A study of foundation models for event classification in collider physics

Tomoe Kishimoto

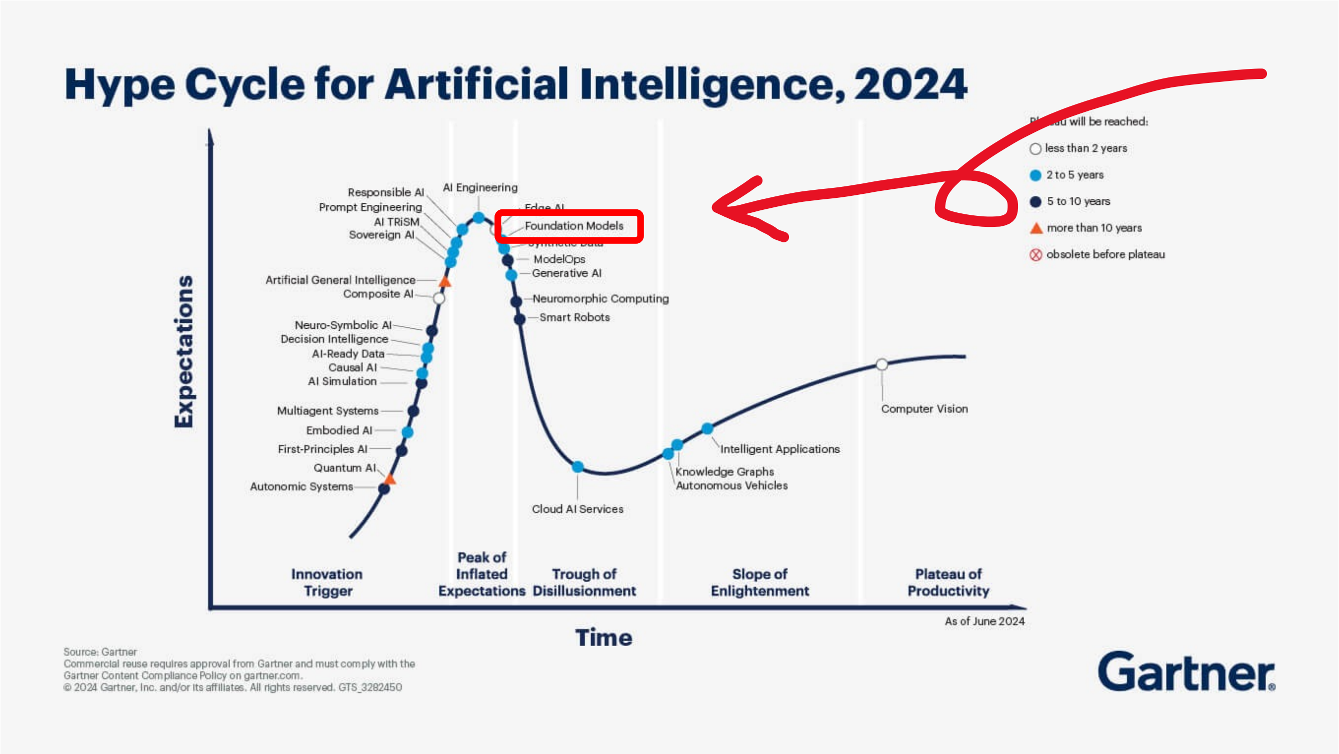
Computing Research Center, KEK

tomoe.kishimoto@kek.jp

加速器だから見える世界。

KEK

Introduction



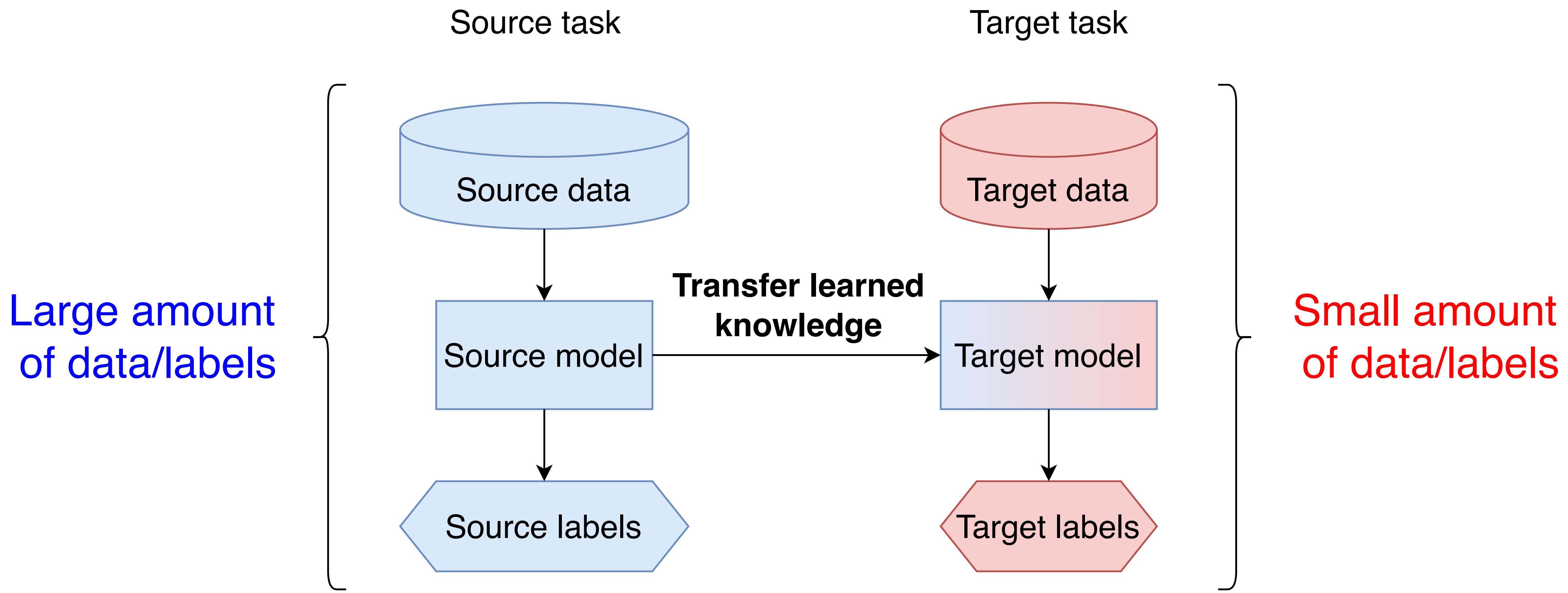
Gartner.com

- "Foundation models" is one of the keywords for AI
  - Pre-training using a large amount of "unlabeled" data
  - Fine-tuning for a target application (transfer learning)

