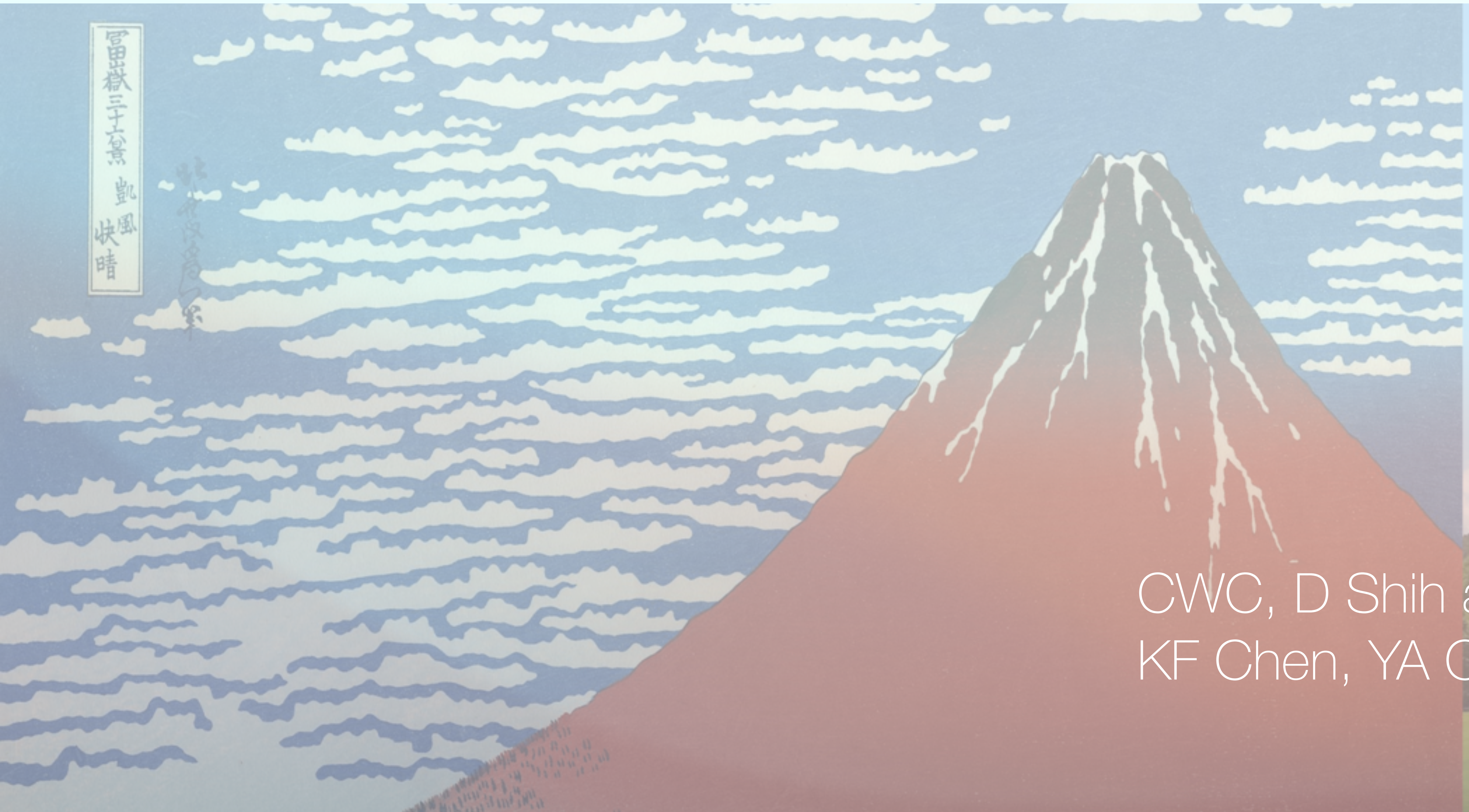
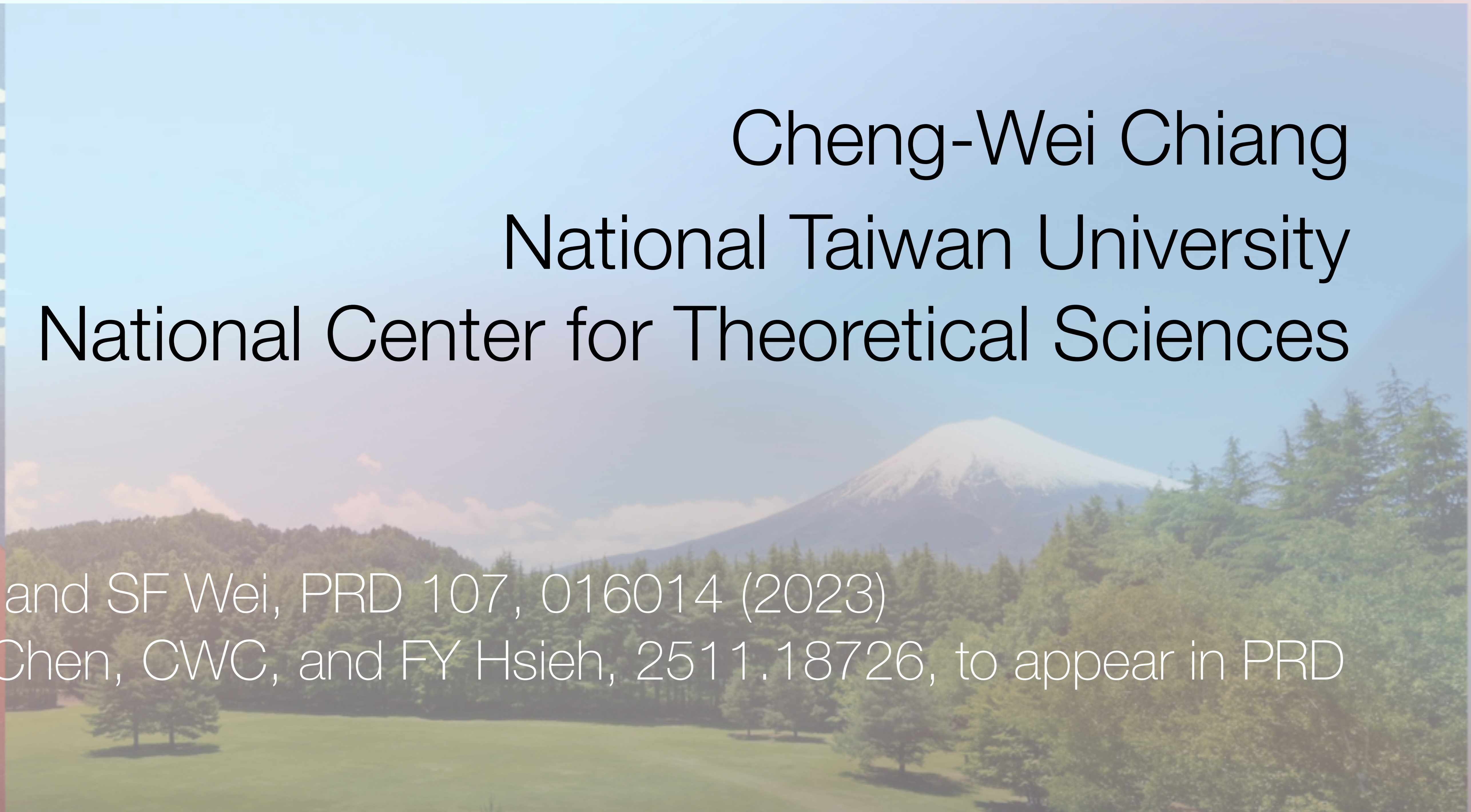
Summer Institute @ Fuji Calm
August 9, 2026

# Higgs Production Classifier using Weak Supervision



Cheng-Wei Chiang
National Taiwan University
National Center for Theoretical Sciences

CWC, D Shih and SF Wei, PRD 107, 016014 (2023)
KF Chen, YA Chen, CWC, and FY Hsieh, 2511.18726, to appear in PRD

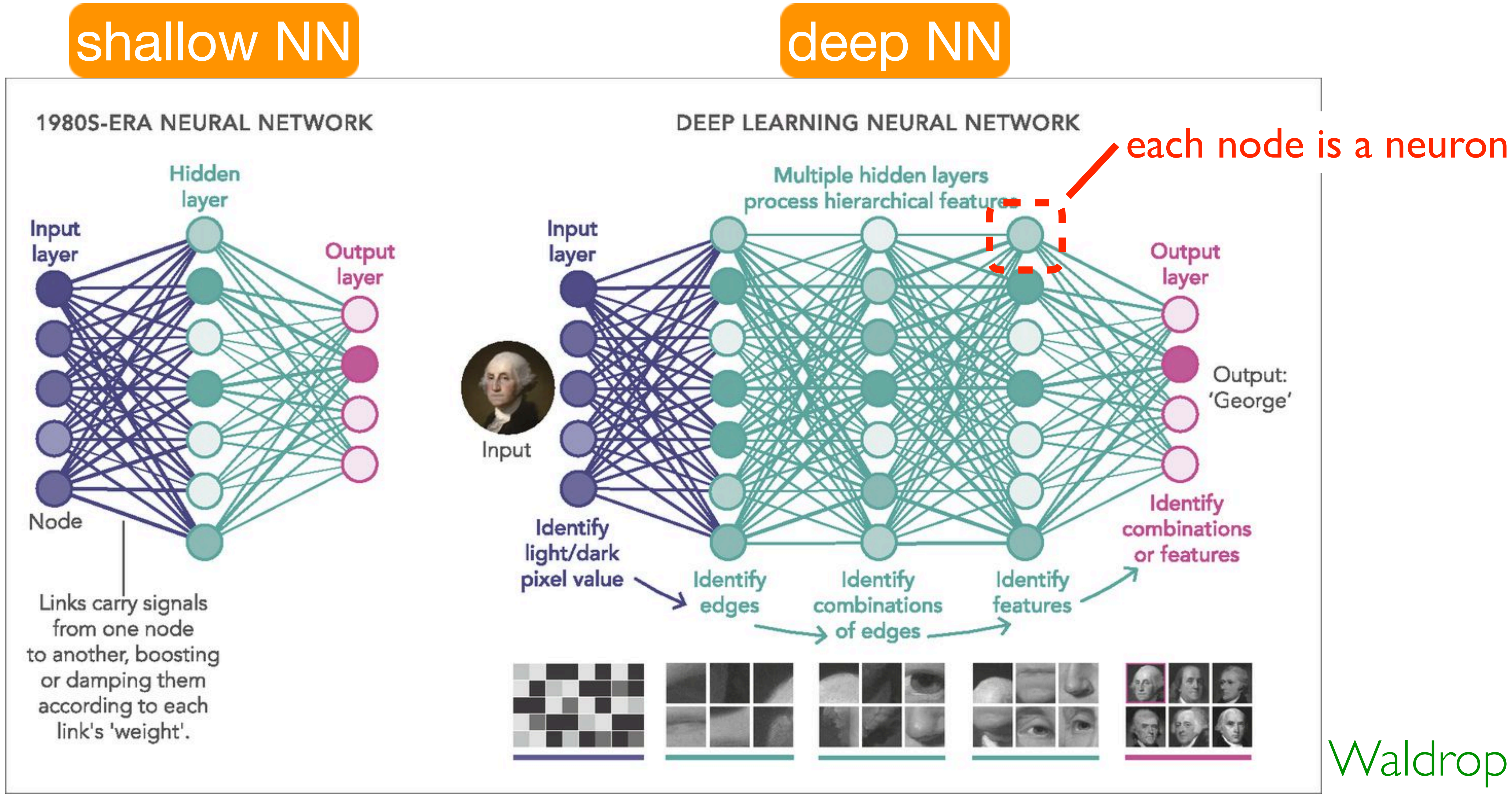
# Revolution is Driven by New Tools

"New directions in science are launched by **new tools** much more often than by **new concepts**. The effect of a concept-driven revolution is to **explain old things in new ways**. The effect of a tool-driven revolution is to **discover new things that have to be explained**."

