Generative models and uncertainty

Gregor Kasieczka
Email: gregor.kasieczka@uni-hamburg.de
Twitter/X: @GregorKasieczka

Artificial Intelligence and the Uncertainty challenge in Fundamental Physics
30.11.2023 — AISSAI / IN2P3

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE
KISS
CDCS
CENTER FOR DATA AND COMPUTING IN NATURAL SCIENCES
FSP CMS
PUNCH 4NFDI
DASHH
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
UHH Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG
PIER
Partnership of Universität Hamburg and DESY
Emmy Noether-Programm
Deutsche Forschungsgemeinschaft DFG

Introduction

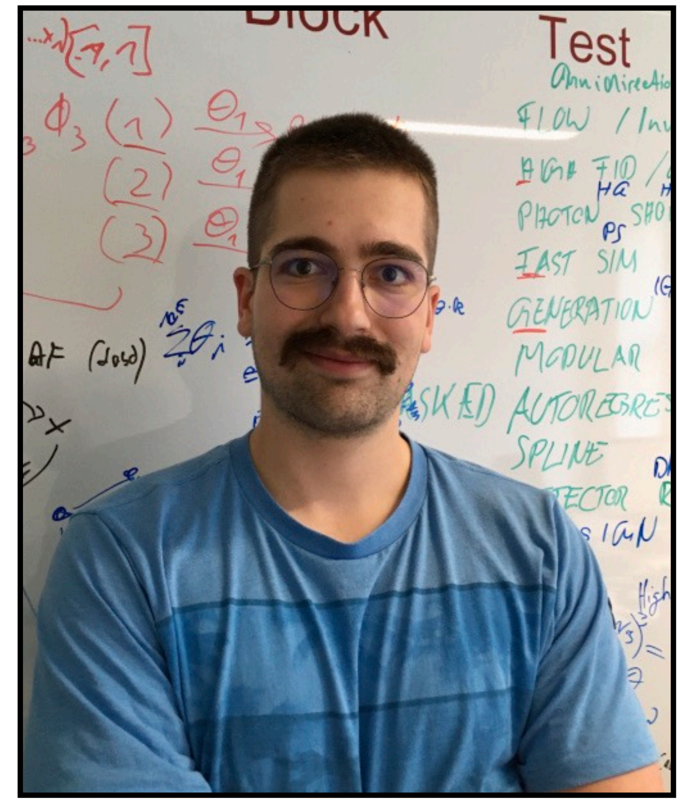
- Increasing use of generative models in different aspects of LHC analysis chain

- Proper treatment of uncertainties is not fully keeping up: interesting problems

- Will discuss 4 examples:
    - Calorimeter Simulation
    - Ephemeral learning
    - Anomaly Detection
    - Surrogate Classifiers

Introduction

- Increasing use of generative models in different aspects of LHC analysis chain
- Proper treatment of uncertainties is not fully keeping up: interesting problems
- Will discuss 3 examples:
  - Calorimeter Simulation
  - Ephemeral learning
  - Anomaly Detection
  - ~~Surrogate Classifiers~~



10:15 **Efficient Sampling from Bayesian Network Posteriors for Optimal Uncertainties** 25m

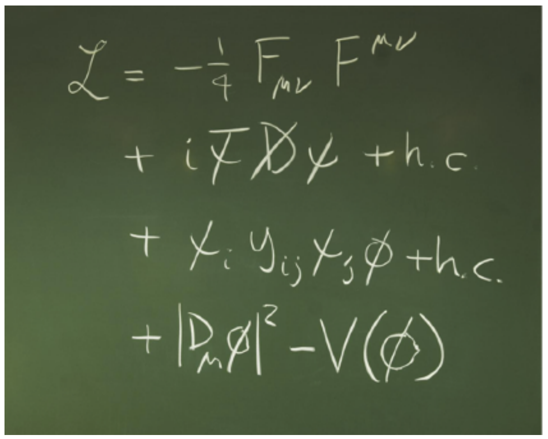
Bayesian neural networks are a key technique when including uncertainty predictions into neural network analysis, be it in classification, regression or generation. Although being an essential building block for classical Bayesian techniques, Markov Chain Monte Carlo methods are seldomly used to sample Bayesian neural network weight posteriors due to slow convergence rates in high dimensional parameter spaces. Metropolis-Hastings corrected chains exhibit two major issues: using a stochastic Metropolis-Hastings term and bad acceptance rates. We present solutions to both problems in form of a correction term to the loss objective and novel proposal distributions based on the Adam-optimizer. The combined algorithm shows fast convergence and good uncertainty estimation for physics use cases without dramatically increasing the cost of computation over gradient descent based optimization.

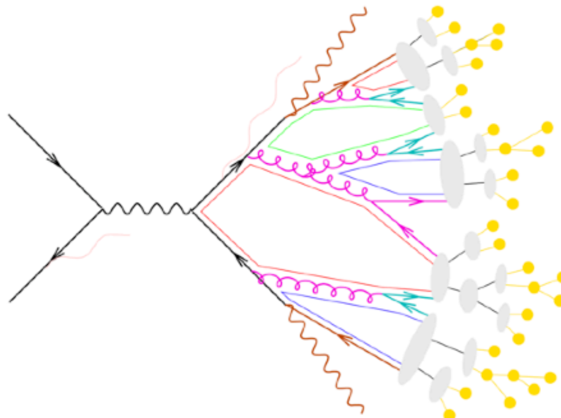
Sprecher: Sebastian Bieringer (Hamburg University, Institute for experimental physics)

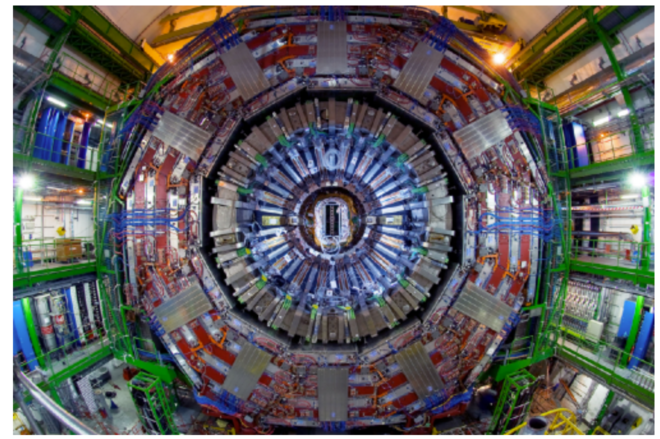
Calorimeter Simulation

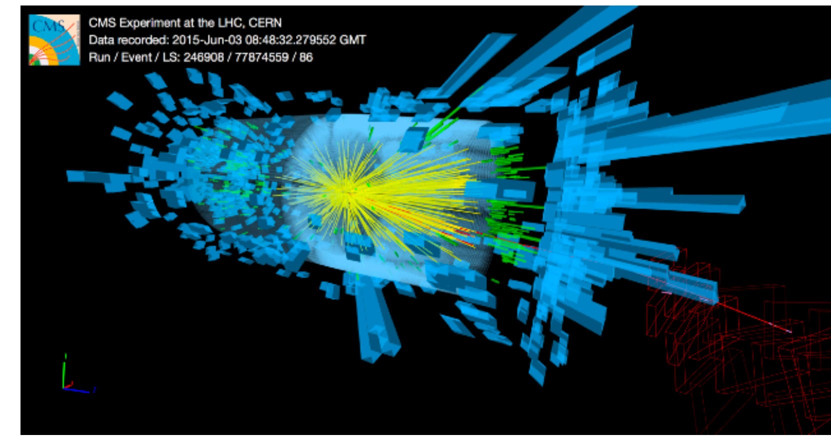
Generative Models

This happens in the experiment







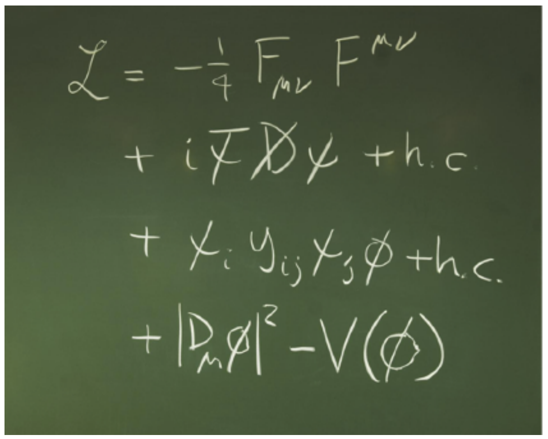


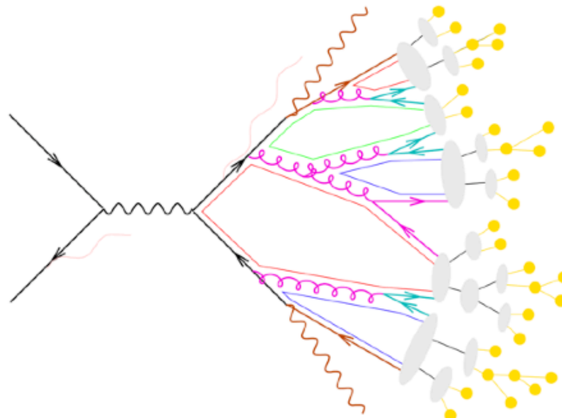
This is what we want to know

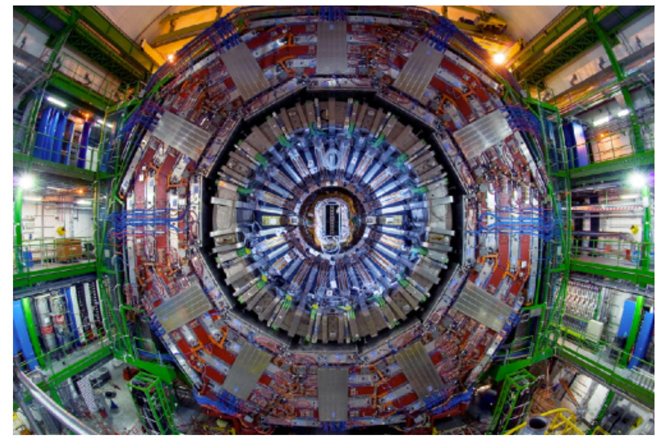
Simulation is crucial to connect experimental data with theory predictions

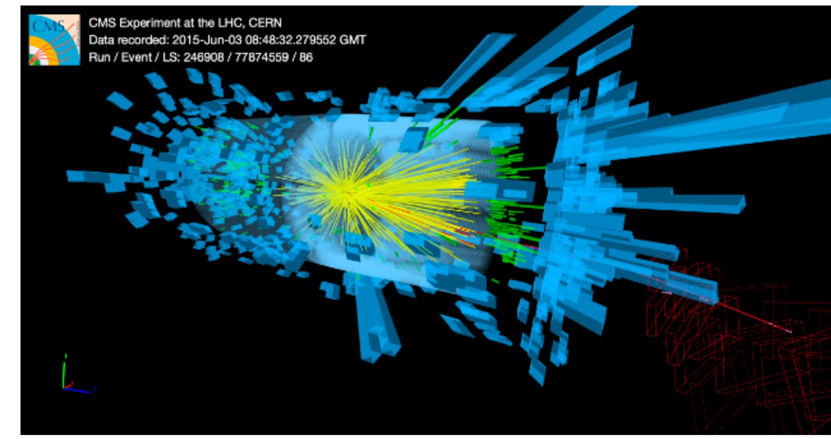
Generative Models

This happens in the experiment



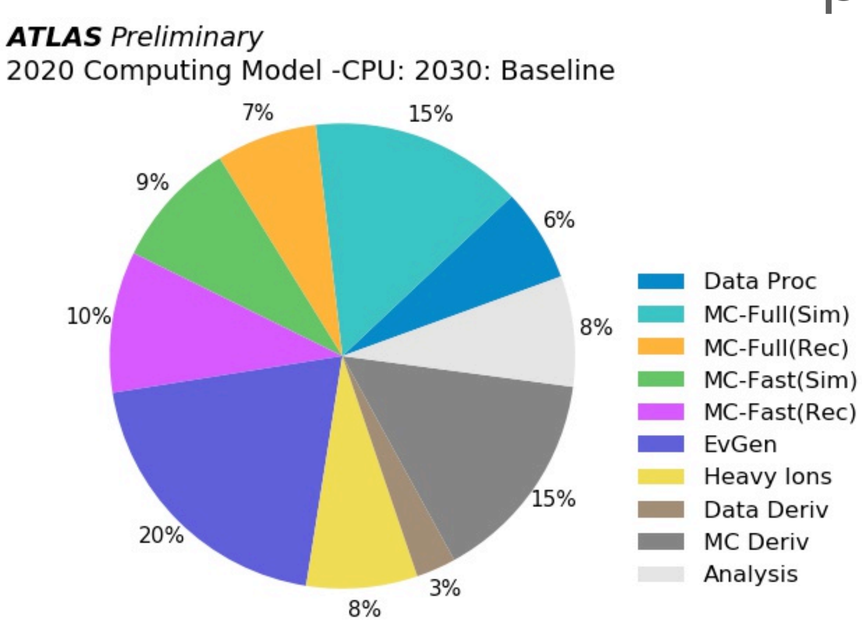


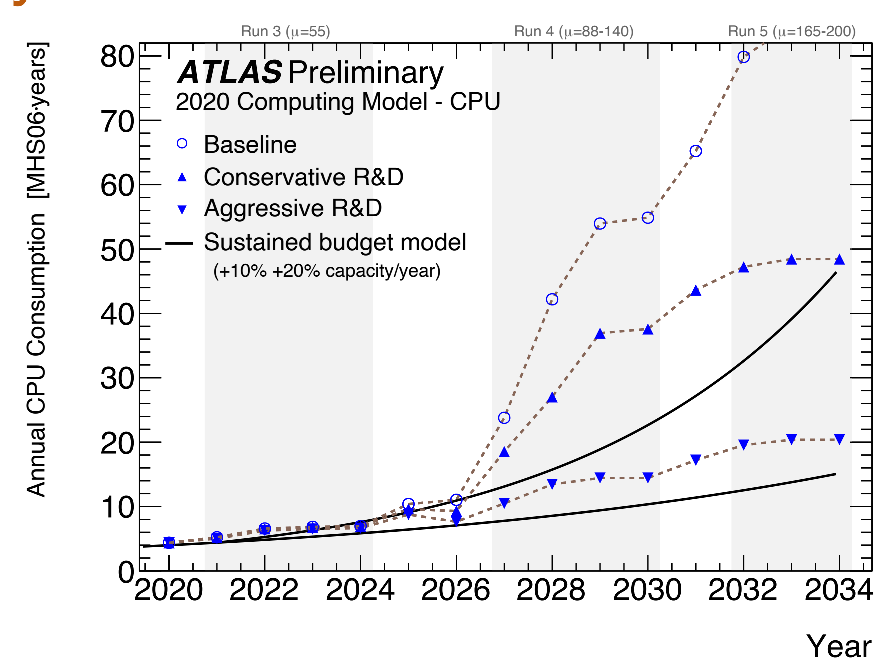




This is what we want to know

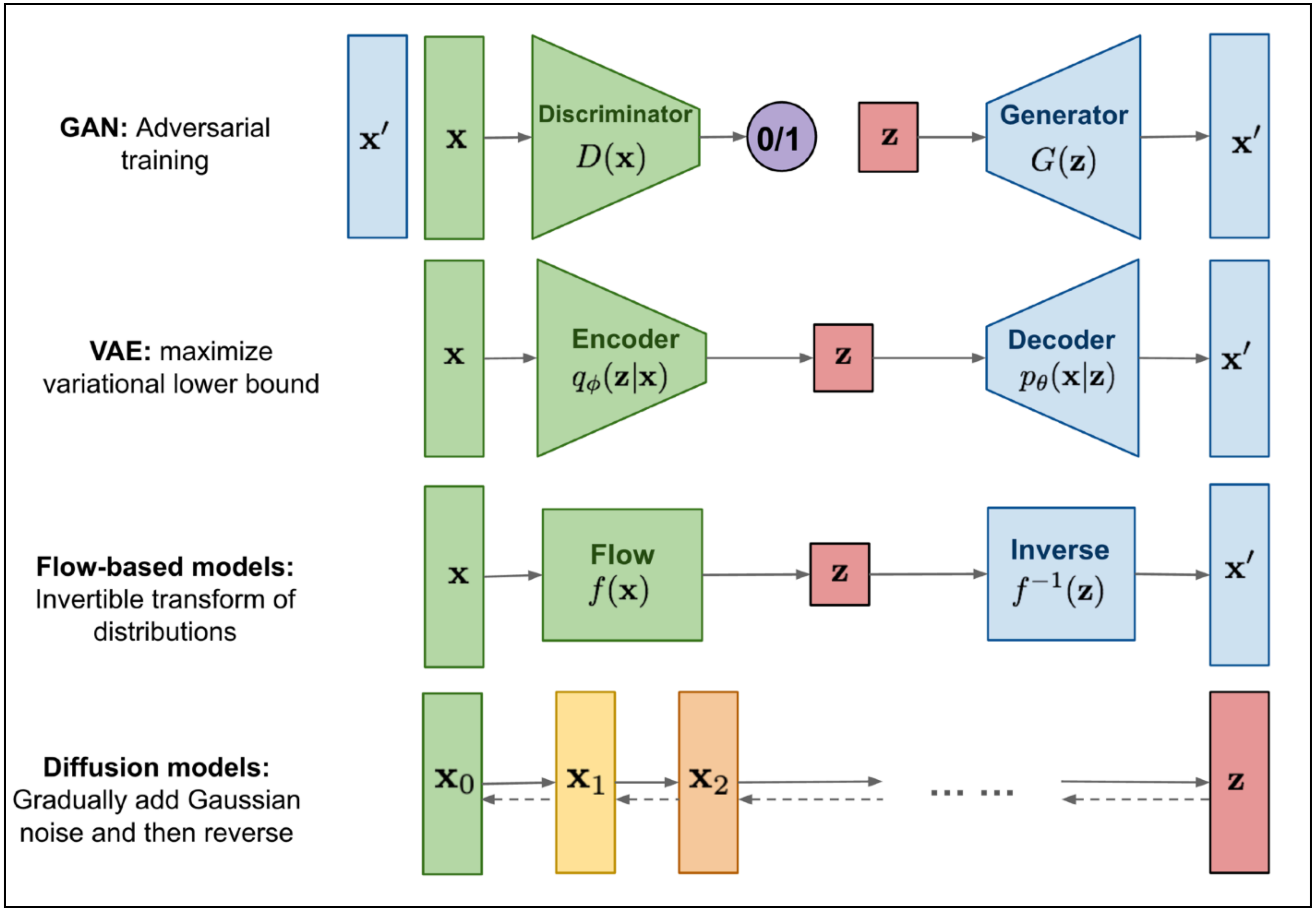
Simulation is crucial to connect experimental data with theory predictions, but computationally very costly





Generative Models

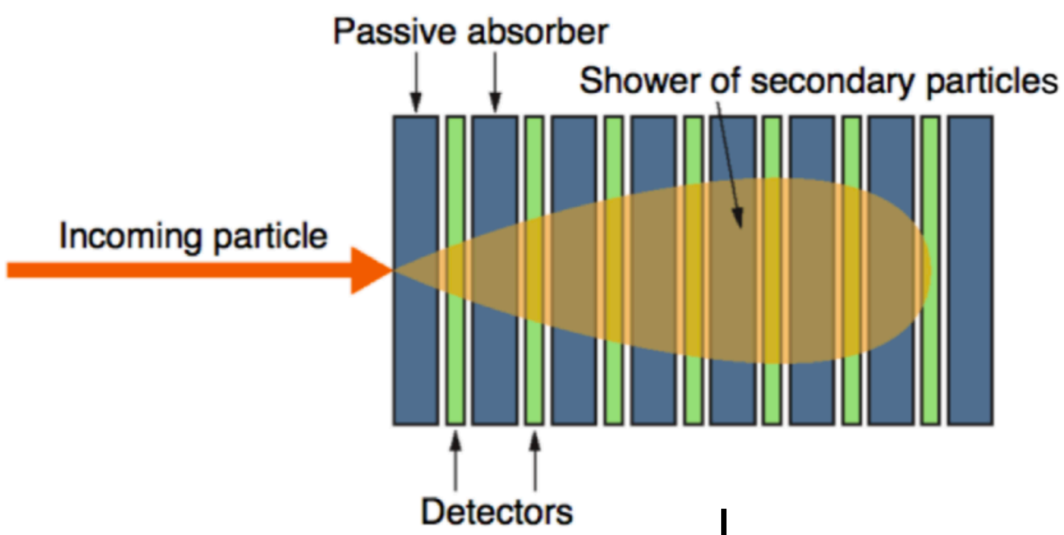
→Use generative models trained on simulation or data as efficient surrogates



