

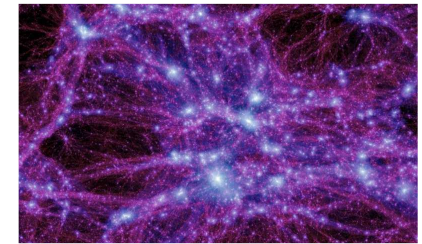
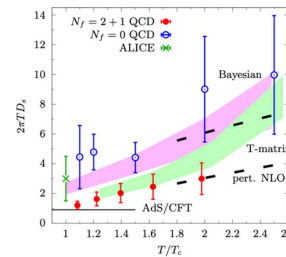
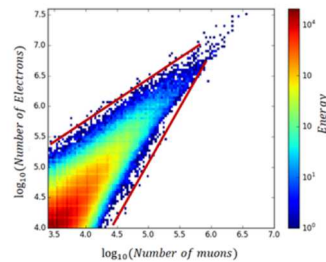
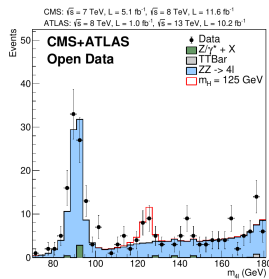
# Examples for work on data management - PUNCH4NFDI use cases and workflows



A.Geiser, DESY, 14.9.23, MU days, KIT, for **PUNCH4NFDI**:

**Particles, Universe, NuClei and Hadrons for the NFDI –**

**the consortium of particle, astroparticle, hadron&nuclear physics and astronomy**

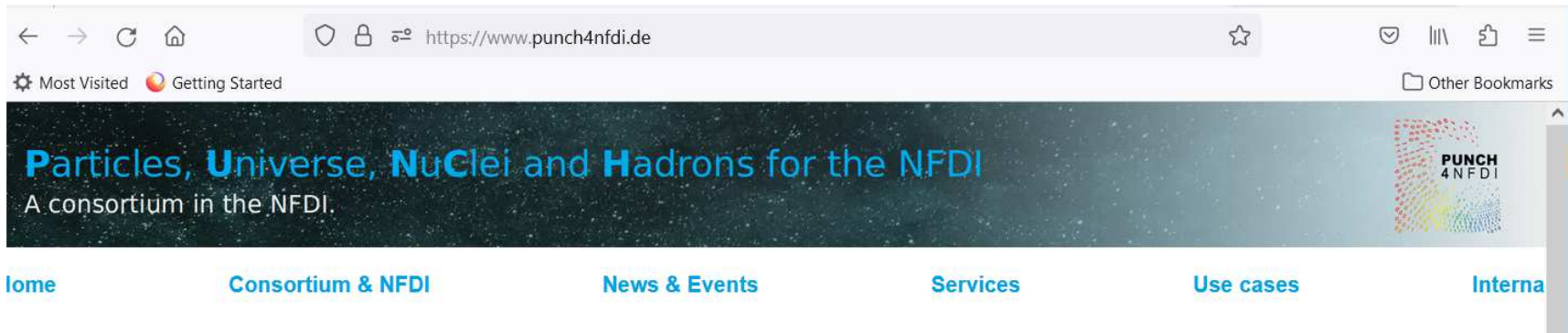


- Scope of the **PUNCH4NFDI** science data platform

**Use case/work flow examples:**

- Heavy Quark Diffusion in **Lattice QCD**
- Reduction of Interferometric **MeerKAT Data**
- Higgs Boson “Rediscovery” using **CERN Open Data**
- Summary and outlook

# Scope of the PUNCH4NFDI Science Data Platform



- via PUNCH4NFDI portal and related AAI, the **PUNCH Science Data Platform** aims at providing **coherent access to the rich data collections** of the PUNCH communities.
- uses **federated infrastructures and resources** (see also previous talks and next talk) like **Compute4PUNCH** (C4P) and **Storage4PUNCH** (S4P), to enable **complex workflows**. **Make them work together coherently also for users without special expertise.**
- special focus on eventual **cross-community uses** of the **heterogeneous data and software** from particle, astroparticle and hadron&nuclear physics and astronomy.

