.7 ± 0.3	2.99 ± 0.09	0.989 ± 0.009
Coverage	0.876 ± 0.002	0.780 ± 0.006	0.872 ± 0.005	0.872 ± 0.004	0.60 ± 0.01	0.406 ± 0.008	0.871 ± 0.002	0.956 ± 0.006

FGD_∞ most promising
(but no sensitivity to higher moments, requires extrapolation)

Jet Study

Kansal et al, 2022

Metric	Truth	Smeared	Shifted	Removing tail	Particle features smeared	Particle η^{rel} smeared	Particle $p_{\mathrm{T}}^{\mathrm{rel}}$ smeared	Particle $p_{\mathrm{T}}^{\mathrm{rel}}$ shifted
$W_1^M \times 10^3$	0.28 ± 0.05	2.1 ± 0.2	6.0 ± 0.3	0.6 ± 0.2	1.7 ± 0.2	0.9 ± 0.3	0.5 ± 0.2	5.8 ± 0.2
Wasserstein EFP	0.02 ± 0.01	0.09 ± 0.05	0.10 ± 0.02	0.016 ± 0.007	0.19 ± 0.08	0.03 ± 0.01	0.03 ± 0.02	0.06 ± 0.02
FGD_∞ EFP $\times 10^3$	0.01 ± 0.02	$\mathbf{21.5 \pm 0.3}$	$\mathbf{26.8 \pm 0.3}$	$\mathbf{2.31 \pm 0.07}$	23.4 ± 0.3	$\mathbf{3.59 \pm 0.09}$	2.29 ± 0.05	28.9 ± 0.2
MMD EFP $\times 10^3$	-0.006 ± 0.005	0.17 ± 0.06	0.9 ± 0.1	0.03 ± 0.02	0.35 ± 0.09	0.08 ± 0.05	0.01 ± 0.02	1.8 ± 0.1
Precision EFP	0.9 ± 0.1	0.94 ± 0.04	0.978 ± 0.005	0.88 ± 0.08	0.7 ± 0.1	0.94 ± 0.06	0.7 ± 0.1	0.79 ± 0.09
Recall EFP	0.9 ± 0.1	0.88 ± 0.07	0.97 ± 0.01	0.92 ± 0.06	0.83 ± 0.05	0.92 ± 0.07	0.8 ± 0.1	0.8 ± 0.1
Wasserstein PN	1.65 ± 0.06	1.7 ± 0.1	2.4 ± 0.4	1.71 ± 0.08	4.5 ± 0.1	1.79 ± 0.05	4.0 ± 0.4	7.6 ± 0.2
FGD_∞ PN $\times 10^3$	0.8 ± 0.7	40 ± 2	193 ± 9	5.0 ± 0.9	$\mathbf{1250 \pm 10}$	20 ± 1	$\mathbf{1230 \pm 10}$	$\mathbf{3640 \pm 10}$
MMD PN $\times 10^3$	-2 ± 2	4 ± 8	80 ± 10	-1 ± 4	500 ± 100	3 ± 2	560 ± 60	1100 ± 40
Precision PN	0.68 ± 0.07	0.64 ± 0.04	0.71 ± 0.06	0.73 ± 0.03	0.09 ± 0.04	0.75 ± 0.08	0.08 ± 0.04	0.39 ± 0.08
Recall PN	0.70 ± 0.05	0.61 ± 0.04	0.61 ± 0.08	0.73 ± 0.06	0.014 ± 0.009	0.7 ± 0.1	0.01 ± 0.01	0.57 ± 0.09
Classifier LLF AUC	0.50	0.52	0.54	0.50	0.97	0.81	0.93	0.99
Classifier HLF AUC	0.50	0.53	0.55	0.50	0.84	0.64	0.74	0.92

- FGD_∞ on EFPs does quite well in these tests
- Would be interesting to see it used and stress tested !

Bayesian Generative Models

Bayesian Generative Models

Butter et al.

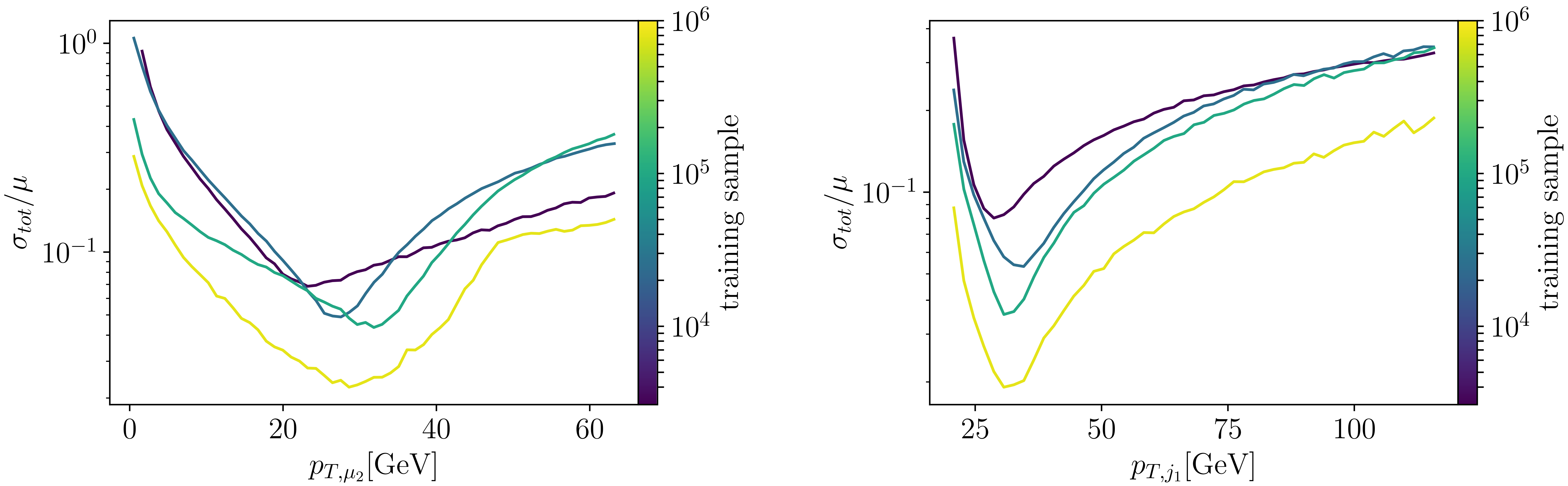


Figure 12: Relative uncertainty from the BINN for the $Z + 1$ jet sample, as a function of the size of the training sample.

Other uncertainty methods

Differentiable Programming: Optimise your final objective directly

Simpson et al.

Following Inferno [de Castro et al.]

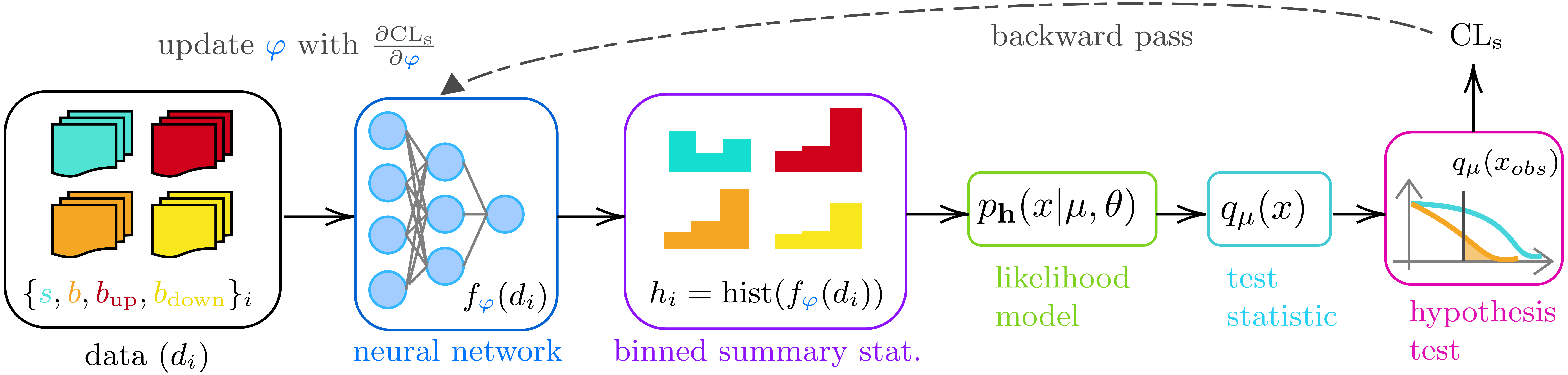


Figure 1. The pipeline for `neos`. The dashed line indicating the backward pass involves updating the weights φ of the neural network via gradient descent.

Unfolding with nuisance parameters

Chan and Nachman arXiv:2302.05390

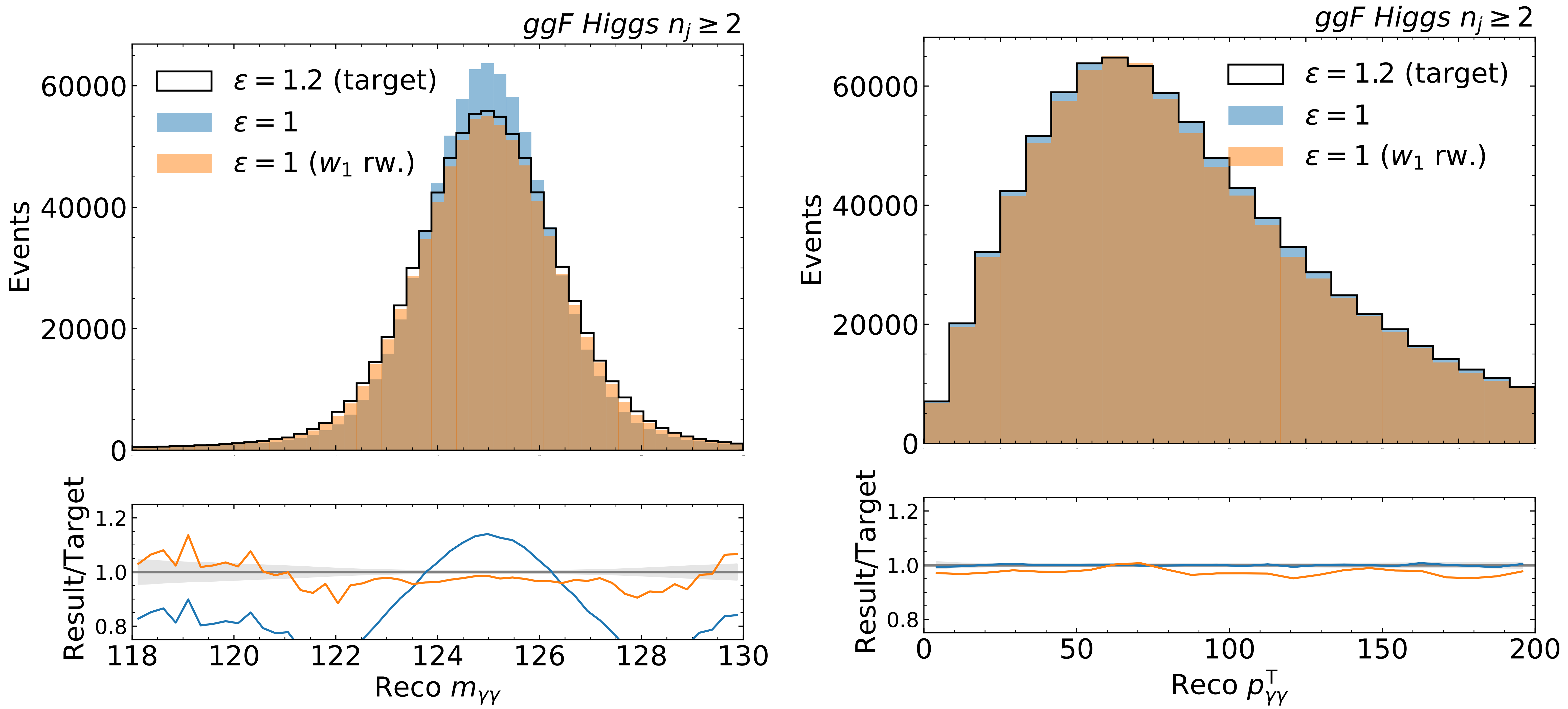
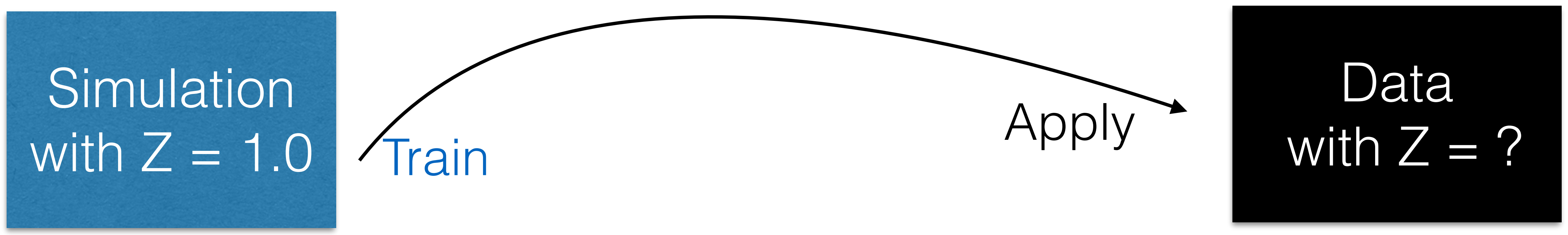


FIG. 6. Higgs boson cross section: the nominal detector-level spectra $m_{\gamma\gamma}$ (left) and $p^{\mathrm{T}}_{\gamma\gamma}$ (right) with $\epsilon_\gamma = 1$ reweighted by the trained w_1 conditioned at $\epsilon_\gamma = 1.2$ and compared to the spectra with $\epsilon_\gamma = 1.2$.

More on uncertainty-aware networks

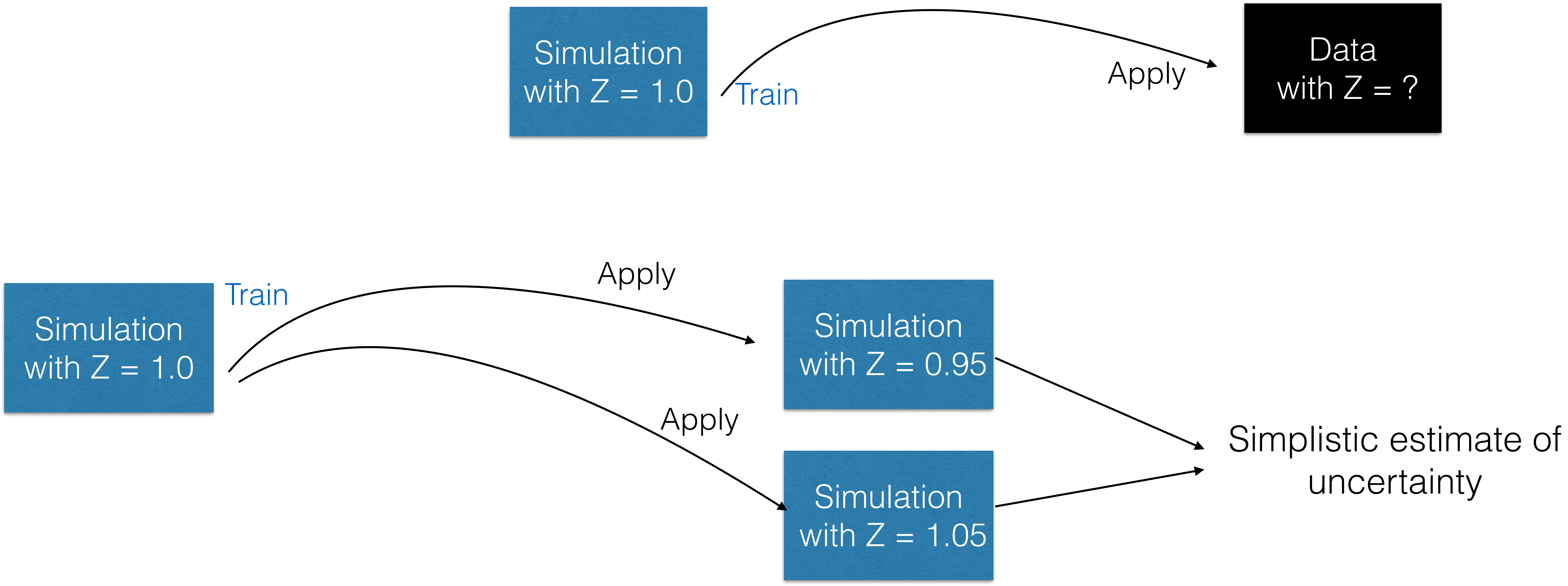
Baseline Approach to Uncertainty Quantification

Train AI classifier on nominal data (assume detector state Z=1) and estimate uncertainties using alternate simulations



Baseline Approach to Uncertainty Quantification

Train AI classifier on nominal data (assume detector state Z=1) and estimate uncertainties using alternate simulations



