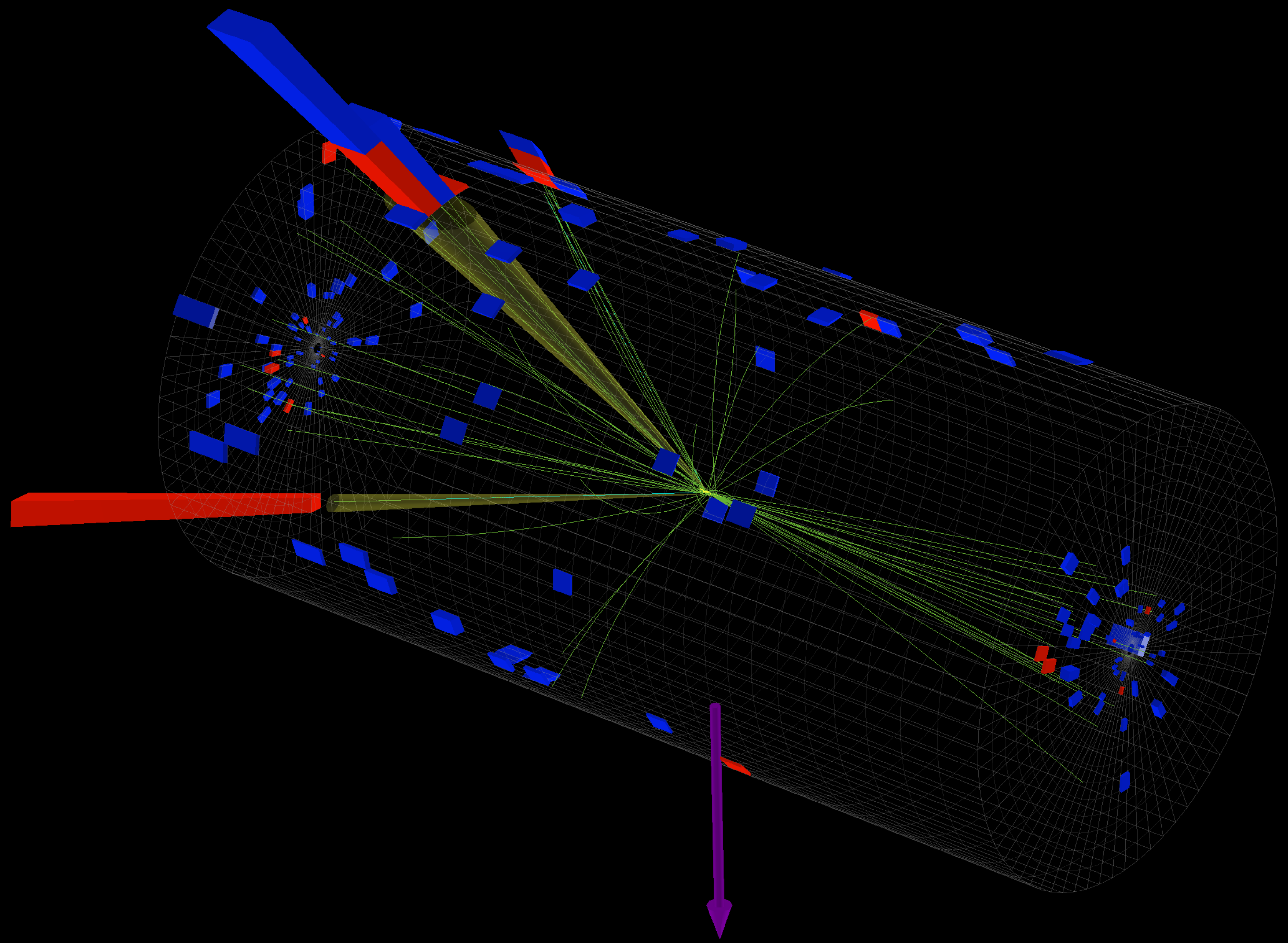
Towards a Foundation Model for Jet Physics

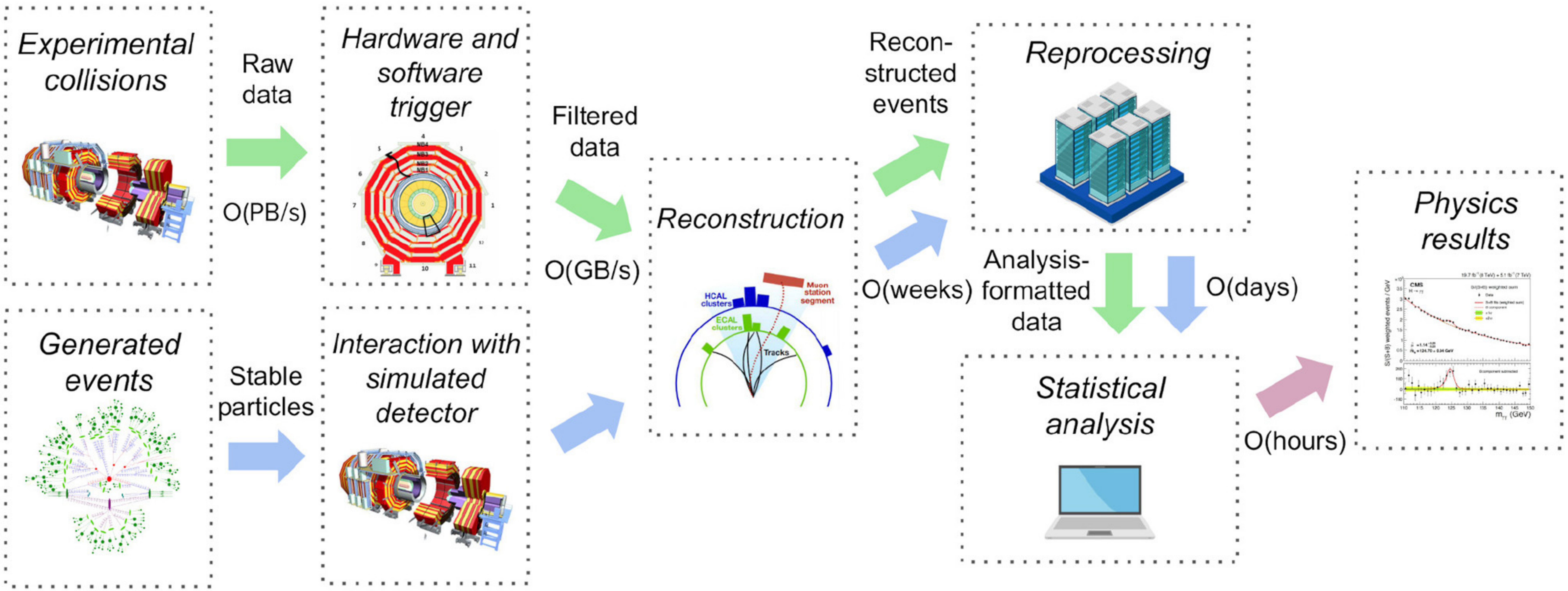
Huilin Qu (曲慧麟)

IHEP EPD Seminar

30.12.2024



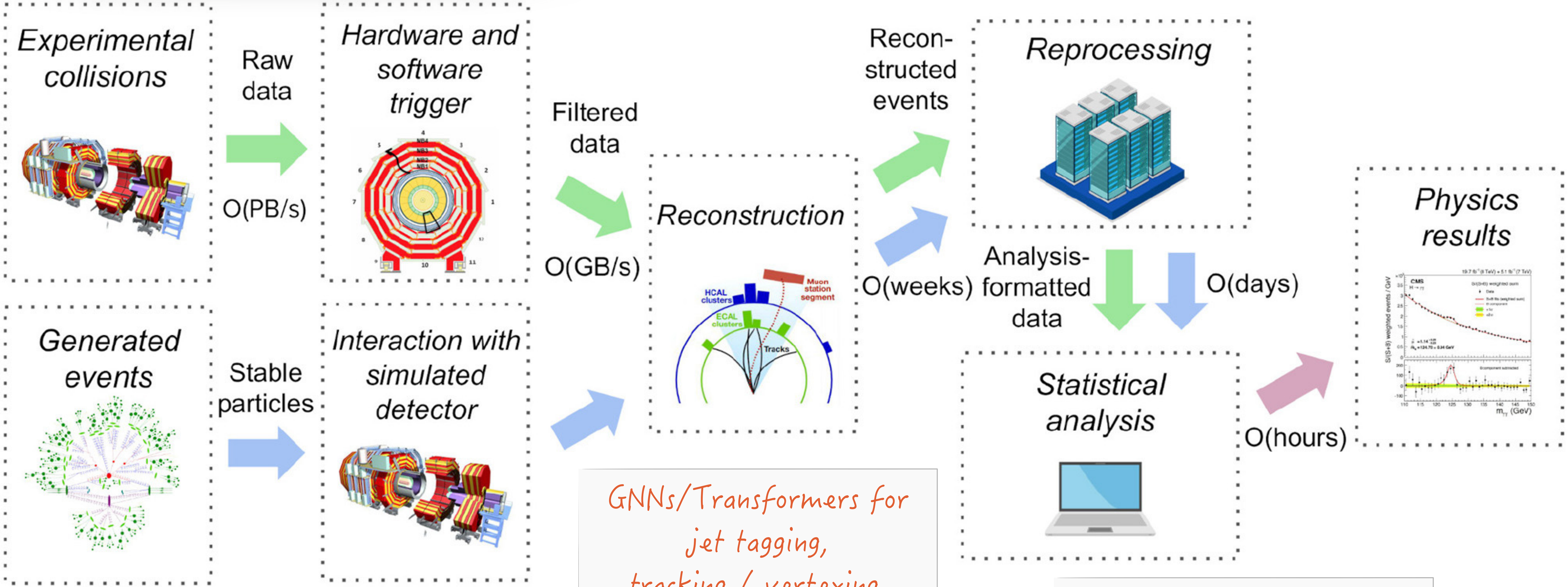
HEP DATA FLOW…



Front.Big Data 4 (2021) 661501

HEP Data Flow… Matched With ML!

Ultrafast inference (FPGA/ASIC), anomaly detection…



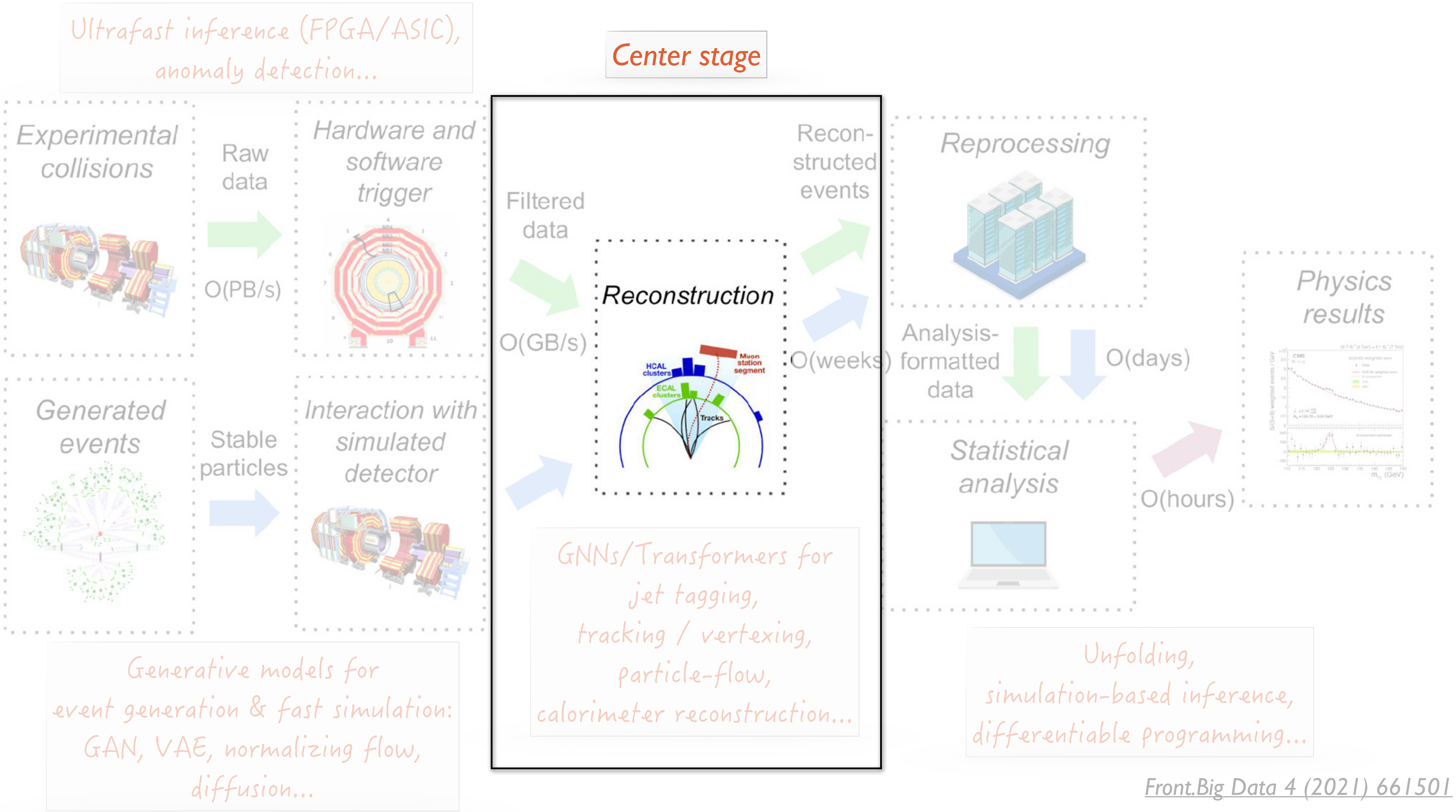
Generative models for event generation & fast simulation: GAN, VAE, normalizing flow, diffusion…

GNNs/Transformers for jet tagging, tracking / vertexing, particle-flow, calorimeter reconstruction…

Unfolding, simulation-based inference, differentiable programming…

Front.Big Data 4 (2021) 661501

HEP Data Flow… Matched With ML!

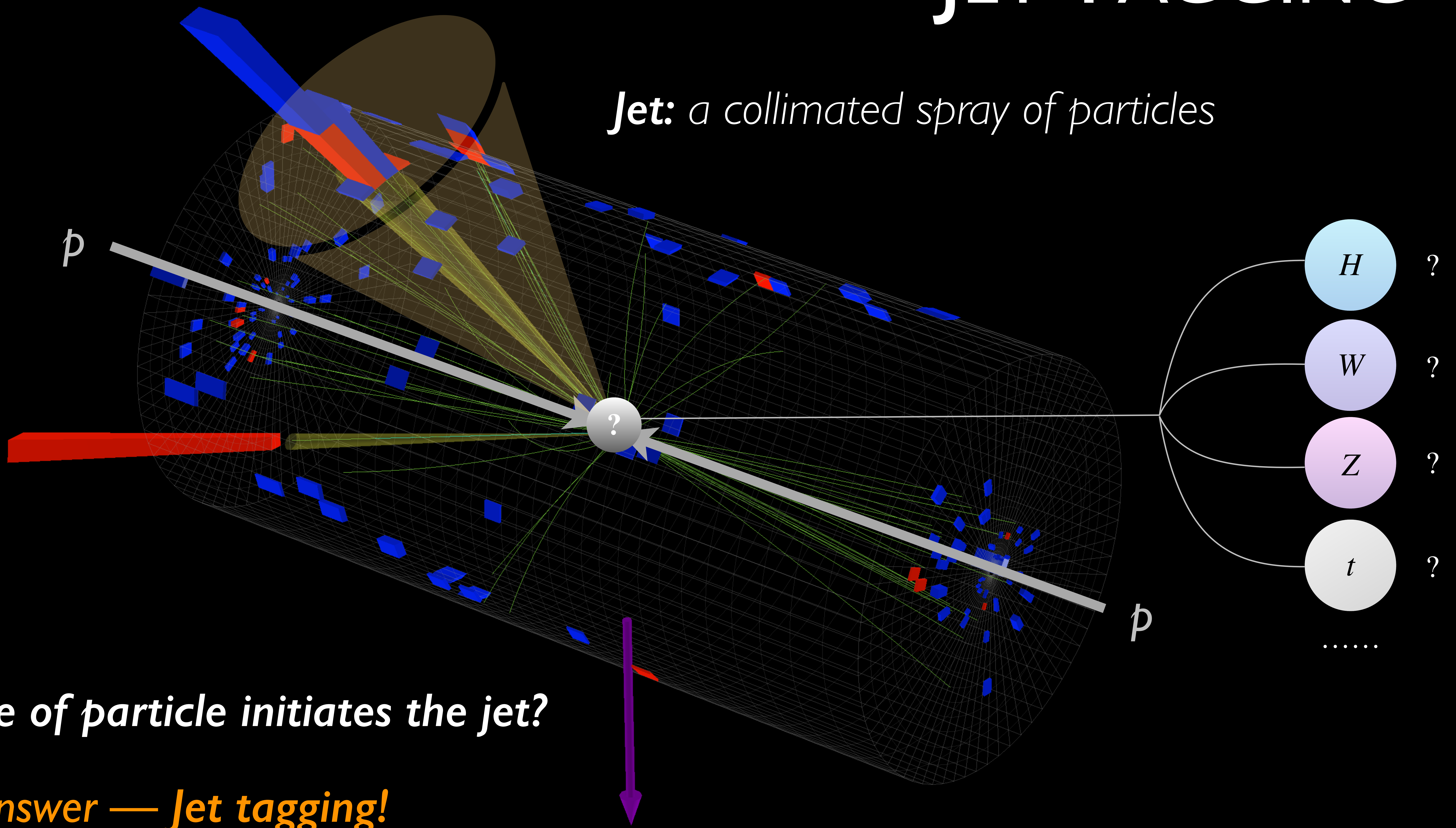


Front.Big Data 4 (2021) 661501

CMS Experiment at LHC, CERN
Data recorded: Sat Aug 5 15:32:22 2017 CEST
Run/Event: 300515 / 205888132

Jet Tagging

***Jet:** a collimated spray of particles*



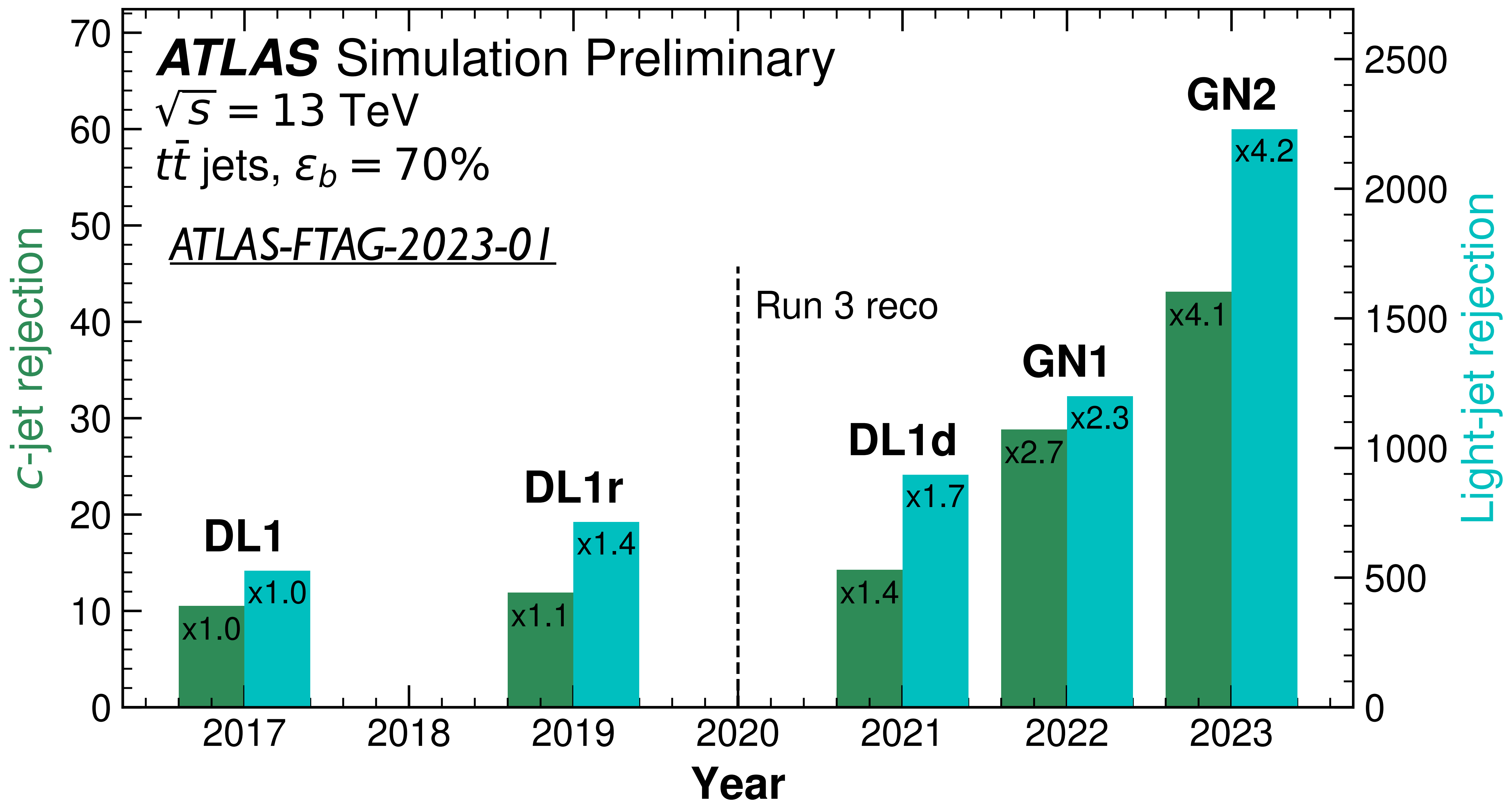
Key question:

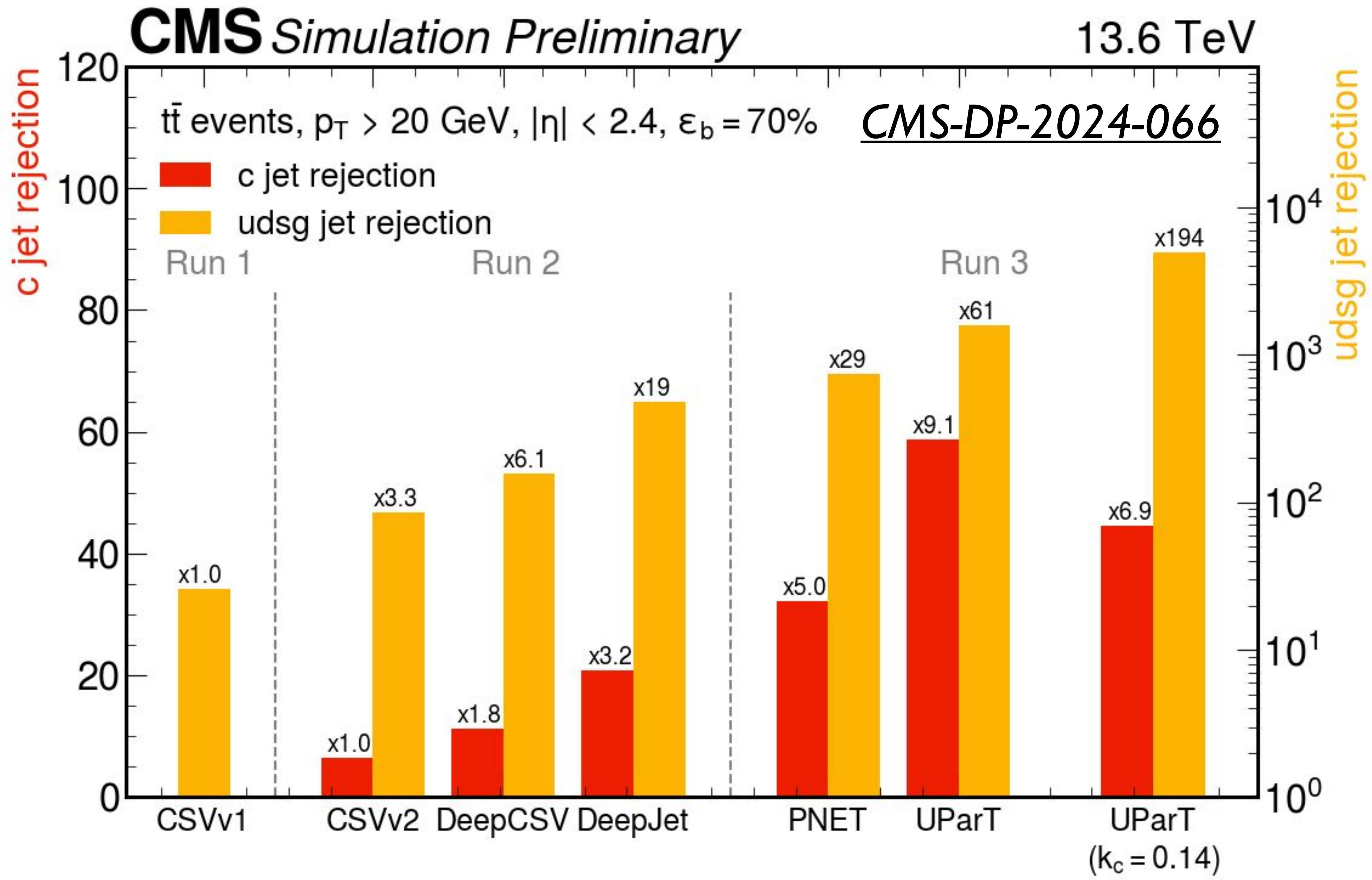
What type of particle initiates the jet?

