The LHCb Open Data Project

Sebastian Neubert[1]

[1]HISKP Bonn

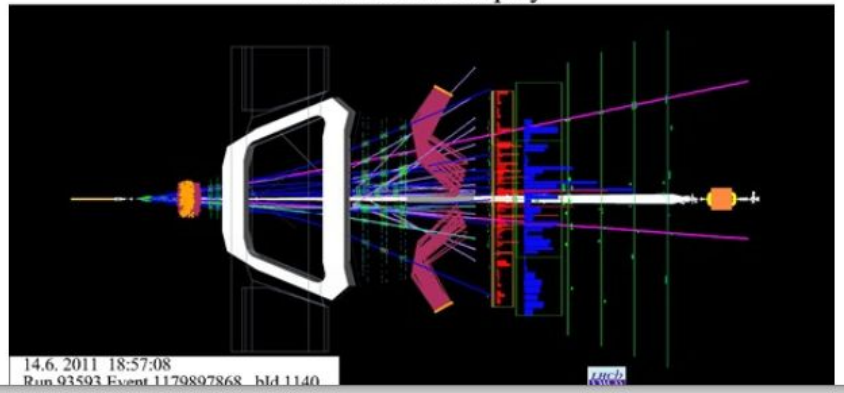
FAIROS-HEP Workshop,
February 2023

LHCb releases first set of data to the public

The LHCb collaboration has released data from Run 1 of the LHC to the public for the first time, allowing research to be conducted by anyone in the world

8 DECEMBER, 2022 | By LHCb collaboration

LHCb Event Display



Related Articles

CMS completes the release of its entire Run 1...
Knowledge sharing | News | 14 December, 2022

CERN publishes comprehensive open science pol...
Knowledge sharing | News | 3 October, 2022

CMS

opendata CERN

Search

Help About

Dataset × Collision × LHCb ×

include on-demand datasets

Filter by type

- Dataset 31
  - Collision 28
  - Derived 3
- Documentation 2149
- Environment 1
- News 1
- Software 1

Filter by experiment

- ALICE 14
- CMS 224
- LHCb 28

Filter by year

- 2011 13
- 2012 14

Filter by file type

- DST 15
- MDST 12
- root 1

Filter by collision type

- pp 27

Filter by collision energy

Sort by: Title A-Z asc. Display: detailed 20 results

Found 28 results.

LHCb 2011 Beam3500GeV MagDown EW Stream Stripping21r1

proton-proton (pp) collision data collected by the LHCb experiment in the year 2011 of Run1 of the LHC....

Dataset Collision LHCb

LHCb 2011 Beam3500GeV MagDown EW Stream Stripping21r1p1

proton-proton (pp) collision data collected by the LHCb experiment in the year 2011 of Run1 of the LHC....

Dataset Collision LHCb

LHCb 2011 Beam3500GeV MagDown EW Stream Stripping21r1p2

proton-proton (pp) collision data collected by the LHCb experiment in the year 2011 of Run1 of the LHC....

Dataset Collision LHCb

LHCb 2011 Beam3500GeV MagDown LEPTONIC Stream Stripping21r1

proton-proton (pp) collision data collected by the LHCb experiment in the year 2011 of Run1 of the LHC....

Dataset Collision LHCb

Level 3 open data release policy

Policy since 27th Feb 2013 updated in CERN Open Data Policy 2020 and CERN Open Science Policy 2022

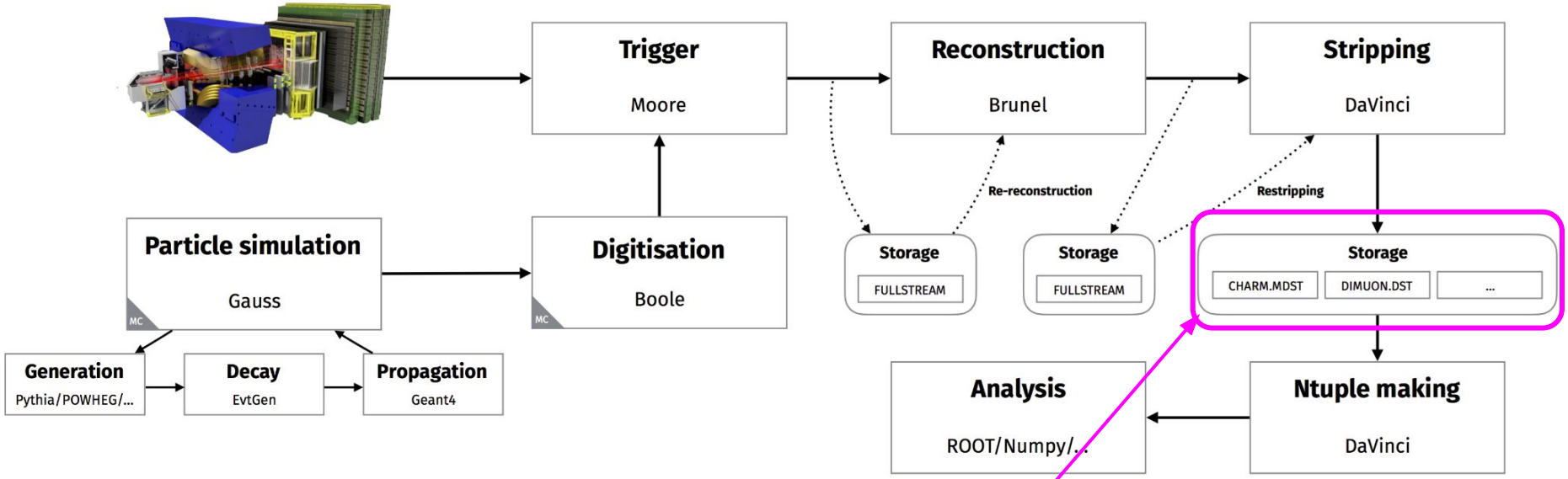
- Level 3 open data: reconstructed events (DPHEP definition)
  - LHCb: Output of stripping / turbo / sprucing
  - MC on demand

- 50 % of data 5 years after end of running period (a.e.r.)
  - Run I: End of 2017
  - **Run II: End of 2023**
- 100 % of data 10 years a.e.r.
  - Run I: End of 2022

- Goal of the OD release is to enable scientific research by 3rd parties
- **Level 3 data releases are addressed at professional users**

LHCb Level 3 Data

Release policy: 50% @ 5yrs, 100% @ 10yrs
after end of running period



- Level 3 data in LHCb **defined as the output of the stripping**
- Same level of abstraction accessed by LHCb members
- Contains **comprehensive set of selections (1620 selections in v21)**
- Organized in ~10 streams, according to physics signature
- Software needed to access data (DaVinci) is open source, available via CVMFS (or container)
- Documentation: LHCb Starterkit openly available

LHCb Run I open data release

- Released 3 Streams:
  - Electroweak EW,
  - Leptonic,
  - Radiative
- ~ 200TB (roughly 20% of RUN I data)
- Data released in LHCb MDST and DST formats
- Needs DaVinci application to read
  - Documentation: Links to LHCb Starterkit

- Detailed description of stripping selections
- Glossary of 960 LHCb specific terms

- Monte Carlo samples on demand

- Missing (planned but LHCb resource limited): Running analysis example

Data to be released next: full 2011/12 Stripping Output

BHADRON.MDST
BHADRONCOMPLETEEVENT.DST
CHARM.MDST
CHARMCOMPLETEEVENT.DST
DIMUON.DST
EW.DST
LEPTONIC.MDST
RADIATIVE.DST
SEMILEPTONIC.DST

Already released

We create **one OD record per stream/year/MagSetting**

Metadata is exported from Dirac BKK

Documentation:
https://lhcb-dpa.web.cern.ch/lhcb-dpa/wp6/open-data-release.html

Release of curation scripts to github in preparation
https://github.com/cernopendata/data-curation/pull/154

Data can only be withheld on a stream by stream basis.
Withholding release of data because of ongoing analyses.

Level 3 Data - Resources

	ALICE	ATLAS	CMS	LHCb
Run 2	2 PB	0.5 PB	2 PB	10 PB (including Run 1)
Run 3	4 PB	1 PB	4 PB	45 PB
Total	**6 PB**	**1.5 PB**	**6 PB**	**55 PB**

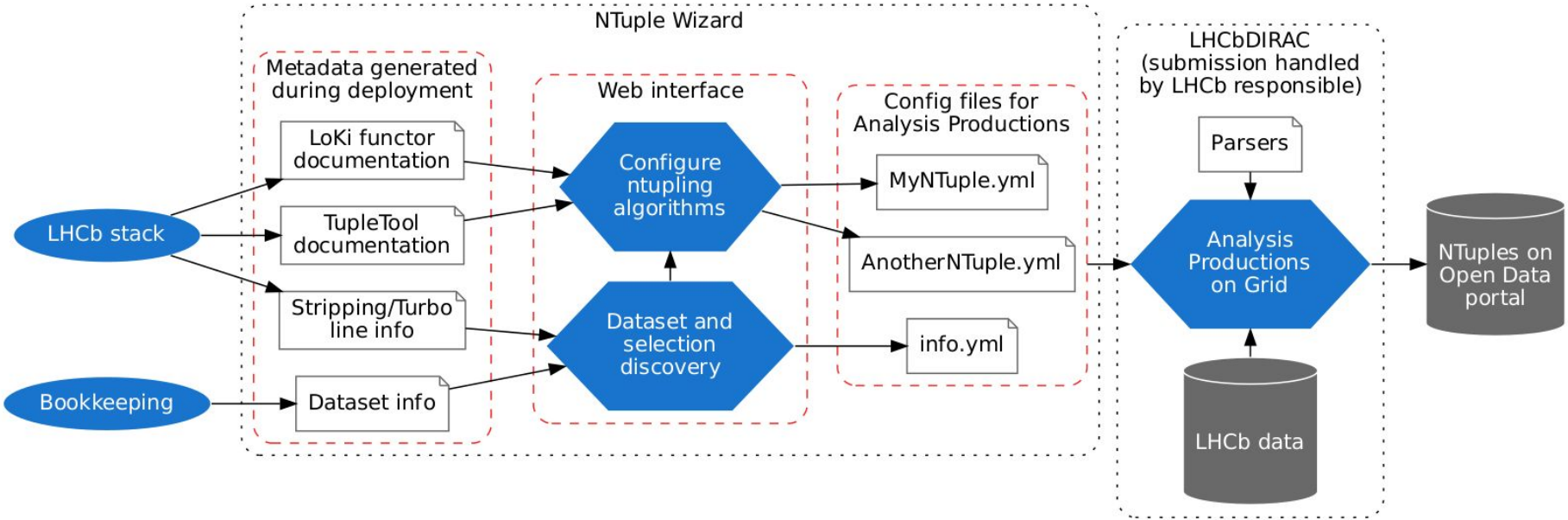
Mitigation Strategies:

- Provide protected access to existing copies of stripping/turbo output via WG-production slots. Needs “ntupling wizard”
- Provide direct access to data on grid storage

Future development: NtupleWizard

Please test and provide feedback! (mattermost)

- NtupleWizard is functional https://lbwizard.web.cern.ch/
- Idea: only keep existing replicas of the data, allow OD users access via dedicated analysis production jobs



Paper in preparation

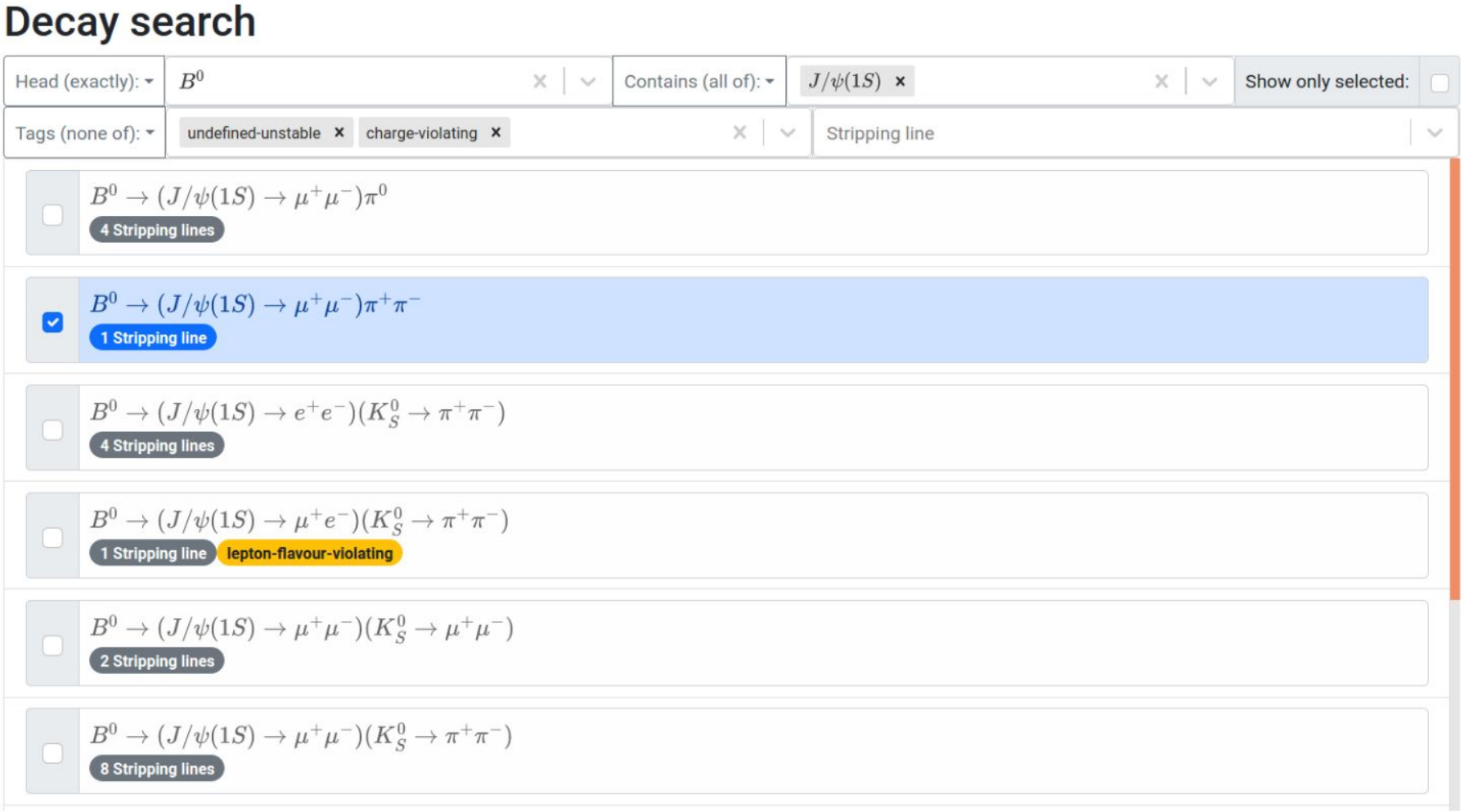


Fig. 3 Example of the decay candidate search function of the Ntuple Wizard.

NTuple Wizard

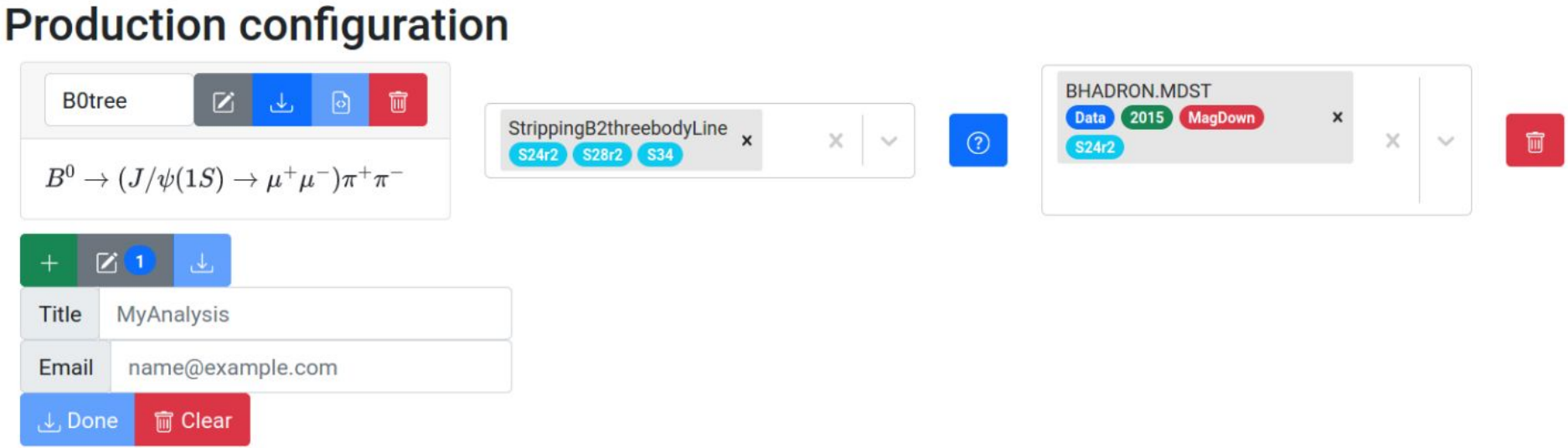
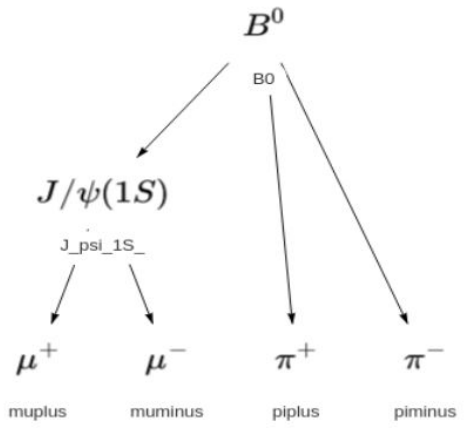


Fig. 4 Example of the data set selection and production configuration step of the Ntuple Wizard.

Configure $B^0 \to (J/\psi(1S) \to \mu^+\mu^-)\pi^+\pi^-$



Select by category

Hadron | Meson | Lepton | X0 | X+ | X- | Down | Beauty | Charm | Up | LongLived | ShortLived | Stable | StableCharged | Scalar | Vector | Spinor

Current selection: $B^0 \to (J/\psi(1S) \to \mu^+\mu^-)\pi^+\pi^-$

5 TupleTools

- TupleToolANNPID
- TupleToolEventInfo
- TupleToolGeometry
- TupleToolKinematic
- TupleToolPid

We are just getting started

Challenges with LHCb Open Data release

Things to improve or add

- Calibration samples + tools
- Documentation on available MC samples
- Analysis example + runtime environment

- MDST and DST are very specialized data formats
  - NTuple wizard will write plain ROOT ntuples
  - NTuple wizard provides much clearer representation of the content of the data
- Integrate NTuple Wizard with Open Data Portal (activity starting now)

- Training for outside users (see CMS Open Data workshops)
- All activity currently severely limited by available resource within LHCb

Going beyond level 3 data

Open science and Open data policies:

We should publish ntuples and statistical models
to make our results more impactful

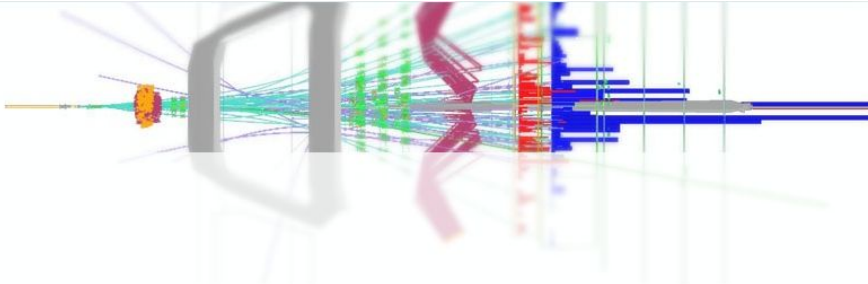
5. Research integrity, reuse and reproducibility

CERN is committed to ensuring the integrity of research. In order to facilitate the reuse of its research products, CERN provides infrastructures to accommodate the scale and complexity of its research outputs. Reuse and reproducibility are facilitated by practising comprehensive analysis preservation to capture relevant research objects, such as research data releases with supporting metadata, auxiliary data, linked software, reproducible analysis workflows, documentation, etc.

what to publish depends on individual analysis

Published Results (Level 1) Policy: Peer-reviewed publications represent the primary scientific output from the experiments. In compliance with the CERN Open Access Policy, all such publications are available with Open Access, and so are available to the public. To maximise the scientific value of their publications, the experiments will make public additional information and data at the time of publication, stored in collaboration with portals such as HEPData,[4] with selection routines stored in specialised tools. The data made available may include simplified or full binned likelihoods, as well as unbinned likelihoods based on datasets of event-level observables extracted by the analyses. Reinterpretation of published results is also made possible through analysis preservation and direct collaboration with external researchers.

