Lecture @ University of Osaka
June 11, 2025

Machine learning and its application to high-energy physics

Cheng-Wei Chiang (蔣正偉)
National Taiwan University
National Center for Theoretical Sciences

Lecture @ University of Osaka
June 11, 2025

Machine learning and its application to high-energy physics — *A very biased introduction*

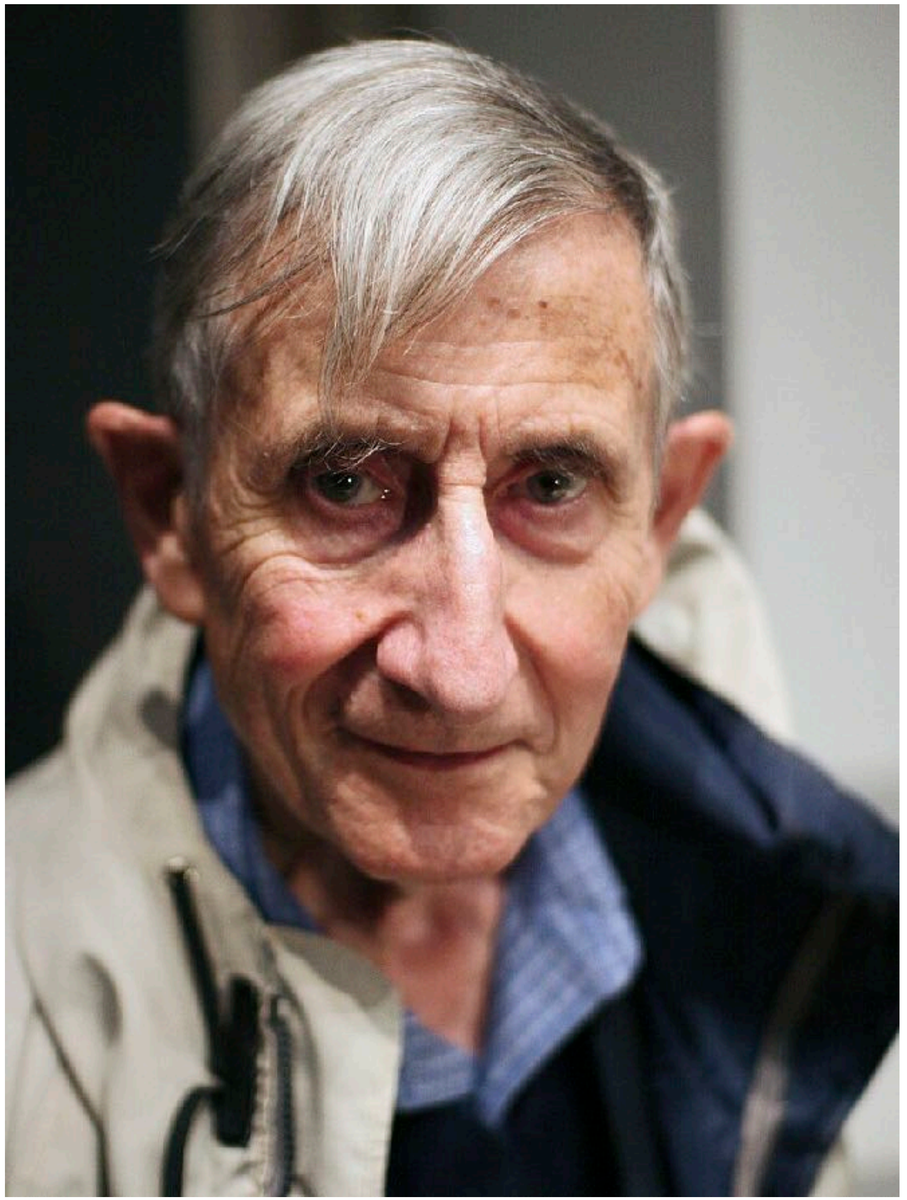
Cheng-Wei Chiang (蔣正偉)
National Taiwan University
National Center for Theoretical Sciences

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 [hep-ph], to appear in JHEP

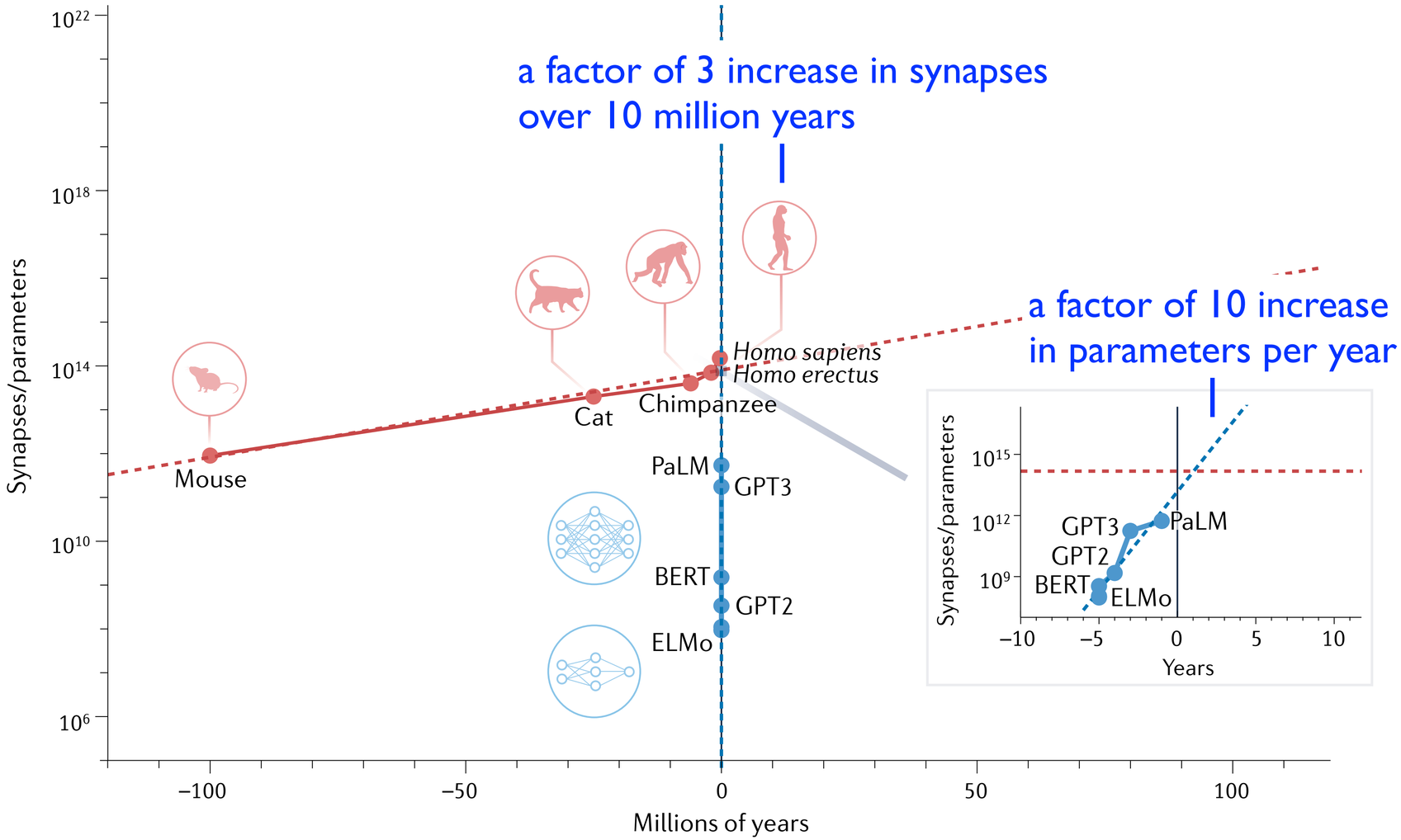
Revolution is Driven by New Tools

"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)



Evolution of Biological Brains and Artificial Intelligence



Schwartz 2022

Fig. 1 | The evolution of biological and artificial intelligence takes place on dramatically different timescales. Any hope of interpreting and understanding AI will exponentially fade. Some example data points are highlighted in the evolution of biological (red) and artificial (blue) intelligence. The dashed lines represent the linear regression to these points. The acronyms in the figure are: Pathways Language Model (PaLM), Embeddings from Language Model (ELMo), Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT).

When Will AI Win Us a Nobel Prize?

When Will AI Win Us a Nobel Prize?

© Nobel Prize Outreach. Photo: Nanaka Adachi

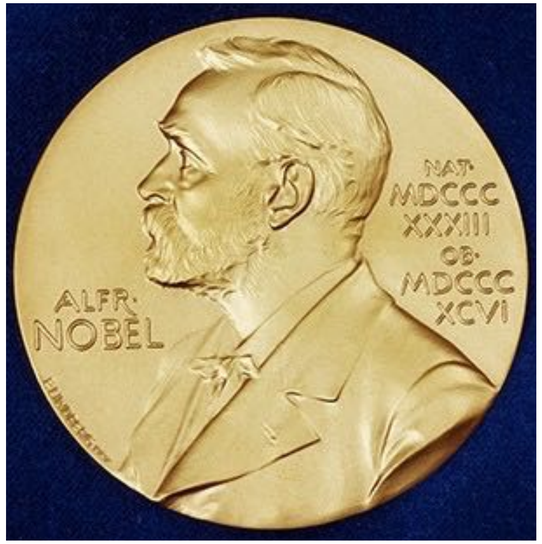
John J. Hopfield

Prize share: 1/2

© Nobel Prize Outreach. Photo: Clément Morin

Geoffrey Hinton

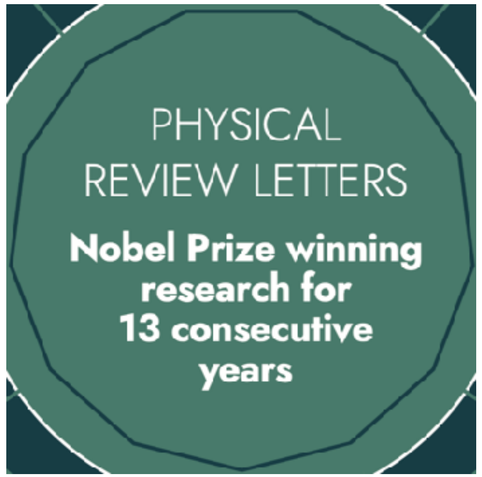
Prize share: 1/2



The Nobel Prize in Physics 2024 was awarded jointly to John J. Hopfield and Geoffrey Hinton "for foundational discoveries and inventions that enable machine learning with artificial neural networks"

When Will AI Win Us a Nobel Prize?

- When will machines be able to
  - come up with a **meaningful** and **original** research project;
  - conduct necessary **calculations** (and perhaps experiment?);
  - write up **arXiv** papers;
  - pass peer reviews and publish them in **leading journals**; and
  - perhaps win a **Nobel Prize** for the breakthrough?



Google's Minerva

- Google builds **Minerva** based upon the **Pathways Language Model (PaLM)** and further train the neural network on a **118-GB dataset** of scientific papers from the **arXiv** preprint server and web pages that contain documents prepared using **LaTeX**, **MathJax**, or **other mathematical typesetting formats**.
- The model is able to communicate using the standard mathematical notation.
  ➠ **multi-step quantitative reasoning** shown to be possible

Question: For every $a, b, b \neq a$ prove that

$$\frac{a^2+b^2}{2} > \left(\frac{a+b}{2}\right)^2.$$

Model output:

$$\frac{a^2+b^2}{2} > \left(\frac{a+b}{2}\right)^2$$

$$\iff \frac{a^2+b^2}{2} > \frac{a^2+b^2+2ab}{4}$$

$$\iff a^2+b^2 > \frac{a^2+b^2+2ab}{2}$$

$$\iff 2a^2+2b^2 > a^2+b^2+2ab$$

$$\iff a^2+b^2 > 2ab$$

$$\iff a^2+b^2-2ab > 0$$

$$\iff (a-b)^2 > 0$$

which is true, because the square of a real number is positive.

https://research.google/blog/minerva-solving-quantitative-reasoning-problems-with-language-models/

Google's Minerva

- If this makes you feel better, Minerva also makes **trivial mistakes**, such as:

Question: If $\sqrt{400} = \sqrt{81} + \sqrt{n}$, then what is the value of n?

Model output: We have that $\sqrt{400} = \sqrt{81} + \sqrt{n}$. Therefore, $400 = 81 + n$, which gives $319 = n$.
Final Answer: The final answer is 319.

https://research.google/blog/minerva-solving-quantitative-reasoning-problems-with-language-models/

Google's Minerva

- If this makes you feel better, Minerva also makes **trivial mistakes**, such as:

Question: If $\sqrt{400} = \sqrt{81} + \sqrt{n}$, then what is the value of n?

Model output: We have that $\sqrt{400} = \sqrt{81} + \sqrt{n}$. Therefore, $400 = 81 + n$, which gives $319 = n$.
Final Answer: The final answer is 319.

https://research.google/blog/minerva-solving-quantitative-reasoning-problems-with-language-models/

- Such a mistake is **not uncommon** for high school (or even college?) students after years of training.

Google's Minerva

- If this makes you feel better, Minerva also makes **trivial mistakes**, such as:

Question: If $\sqrt{400} = \sqrt{81} + \sqrt{n}$, then what is the value of n?

Model output: We have that $\sqrt{400} = \sqrt{81} + \sqrt{n}$. Therefore, $400 = 81 + n$, which gives $319 = n$.
Final Answer: The final answer is 319.

https://research.google/blog/minerva-solving-quantitative-reasoning-problems-with-language-models/

- Such a mistake is **not uncommon** for high school (or even college?) students after years of training.

- In fact, machines already have a passable understanding of **college-level science** (but across many subjects and within a short time of training!).

Should We Worry About Our Future?

Should We Worry About Our Future?

- Over the years, with improved computing **power** and **efficiency**, machines have been shown to **surpass** human beings in analyzing complicated things (such as particle physics data, as we will see).
  ➠ should we be worried?

Should We Worry About Our Future?

- Over the years, with improved computing **power** and **efficiency**, machines have been shown to **surpass** human beings in analyzing complicated things (such as particle physics data, as we will see).
  ➠ should we be worried?
- With a possibly **fundamentally different** way of "**understanding**" the world (input data), it will probably be **impossible to comprehend or interpret** in our way how machines work.
  ➠ should we be worried?

Should We Worry About Our Future?

- Over the years, with improved computing **power** and **efficiency**, machines have been shown to **surpass** human beings in analyzing complicated things (such as particle physics data, as we will see).
  ➠ should we be worried?
- With a possibly **fundamentally different** way of "**understanding**" the world (input data), it will probably be **impossible to comprehend or interpret** in our way how machines work.
  ➠ should we be worried?
- **Probably not**.
  ➠ we admire many **genius** musicians, artists, and scientists around us and do not know why they are so smart, but never worry about their existence
  ➠ just like computers help to revolutionize human life, we should **exploit** and **embrace** the **immense power of AI** to help us explore the Universe

Outline

- Introduction to deep learning
- Full supervision
- Weak supervision — CWoLa
- Dark valley model — a physical model
- Transfer learning
- Data augmentation
- Summary

Outline

- Introduction to deep learning
- Full supervision
- Weak supervision — CWoLa
- Dark valley model — a physical model
- Transfer learning
- Data augmentation
- Summary

Machine Learning

- **Machine learning (ML)** is the tool used for **large-scale data processing** and is 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.

