LHCb ANALYSIS MODEL

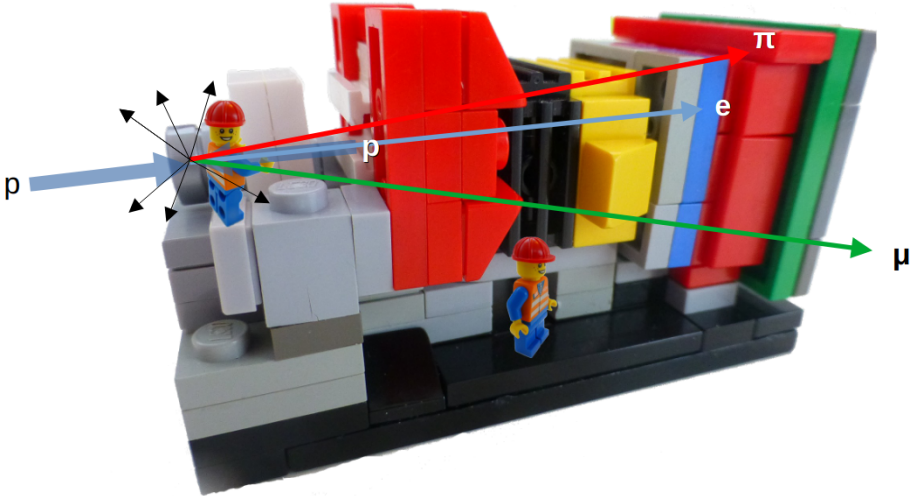
- **LHCb dataflow**
- **Analysis model**

On behalf of the LHCb collaboration

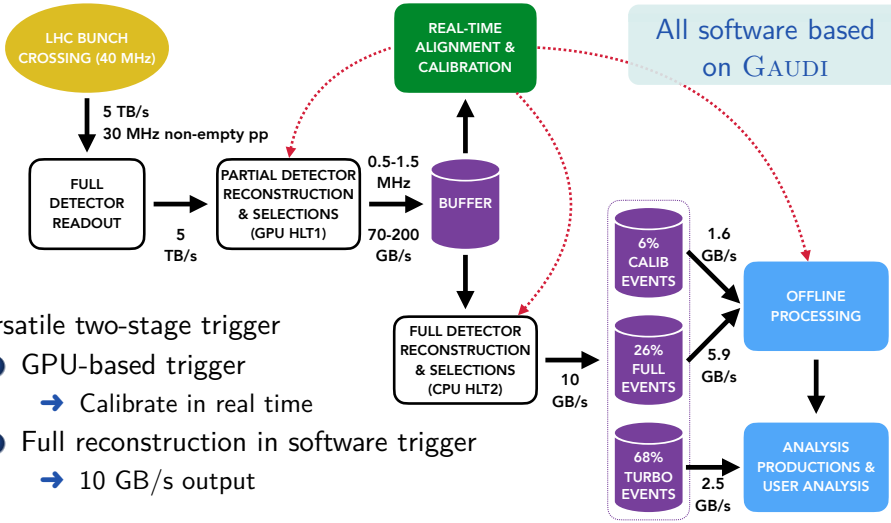
11/05/2022 — ROOT User's workshop
[indico]

Patrick Koppenburg
[@pkoppenburg] [patrick.koppenburg@nikhef.nl]