The LHCb Public results



LHCb publications

[to restricted-access page]

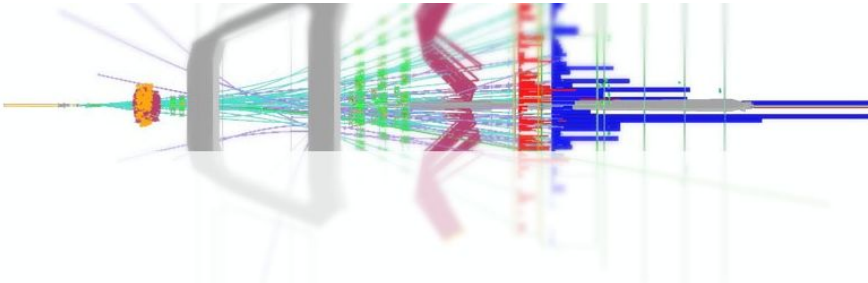
PUBLICATIONS PER WORKING GROUP

- B DECAYS TO CHARMONIUM
- B DECAYS TO OPEN CHARM
- CHARMLESS b-HADRON DECAYS
- b-HADRONS AND QUARKONIA
- CHARM PHYSICS
- FLAVOUR TAGGING
- LUMINOSITY
- QCD, ELECTROWEAK AND EXOTICA
- RARE DECAYS
- SEMILEPTONIC B DECAYS
- DETECTOR PERFORMANCE
- IONS AND FIXED TARGET

List of papers (Total of 655 papers and 50538 citations)

TITLE	DOCUMENT NUMBER	JOURNAL	SUBMITTED ON	CITED
Measurement of Υ production in pp collisions at $\sqrt{s} = 5$ TeV	PAPER-2022-036 arXiv:2212.12664 [PDF]	JHEP	24 Dec 2022	
First observation and branching fraction measurement of the $\Lambda_b^0 \to D_s^- p$ decay	PAPER-2022-038 arXiv:2212.12574 [PDF]	JHEP	23 Dec 2022	
Search for rare decays of D^0 mesons into two muons	PAPER-2022-029 arXiv:2212.11203 [PDF]	PRL	21 Dec 2022	
Measurement of lepton universality parameters in $B^+ \to K^+\ell^+\ell^-$ and $B^0 \to K^{*0}\ell^+\ell^-$ decays	PAPER-2022-045 arXiv:2212.09153 [PDF]	PRD	18 Dec 2022	10
Test of lepton universality in $b \to s\ell^+\ell^-$ decays	PAPER-2022-046 arXiv:2212.09152 [PDF]	PRL	18 Dec 2022	9
Search for the rare decays $W^+ \to D_s^+\gamma$ and $Z \to D^0\gamma$ at LHCb	PAPER-2022-033 arXiv:2212.07120 [PDF]	Chin. Phys. C	14 Dec 2022	
Search for $K^0_{S(L)} \to \mu^+\mu^-\mu^+\mu^-$ decays at LHCb	PAPER-2022-035 arXiv:2212.04977 [PDF]	PRD	09 Dec 2022	
Amplitude analysis of $B^0 \to \overline{D}^0 D_s^+\pi^-$ and $B^+ \to D^- D_s^+\pi^+$ decays	PAPER-2022-027 arXiv:2212.02717 [PDF]	PRD	06 Dec 2022	
First observation of a doubly charged tetraquark and its neutral partner	PAPER-2022-026 arXiv:2212.02716 [PDF]	PRL	06 Dec 2022	
J/ψ and D^0 production in $\sqrt{s_{NN}}$ = 68.5 GeV PbNe collisions	PAPER-2022-011 arXiv:2211.11652 [PDF]	EPJC	21 Nov 2022	
Charmonium production in pNe collisions at $\sqrt{s_{NN}} = 68.5$ GeV	PAPER-2022-014 arXiv:2211.11645 [PDF]	EPJC	21 Nov 2022	
Open charm production and asymmetry in pNe collisions at $\sqrt{s_{NN}} = 68.5$ GeV	PAPER-2022-015 arXiv:2211.11633 [PDF]	EPJC	21 Nov 2022	
Searches for the rare hadronic decays $B^0 \to p\bar{p}p\bar{p}$ and $B_s^0 \to p\bar{p}p\bar{p}$	PAPER-2022-032 arXiv:2211.08847 [PDF]	PRL	16 Nov 2022	

The LHCb Public results



Search for rare decays of D^0 mesons into two muons

[to restricted-access page]

INFORMATION

LHCB-PAPER-2022-029

CERN-EP-2022-273

ARXIV:2212.11203 [PDF]

(SUBMITTED ON 21 DEC 2022)

PRL

INSPIRE 2616985

TOOLS

GET BIBTEX

Abstract

A search for the very rare $D^0 \to \mu^+\mu^-$ decay is performed using data collected by the LHCb experiment in proton-proton collisions at $\sqrt{s} = 7$, 8 and 13TeV, corresponding to an integrated luminosity of 9fb^{-1}. The search is optimised for D^0 mesons from $D^{*+} \to D^0\pi^+$ decays but is also sensitive to D^0 mesons from other sources. No evidence for an excess of events over the expected background is observed. An upper limit on the branching fraction of this decay is set at $\mathcal{B}(D^0 \to \mu^+\mu^-) < 3.1 \times 10^{-9}$ at a 90% CL. This represents the world's most stringent limit, constraining models of physics beyond the Standard Model.

Figures and captions

Distributions of Δm for (left) $D^0 \to K^-\pi^+$ and (right) $D^0 \to \pi^+\pi^-$ normalisation channels candidates for (top) Run 1 and (bottom) Run 2 data. The distributions are superimposed with the fit.

Fig1a.pdf [32 KiB]
HiDef png [192 KiB]
Thumbnail [154 KiB]

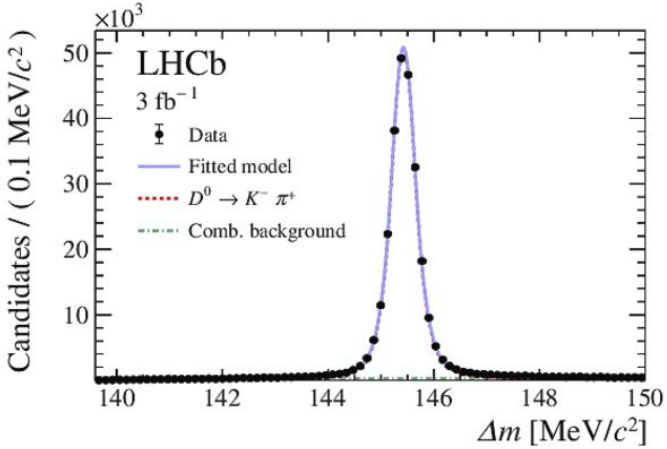


Fig1b.pdf [32 KiB]
HiDef png [195 KiB]
Thumbnail [154 KiB]

LHCb Analysis LifeCycle Management tool

Database tool:

- Organizes complete review workflow
  - Workflow tracker connected to membership database
  - Overview tables for management

- **Collect all kinds of additional information belonging to an analysis**

- Extracts of collected material can be exported
  - Public pages
  - Open science portals, etc

- Will unify/replace WG-databases, EB-database, public pages, …

First Implementation: new LHCb Public FIGURES pages

Old LHCb Public Figures page

Filter figures...

Title	Report number	Keywords	Submitted on
First LHCb upgrade reconstruction results on fixed-target data	LHCb-FIGURE-2023-001	SMOG, Tracking, Real Time Analysis, LHC Run 3	2023-01-04
Di-photon invariant mass	LHCb-FIGURE-2022-019	ECal calibration, Real Time Analysis, LHC Run 3	2022-12-08
Tracking alignment with LHCb Run 3 commissioning data	LHCb-FIGURE-2022-018	SciFi, Tracking, Alignment and Vertexing, VELO, Real Time Analysis, LHC Run 3	2022-11-29
Coarse Time Alignment of the SciFi - Run253101 - LHCb Commissioning	LHCb-FIGURE-2022-017	SciFi, Tracking, LHC Run 3 and 4, LHC Run 3	2022-11-24
VELO alignment with LHCb Run 3 commissioning data	LHCb-FIGURE-2022-016	LHC Run 3, Real Time Analysis, Alignment and Vertexing, ACAT2022, VELO, Tracking	2022-11-14
Recent updates from PV-finder for ACAT 2022	LHCb-FIGURE-2022-015	Alignment and Vertexing, Real Time Analysis, ACAT2022	2022-11-14

Di-photon invariant mass from early Run~3 data

Report Number
LHCb-FIGURE-2022-019

Short abstract
Mass distributions of π^0 candidates reconstructed from Run 243067 and Run 253597.

