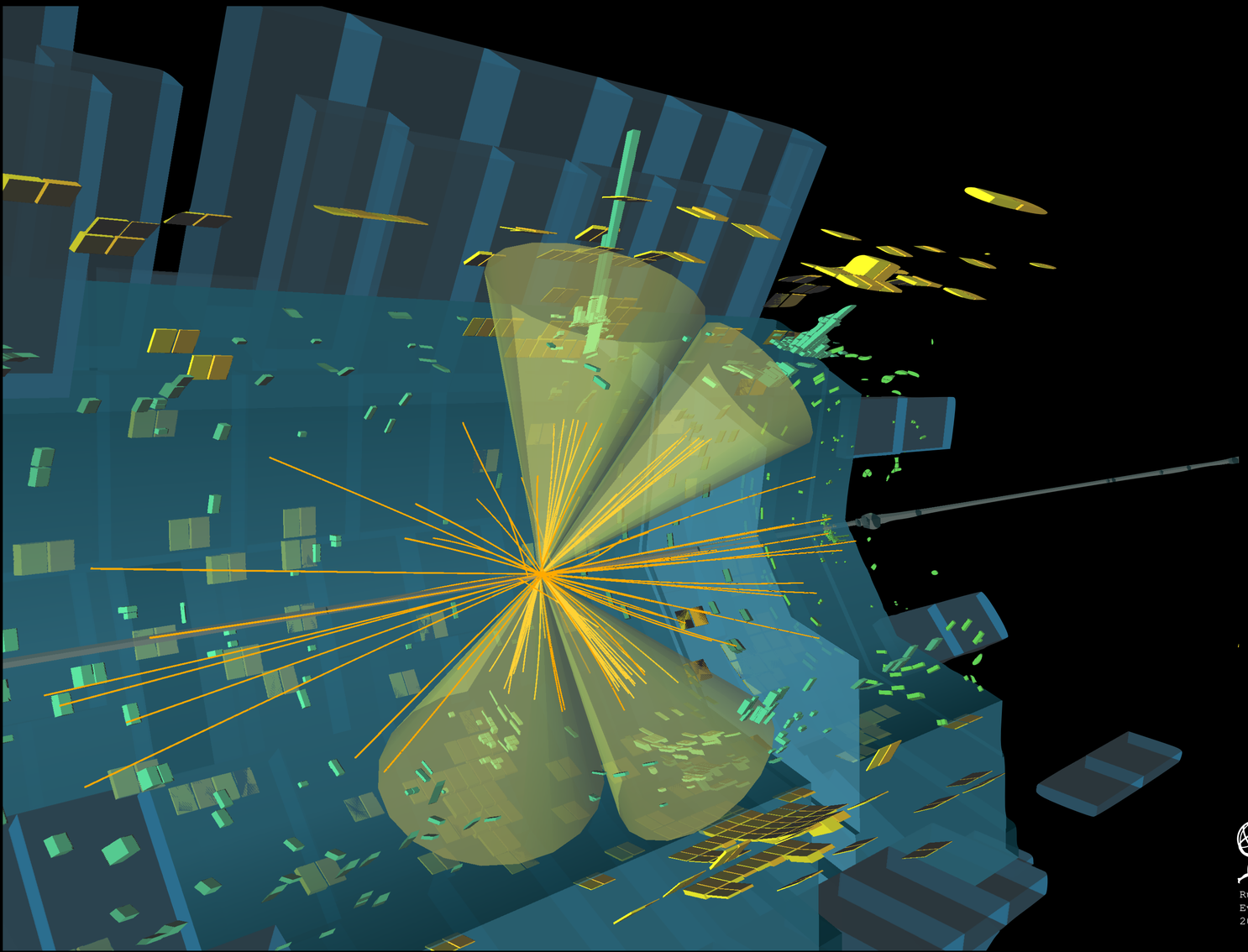
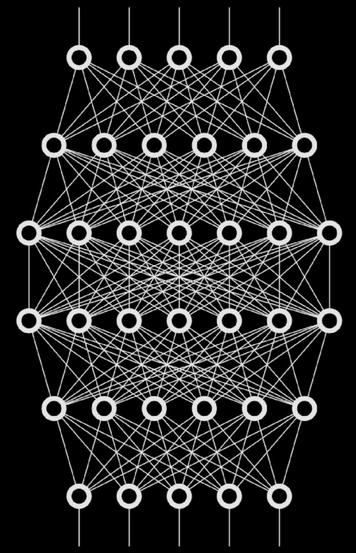
Machine Learning for Particle Physics

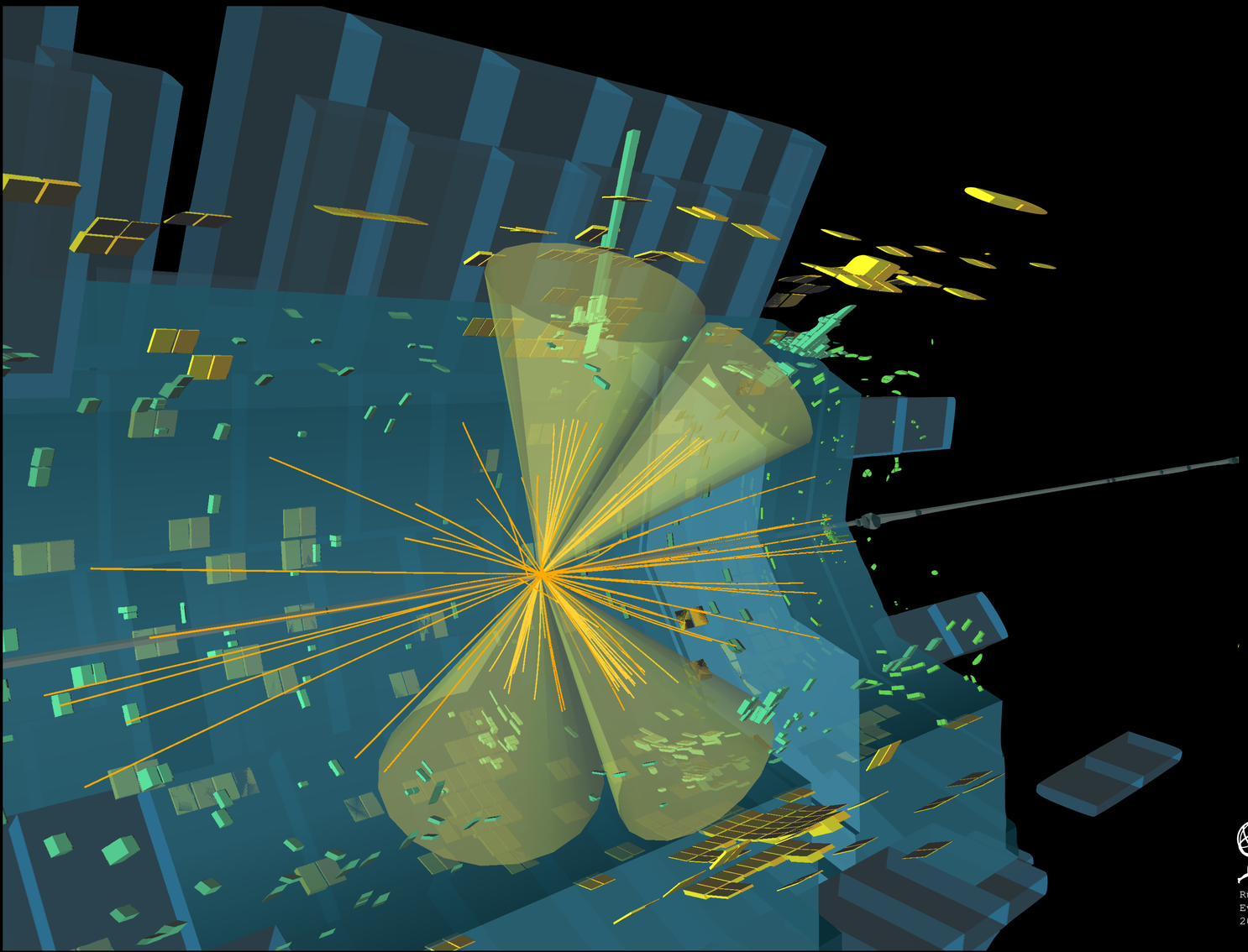


$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\not{D}\psi + h.c + \psi_i y_{ij}\psi_j\phi + h.c + |D_\mu\phi|^2 - V(\phi)$$

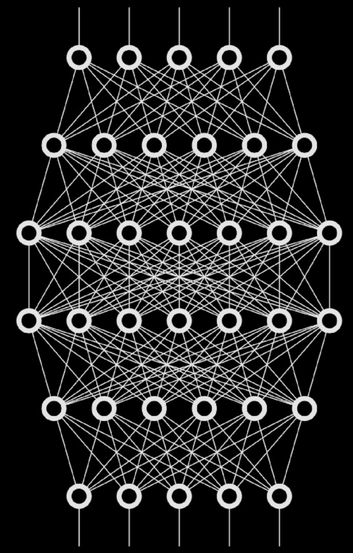


Nicole Hartman
nicole.hartman@tum.de
Belle II Germany meeting
University of Bonn, 8.9.2025

Machine Learning for ~~Particle Physics~~ High Energy Physics

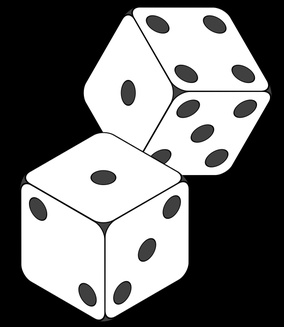


$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\not{D}\psi + h.c + \psi_i y_{ij}\psi_j\phi + h.c + |D_\mu\phi|^2 - V(\phi)$$



Nicole Hartman
nicole.hartman@tum.de
Belle II Germany meeting
University of Bonn, 8.9.2025

Particle physics: Probabilistic model to describe the world we live in.

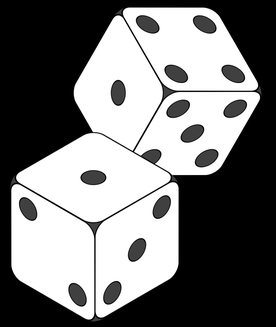


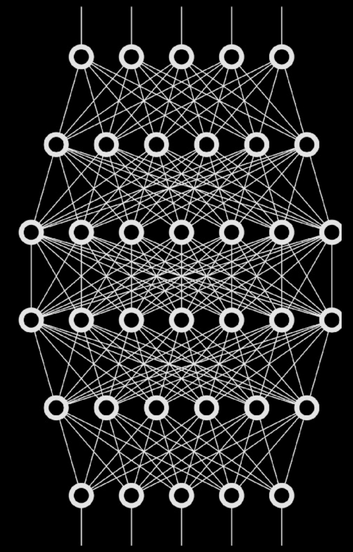
$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i \not{F} \not{D} \chi + h.c + \chi_i y_{ij} \chi_j \phi + h.c + |D_\mu \phi|^2 - V(\phi)$$

$$p(x \mid \theta) = \frac{1}{\sigma} \int dx \, |\mathscr{M}(x)|^2$$

data SM

Machine Learning: Probabilistic model to describe the world we live in.

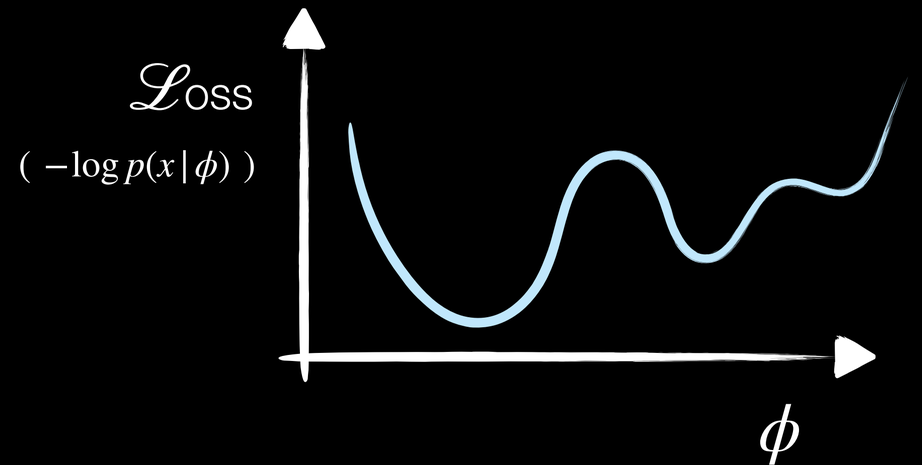




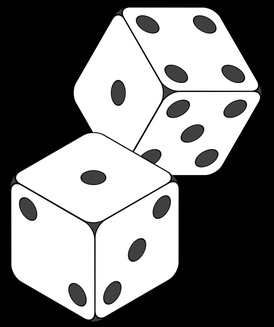
$$p(x \mid \phi)$$

data Parameters

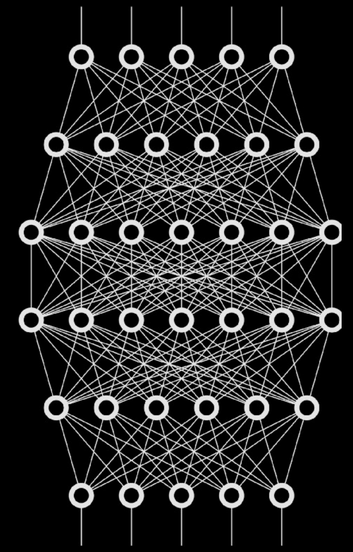
$$\phi \leftarrow \phi - \lambda \nabla_{\phi} \mathcal{L}$$



Machine Learning: Probabilistic model to describe the world we live in.



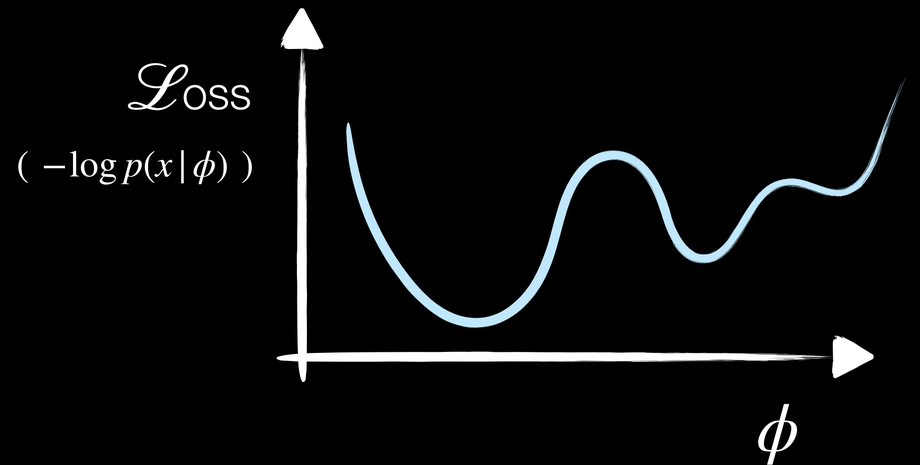
Data drives discoveries



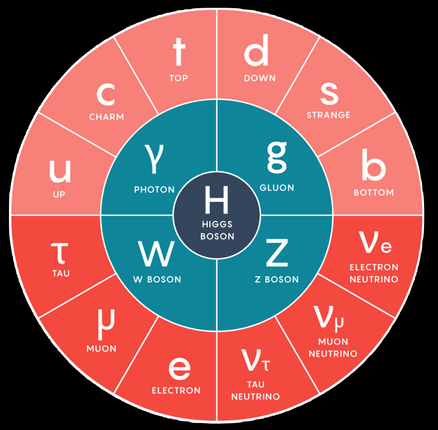
$$p(x \mid \phi)$$

data Parameters

$$\phi \leftarrow \phi - \lambda \nabla_{\phi} \mathscr{L}$$



Machine Learning for particle physics



W / Z 1983

High fidelity MC simulations

Cern Courrier; Discovery of the weak neutral currents

top 1995

Probabilistic reconstruction

Phys. Rev. D 94, 032004 (2016)

A. Matrix Element Technique

This measurement uses the matrix element technique [13]. This method provides the most precise m_t measurement at the Tevatron in the ℓ+jets final state [1, 2], and was applied in previous measurement of m_t in the dilepton final state using 5.3 fb^{-1} of integrated luminosity [39]. The ME method used in this analysis is described below.

B. Event probability calculation

The ME technique assigns a probability to each event, which is calculated as

$$P(x, f_{t\bar{t}}, m_t) = f_{t\bar{t}} \cdot P_{t\bar{t}}(x, m_t) + (1 - f_{t\bar{t}}) \cdot P_{\rm bkg}(x), \quad (2)$$

where $f_{t\bar{t}}$ is the fraction of $t\bar{t}$ events in the data, and $P_{t\bar{t}}$ and $P_{\rm bkg}$ are the respective per-event probabilities calculated under the hypothesis that the selected event is either a $t\bar{t}$ event, characterized by a top quark mass m_t, or background. Here, x represents the set of measured observables, i.e., p_T, η, and ϕ for jets and leptons. We assume that the masses of top quarks and anti-top quarks are the same. The probability $P_{t\bar{t}}(x, m_t)$ is calculated as

$$P_{t\bar{t}}(x, m_t) = \frac{1}{\sigma_{\rm obs}(m_t)} \int f_{\rm PDF}(q_1) f_{\rm PDF}(q_2) \times \times \frac{(2\pi)^4 |\mathscr{M}(y, m_t)|^2}{q_1 q_2 s} W(x, y) d\Phi_6 dq_1 dq_2, \quad (3)$$

B-factories 2000s

Measurements of Branching Fraction, Polarization, and Direct-CP-Violating Charge Asymmetry in $B^+ \to K^{*0}\rho^+$ Decays

The *BABAR* Collaboration

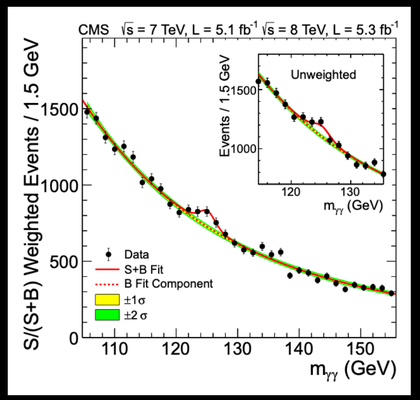
August 18, 2004

To discriminate signal from continuum background, we also use a neural network (NN) combining six variables: a Fisher discriminant made from two event-shape variables (see [18]); the cosine of the angle between the direction of the B and the collision axis (z) in the CM frame; the cosine of the angle between the B-thrust axis and the z axis; the cosine of the angle between the B-thrust axis and the thrust of the particles of the rest of the event; the angle between the direction of the π^0 and that of one of its daughter photons in the π^0 rest frame (π^0 decay angle); and the sum of transverse momenta relative to the z-axis of the particles in the rest of the event.

hep-ex/0408093

Higgs 2012

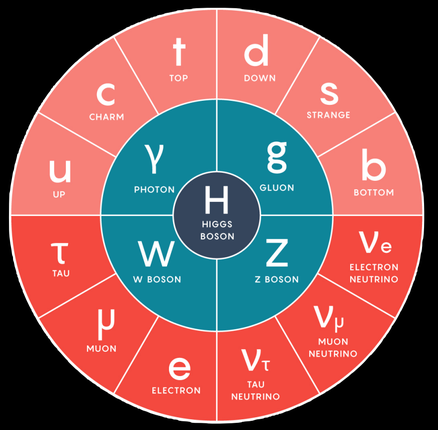
BDTs for Higgs discovery



Nature 560 (2018)

Phys Lett B (2012) 30

Machine Learning for particle physics



W / Z
1983

top
1995

B-factories
2000s

Higgs
2012

What's next??

High fidelity MC simulations

Cern Courrier; Discovery of the weak neutral currents

Probabilistic reconstruction

Phys. Rev. D 94, 032004 (2016)

A. Matrix Element Technique

This measurement uses the matrix element technique [13]. This method provides the most precise m_t measurement at the Tevatron in the ℓ+jets final state [1, 2], and was applied in previous measurement of m_t in the dilepton final state using 5.3 fb^{-1} of integrated luminosity [39]. The ME method used in this analysis is described below.

B. Event probability calculation

The ME technique assigns a probability to each event, which is calculated as

$$P(x, f_{t\bar{t}}, m_t) = f_{t\bar{t}} \cdot P_{t\bar{t}}(x, m_t) + (1 - f_{t\bar{t}}) \cdot P_{\rm bkg}(x), \quad (2)$$

where $f_{t\bar{t}}$ is the fraction of $t\bar{t}$ events in the data, and $P_{t\bar{t}}$ and $P_{\rm bkg}$ are the respective per-event probabilities calculated under the hypothesis that the selected event is either a $t\bar{t}$ event, characterized by a top quark mass m_t, or background. Here, x represents the set of measured observables, i.e., p_T, η, and ϕ for jets and leptons. We assume that the masses of top quarks and anti-top quarks are the same. The probability $P_{t\bar{t}}(x, m_t)$ is calculated as

$$P_{t\bar{t}}(x, m_t) = \frac{1}{\sigma_{\rm obs}(m_t)} \int f_{\rm PDF}(q_1) f_{\rm PDF}(q_2) \times \times \frac{(2\pi)^4 |\mathscr{M}(y, m_t)|^2}{q_1 q_2 s} W(x, y) d\Phi_6 dq_1 dq_2, \quad (3)$$

Measurements of Branching Fraction, Polarization, and Direct-CP-Violating Charge Asymmetry in $B^+ \to K^{*0}\rho^+$ Decays

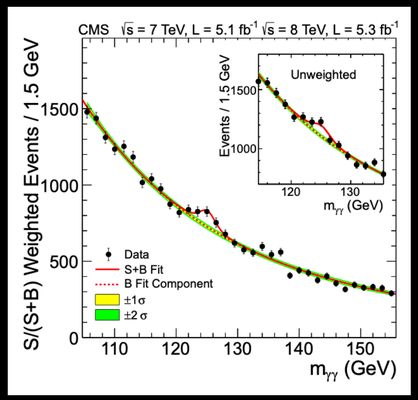
The *BABAR* Collaboration

August 18, 2004

To discriminate signal from continuum background, we also use a neural network (NN) combining six variables: a Fisher discriminant made from two event-shape variables (see [18]); the cosine of the angle between the direction of the B and the collision axis (z) in the CM frame; the cosine of the angle between the B-thrust axis and the z axis; the cosine of the angle between the B-thrust axis and the thrust of the particles of the rest of the event; the angle between the direction of the π^0 and that of one of its daughter photons in the π^0 rest frame (π^0 decay angle); and the sum of transverse momenta relative to the z-axis of the particles in the rest of the event.

hep-ex/0408093

BDTs for Higgs discovery



Nature 560 (2018)

Phys Lett B (2012) 30

The New York Times

Physicists Find Elusive Particle Seen as Key to Universe

Share full article 122

Scientists in Geneva on Wednesday applauded the discovery of a subatomic particle that looks like the Higgs boson. Pool photo by Denis Balibouse

By Dennis Overbye

July 4, 2012

article

The New York Times

Scientists See Promise in Deep-Learning Programs

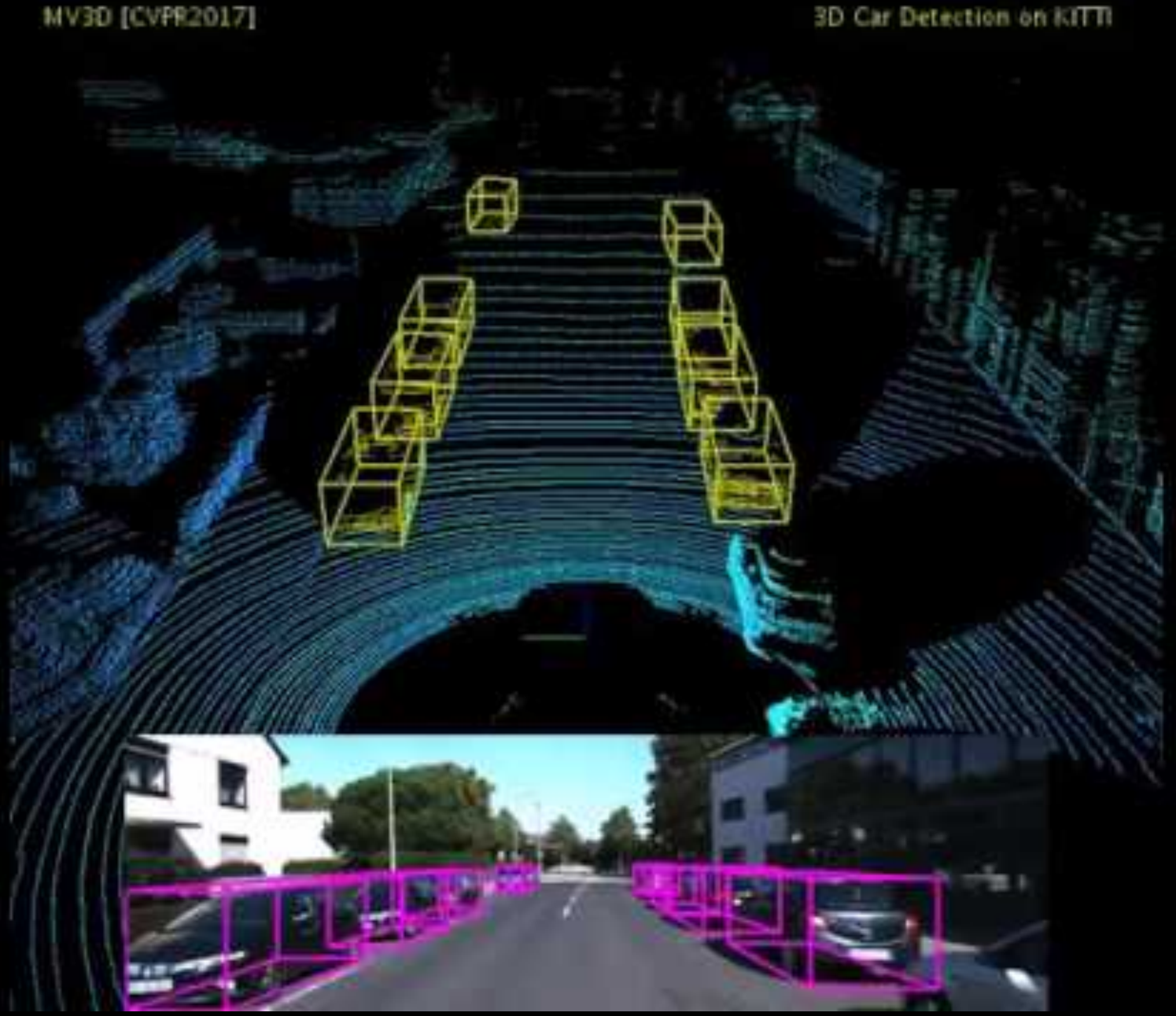
Share full article

A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese. Hao Zhang/The New York Times

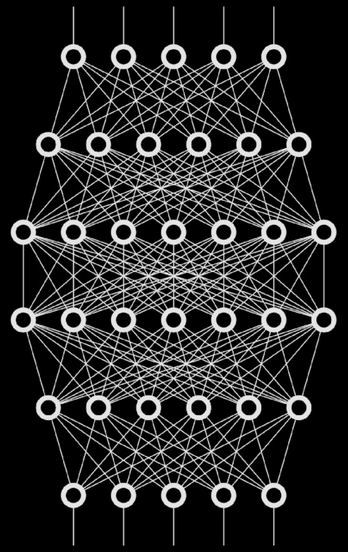
By John Markoff

Nov. 23, 2012

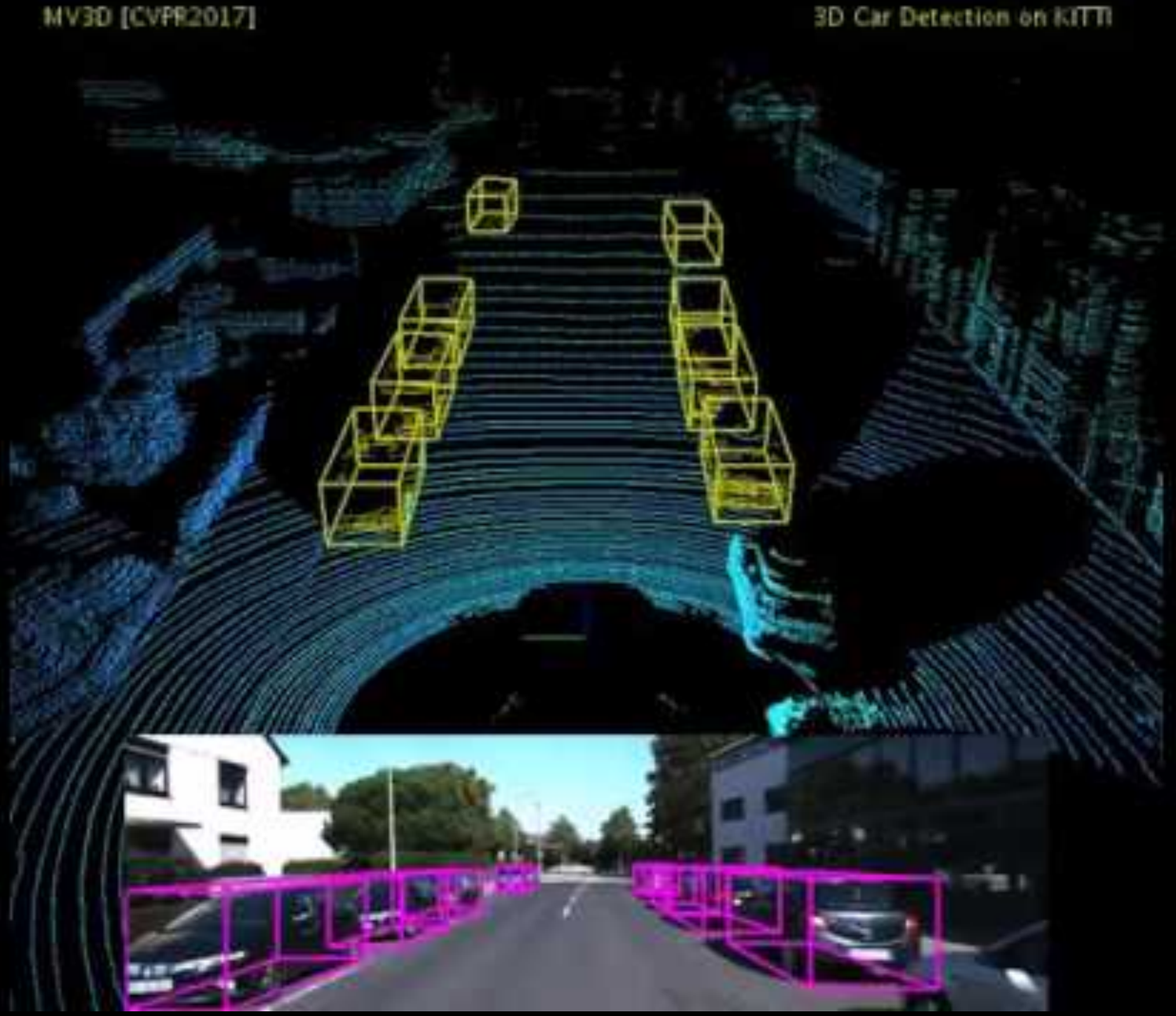
article



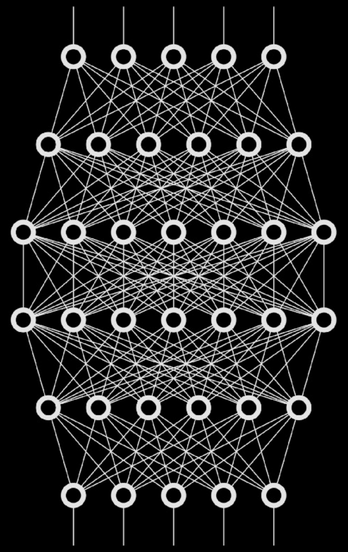
CVPR'17 video; as taken from Stanford CS231n course



Networks get *deep* with CNNs; RNNs; graph networks; transformers

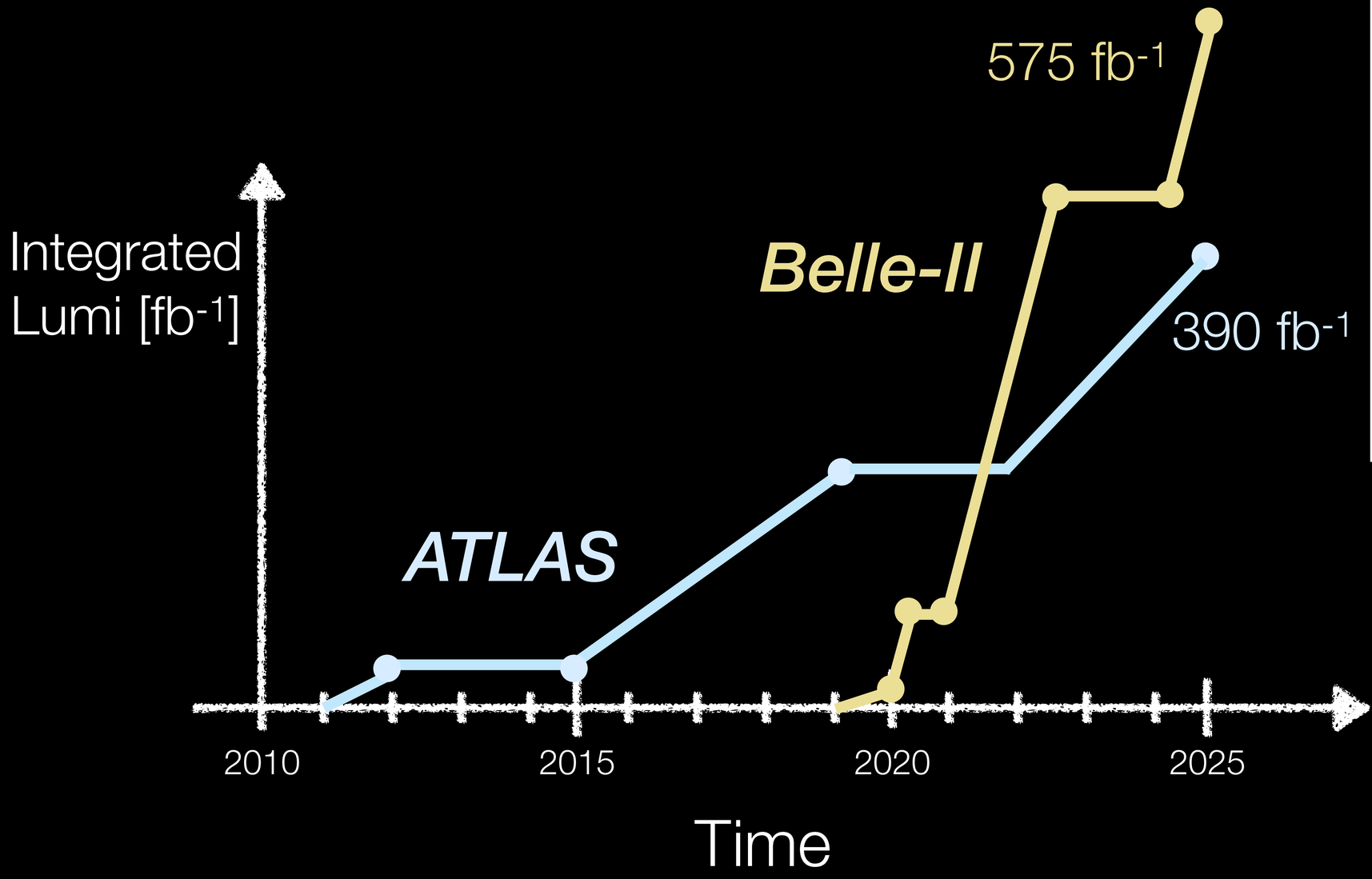


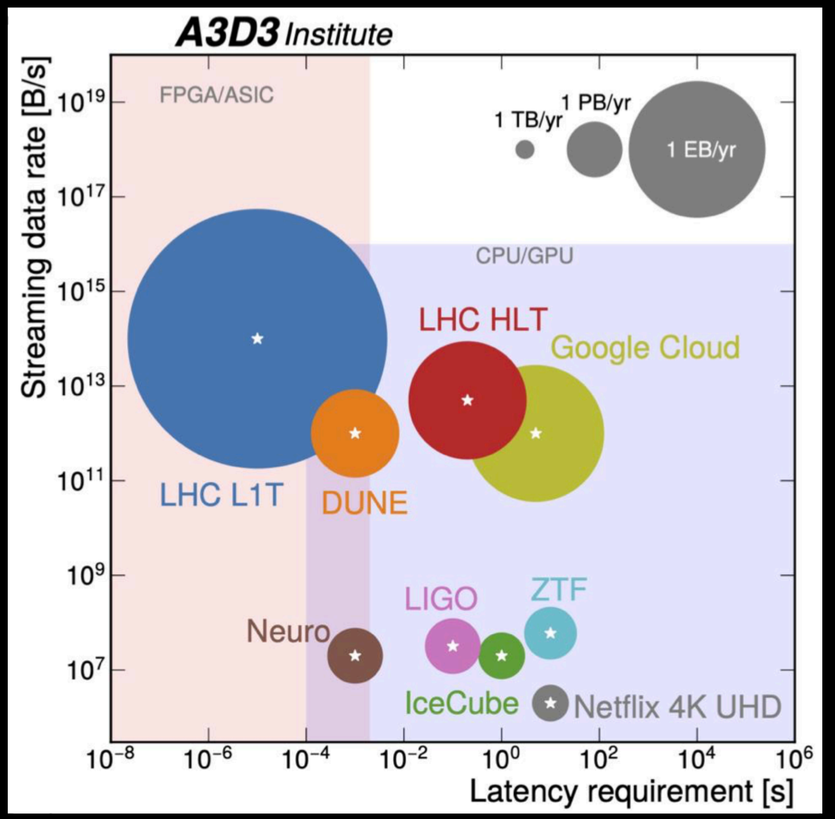
CVPR'17 video; as taken from Stanford CS231n course



Networks get *deep* with CNNs; RNNs; graph networks; transformers

Big science, big data





https://a3d3.ai/about/

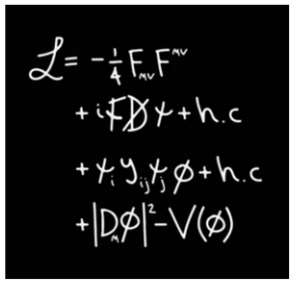
ATLAS: Run 3, HL-LHC
Belle-II lumi

$p(x \mid \theta)$

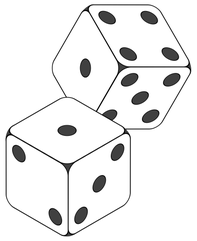
data theory

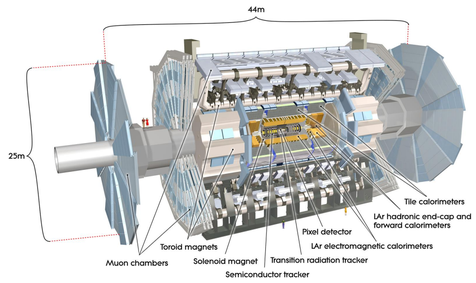
Simulation

Generation / Forward modelling



θ





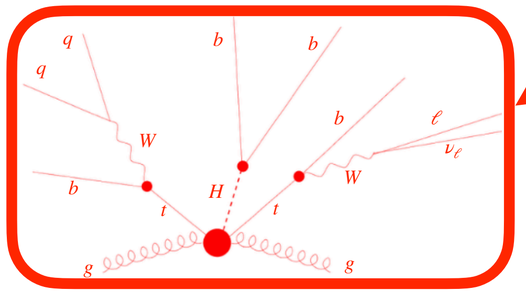
x

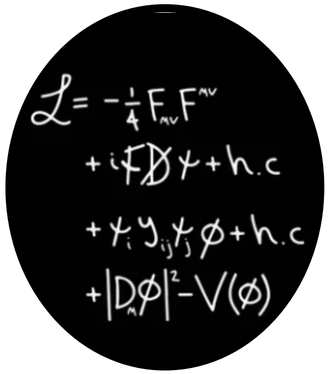
High dimension $\mathcal{O}(10^8)$

Slide adapted from K. Cranmer

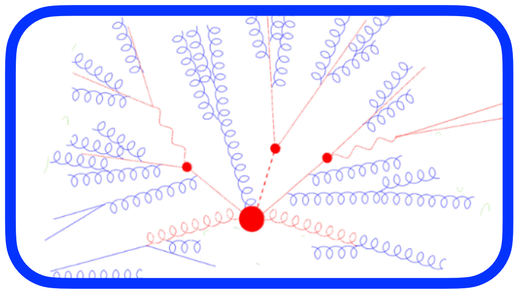
$p(\ x\ |\ \theta\ )$ Simulation

$p(z_1 \,|\, \theta)$

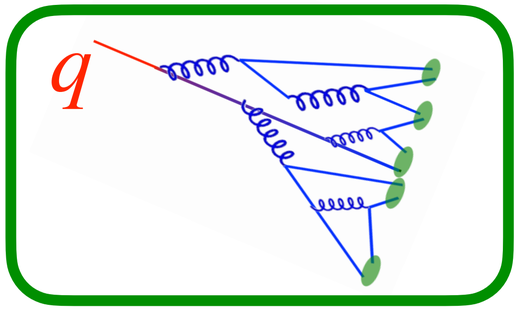


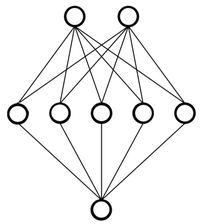


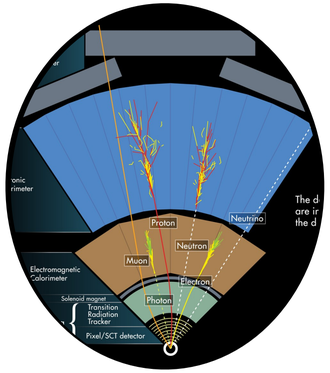
$p(z_1 \,|\, \theta)$



$p(z_3 \,|\, z_2)$







$p(x \,|\, z_3)$

Slide adapted from K. Cranmer

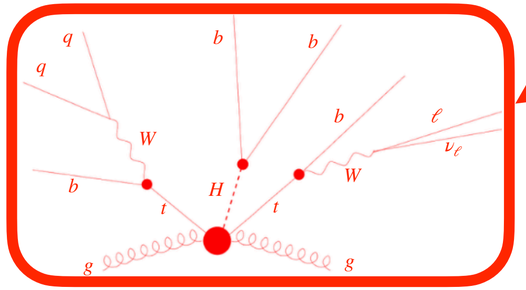
$p(x \mid \theta)$

Simulation

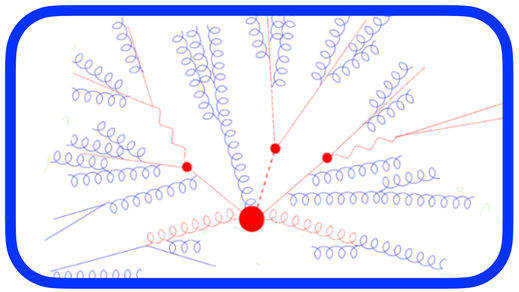
Reconstruction

$p(t(x) \mid \theta)$

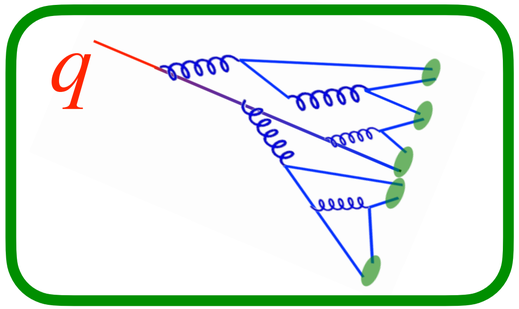
$p(z_1 \mid \theta)$



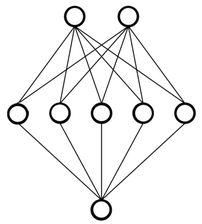
$p(z_1 \mid \theta)$

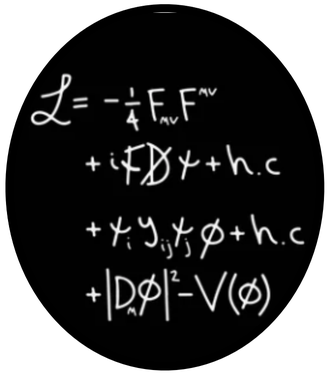


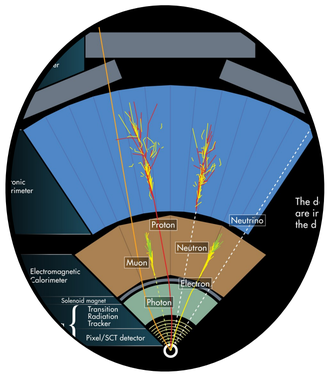
$p(z_3 \mid z_2)$



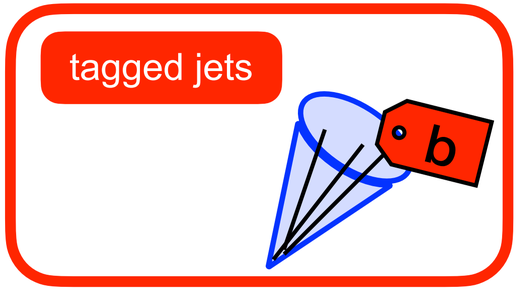
$p(x \mid z_3)$



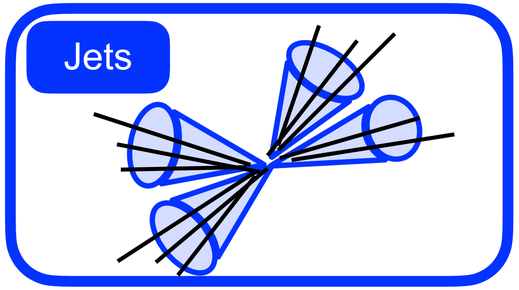




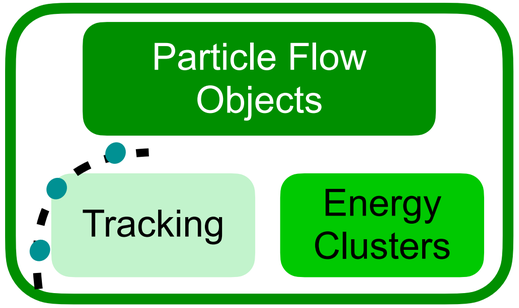
Analysis



tagging



anti-kT



Clustering

Slide adapted from K. Cranmer

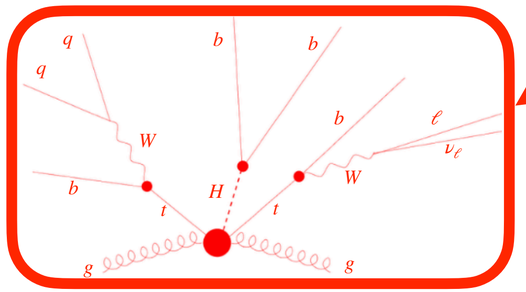
$p(x \mid \theta)$

Simulation

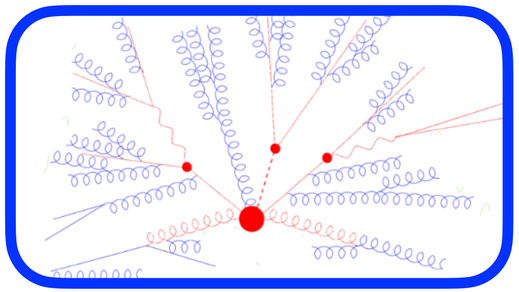
Reconstruction

$p(t(x) \mid \theta)$

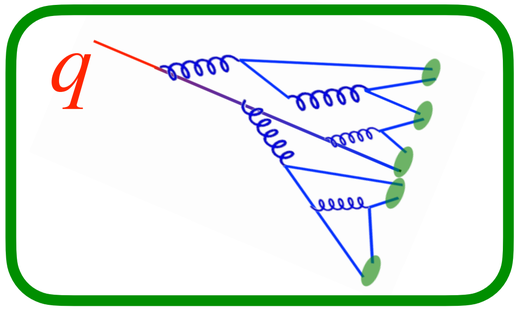
$p(z_1 \mid \theta)$



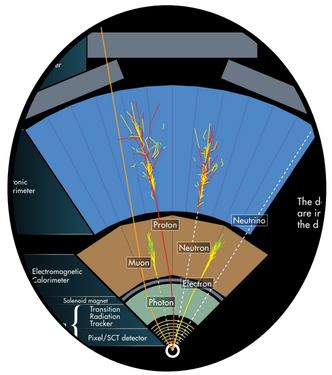
$p(z_1 \mid \theta)$

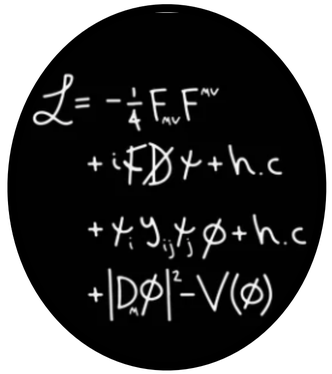


$p(z_3 \mid z_2)$

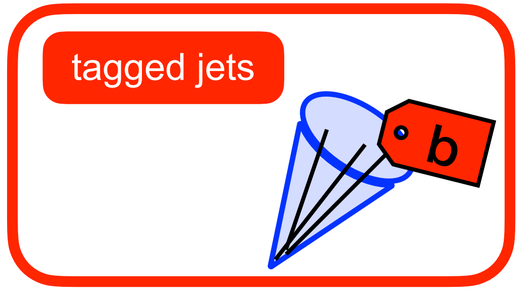


$p(x \mid z_3)$

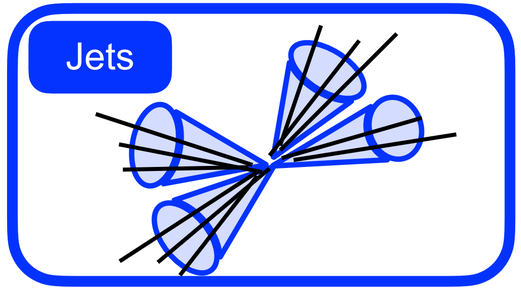




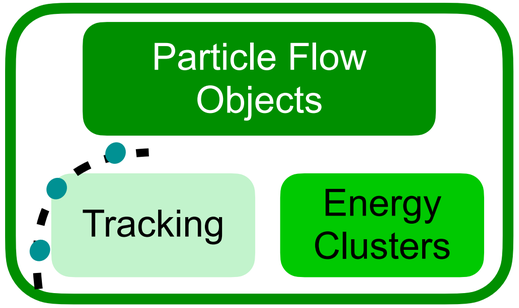
Analysis



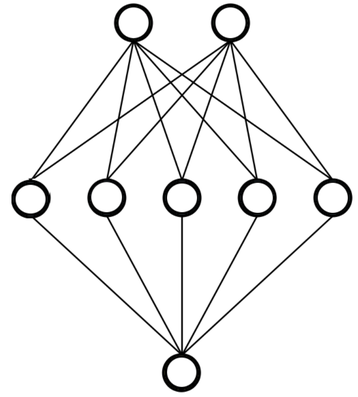
tagging



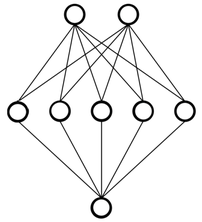
anti-kT

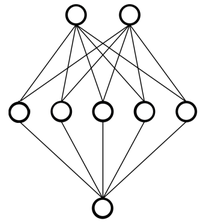


Clustering



NNs bigger & higher dim data.