Overview of generative architectures

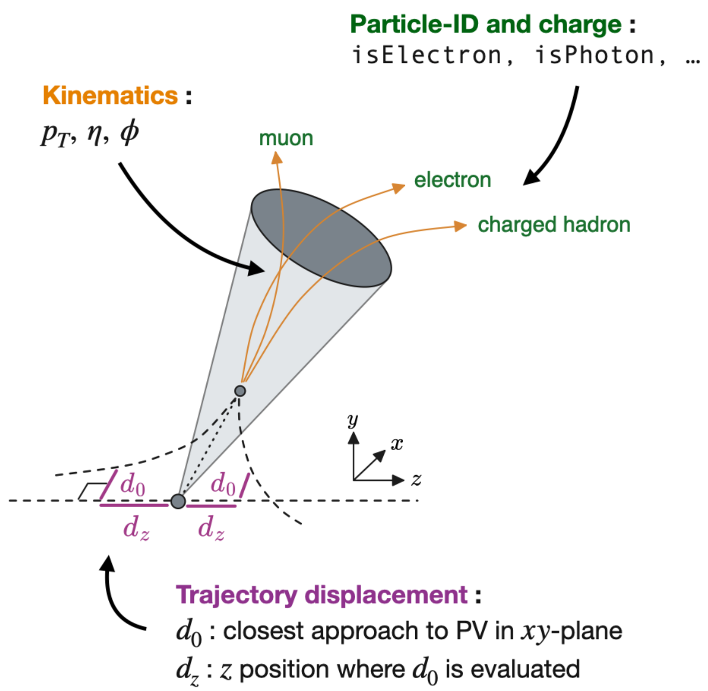
Simulation targets

Calorimeter Showers



Reduce computational bottleneck

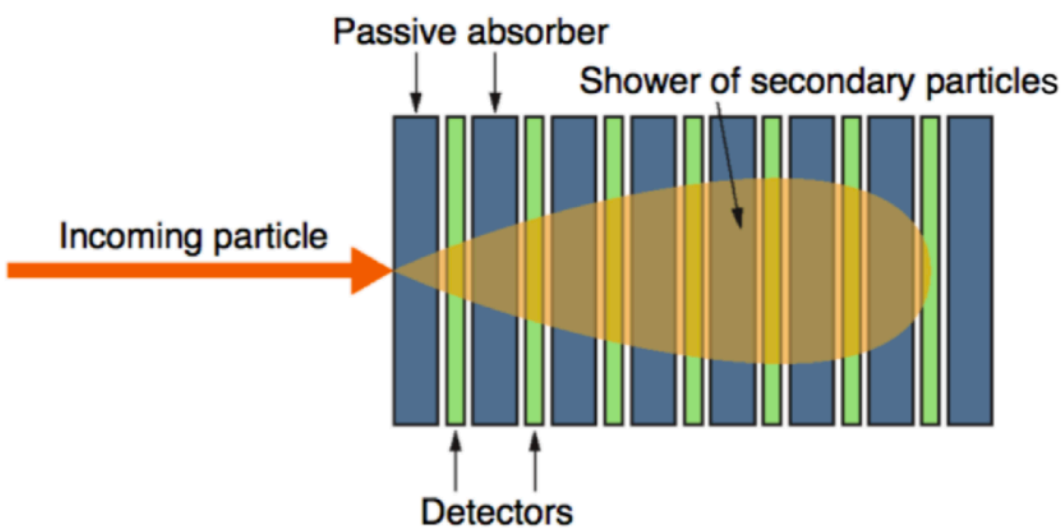
Jet Constituents



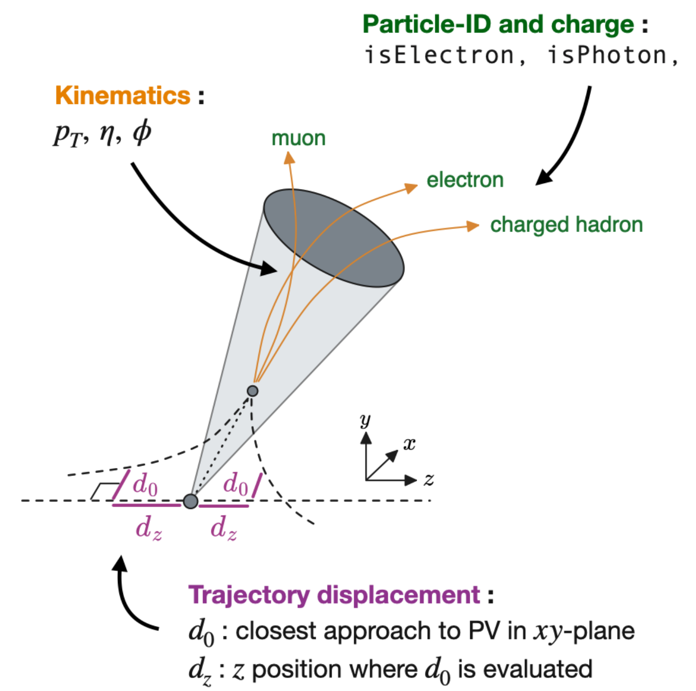
Learn from data

Simulation targets

Calorimeter Showers

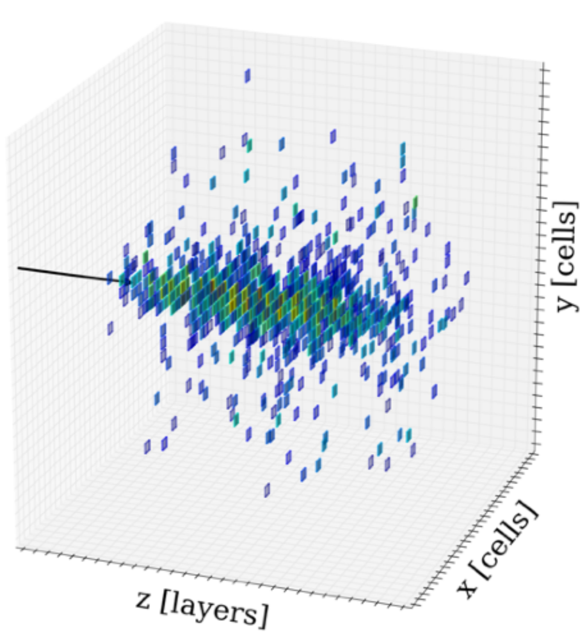


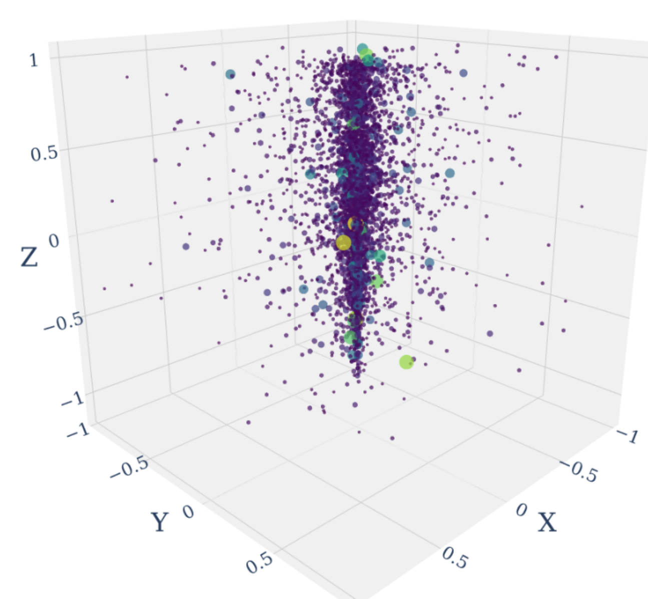
Jet Constituents

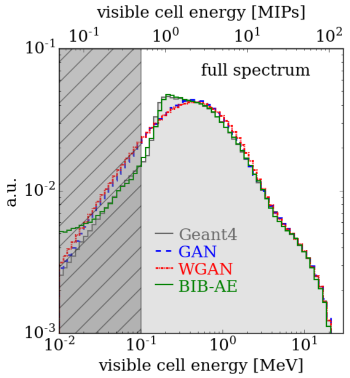


as fixed grid

as point cloud

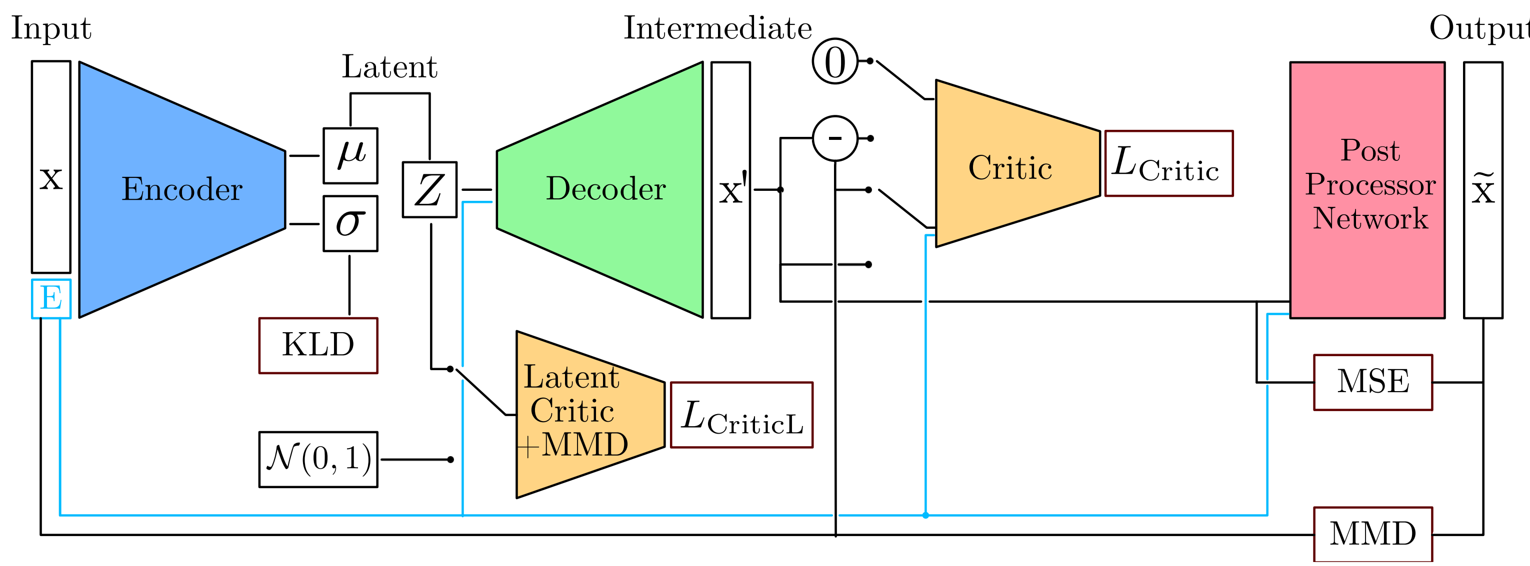




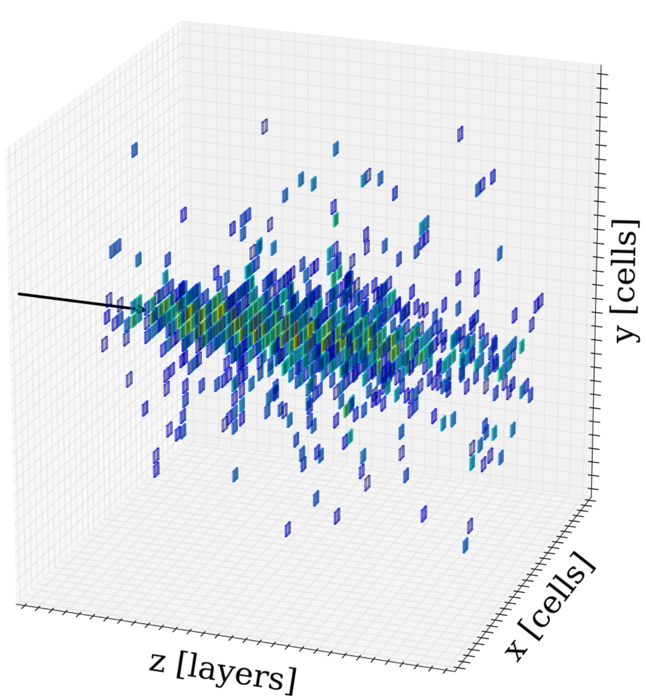


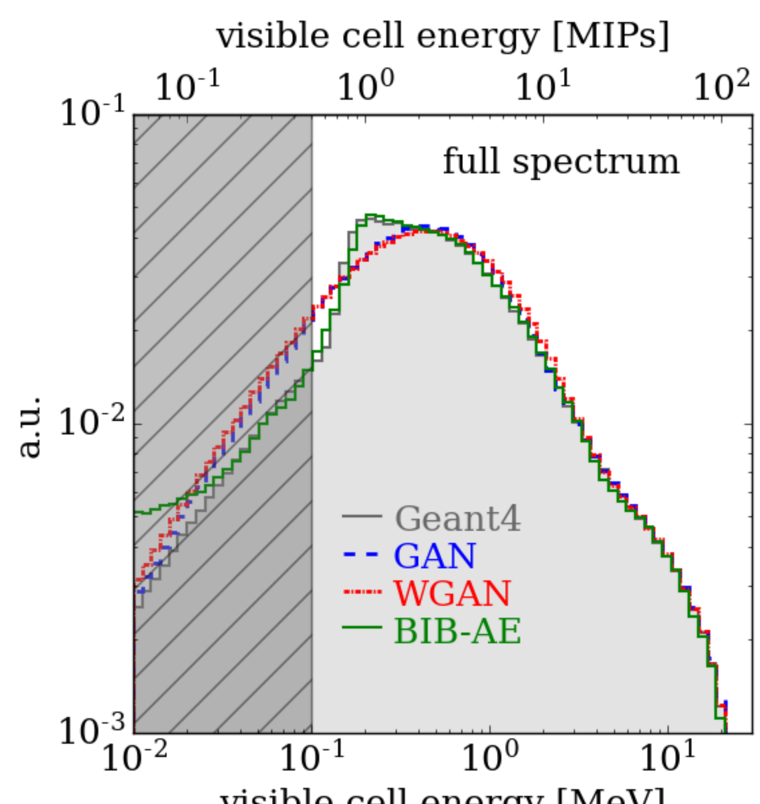
Generative progress

Progress



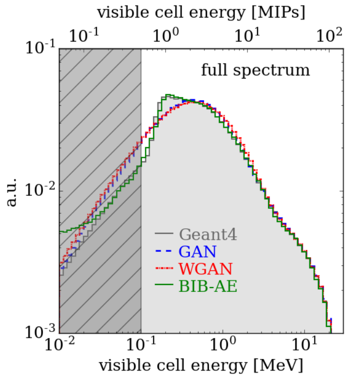
BIB-AE (GAN + VAE):
1st simulation of Photon shower in 27k cell calorimeter

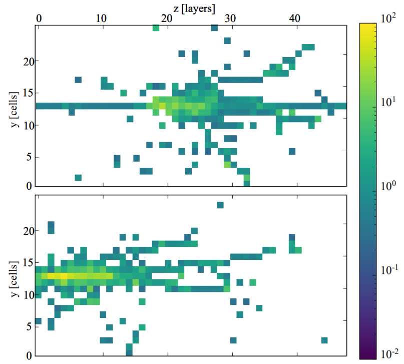




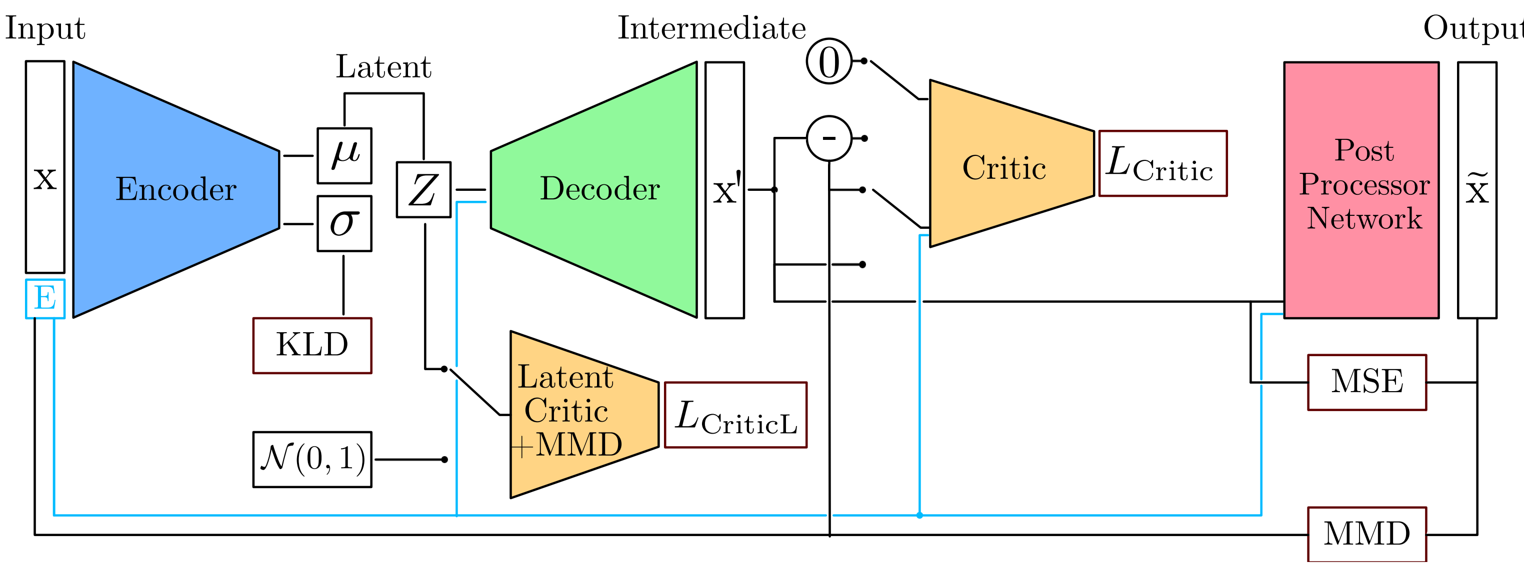
Buhmann, .., GK et al 2005.05334

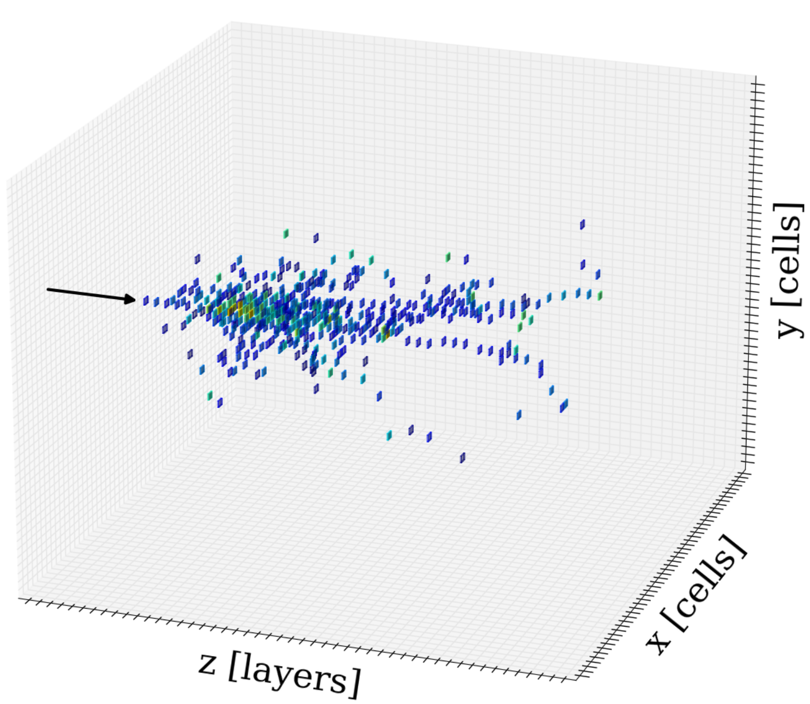
erative progress

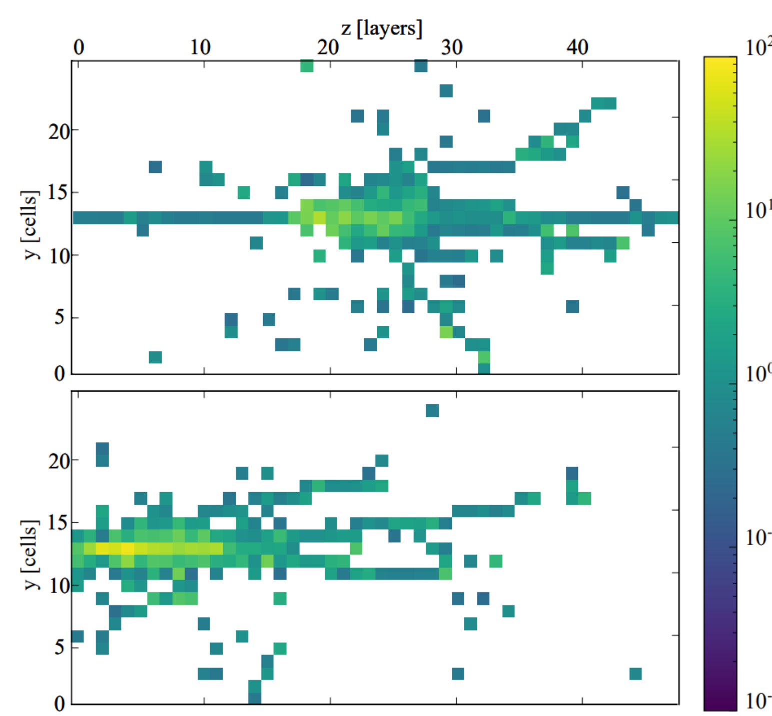




Progress

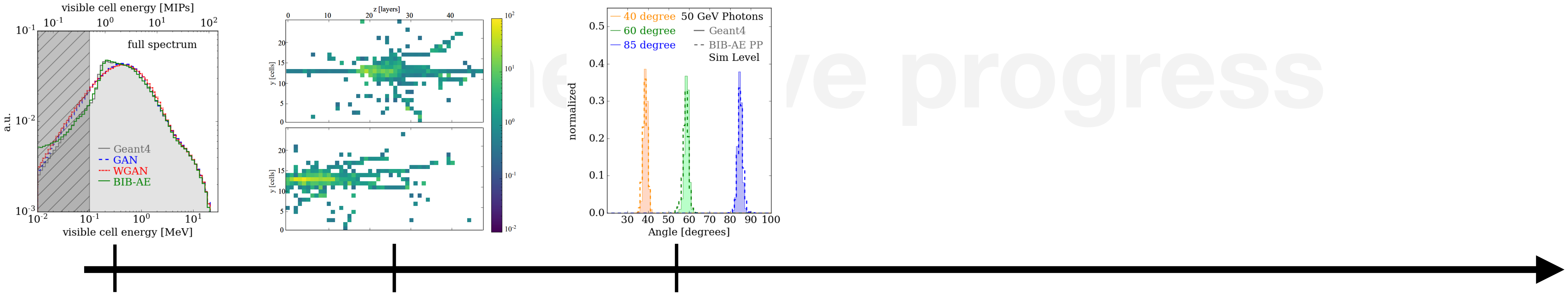




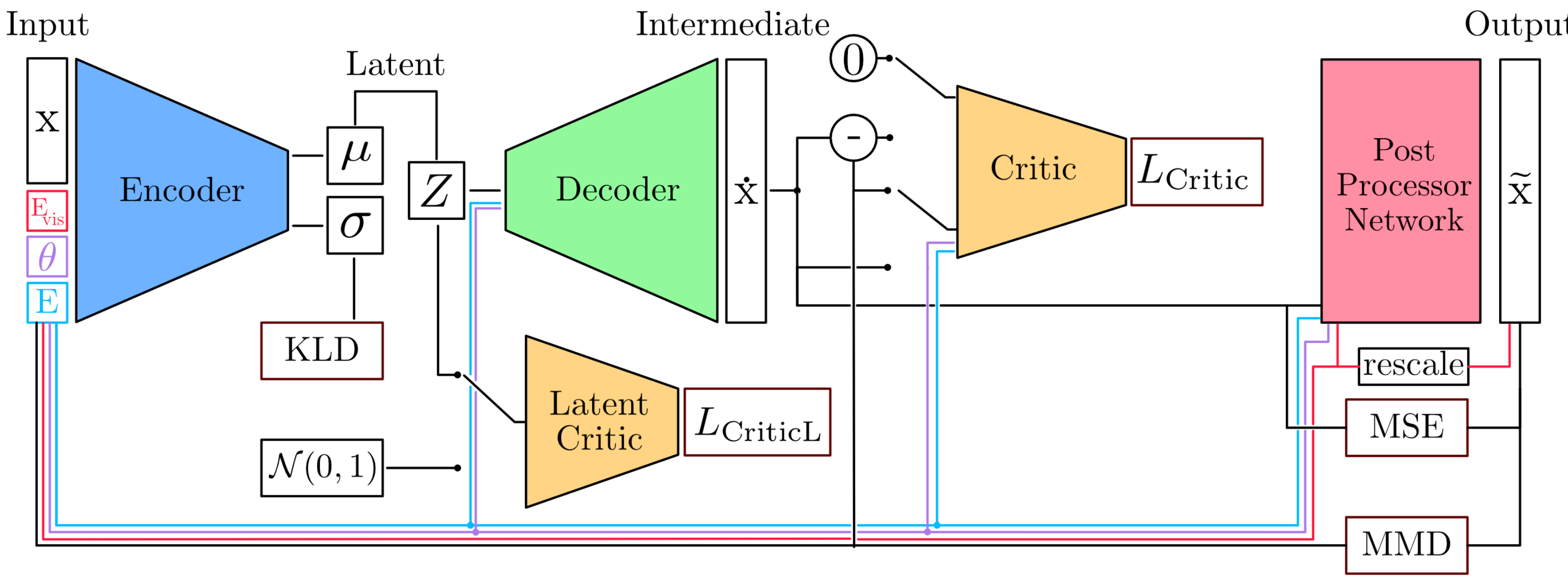


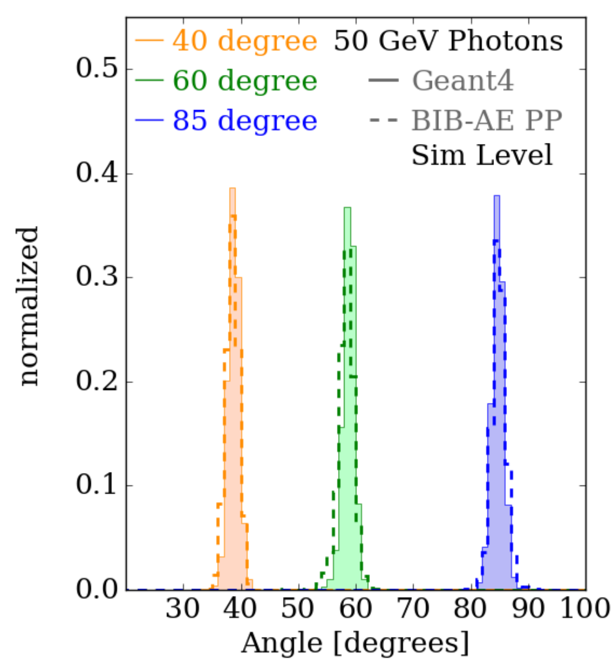
Handle more complex pion showers

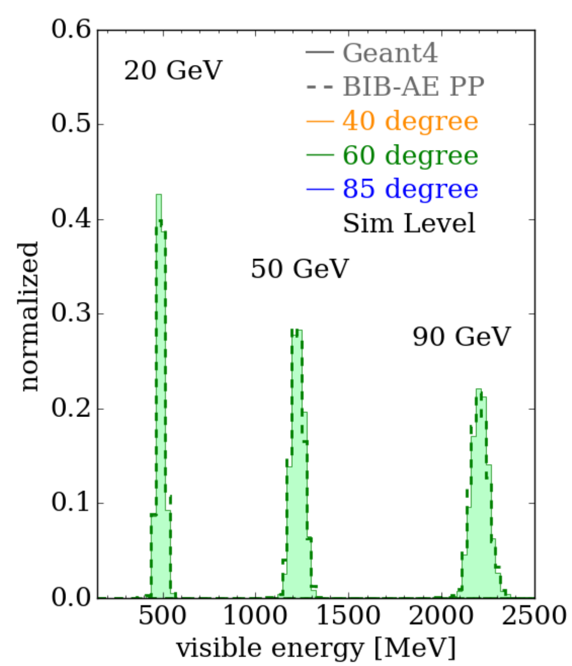
Buhmann, .., GK et al 2112.09709;



Progress

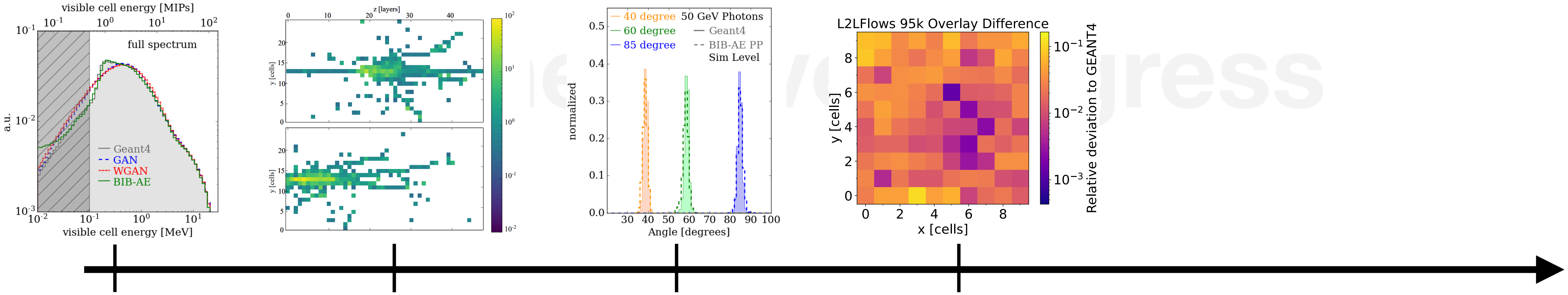




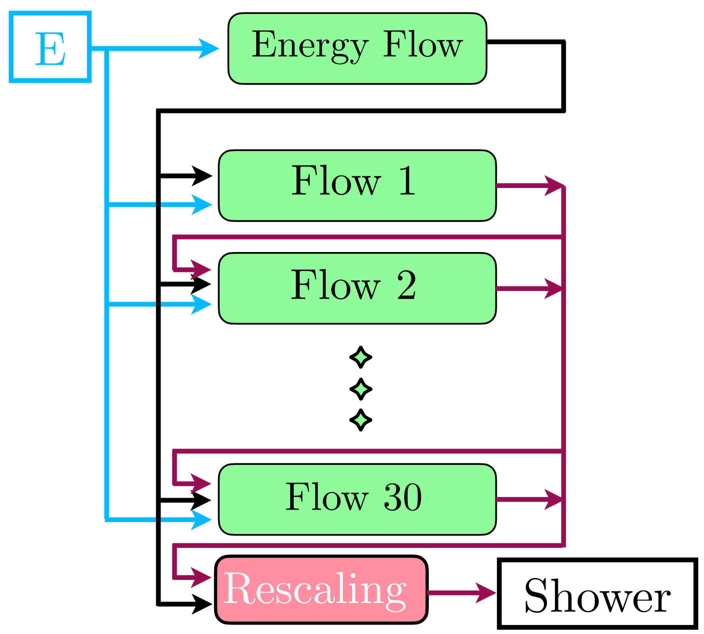


Extend to condition on angles

Diefenbacher, .., GK et al 2303.18150

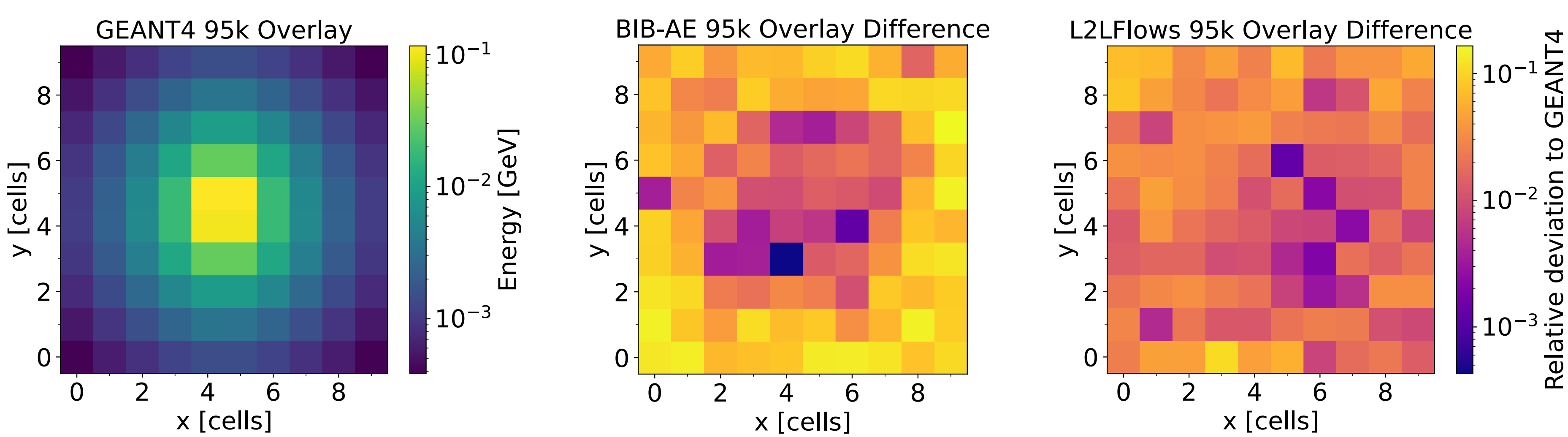


Progress

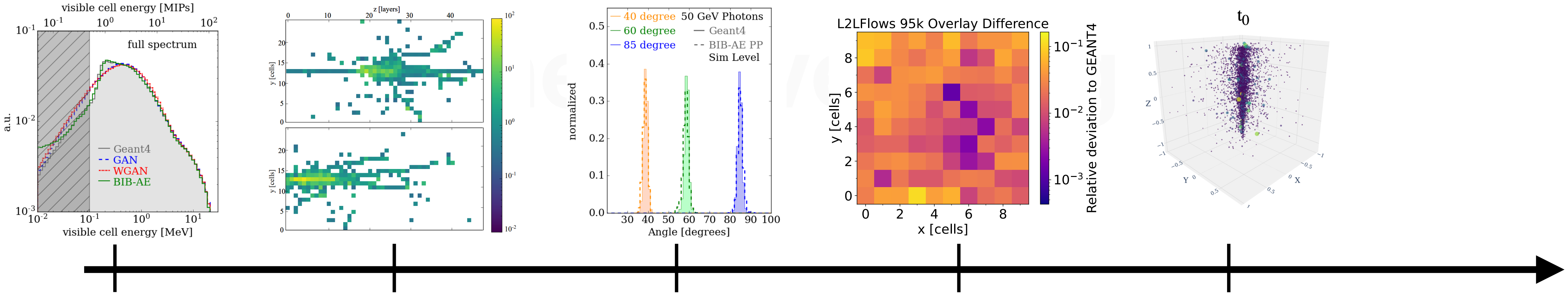


Better convergence of normalising flows:
→better fidelity

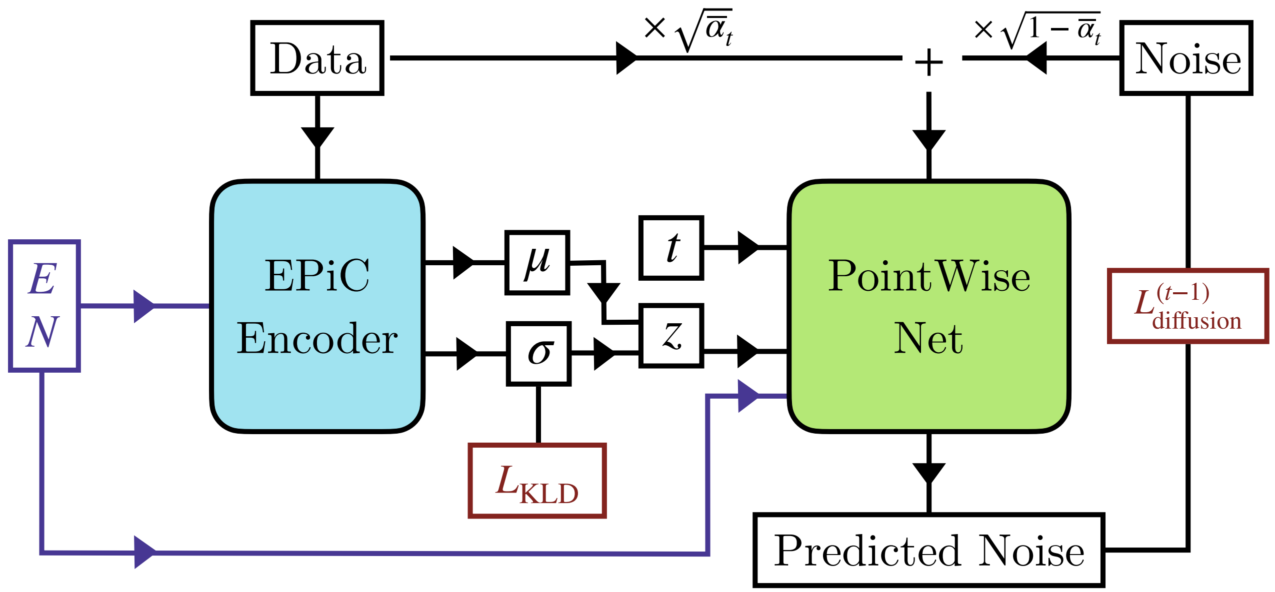
Individual flows per layer for efficiency



