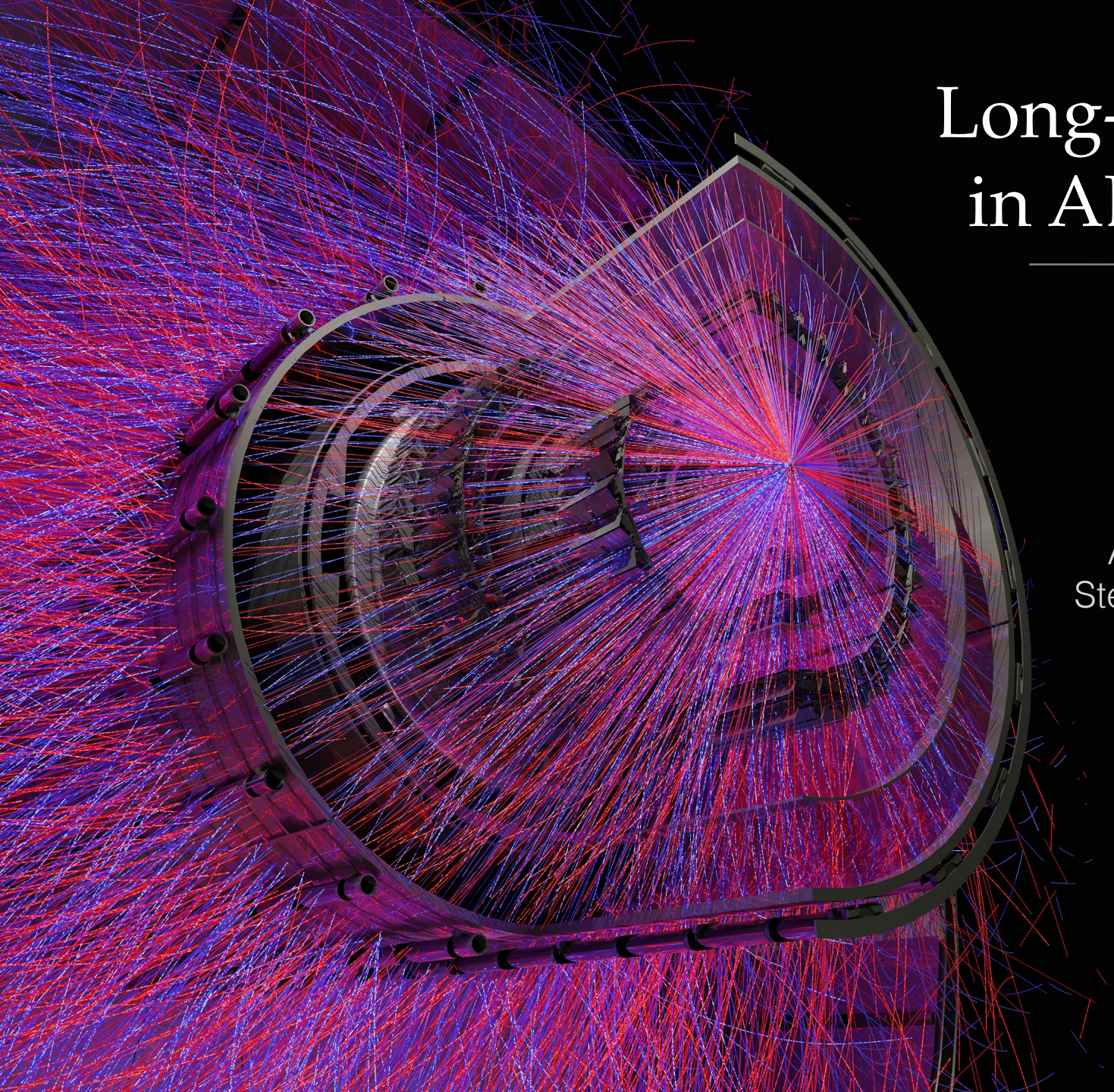


Long-term data preservation in ALICE: status and plans

DPHEP Workshop, 3rd October 2024

Alexandru Florin Dobrin, Peter Hristov,
Stefano Piano, David Dobrigkeit Chinellato
on behalf of ALICE

New ALICE Open Data Format

- **New AO2D format** has been chosen to publish open data
- Based on the new data format and software framework developed by the ALICE O2 project **for Run 3 and Run 4**:
  - It will ensure data preservation of Run 1 and Run 2 data
  - Significantly reduced size per collision (**factor 16** wrt Run 2 ESD and **factor 5** wrt Run 2 AOD)
  - New flat data model optimized for fast IO **(>10x faster** than Run 2 AOD)
  - Possible to adopt (skimmed) derived data set like nanoAOD format to compress further
- Such a refurbishment required a long conversion production of all Run 1 and Run 2 ESDs and AODs into new AOD format for both data and MC
  - Conversion was done in 2022, but the new data format required additional efforts to make the new analysis framework work with the old data in standalone mode (with a reasonable size)
  - Most of the Run 1 and Run 2 results have been published with the old analysis framework

