Based on arXiv:2405.12972

[Github] [Google Colab]

Sophon meets LHC:
Accelerating resonance discovery via signature-oriented pre-training

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

based on the work with our colleagues in the CMS Collaboration:

Antonios Agapitos[1], Jovin Drews[2], Javier Duarte[3], Dawei Fu[1], Leyun Gao[1], Raghav Kansal[3], Gregor Kasieczka[2], Louis Moureaux[2], Huilin Qu[4], Cristina Mantilla Suarez[5], Qiang Li[1]

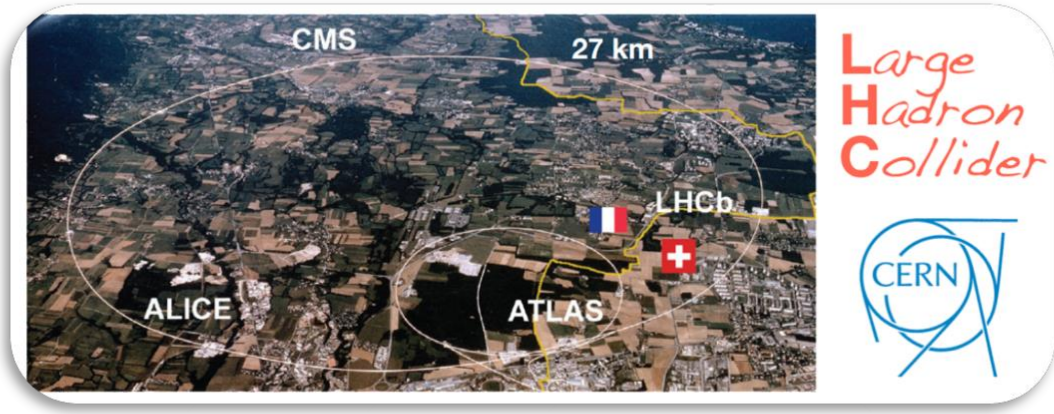
1) Peking U. 2) Hamburg U. 3) UC San Diego 4) CERN 5) FNAL

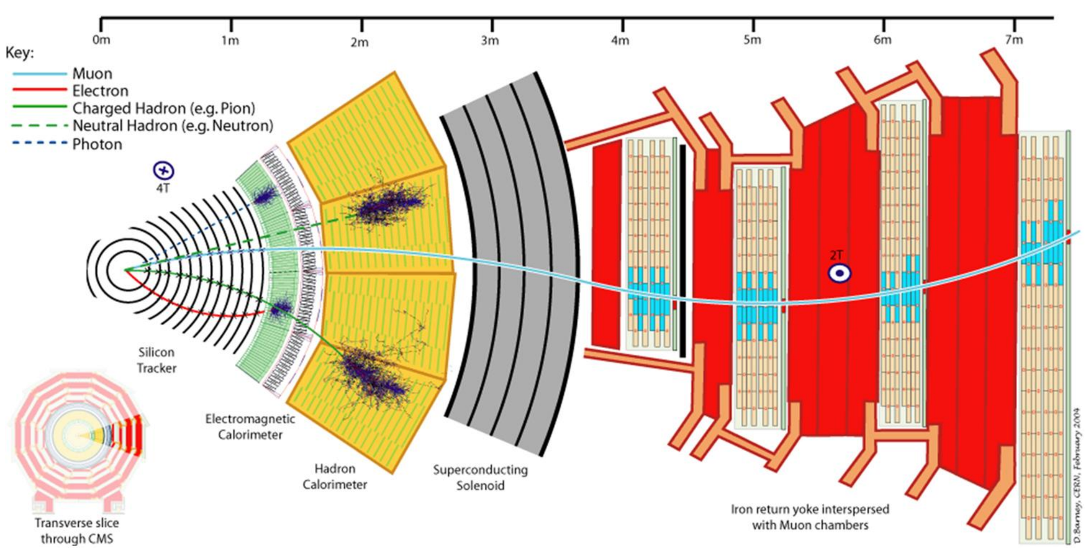
also thanks **Yuzhe Zhao[1]** *for his contribution*

Quantum Computing and Machine Learning Workshop 2024, Changchun

6 August, 2024

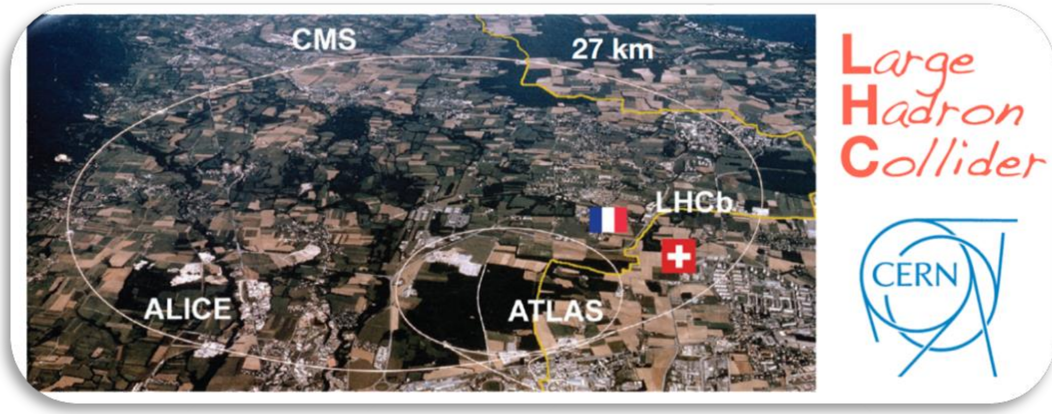
LHC physics × the era of deep learning

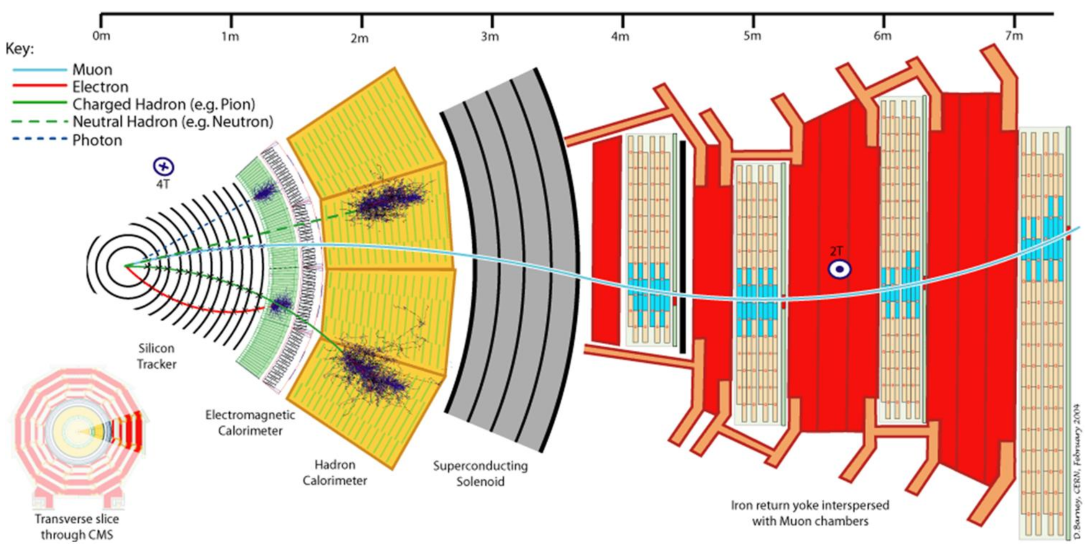




- ➔ The LHC: current world's largest particle collider
- ➔ Physics programs on LHC: reveal the fundamental theory of matters
- ➔ ATLAS and CMS: general-purpose detectors for precise SM measurement and **searching of BSM**

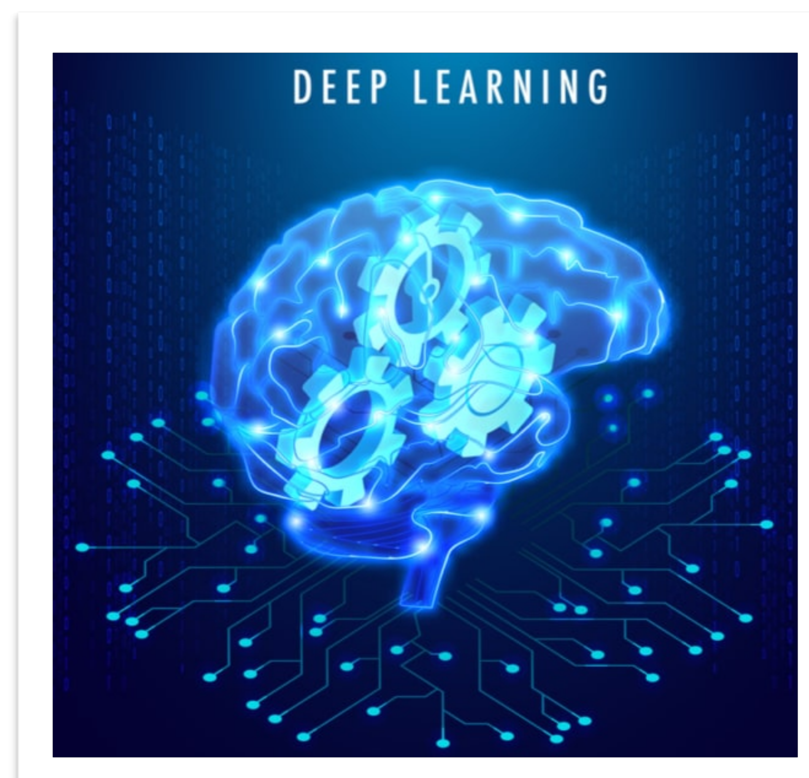
LHC physics × the era of deep learning





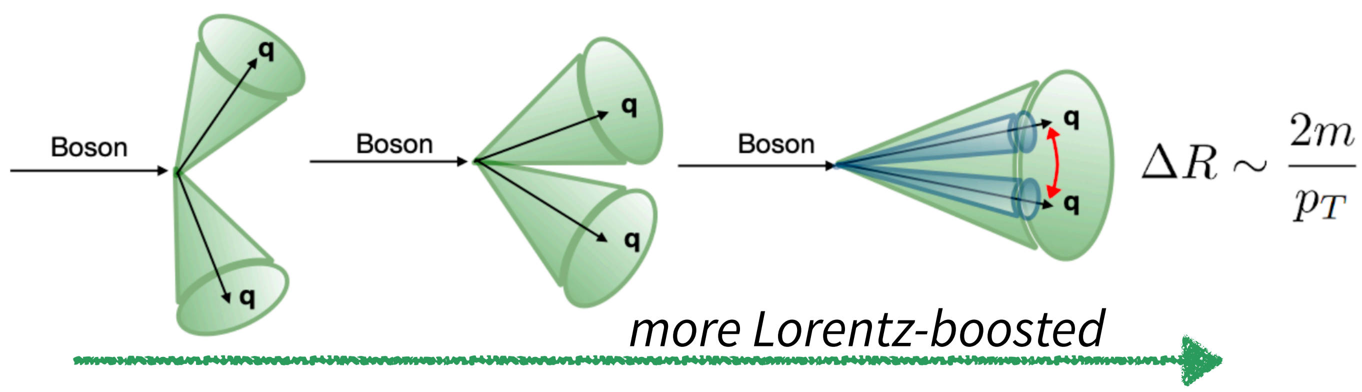
- ➔ The LHC: current world's largest particle collider
- ➔ Physics programs on LHC: reveal the fundamental theory of matters
- ➔ ATLAS and CMS: general-purpose detectors for precise SM measurement and **searching of BSM**

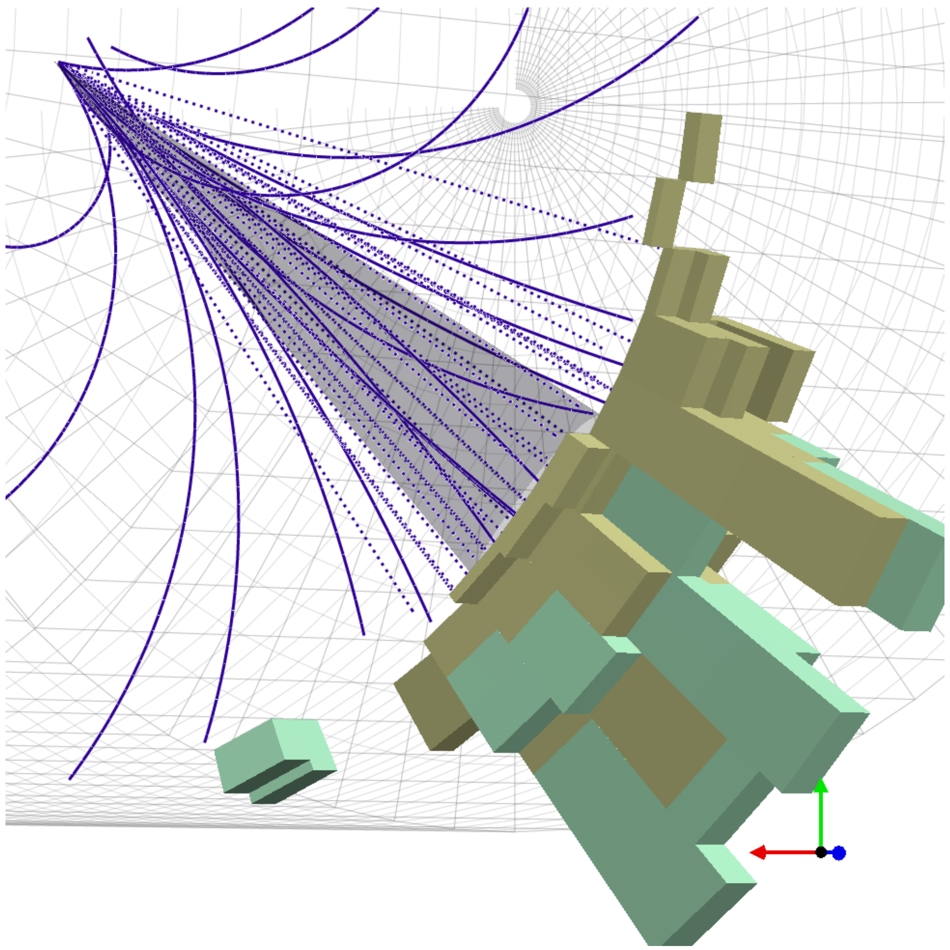
- ✤ *Deep learning (AI) is brining a technological leap in analyzing LHC data*
- ✤ *It is intriguing to think where the future possibility lies*