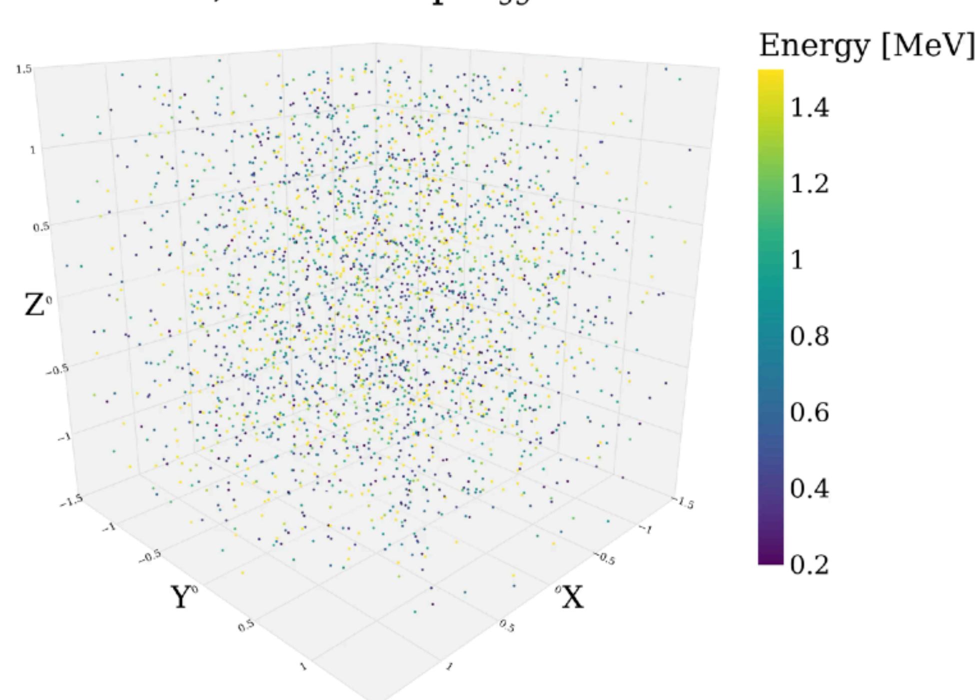
Diefenbacher, .., GK et al 2302.11594



Progress

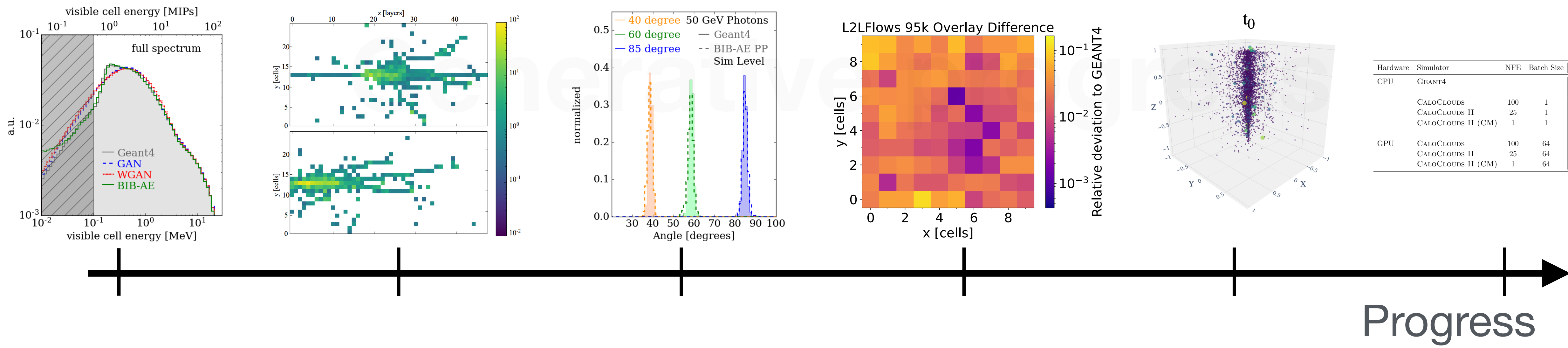


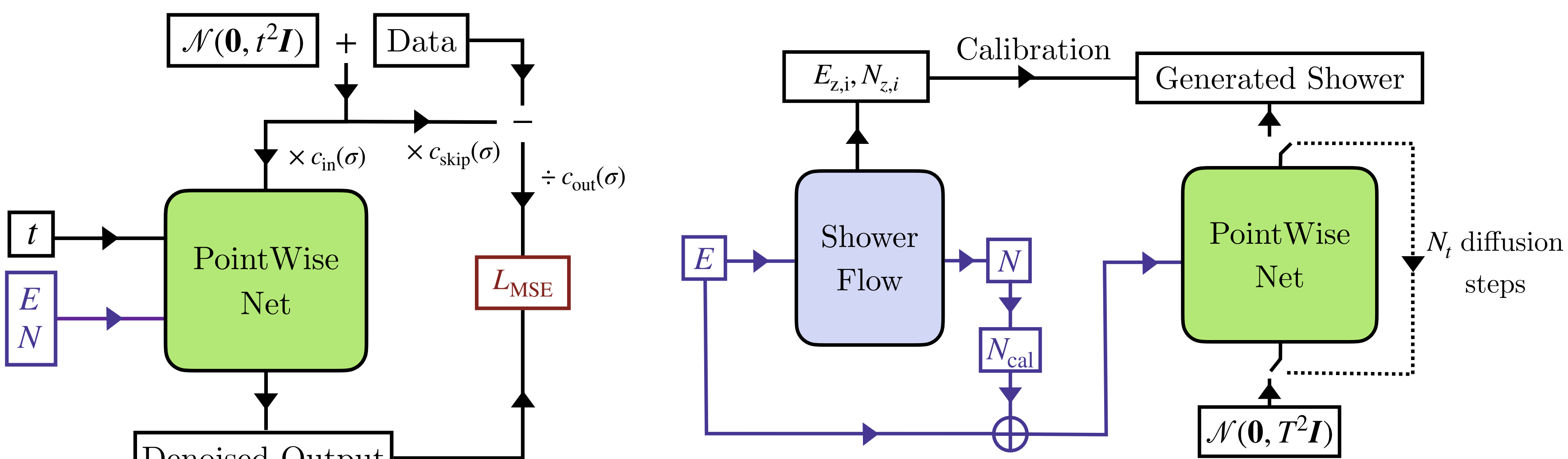
CaloCloud, time stamp: t_{99}



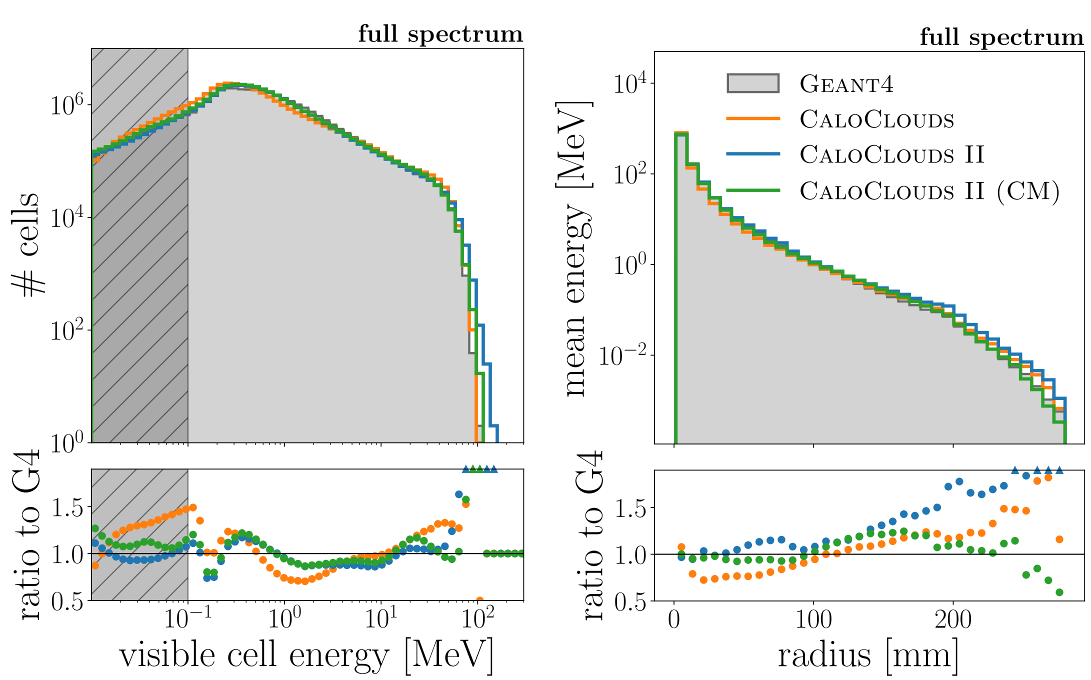
Add permutation invariance & diffusion:
→simulate as point cloud!

Buhmann, .., GK, et al 2305.04847





Speed up diffusion with continuos time and consistency distillation



Hardware	Simulator	NFE	Batch Size	Time / Shower [ms]	Speed-up
CPU	Geant4			3914.80 ± 74.09	×1
	CaloClouds	100	1	3146.71 ± 31.66	×1.2
	CaloClouds II	25	1	651.68 ± 4.21	×6.0
	CaloClouds II (CM)	1	1	84.35 ± 0.22	×46
GPU	CaloClouds	100	64	24.91 ± 0.72	×157
	CaloClouds II	25	64	6.12 ± 0.13	×640
	CaloClouds II (CM)	1	64	2.09 ± 0.13	×1873

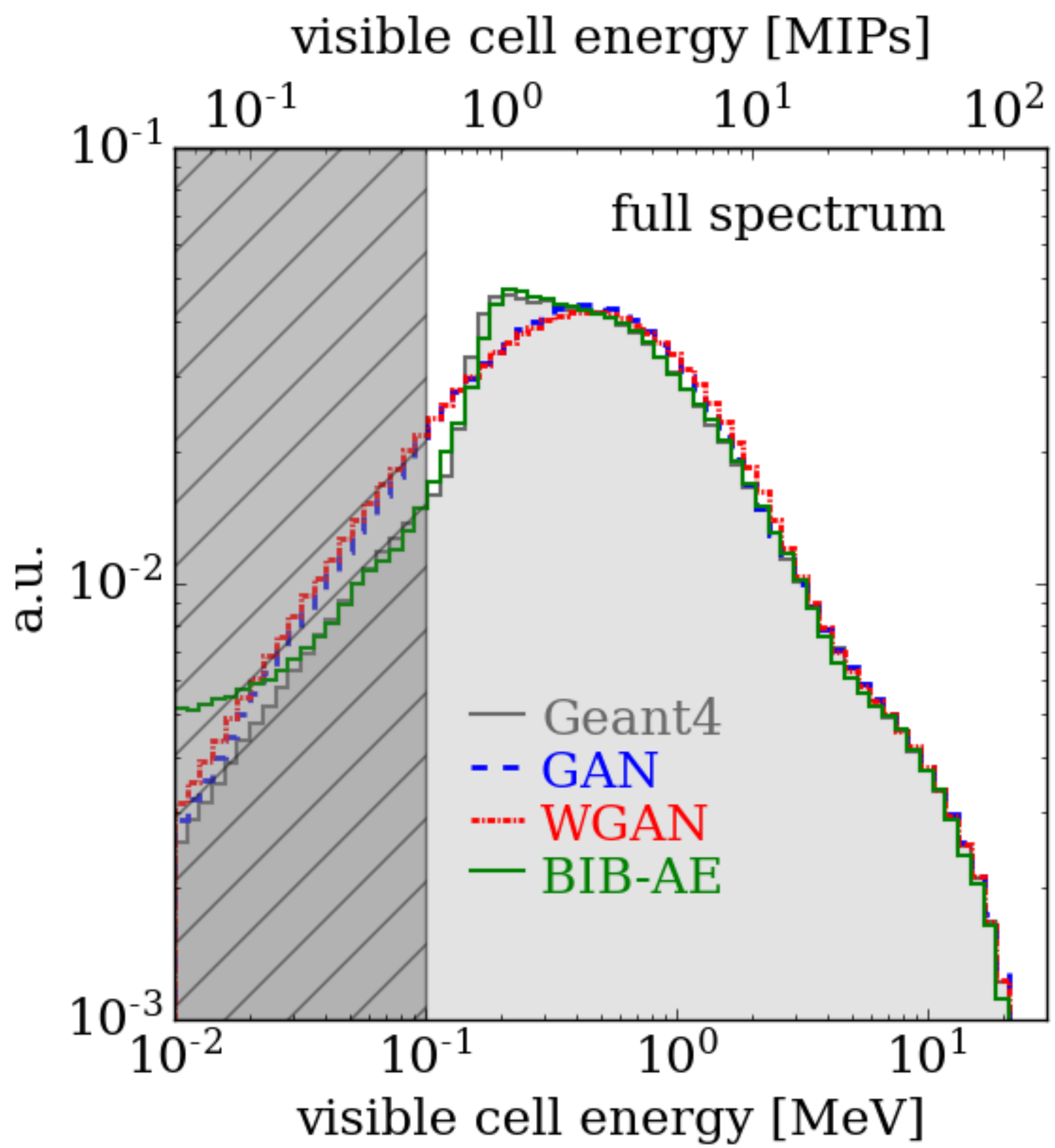
Buhmann, .., GK et al 2309.05704

Quality of simulation

GEANT4 Simulation ⟷ Generative Model

How well does the generative model describe the training data?

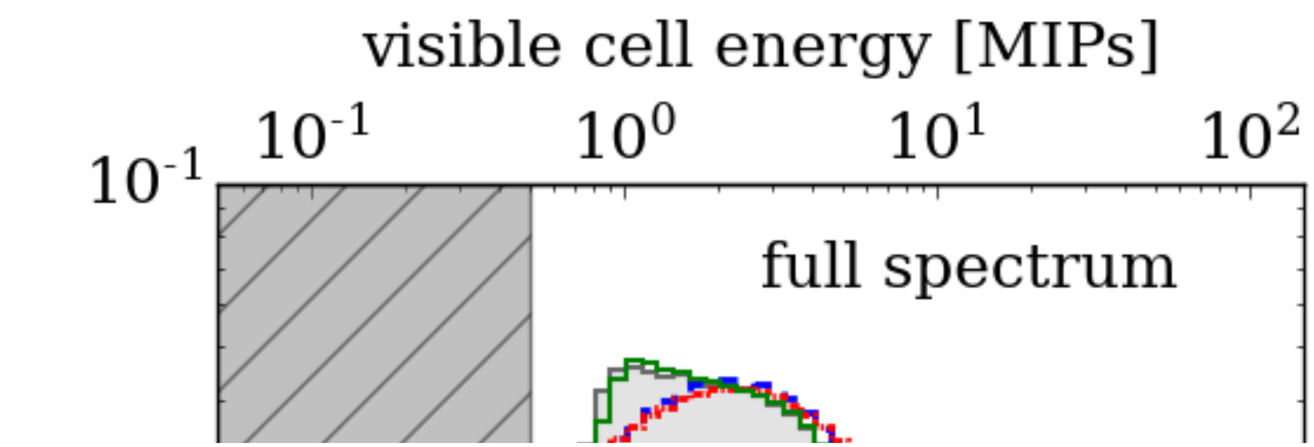
One-dimensional metrics



Visual inspection is insufficient

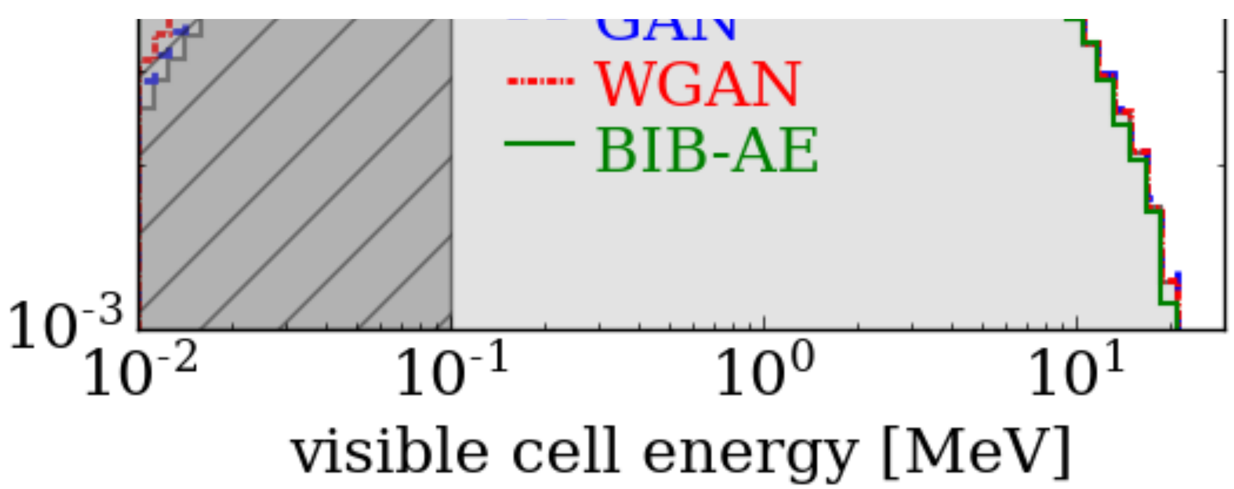
Buhmann, .., GK et al 2005.05334

One-dimensional metrics



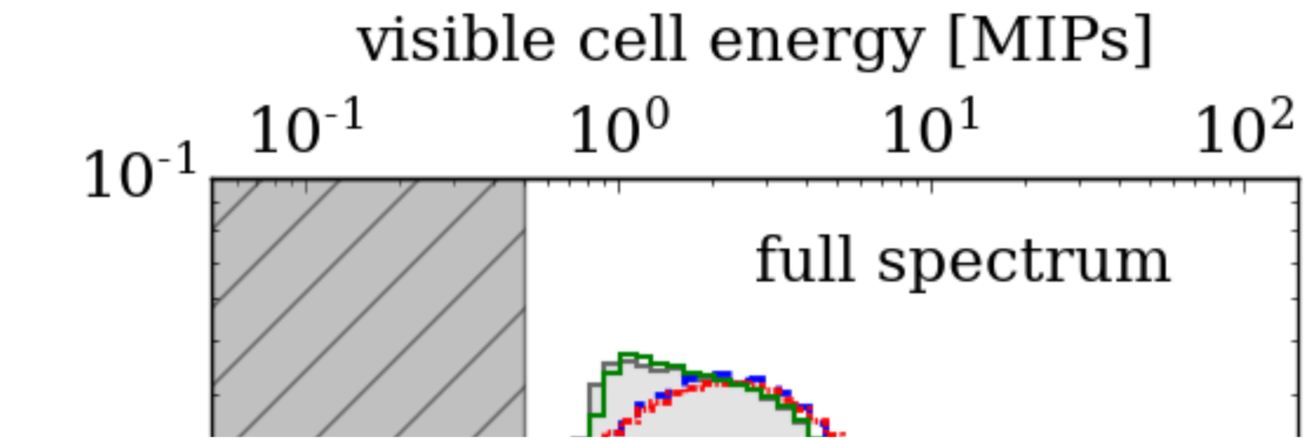
Wasserstein distance

Simulator	$W_1^{N_{\mathrm{hits}}}$ $(\times 10^{-3})$	$W_1^{E_{\mathrm{vis}}/E_{\mathrm{inc}}}$ $(\times 10^{-3})$	$W_1^{E_{\mathrm{cell}}}$ $(\times 10^{-3})$	$W_1^{E_{\mathrm{long}}}$ $(\times 10^{-3})$	$W_1^{E_{\mathrm{radial}}}$ $(\times 10^{-3})$	$W_1^{m_{1,X}}$ $(\times 10^{-3})$	$W_1^{m_{1,Y}}$ $(\times 10^{-3})$	$W_1^{m_{1,Z}}$ $(\times 10^{-3})$
Geant4	0.7 ± 0.2	0.8 ± 0.2	0.9 ± 0.4	0.7 ± 0.8	0.7 ± 0.1	0.9 ± 0.1	1.1 ± 0.3	0.9 ± 0.3
CaloClouds	**2.5 ± 0.3**	11.4 ± 0.4	15.9 ± 0.7	**2.0 ± 1.3**	38.8 ± 1.4	4.0 ± 0.4	8.7 ± 0.3	1.4 ± 0.5
CaloClouds II	3.6 ± 0.5	26.4 ± 0.4	**15.3 ± 0.6**	3.7 ± 1.6	11.6 ± 1.5	**2.4 ± 0.4**	**7.6 ± 0.2**	3.9 ± 0.4
CaloClouds II (CM)	6.1 ± 0.7	**9.8 ± 0.5**	16.0 ± 0.7	**2.0 ± 1.4**	**8.3 ± 1.9**	3.0 ± 0.4	9.5 ± 0.6	**1.2 ± 0.5**



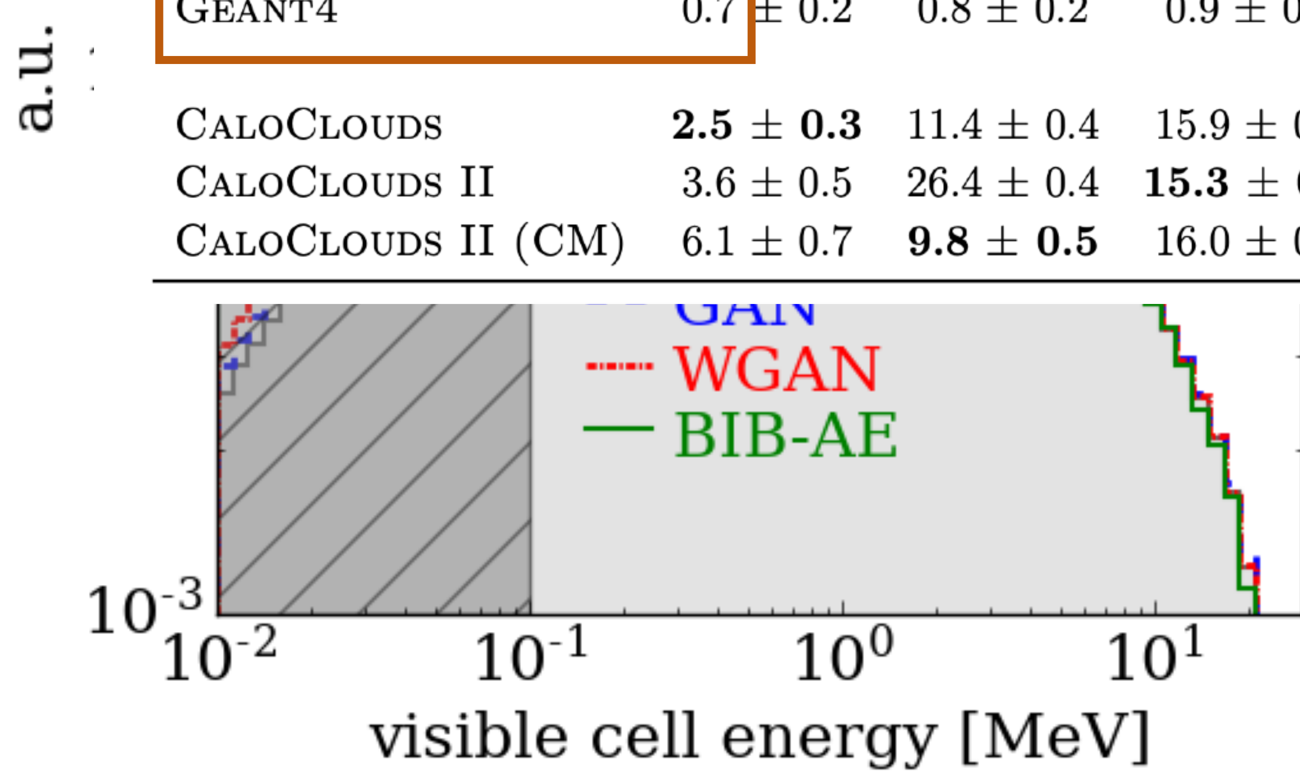
Buhmann, .., GK et al 2309.05704

One-dimensional metrics



Wasserstein distance

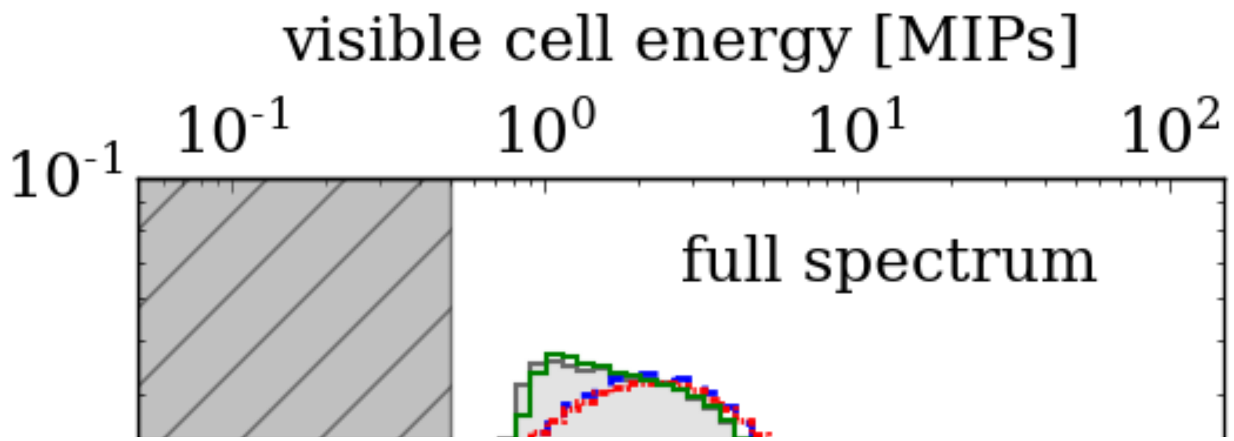
Simulator	$W_1^{N_{\text{hits}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{vis}}/E_{\text{inc}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{cell}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{long}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{radial}}}$ $(\times 10^{-3})$	$W_1^{m_{1,X}}$ $(\times 10^{-3})$	$W_1^{m_{1,Y}}$ $(\times 10^{-3})$	$W_1^{m_{1,Z}}$ $(\times 10^{-3})$
Geant4	0.7 ± 0.2	0.8 ± 0.2	0.9 ± 0.4	0.7 ± 0.8	0.7 ± 0.1	0.9 ± 0.1	1.1 ± 0.3	0.9 ± 0.3
CaloClouds	**2.5 ± 0.3**	11.4 ± 0.4	15.9 ± 0.7	**2.0 ± 1.3**	38.8 ± 1.4	4.0 ± 0.4	8.7 ± 0.3	1.4 ± 0.5
CaloClouds II	3.6 ± 0.5	26.4 ± 0.4	**15.3 ± 0.6**	3.7 ± 1.6	11.6 ± 1.5	**2.4 ± 0.4**	**7.6 ± 0.2**	3.9 ± 0.4
CaloClouds II (CM)	6.1 ± 0.7	**9.8 ± 0.5**	16.0 ± 0.7	**2.0 ± 1.4**	**8.3 ± 1.9**	3.0 ± 0.4	9.5 ± 0.6	**1.2 ± 0.5**



Floor is given by sample-size effects of GEANT4 vs itself

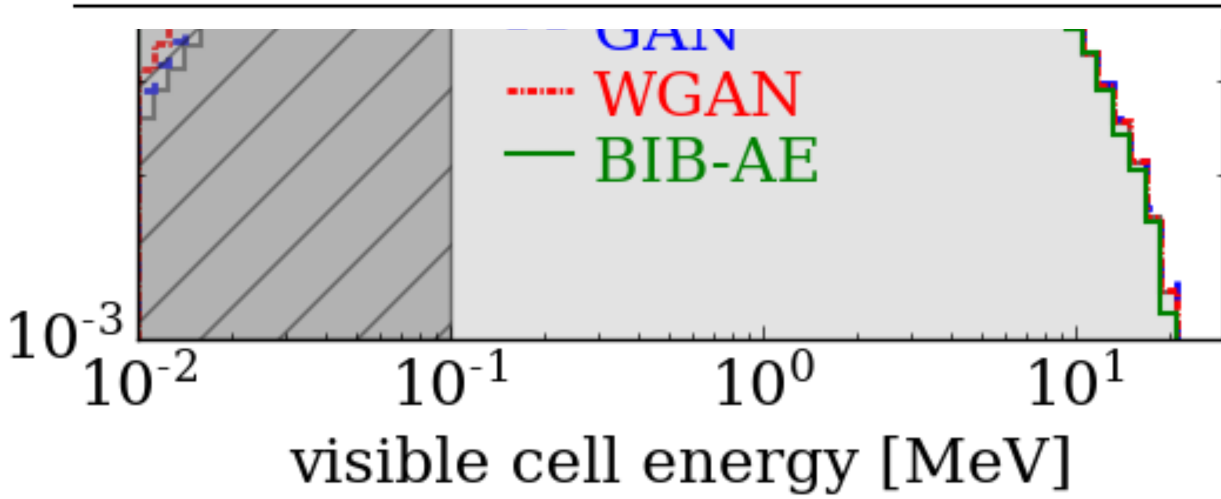
Buhmann, .., GK et al 2309.05704

One-dimensional metrics



Wasserstein distance

Simulator	$W_1^{N_{\text{hits}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{vis}}/E_{\text{inc}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{cell}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{long}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{radial}}}$ $(\times 10^{-3})$	$W_1^{m_{1,X}}$ $(\times 10^{-3})$	$W_1^{m_{1,Y}}$ $(\times 10^{-3})$	$W_1^{m_{1,Z}}$ $(\times 10^{-3})$
GEANT4	0.7 ± 0.2	0.8 ± 0.2	0.9 ± 0.4	0.7 ± 0.8	0.7 ± 0.1	0.9 ± 0.1	1.1 ± 0.3	0.9 ± 0.3
CALOCLOUDS	**2.5 ± 0.3**	11.4 ± 0.4	15.9 ± 0.7	**2.0 ± 1.3**	38.8 ± 1.4	4.0 ± 0.4	8.7 ± 0.3	1.4 ± 0.5
CALOCLOUDS II	3.6 ± 0.5	26.4 ± 0.4	**15.3 ± 0.6**	3.7 ± 1.6	11.6 ± 1.5	**2.4 ± 0.4**	**7.6 ± 0.2**	3.9 ± 0.4
CALOCLOUDS II (CM)	6.1 ± 0.7	**9.8 ± 0.5**	16.0 ± 0.7	**2.0 ± 1.4**	**8.3 ± 1.9**	3.0 ± 0.4	9.5 ± 0.6	**1.2 ± 0.5**

