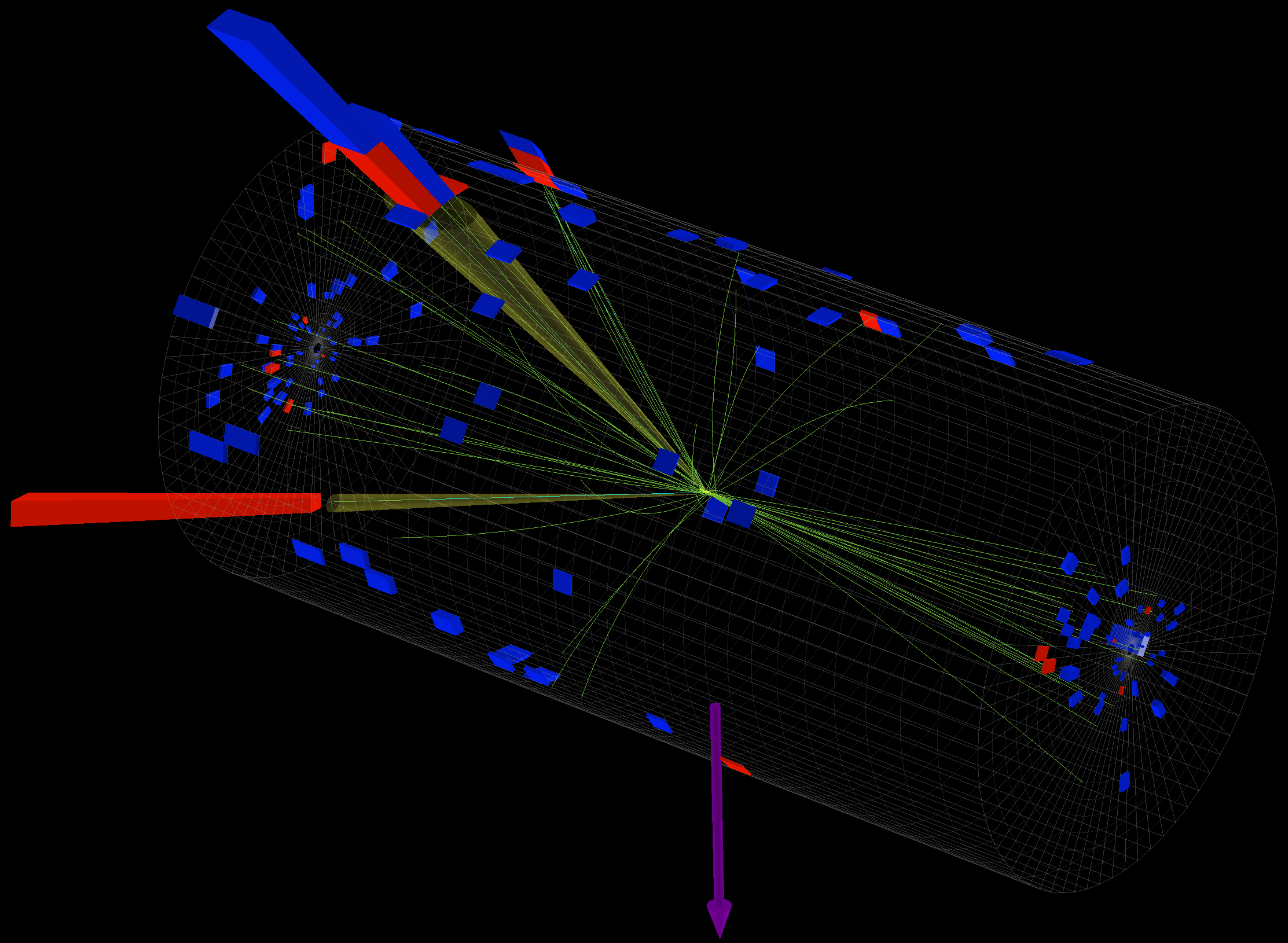
Towards a Foundation Model for Jet Physics

