Welcome to WG1: Foundation Models (FMs)

Tobias Golling (UniGE)

Lukas Heinrich (TUM)

Proposed agenda of today

Introduction to Foundation Models (5’)

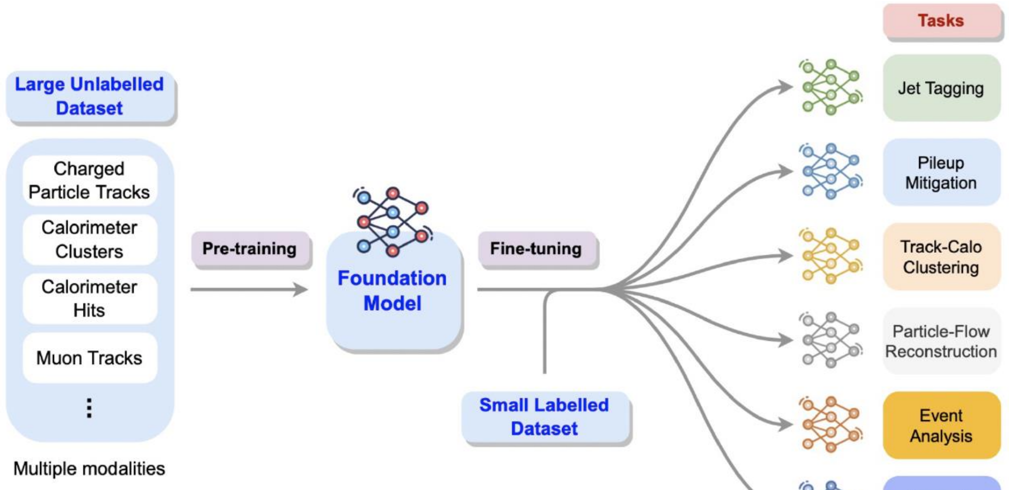
Quick reminder & recap & poll summary (5’)

~~Get to know each other~~ (0’)

What do **you** want to do (15’)

Discussion, next steps & AOB (35’)

FMs so far in EuCAIFCon2025



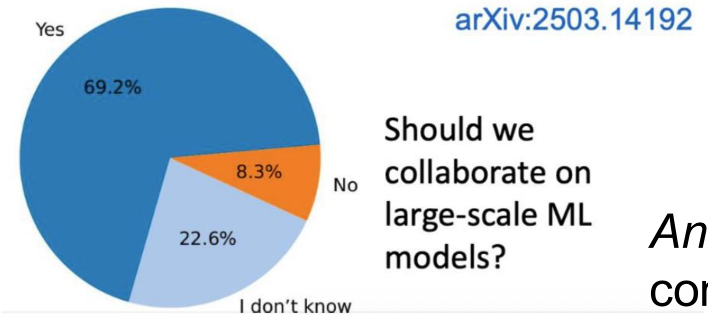
from M. Kagan at H&N 2023

AI in HEP
Jennifer Ngadiuba

Follow-up question by Sascha: learn representation of all events before triggering

“End-to-end” is not such if you only optimize the software, or only the hardware. Doing everything together – that’s what we should aim for. It is called co-design.

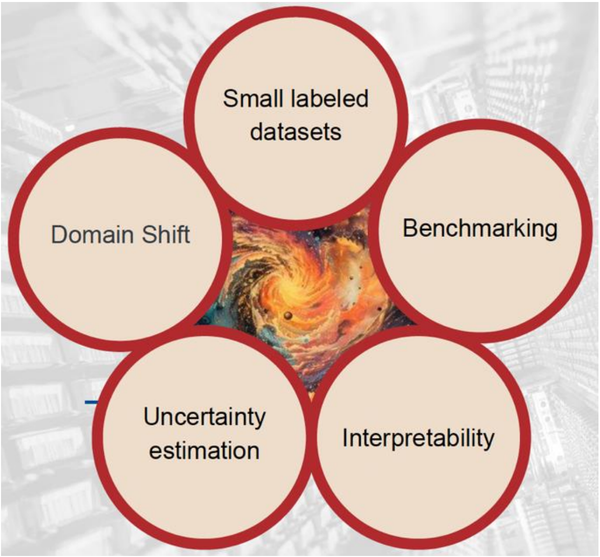
Tommaso Dorigo: connection to WG2



arXiv:2503.14192

Should we collaborate on large-scale ML models?

