Data and software preservation in H1

Daniel Britzger for the H1 Collaboration
Max-Planck-Institut für Physik München, Germany

DPHEP collaboration meeting, 5.3.2026

MAX-PLANCK-INSTITUT
FÜR PHYSIK

Home | CERN — Mozilla Firefox

Home | CERN

https://home.cern

Search

CERN Accelerating science

Sign in Directory

CERN

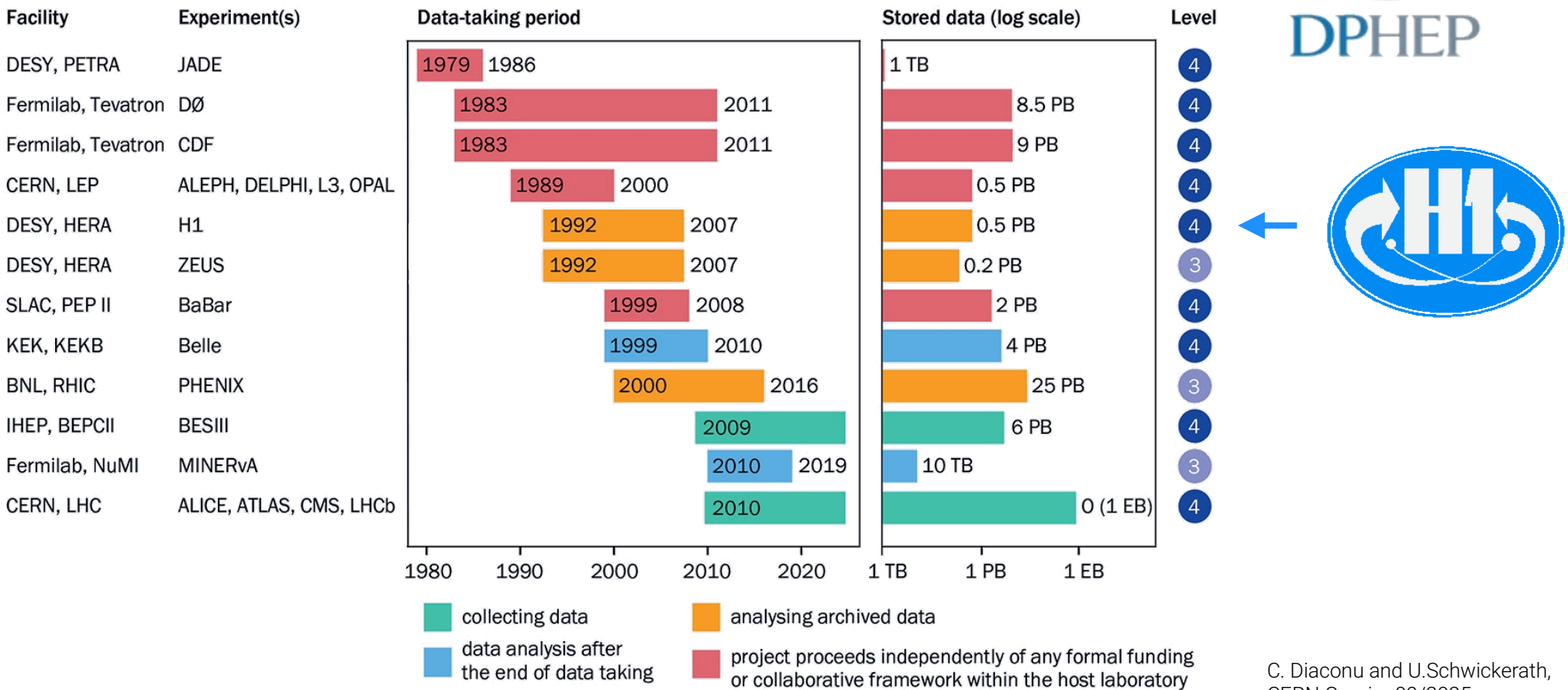
ABOUT NEWS SCIENCE RESOURCES SEARCH | EN

Preserving particle physics data

Data preservation allows physicists to unearth hidden treasures long after collisions end

Read more →

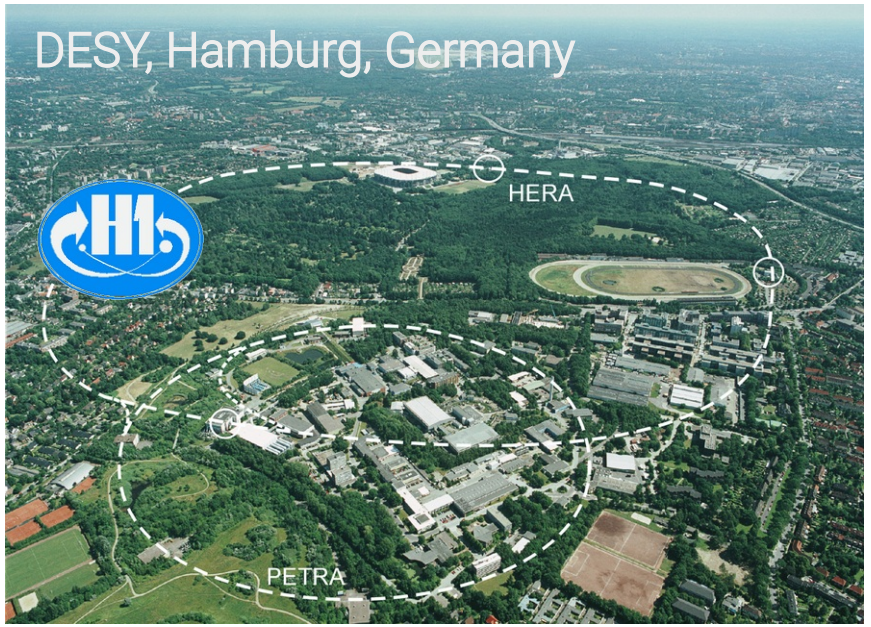
DPHEP



C. Diaconu and U.Schwickerath, CERN Courier 09/2025

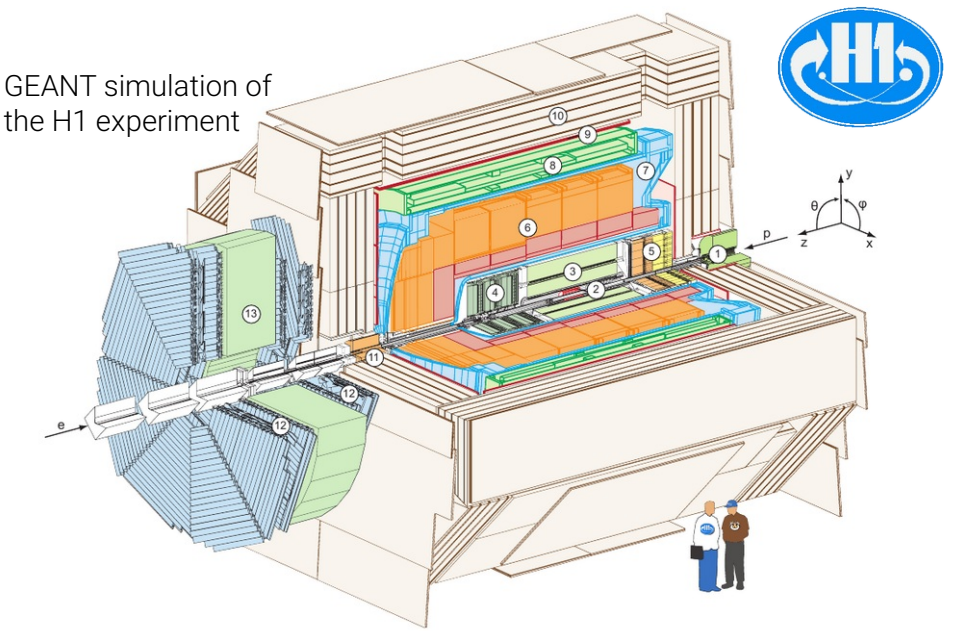
The H1 experiment at HERA

HERA electron-proton collider at DESY



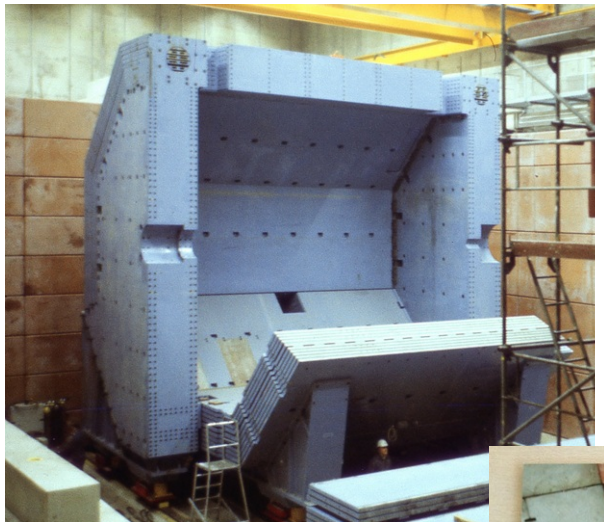
- HERA I: 1994 – 2000
  HERA II: 2003 – 2007
- E_e=27.6 GeV, E_p=920GeV
  √s = 300 or 319 GeV

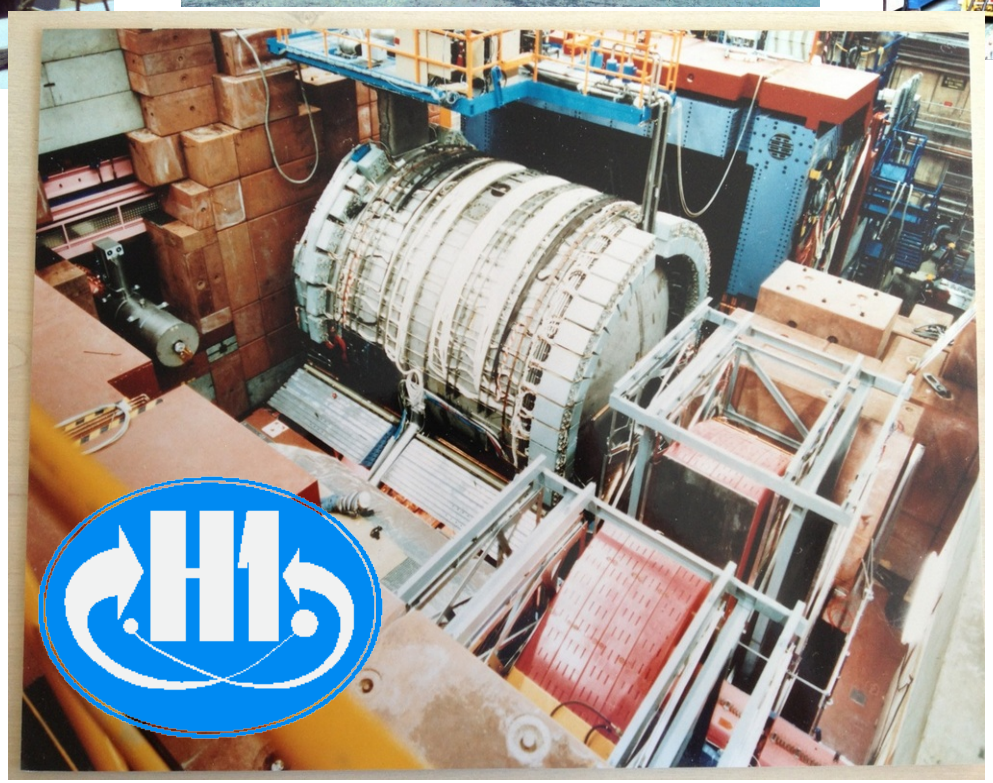
H1 experiment at HERA

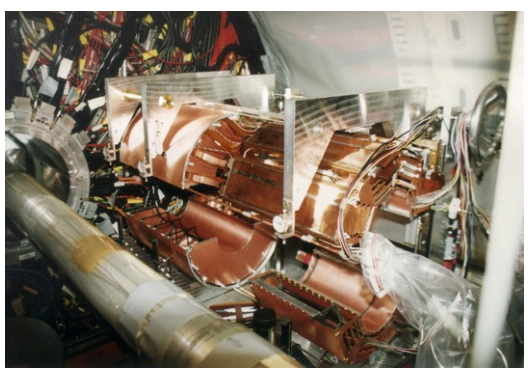


'multi-purpose' detector

- Asymmetric design with trackers, calorimeter, solenoid, muon-chambers, forward & backward detectors, ...
- 270,000 readout channels

