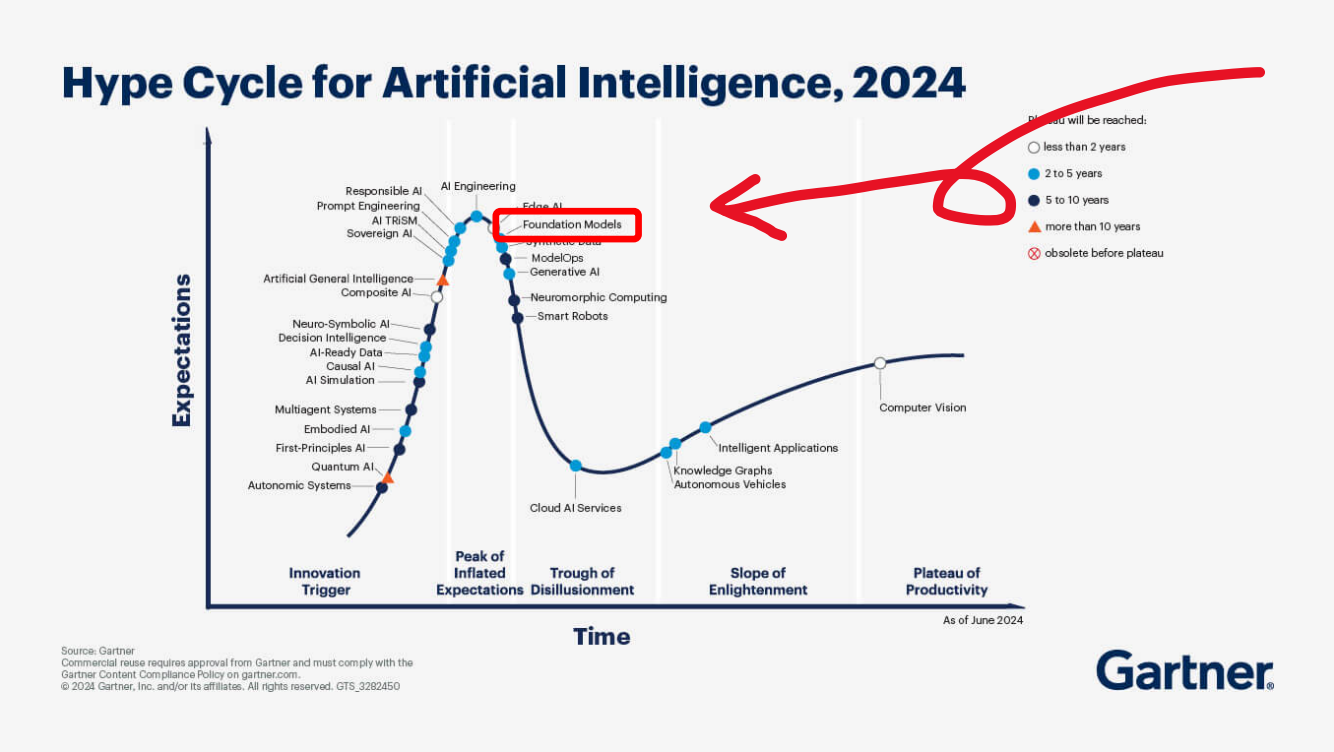
A study of foundation models for event classification in collider physics

Tomoe Kishimoto

Computing Research Center, KEK

tomoe.kishimoto@kek.jp

Introduction

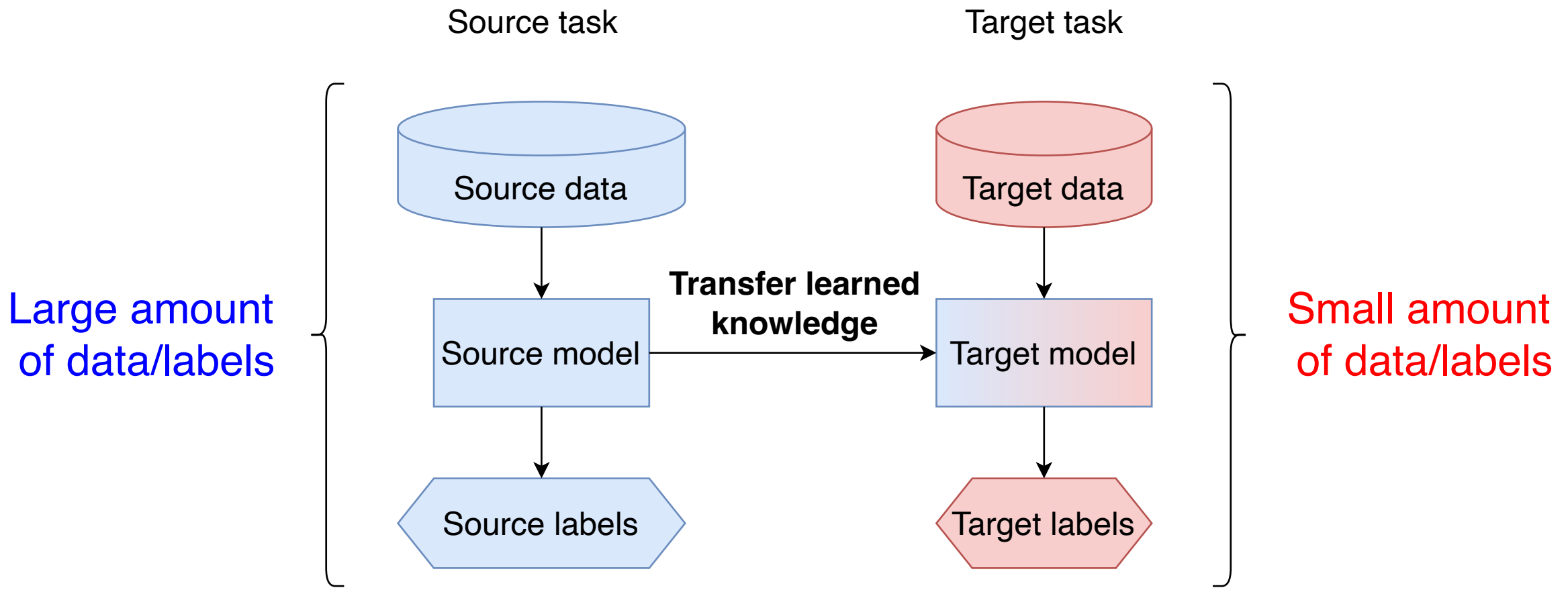


Gartner.com

- “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)