Boosted topology – a booster to sensitivity for LHC physics

➔ Large-*R* jets: an important handle to analyze boosted topologies at the LHC





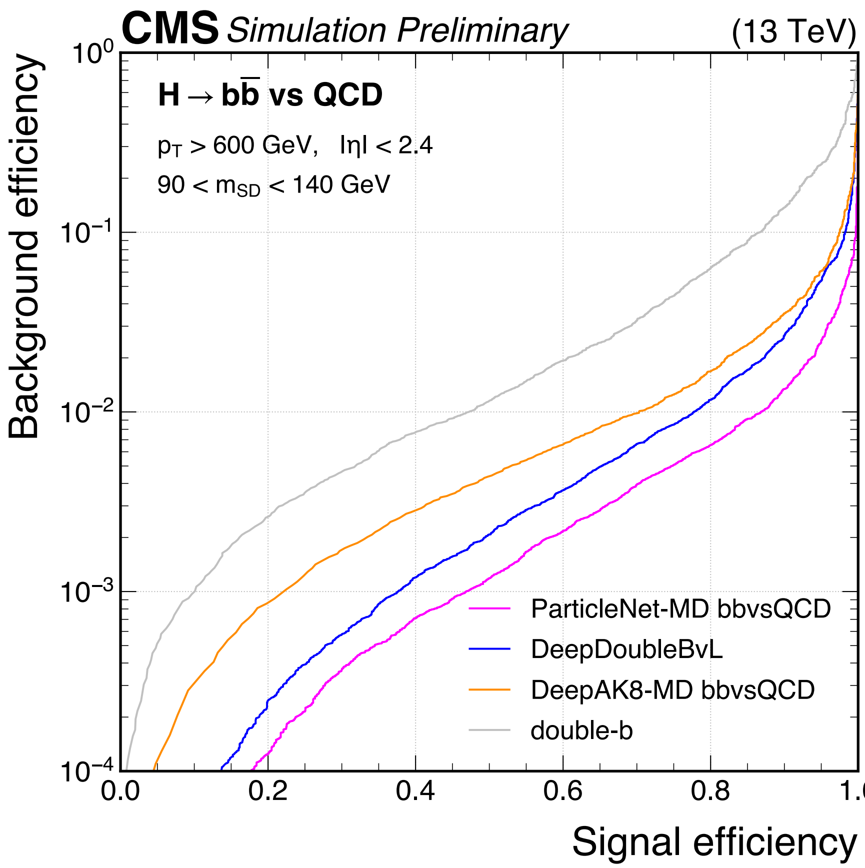
- Applications to Higgs/di-Higgs/BSM searches in boosted **H(X)→b$\bar{b}$/c$\bar{c}$** final states have been a success

➔ *Suitable for deploying cutting-edge deep learning techniques*

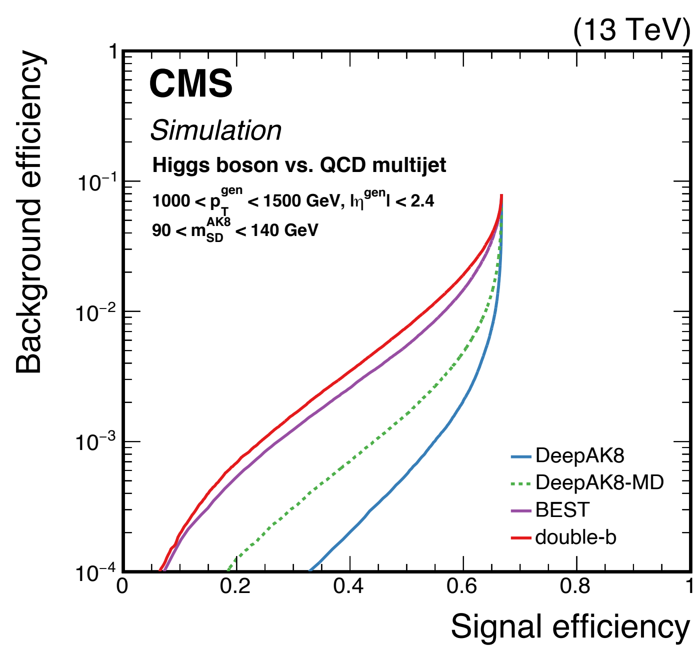
- most complex object to handle at the LHC (up to ~100 constituent particles)
- advanced DNNs greatly boost analysis sensitivity

Inspiring progress on H→b$\bar{b}$/c$\bar{c}$ tagging

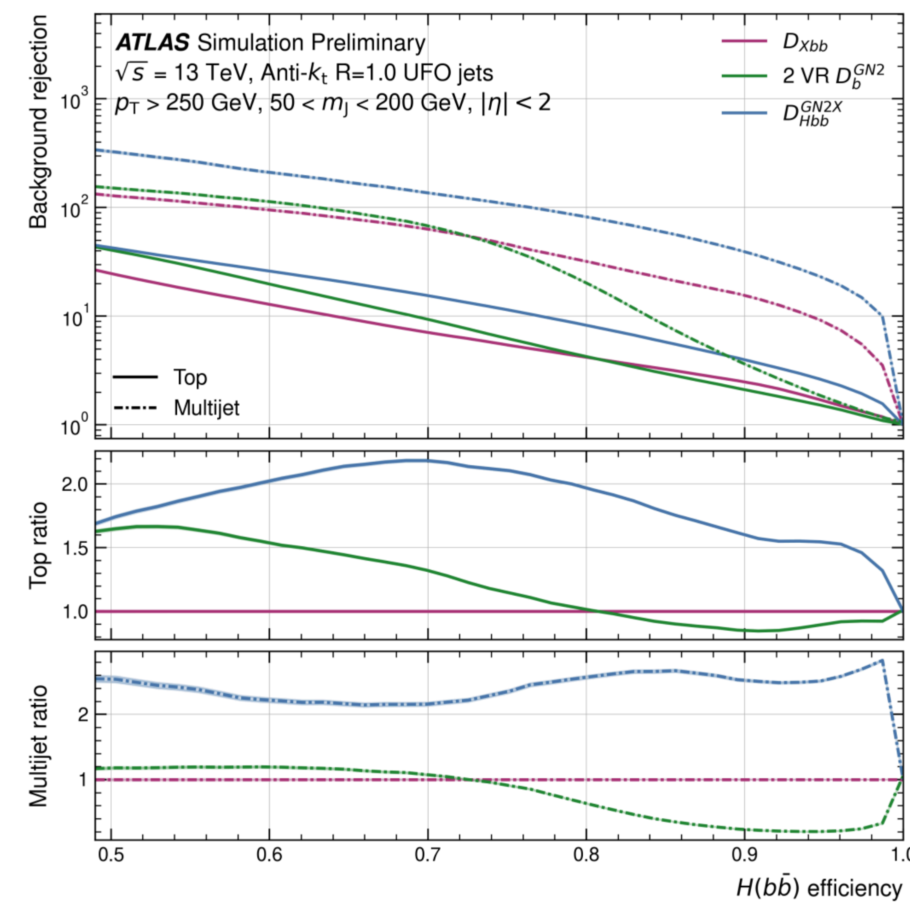
CMS-PAS-BTV-22-001



JINST 15 (2020) P06005



ATL-PHYS-PUB-2023-021



An upgrade of network

DeepAK8 → ParticleNet:

x5 📈 QCD background rejection

Note: back to the results 5 years ago

DeepAK8 tagger already has **~x5** 📈 improved background rejection than early methods

Recent GN2X tagger:

~x3 📈 QCD and x2 top background rejection

Inspiring progress on H→$b\bar{b}$/$c\bar{c}$ tagging

Implications

Advancements in NN design are the true driving force behind the gains in sensitivity!

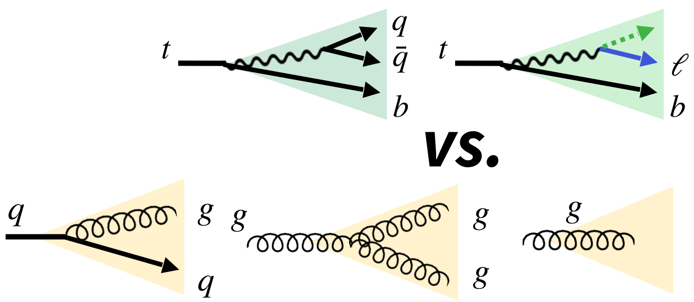
- **However, this tool is available only in limited phase space**
  (e.g. H→bb, W→qq, t→bqq…)

- **Can we extend its usage to all possible boosted phase spaces?**
- **If we can boost the sensitivity by *x10* in many phase spaces, will it accelerate the "next potential discovery" of a new resonance?**

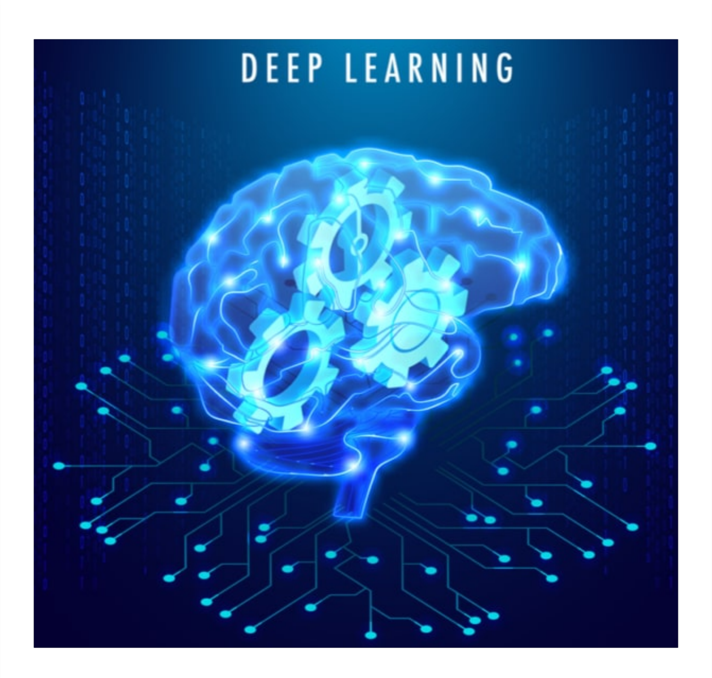
Propose “Large model for large-scale classification”

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 (next slide)
- ➔ 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



- ➔ We own a comprehensive jet dataset, and we hope to pre-train **a foundational 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 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

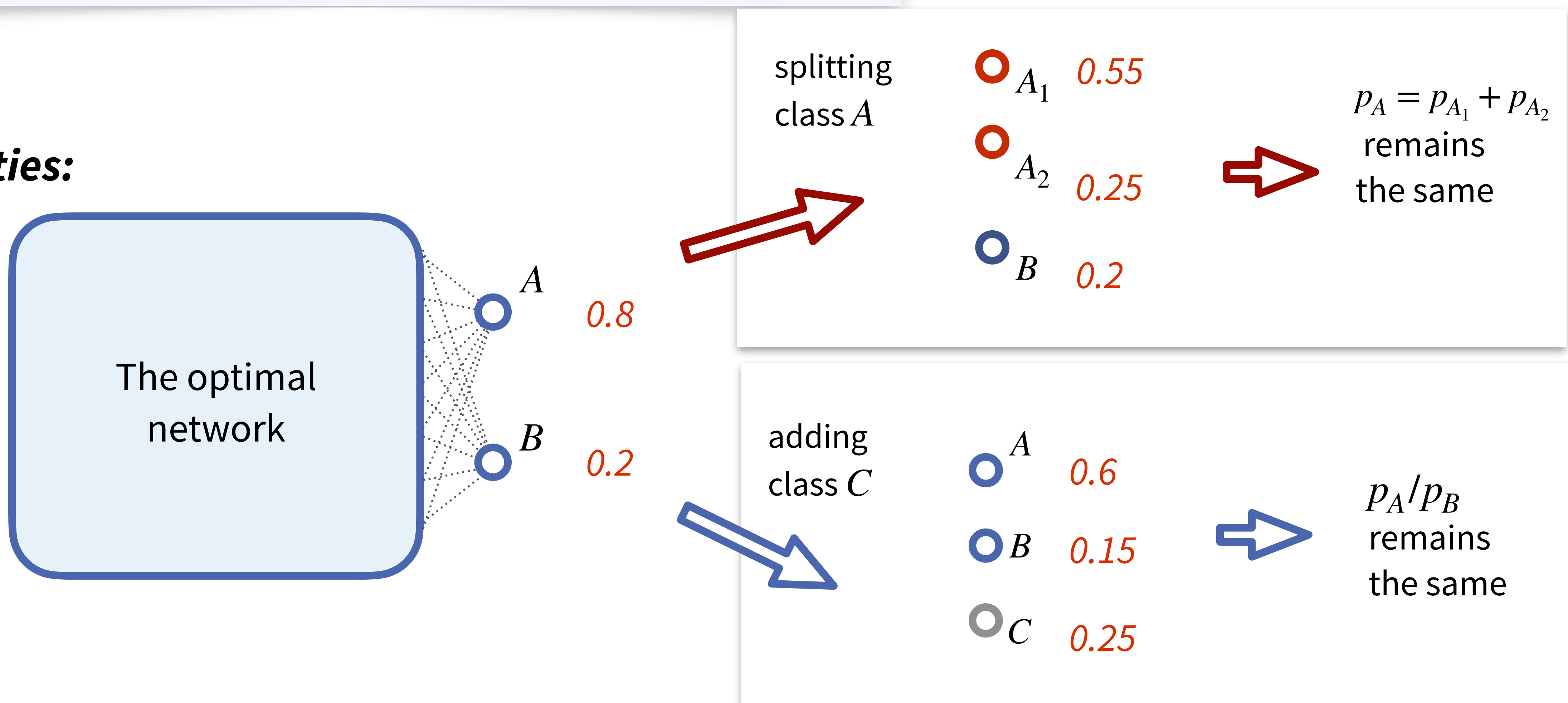
→ Statistical theory shows that:

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:

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

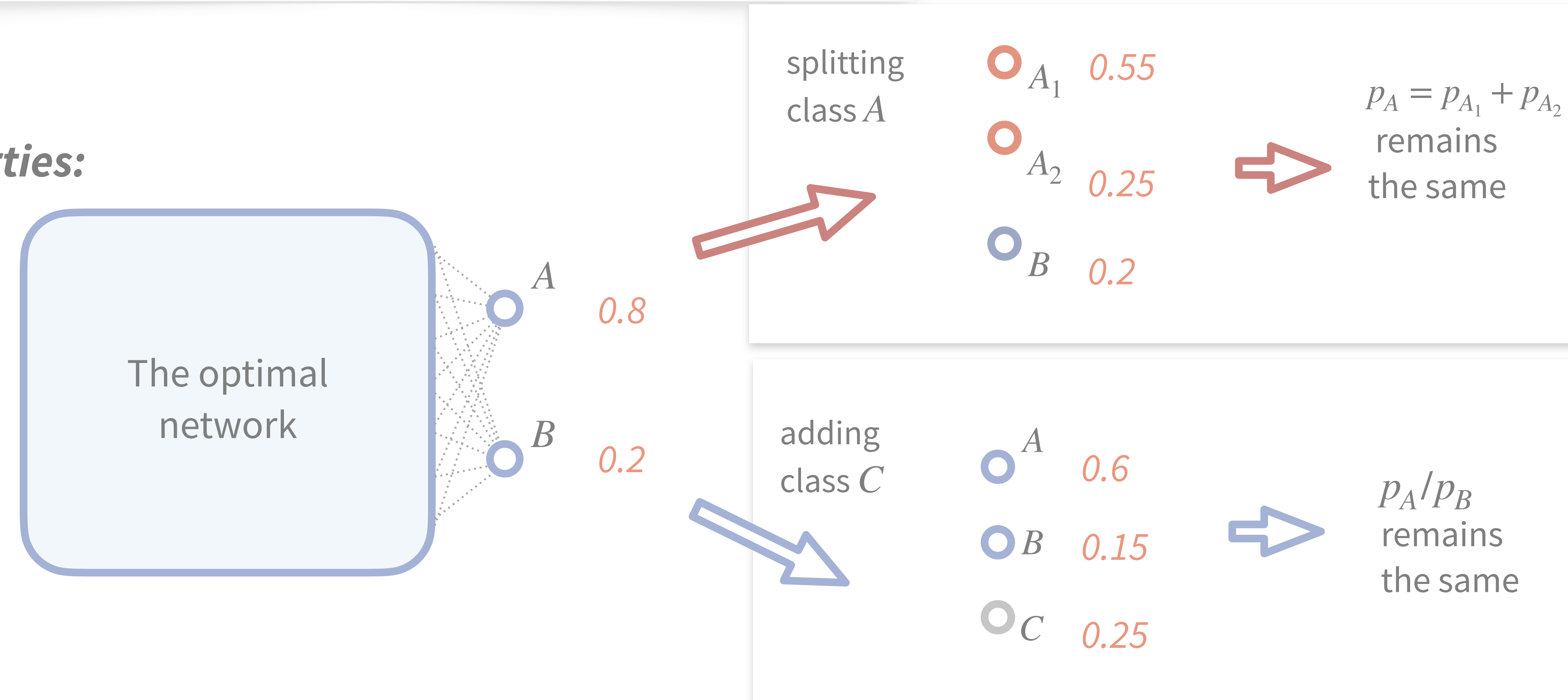
The key question in this context

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

Our result will show: Yes.

