Masked particle modelling

Foundation models for HEP

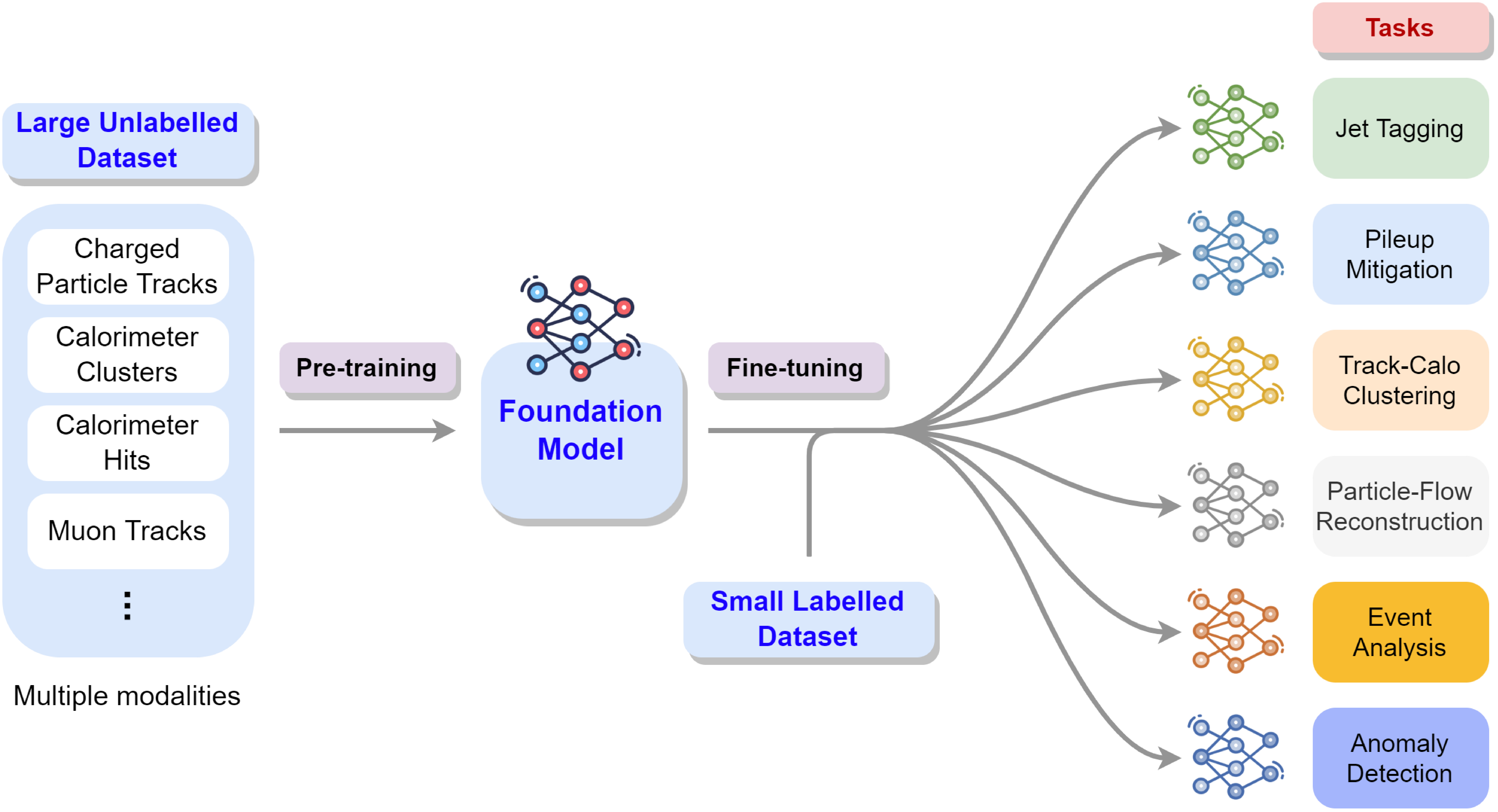






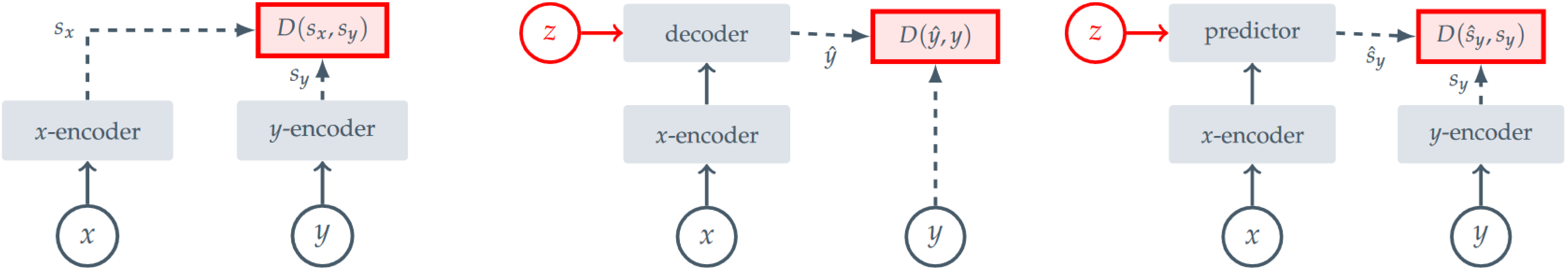
S. Klein **<u>M. Leigh</u>** J. Raine L. Heinrich M. Kagan R. Osadchy T. Golling

Why Foundation Models?



Self-Supervised Learning

Popular Methods



(a) **Joint-Embedding Architecture** (b) **Generative Architecture** (c) **Joint-Embedding Predictive Architecture**

[2301.08243]

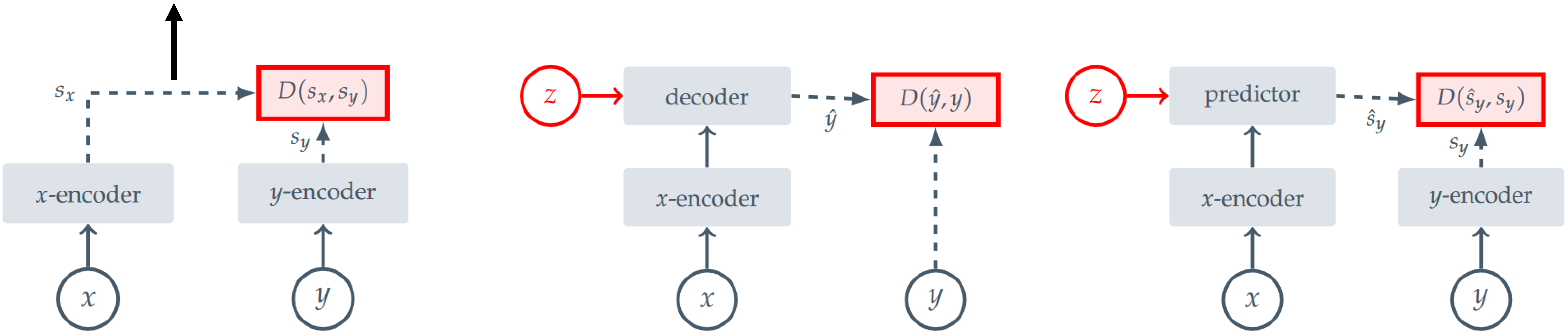
Self-Supervised Learning

Popular Methods

JetCLR - Heidelburg/Hamburg

RS3L - MIT/KIT/SLAC

Detector Replicas - NYU/Weizman



(a) **Joint-Embedding Architecture** (b) **Generative Architecture** (c) **Joint-Embedding Predictive Architecture**

