University of Oxford

MAXENCE DRAGUET
DANIELA BORTOLETTO

TRAINING & OPTIMISATION OF LARGE TRANSFORMERS
ATLAS CASE STUDY ON KUBEFLOW

JET FLAVOUR TAGGING

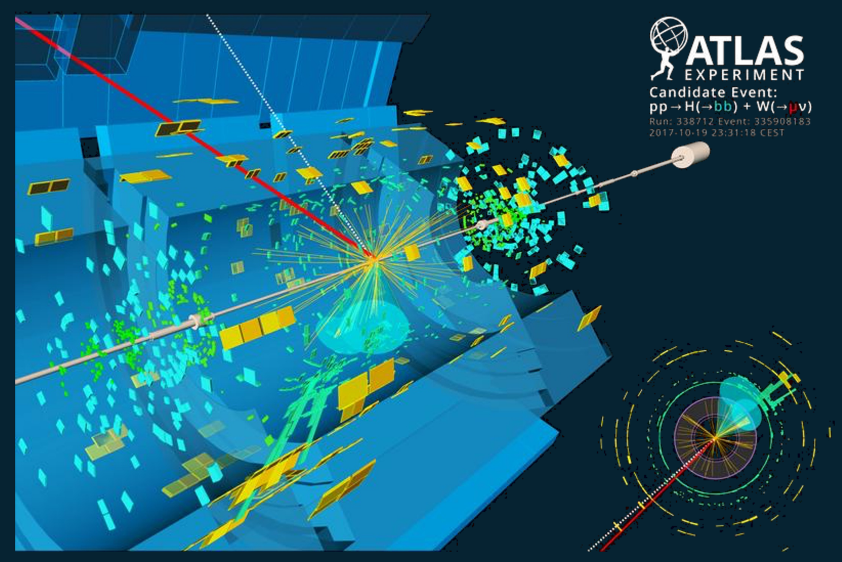
CLASSIFICATION TASK

Labels
b, c, *light*, τ

Inputs
Tracks + jet information

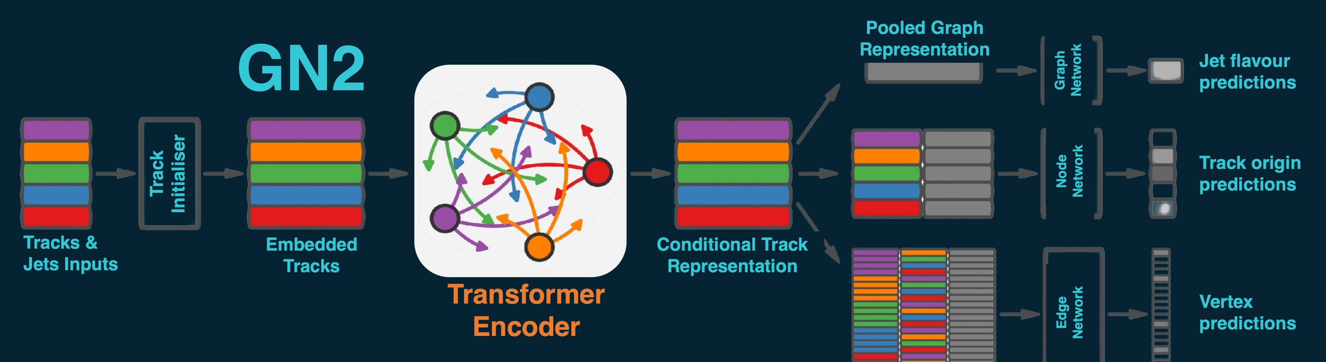
Outputs
Per-flavour probability + Discriminant

Used in ATLAS:
$H\,(\to b\bar{b}\mid c\bar{c})$, di-Higgs, ...



