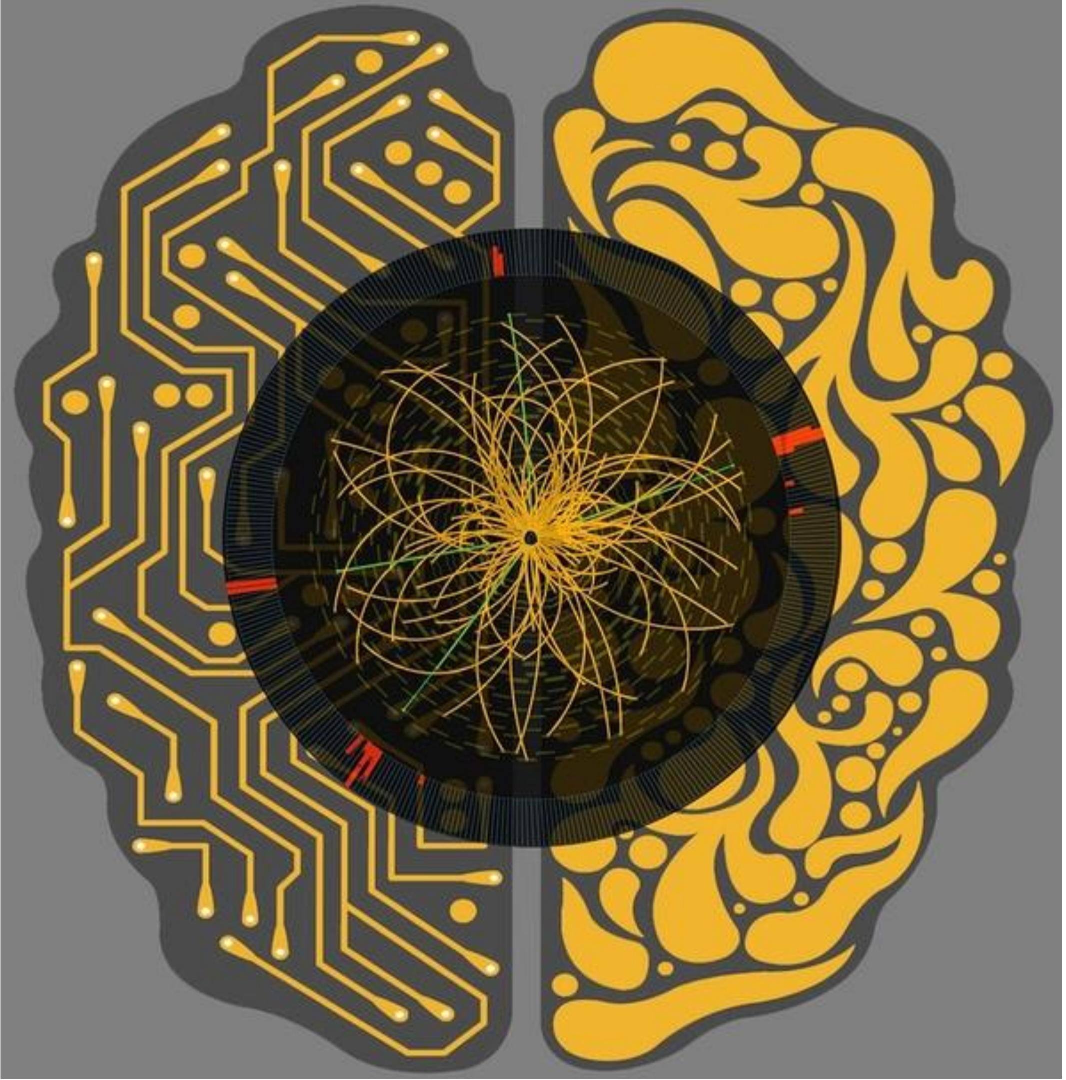


Building Foundation Models with low-level collider data

Aritra Das, Daniel Murnane, <u>Shreya Saha</u>

HAMLET-Physics
Copenhagen 2026

IISER KOLKATA
KØBENHAVNS UNIVERSITET
Adelaide University

Foundation Models

Dataset

Task

Learn missing parts, rotations or context.