2301.08243

Self-Supervised Learning

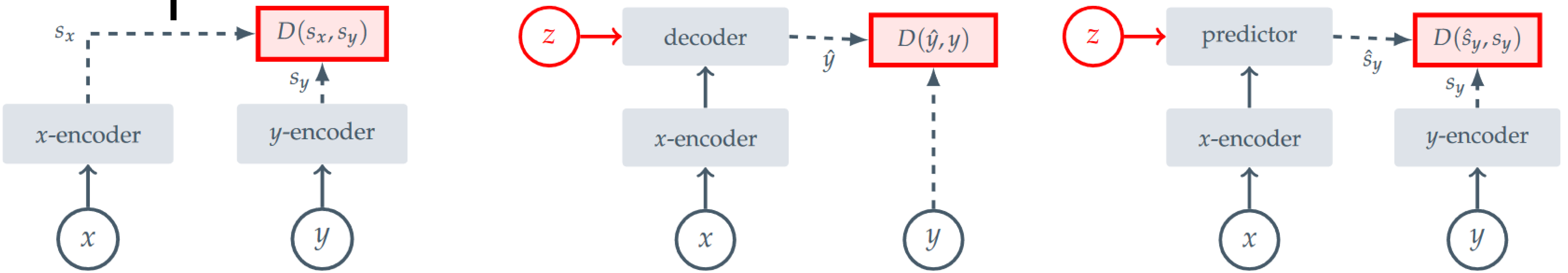
Popular Methods

JetCLR - Heidelburg/Hamburg

RS3L - MIT/KIT/SLAC

Detector Replicas - NYU/Weizman

Ours



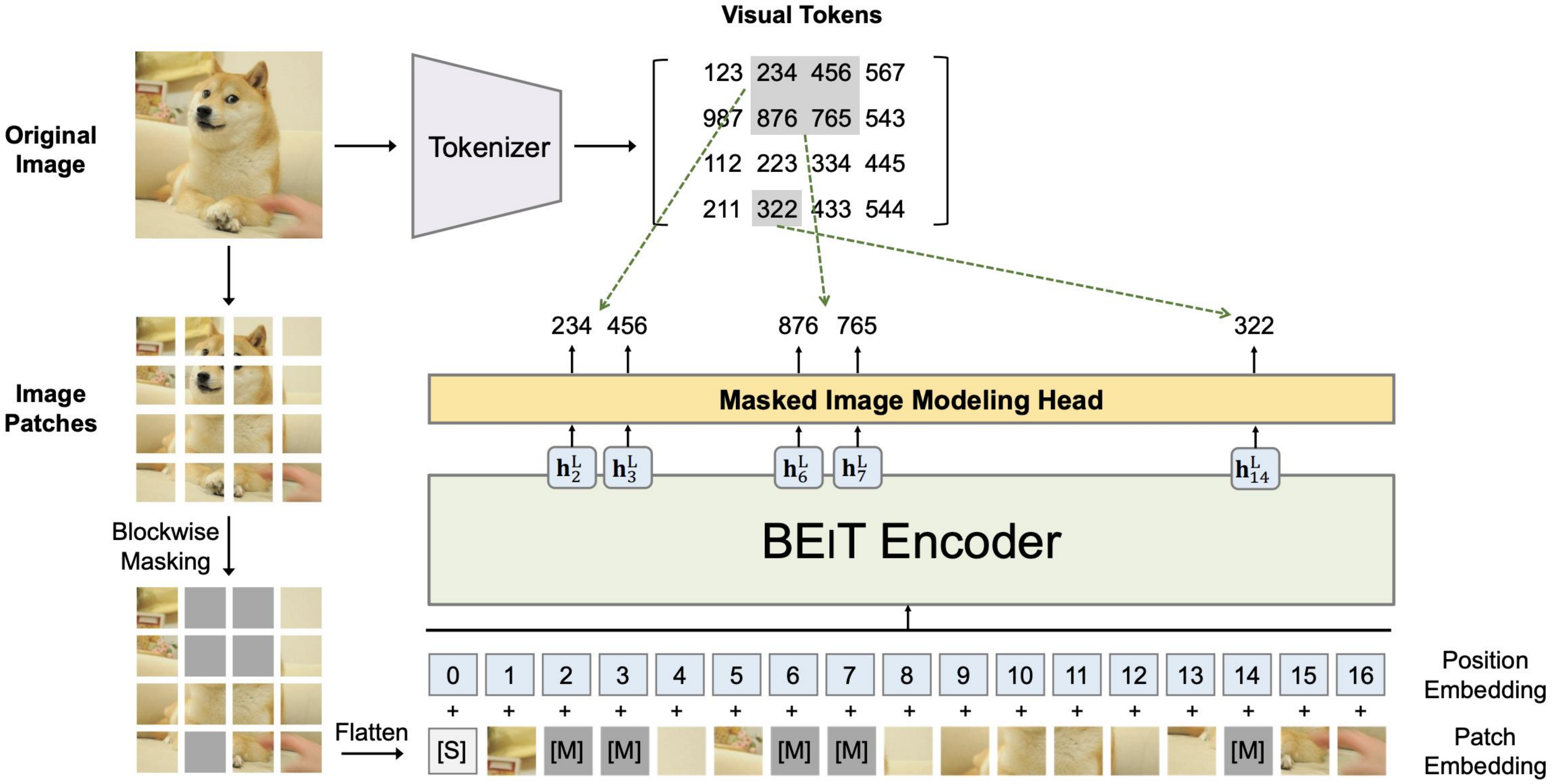
(a) **Joint-Embedding Architecture** (b) **Generative Architecture** (c) **Joint-Embedding Predictive Architecture**

[2301.08243](https://arxiv.org/abs/2301.08243)