— Freeman J. Dyson, *Imagined Worlds*
Harvard University Press (1998)

# Machine Learning

- **Machine learning (ML)** is a **new tool** used for **large-scale** data processing and well-suited for **complex** datasets with huge numbers of **variables** and **features** (patterns and regularities), especially for **deep learning neural networks (NNs)**.
- **The Universal Theorem**: any function can be approximated by a neural network with at least one hidden layer.

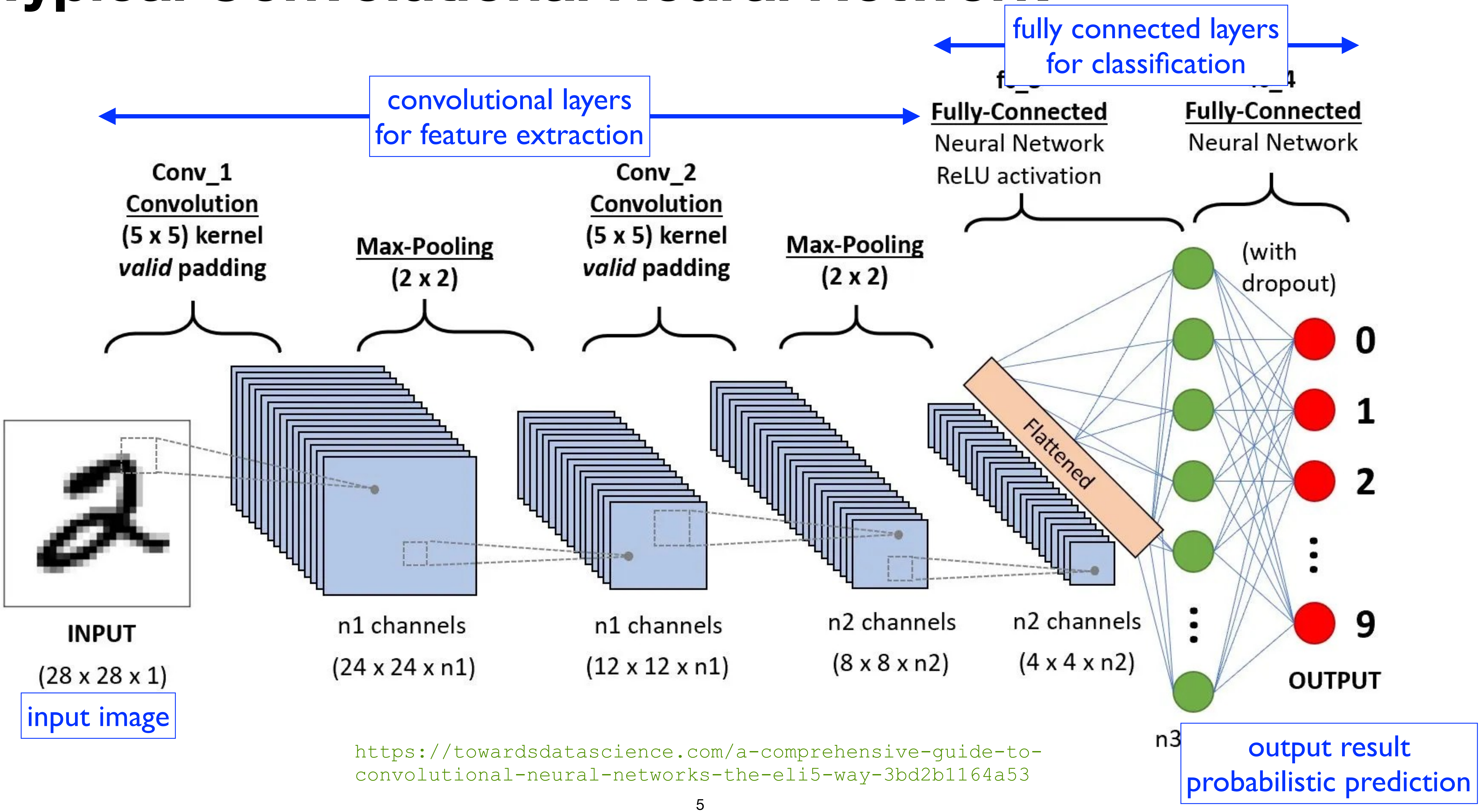



Waldrop 2019

# Types of Machine Learning

- **Fully supervised learning**
  - Training data with labels (e.g., recognizing photos of cats and dogs)
- **Unsupervised learning**
  - Training data without labels (e.g., analyzing and clustering unlabeled datasets)
- **Reinforced learning**
  - Data from interactions with the environment (e.g., chess and Go games)
- **Weakly supervised learning**
  - Training data whose labeling is *infeasible*, *imperfect*, *difficult*, or *expensive* (e.g., medical imaging, identifying celestial objects from low-quality telescope images, anomaly searches)

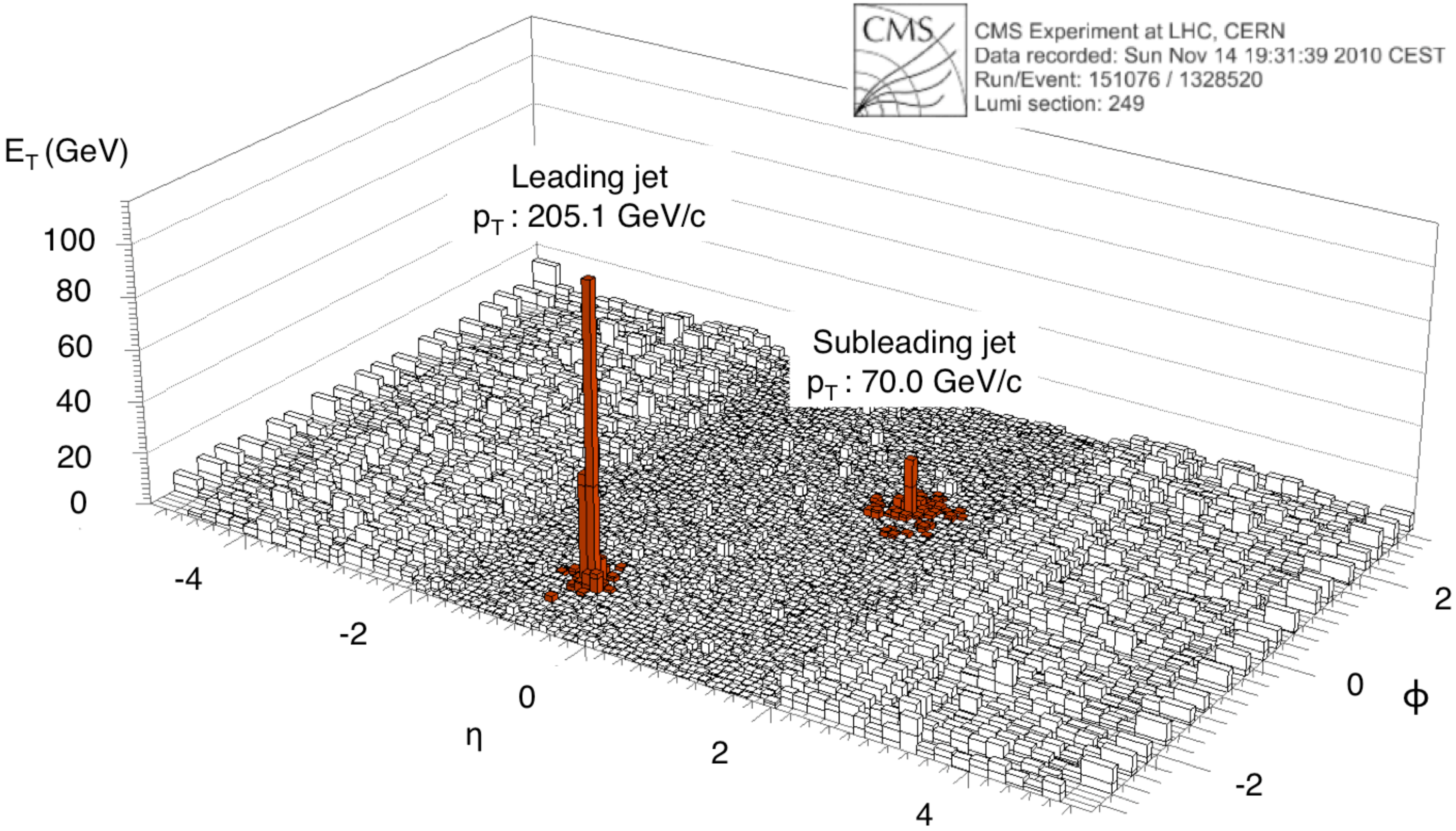
# A Typical Convolutional Neural Network




https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

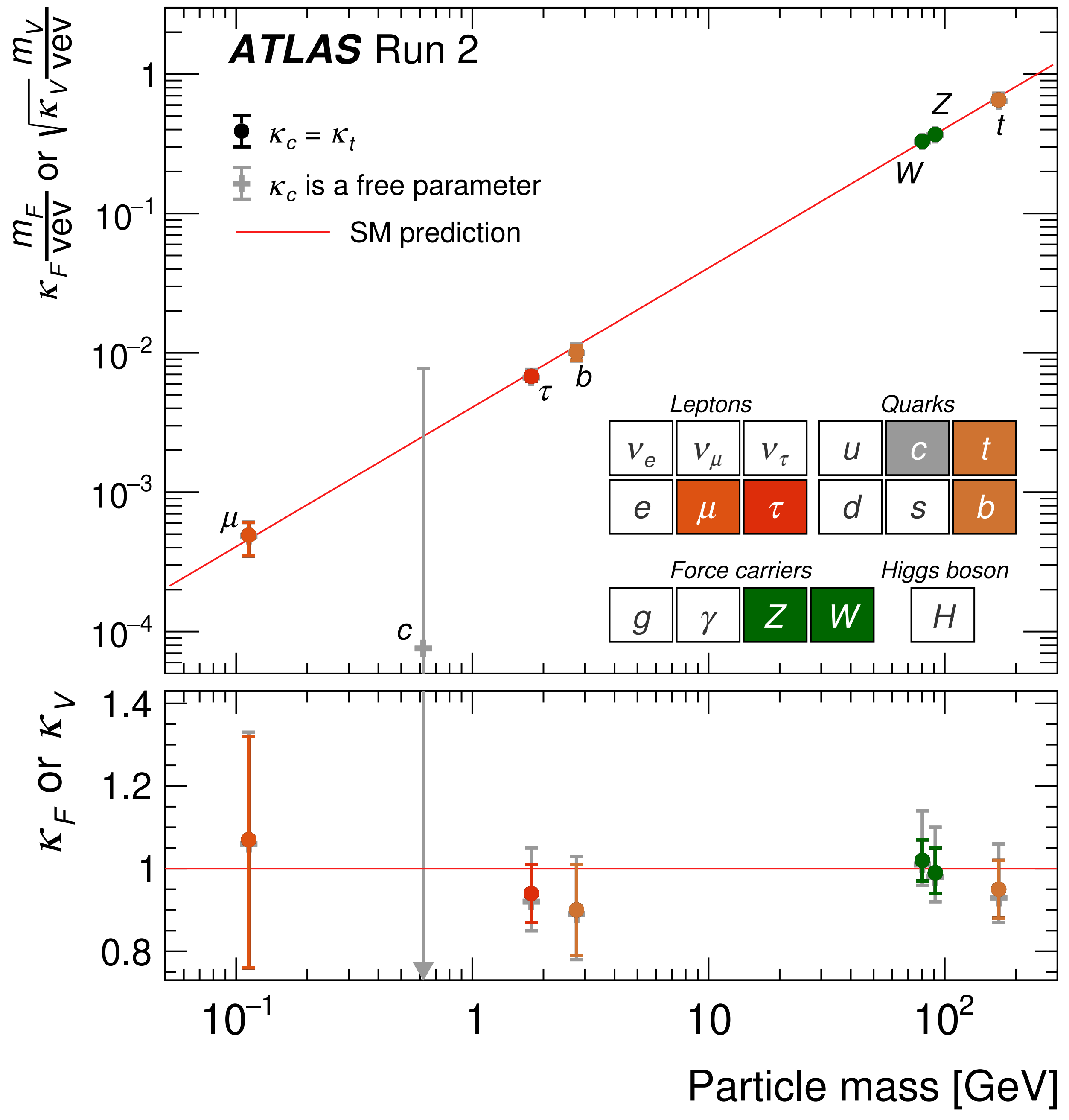
# 2D Image of A Collision Event

- A histogram of jets in the $\eta$-$\phi$ plane.