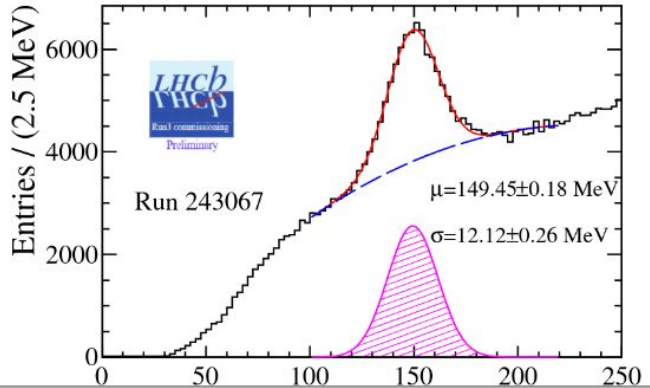
CDS Link

Additional information (only available for LHCb members)

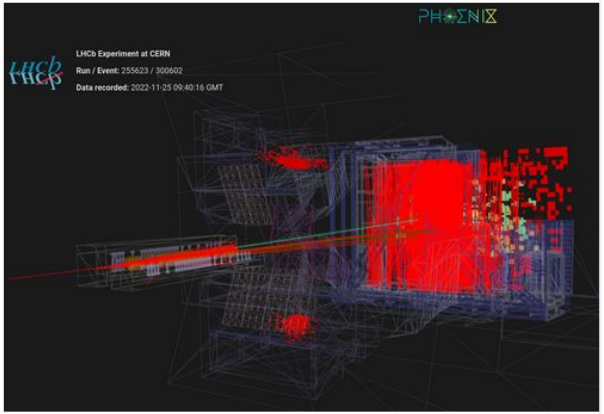
Figures and captions

DOWNLOAD PLOTS

Distribution of the invariant mass of the π^0 candidates (black histogram) reconstructed in Run 243067. The total PDF (red solid line), signal PDF (pink hatched area) and background PDF (blue dashed line) of the fit results are also shown.

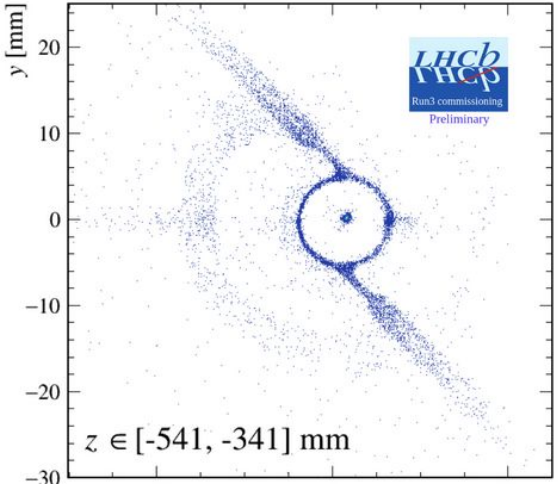


Event display of a proton-argon collision recorded on the 25/11/2022. Data acquired with \velo open with $1\,\mathrm{mm}$ gap, Ar injected in the \smogtwo cell and the reconstruction sequence under commissioning. The same picture with a non-dark theme, different views or different settings for the detector display can be obtained by running on the ProtonArgon.json file in \href{https://lhcb-eventdisplay.web.cern.ch}{phoenix}.



Proton-Argon-zoom_dark.png

Vertices with $z \in [-541, -341]\,\mathrm{mm}$ from beam collisions on the residual gas in \lhc and secondary interactions in the material reconstructed by the \velo open with $1\,\mathrm{mm}$ gap in a run with no injected gas in the \smogtwo cell. The beamspot, the \smogtwo cell in its fully closed position and the support of the injection capillary (right side of the cell) can be clearly distinguished. Note that the \smogtwo closure occurs before that of the \velo, possibly with a misalignment with respect to the beam that cancels when the \velo is also fully closed. Run number 250356.



ALCM

FAQ Help sneubert

First \lhcb upgrade reconstruction results on fixed-target data

Figure approved

SEND E-MAIL NOTIFICATIONS

Report Number
LHCb-FIGURE-2023-001

Associated projects
RTA
VELO

Physics working groups
-

GitLab Repository
Access GitLab Repository page

GitLab CI Artifacts
Access latest pdf document

Proponents
Saverio Mariani

Reviewers
Carla Marin Benito RTA project leader
Michel De Cian RTA deputy project leader
Kazu Akiba VELO deputy project leader
Stefano De Capua VELO deputy project leader
Victor Coco VELO project leader

Approvers
Francesco Polci OPG Chair

Observers
Carla Marin Benito
Michel De Cian
Pasquale Di Nezza

FOLLOW THIS FIGURE

Group Observers
lhcb-rta-rd-paper-review
lhcb-velo

Keywords

Workflow tracker

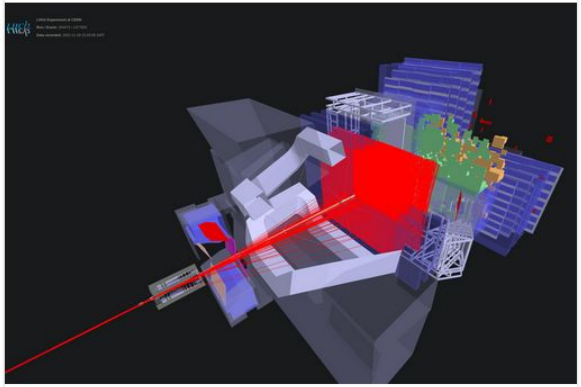
Figure approved

LINK INDICO CONTRIBUTION

Figure graphics

DOWNLOAD PLOTS

Lead-Argon-Collision.jpg



Event display of a lead-argon collision recorded during the test with

ALCM FAQ Help sneubert

first results for the reconstruction of fixed-target beam-gas data with injected hydrogen, helium and argon obtained during the commissioning of the LHCb experiment are discussed.

CDS Link

Workflow history

modified on 2023-01-04 10:17

Figure approved by **Francesco Polci**

modified on 2023-01-02 20:44

Figure marked as ready for approval by **Victor Coco**

modified on 2022-12-14 17:52

Figure marked as ready for review by **Saverio Mariani**

modified on 2022-12-14 14:31

Figure registered by **Saverio Mariani**

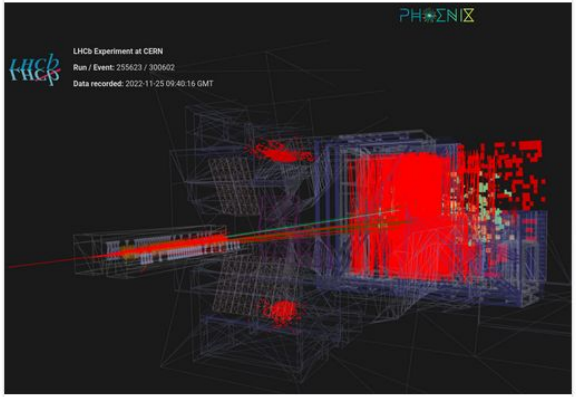
Figure information history

Indico contributions

Filter indico contributions...

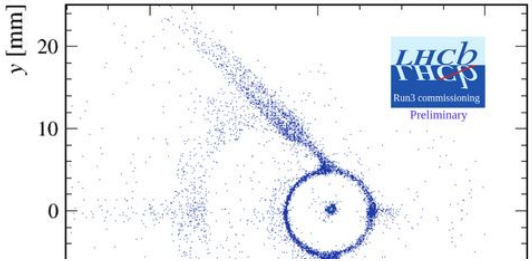
Name	Tags	
RTA: Reconstruction WP2 meeting	EDIT TAGS	

Rows per page: 10 1-1 of 1



Event display of a proton-argon collision recorded on the 25/11/2022. Data acquired with \velo open with 1 mm gap, Ar injected in the \smogtwo cell and the reconstruction sequence under commissioning. The same picture with a non-dark theme, different views or different settings for the detector display can be obtained by running on the ProtonArgon.json file in \href{https://lhcb-eventdisplay.web.cern.ch}{phoenix}.

SMOG2_cell_material_interactions.png



ALCM

FAQ Help sneubert

- Home
- Figure
  - Register figure
  - All figures
  - Figures public pages

Old LHCb Public Figures page

Filter figures

REGISTER FIGURE

Show only ongoing figures

Title	Report number	Stage	Keywords	Registration date	GitLab
Machine-Learnt parametrizations for the Ultra-Fast Simulation of the LHCb detector	LHCb-FIGURE-2022-004	Figure imported	Imported figure	2022-12-19	https://gitlab.ce docs/FIGURE/L FIGURE-2022-0
First LHCb upgrade reconstruction results on fixed-target data	LHCb-FIGURE-2023-001	Figure approved	SMOG, Tracking, Real Time Analysis, LHC Run 3	2022-12-14	https://gitlab.ce docs/FIGURE/ll 2023-001
Di-photon invariant mass	LHCb-FIGURE-2022-019	Figure approved	ECal calibration, Real Time Analysis, LHC Run 3	2022-11-30	https://gitlab.ce docs/FIGURE/ll 2022-019
Mass plots with early Run\,3 data	draft-LHCb-FIGURE-2022-034	Drafting	Real Time Analysis, Trigger, LHC Run 3	2022-11-28	https://gitlab.ce docs/FIGURE/d /draft-lhcb-figu
Coarse Time Alignment of the SciFi - Run253101 - LHCb Commissioning	LHCb-FIGURE-2022-017	Figure approved	SciFi, Tracking, LHC Run 3 and 4, LHC Run 3	2022-11-24	https://gitlab.ce docs/FIGURE/ll 2022-017
Tracking alignment with LHCb Run 3 commissioning data	LHCb-FIGURE-2022-018	Figure approved	SciFi, Tracking, Alignment and Vertexing, VELO, Real Time Analysis, LHC Run 3	2022-11-22	https://gitlab.ce docs/FIGURE/ll 2022-018
Calo Graph Clustering performance	draft-LHCb-FIGURE-2022-031	Marked as ready for review	Real Time Analysis, Computing, LHC Run 3	2022-11-18	https://gitlab.ce docs/FIGURE/d /draft-lhcb-figu
Performance of the Run 3 HLT2 topological triggers	draft-LHCb-FIGURE-2022-030	Drafting	Real Time Analysis, Physics, LHC Run 3, ACAT2022	2022-10-26	https://gitlab.ce docs/FIGURE/d /draft-lhcb-figu
VELO alignment with LHCb Run 3	LHCb-FIGURE-2022-016	Figure approved	Tracking, Alignment and Vertexing, VELO	2022-10-12	https://gitlab.ce docs/FIGURE/ll

ALCM

FAQ Help sneubert

Calo Graph Clustering performance

Marked as ready for review

SEND E-MAIL NOTIFICATIONS

Associated projects
Calorimeter
RTA

Physics working groups
-

GitLab Repository
Access GitLab Repository page

GitLab CI Artifacts
Access latest pdf document

Proponents
Nuria Valls Canudas

Reviewers
Frederic Machefert (Calorimeter project leader)
Iouri Guz (Calorimeter deputy project leader)
Andreas Schopper (Calorimeter deputy project leader)
Carla Marin Benito (RTA project leader)
Michel De Cian (RTA deputy project leader)