New ALICE format: further practicalities

- The Run 1/2 data will be kept in the new AO2D format also for internal ALICE activities
  - **Synergy between efforts**: conversion benefits both Open Data initiative and ALICE processing / storage

  - Prime example of conversion advantage:
    - Pb-Pb 2015 data is **~1.2 PB** in the native Run 2 format
    - Pb-Pb 2015 data in the AO2D format is only **52 TB but also general-purpose**

  - Cross-checks also took some effort and time
  - Minor adjustments in some analysis-specific services will still be carried out for effortless internal use in complex analyses: still some internal work necessary

New ALICE Open Data Analysis Framework

- **Software integration with CERN OD Portal** with docker container:
  - Required some efforts of the offline group to enable standalone local compilation
  - Available at https://github.com/AliceO2Group/O2OpenAccess
  - Documentation for compilation available on README.md
- **Fully operational:** possible to run the analysis on AO2D for both real data and MC:
  - Provided some examples of analysis tasks

- Open points that prevented us from publishing the data earlier:
  - Centrality and multiplicity estimators' conversion (completed in May 23)
  - TOF and TPC PID calibration
    - Not possible to import them but regenerated for all Run 1 and Run 2 runs
    - TOF PID conversion completed in Sept. 23
    - TPC PID conversion with NN developed in Oct. 23
  - Data validation

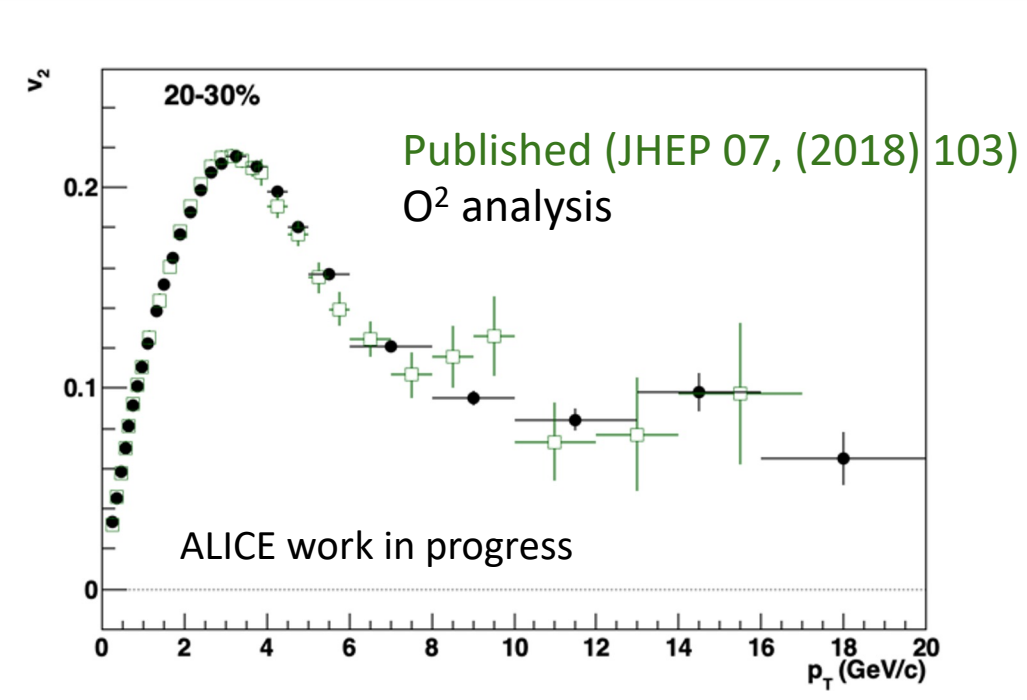
ALICE plans for Open Data implementation