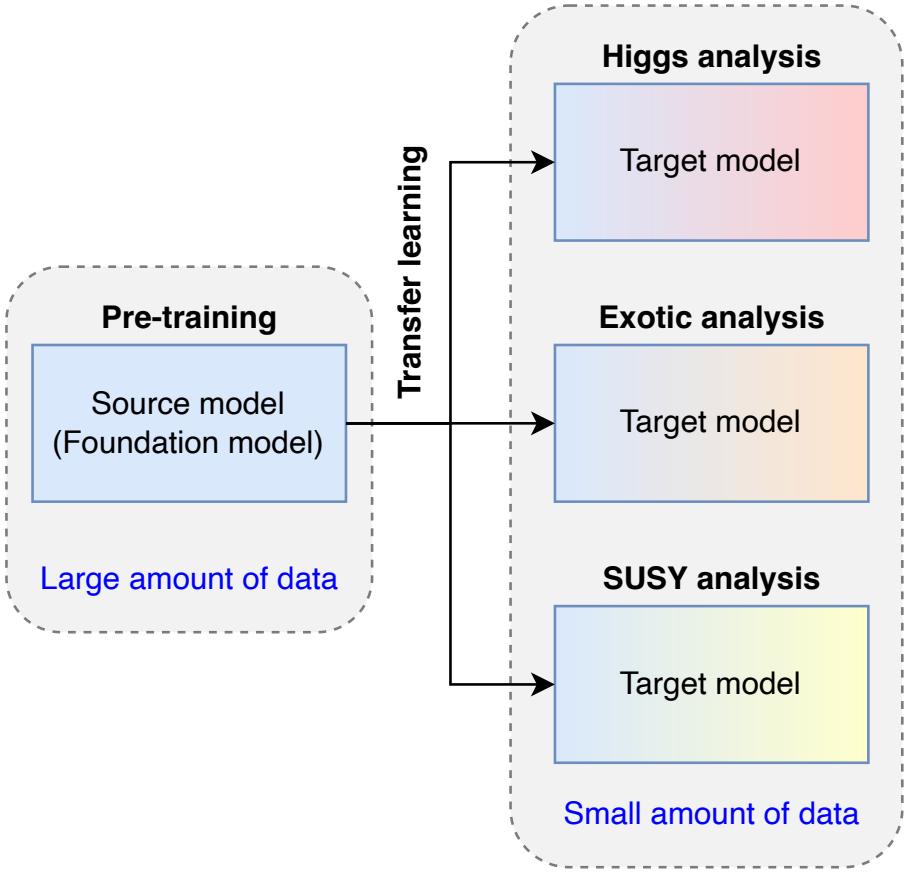
→ Q: Is the concept of foundation models beneficial for collider physics?

Transfer learning



Use cases in physics analysis

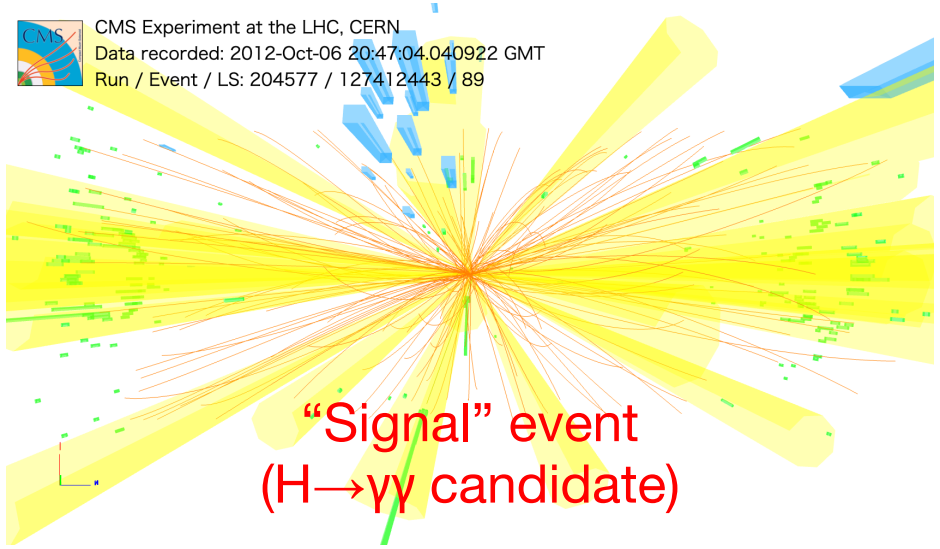
- Multiple analysis channels in collider physics
  - Higgs, Exotic, SUSY, etc
  - Currently, dedicated DL models are trained from scratch for each channel
    ← Large amount of training data (MC) required for each channel

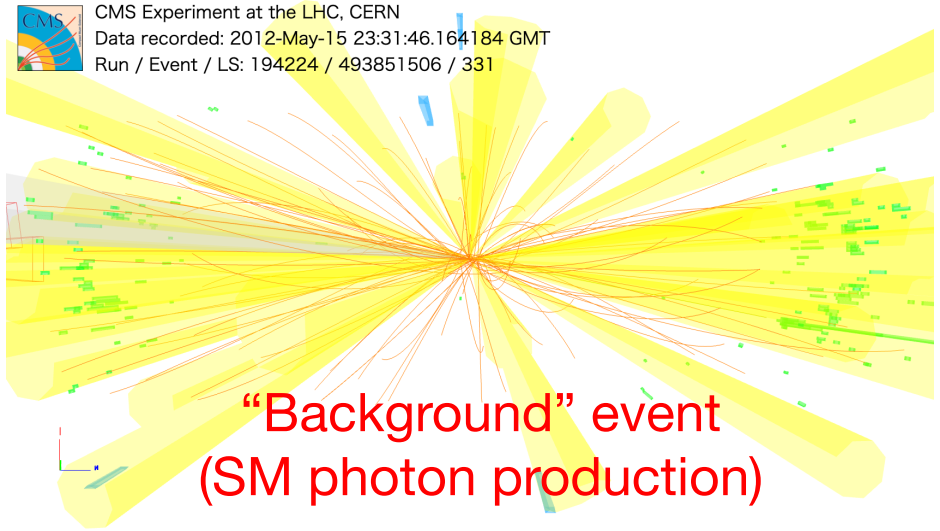


→ If transfer learning can be applied across different analysis channels, the computing resources for MC simulations and DL training could be reduced

Event classification

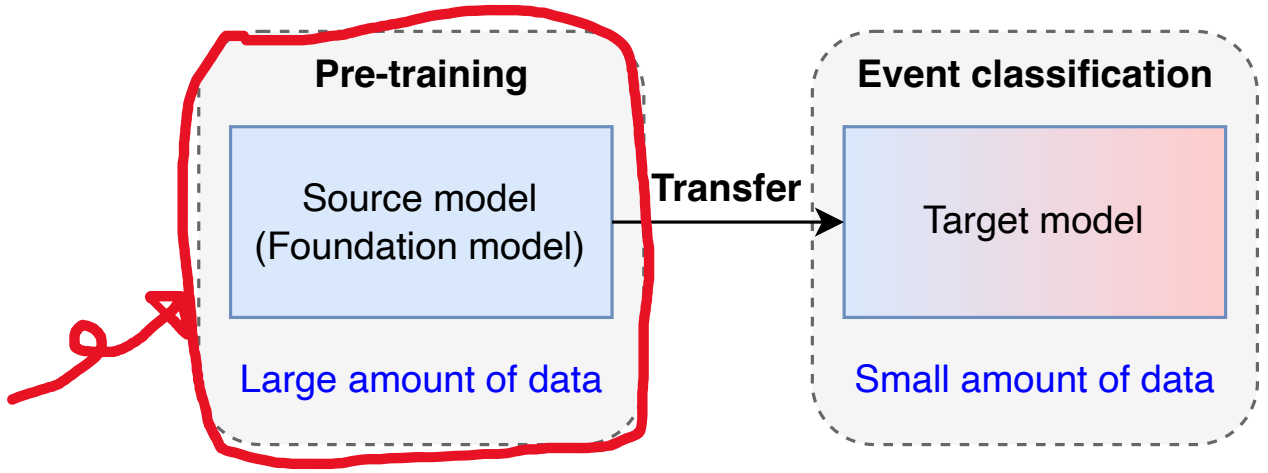
- The concept is examined using “an event classification” problem
  - A typical task in HEP, signal event vs. background events





→ Reconstructed particles (objects), such as electrons, muons, and jets, provide the input basic information for the classification

