Augmented Muon Background Simulation with Normalising Flows

Martina Ferrillo[1], Maxim Borisyak[2], Oliver Lantwin[3], Nicola Serra[1], Andrey Ustyuzhanin[2]

[1] Universität Zürich

[2] Constructor University Bremen

[3] Universität Siegen

SHiP

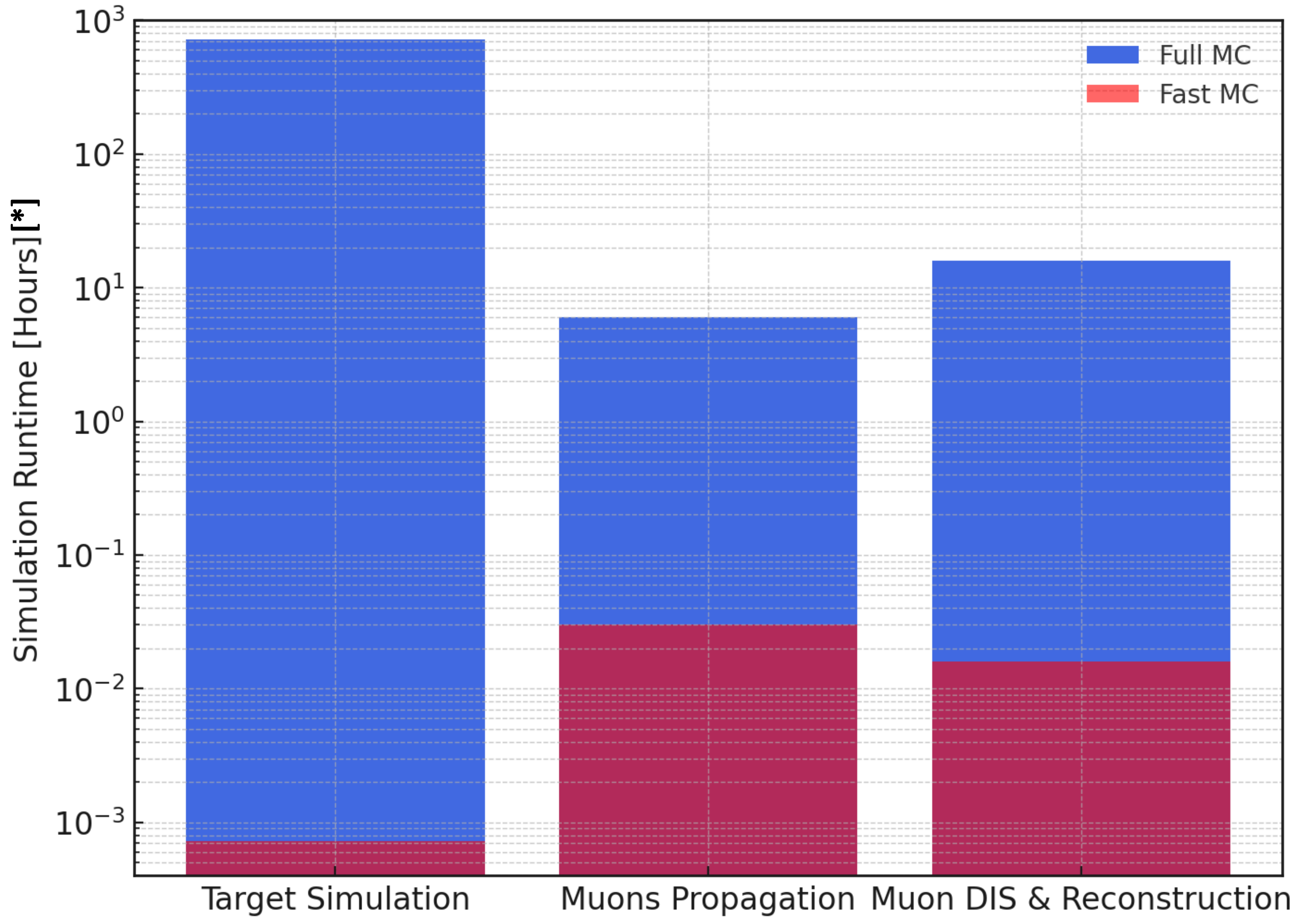
SHiP 33rd Collaboration Meeting

5th June 2025

Problem Statement: Simulation Speedup Needed

[*] for the equivalent of **1 spill**

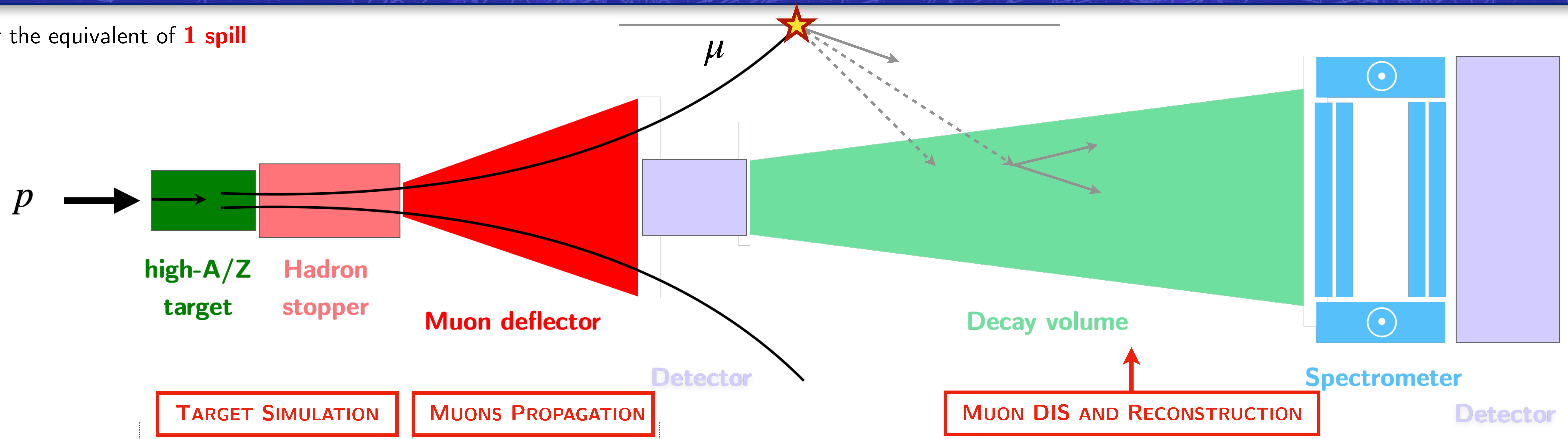
From my talk at SHiP's 32nd Collaboration Meeting



The current setup can be **significantly** optimised, achieving **orders of magnitude speed-up** in each simulation step.

How?

[*] for the equivalent of **1 spill**



	Target Simulation	Muons Propagation	Muon DIS and Reconstruction
Generator(s)	Pythia6/Pythia8 Geant4	Geant4	Pythia8 Geant4
Task	Simulation of **primary collisions of protons**, propagation of decay products	Propagation of **muons** in the **muon shield**	Simulation of **muon DIS**, propagation of decay products, reconstruction
Simulation Runtime [*]	ca. **1 month** on CERN CPU farm (optimistic)	ca. **6 hours** on CERN CPU farm (optimistic)	- ca. **11-16 hours** on CERN CPU farm (optimistic). - Limited generalisation power and potential statistical dependence: a **single muon is reused up to 7000 times** [slides]. - **Only 0.02% of the expected muon DIS interactions have been simulated** (ca. 4.5M), and **most will never contribute to the background** (waste of computing resources)

Limitations Arising from Limited MC Statistics

Limitations

Possibly prone to criticism, case of **Muons DIS** background.

- **Limited generalisation power:** over-usage (x7k) of **narrow phase space** region due to limited statistics muon samples
- **Strong assumptions** requiring validation: **factorisation** trick between veto and selection efficiencies

Solutions

Both could be overcome by simply generating **larger MC** samples, which comes at the **expensive cost of computational resources and unfeasible processing times.**

⇒ A novel approach to **background** simulation is **urgently** needed and should be on the **top of priorities** list.

Action Plan for Muon DIS Fast Simulation

Achievable MC simulation **speedups**:

(1) Accelerating **target simulation** with GANs and exploiting azimuthal symmetry $\rightarrow$ factor $\sim \mathcal{O}(10^6)$

(2) Muons propagation in MS significantly faster in ad-hoc, standalone tools $\rightarrow$ factor $\sim \mathcal{O}(10^2)$

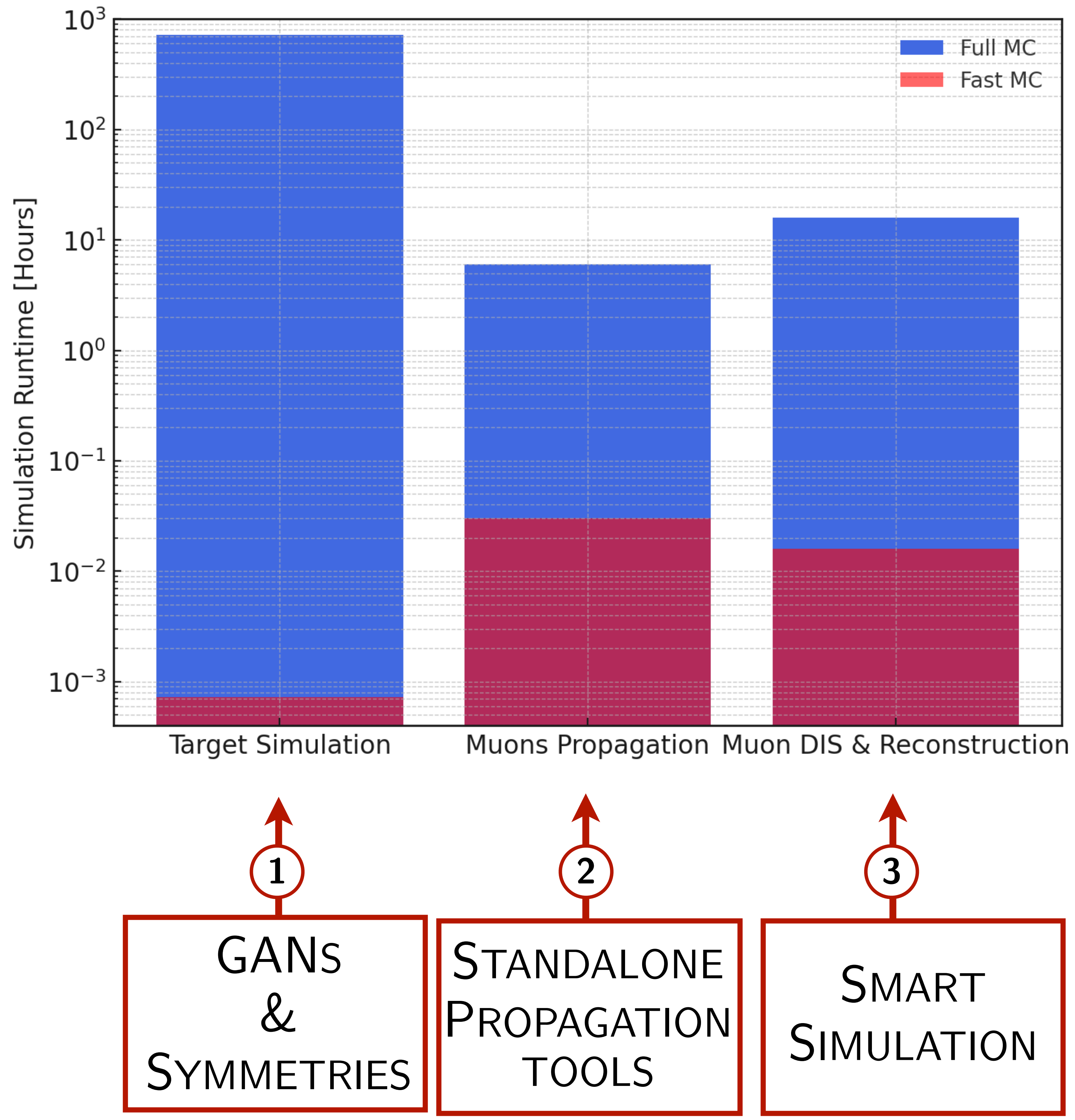
(3) Intelligent simulation strategy

(a) Use **NF** to efficiently generate **synthetic muon data after the MS**, interpolating sparse distribution in phase space