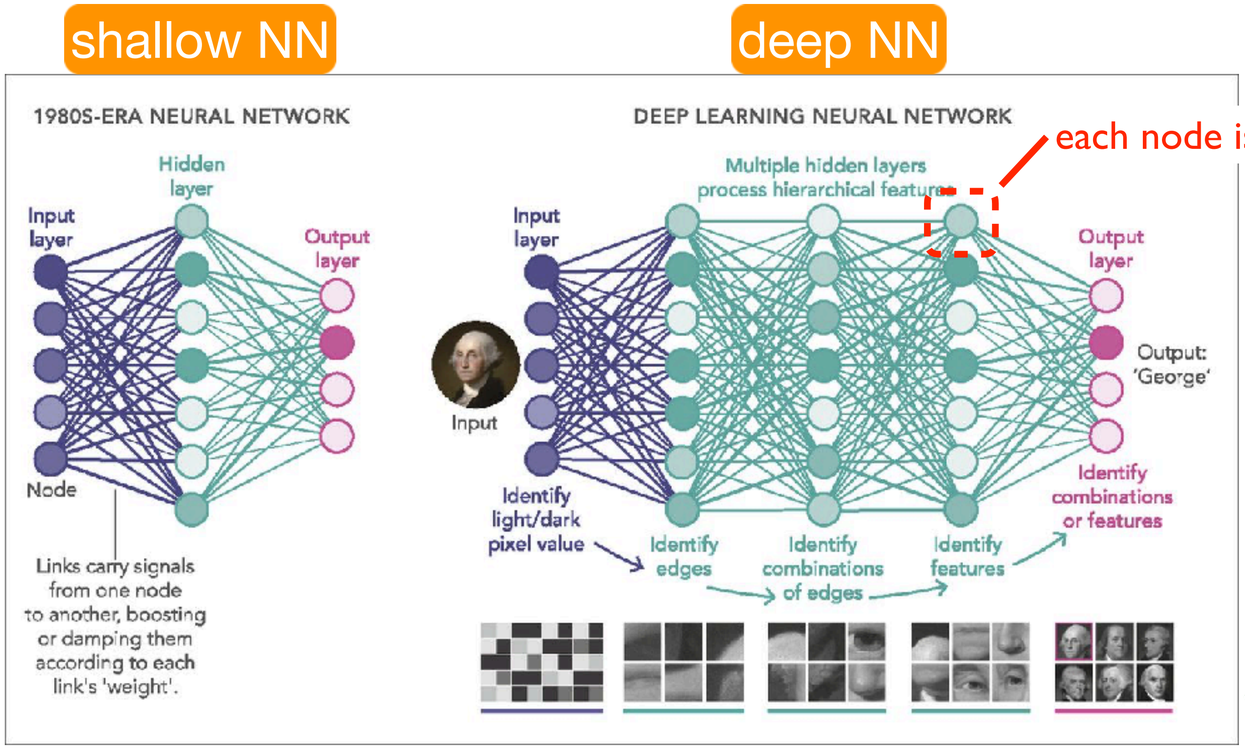
- For a long time, given this theorem and the difficulty in complex networks, people have restricted themselves to **shallow** networks with only **one hidden layer**.
- Recently, people have realized that **deeper**, more **complex** networks with many hidden layers can “understand” higher levels of abstraction better than shallow layers.

Resurgence of Neural Networks

- Neural networks (NNs) have resurged in the last decade partly due to:
  - faster computers, with the **use of GPUs** versus the traditional use of CPUs,
  - better, **deeper algorithms** and NN architecture designs, and
  - increasingly **large datasets** being available for training.

shallow NN

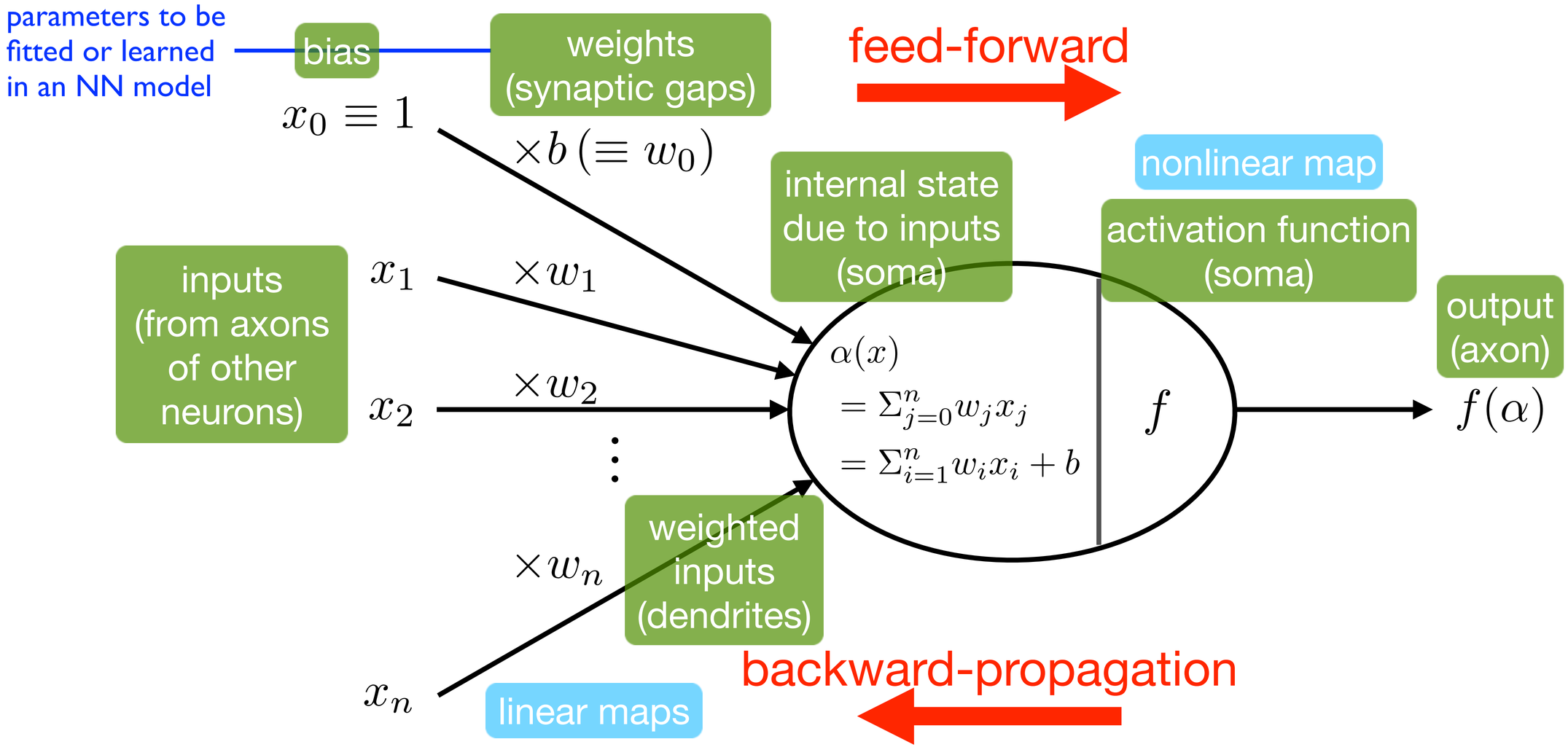
deep NN



Waldrop 2019

Artificial Neuron

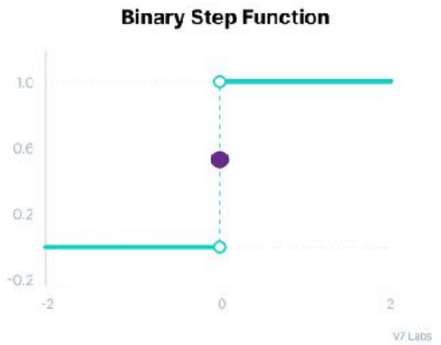
- Different types of artificial neurons are modeled using different **activation functions**, which are required to introduce **nonlinearity** to the process.



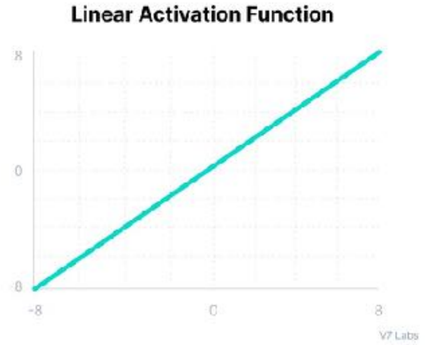
Common Activation Functions

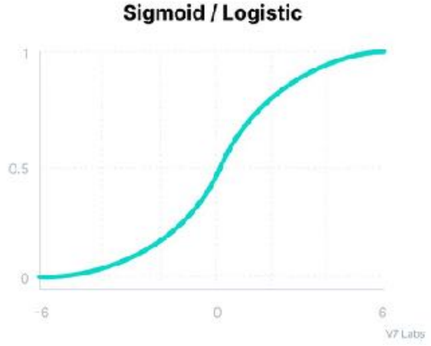
- The choice of activation function is mostly determined by the nature of the problems at hand.

$$f(x) = x$$



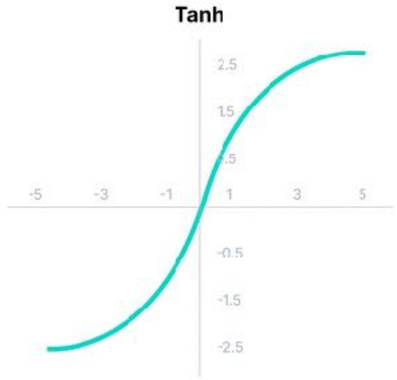
$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geqslant 0 \end{cases}$$

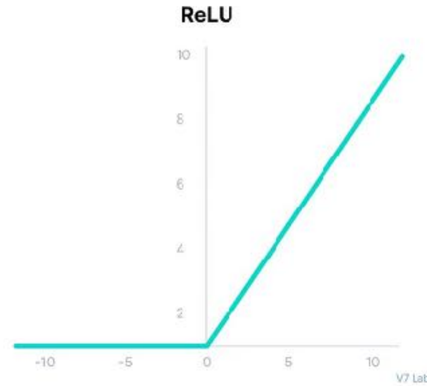




$$f(x) = \frac{1}{1 + e^{-x}}$$

$$f(x) = \tanh x$$





$$f(x) = \max(0, x)$$

https://ai.plainenglish.io/activation-functions-in-neural-networks-9f0170334bf5

Types of Machine Learning

- **Supervised learning**
  - Training data with labels (e.g., recognizing photos of cats and dogs)
- **Unsupervised learning**
  - Training data without labels (e.g., analyze and cluster unlabeled datasets)
- **Reinforced learning**
  - Data from interactions with the environment (e.g., chess and Go games)
- **Weakly supervised learning**
  - When data labeling is infeasible, imperfect, difficult, or expensive.

Simple Types of Neural Networks

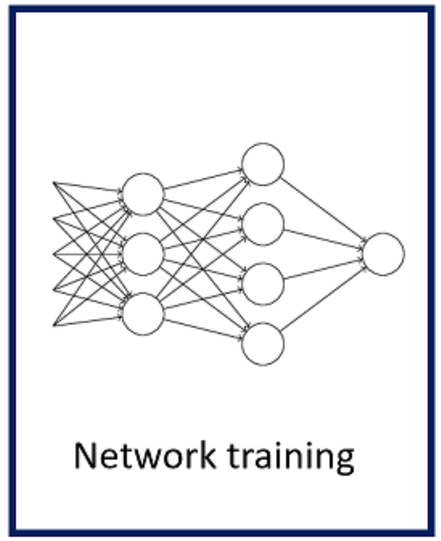
- **Dense neural network** (**Dense-NN**): a network with standard **fully-connected** feed-forward layers that take flattened vectors as the input, prototypical for most tasks; sometimes also called **multi-layer perceptron (MLP)**.
- **Recurrent neural network** (**RNN**): a network that deals with sequences of variable length by defining a recurrence relation over these sequences, suitable for **natural language processing (NLP)** and speech recognition tasks.
- **Convolutional neural network** (**CNN**)*: a network with special layers that **filter** image data, suitable for computer vision.
  ➠ ideal for **jet image recognition** tasks in collider physics

* Some evidence shows that neurons in CNNs are organized in a way similar to biological cells in the visual cortex of the human brain.

Script Digit Recognition

- One classic example of CNN is training the computer to recognize **hand-written digits** (with 60,000 training images and 10,000 testing images, and each image being normalized to 28×28 pixels and having 256 grey levels).

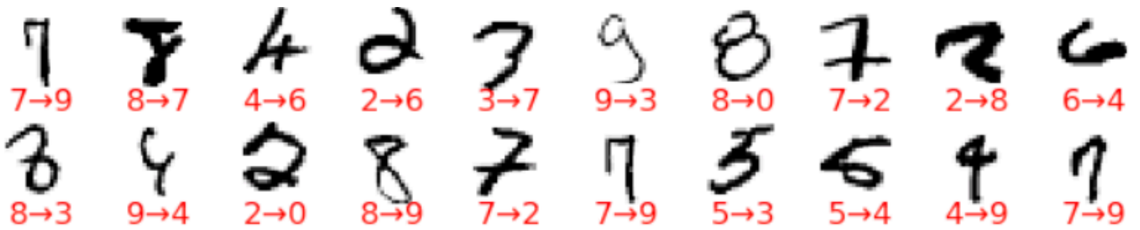
Data & Labels



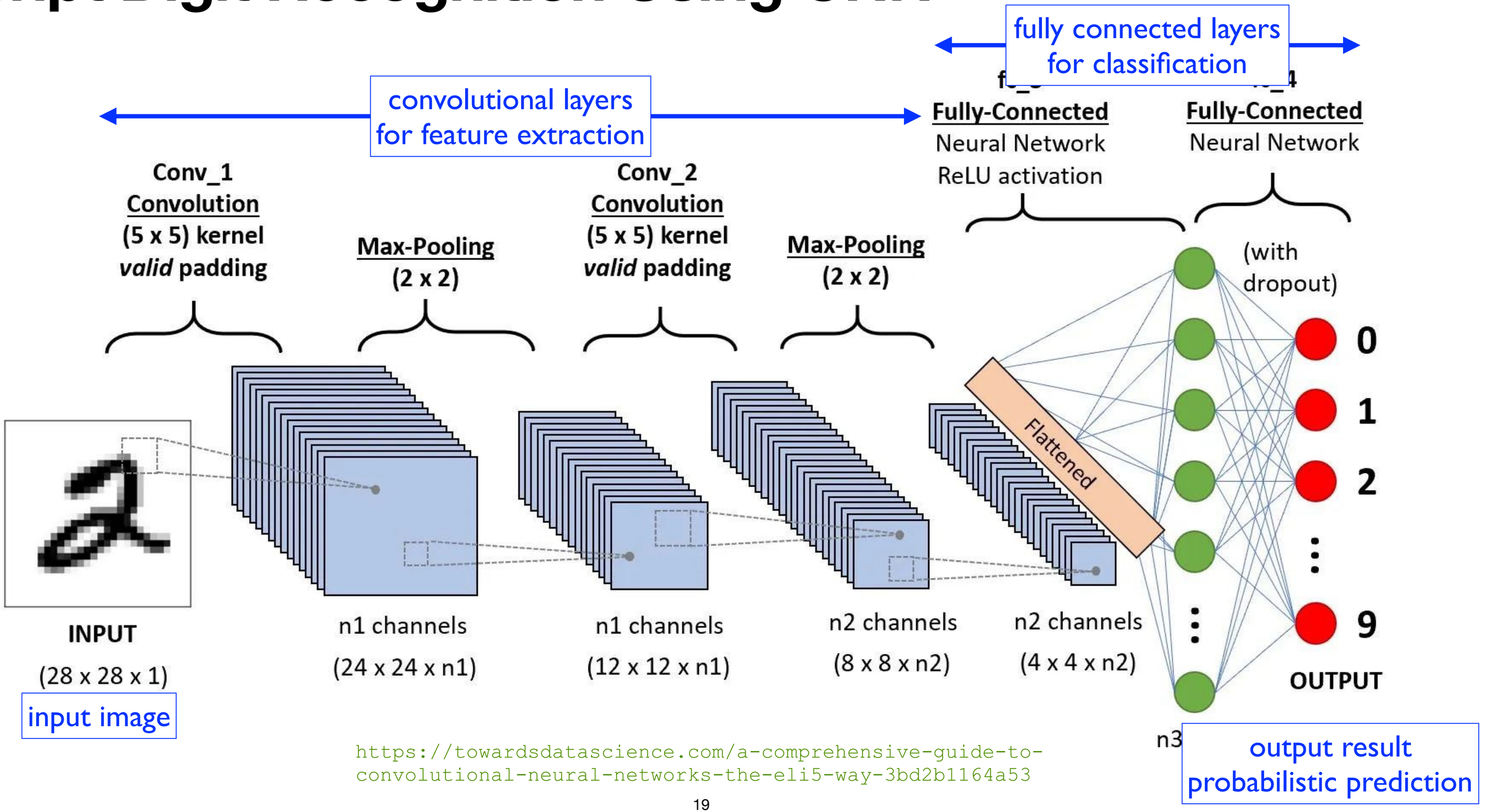
0
1
2
3
4
5
6
7
8
9

https://www.katacoda.com/basiafusinska/courses/tensorflow-getting-started/tensorflow-mnist-beginner

passing test samples to NN gives an accuracy of ~99%, with some mistakes from time to time:

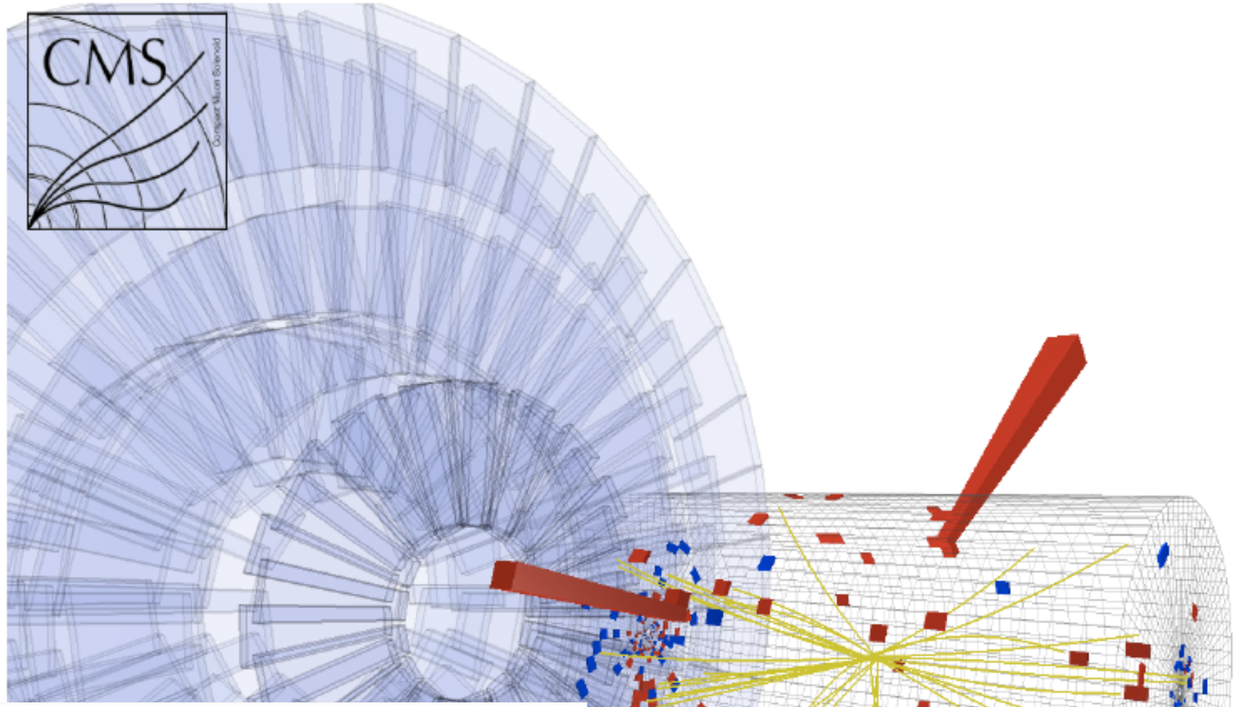


Script Digit Recognition Using CNN



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

A Higgs to Diphoton Event



Event parameters:

$M_{\gamma\gamma}$ = 125.9 GeV

$p_T^{\gamma 1}$ = 89.8 GeV

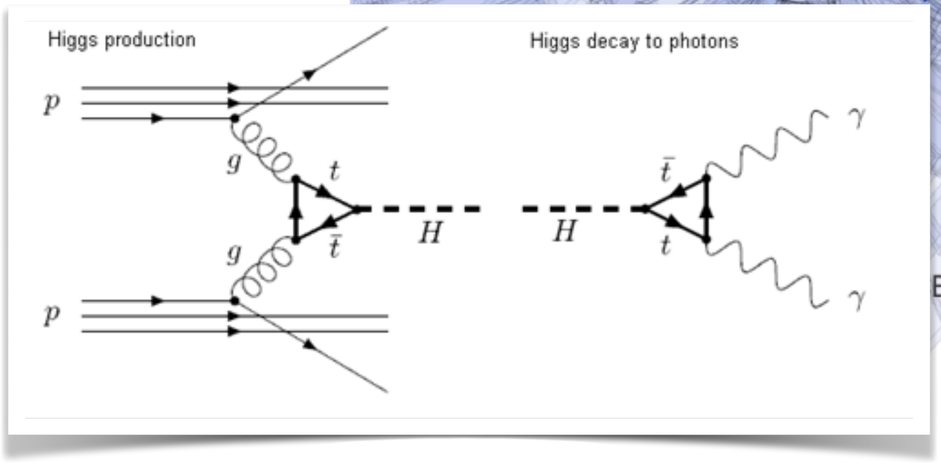
$p_T^{\gamma 2}$ = 46.5 GeV

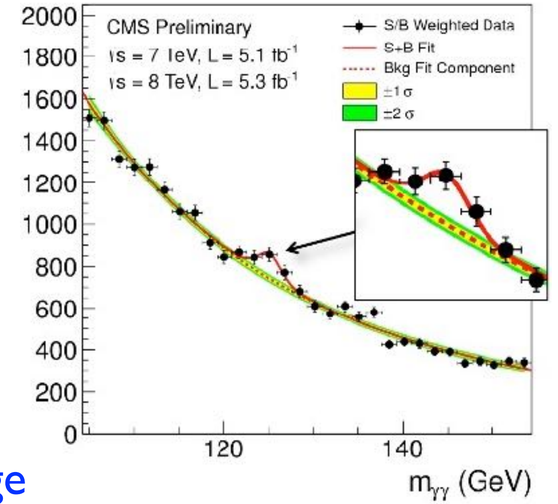
$\eta_{\gamma 1}$ = 0.06

$\eta_{\gamma 2}$ = -0.81

σ_M/M =0.89%

$p_T^{\gamma\gamma}$ =78.4 GeV

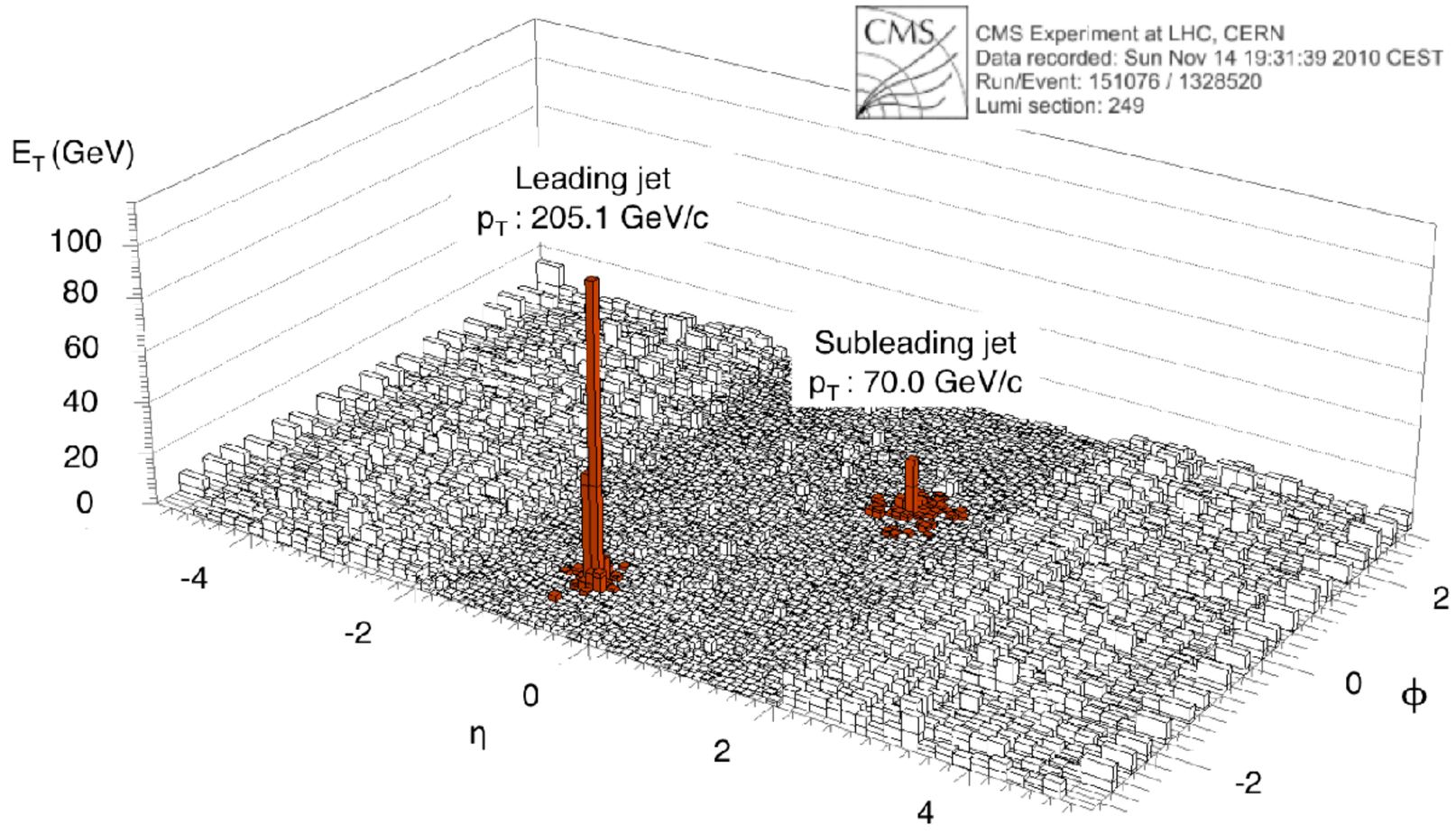




EST open up to render a 2D image

2D Image

- A histogram of jet transverse momentum in the η-ϕ plane from the detector data (cut along a particular ϕ).

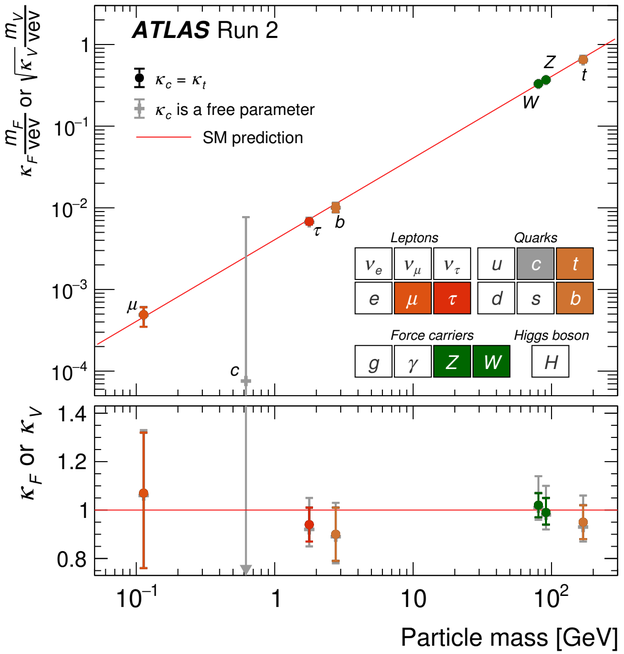


Outline

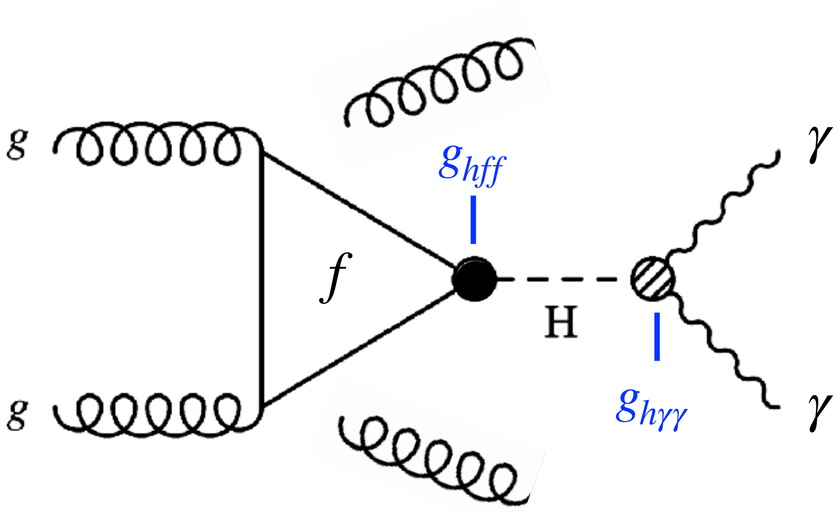
- Introduction to deep learning
- Full supervision
- Weak supervision — CWoLa
- Dark valley model — a physical model
- Transfer learning
- Data augmentation
- Summary

Higgs Physics Program

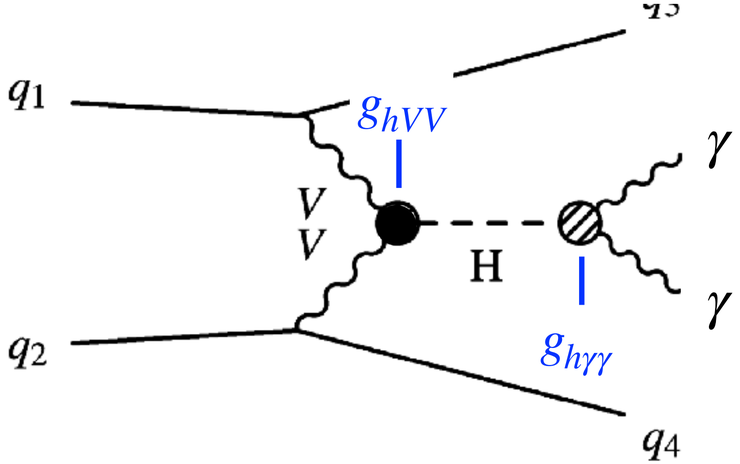
- The current Higgs physics program is to determine all the **Higgs couplings** precisely.
  ➠ look for any deviations and hints of new physics
- This requires the ability to tell apart the **two** dominant production channels (others being even smaller).
  ➠ cf. double-slit experiment



ATLAS 2019



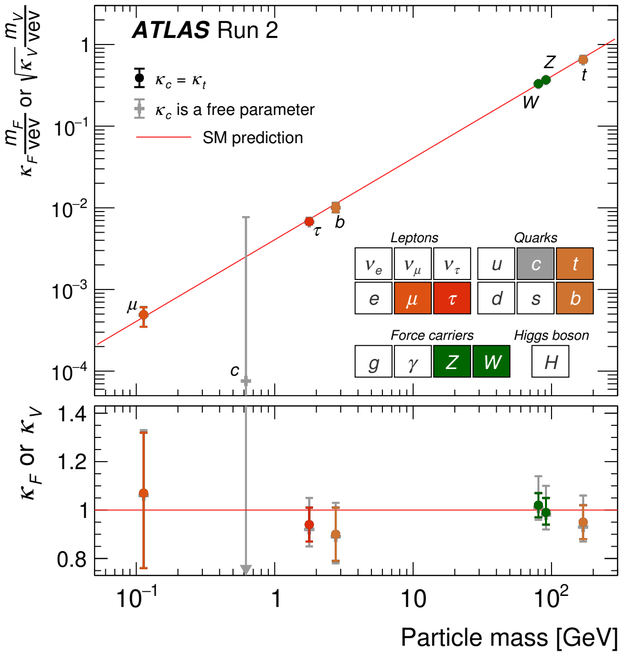
(a) ggF production



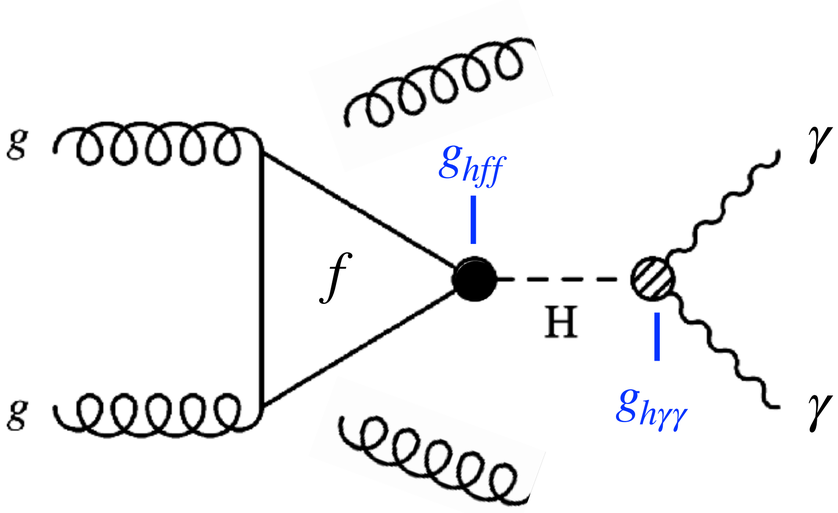
(b) VBF production

VBF/GGF Higgs Production

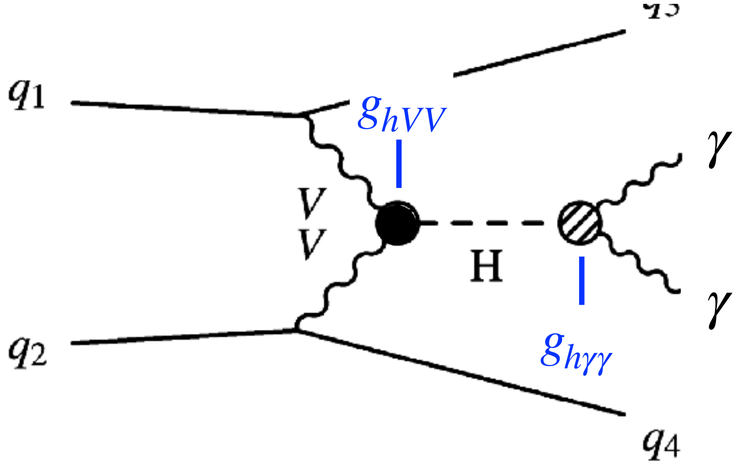
- Questions:
  - For each **detected** Higgs event, how can we **efficiently** and **correctly** determine/label its production mechanism?
  - Can it be **independent** of how the Higgs boson decays?