Pre-training strategy

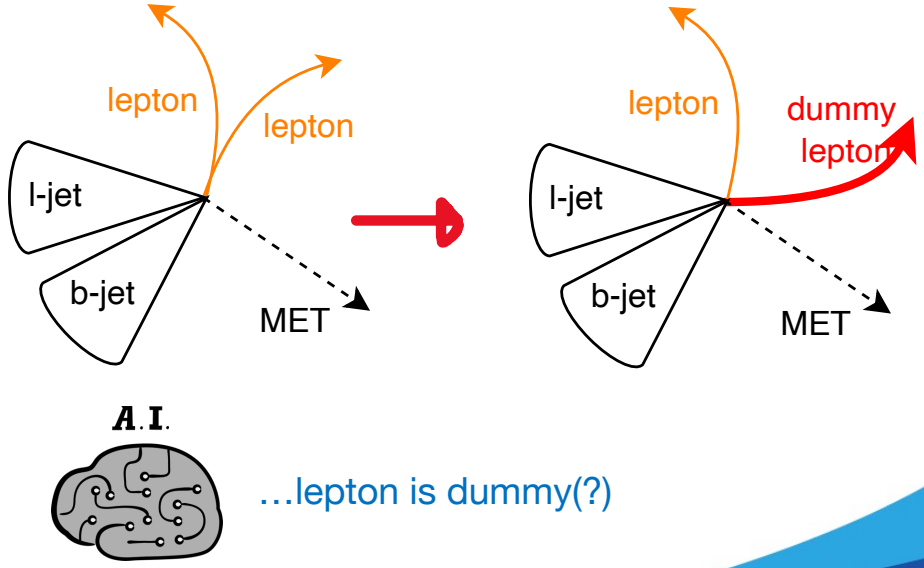


- The key concept of this study:
  - How to perform pre-training?
- Pre-training strategy:
  - Real collision data (CMS Open Data) are used for pre-training
    - No need to generate a large amount of MC data, and
    - Bias from the arbitrary selection of physics processes is mitigated
  - Self-supervised learning is employed to handle the unlabeled collision data

Self-supervised learning

- Labels are automatically generated during the training process
- Self-supervised learning strategy:
  - One object in an event (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

→ DL model needs to learn relationships among objects, which should be useful knowledge for downstream event classification tasks!

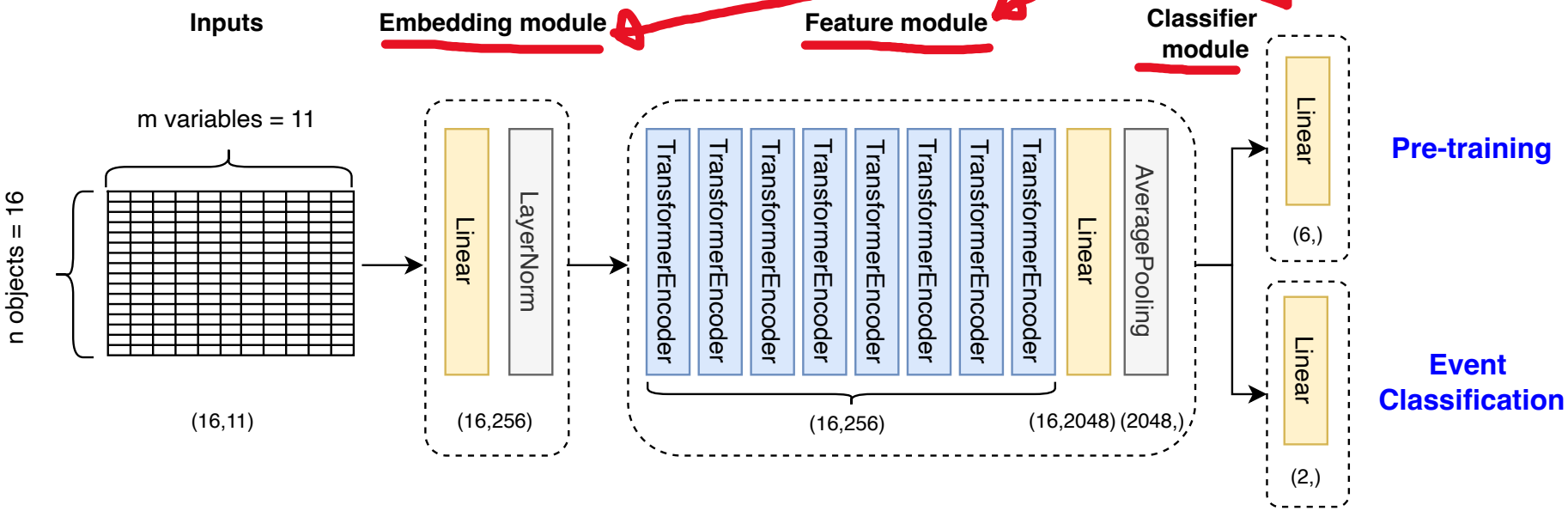


DL model

- Transformer encoder is employed:
  - ~11M trainable parameters
  - Inputs: 4-vector + charge for each object

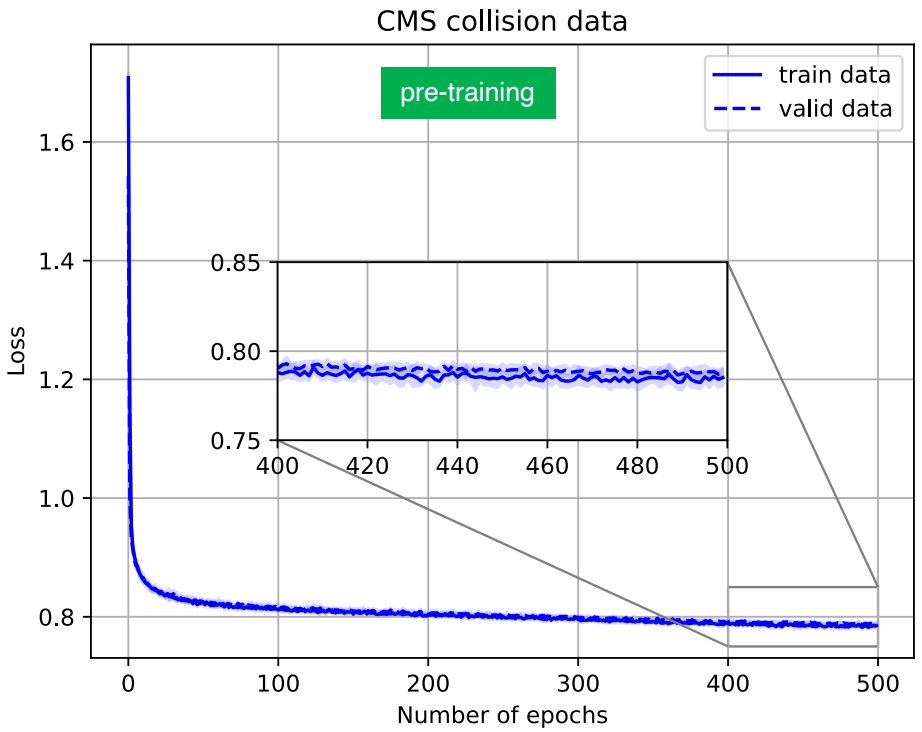
→ Weight parameters of embedding and feature modules are transferred and fine-tuned

