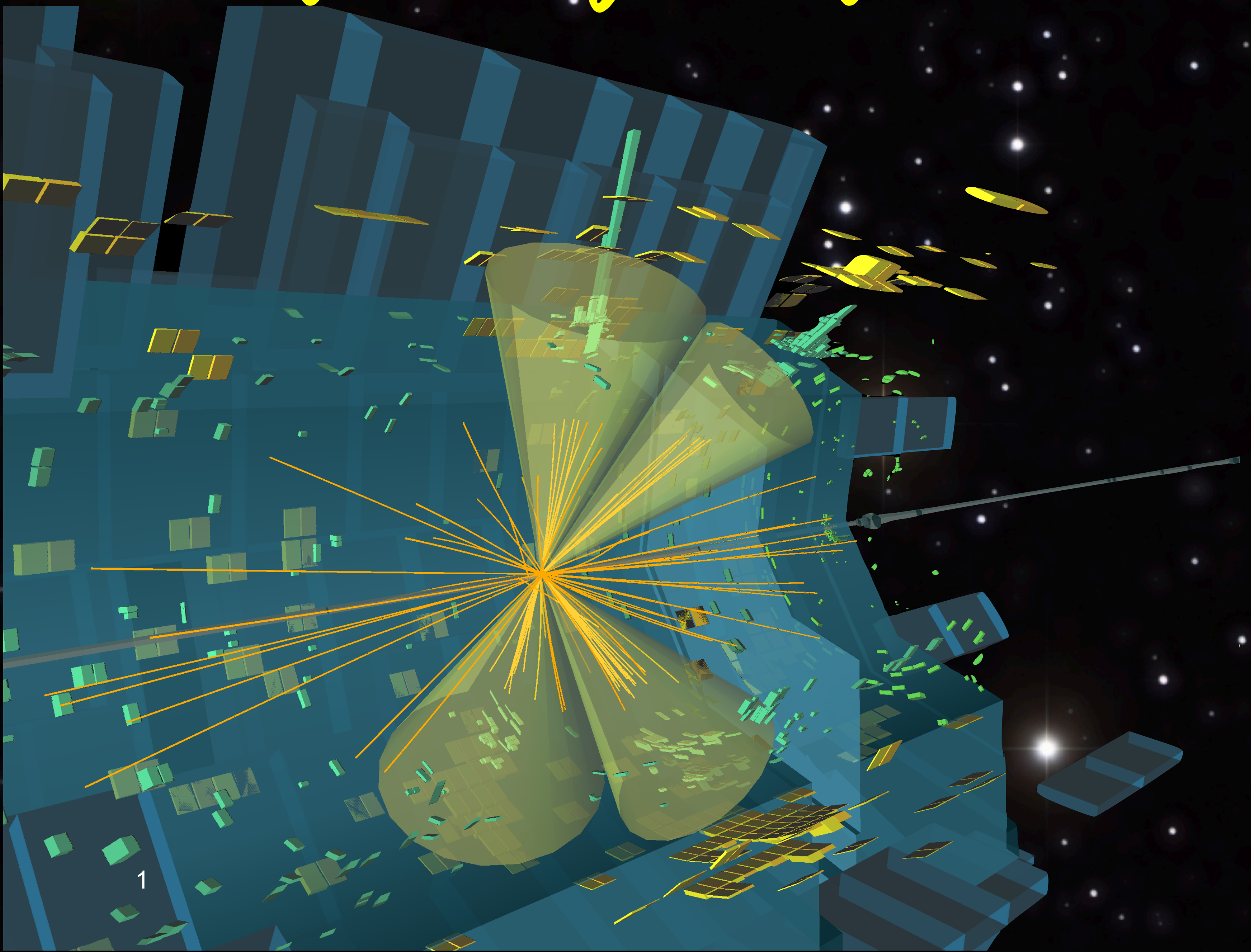
Transforming Jet flavour tagging on ATLAS

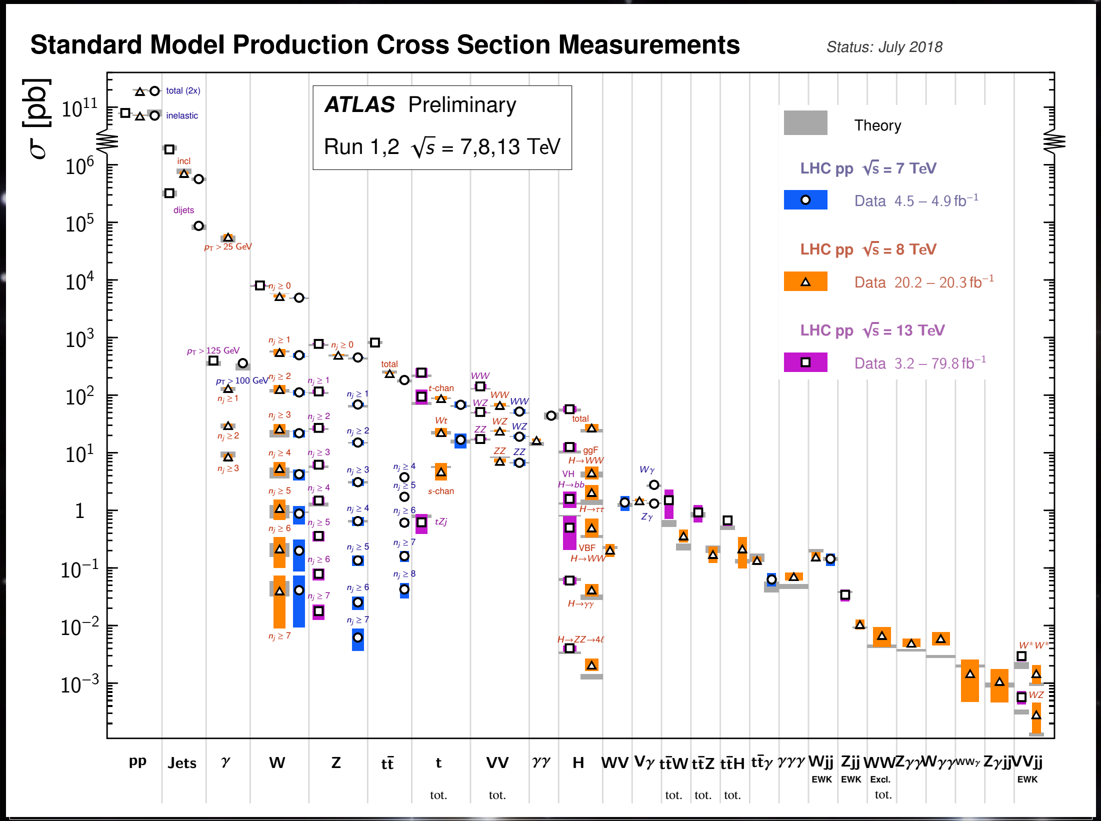


Nicole Hartman
nicole.hartman@tum.de

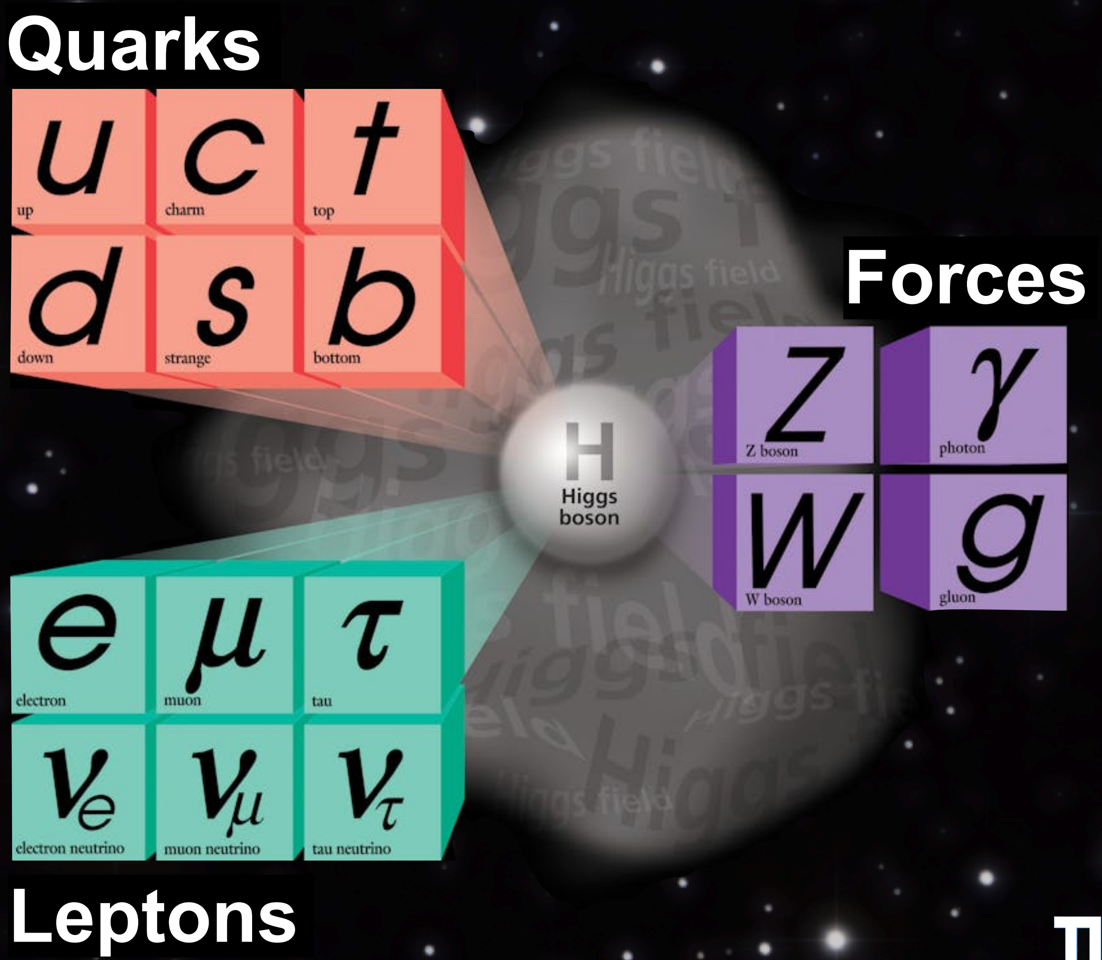
Siegen CPPS seminar
26th Nov 2024

ATLAS EXPERIMENT
Run: 362619
Event: 524614423
2018-10-03 08:06:34 CEST

$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\gamma_\mu D^\mu\psi + |D_\mu\phi|^2 - V(\phi) + (y_{ij}\bar{\psi}_i\psi_j + \text{h.c.})$$



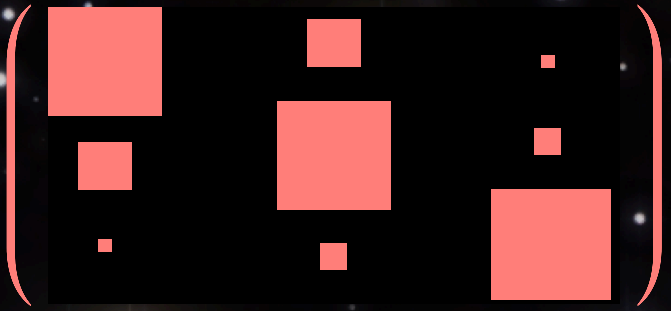
ATLAS SM summary plots

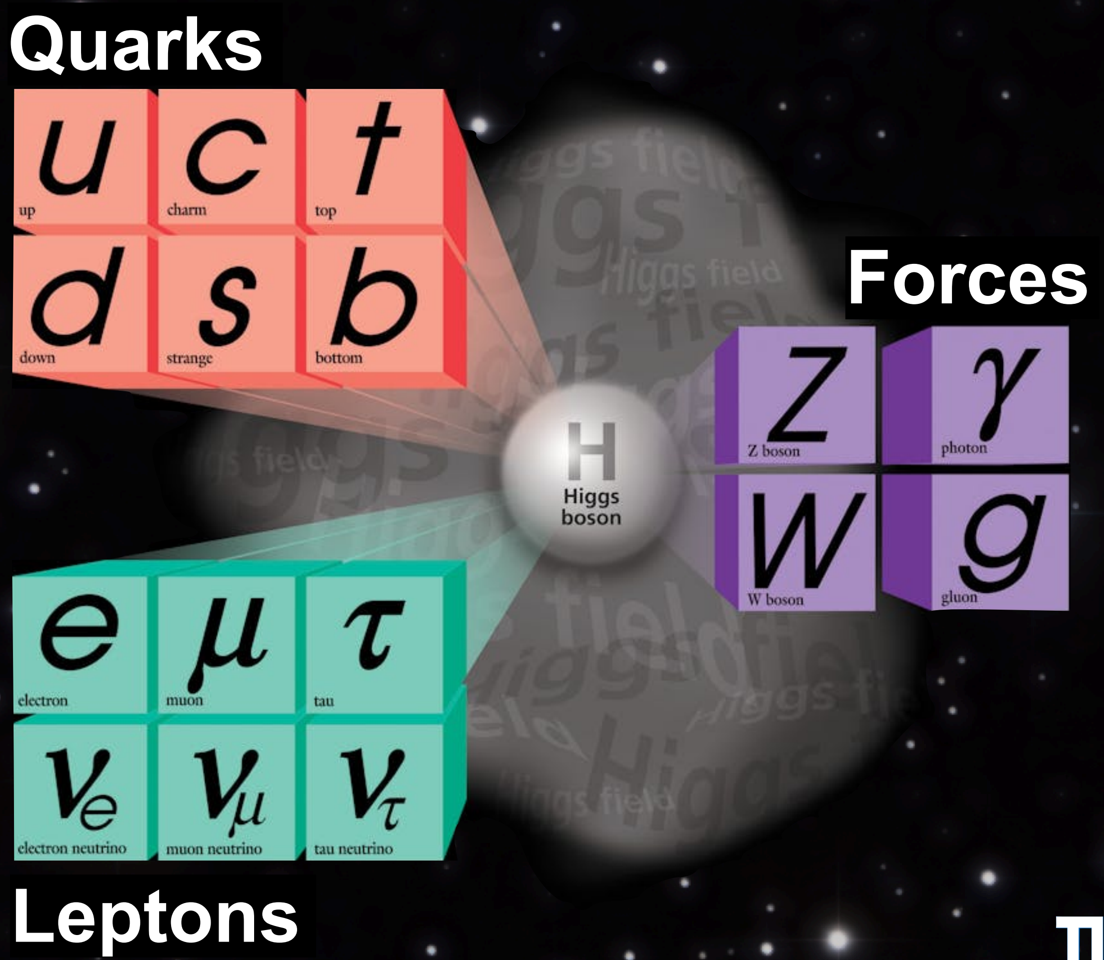


$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\gamma_\mu D^\mu \psi + |D_\mu \phi|^2 - V(\phi) + (y_{ij}\bar{\psi}_i \psi_j + \text{h.c.})$$

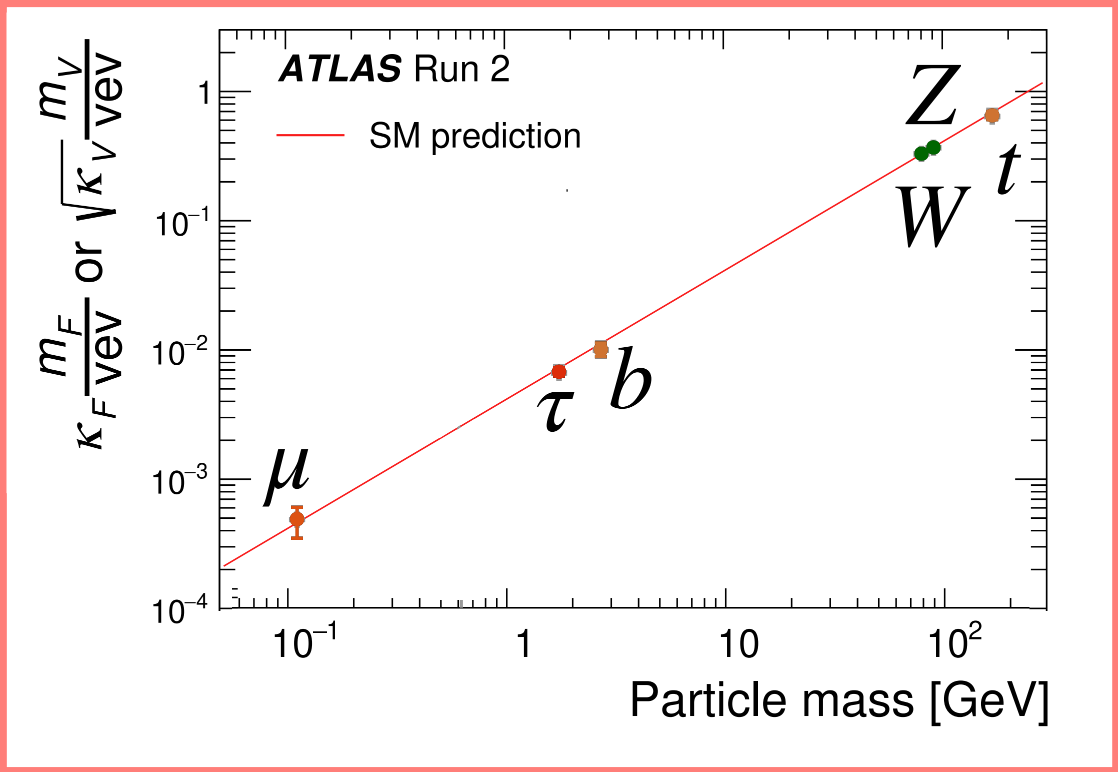
$$\begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix}$$

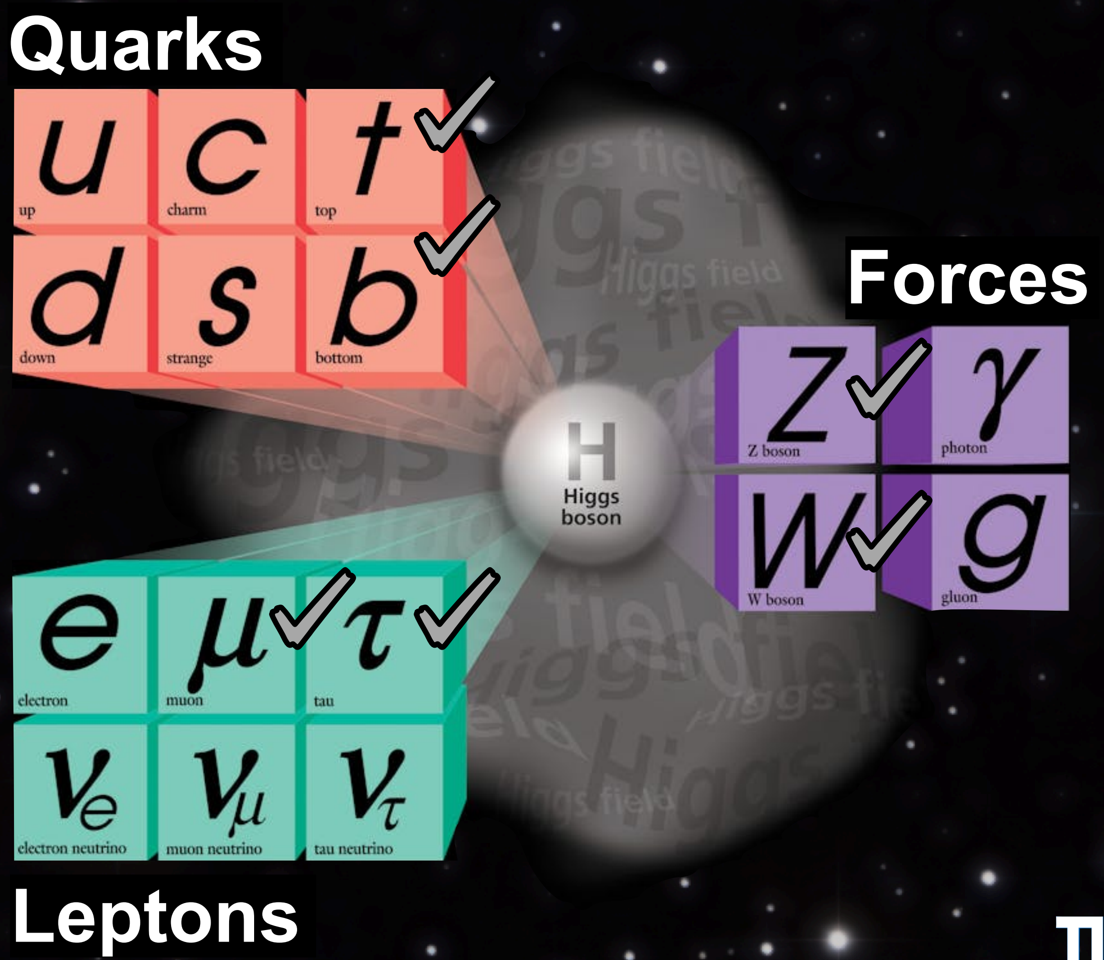
$=$



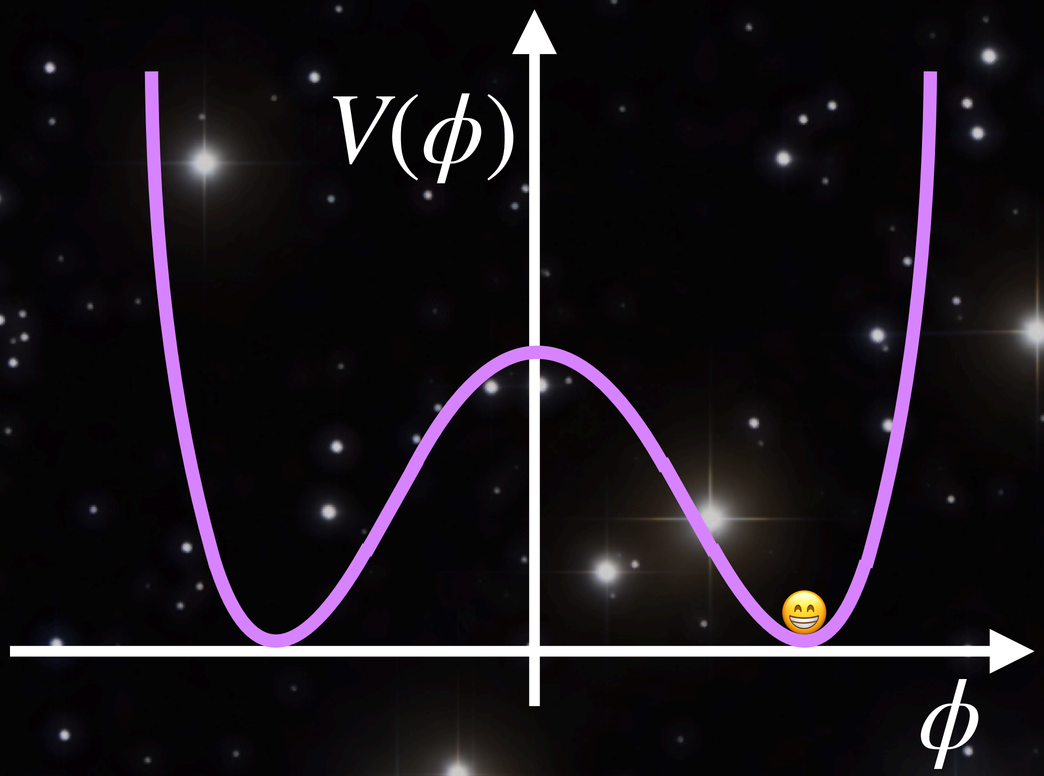


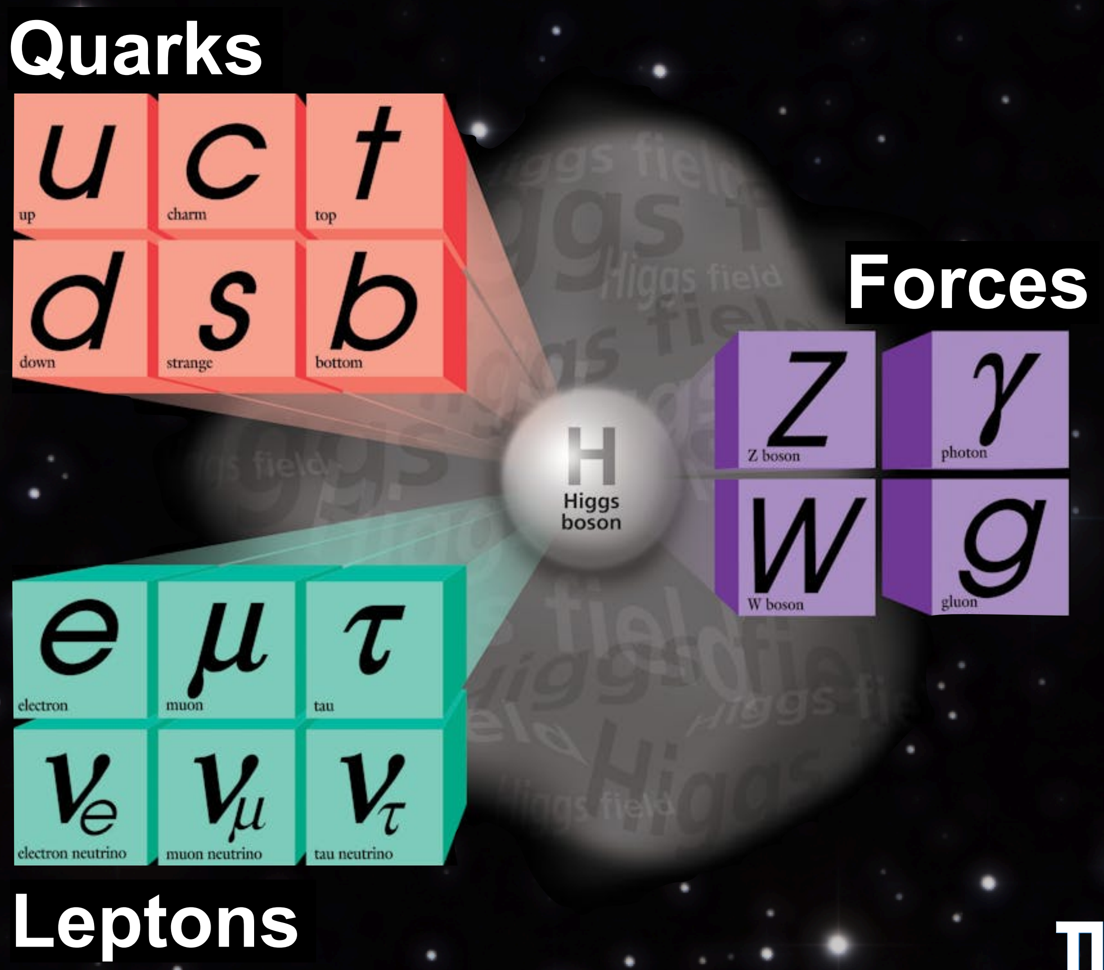
$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\gamma_\mu D^\mu\psi + |D_\mu\phi|^2 - V(\phi) + \boxed{(y_{ij}\bar{\psi}_i\psi_j + \text{h.c.})}$$



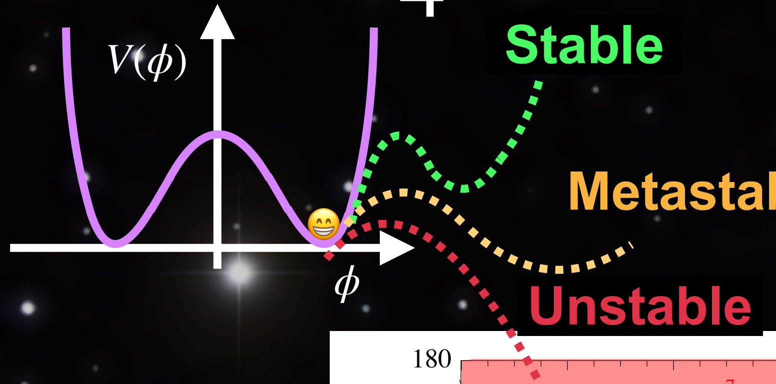


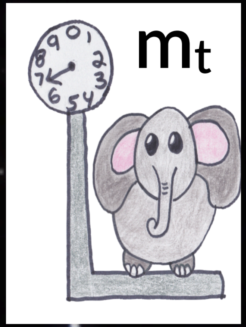
$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\gamma_\mu D^\mu\psi + \boxed{|D_\mu\phi|^2 - V(\phi)} + (y_{ij}\bar{\psi}_i\psi_j + \text{h.c.})$$

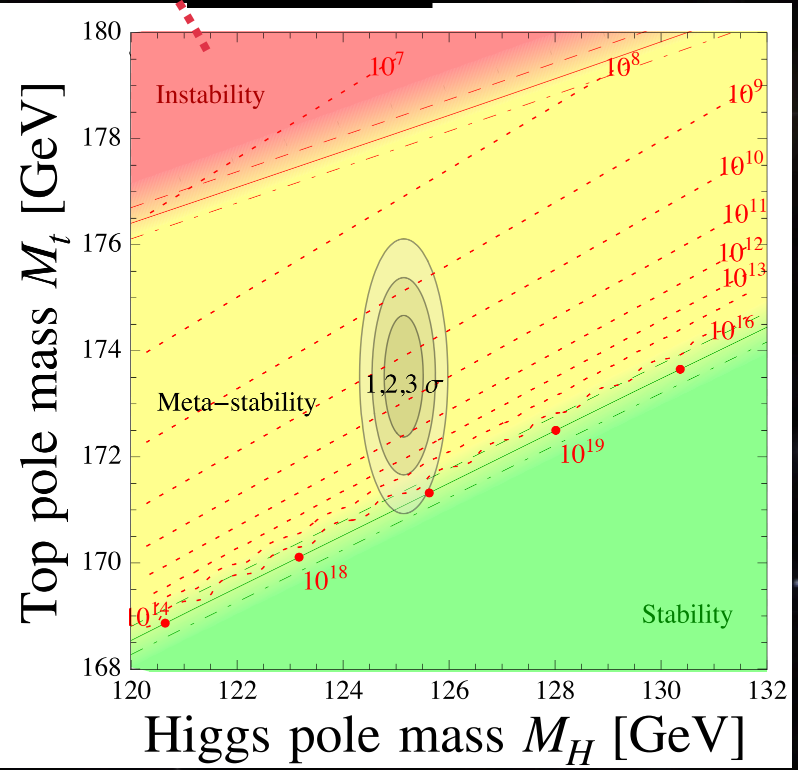


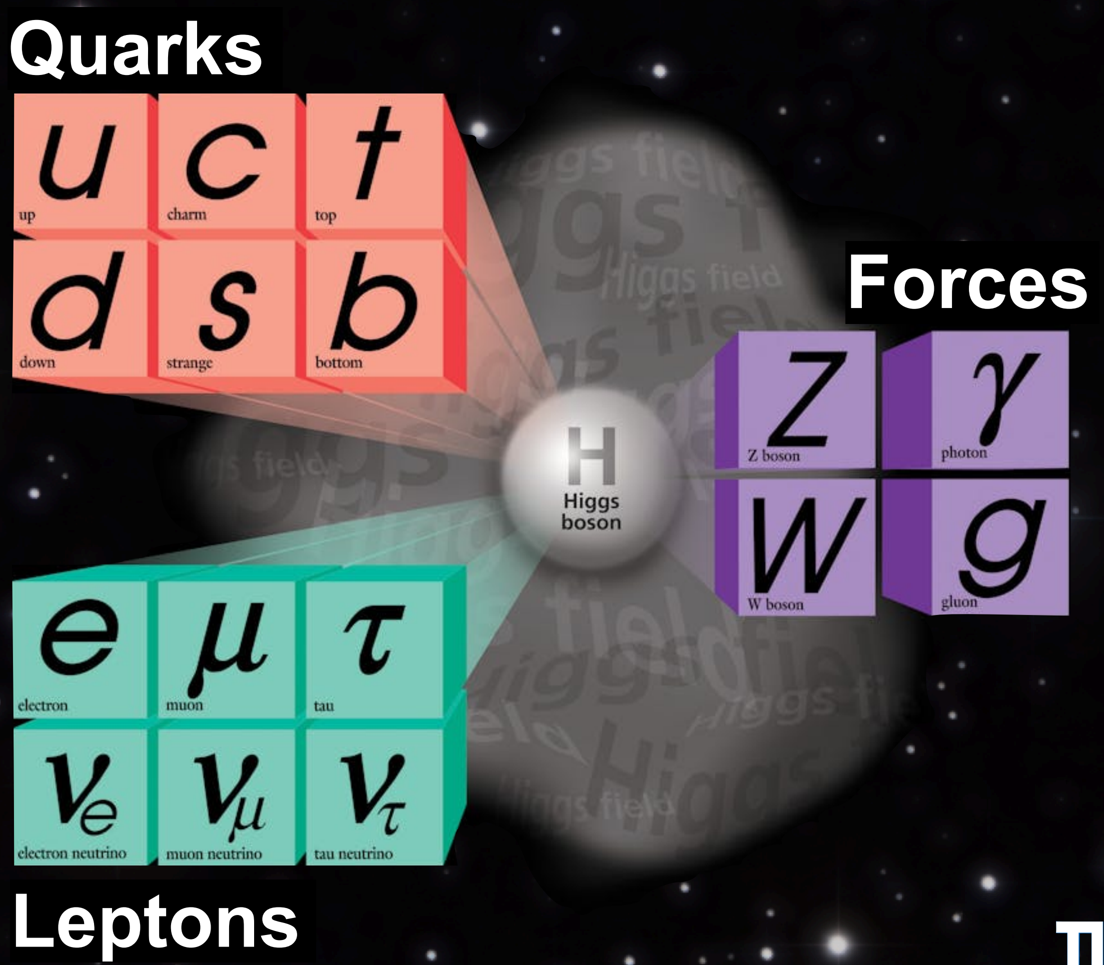


$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\gamma_\mu D^\mu\psi + |D_\mu\phi|^2 - V(\phi) + (y_{ij}\bar{\psi}_i\psi_j + \text{h.c.})$$