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

Two properties:



Introducing Sophon

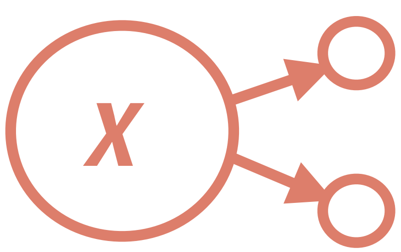
arXiv:2405.12972
https://github.com/jet-universe/sophon

➔ We explore this possibility in the CMS experiment first, and also in a recent pheno work:

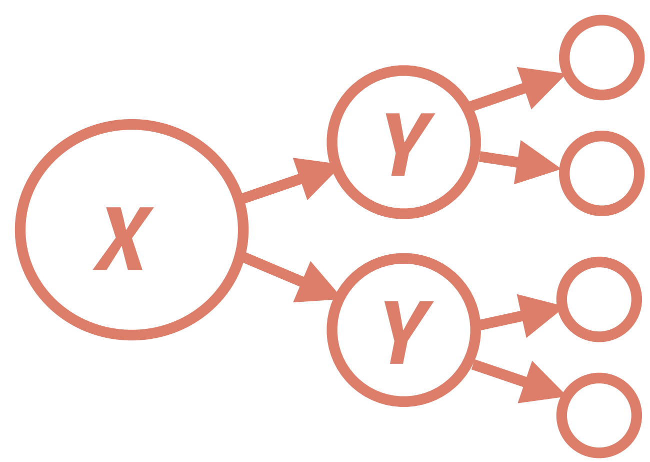
- **Signature-Oriented Pre-training for Heavy-resonant ObservatioN**
- the model is based on Particle Transformer (**ParT**) architecture
- a pre-trained model on a comprehensive dataset: **JetClass-II**
  - ***finely categorized labels:***

[H.Qu, CL, S.Qian. ICML 2022]
The current "state-of-the-art" on large-scale dataset ***(o(100M), close to experimental setup)***

Resonant jet:
X→ 2 prong

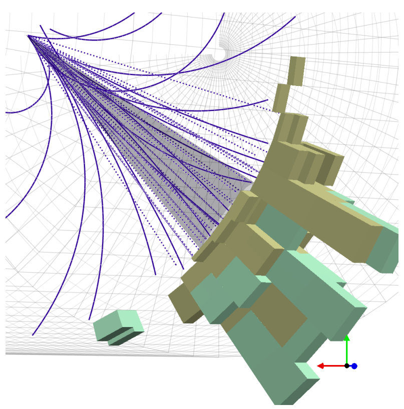


Resonant jet:
X→ 3/4 prong



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

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

➔ We explore this possibility in the CMS experiment first, and also in a recent pheno work:

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

[H.Qu, CL, S.Qian. ICML 2022]

The current "state-of-the-art" on large-scale dataset ***(o(100M), close to experimental setup)***

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, $\mu\mu$, $\tau_h\tau_e$, $\tau_h\tau_\mu$, $\tau_h\tau_h$
Resonant jets: $X \to 3$ or 4 prong	15–160	$bbbb$, $bbcc$, $bbss$, $bbqq$, $bbgg$, $bbee$, $bb\mu\mu$, $bb\tau_h\tau_e$, $bb\tau_h\tau_\mu$, $bb\tau_h\tau_h$, bbb, bbc, bbs, bbq, bbg, bbe, $bb\mu$, $cccc$, $ccss$, $ccqq$, $ccgg$, $ccee$, $cc\mu\mu$, $cc\tau_h\tau_e$, $cc\tau_h\tau_\mu$, $cc\tau_h\tau_h$, ccb, ccc, ccs, ccq, ccg, cce, $cc\mu$, $ssss$, $ssqq$, $ssgg$, $ssee$, $ss\mu\mu$, $ss\tau_h\tau_e$, $ss\tau_h\tau_\mu$, $ss\tau_h\tau_h$, ssb, ssc, sss, ssq, ssg, sse, $ss\mu$, $qqqq$, $qqgg$, $qqee$, $qq\mu\mu$, $qq\tau_h\tau_e$, $qq\tau_h\tau_\mu$, $qq\tau_h\tau_h$, qqb, qqc, qqs, qqq, qqg, qqe, $qq\mu$, $gggg$, $ggee$, $gg\mu\mu$, $gg\tau_h\tau_e$, $gg\tau_h\tau_\mu$, $gg\tau_h\tau_h$, ggb, ggc, ggs, ggq, ggg, gge, $gg\mu$, bee, cee, see, qee, gee, $b\mu\mu$, $c\mu\mu$, $s\mu\mu$, $q\mu\mu$, $g\mu\mu$, $b\tau_h\tau_e$, $c\tau_h\tau_e$, $s\tau_h\tau_e$, $q\tau_h\tau_e$, $g\tau_h\tau_e$, $b\tau_h\tau_\mu$, $c\tau_h\tau_\mu$, $s\tau_h\tau_\mu$, $q\tau_h\tau_\mu$, $g\tau_h\tau_\mu$, $b\tau_h\tau_h$, $c\tau_h\tau_h$, $s\tau_h\tau_h$, $q\tau_h\tau_h$, $g\tau_h\tau_h$, $qqqb$, $qqqc$, $qqqs$, $bbcq$, $ccbs$, $ccbq$, $ccsq$, $sscq$, $qqbc$, $qqbs$, $qqcs$, $bcsq$, bcs, bcq, bsq, csq, $bce\nu$, $cse\nu$, $bqe\nu$, $cqe\nu$, $sqe\nu$, $qqe\nu$, $bc\mu\nu$, $cs\mu\nu$, $bq\mu\nu$, $cq\mu\nu$, $sq\mu\nu$, $qq\mu\nu$, $bc\tau_e\nu$, $cs\tau_e\nu$, $bq\tau_e\nu$, $cq\tau_e\nu$, $sq\tau_e\nu$, $qq\tau_e\nu$, $bc\tau_\mu\nu$, $cs\tau_\mu\nu$, $bq\tau_\mu\nu$, $cq\tau_\mu\nu$, $sq\tau_\mu\nu$, $qq\tau_\mu\nu$, $bc\tau_h\nu$, $cs\tau_h\nu$, $bq\tau_h\nu$, $cq\tau_h\nu$, $sq\tau_h\nu$, $qq\tau_h\nu$
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

resulting to 4-prong or 3-prong

All final states!

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

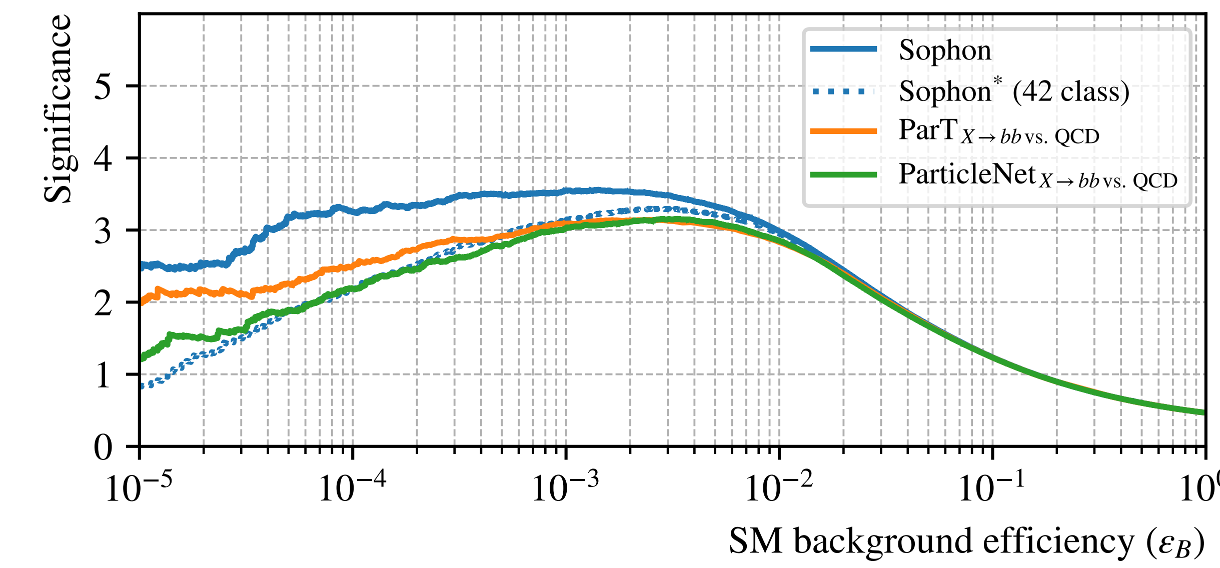
Sophon: performance benchmark

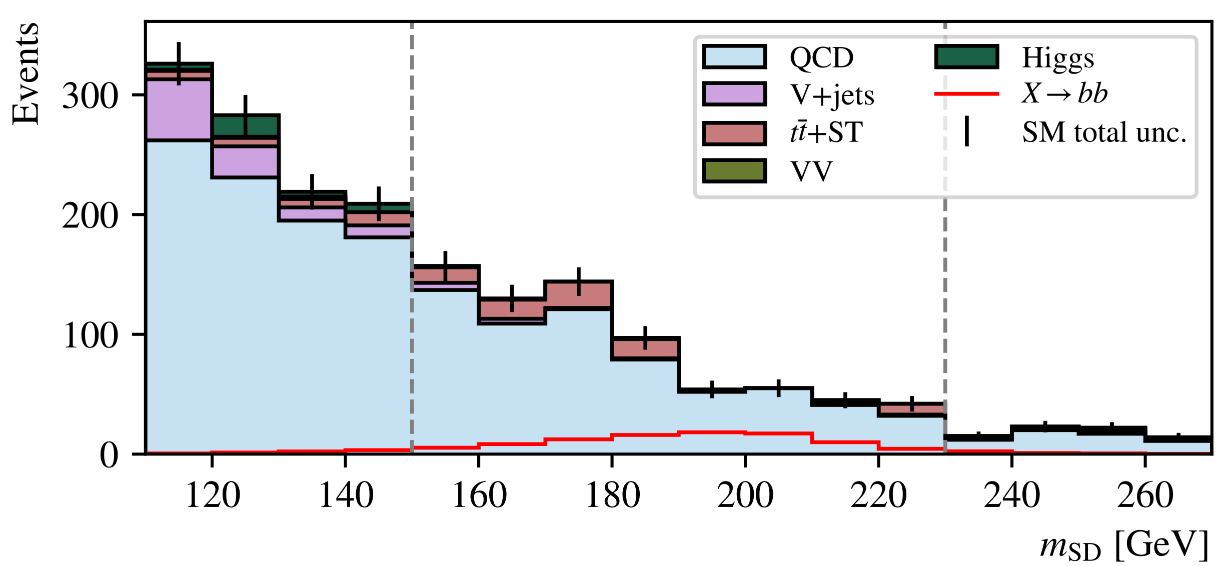
arXiv:2405.12972

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

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

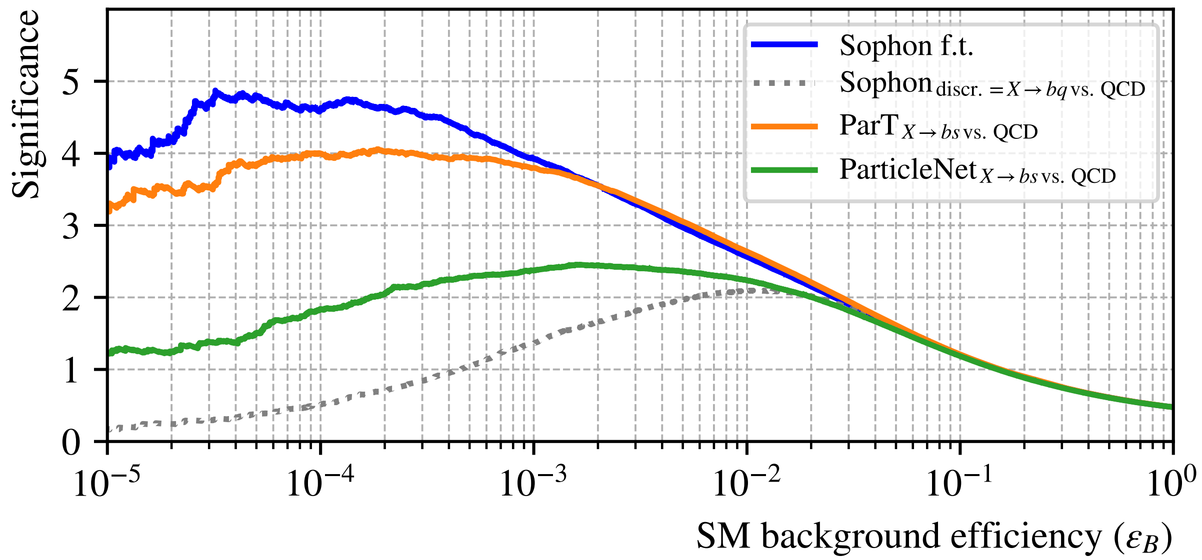
- Performance gain does come from large-scale classification (compared to **Sophon* (42 classes)**)
- **ParT** and **ParticleNet** for binary classification: they represent the best performance we can reach in experiment now

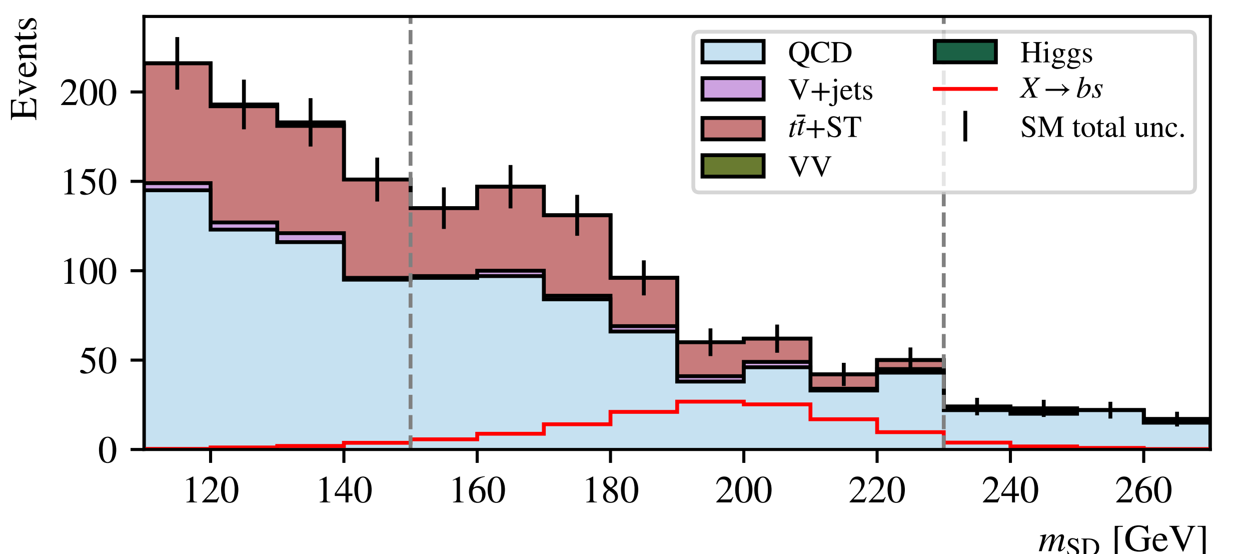
Sophon: performance benchmark

arXiv:2405.12972

Search significance:

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





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

Transfer learning ability

(adapt it to a brand new task)

- **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?