Huilin Qu

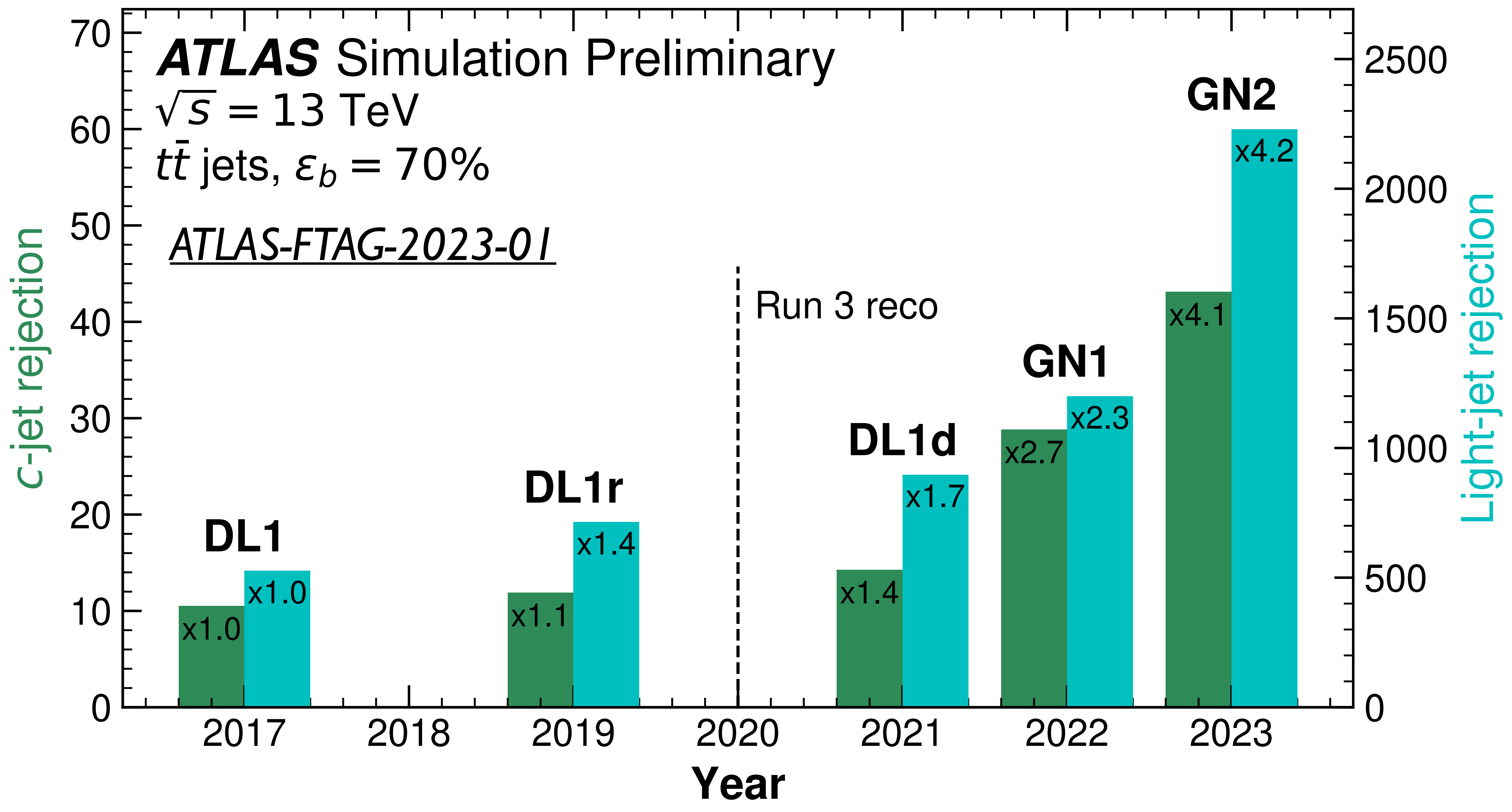


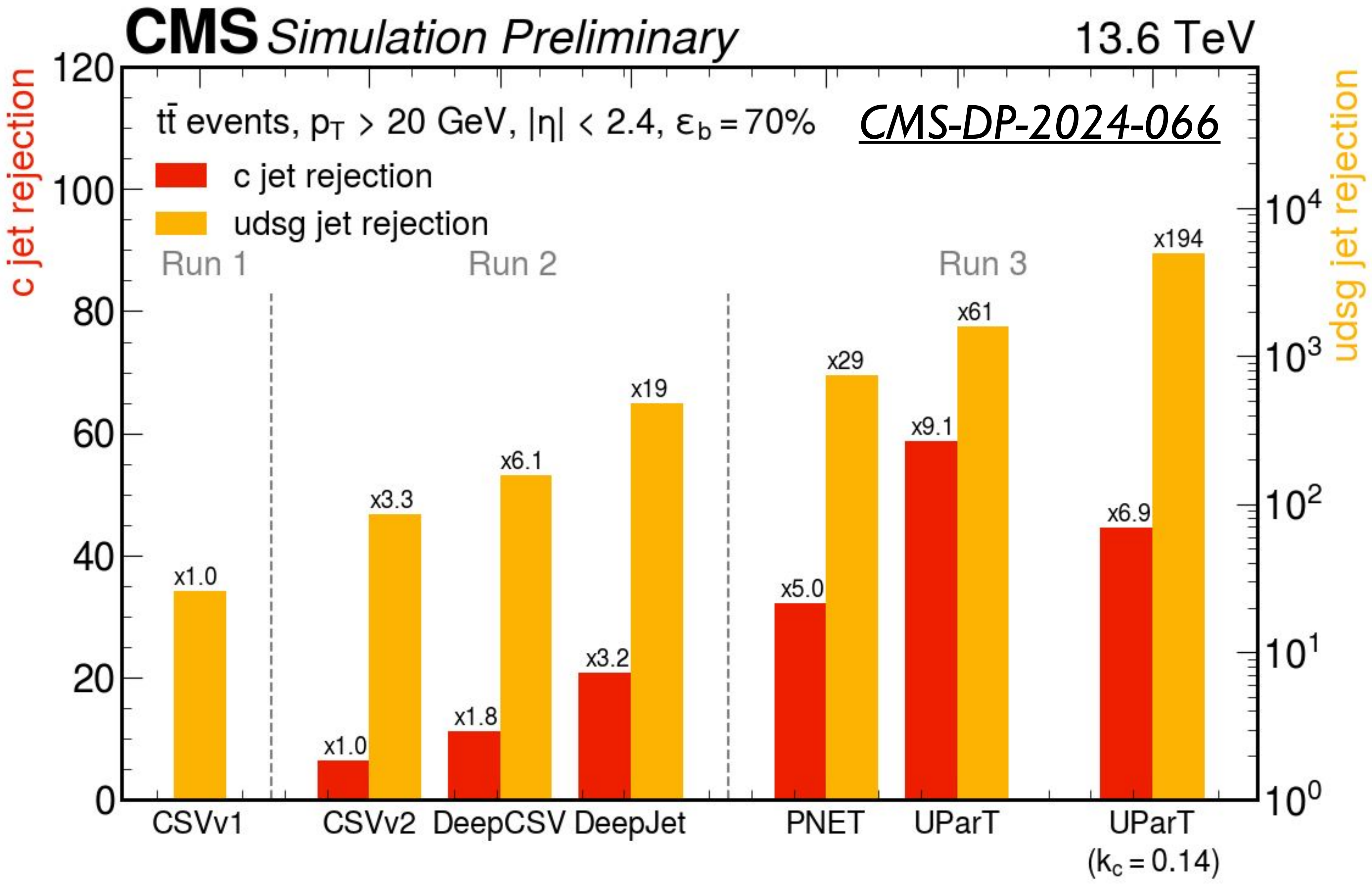
EPS-HEP

July 10, 2025

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

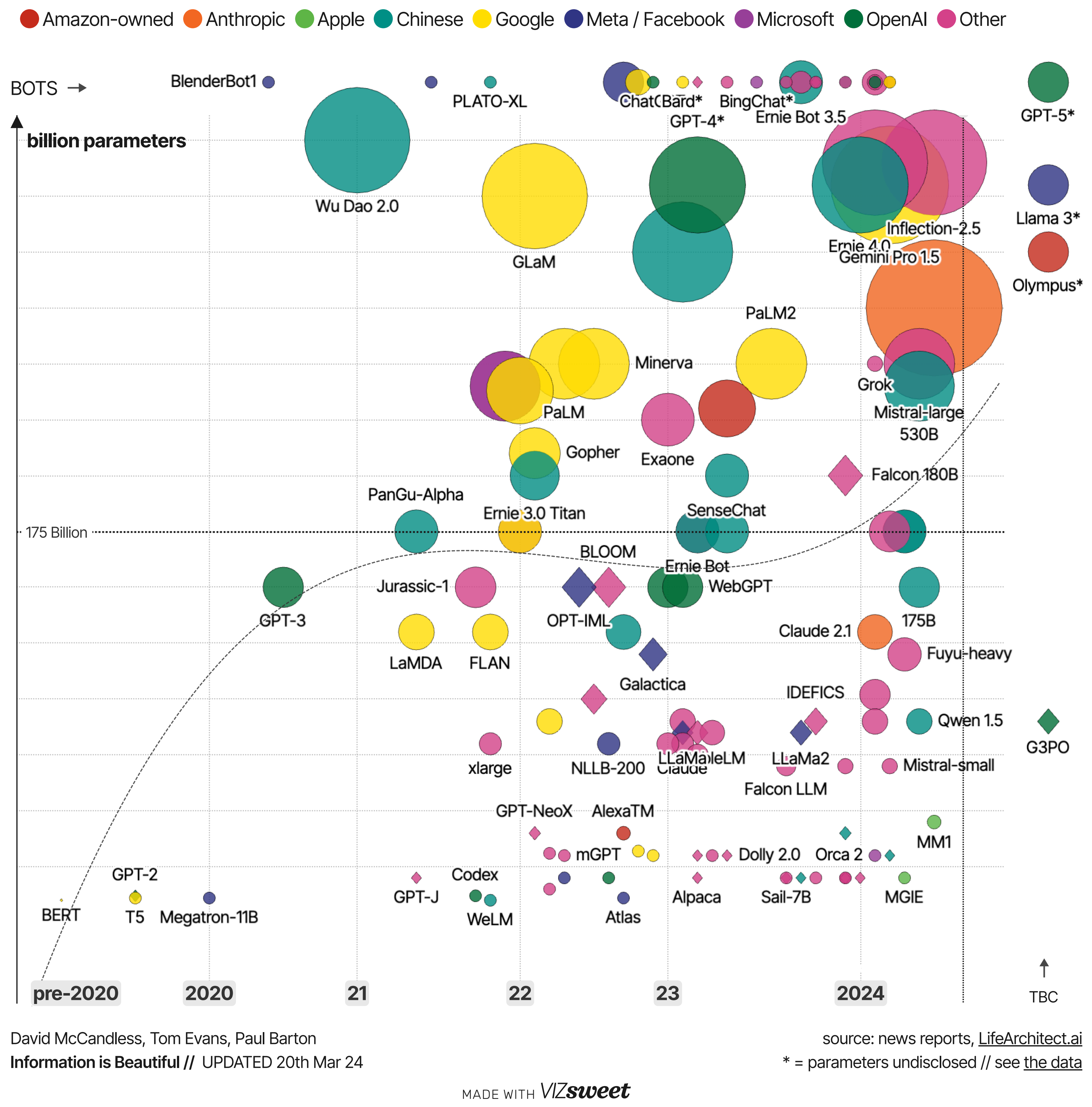




- **A driving force – advanced machine learning (ML) techniques**

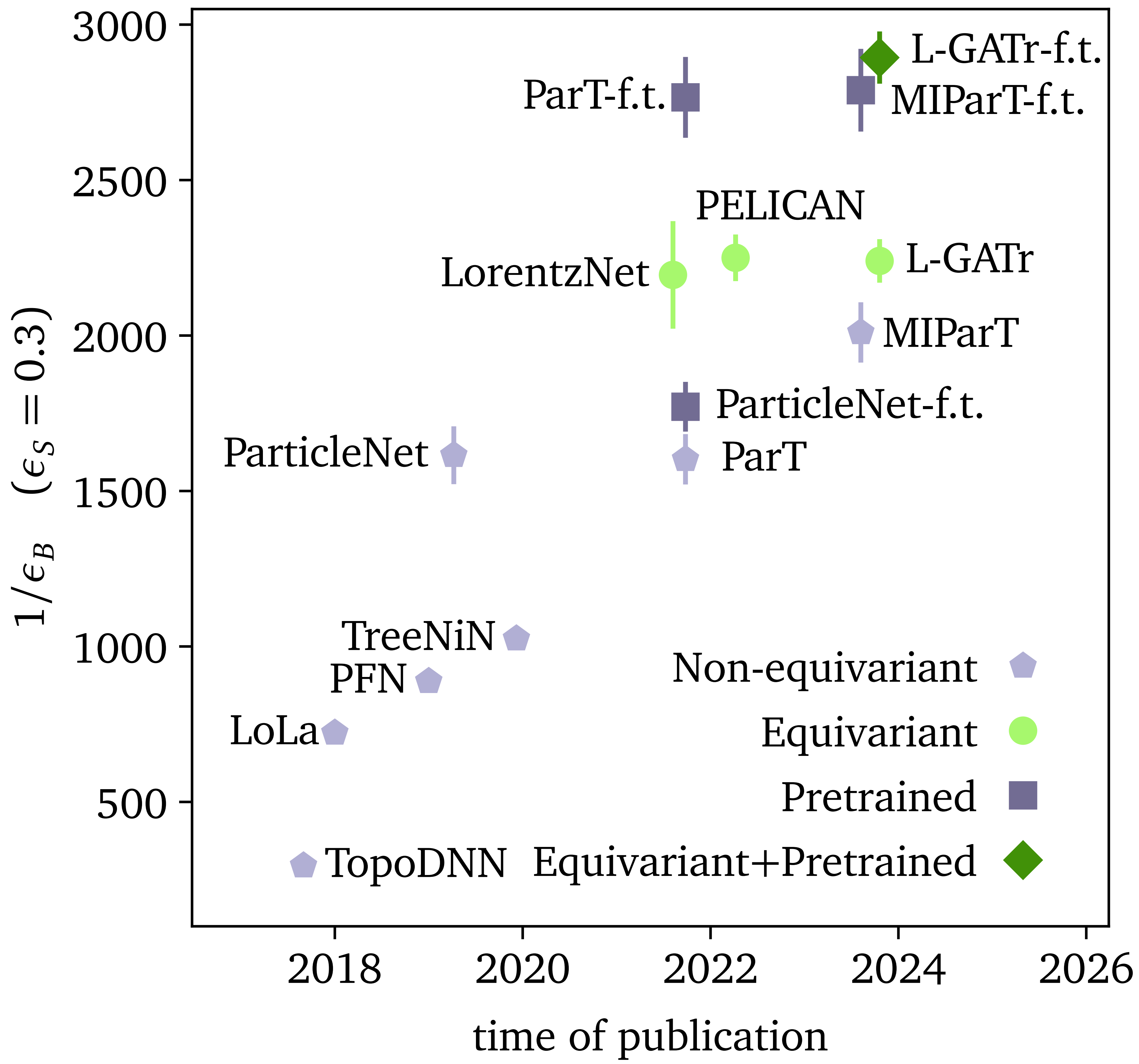
SCALING UP?

Natural language models



Source: informationisbeautiful.net

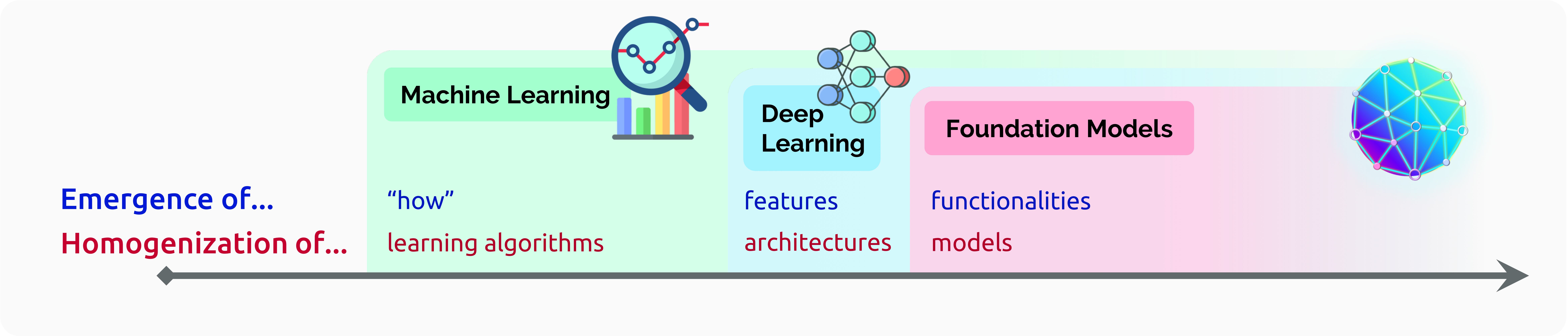
HEP models (jet tagging)



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

Foundation Model

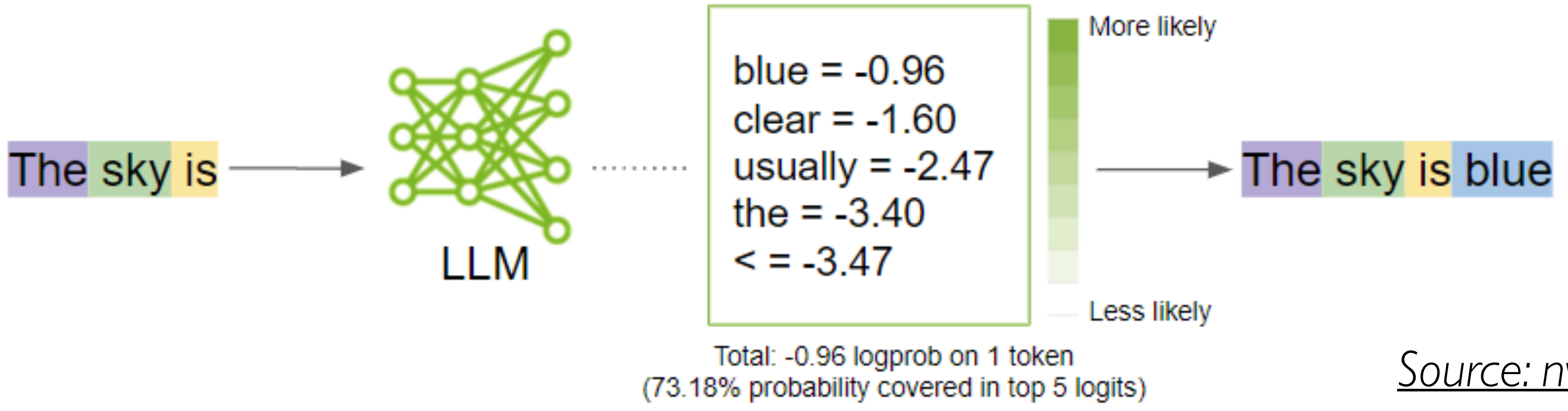
"A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks."



On the Opportunities and Risks of Foundation Models
[arXiv: 2108.07258]

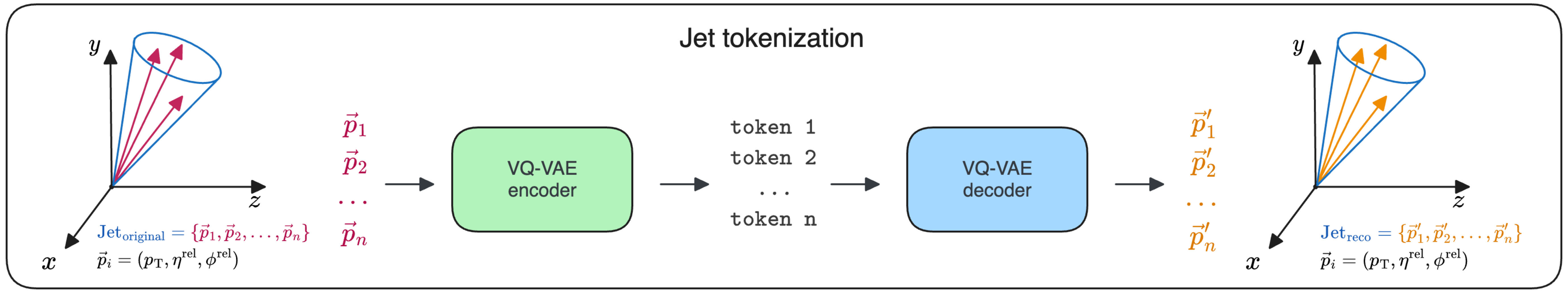
Self-supervision: Next Token Prediction

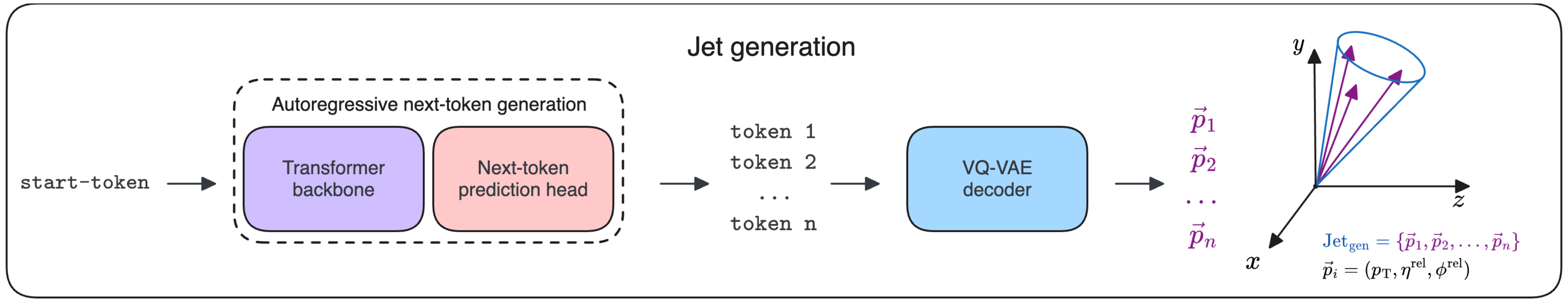
- The LLM way: **(autoregressive) language modeling**
  - i.e., next token prediction