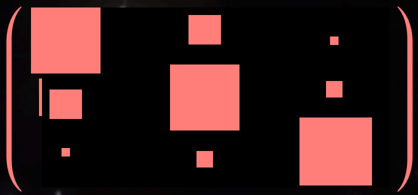


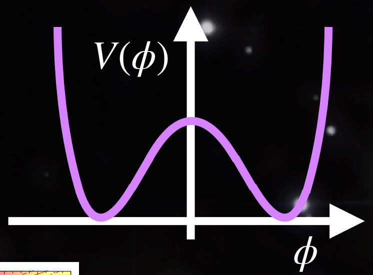


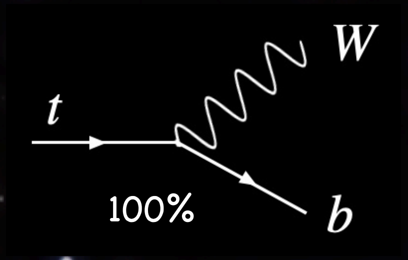
Physics 8, 108
1307.3536

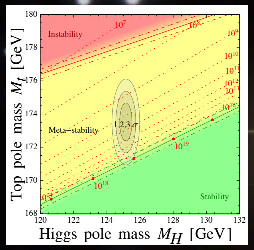
$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\gamma_\mu D^\mu \psi + |D_\mu \phi|^2 - V(\phi) + (y_{ij}\bar{\psi}_i\psi_j + \text{h.c.})$$

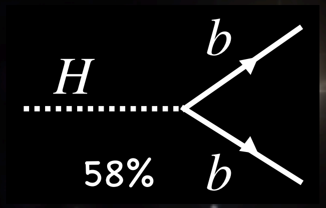
Key questions:

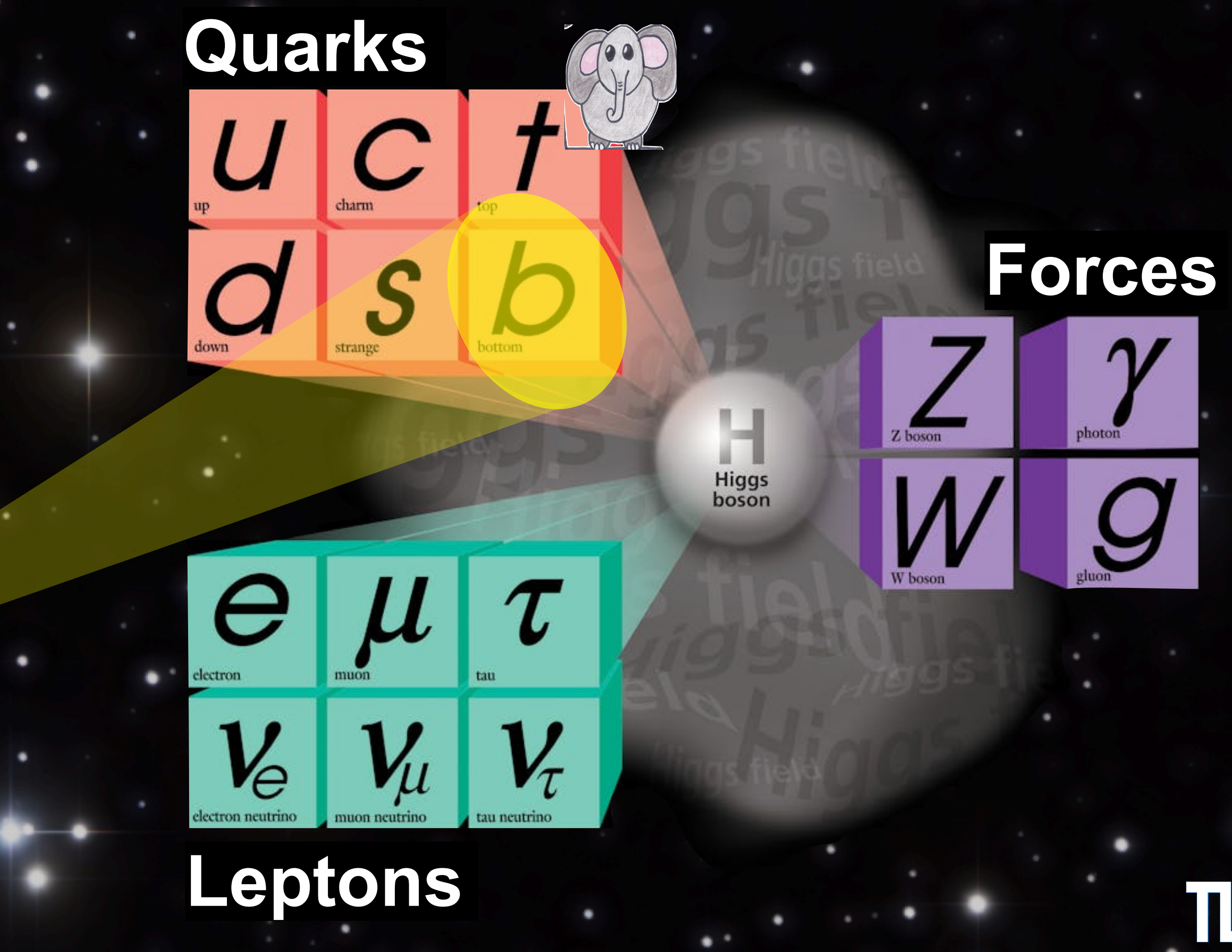




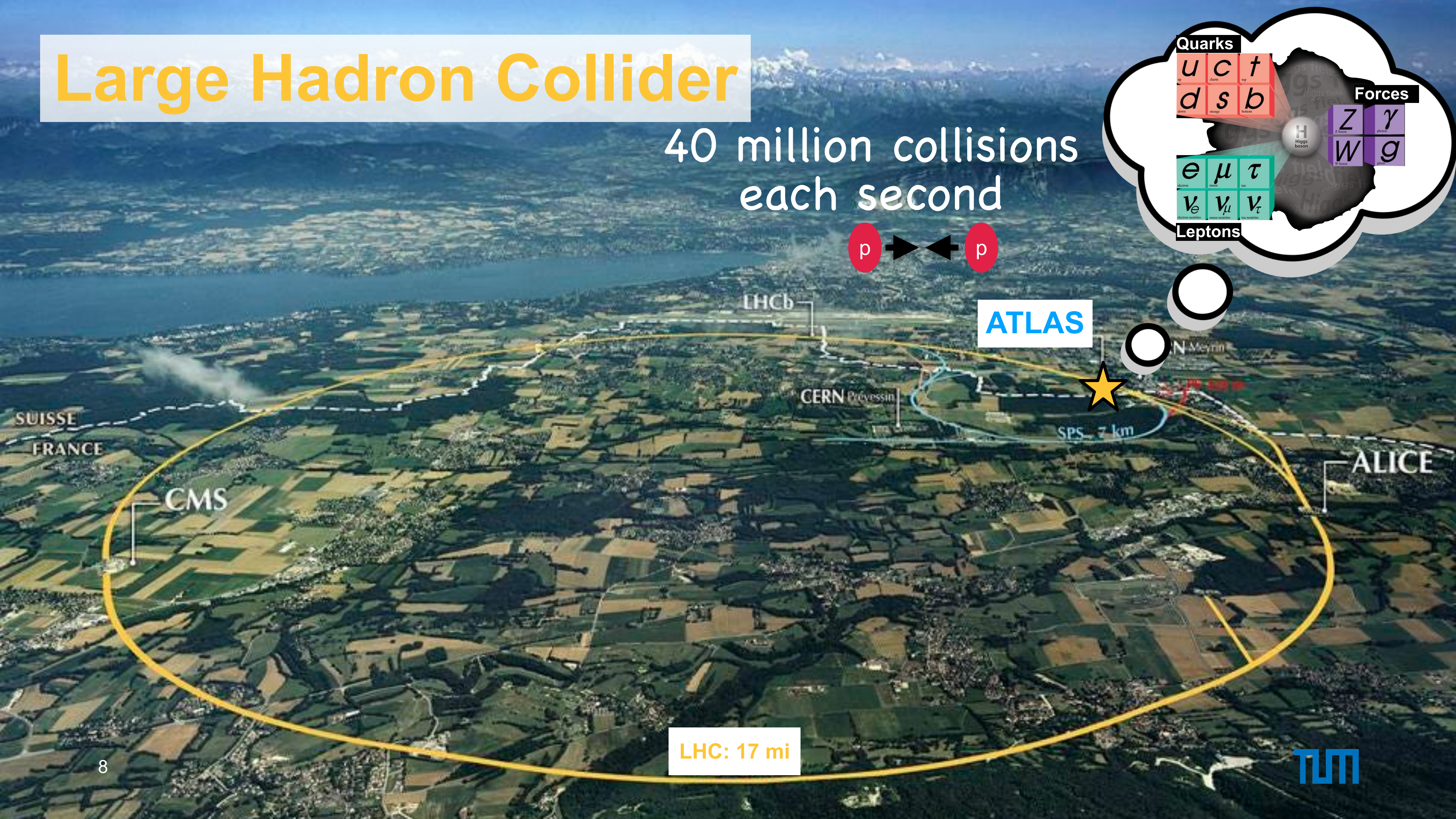




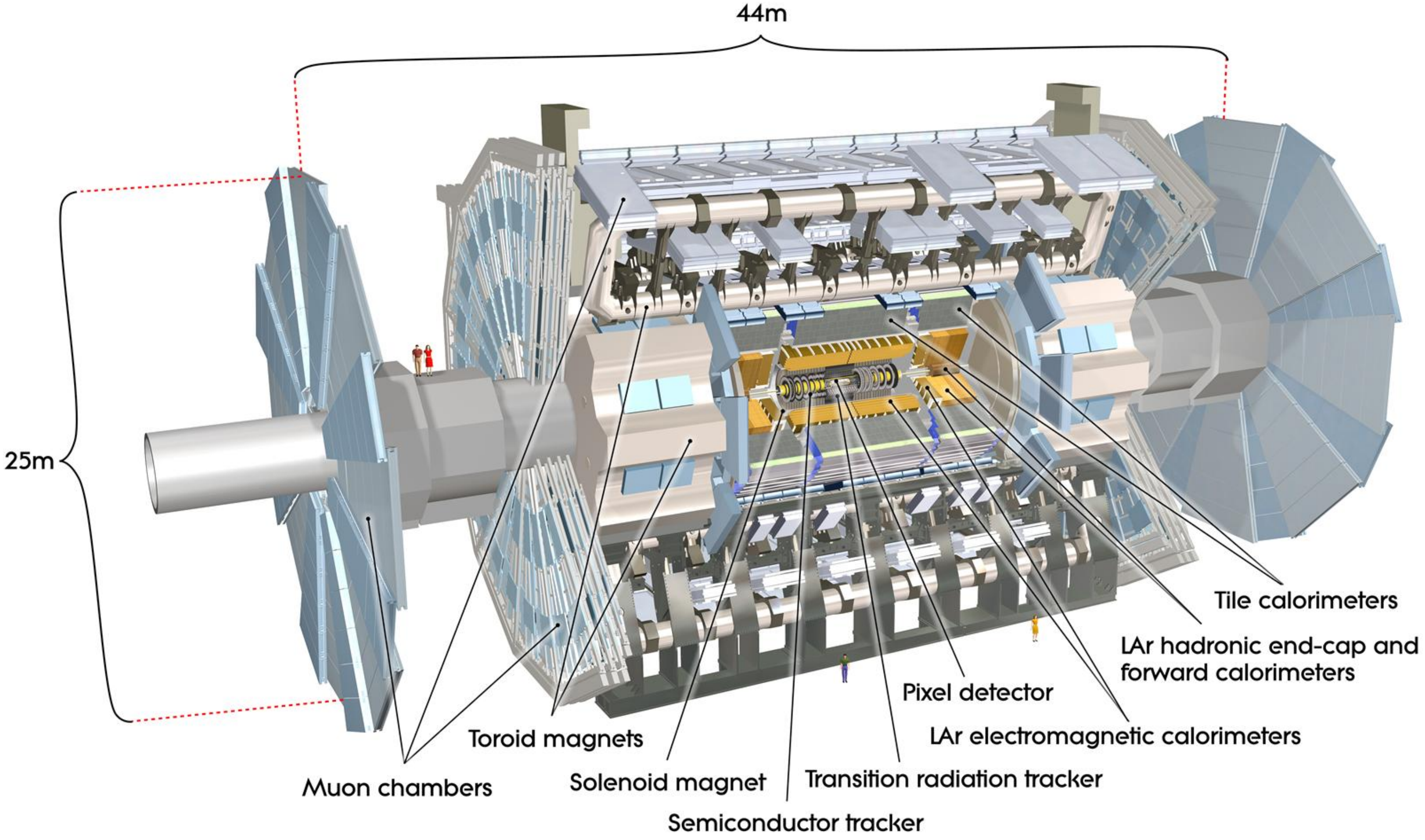




Large Hadron Collider



ATLAS

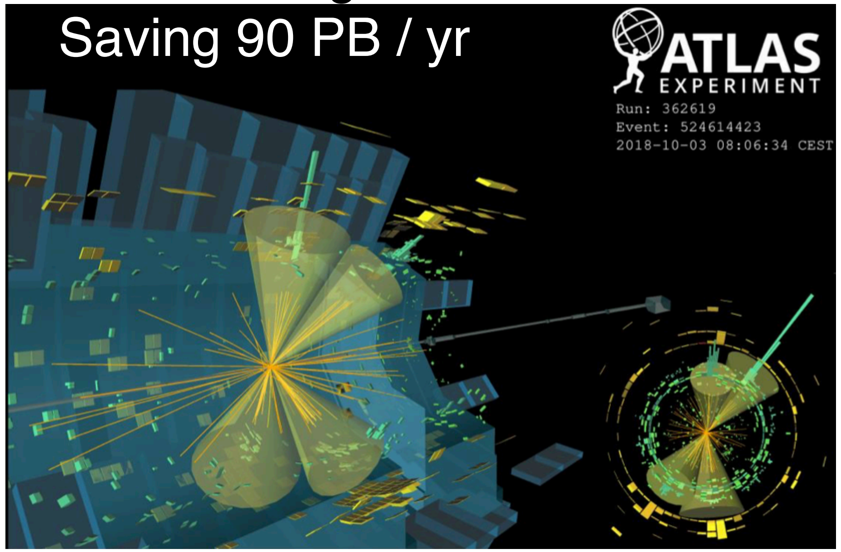


100 million read out channels!

What's exciting about data analysis now?

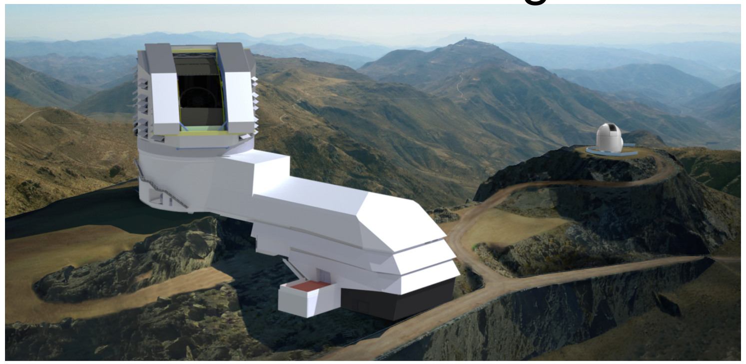
Increasingly large datasets!!

LHC : Collecting PB/ sec

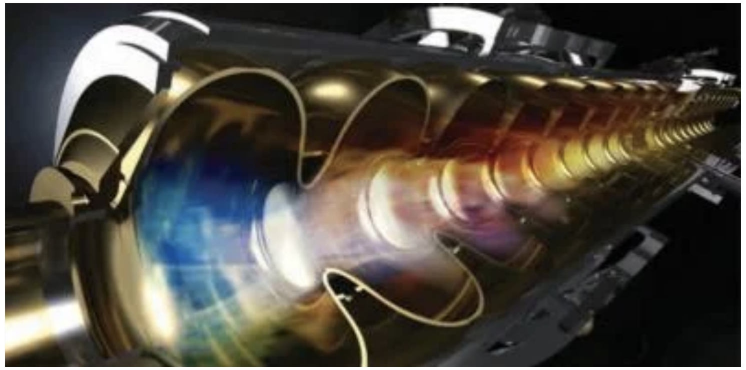


SKA : TB / s

LSST : 20 TB / night



LCLS : Tb /s



What's exciting about methods now?

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]

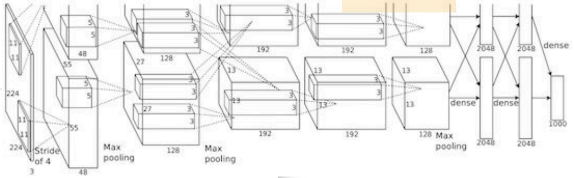
Physics, 2014

The New York Times

Scientists See Promise in Deep-Learning Programs

Share full article

"Alex Net"
Krizhevsky, Sutskever, Hinton, 2012



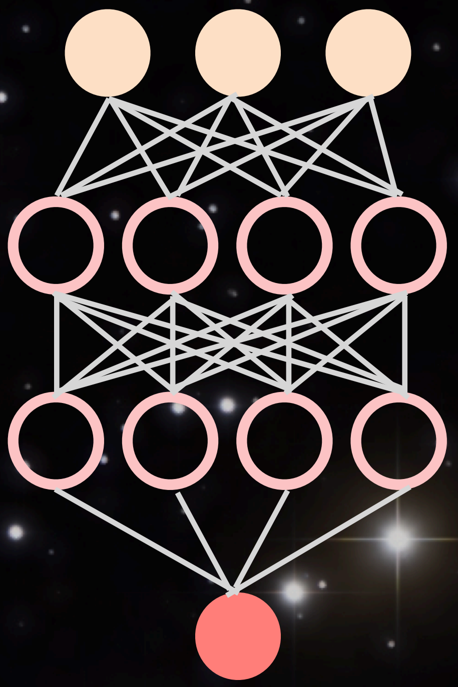
A voice recognition program translated a speech given by Richard F. Rashid, Microsoft's top scientist, into Mandarin Chinese. Hao Zhang/The New York Times

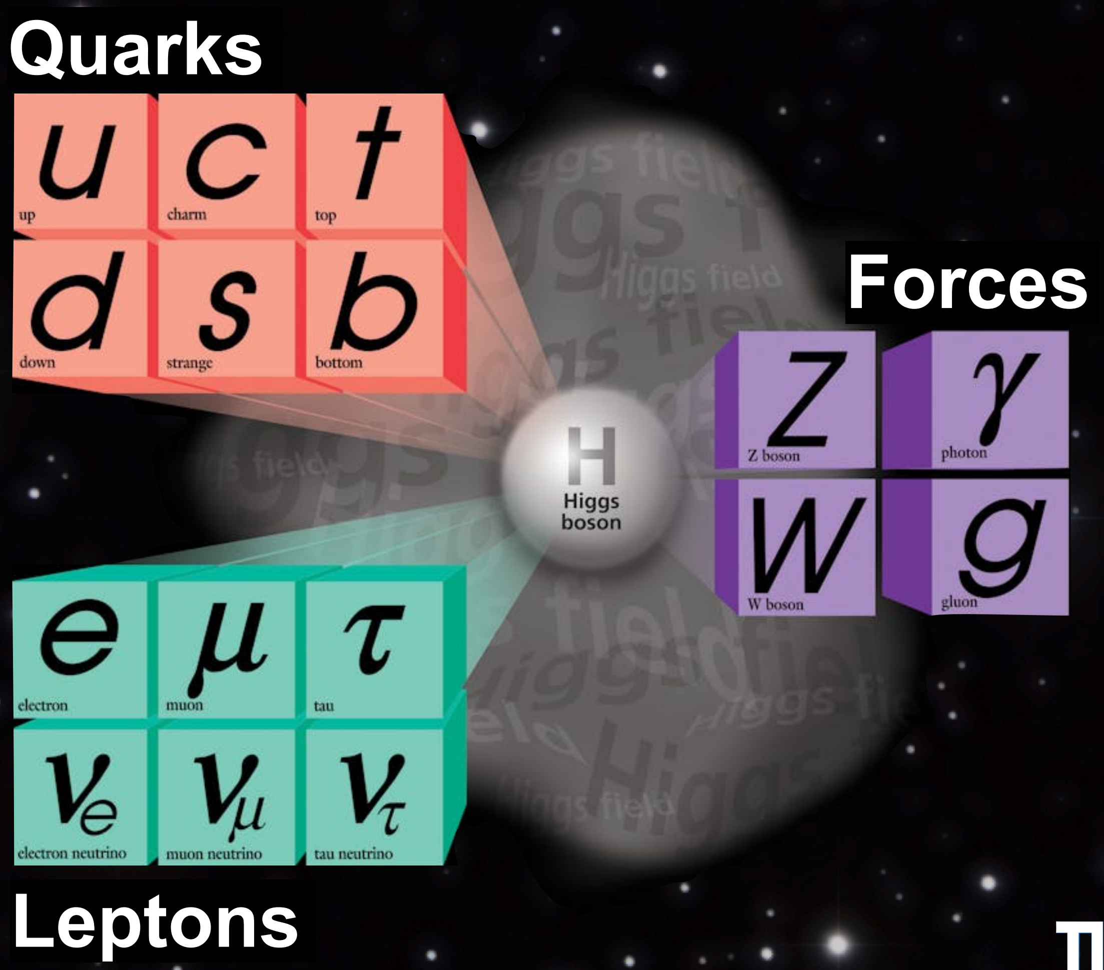
By John Markoff
Nov. 23, 2012

[article]

Physics, 2024

$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + i\bar{\psi}\gamma_\mu D^\mu\psi + |D_\mu\phi|^2 - V(\phi) + (y_{ij}\bar{\psi}_i\psi_j + \text{h.c.})$$

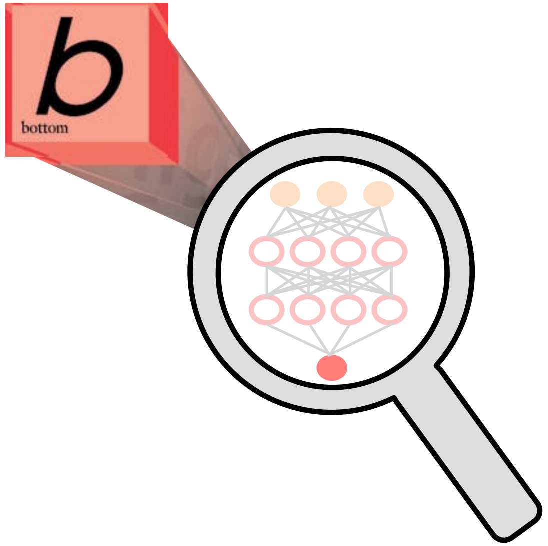




AI for science: How to gain the *most* out of our physics datasets

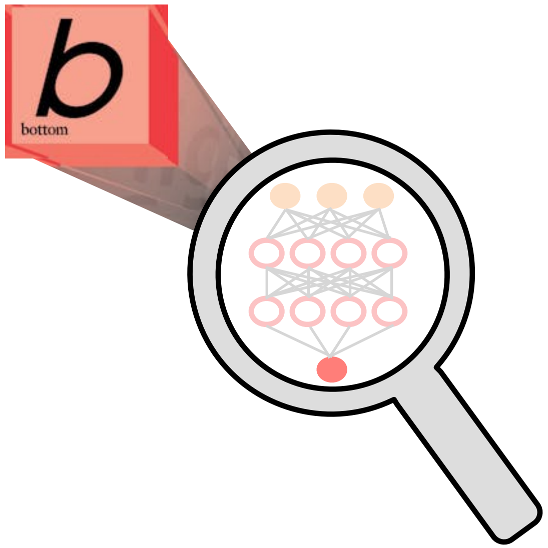
Slide design from [K. Leney]

Future Outlook



Transformer-era

Deep Learning in FTAG



Future Outlook

Slide design from [K. Leney]

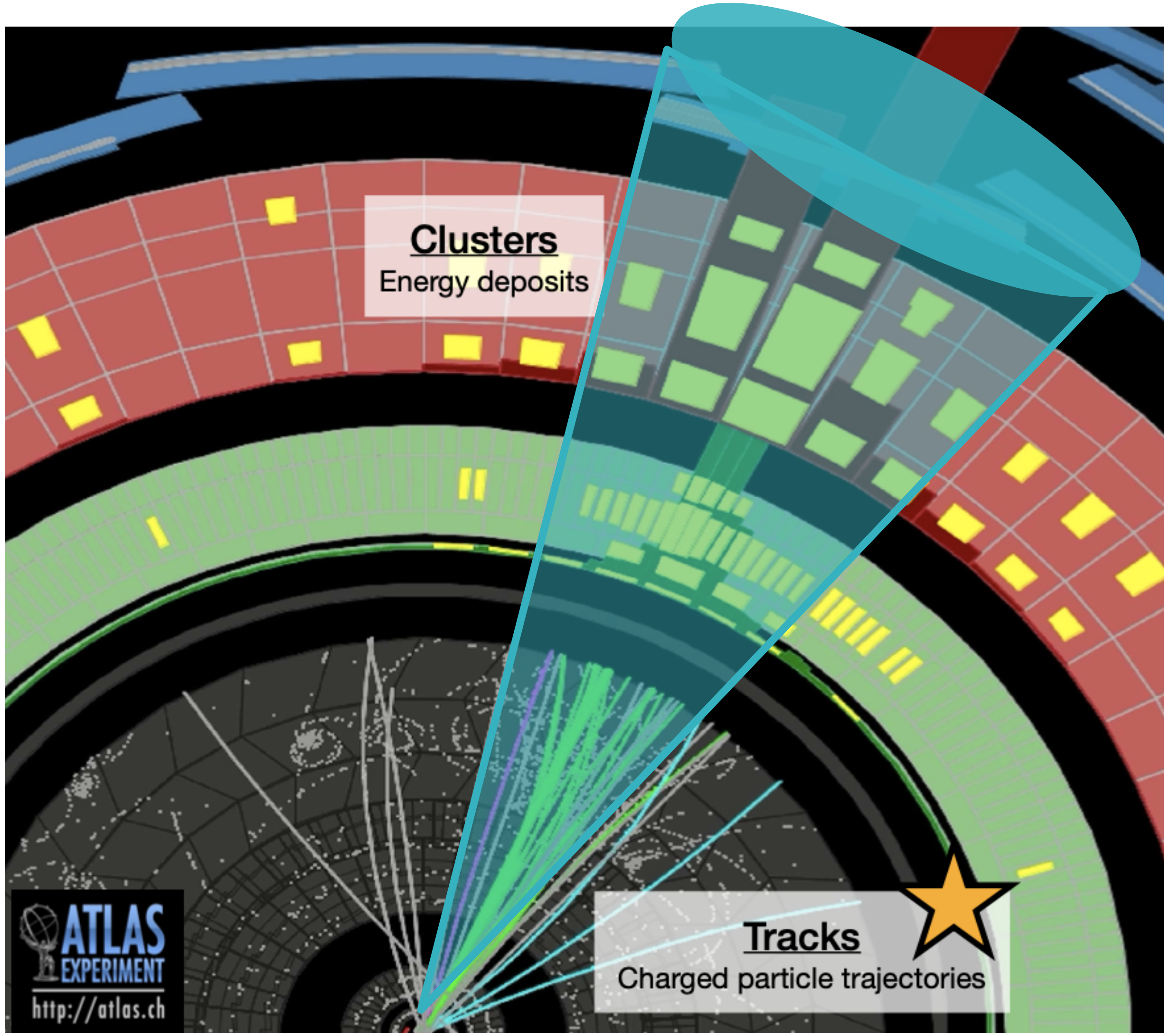
Transformer-era

Deep Learning in FTAG

- b-jets on ATLAS
- Why Deep Learning
  - RNNIP
  - Physics impact
- DIPS

The European Physical Journal C
Vol 73 3 (2013) 2304

Quark signature



Quark ➜ reconstructed as collimated spray of particles

Jet : Unsupervised clustering algorithm

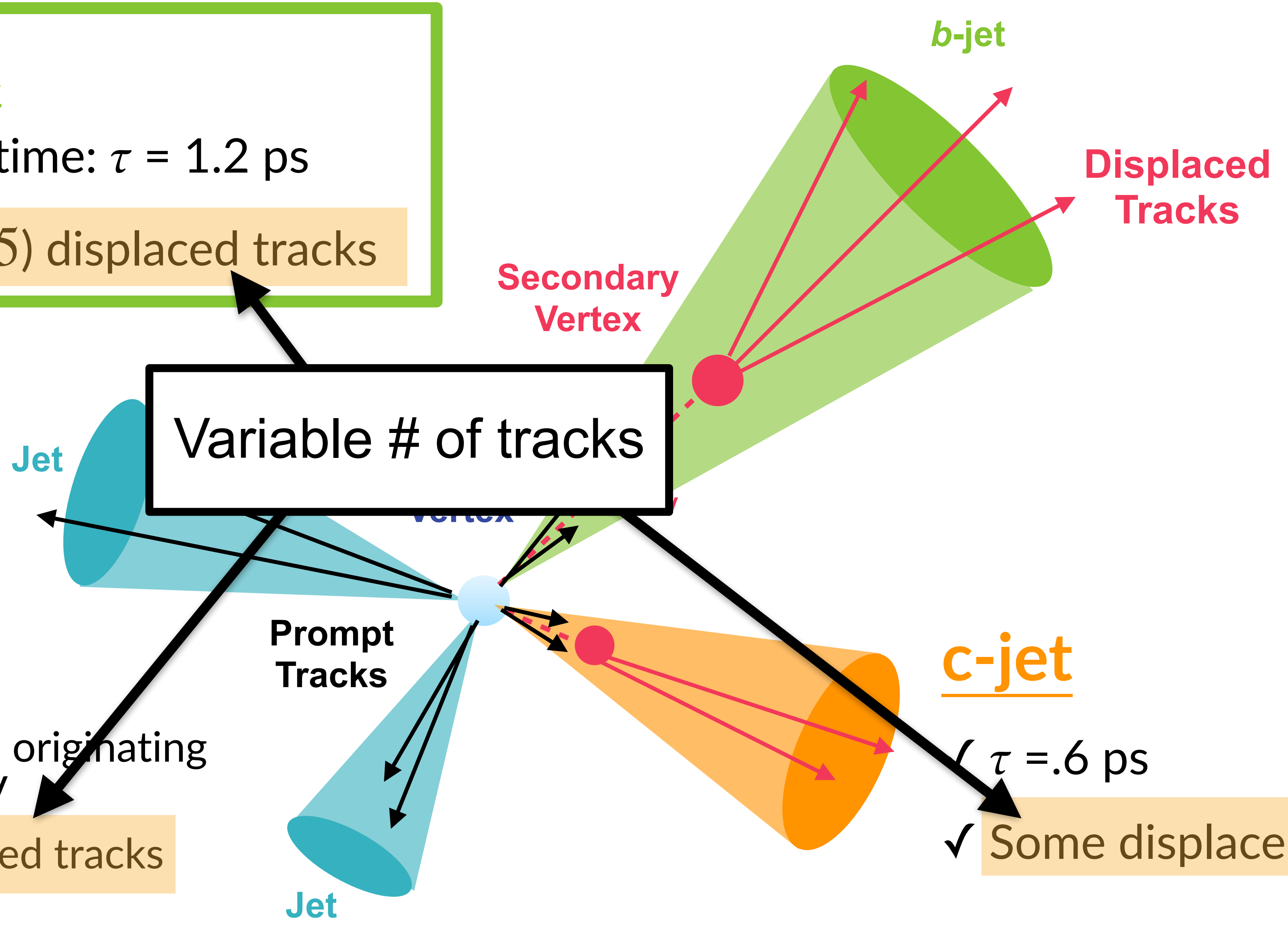
- Proxy for the quark
- Cluster with anti-kT
  - R=0.4 for b-jets ✨
  - R=1.0 for Higgs-jets