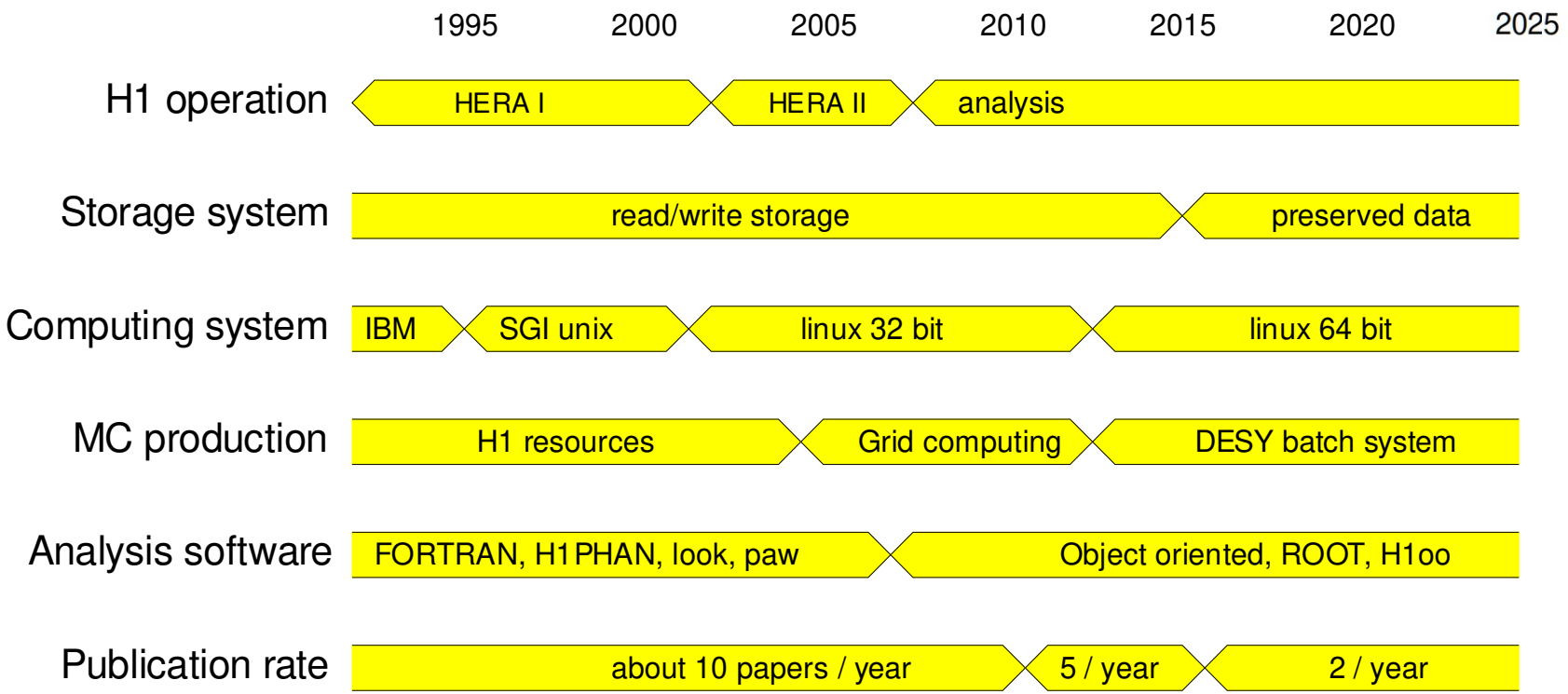



H1

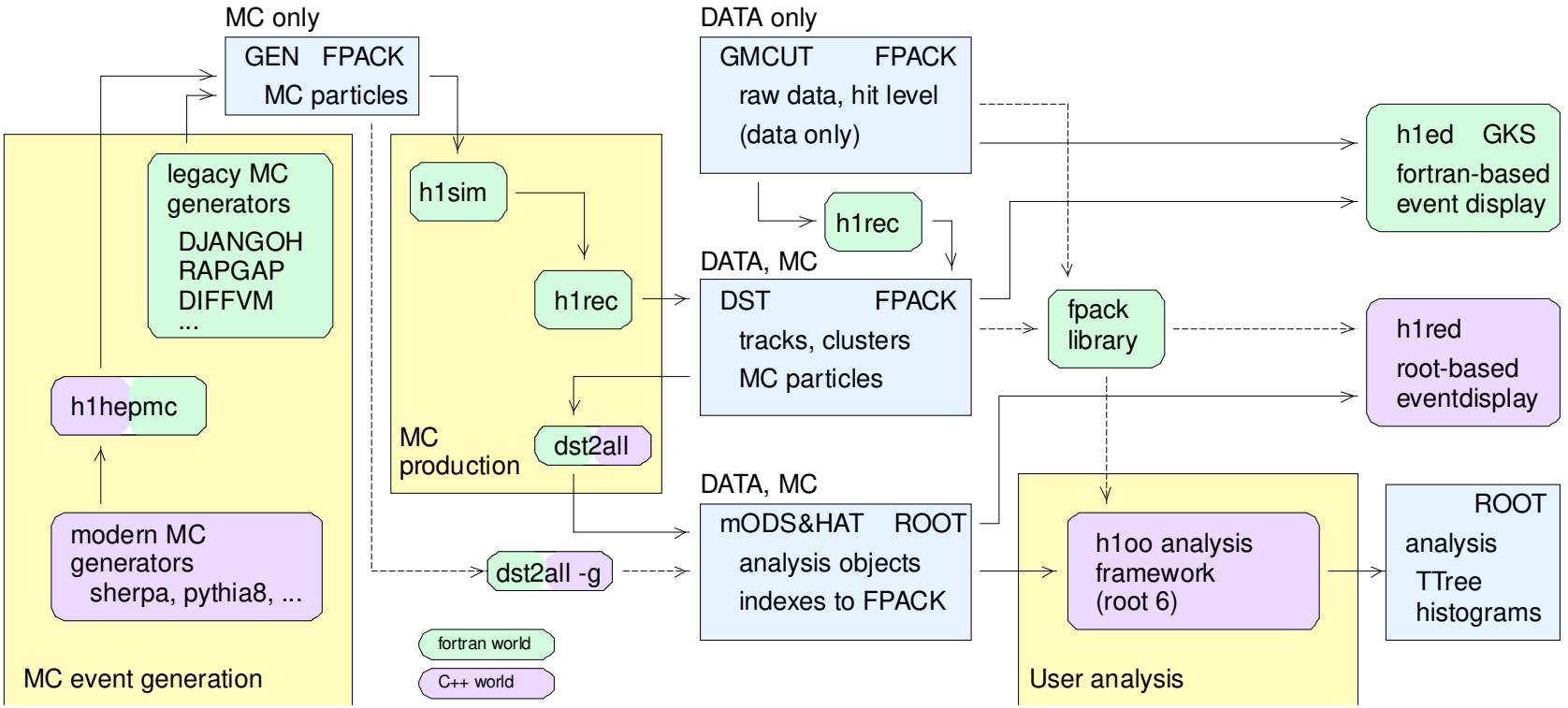


The H1's evolutions over time

- √s=320 GeV *ep* collider HERA at DESY, data taking 1992-2007
- Once 400 collaboration members, today still ~150.
- H1 has seen many different computing environments



H1 storage and analysis model



- Various MC event generators
- Detector simulation “h1sim”: based on GEANT3, written in Fortran
- Reconstruction “h1rec”: written in Fortran
- User analysis in C++, wrapping ROOT and on-the-fly access to Fortran objects (hits, track segments, etc)

Data preservation of the H1 experiment

H1 is involved in DPHEP since the beginning

Preservation Model	Use case
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

taken from DPHEP paper (arXiv:0912.0255)

H1 adopted a 'level 4' preservation model

- Preserve not only analysis level data, but also reconstruction and simulation software as well as the basic level data
- Retain the full flexibility and potential of the experimental data

➞ For full access to the data, the software must also be considered

Data preservation (the H1 data themselves)

- ~1 billion *ep* events, Total RAW data: 75TB; Final re-processed data (DST): 20TB; Analysis "H1oo": 4TB, other special data, full set of MC samples,
  → total data volume about 0.5PB
- Data organised in a dedicated DPHEP storage at DESY (dCache)

The FORTRAN 'core' packages and H1oo

The 'core'-software packages (FORTRAN)

- H1 core software written in FORTRAN 77
  [NIM A A386 (1997) 310]
- first developed in 1988
  already with clear structure:
  highly modular, and based on BOS/FPACK
  [S. Egli 1990; V. Blobel 1990; V. Blobel CHEP1992]
- Programs for: Data storage and I/O, simulation (based on GEANT3), reconstruction and post-processing, visualisation (based on LOOK), data analysis, etc…
- MC generators
- External dependencies:
  CERNLIB, GEANT3, GKS, oracle-instant
- about 950k lines of code

H1oo

- H1**oo**: **o**bject-**o**riented C++ common analysis framework based on ROOT
  [U. Berthon et al. CHEP2000; M. Peez et al. CHEP2003,
  J. Katzy et al. CHEP2005; M. Steder et al. CHEP2012]
- written in C++98, and until recently based on ROOT 5.34
- About 50 packages and 600 classes; analysis environment and data formats for analysis
- External dependencies:
  ROOT, fastjet, neurobayes-expert
- Standardised H1 data analysis and benefit from expert knowledge
- about 300k lines of code