ATLAS event display of a Higgs boson decaying to two b-quarks

SOTA: Multimodal Multitask Transformer Model



Goal Hyperparameter Optimisation (HPO)

Kubeflow

Built on Kubernetes
Open-source container orchestration engine

Designed for MLOps
Training, Inference, Katib HPO + AutoML

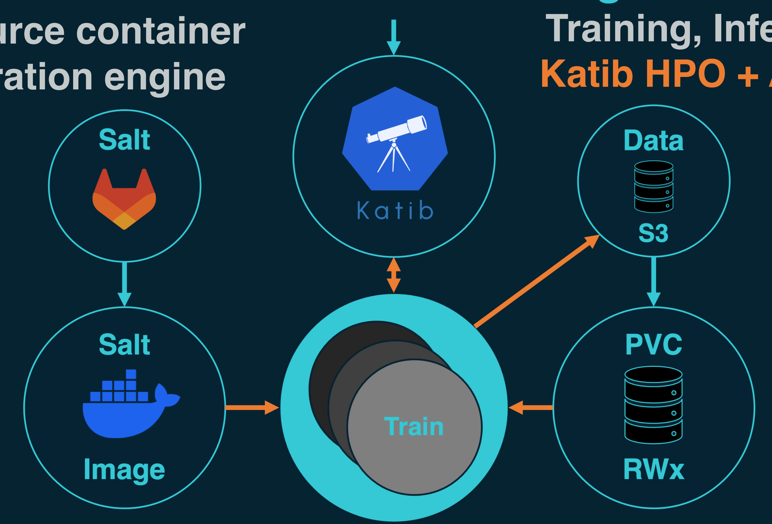
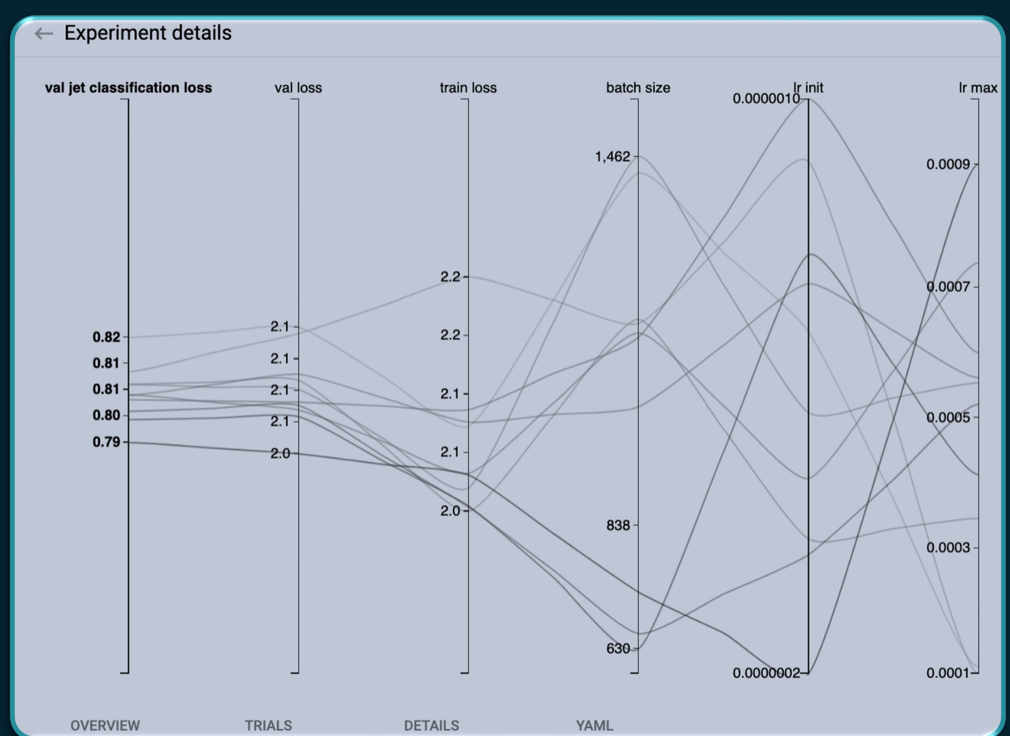


Image Contained Code

Server-based data storage



EFFICIENT DESIGN

1. Multi-platform
2. Resource Usage
3. Monitoring
4. AutoML Algorithms

PARAMETRIZATION MATTERS!