Pretraining

Train model : dataset + task + learning method

Latent Representation

Model learns features (wheels, mirrors,windows...)

Downstream tasks

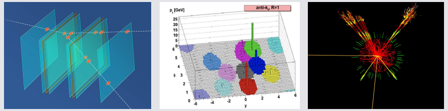
Car?

Truck?

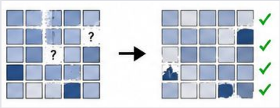
Bike?

Foundation Models for Particle Physics

Dataset



Task



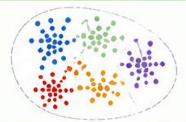
Learn missing parts, rotations or context.

Pretraining

Train model : dataset + task + learning method

Latent Representation

Model learns features (kinematics, signature patterns, correlations…)



MPM JetSSL OmniJet-Alpha

Downstream tasks

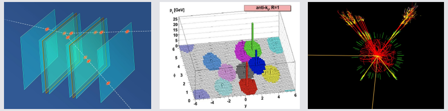
Event Classification

Anomaly Detection

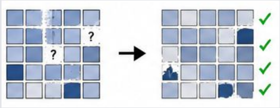
Object Classification

Foundation Models for Particle Physics

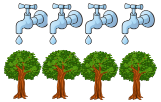
Dataset



Task



Learn missing parts, rotations or context.



Expensive!

Pretraining

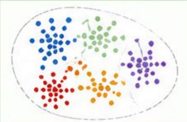
Train model : dataset + task + learning method

Source

~10.8kg CO2e

Latent Representation

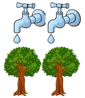
Model learns features (kinematics, signature patterns, correlations…)



Optimize resources...

