Vassily Kandinsky, *Several Circles*, 1926

ICHEP 2024 PRAGUE

Flavour Tagging with Graph Neural Network at ATLAS

19th July 2024

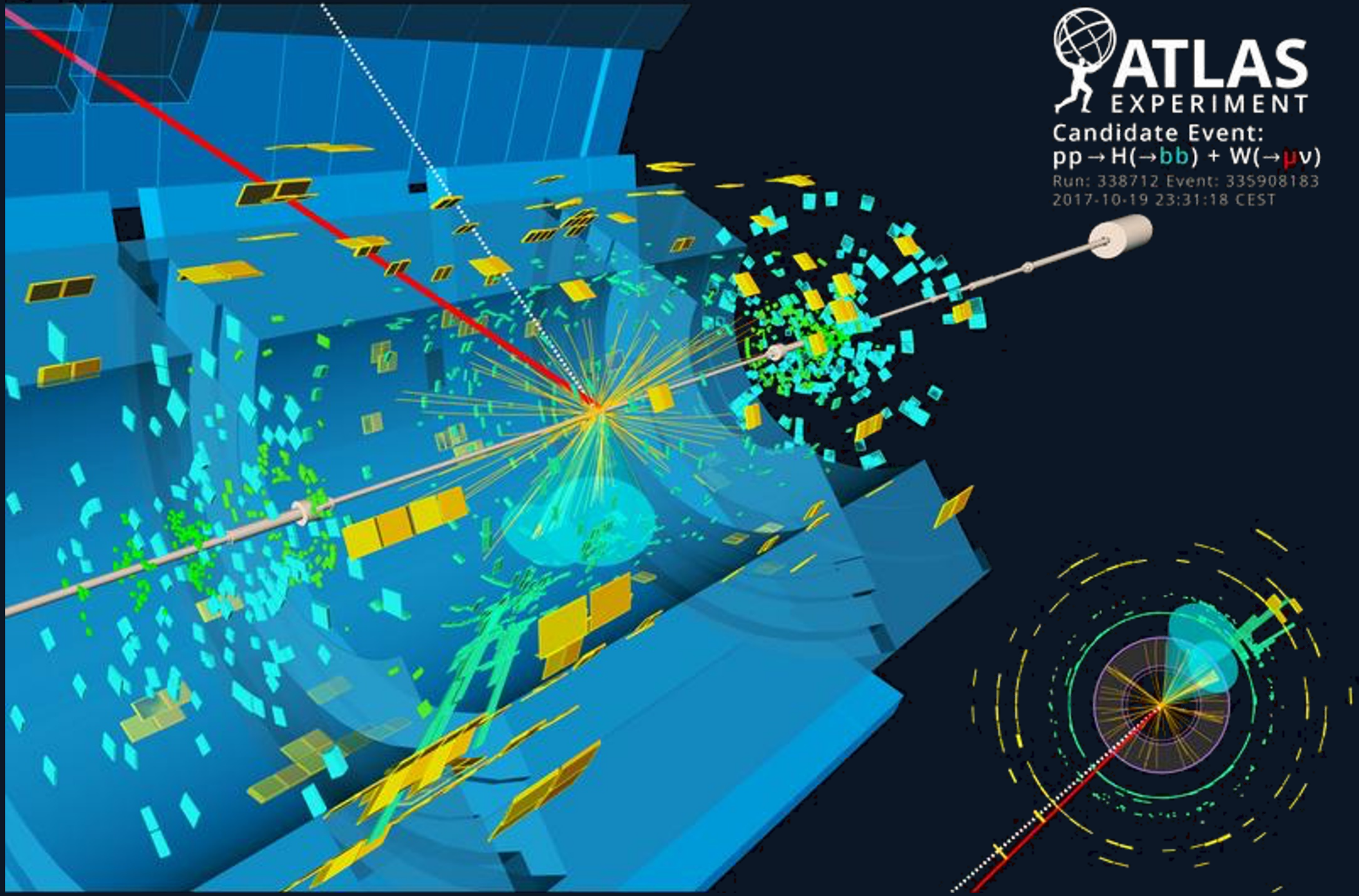
Author	Maxence DRAGUET
Supervisor	Daniela BORTOLETTO

On behalf of the ATLAS Experiment

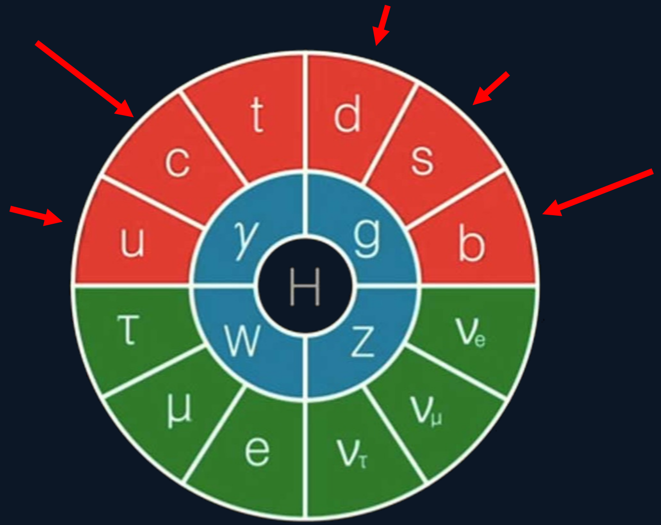
maxence.draguet@physics.ox.ac.uk

INTRODUCTION

ATLAS event display of a Higgs boson decaying to two b-quarks with an associated W boson decaying into a muon and a neutrino



ATLAS Collaboration relies on heavy-flavour jets classifiers



Used in many analyses:

$H \to b\bar{b}$, $H \to c\bar{c}$, di-Higgs, …

These classifiers are **ML TAGGERS**

Example: $VH \to b\bar{b}$ / $c\bar{c}$ presented this morning

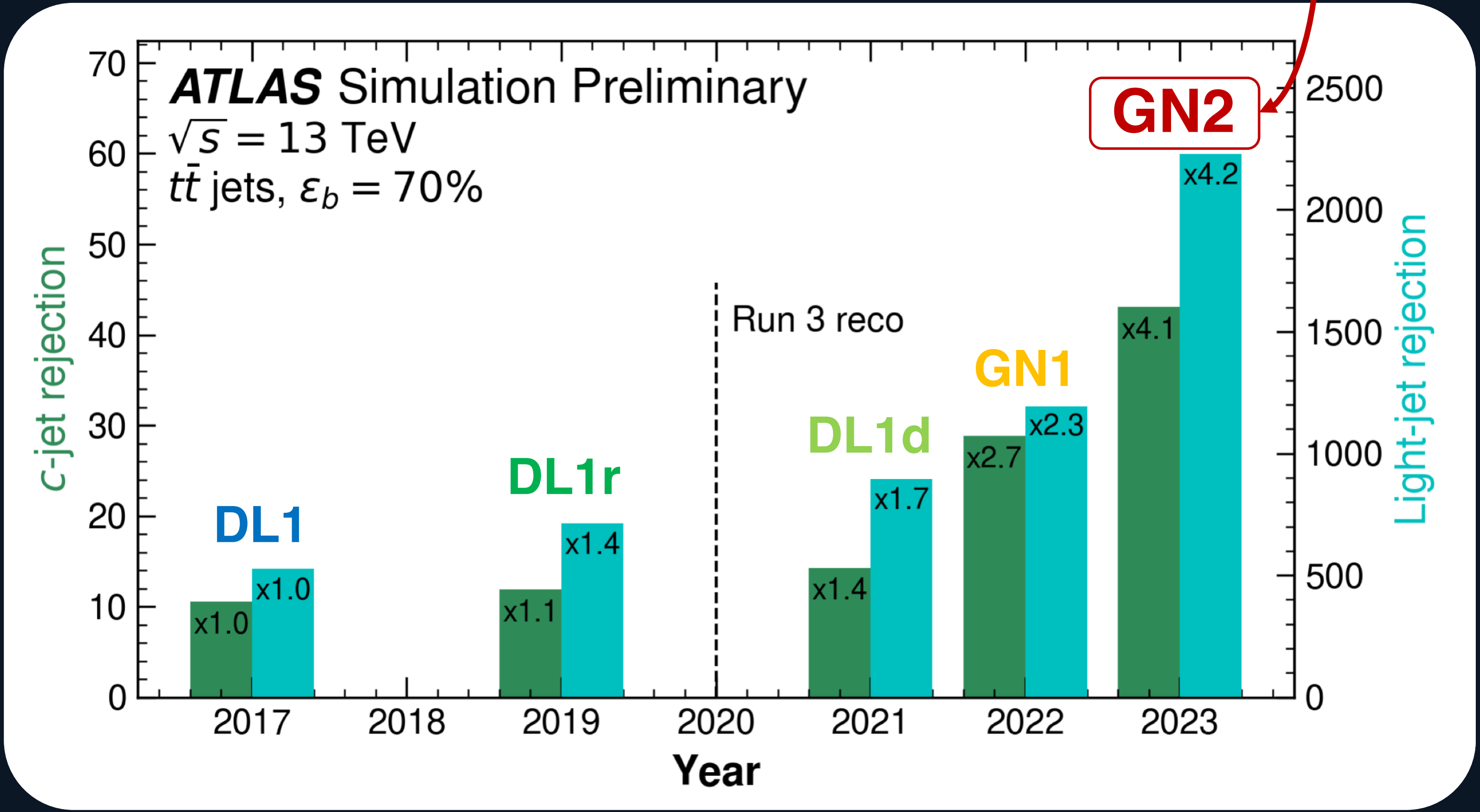
JET FLAVOUR TAGGING

CONTINUOUS EVOLUTION

→ BDT → Deep Network → RNN → DeepSet → Graph Attention → Transformer

Mission

Continuously improve the performance of the ATLAS tagger for {*b*, *c*, *light**, τ**} jet discrimination



rejection = 1 / miss-classification efficiency

* *light = u, d, s, gluon*
** *hadronic τ decays resemble c-jets*

JET FLAVOUR TAGGING

CONTINUOUS EVOLUTION

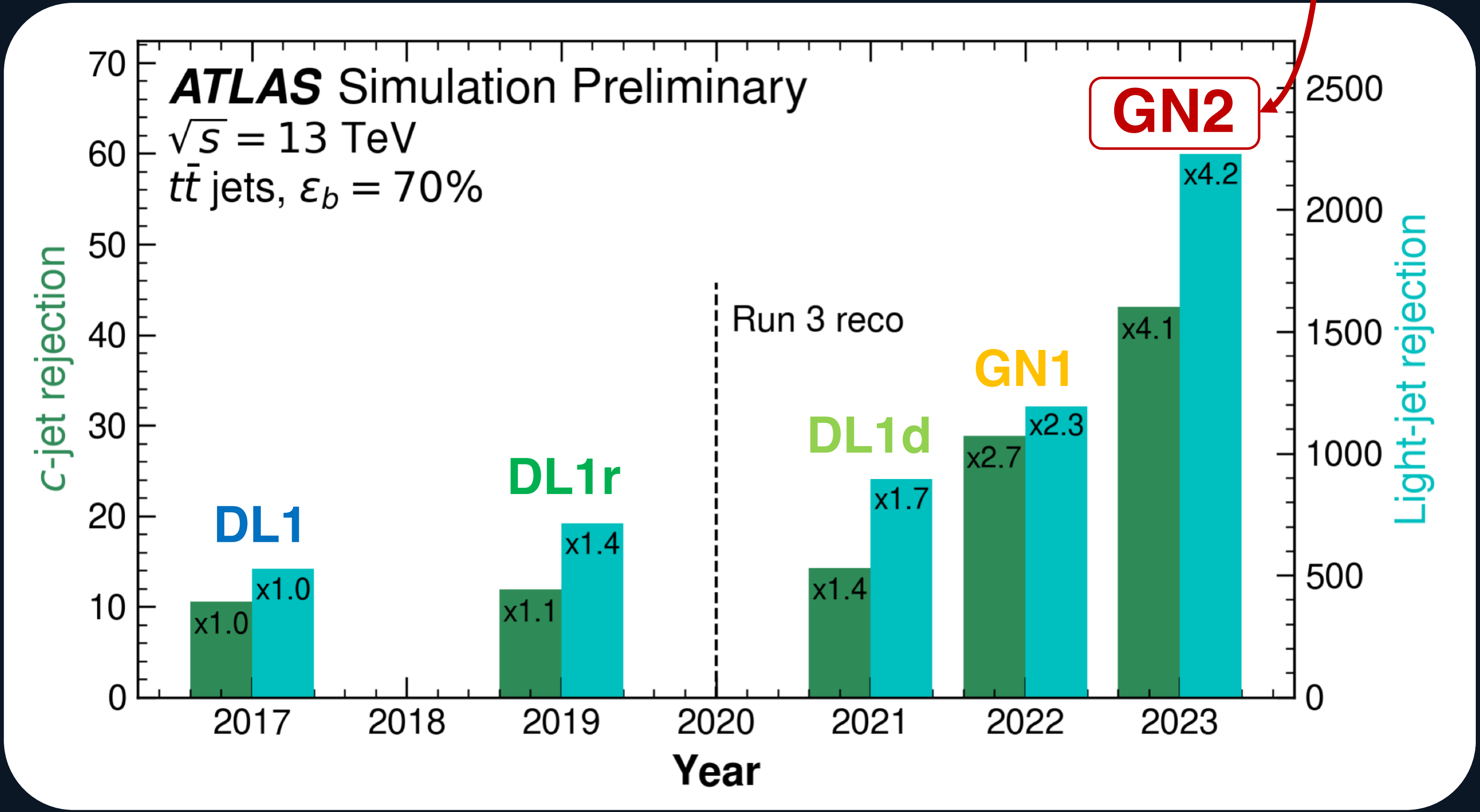
→ BDT → Deep Network → RNN → DeepSet → Graph Attention → Transformer

