Towards a foundational jet model:
Enhancing generalization with contrastive "gen-reco" pre-training

Zixun Kou Congqiao Li Qiang Li

Peking University

Quantum Computing and Machine Learning Workshop 2026, Shenzhen
19 August, 2026

Motivation

Universal model

- Pre-training
  according to scaling law, pre-training large models on comprehensive datasets can push a broad range of final states towards their sensitivity frontier

- Achieve state-of-the-art performance
  reach the best possible accuracy across all established tasks (e.g. A vs. B tagging, mass/p_{T} regression, etc.)

- Ensure strong generalization
  reach as better performances as possible for new tasks

Unsupervised approach

Collected a few examples
(sorry for missing your favourite work)

✤ **Contrastive learning** via pos/neg samples

✤ **BERT-like**: masked XX modelling

✤ **GPT-like**: next-token prediction

✤ **JEPA-like**: predict masked representation

JetCLR SciPost Phys. 12, 188 (2022)

RS3L PRD, 111 (2025) 3, 032010

CLEAR Today's talk

MPM 2401.13537

MPMv2 2409.12589

Bumble-bee 2412.07867

OmniJet-α MLST, 5 035031 (2024)

P-JEPA ML4Jets talk

J-JEPA 2412.05333

HEP-JEPA 2502.03933

Supervised approach (+X)

Sophon w/ JetClass-II

2405.12972

Motivation

Computer Vision

- models trained on ImageNet were among the **earliest pretrained CV models.** (serving to **generalize to other CV domains**)
- Modern self-supervised learning (SSL) methods (e.g. MAE, I-JEPA, …) show strong performance and beat SL.
- but one fact is that the **supervised baseline is relatively weak** (ImageNet-1k only has 1M images)

ImageNet-1k
the cornerstone of modern CV
1M dataset; 1000 labels

HEP dataset

- Establish **a very strong baseline**

modern supervised models (e.g. SoTA taggers in ATLAS/CMS) are already trained on o(100M)-level datasets.

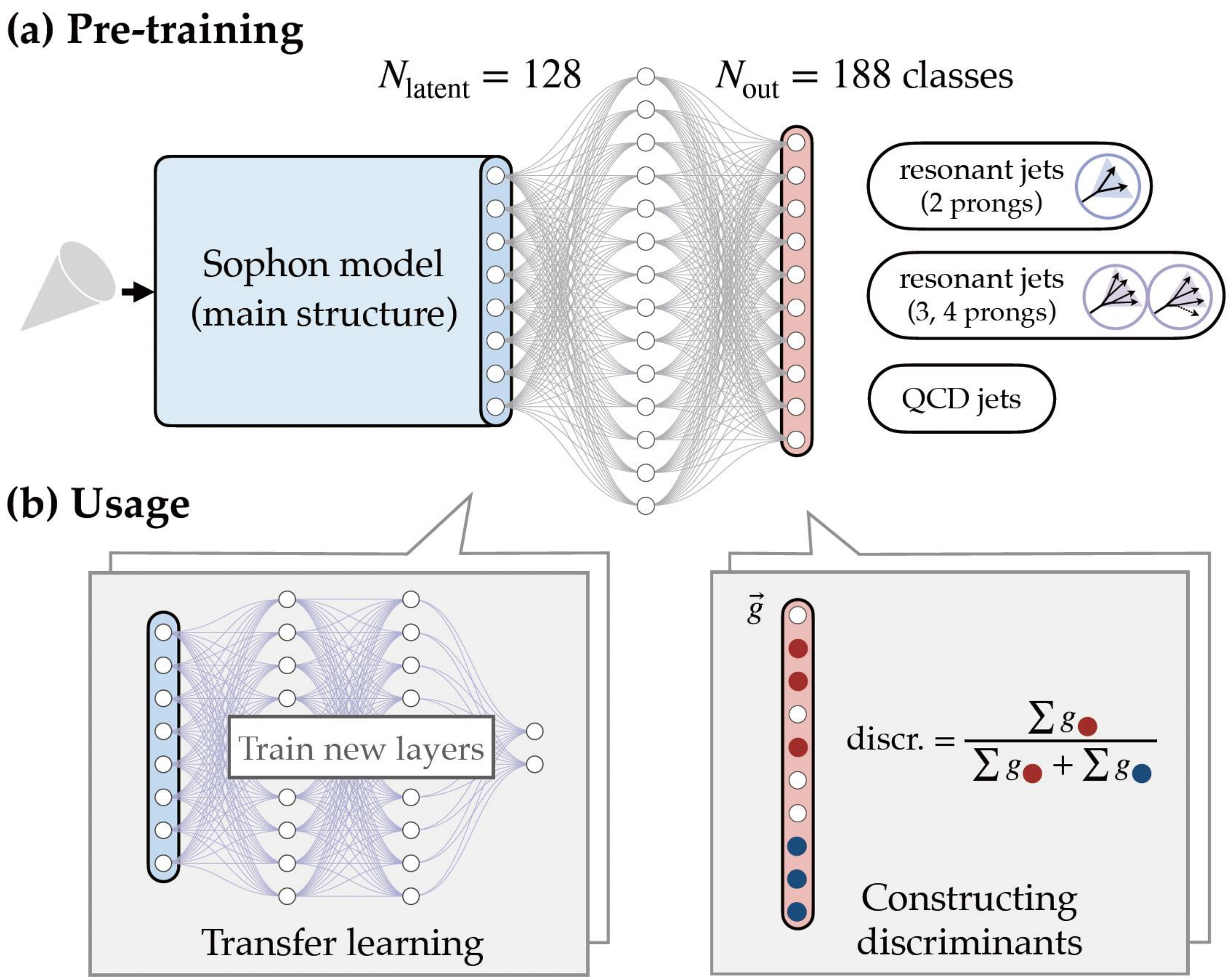
Motivation

Current *Sophon* model

- *Sophon*: Signature-Oriented Pre-training for Heavy-resonance ObservatioN

- Key concept: Pre-training an expressive Transformer model on a wide range of jet phase spaces on a multiclass classification task

- Target boosted-jet phase-space explored simulate datasets
  QCD, resonant jets with 2, 3 or 4 prongs (188 categories)



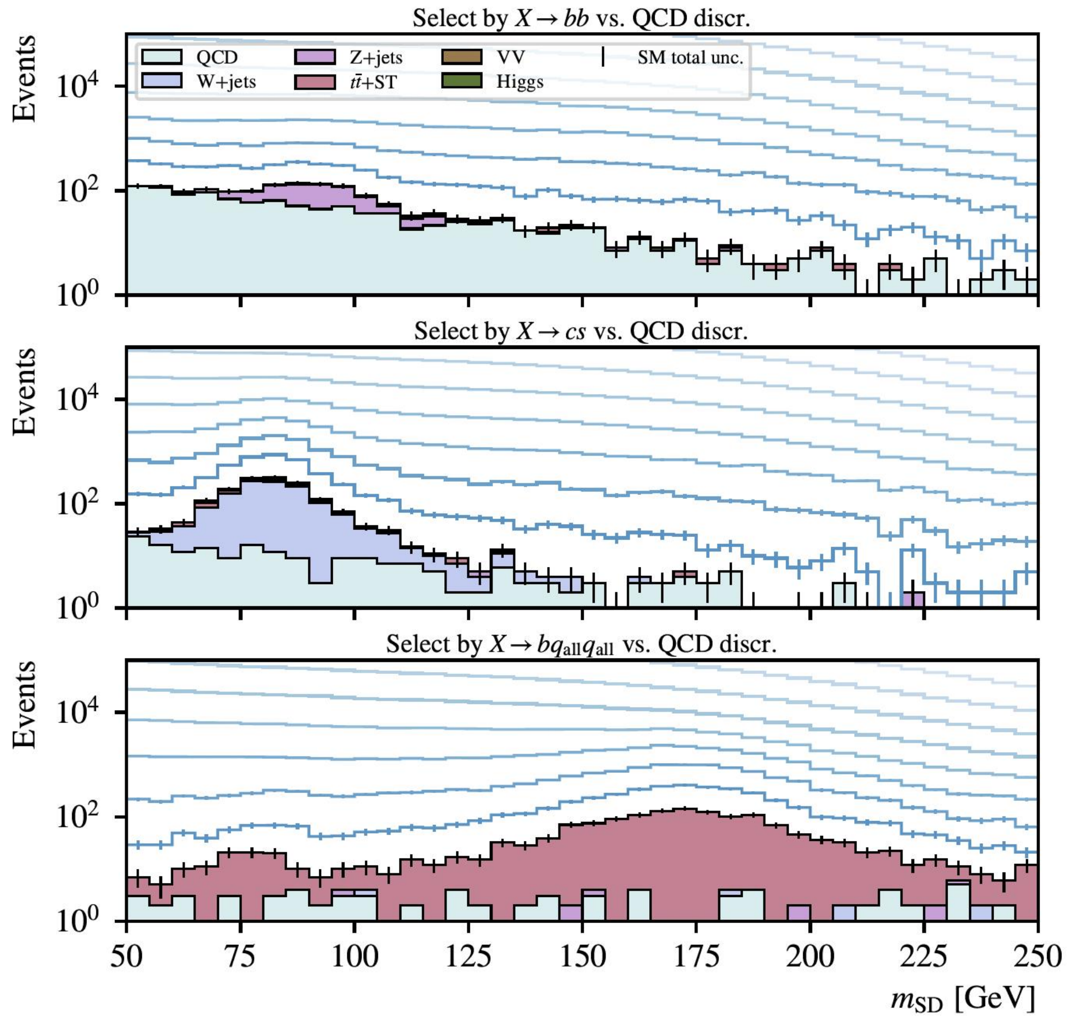
Motivation

- Provides directly usable discriminants and outperforms SOTA results

- Demonstrate broad ability to construct discriminants and sensitively probe resonances with unknown properties (QCD, V+jets, ttbar+single top, di-boson, Higgs production)

Way to improve

- A simple classification approach
- Jet signatures of initial state remain unused
- Enhance transfer-learning ability in other tasks



Next Step

Supervised approach + X?

"Signature-Oriented Pretraining" ↔ pretraining with labels
serving as a pretraining foundation

Further enhance model generalization with some SSL/…

Sophon

Next Step

Supervised approach + X?

"Signature-Oriented Pretraining" ↔ pretraining with labels
serving as a pretraining foundation

Sophon

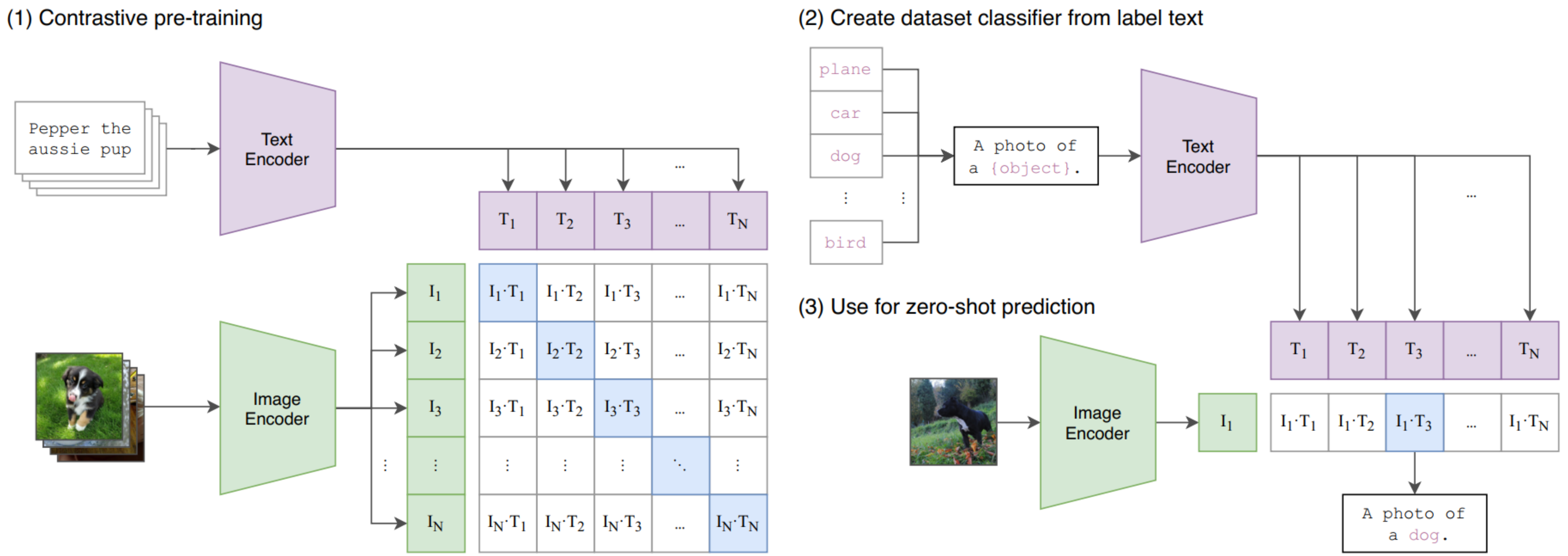
Further enhance model generalization with some SSL/…