Masked modelling

Images and words

- The BERT pretraining strategy has been very successful for NLP
- So has BEiT for images
  - Tokenized targets performed better than direct regression

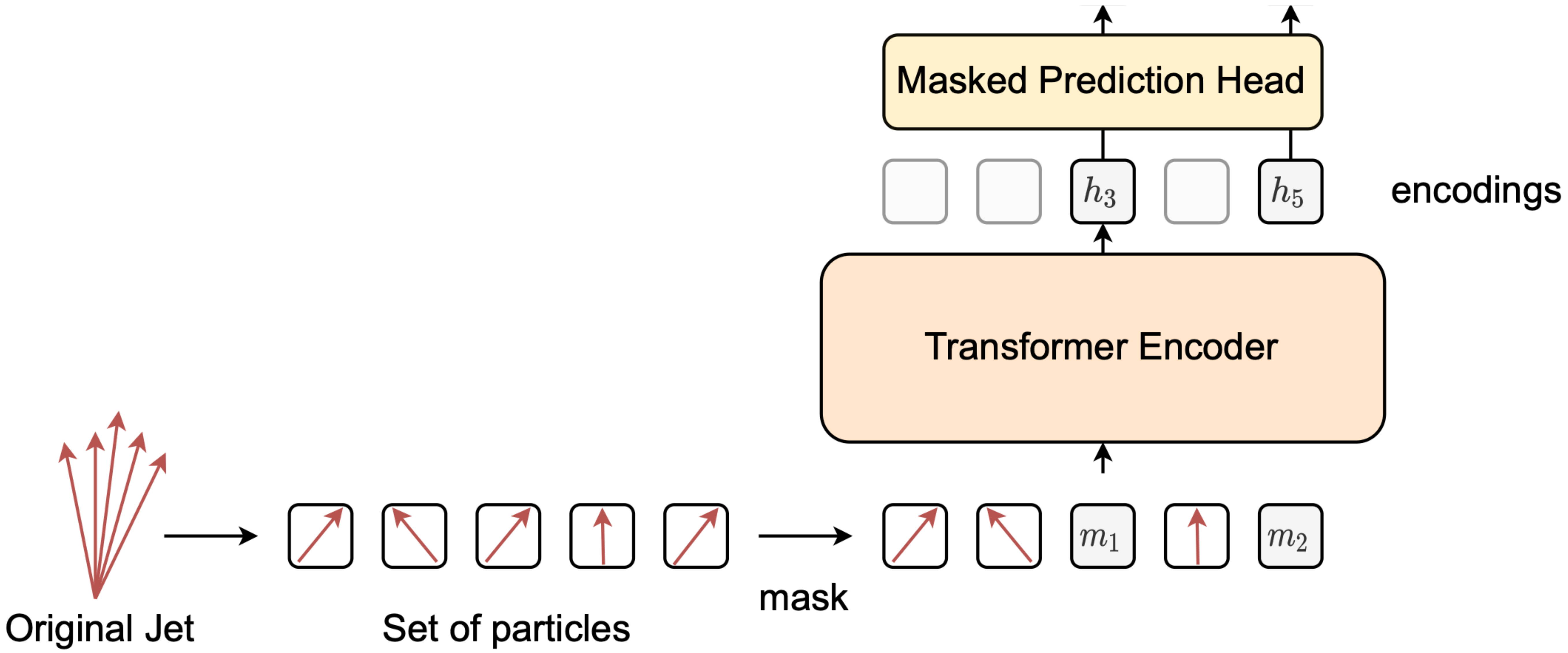


2106.08254

Masked modelling

Does this work for HEP?

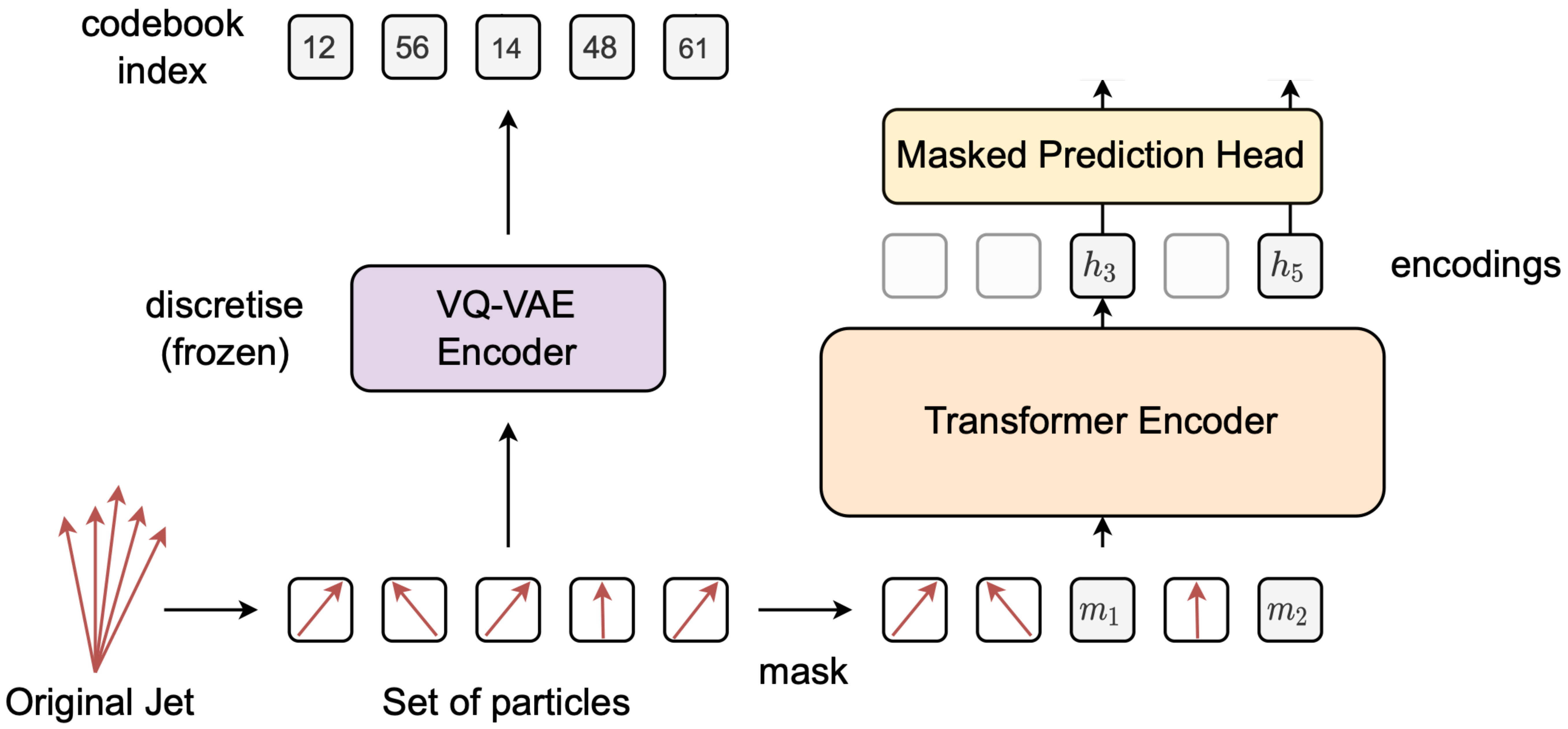
- Like language:
  ‘meaningful’ constituents



Masked modelling

Does this work for HEP?

