A big "class" on jet physics

(Learning rich jet representation via o(100) classes to accelerate the LHC resonance search programme)

Congqiao Li (李聪乔), ***Peking University***

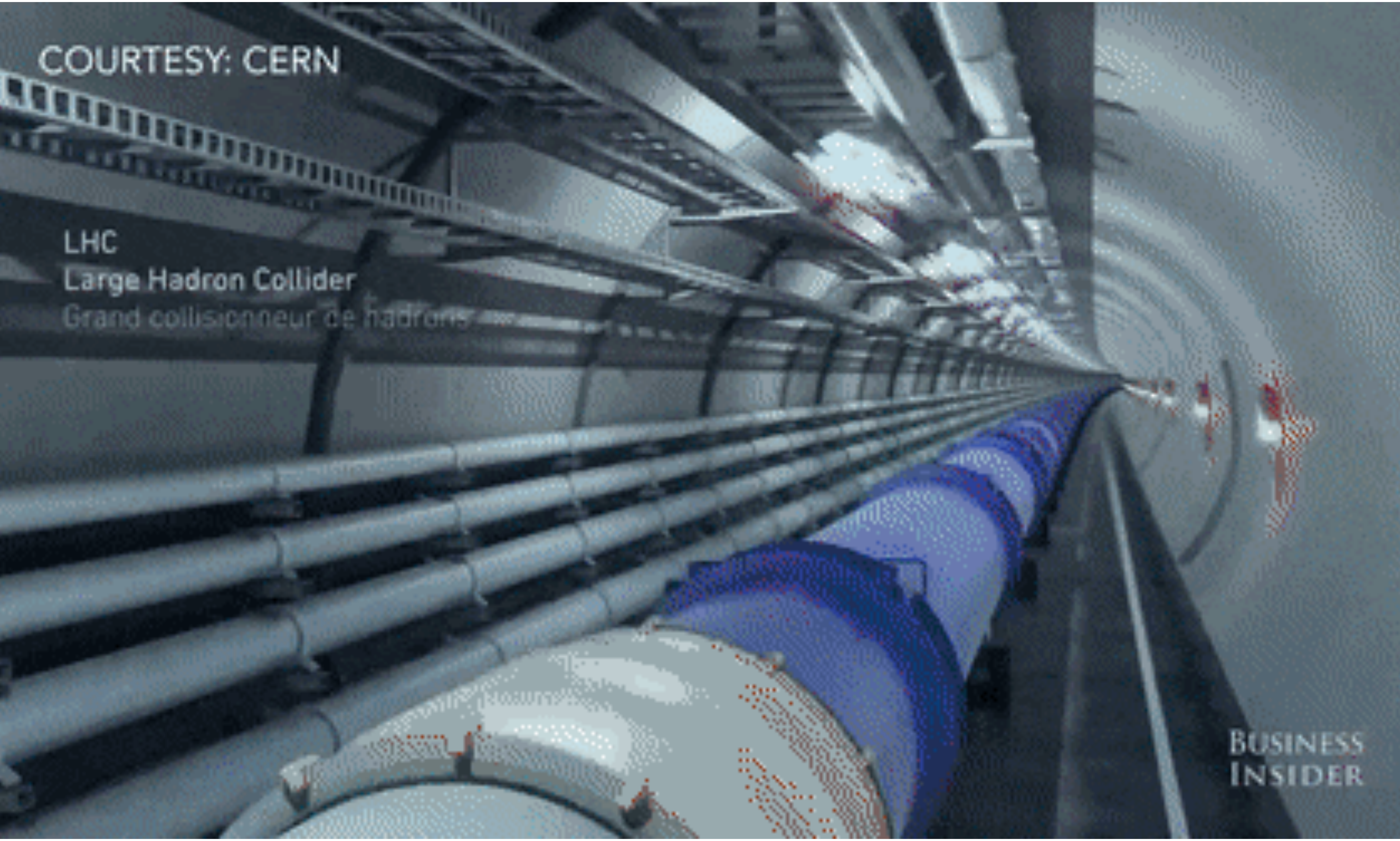
This talk is mainly based on

- Accelerating LHC resonant search via *Sophon*: arXiv:2405.12972 [Github] [Dataset] [Model] [Google Colab]
- Development of *Global Particle Transformer (GloParT) 3* within CMS

Larger than Larger: Large AI Models at the Frontiers of Experimental High-Energy Physics, Beijing
7 January, 2025

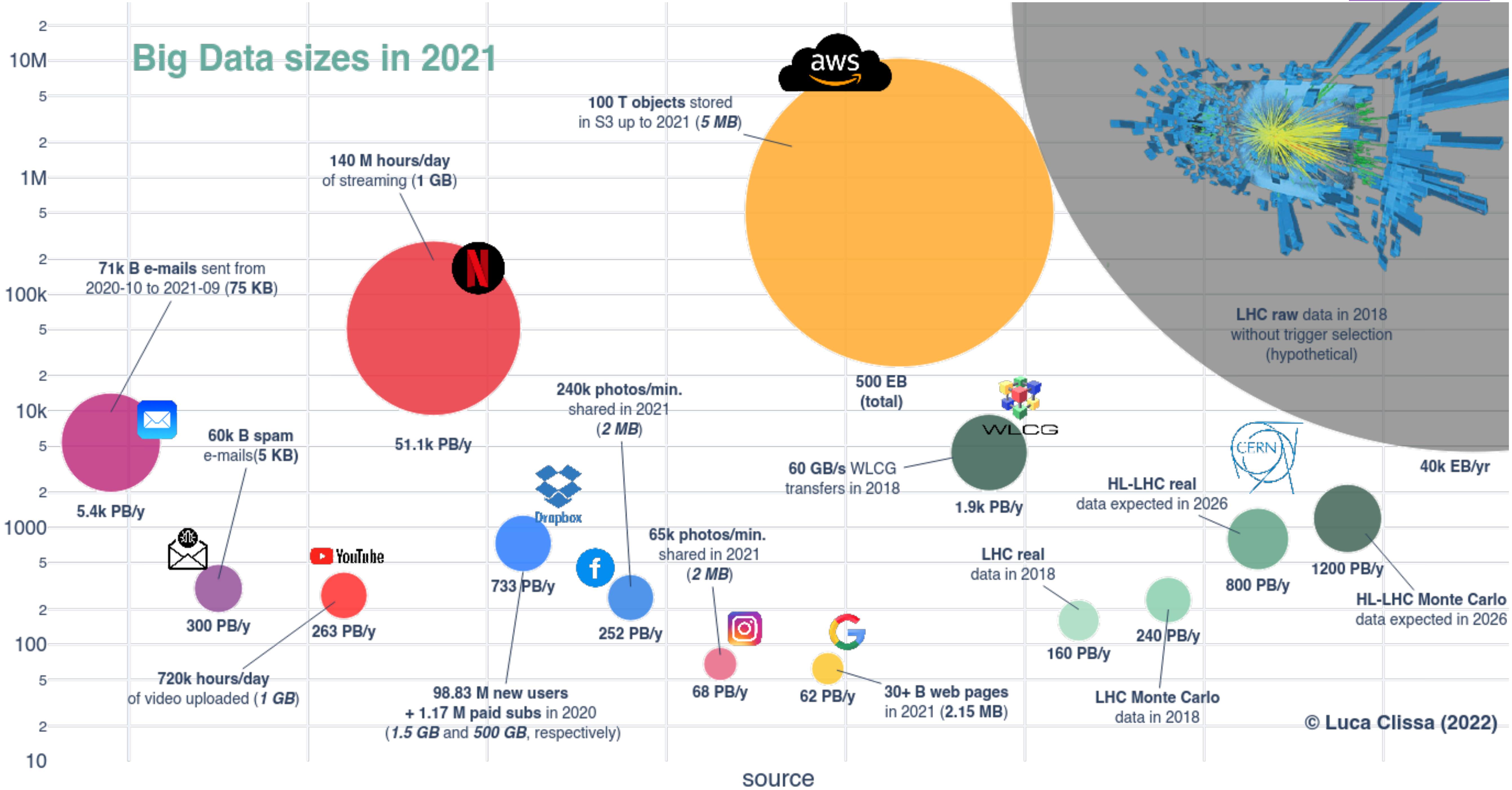
Let's begin with stories …

The Large Hadron Collider



In the age of big data...

Image credit

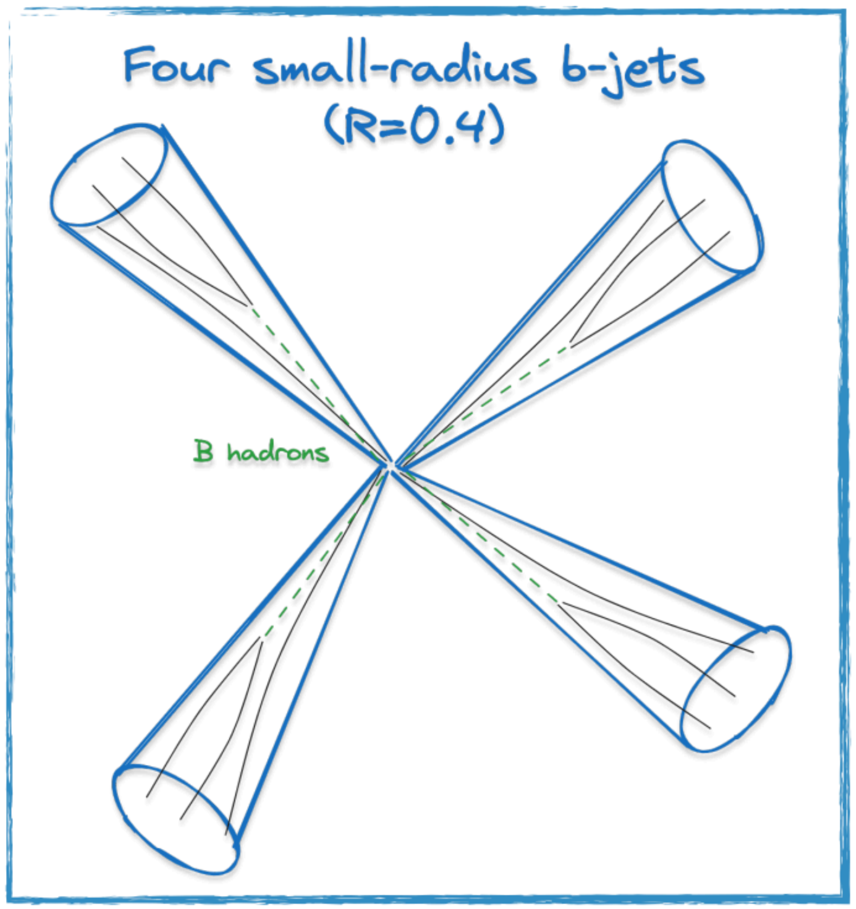


- Data analysis is a crucial subjet in the particle physics of this age
- *Novel engineering solutions are appearing!*
- *The advent of deep learning / AI is expected to play a transformative role*

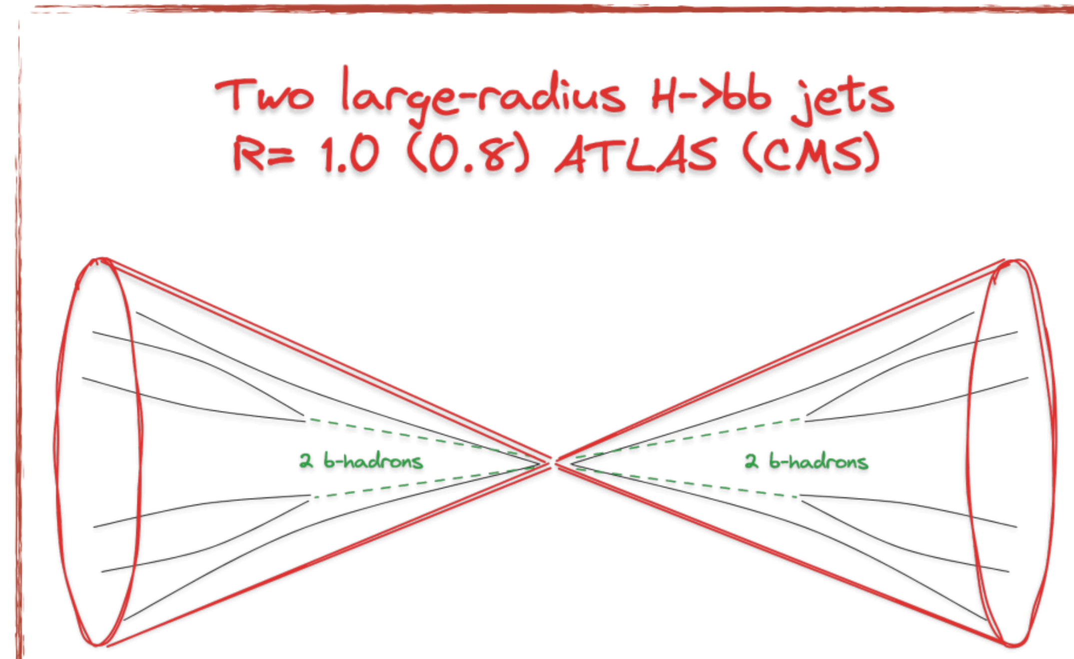
Quick example: how to search for HH→4b

cartoons credit to link

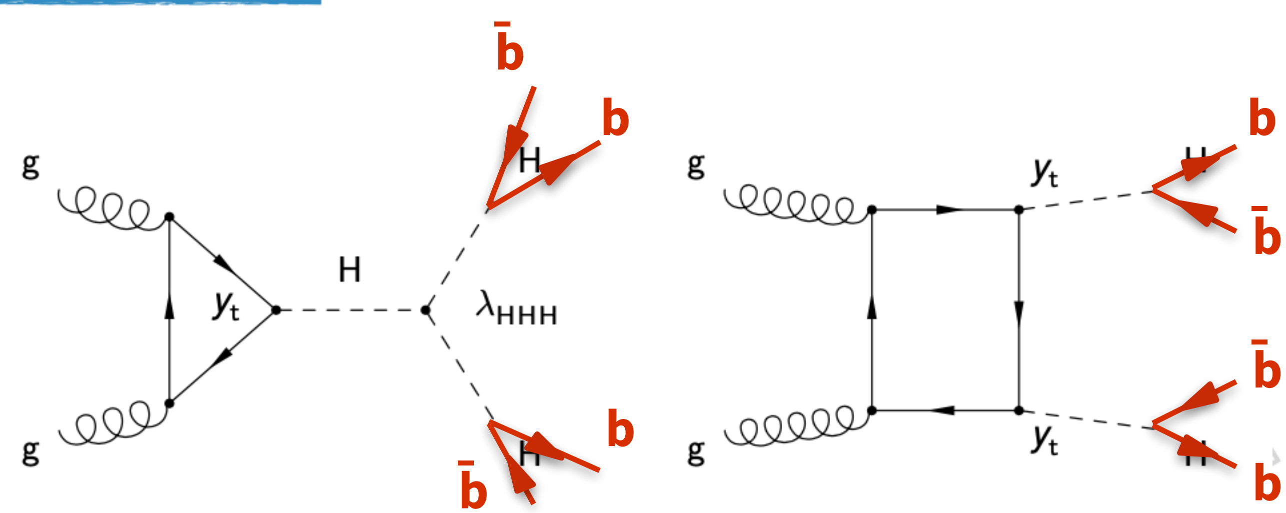
Resolved regime



Boosted regime



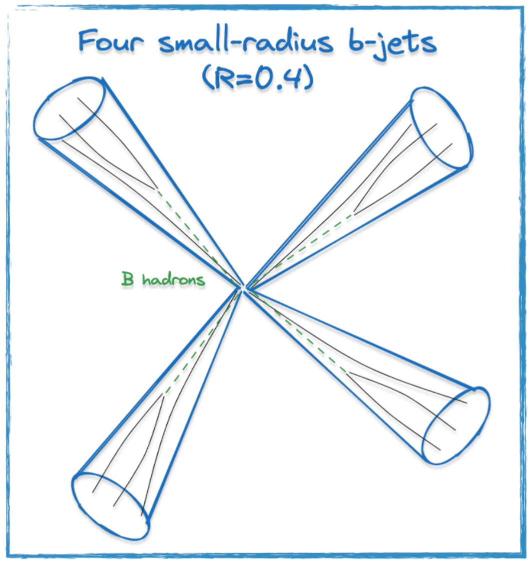
$$p_T \sim \frac{2m}{R} \geq 250 - 300 \text{ GeV}$$



To search for HH→4b signal at the LHC?

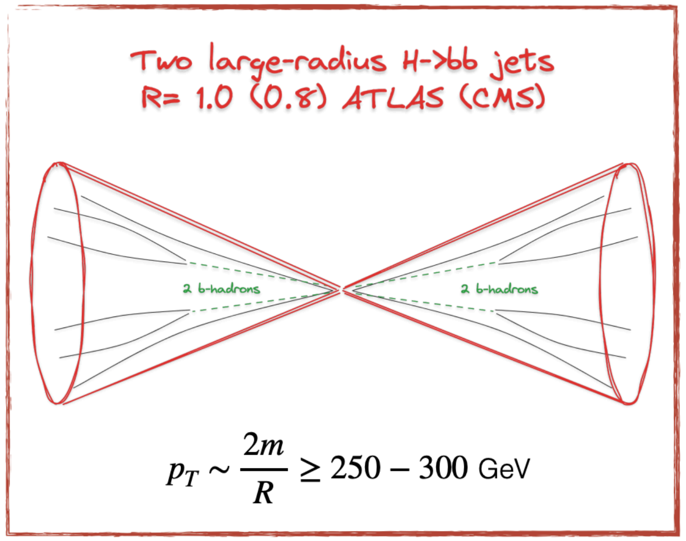
Boosted regime as a booster?

Resolved regime



- Boosted regime can only focus on high-pT behaviour
  - can only detect a tiny amount of signals ($m_{HH} \gtrsim 600$ GeV)
- but their sensitivity is close!

Boosted regime

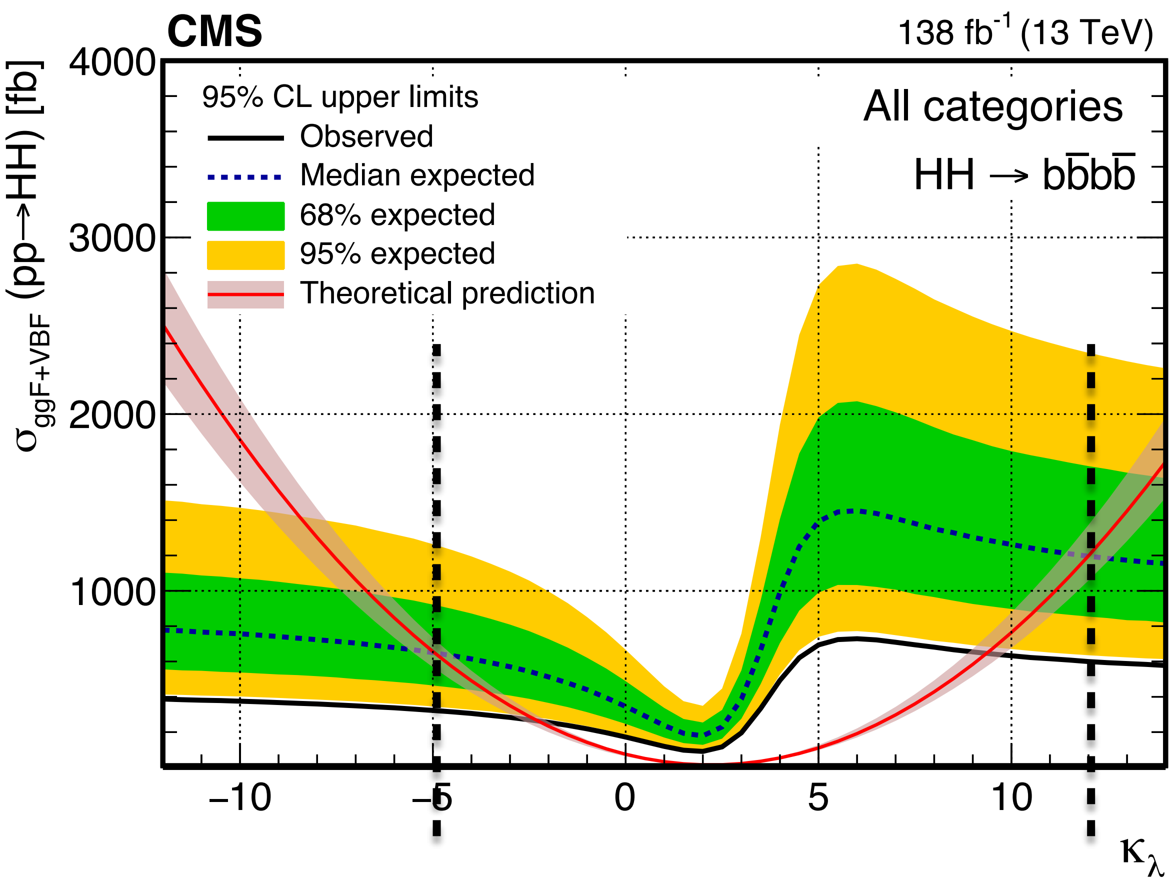


PRL 129 (2022) 081802

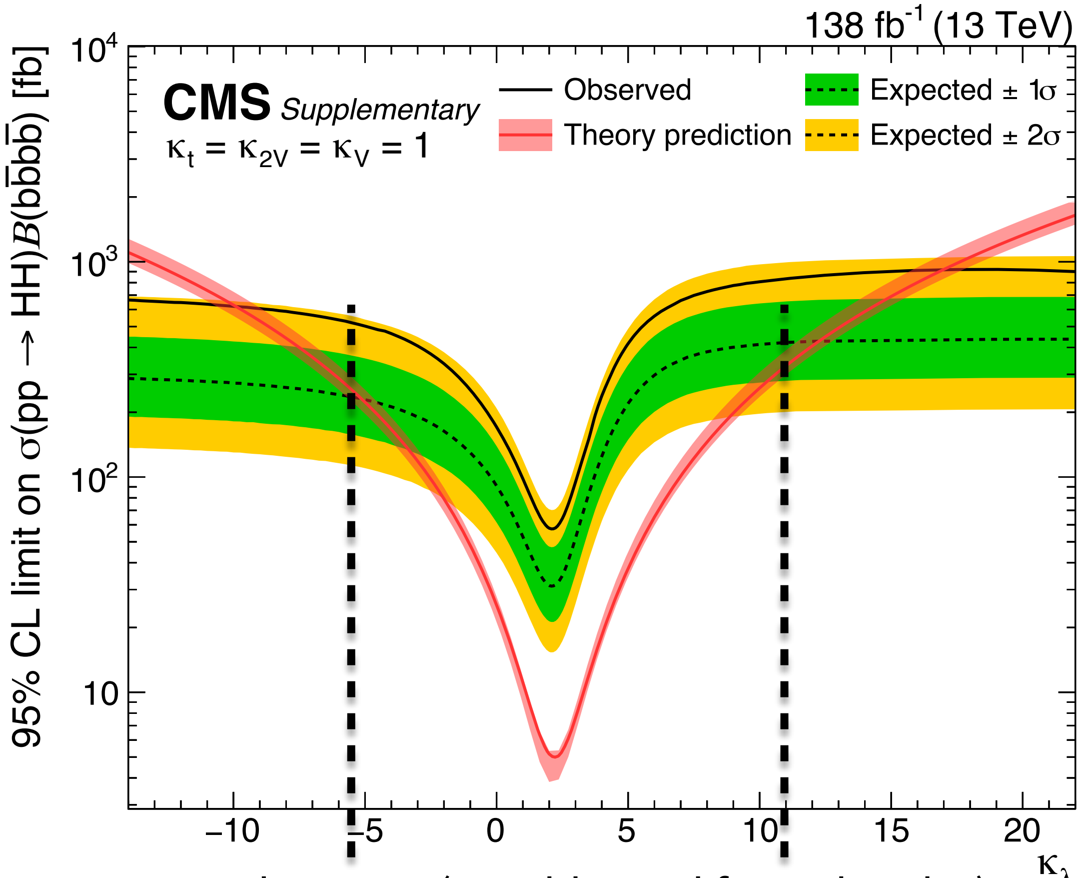
results website

PRL 131 (2023) 041803

results website



Exp (obs) 95% CL: $-5.0\ (-2.3) < \kappa_\lambda < 12.0\ (9.4)$

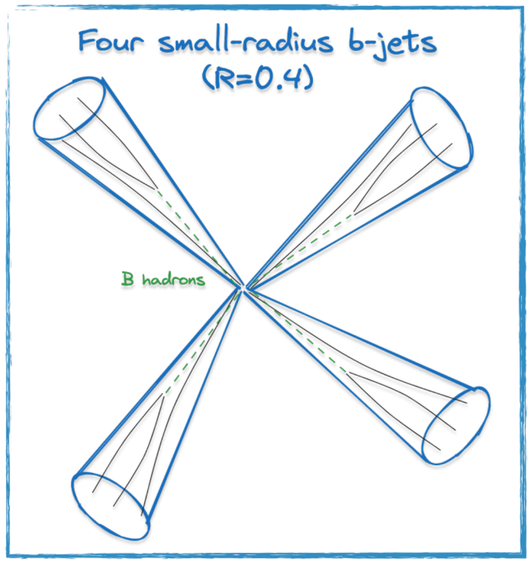


Expected 95% CL (roughly read from the plot):

$-5.0 < \kappa_\lambda < 12.5$

Boosted regime as a booster?

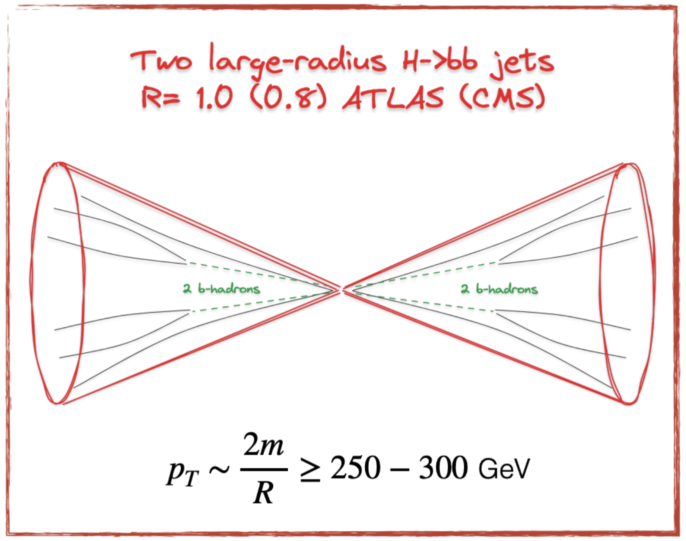
Resolved regime



- Boosted regime can only focus on high-pT behaviour
  - can only detect a tiny amount of signals ($m_{HH} \gtrsim 600$ GeV)
- but their sensitivity is close!

