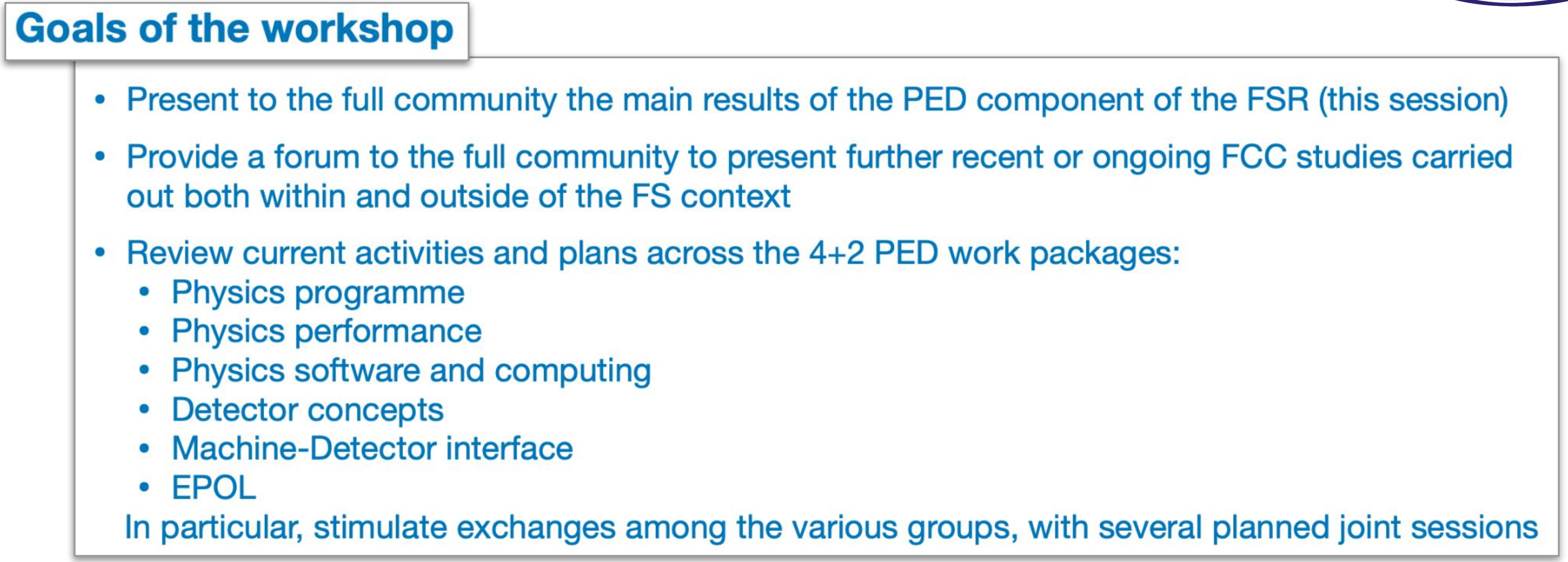
Software for PED studies

Software & Computing Summary

8th FCC Physics Workshop 2025
CERN

January 16, 2025
G Ganis, CERN-EP

Software & Computing Sessions



Goals of the workshop

- Present to the full community the main results of the PED component of the FSR (this session)
- Provide a forum to the full community to present further recent or ongoing FCC studies carried out both within and outside of the FS context
- Review current activities and plans across the 4+2 PED work packages:
  - Physics programme
  - Physics performance
  - Physics software and computing
  - Detector concepts
  - Machine-Detector interface
  - EPOL

  In particular, stimulate exchanges among the various groups, with several planned joint sessions

2 plenary sessions: Main results of the FSR , Workshop Summary
6 parallel sessions: Analysis, Beam-Bkg (w/ MDI), Key4hep & co, Detector Description (w/ Det.Con.),
Digitisation/Reconstruction, Global Reconstruction (w/ Det.Con.&Phys.Perf.)

Software & Computing in the FSR

B. François

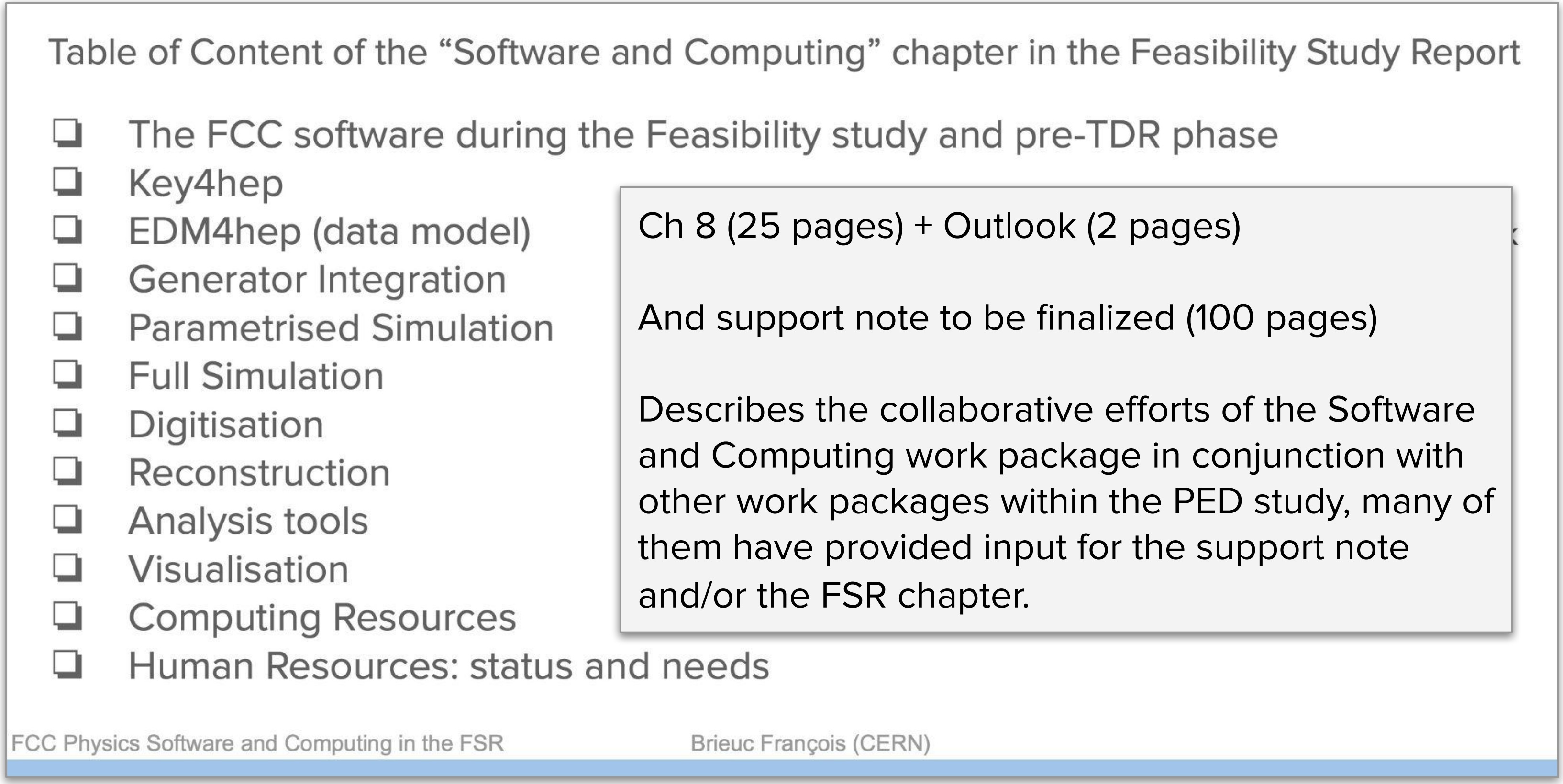


Table of Content of the “Software and Computing” chapter in the Feasibility Study Report

- ☐ The FCC software during the Feasibility study and pre-TDR phase
- ☐ Key4hep
- ☐ EDM4hep (data model)
- ☐ Generator Integration
- ☐ Parametrised Simulation
- ☐ Full Simulation
- ☐ Digitisation
- ☐ Reconstruction
- ☐ Analysis tools
- ☐ Visualisation
- ☐ Computing Resources
- ☐ Human Resources: status and needs

Ch 8 (25 pages) + Outlook (2 pages)