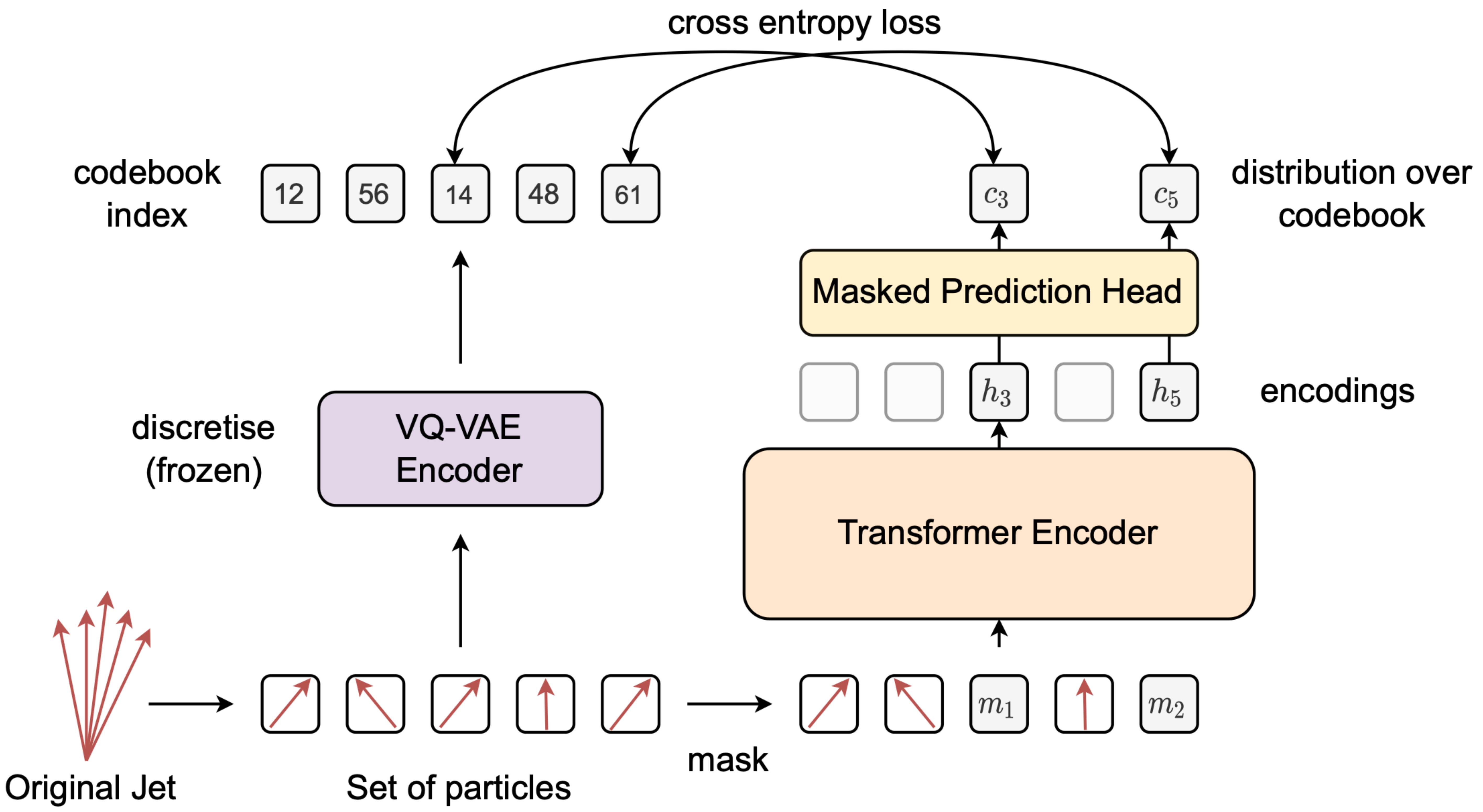
- Like language:
  ‘meaningful’ constituents
- Like images:
  continuous inputs



Masked modelling

Does this work for HEP?

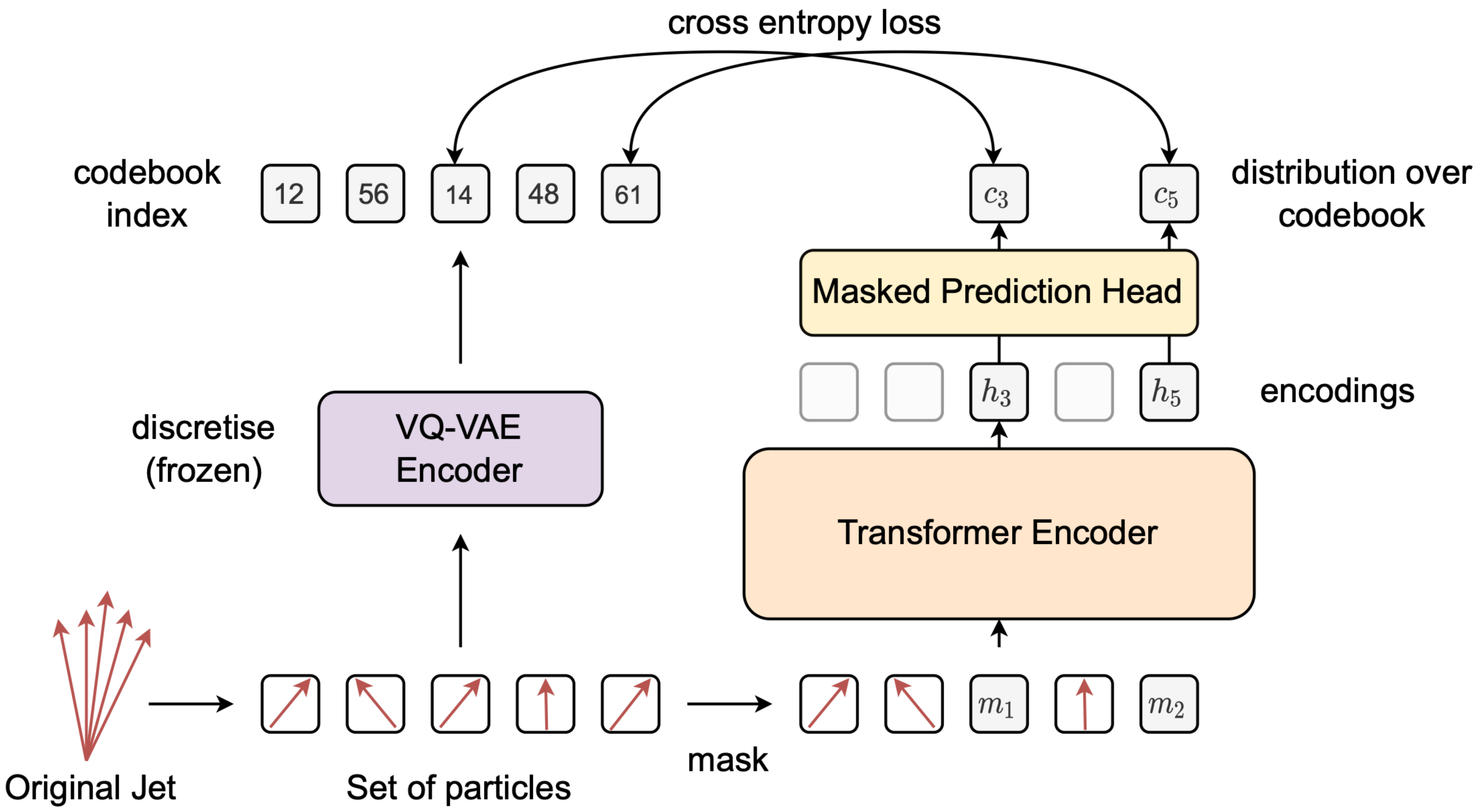
- Like language:
  'meaningful' constituents
- Like images:
  continuous inputs