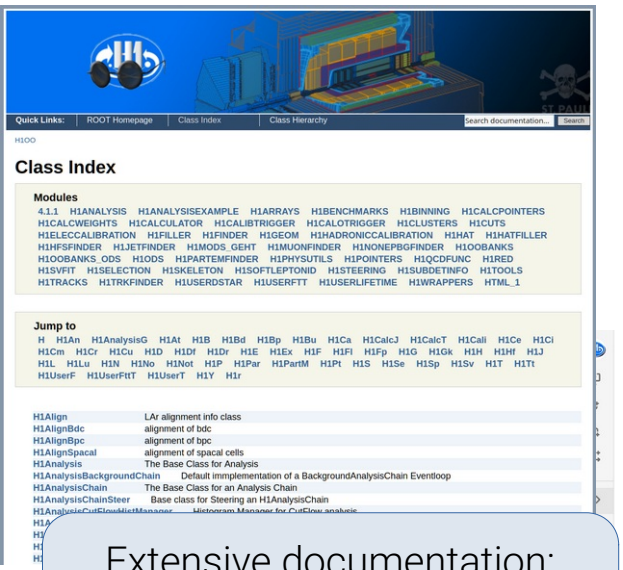
DPHEP developments

- 2015: transition to DPHEP storage (read-only), selected data and MC set,
  Clean-up of internal documentation, digitalization of analog documents, transition to static webpages
- 2017: transition from SL5 to SL6 (32 to 64 bit)
- 2019: revive GKS-based event display
- 2020: large-scale modernisation of software stack, including gcc9, C++20,
  - ROOT 5.34 → ROOT6; CINT → CLING (Root's C++ interpreter, deeply integrated in h1oo)
  - C++20, pythonic analysis, LCG compatibility, transition from SL6 to SL7
  - from svn → git
- 2025: transition to Alma9 completed, gcc14.2 & gfortran 14.2
  git/stash → gitlab, ROOT 6.34
- 2026: Adaptation of public H1 webages to modern data protection and security requirements
- Also can run SL5 - SL7 binaries in containers

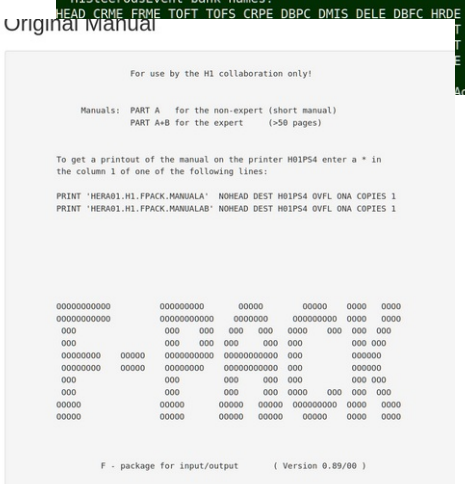
Examples with recent software release

Entire software is relocatable and globally available (here: lxplus@CERN)

```
[lxplus750] ~/test $ minimum_example -f $PWD/minimum_examp

Info in <H1ErrorHandler::H1ErrorHandler>: Using H1ErrorHand
<H1ErrorHandler::SetMaxCount> Errors printed 1 times
H1SteerManager: Searching for file /afs/cern.ch/user/b/brit
H1SteerManager: Searching for file /afs/cern.ch/user/b/britzger/test/minimum_example.steer
H1SteerManager: File /afs/cern.ch/user/b/britzger/test/minimum_example.steer opened for reading

======== H1SteerManager: Reading from file '/afs/cern.ch/user/b/britzger/test/minimum_example.steer' =============
----- H1SteerTree(){
----- fHatFiles="hat.4.0.6.DJANGOH14.NC.EPLUSP.0607.RAD.Q2GT60.CTEQ6L.A.7091.S39200.R97800.ftt.DST.0000.root";
----- fModsFiles="mods.4.0.6.DJANGOH14.NC.EPLUSP.0607.RAD.Q2GT60.CTEQ6L.A.7091.S39200.R97800.ftt.DST.0000.root";
----- }
====== H1SteerManager: Accepted values from file '/afs/cern.ch/user/b/britzger/test/minimum_example.steer' ======
============================================================
Steering for H1SteerTree
============================================================
fModsFiles      = 1 entries
fHatFiles       = 1 entries
fTreeCacheSize  = 1000000
============================================================
======== H1SteerManager: Done reading from file '/afs/cern.ch/user/b/britzger/test/minimum_example.steer' ========

===> Addfile for file mods.4.0.6.DJANGOH14.NC.EPLUSP.0607.RAD.Q2GT60.CTEQ6L.A.7091.S39200.R97800.ftt.DST.0000.root of type MODS
===> Addfile for file hat.4.0.6.DJANGOH14.NC.EPLUSP.0607.RAD.Q2GT60.CTEQ6L.A.7091.S39200.R97800.ftt.DST.0000.root of type HAT
Consistency checks for list H1TreeEventList

====== H1SteerManager: Using default values for class 'H1SteerOdsEvent'  ======
  H1SteerOdsEvent bank names:
HEAD CRME FRME TOFT TOFS CRPE DBPC DMIS DELE DBFC HRDE DT
```



Extensive documentation: README's, Gitlab markdown, int-notes, PDF-manuals, etc...



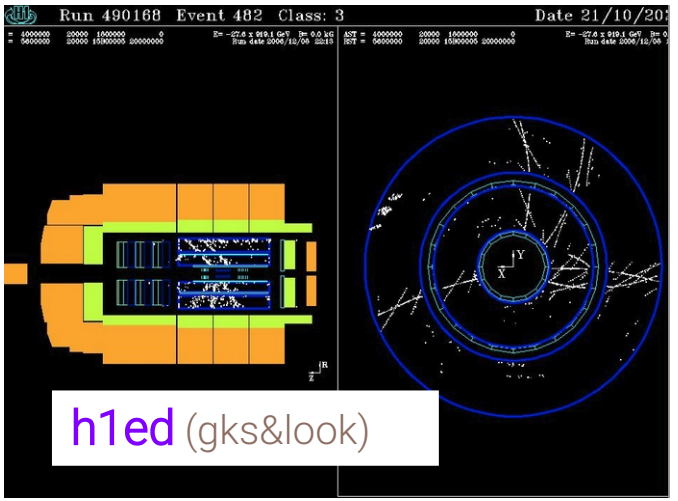

```python
# A minimum example to read H1 MOD files
# with python and write the results into
# a ROOT.TH1D histogram and plot it
def minimum_example():
    tree = H1.H1Tree.Instance()
    tree.AddFile("/pnfs/desy.de/dphep/online/h1/mc2/oo-4.0/djangoh14/7091/hat.4.0.6.DJANGOH14.NC.EPL
    tree.AddFile("/pnfs/desy.de/dphep/online/h1/mc2/oo-4.0/djangoh14/7091/mods.4.0.6.DJANGOH14.NC.EP

    # direct access to MOD quantities using H1 Pointers
    Q2_ptr  = H1.H1FloatPtr("Q2e")           # virtuality
    Wgt_ptr = H1.H1FloatPtr("Weight1")       # Event weight
    PartCands_ptr = H1.H1PartCandArrayPtr()  # Array pointer to particle 'candidates' (can be static

    # book some ROOT-histograms
    hist_Q2 = ROOT.TH1F("Q2", "Q^{2} Detector level;Q^{2} [GeV^{2}];events",  40, 10, 1000);
    hist_Pt = ROOT.TH1F("pt", "P_{T} Particles;P_{T} of all PartCands [GeV];events",  40, 0, 10);
    hist_VtxZ = ROOT.TH1F("VtxZ", "Vertex z-position; Vertex z-position [cm];events",  40, -55, 55);
    hist_Empz = ROOT.TH1F("Empz", "Empz; E - P_{Z} of FS [GeV];events",  40, -0, 100);
    # --- H1Calculator, if requested
    gH1Calc = H1.H1Calculator.Instance()

    # --- event loop
    events = 0;
    while tree.Next() and events < 10000:

        # acces Q2, Wgt, etc...
        Q2  = Q2_ptr[0]
        Wgt = Wgt_ptr[0]
        # fill histogram
        hist_Q2.Fill(Q2,Wgt);
```

Full 'pythonic' H1 analysis possible

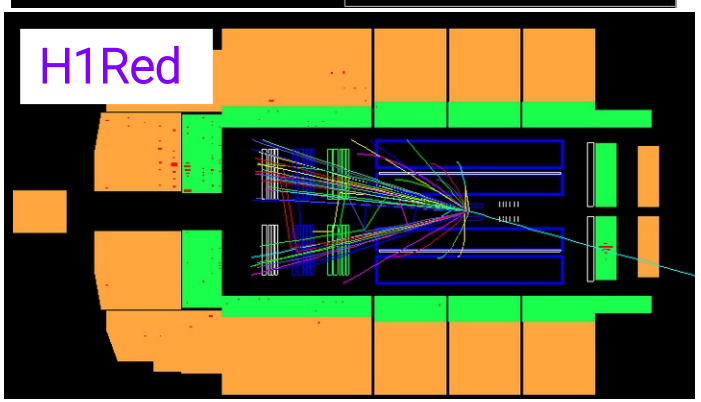
Pythia8.3 with H1 simulation

Proof-of-concept for new generators in H1

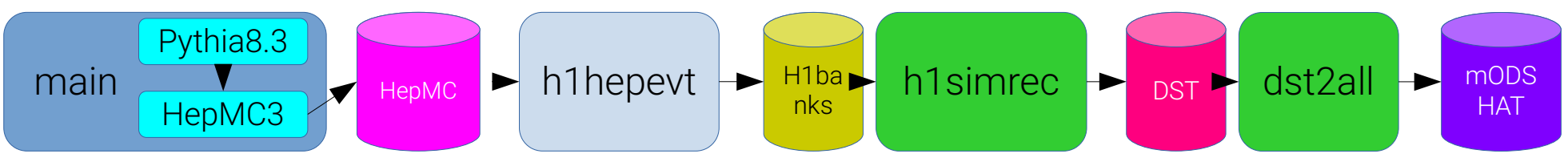
- Generator comes from *LCG* without H1-specific installation
- 'main'-routine generates HepMC3 record
- HepMC-record is converted into H1gen format
- H1gen-format is simulated with h1simrec
- Simulated DST-files are converted to mODS and HAT files with dst2all
- Chain works, but some details get lost (HEPEVT)