→ Q: Is the concept of foundation models beneficial to collider physics

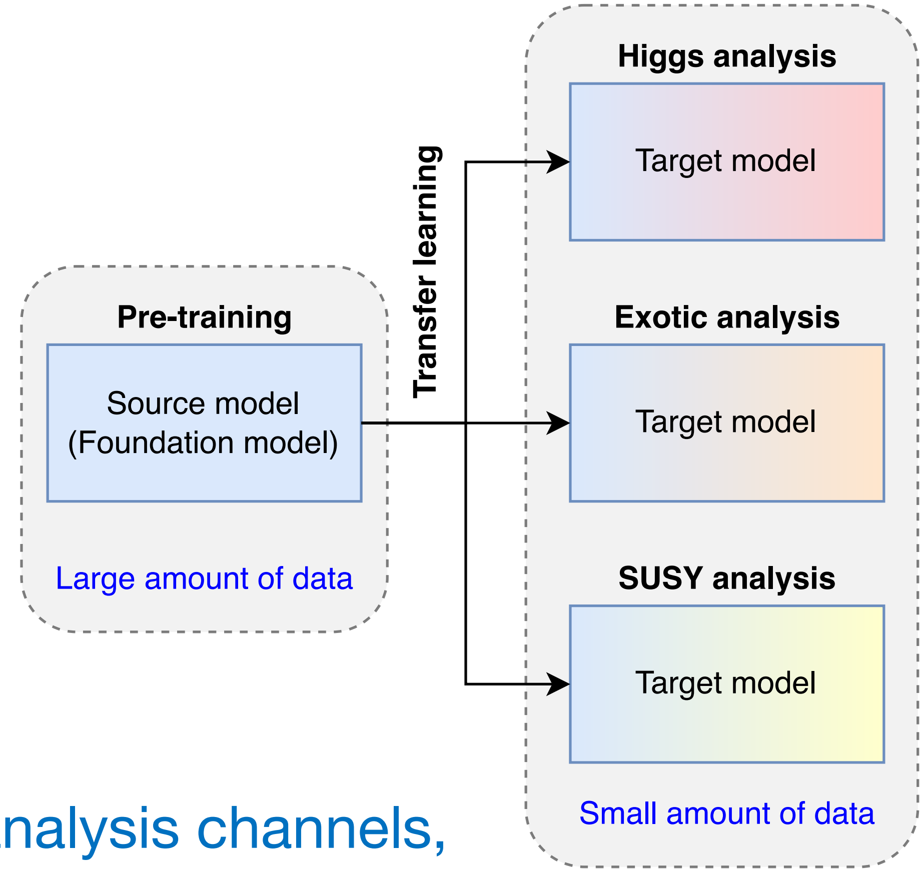
加速器だから見える世界。

Transfer learning



Use case of physics analysis

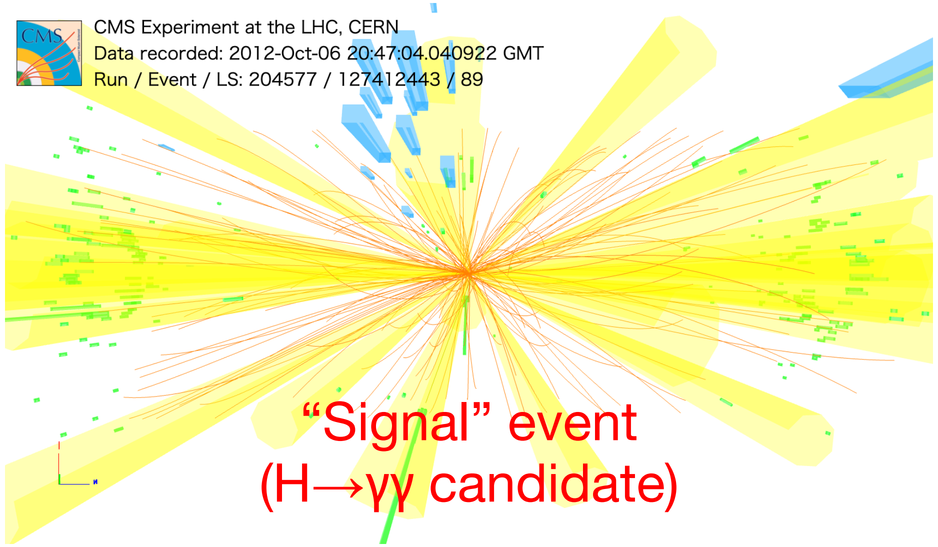
- Many analysis channels in collider physics
  - Higgs, Exotic, SUSY, etc
  - Currently, dedicated DL models are trained from scratch for each channel
    ← Large amount of training data (MC) for each channel

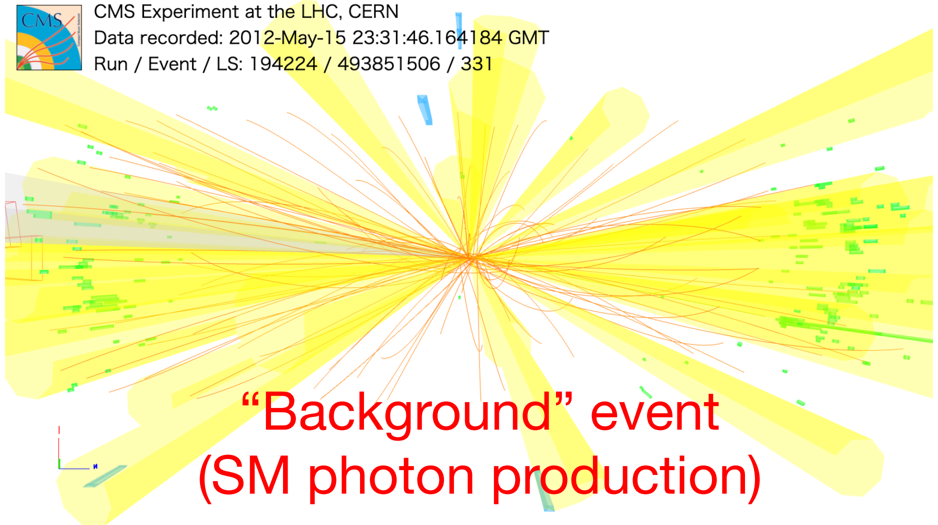


→ If transfer learning can be applied to different analysis channels, computing resources for MC simulations and DL training are saved

Event classification

- The concept is examined using “event classification” problem
  - A typical problem in HEP, signal event vs. background event





→ Reconstructed particles (objects) are the basic information for the classification