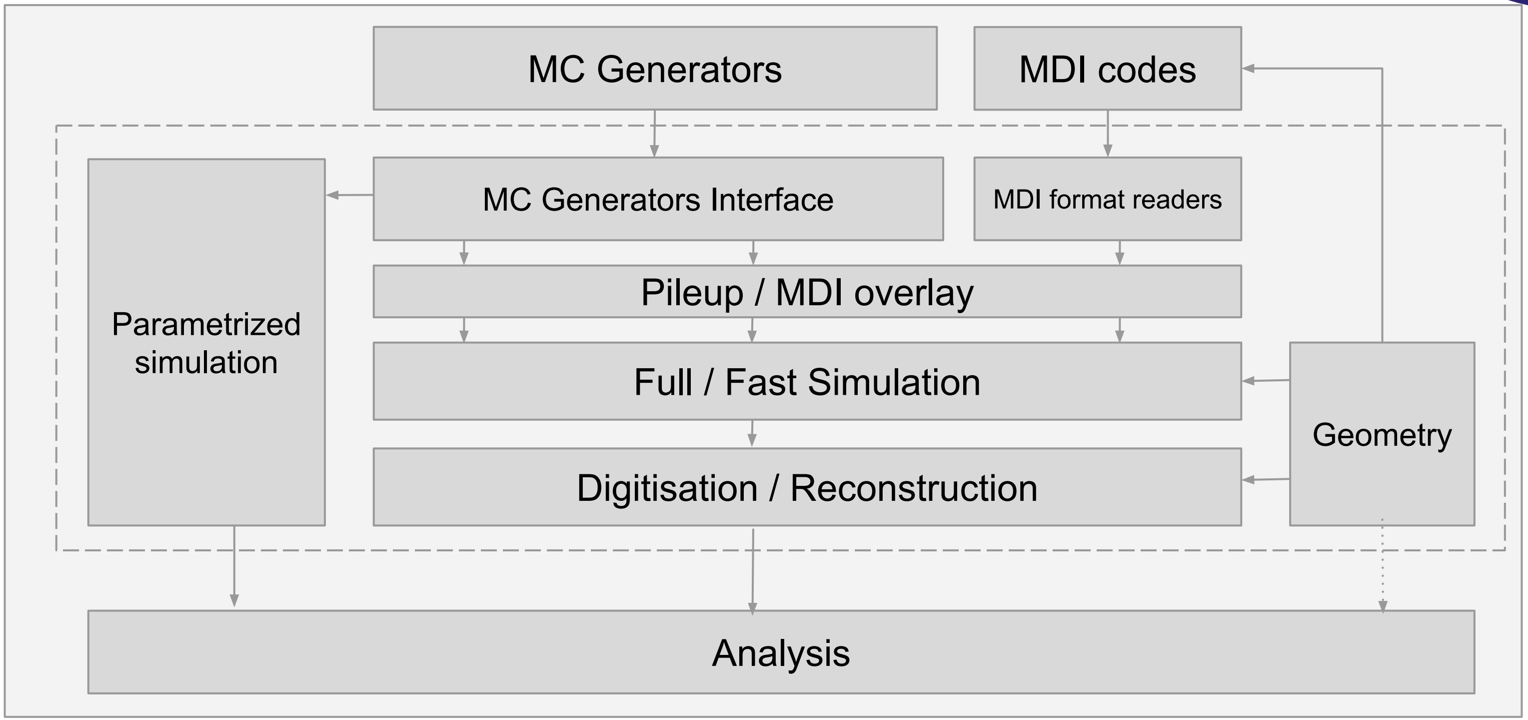
And support note to be finalized (100 pages)

Describes the collaborative efforts of the Software and Computing work package in conjunction with other work packages within the PED study, many of them have provided input for the support note and/or the FSR chapter.

FCC Physics Software and Computing in the FSR

Brieuc François (CERN)

Reminder of the workflows to support


MC Generators

MDI codes

MC Generators Interface

MDI format readers

Parametrized simulation

Pileup / MDI overlay

Full / Fast Simulation

Geometry

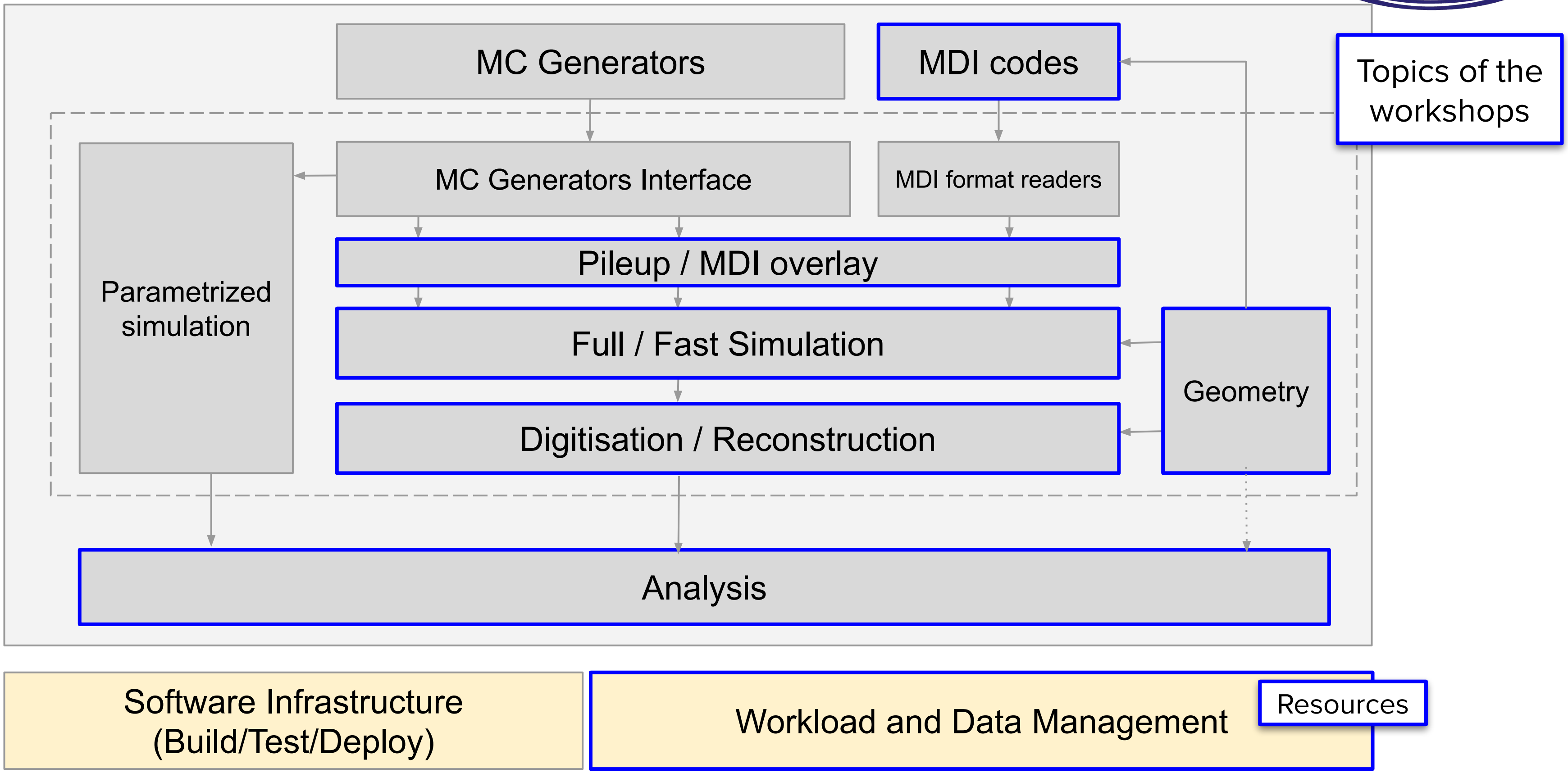
Digitisation / Reconstruction

Analysis

Software Infrastructure (Build/Test/Deploy)

Workload and Data Management

Reminder of the workflows to support



G Ganis, S&C, Summary, 8th FCC Physics Workshop 2025, 16 January 2025

Selected topics

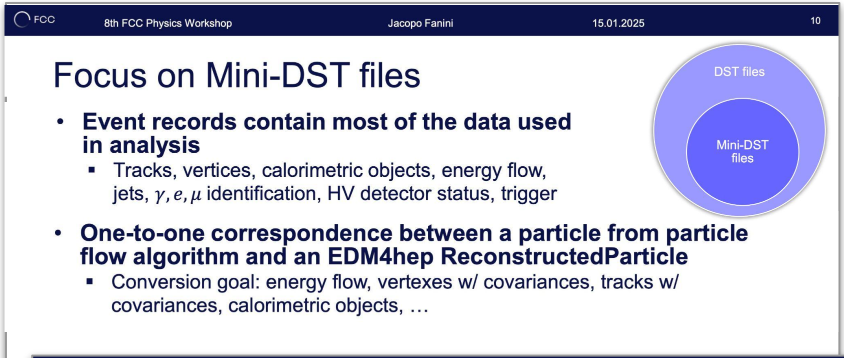
- Key4HEP
- Analysis
- MDI
- Detector Concepts
- Particle Flow