Inputs
Tracks & jet

Outputs
Per-flavour* probabilities & discriminant

Trained on simulated data

Calibrated on real data



*{b, c, light, τ}

rejection = 1 / miss-classification efficiency

JET FLAVOUR TAGGING

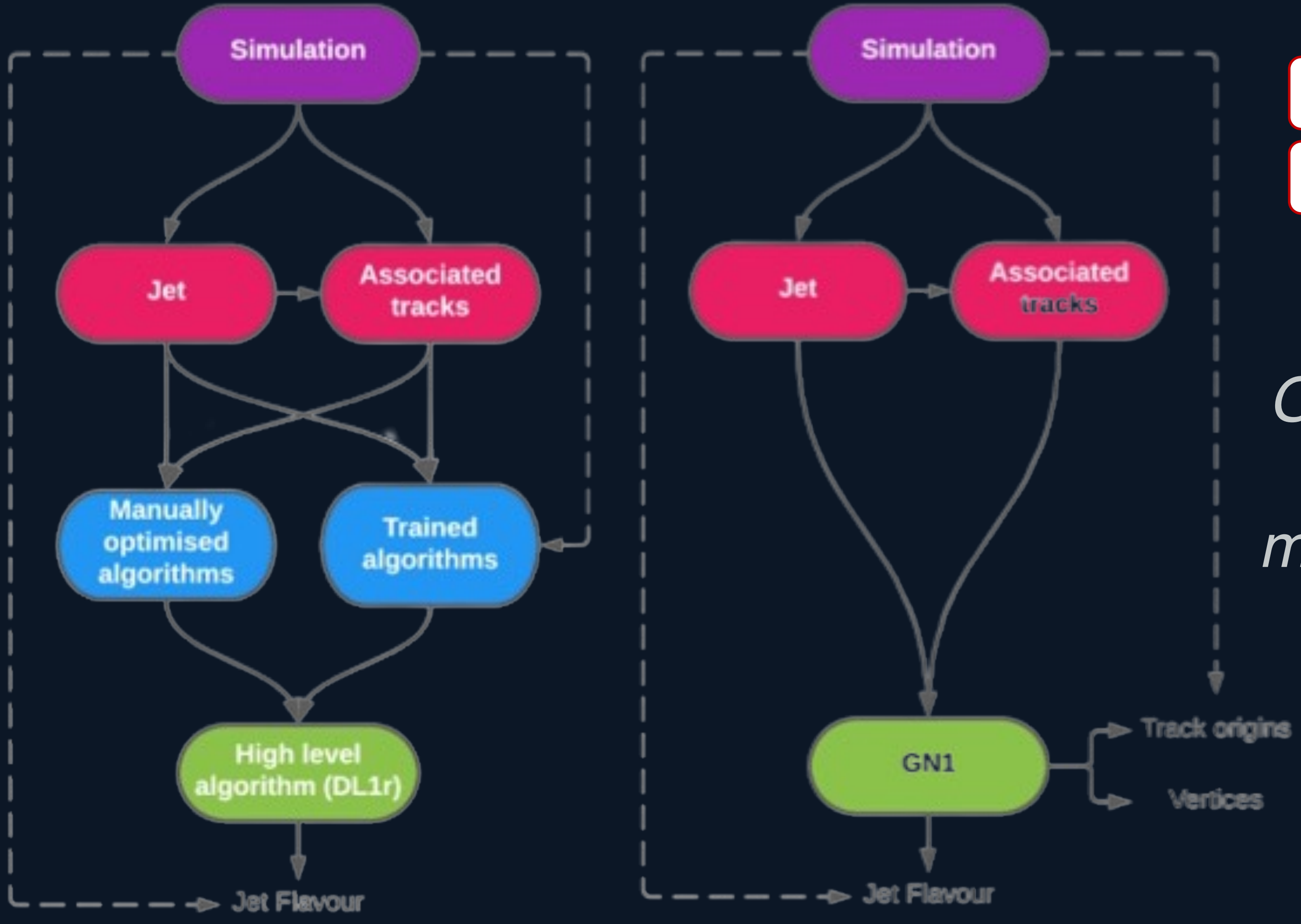
New design adopted by ATLAS for GN1 & GN2

DL1d

DL1r

Hierarchical

Many small submodels, cumbersome to maintain, limited use of Deep Learning



GN2

GN1

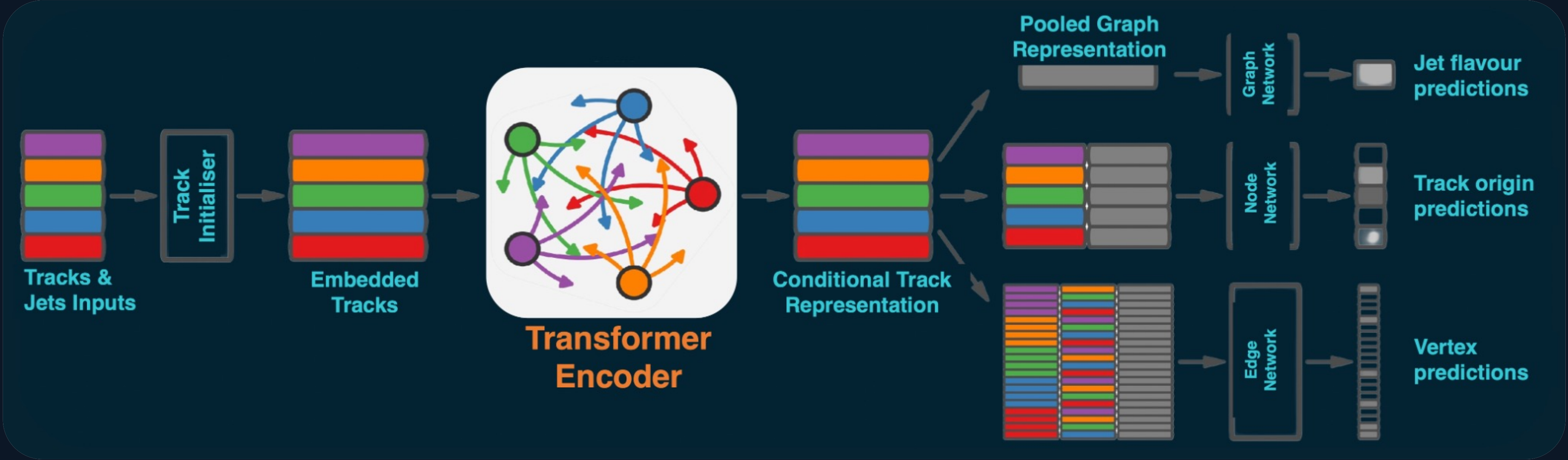
Integrated

One large network to rule them all, multitask, multimodal, agile and easy to update, fully leveraging Deep Learning

GN2

Large Multimodal Multitask Transformer Model

ATLAS GNN



Salt

Multimodal

Combines multiple physics input types

Architecture

Single network with SOTA performance

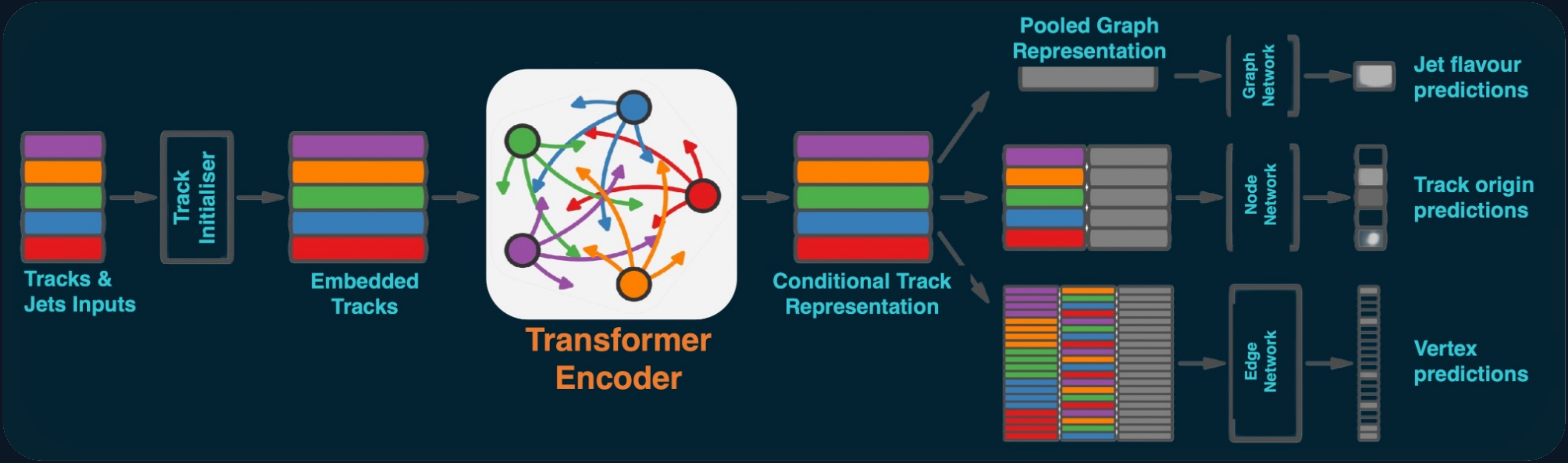
Multitask

Per-flavour probabilities + auxiliary objectives

GN2

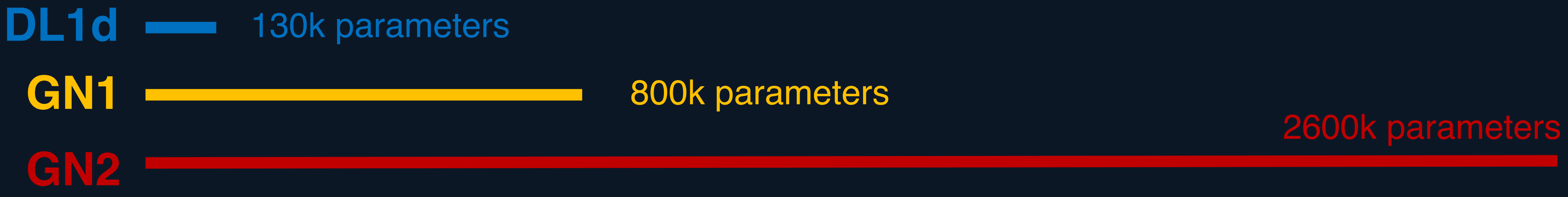
Large Multimodal Multitask Transformer Model

ATLAS GNN



Salt

“Large”



GN2 *Inputs*

Multimodal

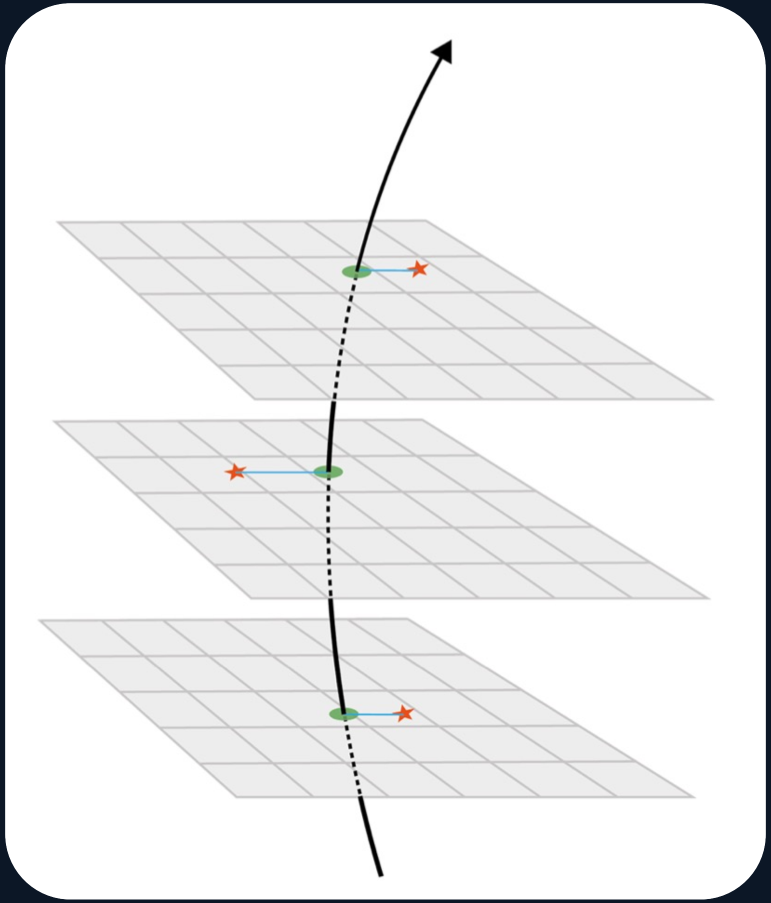
Tracks + Jet Variables

Tracks

- Track parameters
- Uncertainties
- Impact parameters

Jet

- p_T & η
- Resampled / flavour



192,000,000 simulated jets for training

↑

Thanks to new pre-processing software

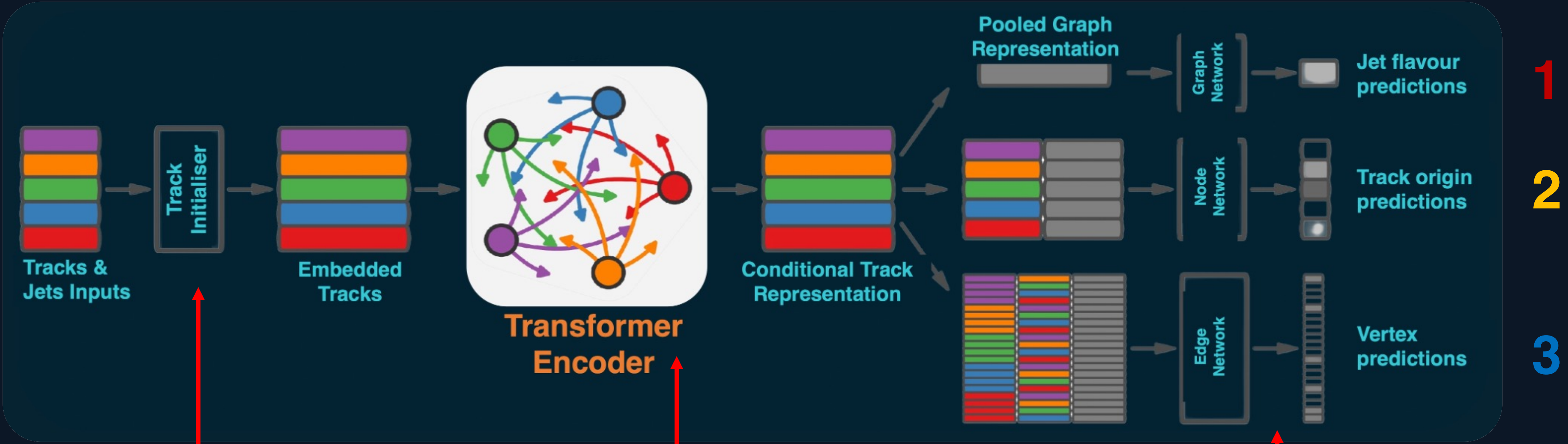
+

Stabilising architecture choices:

Layer Normalisation

Dropout

GN2 *Architecture*

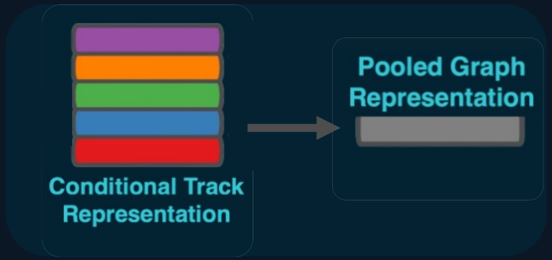


Neural network for combined representations embedding

Transformer Encoder for conditional track representation

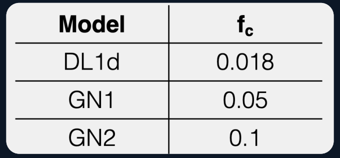
Auxiliary tasks for expert knowledge distillation

GN2 *Loss*

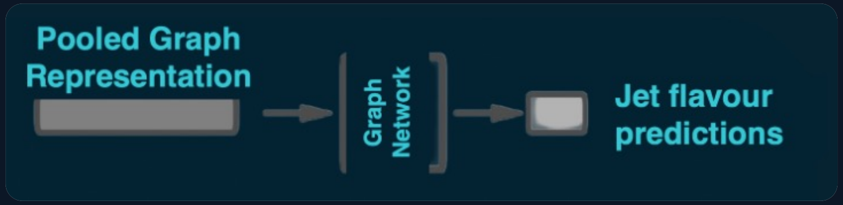


Global Attention Pooling

$$\mathcal{L}_{Total} = \mathcal{L}_{Jet} + \alpha\mathcal{L}_{Track} + \beta\mathcal{L}_{Vertex}$$

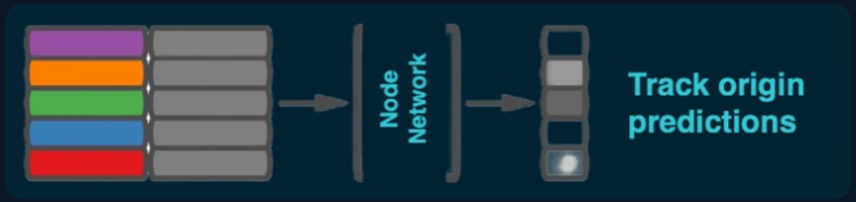


Model	f_c
DL1d	0.018
GN1	0.05
GN2	0.1

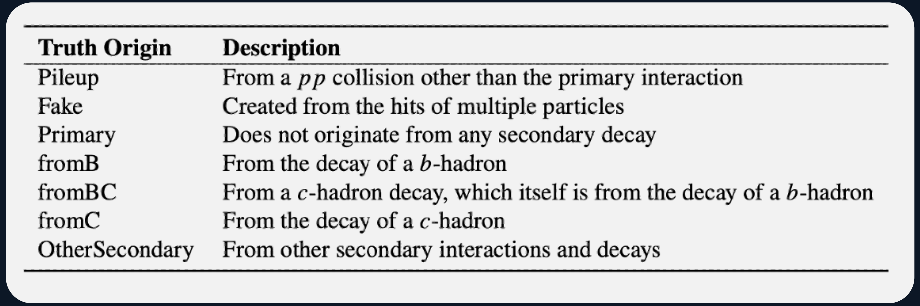


1 Predict jet flavour probabilities + tagging discriminant

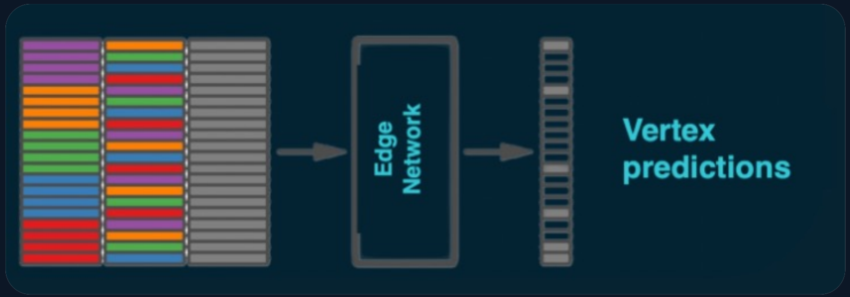
$$D_b = \frac{p_b}{f_c p_c + (1 - f_c)p_{light}}$$



2 Predict origin process of each track



Truth Origin	Description
Pileup	From a *pp* collision other than the primary interaction
Fake	Created from the hits of multiple particles
Primary	Does not originate from any secondary decay
fromB	From the decay of a *b*-hadron
fromBC	From a *c*-hadron decay, which itself is from the decay of a *b*-hadron
fromC	From the decay of a *c*-hadron
OtherSecondary	From other secondary interactions and decays



