11th China LHC Physics Conference (CLHCP2025)
@ Landison ICC Hotel Xinxiang, Henan
October 30, 2025

Higgs Production Classifier using Weak Supervision

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

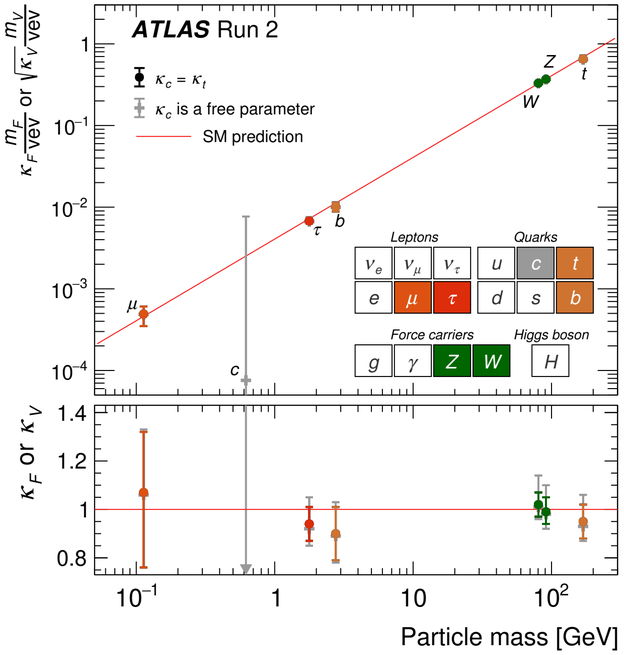
Refs:
- CWC, David Shih, and Shang-Fu Wei, PRD 107, 016014 (2023)
- Hugues Beauchesne, Zong-En Chen, and CWC, JHEP 02 (2024) 138
- Zong-En Chen, CWC, and Feng-Yang Hsieh, 2412.00198
- Kai-Feng Chen, Yi-An Chen, CWC, and Feng-Yang Hsieh, in preparation

Outline

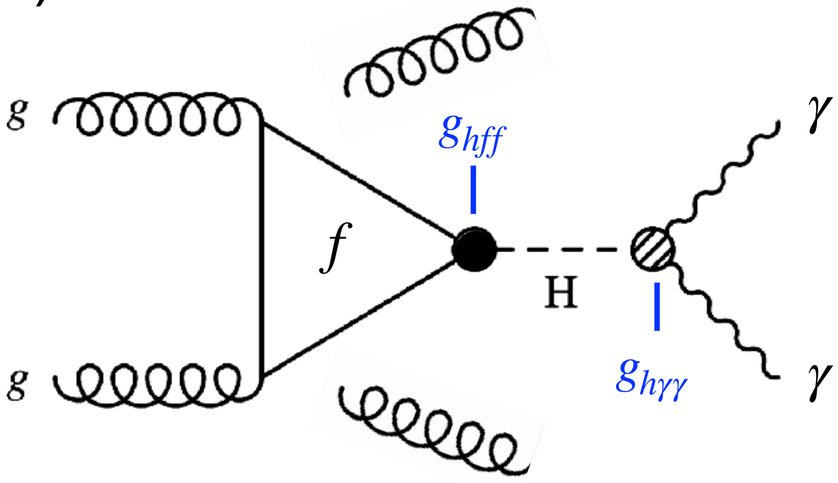
- Introduction — VBF/GGF production
- Classification by full supervision
- Weak supervision — CWoLa
- Classification by weak supervision
  — CNN and Particle Transformer
- Results
- Summary

Higgs Physics Program

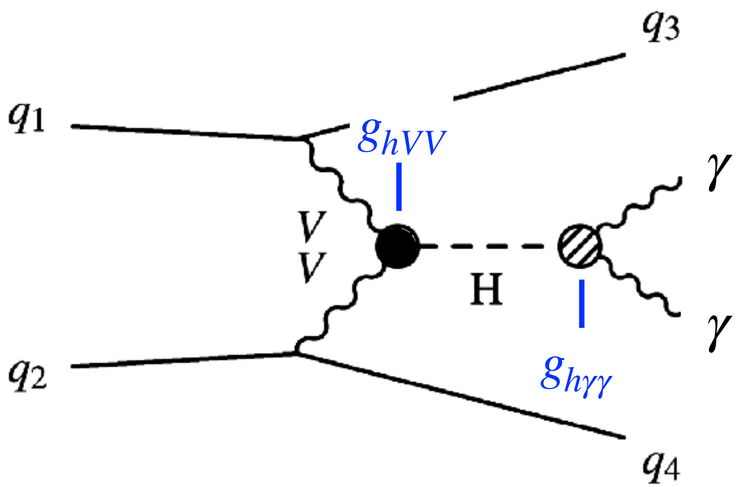
- After the Higgs boson discovery, an urgent physics program is to determine all the **Higgs couplings** precisely.
  - ➠ look for any significant deviations
  - ➠ hints of new physics
- This requires the ability to discriminate the two dominant production channels (others being even smaller).
  - ➠ pinpoint the sources of deviations (production or decay part or both)



ATLAS 2019



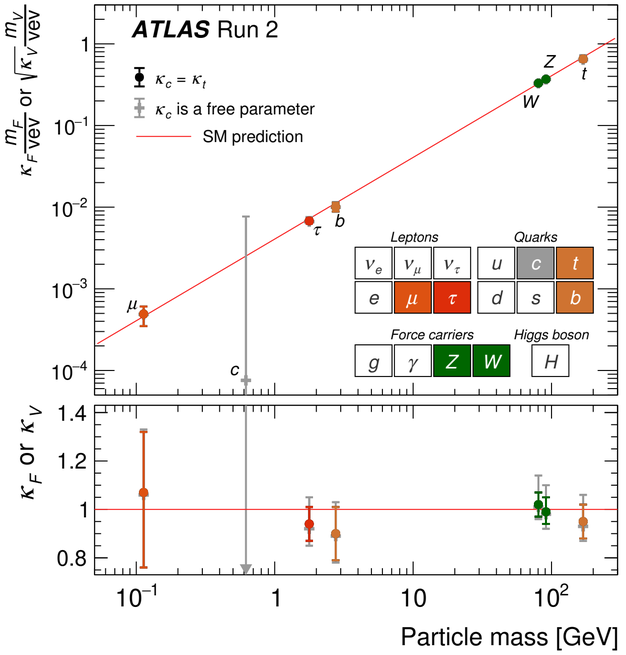
(a) ggF production



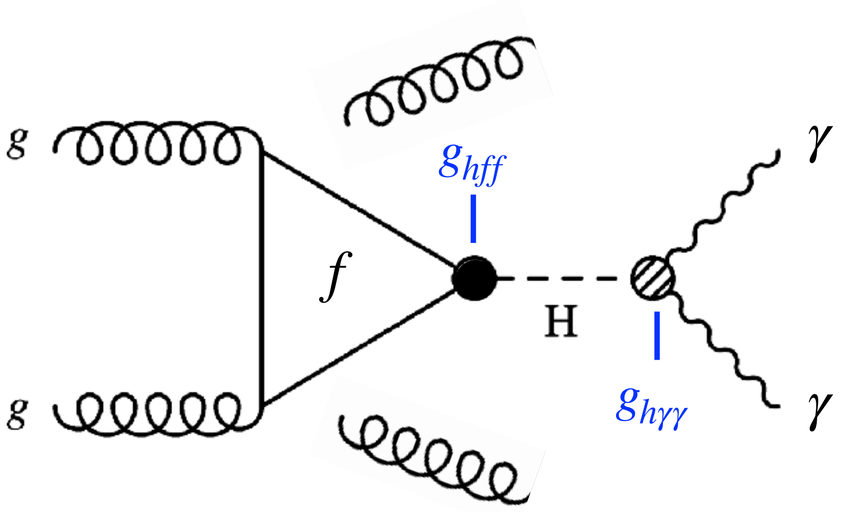
(b) VBF production

VBF vs GGF

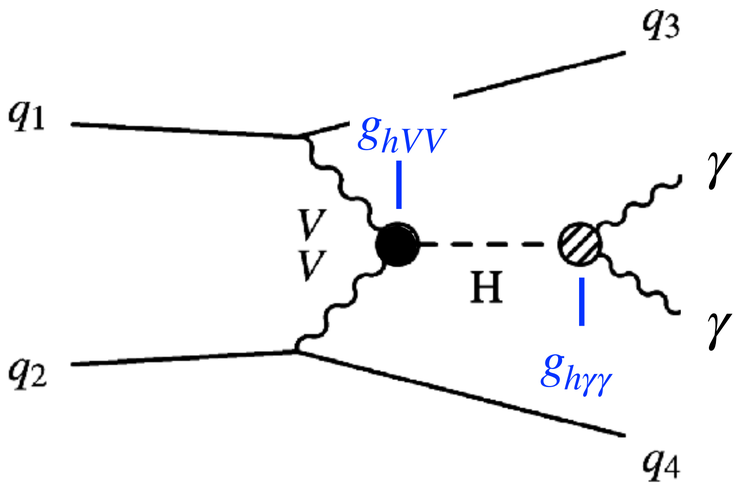
- The **VBF** process or the g_{hVV} coupling is essential for studying the role of the Higgs boson in the EWSB.
- Questions:
  - For any **detected** Higgs event, how can we **efficiently** and **correctly** determine/label its production mechanism?
  - Can it be **independent** of how the Higgs boson decays?



ATLAS 2019



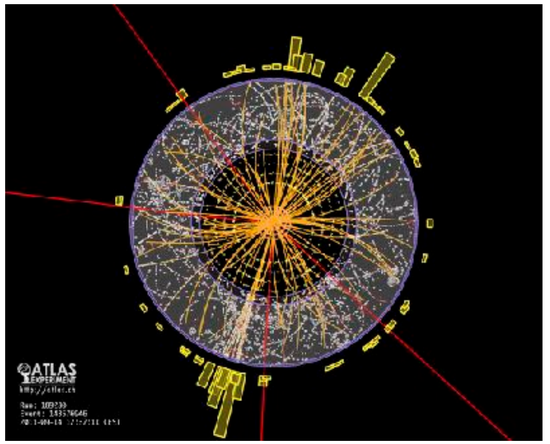
(a) ggF production



(b) VBF production

Our Classifiers

- We construct a **BDT** trained on *high-level features* defined from the leading two jets and the Higgs decay products (the latter to be taken away eventually) as the **baseline** characterizing the prior art.

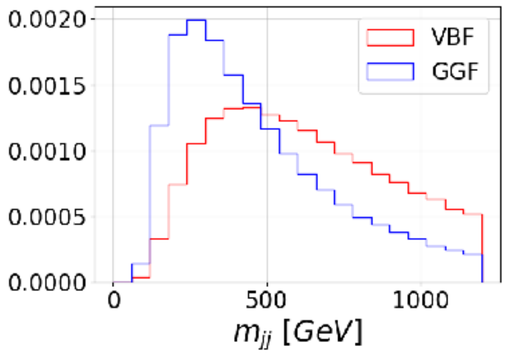


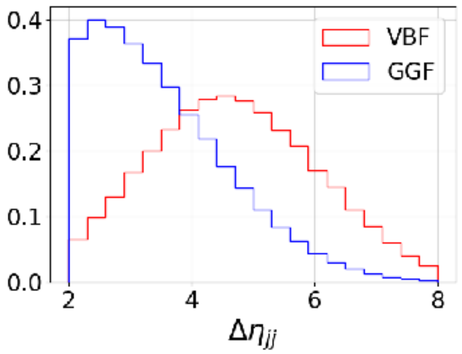
- Beyond it, we consider the following methods:
  - Train a **jet-level CNN** to distinguish the leading two jets (quark vs gluon), and add the jet-CNN scores to the inputs of the BDT for improvement.
  - Train an **event-level CNN** to distinguish full VBF vs GGF events, using full-event images out of the energy deposits of all the reconstructed particles in the event.
  - Train an event-level neural network based on the **self-attention** model, by converting the input event into a sequence that directly records the detector-level information.