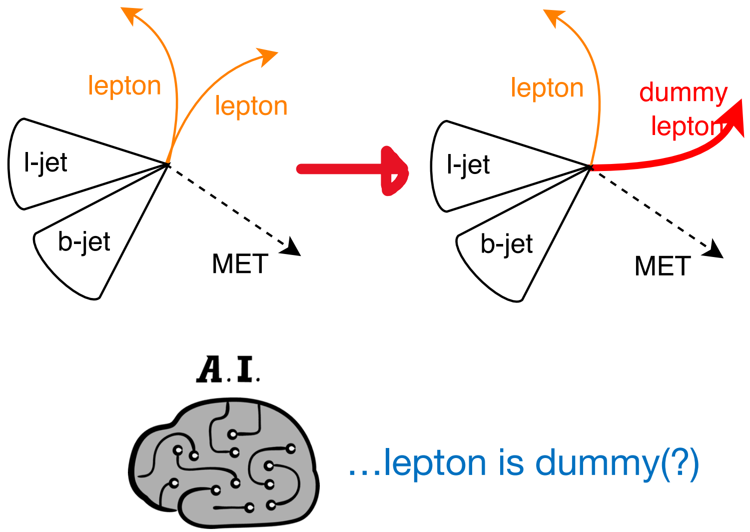
Datasets (CMS Opendata)

		Selections	# of events
Pre-training →	Collision data	lepton ≥ 1 + jets ≥ 2 + bjets ≥ 1	~10^6
Event classification	H^+tb[ref.] vs ttbar+jets	lepton ≥ 1 + jets ≥ 4 + bjets ≥ 1	~10^6
	H^+HW[ref.] vs ttbar+jets	lepton ≥ 1 + tau≥ 1 + jets ≥ 3 + bjets ≥ 1	~10^6
	ttH[ref.] vs ttbar+jets	lepton ≥ 1 + jets ≥ 4 + bjets ≥ 2	~10^6
	ttH[ref.] vs ttbar+jets	lepton ≥ 2 + jets ≥ 2 + bjets ≥ 1	~10^6

- Pre-training is performed using collision data (unlabelled data) based on the foundation model concept
  - ~10^7 events are available after the selection, but only ~10^6 events are used
  - NVIDIA A100: ~10^4 events/sec (10^7 events /10^4 x 500 epochs = 138 hours)

Pre-training strategy

- Only low-level features of each object (4-vector + charge) are used as inputs
- Self-supervised learning is employed to handle the unlabeled collision data
- Strategy:
  - An object (lepton, tau, b-jet, light-jet, or MET) is randomly replaced with a dummy object when preparing a mini-batch
    → DL model is trained to predict what type of object was replaced



Pre-training strategy

Pre-training phase

	pT	η	ϕ	E	charge
lep1	x1	y1	z1	i1	j1
lep2	x0	y0	z0	i0	j0
l-jet1	x3	y3	z3	i3	j3
b-jet1	x4	y4	z4	i4	j4
MET	x5	y5	z5	i5	j5

Dummy object
(from other events)

DL model

What type of objet is dummy ?

→ Random masks increase prediction pattern (data augmentation)

Transfer

Event classification phase

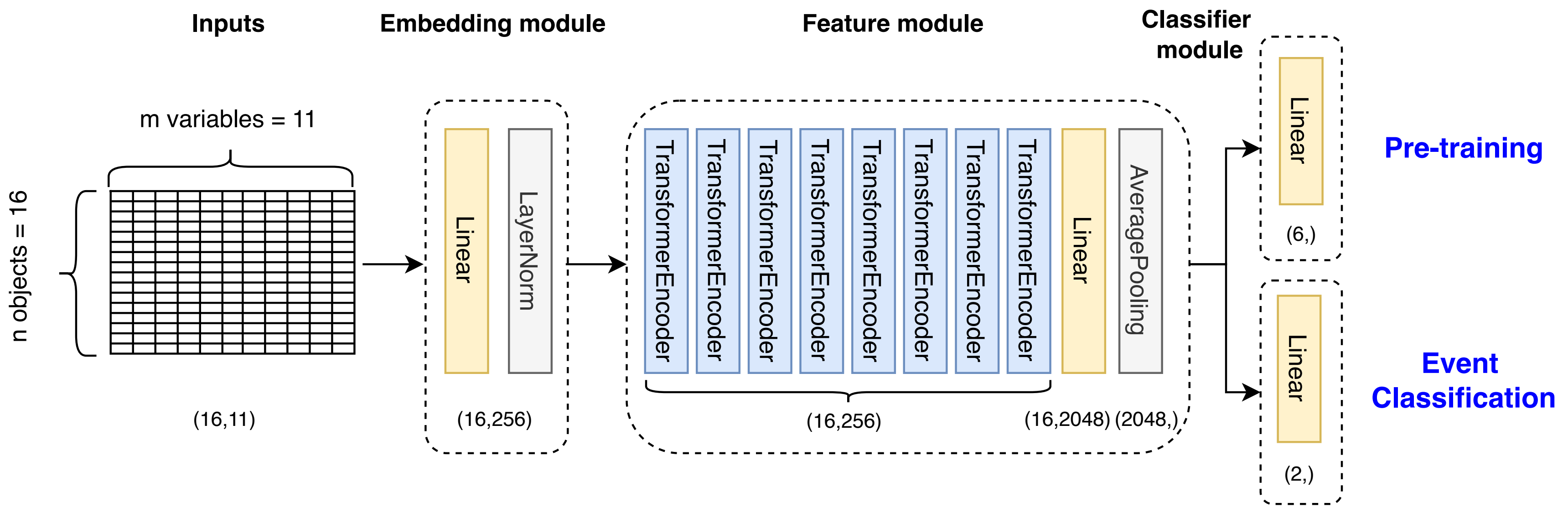
	pT	η	ϕ	E	charge
lep1	x1	y1	z1	i1	j1
lep2	x2	y2	z2	i2	j2
l-jet1	x3	y3	z3	i3	j3
b-jet1	x4	y4	z4	i4	j4
MET	x5	y5	z5	i5	j5