ATLAS 2019



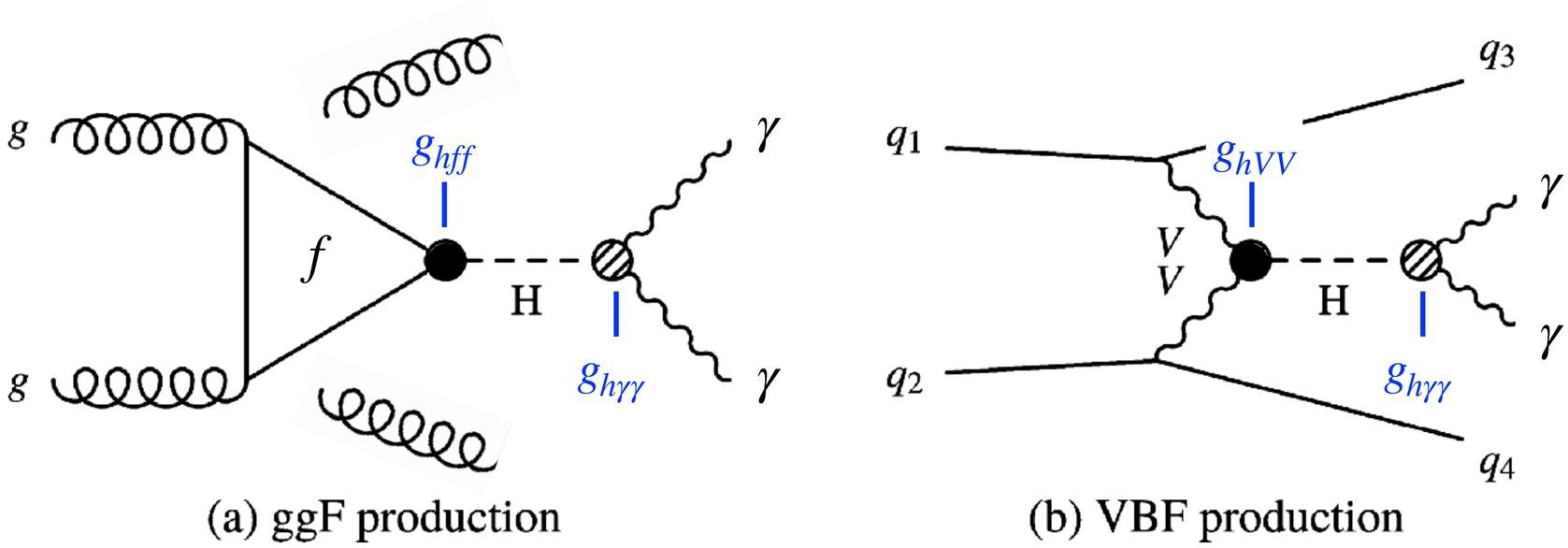
(a) ggF production



(b) VBF production

Two Observations

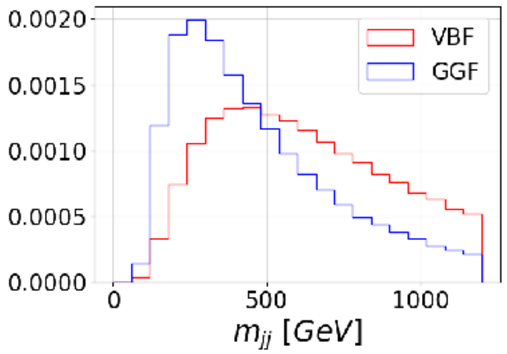
- VBF events come with two **forward spin-1/2 quark-initiated jets** from the hard process, while GGF jets tend to be **spin-1 gluon-initiated initial-state radiation**.
  ➠ *different jet constituent distributions*, particularly soft radiation patterns
- Since the Higgs is a **color singlet scalar with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

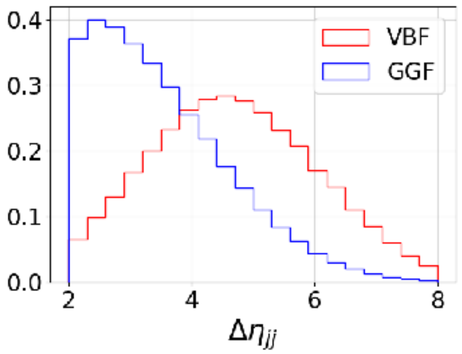


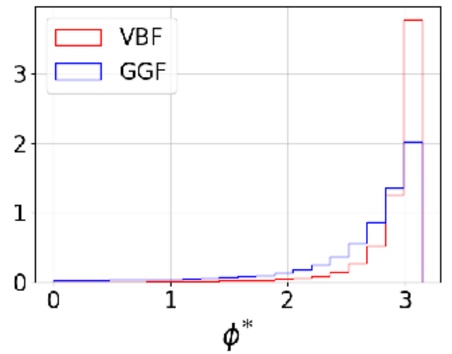
(a) ggF production (b) VBF production

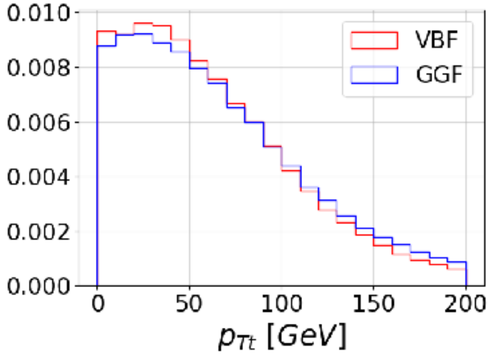
Using BDT With Human-Engineered Variables

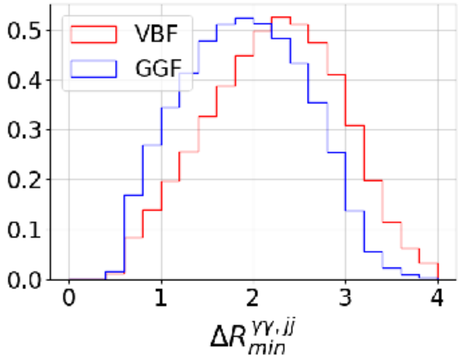
baseline

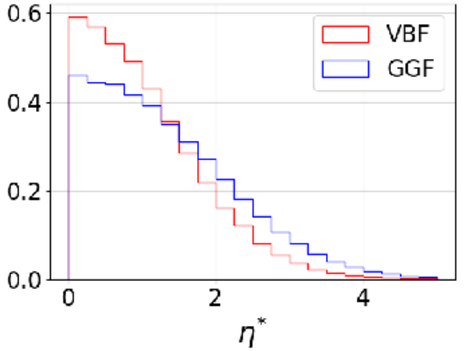




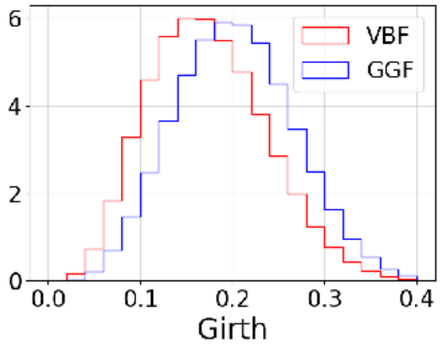


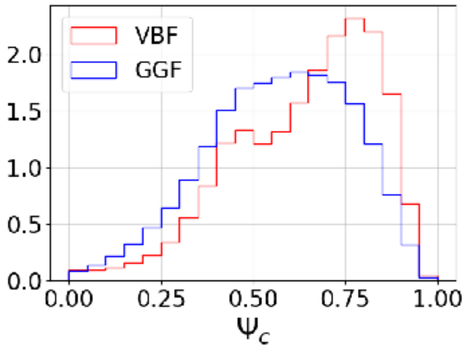


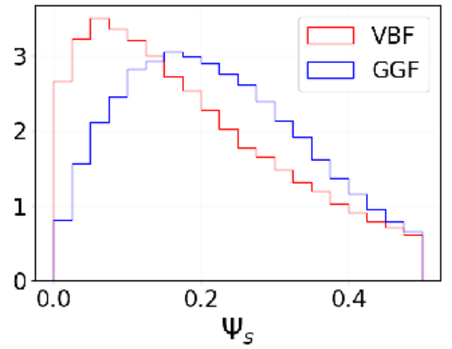




shapes







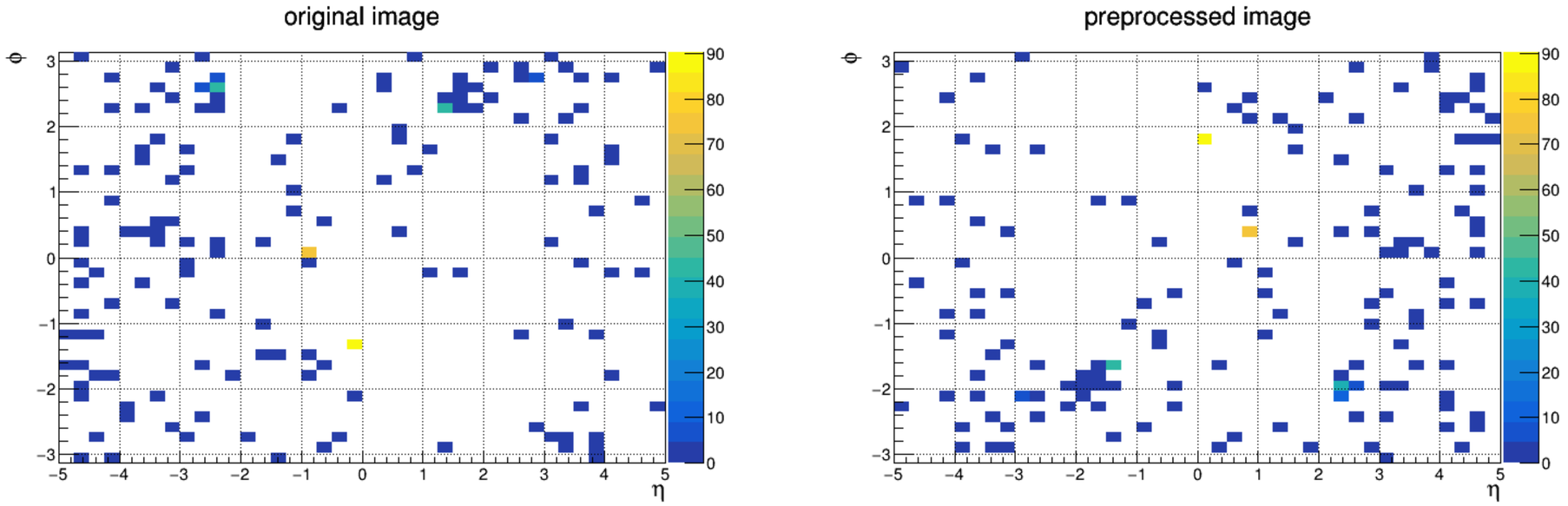
- Cut-based methods cannot reach high purity.
- BDT-based methods can achieve a purity of about 70% for the VBF sample, depending on the decay channel. ATLAS 2019

all histograms normalized to have unit area under the curves

Event-CNN

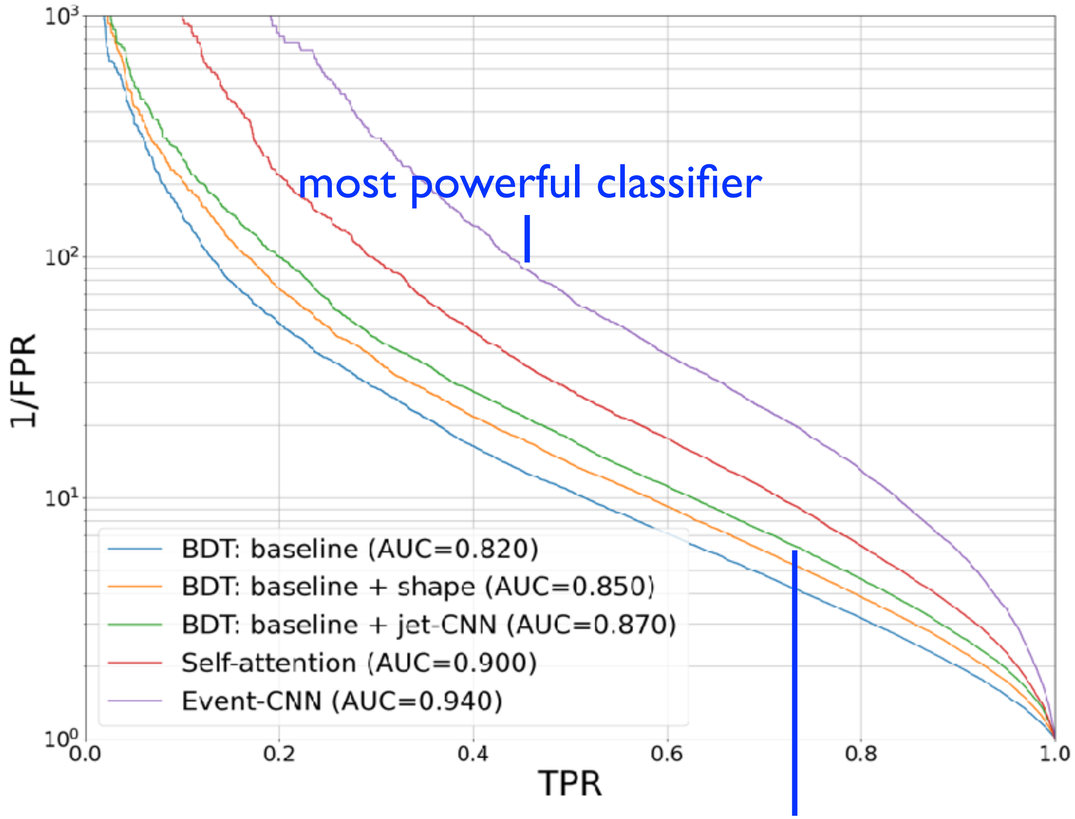
- Train a CNN by **full supervision** to discriminate the two production mechanisms by examining the final-state images.
- A successful training typically requires at least **tens of thousands** of samples.