# Heavy Quark Diffusion in Lattice QCD

**Application bridging particle physics, heavy ion/nuclear physics, theory and high performance computing, and end user analysis**

- Calculation of diffusion coefficient for heavy quarks in the quark-gluon plasma (QGP).
- Ongoing studies towards physical pion masses – important input for hydro / transport models for the study of heavy quarks in the QGP. Also provides information on the thermalization of heavy quarks.

○ Gauge configurations generated from LQCD simulations of  $N_f=2+1$  at pion mass  $320\text{ MeV}$  using SIMULATEQCD [6]. Plan: upload all configurations in ILDG.



Computing resources:

- Bielefeld GPU cluster
- JUWELS at GCS@FZJ
- Marconi 100 at CINECA

Raw gauge configurations used to calculate observable of interest: chromoelectric correlator ( $G_E$ ) for various temperatures. All measured data public [5].

[5] O. Kaczmarek et al., <https://pub.uni-bielefeld.de/record/2979080>

[6] <https://latticeqcd.github.io/SIMULATEQCD>

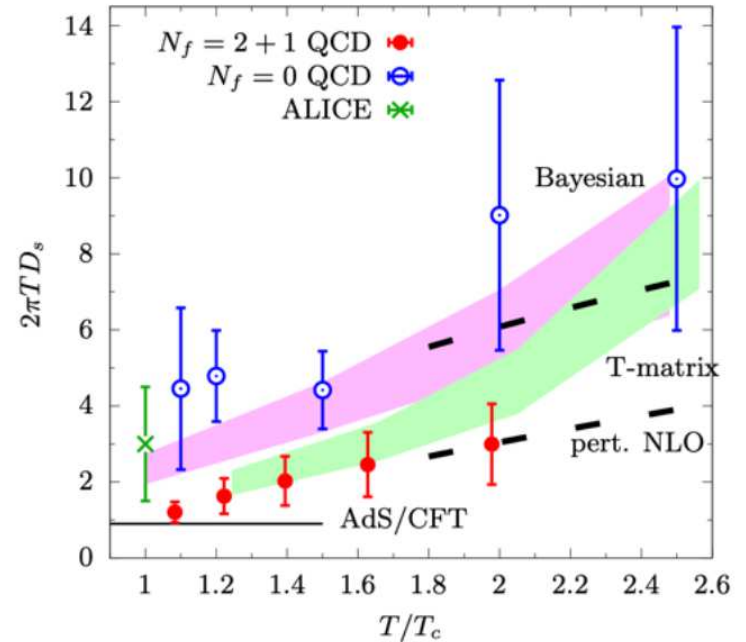
# Heavy Quark Diffusion in Lattice QCD

Raw gauge configurations used to calculate observable of interest: chromoelectric correlator ( $G_E$ ) for various temperatures. All measured data public [5].



Final analysis step

Complete workflow implemented as a single bash / python script running on C4P.