Source: nvidia.com

- An attempt for jets: OmniJet-α [MLST 5 (2024) 035031]



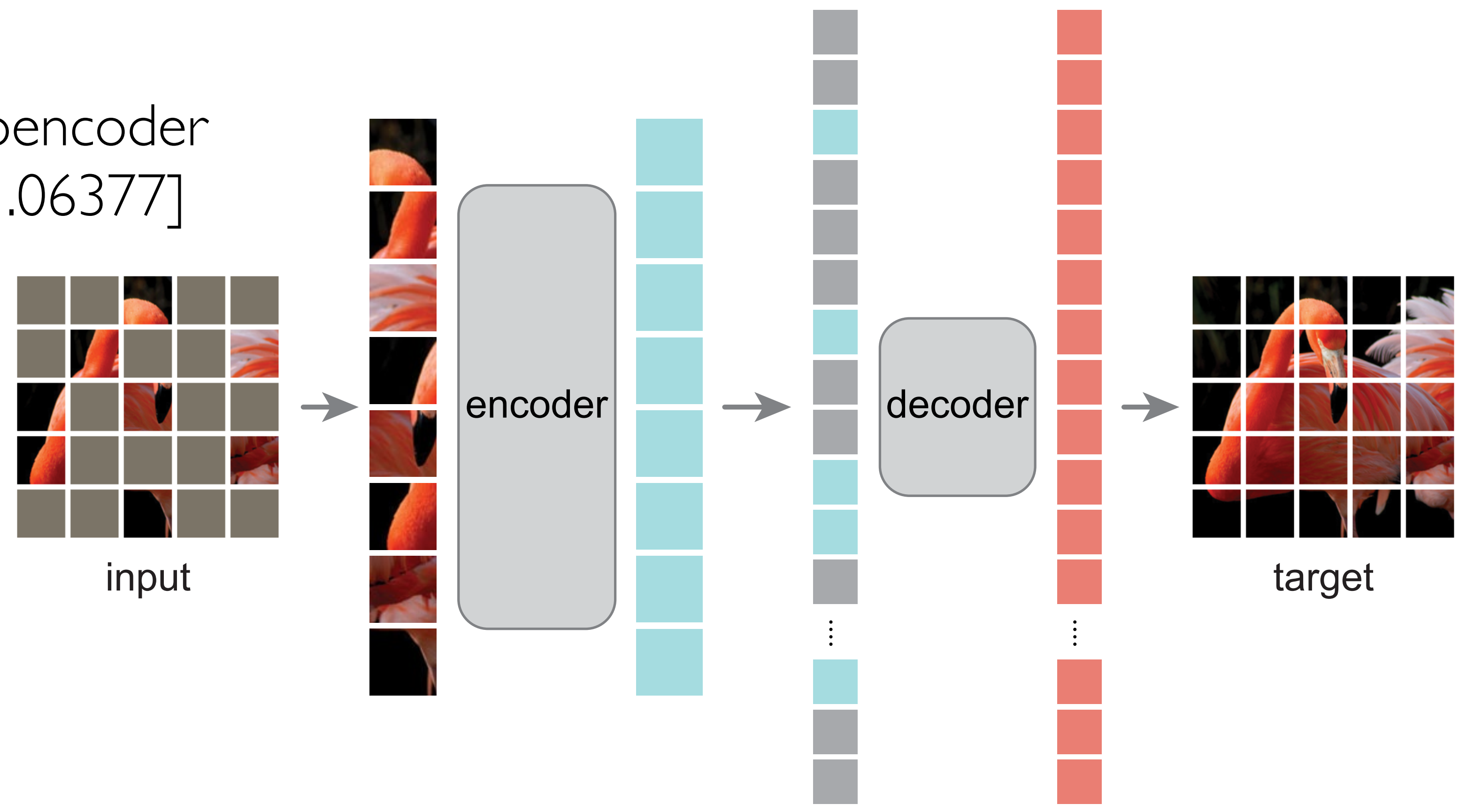


- Probably not the most natural approach:
  - requires (discrete) tokenization of high-dimensional numerical inputs
  - needs to impose an ordering on jet constituent particles, which are intrinsically permutation invariant

SELF-SUPERVISION: MASKED MODELING

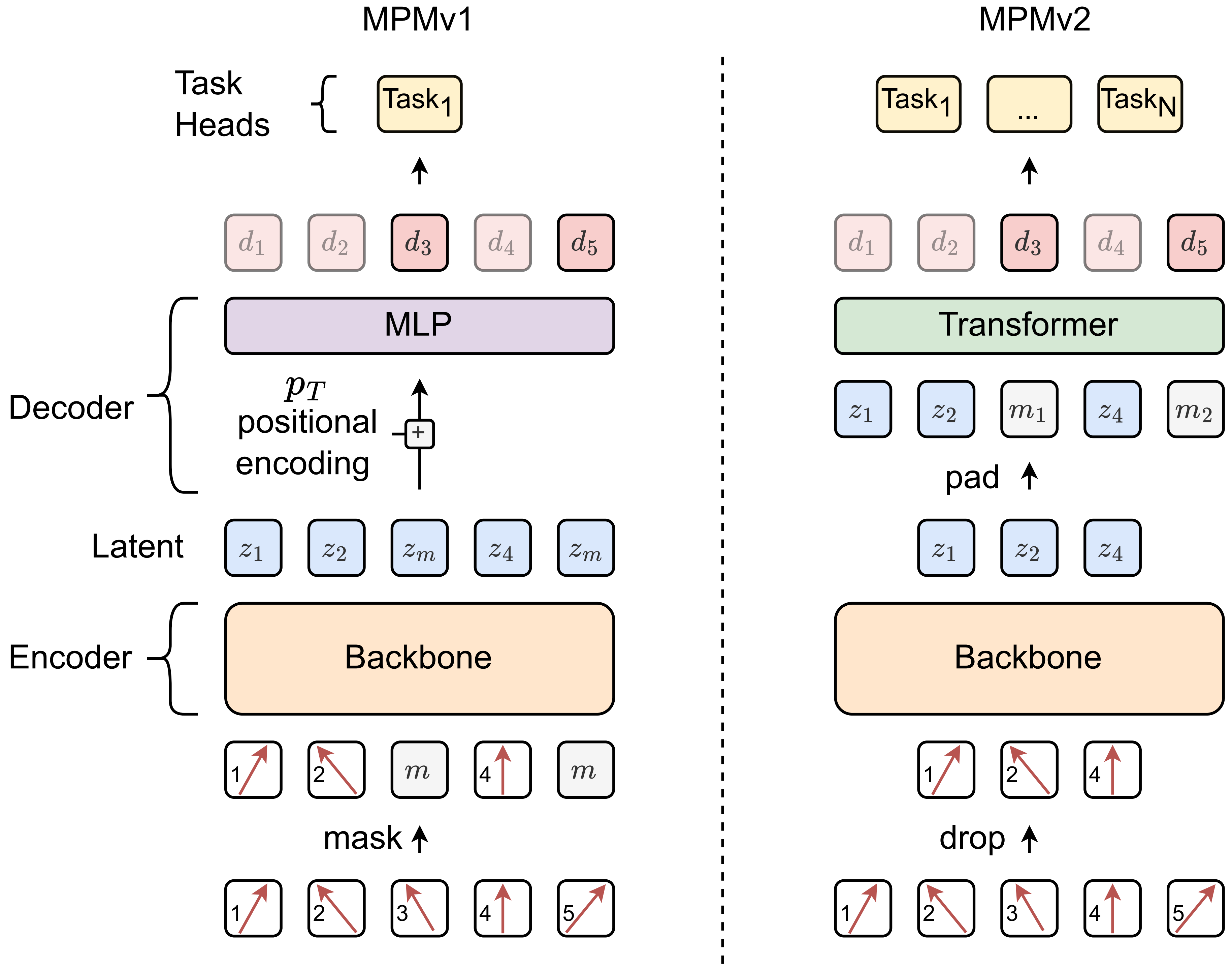
- The CV approach: **“masked modeling”**
  - i.e., mask and reconstruct

Masked Autoencoder
[arXiv: 2111.06377]



- Adapted for particle physics: Masked Particle Modeling

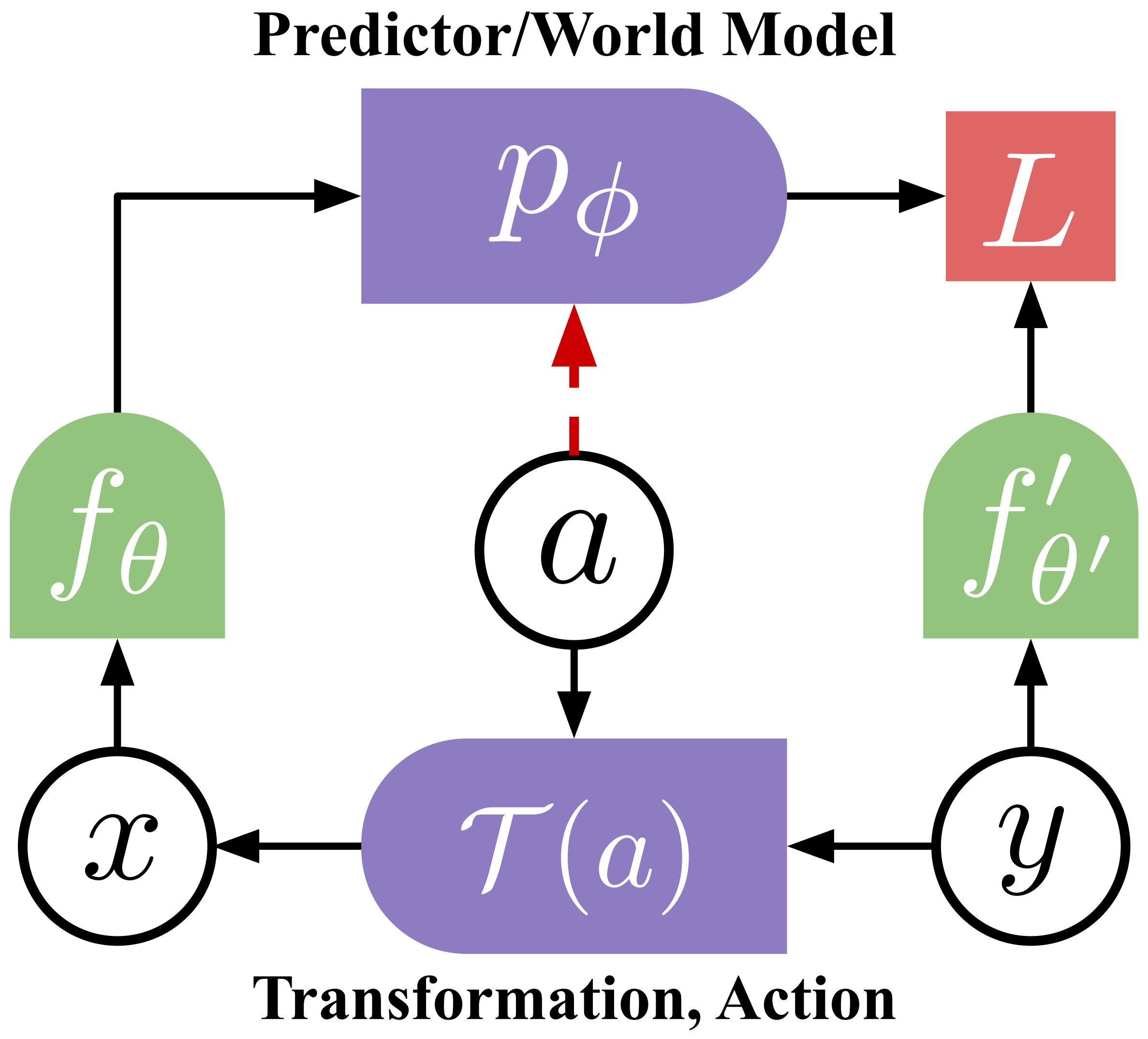
  - MPMv1 [arXiv: 2401.13537]
    - VQ-VAE for particle tokenization
    - predict discrete tokens for masked particles



  - MPMv2 [arXiv: 2409.12589]
    - no need for discrete tokenization
    - multiple reconstruction tasks:
      - PID prediction
      - direct regression
      - conditional generative tasks (via CNF / CFM)

