EXPLORE: A Scalable Infrastructure for LHC Open Data Analysis and FAIR Data Provisioning

Baida Achkar*, Arnulf Quadt, Sebastian Wozniewski

Georg-August-Universität Göttingen

*** Corresponding author: baida.achkar@phys.uni-goettingen.de**

PUNCHLunch Online Seminar

18.09.2025 - 12:30-13:30

PUNCH 4NFDI

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Part I – Motivation & Concept

Why EXPLORE? Lowering barriers to LHC Open Data

Why EXPLORE?
Barriers to Public Use of LHC Open Data

🔧 Technical Barriers

Specialized High Energy Physics (HEP) software & tools required to handle ROOT-format multi-GB datasets efficiently.

Limited runtime & computing resources constrain online analysis environments (e.g. SWAN, Binder).

🔑 Access Barriers

CERN credentials required to access certain advanced tools (e.g. EOS storage, SWAN).

Obscure & complex interfaces make data access & analysis tools difficult for non-experts to use.

Why EXPLORE?
Overcoming Barriers to Public Use of Open Data

EXPLORE's Role

💻 Technical Solution:

Ready-to-use containers with ROOT, libraries, & dependencies (No local install).

Batch mode on EXPLORE's distributed computing (GoeGrid Cluster) enables large-scale analyses.

🌐 Access Solution:

No CERN account required to use computing resources or access datasets & tutorials.

Preconfigured remote data access inside job containers (no manual setup).

Outcome with EXPLORE

Enables public & unaffiliated scientists to run real analyses without hardware, credential, or setup hurdles.

Provides researchers & educators a ready-to-use environment for both training & research.

Supports compute-heavy LHC analyses.

Serves as a bridge from Open Data in theory to Open Data in practice.

EXPLORE within PUNCH4NFDI & Göttingen GoeGrid

EXPLORE, part of **PUNCH4NFDI** initiative, promoting cross-disciplinary **FAIR** data practices in physics.

GoeGrid , a Tier-2 WLCG site at the University of Göttingen, historically serving ATLAS Monte Carlo production & analysis.

EXPLORE repurposes part of GoeGrid as "**Open Analysis Resources**" to serve **non-affiliated users**.

This enables LHC Open Data analysis **without the need for direct CERN/Institution affiliation.**

EXPLORE leverages Göttingen's computing resources, as **Open Analysis Resources**, enabling **LHC–CERN Open Data analysis** without requiring CERN or institutional affiliation.

EXPLORE Objectives

Bridging Open Data from theory to practice

Accessibility: Lower barriers for non-CERN & non-PUNCH users to access & analyze HEP data.

Education & Outreach: Provide hands-on LHC Open Data experience for high-school students, educators.

Scalability: Support diverse workloads with dynamic resource allocation.

FAIR Principles: Ensure data & tools are Findable, Accessible, Interoperable, and Reusable.

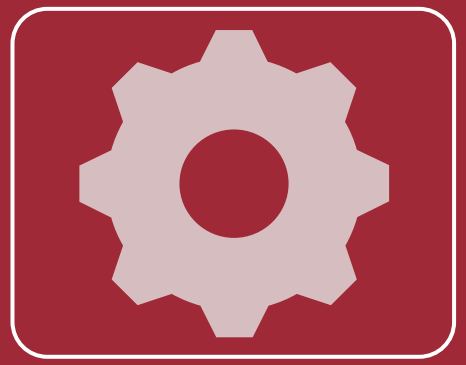
Sustainability: Leverage existing infrastructure with minimal overhead.

EXPLORE: Empowering global participation in high-energy physics research!

Part II – Infrastructure & Operation

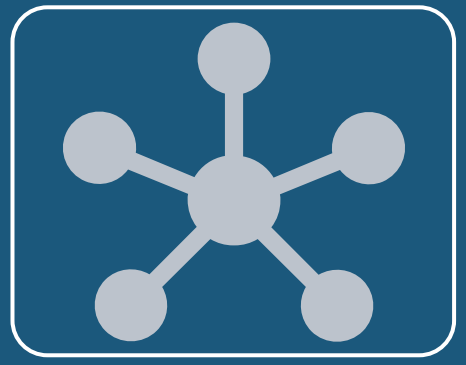
How EXPLORE works: batch system, containers, and monitoring!

Scalable Infrastructure for FAIR Open Data Analysis: Core Components



HTCondor Overlay Batch System (OBS):

- Aggregates distributed compute resources from GoeGrid Cluster into a unified job execution pool with dynamic scheduling.



Dynamic Resource Management (COBalD/TARDIS):

- Continuously monitors real-time workload demands & coordinates automatic scaling of compute resources to ensure optimal performance & cost-efficiency

Drone Architecture & Environments

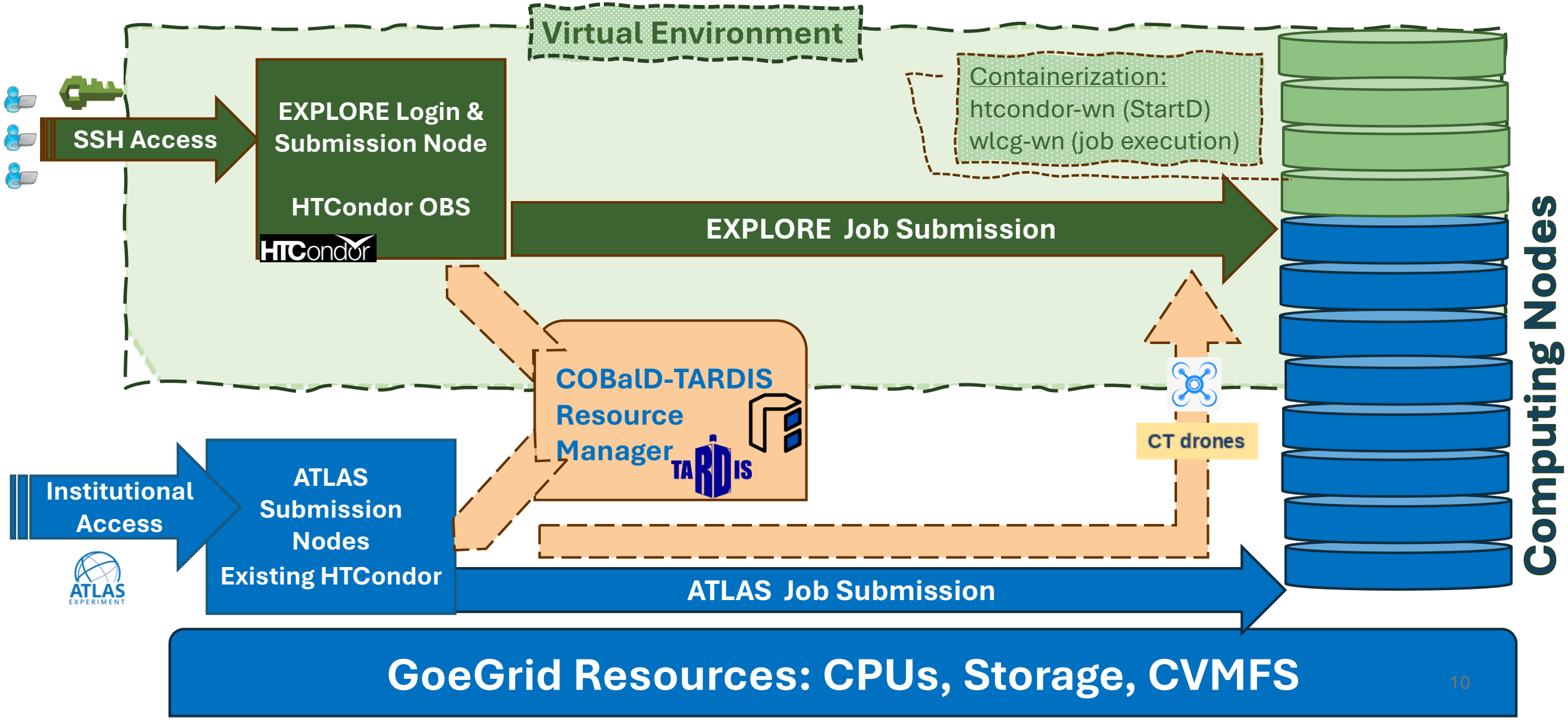
Drone Architecture & Scheduling:

- Placeholder jobs (“drones”) submitted, each runs an HTCondor StartD daemon & automatically registers as a Worker Node.
- Drones integrate into the OBS, forming a seamless, unified execution pool.
- HTCondor Scheduler (Schedd) manages job queuing, matching, & dispatch across dynamically provisioned resources, pairing queued jobs with available drones for efficient execution.

Drone Environments:

- Version-controlled containerized setups with ROOT, libraries, and dependencies, deployed dynamically at runtime for reproducibility using HTCondor tools.

EXPLORE setup within GoeGrid cluster-Göttingen



Software Environment Provisioning

Containerized Analysis Environments with **Apptainer** & **CERN CVMFS** provide standardized setups.

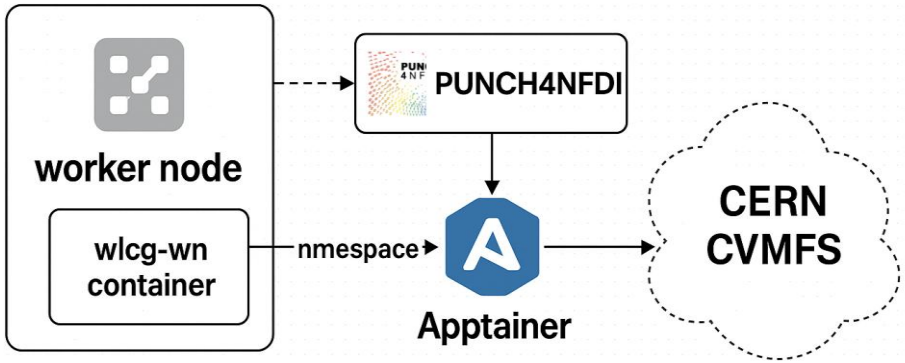
Jobs run inside **wlcg-wn Apptainer container**, defined in **PUNCH4NFDI Container Registry,** delivered unpacked via **CVMFS**.

CVMFS (CernVM File System):

Installed on all GoeGrid worker nodes.

Mounted via namespace to each drone (HTCondor-wn).

Automatic setup: HTCondor job configuration specifies the container, triggering retrieval & environment initialization at runtime.



Real-Time Monitoring with Prometheus & Grafana

Prometheus collects & stores real-time metrics from EXPLORE's Access/Submit Node.

Node_exporter gathers:

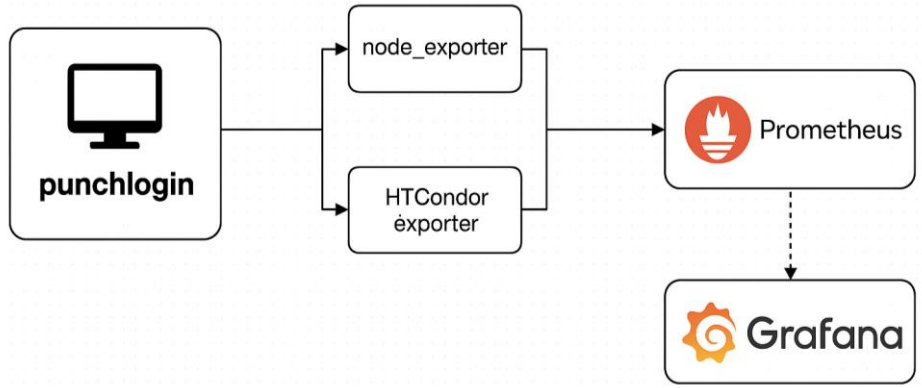
CPU, Memory utilization, Disk I/O, Network stats,..

Custom HTCondor exporter configured to track **HTCondor metrics** (queue status, execution stats,..)

Grafana provides the visualization layer:

Real-time dashboards for system health & HTCondor performance

→ **Proactive monitoring** and **issue detection**



Registration Workflow & User Access

🔍 **Fill out form (with valid Email)**

🔑 **Submit SSH key pair**

✅ **Receive confirmation & access instructions**

💻 **Log in via Submission Node:**

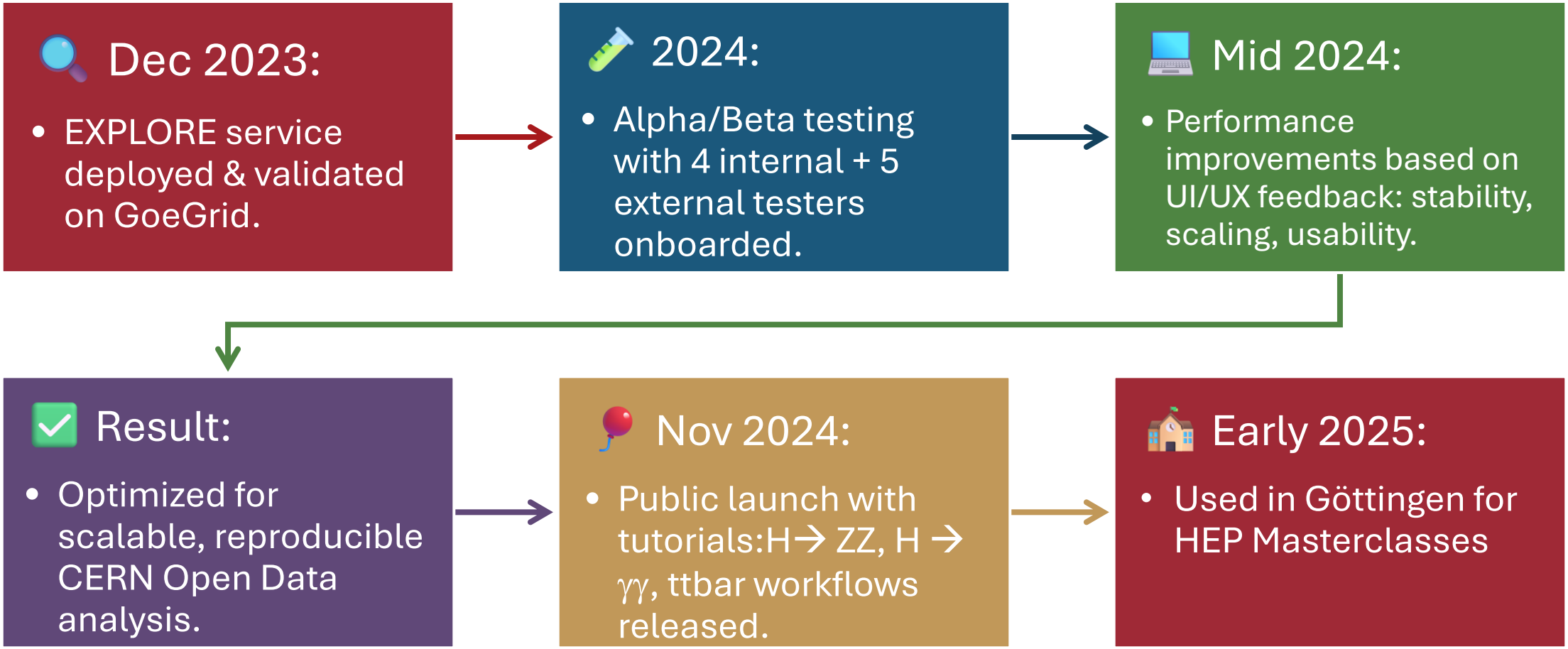
- ssh -i .ssh/id_rsa <username>@punchlogin.goegrid.gwdg.de

🔒 🌍 **Custom Authentication Model**

- Independent of institutional identity providers
- Enables flexible, open access for global users
- Secure, SSH-based login

👉 Register Now: https://punchlogin.goegrid.gwdg.de

From Prototype to Production: EXPLORE Adoption Journey



EXPLORE: From validation to public adoption in under 2 years.

Part III – Showcase: $t\bar{t}$ Analysis

Case study: demonstrating EXPLORE capabilities with ATLAS Open Data!

Why tt̄? Physics Motivation - l+jets Channel

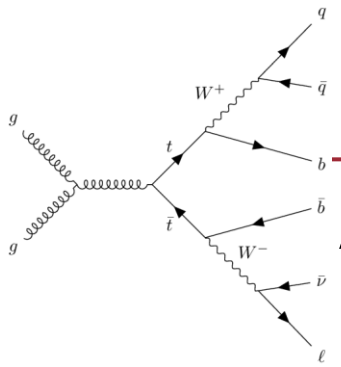
Top–antitop (tt̄) production is a key process at LHC:

- **Test the Standard Model** (QCD at high energies)
- **Search for new physics** beyond the SM
- **Background** for Higgs and BSM studies

Educational value:

- Realistic yet manageable complexity
- Rich variety of physics objects: leptons, jets, MET
- Ideal for teaching event selection, reconstruction, and statistical analysis

The l+jets Channel:



Final state:	1 high-p_t lepton (e/μ)
	≥4 jets (2 b-jets, 2 from W → qq')
	Missing transverse energy (ν from W → ℓν)
Advantages:	Large statistics, clean event signature
	Balance between complexity and yield
Dataset:	From **ATLAS Open Data 2025 Beta Release** (>100M events across samples)
	Includes MC simulations for validation and comparison

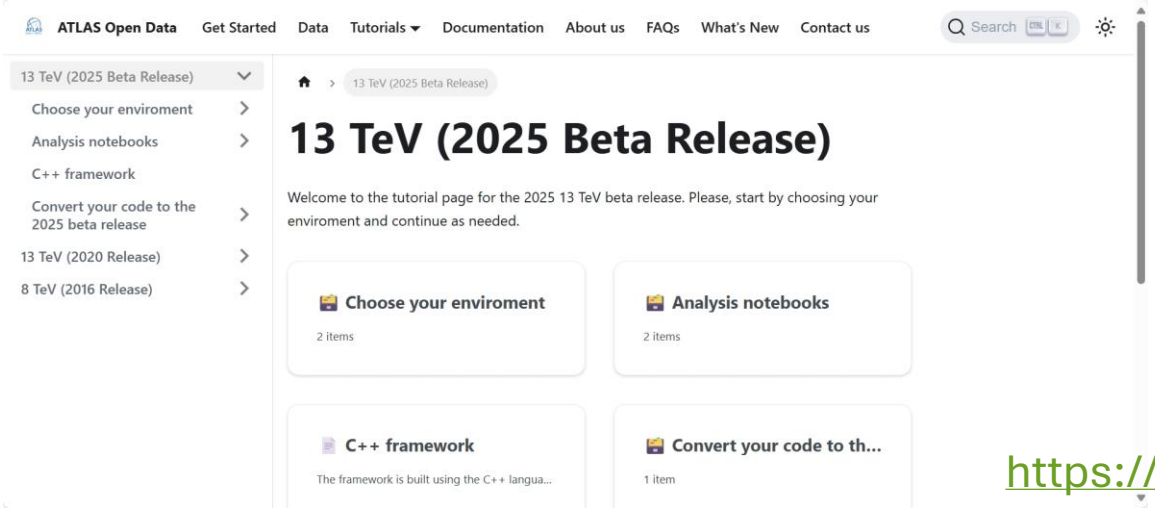
This analysis is implemented on EXPLORE-GoeGrid resources to demonstrate the platform's ability to process large datasets efficiently, reproducibly, and at scale.