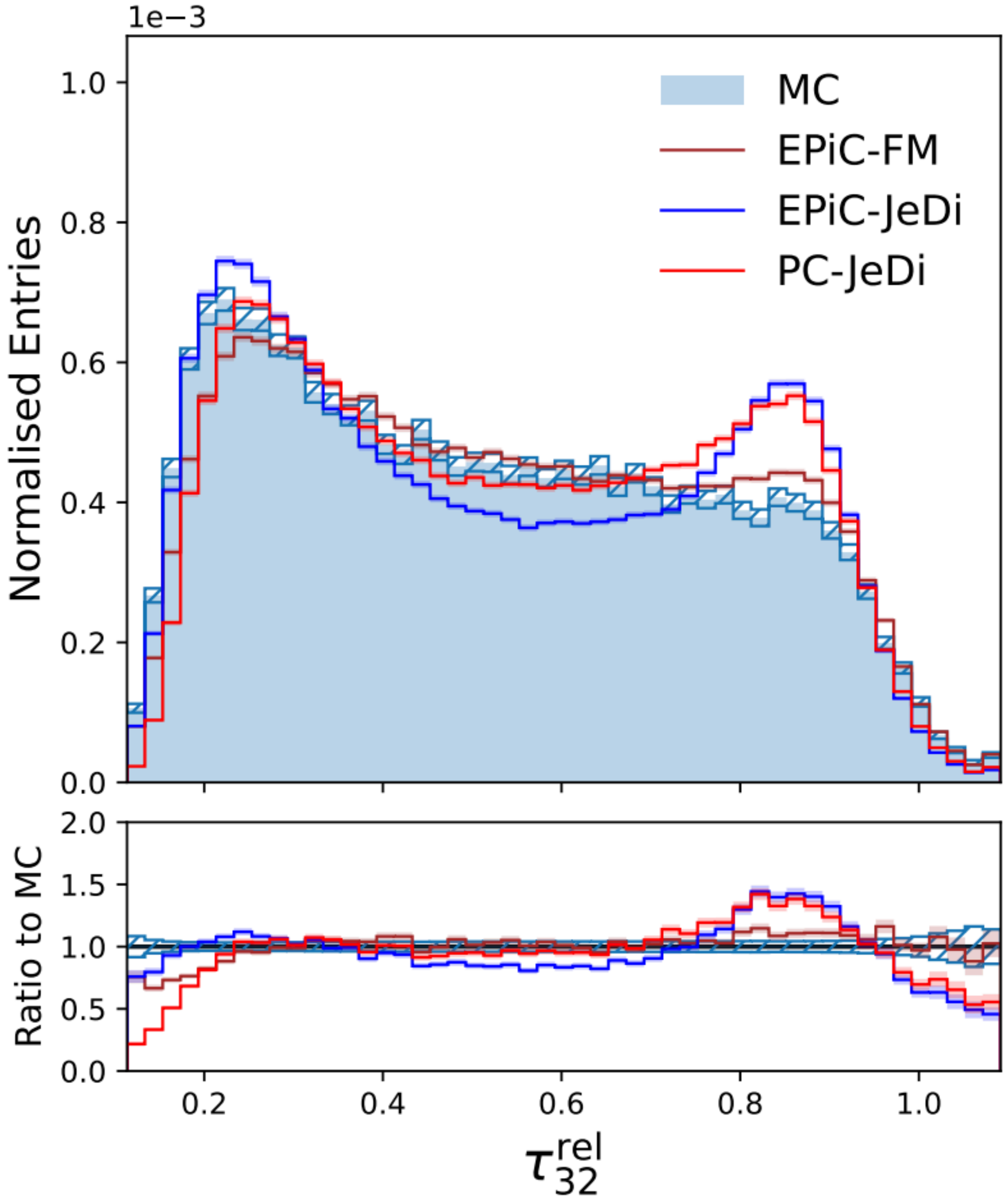


Floor is given by sample-size effects of GEANT4 vs itself

Uncertainty is standard deviation of 10 independent samples

Buhmann, .., GK et al 2309.05704

One-dimensional metrics



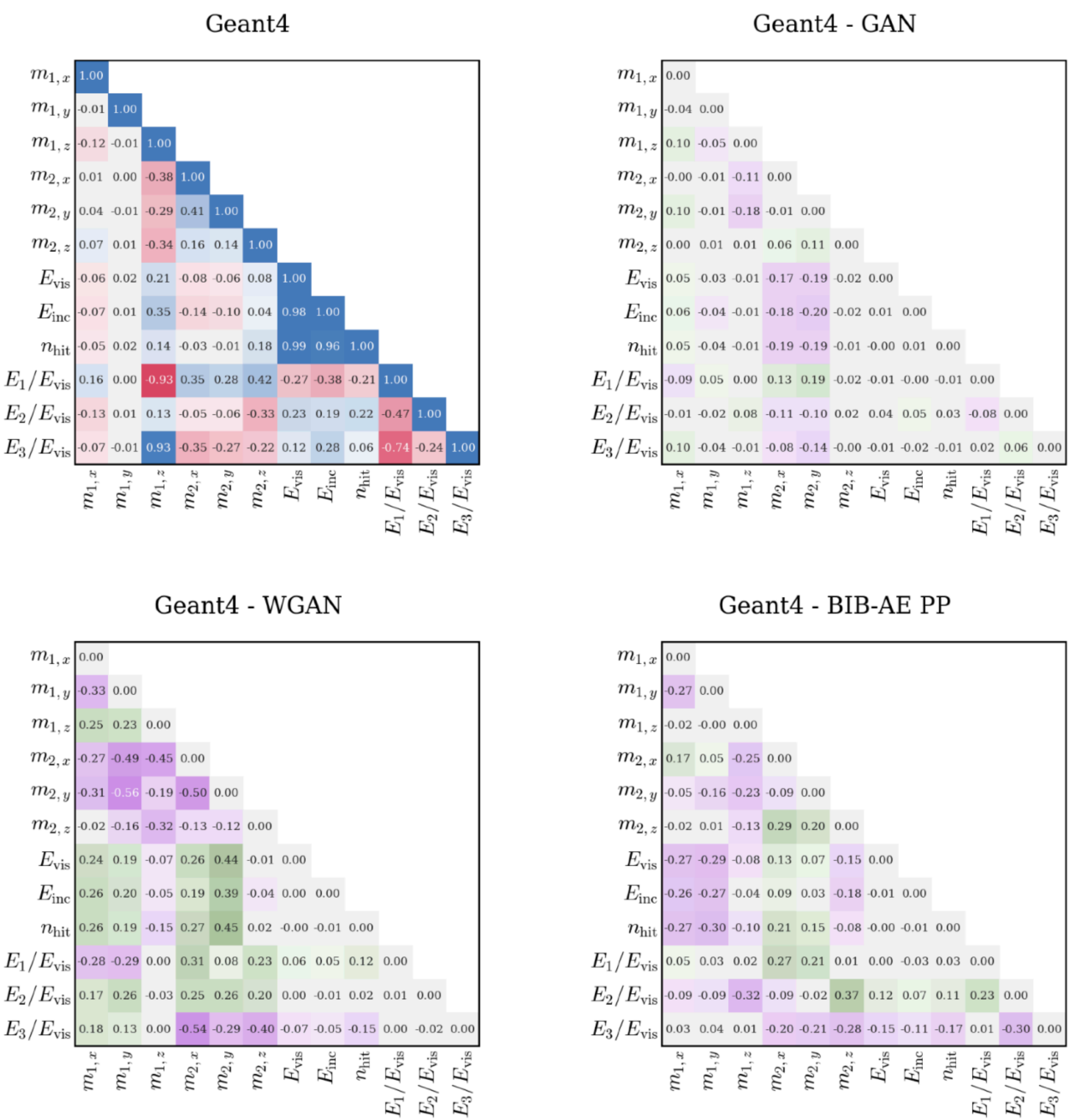
Wasserstein distance

More robust, well defined also for non-overlapping distributions

Kullback-Leibler divergence

More intuitive for shape-mismodelling

Two-dimensional metrics



Pair-wise correlations contain more information

Multi-dimensional metrics

# Showers per simulator	AUC Geant4 vs L2LFlows	AUC Geant4 vs BIB-AE
95k	0.8518 ± 0.0042	0.9947 ± 0.0025
190k	0.8768 ± 0.0029	–
380k	0.8962 ± 0.0024	–
760k	0.9402 ± 0.0011	–

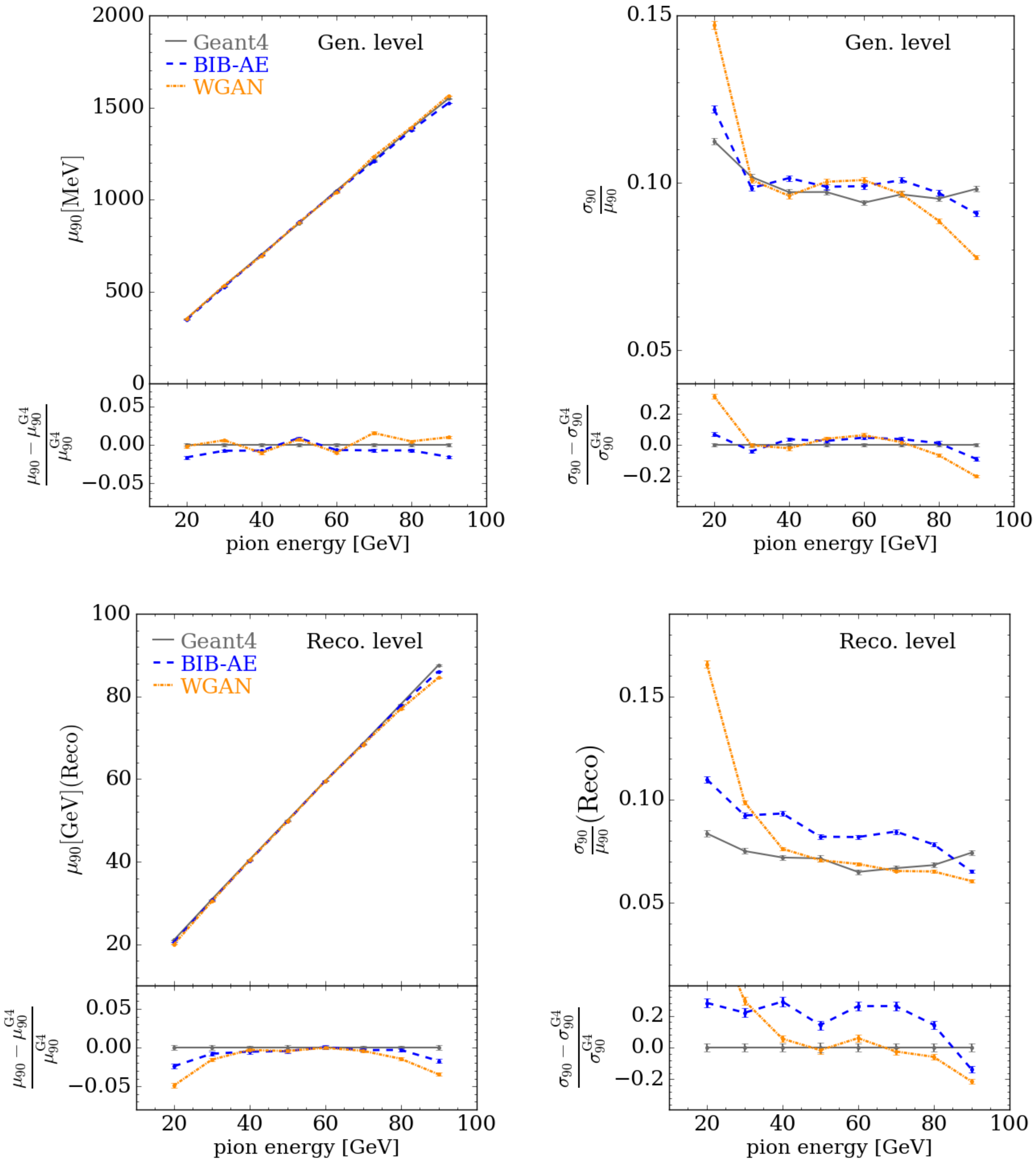
Capture full phase space information with classifiers

Still depends on training data

Choice of classifier

How good, is good enough really?

Adding reconstruction



Without reconstruction

With reconstruction

→ Non-linear effects of reconstruction can change relative performance

Taking stock so far

GEANT4 Simulation ⟷ **Generative Model**

Extensive set of metrics to judge quality

Useful for ranking generators

Less useful to make an absolute decision “good enough”

Taking stock so far

GEANT4 Simulation ⟷ **Generative Model**

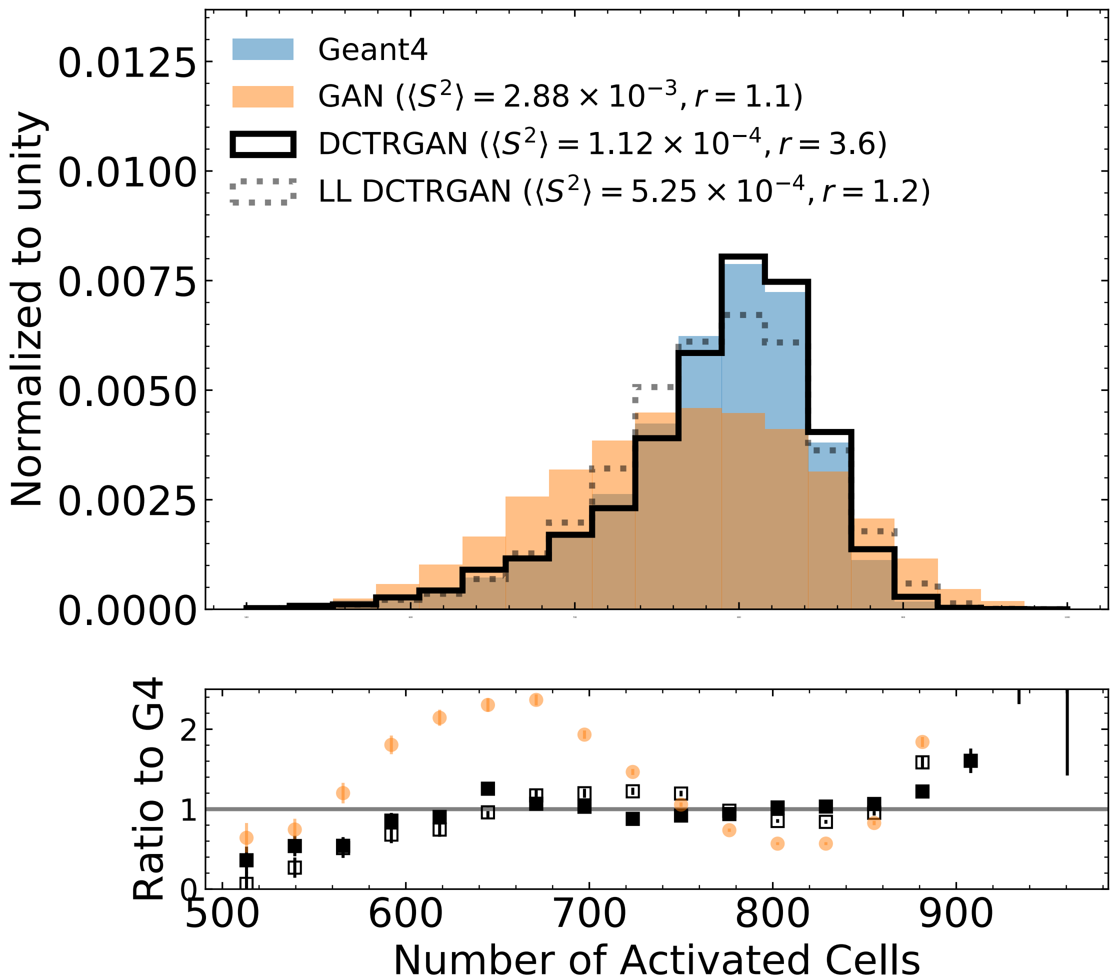
Extensive set of metrics to judge quality

Useful for ranking generators

Less useful to make an absolute decision “good enough”

But can additionally correct

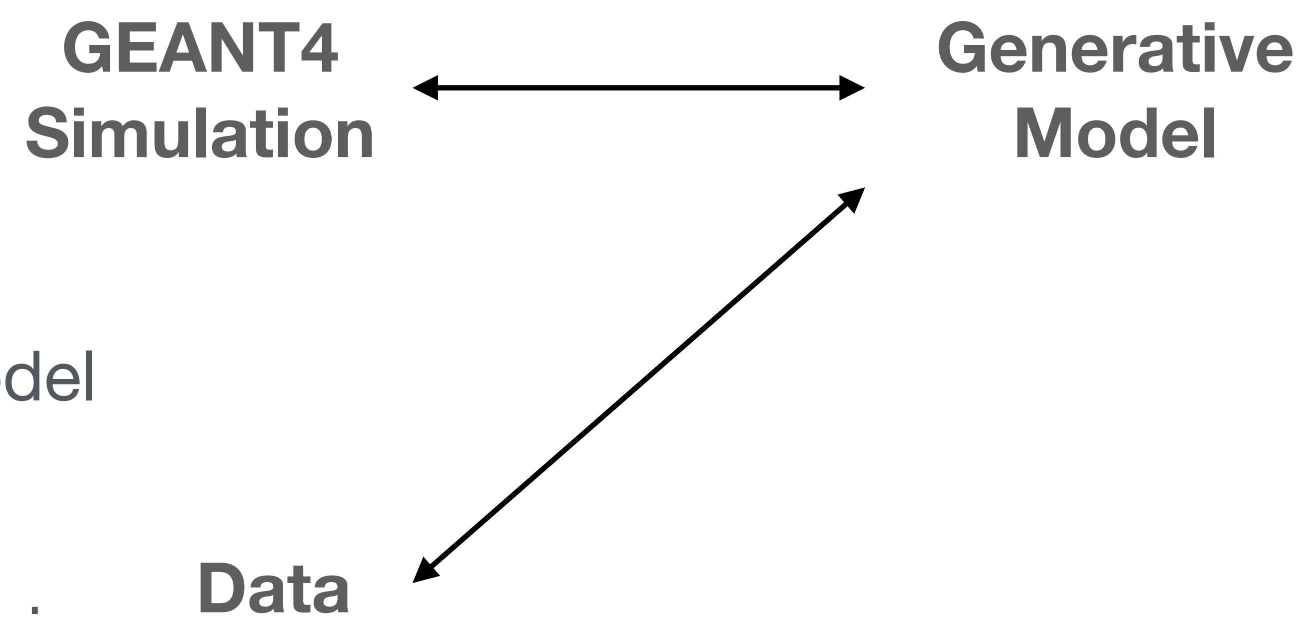
DCTRAN



Train classifiers to reweight distributions

Diefenbacher, .., GK et al 2009.03796

Calorimeter Summary

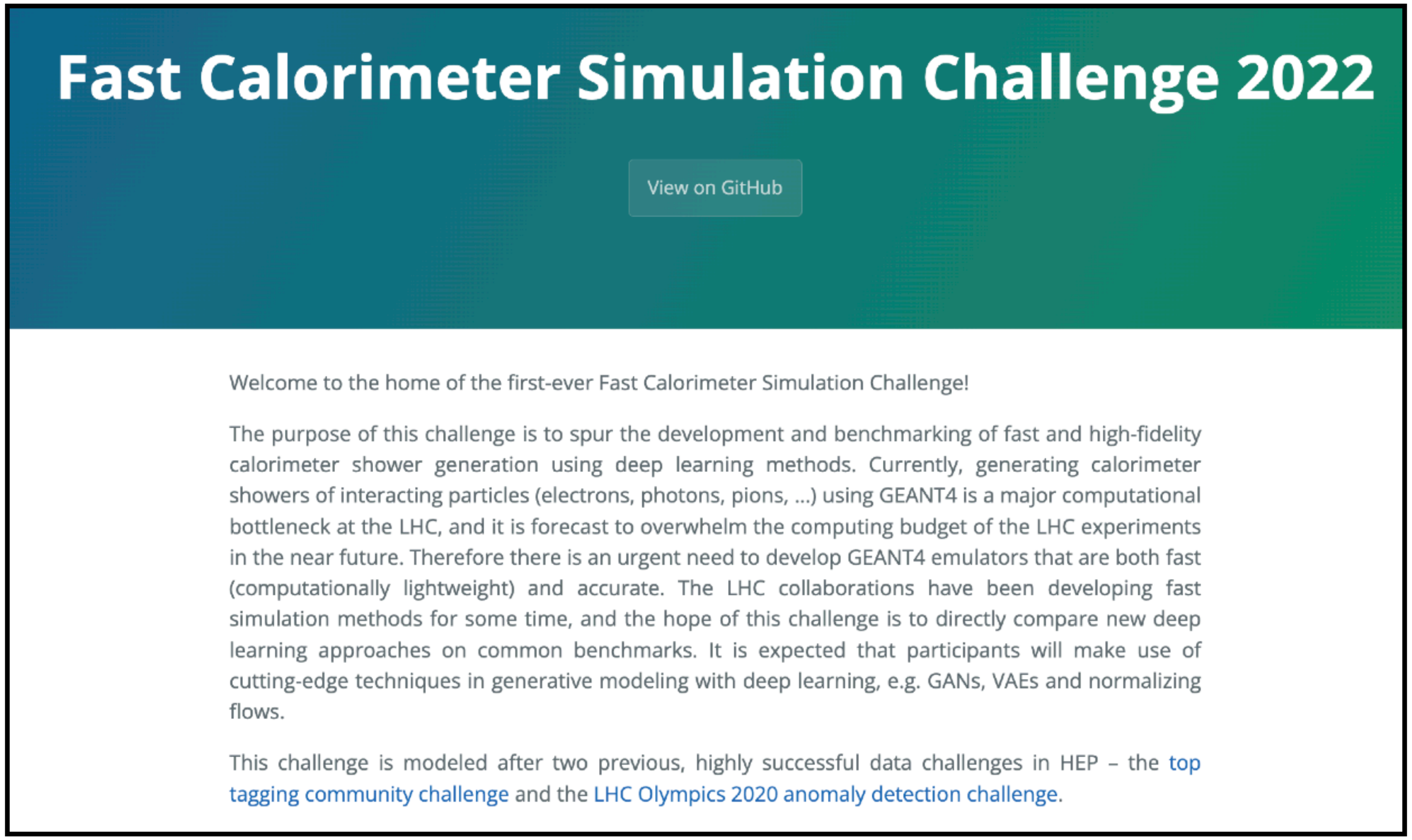


Option to calibrate generative model directly against data?

Standard way: Calibrate final physics observables.

Alternative idea: Per-shower uncertainty?
How to measure from data? (Different from training uncertainty!)
How to combine?
Still need propagate effect to physics observables.

Calo Challenge



Fast Calorimeter Simulation Challenge 2022

View on GitHub

Welcome to the home of the first-ever Fast Calorimeter Simulation Challenge!

The purpose of this challenge is to spur the development and benchmarking of fast and high-fidelity calorimeter shower generation using deep learning methods. Currently, generating calorimeter showers of interacting particles (electrons, photons, pions, ...) using GEANT4 is a major computational bottleneck at the LHC, and it is forecast to overwhelm the computing budget of the LHC experiments in the near future. Therefore there is an urgent need to develop GEANT4 emulators that are both fast (computationally lightweight) and accurate. The LHC collaborations have been developing fast simulation methods for some time, and the hope of this challenge is to directly compare new deep learning approaches on common benchmarks. It is expected that participants will make use of cutting-edge techniques in generative modeling with deep learning, e.g. GANs, VAEs and normalizing flows.

This challenge is modeled after two previous, highly successful data challenges in HEP – the top tagging community challenge and the LHC Olympics 2020 anomaly detection challenge.

Final phase of generative calorimeter challenge

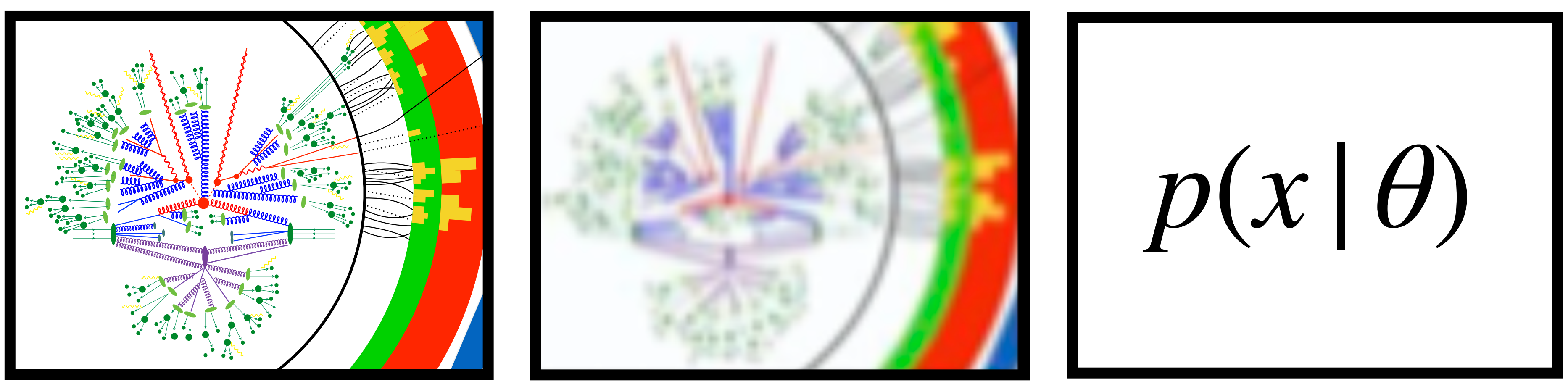
See Claudius' talk at ML4Jets for latest

Emphemeral Learning

Emphemeral Learning

- Trigger:
  - Only able to store a subset (<1 in 10.000) of events
- Possible alternative:
  - Train a generative model online during data taking

No compression | Compress per event | Compress entire dataset

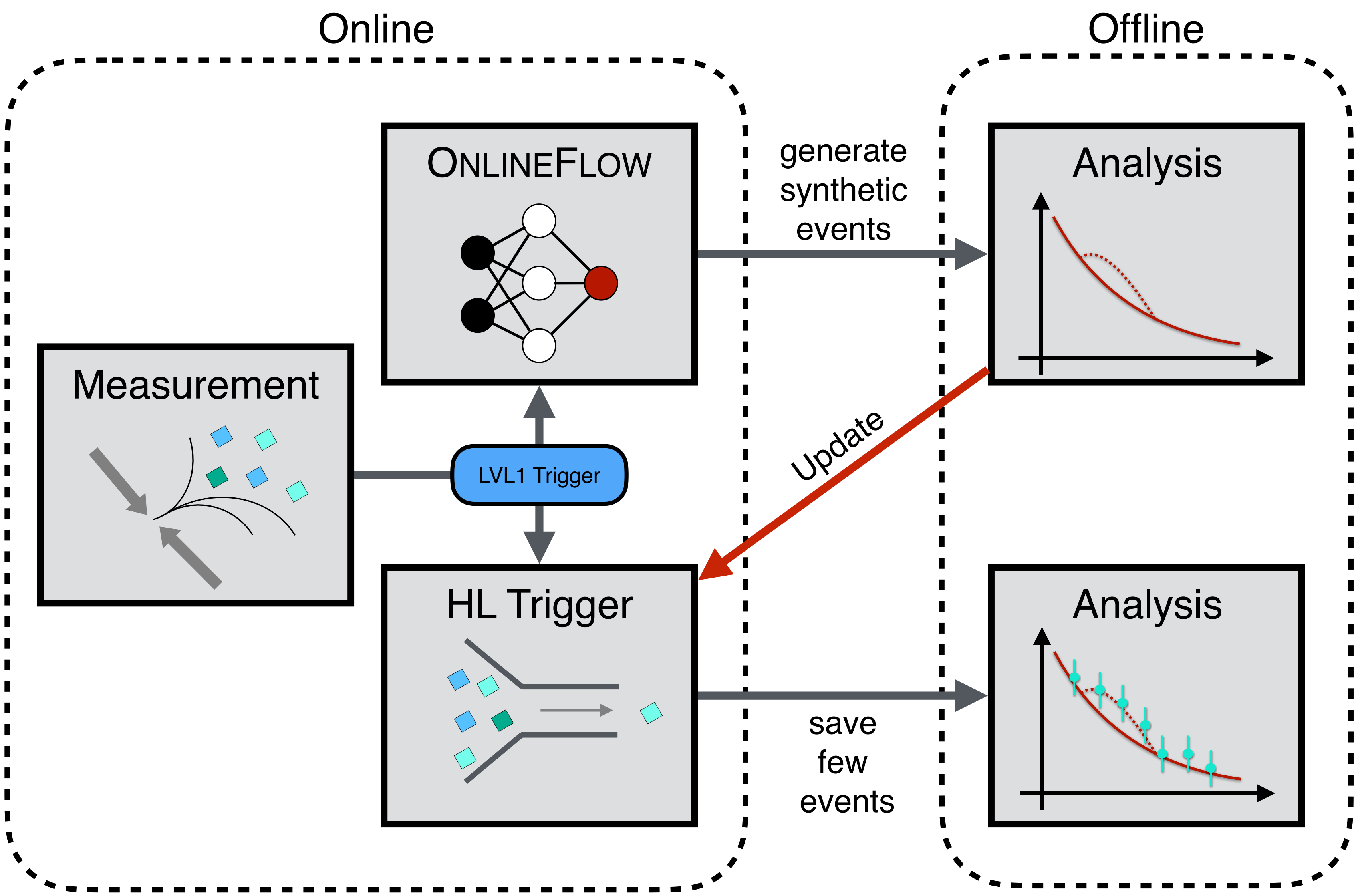


Many numbers per **event** | Small set of numbers per **event** | Small set of numbers per **dataset**

- Fixed size, independent of training data amount
- Radically different format from usual way of storing data, but might open up new approaches

Diefenbacher, .., GK et al 2202.0937

OnlineFlows



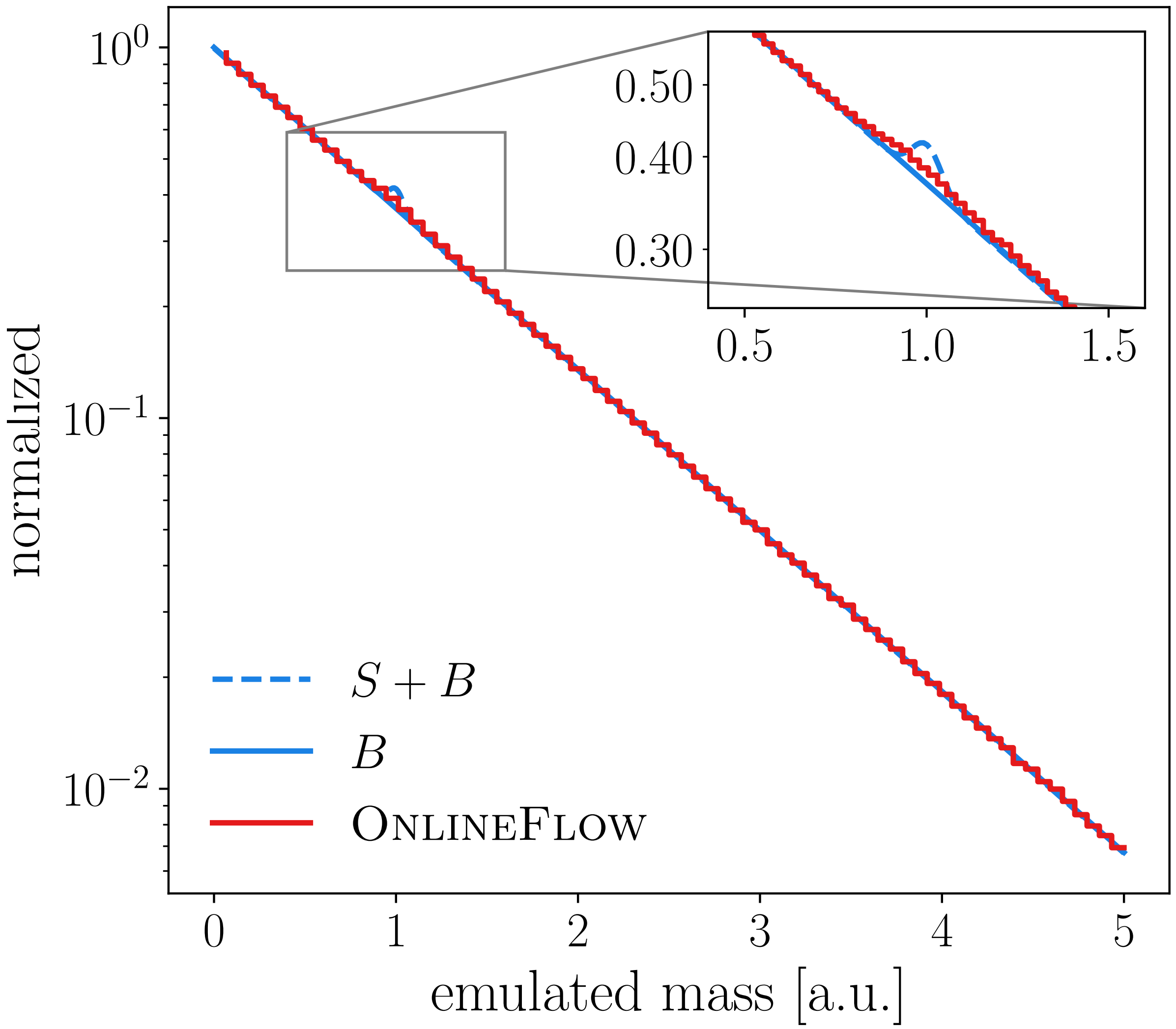
Schematic of proposed approach.