$\rightarrow$ factor $\sim \mathcal{O}(10^6)$

(b) Focus full simulation effort exclusively on potential muon DIS background events

$\rightarrow$ factor $\sim \mathcal{O}(10^3)$

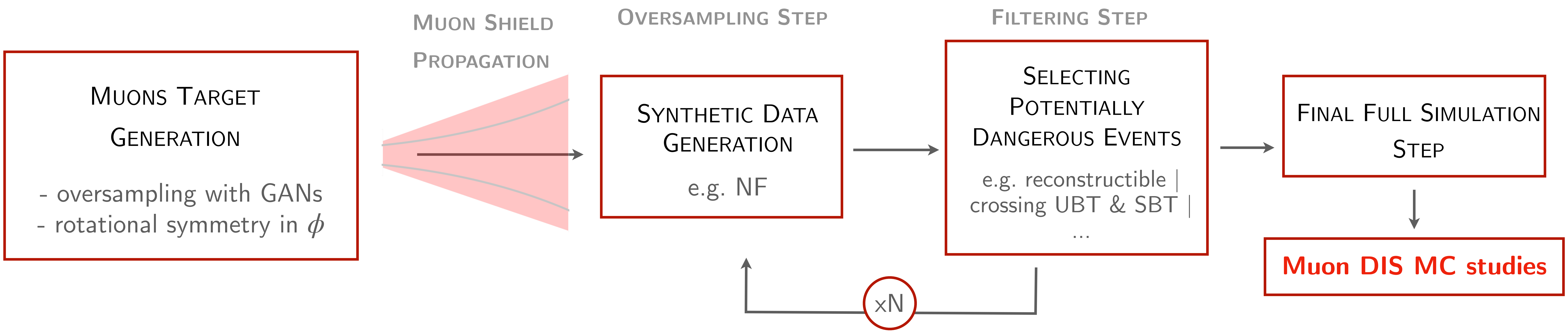


Smart Simulation Approach

Efficiently generate fewer events by ensuring **simulations focus exclusively on relevant background**

Iterative Augmentation Strategy

- Use generative surrogate models to create **synthetic muon samples**, *interpolating* between scarcely populated bins
- Apply cascade of acceptance requirements, *a.k.a.* "**Filters**", corresponding to a certain selection stage in the background generation step



Efficiently generate fewer events by ensuring **simulations focus exclusively on relevant background**

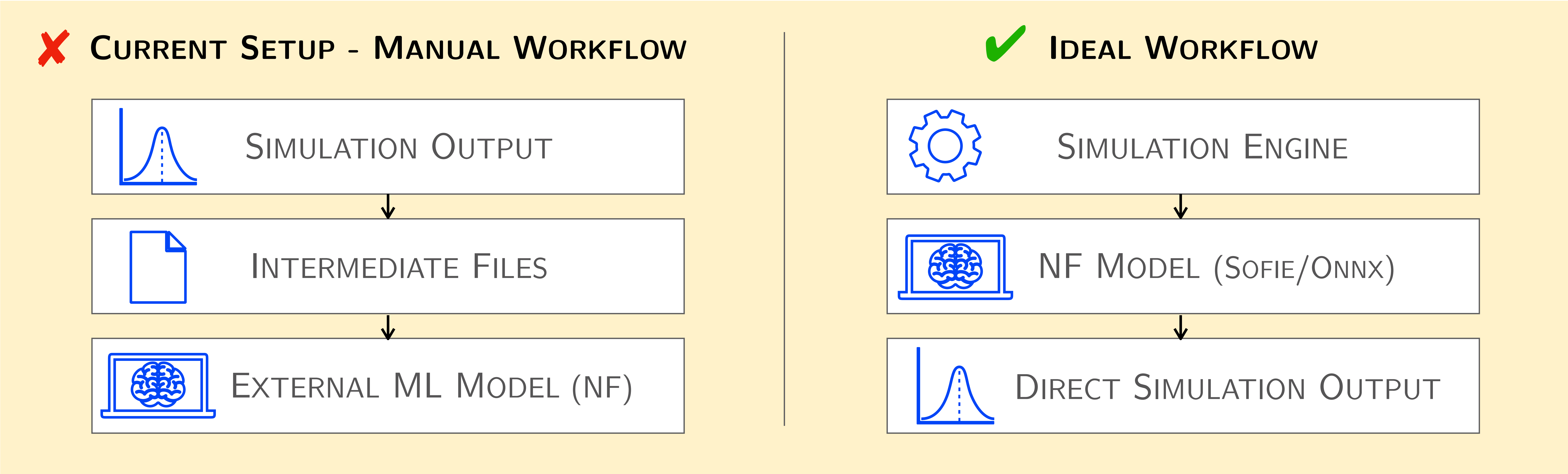
Iterative Augmentation Strategy

- Use generative surrogate models to create **synthetic muon samples**, *extrapolating* between scarcely populated bins
- Apply cascade of acceptance requirements, *a.k.a.* "**Filters**", corresponding to a certain selection stage in the background generation step

Integration of Fast Simulation Within Common Framework

[O. Lantwin]

- Any fast simulation mechanisms developed should be available to the whole collaboration, seamlessly
- **FairShip** can already read **Muons & Matter** input data from NTuples (and adding support for new formats trivial)
- Closer **integration** using runtime inference from **fast simulation model using SOFIE/ONNX** feasible once proof-of-concept using normalising flows is ready (no intermediate files needed, fully integrated)
  - Inference performance on CPU will require study
  - Dependencies available via LCG, once common software migrates to use LCG



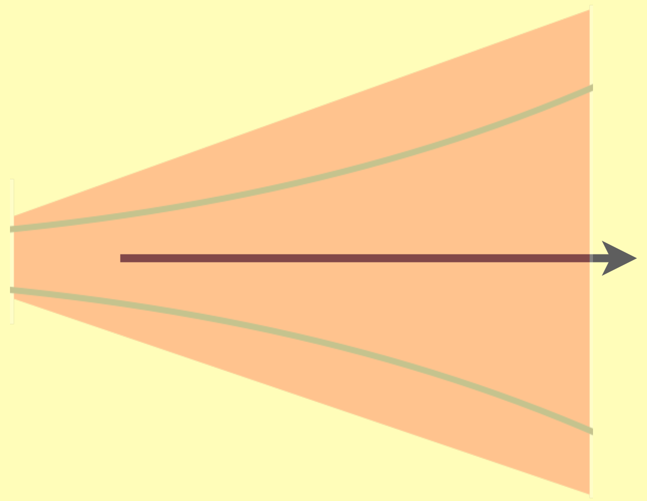
(*) diagram for a single iteration

Step-By-Step Procedure

Muons Target Generation

- oversampling with GANs
- rotational symmetry in ϕ

Muon Shield Propagation



next slides

Oversampling Step

Synthetic Data Generation

e.g. NF

Filtering Step

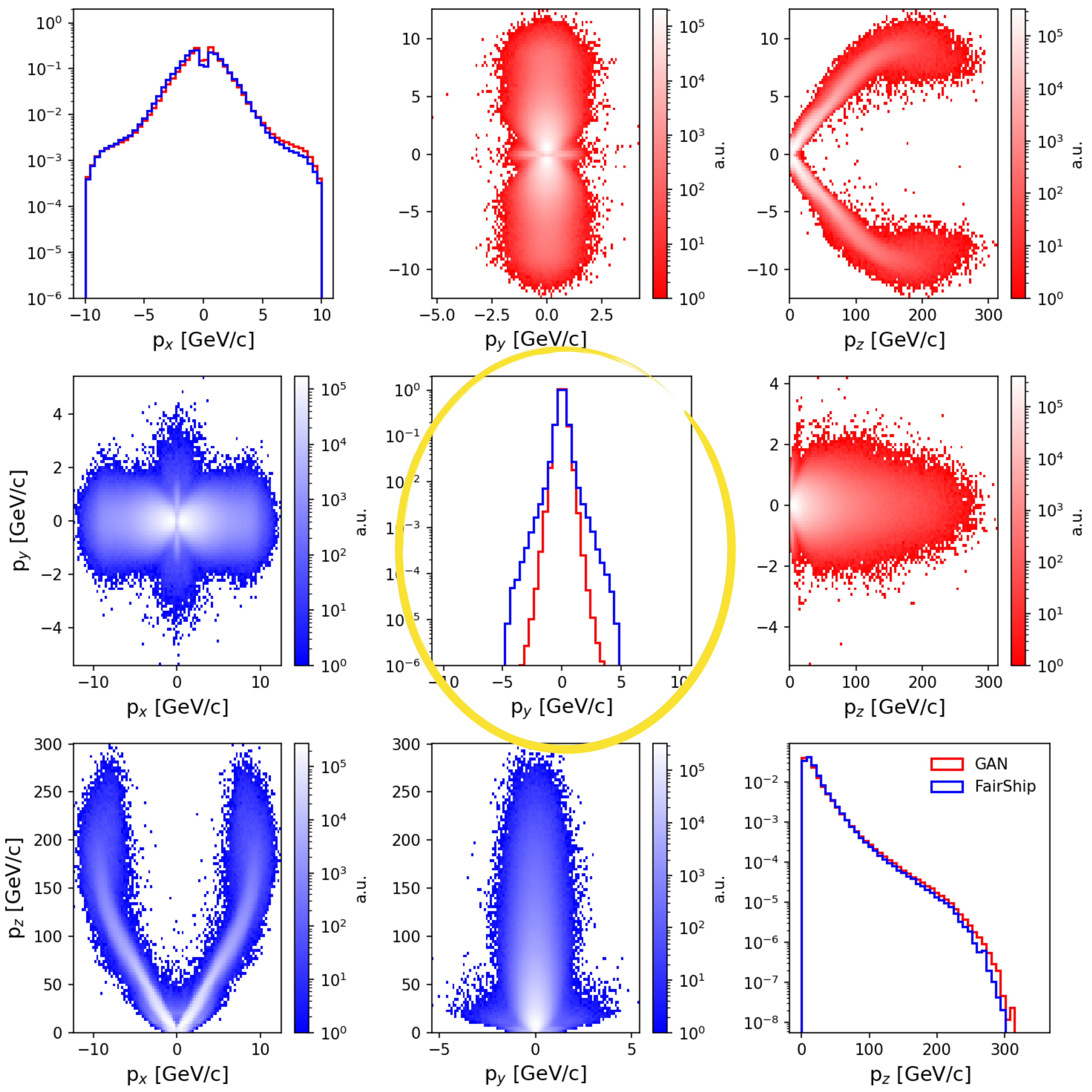
Selecting Potentially Dangerous Events

e.g. reconstructible | crossing UBT & SBT | ...

xN

Final Full Simulation Step

Muon DIS MC studies



[O. Lantwin, MF]