Andreas Ipp: connection to WG4



AI in astrophysics
Aleksandra Ćiprijanović

Munich Workshop on Foundation Models

4-week program to discuss “inductive bias vs scale”

→ 2-day Workshop embedded for EuCAIF **Sept 4 – Sept 5**

Contact | Press | Publications

MIAPbP

About MIAPbP Activities Registration For Visitors Propose

Image Credit: iStock

BUILD BIG OR BUILD SMART: EXAMING SCALE AND DOMAIN KNOWLEDGE IN MACHINE LEARNING FOR FUNDAMENTAL PHYSICS

25 August - 19 September 2025

Lukas Heinrich, Michael Kagan, Margarita Osadchy, Tobias Golling, Siddarth Mishra-Sharma

https://www.munich-iapbp.de/activities/activities-2025/machine-learning

Geneva Workshop on FMs & Co-design

Villa Boninchi at lake Geneva **Oct 24 – Oct 28**

– If interested, please contact Tobias.Golling@unige.ch

Do we all know what we mean when we say **Foundation Model**?

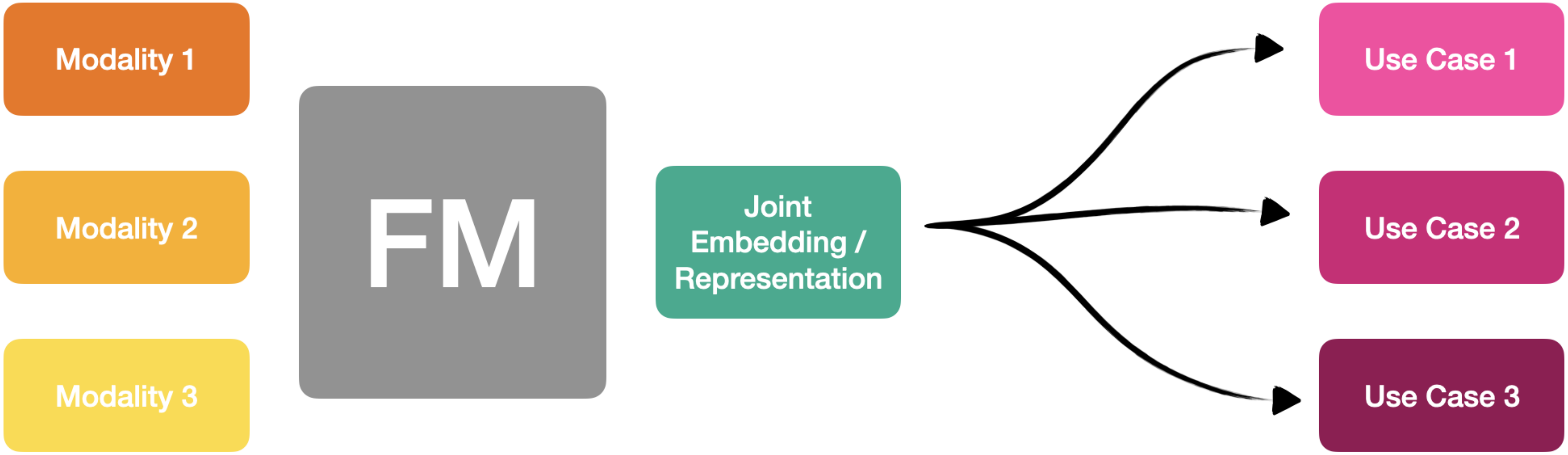
It has become an overloaded term

WG1: Foundation models (FMs)

Background: Pioneered in LLMs like ChatGPT or image generators like DALLE

What is a FM? Multimodal FMs centralize information from various data modalities & domains & encode them in a common meaningful latent representation [pre-training] + multi-head fine-tuning [post-training]

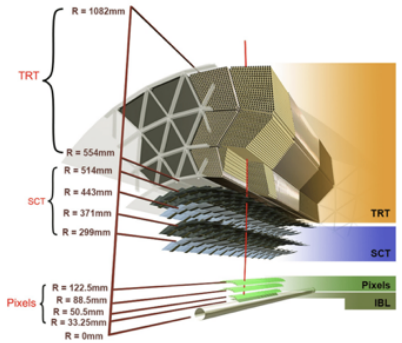
Why a FM? Amortization, automation, acceleration [compute & person-power]: Narrow task-centric → multi-task, reusable, data+MC-trained backbone, reduce uncertainties