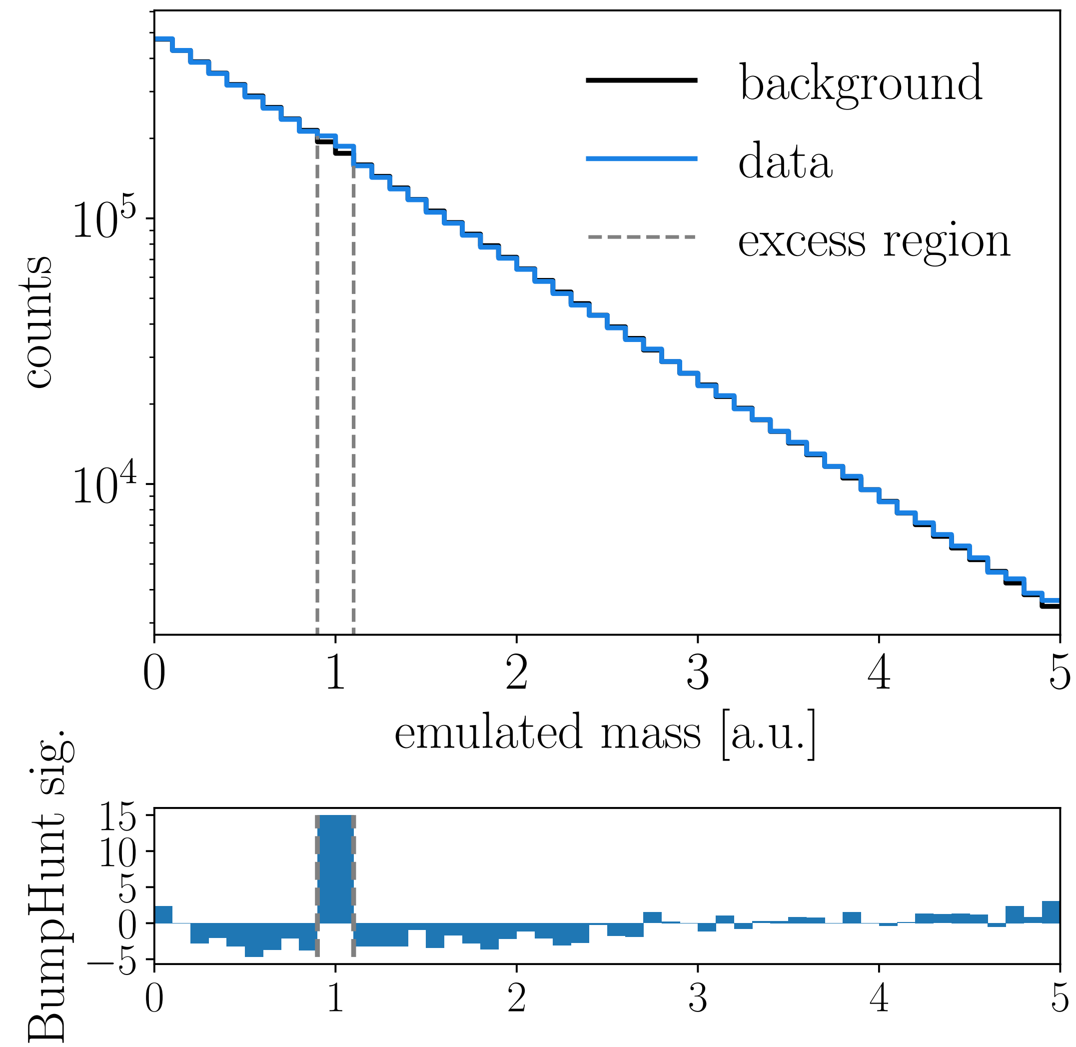
Focus on HLT, more technical challenges for use in hardware Trigger

Main problem:
How to make training work if each event is only available for short time?

Diefenbacher, .., GK et al 2202.0937

Classical Bump Hunt





$$\text{significance} = \frac{\mathcal{O} - B}{\sqrt{B}} \equiv \frac{S}{\sqrt{B}}$$

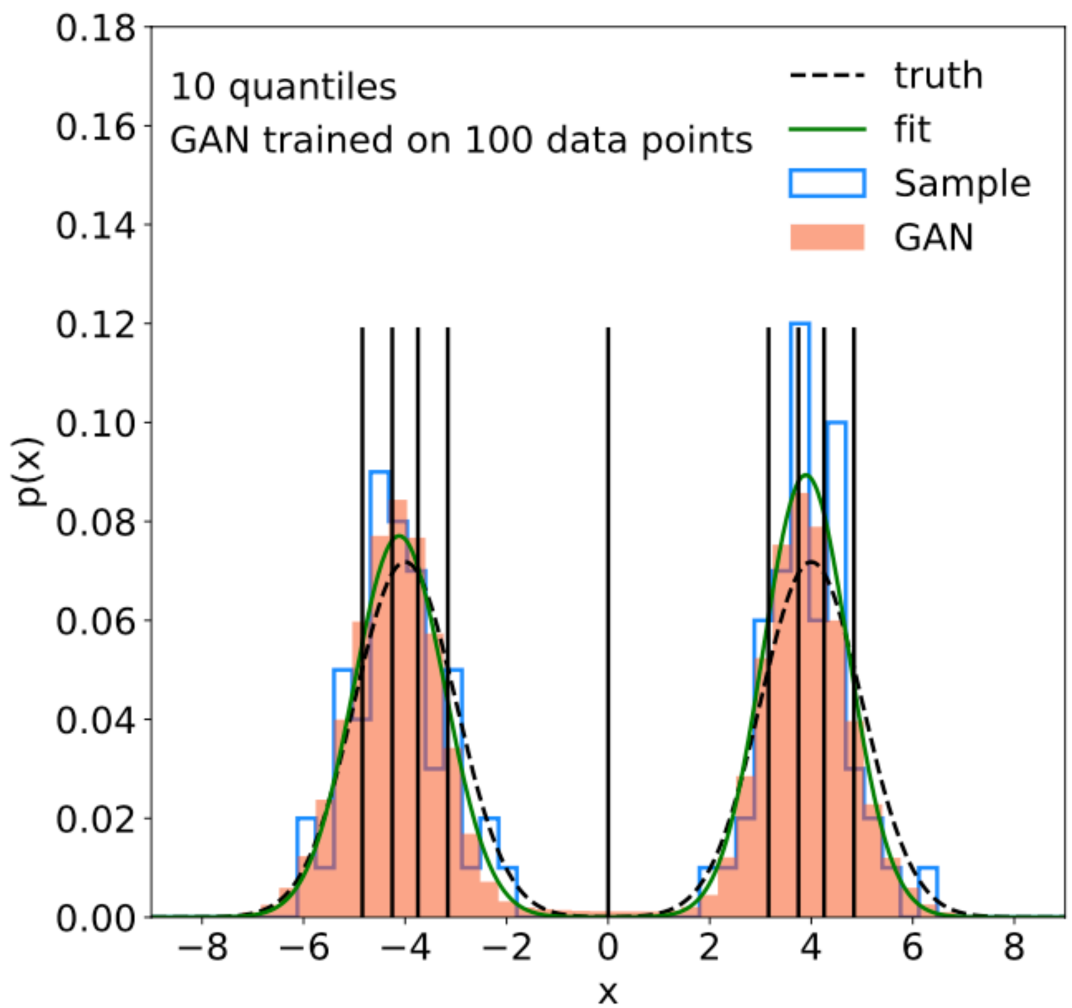
Diefenbacher, .., GK et al 2202.0937

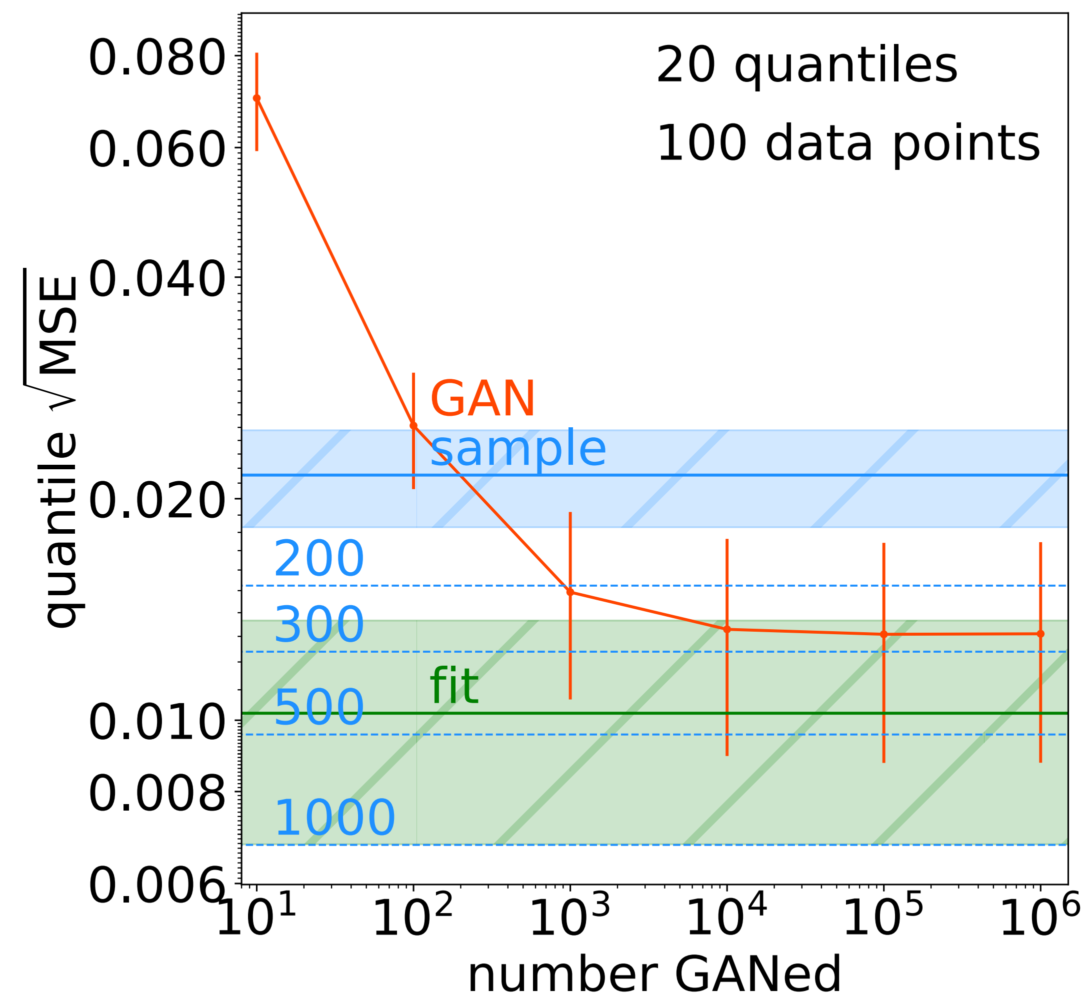
GANplification aside

Statistics

If we train a generator on N data points, and use it to produce M>>N examples, what is the statistical power of the M points?

Compare (known) truth distribution to sample and oversampled data from GAN

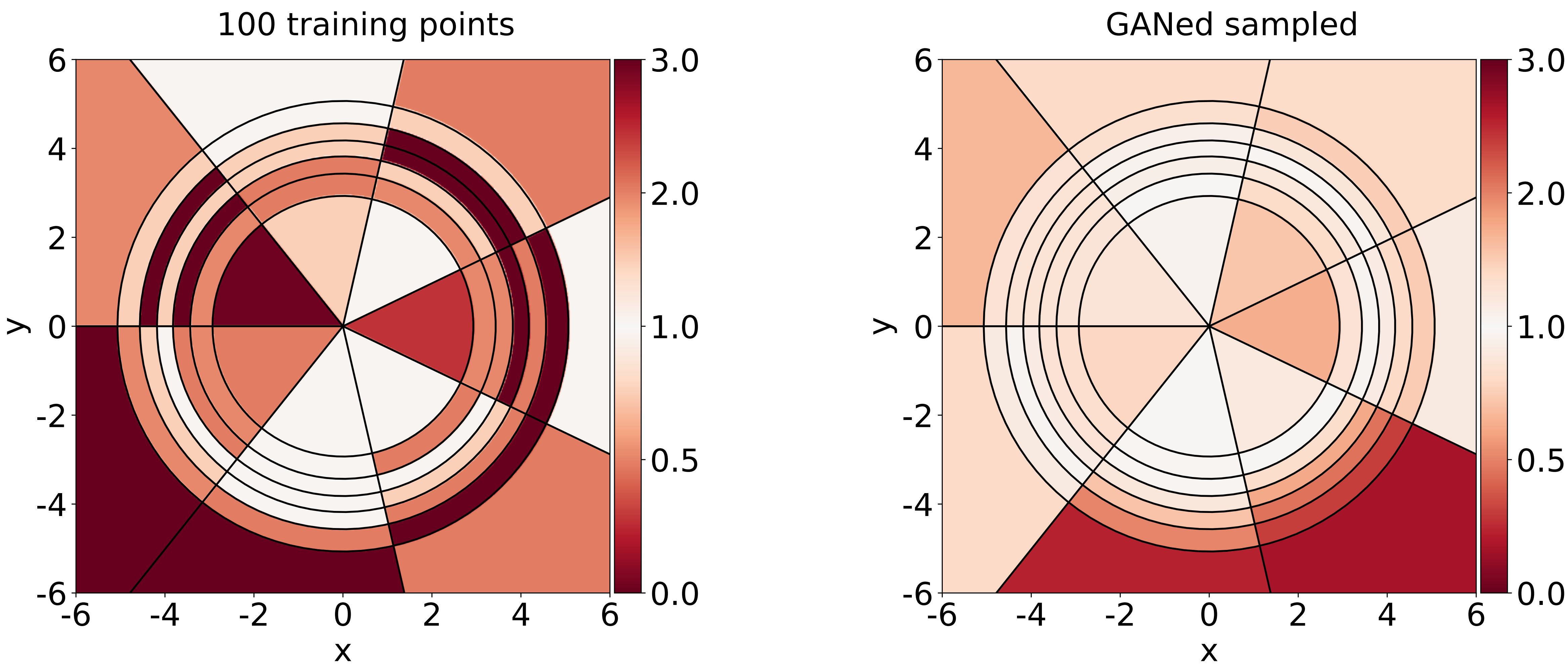




$$\mathrm{MSE} = \frac{1}{N_{\mathrm{quant}}} \sum_{j=1}^{N_{\mathrm{quant}}} \left(x_j - \frac{1}{N_{\mathrm{quant}}} \right)^2$$

Diefenbacher, .., GK et al 2008.06545

Statistics - 2D

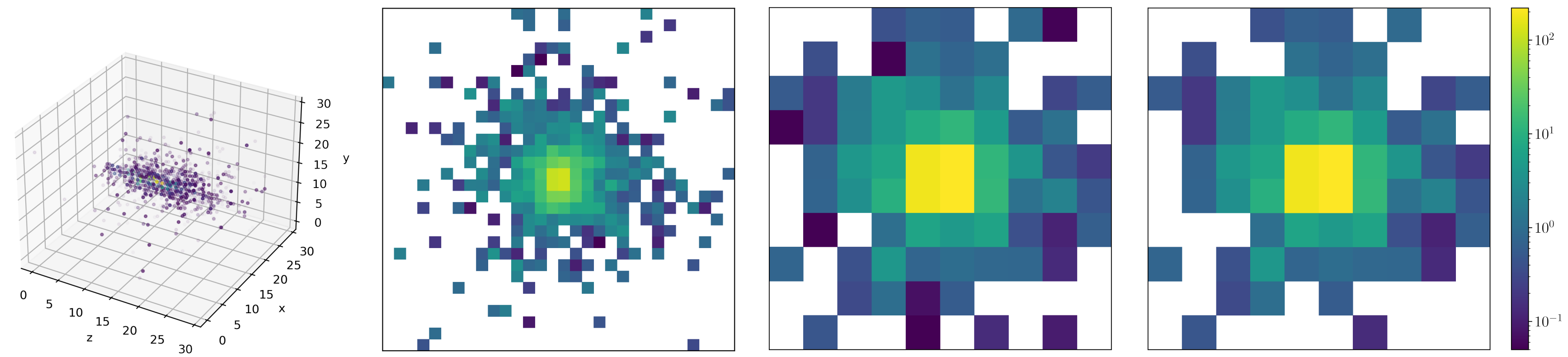


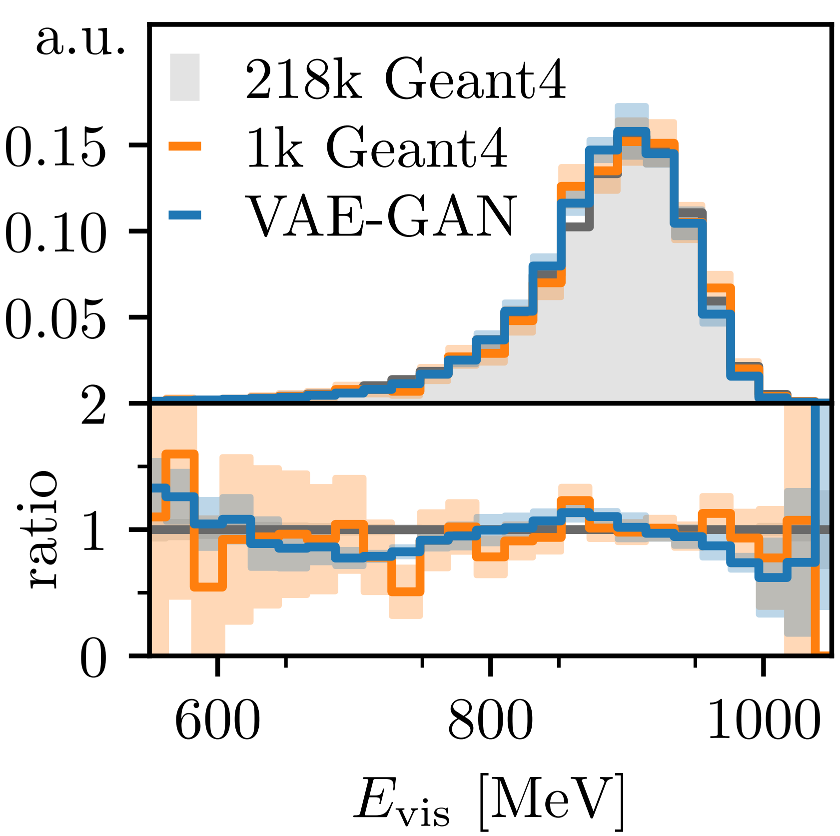
Relative deviation from Gaussian ring distribution

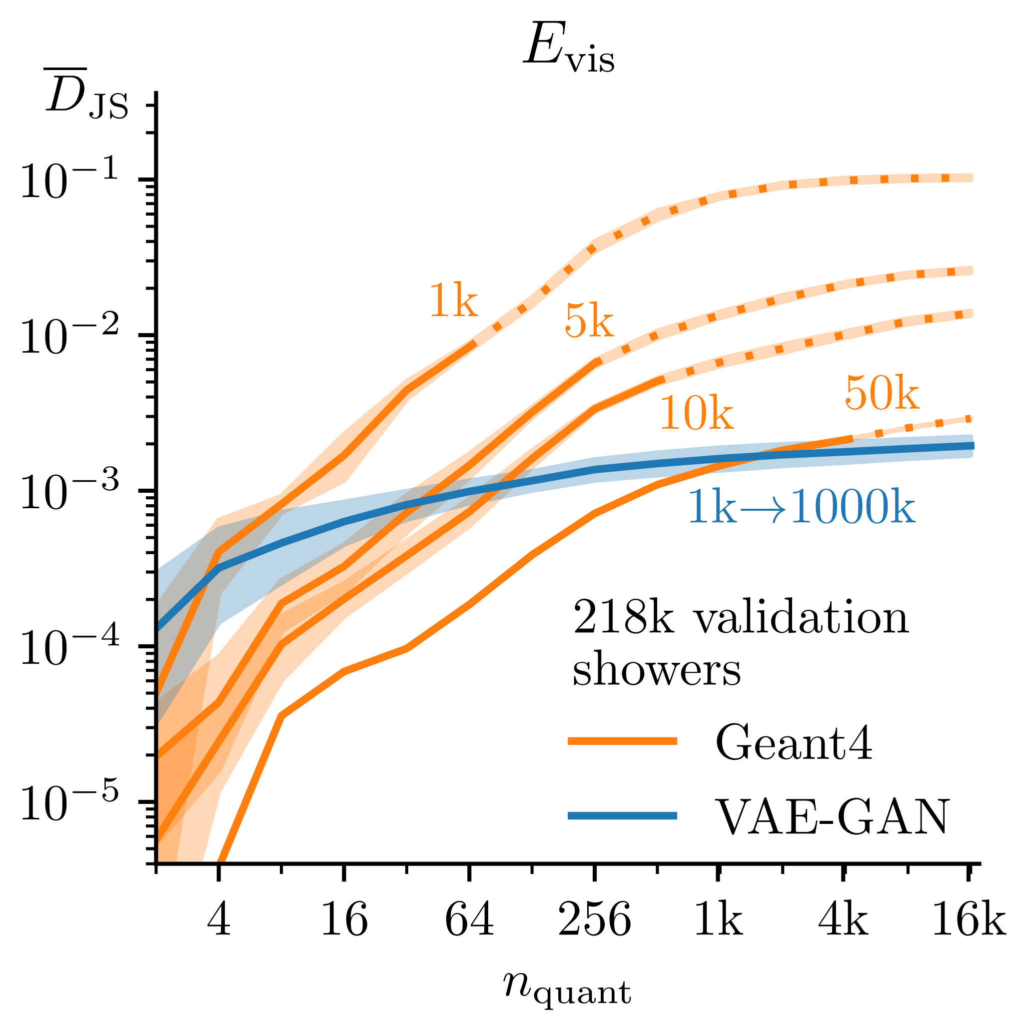
Diefenbacher, .., GK et al 2008.06545

Statistics - Physics

Test the statistical properties of simplified calorimeter showers.







Scaling of difference to ground truth with resolution again better for the generative model.

Bieringer, .., GK et al 2202.07352

Back to our problem

Generative Bump Hunt

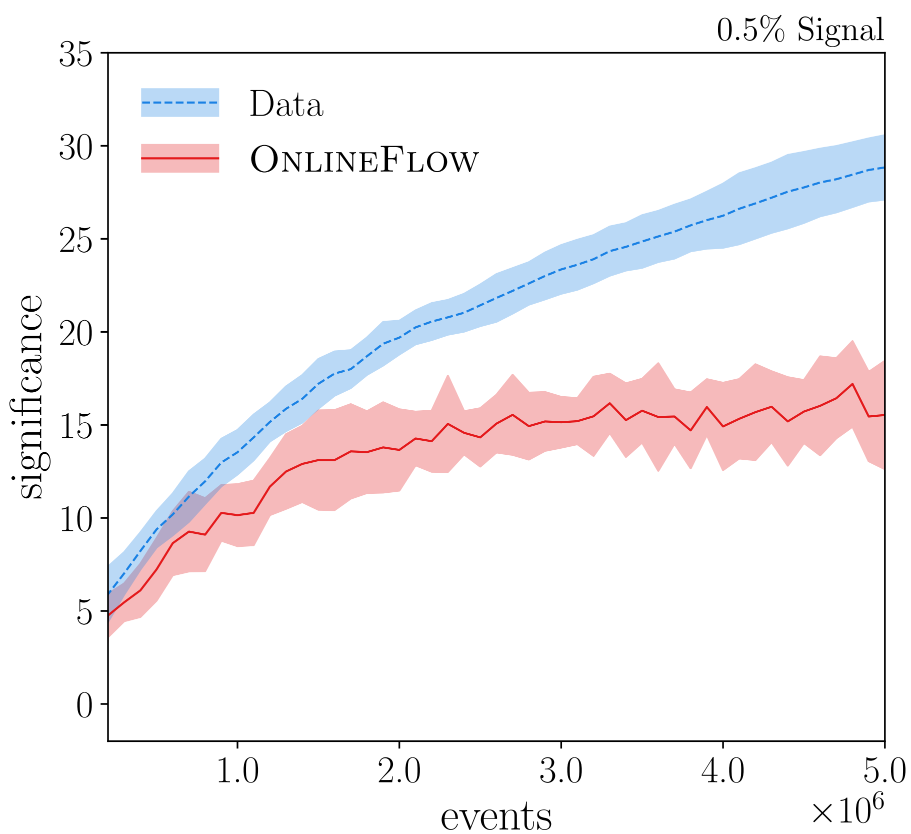
$$B = \frac{2}{N_{\text{ens}}} \sum_{i}^{N_{\text{ens}}/2} B_i \qquad \delta_B = \sqrt{\frac{2}{N_{\text{ens}}}}\,\sigma(B)$$

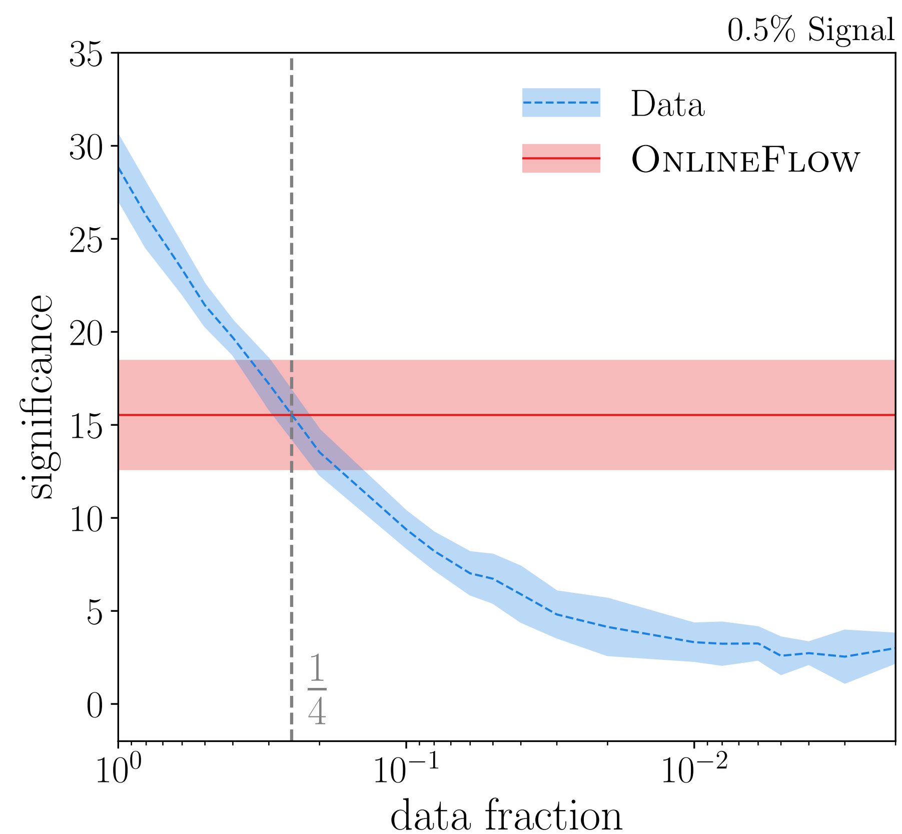
$$\mathcal{O} = \frac{2}{N_{\text{ens}}} \sum_{i}^{N_{\text{ens}}/2} \mathcal{O}_i \qquad \delta_{\mathcal{O}} = \sqrt{\frac{2}{N_{\text{ens}}}}\,\sigma(\mathcal{O}),$$

$$\text{significance} = \frac{S}{\sqrt{\delta_S^2 + (\sqrt{B})^2}}$$

$$S = \mathcal{O} - B \qquad \delta_S^2 = \delta_{\mathcal{O}}^2 + \delta_B^2$$

Include ensemble uncertainty of background predictions



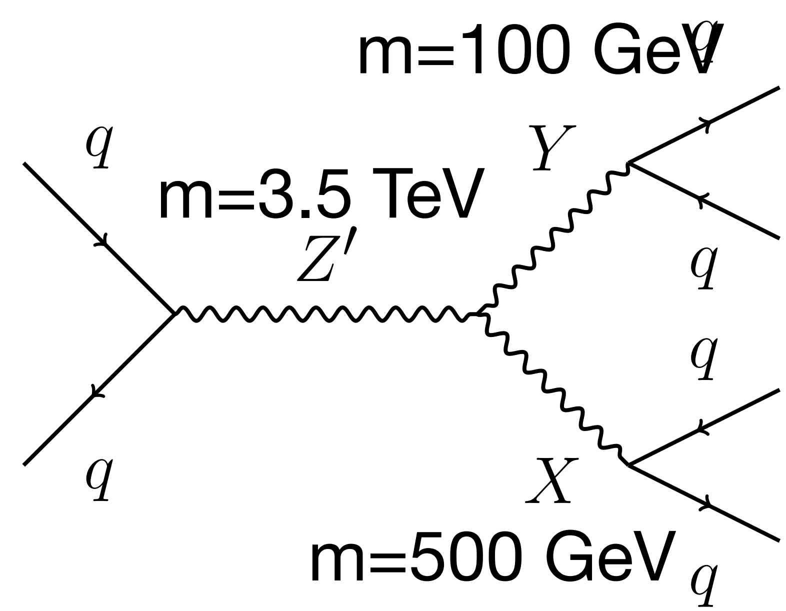


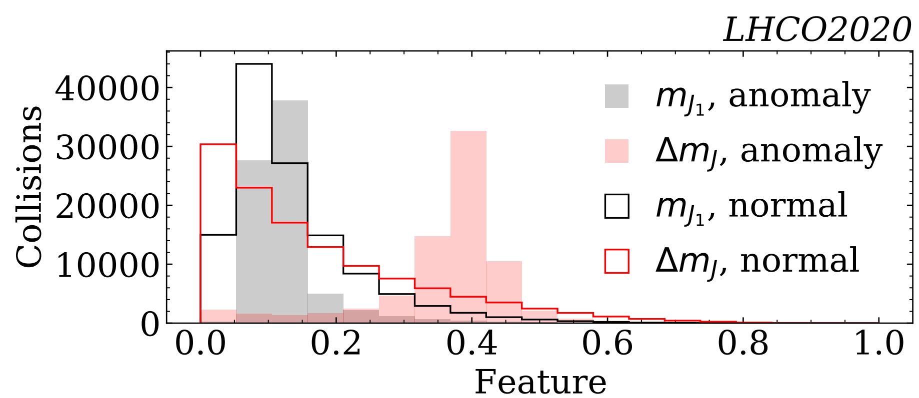
Diefenbacher, .., GK et al 2202.0937

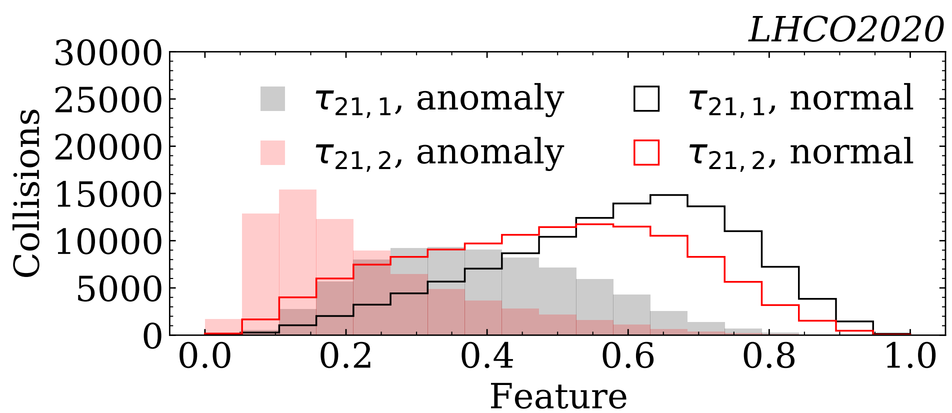
LHC Olympics Dataset

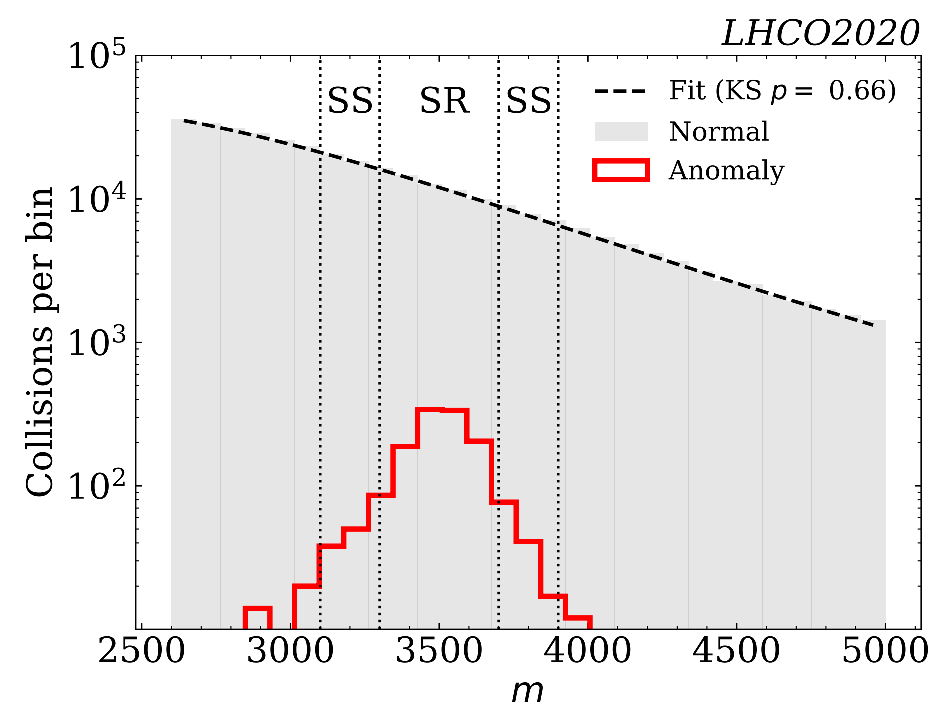
- For building and testing methods
- 1M background examples (Standard Model), 100k signal examples (signal, see Feynman diagram on the right)
- Labels provided