The answer — Jet tagging!

THE EVOLUTION OF JET TAGGERS

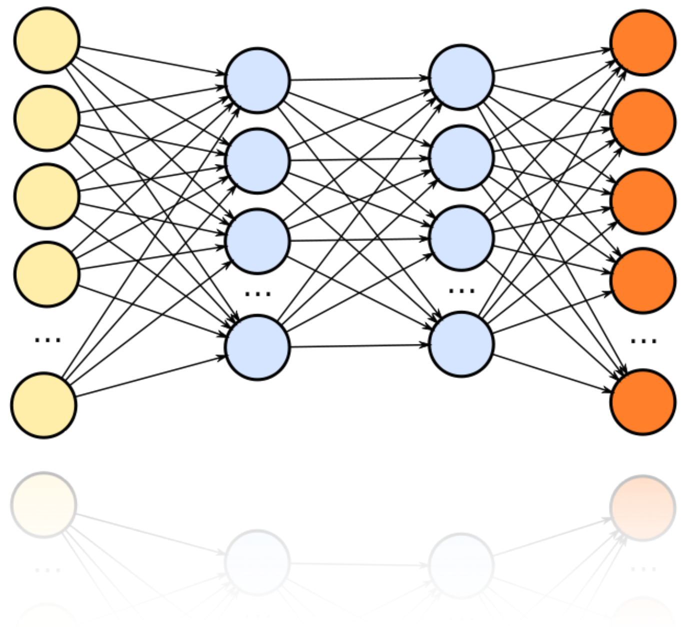
- **Tremendous progress in jet tagging in the past few years**
  - more than an order of magnitude improvement in light jet rejection





- **A driving force – advanced machine learning (ML) techniques**

The Evolution of Jet Taggers

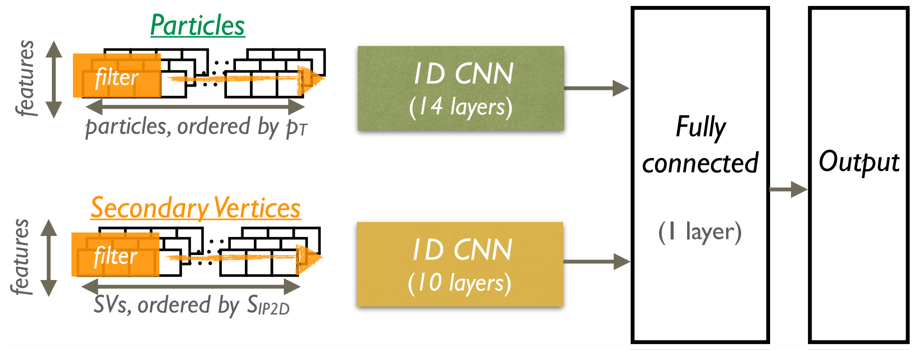


<u>"Shallow" ML</u>

- Inputs: O(10) hand-crafted features
  - tracks, SVs, (soft leptons)
- Model: BDTs or feedforward NNs

2015 2017 2019 2021 2023

The Evolution of Jet Taggers



"Shallow" ML

- Inputs: O(10) hand-crafted features
  - tracks, SVs, (soft leptons)
- Model: BDTs or feedforward NNs

"Deep" ML

- Inputs:
  - O(10-100) particles
  - O(1-10) SVs
  - O(~1000) low-level features in total
- Model: sequence-based deep NNs
  - 1D CNNs, RNNs, …

2015 2017 2019 2021 2023

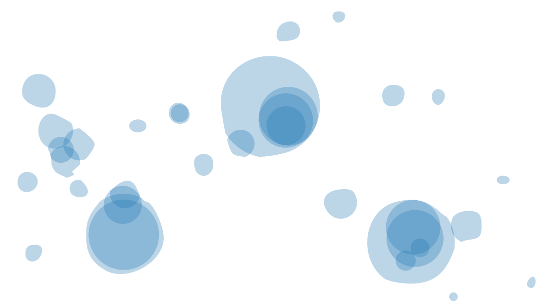
The Evolution of Jet Taggers

"Shallow" ML

- Inputs: O(10) hand-crafted features
  - tracks, SVs, (soft leptons)
- Model: BDTs or feedforward NNs

"Deep" ML

- Inputs:
  - O(10-100) particles
  - O(1-10) SVs
  - O(~1000) low-level features in total
- Model: sequence-based deep NNs
  - 1D CNNs, RNNs, …



Particle Cloud / GNNs

- Inputs:
  - O(10-100) particles
  - O(1-10) SVs
  - but viewed as an unordered "cloud"
- Model:
  - Graph Neural Networks (e.g., ParticleNet)

2015 2017 2019 2021 2023

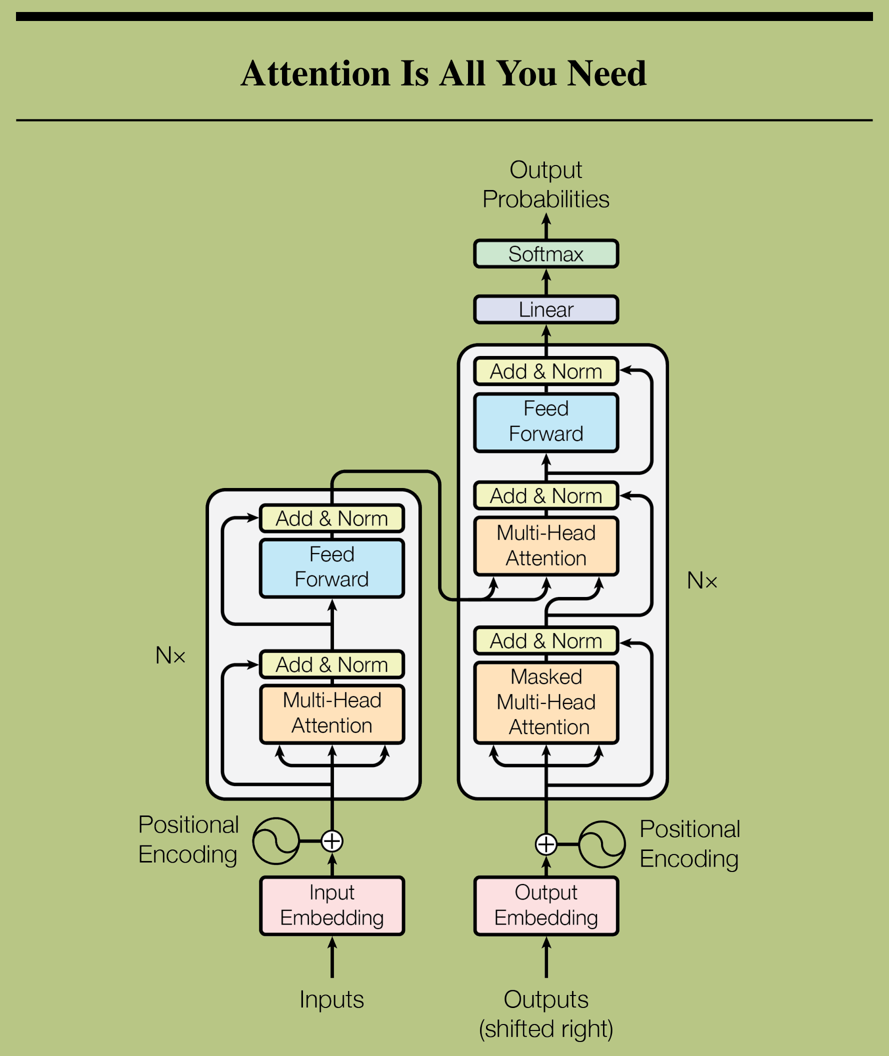
The Evolution of Jet Taggers

"Shallow" ML
- Inputs: O(10) hand-crafted features
  - tracks, SVs, (soft leptons)
- Model: BDTs or feedforward NNs

"Deep" ML
- Inputs:
  - O(10-100) particles
  - O(1-10) SVs
  - O(~1000) low-level features in total
- Model: sequence-based deep NNs
  - 1D CNNs, RNNs, …

Particle Cloud / GNNs
- Inputs:
  - O(10-100) particles
  - O(1-10) SVs
  - but viewed as an unordered "cloud"
- Model:
  - Graph Neural Networks (e.g., ParticleNet)

Transformers
- Inputs:
  - O(10-100) particles
  - O(1-10) SVs
- Model:

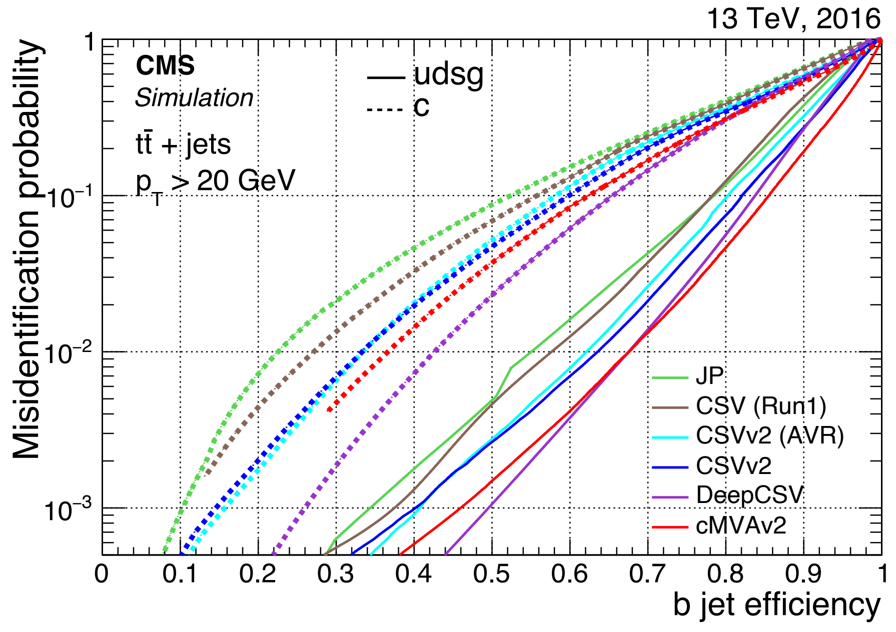


2015 2017 2019 2021 2023

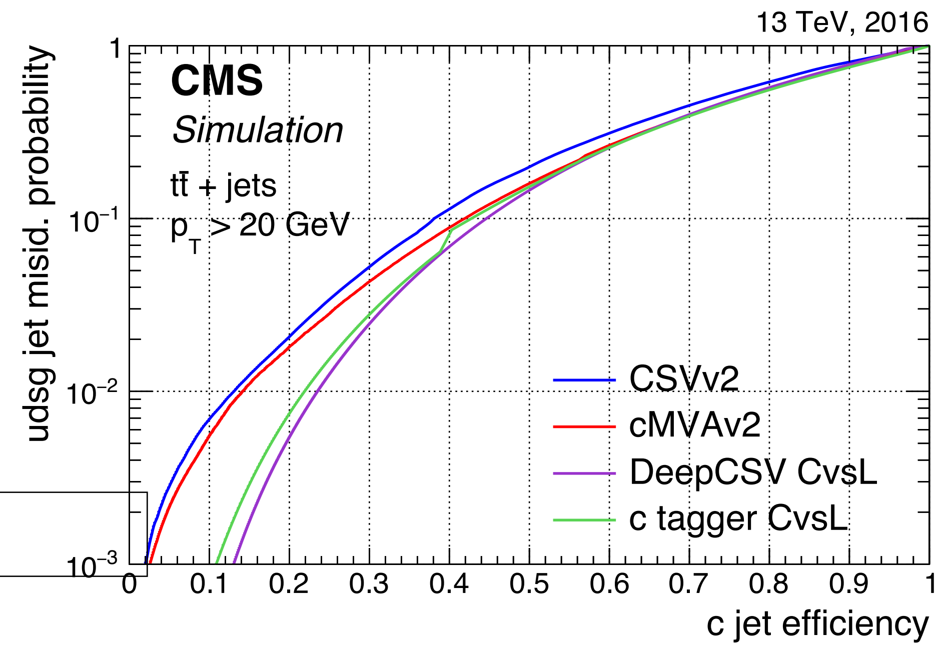
Towards a Universal Tagger

- For small-R jets: from individual taggers …

b-tagger

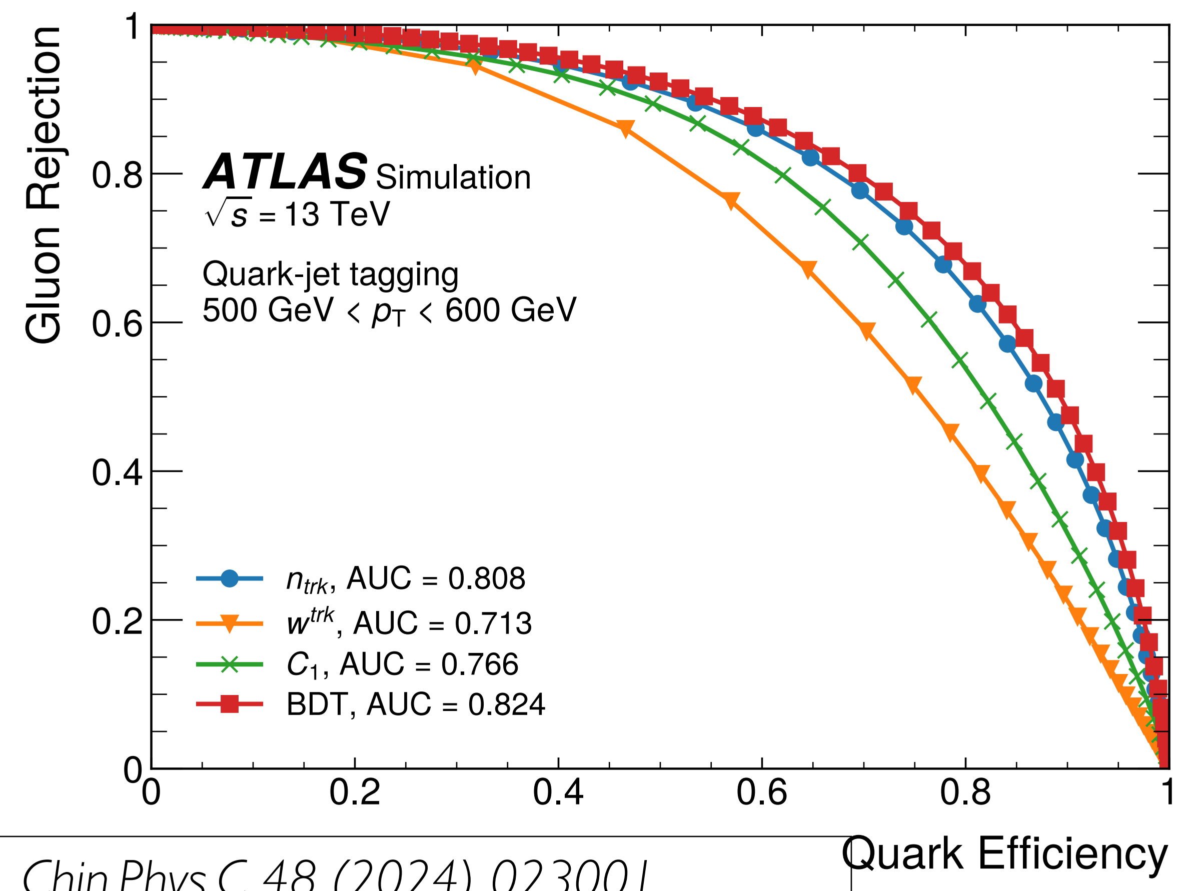


c-tagger



JINST 13 (2018) P05011

quark/gluon tagger



Chin.Phys.C 48 (2024) 023001

TOWARDS A UNIVERSAL TAGGER (I)

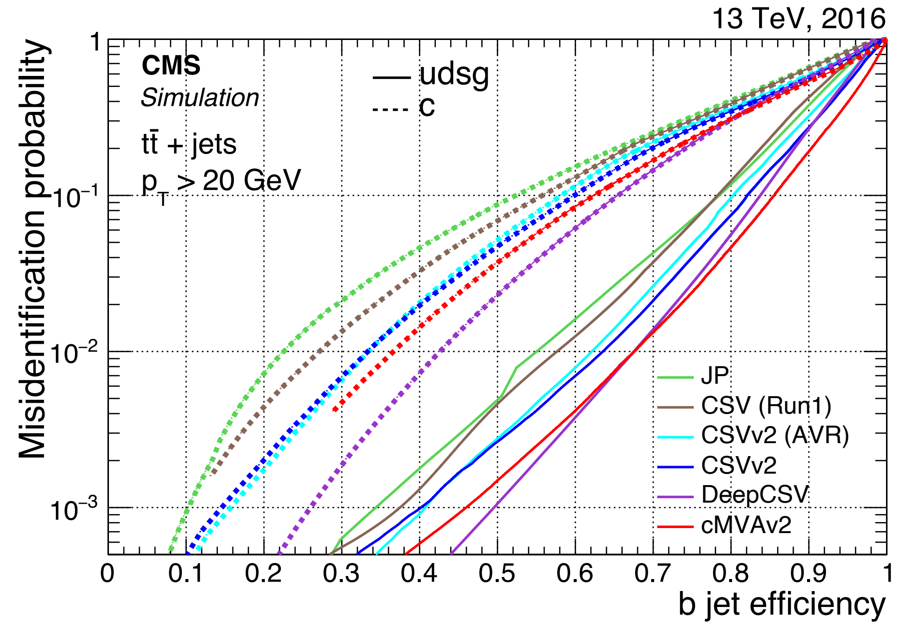
- For small-R jets: from individual taggers … to **a unified approach**