Sophon++

Next Step

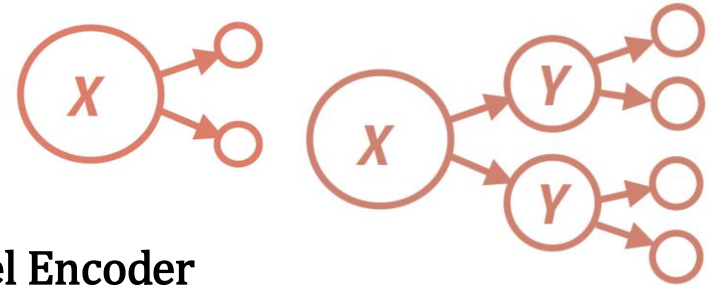
CLIP

- Contrastive Language-Image Pre-training
- Multi-Modality(language, image)

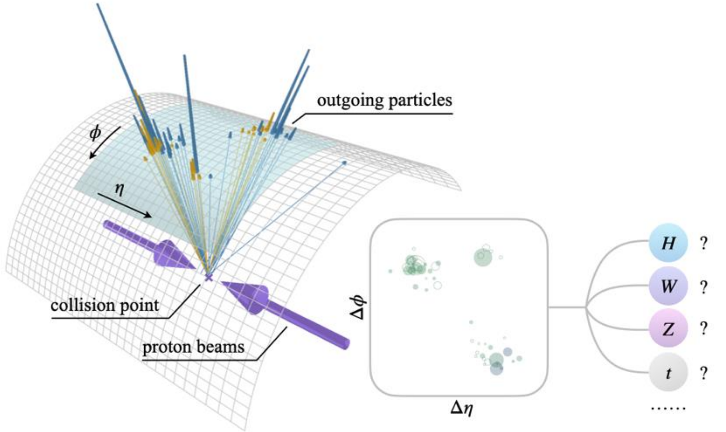


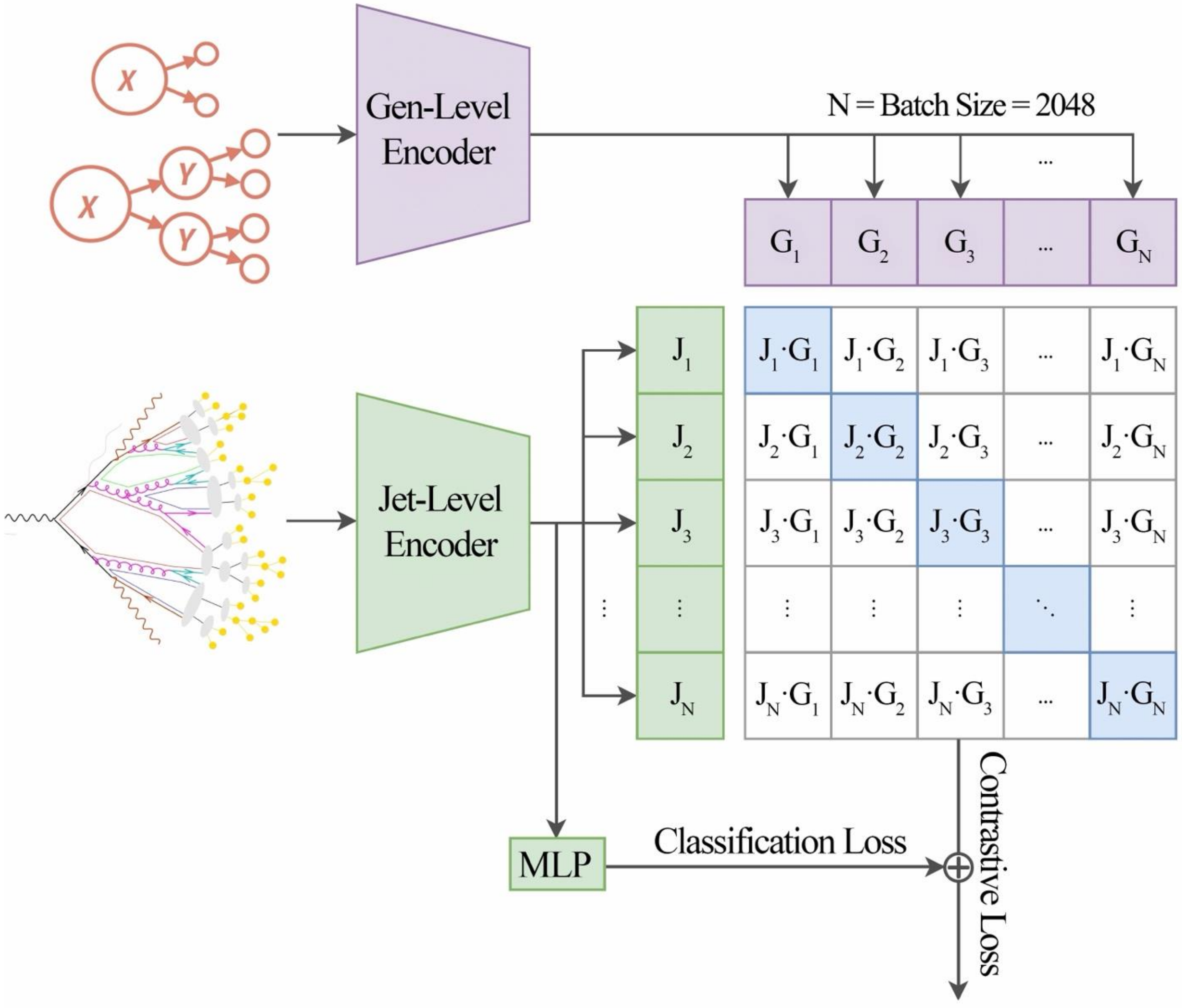
Pre-training Setup

- **Gen-Level Encoder**
  Input: features of dedicated final-state quarks and leptons



- **Jet-Level Encoder**
  Input: features of jet constituents (same as original *Sophon*)

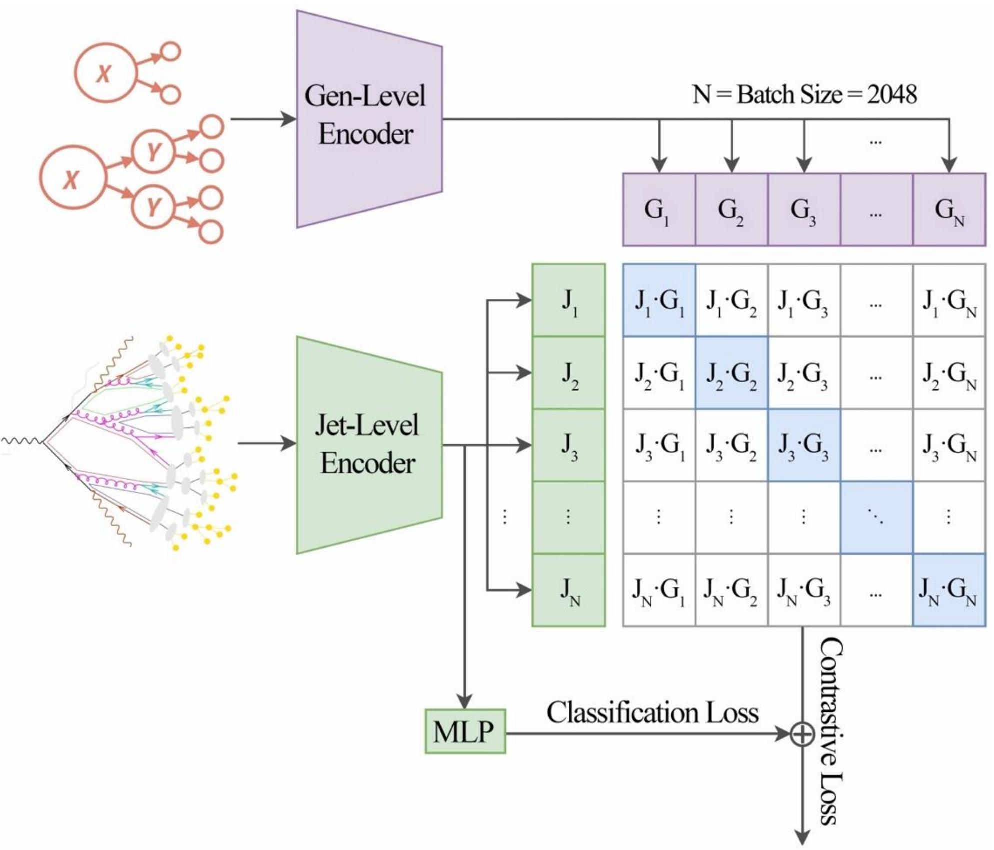


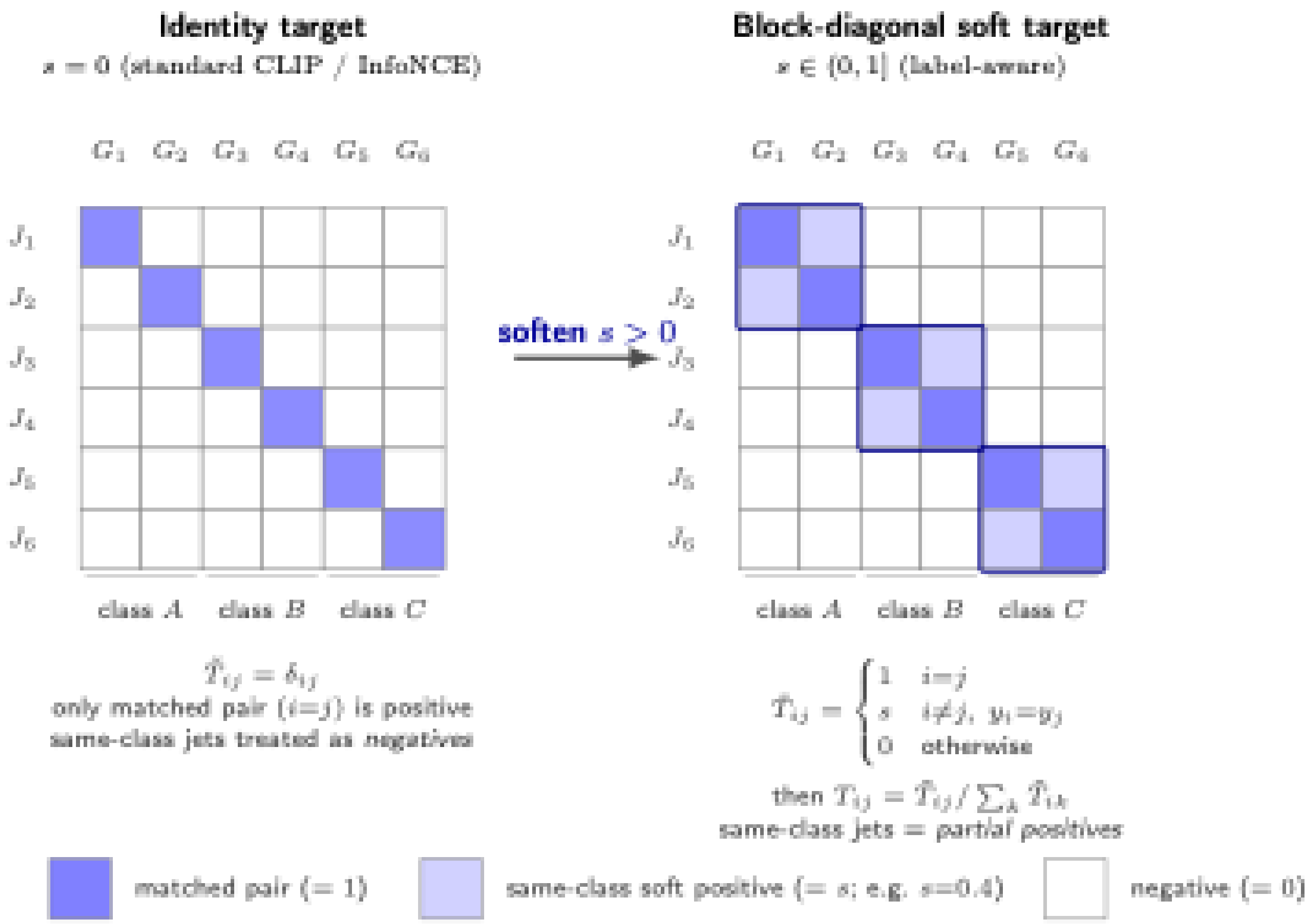


Class tokens: separated token for each task
Network structure: ParT for both

loss = $\alpha *$ contrastive loss + $\beta *$ classification loss

Pre-training Setup





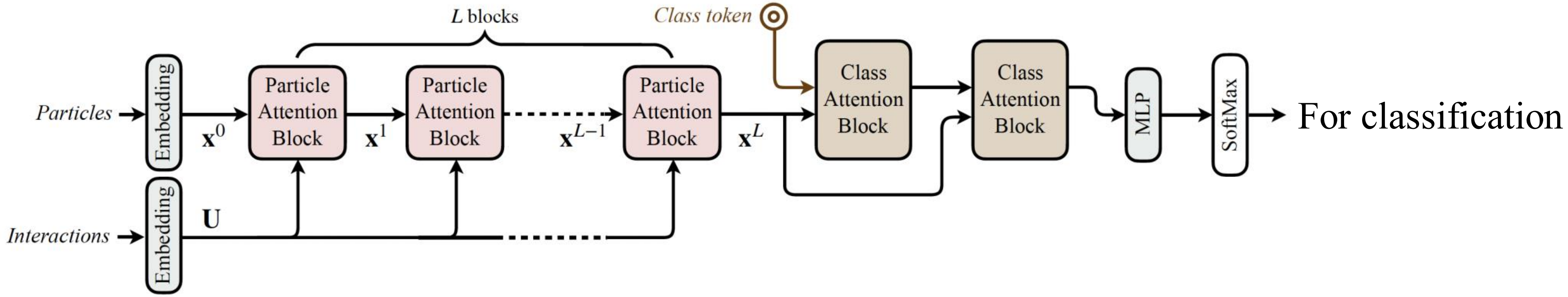
loss = $\alpha *$ contrastive loss + $\beta \star$ classification loss