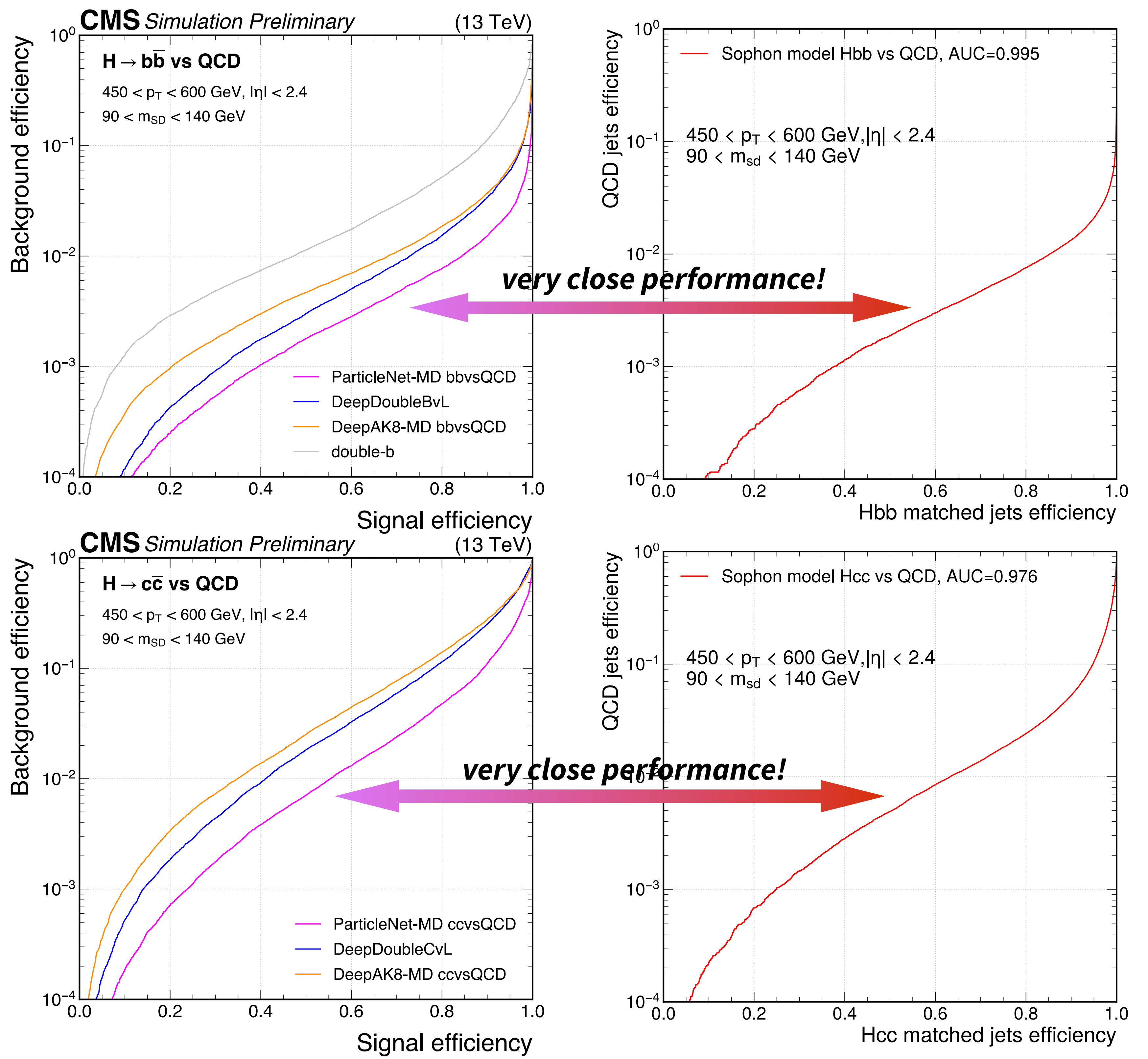
Thanks Yuzhe for his input

Take recent **CMS performance plots** as an example

Benchmark performance on SM H→$b\bar{b}$/$c\bar{c}$ jets vs QCD background jets

- Demonstrate that **applying "Sophon" on Delphes dataset for pheno study** is pretty realistic

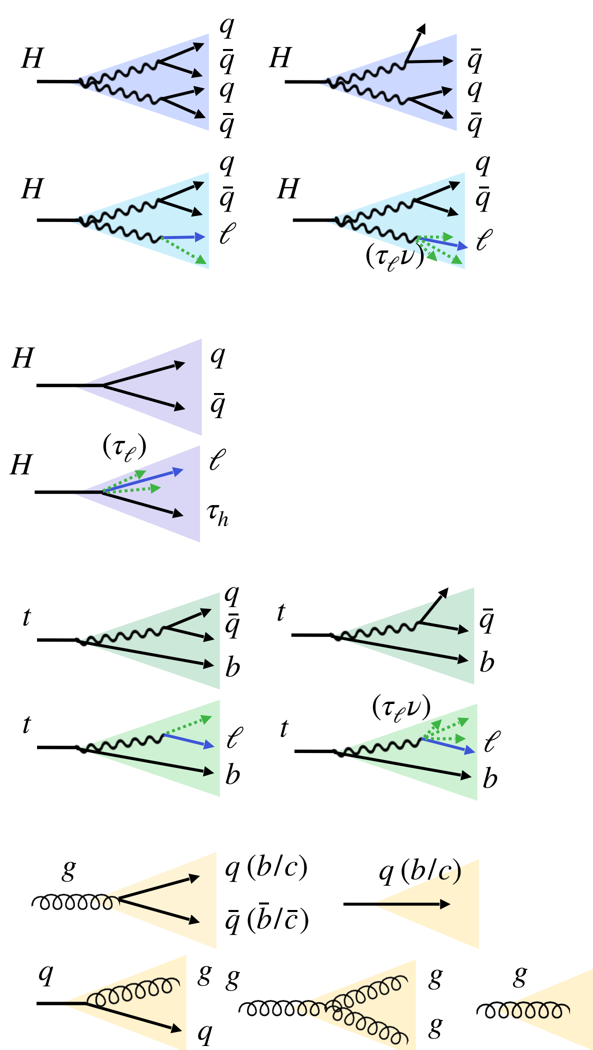
CMS results CMS-PAS-BTV-22-001 ***Sophon results*** *(on Delphes dataset)*



Highlight: CMS's Global ParT tagger

CMS-PAS-HIG-23-012

Process	Final state/ prongness	heavy flavour	# of classes
H→VV (full-hadronic)	qqqq	0c/1c/2c	3
	qqq		3
H→WW (semi-leptonic)	eνqq	0c/1c	2
	μνqq		2
	τ_eνqq		2
	τ_μνqq		2
	τ_hνqq		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)	beν	1b	1
	bμν		1
	bτ_eν		1
	bτ_μν		1
	bτ_hν		1
QCD		b	1
		bb	1
		c	1
		cc	1
		others (light)	1



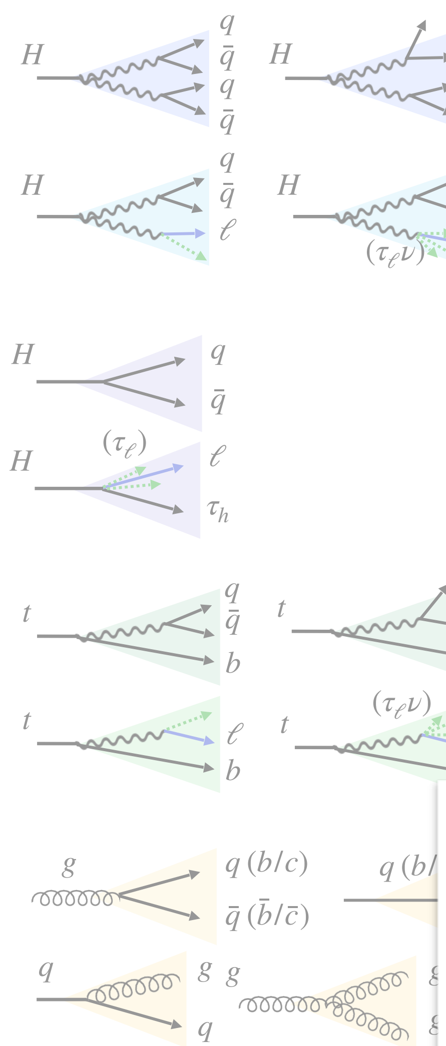
A global large-*R* mass-decorrelated tagger for **37-category classification**

- First time identifying the H→WW→4q signature with a jet tagger
- set a strong limit to κ_{2V} in the search of HH→bbVV signal

Highlight: CMS's Global ParT tagger

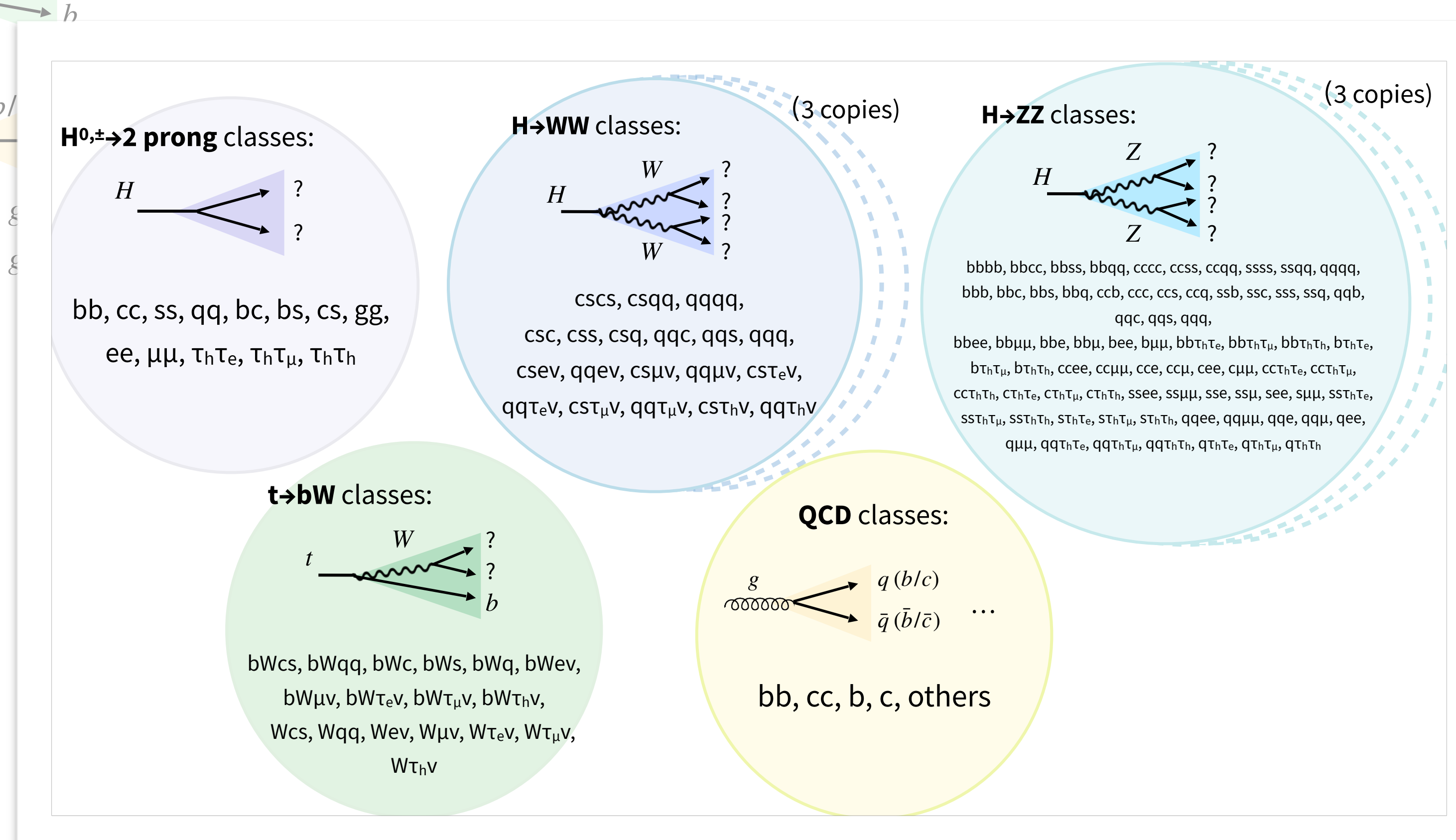
CMS-PAS-HIG-23-012

Process	Final state/ prongness	heavy flavour	# of classes
H→VV (full-hadronic)	qqqq	0c/1c/2c	3
	qqq		3
H→WW (semi-leptonic)	eνqq	0c/1c	2
	μνqq		2
	τ_eνqq		2
	τ_μνqq		2
	τ_hνqq		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)	beν	1b	1
	bμν		1
	bτ_eν		1
	bτ_μν		1
	bτ_hν		1
QCD		b	1
		bb	1
		c	1
		cc	1
		others (light)	1



A global large-R mass-decorrelated tagger for **37-category classification**

- First time identifying the H→WW→4q signature with a jet tagger
- set a strong limit to κ_{2V} in the search of HH→bbVV signal