And even better results on κ_{2V} (where boosted regime is posed to give better sensitivity)

Boosted regime

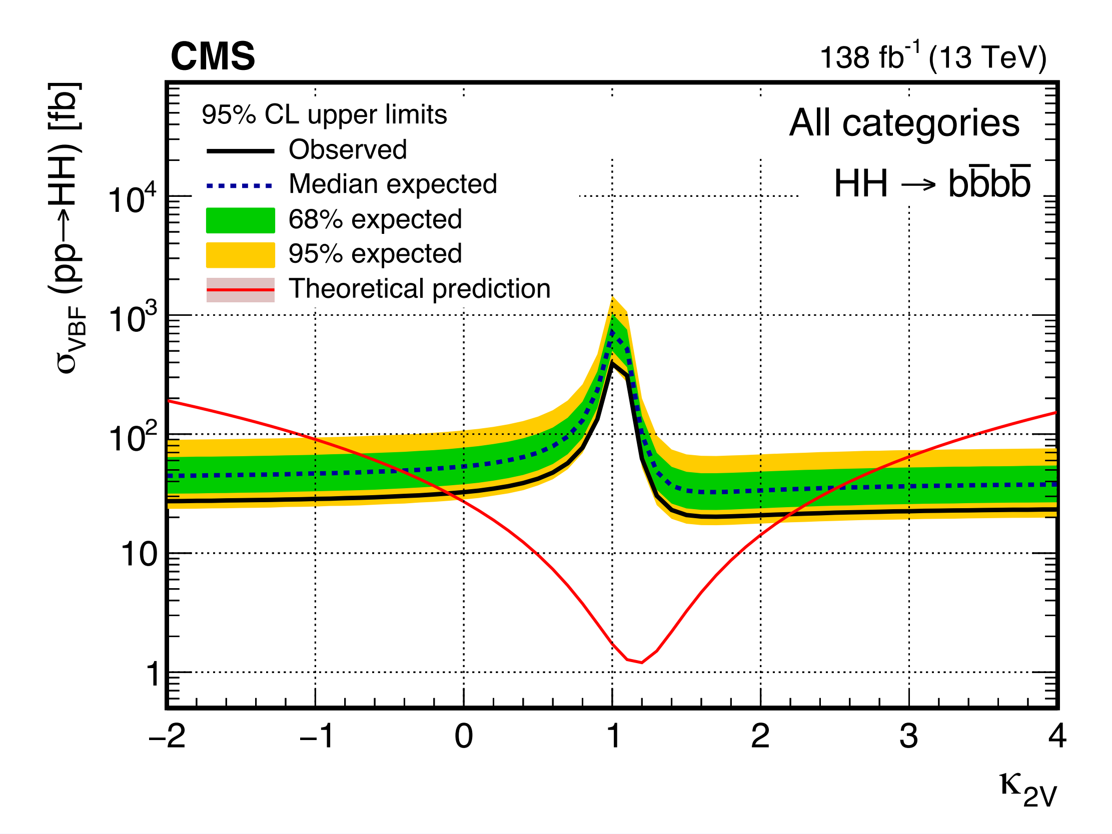


PRL 129 (2022) 081802

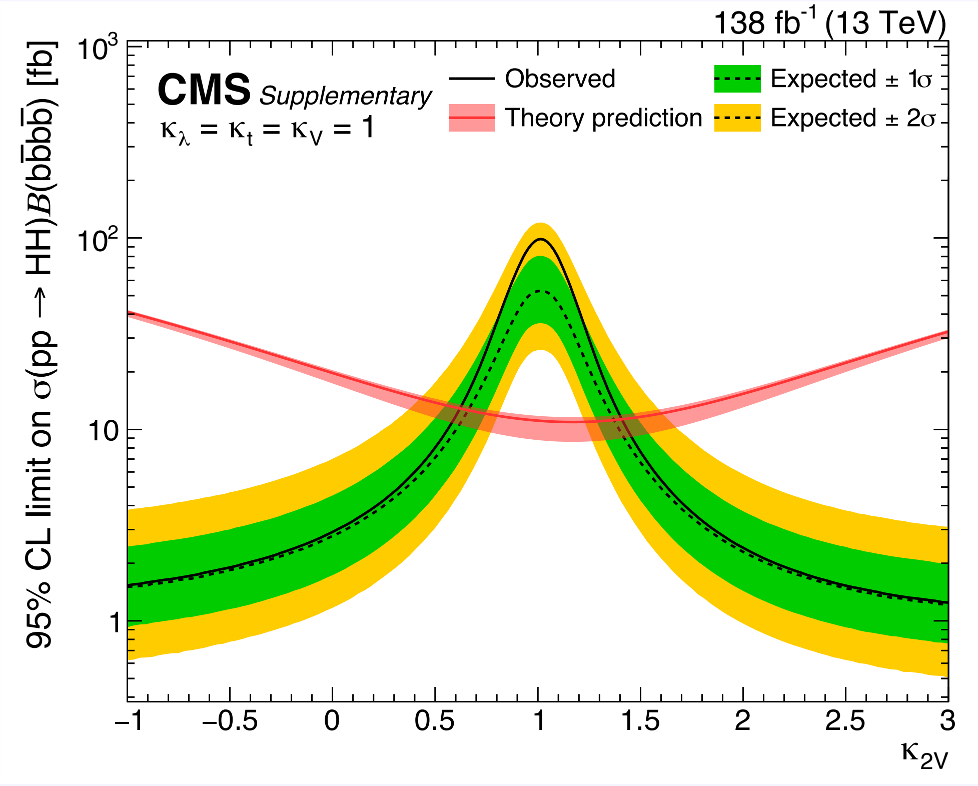
results website

PRL 131 (2023) 041803

results website



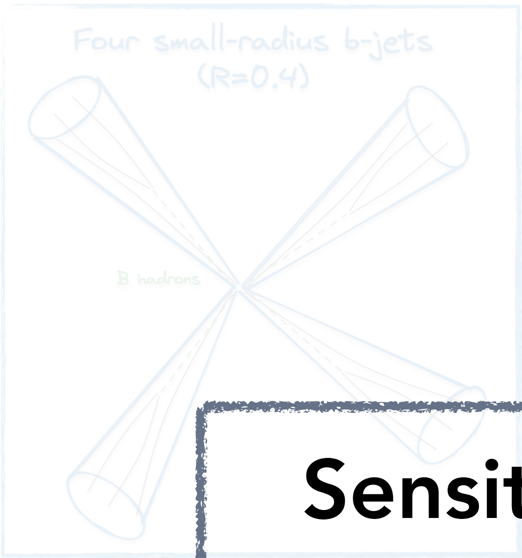
Exp (obs) 95% CL: $-0.1\ (-0.4) < \kappa_{2V} < 2.2\ (2.5)$



Exp (obs) 95% CL: $0.66\ (0.64) < \kappa_{2V} < 1.37\ (1.41)$

Boosted regime as a booster?

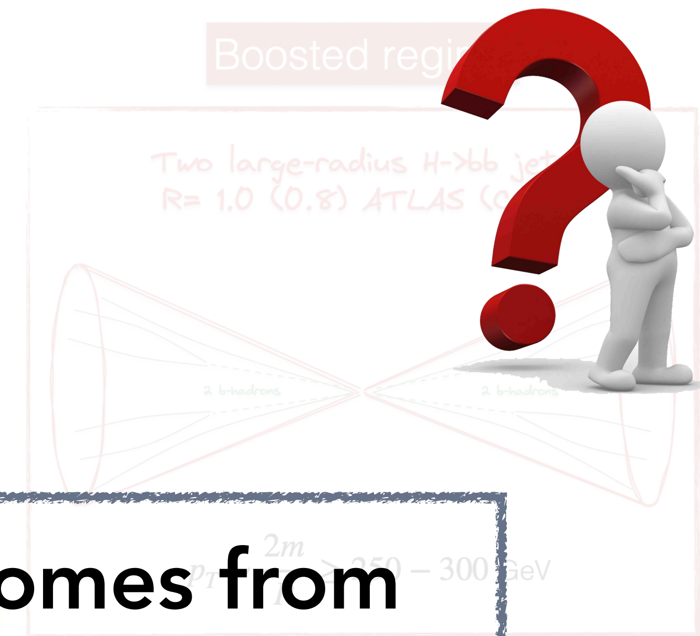
Resolved regime



- Boosted regime can only focus on high-pT behaviour
  - can only detect a tiny amount of signals ($m_{HH} \gtrsim 600$ GeV)
- but their sensitivity is close!

And even better results on κ_{2V} (where boosted regime is posed to give better sensitivity)

Boosted regime

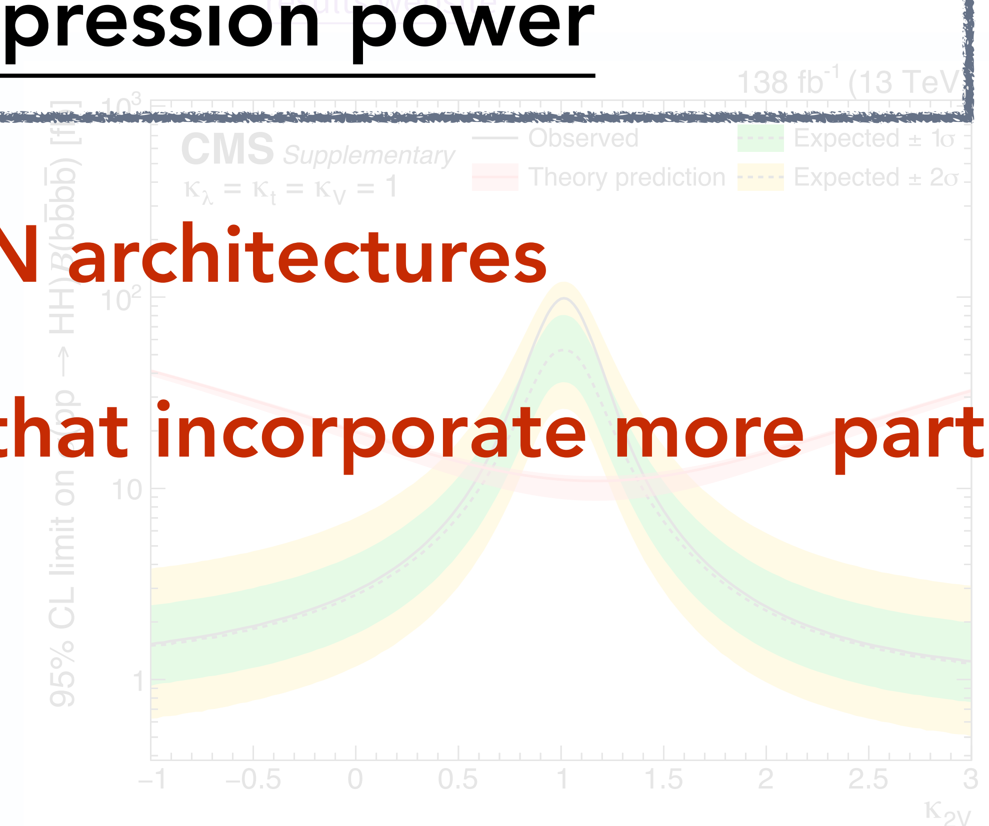


$p_T > \frac{2m}{\Delta R} \sim 250 - 300$ GeV

PRL 129 (2022) 081802 | PRL 131 (2023) 041803

results website | results website

Sensitivity in boosted topology largely comes from <u>the superior BKG suppression power</u>



More expressive NN architectures

+

Model operating on “large-R jets” that incorporate more particles

Exp (obs) 95% CL: $-0.1\ (-0.4) < \kappa_{2V} < 2.2\ (2.5)$

Exp (obs) 95% CL: $0.66\ (0.64) < \kappa_{2V} < 1.37\ (1.41)$

Outline

➔ I. Backgrounds

- overview of boosted-jet taggers at the LHC
- what's next?

➔ II. Introducing Sophon

- "large model for large-scale classification"; how are we led there?
- Sophon details & performance benchmark

➔ III. Implications for LHC resonance search

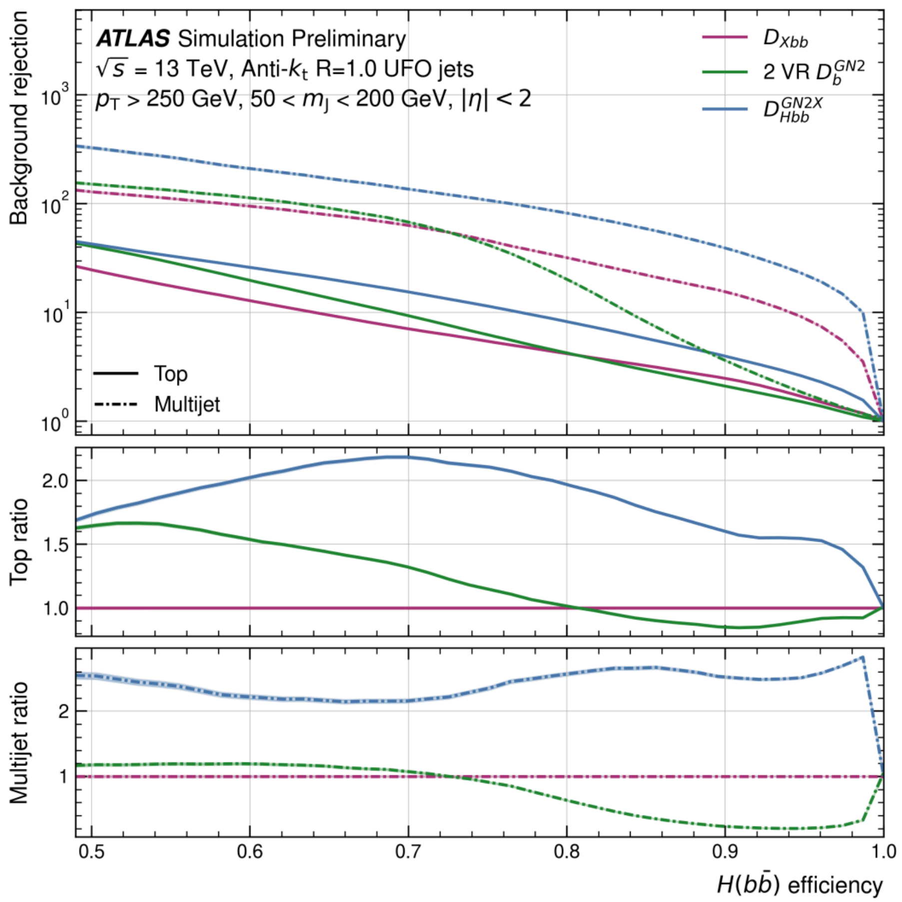
- model-specific and model-agnostic approaches
- *Global Particle Transformer*, the next-generation model in CMS
- More opportunities by Sophon/GloParT

➔ IV. Open discussion

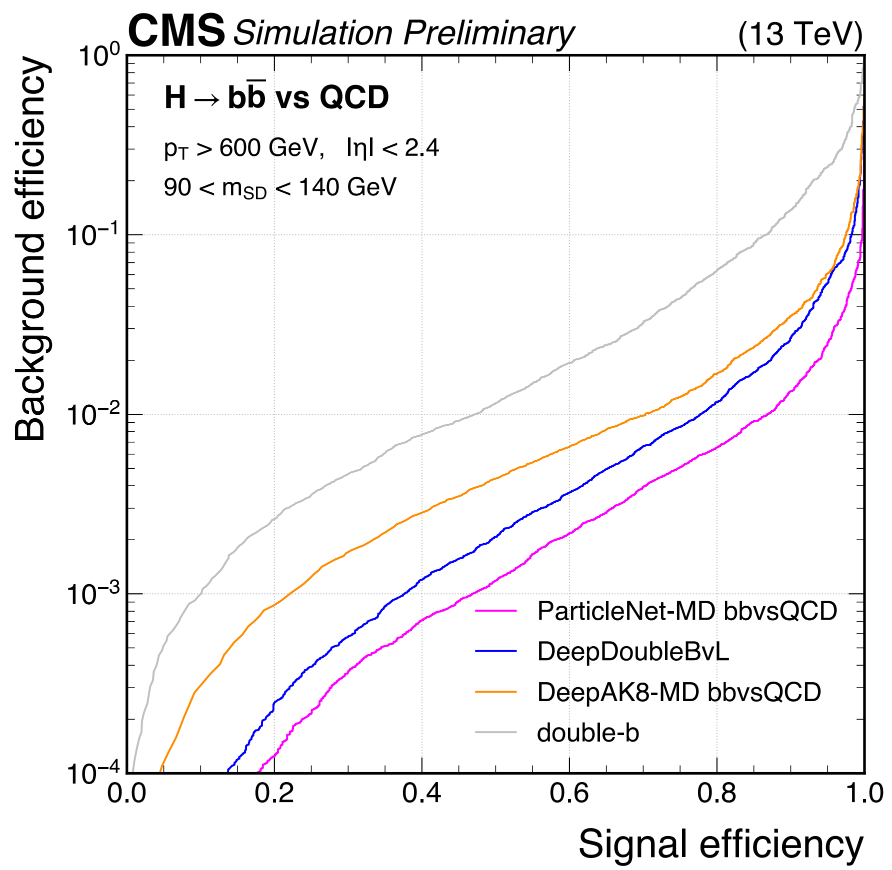
- datasets, training targets and scaling capabilities

Inspiring progress on H→bb̄/cc̄ tagging

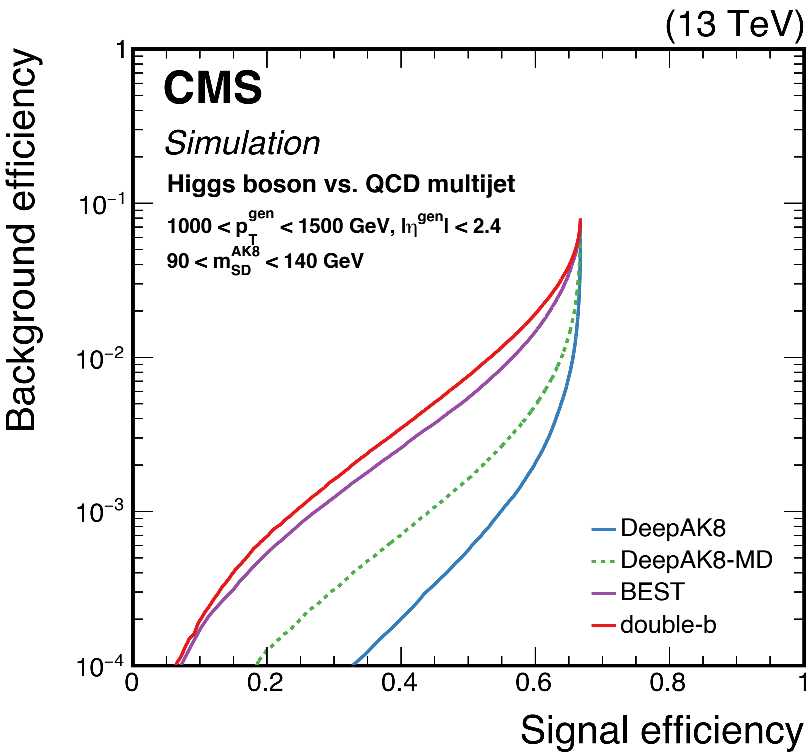
ATL-PHYS-PUB-2023-021



CMS-PAS-BTV-22-001



JINST 15 (2020) P06005



Transformer-based GN2X tagger:
~x3 📈 QCD and x2 top background rejection

DeepAK8 → ParticleNet:
x5 📈 QCD background rejection

Comparing with early approaches

Another **~x5 📈 improvement achieved**

Current boosted taggers in ATLAS/CMS

CMS: DeepAK8 and ParticleNet algorithms

JINST 15 (2020) P06005

Output

Category	Label
Higgs	H (bb)
	H (cc)
	H (VV*→qqqq)
Top	top (bcq)
	top (bqq)
	top (bc)
	top (bq)
W	W (cq)
	W (qq)
Z	Z (bb)
	Z (cc)
	Z (qq)
QCD	QCD (bb)
	QCD (cc)
	QCD (b)
	QCD (c)
	QCD (others)

DeepAK8-MD, ParticleNet-MD, and DeepDoubleX algorithms

CMS-PAS-BTV-22-001

- Focus on variable-mass resonance decays
- X→bb, cc, qq and QCD (5 subclasses)

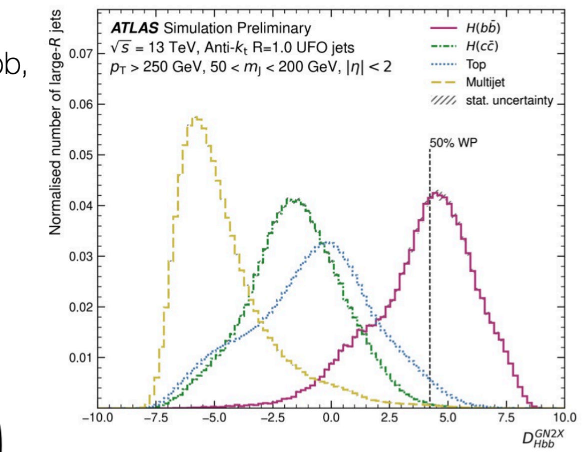
GN2X tagger *ATL-PHYS-PUB-2023-021*

- including flat-mass H→bb, cc and t→bqq samples, with QCD

Jackson's slides

GN2X Outputs

- GN2X adds a H → cc output class in addition to the H → bb, top and QCD classes from the previous tagger
- A discriminant score is built using a weighted log likelihood ratio similar to what's used for small-R tagging
- GN2X also includes the same auxiliary vertexing and track origin classification tasks present in GN1/GN2



$$D_{\mathrm{Hbb}}^{\mathrm{GN2X}} = \ln\left(\frac{p_{\mathrm{Hbb}}}{f_{\mathrm{Hcc}}\cdot p_{\mathrm{Hcc}} + f_{\mathrm{top}}\cdot p_{\mathrm{top}} + \left(1 - f_{\mathrm{Hcc}} - f_{\mathrm{top}}\right)\cdot p_{\mathrm{QCD}}}\right)$$

How can we accelerate the pace?

CMS: DeepAK8 and ParticleNet algorithms

JINST 15 (2020) P06005

Output

Category	Label
Higgs	H (bb)
	H (cc)
	H (VV*→qqqq)
Top	top (bcq)
	top (bqq)
	top (bc)
	top (bq)
W	W (cq)
	W (qq)
Z	Z (bb)
	Z (cc)

DeepAK8-MD, ParticleNet-MD, and DeepDoubleX algorithms

CMS-PAS-BTV-22-001

- Focus on variable-mass resonance decays
- X→bb, cc, qq and QCD (5 subclasses)

GN2X tagger *ATL-PHYS-PUB-2023-021*

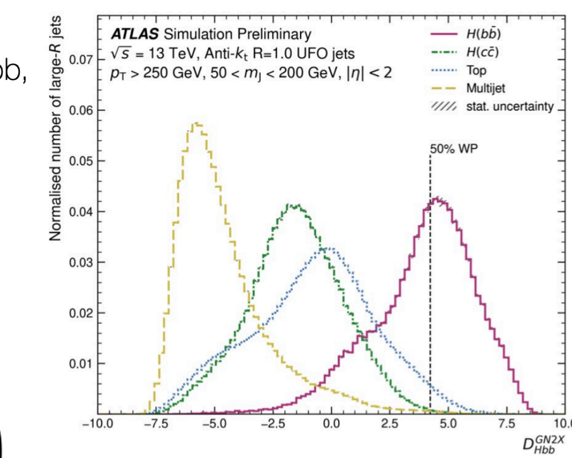
- including flat-mass H→bb, cc and t→bqq samples, with QCD

Jackson's slides

GN2X Outputs

- GN2X adds a H → cc output class in addition to the H → bb, top and QCD classes from the previous tagger
- A discriminant score is built using a weighted log likelihood ratio similar to what's used for small-R tagging
- GN2X also includes the same auxiliary vertexing and track origin classification tasks present in GN1/GN2

$$\ln\left(\frac{p_{\mathrm{Hbb}}}{f_{\mathrm{Hcc}}\cdot p_{\mathrm{Hcc}} + f_{\mathrm{top}}\cdot p_{\mathrm{top}} + (1 - f_{\mathrm{Hcc}} - f_{\mathrm{top}})\cdot p_{\mathrm{QCD}}}\right)$$



All final states!