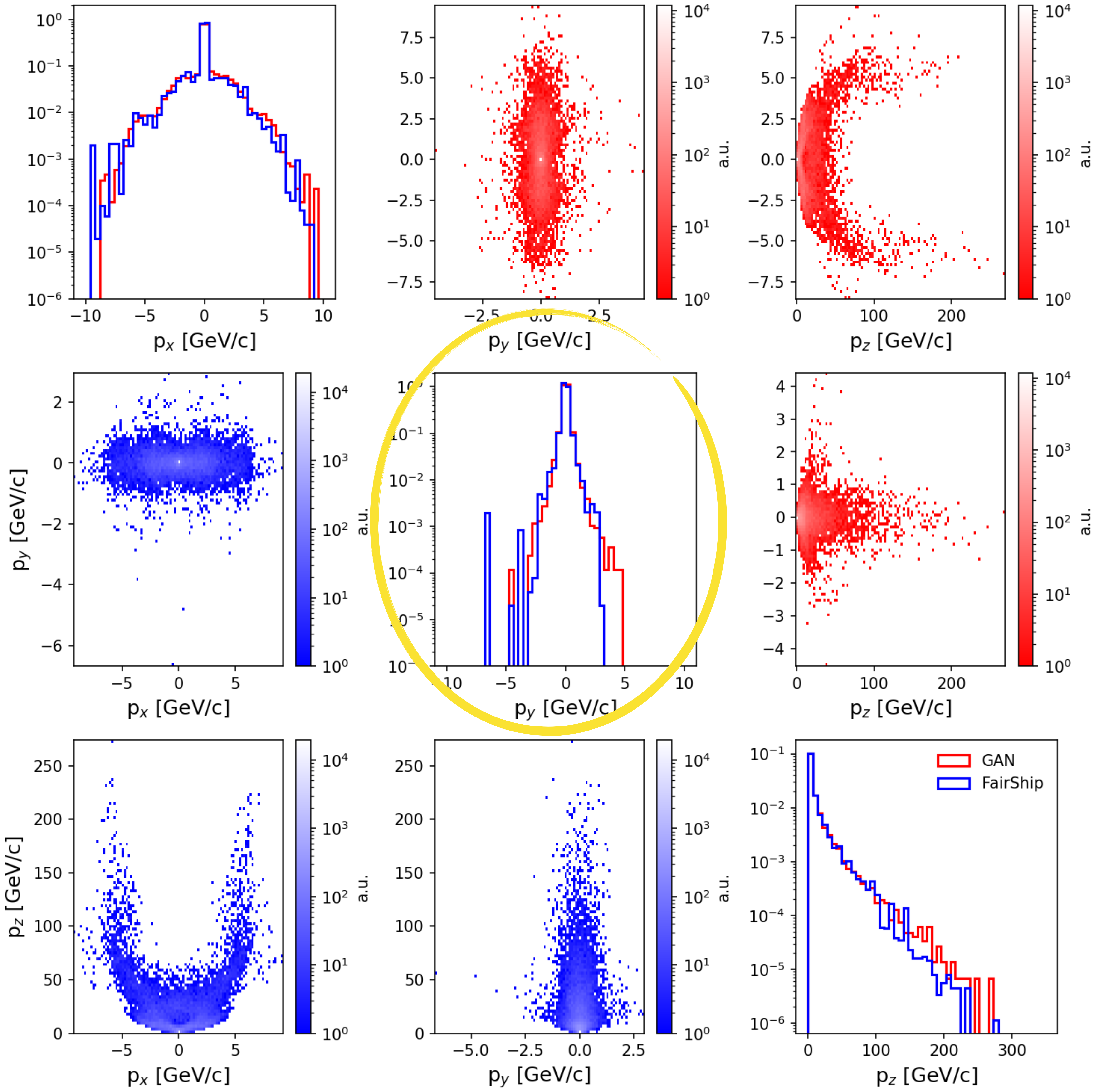
MUON KINEMATIC DISTRIBUTION AT THE END OF THE MUON SHIELD

USING MUONS & MATTER

- **Propagating muons** inside the **MS** (optimal design for the warm magnet option) using standalone tool [Luis Felipe's talk]
- Good match, **except for** p_y
- Floating point differences btw M&M/GAN samples?

⇒ Let's test this within the Fairship framework

USING FAIRSHIP



[O. Lantwin, MF]

MUON KINEMATIC DISTRIBUTION AT THE END OF THE MUON SHIELD

USING FAIRSHIP

- Despite the lower stats, the effect on the p_y tails seem to vanish

Step-by-Step Procedure

Muons Target Generation
- oversampling with GANs
- rotational symmetry in ϕ

Muon Shield Propagation

Oversampling Step

Synthetic Data Generation

e.g. NF

Filtering Step

Selecting Potentially Dangerous Events

e.g. reconstructible | crossing UBT & SBT | ...

xN

Final Full Simulation Step

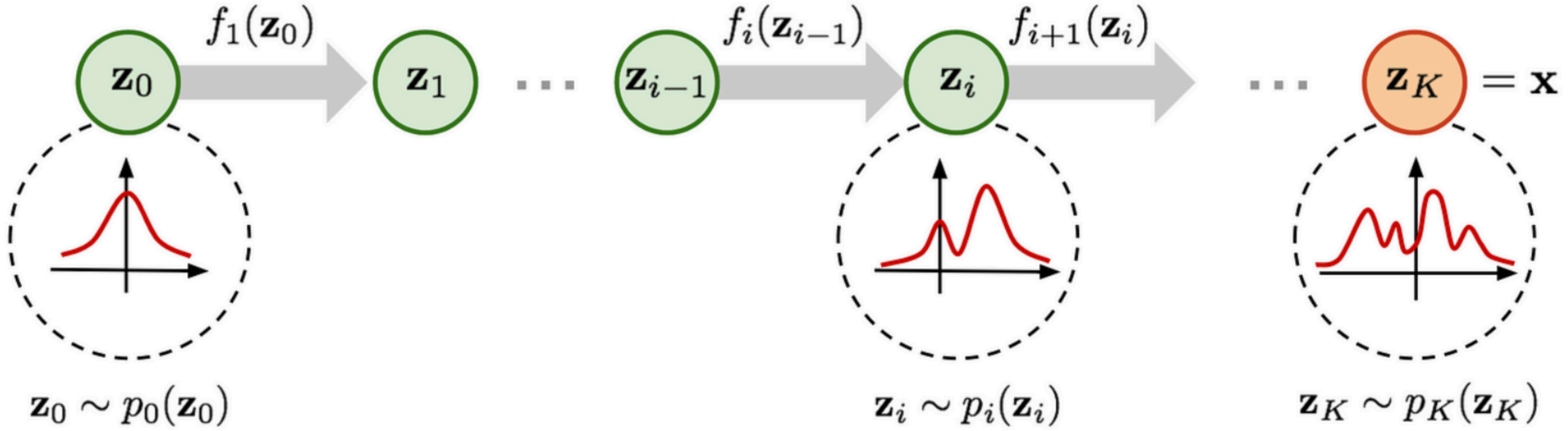
Muon DIS MC studies

next slides

Step 4 Use **fast** **deep-generative models** to generate **synthetic muon data** *after* the MS to interpolate between sparse simulation bins

(e.g. Normalising Flows, but any other Deep Generative Model could fit here)

Normalising Flows (NF) are **probabilistic generative models** that learn an **invertible mapping** between a simple distribution (*e.g.* Gaussian) and a complex target distribution

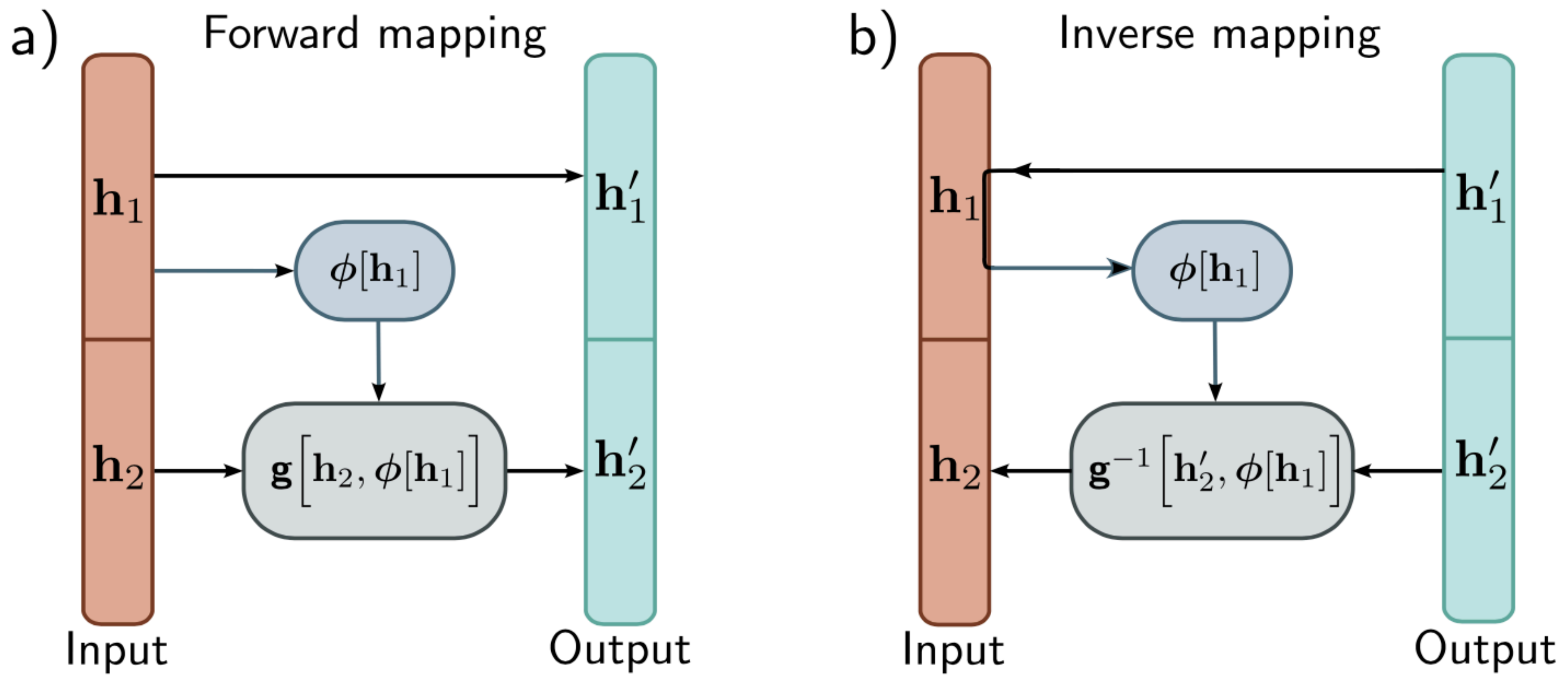


NF at work (in a nutshell)

- Starting with an easy to sample distribution, usually Gaussian, $z \sim \Pi(z)$
- Applying an **invertible**, **parameterised transformation** $x = f(x \,|\, \phi)$, typically a **NN** with parameters ϕ
- After training, the **NF approximates the target** distribution $P(x \,|\, \phi) = \left| \frac{\partial f(z, \phi)}{\partial z} \right|^{-1} \Pi(z)$

Results with Coupling Normalising Flows

Architecture: Coupling NF



- Good match between **True** VS **Generated** muons, after the muon shield, for a relatively simple and small NF model (16k parameters)
- Fast inference time per NF-generated event $\simeq 10^{-7}$ s [from talk]