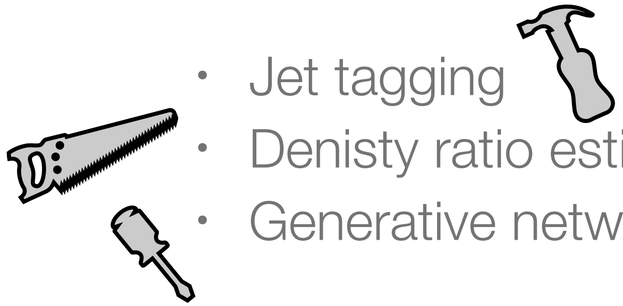


Slide adapted from K. Cranmer

1) ML for our tools

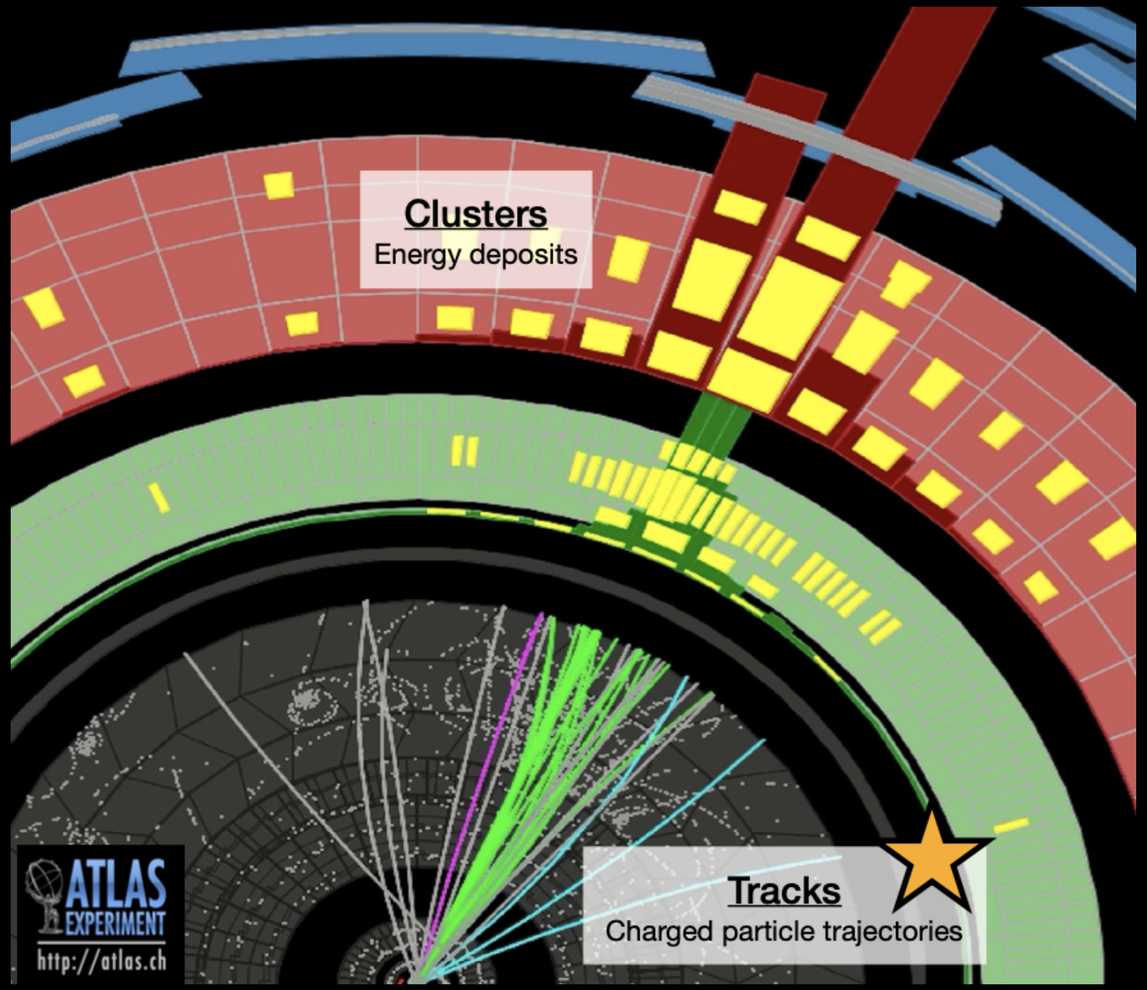
2) ML for our future

1) ML for our tools

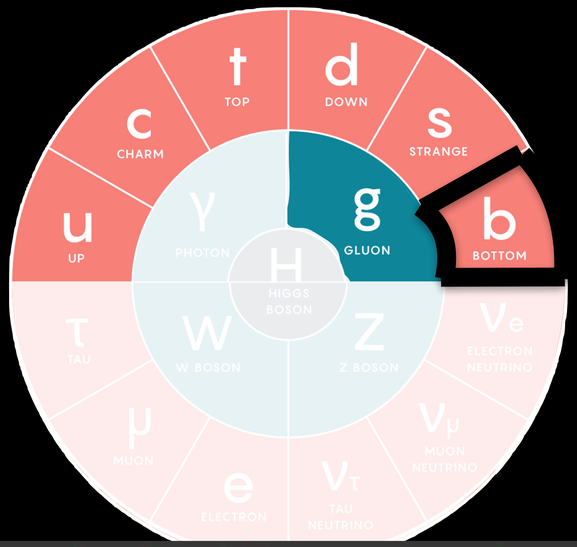


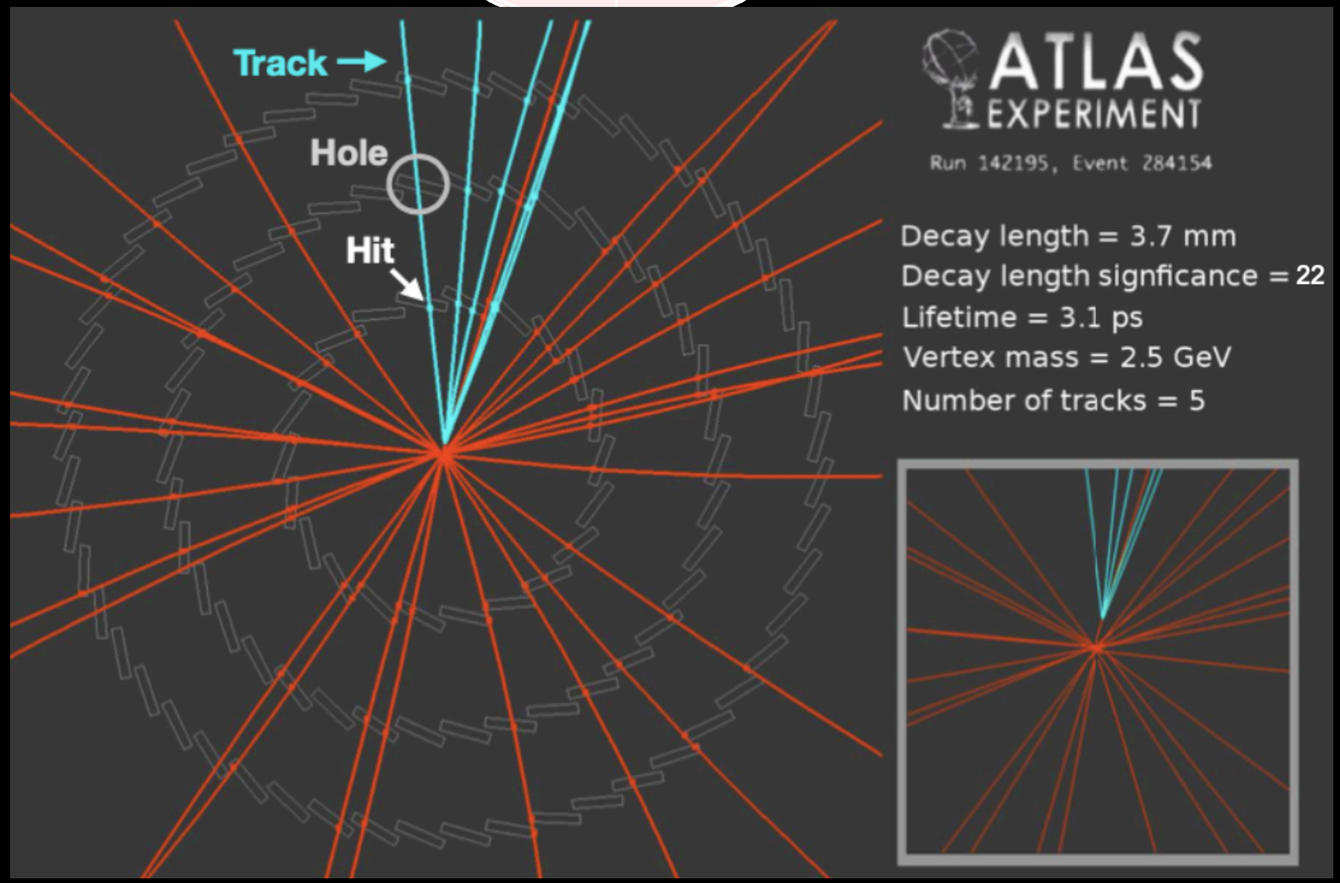
- Jet tagging
- Denisty ratio estimation
- Generative networks

2) ML for our future



Eur. Phys Journal C
Vol 73 3 (2013) 2304





1005.5254

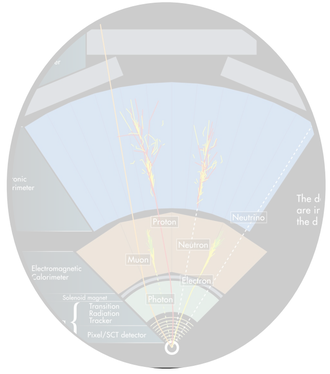
Reconstruction

$p(z_1 \mid \theta)$

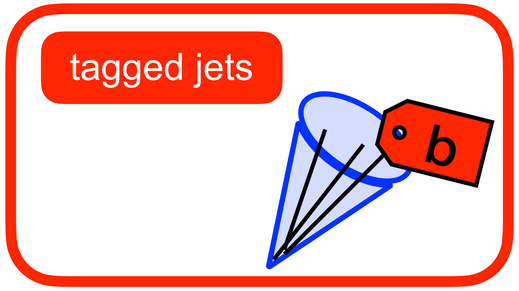
$p(z_1 \mid \theta)$

$p(z_3 \mid z_2)$

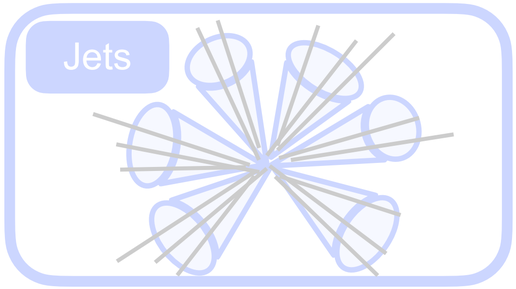
$p(x \mid z_3)$



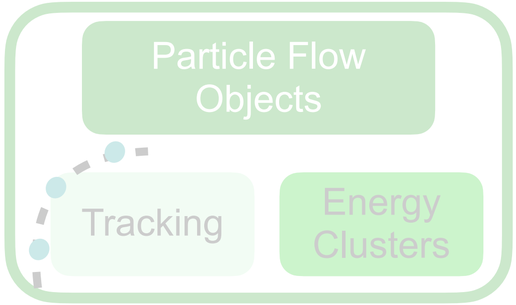
Clustering

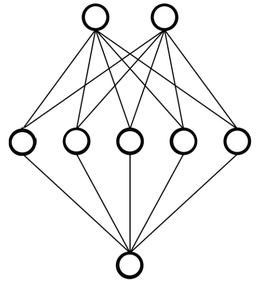


tagging



anti-kT





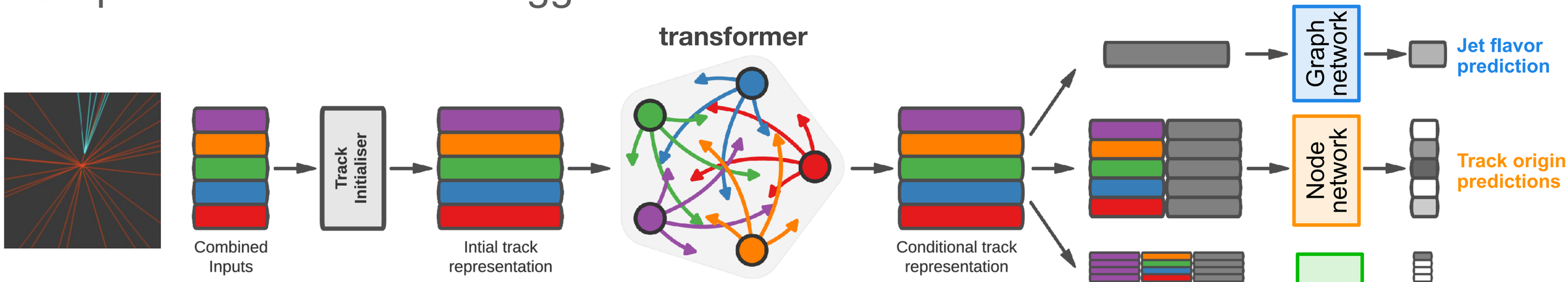
Deep learning for HEP

- CNNs (2015)
- RNNs (2017)
- Deep Sets / Graphs (2018)
- Transformers (2020)

Set-to-graph: 2002.08772
ATL-PHYS-PUB-2022-027
S. Stroud's GN* seminar

Jet tagging: multi-task learning!

Graph-net based flavour tagger

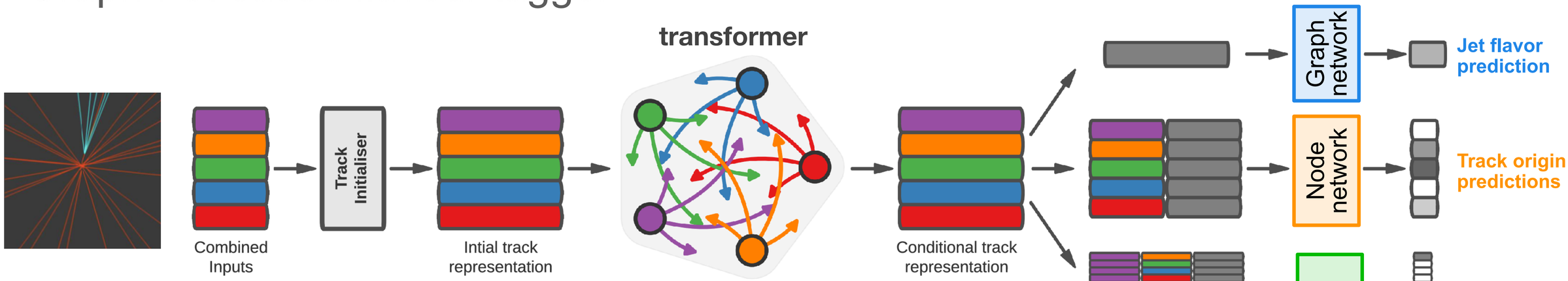


$$\mathscr{L}_{tot} = \mathscr{L}_{jet} + \alpha \mathscr{L}_{trk} + \beta \mathscr{L}_{vtx}$$

Set-to-graph: 2002.08772
ATL-PHYS-PUB-2022-027
S. Stroud's GN* seminar

Jet tagging: multi-task learning!

Graph-net based flavour tagger



$$\mathscr{L}_{tot} = \mathscr{L}_{jet} + \alpha \mathscr{L}_{trk} + \beta \mathscr{L}_{vtx}$$

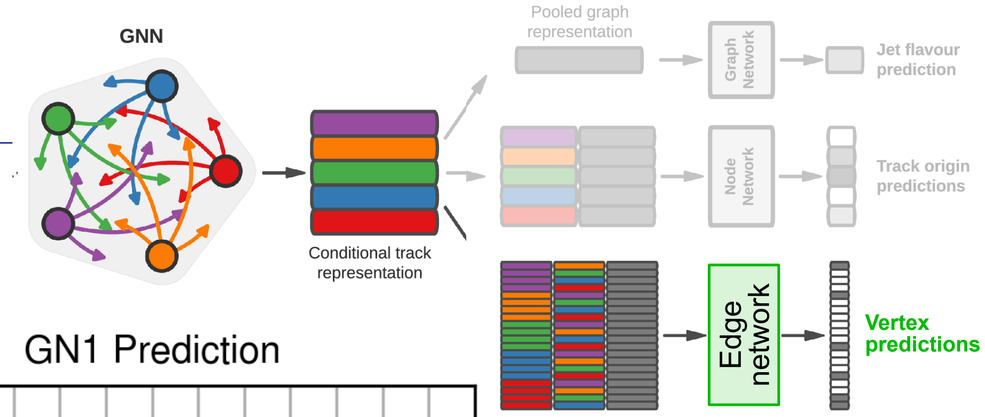
GN1 (2021): helped a lot

- 30m training jets
- 100% improvement

ATL-PHYS-PUB-2022-027

Vertex finding

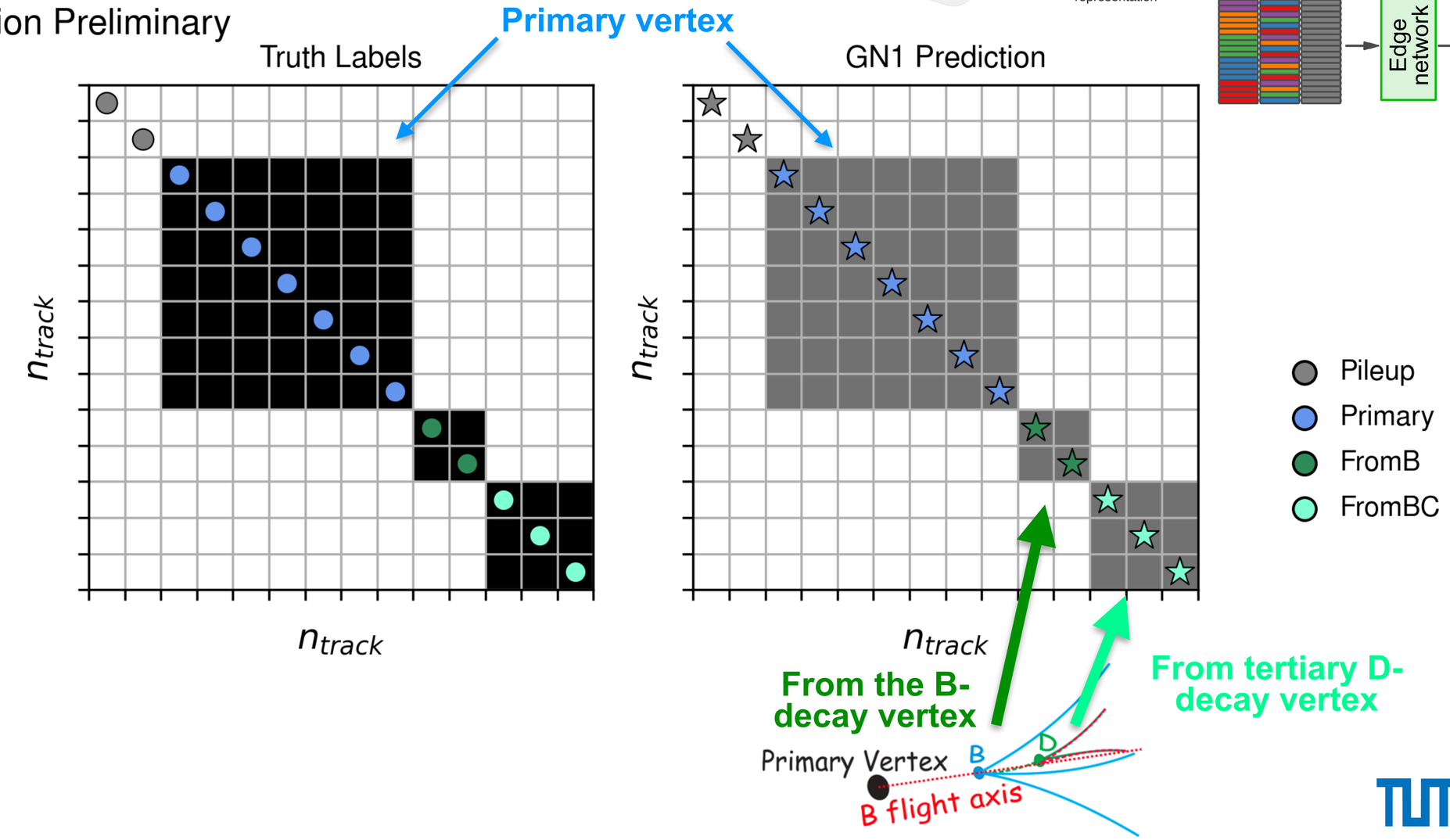
Group tracks with pair-wise compatibility > 0.5



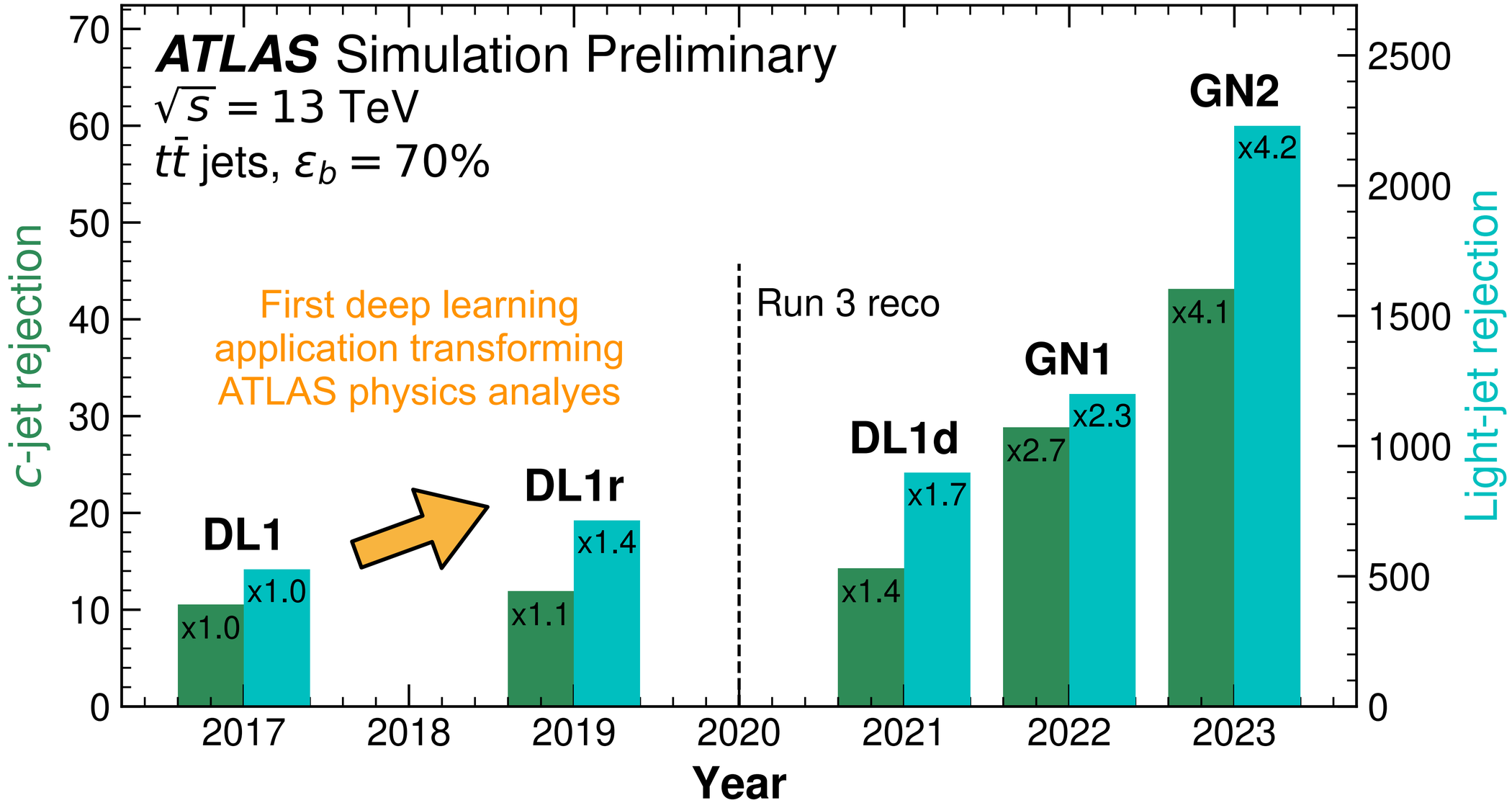
ATLAS Simulation Preliminary
$\sqrt{s} = 13$ TeV
$t\bar{t}$ jets

Truth *b*-jet
$p_T = 134.1$ GeV

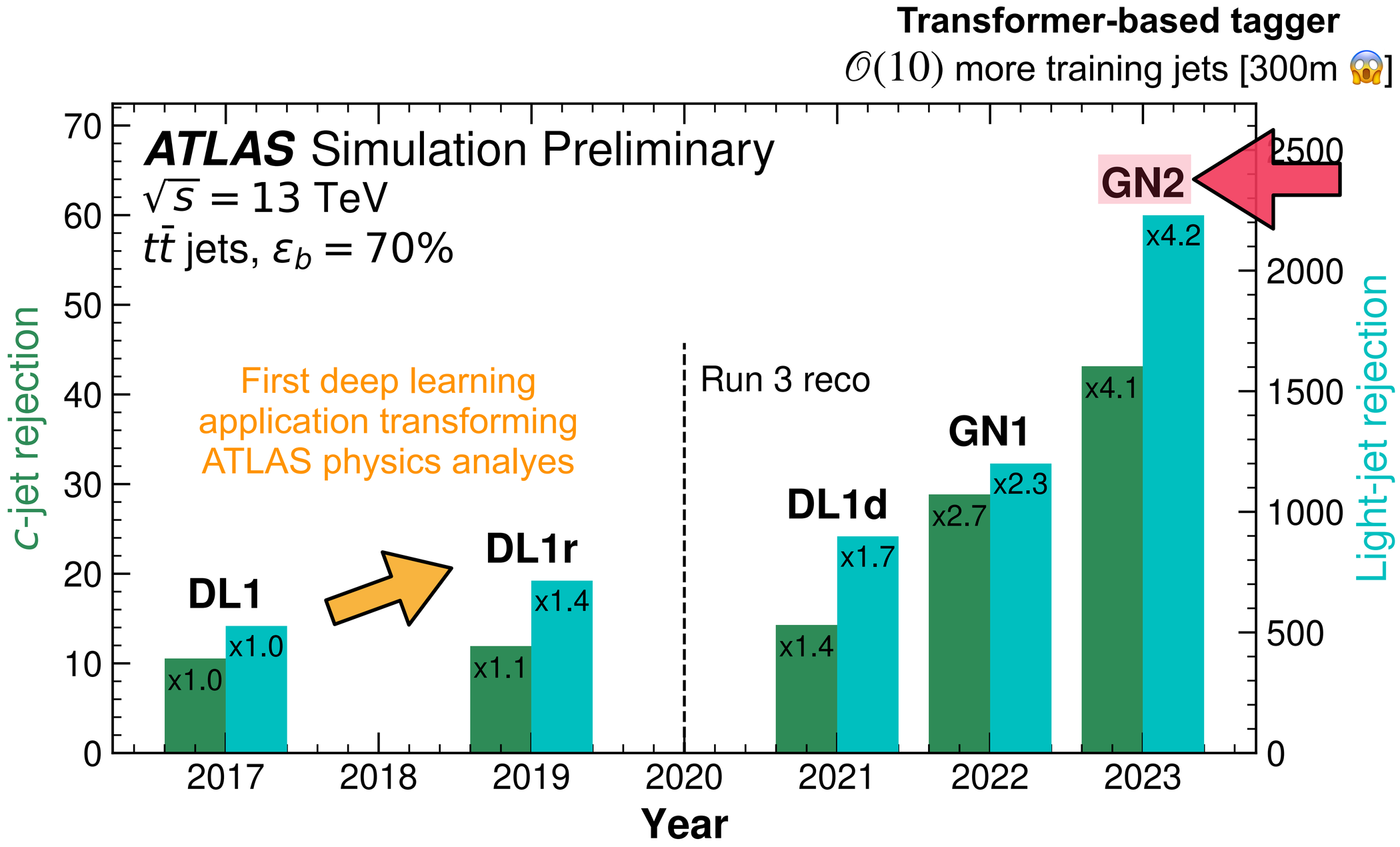
$p_b = 0.995$
$p_c = 0.005$
$p_u = 0.000$



FTAG-2023-01

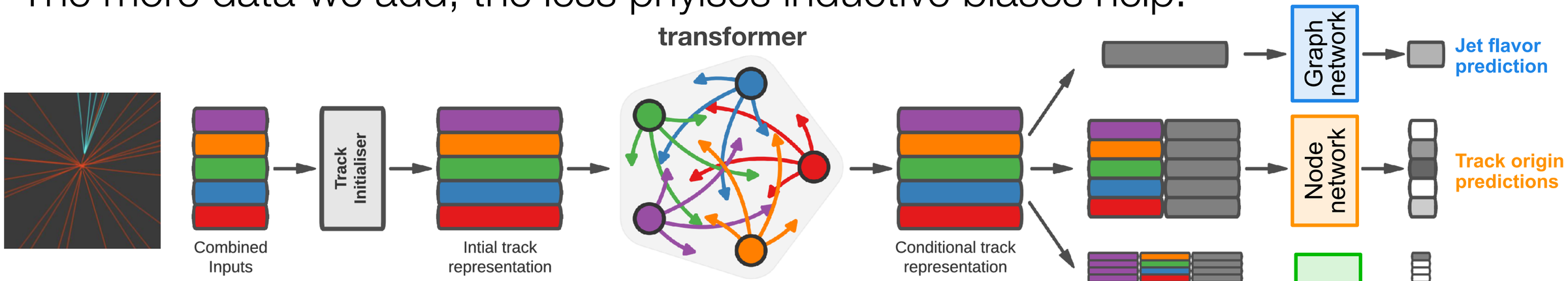

ATLAS Simulation Preliminary
√s = 13 TeV
tt̄ jets, ε_b = 70%
First deep learning application transforming ATLAS physics analyes
Run 3 reco
DL1
DL1r
DL1d
GN1
GN2
x1.0
x1.0
x1.1
x1.4
x1.4
x1.7
x2.7
x2.3
x4.1
x4.2
c-jet rejection
Light-jet rejection
Year
0
10
20
30
40
50
60
70
0
500
1000
1500
2000
2500
2017
2018
2019
2020
2021
2022
2023

FTAG-2023-01


Transformer-based tagger
$\mathcal{O}(10)$ more training jets [300m 😱]
ATLAS Simulation Preliminary
$\sqrt{s} = 13$ TeV
$t\bar{t}$ jets, $\varepsilon_b = 70\%$
First deep learning application transforming ATLAS physics analyes
Run 3 reco
DL1
DL1r
DL1d
GN1
GN2
x1.0
x1.0
x1.1
x1.4
x1.4
x1.7
x2.7
x2.3
x4.1
x4.2
c-jet rejection
Light-jet rejection
Year
2017
2018
2019
2020
2021
2022
2023
0
10
20
30
40
50
60
70
0
500
1000
1500
2000
2500

More data, less physics

The more data we add, the less phyiscs inductive biases help.



$$\mathscr{L}_{tot} = \mathscr{L}_{jet} + \alpha \mathscr{L}_{trk} + \beta \mathscr{L}_{vtx}$$

GN1 (2021): helped a lot

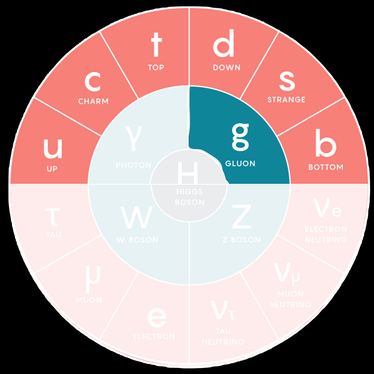
- 30m training jets
- 100% improvement

ATL-PHYS-PUB-2022-027

GN2 (2023): help a little

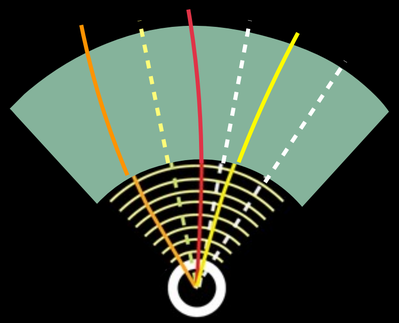
- 300m training jets
- 15% improvement

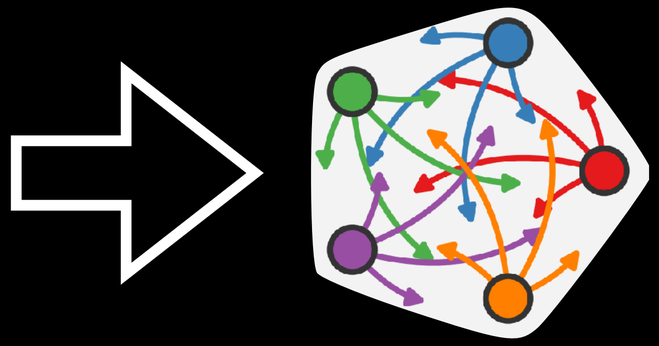
2505.19689



Classify b-jets

GN2



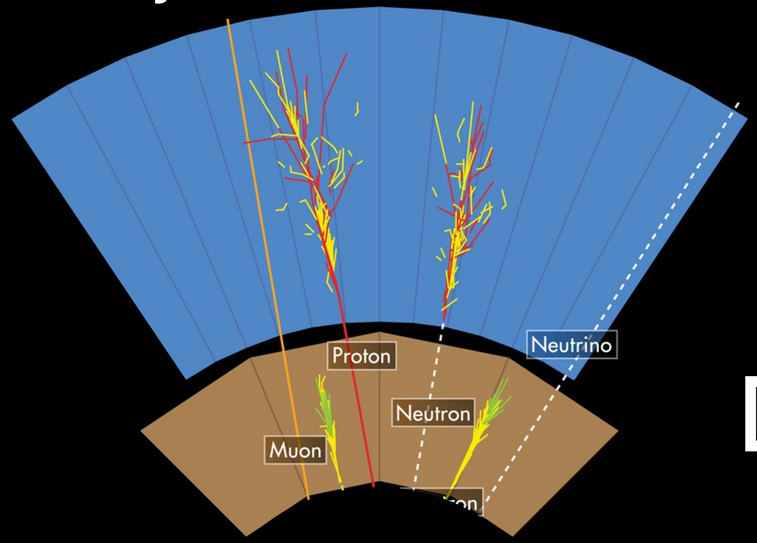


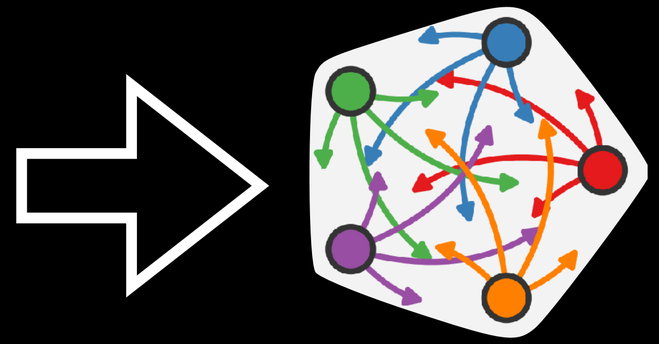
b

c

light
(ud,s,gluon)

Classify other jets



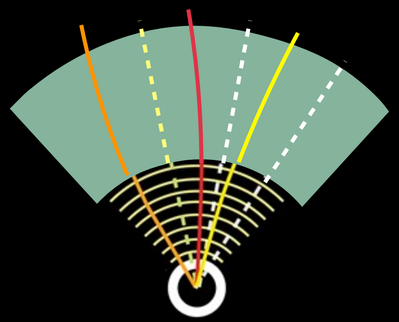


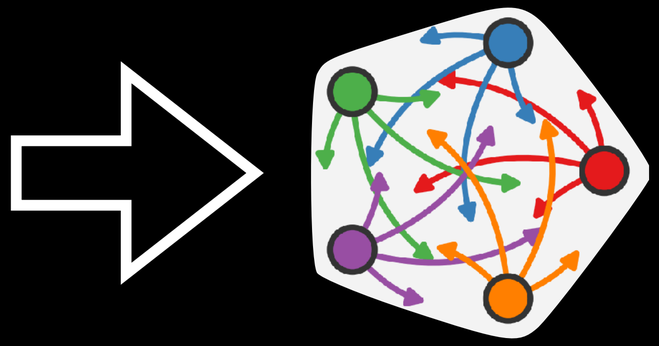
q

g

Classify b-jets

GN2

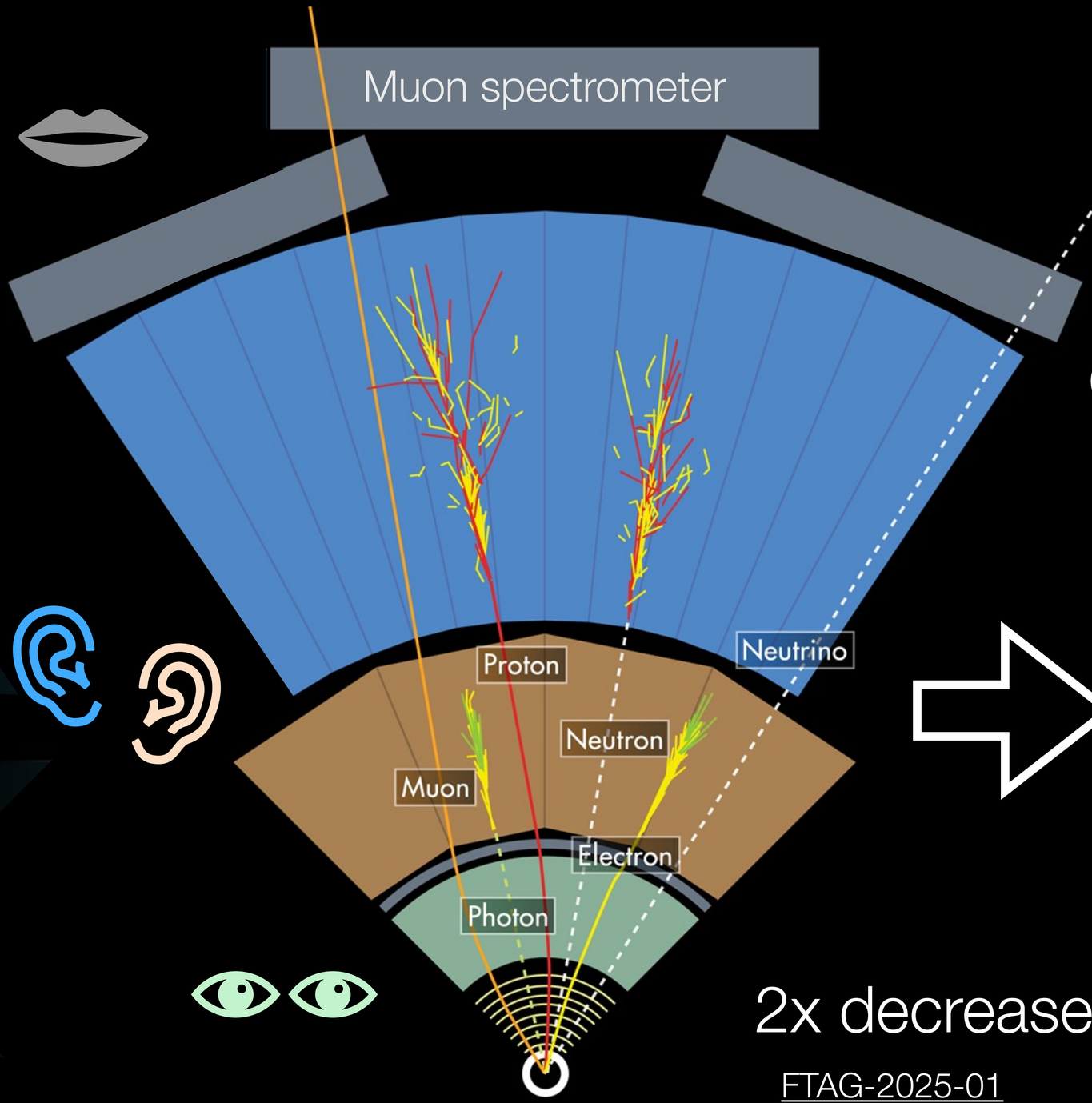


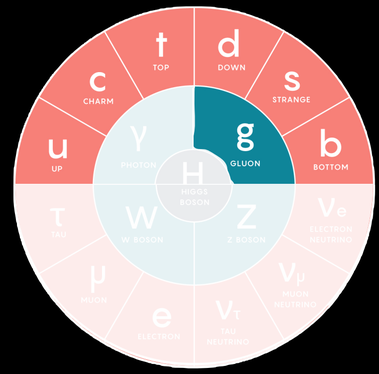


b

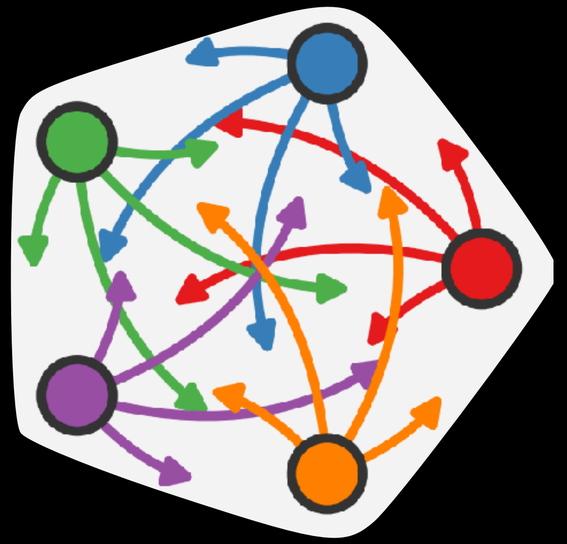
c

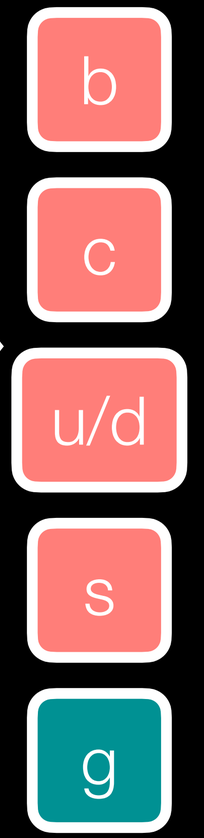
light
(ud,s,gluon)





GN3: multi-modal, multi-task

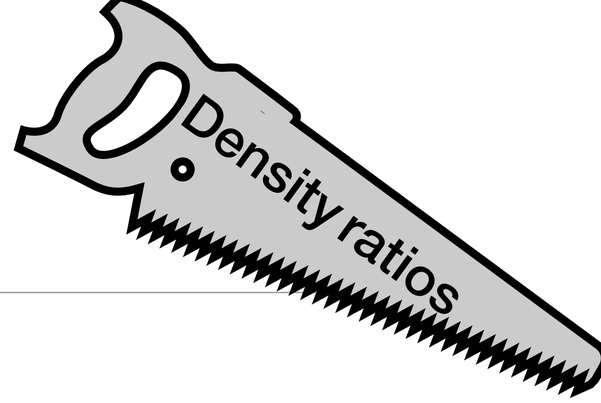




2x decrease in bkg mistag rate 😱

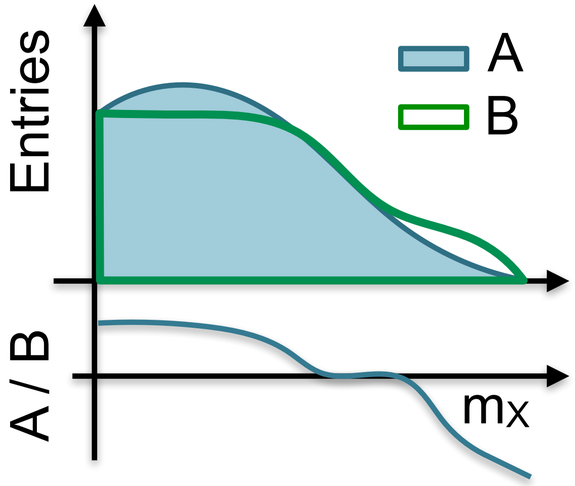
FTAG-2025-01

Density ratio intro



Generalizes the ratio of two histograms to high dimensions

Histograms ⇒ **Neural density ratio estimation**



$$\mathrm{NN}_\Theta(x) = \frac{p_A(x)}{p_B(x)}, \quad x \in \mathbb{R}^d$$

Loss: binary cross entropy

$$\mathscr{L} = -\sum_{x_i \in A} \log \mathrm{NN}_\theta(x_i) - \sum_{x_i \in B} \log\left(1 - \mathrm{NN}_\theta(x_i)\right)$$

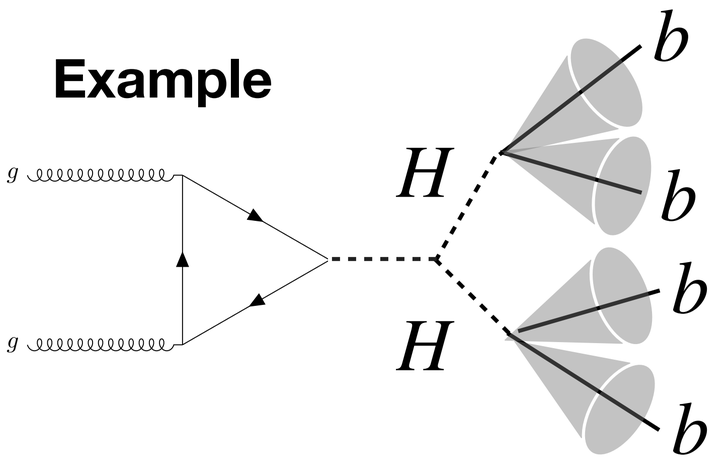
Background estimation

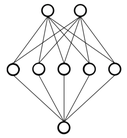
1911.00405
2202.07288
2301.03212

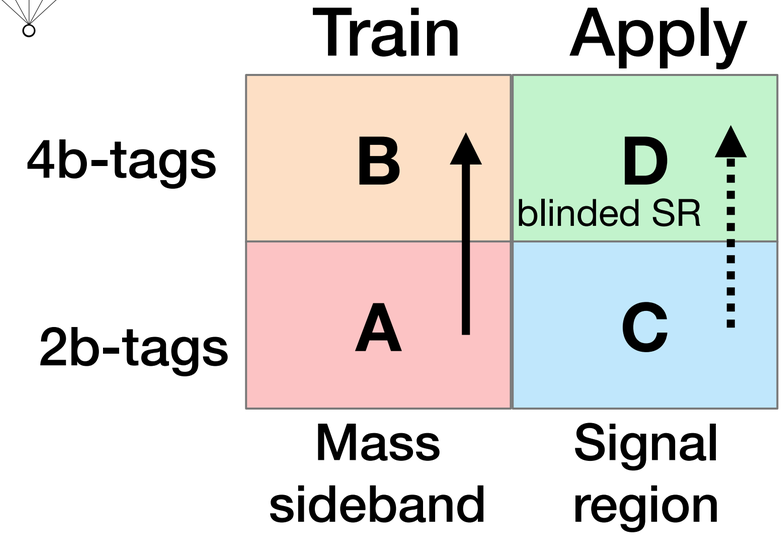
Density ratios

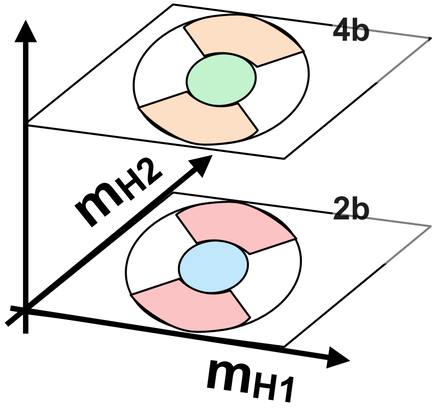
Generalized ABCD method

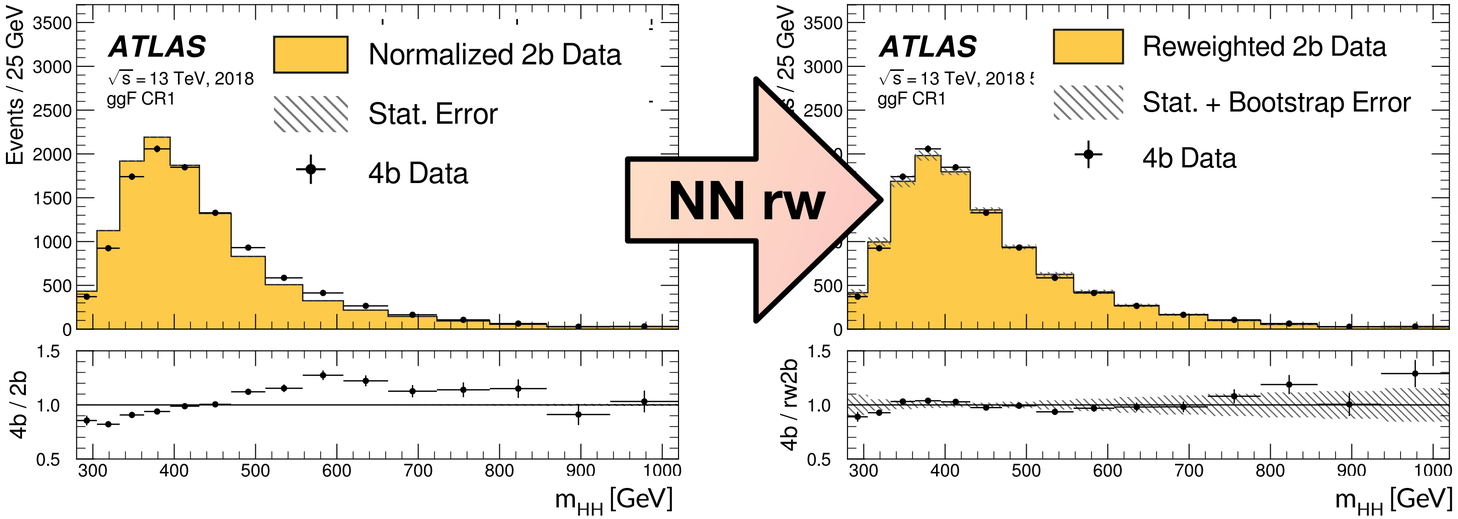
Example











$$p_{4b} = w(x) \cdot p_{2b}(x)$$

Key assumption:

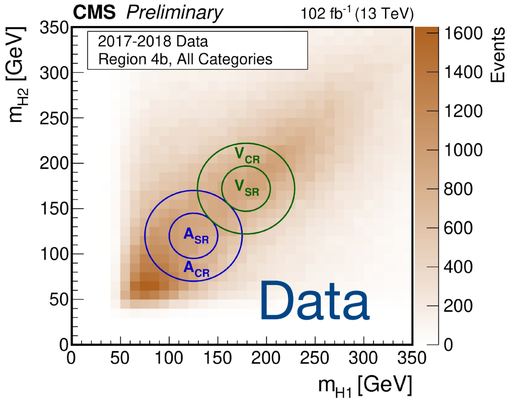
$w_{CR}(X)$ is a valid ≈ of $w_{SR}(X)$.

Wide spread adoption

Density ratios

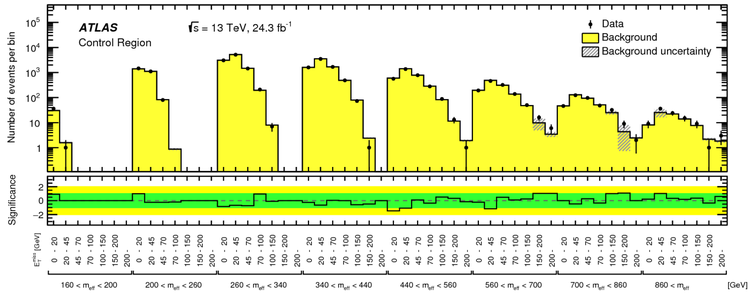
CMS HH4b NR

BDT rw: 2202.09617 + Higgs Pairs 2022 talk



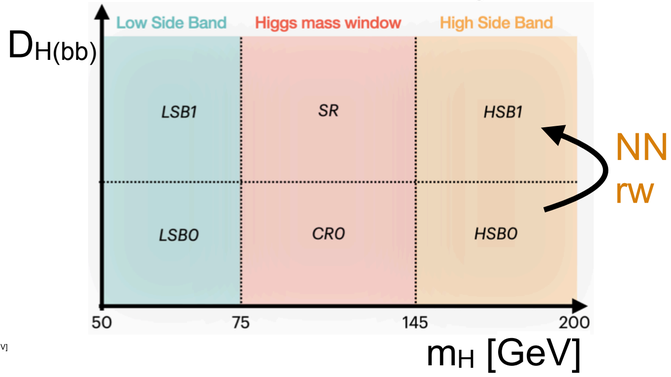
HH4b + MET

BDT rw 1806.04030 and ATLAS-CONF-2023-048



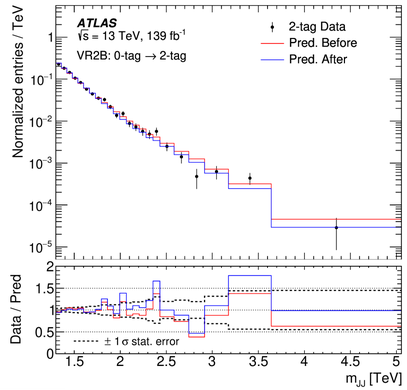
Y -> XH 2306.03637

NN rw for anomalous jets



VH(qqbb)

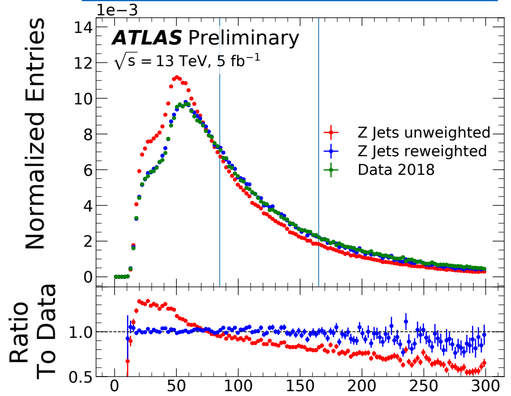
BDT 2007.05293



gg tagger

Match MC -> data

ATL-PHYS-PUB-2021-027



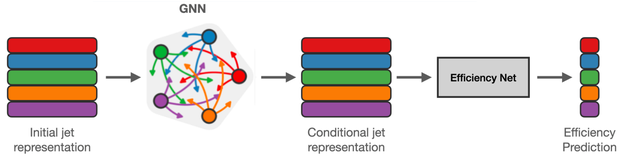
Search for generic resonances

MC -> data reweighting in sideband

2502.09770

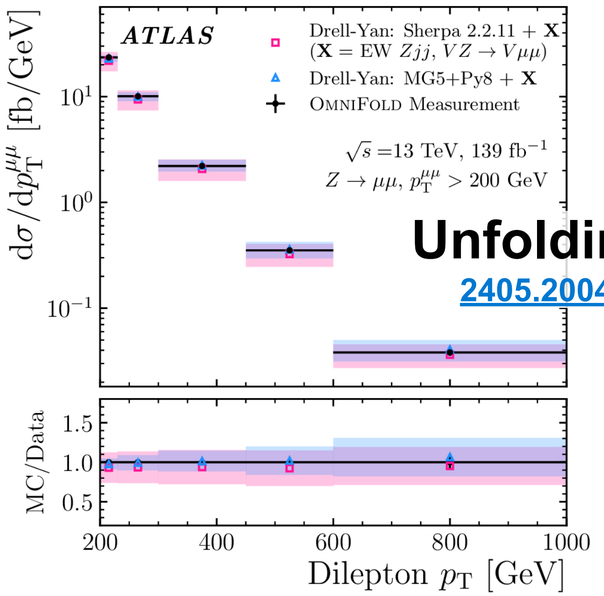
FTAG truth -> reco

ATL-PHYS-PUB-2022-041



Unfolding Z+jets

2405.20041; Mariel's talk



+ *others!*

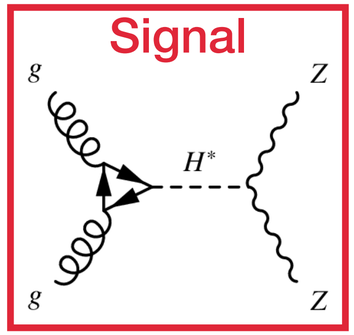
Method 2412.01600
result 2412.01548

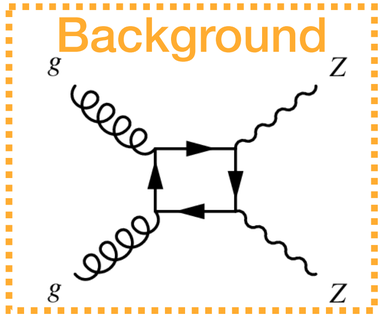
NEW!