→ Classifier module is always trained from scratch, depending on tasks



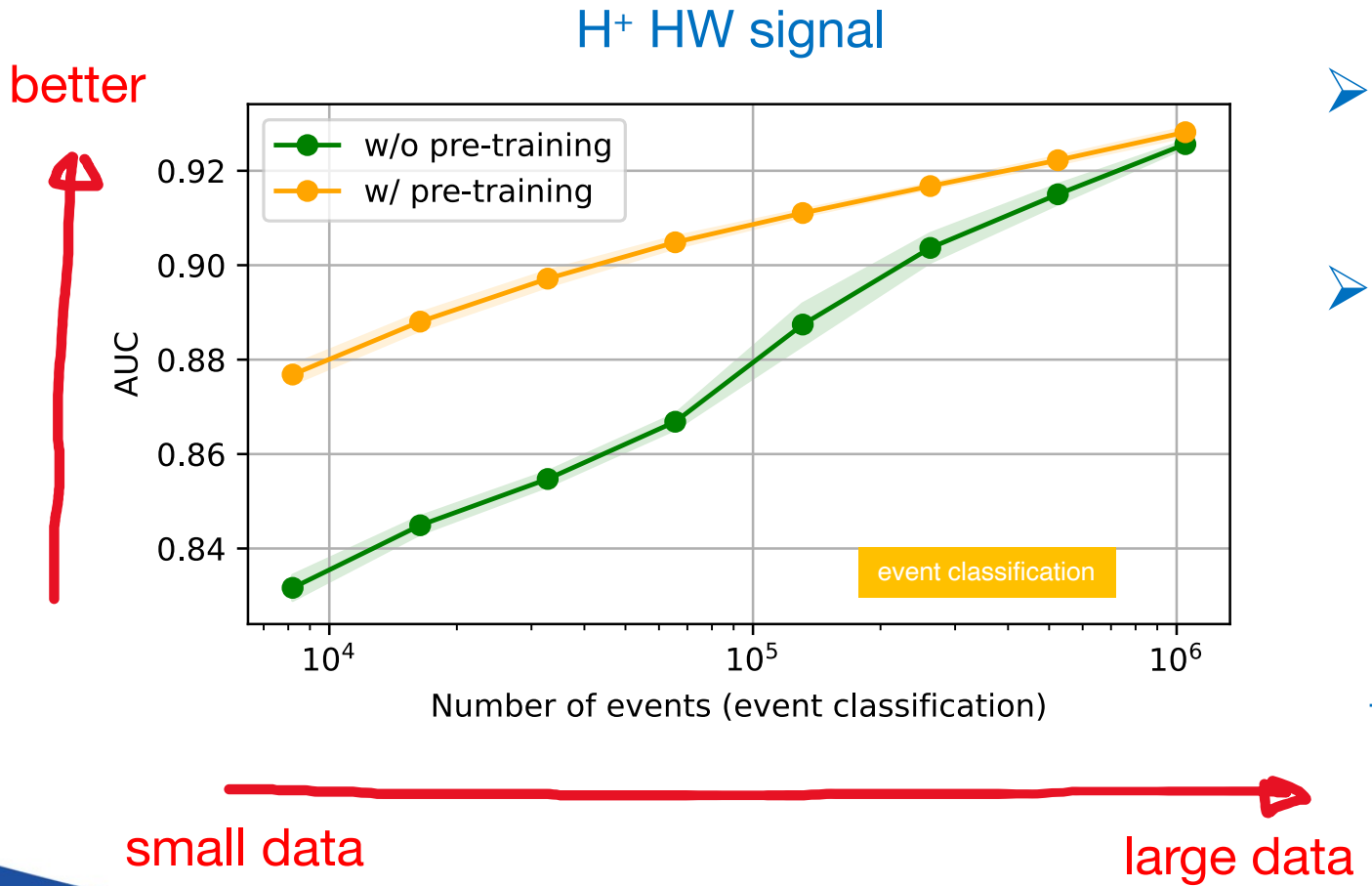
Training details

- 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:
- NVIDIA A100: ~20 batches/s
  - ~13 hours for one training



~1M events used

AUC of event classification

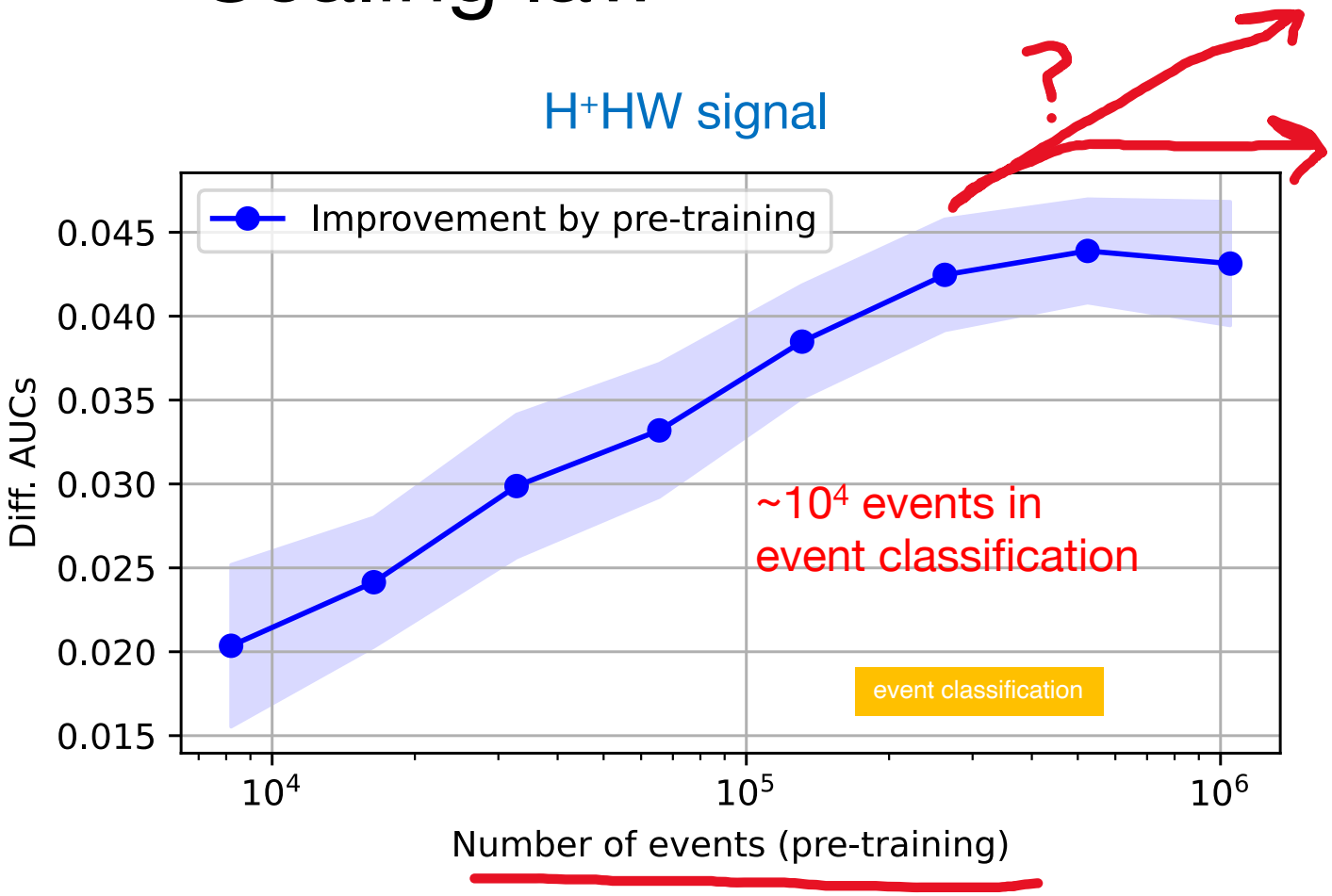


- Charged Higgs (H^+) is selected as signal events
- Significant improvements by introducing the pre-training
  - The performance converged when # of events increased to ~10^6