LHCb



LHCb dataflow in Run 3

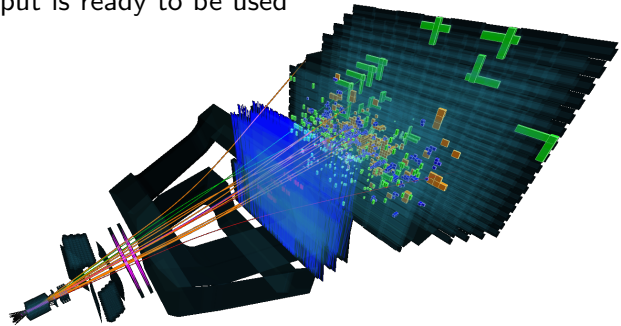


Versatile two-stage trigger

1. GPU-based trigger
   - ➜ Calibrate in real time
2. Full reconstruction in software trigger
   - ➜ 10 GB/s output

TURBO

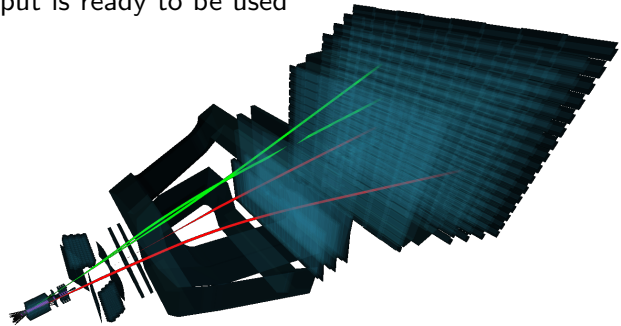
We perform a full calibration in real time. The output is ready to be used for physics.



Plenty of collision events discarded, while the interesting are kept.

TURBO

We perform a full calibration in real time. The output is ready to be used for physics.

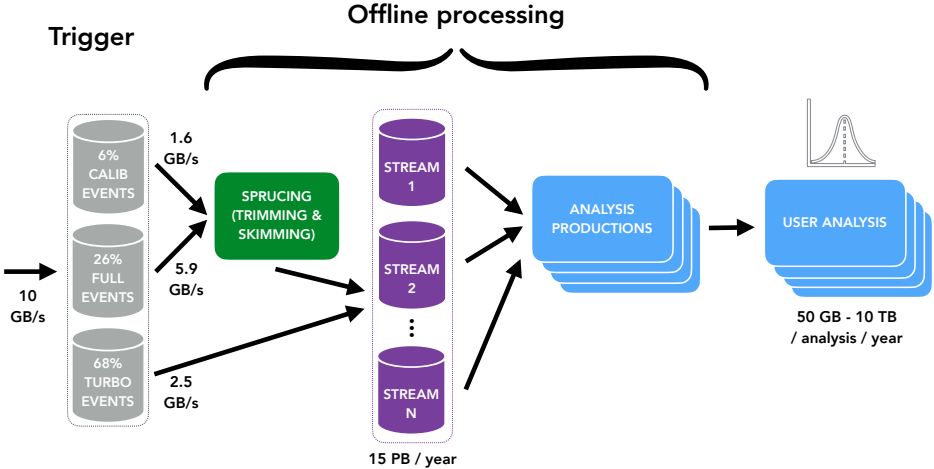


TURBO stores only the information needed for the analysis

➜ Huge savings in time and cost

[FIGURE-2020-016]

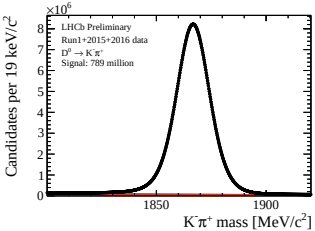
LHCb dataflow in Run 3

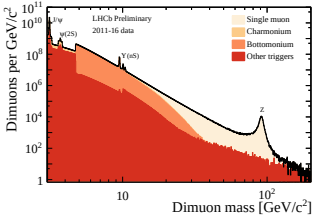


LHCb analysis model

Workflow for typical Run 3 analysis

1. Obtain Tuples from analysis production job ➜ Slide
2. Refine candidates ➜ will be $> 10^{10}$ for charm
3. Fit distributions of interest ➜ Slide
4. Determine efficiencies on simulation weighted by common calibration samples
5. Run pseudoexperiments to properly assess uncertainties ➜ Slide





Always: Make sure analysis is preserved and reproducible

- Use `conda` to get shared analysis environment ➜ Chris Burr

[6]

Analysis productions

LHCb Analysis Productions

Log out (pkoppenb)

- Home
- Productions
- Pipelines
- Submissions
- Mattermost
- Documentation
- LHCbDIRAC

fest / spruce_exclusive_feb_2022

DPA

Productions / DPA / Test / spruce_exclusive_feb_2022

State	ACTIVE
Version	v0r0p3657063
Size	— (NaN% ready on disk)
Ownership	christopher.burr@cern.ch
Merge Request	https://gitlab.cern.ch/lhcb-datapkg/AnalysisProductions/-/merge_requests/219
JIRA Task	https://its.cern.ch/jira/browse/WGP-288

Tags

config	fest
eventtype	90000001

DIRAC Production Request 96783

is assigned sample ID 6929 and comprises the following transformations:

Transformation 157185

comprises 1 step - output is not kept

Step ID	154271
Application	Moore/v53r4
Options	$ANALYSIS_PRODUCTIONS_BASE/FEST/sprucing_excl.py $APPCONFIGOPTS/Persistency/Compression-ZLIB-1.py
Extra Data Packages	AnalysisProductions.v0r0p3657063 ProdConf

Transformation 157186

ACTIVE

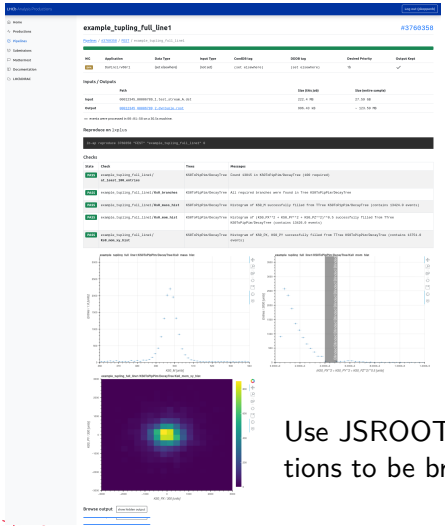
comprises 1 step - output is kept

Step ID	154272
Application	Noether/v1r4

Analysis productions web interface

[6]

Analysis productions



Analysis productions web interface

Use JSROOT for allowing the output of test productions to be browsed.

[6]

[Ferillo, LHCb-PROC-2022-004]

Ntuples for analysts "FunTuple"

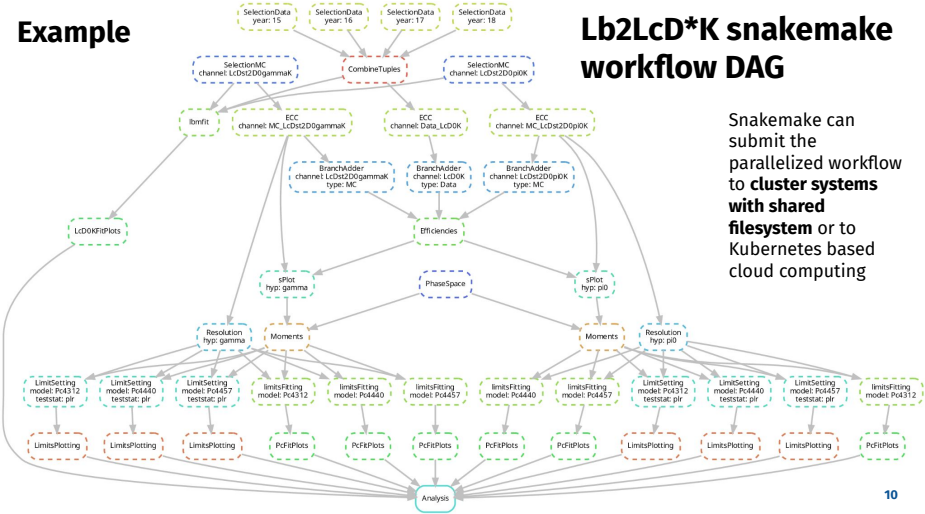
1. Access **exactly** the information that is used in Hlt2 and sprucing
   - ➜ Use same `functors`
2. Have a fine granularity
   - ➜ The user defines variables one-by-one
   - Our Tuples are essentially flat
   - Now use `TTree`. Would love to try `RNTuple`!
3. Be smart at configuration time
   - ➜ Use `Gaudi::functional`
   - ➜ Put all intelligence in PYTHON (*e.g.* classes that generate lists of functors)
   - ➜ C++ should just ship data around

[8]