Pre-training Setup

Network parameters:

Encoder	Particle Embedding	Pairwise Embedding	Particle Attention Block	Heads	Class Attention Block	FC	Projection
Jet-level	(512, 128, 512)	(64, 64, 64, 8)	8	8	2 (for classification)	(512, 188)	-
					2 (for contrastive)	(512,512,512)	512
gen-level	(64, 64, 64)	(32, 32, 32, 4)	4	4	2	(256)	512

Structure of Jet-level:



Training parameters(*Sophon++*):
α=0.5, β=1, s=0.4
Batch size=2048
Learning rate=5×10^{-3}
Steps per epoch=5000 (1024 ∗ 10,000/2048)
Epoch=160
Use NCCL on 4 GPU

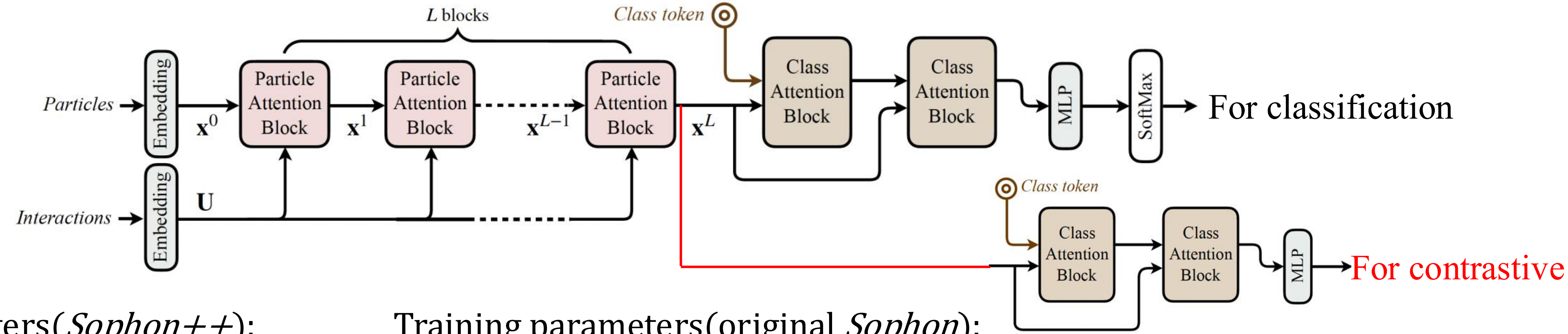
Training parameters(original *Sophon*):
α=0, β=1
Batch size=2048
Learning rate=3×10^{-3}
Steps per epoch=5000
Epoch=160
Use NCCL on 4 GPU

Pre-training Setup

Network parameters:

Encoder	Particle Embedding	Pairwise Embedding	Particle Attention Block	Heads	Class Attention Block	FC	Projection
Jet-level	(512, 128, 512)	(64, 64, 64, 8)	8	8	2 (for classification)	(512, 188)	-
					2 (for contrastive)	(512,512,512)	512
gen-level	(64, 64, 64)	(32, 32, 32, 4)	4	4	2	(256)	512

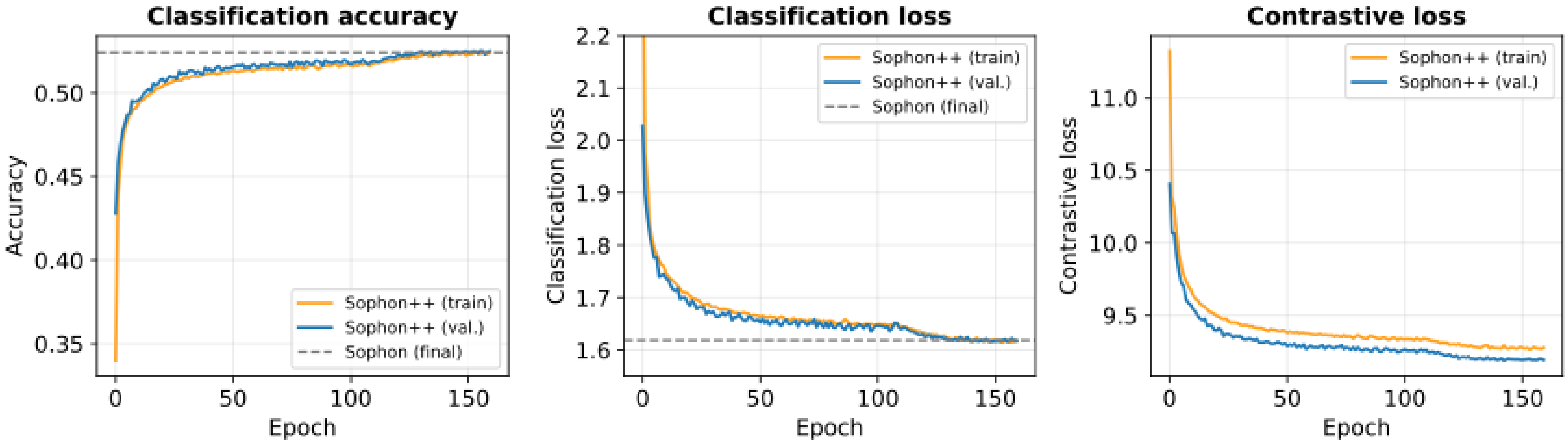
Structure of Jet-level:



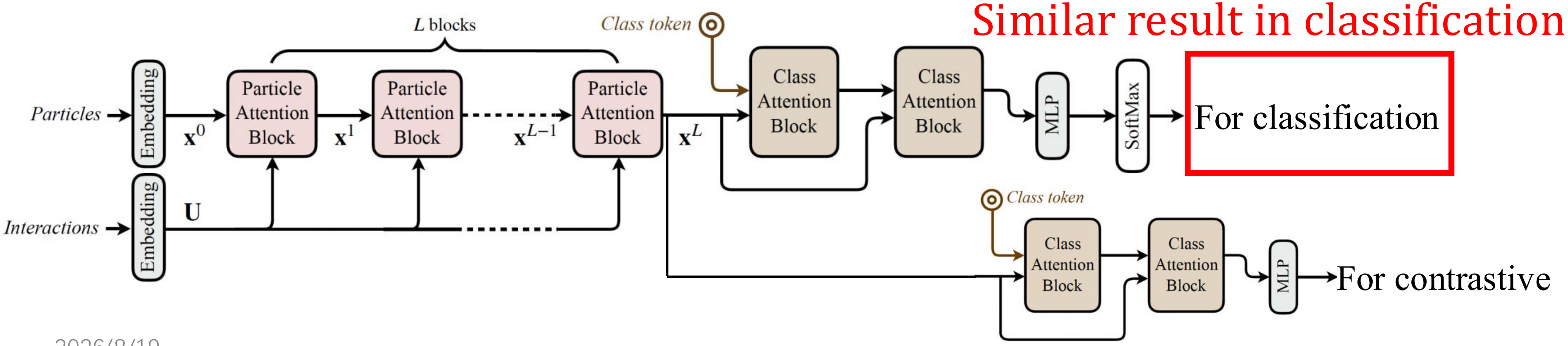
Training parameters(*Sophon++*):
α=0.5, β=1, s=0.4
Batch size=2048
Learning rate=5×10^{-3}
Steps per epoch=5000 (1024 ∗ 10,000/2048)
Epoch=160
Use NCCL on 4 GPU

Training parameters(original *Sophon*):
α=0, β=1
Batch size=2048
Learning rate=3×10^{-3}
Steps per epoch=5000
Epoch=160
Use NCCL on 4 GPU

Result of classification (188 categories)

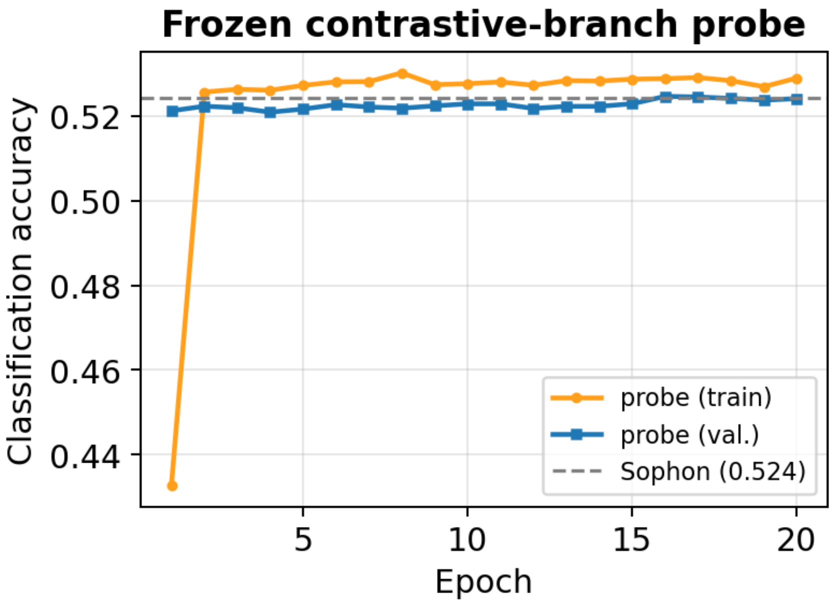


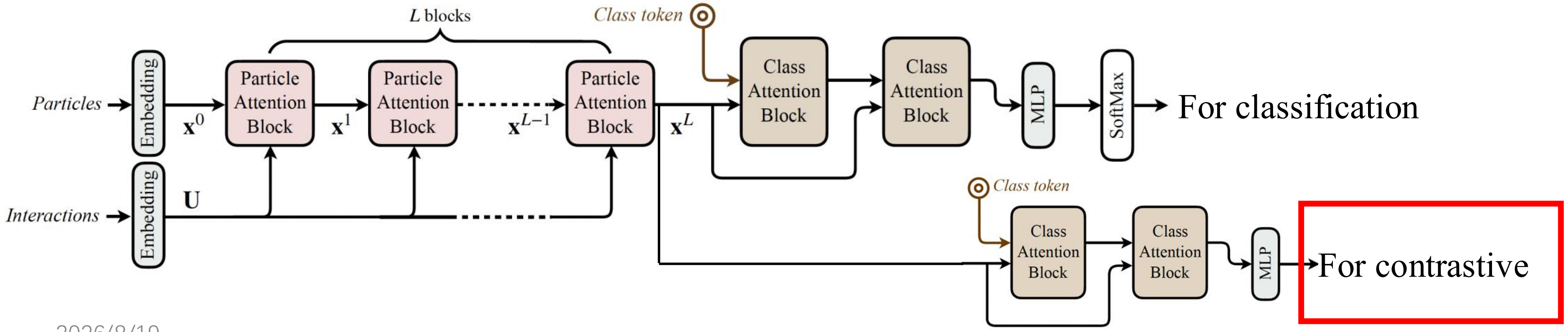
Similar result in classification task



Result of classification (188 categories)

- Transfer-learn the contrastive branch onto the 188-class classification task
- Freeze the full backbone; train only a Sophon-identical FC layer (512, dropout 0.1) + classification head
- Probe accuracy reaches 0.5246, matching Sophon (0.524) and the Sophon++ classification branch (0.525), at the same loss (~1.62)
- Contrastive representation is 'classification-complete': full supervised info is pretained.