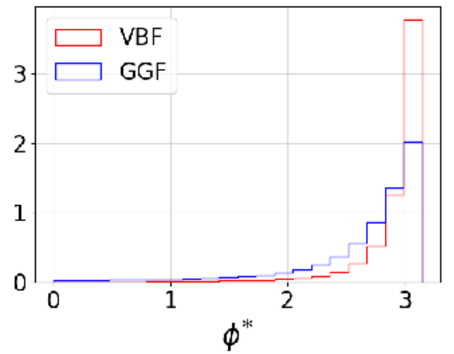
Lin, Feng, dos Santos, Yu, Xiang, Zhou and Bengio 2017
Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin 2017

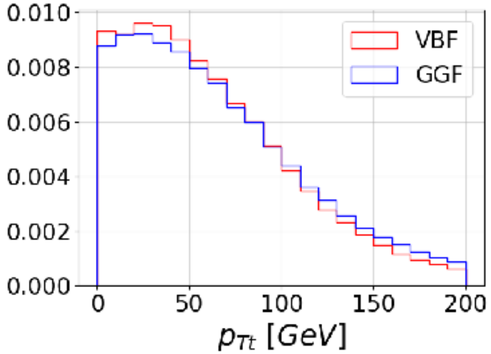
Distributions of BDT Input Variables

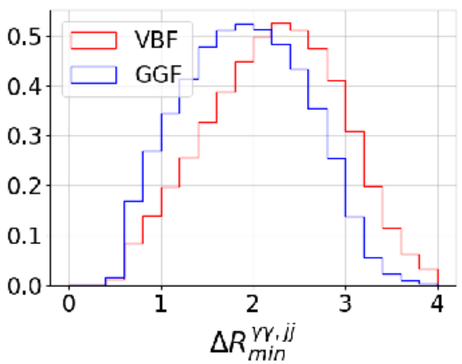
baseline

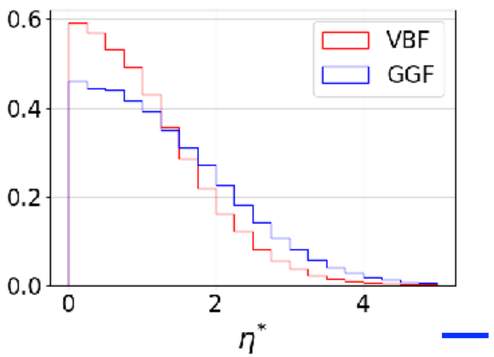




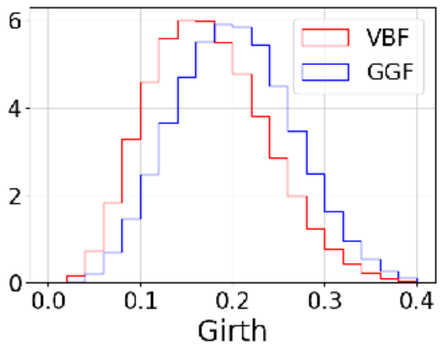


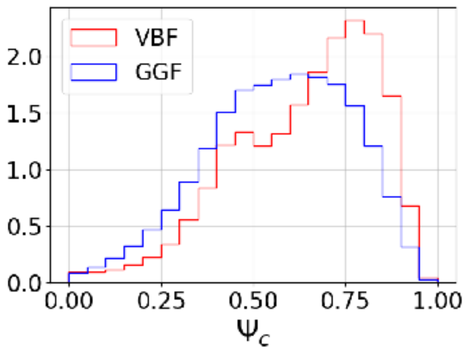


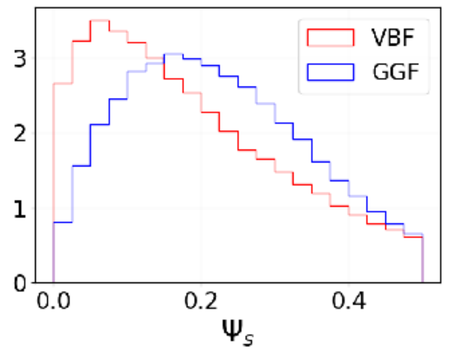




shapes







- Cut-based methods cannot reach high purity.
- BDT-based methods can achieve a purity of about 70% for the VBF sample, depending on the decay channel. ATLAS 2019

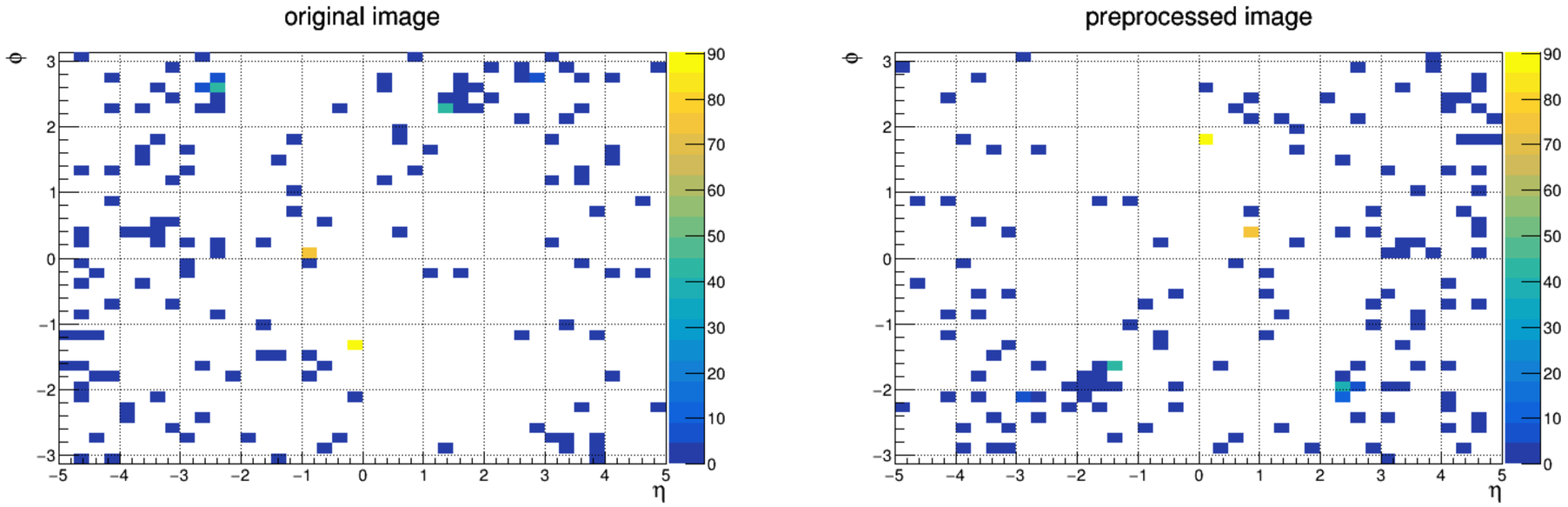
human-engineered kinetic variables

all histograms normalized to have unit area under the curves

Event-CNN

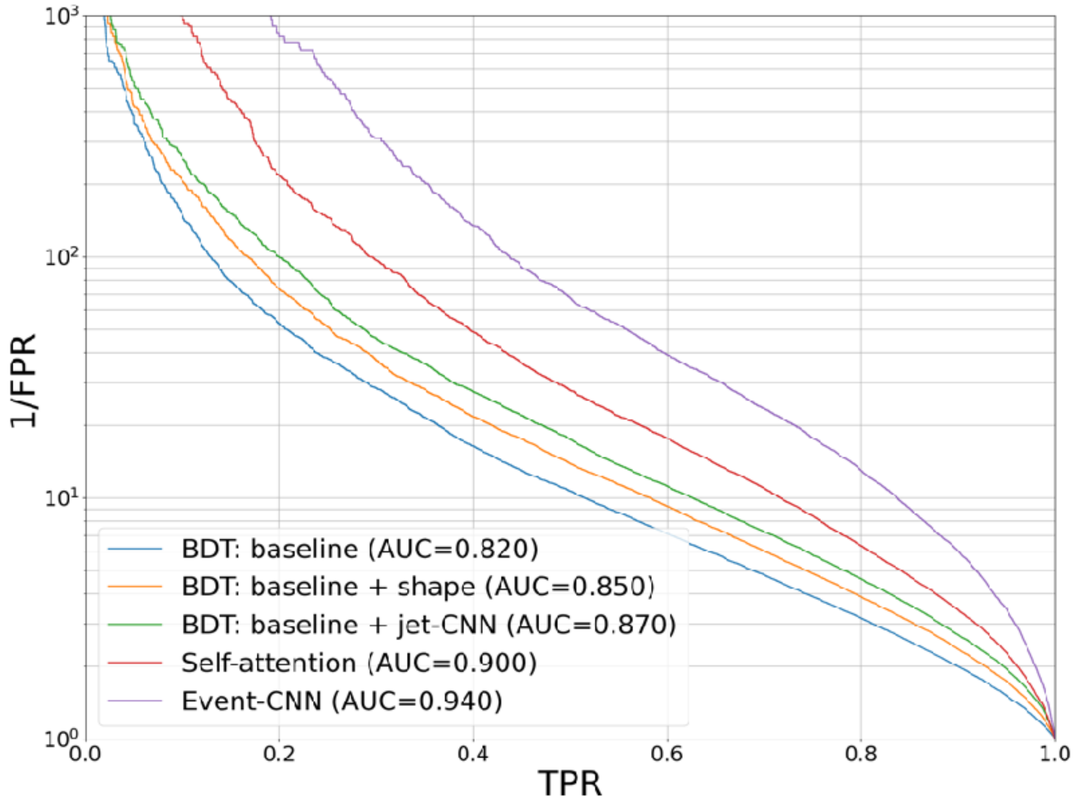
- Train a **convolutional neural network** (**CNN**) by **full supervision** to discriminate between the two production mechanisms by examining the final-state image.
- A successful training typically requires at least **tens of thousands** of samples.

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

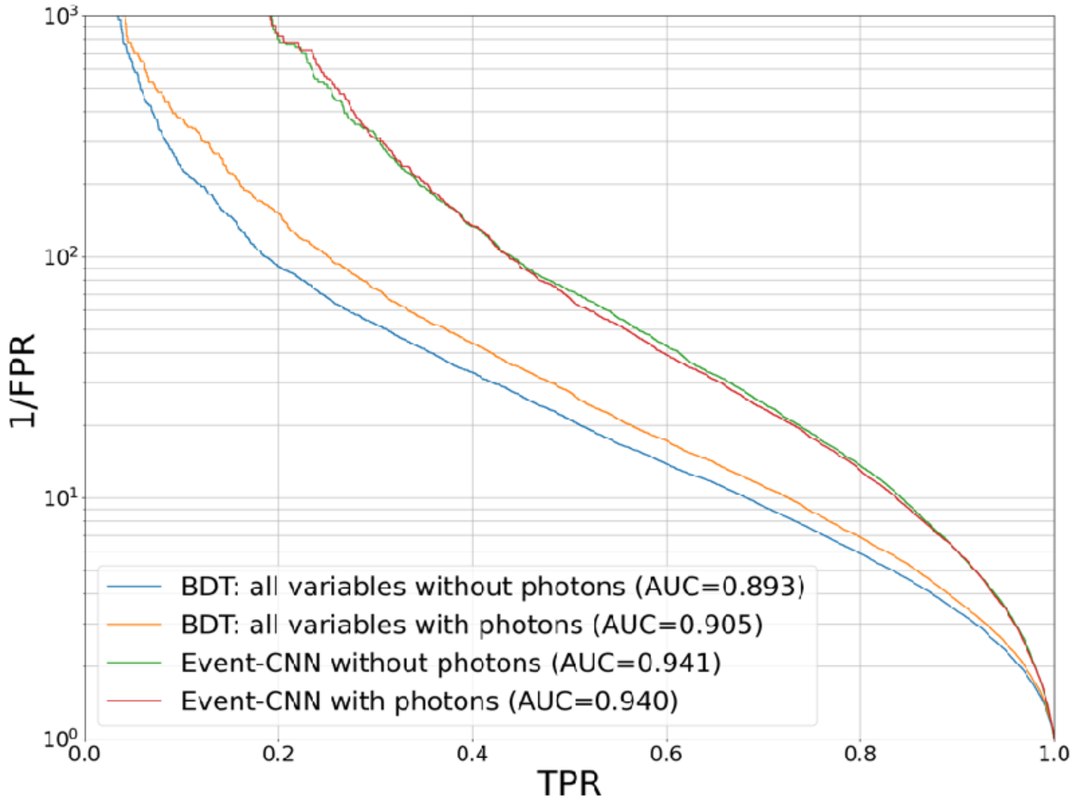


Comparison of Classifiers

ROC curves (Receiver Operating Characteristic curves)