Unified Tagger

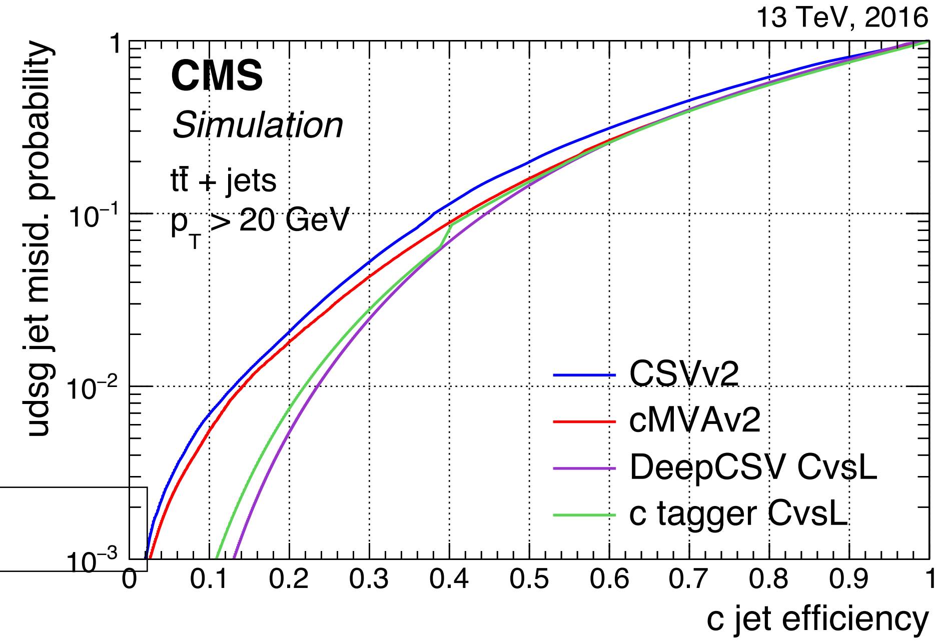
Attention Is All You Need

CMS-DP-2024-066

b-tagger

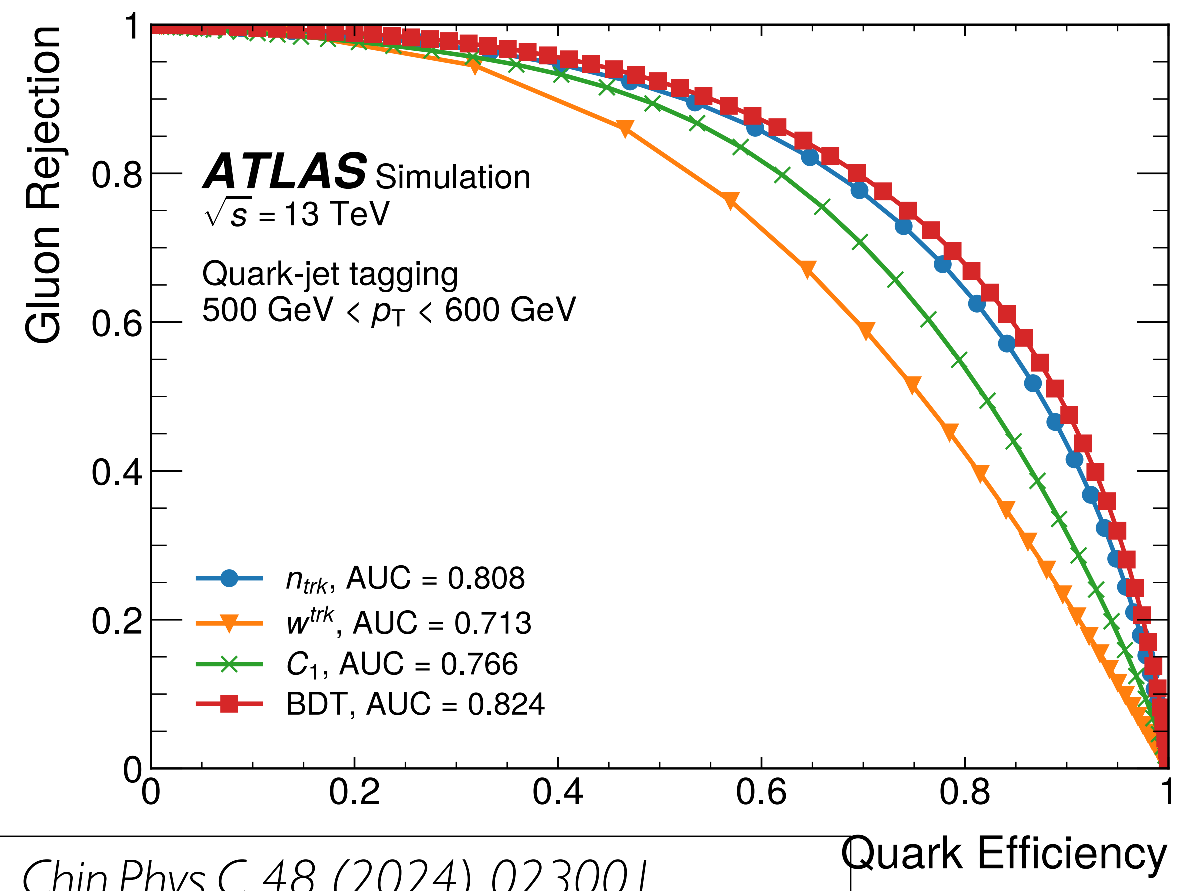


c-tagger



JINST 13 (2018) P05011

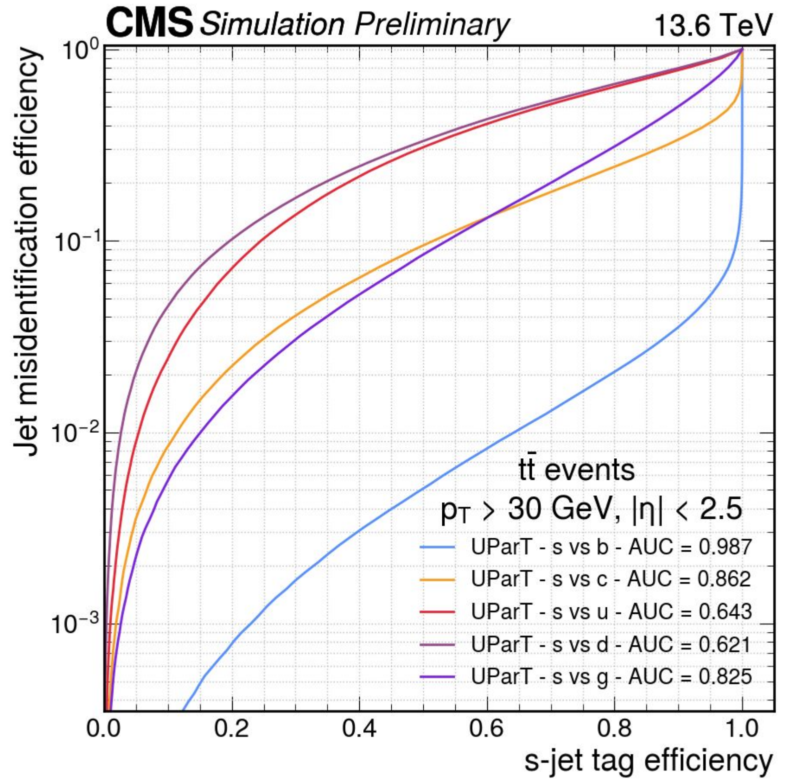
quark/gluon tagger



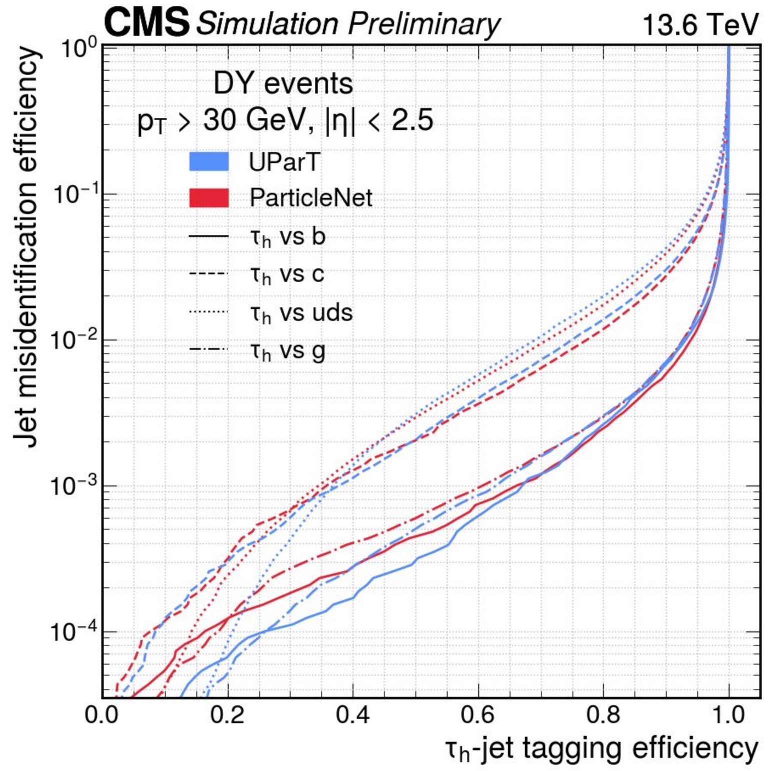
Chin.Phys.C 48 (2024) 023001

heavy flavor tagging, q/g discrimination, plus…

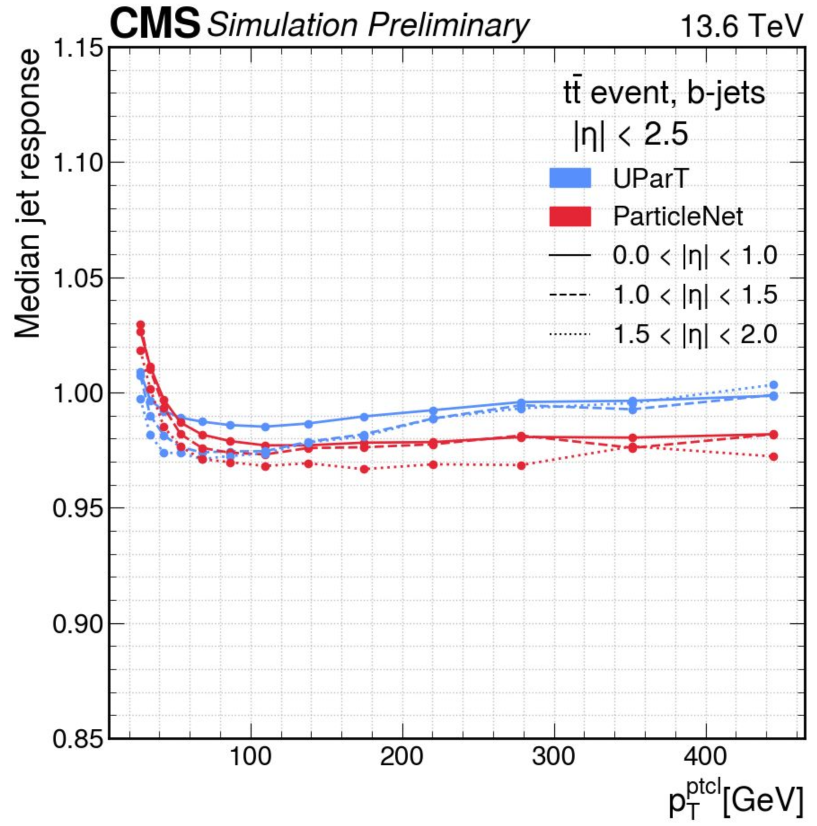
s-tagging



τ identification



jet energy regression



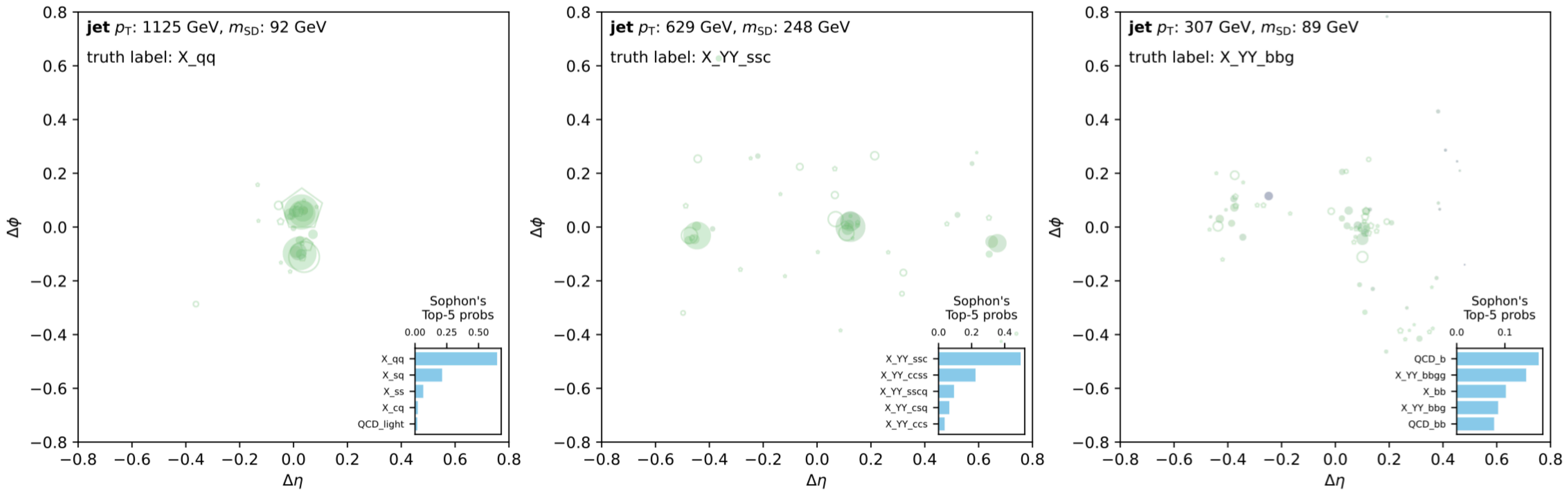
TOWARDS A UNIVERSAL TAGGER (II)

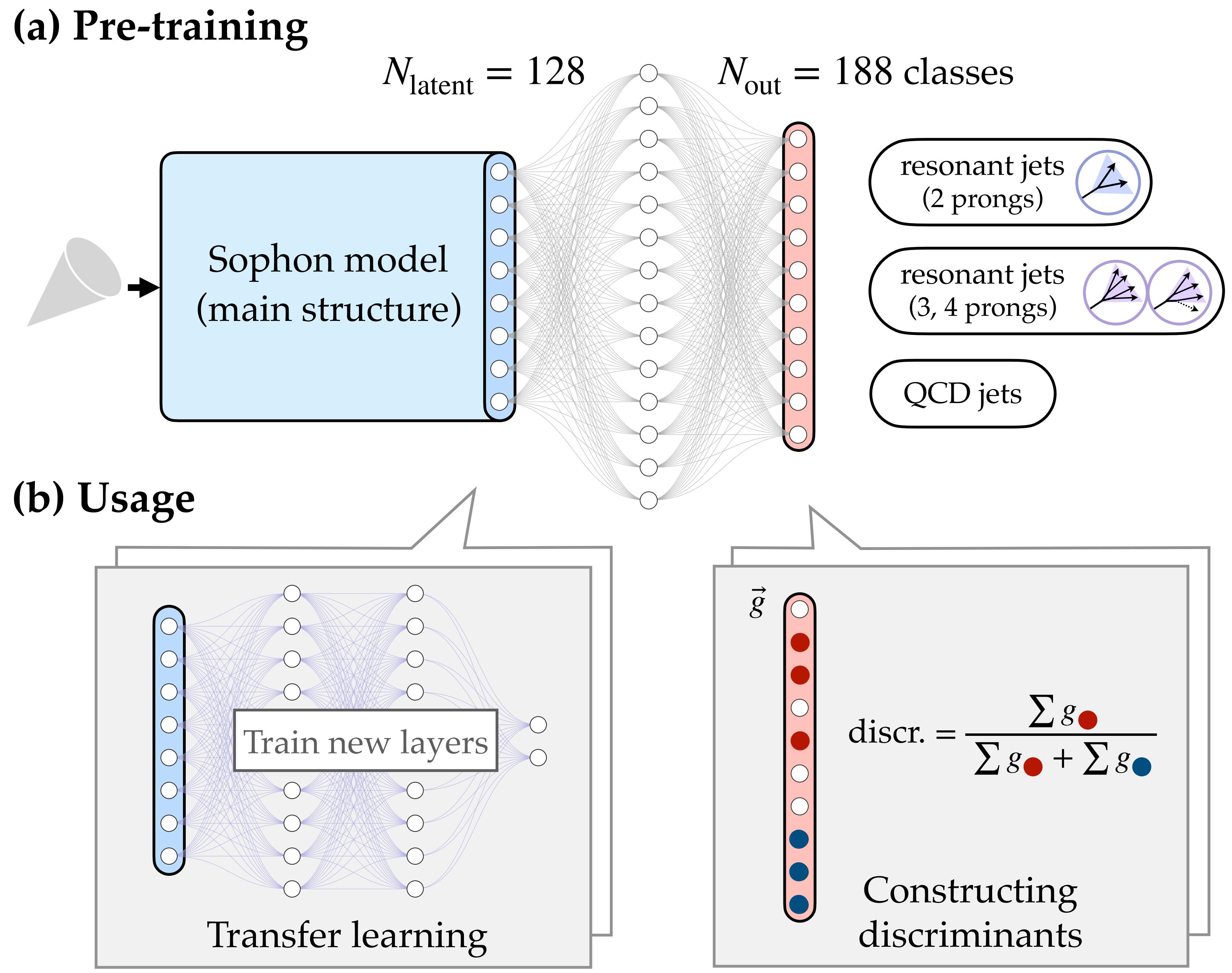
C. Li (李聡乔) et. al., arXiv:2405.12972

- For large-R jets: from specific SM resonance (W/Z/H/top) tagging to **generic signature-based tagging**
  - a proof-of-concept "*Sophon*": Particle Transformer trained on a wide range of boosted jet signatures (QCD + 2-, 3-, and 4-prong, 188 classes in total), decay modes, and resonance masses (up to 500 GeV)

TABLE I. Summary of the 188 jet labels in the JETCLASS-II dataset.

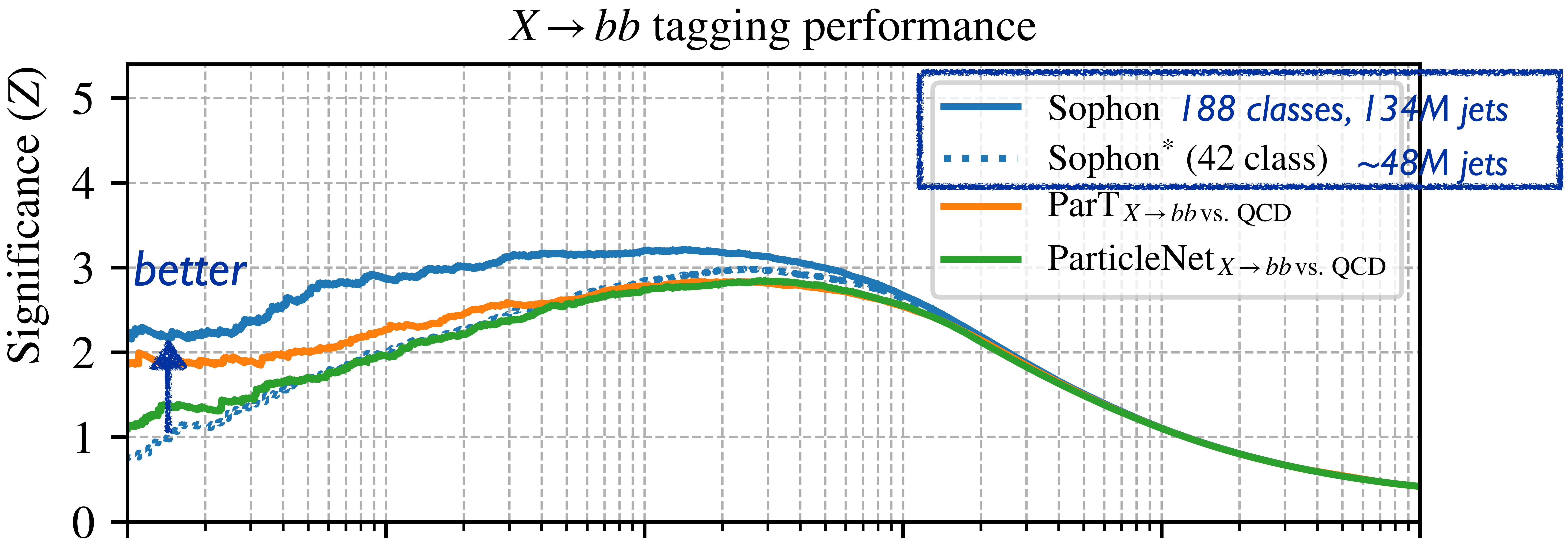
Major types	Index range	Label names
Resonant jets: $X \to 2$ prong	0–14	bb, cc, ss, qq, bc, cs, bq, cq, sq, gg, ee, $\mu\mu$, $\tau_h\tau_e$, $\tau_h\tau_\mu$, $\tau_h\tau_h$
Resonant jets: $X \to 3$ or 4 prong	15–160	$bbbb$, $bbcc$, $bbss$, $bbqq$, $bbgg$, $bbee$, $bb\mu\mu$, $bb\tau_h\tau_e$, $bb\tau_h\tau_\mu$, $bb\tau_h\tau_h$, bbb, bbc, bbs, bbq, bbg, bbe, $bb\mu$, $cccc$, $ccss$, $ccqq$, $ccgg$, $ccee$, $cc\mu\mu$, $cc\tau_h\tau_e$, $cc\tau_h\tau_\mu$, $cc\tau_h\tau_h$, ccb, ccc, ccs, ccq, ccg, cce, $cc\mu$, $ssss$, $ssqq$, $ssgg$, $ssee$, $ss\mu\mu$, $ss\tau_h\tau_e$, $ss\tau_h\tau_\mu$, $ss\tau_h\tau_h$, ssb, ssc, sss, ssq, ssg, sse, $ss\mu$, $qqqq$, $qqgg$, $qqee$, $qq\mu\mu$, $qq\tau_h\tau_e$, $qq\tau_h\tau_\mu$, $qq\tau_h\tau_h$, qqb, qqc, qqs, qqq, qqg, qqe, $qq\mu$, $gggg$, $ggee$, $gg\mu\mu$, $gg\tau_h\tau_e$, $gg\tau_h\tau_\mu$, $gg\tau_h\tau_h$, ggb, ggc, ggs, ggq, ggg, gge, $gg\mu$, bee, cee, see, qee, gee, $b\mu\mu$, $c\mu\mu$, $s\mu\mu$, $q\mu\mu$, $g\mu\mu$, $b\tau_h\tau_e$, $c\tau_h\tau_e$, $s\tau_h\tau_e$, $q\tau_h\tau_e$, $g\tau_h\tau_e$, $b\tau_h\tau_\mu$, $c\tau_h\tau_\mu$, $s\tau_h\tau_\mu$, $q\tau_h\tau_\mu$, $g\tau_h\tau_\mu$, $b\tau_h\tau_h$, $c\tau_h\tau_h$, $s\tau_h\tau_h$, $q\tau_h\tau_h$, $g\tau_h\tau_h$, $qqqb$, $qqqc$, $qqqs$, $bbcq$, $ccbs$, $ccbq$, $ccsq$, $sscq$, $qqbc$, $qqbs$, $qqcs$, $bcsq$, bcs, bcq, bsq, csq, $bce\nu$, $cse\nu$, $bqe\nu$, $cqe\nu$, $sqe\nu$, $qqe\nu$, $bc\mu\nu$, $cs\mu\nu$, $bq\mu\nu$, $cq\mu\nu$, $sq\mu\nu$, $qq\mu\nu$, $bc\tau_e\nu$, $cs\tau_e\nu$, $bq\tau_e\nu$, $cq\tau_e\nu$, $sq\tau_e\nu$, $qq\tau_e\nu$, $bc\tau_\mu\nu$, $cs\tau_\mu\nu$, $bq\tau_\mu\nu$, $cq\tau_\mu\nu$, $sq\tau_\mu\nu$, $qq\tau_\mu\nu$, $bc\tau_h\nu$, $cs\tau_h\nu$, $bq\tau_h\nu$, $cq\tau_h\nu$, $sq\tau_h\nu$, $qq\tau_h\nu$
QCD jets	161–187	$bbccss$, $bbccs$, $bbcc$, $bbcss$, $bbcs$, bbc, $bbss$, bbs, bb, $bccss$, $bccs$, bcc, $bcss$, bcs, bc, bss, bs, b, $ccss$, ccs, cc, css, cs, c, ss, s, others





Towards a Universal Tagger (II)

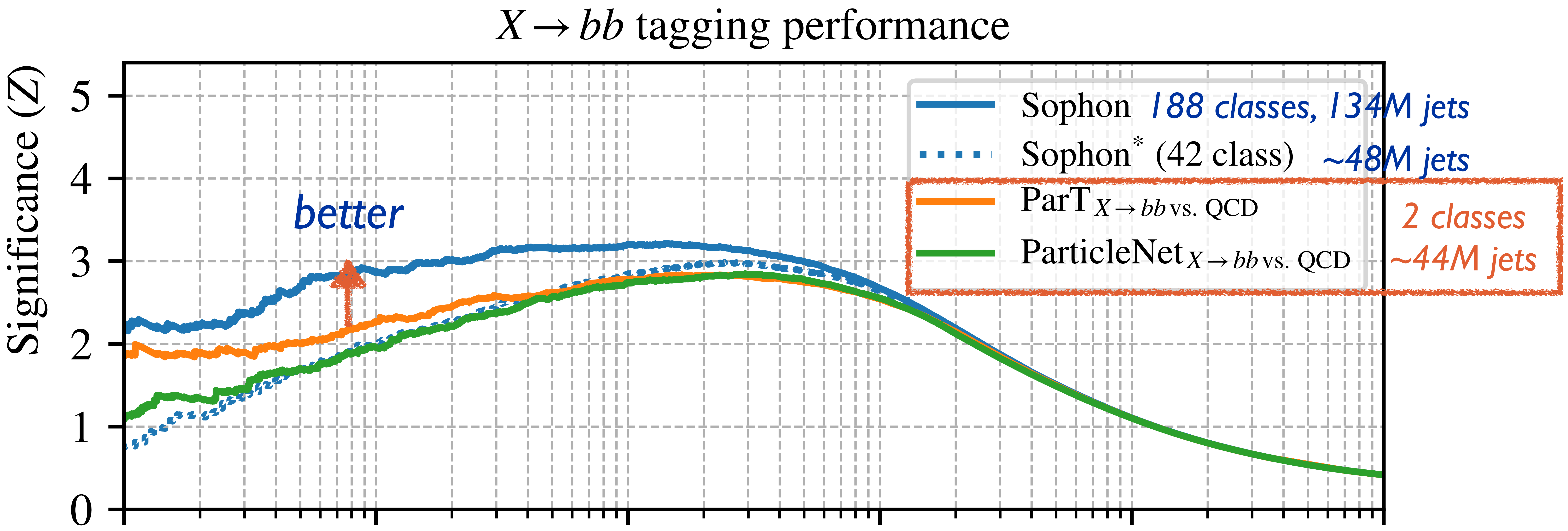
- For large-R jets: from specific SM resonance (W/Z/H/top) tagging to **generic signature-based tagging**
- A few observations:
  - larger dataset helps – even if not directly adding the target classes



C. Li et. al., arXiv:2405.12972

Towards a Universal Tagger (II)

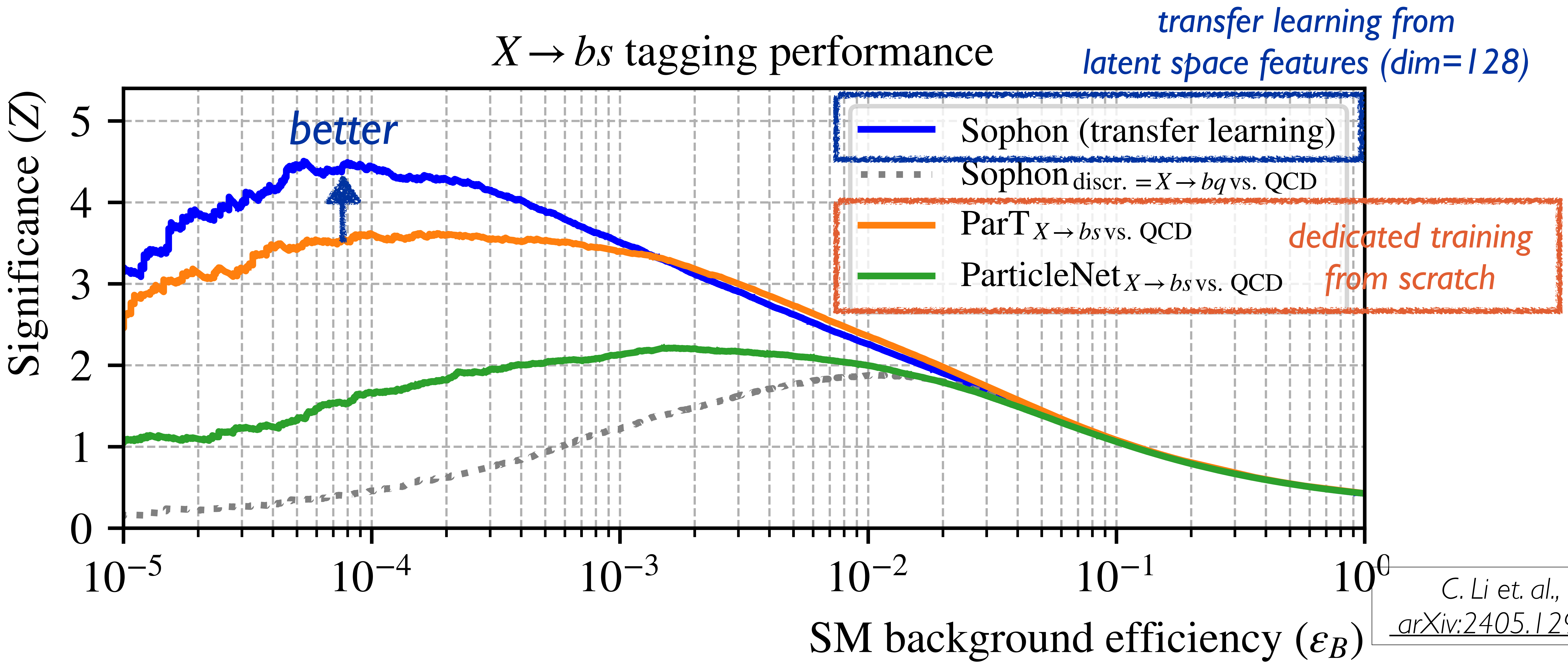
- For large-R jets: from specific SM resonance (W/Z/H/top) tagging to **generic signature-based tagging**
- A few observations:
  - larger dataset helps – even if not directly adding the target classes



C. Li et. al., arXiv:2405.12972

TOWARDS A UNIVERSAL TAGGER (II)