➢ **Standard Parametrization (SP)**

LeCun et al; 1998

Initialisation: $W^{L_{in}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{in}}}\right), W^{L_{hid}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{hid}}}\right), W^{L_{out}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{out}}}\right), b^{L_{...}} = 0$

SGD & Adam LR: $\forall$ weights η

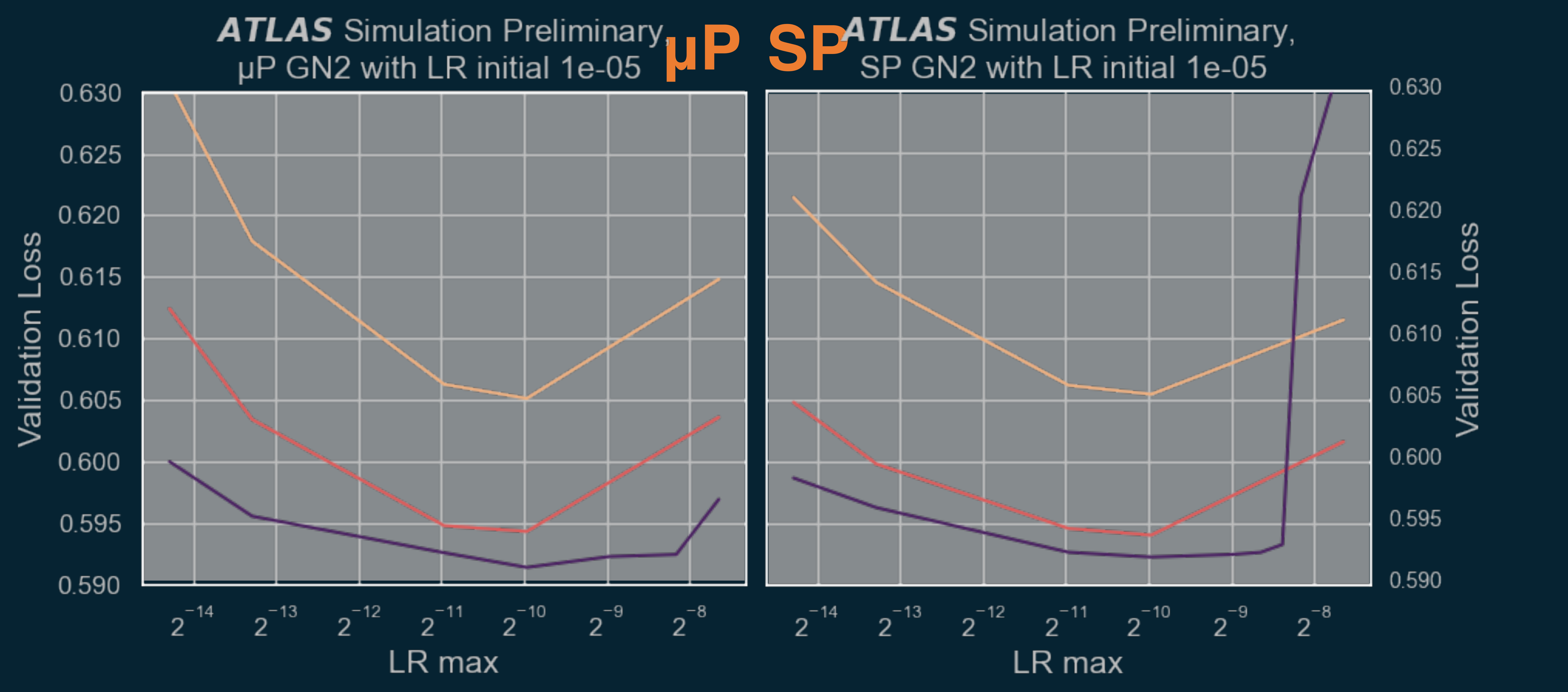
➢ **Maximal Update Parametrization (µP)**

Yang et Hu; 2022

Initialisation: $W^{L_{in}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{in}}}\right), W^{L_{hid}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{hid}}}\right), W^{L_{out}} \sim \mathcal{N}\left(0, \frac{1}{d^{in}_{L_{out}} \times d^{in}_{L_{out}}}\right), b^{L_{...}} = 0$