Key4hep

- Project started in 2019, essential for FCC (also for workforce)
  - Role in the community consolidating
- Status and recent developments/improvements J.M. Carceller
  - Consolidation of EDM4hep (v1.00 in line of view)
  - Use of multi-threading aware components in Gaudi (Functional, IOSvc, …)
  - Adoption of Marlin algorithms from iLCSoft as needed using k4MarlinWrapper
    - OverlayTiming: ovary background events (e.g. from beams) to signal events (see later)
    - Porting to Gaudi still not much advanced
  - Improve quality and availability of builds, increase OS support
- Recent workshop to discuss status and future of the project A. Sailer
  - Sustainability after R&D (consortium?), EPPSU contribution
  - Discussion on broader technical issues
    - Build system, EDM aspects (extension of datatypes, backward compatibility), interfaces with common packages (Pandora, ACTS, …), analysis (line between Key4hep and other ecosystems)
  - Message for the management
    - "Support for core Key4hep is crucial to keep momentum going"

ALEPH data in EDM4HEP

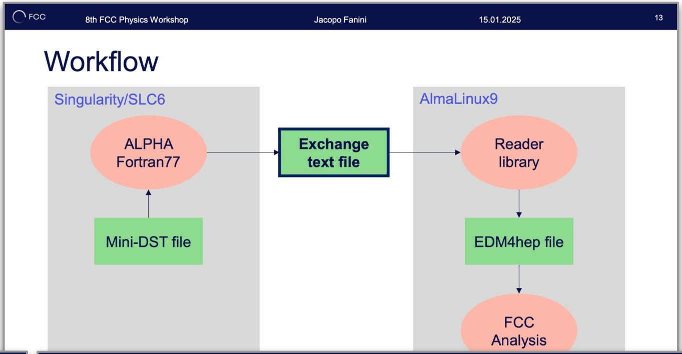
J. Fanini



FCC 8th FCC Physics Workshop Jacopo Fanini 15.01.2025 10

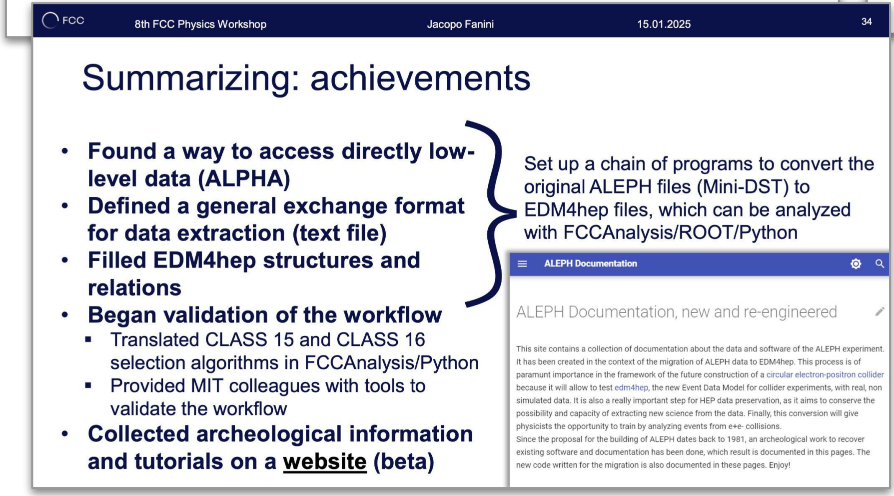
Focus on Mini-DST files

- **Event records contain most of the data used in analysis**
  - Tracks, vertices, calorimetric objects, energy flow, jets, γ, e, μ identification, HV detector status, trigger
- **One-to-one correspondence between a particle from particle flow algorithm and an EDM4hep ReconstructedParticle**
  - Conversion goal: energy flow, vertexes w/ covariances, tracks w/ covariances, calorimetric objects, …



FCC 8th FCC Physics Workshop Jacopo Fanini 15.01.2025 13

Workflow



FCC 8th FCC Physics Workshop Jacopo Fanini 15.01.2025 34

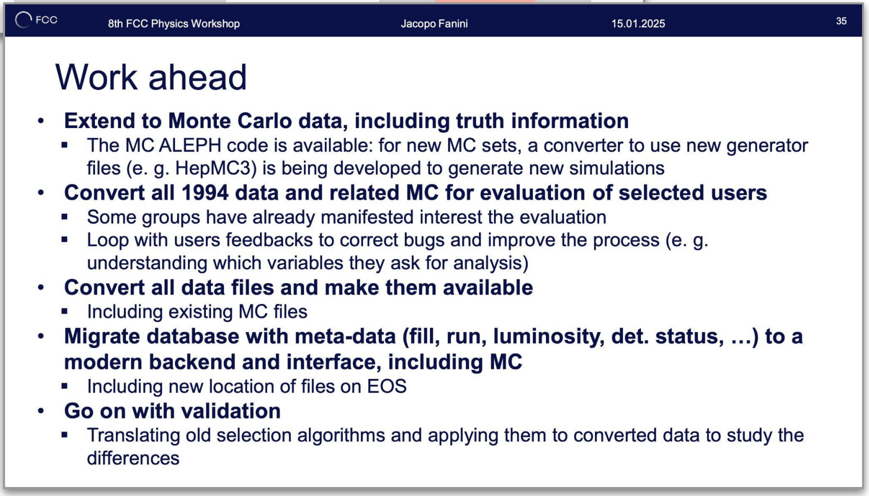
Summarizing: achievements

- **Found a way to access directly low-level data (ALPHA)**
- **Defined a general exchange format for data extraction (text file)**
- **Filled EDM4hep structures and relations**
- **Began validation of the workflow**
  - Translated CLASS 15 and CLASS 16 selection algorithms in FCCAnalysis/Python
  - Provided MIT colleagues with tools to validate the workflow
- **Collected archeological information and tutorials on a website (beta)**

Set up a chain of programs to convert the original ALEPH files (Mini-DST) to EDM4hep files, which can be analyzed with FCCAnalysis/ROOT/Python

ALEPH Documentation, new and re-engineered

This site contains a collection of documentation about the data and software of the ALEPH experiment. It has been created in the context of the migration of ALEPH data to EDM4hep. This process is of paramount importance in the framework of the future construction of a circular electron-positron collider because it will allow to test edm4hep, the new Event Data Model for collider experiments, with real, non simulated data. It is also a really important step for HEP data preservation, as it aims to conserve the possibility and capacity of extracting new science from the data. Finally, this conversion will give physicists the opportunity to train by analyzing events from e+e- collisions.
Since the proposal for the building of ALEPH dates back to 1981, an archeological work to recover existing software and documentation has been done, which result is documented in this pages. The new code written for the migration is also documented in these pages. Enjoy!



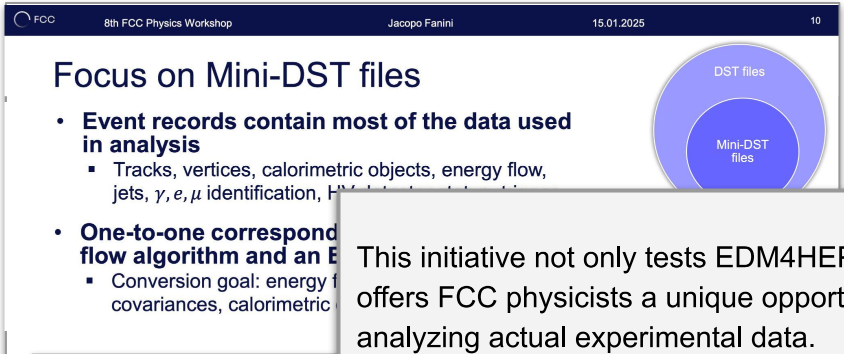
FCC 8th FCC Physics Workshop Jacopo Fanini 15.01.2025 35

Work ahead

- **Extend to Monte Carlo data, including truth information**
  - The MC ALEPH code is available: for new MC sets, a converter to use new generator files (e. g. HepMC3) is being developed to generate new simulations
- **Convert all 1994 data and related MC for evaluation of selected users**
  - Some groups have already manifested interest the evaluation
  - Loop with users feedbacks to correct bugs and improve the process (e. g. understanding which variables they ask for analysis)
- **Convert all data files and make them available**
  - Including existing MC files
- **Migrate database with meta-data (fill, run, luminosity, det. status, …) to a modern backend and interface, including MC**
  - Including new location of files on EOS
- **Go on with validation**
  - Translating old selection algorithms and applying them to converted data to study the differences

