

Umami: A Python toolkit for jet flavour tagging

Jackson Barr¹, Joschka Birk¹³, Maxence Draguet³, Stefano Franchellucci⁴, Alexander Froch⁴, Philipp Gadow⁵, Daniel Hay Guest⁶, Manuel Guth⁴, Nicole Michelle Hartman⁷, Michael Kagan⁸, Osama Karkout⁹, Dmitrii Kobylanskii¹⁰, Ivan Oleksiyuk⁴, Nikita Ivvan Pond¹, Frederic Renner¹¹, Sebastien Rettie⁵, Victor Hugo Ruelas Rivera⁶, Tomke Schröer⁴, Martino Tanasini¹², Samuel Van Stroud¹, and Janik Von Ahnen¹¹

¹ University College London, United Kingdom ² Albert-Ludwigs-Universität Freiburg, Germany ³ University of Oxford, United Kingdom ⁴ Université de Genève, Switzerland ⁵ European Laboratory for Particle Physics CERN, Switzerland ⁶ Humboldt University Berlin, Germany ⁷ Technical University of Munich, Germany ⁸ SLAC National Accelerator Laboratory, United States of America ⁹ Nikhef National Institute for Subatomic Physics and University of Amsterdam, Netherlands ¹⁰ Department of Particle Physics and Astrophysics, Weizmann Institute of Science, Israel ¹¹ Deutsches Elektronen-Synchrotron DESY, Germany ¹² Stony Brook University, United States of America ¹³ University of Hamburg, Germany

DOI: [10.21105/joss.05833](https://doi.org/10.21105/joss.05833)

Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Editor: [Matthew Feickert](#) ↗ 

Reviewers:

- [@jpata](#)
- [@hqcms](#)

Submitted: 22 August 2023

Published: 02 October 2024

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Summary

Flavour-tagging, the identification of collimated sprays of particles (“jets”) originating from bottom and charm quarks, is a critically important technique in the data analysis of the ATLAS experiment ([ATLAS Collaboration, 2008](#)) at the Large Hadron Collider ([Evans & Bryant, 2008](#)). It is applied in precision measurements of the Standard Model of particle physics, which is the theory describing fundamental forces and classifying all known elementary particles, as well as in searches for yet unknown phenomena. The long lifetime, high mass, and large decay multiplicity of hadrons containing bottom and charm quarks provide distinct signatures in particle detectors which can be exploited by machine learning algorithms. Excellent knowledge of the detector and the physics processes at hand enables simulations to provide a high-quality training dataset representative of recorded ATLAS data. The Umami software toolkit provides a unified data pipeline, definition of the algorithms, training and performance evaluation with a high degree of automation.

Statement of need

Umami is a Python toolkit for training and evaluating machine learning algorithms used in high energy physics for jet flavour tagging. The creation and training of production-grade machine learning models is supported by the TensorFlow and keras packages. The training datasets feature highly imbalanced distributions among the target classes and input features of vastly different magnitude. Consequentially, the preprocessing of the training data requires resampling to provide balanced datasets and transformation of the input features by scaling and shifting.

Umami provides a class-based and user-friendly interface with yaml configuration files to steer the data preprocessing and the training of deep neural networks. It is available as a Python application and is also distributed via Linux container images. Umami was designed to be used by researchers in the ATLAS collaboration and is open to be applied in a more general context.

Related work

The application of machine learning in high energy physics, particularly for the classification of jets, is a common and critically important technique (Cagnotta et al., 2022; Guest et al., 2018). In contrast to previous efforts in jet flavour tagging (ATLAS Collaboration, 2023; Bols et al., 2020), the current state-of-the-art algorithms (Qu et al., 2022a) rely on specialised toolkits, such as the Weaver framework (Qu & Gouskos, 2020). These toolkits enable the design of algorithms by taking care of input processing, steering the training on large datasets and providing performance metrics. Umami provides the required functionality to define, train and evaluate the algorithms used in ATLAS data analysis.

Software description

The Umami toolkit provides an integrated workflow including input data preprocessing, algorithm training, and performance evaluation.

The algorithms are trained on simulated physics processes which provide jets originating from bottom and charm quarks, as well as the background processes which produce jets originating from other sources, such as light-flavour quarks, gluons, or hadronically decaying tau leptons. The input features to the algorithm provide discrimination between the processes. A more detailed discussion of the input features for jet flavour tagging is provided in Ref. (ATLAS Collaboration, 2023). The preprocessing in Umami addresses several challenges provided both by the nature of the training datasets and the input features.

Using Umami is not limited to jet flavour tagging but provides support for a broad range of applications. The preprocessing capabilities are demonstrated with simulated physics processes from the JetClass dataset (Qu et al., 2022b) to distinguish jets originating from Higgs boson decays from jets originating from top quark decays. This represents a similar but slightly different use of machine learning algorithms for jet classification. The software is flexible enough to address this task with only minimal modifications in configuration files.

Figure 1 shows the absolute value of the pseudorapidity η of the jets from Higgs boson decays to b-quarks, Higgs boson decays to c-quarks, and to top quarks before and after the re-sampling step in the preprocessing. The total number of events in each class is equalised and the shape differences between classes are removed by the resampling.