SGD LR: $\eta_{W_{L_{in}}} = \eta_{b_{L_{...}}} = \eta \times d^{out}_{L_{in}\,|\,L_{...}}$ $\quad \eta_{W_{L_{hid}}} = \eta$ $\quad \eta_{W_{L_{out}}} = \eta\,/\,d^{in}_{L_{out}}$

Adam LR: $\eta_{W_{L_{in}}} = \eta_{b_{L_{...}}} = \eta$ $\quad \eta_{W_{L_{hid}}} = \eta\,/\,d^{in}_{L_{hid}}$ $\quad \eta_{W_{L_{out}}} = \eta\,/\,d^{in}_{L_{out}}$

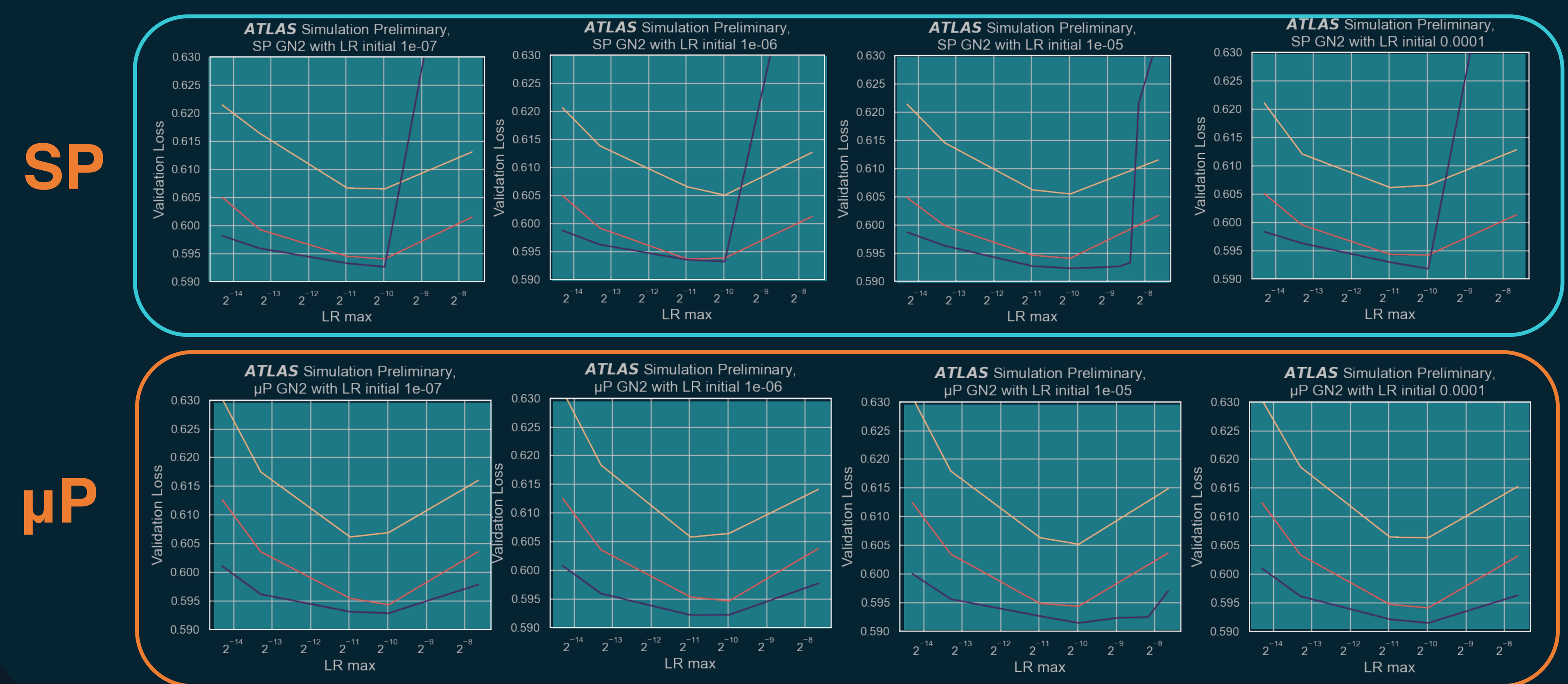


µTransfer Algorithm