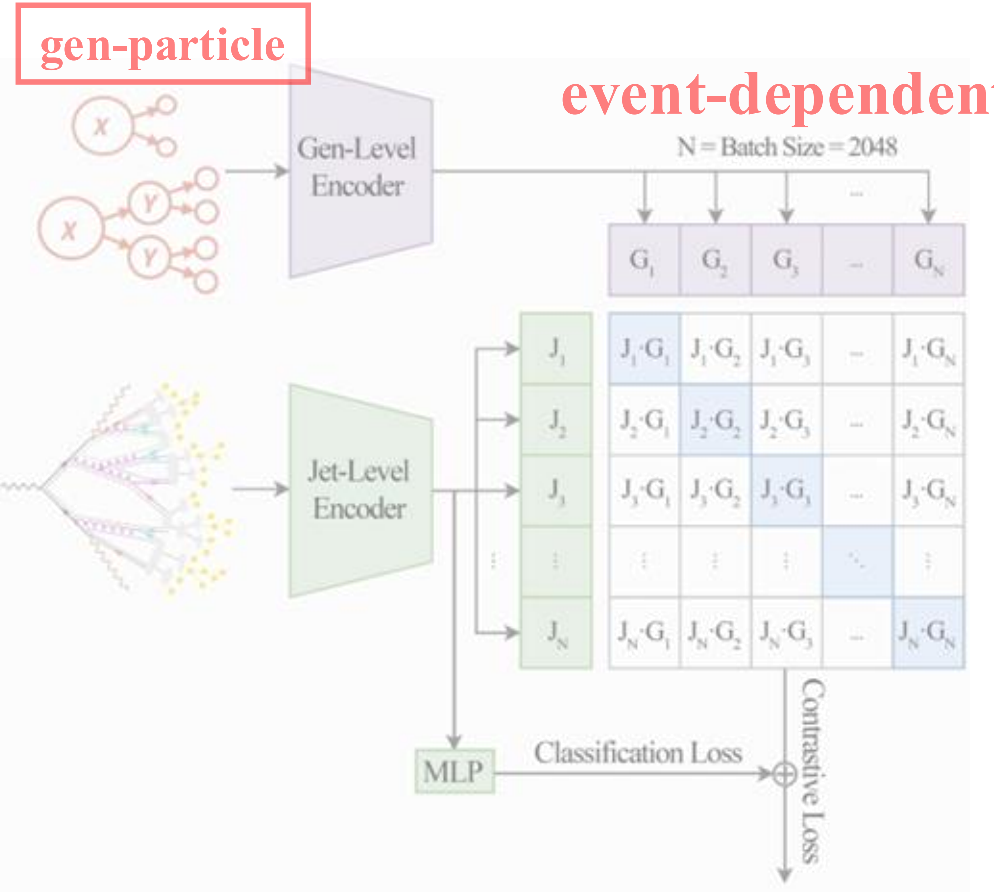


contrastive branch is 'classification-complete'

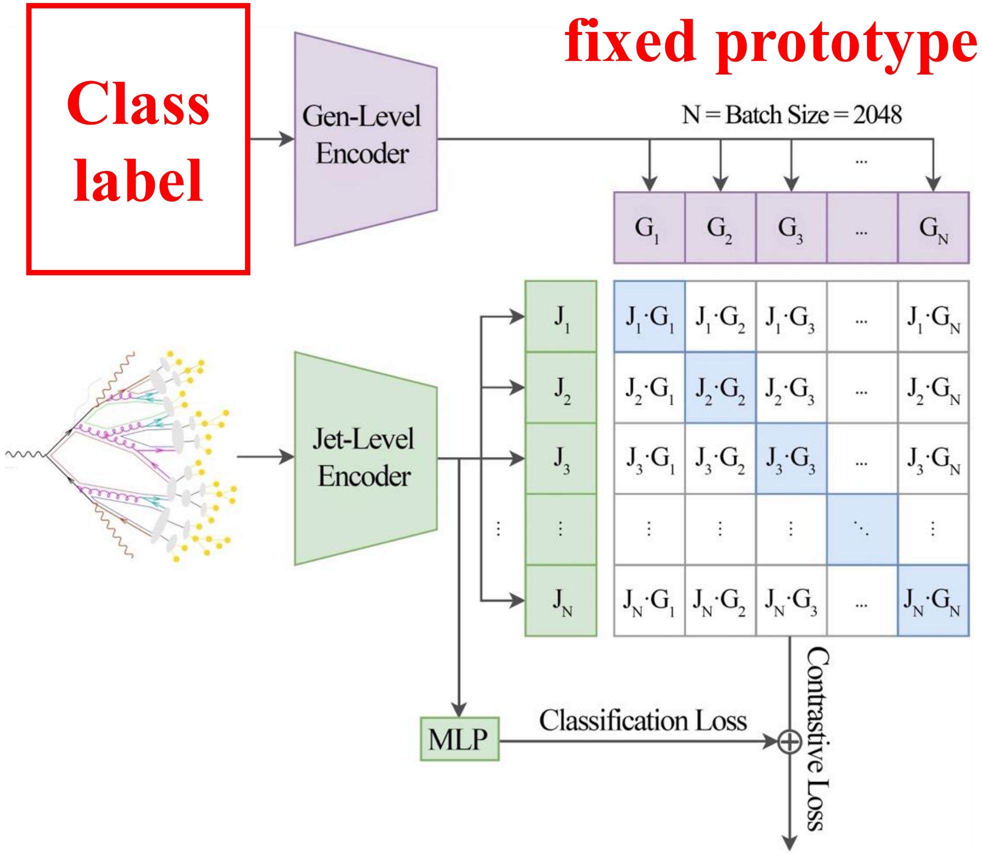
When Contrastive Learning Becomes Classification

(The label-only limit: theory and experimental validation)

What remains if the gen input contains only the label?



Sophon++



Label-only

Is the model still learning gen–reco correspondence, or has it become a classifier?

➡ **“matching a jet to a gen event” becomes “matching a jet to a class prototype”**

Label-only limit

Standard classification

$$h^*(x, y) = \log p\,(y|x) + a(x)$$

Label−only contrastive learning

$$s^*(x, y) = \log p\,(y|x) - \log \nu\,(y) + C$$

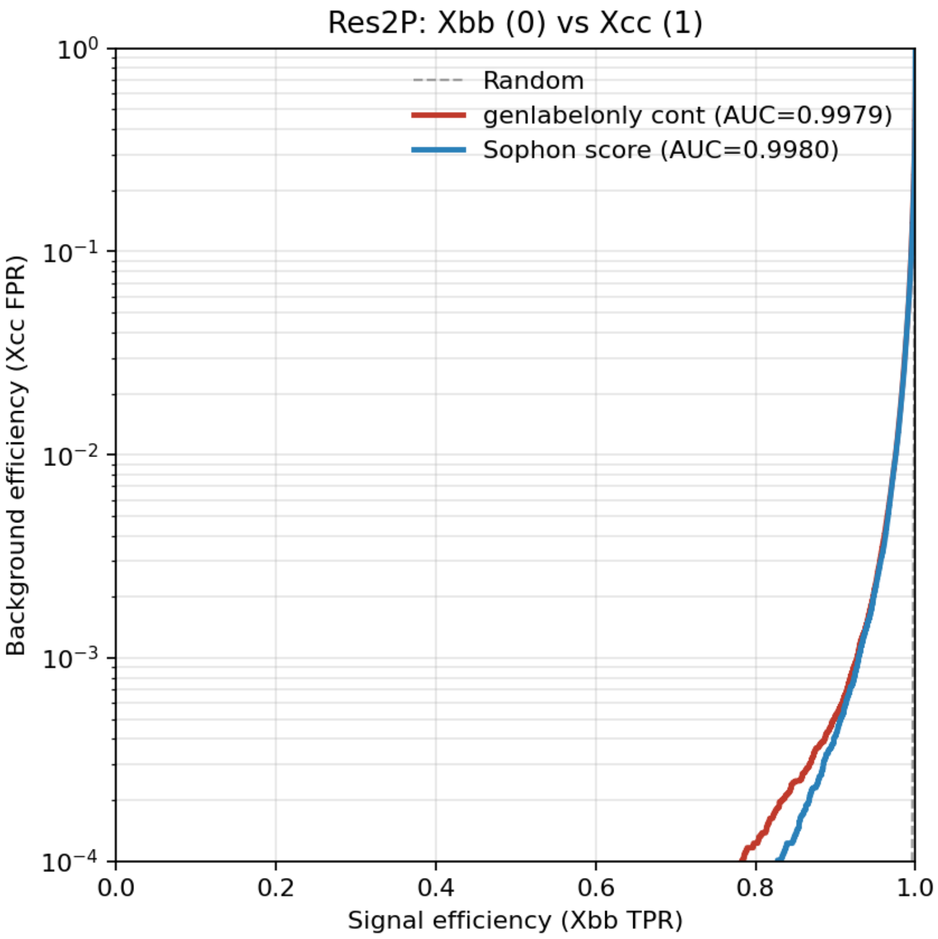
Notation

(x): reconstructed jet input.
$(y \in 1, \ldots, K)$: jet class label.
$(\nu(y))$: frequency with which class (y) is sampled as a candidate in the training batches.
$(h^*(x, y))$: theoretically optimal classification logit for class (y).
$(s^*(x, y))$: theoretically optimal contrastive similarity score between jet (x) and label (y).
$(a(x))$: an additive term shared by all classes for the same jet; it does not affect the class softmax.

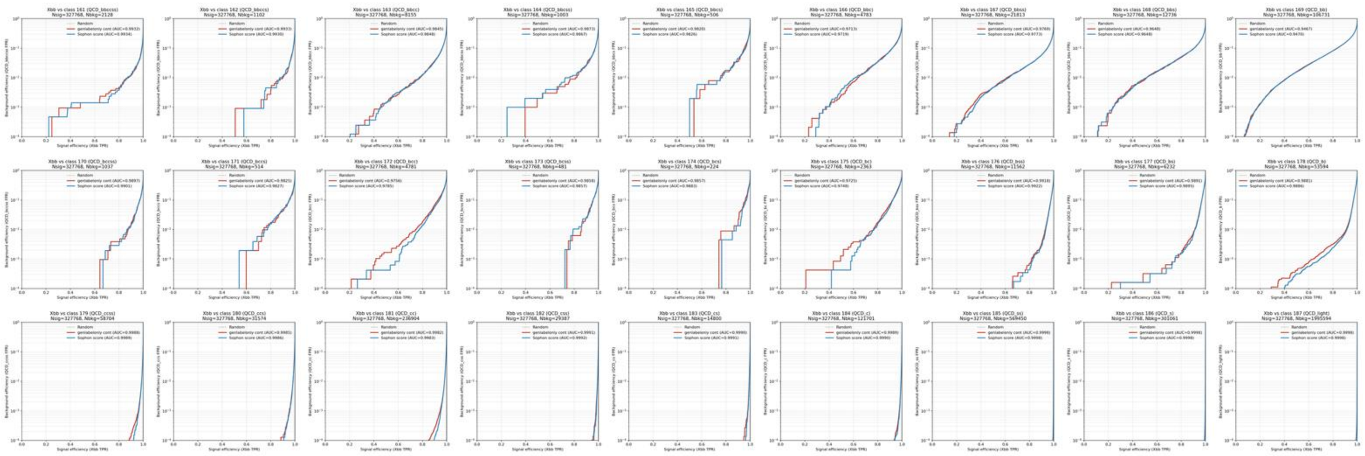
1. Both objectives learn essentially the same class
2. Contrastive learning removes the effective class prior
3. With class-balanced sampling, the two optimal scores differ only by an irrelevant constant.
4. The reverse contrastive objective removes $a(x)$, leaving only a global additive constant.

Therefore, the two models should give very similar event rankings,
but their class-to-class calibration may differ.