DL model

Signal
vs
Backdgloud

DL model

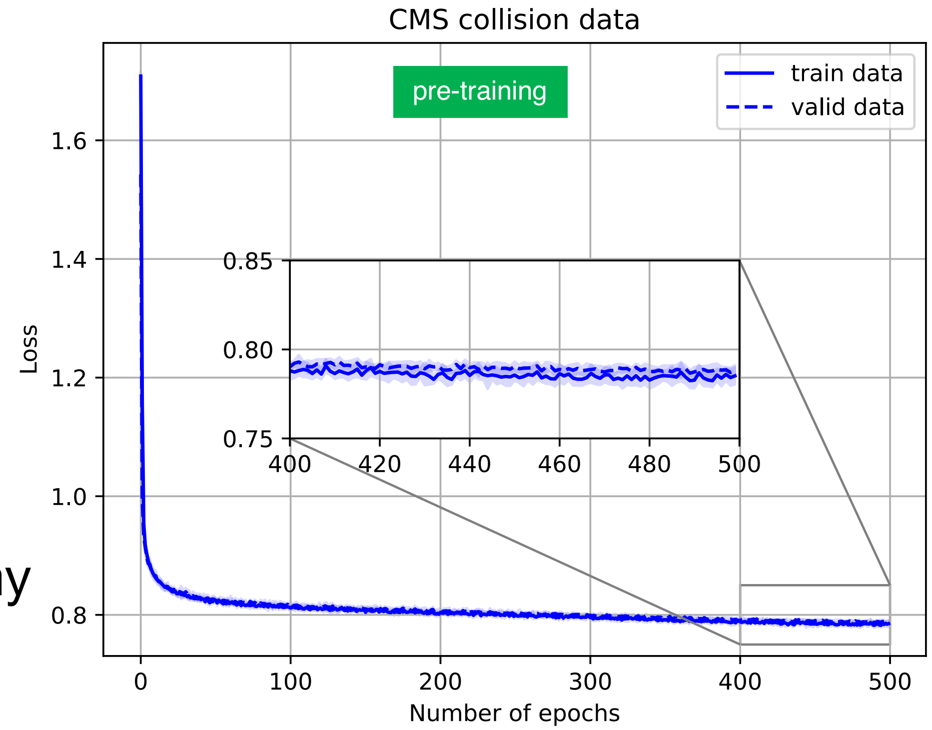
- Transformer encoder is employed:
  - ~11M trainable parameters



→ Weight parameters of embedding and feature modules are transferred and fine-tuned
→ Classifier module is always trained from scratch

Training details

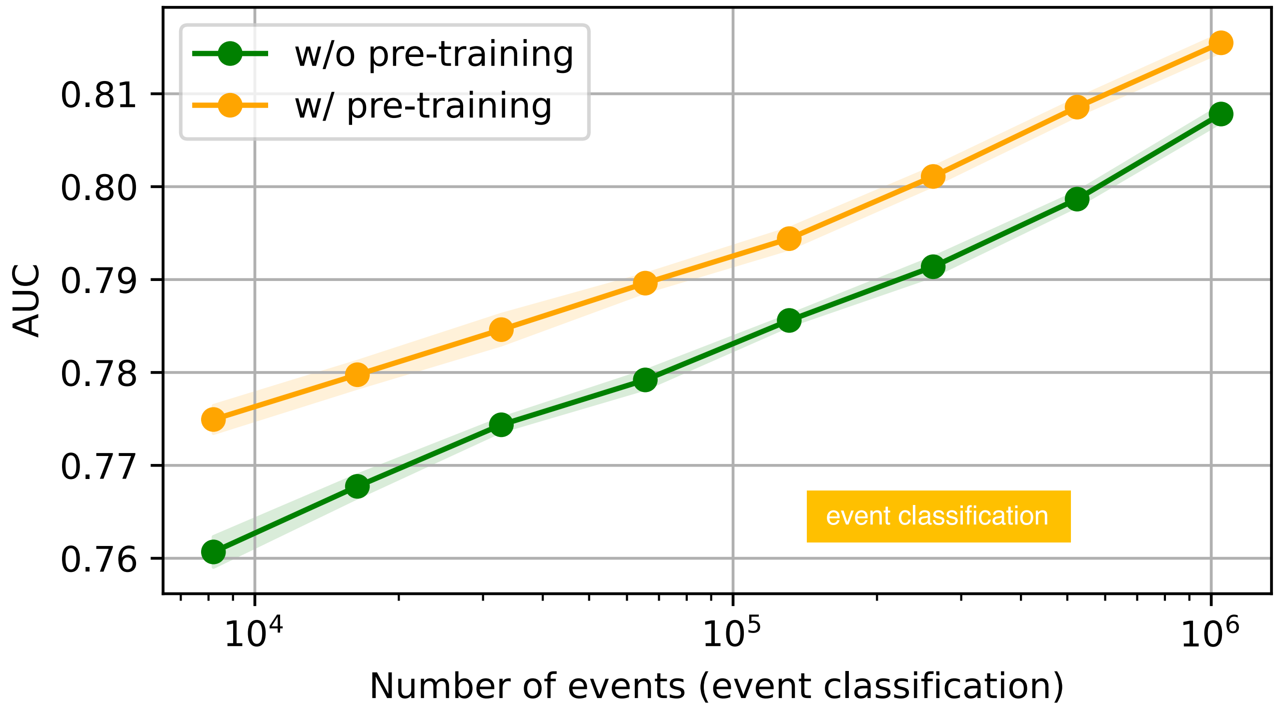
- Basically, the same setting between the pre-training and event classification phases:
  - SGD optimizer:
    - Learning rate: 10^{-2}-10^{-4} (CosineAnnealingLR)
  - Batch size: 512, Epochs: 500
  - Cross entropy loss:
    - Pre-training: lepton, b-jet, l-jet, MET, or No dummy
    - Event classification: signal or background
- NVIDIA A100: ~20 batches/s
  - ~13 hours for one training



~1M events used

AUC of event classification

H⁺tb signal



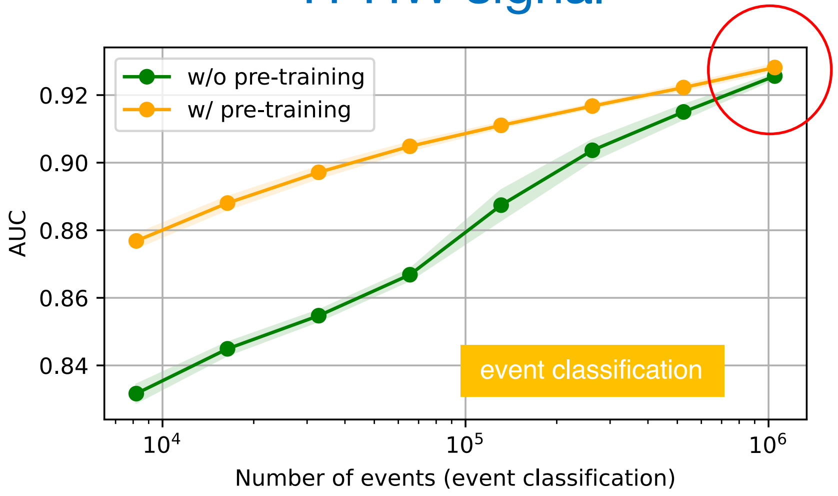
small data

large data

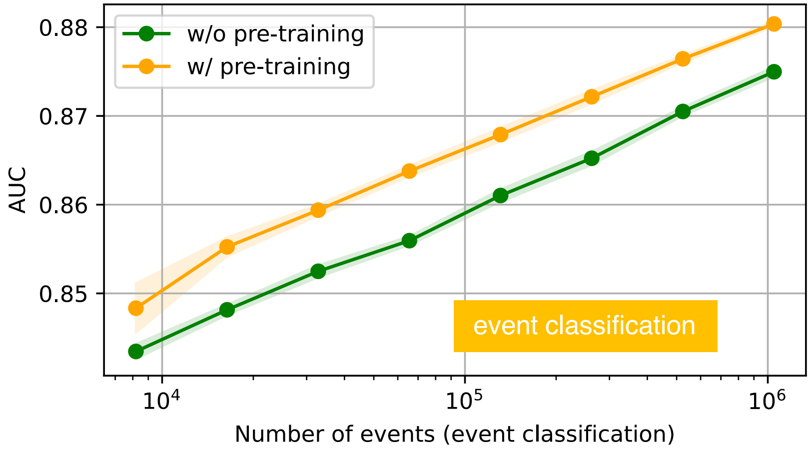
- Significant improvements by introducing the pre-training
- Future work: need to check if the performances converge when more data (>10^6) are added