3 Predict track-pairs vertex compatibility

GN2 *Performance*

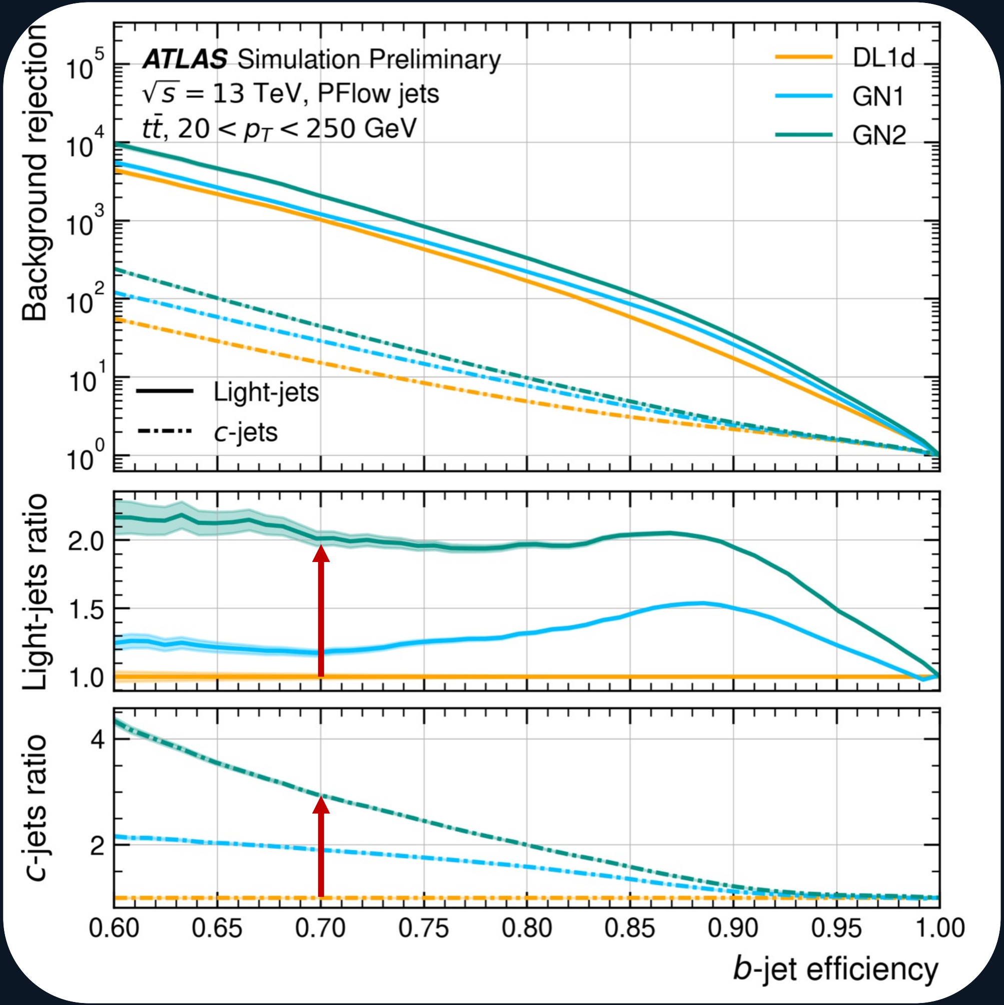
b-tagging

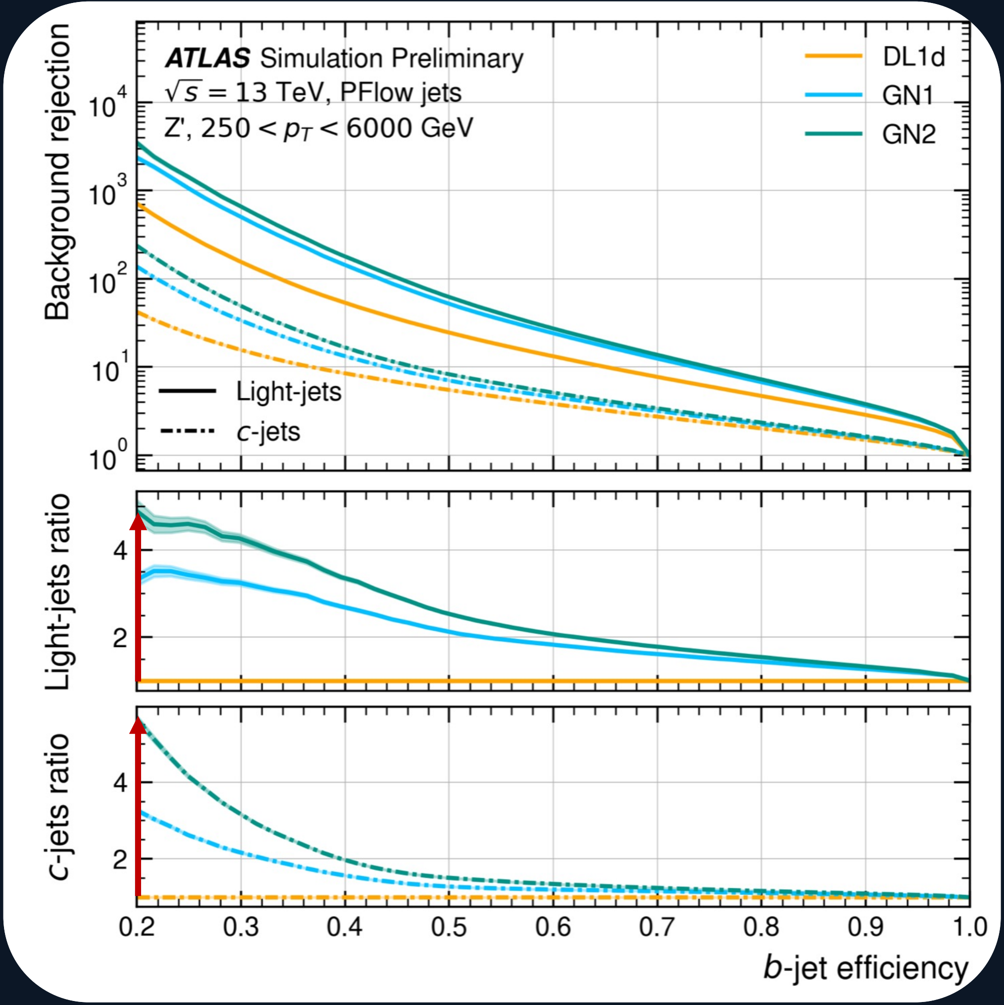
Significant improvement with GN2

Low p_T $(t\bar{t})$ 70* b-eff

Light-rej x2

c-rej x2.8





High p_T (Z') 20* b-eff

Light-rej x4.8

c-rej x5.5

Left working point corresponds to right one

GN2 *Performance*

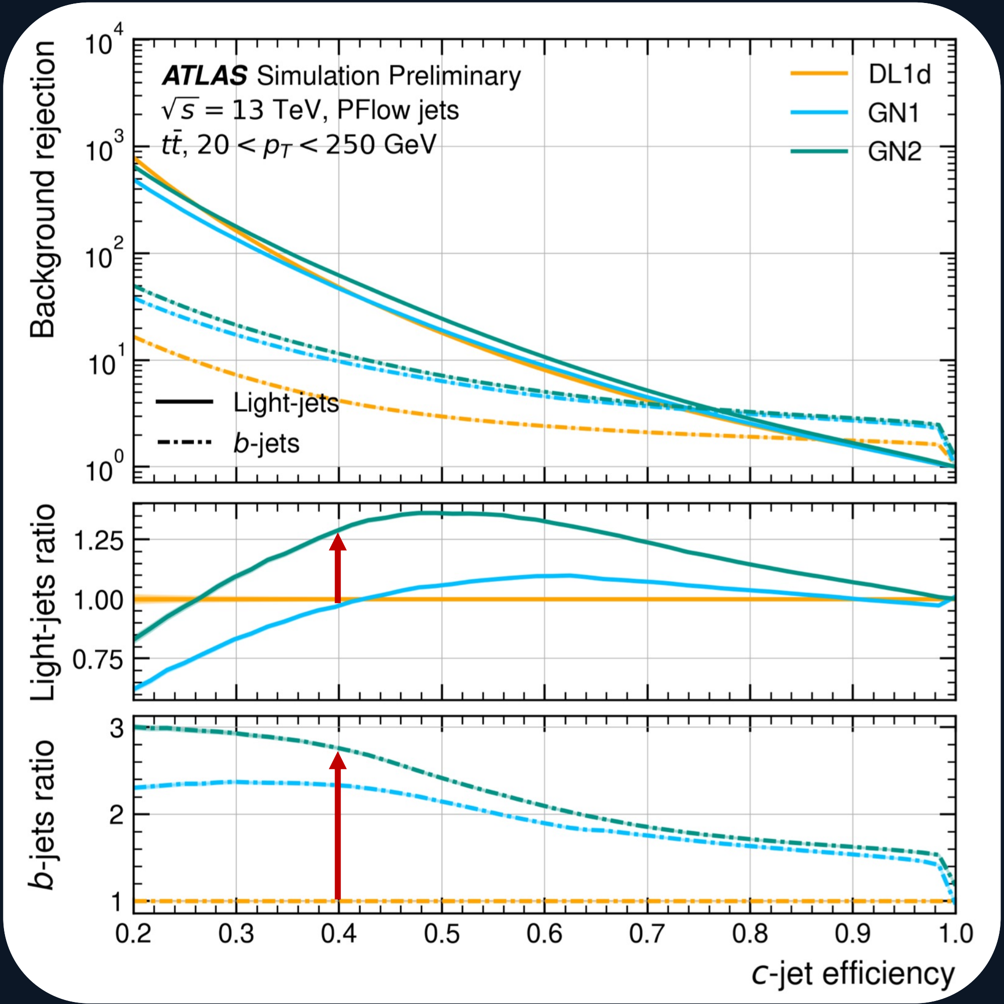
c-tagging

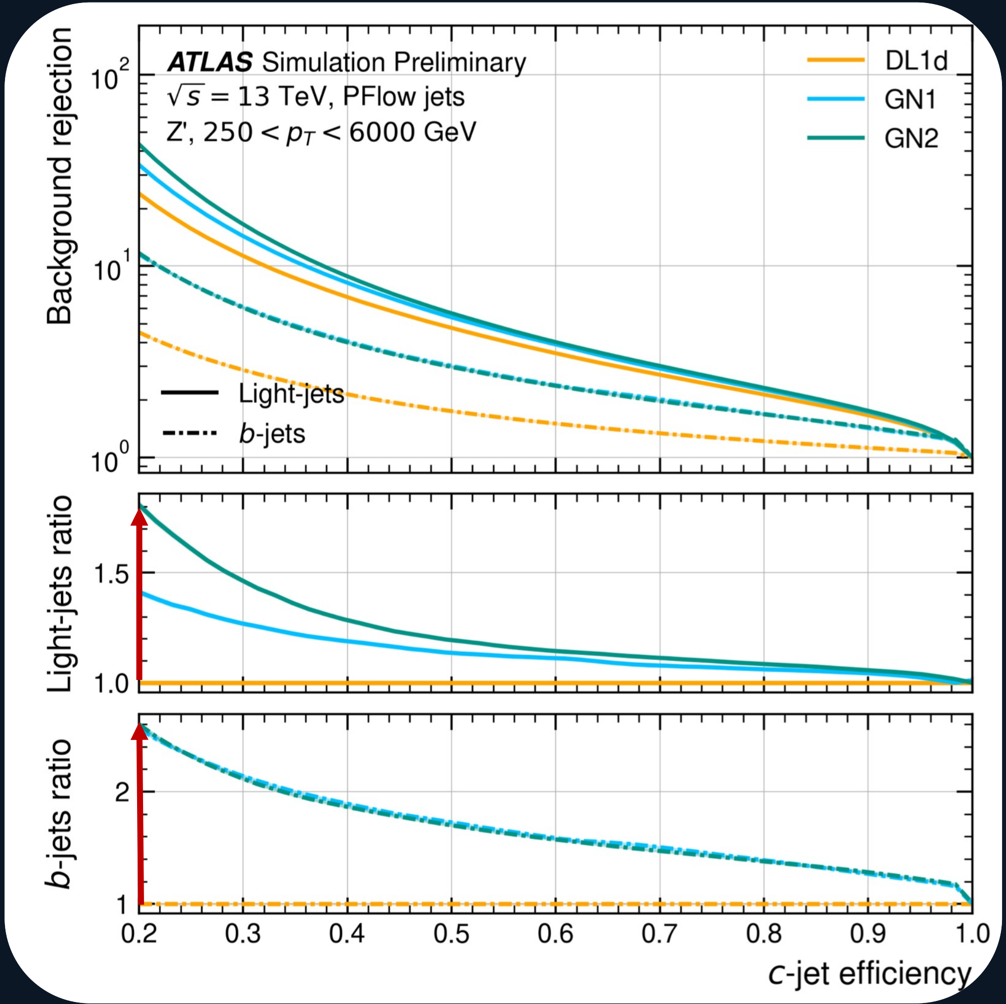
Significant improvement with GNN2

Low p_T ($t\bar{t}$) 40* c-eff

Light-rej x1.3

c-rej x2.7





High p_T (Z') 20* c-eff

Light-rej x1.8

c-rej x2.6

Left working point corresponds to right one

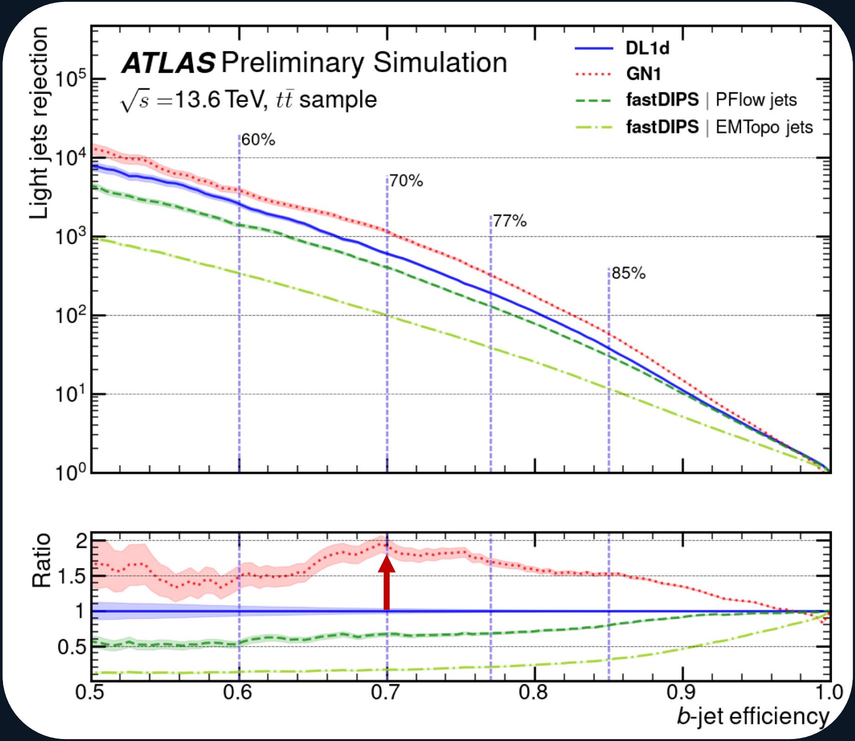
GN2 *Aftermath*

GN models bring a remarkable improvement to ATLAS

Quickly Proliferating

Inference time (ms)

	$t\bar{t}$	Z'
DL1d	0.07	0.08
GN1	0.40	0.78



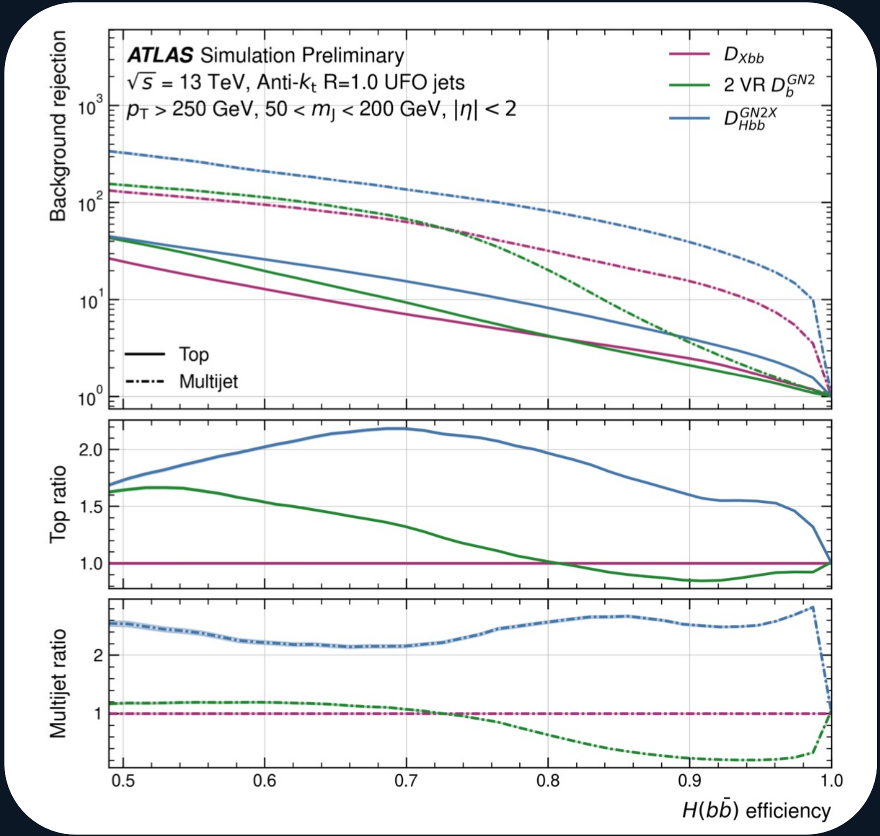
TRIGGERS

Boosted X → $b\bar{b}$ / $c\bar{c}$

For Higgs tagging

Top-rej

Multi-jet-rej



GN2X

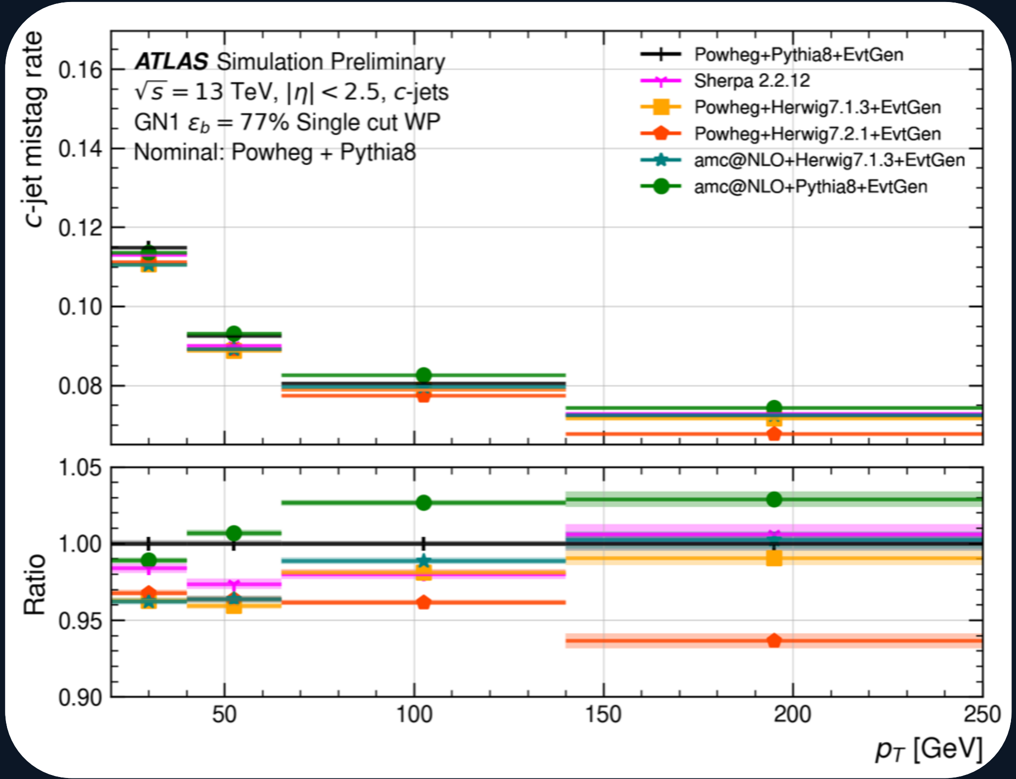
+ HL-LHC forecast, application to other part of the Collaboration, …