ALEPH data in EDM4HEP

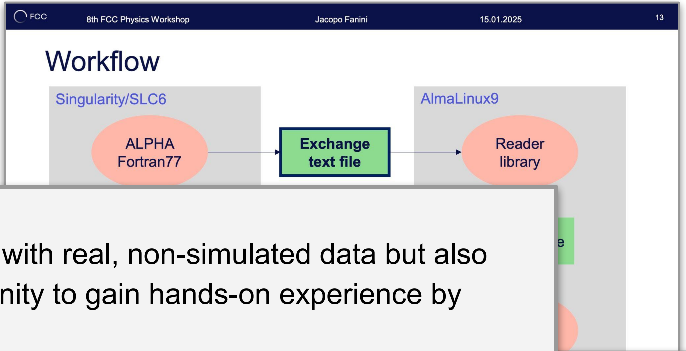
J. Fanini



FCC 8th FCC Physics Workshop Jacopo Fanini 15.01.2025 10

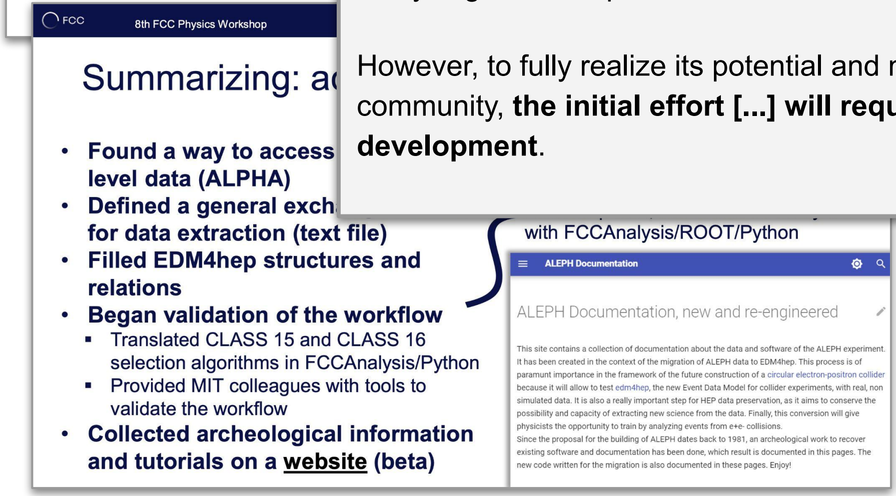
Focus on Mini-DST files

- **Event records contain most of the data used in analysis**
  - Tracks, vertices, calorimetric objects, energy flow, jets, γ, e, μ identification, H
- **One-to-one correspond flow algorithm and an E**
  - Conversion goal: energy f covariances, calorimetric



FCC 8th FCC Physics Workshop Jacopo Fanini 15.01.2025 13

Workflow



FCC 8th FCC Physics Workshop

Summarizing: a

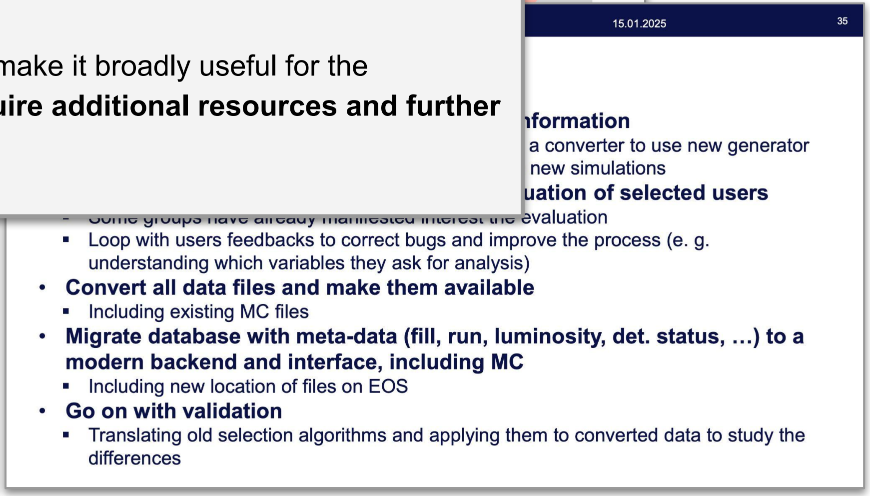
- **Found a way to access level data (ALPHA)**
- **Defined a general exch for data extraction (text file)**
- **Filled EDM4hep structures and relations**
- **Began validation of the workflow**
  - Translated CLASS 15 and CLASS 16 selection algorithms in FCCAnalysis/Python
  - Provided MIT colleagues with tools to validate the workflow
- **Collected archeological information and tutorials on a website (beta)**

with FCCAnalysis/ROOT/Python

ALEPH Documentation

ALEPH Documentation, new and re-engineered

This site contains a collection of documentation about the data and software of the ALEPH experiment. It has been created in the context of the migration of ALEPH data to EDM4hep. This process is of paramunt importance in the framework of the future construction of a circular electron-positron collider because it will allow to test edm4hep, the new Event Data Model for collider experiments, with real, non simulated data. It is also a really important step for HEP data preservation, as it aims to conserve the possibility and capacity of extracting new science from the data. Finally, this conversion will give physicists the opportunity to train by analyzing events from e+e- collisions.
Since the proposal for the building of ALEPH dates back to 1981, an archeological work to recover existing software and documentation has been done, which result is documented in this pages. The new code written for the migration is also documented in these pages. Enjoy!



15.01.2025 35

nformation
a converter to use new generator
new simulations
uation of selected users

- Some groups have already manifested interest the evaluation
- Loop with users feedbacks to correct bugs and improve the process (e. g. understanding which variables they ask for analysis)

- **Convert all data files and make them available**
  - Including existing MC files
- **Migrate database with meta-data (fill, run, luminosity, det. status, …) to a modern backend and interface, including MC**
  - Including new location of files on EOS
- **Go on with validation**
  - Translating old selection algorithms and applying them to converted data to study the differences

This initiative not only tests EDM4HEP with real, non-simulated data but also offers FCC physicists a unique opportunity to gain hands-on experience by analyzing actual experimental data.

However, to fully realize its potential and make it broadly useful for the community, **the initial effort [...] will require additional resources and further development**.

Analysis

- FCCAnalysis: framework based on RDataFrame, ROOT current analysis technology
- Complications arise by the fact that EDM4HEP is not a FLAT ntuple
  - FCCAnalysis provides helper tools, not always user-friendly
    - New development based on RDataSource helps
      - But performance needs optimisation
  - Room for alternative approaches
    - À la LHC (CMS, …)
      - Coffea: fully Python based system
    - Look to new languages
      - Julia, developed for scientific calculations, bringing the best out of Python and C++
        - Ready for analysis today

J. Smiesko

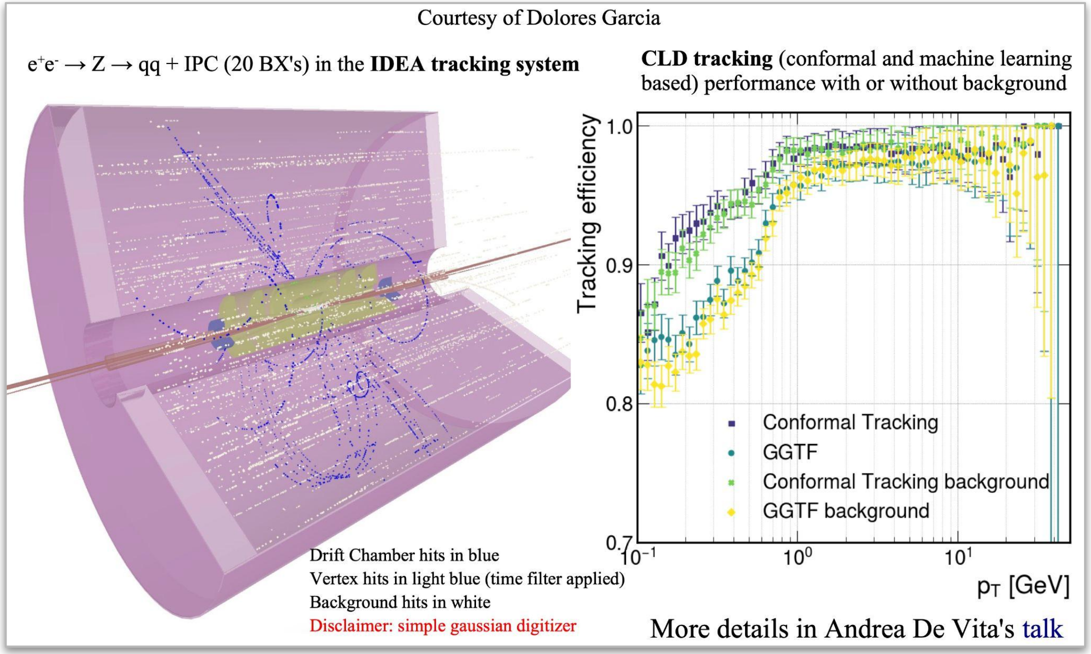
P. Yadav

P. Mato

MDI

- OverlayTiming in action

  B. François



Courtesy of Dolores Garcia

$e^+e^- \rightarrow Z \rightarrow qq$ + IPC (20 BX's) in the **IDEA tracking system**