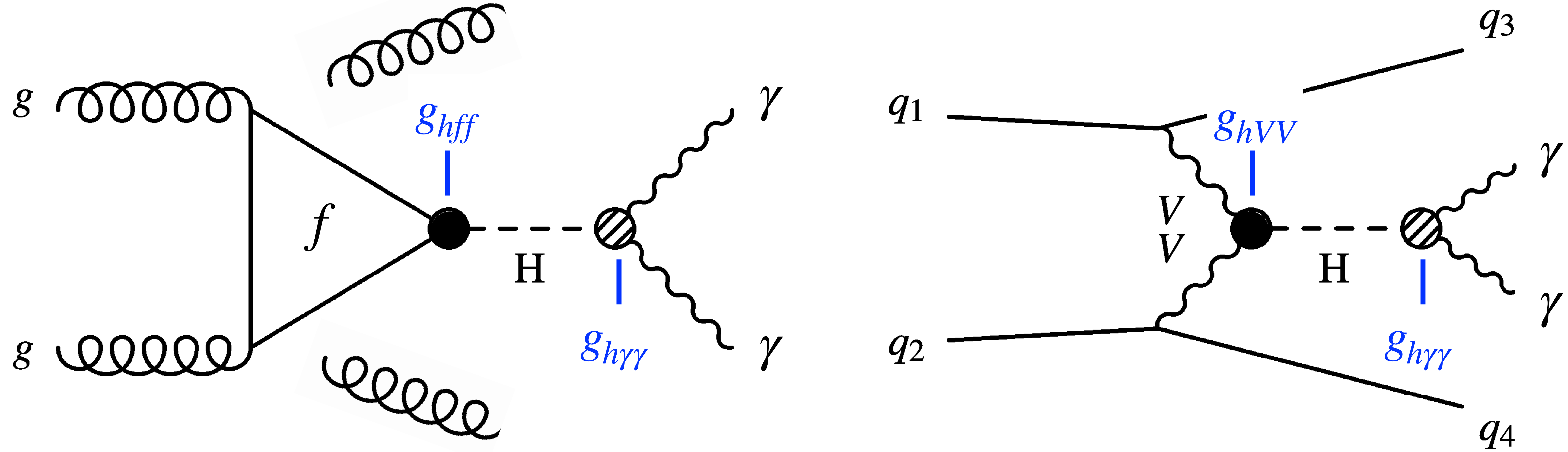


# VBF/GGF Higgs Production

- An important physics program is to determine all the **Higgs couplings** precisely, which requires the ability to discriminate the two dominant production channels.
- Questions:
  - For each **detected** Higgs event, how can we **efficiently** and **correctly** determine/label its production mechanism?
  - Can it be **independent** of how the Higgs boson decays?



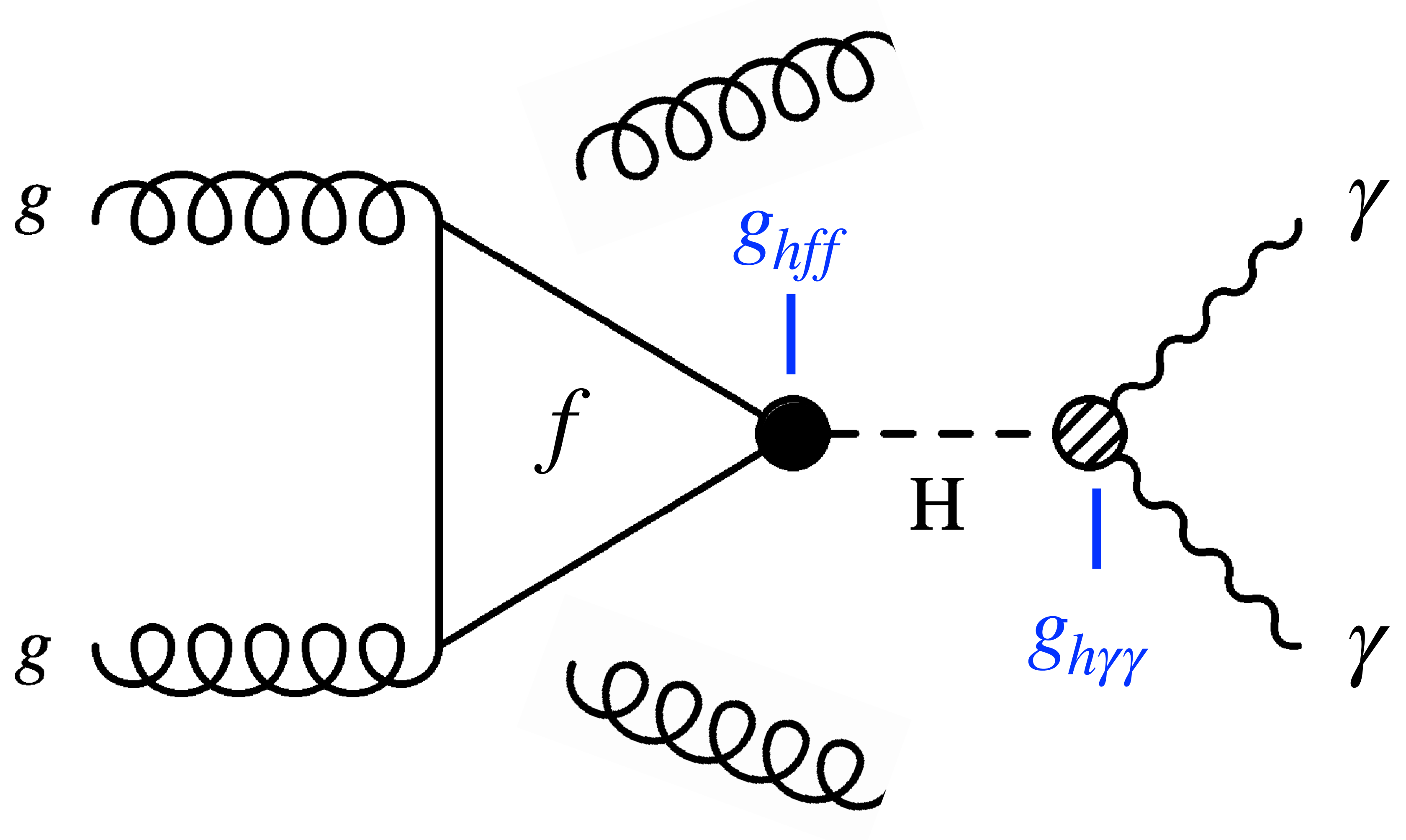

ATLAS 2023



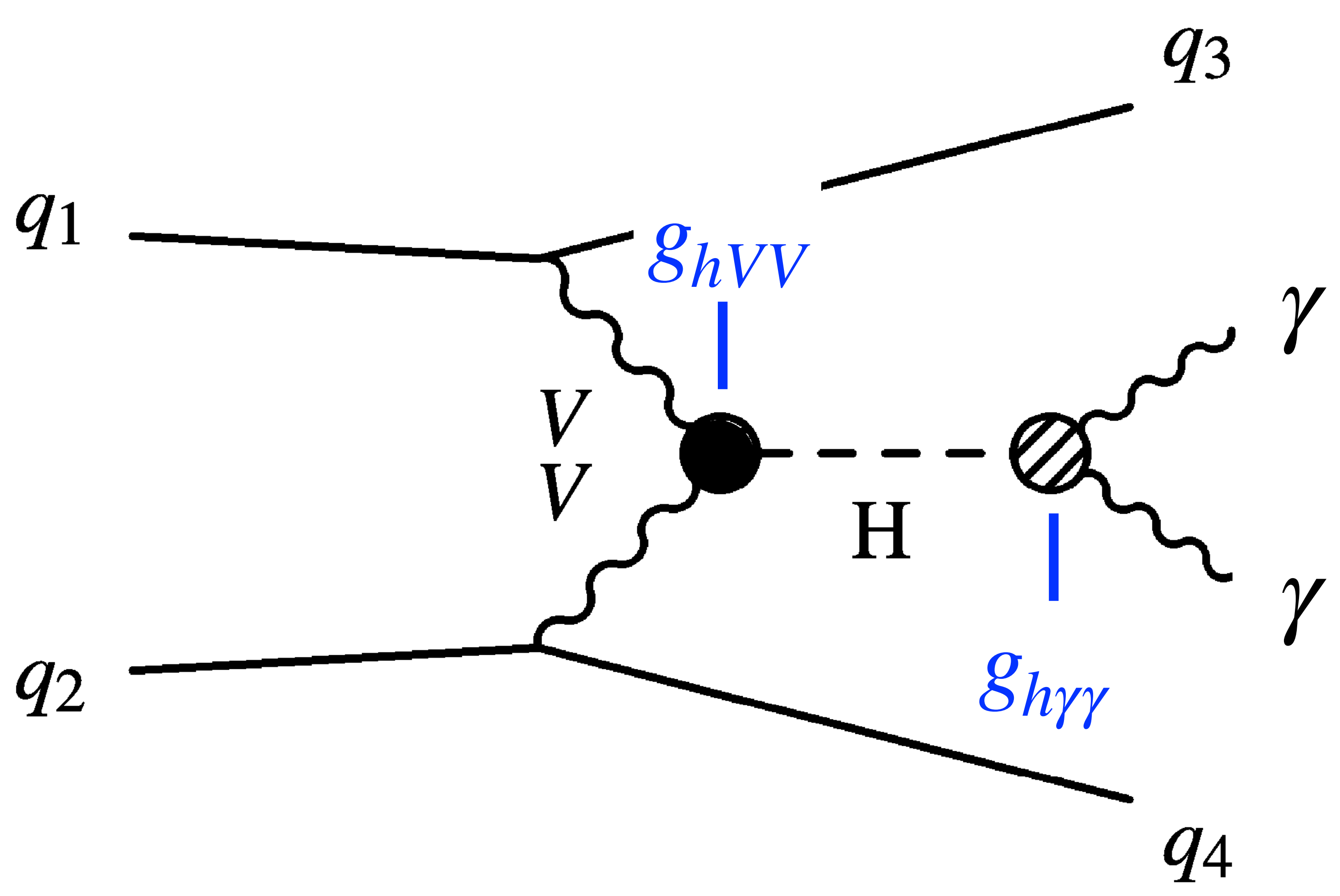

(a) ggF production (b) VBF production

# Two Observations

- VBF events come with two **forward spin-1/2 quark-initiated jets** from the hard process, while GGF jets tend to be **spin-1 gluon-initiated initial-state radiation.**
  ➠ **different jet constituent distributions**, particularly in soft radiation patterns
- Since the Higgs is a **color singlet scalar with a narrow width**, the Higgs decay should be **factorizable** from the VBF or GGF initial state jets, especially for **colorless** final states.
  ➠ **Higgs decay-independent**



(a) ggF production



(b) VBF production

# Fully Supervised Learning

- Under the framework of **full supervision**, we have trained the following classifiers:
  - a **Boosted Decision Tree (BDT),** trained on **high-level features** defined from the **two leading jets** and the Higgs **decay products** as the **baseline** characterizing the prior art.