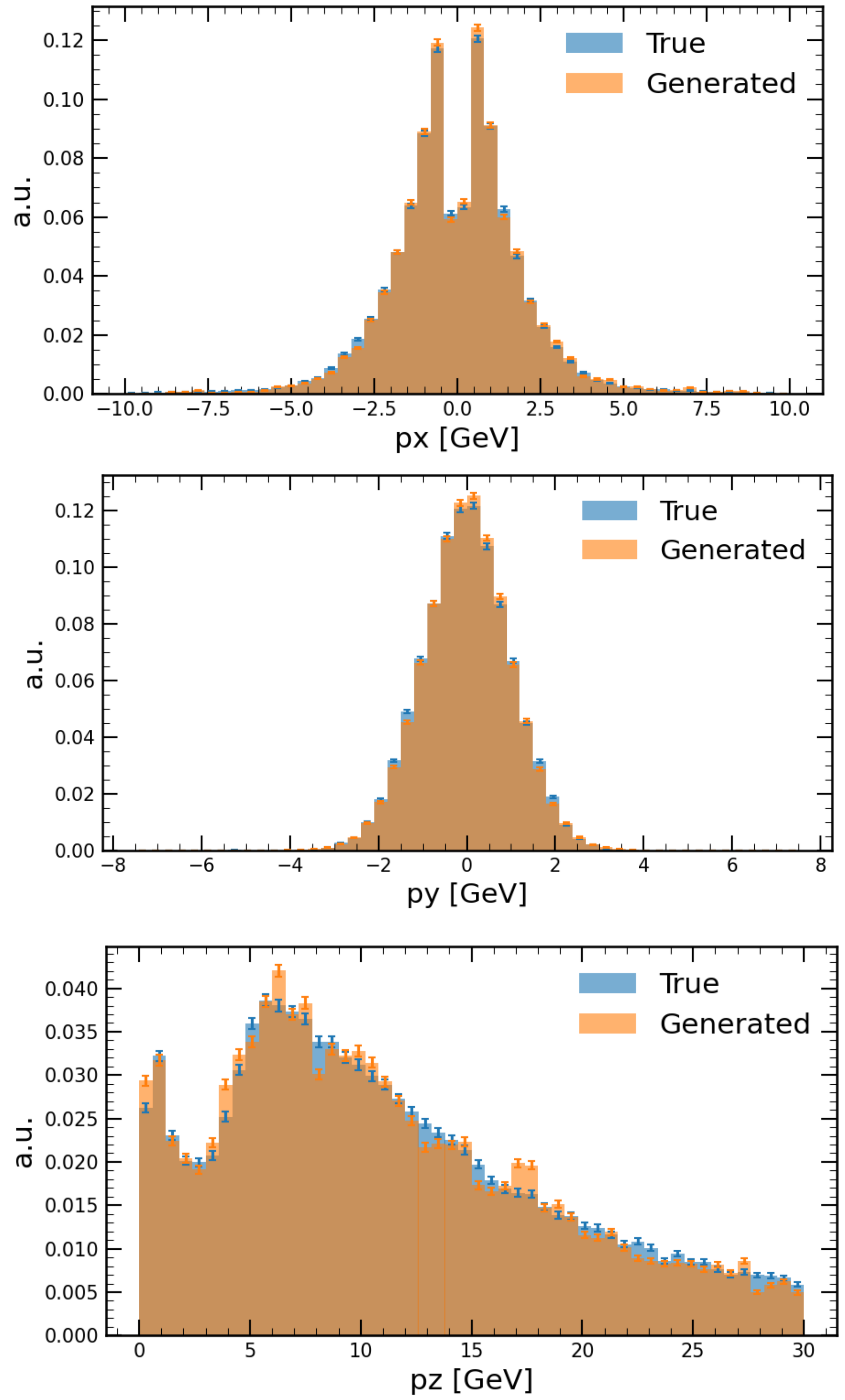
  Estimated time to generate statistically independent events, equivalent to the 1 spill benchmark (4.5 M): **~0.58 s** on **1 GPU**, NVIDIA L40S

$\Rightarrow$

Speedup
Factor $\mathcal{O}(10^6)$

Data Augmentation in
Scarcely Populated PhSp Area

Muon Spatial Momentum - After MS



Structured Normalising Flows

[M. Borysiak]

Propagation and selection of muons in SHiP is a **sequential procedure** of clearly defined steps :

- Muons get produced in the target area

$i-1$

- After deflection in the MS, a fraction of muons enters the Decay Volume

i

- Muons interacting inelastically within acceptance

$i+1$

- Muon DIS events forming a "good candidate"

...

- Muon DIS events eluding the veto system

n

- ...

⇒ **Proposal: structure the latent space of the Normalising Flow to reflect this hierarchy**

Structured NF Loss (I)

[M. Borysiak]

Loss

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim P_0}\, l(x \mid \theta, 1) + \lambda_1 \cdot \mathbb{E}_{x\sim P_1}\, l(x \mid \theta, \sigma_1^2) + \lambda_2 \cdot \mathbb{E}_{x\sim P_2}\, l(x \mid \theta, \sigma_2^2) + \dots$$

- $l(x \mid \theta, \sigma^2)$ is the likelihood of the sample x given the parameters θ and variance σ^2 of the prior distribution of the latent variable
- $\lambda_i > 0$ hyperparameters

Nested Structure

$$x = f(z), z \sim \mathcal{N}(0, \mathbb{I}), \text{all muons;}$$

$$x = f(z), z \sim \mathcal{N}(0, \sigma_1^2\mathbb{I}), \text{muons entering the decay volume;}$$

$$x = f(z), z \sim \mathcal{N}(0, \sigma_2^2\mathbb{I}), \text{muons entering spectrometer;}$$

$$\dots$$

$$\text{where} : 1 > \sigma_1 > \sigma_2 > \dots$$

- Minimising this Loss is equivalent to minimising the weighted Kullback-Leibler divergence

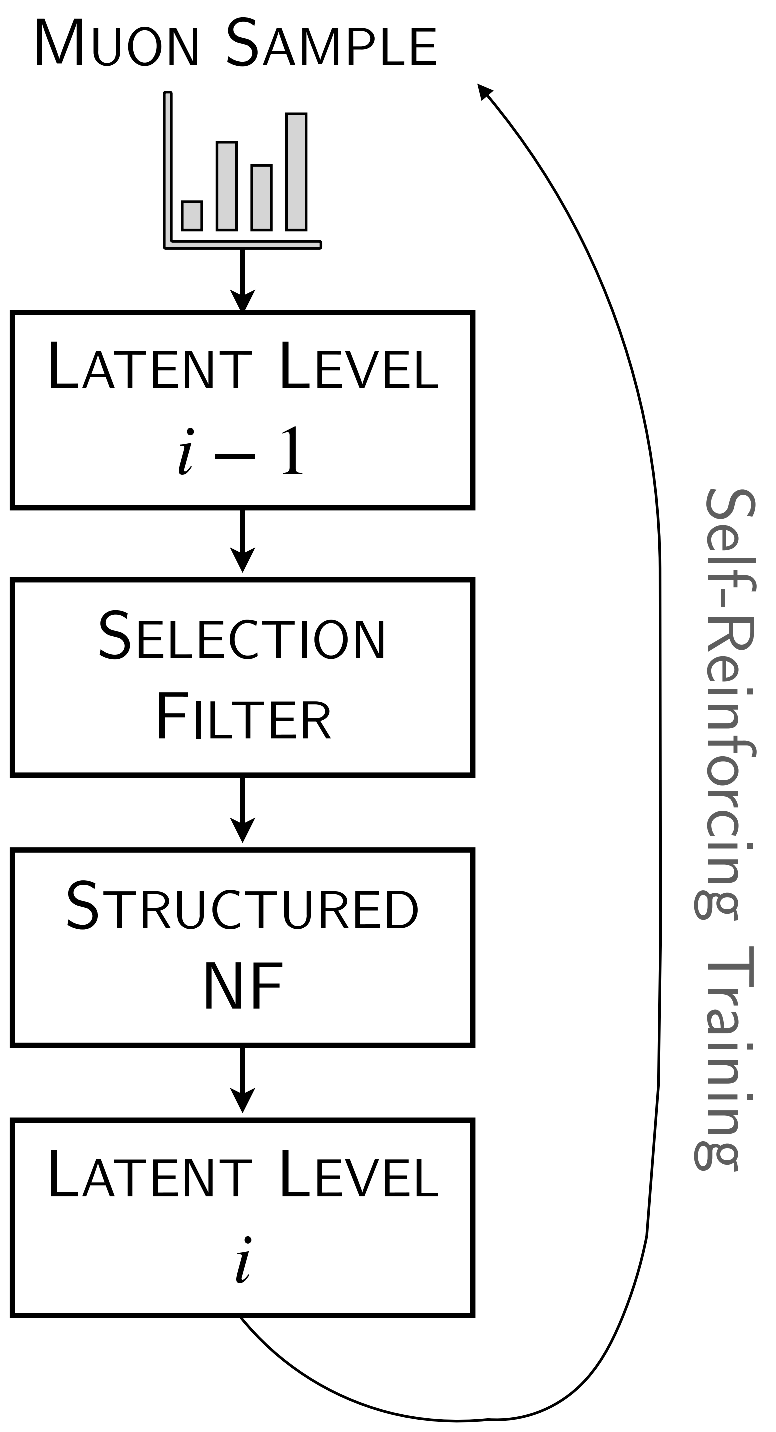
  ⇒ Special structure allowing to **generate both inclusive samples** or specific, **dangerous background samples**

Structured NF Loss (II)

[M. Borysiak]

The proposed loss:

- Enables fine control of the surrogate
- Equivalent to weighted Kullback-Leibler divergence
  - NF converges to the true distribution
- **Forces precise simulation of rare but interesting events (“*dangerous backgrounds*”)**
  - Increased penalty for such events
- Can be trained in a self-reinforcing manner



Planned as a Summer Student project, potentially in Bremen

Open Call to Project Collaborators

- Given the **urgency** and **importance** of the task, we **encourage** the **contribution** from **diverse groups** with complementary expertise to **quickly reach the completion stage**.
- The **project** is already **beyond** the **proof-of-principle** phase:
  - Ready to build, test, and deploy the final working solution, including the last *iteration steps*
  - Dedicated **technical paper** planned

Great opportunity (especially for PhDs!) to help the SHiP Collaboration and gain visibility

For questions and project collaboration requests, please get in touch with me or Nico

martina.ferrillo@cern.ch

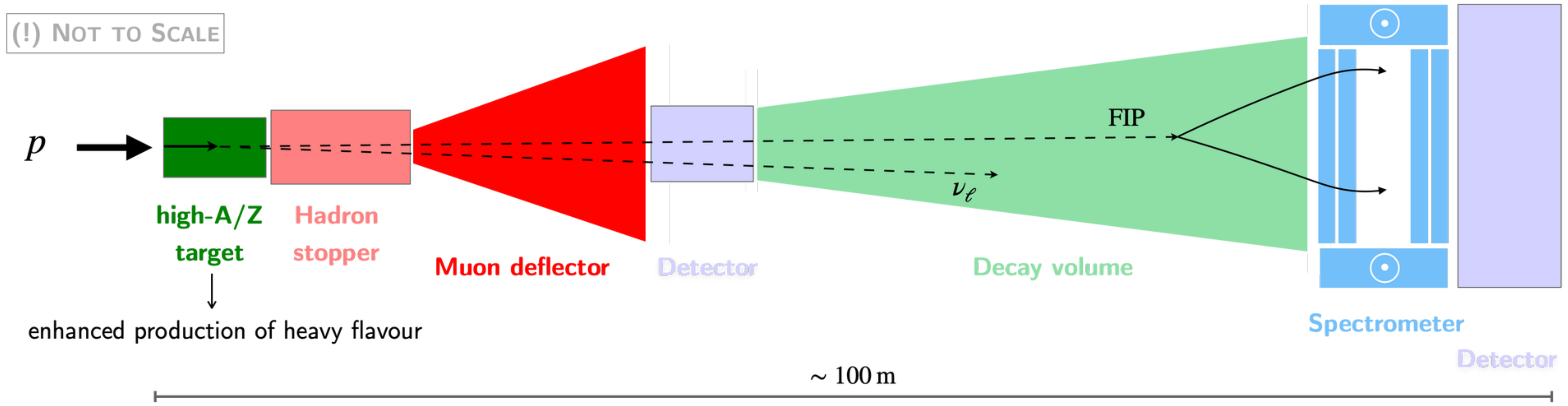
Thank you

Backup Slides

How Much MC Simulation is "Enough"?