H1banks



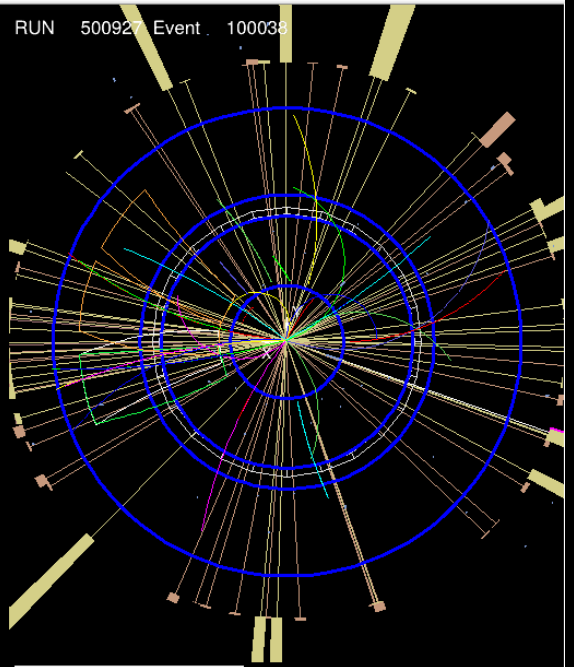
mODS HAT

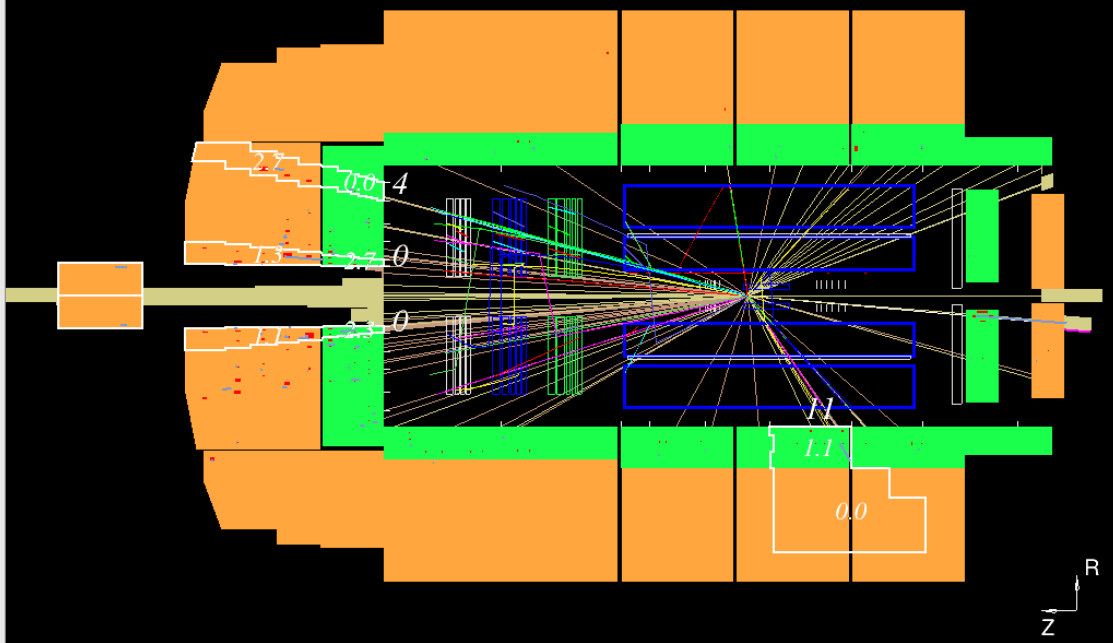


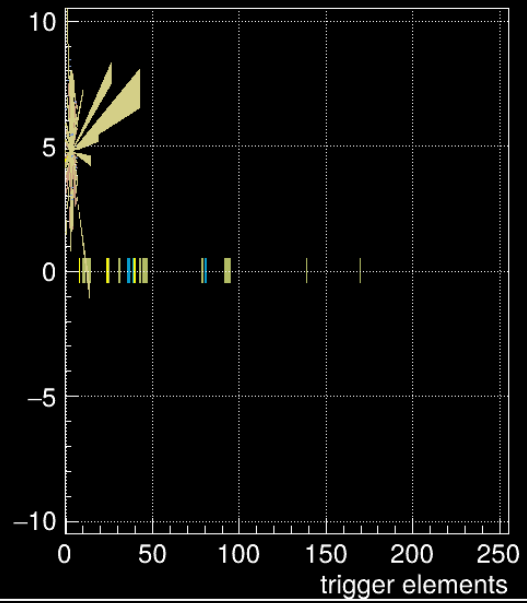
RUN 500927 Event 100038 MC date 25/04/13 14:17
DSN=

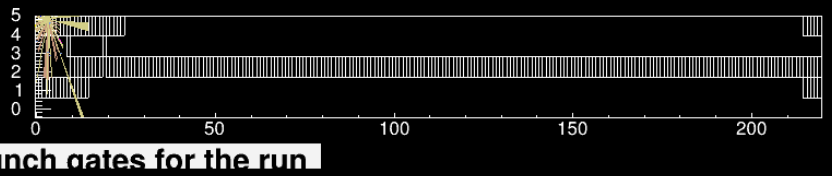
E= -27.6 x 920.0 GeV B=11.6 kG

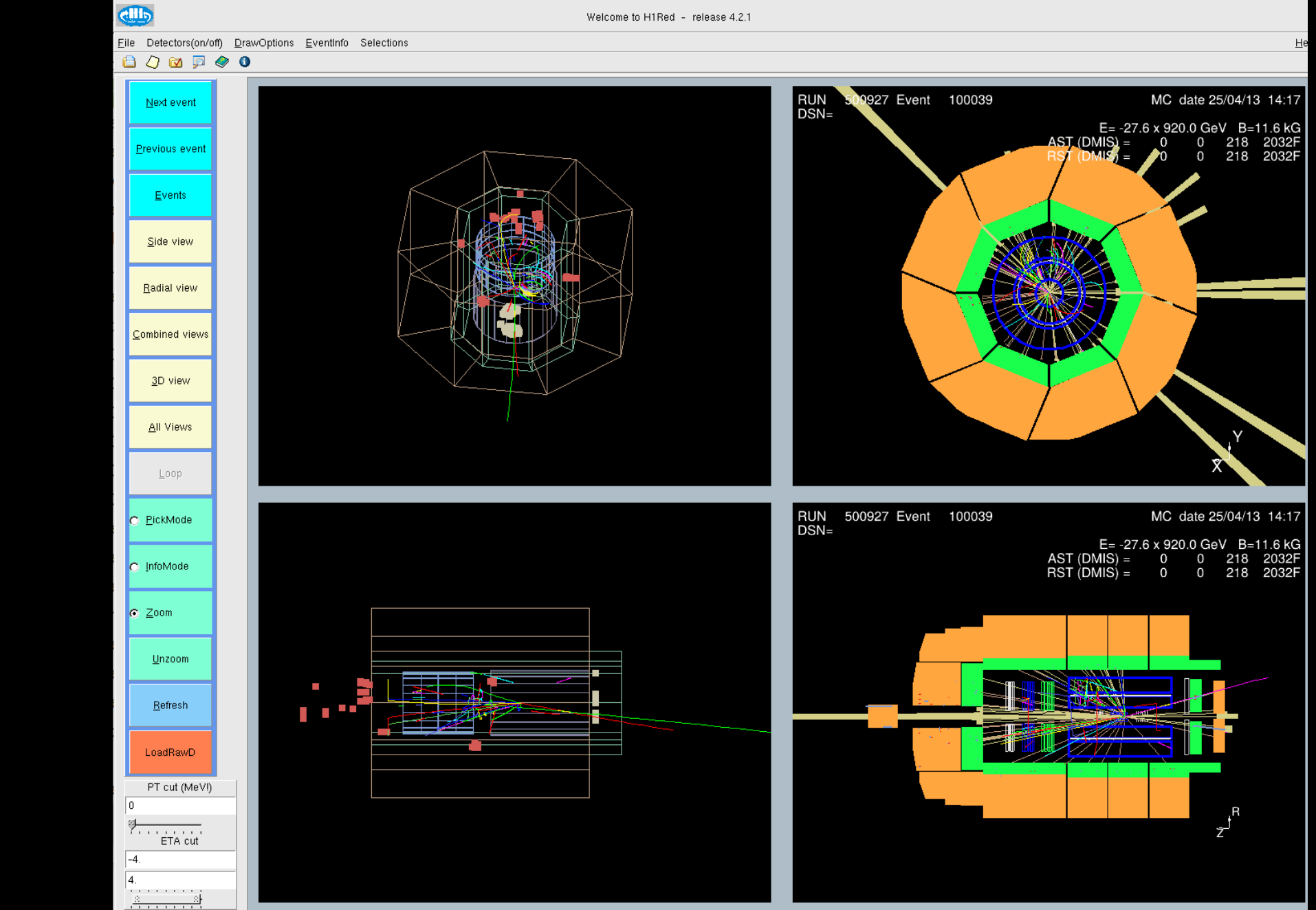
Trigger information (DMIS)
AST bits = 1 17 36 41
RST bits = 1 17 36 41
L2 bits = 2