AUC of event classification

H⁺HW signal



ttH (1lep) signal



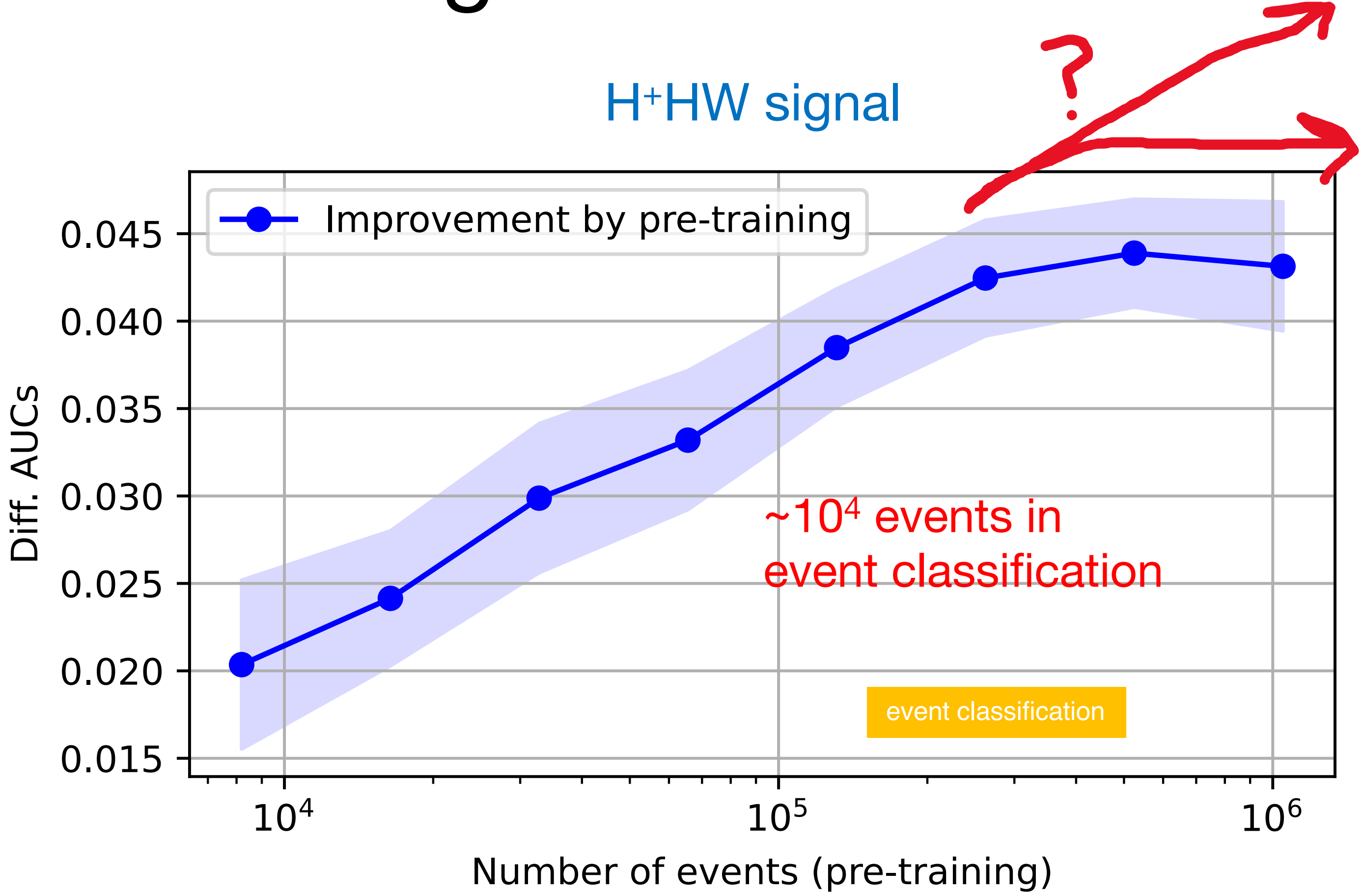
ttH (2lep) signal

➢ The improvements are confirmed for all signal events

→ The pre-trained model (foundation model) is well generalized

Scaling raw

H+HW signal

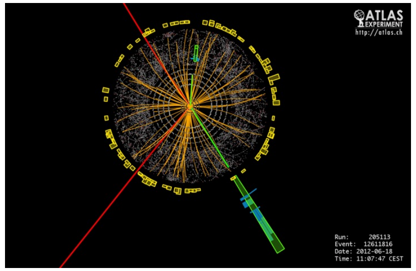


- ➢ The scaling behavior encourages a pre-training with a larger data
  - ➢ However, the number of events in the CMS open data itself and computing resources are limited

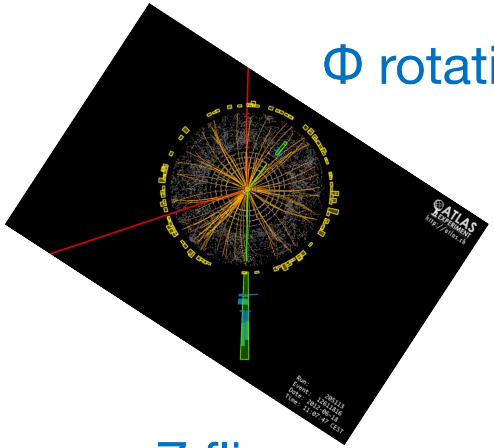
→ Data augmentation is examined

Lorentz transformation

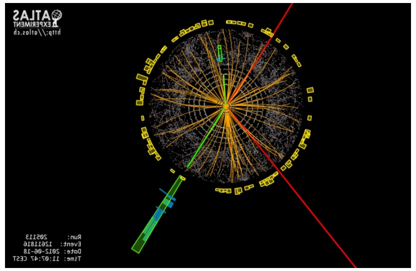
Original event
(Higgs candidate)