CLD tracking (conformal and machine learning based) performance with or without background

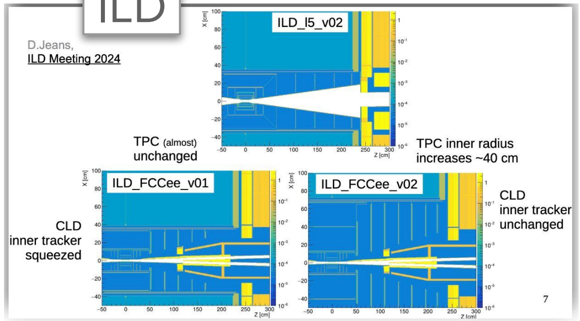
Drift Chamber hits in blue
Vertex hits in light blue (time filter applied)
Background hits in white
Disclaimer: simple gaussian digitizer

More details in Andrea De Vita's talk

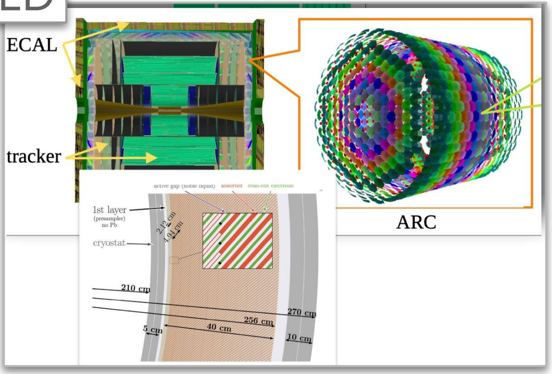
- First evaluation of fluences, doses and detector backgrounds w/ FLUKA A. Frasca et al.
  - Requiring re-implementations of (approximate) detector geometries
  - Should investigate the possibility of using Geant4 for this to try to minimize duplications
    - See V. Ivantchenko about Geant4 capabilities

All detector concepts ready for full sim

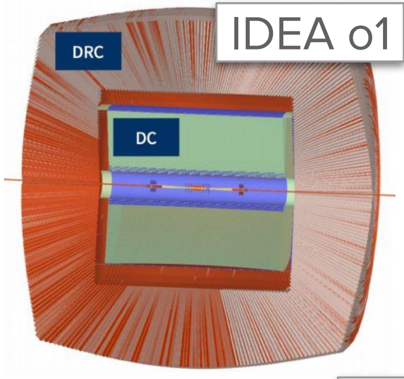
ILD



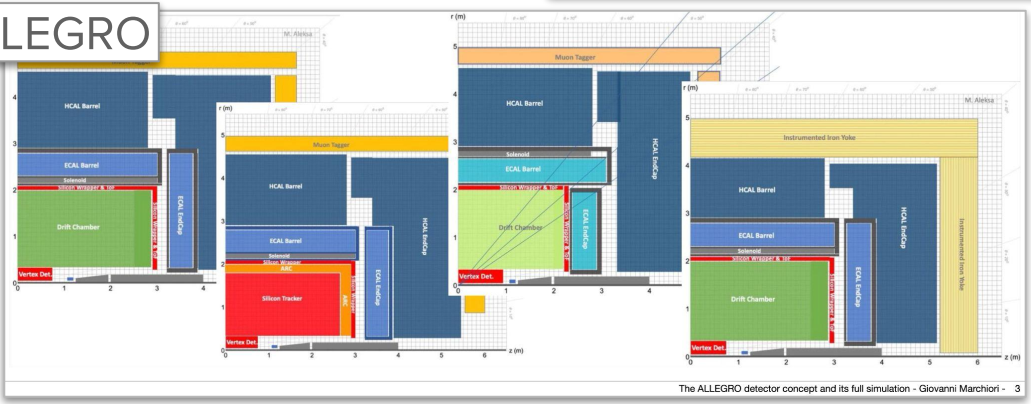
CLD



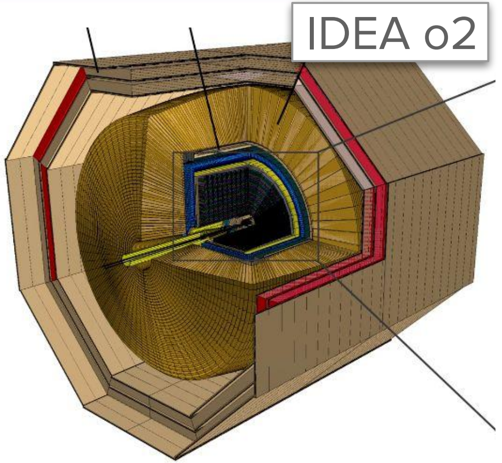
IDEA o1



ALLEGRO

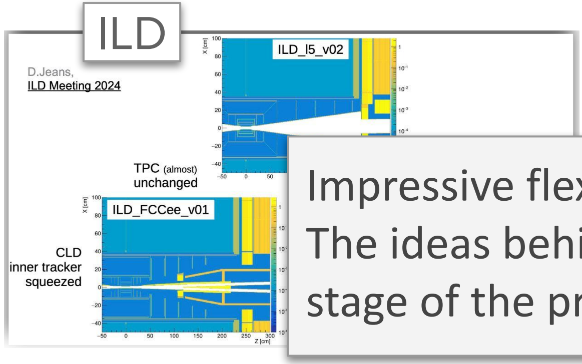


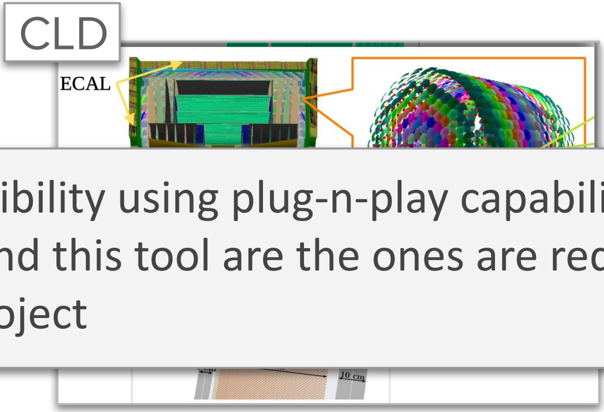
IDEA o2

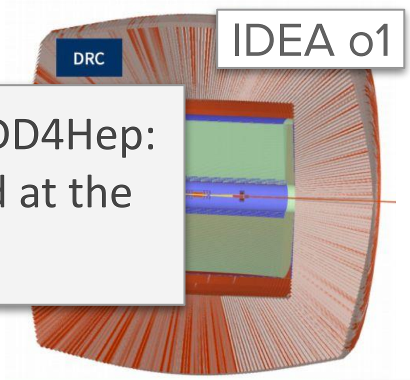


F Gaede, L Pezzotti, N Morange

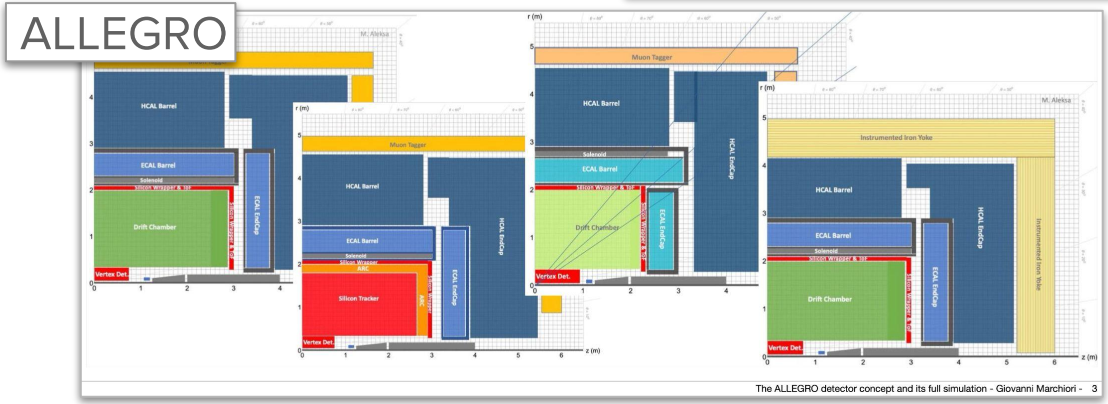
All detector concepts ready for full sim

