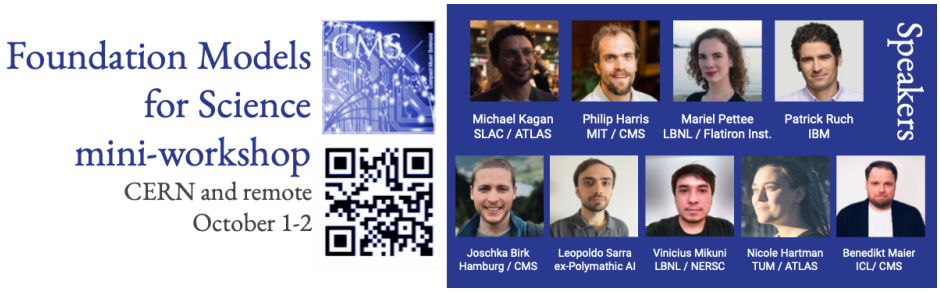
Foundation Models for Science Mini Workshop



Report of Contributions

https://indico.cern.ch/e/fm4s

Contribution ID: **1** Type: **not specified**

OmniLearn: A Method to Simultaneously Facilitate All Jet Physics Tasks

Wednesday 2 October 2024 17:30 (1 hour)

Machine learning has become an essential tool in jet physics. Due to their complex, high-dimensional nature, jets can be explored holistically by neural networks in ways that are not possible manually. However, innovations in all areas of jet physics are proceeding in parallel. We show that specially constructed machine learning models trained for a specific jet classification task can improve the accuracy, precision, or speed of all other jet physics tasks. This is demonstrated by training on a particular multiclass generation and classification task and then using the learned representation for different generation and classification tasks, for datasets with a different (full) detector simulation, for jets from a different collision system (pp versus ep), for generative models, for likelihood ratio estimation, and for anomaly detection. Our OmniLearn approach is thus a foundation model and is made publicly available for use in any area where state-of-the-art precision is required for analyses involving jets and their substructure.

Presenter: MIKUNI, Vinicius Massami (Lawrence Berkeley National Lab. (US))

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **2** Type: **not specified**

Foundation models for HEP

Tuesday 1 October 2024 15:45 (45 minutes)

Presenter: HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US))

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **3** Type: **not specified**

Introduction

Tuesday 1 October 2024 14:00 (10 minutes)

Author: KANSAL, Raghav (Univ. of California San Diego (US))

Co-author: Dr GROSSO, Gaia (IAIFI, MIT)

Presenters: Dr GROSSO, Gaia (IAIFI, MIT); KANSAL, Raghav (Univ. of California San Diego (US))

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **4** Type: **not specified**

Introduction to Day 2

Wednesday 2 October 2024 14:00 (15 minutes)

Presenters: Dr GROSSO, Gaia (IAIFI, MIT); KANSAL, Raghav (Univ. of California San Diego (US))

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **5** Type: **not specified**

OmniJet-α: The first cross-task foundation model for particle physics

Wednesday 2 October 2024 16:00 (45 minutes)

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. The successful development of such general-purpose models for physics data would be a major breakthrough as they could improve the achievable physics performance while at the same time drastically reduce the required amount of training time and data.
We report significant progress on this challenge on several fronts. First, a comprehensive set of evaluation methods is introduced to judge the quality of an encoding from physics data into a representation suitable for the autoregressive generation of particle jets with transformer architectures. These measures motivate the choice of a higher-fidelity tokenization compared to previous works. Finally, we demonstrate transfer learning between an unsupervised problem (jet generation) and a classic supervised task (jet tagging) with our new OmniJet-α model. This is the first successful transfer between two different and actively studied classes of tasks and constitutes a major step in the building of foundation models for particle physics.

Presenter: BIRK, Joschka (Hamburg University (DE))

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **6** Type: **not specified**

Invisible Cities: Towards a multi-modal era of fundamental physics research

Tuesday 1 October 2024 16:30 (1 hour)

To achieve some of the biggest physics discoveries in the last decade – e.g. finding definitive evidence of the Higgs boson, gravitational waves, and black holes – physicists had to radically re-imagine the paradigm of working in small teams and instead construct large-scale experimental collaborations of hundreds or even thousands of scientists. The recent success of foundation models in various domains begs the question: could our scientific conventions yet again be restricting our access to major discoveries? In this talk, I propose that an interdisciplinary, multi-modal approach to fundamental physics research will be critical to finally answering the grand scientific mysteries about our Universe that have thus far eluded our usual strategies. In particular, I will present some recent work from my team at Polymathic AI exploring how we might form our first scientific foundation models, and I'll also share my perspectives on how we should strive to shape such models to reflect our highest priorities as scientists.

