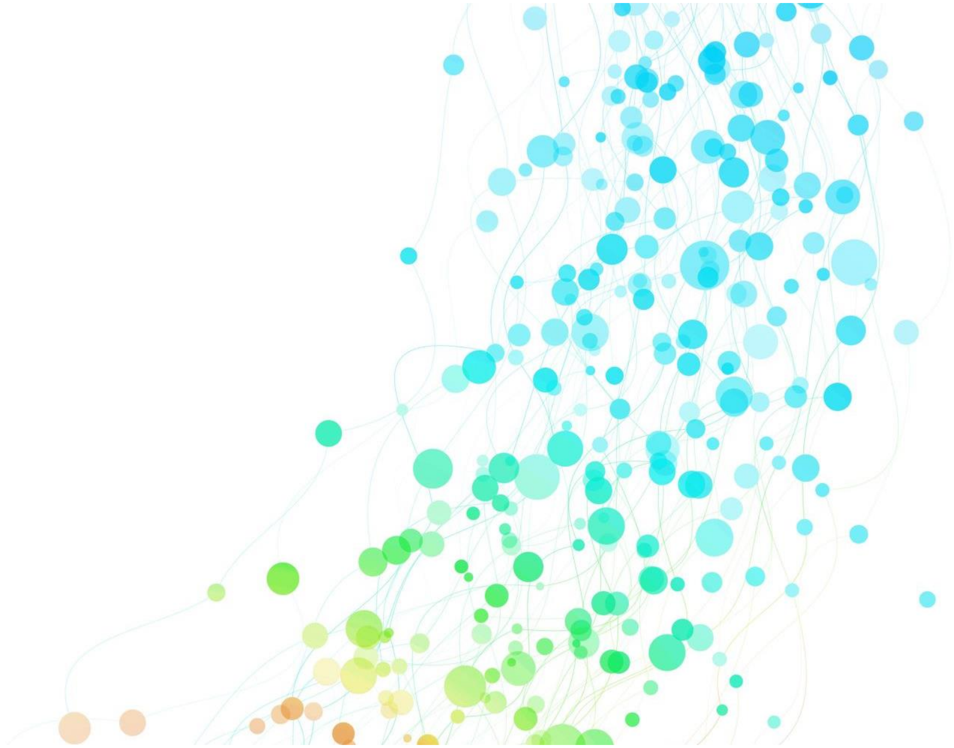
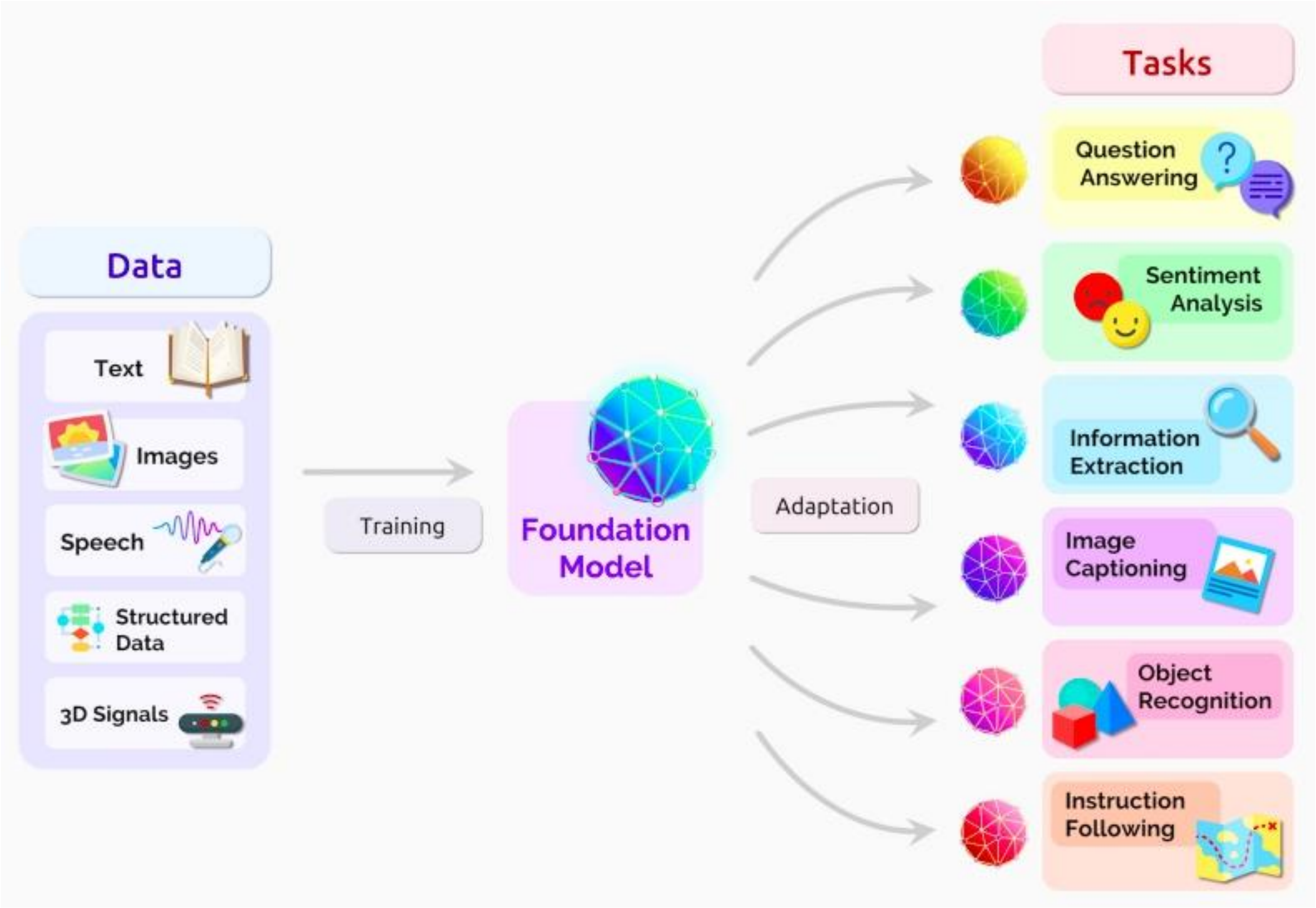
Foundation Models In Particle PHYSICS

