WORKSHOP ON
SOFTWARE INFRASTRUCTURE
FOR ADVANCED
NUCLEAR PHYSICS
COMPUTING JUNE 20-22, 2024

BERKELEY LAB

Archived data reanalysis: lessons from a user

Yi Chen (Vanderbilt), Peter Jacobs (LBNL)

2024 Jun 21, SANPC Workshop

Vanderbilt HENP's group is supported by US DOE-NP

Main takeaways

In case we don’t have time in the parallel session

Data reanalysis: hindsight

- Foresight from the **collaborations** for the data preservation
- Incredible support from members in the collaboration on the **technical aspects and knowledge**
- Many **bright young students** who dug into the data collected before they were born
- **Reproduction of published physics results** using identical event selections
- Development of **data-driven checks** to understand the data
- Ability to **rerun key software** is crucial

Lessons for future

- Mileage vary ***a lot*** depending on experiments
  - Make sense of the format: **knowledge** needed from members
  - Not easy to gain **control of stored information** — more lower-level information will be useful
  - Good to have more **sets of fully simulated MCs** available
  - Ability to **rerun key software** is crucial (as we see in H1)
- Many lessons for current & future experiments
  - Enough information for end-to-end measurements?
  - Best to do some "**user tests**" for open data as we go

Full set of slides

Why archived data

- Reanalysis with old data. e.g. ALEPH (e^+e^-) and H1 (ep)
- **Huge amount of exciting things to explore!**
    - **Modern algorithms** show up long after LEP/HERA (e.g. anti-kT jet 2008, Centauro 2021)
        - e^+e^- and ep much cleaner than others → fundamental QCD studies, complementary to hadron colliders (LHC/RHIC)
    - **New ideas** (e.g. ridge in 2-particle correlation in e^+e^-?)
    - **Testing ground** for new algorithm developments (e.g. EIC)
- Capitalize on what we have accumulated already and prepare for new endeavors

JHEP 0804 (2008) 063

How ALEPH reanalysis started

2017

February: **Yen-Jie Lee** connected to **Gigi Rolandi** and later to spokesperson **Roberto Tenchini** about the use of archived data

Marcello Maggi help extract the energy flow information and archived simulation/data

Mid-2017: all samples converted to the MIT open-data format

Bibek Pandit & **Anthony Badea** (Yen-Jie's undergraduate student) started working on event selection validation

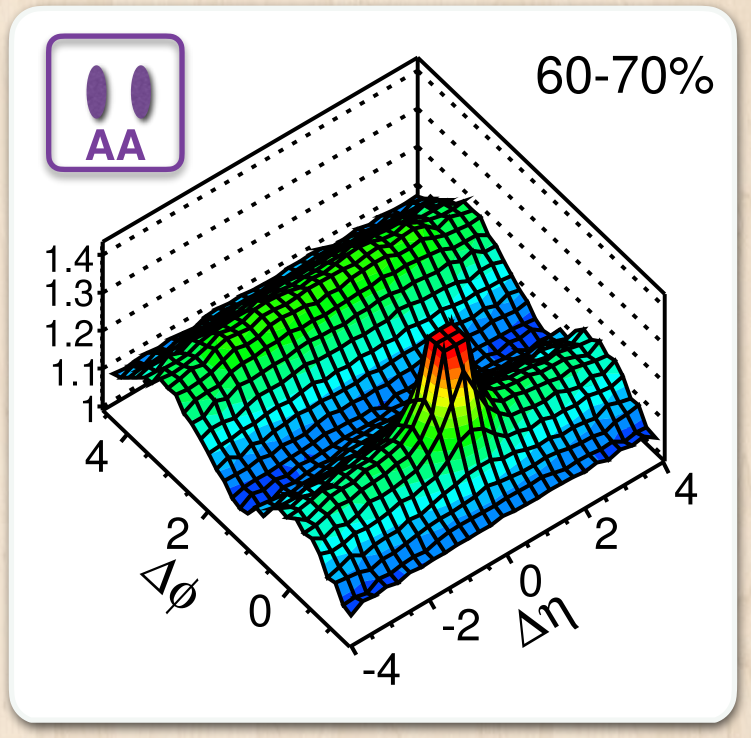
Guenther Dissertori provided analysis code from the QCD paper