Simulation-based inference

Density ratios

$$\Gamma_H \sim \frac{\sigma^{off-shell}}{\sigma^{on-shell}}$$



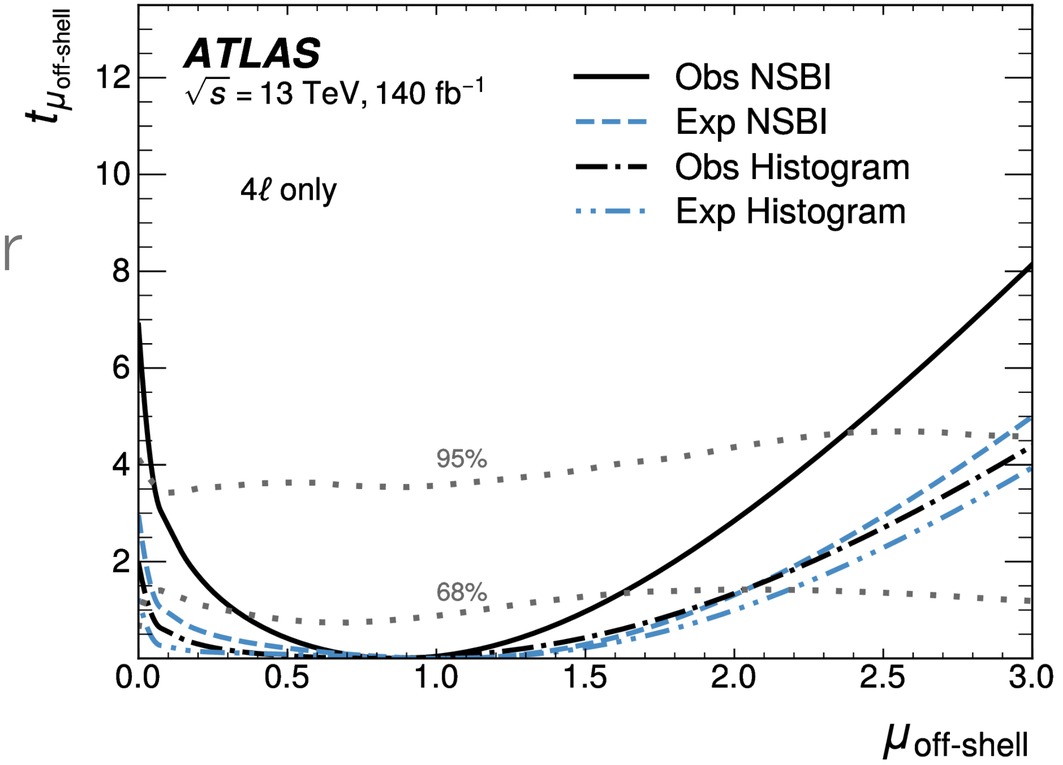


$$P_{off-shell}(\mu) = \mu P_S + \sqrt{\mu} P_I + P_B$$

Kinemmatics depend on μ… optimal classifier changes for each μ.

$$\frac{p(x\,|\,\mu)}{p'(x)} = \frac{1}{\sigma(\mu)}\left(\mu\sigma_s\frac{p_S(x)}{p'(x)} + \sqrt{\mu}\sigma_I\frac{p_I(x)}{p'(x)} + \sigma_B\frac{p_B(x)}{p'(x)}\right)$$

Neural Networks

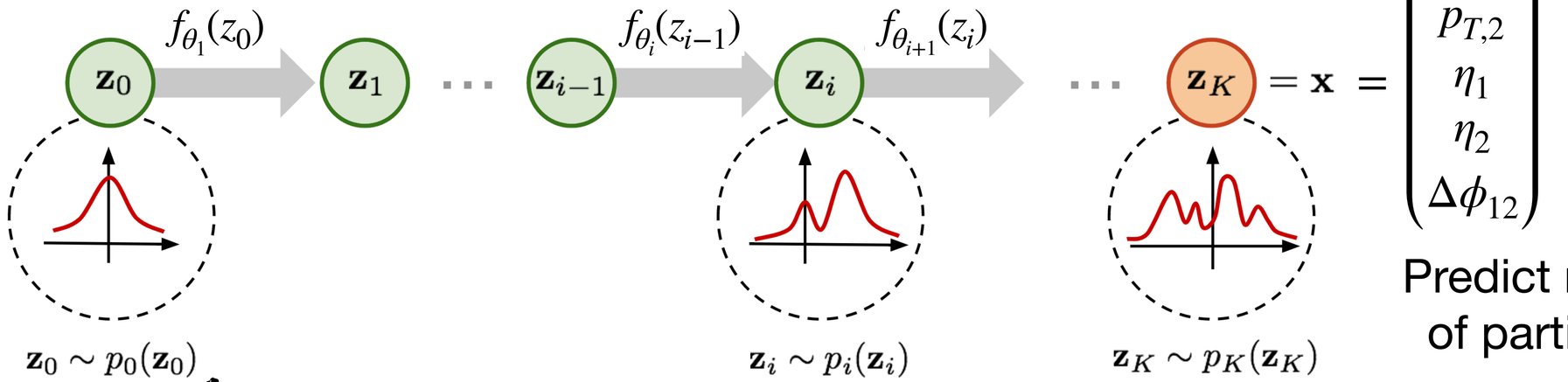


Backup

Learn $p_\phi(x)$

blog post

Gaussian,
e.g, $z \in \mathbb{R}^5$



$$\mathbf{x} = \begin{pmatrix} p_{T,1} \\ p_{T,2} \\ \eta_1 \\ \eta_2 \\ \Delta\phi_{12} \end{pmatrix}$$

Predict momenta of particles 1,2

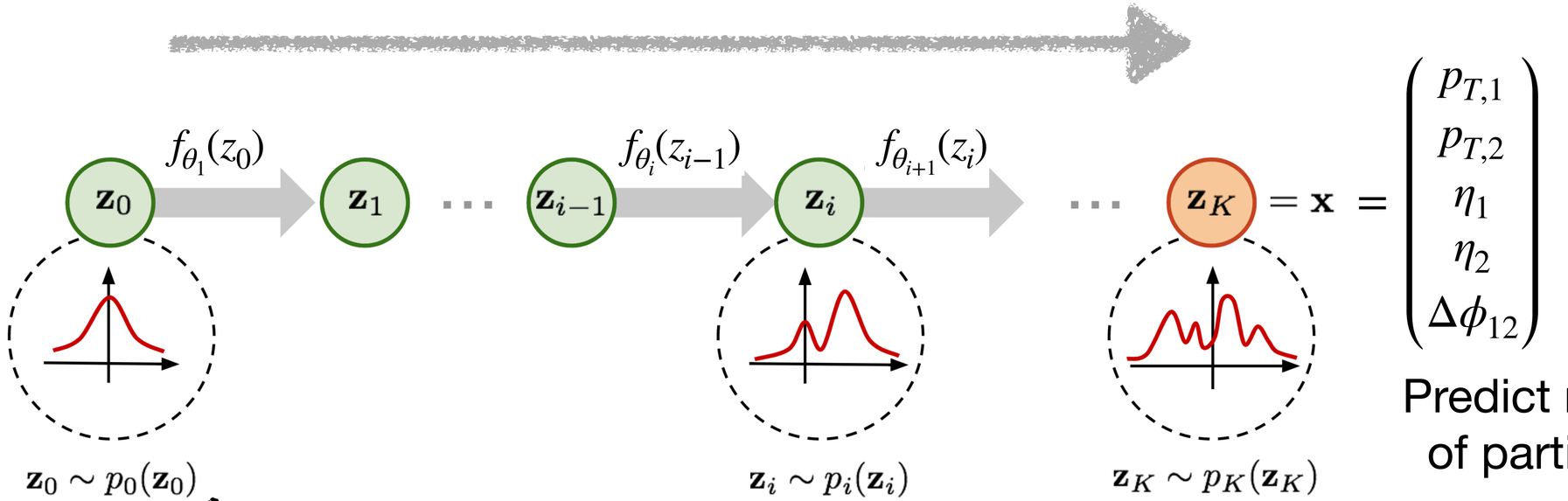
$$f_\phi^{-1} = f_{\phi_1}^{-1} \circ f_{\phi_2}^{-1} \circ \cdots \circ f_{\phi_L}^{-1}$$

Invertible f_{ϕ_i} for density of training samples

Learn $p_\phi(x)$

blog post

Gaussian,
e.g, $z \in \mathbb{R}^5$



$$\mathbf{x} = \begin{pmatrix} p_{T,1} \\ p_{T,2} \\ \eta_1 \\ \eta_2 \\ \Delta\phi_{12} \end{pmatrix}$$

Predict momenta of particles 1,2

$$f_\phi^{-1} = f_{\phi_1}^{-1} \circ f_{\phi_2}^{-1} \circ \cdots \circ f_{\phi_L}^{-1}$$

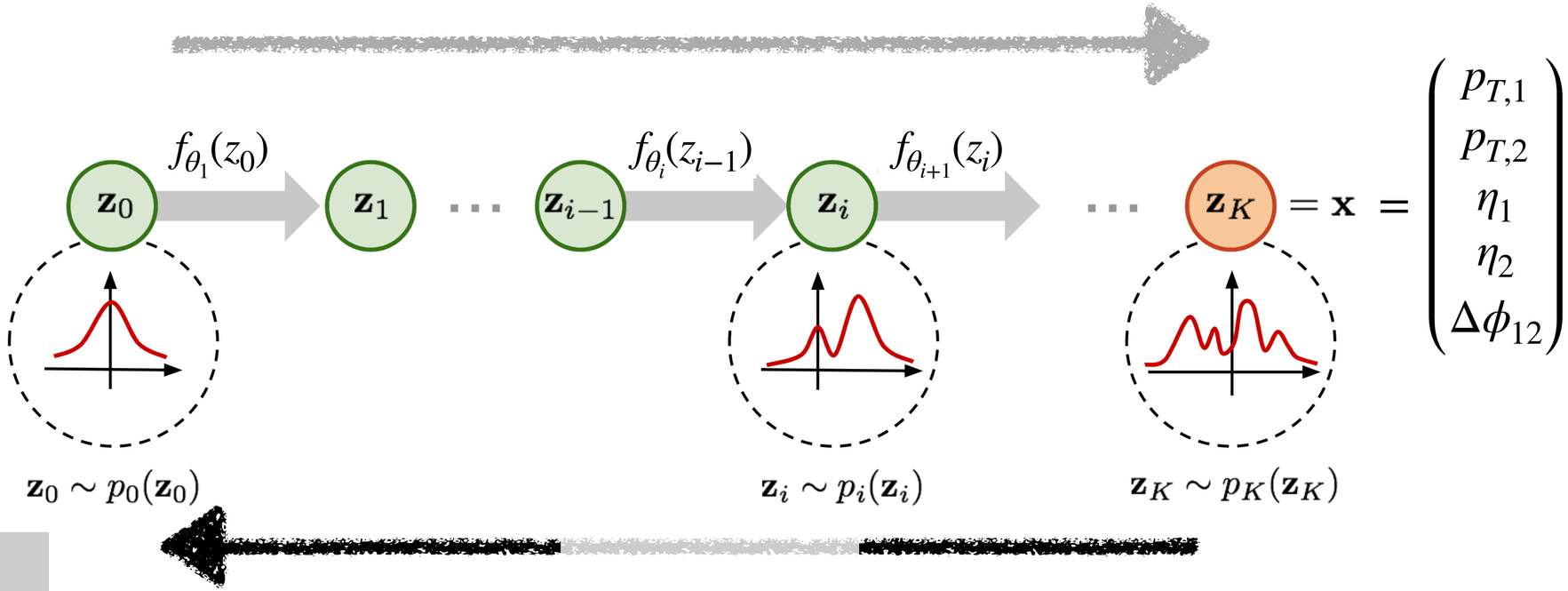
Invertible f_{ϕ_i} for density of training samples

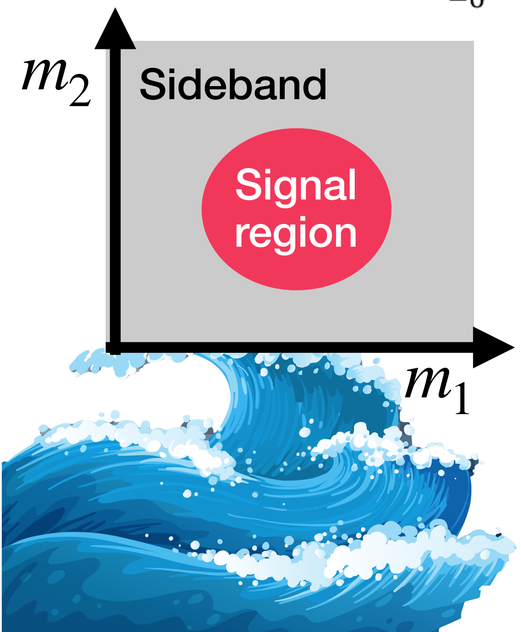
$$\mathcal{L}oss = -\log p_\phi(x) = -\log p_z(f_\phi^{-1}(x)) - \sum_{i=1}^{K} \left| \frac{\partial f_{\phi_i}^{-1}}{\partial z_i^T} \right|$$

blog post

Learn $p_\phi(\ x \mid y\ )$

Gaussian,
e.g, $z \in \mathbb{R}^5$





$$f_\phi^{-1} = f_{\phi_1}^{-1} \circ f_{\phi_2}^{-1} \circ \cdots \circ f_{\phi_L}^{-1}$$

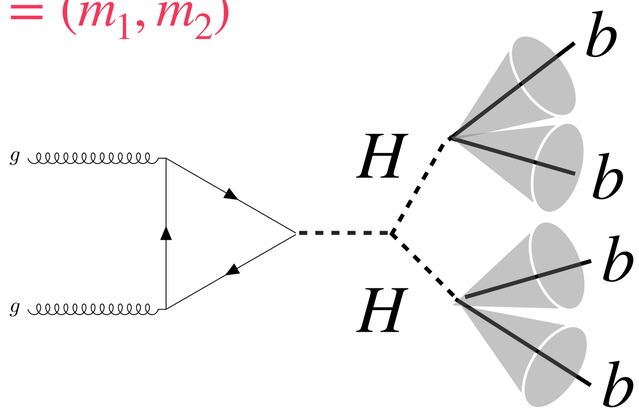
Conditional density: $p_\theta(x \mid y),\ y = (m_1, m_2)$

$$\mathcal{L}oss = -\log p_\phi(x \mid y) = -\log p_z(f_\phi^{-1}(x \mid y)) - \sum_{i=1}^{K} \left| \frac{\partial f_{\phi_i}^{-1}}{\partial z_i^T} \right|$$

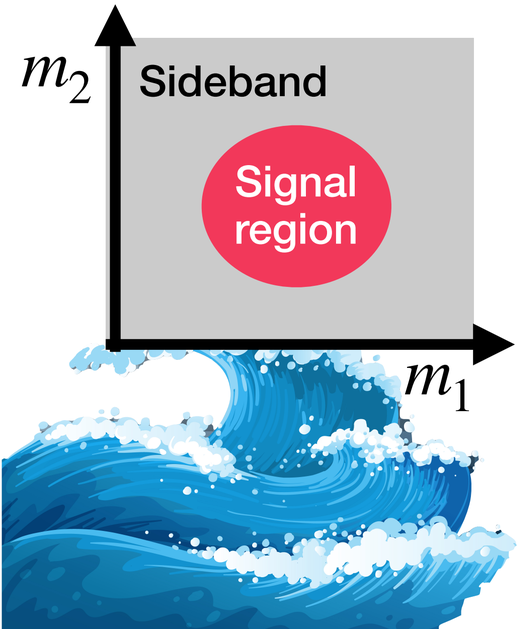
NH thesis

Learn $p_\phi(\, x \mid y \,)$

$y = (m_1, m_2)$



Learns about preprocessing cut ($\Delta\eta_{HH} < 1.5$)



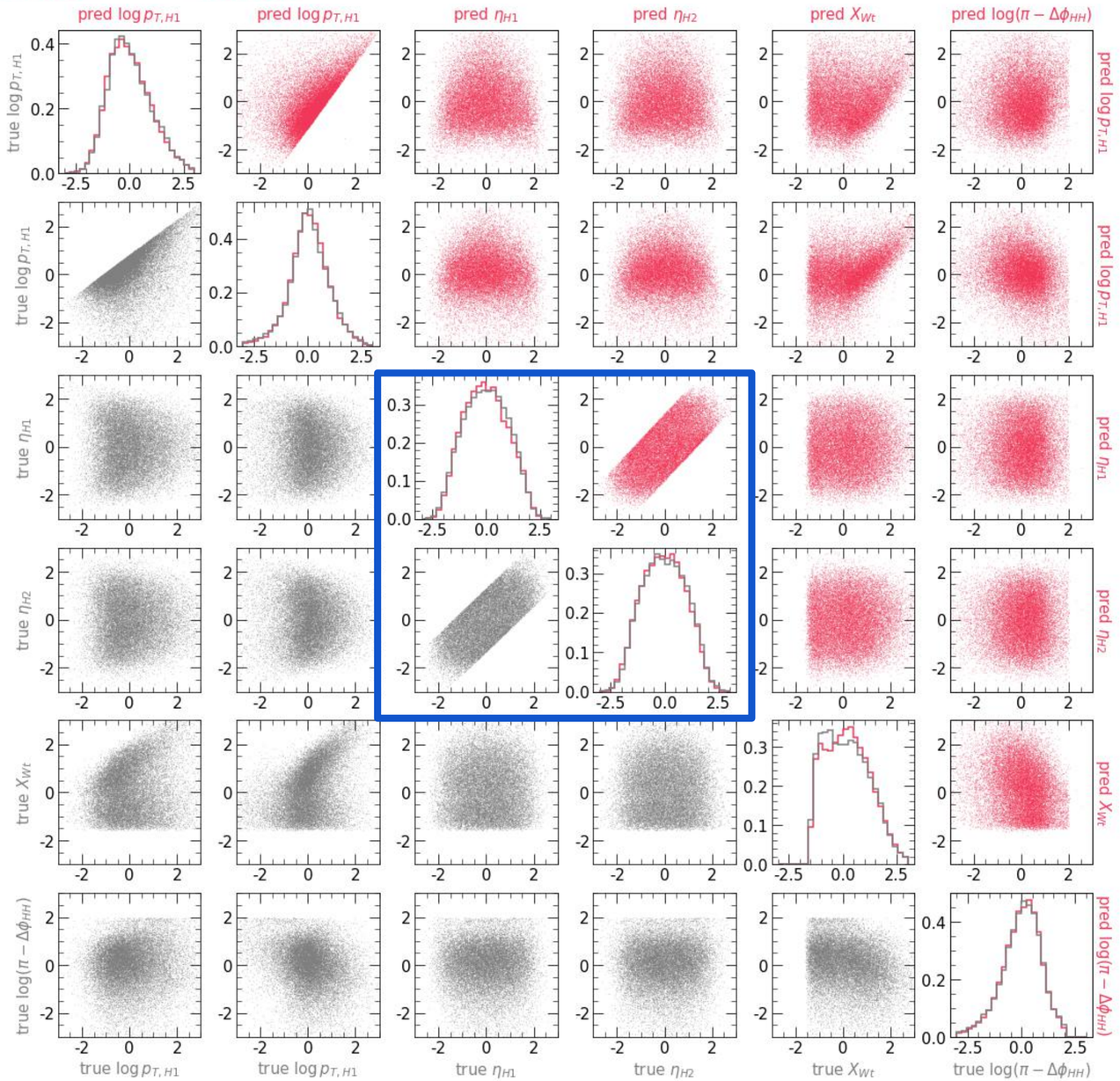
data

ATLAS Thesis

$\sqrt{s}$ = 13 TeV, 126 fb^{-1}

4b SR

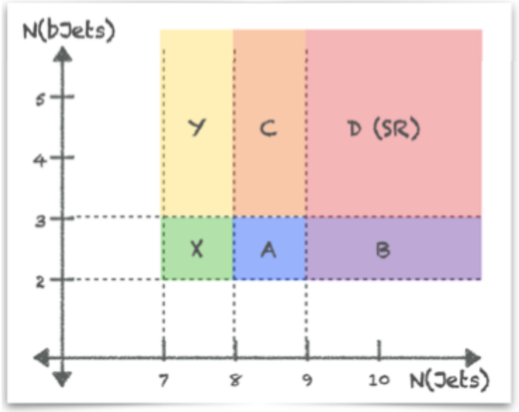
flow pred



Bkg estimation (when we don't trust QCD simulation)

- CMS 4 tops evidence (2023), M. Quinnan's ML4Jets slides
  Methods / pheno papers [1], [2]
- ATLAS genric search, 2502.09770
- CMS genric search, 2412.03747, seminar

Learn how to interpolate into the SR



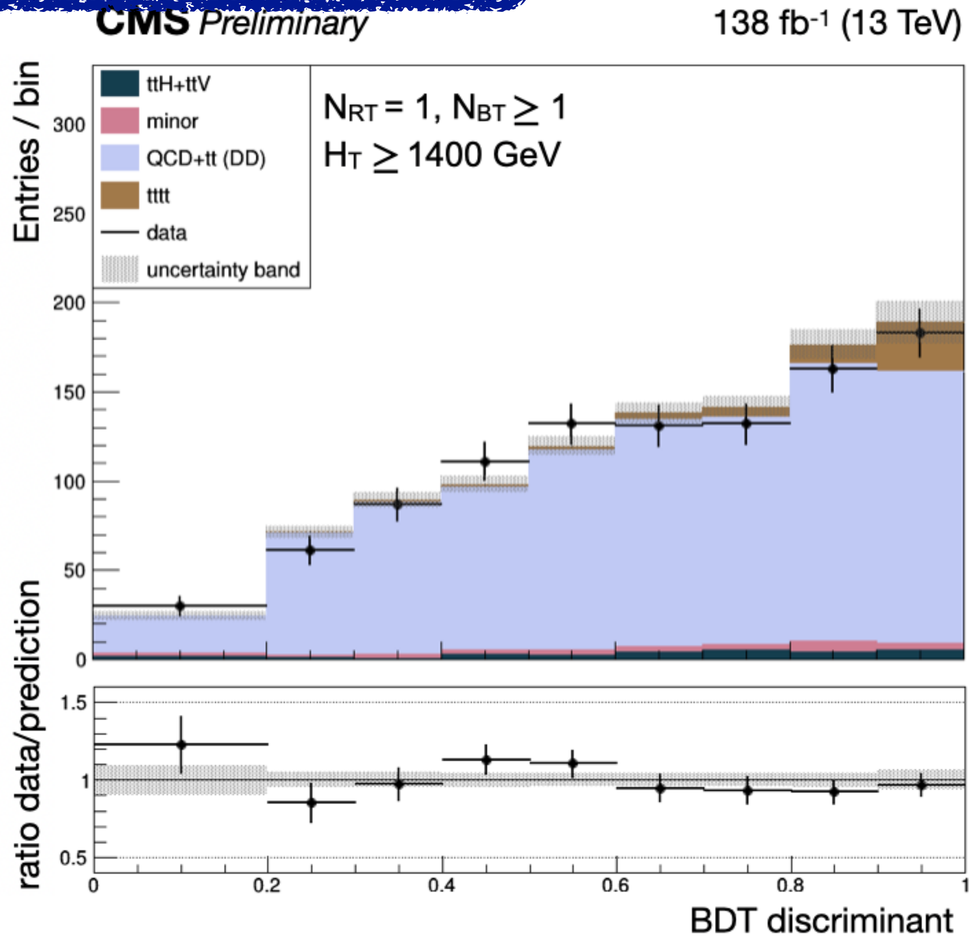
Shape

Based on a **flow**:

$t\bar{t}$ sim → $t\bar{t}$ +QCD data

NEW

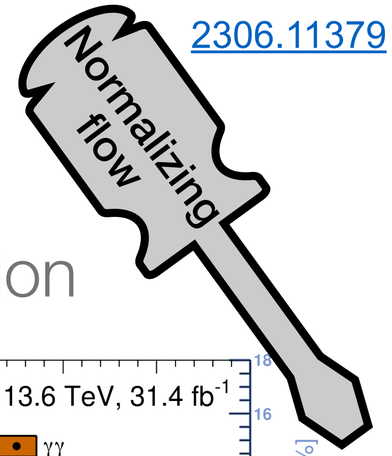
First public result with LHC data using flows!



Simulation

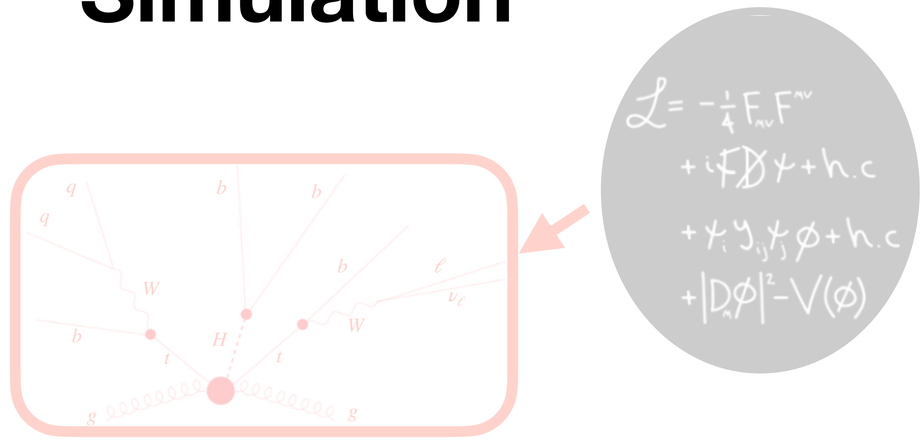
ATLAS H → yy

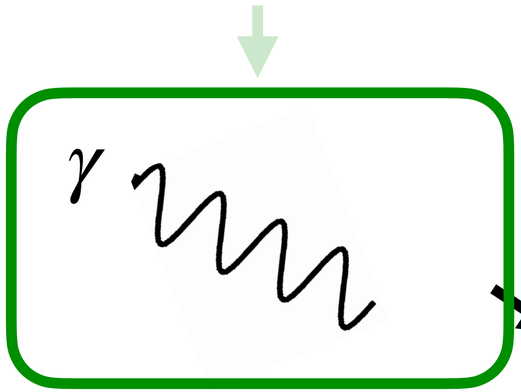
2306.11379

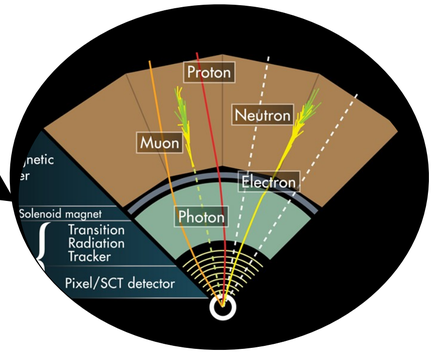


GEANT4 most $$$ part of the simulation

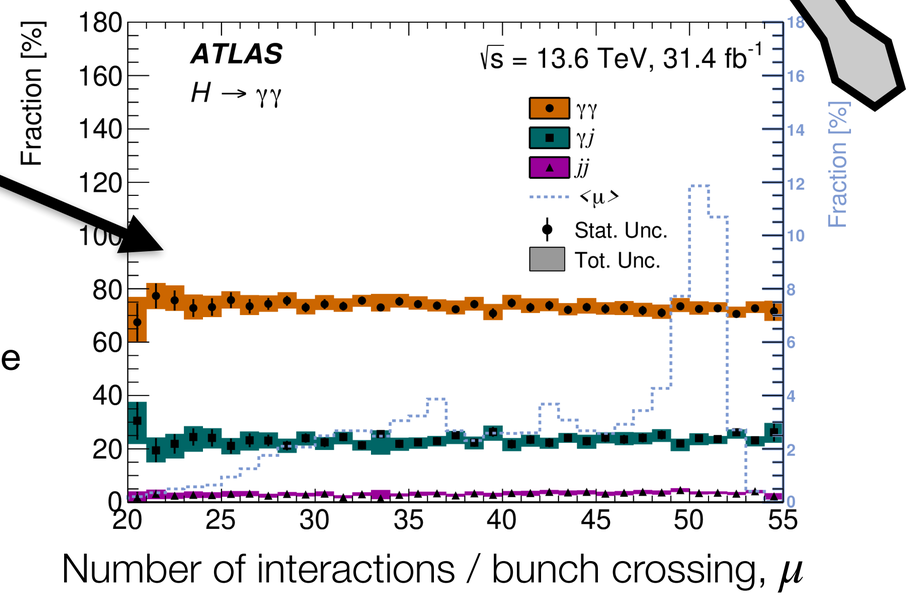
LHCC-G-182





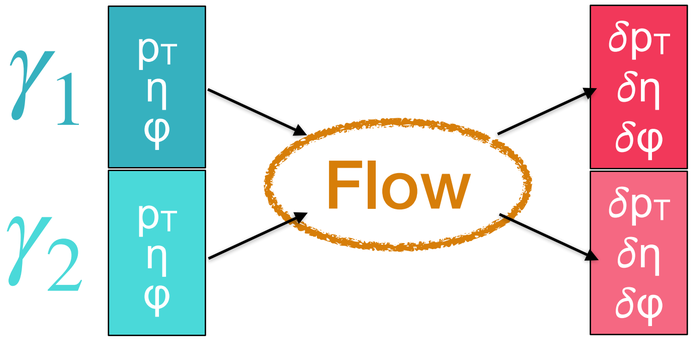


Flow to estimate the $\gamma\gamma$ background



Condition
Truth kinematics

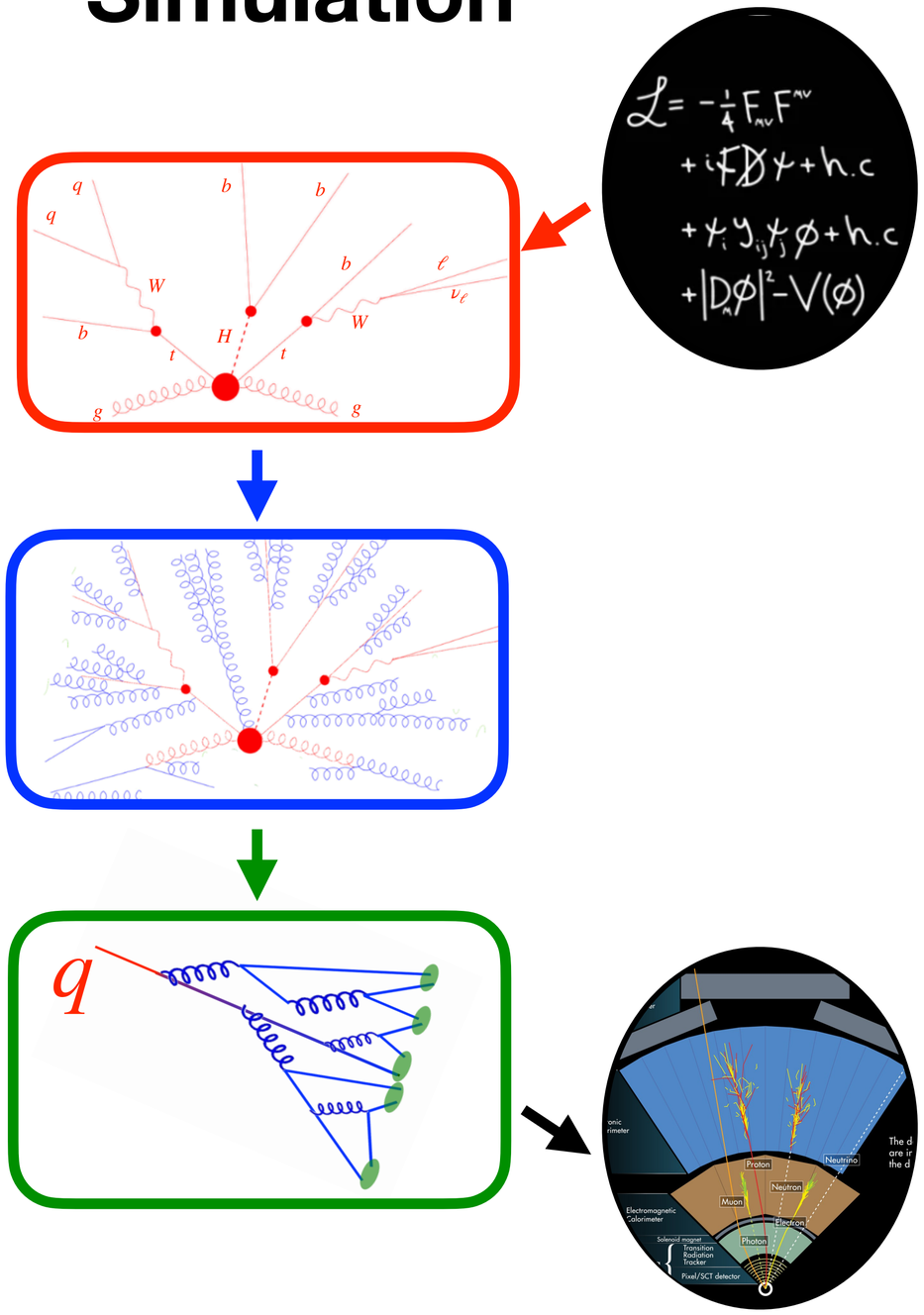
Predict
Detector response



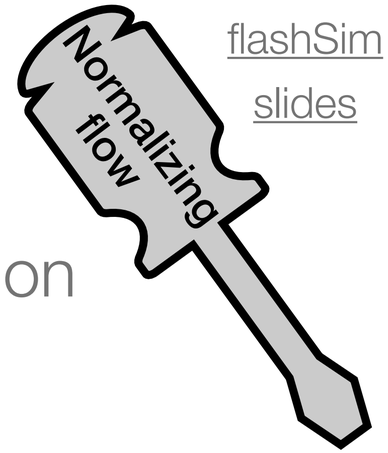
	MC sample size
GEANT4	1.3 ab^{-1}
Flow	3.5 ab^{-1}

2.7x more effective stats!

Simulation

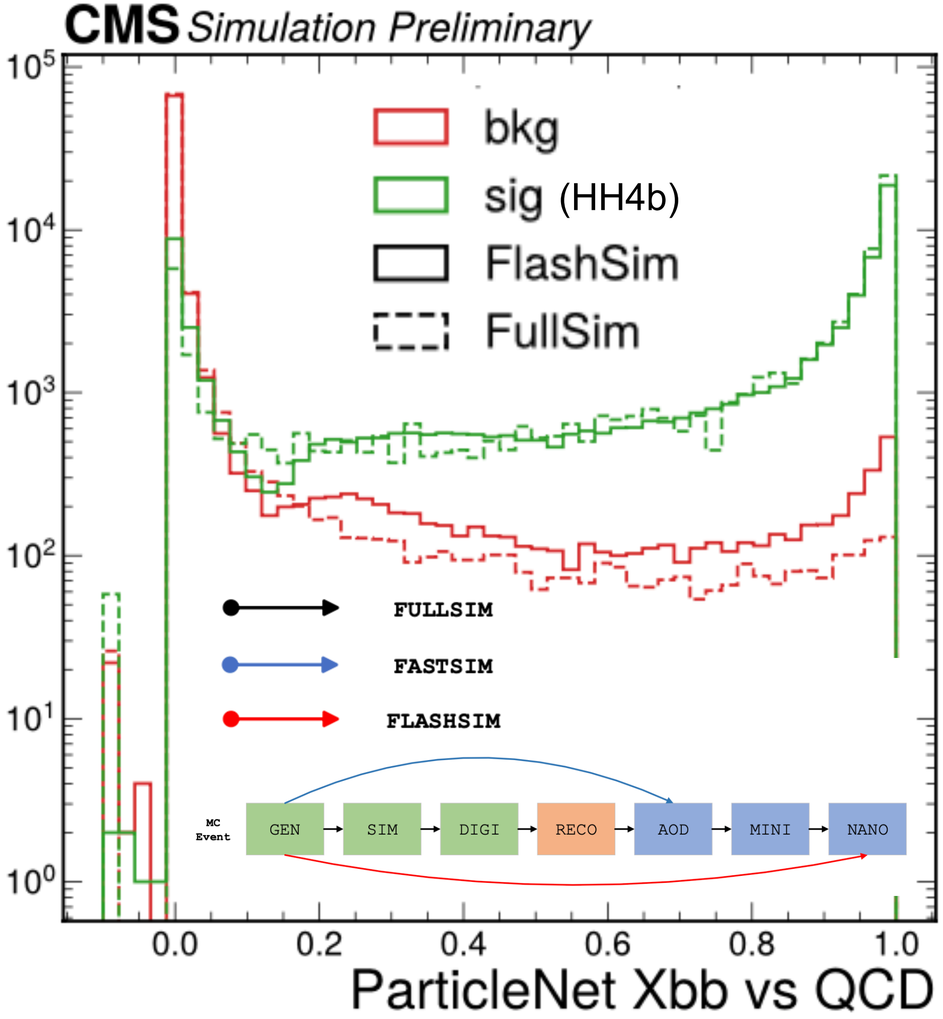


CMS “flash sim”



flashSim slides

GEANT4 most $$$ part of the simulation



1) ML for our tools

2) ML for our future

- Foundation Models
- Reconstruction = HEP foundation model
- End-to-end analysis
- Self supervised learning
- Incorporating domain expertise

Computer vision

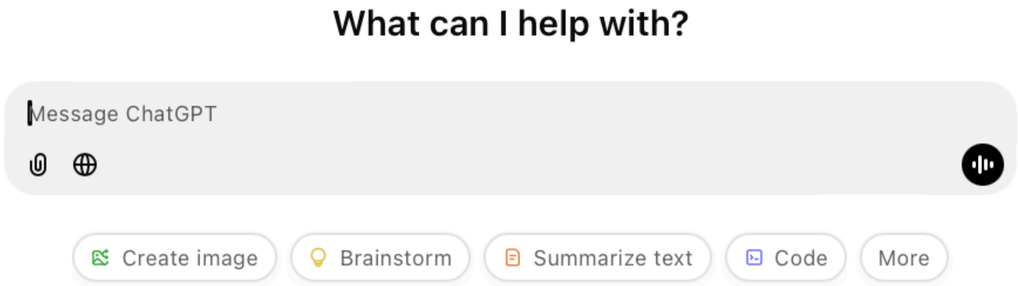
2204.06125

a teddy bear on a skateboard in times square

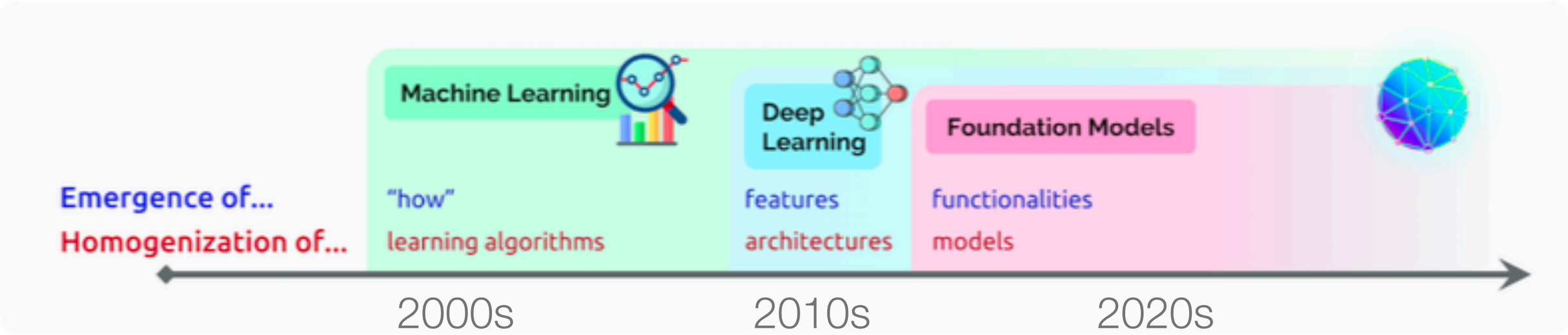
Natural language

$$p_\phi(x \mid y)$$

answer prompt



2108.07258

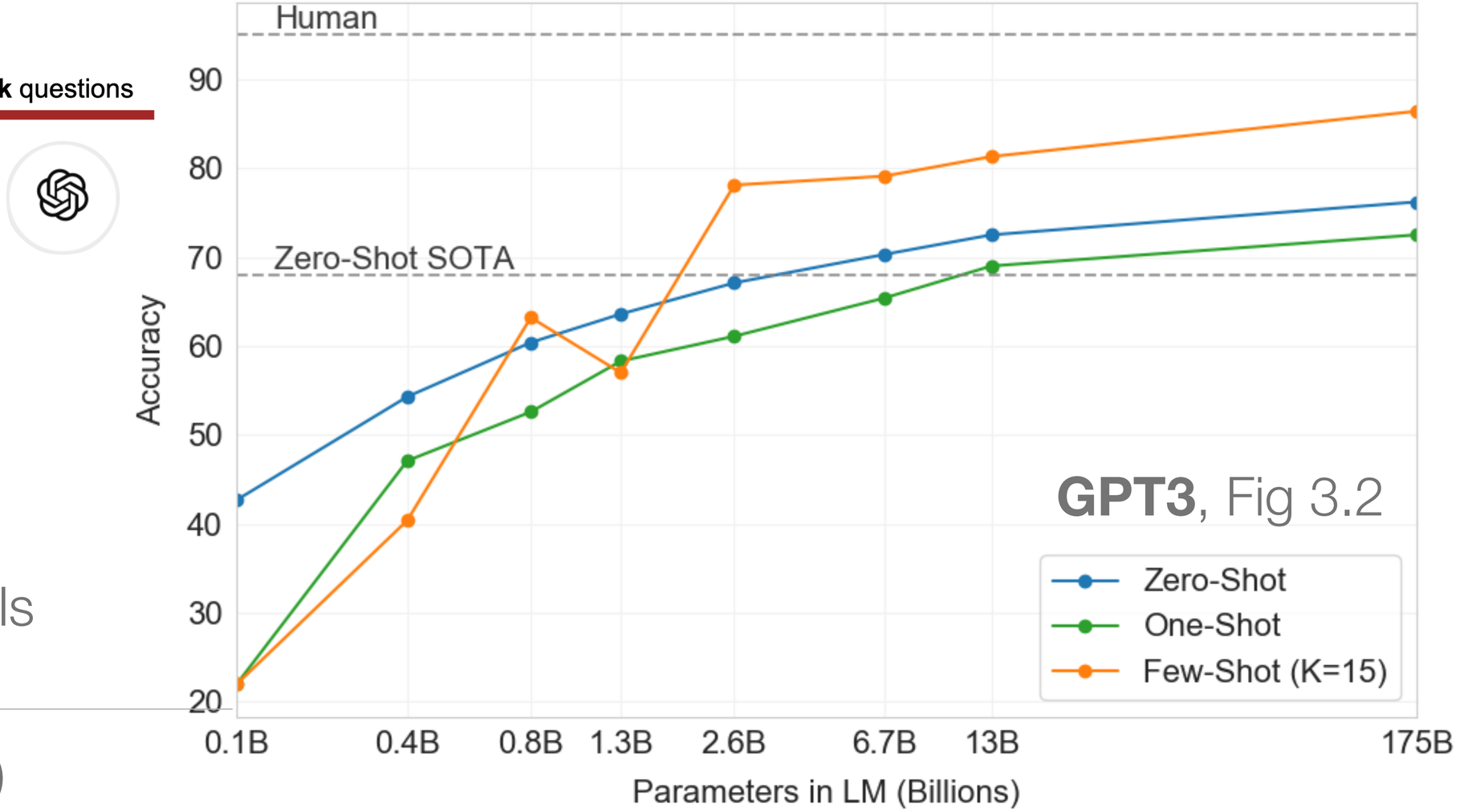


"From a technological point of view, **foundation models** are not new… however, the sheer scale and scope of foundation models from the last few years have stretched our imagination of what is possible."

Tell information

Ask questions

Generative
Pre
Training



GPT3, Fig 3.2

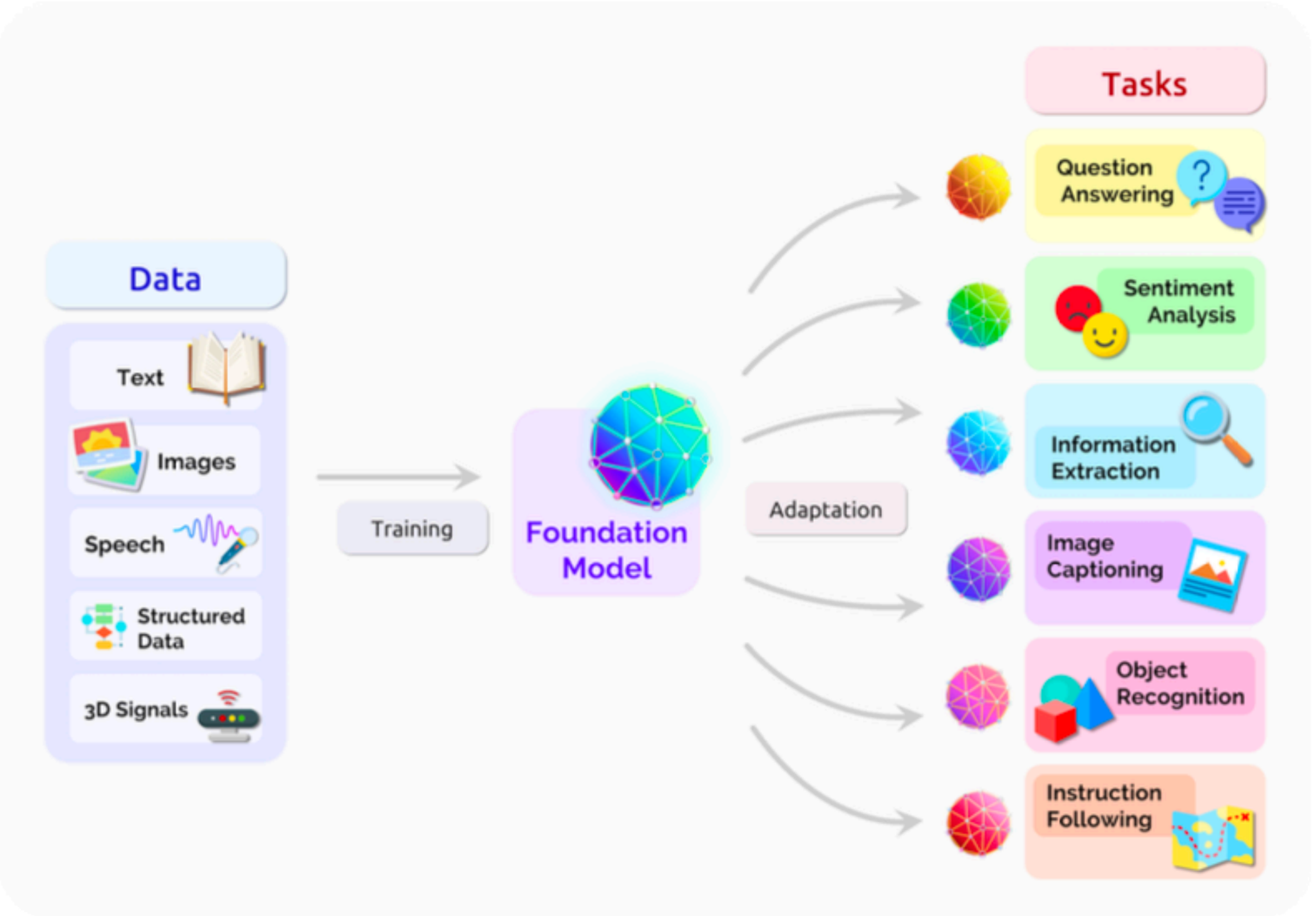
Comparing w/ size of
SOTA jet tagging models

GN2 (2m)
2023

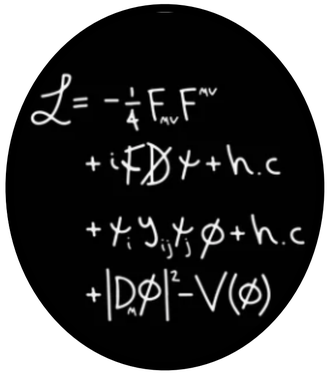
GN3 (13m)
2025

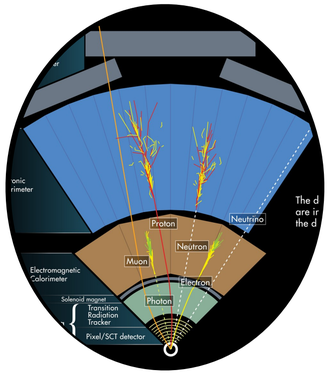
GPT4 . . . ▶
1.8T param

2108.07258

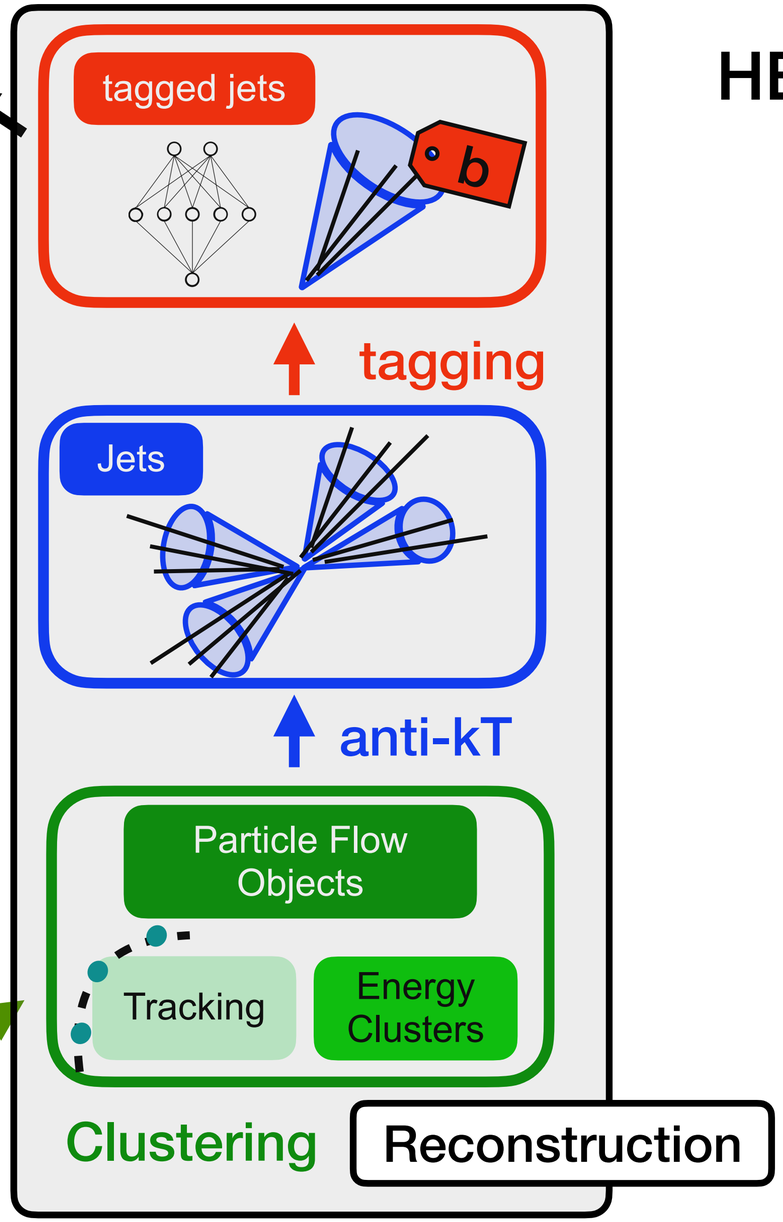


"A **foundation model** is any model that is trained on broad data (generally self-supervision at scale) that can be adapted (e.g, fine-tuned) to a wide range of downstream tasks."

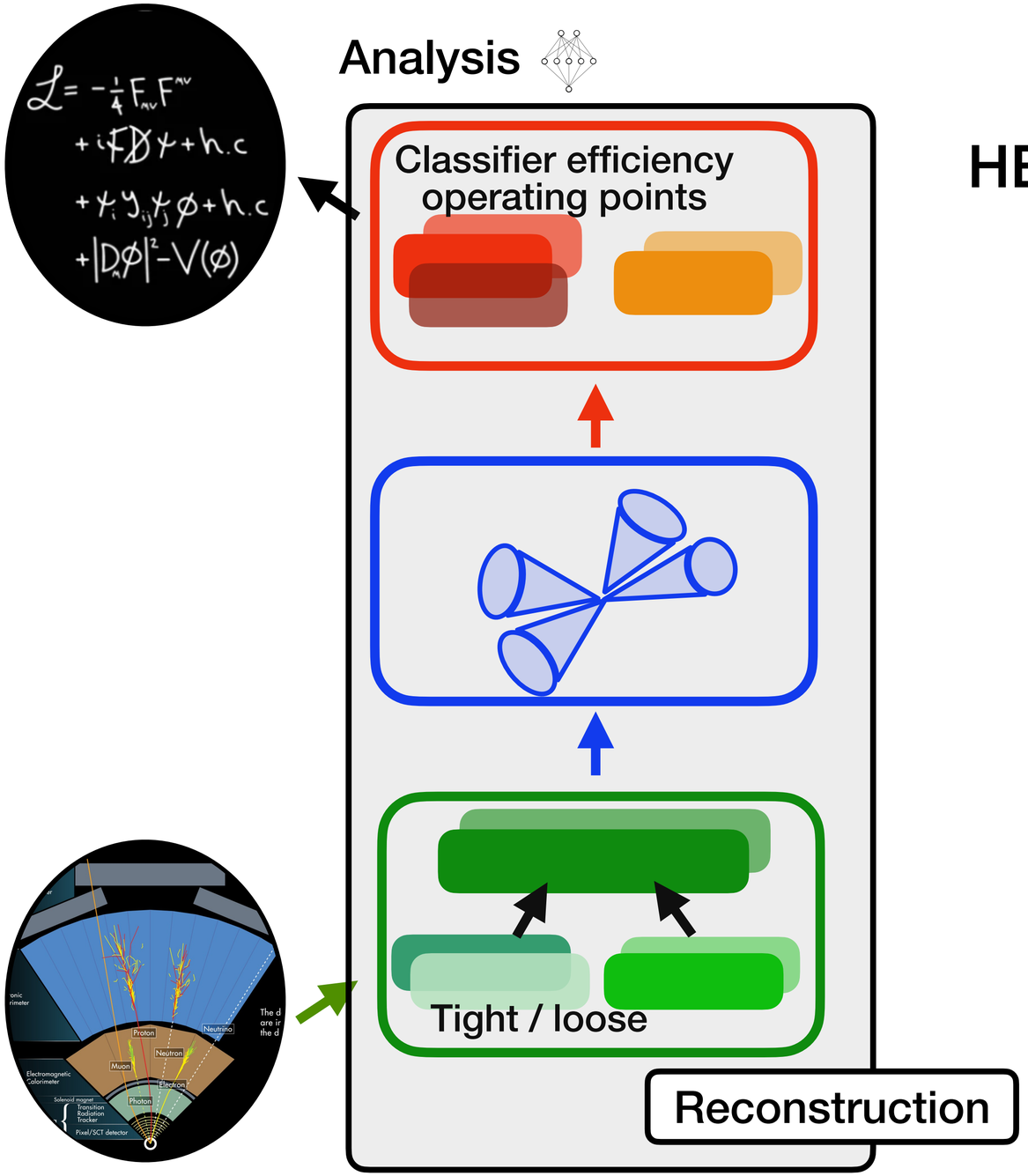




Analysis

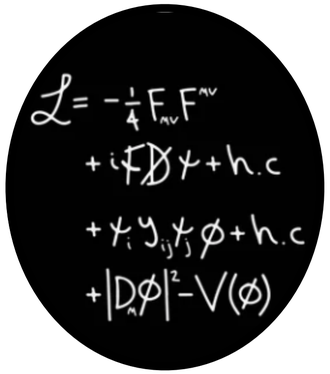


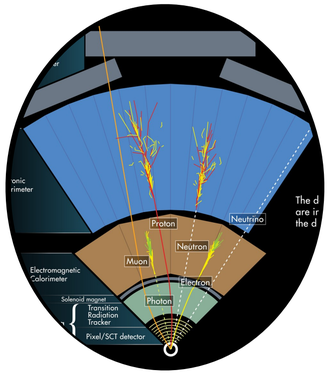
HEP reconstruction optimized on broad data



HEP reconstruction is trained on broad data

Each of analysis chooses their own operating points (grad student descent).





Analysis

Classifier efficiency operating points

Tight / loose

Reconstruction

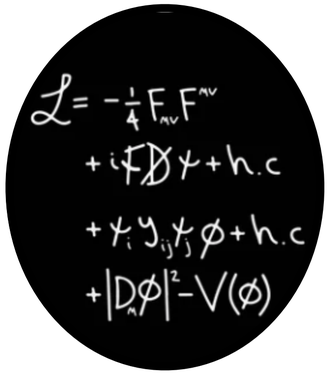
$\nabla_\phi \mathscr{L}$

HEP reconstruction is trained on broad data

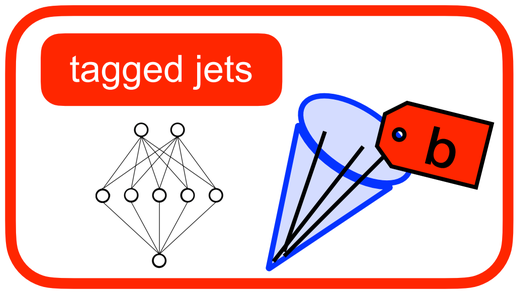
Each of analysis chooses their own operating points (grad student descent).