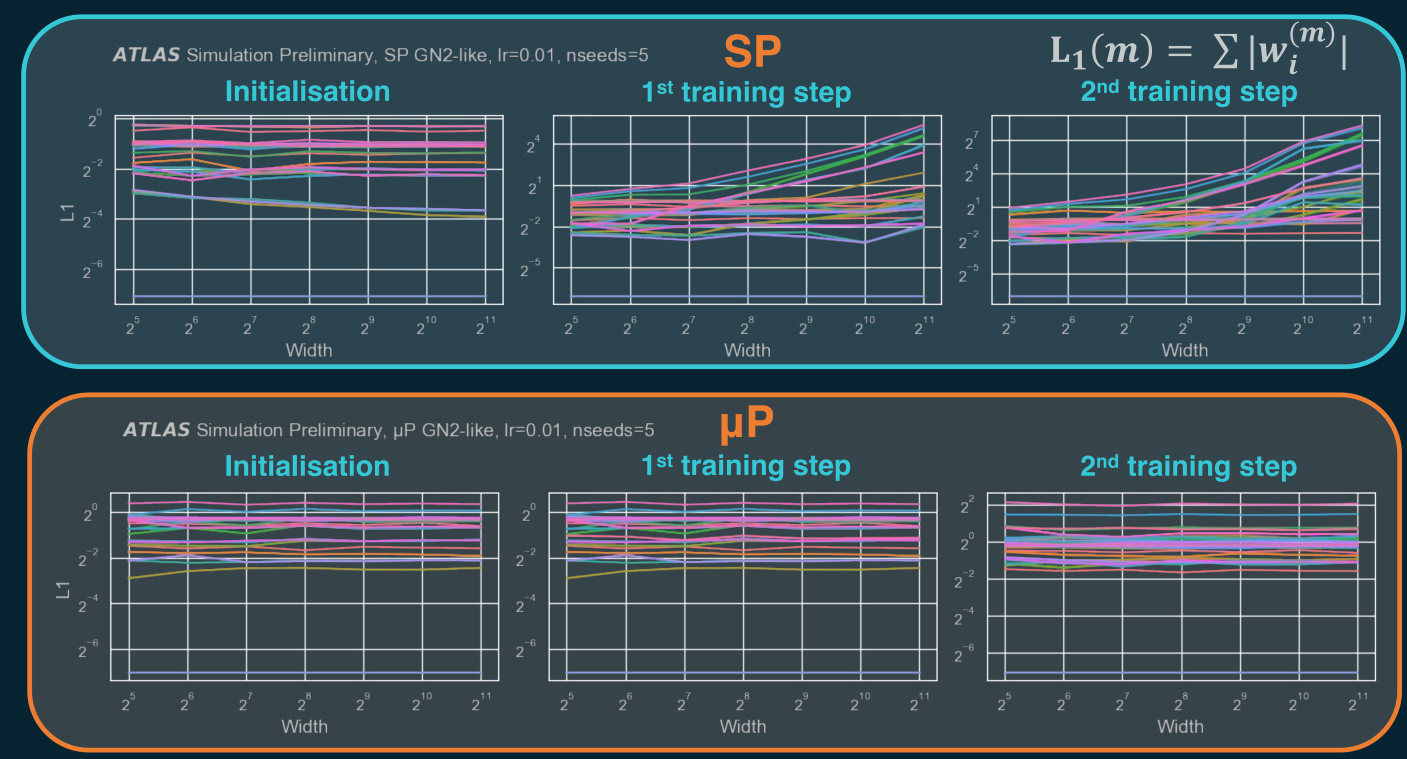
Yang et al; 2022

1. Similar performance hierarchy between small and large models in µP
2. Hyperparameter optimisation on a smaller model
3. Zero-shot transfer: best low-complexity = best high-complexity model
4. Wider is always better

Real Deployment: LR Scheduler Optimisation (initialisation & maximal values)



VERIFICATION A look at pre-activation weights



Stable for µP, blows up for SP!