Joint-Embedding Predictive Architecture

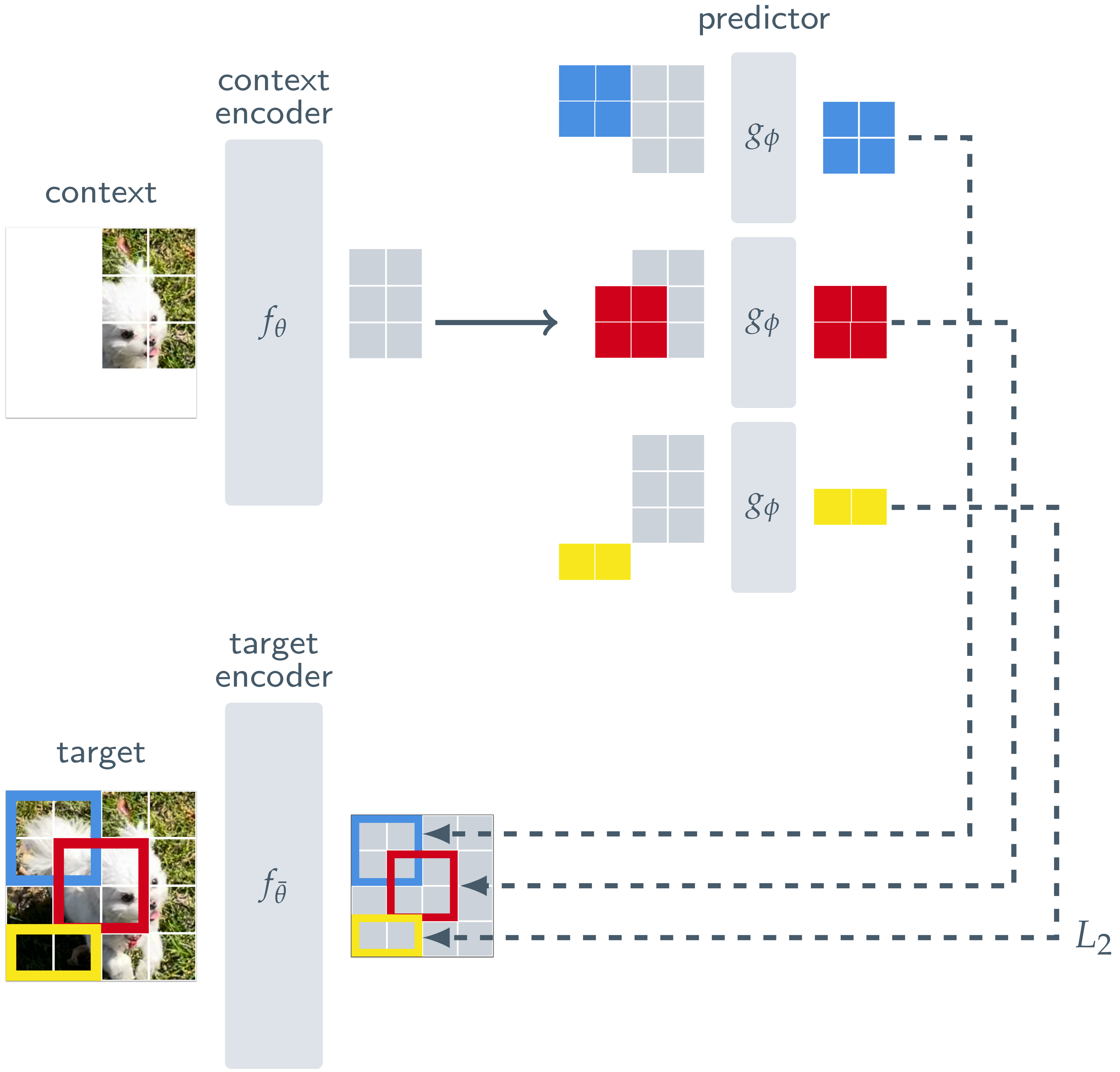
Joint-Embedding
Predictive Architecture (JEPA)



Learns to predict the embeddings in the ***latent*** *space.*
A path towards "World Models".

arXiv: 2403.00504

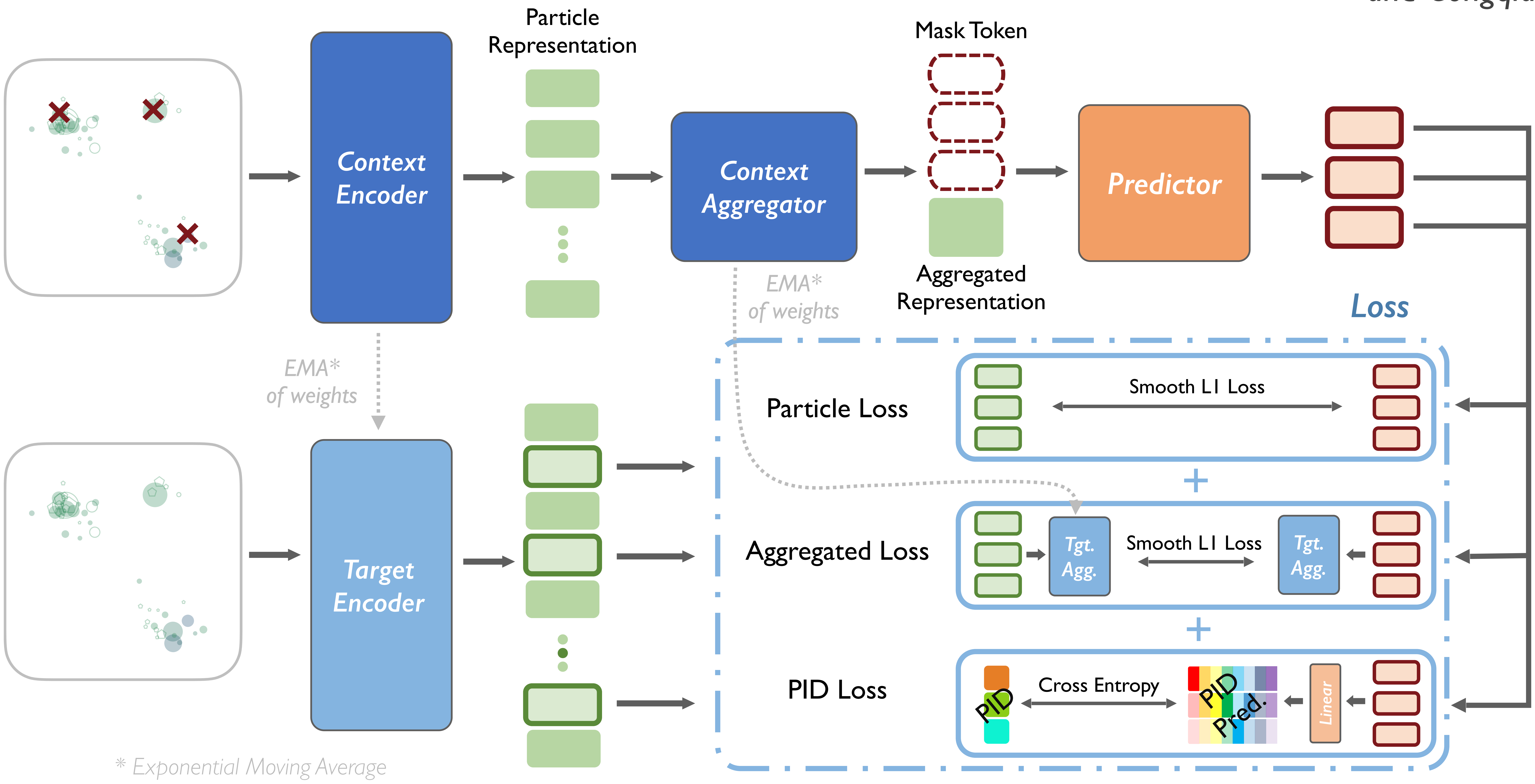
I-JEPA [arXiv: 2301.08243]



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

INTRODUCING P-JEPA

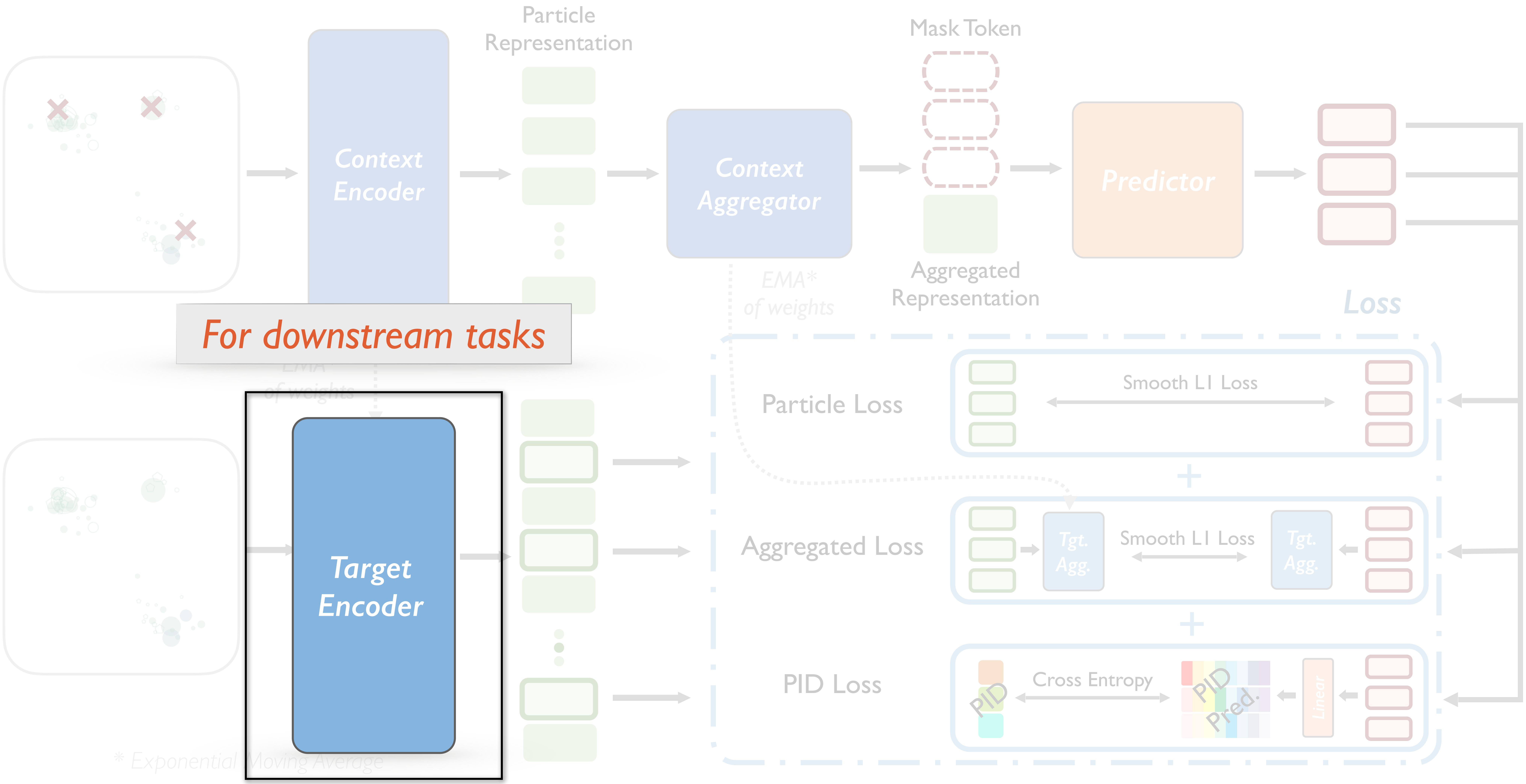
Work in progress with Qibin Liu, Shudong Wang and Congqiao Li



* Exponential Moving Average

See also: "J-JEPA" [S. Katel, H. Li, Z. Zhao, F. Mokhtar, J. Duarte and R. Kansal, arXiv: 2412.05333], "HEP-JEPA" [J. Bardhan, R. Agrawal, A. Tilak, C. Neeraj and S. Mitra, arXiv: 2502.03933]