MPM JetSSL OmniJet-Alpha

Downstream tasks



Event Classification

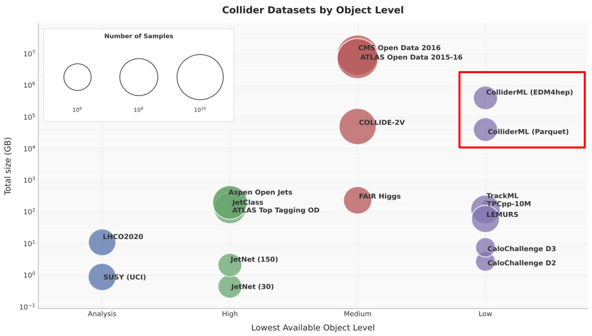
Anomaly Detection

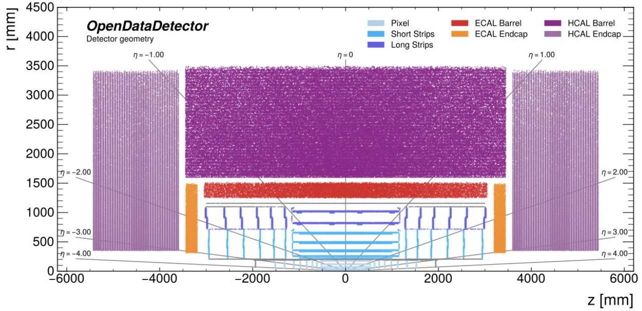
Object Classification

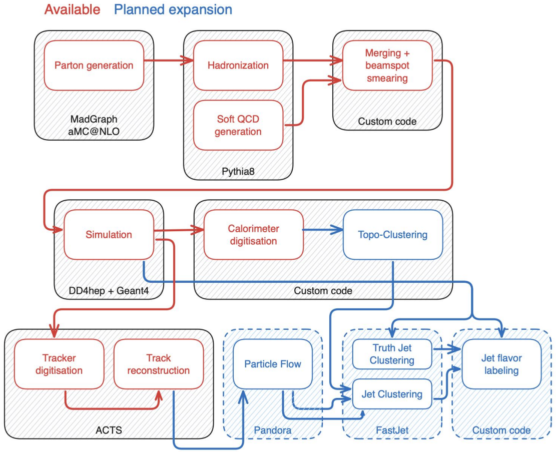
ColliderML

GitHub Paper CTD 2025

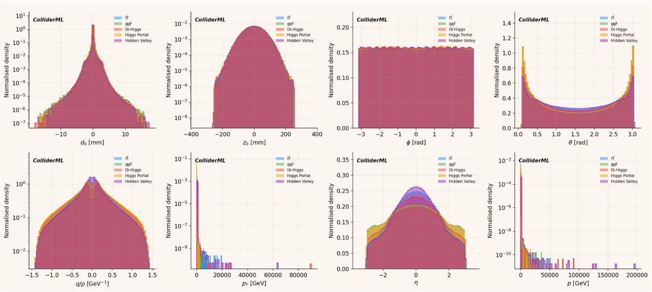
- Benchmark **full simulation** dataset with **low level information** (tracks, hits, calorimeter deposits) for High Luminosity LHC.
- Reconstruction based on the Open Data Detector and *A Common Tracking Software* (ACTS).
- **Standard Model processes -** Top pair production (pp→tt), gluon-gluon fusion (ggF, pp→H), Di-Higgs (gg→HH)
- **Beyond Standard Model processes -** Higgs Portal production of a long-lived scalar (H→aa→bbbb) and Hidden Valley (pp→Z′→ Dark QCD)
- Zero pileup option used for this study.

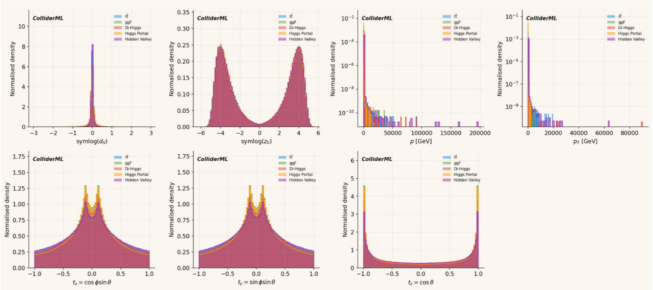


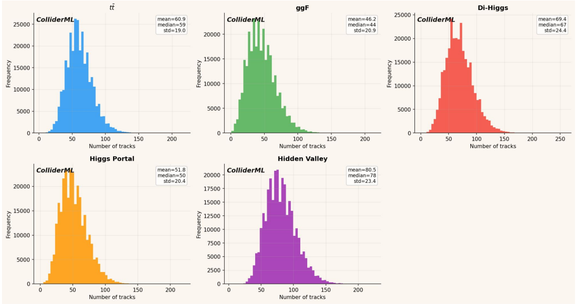




Feature distribution (tracks)







Raw track variables

Transverse impact parameter (**d_0**), longitudinal impact parameter **(z_0),** azimuthal angle **(ϕ),** polar angle **(θ),** charge to momentum ratio **(q/p),** transverse momentum **(p_T),** pseudorapidity **(η)**, momentum **(p)**

Engineered track variables (pretraining)

Log transformations for **d_0** and **z_0**,
Polar to Cartesian coordinates
→ **t_x** = cos(ϕ) sin(θ), **t_y** = sin(ϕ) sin(θ),
t_z = cos(θ)
p, **p_T**

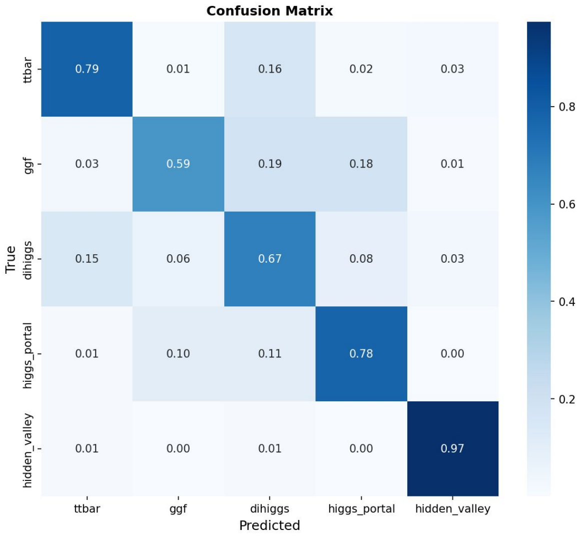
Number of tracks per event for each physics processes