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



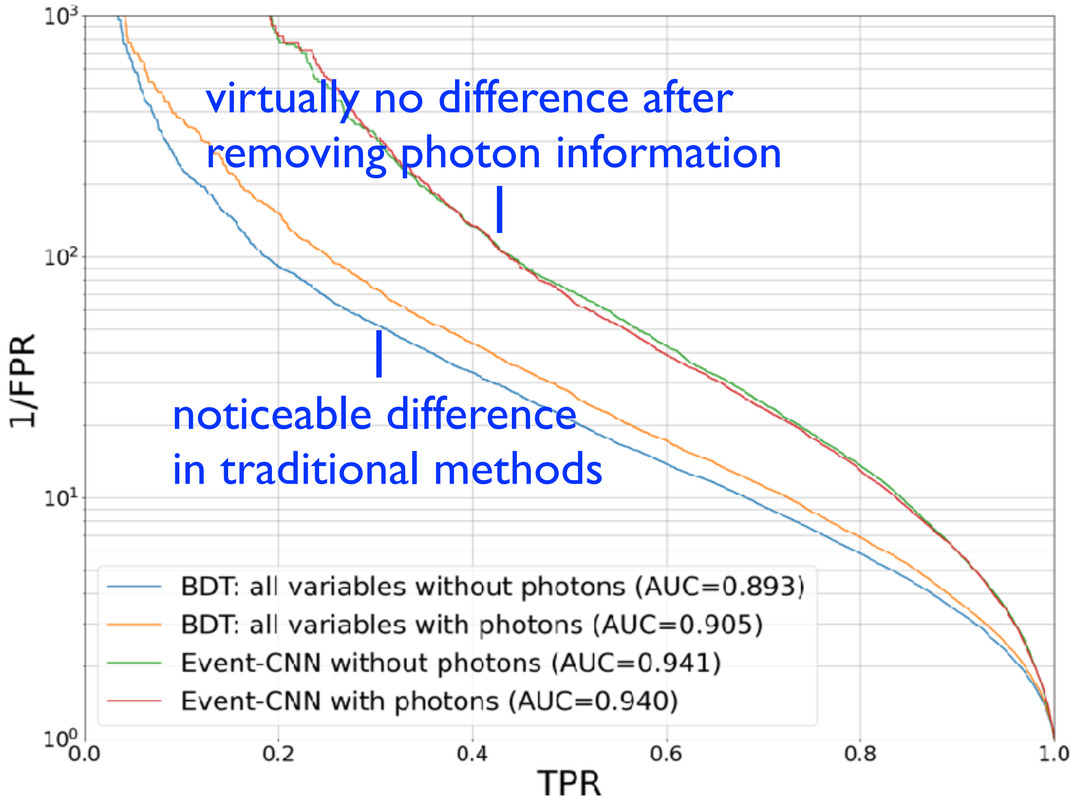
Comparison of Classifiers

ROC curves (Receiver Operating Characteristic curves)



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

ROC curves



CWC, Shih, Wei 2023

Requirements on Training Data

- **High-Quality Data**: The dataset should be **representative** of the problem domain and free of noise or irrelevant features. **Preprocessing** steps like removing outliers, handling missing values, standardization by utilizing symmetries, and balancing class distributions are crucial.

- **Sufficient Data**: Neural networks typically require **large amounts of labeled data** to learn meaningful patterns. When the dataset is small, techniques like **transfer learning** or **data augmentation** can mitigate data scarcity.

- **Data Diversity**: Samples in the datasets should be sufficiently **diverse** in properties in order to help the model **generalize** better and **avoid overfitting** to specific patterns.

Outline

- Introduction to deep learning
- Full supervision
- Weak supervision — CWoLa
- Dark valley model — a physical model
- Transfer learning
- Data augmentation
- Summary

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

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

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)

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?

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**…

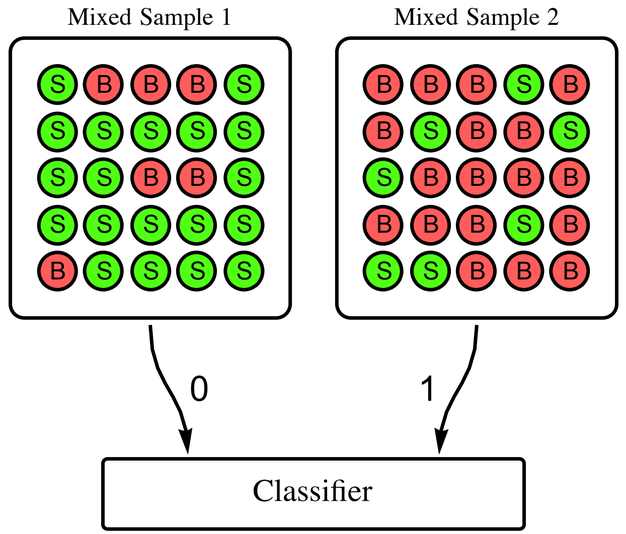
Can We Be More Realistic?

- 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 the likelihoods

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

  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

- An important feature of CWoLa is that, unlike the learning from label proportions (LLP) weak supervision, the label proportions f_1 and f_2 are **not required** for training as long as they are **different**.
- This theorem only guarantees that the optimal classifier from CWoLa, if reached, is the **same** as the optimal classifier from fully-supervised learning.
- Just like most cases, successful/optimal training for CWoLa also requires **a large amount of samples**.

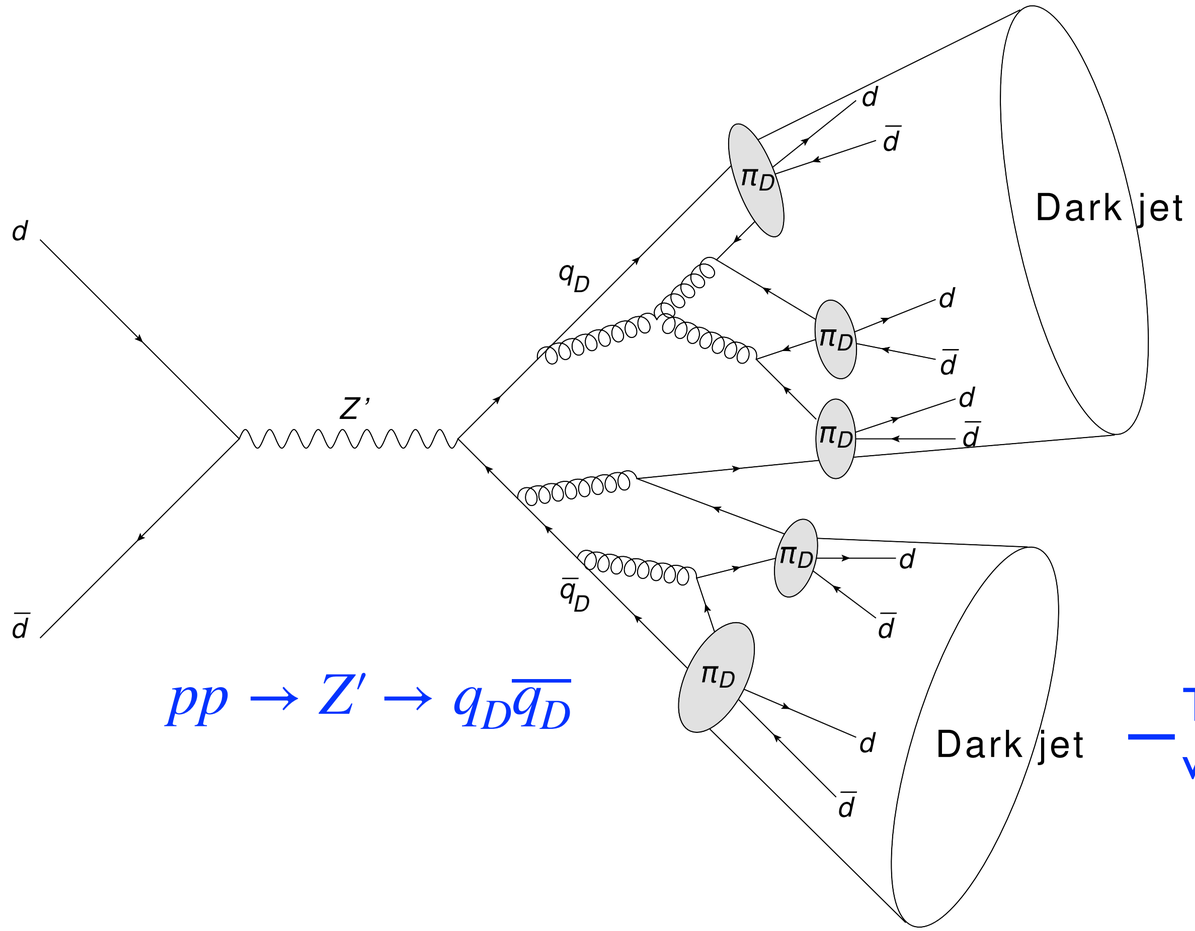
- What happens if available data for the mixed samples are **insufficient or limited**, as is often the case of **real data for BSM searches**?

Outline

- Introduction to deep learning
- Full supervision
- Weak supervision — CWoLa
- Dark valley model — a physical model
- Transfer learning
- Data augmentation
- Summary

Dark Valley Model and Dark Jets

- Assume the existence of a **dark confining sector** that communicates with the visible sector via a **heavy Z' portal**:



$pp \to Z' \to q_D\overline{q_D}$

Courtesy of Hugues Beauchesne

dark quarks

$$\mathcal{L} \supset -Z'_\mu \left(g_q \overline{q_i} \gamma^\mu q_i + g_{q_D} \overline{q_{D\alpha}} \gamma^\mu q_{D\alpha} \right)$$

respective effective coupling constants

— The LHC signature is **a pair of dark jets** with invariant mass consistent with $m_{Z'}$.

Dark Sector Parameter Choices

- We fix the Z' mass at 5.5 TeV and its width at 10 GeV.
- Try **seven dark confining scales** $\Lambda_D \in \{1,\ 5,\ 10,\ 20,\ 30,\ 40,\ 50\}$ GeV.
- Dark vector ρ_D and pseudoscalar π_D masses and **two decay scenarios**:

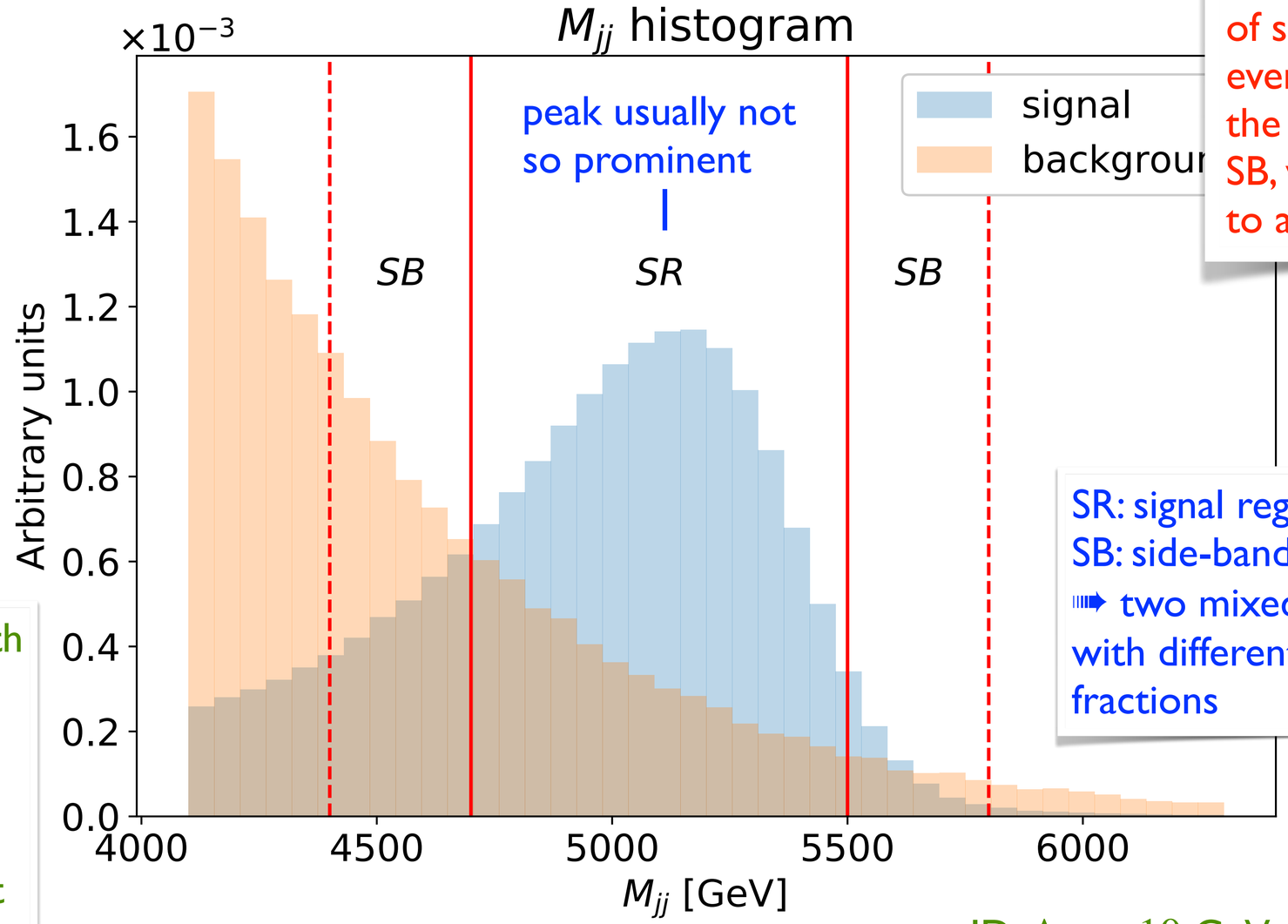
$$\frac{m_{\rho_D}}{\Lambda_D} = \sqrt{5.76 + 1.5\frac{m_{\pi_D}^2}{\Lambda_D^2}}$$

Albouy et al 2022

  - **Indirect Decay (ID)**: $\rho_D \to \pi_D\pi_D$ followed by $\pi_D \to d\bar{d}$ for $m_{\pi_D}/\Lambda_D = 1.0$
  - **Direct Decay (DD)**: $\rho_D,\ \pi_D \to d\bar{d}$ for $m_{\pi_D}/\Lambda_D = 1.8$

- Totally **14 similar "models"** from different combinations of the above parameters.

Dijet Invariant Mass Distributions



Probability distributions of signal and background events are assumed to be the **same** in both SR and SB, which should be valid to a good approximation.

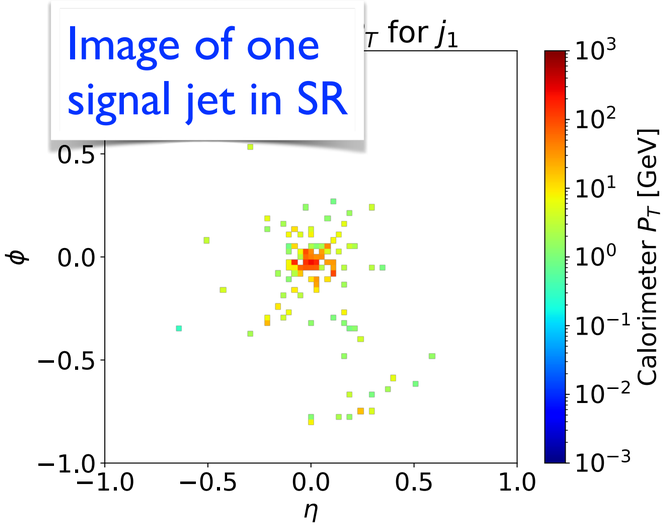
SR: signal region
SB: side-band region
➠ two mixed samples (M_1 and M_2) with different signal/background fractions

- `Madgraph 2.7.3` with PDF = `NN23LO1`
- `Pythia 8.307` with default settings
- `Delphes 3.4.2` with default CMS card and jet radius $R = 0.8$

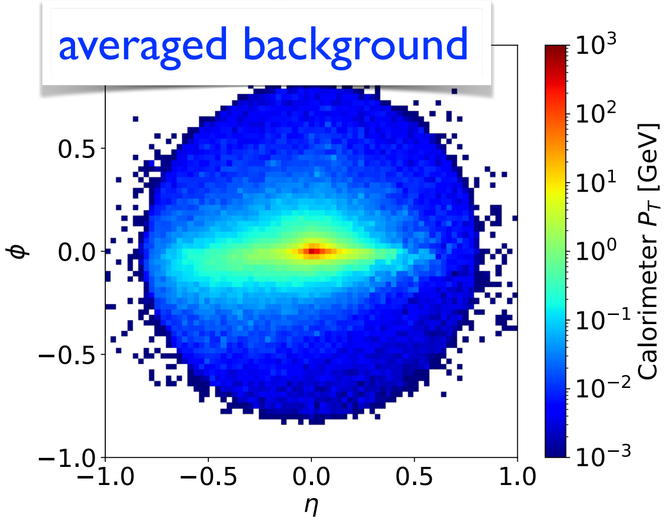
ID; $\Lambda_D = 10$ GeV

Convolutional + Dense Layers

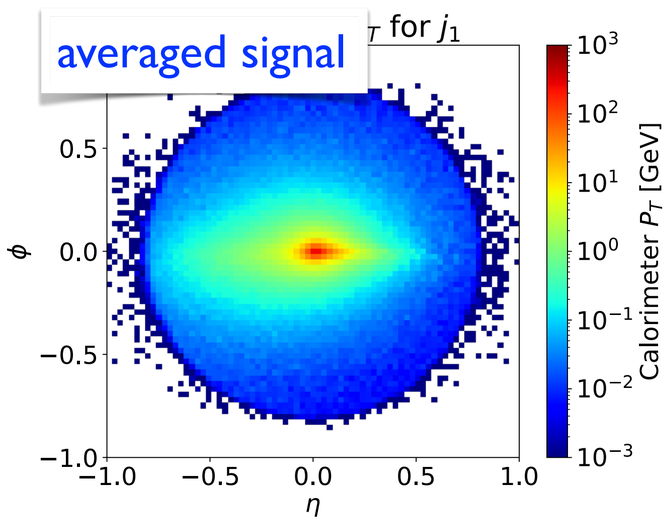
- Prepare each jet image in **three resolutions**: 25×25, 50×50, 75×75.
- Use the **images of the two leading jets** as input data.
- Pass each image through a **common** CNN*, and each returns a score $\in [0,1]$.
- Take the **product** of these two scores as the output of the full NN.