- **Ideal Scenario:** generating enough MC samples to ensure statistical uncertainties are *negligible* compared to systematic effects
- **Current Practice:** typically generating $\sim$ 10x more MC than the amount of expected data

Example: Target Simulation



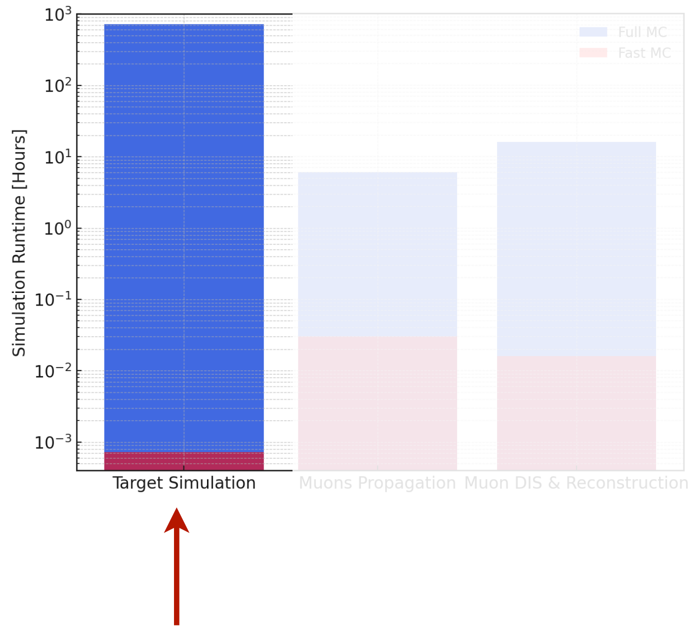
- Currently generated MC sample corresponds to only about **1 spill** [*] of real data
- **Generation** from target simulation to reconstruction in HSDS takes **months** on dedicated CPU infrastructure

These statistical limitations severely restrict our ability to perform optimisations/estimate backgrounds accurately

⇒ Can't we just generate more?

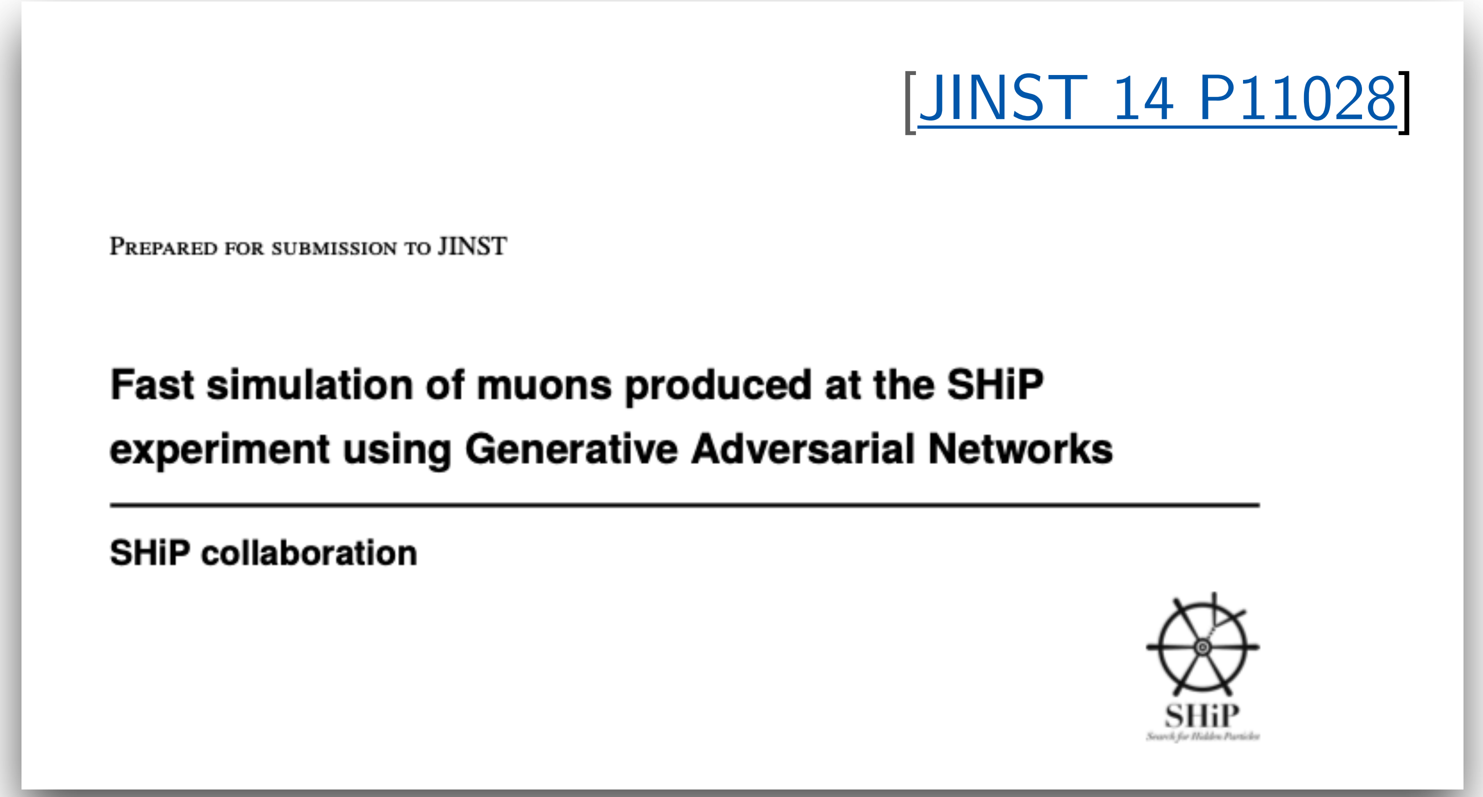
[*] Simulated sample statistics for the *minimum bias* corresponding to 6.5×10^{10} PoT, *i.e.* 2.5×10^{-3} spills [/eos/experiment/ship/data/Mbias/background-prod-2018/README]

ACCELERATING TARGET SIMULATION WITH GANs AND SYMMETRY

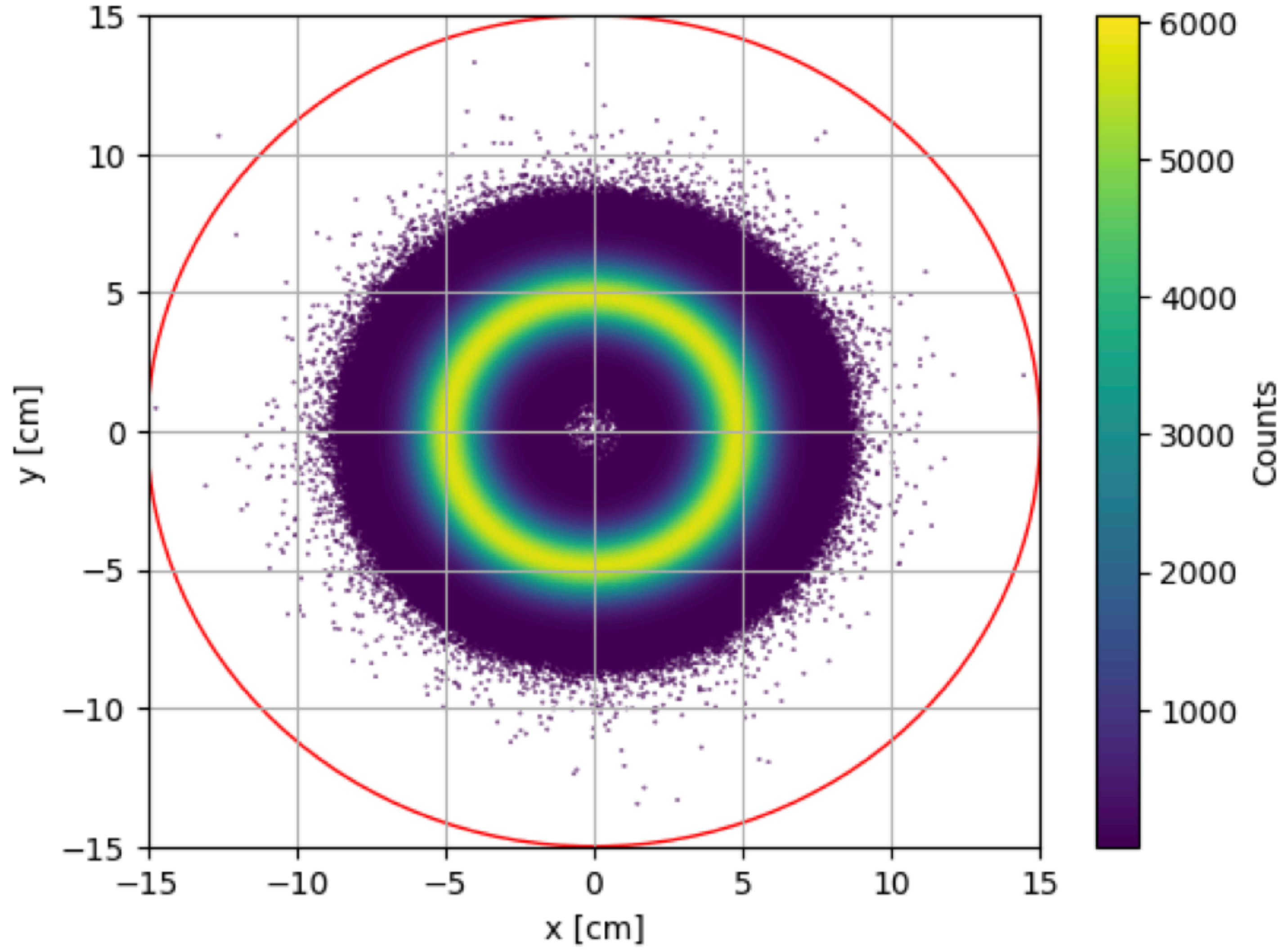


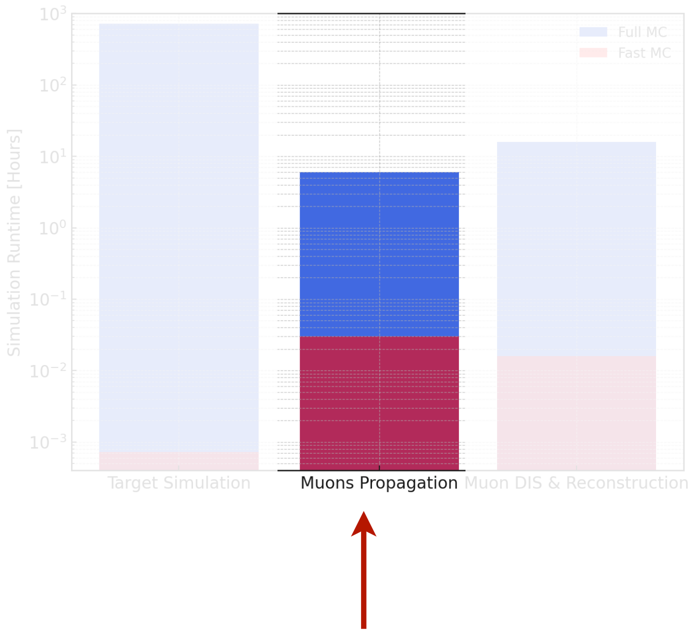
- Custom **SHiP-GAN** model produces muons $\sim \mathcal{O}(10^6)$ faster than the current Pythia + GEANT4 combo, preserving **physics fidelity**
- **Speed-up** applies primarily to already **sampled** regions of **phase space** produced in the full simulation (interpolation vs extrapolation)
- Additional $\sim \mathcal{O}(10^3)$ **increase** in **statistics** by exploiting rotational **symmetry** in the **azimuthal plane**

[JINST 14 P11028]



[from Luis Felipe's talk]





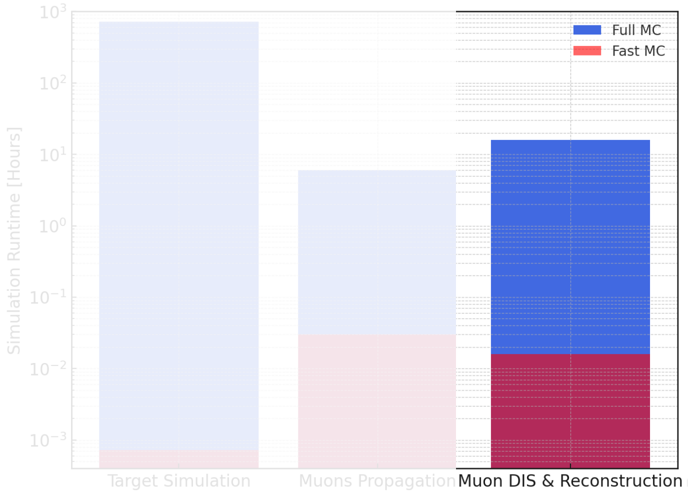
- Significant speedup with **standalone propagation tools**
  - **(1)** CPU and GEANT4-based: **2x faster** compared to FairShip implementation
  - **(2)** GPU-based using pre-cached GEANT4 losses: $\sim \mathcal{O}(10^2)$ **faster** than CPU-based implementations