Masked modelling

Does this work for HEP?

- Like language:
  ‘meaningful’ constituents
- Like images:
  continuous inputs
- Unlike both:
  no positional information



Masked modelling

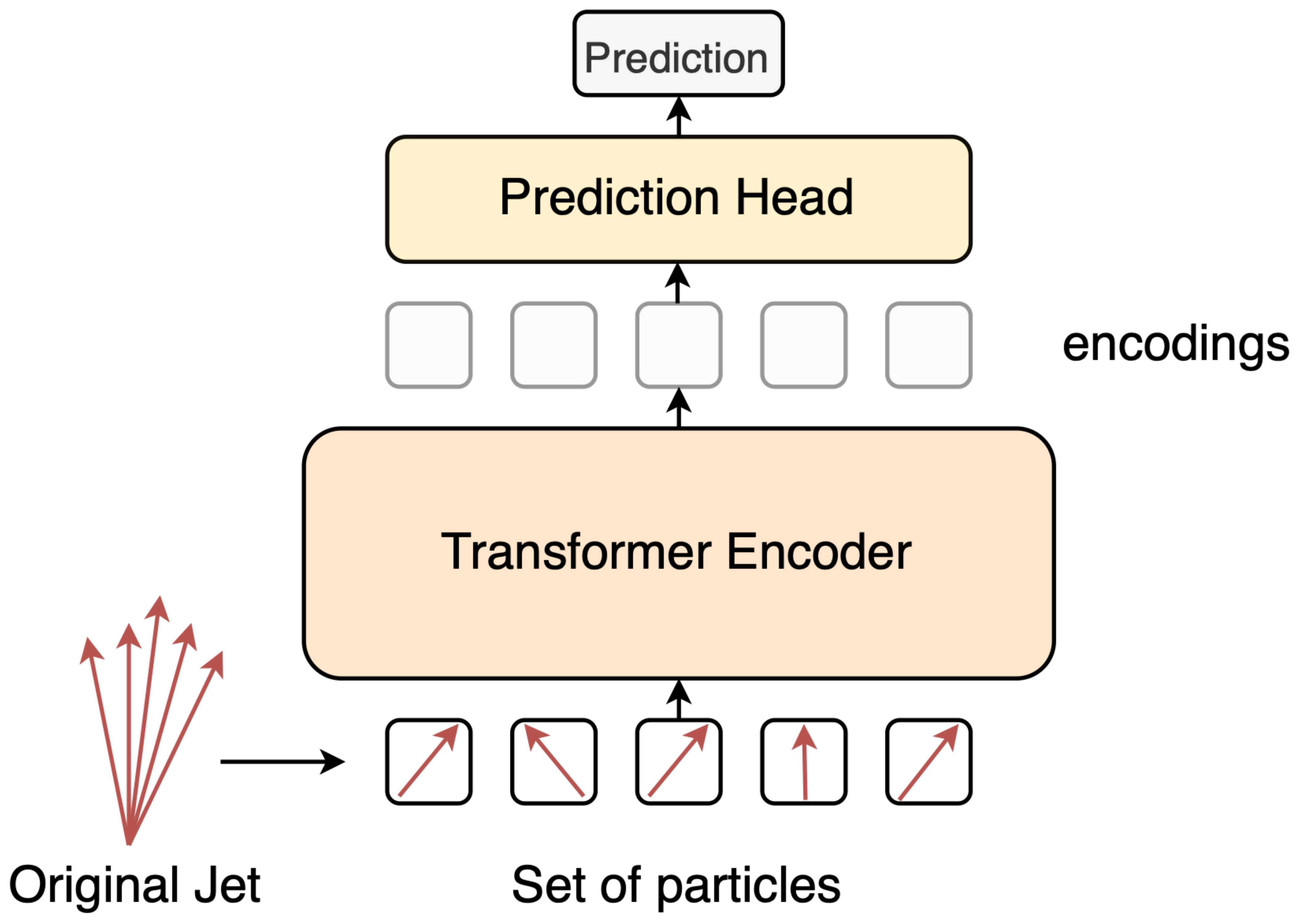
Strengths

- Very simple training objective and data pipeline yet
- Proven to be very effective in NLP and computer vision
- Requires no augmentation / re-simulation
- Can train the backbone **directly on data**
  - Pretraining at unprecedented scale

Masked modelling

Performance

- Pretraining on 100M Jets from [JetClass]
  - 10 classes
- How to quantify the performance of a pretrained model?
  - Array of downstream tasks — fine tuning



