EUROPEAN AI FOR
FUNDAMENTAL PHYSICS
CONFERENCE
EuCAIFCon 2025

Contribution ID: **123**

Type: **Parallel talk**

Challenges and Innovations in Learning from Heterogeneous Data in Fundamental Physics

Learning from heterogeneous data is one of the major challenges for AI in the coming years, particularly for what are called ‘foundation models’. In fundamental physics, the heterogeneity of data can come from the instruments used to acquire them (the subsystems of large detectors in particle physics or the crossing of modalities in multi-messenger, multi-instrument systems in astrophysics, for example) or from the data themselves when the signal is a superposition of many sources of different nature (as is the case in the forthcoming LISA detector for gravitational waves that will observe the superposition of signals from a large, a priori unknown number of sources of different types). Models capable of learning from these heterogeneous data will need to be able to integrate a common representation of these data in shared latent spaces. We discuss the problems posed by such learning and why it is crucial to solve them for future AI-assisted research in fundamental physics. We provide an overview of the significant work in this area, in particular different integration architectures considered and techniques for latent alignment in AI in general and in fundamental physics in particular. We will cite current projects and summarise the main contributions made and key messages identified during the workshop “Heterogeneous Data and Large Representation Models in Science” [1] held as part of the “Artificial Intelligence for the two infinites [2]” initiative of the AISSAI centre [3] of the French National Centre for Scientific Research (CNRS). The results obtained, key questions and challenges for enabling these models to develop and become fully operational in the near future will be discussed.

[1] https://indico.in2p3.fr/e/AISSAI-TLS

[2] the infinitely small (particle physics) and the infinitely large (cosmology)

[3] https://aissai.cnrs.fr/en/

AI keywords

heterogeneous data; latent representations, and their alignment; model architectures; training techniques

Primary authors: BISCARAT, Catherine (L2I Toulouse, CNRS/IN2P3, Université de Toulouse); STARK, Jan (L2I Toulouse, CNRS/IN2P3, Université de Toulouse); CAILLOU, Sylvain (L2I Toulouse, CNRS/IN2P3, Université de Toulouse)

Presenter: CAILLOU, Sylvain (L2I Toulouse, CNRS/IN2P3, Université de Toulouse)

Track Classification: Datasets & Ethics