Towards (AI/ML) Sample Generation

A brief intro & discussion

Karolos Potamianos

Polarized Perspectives: Tagging and Learning in the SM
February 20, 2025

Towards Sample Generation for AI/ML/DL

From the COMETA proposal …

WG2: Technological innovation in data analysis

D2.1 Preparation of material contributing to ML research: survey of existing ML-based tools in high energy physics, Action-specific public datasets with documentation

Objective is to facilitate the inclusion of “AI” practitioners in our workflows, without them requiring to be physicists or have a deep physics understanding, and hence reap benefits from advances in the field.

Existing Datasets

Several datasets have been produced in HEP context.

The IML (Inter-Experimental LHC ML) WG maintains a list of public datasets used for ML studies at the LHC: https://iml.web.cern.ch/public-datasets, covering three main areas:

- **Simplified datasets for benchmarking**
  - Top tagging without heavy flavour & pileup: arXiv:1707.08966 [Doc]
  - Jet substructure: arXiv:1607.08633 [MLPhysics]
  - Flavour tagging without pileup: arXiv:1603.09349 [MLPhysics]
- **Datasets for developing simulation**: jet images, LAGAN/CaloGan
- **Challenge datasets:** Kaggle, IML challenges (2017 & 2018)

And there's the Open Data from ATLAS and CMS

Goal: aim for survey of existing datasets (and tools) and their suitability for the purpose of COMETA's goals, e.g. (polarised) vector boson tagging.

Existing Datasets (examples)

Jet substructure: Data set consists of 10M training examples, 5M test examples. There are two sets of features: the l**ow-level calorimeter images**, and the **high-level derived features**. There are also two versions of the datasets, one **with pile-up and one without pile-up**.

Flavour tagging without pileup: 11.5M samples. Contains variables related to **jet kinematics, tracks, vertex and high level features**. Each track contains 20 variables (+ # of tracks). Each vertex contains 8 variables (+# of vertices). There are 14 high level variables. Labels for light, charm and bottom jets.

Towards a Sample

Preparing a little survey (see questions below) about this dataset: https://forms.gle/HNFrcrU6n8X6Raq36

Aim is to collect experience with this kind of dataset and wishes for a COMETA-specific dataset

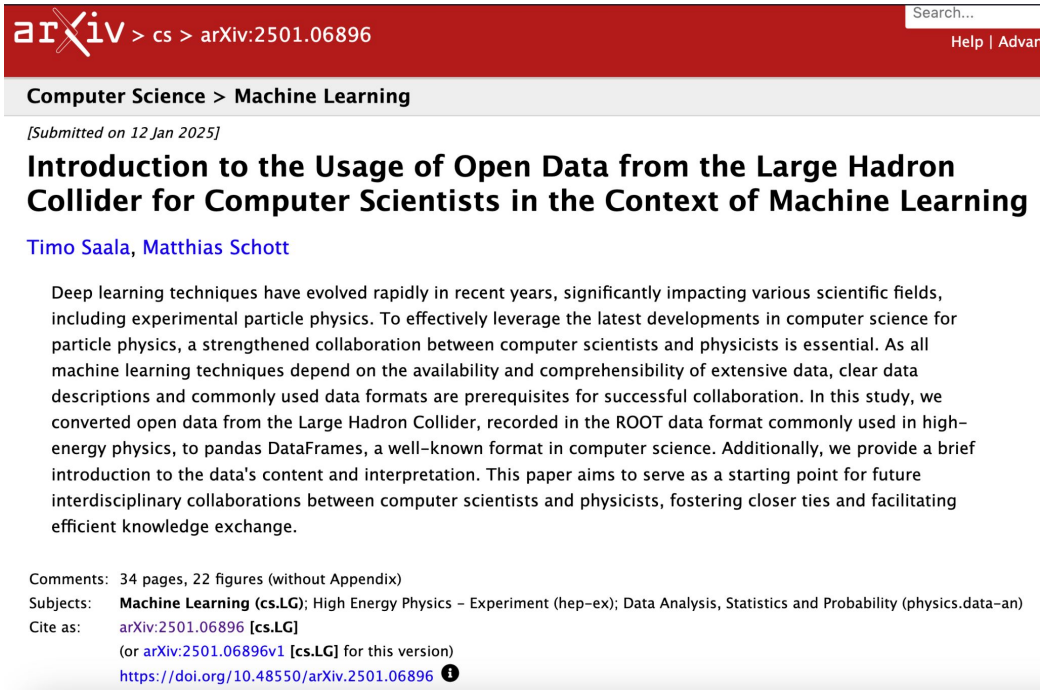
Goal is to share it within COMETA and to interested people

Please have a look and send me comments/feedback!

Aim to circulate the survey in the coming days.

Suitability for AI/ML Practitioners

ML Practitioners often don't know enough/at all about the physics behind our datasets … **we should help them understand**



arXiv > cs > arXiv:2501.06896

Search...