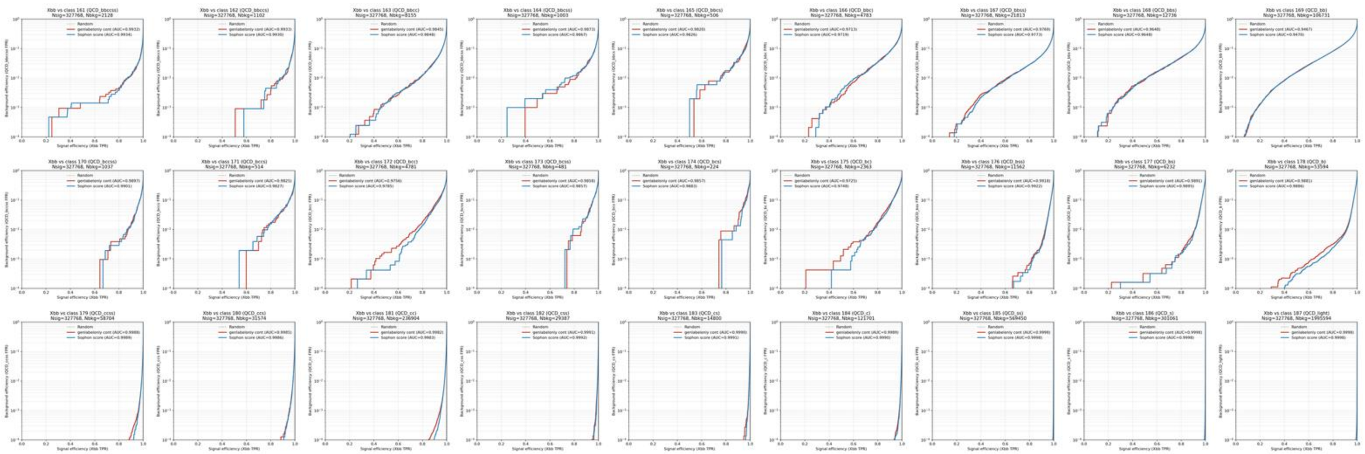
Label-only ROC per-class performance



X_{bb} vs X_{cc}



X_{bb} vs QCD sub-class (27 classes)

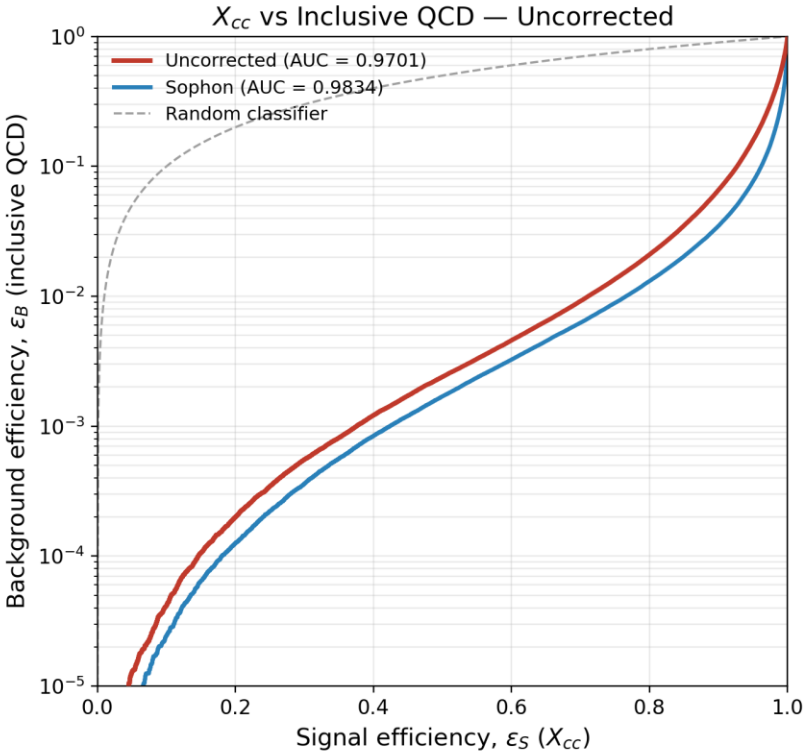


X_{cc} vs QCD sub-class (27 classes)

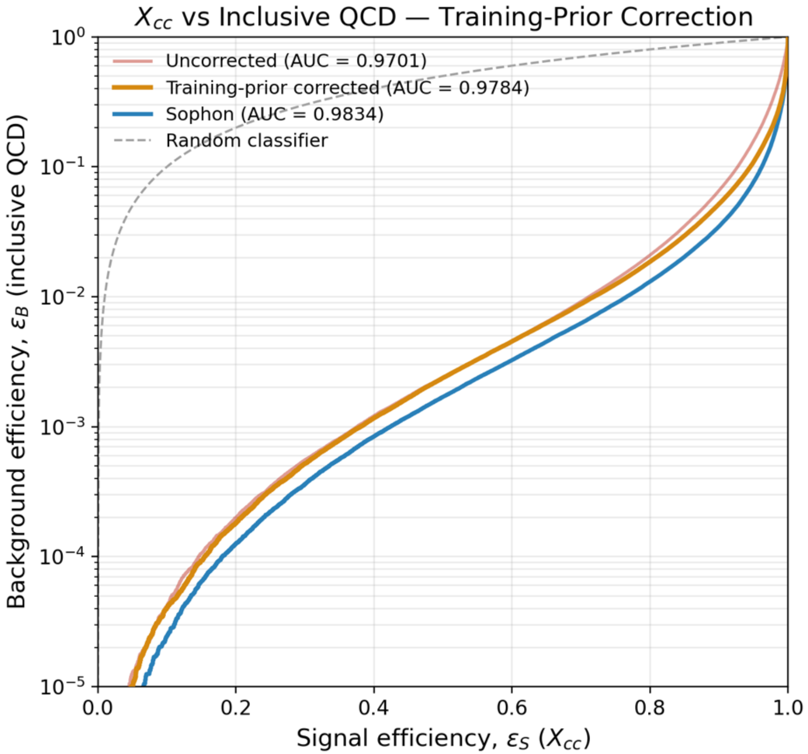
Full ROC shapes remain consistent across all tested QCD subclasses

Label-only ROC discriminant performance

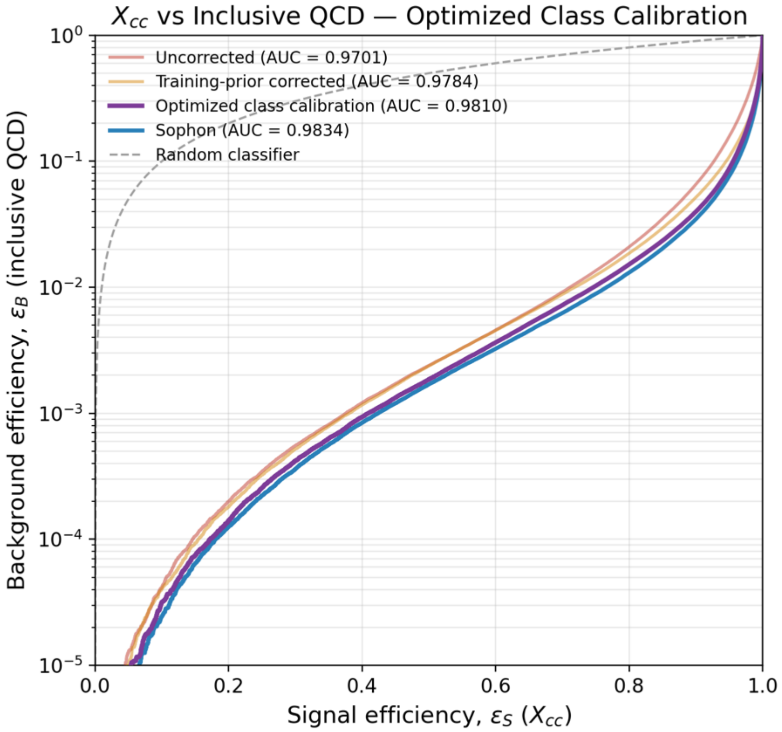
$$D_{\text{Sophon}} = \frac{p_s}{p_s + \sum_q p_q} \quad \text{vs} \quad D_{\text{cont}} = \frac{e^{s_s}}{e^{s_s} + \sum_q \omega_q\, e^{s_q}}$$



Uncorrected: $\omega_q = 1$



Theoretical-prior correction: $\omega_q = \nu_q/\nu_{Xcc}$



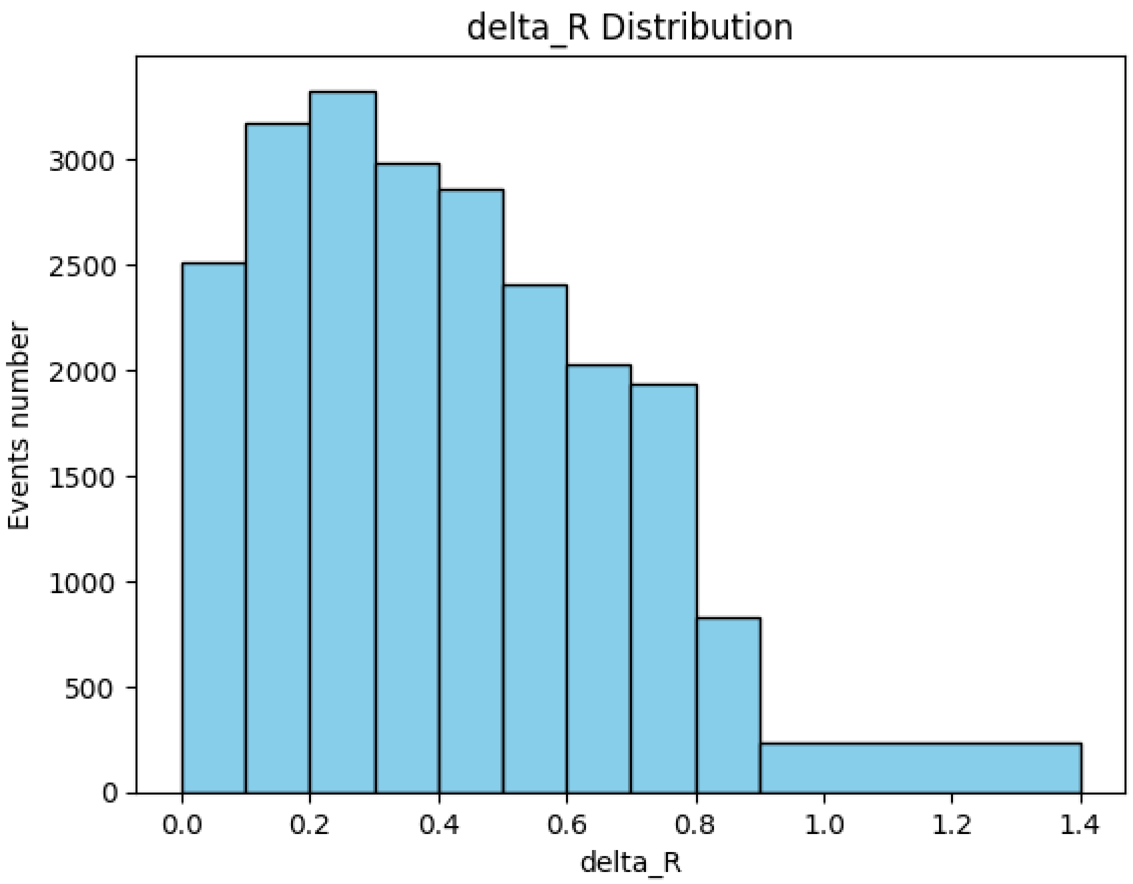
Optimized class calibration： $\omega_q = \omega_q^{\text{opt}}$

Label-only contrastive learning behaves as an approximately calibrated form of multiclass classification

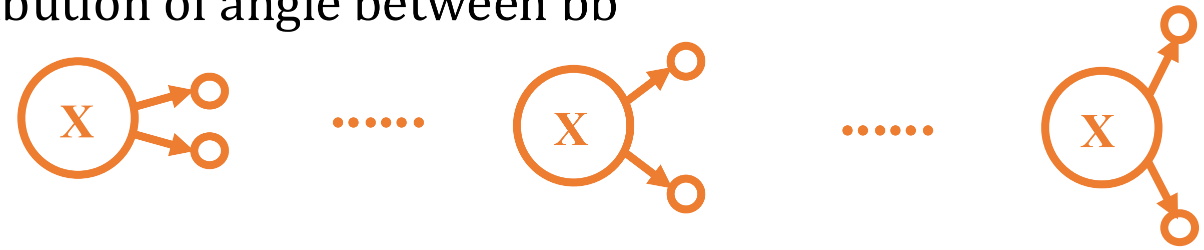
Downstream Verification Experiments

(Generalization: To testify global optimization of *Sophon++*)

Verification Experiment 1



Distribution of angle between bb



Modifying data config(yaml card):

- Only input events labeled as 'X to bb'
- Define new variables to calculate bb_deltaR
- Assign indices for each event according to bb_deltaR
- Set the indices as the new label

0.0~0.1: 0
0.1~0.2: 1
...
0.8~0.9: 8
>0.9: 9

A 10-class classification model

Verification Experiment 1

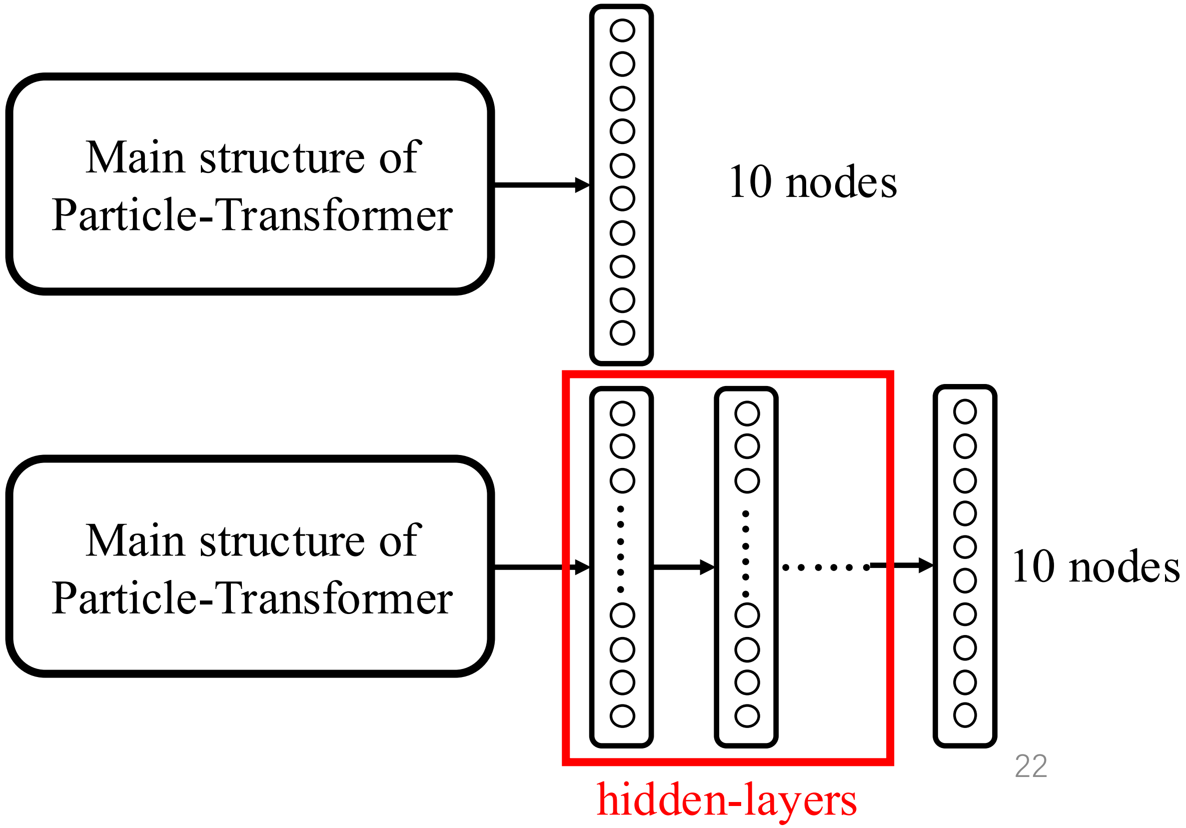
Universal model selection:

	Model	Epoch Selection	Accuracy	Classification Loss	Contrastive Loss
	Sophon	158	0.5241	1.619	-
Sophon++	classification branch	159	0.5249	1.618	-
	contrastive branch		-	-	9.191

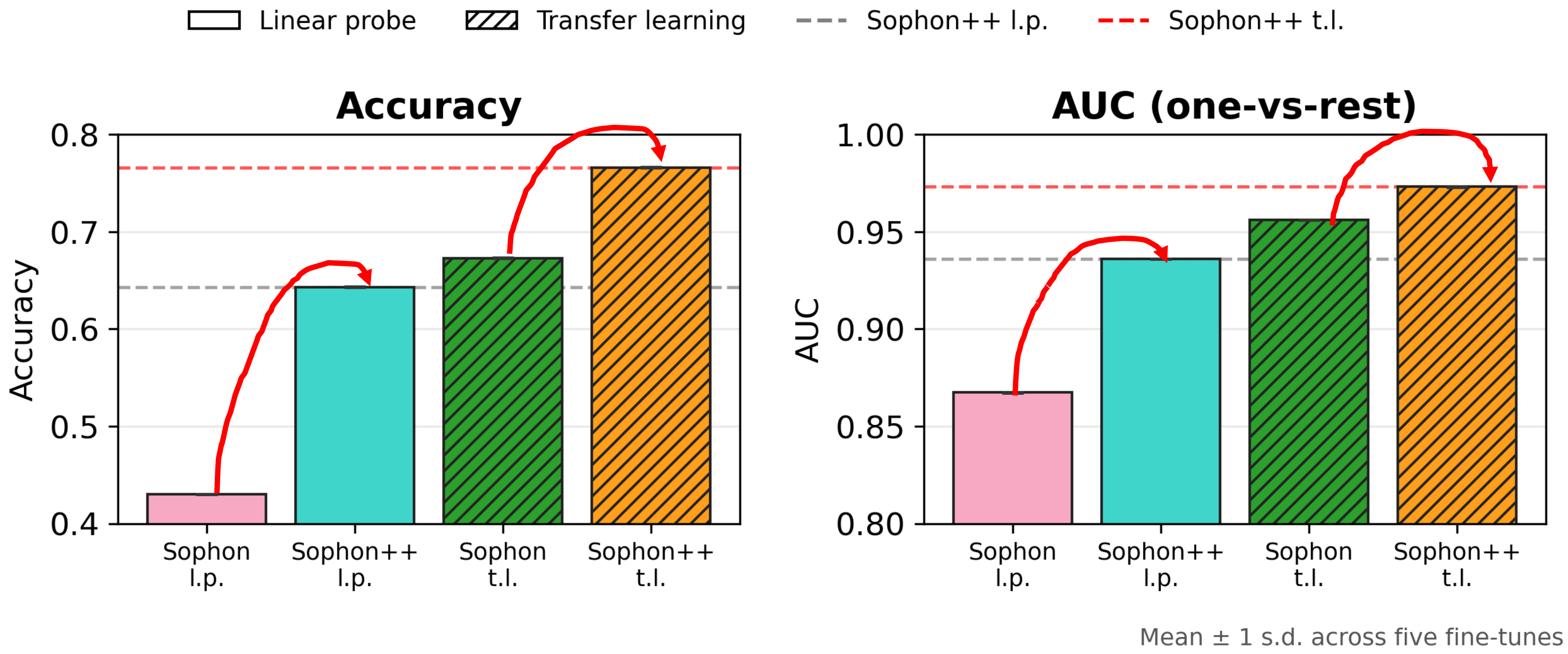
Training parameters:
Batch size=1024
Learning rate=1×10^{-3}
Steps per epoch=5000 ($1024 * 5{,}000/1024$)
Epoch=20 (10 for Linear-probe)
Linear-probe (train a one-layer FC with 10 nodes)
Transfer-learning (train a best FFN to get highest accuracy)
1 GPU per training

Linear-probe

Transfer-learning

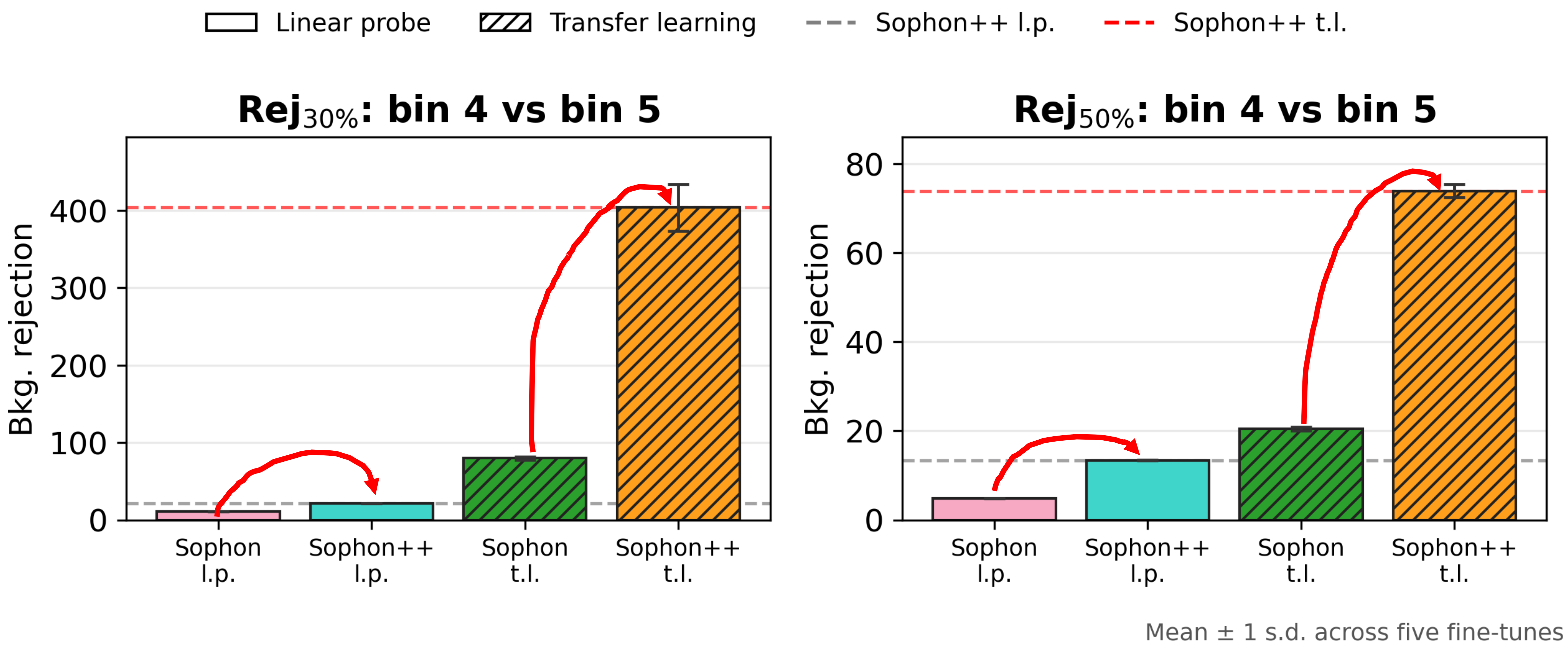


Verification Experiment 1



performance of *Sophon++* is much better than Original Sophon!

Verification Experiment 1



performance of *Sophon++* is much better than Original Sophon!

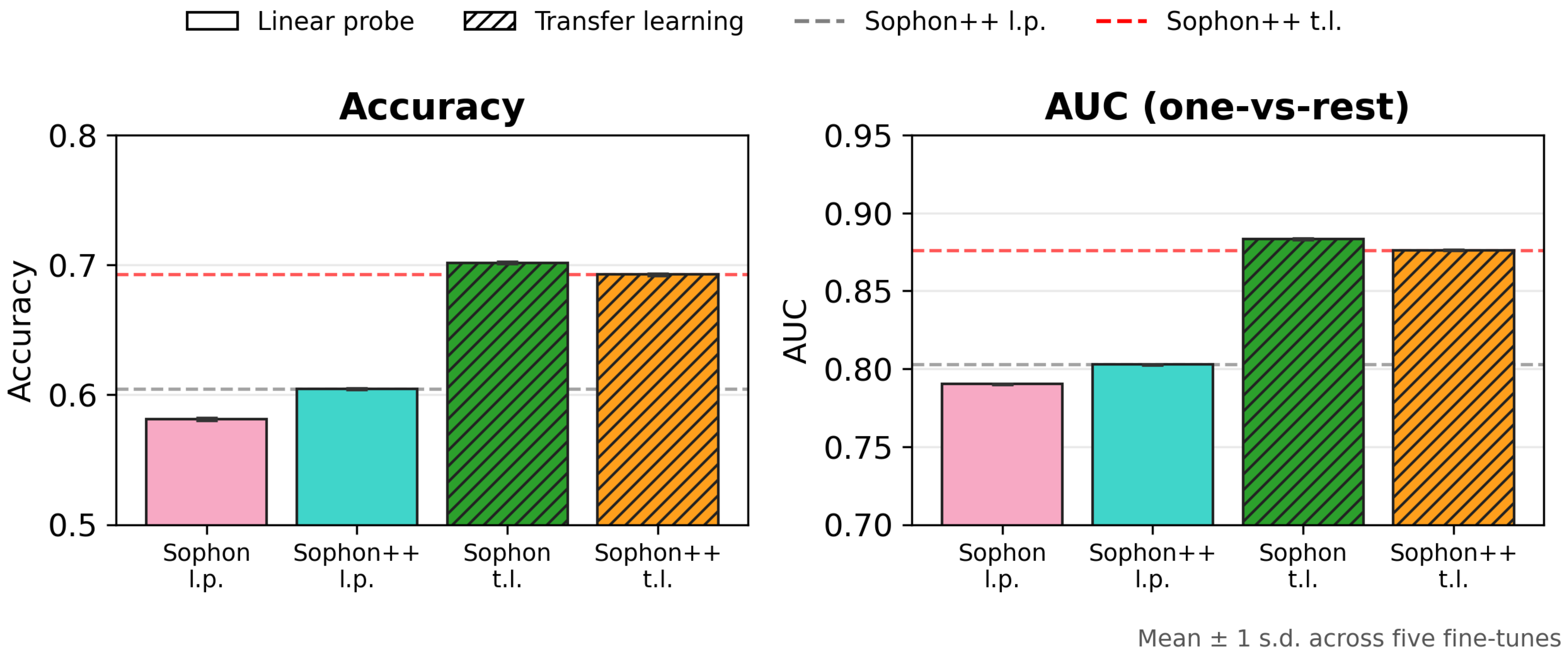
Verification Experiment 2: hadronic $W^{\pm}$ tagging

- While Lorentz-boosted hadronic W/Z jet tagging has been considered in Run 1, recent deep learning advances now enable discrimination of the W boson charge

- traditional approach: uses jet charge (p_T weighted particle charge) as a discriminant
- opportunity: integrate charge determination into the standard deep learning jet tagging framework

  e.g. as explored in [PRD 101, 053001 (2020)]