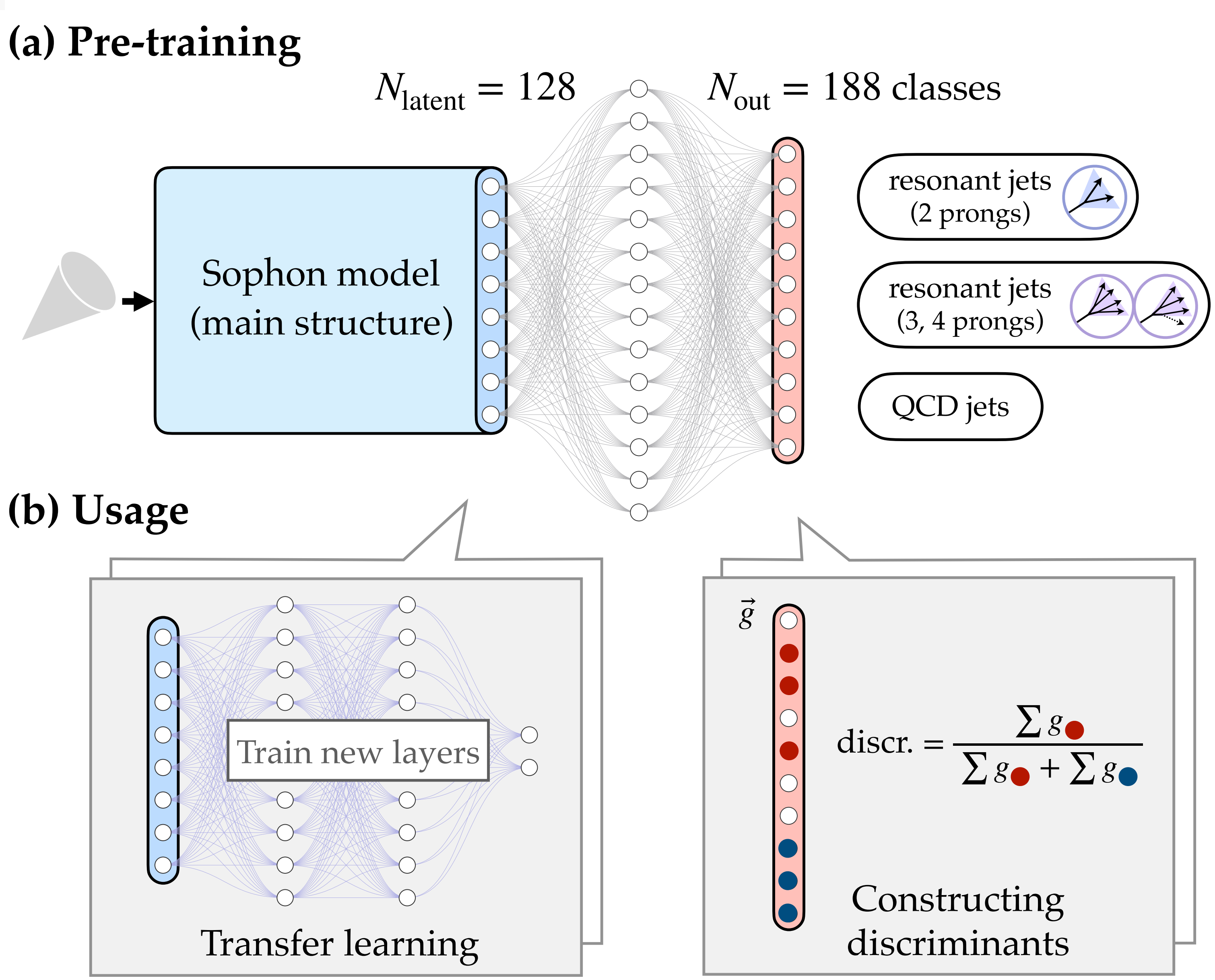


and we plan to update the CMS tagger under Sophon's philosophy

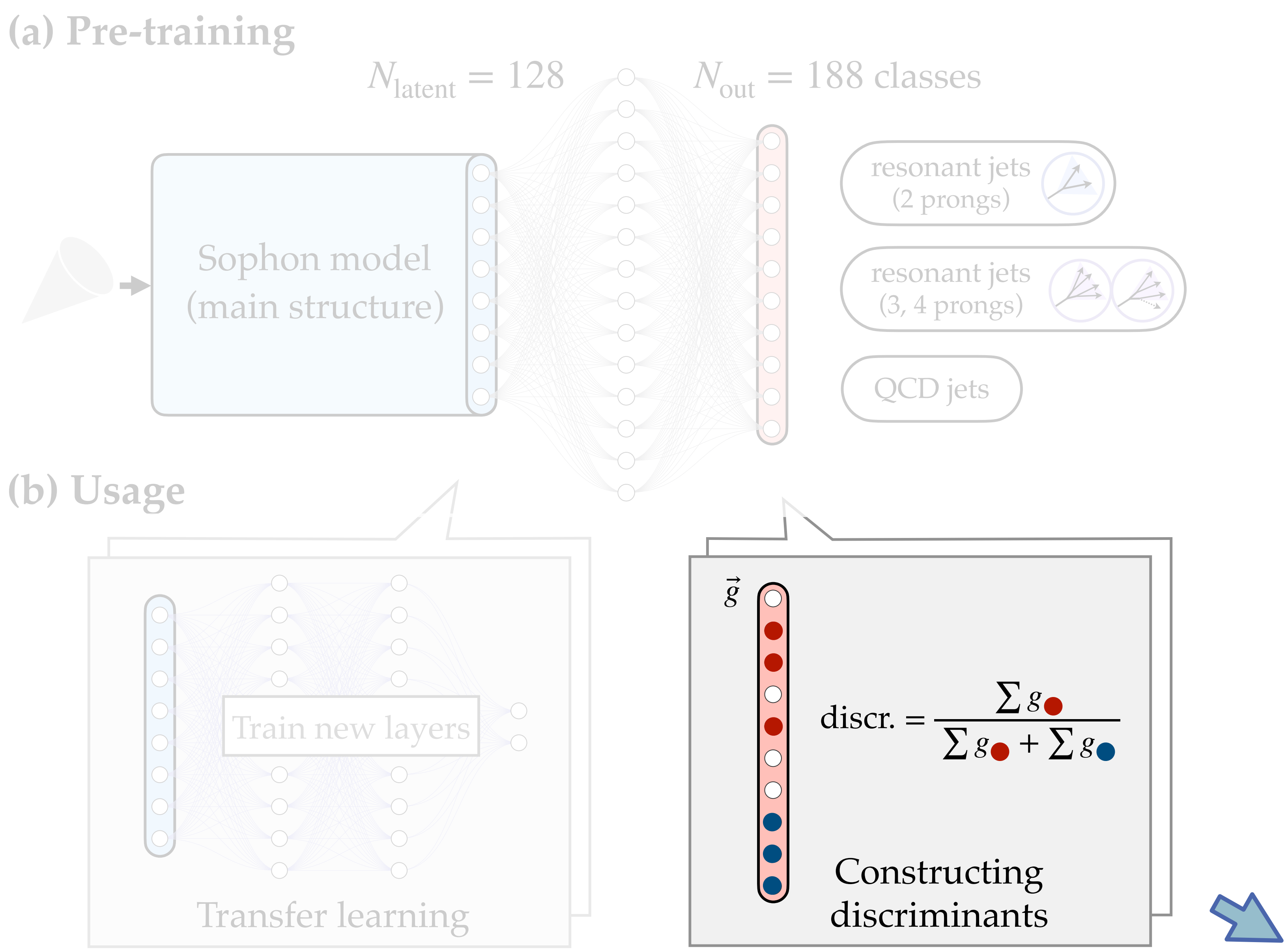
please stay tuned!

Implications for LHC resonance search

Using Sophon



Using Sophon

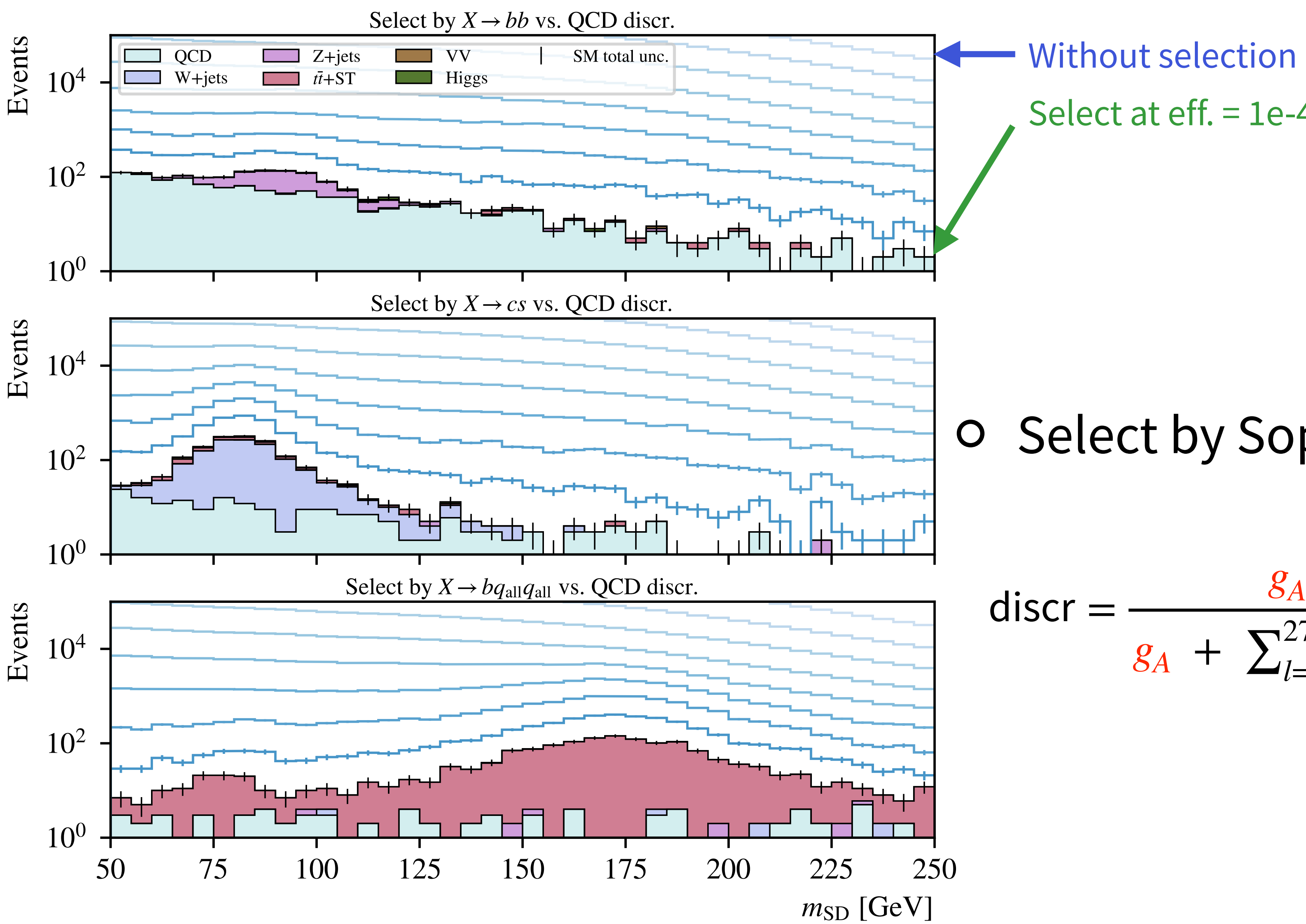


Use it out of the box!

Construct a dedicated discr.
→ perform a bump hunt

Can we rediscover the SM particles?

- ➔ Simulate 40fb^{-1} LHC collision events, $\sqrt{s}$ = 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**

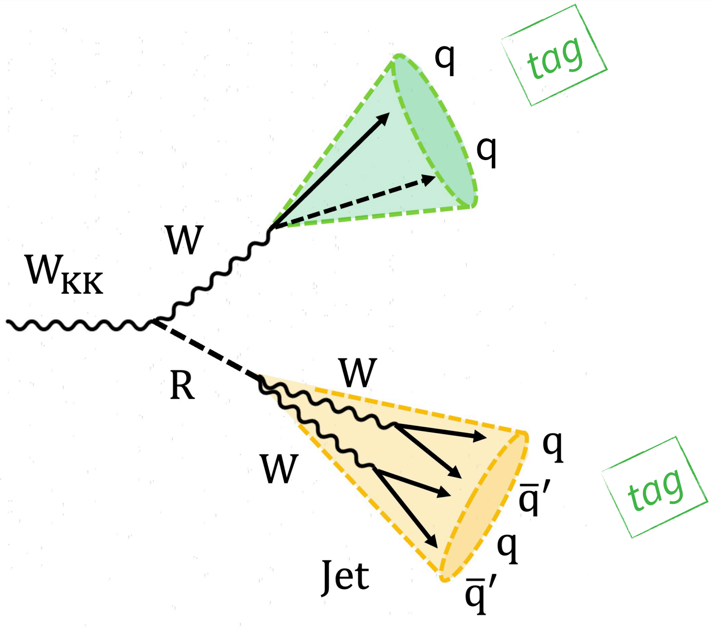


- Select by Sophon's different discriminants

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

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.

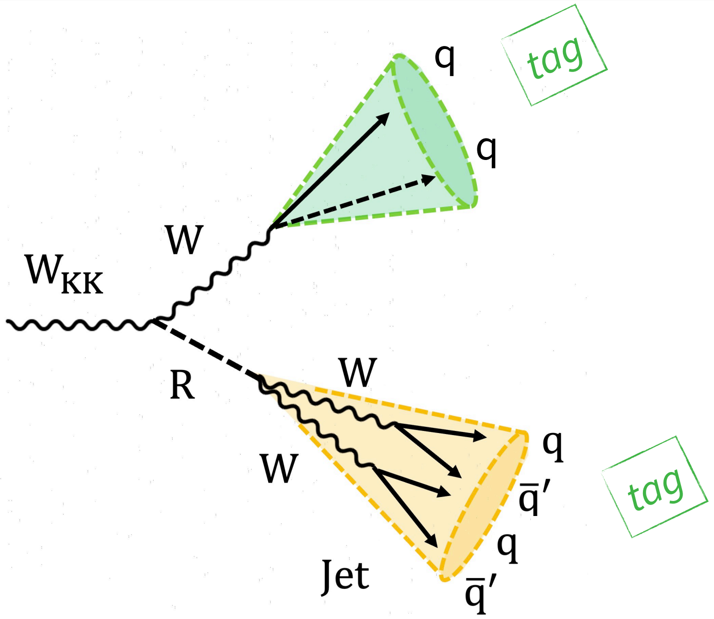


$$\text{discr} = \sum_{\text{jet}=1,2} \frac{g_{A,\text{jet}}}{g_{A,\text{jet}} + \sum_{l=1}^{27} g_{\text{QCD}_l,\text{jet}}}$$ (sum for jets 1, 2)

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

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.