with photon

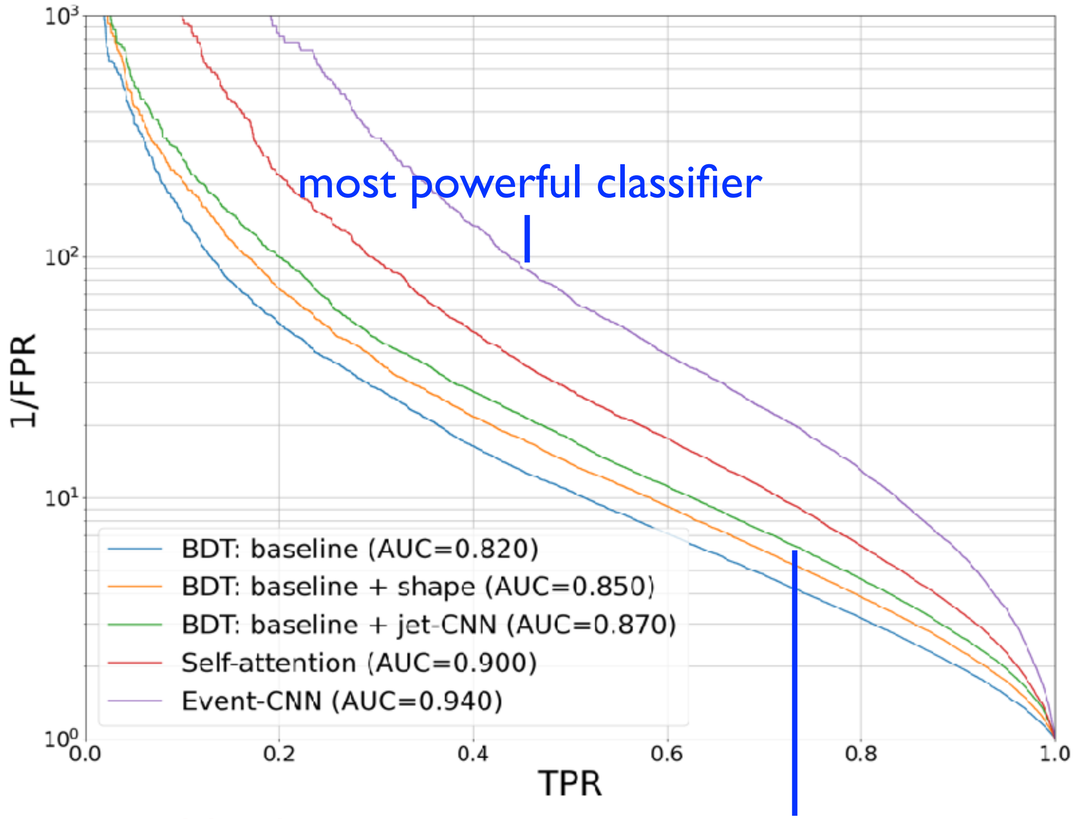


w/ vs w/o photon

CWC, Shih, Wei 2023

Comparison of Classifiers

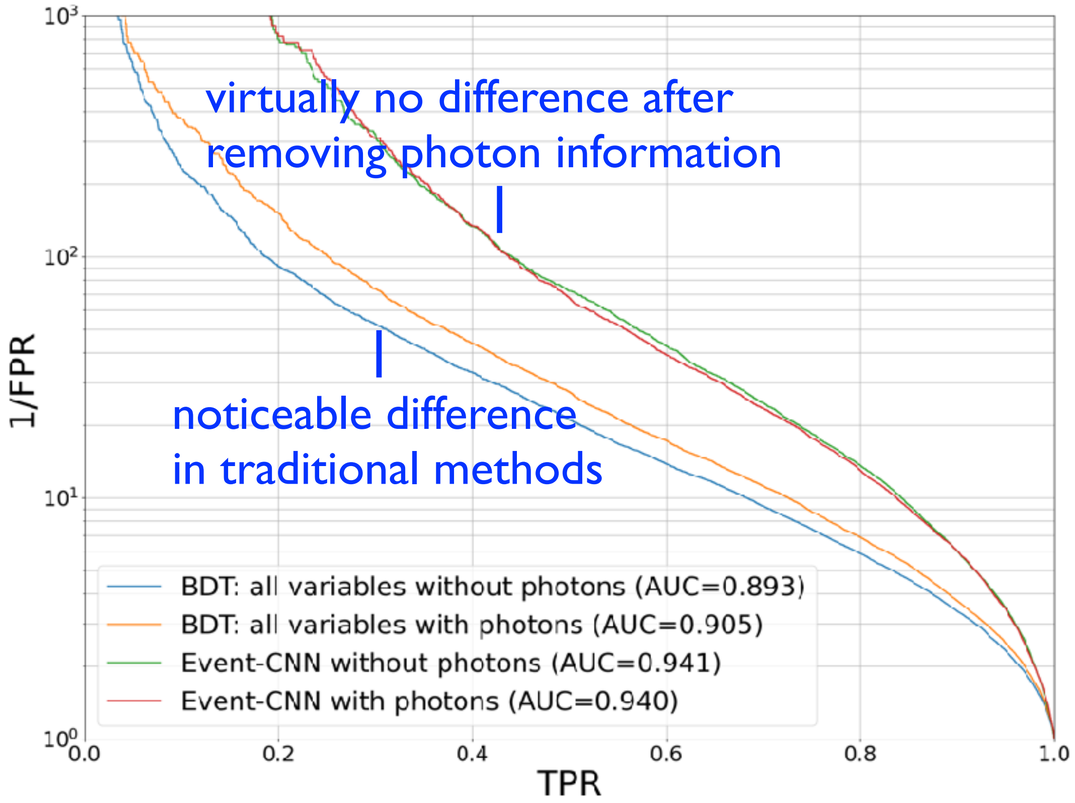
ROC curves (Receiver Operating Characteristic curves)



with photon

jet-CNN has learned the information contained in the human-engineered jet shape variables

ROC curves



w/ vs w/o photon

CWC, Shih, Wei 2023

Collider Simulations

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

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

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

- 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, GEANT)

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

- 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, GEANT)

Can We Be More Realistic?

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*

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**…

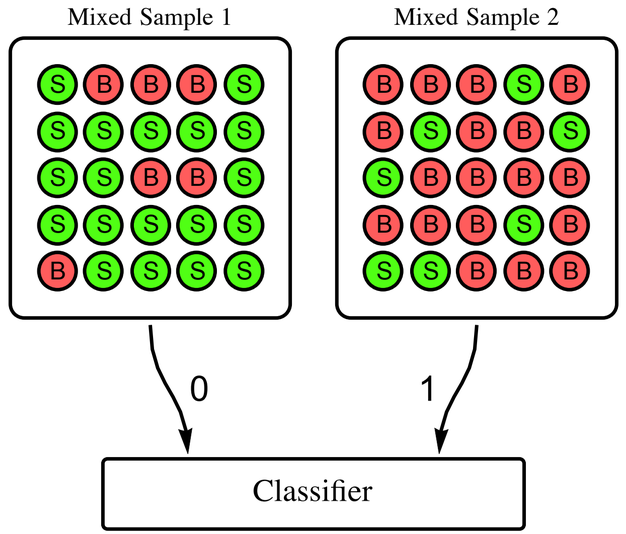
Can We Be More Realistic?

- 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**…

- Introduce **classification without labels** (**CWoLa**). Metodiev, Nachman, Thaler 2017
  ➠ belonging to a broad framework called **weak supervision**, whose goal is to learn from **partially** and/or **imperfectly labeled** data Herna´ndez-Gonz´alez, Inza, Lozano 2016
  ➠ first weak supervision application in particle physics for **quark vs gluon** tagging using *only* **class proportions** during training; shown to match the performance of fully supervised algorithms Dery, Nachman, Rubbo, Schwartzman 2017

A Theorem for 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.

this can be trained with full supervision

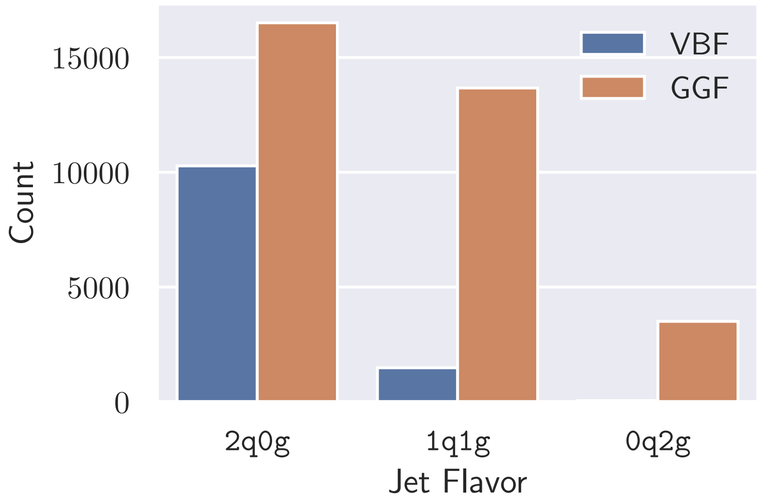
Sample Preparation

- SM Higgs boson events produced via VBF and GGF processes are simulated for a 14-TeV LHC.
  - Parton-level event generation is performed using **MadGraph** 3.3.1 for both production modes, with Higgs decays into $H \to \gamma\gamma$ and $H \to ZZ \to 4\ell$.
  - The parton showering and hadronization are simulated by **Pythia** 8.306.
  - The detector simulation is conducted by **Delphes** 3.4.2.
  - Jet reconstruction is carried out using **FastJet** 3.3.2 with the anti-k_t algorithm and a jet radius parameter of $R = 0.4$.
  - Jets are required to have transverse momentum $p_{\mathrm{T}} > 25$ GeV.

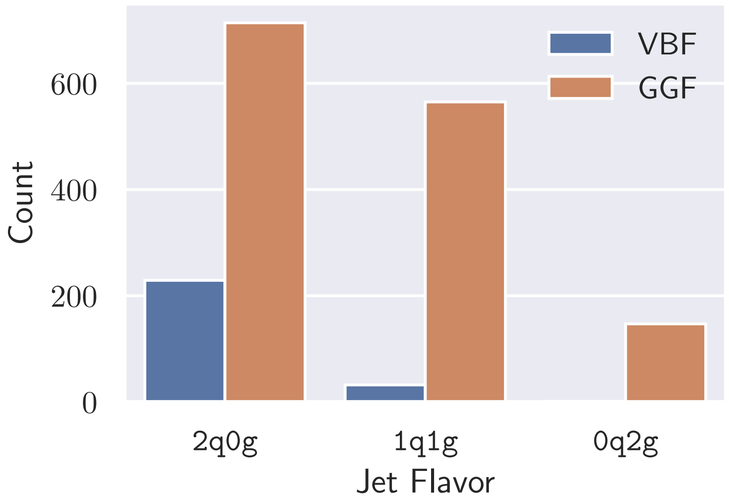
Signal Region and Background Region

- Here, **VBF** events are treated as the **signal** and **GGF** events as the **background**.