2018

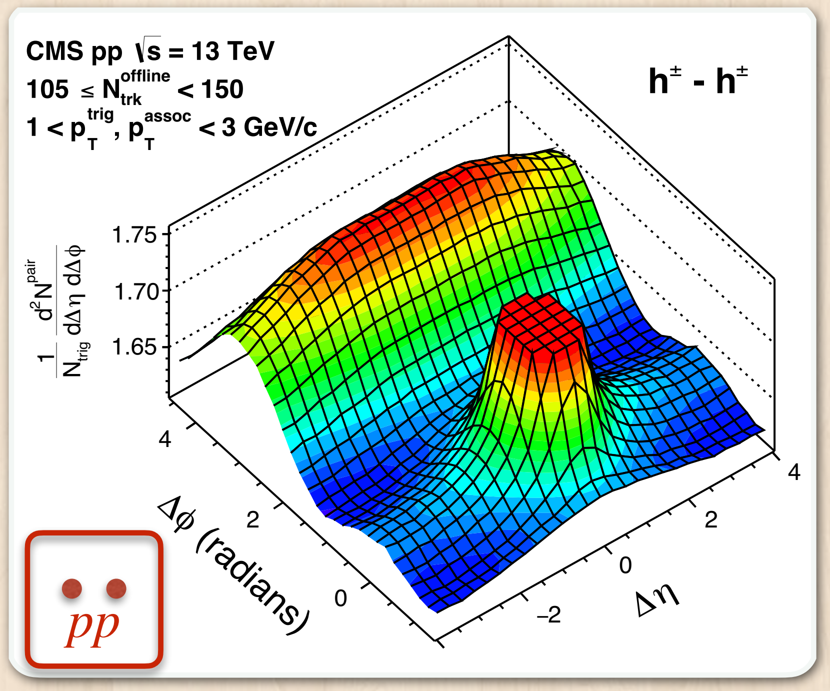
March: Successfully reproduced unfolded thrust distribution

On H1 side, all software (e.g. MC, GEANT) has been kept current, can run recent MC tunes through GEANT to do LHC-quality analyses including sys uncertainty studies

Example: e^+e^- 2-particle correlation



In PbPb collectivity is seen as a potential sign of QGP

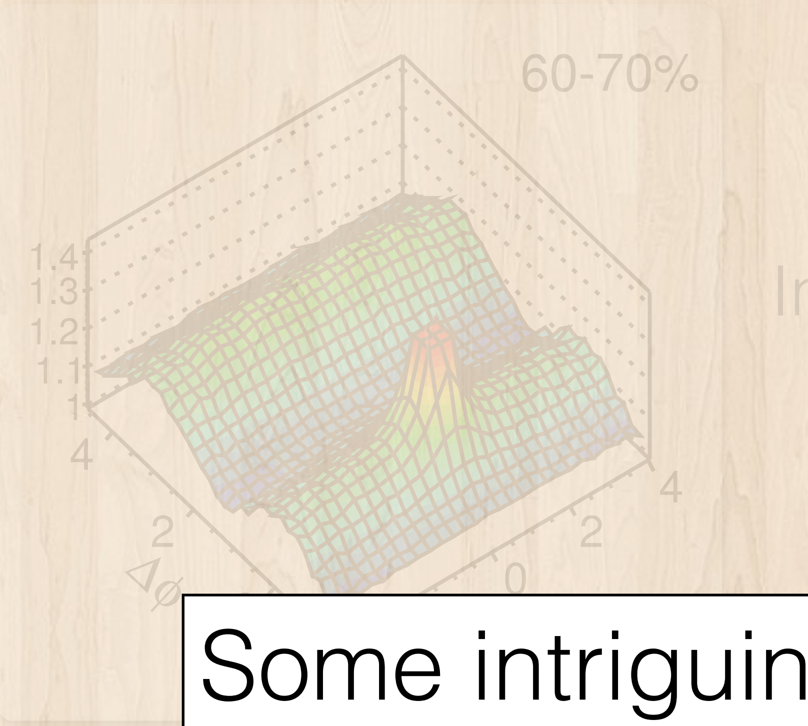


But we see it somehow in high multiplicity pp!?