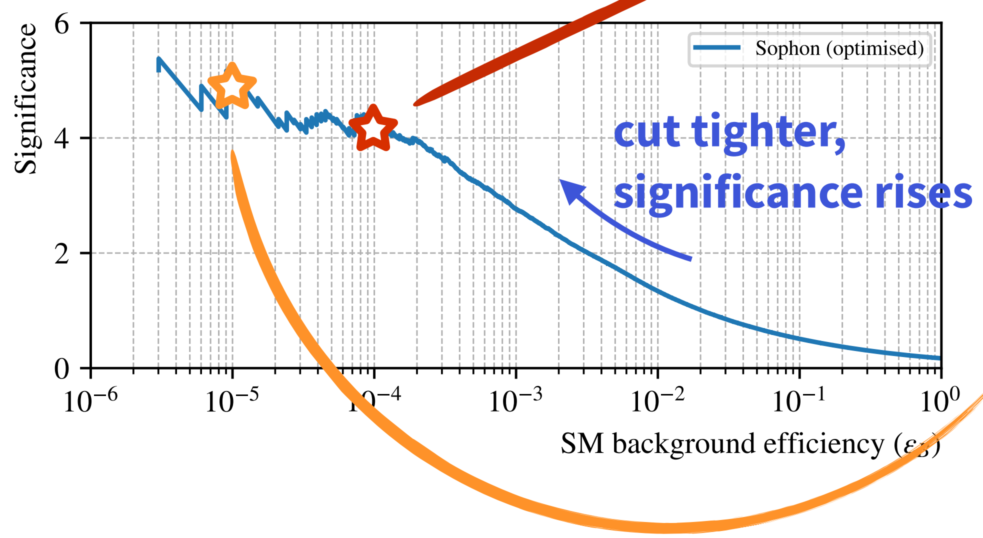
$$\text{discr} = \sum_{\text{jet}=1,2} \frac{g_{A,\text{jet}}}{g_{A,\text{jet}} + \sum_{l=1}^{27} g_{\text{QCD}_l,\text{jet}}}$$ (sum for jets 1, 2)

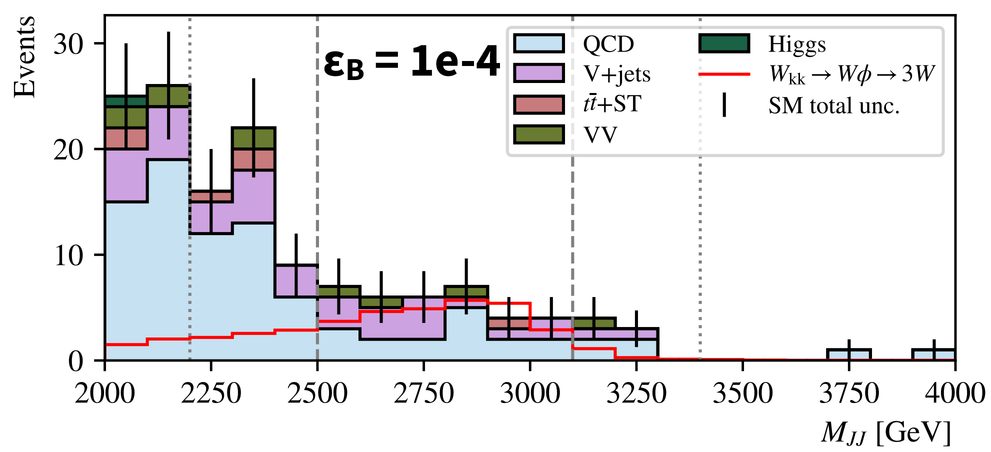
$$A = \begin{cases} 0.3 \times \{cs, qq\} \\ + 0.1 \times \{ccss, qqcs, qqqq\} \\ + 0.6 \times \{ccs, ccq, ssc, ssq, qqc, qqs, qqq\} \end{cases}$$

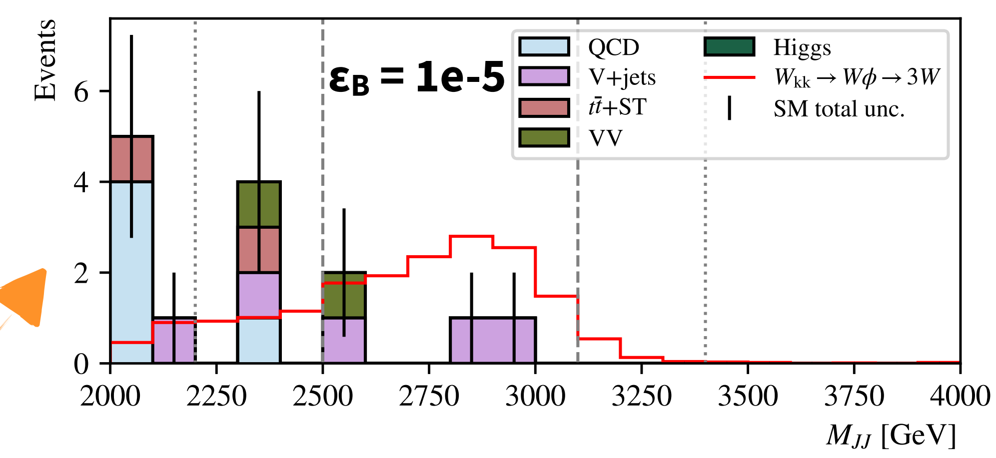
Search significance

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

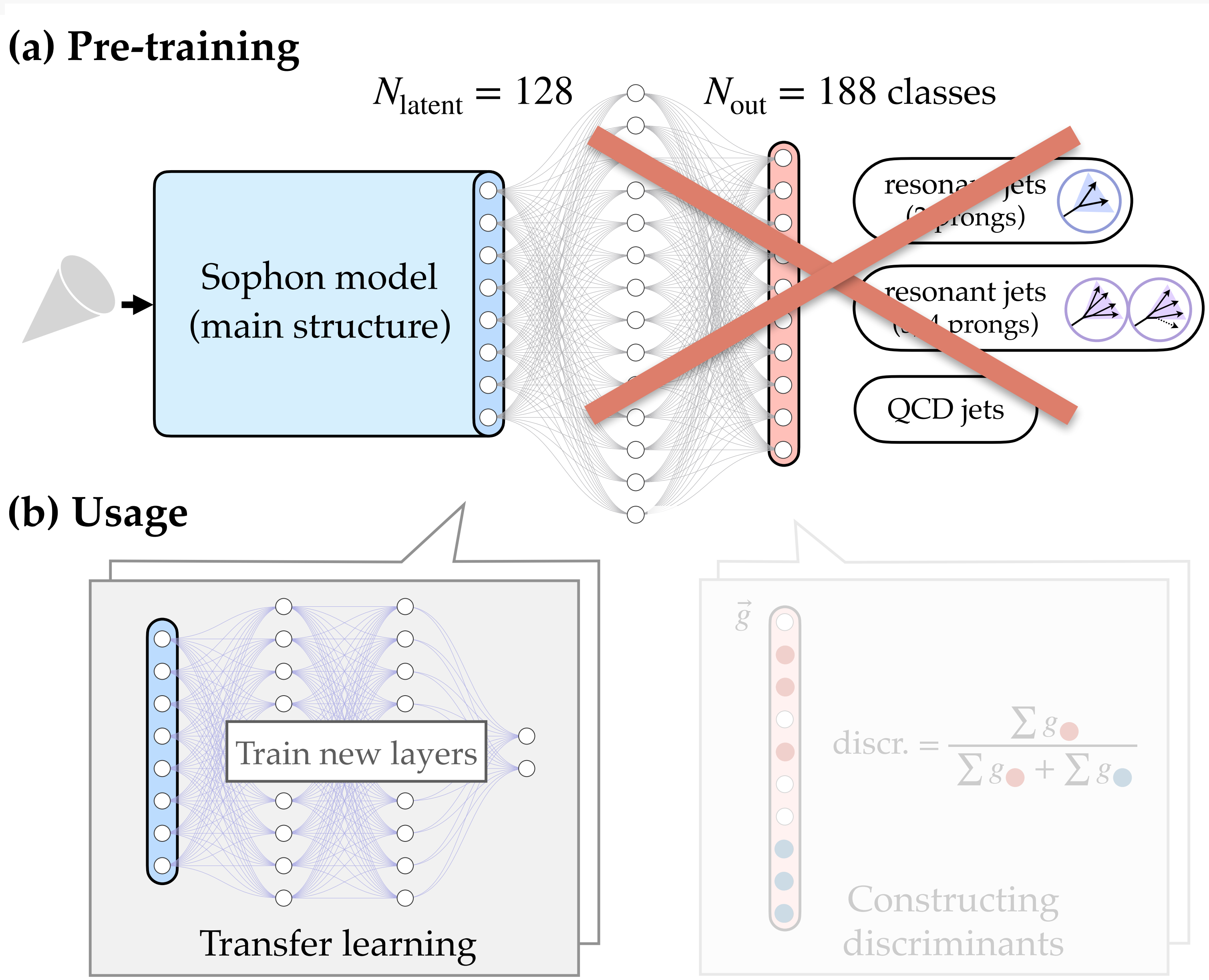
in dijet inv. mass window **2500–3100 GeV**







Sophon's transfer learning



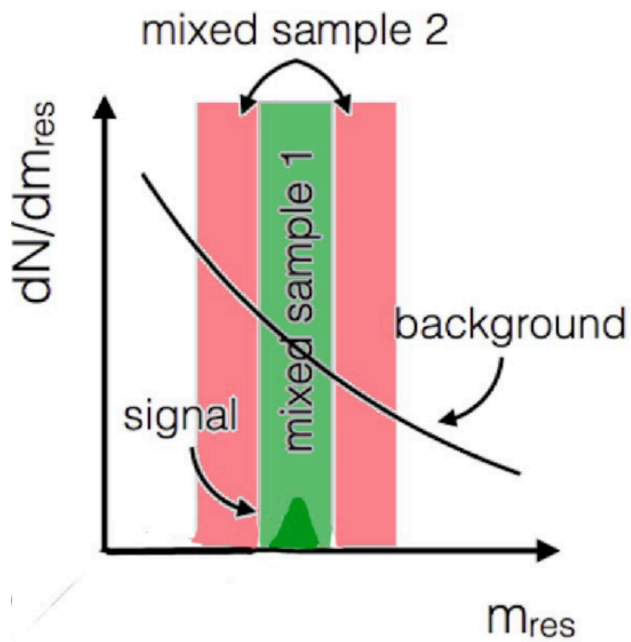
- Transfer to uncovered tagging scenarios…
- facilitate anomaly detection (weakly-supervised, autoencoder)…
- *more potential to unlock!*

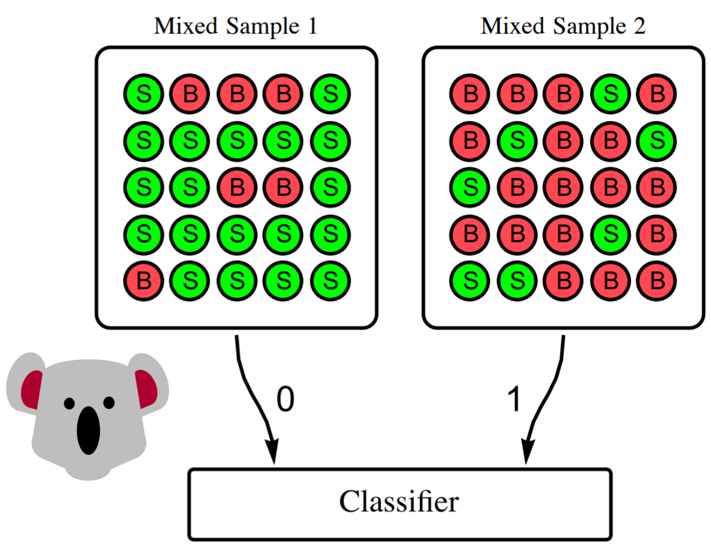
Use it out of the box!

Construct a dedicated discr.
→ perform a bump hunt

Background: anomaly detection in weakly-supervised approach

JHEP 10 (2017) 174





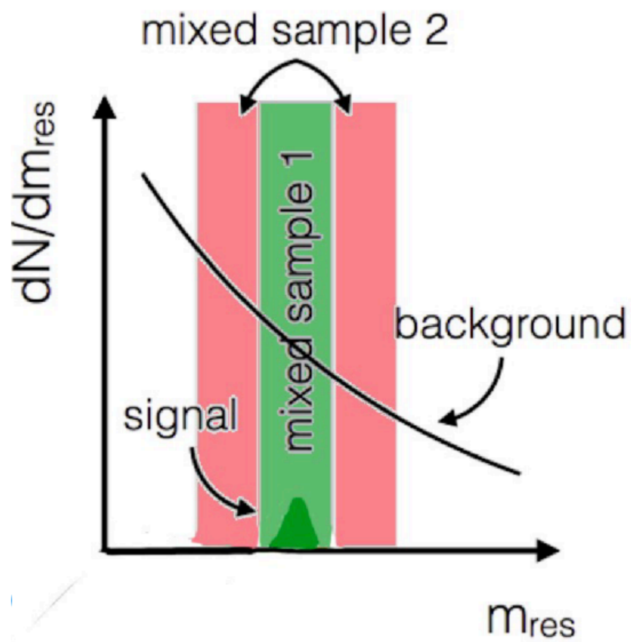
Equivalent effect for training **S** vs **B**

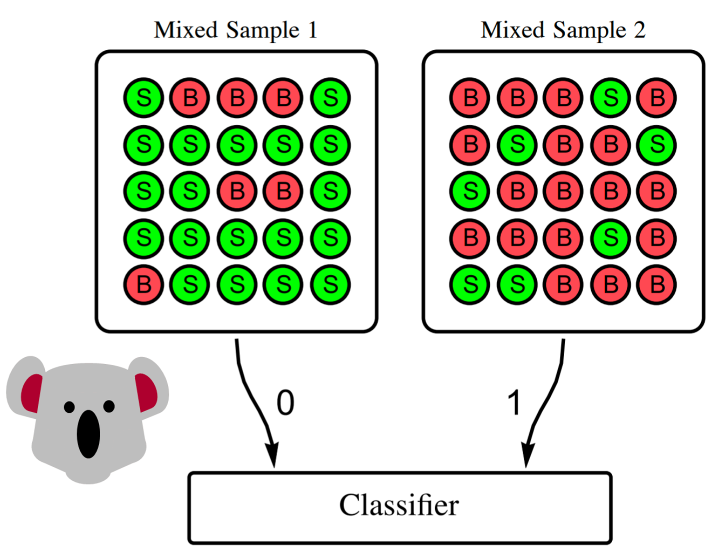
➔ Recall the early work: CWoLa (classification without labels) Hunting