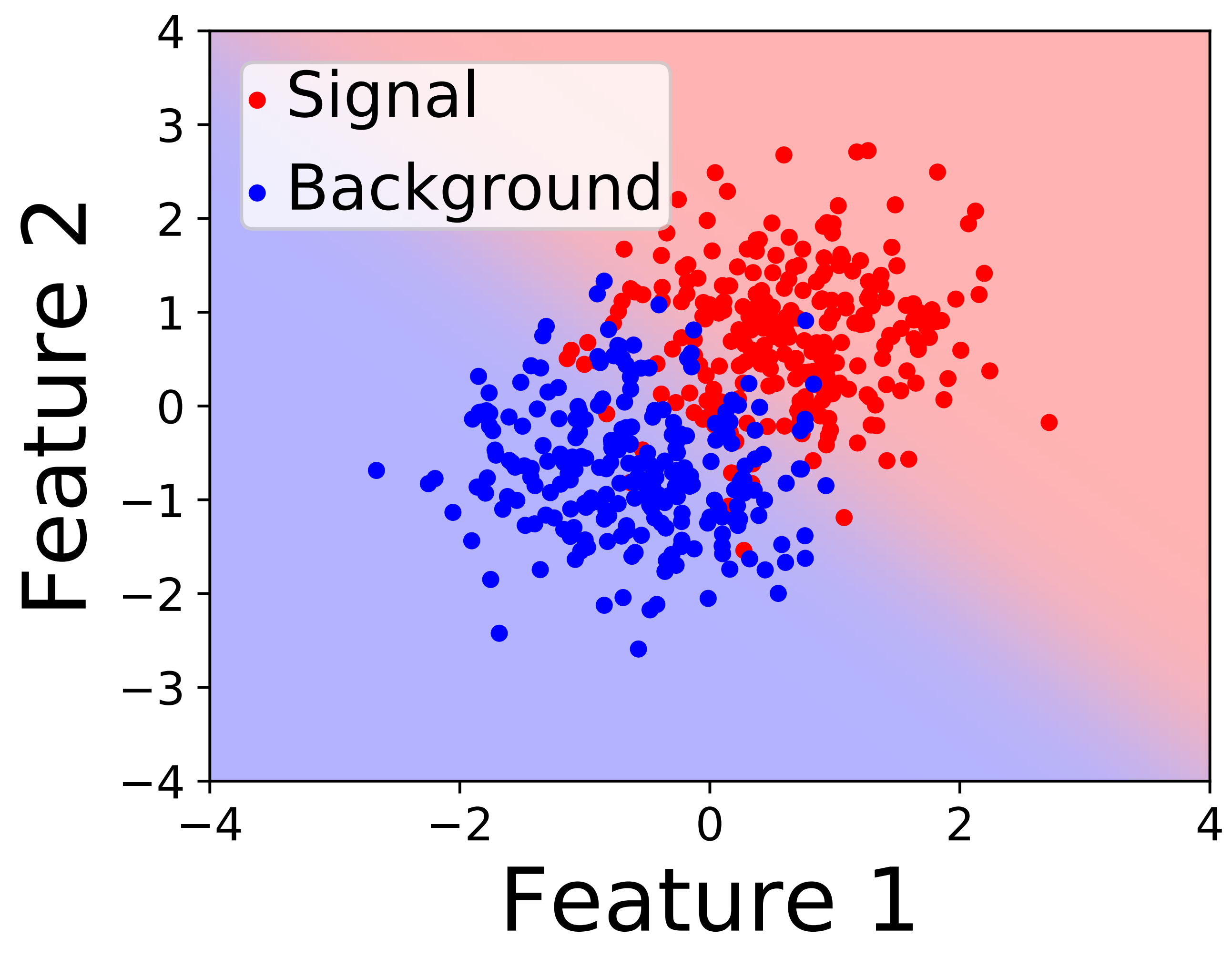
Full statistical treatment → Expensive 'Profile Likelihood'

Nominal and Systematic Up Examples

Baseline Classifier

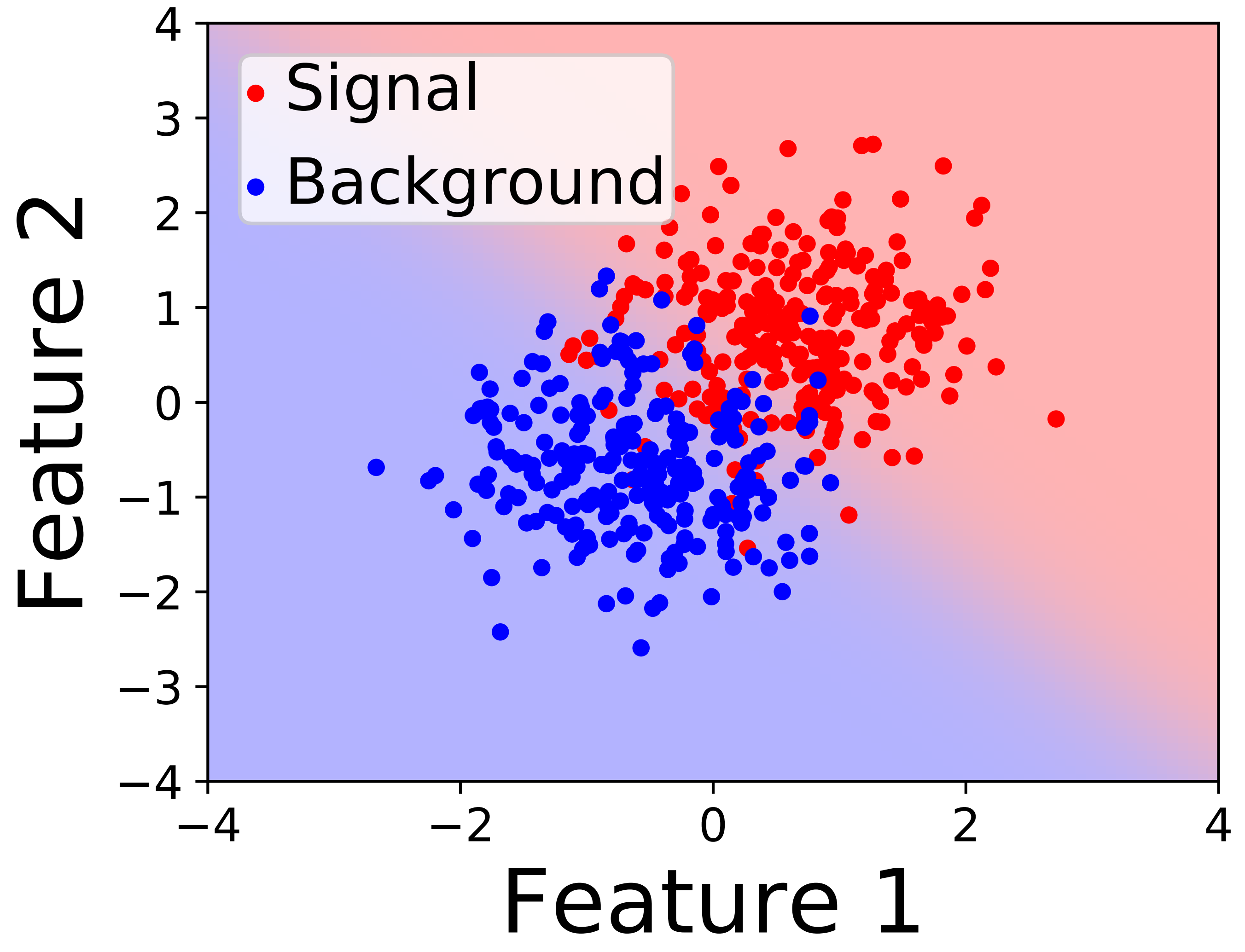
Uncertainty-Aware Classifier

Nominal "Data"



AUC=0.978

Optimal

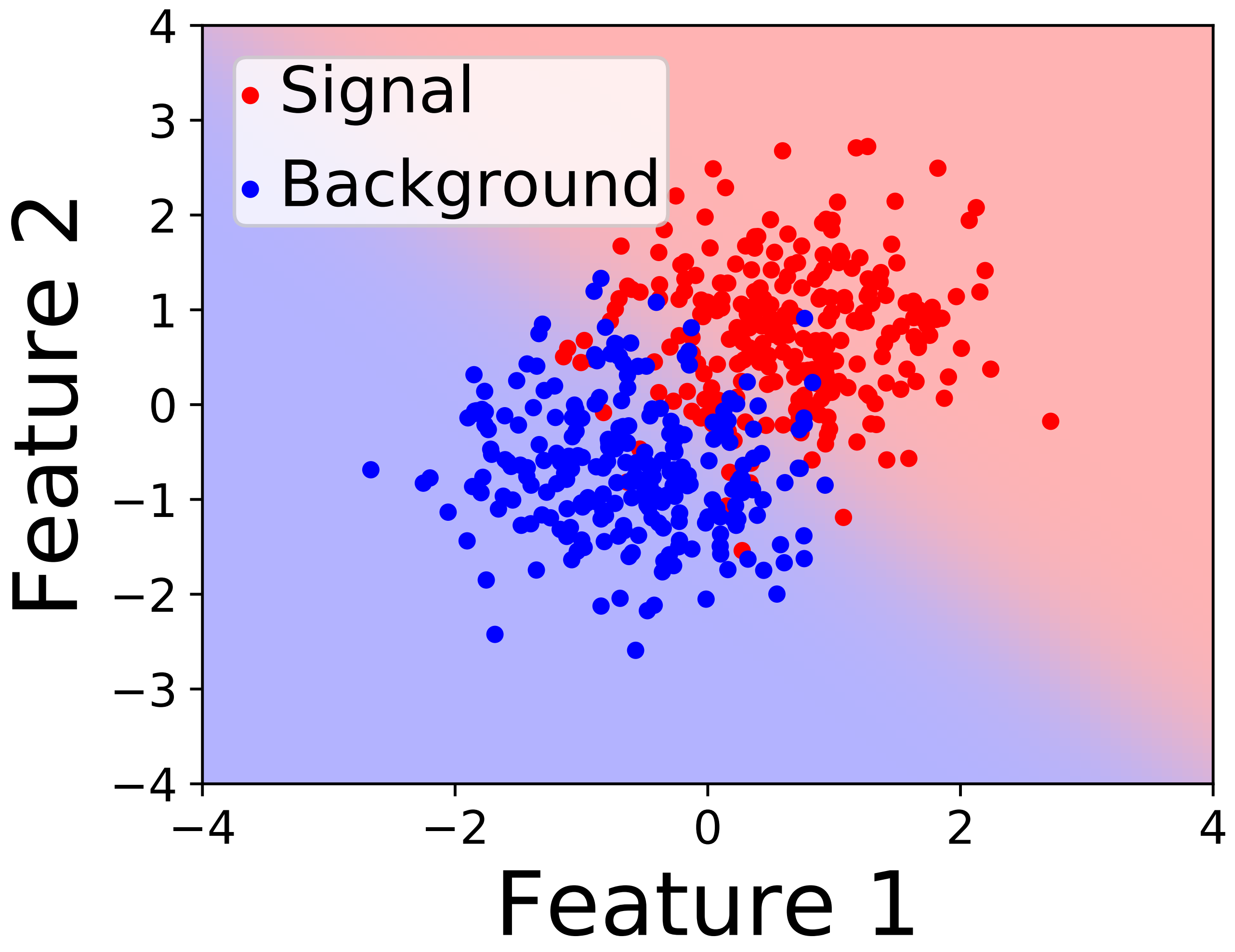
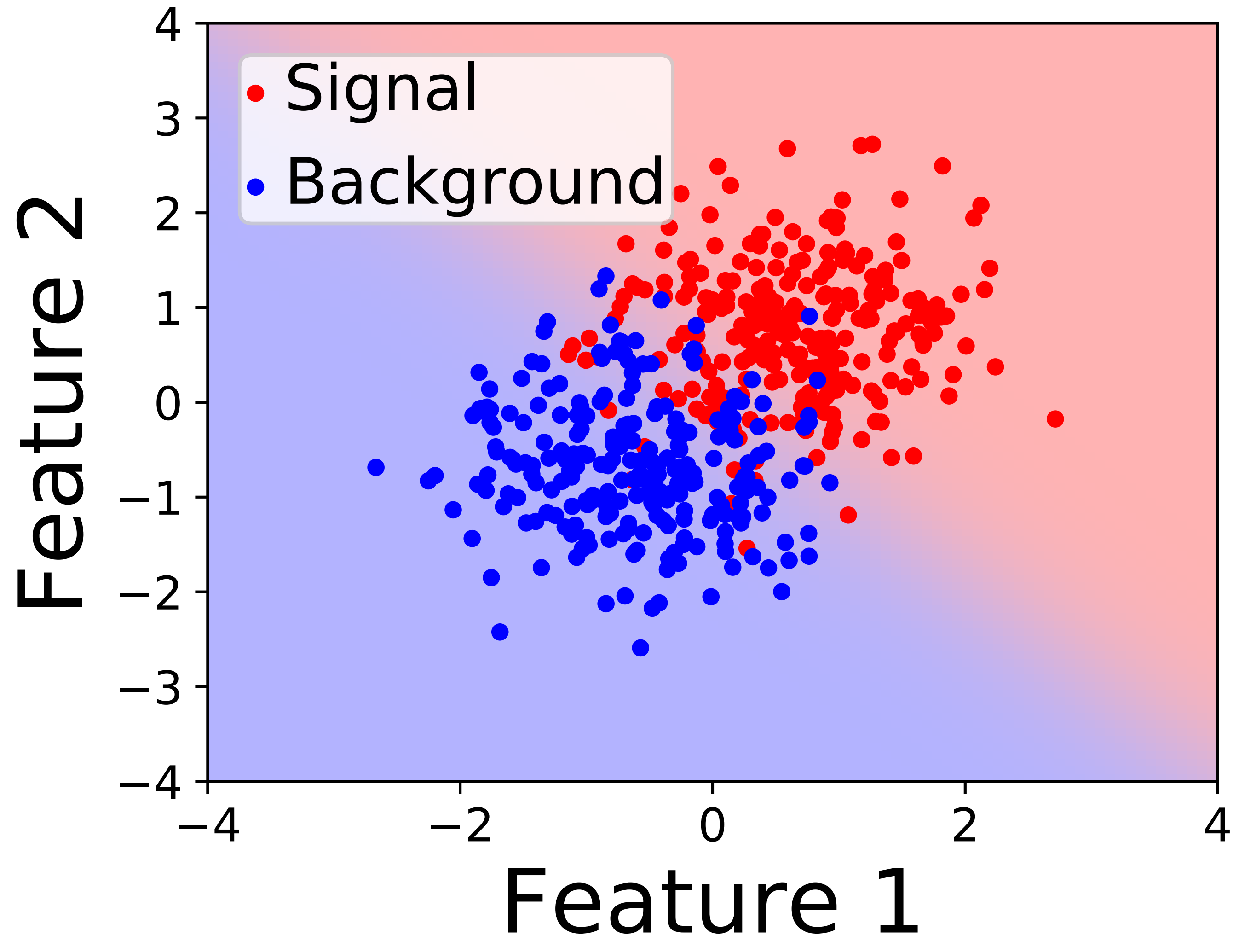
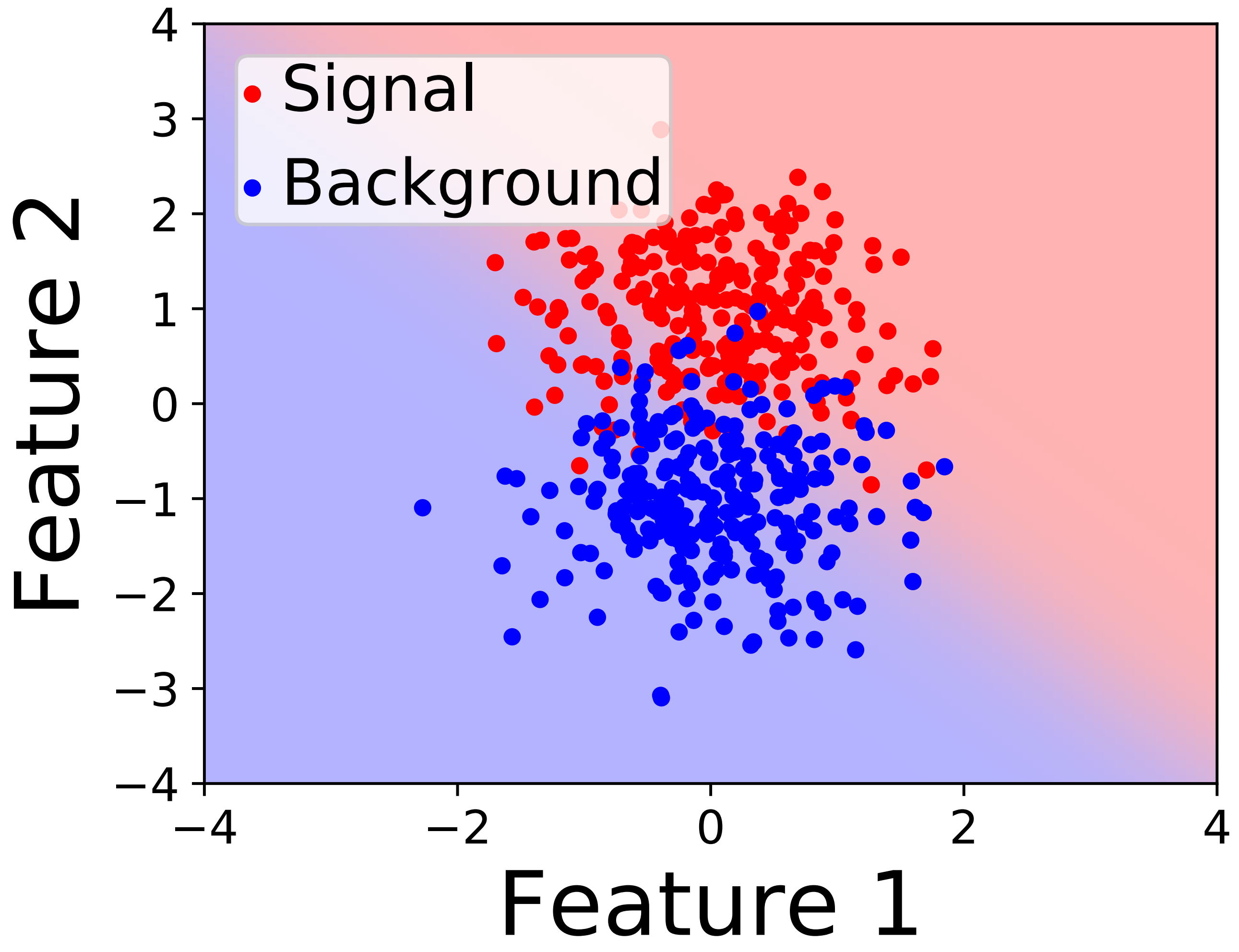