Major types	Index range	Label names
Resonant jets: $X \to 2$ prong	0–14	$bb, cc, ss, qq, bc, cs, bq, cq, sq, gg, ee, \mu\mu, \tau_{\mathrm{h}}\tau_e, \tau_{\mathrm{h}}\tau_\mu, \tau_{\mathrm{h}}\tau_{\mathrm{h}}$
Resonant jets: $X \to 3$ or 4 prong	15–160	$bbbb, bbcc, bbss, bbqq, bbgg, bbee, bb\mu\mu, bb\tau_{\mathrm{h}}\tau_e, bb\tau_{\mathrm{h}}\tau_\mu, bb\tau_{\mathrm{h}}\tau_{\mathrm{h}}, bbb, bbc, bbs, bbq, bbg, bbe, bb\mu, cccc, ccss, ccqq, ccgg, ccee, cc\mu\mu, cc\tau_{\mathrm{h}}\tau_e, cc\tau_{\mathrm{h}}\tau_\mu, cc\tau_{\mathrm{h}}\tau_{\mathrm{h}}, ccb, ccc, ccs, ccq, ccg, cce, cc\mu, ssss, ssqq, ssgg, ssee, ss\mu\mu, ss\tau_{\mathrm{h}}\tau_e, ss\tau_{\mathrm{h}}\tau_\mu, ss\tau_{\mathrm{h}}\tau_{\mathrm{h}}, ssb, ssc, sss, ssq, ssg, sse, ss\mu, qqqq, qqgg, qqee, qq\mu\mu, qq\tau_{\mathrm{h}}\tau_e, qq\tau_{\mathrm{h}}\tau_\mu, qq\tau_{\mathrm{h}}\tau_{\mathrm{h}}, qqb, qqc, qqs, qqq, qqg, qqe, qq\mu, gggg, ggee, gg\mu\mu, gg\tau_{\mathrm{h}}\tau_e, gg\tau_{\mathrm{h}}\tau_\mu, gg\tau_{\mathrm{h}}\tau_{\mathrm{h}}, ggb, ggc, ggs, ggq, ggg, gge, gg\mu, bee, cee, see, qee, gee, b\mu\mu, c\mu\mu, s\mu\mu, q\mu\mu, g\mu\mu, b\tau_{\mathrm{h}}\tau_e, c\tau_{\mathrm{h}}\tau_e, s\tau_{\mathrm{h}}\tau_e, q\tau_{\mathrm{h}}\tau_e, g\tau_{\mathrm{h}}\tau_e, b\tau_{\mathrm{h}}\tau_\mu, c\tau_{\mathrm{h}}\tau_\mu, s\tau_{\mathrm{h}}\tau_\mu, q\tau_{\mathrm{h}}\tau_\mu, g\tau_{\mathrm{h}}\tau_\mu, b\tau_{\mathrm{h}}\tau_{\mathrm{h}}, c\tau_{\mathrm{h}}\tau_{\mathrm{h}}, s\tau_{\mathrm{h}}\tau_{\mathrm{h}}, q\tau_{\mathrm{h}}\tau_{\mathrm{h}}, g\tau_{\mathrm{h}}\tau_{\mathrm{h}}, qqqb, qqqc, qqqs, bbcq, ccbs, ccbq, ccsq, sscq, qqbc, qqbs, qqcs, bcsq, bcs, bcq, bsq, csq, bce\nu, cse\nu, bqe\nu, cqe\nu, sqe\nu, qqe\nu, bc\mu\nu, cs\mu\nu, bq\mu\nu, cq\mu\nu, sq\mu\nu, qq\mu\nu, bc\tau_e\nu, cs\tau_e\nu, bq\tau_e\nu, cq\tau_e\nu, sq\tau_e\nu, qq\tau_e\nu, bc\tau_\mu\nu, cs\tau_\mu\nu, bq\tau_\mu\nu, cq\tau_\mu\nu, sq\tau_\mu\nu, qq\tau_\mu\nu, bc\tau_{\mathrm{h}}\nu, cs\tau_{\mathrm{h}}\nu, bq\tau_{\mathrm{h}}\nu, cq\tau_{\mathrm{h}}\nu, sq\tau_{\mathrm{h}}\nu, qq\tau_{\mathrm{h}}\nu$
QCD jets	161–187	$bbccss, bbccs, bbcc, bbcss, bbcs, bbc, bbss, bbs, bb, bccss, bccs, bcc, bcss, bcs, bc, bss, bs, b, ccss, ccs, cc, css, cs, c, ss, s,$ others

How can we accelerate the pace?

Sophon (智子): *Signature-Oriented Pre-training for Heavy-resonant ObservatioN*

	Index range	Label names
$X \to 2$ prong	0–14	$bb, cc, ss, qq, bc, cs, bq, cq, sq, gg, ee, \mu\mu, \tau_{\mathrm{h}}\tau_e, \tau_{\mathrm{h}}\tau_\mu, \tau_{\mathrm{h}}\tau_{\mathrm{h}}$
Resonant jets: $X \to 3$ or 4 prong	15–160	$bbbb, bbcc, bbss, bbqq, bbgg, bbee, bb\mu\mu, bb\tau_{\mathrm{h}}\tau_e, bb\tau_{\mathrm{h}}\tau_\mu, bb\tau_{\mathrm{h}}\tau_{\mathrm{h}}, bbb, bbc, bbs, bbq, bbg, bbe, bb\mu, cccc, ccss, ccqq, ccgg, ccee, cc\mu\mu, cc\tau_{\mathrm{h}}\tau_e, cc\tau_{\mathrm{h}}\tau_\mu, cc\tau_{\mathrm{h}}\tau_{\mathrm{h}}, ccb, ccc, ccs, ccq, ccg, cce, cc\mu, ssss, ssqq, ssgg, ssee, ss\mu\mu, ss\tau_{\mathrm{h}}\tau_e, ss\tau_{\mathrm{h}}\tau_\mu, ss\tau_{\mathrm{h}}\tau_{\mathrm{h}}, ssb, ssc, sss, ssq, ssg, sse, ss\mu, qqqq, qqgg, qqee, qq\mu\mu, qq\tau_{\mathrm{h}}\tau_e, qq\tau_{\mathrm{h}}\tau_\mu, qq\tau_{\mathrm{h}}\tau_{\mathrm{h}}, qqb, qqc, qqs, qqq, qqg, qqe, qq\mu, gggg, ggee, gg\mu\mu, gg\tau_{\mathrm{h}}\tau_e, gg\tau_{\mathrm{h}}\tau_\mu, gg\tau_{\mathrm{h}}\tau_{\mathrm{h}}, ggb, ggc, ggs, ggq, ggg, gge, gg\mu, bee, cee, see, qee, gee, b\mu\mu, c\mu\mu, s\mu\mu, q\mu\mu, g\mu\mu, b\tau_{\mathrm{h}}\tau_e, c\tau_{\mathrm{h}}\tau_e, s\tau_{\mathrm{h}}\tau_e, q\tau_{\mathrm{h}}\tau_e, g\tau_{\mathrm{h}}\tau_e, b\tau_{\mathrm{h}}\tau_\mu, c\tau_{\mathrm{h}}\tau_\mu, s\tau_{\mathrm{h}}\tau_\mu, q\tau_{\mathrm{h}}\tau_\mu, g\tau_{\mathrm{h}}\tau_\mu, b\tau_{\mathrm{h}}\tau_{\mathrm{h}}, c\tau_{\mathrm{h}}\tau_{\mathrm{h}}, s\tau_{\mathrm{h}}\tau_{\mathrm{h}}, q\tau_{\mathrm{h}}\tau_{\mathrm{h}}, g\tau_{\mathrm{h}}\tau_{\mathrm{h}}, qqqb, qqqc, qqqs, bbcq, ccbs, ccbq, ccsq, sscq, qqbc, qqbs, qqcs, bcsq, bcs, bcq, bsq, csq, bce\nu, cse\nu, bqe\nu, cqe\nu, sqe\nu, qqe\nu, bc\mu\nu, cs\mu\nu, bq\mu\nu, cq\mu\nu, sq\mu\nu, qq\mu\nu, bc\tau_e\nu, cs\tau_e\nu, bq\tau_e\nu, cq\tau_e\nu, sq\tau_e\nu, qq\tau_e\nu, bc\tau_\mu\nu, cs\tau_\mu\nu, bq\tau_\mu\nu, cq\tau_\mu\nu, sq\tau_\mu\nu, qq\tau_\mu\nu, bc\tau_{\mathrm{h}}\nu, cs\tau_{\mathrm{h}}\nu, bq\tau_{\mathrm{h}}\nu, cq\tau_{\mathrm{h}}\nu, sq\tau_{\mathrm{h}}\nu, qq\tau_{\mathrm{h}}\nu$
QCD jets	161–187	$bbccss, bbccs, bbcc, bbcss, bbcs, bbc, bbss, bbs, bb, bccss, bccs, bcc, bcss, bcs, bc, bss, bs, b, ccss, ccs, cc, css, cs, c, ss, s$, others

➔ Major concerns:

- ❖ Will the model achieve the best performance for each specific task?
- ❖ What can we use this model for, beyond its identification task supported by its final states?

How can we accelerate the pace?

Sophon (智子): *Signature-Oriented Pre-training for Heavy-resonant ObservatioN*

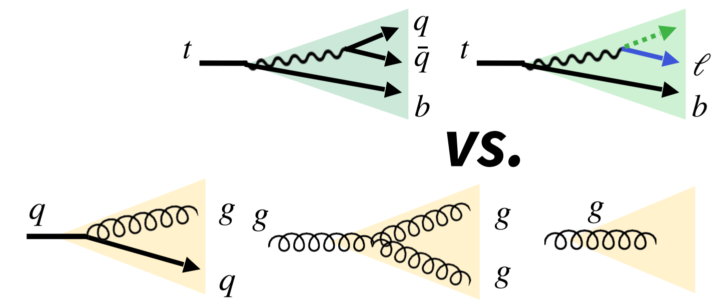
	Index range	Label names
Resonant jets: $X \to 2$ prong	0–14	bb, cc, ss, qq, bc, cs, bq, cq, sq, gg, ee, $\mu\mu$, $\tau_\mathrm{h}\tau_e$, $\tau_\mathrm{h}\tau_\mu$, $\tau_\mathrm{h}\tau_\mathrm{h}$
Resonant jets: $X \to 3$ or 4 prong	15–160	$bbbb$, $bbcc$, $bbss$, $bbqq$, $bbgg$, $bbee$, $bb\mu\mu$, $bb\tau_\mathrm{h}\tau_e$, $bb\tau_\mathrm{h}\tau_\mu$, $bb\tau_\mathrm{h}\tau_\mathrm{h}$, bbb, bbc, bbs, bbq, bbg, bbe, $bb\mu$, $cccc$, $ccss$, $ccqq$, $ccgg$, $ccee$, $cc\mu\mu$, $cc\tau_\mathrm{h}\tau_e$, $cc\tau_\mathrm{h}\tau_\mu$, $cc\tau_\mathrm{h}\tau_\mathrm{h}$, ccb, ccc, ccs, ccq, ccg, cce, $cc\mu$, $ssss$, $ssqq$, $ssgg$, $ssee$, $ss\mu\mu$, $ss\tau_\mathrm{h}\tau_e$, $ss\tau_\mathrm{h}\tau_\mu$, $ss\tau_\mathrm{h}\tau_\mathrm{h}$, ssb, ssc, sss, ssq, ssg, sse, $ss\mu$, $qqqq$, $qqgg$, $qqee$, $qq\mu\mu$, $qq\tau_\mathrm{h}\tau_e$, $qq\tau_\mathrm{h}\tau_\mu$, $qq\tau_\mathrm{h}\tau_\mathrm{h}$, qqb, qqc, qqs, qqq, qqg, qqe, $qq\mu$, $gggg$, $ggee$, $gg\mu\mu$, $gg\tau_\mathrm{h}\tau_e$, $gg\tau_\mathrm{h}\tau_\mu$, $gg\tau_\mathrm{h}\tau_\mathrm{h}$, ggb, ggc, ggs, ggq, ggg, gge, $gg\mu$, bee, cee, see, qee, gee, $b\mu\mu$, $c\mu\mu$, $s\mu\mu$, $q\mu\mu$, $g\mu\mu$, $b\tau_\mathrm{h}\tau_e$, $c\tau_\mathrm{h}\tau_e$, $s\tau_\mathrm{h}\tau_e$, $q\tau_\mathrm{h}\tau_e$, $g\tau_\mathrm{h}\tau_e$, $b\tau_\mathrm{h}\tau_\mu$, $c\tau_\mathrm{h}\tau_\mu$, $s\tau_\mathrm{h}\tau_\mu$, $q\tau_\mathrm{h}\tau_\mu$, $g\tau_\mathrm{h}\tau_\mu$, $b\tau_\mathrm{h}\tau_\mathrm{h}$, $c\tau_\mathrm{h}\tau_\mathrm{h}$, $s\tau_\mathrm{h}\tau_\mathrm{h}$, $q\tau_\mathrm{h}\tau_\mathrm{h}$, $g\tau_\mathrm{h}\tau_\mathrm{h}$, $qqqb$, $qqqc$, $qqqs$, $bbcq$, $ccbs$, $ccbq$, $ccsq$, $sscq$, $qqbc$, $qqbs$, $qqcs$, $bcsq$, bcs, bcq, bsq, csq, $bce\nu$, $cse\nu$, $bqe\nu$, $cqe\nu$, $sqe\nu$, $qqe\nu$, $bc\mu\nu$, $cs\mu\nu$, $bq\mu\nu$, $cq\mu\nu$, $sq\mu\nu$, $qq\mu\nu$, $bc\tau_e\nu$, $cs\tau_e\nu$, $bq\tau_e\nu$, $cq\tau_e\nu$, $sq\tau_e\nu$, $qq\tau_e\nu$, $bc\tau_\mu\nu$, $cs\tau_\mu\nu$, $bq\tau_\mu\nu$, $cq\tau_\mu\nu$, $sq\tau_\mu\nu$, $qq\tau_\mu\nu$, $bc\tau_\mathrm{h}\nu$, $cs\tau_\mathrm{h}\nu$, $bq\tau_\mathrm{h}\nu$, $cq\tau_\mathrm{h}\nu$, $sq\tau_\mathrm{h}\nu$, $qq\tau_\mathrm{h}\nu$
QCD jets	161–187	$bbccss$, $bbccs$, $bbcc$, $bbcss$, $bbcs$, bbc, $bbss$, bbs, bb, $bccss$, $bccs$, bcc, $bcss$, bcs, bc, bss, bs, b, $ccss$, ccs, cc, css, cs, c, ss, s, others

- ➔ Major concerns:
  - ❖ Will the model achieve the best performance for each specific task?
    - ⇨ *Yes, it does. The Particle Transformer can support each task to reach its optimal performance*
  - ❖ What can we use this model for, beyond its identification task supported by its final states?
    - ⇨ *We can regard it as a true **"based model"** and fine-tune it for wider ranges of downstream tasks*

Propose "Large model for large-scale classification"

View from jet tagging

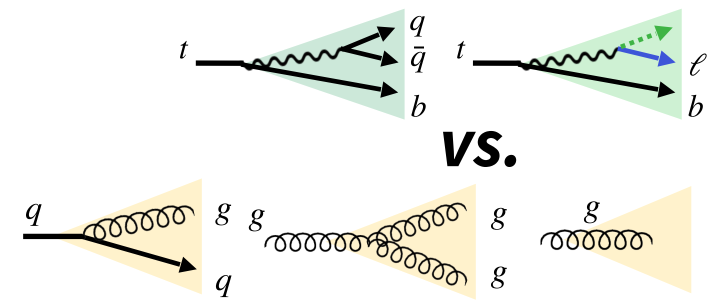
- ➔ Instead of training dedicated jet taggers, we consider **multi-class classification** with N(class) reaches $o(100)$
  - ❖ statistical insights: an ideal multi-class classifier is a stack of ideal binary classifiers
- ➔ The model should be **large** → carry enough capacity
- ➔ The classes should be comprehensive → **tagging ability can be further generalized by fine-tuning**



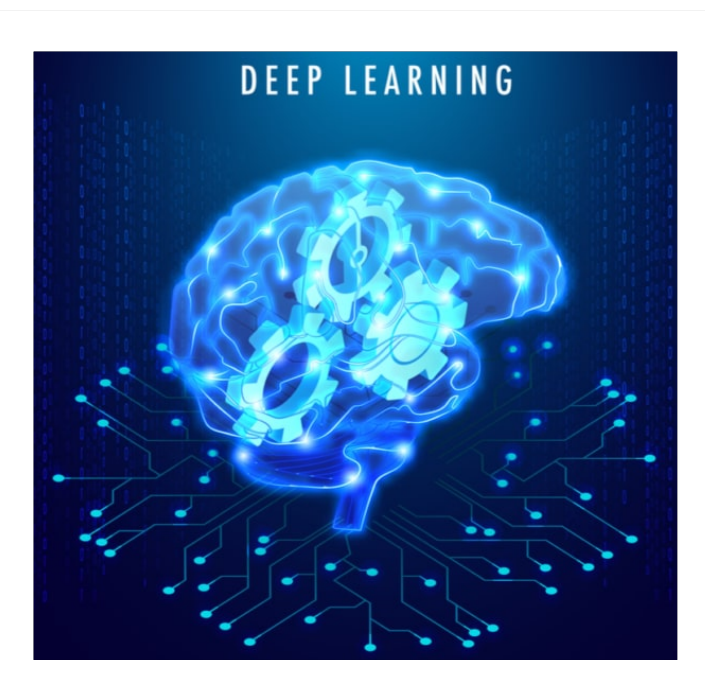
Propose "Large model for large-scale classification"

View from jet tagging

- ➔ Instead of training dedicated jet taggers, we consider **multi-class classification** with *N*(class) reaches *o*(100)
  - ❖ statistical insights: an ideal multi-class classifier is a stack of ideal binary classifiers
- ➔ The model should be **large** → carry enough capacity
- ➔ The classes should be comprehensive → **tagging ability can be further generalized by fine-tuning**



View from a pre-training solution

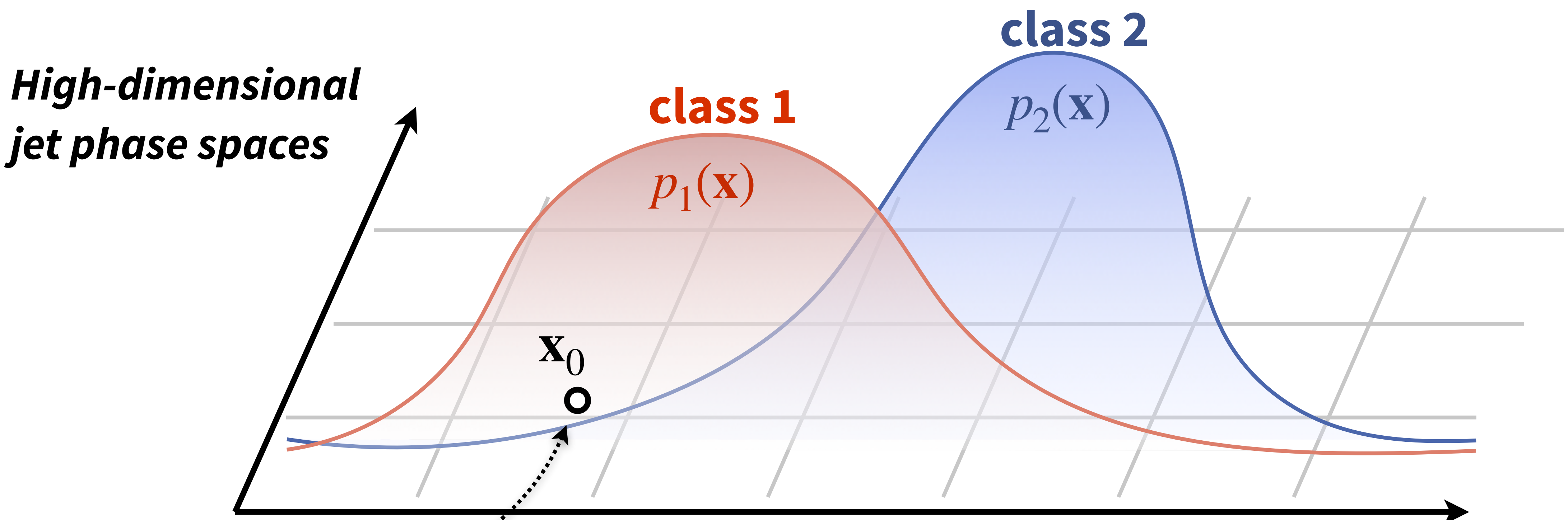


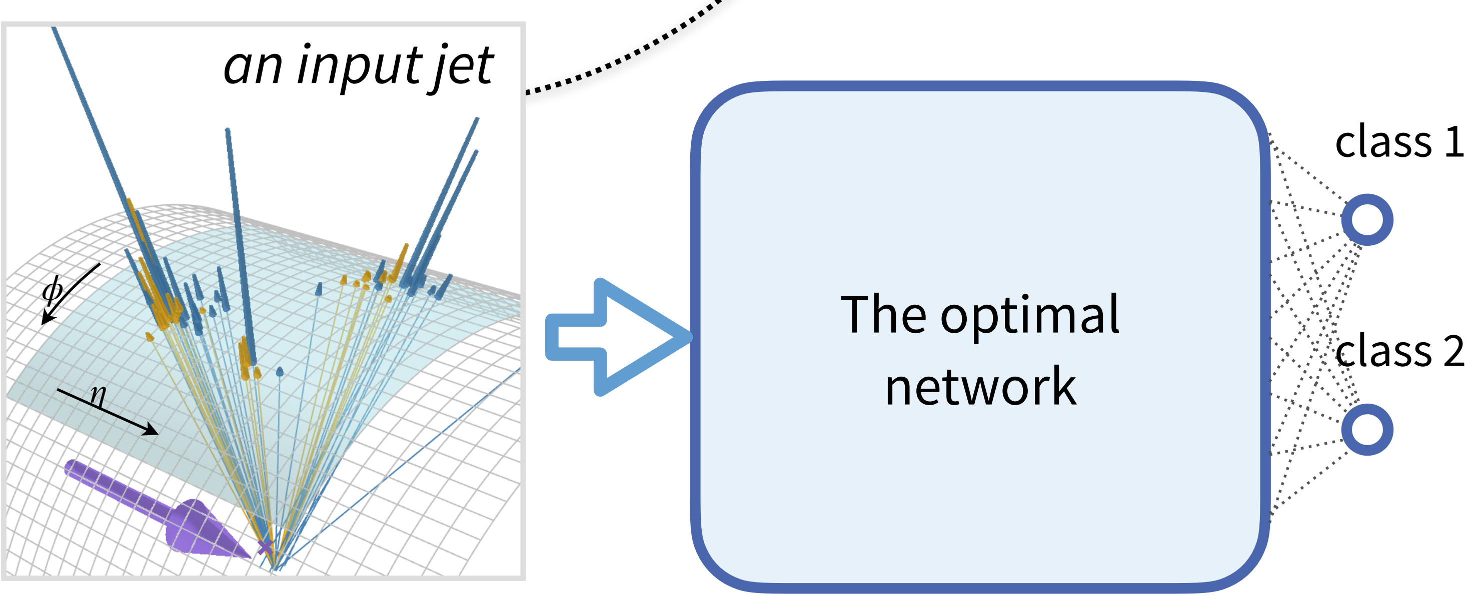
- ➔ Based on a comprehensive jet dataset, we hope to pre-train **a base model** to facilitate **all** LHC analyses exploring the large-*R* jet
- ➔ Set the training task: let the model learn to connect **"what a jet is like"** to **"which truth signature the jet reveals"** (= jet label in our case)
  - ❖ "jet labels" are simple signatures to explore → pre-training it as a classifier is just a starting point in this sense!

Statistical essence of jet tagging problem

→ **Question: where is the limit of jet tagging?**

→ Answer: the probability density ratio of two classes provides the optimal tagging





- Ideal classifier network results in $g_1 : g_2 : \ldots = p_1(\mathbf{x}_0) : p_2(\mathbf{x}_0) : \ldots$
- It is a direct estimation of p
- The **network capacity** decides how close the estimation is

Statistical property of multi-class classifier

→ Statistical theory shows that:

A **multi-class** classifier with minimum **cross-entropy loss** **estimates the probability ratios** on the input classes:

$$g_i(\mathbf{x}) = \frac{p(\text{class} = i\,|\,\mathbf{x})}{\sum_{j=1}^{N_{\text{out}}} p(\text{class} = j\,|\,\mathbf{x})}$$

→ hence it contains all the information the ideal $N(N-1)$ binary classifiers can do

Statistical property of multi-class classifier

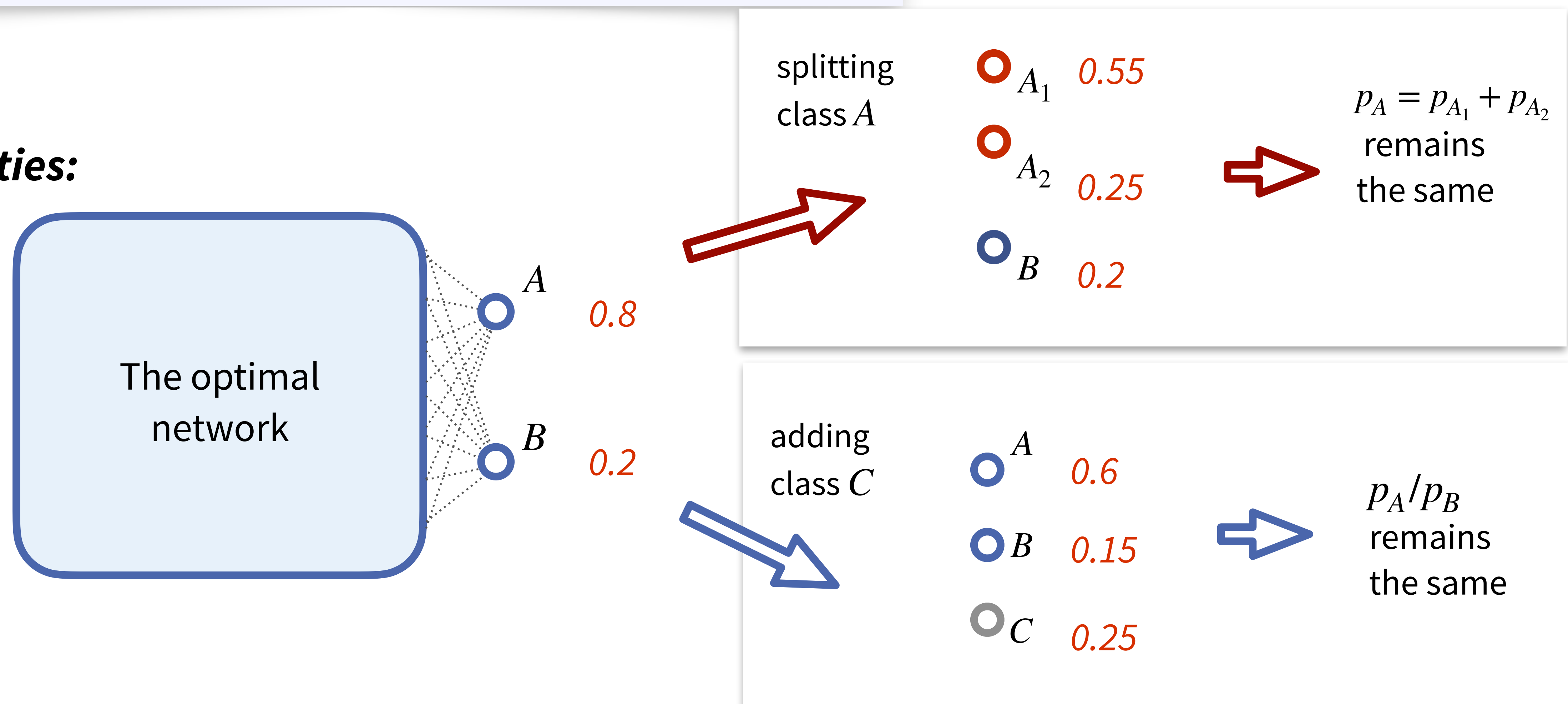
→ Statistical theory shows that:

A **multi-class** classifier with minimum **cross-entropy loss** **estimates the probability ratios** on the input classes:

$$g_i(\mathbf{x}) = \frac{p(\text{class} = i \,|\, \mathbf{x})}{\sum_{j=1}^{N_{\text{out}}} p(\text{class} = j \,|\, \mathbf{x})}$$

hence it contains all the information the ideal $N(N-1)$ binary classifiers can do

Two properties:



Statistical property of multi-class classifier

→ Statistical theory shows that:

A **multi-class** classifier with minimum **cross-entropy loss** estimates the probability ratios on the input classes:

$$g_i(\mathbf{x}) = \frac{p(\text{class} = i\,|\,\mathbf{x})}{\sum_{j=1}^{N_\text{out}} p(\text{class} = j\,|\,\mathbf{x})}$$

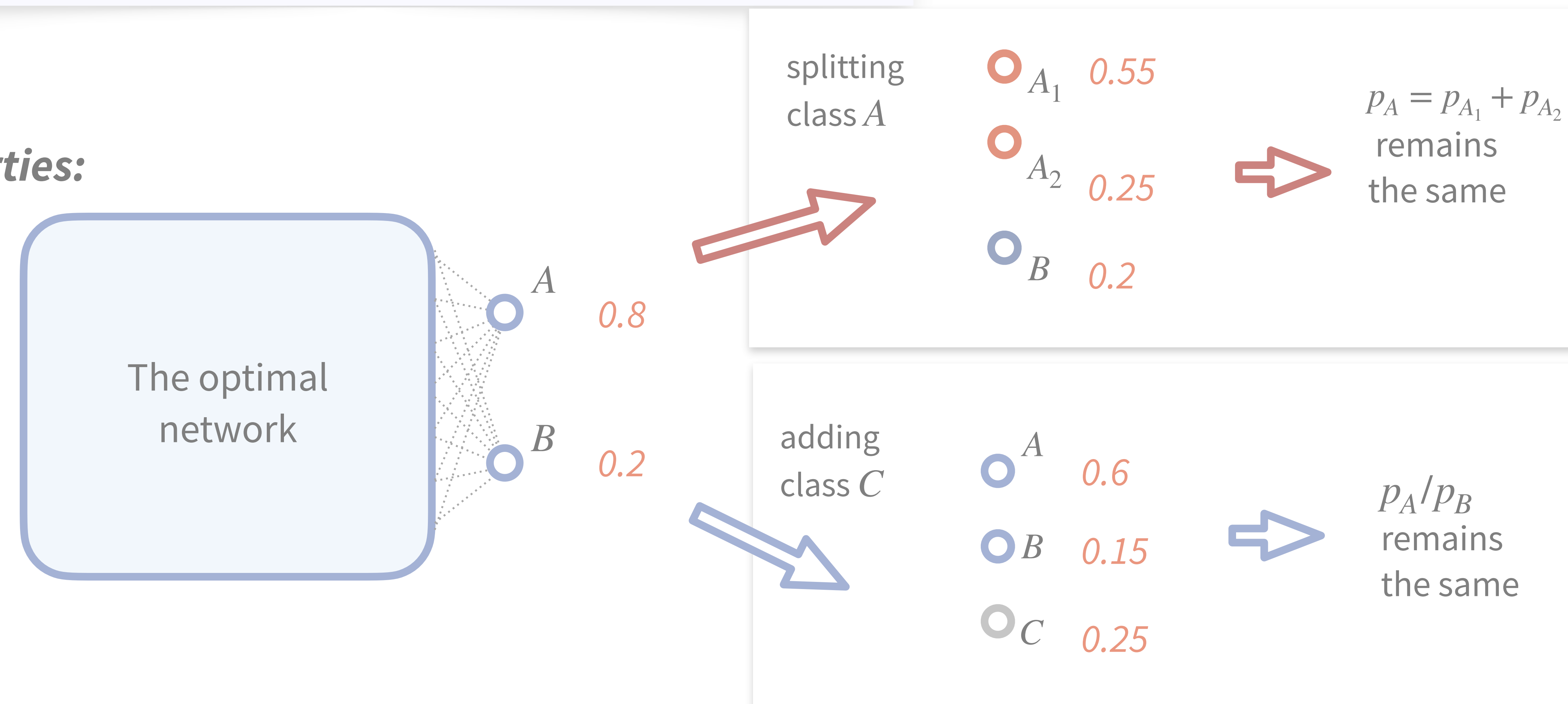
hence it contains all the information the ideal $N(N-1)$ binary classifiers can do

The key question in this context

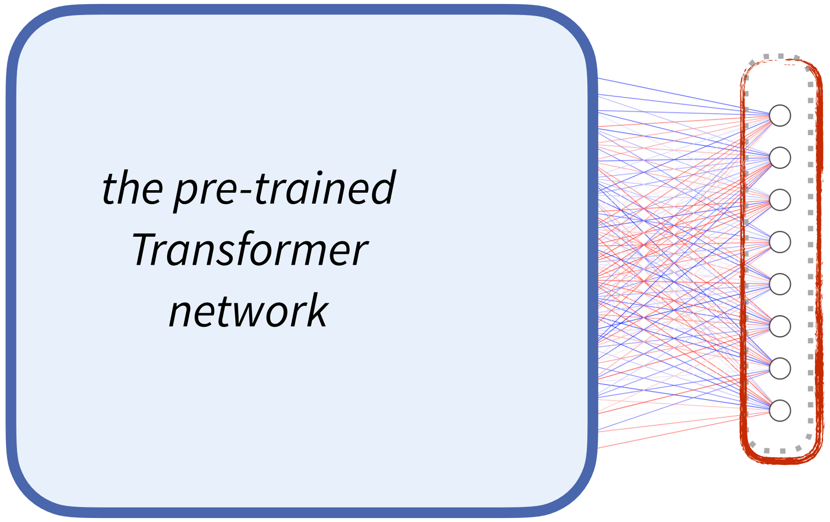
Does the model's capacity still enable us to reach the best achievable performance in existing tasks?

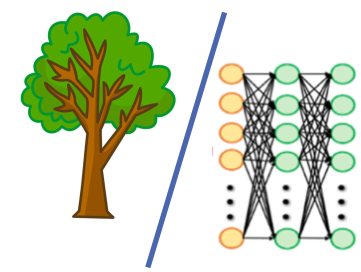
Our result will show: Yes.

Two properties:



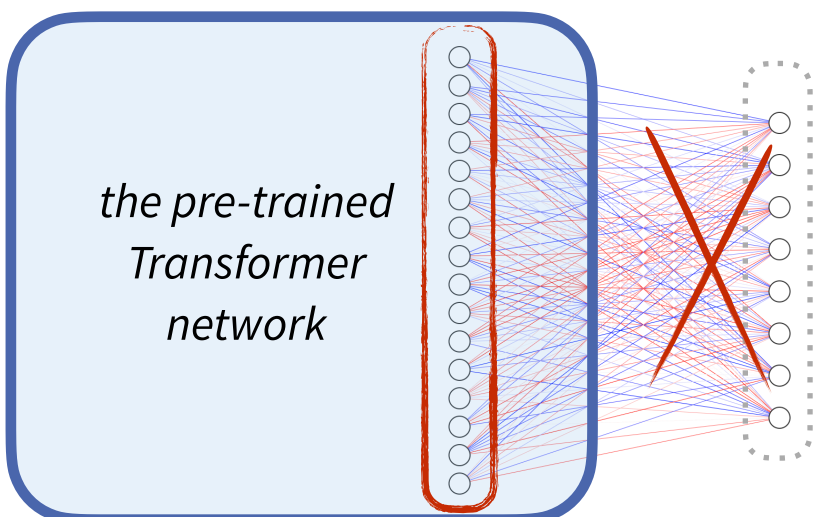
A glance into fine-tuning spirits



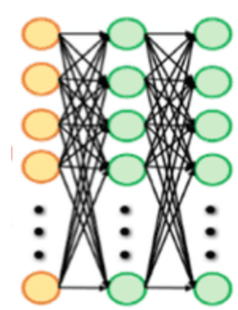


train a BDT or NN

customized scores!
(optimized for analysis)



use the
hidden layer



train an NN

equivalently, this means to replace then retrain the last layer

customized scores!
(optimized for analysis)

This is a fine-tuning approach (specifically, transfer learning) in its equivalent form

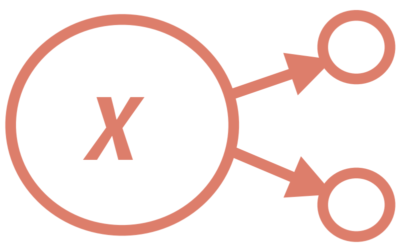
Introducing Sophon (智子)

arXiv:2405.12972
https://github.com/jet-universe/sophon

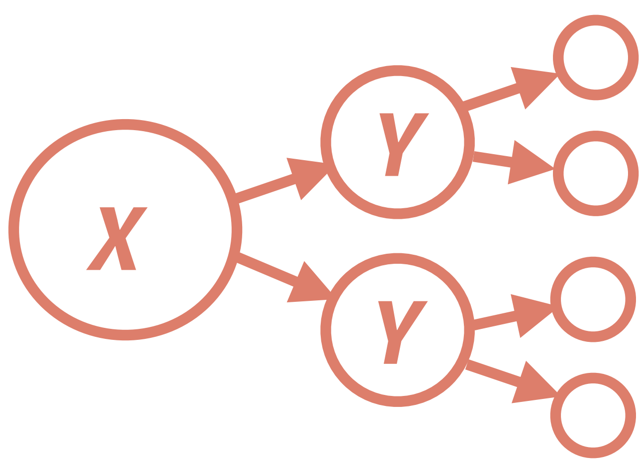
THREE-BODY

- ➔ **Signature-Oriented Pre-training for Heavy-resonant ObservatioN**
- ➔ the model is based on Particle Transformer (**ParT)** architecture
- ➔ a pre-trained model on a newly developed comprehensive dataset: **JetClass-II**
  - ***finely categorized labels:***

Resonant jet:
X→ 2 prong

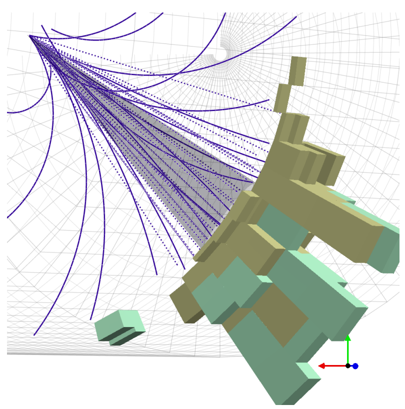


Resonant jet:
X→ 3/4 prong



bb/cc/ss/qq/gg/ee/μμ/ττ
bc/bq/cs/cq
eν/μν/τν

QCD jets



contributed final states:

bb/cc/ss/qq/gg/ee/μμ/ττ
bc/bq/cs/cq

all combination of Y decays, resulting to 4-prong or 3-prong

Key property: we do not focus on any specific *X* and *Y* masses
Their masses are variables: ranges from 20-500 GeV

Introducing Sophon (智子)

[arXiv:2405.12972](https://arxiv.org/abs/2405.12972)
https://github.com/jet-universe/sophon

→ **Signature-Oriented Pre-training for Heavy-resonant ObservatioN**

→ the model is based on Particle Transformer (**ParT)** architecture

→ a pre-trained model on a newly developed comprehensive dataset: **JetClass-II**

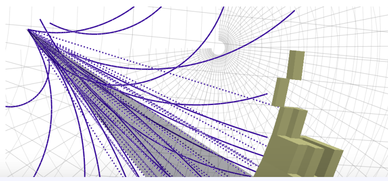
- ***finely categorized labels:***

Resonant jet:
X→ 2 prong

Resonant jet:
X→ 3/4 prong

QCD jets

bb/cc/ss/qq/gg/ee/μμ/ττ
bc/bq/cs/cq
eν/μν/τν



All final states!

Major types	Index range	Label names
Resonant jets: $X \to 2$ prong	0–14	bb, cc, ss, qq, bc, cs, bq, cq, sq, gg, ee, $\mu\mu$, $\tau_\mathrm{h}\tau_e$, $\tau_\mathrm{h}\tau_\mu$, $\tau_\mathrm{h}\tau_\mathrm{h}$
Resonant jets: $X \to 3$ or 4 prong	15–160	$bbbb$, $bbcc$, $bbss$, $bbqq$, $bbgg$, $bbee$, $bb\mu\mu$, $bb\tau_\mathrm{h}\tau_e$, $bb\tau_\mathrm{h}\tau_\mu$, $bb\tau_\mathrm{h}\tau_\mathrm{h}$, bbb, bbc, bbs, bbq, bbg, bbe, $bb\mu$, $cccc$, $ccss$, $ccqq$, $ccgg$, $ccee$, $cc\mu\mu$, $cc\tau_\mathrm{h}\tau_e$, $cc\tau_\mathrm{h}\tau_\mu$, $cc\tau_\mathrm{h}\tau_\mathrm{h}$, ccb, ccc, ccs, ccq, ccg, cce, $cc\mu$, $ssss$, $ssqq$, $ssgg$, $ssee$, $ss\mu\mu$, $ss\tau_\mathrm{h}\tau_e$, $ss\tau_\mathrm{h}\tau_\mu$, $ss\tau_\mathrm{h}\tau_\mathrm{h}$, ssb, ssc, sss, ssq, ssg, sse, $ss\mu$, $qqqq$, $qqgg$, $qqee$, $qq\mu\mu$, $qq\tau_\mathrm{h}\tau_e$, $qq\tau_\mathrm{h}\tau_\mu$, $qq\tau_\mathrm{h}\tau_\mathrm{h}$, qqb, qqc, qqs, qqq, qqg, qqe, $qq\mu$, $gggg$, $ggee$, $gg\mu\mu$, $gg\tau_\mathrm{h}\tau_e$, $gg\tau_\mathrm{h}\tau_\mu$, $gg\tau_\mathrm{h}\tau_\mathrm{h}$, ggb, ggc, ggs, ggq, ggg, gge, $gg\mu$, bee, cee, see, qee, gee, $b\mu\mu$, $c\mu\mu$, $s\mu\mu$, $q\mu\mu$, $g\mu\mu$, $b\tau_\mathrm{h}\tau_e$, $c\tau_\mathrm{h}\tau_e$, $s\tau_\mathrm{h}\tau_e$, $q\tau_\mathrm{h}\tau_e$, $g\tau_\mathrm{h}\tau_e$, $b\tau_\mathrm{h}\tau_\mu$, $c\tau_\mathrm{h}\tau_\mu$, $s\tau_\mathrm{h}\tau_\mu$, $q\tau_\mathrm{h}\tau_\mu$, $g\tau_\mathrm{h}\tau_\mu$, $b\tau_\mathrm{h}\tau_\mathrm{h}$, $c\tau_\mathrm{h}\tau_\mathrm{h}$, $s\tau_\mathrm{h}\tau_\mathrm{h}$, $q\tau_\mathrm{h}\tau_\mathrm{h}$, $g\tau_\mathrm{h}\tau_\mathrm{h}$, $qqqb$, $qqqc$, $qqqs$, $bbcq$, $ccbs$, $ccbq$, $ccsq$, $sscq$, $qqbc$, $qqbs$, $qqcs$, $bcsq$, bcs, bcq, bsq, csq, $bce\nu$, $cse\nu$, $bqe\nu$, $cqe\nu$, $sqe\nu$, $qqe\nu$, $bc\mu\nu$, $cs\mu\nu$, $bq\mu\nu$, $cq\mu\nu$, $sq\mu\nu$, $qq\mu\nu$, $bc\tau_e\nu$, $cs\tau_e\nu$, $bq\tau_e\nu$, $cq\tau_e\nu$, $sq\tau_e\nu$, $qq\tau_e\nu$, $bc\tau_\mu\nu$, $cs\tau_\mu\nu$, $bq\tau_\mu\nu$, $cq\tau_\mu\nu$, $sq\tau_\mu\nu$, $qq\tau_\mu\nu$, $bc\tau_\mathrm{h}\nu$, $cs\tau_\mathrm{h}\nu$, $bq\tau_\mathrm{h}\nu$, $cq\tau_\mathrm{h}\nu$, $sq\tau_\mathrm{h}\nu$, $qq\tau_\mathrm{h}\nu$
QCD jets	161–187	$bbccss$, $bbccs$, $bbcc$, $bbcss$, $bbcs$, bbc, $bbss$, bbs, bb, $bccss$, $bccs$, bcc, $bcss$, bcs, bc, bss, bs, b, $ccss$, ccs, cc, css, cs, c, ss, s, others

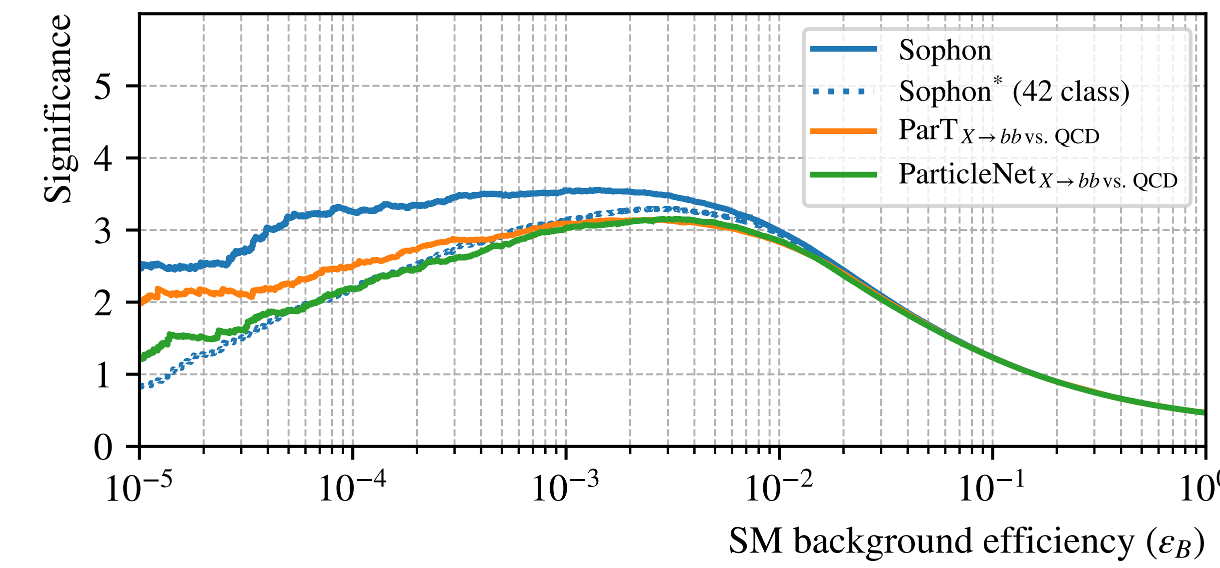
Sophon: performance benchmark

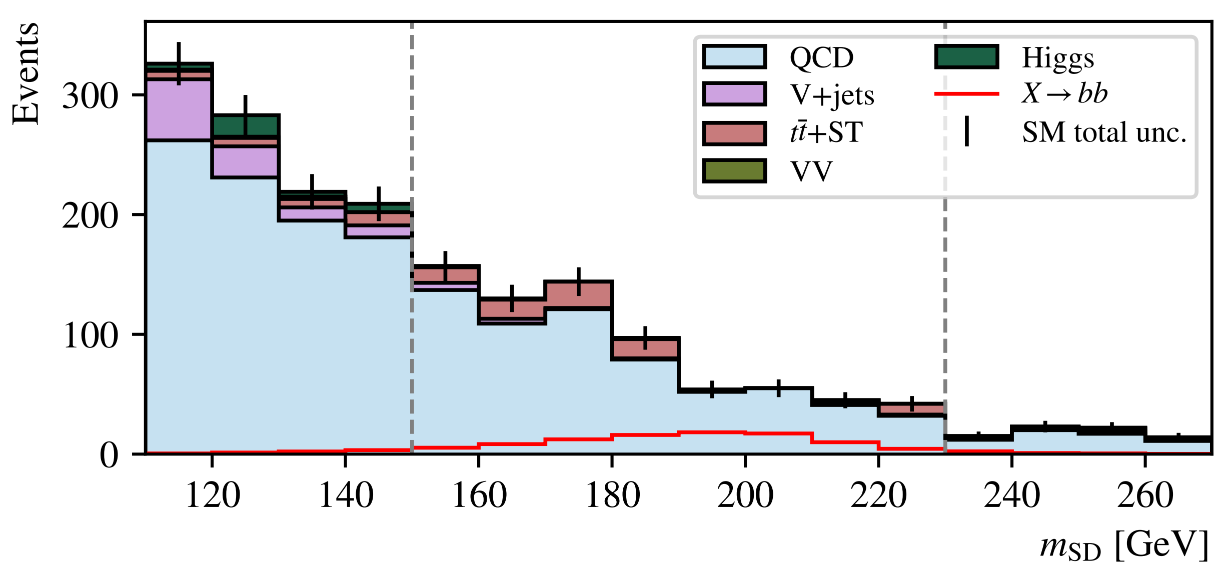
arXiv:2405.12972

Search significance:

$Z = \sqrt{2((s+b)\log(1+s/b)-s)}$

Direct tagging ability





- Apply tagger selection
- Check discrimination power of
  X (200 GeV) → **bb** signal vs. all backgrounds

- **Sophon** (training on 188 classes) has best performance

$$\mathrm{discr}(X \to bb\ \text{vs. QCD}) = \frac{g_{X\to bb}}{g_{X\to bb} + \sum_{l=1}^{27} g_{\mathrm{QCD}_l}}$$

- Performance gain does come from large-scale classification (compared to **Sophon* (42 classes)**)
- **ParT** and **ParticleNet** for binary classification: they represent the best performance we can reach in experiment now

Sophon: performance benchmark

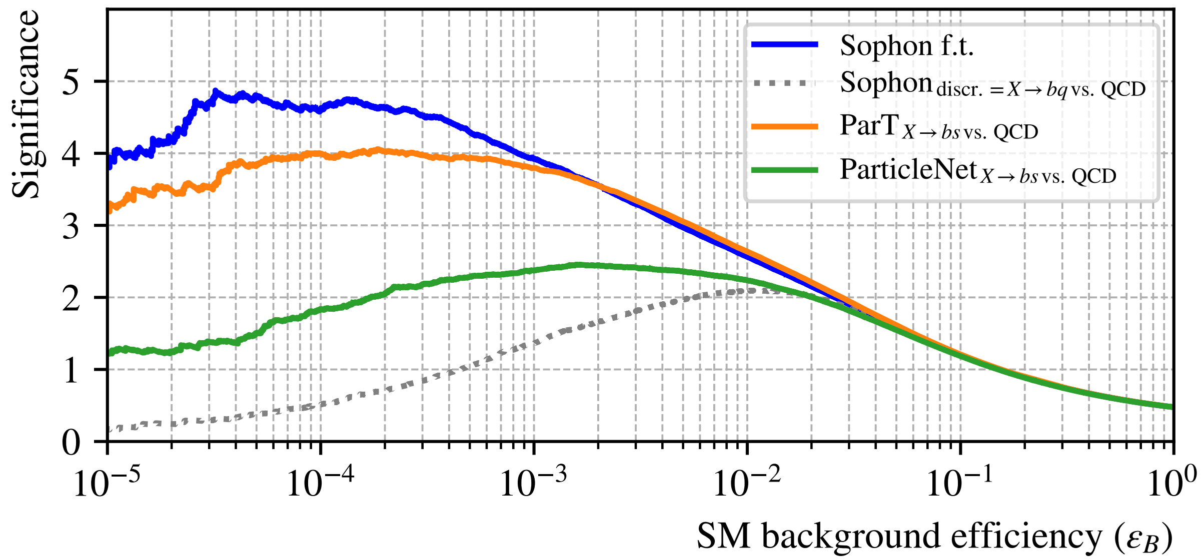
arXiv:2405.12972

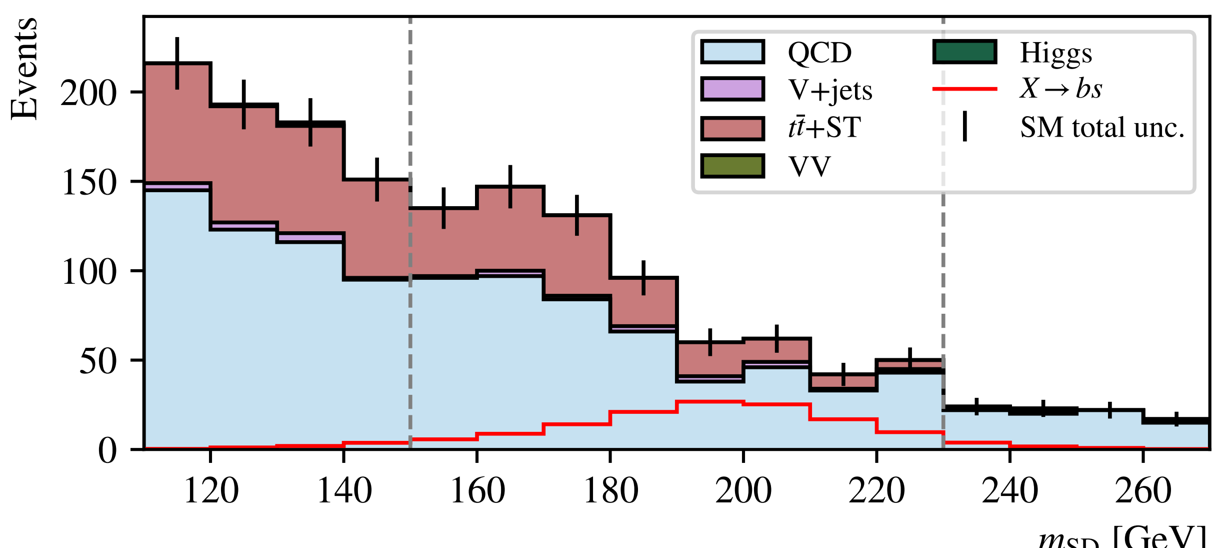
Search significance:

$Z = \sqrt{2((s+b)\log(1+s/b)-s)}$

Transfer learning ability

(adapt the tagger to a new classification task)