$H \to \gamma\gamma$



$H \to ZZ \to 4\ell$

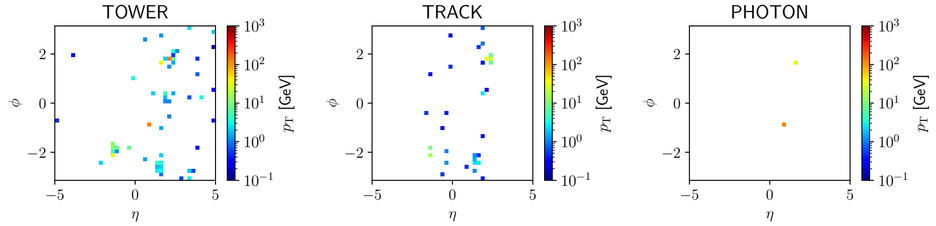


- Distribution of **jet flavor compositions** at $\mathscr{L} = 3000\ \text{fb}^{-1}$.
- **SR** is the 2q0g category, while the **BR** includes both 1q1g and 0q2g events.
- Although jet flavor is available from the Monte Carlo truth, it is assumed that in a realistic experimental setting, such information could be obtained from an **auxiliary jet-flavor tagging algorithm**.

Data Augmentation by ϕ-shifting

- While there are numerous augmentation methods in the field of computer vision, we focus on a **physics-inspired** technique related to our study, based on **azimuthal symmetry**, referred to as **ϕ-shifting**.
- This property allows the generation of additional **statistically independent samples** without modifying the event **kinematics** or **topology**.
- The augmentation is applied **before data representation**, ensuring **consistency** between image-based (CNN) and set-based (Transformer) representations.

CNN and Transformer



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

- Two types of models are considered in this work:
  - For the **image-based** models (**CNN**), the event information is converted into a three-channel (calorimeter towers, tracks, and Higgs decay products) image defined on a 40×40 grid covering $\phi \in [-\pi, \pi]$ and $\eta \in [-5,5]$.
  - For **set-based** architectures (**Particle Transformer**), each event is represented as a collection of reconstructed objects or particles. Each object is described by a six-dimensional feature vector, with the first three components being the **kinematic variables** (p_T, η, ϕ) and the remaining components being the **one-hot encoded type identifiers** (tower, track, or decay product).
    ➠ this representation allows the model to process heterogeneous object types within a unified feature space
- To mitigate the potential **sculpting** effect in CWoLa training, p_T **normalization** is applied to both image- and set-based data.

CNN and Transformer

- Employ **two** NN architectures: **CNN** and **Transformer**, both optimized using the `Adam` optimizer with `binary cross-entropy` as the loss function and a batch size of 512. Early stopping is implemented by monitoring the validation AUC with a patience of 10 epochs.
- The CNN follows the design of **Event-CNN** proposed in [CWC, Shih, and Wei 2023], with a learning rate of 10^{-4}. It consists of several 2D convolutional layers with `ReLU` activations and residual connections, followed by fully connected layers, culminating in a `sigmoid` output layer for binary classification. The total number of trainable parameters in this model is approximately **270K**.
- We utilize the **Particle Transformer (ParT)** introduced in [Qu, Li, and Qian 2024], with a learning rate of 4×10^{-4}. Unlike the original ParT, we omit the interaction matrix, include **one particle attention block** and **one class attention block**, and use a **simplified architecture** with approximately **9.5K** trainable parameters that avoids overfitting and offers stable performance across repeated trials.

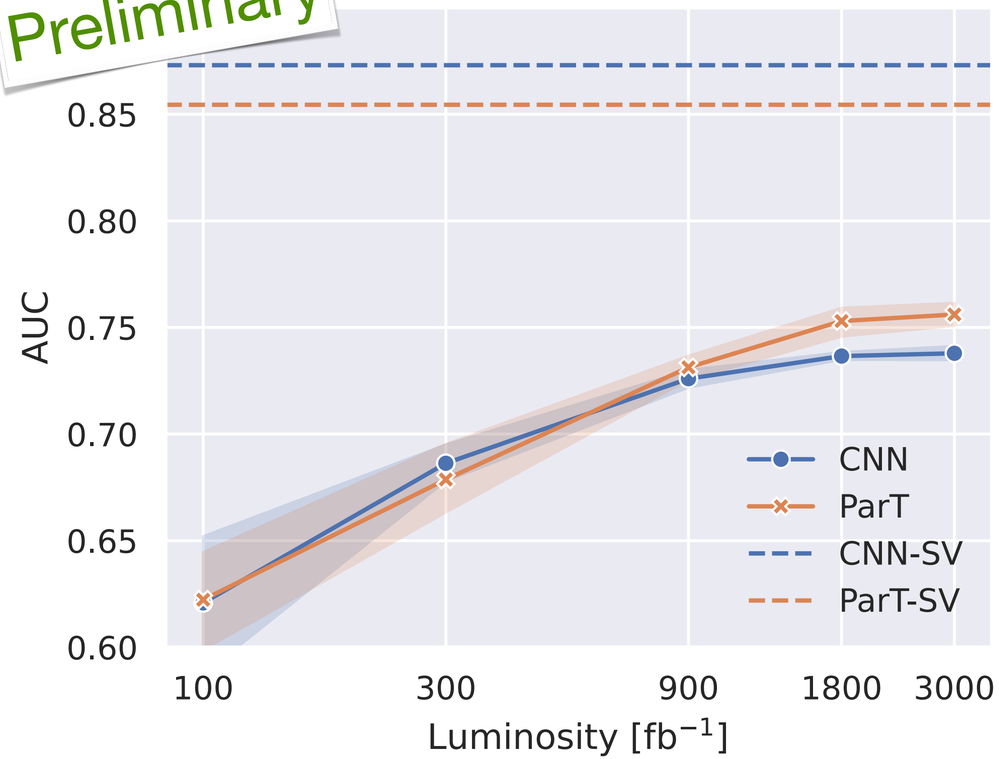
Objectives

- We perform two primary studies:
  - model training on the $H \to \gamma\gamma$ and $H \to ZZ \to 4\ell$ datasets **with and without** the inclusion of decay-product information, and
  - evaluation of the **transferability** of models pretrained on $H \to \gamma\gamma$ to $H \to ZZ \to 4\ell$ events, where decay-product information is removed.
- For each experimental setup, we consider multiple training luminosities: $\mathscr{L} \in \{100, 300, 900, 1800, 3000\}\ \text{fb}^{-1}$.
- In all setups, **ϕ-shifting augmentation** is applied to enhance **rotational invariance** by randomly shifting the ϕ coordinates of all constituents.
- The performance of the NNs is quantified using the **area under the receiver operating characteristic curve (AUC)**.

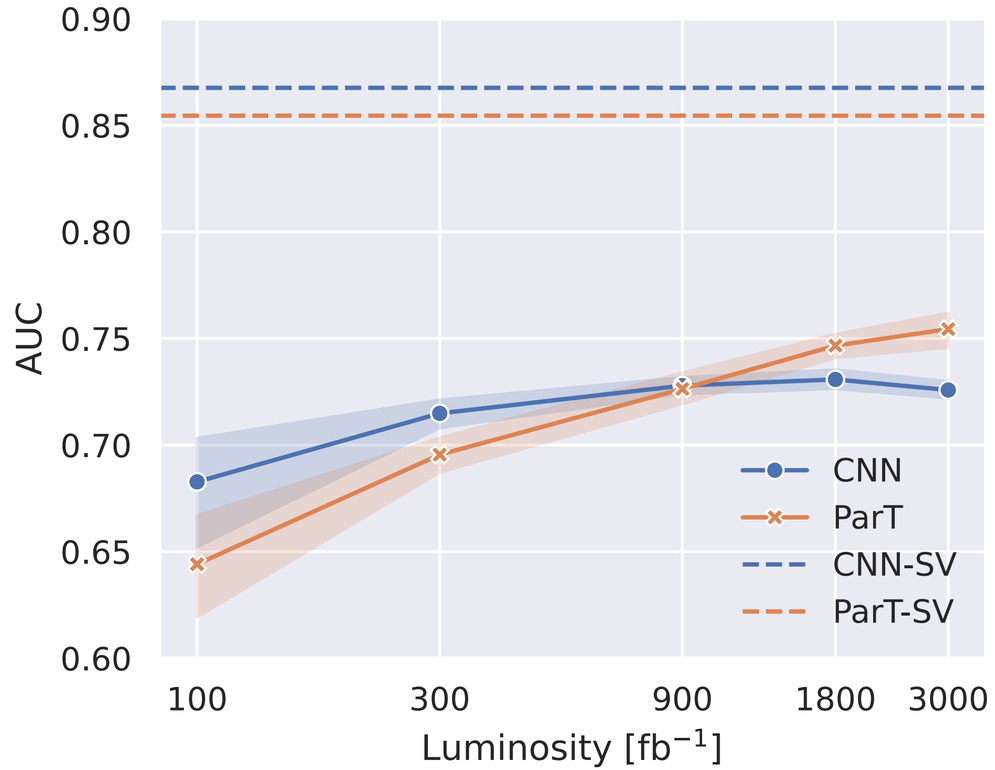
Performance on $H \to \gamma\gamma$

Chen, Chen, CWC, Hsieh, in preparation

Preliminary



with photon



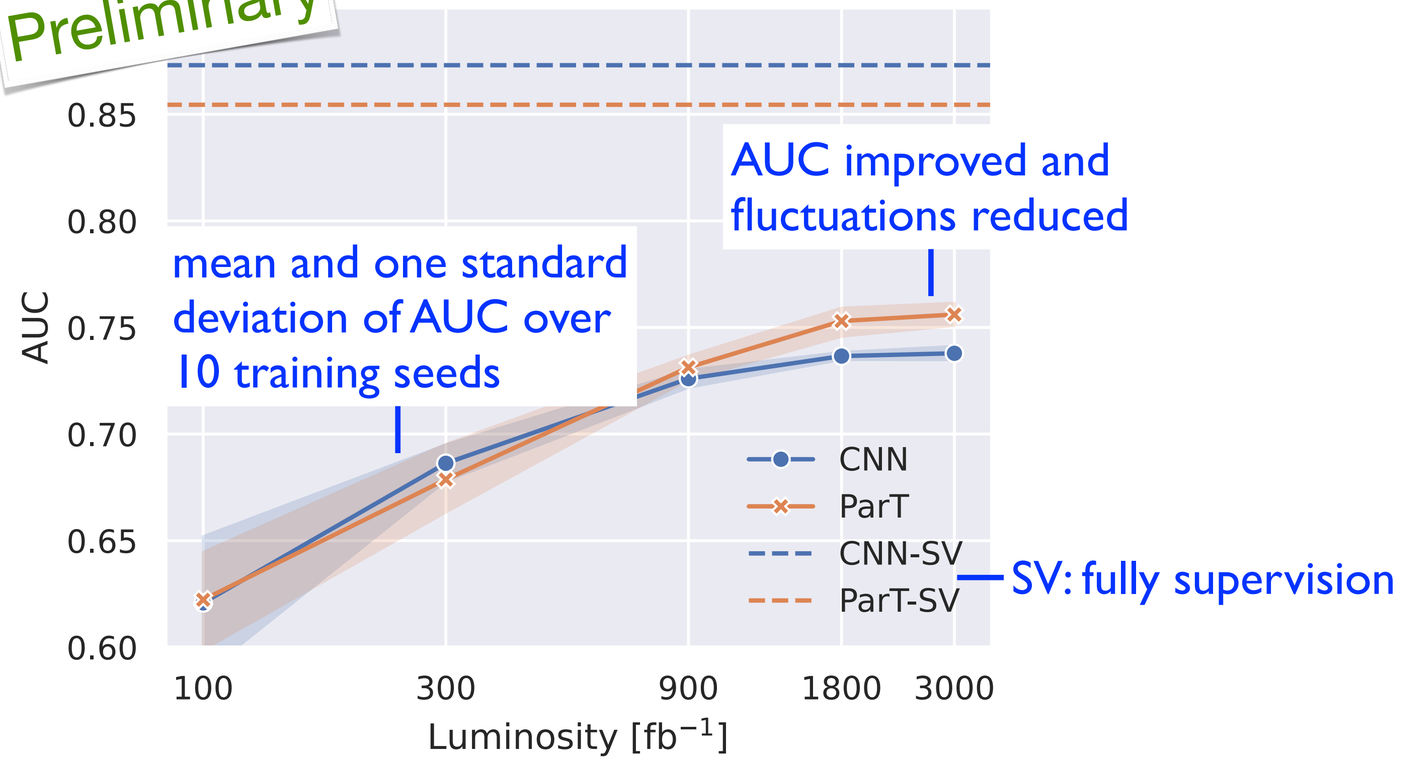
without photon

- CNN and ParT have comparable AUCs for $\mathscr{L} \lesssim 900\ \text{fb}^{-1}$. ParT outperforms for higher luminosities, demonstrating its stronger capacity to exploit complex correlations when more events are available for training.