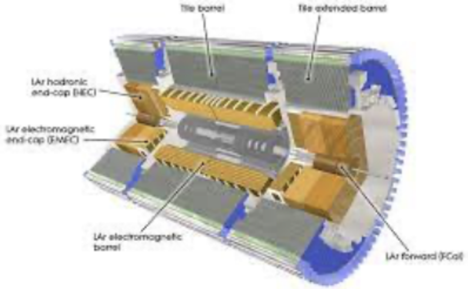
[Credit: Lukas Heinrich]

Example: HEP

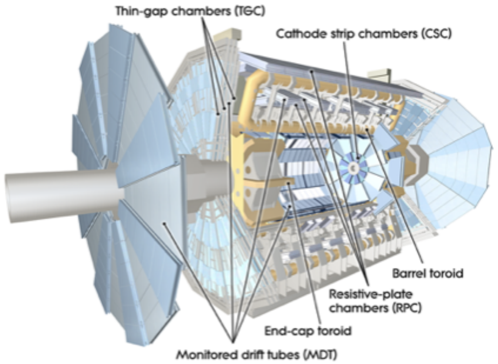
Tracking Data

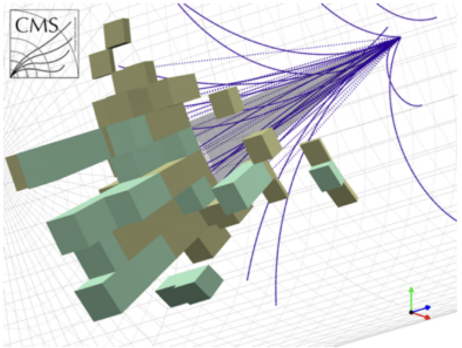


Calorimeter



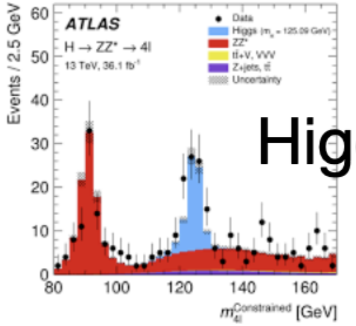
Muon Data



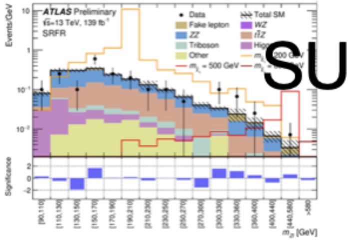


Reco

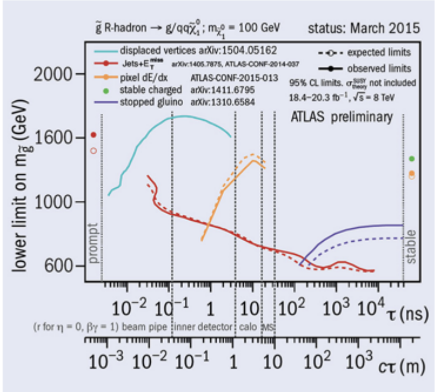
Reconstructed Event



Higgs



SUSY



Exotic Particles

In many ways we've always had a foundation model for general purpose experiments..

[Credit: Lukas Heinrich]

FM to address bottlenecks in HEP

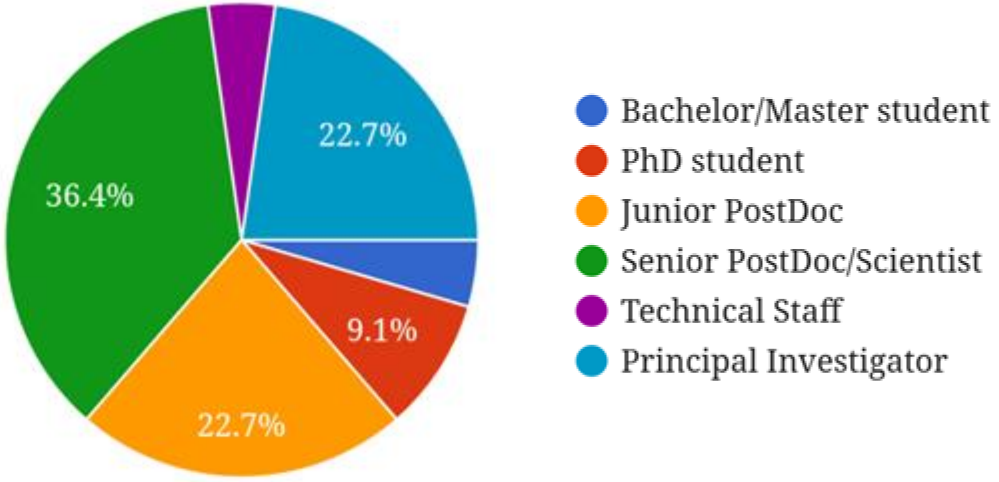
Vast amounts of **synthetic data** needed to model S and B (compute bottleneck)