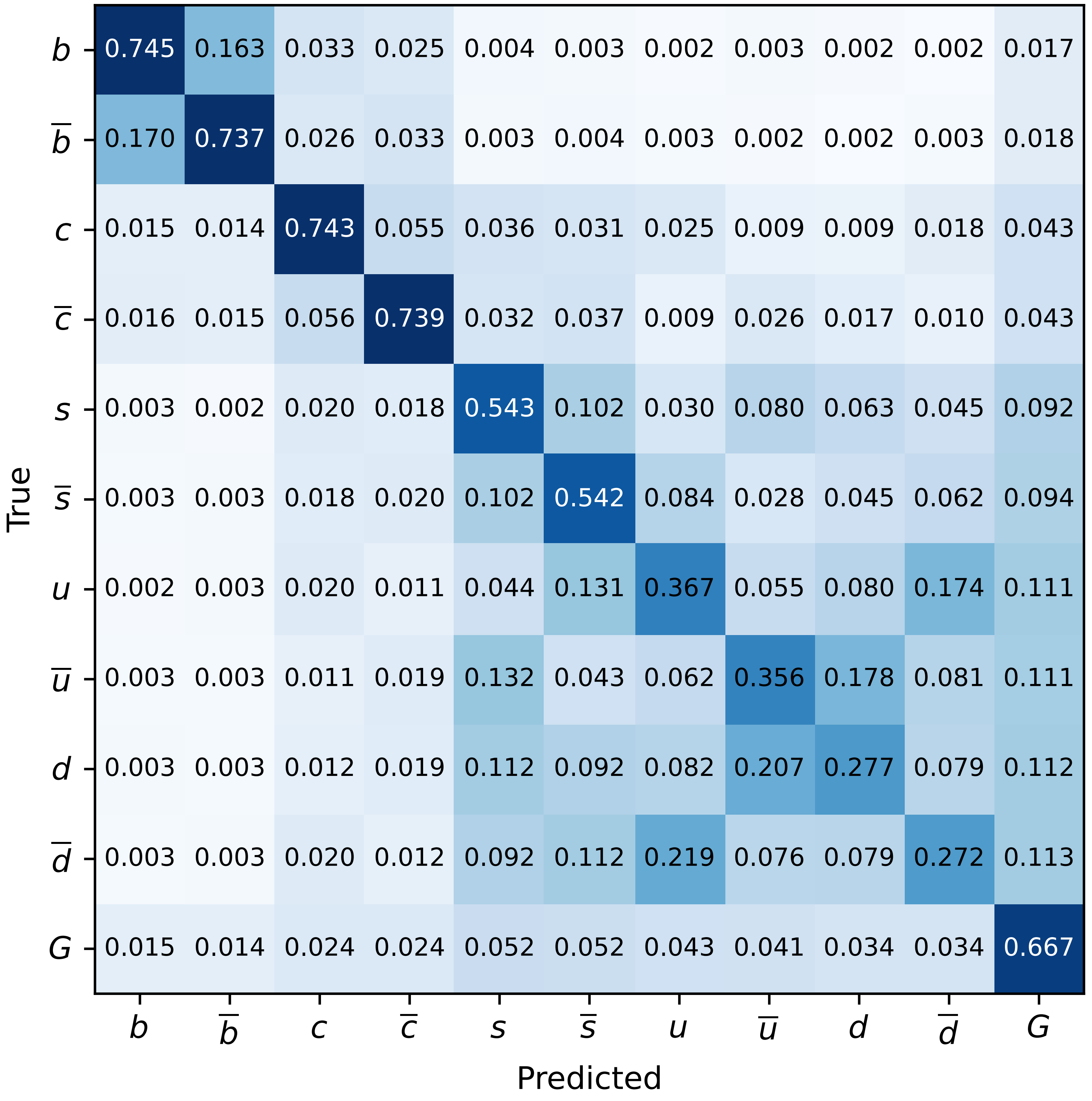
- For large-R jets: from specific SM resonance (W/Z/H/top) tagging to **generic signature-based tagging**
- A few observations:
  - larger dataset helps – even if not directly adding the target classes
  - large model -> **stronger transfer learning capability**



C. Li et. al., arXiv:2405.12972

TOWARDS A UNIVERSAL TAGGER (III)

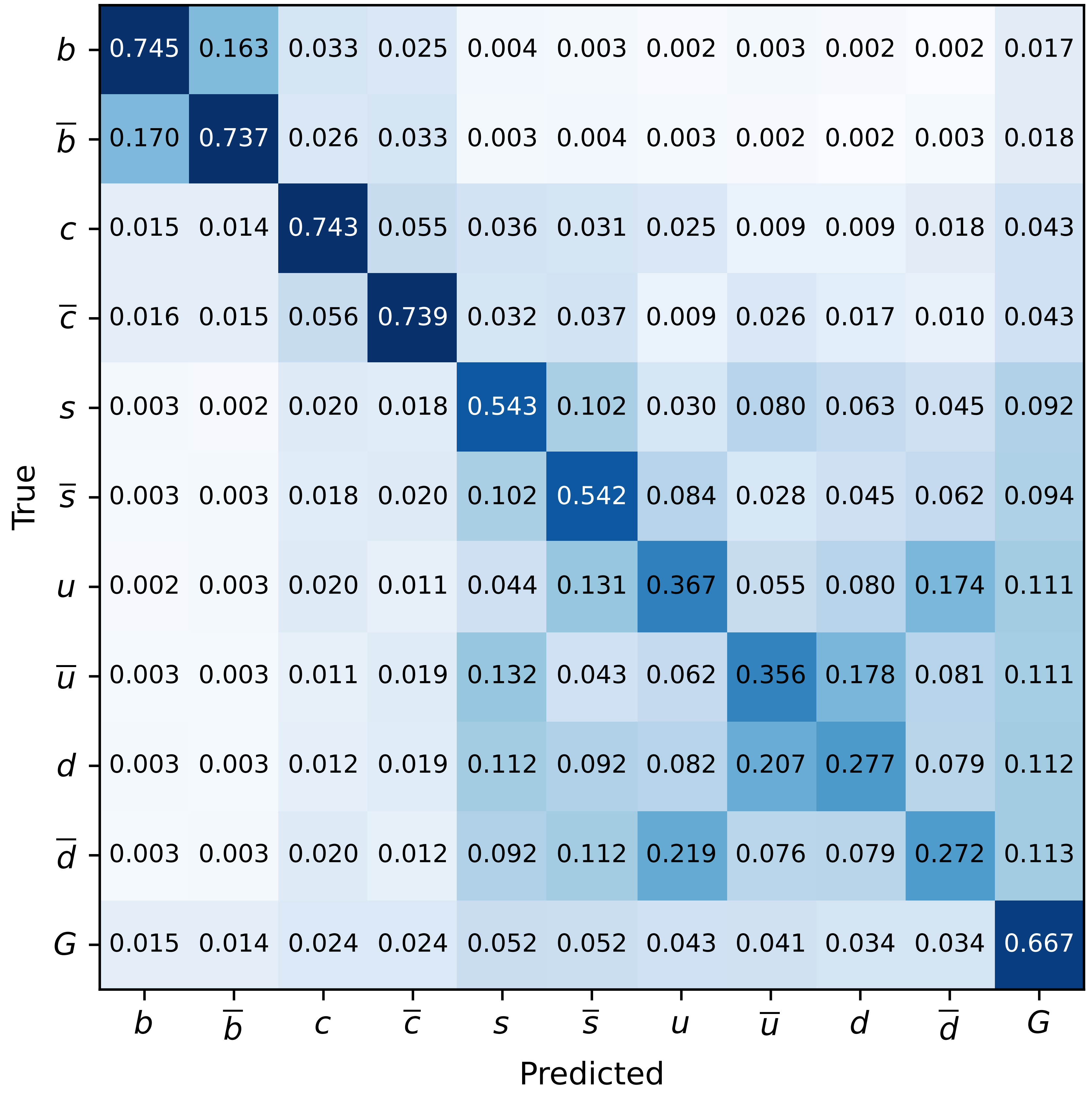
- At future e+e- collider…
  - resolving all 11 species of colored particles:
    - $b, \bar{b}, c, \bar{c}, s, \bar{s}, u, \bar{u}, d, \bar{d}, g$

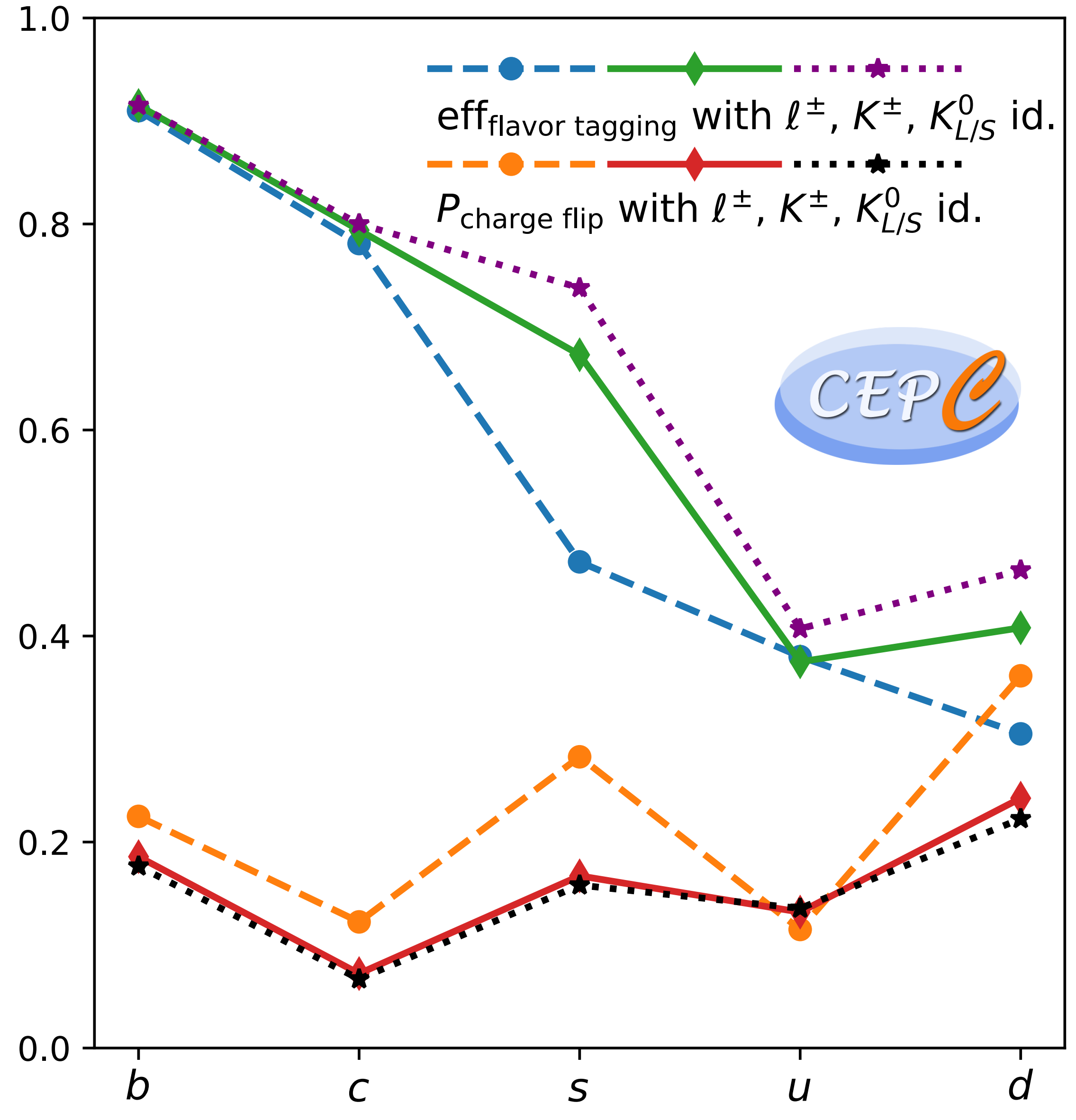


H. Liang, Y. Zhu, Y. Wang, Y. Che, M. Ruan, C. Zhou, and HQ
PRL 132 (2024), 221802
Eur.Phys.J.C 84 (2024), 152

Towards a Universal Tagger (III)

- At future e+e- collider…
  - resolving all 11 species of colored particles:
    - $b, \bar{b}, c, \bar{c}, s, \bar{s}, u, \bar{u}, d, \bar{d}, g$

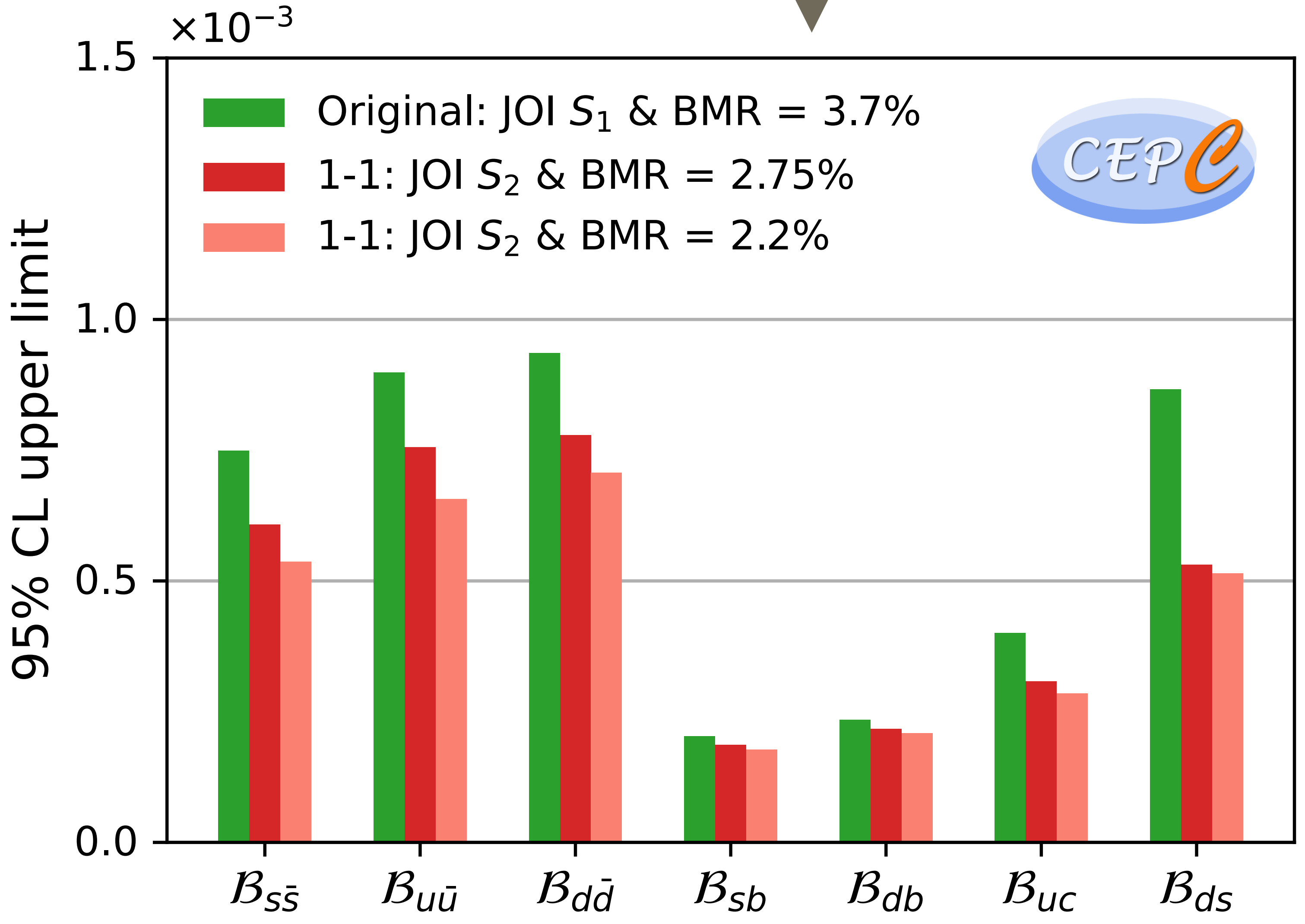




ML taggers provide an end-to-end assessment of key detector designs choices

H. Liang, Y. Zhu, Y. Wang, Y. Che, M. Ruan, C. Zhou and HQ
[PRL 132 (2024), 221802](#)
[Eur.Phys.J.C 84 (2024), 152](#)

Further enhancement from AI-assisted 1-1 correspondence reconstruction

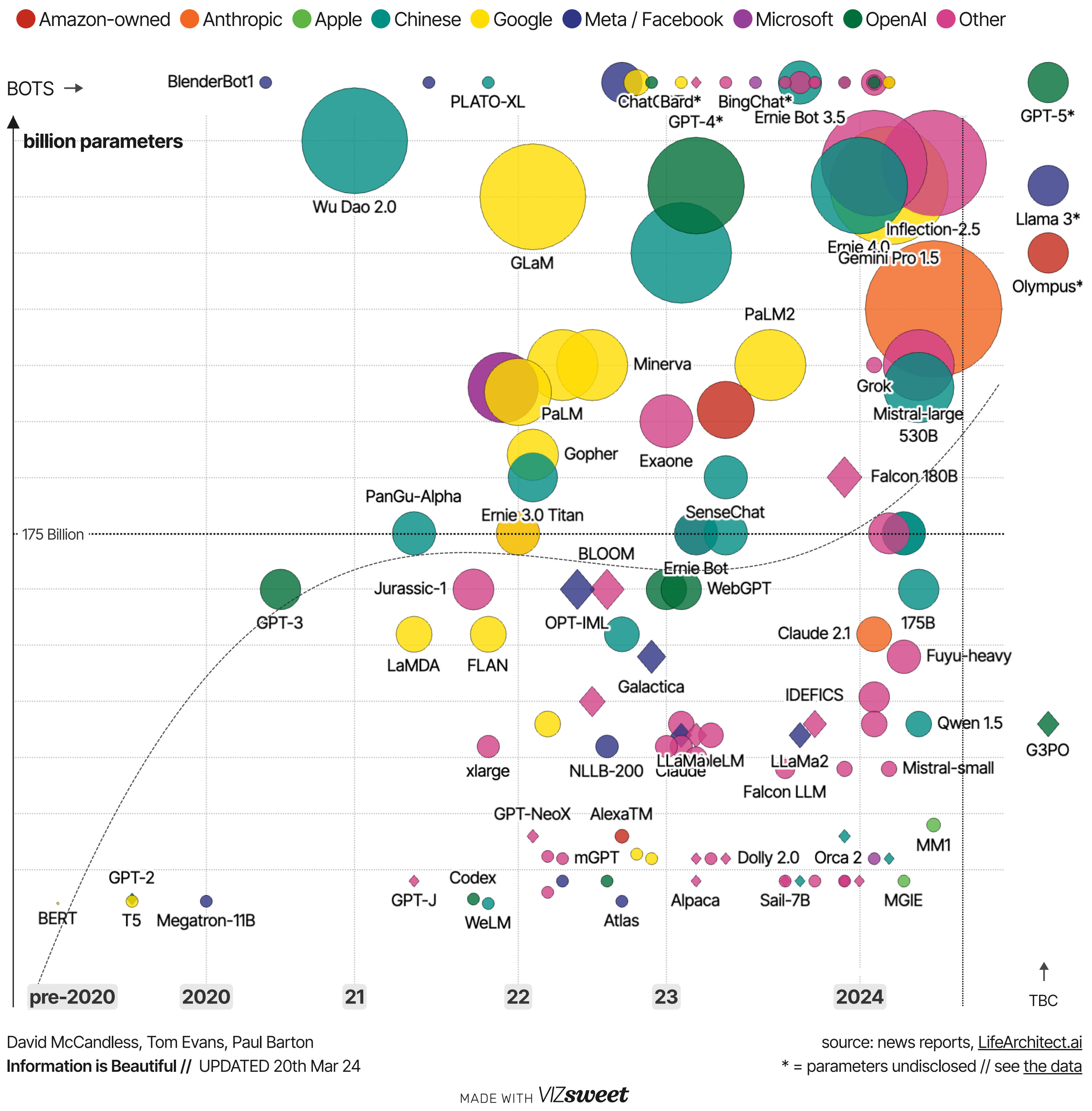


Y. Wang, H. Liang, Y. Zhu, Y. Che, X. Xia, HQ, C. Zhou, X. Zhuang and M. Ruan, [arXiv:2411.06939](#)

What's Next?

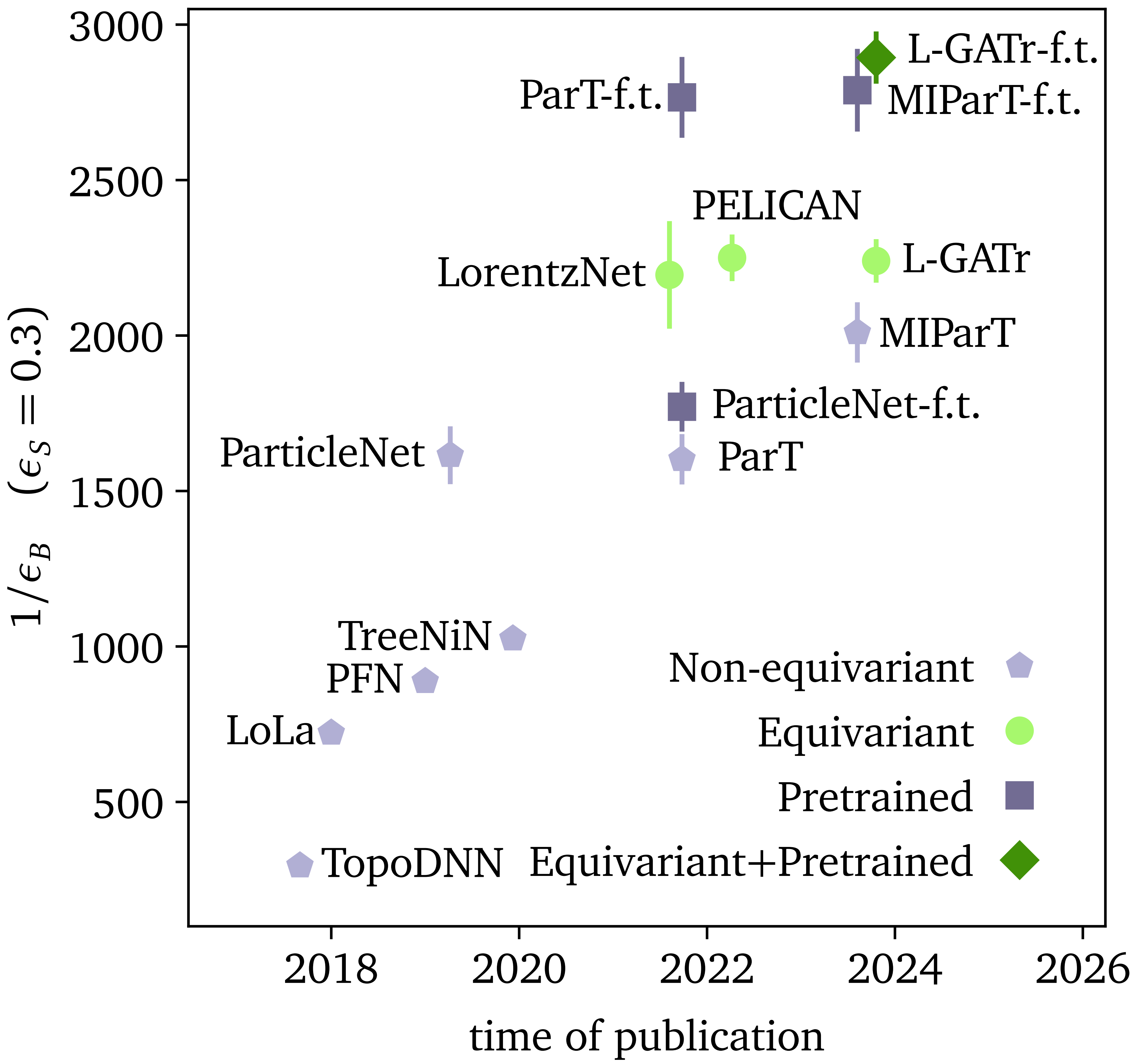
Scaling Up?