- 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

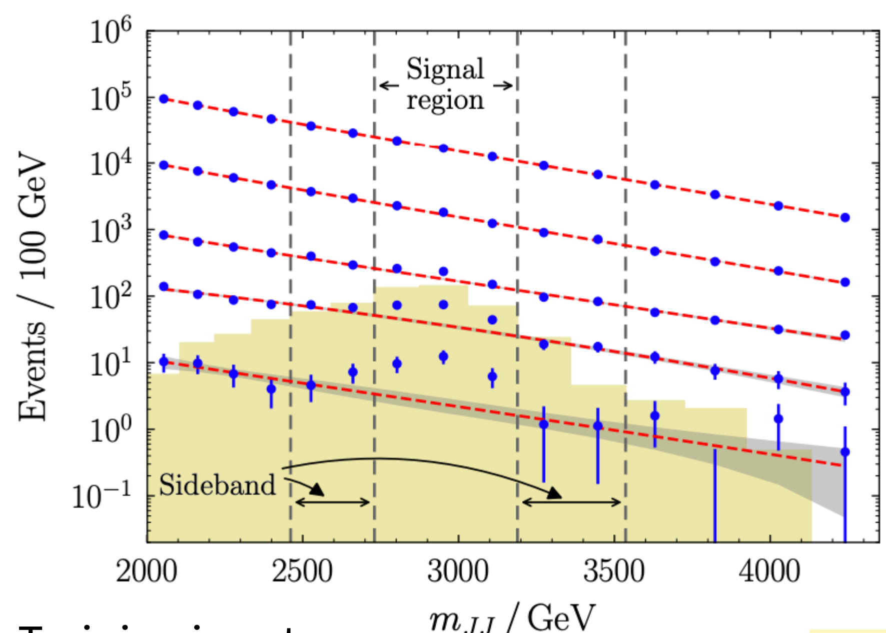




Equivalent effect for training **S** vs **B**

➔ 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

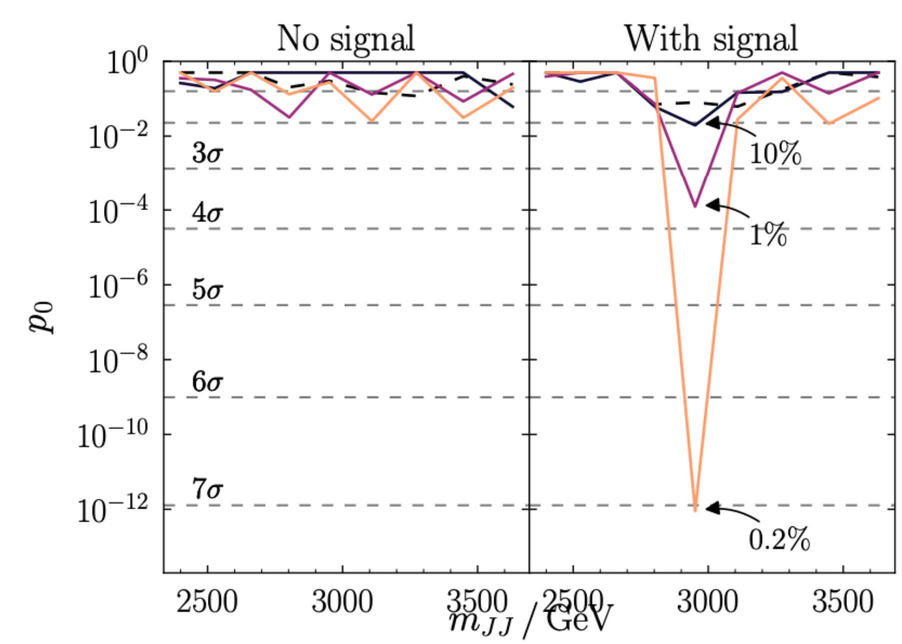


Training input

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

can discover $W' \to W\phi \to WWW$ signals

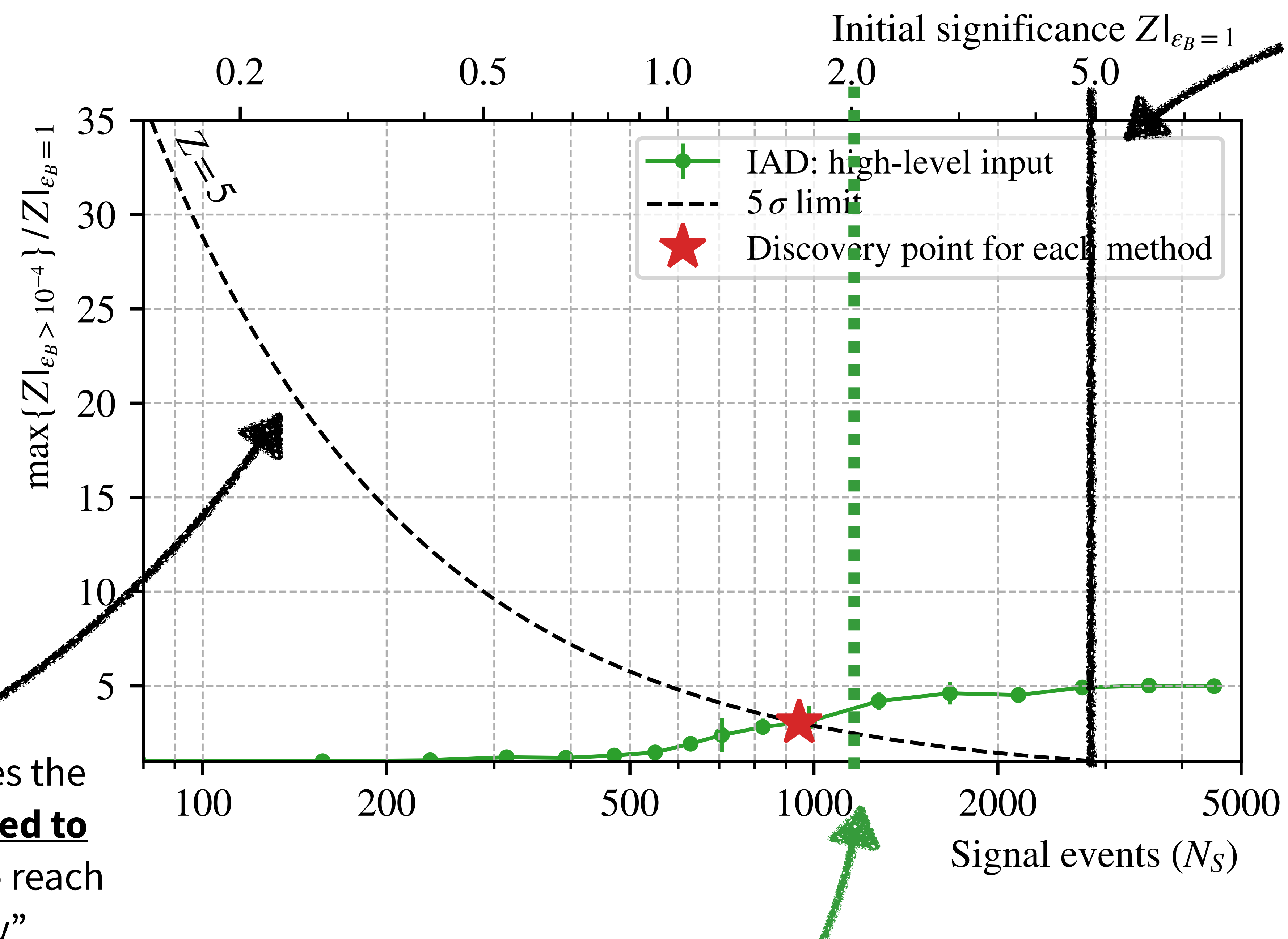
see 2σ → 7σ improvement



PRL, 121 (2018) 24, 241803

PRD, 99 (2019) 1, 014038

Dijet search capabilities

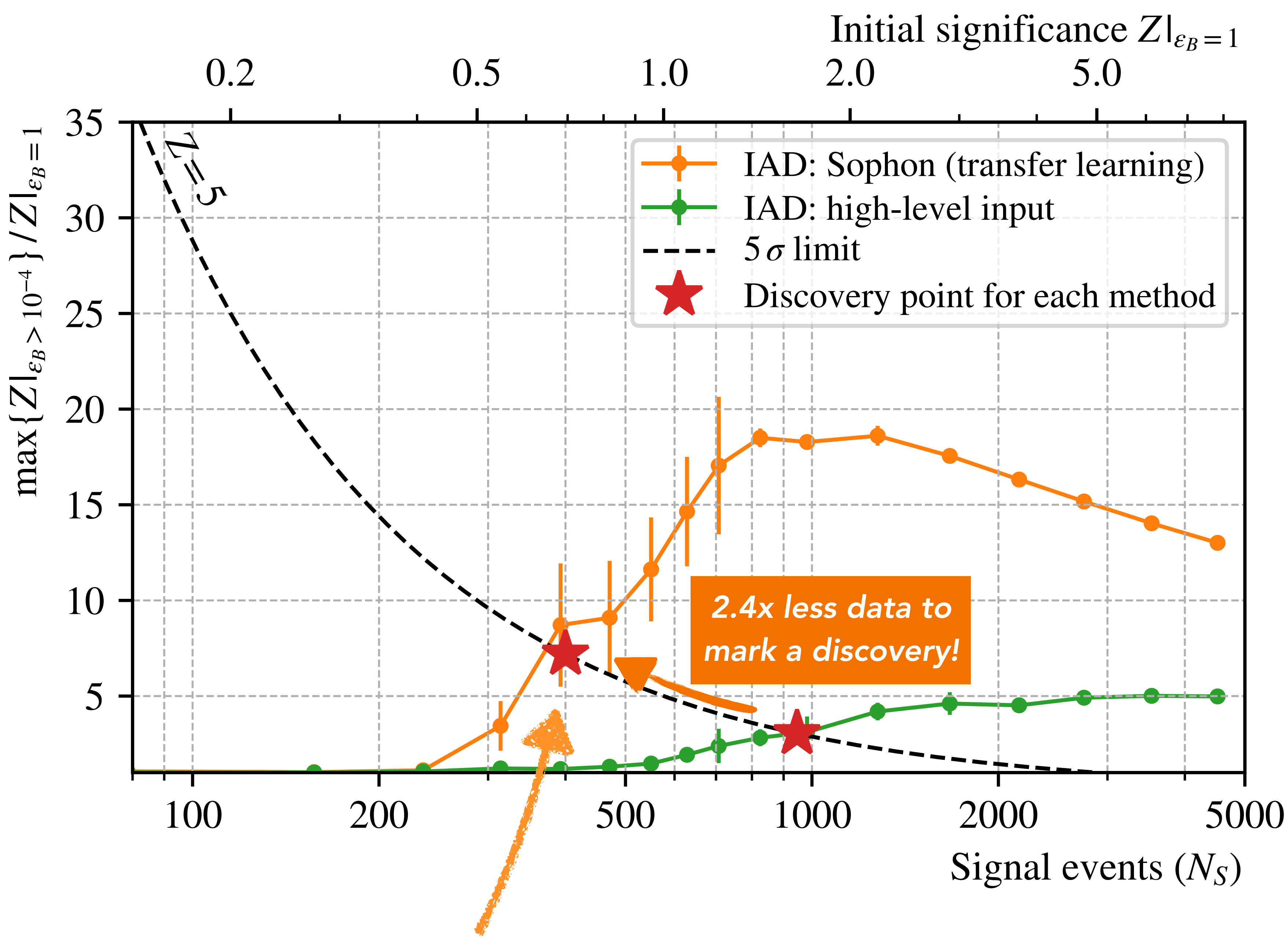


"If signal events reach this point, **with initial *Z*=5**, then we have already discovered the signal without needing to make a cut"