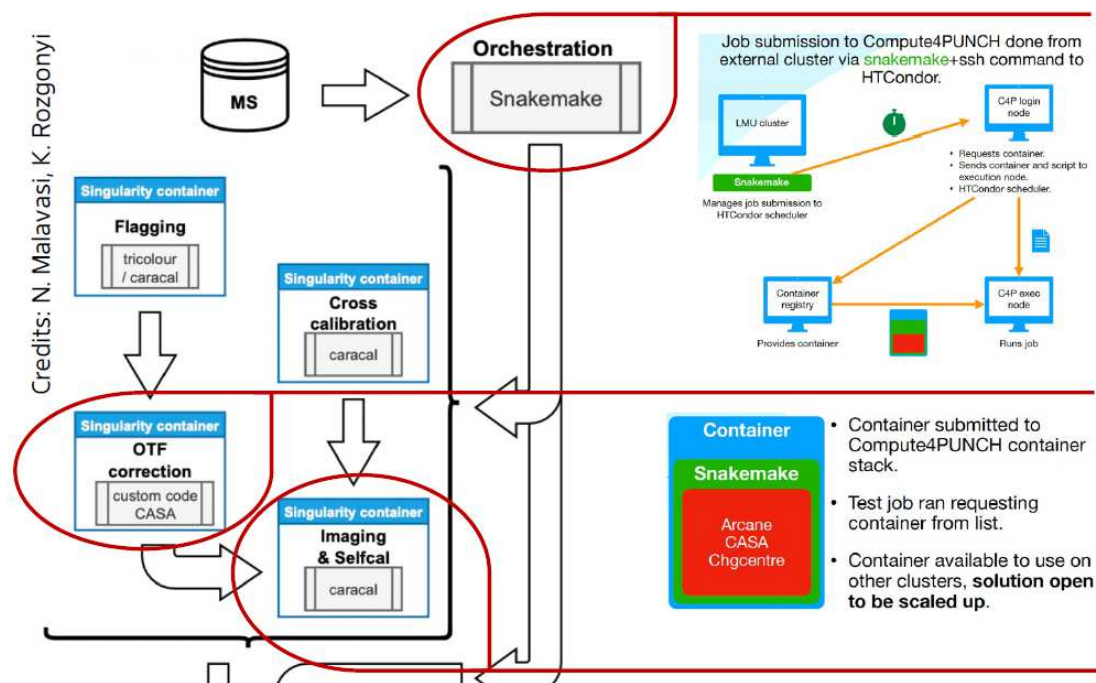


- First results on heavy quark diffusion coefficients using dynamical light quarks from the lattice!  
L. Altenkort et al., PRL 130 (2023) 231902
- Next steps:
  - C4P to provide infrastructure for running the container that includes the entire workflow
  - Metadata of the data sets to be designed and extracted
  - Metadata and data sets of the analysis to be stored in S4P
- End goal: Include the full analysis in one digital research product running on the PUNCH4NFDI science data platform

# Reduction of Interferometric MeerKAT Data

## Application of tools from both astronomy and particle physics to Radioastronomy data analysis.

- MeerKLASS [1] is a radio-astronomy observation campaign performed with MeerKAT, involving massive data volumes, requiring on-the-fly (OTF) data reduction and generation of radio images. The analysis chain [2] involves numerous software packages. Automation and resources will be dealt with using the PUNCH4NFDI resources C4P and S4P [3].