😍 b-jet

✓ "Long" lifetime: τ = 1.2 ps

✓ Many ($\approx$ 5) displaced tracks

Variable # of tracks



light jet

u,d,s g

✓ Most tracks originating from the PV

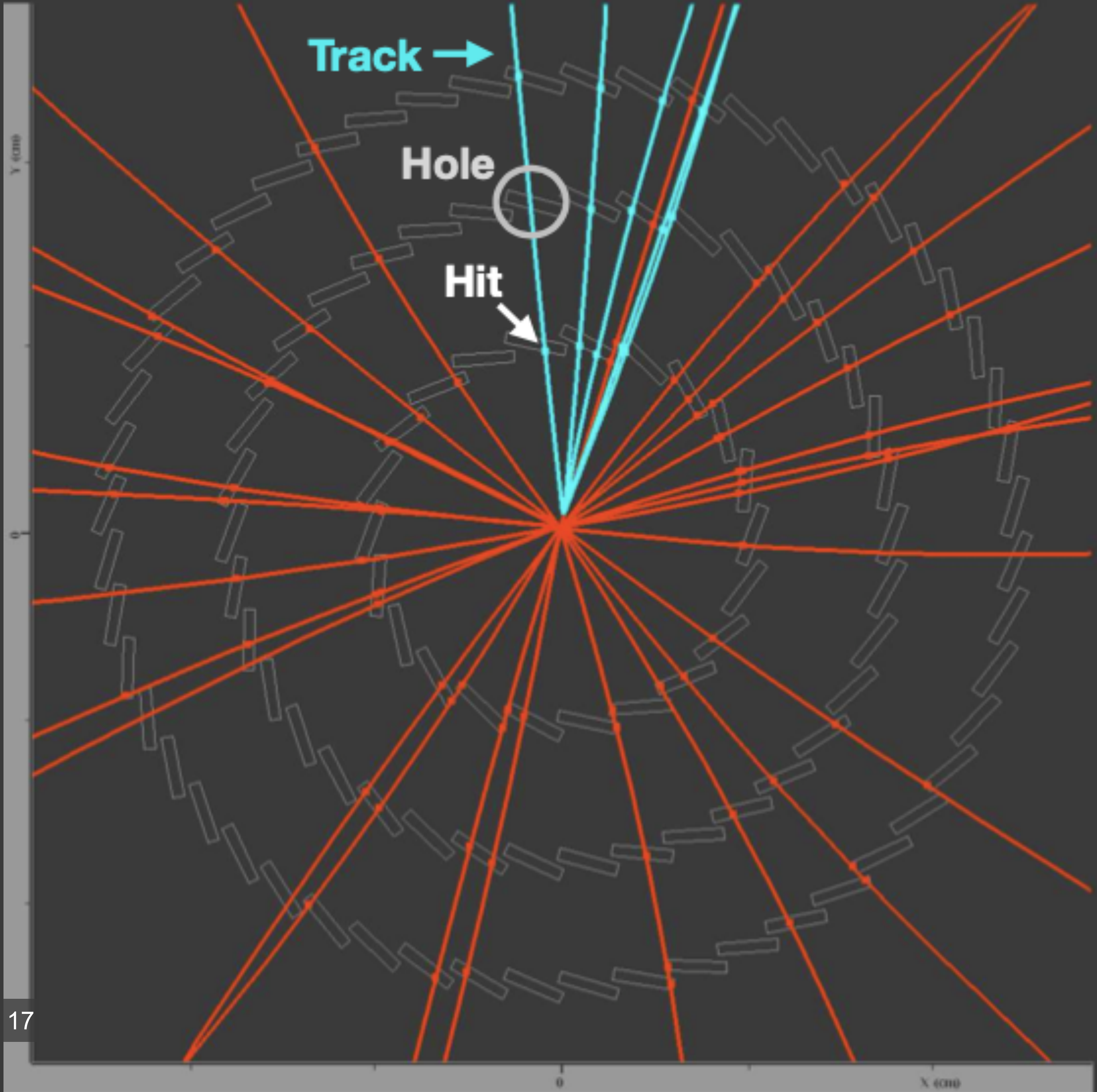
✓ Few displaced tracks

c-jet

✓ τ =.6 ps

✓ Some displaced tracks

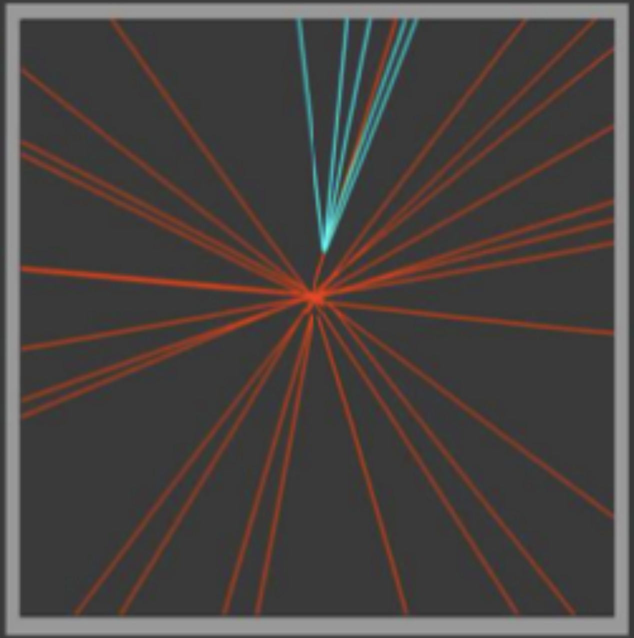
1005.5254



ATLAS EXPERIMENT

Run 142195, Event 284154

Decay length = 3.7 mm
Decay length signficance = 22
Lifetime = 3.1 ps
Vertex mass = 2.5 GeV
Number of tracks = 5



ATLAS b-jet classifiers

Low level

Impact Parameter based

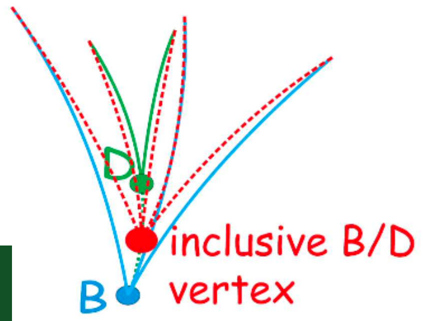
IP2D

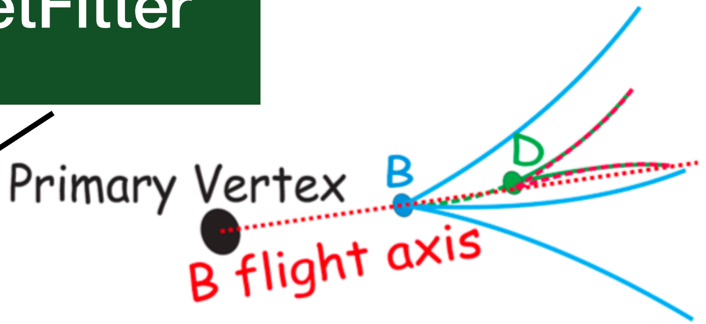
IP3D

Vertex based

SV1

JetFitter





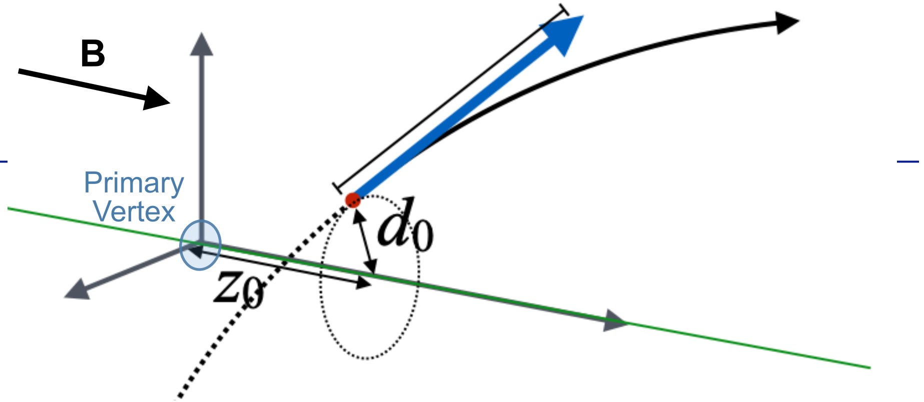
High level

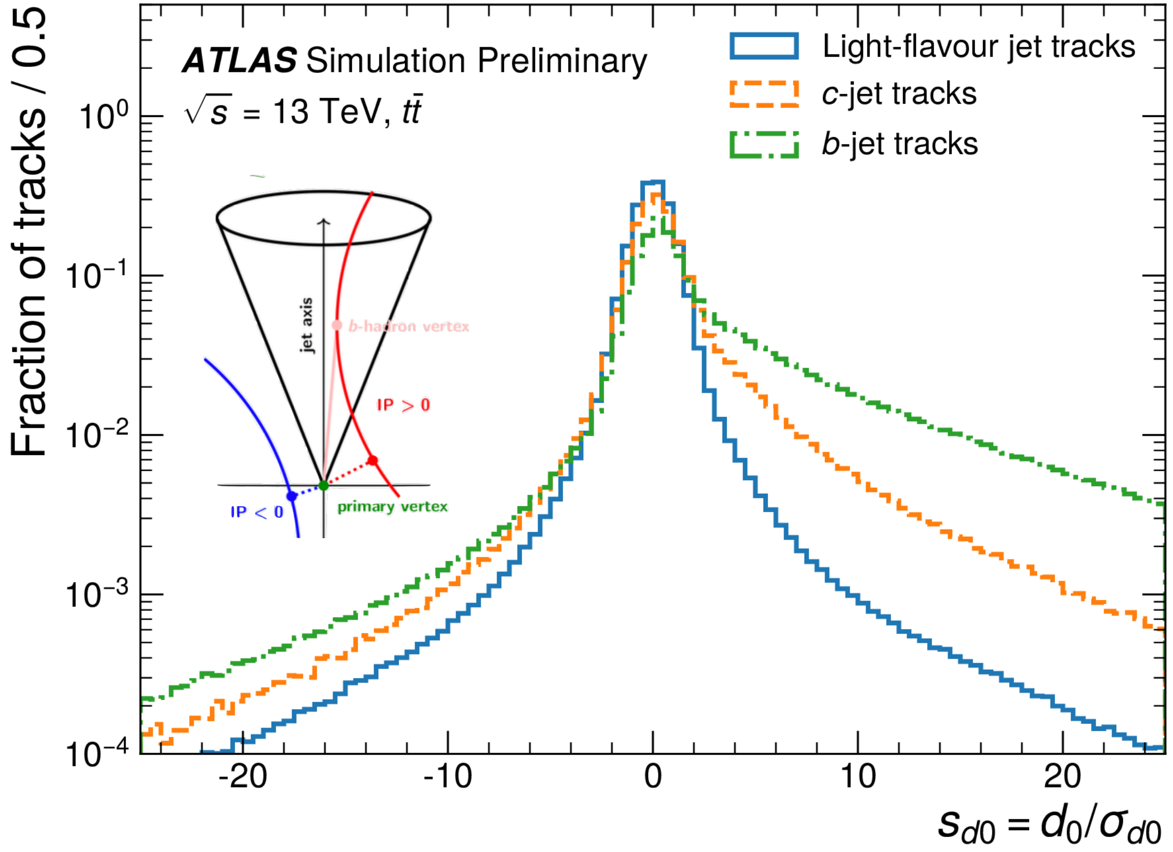
DL1

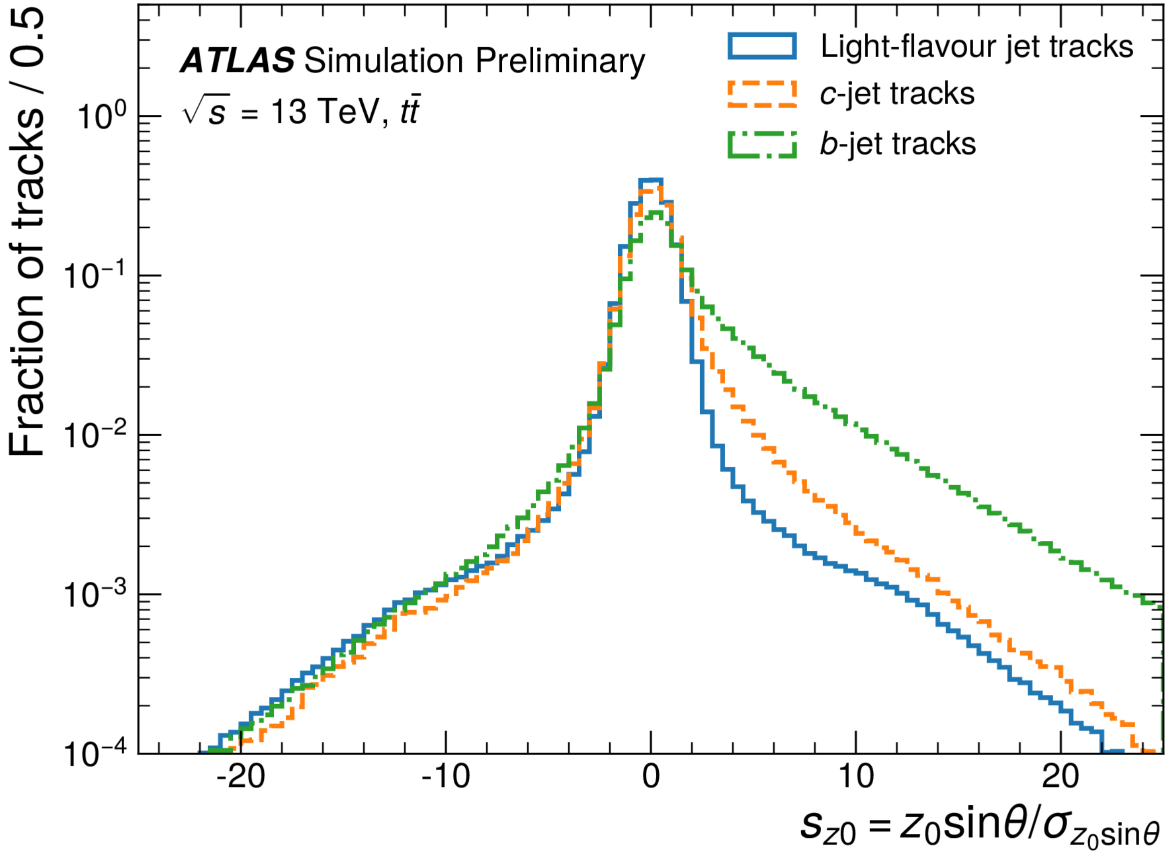
NN trained on outputs
of other classifiers

b-tagging

Key variable: impact parameter





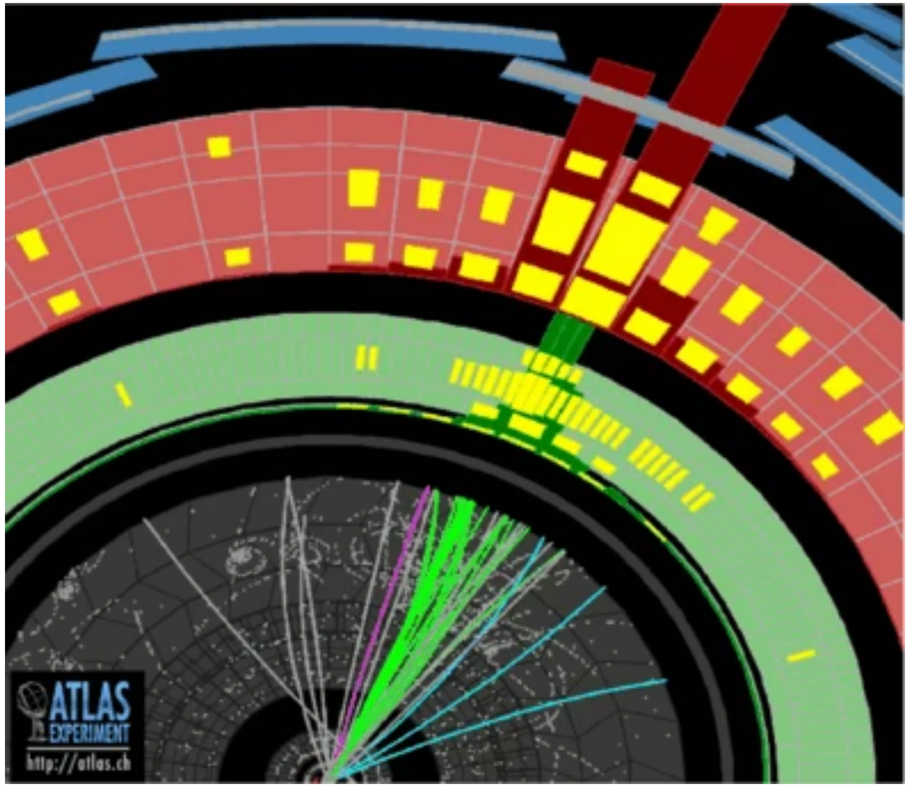


How do we aggregate this information?

What we have:

- Collection of tracks
  - $X_i : i = \{1, \dots, \underline{n}\}$

    variable # of tracks
- Each track has features
  - $X_i \in \mathbb{R}^m$ ← E.g, impact parameters momenta, quality
- Jet has labels
  - Y: {b, c, light} – or –
  - Y: {bb, cc, top, QCD}



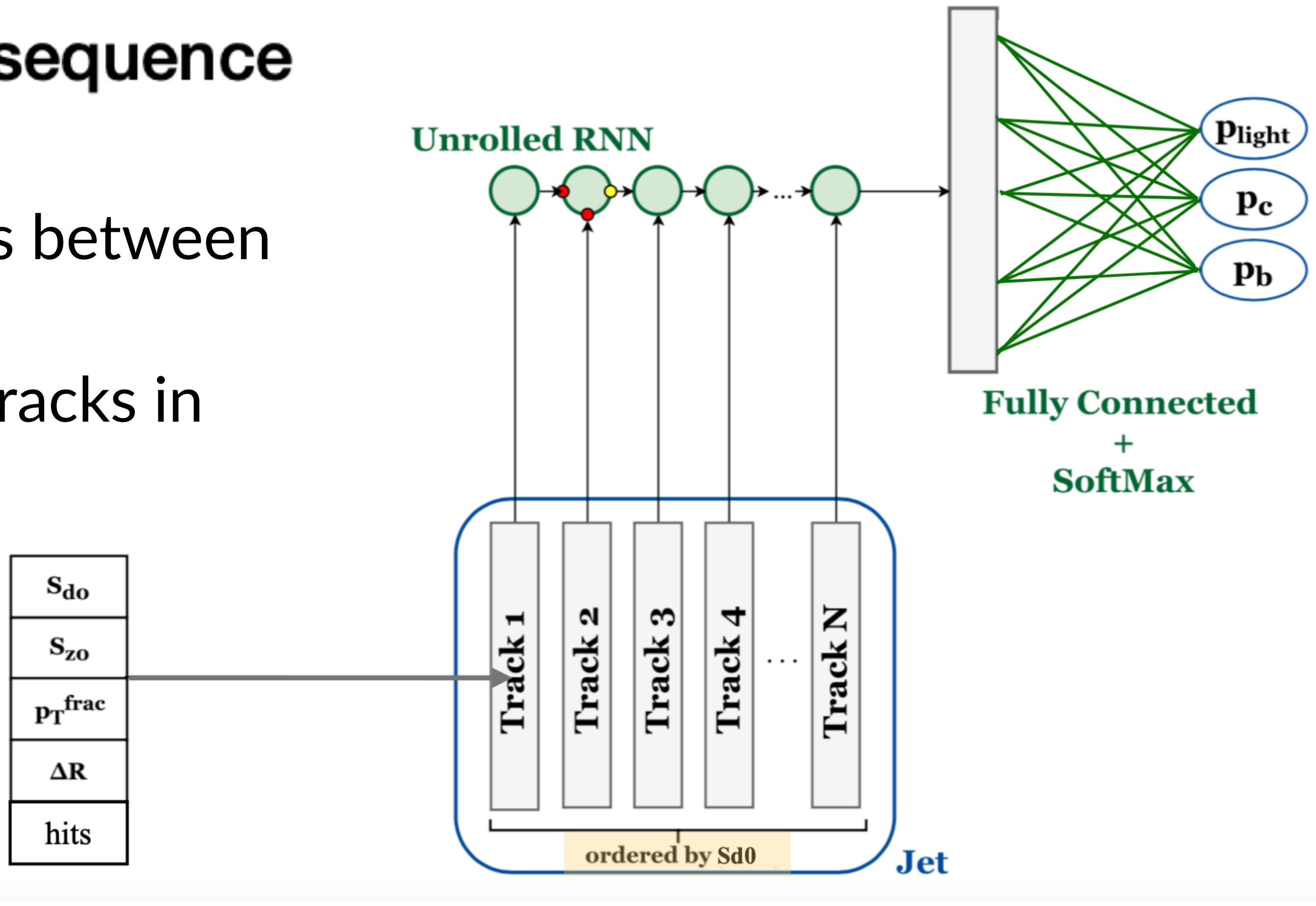
High dimensional problem
$n \cdot m \sim \mathcal{O}(10^3)$

What we want: $p(Y \mid X_1, \dots, X_n)$

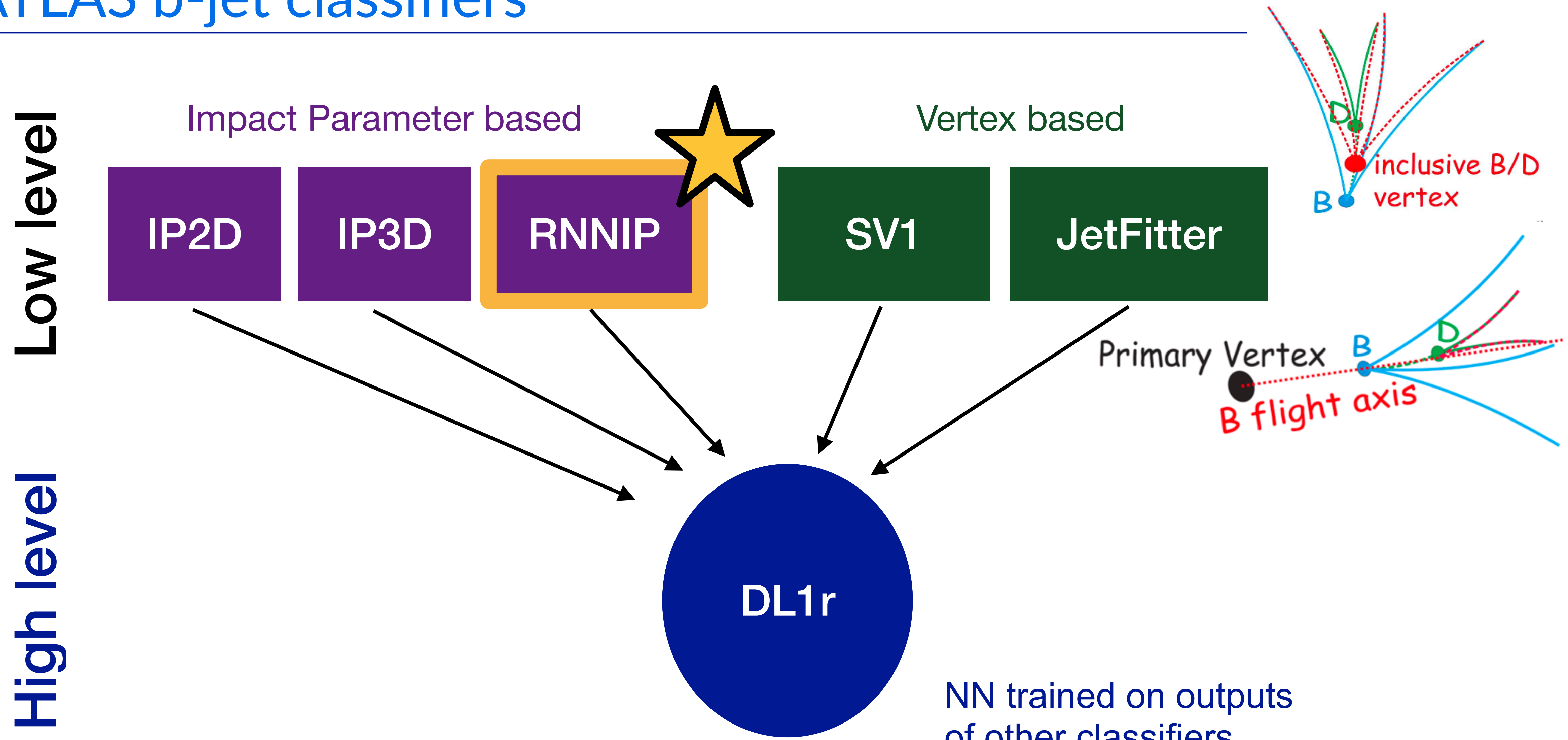
Recurrent Neural Network

Model the jet as a sequence

- ☑ Account for correlations between tracks
- ☑ Allow for variable # of tracks in the jet
- ☑ Avoids curse of dimensionality - add more features

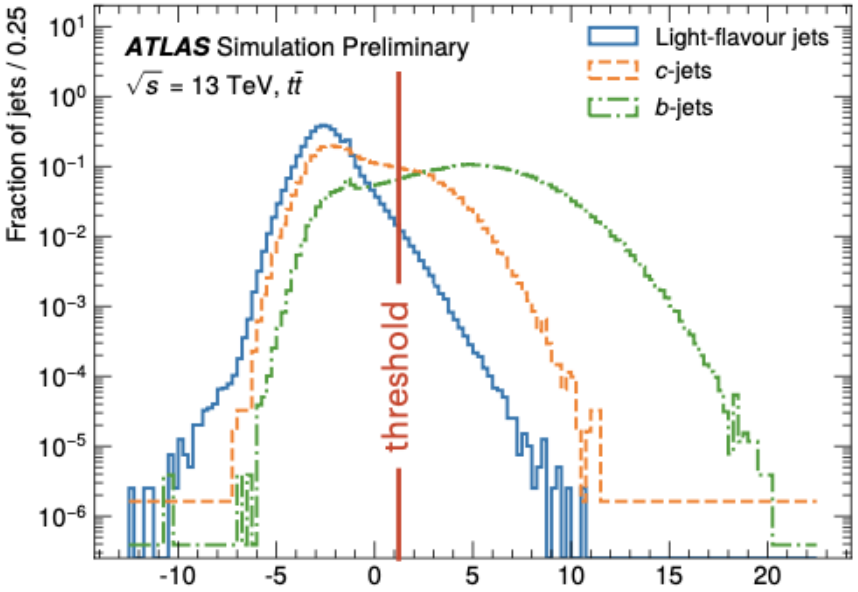


ATLAS b-jet classifiers



NN trained on outputs
of other classifiers

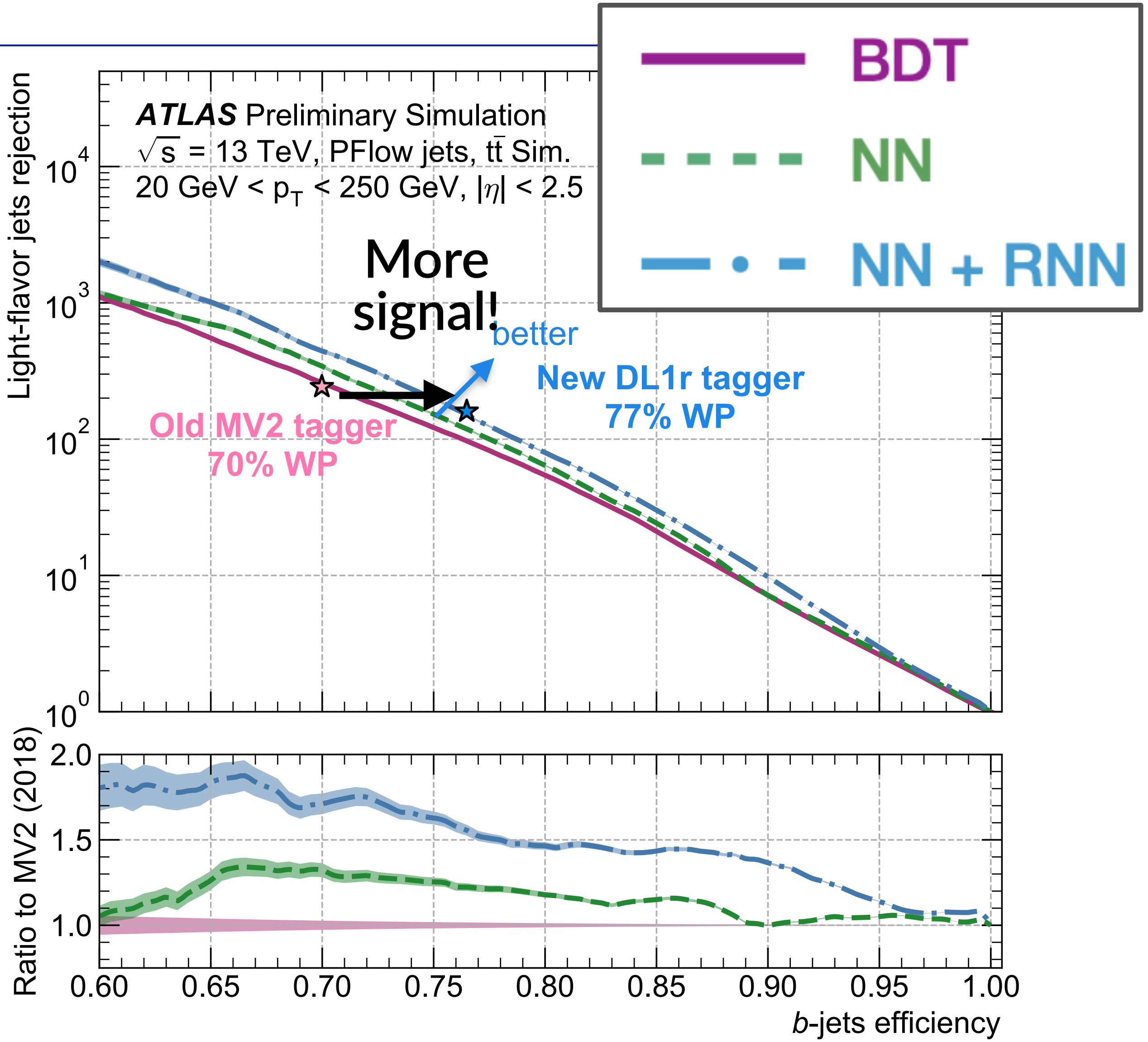
DL1r classifier



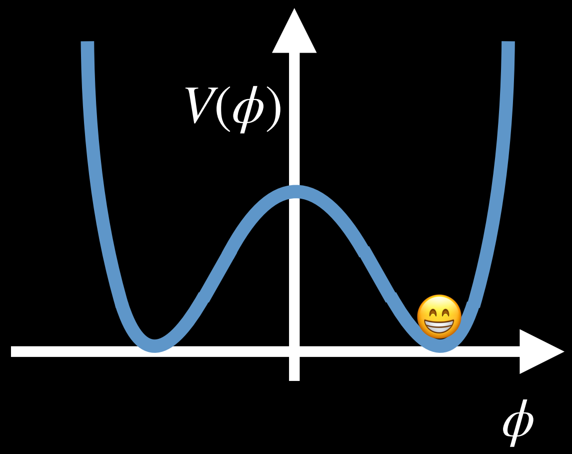
$$D_b = \log \frac{p_b}{f_c p_c + (1 - f_c) p_l}$$

$$\text{b-jet efficiency} = \frac{\#\ \text{b-jets} > \text{threshold}}{\#\ \text{b-jets}}$$

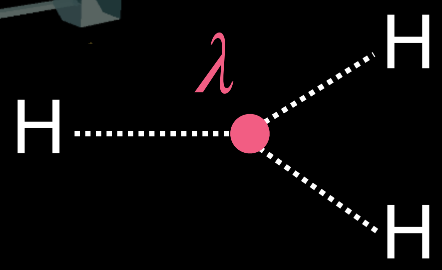
$$\text{Background rejection} = \frac{1}{\text{Background efficiency}}$$



How did b-tagging improvements help our HH analyses?