    ATLAS 2019
  - a **jet-level CNN**, trained to distinguish the two leading jets (**quark vs gluon**), and add the jet-CNN scores to the inputs of the BDT for improvement.
  - an **event-level CNN** to distinguish VBF vs GGF events, using **full-event images** of all the reconstructed particles in the event.
  - an **event-level self-attention** model.

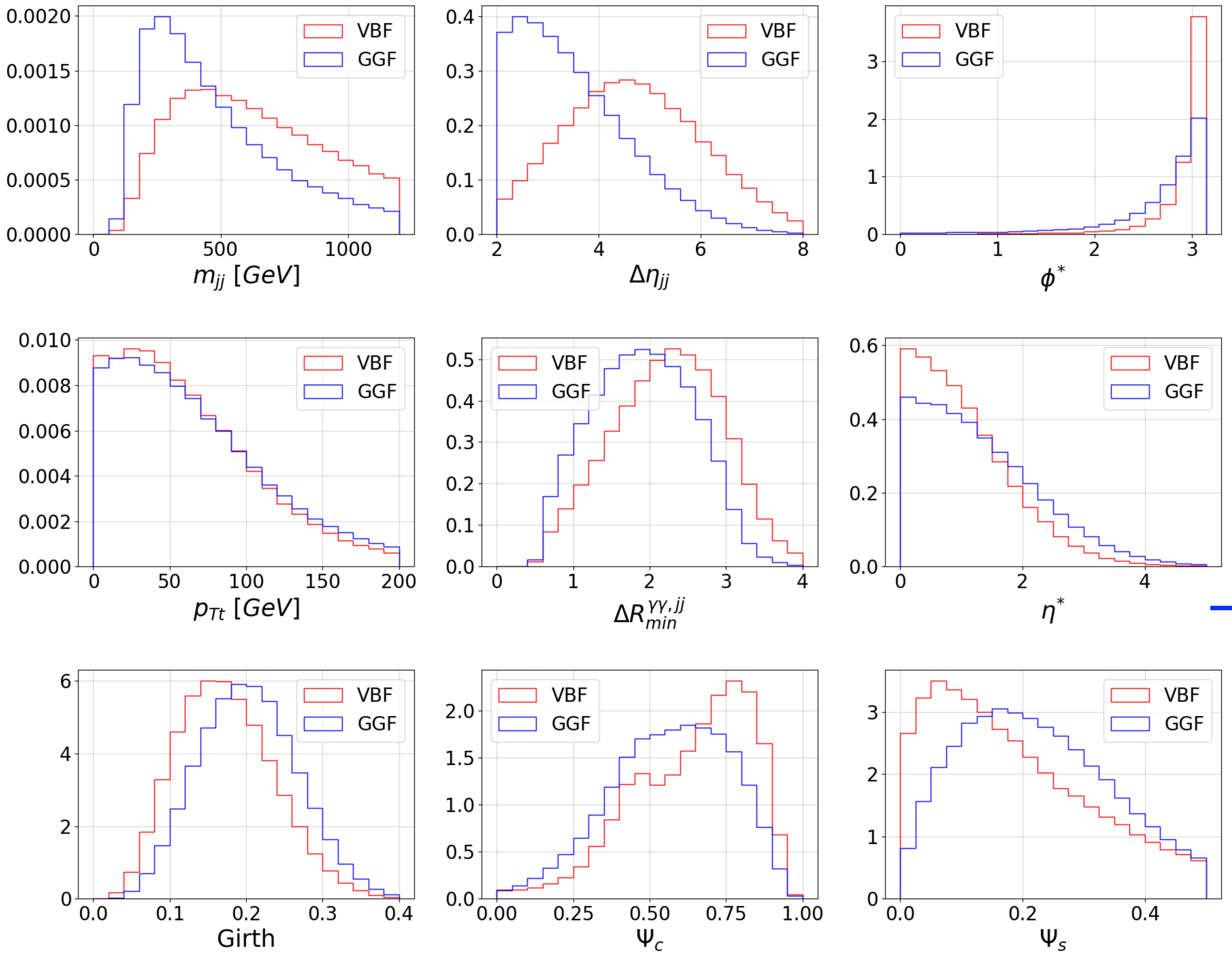
    Lin, Feng, dos Santos, Yu, Xiang, Zhou and Bengio 2017

    Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017

# Distributions of BDT Input Variables

baseline

shapes




- Cut-based methods cannot reach high purity.
- BDTs can achieve a purity of about 70% for the VBF sample and depend on the decay channel.

ATLAS 2019

human-engineered kinetic variables

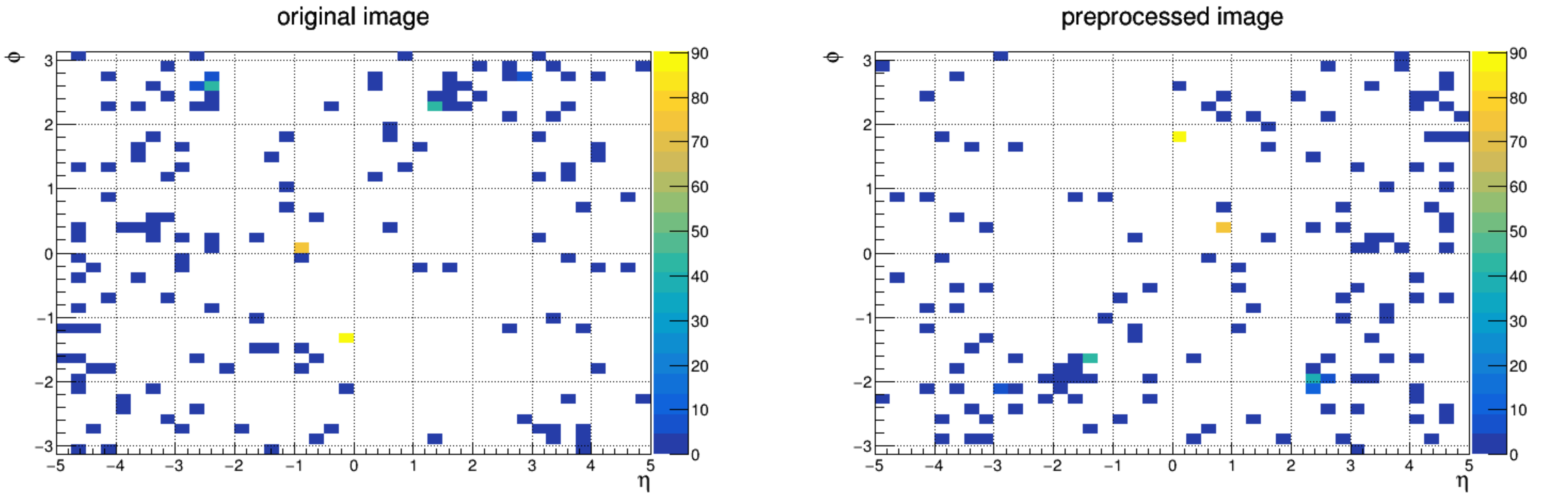
all histograms normalized to have unit area under the curves

# Event-CNN

- Train a **convolutional neural network** (**CNN**) by **full supervision** to discriminate the two production mechanisms by examining the final-state image.
- A successful training typically requires at least **tens of thousands** of samples.

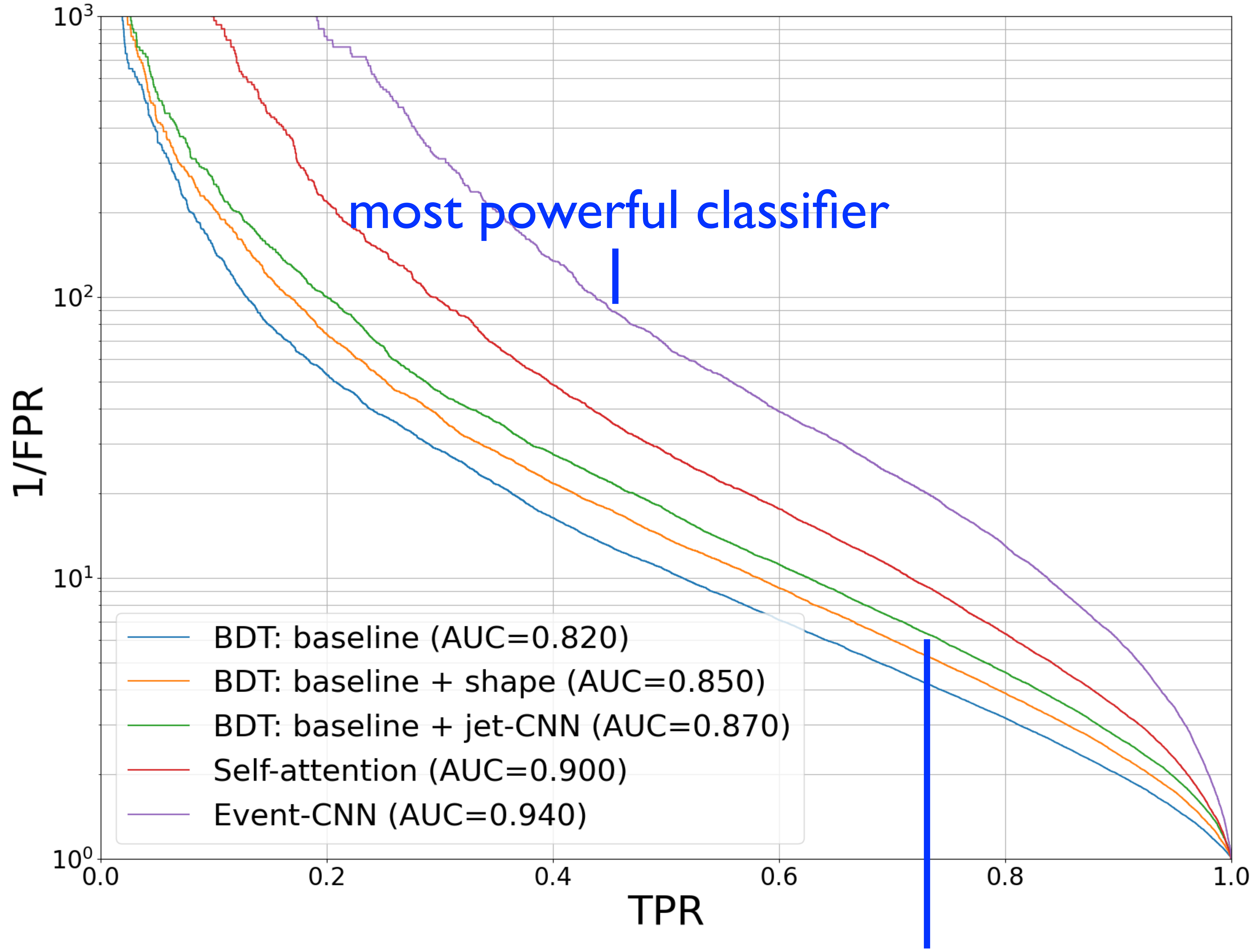
| | training | validation | testing |
|---|---|---|---|
| VBF events | 105k | 26k | 33k |
| GGF events | 83k | 21k | 26k |