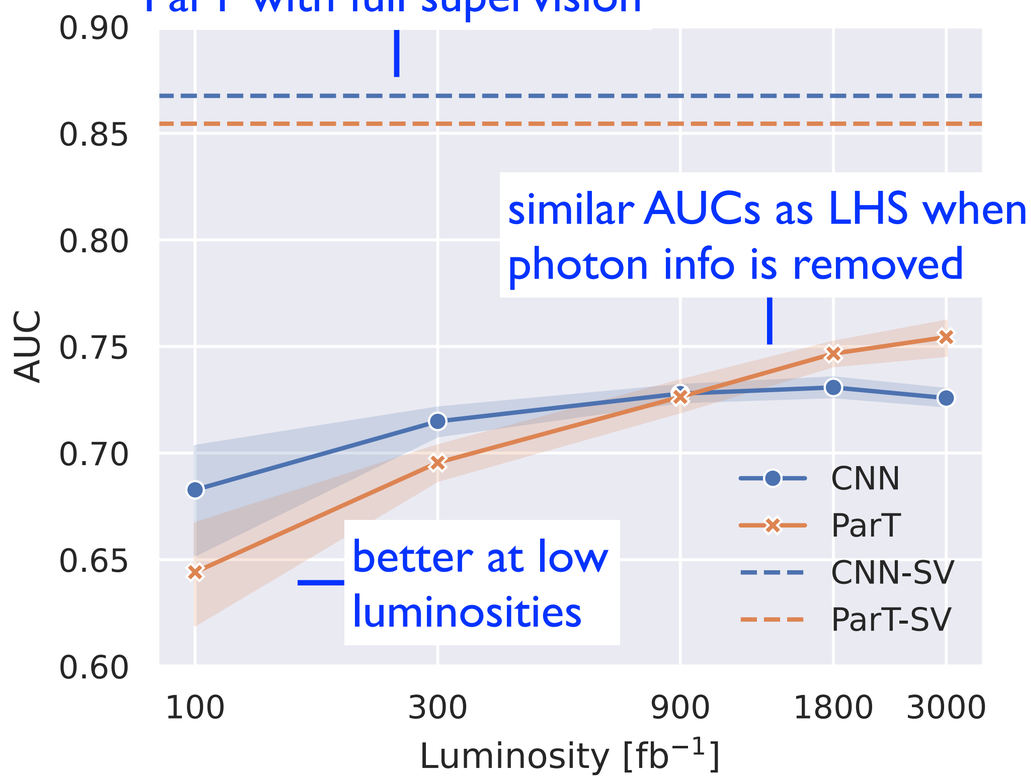
Performance on $H \to \gamma\gamma$

Chen, Chen, CWC, Hsieh, in preparation

Preliminary



with photon



without photon

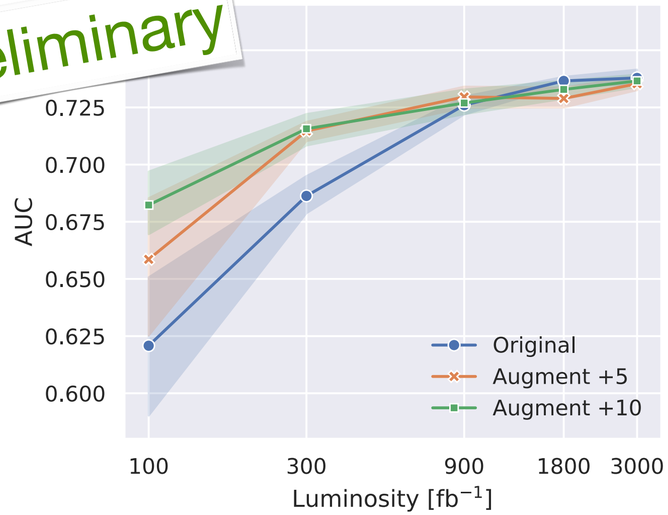
- CNN and ParT have comparable AUCs for $\mathscr{L} \lesssim 900\ \text{fb}^{-1}$. ParT outperforms for higher luminosities, demonstrating its stronger capacity to exploit complex correlations when more events are available for training.

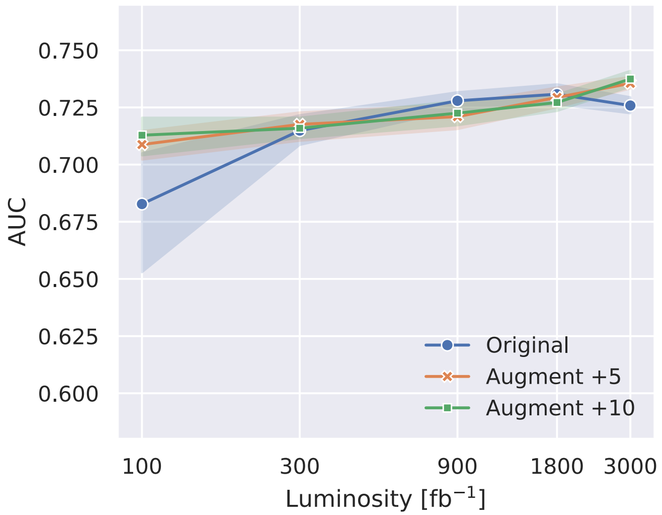
With Data Augmentation

Chen, Chen, CWC, Hsieh, in preparation

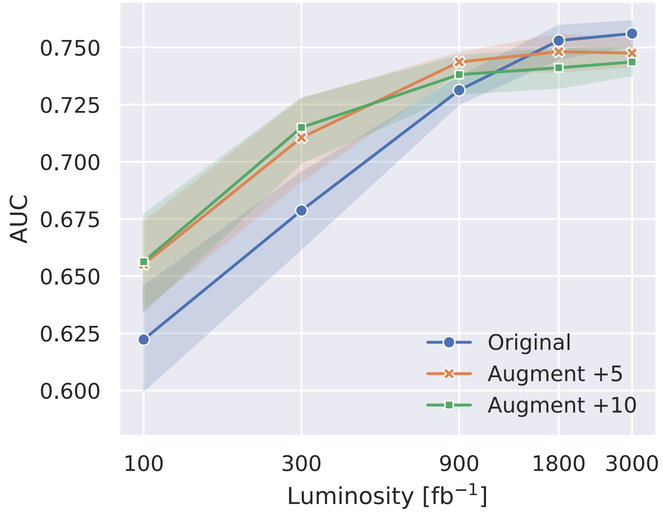
Preliminary

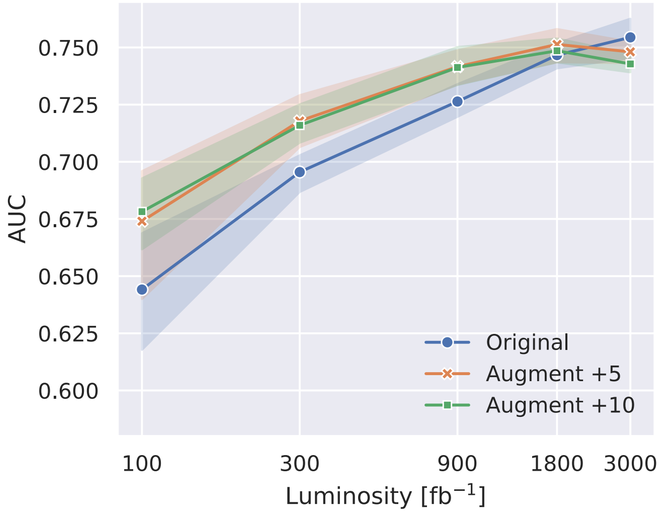
CNN





ParT





with photon

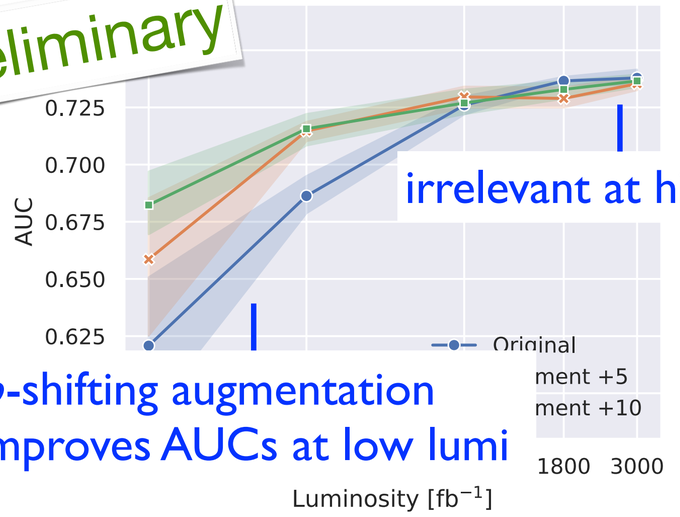
without photon

With Data Augmentation

Chen, Chen, CWC, Hsieh, in preparation

Preliminary

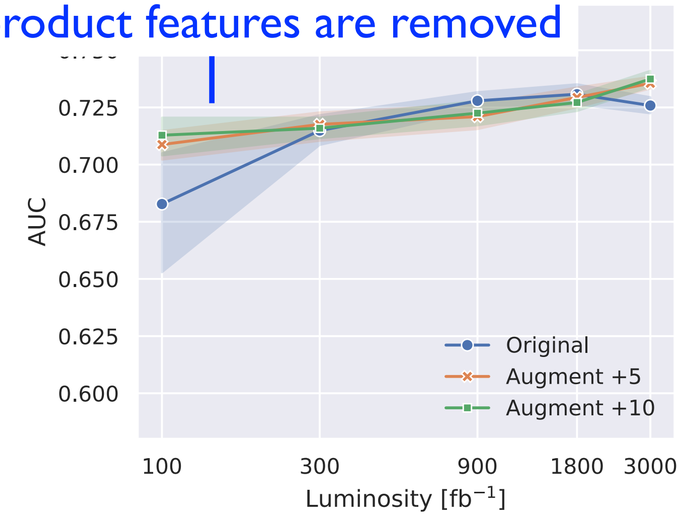
CNN



irrelevant at high lumi

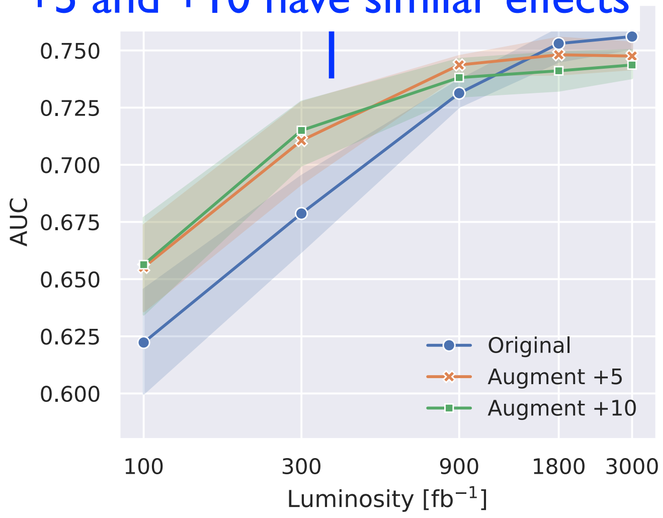
ϕ-shifting augmentation improves AUCs at low lumi