ATLAS Open Data 2025 Beta

Released: February 2025

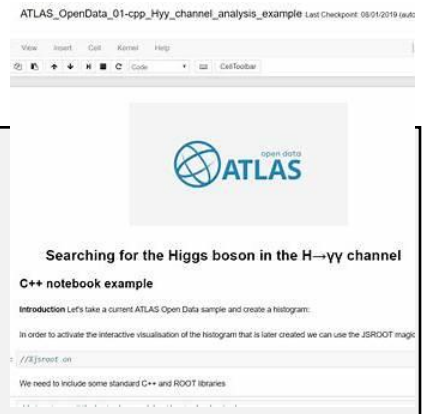
Release Overview

- 36 fb^{-1} of pp collisions at $\sqrt{s}$ = 13 TeV (2015–2016)
- For Education & Outreach, largest ATLAS educational release to date
- Provided as flat ROOT Ntuples (~80 branches) for ease of use
- Includes real collision data and Monte Carlo simulations
- Fully calibrated and ready for analysis



https://opendata.atlas.cern/docs/category/13-tev-2025-beta-release

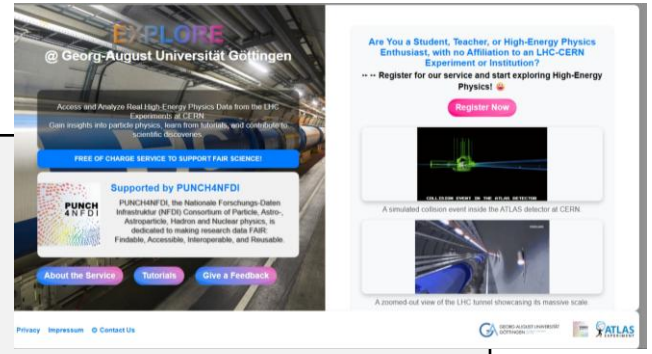
ATLAS Open Data & EXPLORE: Complementary Tools for Education and Research



ATLAS Open Data

Jupyter Notebooks (Interactive)

- Developed by the ATLAS Open Data group
- Designed for teaching and outreach
- Runs in interactive environments: CERN SWAN, Google Colab, Binder
- Ideal for small datasets and quick demonstrations
- Supports Python/C++ with ROOT libraries
- Covers Higgs, Z bosons, top quarks, W bosons, etc.
- Purpose: Learn concepts, data formats, analysis basics



EXPLORE Service

(Batch / Large-scale Analysis)

- Provided by Göttingen computing resources (GoeGrid → Emmy HPC by 2030)
- Designed for large datasets and complex workflows
- Uses HTCondor batch submission, no interactive execution
- Pre-configured environment and software stack for LHC analysis
- Scales to many CPU cores and long jobs without local setup
- Purpose: Realistic, high-statistics research-grade analyses

Workflow Summary & Execution

- Analysis: TTbarAnalysis (lepton + jets channel) based on C++ framework (latest)
- Data: ATLAS Open Data 2025 Beta release
- Platform: EXPLORE service – HTCondor Overlay Batch System
- Environment: punch4nfdi/wlcg-wn container with XRootD-based data access
- Scripts: Executable & Job description file for HTCondor submission
- Submission: 1 sample per job, 58 jobs in parallel

✓All 58 jobs completed successfully

✓First complete workflow validated for EXPLORE using the 2025 Beta release

✓Added to the EXPLORE material/resources portfolio

Quantitative Job Summary

Total jobs: 58, 35 Jobs with analyzed events ≥ 1,000,000

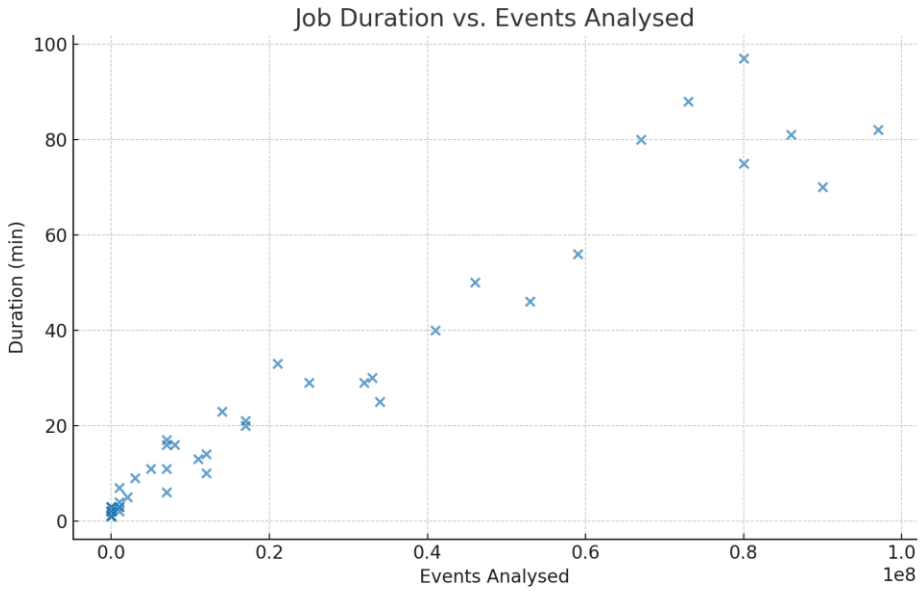
Jobs with <1,000,000 events: 23 (shown as 0)

Average duration: 20.0 min, Duration range: 1 to 97 min

Event range: 1,000,000 to 97,000,000 events

Outputs stored on EXPLORE storage (User's home directory)

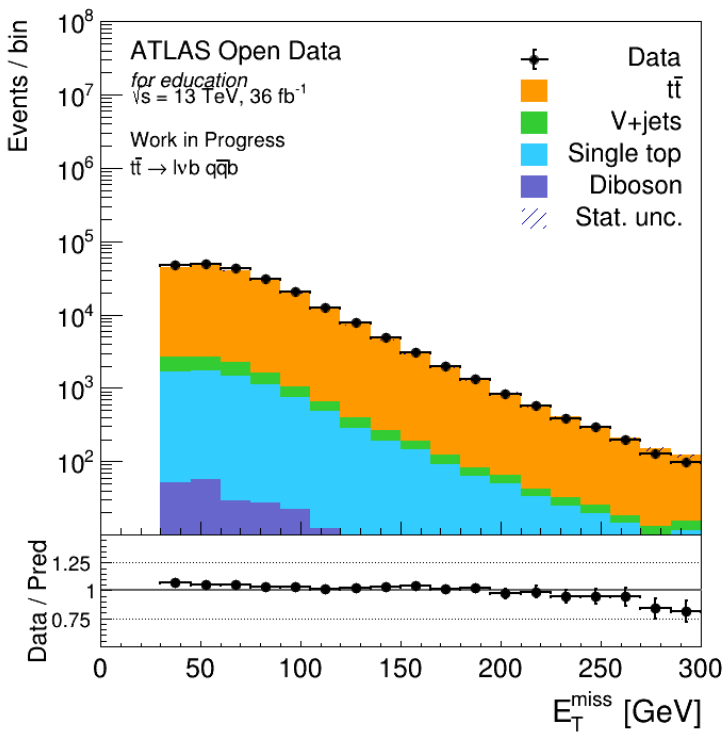
Output includes **histograms(396K)** and **ROOT files (1.1M)**