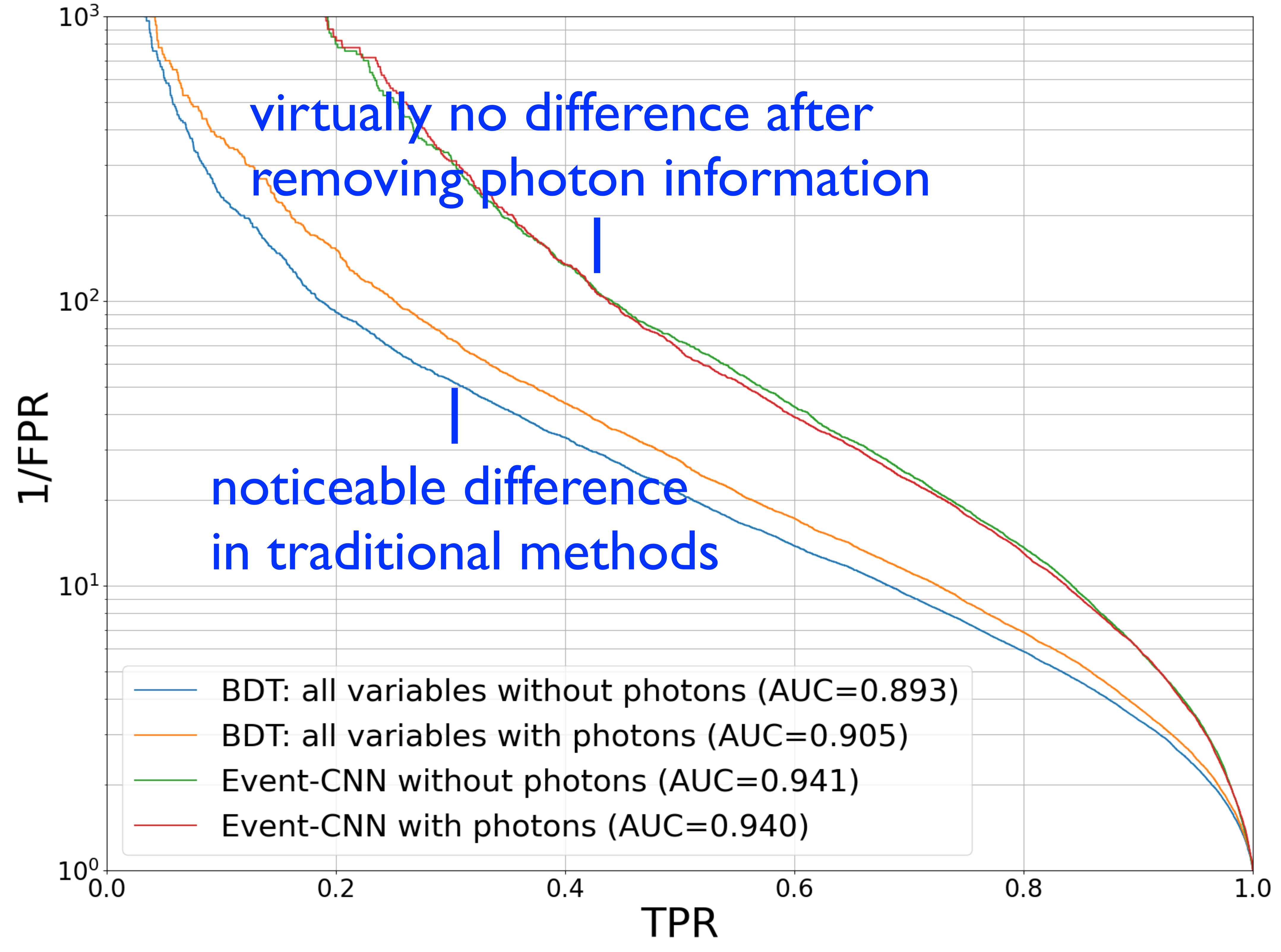
prepared by event simulations using MadGraph, Pythia, Delphes



# Comparison of Classifiers

ROC curves (Receiver Operating Characteristic curves) ROC curves




jet-CNN can supercede the human-engineered jet shape variables

CWC, Shih, Wei 2023

# Collider Simulations

- Particle experimentalists deal with **real data** collected by detectors around colliders.
  ➠ just like analyzing real images for CS people
  ➠ even current multivariate approaches for classification rely on simulations and must be corrected later on using data-driven techniques

https://www.catbreedslist.com/stories/what-breed-of-cat-is-garfield.html

- As particle theorists, we think we are simulating verisimilar data using various packages.
  ➠ in fact, we have been generating **fake data** all along
  ➠ problems: fixed-order in perturbation (e.g., CalcHEP, MadGraph), model-dependent showering/hadronization (e.g., Pythia, Herwig), crude detector simulations (e.g., Delphes)

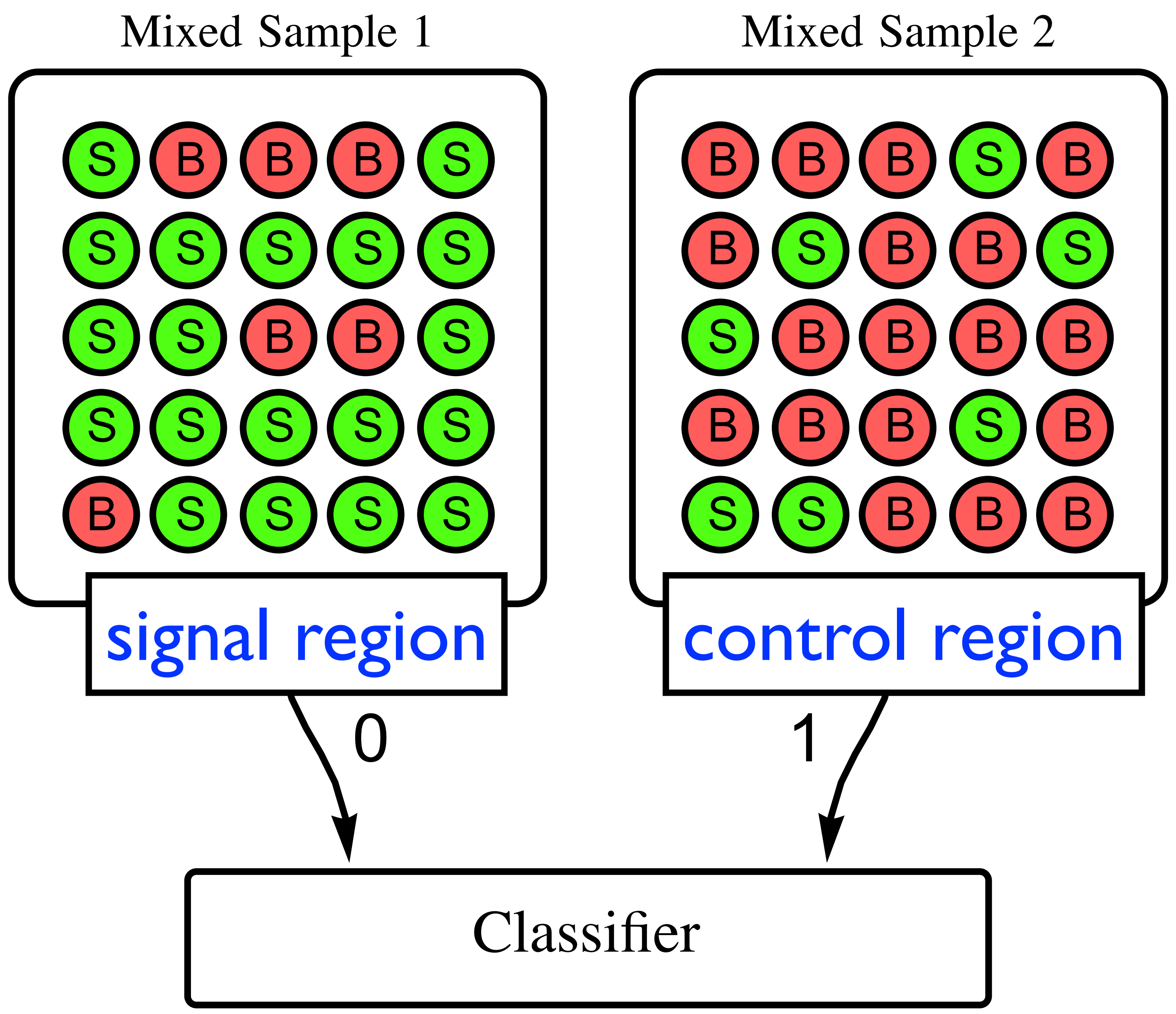
https://en.wikipedia.org/wiki/Garfield_(character)

# Can We Be More Realistic?

- Use a **generative adversarial network** (so-called **GAN**). Louppe, Kagan, Cranmer 2016
  ➠ can alleviate model dependence during training, but at the cost of **algorithmic performance** and **computational resources**

- It would be nice to train directly using **real data.**
  ➠ but real data are **unlabeled**…

# Classification Without Labels (CWoLa)




Metodiev, Nachman, Thaler 2017

- Let $\vec{x}$ represent a list of observables or an image, used to distinguish signal $S$ from background $B$, and define:
  - $p_S(\vec{x})$: probability distribution of $\vec{x}$ for the signal,
  - $p_B(\vec{x})$: probability distribution of $\vec{x}$ for the background.
- Given mixed samples $M_1$ and $M_2$ defined in terms of pure events of $S$ and $B$ (both being **identical** in the two mixed samples) using

$$p_{M_1}(\vec{x}) = f_1 p_S(\vec{x}) + (1 - f_1)\, p_B(\vec{x})$$
$$p_{M_2}(\vec{x}) = f_2 p_S(\vec{x}) + (1 - f_2)\, p_B(\vec{x})$$

with **different** signal fractions $f_1 > f_2$, an **optimal classifier** (most powerful test statistic) trained to distinguish samples in $M_1$ and $M_2$ is also **optimal** for distinguishing $S$ from $B$.

# Remarks

- In the asymptotic limit of **infinite statistics** and **model capacity**, the CWoLa-trained classifier is **statistically equivalent to** the fully supervised classifier.
- Nonetheless, this still requires **a large number** of training samples. What happens if the available data for the mixed samples are **insufficient or limited**, as is often the case with **real data for BSM searches**?