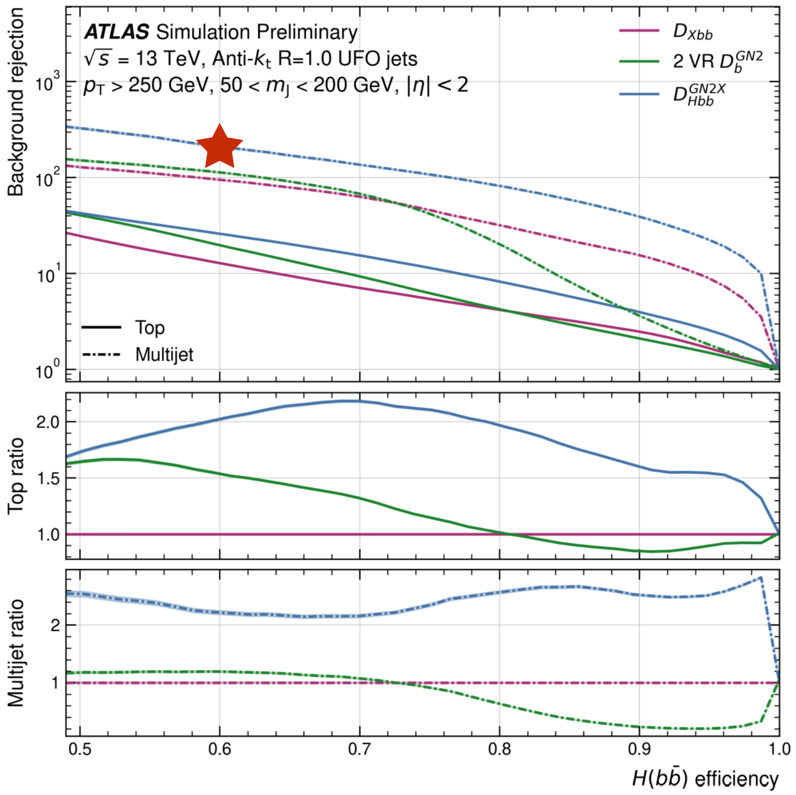
- Check discrimination power of X (200 GeV) → **bs** signal vs. all backgrounds

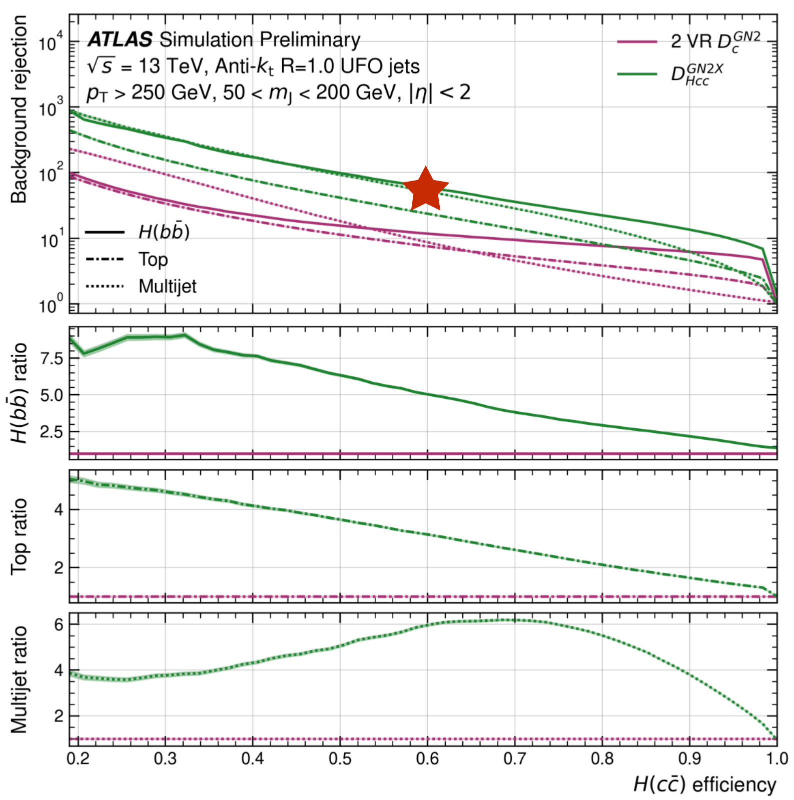
- **Sophon** (training on 188 classes) reaches the best performance **after fine-tuned (via transfer learning)**
- **ParT** and **ParticleNet** for binary X→bs vs QCD classification: they reveal the best performance we can reach in the experiment now

Sophon: close to real experimental performance?

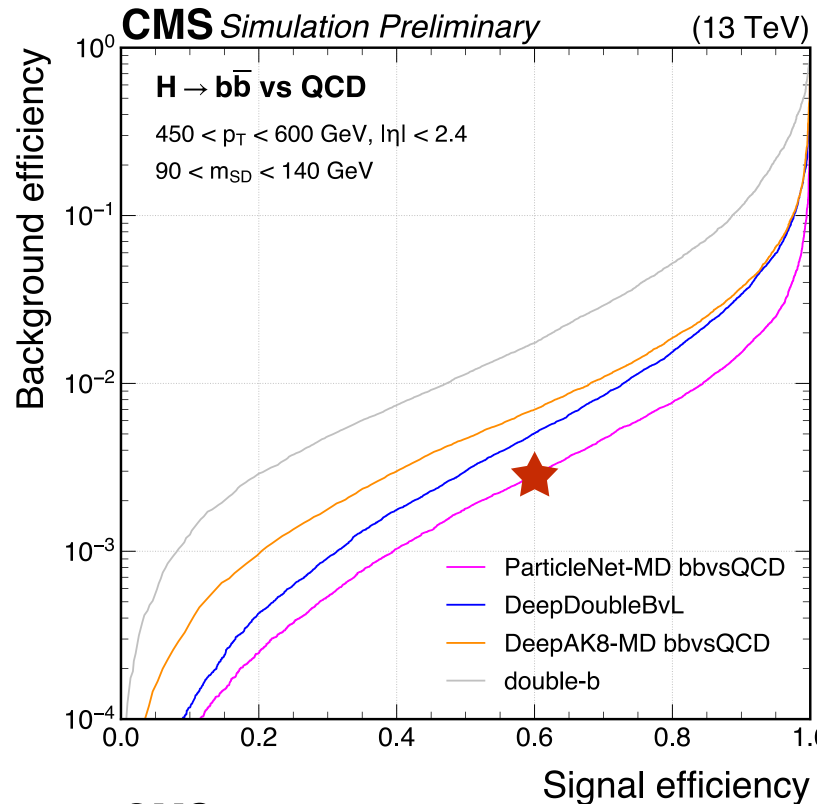
★ marks
QCD BKG rej at signal eff. = 60%

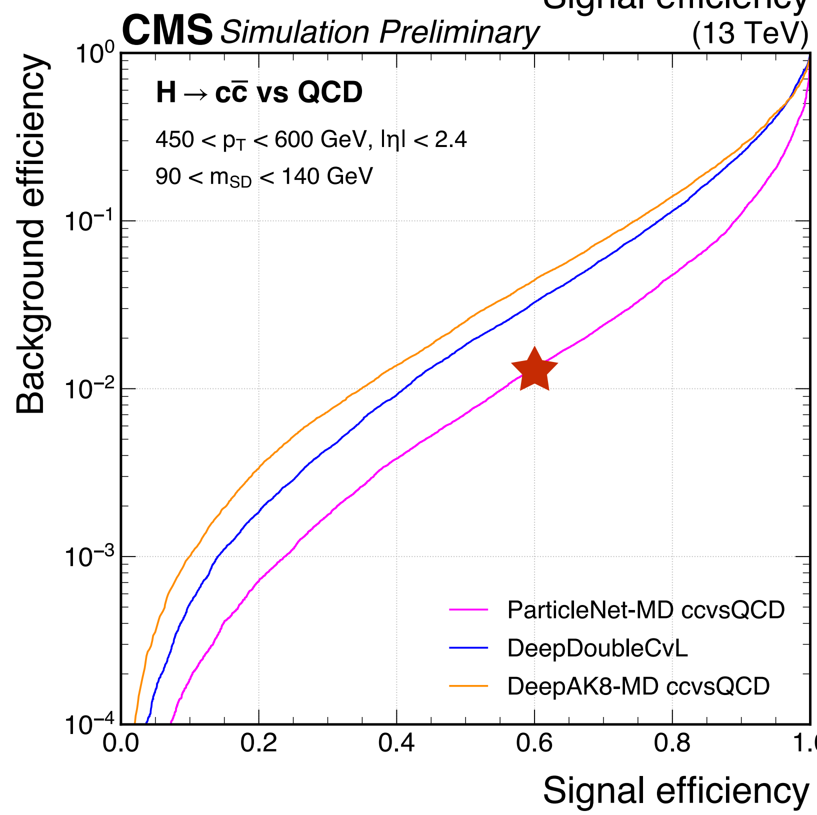
ATLAS results ATL-PHYS-PUB-2023-021



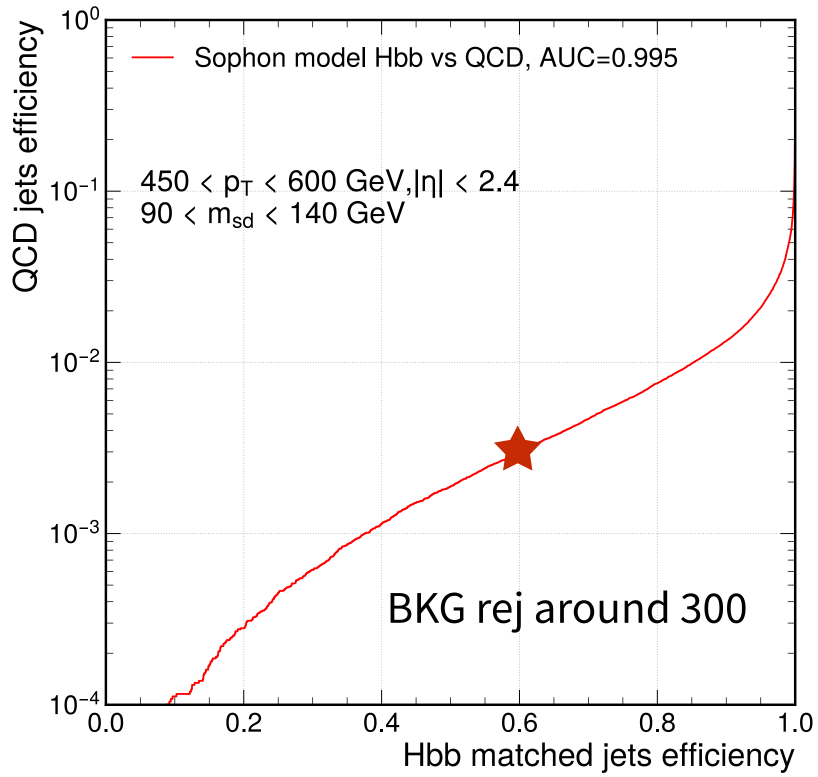


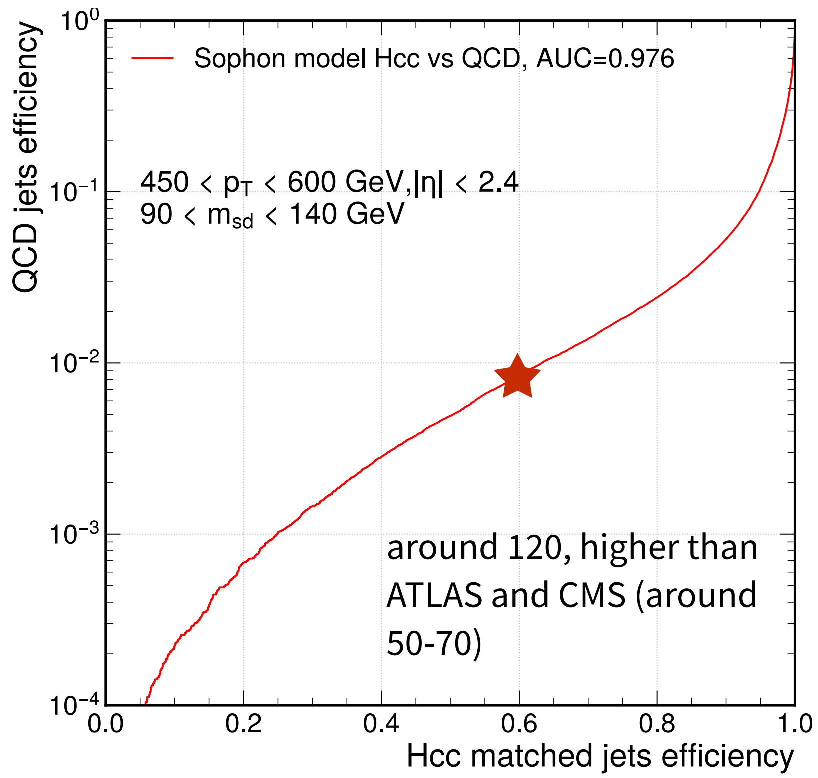
CMS results CMS-PAS-BTV-22-001





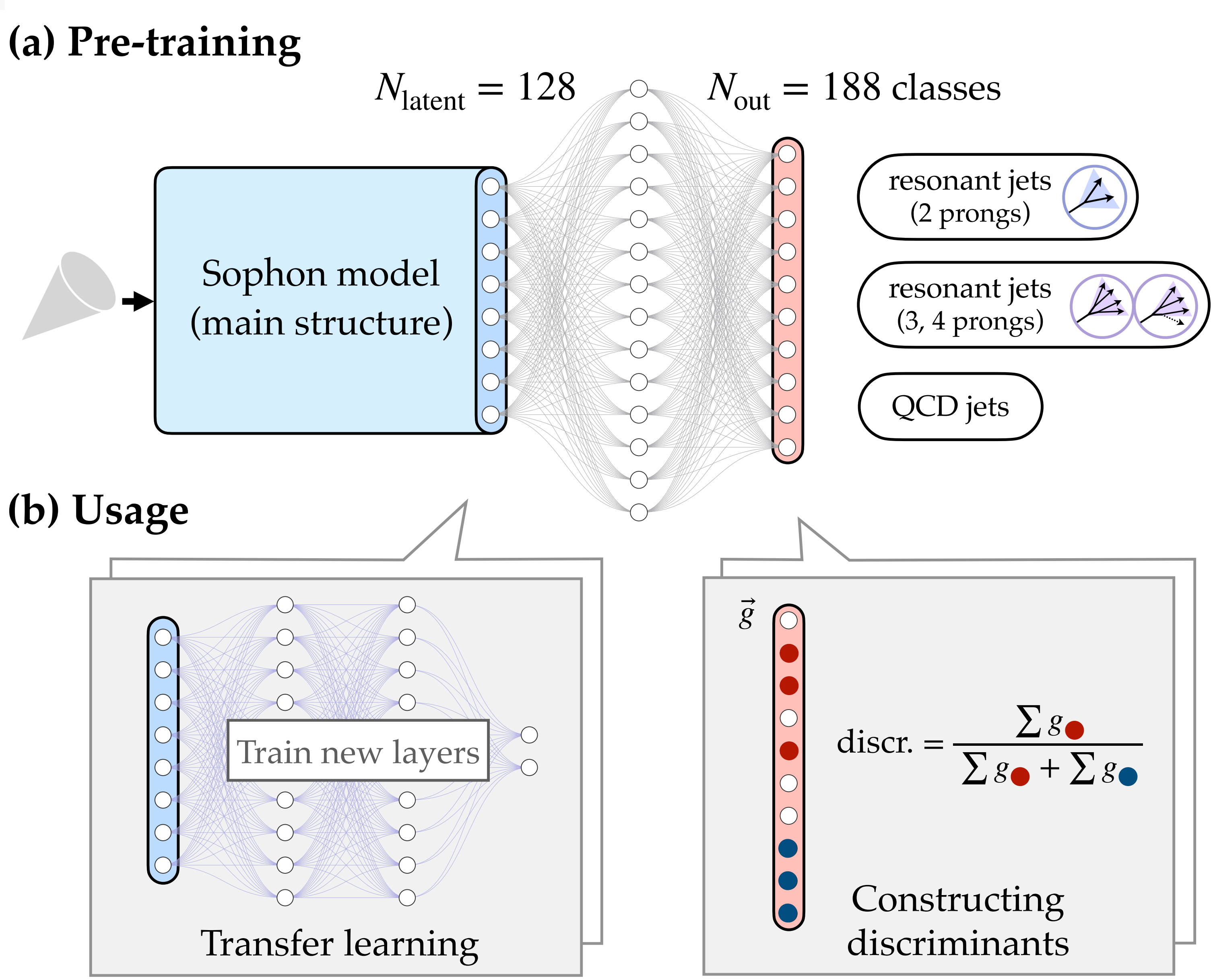
Sophon results
(performance on Delphes)



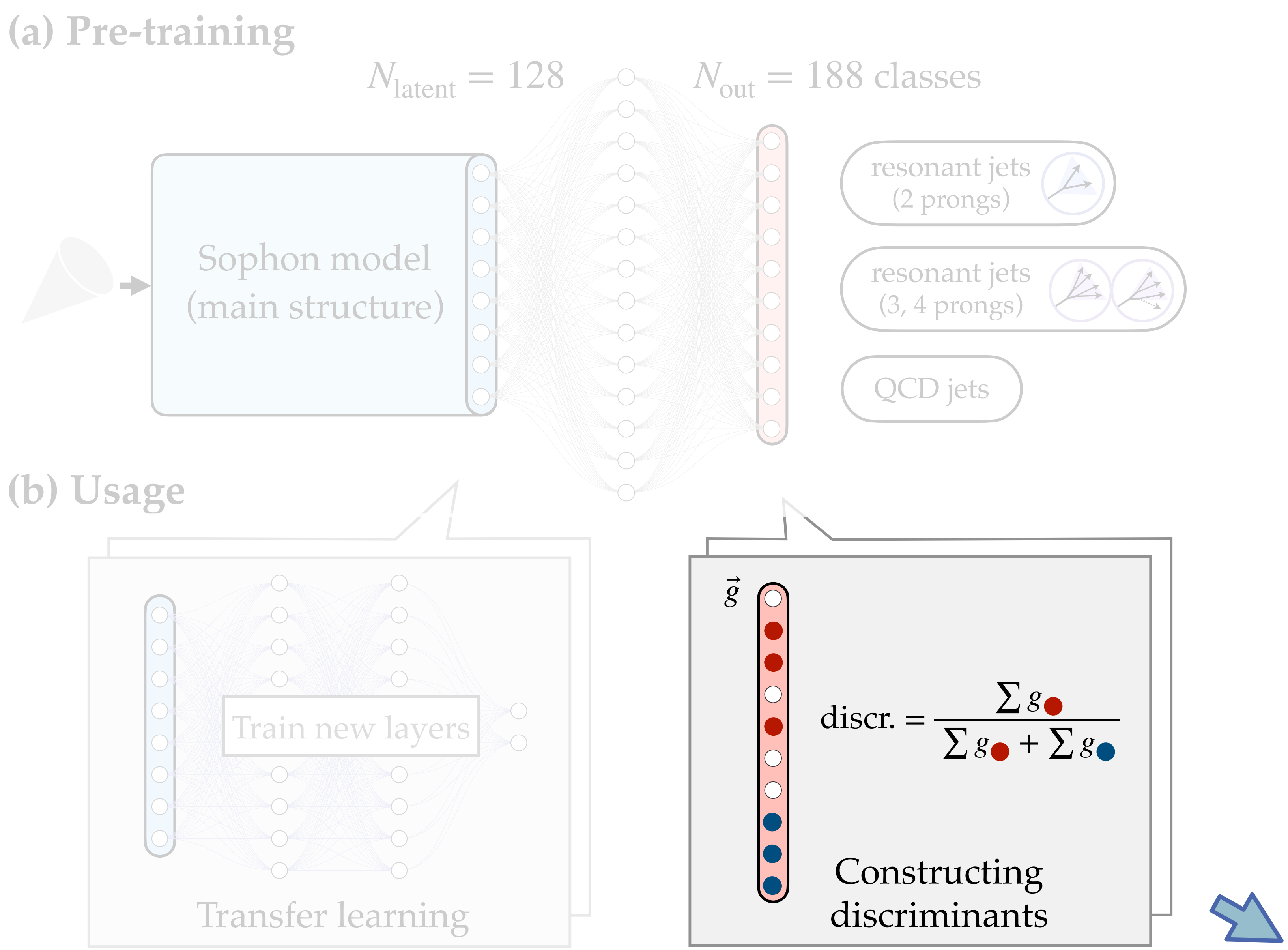


Implications for LHC resonance search

Using Sophon



Using Sophon



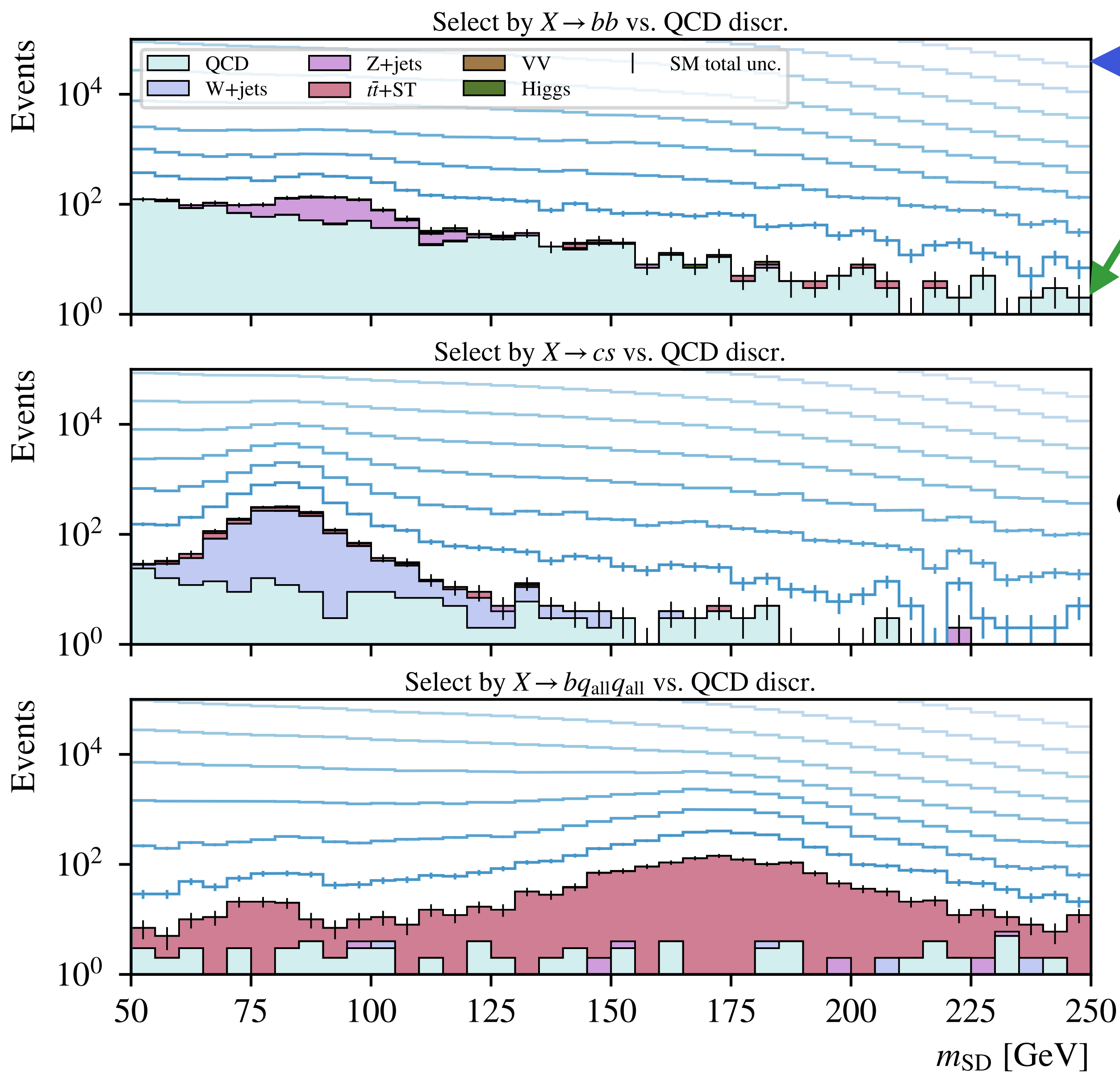
Use it out of the box!

Construct a dedicated discr.
→ perform a bump hunt

Can we rediscover the SM particles?

➔ Simulate 40fb-1 LHC collision events, √s = 13 TeV, nPU=50

- focus on the large-*R* jet trigger (triggered with Σp_T threshold and trimmed mass)
- abundant QCD backgrounds
- **rediscover Z/W/t particles** simply from the large-*R* jet's **mass spectrum**



Without selection

Select at eff. = 1e-4

- Select by Sophon's different discriminants

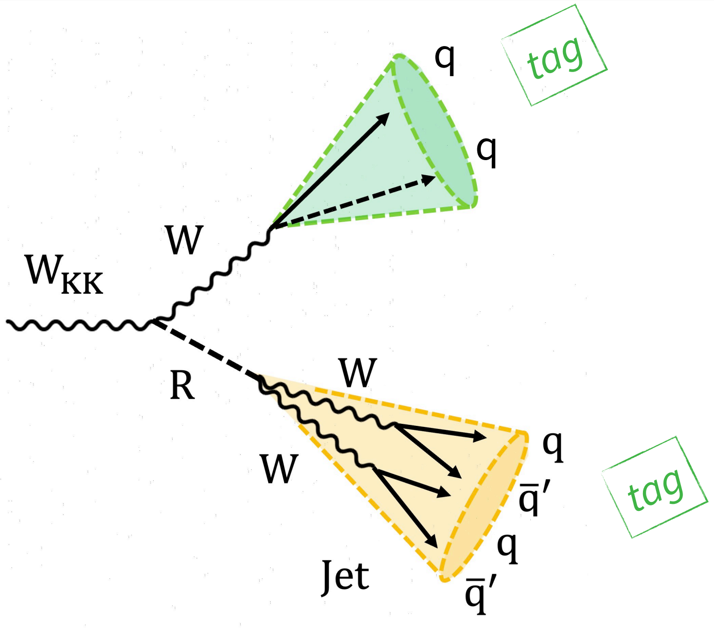
$$\text{discr} = \frac{g_A}{g_A + \sum_{l=1}^{27} g_{\text{QCD}_l}} \quad \begin{cases} ①: A = \{bb\} \\ ②: A = \{cs\} \\ ③: A = \{ccb, ssb, qqb, bcs, bcq, bsq\} \end{cases}$$

More heavy resonances

➔ Consider triboson signal:

W' ($m_{W'}$ = 3 TeV) → $W\phi$ (m_ϕ = 400 GeV) → WWW (fully hadronic decays)

➔ Optimize an event-level discr. from tagger discr.