What about e^+e^-?

PLB 724 (2013) 213, PLB 765 (2017) 193

Example: e^+e^- 2-particle correlation

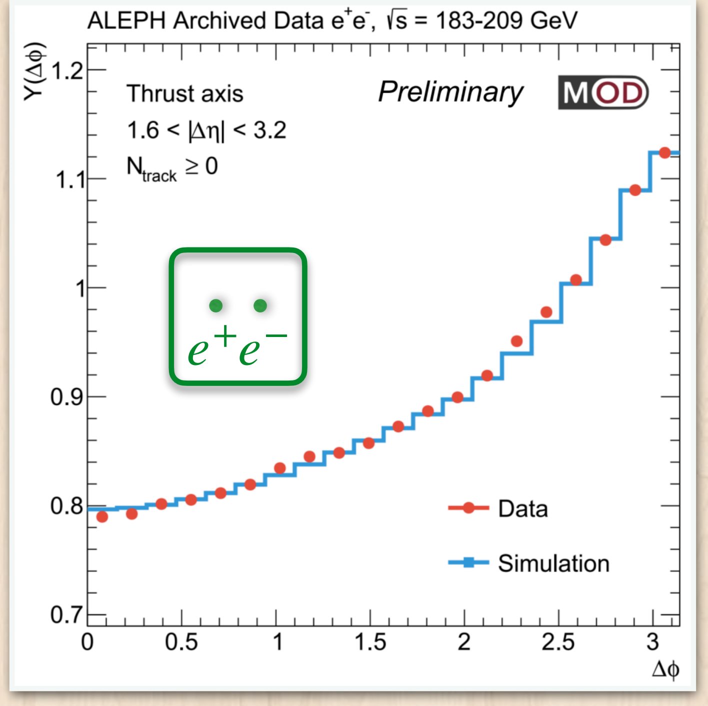


In PbPb collectivity is seen as a potential sign of QGP

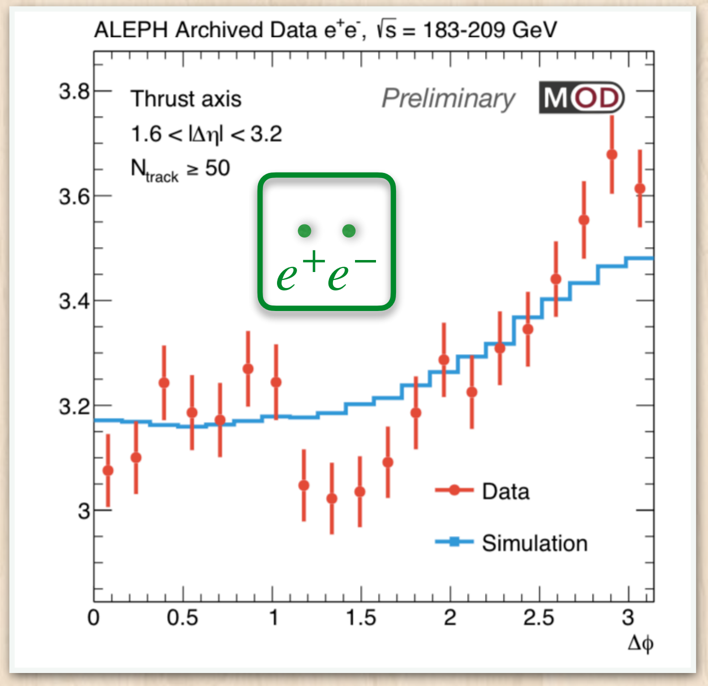
Some intriguing structure in high multiplicity e^+e^-

But we see it somehow in high multiplicity pp!?