- Relevant also for the evaluation of **muon-induced background** on the electronics of the **SND** detector **within the muon shield**

STANDALONE OPTIMISATION FRAMEWORK

GPU-BASED MUON PROPAGATION

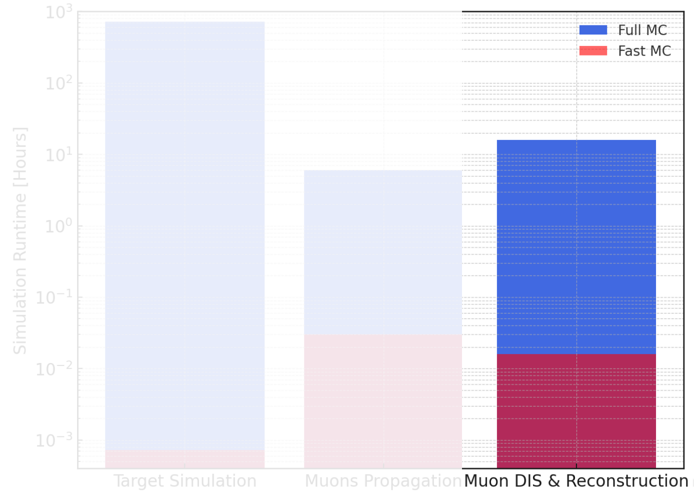
Smarter Muon DIS Event Generation



Simulating **all** possible **muon DIS** interactions ($\sim 2 \times 10^{10}$) so far **inefficient** and **computationally impractical**

- Non-representative input muon sample, given the limited portion of sampled phase space (only 6k seeds)
- If 16h needed for ca. 4.5M events, the **full sample** would require **ca. 8 years**

Smarter Muon DIS Event Generation



Simulating **all** possible **muon DIS** interactions ($\sim 2 \times 10^{10}$) so far **inefficient** and **computationally impractical**

Proposed solution

(1) Use **fast deep-generative models** to generate **synthetic muon data** *after* the MS to interpolate between sparse simulation bins

e.g. **Normalising Flows**, see next slides, or using gaussian smearing of muon kinematics (W. Bonivento)

(2) **Identify** and select **only dangerous events** that realistically contribute to **muon DIS background**

e.g. using simple MLP/DNN for a classification task, identifying **only reconstructible events** (reduction factor $\sim \mathcal{O}(10^3)$)

Step-By-Step Procedure

Muons Target Generation

- oversampling with GANs
- rotational symmetry in ϕ

next slides

Muon Shield Propagation

Oversampling Step

Synthetic Data Generation

e.g. NF

Filtering Step

Selecting Potentially Dangerous Events

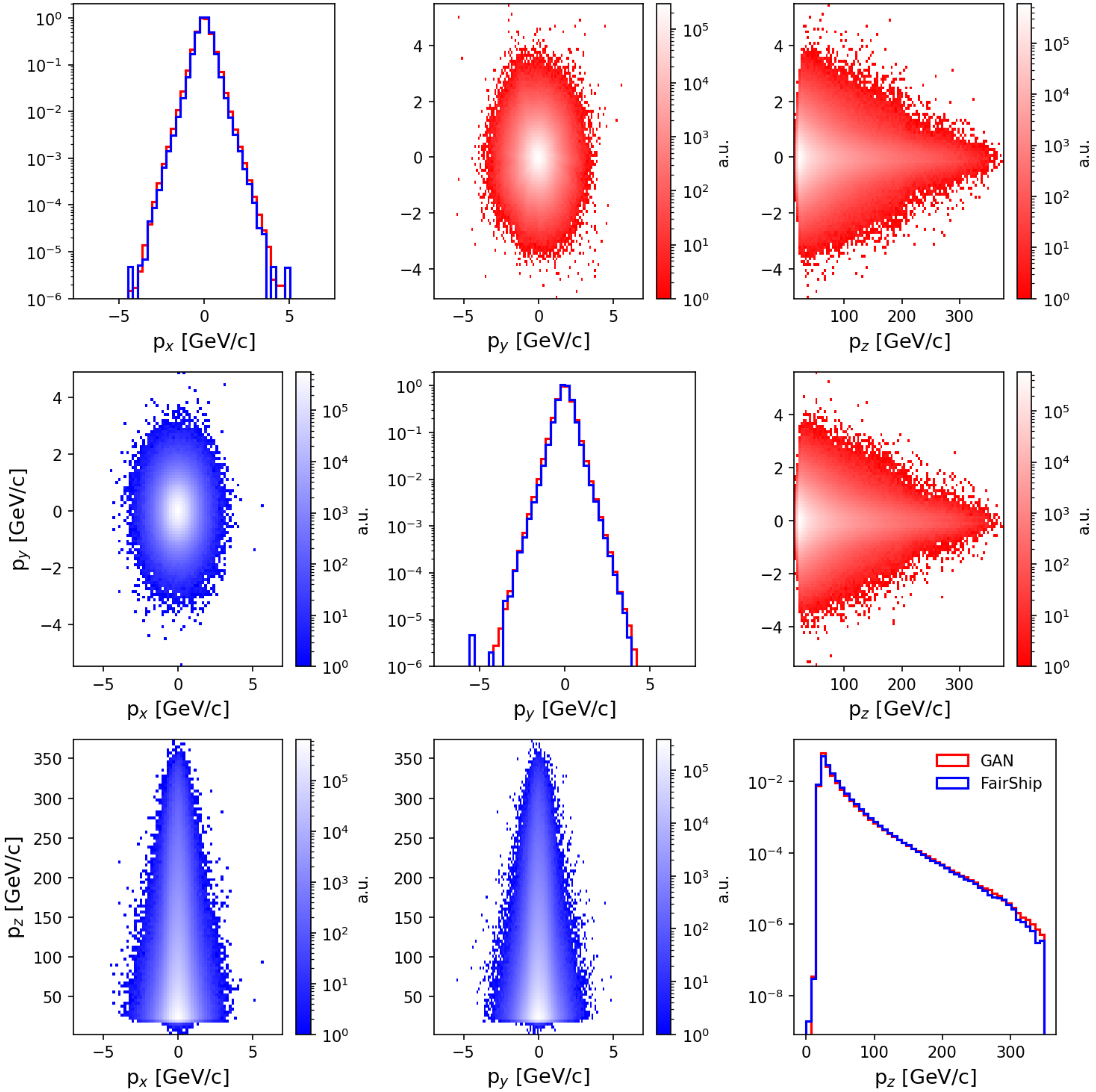
e.g. reconstructible | crossing UBT & SBT | …

xN

Final Full Simulation Step

Muon DIS MC studies

Target Simulation with GANs - Before MS

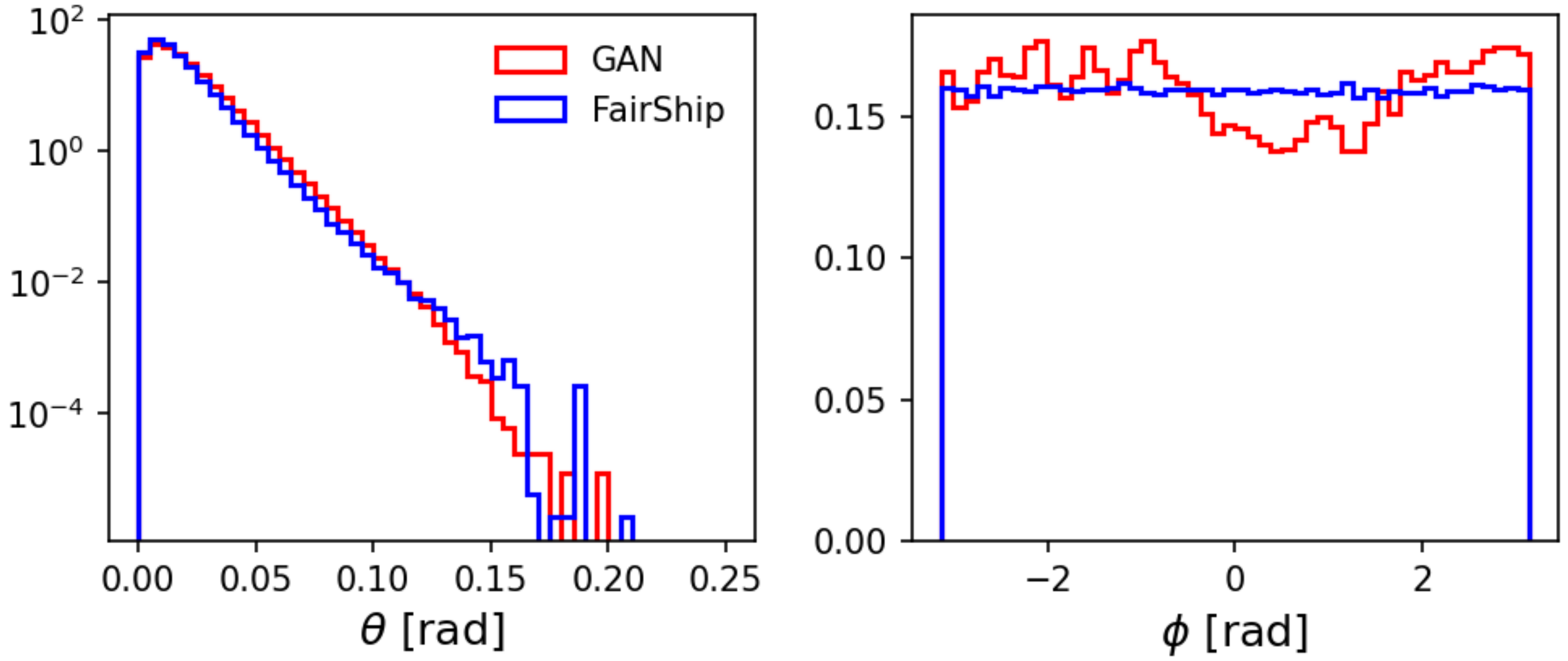


Muon Kinematic Distribution in the Target (Production Point)

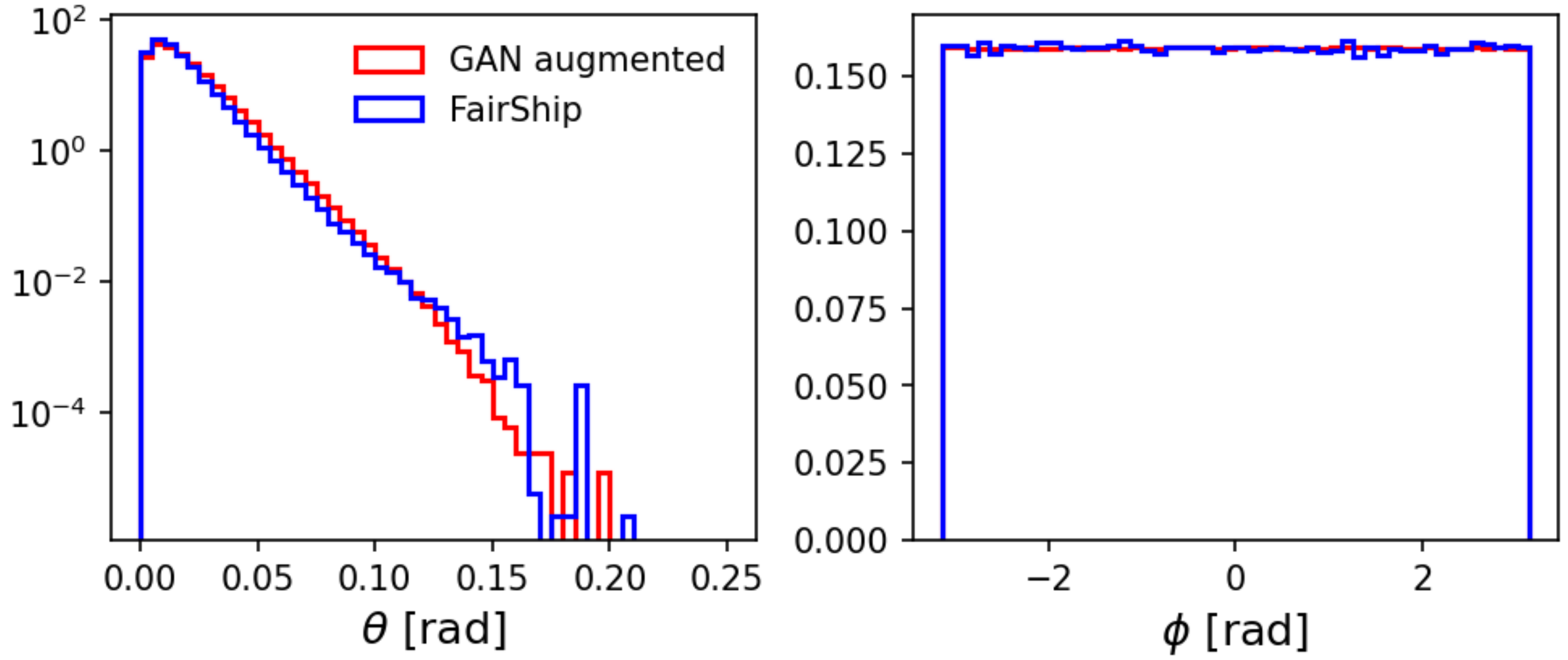
Step 1

- **Oversampling** full MC (FairShip) MC with the **GAN tool** [JINST 14 P11028]
- Sanity check between **FairShip** and **GAN**-generated muons
- Comparison btw kinematic distribution before the propagation in the Muon Shield