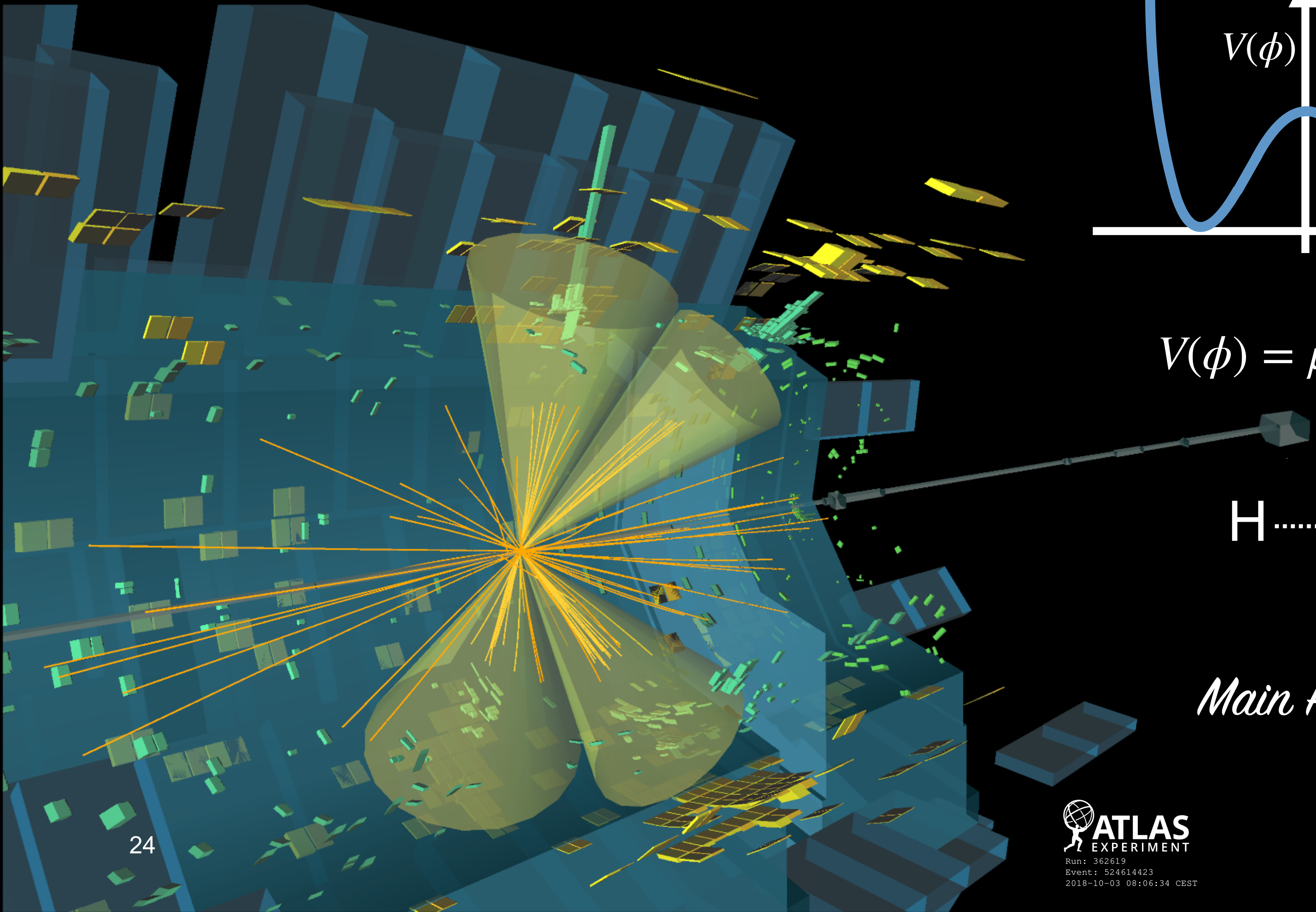


$$V(\phi) = \mu^2 h(x)^2 + \lambda v h(x)^3 + \frac{1}{4}\lambda h(x)^4$$

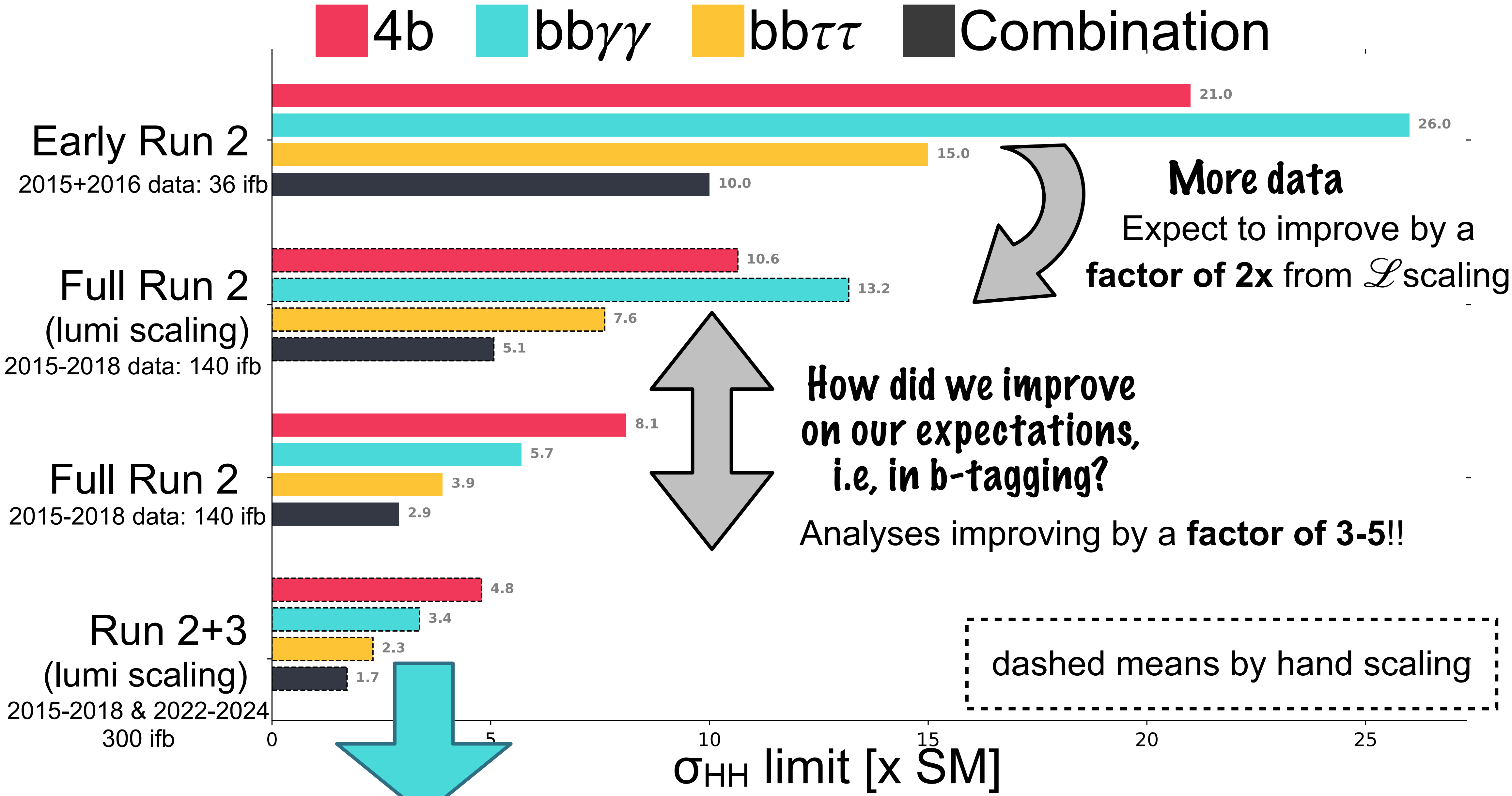


$$\frac{\text{signal}}{\text{background}} \sim 10^{-13}$$

Main HH channels 4b, $bb\gamma\gamma$, $bb\tau\tau$, all need b-tagging.



V. Cairo's slide

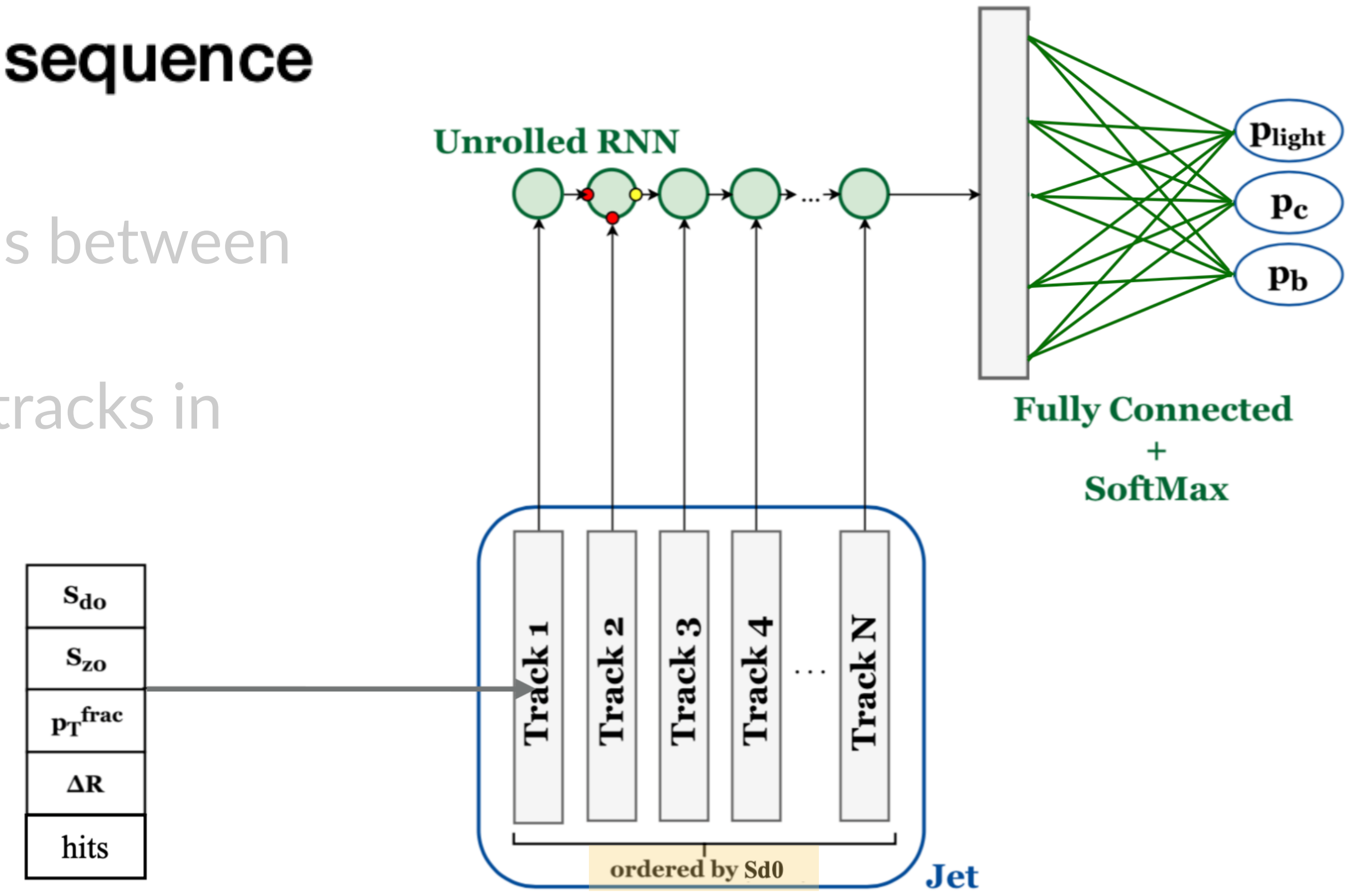


How will b-tagging improvements help bbγγ again exceed expectations?

Issue: ordering

Model the jet as a sequence

- ☑ Account for correlations between tracks
- ☑ Allow for variable # of tracks in the jet
- ☑ Avoids curse of dimensionality - add more features



Challenge: How to order??

When does order matter (?)

Natural language

(1) Mary likes John

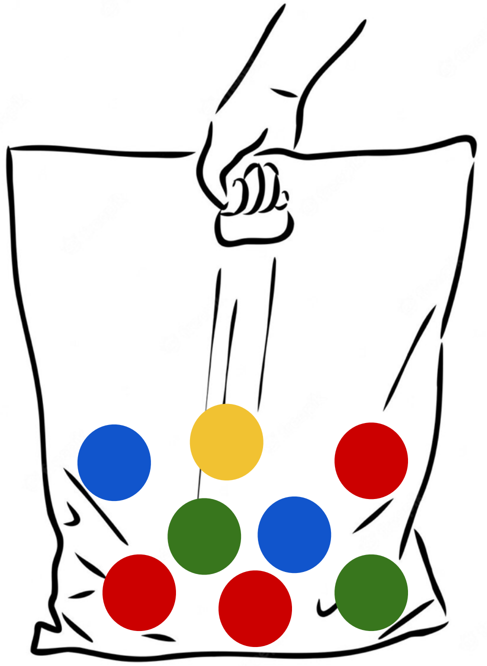
(2) John likes Mary

Same words… the order changes the meaning.

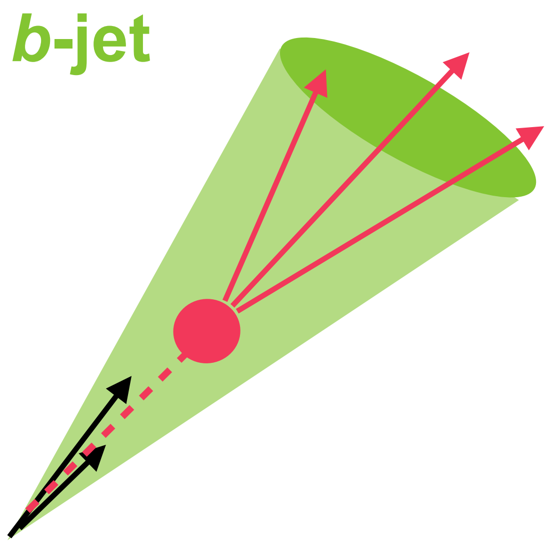
Permutation invariance

Set: Collection of objects without any specified order

Ex 1: # of colored balls in a bag



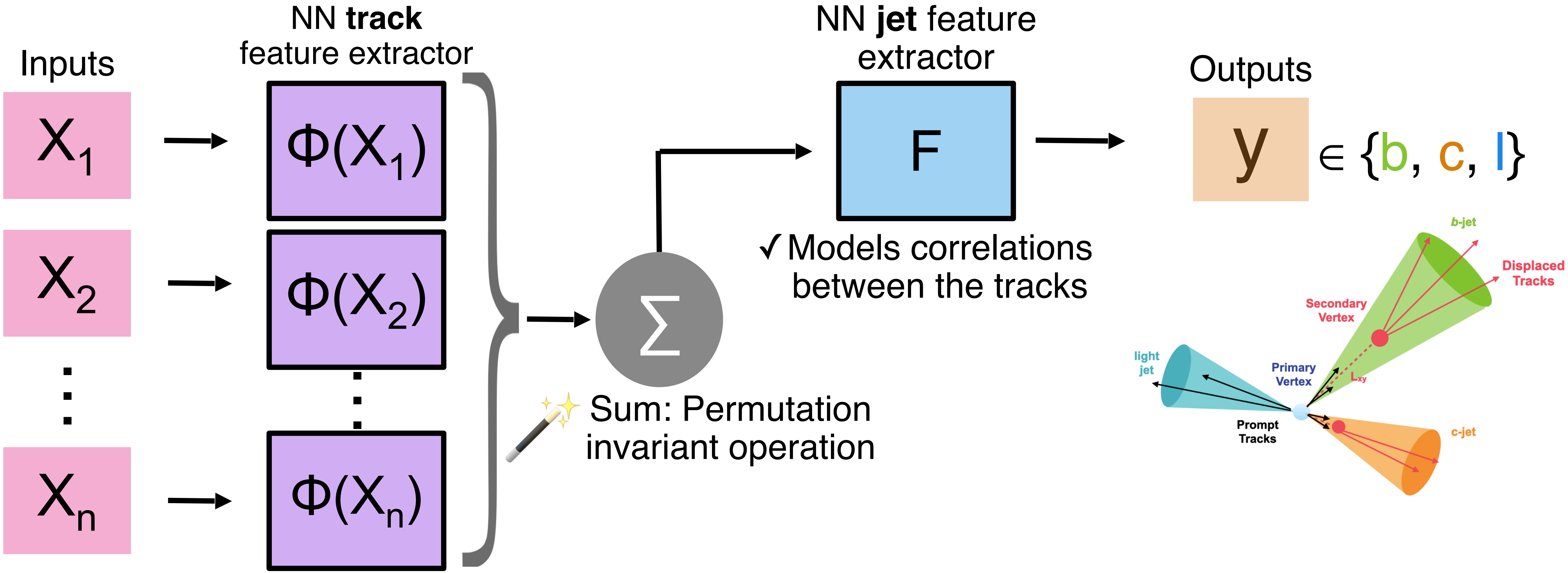
Ex 2: tracks in heavy flavour decay



Deep Set: Neural Network designed to operate on sets

Deep Sets

Network of networks!!

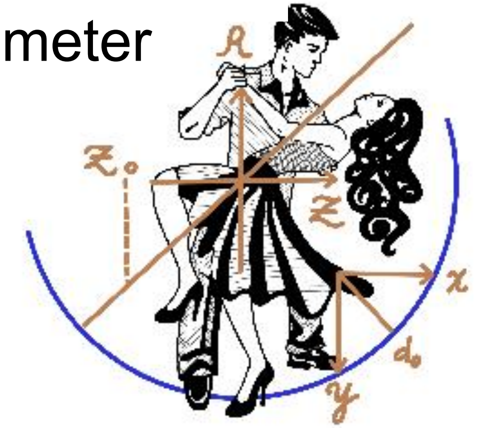


✓ Computationally efficient
✓Allows for variable # of tracks

Energy Flow networks: 1810.05165
& python package

DIPS

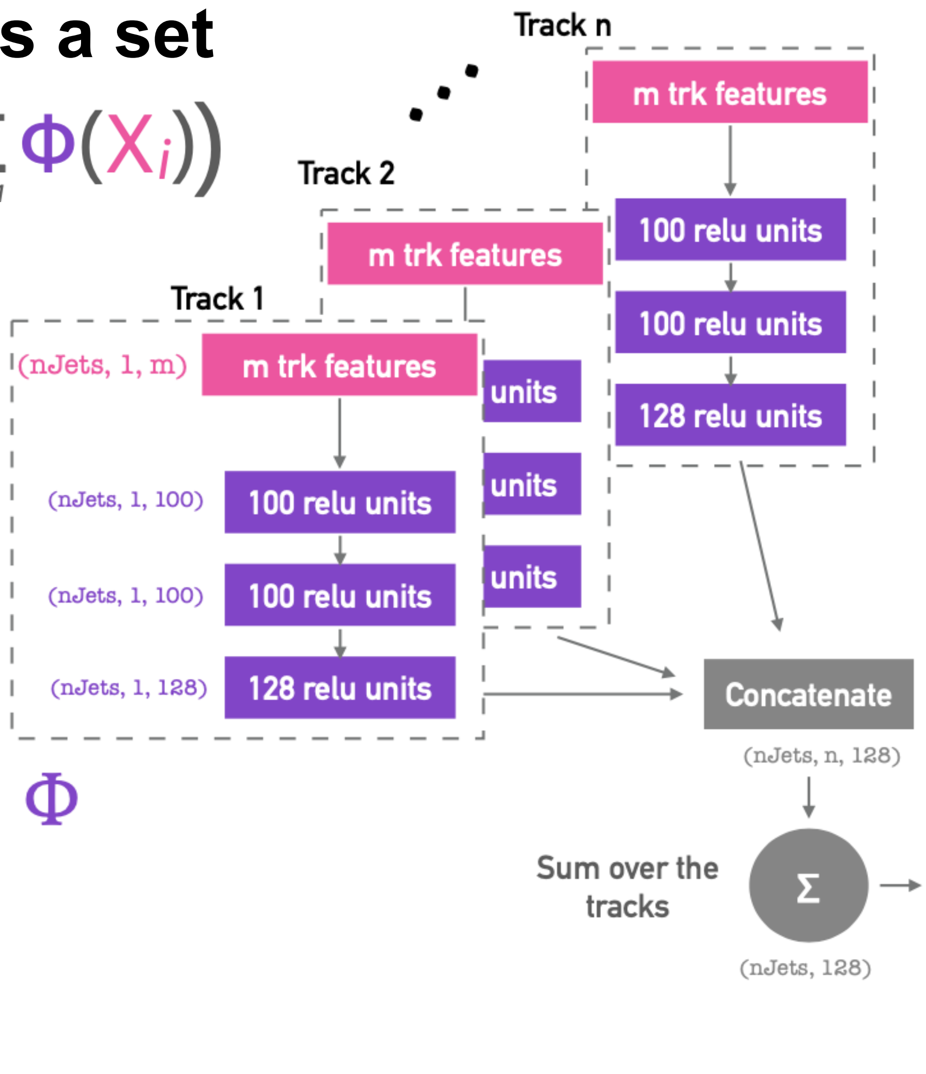
Deep Impact Parameter Sets

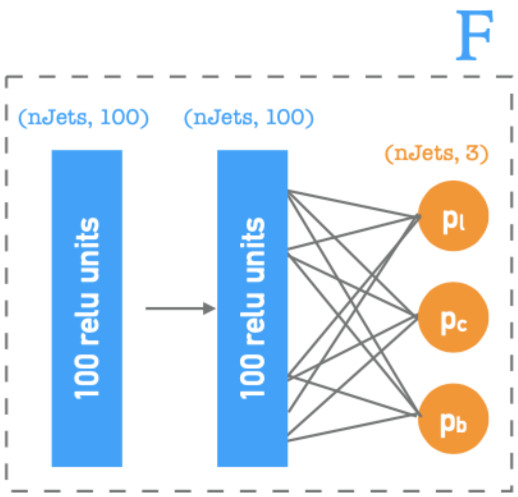


Model the jet as a set

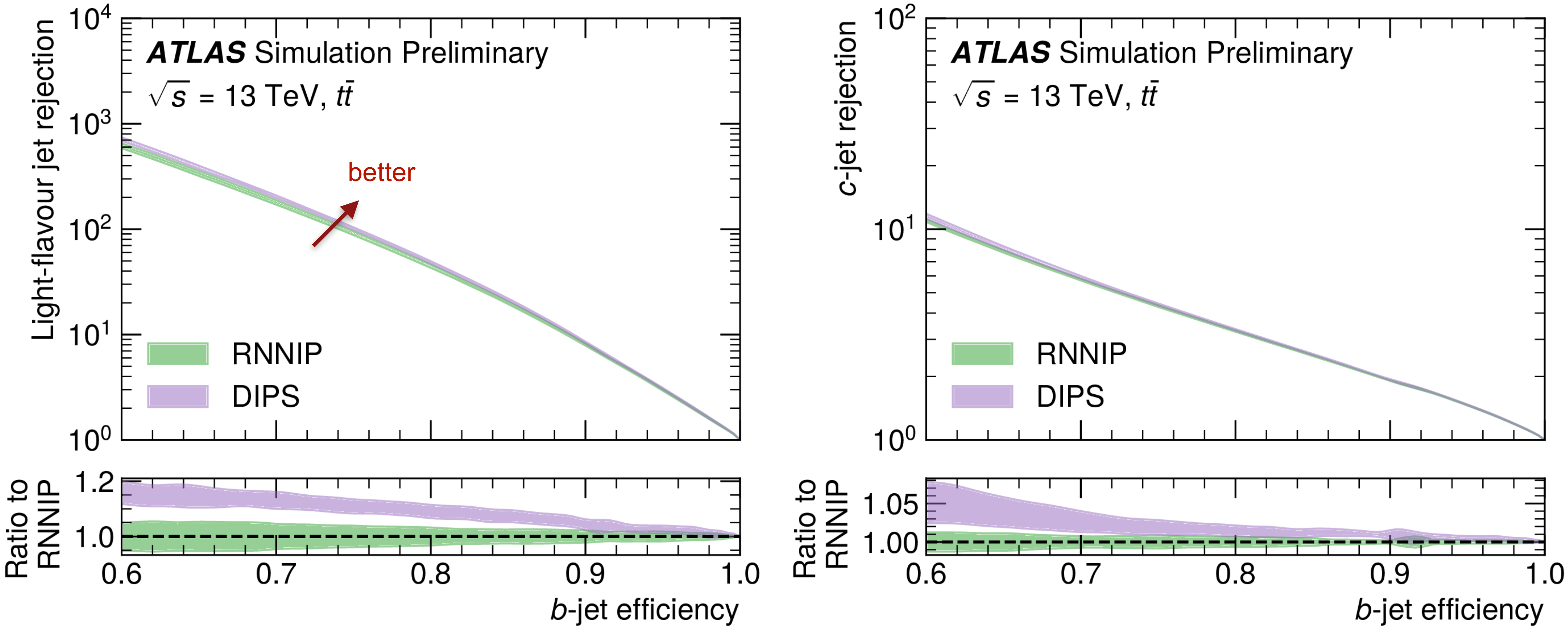
$$O(\{X_1,..X_n\}) = F\left(\sum_{i=1}^{n}\Phi(X_i)\right)$$

Same inputs as b-tagging RNN

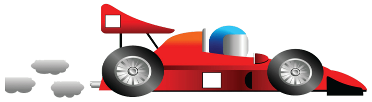




DIPS



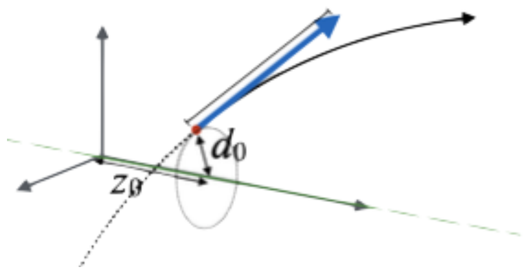
Similar performance with the same inputs



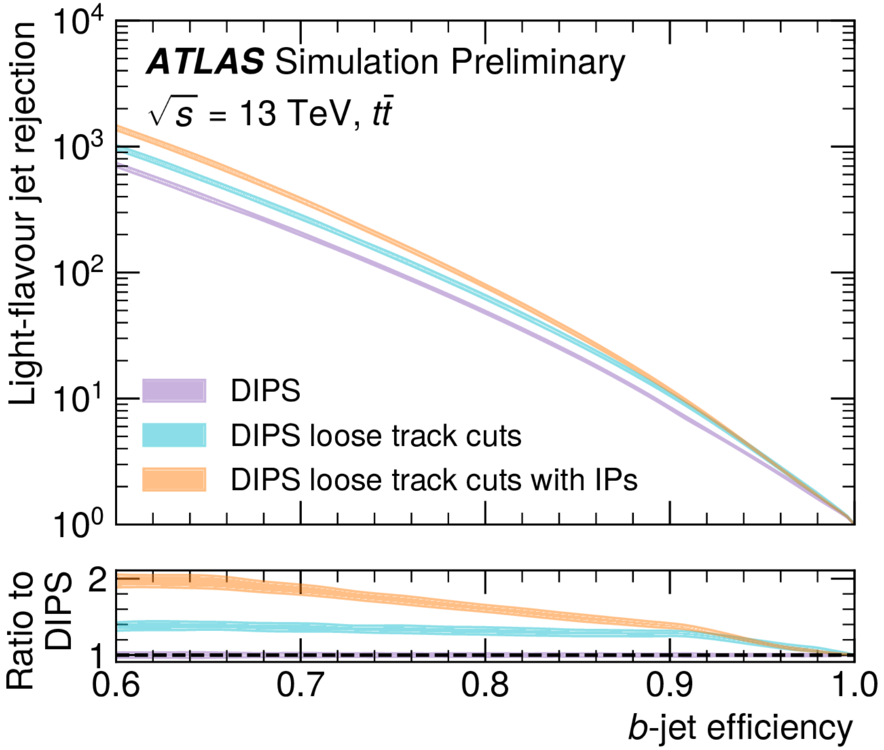
4x speed-up in the training time!!!