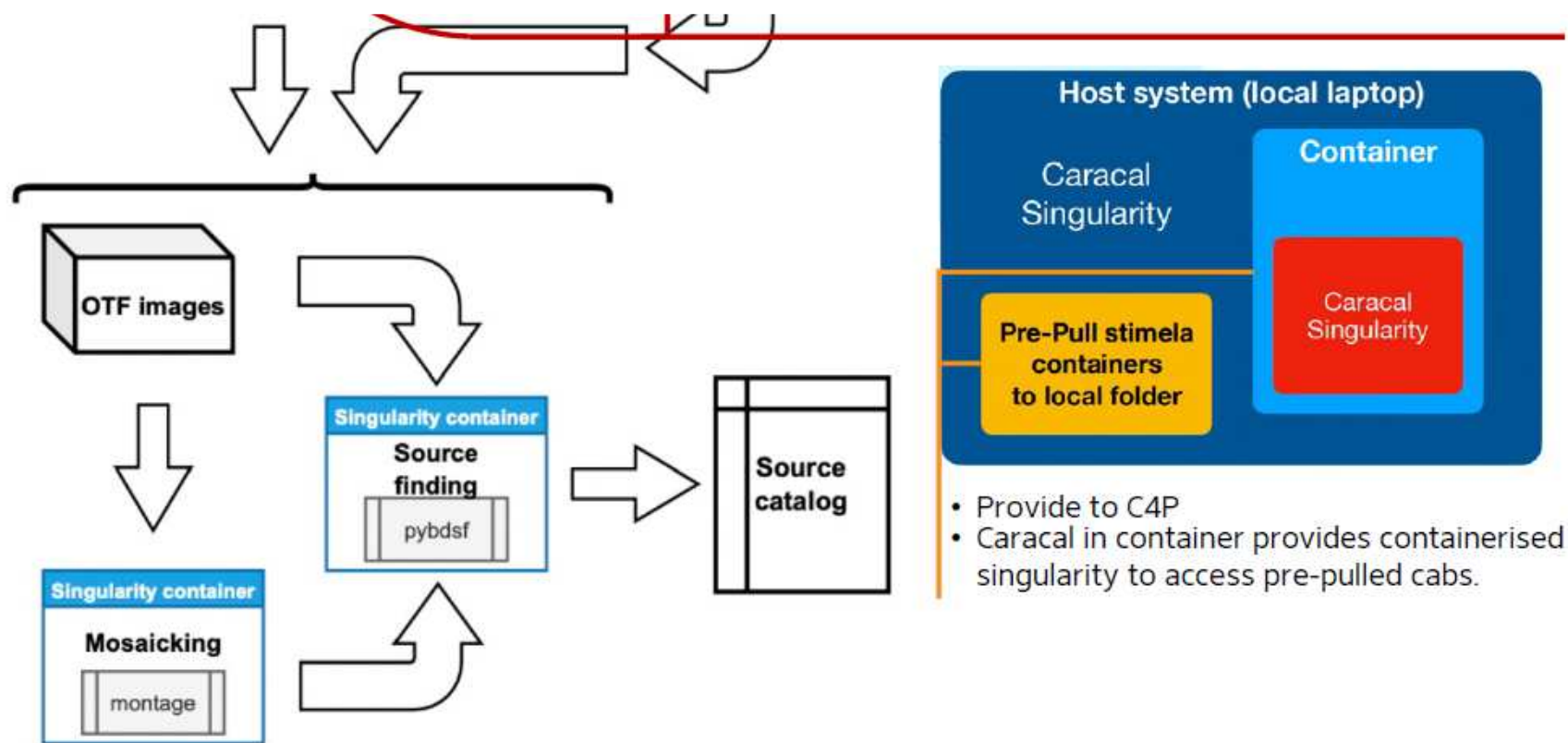


[1] MeerKLASS collaboration,  
<https://www.glowconsortium.de/index.php/en/meerkat-about/science>

[2] K. Rozgonyi et al., in preparation

[3] N. Malavasi et al., in preparation

# Reduction of Interferometric MeerKAT Data



## ○ Next steps:

- **Scale up OTF step** to large volumes on C4P, implement **imaging step** on C4P
- Use S4P to **store data** (streaming data from workflow)
- Orchestration scheme to be replaced by **REANA** once interfaced with C4P and S4P

# Combined CMS + ATLAS Higgs “Rediscovery”



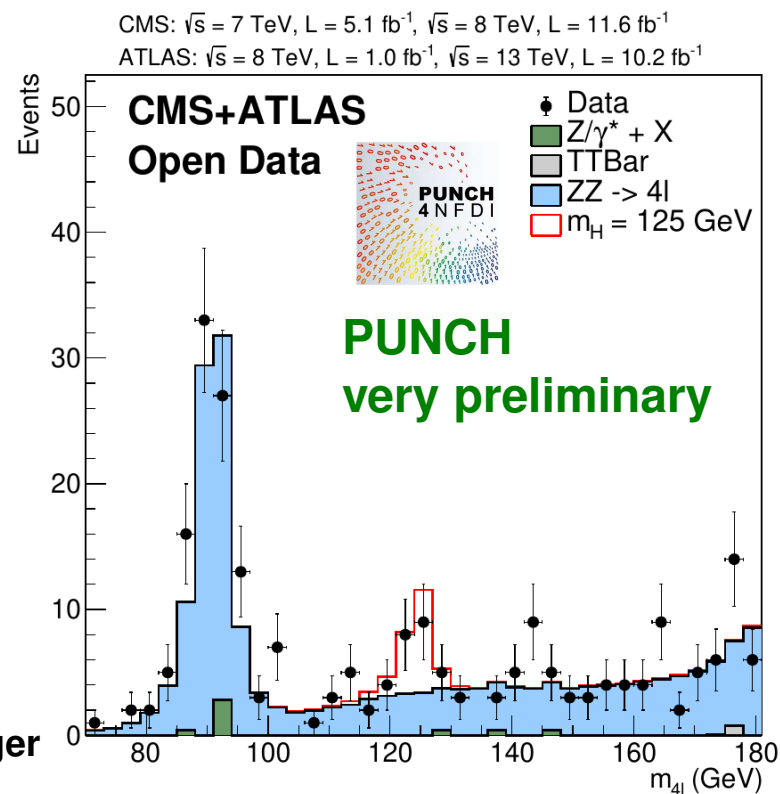
**First example of analysis of low level (open) data from two different HEP experiments within the same analysis framework.**

Main purpose: **demonstrate practical feasibility of a PUNCH use case on the PUNCH Science Data Platform going significantly beyond what is already available outside PUNCH** (i.e. not just an import of things already available elsewhere), **using PUNCH resources** already now wherever possible.