- Relatively simple signal
    - Known to differ in previously mentioned features from background distribution
- Unrealistically high S/B





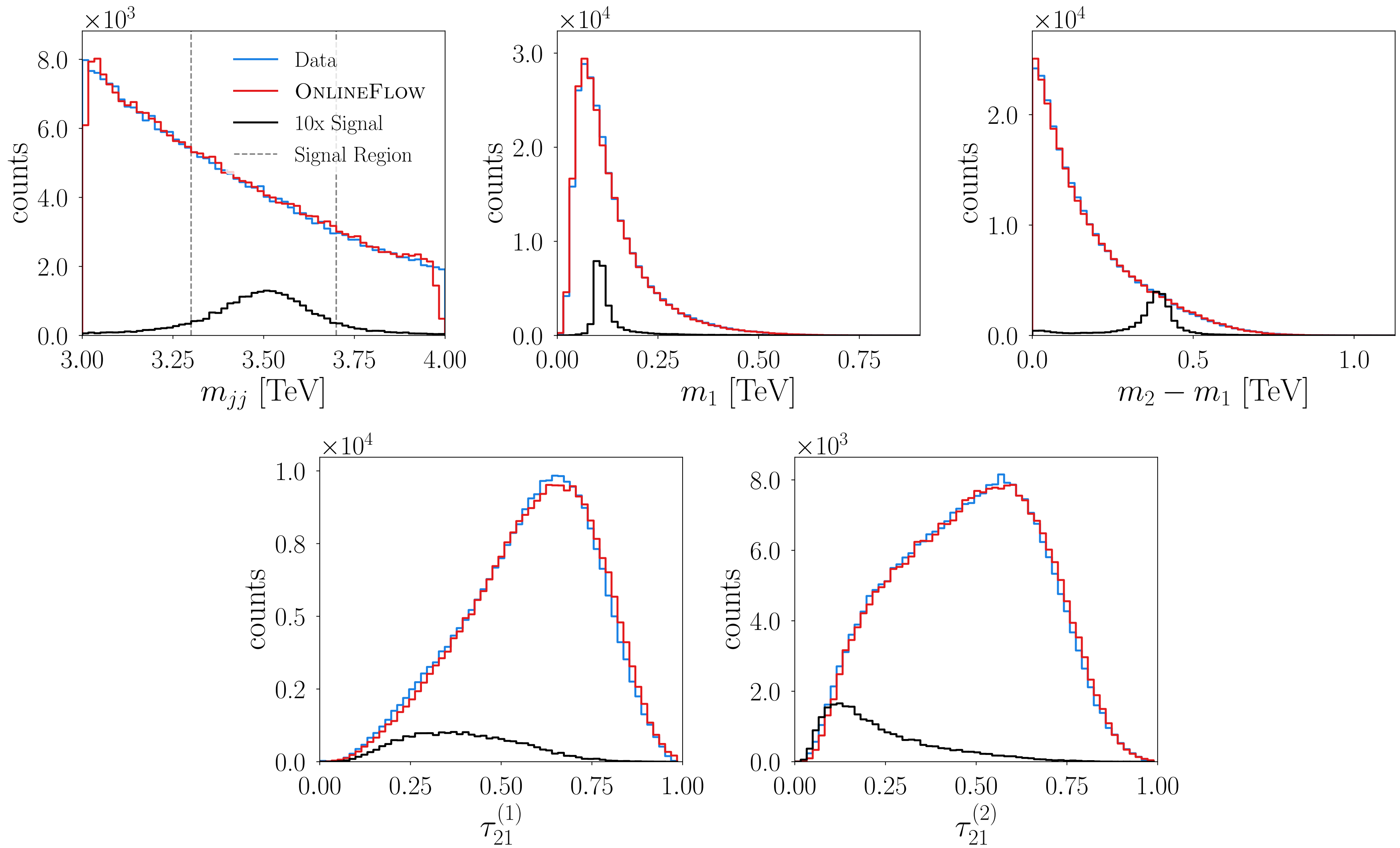




GK, Nachman, Shih, et al, 2107.02821

More realistic example

Use LHCO dataset, train on high-level features on a mixture of background (99%) and signal (1%).



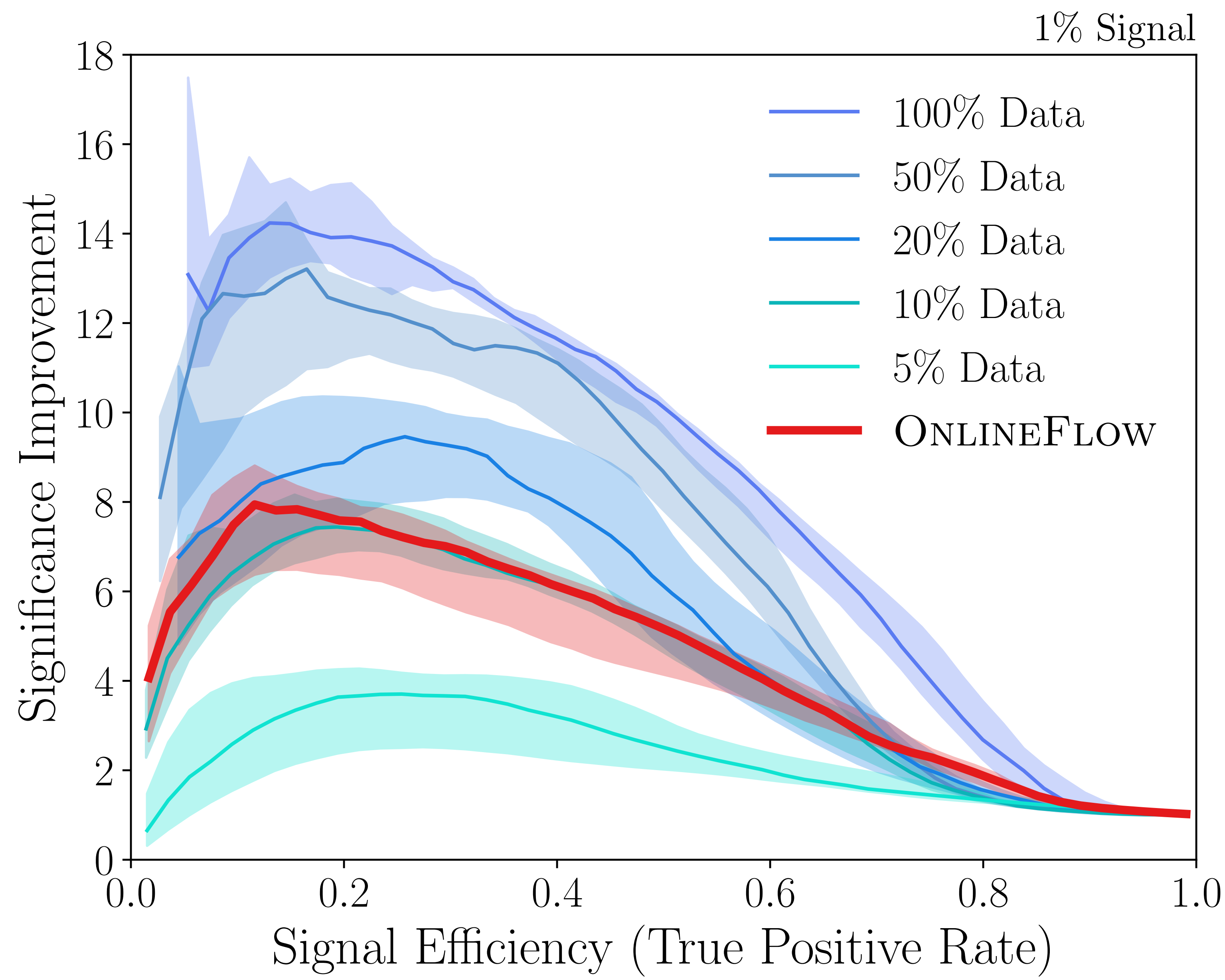
Diefenbacher, .., GK et al 2202.0937

More realistic example

Use LHCO dataset, train on high-level features on a mixture of background (99%) and signal (1%).

Train classifier to distinguish a signal region and sideband (CWoLA appaorach)

Compare procedure directly carried out on data with output of flow.

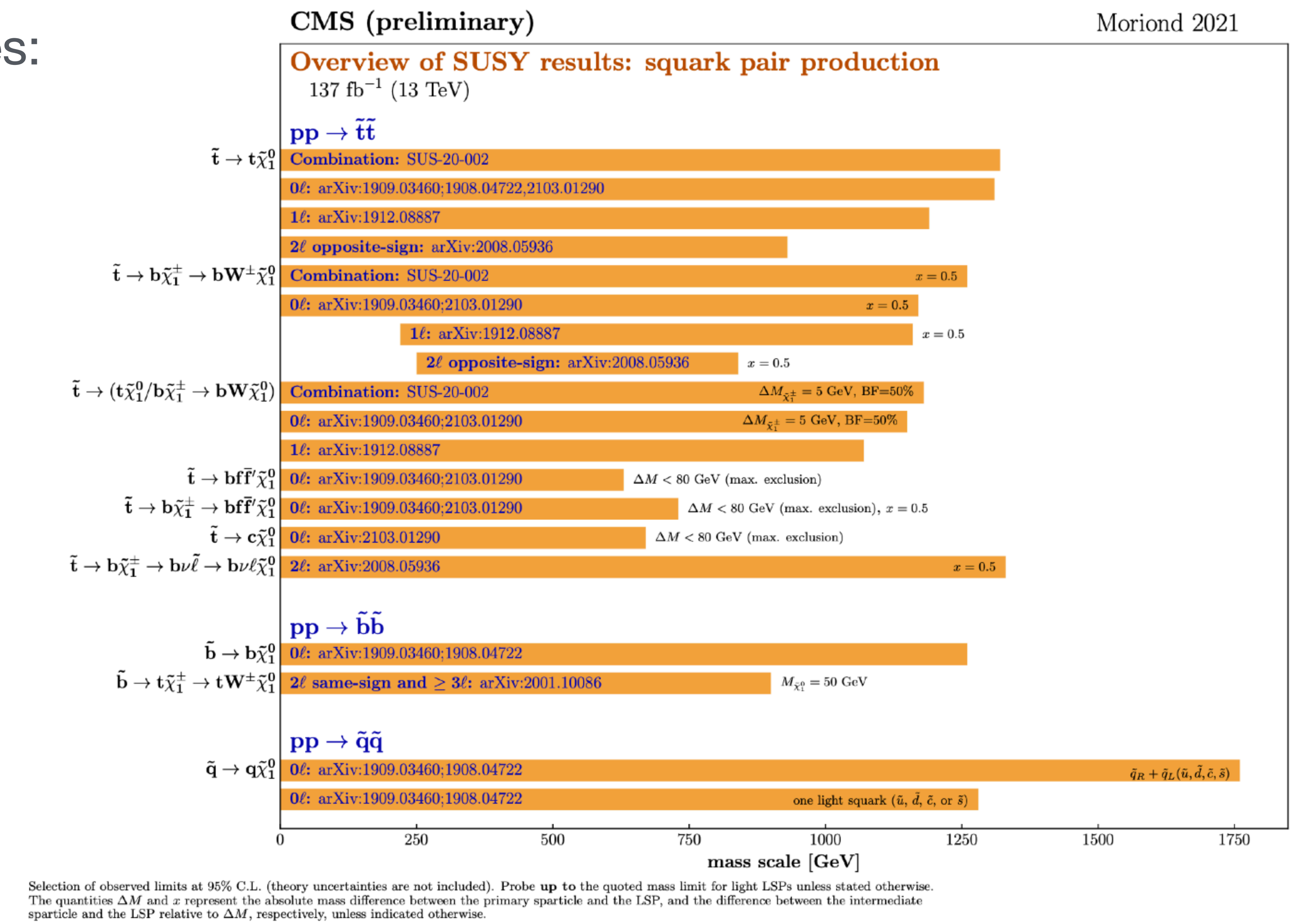


Diefenbacher, .., GK et al 2202.0937

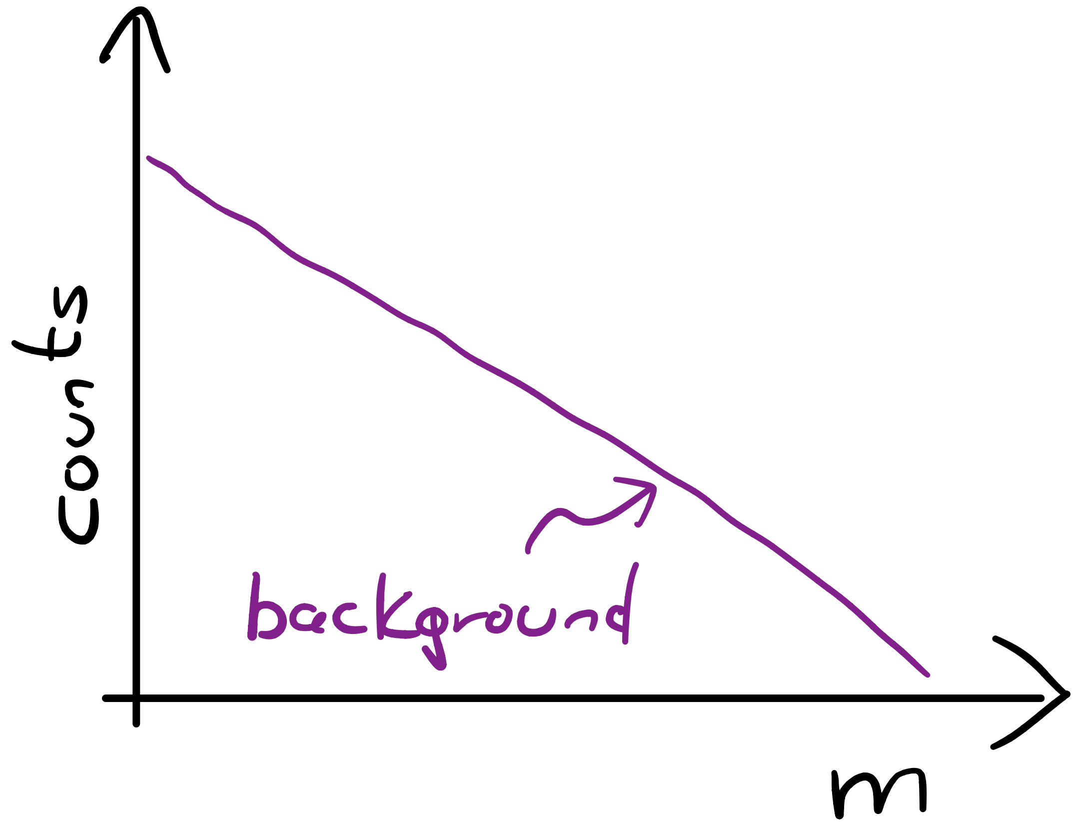
Anomaly Detection

Motivation

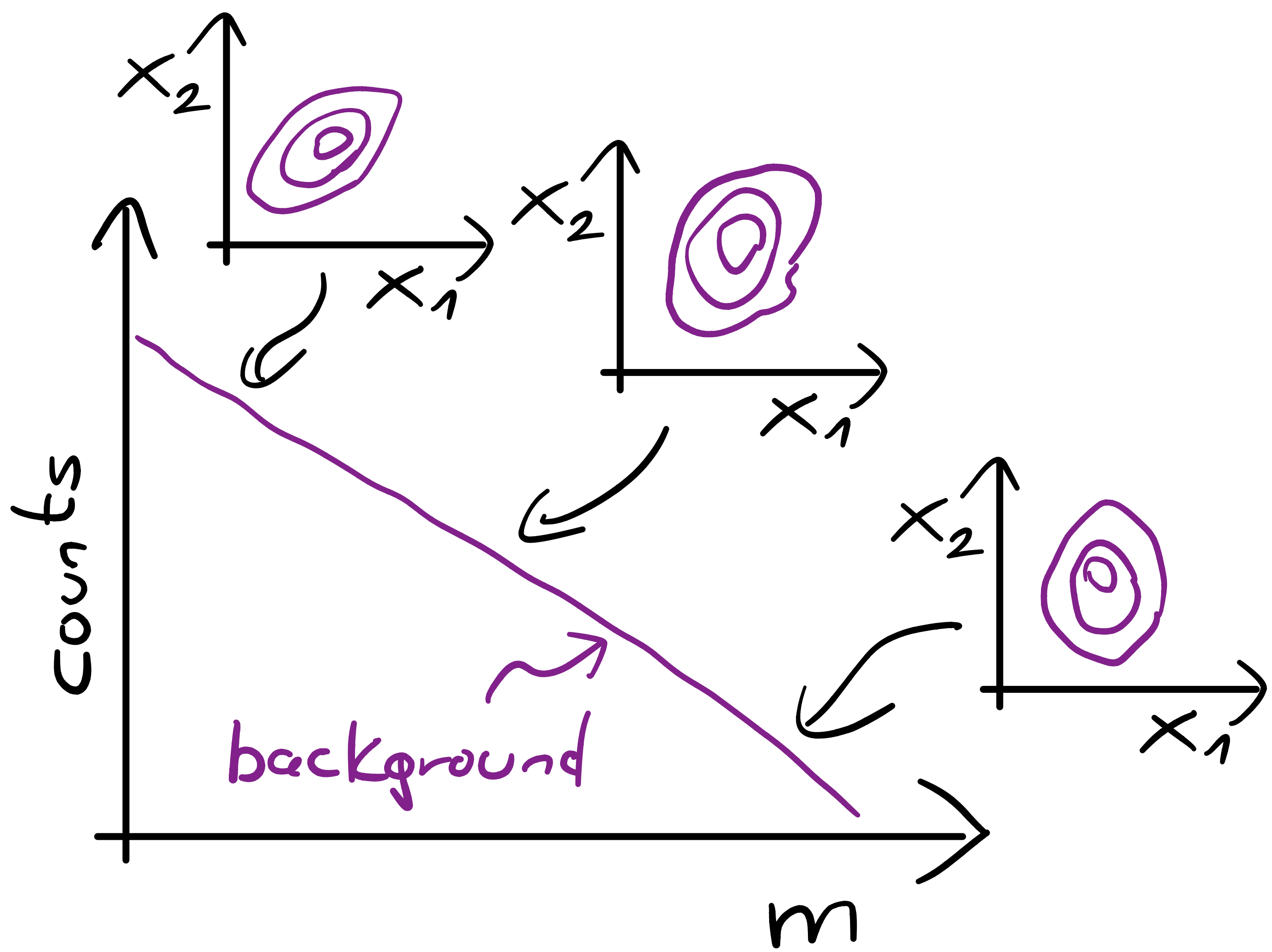
- Expect physics beyond the Standard Model
- Only negative results in searches
- Two discovery strategies:
  - Model-specific
  - Model independent
- Trade off:
  Sensitivity to
  specific model vs
  broad coverage



Resonant Anomaly Detection

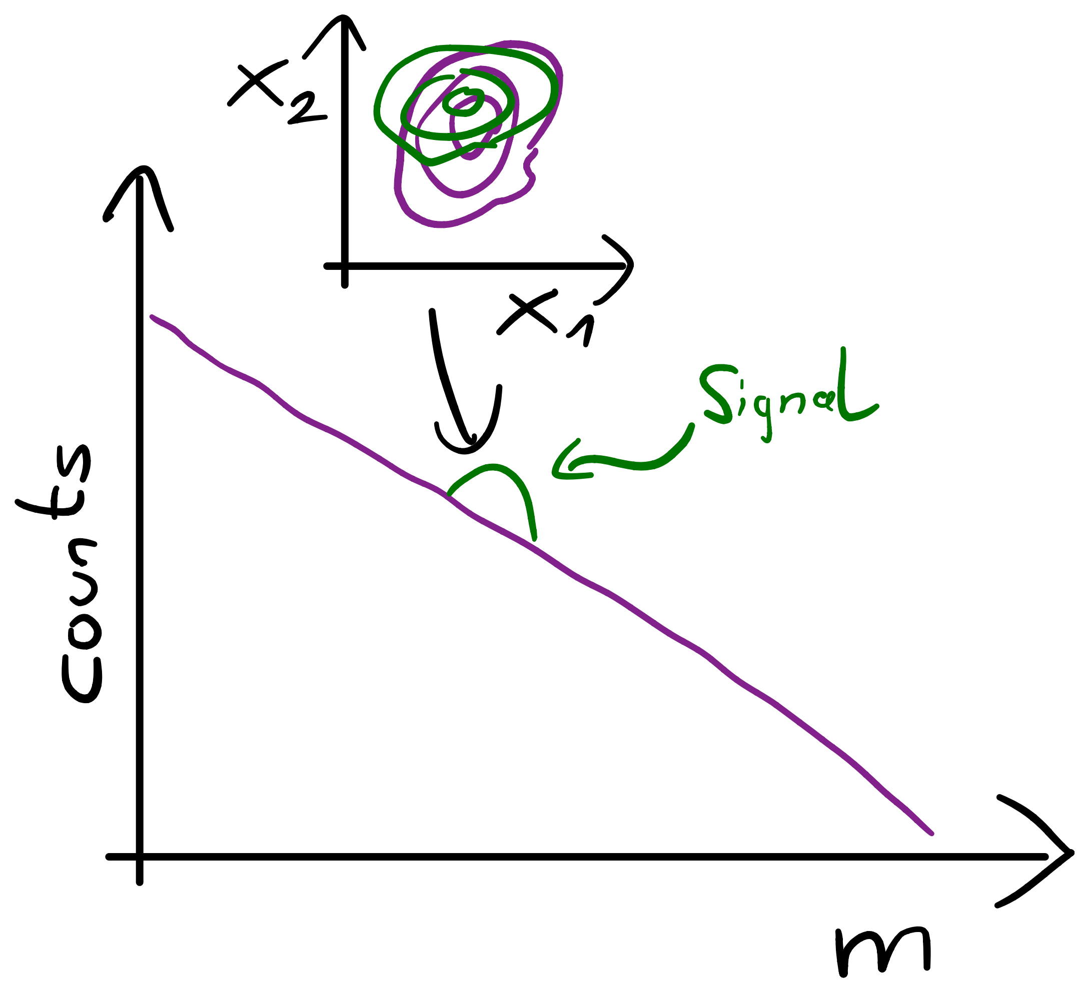


Resonant Anomaly Detection



+ additional dimensions
(in general correlated with m)

Resonant Anomaly Detection



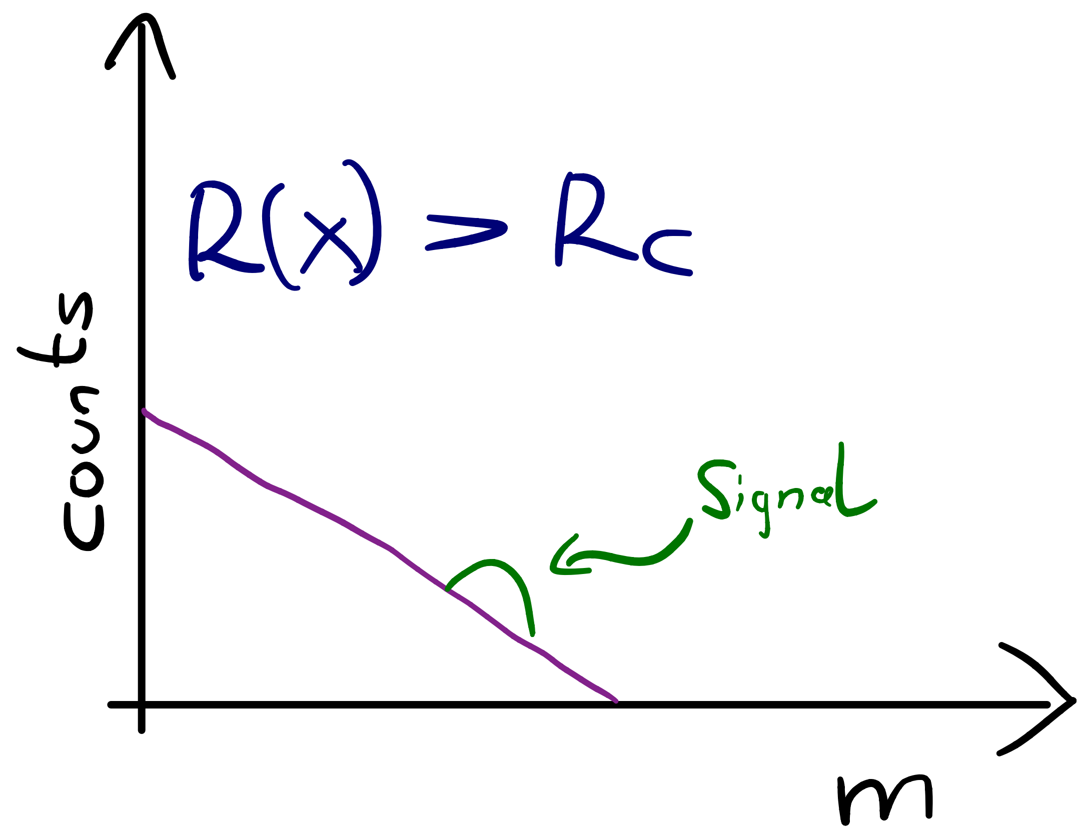
Look for a small signal, localised in m, and different shape in other features

Need to find a feature in which signal is resonant and background smooth.

No assumptions in other features.