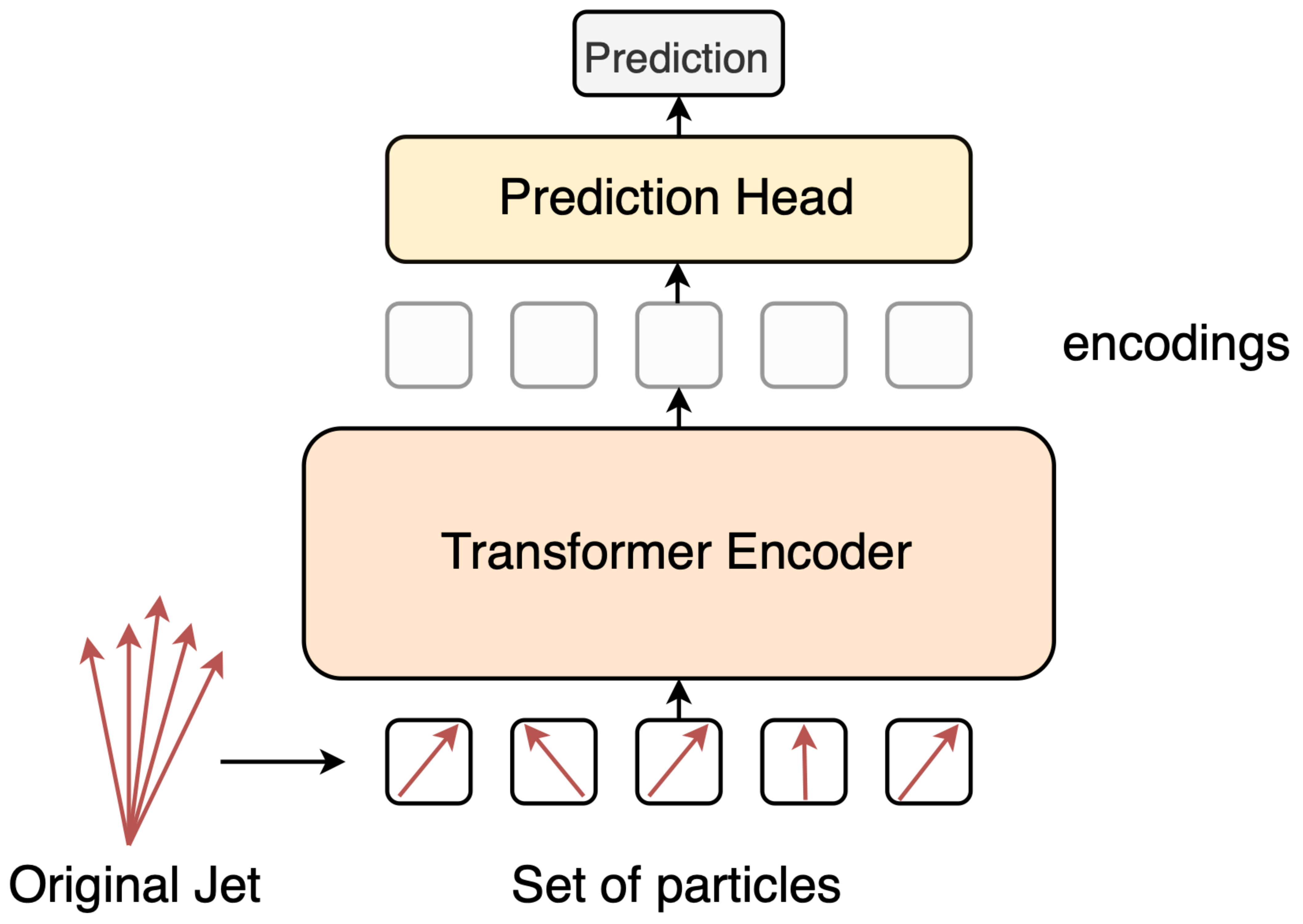
Masked modelling

Downstream training strategies

- Train encoder and head
  Fine-Tuned

- Freeze encoder, only train head
  Fixed

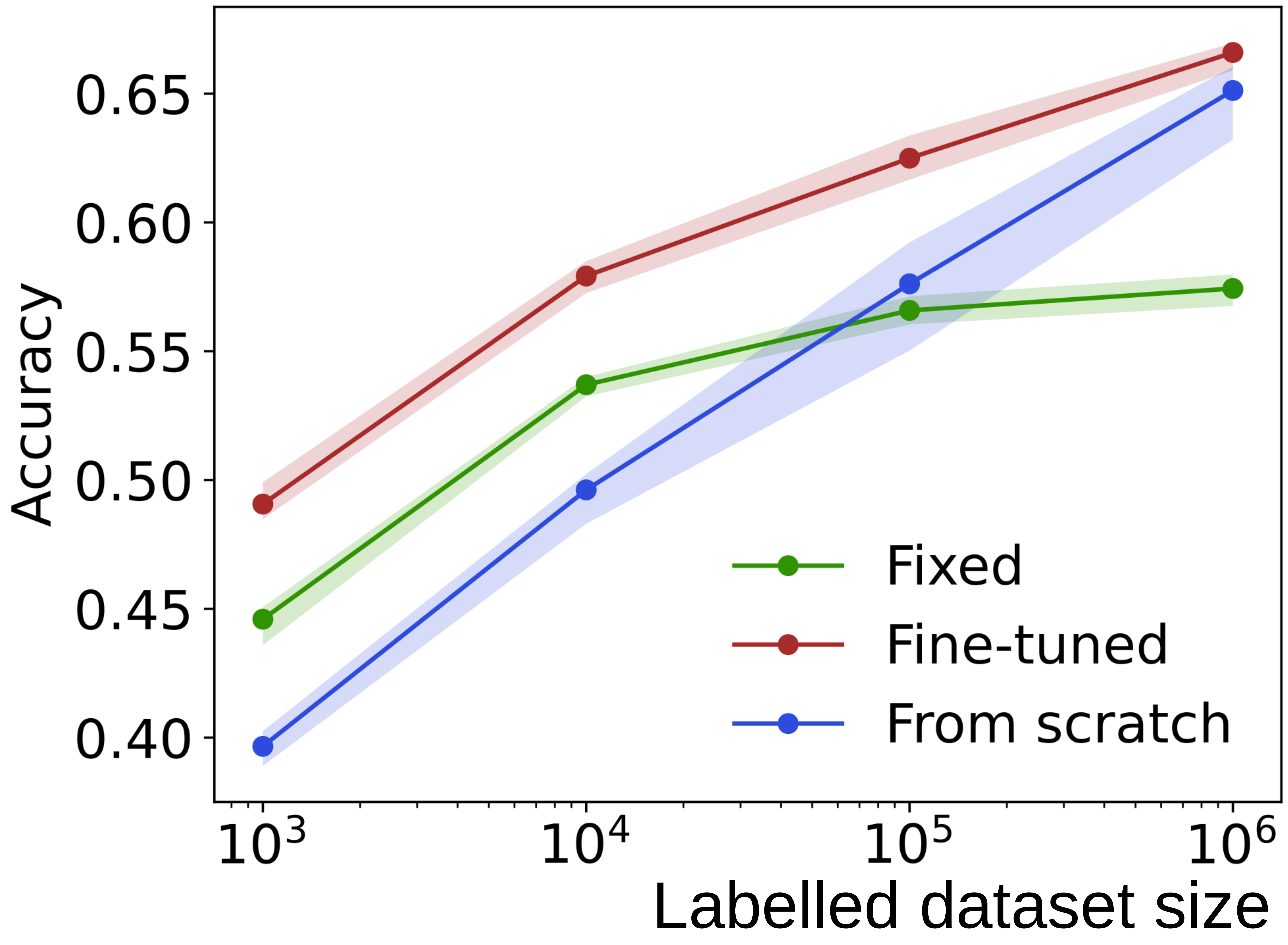
- Reinitialize model, train from scratch
  From scratch