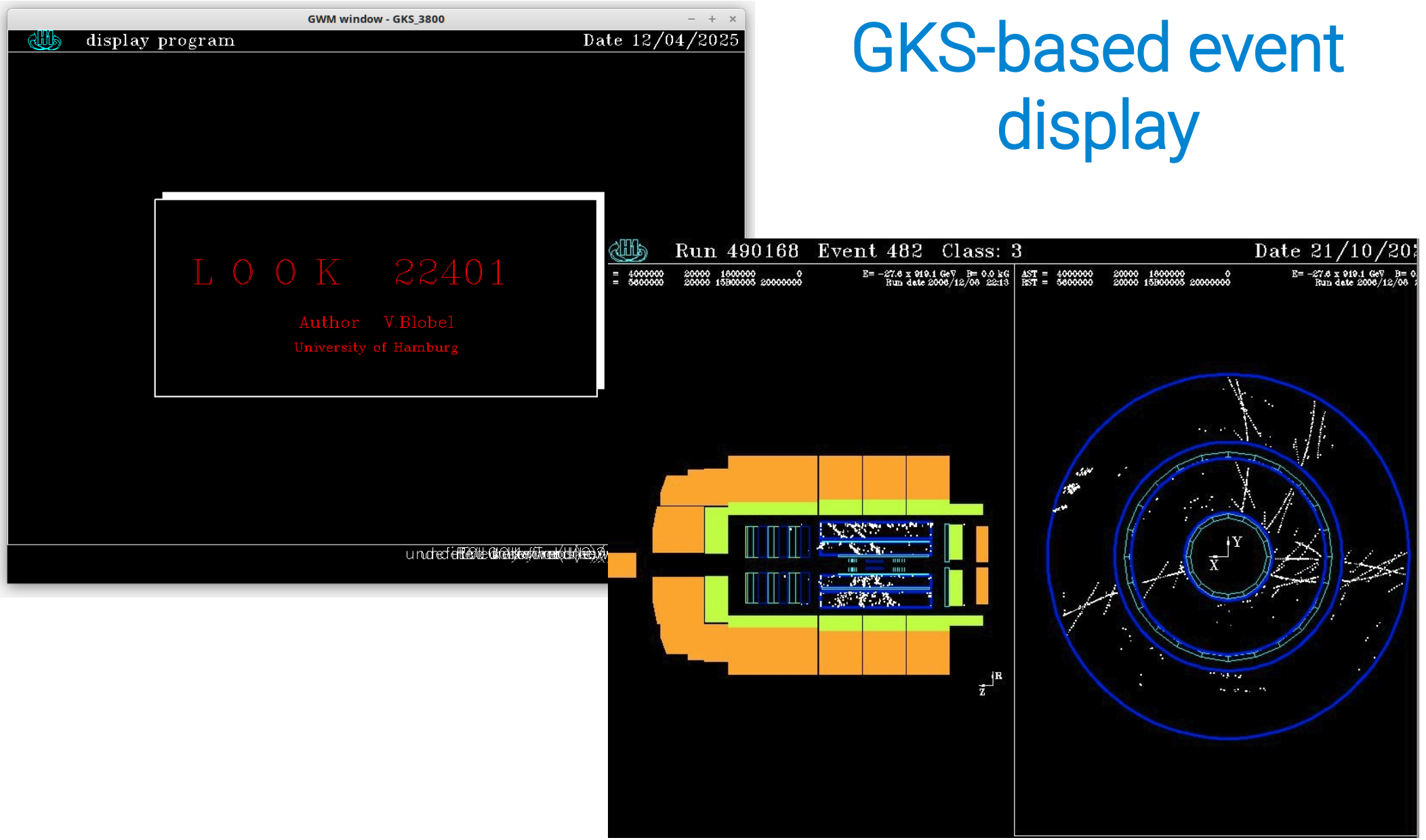







Welcome to H1Red - release 4.2.1
File
Detectors(on/off)
DrawOptions
EventInfo
Selections
Next event
Previous event
Events
Side view
Radial view
Combined views
3D view
All Views
Loop
PickMode
InfoMode
Zoom
Unzoom
Refresh
LoadRawD
PT cut (MeV!)
0
ETA cut
-4.
4.
RUN 500927 Event 100039
MC date 25/04/13 14:17
DSN=
E= -27.6 x 920.0 GeV B=11.6 kG
AST (DMIS) = 0 0 218 2032F
RST (DMIS) = 0 0 218 2032F
Y
X
RUN 500927 Event 100039
MC date 25/04/13 14:17
DSN=
E= -27.6 x 920.0 GeV B=11.6 kG
AST (DMIS) = 0 0 218 2032F
RST (DMIS) = 0 0 218 2032F
R
Z

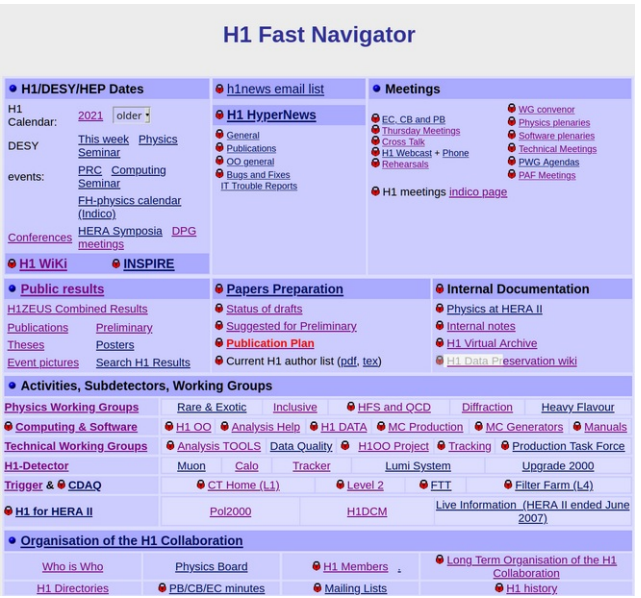
GKS-based event display



Documention beyond software

H1 maintains a webserver as the central resource for documentation

- working group webpages
- technical documentation (s/w, hardware)
- Analysis notes, internal notes (~100)
- All meetings and wiki pages
  → these were converted to a static format
- MC database access
- Theses, slides, meeting minutes, etc...
- Scanned documents
- content from linked private homepages were copied
- mails from mailing lists preserved
- etc...



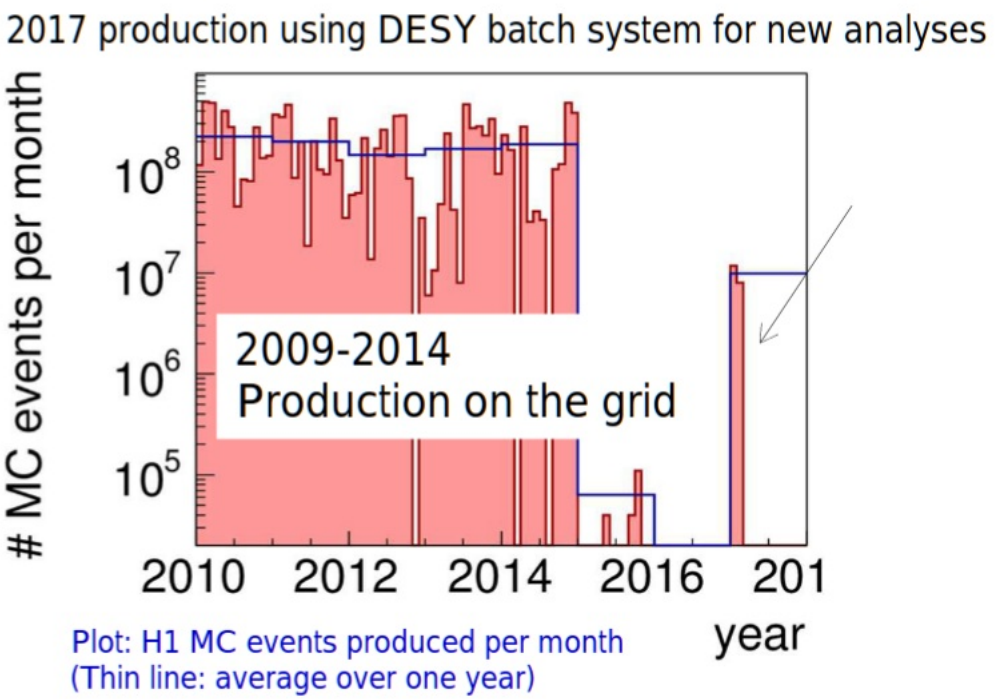
Shared working directories (/afs, /cvmfs, /nfs) for H1 members

- ongoing analyses, webpages, software releases, etc...

H1 archive with paper documents [level-1 documentation] → Scanned: **H1 virtual archive**

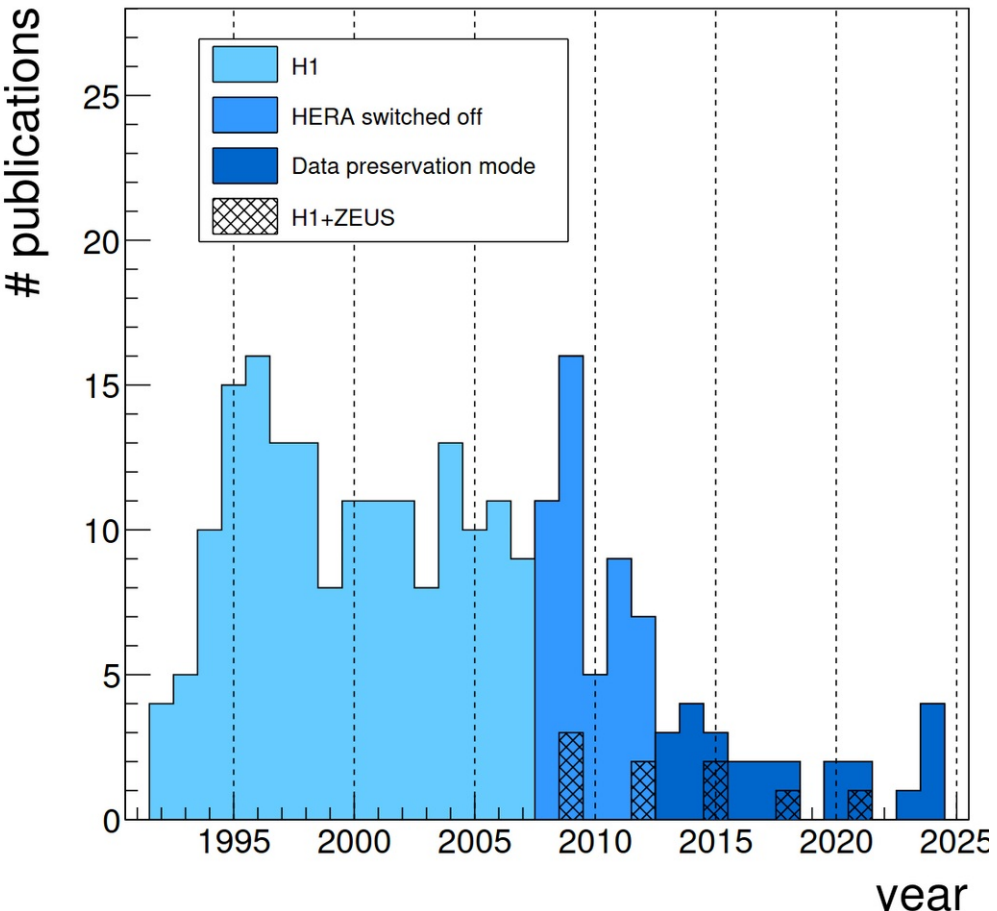
Physics results

Monte Carlo event production



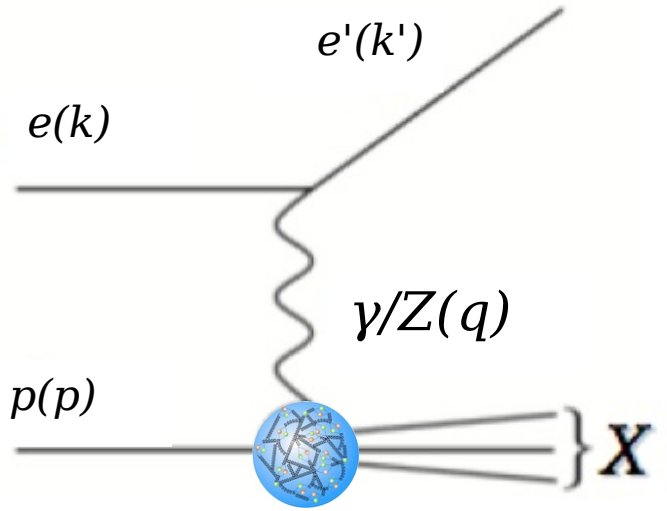
Plot: H1 MC events produced per month
(Thin line: average over one year)