→ Expected behavior of transfer learning

加速器だから見える世界。
KEK

Scaling law



- The scaling behavior encourages a pre-training with a larger data
  - Future work: need to check the scaling behavior with larger data and a larger model
    - (One training with 10^{10} events will require (A100 x 8) x 700 days)
  - ATLAS Open Data are now also available

Summary

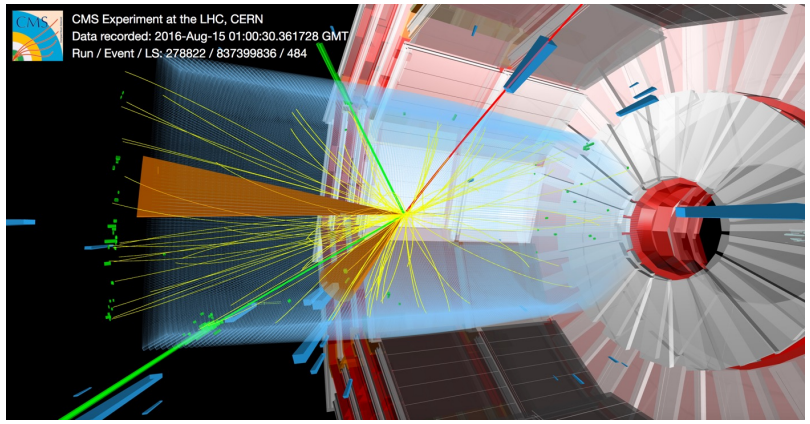
- Focusing on foundation models (transfer learning) and studying their applications to collider physics
  - Motivated by reduction of computing resources for future experiments
- Developed a self-supervised learning using real data in pre-training
  - The pre-trained model provides significant improvements in event classification when the # of events is small
  - The scaling behavior encourages pre-training with a larger data
- (But, my interests are moving toward LLM studies…)

Backup

CMS open data

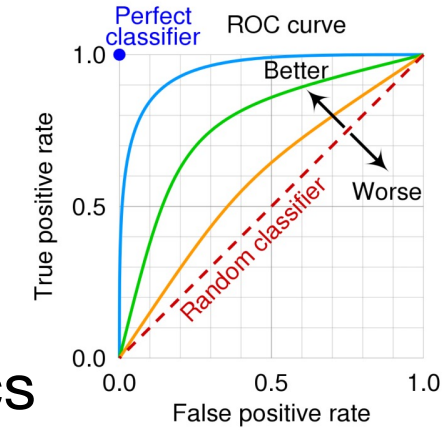
- LHC-CMS released new open data in 2024
  - 70 TB of 13 TeV collision data in 2016 and 830 TB of MC simulations
  - 16.4 fb^{-1} collision data (the Higgs discovery required 10.4 fb^{-1})
  - Nano AOD format
    - Possible to analyse by pure ROOT (and RDataFrame) 😀
    - (Previous open data requires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



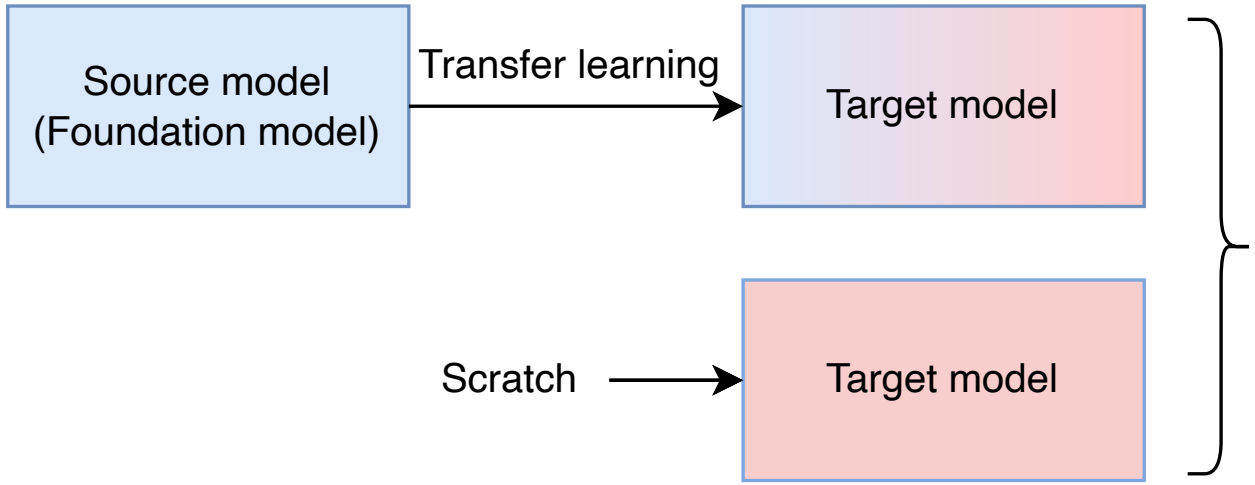
- Event classification performances are evaluated with AUC metrics

Pre-training phase

(Self-supervised learning)

Event classification phase

(signal vs background)



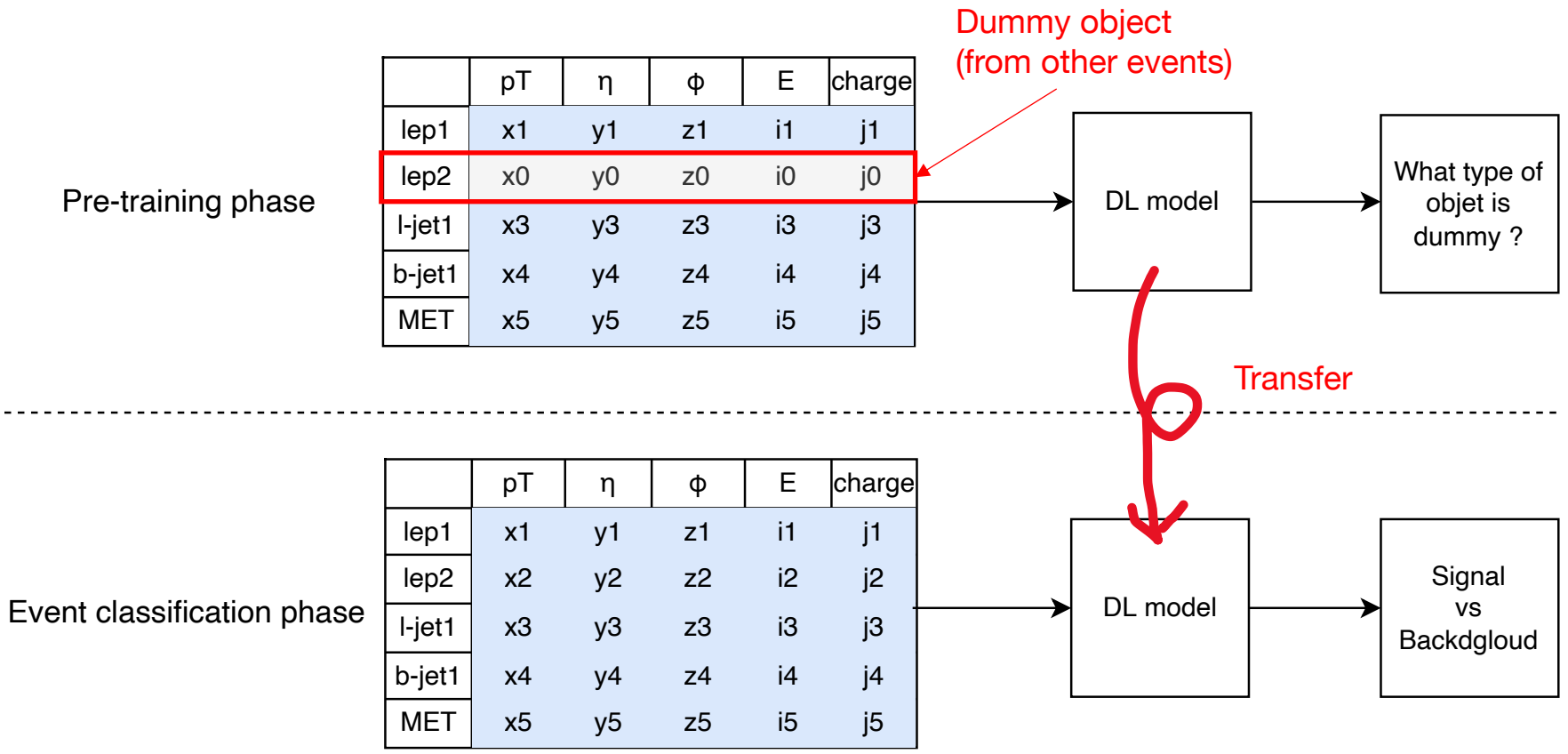
→ AUC values of event classifications are compared with and without a foundation model