Foundation models: can we let gradients do our work for us? **“Fine-tuning”**

M. Vigl, **NH**, L. Heinrich
2401.13536

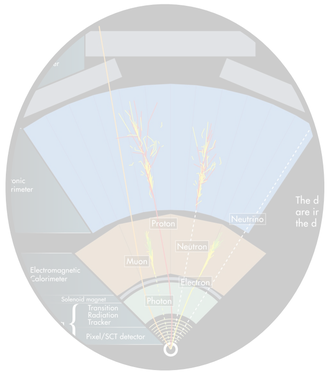


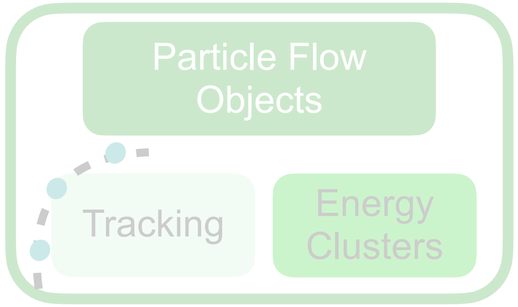
Analysis



tagging

anti-kT

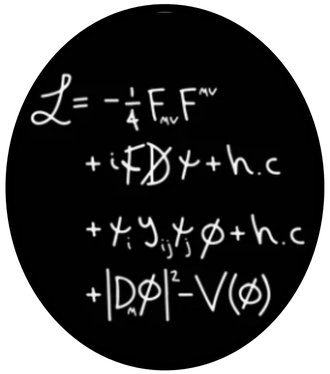




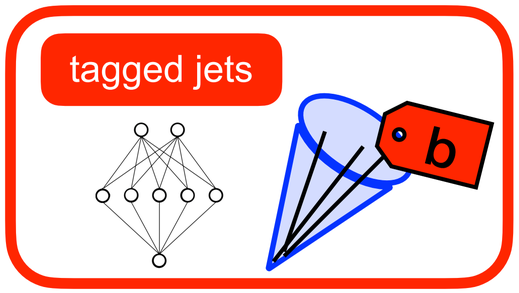
Clustering

Foundation models: can we let gradients do our work for us? **“Fine-tuning”**

M. Vigl, **NH**, L. Heinrich
2401.13536

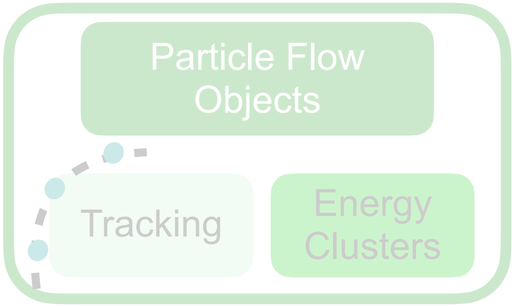


Analysis

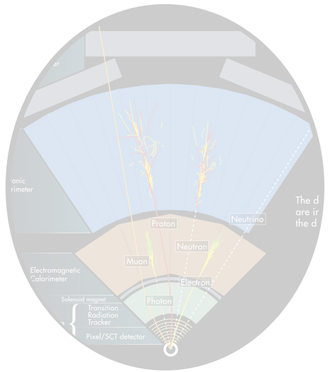


tagging

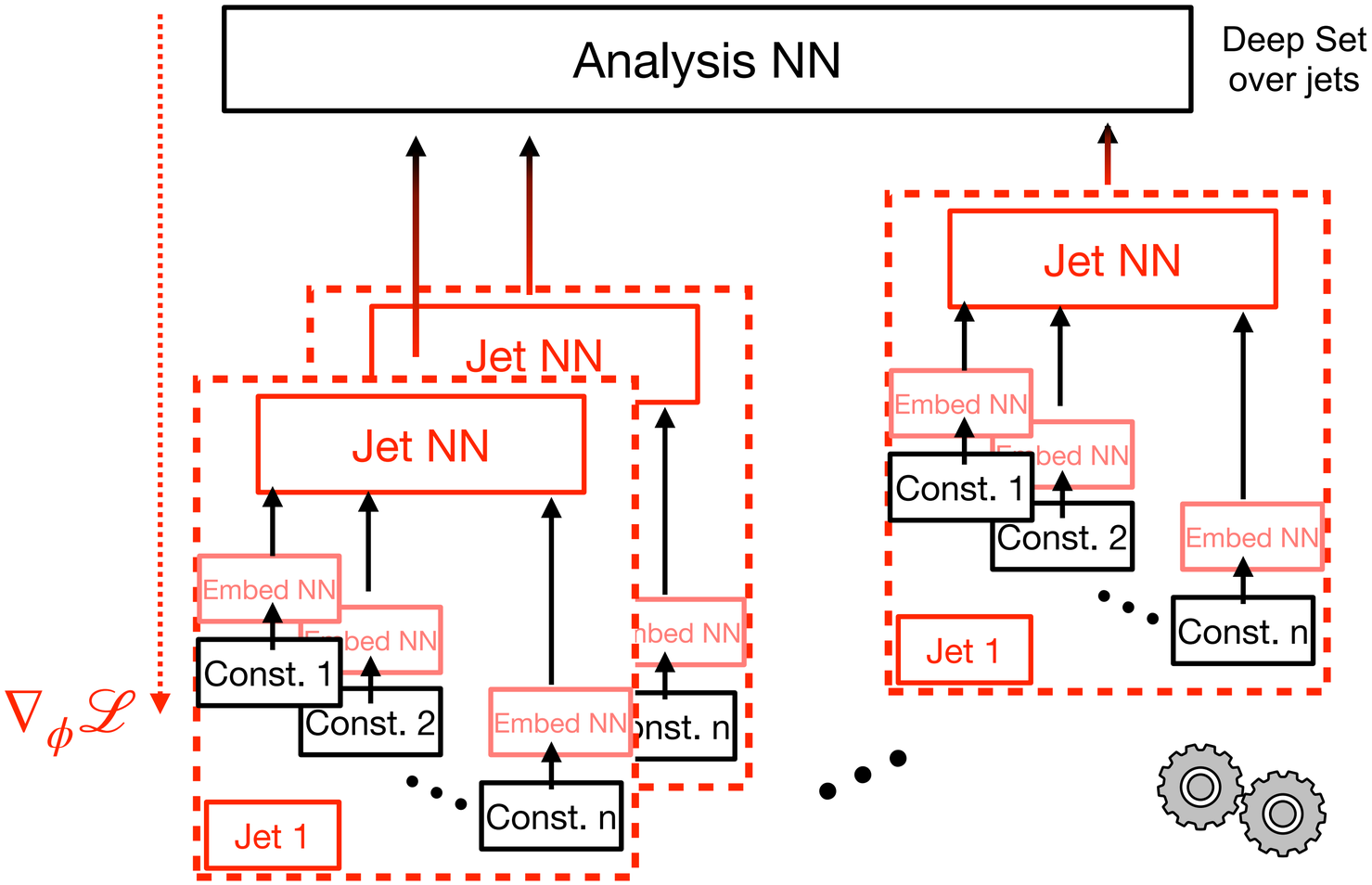
anti-kT



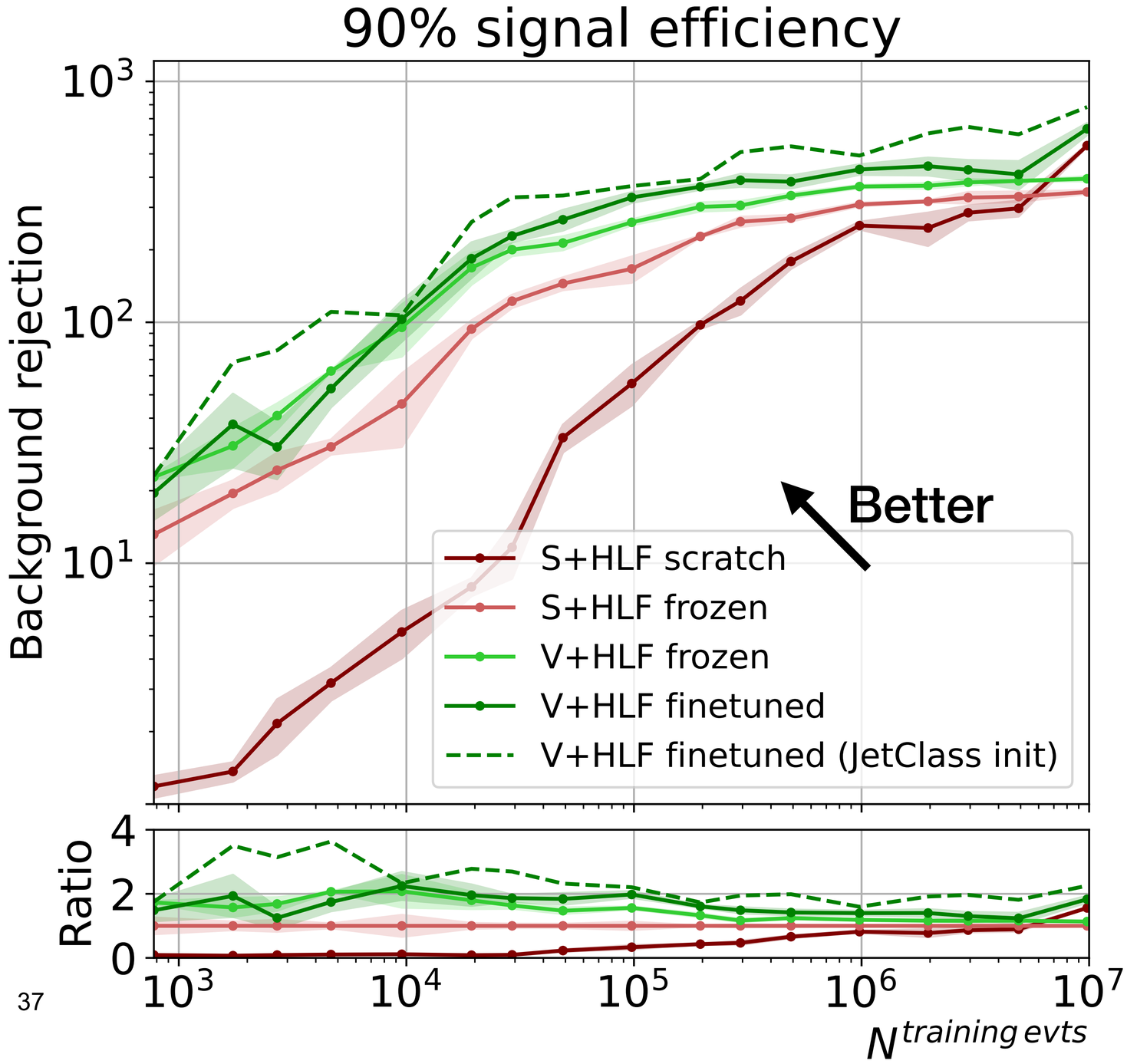
Clustering

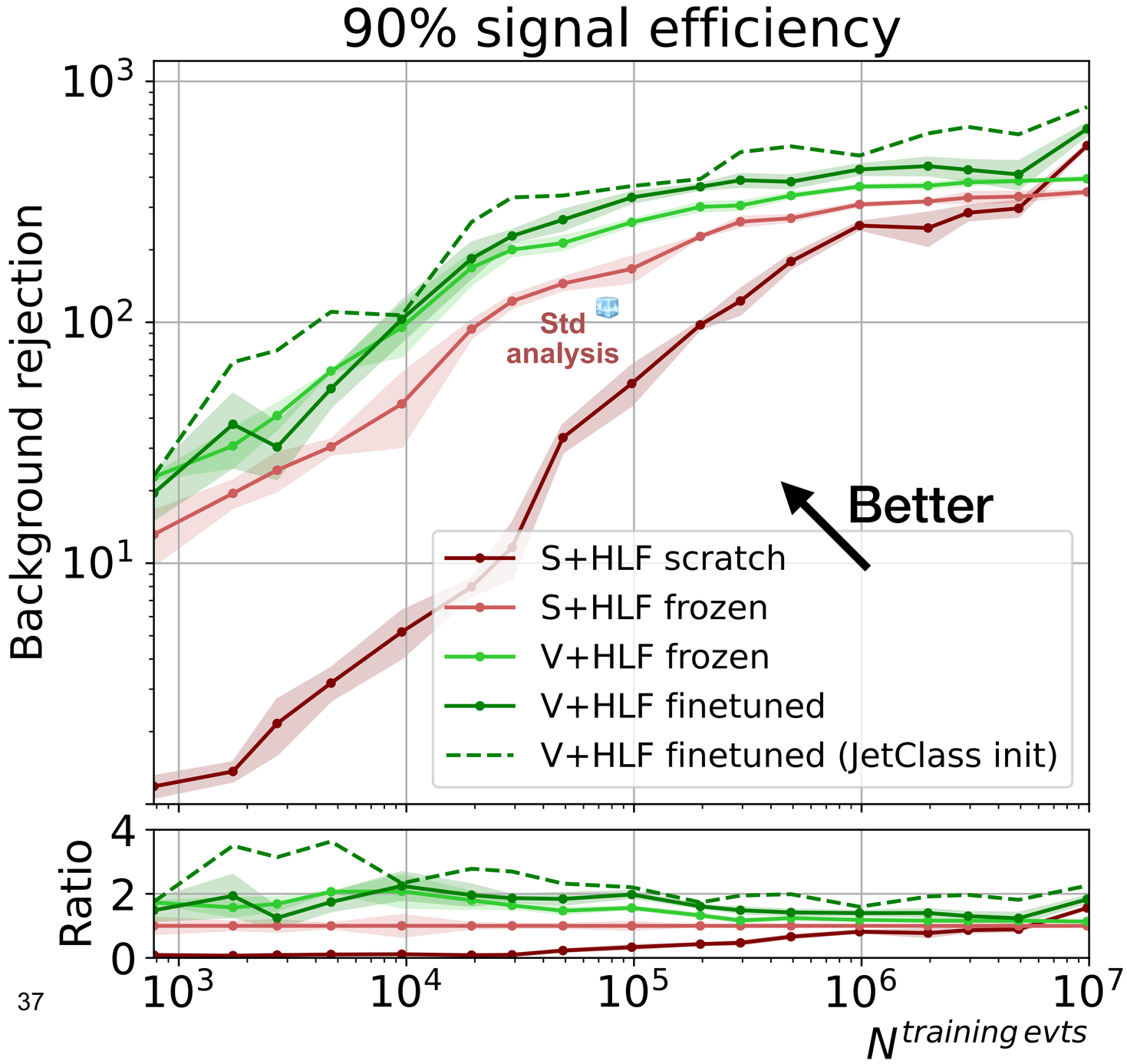


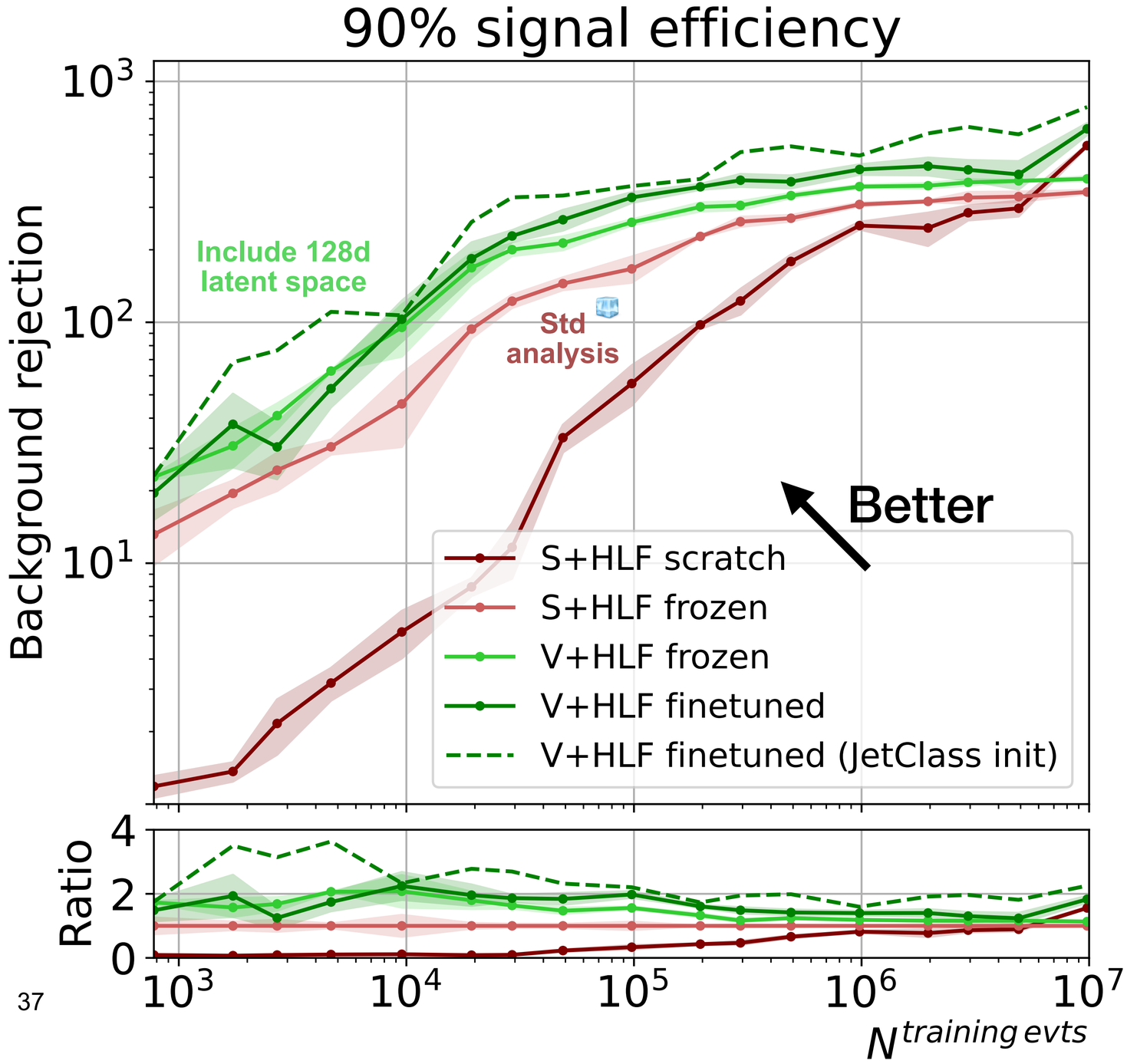
HH vs bkg

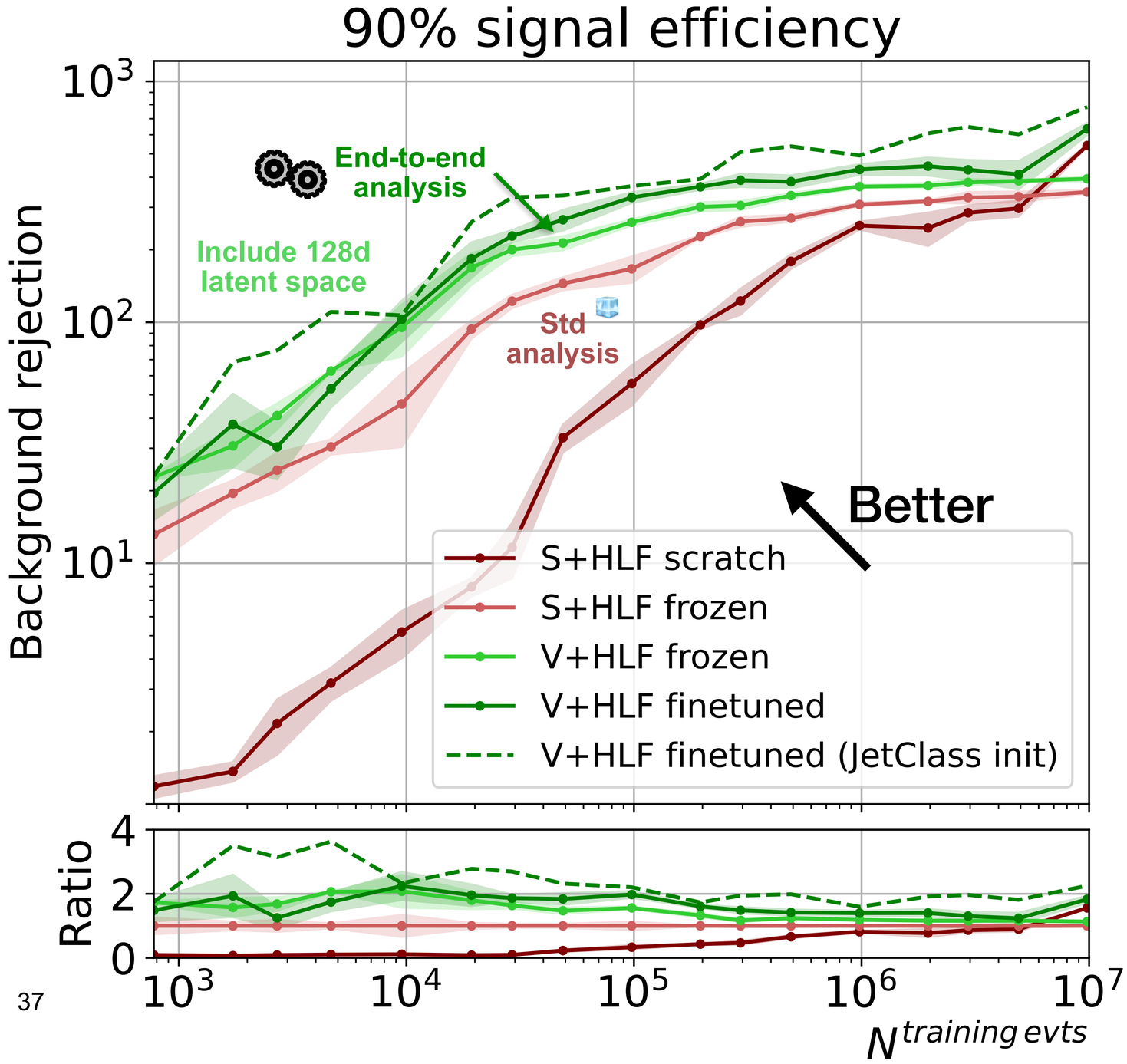


Foundation models: can we let gradients do our work for us? **“Fine-tuning”**


90% signal efficiency
Background rejection
10^3
10^2
10^1
Better
S+HLF scratch
S+HLF frozen
V+HLF frozen
V+HLF finetuned
V+HLF finetuned (JetClass init)
Ratio
4
2
0
10^3
10^4
10^5
10^6
10^7
$N^{training\ evts}$

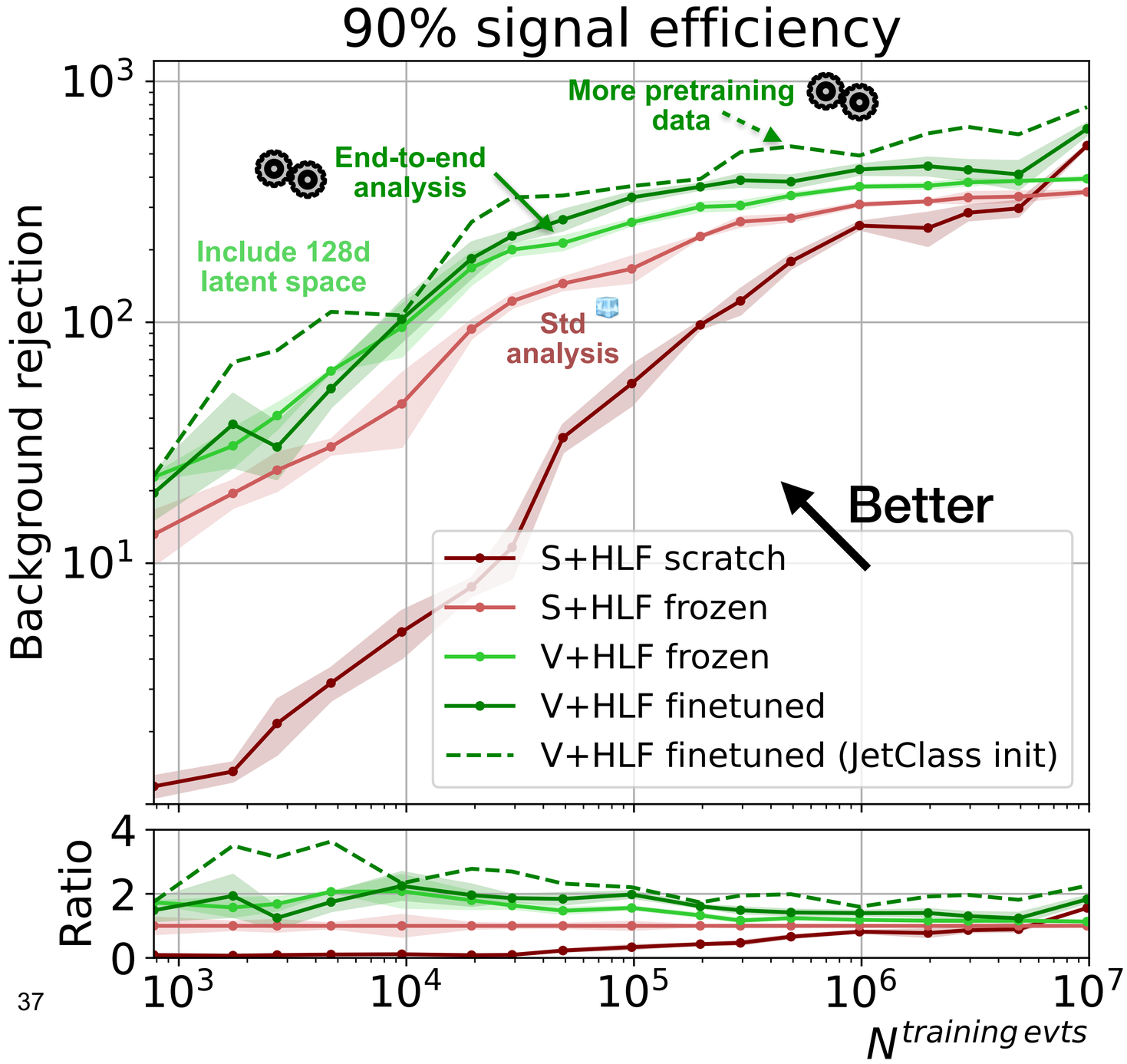

90% signal efficiency
Background rejection
10^3
10^2
10^1
Std analysis
Better
S+HLF scratch
S+HLF frozen
V+HLF frozen
V+HLF finetuned
V+HLF finetuned (JetClass init)
Ratio
4
2
0
10^3
10^4
10^5
10^6
10^7
$N^{training\ evts}$


90% signal efficiency
Background rejection
10^3
10^2
10^1
Include 128d latent space
Std analysis
Better
S+HLF scratch
S+HLF frozen
V+HLF frozen
V+HLF finetuned
V+HLF finetuned (JetClass init)
Ratio
4
2
0
10^3
10^4
10^5
10^6
10^7
$N^{training\ evts}$



Decreases bkg by 2x …

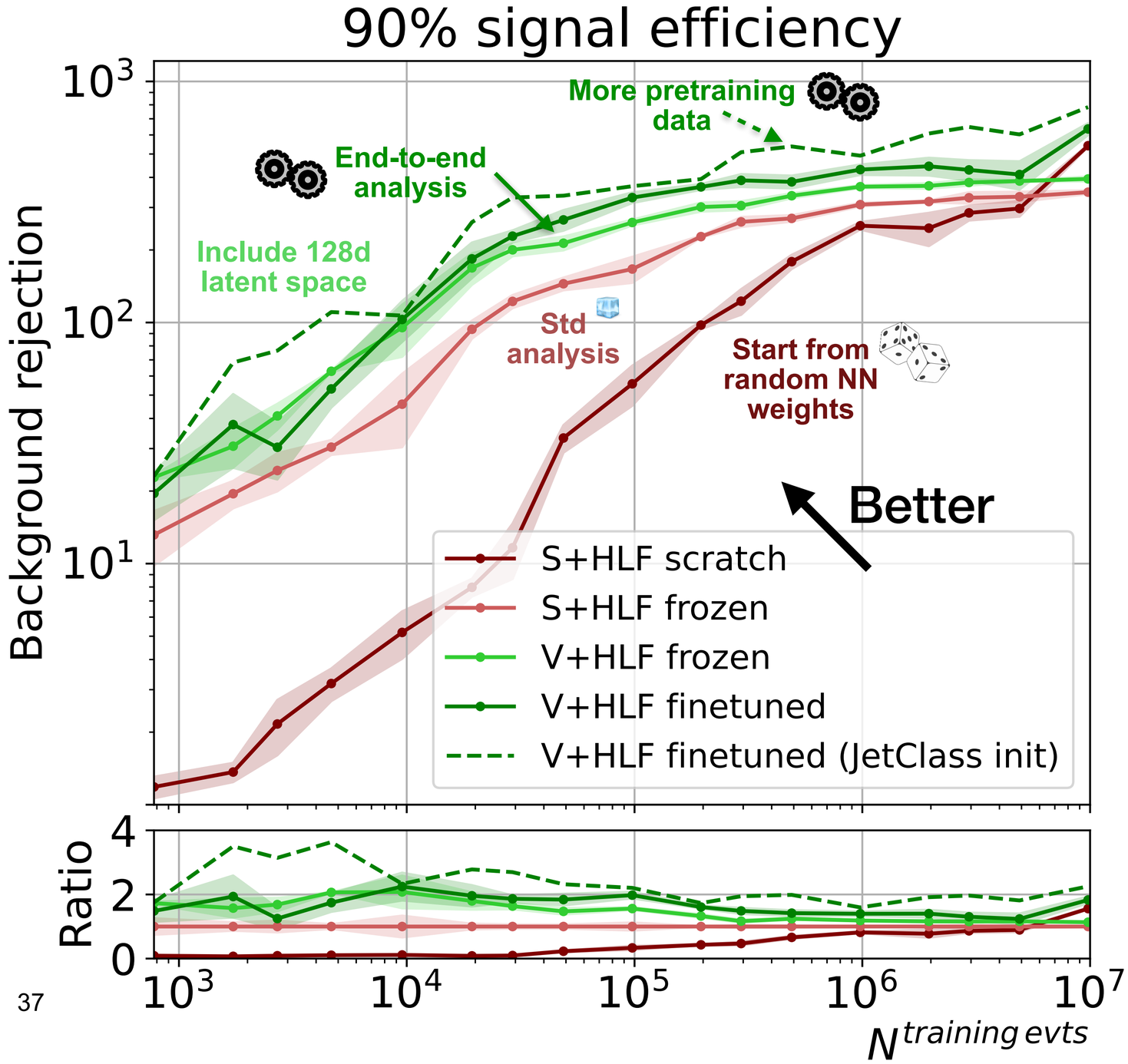
$S/\sqrt{B}$: increases significance by **40%!**



A better Higgs tagger helps analysis performance

Decreases bkg by 2x …

$S/\sqrt{B}$: increases significance by **40%!**



A better Higgs tagger helps analysis performance

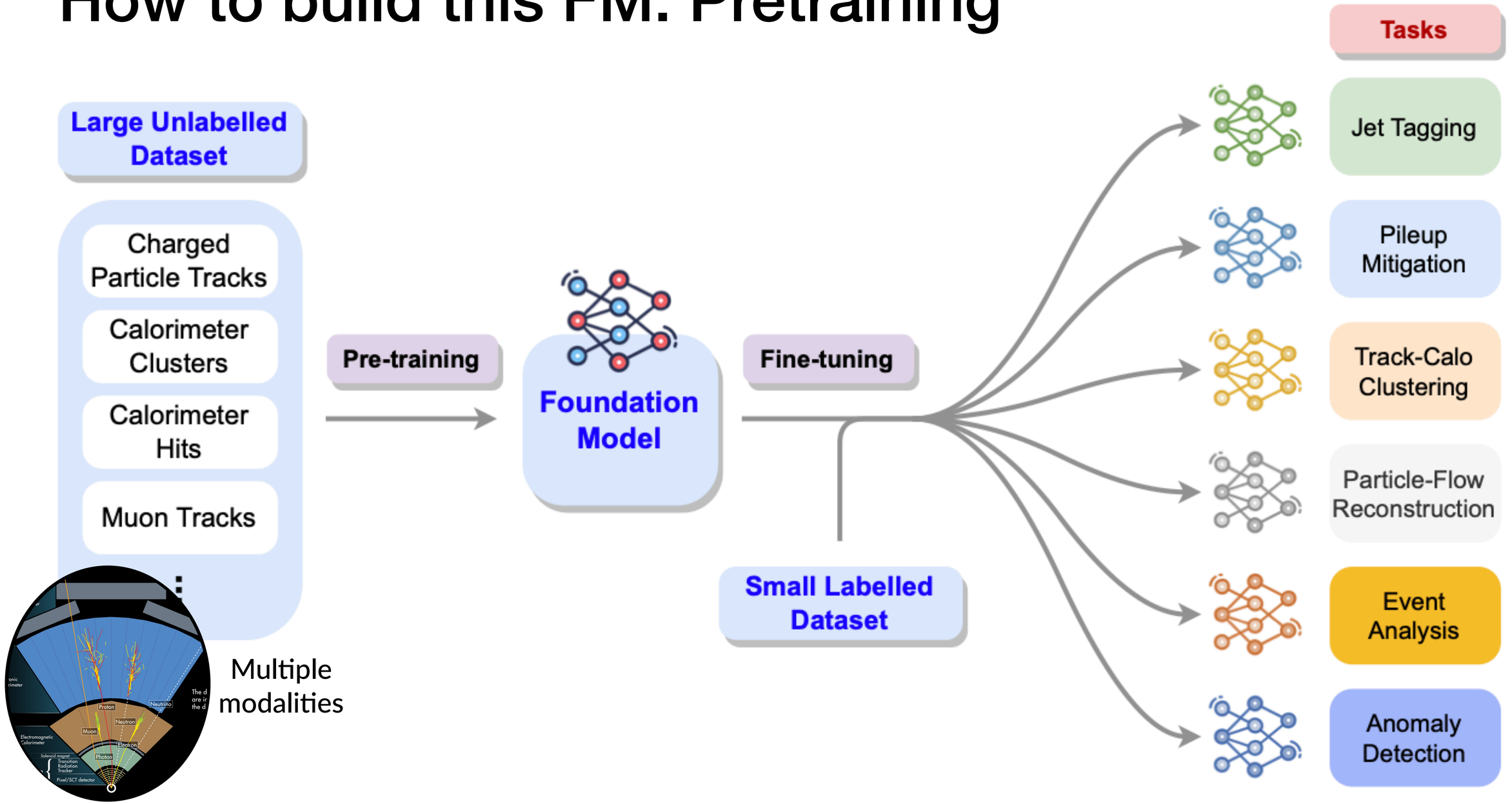
But training from scratch, with enough data, will surpass traditional analyses.

Decreases bkg by 2x …

$S/\sqrt{B}$: increases significance by **40%!**

Graphic from M. Kagan

How to build this FM: Pretraining



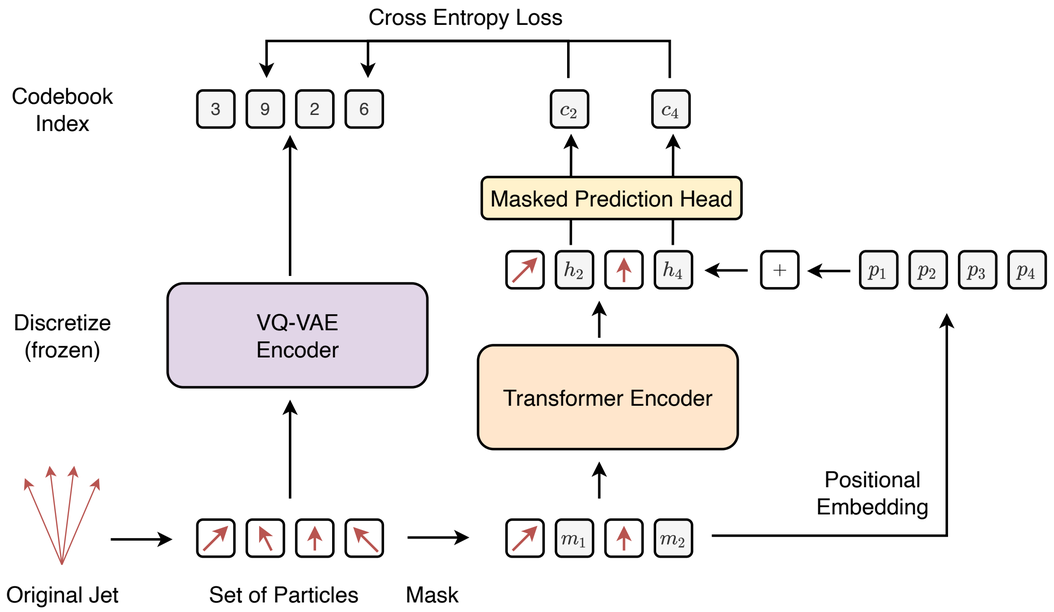
2401.13537; ML4Jets; IML

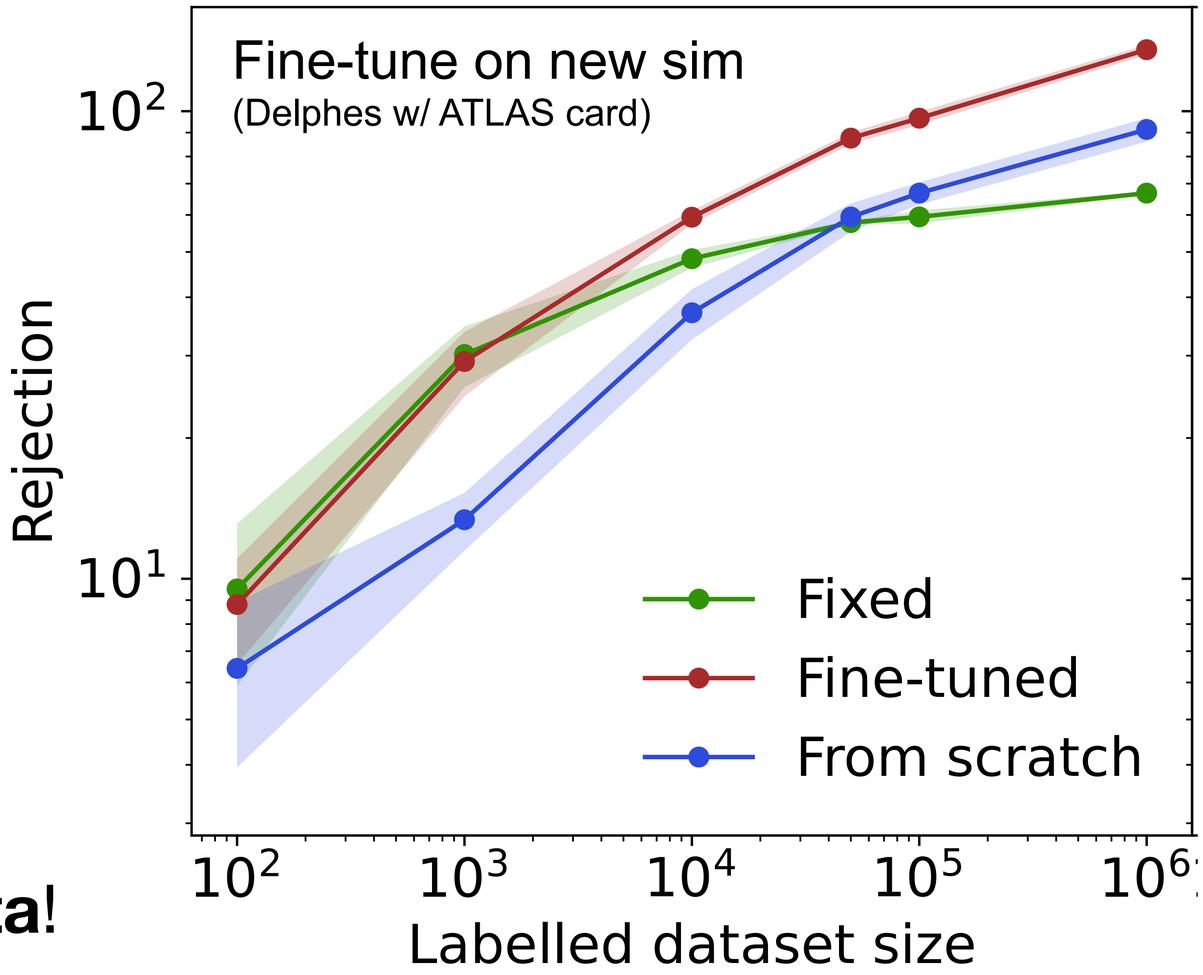
More self-supervised training proposals, see M. Kagan & A. Hallin

Masked Particle Modelling

Pre-train in an “unsupervised learning” setting, a.k.a, “generative model”

→ **Jet completion**





Pretrain Delphes 100m jets.

Ultimately… **could be on ATLAS data**!

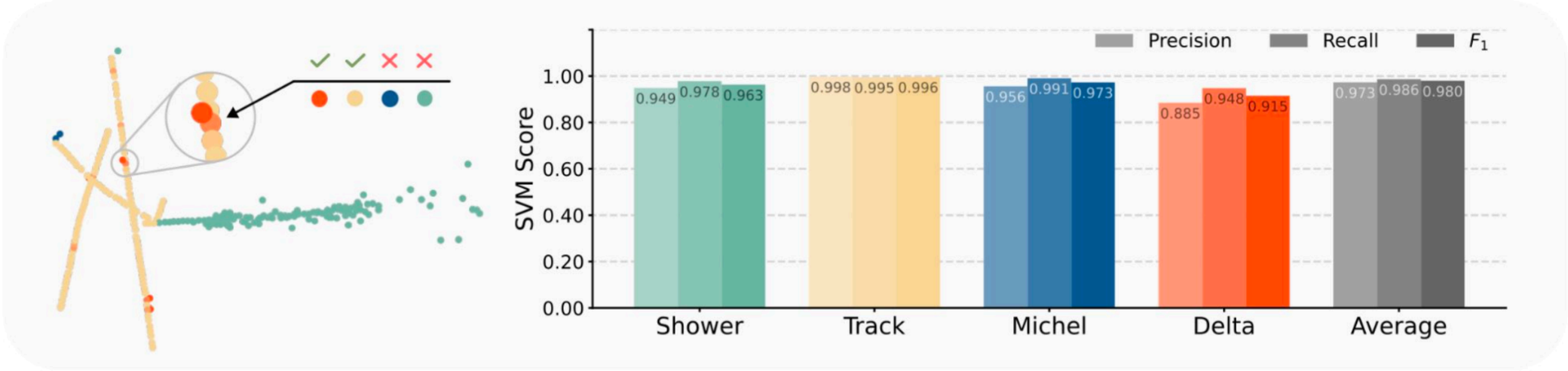
K. Terao slide

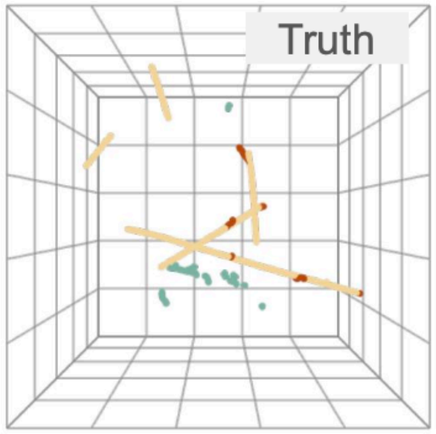
2502.02558

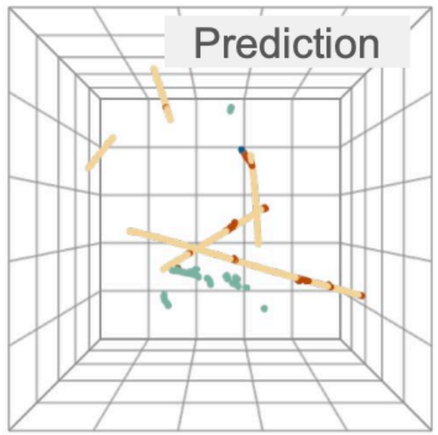
Also in neutrino physics (DUNE)

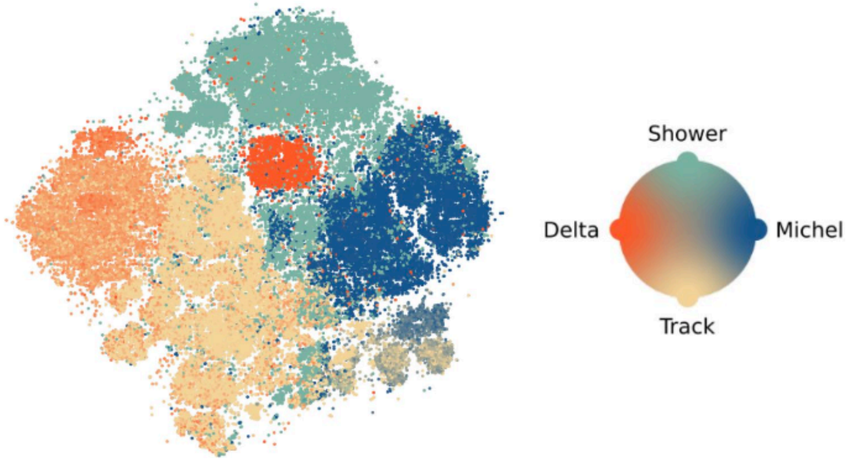
Linear Probing: Semantic Segmentation

Semantic Segmentation





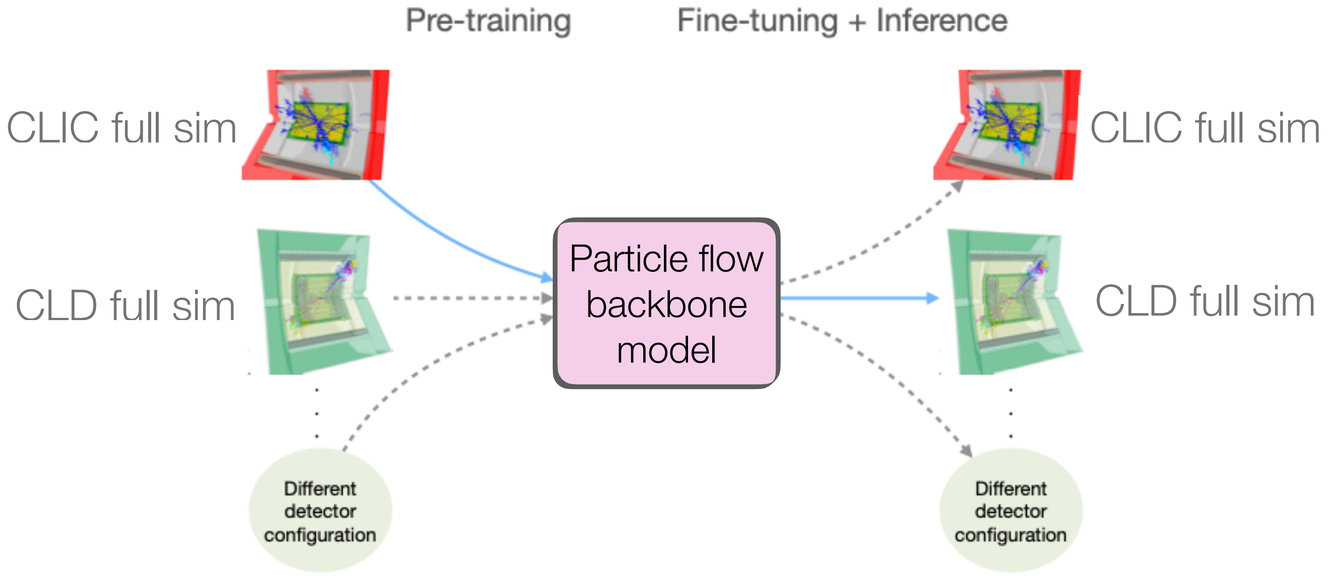


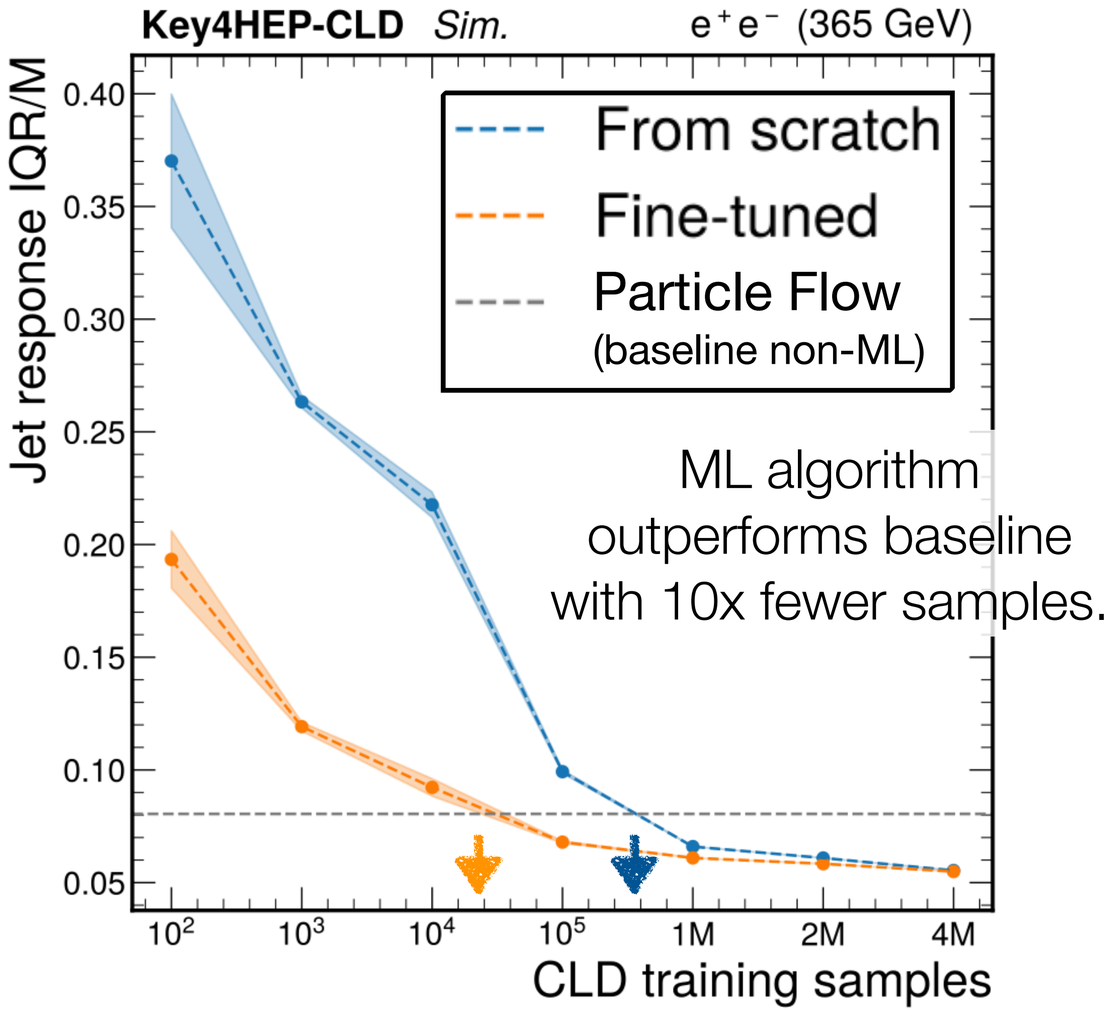


Phys. Rev. D 111, 092015 (2025)

What does this enable?

➜ Sharing reconstruction across experiments

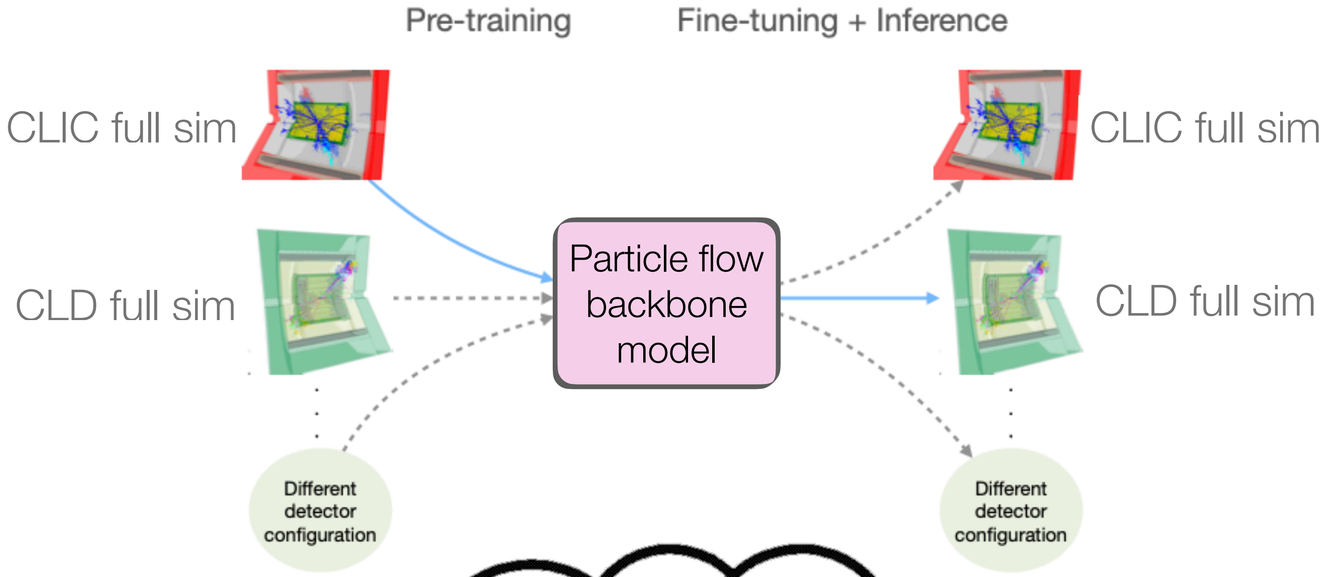




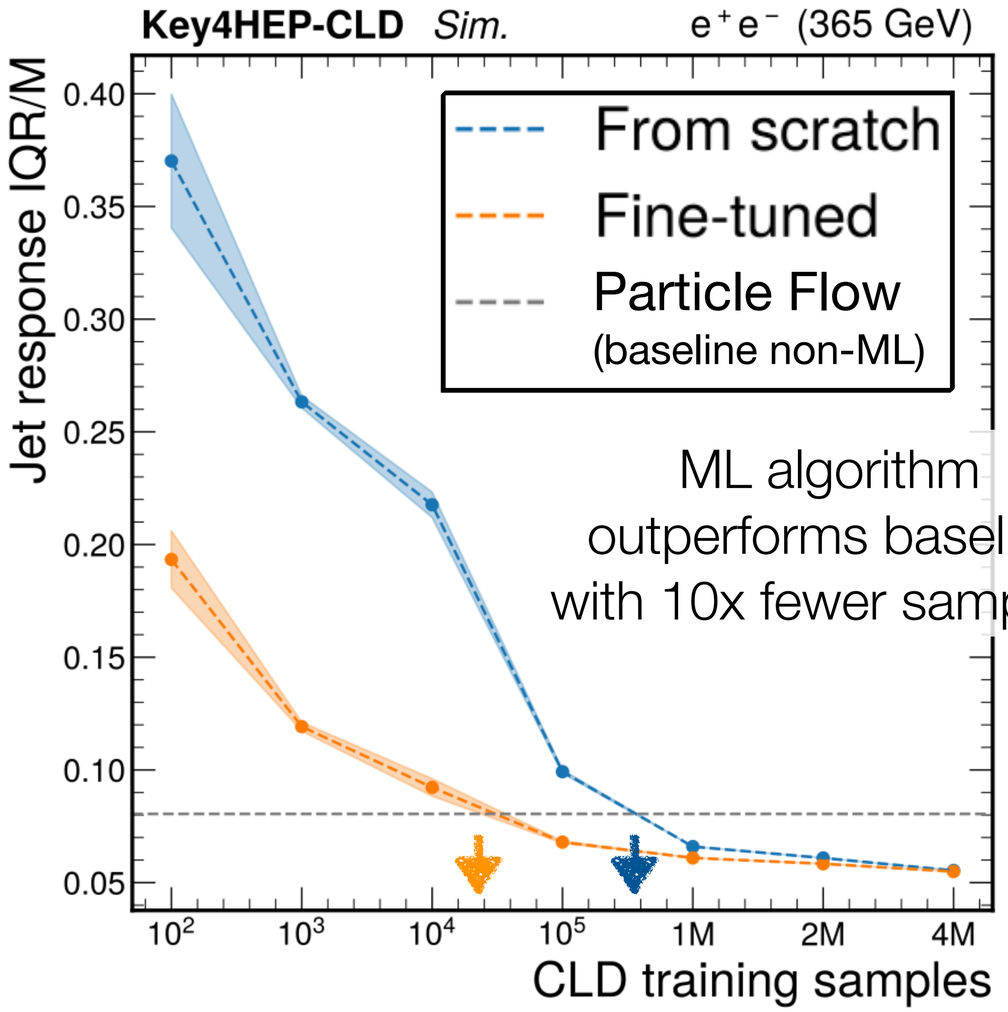
Phys. Rev. D 111, 092015 (2025)

What does this enable?

➜ Sharing reconstruction across experiments



Could we share a reconstruction backbone across ATLAS and Belle II?



The Q: Build big or build smart?

Workshop ongoing in Munich, 25.8 — 19.9

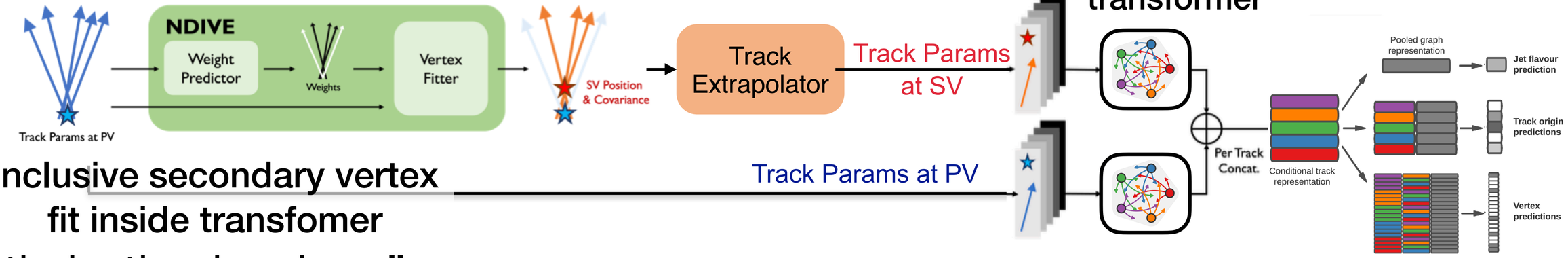
BUILD BIG OR BUILD SMART: EXAMING SCALE AND DOMAIN KNOWLEDGE IN MACHINE LEARNING FOR FUNDAMENTAL PHYSICS

25 August - 19 September 2025

Lukas Heinrich, Michael Kagan, Margarita Osadchy, Tobias Golling, Siddarth Mishra-Sharma

https://www.munich-iapbp.de/activities/activities-2025/machine-learning

Neural Differentiable Vertex Fitter

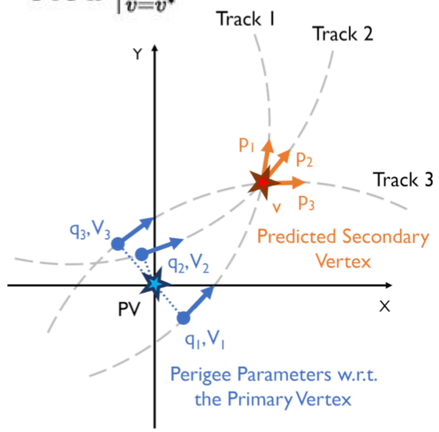


Inclusive secondary vertex fit inside transfomer

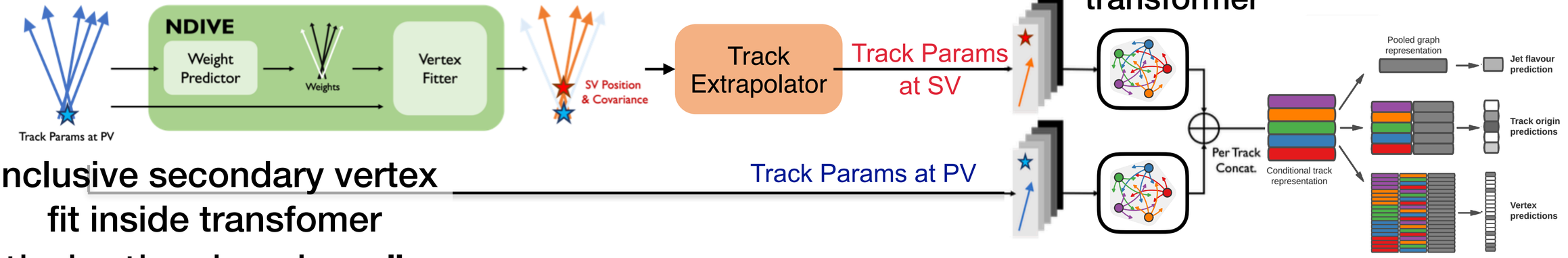
“Optimization in a loop”

Gradients of fitted vertex to optimize end-to- end trained b-tagger

$$\frac{\mathrm{d}v^*}{\mathrm{d}w} = -\left(\frac{\partial^2\chi^2}{\partial v^2}\right)^{-1}\frac{\partial^2\chi^2}{\partial v \partial w}\bigg|_{v=v^*}$$



Neural Differentiable Vertex Fitter

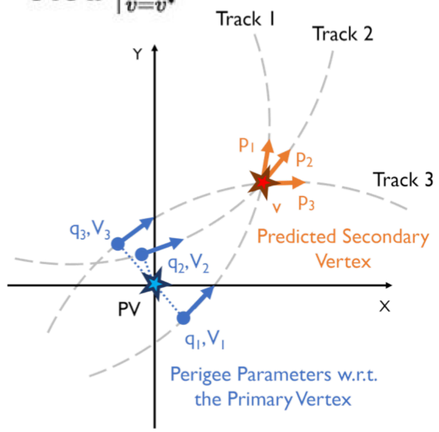


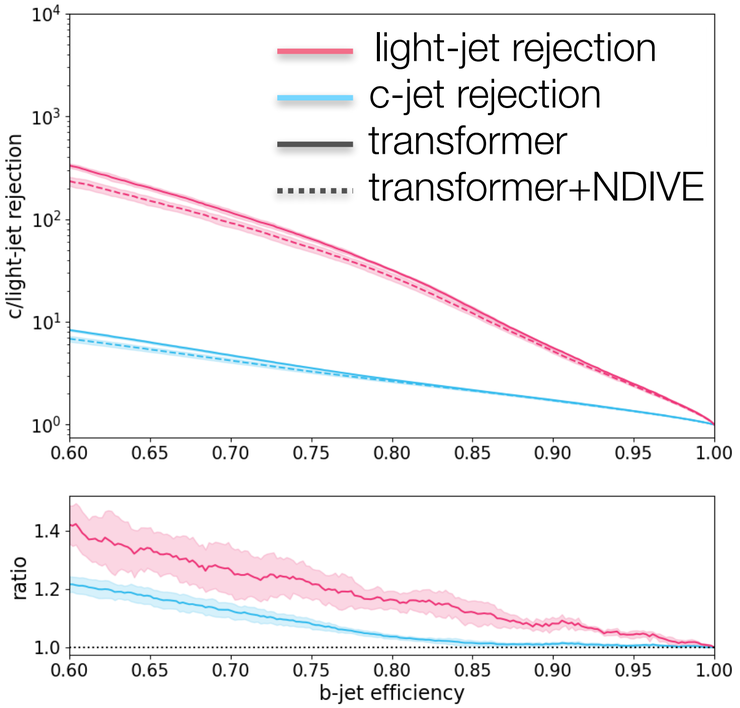
Inclusive secondary vertex fit inside transfomer

"Optimization in a loop"

Gradients of fitted vertex to optimize end-to- end trained b-tagger

$$\frac{\mathrm{d}v^*}{\mathrm{d}w} = -\left(\frac{\partial^2\chi^2}{\partial v^2}\right)^{-1}\frac{\partial^2\chi^2}{\partial v\partial w}\bigg|_{v=v^*}$$

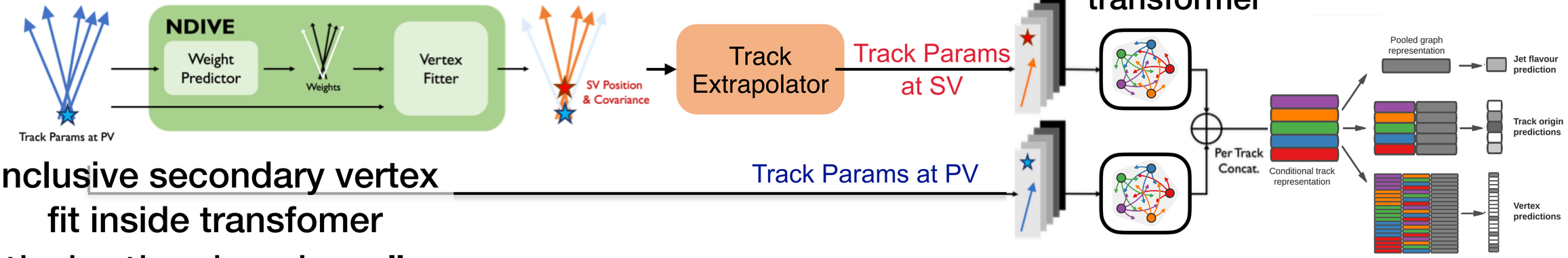




⬅ 50% improvement in bkg rejection

(Up to 10x improvement 😱 in jet tagging from vtx fit if perfect track selection)

Neural Differentiable Vertex Fitter

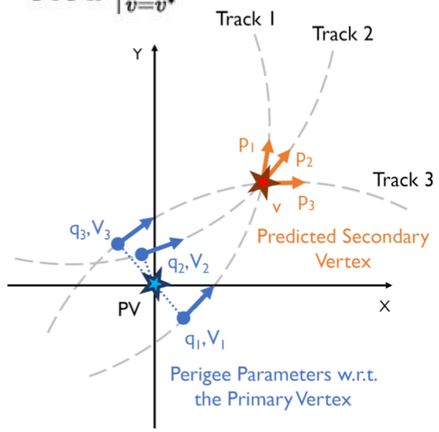


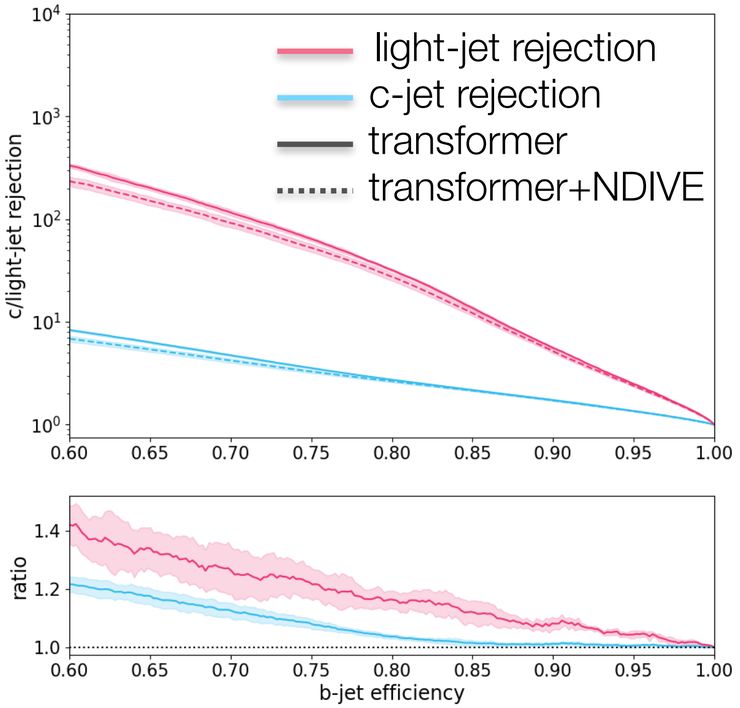
Inclusive secondary vertex fit inside transfomer

"Optimization in a loop"

Gradients of fitted vertex to optimize end-to- end trained b-tagger

$$\frac{\mathrm{d}v^*}{\mathrm{d}w} = -\left(\frac{\partial^2\chi^2}{\partial v^2}\right)^{-1}\frac{\partial^2\chi^2}{\partial v \partial w}\bigg|_{v=v^*}$$





⬅ 50% improvement in bkg rejection

(Up to 10x improvement 😱 in jet tagging from vtx fit if perfect track selection)

Interesting for Belle 2? 1808.10567

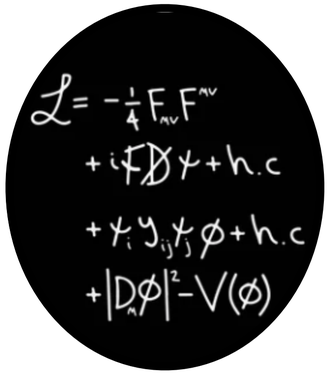
6.2.1. Vertex finding algorithms. The Belle II experiment has deployed three implementations of a vertex fit: KFitter, developed for the Belle experiment, RAVE [59], a standalone package originating from the CMS vertex fitting libraries, and TreeFitter [71], initially conceived by the BaBar collaboration. We use both KFitter and RAVE for kinematic fits and RAVE for geometric fits. TreeFitter is used for the fitting entire decay chains.