Natural language models



Source: informationisbeautiful.net

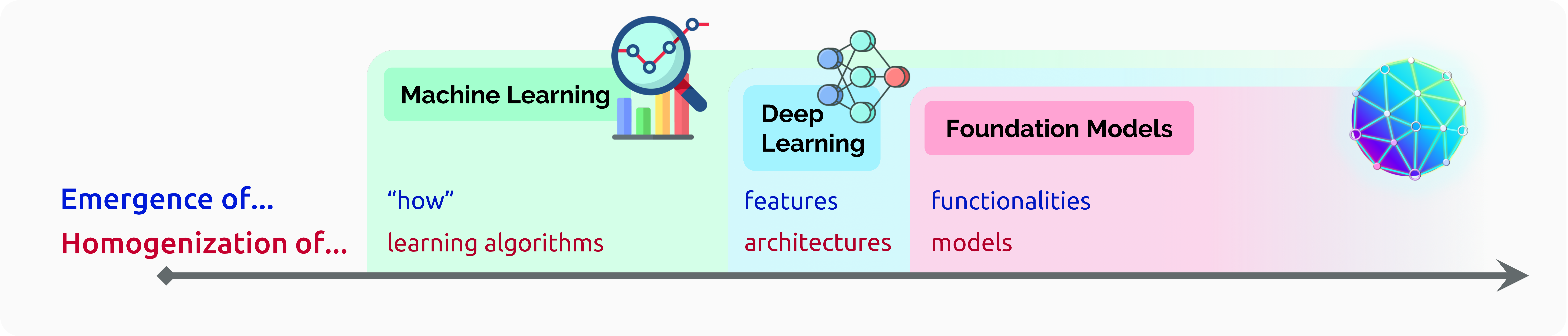
HEP models (jet tagging)



J. Brehmer, V. Bresó, P. Haan, T. Plehn, HQ, J. Spinner and J. Thaler, arXiv: 2411.00446

Foundation Model

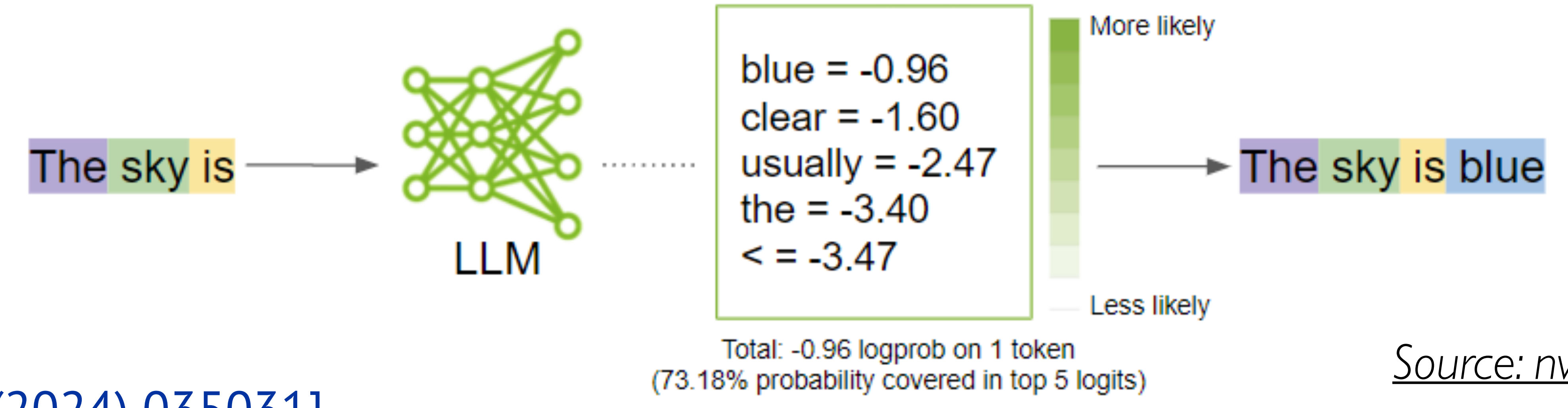
"A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks."



[arXiv: 2108.07258]

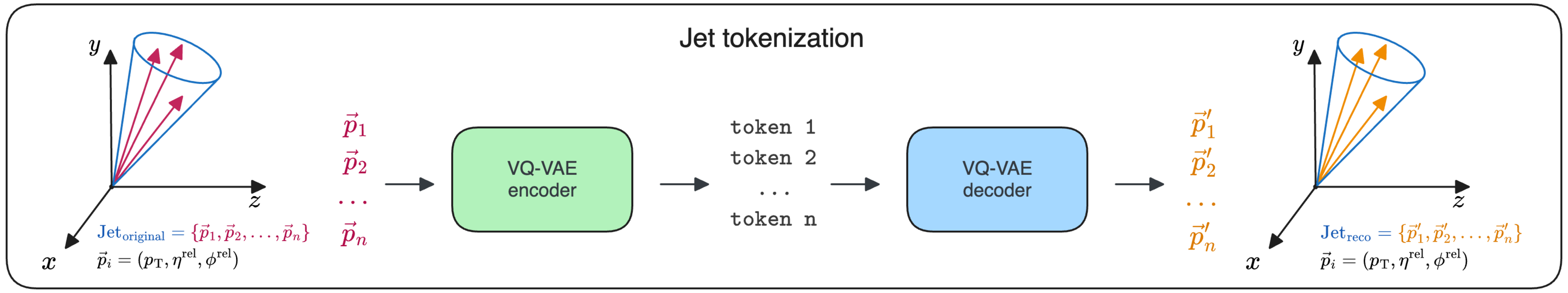
Self-supervision: The NLP Approach

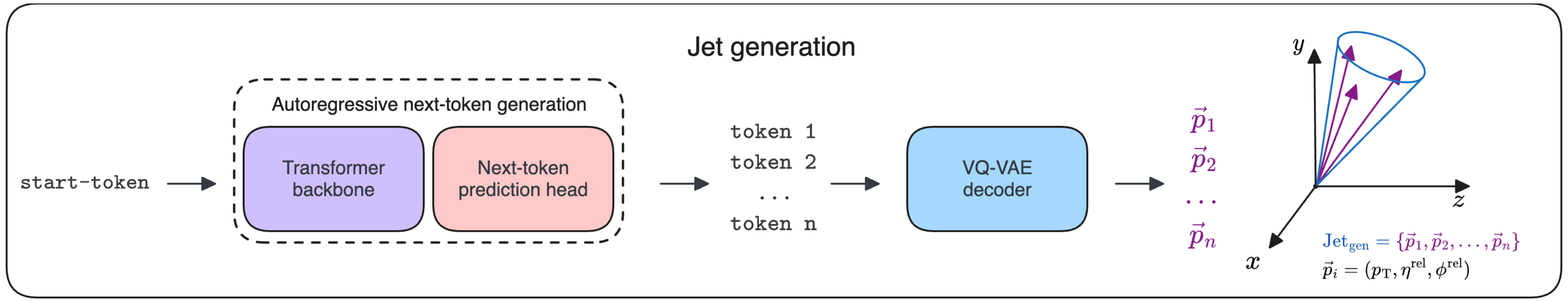
- The NLP way: **(autoregressive) language modeling**
  - i.e., next token prediction



Source: nvidia.com

- An attempt for jets: OmniJet-α [J. Birk, A. Hallin and G. Kasieczka, MLST 5 (2024) 035031]

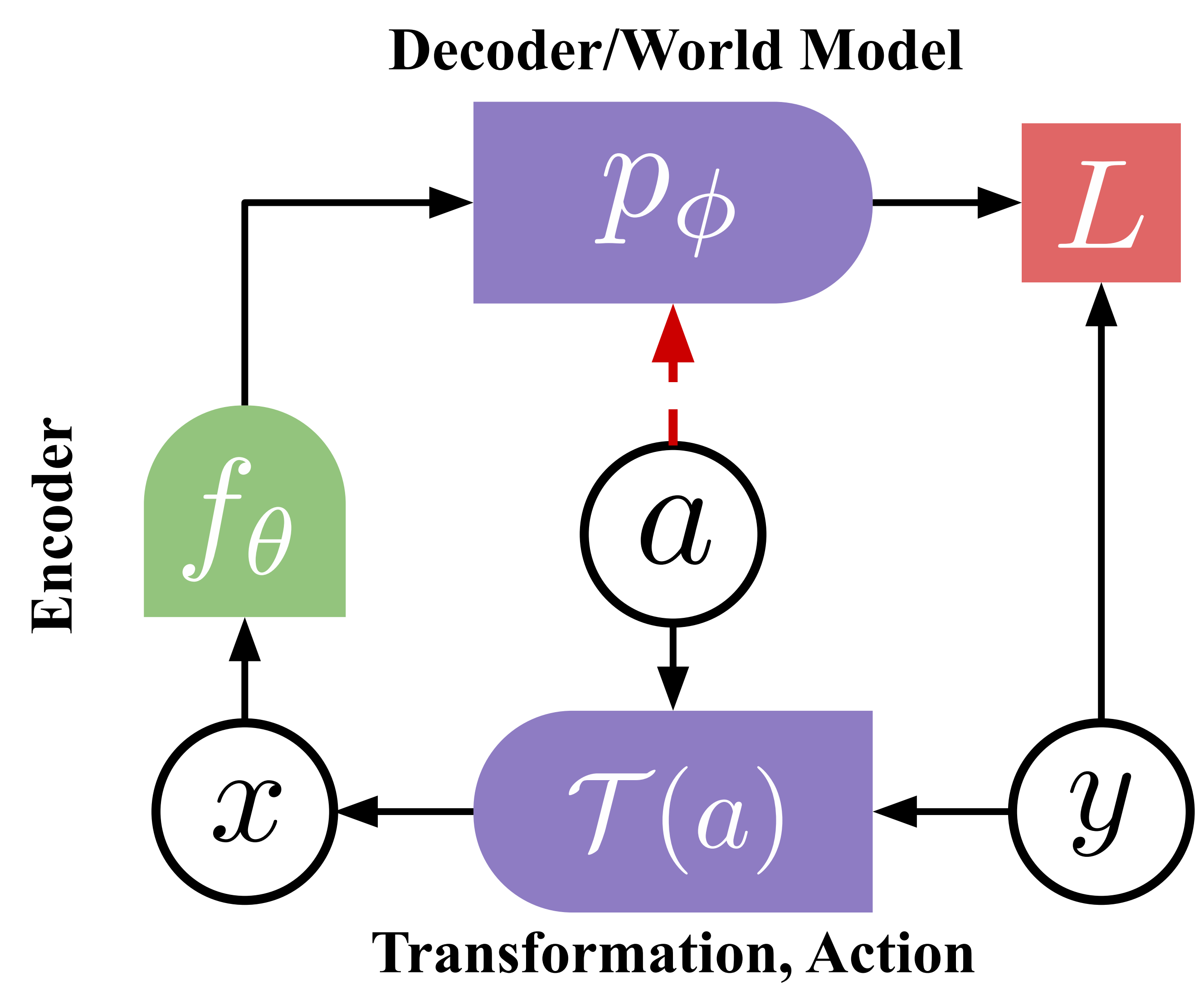




- Not the most natural approach though:
  - requires (discrete) tokenization of high-dimensional numerical inputs
  - needs to impose an ordering on jet constituent particles, which are intrinsically permutation invariant

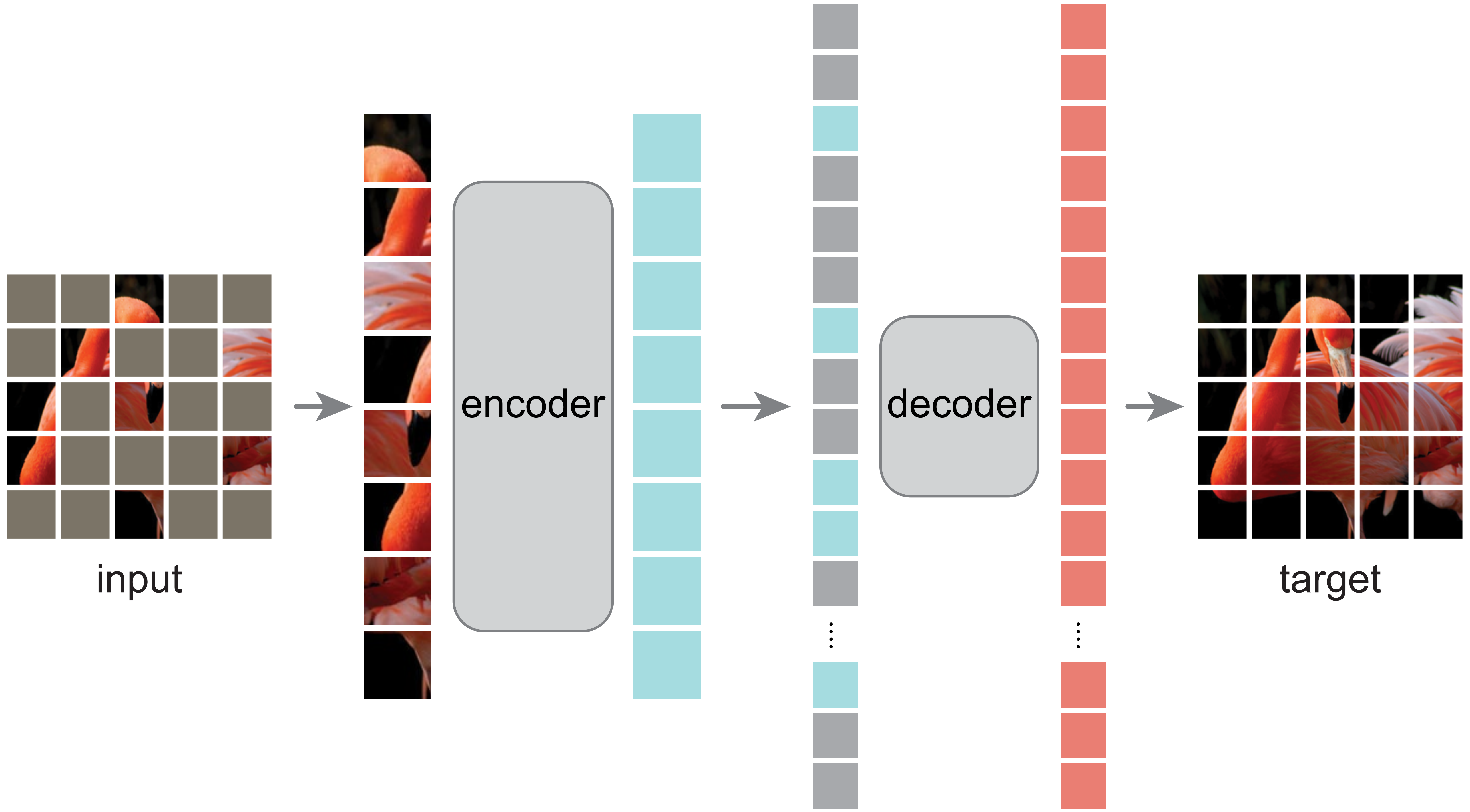
SELF-SUPERVISION: THE CV APPROACH

Generative Architecture



Learns to invert a transformation in the ***input*** *space.*
e.g., Masked Autoencoder (MAE)

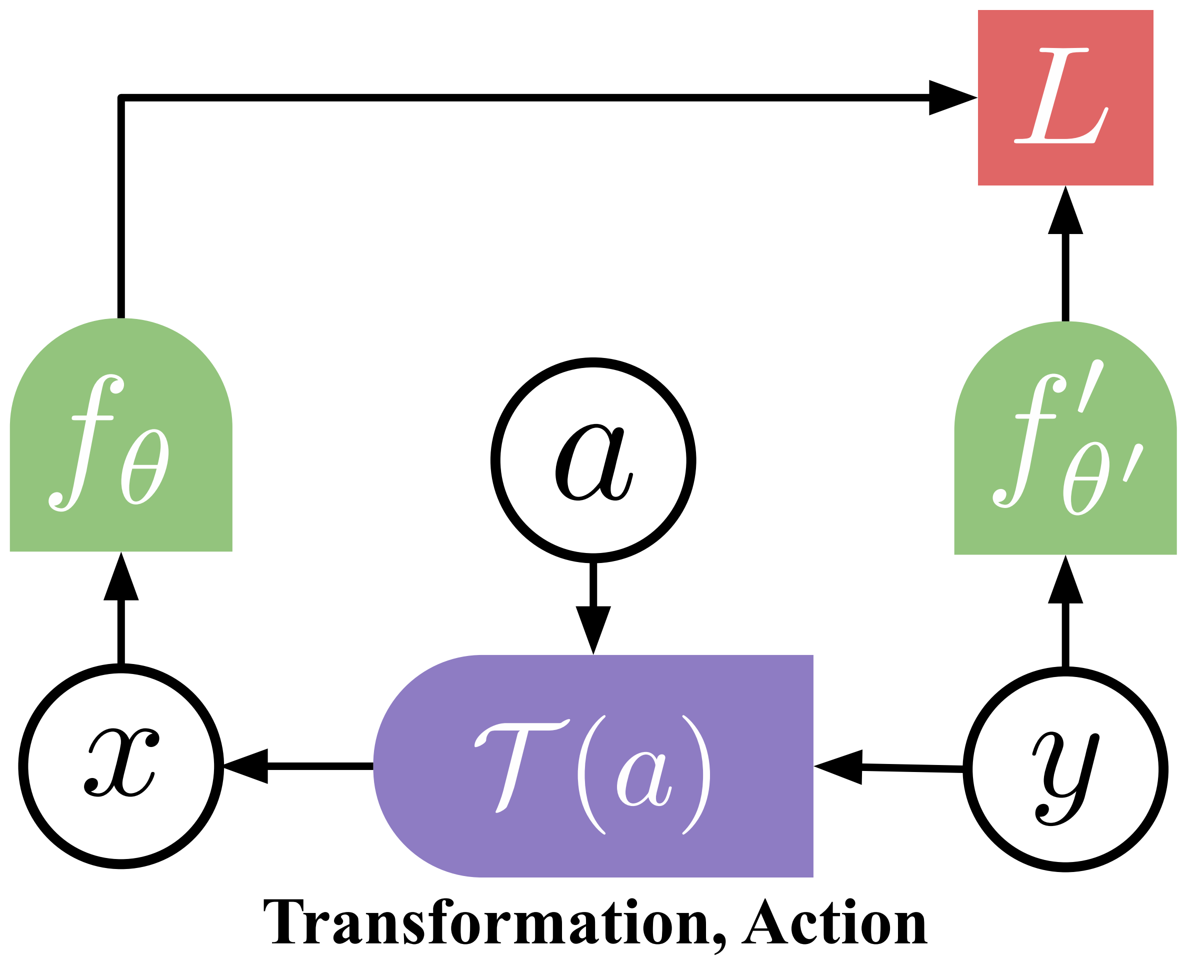
Masked Autoencoder [arXiv: 2111.06377]



… masks random patches of the input image and reconstructs the missing pixels

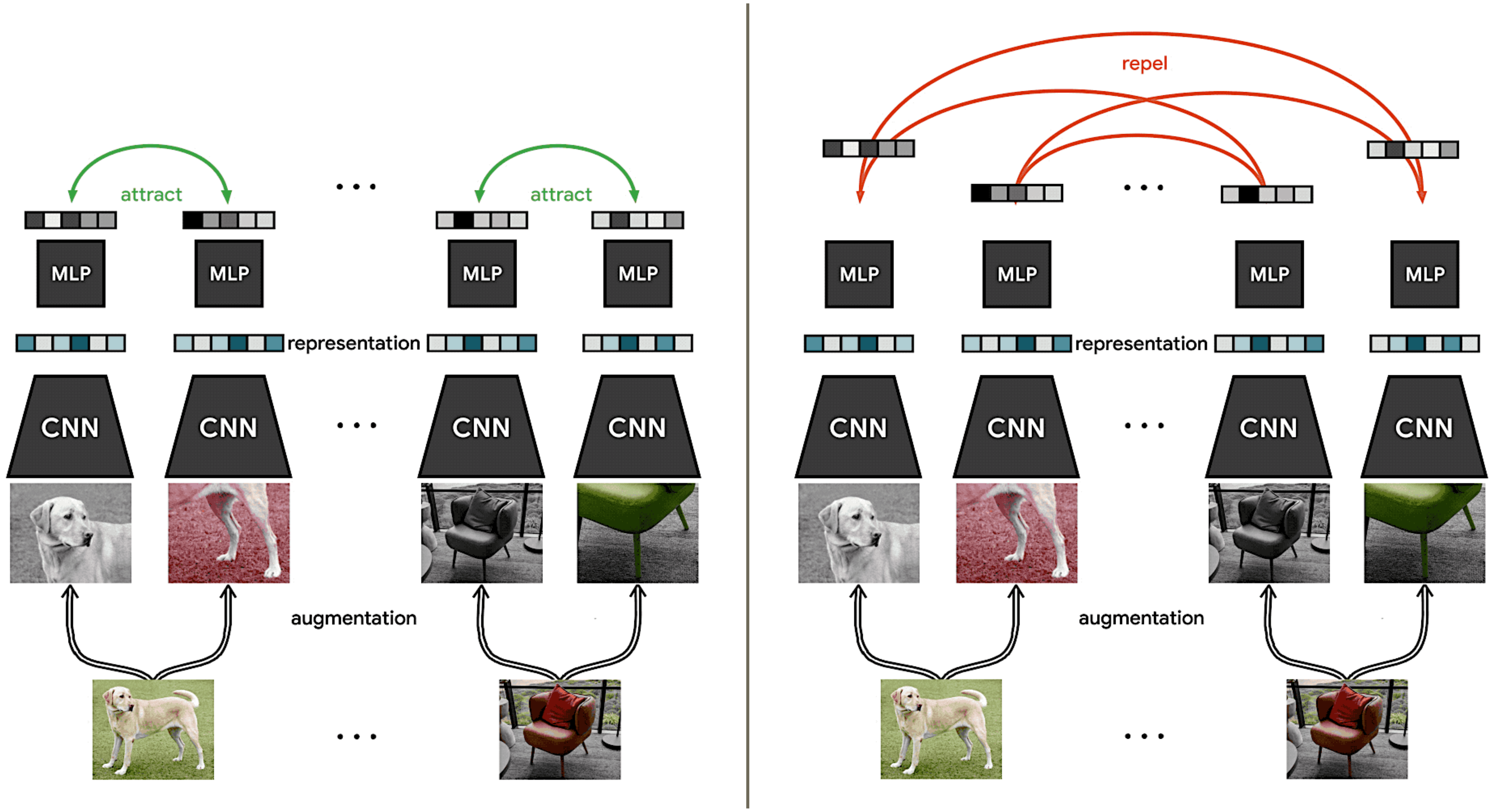
SELF-SUPERVISION: THE CV APPROACH

Joint-Embedding Architecture



*Learns an invariant representation in the **latent** space.*
e.g., SimCLR / DINO

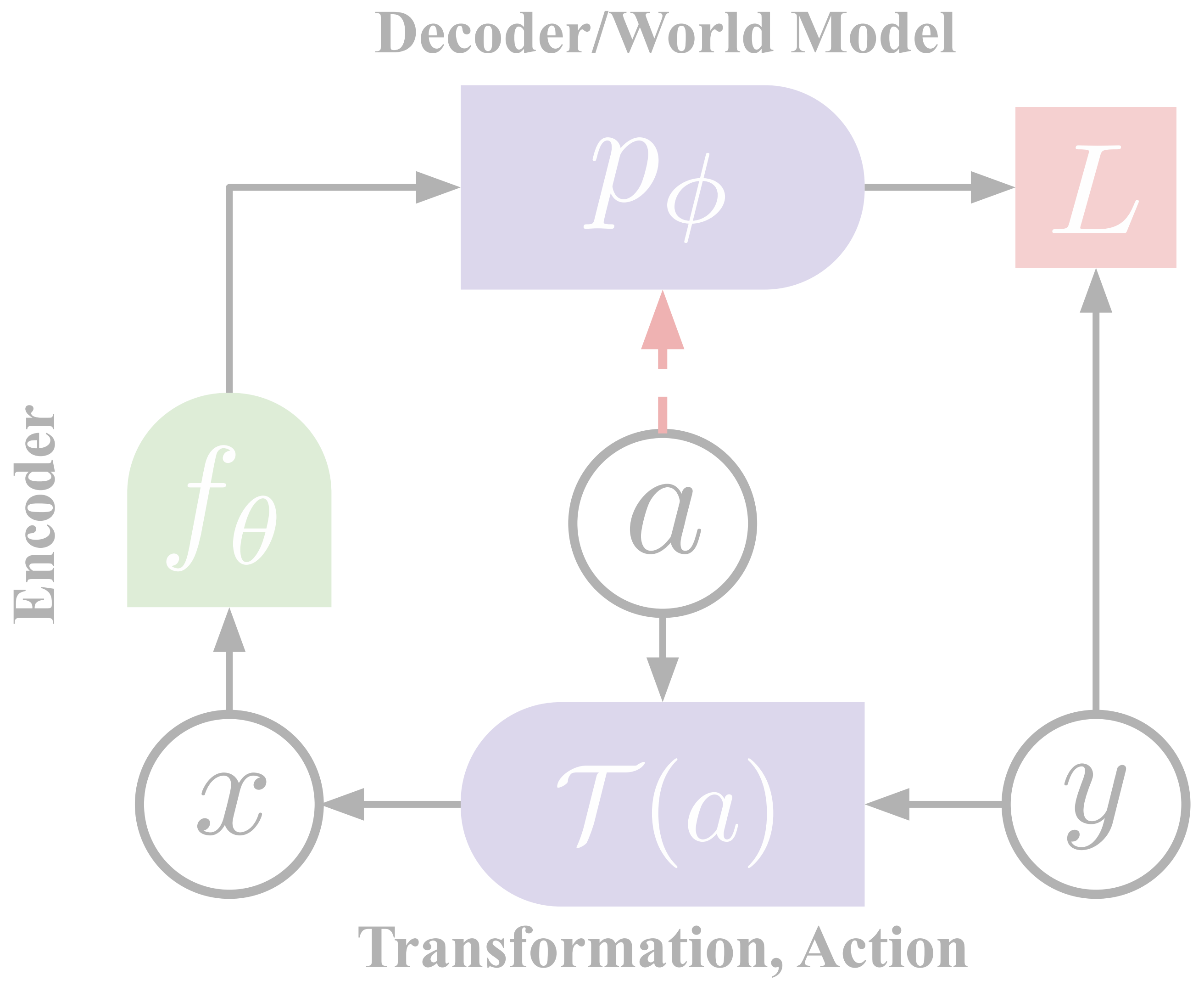
SimCLR [arXiv: 2002.05709]



… *maximizes similarity between positive pairs (same images after transformations) and minimizes that between negative pairs (different images)*