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)

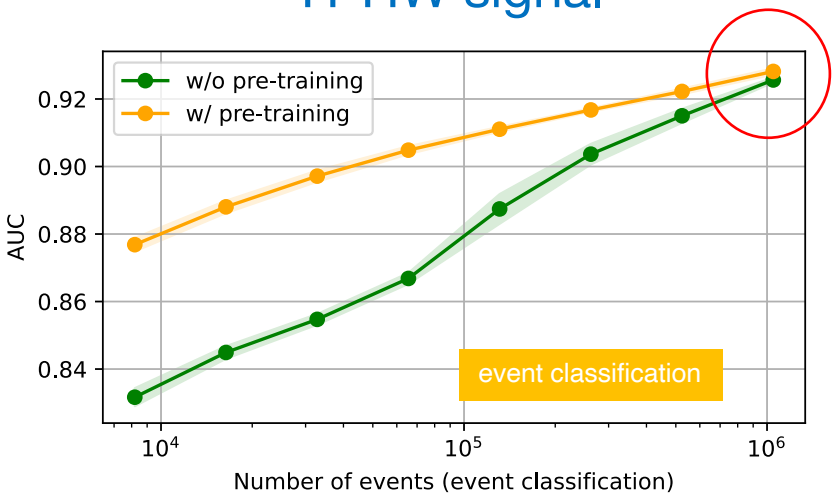
Pre-training strategy



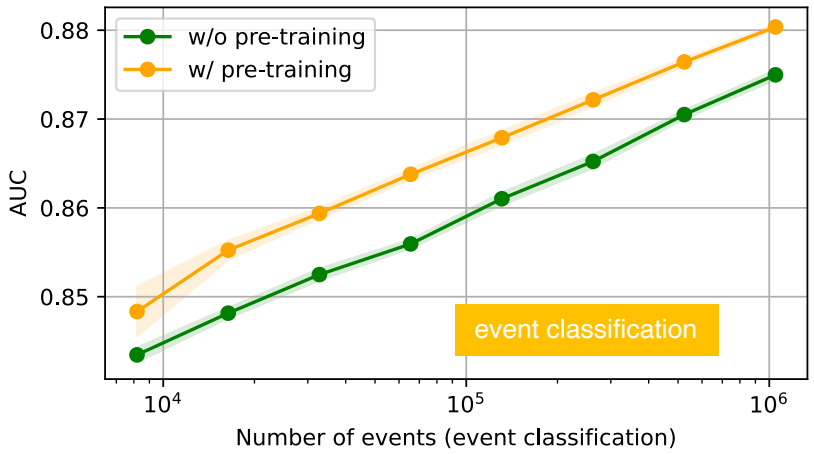
→ Random masks increase prediction pattern (data augmentation)

AUC of event classification

H⁺HW signal



ttH (1lep) signal



ttH (2lep) signal

- The improvements are confirmed for all signal events

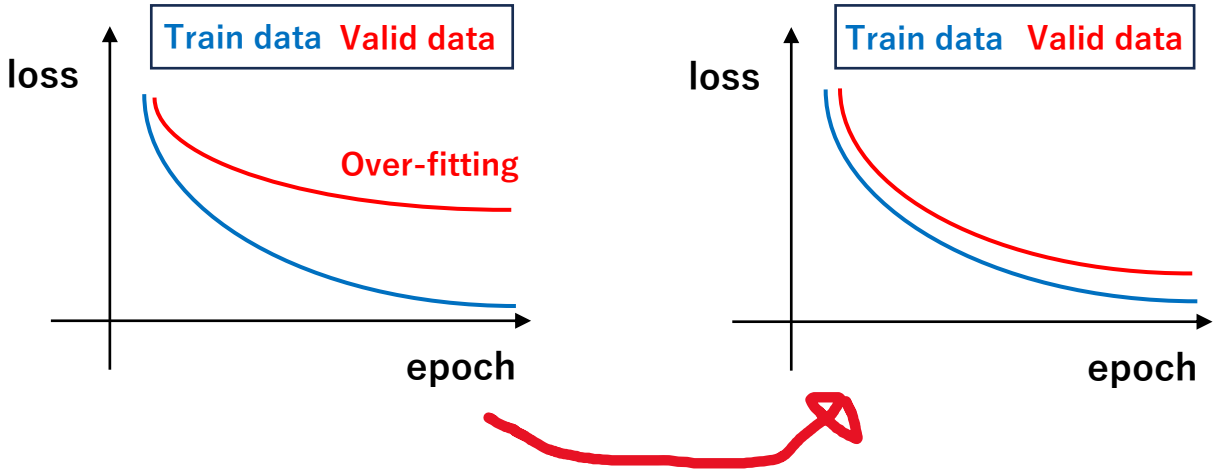
  → The pre-trained model (foundation model) is well generalized

Data augmentation

- Data augmentation is well established technique in computer vision field

albumentations

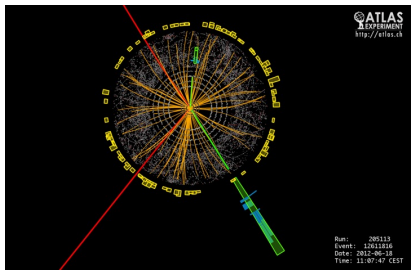
→ Easy to increase data with low computing cost, and effective to suppress over-fitting