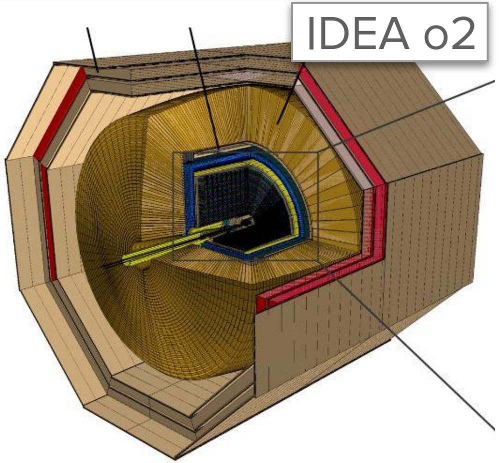

ILD


CLD


IDEA o1

Impressive flexibility using plug-n-play capabilities DD4Hep: The ideas behind this tool are the ones are required at the stage of the project


ALLEGRO


IDEA o2

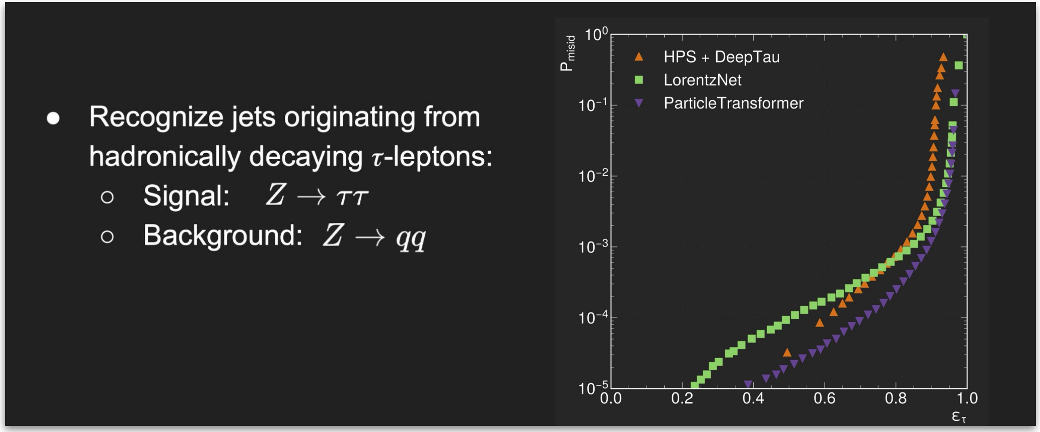
F Gaede, L Pezzotti, N Morange

Reconstruction (selected)

- Digitisers for the IDEA Drift chamber

  N. De Filippis
  - A path to to a realistic digitisation has been defined
  - Lower steps (L0 digitisation) rely on codes Garfield++/HEED on fragile support path
- ML based tau identification

  L. Tani
  - ClicDet Fullsim datasets via k4MarlinWrappers
  - End-to-end ML, easy for retraining, CMS inspired

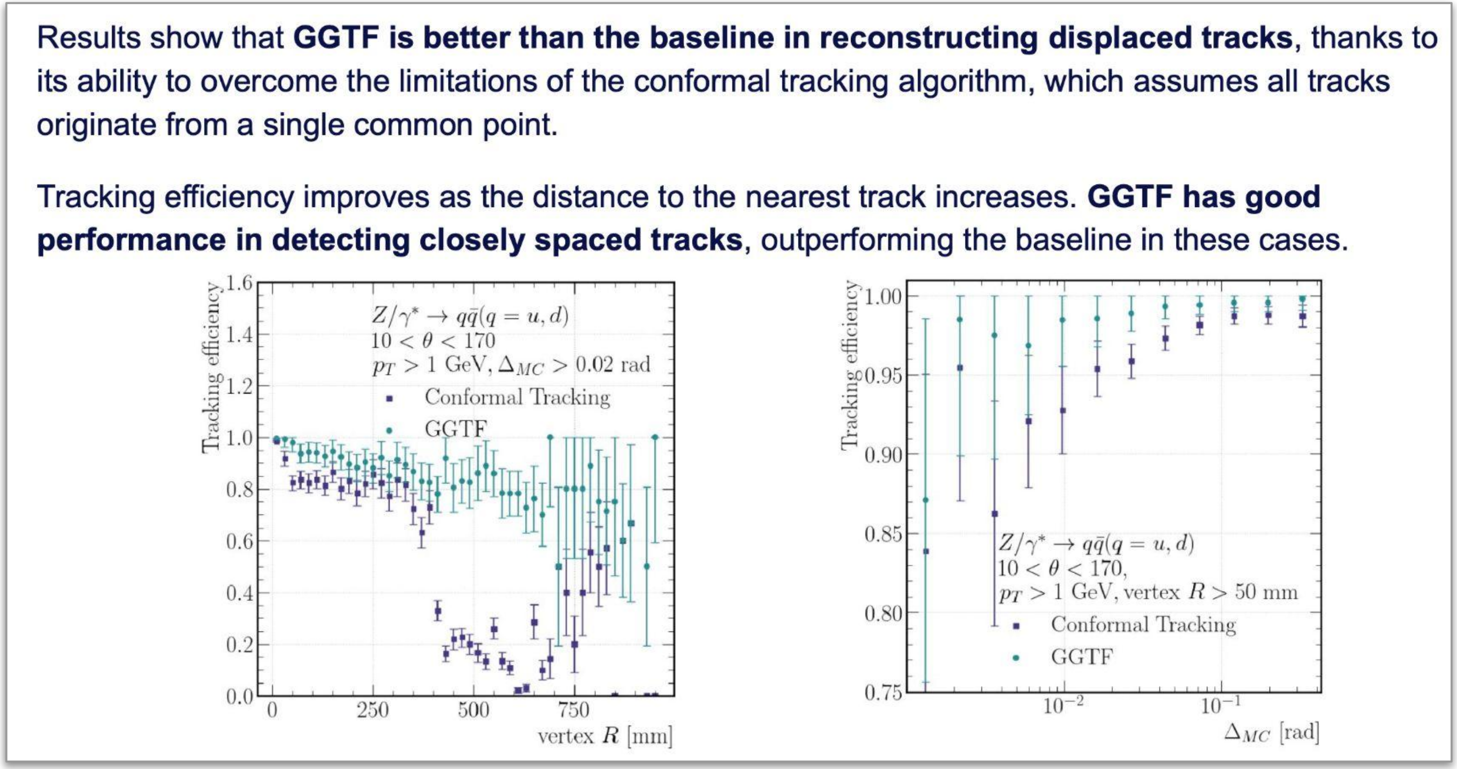


- Recognize jets originating from hadronically decaying τ-leptons:
  - Signal: $Z \to \tau\tau$
  - Background: $Z \to qq$

Reconstruction (selected)

- ML based tracking

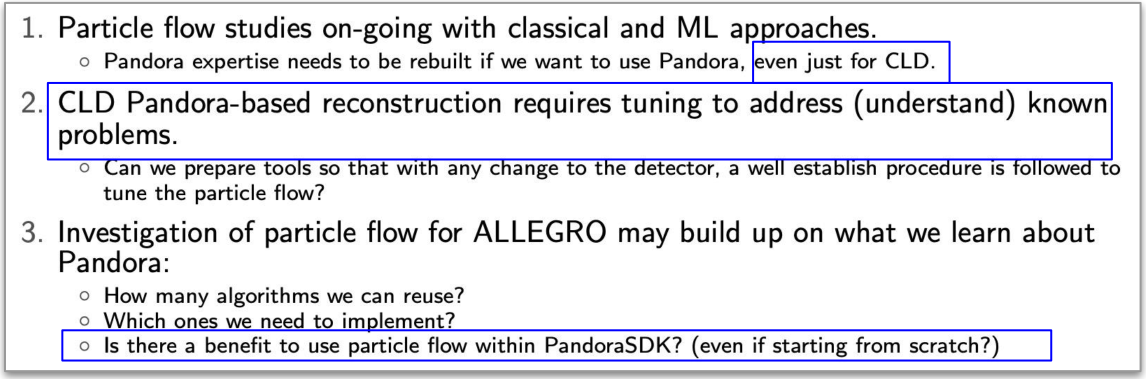
A. De Vita



Results show that **GGTF is better than the baseline in reconstructing displaced tracks**, thanks to its ability to overcome the limitations of the conformal tracking algorithm, which assumes all tracks originate from a single common point.