Verification Experiment 2: hadronic $W^{\pm}$ tagging

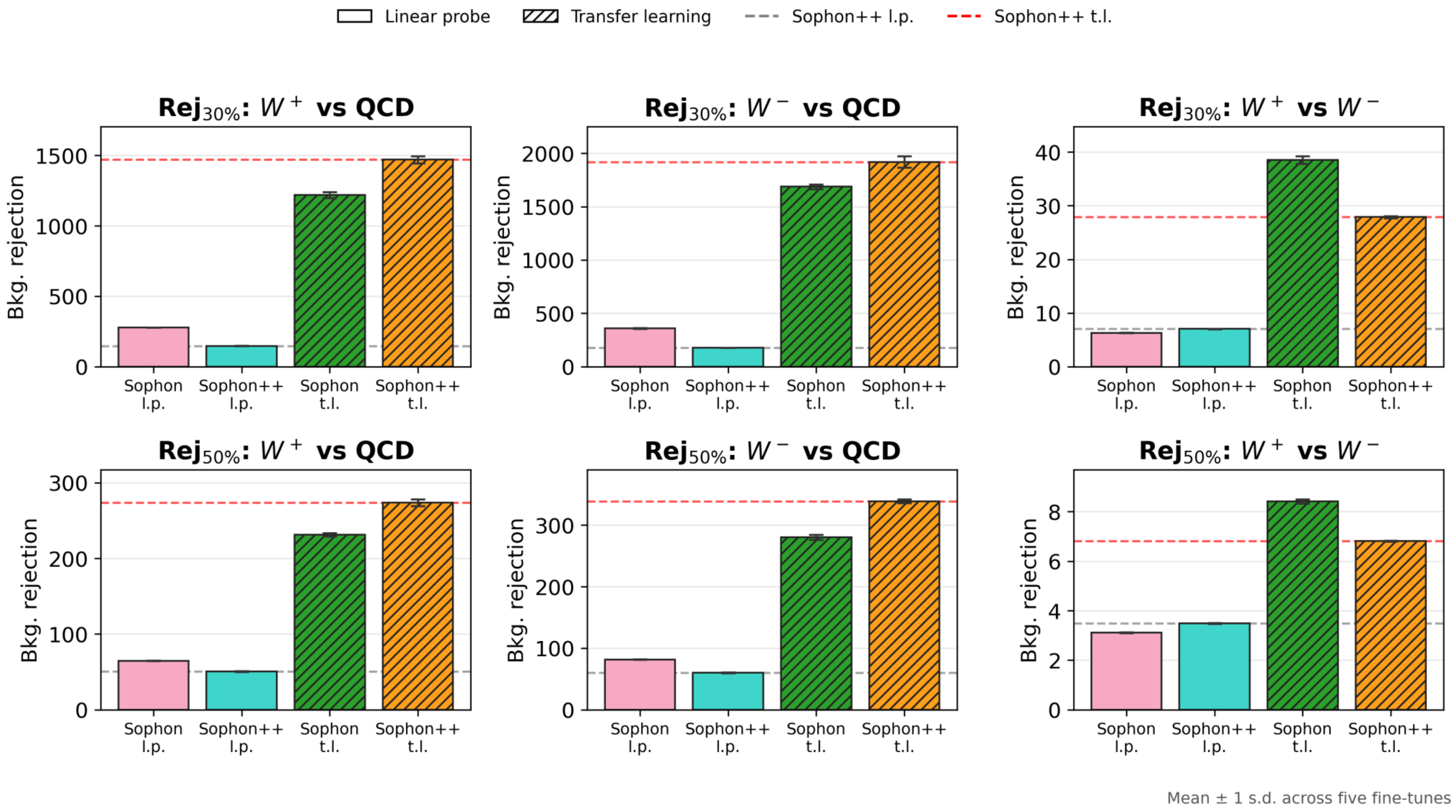


performance of *Sophon++* is better than Original Sophon in linearprobe

Verification Experiment 2: hadronic $W^{\pm}$ tagging

Rej30%

Rej50%



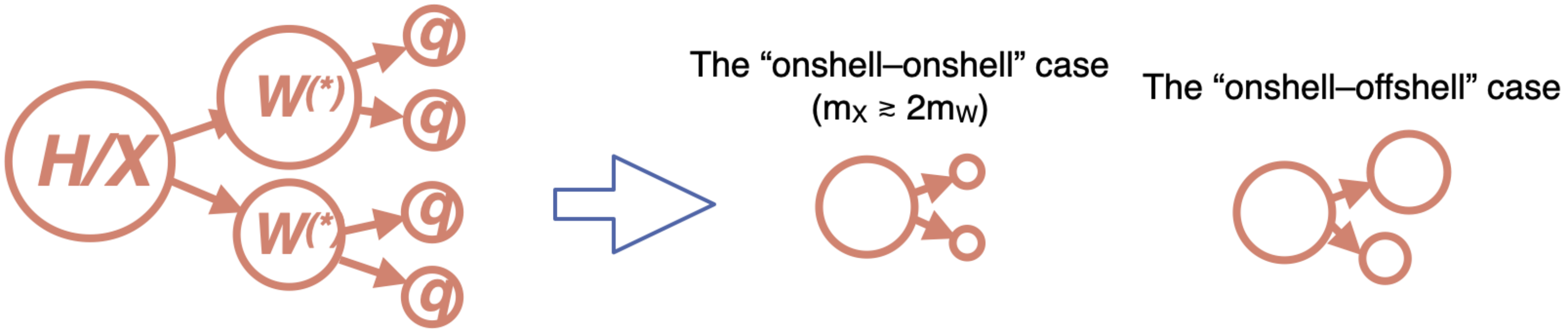
W^{+} vs QCD

W^{-} vs QCD

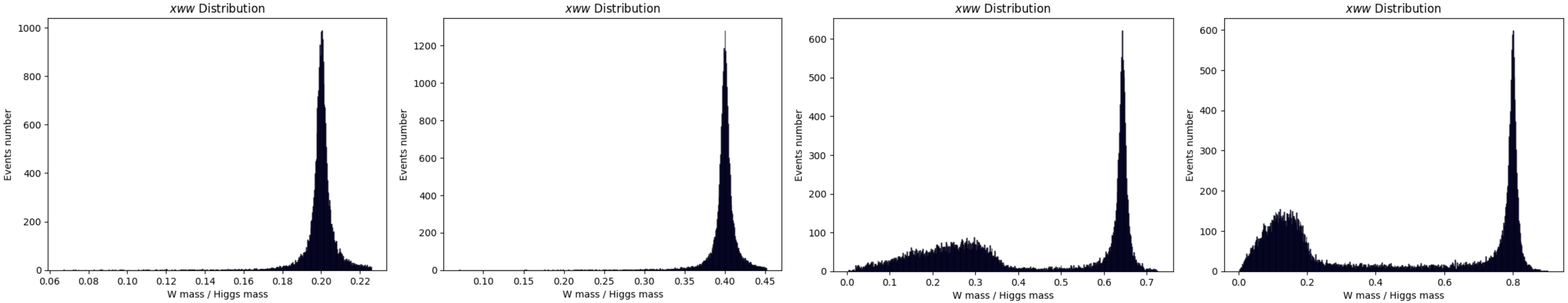
W^{+} vs W^{-}

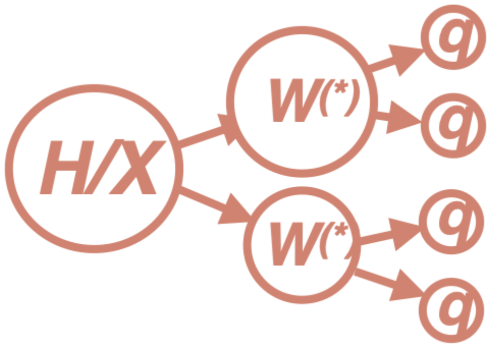
Verification Experiment 3: X→WW(*) tagging

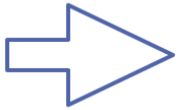
- Discriminating $X \to WW^{(*)}$ signals across varying m_W/m_X and against QCD backgrounds
- Exploring $H/X \to WW^{(*)}$ in a fully-merged topology is an emerging LHC direction
  - multiple CMS results now available: SM (single/di-Higgs), and BSM searches
    - SM HH→bbWW*(4q) [CMS-PAS-HIG-23-012], H→WW* [CMS-PAS-HIG-24-008]
    - Resonant X→H(bb)Y(WW) [CMS-PAS-B2G-23-007]
- Different m_W/m_X ratios probe distinct phase space regions → require separately optimized discriminants
- We consider four benchmark scenarios
  - $m_W/m_X = 0.2, 0.4, 0.6432$ (SM case), 0.8



Verification Experiment 3: X→WW(*) tagging

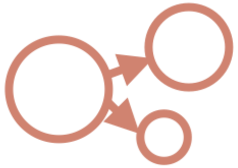




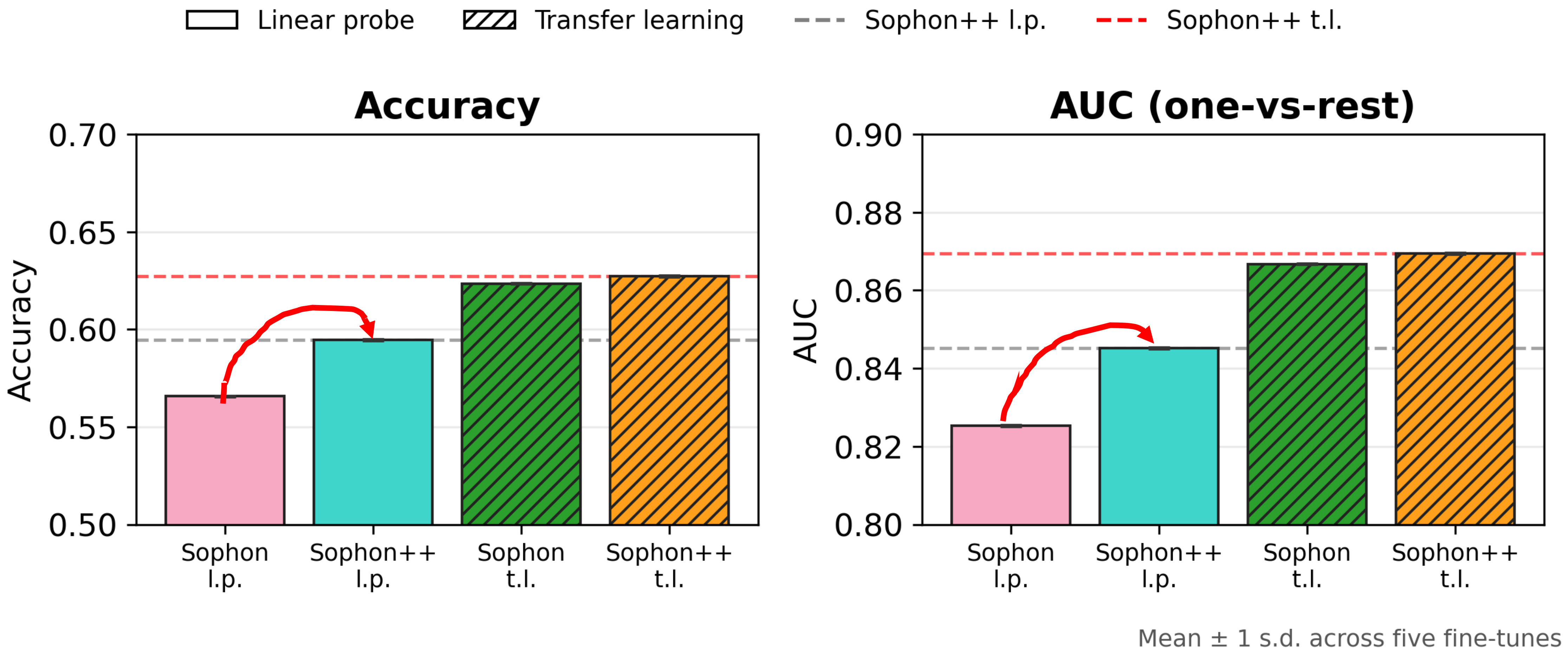


The “onshell–onshell” case
($m_X \gtrsim 2m_W$)