H1 Publications

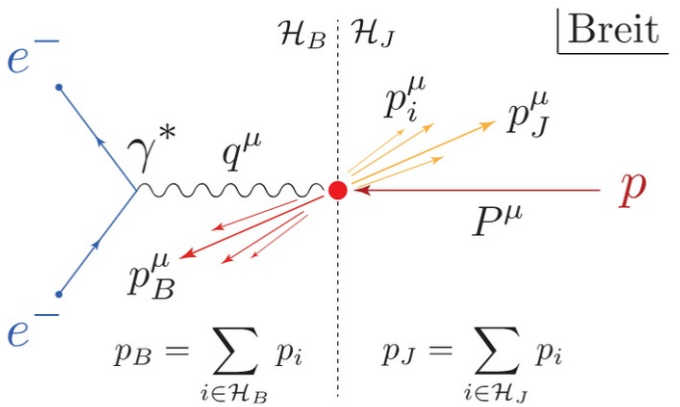


Physics motivation for H1 data preservation and analysis

Deep-inelastic scattering (DIS)



DIS in the Breit frame

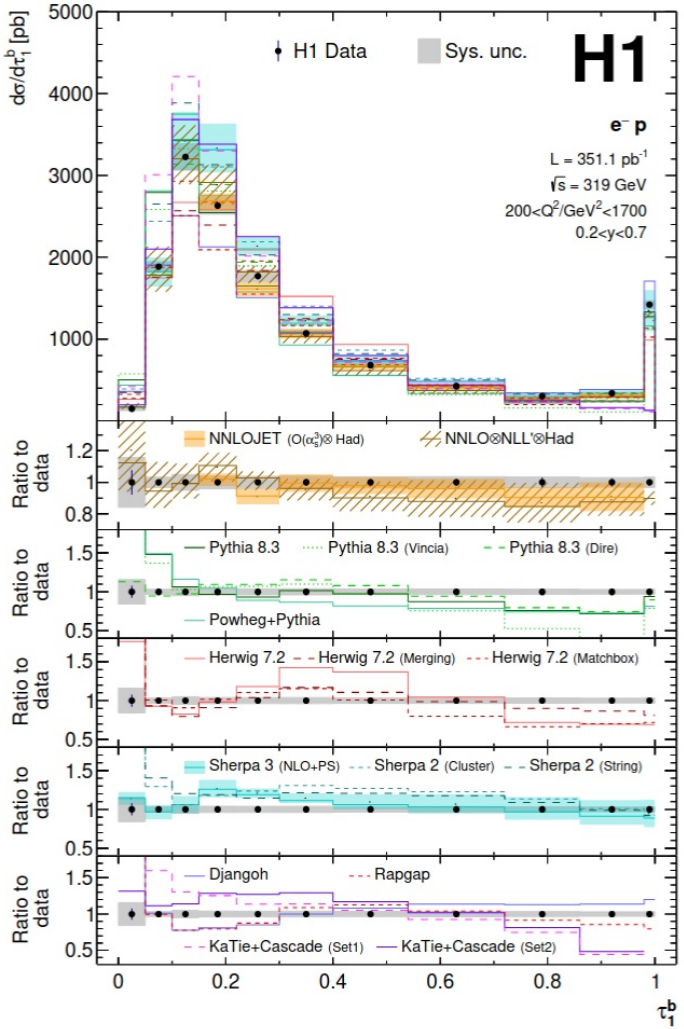


Broad physics programme

- Proton structure, QCD, heavy flavors, electroweak physics, exclusive processes, diffraction, BSM, etc...

- In the Breit reference frame, a virtual photon collides head on with the proton
- In the photoproduction limit, the photon becomes quasi-real with an internal (hadronic) structure

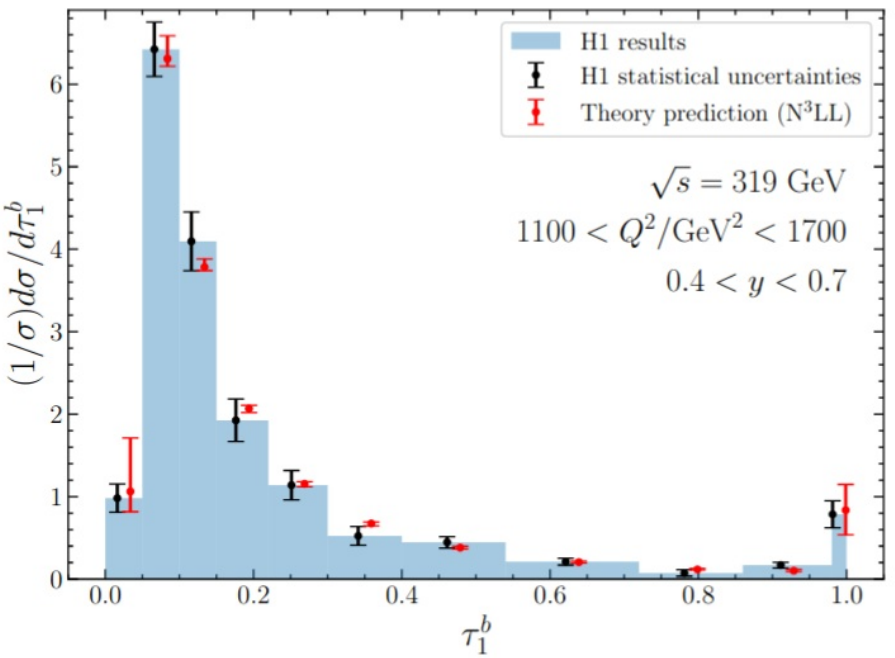
Recent results



- Single-, double- and triple-differential distributions were measured (Q^2, y, τ_1^b)

Measurement of the 1-jettiness event shape observable

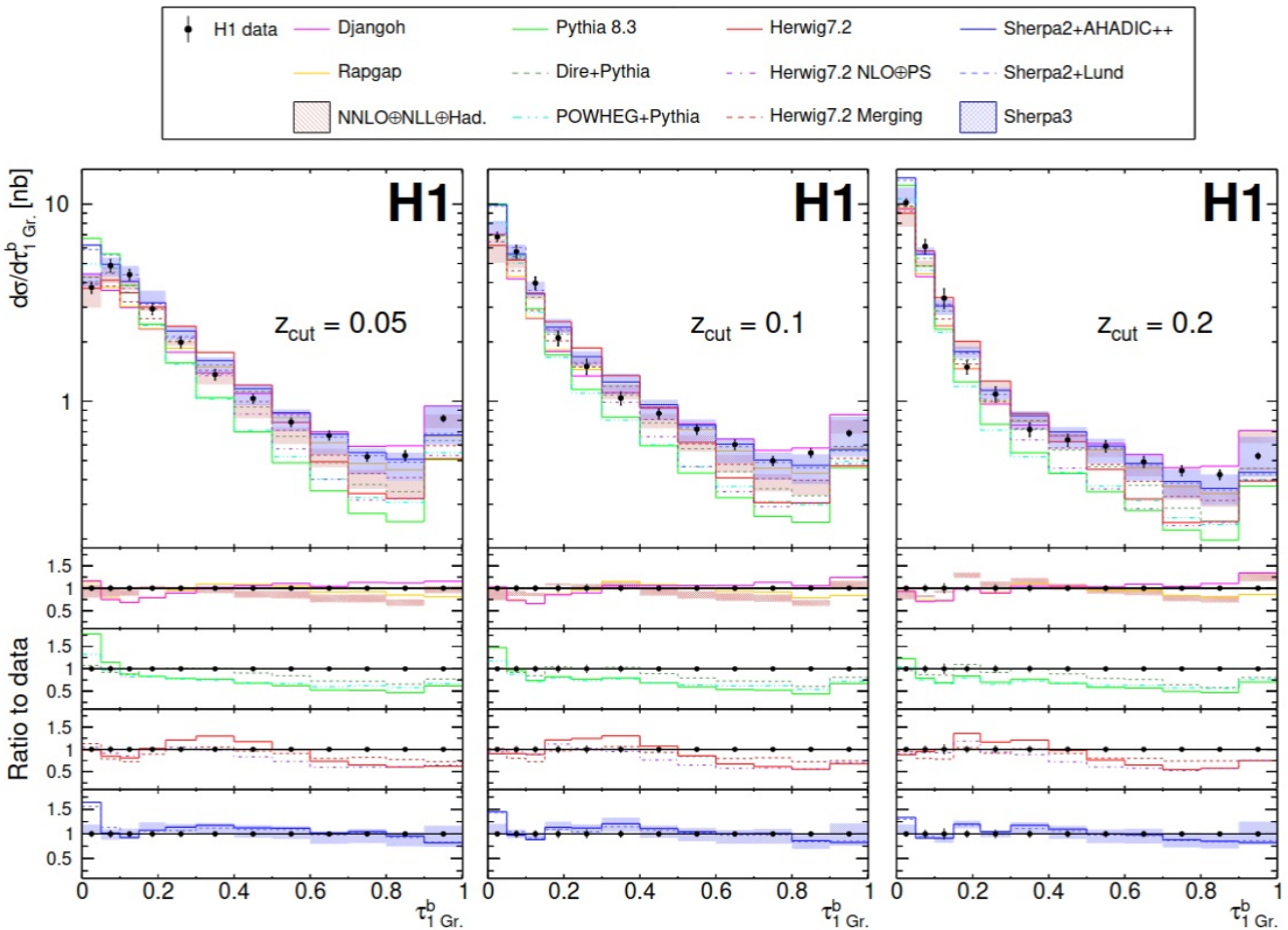
Eur.Phys.J.C84 (2024), 785 [arxiv:2403.10109]



Ee ,Kang, Lee, Stewart, NNLO+N3LL predictions,
JHEP 07 (2025) 240

- New data triggered new theoretical developments

Recent results

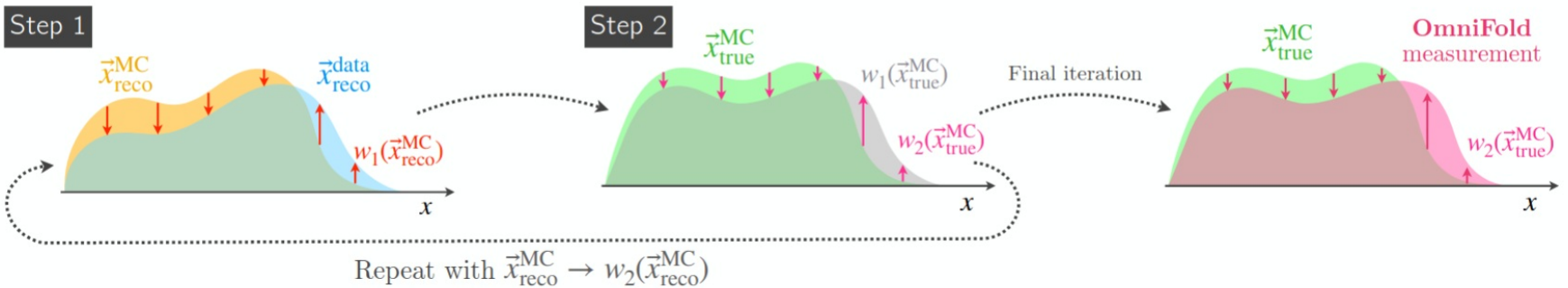


Groomed event shape observables in DIS

Eur.Phys.J.C84 (2024), 718 [arxiv:2403.10134]