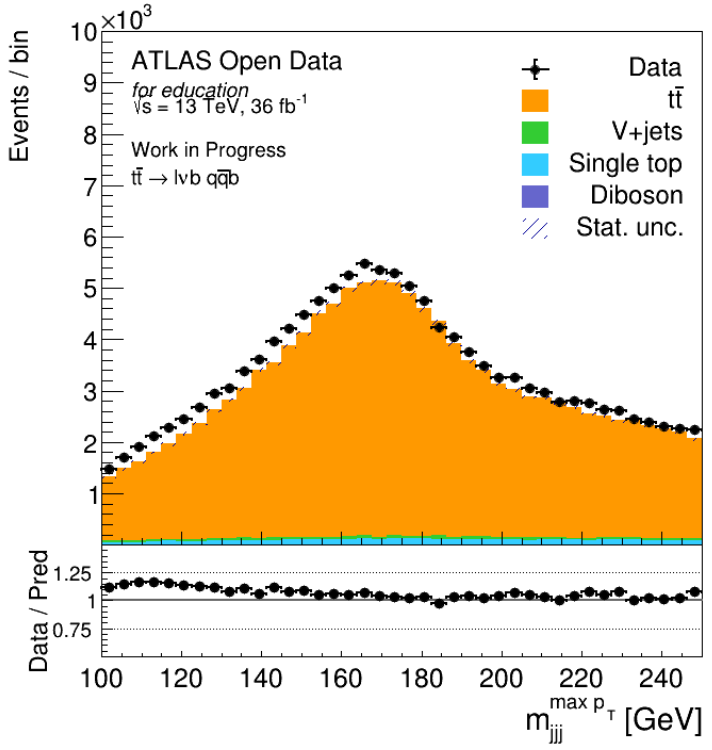
📌 This workflow demonstrates **reproducibility, scalability,** and **compatibility of ATLAS Open Data Beta Release with EXPLORE infrastructure**

Top-Antitop Analysis Plots

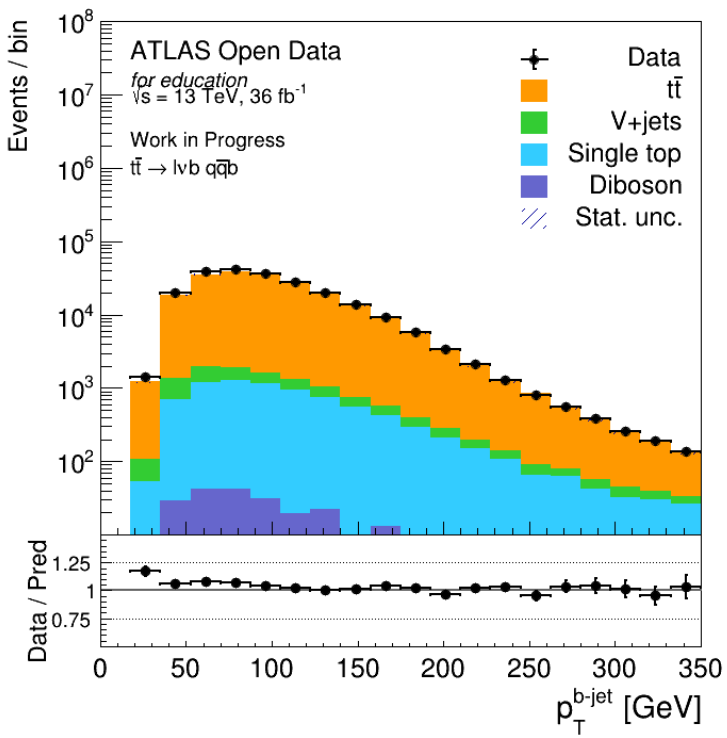
Missing transverse energy



Invariant mass of the three leading jets



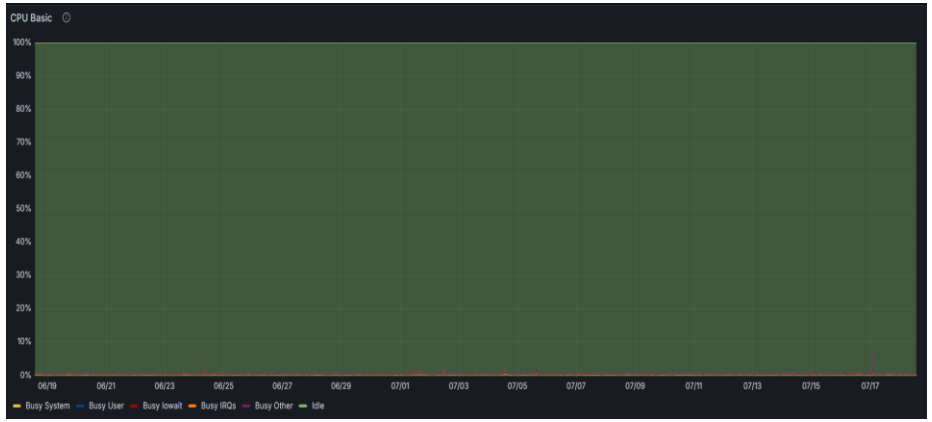
Transverse momentum of the leading b-jet



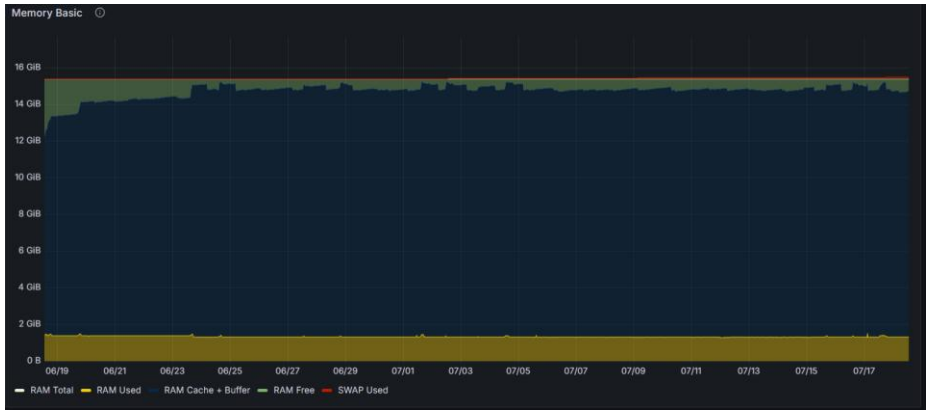
Part IV – Operational Insights & Outreach

What we learned: usage, performance, dissemination!