Presenter: PETTEE, Mariel (Lawrence Berkeley National Lab. (US))

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **7** Type: **not specified**

Towards Foundation Models in HEP with Self-Supervised Learning

Tuesday 1 October 2024 14:10 (50 minutes)

Can Foundation Models, which rely on massive parameter counts, data sets, and compute, and have proven extremely powerful in computer vision and natural language systems, be built for High Energy Physics? To do so, several challenges must be addressed, including understanding how the training strategies, which are often data-type specific, can be developed for HEP data. In this talk, we will discuss our first steps towards building HEP foundation models using Self-Supervised training methods, such as masking strategies. We will also explore how pre-trained models can encode general knowledge of high utility when adapted for a variety of tasks.

Presenter: KAGAN, Michael (SLAC National Accelerator Laboratory (US))

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **8** Type: **not specified**

Finetuning Foundation Models for Joint Analysis Optimization

Wednesday 2 October 2024 15:00 (45 minutes)

This talk highlights the significant gains in performance and data efficiency that can be achieved in HEP by moving away from the standard paradigm of separate reconstruction and analysis optimization. We introduce the key idea of fine-tuning a foundation model as a generalization of choosing working points in a physics analysis. The sensitivity gains achievable from end-to-end pipelines are demonstrated in an example with a heavy resonance decaying to a di-Higgs to four b-quark system from the Open CMS dataset with a ParT backbone taken as the foundation model.

Presenter: HARTMAN, Nicole Michelle (TUM (DE))

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **9** Type: **not specified**

Re-simulation-based self-supervised learning

Wednesday 2 October 2024 14:15 (45 minutes)

Self-Supervised Learning (SSL) is at the core of training modern large machine learning models, providing a scheme for learning powerful representations that can be used in a variety of downstream tasks. We propose RS3L ("Re-simulation-based self-supervised representation learning"), a novel simulation-based SSL strategy that employs a method of re-simulation to drive data augmentation for contrastive learning in the physical sciences, particularly, in fields that rely on stochastic simulators. By intervening in the middle of the simulation process and re-running simulation components downstream of the intervention, we generate multiple realizations of an event, thus producing a set of augmentations covering all physics-driven variations available in the simulator. Using experiments from high-energy physics, we explore how this strategy may enable the development of a foundation model; we show how RS3L pre-training enables powerful performance in downstream tasks such as discrimination of a variety of objects and uncertainty mitigation.

Presenter: MAIER, Benedikt (Imperial College (GB))

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **10** Type: **not specified**

Foundation Models for Astrophysics and Representation Learning

Wednesday 2 October 2024 16:45 (45 minutes)

Given the remarkable success of foundation models in language and vision, it is worth exploring whether a similar approach can be applied to scientific domains. These models have the potential to improve computational efficiency, generalize better to low-data regimes, and significantly amortize training costs. However, many questions remain open regarding architectures, data selection, preprocessing techniques, and evaluation strategies. In this talk, I will focus on two foundation model approaches for astrophysics. The first, AstroCLIP, uses contrastive learning to build a shared latent space by aligning two models representing different views of the same phenomenon. The second, AstroOBS (work in progress), uses latent masked modeling to construct a unified multimodal representation capable of integrating diverse observational data. I will also discuss the importance of representation learning and briefly mention our ongoing work on time-series modeling as a starting point for future modality encoders.

Presenter: SARRA, Leopoldo

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **11** Type: **not specified**

Foundation Models for Scientific Discovery in the Lab of the Future

Tuesday 1 October 2024 17:30 (1 hour)

We reflect on the changes imparted by foundation models (FMs) to data-driven exploration and discovery in traditional scientific fields such as chemistry, spanning from hypothesis generation to experimental planning and validation. The emergence of multi-modal FMs also opens up new opportunities for data capture during manual experimentation, for which we present examples and lessons learned. Finally, we discuss recent work on how generative FMs can help accelerate simulation of stochastic particle shower events.

Presenter: RUCH, Patrick

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **12** Type: **not specified**

Talk

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **13** Type: **not specified**

Summary and Roundtable

Wednesday 2 October 2024 18:30 (30 minutes)

Presenters: Dr GROSSO, Gaia (IAIFI, MIT); KANSAL, Raghav (Univ. of California San Diego (US))

Session Classification: Foundation Model MiniWorkshop

Contribution ID: **14** Type: **not specified**

Automatic Differentiation in RooFit and Community Needs

Tuesday 1 October 2024 15:00 (45 minutes)

Presenters: REMBSER, Jonas (CERN); VASILEV, Vassil (Princeton University (US)); VASSILEV, Vassil (CERN)

Session Classification: CMS/RooFit team discussion