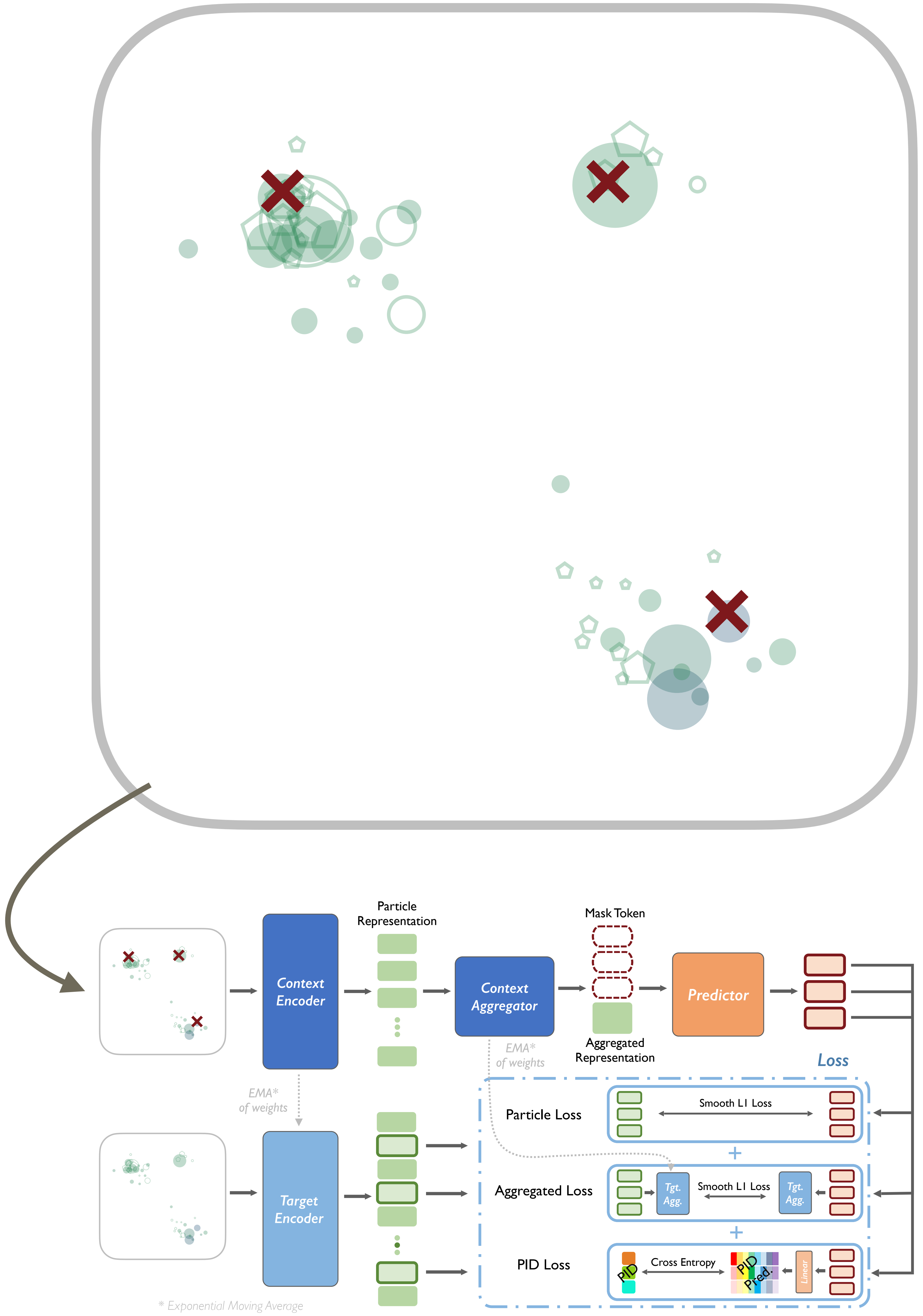
Introducing P-JEPA



Particle Masking

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

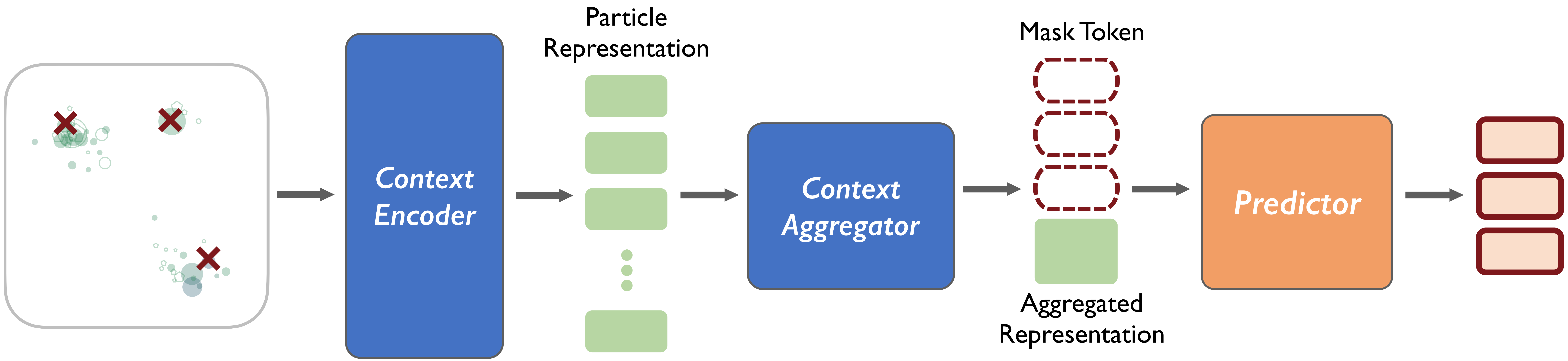
- Masking strategy:
  - randomly mask **30–50%** of the particles in a jet
  - the remaining particles serve as the context for the prediction
    - ==> input to the context encoder & predictor
  - the masked particles become the **target** to be predicted
    - ==> NOT seen by the context encoder &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



Context Encoder and Predictor

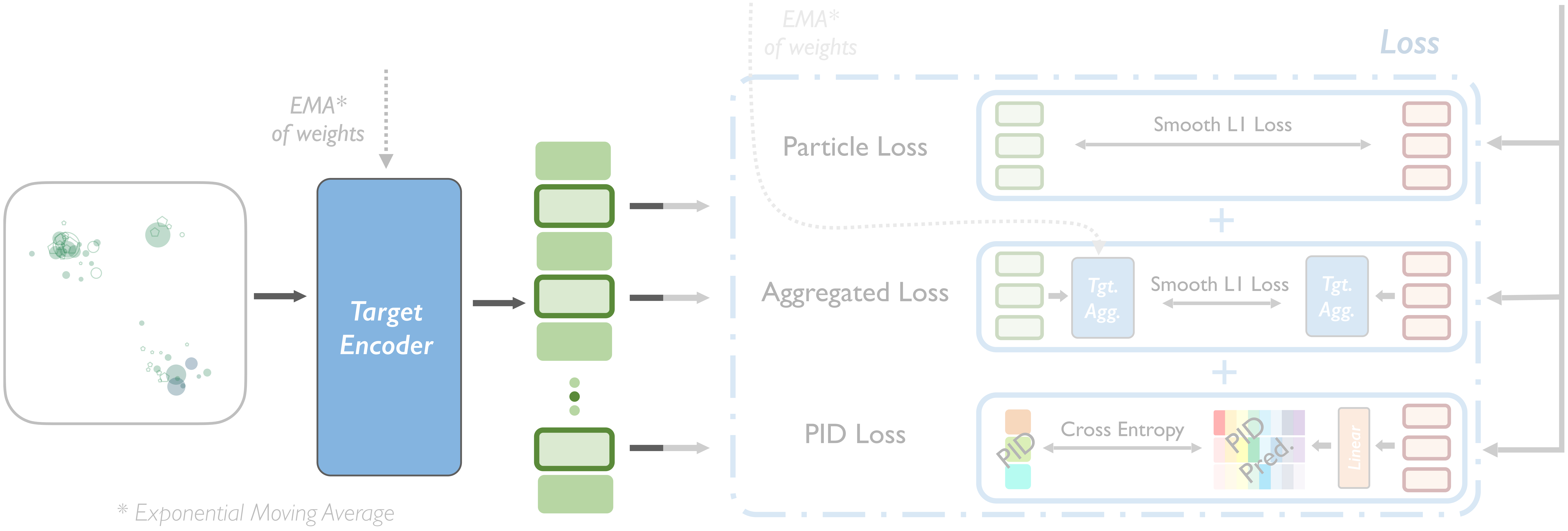
- Context encoder
  - a larger Particle Transformer (w/ pairwise features)
- Context aggregator
  - aggregates all context particles into a single token
- Predictor
  - plain Transformer, smaller than encoder
  - predicts the masked particles from the aggregated representation + mask tokens w/ pos. emb.

	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	/
Num Params	76M	2.6M



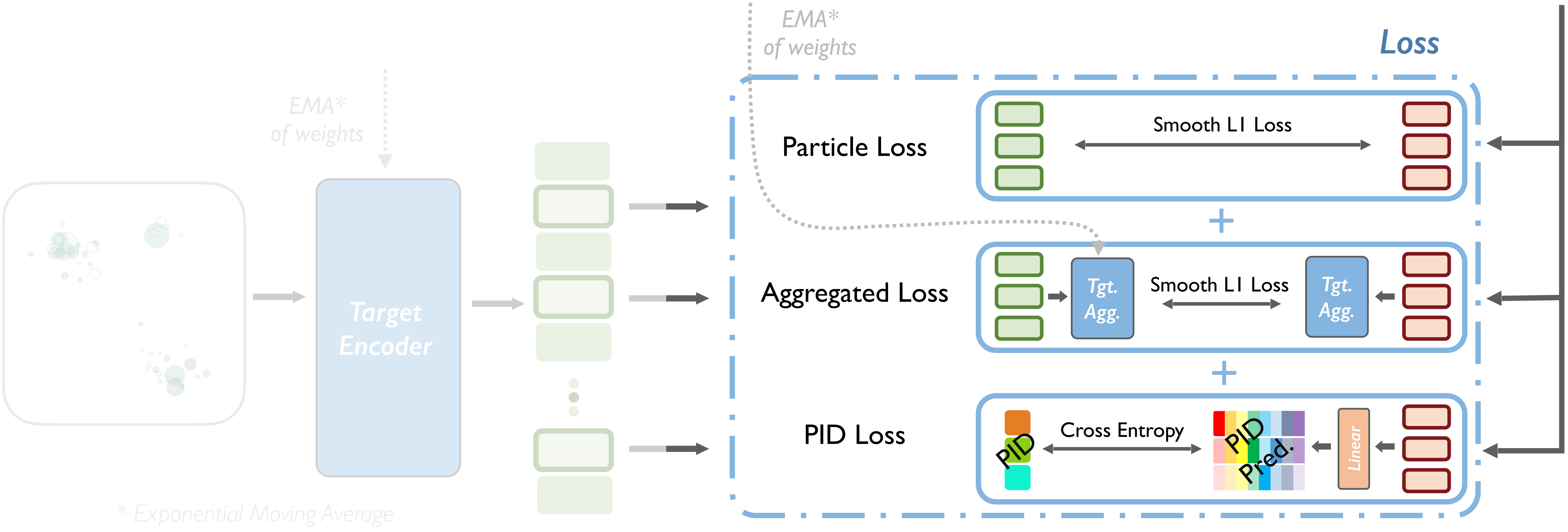
Target Encoder

- 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 up for loss computation
  - updated by "copying" the weights from the context encoder (via exponential moving average)