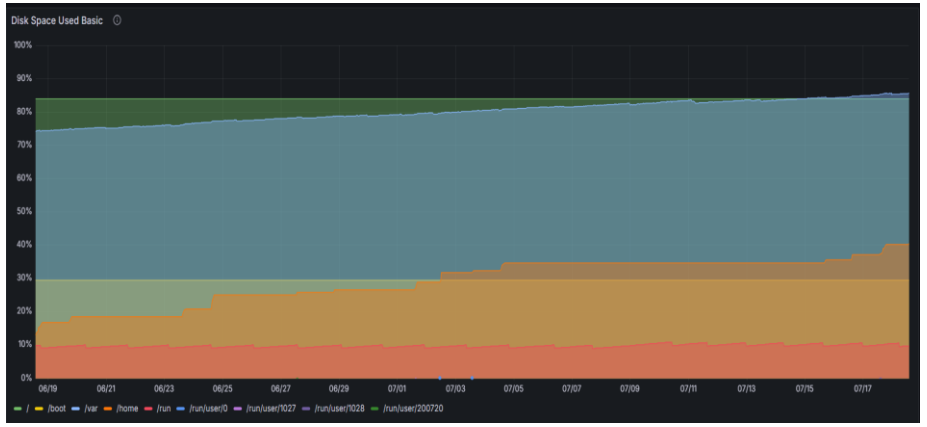
Submit Node Resource Footprint (CPU, Memory, Disk, Network)



CPU: efficient job scheduling and no overloads.



Memory: stability across time.



Disk: growing gradually, as users run jobs and store outputs.

Network traffic: healthy data transfer with occasional peaks, reflecting data access bursts.

Preliminary Usage Statistics & Operational Insights

- **Registration Page**: ~200 visitor
- **Users**: 14 (students from different fields, scientists)
- **Total Clusters Submitted @EXPLORE Access Point**: ~4,000
- **Jobs per Cluster**: 9 to 33 (varies by workflow)
- **Estimated Total Jobs**: ~36,000 to 132,000
- **Design**: Modular & parallel job execution = scalable performance
- **Datasets**: 130 real ATLAS data samples made available!
- **Key Insight**: User feedback has shaped UI, documentation, and system stability

Promotional Contribution: *EXPLORE Dissemination*

ATLAS Open Data Weekly Meetings

Regular service updates to the collaboration

Key contributions:

- July 18, 2024
- December 12, 2024

ATLAS Week Outreach

Outreach Parallel Session

Speaker: *Miguel Ángel García Ruíz* (for ATLAS Open Data team)

February 19, 2025

Link

Advertising & Outreach

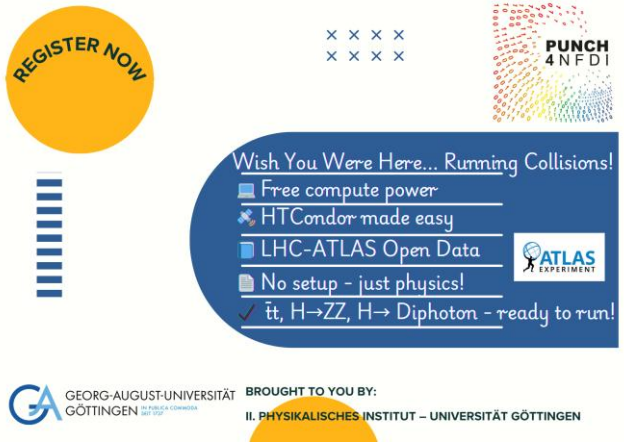
- PUNCH4NFDI Website.
- NFDI Newsletters.
- GAU Newsletters
- DPG Meetings
- Email/Letter Campaigns: Targeting Lower Saxony High Schools (# 59 Gymnasien) to expand accessibility.

Conference Contributions

CoRDI 2025 *(Conference on Research Data Infrastructure)*

Aachen, August 26–28, 2025

- Abstract



Part V – Conclusion & Impact

Takeaways: from open data in theory to open science in practice!

Conclusion & Impact

Conclusion

- EXPLORE removes long-standing technical and access barriers to ATLAS Open Data.
- By eliminating the need for CERN or university credentials, it opens participation to high school students, educators, and independent researchers.
- Ready-to-use containers, distributed computing, and remote data access make real analyses possible without specialist infrastructure.

Impact for Target Users

- **High School & University Students:** Early exposure to real HEP data and analysis workflows.
- **Educators:** Ready-made environment for teaching particle physics data analysis.
- **Researchers Without Institutional Affiliation:** Access to computing resources and datasets without credential barriers.

Key Takeaways – EXPLORE Service

✓ Simplifies complex technical setups into accessible, batch-based workflows.
✓ Supports compute-heavy analyses beyond online service limits.
✓ Acts as a bridge between Open Data in theory and practical, hands-on analysis.
✓ Positions LHC Open Data as an inclusive resource for education, outreach, and independent research.

Part VI – Future Scope

Sustainability & Integration of EXPLORE into PUNCH-2.0

Migration & Integration Roadmap

Transition contingent on PUNCH-2.0 approval & funding

- EXPLORE @ GoeGrid → phase-out by **2030**
- Transition to **NHR EMMY (GWDG Göttingen, NHR Alliance)**
  - *Provided PUNCH-2.0 gets approved and funded..*
- Ensures:
  - **Scalability** – national HPC resources
  - **Sustainability** – long-term, federated ops
  - **Integration** – PUNCH-2.0 & FAIR ecosystem

Towards Inclusive & Sustainable Access

AAI migration contingent on new features for unaffiliated & under-18 users

- Migration to **PUNCH AAI infrastructure**
  - *Provided AAI extensions for unaffiliated & young users are implemented*
- **Requirements** before migration:
  - ***Email validation*** *for unaffiliated users*
  - ***Parental consent & GDPR compliance*** *for <18*
- Renewable accounts beyond 3-month limit

Goal: EXPLORE accessible via PUNCH Portal → lasting, open entry point to LHC Open Data