AUC=0.978

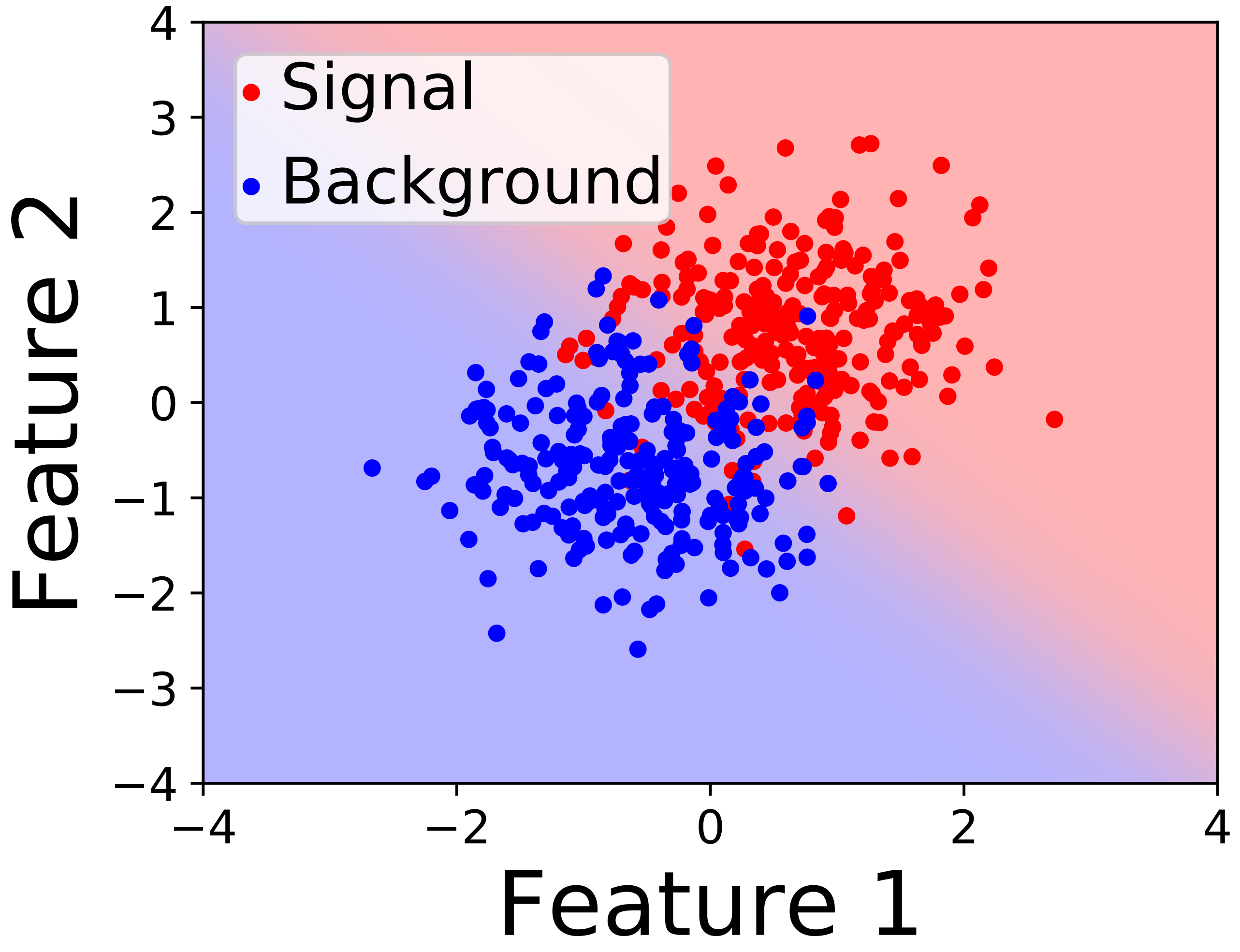
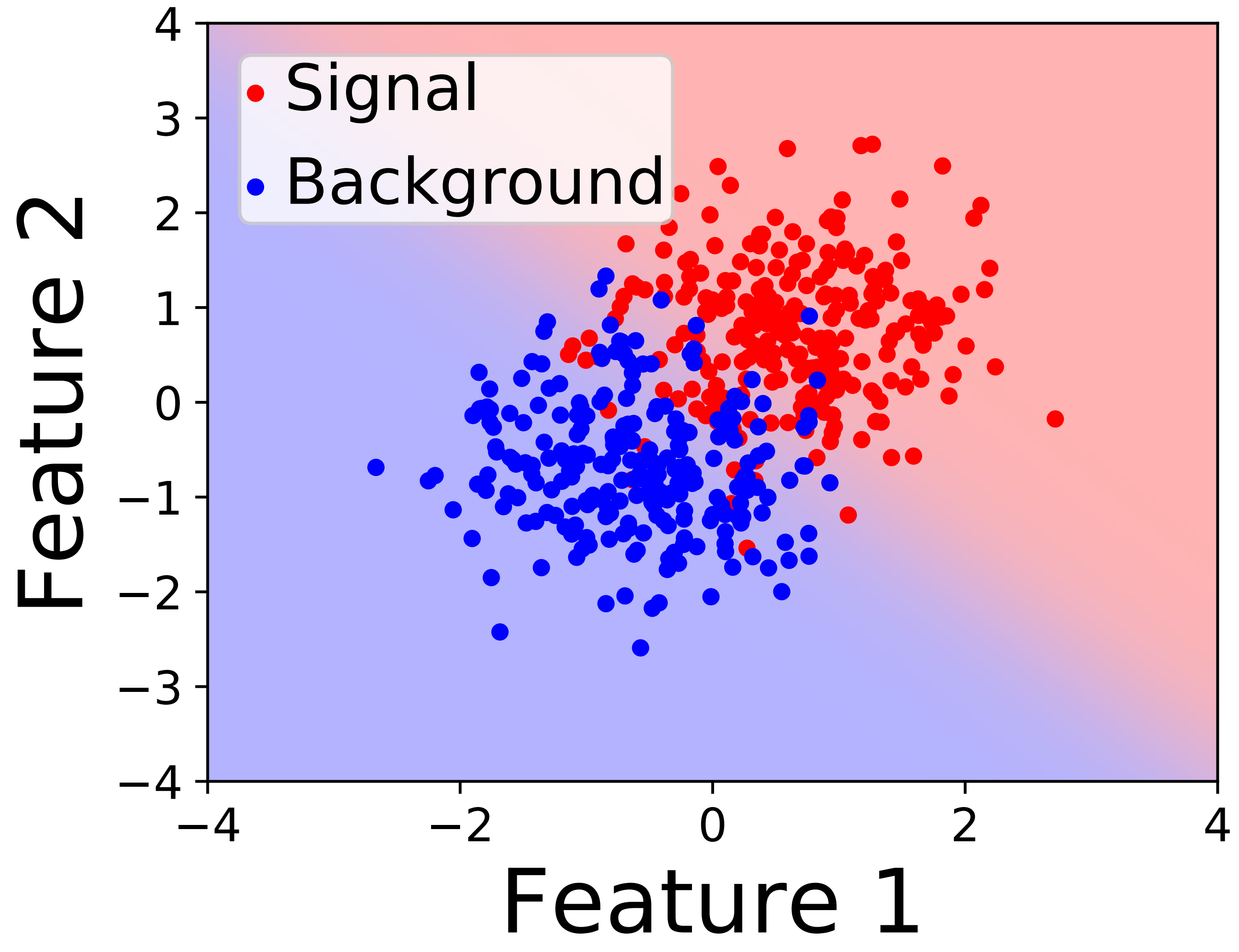
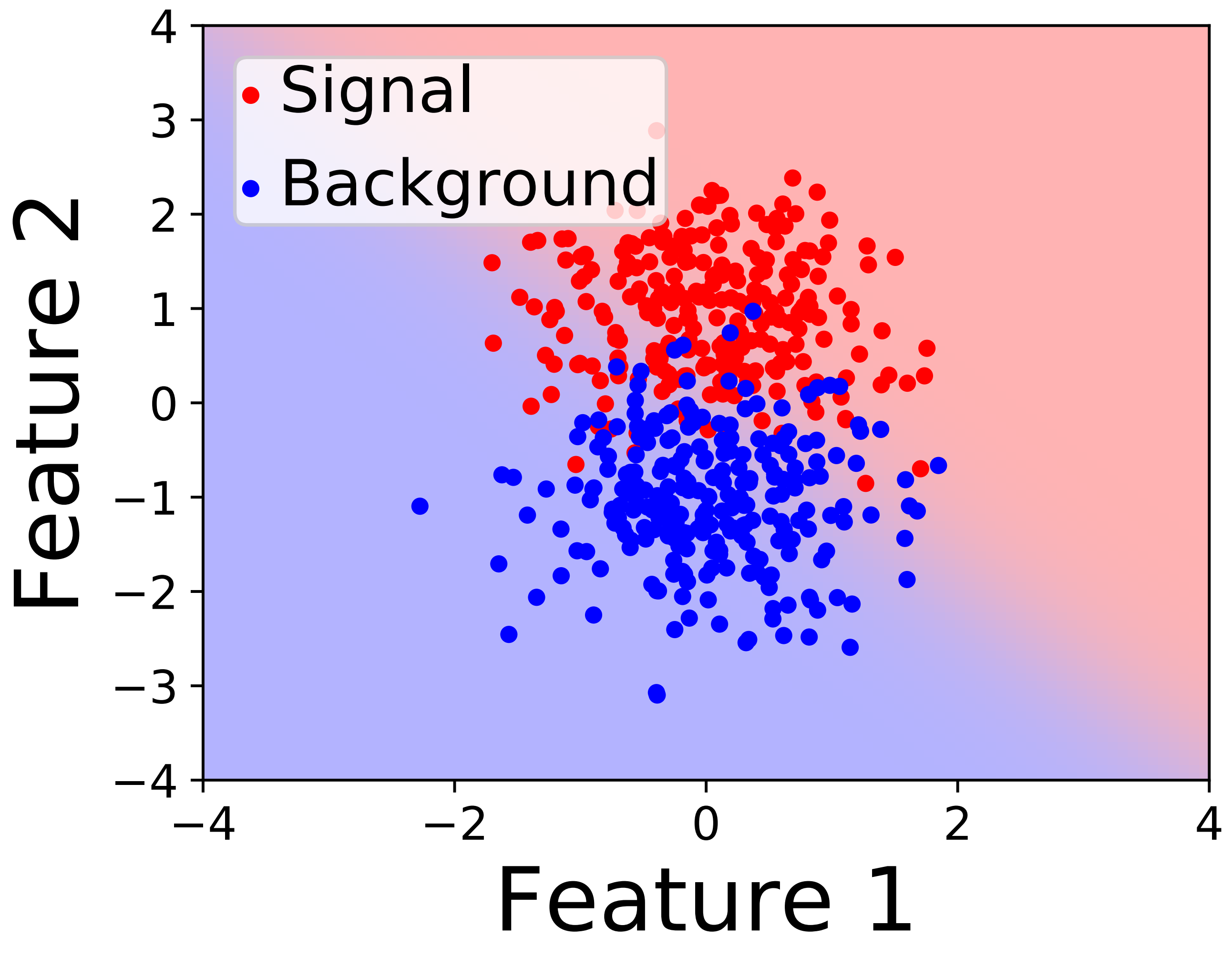
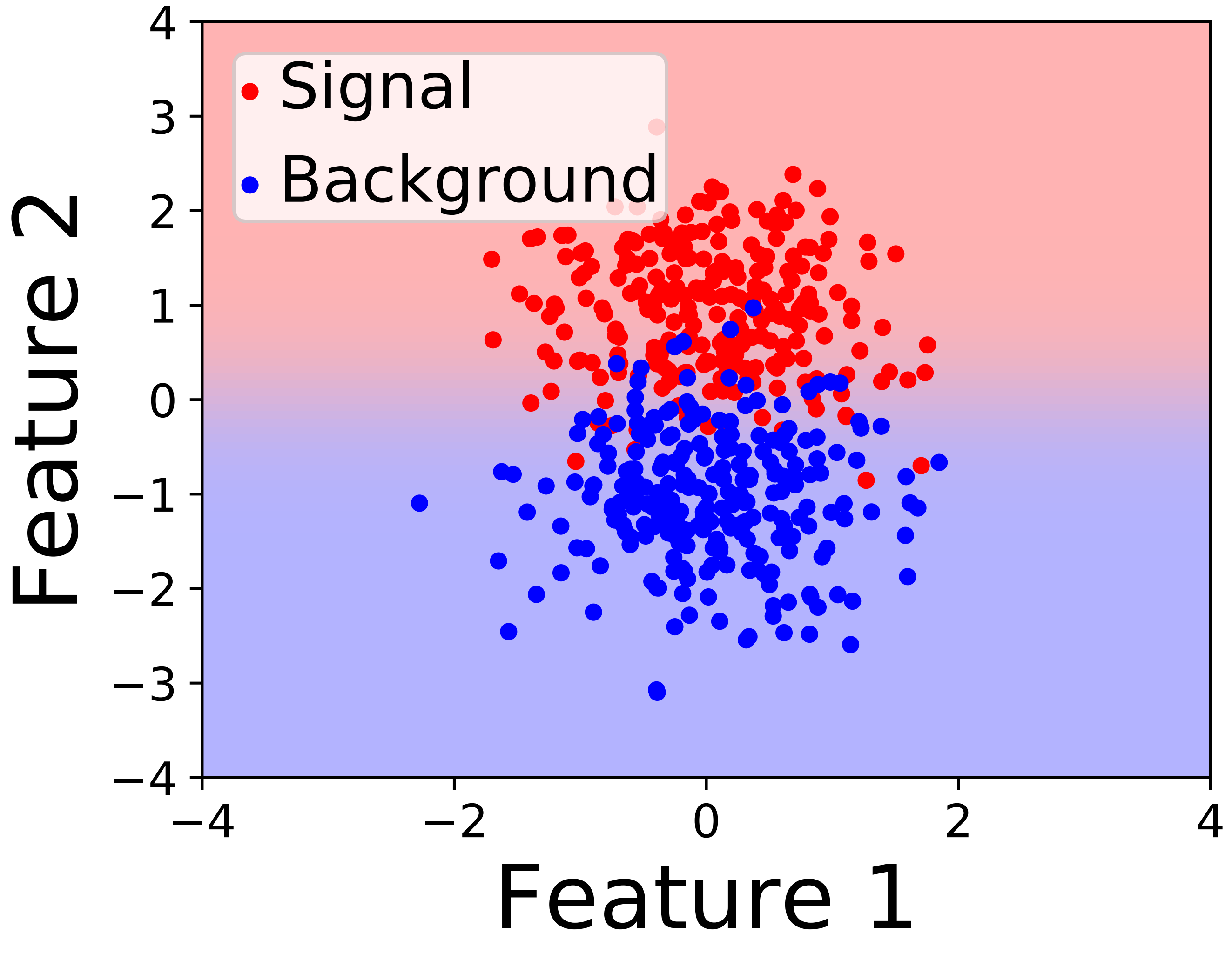
Optimal

SystUp "Data"

Nominal and Systematic Up Examples

	Baseline Classifier		Uncertainty-Aware Classifier	
Nominal "Data"		AUC=0.978 Optimal		AUC=0.978 Optimal
SystUp "Data"		AUC=0.924 **Sub-Optimal**		

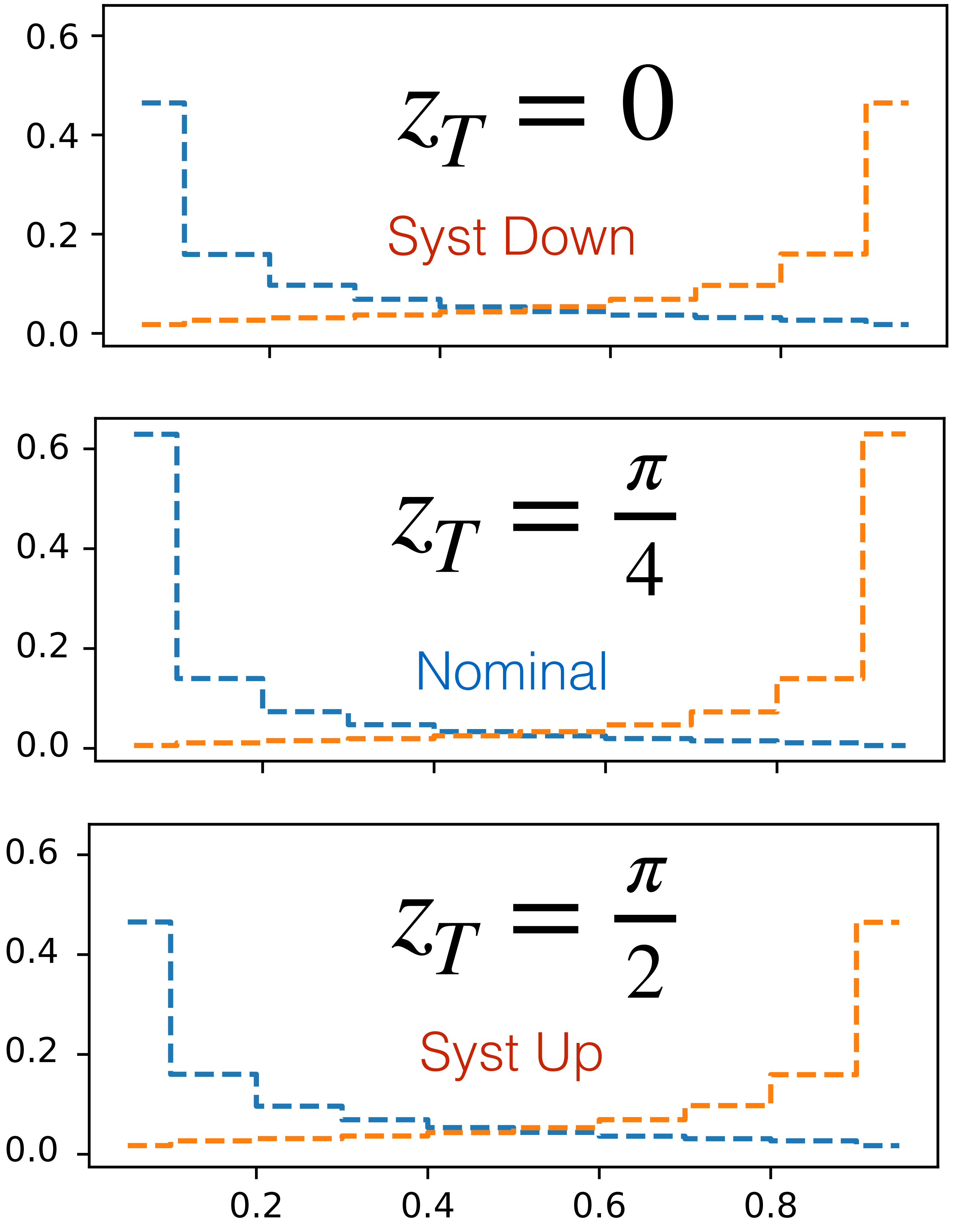
Nominal and Systematic Up Examples

	Baseline Classifier		Uncertainty-Aware Classifier	
Nominal "Data"		AUC=0.978 Optimal		AUC=0.978 Optimal
SystUp "Data"		AUC=0.924 **Sub-Optimal**		AUC=0.978 **Optimal**