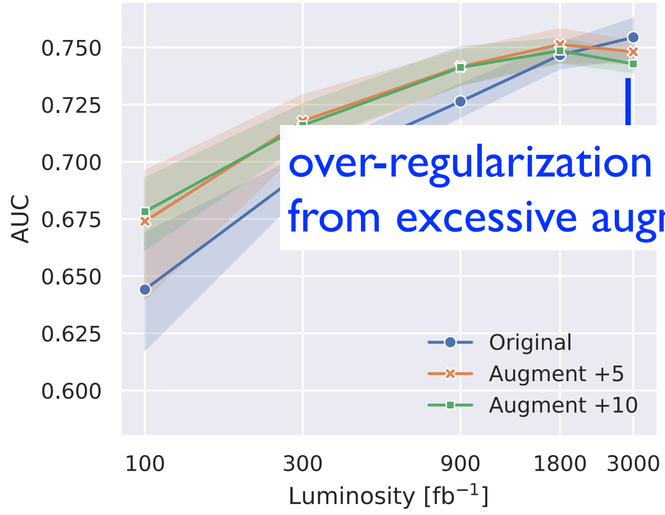
ϕ-shifting also beneficial when decay-product features are removed



+5 and +10 have similar effects

ParT





over-regularization effects resulted from excessive augmentation

with photon

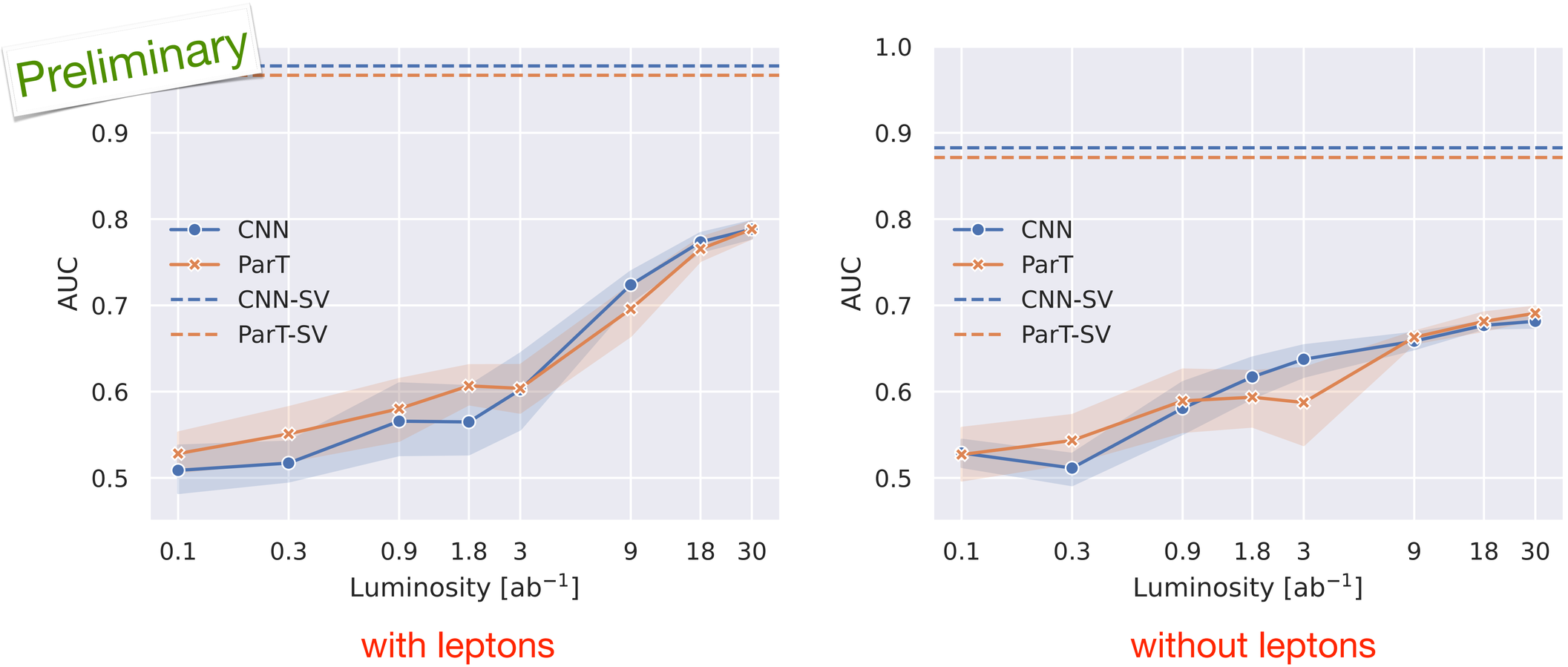
without photon

Remarks

- In general, CNN performs better [worse] than ParT when the datasets are small [large]
- Interestingly, at lower luminosities, models trained **without photon** information generally achieve higher AUC values.
- This suggests that, when the available training data are more limited, the networks tend to **overfit to the simpler photon features**, which has little to do with the initial state, rather than learning the more complex hadronic structures.
- When photon information is removed, the models are forced to focus on **hadronic activity patterns**, thereby achieving better generalization and higher performance in the low-statistics regime.
- Overall, these results demonstrate that **the primary discriminative power arises from the hadronic activities** in the event, and that explicit inclusion of photon information is not essential for achieving optimal classification performance.

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

Chen, Chen, CWC, Hsieh, in preparation



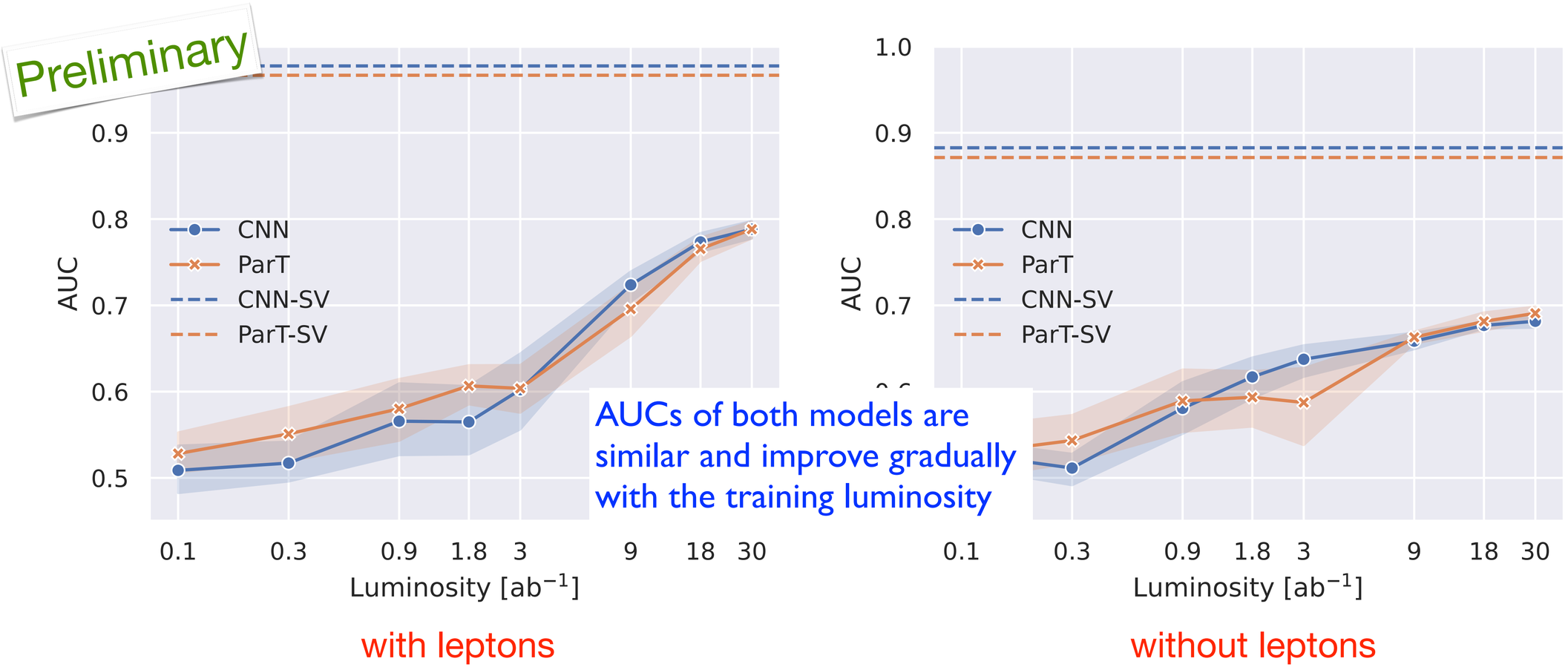
with leptons

without leptons

- Even at $\mathscr{L} = 3\ \text{ab}^{-1}$, the AUCs remain modest and the fluctuations are substantial, reflecting the **severe data scarcity** in this decay channel.
  ➠ luminosities required for stable training are well beyond those achievable in realistic experiments

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

Chen, Chen, CWC, Hsieh, in preparation



with leptons

without leptons

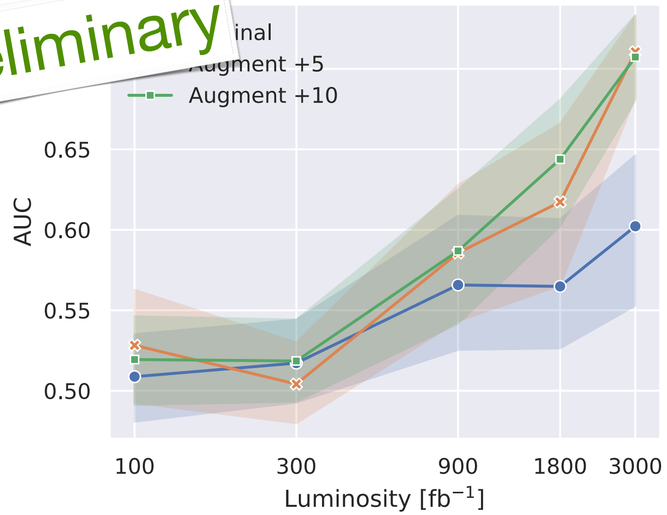
- Even at $\mathscr{L} = 3\ \mathrm{ab}^{-1}$, the AUCs remain modest and the fluctuations are substantial, reflecting the **severe data scarcity** in this decay channel.
  ➠ luminosities required for stable training are well beyond those achievable in realistic experiments