Approvers
Francesco Polci (OPG Chair)

Observers FOLLOW THIS FIGURE

Group Observers
lhcb-rta-rd-paper-review

Keywords
Computing
LHC Run 3
Real Time Analysis

Short Abstract
The Graph Clustering algorithm is the default reconstruction method for the ECAL detector in Run 3. It implements the same

Workflow tracker

1 2 3 4

Required actions to go to Marked as ready for approval

☐ Mark figure as ready for approval

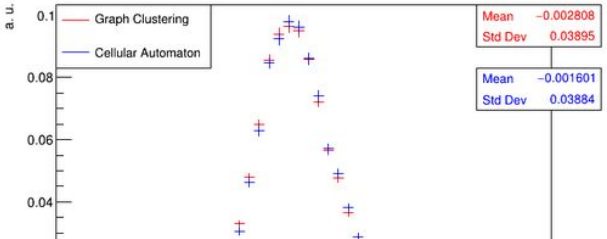
The next action should be performed by one of the following:
Frederic Machefert
Iouri Guz
Andreas Schopper
Carla Marin Benito
Michel De Cian

+ LINK INDICO CONTRIBUTION

Figure graphics

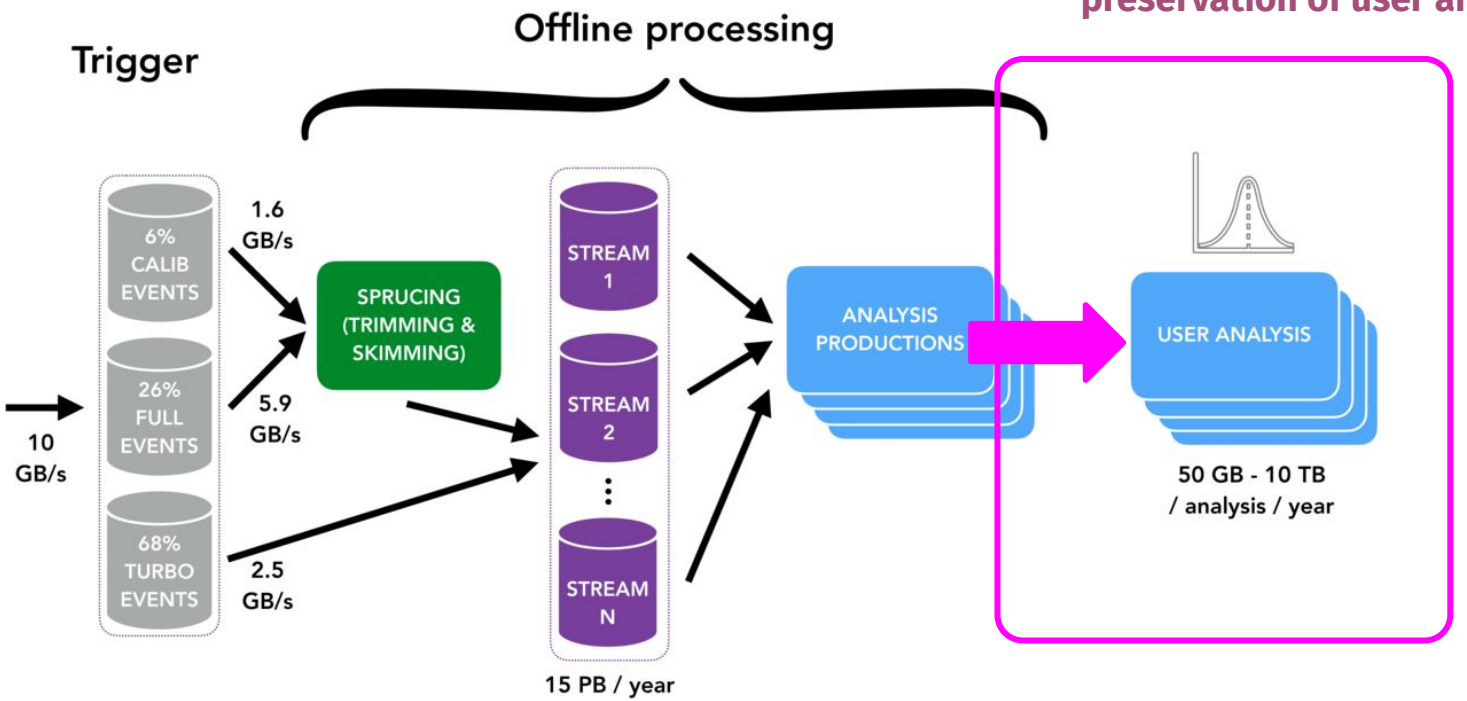
DOWNLOAD PLOTS

histoE.png



Preservation of User Analyses

This talk:
preservation of user analyses



Goals:
Allow analysts to creatively solve their analysis.

Flexible choice of tools and methods.

Preserve ingredients needed for **interpretation of data**

centrally managed and preserved | managed by proponents / PWG

data preparation | data interpretation

Analysis preservation domains

Full Analysis Preservation

CERN analysis preservation supports each of these domains (see later in this talk)

Best Practices

Analysis scripting
(automation)

Runtime Environment

Foundation

Analysis Documentation

- ANA note
- Twiki → future ALCM
- Paper
- Plots, tables
  (as submitted to publication database)
- HEPdata entry, RIVET plugin etc

Input data

- **Ntuples**
- Calibrations
- Classifiers, efficiency maps

Link between ntuple production and analysis

Analysis Code

- source code and scripts implementing the analysis

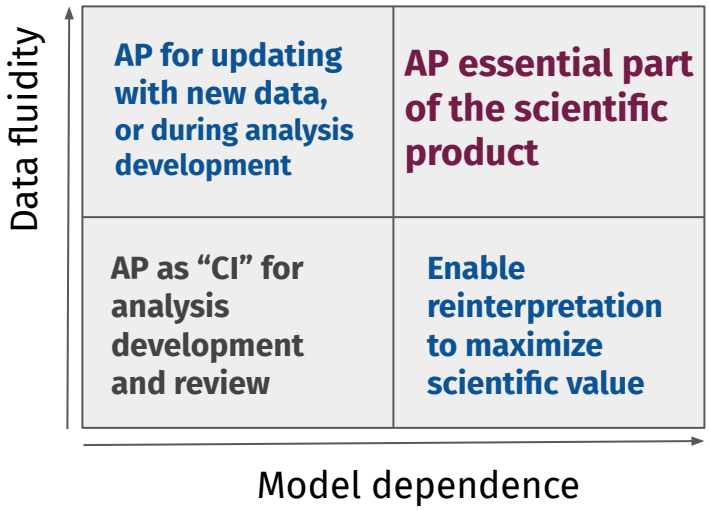
Minimal Analysis Preservation

Who should use full analysis preservation (AP)?

Data fluidity

- updating analysis with new data
  - **e.g. early measurements**
- control channels and their analysis for
  calibrations and efficiencies
  - during commissioning
  - precision measurements