Little room for **re-utilization** due to highly specific approaches (person-power bottleneck)

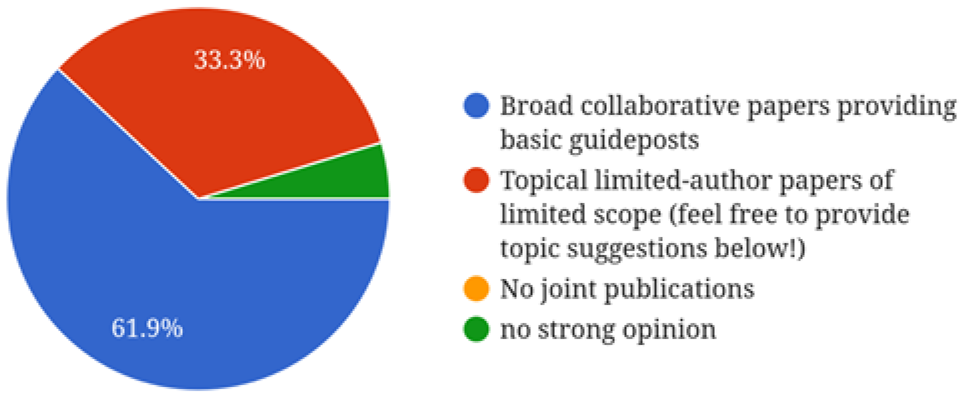
Domain shifts between real and synthetic data lead to a bottleneck of systematic uncertainties (compute+human)

https://arxiv.org/abs/2503.14192

Poll mini-summary



Publication Preferences



Whole community represented

Various pre-training strategies & downstream tasks

Need for more high-stats public data in more suitable format

Full support of FM challenge

Interests, network, collaborate, training, topical meetings

For full poll results see: https://indico.cern.ch/event/1537738/

Mini-minutes from last WG1 meeting

- **Diverse** attendants (33 max), diverse downstream tasks
- Include **theory modality** → symmetries,…?
- Action items
  - Kick-off **seminar series** + invite other domains (weather, climate,…)
  - Define sub-topics for **topical meetings** + volunteers to help coordinate
  - Do **match-making**: who is interested in XYZ / what is the FTE
  - Incentives: address **your bottlenecks & needs** = your win
  - Can we make the case for **scaling** → supercomputing case
  - Organise a **data day**: what data exists for which domain [need multi-modal & raw data]
  - **Models across experiments**: ATLAS/CMS or LISA/LIGO,...
  - **Modular training**: extend FM knowledge by adding new data
  - Approach: **end-to-end** *hits-to-Higgs* vs alternative: start from LLM & inject science
  - **Interpretability**: numerical and symbolic data

For slides from April 11 2025 see: https://indico.cern.ch/event/1537738/

What do **you** want to do?

Bottom-up, you pitch & we discuss/develop, cluster & link efforts to maximize **progress as a community** [local *career loss* → community]

What we can do now

Sketch common (living) strategy → white paper

Breaking it down into topics / work packages [next slide]
→ limited-author publications

Timeline + FTE + attach names

Seminars, data-day, match-making,…

Possible work packages

Pre-training: contrastive, superv., self-supervised, masking, autoregr.
Post-training: downstream task definition
Domain shift
Physics encoding
Automation
Explainability & interpretability
Define metric(s) [e.g. 95% of asymptotic performance]
Scaling
Language & symbolic encoding (+ theory modalities)
…

Show of hands 🖐 please

Define common benchmark / challenge

Define **benchmark data** (future-proof & extendable)
– REFERENCE in community – to develop & compare models

Inclusive: involve whole community

Where to **host** [ML4Jets was pitched to us]