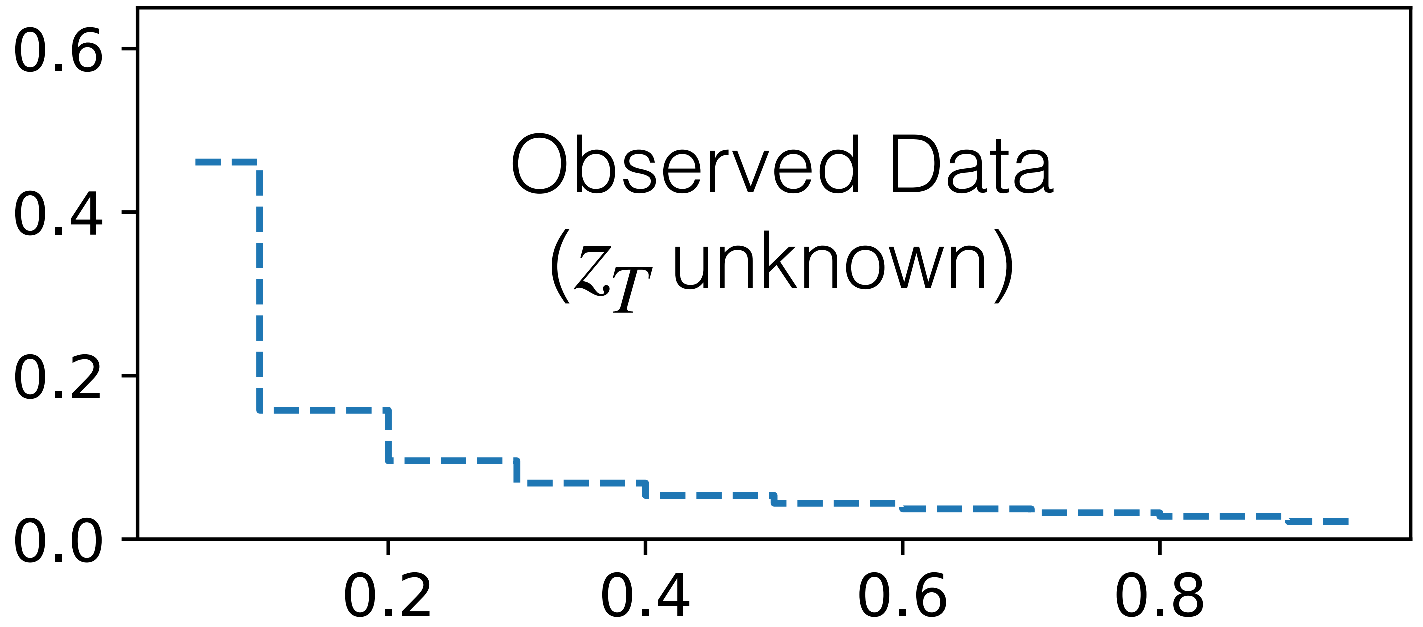
Uncer-Aware Classifier is able to rotate its decision function based on Z while the Baseline Classifier decision function remains frozen

Scan the 2D Likelihood space in Z vs μ

Template **Baseline Classifier** Score Histograms for various Z

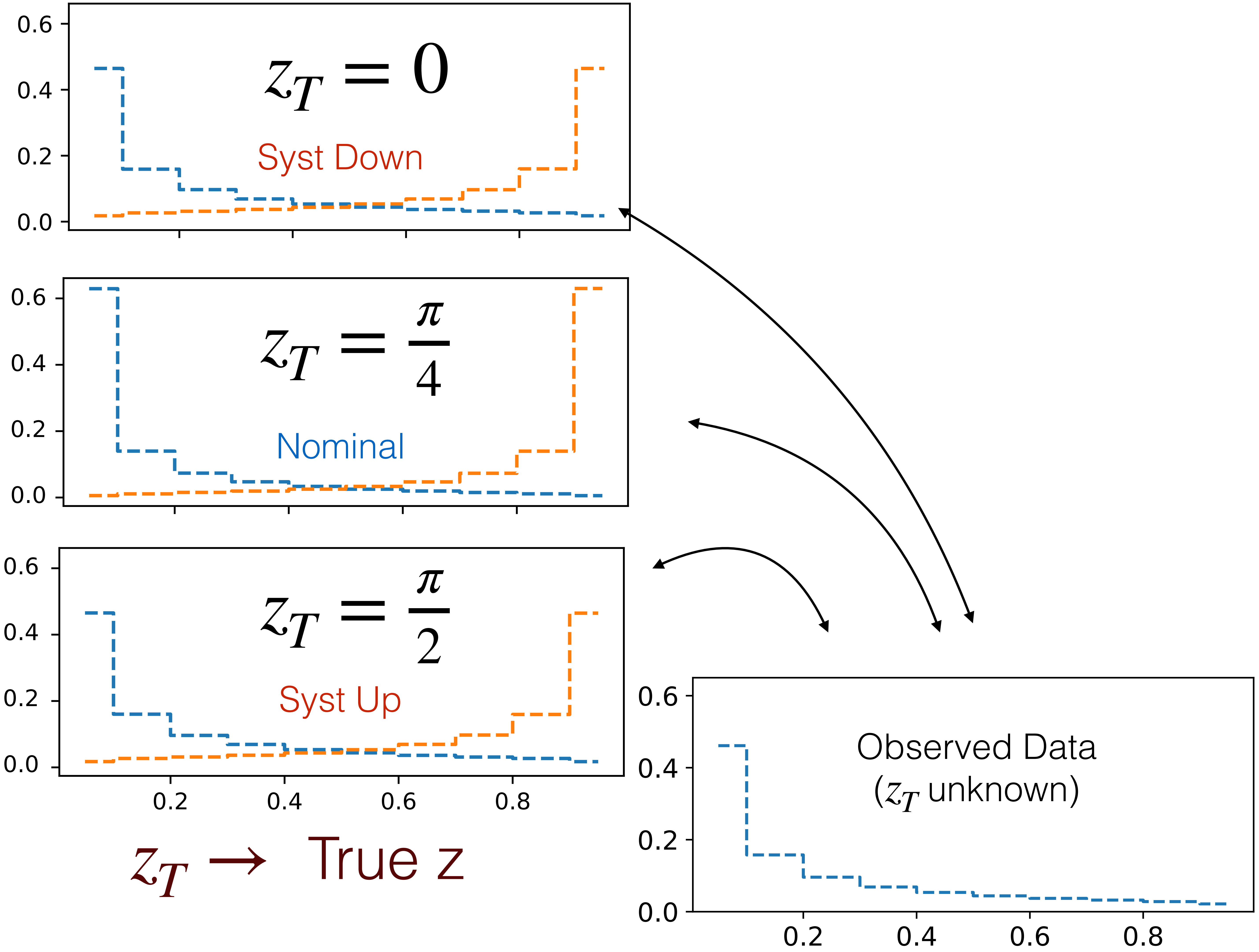


$z_T \rightarrow$ True z



Scan the 2D Likelihood space in Z vs μ

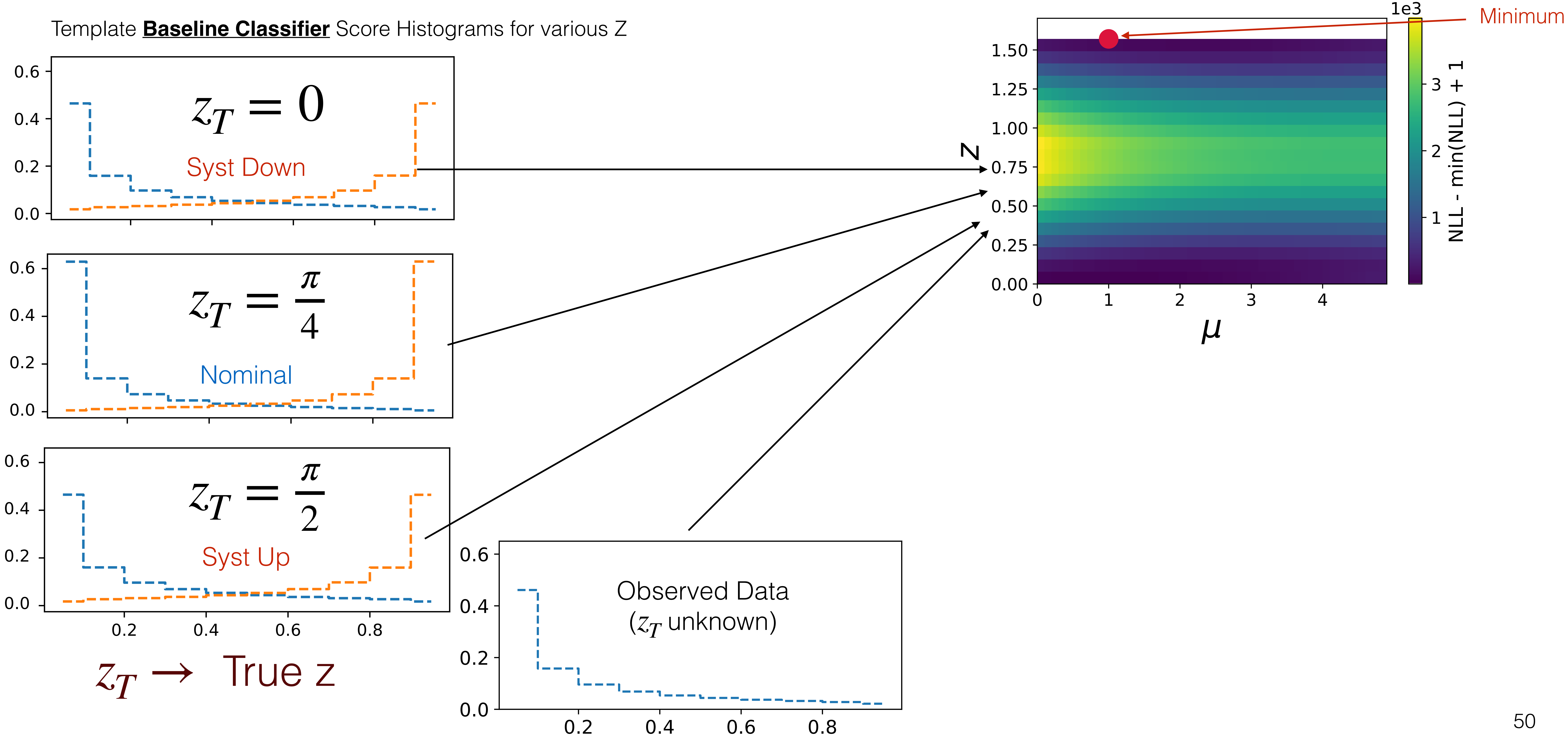
Template **Baseline Classifier** Score Histograms for various Z



$z_T \rightarrow$ True z

Scan the 2D Likelihood space in Z vs μ

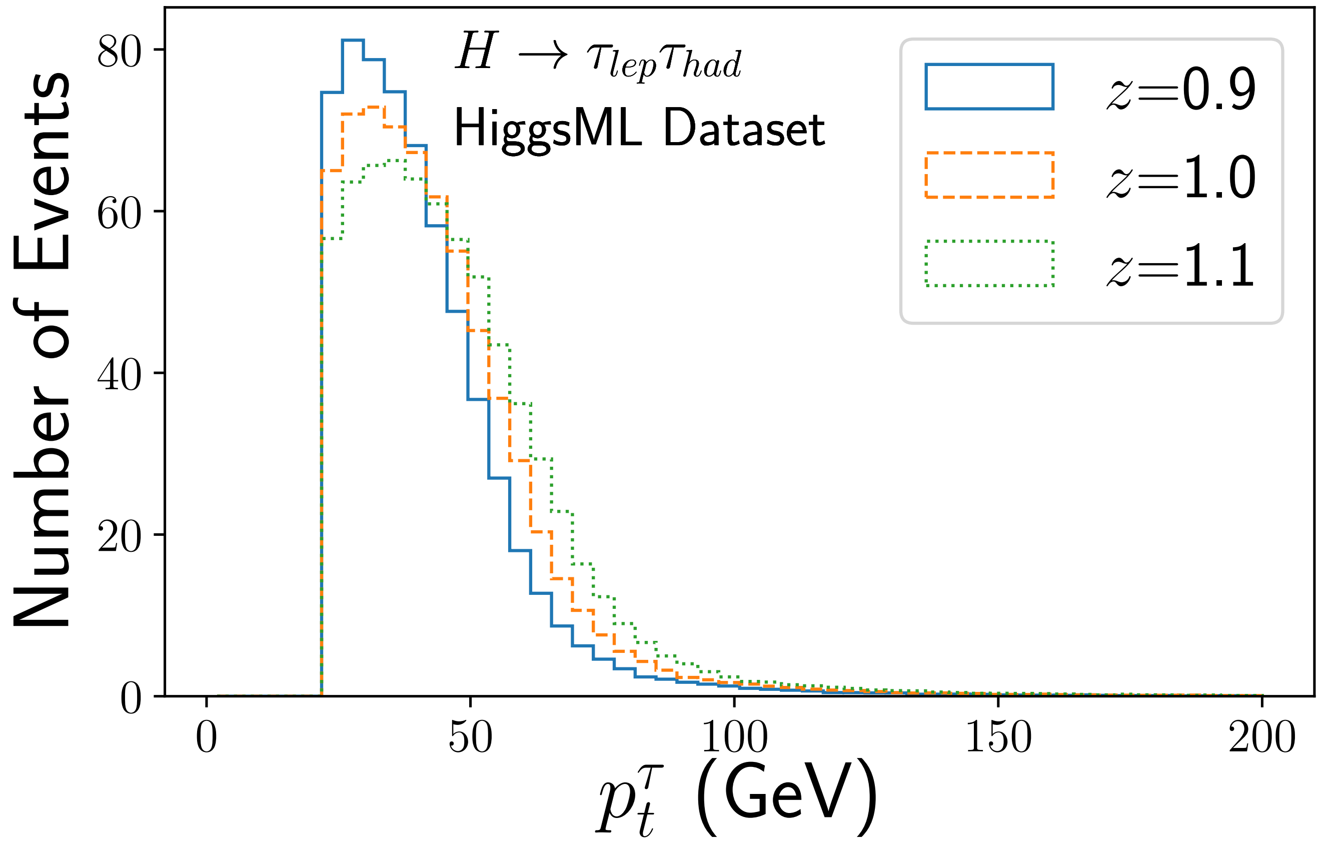
Template **Baseline Classifier** Score Histograms for various Z