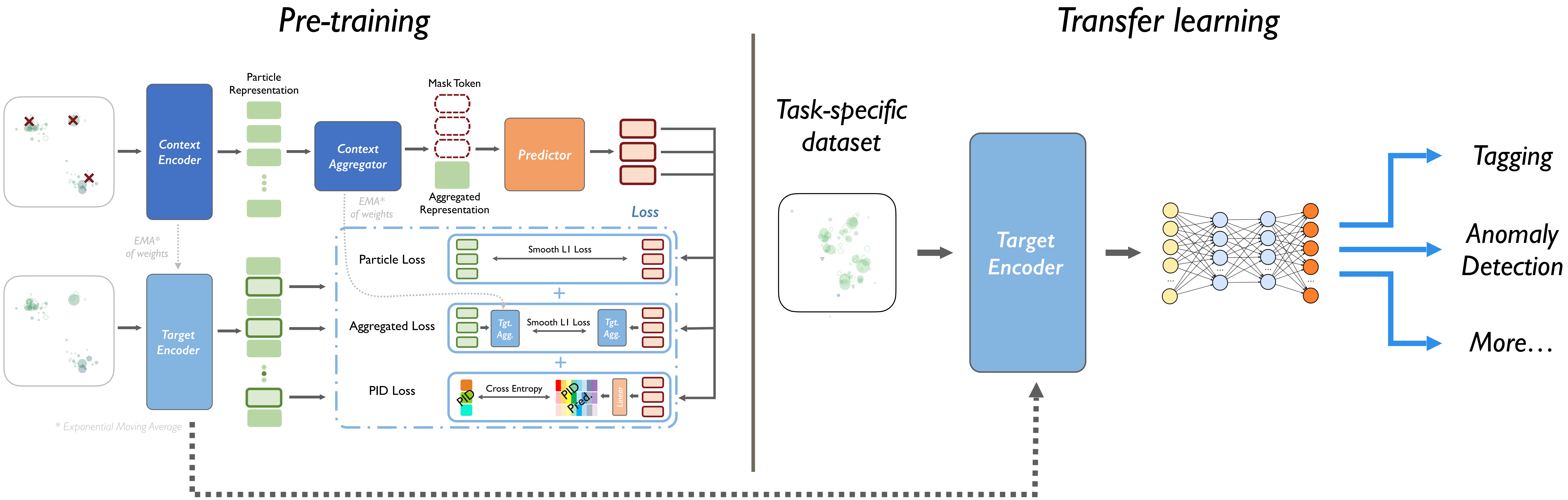
Pre-training Loss

- 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



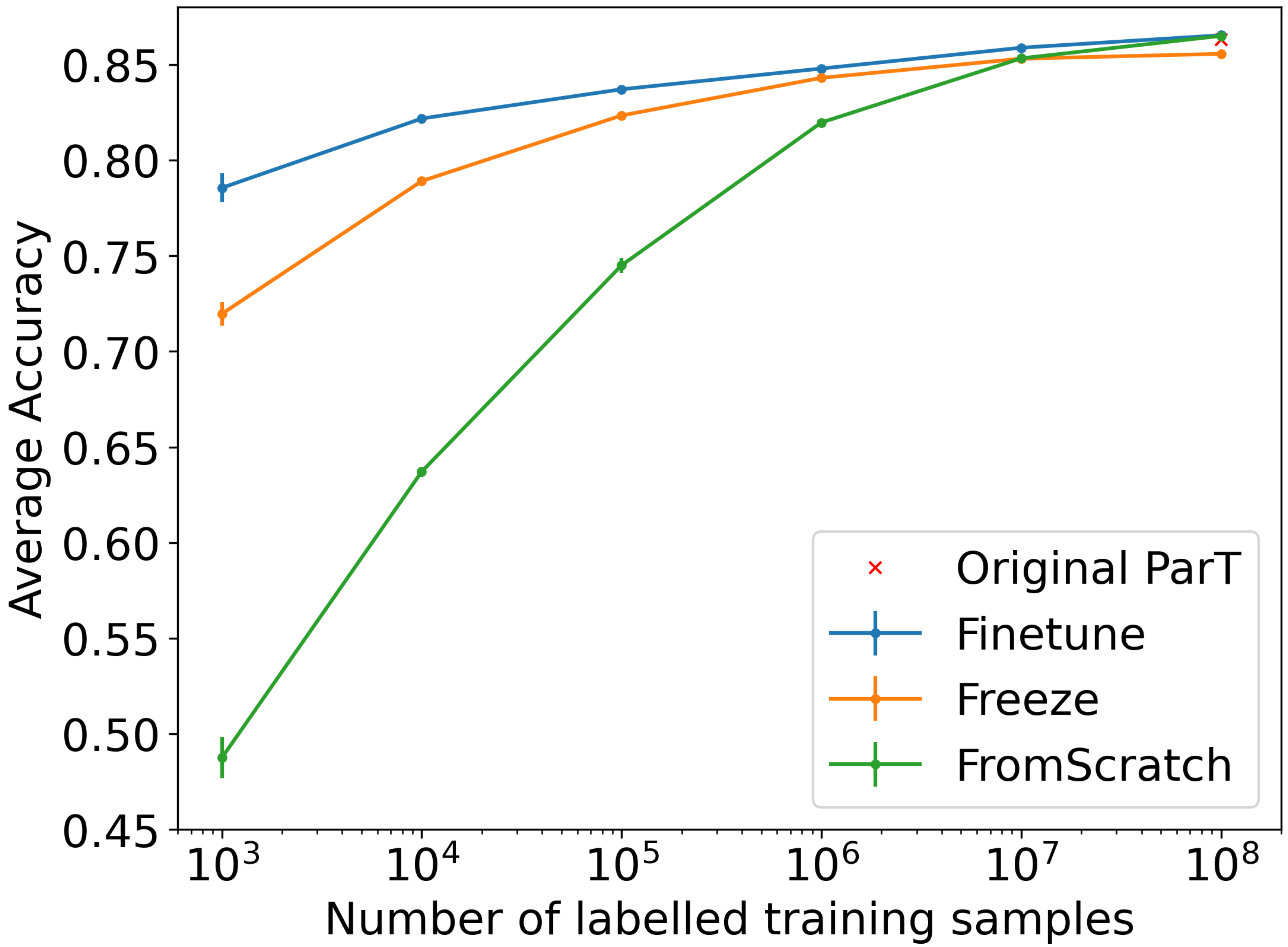
PRE-TRAINING AND TRANSFER LEARNING

- 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:

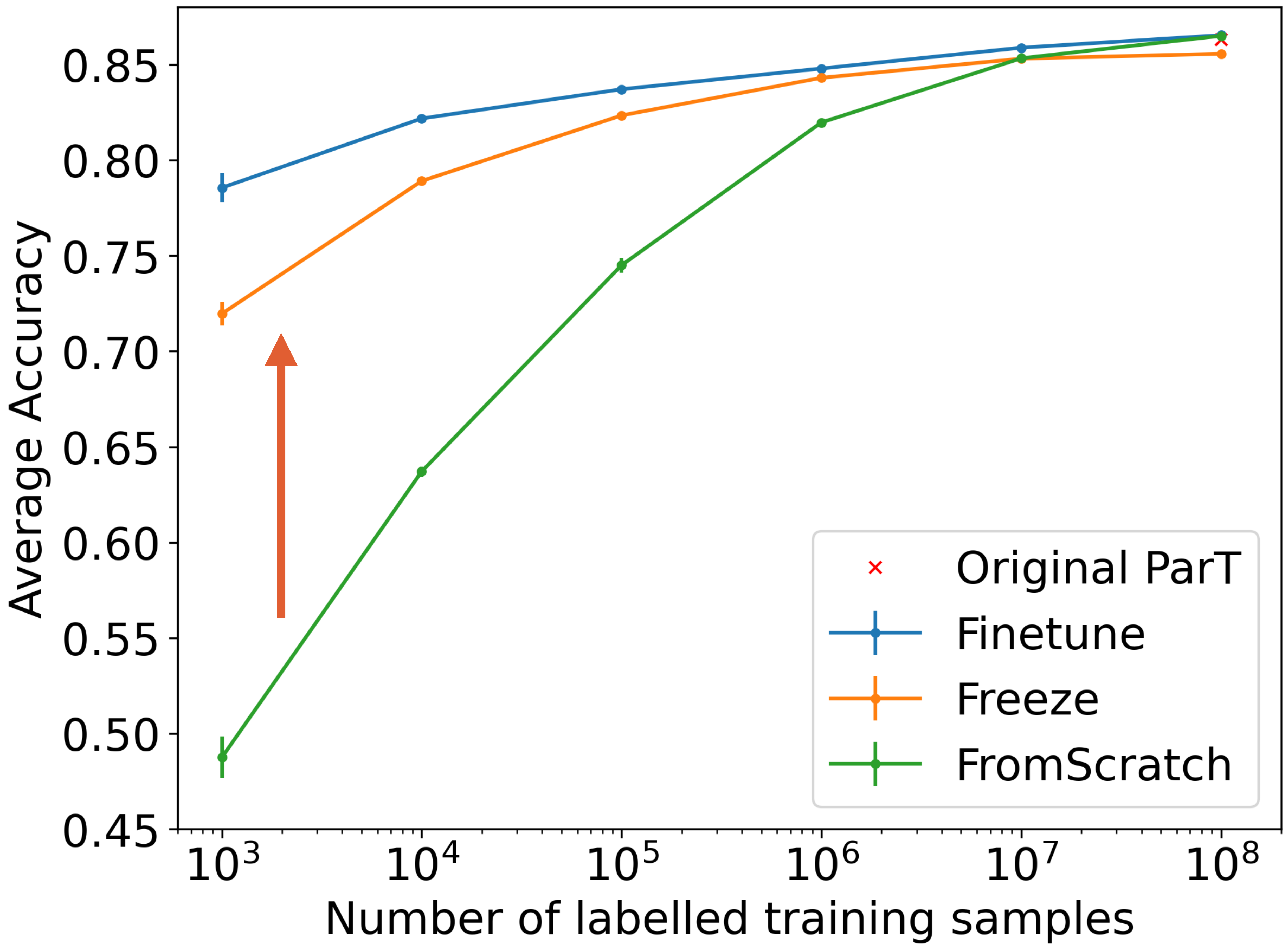
Encoder fixed when trained with labelled jets for tagging

FromScratch:

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

Transfer Learning: Jet Tagging

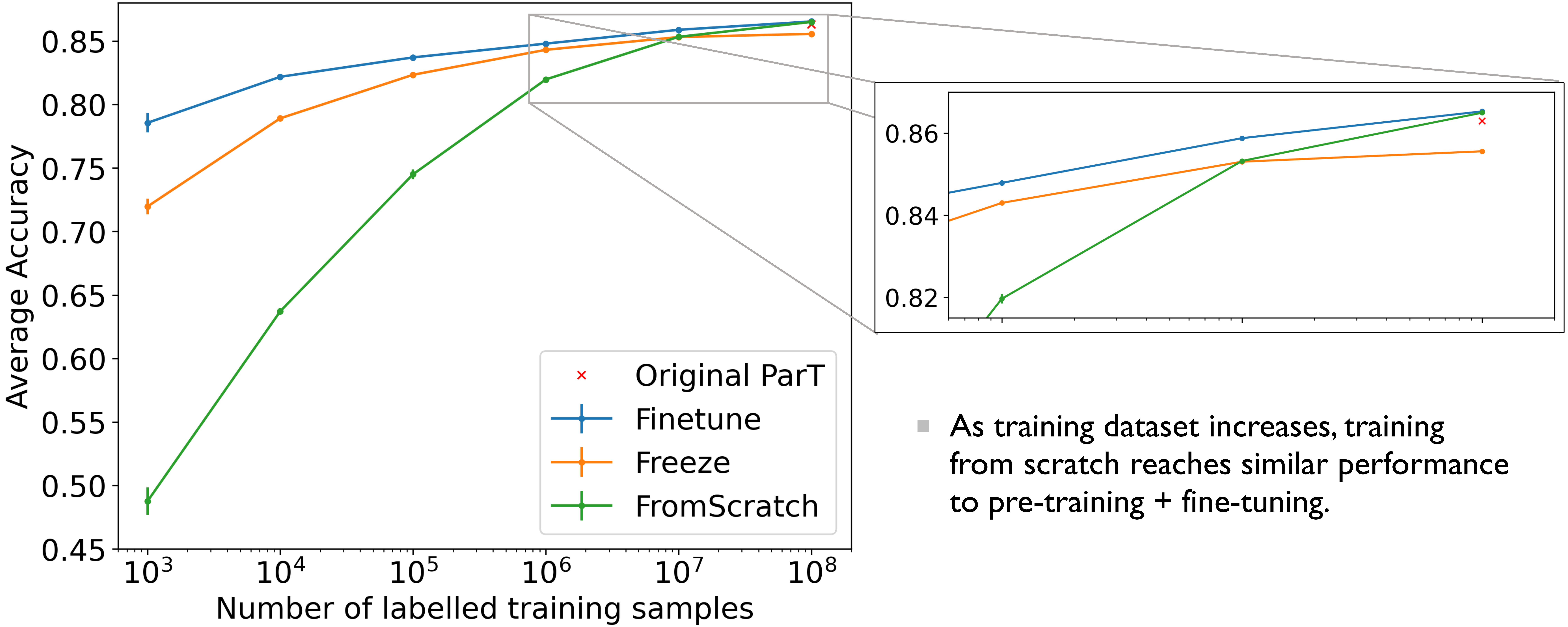
- Benchmark: 10-class jet classification on JetClass



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

TRANSFER LEARNING: JET TAGGING

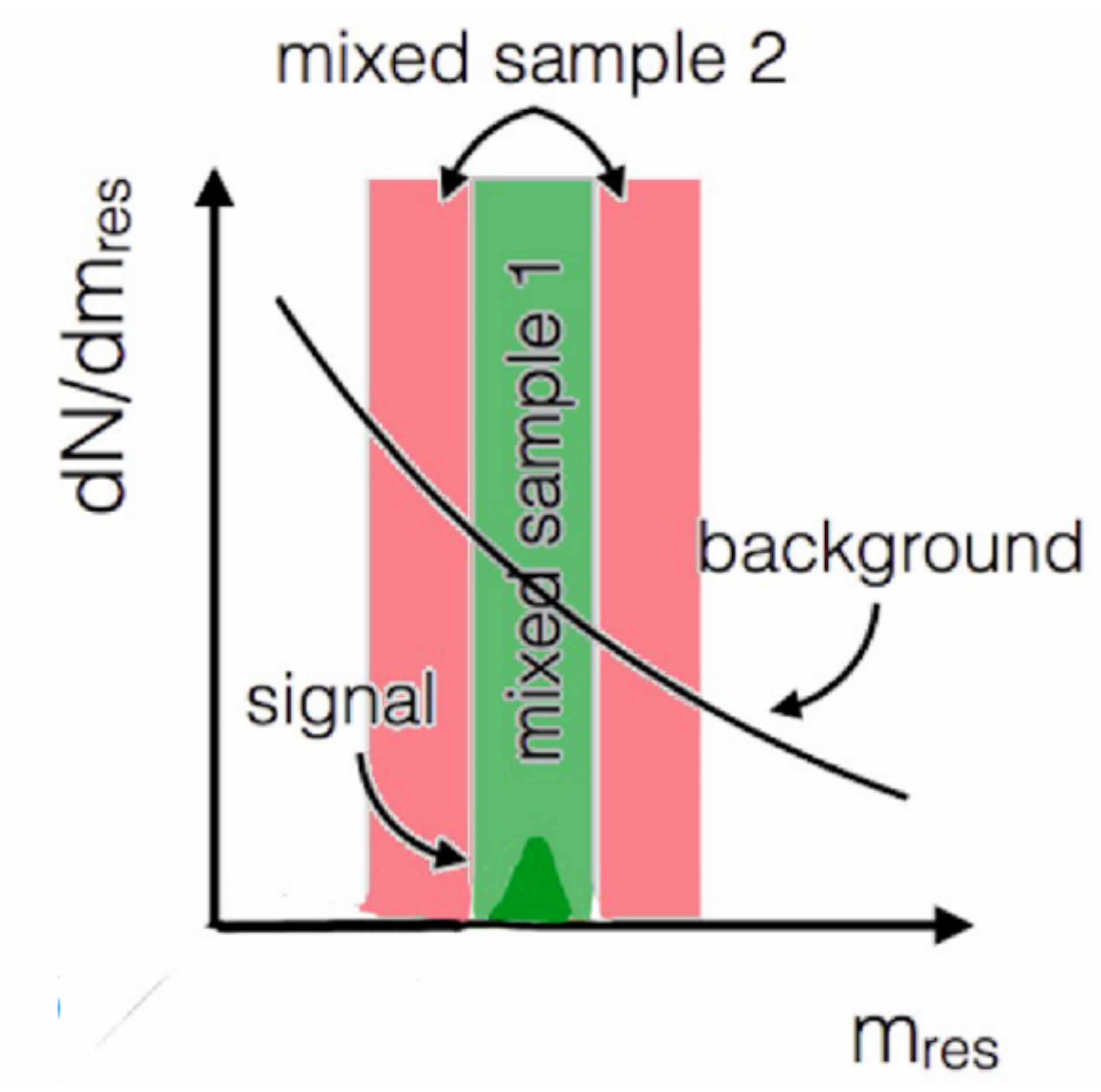
- Benchmark: 10-class jet classification on JetClass