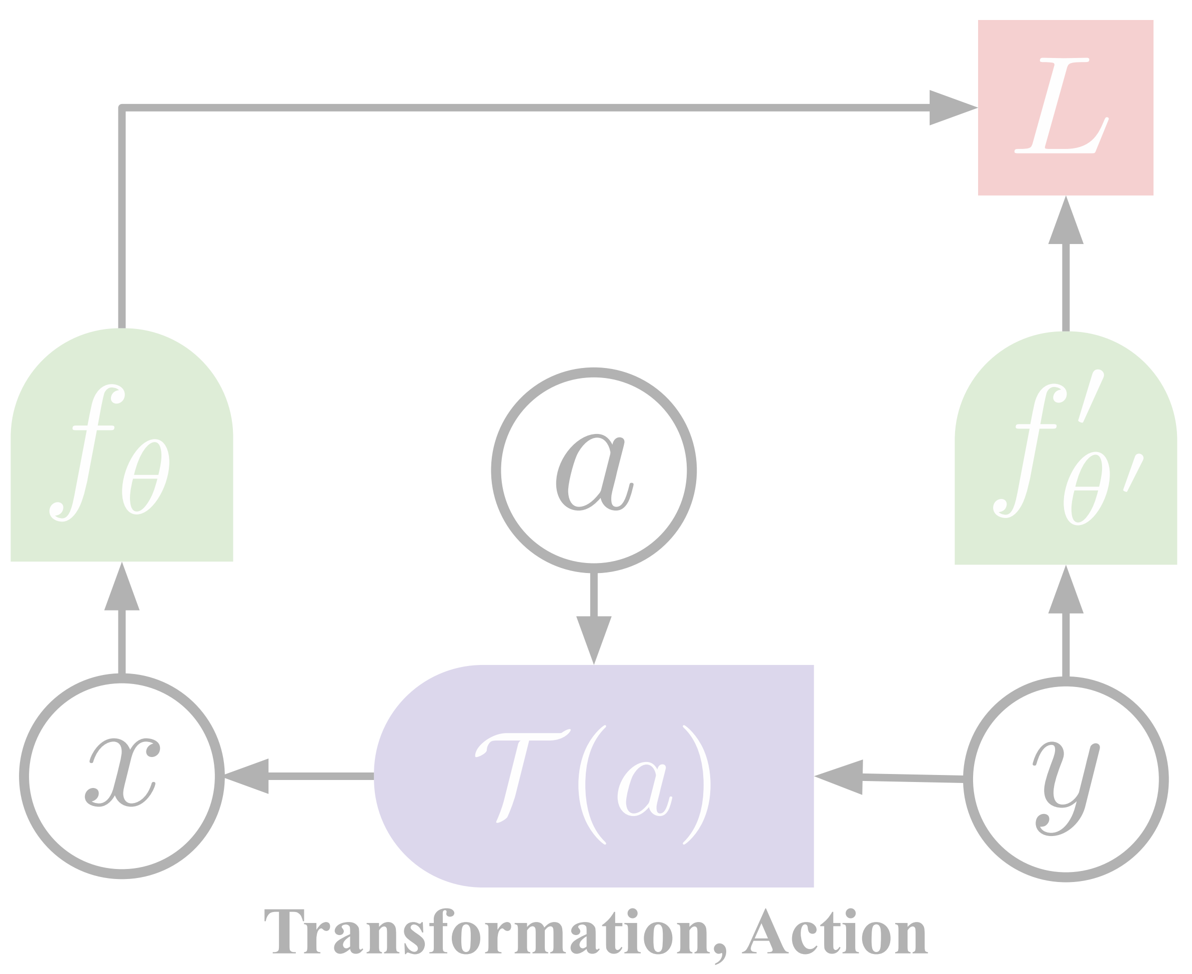
Self-Supervision: The CV Approach

Generative Architecture



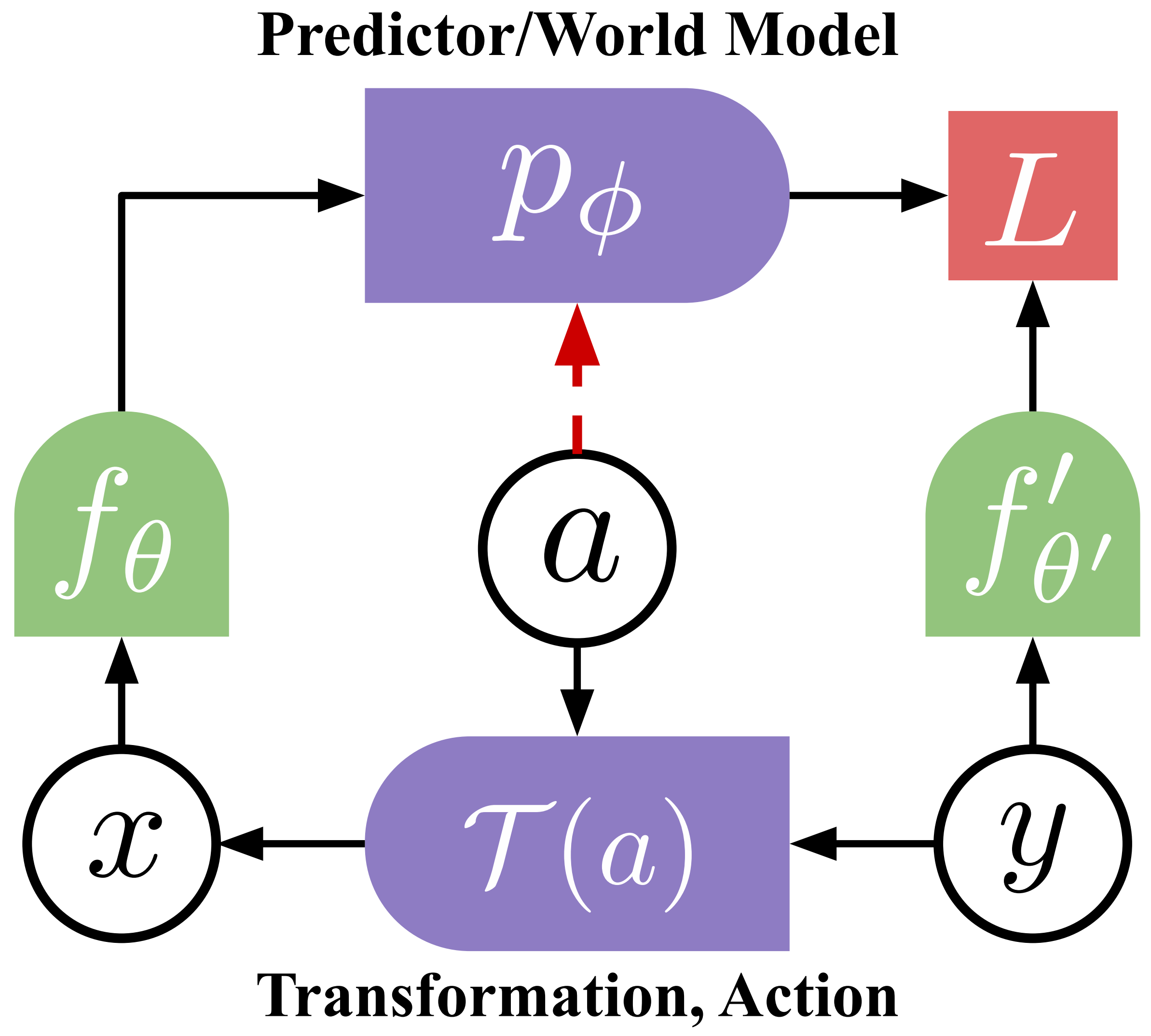
*Learns to invert a transformation in the **input** space.*
e.g., Masked Autoencoder (MAE)

Joint-Embedding Architecture



*Learns an invariant representation in the **latent** space.*
e.g., SimCLR / DINO

Joint-Embedding Predictive Architecture (JEPA)

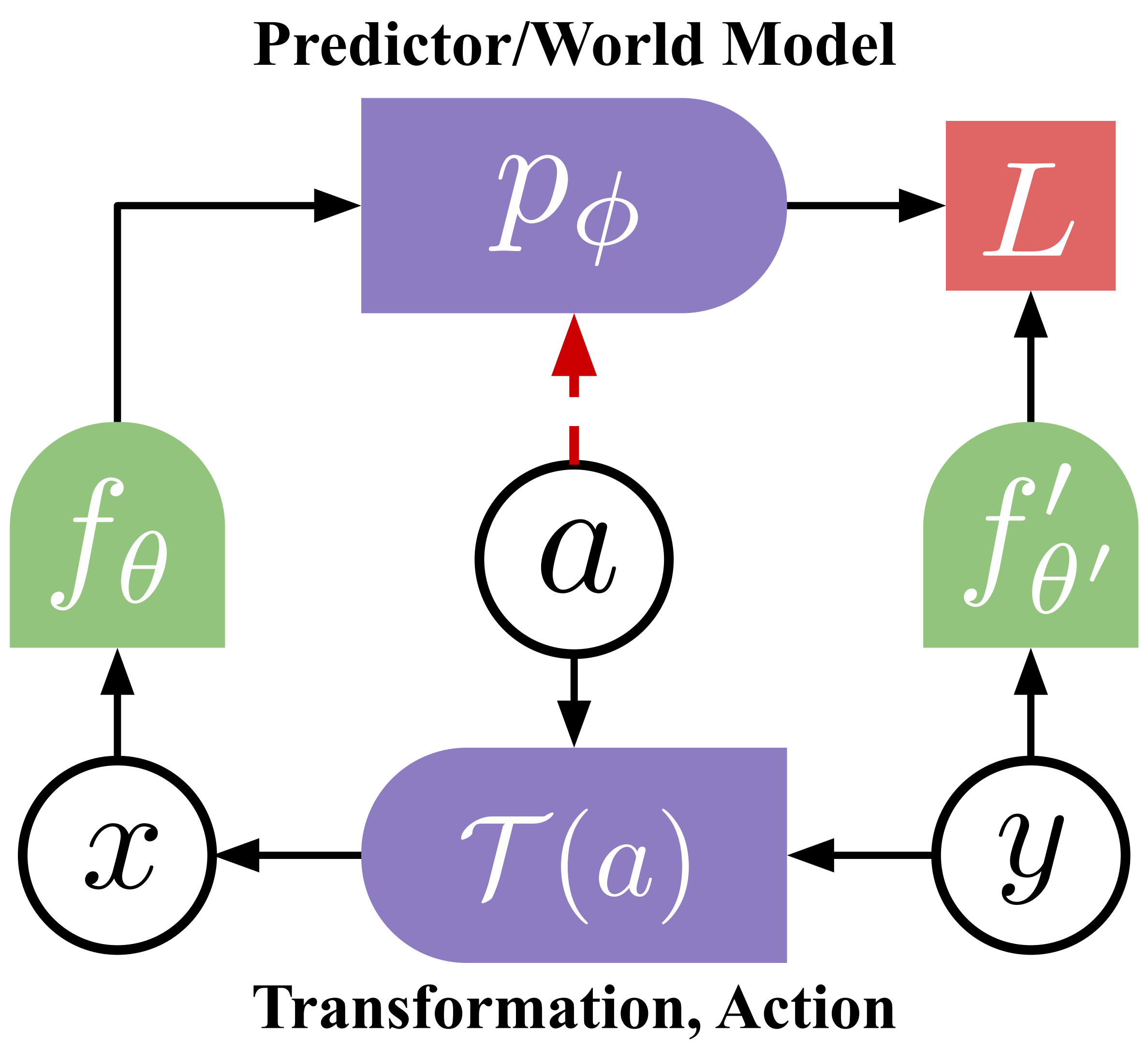


*Learns to predict the embeddings in the **latent** space.*
LeCun's vision for World Models.

arXiv: 2301.08243
arXiv: 2403.00504

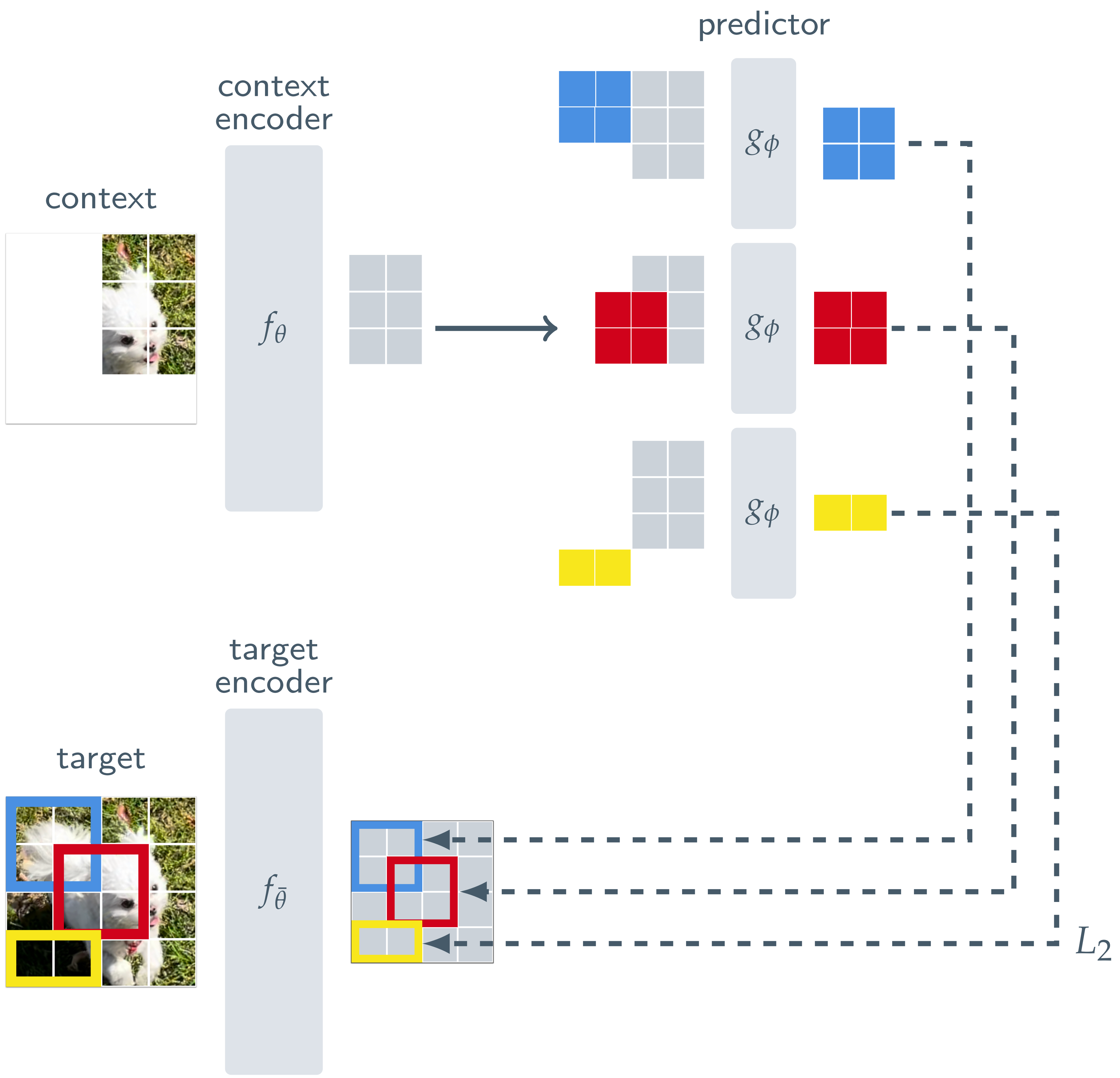
Self-supervision: The CV Approach

Joint-Embedding Predictive Architecture (JEPA)



Learns to predict the embeddings in the ***latent*** *space.*
LeCun's vision for World Models.

I-JEPA [arXiv: 2301.08243]

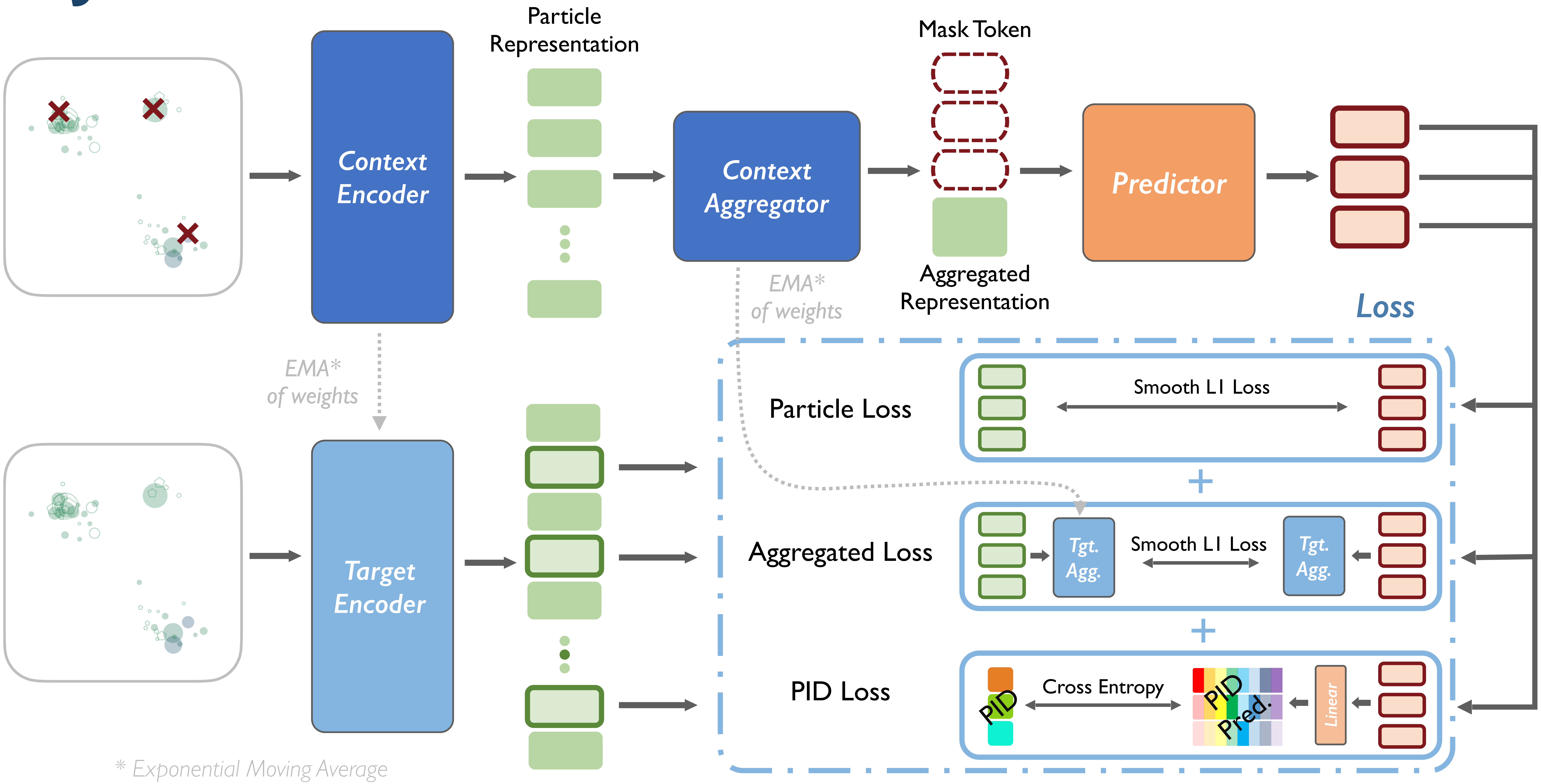


... predicts the embeddings of masked image patches in a (learned) latent space

P-JEPA

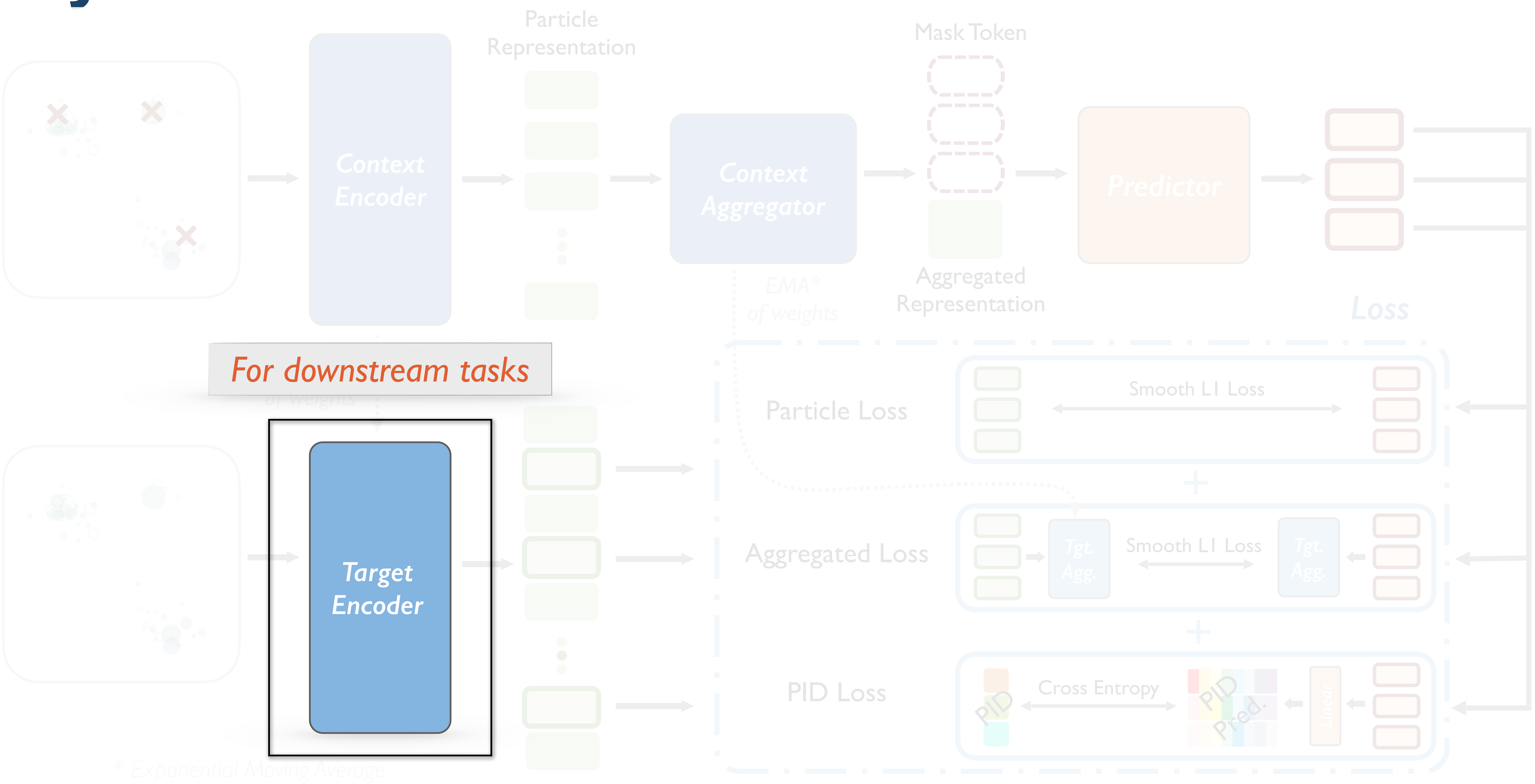
Work in progress with Q. Liu (刘齐斌), S. Wang (王书栋) and C. Li (李聪乔)

P-JEPA



* Exponential Moving Average

P-JEPA


Particle Representation
Mask Token
Context Encoder
Context Aggregator
Predictor
Aggregated Representation
EMA* of weights
Loss
For downstream tasks
Target Encoder
Particle Loss
Smooth L1 Loss
Aggregated Loss
Tgt. Agg.
PID Loss
Cross Entropy
PID
PID Pred.
* Exponential Moving Average

Particle Masking

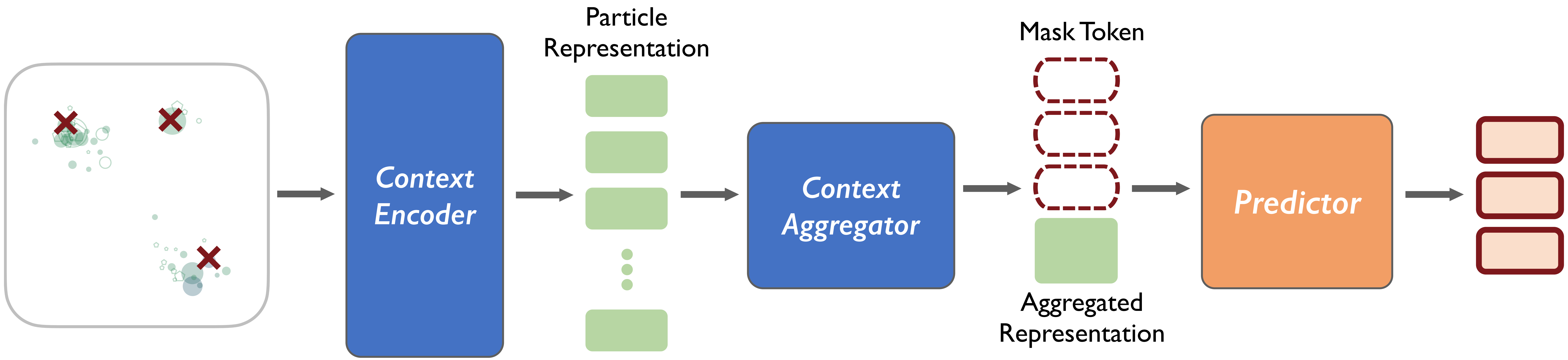
- The pre-training task in a nutshell:
  - predict the masked particles from the remaining ones
  - … but in the latent space

- Masking strategy:
  - randomly mask **30–50%** of the particles in a jet
  - the remaining particles serve as the context for the prediction
    - ==> input to the context encoder & predictor
  - the masked particles become the **target** to be predicted
    - ==> NOT seen by the context encoder & predictor
    - ==> the loss is computed only for the **target** particles

Context Encoder and Predictor

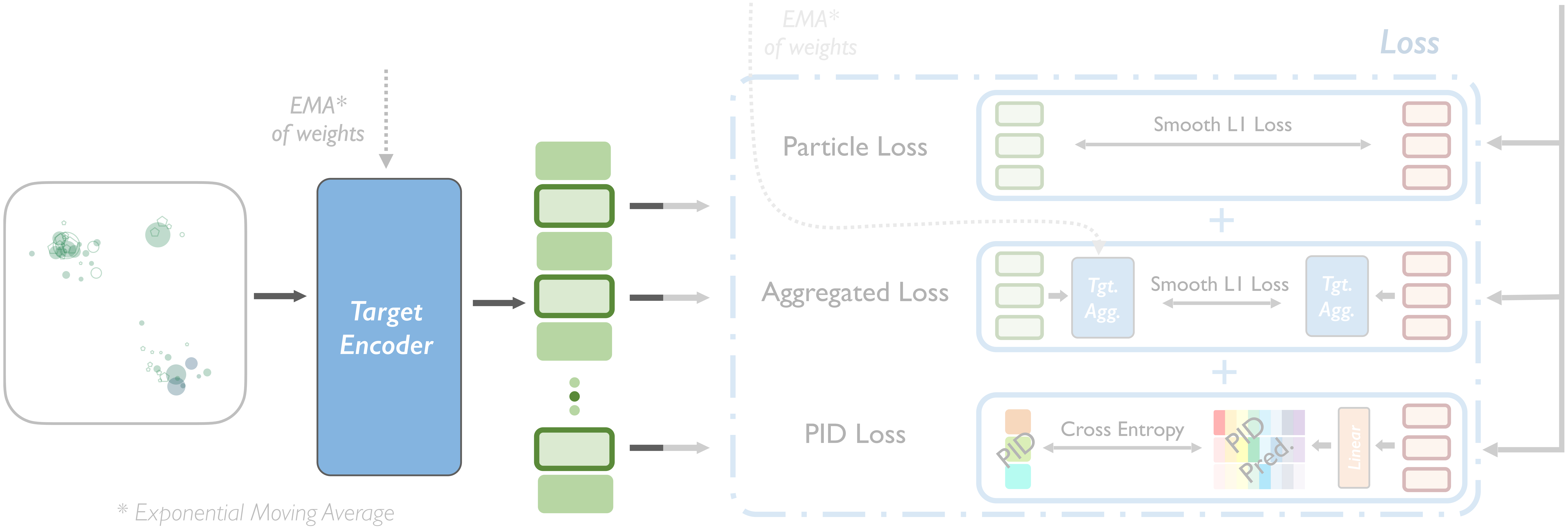
- Context encoder
  - large Particle Transformer (w/ pairwise features between context particles)
- Context aggregator
  - aggregates all context particles into a single token
- Predictor
  - plain Transformer, smaller than encoder
  - predicts the masked particles from the aggregated representation + mask tokens w/ pos. emb.