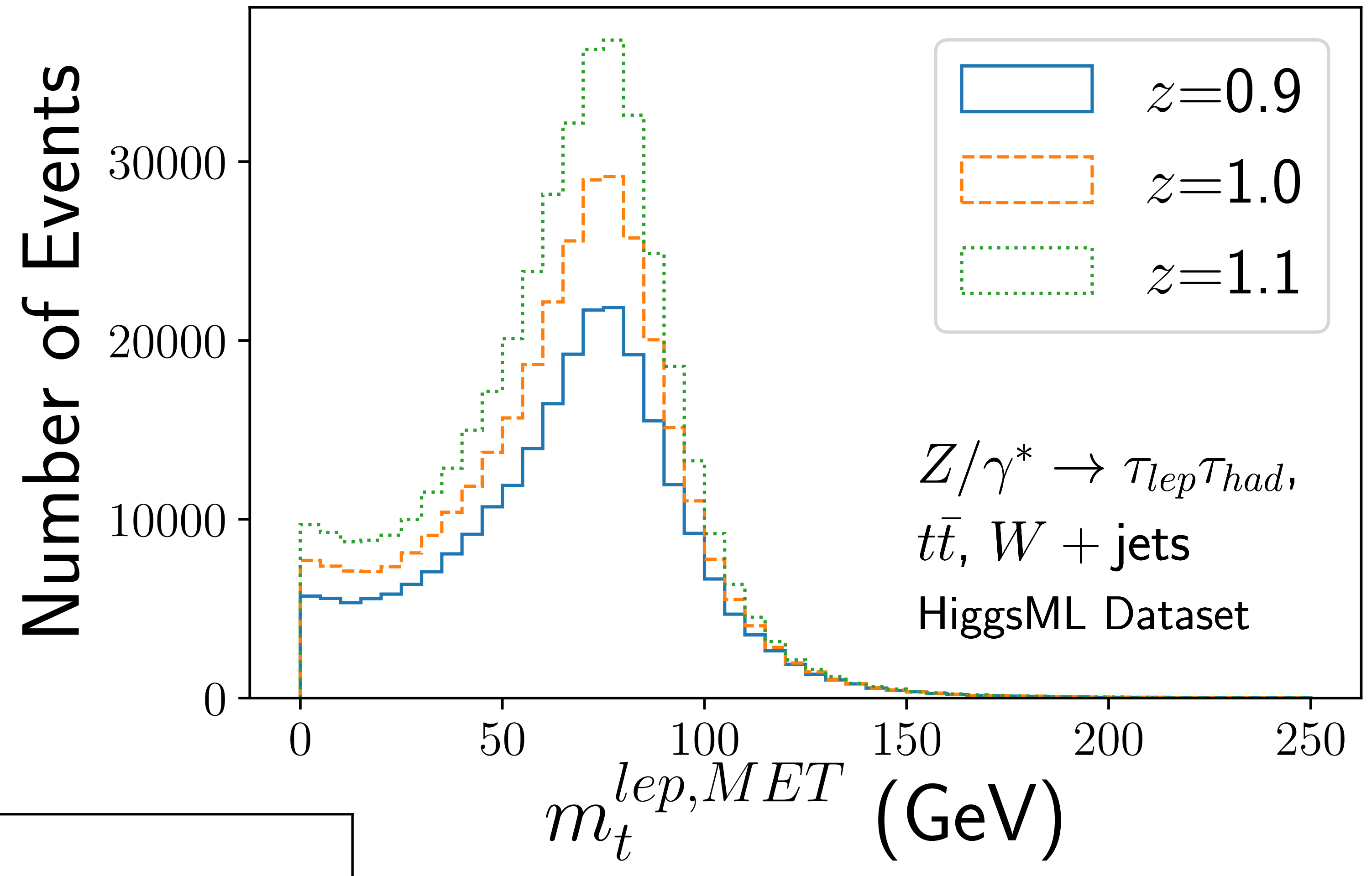


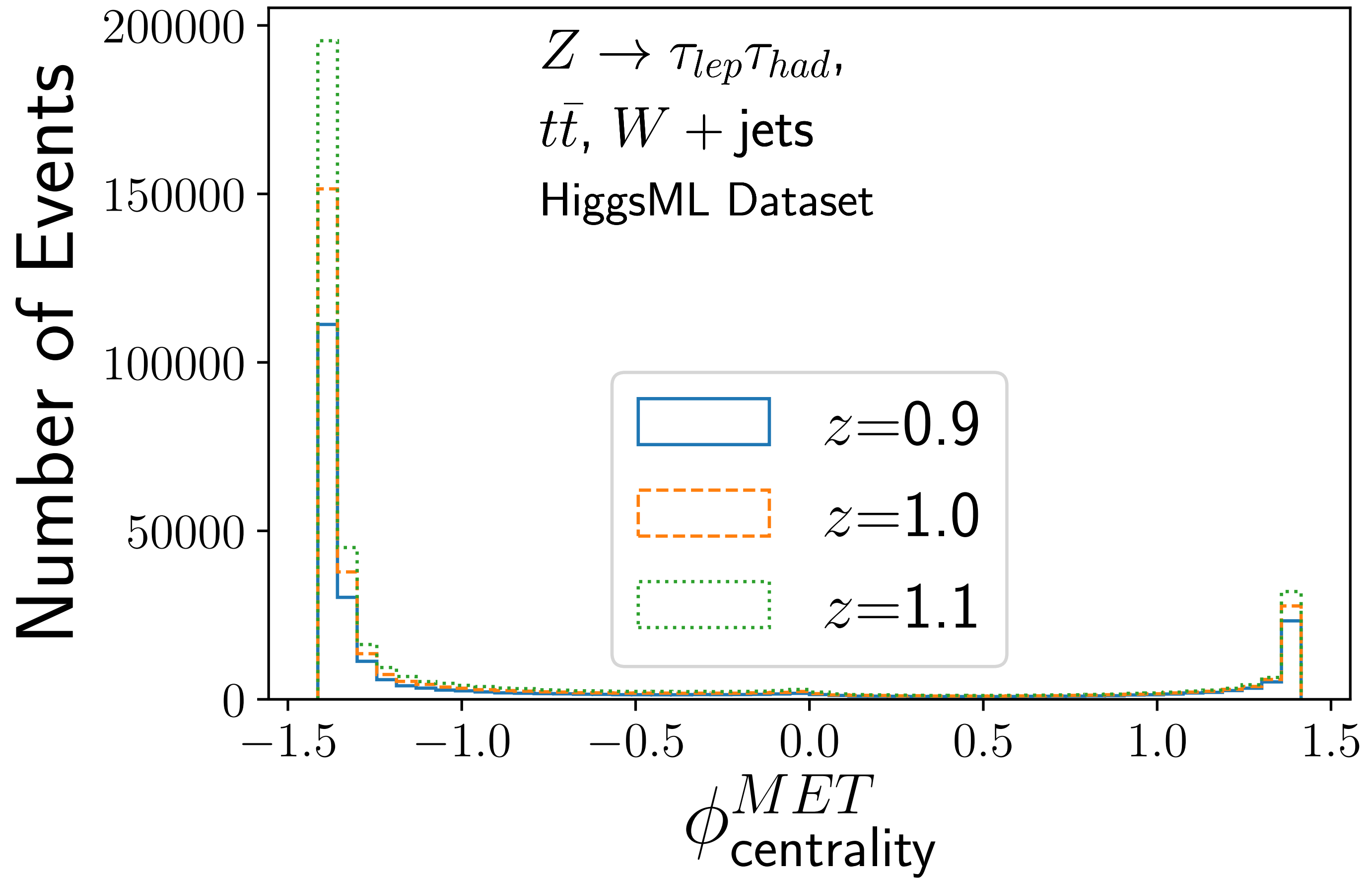
$z_T \rightarrow$ True z

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

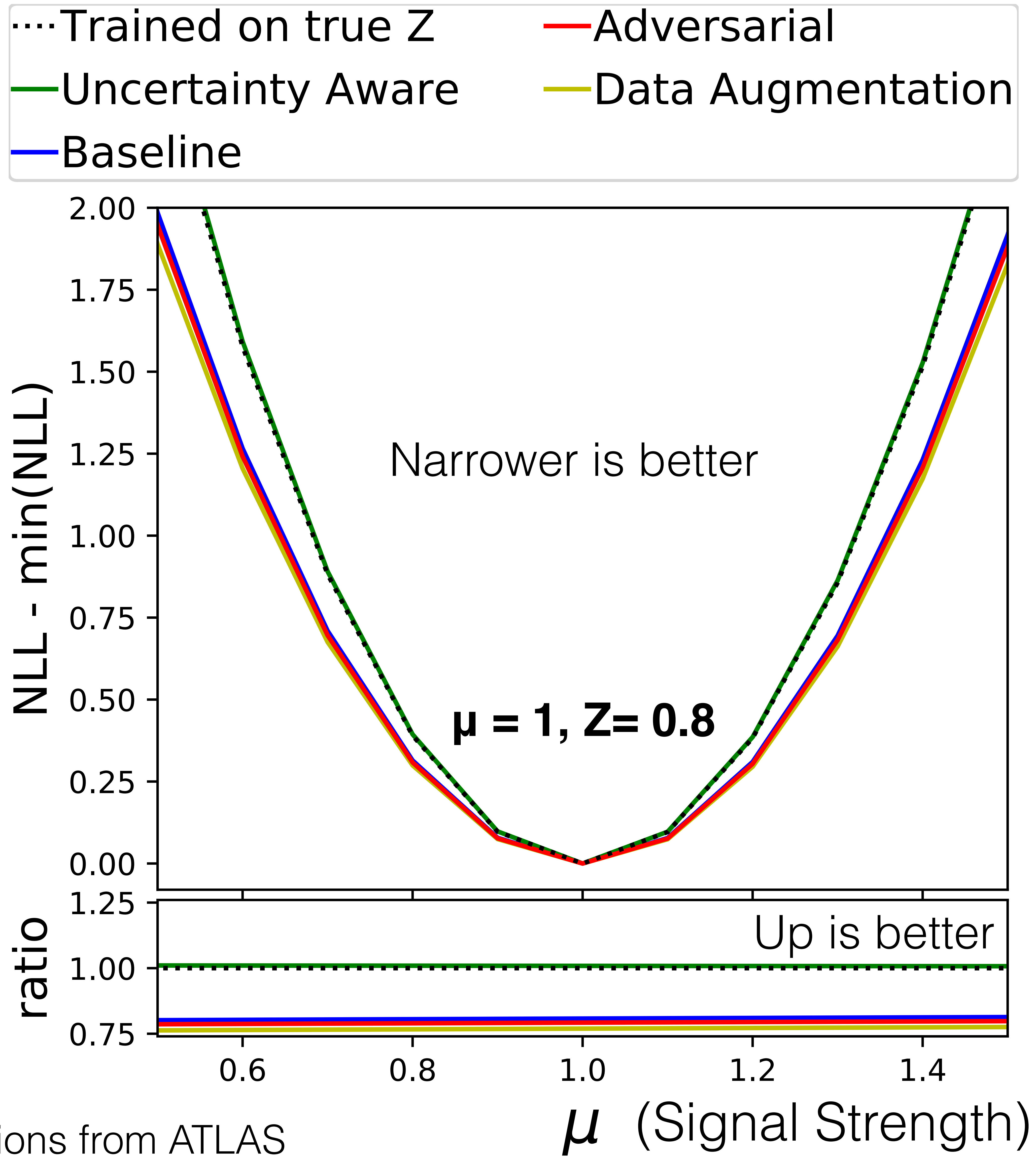
Parameter of Interest is Higgs signal strength μ, and TES is the nuisance parameter Z







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

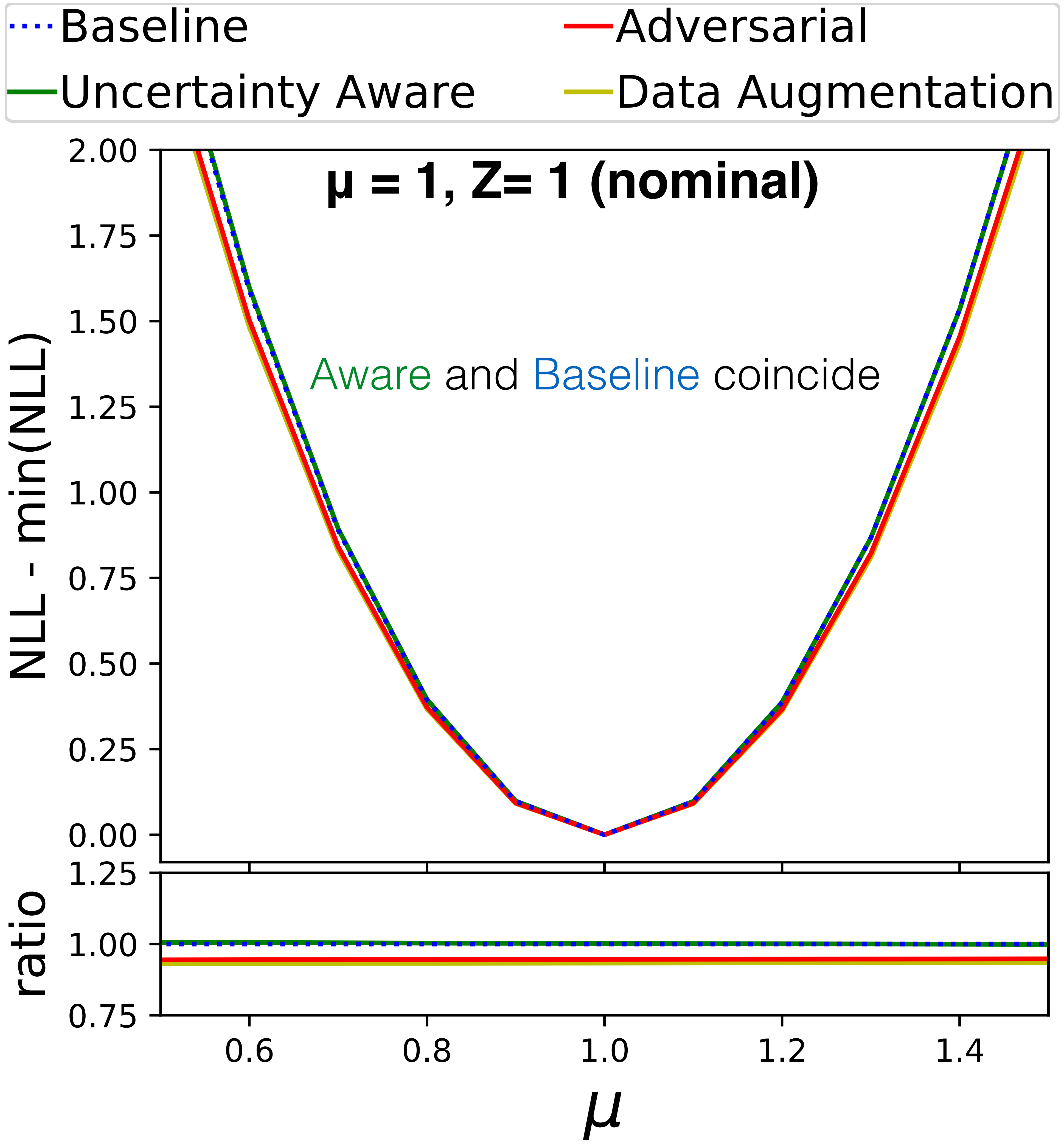


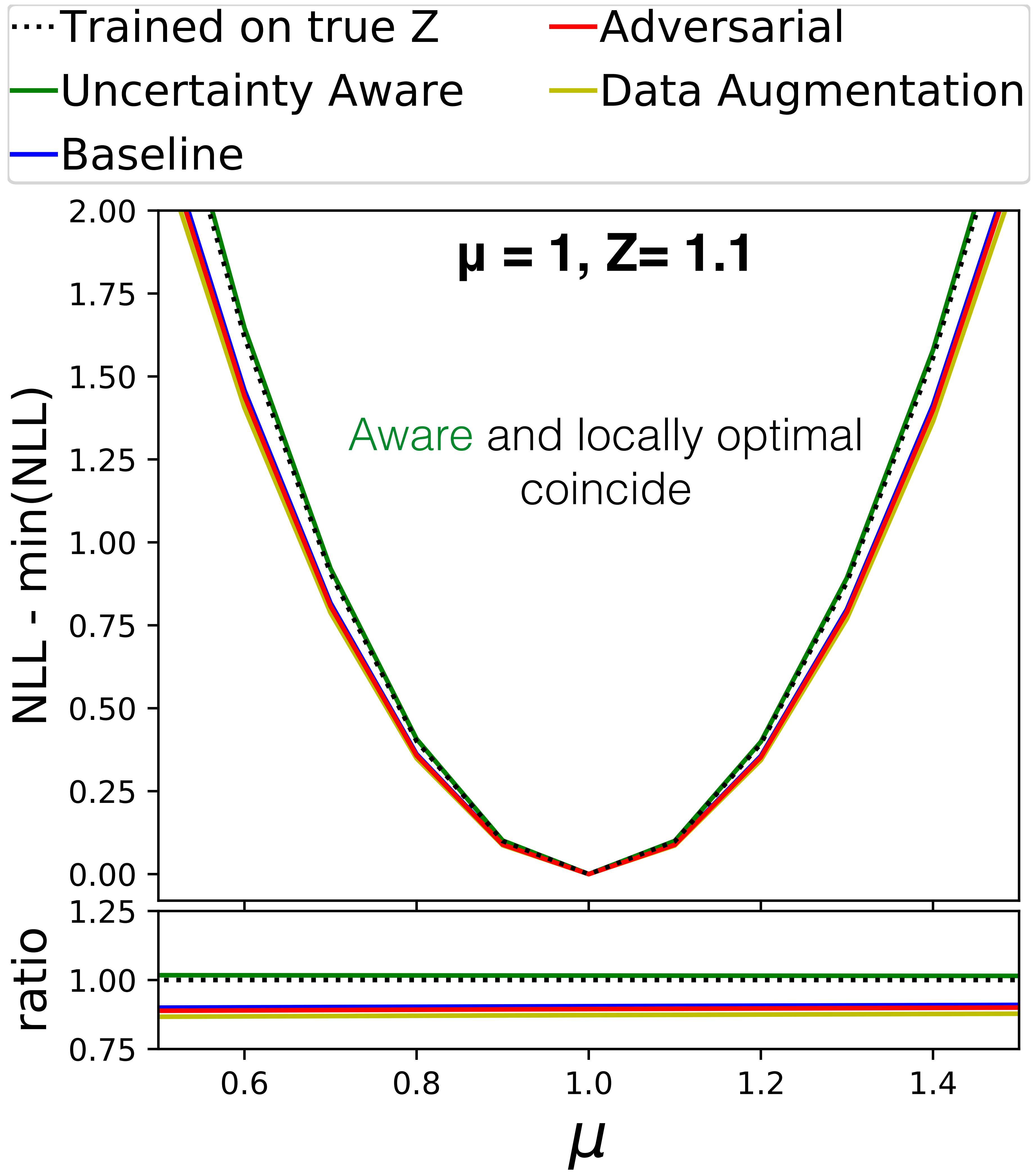
Uncertainty-Aware coincides with classifier trained on true Z

⇒ 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

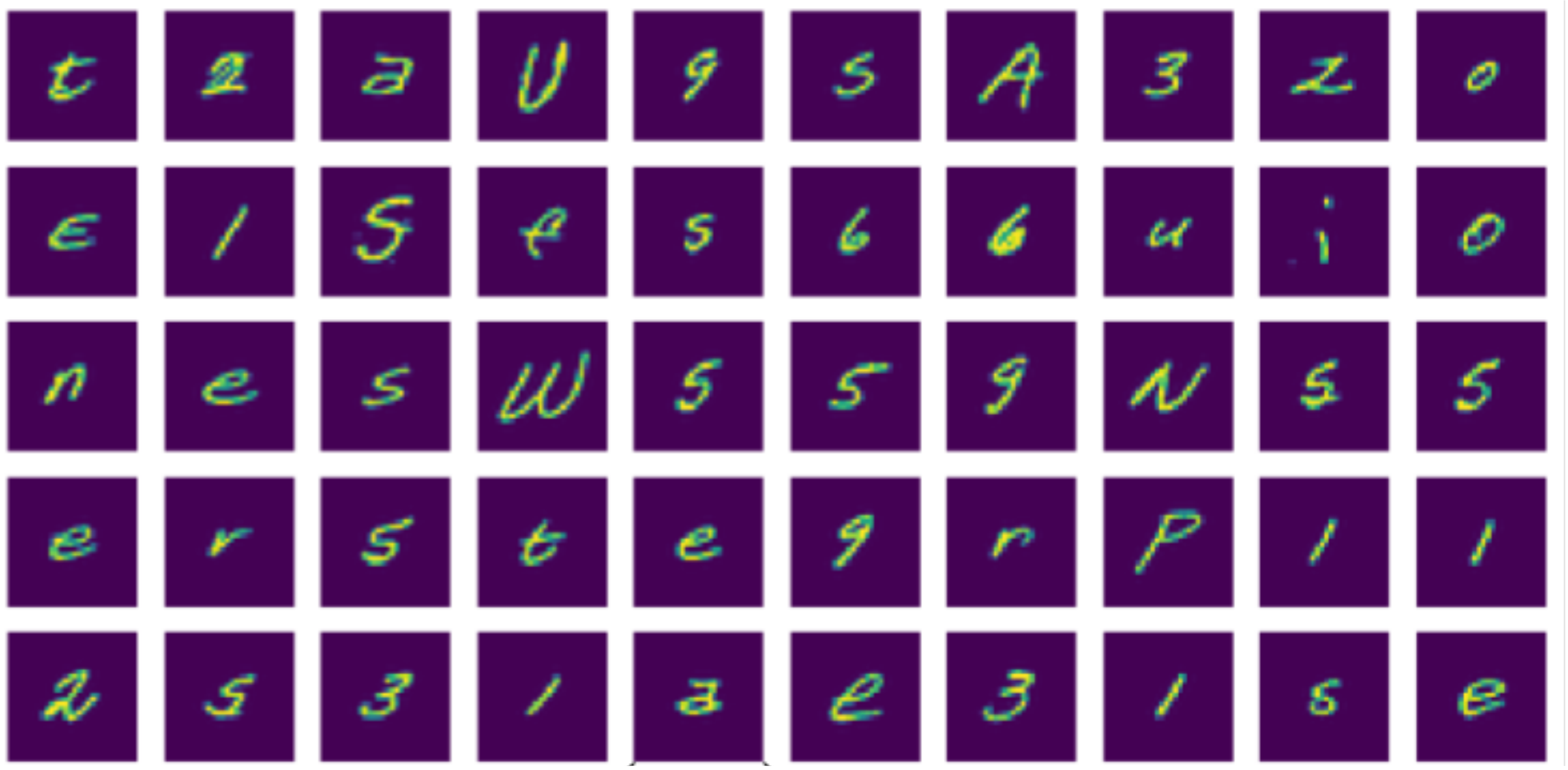
Idea fascinating also to ML researchers !

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

Idea fascinating also to ML researchers !