Kinematic fits. Kinematic fitting uses the known properties of a specific decay chain to improve the measurements of the process. Lagrangian multipliers are used in order to impose the kinematic constraints to the fit. Given the measurements, $\mathbf{q} = (q_1, ..., q_n)$, with a covariance matrix, V, and kinematic constraints, $\mathbf{h}(\mathbf{q})$, the function to be minimised in terms of the most suitable vertex is:

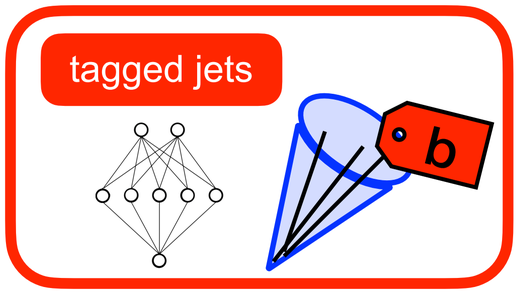
$$\chi^2 = (\mathbf{q} - \bar{\mathbf{q}})^T V^{-1} (\mathbf{q} - \bar{\mathbf{q}}) + +2\lambda^T (\boldsymbol{D}\delta\boldsymbol{y} + \boldsymbol{h}(boldsymbol{q})) \quad (15)$$

where λ is the Lagrange multiplier, $h(\bar{\mathbf{q}}) = 0$ and $\mathbf{D} = \partial\mathbf{h}/\partial\mathbf{y}$. Here $\bar{\mathbf{q}}$ represents the improved measurements, $\mathbf{d}$ is the kinematic constraint at the starting value and δyis the difference between the improved measurement and the starting value.

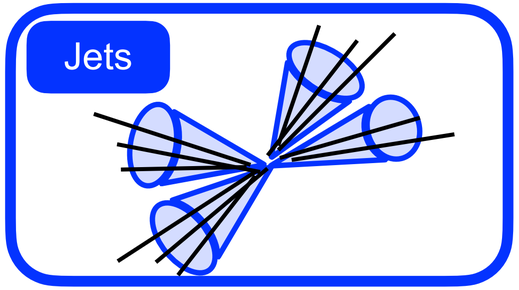
How to use gradients *MORE*?



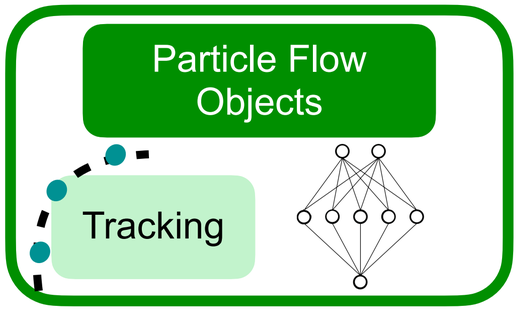
Analysis

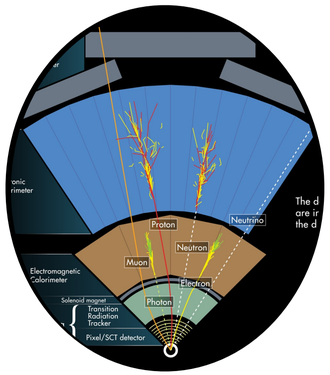


tagging



anti-kT



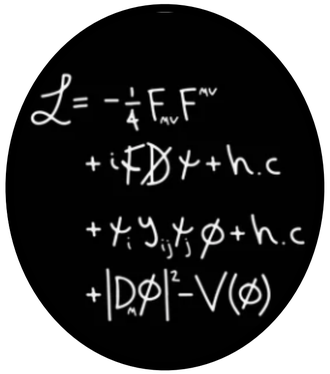


Clustering

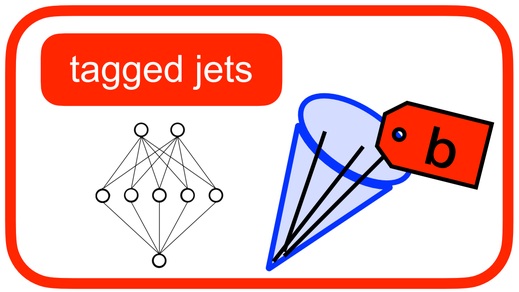
$\frac{d\ \text{Analysis}}{d\ \text{b-tag}}$	$\frac{d\ \text{b-tag}}{d\ \text{jet}}$	$\frac{d\ \text{jet}}{d\ \text{pflow}}$	$\frac{d\ \text{pflow}}{d\ \text{tracks}}$
✅	✅		✅

W/ A. Kofler, M. Kagan and L. Heinrich, in prep

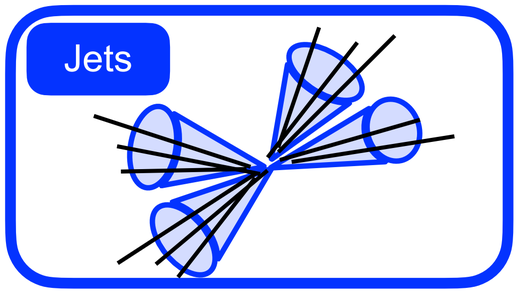
How to use gradients *MORE*?



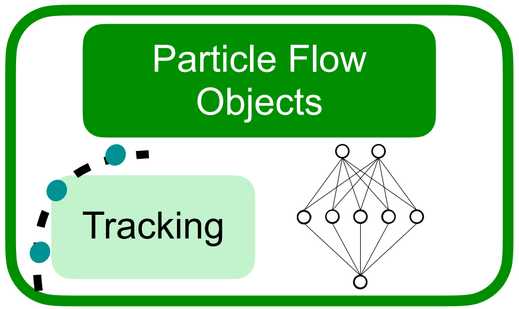
Analysis

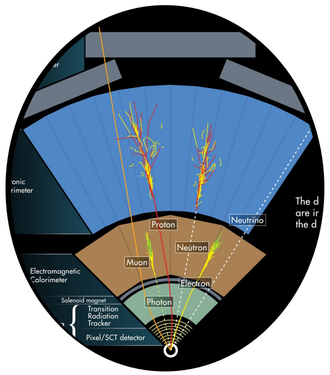


tagging



anti-kT





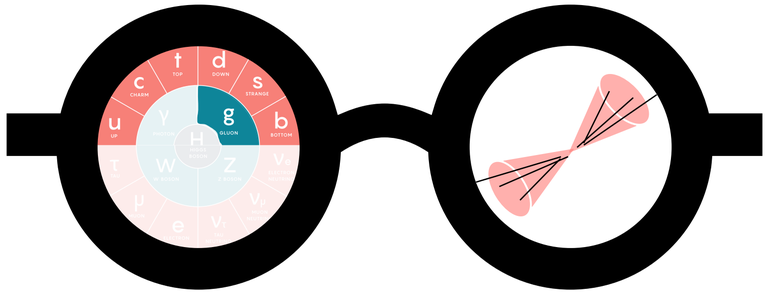
Clustering

$$\frac{d\,\text{Analysis}}{d\,\text{b-tag}} \quad \frac{d\,\text{b-tag}}{d\,\text{jet}} \quad \frac{d\,\text{jet}}{d\,\text{pflow}} \quad \frac{d\,\text{pflow}}{d\,\text{tracks}}$$

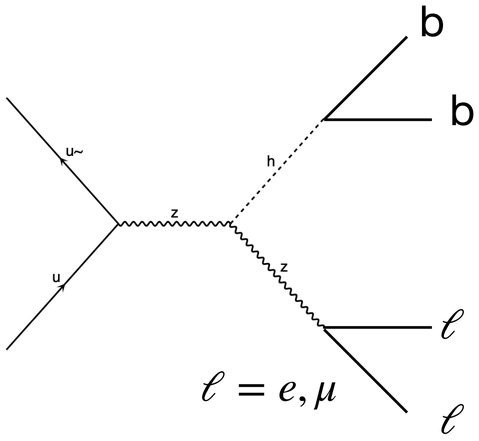
✅ ✅ ❌ ✅

Clustering *not* differentiable… how to fully exploint the richness of our high dim data (**tracks** / hits 📸).

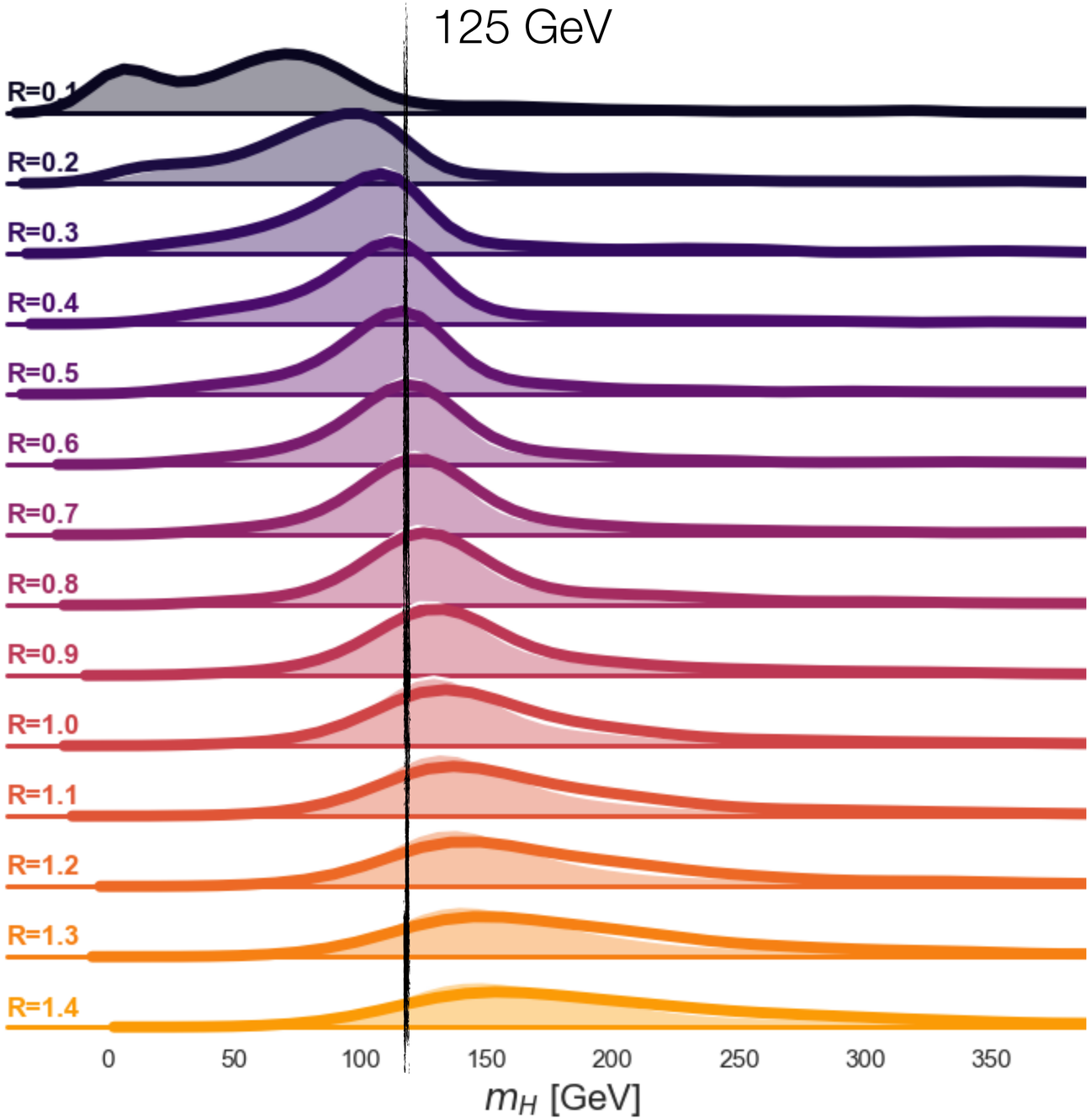
W/ A. Kofler, M. Kagan and L. Heinrich, in prep



How to see a Higgs: wearing the right glasses



Madgraph + pythia, pp collisions @ 14 TeV



With A. Kofler, M. Kagan and L. Heinrich, in prep

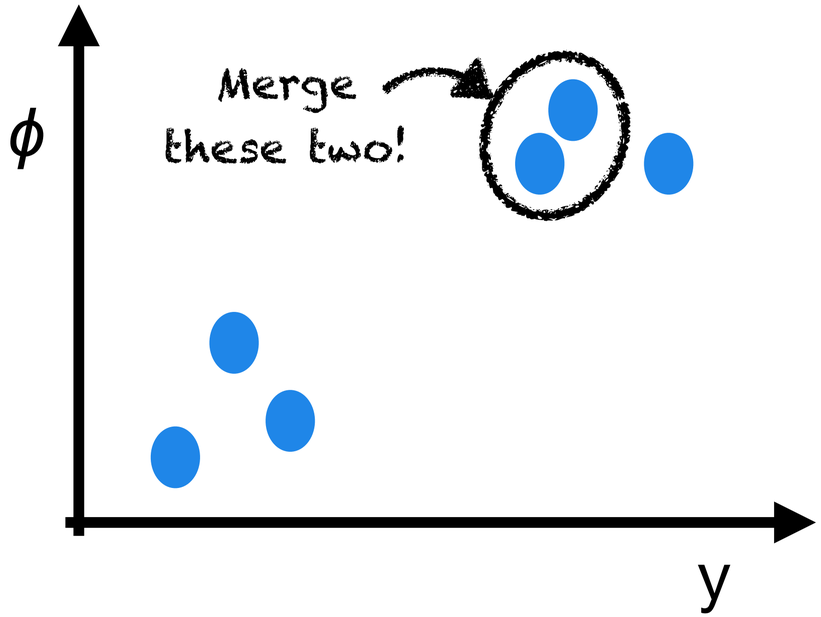
Differentiating clustering

Step 1: Interpret clustering decision probabilistically

Qanti-kT, M. Schwartz + collab: 1201.1914, 1304.2394

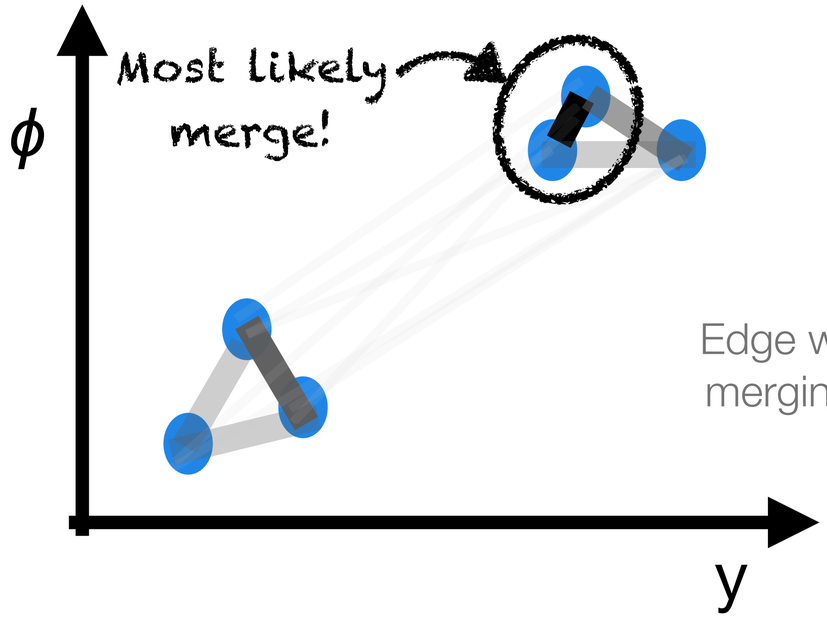
Deterministic

Merge closest two particles



Probailistic

Sample which particles to merge



Edge weights show merging probability

With A. Kofler, M. Kagan and L. Heinrich, in prep

Differentiating clustering

Step 1: Interpret clustering decision probabilistically

Step 2: Gradient with score based estimate

As in 2308.16680

$$\nabla_\theta \mathbb{E}_{x\sim p(\theta)}[f(x)]$$
$$= \mathbb{E}_{x\sim p(\theta)}[f(x)\,\nabla_\theta \log p(\theta)]$$

With A. Kofler, M. Kagan and L. Heinrich, in prep

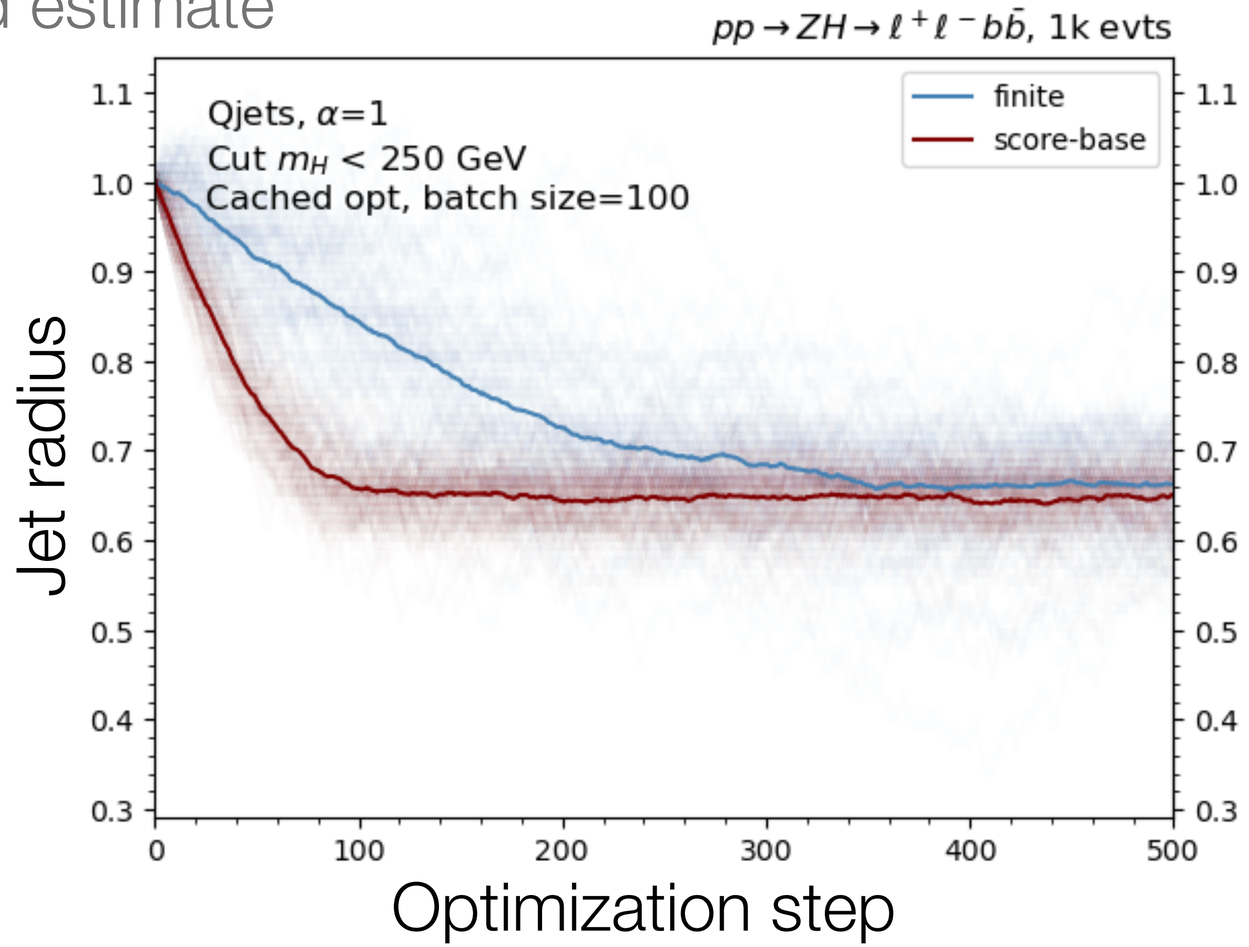
Differentiating clustering

Step 1: Interpret clustering decision probabilistically

Step 2: Gradient with score based estimate

NEW!

Step 3: Gradient based opt for clustering radius parameter



With A. Kofler, M. Kagan and L. Heinrich, in prep

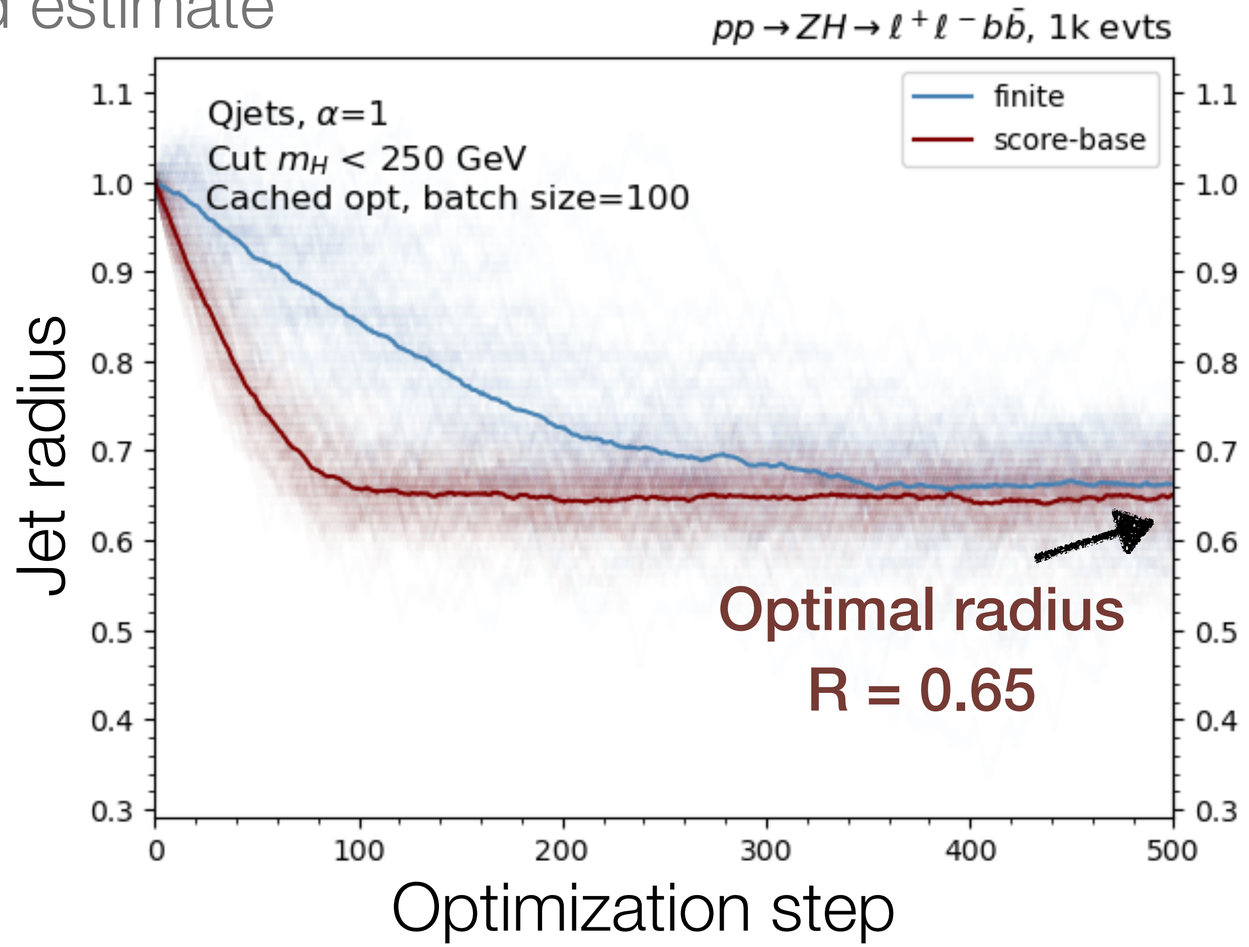
Differentiating clustering

Step 1: Interpret clustering decision probabilistically

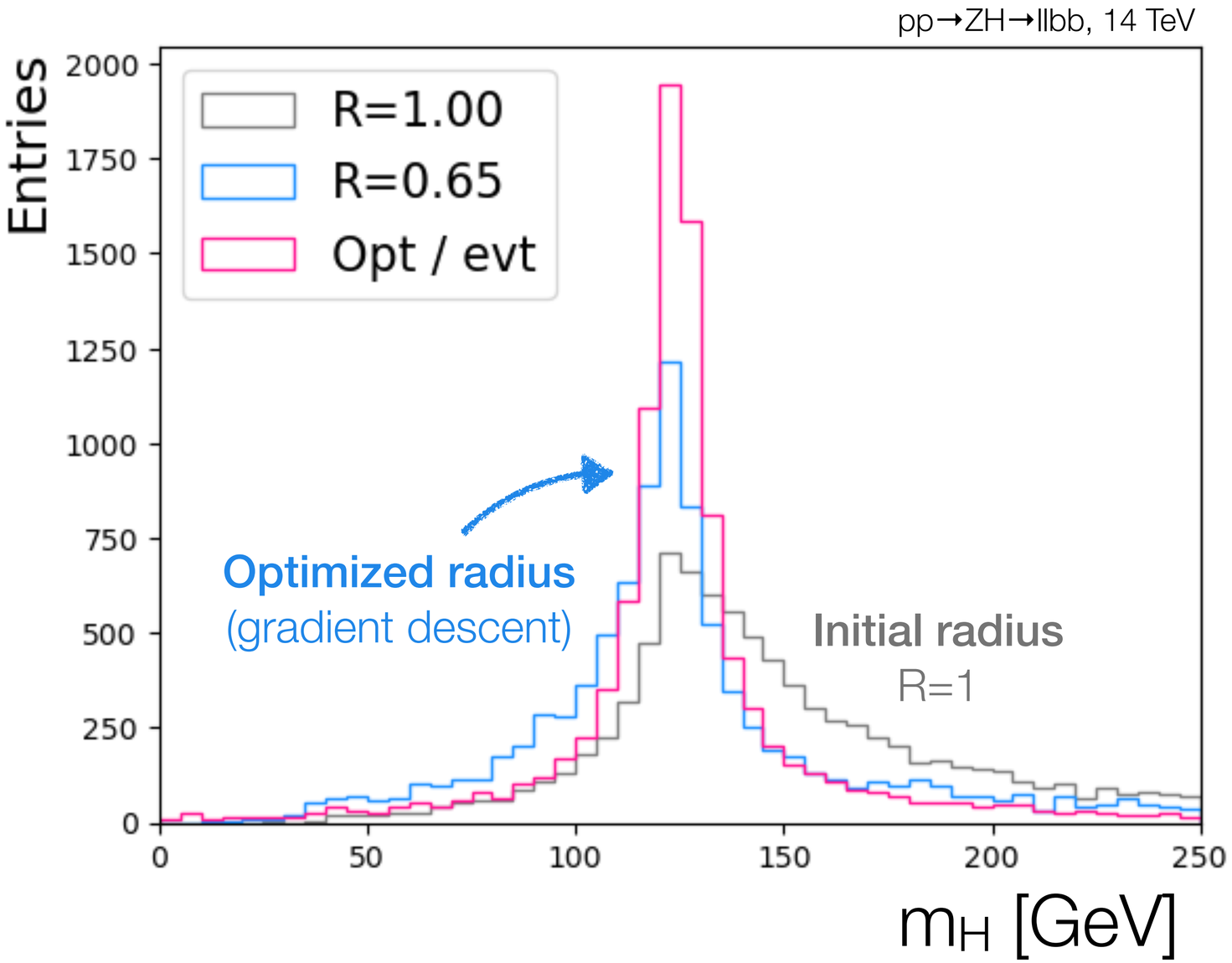
Step 2: Gradient with score based estimate

NEW!

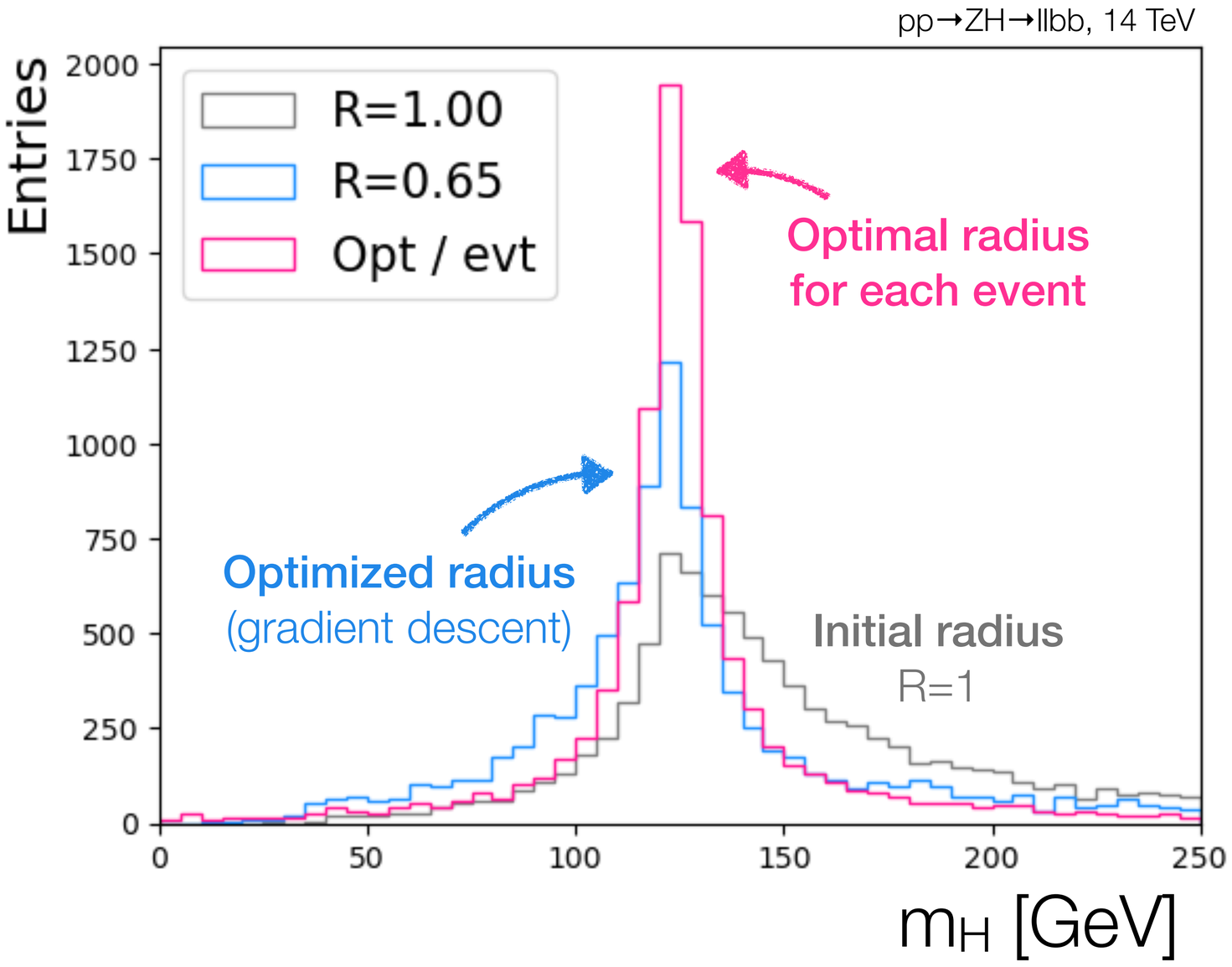
Step 3: Gradient based opt for clustering radius parameter



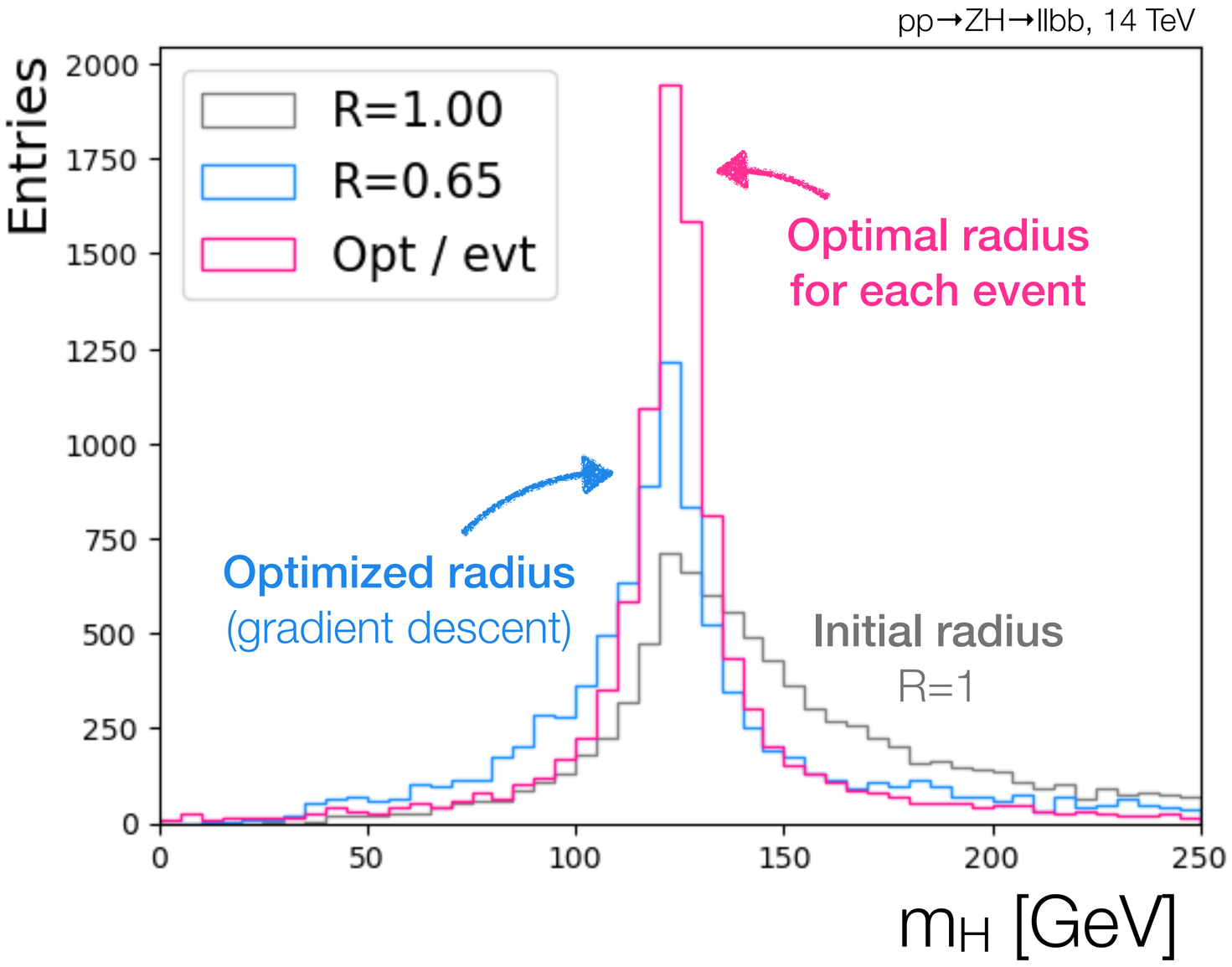
With A. Kofler, M. Kagan and L. Heinrich, in prep

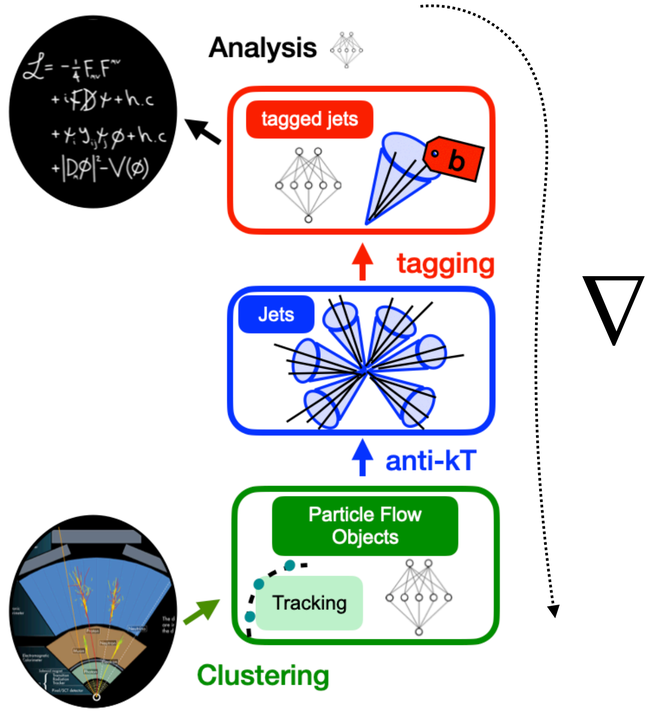


With A. Kofler, M. Kagan and L. Heinrich, in prep



With A. Kofler, M. Kagan and L. Heinrich, in prep

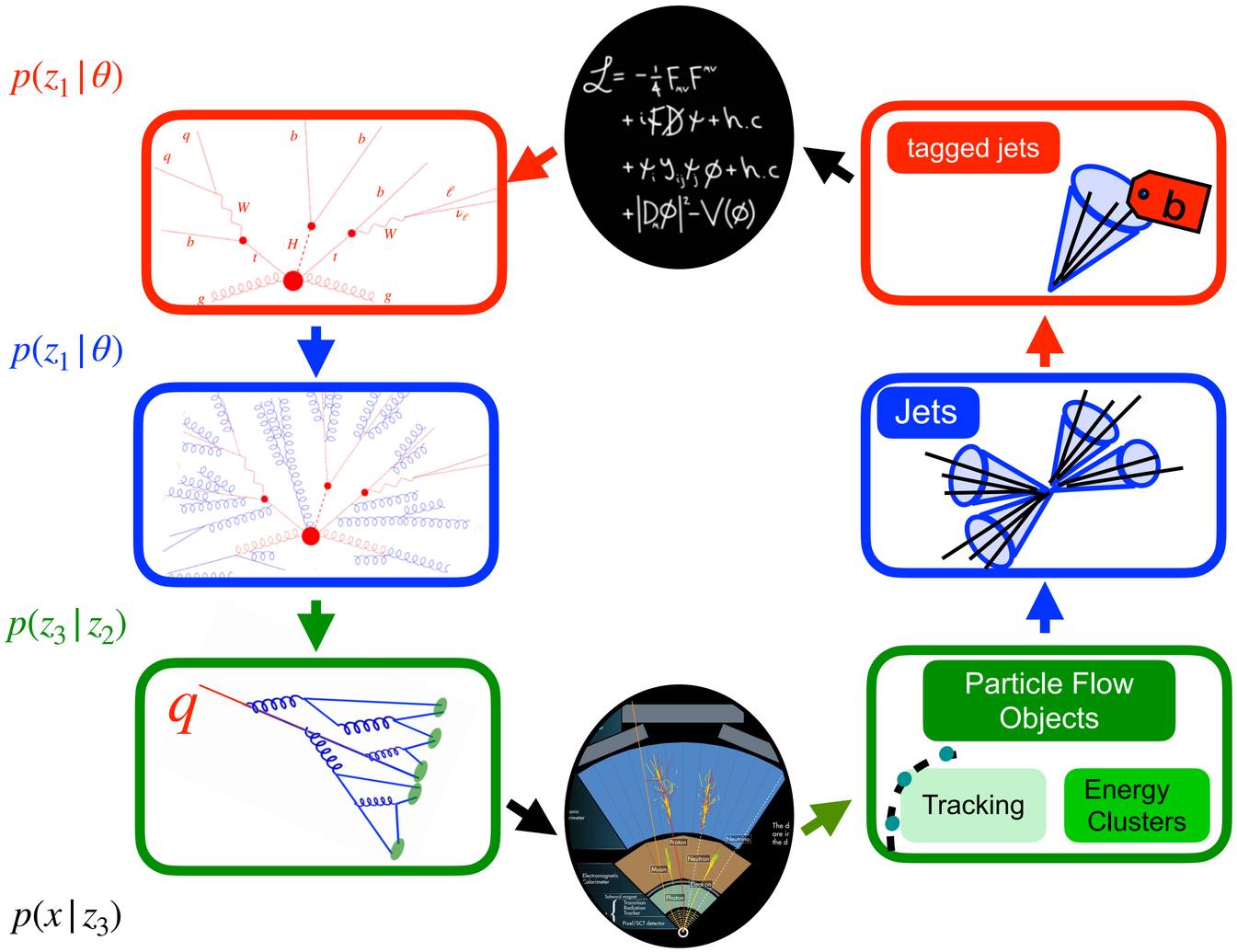




Promising for end-to-end models…
holy grail for physics data anlaysis!