(b) After preprocessing.



(c) Average histogram of background.



(d) Average histogram of signal.

$\Lambda_D = 10$ GeV
Resolution = 75×75

* All NNs are implemented using `Keras` with `TensorFlow` backend. Also, using two distinct networks for the two jets would give slightly inferior results, possibly caused by the lack of signal.

Convolutional + Dense Layers

- The convolutional part of the NN is referred to as the **feature extractor**, and its weights and biases are collectively labeled as Θ.
  ➠ to be **transferred** later
- The dense layer part of the NN is referred to as the **classifier**, and its weights and biases of the dense layers are collectively labeled as θ.
  ➠ to be **fine-tuned** later

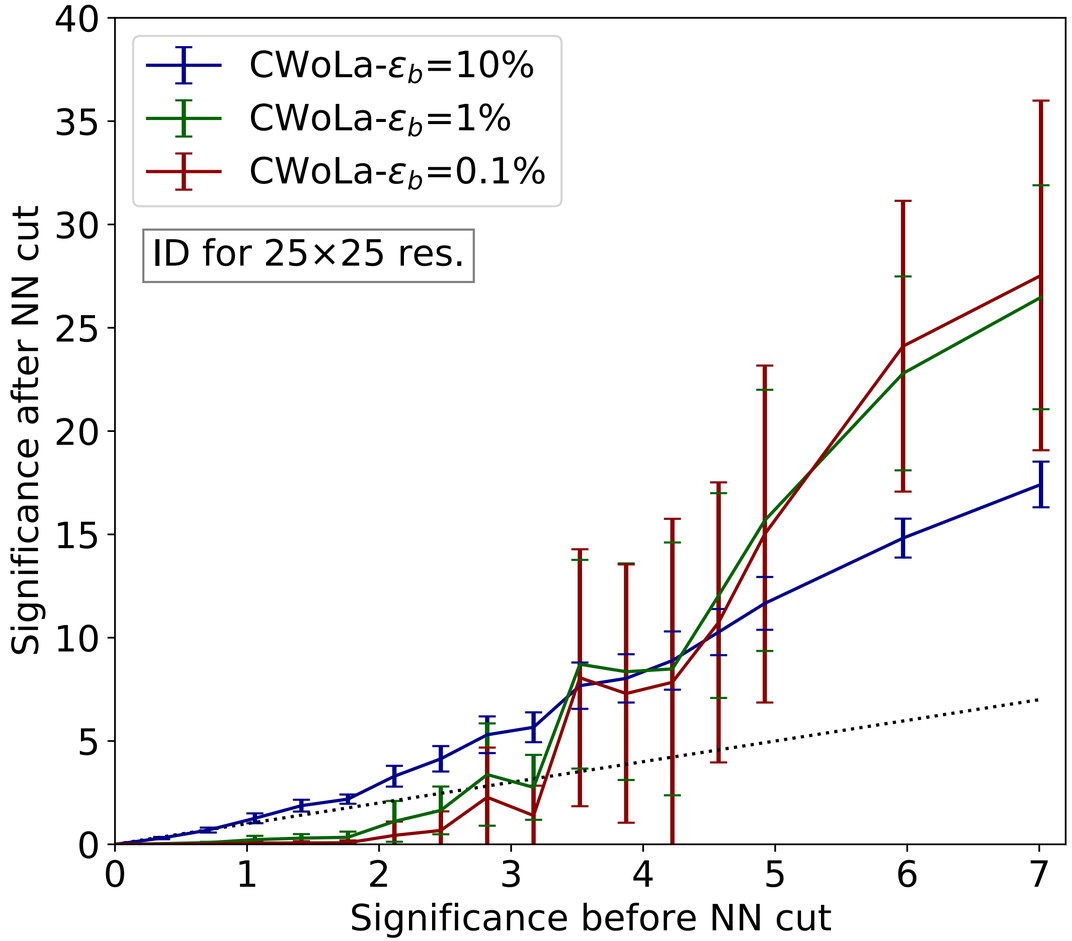
Layers of CNN subnetwork	$\begin{pmatrix}\text{convolutional 2D layer: 64 filters with } 5\times 5 \text{ kernel size}\\ \text{maxpooling layer: } 2\times 2 \text{ pool size}\end{pmatrix}\times 2$ convolutional 2D layer: 128 filters with 3×3 kernel size maxpooling layer: 2×2 pool size convolutional 2D layer: 128 filters with 3×3 kernel size flatten layer (dense layer: 128 units) $\times 3$ dense layer (output): 1 unit

Θ

θ

Results of Regular CWoLa

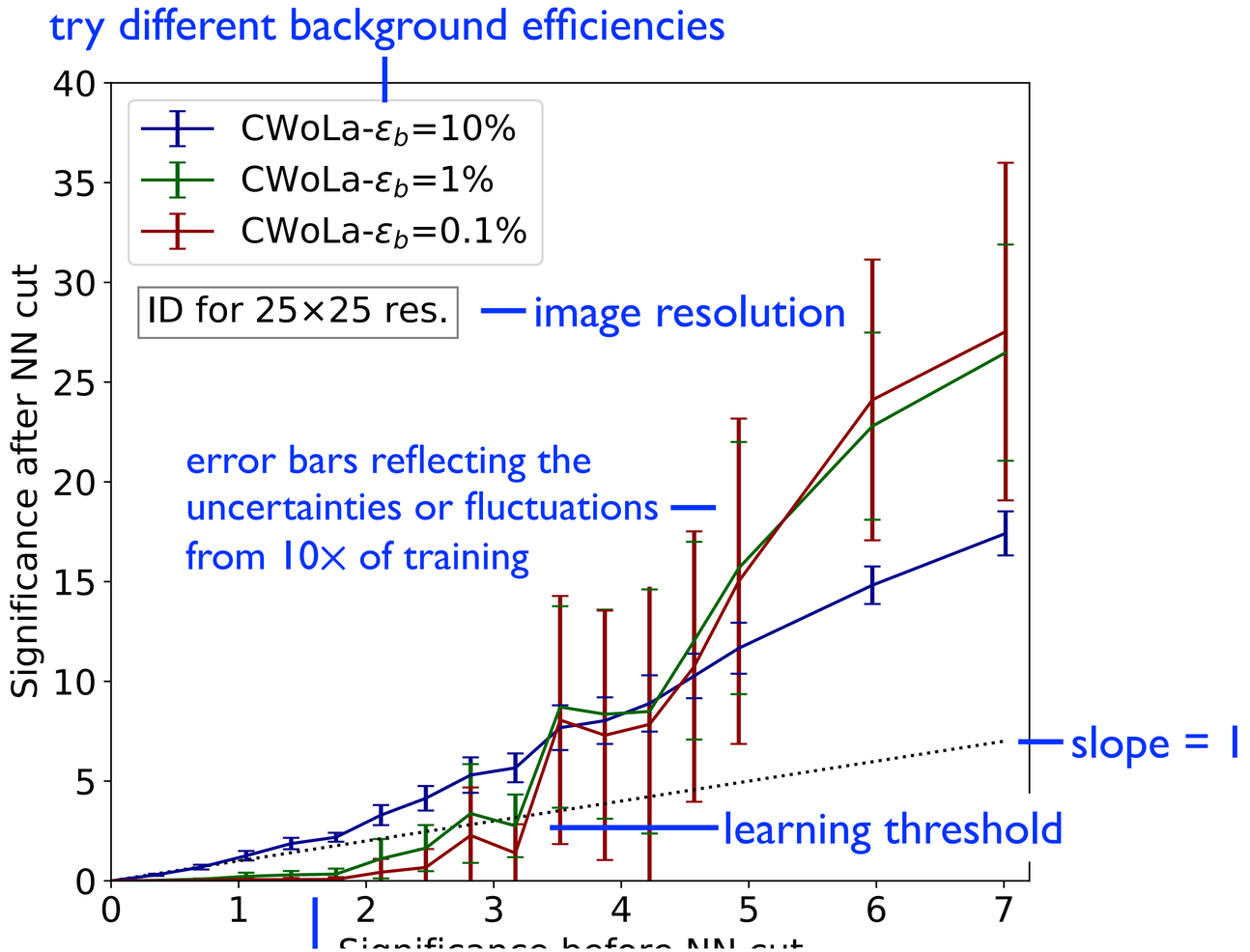
Beauchesne, Chen, CWC 2024



ID; $\Lambda_D = 10$ GeV

Results of Regular CWoLa

Beauchesne, Chen, CWC 2024



below learning thresholds, NN fails to learn from data as it cuts background and signal indiscriminately

Outline

- Introduction to deep learning
- Full supervision
- Weak supervision — CWoLa
- Dark valley model — a physical model
- Transfer learning
- Data augmentation
- Summary

Introduction to Transfer Learning

- The phrase "**transfer learning (TL)**" comes from **psychology**.
  ➠ a learner new to a fresh topic (e.g., riding a motorcycle or playing guitar) typically has a higher learning threshold, while a learner experienced in related topics (e.g., riding a bicycle or playing violin) usually has less difficulty in quickly picking it up
- As an ML technique, TL reuses a **pre-trained model** developed for one task as the starting point of a new model for a new task.
  ➠ transferring knowledge or experience extracted in the pre-trained model for a **source task/domain** to a new model for a **target task/domain**
  ➠ weights from the pre-trained model used to initialize those of the new model
- TL would only be successful when the features learned from the first model trained on its task can be **generalized** and **transferred** to the second task.
  ➠ dataset in the second training should be **sufficiently similar** to those in the first training

Transfer Learning by Pre-training and Fine-tuning

- **Step 1**: The NN is first trained to distinguish a sample of **pure background** from a **pure combination of different signals**, which includes all the models mentioned before, except the benchmark model to be tested.
  - ➠ **pre-training** on a large set of simulations as the **source data**
  - ➠ 200k S and 200k B events in the SR for training
    + 50k S and 50k B events for validation
  - ➠ training both Θ (from convolutional layers) and θ (from dense layers)

Layers of CNN subnetwork	$\begin{pmatrix}\text{convolutional 2D layer: 64 filters with } 5\times 5 \text{ kernel size}\\ \text{maxpooling layer: } 2\times 2 \text{ pool size}\end{pmatrix}\times 2$ convolutional 2D layer: 128 filters with 3×3 kernel size maxpooling layer: 2×2 pool size convolutional 2D layer: 128 filters with 3×3 kernel size flatten layer (dense layer: 128 units) $\times 3$ dense layer (output): 1 unit

Θ

θ

Transfer Learning by Pre-training and Fine-tuning

- **Step 2**: The NN is then trained to distinguish the mixed samples (i.e., the SR and SB regions) using the **actual** data of the benchmark signal (of the true model) plus the SM background.
  - ➠ **fine-tuning** on the small set of actual data as **target data**
  - ➠ freezing Θ in the convolutional layers and reinitializing and training θ in the dense layers
  - ➠ fixing the feature extraction part while training the classification part

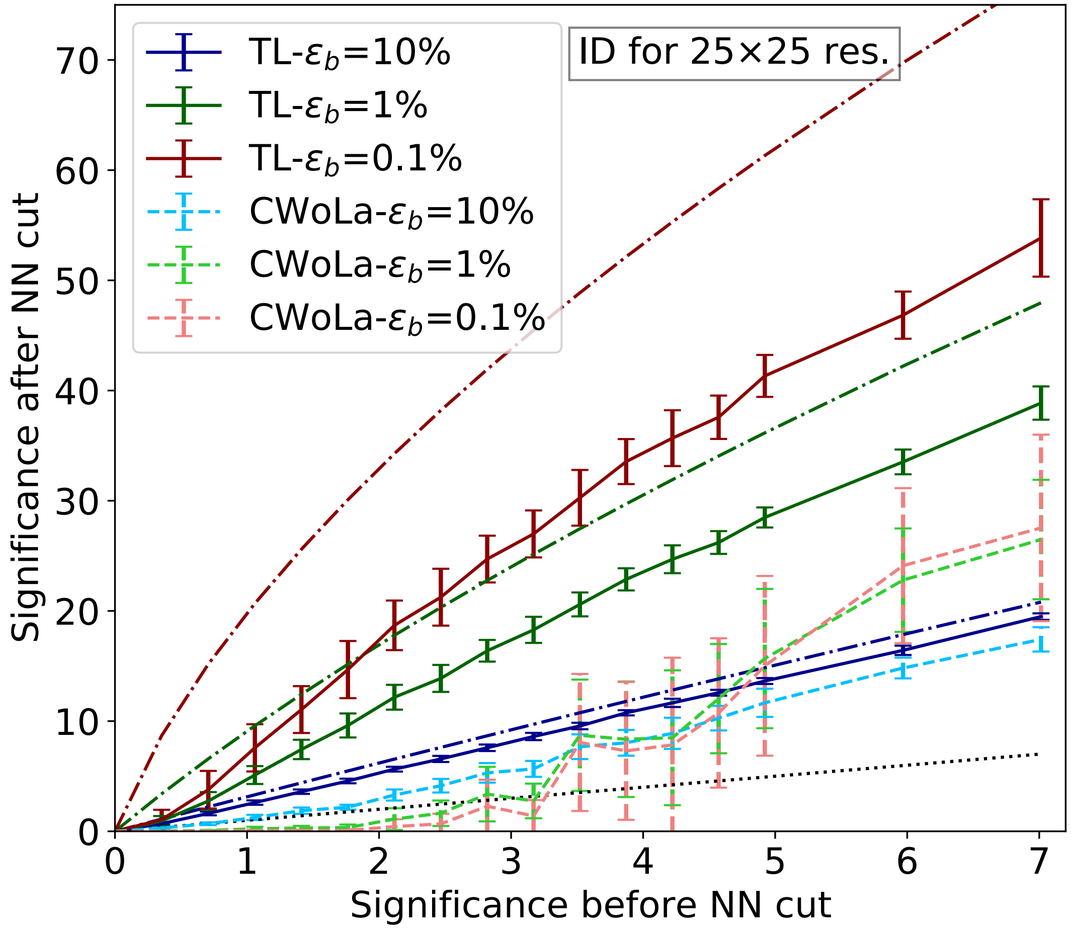
Layers of CNN subnetwork	$\begin{pmatrix}\text{convolutional 2D layer: 64 filters with } 5\times 5 \text{ kernel size}\\ \text{maxpooling layer: } 2\times 2 \text{ pool size}\end{pmatrix}\times 2$ convolutional 2D layer: 128 filters with 3×3 kernel size maxpooling layer: 2×2 pool size convolutional 2D layer: 128 filters with 3×3 kernel size flatten layer (dense layer: 128 units) $\times 3$ dense layer (output): 1 unit