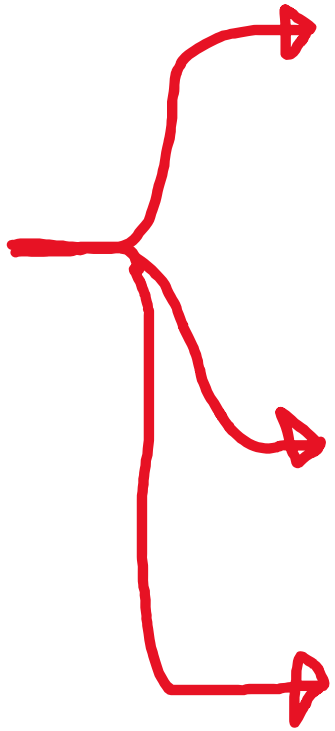


加速器だから見える世界。
KEK

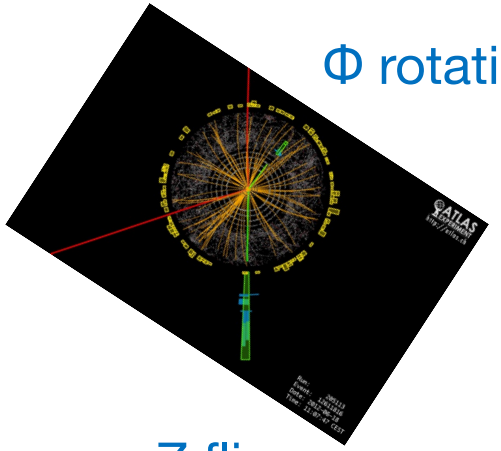
Lorentz transformation

Original event
(Higgs candidate)

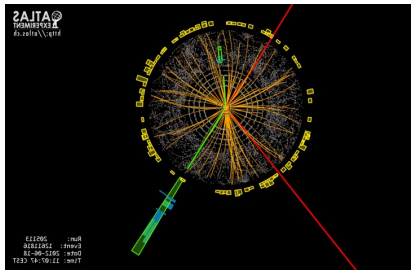




Φ rotation



Z flip



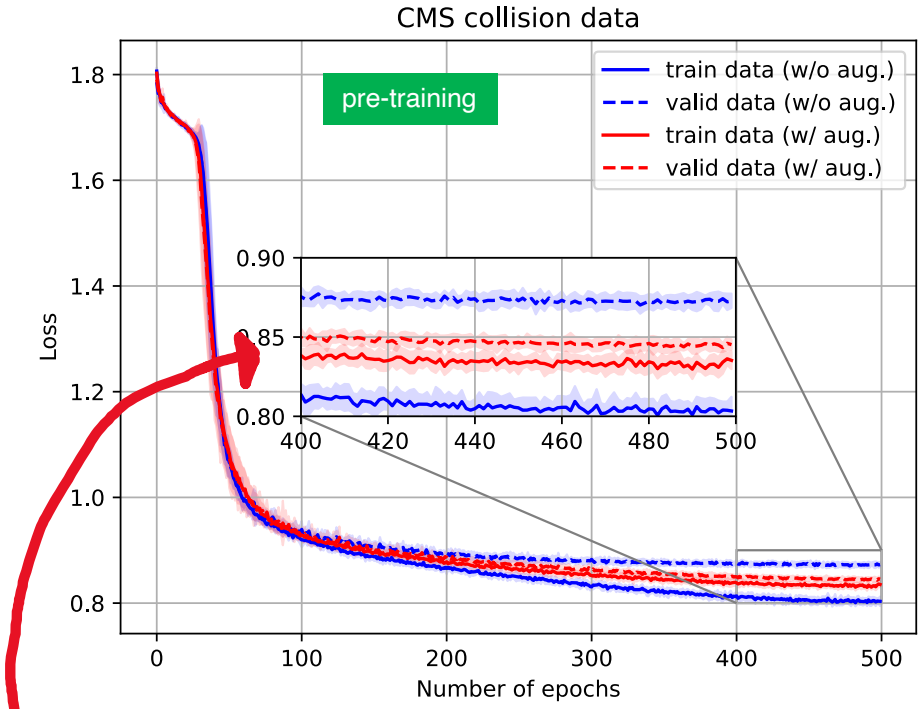
Lorentz boost
(z direction)

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

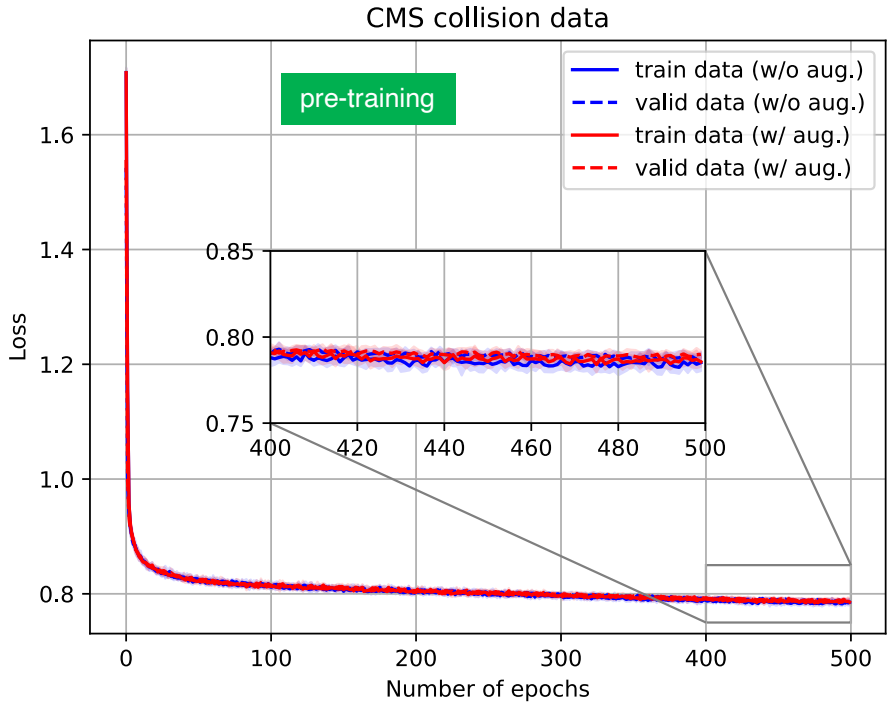
- These transformations are applied randomly before being fed into the DL model (pre-training phase)

DA (pre-training phase)