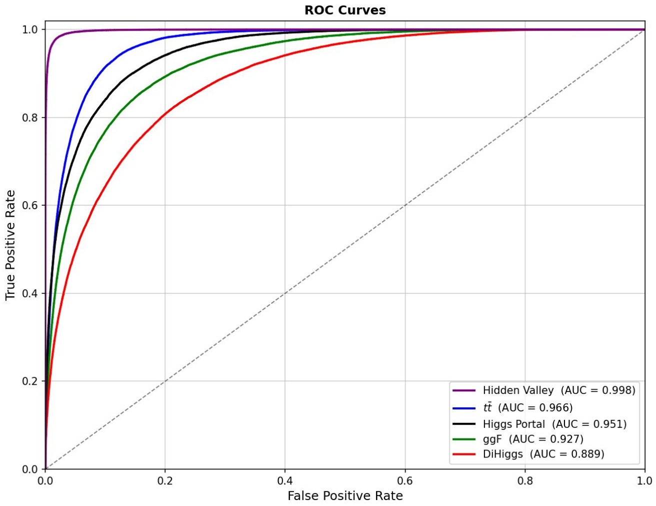
Highest number of tracks - **Hidden Valley**

Lowest number of tracks - **ggF**

Vanilla Transformer

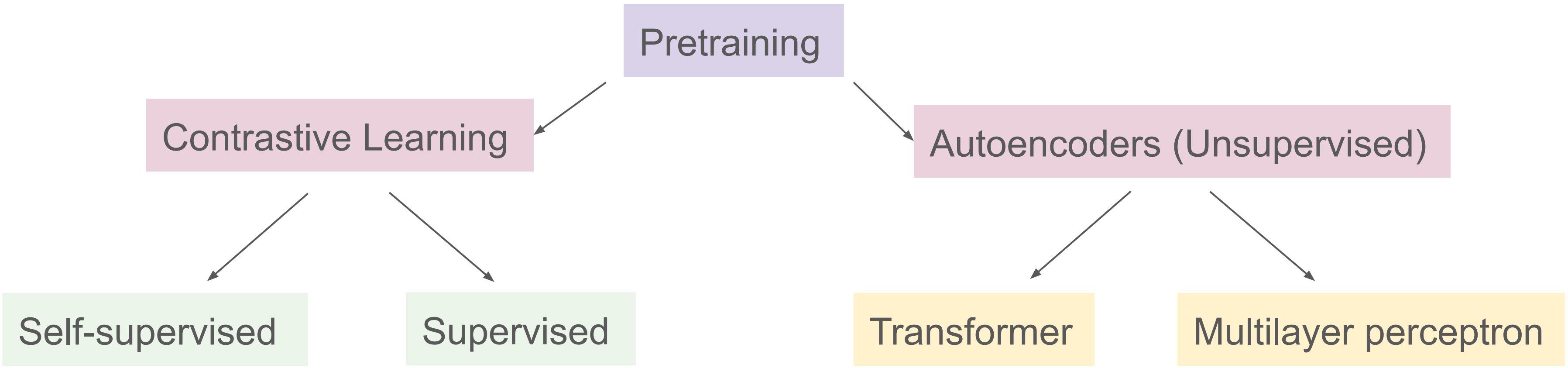
- **Input -** 300,000 events per process, raw track variables.
- Tracks → 128-dimensional embeddings → 8 transformer layers X 8 attention heads.
- **Classification head**: 64-dimensional multi-layer perceptron (MLP) maps the latent representation to classes.
- **Highest efficiency -** Hidden Valley
- **Lowest efficiency** – ggF
- **Largest misclassification** for ggF is with the Di-Higgs process, due to overlapping final state signatures.