Before Augmentation in ϕ



After Augmentation in ϕ



Muon Angular Distribution in the Target (Production Point)

Step 2

- Comparison btw angular distribution before the propagation in the Muon Shield
- **ϕ augmentation**

  Exploiting rotation invariance in the azimuthal plane by introducing a random rotation

Reconstructible events survive the selection criteria below:

[from CERN-SHiP-INT-2024-001]

Cut	Value
Good daughters	nDoF>25, χ^2/nDoF < 5, $p_{\text{track}} > 1$ GeV
Number of "good" candidates per event	1
DOCA	< 1 cm
Vertex distance from vessel's wall	> 5 cm (transverse), 20 cm (longitudinal)
IP (f.r.)	< 10 cm
IP (p.r.)	< 250 cm

Table 5: Event selection criteria on the muon-DIS background.

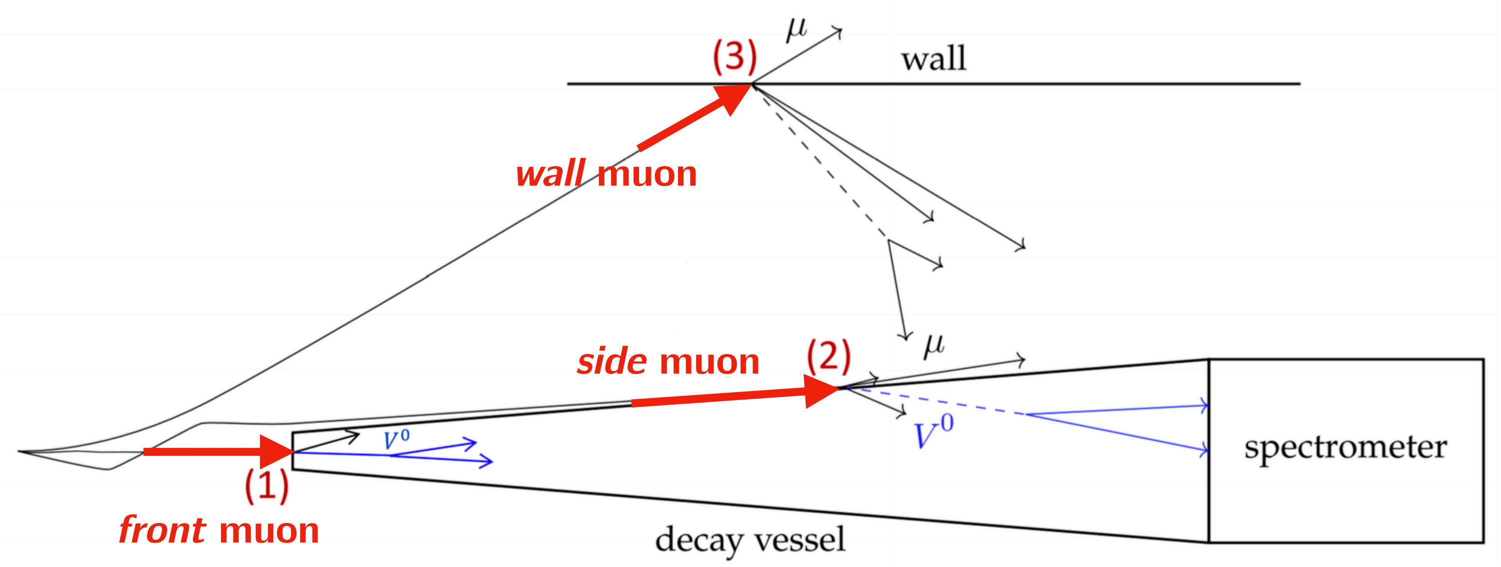
Setup	$\boxed{N_{\text{DIS}}^{\text{pre-sel}}}$	$N_{\text{DIS}}^{\text{pre-sel+SBT}}$	$N_{\text{DIS}}^{\text{pre-sel+SBT+UBT}}$
2024_{vac}	$1.9 \cdot 10^5$	73	0
2024_{He}	$4 \cdot 10^5$	718	0

Table 7: DIS events per full timeline; effect of veto.

Results: Coupling Normalising Flows

Goal of the Project

Speed up the ***SHiP muon background simulation*** by ***orders of magnitude*** ($\sim \mathcal{O}(10^6)$) and, in turn, generate a much larger and representative muon sample.

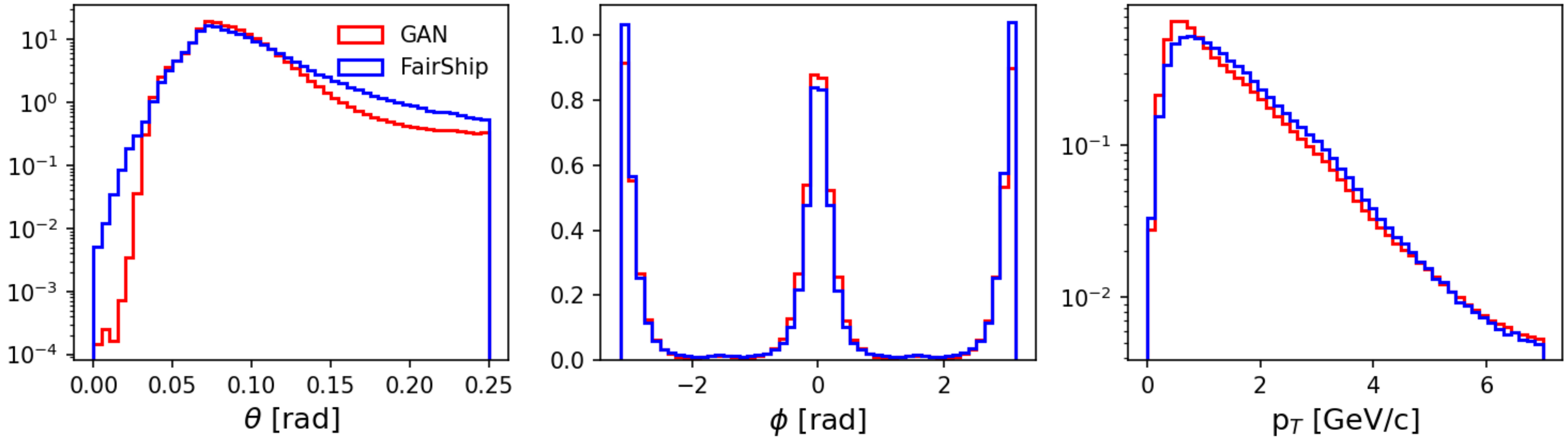


$\Rightarrow$ ***NF*** to ***learn the kinematic distribution*** of ***muons*** producing hard scattering in the proximity of the SHiP decay vessel.

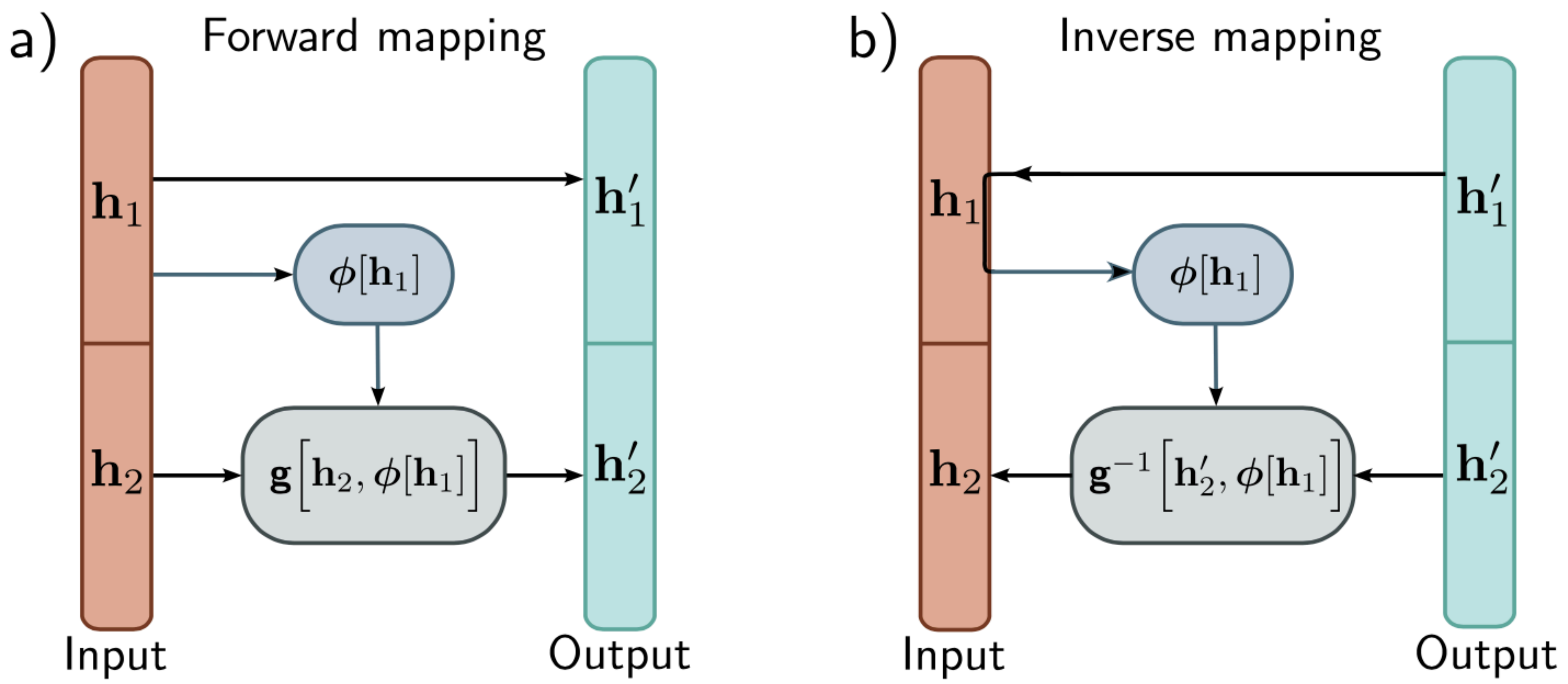
Architecture

- **Model type**: Coupling Normalising Flows (RealNVP)
- **Input features**: muon 4-momentum (p_x, p_y, p_z, E)
- **Generation time per event with NF** (on 1 GPU, NVIDIA L40S): $\simeq 1.47 \times 10^{-7}$ s
  - **Benchmark time** corresponding to the MC simulation of 10^{11} PoT (2.5×10^{-3} spills): "months of running on dedicated CPU farms" [arXiv:1909.04451]