~10^4 events used



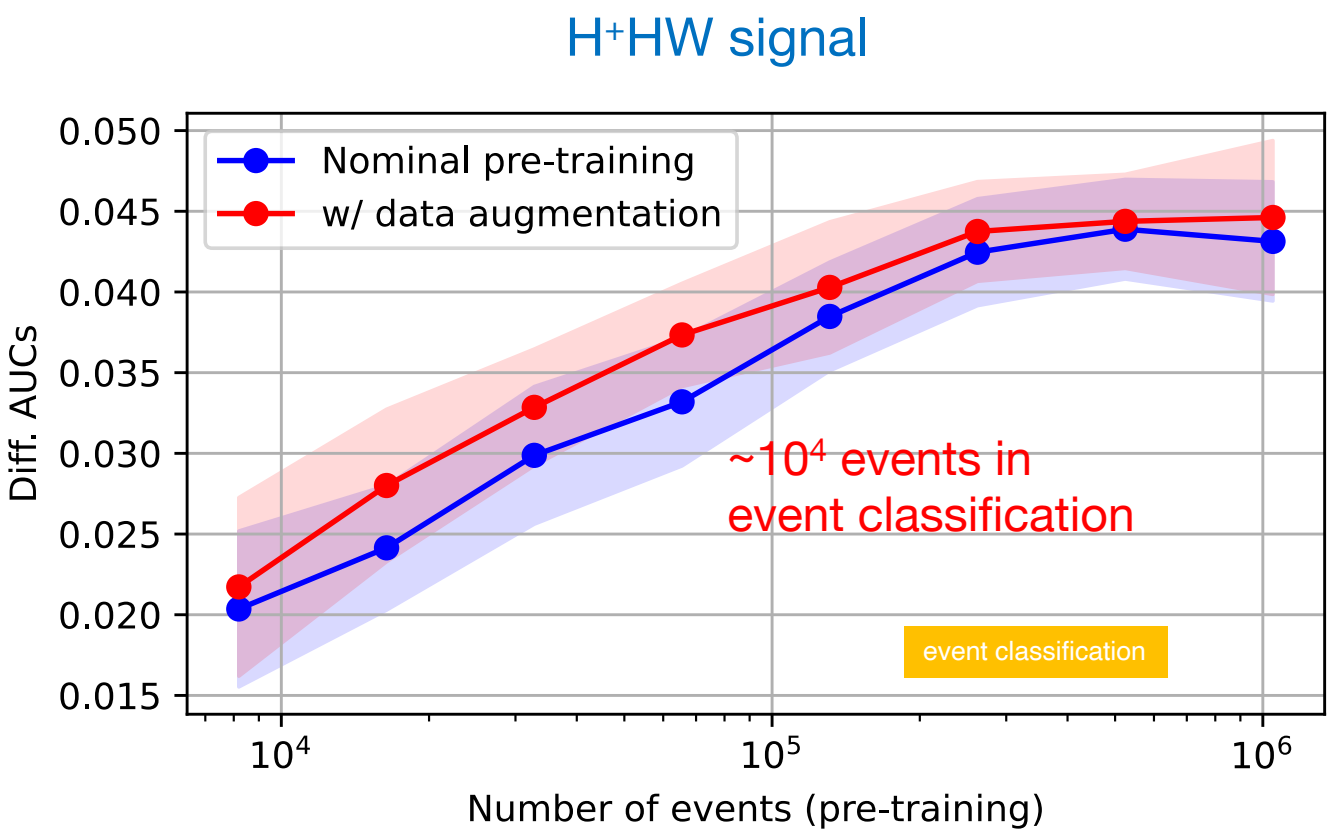
~10^6 events used



No effect(?)

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

Improvements for event classification

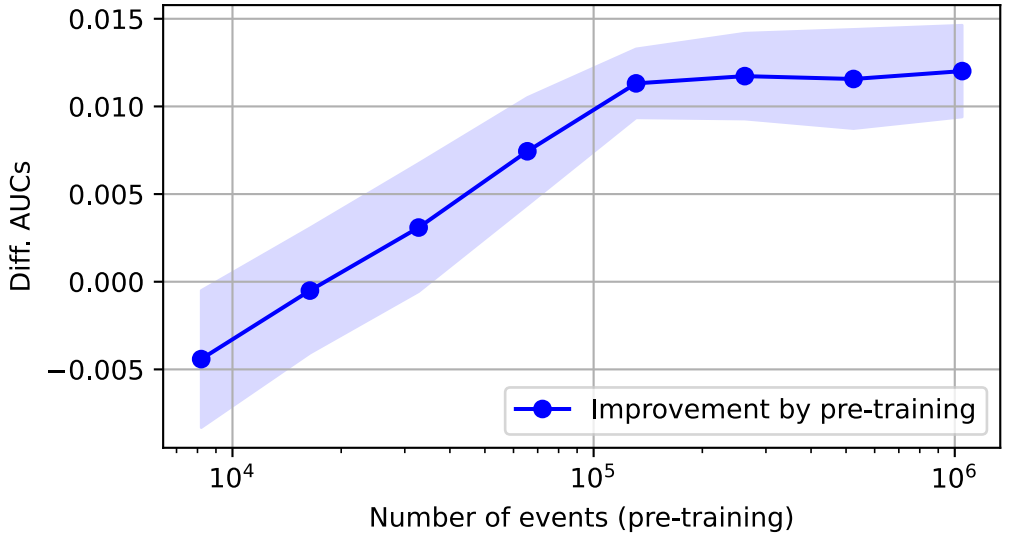


- Improvements for the downstream event classification are not so visible (within the standard deviation)

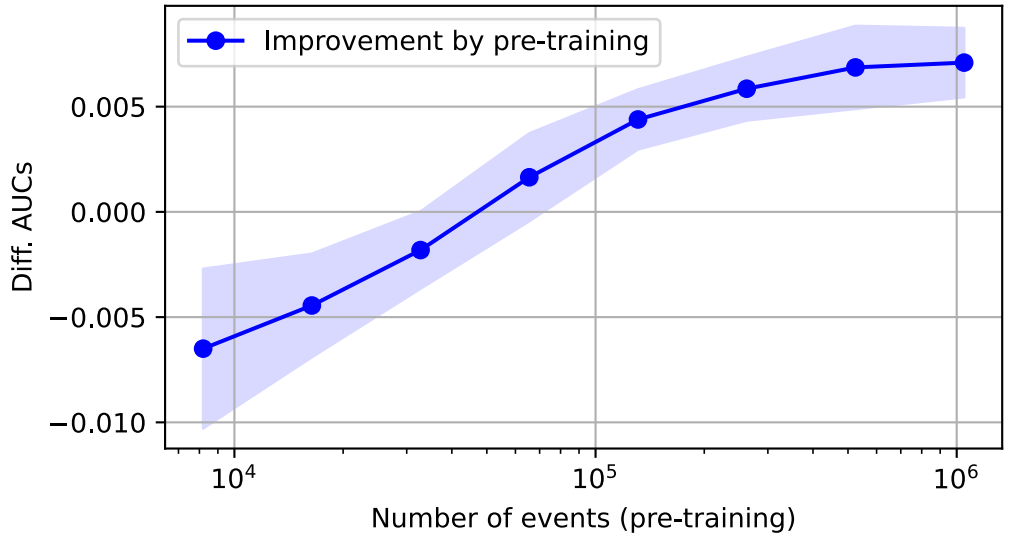
→ Do you have any other data augmentation ideas?

Scaling law

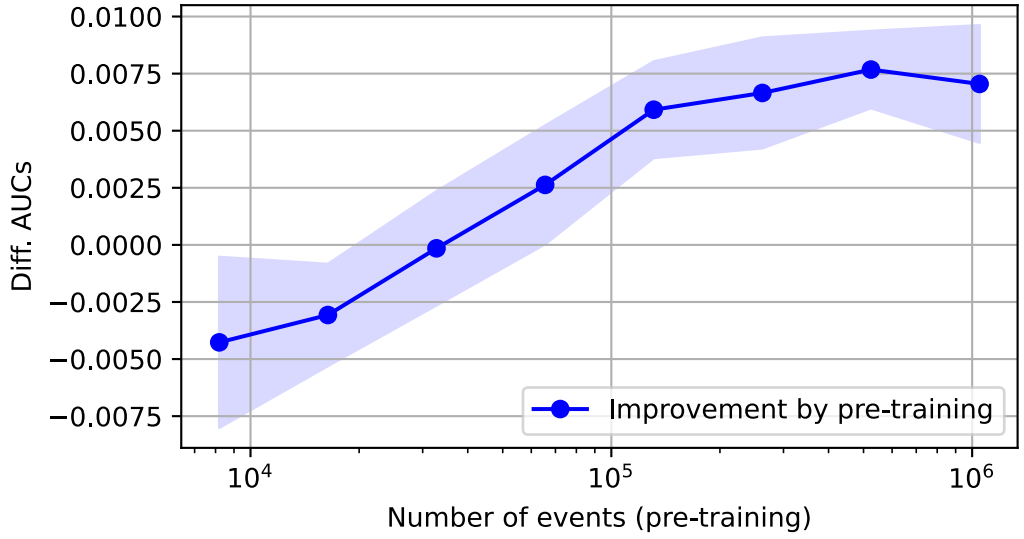
c1



c3



c4



加速
KEK