Help | Advar

Computer Science > Machine Learning

[Submitted on 12 Jan 2025]

Introduction to the Usage of Open Data from the Large Hadron Collider for Computer Scientists in the Context of Machine Learning

Timo Saala, Matthias Schott

Deep learning techniques have evolved rapidly in recent years, significantly impacting various scientific fields, including experimental particle physics. To effectively leverage the latest developments in computer science for particle physics, a strengthened collaboration between computer scientists and physicists is essential. As all machine learning techniques depend on the availability and comprehensibility of extensive data, clear data descriptions and commonly used data formats are prerequisites for successful collaboration. In this study, we converted open data from the Large Hadron Collider, recorded in the ROOT data format commonly used in high-energy physics, to pandas DataFrames, a well-known format in computer science. Additionally, we provide a brief introduction to the data's content and interpretation. This paper aims to serve as a starting point for future interdisciplinary collaborations between computer scientists and physicists, fostering closer ties and facilitating efficient knowledge exchange.

Comments: 34 pages, 22 figures (without Appendix)
Subjects: **Machine Learning (cs.LG)**; High Energy Physics - Experiment (hep-ex); Data Analysis, Statistics and Probability (physics.data-an)
Cite as: arXiv:2501.06896 **[cs.LG]**
(or arXiv:2501.06896v1 **[cs.LG]** for this version)
https://doi.org/10.48550/arXiv.2501.06896

Suitability for AI/ML Practitioners

ML Practitioners often don't know enough/at all about the physics behind our datasets …

On the other hand, they don't care so much about the underlying aspects … we might as well provide them data with unlabelled features/columns (was clearly expressed during the October 2024 COMETA Workshop in Amsterdam)

We should keep these seemingly contradictory aspects in mind and find a middle ground that works in mosts cases

Note: there's also the educational aspect that also may enter the picture, to motivate students to pursue research in HEP

What are you ? *

- AI/ML/DL practitioner (computer scientist)
- Particle physicist with no experience in AI/ML/DL
- Particle physicist with some experience in AI/ML/DL
- Other: _____

Have you previously used any of the HEP datasets specifically targeting AI/ML/DL * applications ?

- Yes
- No

If so, could you please list which datasets you've used and how useful you found them ? (Scope, usage, etc.)

Your answer

Would you be interested in using a dataset targeting polarised vector boson * tagging ?

- Yes
- No

What features would you like to see in a COMETA dataset for polarised vector boson tagging ? Why ?

Your answer

What level of documentation should there be with this dataset ? *

- Every variable should be explained thoroughly
- I don't care: I just want the features; don't even need to know what you call them

Is there a technology stack that you'd recommend for this COMETA dataset ? Please try to explain what makes the tools you suggest appealing for the purpose.

Your answer

Would you be interested in taking part in a (HANDS-ON) Hackathon around DESIGNING such a new dataset on polarised vector boson tagging ?

- Yes
- No
- Maybe

Would you be interested in taking part in a Hackathon around USING such a new dataset on polarised vector boson tagging ? *

- ◯ Yes
- ◯ No
- ◯ Maybe

Would you be interested in taking part in an Online Challenge (Kaggle, etc.) around such a new dataset on polarised vector boson tagging ? *

- ◯ Yes
- ◯ No
- ◯ Maybe

How much material (tutorials, examples, etc.) should accompany the dataset ? *

- ☐ Nothing: if you need to show examples, it looses it's appeal to simplicity
- ☐ A little bit: one should know the basics of the data layout and how to get started
- ☐ Lots of details are needed for students and other newcomers to understand what they're doing and not have everything be a black box!
- ☐ Other: _____

Would you use such a dataset for students to learn about AI/ML/DL methods in HEP ?

- ◯ Yes, any dataset is good
- ◯ No, it's too specific
- ◯ Maybe

How important is it to you to have a realistic detector simulation around this dataset ? *

	1	2	3	4	5	
Not important at all	◯	◯	◯	◯	◯	Essential

What do detector aspects you consider important to have modelled realistically, and can this be achieved (in your opinion) with publicly available tools (i.e. without the need to run through the experiments' detector simulation suites) ? *

Your answer

Should the dataset include HL-LHC detector features ? *

- ◯ Yes
- ◯ No
- ◯ It would be useful, but not essential

Should the dataset include features specific to a certain LHC experiment (e.g. timing layers) ? *

- ◯ Yes, planned or hypothetical upgrades, the more the merrier
- ◯ It would be useful, but not essential
- ◯ No, we should only stick to what can be applied generally

How important is it to have lots of data ? *

	1	2	3	4	5	
Not important at all	◯	◯	◯	◯	◯	Essential

What sample size (events) would be a minimum, and ideal, in your opinion ? *

Your answer

Would it be important to have a data generator ? i.e. the ability to generate a almost arbitrarily large dataset on demand ?

	1	2	3	4	5	
Not important at all	◯	◯	◯	◯	◯	Essential

What possible pitfalls do you see in preparing such a dataset ?

Your answer

Any final thoughts or comments ?

Your answer

COMETA AI/ML Hackathon

Would like to work towards a COMETA AI/ML Hackathon next year, following the release of a dataset for AI/ML, covering, at least the topic of **polarised vector boson tagging** in hadronic decays at the LHC (to be agreed/discussed).