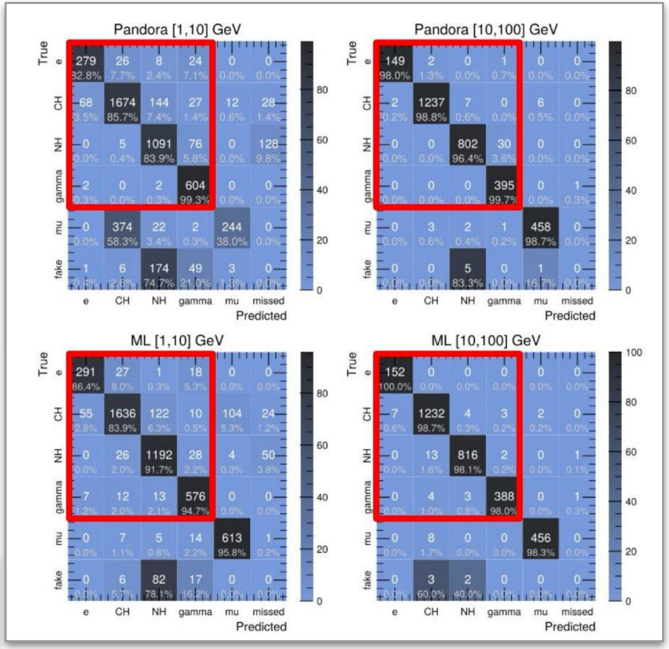
Tracking efficiency improves as the distance to the nearest track increases. **GGTF has good performance in detecting closely spaced tracks**, outperforming the baseline in these cases.

Particle Flow: Pandora and Co.



1. Particle flow studies on-going with classical and ML approaches.
    - Pandora expertise needs to be rebuilt if we want to use Pandora, even just for CLD.
2. CLD Pandora-based reconstruction requires tuning to address (understand) known problems.
    - Can we prepare tools so that with any change to the detector, a well establish procedure is followed to tune the particle flow?
3. Investigation of particle flow for ALLEGRO may build up on what we learn about Pandora:
    - How many algorithms we can reuse?
    - Which ones we need to implement?
    - Is there a benefit to use particle flow within PandoraSDK? (even if starting from scratch?)

A. Zaborowska



Pandora [1,10] GeV

	e	CH	NH	gamma	mu	missed
e	279 82.8%	26 7.7%	8 2.4%	24 7.1%	0 0.0%	0 0.0%
CH	68 3.5%	1674 85.7%	144 7.4%	27 1.4%	12 0.6%	28 1.4%
NH	0 0.0%	5 0.4%	1091 83.9%	76 5.8%	0 0.0%	128 9.8%
gamma	2 0.3%	0 0.0%	2 0.3%	604 99.3%	0 0.0%	0 0.0%
mu	0 0.0%	374 58.3%	22 3.4%	2 0.3%	244 38.0%	0 0.0%
fake	1 0.4%	6 2.6%	174 74.7%	49 21.0%	3 1.3%	0 0.0%

Pandora [10,100] GeV

	e	CH	NH	gamma	mu	missed
e	149 98.0%	2 1.3%	0 0.0%	1 0.7%	0 0.0%	0 0.0%
CH	2 0.2%	1237 98.8%	7 0.6%	0 0.0%	6 0.5%	0 0.0%
NH	0 0.0%	0 0.0%	802 96.4%	30 3.6%	0 0.0%	0 0.0%
gamma	0 0.0%	0 0.0%	0 0.0%	395 99.7%	0 0.0%	1 0.3%
mu	0 0.0%	3 0.6%	2 0.4%	1 0.2%	458 98.7%	0 0.0%
fake	0 0.0%	0 0.0%	5 83.3%	0 0.0%	1 16.7%	0 0.0%

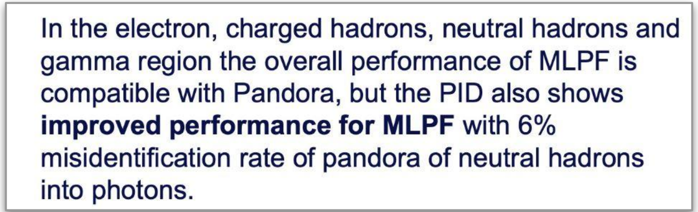
ML [1,10] GeV

	e	CH	NH	gamma	mu	missed
e	291 86.4%	27 8.0%	1 0.3%	18 5.3%	0 0.0%	0 0.0%
CH	55 2.8%	1636 83.9%	122 6.3%	10 0.5%	104 5.3%	24 1.2%
NH	0 0.0%	26 2.0%	1192 91.7%	28 2.2%	4 0.3%	50 3.8%
gamma	7 1.2%	12 2.0%	13 2.1%	576 94.7%	0 0.0%	0 0.0%
mu	0 0.0%	7 1.1%	5 0.8%	14 2.2%	613 95.8%	1 0.2%
fake	0 0.0%	6 5.7%	82 78.1%	17 16.2%	0 0.0%	0 0.0%

ML [10,100] GeV

	e	CH	NH	gamma	mu	missed
e	152 100.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
CH	7 0.6%	1232 98.7%	4 0.3%	3 0.2%	2 0.2%	0 0.0%
NH	0 0.0%	13 1.6%	816 98.1%	2 0.2%	0 0.0%	1 0.1%
gamma	0 0.0%	4 1.0%	3 0.8%	388 98.0%	0 0.0%	1 0.3%
mu	0 0.0%	8 1.7%	0 0.0%	0 0.0%	456 98.3%	0 0.0%
fake	0 0.0%	3 60.0%	2 40.0%	0 0.0%	0 0.0%	0 0.0%

A. De Vita

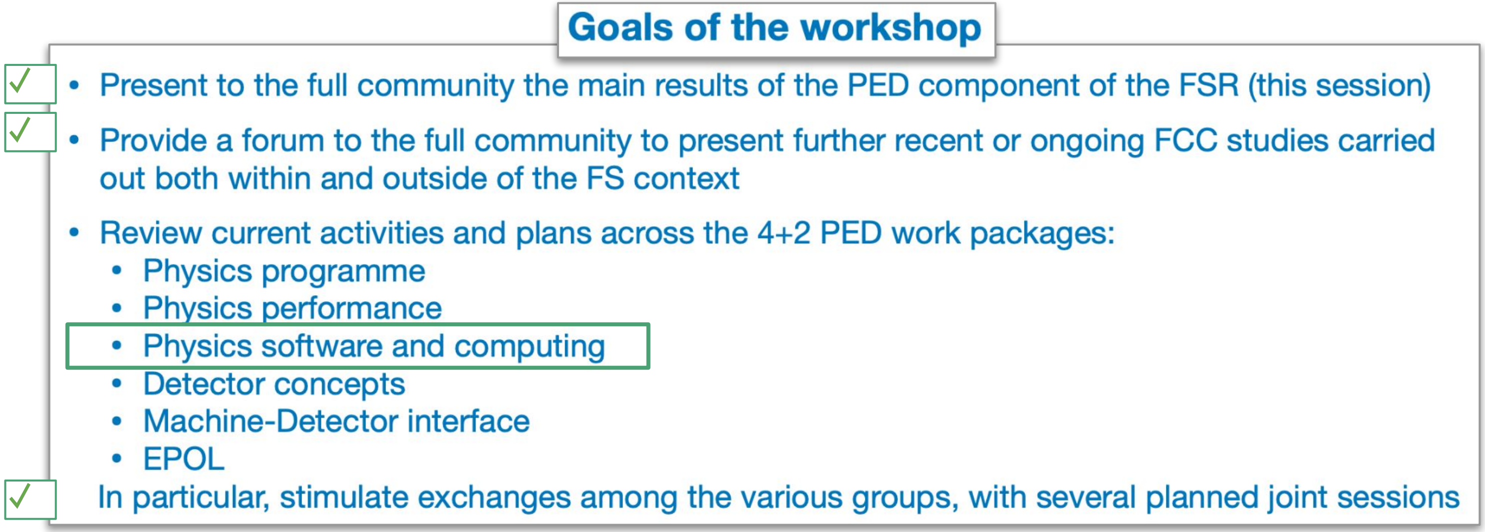


In the electron, charged hadrons, neutral hadrons and gamma region the overall performance of MLPF is compatible with Pandora, but the PID also shows **improved performance for MLPF** with 6% misidentification rate of pandora of neutral hadrons into photons.