Low multiplicity



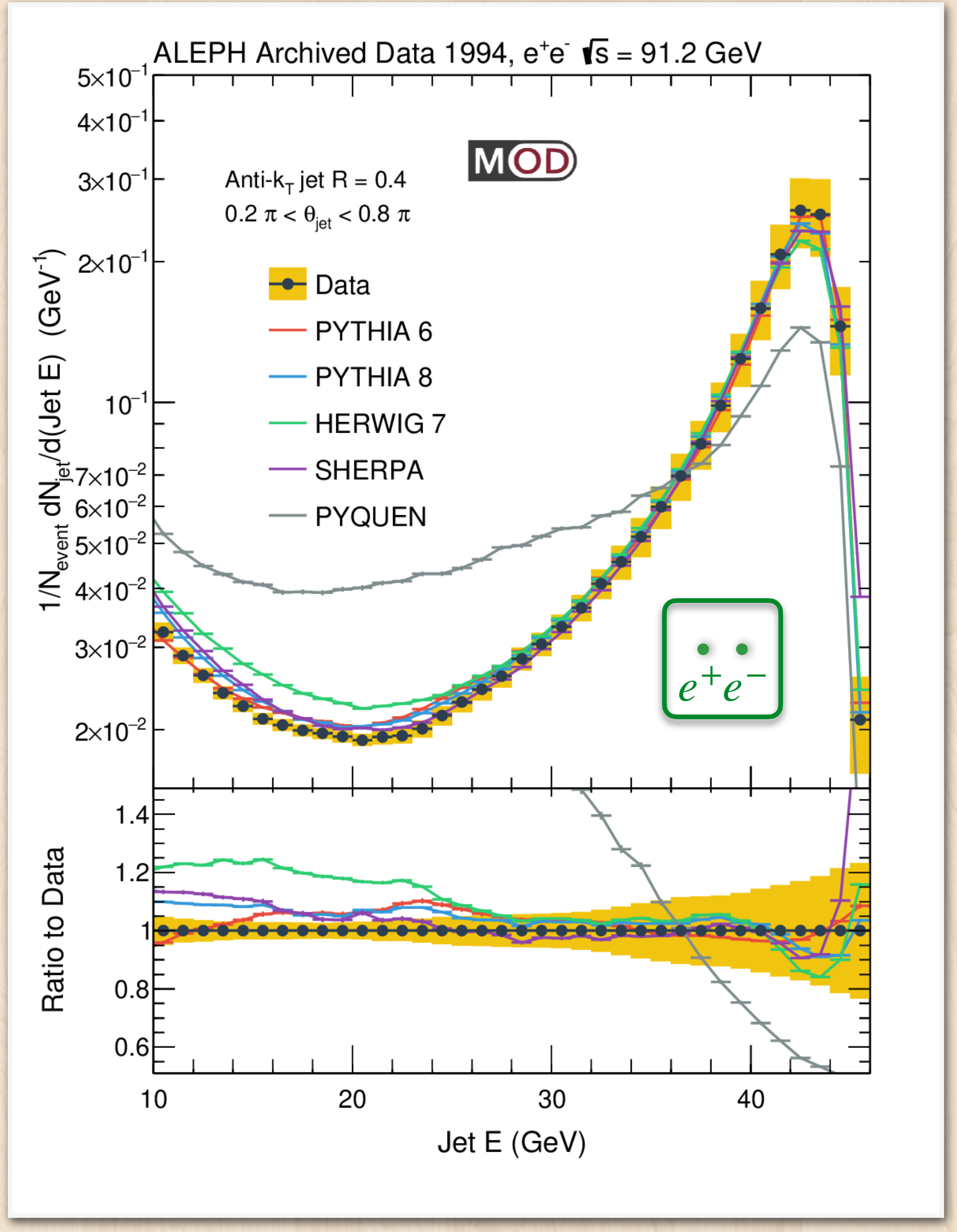
High multiplicity



arXiv 2312.05084, 2306.04808

Example: e^+e^- jet measurement

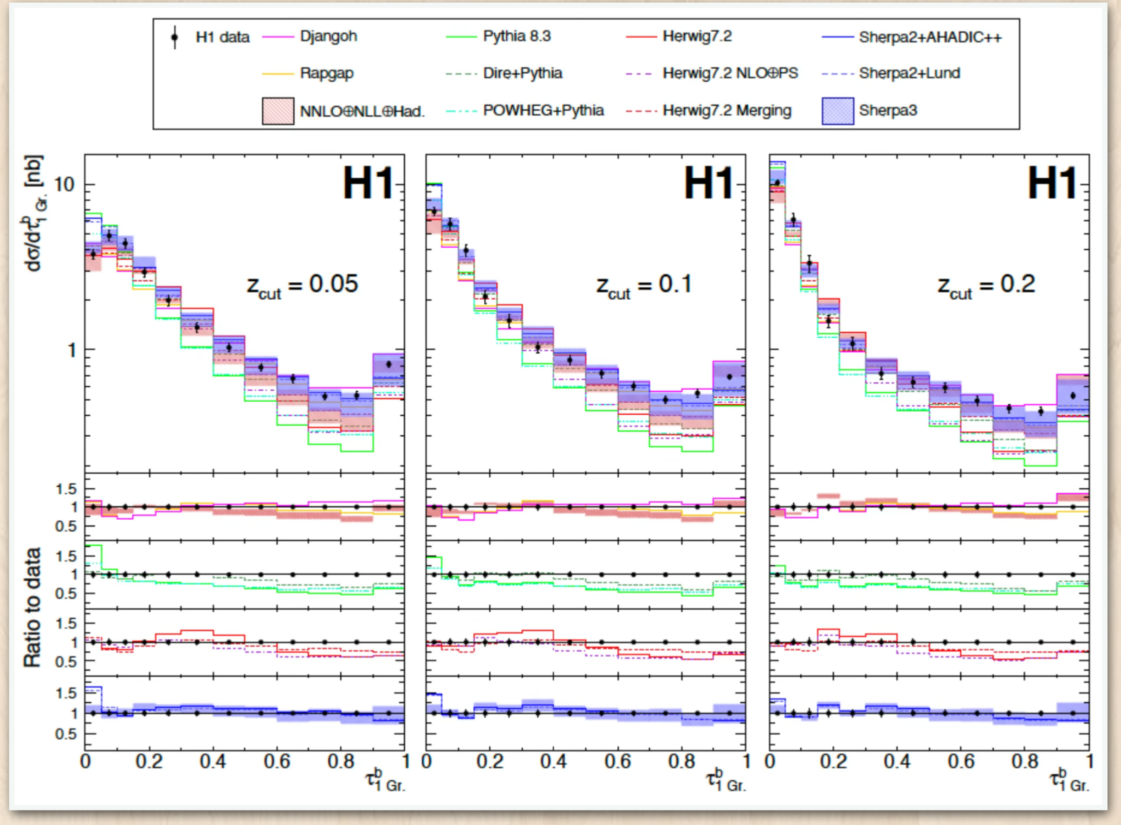
- Measured jet spectra and jet substructure (not shown)
- This result: 91 GeV data
- Unique peaking structure access to the rising edge



JHEP 06 (2022) 008

Example: *ep* groomed event shape (H1)

- Clean up event using event-wide grooming algorithms (Centauro clustering + soft drop idea)
    - Experimental handle to control amount of non-perturbative effect
- Measure invariant mass (not shown) and 1-jettiness
- Rich dataset for precision MC tuning