- As training dataset increases, training from scratch 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 is effectively a signal vs background discriminator, thus can be used to enhance signal purity
  - allows to detect unknown signals purely from data



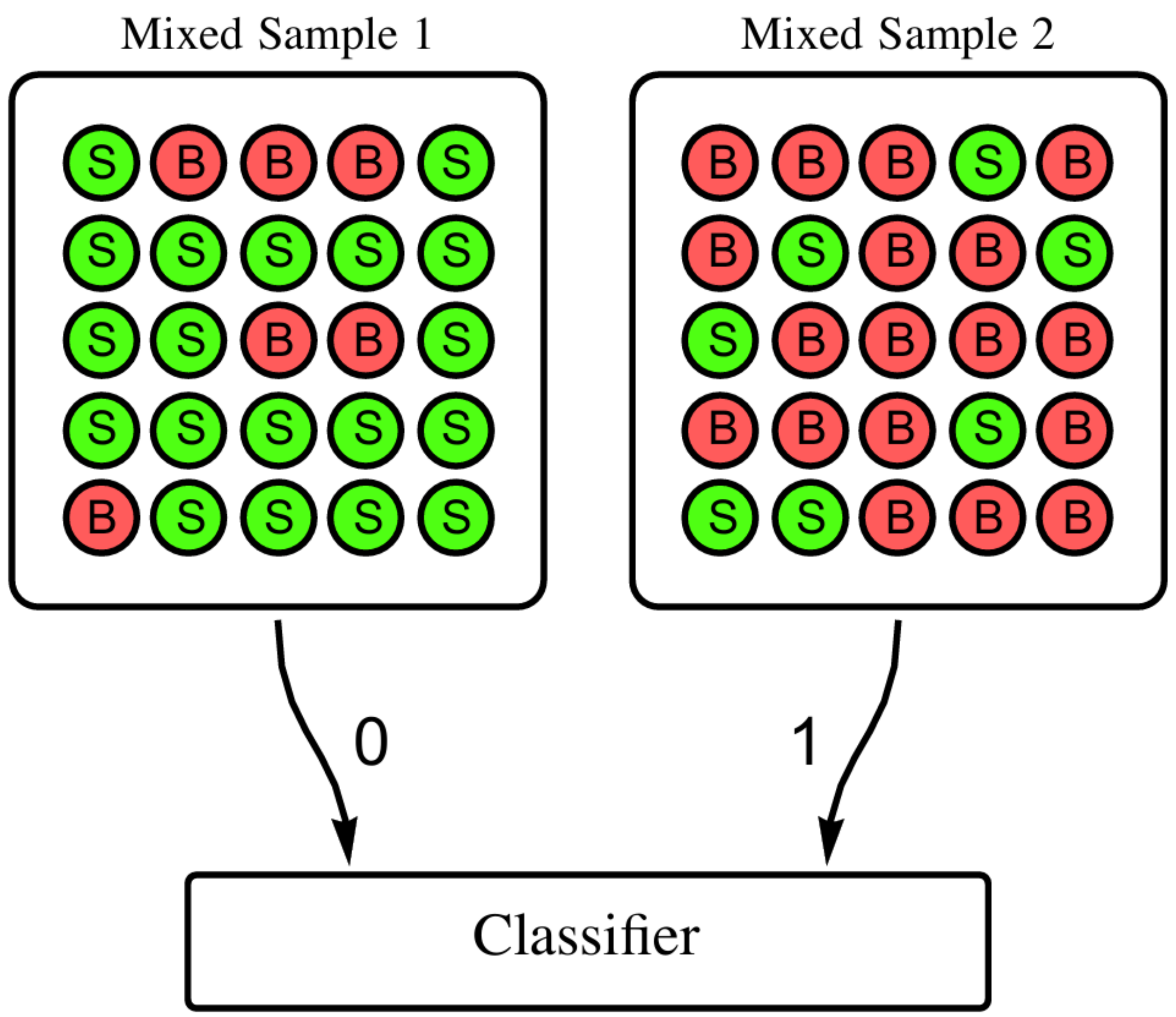
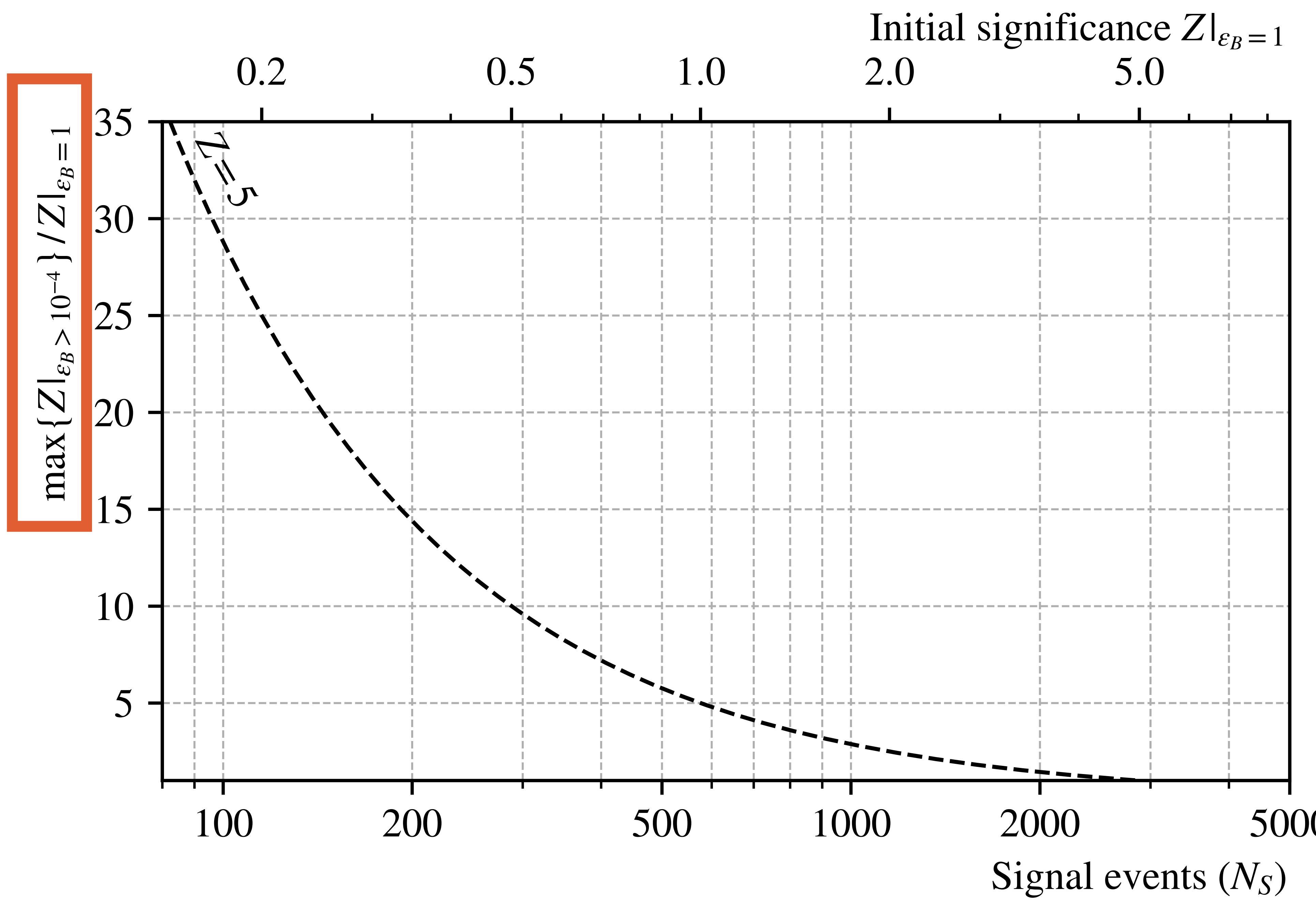


Figure Credit

Transfer Learning: Anomaly Detection

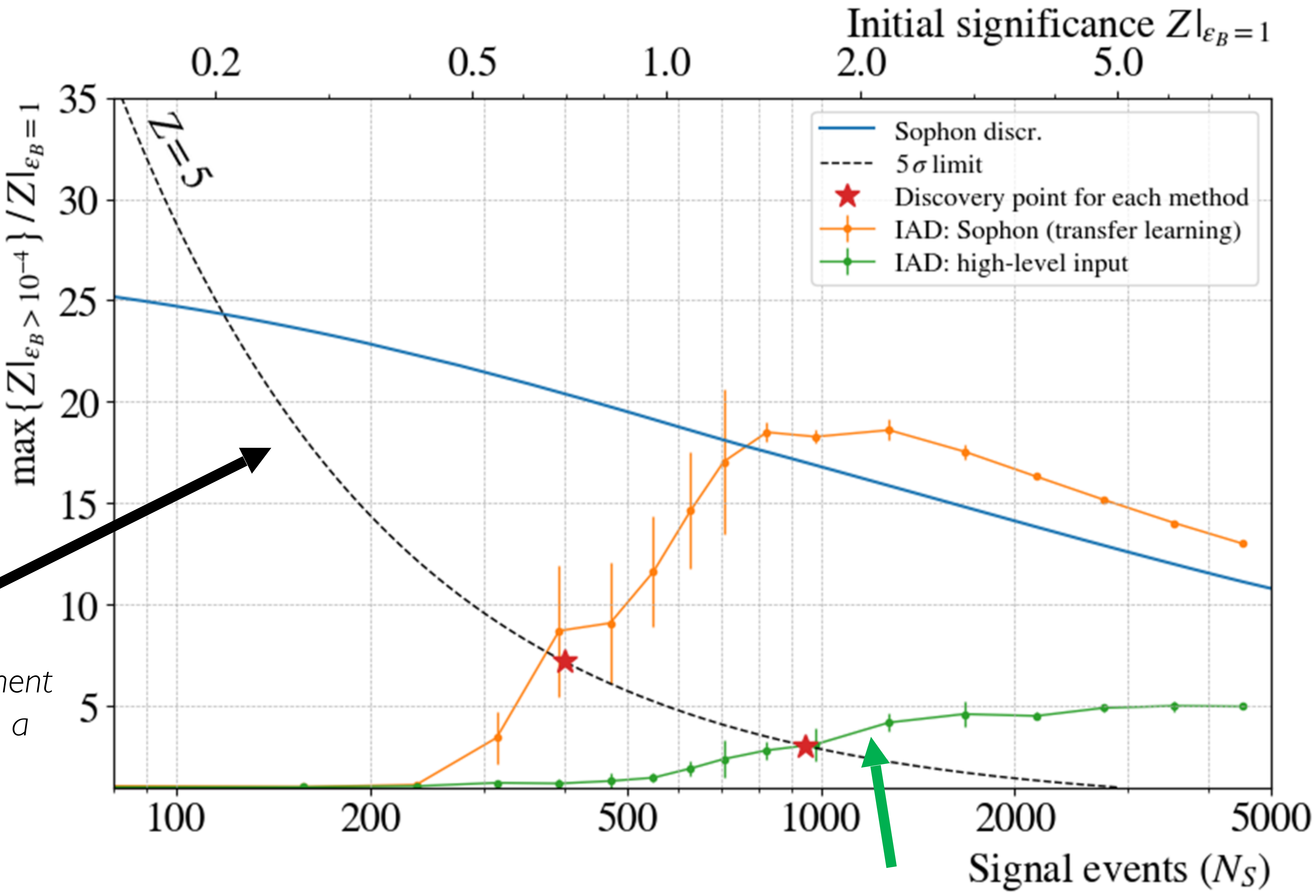
- Traditionally AD was performed using only high-level features (e.g., jet mass, substructures) as inputs
- Machine-learned representations captures richer information of a jet, thus can improve the performance of AD
  - see e.g., the "Sophon" approach [arXiv: 2405.12972]
- We benchmark the P-JEPA extracted features 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 evaluated by the **significance improvement** metric
    - i.e., **ratio** of the *maximal* significance (at an optimal classifier cut) over the *initial* significance (i.e., inclusive w/o classifier cut)



TRANSFER LEARNING: ANOMALY DETECTION

Significance improvement

(better)



How much improvement is required to achieve a 5σ discovery?