Relevant **PUNCH-transformed data sets** already produced  
and **stored on S4P** (including hifis)

**Full PUNCH workflow** (series of ROOT scripts)

being implemented on **PUNCH instance of REANA** workflow manager

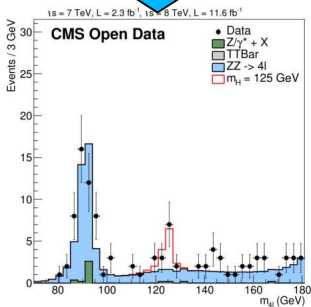


original CMS legacy research data  
(2 PB on CERN /eospublic via  
CERN Open Data portal)  
2010 data (100%, legacy format 1)  
and 2011/12 (70%, legacy format 2)

original CMS legacy software  
(from public github via  
CERN Open Data portal)  
(2 different versions, run on two  
different legacy VMs or containers)

produce  
histograms

many  
CPU  
months



original ATLAS legacy research data (not public)

by ATLAS collaboration

simplified educational  
ATLAS Open Data 2012  
(on CERN /eospublic via  
CERN Open Data portal)  
(10%, simplified format 3)

VM with dedicated  
software package 1

simplified educational  
ATLAS Open Data 2016  
(on separate ATLAS  
Open Data portal)  
(25%, simplified format 4)

VM with dedicated  
software package 2  
or Jupyter notebook

download

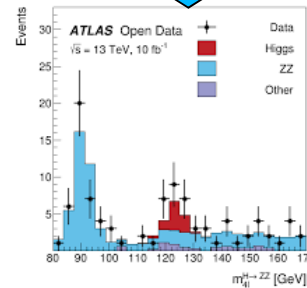
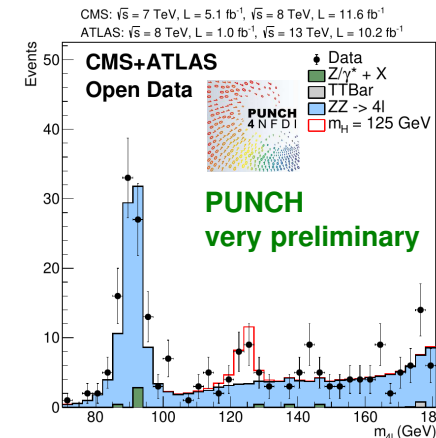
## PUNCH

apply data transformation  
interface (versions 1 & 2)

apply data transformation  
interface (versions 3 & 4)

76 different samples with **common  
unified & simplified research level  
data format**

“single” script, < 1 CPU day  
documentation+metadata



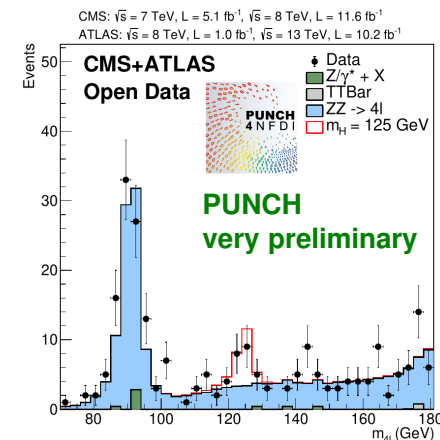
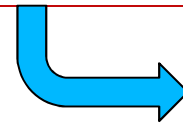


# What a PUNCH user will see on the platform once fully completed



76 different samples with common unified & simplified research level data format, via the PUNCH platform

“single” script, < 1 CPU day, documentation + metadata



details being finalized

# Summary and Outlook



- via PUNCH4NFDI portal and related AAI, the **PUNCH Science Data Platform** aims at providing **coherent access to the rich data collections** of the PUNCH communities, including **cross-community applications** on **federated infrastructures**.
- some **first example usecases/workflows** are **already being implemented, in parallel to the gradual implementation of the data management infrastructures**. Many more, and more ambitious ones, to come.
- these use-cases are intended as **examples and inspirations for users of the platform to design and implement their own applications on the platform** in a greatly simplified/unified environment which is so far not available anywhere else.