Pretraining the model

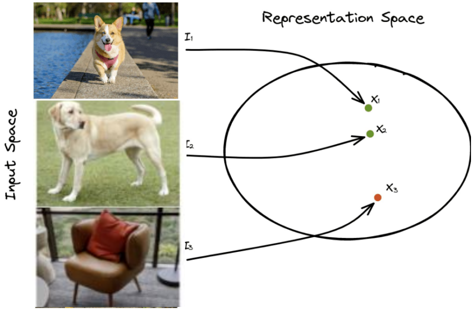
The model is pretrained using **track** information, and divided into various types depending on the **learning method** (*Unsupervised, Self-supervised, Supervised*) and the **model architecture**.

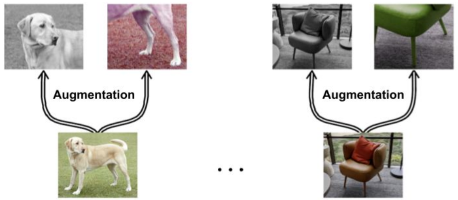


Contrastive Learning

Contrastive learning (CL) learns representations of data by **contrasting** between **similar** and **dissimilar samples.**
Input - 7 engineered track variables per track.

Metzger, Kyle, et al





Self Supervised CL (No labels)

- Two independent **physics-invariant augmentations** (Random rotation and Gaussian smearing) are applied to each event.
- Augmentations from the **same event are grouped together** in the latent space.
- Loss **maximizes** the **similarity** between the **event and its positive counterpart**, and **minimizes** the **similarity** to all **other representations**.

Supervised CL (Labels)

- Similar machinery as above, but uses **physics process labels** instead of augmentations.
- Loss function is maximized to **group events** from the **same process together**.

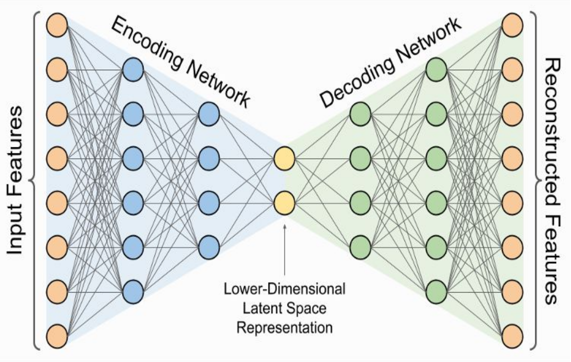
Output - Pretrained models with D = 8/32/64 dimensional latent space.

Autoencoders

The goal of the autoencoder is to compress every event **(encoder)** into a **latent representation,** to capture the most important event features, and further reconstruct the original event **(decoder)**. **No labels** provided.

Transformer-based architecture

- **Input -** Engineered track variables.
- **Reconstruction target -** Summary statistics (mean, standard deviation, minimum and maximum value) per track variable.
- **Encoder -** 4 layers X 4 **attention heads**/ layer, **decoder** is a three-layer MLP (Dimensions: 128→ 256→ 28).
- **Loss function** = Mean squared error.