- Grooming:
  Iteratively de-cluster until grooming condition is passed
  → Procedure removes 'soft' particles and effectively excludes regions that are not well measured
- Most MC models do well
- Sherpa 3 agrees well with data

New results with ML-assisted unbinned unfolding



Experiment	Paper Link	Dimensions	Final State	Momentum Selection
ATLAS [23]	2405.20041	24	Z+jets	$p_{\mathrm{T}}^{\ell\ell} > 200$ GeV
ATLAS [24]	2502.02062	6	Dijets	$p_{\mathrm{T}}^{j1} > 240$ GeV & $p_{\mathrm{T}}^{j1} < 1.5\ p_{\mathrm{T}}^{j2}$
CMS [25]	2505.17850	8	Minimum bias	> 2 charged particles with $p_{\mathrm{T}} > 0.5$ GeV
H1 [26]	2108.12376	8*	High Q^2 DIS	$Q^2 > 150$ GeV^2
H1 [27]	2303.13620	10	High Q^2 DIS	$Q^2 > 150$ GeV^2
H1 [28]	2412.14092	8*	High Q^2 DIS	$Q^2 > 150$ GeV^2
H1 [29]	H1prelim-25-031	Variable	High Q^2 DIS	$Q^2 > 150$ GeV^2
LHCb [30]	2208.11691	4	Z+hadrons in jets	$20 < p_{\mathrm{T}}^{j} < 100$ GeV and $p_{\mathrm{T}}^{h} > 0.25$ GeV
STAR [31]	2307.07718	6	Jets	$20 < p_{\mathrm{T}}^{j} < 50$ GeV
STAR [32]	2403.13921	7	Jets in heavy ions	$20 < p_{\mathrm{T}}^{j} < 45$ GeV
T2K [33]	2504.06857	6	Muon + Proton	$p^{p} > 450$ MeV for single transverse variables

TABLE I. An overview of the recently-published experimental results that use unbinned unfolding methods. For most results, the unfolded dimensionality is the same at reconstruction-level and at truth-level, but * indicates the analyses for which the unfolded dimensionality at reconstruction-level was the full phase space, but the truth-level unfolded result is 8-dimensional. Full details of the phase space for each measurement, including η selections, are listed in the individual papers.

A practical guide to unbinned unfolding [arXiv:2507.09582]

New measurements based on unbinned unfolding approach

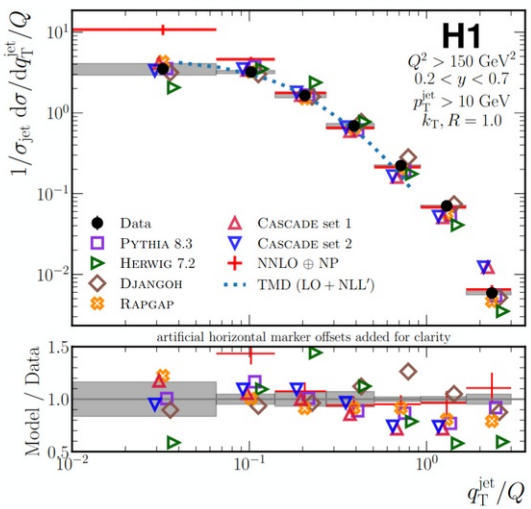
lepton-jet correlation in DIS,

PRL 128 (2022), 132002
[arxiv:2108.12376]

same unfolded data

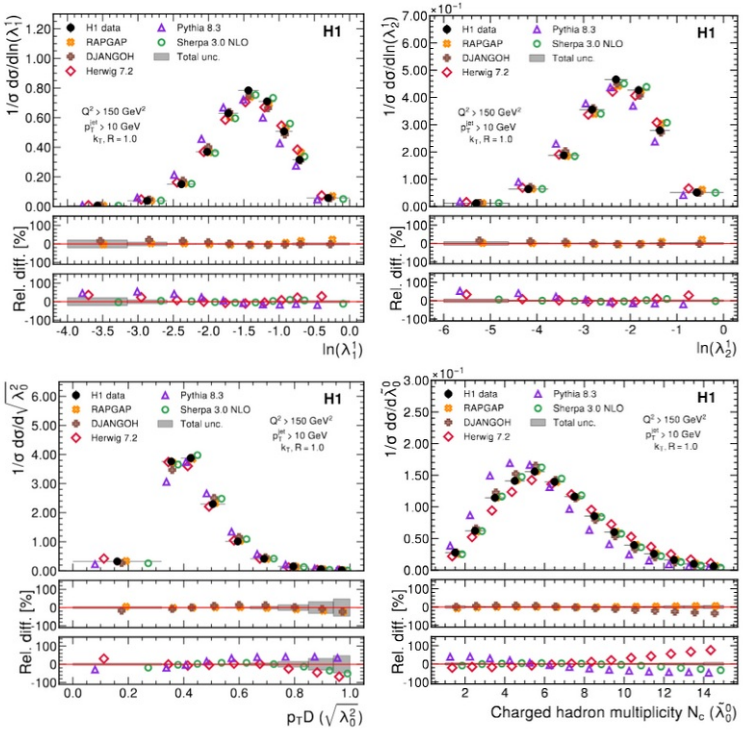
Lepton-Jet Azimuthal Angular Asymmetries in DIS,

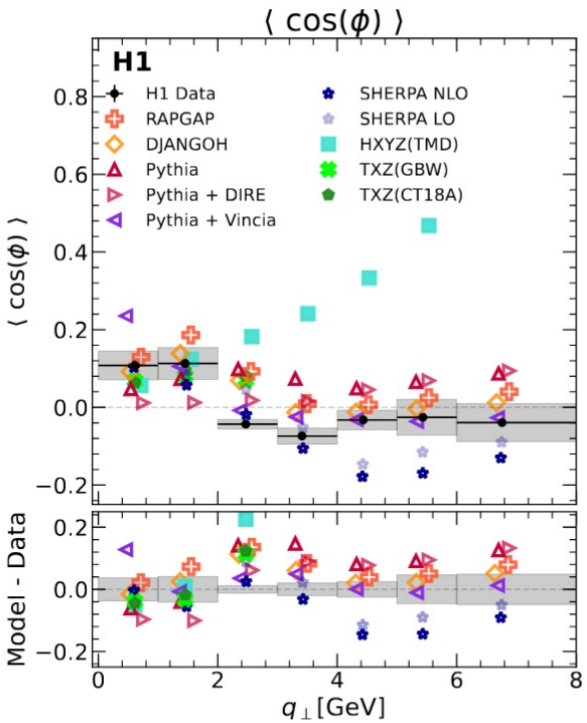
Subm. to PLB [arxiv:2412.14092]



Jet Substructure Measurement in High Q²

PLB 844 (2023) 138101 [arxiv:2303.13620]

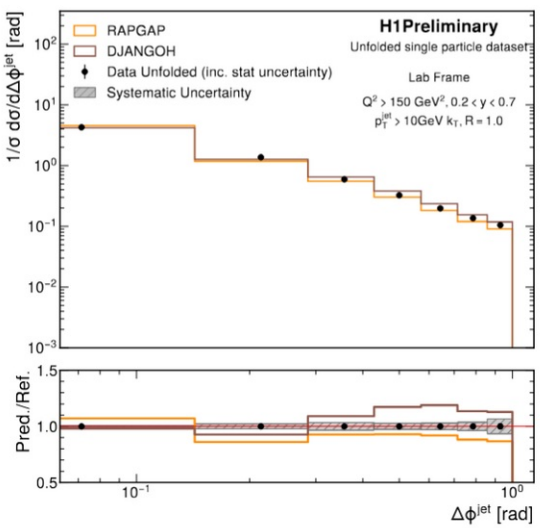


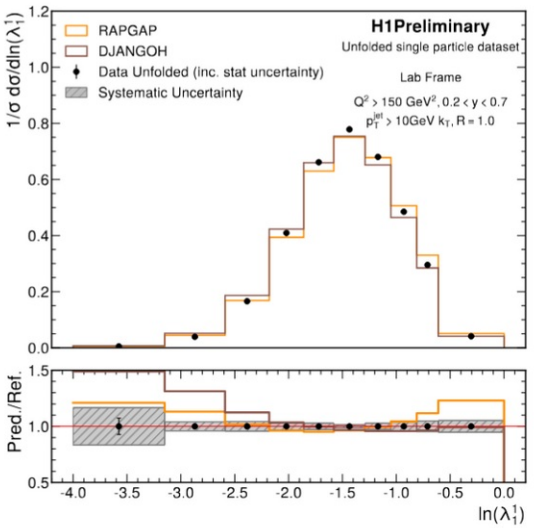


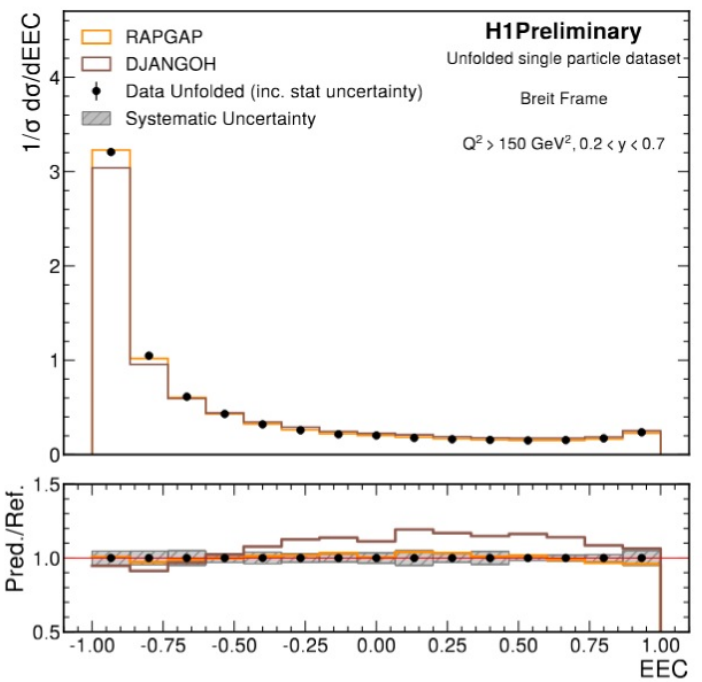
H1prelim-25-031 - Towards unfolding all particles

Unbinned unfolding of all particles

- Full granularity of the data is unfolded and available at the (unfolded) particle level
- Allows for a easy and quick measurmenet of new observables







k_T jets and their properties: constructed from unbinned unfolded single particles

New observables, like energy-energy-correlators can be obtained without extensive analysis efforts !