# Processes and Samples

- Two mixed datasets are constructed from MC simulations and denoted as the **SR**, events with **two quark jets**, and the **CR**, events with **at least one gluon jet**.
- Although such information is available in simulations, in a realistic experimental setting, it could be inferred using **auxiliary quark–gluon tagging algorithms.**

Gallicchio and Schwartz 2011

(a) $H \to \gamma\gamma$

| | VBF | GGF |
|---|---|---|
| SR | 10229 | 16828 |
| CR | 1596 | 16865 |

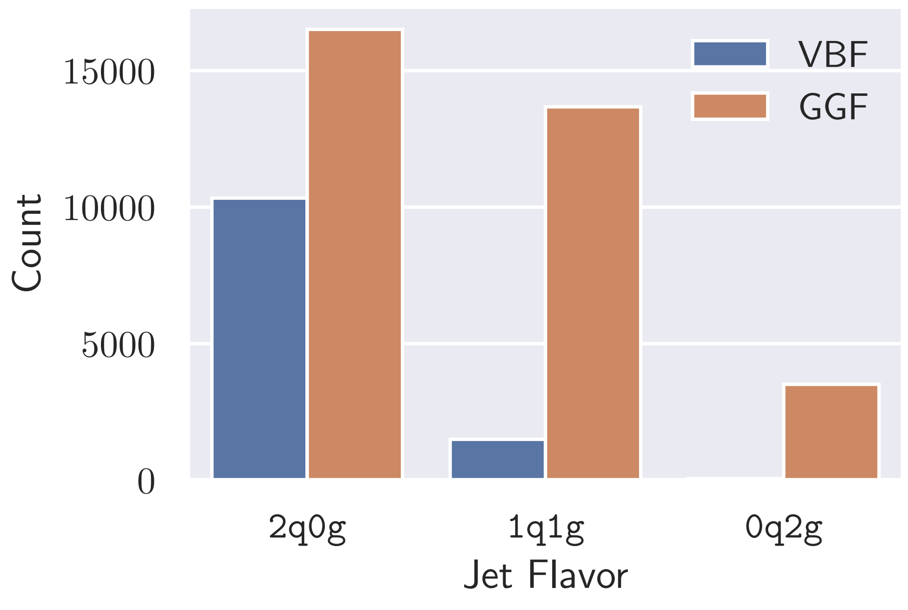
$f_1 > f_2$

(b) $H \to ZZ \to 4\ell$

| | VBF | GGF |
|---|---|---|
| SR | 228 | 722 |
| CR | 34 | 704 |

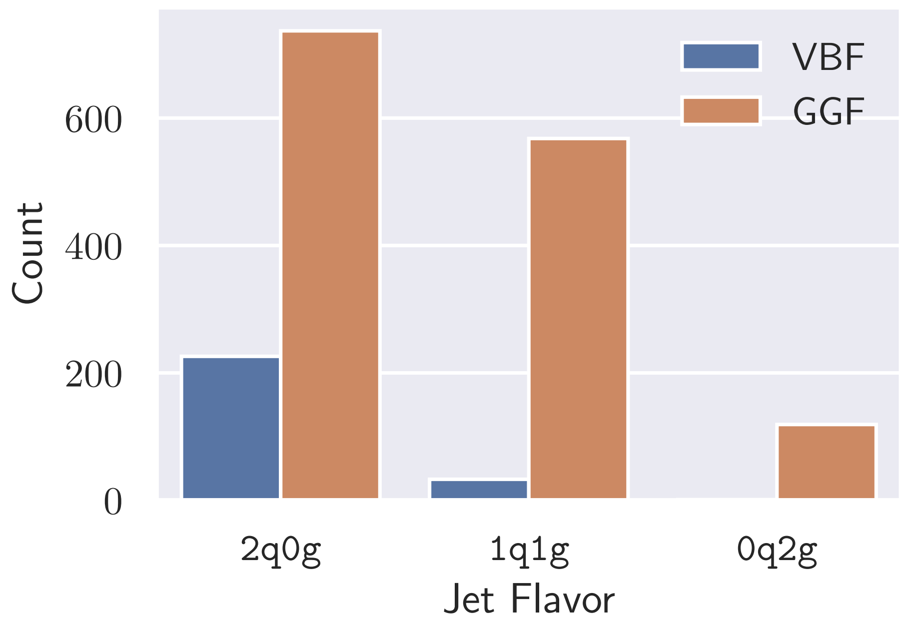
(c) $H \to Z\gamma \to 2\ell\gamma$

| | VBF | GGF |
|---|---|---|
| SR | 275 | 402 |
| CR | 42 | 437 |



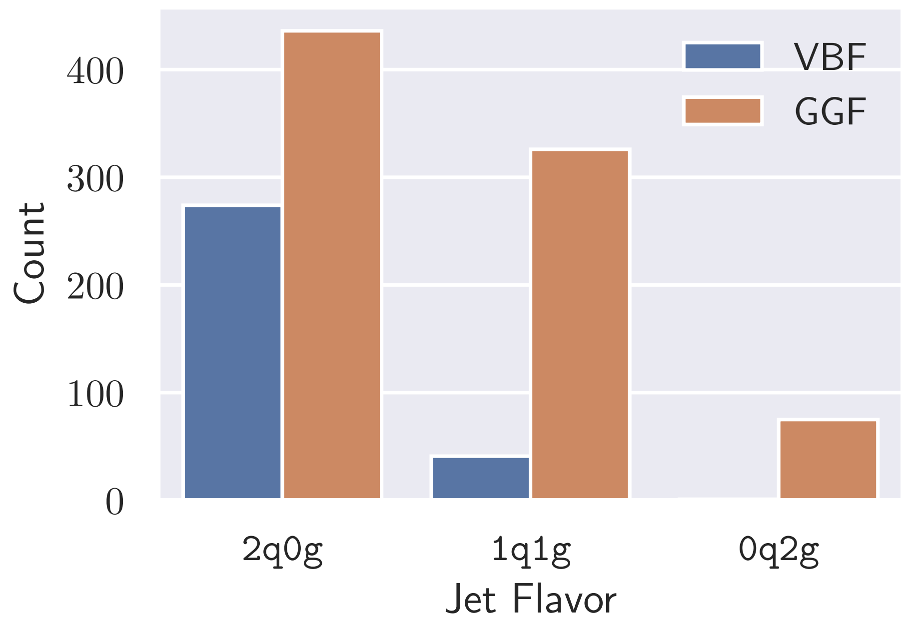

(a) $H \to \gamma\gamma$

$\mathrm{BR}(H \to \gamma\gamma) = 2.27 \times 10^{-3}$




(b) $H \to ZZ \to 4\ell$

$\mathrm{BR}(H \to ZZ \to 4\ell) = 1.24 \times 10^{-4}$




(c) $H \to Z\gamma \to 2\ell\gamma$

$\mathrm{BR}(H \to Z\gamma \to 2\ell\gamma) = 1.03 \times 10^{-4}$

Simulated environment:
- 14-TeV LHC
- $\mathscr{L} = 3000$/fb
- MadGraph 3.3.1 with NNPDF23_nlo_as_0119 PDF
- Pythia 8.306 with NNPDF2.3 LO PDF
- Delphes 3.4.2 with default CMS card
- FastJet 3.3.2 using anti-$k_t$ algorithm
- jet radius $R = 0.4$
- $p_T > 25$ GeV
- $\sigma_{\mathrm{VBF}} = 4.278$ pb, $\sigma_{\mathrm{GGF}} = 54.67$ pb

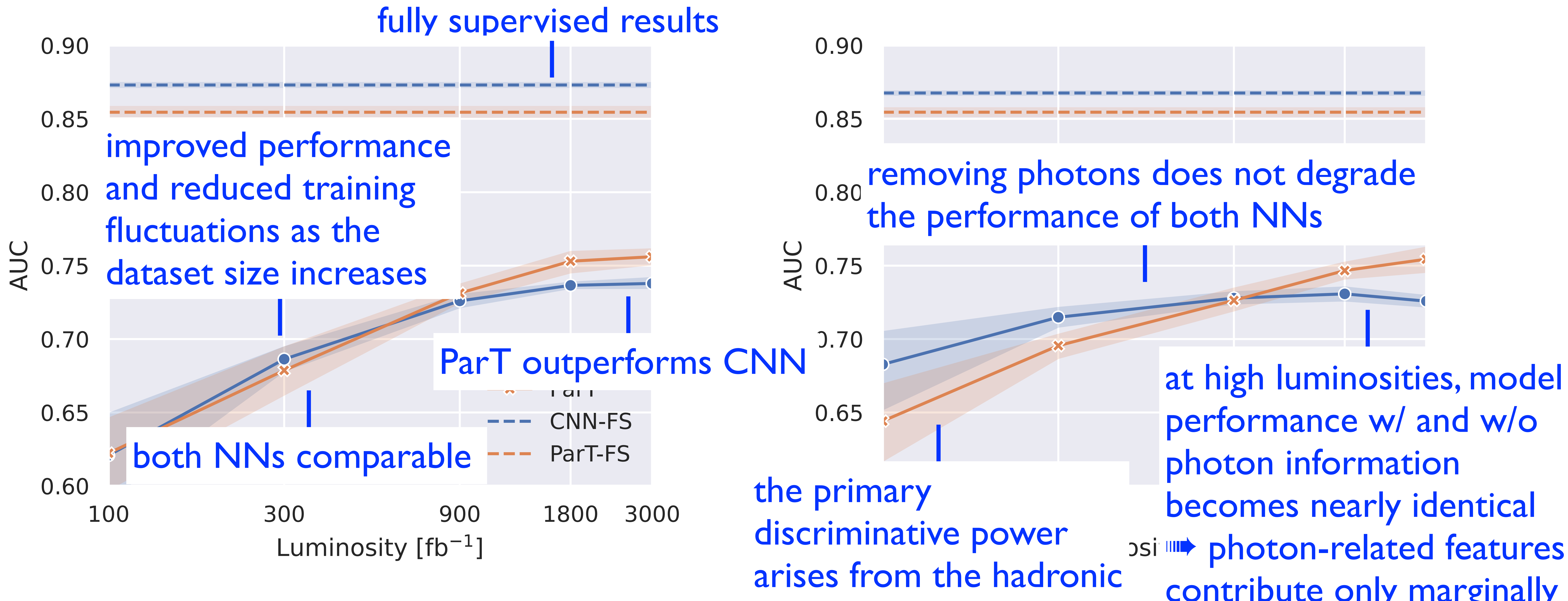
LHCXSWG 2017

# CNN and ParT

- Two types of deep neural network are employed under the CWoLa framework:
  - a **CNN**, designed to capture spatial correlations in hadronic activity, $\sim 2.7 \times 10^5$ trainable parameters, and
  - a **Particle Transformer (ParT)**, designed to process unordered sets of event constituents, $\sim 9.5 \times 10^3$ trainable parameters. Qu, Li and Qian 2022
- Both models are trained on identical datasets and under the same weak supervision conditions, differing only in their **data representation** (image-based or set-based) and **network architectures**.
- To estimate the **upper bound** on classification performance achievable under **ideal labeling**, we construct **fully supervised** datasets where each event carries its truth label (VBF or GGF). For each production mode, 100k, 25k, and 25k events are used for training, validation, and testing, respectively.

# The $H \to \gamma\gamma$ Channel

- Offering a larger event yield and a simpler topology compared to the $H \to ZZ \to 4\ell$ and $H \to Z\gamma \to 2\ell\gamma$ channels, the $H \to \gamma\gamma$ channel serves as the **benchmark** for evaluating the weakly supervised training, enabling robust statistical comparisons between architectures and data representations.




(a) With photon information

# Data Augmentation Methods

- **Data augmentation** is a technique to enlarge the training dataset without acquiring more data through real experiments.