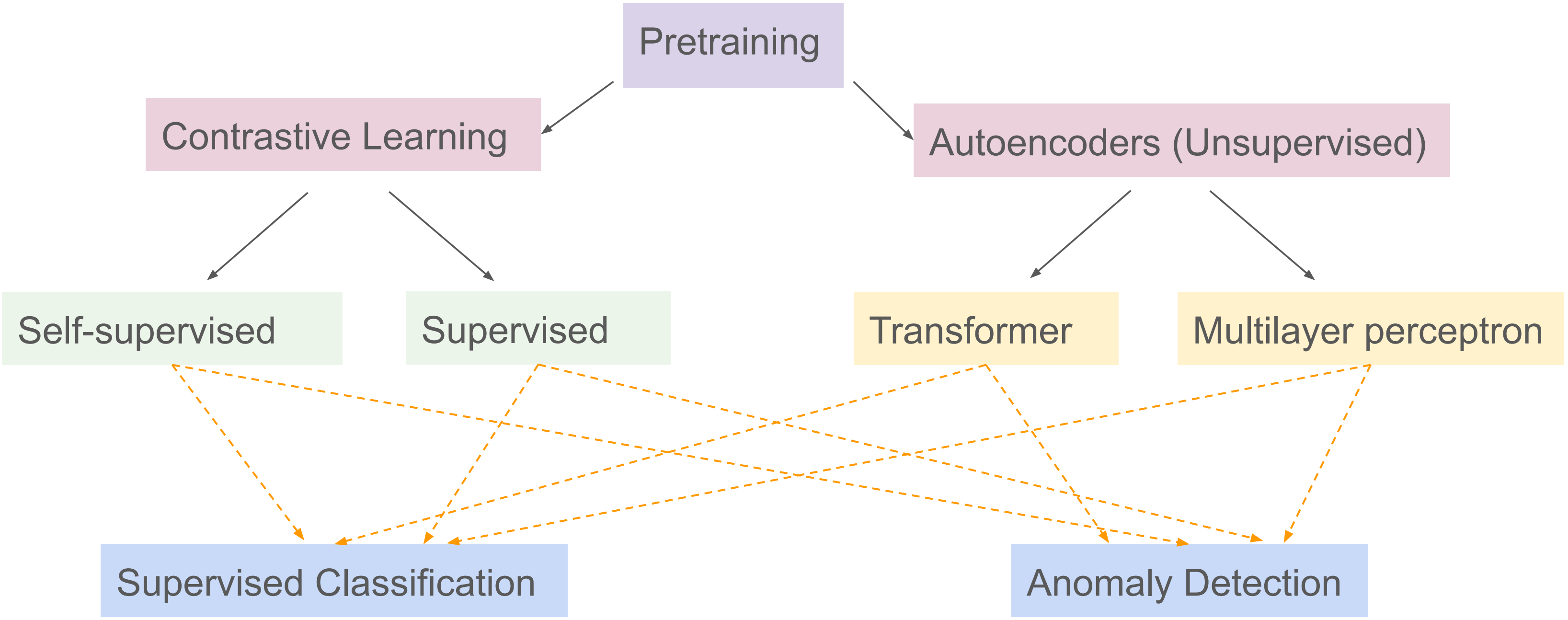
Multilayer Perceptron (MLP) based architecture

- **Input -** Summary statistics
- **Reconstruction target -** Summary statistics
- **Encoder -** Three- layer MLP (28 → 256 → 128 → D), decoder (above).
- **Loss function** = Mean squared error.

Output - Pretrained models with D = 8/32/64 dimensional latent space.