as of November 2024



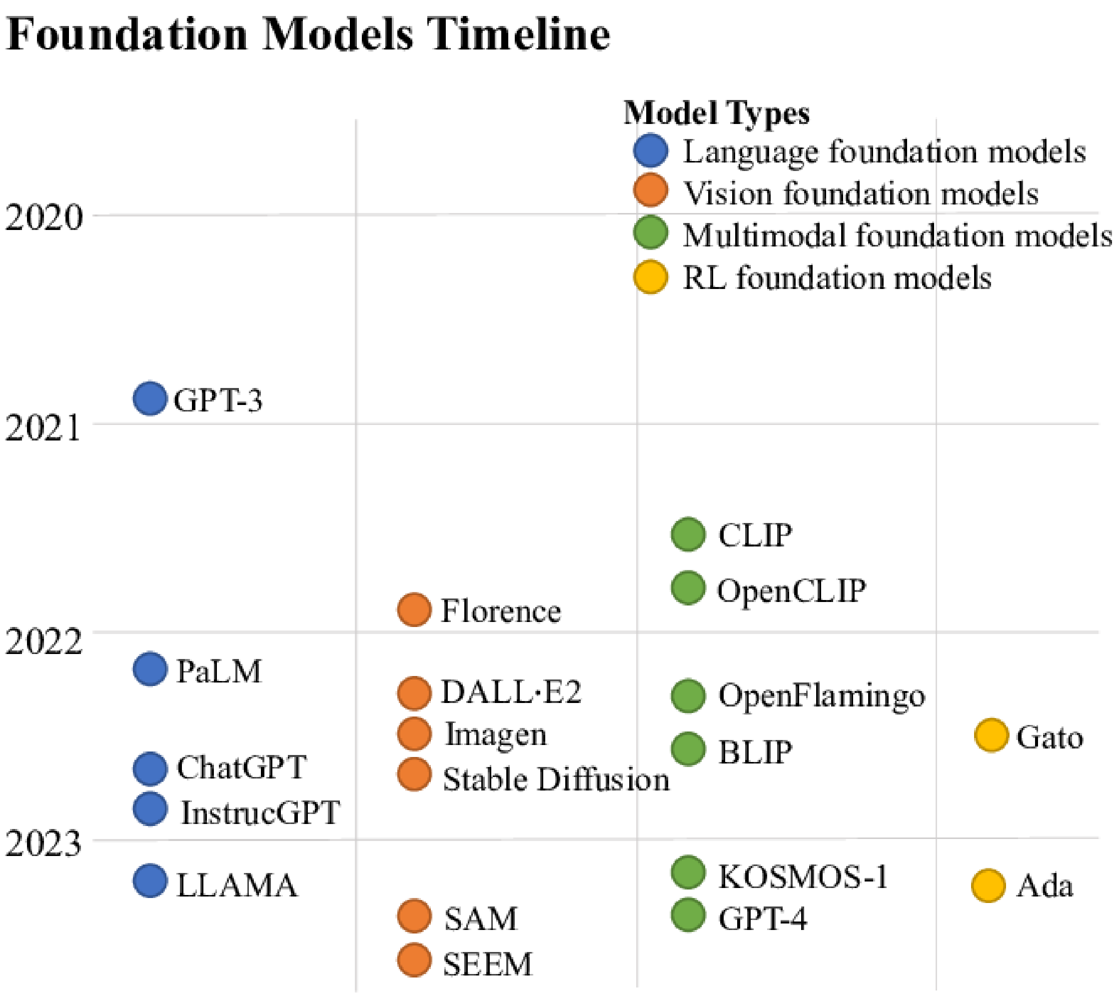


Source

Foundation models

Foundation models are multi-dataset and multi-task machine learning methods that once pre-trained can be fine-tuned for a large variety of downstream applications.

Popular models in other fields (Vision, Language)



[Source]

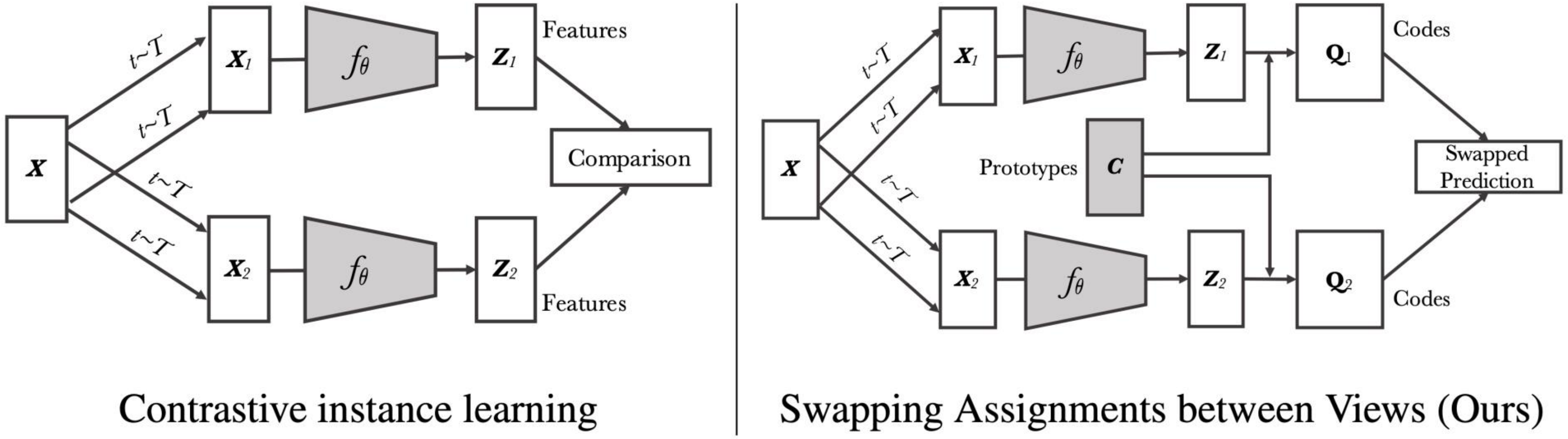
Different Methods in Self-Supervised learning

1. Contrastive Learning Representations (CLR)

Contrastive Learning is a machine learning technique that teaches computers to understand similarities and differences by comparing pairs of data points. Similar things should be close together in the computer's understanding, while different should be apart.