	Context Encoder + Aggregator	Predictor
Embed Dims	(512, 512, 512)	192
Pair Embed Dims	(64, 64, 64)	/
Num Heads	8	6
Num Blocks	16	4
Num Class Blocks	2	/



Target Encoder

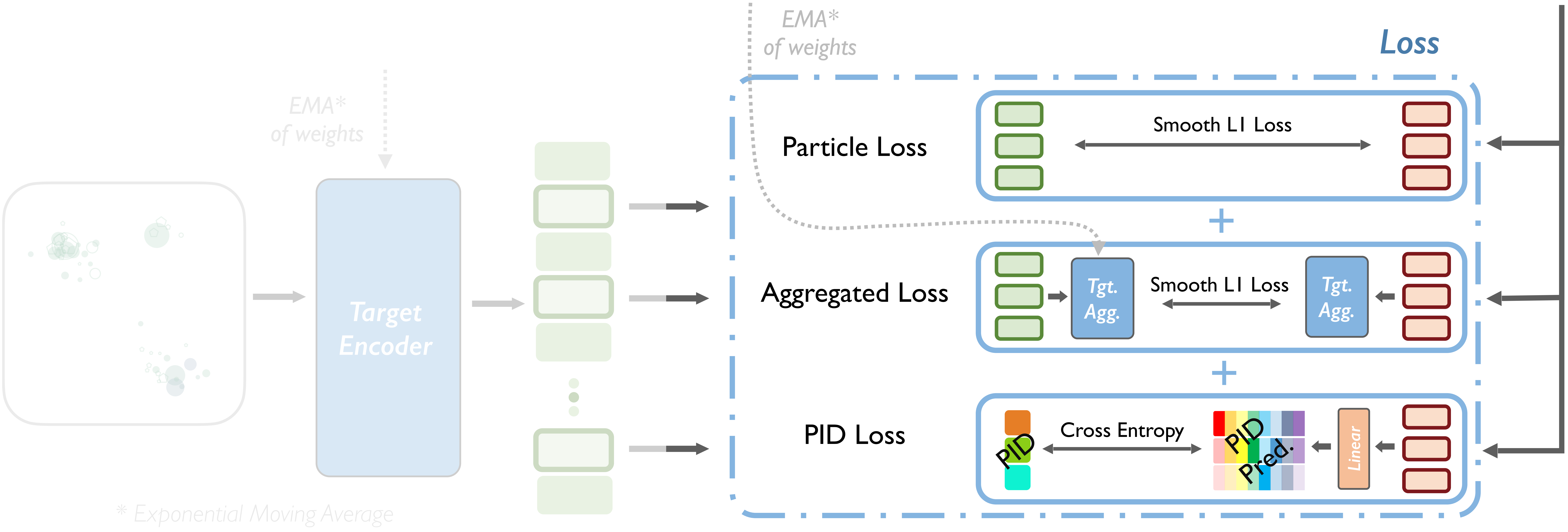
- A target encoder is used to derive the particle embeddings in the latent space for loss computation
  - processes the complete set of particles in a jet (i.e., context + target)
    - then only the embeddings of the target particles are picked for loss computation
  - updated via an exponential moving average of the context encoder's weights



* Exponential Moving Average

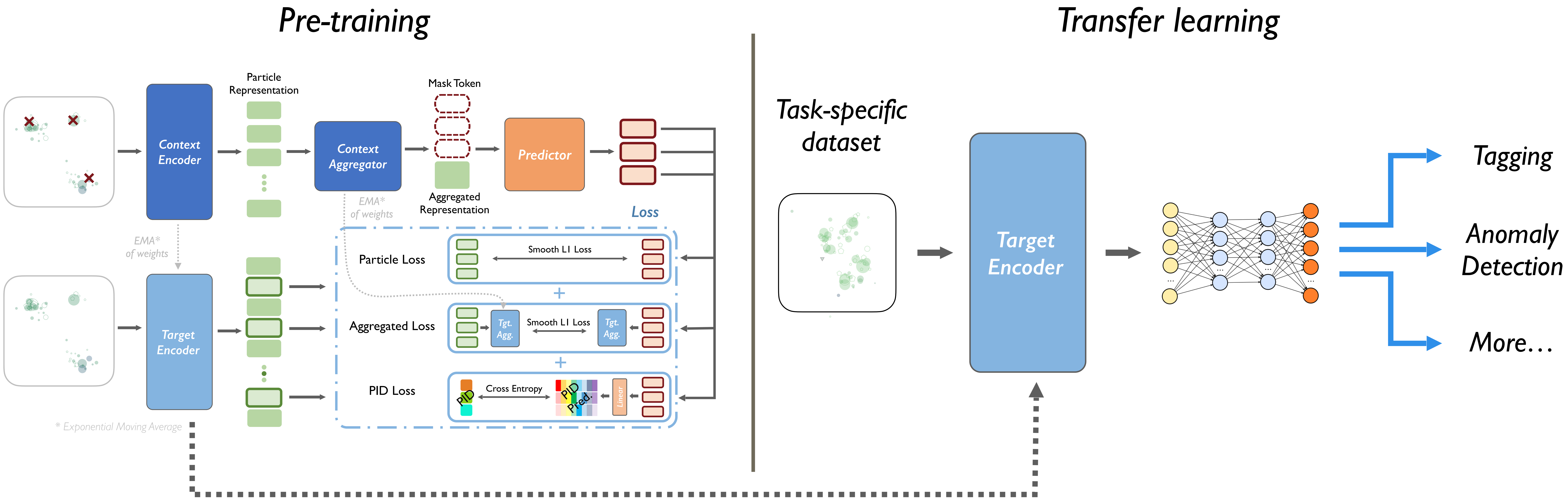
Pre-training Loss

- Loss = Particle Loss + Aggregated Loss + PID loss
  - Particle Loss: smooth L1 loss between the predicted embeddings and those from target encoder
  - Aggregated Loss: computed on the aggregated representations of target particles using the target aggregator
  - PID Loss: auxiliary task to predict the reconstructed PID of each masked particle from the predicted embeddings



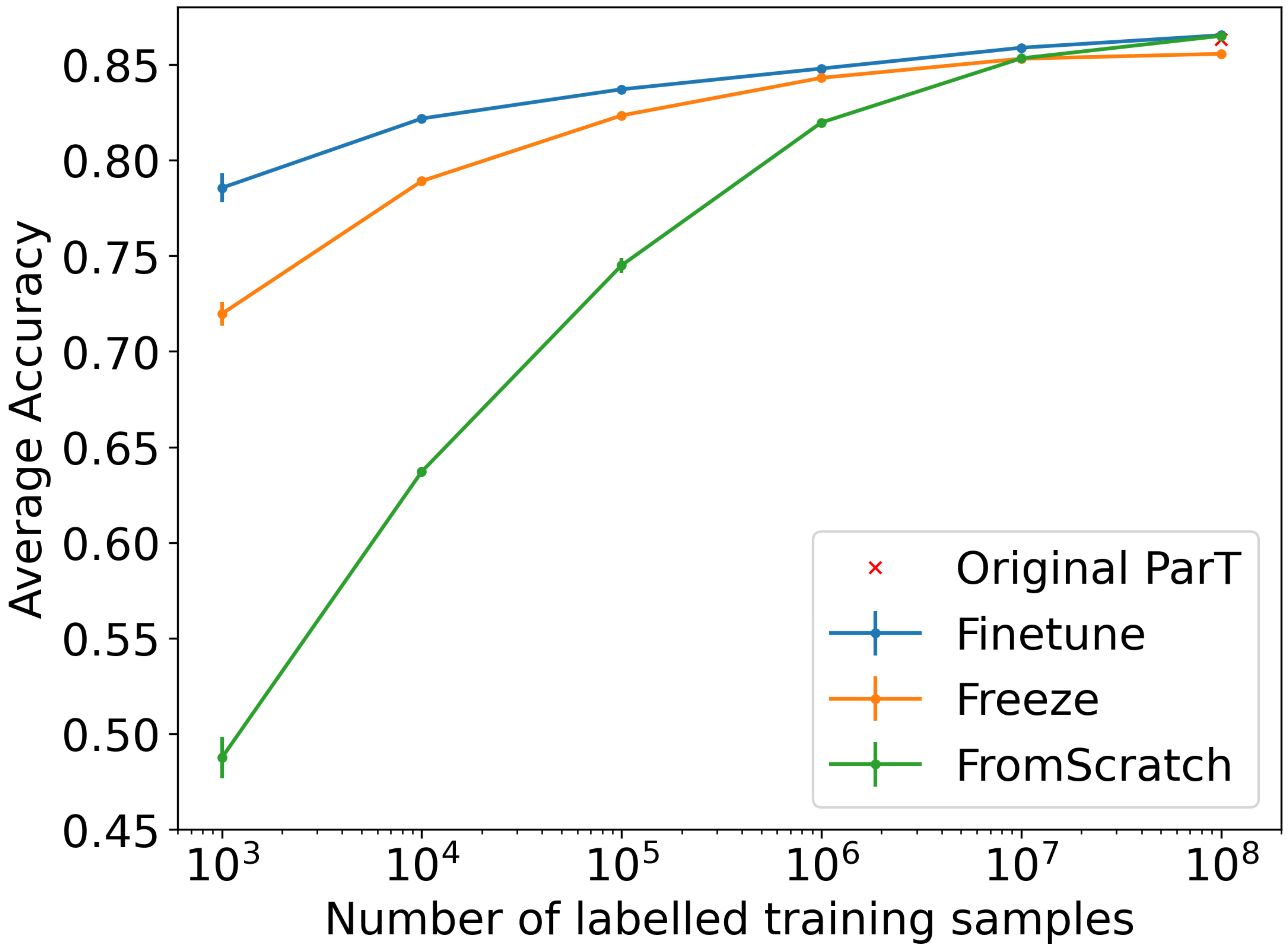
Pre-training and Transfer Learning

- The pre-training of the P-JEPA model can be performed on large-scale real data
  - we demonstrate this by pre-training P-JEPA on the JetClass dataset (100M jets) without using any truth labels
- Once pre-trained, the target encoder can be viewed as a foundation model
  - transfer learning to specific downstream tasks



Transfer Learning: Jet Tagging

- Benchmark: 10-class jet classification on JetClass



FineTune:

Encoder allowed to be slightly updated when trained with labelled jets for tagging

Freeze:

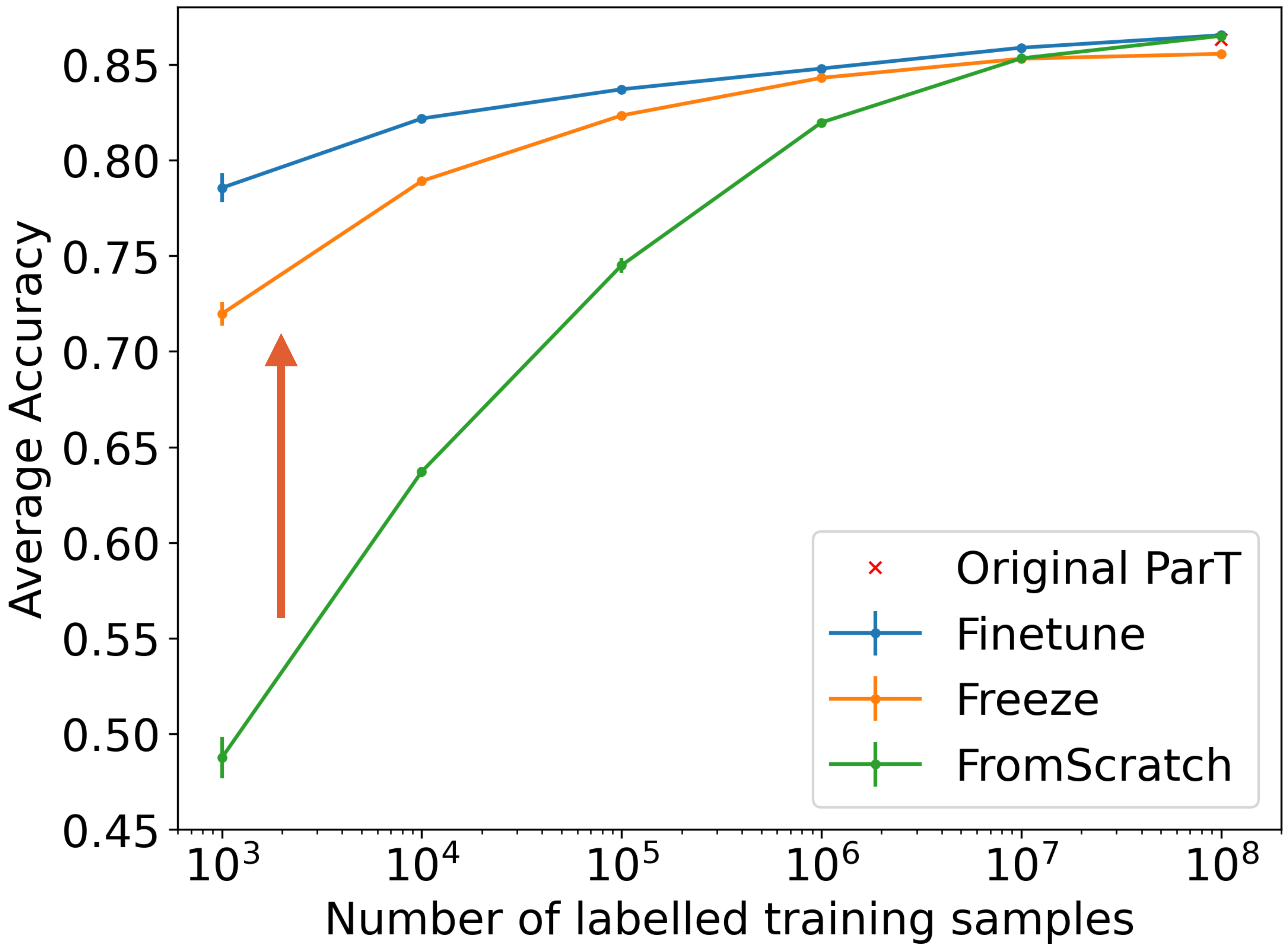
Encoder fixed when trained with labelled jets for tagging

FromScratch:

Same network architecture, but trained with labelled jets starting from randomly initialized weights

Transfer Learning: Jet Tagging

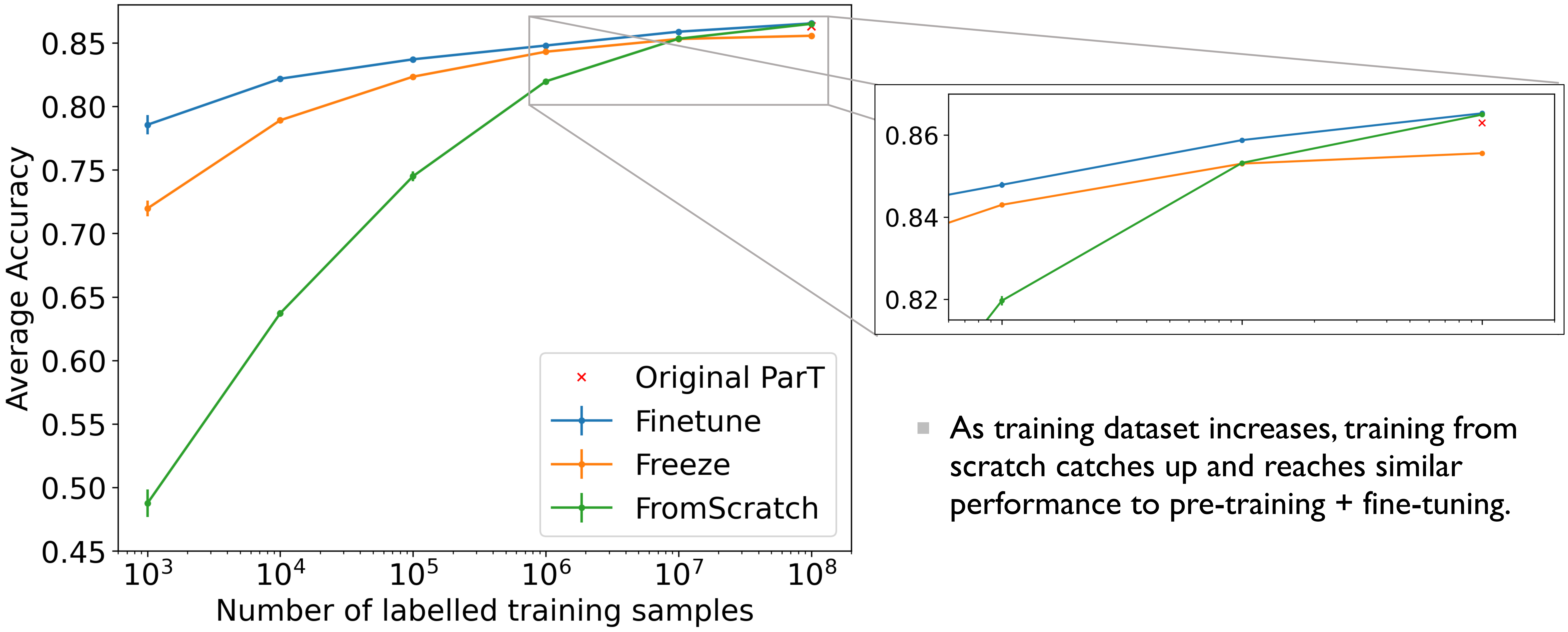
- Benchmark: 10-class jet classification on JetClass



- Pre-training + transfer learning shows a significant performance boost when labelled samples are limited.

Transfer Learning: Jet Tagging

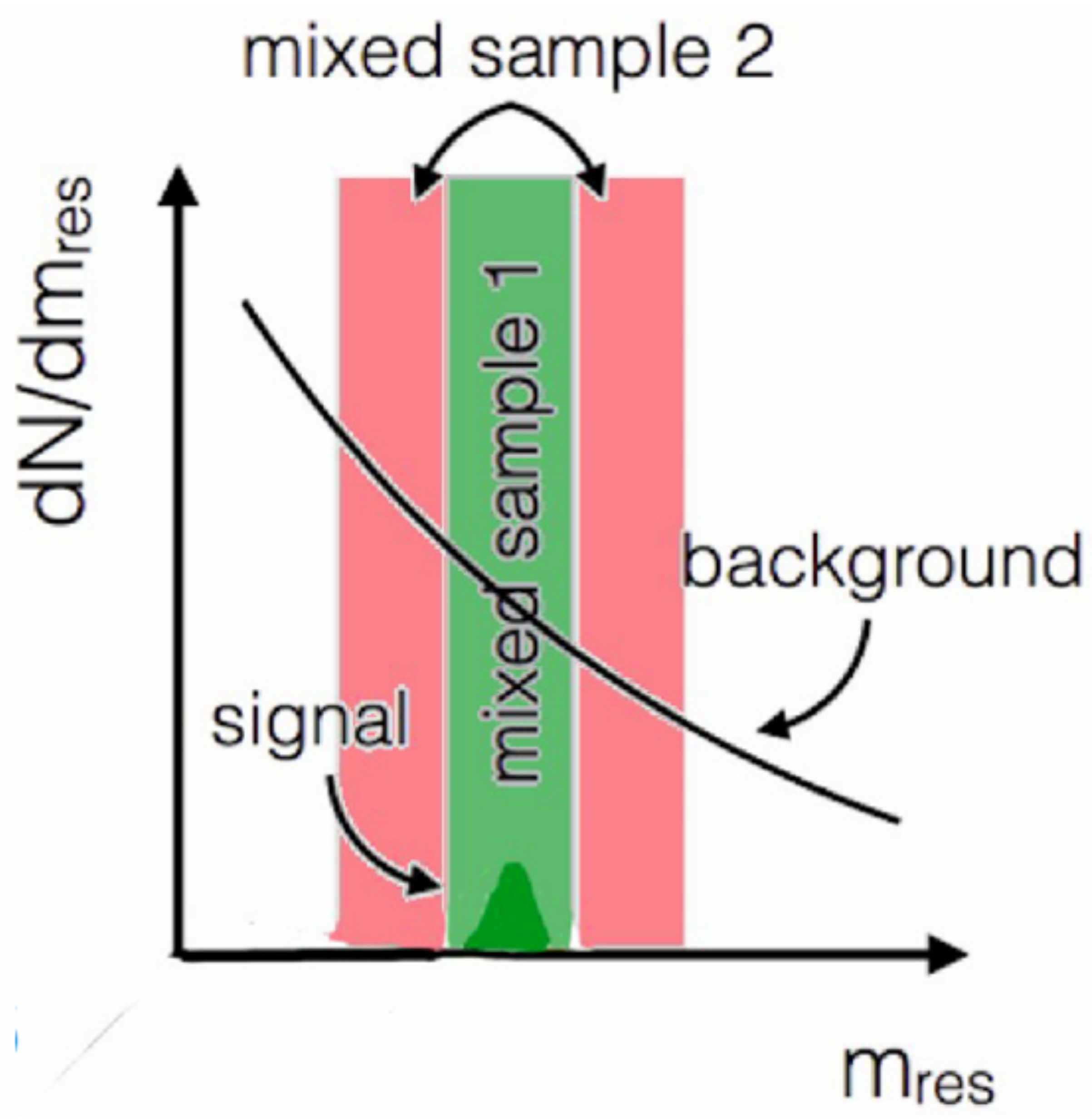
- Benchmark: 10-class jet classification on JetClass



- As training dataset increases, training from scratch catches up and reaches similar performance to pre-training + fine-tuning.