- **combining measurements**

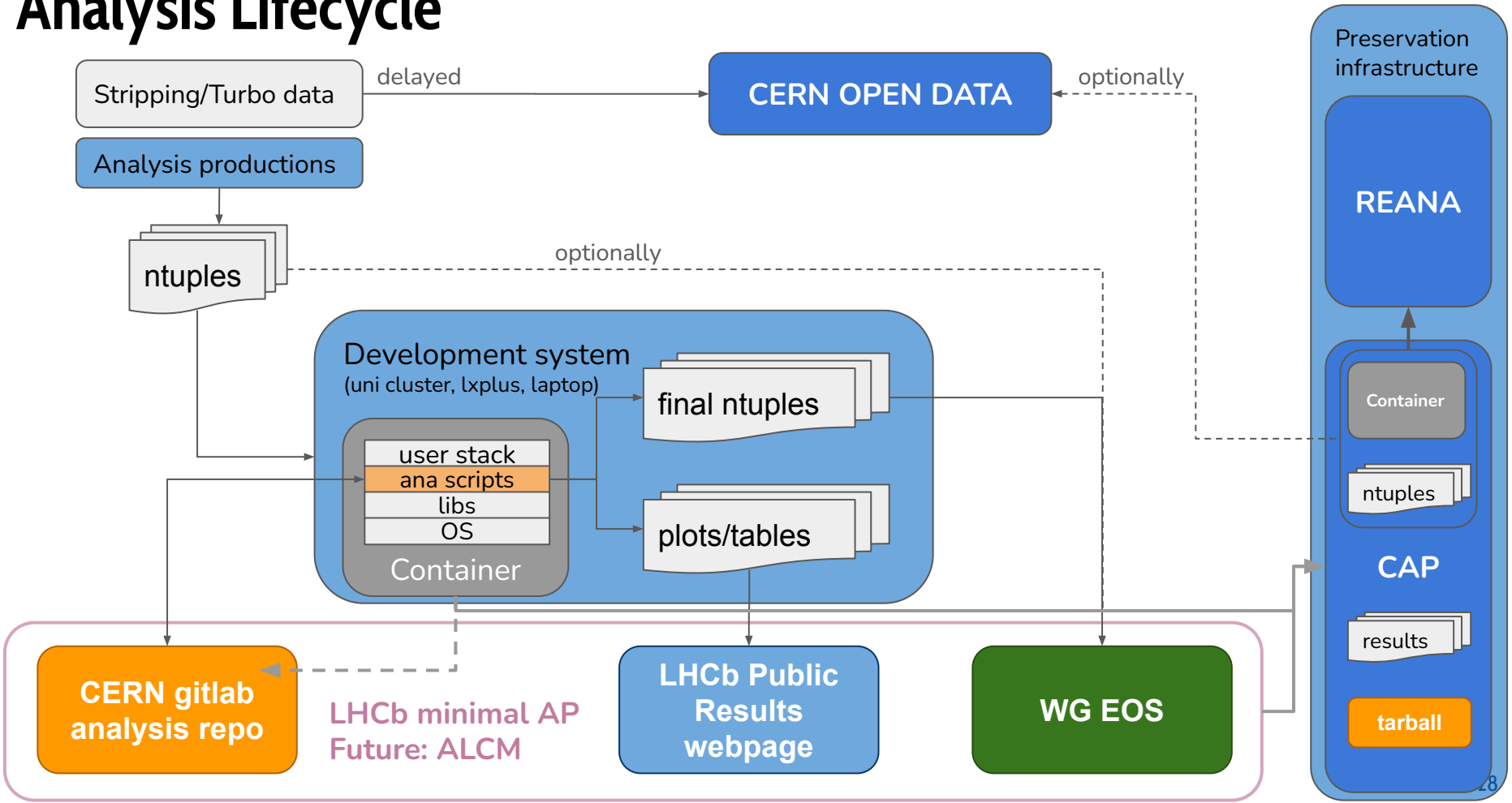


Model dependence

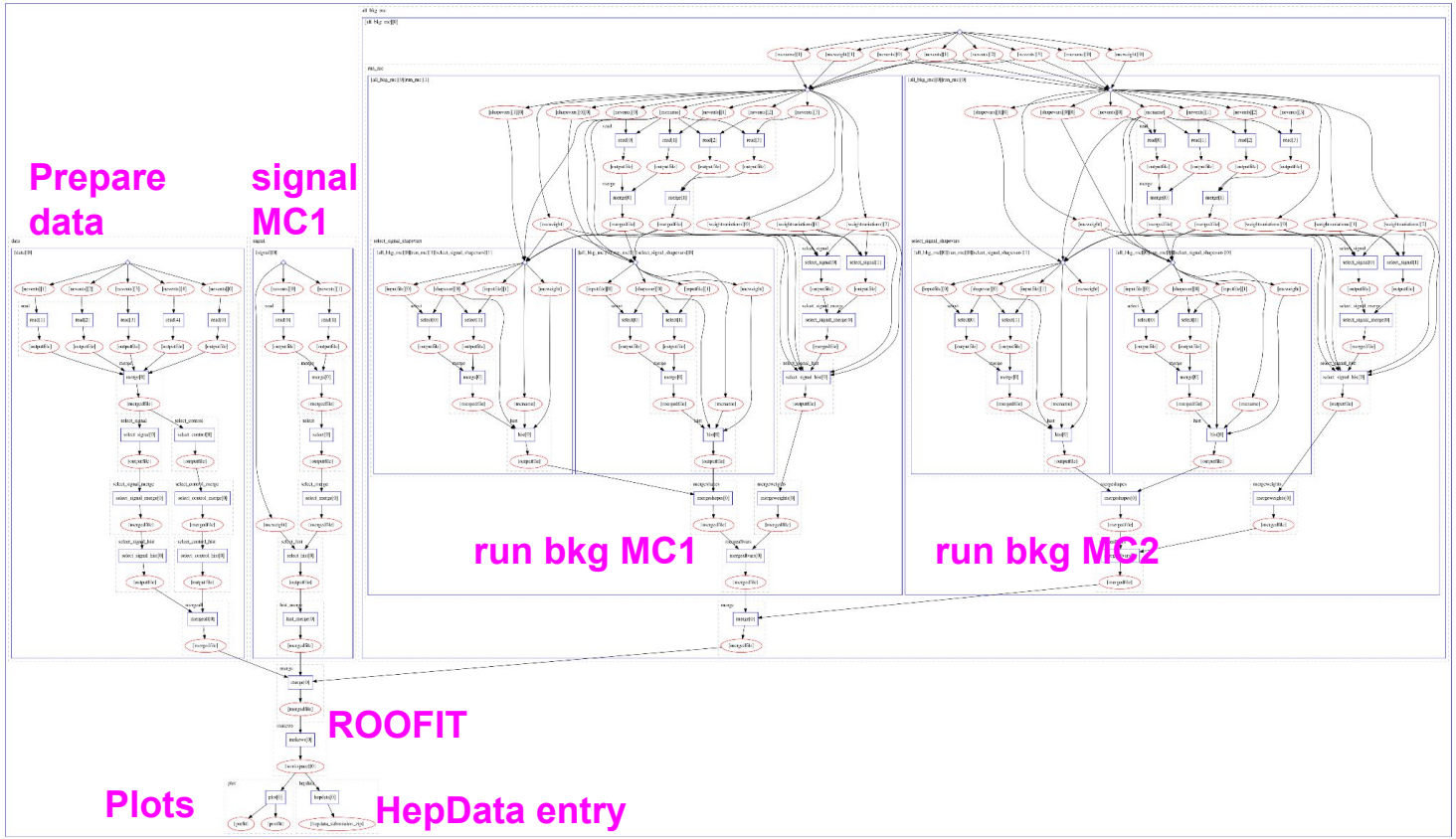
- significant phenomenology input
  - **amplitude analyses!**
- choice of observables based on theory input
- **auxiliary inputs**: e.g. Formfactors
- MC generators / samples
- statistical methodology

The level of detail of analysis preservation and published research products need to be decided case-by-case

Analysis Lifecycle



https://github.com/reanahub/reana-demo-bsm-search



Analysis Productions

Starterkit Lesson

- Ntuple production metadata preserved automatically!
- Not yet supporting Run 3 DaVinci
  - Conversion will be simple from

```
lb-run DaVinci/vXrY \
      gaudirun.py my_options.py
```

- Need to maintain link between ntuple production and analysis
- Will be able to use apd (Analysis Production Data)
- Provides PFN(s) for datasets
  - Designed to allow analyses to be rerunnable long-term

```
import apd
datasets = apd.AnalysisData("MyWG", "MyAnalysis")

rule train_bdt:
    input:
        data = datasets(datatype="2022", mc=False),
        mc = datasets(datatype="2022", mc=True)
    output:
        fn = "classifier.pkl"
    shell:
        "scripts/train_bdt.py --data {' '.join(input.data)} --mc {' '.join(input.mc)}"
```

ALCM for Papers and Analysis Preservation

ALCM

FAQ Help sneubert

Testing analysis registration on integration

WG discussion

Physics working group
B decays to Charmonia

Proponents
Gabriel Jose Souza E Silva
Carlos Brito Analysis contact

Approvers
Yasmine Sara Amhis Physics coordinator
John Walsh Editorial board chairperson
Mika Anton Vesterinen Editorial board chairperson deputy

Observers
Joel Closier

FOLLOW THIS ANALYSIS

Group Observers
lhcb-glance-admins

Keywords
detector plot

WG Reader
-

Workflow tracker

Required actions to go to WG review

Assign working group reader

ASSIGN WORKING GROUP READER

+ LINK GITLAB REPOSITORY

+ LINK INDICO CONTRIBUTION

+ DOCUMENT

+ STORAGE LOCATION

+ TAG ATTACHMENT

Future: add link to computational elements

Key Attachments

-

find analysis code here

find ntuples here

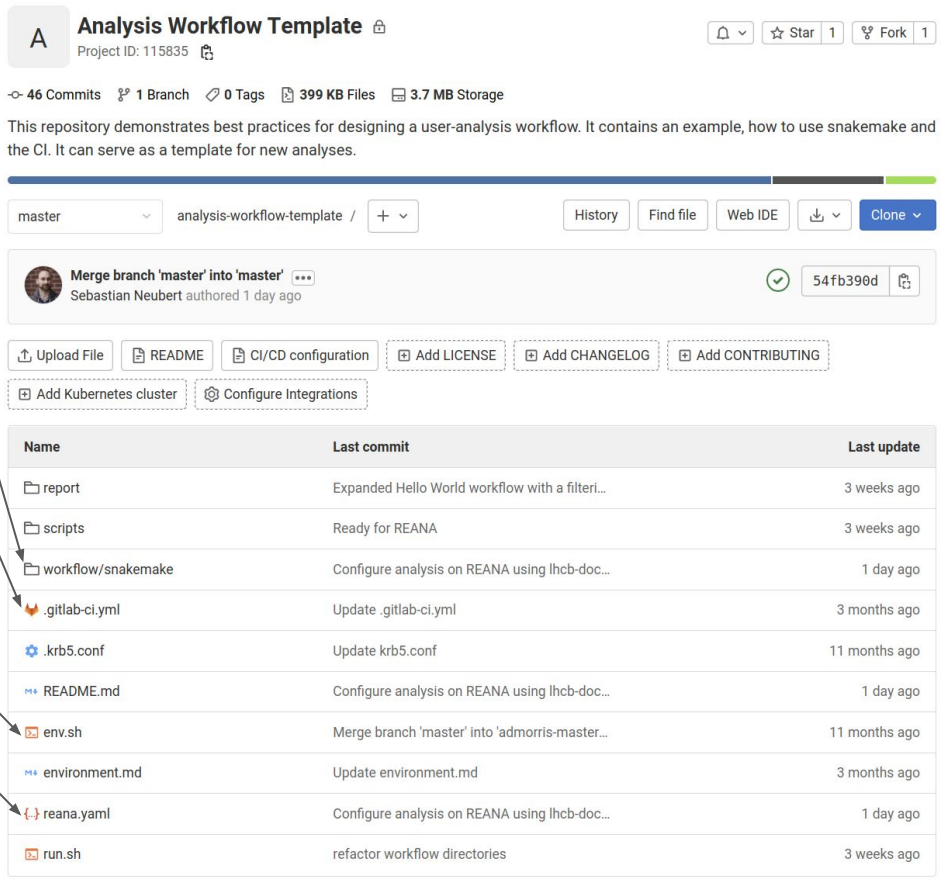
DOCUMENTS GITLAB REPOSITORIES INDICO CONTRIBUTIONS STORAGE LOCATIONS Analysis Workflows

Attachments

LHCb Analysis Workflow Template

https://gitlab.cern.ch/lhcb-dpa/wp6-analysis-preservation-and-open-data/analysis-workflow-template

- Basic skeleton snakemake workflow (not a snakemake tutorial)
- Demonstrates **snakemake reports**
- Demonstrates running analysis in the **gitlab-ci**
- Runtime environment config through **lb-conda**
- Demonstrates **deploy to REANA**
- Can be used to initialize a new analysis repo
- .... or as tutorial what to add to your existing project



Snakemake workflow description

Set of analysis scripts, input data, and parameters +
tacit knowledge how and in what order to run them

→

Machine readable description of workflow (similar to Makefile for software build)

- Snakemake selected as top recommendation after comparative review in 2017
  (see LHCb-INT-2017-021)