The process involves three main steps:

- Data augmentation
- Feature extraction
- Projection



Popular Models based on Contrastive Learning:

- **MoCo (Nov 2019) (Momentum Contrast for unsupervised Visual Representation Learning):**

  MoCo treats contrastive learning as a dynamic dictionary lookup process using two encoders - a query encoder and a momentum encoder - achieving 60.6% accuracy with ResNet-50 and scaling up to 68.6% with R50w4×, making it competitive with larger models while using standard architectures.

- **SimCLR (Feb 2020) (A simple framework for contrastive learning of visual representation):**

  SimCLR is a simplified framework for self-supervised visual learning that outperforms previous methods by data augmentations and non-linear transformations, achieving 76.5% accuracy with linear classifier and 85.8% on fine-tuned on ResNet-50.

- **CLIP (Feb 2021) (Contrastive Language-Image Pre-training)**

  CLIP (Contrastive Language-Image Pre-training) pairs images with their text captions using dual encoders (ViT for images, text encoder for captions), matching images to their descriptions in a shared embedding space, achieving notable zero-shot capabilities in image classification tasks when trained on 400M image-text pairs.

Particle Physics Adaptations of CLRs

1. **JetCLR (Aug 2022):**
   - Focuses on creating symmetry-aware representations for jets
   - Emphasizes built-in physical symmetries (rotation, translation, permutation) and soft/collinear safety
   - Uses transformer-encoder network and evaluates performance through linear classifier tests

2. **Dark CLR (Oct 2024):**
   - Focuses specifically on detecting semi-visible jets from dark sector models
   - Uses contrastive learning with "anomalous enhancements" specifically designed for dark sector physics
   - Combines representation learning with normalized autoencoder for anomaly detection

3. **AnomalyCLR (Aug 2024):**
   - More general approach for model-agnostic anomaly detection at LHC
   - Introduces "anomaly-augmentations" that mimic generic features of anomalous events (high multiplicity, large MET, large pT)
   - Uses two-step process: contrastive learning followed by autoencoder on learned representations

2. Generative Models

A generative model is a type of machine learning model that aims to learn underlying patterns or distributions of data to generate new, similar data.

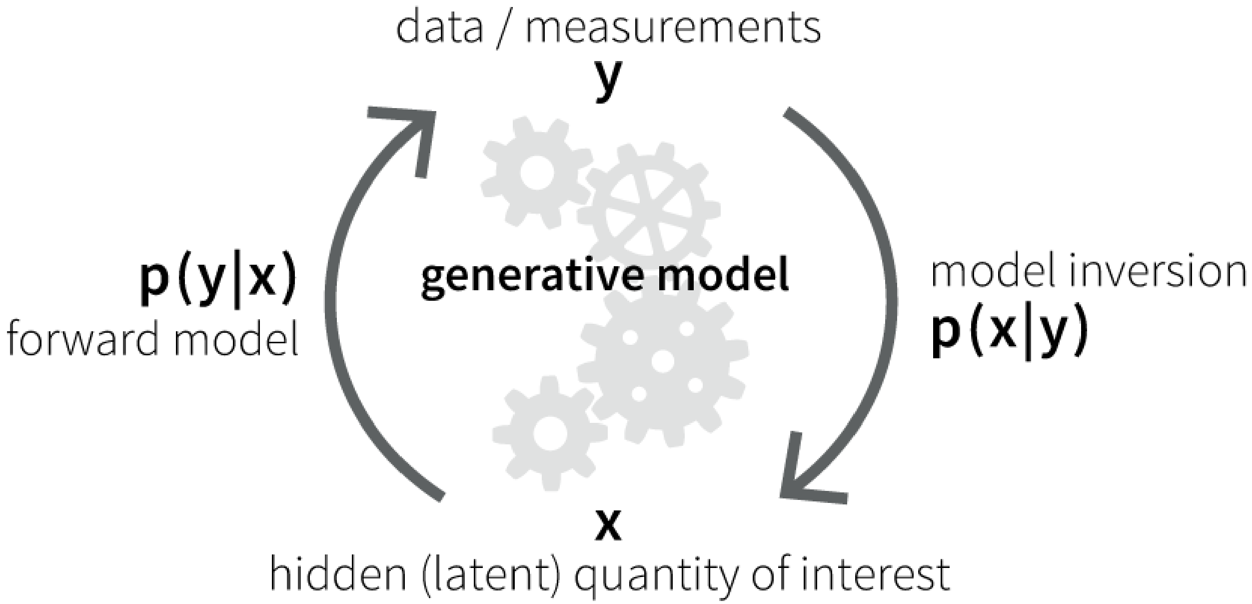
This is often used in the context of enabling computers to understand the real world.