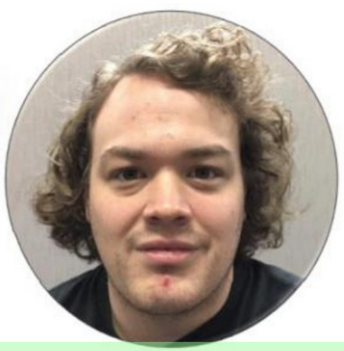
(Young) scientists' perspective

H1 was more or less perfect since it allowed me to get hands-on experience with all aspects of a cross-section analysis but in a relatively short time. I was working almost entirely on hardware throughout my Ph.D, so I wouldn't have been in the best position career-wise without it.

H1 also provided me super valuable experience toward the EIC. Relatively few of the people working on EIC have the e+p collider experience, so it's really useful. I think it was largely for this reason that I was hired for a permanent position after only a year and a half of postdoc.

I would say the analysis approach is as up-to-date as it can be and the analysis feels modern.

Henry Klest,
PhD in 2023, now
Assistant Physicist, ANL



Gage Tustin,
Master in 2025, now
Physics Associate, BNL

Working with H1 is the first experience I have had performing a physics analysis and I am very grateful for the opportunity and have learned so much! I do enjoy the physics and the H1 collaboration has been extremely helpful sharing their wealth of knowledge and are always very helpful in guiding me through physics/software issues that I have encountered.

H1 is a treasure trove of data and retaining its full potential for future physics analyses in the context of advancing theoretical understanding is important, especially with EIC on the horizon. Our analysis was in part motivated by potential physics of interest to study at EIC and H1 is a place to get an initial look into these ideas for DIS and in the long term would be nice to have for comparing/replicating future EIC DIS results.

Summary and conclusion

The H1 experiment at HERA took a unique set electron-proton collision data

- All data preserved and software stack is continuously evolving

H1 data and software are kept in DPHEP mode 'level 4'

- Full offline and online documentation
- Full analysis capability: recompilation of software and continuous migrations to newer OS
- Since 2012: migrations from SLD5-32bit to SLD5-64bit, to SLD6, to CentOS7 and to Alma9
- Bonus: all previous releases can be executed within default Singularity images

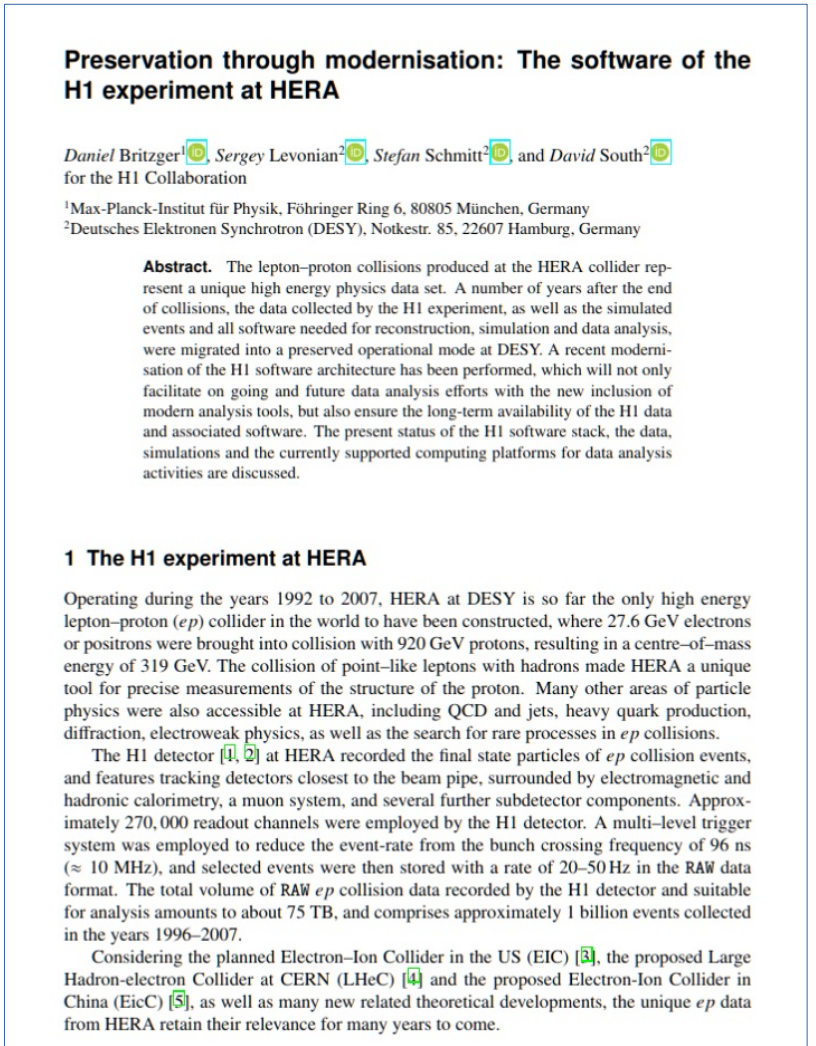
Modernisation of the H1 software architecture in 2020 & 2024

- Introduction of LCG dependencies, and DESY-IT standards
  → reduce maintenance efforts
- Latest dependencies (gcc14, ROOT6, C++22, git, …)
  → Modern analysis and computing environment → attractive for young physicists
- Data are actively analysed and new collaborators are welcomed and are joining

vCHEP2021 contribution

After about 10 years of data preservation…

- Summary of the DP status in H1
- Focus on 'software' preservation
  - H1oo C++ analysis framework
  - Fortran core software
- Report experiment specific experience on DPHEP level-4
- Discussion about H1's computing model in DPHEP mode



Preservation through modernisation: The software of the H1 experiment at HERA

Daniel Britzger[1], *Sergey* Levonian[2], *Stefan* Schmitt[2], and *David* South[2] for the H1 Collaboration

[1]Max-Planck-Institut für Physik, Föhringer Ring 6, 80805 München, Germany
[2]Deutsches Elektronen Synchrotron (DESY), Notkestr. 85, 22607 Hamburg, Germany

Abstract. The lepton–proton collisions produced at the HERA collider represent a unique high energy physics data set. A number of years after the end of collisions, the data collected by the H1 experiment, as well as the simulated events and all software needed for reconstruction, simulation and data analysis, were migrated into a preserved operational mode at DESY. A recent modernisation of the H1 software architecture has been performed, which will not only facilitate on going and future data analysis efforts with the new inclusion of modern analysis tools, but also ensure the long-term availability of the H1 data and associated software. The present status of the H1 software stack, the data, simulations and the currently supported computing platforms for data analysis activities are discussed.

1 The H1 experiment at HERA

Operating during the years 1992 to 2007, HERA at DESY is so far the only high energy lepton–proton (ep) collider in the world to have been constructed, where 27.6 GeV electrons or positrons were brought into collision with 920 GeV protons, resulting in a centre–of–mass energy of 319 GeV. The collision of point–like leptons with hadrons made HERA a unique tool for precise measurements of the structure of the proton. Many other areas of particle physics were also accessible at HERA, including QCD and jets, heavy quark production, diffraction, electroweak physics, as well as the search for rare processes in ep collisions.

The H1 detector [1, 2] at HERA recorded the final state particles of ep collision events, and features tracking detectors closest to the beam pipe, surrounded by electromagnetic and hadronic calorimetry, a muon system, and several further subdetector components. Approximately 270,000 readout channels were employed by the H1 detector. A multi–level trigger system was employed to reduce the event-rate from the bunch crossing frequency of 96 ns ($\approx$ 10 MHz), and selected events were then stored with a rate of 20–50 Hz in the `RAW` data format. The total volume of `RAW` ep collision data recorded by the H1 detector and suitable for analysis amounts to about 75 TB, and comprises approximately 1 billion events collected in the years 1996–2007.

Considering the planned Electron–Ion Collider in the US (EIC) [3], the proposed Large Hadron-electron Collider at CERN (LHeC) [4] and the proposed Electron-Ion Collider in China (EicC) [5], as well as many new related theoretical developments, the unique ep data from HERA retain their relevance for many years to come.

https://indico.cern.ch/event/948465/contributions/4324171/

Modernisation of the H1 software

2020: Successful migration to CentOS7, but a few shortcomings now evident in the H1 software

- The programming languages (C++98) and standards are unattractive for new (young) people to learn
- Outdated dependencies, such as ROOT 5, complicate the usage of modern data analysis techniques
- New dependencies may be incompatible (different compilers standards or MC-generator formats)

→ Modern tools cannot or have not in general been introduced

- Relevant maintenance effort for external (outdated) dependencies

Component	Responsible	Maintained packages	Discontinued packages
H1 software	H1	H1 core software, H1OO	–
OS dependencies (continuous updates)	DESY–IT	Oracle, dCache, web–services, compilers, GNU utilities, gmake, system libraries	CVS
External dependencies (selected fixed releases)	H1	fastjet, neurobayes–expert, MC generators	CERNLIB, GKS, GEANT3, ROOT5, LHAPDF5, MC generators

2020: Restructuring the software

- Make use of 'modern' tools and dependencies, and recent releases of external packages

➞ Introduction of dependence on the LCG package repository

- Previously: no externally maintained package repository: packages provided manually
- Two effects: reduction of H1 maintenance and bring in newer versions of existing software dependencies and compilers

Modernisation of the H1 software (cont'd)

All code repositories migrated to git (DESY-IT service)

- H1 used CMZ and CVS (H1 did not get to SVN)
- New build instructions for entire H1 s/w stack
  → Less reliance on historic development

Using recent dependencies from LCG release (97a)

- Entire FORTRAN software stack was migrated
  (huge jump in GNU compiler collection 4.8 to 9.2)

Component	Responsible	Maintained packages	Discontinued packages
H1 software	H1	H1 core software, H1OO	–
OS dependencies (continuous updates)	DESY–IT	Oracle, dCache, web–services, GNU utilities, git, gmake, system libraries	–
External dependencies (selected fixed releases)	H1	–	CERNLIB, GKS, GEANT3 (selected) MC generators
External dependencies (selected regular updates)	LCG	LHADPF6, ROOT6, compilers, fastjet, neurobayes–expert, MC generators, (and as back up option: Oracle, dCache, git)	–

H1oo analysis framework updated to ROOT 6 and C++17; CLING replaces CINT

- Original production of data and MC files remain compatible
- New C++ standard allowed s/w improvements, for example range-based for loops in H1Arrays
- Another benefit of ROOT 6 is PyROOT: Fully pythonic analysis of H1 data now possible, incl. interactive

Complete release of all H1 software now on /afs and /nfs at DESY (to be distributed on /cvmfs)

- H1 core packages were previously bound to the DESY–IT infrastructure; now can be relocated
- H1 s/w now runs (in principle) without problems e.g on CentOS7 lxplus at CERN

Bonus: SLD5, SLD6 container builds using Singularity as retrospective “DPHEP level 3” preservation