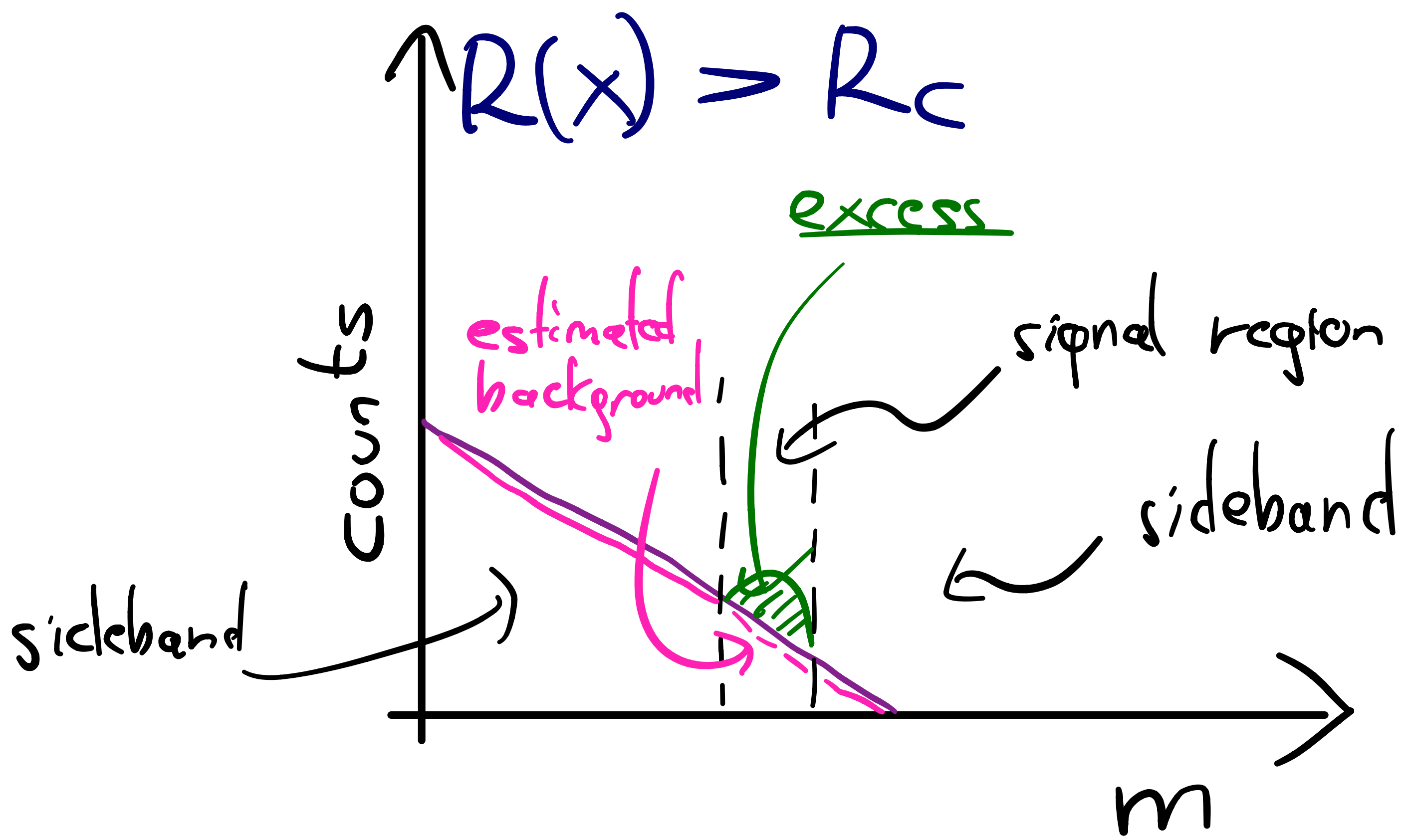
Further generalisation as open issue.

Resonant Anomaly Detection



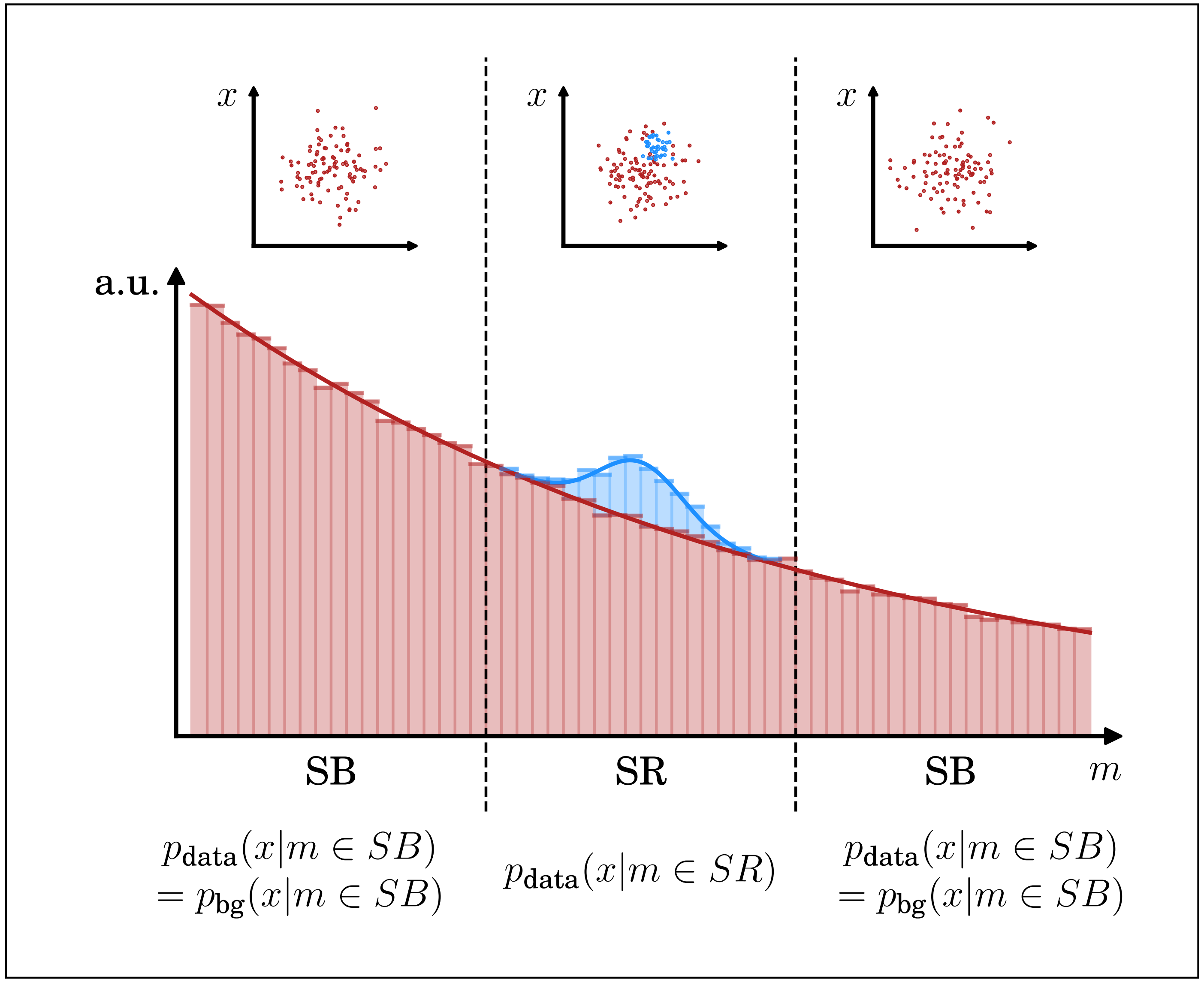
Enhanced bump-hunt: Use ML to build classifier $R(x)$ so that selecting $R(x) > c$ enhances signal fraction

Resonant Anomaly Detection



Enhanced bump-hunt: Then fit background from sidebands, compare to data in signal region

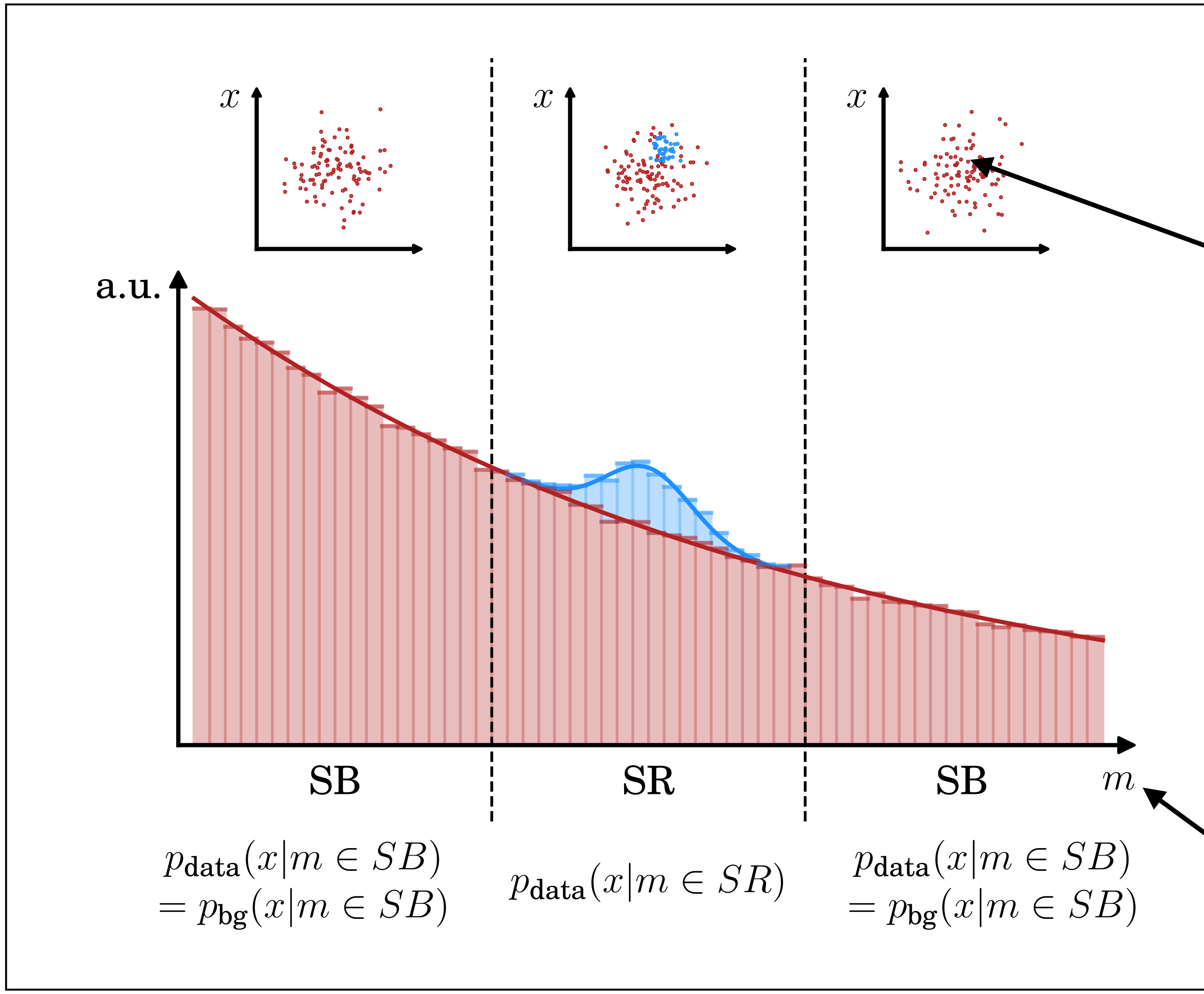
CATHODE



Consider resonant anomalies: slightly reduces generality, but allows fully data-based construction of anomaly detection score

GK, Nachmann, Shih et al 2101.08320;
Hallin, .., GK et al 2109.00546;

CATHODE

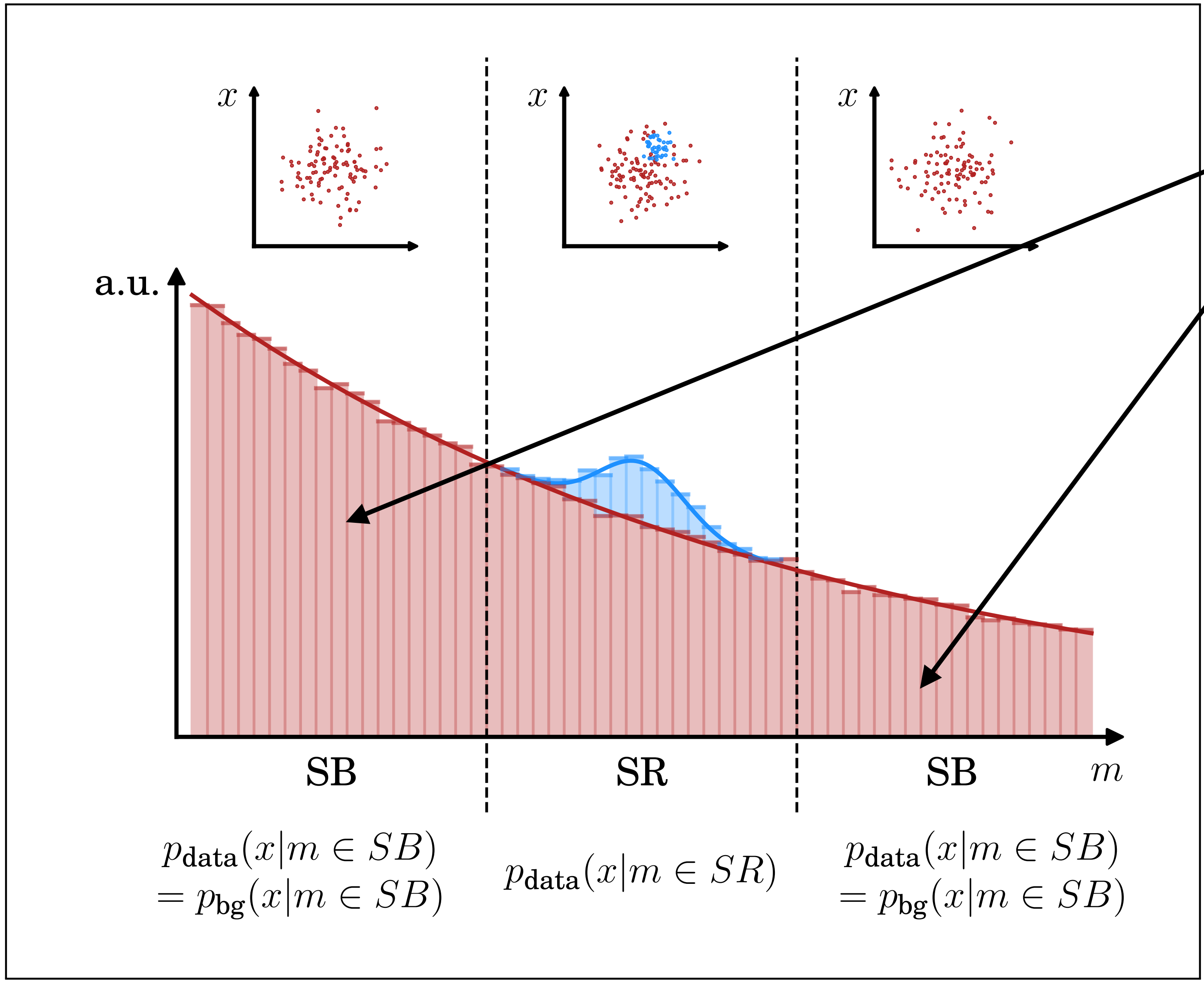


4 or (many) more additional features

one resonant feature

GK, Nachmann, Shih et al 2101.08320;
Hallin, .., GK et al 2109.00546;

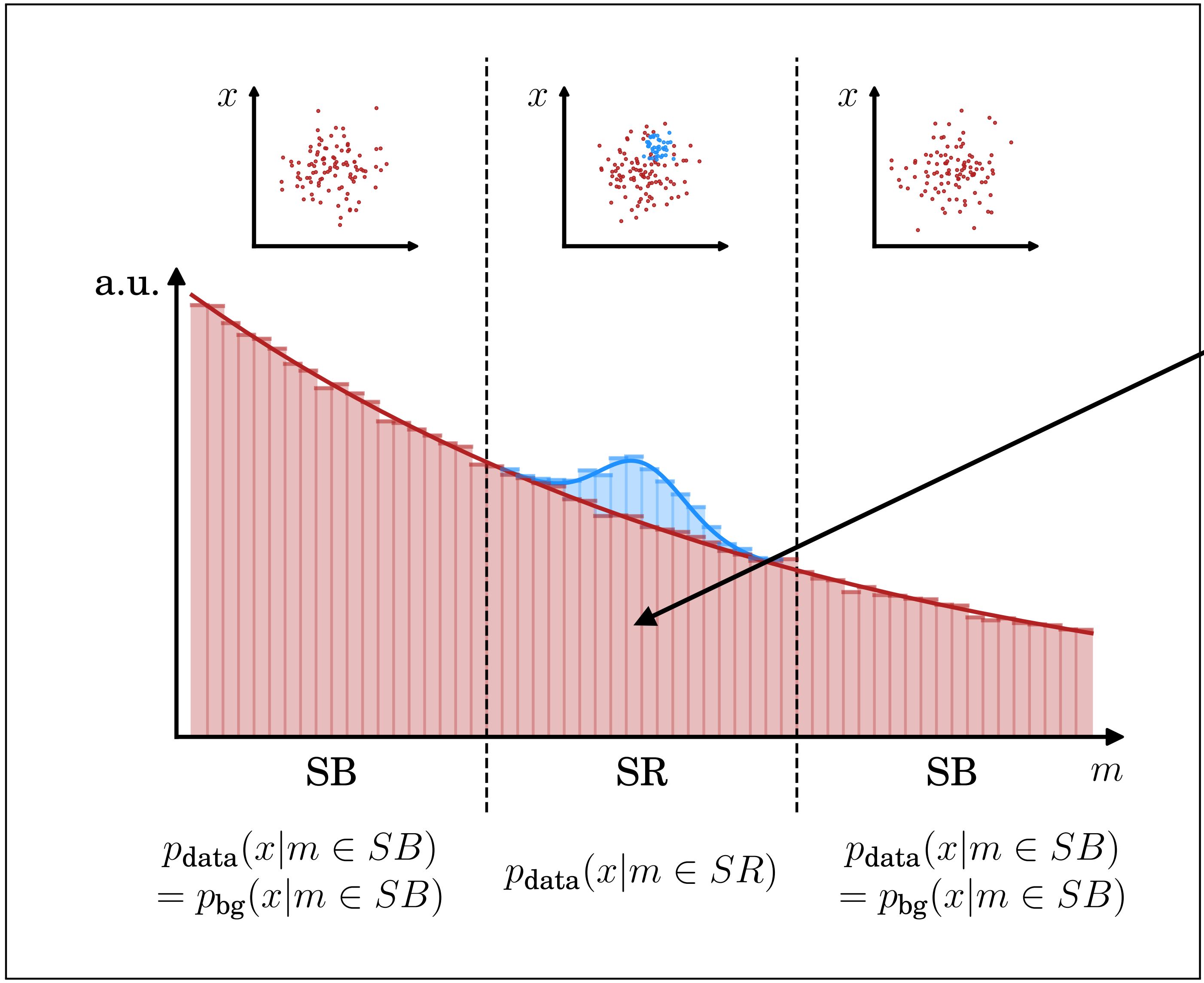
CATHODE



Train generative model (conditional normalising flow)

GK, Nachmann, Shih et al 2101.08320;
Hallin, .., GK et al 2109.00546;

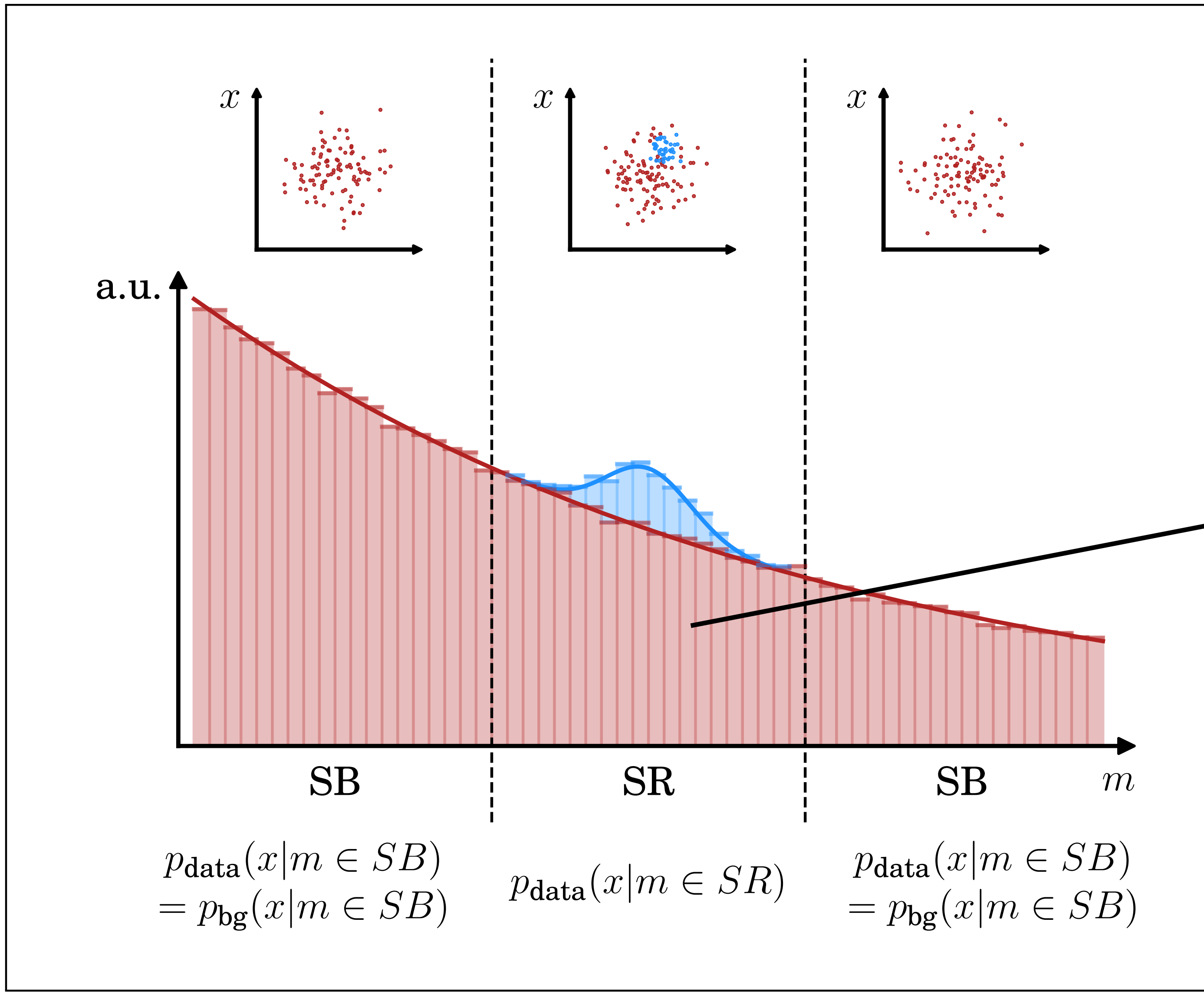
CATHODE



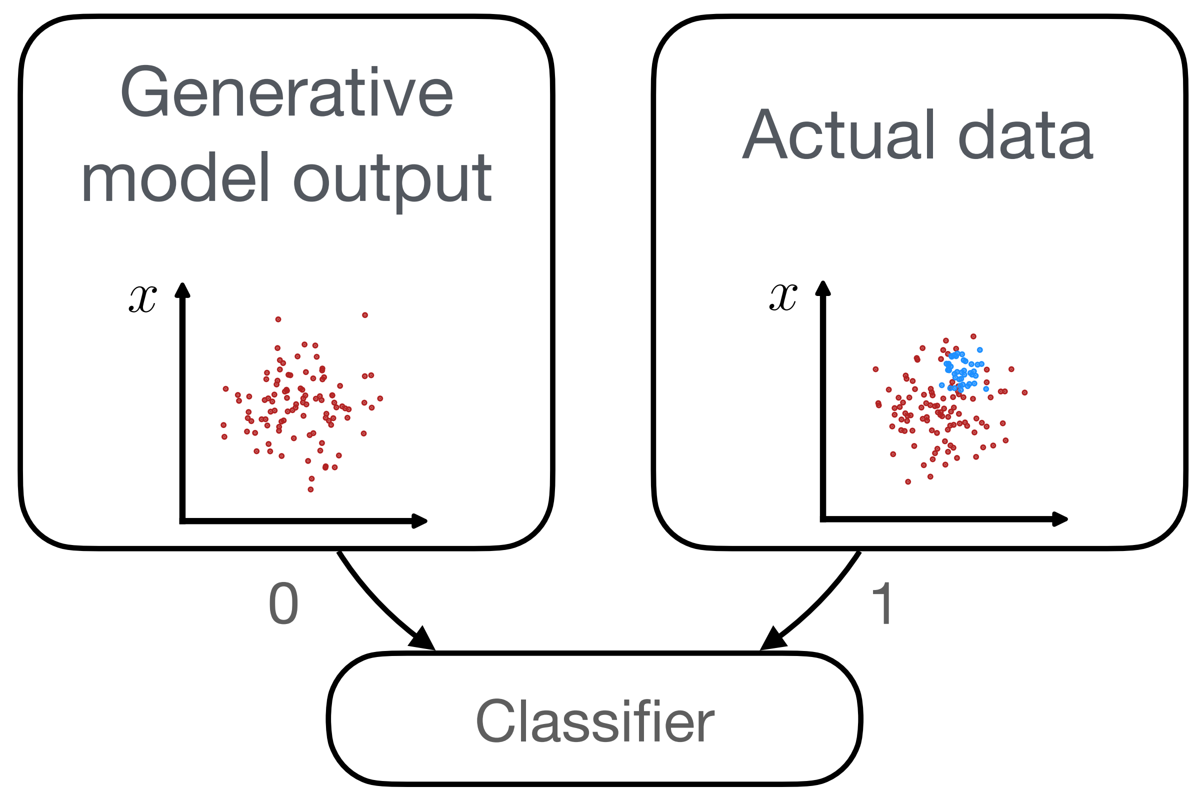
Interpolate &
and sample here

GK, Nachmann, Shih et al 2101.08320;
Hallin, .., GK et al 2109.00546;

CATHODE

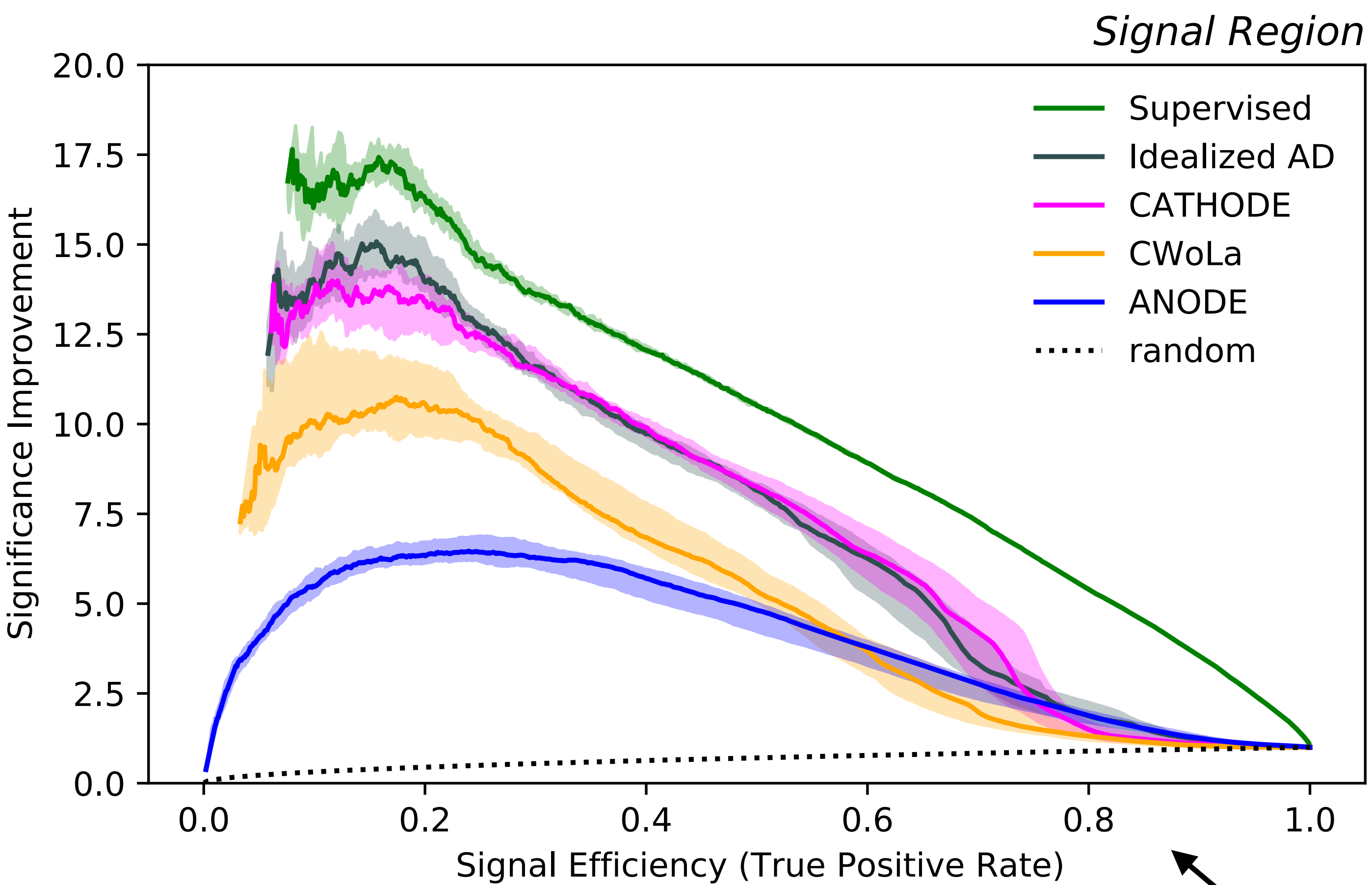


Train a classifier between
prediction vs data

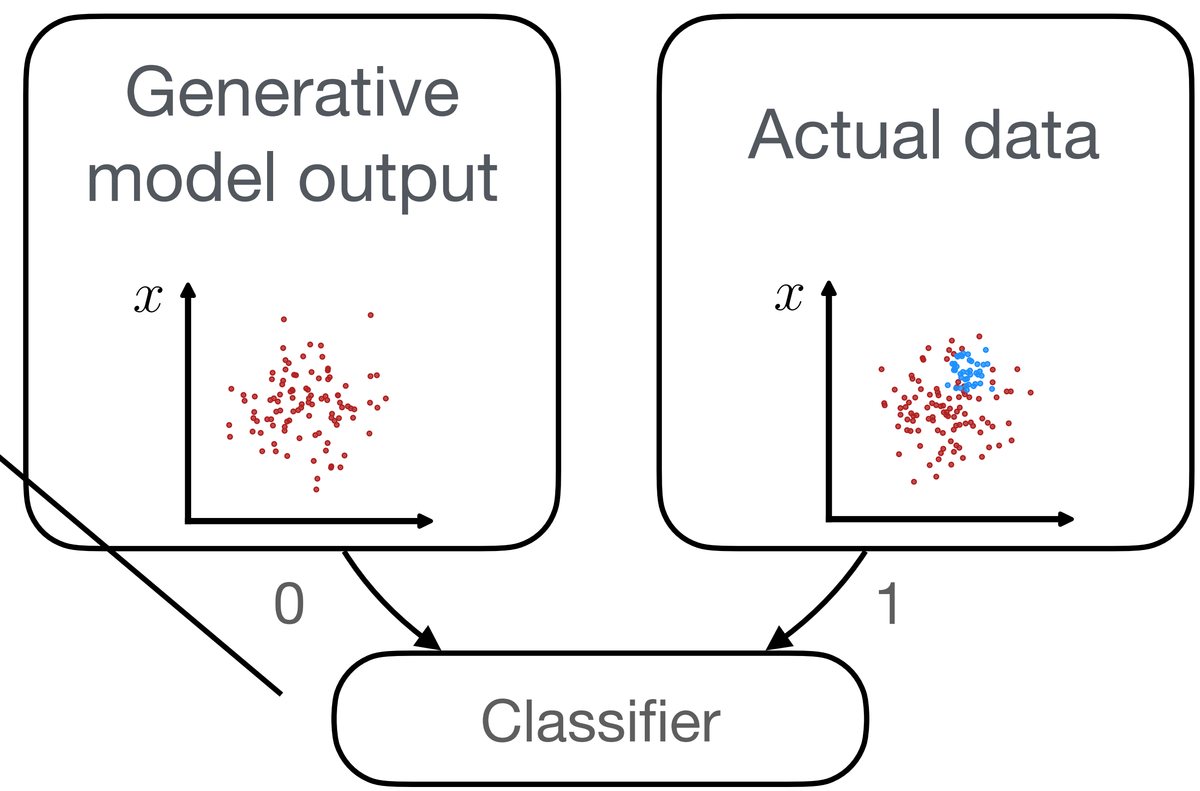


GK, Nachmann, Shih et al 2101.08320;
Hallin, .., GK et al 2109.00546;