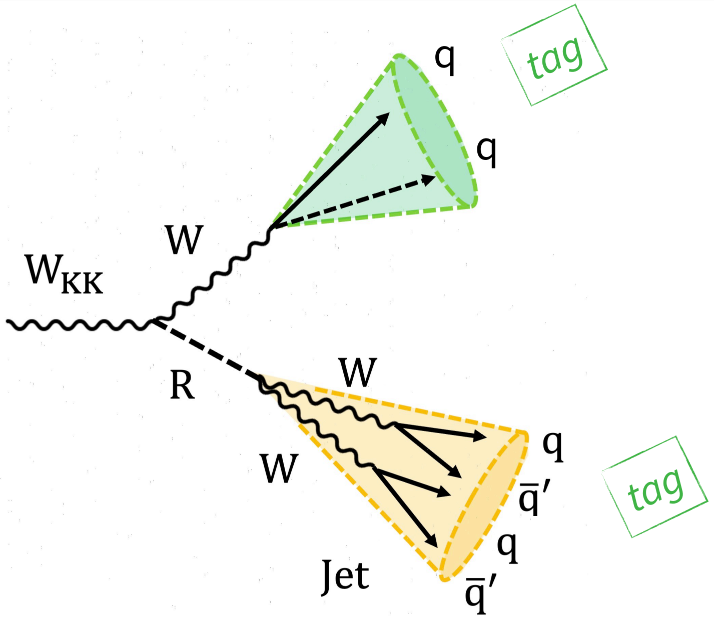


$$\text{discr} = \sum_{\text{jet}=1,2} \frac{g_{A,\text{jet}}}{g_{A,\text{jet}} + \sum_{l=1}^{27} g_{\text{QCD}_l,\text{jet}}} \quad \text{(sum for jets 1, 2)}$$

$$A = \begin{cases} 0.3 \times \{cs, qq\} \\ + 0.1 \times \{ccss, qqcs, qqqq\} \\ + 0.6 \times \{ccs, ccq, ssc, ssq, qqc, qqs, qqq\} \end{cases}$$

More heavy resonances

- ➔ Consider triboson signal:
  W' ($m_{W'}$ = 3 TeV) → $W\phi$ (m_ϕ = 400 GeV) → WWW (fully hadronic decays)
- ➔ Optimize an event-level discr. from tagger discr.



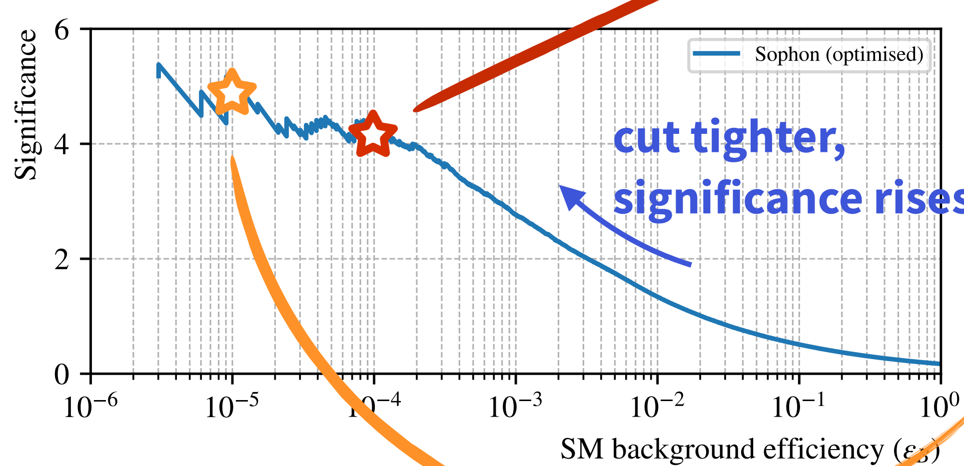
$$\text{discr} = \sum_{\text{jet}=1,2} \frac{g_{A,\text{jet}}}{g_{A,\text{jet}} + \sum_{l=1}^{27} g_{\text{QCD}_l,\text{jet}}}$$ (sum for jets 1, 2)

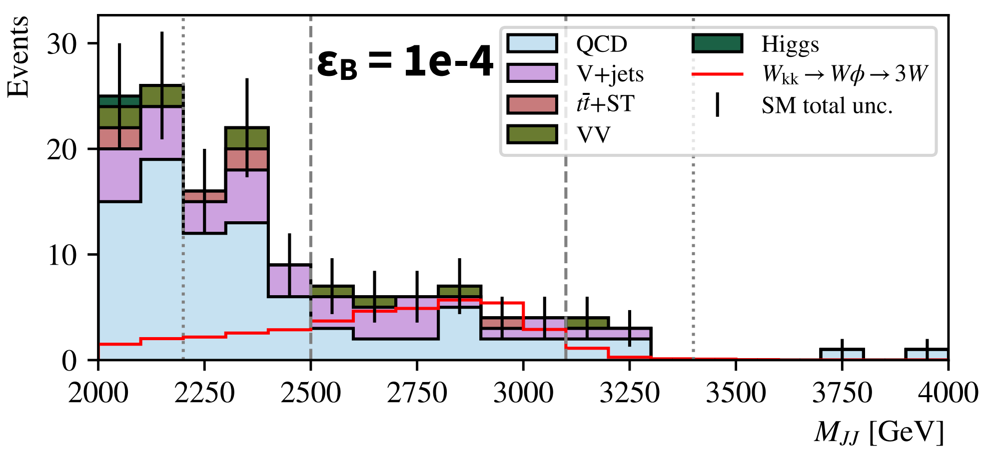
$$A = \begin{cases} 0.3 \times \{cs, qq\} \\ +\ 0.1 \times \{ccss, qqcs, qqqq\} \\ +\ 0.6 \times \{ccs, ccq, ssc, ssq, qqc, qqs, qqq\} \end{cases}$$

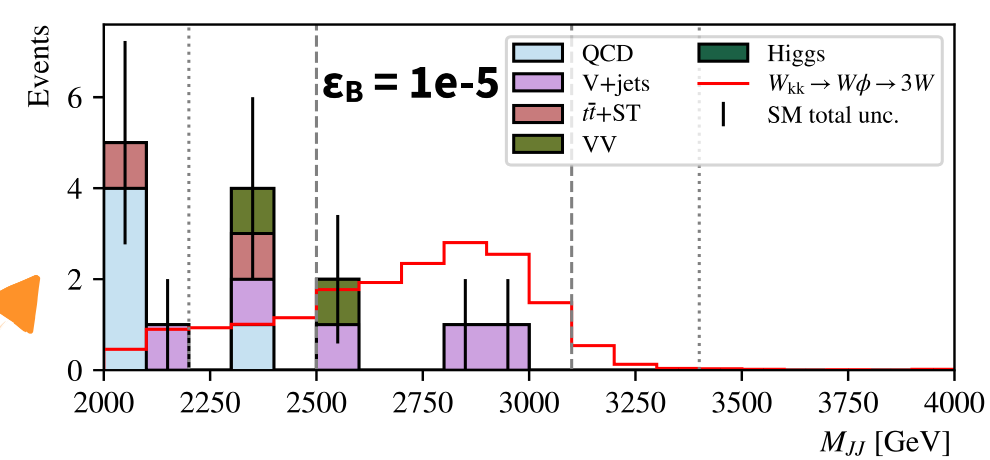
Search significance

$$Z = \sqrt{2((s+b)\log(1+s/b) - s)}$$

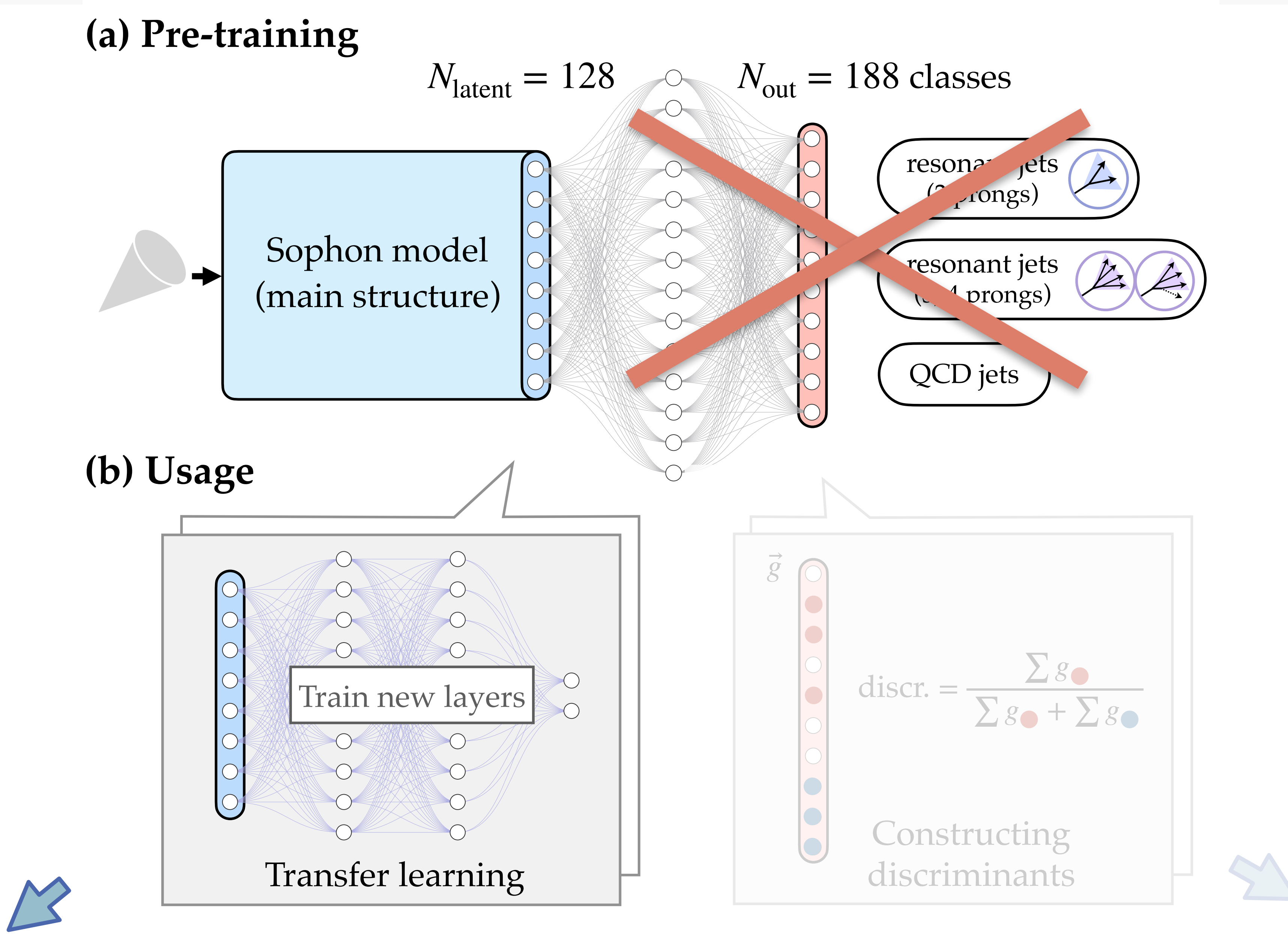
in dijet inv. mass window
2500–3100 GeV







Sophon's transfer learning



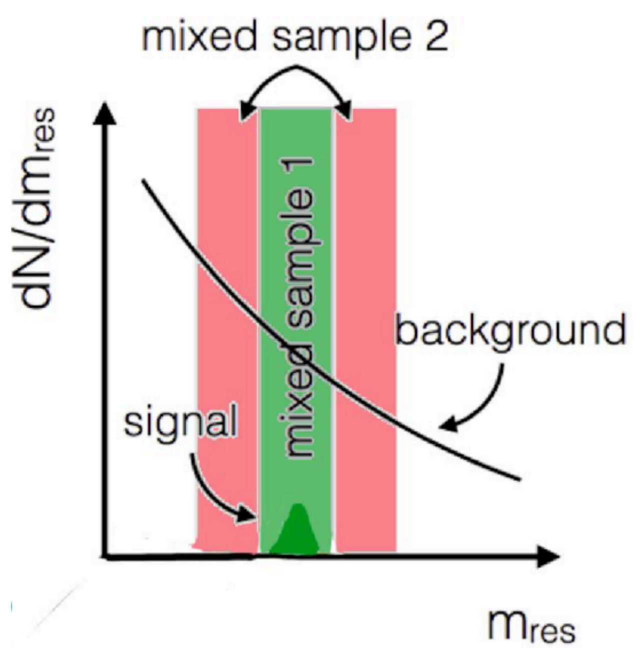
Use it out of the box!

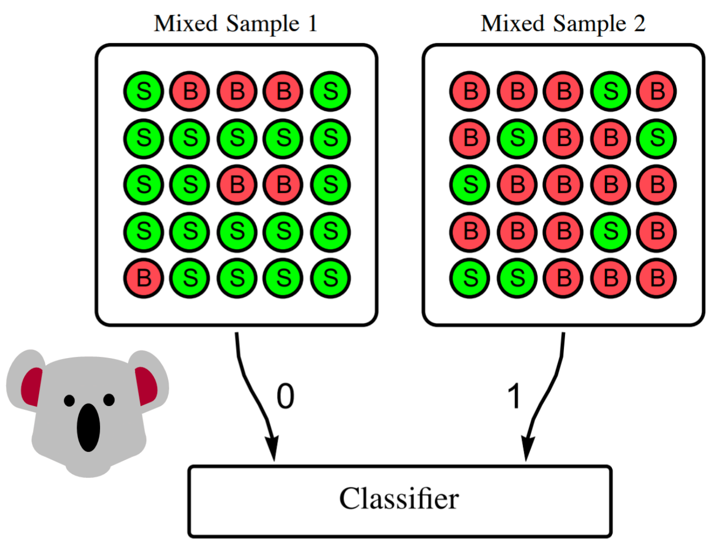
- Transfer to uncovered tagging scenarios…
- facilitate anomaly detection (weakly-supervised, autoencoder)…
- *more potential to unlock!*

Construct a dedicated discr.
→ perform a bump hunt

Background: anomaly detection in weakly-supervised approach

JHEP 10 (2017) 174



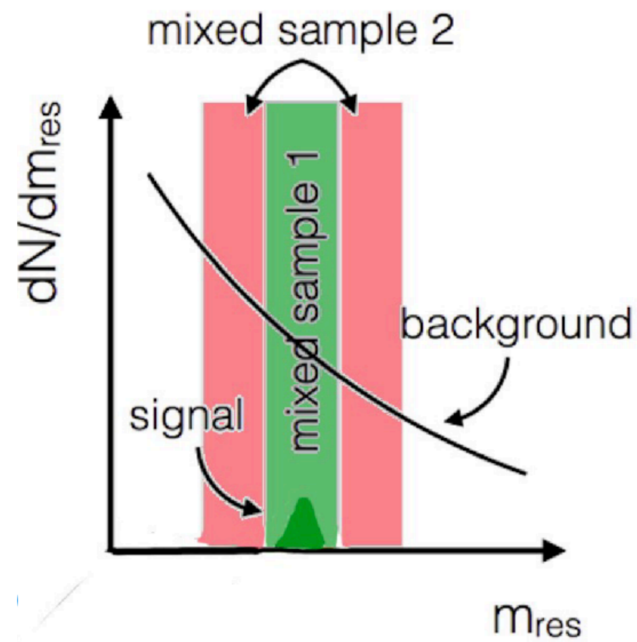


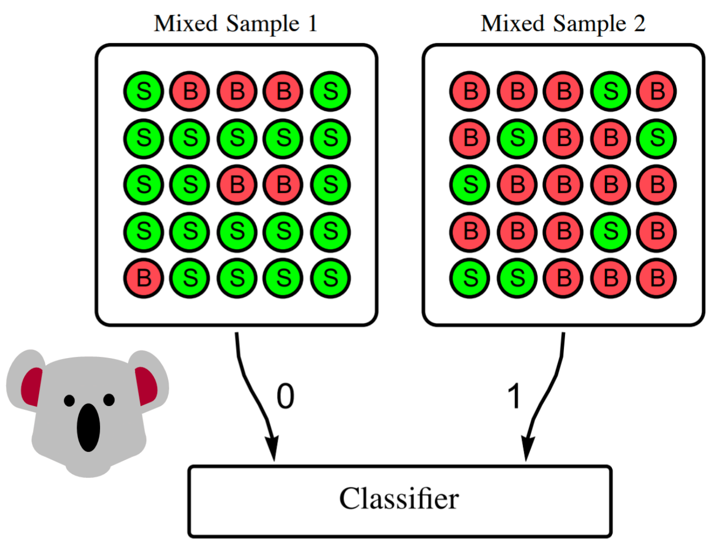
Equivalent effect for training **S** vs **B**

- Recall the early work: CWoLa (classification without labels) Hunting
  - allow to detect anomalies purely from data
  - train a classifier for mass window vs mass sideband (mixed sample 1 vs 2)
  - many improved approaches in recent years → very active field

Background: anomaly detection in weakly-supervised approach

JHEP 10 (2017) 174

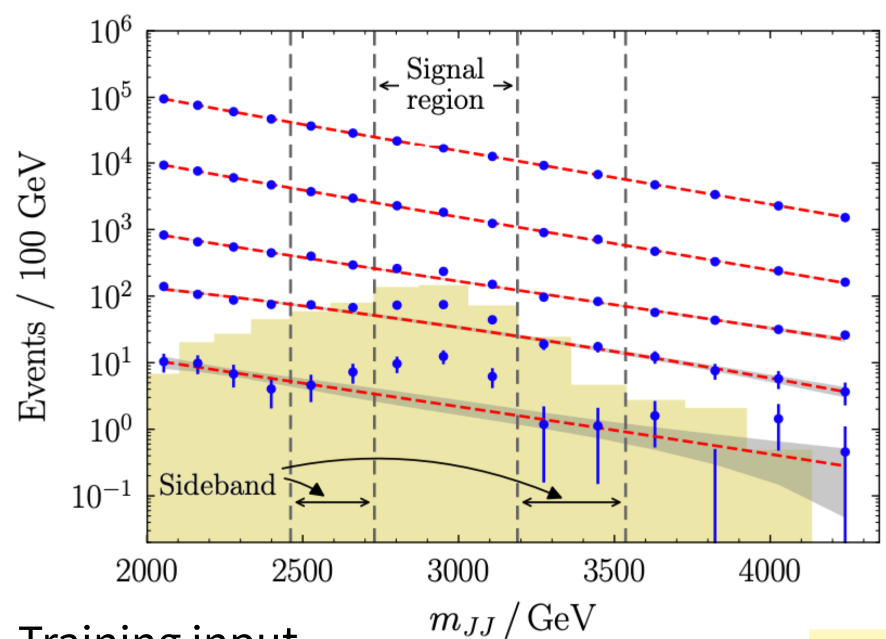




Equivalent effect for training **S** vs **B**

➔ Recall the early work: CWoLa (classification without labels) Hunting

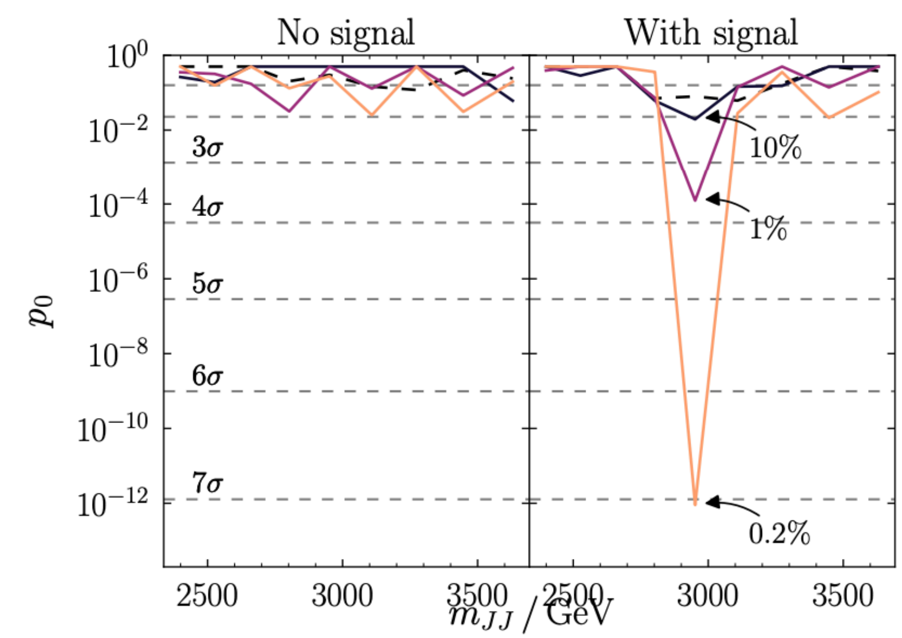
- allow to detect anomalies purely from data
- train a classifier for mass window vs mass sideband (mixed sample 1 vs 2)
- many improved approaches in recent years → very active field



Training input

$m_J, \sqrt{\tau_1^{(2)}/\tau_1^{(1)}}, \tau_{21}, \tau_{32}, \tau_{43}, n_{\mathrm{trk}},$

can discover $W' \to W\phi \to WWW$ signals
see 2σ → 7σ improvement

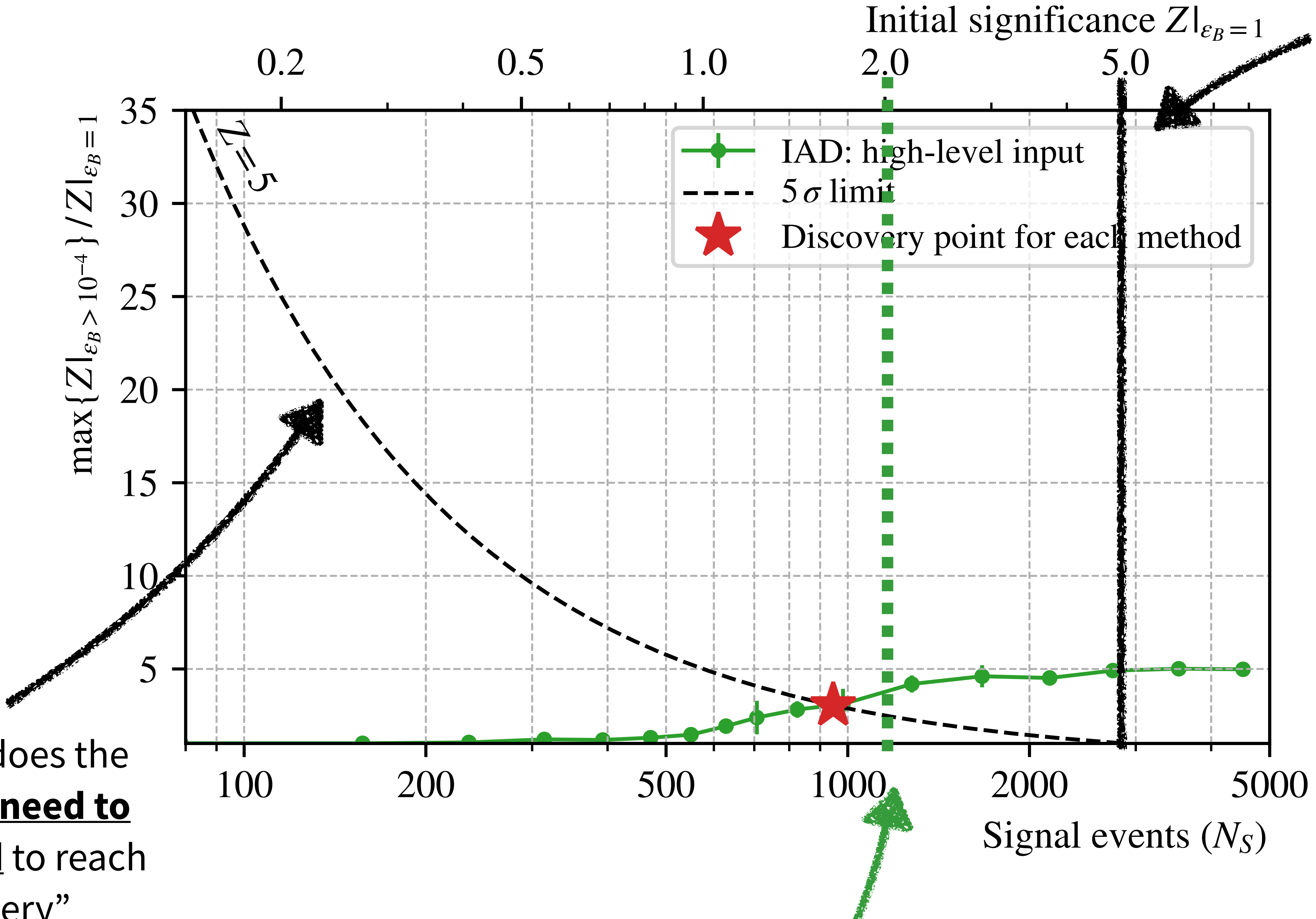


PRL, 121 (2018) 24, 241803

PRD, 99 (2019) 1, 014038

Dijet search capabilities

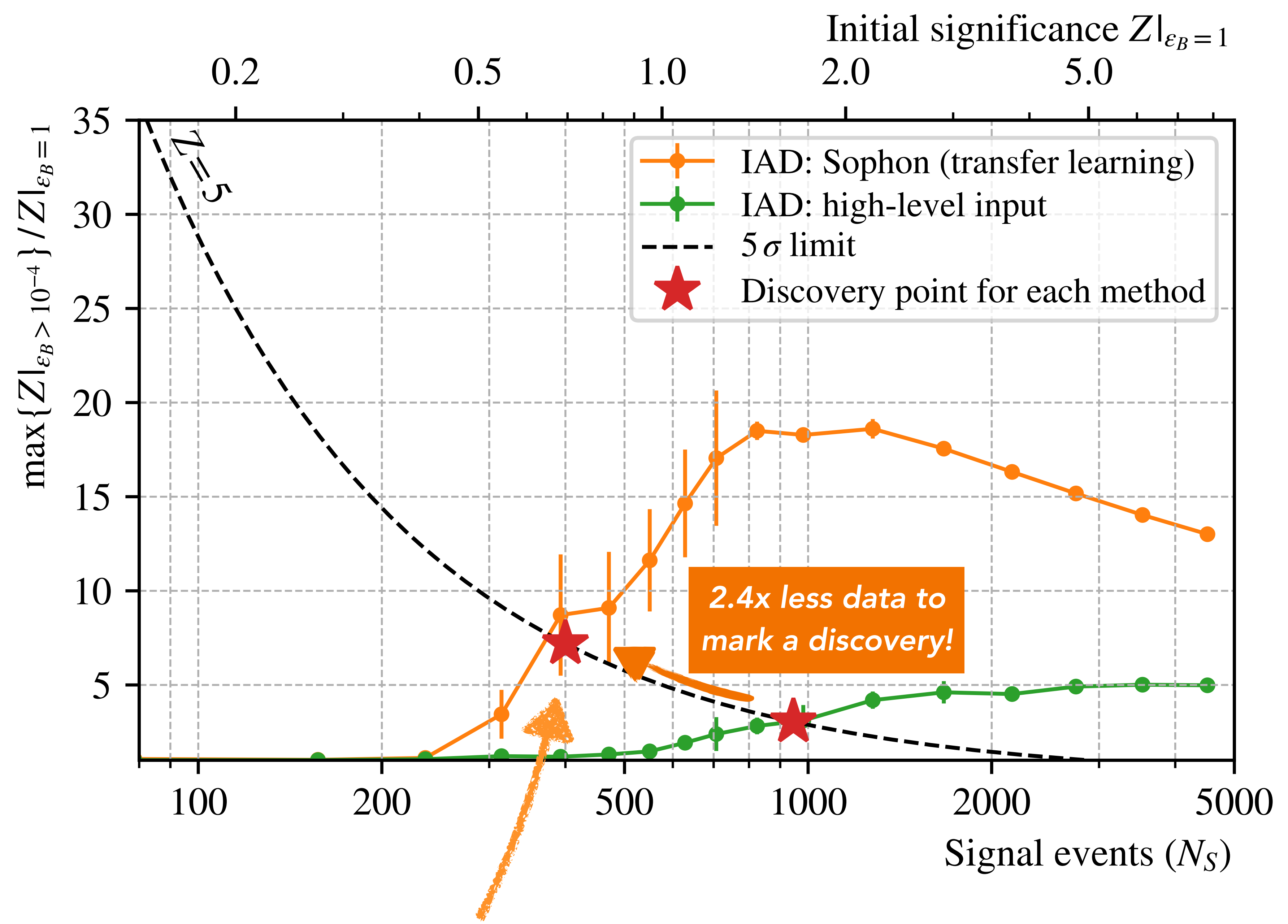
"If signal events reach this point, **with initial *Z*=5**, then we have already discovered the signal without needing to make a cut"