LHCb analysis model

Example

Lb2LcD*K snakemake workflow DAG



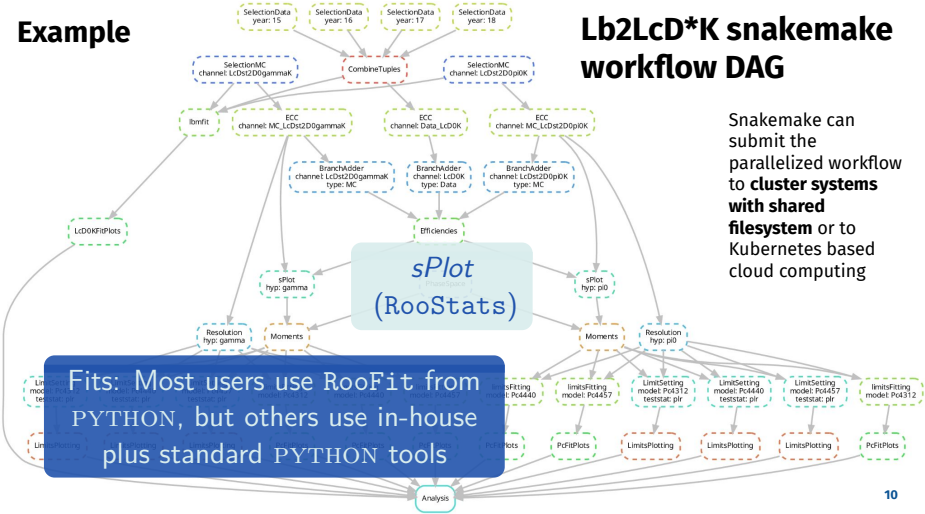
Snakemake can submit the parallelized workflow to **cluster systems with shared filesystem** or to Kubernetes based cloud computing

[6]

LHCb analysis model

Example

Lb2LcD*K snakemake workflow DAG



sPlot
(RooStats)

Fits: Most users use RooFit from PYTHON, but others use in-house plus standard PYTHON tools

Snakemake can submit the parallelized workflow to **cluster systems with shared filesystem** or to Kubernetes based cloud computing

[8]

Innovative Analysis Techniques

aka everything beyond RooFit.

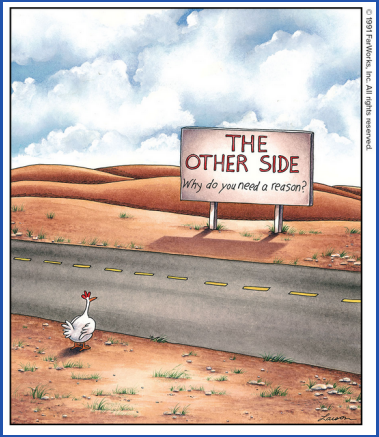
- GPU resources usage in analyses:
  - DNN for b- vs c- jet tagging, Cineca Bologna
  - Zfit and likelihood inference, Zürich
  - DNN and ultra-fast simulation Florence
  - Amplitude Analyses, Aix Marseille, CNRS/IN2P3
  - Charm Analyses, Manchester
  - Amplitude Analysis of Λ_b^0, Tsinghua
  - φ_s, Santiago
- Analyses usually done on local facilities. Availability of clusters defines implementation choices.
  - ➜ Use of the powerful Hlt1 GPU farm is considered
- Quantum Computing applications to HEP (Padova, CERN, Maastricht. . .) [Gianelle et al., arXiv:2202.13943]

[8]

Conclusion

- The LHCb detector has undergone a major upgrade
- Most analyses will be based on triggered candidates
  - ➜ Share of on- and offline software
- Analysis workflows will change to cope with the high rate
- Many developments to exploit GPUs
  - ➜ And thinking about quantum computing

Contact: [@pkoppenburg] [patrick.koppenburg@nikhef.nl]


THE
OTHER SIDE
Why do you need a reason?
© 1991 FarWorks, Inc. All rights reserved.