GN2 *Aftermath*

GN models bring a remarkable improvement to ATLAS

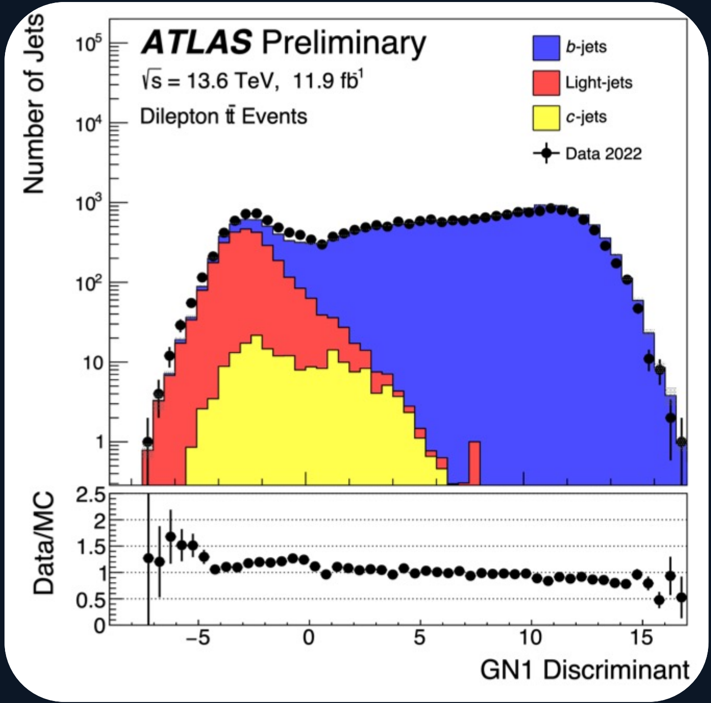
Ongoing Calibration

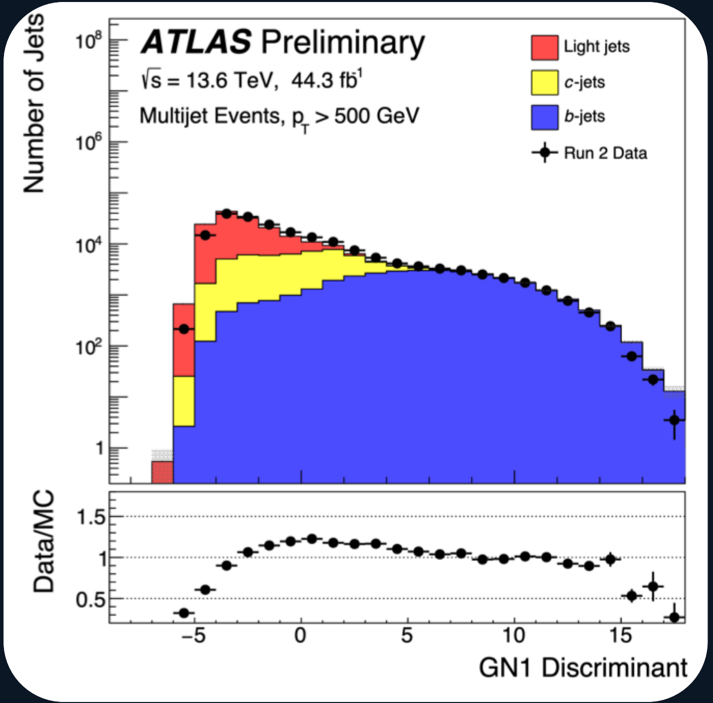
MC Dependence



✓ Overall generator dependence ~ O(3-6%)

Data / MC Agreement





✓ Good agreement in the bulk

GN2 *HPO*

GN models bring a remarkable improvement to ATLAS

Ongoing Hyperparameters Tuning

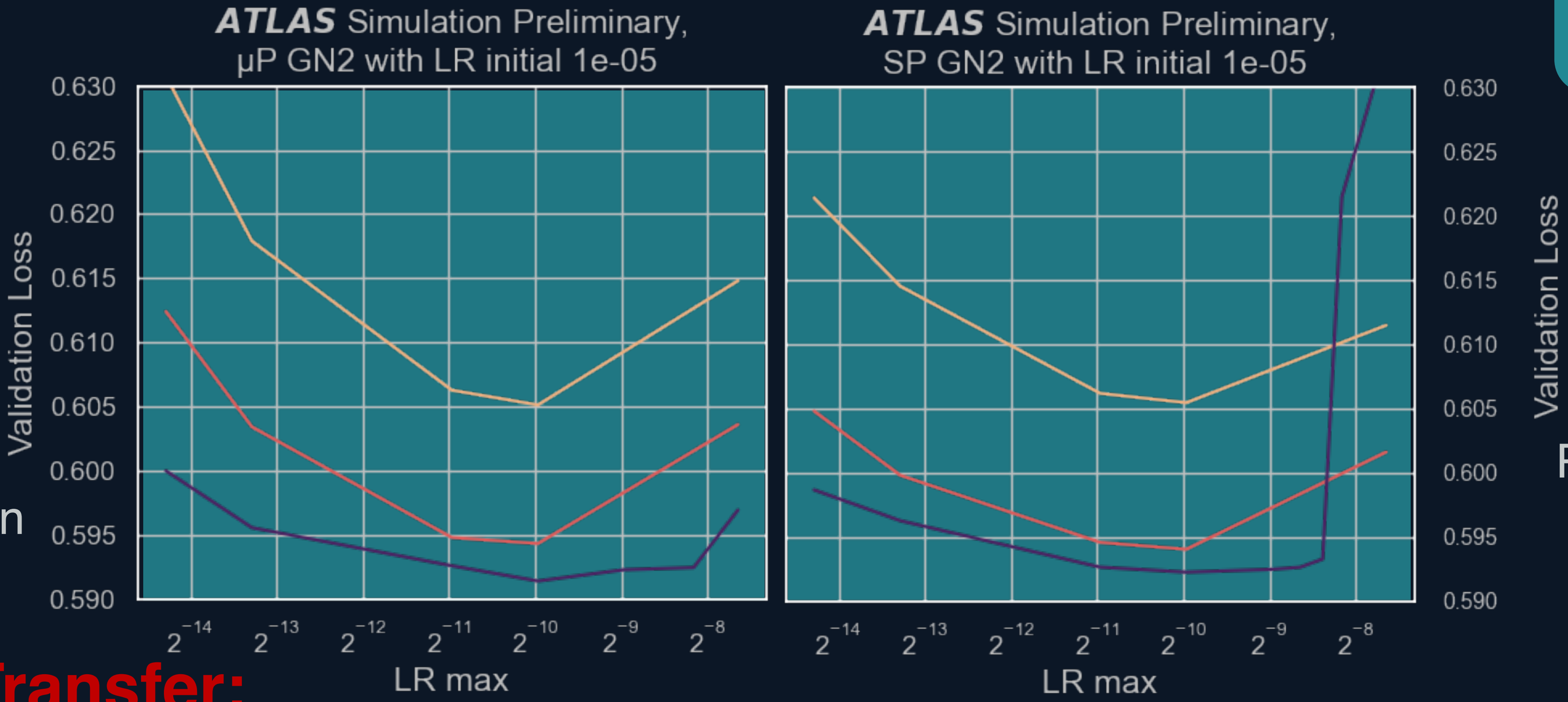
Embedding Width

- 64
- 128
- 256

µP

µP

Maximal Update Parametrisation



SP

Standard Parametrisation (LeCun)

µTransfer

µTransfer:

- ➢ HPO on the smaller model → faster training → better coverage
- ➢ Zero-shot transfer: best HP small model = best HP large model

GN2 *HPO*

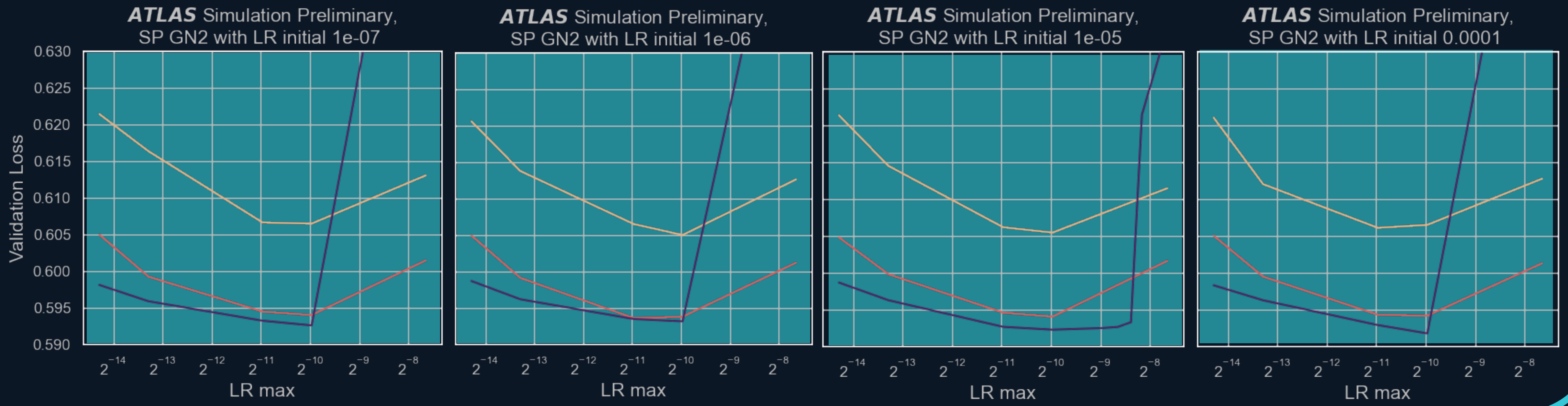
Real Deployment

LR Scheduler Optimisation: LR Max per LR Initial

Embedding Width

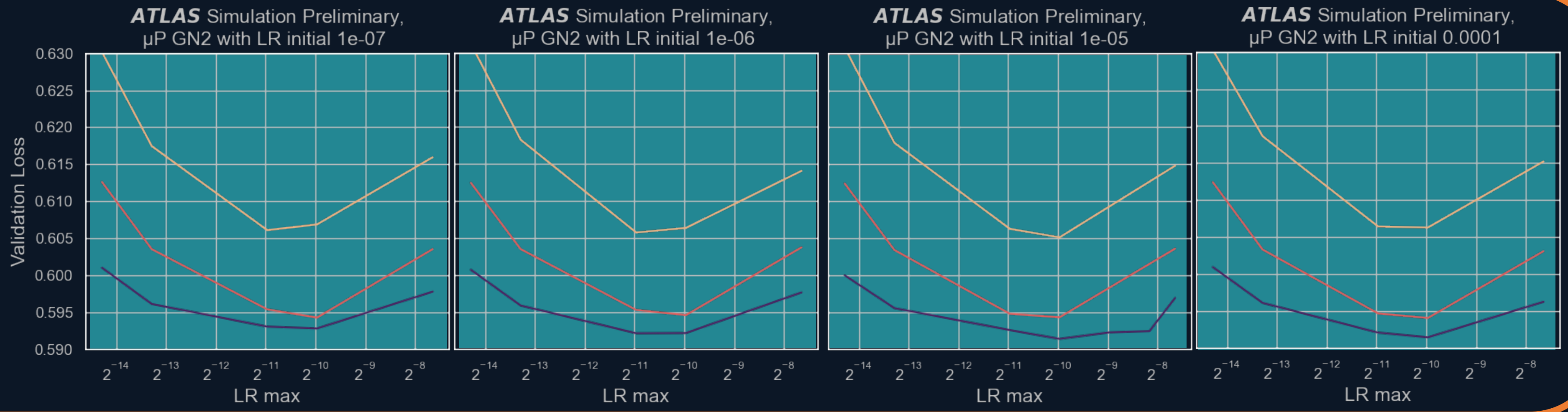
—	64
—	128
—	256

SP



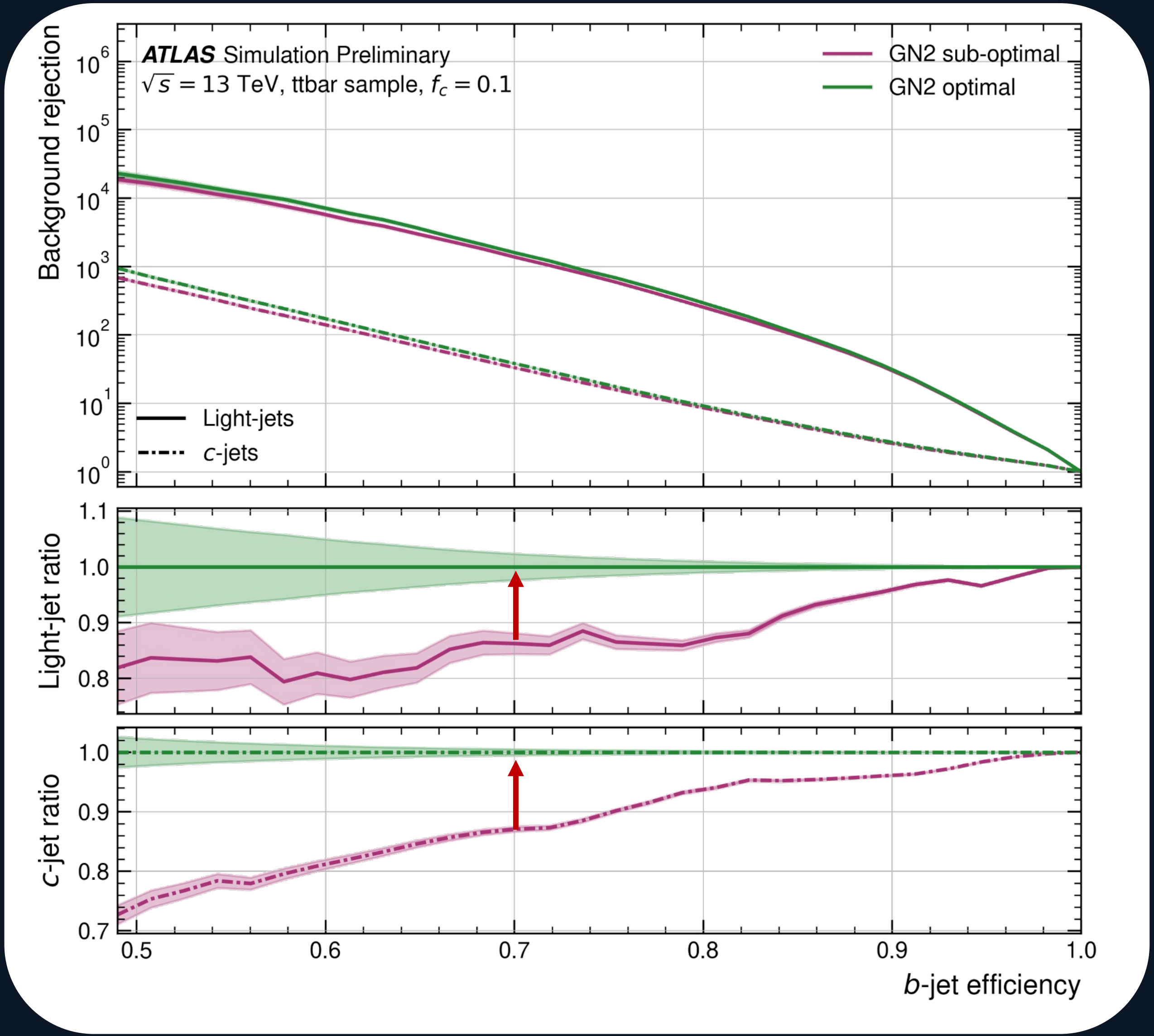
- Some SP trainings unstable
- No guarantee optimum shared across width