With A. Kofler, M. Kagan and L. Heinrich, in prep

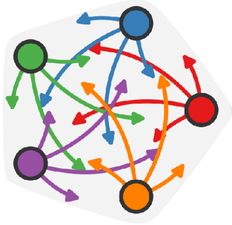
In conclusion…

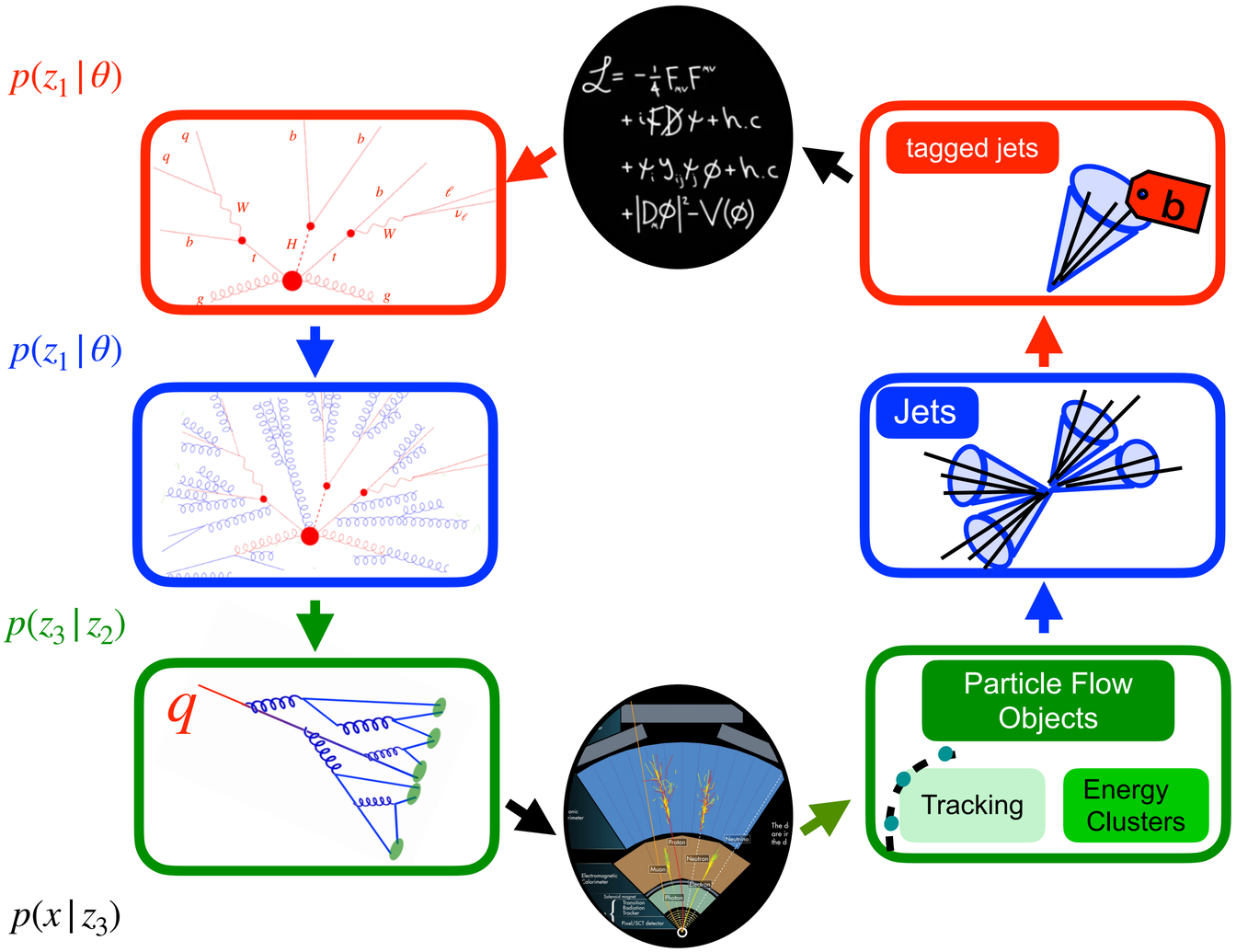


In conclusion…

b-tagging

more inputs;
bigger model, better perf

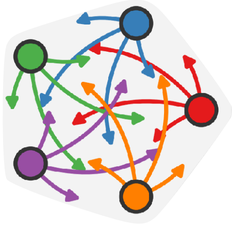


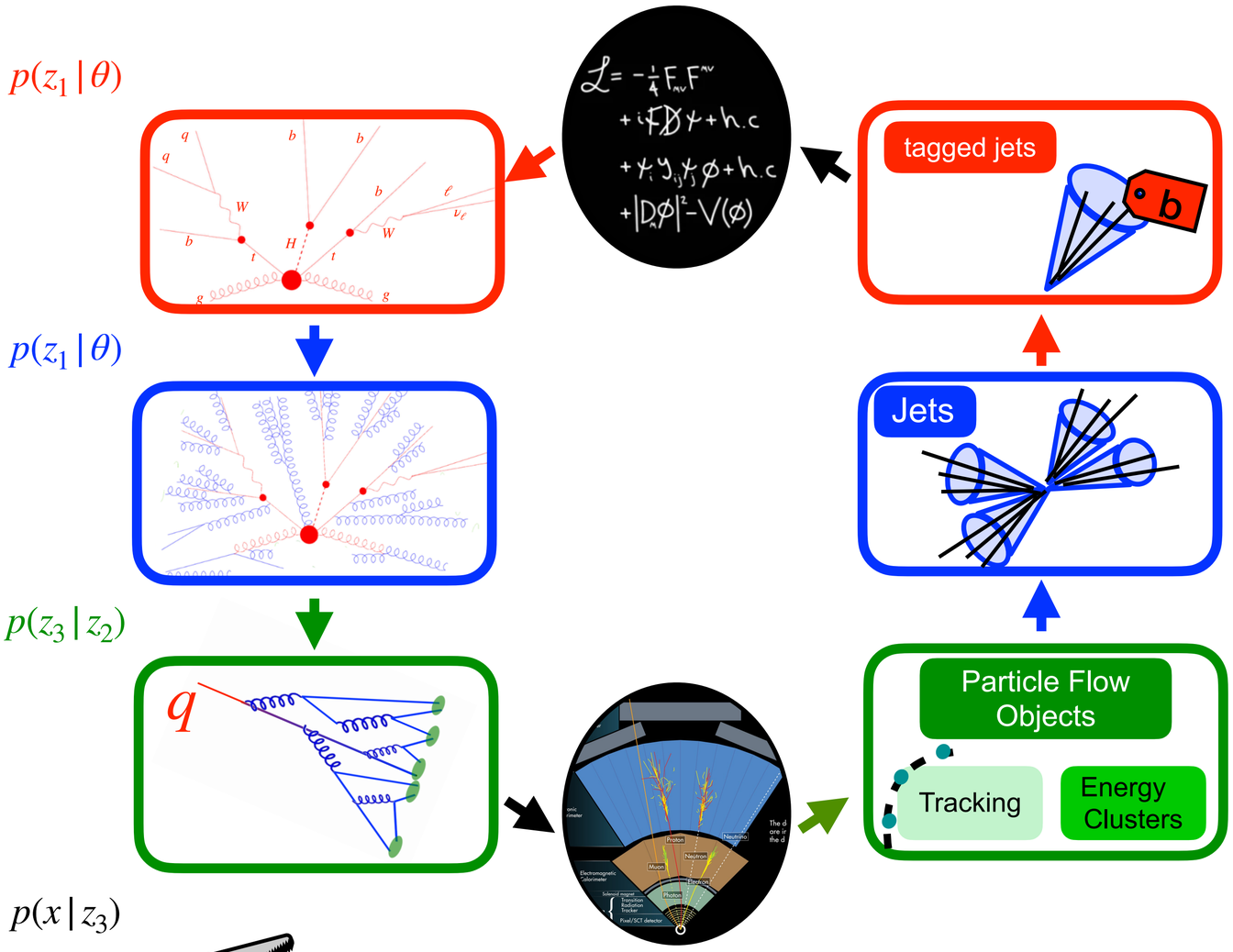


In conclusion…

b-tagging

more inputs;
bigger model, better perf

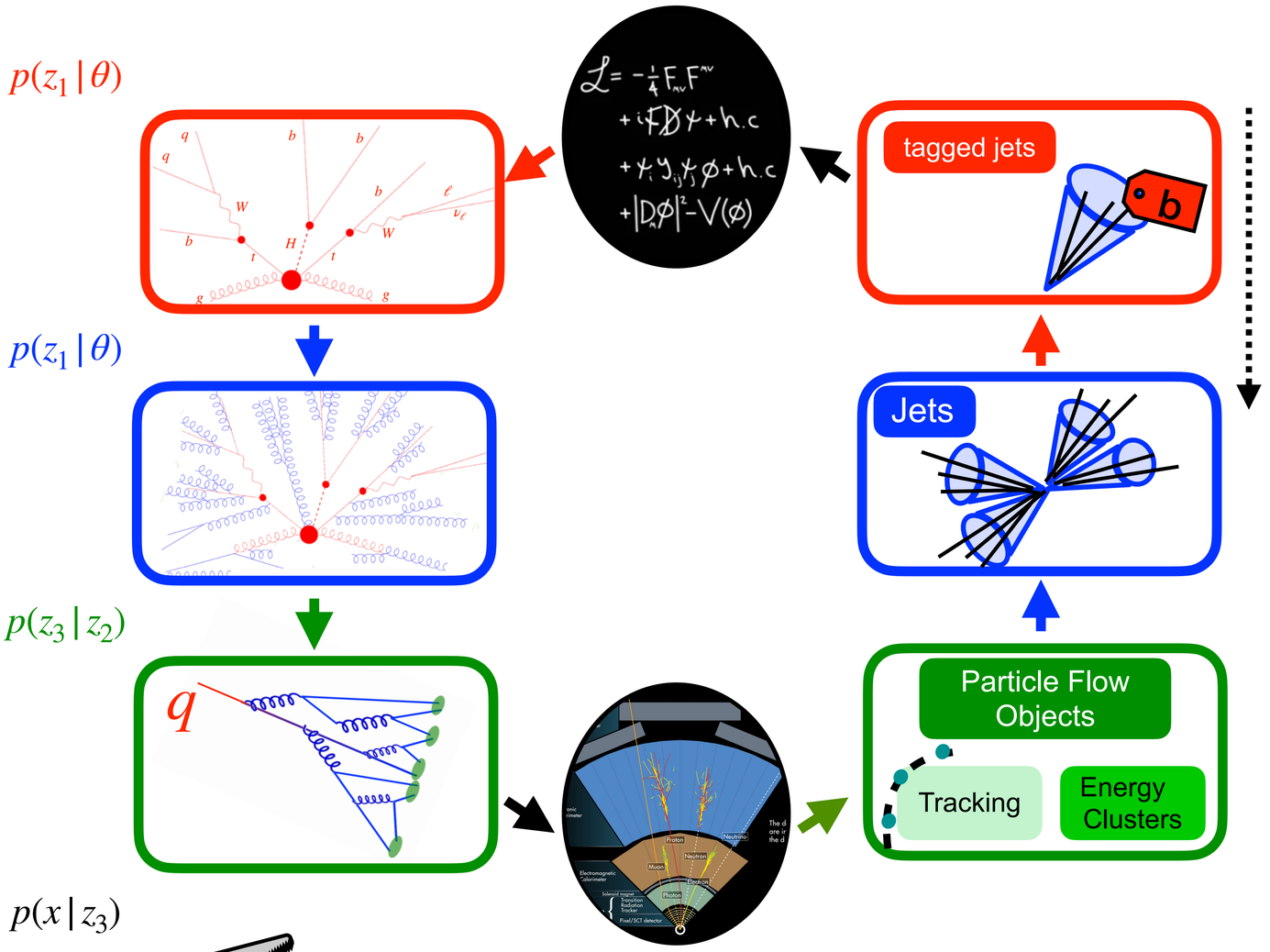




reweighting

flows

In conclusion…



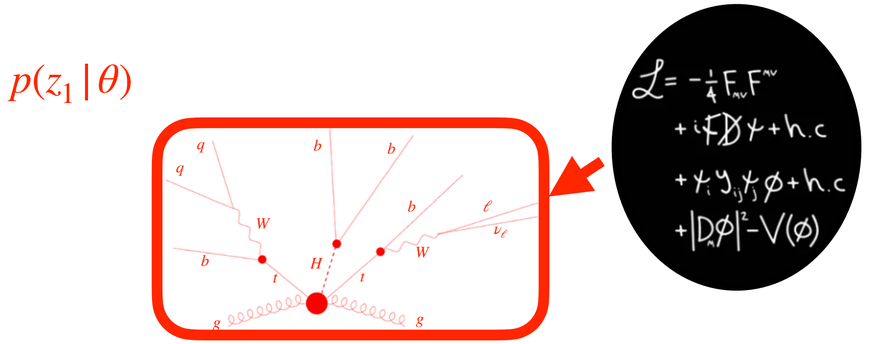
b-tagging

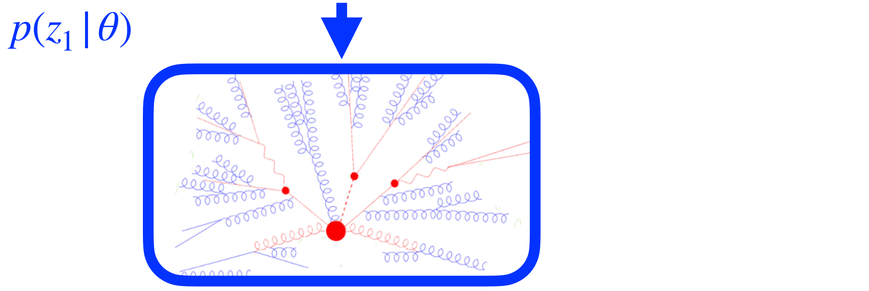
more inputs;
bigger model, better perf

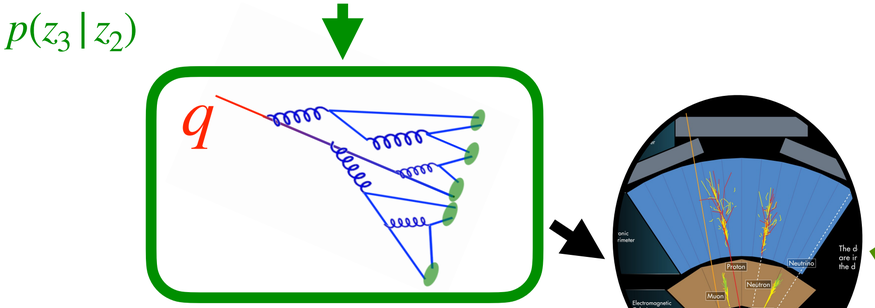
reweighting

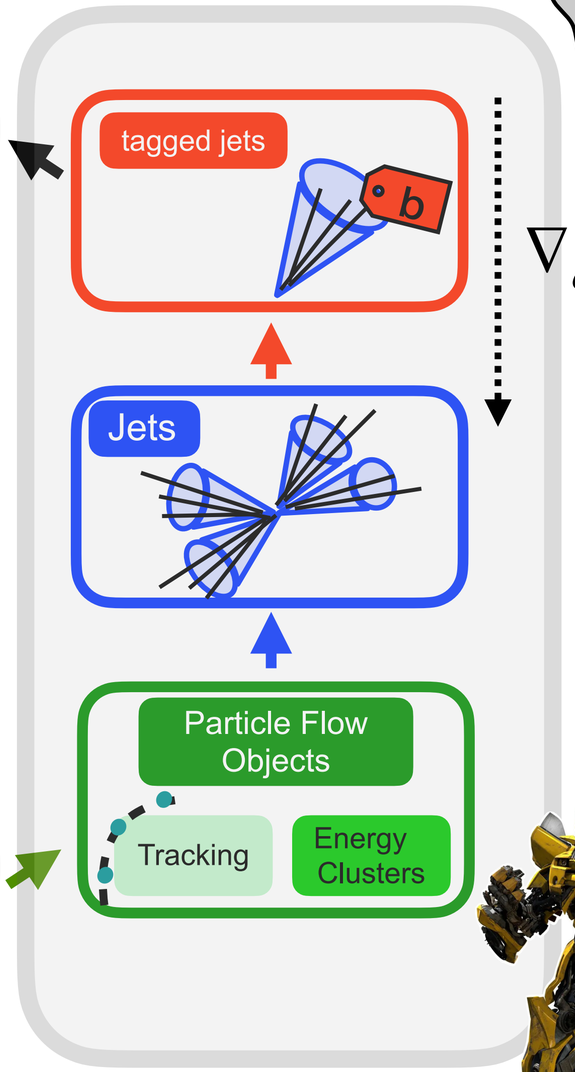
flows

In conclusion…





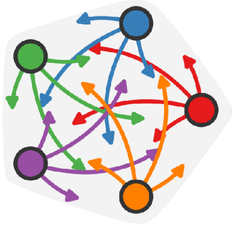




$\nabla_\phi \mathscr{L}$

b-tagging

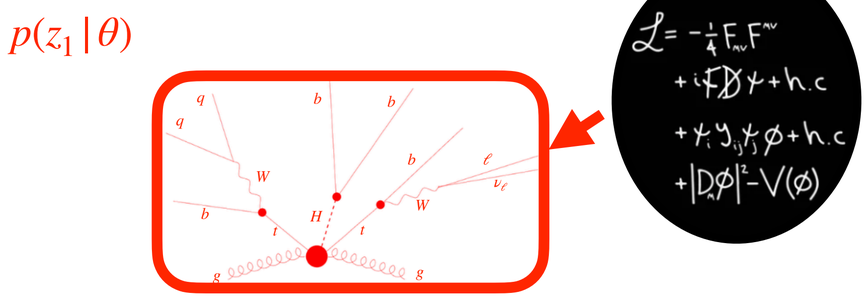
more inputs;
bigger model, better perf

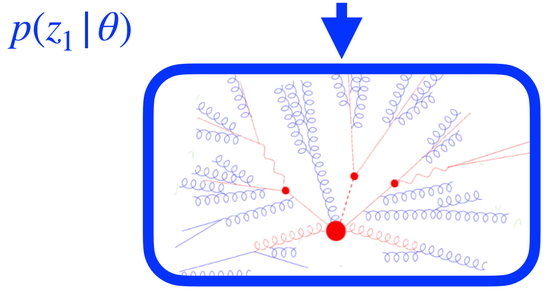


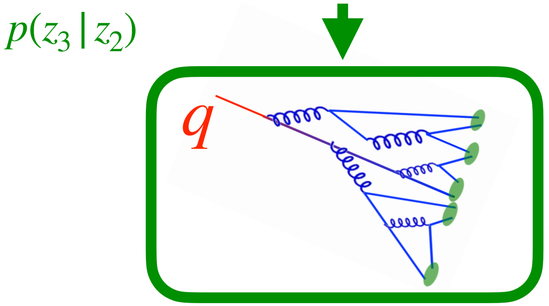
reweighting

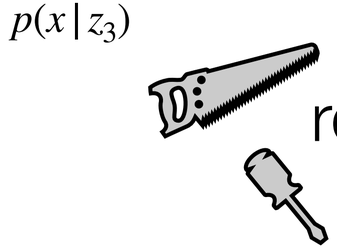
flows

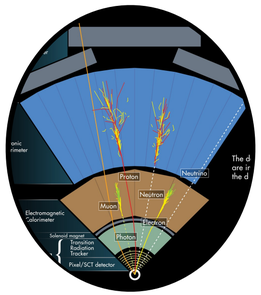
In conclusion…

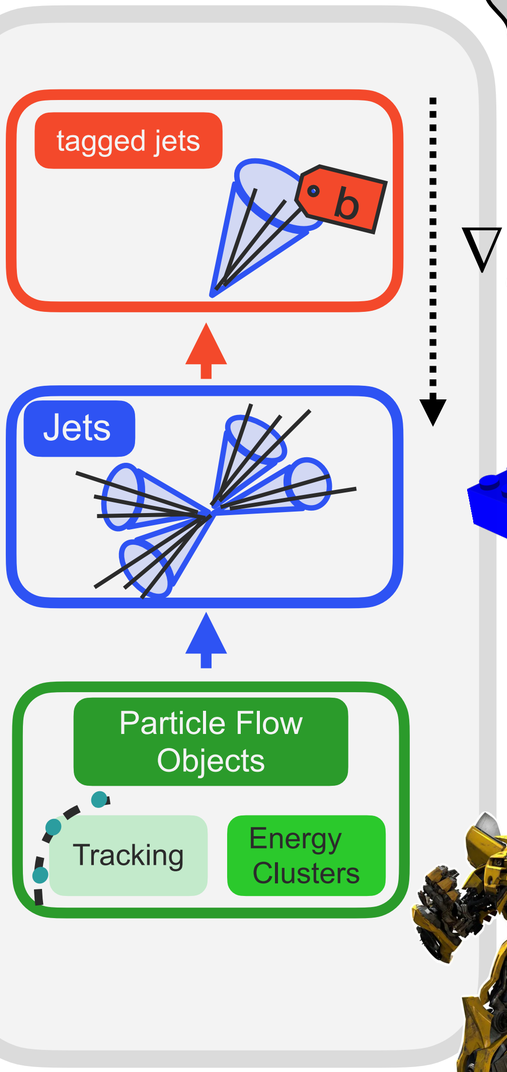












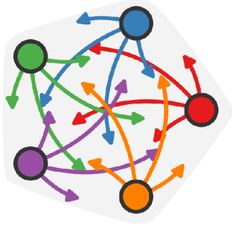
reweighting

flows

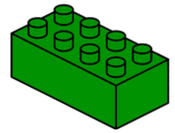
b-tagging

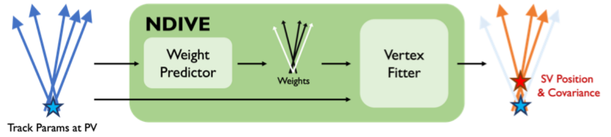
more inputs;

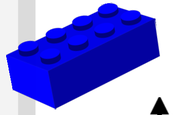
bigger model, better perf



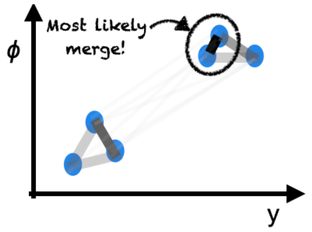
$\nabla_\phi \mathcal{L}$



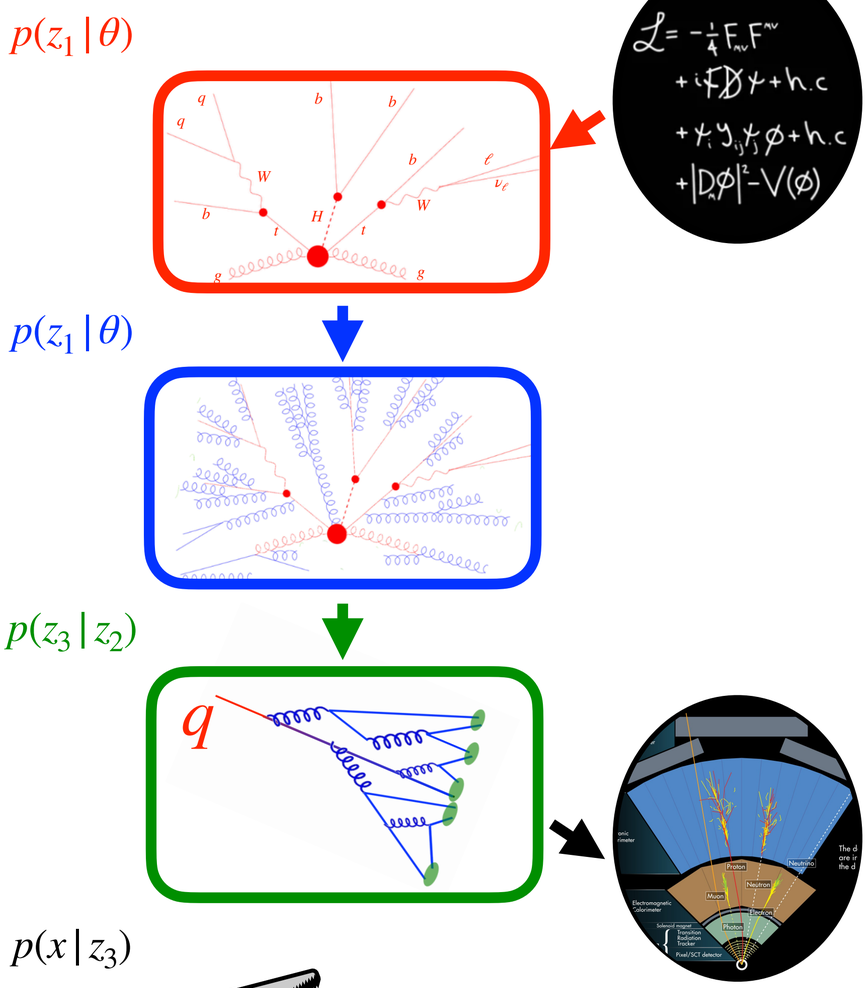


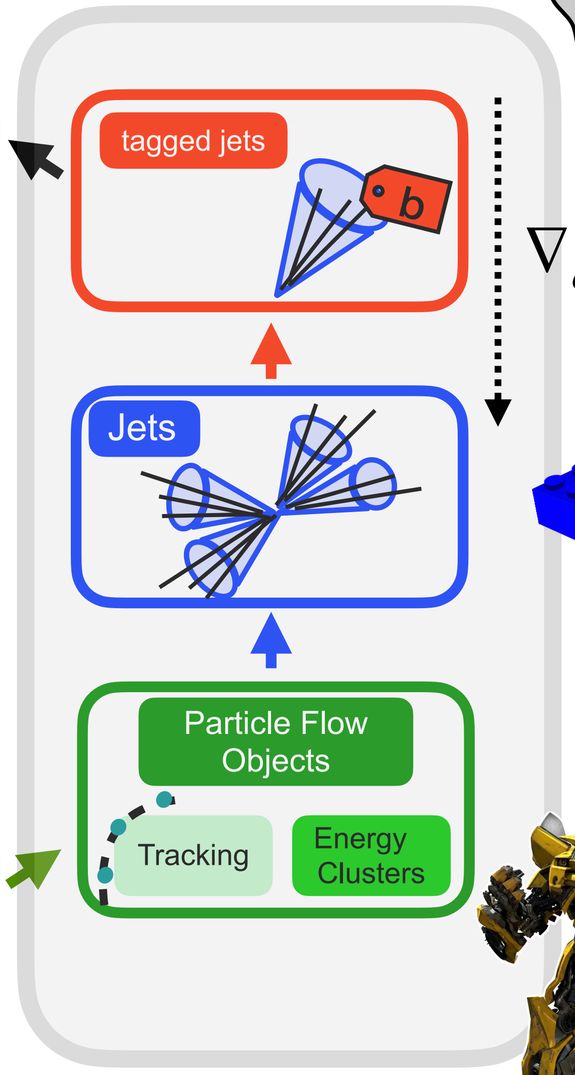


$\nabla_\theta \log p(x;\theta)$

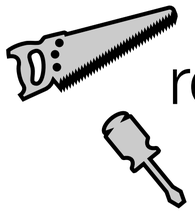


In conclusion…





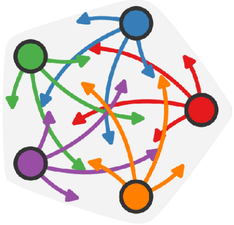
$\nabla_\phi \mathcal{L}$

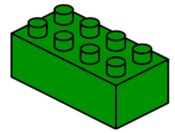


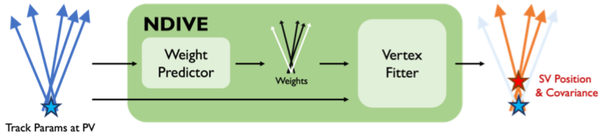
reweighting flows

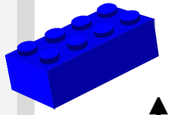
b-tagging

more inputs;
bigger model, better perf

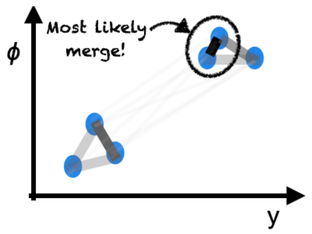








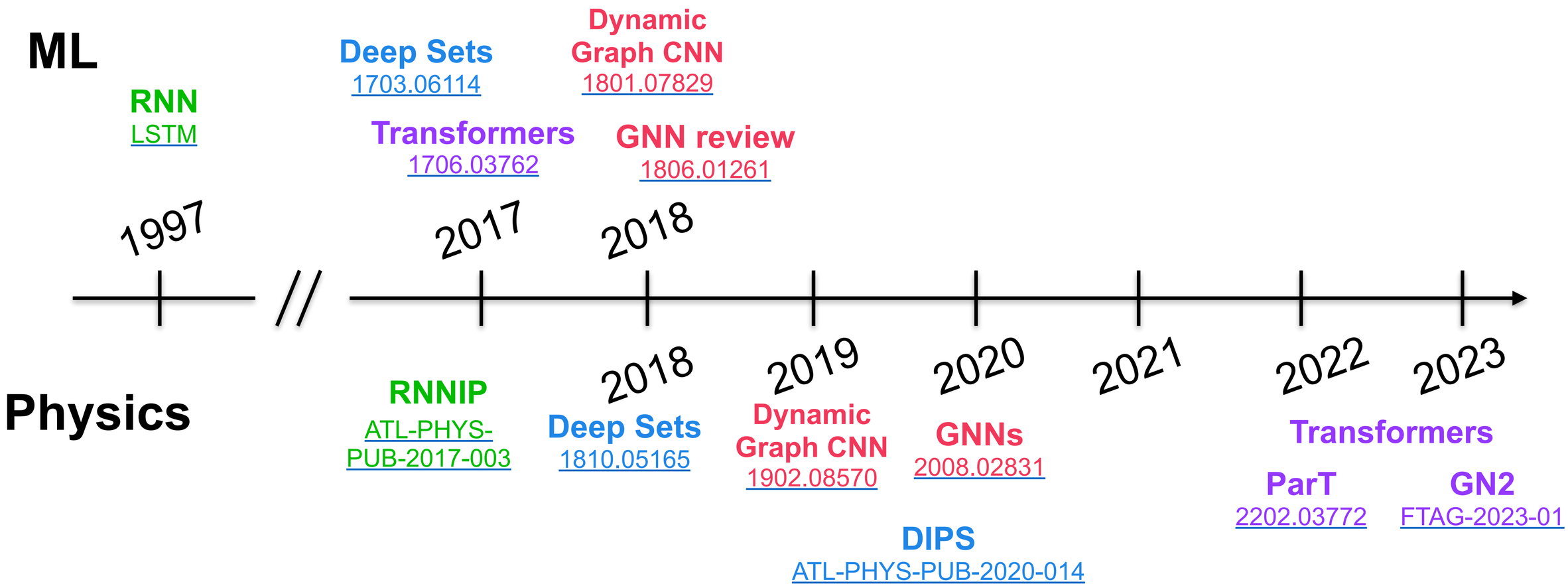
$\nabla_\theta \log p(x;\theta)$



Thanks!

Backup

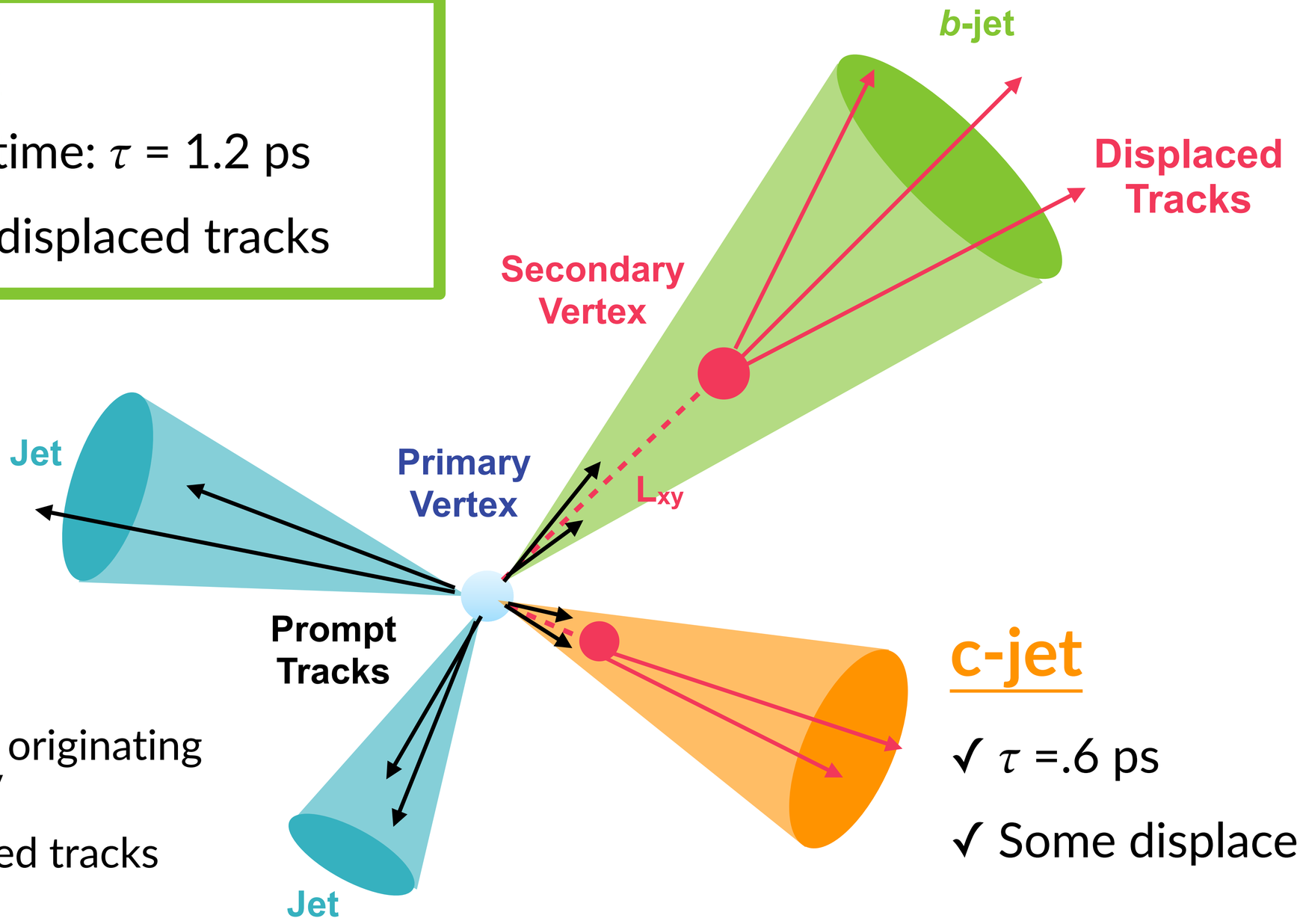
Timeline


ML
RNN
LSTM
Deep Sets
1703.06114
Dynamic
Graph CNN
1801.07829
Transformers
1706.03762
GNN review
1806.01261
1997
2017
2018
2018
2019
2020
2021
2022
2023
Physics
RNNIP
ATL-PHYS-PUB-2017-003
Deep Sets
1810.05165
Dynamic
Graph CNN
1902.08570
GNNs
2008.02831
DIPS
ATL-PHYS-PUB-2020-014
Transformers
ParT
2202.03772
GN2
FTAG-2023-01

😍 b-jet

✓ “Long” lifetime: τ = 1.2 ps

✓ Many (≈ 5) displaced tracks



light jet

u,d,s g

✓ Most tracks originating from the PV

✓ Few displaced tracks

c-jet

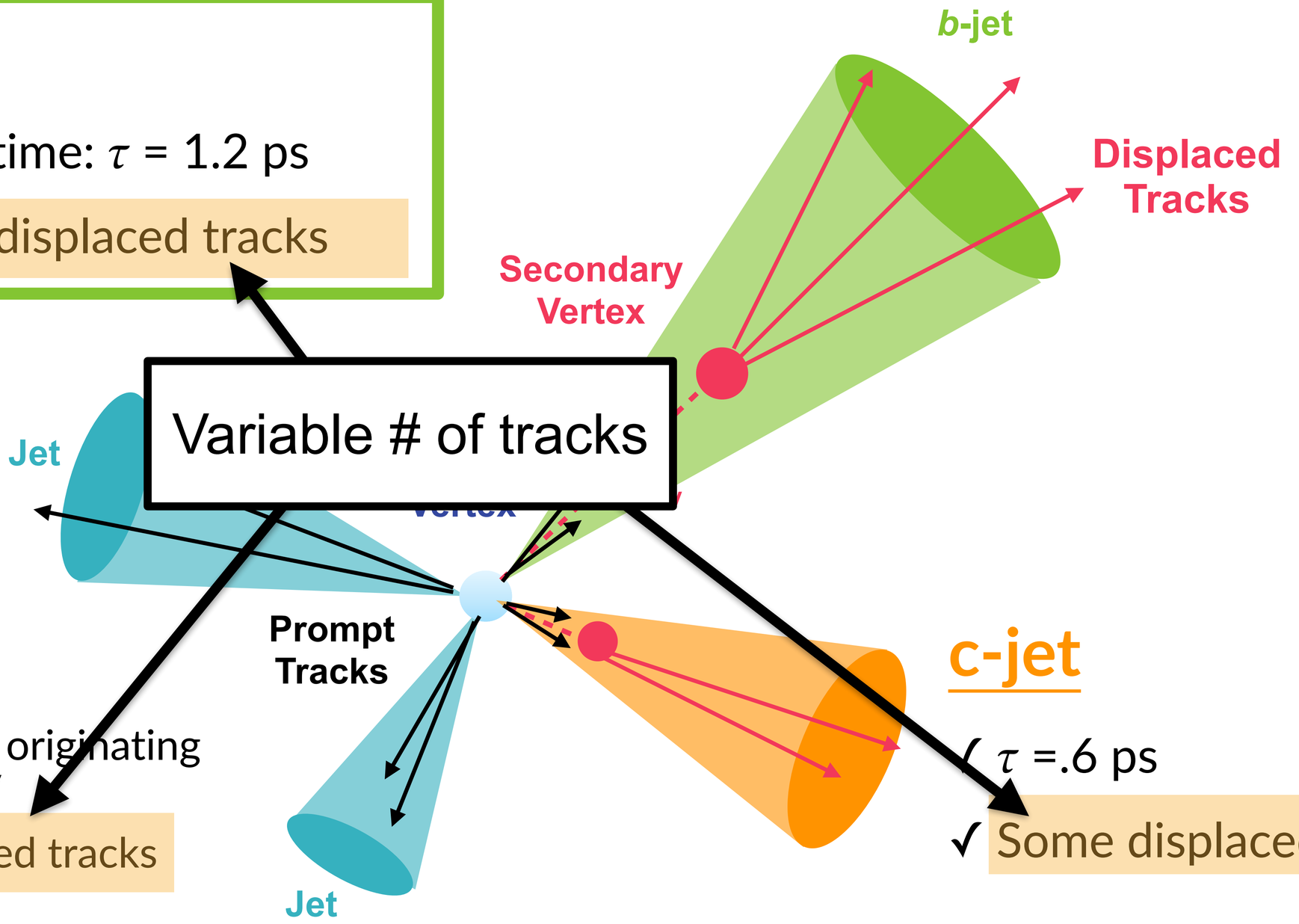
✓ τ =.6 ps

✓ Some displaced tracks

😍 b-jet

✓ “Long” lifetime: τ = 1.2 ps

✓ Many (≈ 5) displaced tracks



light jet

u,d,s g

✓ Most tracks originating from the PV

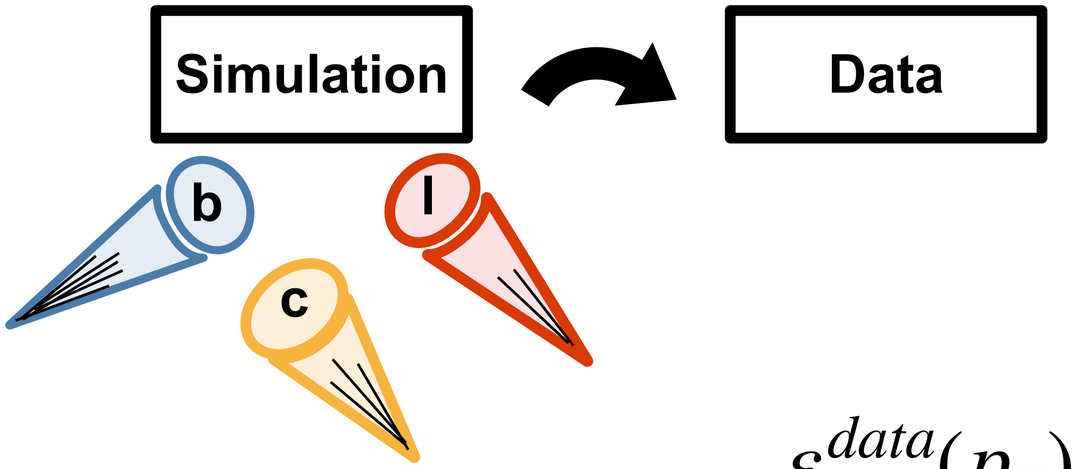
✓ Few displaced tracks

c-jet

✓ τ =.6 ps

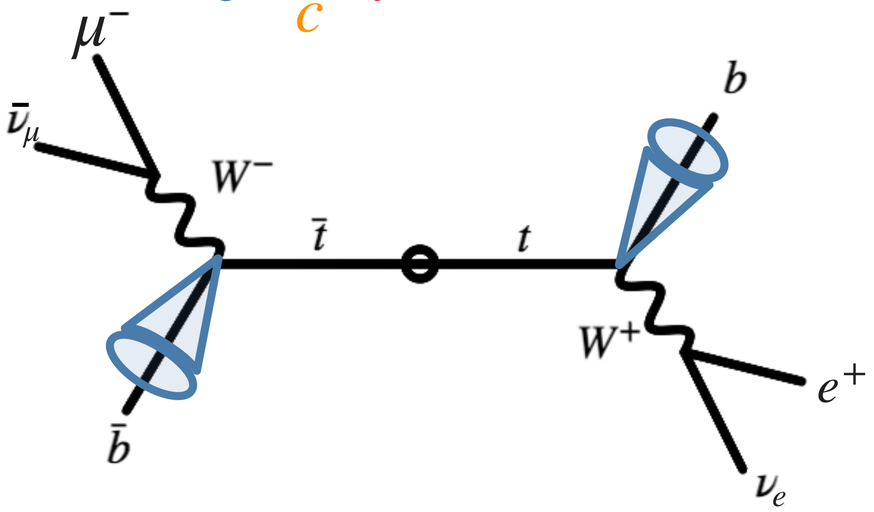
✓ Some displaced tracks

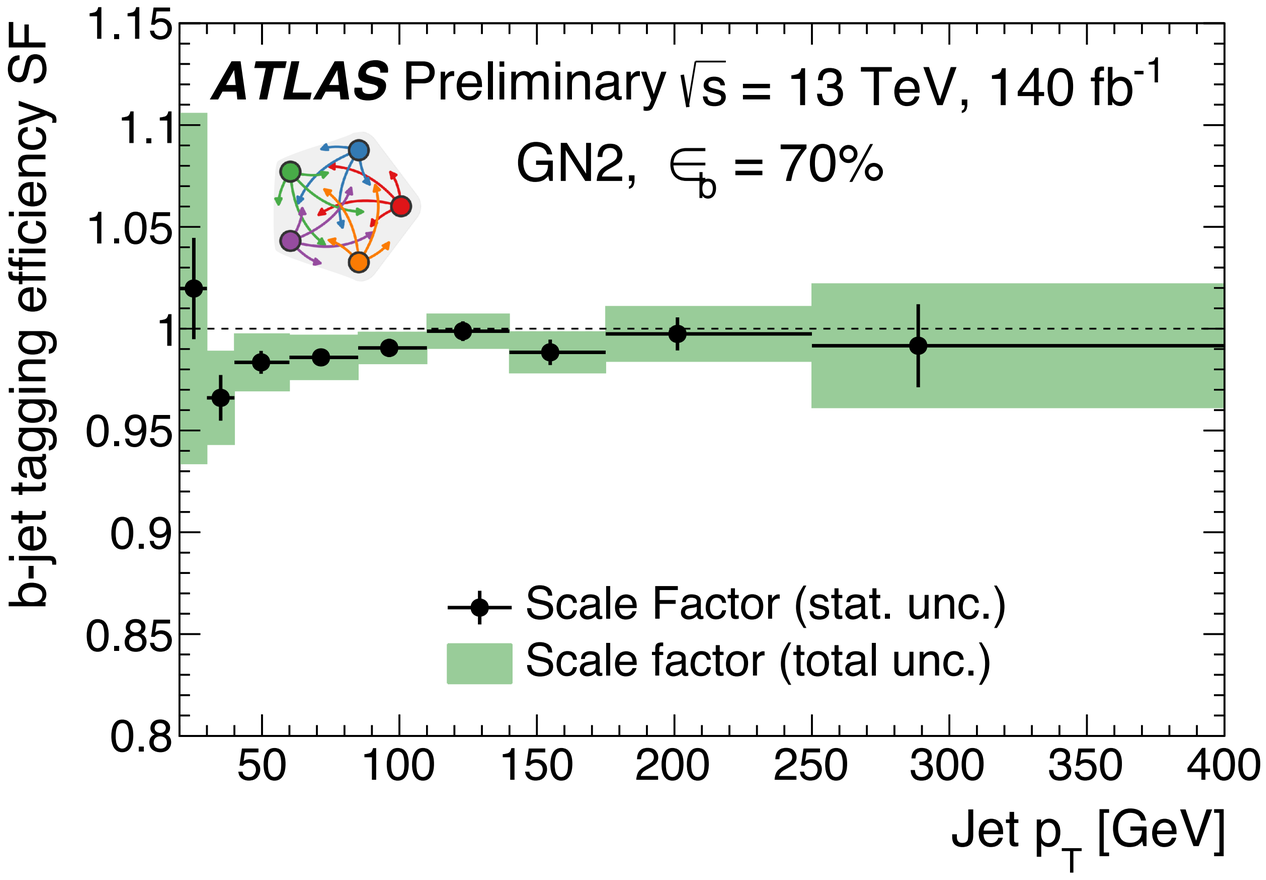
Calibration



$$SF_f(p_T) = \frac{\varepsilon_f^{data}(p_T)}{\varepsilon_f^{MC}(p_T)}$$

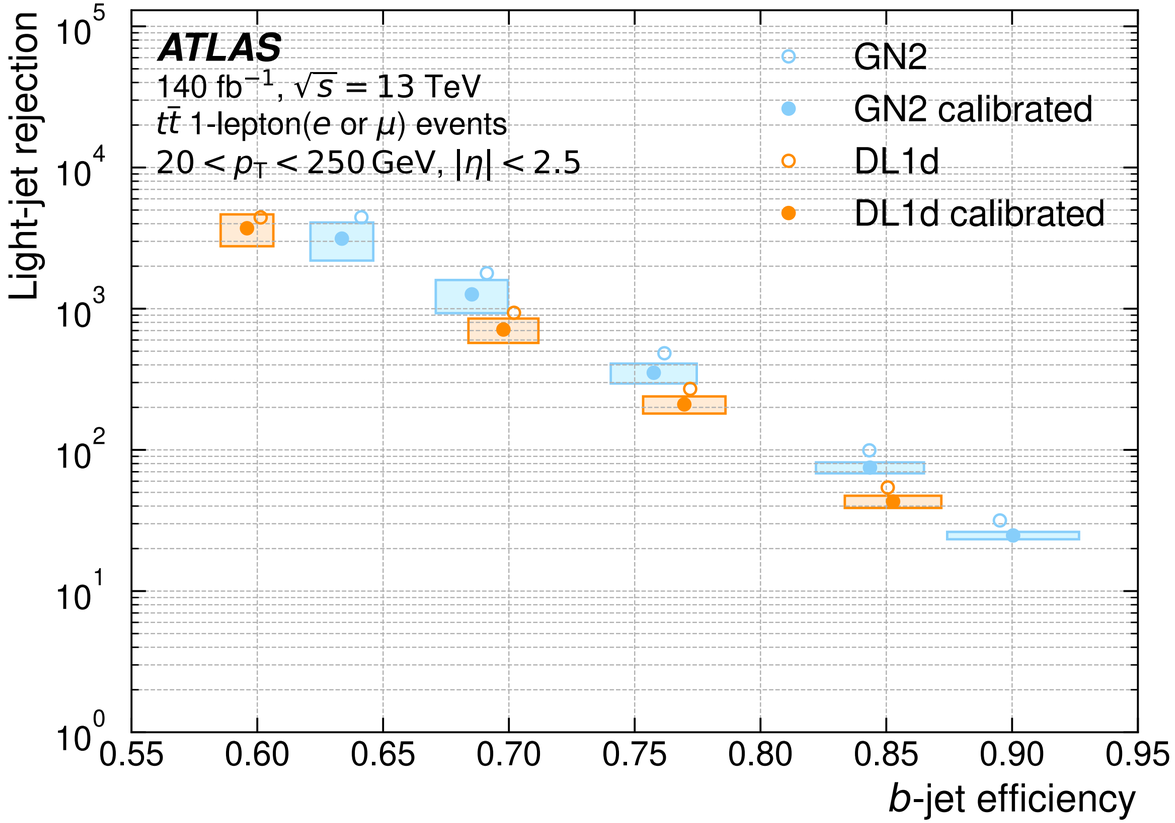
b c l

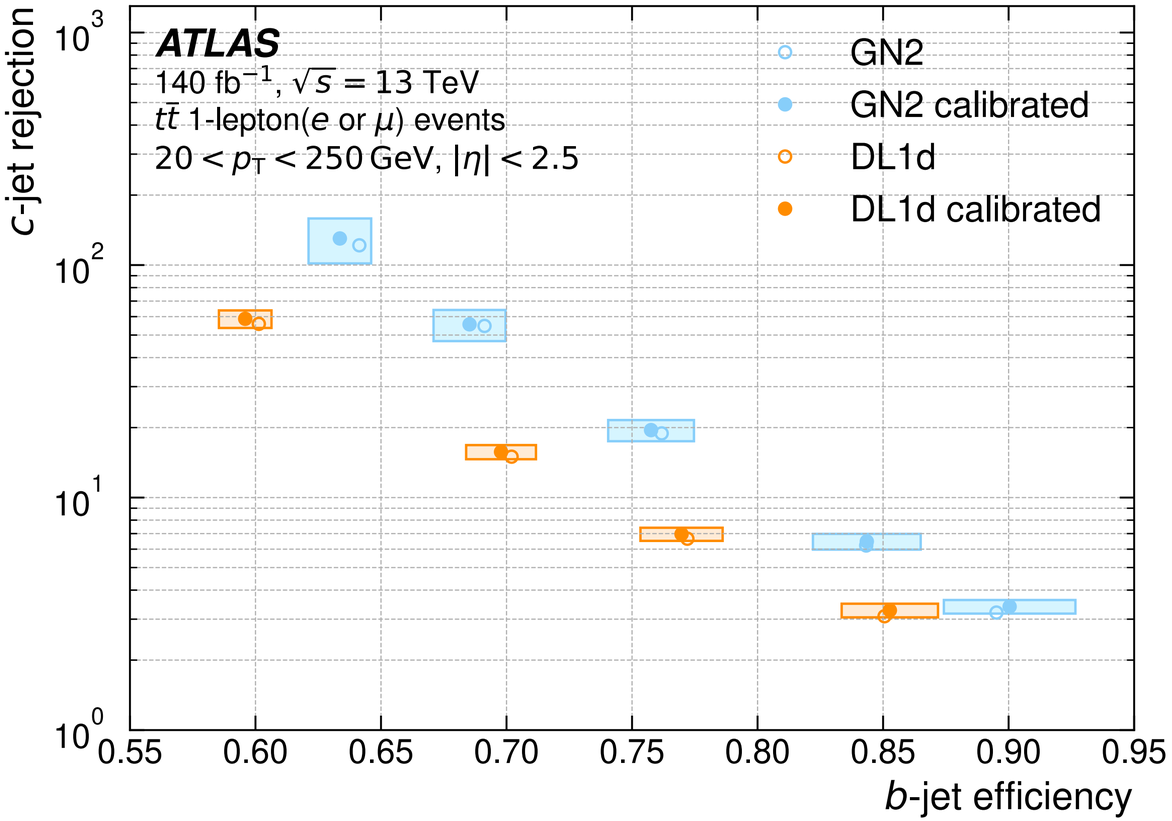




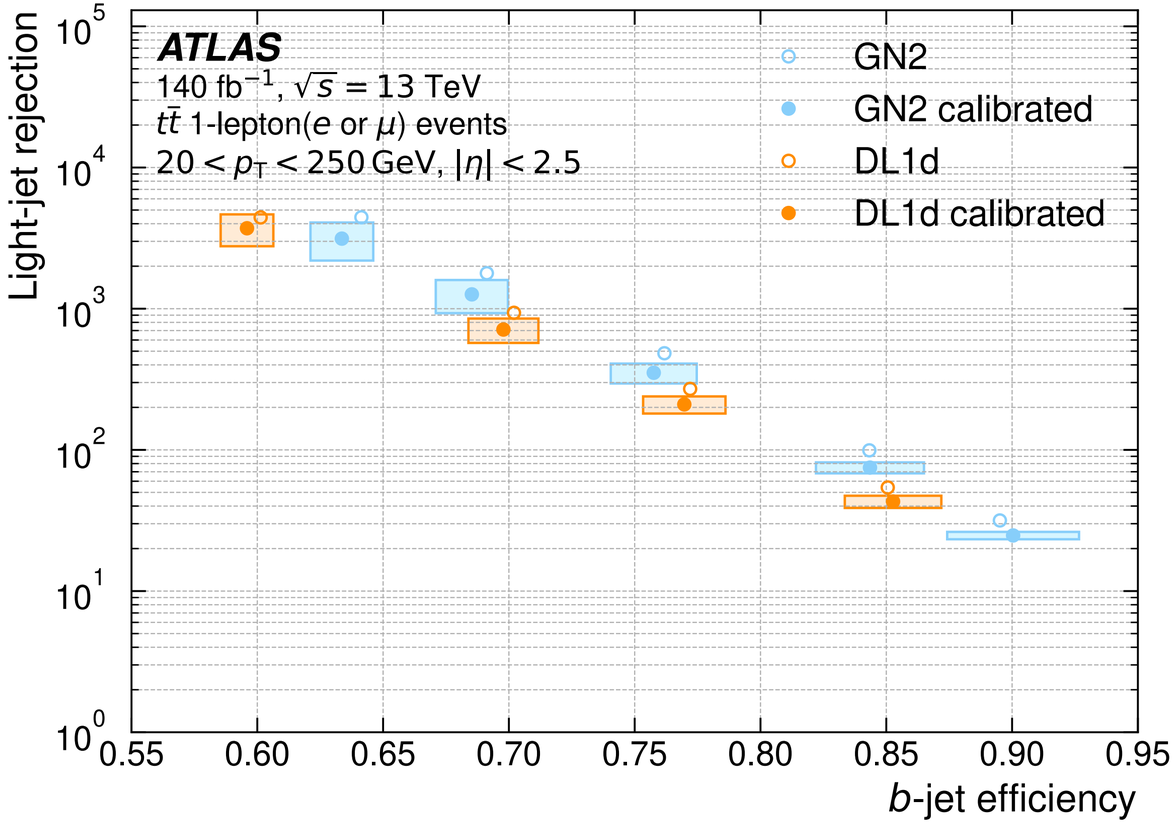
😀 % level non-closure

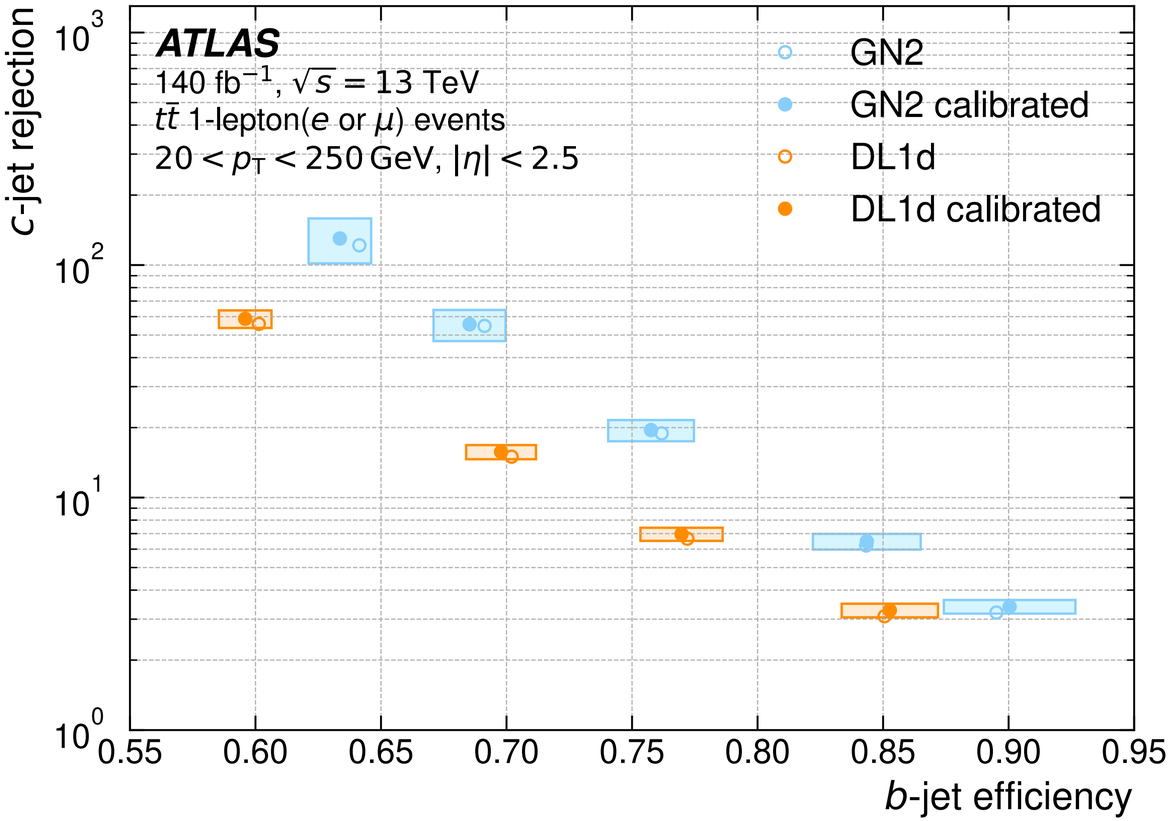
And… it translates to the physics (!)





And… it translates to the physics (!)

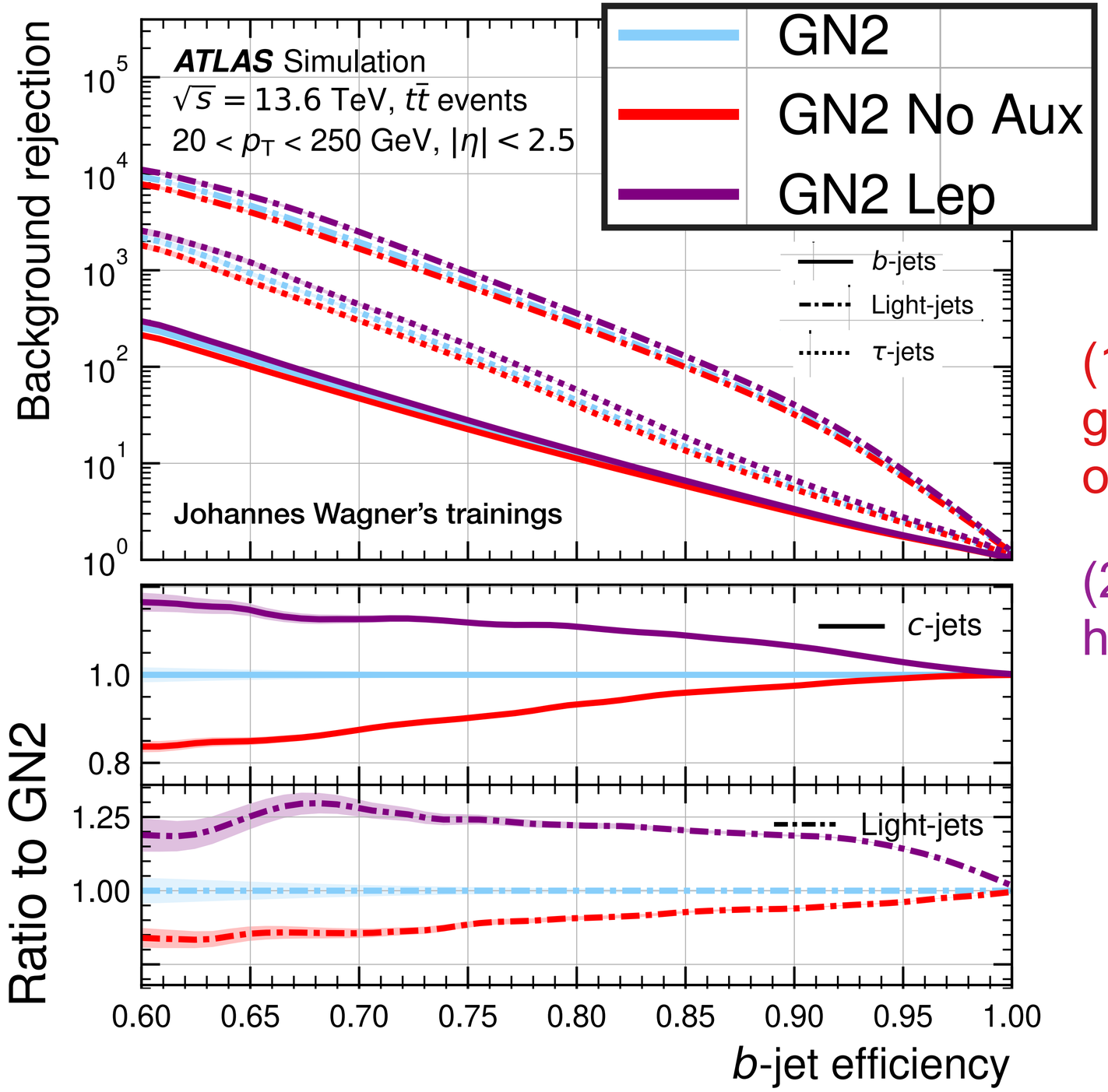




GN2 used in ***new*** $HH \to bb\gamma\gamma$ analysis (with 2022 — 2024 dataset)

Physcics briefing

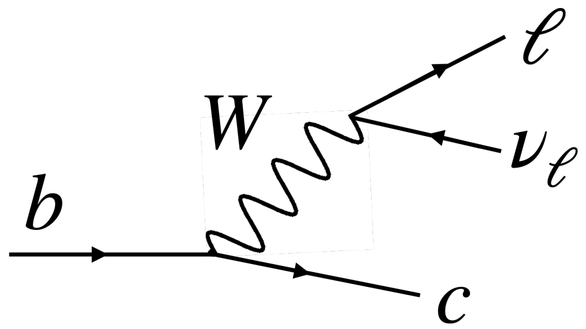
2505.19689



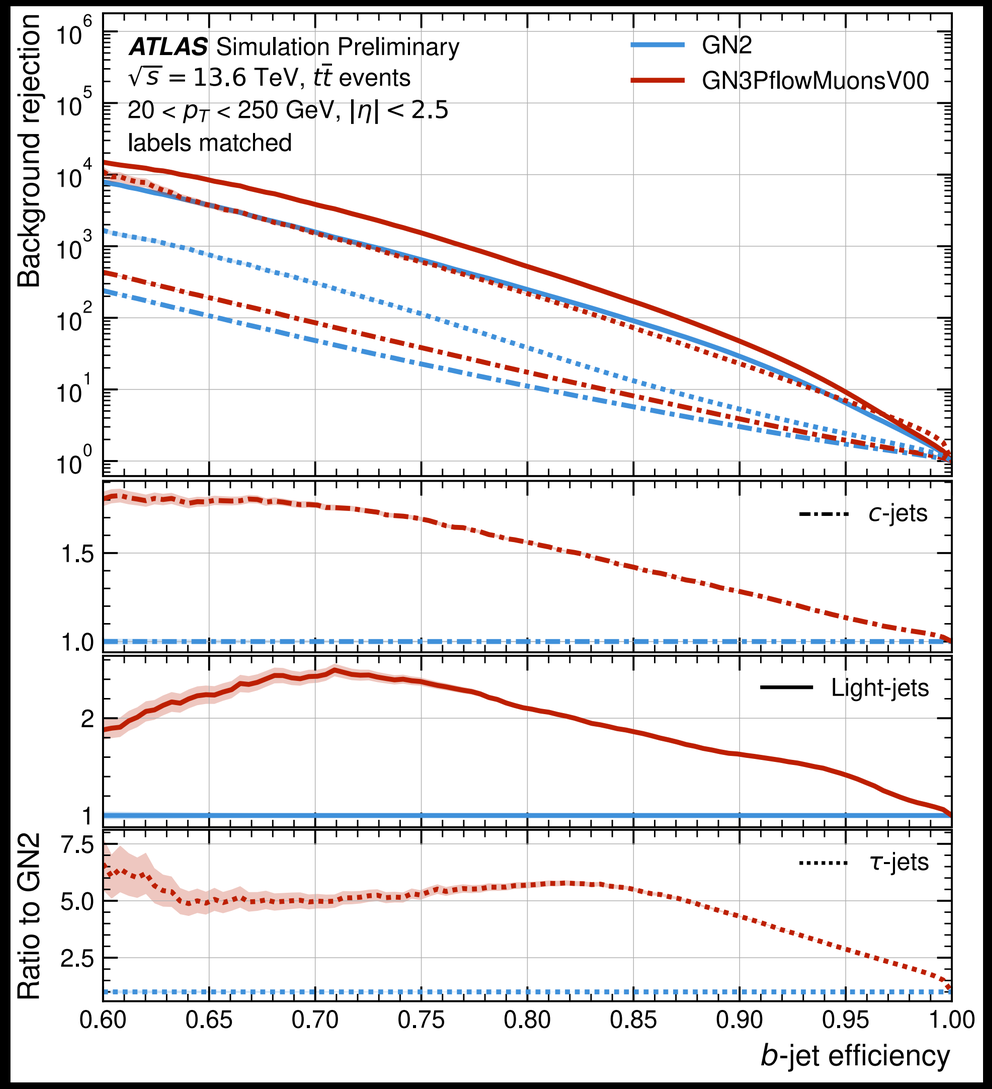
(1) The auxiliary tasks guide the model to more optimal solutions

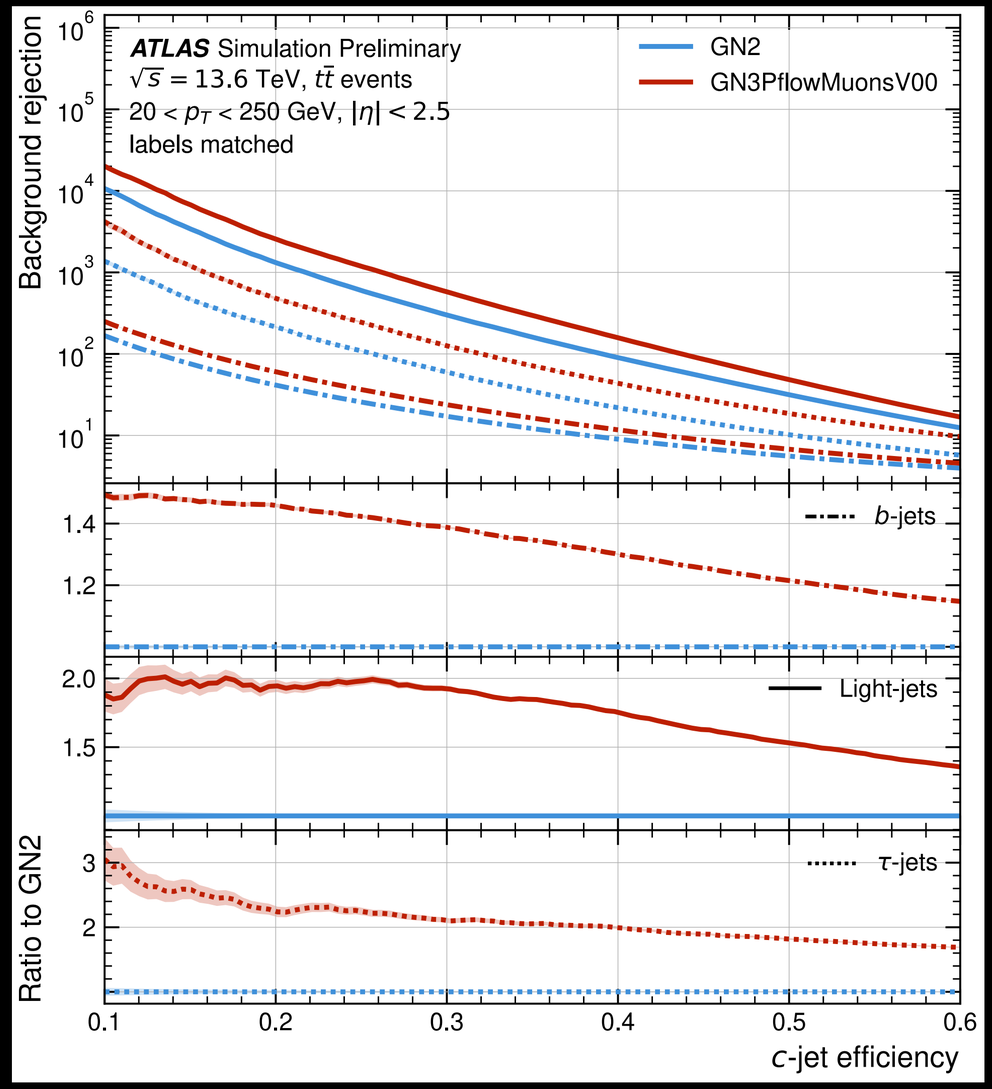
(2) Extra physics info helps more

GN3



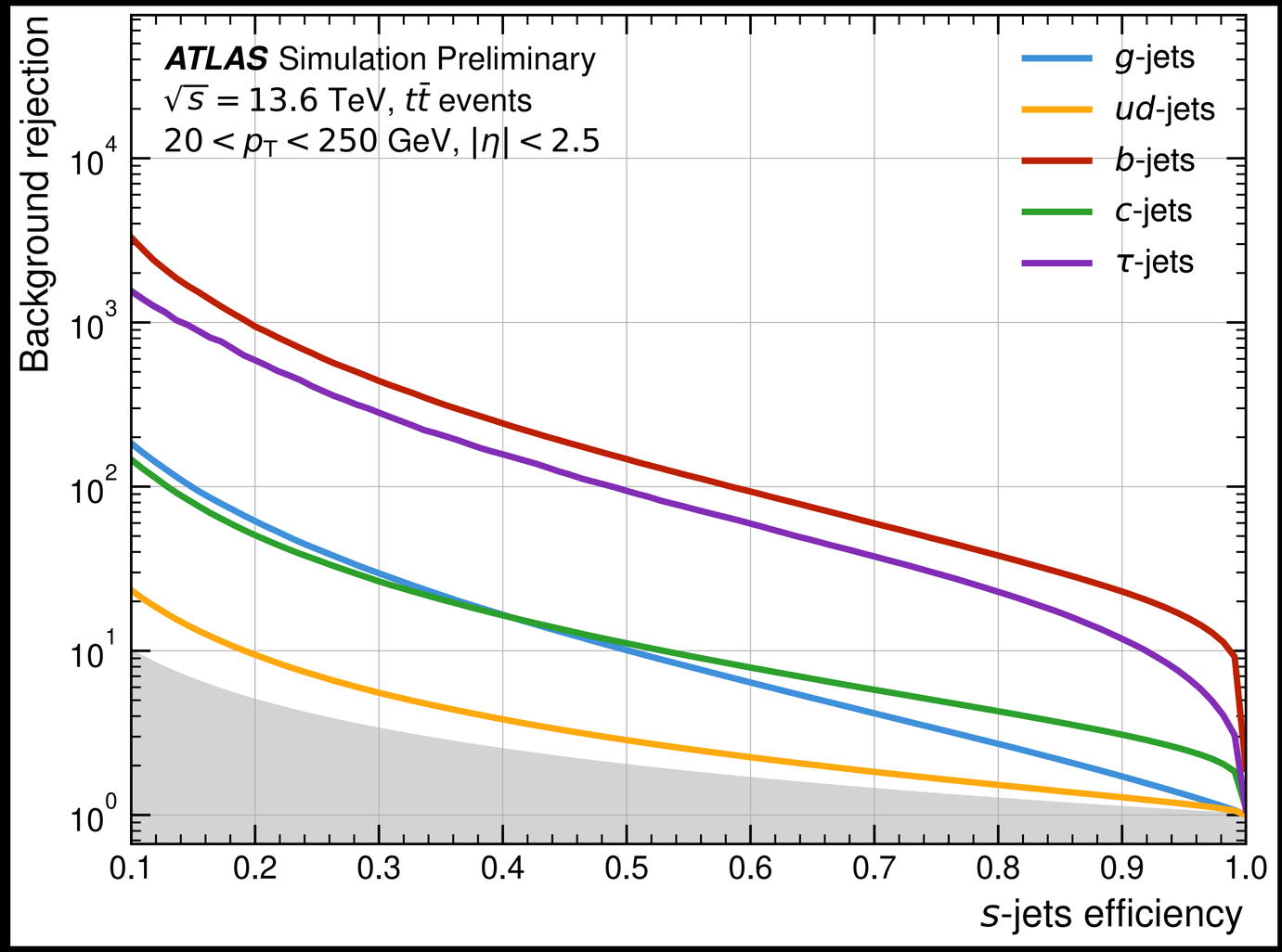
GN3: Multi-task, multi-modal b-tagger





GN3: Multi-task, multi-modal b-tagger

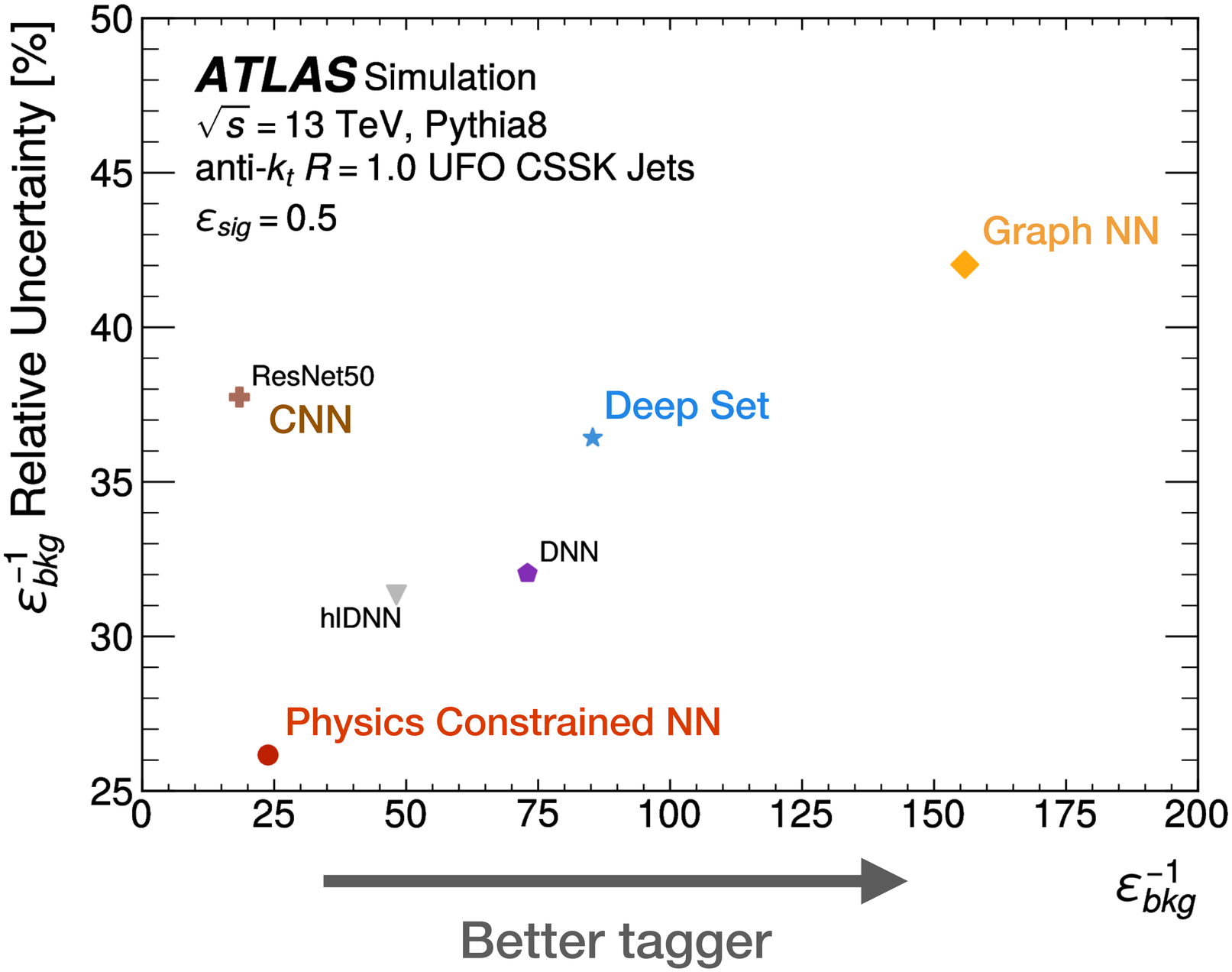
Component	GN2	GN3
Track–jet association	ΔR–based association	Ghost-association
Track selection	$d_0 < 3.5$ mm, 40 tracks	$d_0 < 5$ mm, 50 tracks
Inputs	Jets and tracks	Jets, ghost-associated tracks, soft muons, PFlow objects
Activation function	ReLU	SiLU
Initialisation layers	256	512
Transformer encoder	4 scaled dot-product attention	4 Flash Attention, Gated Linear Units, 8 register tokens
Embedding dimension	256	512
Jet classification	4 classes: b, c, ℓ, τ	6 classes: b, c, ud, s, g, τ
Loss balancing	Fixed weights per task	Geometric mean of losses (GLS)
Optimiser	AdamW	Lion



Rob Les slides
2407.20127

A word of caution

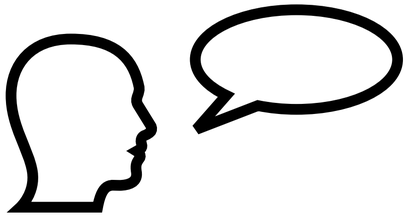
Larger uncertainties



2019: The bitter lesson (Richard Sutton)

"The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin."