TRANSFER LEARNING: ANOMALY DETECTION

- Anomaly Detection (AD): model-agnostic search for new physics signals
- A classic paradigm for AD: CWoLa (classification without labels)
  - trains a classifier to distinguish two mixed samples
    - e.g., mass window (signal enriched) vs mass sideband (background enriched)
    - the classifier effectively becomes a signal vs background discriminator, thus can be used to enhance signal purity
  - allows to detect unknown signals purely from data



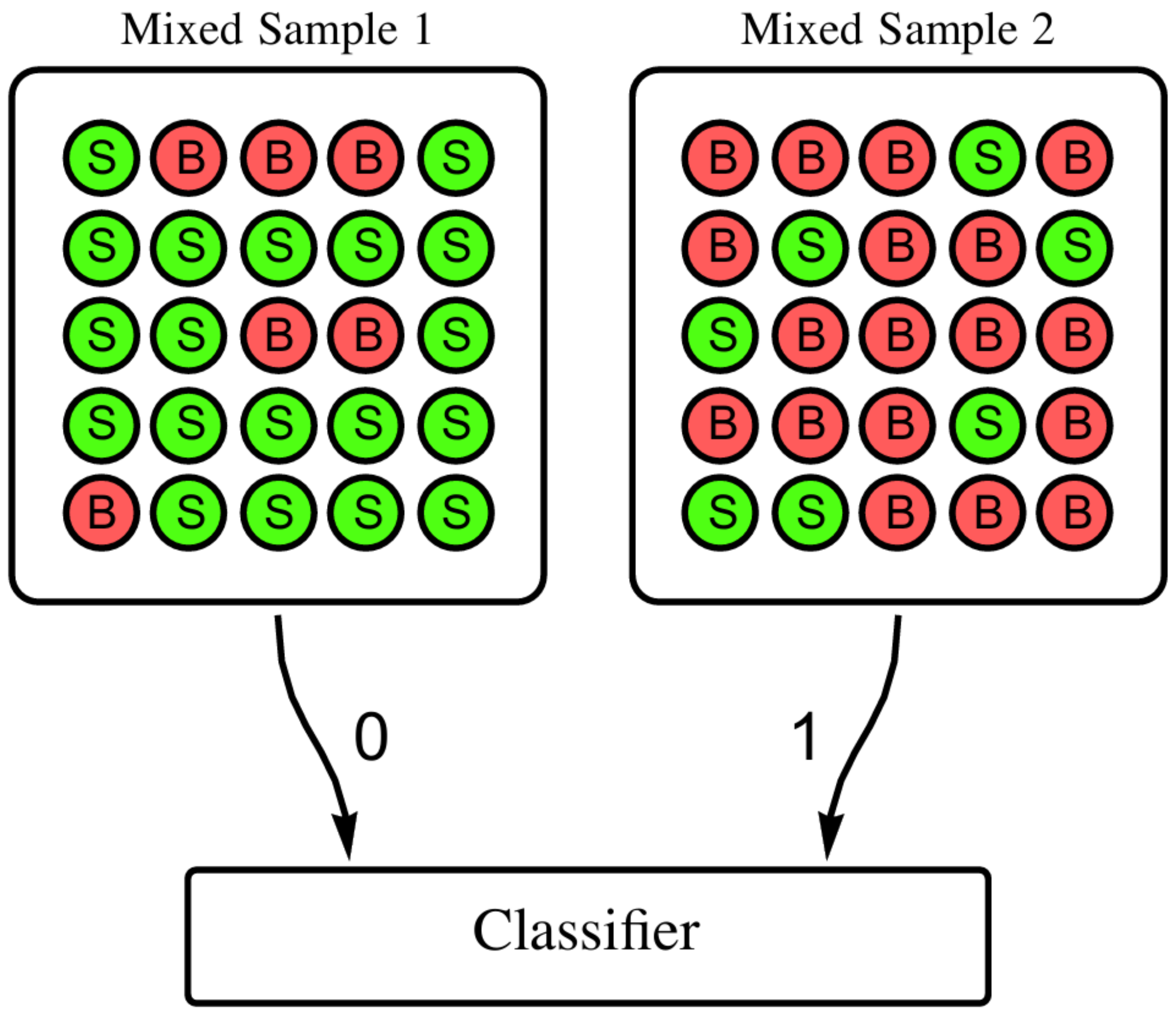


Figure Credit

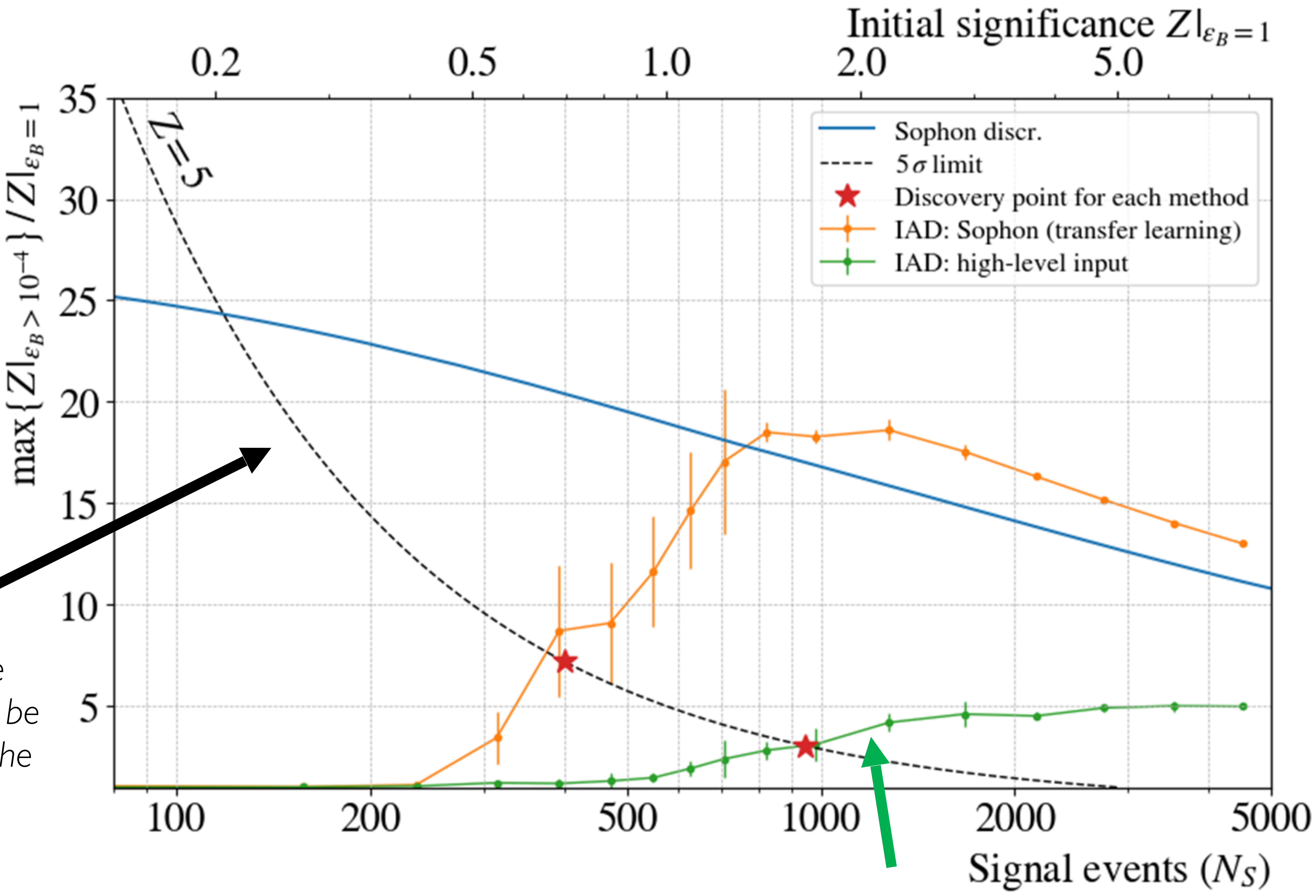
Transfer Learning: Anomaly Detection

- Traditionally AD was performed using only high-level features (e.g., jet mass, substructures) as inputs
- Machine-learned representations captures richer information of a jet, thus can improve the performance of AD
- We benchmark this using the IAD [arXiv:2210.14924] framework
  - idealized setup for the mixed samples: background only vs background + signal
    - background in the two mixed samples are drawn from the same distribution, no need to worry about e.g., mass dependency and interpolation into the mass window etc.
  - performance of the learned features evaluated by the significance improvement metric
    - i.e., the maximal gain in significance by varying the classifier cut

Transfer Learning: Anomaly Detection

Significance improvement

(better)



How much does the significance need to be increased to reach the 5σ discovery

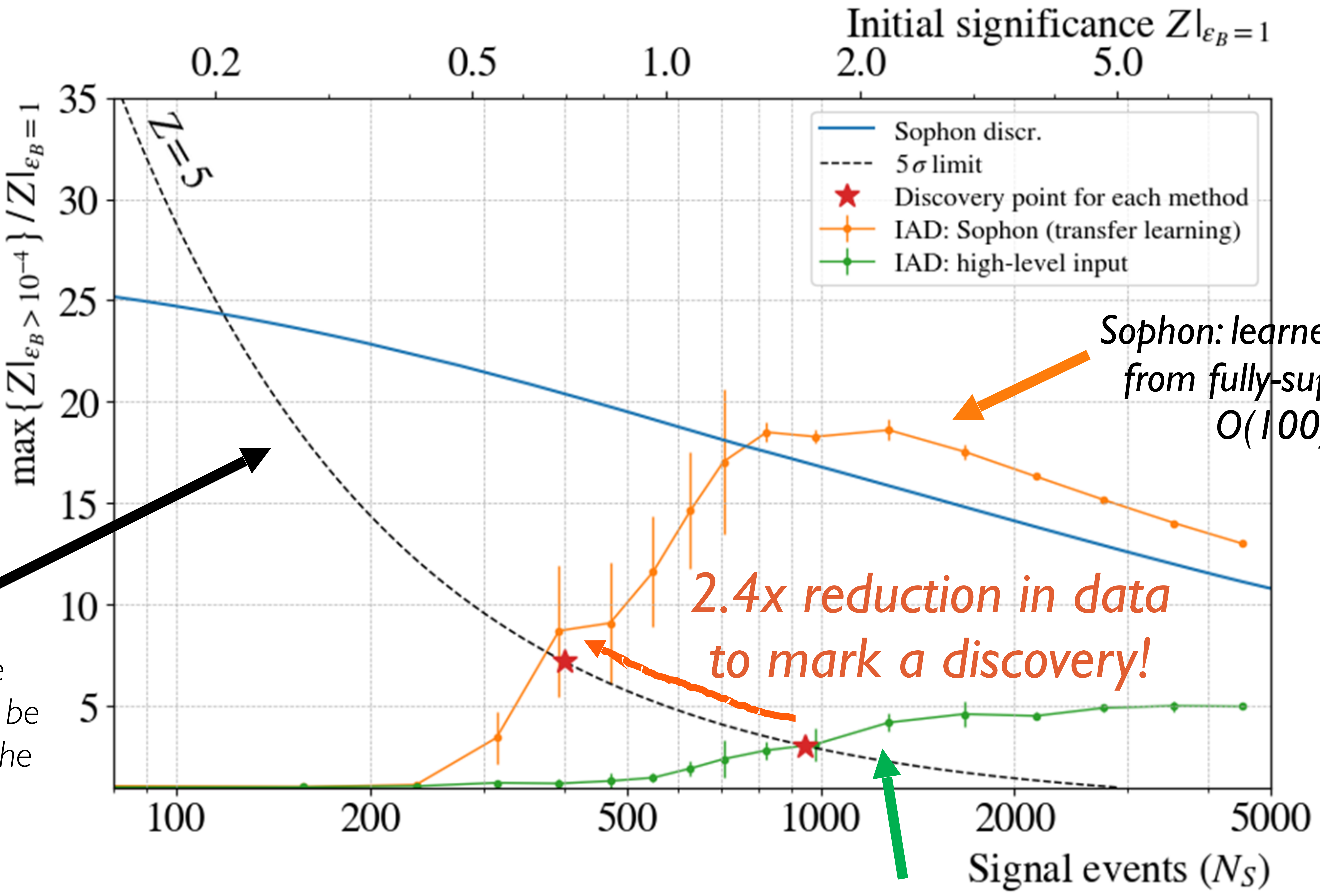
Classifier using high-level observables

[PRL, 121 (2018) 24, 241803]

Transfer Learning: Anomaly Detection

Significance improvement

(better)



Sophon: learned jet representations from fully-supervised training on O(100) types of jets

How much does the significance need to be increased to reach the 5σ discovery

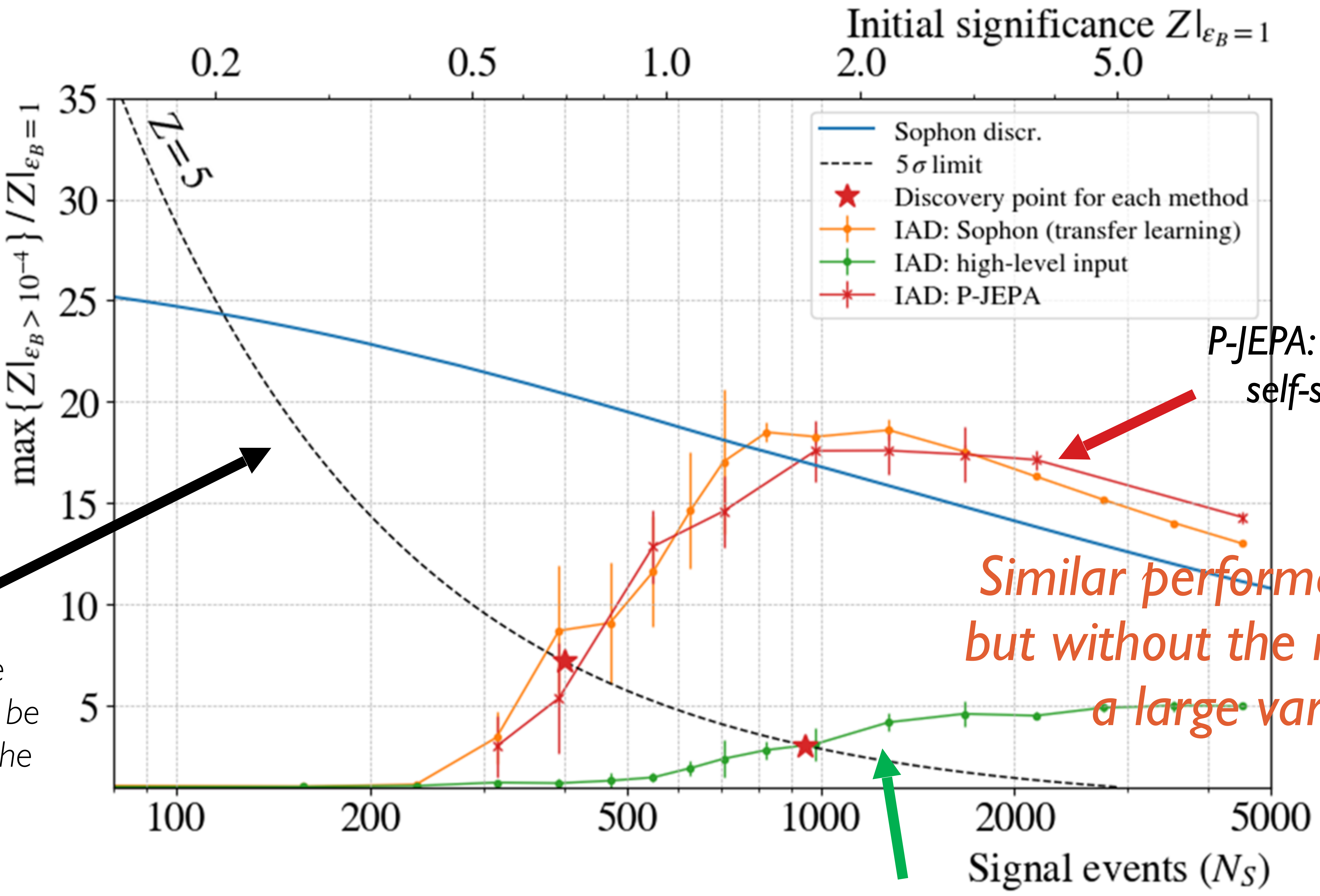
Classifier using high-level observables

[PRL, 121 (2018) 24, 241803](https://doi.org/10.1103/PhysRevLett.121.241803)

Transfer Learning: Anomaly Detection

Significance improvement

(better)



P-JEPA: representations from self-supervised learning

How much does the significance need to be increased to reach the 5σ discovery

Similar performance to Sophon, but without the need to simulate a large variety of jets

Classifier using high-level observables

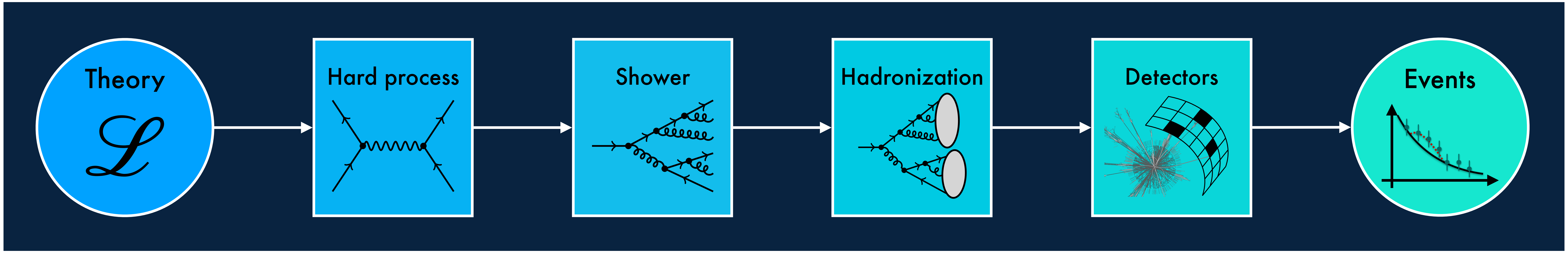
[PRL, 121 (2018) 24, 241803](PRL, 121 (2018) 24, 241803)

Summary & Outlook

- Tremendous progress in machine learning for jet physics in recent years
- Towards a foundation model for jet physics
  - Sophon: pre-training via fully supervised classification over a large variety of jets
  - P-JEPA: pre-training via self-supervised learning on unlabelled dataset
- Outlook: a foundation model for all of the LHC?

Generation, Simulation, …

Forward



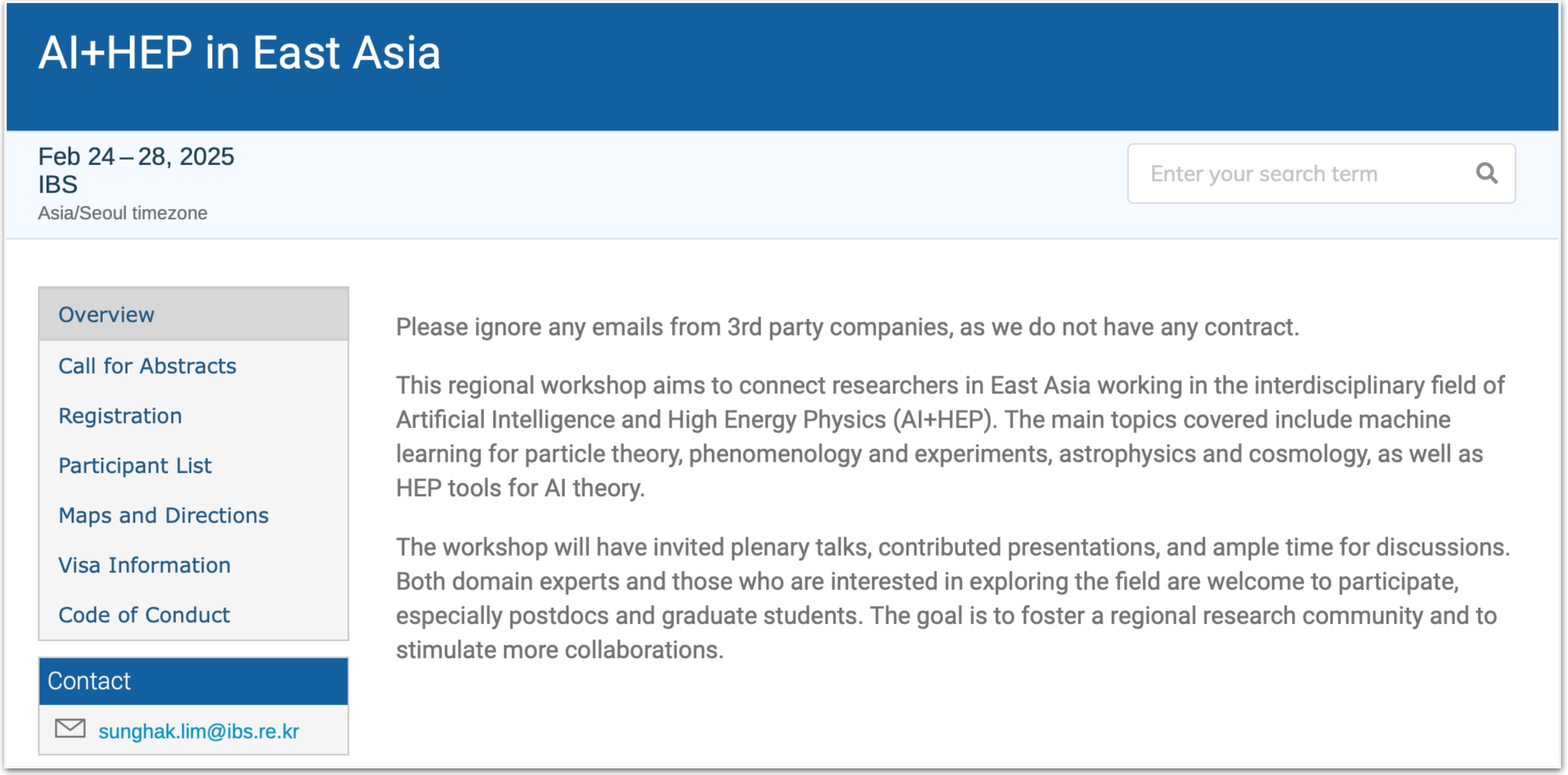
Inverse

Reconstruction, Unfolding, …

Credits: R. Winterhalder

ADVERTISEMENT

- AI+HEP workshop in East Asia
  - Feb. 24–28, 2025 at IBS (Korea)
  - Indico:
    - https://indico.ibs.re.kr/event/789/
  - Organizing Committee:
    - Tianji Cai (蔡恬吉, SLAC)
    - Sung Hak Lim (CTPU-PTC, IBS)
    - Huilin Qu (CERN)
  - Advisory Committee:
    - Mihoko M. Nojiri (KEK)
    - David Shih (Rutgers)

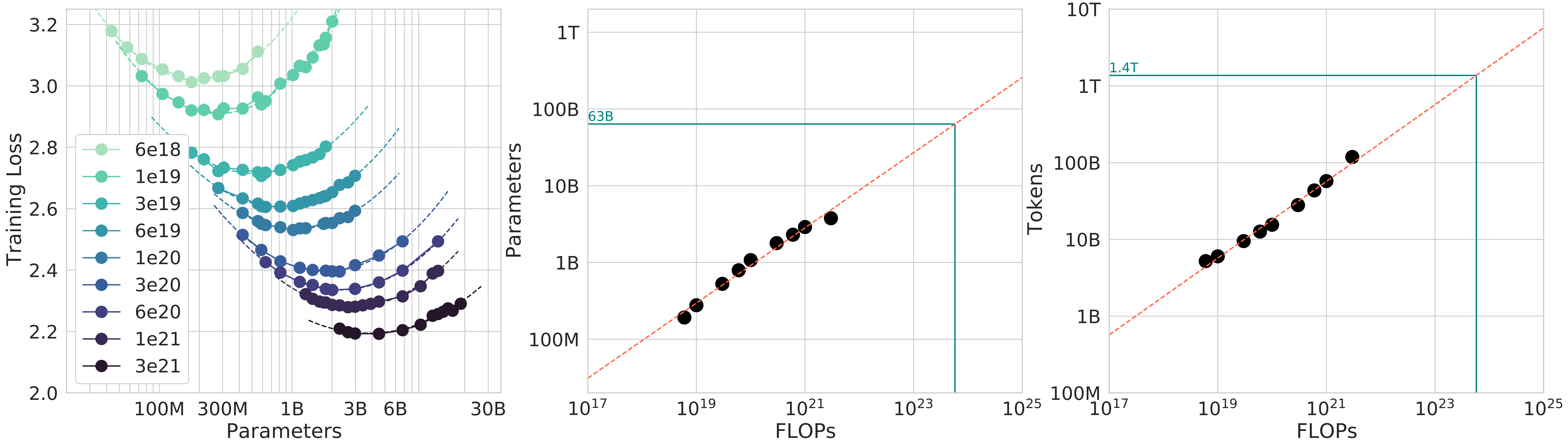


Hope to see many of you there!

Extras

Scaling Law

- How far can we push the performance with bigger models, larger datasets, and more computing power?
- For language models — neural scaling law [arXiv: 2001.08361, 2203.15556]



  - empirical power law scaling of the loss as a function of the compute (C), dataset size (D) and model parameters (N)
  - once established, can be extrapolated to determine the best dataset size & parameter combination under a fixed compute budget
- Would be interesting to see the scaling law for jets — but very computation intensive…