µP

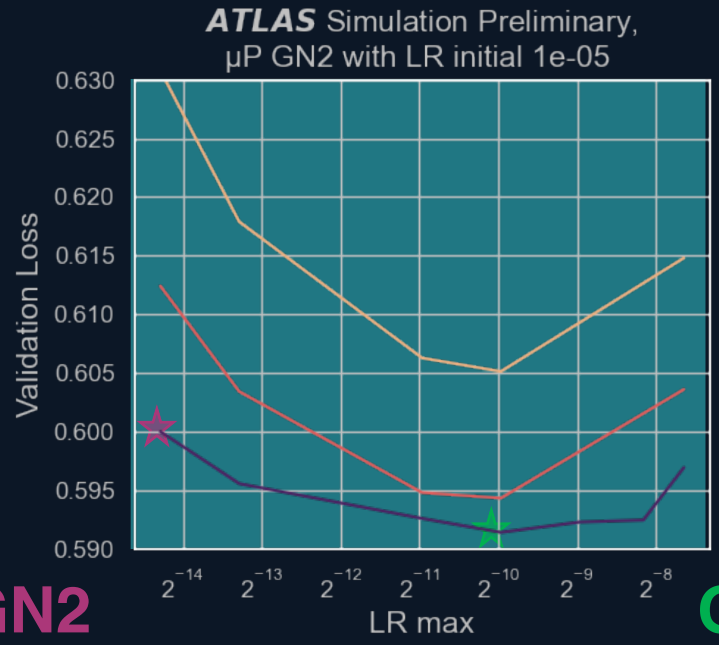


- ✓ µP trainings always stable
- ✓ Guaranteed shared optimum for sufficient width

HPO matters …



ROC curves on $t\bar{t}$



GN2 Sub-optimal

GN2 Optimal

Significant performance dependency on HP

$$D_b = \frac{p_b}{f_c p_c + (1 - f_c) p_{light}}$$

Vassily Kandinsky, *Several Circles*, 1926

Thank you for your attention!

19th July 2024

Author	Maxence DRAGUET
Supervisor	Daniela BORTOLETTO

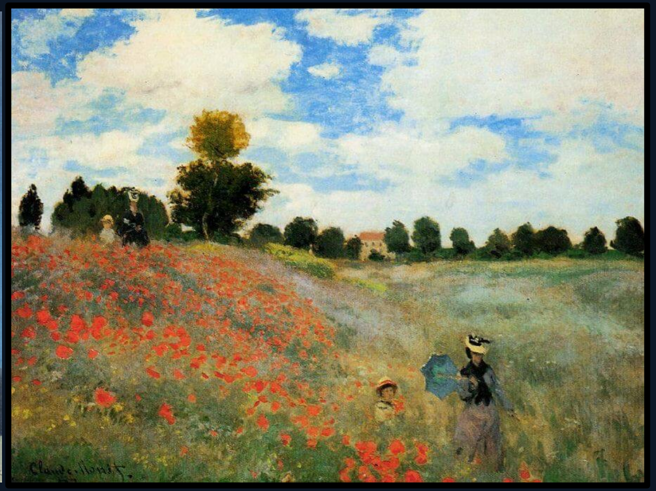
On behalf of the ATLAS Experiment

maxence.draguet@physics.ox.ac.uk

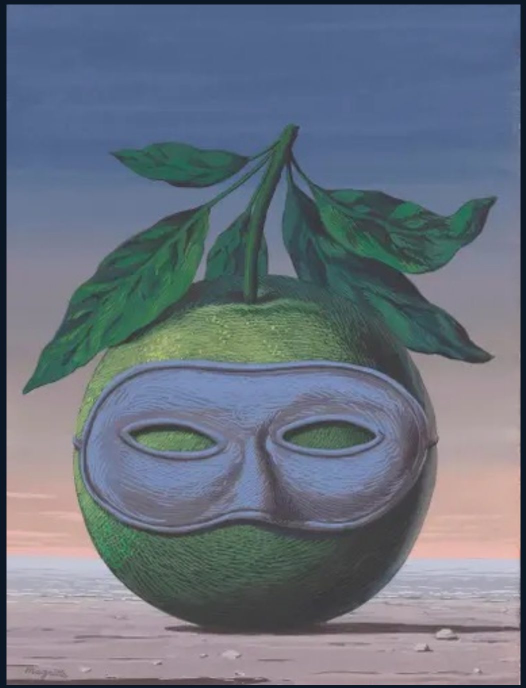
BACKUP

Vassily Kandinsky, *Several Circles*, 1926

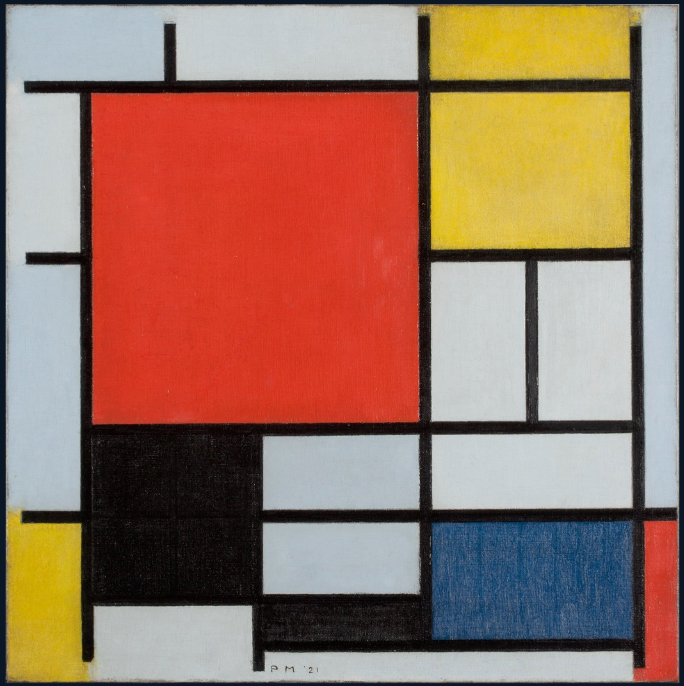
René Magritte — Le Retour



Monet — Les Coquelicots

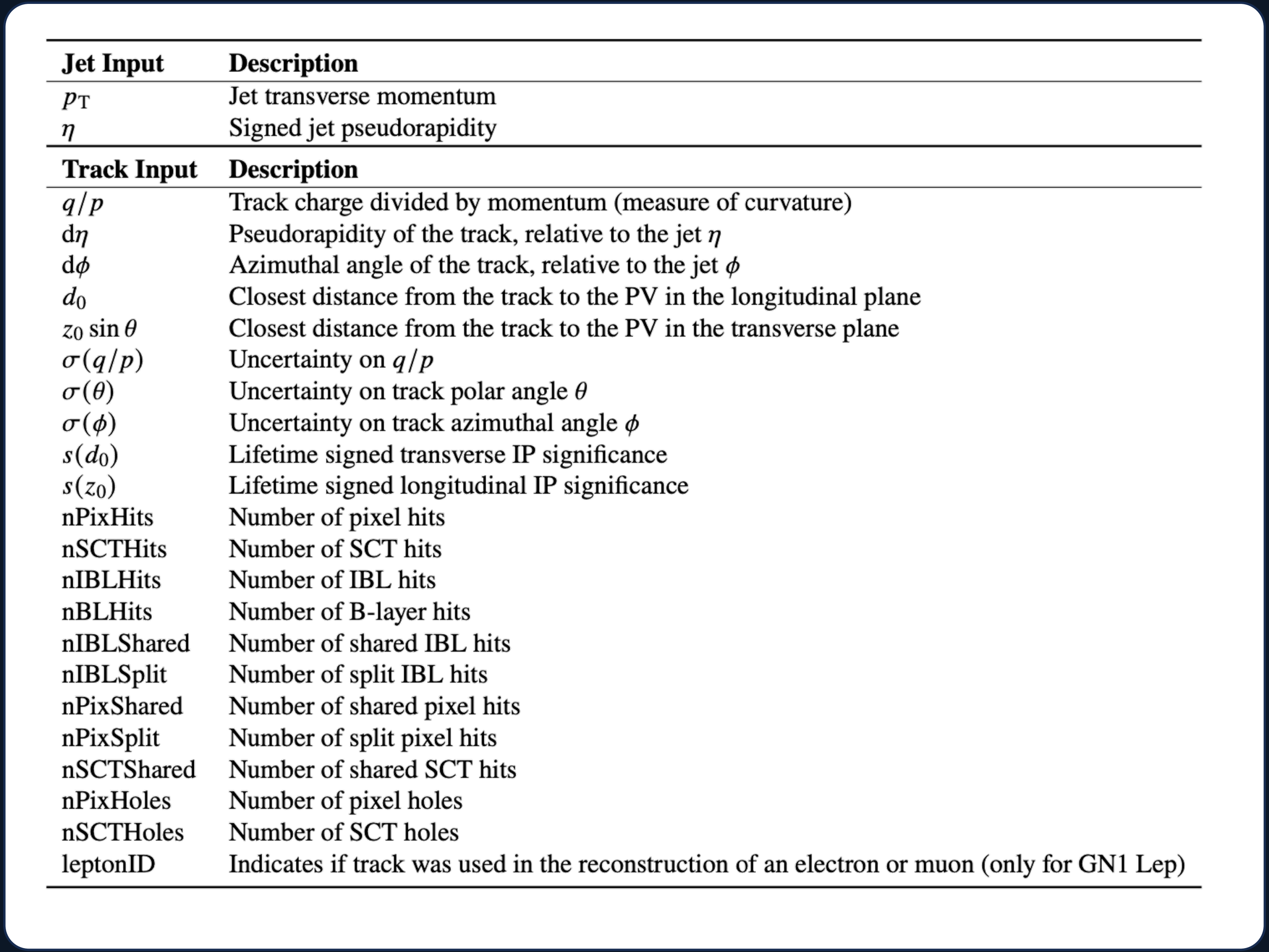


René Magritte — Souvenir de Voyage



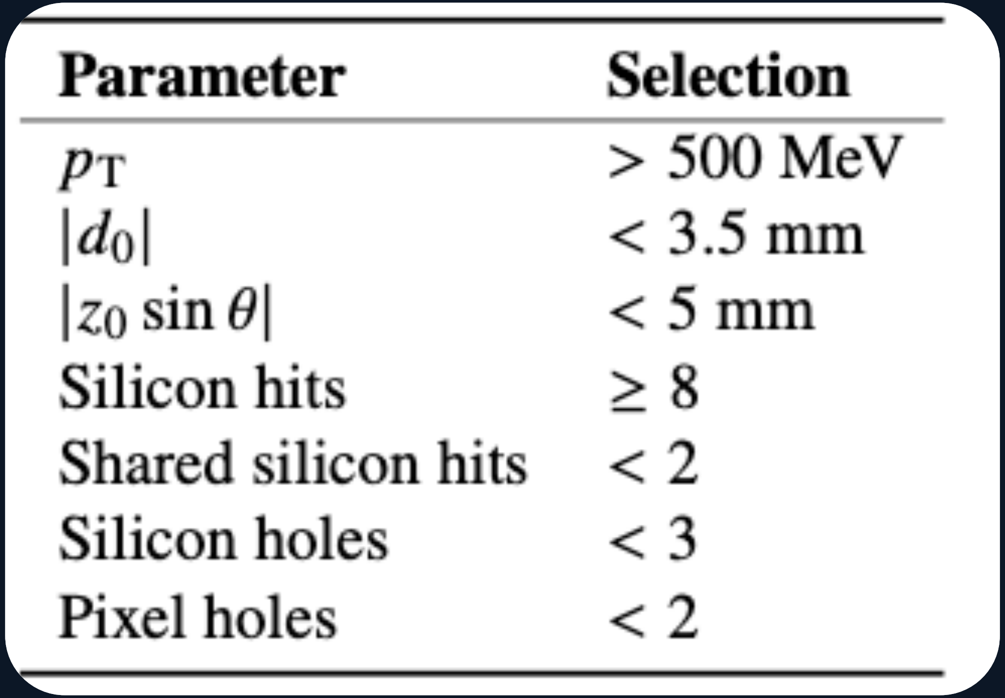
Piet Mondriaan
— Composition with Red, Yellow, and Blue