"How much does the **significance need to be increased** to reach the 5σ discovery"

a similar 2σ→7σ is reached with conventional AD approach; roughly reproduce the result in *PRL, 121 (2018) 24, 241803*

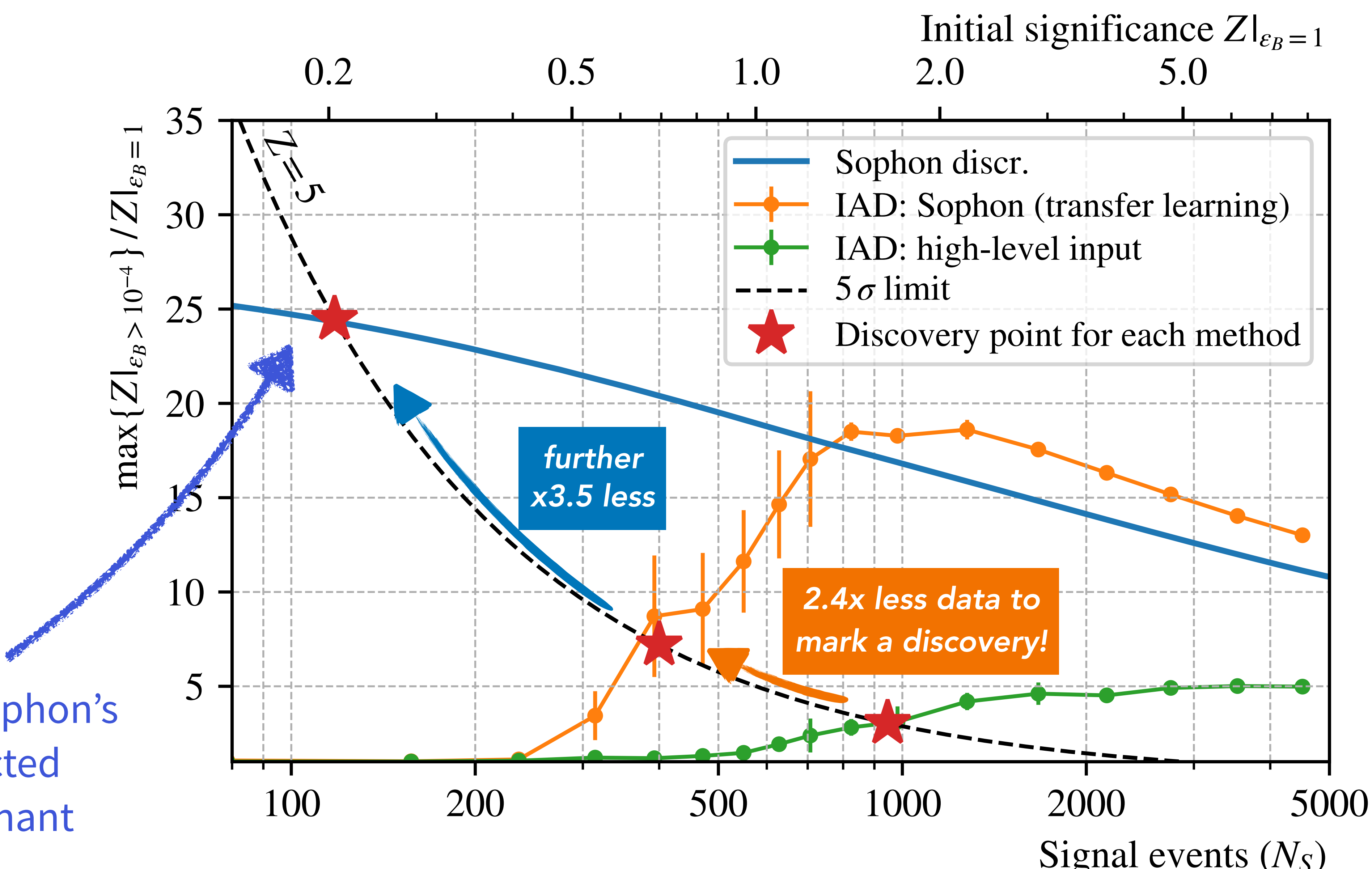
Dijet search capabilities



Combining Sophon's transfer learning (using Sophon's "knowledge") with AD marks a success

- More sensitive at low signal injection (**even starting at ~0.6σ**)
- Much improved S vs B distinguishability than using high-level input

Dijet search capabilities



using Sophon's constructed discriminant

$$\text{discr} = \sum_{\text{jet}=1,2} \frac{g_{A,\text{jet}}}{g_{A,\text{jet}} + \sum_{l=1}^{27} g_{\text{QCD}_l,\text{jet}}}$$

$$A = \begin{cases} 0.3 \times \{cs,\ qq\} \\ +\ 0.1 \times \{ccss,\ qqcs,\ qqqq\} \\ +\ 0.6 \times \{ccs,\ ccq,\ ssc,\ ssq,\ qqc,\ qqs,\ qqq\} \end{cases}$$

CMS's path to develop Global Particle Transformer

Philosophy to develop Global Particle Transformer (GloParT) in CMS

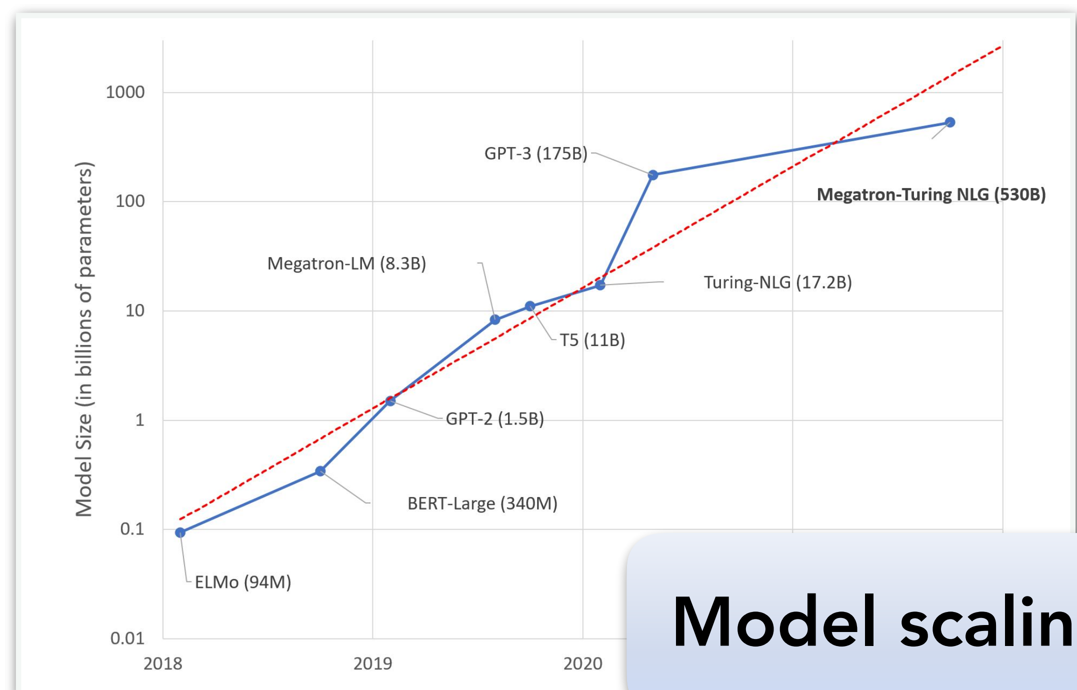
Good probability density estimators

- What is p? - the "differential cross section" of a process A on very high-dim space
- discriminating process A vs. B: estimate $p_A(\mathbf{x})/p_B(\mathbf{x})$ as best as we can
- need a model to **cover a variety of processes** $A, B, C, D, \ldots$.

$A \to BC$	$B=$ SM: e	μ	τ	q/g	b	t	γ	Z/W	H	$B=$ BSM
$C=$ SM: e	Z'	$\not{R}$	$\not{R}$	LQ	LQ	LQ	L^*	L^*	L^*	Many
μ		Z'	$\not{R}$	LQ	LQ	LQ	L^*	L^*	L^*	
τ			Z'	LQ	LQ	LQ	L^*	L^*	L^*	
q/g				Z'	W'	T'	Q^*	Q^*	Q'	
b					Z'	W'	Q^*	Q^*	B'	
t						Z'	Q^*	T'	T'	
γ							H	H	Z_{KK}	
Z/W								H	$H^{\pm}/A$	
H									H	
$C=$ BSM	Consider just the di-object search for resonant A → B C									Many

J.Kim *et al.* JHEP 04 (2020) 30 1907.06659

Generalization ability

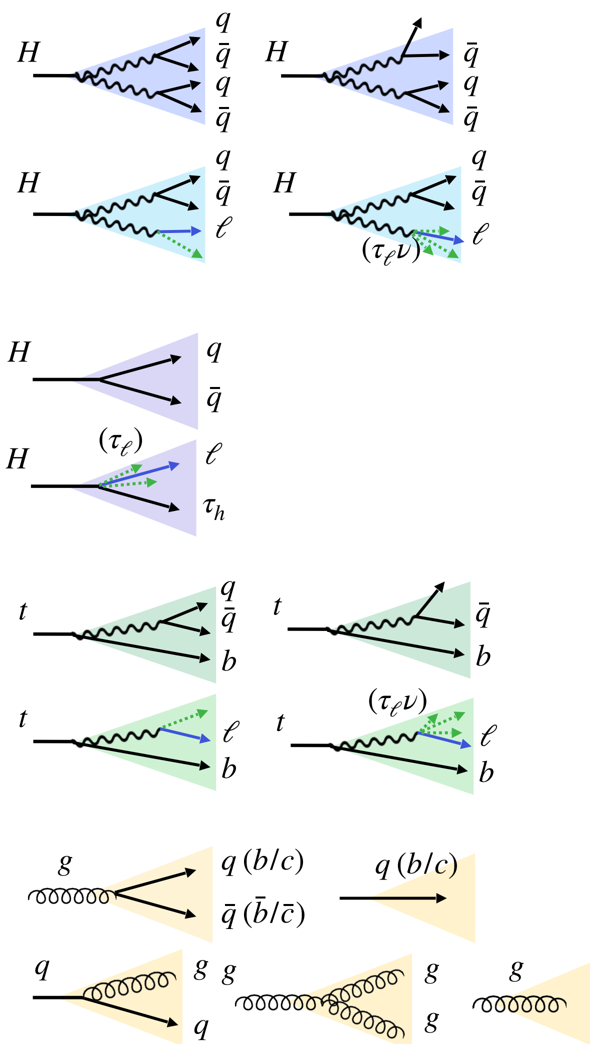


Model scaling up

- one upstream pre-training, broad downstream applicability

CMS’s path to develop Global Particle Transformer

Process	Final state/ prongness	heavy flavour	# of classes
H→VV (full-hadronic)	qqqq	0c/1c/2c	3
	qqq		3
H→WW (semi-leptonic)	$e\nu qq$	0c/1c	2
	$\mu\nu qq$		2
	$\tau_e\nu qq$		2
	$\tau_\mu\nu qq$		2
	$\tau_h\nu qq$		2
H→qq		bb	1
		cc	1
		ss	1
		qq (q=u/d)	1
H→ττ	$\tau_e\tau_h$		1
	$\tau_\mu\tau_h$		1
	$\tau_h\tau_h$		1
t→bW (hadronic)	bqq	1b + 0c/1c	2
	bq		2
t→bW (leptonic)	$be\nu$	1b	1
	$b\mu\nu$		1
	$b\tau_e\nu$		1
	$b\tau_\mu\nu$		1
	$b\tau_h\nu$		1
QCD		b	1
		bb	1
		c	1
		cc	1
		others (light)	1



CMS-PAS-HIG-23-012

The early version **(GloParT stage-1)** has been released with the HH→bbWW search

✤ **GloParT stage-1:** a fatjet tagger for **37-category classification**

- bbWW is the second work in the series of “boosted HH search”
- tagging **boosted H→WW→4q signature** for the first time
- set a tight constraint to κ_{2V}

CMS's path to develop Global Particle Transformer

GloParT-3 in CMS:

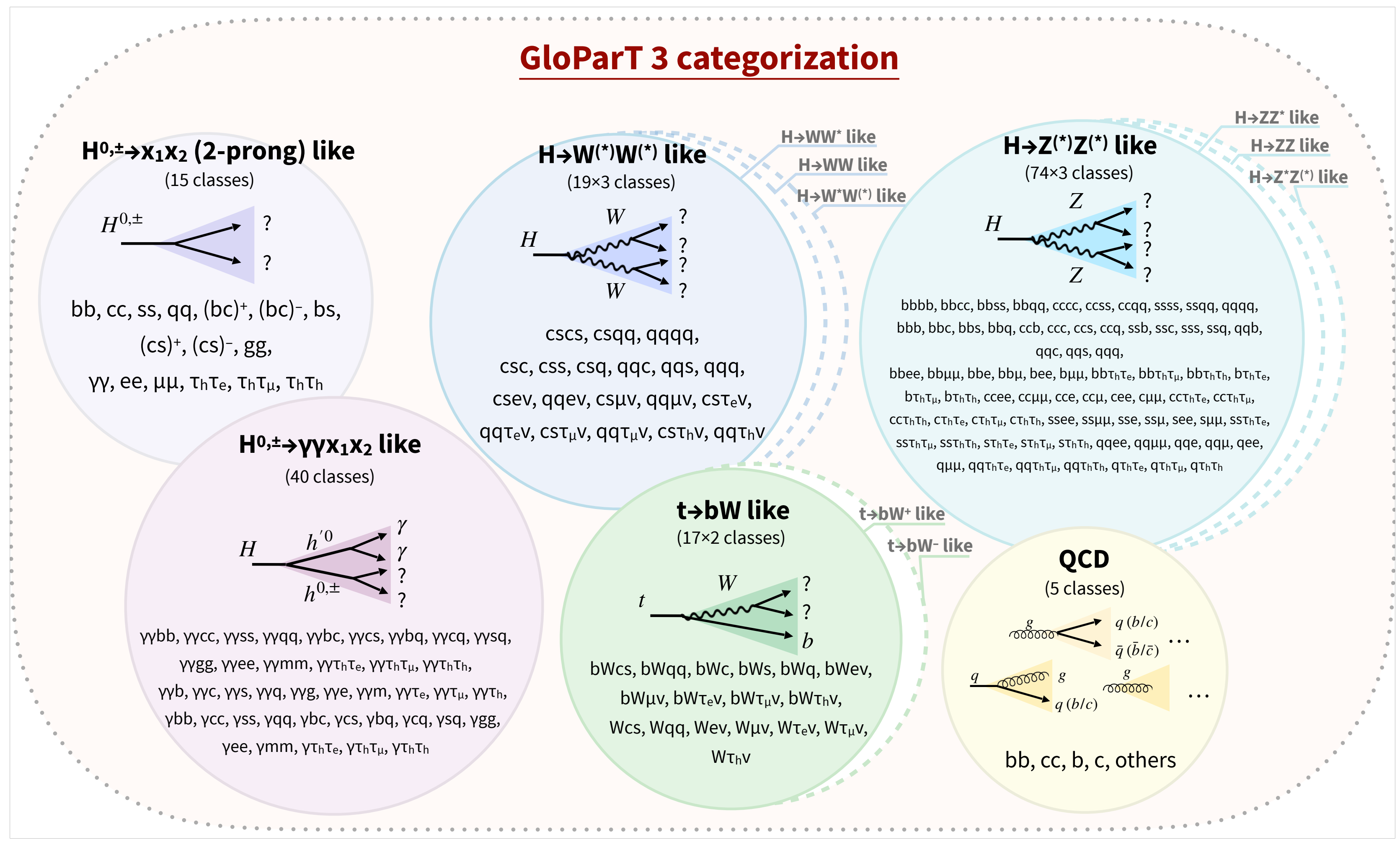
✅ A universal model that outperforms all existing taggers across existing tasks

✅ Manifest strong fine-tuning capability in various downstream tasks

A more powerful model is now in place

- **the 3rd version** of the Global Particle Transformer **(GloParT-3)** now integrated in CMS

- ***750 nodes in total***



Discussion: Implications to ATLAS/CMS experiments?

- "Sophon/GloParT methodology" releases a lot of new opportunities for future LHC experiments
  - it creates a **"global large-*R* jet tagger"** → bring benefits of the advanced NN to ~all hadronic final-state searches
  - **also viewed as a pre-trained jet model**: a base model tailored for a broad range of LHC analyses
- How to use the experimental version of the Sophon model?
  - used in conventional analyses: except for some well-calibrated nodes, the major challenge will be the calibration of peculiar signals (not easy to find proxies)
  - **analyses that only use data (simulation-free):** develop discriminants dedicated to different signals → cut tight on the data events → peak finding on some mass observable (single jet / di-jet / jet+lepton…)
    - **could be helpful in broadly searching for BSM resonance!**
  - anomaly detection: weakly-supervised approaches / further improvements? *(see backup for recent ATLAS/CMS results)*

Discussion: JetClass-II and Sophon

arXiv:2405.12972

- Developed the **JetClass-II** dataset and the **Sophon** model
- **JetClass-II** **[Hugging Face dataset]** covers more comprehensive phase spaces and can be a good playground to develop future foundation models
  - can be used to train models for various jet-related tasks, e.g. jet classification, regression, generation or reconstruction…
  - its extensive phase space coverage and high statistics enable model developers to **focus on specific regions of interest**, or **work with the entire dataset**
  - generation details can be found in this repository
- The **Sophon model** **[Hugging Face]** can be helpful in delivering future LHC pheno research
  - optimizing sensitivity for dedicated searches/anomaly detection/ novel paradigms
  - performing studies on the pheno dataset/model can inspire how we do real experiments at the LHC

Discussion: Close-up & Future Fantasies...

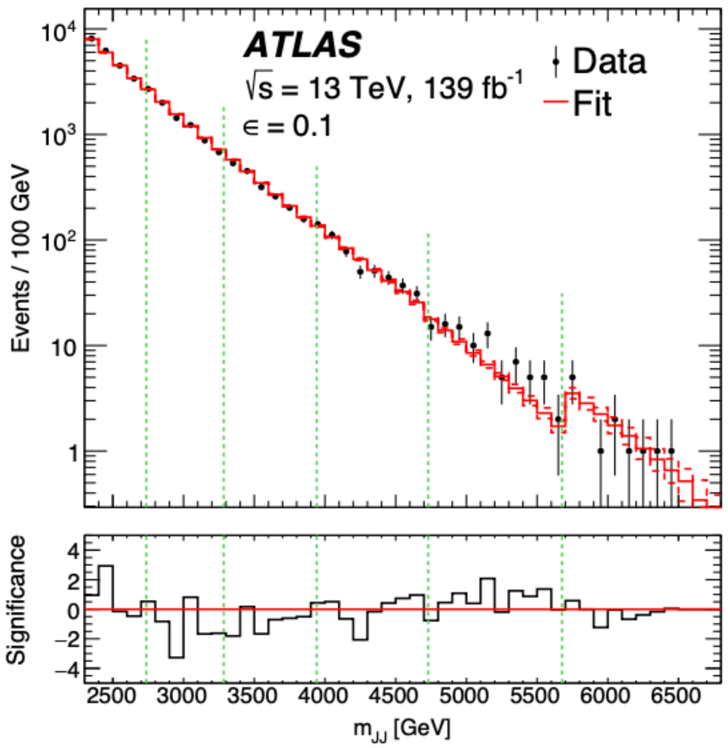
- Future foundation/base model at LHC:
  - which dataset will it be trained on (data/simulation)?
  - which training targets (generation/classification/embedding prediction)?
  - maybe, most importantly, which goal do we want to achieve?

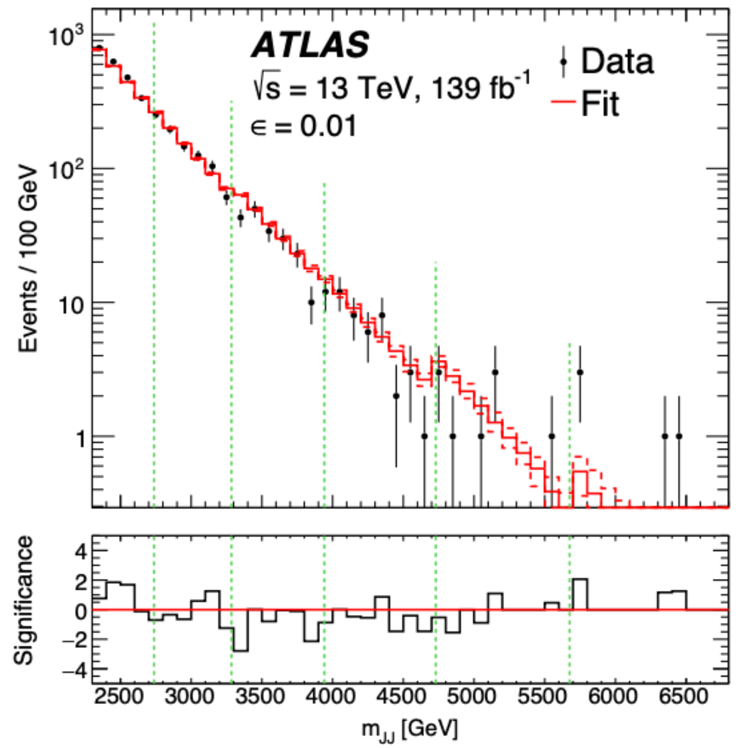
- Some specific (maybe preliminary) points to discuss
  - SSL or signature-oriented pre-training?
    - In HEP, we do not lack data labels; simulation will also reach expected accuracy in future
    - should we do SSL, or supervised pre-training exploring all GEN labels, or both?
  - foundation model requires a "foundation dataset"; which will be the foundation dataset in HEP?
    - should cover a vast phase-space (beyond SM simulation/real data): produced by philosophies like JetClass-II?
    - should be in the most general form, e.g. independent of experiments.. - GEN-level data be the most suitable playground?
  - The limit of scaling capabilities?
    - are we reaching the limit for classification tasks? Will fine-tuning from a large pre-trained model provide better results?