- Wide use inside collaboration
- Feature complete
- Easy to get started
- Supported by CERN REANA

Snakemake is very well documented

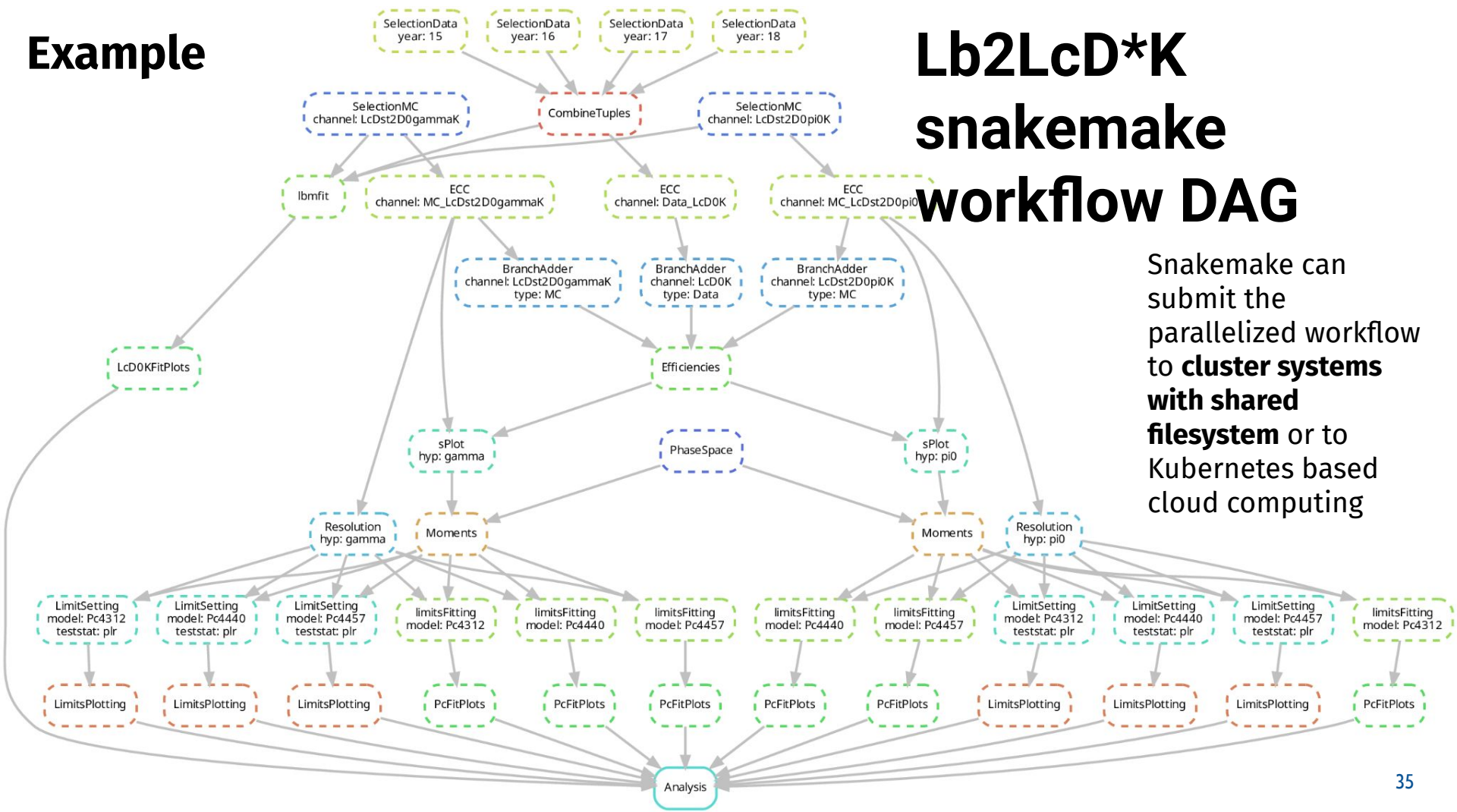
https://snakemake.readthedocs.io/en/stable/

https://snakemake.readthedocs.io/en/stable/snakefiles/best_practices.html

https://hsf-training.github.io/analysis-essentials/snakemake/README.html

Example

Lb2LcD*K snakemake workflow DAG



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

Provenance tracking: snakemake reports

Static html generated by snakemake

Register important results for reporting

Shows rules, and parameters how each plot was produced

Snakemake Report

Fri Apr 29 17:35:15 2022 CET
Snakemake 6.8.0

- Workflow
- Statistics
- Configuration

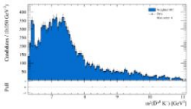
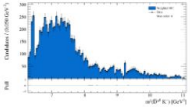
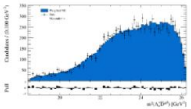
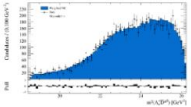
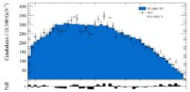
RESULTS

- Efficiencies
- Fits_LcD0K
- Legendre Moments
- Resolution

Legendre Moments

Show 10 entries

Search:

File	Description	Job properties		Thumbnail
Moments_DstK_ECCgamma_4.png		Rule	MomentsPlotting	
		Wildcards	hyp=gamma, max_order=4	
		Parameters	-	
Moments_DstK_ECCpi0_4.png		Rule	MomentsPlotting	
		Wildcards	hyp=pi0, max_order=4	
		Parameters	-	
Moments_LcDst_ECCgamma_4.png		Rule	MomentsPlotting	
		Wildcards	hyp=gamma, max_order=4	
		Parameters	-	
Moments_LcDst_ECCpi0_4.png		Rule	MomentsPlotting	
		Wildcards	hyp=pi0, max_order=4	
		Parameters	-	
Moments_LcK_ECCgamma_4.png		Rule	MomentsPlotting	
		Wildcards	hyp=gamma, max_order=4	
		Parameters	-	

https://snakemake.readthedocs.io/en/stable/snakefiles/reporting.html

Snakemake Report

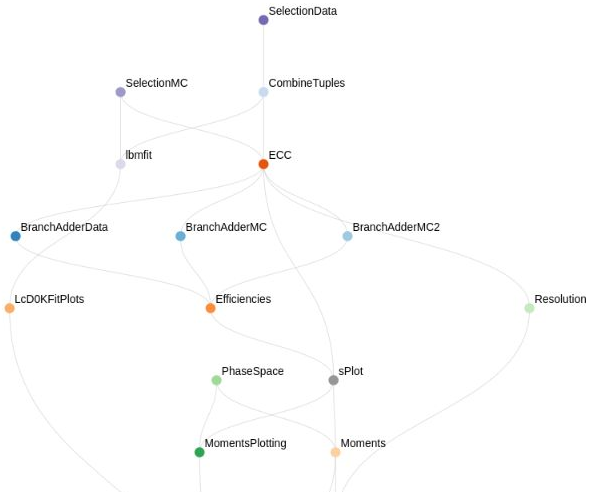
Fri Apr 29 17:35:15 2022 CET
Snakemake 6.8.0

- Workflow
- Statistics
- Configuration

RESULTS

- Efficiencies
- Fits_LcD0K
- Legendre Moments
- Resolution

Workflow



Explore workflow interactively

```
35  rule hello_world:
36      """
37      Run a helloworld script, which  takes the filtered tree and makes a histogram
38      """
39      input:
40          script = 'scripts/helloworld.C',
41          tree = 'filtered_tree.root'
42      container:
43          "docker://gitlab-registry.cern.ch/lhcb-docker/os-base/centos7-devel"
44      output:
45          log='output/logs/logfile_helloworld',
46          plot=report('x.pdf',category='Histograms',caption='report/caption_x.rst')
47      shell:
```

Deploy analysis to REANA via gitlab https://reana.cern.ch/profile

reana.yaml 415 Bytes

```yaml
# reana-snakemake.yaml
version: 0.1.0
inputs:
  files:
    - scripts/filter_tree.C
    - scripts/helloworld.C
  directories:
    - workflow/snakemake
    - output/logs
  parameters:
    input: workflow/snakemake/config.yaml
workflow:
  type: snakemake
  file: workflow/snakemake/Snakefile
  resources:
    cvmfs:
      - lhcb.cern.ch
      - lhcbdev.cern.ch
      - lhcb-condb.cern.ch
outputs:
  files:
    - x.pdf
```

Your GitLab projects

- couchWGDB — sneubert/couchWGDB
- cookietest — sneubert/cookietest
- PhiKKmatrix — sneubert/phikkmatrix
- Analysis Workflow Template — sneubert/analysis-workflow-template
- Snakemake Best Practices — sneubert/snakemake-best-practices
- IR3Detector — sneubert/ir3detector

Switch on to deploy to REA
next gitlab-ci job

Pipeline | Needs | Jobs 2 | Tests 0

Run: analysis

External: default

REANA Webinterface https://reana.cern.ch

Analysis Workflow Template #17
Finished 5 days ago

finished in 2 min 42 sec
step 2/2

Open Jupyter Notebook
Delete workflow

Logs | Workspace | Specification

Step hello_world | finished | Kubernetes | gitlab-registry.cern.ch/lhcb-docker/os... | $ ZSH_VERSION= VIRTUAL_ENV= PYT...

```
job: :
    ------------------------------------------------------------------
  | Welcome to ROOT 6.26/00                        https://root.cern |
  | (c) 1995-2021, The ROOT Team; conception: R. Brun, F. Rademakers |
  | Built for linuxx8664gcc on Mar 05 2022, 12:03:00                 |
  | From tag , 3 March 2022                                          |
  | With                                                             |
  | Try '.help', '.demo', '.license', '.credits', '.quit'/'.q'       |
   ------------------------------------------------------------------


Processing scripts/helloworld.C("filtered_tree.root")...
RooRealVar::Hello World from Sebastian Neubert = 0  L(-42 - 42)
TFile**         filtered_tree.root
 TFile*         filtered_tree.root
  KEY: TTree    tree;1  tree
Info in <TCanvas::Print>: pdf file x.pdf has been created
(int) 0
```

Interactive session possible via Jupyter

Don’t make perfect the enemy of good

Ideally:

Produce every plot and (result) number in the ANA note in a workflow on REANA and document their provenance in a snakemake report.

but a more realistic goal might be to ensure that:

The central value of the result can be computed in a (snakemake) workflow running on REANA.