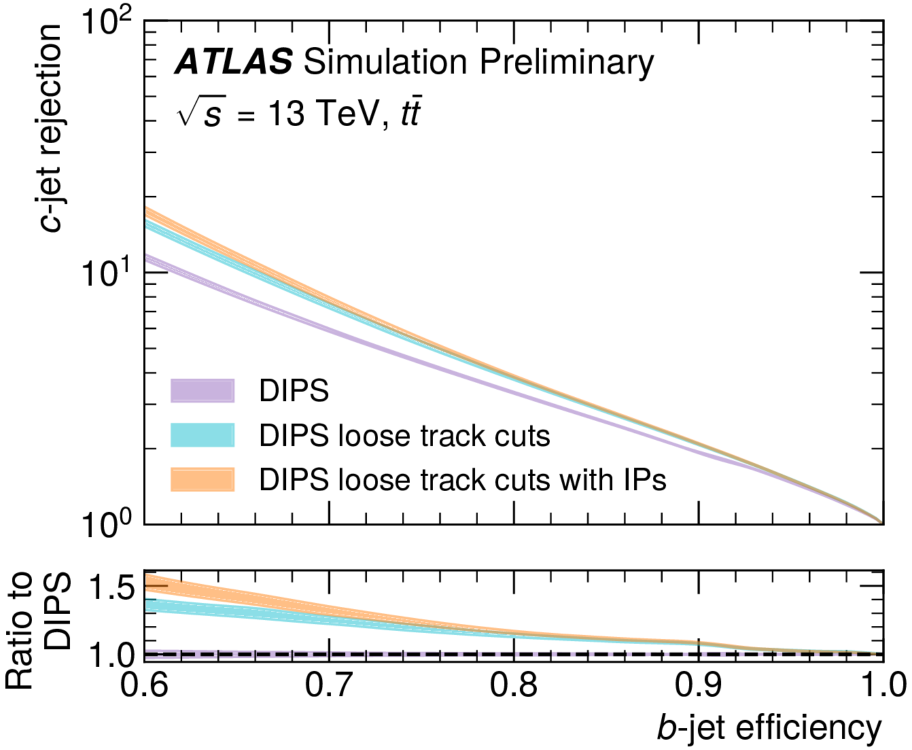
Faster turn around time for physics optimizations

nominal: from before (p_T > 1 GeV, $|d_0|$ < 1 mm, $|z_0 \sin\theta|$ < 1.5 mm)
loose: poorer quality (p_T > 1 GeV, $|d_0|$ < 1 mm, $|z_0 \sin\theta|$ < 1.5 mm)
Optimized DIPS: loose selection + new inputs (d_0, $z_0 \sin\theta$)



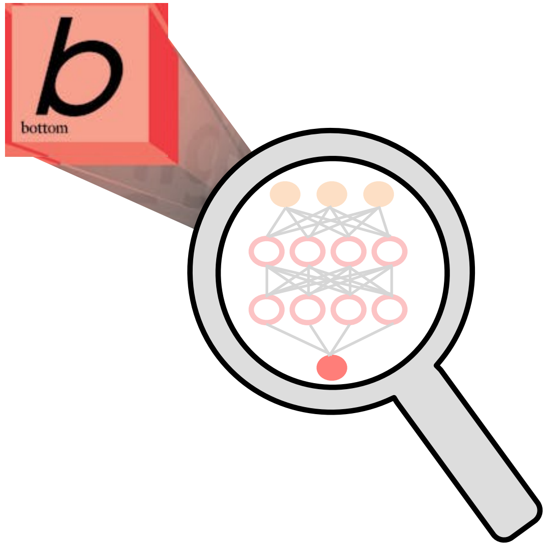
DIPS more performant for a high efficiency track selection





Slide design from
K. Leney

Future Outlook

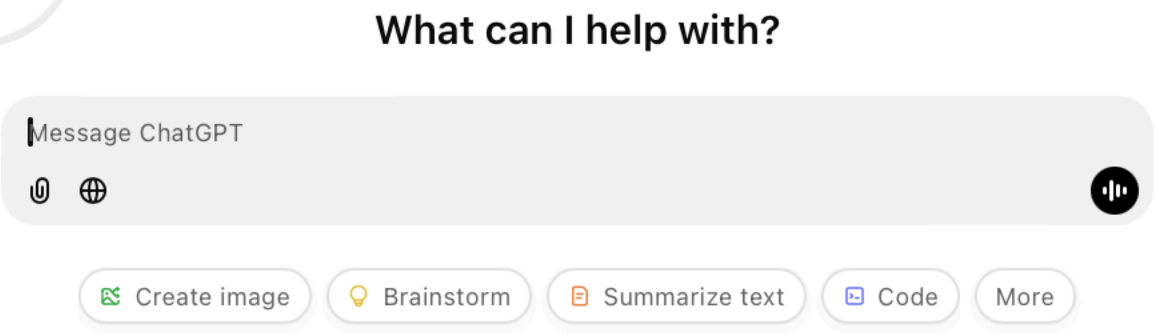


Deep Learning in FTAG

Transformer-era

- GN1
- Vertexing plots
- Performance in data
- Other similar applications
  - pT regression
  - Higgs tagging

Transformers are transforming our day-to-day lives!



- **Math**

how do you integrate a multidimensional gaussian with a source term

To integrate a multidimensional Gaussian with a source term, you are typically dealing with an integral of the form:

$$I = \int d^n\mathbf{x} \, \exp\left(-\frac{1}{2}\mathbf{x}^T A \mathbf{x} + \mathbf{J}^T \mathbf{x}\right)$$

where:

- $\mathbf{x}$ is an n-dimensional vector,
- A is an $n \times n$ symmetric positive-definite matrix (the covariance or precision matrix),
- $\mathbf{J}$ is an n-dimensional source term (linear term),
- $d^n\mathbf{x}$ is the measure for the n-dimensional integral.

This integral represents a Gaussian function in n-dimensions with a linear source term $\mathbf{J}$. The steps to solve this are outlined below:

Step 1: Completing the square

First, you need to complete the square in the exponent. The exponent is:

$$-\frac{1}{2}\mathbf{x}^T A \mathbf{x} + \mathbf{J}^T \mathbf{x}$$

You can rewrite it by factoring terms:

$$-\frac{1}{2}\left(\mathbf{x}^T A \mathbf{x} - 2\mathbf{J}^T \mathbf{x}\right)$$

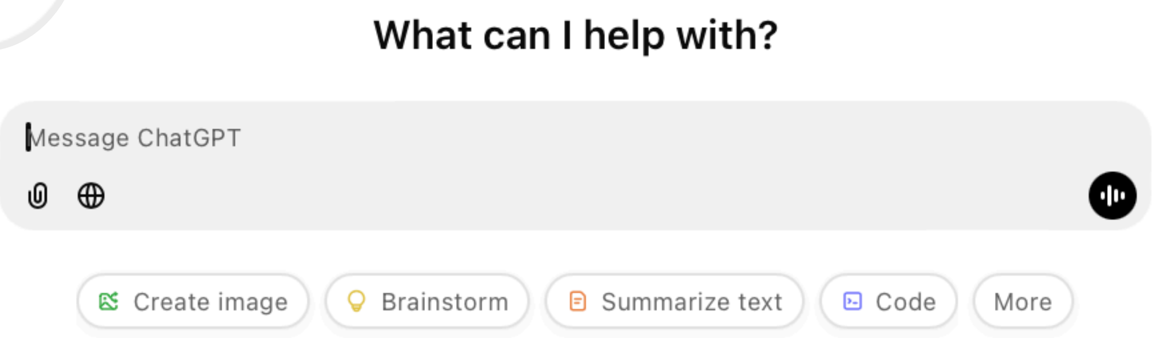
Next, complete the square by introducing a new shifted variable $\mathbf{y} = \mathbf{x} - A^{-1}\mathbf{J}$, which removes the linear term from the exponent.

$$-\frac{1}{2}\left(\mathbf{x} - A^{-1}\mathbf{J}\right)^T A \left(\mathbf{x} - A^{-1}\mathbf{J}\right) + \frac{1}{2}\mathbf{J}^T A^{-1}\mathbf{J}$$

Now, the exponent is:

$$-\frac{1}{2}\mathbf{y}^T \cdots + \frac{1}{2}\mathbf{J}^T A^{-1}\mathbf{J}$$

Transformers are transforming our day-to-day lives!



- Math
- **Poetry**

Write me a song for the ATLAS flavour tagging group.

(Verse)
Impact parameter, a tiny shift,
Lifetime of beauty, a precious gift.
Secondary vertices, a hidden clue,
Decays revealing what particles flew.
Machine learning, a powerful friend,
Boosting performance, till the very end.
(Chorus)
Flavour tagging, that's our game,
Sorting through the chaos, calling out the name.
b-jets, c-jets, light quark too,
Reconstructing truth in all we do.
Algorithms humming, a digital art,
Unraveling secrets, playing our part.

Curtesy of Diptaparna
* Gemini 1.5 Pro 002

Transformers are transforming our day-to-day lives!

What can I help with?

Message ChatGPT

Create image | Brainstorm | Summarize text | Code | More

Could you draw me a 16x9 image of Siegen in impressionist style

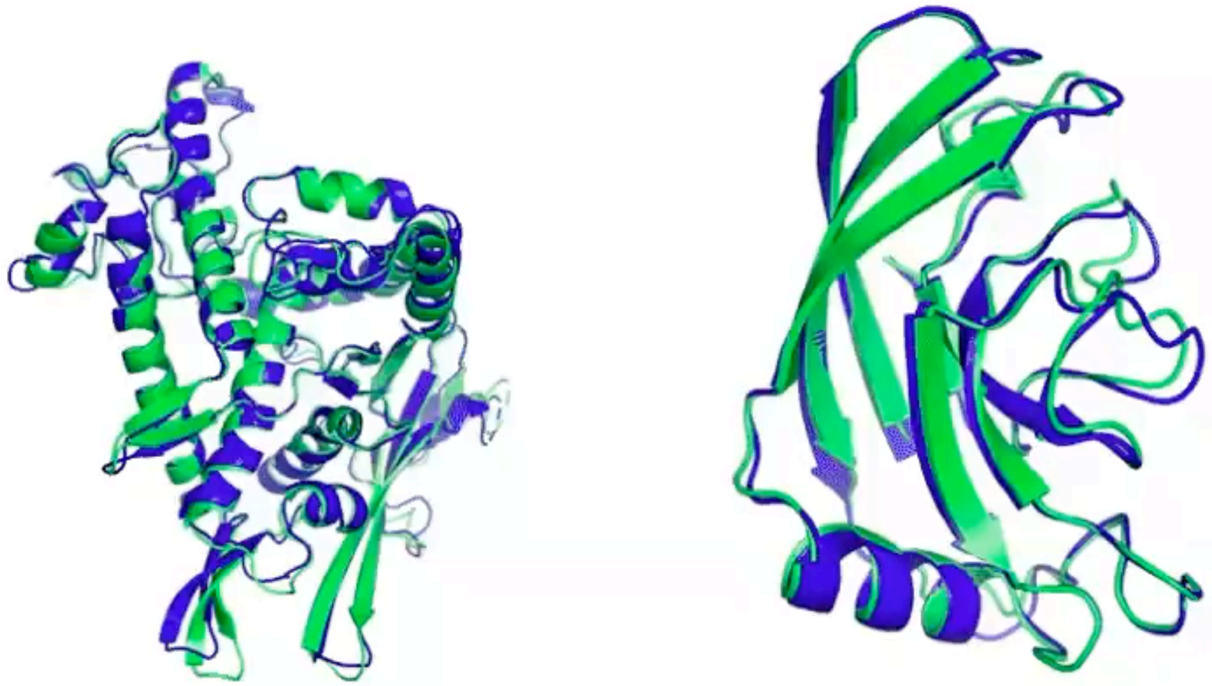
- Math
- Poetry
- **Art**

Here is an impressionist-style illustration of Siegen, Germany, showcasing its landmarks and natural beauty with soft, textured brushstrokes. Let me know if you need further adjustments!

Transformers for collecting your Nobel prize

Chemistry, 2024

AlphaFold2



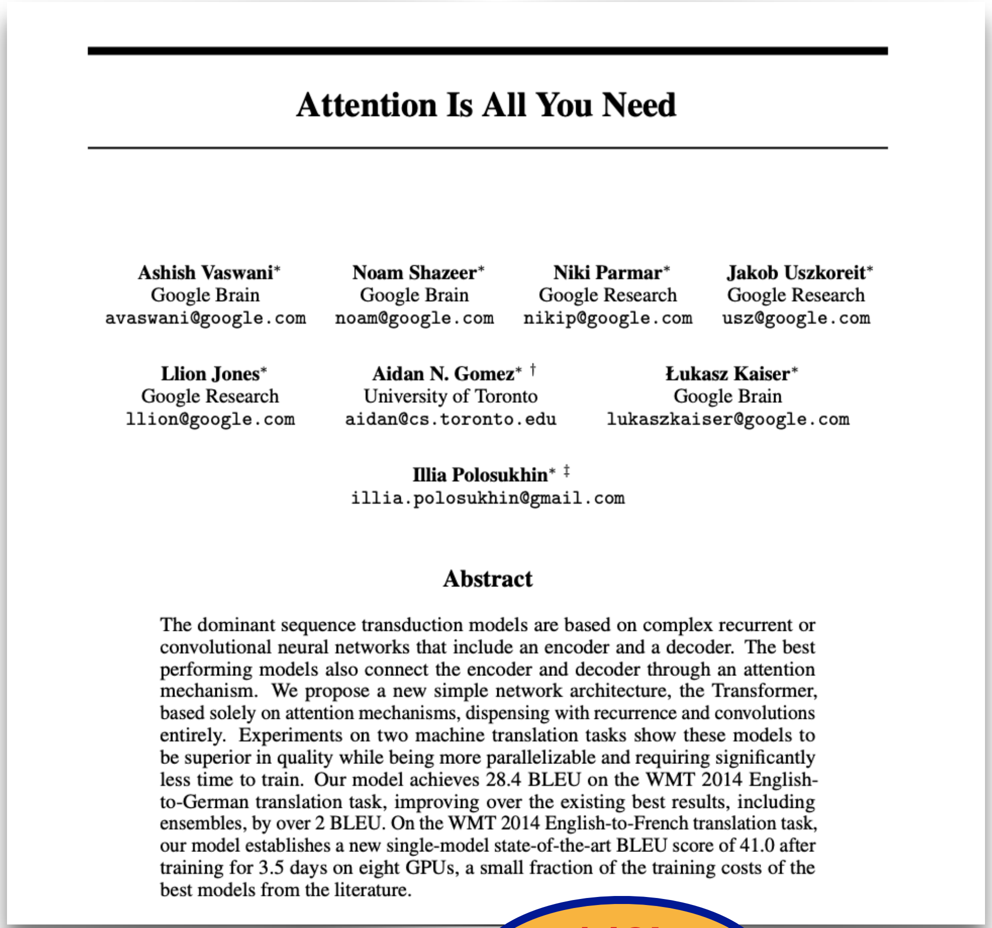
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

- Experimental result
- Computational prediction

1706.03762

Transformers 101



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

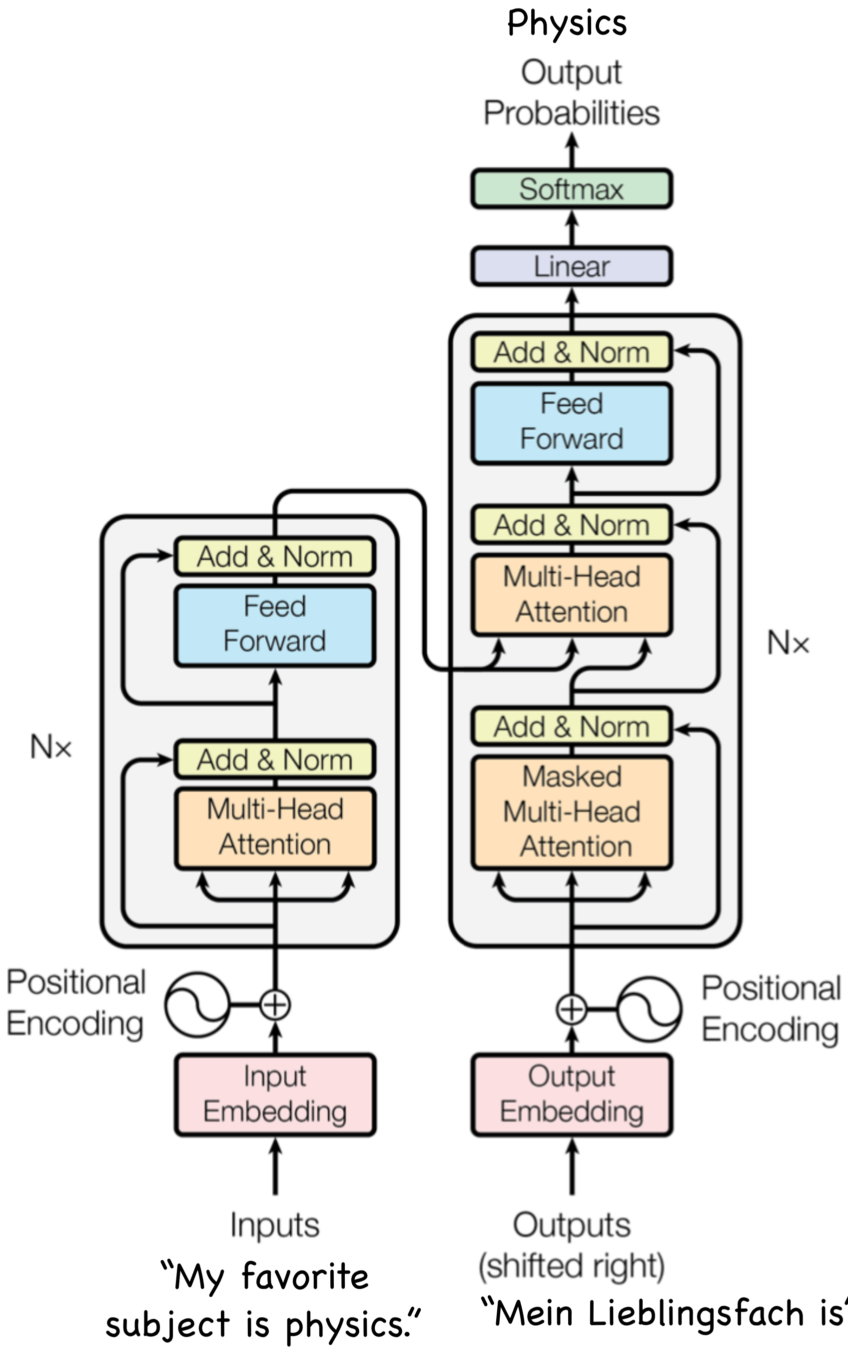
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

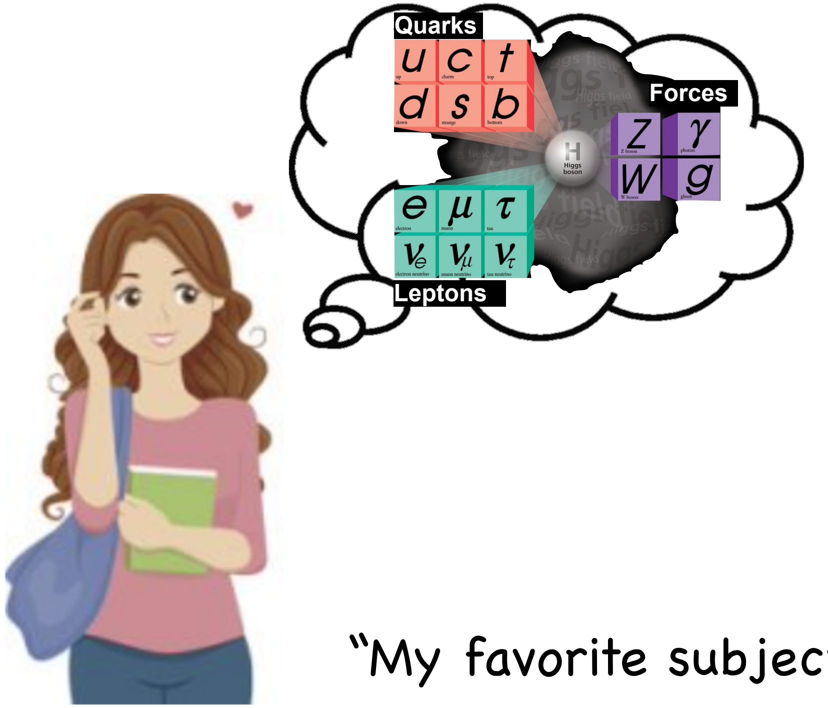
142k citations



[1706.03762](https://arxiv.org/abs/1706.03762)

Transformers 101

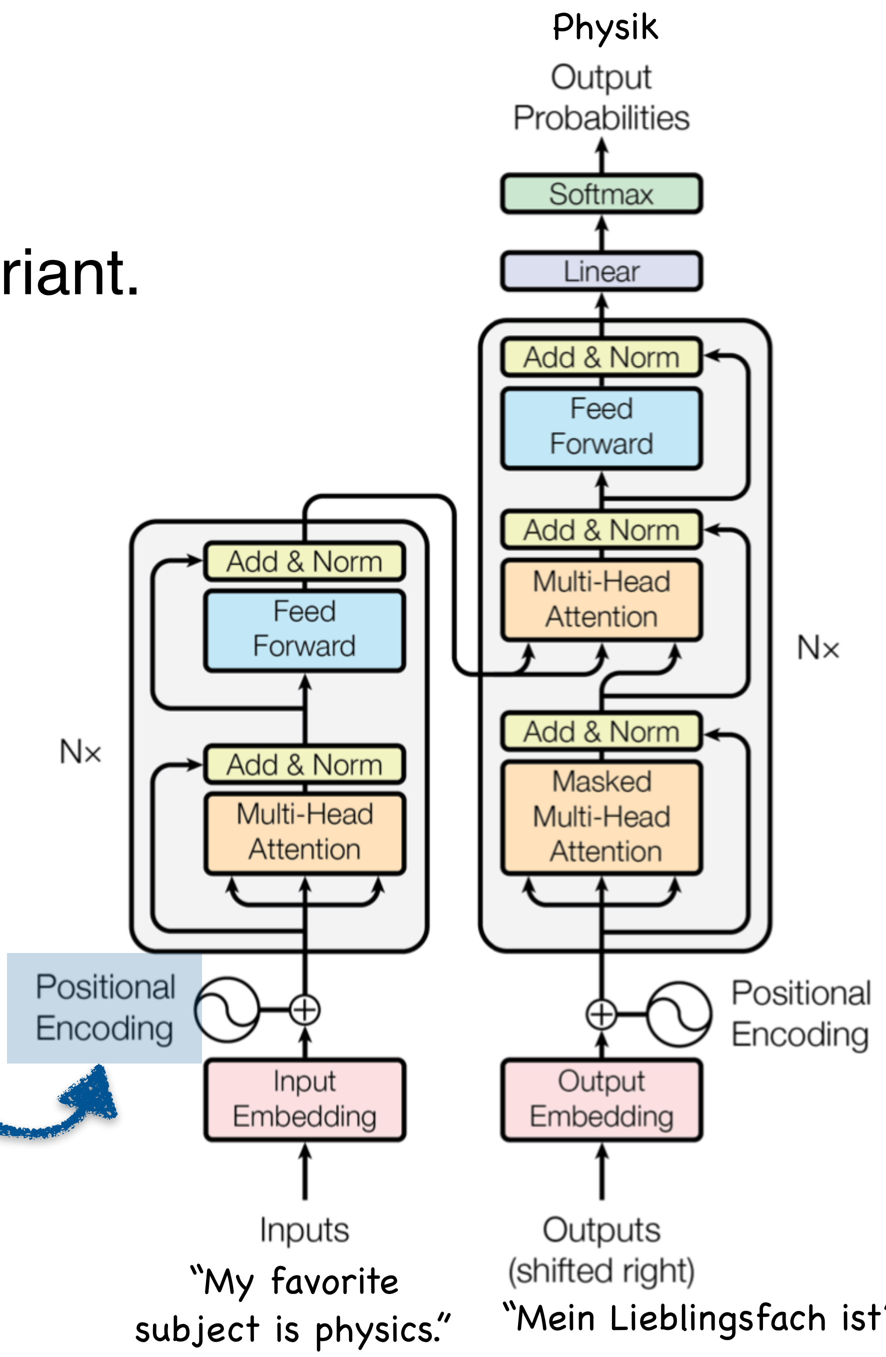
Architecture natively permutation invariant.



"My favorite subject is physics."

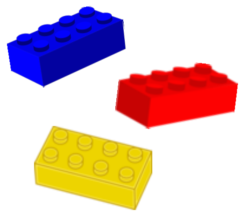
<START> 0 1 2 3 4 <END>

Sequence information encoded via an additional input



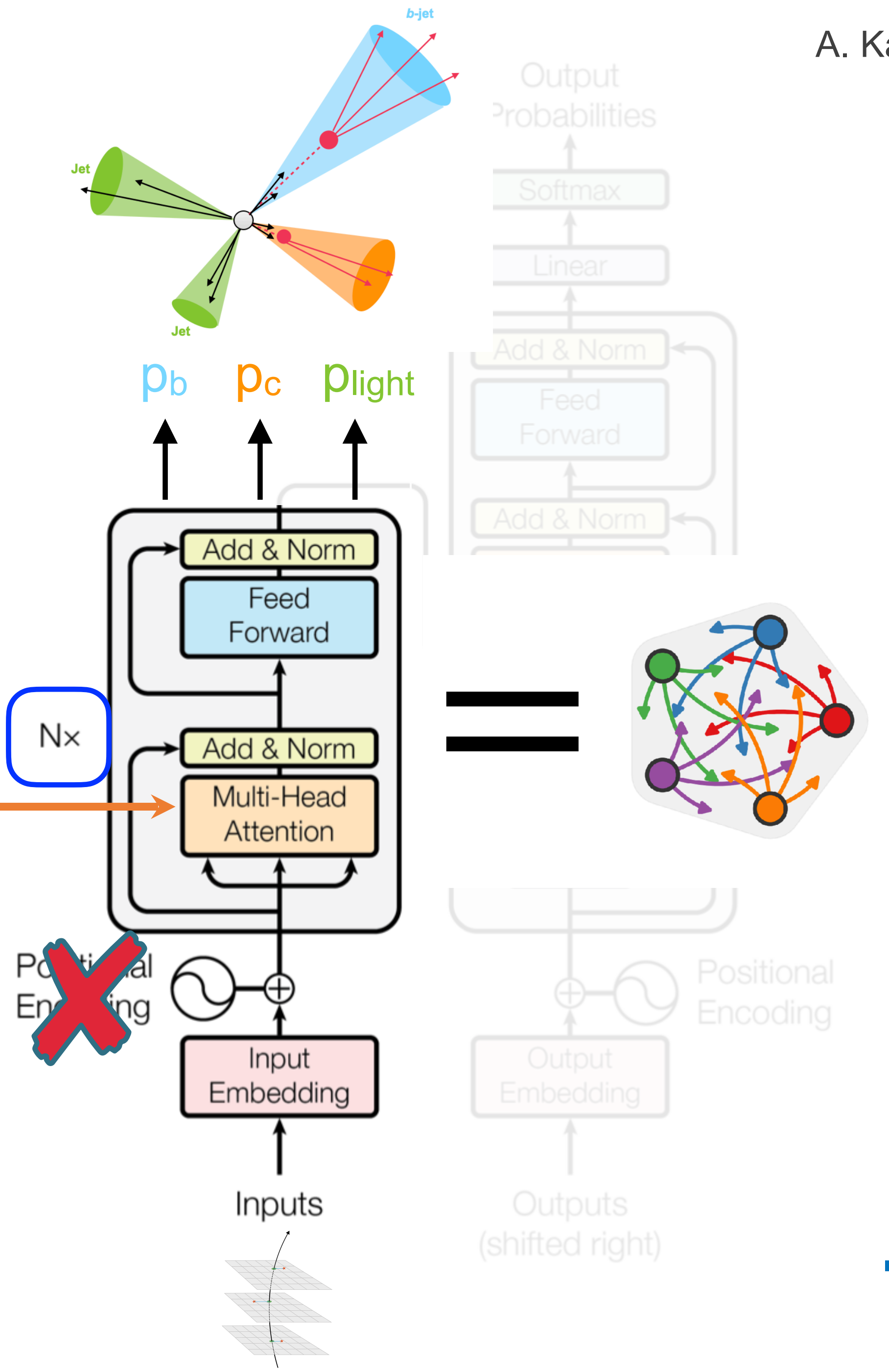
[1706.03762](1706.03762)
A. Karpathy's lecture

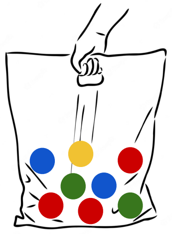
Transformers for FTAG



And do this lots of times!

Replace the Deep Sets Sum with a weighted sum (still permutation invariant)



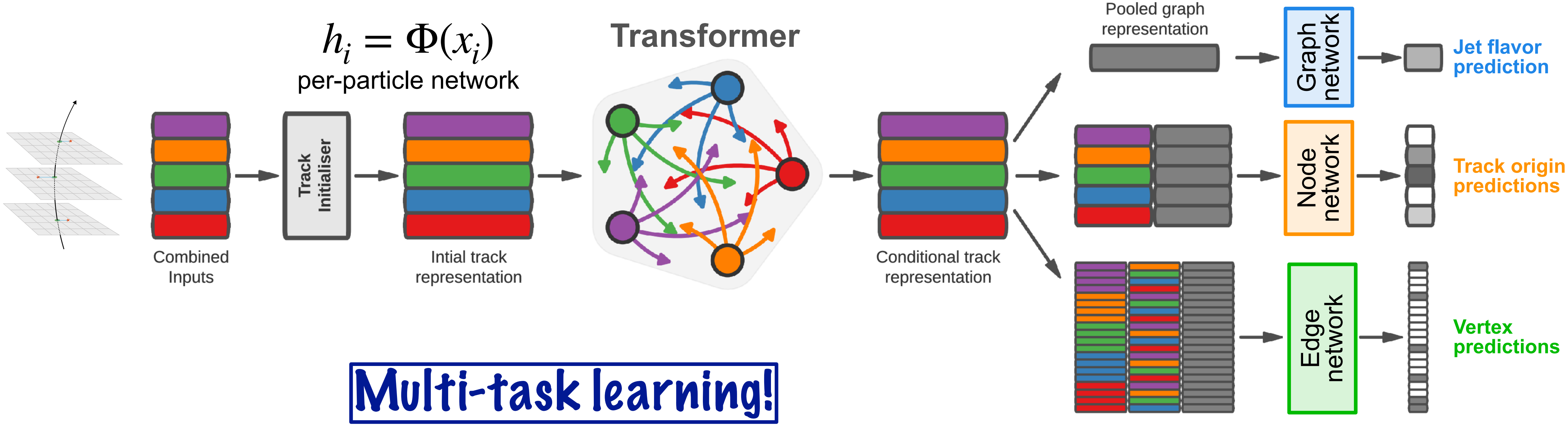


No need for positional encoding
(permutation invariance over the tracks)

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

GN2: architecture

Transformer-based flavour tagger

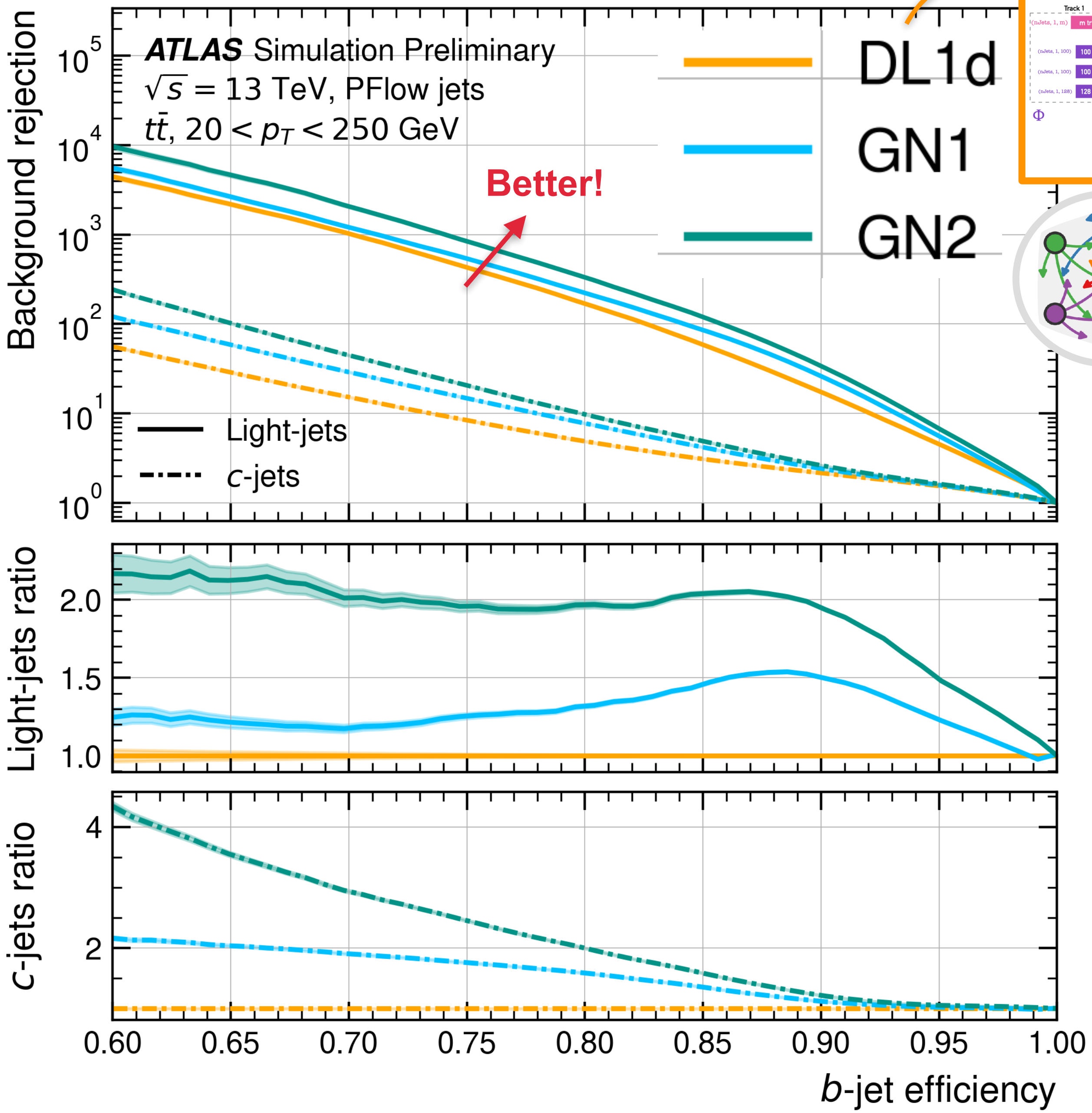


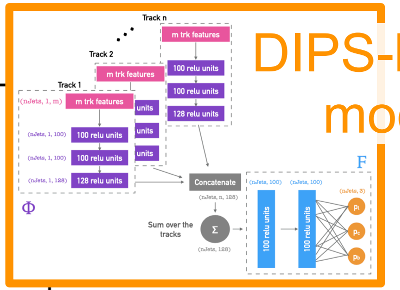
Multi-task learning!