Backup

Recent ATLAS/CMS anomaly detection results

Weakly supervised approach

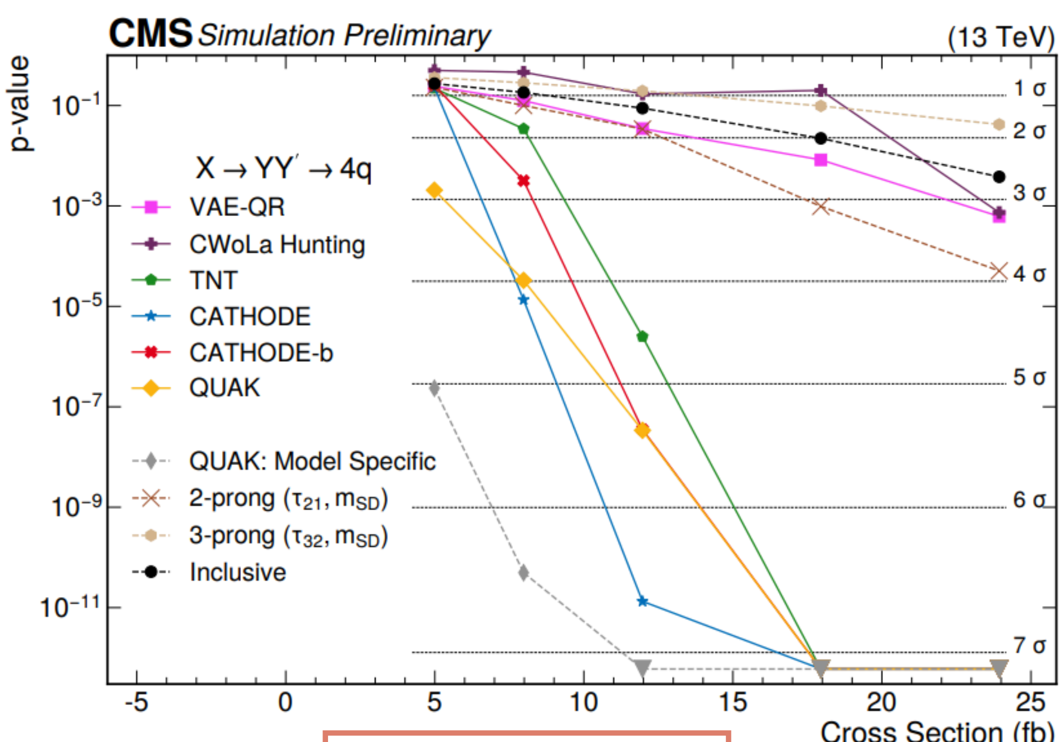




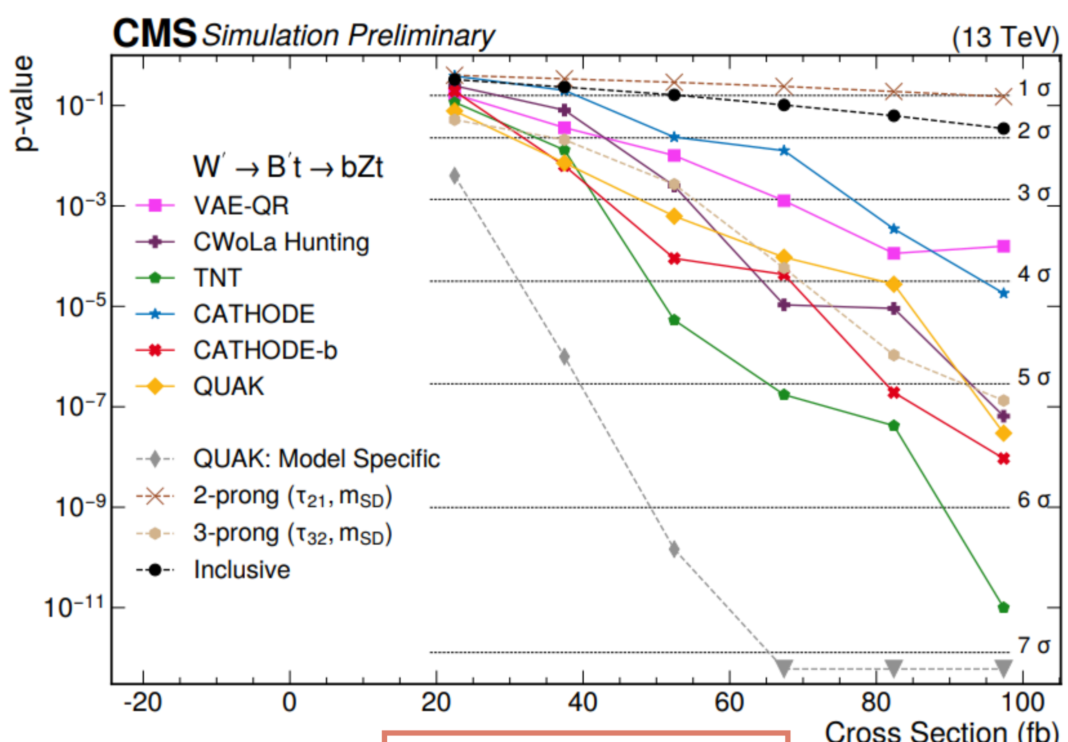
Application in ATLAS

PRL 125, 131801 (2020)

Select on the trained weakly supervised classifier in each signal regions, and search for the peak



2-prong + 2-prong



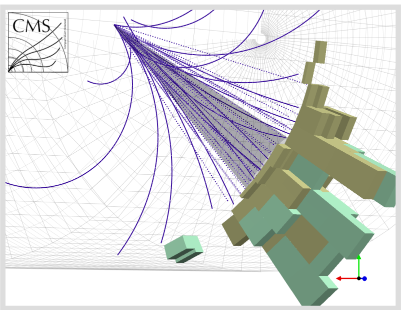
3-prong + 3-prong

CMS-PAS-EXO-22-026

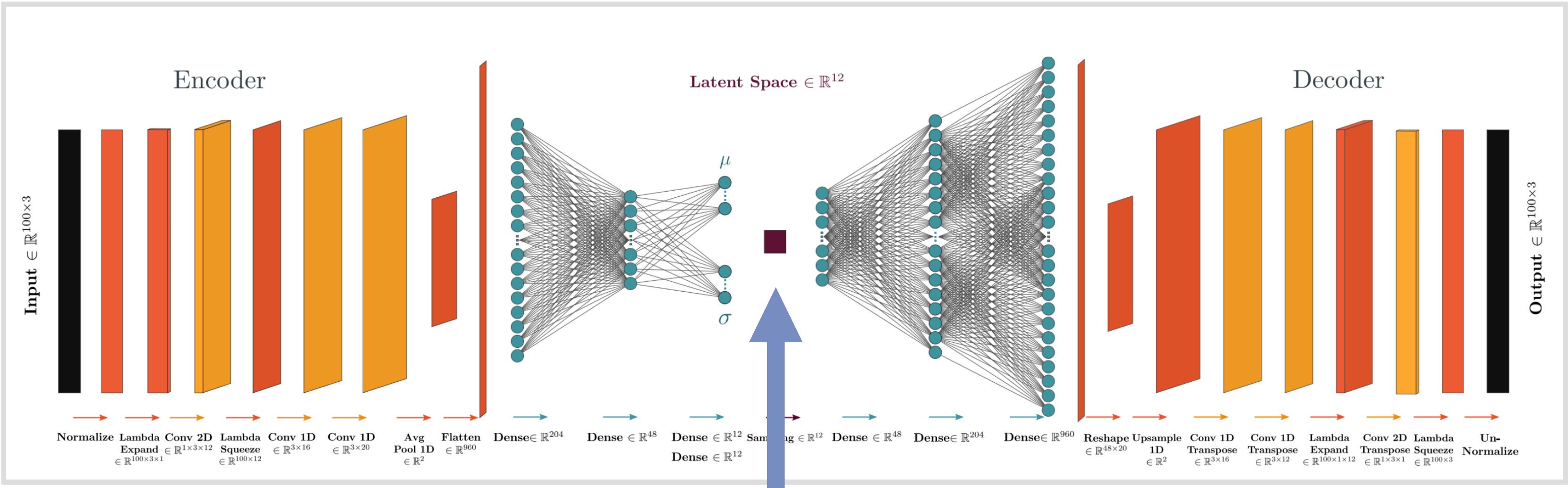
In a recent work, CMS systematically test all model-agnostic approaches to search for resonance

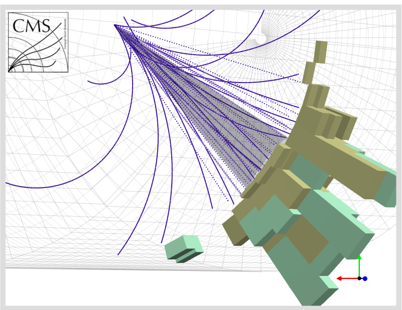
Recent ATLAS/CMS anomaly detection results

Autoencoder approach



input jet

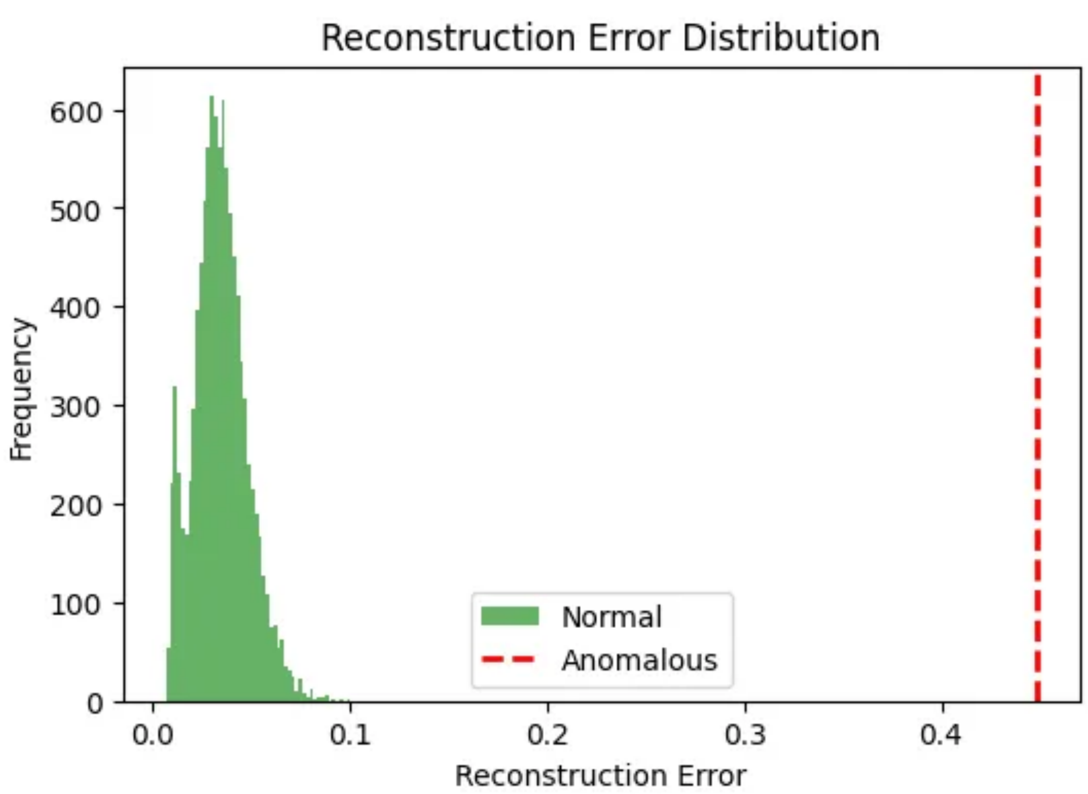




reconstructed jet

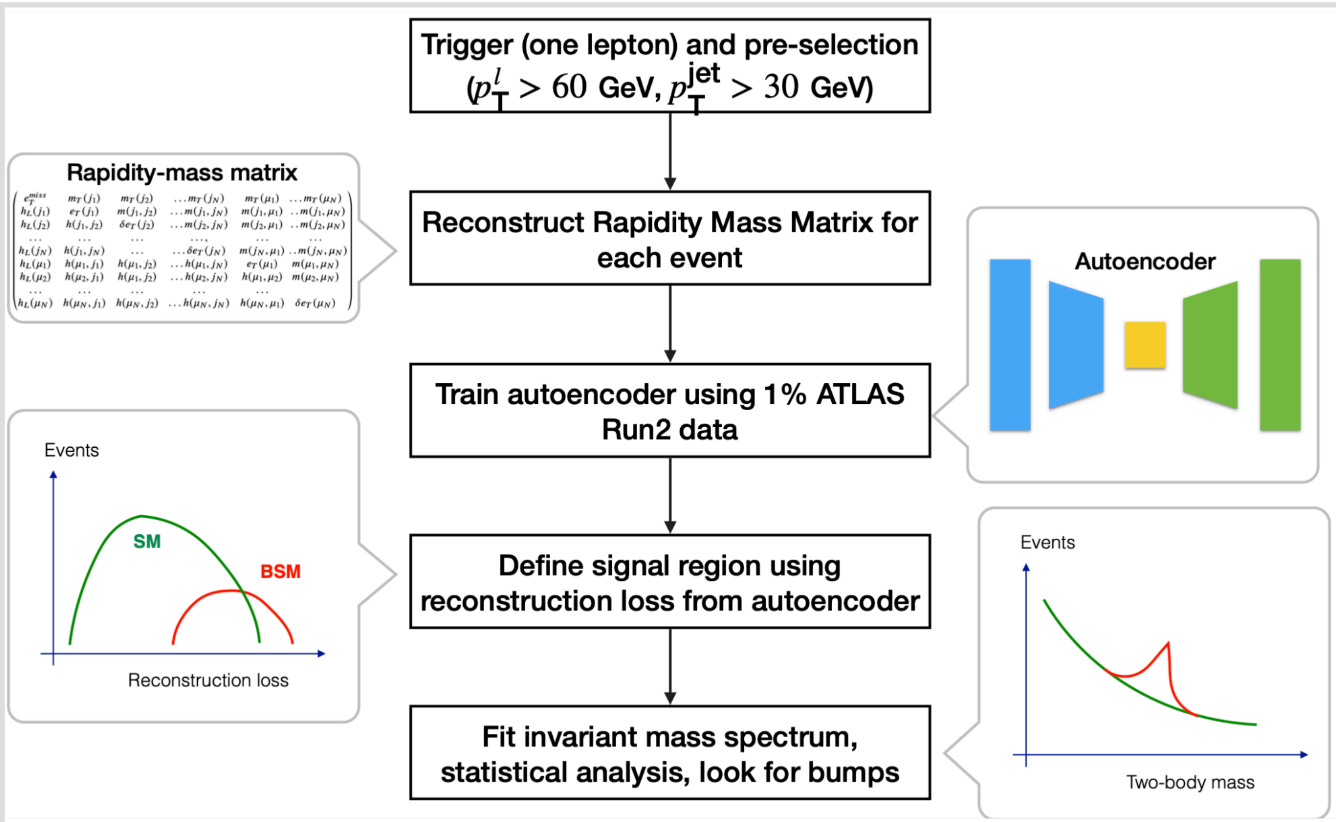
a compressed jet representation

- ➔ A view on (variational) autoencoder for anomaly detection
  - ❖ Training on SM background jet → **anomalous jet will produce outlier latent scores** → make selection on the score
- ➔ Use autoencoder for anomaly detection has industry basis

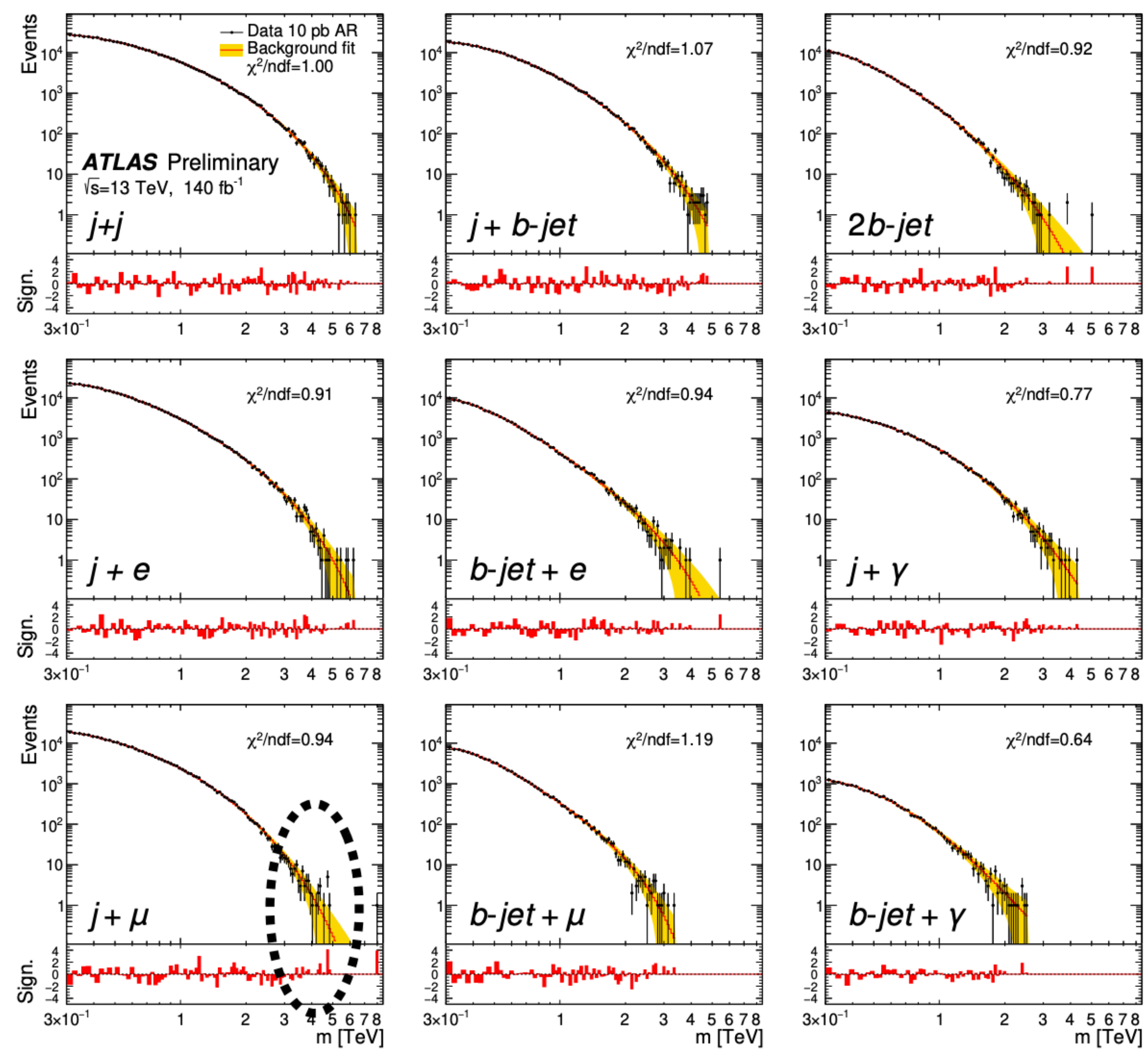


Recent ATLAS/CMS anomaly detection results

Autoencoder approach



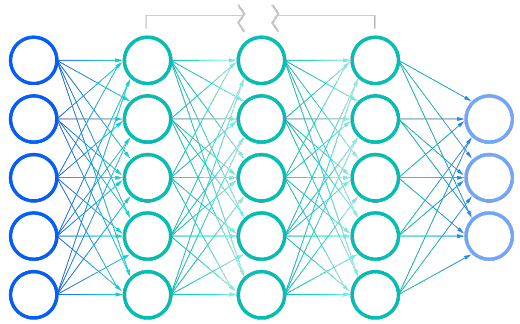
PRL 132 (2024) 081801



- ➔ ATLAS applies full-event-level anomaly detection
- ➔ Train "autoencoder" and select on the score
- ➔ Search in 9 invariant masses including di-jet, di-b-jet, with three anomaly regions

Evolution of jet NNs

feed-forward NN *(high-level inputs)* ····▶···· *1D/2D CNN, RNN* *(low-level inputs)* ····▶··· *graph NN, Transformers* *(low-level inputs)* ····▶··· *??*

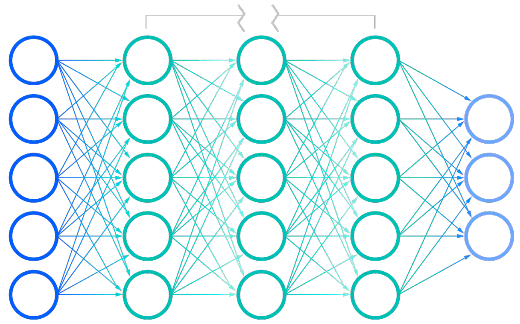


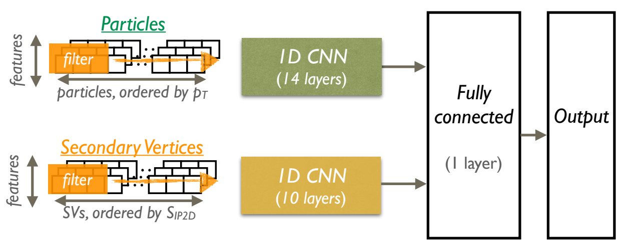
Shallow networks

✦ Using high-level features directly as input to a shallow network

Evolution of jet NNs

feed-forward NN *(high-level inputs)* ••••▶••• *1D/2D CNN, RNN* *(low-level inputs)* ••••▶••• *graph NN, Transformers* *(low-level inputs)* ••••▶••• *??*





Shallow networks

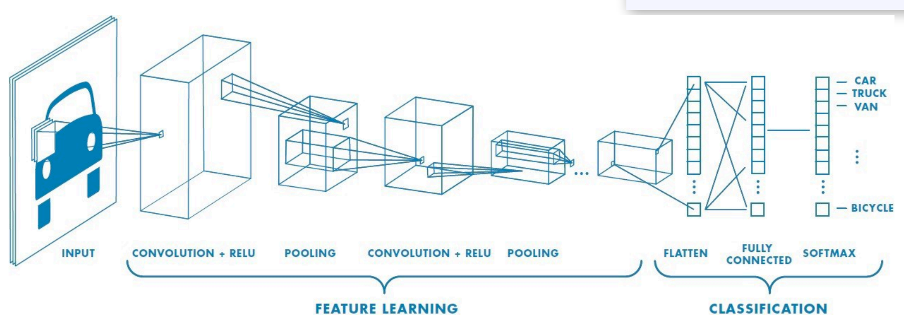
✦ Using high-level features directly as input to a shallow network

Deep NN with low-level inputs

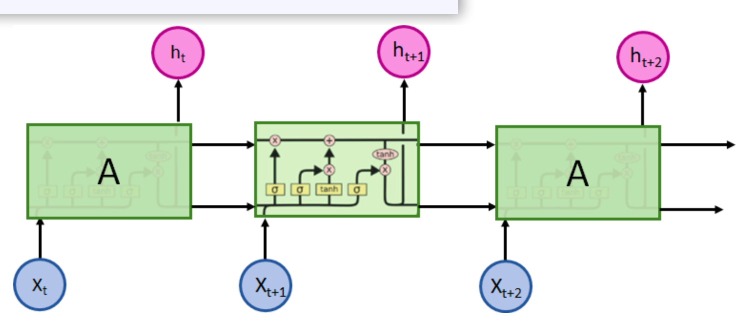
✦ Using particle-level features

✦ Input data structure determines the type of networks

- jet as a *image (fixed-grid data structure)*
- jet as a *sequence* → 1D CNN or RNN



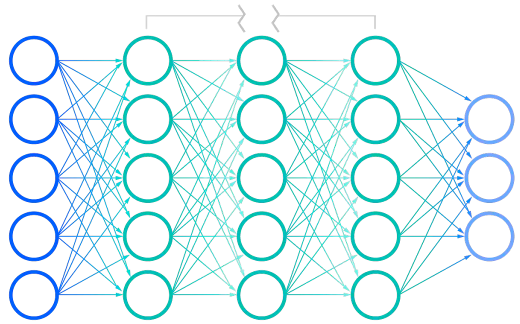
Typical CNN



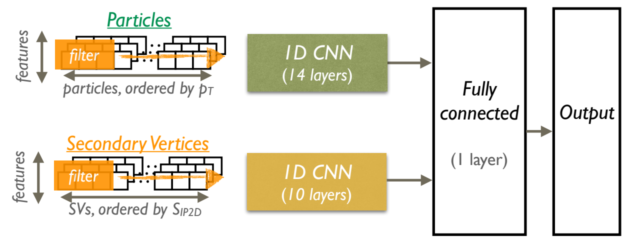
Typical RNN

Evolution of jet NNs

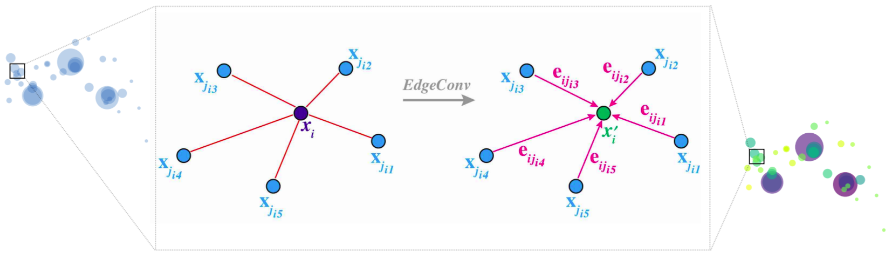
feed-forward NN *(high-level inputs)* ····▶···· *1D/2D CNN, RNN* *(low-level inputs)* ····▶··· *graph NN, Transformers* *(low-level inputs)* ····▶··· *??*



Shallow networks



Deep NN with low-level inputs



Graph structure

✦ Using high-level features directly as input to a shallow network

✦ Using particle-level features

✦ Input data structure determines the type of networks

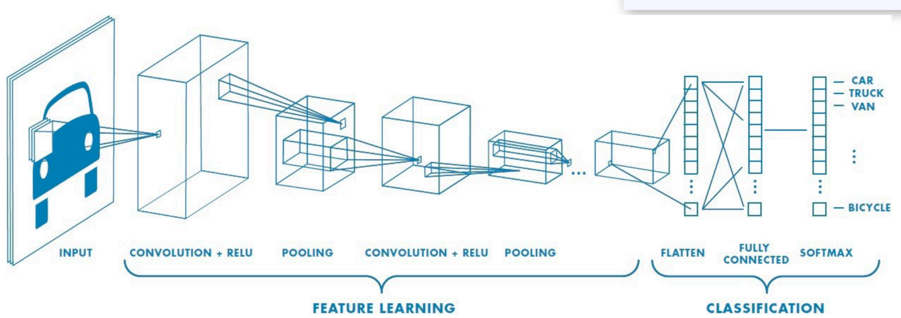
- jet as a *image (fixed-grid data structure)*
- jet as a *sequence* → 1D CNN or RNN

✦ Graph neural networks

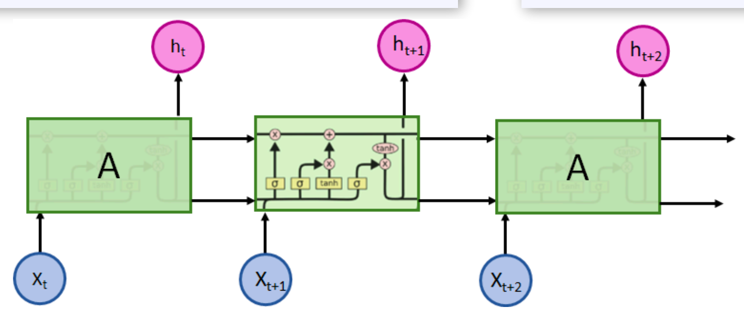
- treat a jet as a permutational-invariant set of particles (or, point cloud)
- build "edges" between particles

✦ Transformer networks

- modern architectural designs; like a full-connected graph



Typical CNN



Typical RNN

Typical graph

GNNs and Transformers

→ **Modern architectures done right**: (which types of DNNs better suit the particle-format data?)

- **inductive bias**: particle-format data has their intrinsic symmetries
  - **permutational-invariant symmetry**: GNN is better than CNN/RNN; native Transformer (w/o positional encoding)
  - Lorentz symmetry: adding "pairwise particle masses" to input features
- **let particles interact**:
  - "message passing" in GNNs and attention mechanism in Transformers
- **scale better with data and model size**
  - Transformers!