Exs:



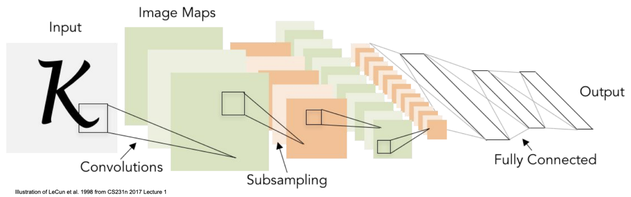
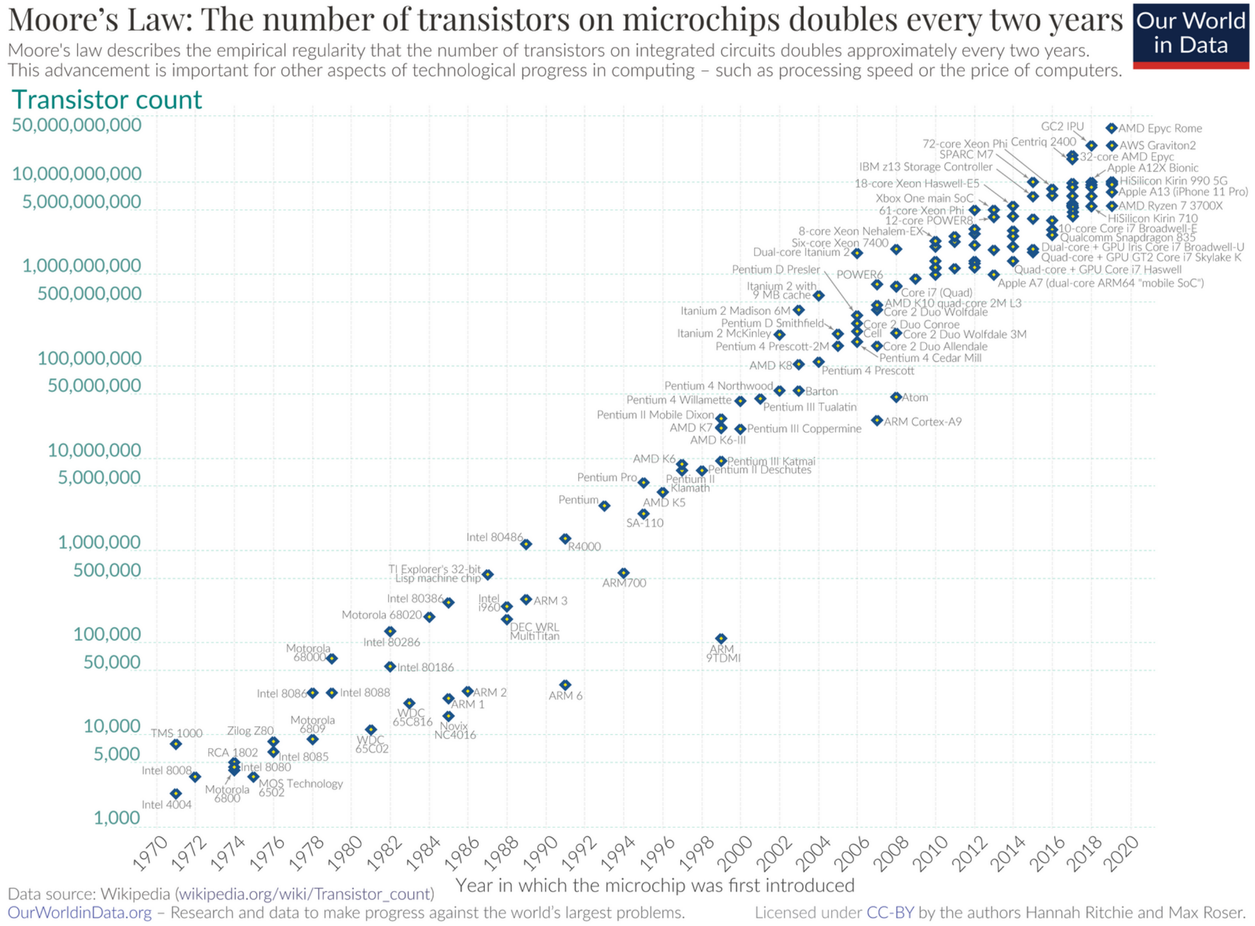


Illustration of LeCun et al. 1998 from CS231n 2017 Lecture 1



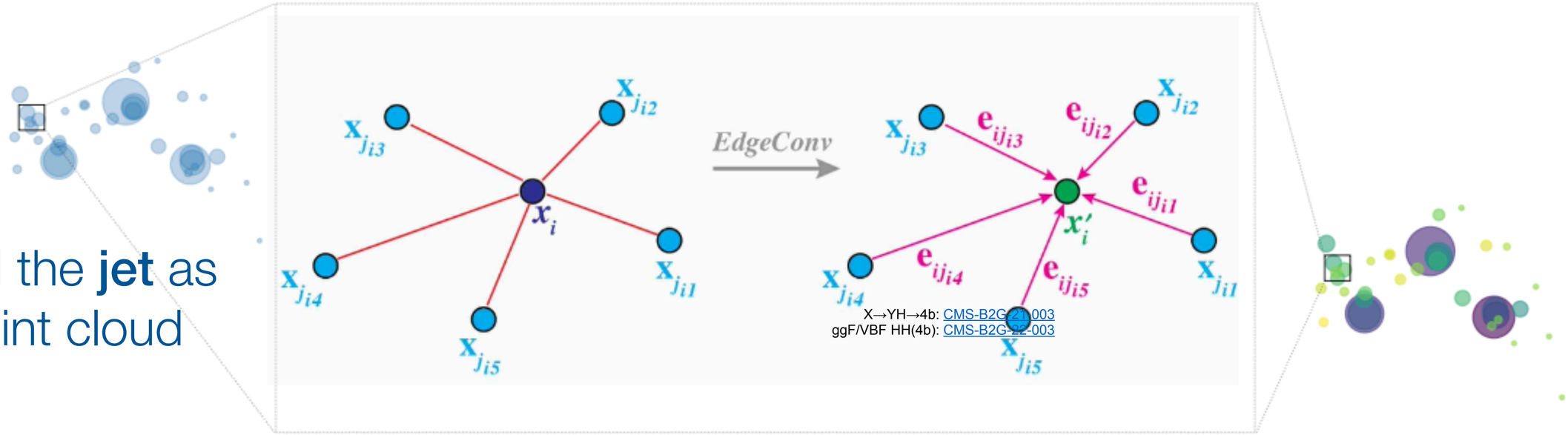
Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org – Research and data to make progress against the world's largest problems.

Huilin Qu's slide
Phys.Rev.D101(2020)5,056019

Dynamic Graph CNN

"ParticleNet" [CMS]

Model the **jet** as a point cloud

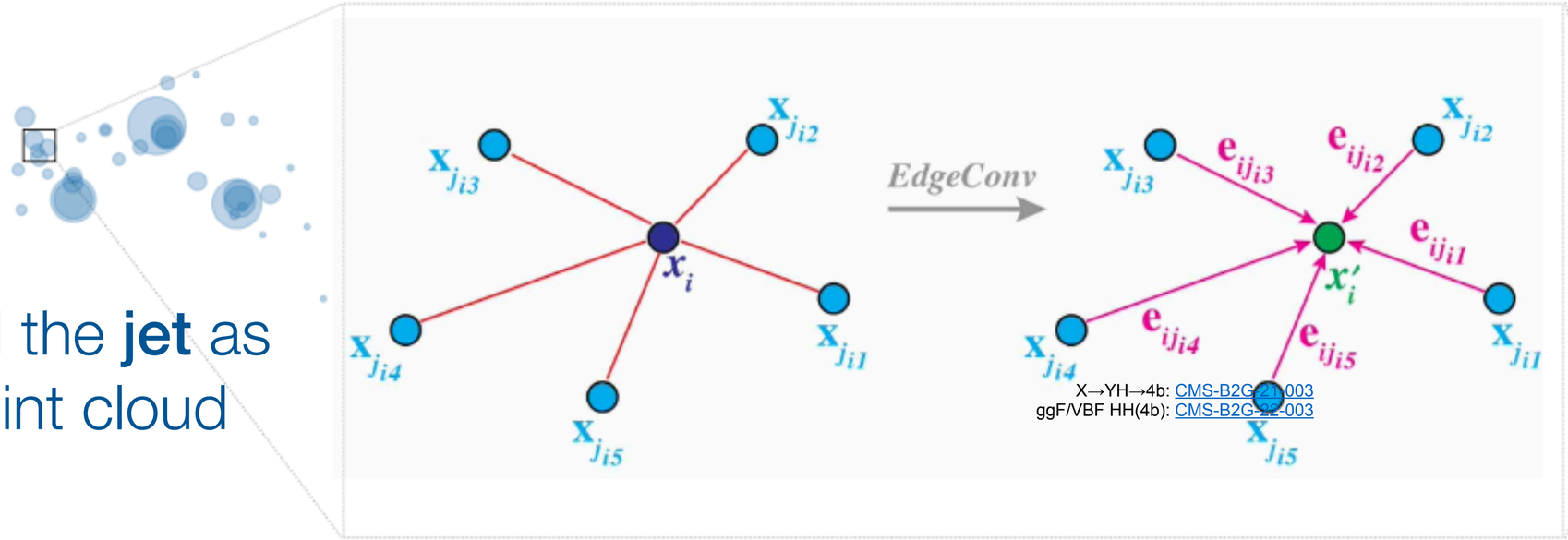


Huilin Qu's slide
Phys.Rev.D101(2020)5,056019

Dynamic Graph CNN

"ParticleNet" [CMS]

Model the **jet** as a point cloud

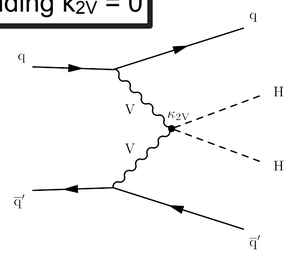


Very impressive physics results

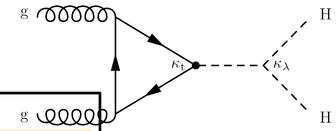
But was the graph representation needed?

VBF HH→4b

First analysis excluding $\kappa_{2V} = 0$



ggF HH→4b



Boosted:

Obs (exp): 9.9 (5.1)

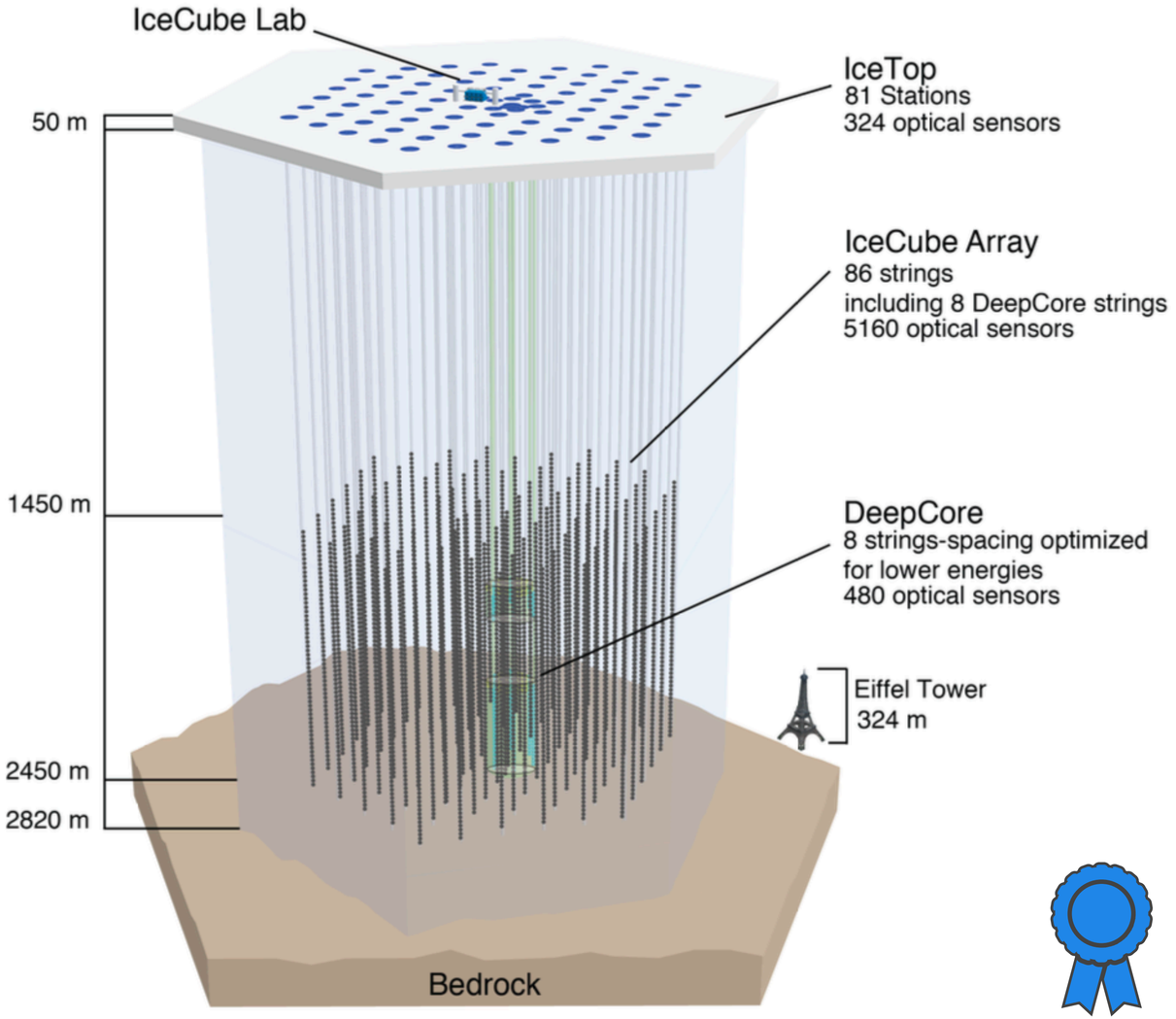
Resolved:

ATLAS Obs (exp): 5.4 (8.1)

CMS Obs (exp): 3.8 (7.8)


Quarks
u c t
up charm top
d s b
down strange bottom
Forces
Z γ
Z boson photon
W g
W boson giuon
H
Higgs boson
e μ τ
electron muon tau
νe νμ ντ
electron neutrino muon neutrino tau neutrino
Leptons
How did this story evolve for neutrino identification?

2310.15674



kaggle competition

3.1.4 Edge Selection

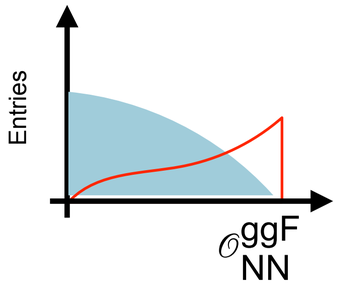
In the original EdgeConv implementation, edges are calculated in each layer dynamically by k-Nearest Neighbors (kNN). However, this edge selection scheme is not differentiable in itself and therefore does not have gradients. This would still work well for the segmentation task in the original paper, as points in the same segment are trained to be close in the latent space. However, the situation is different in this task, where it would not make sense to dynamically select edges. Therefore, the edges used in EdgeConv are calculated from the input features only once in our model.

All three of the winning solutions used a transformer architecture.

Also in jet tagging (ATLAS and CMS), transformers outperform GNN architectues.

Simulation based inference

Traditionally, train MVA & histogram-ize



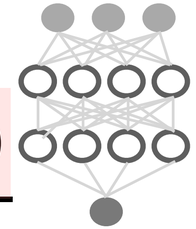
$$t_\mu = -2\log\frac{\mathscr{L}(\mu)}{\mathscr{L}(\hat{\mu})}$$

W/ interference, no single classifier is optimal $\forall\ \mu$

NEW!

$$t_\mu = -2\sum_{i=1}^{N_{evt}}\frac{p(x_i\,|\,\mu)}{p(x_i\,|\,\hat{\mu})} = -2\sum_{i=1}^{N_{evt}}\frac{p(x_i\,|\,\mu)/p'(x_i)}{p(x_i\,|\,\hat{\mu})/p'(x_i)}$$

NN(μ)



Trick: don't need NN(μ)

$$\frac{p(x\,|\,\mu)}{p'(x)} = \frac{1}{\sigma(\mu)}\left(\mu\sigma_s\frac{p_S(x)}{p'(x)} + \sqrt{\mu}\sigma_I\frac{p_I(x)}{p'(x)} + \sigma_B\frac{p_B(x)}{p'(x)}\right)$$

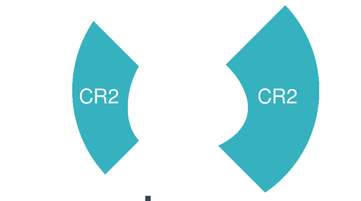
Neural Networks

Uncertainty on background model

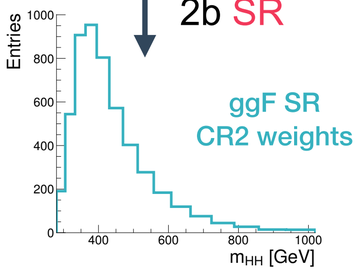
1. Choice of CR

Alternative estimate

Train NNs in CR2



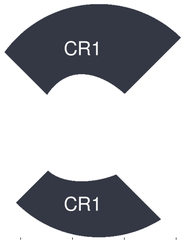
Apply to 2b SR



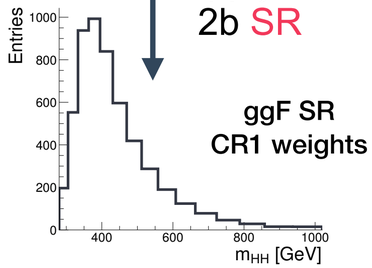
ggF SR
CR2 weights

Nominal estimate

Train NNs in **CR1**



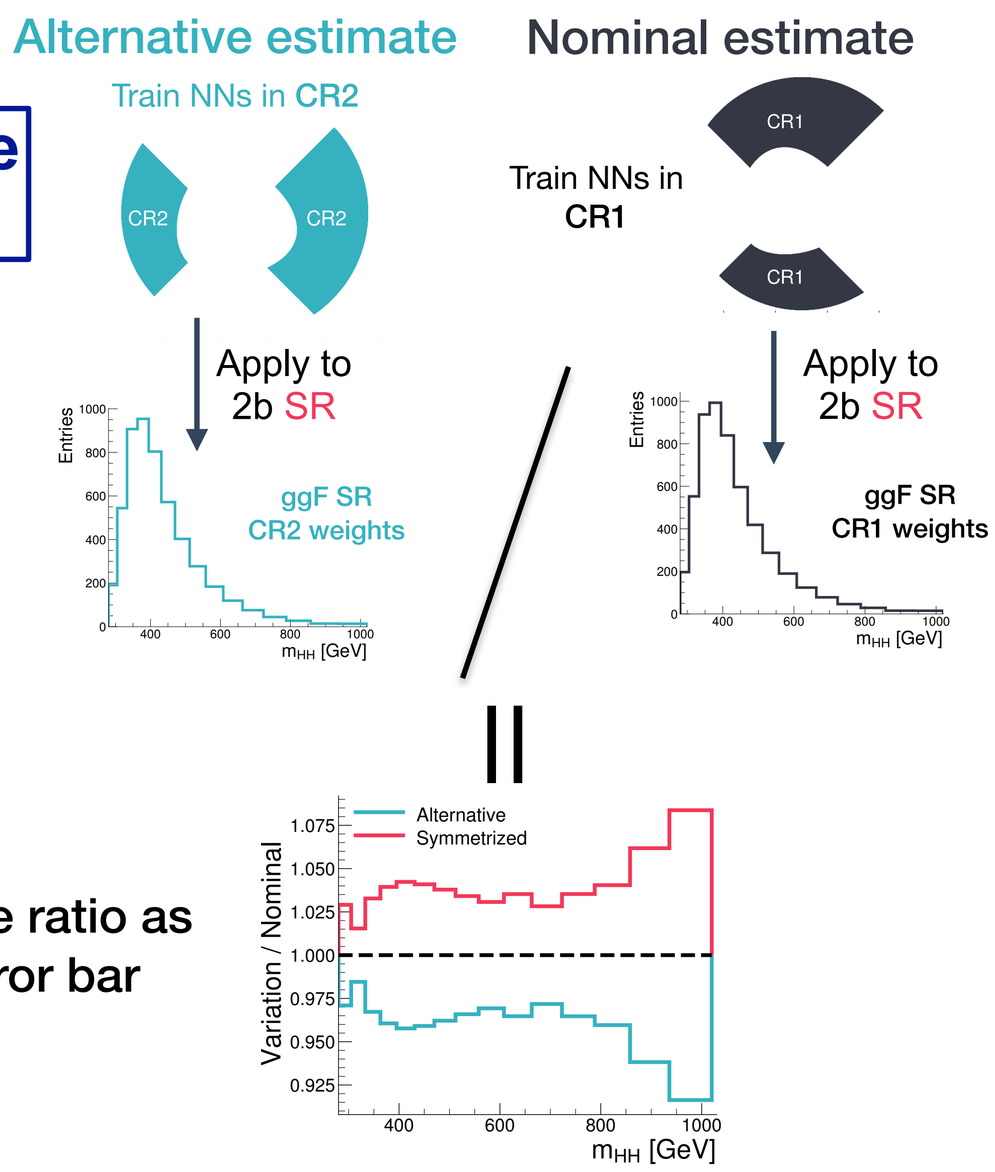
Apply to 2b SR



ggF SR
CR1 weights

Uncertainty on background model

1. Choice of CR



Take the ratio as an error bar

1612.01474 [3.6k citations]

Fig modified from 1912.02757

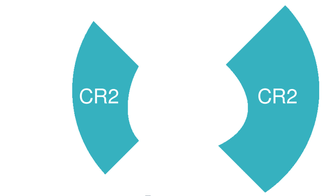
2301.03212

Uncertainty on background model

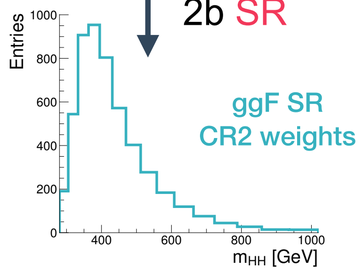
1. Choice of CR

Alternative estimate

Train NNs in CR2



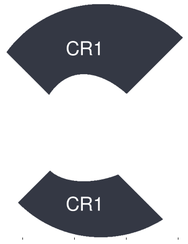
Apply to 2b SR



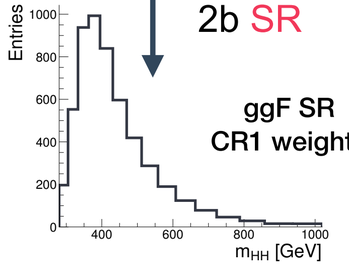
ggF SR CR2 weights

Nominal estimate

Train NNs in **CR1**



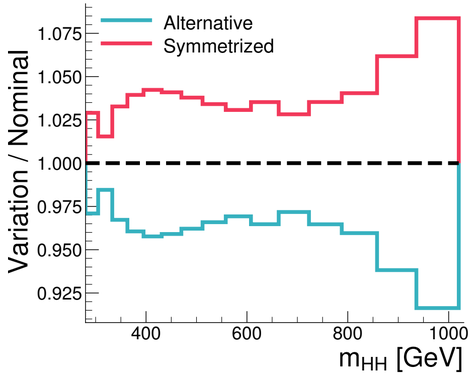
Apply to 2b SR



ggF SR **CR1 weights**

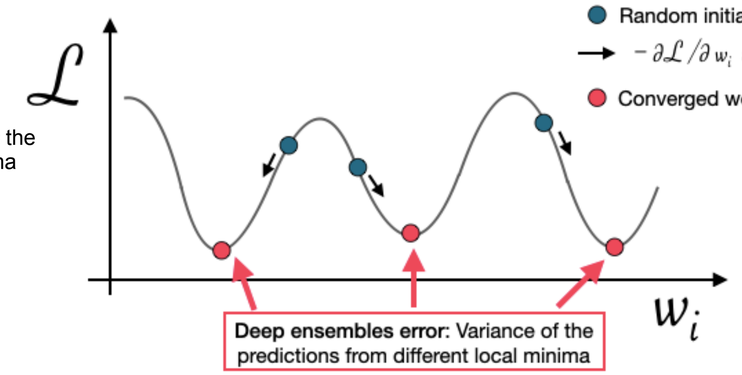
=

Take the ratio as an error bar



2. NN initialization

Retraining the NN 100x captures the uncertainty from multiple minima



1612.01474 [3.6k citations]

Fig modified from 1912.02757

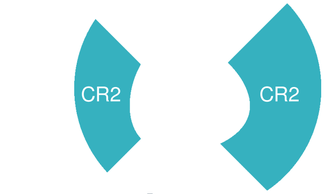
2301.03212

Uncertainty on background model

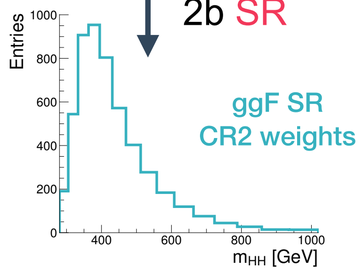
1. Choice of CR

Alternative estimate

Train NNs in CR2



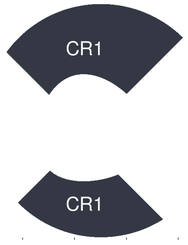
Apply to 2b SR



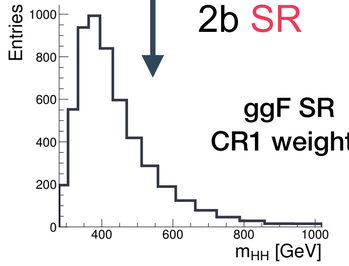
ggF SR CR2 weights

Nominal estimate

Train NNs in **CR1**



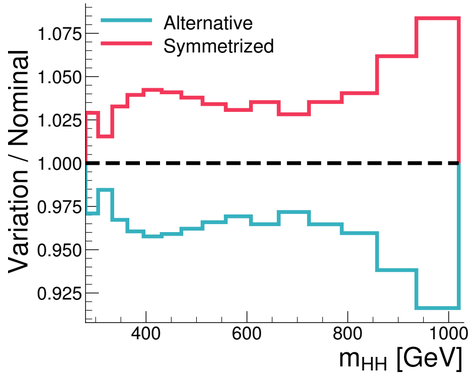
Apply to 2b SR



ggF SR CR1 weights

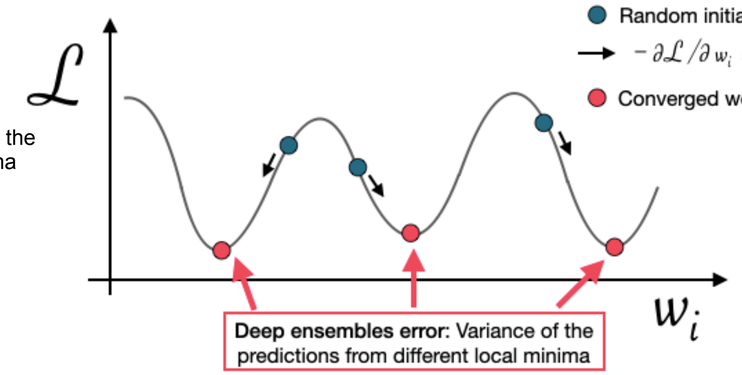
=

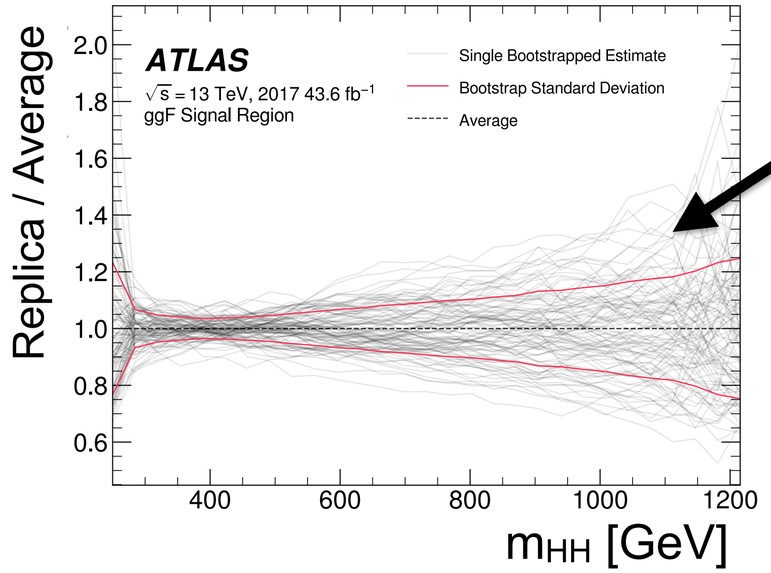
Take the ratio as an error bar



2. NN initialization

Retraining the NN 100x captures the uncertainty from multiple minima





Use the variation of trainings as a nuisance parameter

Flows for high dimensional interpolation

Hierarchical model: $p(x, m_{H1}, m_{H2}) = p(x|\, m_{H1}, m_{H2})\; p(m_{H1}, m_{H2})$

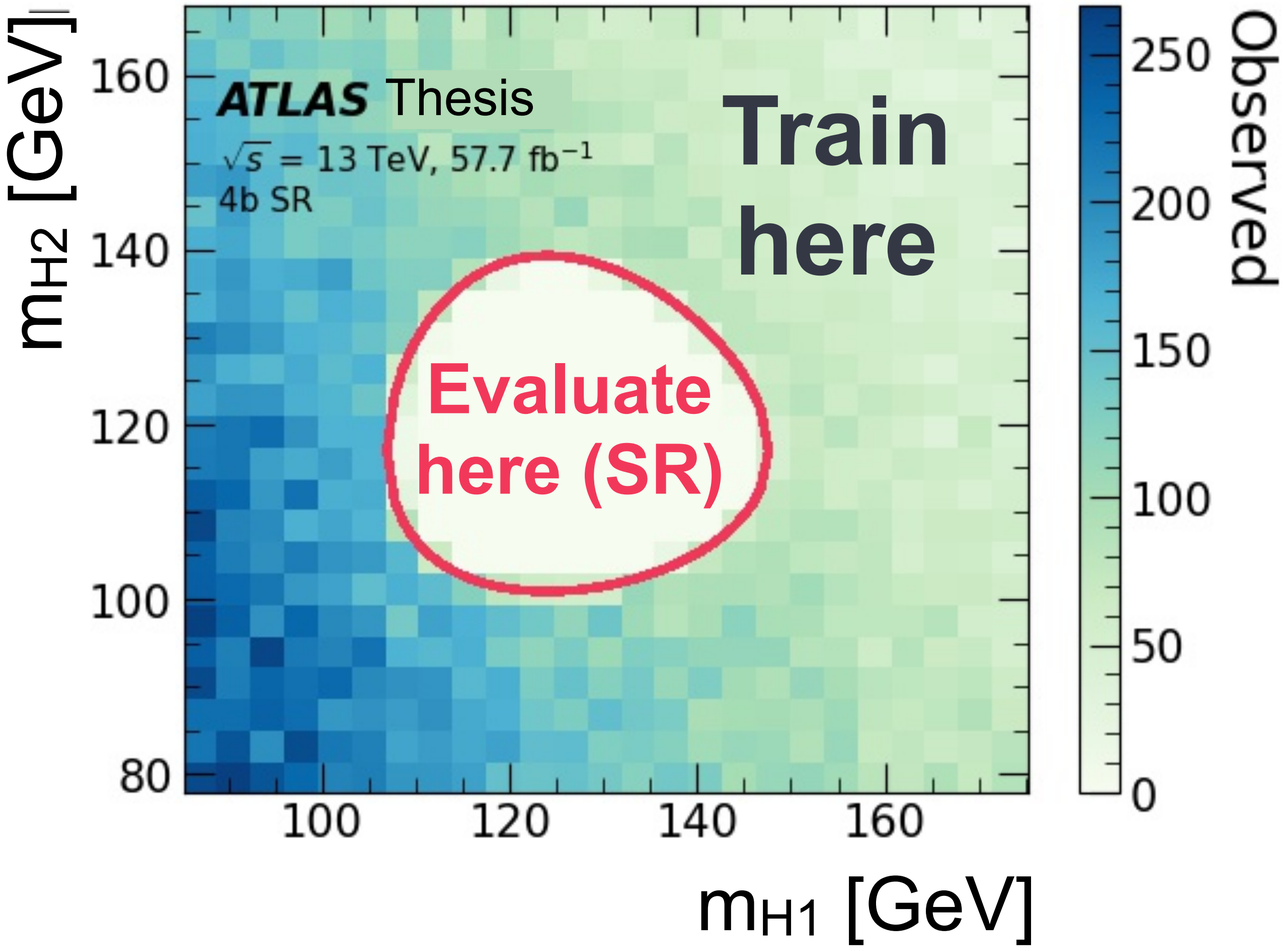
Normalizing flow — Gaussian process

x: Event kinematics

$p_{T,H1}$, $p_{T,H2}$, η_{H1}, η_{H2}, $\Delta\phi_{HH}$, X_{Wt} [top veto]

Use the smoothly varying (m_{H1}, m_{H2}) to predict SR kinematics.

Conceptually identical to CATHODE

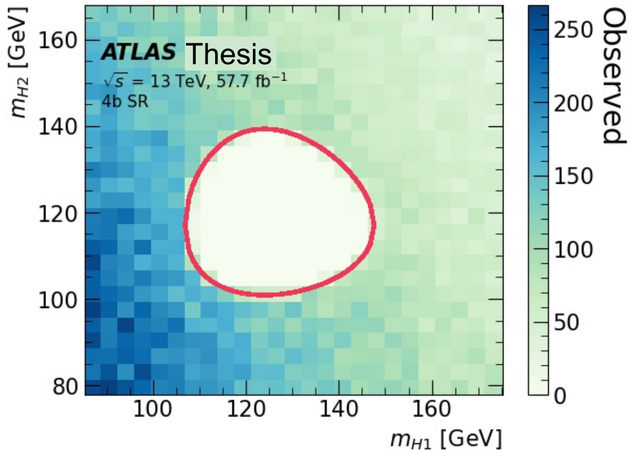


1) GP fits

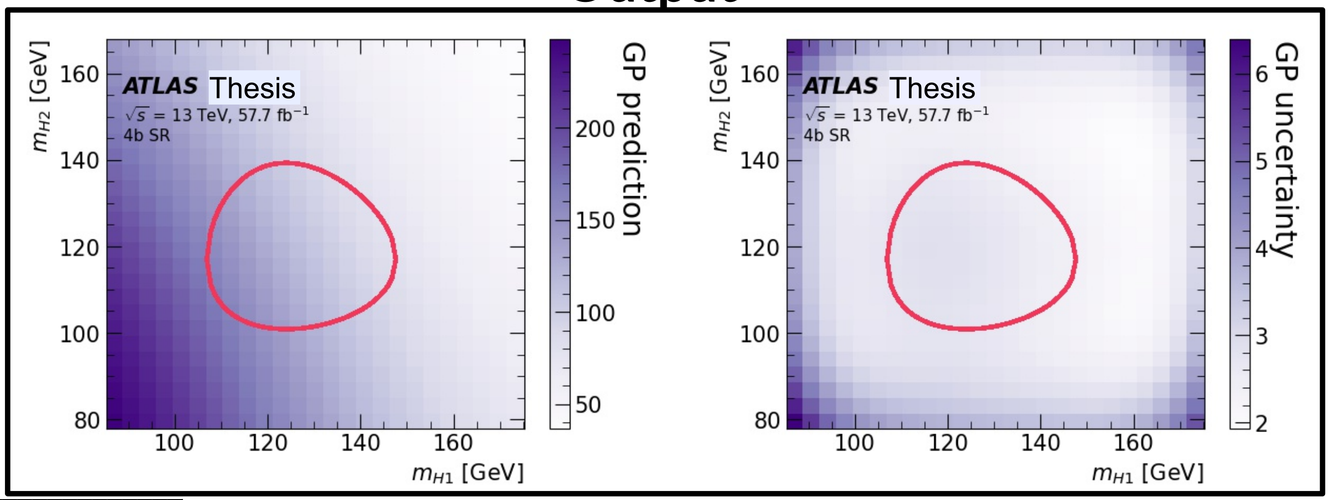
$$p(x, m_{H1}, m_{H2}) = p(x \mid m_{H1}, m_{H2})\ p(m_{H1}, m_{H2})$$

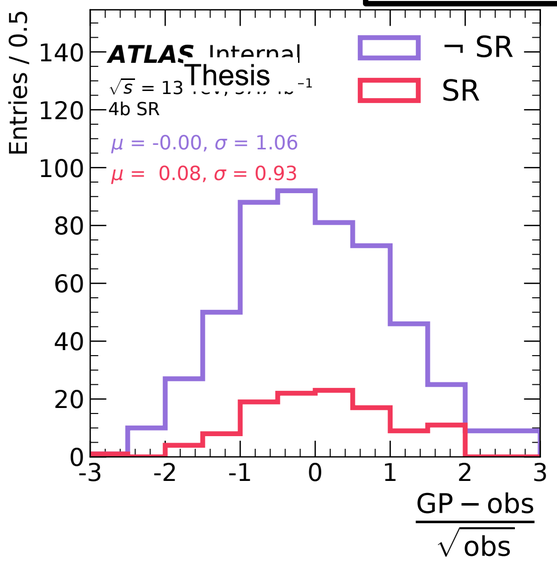
Fit a GP to the 2d (m_{H1}, m_{H2}) histogram
Radial basis function kernel, 2d length scale

Input



Output





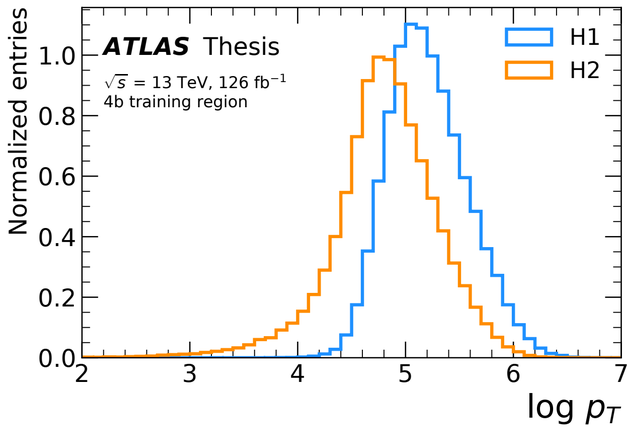
Set the normalization from massplane fit: ★

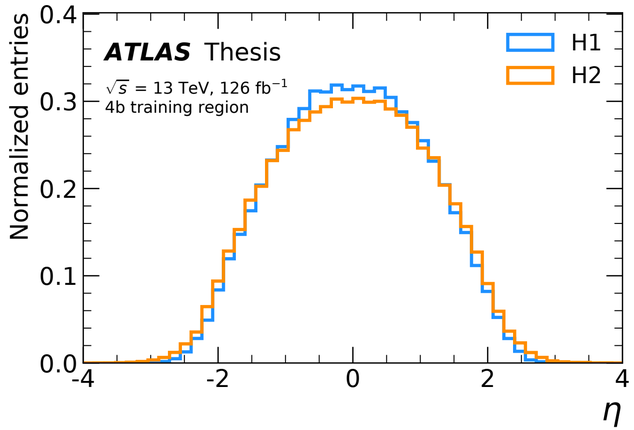
$$\text{SR yield} = \text{obs(not SR)}\ \frac{\text{pred(SR)}}{\text{pred(} \neg \text{ SR)}}$$

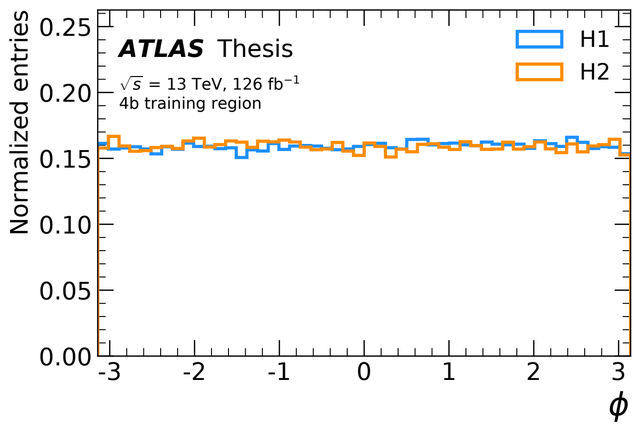
Input processing: HH -> 4b background modeling

Conditioning on m_{h1}, m_{h2} → ~~5~~ variables left to model

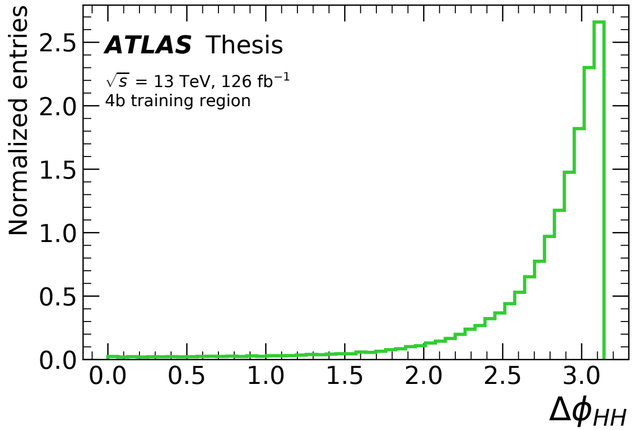
Concerned about modeling b/c no way to know that -π → π.

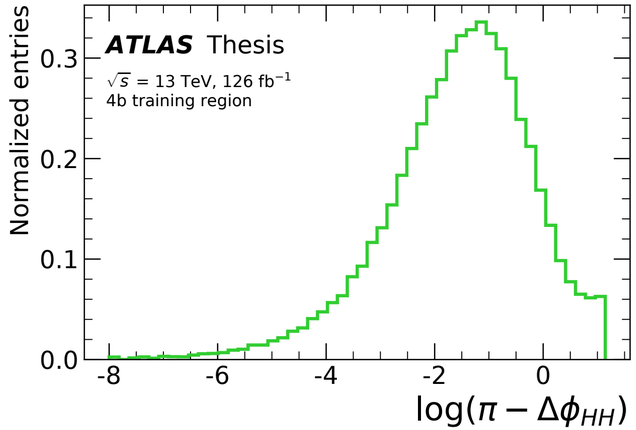






Constant $\Delta\Phi_{HH}$ - will give the same m_{HH}



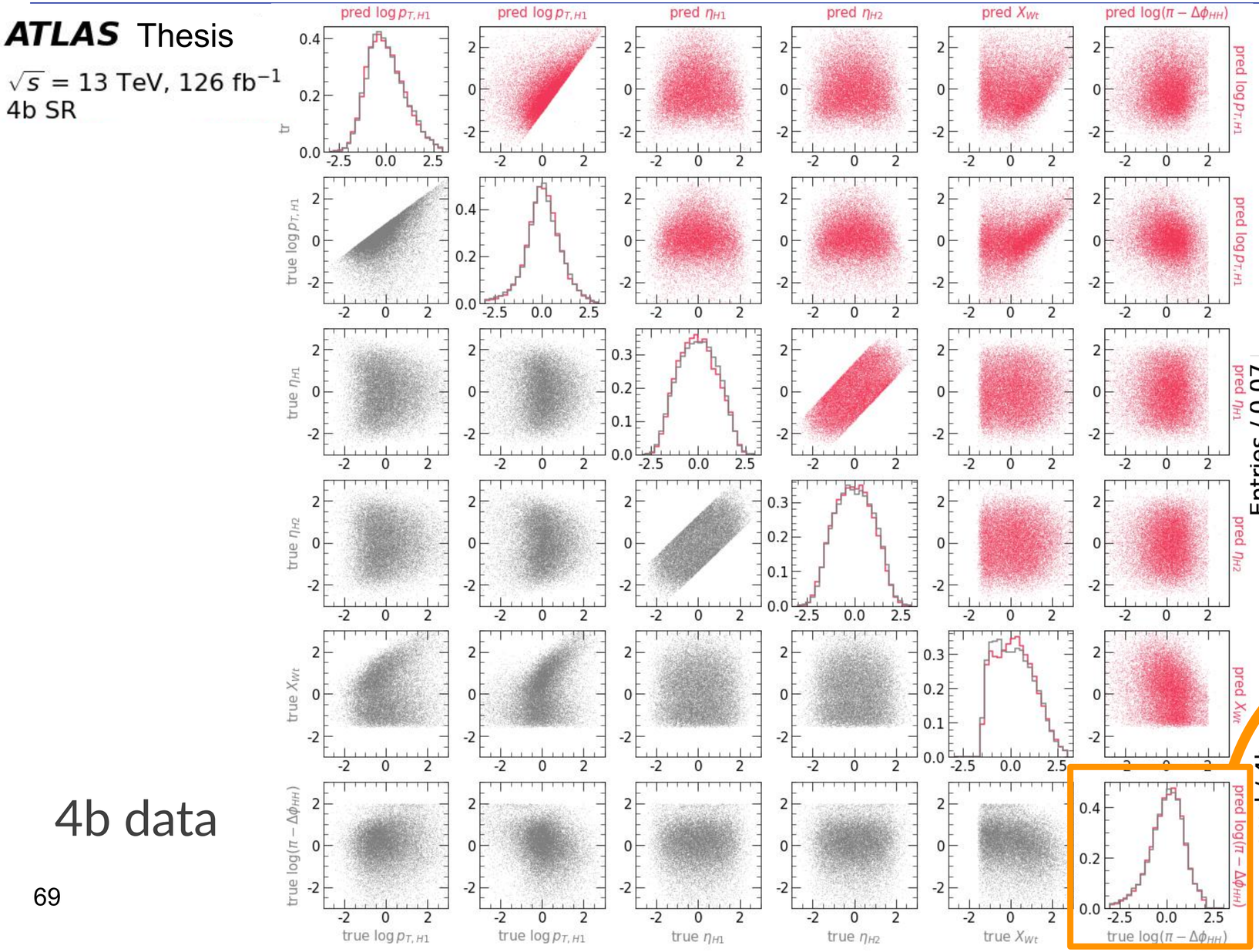


Azimuthal symmetry baked into the model

Symmetry

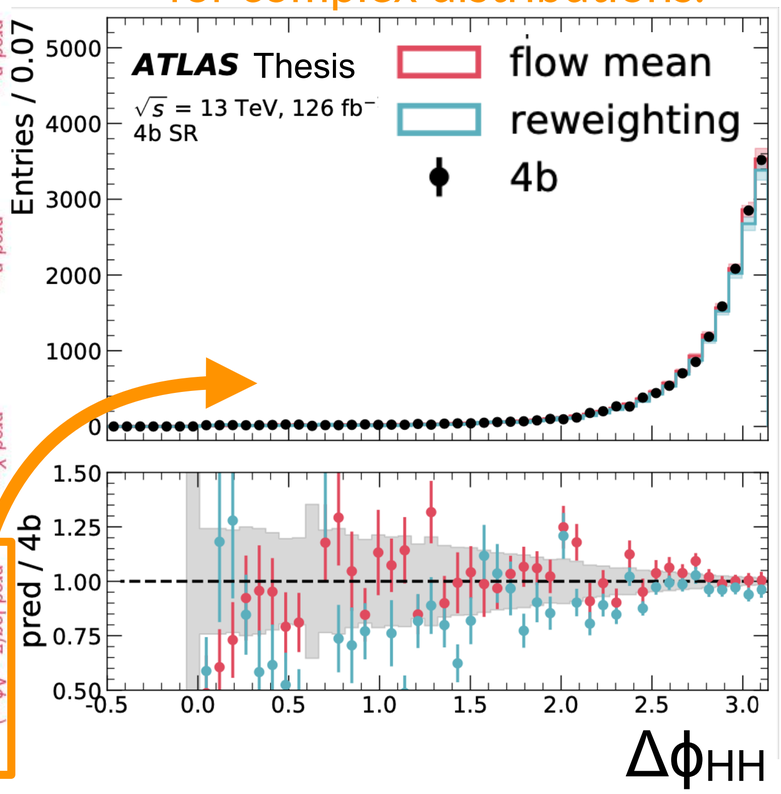
model with symmetry

Input processing : HH -> 4b background modeling



flow prediction

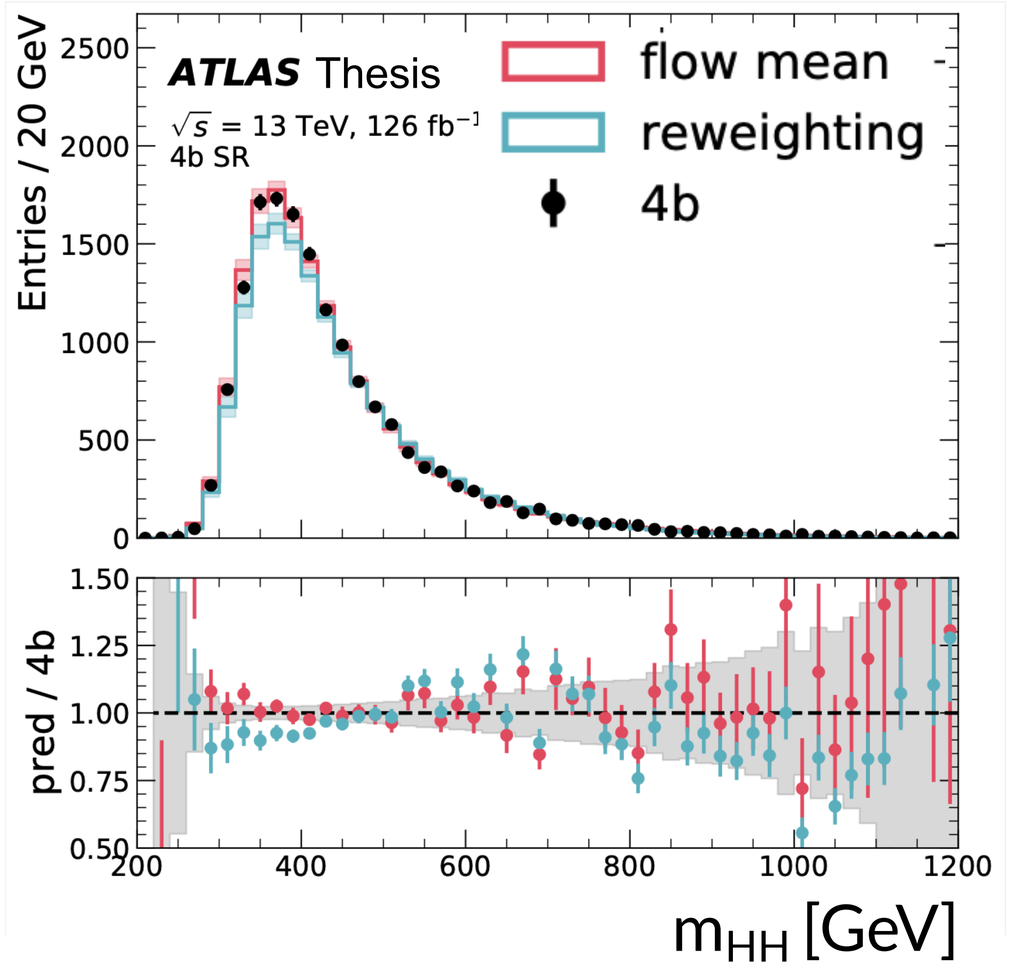
Good transformations help for complex distributions.