Muon Angular Distribution - After MS



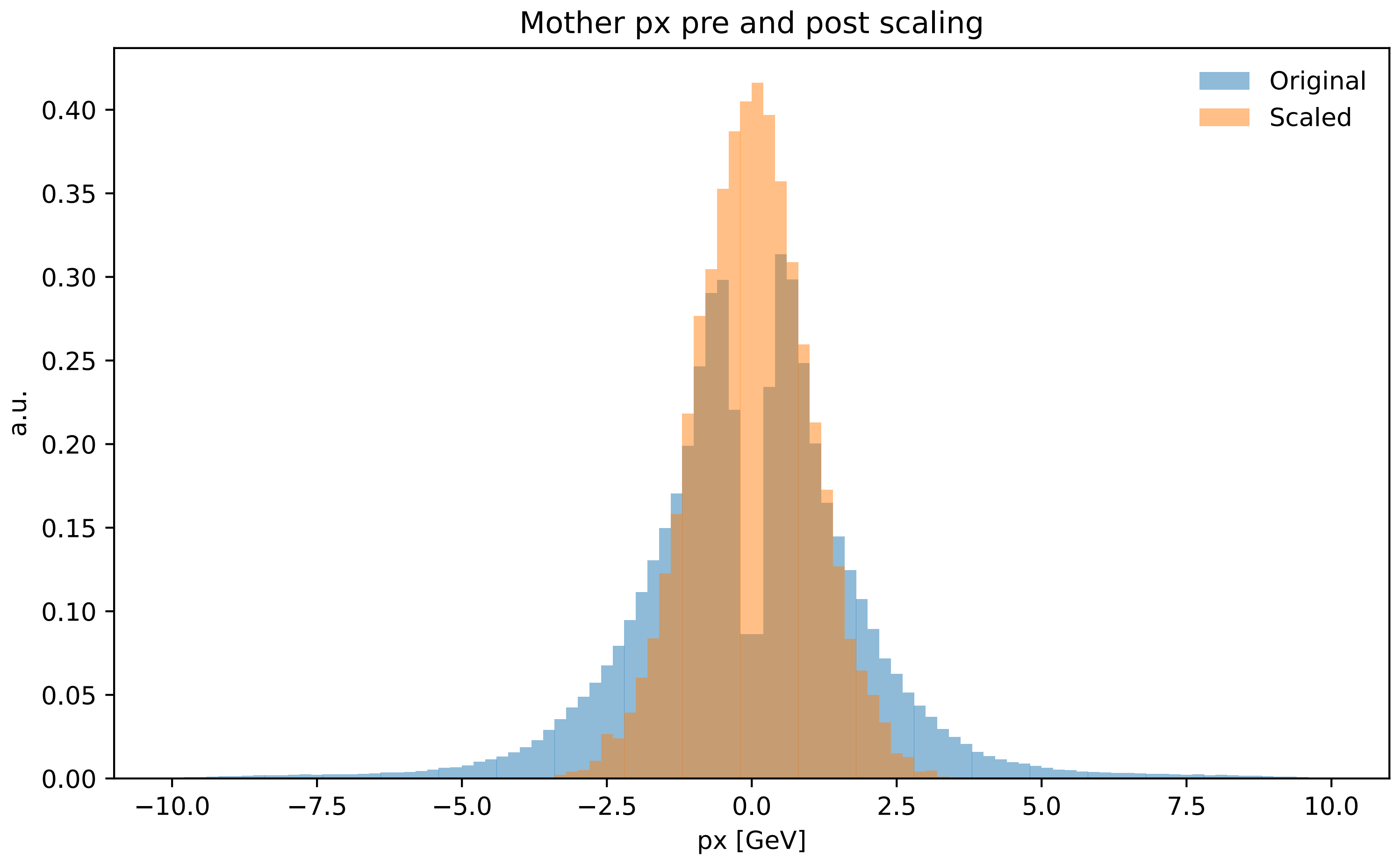
Coupling Normalising Flows

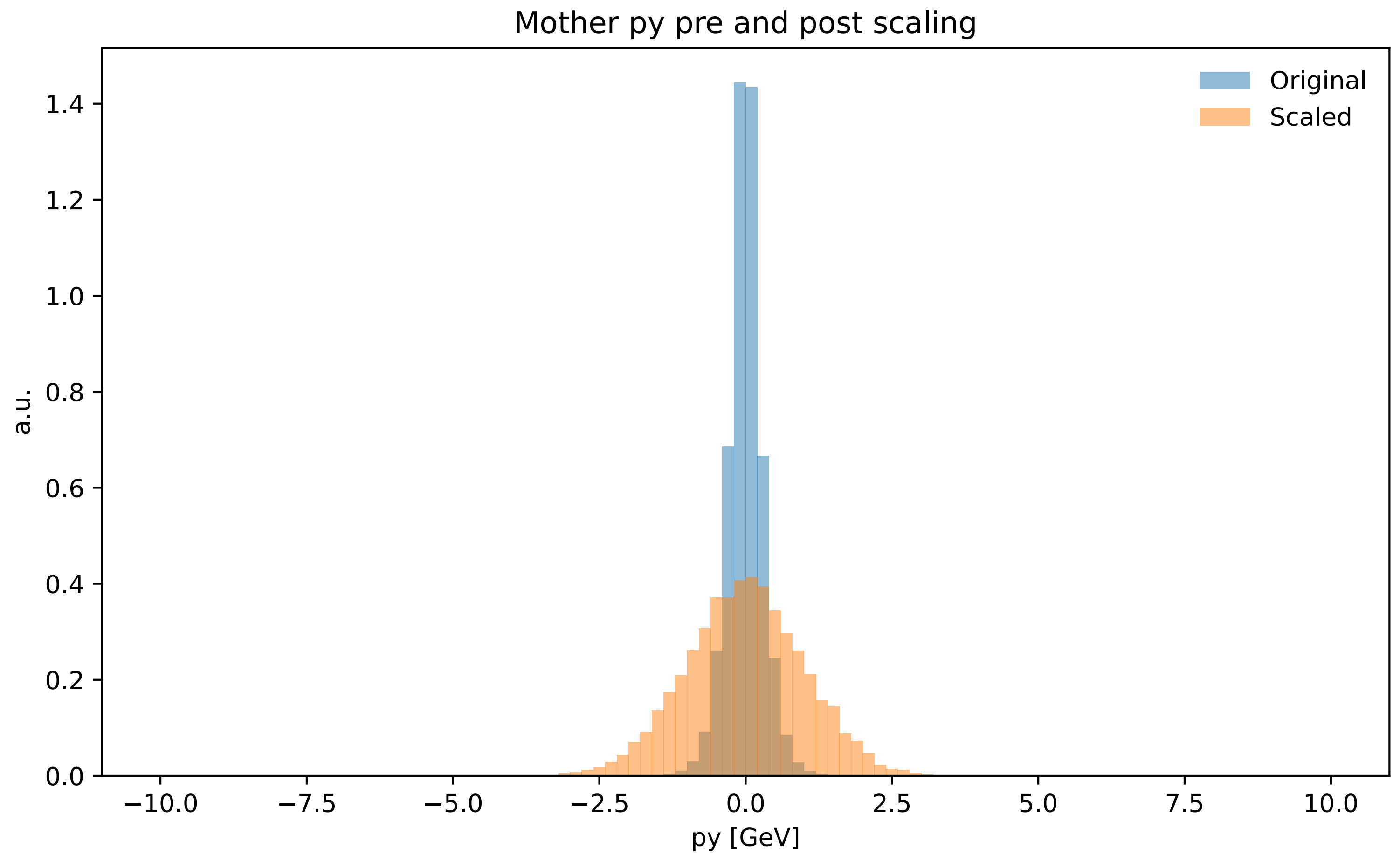


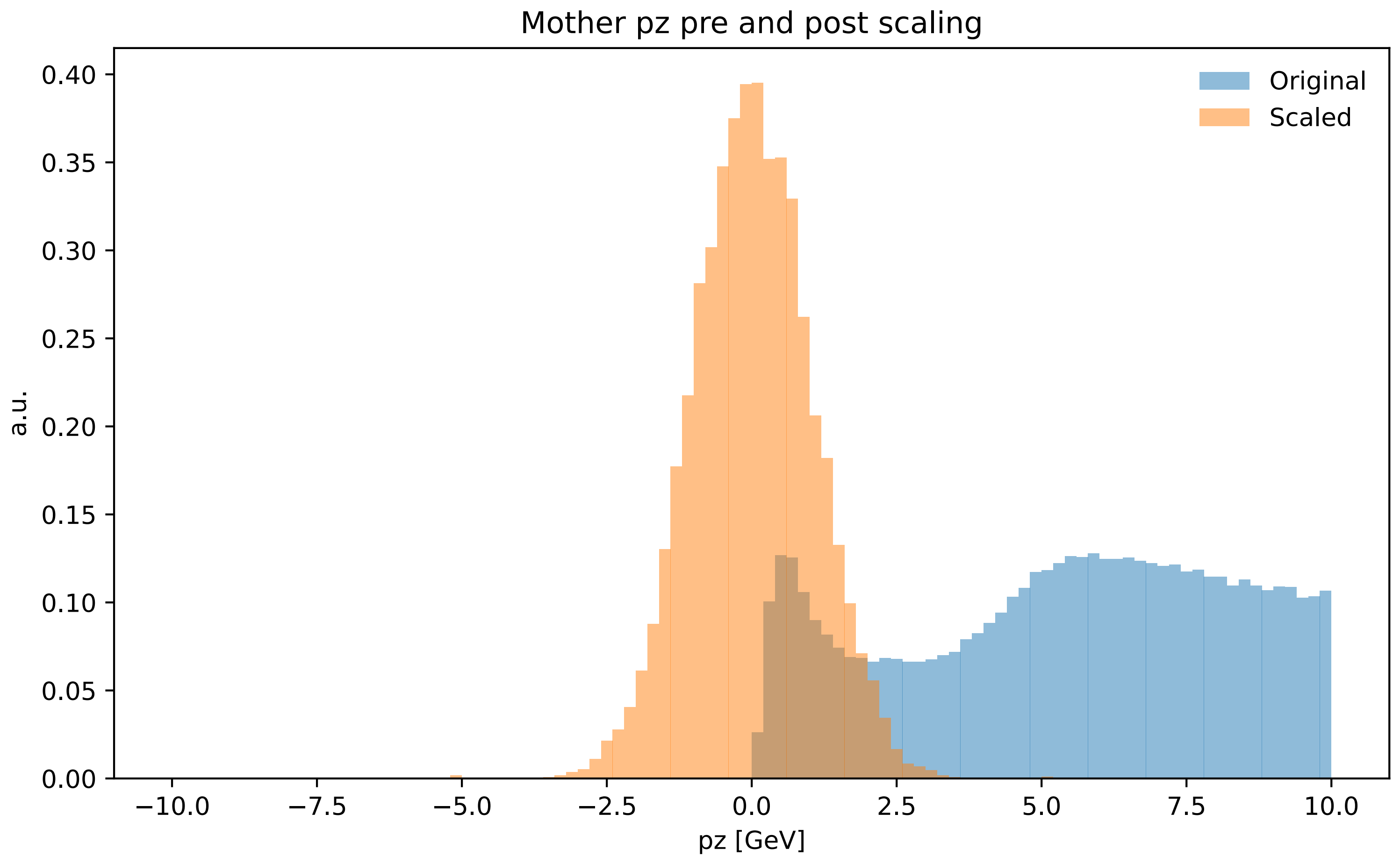
Coupling Normalising Flows: 1 Layer

- Split the input vector h in two parts, $[h_1, h_2]$
- Keep $h_1 = h'_1$ fixed and compute parameters of transformation $\phi(h_1)$
- Apply an *invertible* transformation on h_2, based on $\phi(h_1)$, $g(h_2, \phi(h_1))$
- Shuffle the elements before the next layer to ensure all the elements are transformed

Input Features Preprocessing







Angular Variables Generation