The “onshell–offshell” case



Verification Experiment 3: X→WW(*) tagging

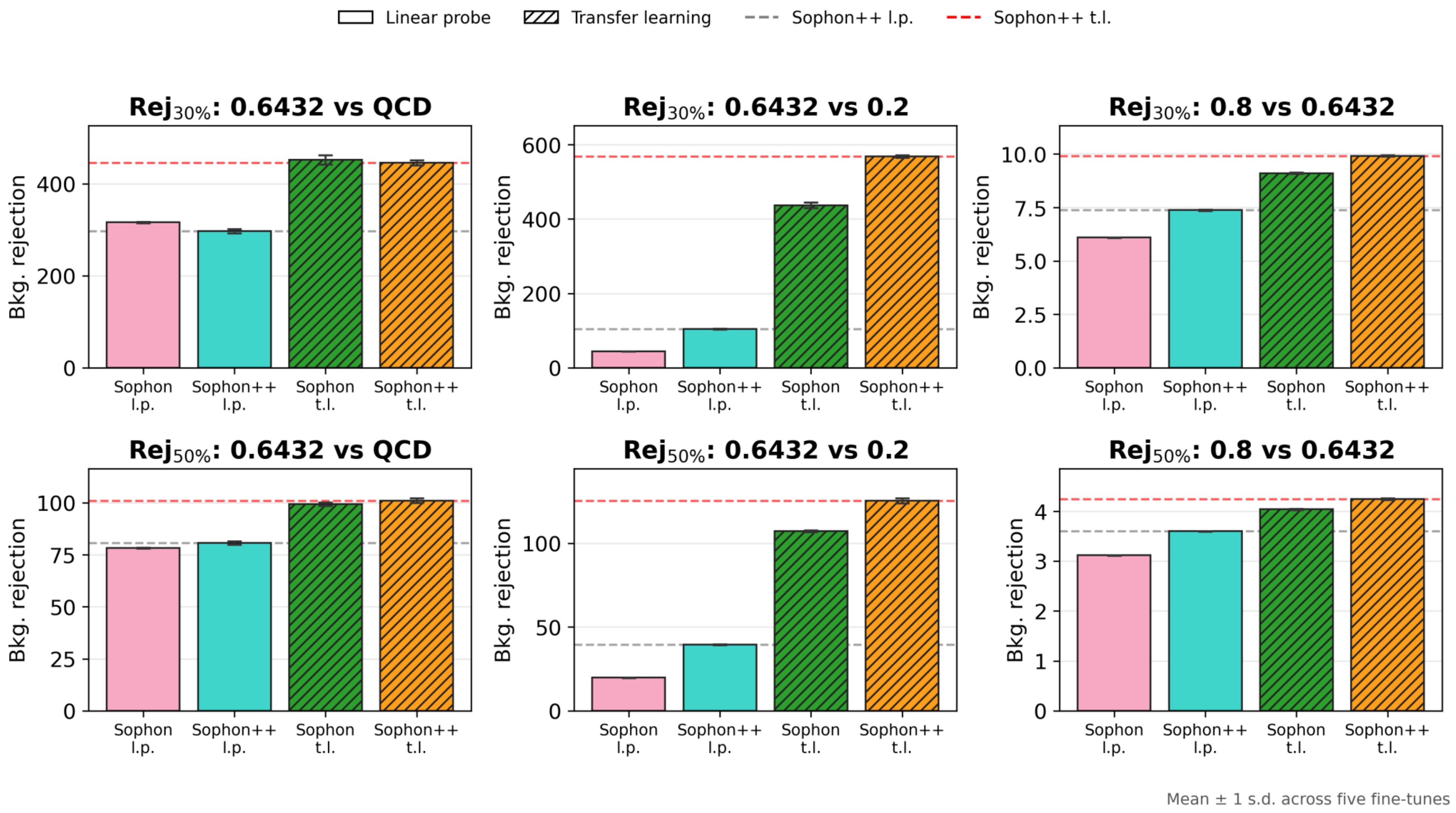


performance of *Sophon++* is better than Original Sophon

Verification Experiment 3: X→WW(*) tagging

Rej30%

Rej50%



0.6432 vs QCD

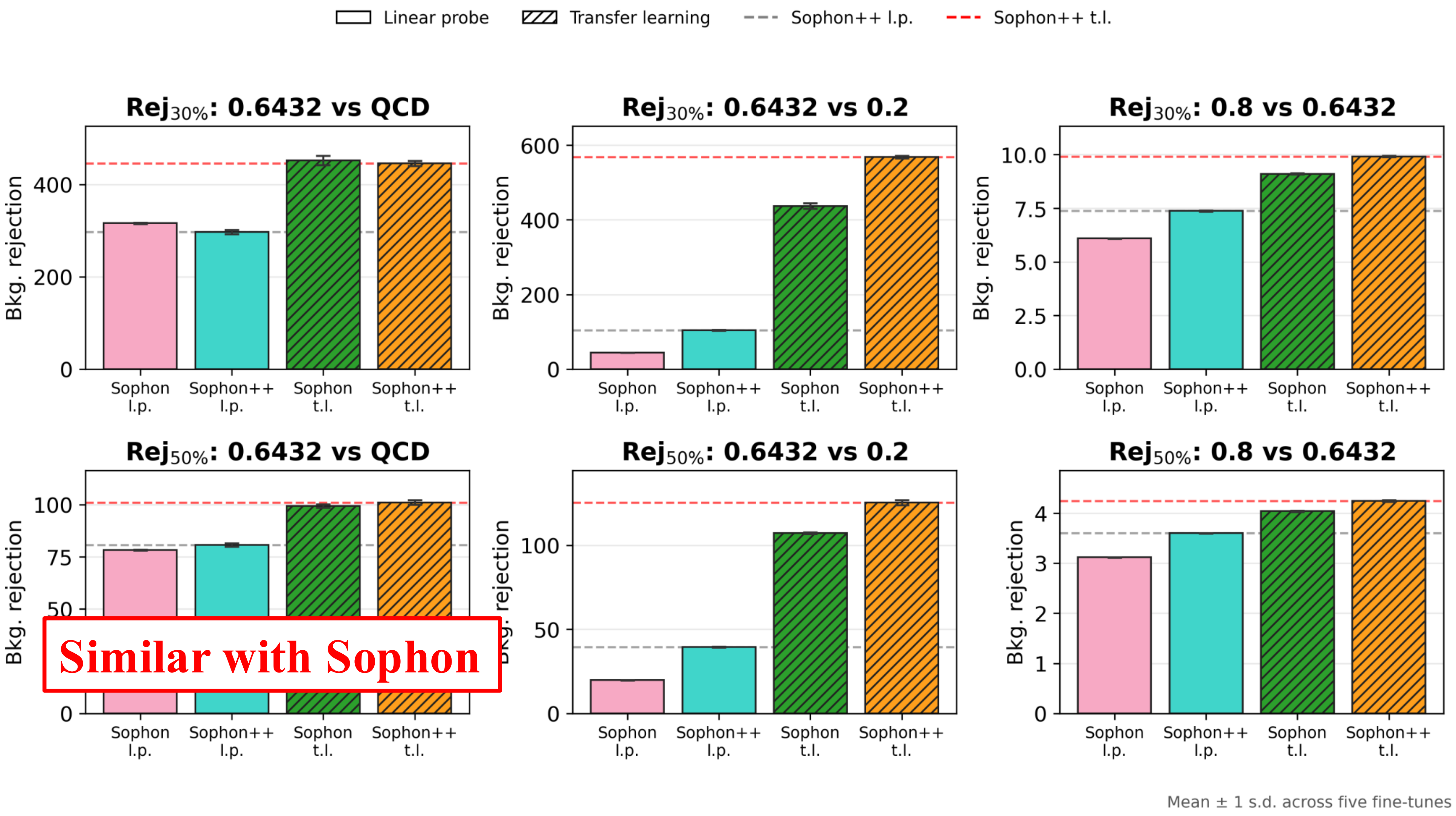
0.6432 vs 0.2

0.8 vs 0.6432

Verification Experiment 3: X→WW(*) tagging

Rej30%

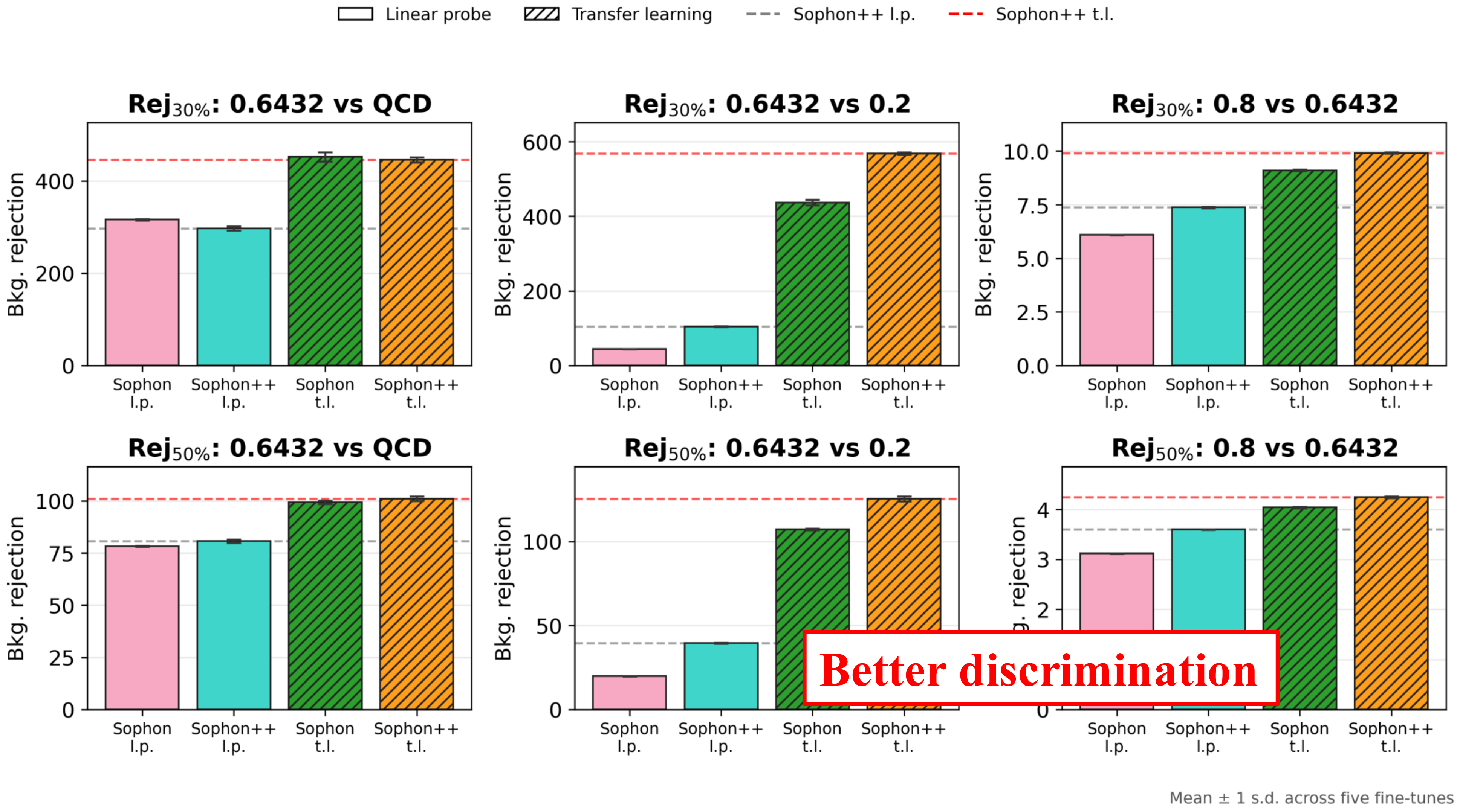
Rej50%



Similar with Sophon

0.6432 vs QCD | 0.6432 vs 0.2 | 0.8 vs 0.6432

Verification Experiment 3: X→WW(*) tagging



Rej30%

Rej50%

0.6432 vs QCD

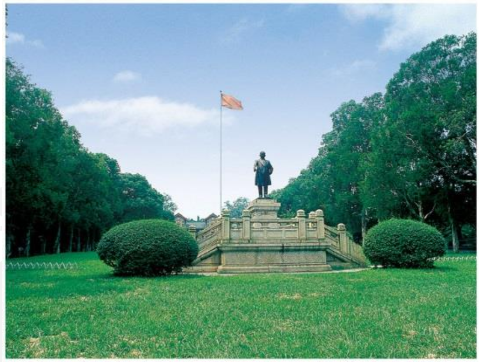
0.6432 vs 0.2

0.8 vs 0.6432

Conclusion

- ***Sophon***, as a large-scale multi-classifier, has achieved optimal performance across various specific tagging tasks.
  - note: It is implemented in CMS as part of GloParT and has been utilized for various boosted jet analyses in Run 3.
- How can we enhance its generalization capabilities?
  - we are exploring contrastive "gen-reco" matching.
  - while pure unsupervised models have been developed in recent years to create jet foundation models, we propose insights derived from computer vision history.
    - a "supervised + X approach" can be developed.
    - "X" refers to a methodology akin to OpenAI's CLIP: it continuously encodes different configurations of gen-level information into Sophon's latent space, thereby achieving stronger generalization.
  - demonstrated superior performance in specific fine-tuning tasks; further optimizations are underway.
- ***Sophon++*** provides a promising pathway to upgrade GloParT in CMS!

Thanks for your attention!



Backup: Core value of Sophon/GloParT at LHC physics

- *Sophon* method implies a pre-training philosophy: "large models for large-scale classification"
  - It has been applied in CMS as GloParT
- A brief summary: GloParT's role for CMS's joint Run2+Run3 analyses in the coming years:
  - **Greater sensitivity improvements for the planned analyses**
    - H/HH/BSM search related to H→bb/cc/WW*/ττ…, analyses requiring W/Z/t tagging, …(several works now using GloParTv1/2/3)
  - **Expanding the landscape of boosted-topology search**
    - Novel final states to be explored! - a reminder to investigate calibration feasibility and discuss with BTV+JME
  - **Creating new paradigms in exploiting the AK8 jet model**
    - Fine-tune GloParT at the analyses level (design custom taggers); support broad MC-free searches
    - Facilitate anomaly detection via GloParT fine-tuning will be a great complement to current BSM programs

Backup: Visualization of output vectors