Masked modelling

Fine tune on pretraining set

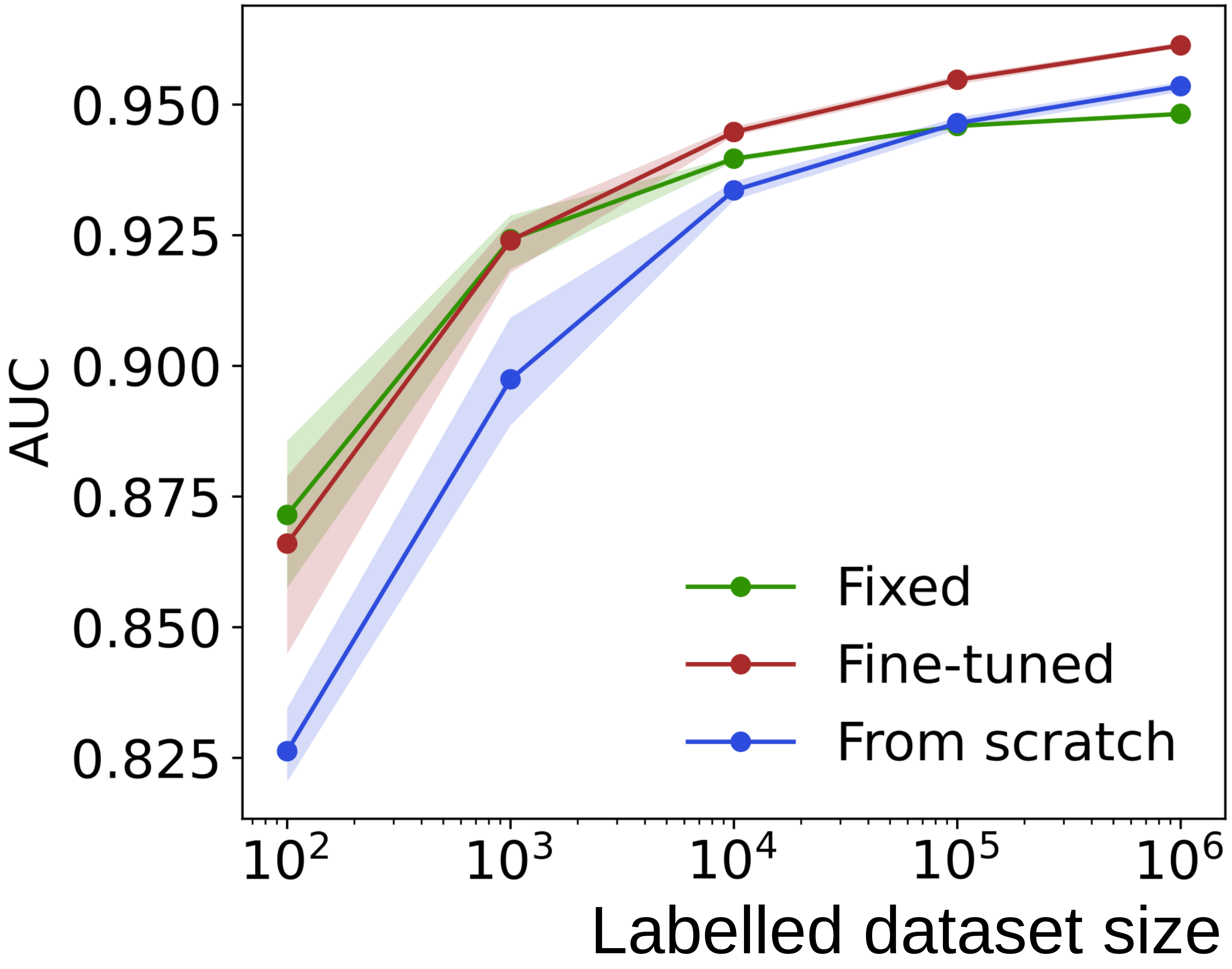
- Select N events and fine tune
- The backbone model outperforms from scratch
    - 10x more data efficient at 60%
- For reference ParT on full 10^8 samples gets around 85%