Time for Questions

THANK YOU

Acknowledgements

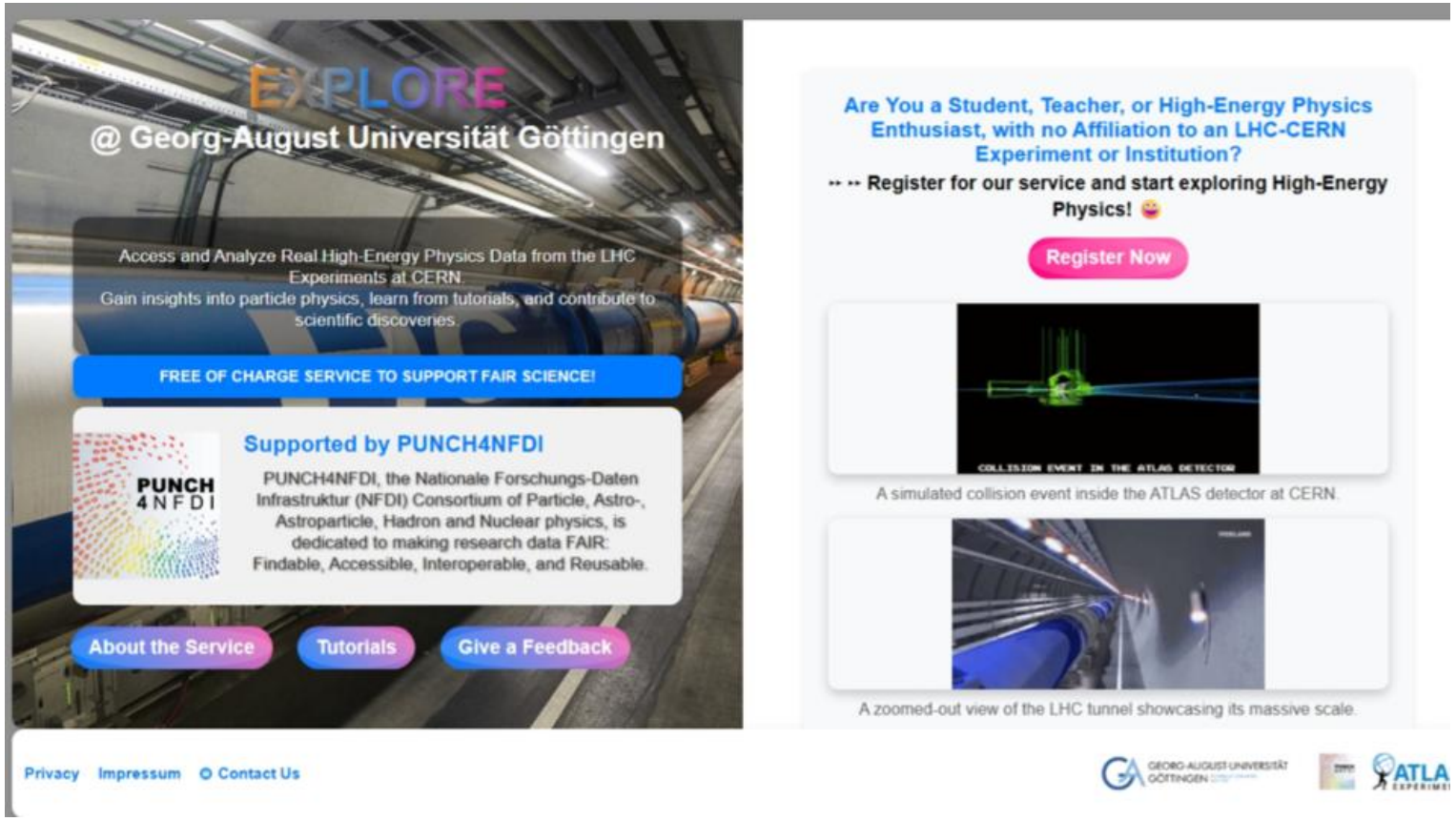
This work was [in part] supported by DFG fund „NFDI 39/1“
for the PUNCH4NFDI consortium.

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Backup

EXPLORE: A Scalable Infrastructure for LHC Open Data Analysis and FAIR Data Provisioning

Access and Analyze LHC Open Data Using EXPLORE-GoeGrid Resources for All Users



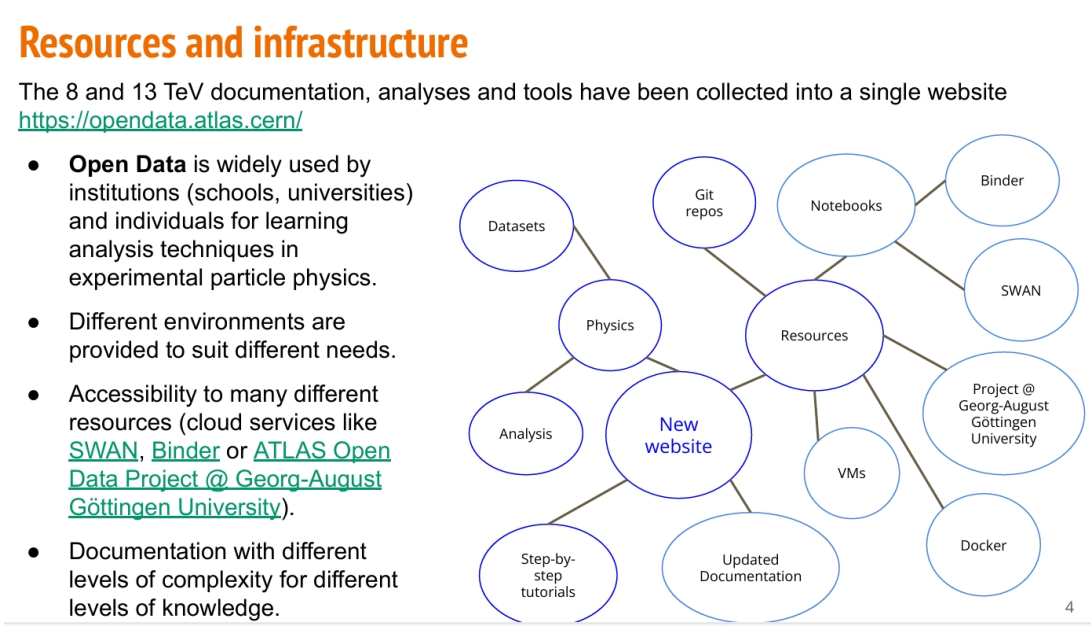
EXPLORE Service @ GAU Göttingen Promoting!

ATLAS Week & Open Data Weekly Meetings

- **ATLAS Open Data Weekly Meetings:** Regular Updates
  - Contribution on 18.07.2024
  - Contribution on 12.12.2024
- **ATLAS Week Outreach:** The service was presented during the **ATLAS Week Outreach parallel session** *(Speaker: Miguel Ángel García Ruíz on behalf of the ATLAS Open Data team)*.
- ATLAS WEEK Outreach Parallel Session February 19, 2025

Advertising & Outreach

- **GAU Newsletters:** Promoting the service within academic networks (**Approximately 2,000 recipients**.).
- **Email/Letter Campaign:** Targeting **Lower Saxony High Schools (# 59 Gymnasien)** to expand accessibility.