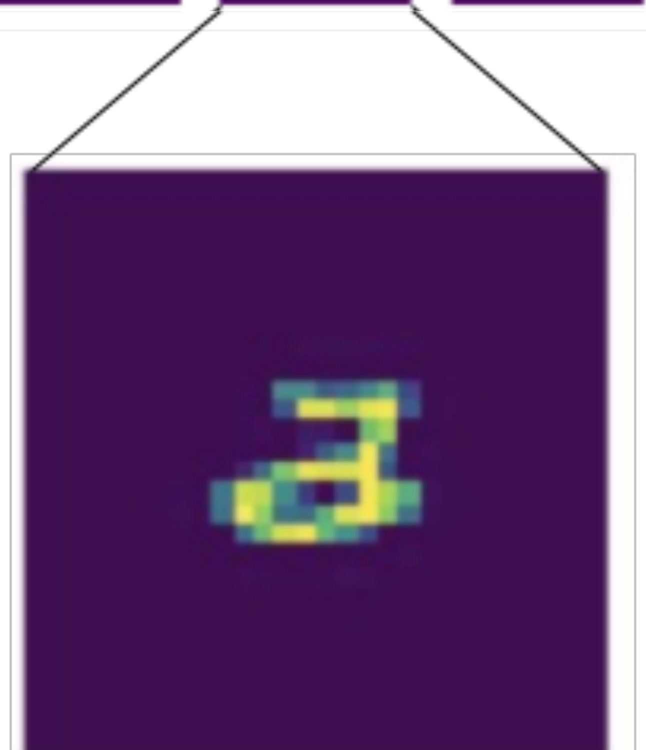
- ML researchers assume i.i.d
- This technique exploits correlations between samples – a different paradigm
- Interesting applications outside of physics



arXiv:2007.02931

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

ERM → 2
ARM → a

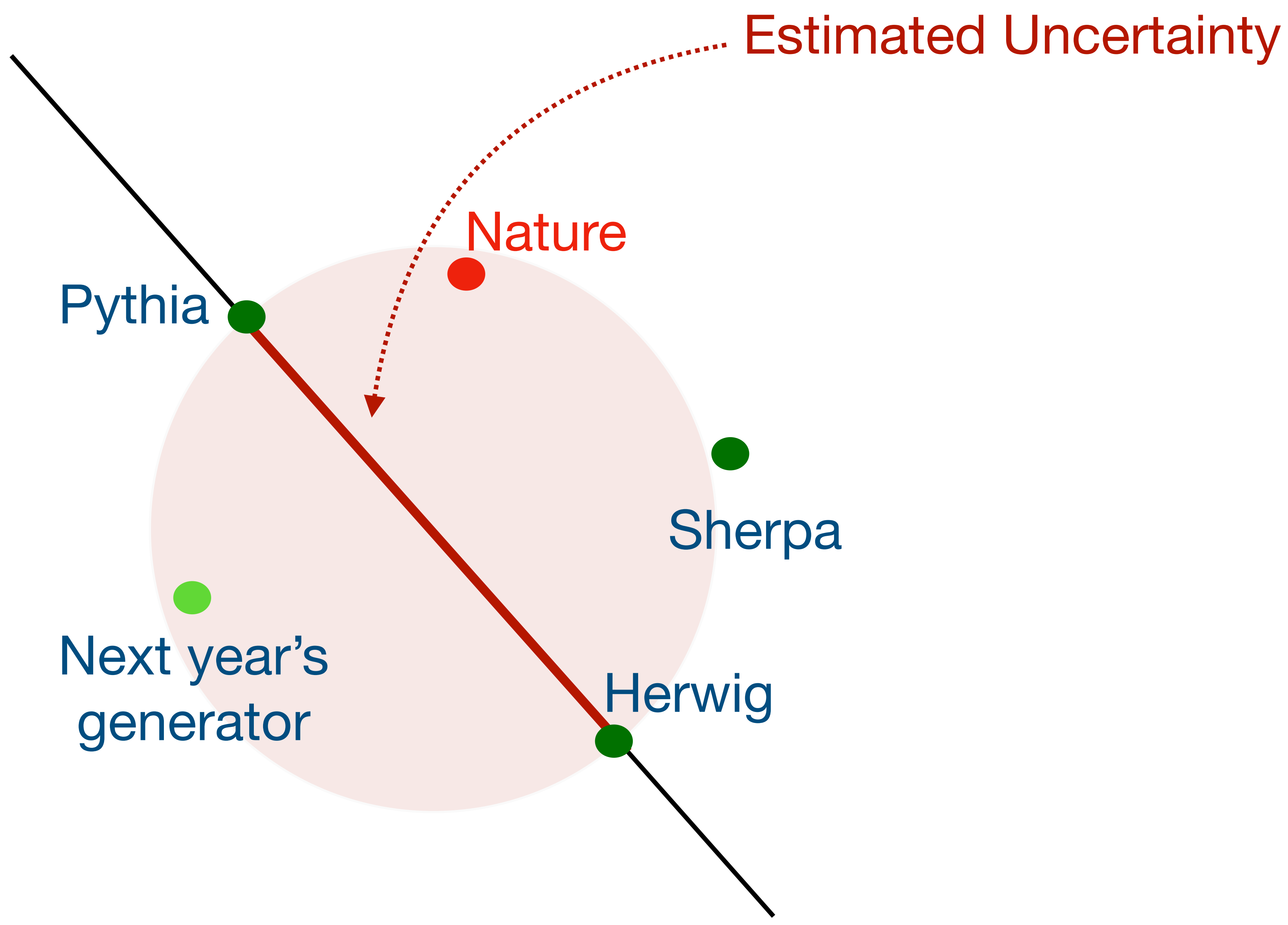
For my handwriting this is ‘2’, for yours it might be ‘a’
ARM: Adapt to the individual + classify

Theory Uncertainties

What are they ?

Theory uncertainties often describe our lack of understanding / ability to calculate

No statistical origin for them (such as auxiliary measurement)



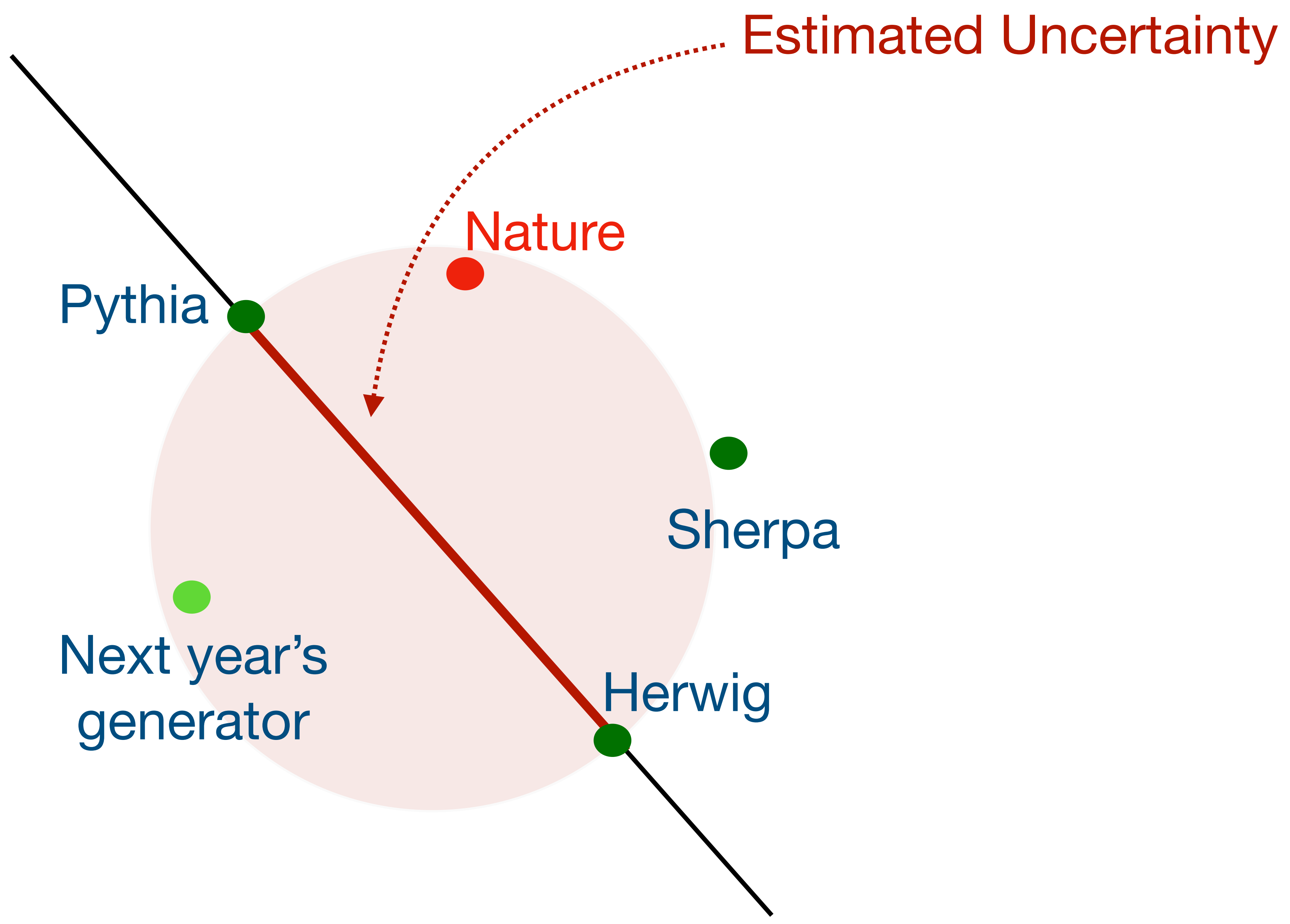
What are they ?

Theory uncertainties often describe our <u>lack of understanding / ability to calculate</u>

No statistical origin for them (such as auxiliary measurement)

Eg. <u>Hadronisation</u>:

- Few different packages to simulate it
- None are correct!
- Use difference in performance of your data analysis algorithm on **Pythia simulator** vs **Herwig simulator** ad-hoc estimate of uncertainty



Goodhart's Law

When a measure becomes a target, it ceases to be a good measure

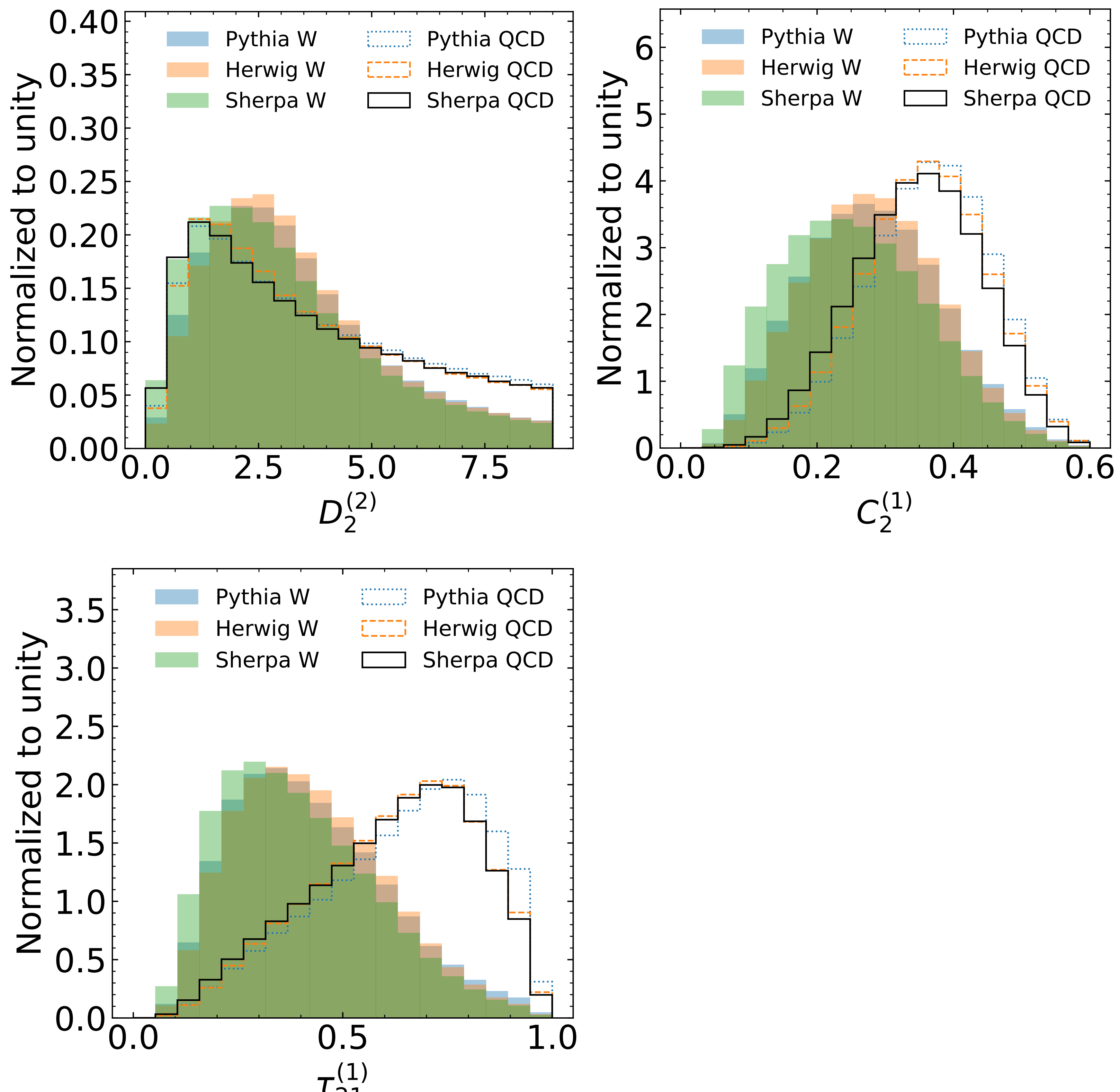
=> Dangerous to optimise proxy metrics of uncertainty

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

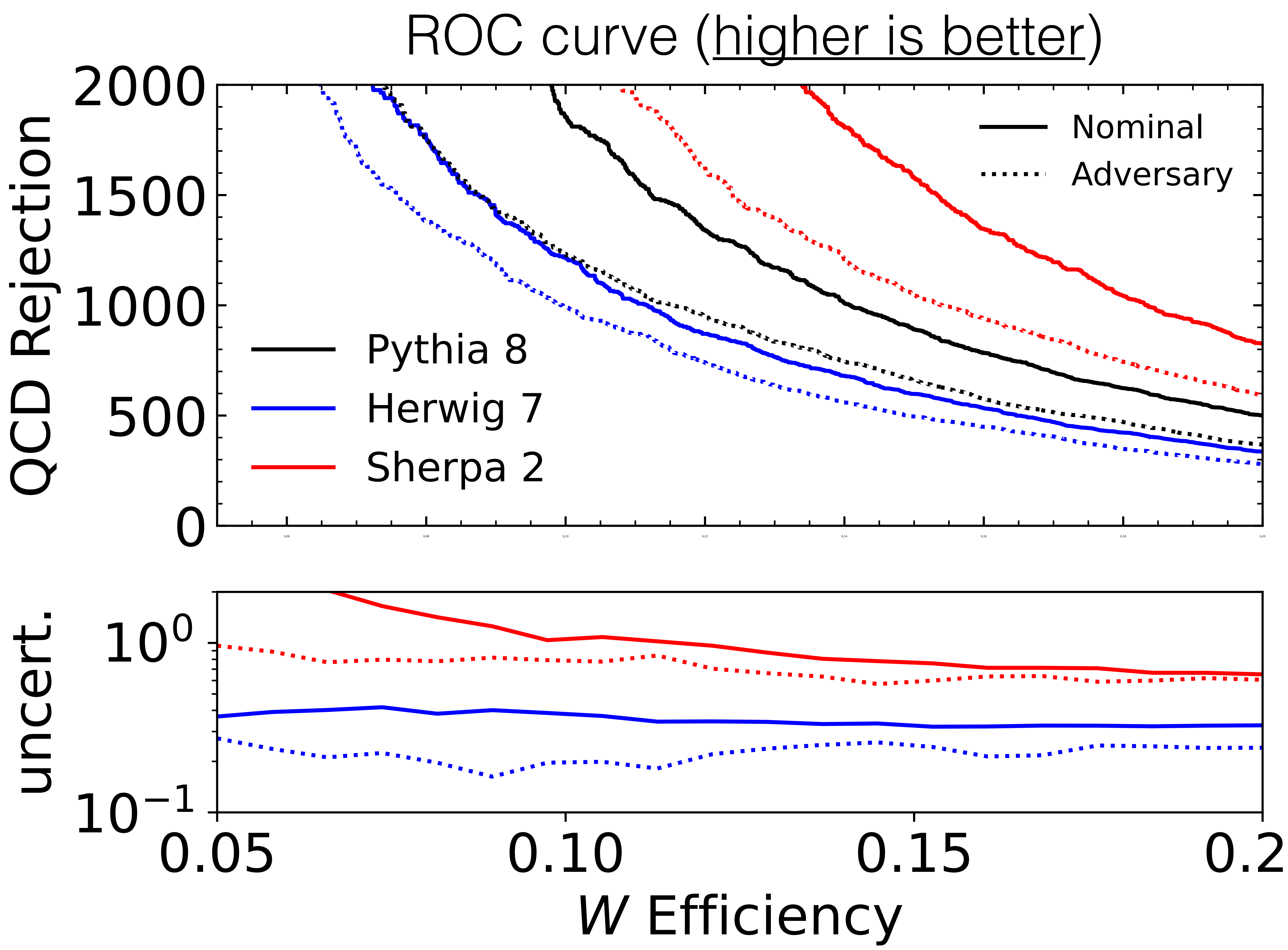
Goal: W jets vs QCD jets

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

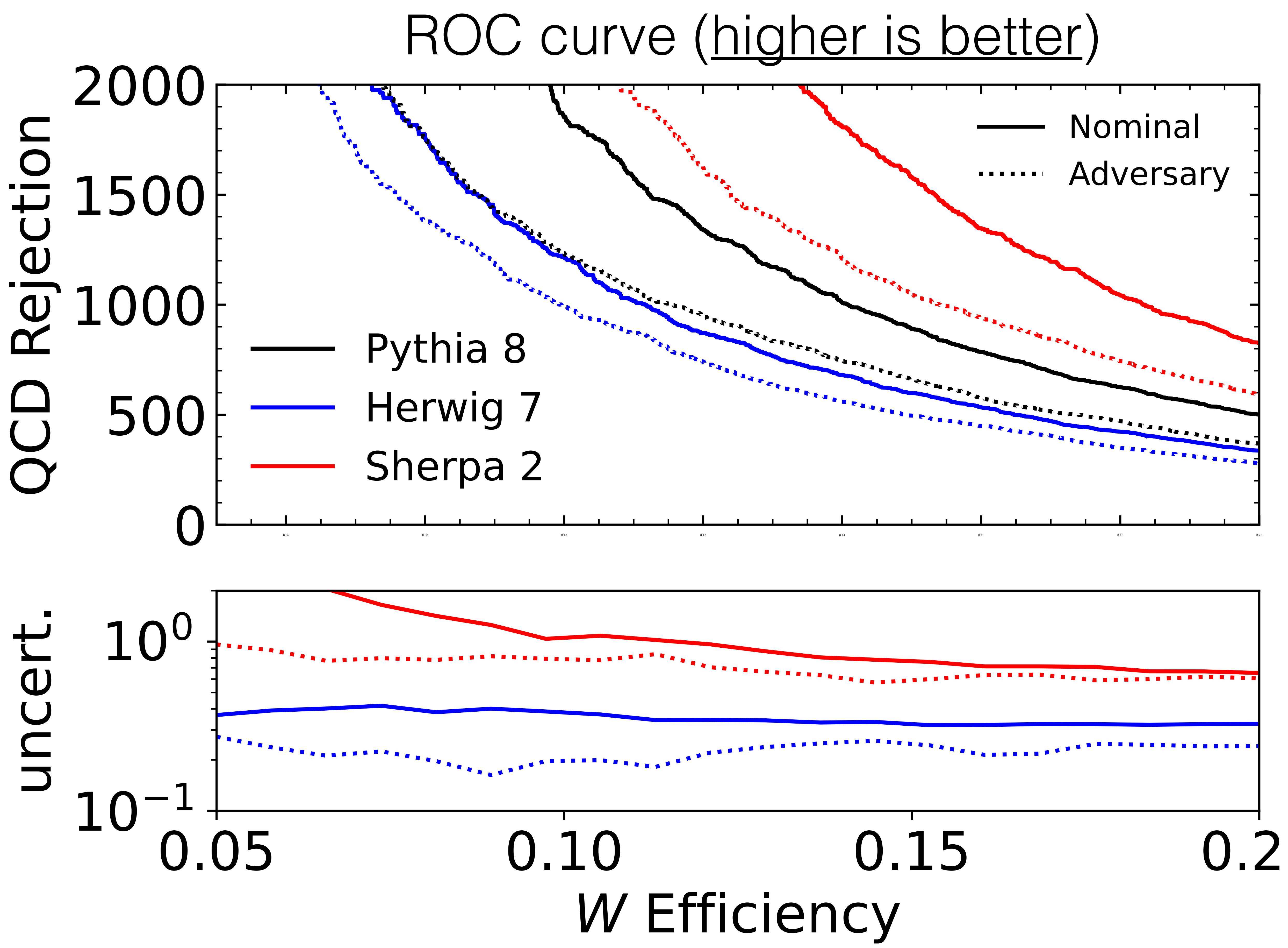


Case Study 1: Two-point uncertainty - Result

Adversary successfully <u>sacrifices separation power</u> in order to reduce difference in performance between Herwig and **Pythia**