Personal suggestion would be to have it open, and part of a broader challenge (always a good motivator for some).

We might also want to have a (less open) hackathon or STSM to design/prepare the dataset in question.

Discussion to follow at the end of today's morning session

(Incomplete) ToDo/Shopping List

- Define processes and phase space
- Determine appropriate tools: generators, detector simulation, etc.
  - How future-oriented do we want o be ?
- Determine nature of features to be reported
- Prepare documentation with code & tutorials

Extras (if wished and possible):

- Define meaningful data challenge conditions
- Prepare hackathons (for generating dataset and usage)
- Produce recipes for generating more data or even new datasets (e.g. different generator or detector conditions)

Follow-up (and regular) meetings w/ interested parties: reach out!!

Challenges

Ensure to clearly define our objectives:

- Do we just want to achieve single object (vector boson) classification or do we want to label events as a whole (e.g. LL, LT, TT in the case of VVjj)
- Determine what detector-level information is essential to include (and what is optional), and ensure that simulation is doing the job we need it (we don't want models to pick up on features of the simulation that aren't present in a real-world scenario)

Start with a simple dataset which can fill a single/simple purpose and iterate towards refined solution in the remaining time of the Action and beyond

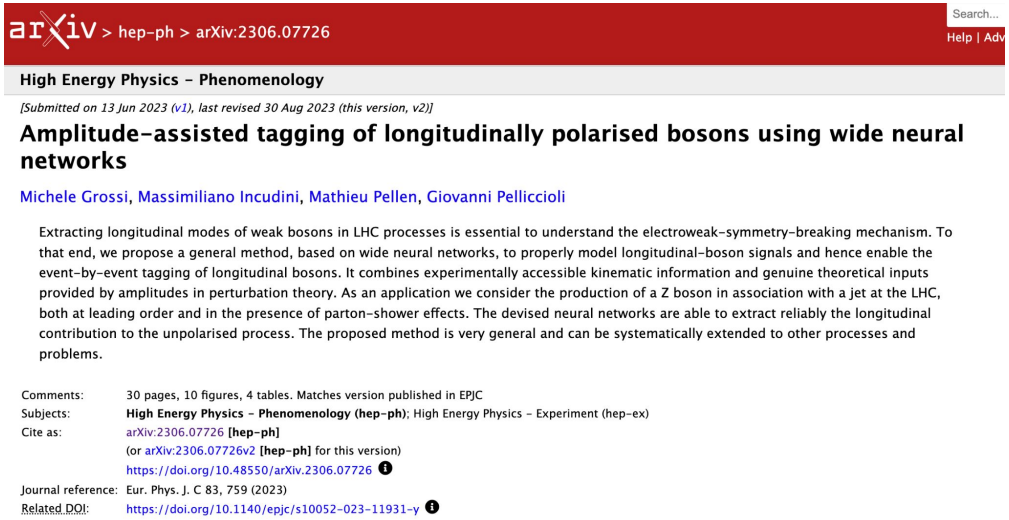
- Much better to have something rather than a perfect "to be ready soon"

Additional Thoughts

Initial proposal is on experimental aspect of tagging/enhancing polarisation

But there are other aspects for which AI/ML/DL can be used as well, e.g.

Do we need/want datasets that cover these aspects too ?



arXiv > hep-ph > arXiv:2306.07726

Search... Help | Adv

High Energy Physics – Phenomenology

[Submitted on 13 Jun 2023 (v1), last revised 30 Aug 2023 (this version, v2)]

Amplitude-assisted tagging of longitudinally polarised bosons using wide neural networks

Michele Grossi, Massimiliano Incudini, Mathieu Pellen, Giovanni Pelliccioli

Extracting longitudinal modes of weak bosons in LHC processes is essential to understand the electroweak-symmetry-breaking mechanism. To that end, we propose a general method, based on wide neural networks, to properly model longitudinal-boson signals and hence enable the event-by-event tagging of longitudinal bosons. It combines experimentally accessible kinematic information and genuine theoretical inputs provided by amplitudes in perturbation theory. As an application we consider the production of a Z boson in association with a jet at the LHC, both at leading order and in the presence of parton-shower effects. The devised neural networks are able to extract reliably the longitudinal contribution to the unpolarised process. The proposed method is very general and can be systematically extended to other processes and problems.

Comments:	30 pages, 10 figures, 4 tables. Matches version published in EPJC
Subjects:	**High Energy Physics – Phenomenology (hep-ph)**; High Energy Physics – Experiment (hep-ex)
Cite as:	arXiv:2306.07726 **[hep-ph]**
	(or arXiv:2306.07726v2 **[hep-ph]** for this version)
	https://doi.org/10.48550/arXiv.2306.07726
Journal reference:	Eur. Phys. J. C 83, 759 (2023)
Related DOI:	https://doi.org/10.1140/epjc/s10052-023-11931-y

Talk by M. Pellen at Toulouse Workshop on Polarisation

Discussion

Document