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

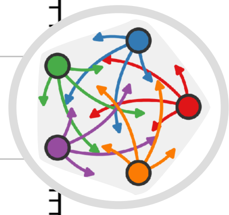
$\alpha = 0.5$ $\beta = 1.5$

GN1: graph neural network
GN2: transformer





DIPS-based model

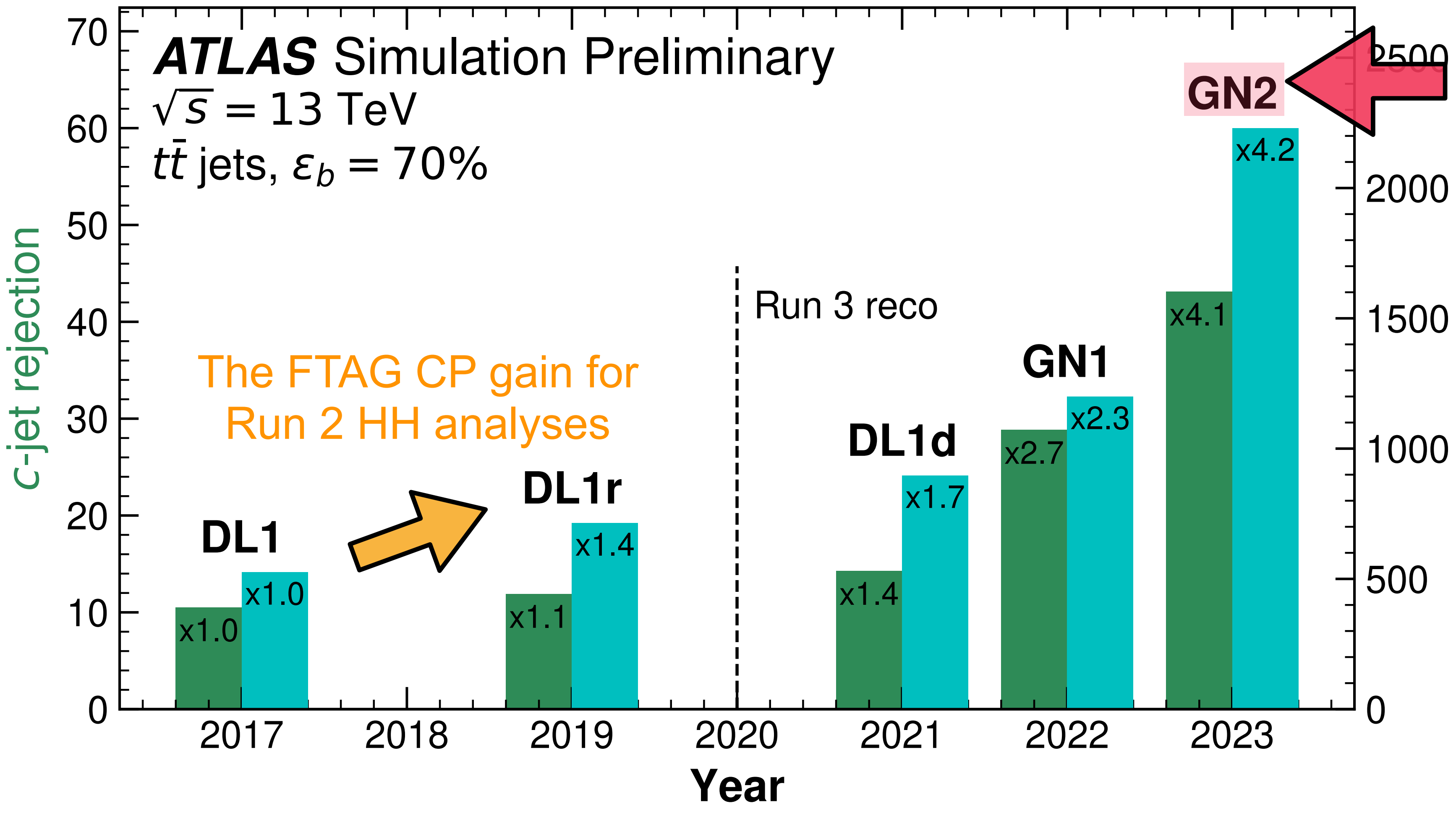


transformer model

2x increase in light rejection

4x increase in charm rejection

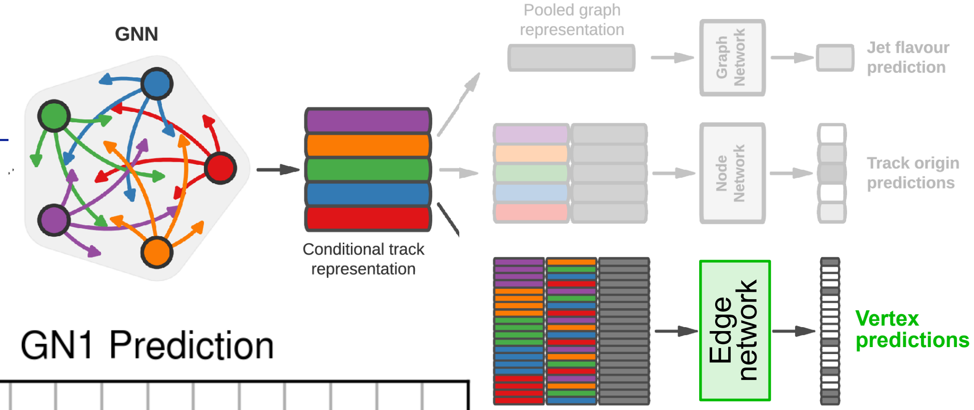
FTAG over time

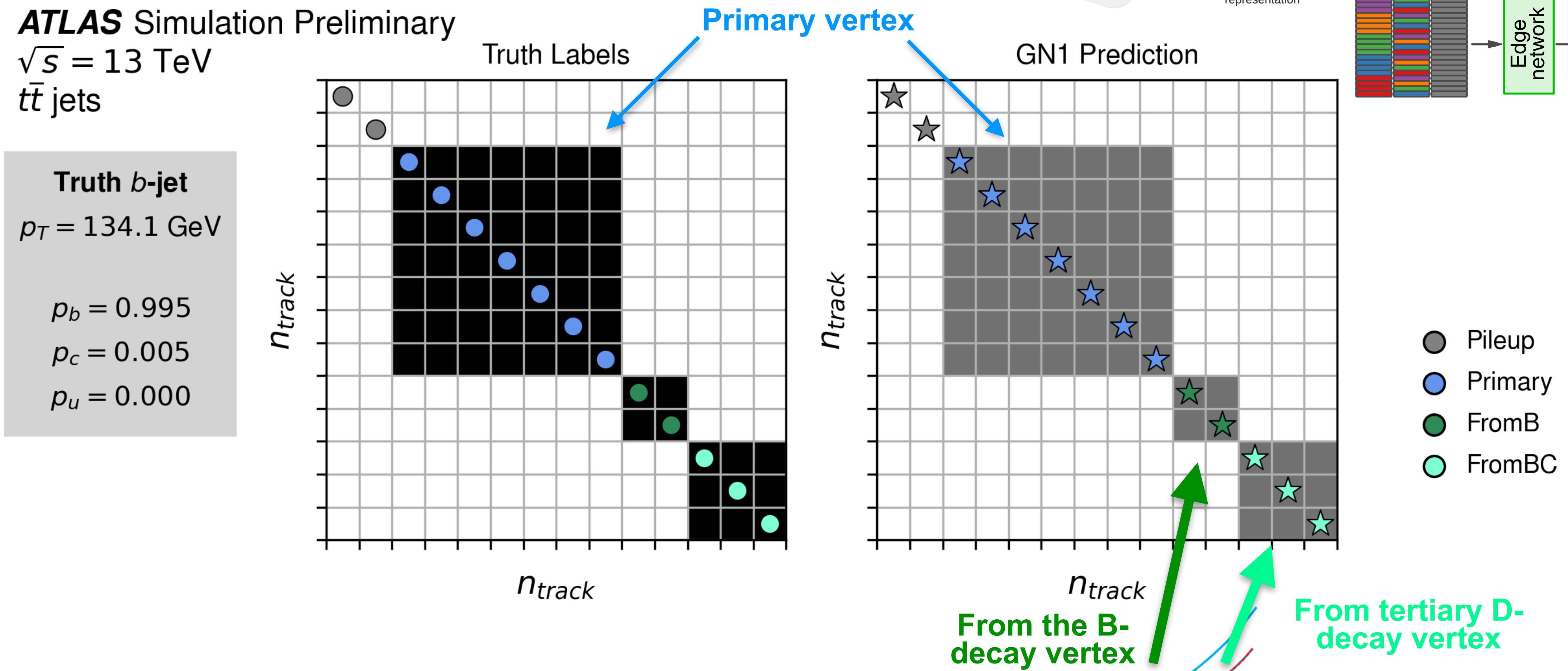


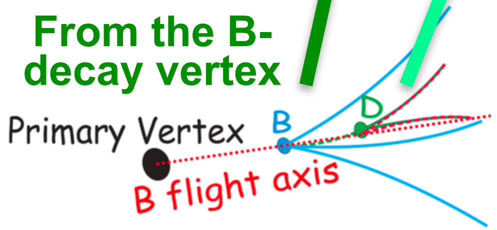
Transformer-based tagger
$\mathcal{O}(10)$ more training jets
[300m 😱]

Vertex finding

Group tracks with pair-wise compatibility > 0.5

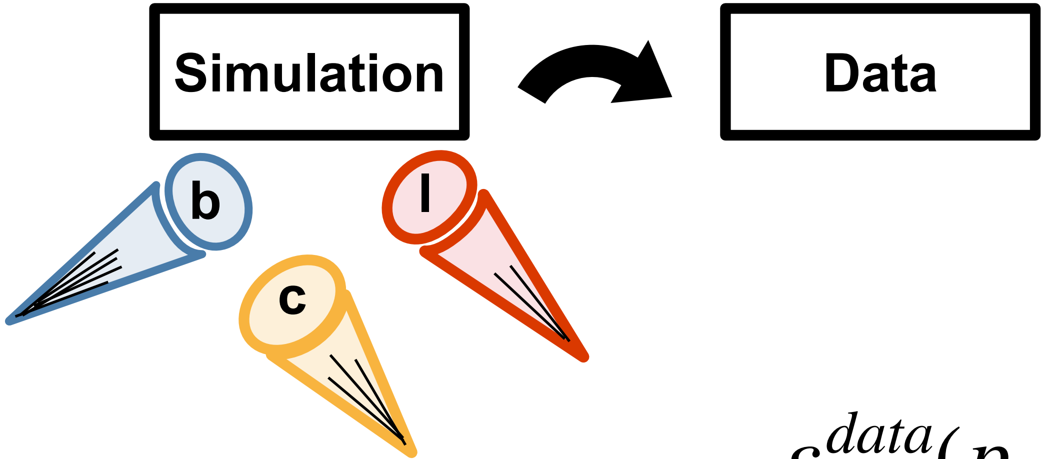






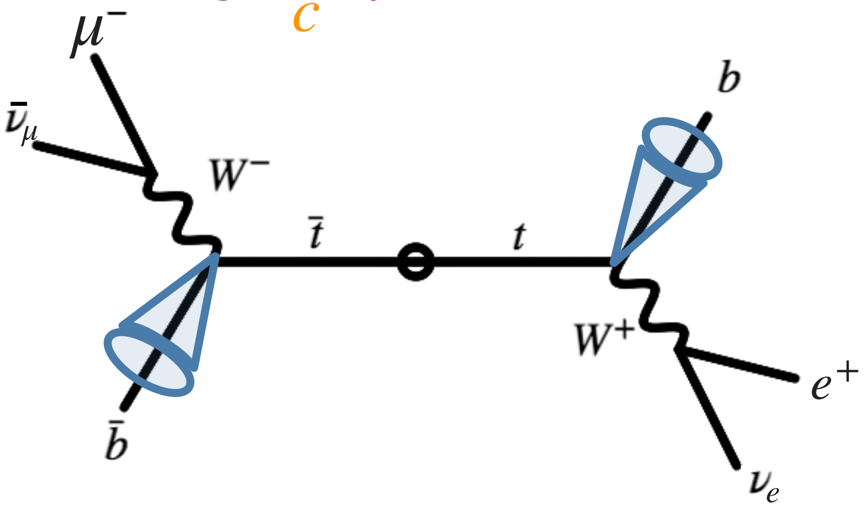
 FTAG-2023-01

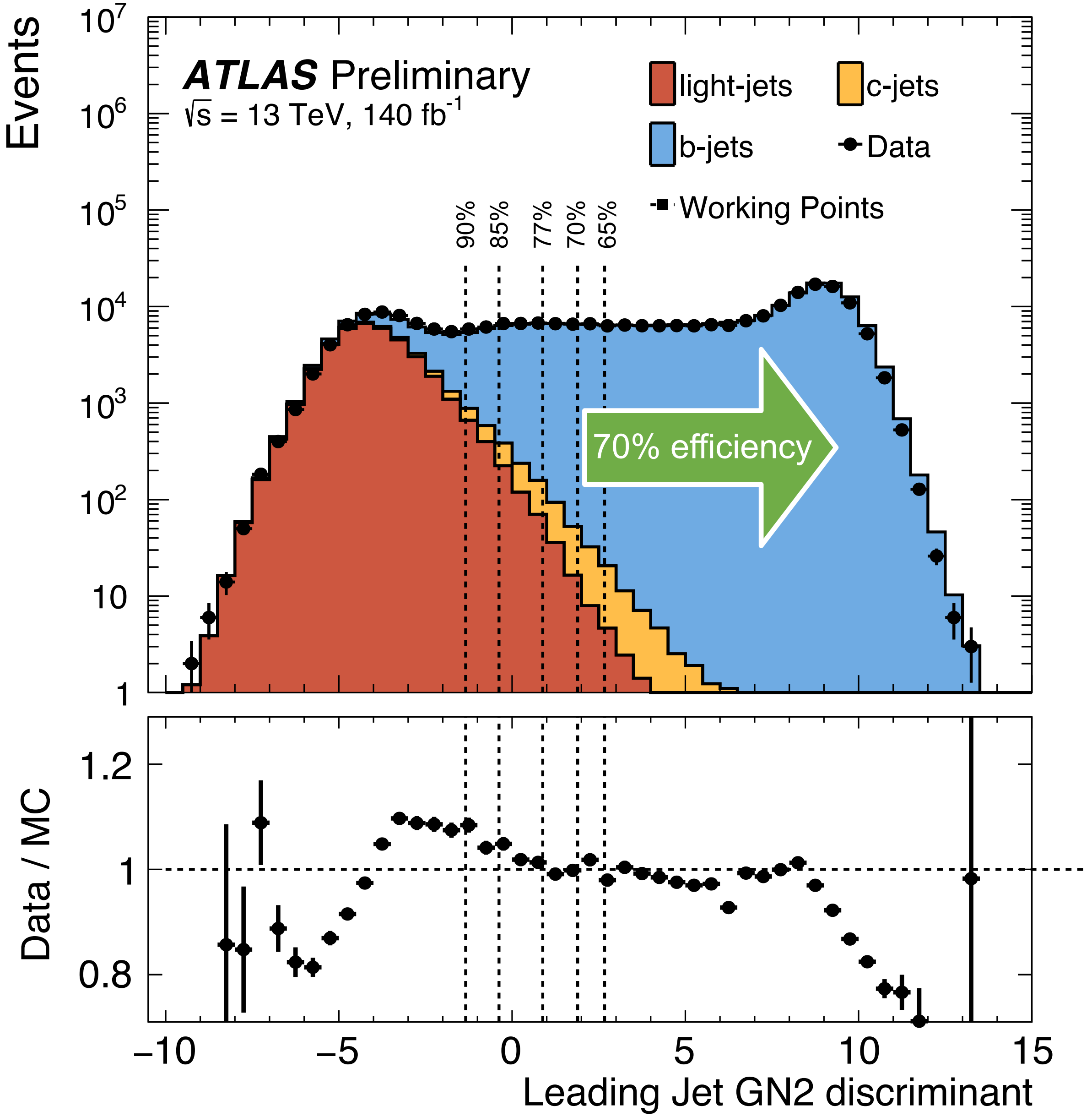
Calibration



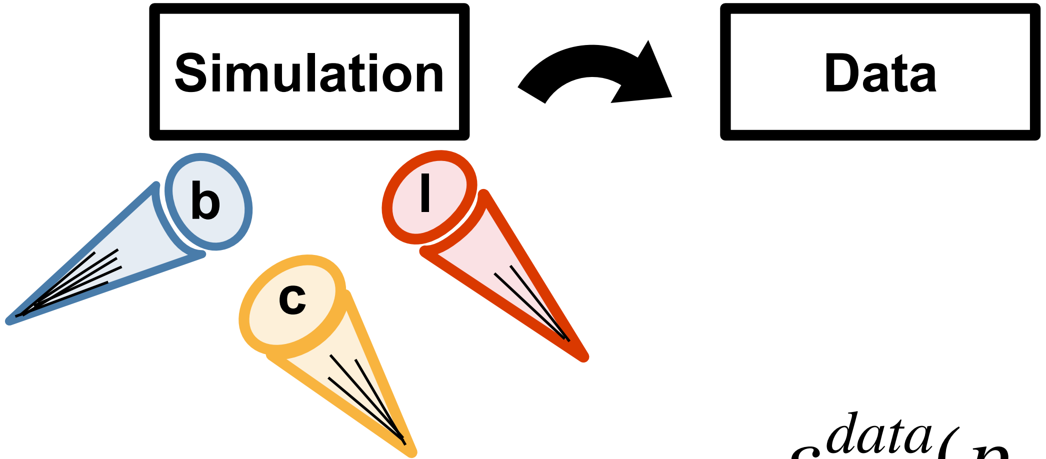
$$SF_f(p_T) = \frac{\varepsilon_f^{data}(p_T)}{\varepsilon_f^{MC}(p_T)}$$

$f \rightarrow b, c, l$



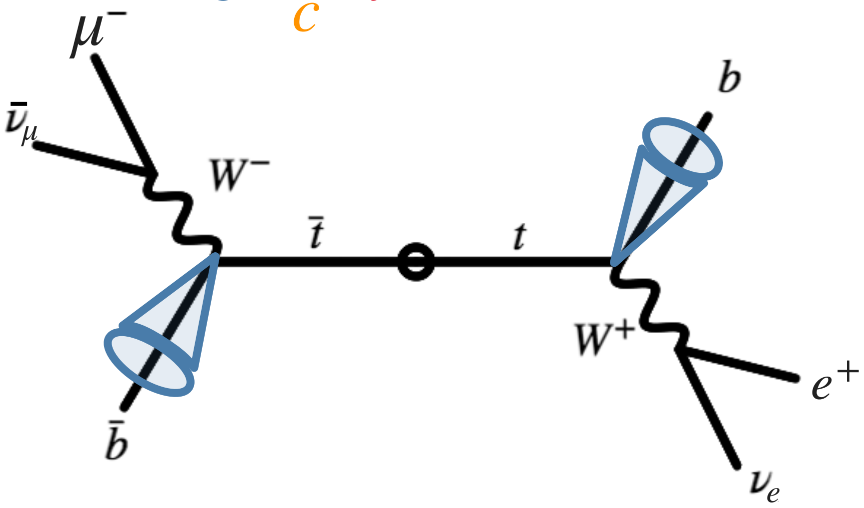


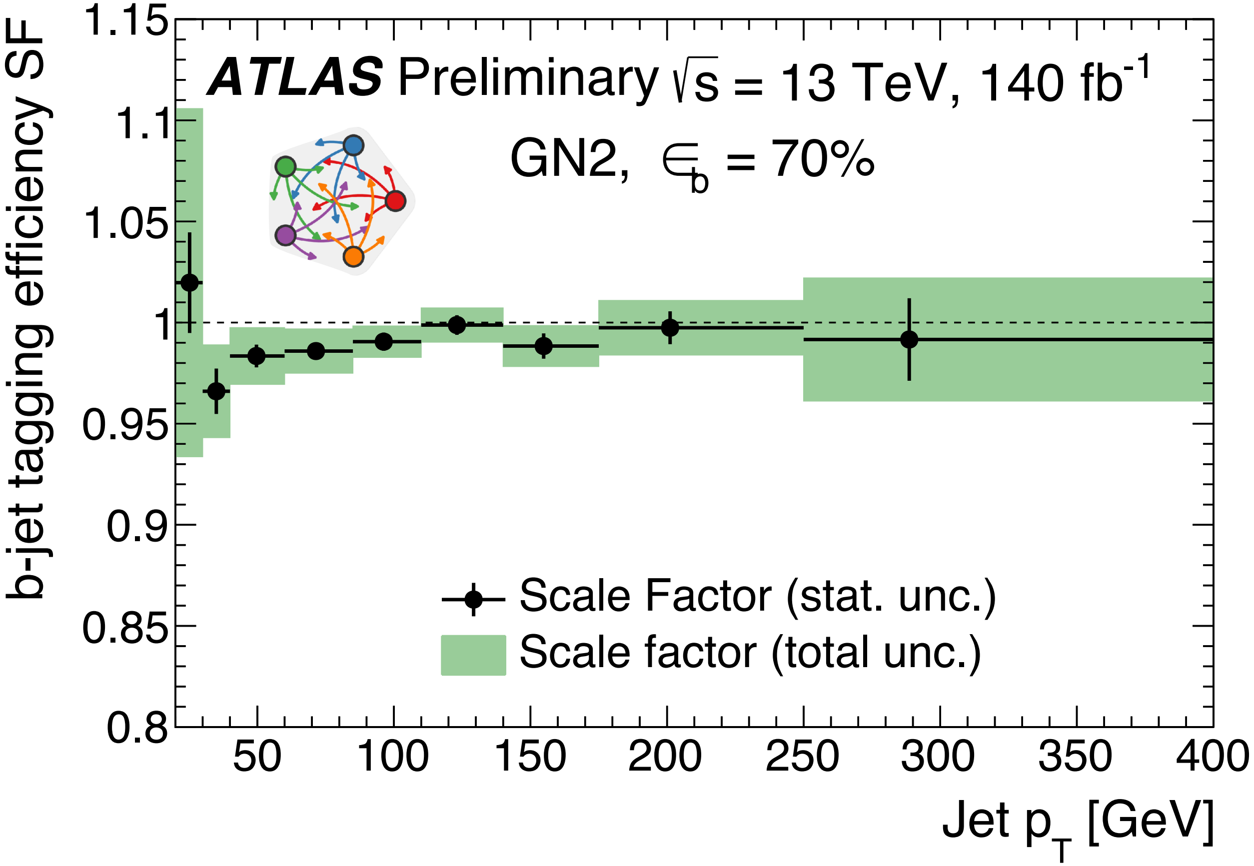
Calibration



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

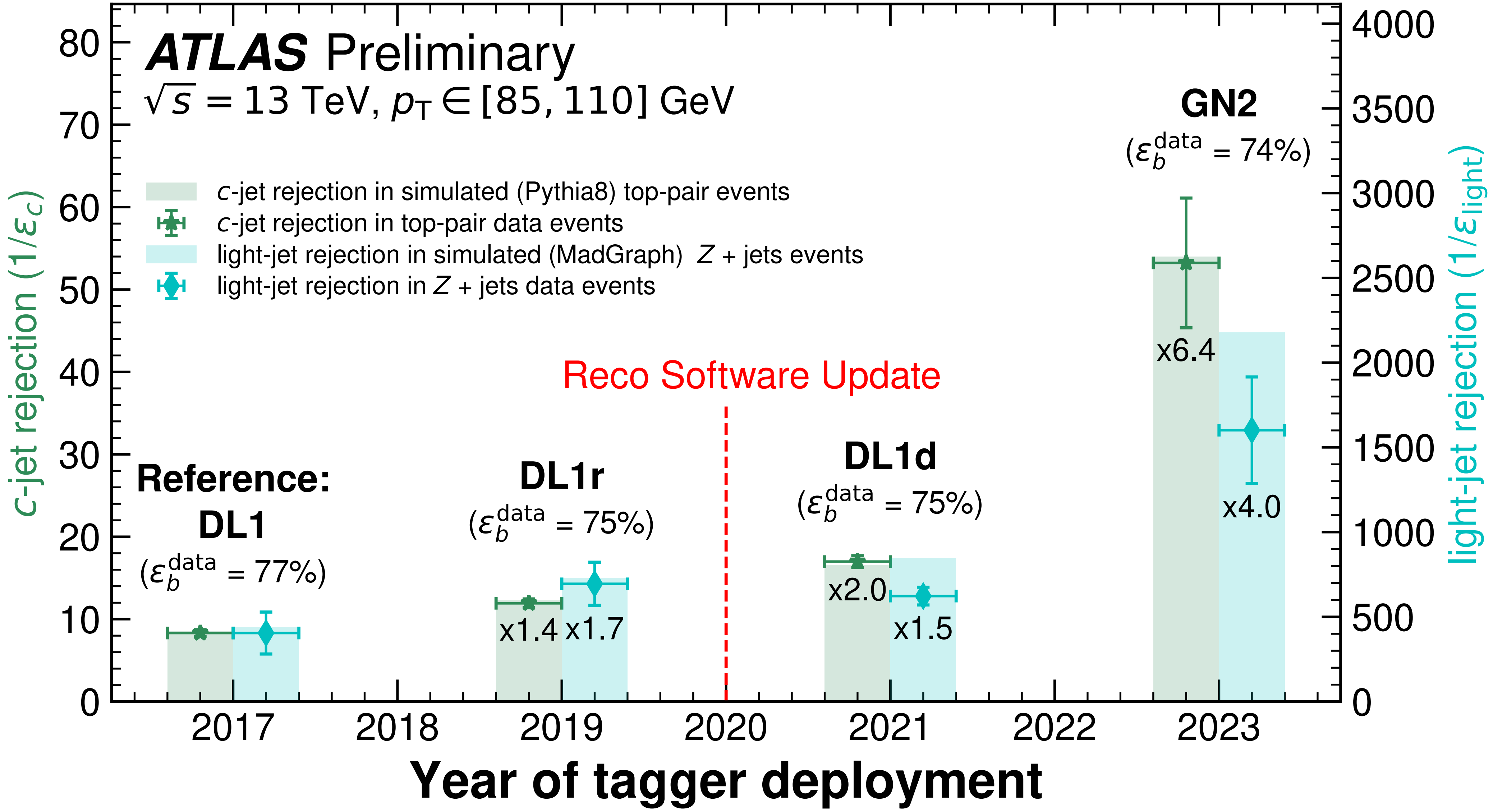
$b \quad c \quad l$






% level non-closure

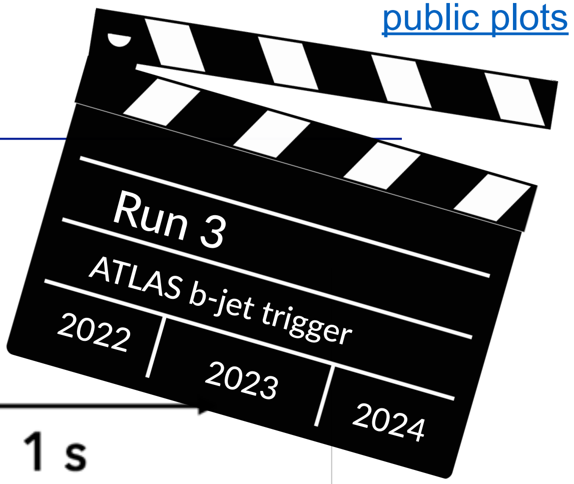
And… it translates to the physics (!)

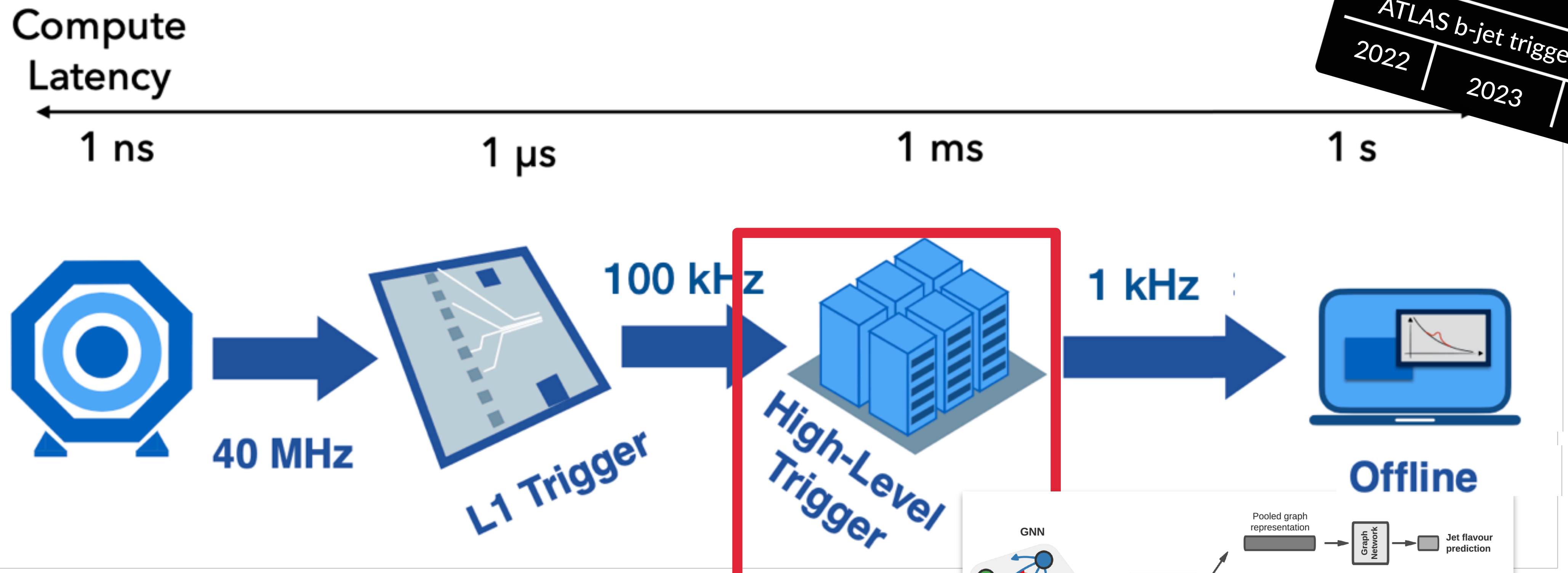


b-jet trigger
public plots

Impacting the physics...