Generative models can be both probabilistic or Neural Network based. Currently generative models can generate:

- Images (Dalle, Stable Diffusion, MidJourney)
- Text Generation (LLaMA, GPT, Mistral)
- Audio and Music (WaveNet, BachBOt)

Generative models are too broad of a topic.

We can cover some of the popular approaches and efforts in adapting them to physics.



parT and OmniJet-α

- **ParT (Particle Transformer) (July 2023) :**
  ParT is based on the traditional Transformer architecture and a successor to ParticleNet, but introduces a modified attention mechanism specifically for jet tagging. Its major innovation is adding a new term to capture pairwise particle interactions within the attention mechanism. The model processes jets as particle clouds (unordered, variable-sized sets of outgoing particles) and was trained on the massive JETCLASS dataset, which is significantly larger than previous datasets. ParT's training approach is straightforward supervised learning for jet classification, contrasting with more complex pre-training strategies.
  Note: There is a new model MiParT which is works and in version 3, which accommodates the More-Interaction Attention (MIA) mechanism.

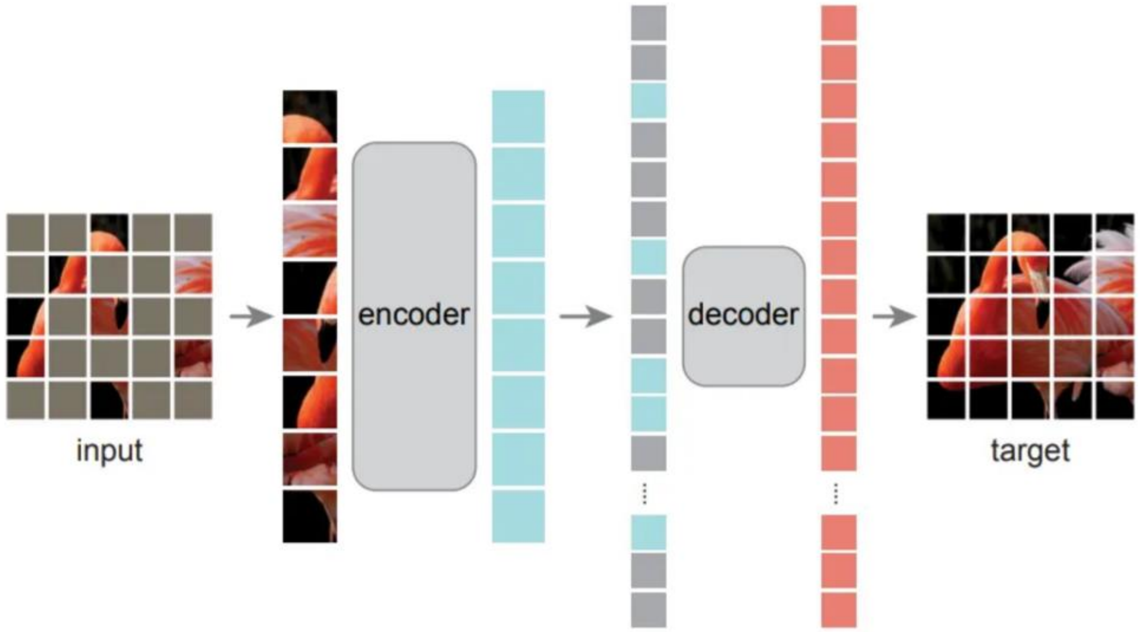
- **OmniJet (Mar 2024):**
  OmniJet is based on the GPT but adapted for continuous physics data. The model's key innovation is its tokenization strategy for converting continuous particle physics data into discrete tokens, expanding the codebook size to 8192 tokens (up from 512 in previous approaches). Unlike ParT's direct classification approach, OmniJet first trains in an unsupervised manner to generate jets as tokens, following the successful pre-training paradigm of language models. The model demonstrates transfer learning capabilities, where the knowledge gained during generative pre-training can be applied to downstream classification tasks. Finally, OmniJet aims to be a foundation model for multiple jet physics tasks, contrasting with ParT's focus on classification specifically.