Masked modelling

Fine tune on new dataset

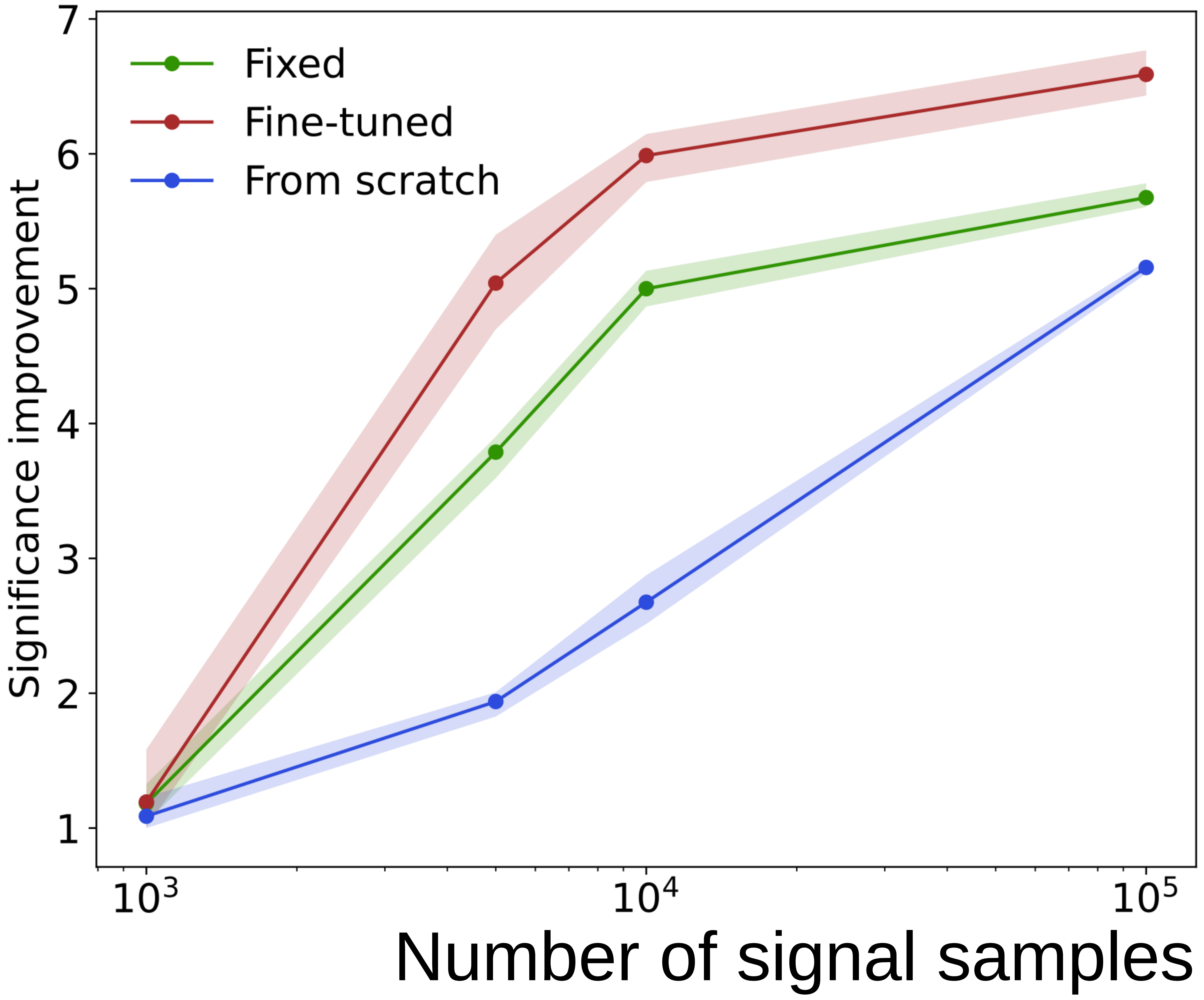
- The learned features are generically useful
- The performance gain applies to data generated with a different simulator
  - Change card to Atlas and fine-tune (JetClass is CMS)



Masked modelling

Fine tune on weak supervision

- Fine tuning with CWoLa
  - Take two QCD samples
  - Add x top jets to one sample and label ‘signal’
  - Fine-tune model on noisy labels



Summary

Masked particle modelling

- Masked particle modelling is a very useful pretraining task for HEP
- Shows great promise in example downstream classification tasks
  - More data efficient
  - Ability to extrapolate to new datasets
  - Better performance in weak supervision
- More to come!

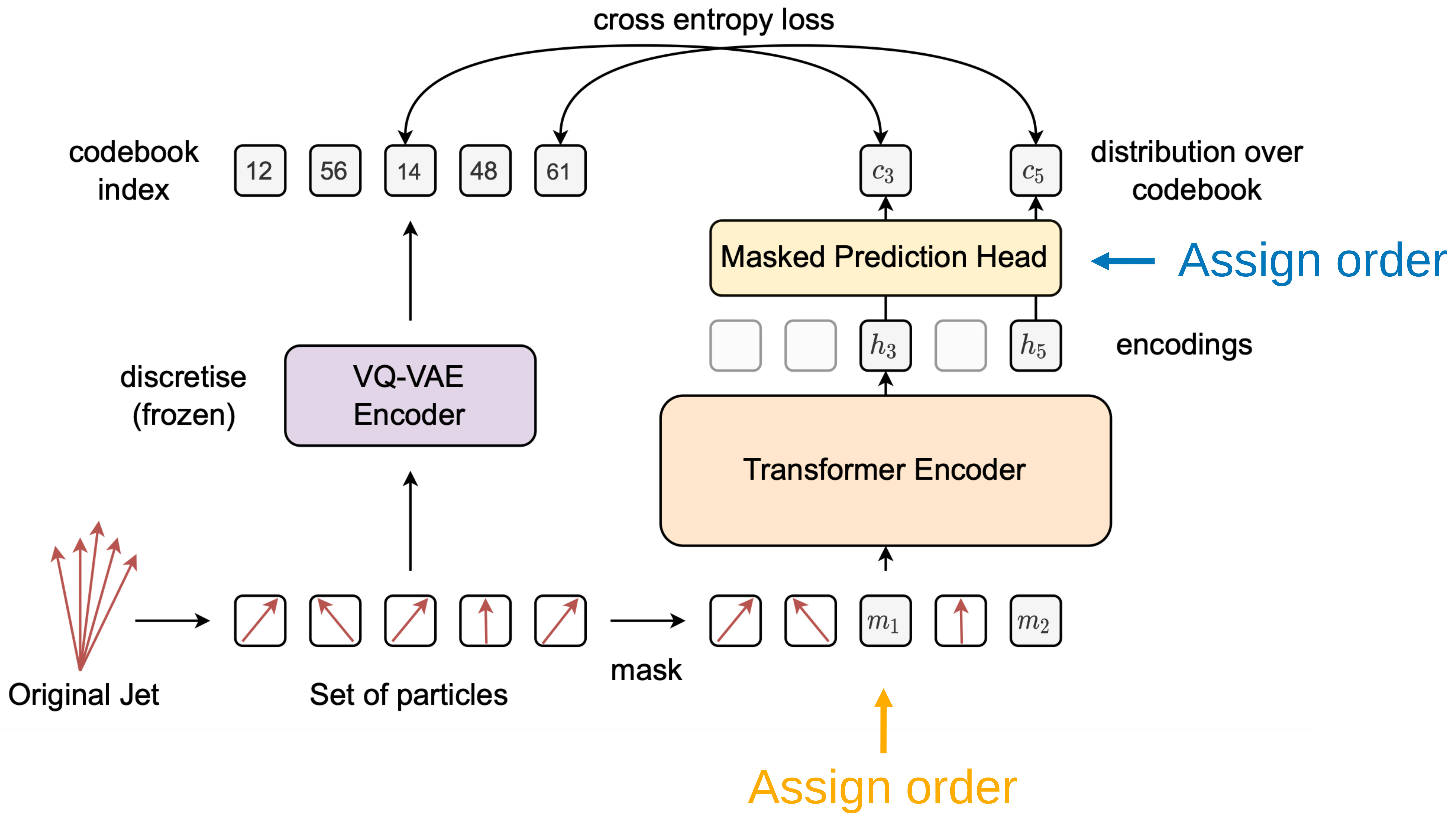
Thank You

Backup

Masked modelling

Permutation invariance

- Three approaches to permutation invariance
  - Don't worry about it
  - Input to backbone
  - Input to masked prediction head



Masked modelling

Permutation invariance

- Three approaches to permutation invariance
  - Don’t worry about it
  - Input to backbone
  - Input to masked prediction head