Θ

θ

Transfer Learning vs Regular CWoLa

Beauchesne, Chen, CWC 2024

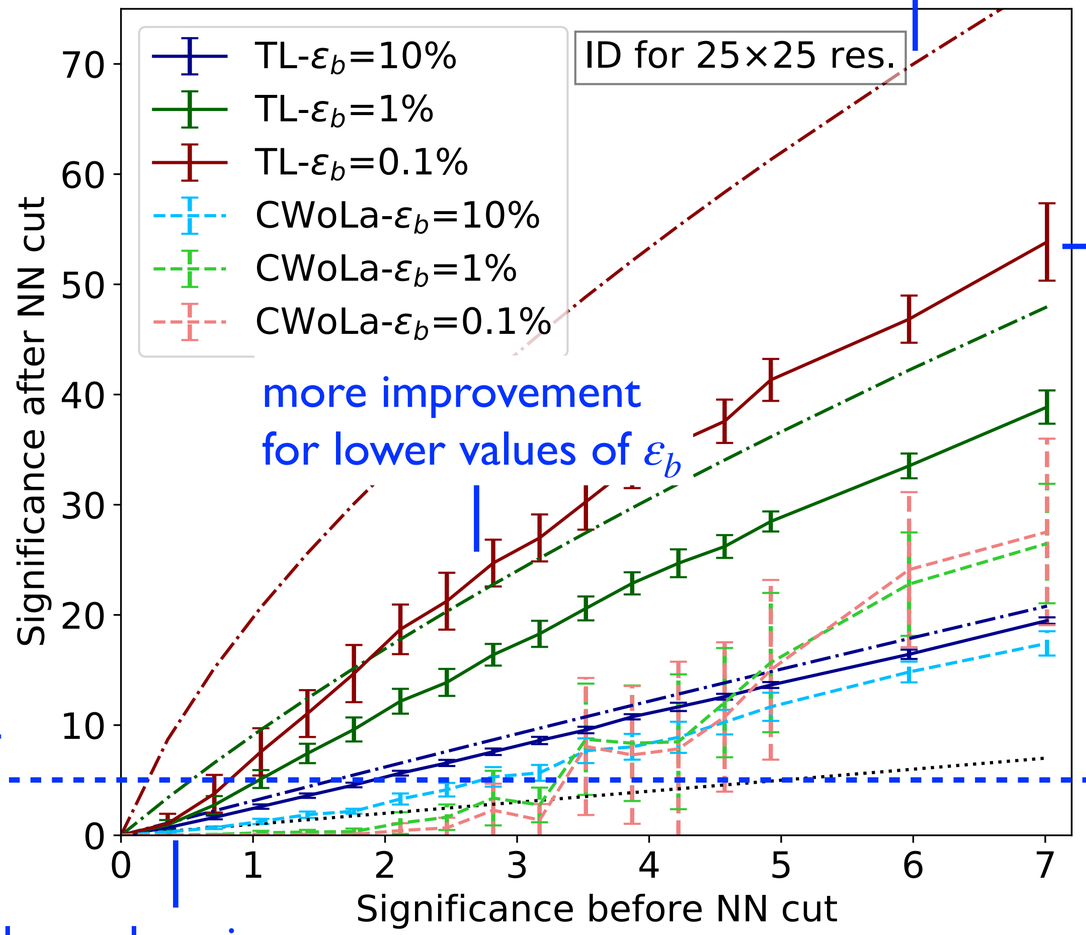


ID; $\Lambda_D = 10$ GeV

Transfer Learning vs Regular CWoLa

Beauchesne, Chen, CWC 2024

dash-dotted curves for full supervision



stabler solid curves and smaller fluctuations than dashed curves

more improvement for lower values of ϵ_b

amount of signal for a 5σ discovery reduced by a factor of a few, due to the fact that NN can better reject backgrounds

lower learning thresholds for TL

ID; $\Lambda_D = 10$ GeV

Outline

- Introduction to deep learning
- Full supervision
- Weak supervision — CWoLa
- Dark valley model — a physical model
- Transfer learning
- Data augmentation
- Summary

Augmentation Methods

- 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 augmentations that capture the **symmetries** of the physical events and the experimental **resolution** or statistical **fluctuations** in the detector, we implement three methods*:
  - p_{T} **(transverse momentum) smearing**;
  - **jet rotation**; and
  - a **combination** of the two.

- Additionally, we have applied $\eta - \phi$ **smearing** and **Gaussian noise** to jet images and observed essentially no improvement.

p_{T} Smearing and Jet Rotation

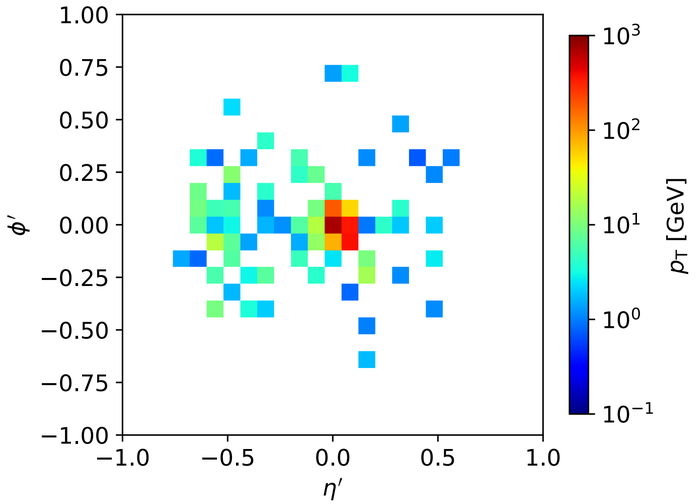
- The p_{T} **smearing method** is used to simulate **detector effects** in **resolution/ response fluctuation** effects on the transverse momentum of jet constituents by resampling the p_{T} of jet constituents according to the **normal distribution**:

$$p'_{\mathrm{T}} \sim \mathcal{N}\left(p_{\mathrm{T}}, f\left(p_{\mathrm{T}}\right)\right), \quad f\left(p_{\mathrm{T}}\right) = \sqrt{0.052 p_{\mathrm{T}}^2 + 1.502 p_{\mathrm{T}}}$$

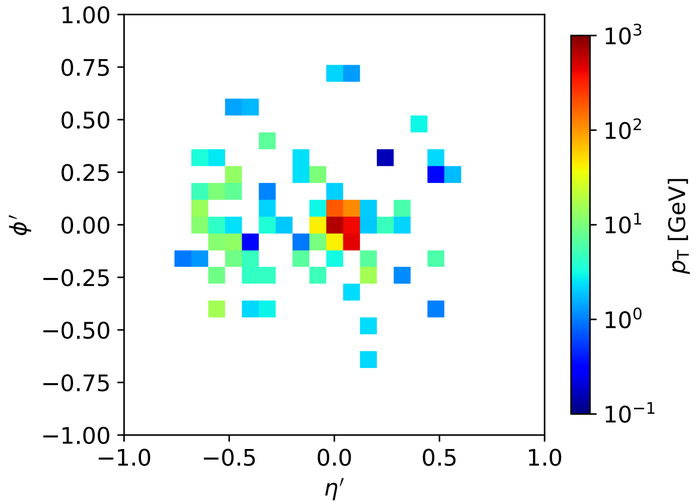
  where $f(p_{\mathrm{T}})$ is the **energy smearing function** applied by Delphes.

- The **jet rotation method** rotates each jet with respect to its center by a random angle $\theta \in [-\pi, \pi]$ to enlarge the **diversity** of training datasets.
- Other angle ranges are also studied and the training performance is found to **improve with the range** of rotation angles.