Masked Autoencoders (MAE)

The idea is to mask random patches of input images and reconstruct them back. This is a computer vision adaptation of the Masked Language Model (or general Masked Token Prediction) in NLP, which are strongly built on the concept of Next Word Predictors.

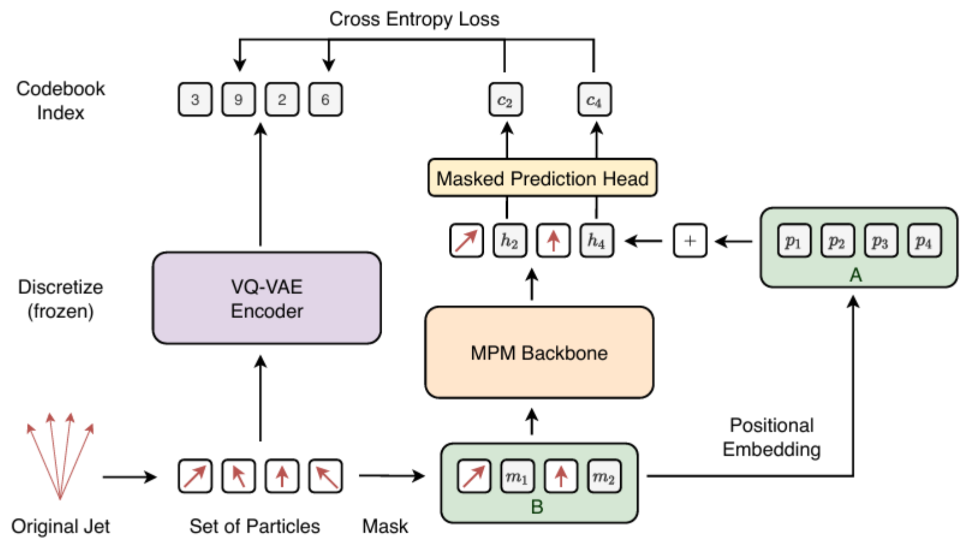
Thus MAE is a self-supervised learning technique that masks (removes) portions of an image and learns to reconstruct them.

- Randomly masks a very high portion of image patches (typically 75%)
- Uses much higher masking ratio than NLP because images have more redundant information
- This forces the model to understand the overall image context rather than just copy nearby pixels



MPM (July 2024) and MPMv2

- Unlike MAE which masks image patches, MPM masks individual particles in a set
- Combines two key techniques:
  - Vector Quantized Variational Autoencoder (VQ-VAE) for token representation
  - Masked modeling similar to BERT/MAE but adapted for particle sets
- Critically maintains permutation invariance, which is essential for particle physics but not relevant for MAE's image patches



Note: There is currently a newer version of Masked Particle Modeling in works. This uses a more powerful decoder and tries to use conditional generative models without the tokenization.

THANK YOU