Φ rotation



Z flip



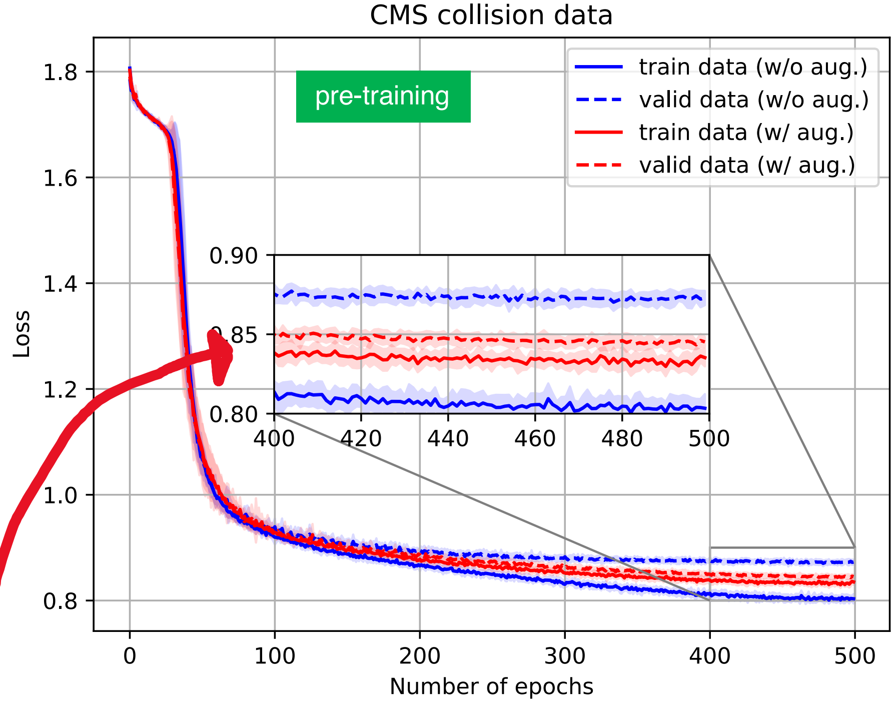
Lorentz boost
(z direction)

← This data is still a Higgs candidate, and should occur with the same probability as the original event

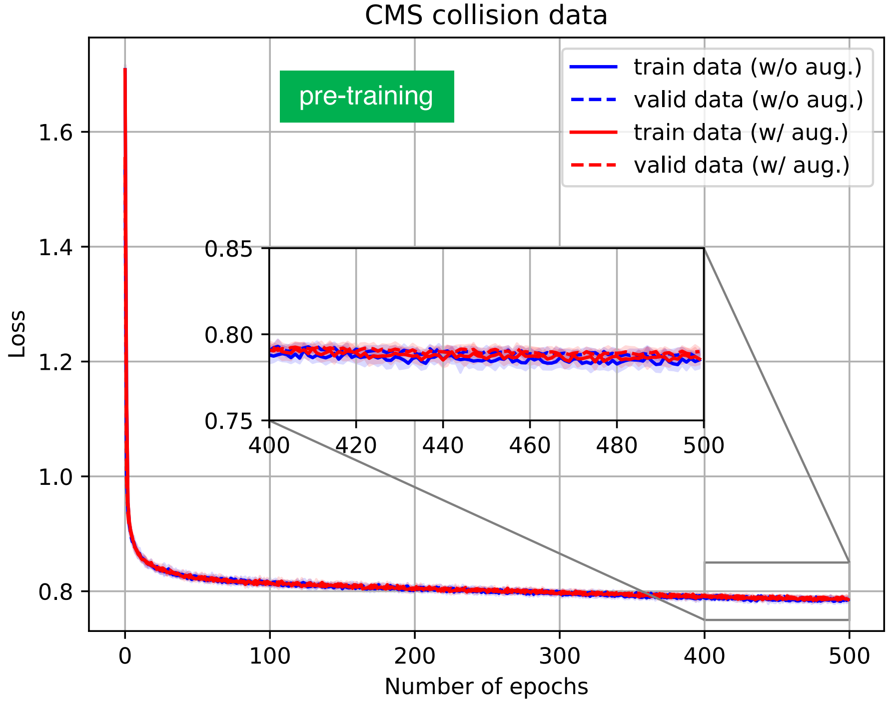
- These transformations are applied randomly before being fed into the DL model (pre-training phase)

DA (pre-training phase)

~10^4 events used



~10^6 events used

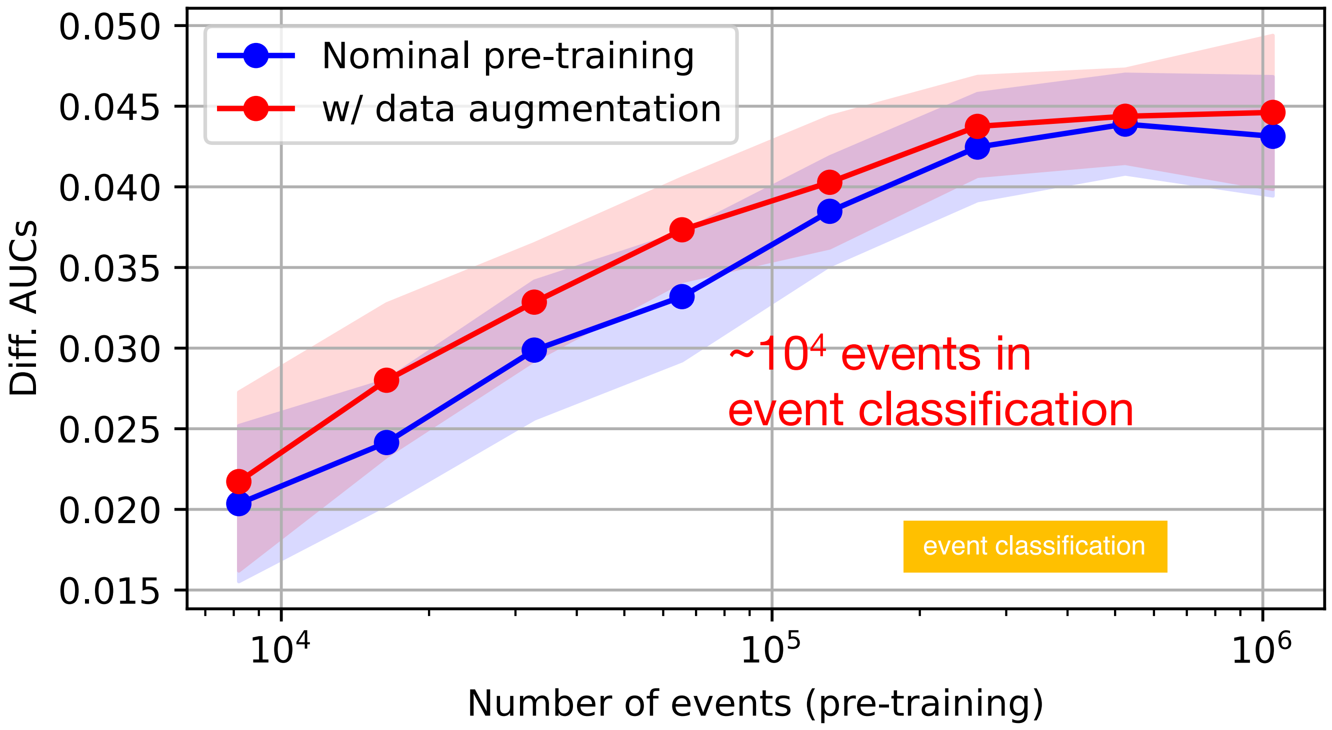


No effect(?)