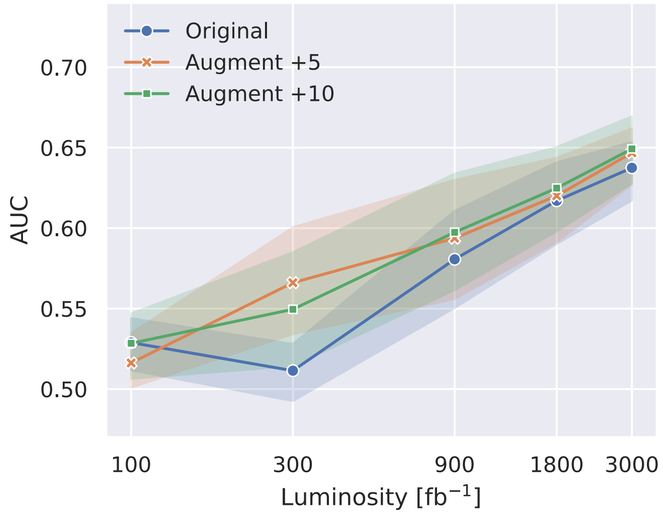
With Data Augmentation

Chen, Chen, CWC, Hsieh, in preparation

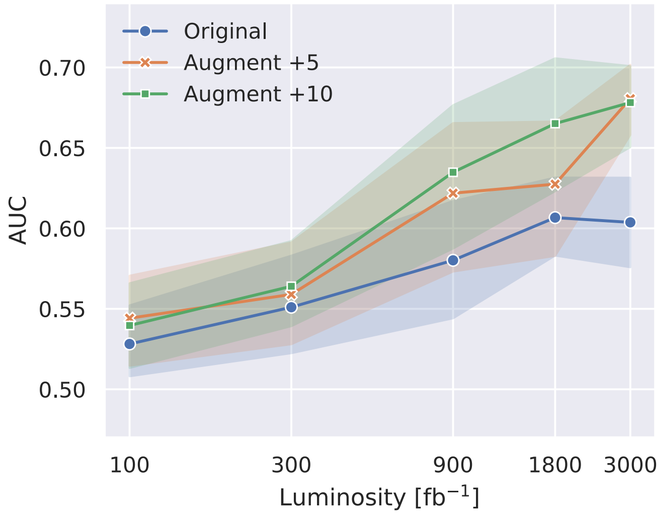
Preliminary

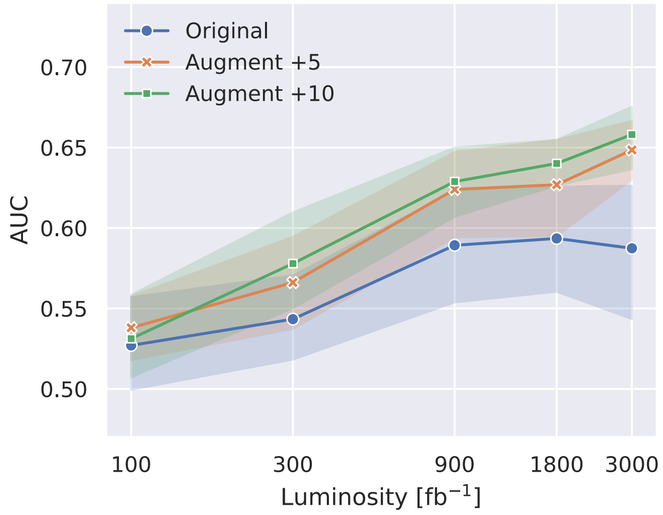
CNN





ParT





with leptons

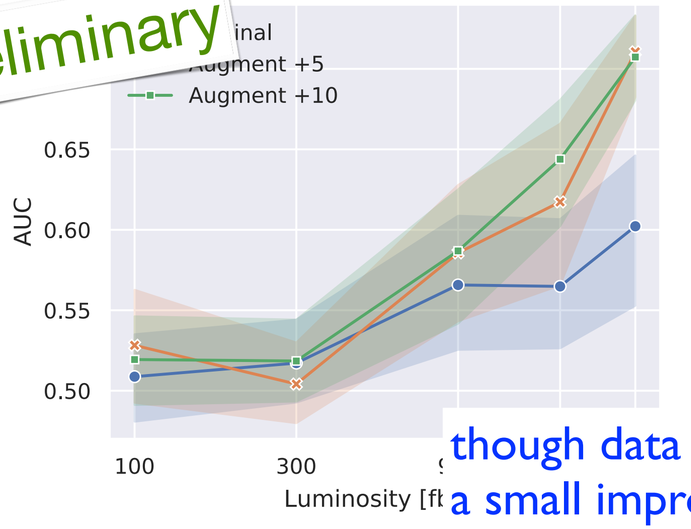
without leptons

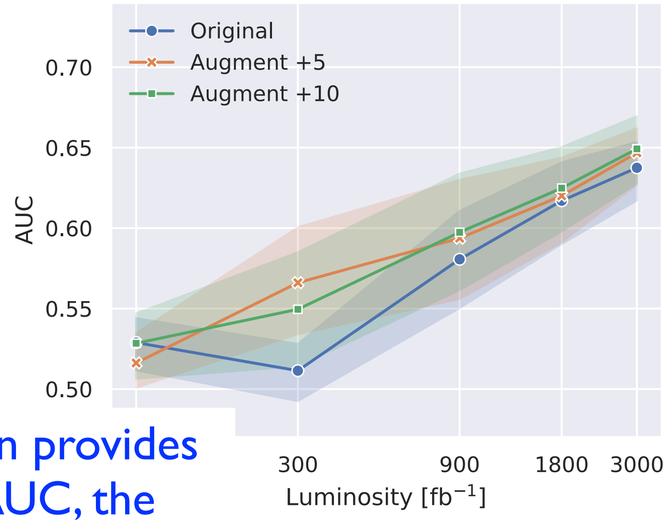
With Data Augmentation

Chen, Chen, CWC, Hsieh, in preparation

Preliminary

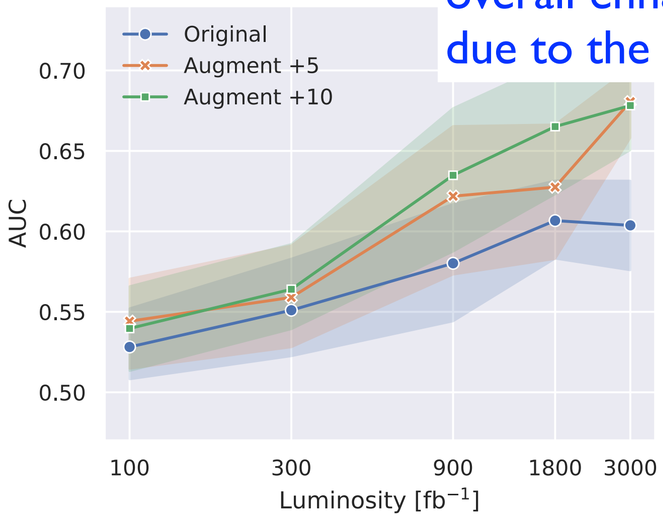
CNN

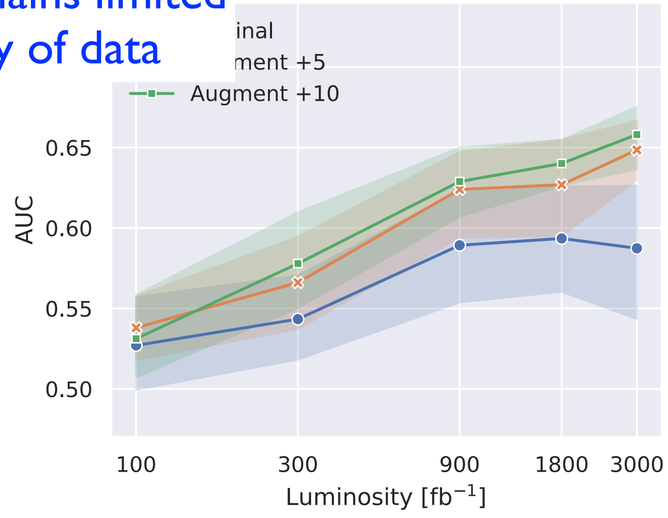




though data augmentation provides a small improvement in AUC, the overall enhancement remains limited due to the extreme rarity of data