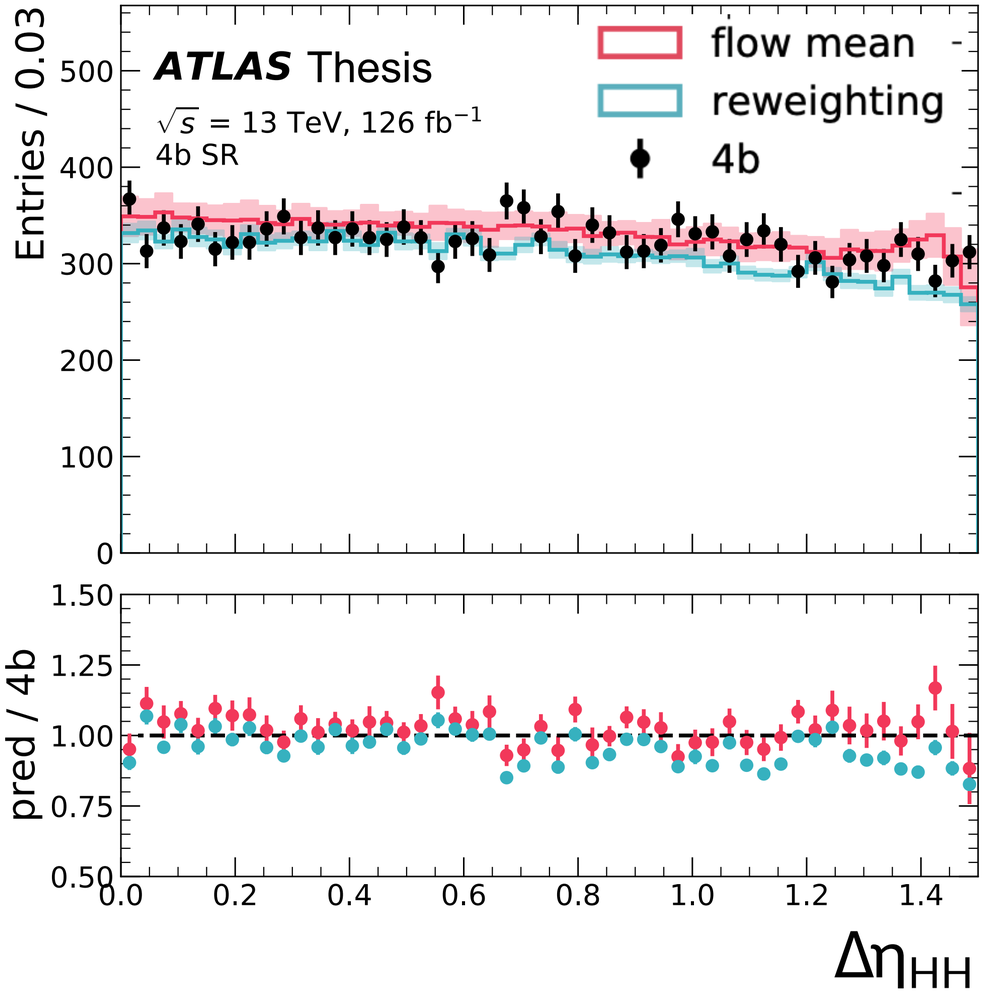


4b data

Input processing : HH → 4b background modeling

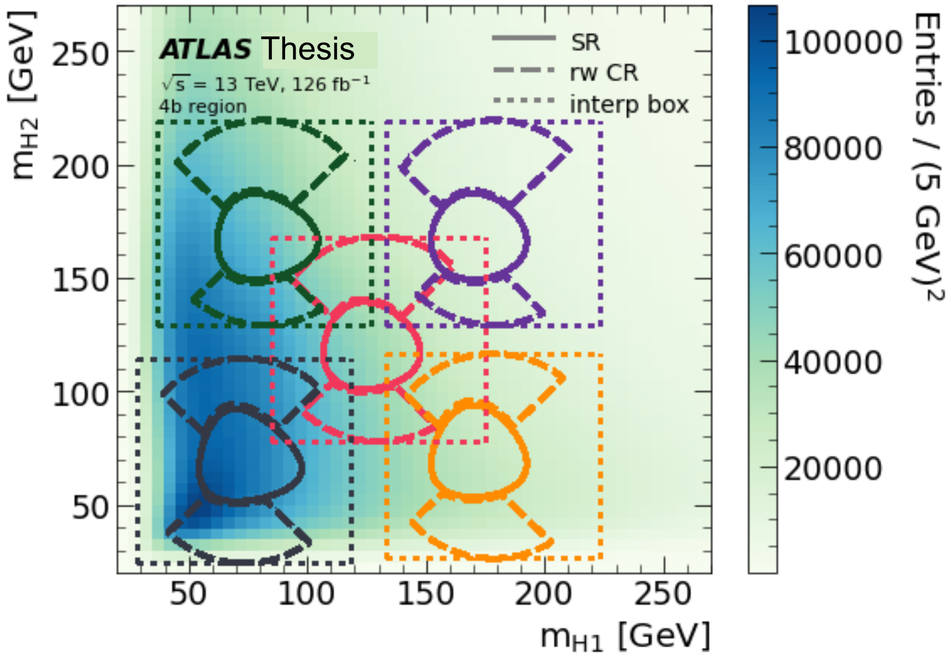
Variable transformations

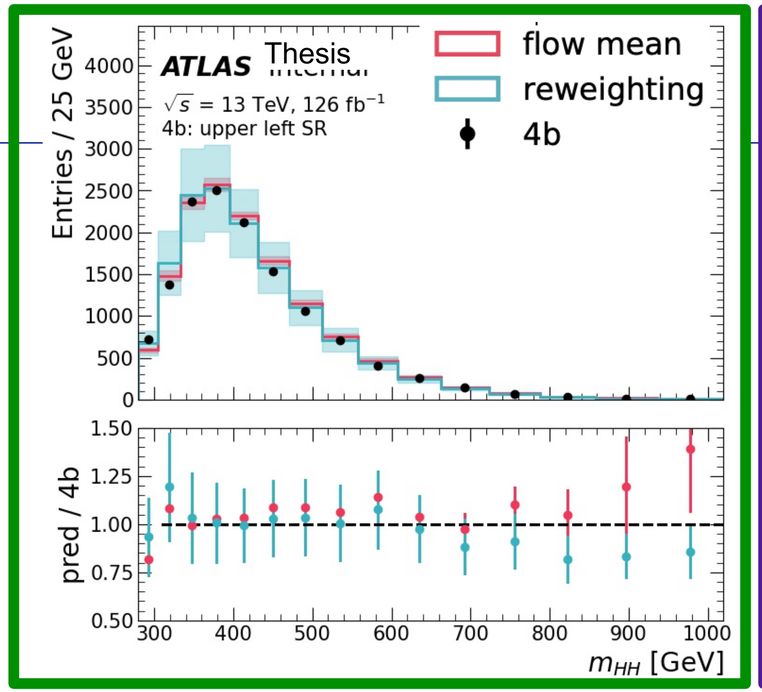


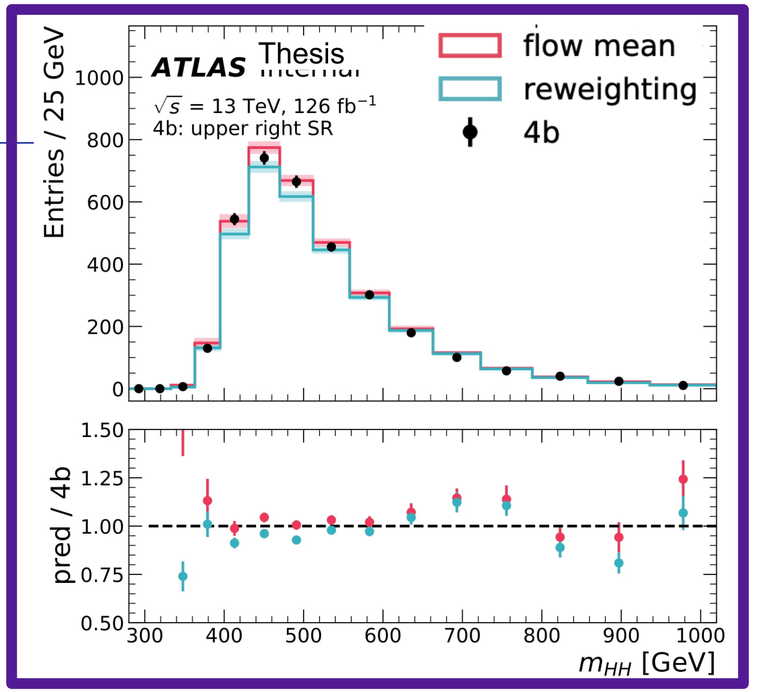


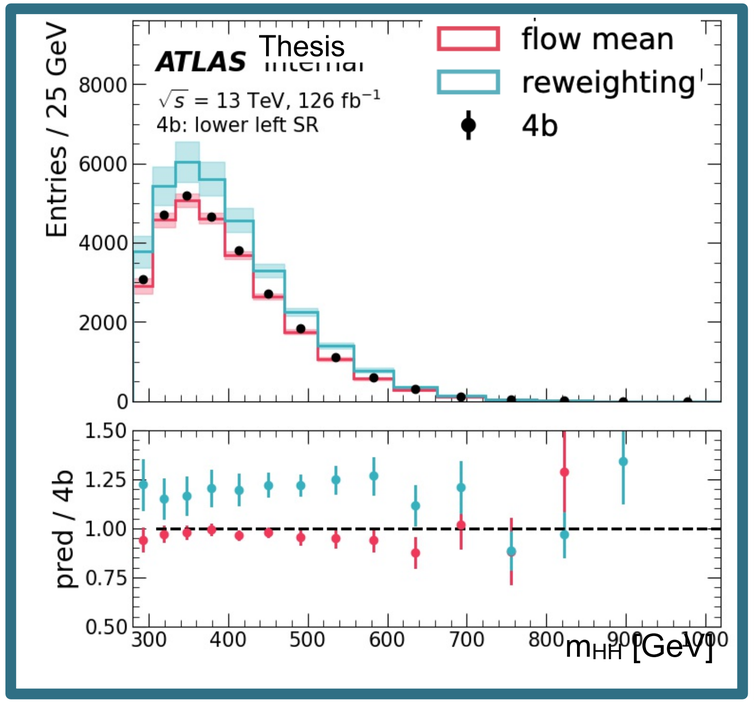
Promising… better closure.

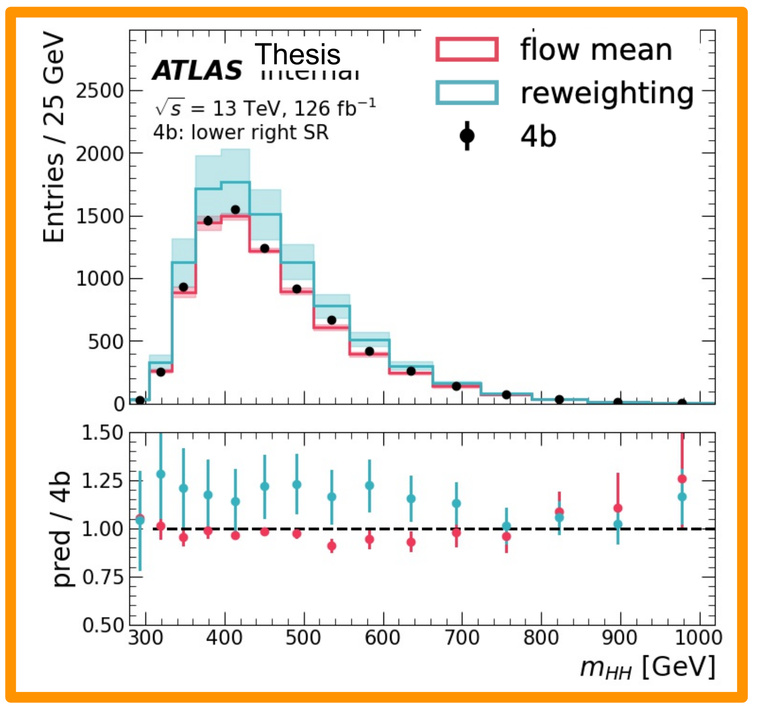
Shifted regions





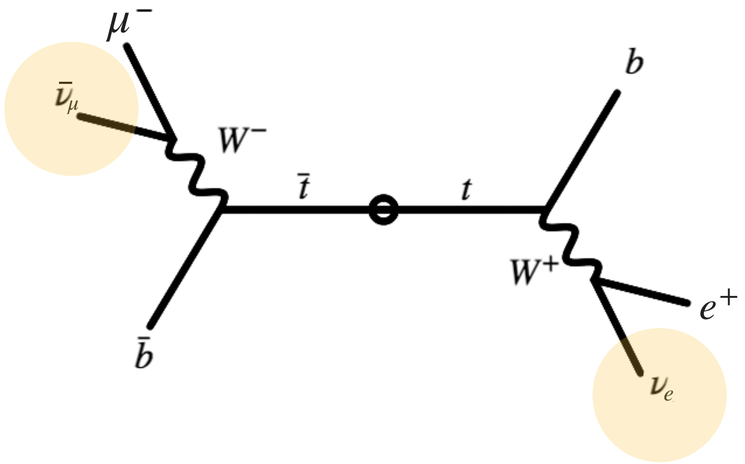






2207.00664
2307.02405

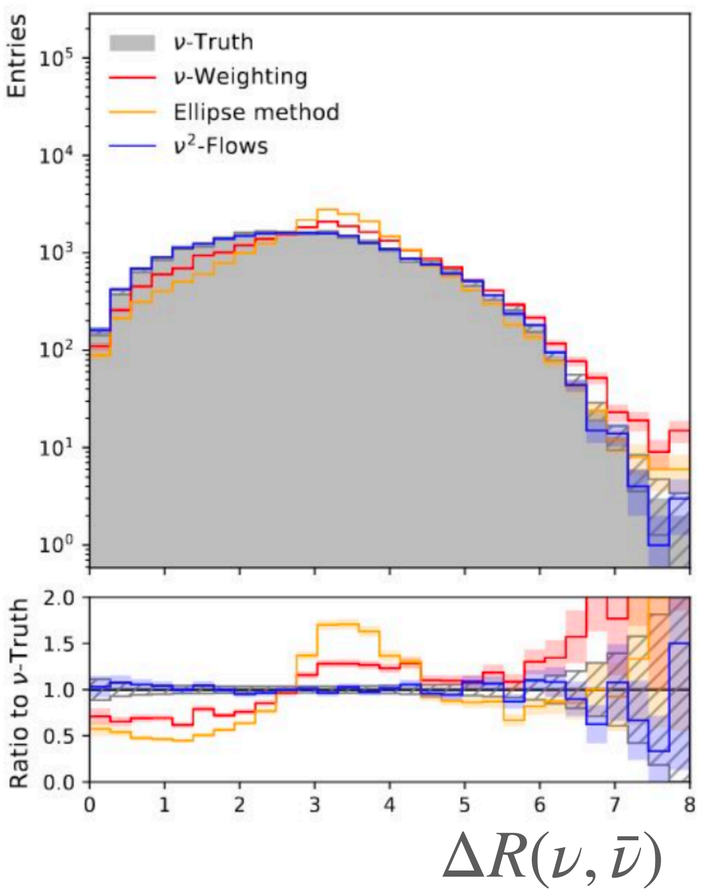
Probailistic reconstruction of two neutrinos

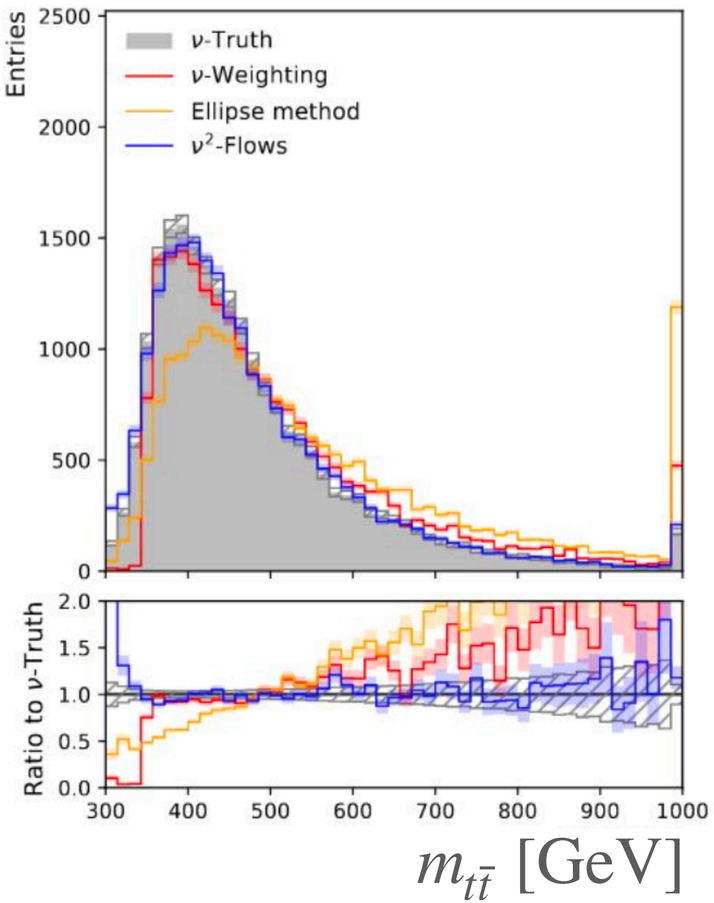


Useful for Belle 2? 1808.10567

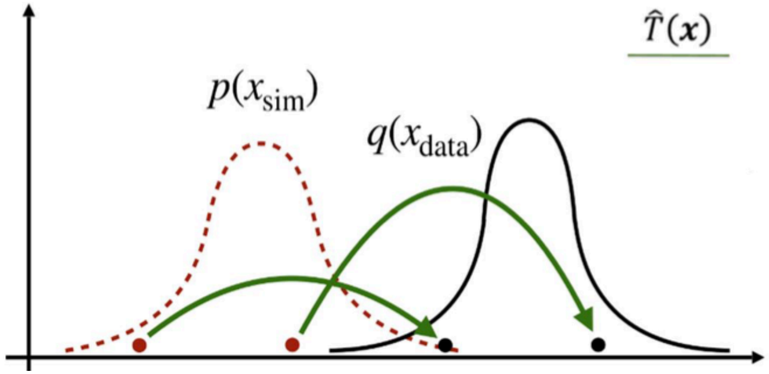
Table 5: Expected errors on several selected observables in radiative and electroweak penguin B decays. Note that 50 ab^{-1} projections for B_s decays are not provided as we do not expect to collect such a large $\Upsilon(5S)$ data set.

Observables	Belle (2017)	Belle II 5 ab^{-1}	Belle II 50 ab^{-1}
$\mathcal{B}(B \to K^{*+}\nu\bar{\nu})$	$< 40 \times 10^{-6}$	25%	9%
$\mathcal{B}(B \to K^{+}\nu\bar{\nu})$	$< 19 \times 10^{-6}$	30%	11%



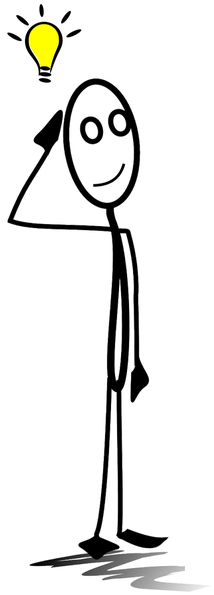


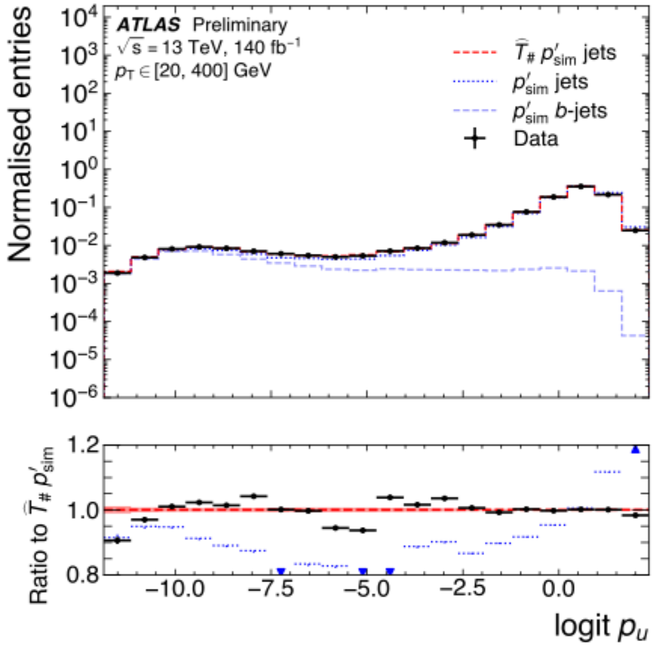
Calibration… in a differentiable way

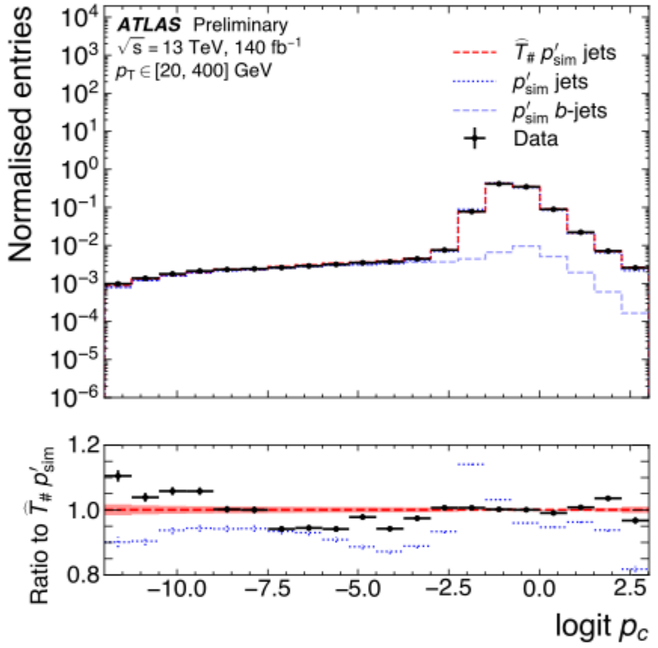


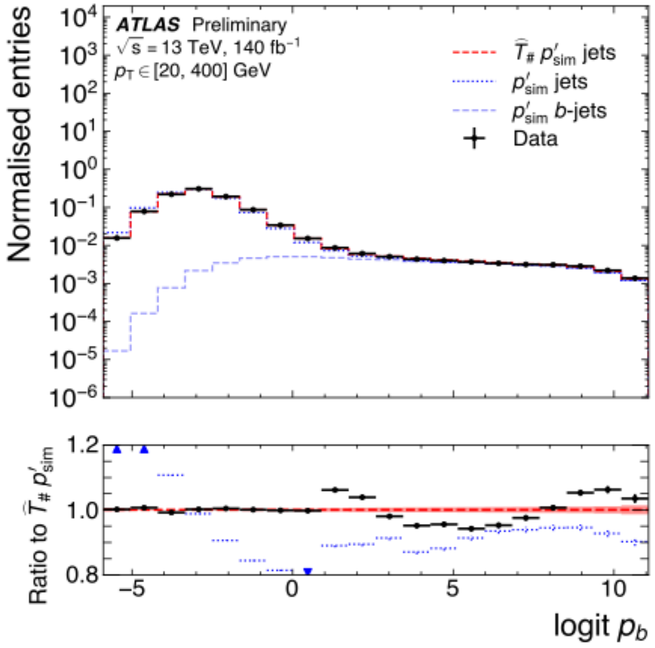
Idea: How do you have a mapping from $p_b^{MC} \rightarrow p_b^{data}$

Architecture: Normalizing flow with a constraint to ensure the transport map is minimal.









ML

Foundation Model

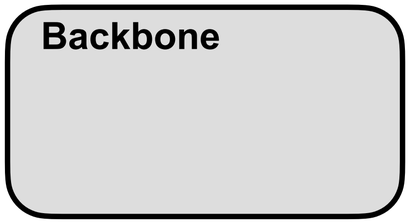
HEP

Reconstruction

ML

Foundation Model

Pretext tasks
(next word prediction)

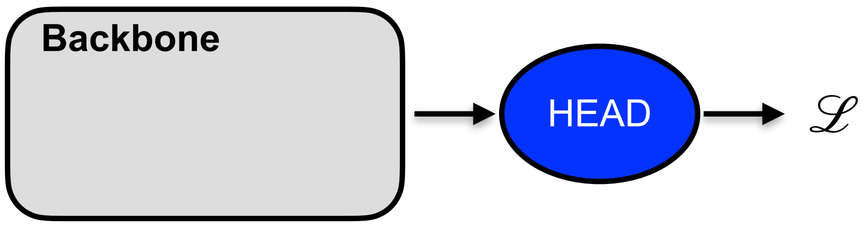


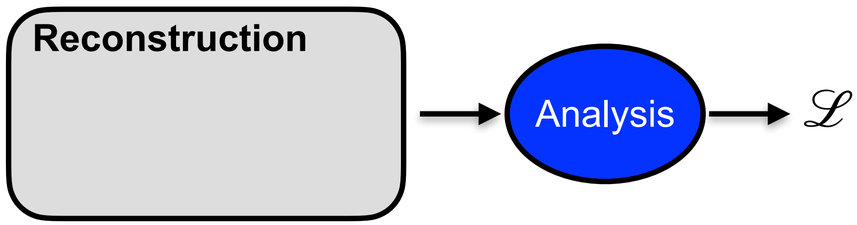
HEP

Reconstruction

Reconstruction closure

ML	**HEP**
Foundation Model	Reconstruction
Pretext tasks (next word prediction)	Reconstruction closure
Downstream head	Analysis





ML

Foundation Model

Pretext tasks
(next word prediction)

Downstream head

Finetuning

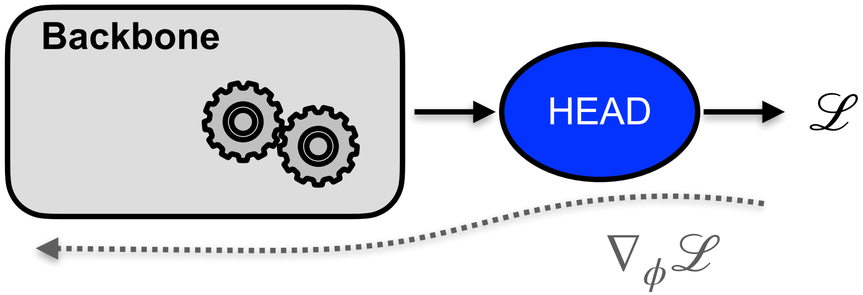
HEP

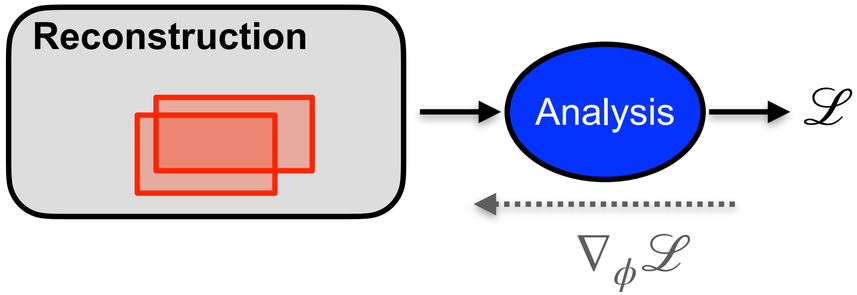
Reconstruction

Reconstruction closure

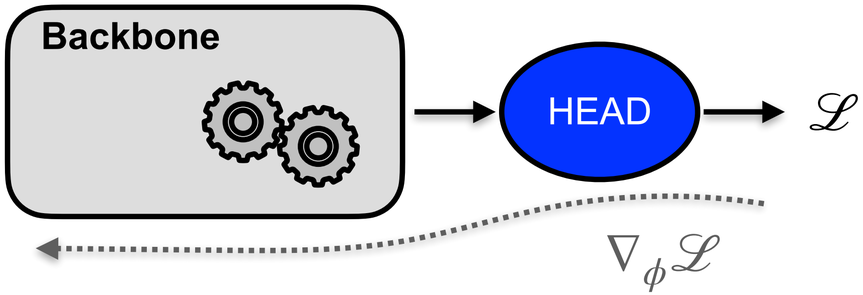
Analysis

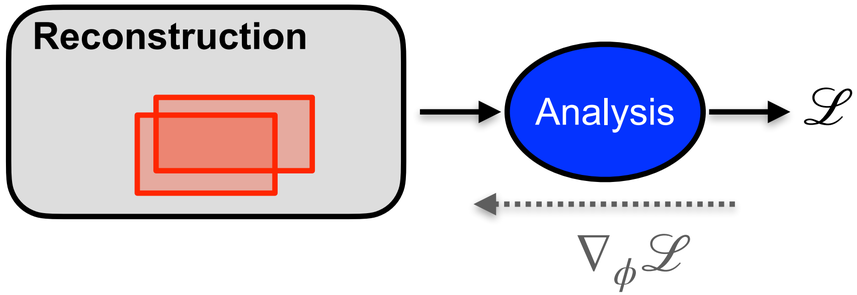
Analysis specific reconstruction, e.g, operating points

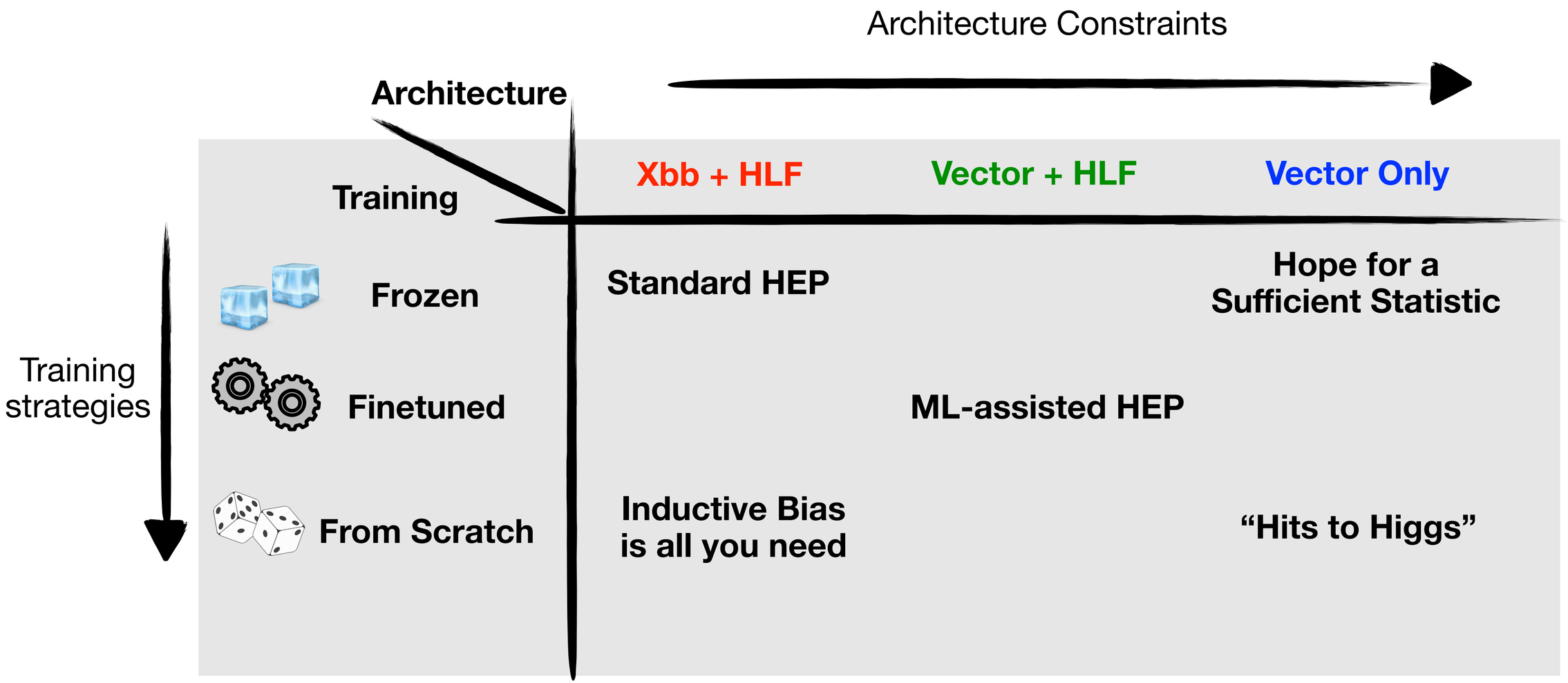


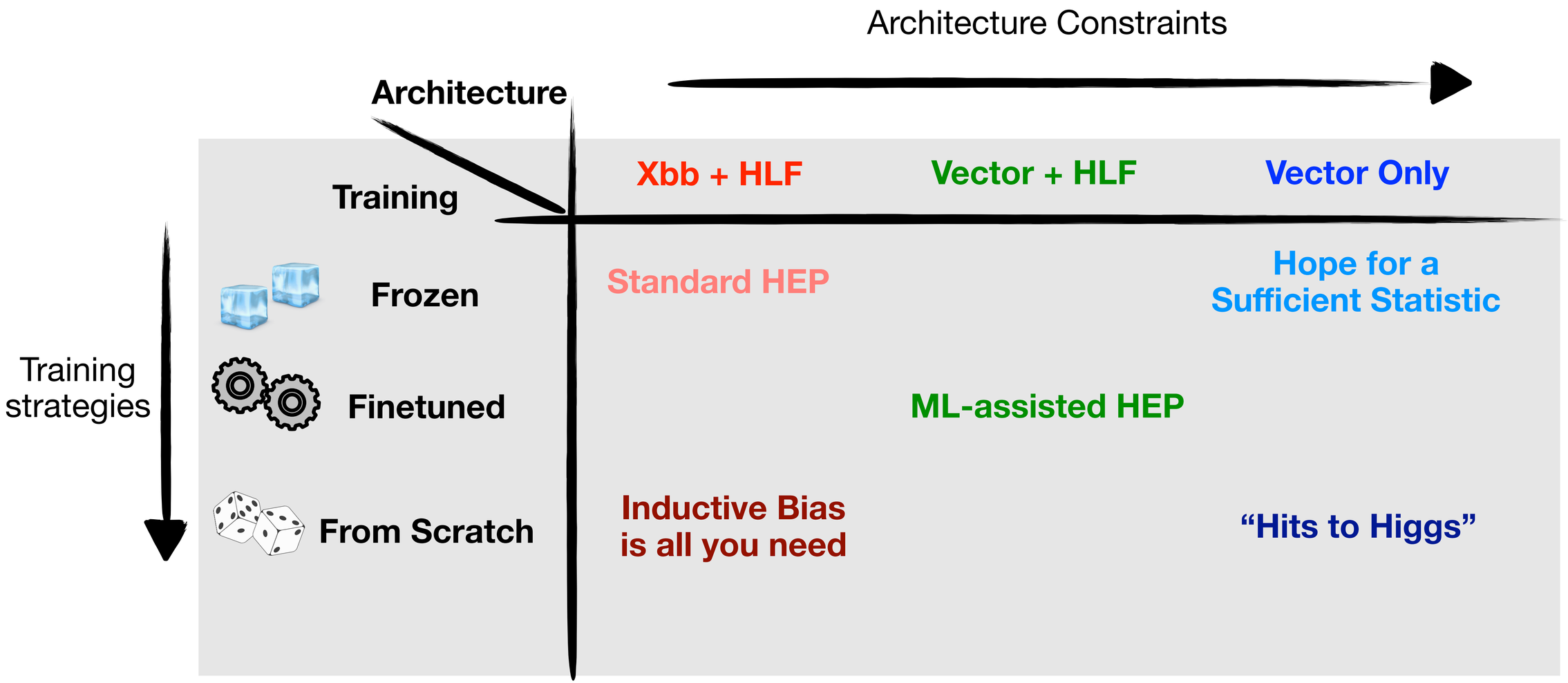


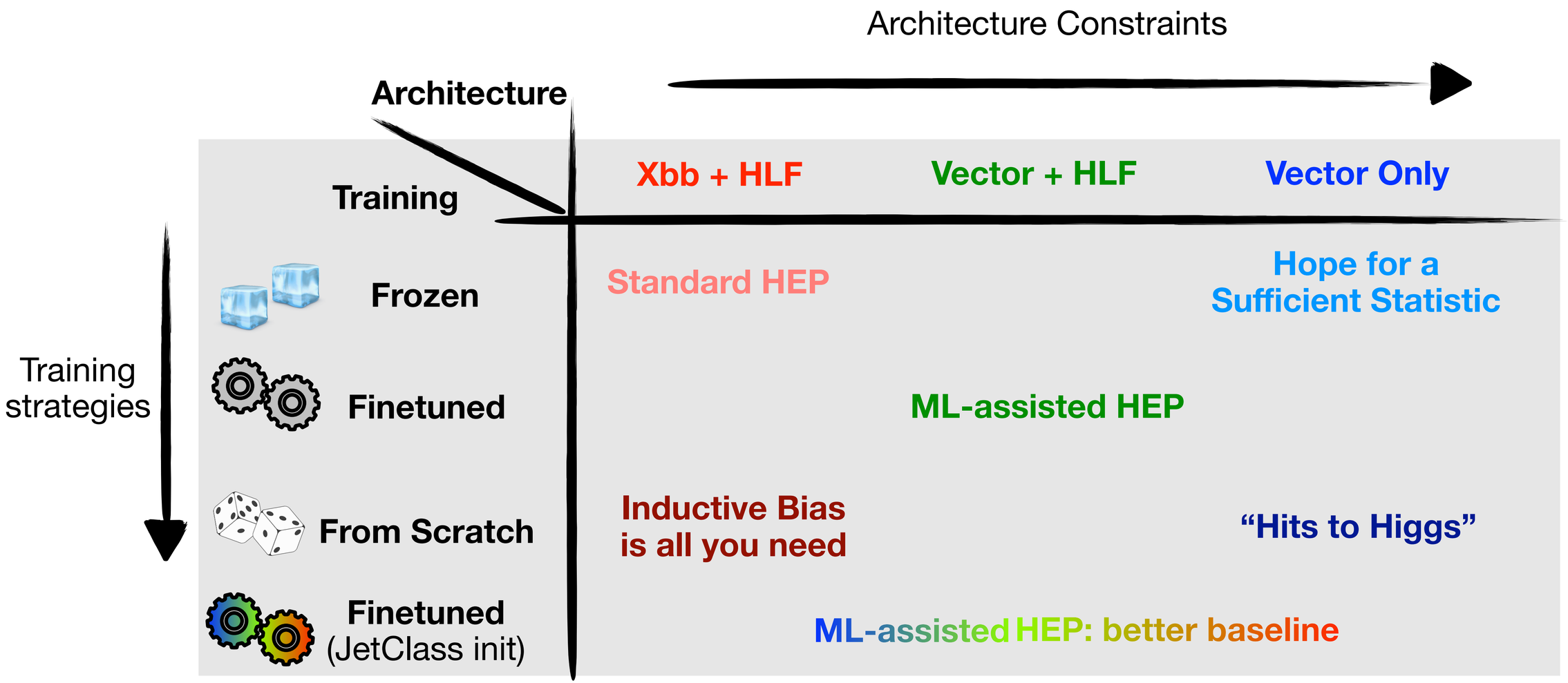
ML	HEP
Foundation Model	Reconstruction
Pretext tasks (next word prediction)	Reconstruction closure
Downstream head	Analysis
Finetuning	Analysis specific reconstruction, e.g, operating points
Embedding	Object observables (jets)






Architecture Constraints
Architecture
Training
Xbb + HLF
Vector + HLF
Vector Only
Training strategies
Frozen
Standard HEP
Hope for a Sufficient Statistic
Finetuned
ML-assisted HEP
From Scratch
Inductive Bias is all you need
“Hits to Higgs”


Architecture Constraints
Architecture
Training
Xbb + HLF
Vector + HLF
Vector Only
Training strategies
Frozen
Standard HEP
Hope for a Sufficient Statistic
Finetuned
ML-assisted HEP
From Scratch
Inductive Bias is all you need
“Hits to Higgs”

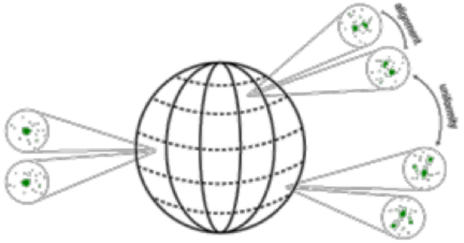

Architecture Constraints
Architecture
Training
Xbb + HLF
Vector + HLF
Vector Only
Training strategies
Frozen
Standard HEP
Hope for a Sufficient Statistic
Finetuned
ML-assisted HEP
From Scratch
Inductive Bias is all you need
“Hits to Higgs”
Finetuned (JetClass init)
ML-assisted HEP: better baseline

M. Kagan's slide

& A. Hallin for up-to-date list

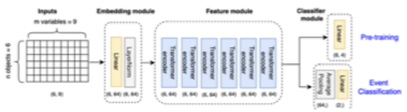
How to build this HEP foundation model

Contrastive Learning: Symmetry Augmentation



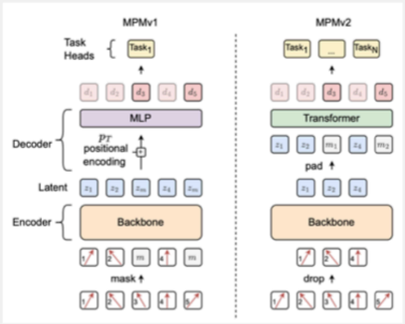
Dillon, Kasieczka, Olischlager Plehn, Sorrenson, Vogel, 2108.04253

Masked Particle Type Prediction



Kishimoto, Morinaga, Saito Tanaka, 2312.06909

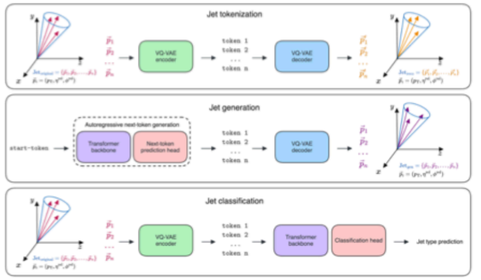
Masked Particle Modeling



Golling, Heinrich, **MK**, Klein, Leigh, Osadchy, Raine, 2401.13537

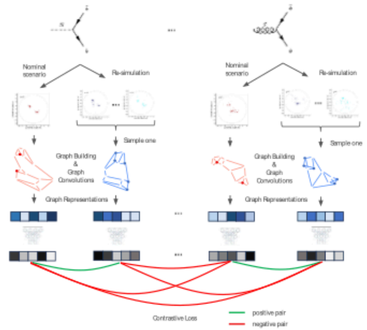
Leigh, Klein, Charton, Golling, Heinrich, **MK**, Ochoa, Osadchy, 2409.12589

Next Token Predictoin



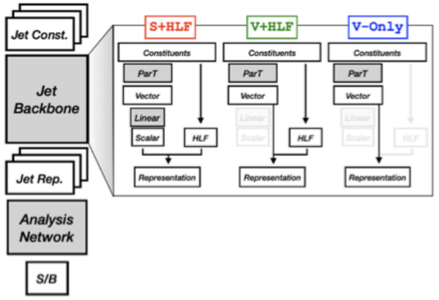
Birk, Hallin, Kasieczka, 2403.05618

Contrastive Learning: Re-Simulation



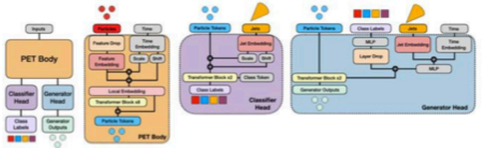
Harris, **MK**, Krupa, Maier, Woodward, 2403.07066

Supervised Pre-training and Joint Optimization



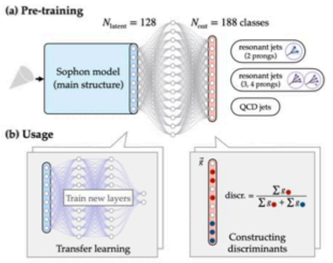
Vigl, Hartman, Heinrich, 2401.13536

Supervised Classification and Generation



Mikuni, Nachman 2404.16091

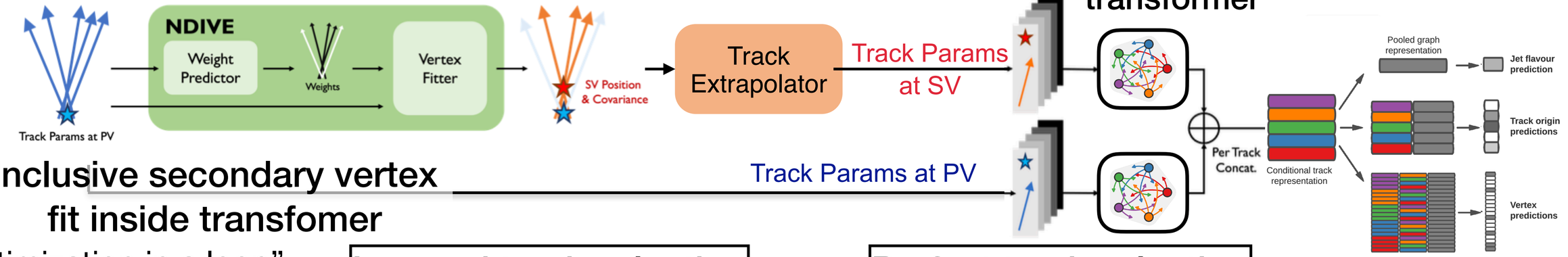
Large-Scale Fine-Grained Classification



...

Li, Li, et al. 2405.12972

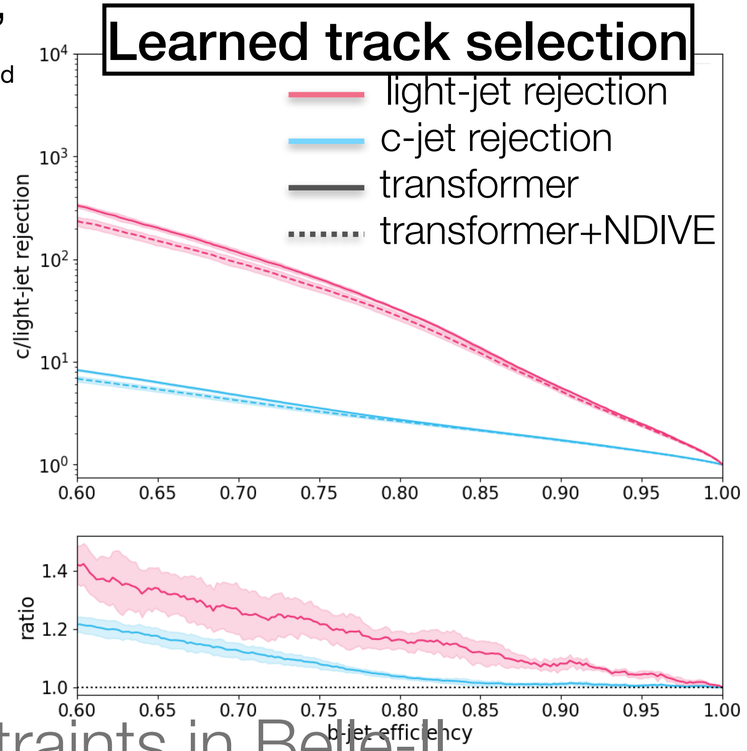
Neural Differentiable Vertex Fitter

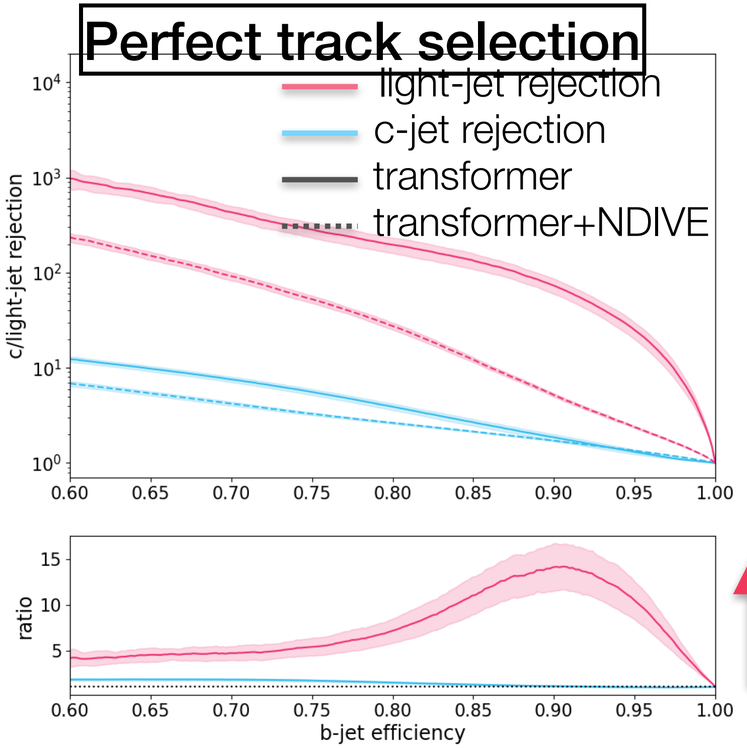


Inclusive secondary vertex fit inside transfomer

"Optimization in a loop"

Gradients of fitted vertex to optimize end-to- end trained b-tagger

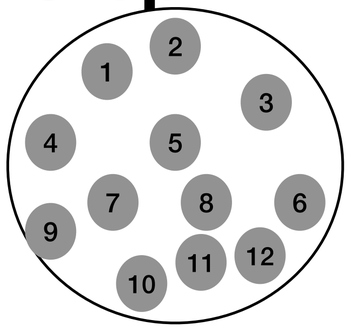




10x imp 😱

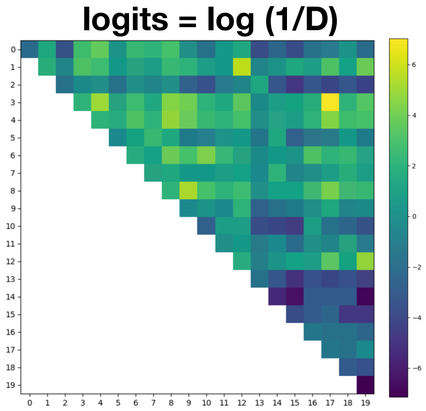
+ kinematic constraints in Belle-II

Step 1



Distance matrix D

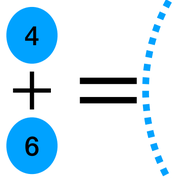
Depends on R

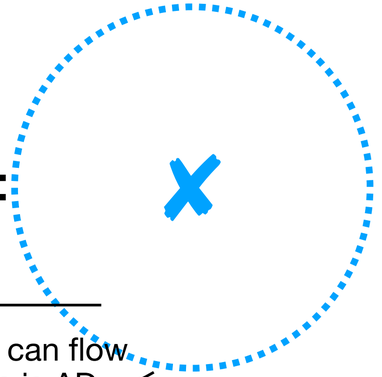


Sample $\sim$ (4,6)

Gradient can flow b/c this is GS

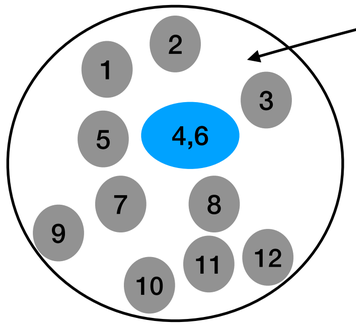
Merge





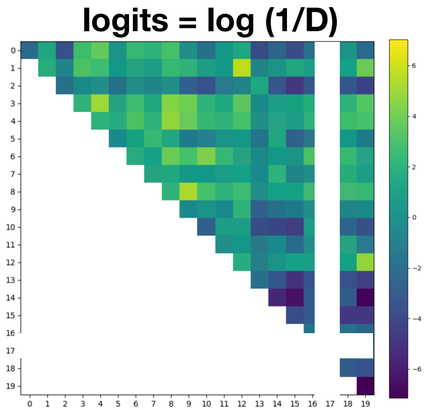
Gradient can flow b/c this is AD

Step 2



Distance matrix D

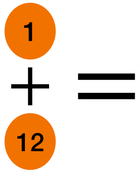
Depends on R

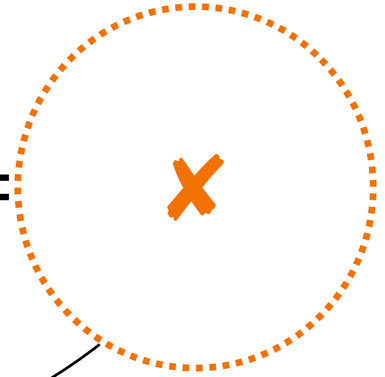


Sample $\sim$ (1,12)

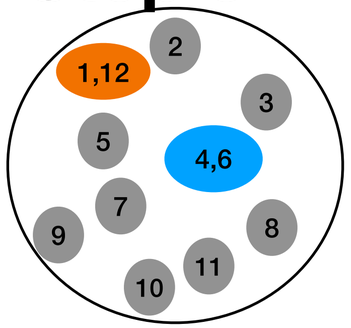
Gradient can flow b/c this is GS

Merge





Step 3



● ● ●

Differentiating clustering

Step 1: Interpret clustering decision probabilistically

Step 2: Gradient with score based estimate

As demonstrated by M. Kagan and L. Heinrich for particle interactions in material: 2308.16680

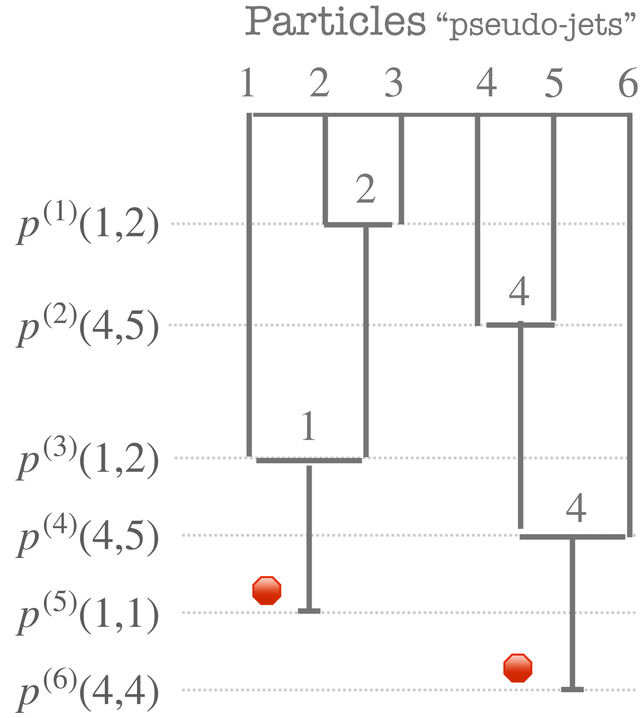
In general:

$$\nabla_\theta \mathbb{E}_{x\sim p(\theta)}[f(x)] = \mathbb{E}_{x\sim p(\theta)}[f(x)\,\nabla_\theta \log p(\theta)]$$

For jet clustering:

$$\nabla_R \mathbb{E}_{p(R)}[m_H] = \mathbb{E}_{p(R)}[\underbrace{m_H \nabla_R \log p(R)}]$$

Ex: sample i through event w/ 6 p'cles



$$m_H^{(i)} \left(\begin{array}{c} \nabla p^{(1)}(1,2) + \nabla p^{(2)}(4,5) + \nabla p^{(3)}(1,2) \\ + \nabla p^{(4)}(4,5) + \nabla p^{(5)}(1,1) + \nabla p^{(6)}(4,4) \end{array} \right)$$

With A. Kofler, M. Kagan and L. Heinrich, in prep