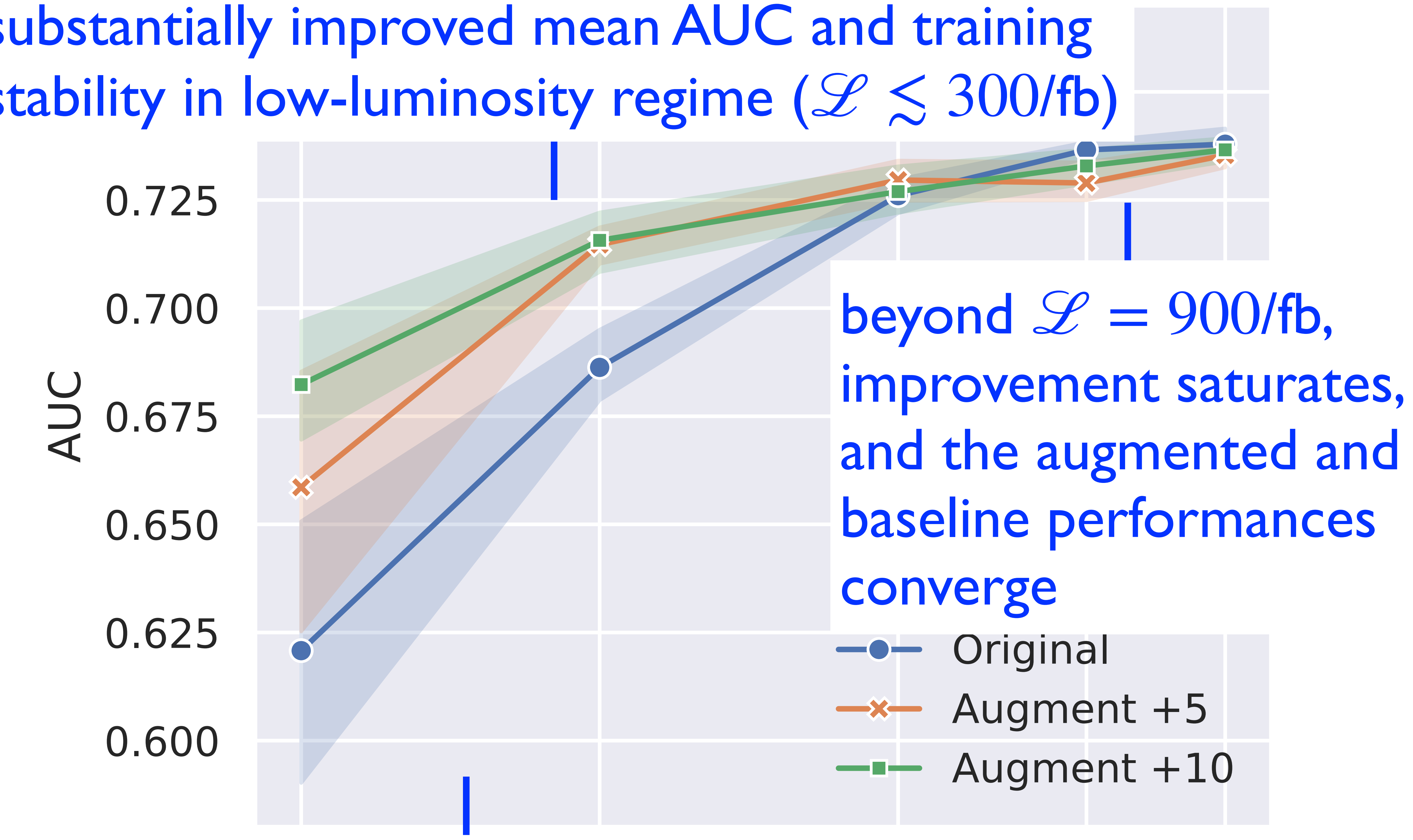
- While there are numerous augmentation methods in the field of computer vision, we focus on **physics-inspired** techniques related to our study.

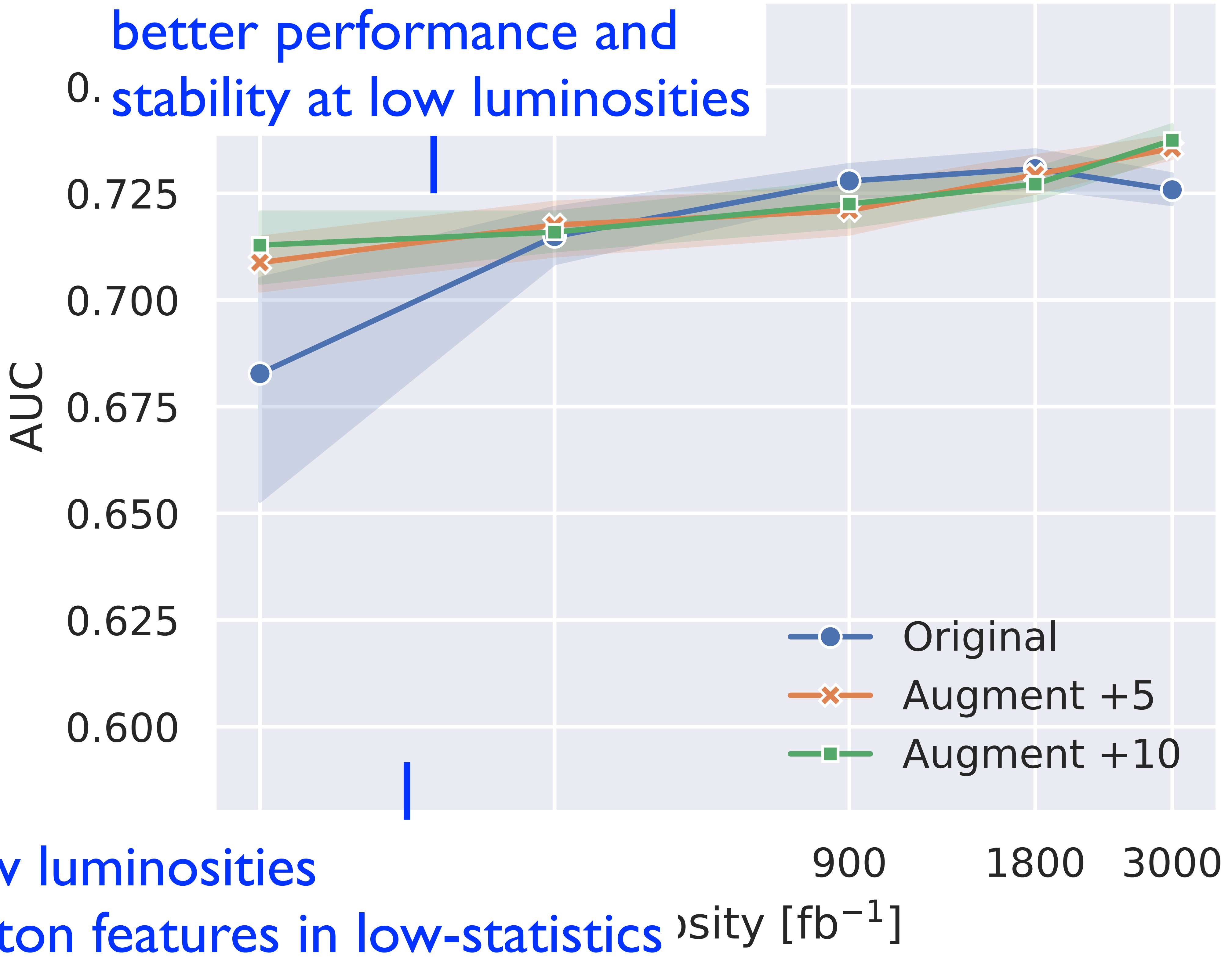
Wang et al 2024
Dillon, Favaro, Feiden, Modak, and Plehn 2024

- Considering **symmetries** in the physical events and experimental **resolution** or statistical **fluctuations** in the detector, common choices include: $p_\mathrm{T}$ smearing; $\phi$ -shifting; jet rotation (not for entire event); $(\eta, \phi)$-smearing; Gaussian noises; etc.
  ➠ only **$\phi$-shifting** found to be effective in our problem

# $\phi$-Shifting Augmentation — CNN

- Due to the cylindrical symmetry of the detectors, rotating all event constituents by a **common azimuthal offset** does not alter the underlying physics of the event.
  ➠ **$\phi$-shifting**, with other kinematic quantities (e.g., $p_T$, $\eta$) remaining unchanged



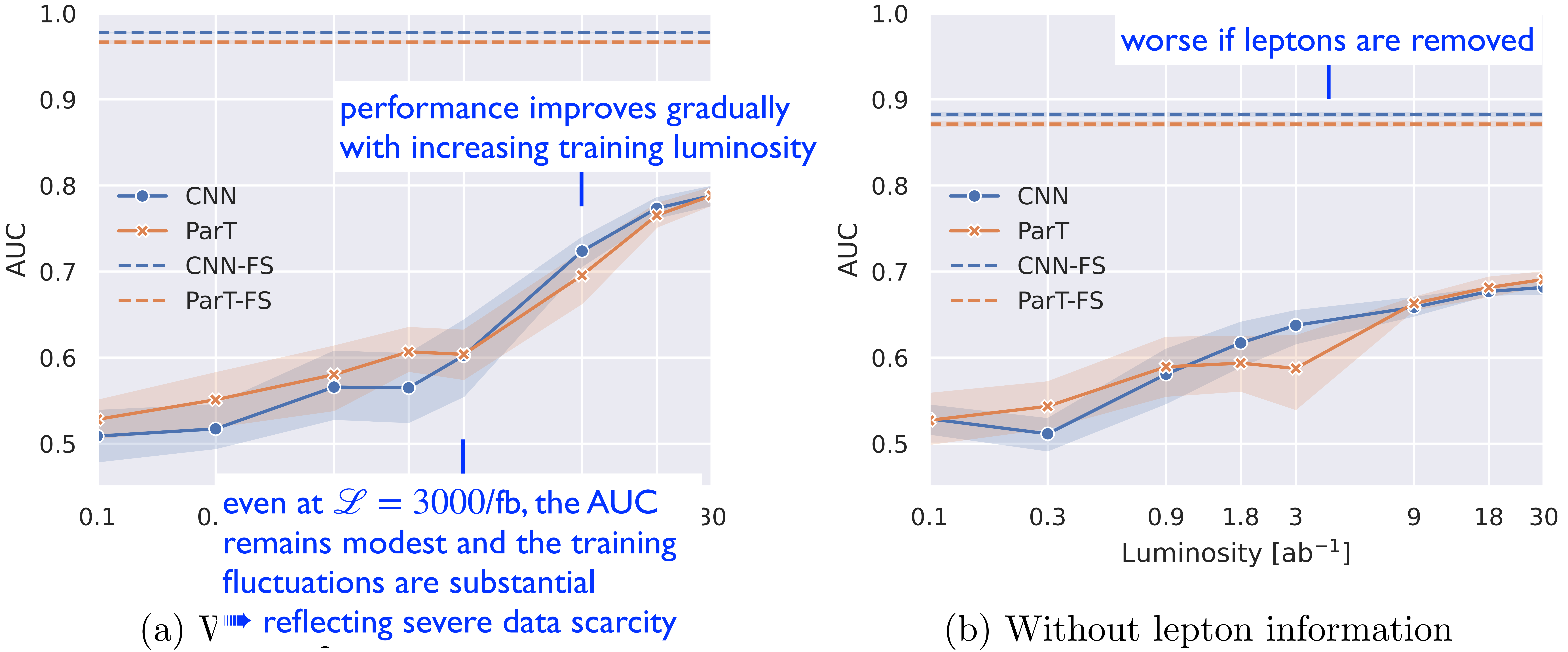

more improvement when photons are included at low luminosities
➠ augmentation helps mitigate overfitting to the photon features in low-statistics

(a) CNN with photon information

(b) CNN without photon information

# Performance on $H \to ZZ \to 4\ell$

- Due to the smaller $BR(H \to 4\ell) = 1.24 \times 10^{-4}$, we had to try larger (unrealistic) luminosities $\mathscr{L} \in \{9000,\ 18000,\ 30000\}$/fb.