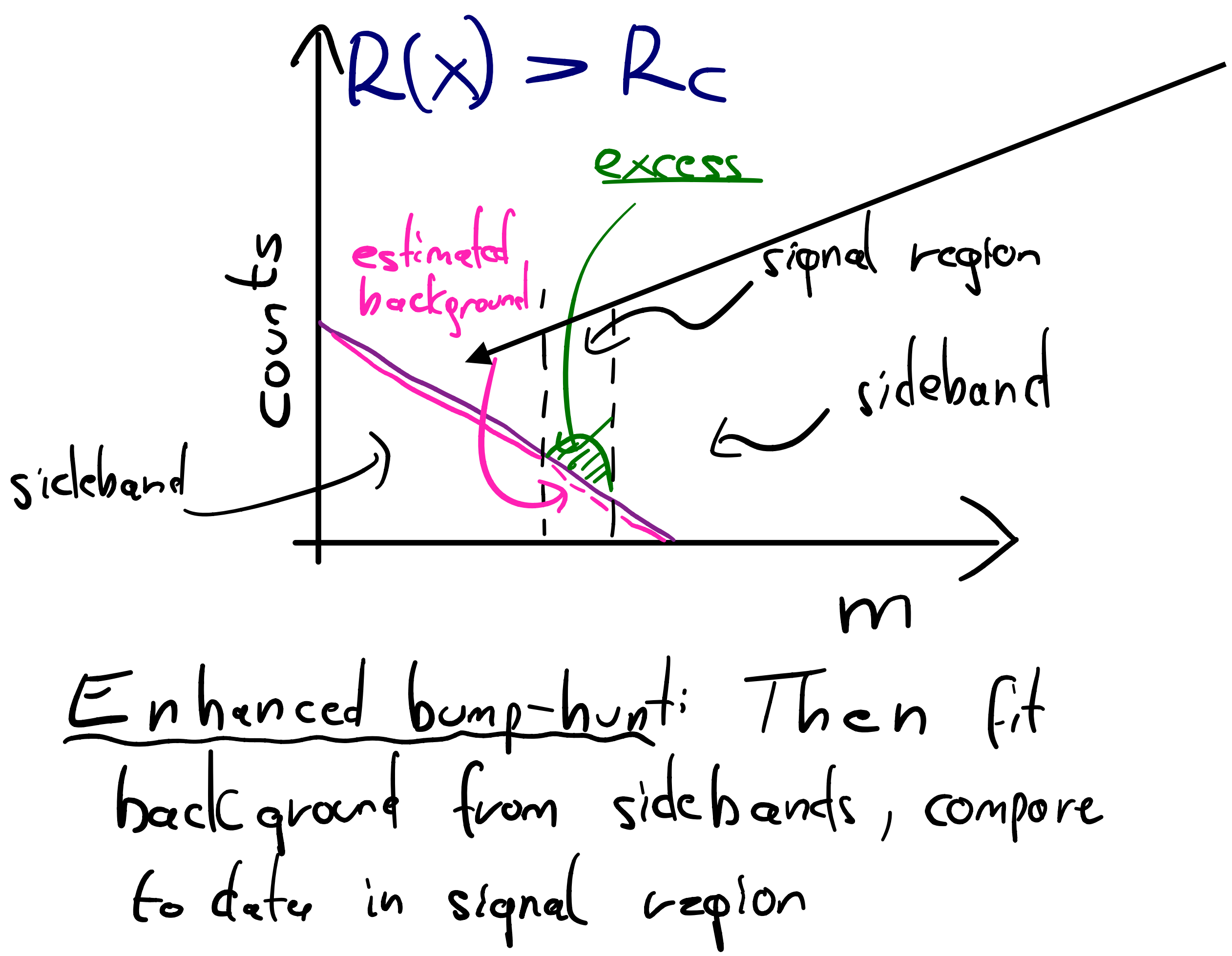
CATHODE



Use classifier to identify anomalies

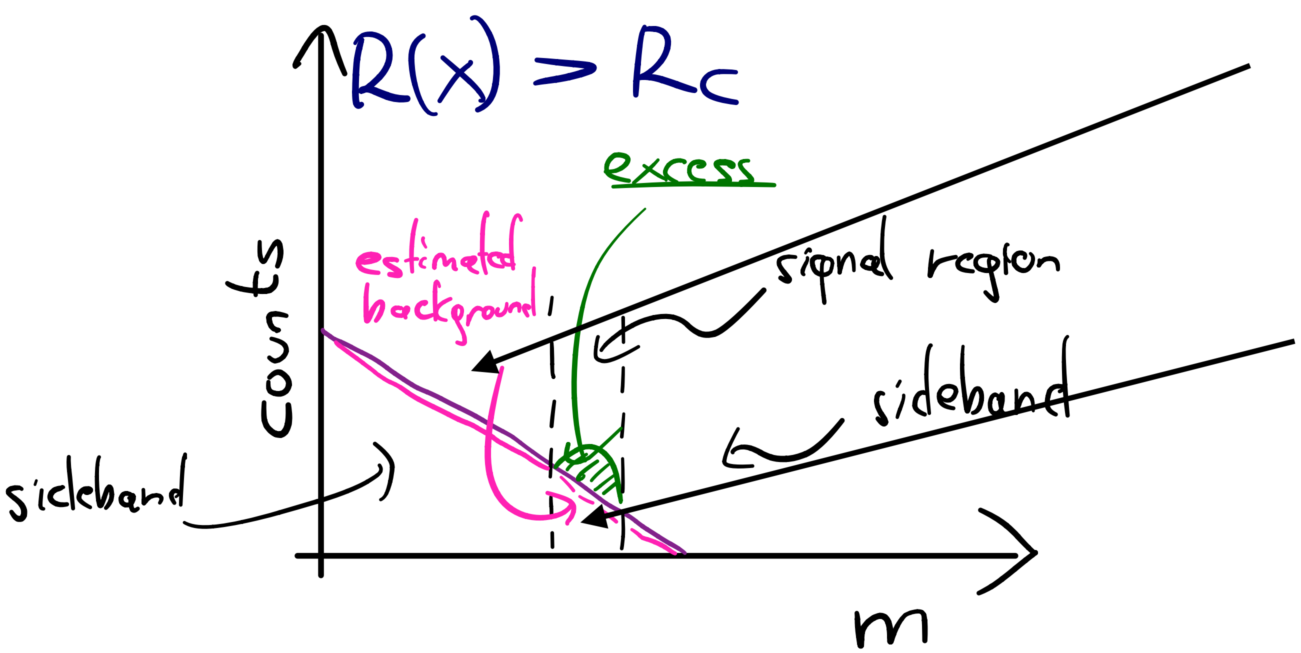


What are the crucial uncertainties?



Uncertainty from the 1d fit (including parameter choice)

What are the crucial uncertainties?

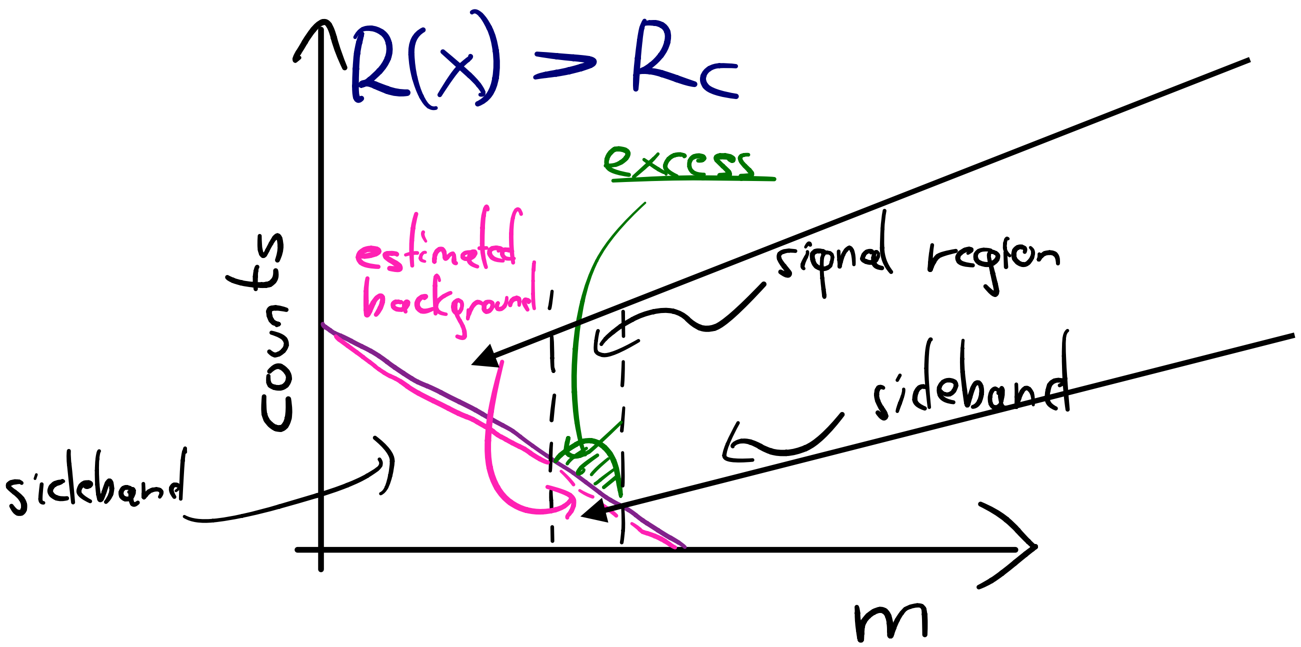


Uncertainty from the 1d fit (including parameter choice)

Statistical uncertainty in signal region

Enhanced bump-hunt: Then fit background from sidebands, compare to data in signal region

What are the crucial uncertainties?



Enhanced bump-hunt: Then fit background from sidebands, compare to data in signal region

Uncertainty from the 1d fit (including parameter choice)

Statistical uncertainty in signal region

Non-closure of generative model!

Generative non-closure

Inclusive is sideband; other distributions after classifier cut

True sideband/ Generated sideband

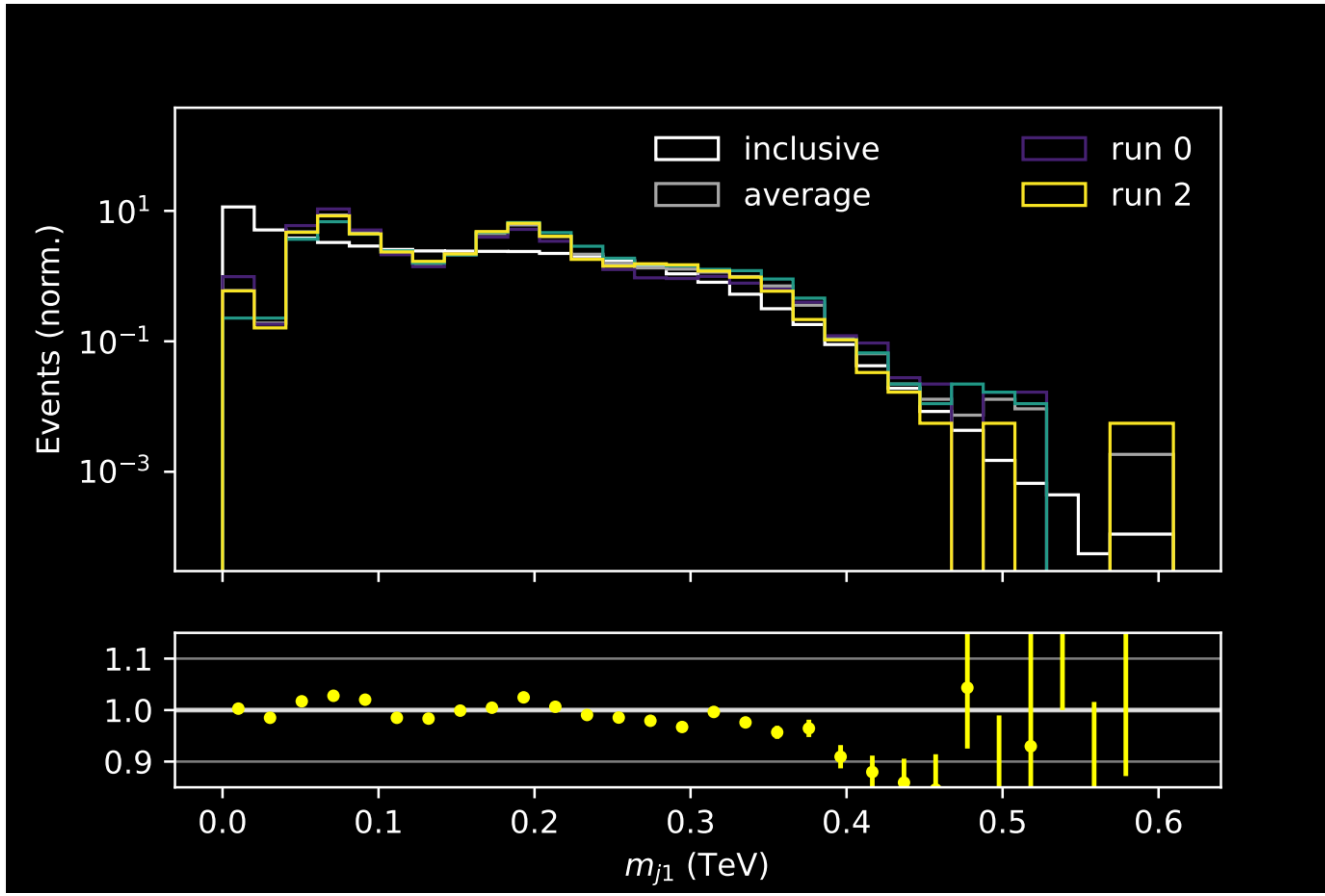
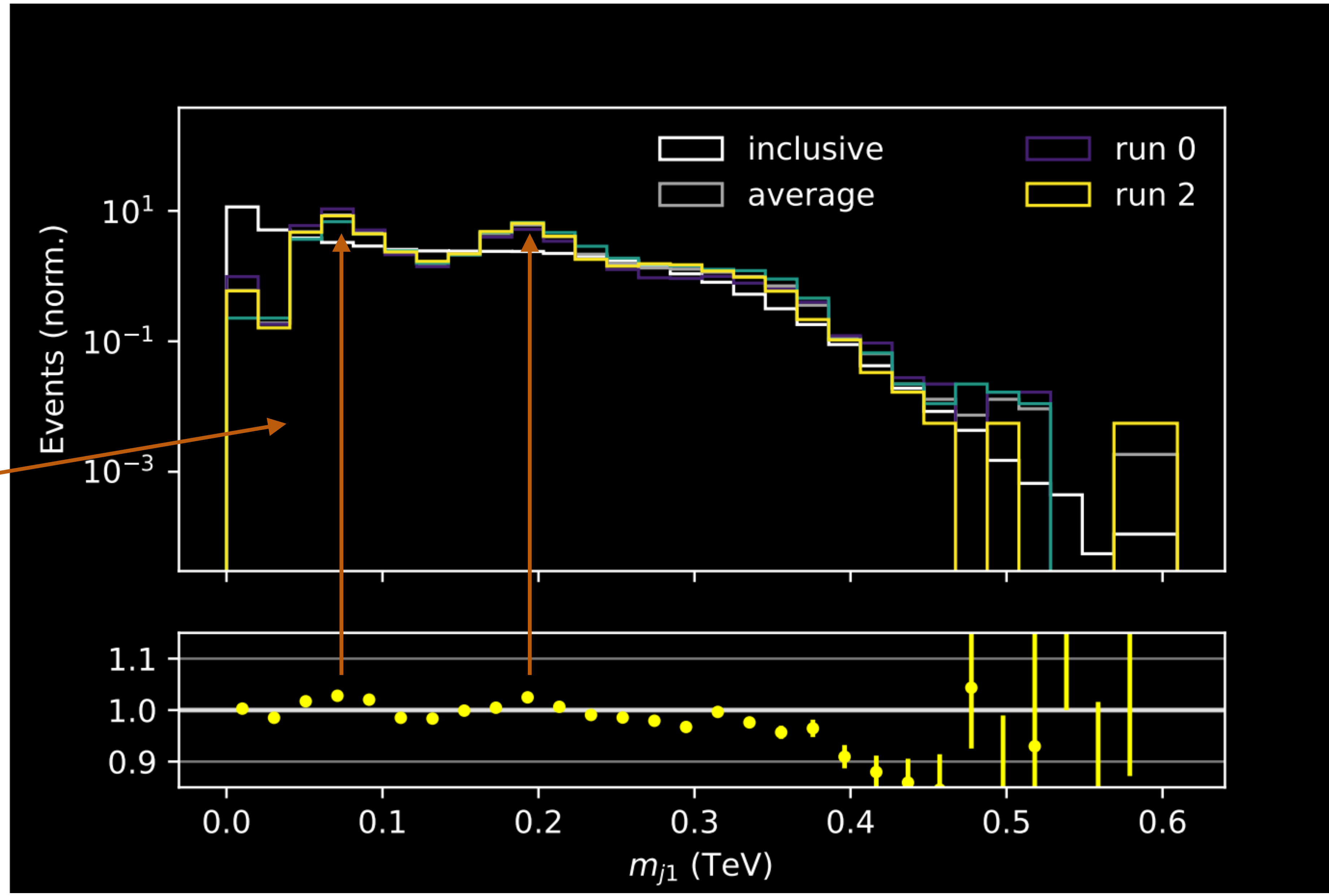


Figure by Sommerhalder

Generative non-closure

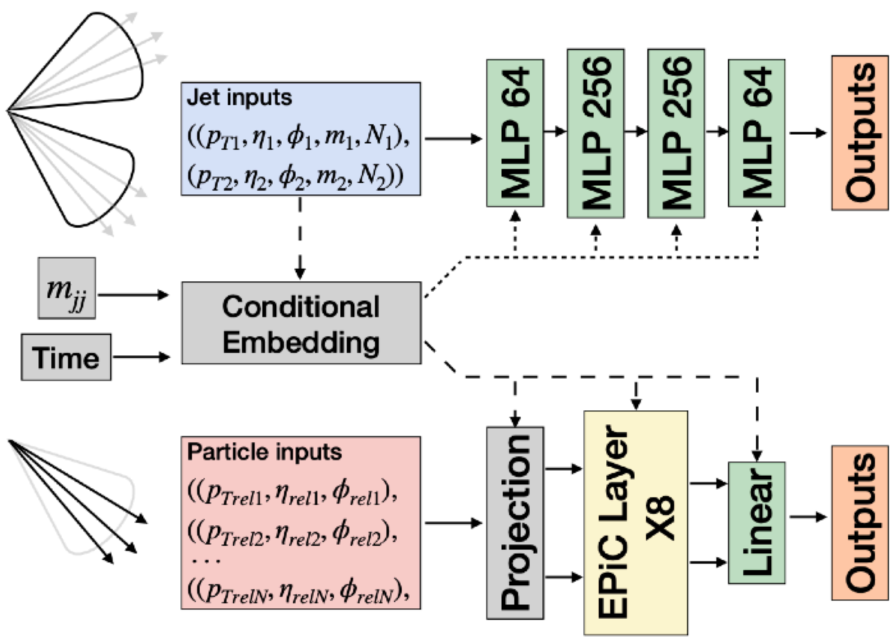


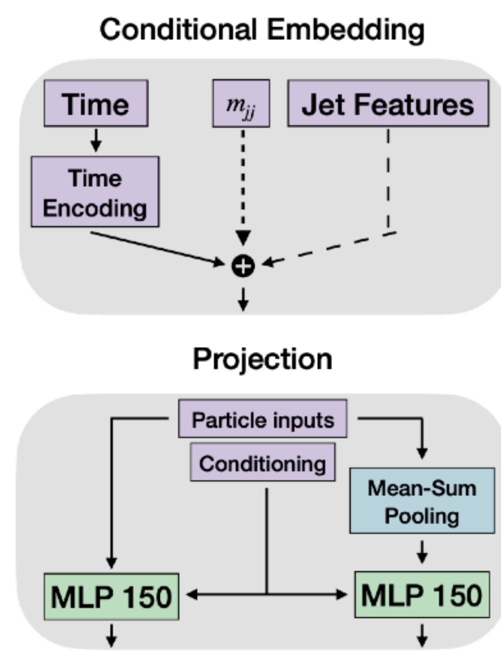
Sensitive to percent-level differences

Might benefit (highly) from clever uncertainty ideas

Figure by Sommerhalder

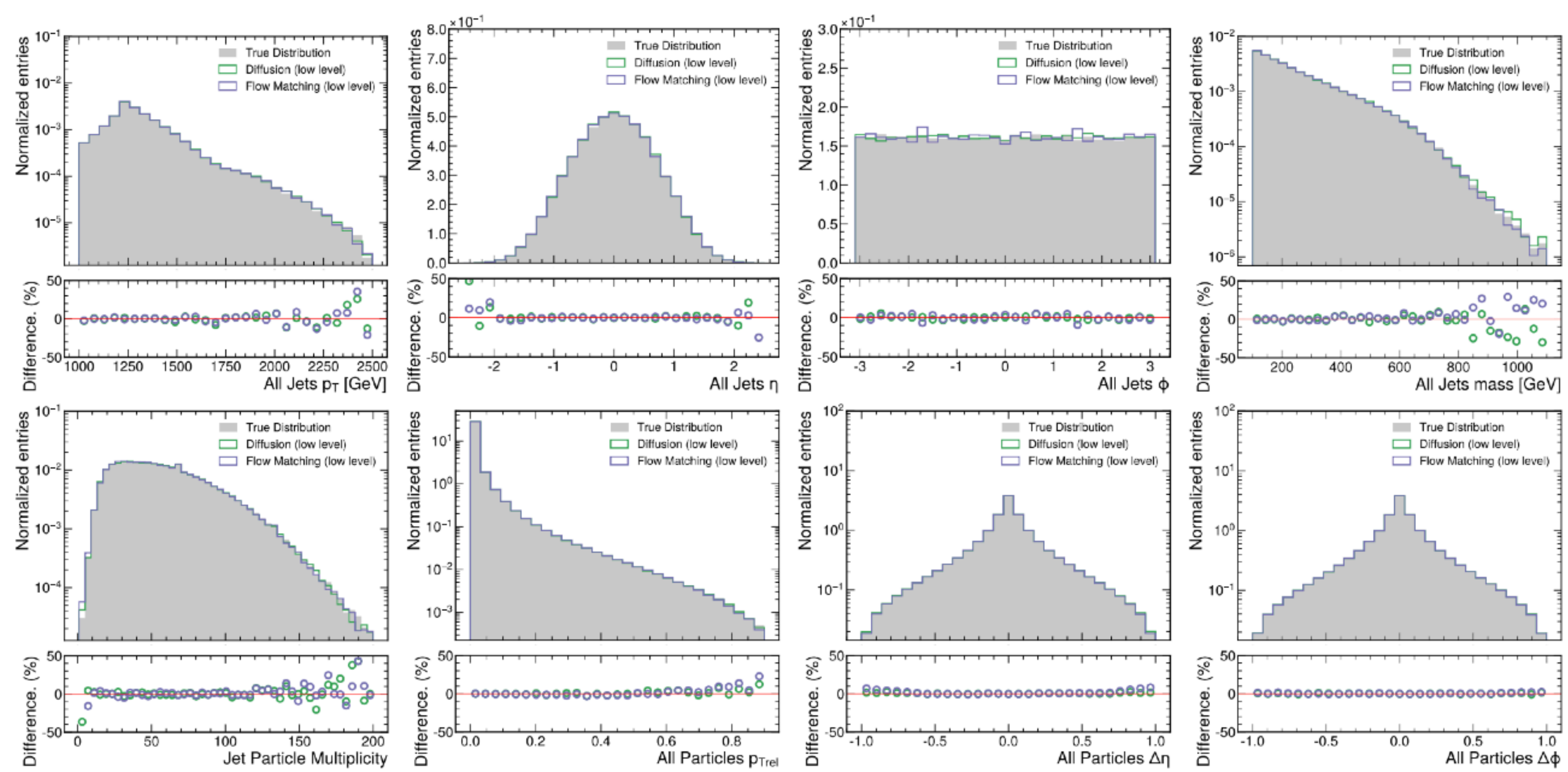
Gain in high dimensions?





Can extend to jet constituents

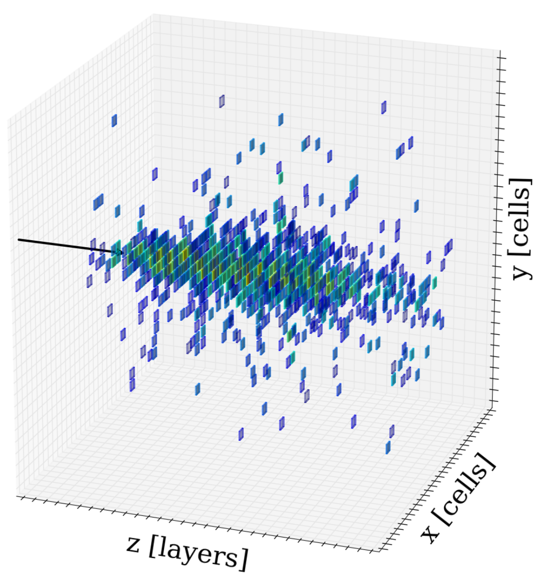
Improves performance, still need to understand generator closure

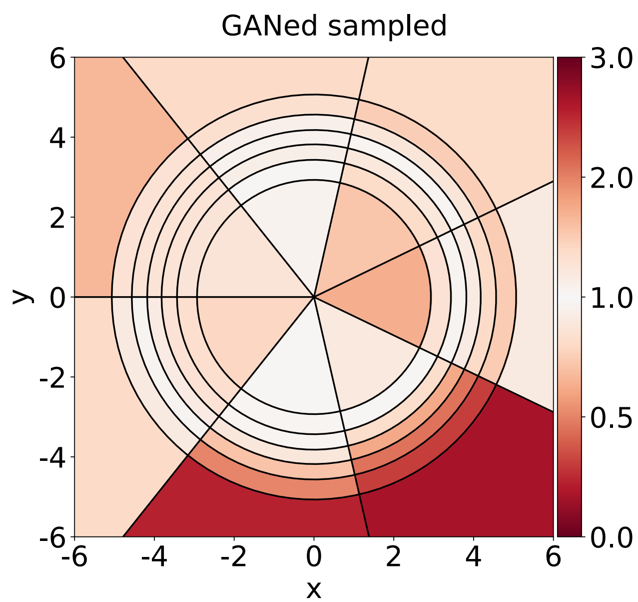


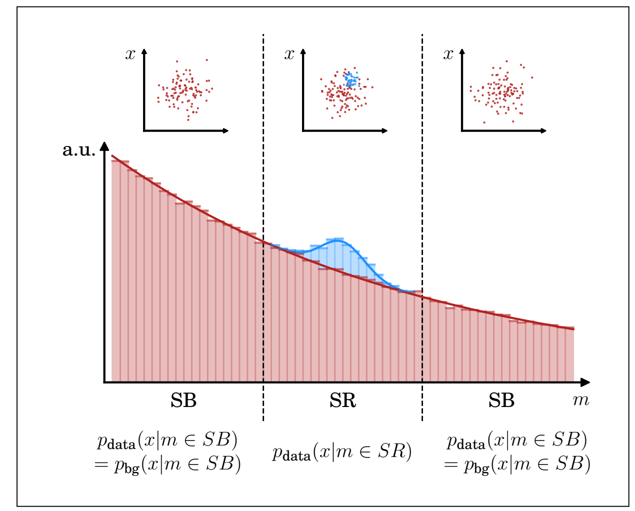
Buhmann, .., GK, et al 2310.06897

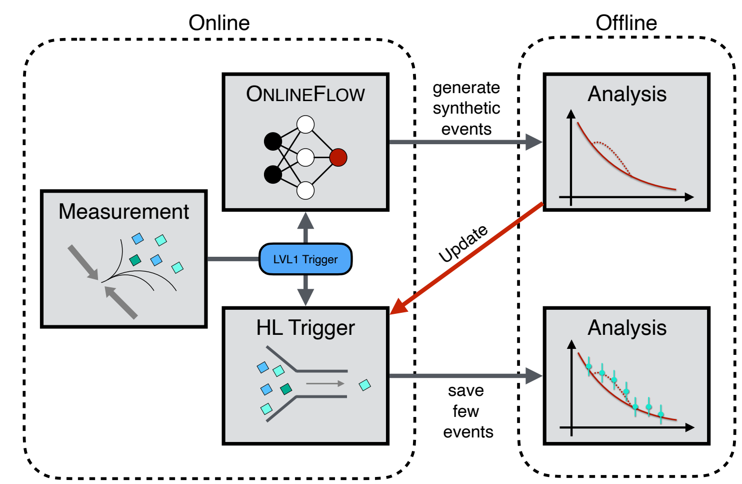
Closing

Closing









- Rapid progress in calorimeter simulation with generative models, including sophisticated benchmarks
  Chance to augment them with uncertainties?
- Anomaly detection as powerful technique to detect new physics.
  Inclusion of generative uncertainty might be crucial
- Demonstrate statistical gain from generative models
- Plays direct role generative model replaces data

Thank you!