ParT



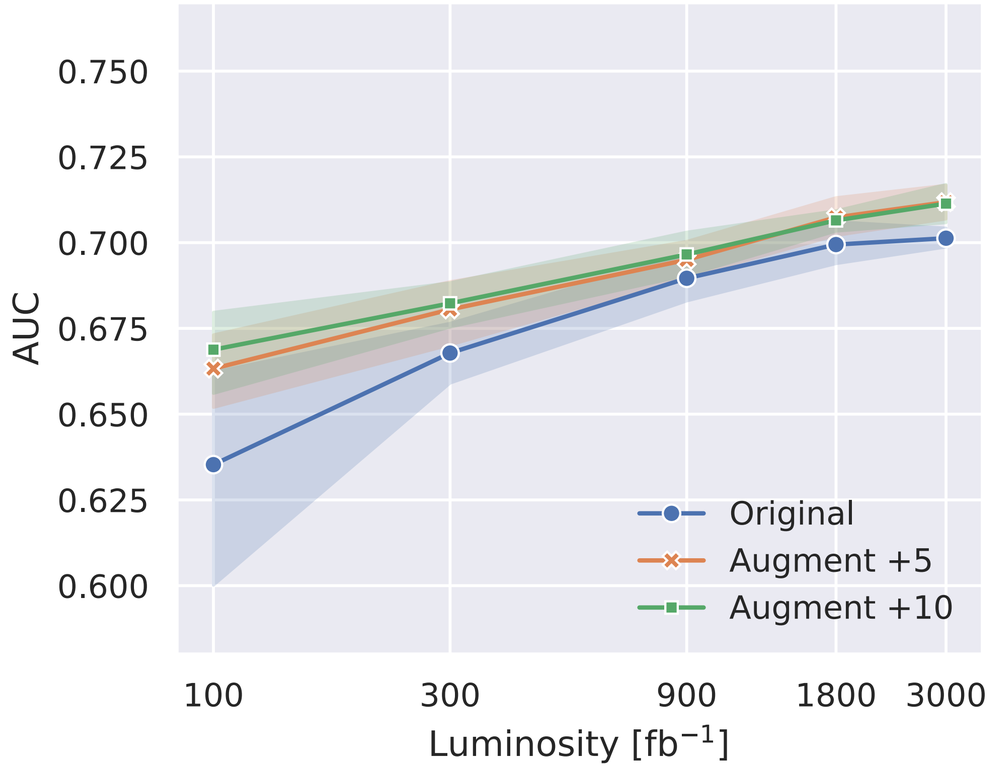


with leptons

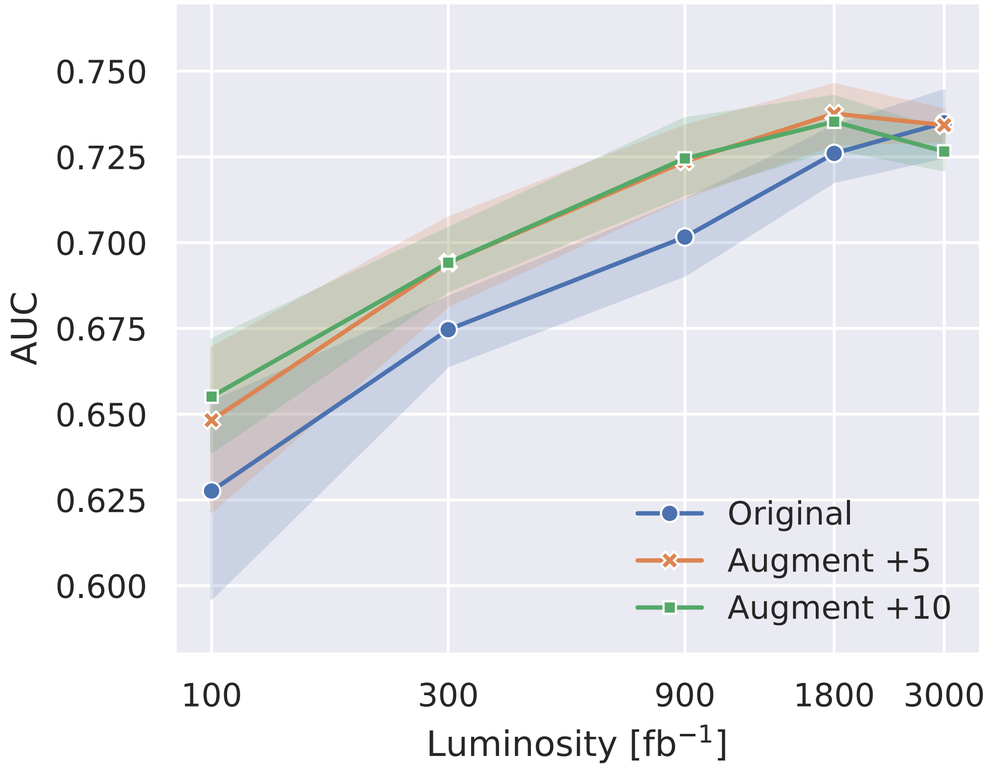
without leptons

Transfer Learning

Chen, Chen, CWC, Hsieh, in preparation



CNN

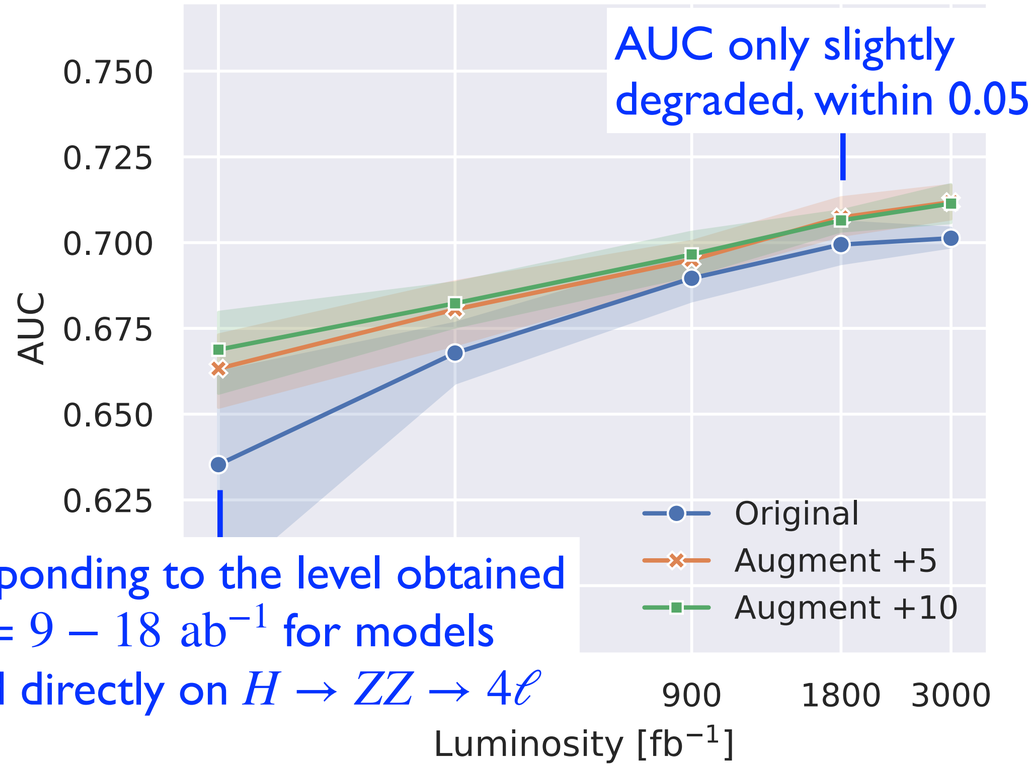


ParT

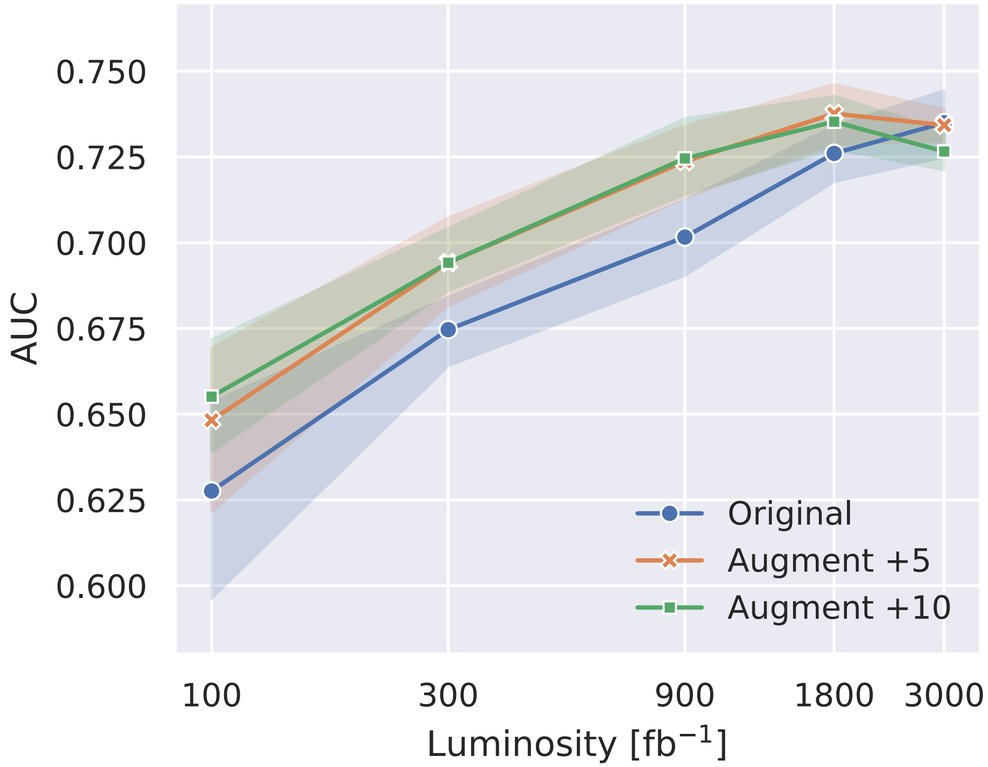
- Explore a **transfer learning** strategy by applying models trained on the high-statistics $H \to \gamma\gamma$ dataset to the $H \to ZZ \to 4\ell$ events, both removing the decay-product information.

Transfer Learning

Chen, Chen, CWC, Hsieh, in preparation



CNN



ParT

- Explore a **transfer learning** strategy by applying models trained on the high-statistics $H \to \gamma\gamma$ dataset to the $H \to ZZ \to 4\ell$ events, both removing the decay-product information.

Summary

- To achieve a reliable determination of the Higgs production mechanism in hadron collider experiments, we employ **weak supervision**, **CWoLa** in particular, to train deep neural networks using real data of the diphoton events, in the hope of reducing biases resulting from Monte Carlo simulations.
- Models based on the **convolutional neural network** and the **transformer** are tested and compared.
- The classification performance gets slightly better when the **photon information is removed** from training in the **low-luminosity region**.
- We show that the performance can be improved when the training dataset is enlarged by **data augmentation** using physics-motivated methods.
- We further demonstrate that the trained model can be successfully applied to the $H \to ZZ$ events, showing that such classifiers are **agnostic to Higgs decay modes** provided they do not involve strong QCD corrections.

Thank You!

Backup Slides

Data Preprocessing

- Each event goes through the following preprocessing steps:
  - Compute the **variance** of ϕ of all event constituents. If this variance exceeds 0.5, all ϕ values are shifted by π to **center the ϕ distribution**. This procedure prevents a significant fraction of event constituents from crossing the $\pm\pi$ boundary.
  - The ϕ coordinates of all event constituents are **centered with respect to the p_{T}-weighted mean**:

transverse momentum of the i-th constituent

$$\phi \to \phi - \frac{1}{p_{\mathrm{T}}^{\mathrm{sum}}} \sum_i p_{\mathrm{T},i}\phi_i, \quad p_{\mathrm{T}}^{\mathrm{sum}} = \sum_i p_{\mathrm{T},i}$$

azimuthal angle of the i th constituent

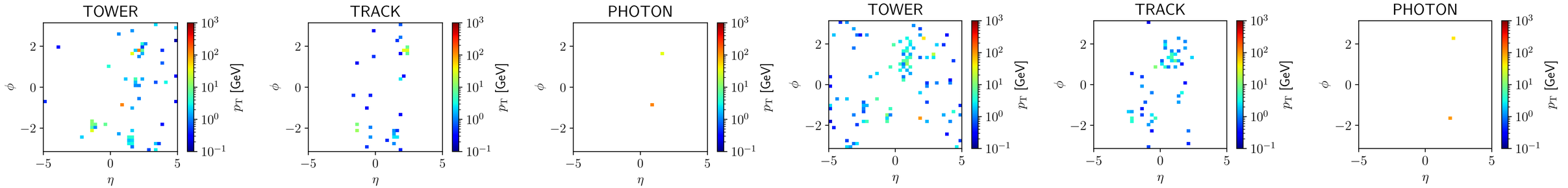
ensuring that the event is rotationally aligned around its p_{T}-weighted centroid.

Data Preprocessing

- Each event goes through the following preprocessing steps:
  - Divide the event into **four quadrants** based on the signs of ϕ and η : $(\phi > 0, \eta > 0)$, $(\phi > 0, \eta < 0)$, $(\phi < 0, \eta > 0)$, and $(\phi < 0, \eta < 0)$. The quadrant with the highest total transverse momentum is identified and reflected into the first quadrant, i.e., the $(\phi > 0, \eta > 0)$ region, by mirroring along the $\phi = 0$ and $\eta = 0$ axes.
  - To **remove decay-product information**, all particles within a region of $|\Delta\phi| < \pi/40$ and $|\Delta\eta| < 5/40$ around each decay product are excluded. These thresholds are chosen to match the corresponding ϕ and η grid divisions of the image representation, ensuring consistency between the image-based and set-based removal procedures.

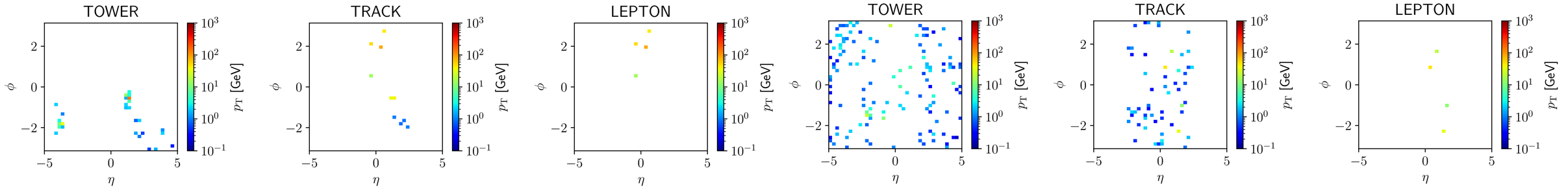
Comparison of p_{T} Distributions

- Comparison of p_{T} distributions for VBF (left triplet) and GGF (right triplet) processes with $H \to \gamma\gamma$ (upper rows) and $H \to ZZ \to 4\ell$ (lower row) events.
- Each has three channels: **tower**, **track**, and **decay-product information**.



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

(c) GGF with $H \to \gamma\gamma$



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

(d) GGF with $H \to ZZ \to 4\ell$

Simplified Particle Transformer

- Unlike the original ParT, we **omit the interaction matrix** for two reasons.
  - First, our dataset does not provide full four-momentum information for all inputs, reducing the utility of **pairwise interaction modeling**.
  - Second, the inputs consist of heterogeneous objects (calorimeter towers, tracks, and decay products) rather than fully reconstructed particles, making it **difficult to define a consistent and physically meaningful interaction representation**.
- Our implementation includes **one particle attention block** and **one class attention block**. The original ParT configuration was found to **overfit** on our dataset; the current, simplified architecture offers stable performance across repeated trials. The model has approximately **9.5K** trainable parameters. A detailed description of the hyperparameters used in this setup is provided in Appendix A of our paper.

Hyperparameters of Particle Transformer

- We **omit the interaction embedding** used in the original implementation. The final configuration used in this study is outlined below:
  - **Particle Embedding**: Input particle features are embedded into a latent space of dimension $d = 16$ using a **three-layer multilayer perceptron** with hidden dimensions of 16, 64, and 16. Each layer uses `GeLU` activation functions, and **layer normalization** is applied between layers to ensure stable training.
  - **Particle Attention Block**: One particle attention block with a **dropout rate** of 0.1 is used, with 4 attention heads. The feedforward network consists of two linear layers with 64 and 16 hidden units, respectively.
  - **Class Attention Block**: One class attention block with no dropout is included, also using 4 attention heads. Its feedforward layers mirror those of the particle attention block (64 and 16 hidden units).
- All remaining components and architectural details follow the original one.