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

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

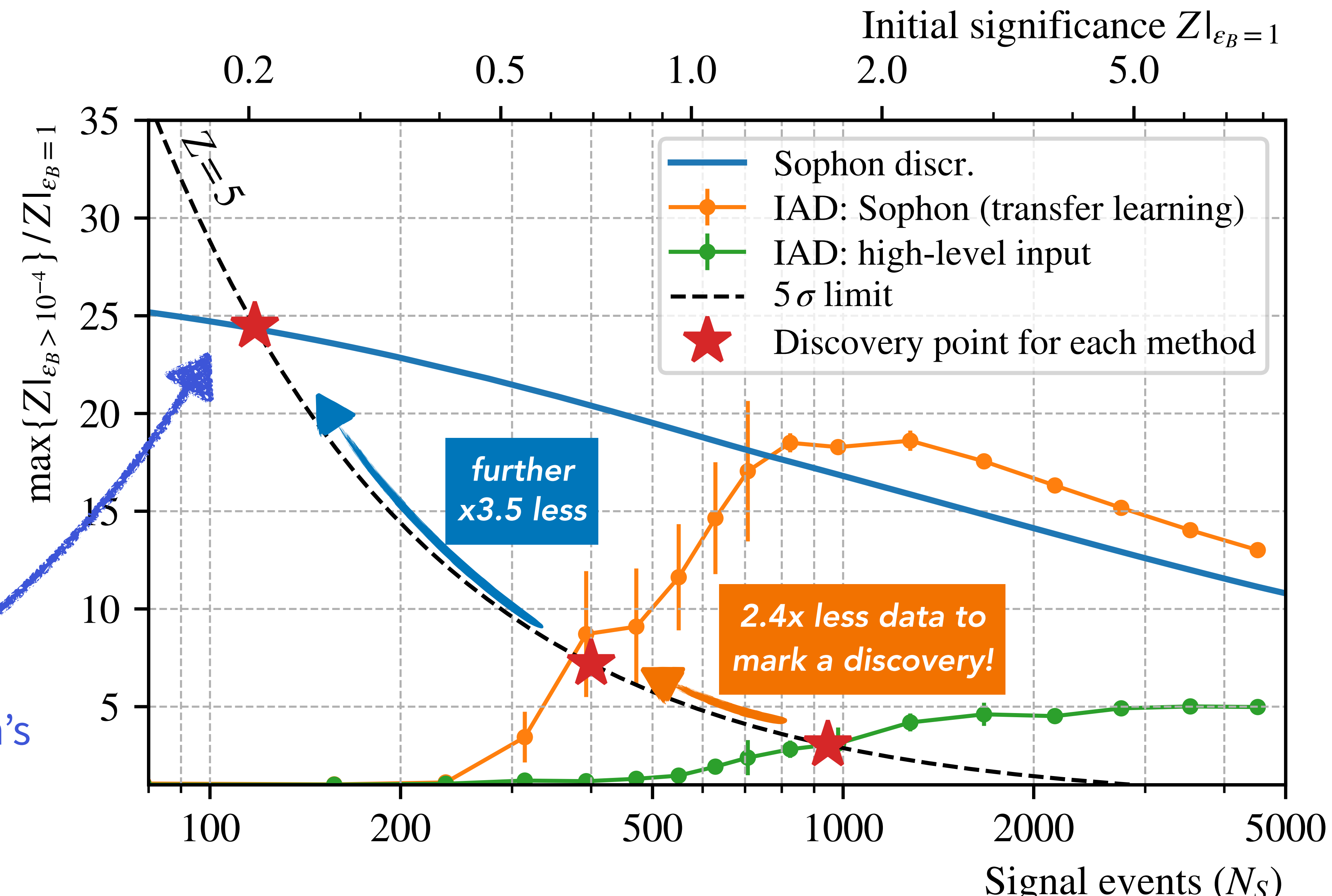
Dijet search capabilities



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

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

Dijet search capabilities



using Sophon's constructed discriminant

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

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

Summary and outlook

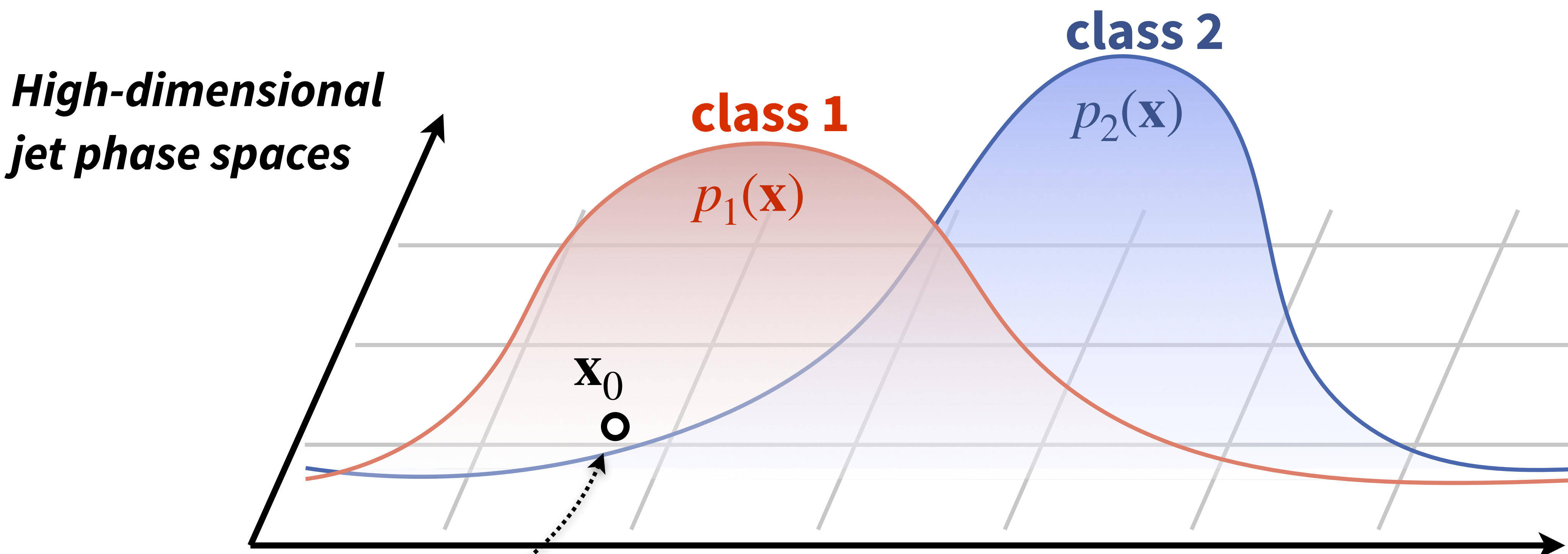
arXiv:2405.12972

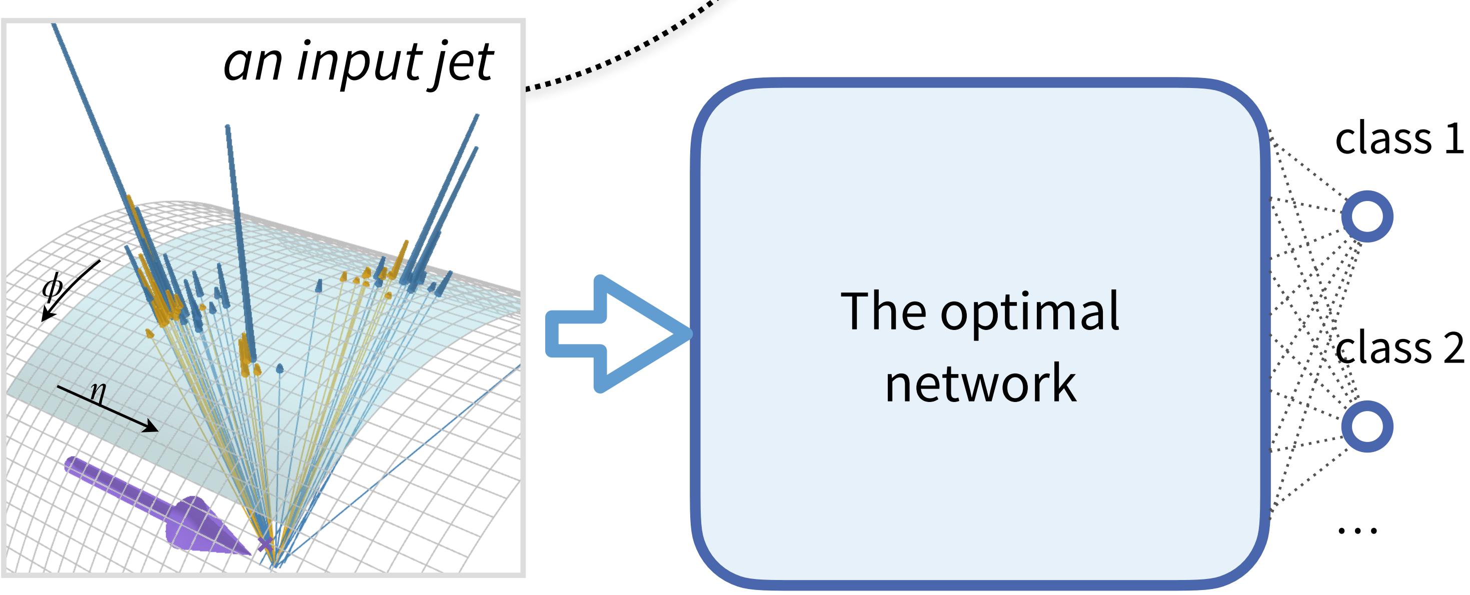
- Sophon releases a lot of new opportunities for future LHC experiments
  - simply viewed as a **"global large-*R* jet tagger"** → should bring benefits of the advanced NN to ~all hadronic final-state searches
  - also viewed as a pre-trained jet model: a foundation model tailored for LHC analyses
- Proposed the **JetClass-II** dataset and the **Sophon** model
  - **JetClass-II** covers more comprehensive phase spaces and can be a good playground to develop future foundation models
  - the **Sophon** model can be helpful in delivering future LHC pheno research! [see implementation details on our Github repo]
    - optimizing sensitivity for dedicated searches/anomaly detection/novel paradigms…
  - this work demonstrates that it can be a great booster to LHC's broad resonance search programs
- Stay tuned to their applications to real LHC experiments!

Backup

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



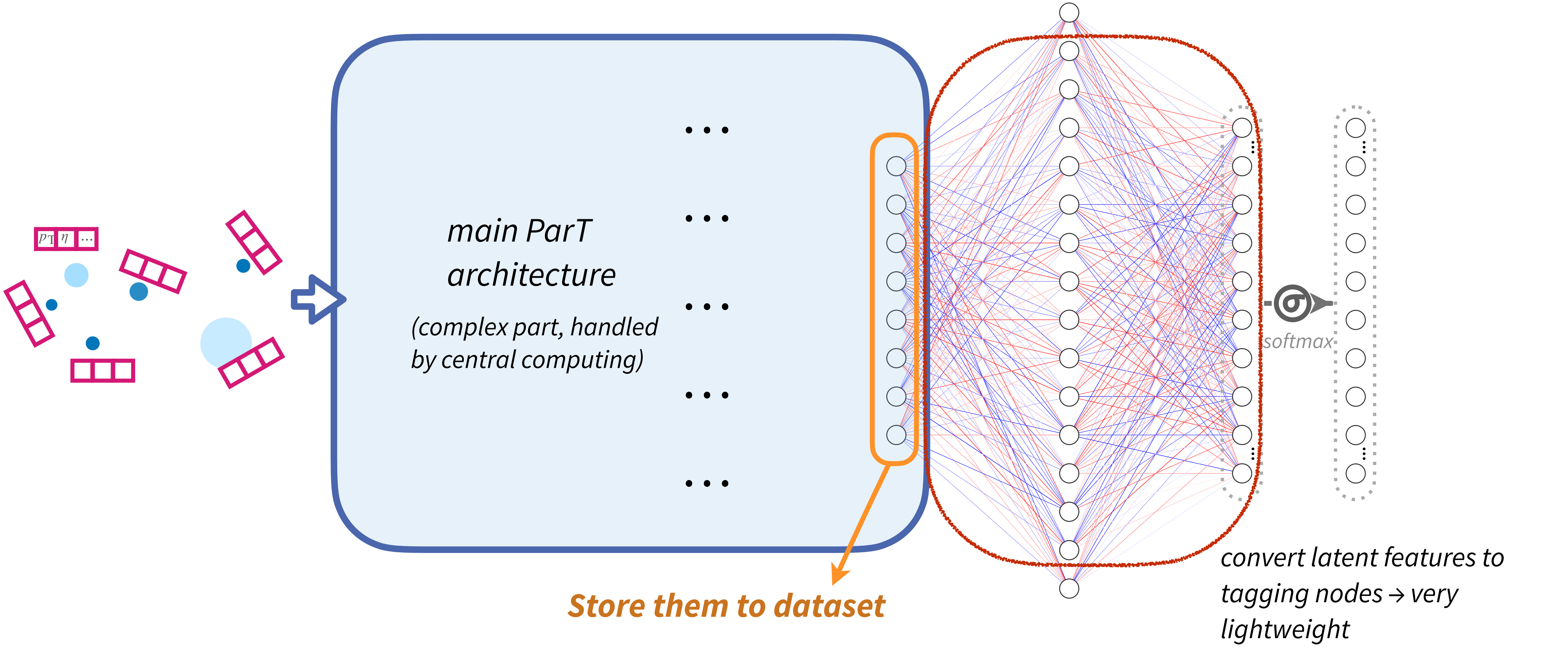


- Ideal classifier network results in $g_1 : g_2 : \ldots = p_1(\mathbf{x}_0) : p_2(\mathbf{x}_0) : \ldots$
- It is a direct estimation of p
- The **network capacity** decides how close the estimation is

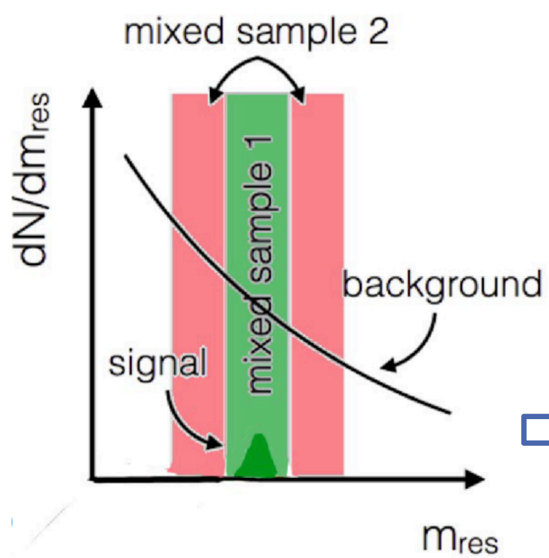
How to deploy the model to LHC experiments?

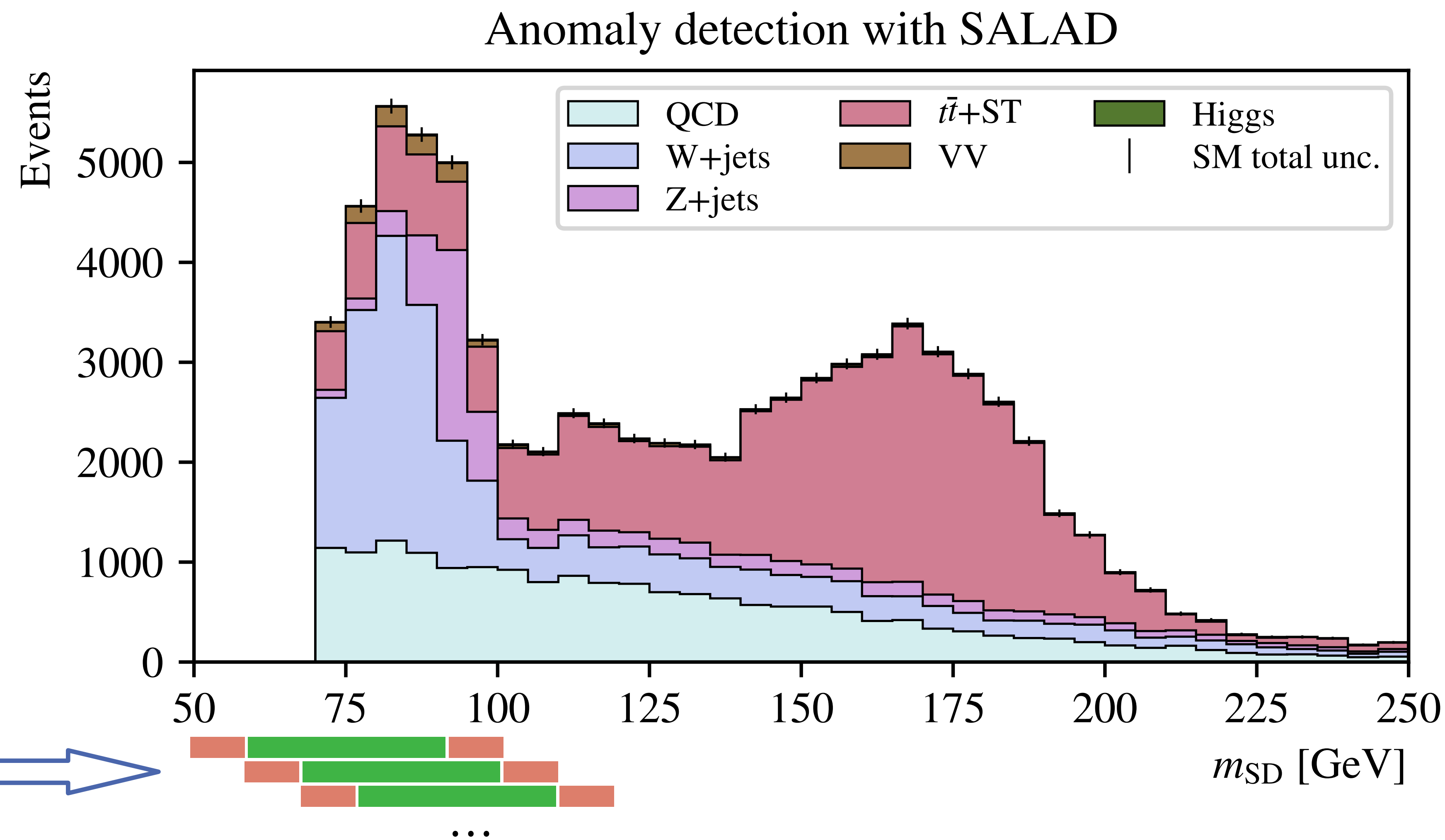
→ Implies how we can do future analysis

- hidden layer neurons values are stored in official sample
- analysis can use them for fine-tuning (equivalently, just think that they are special jet variables)
- easy to implement & integrate into existing workflow



Sophon's transfer learning × anomaly detection





➔ Do SALAD (similar to CWOLA Hunt) in each sliding window

- ❖ purify those peculiar jets in that mass window

➔ Sophon's latent space has encoded fruitful knowledge on "final-state properties"