ADVANTAGES

Kubeflow

1. Multi-platform & flexible
2. Hardware agnostic MLOps & optimised resource usage
3. Advanced AutoML algorithms for improved HPO
4. Powerful visualisation
5. CERN-wide access

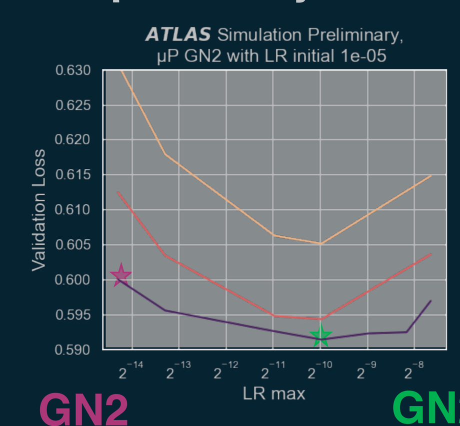
Maximal Update Parametrization (µP)

1. Wider is always better → simple neural architecture search
2. µTransfer HP zero-shot small → large models
3. Reduced computing requirement per HP set test
   ➢ Width 256 (2.30 M params): 2 GPUs → 39 min / epoch
   ➢ Width 64 (0.29 M params): 1 GPU → 20 min / epoch
4. Better coverage of the HP search space
   ➢ With µP, 4 small-model tests ≈ 1 full-model test

COMBINE

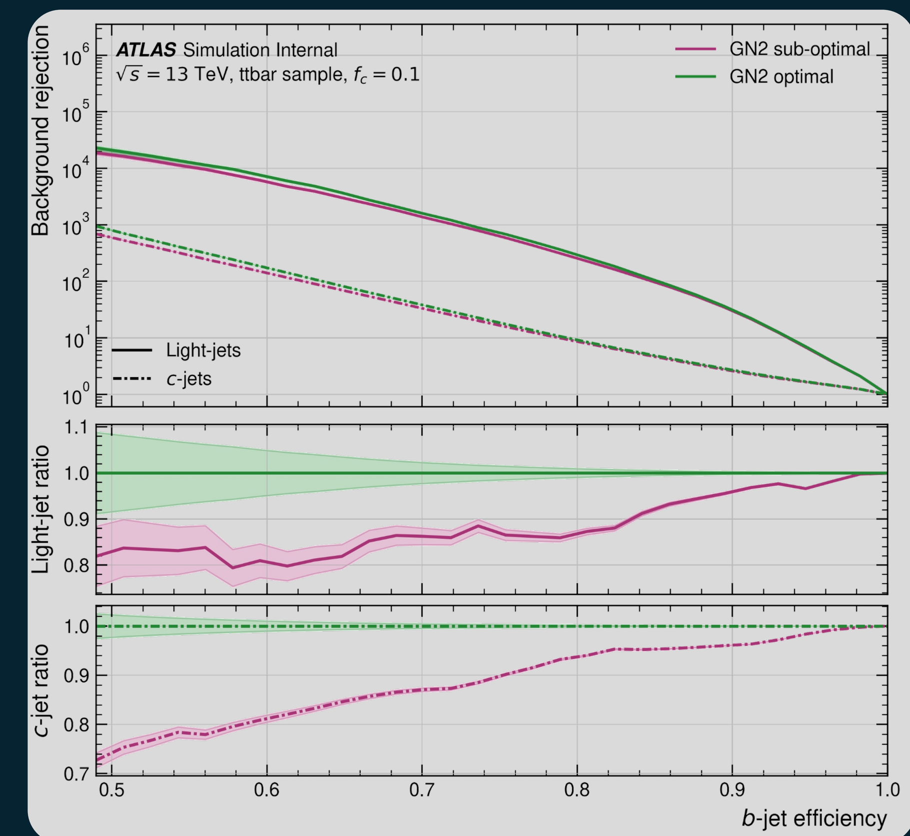
HPO MATTERS

Significant performance dependency on HP



GN2 Sub-optimal

GN2 Optimal



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