H1 Collaboration, arXiv 2403.10134

Data reanalysis: hindsight

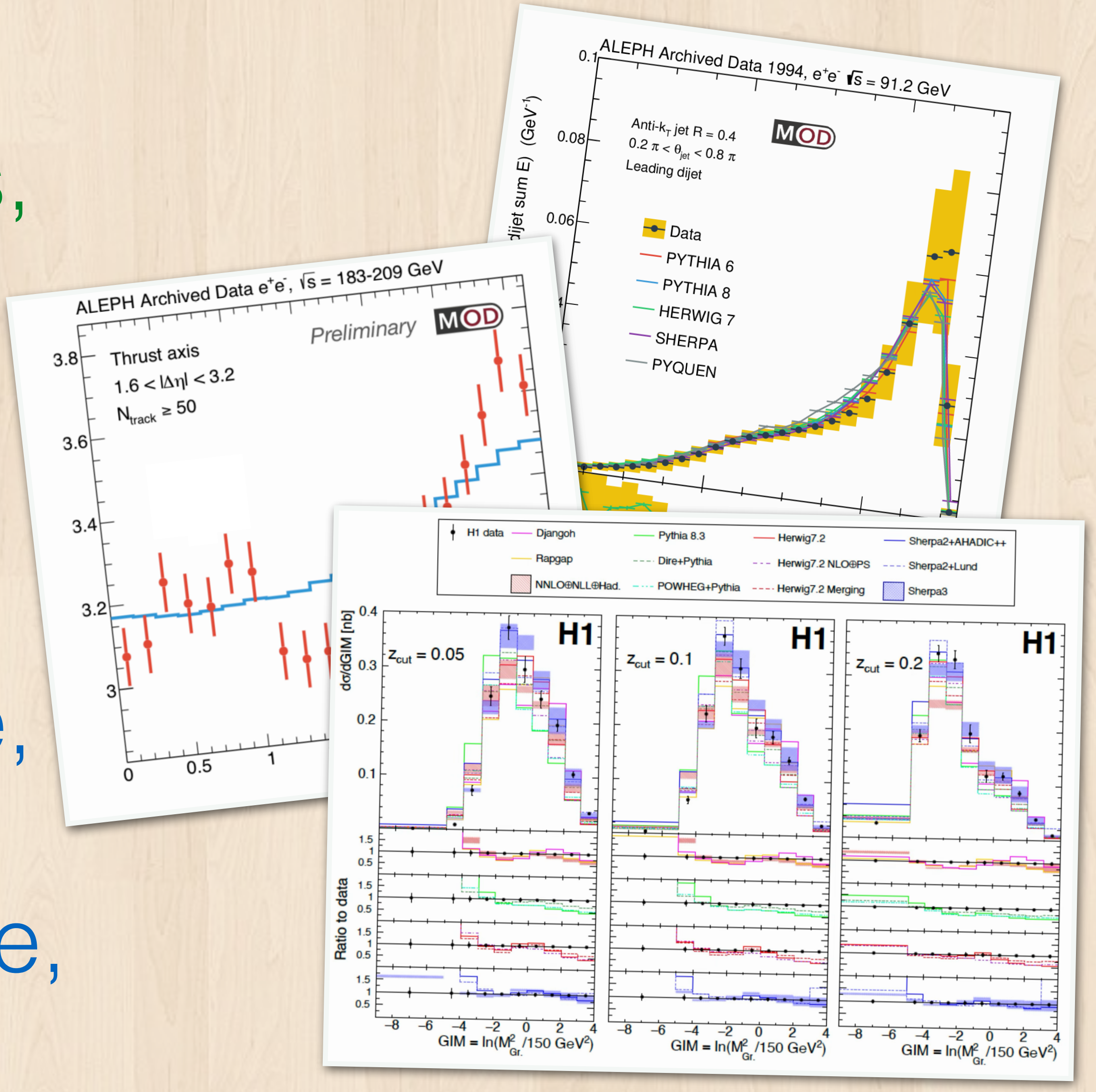
- Foresight from the **collaborations** for the data preservation
- Incredible support from members in the collaboration on the **technical aspects and knowledge**
- Many **bright young students** who dug into the data collected before they were born
- **Reproduction of published physics results** using identical event selections
- Development of **data-driven checks** to understand the data
- Ability to **rerun key software** is crucial