GN2 *Inputs*



Jet Input	**Description**
p_{T}	Jet transverse momentum
η	Signed jet pseudorapidity
Track Input	**Description**
q/p	Track charge divided by momentum (measure of curvature)
$\mathrm{d}\eta$	Pseudorapidity of the track, relative to the jet η
$\mathrm{d}\phi$	Azimuthal angle of the track, relative to the jet ϕ
d_0	Closest distance from the track to the PV in the longitudinal plane
$z_0 \sin\theta$	Closest distance from the track to the PV in the transverse plane
$\sigma(q/p)$	Uncertainty on q/p
$\sigma(\theta)$	Uncertainty on track polar angle θ
$\sigma(\phi)$	Uncertainty on track azimuthal angle ϕ
$s(d_0)$	Lifetime signed transverse IP significance
$s(z_0)$	Lifetime signed longitudinal IP significance
nPixHits	Number of pixel hits
nSCTHits	Number of SCT hits
nIBLHits	Number of IBL hits
nBLHits	Number of B-layer hits
nIBLShared	Number of shared IBL hits
nIBLSplit	Number of split IBL hits
nPixShared	Number of shared pixel hits
nPixSplit	Number of split pixel hits
nSCTShared	Number of shared SCT hits
nPixHoles	Number of pixel holes
nSCTHoles	Number of SCT holes
leptonID	Indicates if track was used in the reconstruction of an electron or muon (only for GN1 Lep)

GN2 *Track Selection*



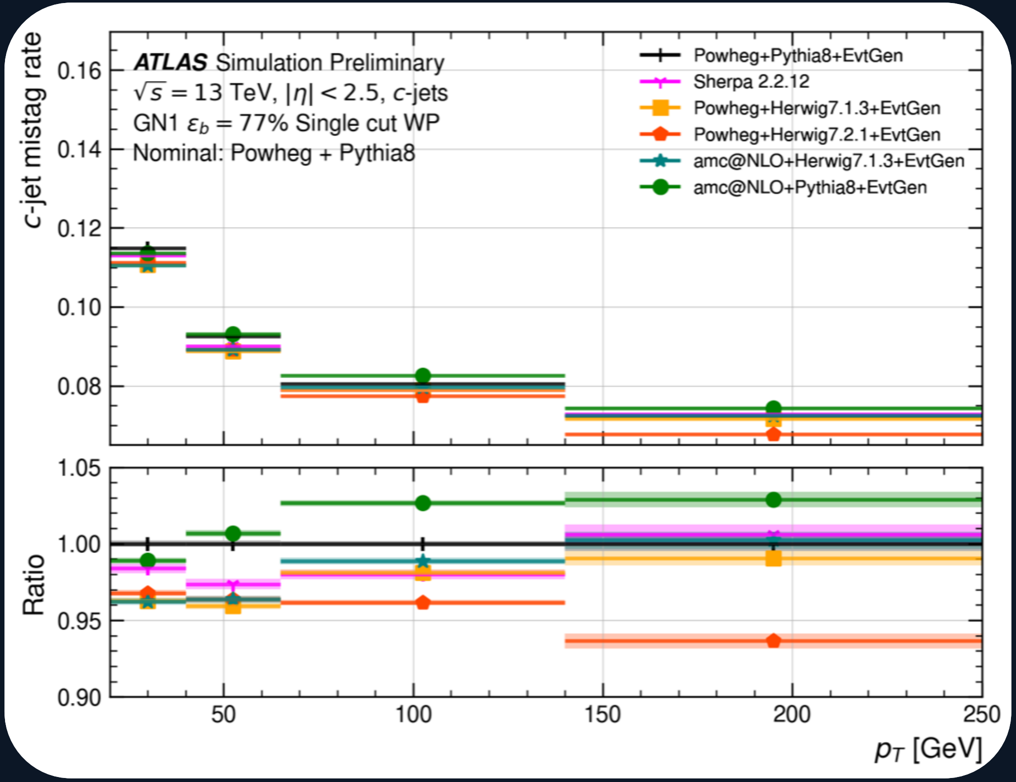
Parameter	Selection
p_T	> 500 MeV
$\|d_0\|$	< 3.5 mm
$\|z_0 \sin\theta\|$	< 5 mm
Silicon hits	≥ 8
Shared silicon hits	< 2
Silicon holes	< 3
Pixel holes	< 2

“Quality selections applied to tracks, where d_0 is the transverse IP of the track, z_0 is the longitudinal IP with respect to the PV and θ is the track polar angle. Shared hits are hits used on multiple tracks which have not been classified as split by the cluster-splitting neural networks. Shared hits on pixel layers are given a weight of 1, while shared hits in the SCT are given a weight of 0.5. A hole is a missing hit, where one is expected, on a layer between two other hits on a track”.

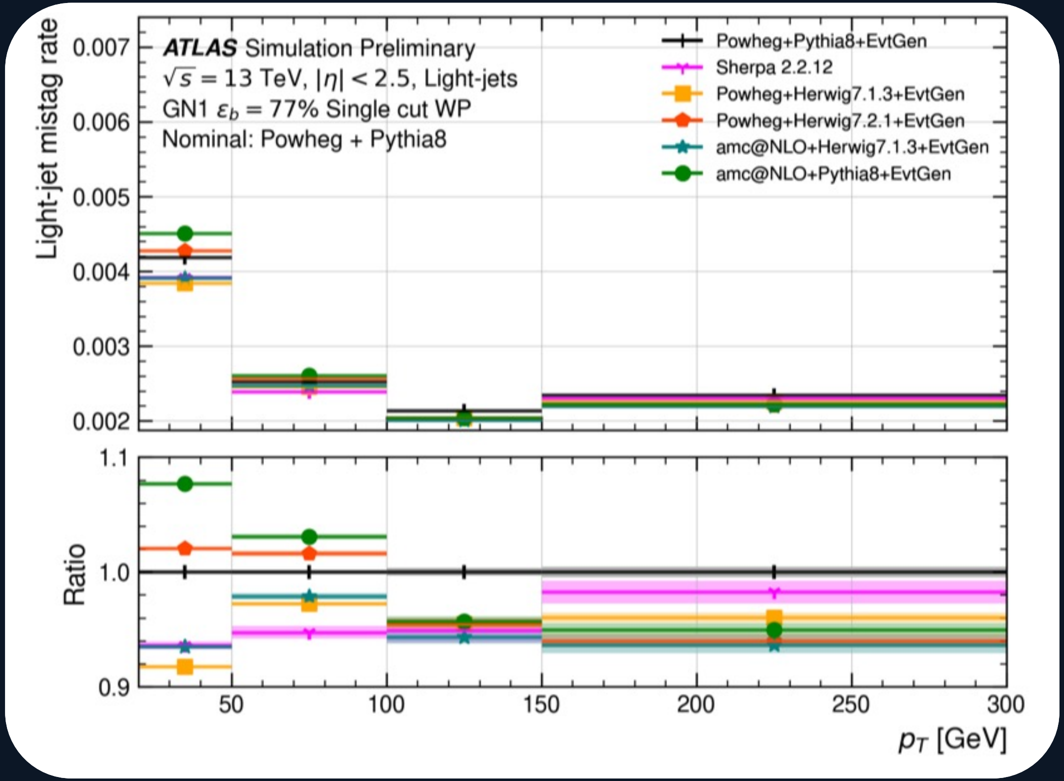
ATLAS GNN

GN2 *MC Dependency*

C-jets



Light-jets



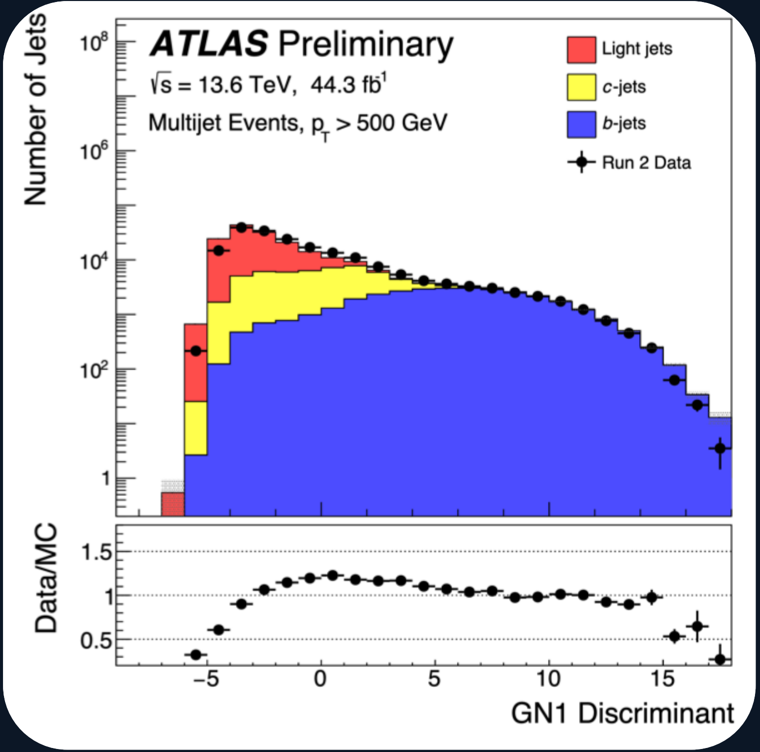
GN2 *Data / MC*

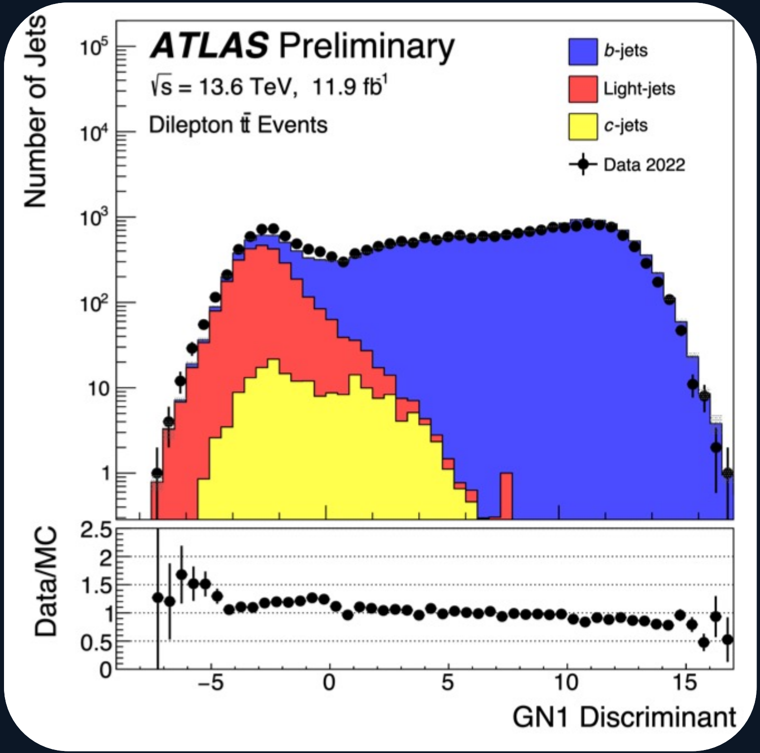
- ➢ "Tag and probe" event selection: tag jet must pass 85% b-eff
- ➢ Simulations are scaled to match the total yield in data

Tag jet pass 85% b-eff

$p_T > 200$ GeV

Event $p_T > 500$ GeV





- 2-lepton 2-jet selection, 1 lepton trigger
- Opposite sign electron and muon
- Invariant mass of each lepton-jet pair < 175 GeV
- Tagger discriminant for leading jet

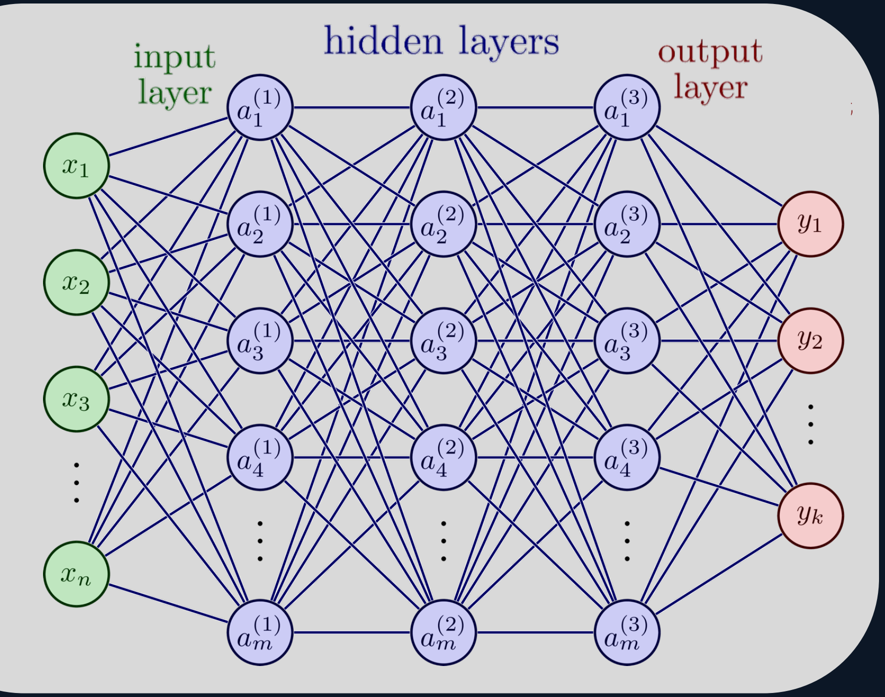
Maximal Update Parametrization

What if we could do HPO on a smaller model instead?

PARAMETRIZATION MATTERS!