Latest releases of ATLAS Open Data for Education and Research by Prof. Farid Ould-Saada (University of Oslo (NO) at **The fourth African Conference of Fundamental and Applied Physics, ACP2025, will be a hybrid event, jointly organized by ASP and the University of Lomé, Togo, from September 14 to 20, 2025:** https://indico.cern.ch/event/1458227/

New release of ATLAS open data

- ATLAS Open Data for Research
  - Particularly useful for theoreticians but not only
  - DAOD_PHYSLITE format 2015-2016 Open Data for Research
    - CERN Open Data Portal. DOI:10.7483/OPENDATA.ATLAS.9HK7.P5SI
      - There are 3 publications referring to these data
  - Dataset characteristics: **9'058'437'931 events**. **70'611 files**. **65.3 TiB** in total.
  - A framework allows processing the research data for analysis, see also Zenodo

Heavy Ion data stored as DAOD_HION14 (similar to PHYSLITE)

PHYSLITE format:

- *designed to efficiently manage and analyze large datasets generated in ATLAS at high LHC luminosities*
- *same format (ROOT xAOD) as all reconstructed data & simulations from the experiment*

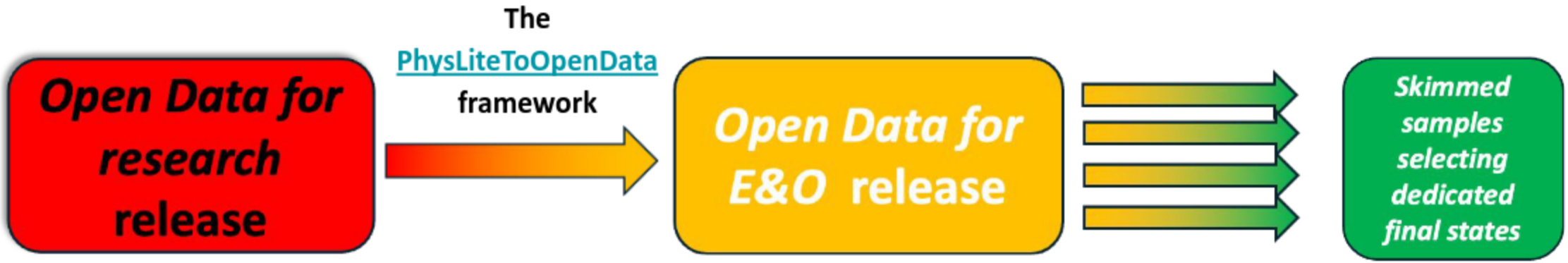
Latest releases of ATLAS Open Data for Education and Research by Prof. Farid Ould-Saada (University of Oslo (NO) at **The fourth African Conference of Fundamental and Applied Physics, ACP2025, will be a hybrid event, jointly organized by ASP and the University of Lomé, Togo, from September 14 to 20, 2025:** https://indico.cern.ch/event/1458227/

New release of ATLAS open data

- ATLAS Open data for education
  - ROOT ntuple format 2015-2016 proton-proton Open Data for Education and Outreach beta release from the ATLAS experiment
  - A framework allows processing the research data to produce ROOT ntuples, see also Zenodo
  - CERN Open Data Portal. DOI:10.7483/OPENDATA.ATLAS.B5M9.44TN
  - Dataset characteristics: **9'837'961'169 events**, **4668 files**, **2.5 TiB** in total.

Latest releases of ATLAS Open Data for Education and Research by Prof. Farid Ould-Saada (University of Oslo (NO) at **The fourth African Conference of Fundamental and Applied Physics, ACP2025, will be a hybrid event, jointly organized by ASP and the University of Lomé, Togo, from September 14 to 20, 2025:** https://indico.cern.ch/event/1458227/

Open Data for Education and Outreach



- Total: **65 TB**
  - 36 fb^{-1} of data recorded in 2015 and 2016
  - 2 billions of simulated events
- **Data format**: PHYSLITE

- Total: **2 TB**
- Research samples with only very basic cuts such as event cleaning, good run lists, overlap removal, … 36 fb^{-1} of data recorded in 2015 and 2016
- **Data format**: ROOT Ntuples

- Selected data samples from **~1.5 GB to ~350 GB**
- Targeted analyses
- **Data format**: ROOT Ntuples

Datasets available for Open Data for Education

- Several final-state collections specifically tailored towards physics analysis example notebooks
- Original full 36 fb^{-1} samples are also available

Selection	Collection Name
At least one lepton with at least 7 GeV of p_T and 30 GeV of missing transverse momentum (i.e. a leptonically-decaying W-boson enhanced selection)	1LMET30
Two to four leptons with at least 7 GeV of p_T each	2to4lep
At least two muons with at least 10 GeV of p_T (i.e. a leptonically-decaying Z-boson enhanced selection)	2muons
At least three jets with at least 20 GeV of p_T, at least one lepton passing tight identification requirements with at least 7 GeV of p_T, and 30 GeV of missing transverse momentum (i.e. a semi-leptonic top-quark enhanced selection)	3J1LMET30
At least two photons with at least 25 GeV of p_T each (i.e. a Higgs boson decaying to two photons enhanced selection)	GamGam
At least two jets with at least 20 GeV of p_T, at least two leptons passing tight identification requirements with at least 7 GeV of p_T, and 30 GeV of missing transverse momentum (i.e. a di-leptonic top-quark enhanced selection)	2J2LMET30
At least two jets with at least 20 GeV of p_T identified as containing at least one heavy flavor hadron using the 70% working point (i.e. a Higgs boson decaying to b-quarks enhanced selection)	2bjets
At least three leptons with at least 7 GeV of p_T each	3lep
Exactly three leptons with at least 7 GeV of p_T (i.e. a leptonically-decaying W+Z boson enhanced selection)	exactly3lep
At least four leptons with at least 7 GeV of p_T each	4lep
Exactly four leptons with at least 7 GeV of p_T (i.e. a leptonically-decaying ZZ boson or Higgs to four leptons enhanced selection)	exactly4lep

Latest releases of ATLAS Open Data for Education and Research by Prof. Farid Ould-Saada (University of Oslo (NO) at **The fourth African Conference of Fundamental and Applied Physics, ACP2025, will be a hybrid event, jointly organized by ASP and the University of Lomé, Togo, from September 14 to 20, 2025:** https://indico.cern.ch/event/1458227/

Tutorials, Repositories, and Analysis Framework

Public GitLab Repository
Hosts code, frameworks, and step-by-step analysis guides

Pre-built Analysis Templates:
HZZ, TTbar, Hyy
C++ framework with ready-made scripts

Guides Included:
- Accessing ATLAS Open Data Portal
- Running analysis scripts
- Plotting & interpreting results

Educational Focus:
Tailored for both novices and experienced users