It is also possible to split the analysis into several smaller workflows.
In this case it would be good to preserve intermediate data products.

Concluding remarks: Curating Research Products

- Different scientific questions require different levels of detail in the empirical evidence.
  - **Level of model-dependence** will influence how much the experimental data can be "compressed" into a few numbers.
  - Techniques that allow reinterpretation of the data are the same as those needed to adapt to a fluid dataset

- Decisions on the level of detail of analysis preservation have to be tuned to the individual study - IMHO: avoid one-fits-all solutions
  - It is possible to support this with a small number of generic tools,practices, and standards

- This **data curation** requires dedicated resources.
  - Maximizing scientific value is not for free

- The technologies used to support the effort are very useful beyond fundamental science. Come join us!

Backup

How to make sure the workflow can be run later?

- Capture all dependencies: docker or singularity container
  - Recommended: standard CERN / LHCb containers
  - e.g. gitlab-registry.cern.ch/lhcb-docker/os-base/centos7-devel

- Configure additional software from CVMFS
- For non-LHCb software use lb-conda to load environments from cvmfs
  - https://gitlab.cern.ch/lhcb-core/lbcondawrappers
  - `default` environment provides: Python 3, ROOT, Snakemake, jupyterlab, matplotlib, scikit-learn, tensorflow and many more.

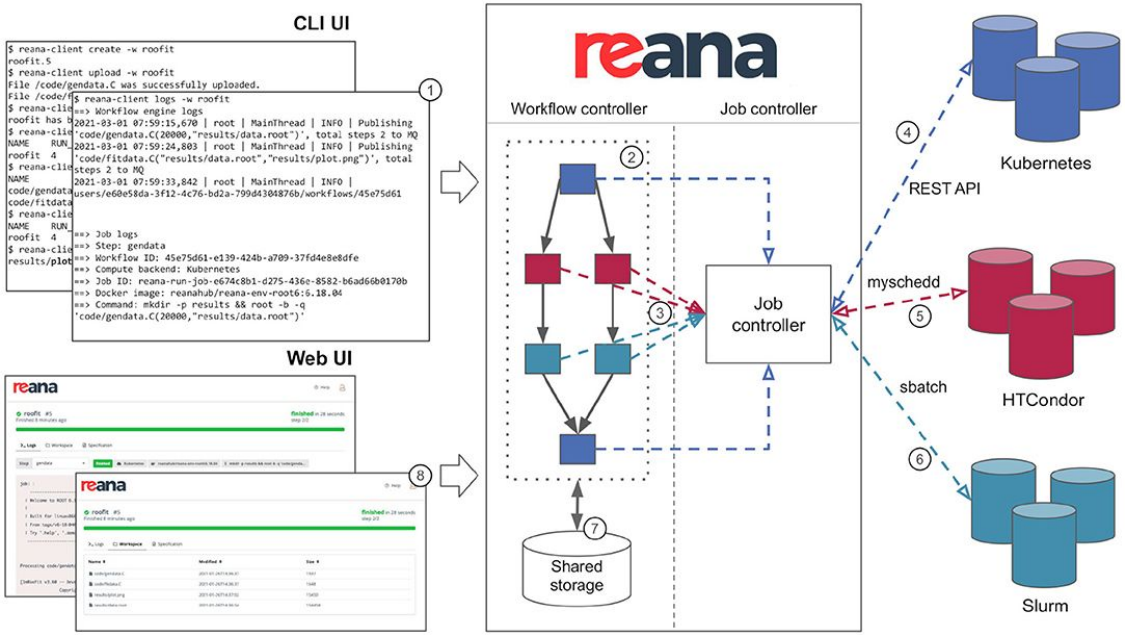
- **Deploy analysis to REANA** to test if everything is preserved appropriately

- EMTF analysis are natural candidates for early adopters!

REANA: https://docs.reana.io/

Scalable Declarative HEP Analysis Workflows for Containerised Compute Clouds" published in Frontiers in Big Data (2021).

[PJ Web Conf 2014 (2019) 06034]



Supported:

Workflow description:
Snakemake, Common Workflow Language, Yadage

Compute backends:
Kubernetes, HTCondor, Slurm

The Open Science Philosophy (at CERN)

Recognize the universal importance of the fundamental scientific knowledge produced at CERN and the key role of openness in the pursuit of CERN organisational mission.

Commits to the advancement of science and wide dissemination of knowledge by adopting practices to make scientific research more open, global, collaborative and responsive to societal changes.

In fulfilment of the collective moral and fiduciary responsibility to member states and the broader global scientific community

Data collected at the LHC is a heritage to humanity.

It has been obtained through collaborative work using public funds.

Therefore, CERN is committed to preserve, curate, steward and share the data with the public.

Goals of Open Data - Maximizing Scientific Value

- Validation / reproduction of published results
- Reinterpretation of the data
  - test future theories
  - refine phenomenological models
  - use different statistical tools
- Reuse of data sets
  - Combined analyses
  - Use collected data as input for future studies
  - Algorithm development (e.g. machine learning community)
- Data mining
  - search for interesting physics in unexplored parts of the data
  - use new techniques to (re-)select data

We cannot anticipate the questions future generations might ask of this data.

require different levels of data complexity

Open Science Landscape - Recent Trends

- Funding agencies: requests for data management plans
- Publishers: requests for data products allowing to
  - validate / reproduce results
  - reuse data for further studies

Science Community: "**Data is not enough**":

- Papers with code https://paperswithcode.com/
- Interactive publications
- Federated infrastructures and computing/science portals (e.g. NFDI)
- Not a new realization (see e.g. DPHEP study group 2013 status report) but technology (esp cloud computing, containerization) has made progress!
- Development driven especially through bioinformatics and machine learning / AI community

Policies

the CERN experiments have given themselves

CERN Open Data Policy 2020

Initiated beginning 2020 by the chair of the European Commission

CERN director of research: Mandate for a working group to draft a common policy for all LHC experiments

Endorsed by the Collaboration Boards of ALICE, ATLAS, CMS and LHCb

CERN Open Science Policy 2022

Includes all experiments at CERN

Working group formed
https://openscience.cern/

Includes a wider scope of topics:

- Open access, open data, open source, open hardware
- Research integrity, research assessment
- Open infrastructure
- Training and outreach, citizen science

Open data policy

The CERN Open Data Policy reflects values that have been enshrined in the CERN Convention for more than sixty years that were reaffirmed in the European Strategy for Particle Physics (2020)[1], and aims to empower the LHC experiments to adopt a consistent approach towards the openness and preservation of experimental data. Making data available responsibly (applying FAIR standards[2]), at different levels of abstraction and at different points in time, allows the maximum realisation of their scientific potential and the fulfillment of the collective moral and fiduciary responsibility to member states and the broader global scientific community. CERN understands that in order to optimise reuse opportunities, immediate and continued resources are needed. The level of support that CERN and the experiments will be able to provide to external users will depend on available resources.

FAIR Data Principles

[The FAIR Guiding Principles for scientific data management and stewardship. Nature *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18]

Findable: Metadata and data should be easy to find for both humans and computers.

Accessible: The exact conditions under which the data is accessible should be provided in such a way that humans and machines can understand them.

Interoperable: The (meta)data should be based on standardized vocabularies, ontologies, thesauri etc. so that it integrates with existing applications or workflows.

Reusable: Metadata and data should be well-described so that they can be replicated and/or combined in different research settings.

https://go-fair.org

Solved by https://opendata.cern.ch

Needs dedicated work by the experimental collaborations (here efforts in HEP are in their infancy)

DPHEP Levels of Data Complexity

[https://arxiv.org/abs/1205.4667]

1. **Published results + additional information**
   - supplemental data tables, ntuples
   - HEPData entries, rivet plugins
   - notes, technical information
   - documentation, slides
   - analysis code, jupyter notebooks

2. **Education and Outreach**
   - simplified data formats, e.g. highly preprocessed ntuples

3. **Reconstructed data + analysis level software**
   - Calibrated reconstructed data with the level of detail useful for algorithmic, performance and physics studies
   - preservation of analysis level experiment-specific software

4. **Raw data + reconstruction software**
   - **Not released for LHC data**

Open data policy: Level 3 data releases

Reconstructed Data (Level 3) Policy: The LHC experiments will release calibrated reconstructed data with the level of detail useful for algorithmic, performance and physics studies. The release of these data will be accompanied by provenance metadata, and by a concurrent release of appropriate simulated data samples, software, reproducible example analysis workflows, and documentation. Virtual computing environments that are compatible with the data and software will be made available. The information provided will be sufficient to allow high-quality analysis of the data including, where practical, application of the main correction factors and corresponding systematic uncertainties related to calibrations, detector reconstruction and identification. A limited level of support for users of the Level 3 Open Data will be provided on a best-effort basis by the collaborations.

Level 3 data is addressed at professional researchers

Analysis workflow template documentation

The repository is designed to be used as a template and only contains basic building blocks. For more complex examples how to use snakemake see https://snakemake.readthedocs.io/en/stable/snakefiles/best_practices.html.

LHCb analyses using snakemake (examples)

- https://gitlab.cern.ch/LHCb-QEE/WmassMeasurement
- https://gitlab.cern.ch/LHCb-RD/rad-lb2pkgamma
- https://gitlab.cern.ch/lhcb-b2cc/sin2beta_b2ccks_run2
- https://gitlab.cern.ch/LHCb-RD/ewp-rkstz
- https://gitlab.cern.ch/lhcb-b2cc/Bs2JpsiPhi-FullRun2
- https://gitlab.cern.ch/mstahl/LcD0KRun2
- https://gitlab.cern.ch/LHCb-RD/rad-lb2l0gamma-angular
- https://gitlab.cern.ch/RECEPT/Semileptonics
- https://gitlab.cern.ch/LHCb-RD/vrd-b2xemu-aachen
- https://gitlab.cern.ch/lhcb-b2oc/analyses/b2oc-deltams-run2
- https://gitlab.cern.ch/lhcb-slb/rdstar-hadronic-run2
- https://gitlab.cern.ch/lhcb-charm/charm-production-run-3
- https://gitlab.cern.ch/lhcb-charm/d2hll-analysis