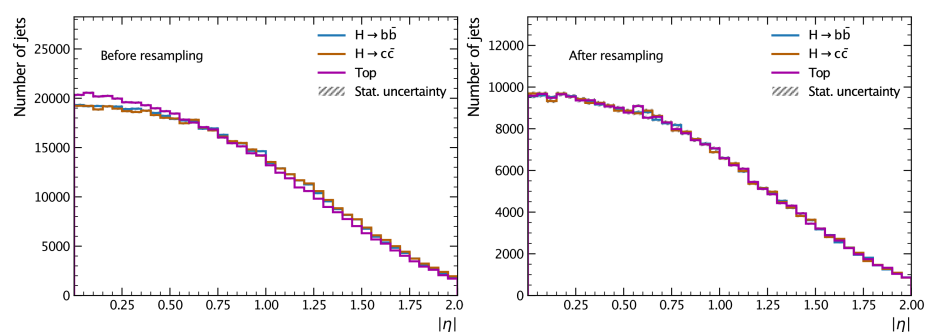


Figure 1: Distributions of the pseudorapidity η of jets from Higgs boson decays to b-quarks ($H \rightarrow b\bar{b}$ -jets), Higgs boson decays to c-quarks ($H \rightarrow c\bar{c}$ -jets), and to top quarks (Top) before and after resampling.

Different architectures of neural networks, including Deep Multi-Layer-Perceptrons (LeCun et al., 2015) and Deep Sets, are supported in Umami for definition with configuration files. The training is performed with keras using the TensorFlow back-end and the Adam optimiser (Kingma & Ba, 2015), supporting the use of GPU resources to shorten the required time to train the networks by an order of magnitude.

The performance of the chosen model can be evaluated in publication-grade plots, which are steered with configuration files. The plots are created using the matplotlib (Hunter, 2007) and puma (Birk et al., 2023) Python libraries.

Conclusions and future work

We present Umami, a Python toolkit designed for training machine learning algorithms for jet flavour tagging. Its strong point is that it unifies the steps for preprocessing of the training samples, the training and validation of the resulting models in a mostly automated and user-friendly way. The software is widely used within the ATLAS collaboration to design neural networks which classify jets originating from bottom quarks, charm quarks or other sources. While the software is customised for this application, it is not limited to it. It is straightforward to modify the expected input features and target classes, such that the general preprocessing and training capabilities can be used in wider contexts. The identification of charged particle tracks or classification of hadronically decaying tau leptons present relevant and adequate possible use-cases.

Acknowledgements

This work was done as part of the offline software research and development programme of the ATLAS Collaboration, and we thank the collaboration for its support and cooperation.

References

- ATLAS Collaboration. (2008). The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3, S08003. <https://doi.org/10.1088/1748-0221/3/08/S08003>
- ATLAS Collaboration. (2023). ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset. *Eur. Phys. J. C*, 83, 681. <https://doi.org/10.1140/epjc/s10052-023-11699-1>
- Birk, J., Froch, A., VS, S., Guth, M., Gadow, P., Schröer, T., Kobylanskii, D., Rettie, S., & Streblor, T. (2023). *Umami-hep/puma: v0.2.4*. Zenodo. <https://doi.org/10.5281/ZENODO.7806395>
- Bols, E., Kieseler, J., Verzetti, M., Stoye, M., & Stakia, A. (2020). Jet flavour classification using DeepJet. *Journal of Instrumentation*, 15(12), P12012–P12012. <https://doi.org/10.1088/1748-0221/15/12/p12012>
- Cagnotta, A., Carnevali, F., & Iorio, A. D. (2022). Machine learning applications for jet tagging in the CMS experiment. *Applied Sciences*, 12(20), 10574. <https://doi.org/10.3390/app122010574>
- Evans, L., & Bryant, P. (2008). LHC Machine. *JINST*, 3, S08001. <https://doi.org/10.1088/1748-0221/3/08/S08001>
- Guest, D., Cranmer, K., & Whiteson, D. (2018). Deep learning and its application to LHC physics. *Annual Review of Nuclear and Particle Science*, 68(1), 161–181. <https://doi.org/10.1146/annurev-nucl-101917-021019>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, CA, USA, may 7-9, 2015, conference track proceedings*. <https://doi.org/10.48550/arXiv.1412.6980>

- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- Qu, H., & Gouskos, L. (2020). Jet tagging via particle clouds. *Physical Review D*, 101(5). <https://doi.org/10.1103/physrevd.101.056019>
- Qu, H., Li, C., & Qian, S. (2022a). Particle transformer for jet tagging. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, & S. Sabato (Eds.), *Proceedings of the 39th international conference on machine learning* (Vol. 162, pp. 18281–18292). PMLR. <https://doi.org/10.48550/arXiv.2202.03772>
- Qu, H., Li, C., & Qian, S. (2022b). *JetClass: A large-scale dataset for deep learning in jet physics*. Zenodo. <https://doi.org/10.5281/ZENODO.6619768>