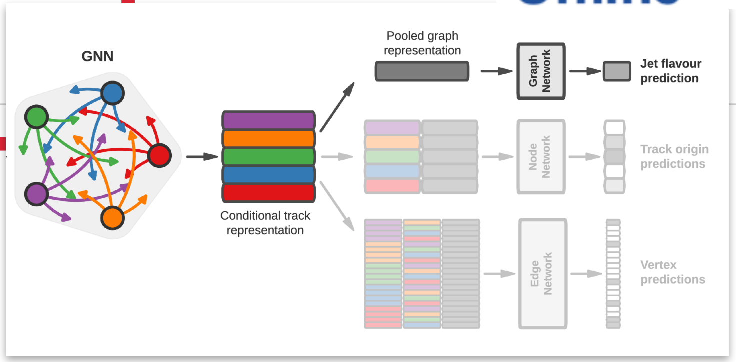
Which events we save?





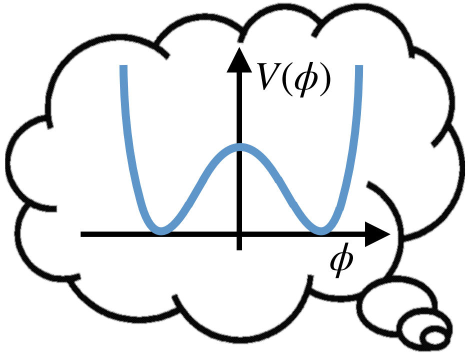
2022: DL1d
2023: GN1
2024: GN2

New b-taggers in trigger each year of Run 3!