µP

- Standard Parametrization* (SP)
- Maximal Update Parametrization (µP)



$d^{in}_{L_{in}} = n \quad d^{in}_{L_{hid}} = m \quad d^{in}_{L_{out}} = m$

INITIALISATION

Standard Parametrization* (SP)	Maximal Update Parametrization (µP)
$w^{L_{in}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{in}}}\right)$	$w^{L_{in}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{in}}}\right)$
$w^{L_{hid}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{hid}}}\right)$	$w^{L_{hid}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{hid}}}\right)$
$w^{L_{out}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{out}}}\right)$	$w^{L_{out}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{out}} \times d^{in}_{L_{out}}}\right)$
$b^{L_{...}} = 0$	$b^{L_{...}} = 0$

LR Adam

Standard Parametrization* (SP)	Maximal Update Parametrization (µP)
∀ weights η	$\eta_{W_{L_{in}}} = \eta$ $\eta_{W_{L_{hid}}} = \eta \,/\, d^{in}_{L_{hid}}$ $\eta_{W_{L_{out}}} = \eta \,/\, d^{in}_{L_{out}}$

*LeCun et al; 1998

With attention scale $1 / d^{in}$ instead $1 / \sqrt{d^{in}}$

Maximal Update Parametrization

“Effect of updates on activations becomes roughly independent of width”

μTransfer

“Each weight matrix is *maximally* updated without blowing up”

μP

*Sum of n squared Gaussians of variance 1 is a Chi distribution of degree n

$$f(x) = V^T U\, x \text{ with } x \in \mathbb{R}\ \&\ V^T, U \in \mathbb{R}^{n \times 1}$$

SP

$$V_i \sim \mathcal{N}\left(0, \frac{1}{n}\right) \qquad U_i \sim \mathcal{N}(0, 1)$$

$$V' \leftarrow V + \theta\, U \qquad U' \leftarrow U + \theta\, V$$

$$f(x) \leftarrow (V^T U + \theta U^T U + \theta V^T V + \theta^2 U^T V)x$$

By LLN* $\Theta(n)$ ... blows up

μP

$$V_i \sim \mathcal{N}\left(0, \frac{1}{n^2}\right) \qquad U_i \sim \mathcal{N}(0, 1)$$

$$V' \leftarrow V + \frac{\theta}{n} U \qquad U' \leftarrow U + \theta\, V$$

$$f(x) \leftarrow (V^T U + \frac{\theta}{n} U^T U + \theta V^T V + \frac{\theta^2}{n} U^T V)x$$

$\Theta(x)$ ... stable ✓

*Sum of n squared Gaussians of variance 1 is a Chi distribution of degree n

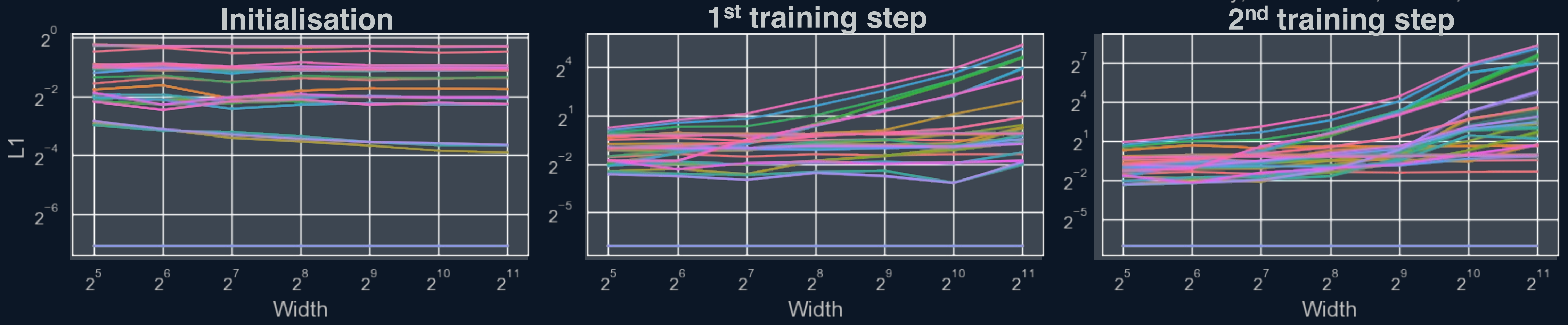
Maximal Update Parametrization

Pre-activation weights $L_1(m) = \sum |w_i^{(m)}|$

Blows up for SP, stable for µP!

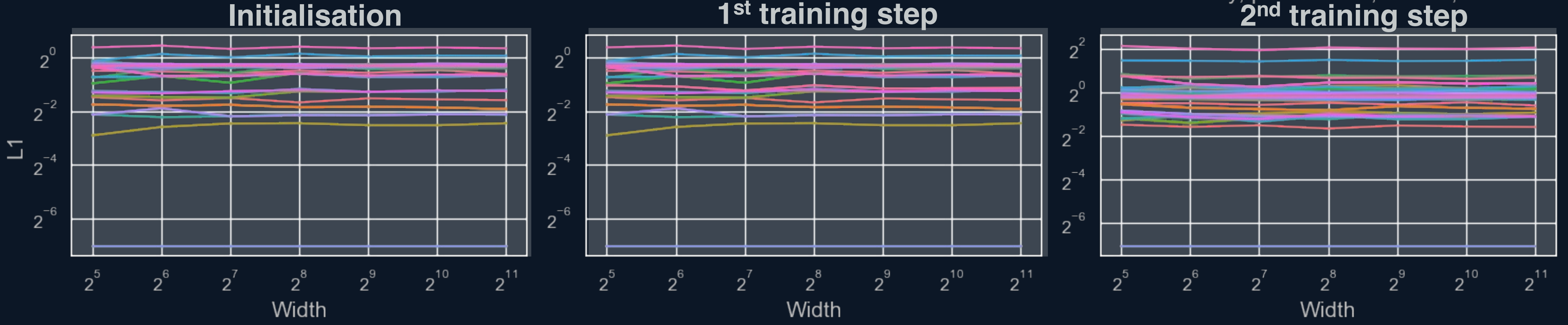
SP

ATLAS Simulation Preliminary, SP GN2-like, lr=0.01, nseeds=5



µP

ATLAS Simulation Preliminary, µP GN2-like, lr=0.01, nseeds=5

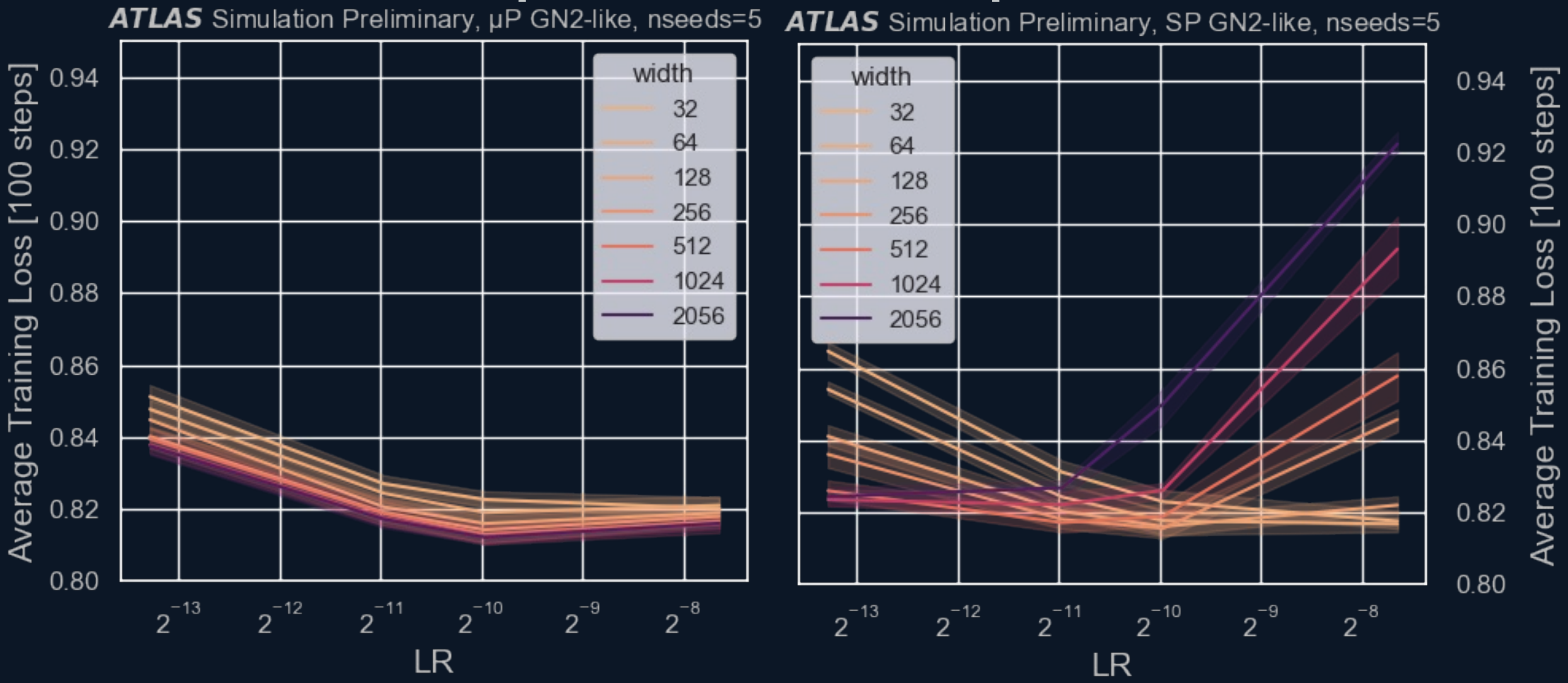


μTransfer

Pathologic Test Case

Fixed LR Value Optimisation on Simplified Architecture

μP


ATLAS Simulation Preliminary, μP GN2-like, nseeds=5

ATLAS Simulation Preliminary, SP GN2-like, nseeds=5

SP

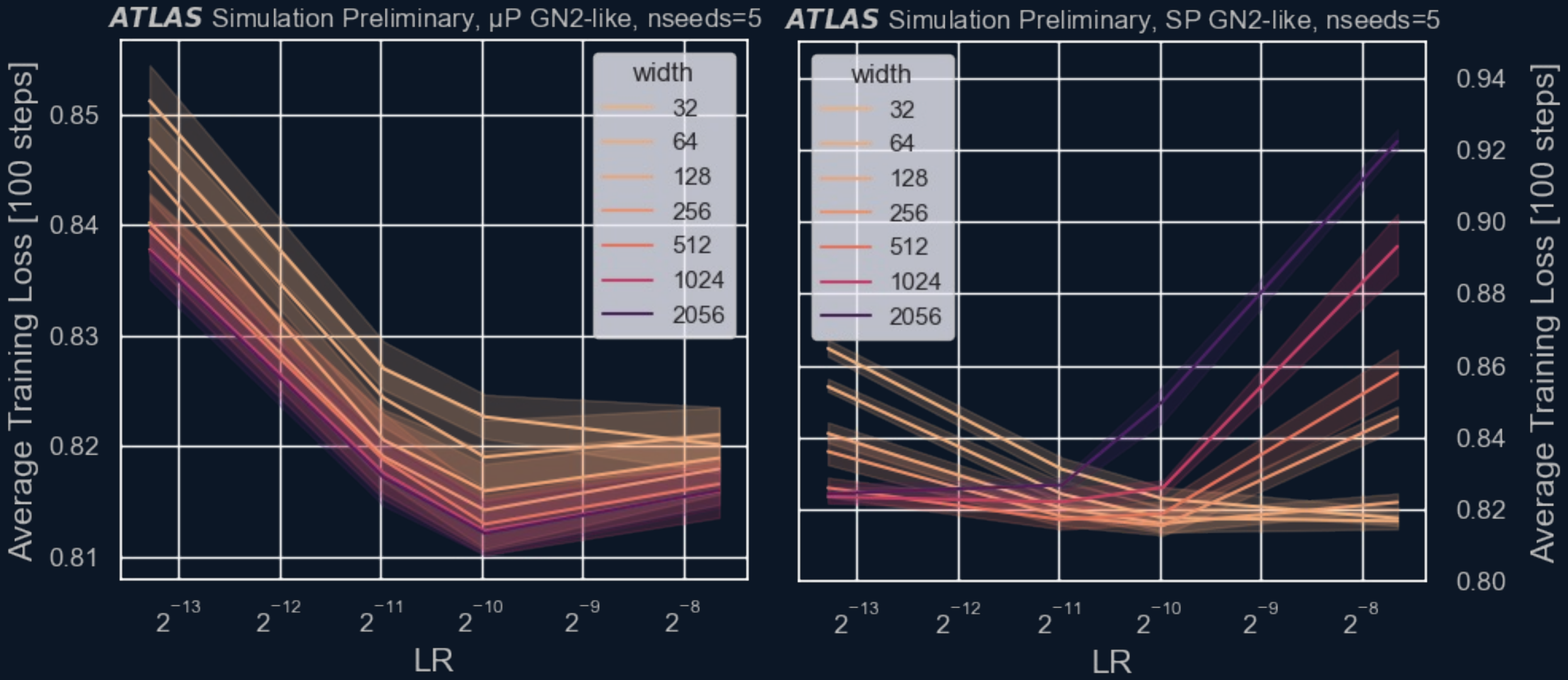
- ✓ μP models share LR optimum across large widths differences
- ✓ Wider μP models are better
- ✓ μP models outperforms SP equivalent
- ✓ SP models do not share LR optimum at different widths

μTransfer

Pathologic Test Case

Fixed LR Value Optimisation on Simplified Architecture

μP



SP

- ✓ μP models share LR optimum across large widths differences
- ✓ Wider μP models are better
- ✓ μP models outperforms SP equivalent
- ✓ SP models do not share LR optimum at different widths

GN2 *Ongoing*

- Many variants under consideration (WIP):
  - lepton,
  - more tracks,
  - hadronic taus,
  - neutral constituents
  - trackless b-tagging with hits
  - more output classes (taus, lep/had b, ...)
  - full vertex reconstruction,
  - mass / energy regression

- Now training on combined MC data for Run3: ~ 300M jets

- Calibration & Trigger ongoing

- GN2X for boosted Higgs tagging (H(bb) & H(cc)) vs top and QCD

- Synergies with other groups: Tau tagging, emerging jet tagging, ...