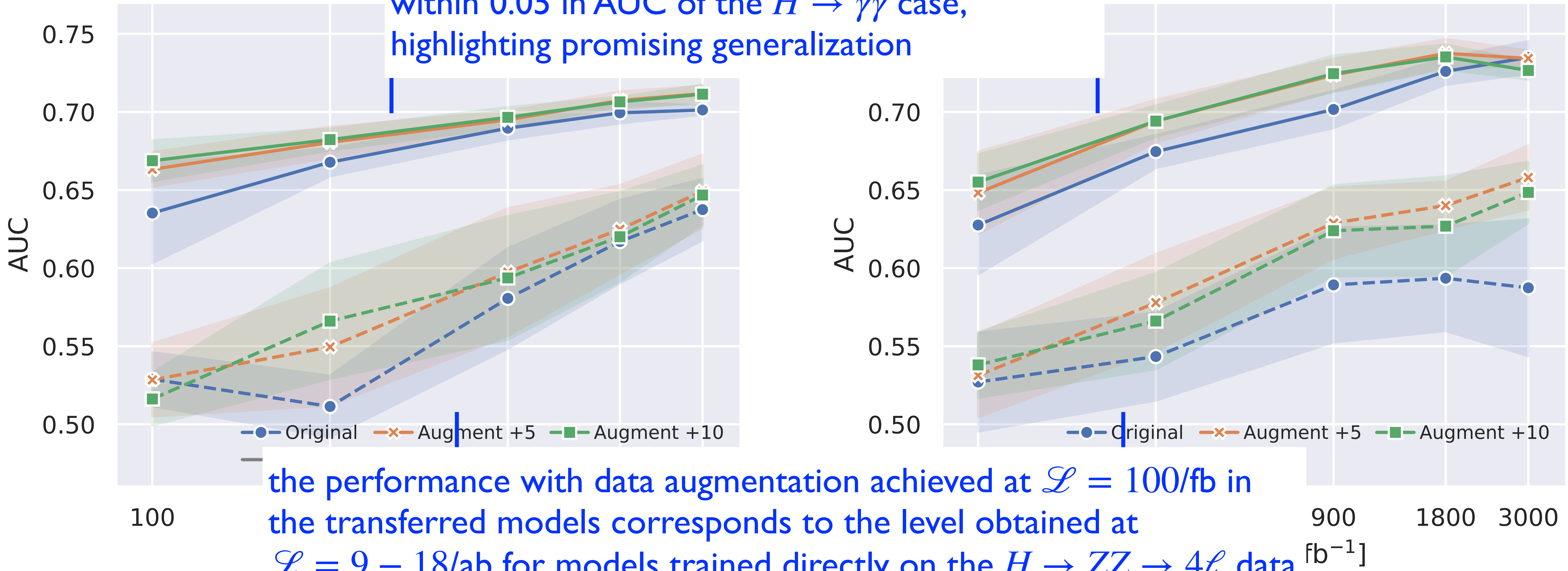

(a) V

(b) Without lepton information

# Transfer Learning to $H \to ZZ \to 4\ell$

- Transfer models trained on the high-statistics $H \to \gamma\gamma$ dataset to the low-statistics ones: $H \to ZZ \to 4\ell$ by removing the decay-product information for all samples.

both CNN and ParT retain strong discriminative power with only modest degradation, typically within 0.05 in AUC of the $H \to \gamma\gamma$ case, highlighting promising generalization



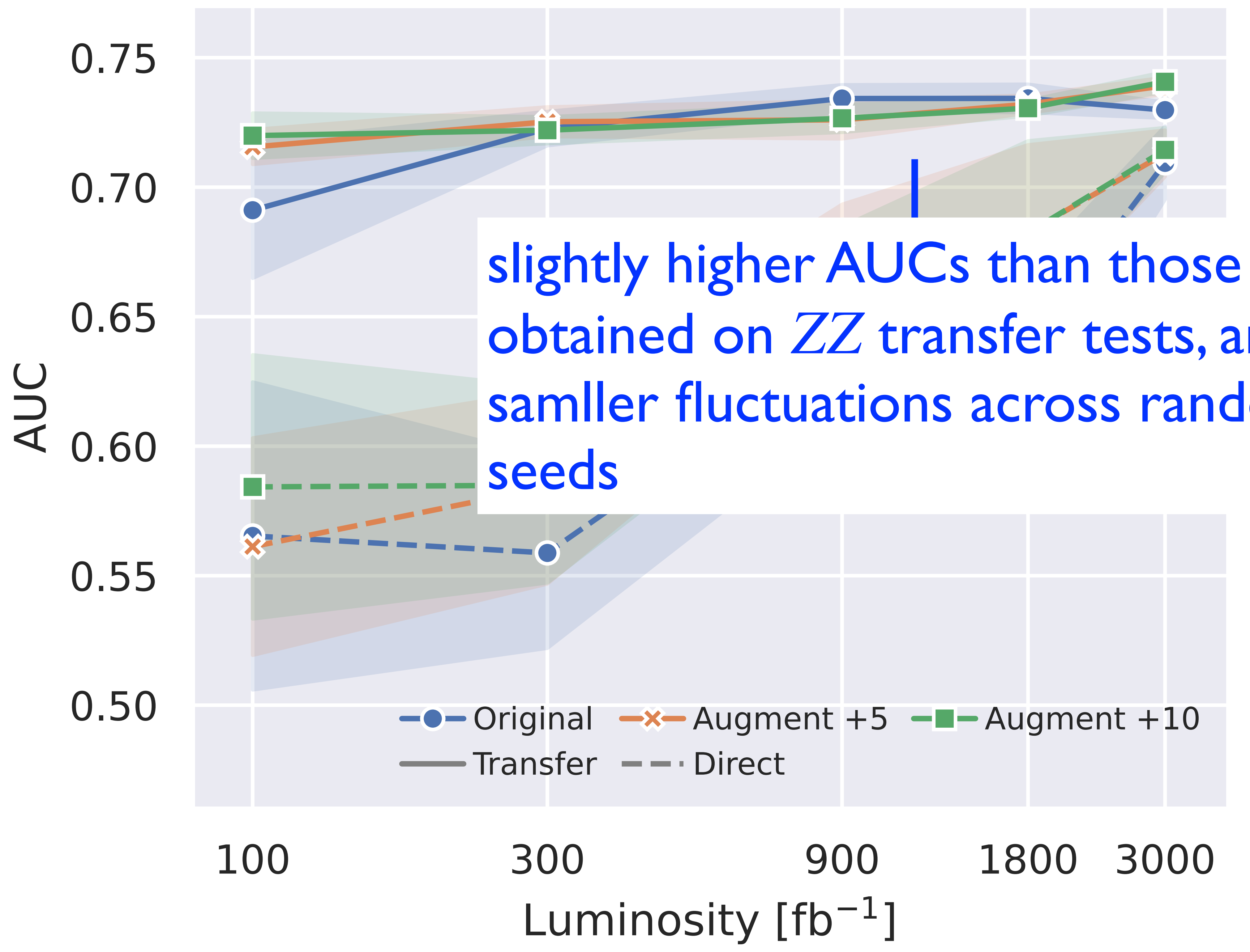

the performance with data augmentation achieved at $\mathscr{L} = 100$/fb in the transferred models corresponds to the level obtained at $\mathscr{L} = 9 - 18$/ab for models trained directly on the $H \to ZZ \to 4\ell$ data

➠ an effective gain of more than two orders of magnitude in statistics

(a) CNN … $\to ZZ \to 4\ell$

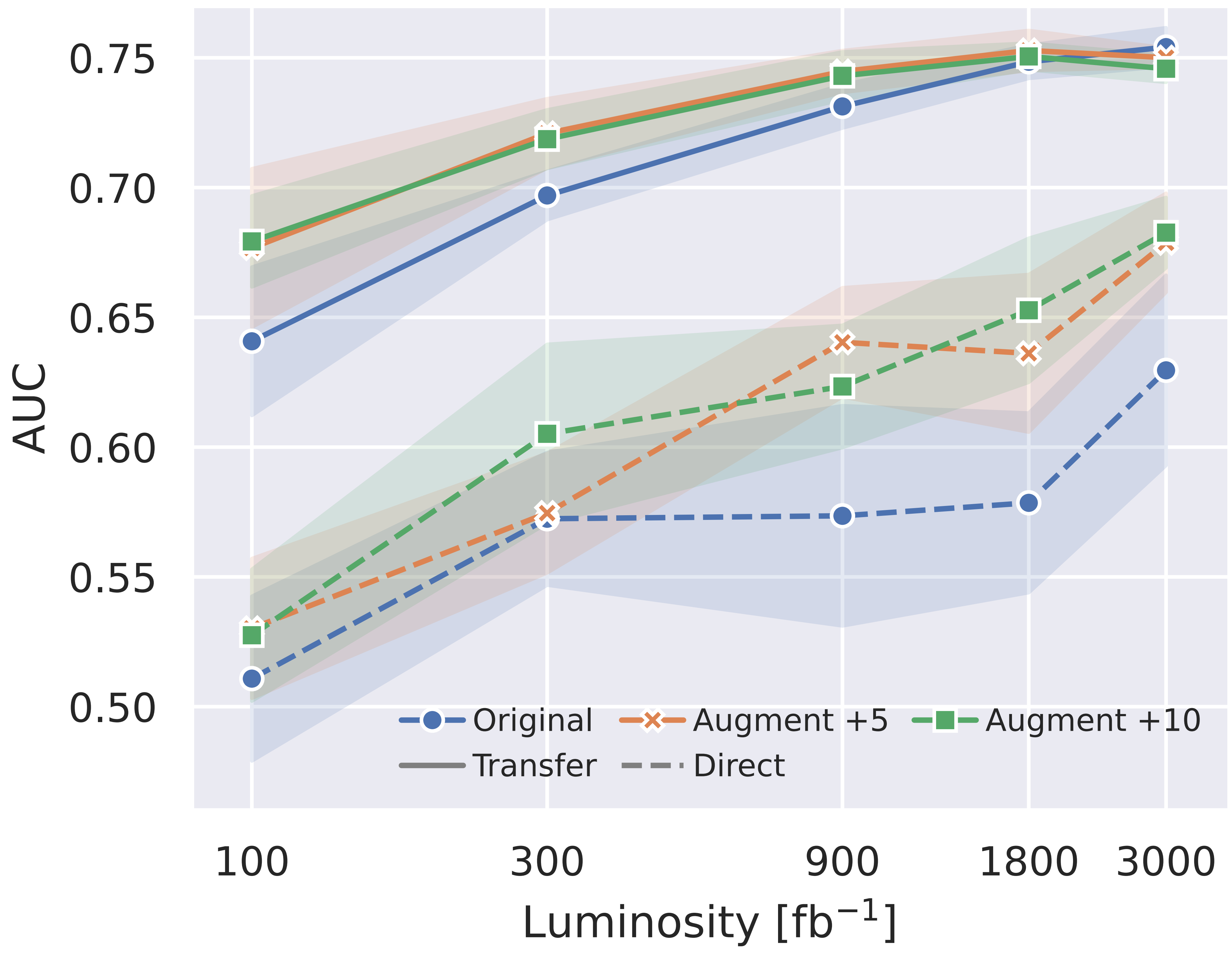
# Transfer Learning to $H \to Z\gamma \to 2\ell\gamma$

- Extending this approach to the $H \to Z\gamma \to 2\ell\gamma$ channel yields **even stronger results**, presumably attributed to the **intermediate event topology** of $Z\gamma$




slightly higher AUCs than those obtained on $ZZ$ transfer tests, and samller fluctuations across random seeds

(c) CNN transferred to $H \to Z\gamma \to 2\ell\gamma$




(d) ParT transferred to $H \to Z\gamma \to 2\ell\gamma$

# Summary

- The **CWoLa** framework, a type of **weak supervision**, is demonstrated to distinguish VBF/GGF Higgs production mechanisms, using **purely experimental data**.
- **CNN** and **ParT** are studied using simulated datasets, allowing us to systematically investigate the effects of **data representation**, **augmentation**, and **transferability** across three Higgs decay channels.
- The $\phi$-shifting augmentation improves performance in the low-data regime, enhancing both the **AUC** and the **training stability**, and **saturates** at higher luminosities.
- The trained NNs primarily rely on global event patterns (particularly **hadronic activities**) rather than specific decay product information, revealing that the models could be **agnostic to Higgs decay modes,** substantially enhancing the **practical utility** of the CWoLa framework in **data-limited scenarios**.

  ➠ train in the $H \to \gamma\gamma$ channel and apply to $H \to ZZ, Z\gamma$

# Thank You!