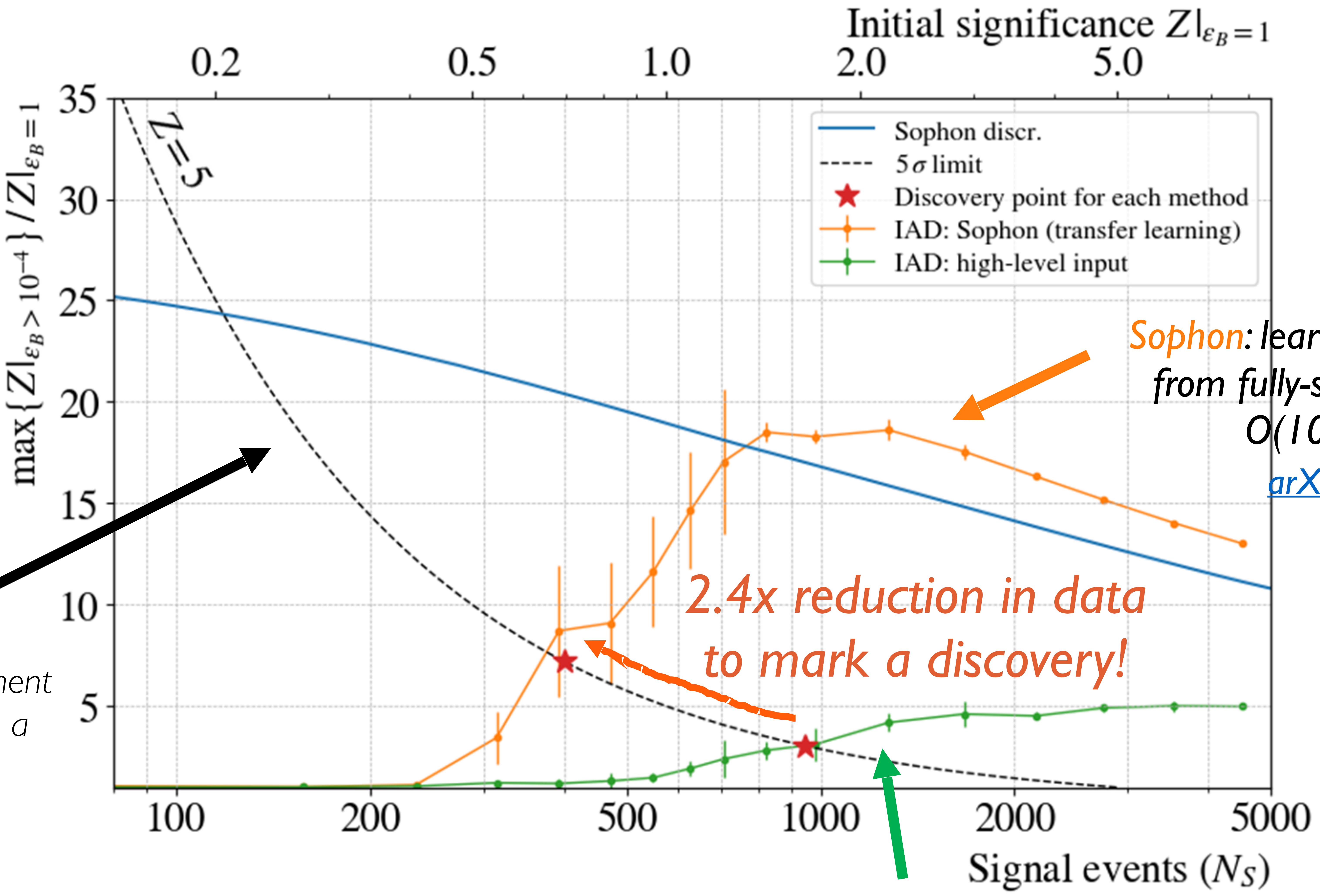
Classifier using high-level observables

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

arXiv: 2405.12972

2.4x reduction in data to mark a discovery!

How much improvement is required to achieve a 5σ discovery?

Classifier using high-level observables

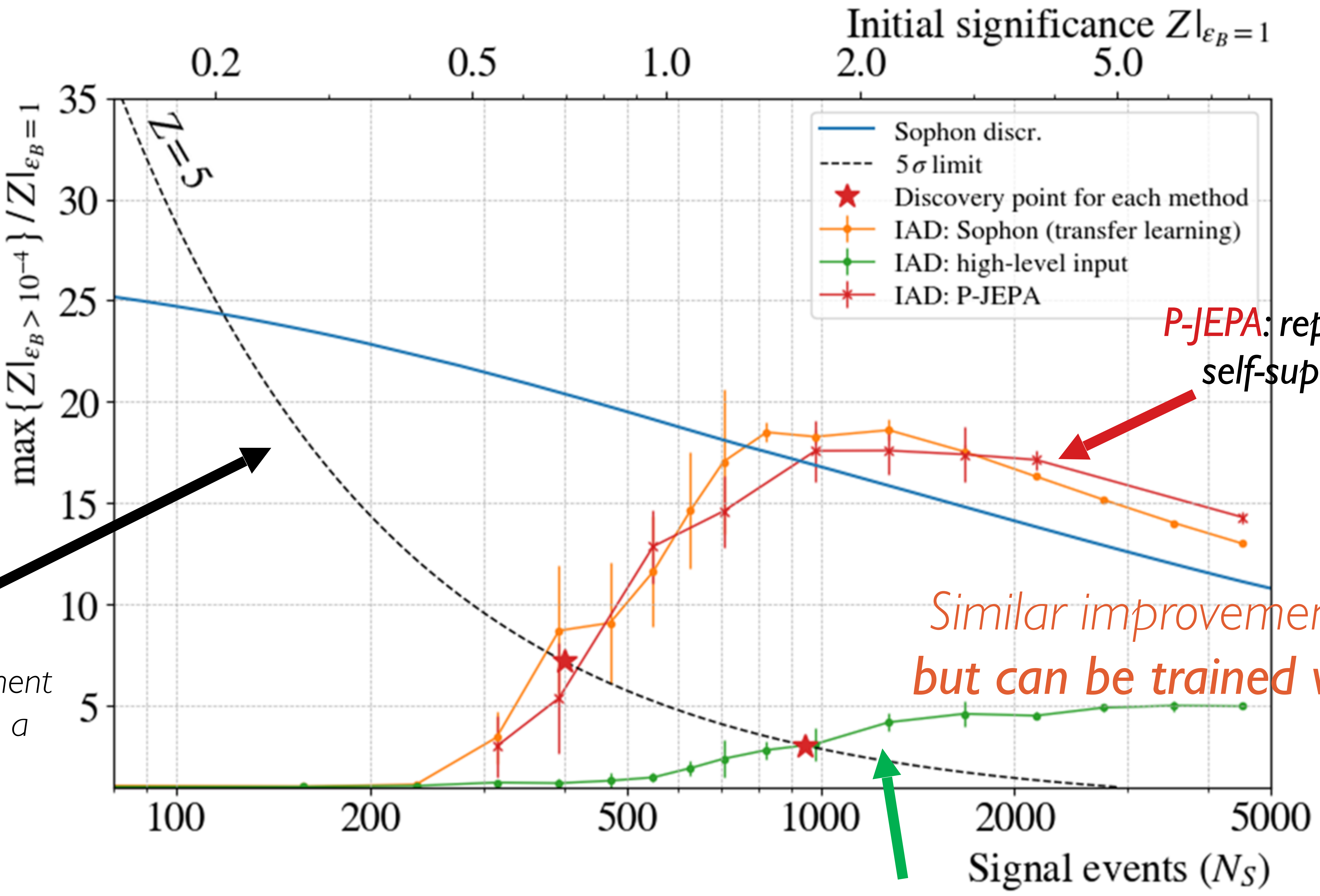
PRL, 121 (2018) 24, 241803

Towards a Foundation Model for Jet Physics - Jlu. 10, 2025 - Huilin Qu (CERN)

Transfer Learning: Anomaly Detection

Significance improvement

(better)



How much improvement is required to achieve a 5σ discovery?

P-JEPA: representations from self-supervised learning

Similar improvements to Sophon, **but can be trained with only data.**

Classifier using high-level observables

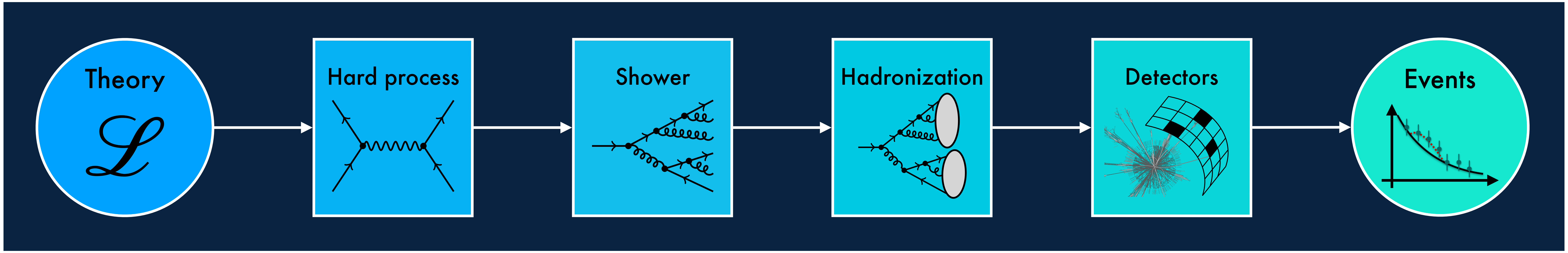
[PRL, 121 (2018) 24, 241803](PRL, 121 (2018) 24, 241803)

SUMMARY & OUTLOOK

- Tremendous progress in machine learning for jet physics in recent years
- P-JEPA: Towards a foundation model for jet physics
  - pre-training via self-supervised learning on unlabelled dataset
  - powerful learned representations for tagging, anomaly detection, and more
- Outlook: a foundation model for all of the LHC?

Generation, Simulation, ...

Forward



Inverse

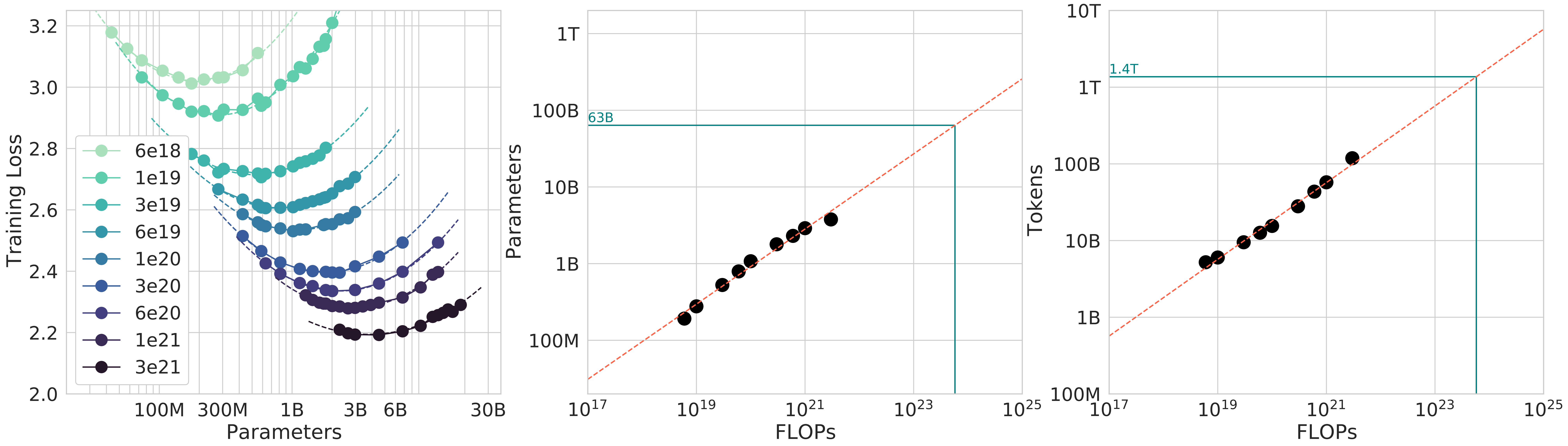
Reconstruction, Unfolding, ...

Credits: R. Winterhalder

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…