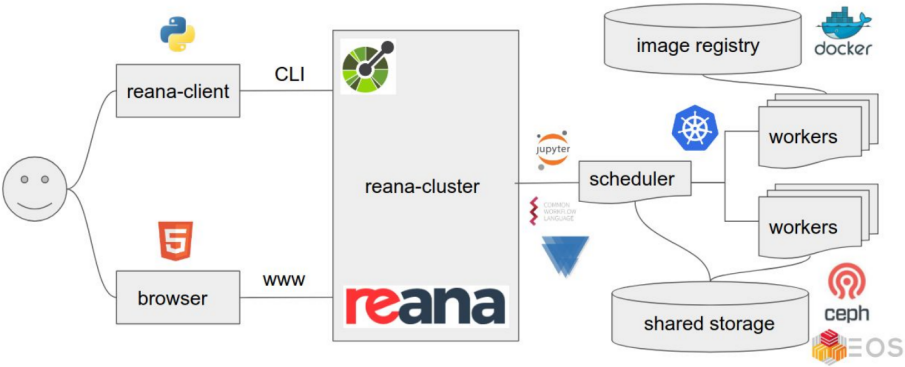
Backup

Steps to Analysis Facilities

- Several Tier-2, Tier-3, medium-large centers in US and in Europe are proposing AF prototypes developed and supported by ATLAS/CMS $\Longrightarrow$ LHCb may want to participate locally to these R&D
- LHCb will leverage the very powerful GPU HLT1 farm
- For the analysis, users want something easy to use, flexible and powerful enough… needless to say
- Given the progress in HSF, the papers submitted to Snowmass and the proposed prototypes, within LHCb starting:
  - Collect all the use cases, available and used resources, code developed, etc.
  - Identify the users needs for next round of analysis
  - Proceed with a structured activity that may led:
    - Different AF configurations depending on the site: CERN may need something dedicated to incorporate HLT1, exploit what available in country, design a dedicated configuration, ect.
    - Definition and identification of mandatory LHCb-specific requests

[6]

LHCb analysis model



Workflow could be deployed to REANA to check it works.

[6]

Operations Statistics Run 3

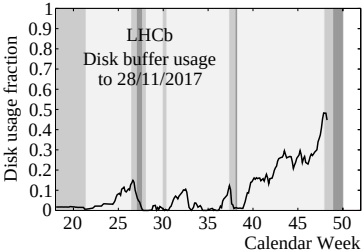
Quantity		unit	Run 1	Run 2	UTDR	2022	Tot/Avg
Peak Luminosity	$\mathcal{L}_{\text{peak}}$	$\mu\text{b}^{-1}/\text{s}$	480	447	2000		
Average Luminosity	$\mathcal{L}_{\text{avg}}$	$\mu\text{b}^{-1}/\text{s}$	298	268	2000		
Seconds of running	t	10^6 s	10.5	20.0	5.0		
Integrated luminosity	$\int\mathcal{L}\mathrm{d}t$	fb^{-1}	3.2	6.4	5.0		
Bunches	N_{b}		1320	2193	2808		
Energy	E	TeV	8	13	14		
Inelastic cross-section	σ_{inel}	mb	66	77	80		
Charged multiplicity	$\frac{\mathrm{d}N_{\text{ch}}}{\mathrm{d}\eta}$		6	6			
$b\bar{b}$ cross-section (acc.)	$\sigma_{b\bar{b}}$	μb	79	144			
pp interactions/BX	$\mu = \frac{\mathcal{L}\sigma_{\text{inel}}}{f_{\text{LHC}}N_{\text{b}}}$		1.32	0.83	5.07		
Non-empty rate	$f_{\text{LHC}}N_{\text{b}}(1-e^{-\mu})$	MHz	10.8	14.0	31.4		
Avg. MB rate	$\sigma_{\text{inel}}\mathcal{L}_{\text{avg}}$	MHz	19.7	20.7	160.0		
Peak particle flow	$\frac{\mathrm{d}N_{\text{ch}}}{\mathrm{d}\eta}\sigma_{\text{inel}}\mathcal{L}_{\text{peak}}$	10^6	189	207			
Irradiation	$\frac{\mathrm{d}N_{\text{ch}}}{\mathrm{d}\eta}\sigma_{\text{inel}}\int\mathcal{L}\mathrm{d}t$	10^{15}	1.3	3.0			
$b\bar{b}$ yield	$\sigma_{b\bar{b}}\int\mathcal{L}\mathrm{d}t$	10^9	254	922			
Output rate	λ_{HLT}	kHz	3.7	6.6	10.8		
Stored events (bkk)	$\lambda_{\text{HLT}}t$	10^9	39	133	54		
Event size	S_{ev}	kB	56	56	61		
HLT B/W	$S_{\text{ev}}\lambda_{\text{HLT}}$	MB/s	212	371	659		
Total storage	$S_{\text{ev}}\lambda_{\text{HLT}}t$	EB	2.2	7.4	3.3		

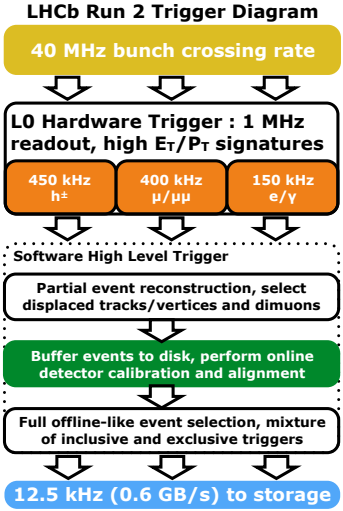
TDR values for 1 typical year of running