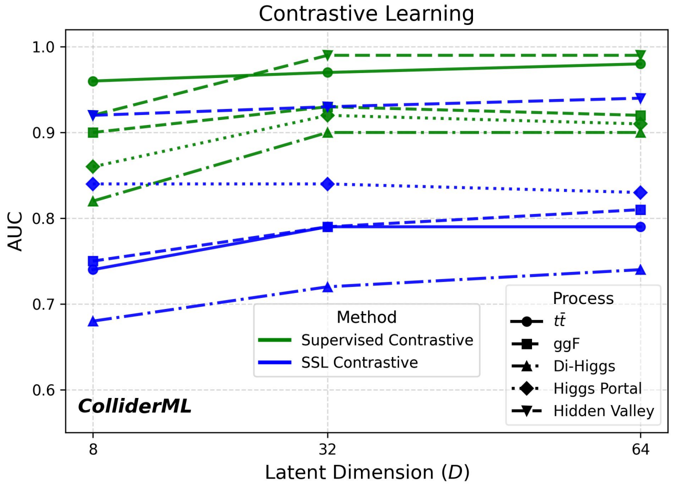
Downstream tasks

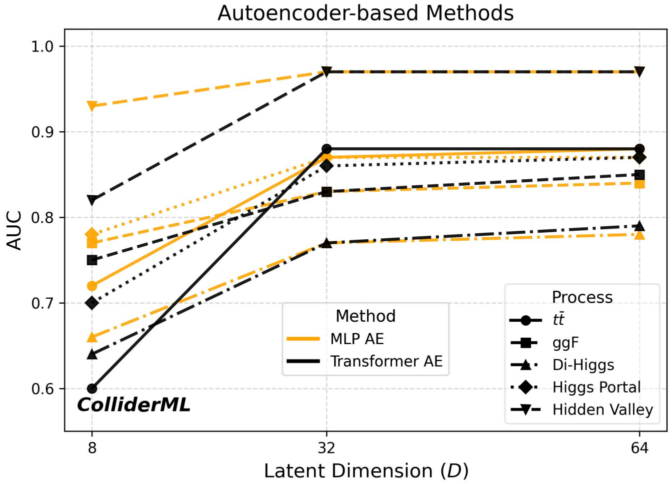
- Each pretrained model is used to produce event embeddings from the frozen encoder.
- Performance on downstream tasks → **quality of the latent representations** learned by each model.
- Supervised Classification, Anomaly Detection.



Supervised Classification

- **Frozen encoder event embeddings + labels** are used as input.
- A 5-dimensional linear probe is trained, keeping the model frozen.
- Labels during pretraining improve downstream classification, with **Supervised CL** (D= 64) achieving **highest accuracy**.
- **Transformer AE** (D = 8) gives **lowest accuracy**.
- Hidden Valley classification consistently shows the highest efficiency across models, owing to its distinct final-state features.



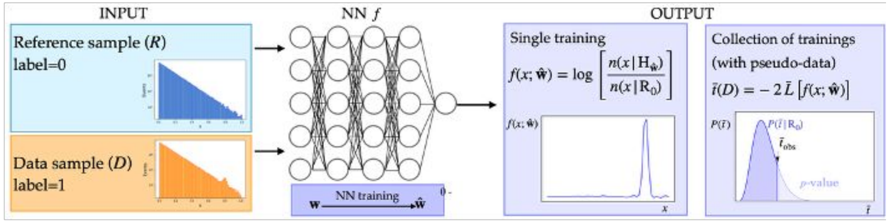


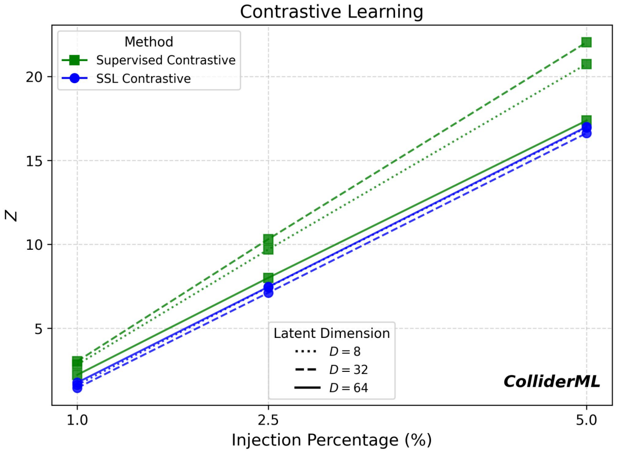
Anomaly Detection

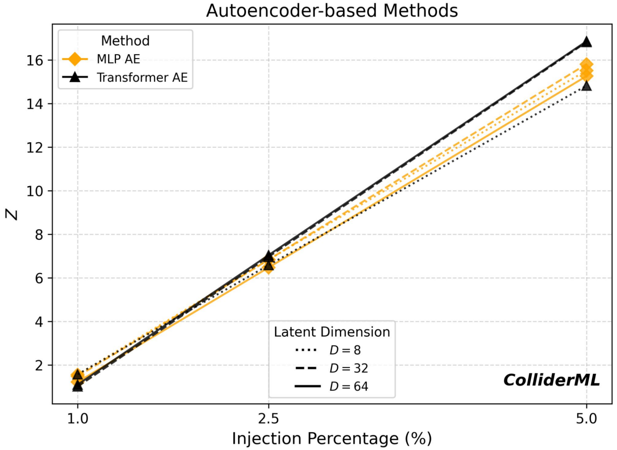
NPLM talk by J. Harrison today

- **New Physics Learning Machine** : Compare the background hypothesis (R) to an unknown alternative (H_w) learnt from the data using a neural network.
- Perform 5000 pseudo-experiments and evaluate **asymptotic Z value** at varying signal injection levels.