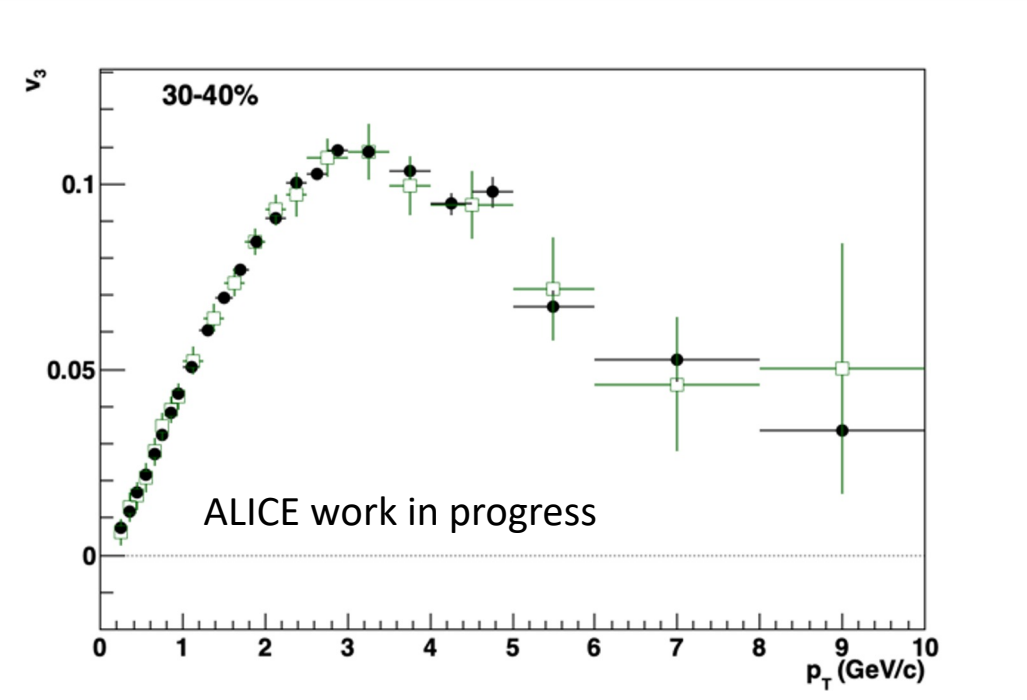
- Set up CERN Open Data Portal with Run 1 ALICE data:
  - Status of previous year: 5% (7%) of Pb-Pb (pp) 2010 ESD datasets released, totaling 6.5 TiB
  - Ready to transfer of ~50% of Run 1 data and related simulations with the new data format:

    Metadata preparation is in progress, and data copy is forthcoming
    LHC10b/c/d/e (pp 7 TeV); LHC10h (Pb-Pb 2.76 TeV); LHC13b/c (p-Pb 5.02 TeV)
    Together with the general purpose MC productions
    (Simple/Run 3) ALICE analysis demonstrator in CERN Open Data portal Docker container to ease portability
  - Documentation with some examples of analysis tasks

- Validation of 10% of Run 2 data
- Collect feedback on the Run 1 released datasets, software framework and documentation before making the Run 2 data public

Lightweight analysis test: flow measurement





- v_n coefficients measured using the scalar product method
- Good agreement with published results
- Lightweight analysis framework being finalised

$$v_n = \frac{\langle u^{\eta<-0.5} Q_n^{\eta>0.5} / M_Q^{\eta>0.5} \rangle}{\sqrt{\langle Q_n^{\eta<-0.5} Q_n^{\eta>0.5} / M_Q^{\eta<-0.5} / M_Q^{\eta>0.5} \rangle}}$$

Expected Open Data release in the next years

- Updated ALICE Open Data release plan :
  - 10% of Run 2 data by 2025 (one-year delay relative to the implementation document)
    - ALICE will gradually reach 50% of Run 2 data by 2029
  - In 2030 ALICE will start releasing 10% of Run 3 data as expected
- Based on the new AO2D format data volume, we expect:
  - 2024: 35 TB (Run 1)
  - 2025: 105 TB (10% Run 2)
  - 2026 - 2029: 105 TB/year to reach 50% of Run 2 in 2028
- In Run 2 ALICE inspected ~1 nb^{-1} Pb-Pb data, while for Run 3 & Run 4 ALICE plans to collect 13 nb^{-1} of Pb-Pb collisions
  - In 2030 we will start releasing Run 3 data
    - ~2 PB/year publishing AO2Ds ⇒ crucial to publish skimmed derived data sets instead

Summary & Outlook

- ALICE Open Data benefit from the new Analysis Framework improvements:
  - New Run 3 AOD format suitable for Run 1 and 2 Open Data
  - Skimmed derived data sets will provide
    - Further data compression crucial to make public Run 3 and 4 data
    - Higher event throughput and reduction of needed CPU wall time
- Drawbacks of the adoption of Run 3 framework for Run 1 and 2 data:
  - Calibrations no longer compatible with the Run 1 and Run 2 framework
  - New calibrations and data validation requires (a lot of) time
- ALICE committed to publicly releasing level 3 scientific data
  - Dedicated human resources for the ALICE Open Data have been allocated
  - Run 1 data validated; preparing transfer to open data portal
  - More feedback on OD workflow in the coming months

Thank you!

ALICE