Cross-check with Sherpa reveals <u>uncertainty severely underestimated</u> 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

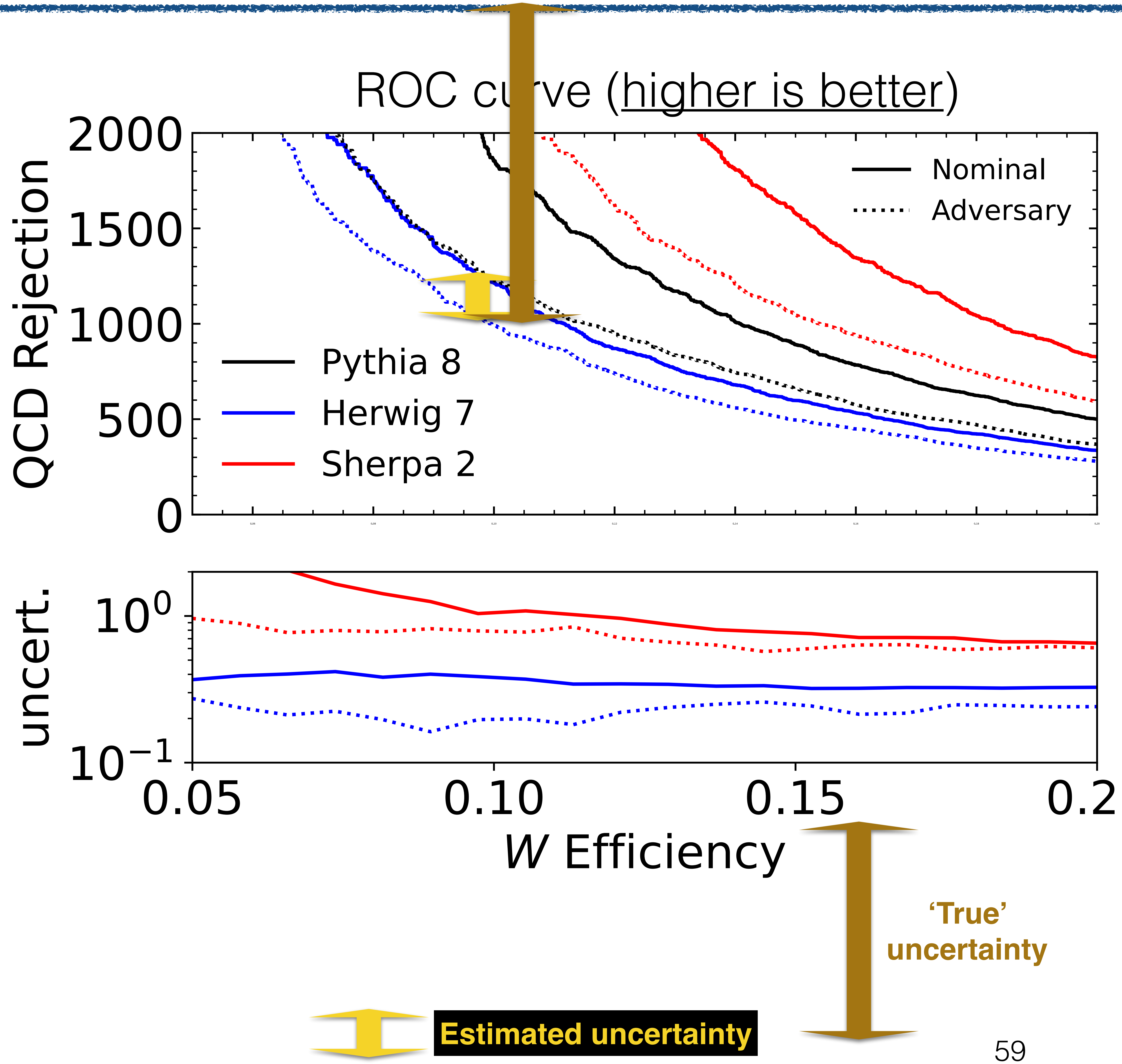


Case Study 1: Two-point uncertainty - Result

Adversary successfully <u>sacrifices separation power</u> in order to reduce difference in performance between Herwig and **Pythia**

Cross-check with Sherpa reveals <u>uncertainty severely underestimated</u> 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

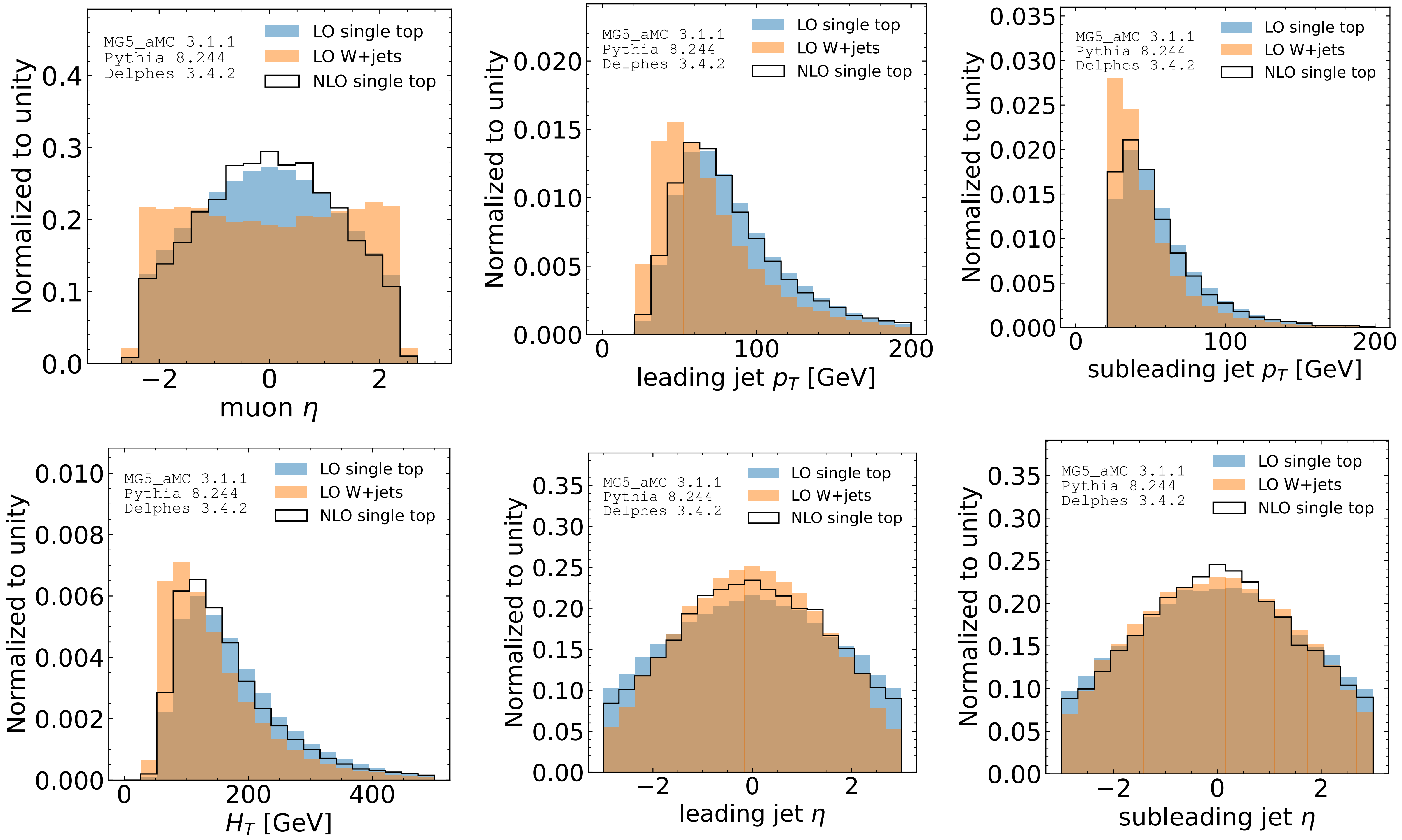
- We can’t calculate QFT to infinite order
- Artefact of truncation of series: Varying certain unphysical scales changes predictions
- Uncertainty quantification: Vary scales (renormalization scale, factorisation scale) between 1/2 to 2 in MC, see change in prediction

Scale uncertainty – Problem Setup

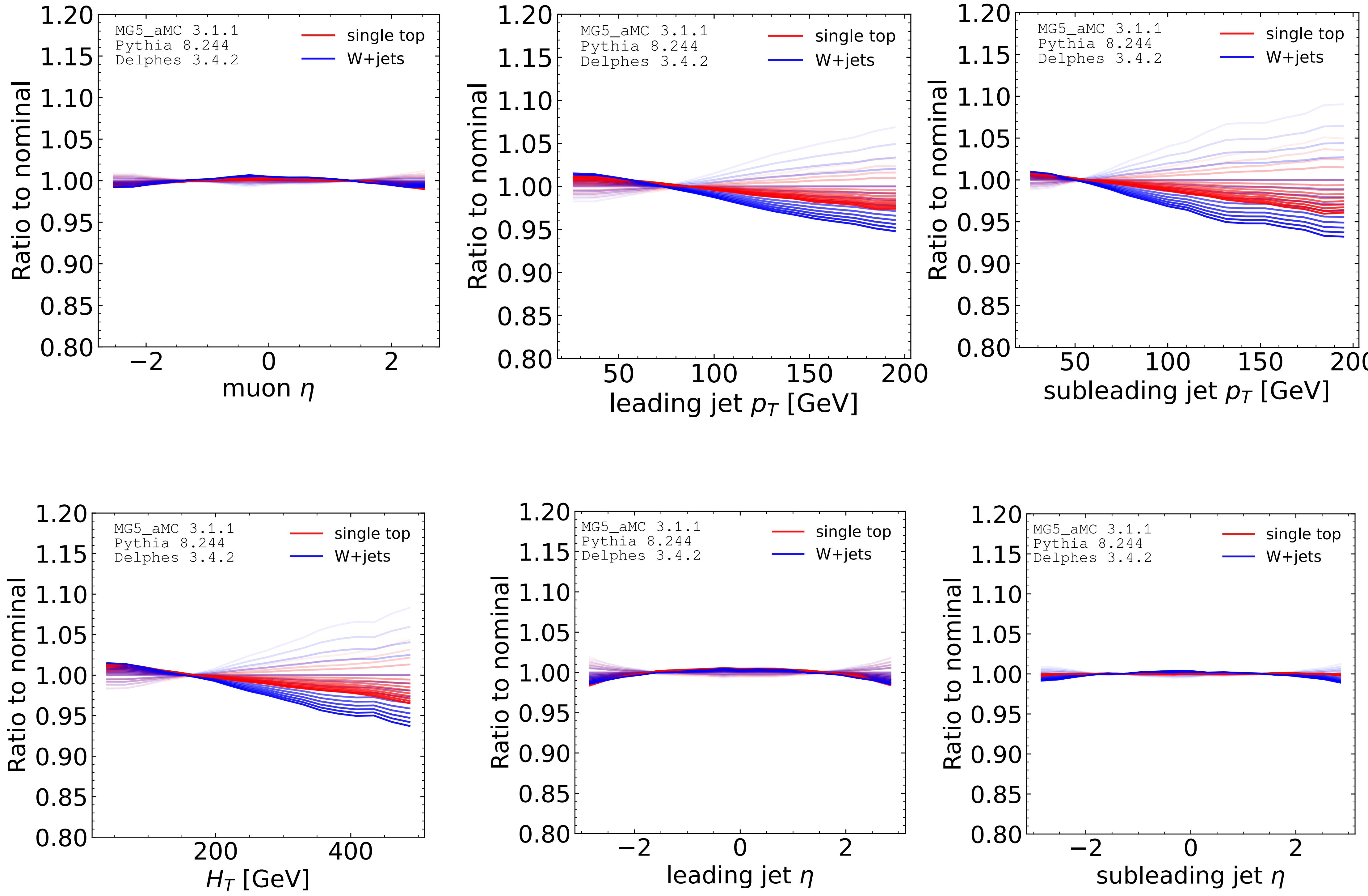
Goal: Single top vs W+Jets

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

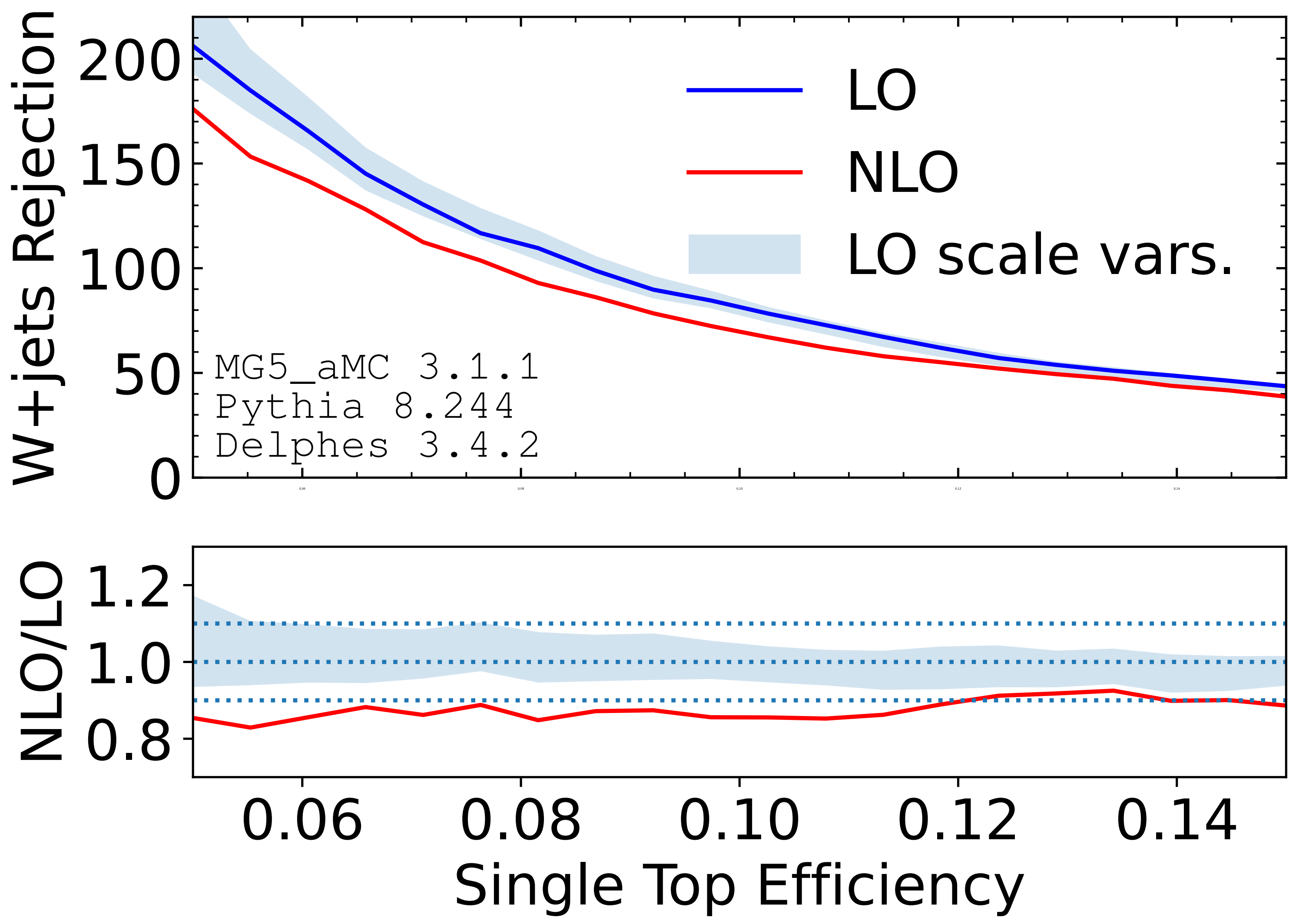


Factorisation scale variations going from 1/2 to 2

Case Study 2: Continuous uncertainty - Result

ROC curve (higher is better)