G.Grosso







Supervised CL shows better performance, highlighting the **benefit of label** information provided during pretraining.

Conclusion and Outlook

- Low-level collider data can be used to build robust FMs, showing promising performance in classification and anomaly detection tasks.
- Supervised Contrastive Learning performs better than other pretraining methods, due to the inclusion of labels in the training process.
- Sustainable computing pratices with ML research - **SC4RC** **SUSHep**, **Green Algorithms**

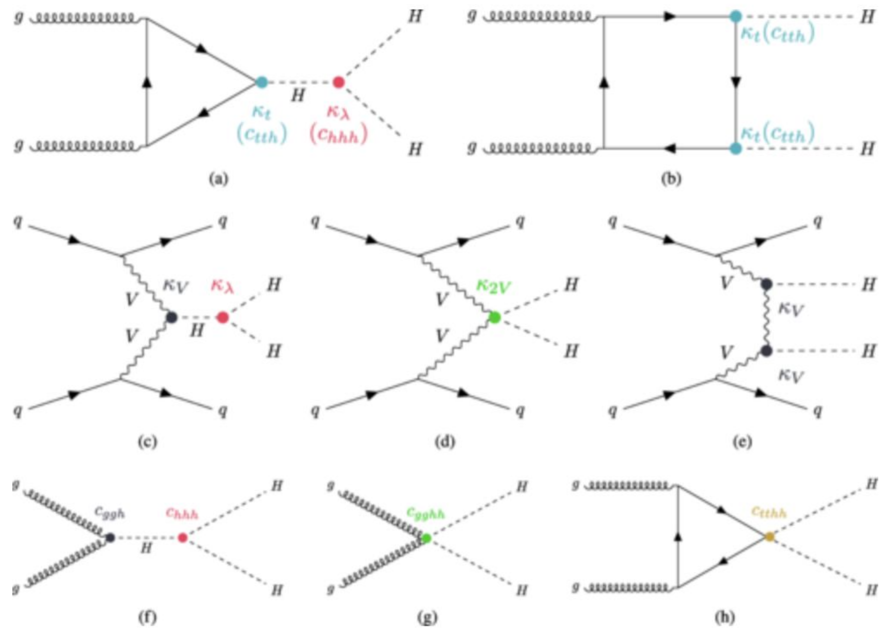
Future Work

- Compare fast vs. full simulation versions of ColliderML for more efficient models for the High-Luminosity LHC.
- Explore various downstream tasks, such as object-level anomaly classification.
- Include additional Standard Model processes (diboson and triboson events).
- Study model performance by incorporating calorimeter information alongside tracks.

Thank you!
Questions?

Backup

ColliderML dataset



ID	Process Label	Process Type	Comments
1	`ttbar`	$pp \to t\bar{t}$	Top quark production; important for SM tests and complex final states.
2	`zee`	$pp \to Z \to e^+e^-$	Drell-Yan process; detector calibration and electroweak probe.
3	`zmumu`	$pp \to Z \to \mu^+\mu^-$	Drell-Yan process; detector calibration and tracking probe.
4	`dihiggs`	$gg \to HH$	Gluon fusion di-Higgs; benchmark for high-Luminosity studies.
5	`diphoton`	$pp \to \gamma\gamma$	Diphoton production; loop-induced QCD/EW process.
6	`multijet`	$pp \to$ jets	QCD production; dominant background for hadronic signatures.
7	`ggf`	$pp \to H$	Gluon-gluon fusion; main Higgs production channel.
8	`susy`	$pp \to \tilde{g}\tilde{g}$	GMSB gluino production; $\tilde{\chi}_1^0$ NLSP provides displaced vertex benchmark.
9	`zprime`	$pp \to Z'_{SSM}$	Heavy resonance; standard candle for bump hunting.
10	`hiddenvalley`	$pp \to Z' \to$ dark	Dark QCD sector; semi-visible jets for anomaly detection.