→ Over-fitting is suppressed by the data augmentation if the number of events is small

Improvements for event classification

H⁺HW signal



➤ Improvements for the downstream event classification are not so visible (within the standard deviation)

→ Do you have any other data augmentation ideas?

Summary

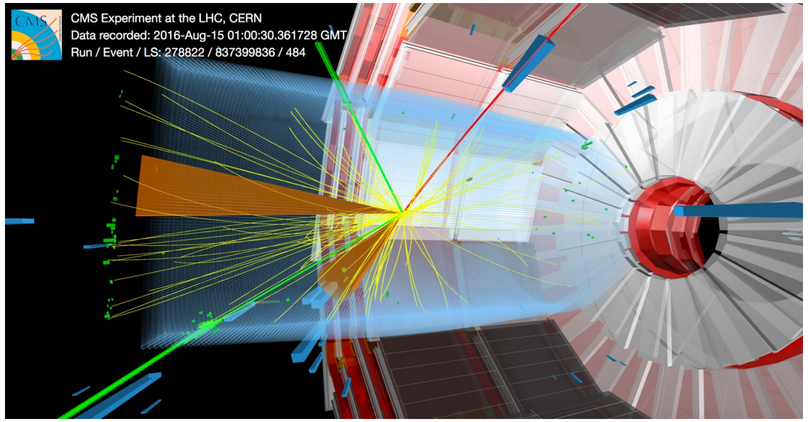
- Focusing on foundation models (transfer learning) and studying their applications to collider physics
  - Motivated by reduction of computing resources for future experiments
- Developed a self-supervised learning using real data in pre-training
  - The pre-trained model provides significant improvements in event classification when the # of events is small
  - The scaling behavior encourages pre-training with a larger data
    → Data augmentation technique in our physics data was discussed
  - (Need to check the scalability with larger models and larger data)

Backup

CMS open data

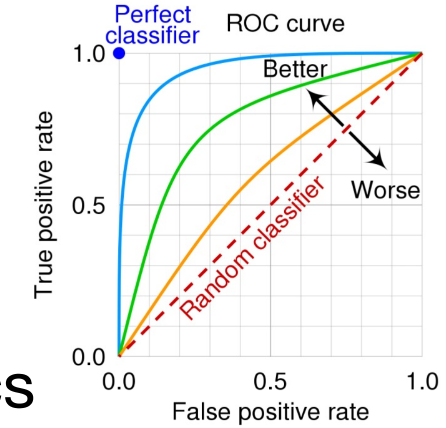
- LHC-CMS released new open data in 2024
  - 70 TB of 13 TeV collision data in 2016 and 830 TB of MC simulations
  - 16.4 fb^{-1} collision data (the Higgs discovery required 10.4 fb^{-1})
  - Nano AOD format
    - Possible to analyse by pure ROOT (and RDataFrame) 😀
    - (Previous open data requires the CMS software…)

→ This study should be reproducible



A candidate event in which a top quark is produced in association with a Z boson.

AUC metric



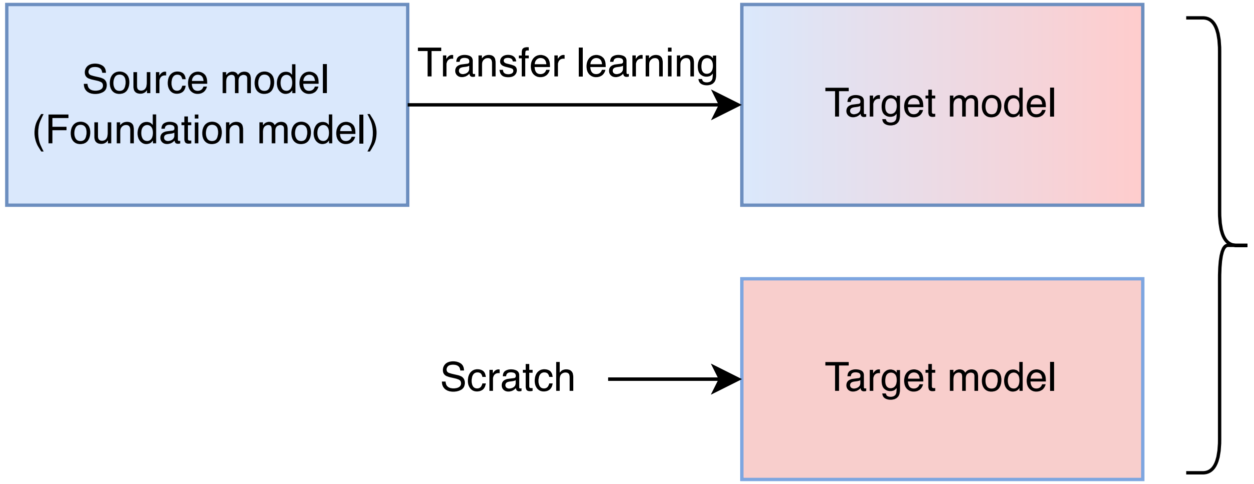
- Event classification performances are evaluated with AUC metrics

Pre-training phase

(Self-supervised learning)

Event classification phase

(signal vs background)