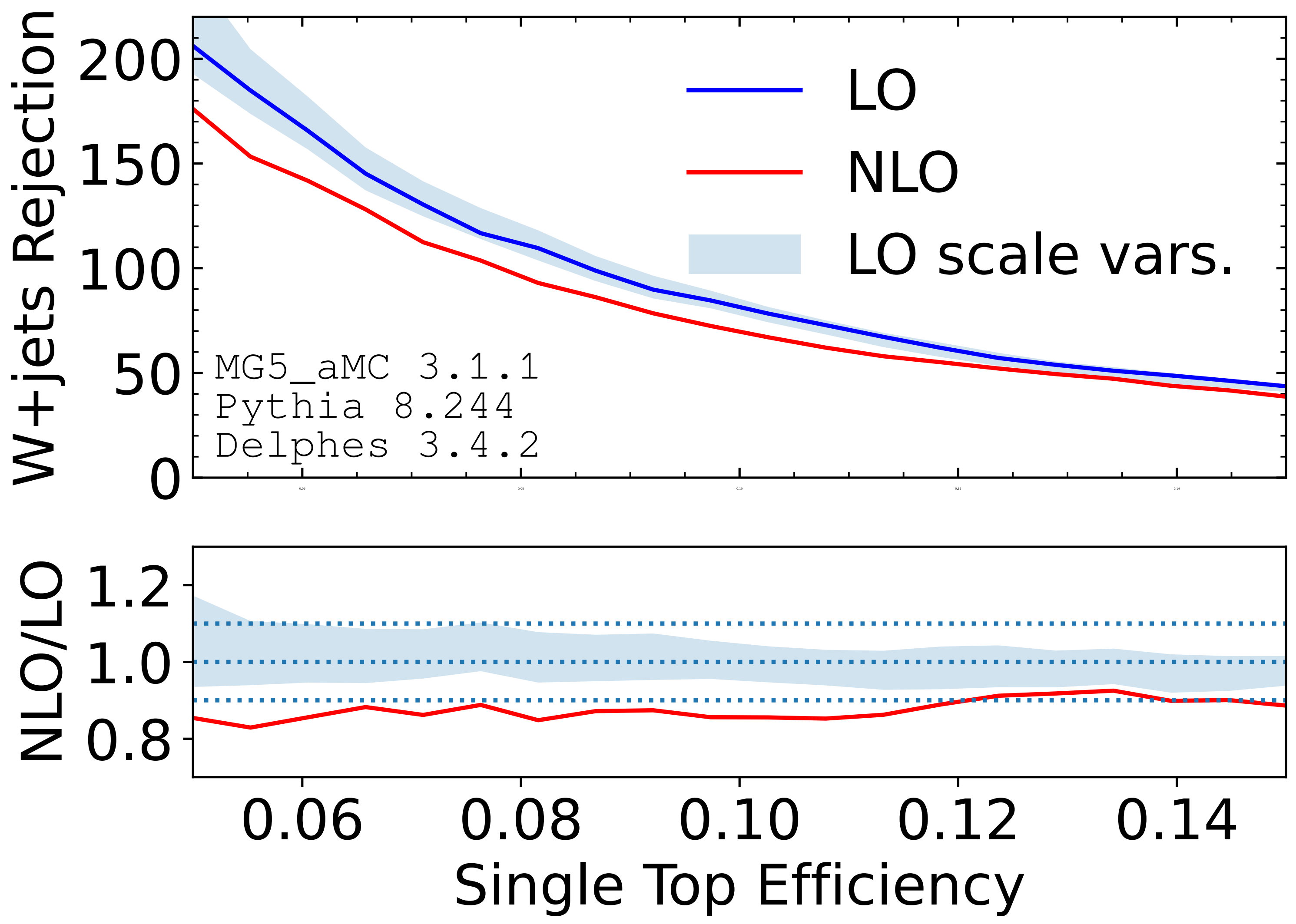
No adversary



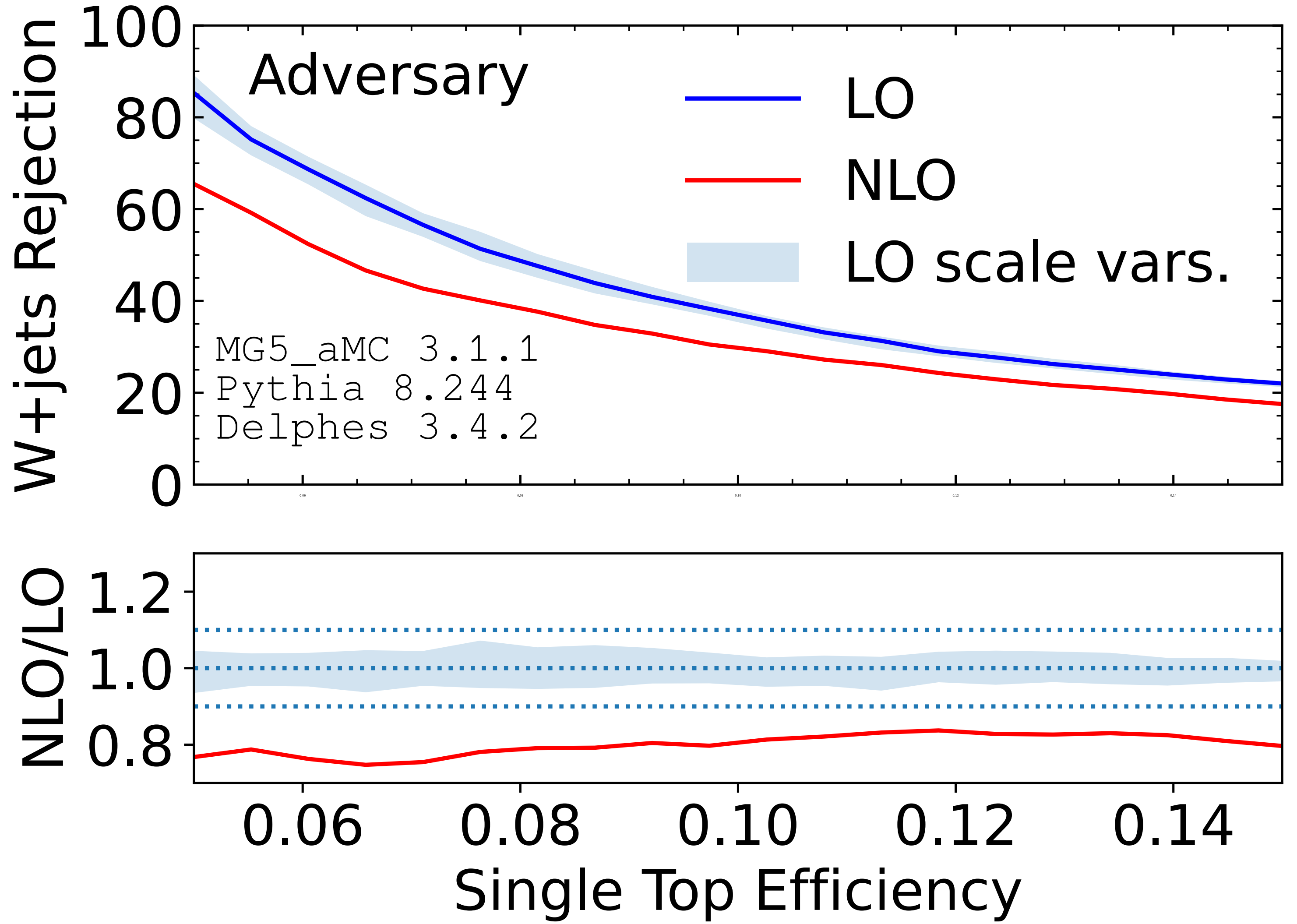
Case Study 2: Continuous uncertainty - Result

ROC curve (<u>higher is better</u>)

No adversary

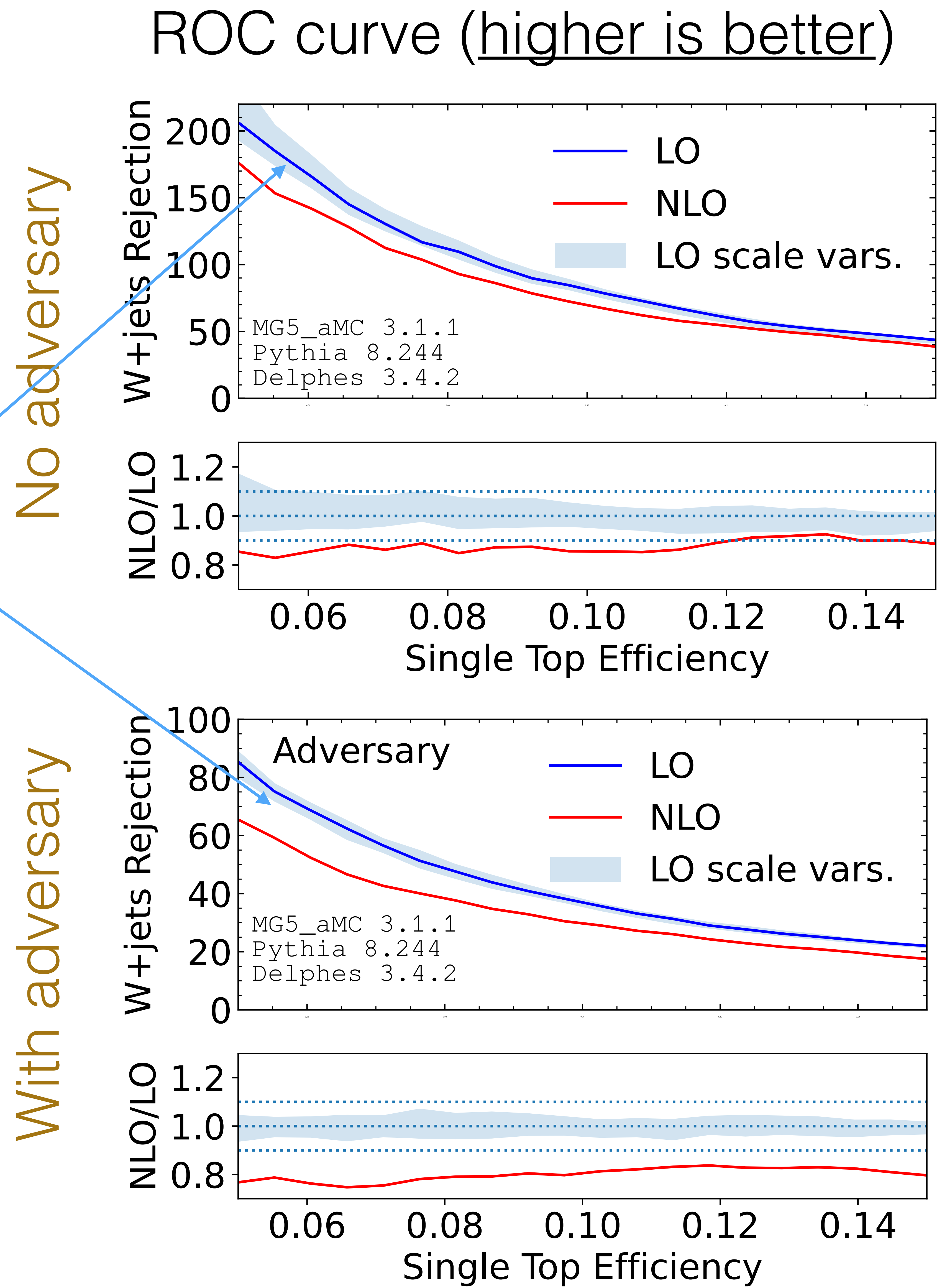


With adversary



Case Study 2: Continuous uncertainty - Result

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



Case Study 2: Continuous uncertainty - Result

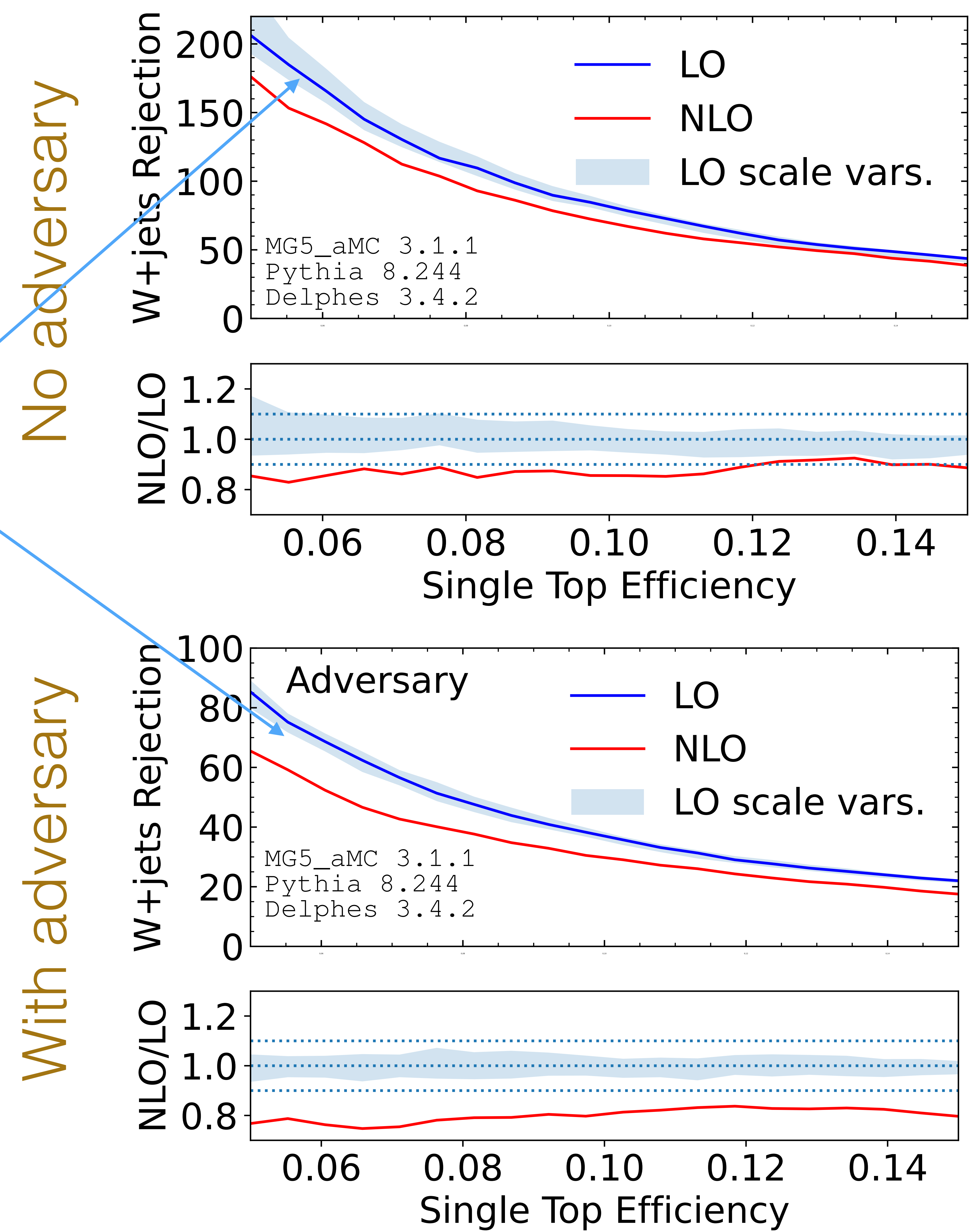
Adversary successfully **sacrifices separation power** in order to reduce difference in performance between scale variations

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)



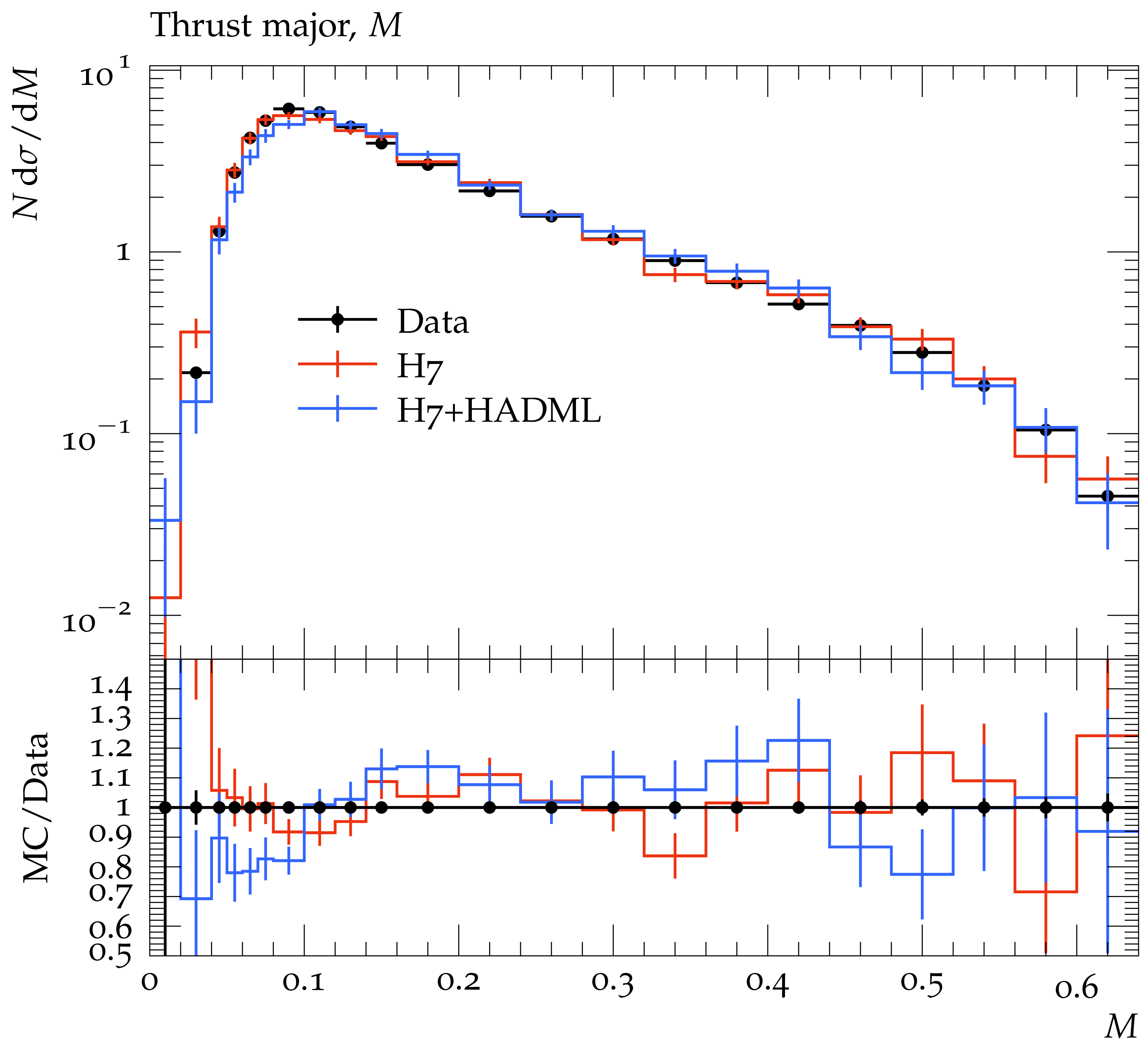
Universe is a perfect simulator

PRD.106.096020: **Aishik Ghosh,** Xiangyang Ju, Benjamin Nachman, and Andrzej Siodmok

Universe is a perfect simulator

PRD.106.096020: **Aishik Ghosh,** Xiangyang Ju, Benjamin Nachman, and Andrzej Siodmok

Bypass theory, can we learn hadronization directly from data ?



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!