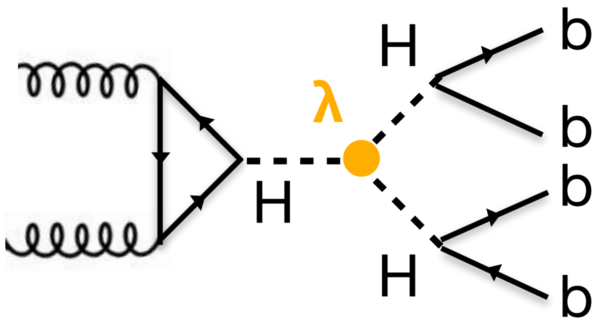


b-jet trigger public plots

FTAG in trigger



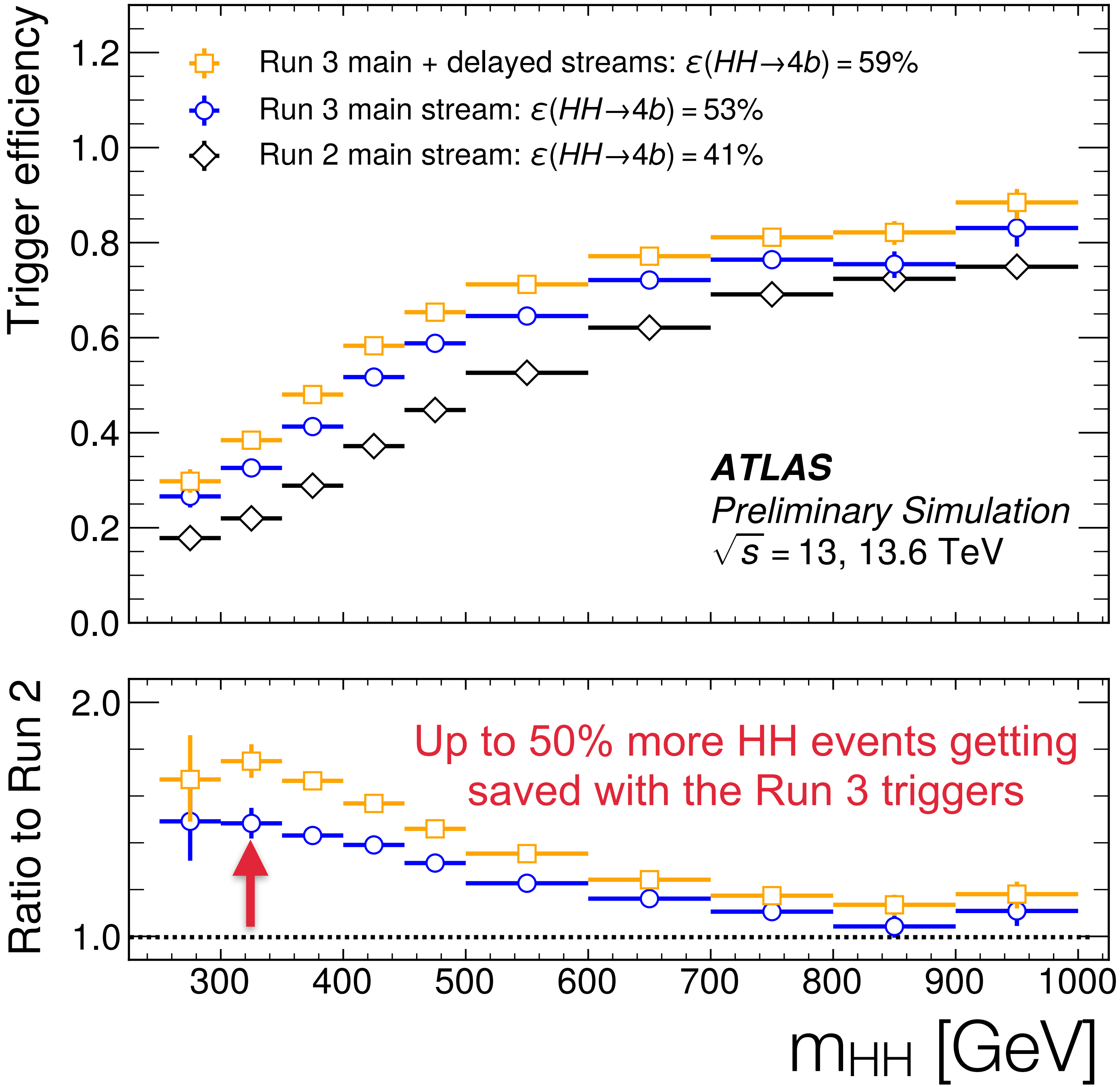
Ex: HH4b efficiency



2022: DL1d

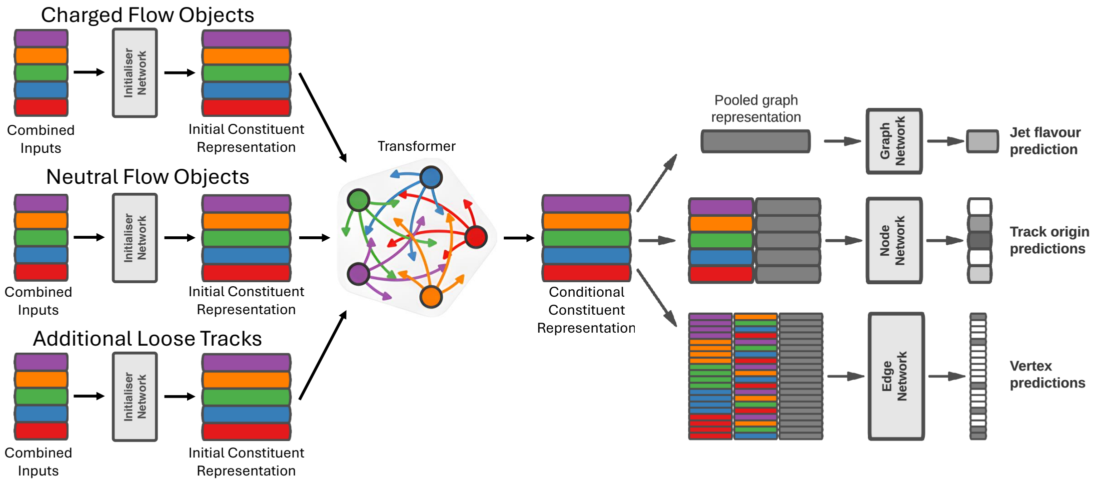
2023: GN1

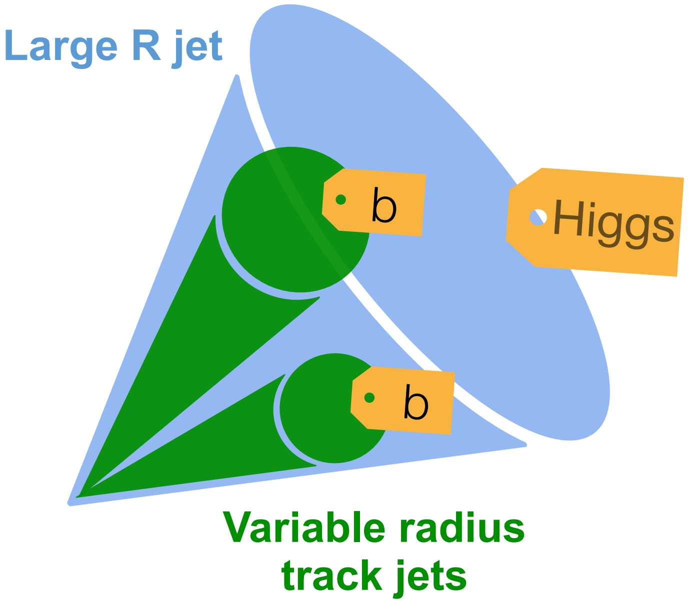
2024: GN2

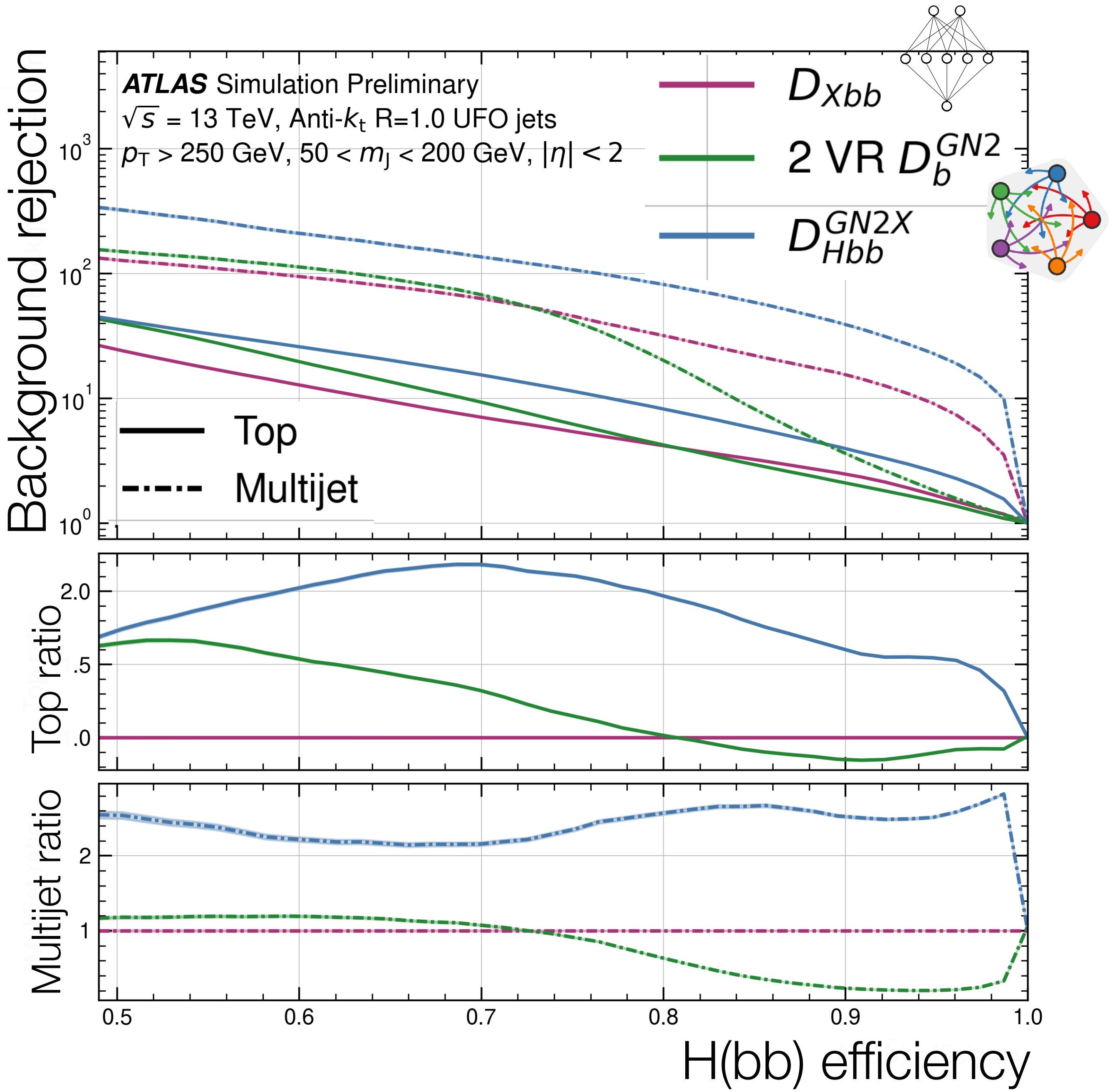


ATL-PHYS-PUB-2023-021

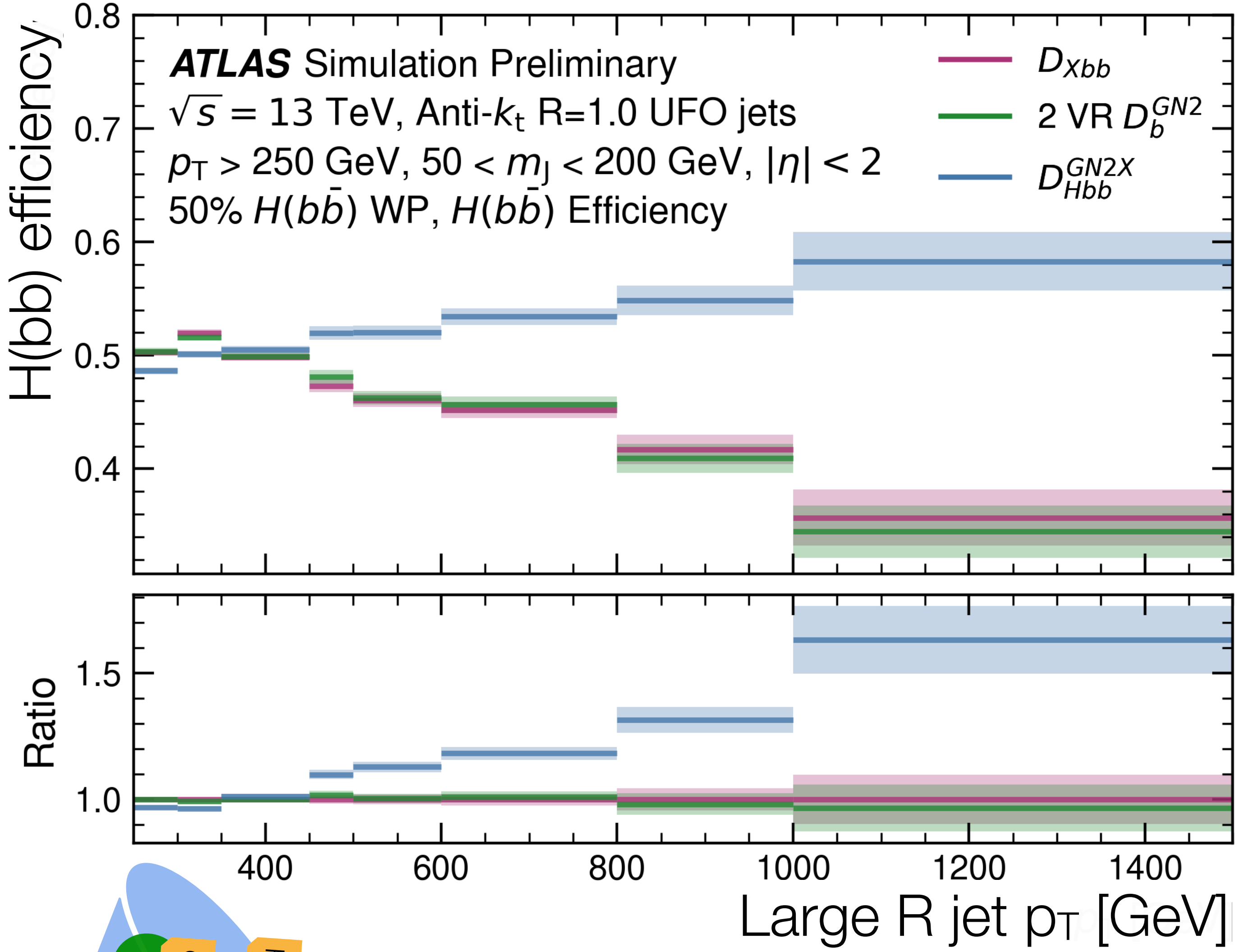
Versatile: Xbb tagging

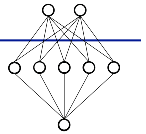




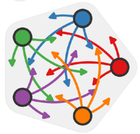


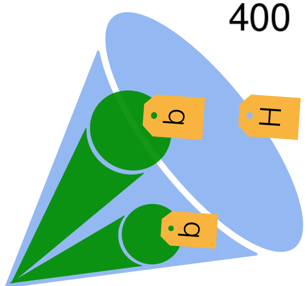
Versatile: Xbb tagging



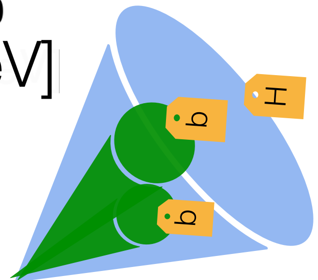
baseline NN 

tagging subjets 

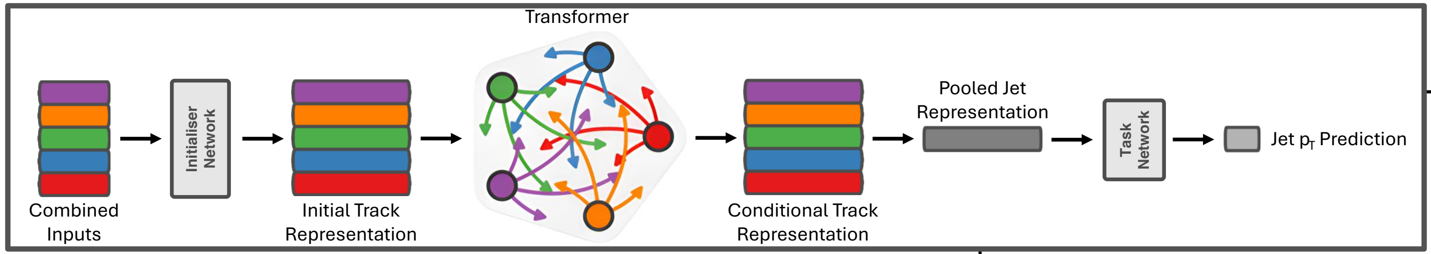
large-R jet transformer 

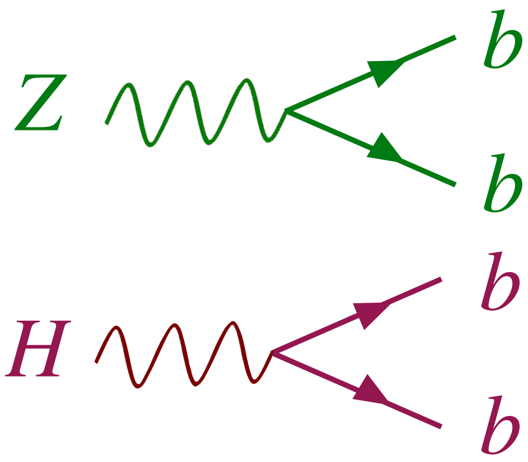


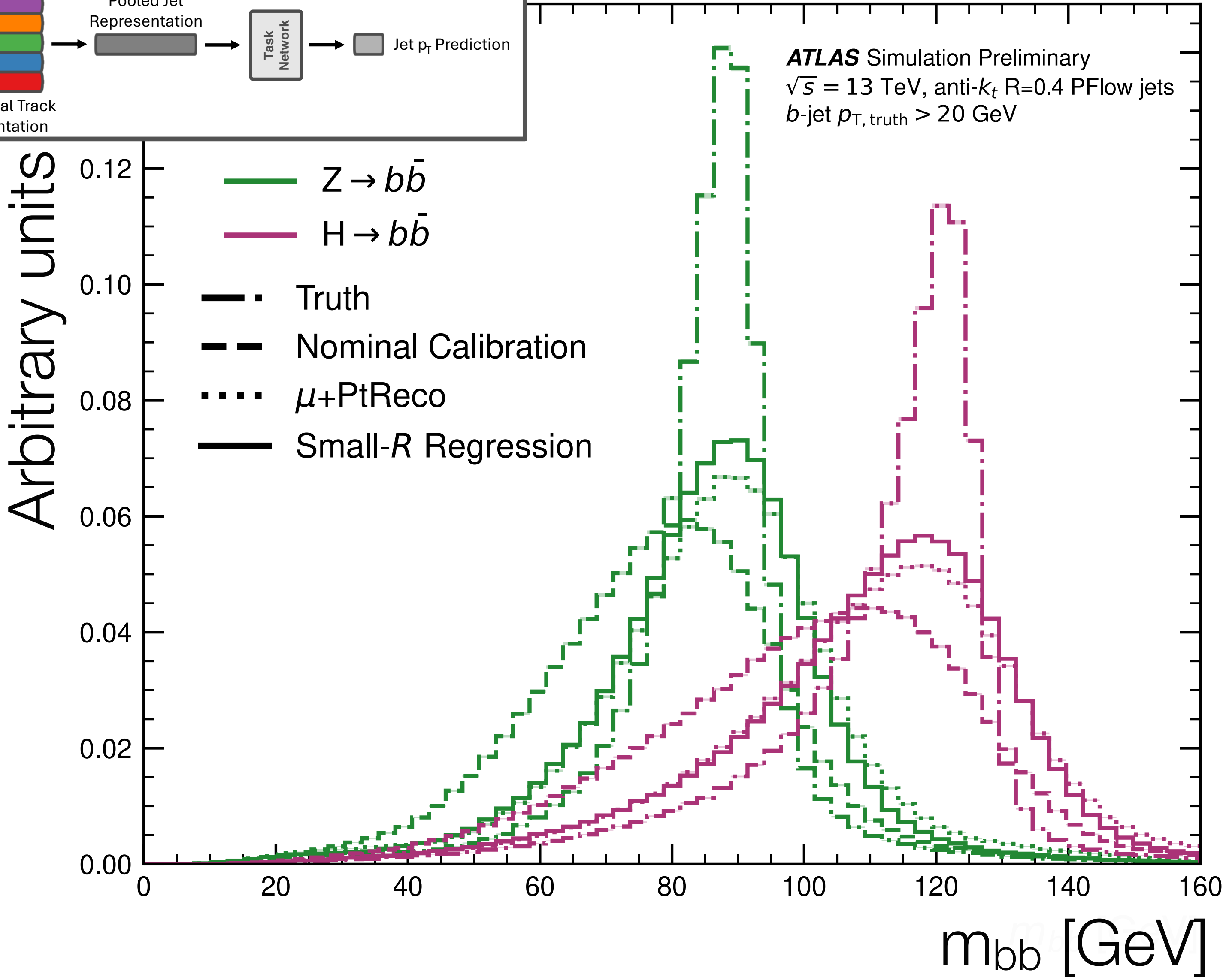
Increasing Higgs p_T



Versatile: regression



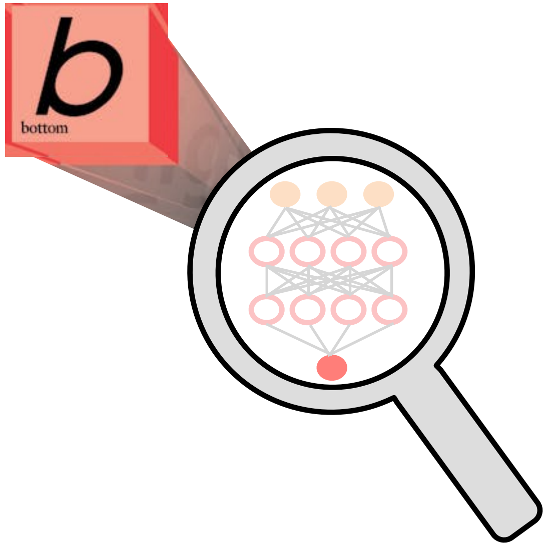




Slide design from
K. Leney

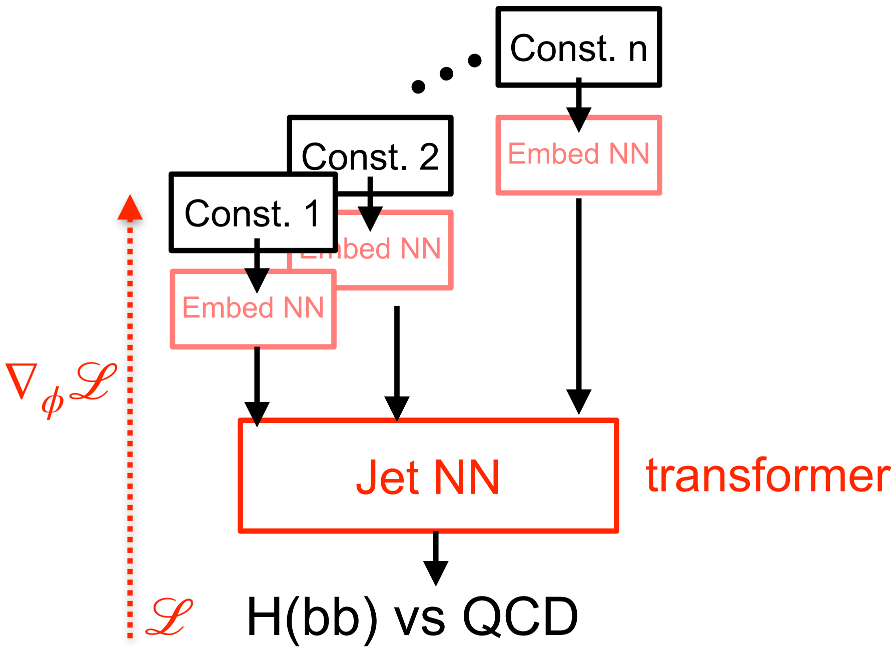
Future Outlook

- End-to-end optimization
- Use with novel calibrations

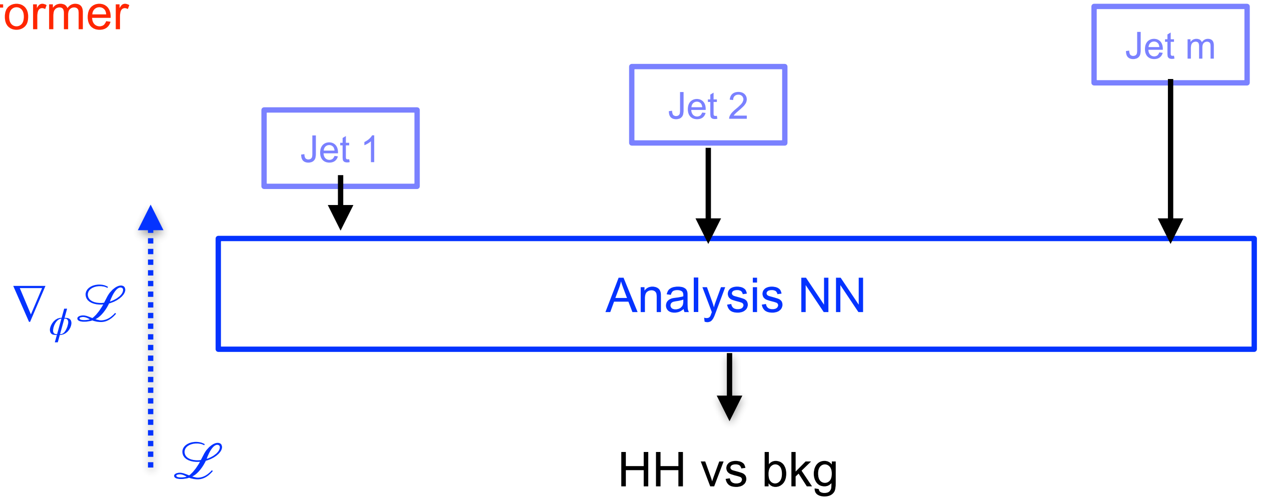


Traditional Analysis

Step 1: Train the tagger

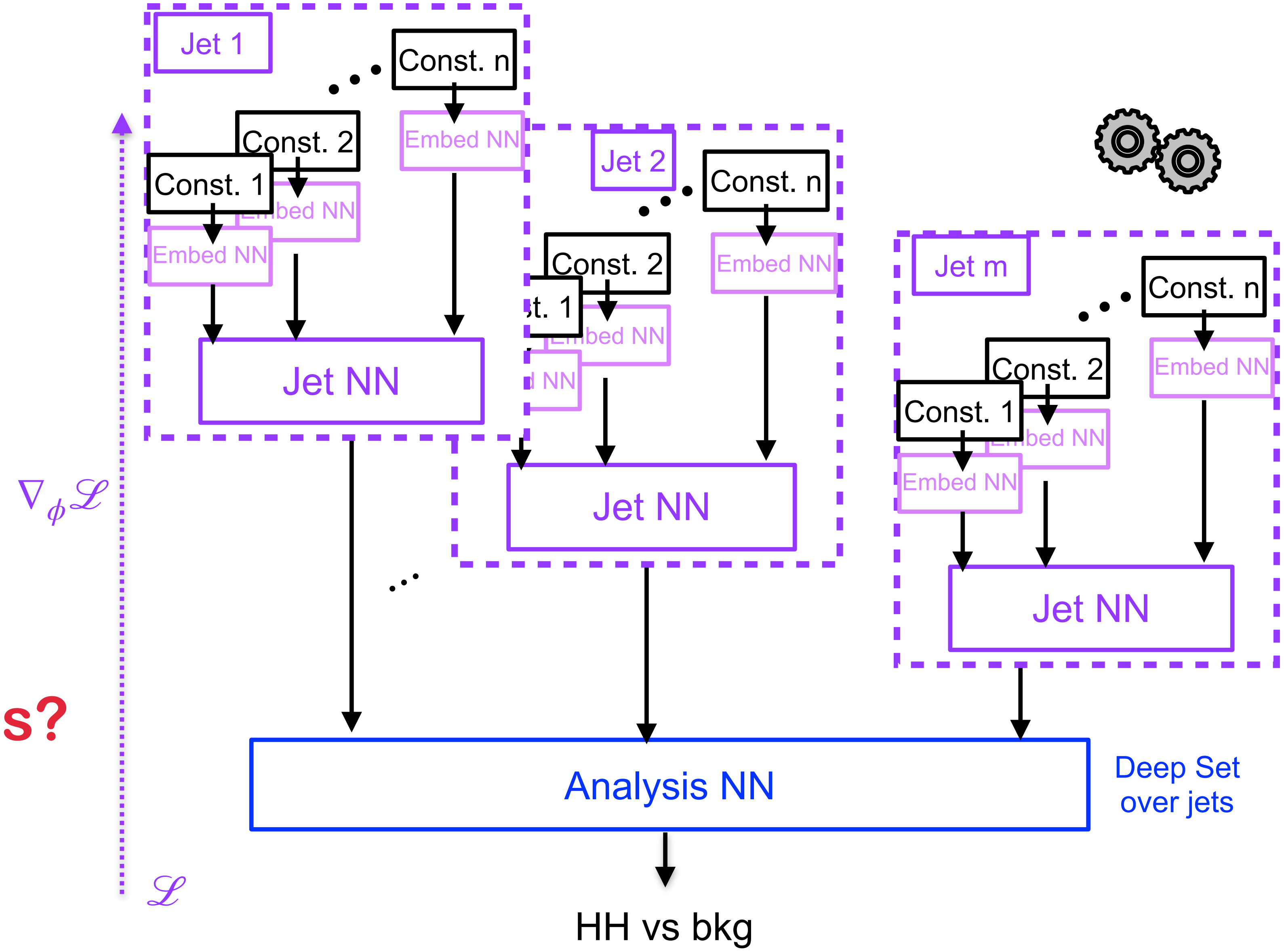


Step 2: Optimize the analysis



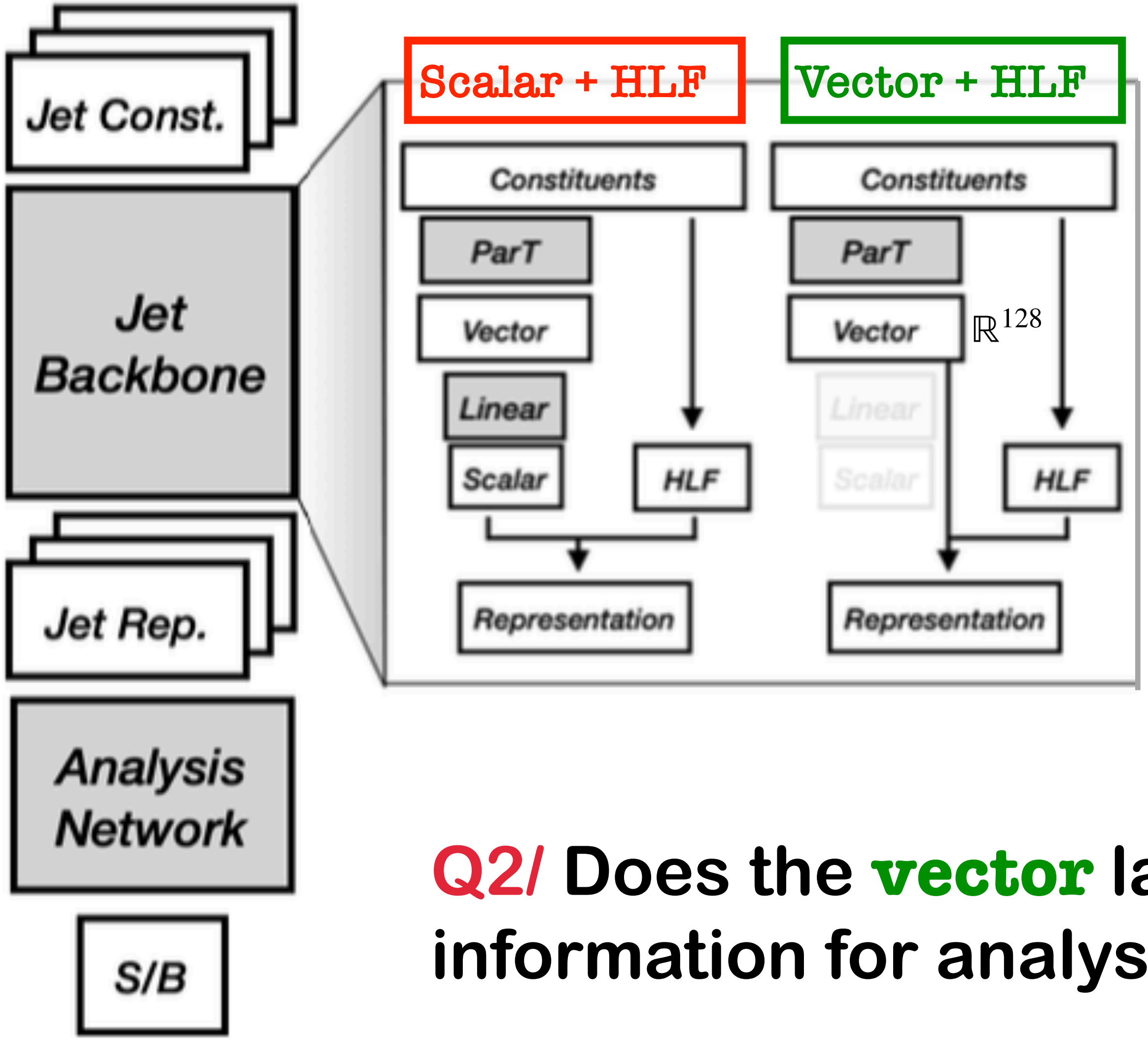
New paradigm

End to end



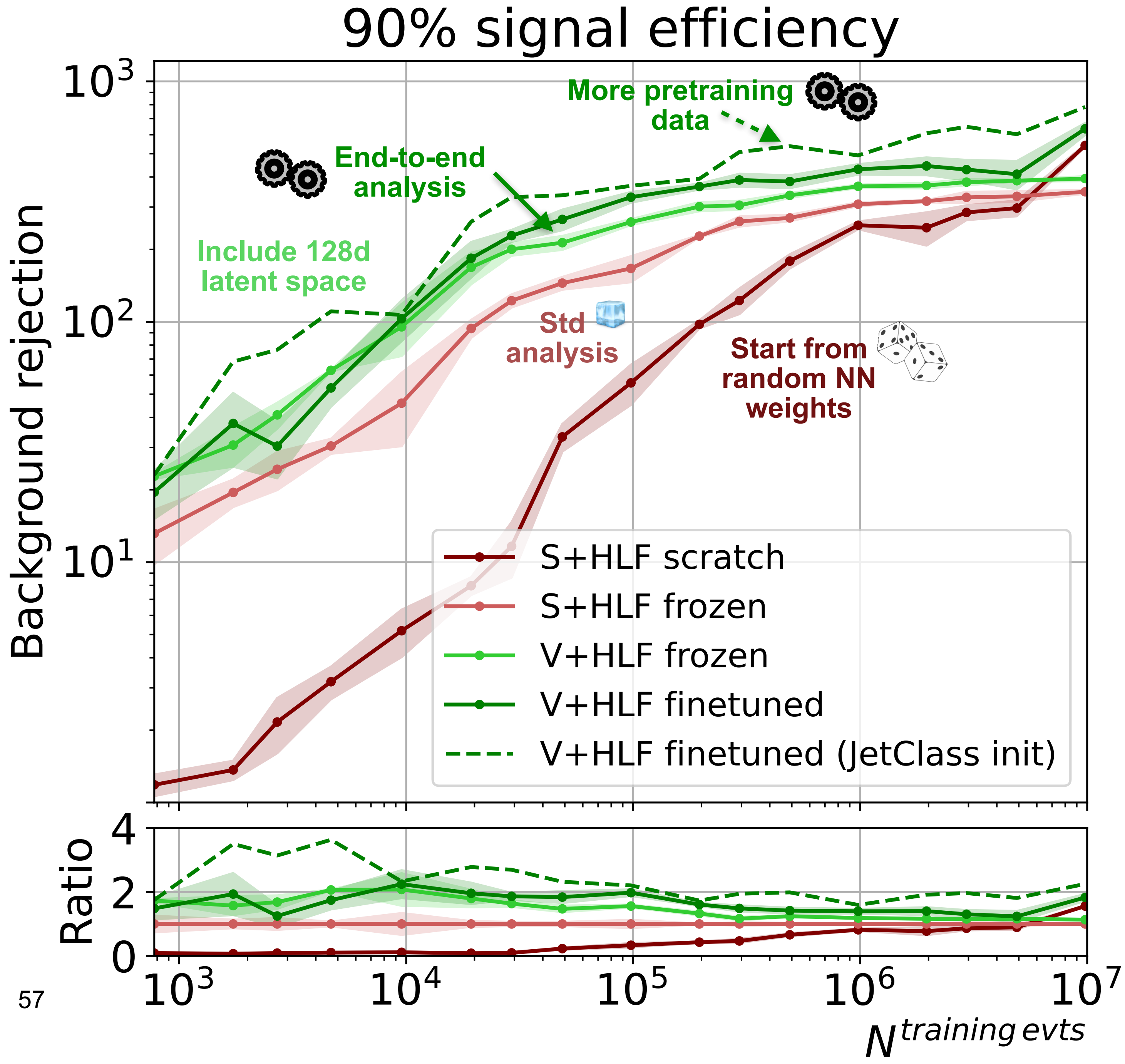
Q1/ Does this help us?

Backbone architectures



High Level Features (HLF):
p_T, η, φ m, soft drop m

Q2/ Does the vector latent dimensions hold extra information for analyses?



A better Higgs tagger helps analysis performance

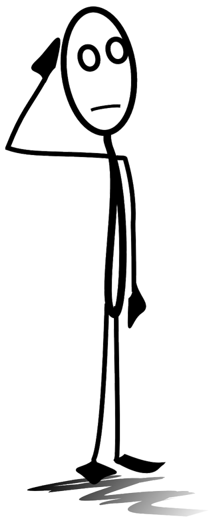
But training from scratch, with enough data, will surpass traditional analyses.

Decreases bkg by 2x …

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

Xbb + 4-vec < **vector + 4-vec** < **Xbb + 4-vec**

Standard ML HEP



More info from latent

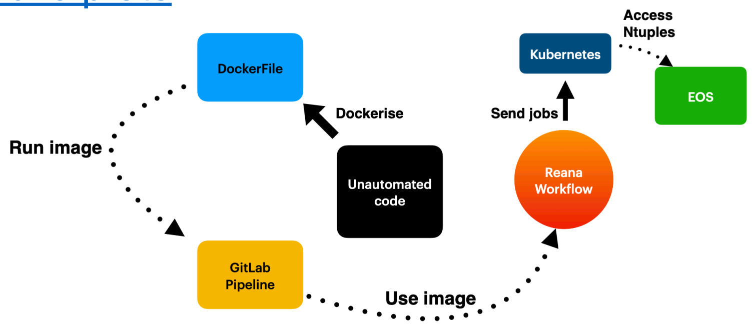
➡ Need a continuous, multi-dim calibration

Recent work

Custom “Xbb” (scalar) for each analysis

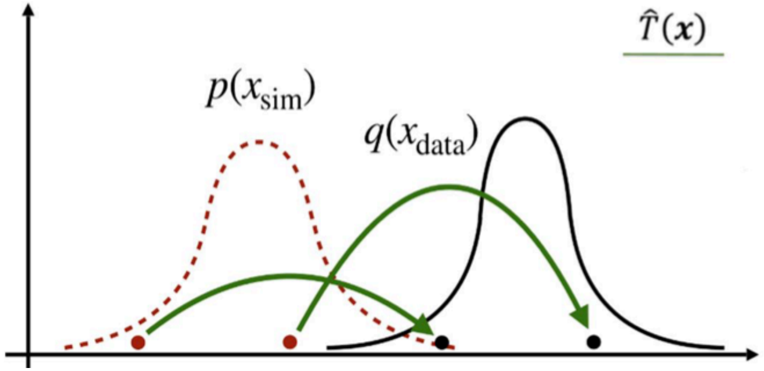
➡ Need custom calibration for each analysis

Will need automated calibrations
[public plots]



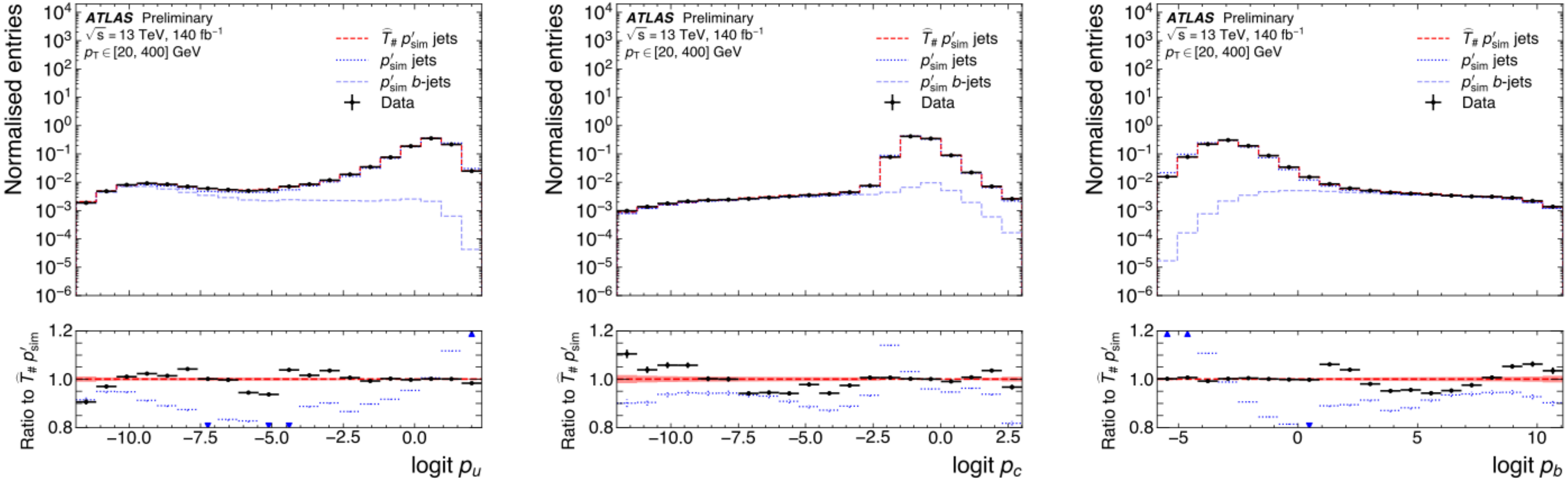
P. Windischhofer + C. Pollard
M. Algren talk; ATLAS-CONF-2024-014
P. Gadow for connection w/ our project

What about calibration …



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.

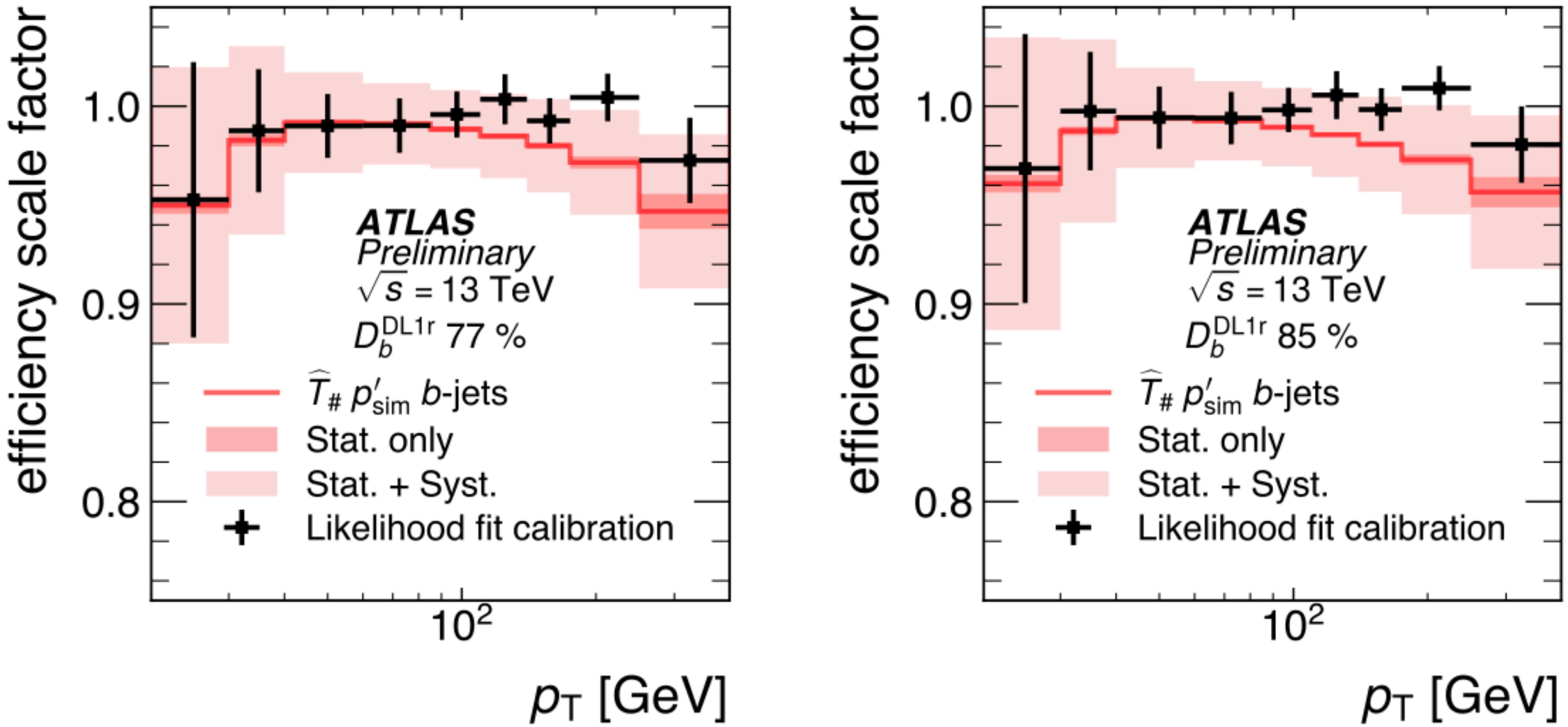


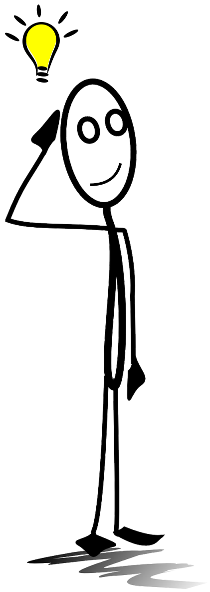
And result consistent with the standard calibration backup

P. Windischhofer + C. Pollard
M. Algren talk; ATLAS-CONF-2024-014
P. Gadow for connection w/ our project

What about calibration …

And matches the “standard calibration” with the same WPs

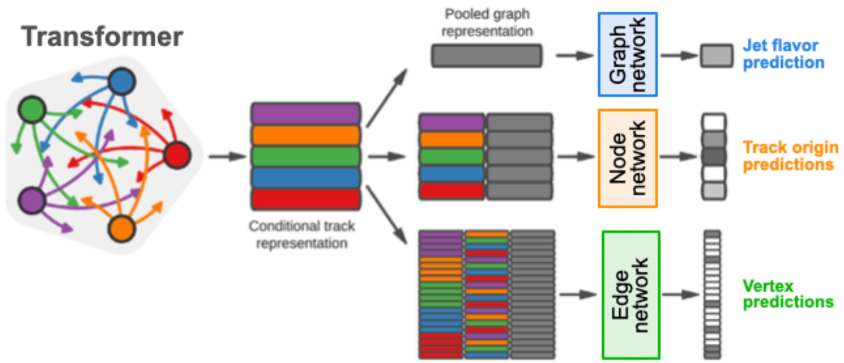


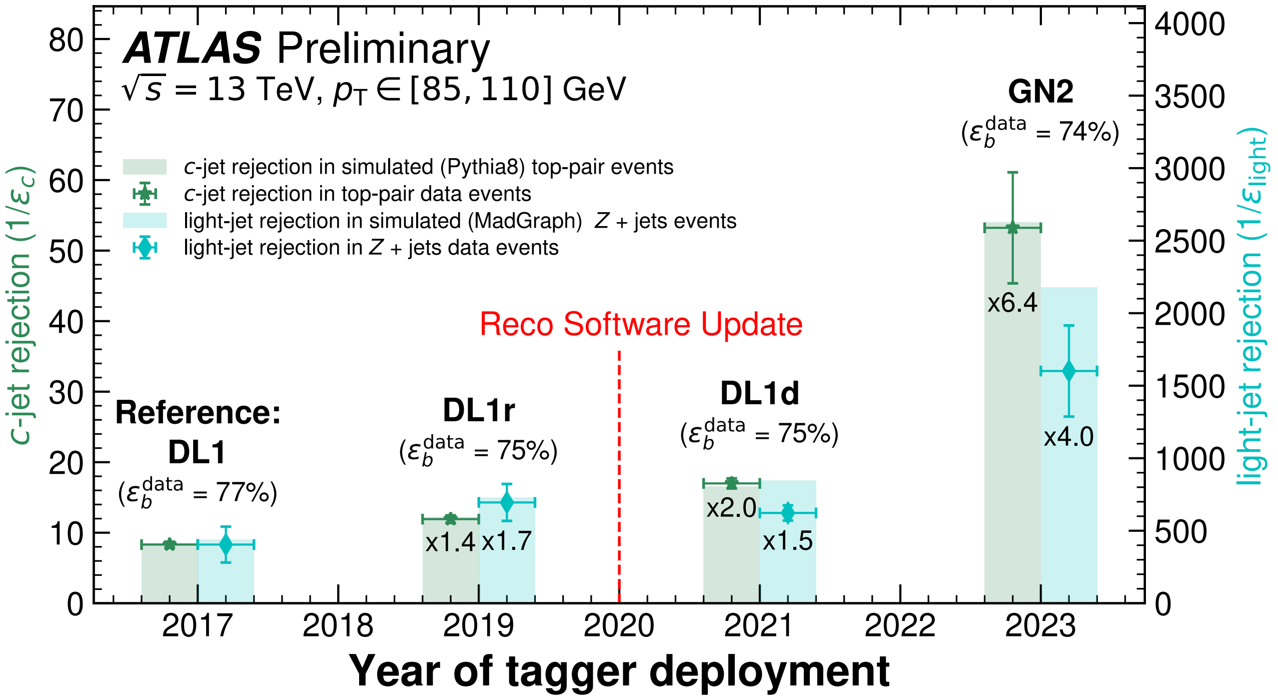


In ∑-mary

Deep Learning has transformed ATLAS FTAG our physics program in the past 7 years

- **RNNs**: model the jet as a SEQUENCE
- **Deep Sets** (DIPS): model the jet as a SET
- **Transformers** (GN2): Monolithic all-in-one architecture
  - Calibrated, ready for physics
  - Now collecting HH events in Run 3 triggers





Backup

Timeline

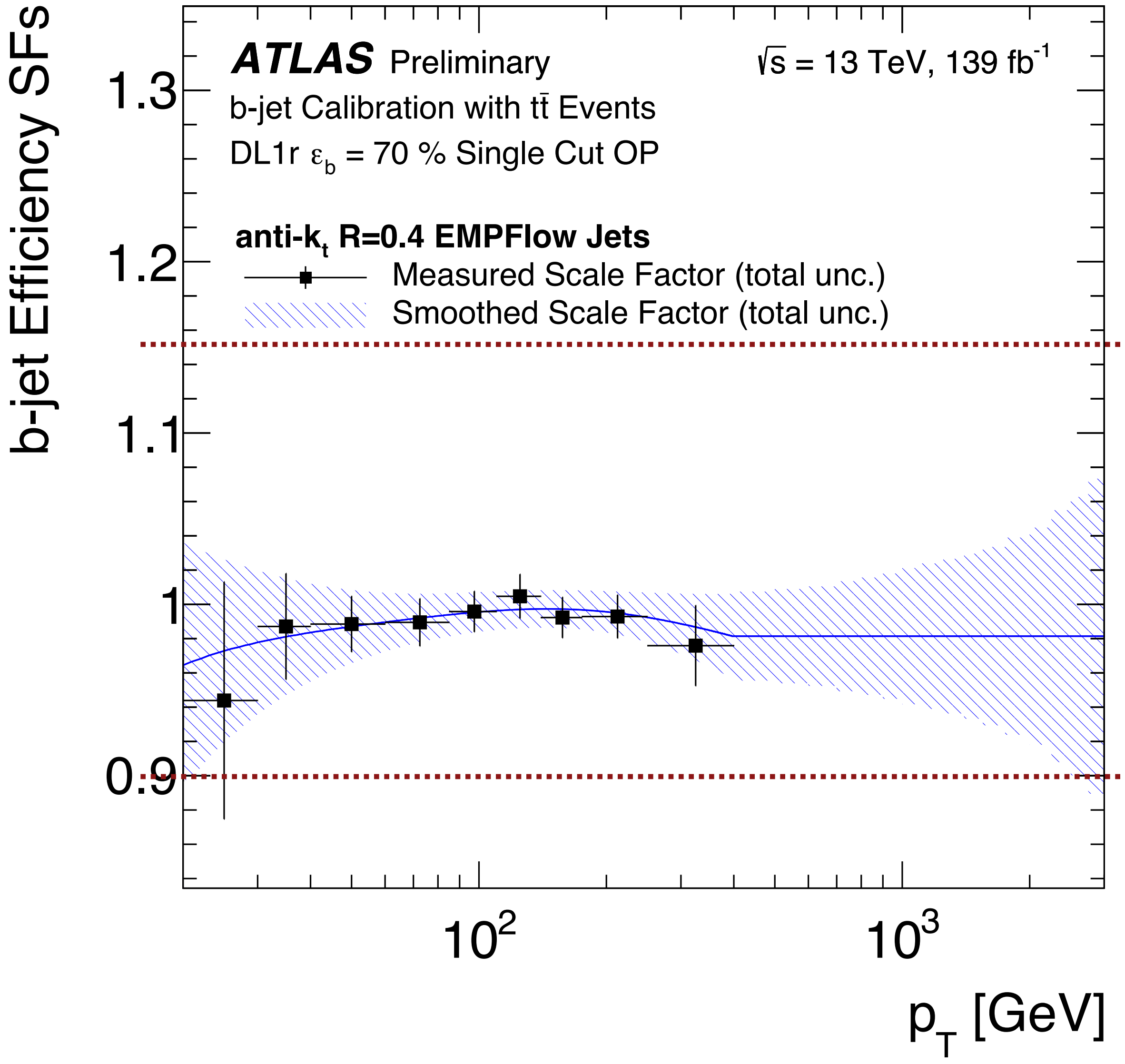
ML

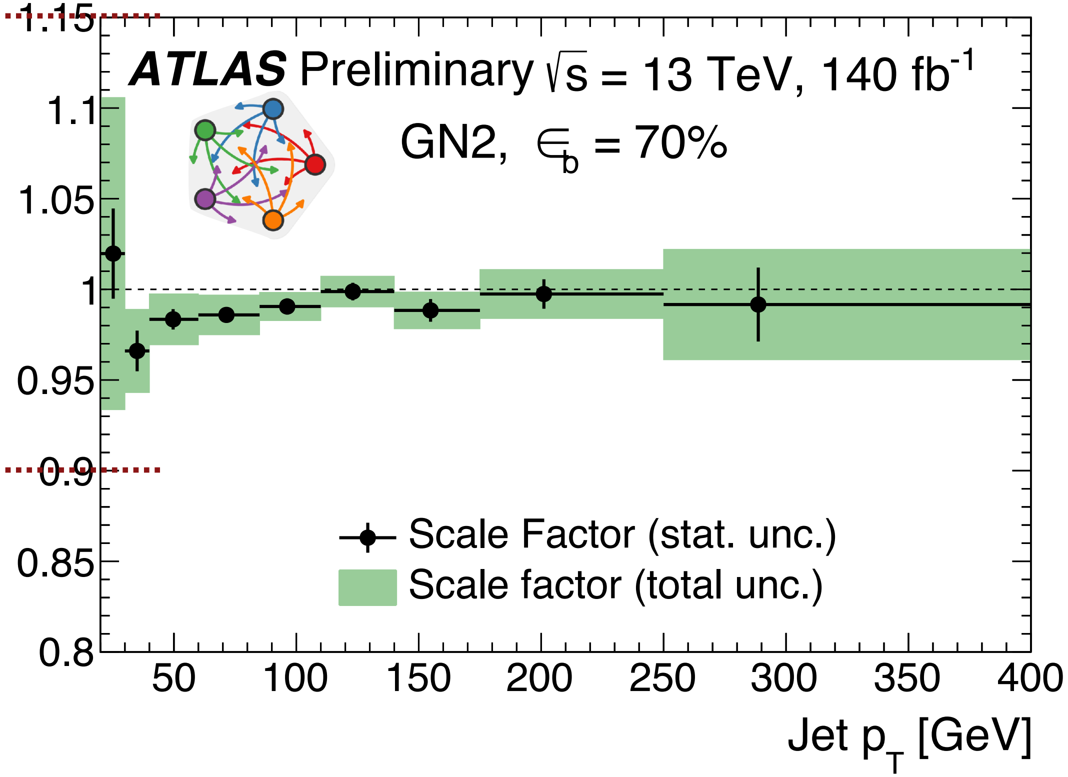
- 1997: **RNN** — LSTM
- 2017: **Deep Sets** — 1703.06114; **Transformers** — 1706.03762
- 2018: **Dynamic Graph CNN** — 1801.07829; **GNN review** — 1806.01261

Physics

- **RNNIP** — ATL-PHYS-PUB-2017-003
- 2018: **Deep Sets** — 1810.05165
- 2019: **Dynamic Graph CNN** — 1902.08570
- 2020: **GNNs** — 2008.02831; **DIPS** — ATL-PHYS-PUB-2020-014
- 2021
- 2022: **Transformers** — **ParT** — 2202.03772
- 2023: **GN2** — FTAG-2023-01

Calibration

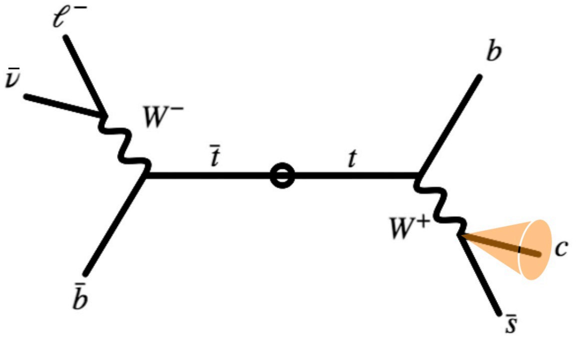


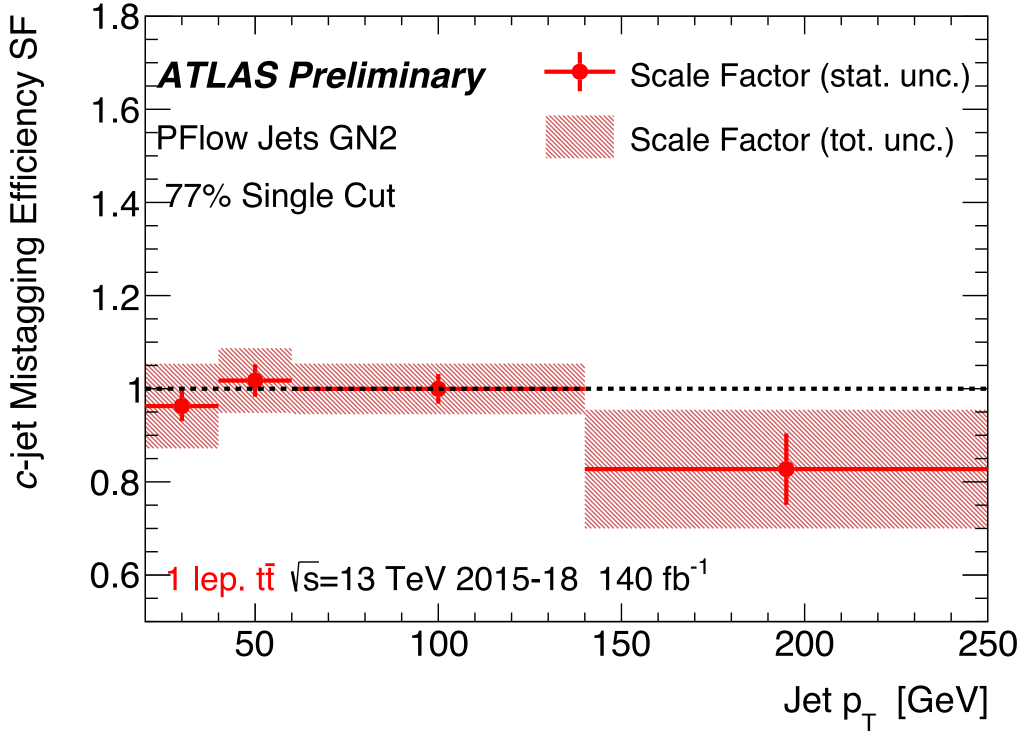


And no worse than the previous DL1r tagger 🥳

Calibration

c-jet mistag





light-jet mistag