Define **tasks & metrics**

Need **volunteers** [potential high impact]

Example common benchmark dataset

HEP Example since last EuCAIF: Jets

Aspen Open Jets: Unlocking LHC Data for Foundation Models in Particle Physics

Oz Amram,[1,*] Luca Anzalone,[2,3,†] Joschka Birk,[4,‡] Darius A. Faroughy,[5,§] Anna Hallin,[4,¶] Gregor Kasieczka,[4,6,‖] Michael Krämer,[7,**] Ian Pang,[5,††] Humberto Reyes-Gonzalez,[7,‡‡] and David Shih[5,§§]

[1] *Fermi National Accelerator Laboratory, Batavia, IL 60510, USA*
[2] *Department of Physics and Astronomy (DIFA), University of Bologna, 40127 Bologna, Italy*
[3] *Istituto Nazionale di Fisica Nucleare (INFN), 40127 Bologna, Italy*
[4] *Institut für Experimentalphysik, Universität Hamburg, 22761 Hamburg, Germany*
[5] *NHETC, Dept. of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854, USA*
[6] *Center for Data and Computing in Natural Sciences (CDCS), 22607 Hamburg, Germany*
[7] *Institut für Theoretische Teilchenphysik und Kosmologie, RWTH Aachen University, 52074 Aachen, Germany*

Foundation models are deep learning models pre-trained on large amounts of data which are capable of generalizing to multiple datasets and/or downstream tasks. This work demonstrates how data collected by the CMS experiment at the Large Hadron Collider can be useful in pre-training foundation models for HEP. Specifically, we introduce the AspenOpenJets dataset, consisting of approximately 180M high p_T jets derived from CMS 2016 Open Data. We show how pre-training the OmniJet-α foundation model on AspenOpenJets improves performance on generative tasks with significant domain shift: generating boosted top and QCD jets from the simulated JetClass dataset. In addition to demonstrating the power of pre-training of a jet-based foundation model on actual proton-proton collision data, we provide the ML-ready derived AspenOpenJets dataset for further public use.

p-ph] 13 Dec 2024

I. INTRODUCTION

While particle physics has long used machine learning techniques and is leading the way in adopting modern

data. In addition, a foundation model can save both human and computational resources. While pre-training may be a resource-intensive task, the downstream models would require less training, less data, and less time spent

Opportunity for junior researchers

Postdocs/junior faculty – the obvious volunteers 😎

- Opportunity: drive community, empower others – crucial skill
- Challenging & rewarding: “convener” on exciting R&D project
- Lukas, Tobias and other seniors here to guide and support

Everyone welcome to volunteer !!!

Backup

What to do after this one hour we have

- Join the EUCAIF-wg-FOUNDATIONMODELS e-group
  - >50 members

- Let’s make the best out of EuCAIFCon2025
  - **You** are the WG: find us & find each other during the breaks to continue the discussions
  - Follow up after EuCAIFCon2025

Examples of recent papers

arXiv ...v1 [cs.LG] 13 May 2025

Towards Foundation Models for Experimental Readout Systems Combining Discrete and Continuous Data

J. Giroux[1,⋆], C. Fanelli[1,

[1] William & Mary, Department

⋆ Author to whom any correspo

E-mail: jgiroux@wm.edu, cfan

8 June 2025

Abstract.
We present a (proto) Fou
operating on low-level detecto
the future Electron Ion Collide
prediction approaches—namely
lack of conditional generation—
vocabularies for discrete spati

PHYSICAL REVIEW D **111,** 092015 (2025)

Fine-tuning machine-learned particle-flow reconstruction for new detector geometries in future colliders

Farouk Mokhtar,[1,*] Joosep Pata,[2,†] Dolores Garcia,[3] Eric Wulff,[3] Mengke Zhang,[1] Michael Kagan,[4] and Javier Duarte[1]

an Diego, La Jolla, California 92093, USA
Physics and Biophysics, Tallinn 12618, Estonia
r Research (CERN), Geneva 1211, Switzerland
aboratory, Menlo Park, California 94025, USA