Lessons for future

- Mileage vary ***a lot*** depending on experiments
  - Make sense of the format: **knowledge** needed from members
  - Not easy to gain **control of stored information** — more lower-level information will be useful
  - Good to have more **sets of fully simulated MCs** available
  - Ability to **rerun key software** is crucial (as we see in H1)
- Many lessons for current & future experiments
  - Enough information for end-to-end measurements?
  - Best to do some “**user tests**” for open data as we go

Summary

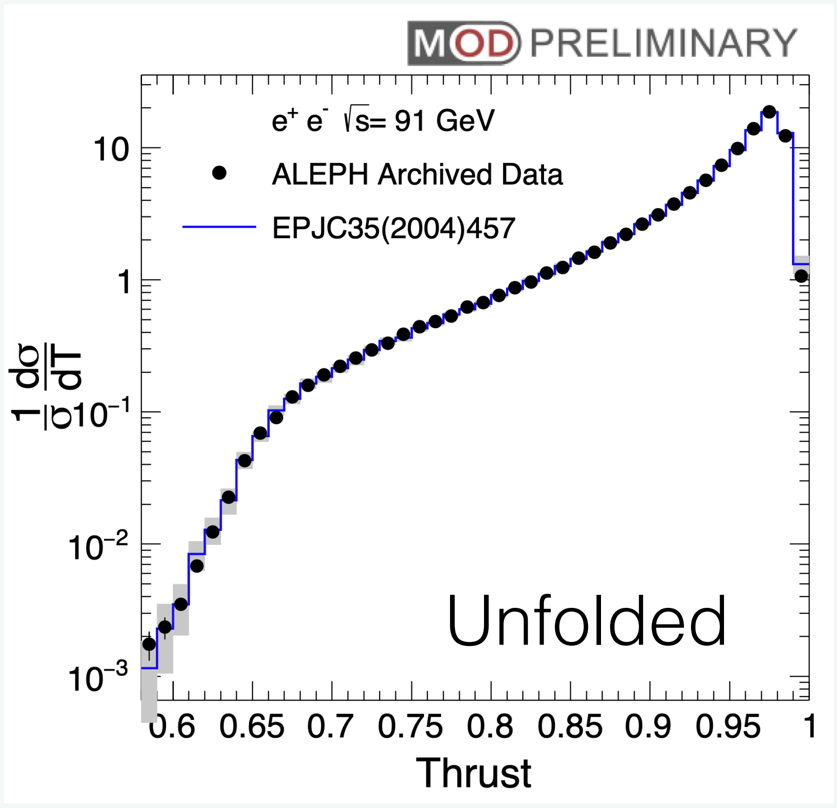
- Archived data is a gold mine with many exciting opportunities
    - QCD studies, new ideas, new algorithms, …
- Food for thought for ongoing experiments: preservation of knowledge, multiple MC samples, ability to rerun key software, low-level information, …



Backup Slides Ahead

Reproducing published results

- Comprehensive data/MC comparisons
- Convince ourselves that we understand the data
- Exact selection as QCD paper
- Thrust $T \equiv \max_{\hat{n}} \frac{\Sigma_i |\vec{p}_i \cdot \hat{n}\,|}{\Sigma_i |\vec{p}_i\,|}$
- Global event shape
    - Back to back dijet: T ~ 1


MOD PRELIMINARY
$e^+ e^-$ $\sqrt{s}$= 91 GeV
● ALEPH Archived Data
EPJC35(2004)457
Unfolded
$\frac{1}{\sigma}\frac{d\sigma}{dT}$
10^{-1}
10^{-2}
10^{-3}
Thrust