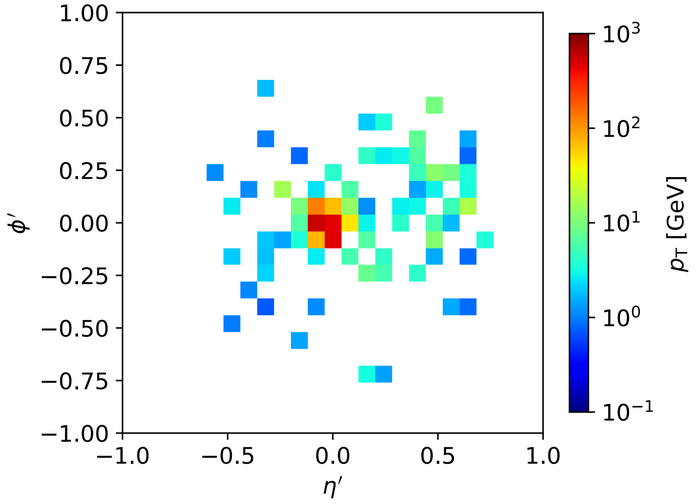
Example of A Jet Image



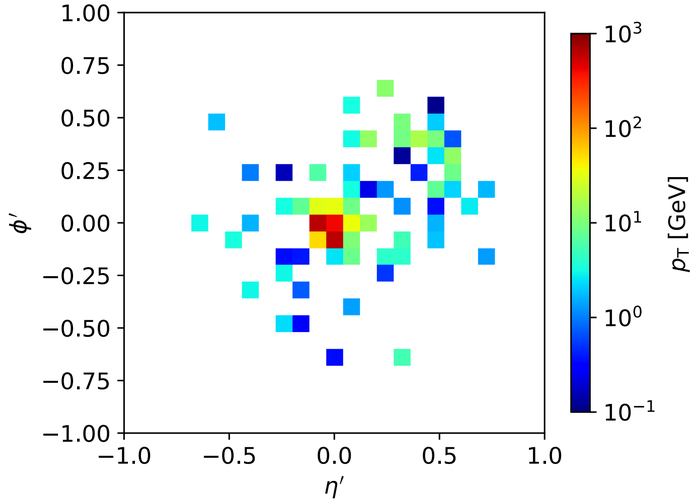
(a) Original jet image



(b) p_{T} smearing

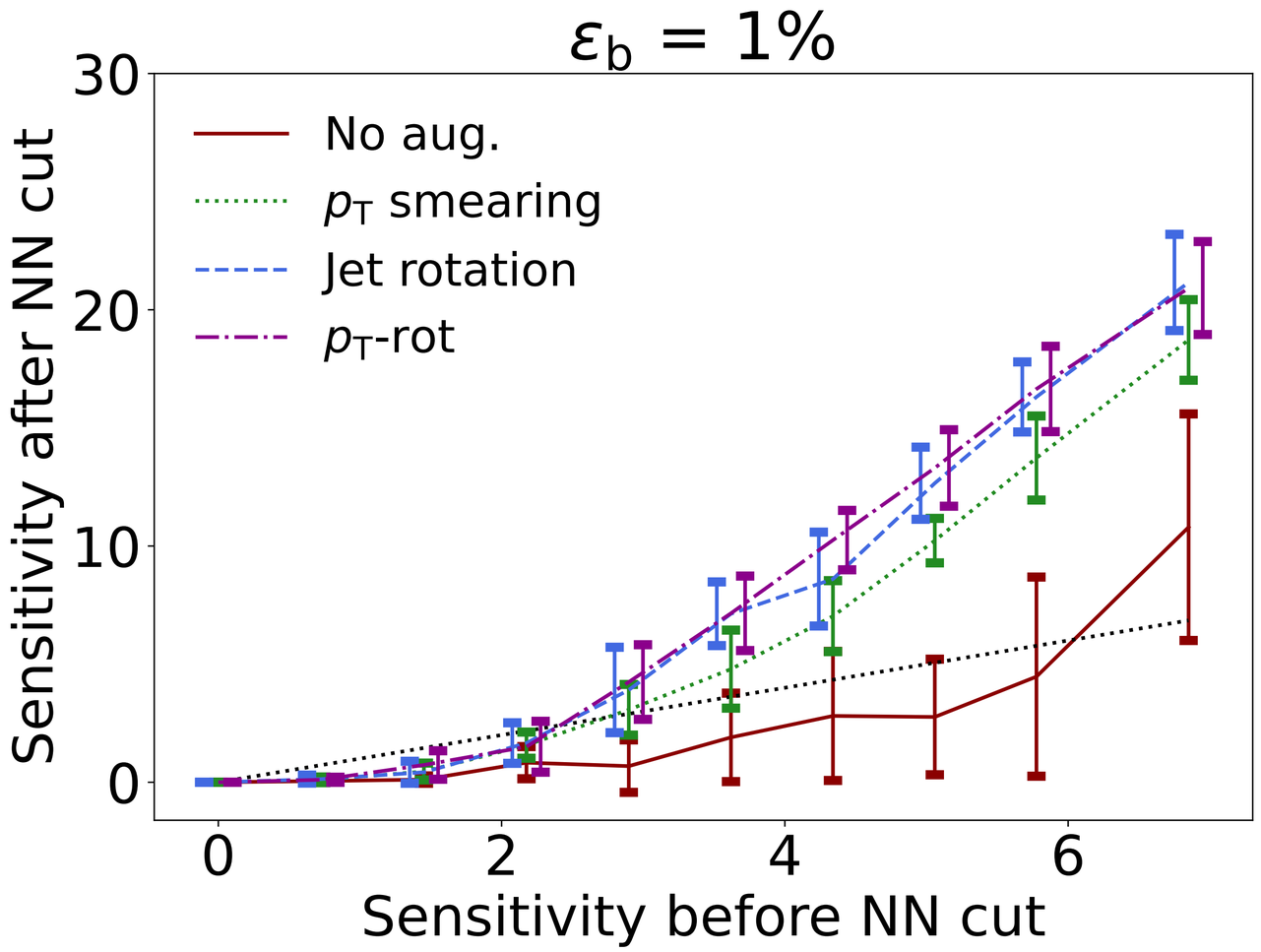


(c) Jet rotation



(d) p_{T} smearing + jet rotation

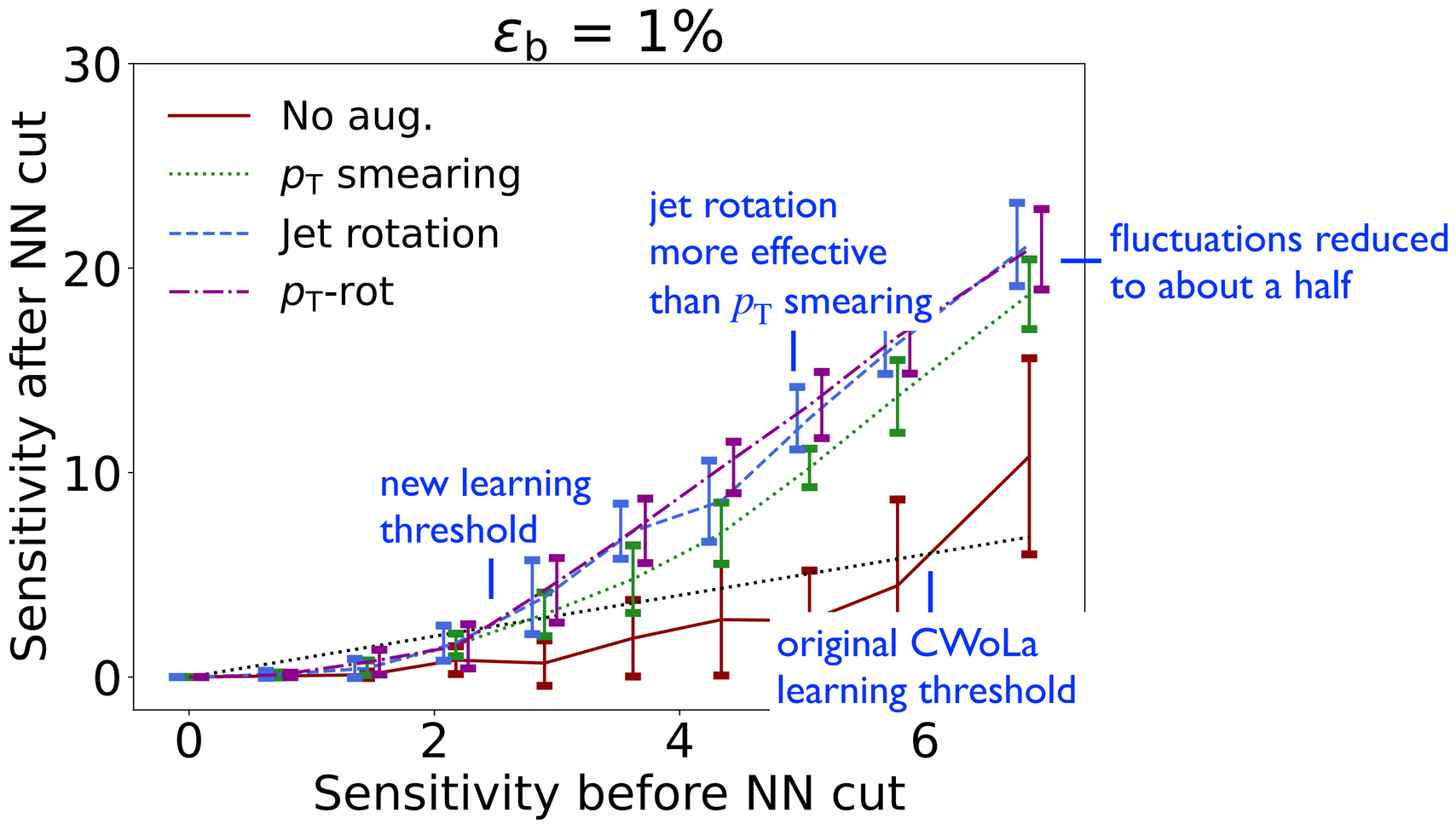
Sensitivity Improvement



ID; $\Lambda_D = 10$ GeV

Chen, CWC, Hsieh 2024

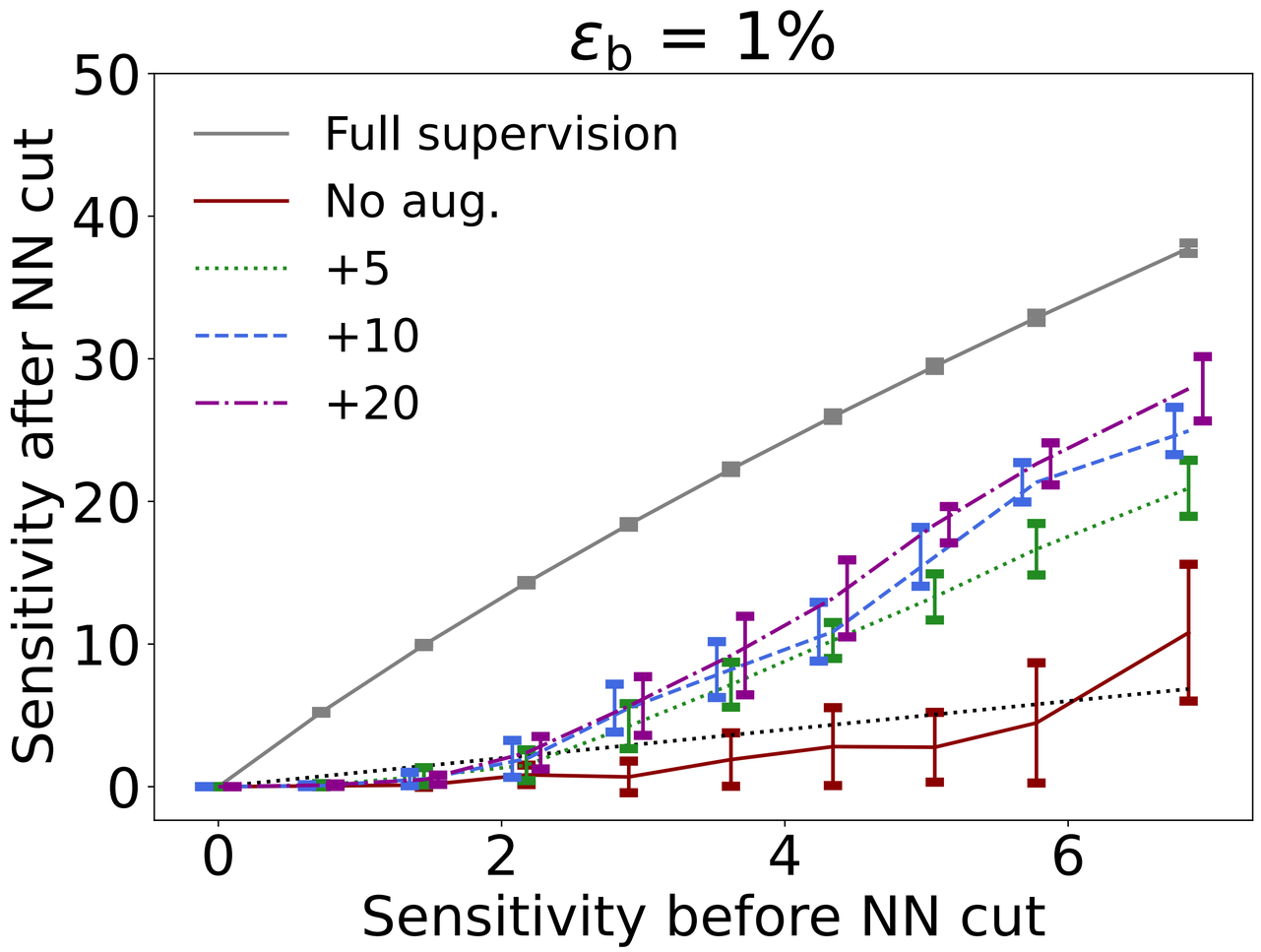
Sensitivity Improvement



ID; $\Lambda_D = 10$ GeV

Chen, CWC, Hsieh 2024

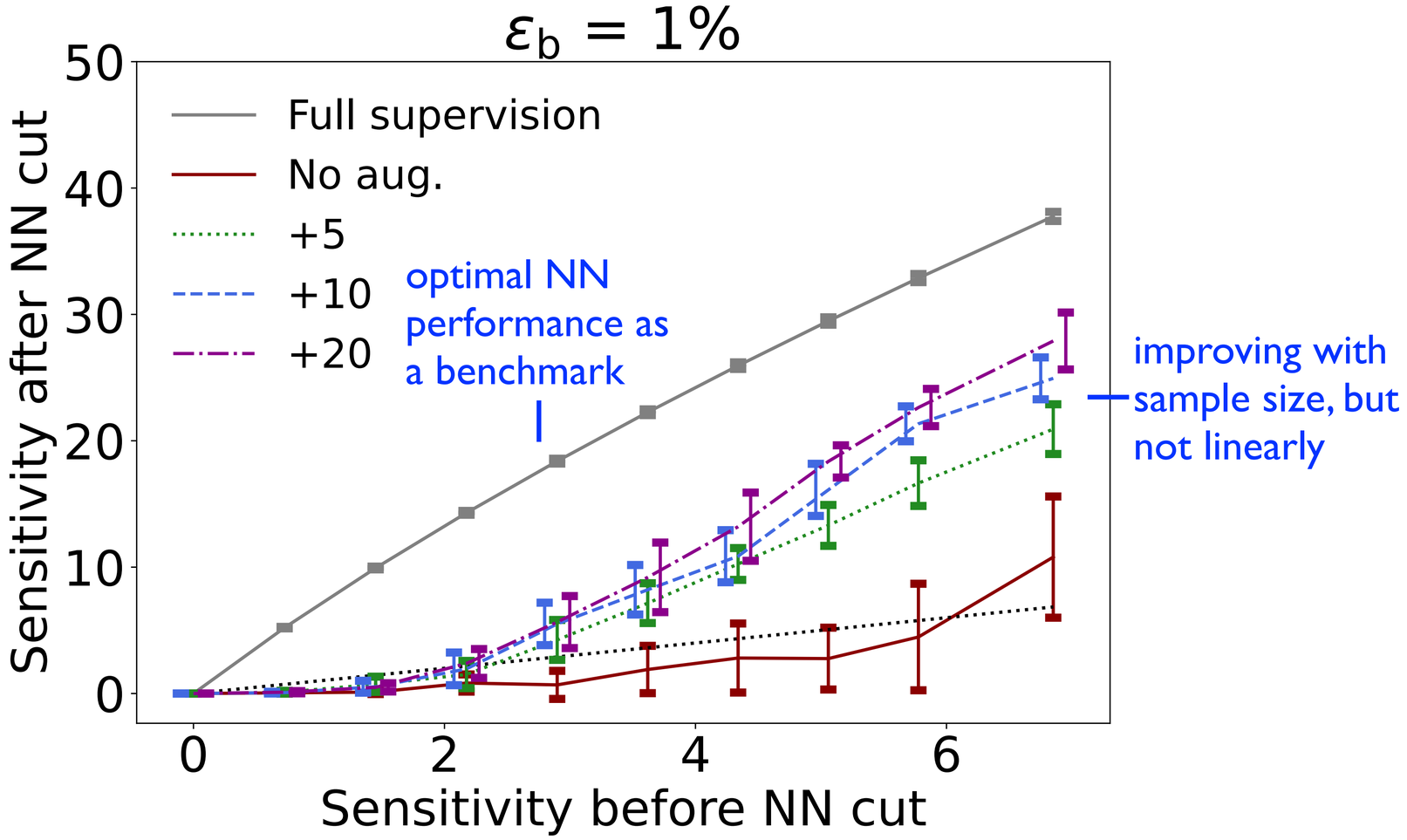
Dependence on Augmentation Size



ID; $\Lambda_D = 10$ GeV

Chen, CWC, Hsieh 2024

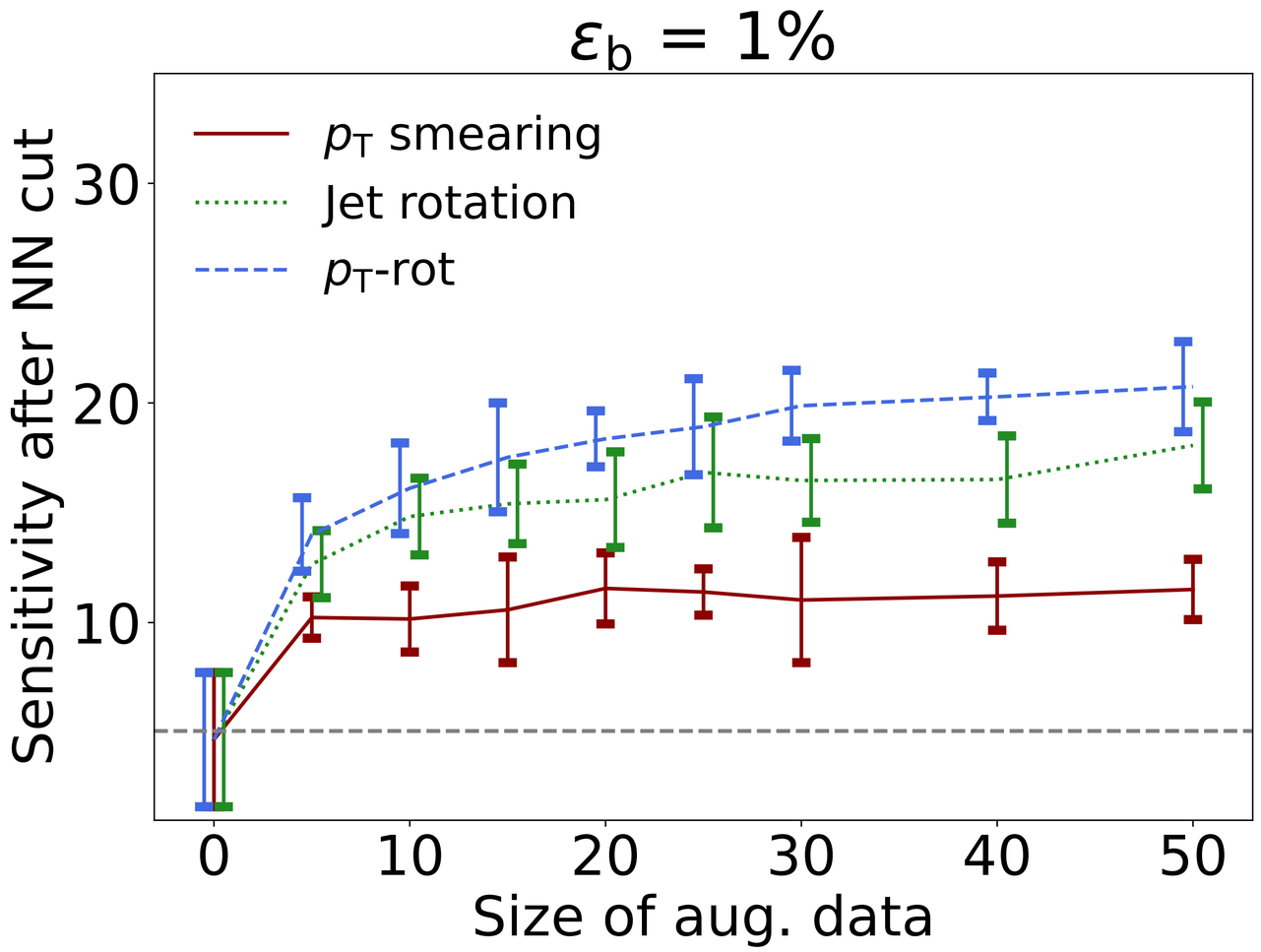
Dependence on Augmentation Size



ID; $\Lambda_D = 10$ GeV

Chen, CWC, Hsieh 2024

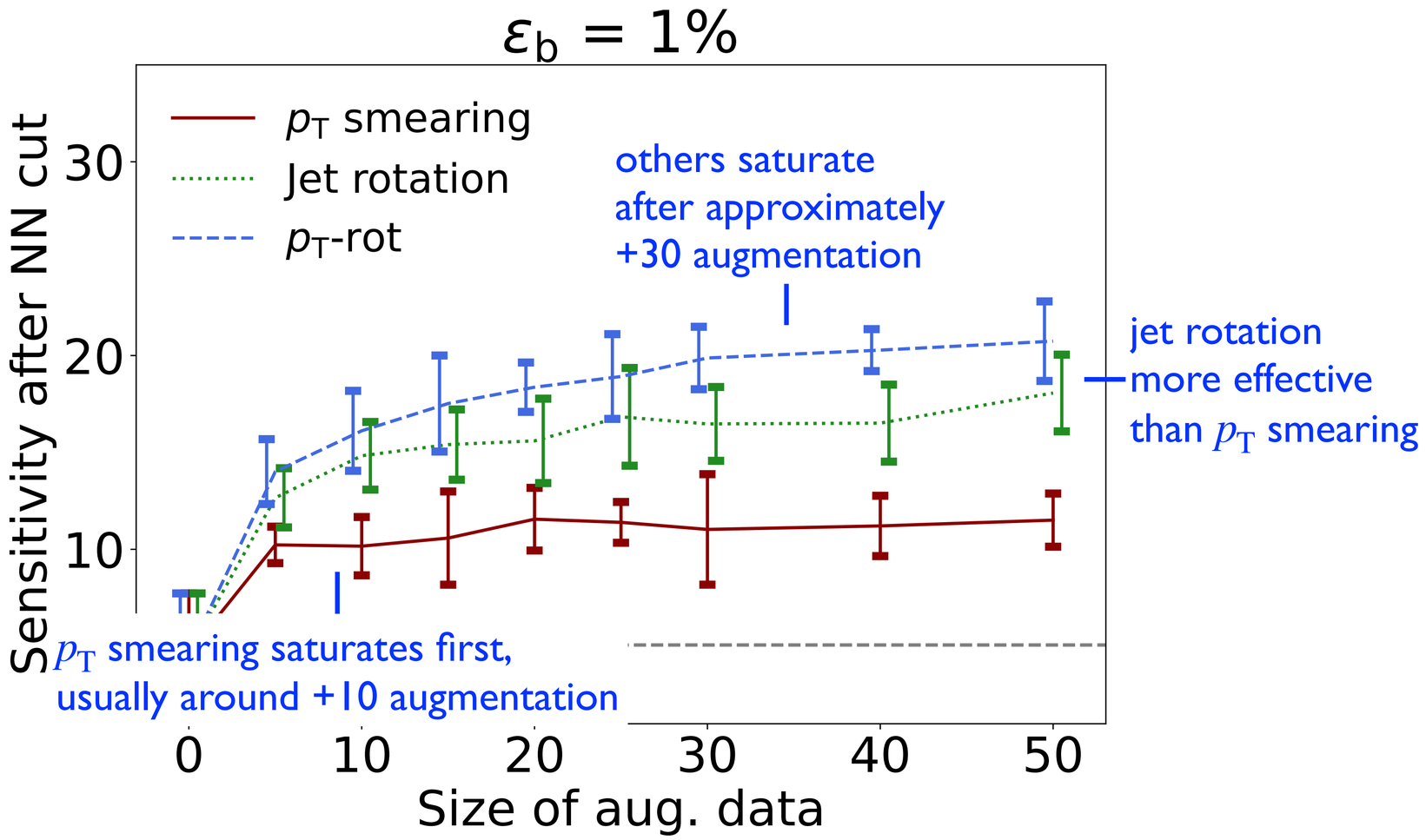
Asymptotic Behavior of Augmentation Size



ID; $\Lambda_D = 10$ GeV

Chen, CWC, Hsieh 2024

Asymptotic Behavior of Augmentation Size



ID; $\Lambda_D = 10$ GeV

Chen, CWC, Hsieh 2024

Summary

- **Deep machine learning** in particle physics has become an unstoppable trend and surpassed traditional data analysis methods.
  ➠ new tools for us to explore the Universe
- **Weak supervision** (CWoLa) is an advantageous techniqe being able to **train on real data** and exploiting distinctive signal properties.
  ➠ ideal tools for **anomaly searches** but fail when signals are **limited**
- We propose to employ the **transfer learning** (TL) technique and show that it can **drastically improve** the performance of CWoLa searches, particularly in the **low-significance region** (because of better identification to exclude background).
- We also propose to employ the **data augmentation** technique and show that jet rotation is more effective than p_{T} smearing, that a **+5 augmentation** can already achieve great results.

Thank You!