Muons in the range of 1-10 GeV are correctly **classified by the MLPF 95%** of the times (38% for pandora).

Final Remarks

- The workshop provided a good overview of the current status of things
- Visibly increased interest and contributions to software
- Lot of progress since Annecy, especially in sim and reco, now reported in the FSR
    - Thanks to collaborative efforts with other work packages within the PED study



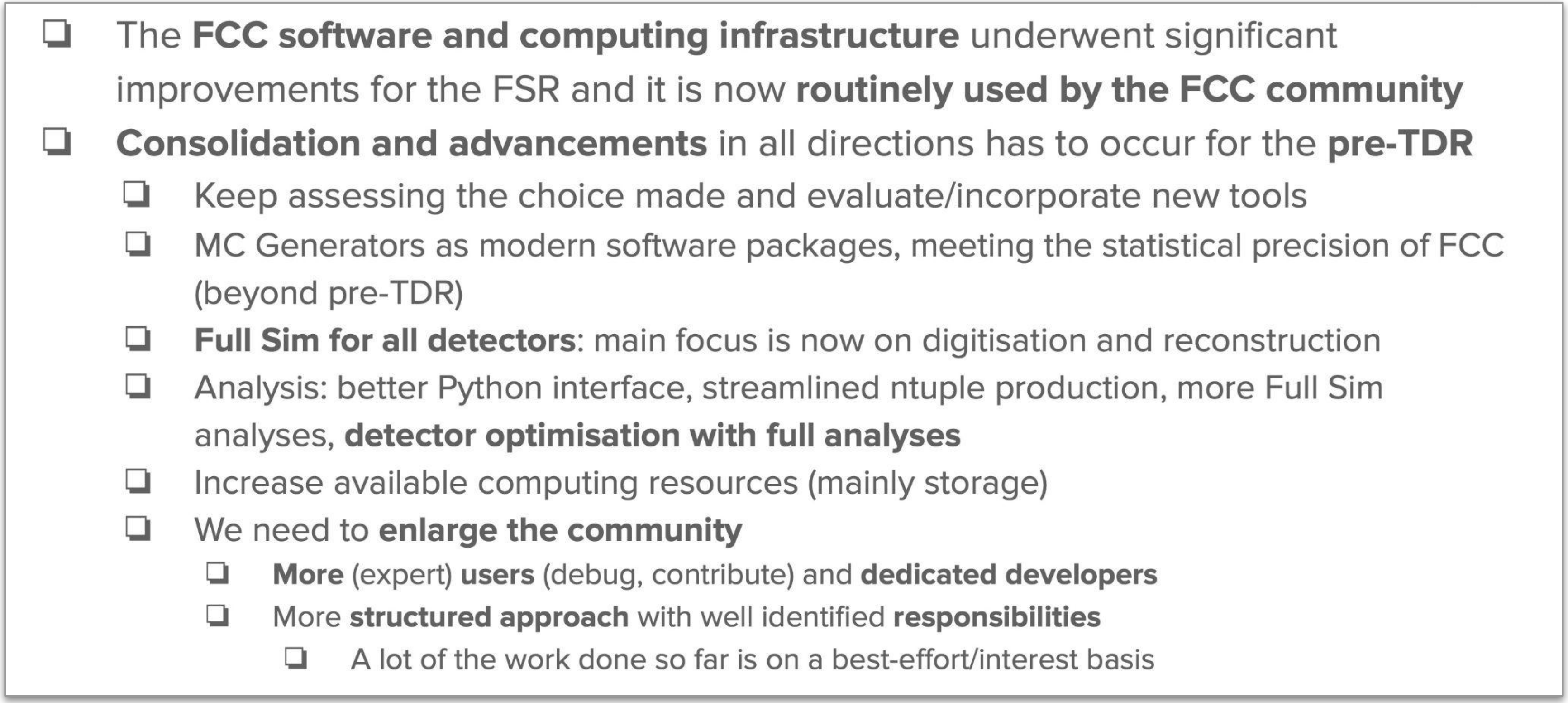
Goals of the workshop

- ✓ Present to the full community the main results of the PED component of the FSR (this session)
- ✓ Provide a forum to the full community to present further recent or ongoing FCC studies carried out both within and outside of the FS context
- Review current activities and plans across the 4+2 PED work packages:
    - Physics programme
    - Physics performance
    - Physics software and computing
    - Detector concepts
    - Machine-Detector interface
    - EPOL

✓ In particular, stimulate exchanges among the various groups, with several planned joint sessions

Outlook (from initial talk)

B. François

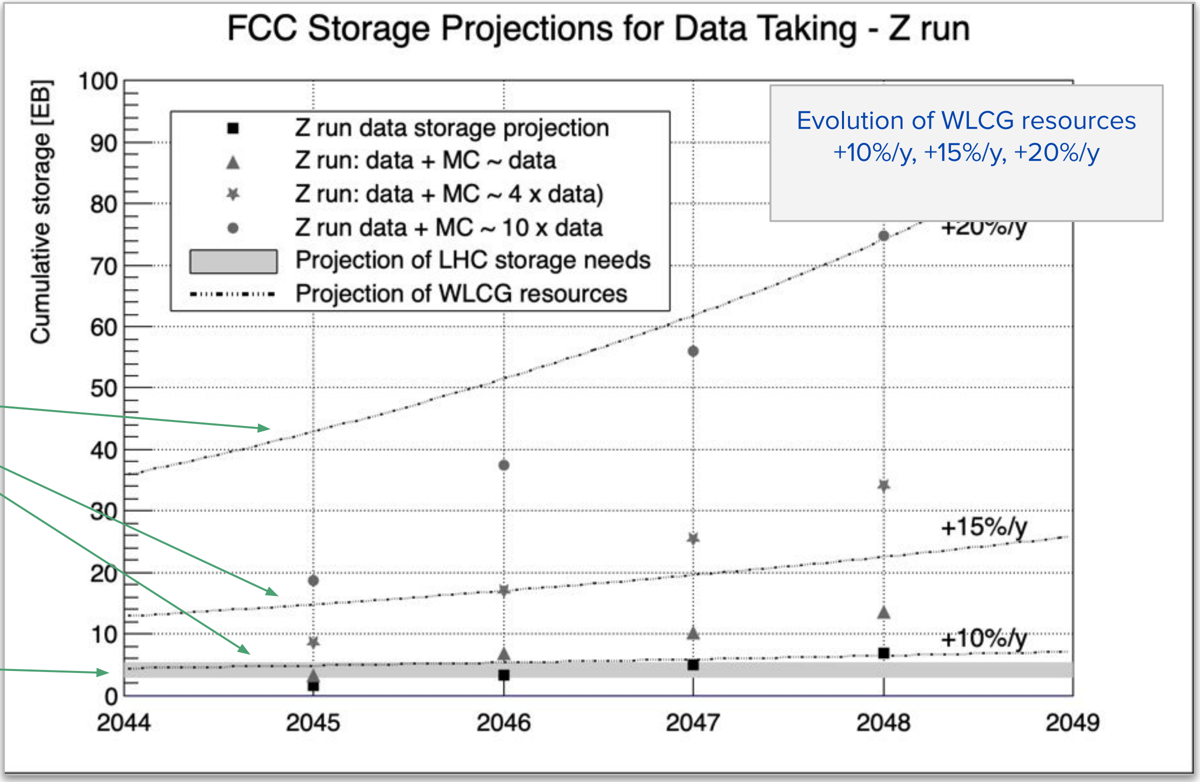


- The **FCC software and computing infrastructure** underwent significant improvements for the FSR and it is now **routinely used by the FCC community**
- **Consolidation and advancements** in all directions has to occur for the **pre-TDR**
  - Keep assessing the choice made and evaluate/incorporate new tools
  - MC Generators as modern software packages, meeting the statistical precision of FCC (beyond pre-TDR)
  - **Full Sim for all detectors**: main focus is now on digitisation and reconstruction
  - Analysis: better Python interface, streamlined ntuple production, more Full Sim analyses, **detector optimisation with full analyses**
  - Increase available computing resources (mainly storage)
  - We need to **enlarge the community**
    - **More** (expert) **users** (debug, contribute) and **dedicated developers**
    - More **structured approach** with well identified **responsibilities**
      - A lot of the work done so far is on a best-effort/interest basis

Backup

Illustrative Storage Projection for Z Run

4 experiments
4 equal runs {2045, 2046, 2047, 2048}



Evolution of WLCG resources
+10%/y, +15%/y, +20%/y

Evolution of WLCG resources
+10%/y, +15%/y, +20%/y

(starting point: 500 PB in 2020
≂ ATLAS+CMS + 10%)

LHC at the end of HL (≂ 5 EB)