LHCb Trigger in Run 2

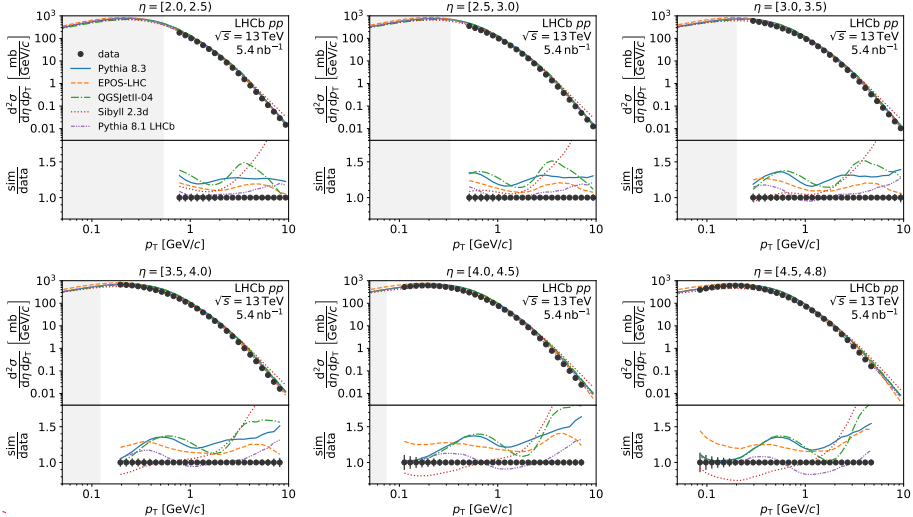
Events are buffered on disk (10 PB) while calibrations are being run.

- ➜ Offline-quality trigger objects available for analysis.
- Disk ➜ more CPU. The full reconstruction can also be run during LHC downtime.





Charged multiplicity at $\sqrt{s} = 13$ TeV



Charged multiplicity at $\sqrt{s} = 13$ TeV

[28] R. Brun and F. Rademakers, *ROOT – An object oriented data analysis framework*, Nucl. Instrum. Meth. **A389** (1997) 81.

[29] G. Corti *et al.*, *Software for the LHCb experiment*, IEEE Trans. Nucl. Sci. **53** (2006) 1323.

[30] A. Tsaregorodtsev *et al.*, *DIRAC3: The new generation of the LHCb grid software*, J. Phys. Conf. Ser. **219** (2010) 062029.

[31] J. D. Hunter, *Matplotlib: A 2D graphics environment*, Comput. Sci. Eng. **9** (2007) 90.

[32] S. K. Lam *et al.*, *numba/numba: Version 0.52.0*, 2020. doi: 10.5281/zenodo.4343231.

[33] C. R. Harris *et al.*, *Array programming with NumPy*, Nature **585** (2020) 357, arXiv:2006.10256.

[34] P. Virtanen *et al.*, *SciPy 1.0: Fundamental algorithms for scientific computing in Python*, Nat. Methods **17** (2020) 261, arXiv:1907.10121.

[35] J. Köster and S. Rahmann, *Snakemake—a scalable bioinformatics workflow engine*, Bioinformatics **28** (2012) 2520.

[36] A. Meurer *et al.*, *SymPy: Symbolic computing in Python*, PeerJ Comput. Sci. **3** (2017) e103.

[37] H. Schreiner *et al.*, *scikit-hep/boost-histogram: Version 0.12.0*, 2021. doi: 10.5281/zenodo.4476368.

[38] H. Dembinski *et al.*, *scikit-hep/iminuit: v2.0.0*, 2020. doi: 10.5281/zenodo.4310361.

[39] E. Rodrigues *et al.*, *scikit-hep/particle: Version 0.14.0*, 2020. doi: 10.5281/zenodo.4292256.

[40] J. Pivarski *et al.*, *scikit-hep/uproot3: 3.14.2*, 2020. doi: 10.5281/zenodo.4321705.

[41] E. Rodrigues *et al.*, *The Scikit HEP project – Overview and prospects*, EPJ Web Conf. **245** (2020) 06028, arXiv:2007.03577.

This paper cites all high-level software used in the analysis. Why don't we always do this?