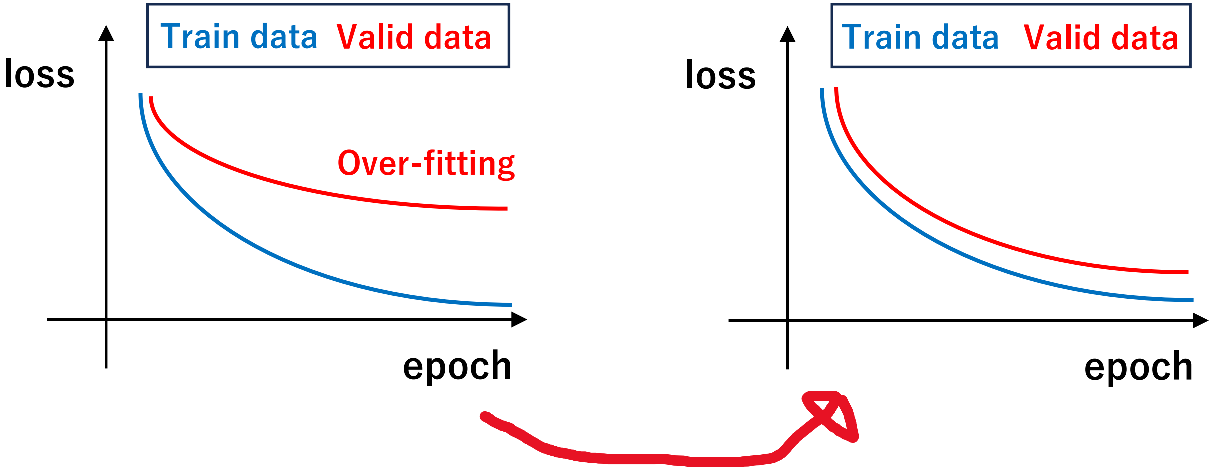
→ AUC values of event classifications are compared with and without a foundation model

Data augmentation

- Data augmentation is well established technique in computer vision field

[albumentations]()

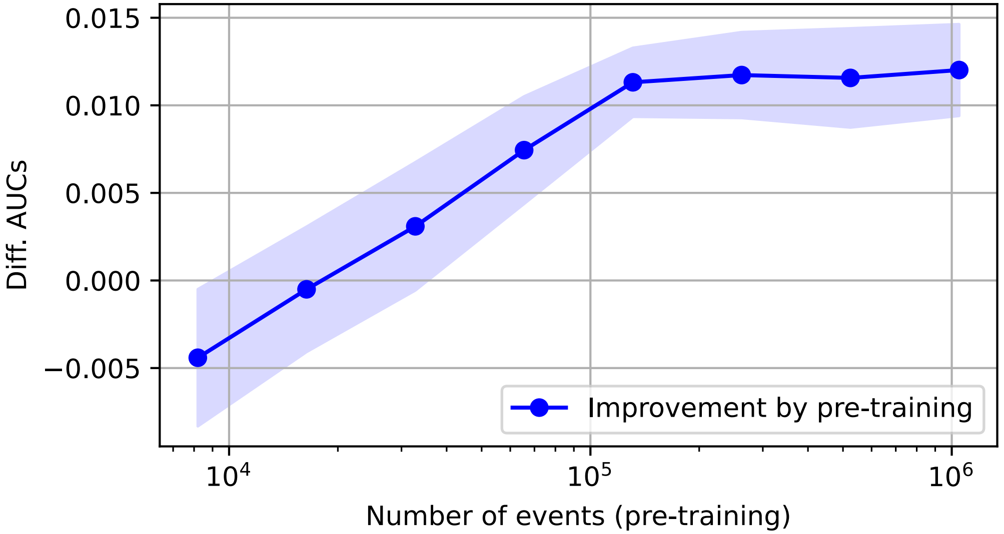
→ Easy to increase data with low computing cost, and effective to suppress over-fitting

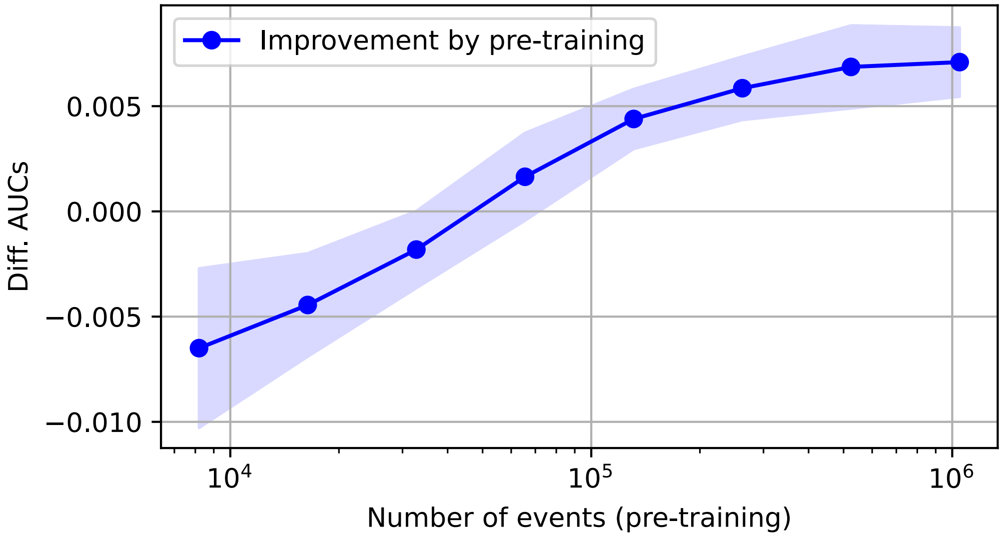


Data augmentation(?)

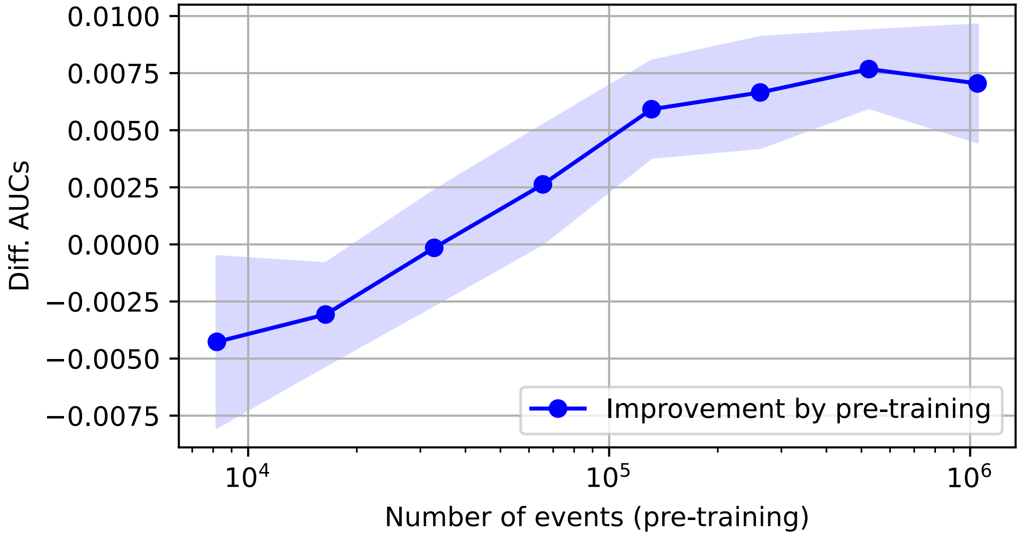
Scaling raw

c1





c3



c4