; accepted 1 May 2025; published 29 May 2025)

lities in a machine-learned algorithm trained for particle-flow
iders. This paper presents a cross-detector fine-tuning study,
large full simulation dataset from one detector design, and
le with a different collider and detector design. Specifically, we
(CLICdet) model for the initial training set and demonstrate
ike detector (CLD) proposed for the Future Circular Collider in
n order of magnitude less samples from the second dataset, we
tly training from scratch, across particle-level and event-level
ng transverse momentum resolution. Furthermore, we find that
erformance to the traditional rule-based particle-flow approach
00 CLD events, whereas a model trained from scratch requires
similar reconstruction performance. To our knowledge, this
tector transfer learning study for particle-flow reconstruction.
rds building large foundation models that can be fine-tuned
etries, helping to accelerate the development cycle for new
tector design and optimization using machine learning.

xperi-

series of hits as they traverse the detector. For instance, ATLAS or CMS track reconstruction algorithms rely on the

arXiv ...v1 [cs.LG] 6 Feb 2025

HEP-JEPA: A foundation model for collider physics using joint embedding predictive architecture

Jai Bardhan [1] **Radhikesh Agrawal** [*1] **Abhiram Tilak** [*1] **Cyrin Neeraj** [1] **Subhadip Mitra** [1]

Abstract

We present a transformer architecture-based foundation model for tasks at high-energy particle colliders such as the Large Hadron Collider. We train the model to classify jets using a self-supervised strategy inspired by the Joint Embedding Predictive Architecture (Assran et al., 2023). We use the JetClass dataset (Qu et al., 2022b) containing 100M jets of various known particles to pre-train the model with a data-centric approach — the model uses a fraction of the jet constituents as the context to predict the embeddings of the unseen target constituents. Our pre-trained model fares well with other datasets for standard classification benchmark tasks. We test our model on two additional downstream tasks: top tagging and differentiating light-quark jets from gluon jets. We

description could show better performance for individual tasks even in this case.

Large Language Models (LLMs), such as the generative pre-trained transformer (GPT) models (Brown et al., 2020) and BERT (Devlin et al., 2018), have shown remarkable capabilities in learning generalised language representations by pre-training on vast amounts of texts. Similarly, a foundation model (FM) in HEP could learn to encode the 'language' of particle interactions, detector responses, and physical laws, providing a versatile tool for analysing and interpreting experimental data. The success of LLMs indicates that FMs could revolutionise data-driven discovery in HEP, offering a scalable approach to uncovering new physics. The LHC has already started collecting massive amounts of data to probe new/rare processes. More data implies that the training time for models deployed at the collider will increase. Since FMs are pre-trained on huge amounts of data and

Up to us to define what a physics FM is

Large-scale ML model = **backbone**

Broad range of **downstream tasks**

Two-stage training:

- **Pre-training** on vast, diverse, and often multi-modal datasets to "learn meaningful representation"
- **Post-training** fine-tune for specific tasks with **relatively small amounts of additional data**

Train on **domain-specific** data (4-vectors, detector hits, astrophysical images…)

From:

Strategic White Paper on AI Infrastructure for Particle, Nuclear, and Astroparticle Physics: Insights from JENA and EuCAIF

https://arxiv.org/abs/2503.14192

What can we do today

- Figure out what **you** want to do?

- EuCAIF WG1 as a **service**: how can this group help **you**?

- **Bottom-up** clustering of efforts
  – Needs, interests, synergies, complementarity,…

- **Not** to *create* new work
- Give visibility to your work & make **meaningful & useful connections** between efforts
- Goal: to facilitate maximum **progress as a community**

Your vision, your needs, your worries?

Your vision, your needs?

What can we do in WG1? [bottom-up]

- Define potential **transformative impact** of FMs for our community
- Design **strategic roadmap** to foster progress as a community
- **Benchmark data set**(s), data challenge, success metric, downstream tasks
- Physics-encoding, mitigate domain shifts, explainability, scalability, language & symbolic encoding
- Facilitate **collaboration**, network, training, topical meetings, **funding**

What are your worries

IP: this is **my** idea, **my** paper,...

Very understandable: it's not you, it's the system!

This system is not working well, it gives the wrong incentives

Apparent dichotomy of loss functions:

- best for one's career ≠ best for advancing science

Our pitch

- Change of mentality: lone warrior of your *tiny* goal → bring value to ambitious project with long-term vision
  - More satisfying & much higher reach of your research
  - **Ultimately, this pays off [visibility, credit]**
- Define work packages & interfaces & metrics
- **Living** strategy – update & amend as needed
- Topical meetings to make progress on “work packages”
- Strategy meetings of all members to keep the ultimate goal in view + everyone is aware of *big picture*