Progress in Data and Analysis Preservation in the PHENIX Experiment at RHIC

M.Potekhin for PHENIX Collaboration

Brookhaven National Laboratory

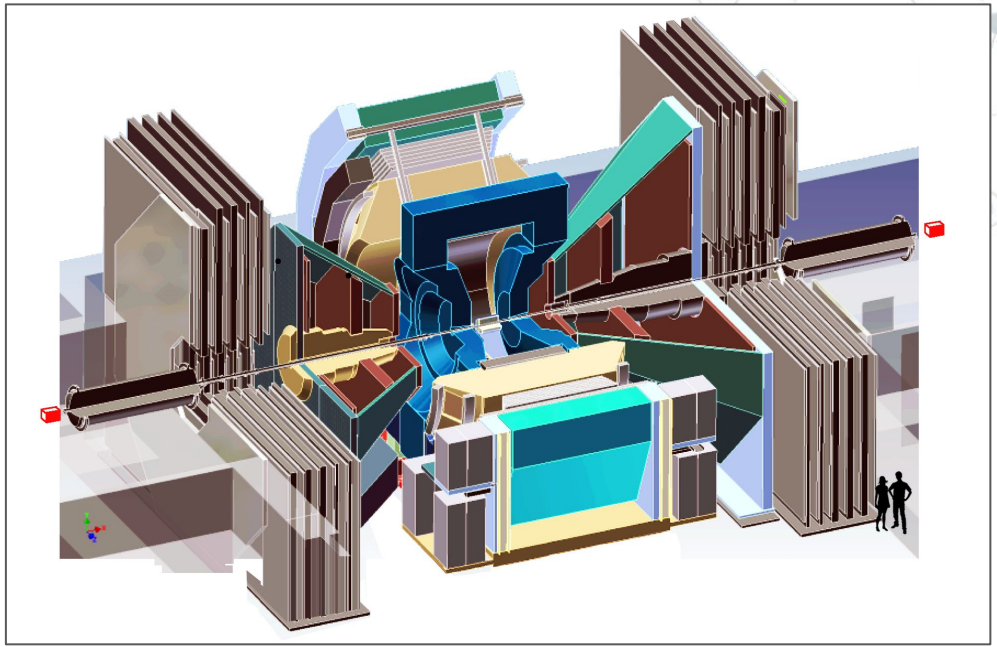
DPHEP 2024

About this presentation

- PHENIX has started its Data and Analysis Preservation (DAP) effort in 2019, and periodic status updates have been presented in the previous DPHEP meetings, PV2023 and two ACAT conferences (2021 and 2024). Useful links:
  - https://indico.cern.ch/event/1043155/timetable/#9-bnl-rhic
  - https://iopscience.iop.org/article/10.1088/1742-6596/2438/1/012020
  - https://doi.org/10.5281/zenodo.7905555
- This presentation is an update on the current status of work in this area and our plans going forward, with a quick summary of previous work.

PHENIX

- RHIC at Brookhaven National Laboratory – “Relativistic Heavy Ion Collider” – is one of only two operating heavy-ion colliders (and the only one in the US), built for the purpose of study of nuclear matter at extremely high temperature and/or density, and spin physics
- PHENIX – “Pioneering High Energy Nuclear Interaction eXperiment”: a large, complex detector with a considerable physics reach
- Data taking finished in 2016, many data analyses are still ongoing and results are being published



DAP Challenges in PHENIX

- Since the end of the data taking, the PHENIX membership base has been diminishing, leading to less available effort and loss of know-how; sometimes, DAP is an unfunded mandate
- In many cases, the record of specific analyses (e.g. in the form of the “analysis notes”) was insufficient for the purposes of reliable reproduction
- A large part of the software environment and information systems are of legacy type (e.g. PHP applications built in-house)
- More recently, the OS environment started to change

Work Areas + Solutions

- Web presence
  - We consolidated previously fragmented PHENIX web resources into a simple but comprehensive website with emphasis on ease of long-term maintenance
- Publication data (data points)
  - Substantial progress has been made in 2023, in publishing the data on HEPDATA (vast majority of all papers, 213)
- Knowledge management and research document management
  - Continued to leverage the Zenodo platform at CERN, complemented with a list of curated keywords with direct links from the PHENIX website
  - ~700 PHENIX items published on Zenodo
- Software environment preservation
  - Version control, containerization, REANA

Zenodo + the Website

- The website https://www.phenix.bnl.gov/:
  - Built with a static website generator
  - Fast and secure
  - Extensive use of YAML for structured data
- Curated keywords: 311 at the time of writing, automatically translated into functional Zenodo links to the PHENIX materials committed to this platform

The website: the catalog of conference presentations, with auto-generated links to Zenodo

Conferences (132 items)

Keyword	Description
acat24	ACAT 2024
aum16	RHIC & AGS Annual Users Meeting (2016)
aum17	RHIC & AGS Annual Users Meeting (2017)
aum18	RHIC & AGS Annual Users Meeting (2018)
aum19	RHIC & AGS Annual Users Meeting (2019)
aum20	RHIC & AGS Annual Users Meeting (2020)
aum21	RHIC & AGS Annual Users Meeting (2021)
aum22	RHIC & AGS Annual Users Meeting (2022)
aum23	RHIC & AGS Annual Users Meeting (2023)
aum24	RHIC & AGS Annual Users Meeting (2024)
charm21	10th International Workshop on CHARM Physics
cipanp18	Conf. on the Intersections of Particle And Nuclear Physics (2018)
cipanp22	Conf. on the Intersections of Particle And Nuclear Physics (2022)
cpod17	Critical Point and Onset of Deconfinement (2017)
cpod18	Critical point and Onset of Deconfinement (2018)
cpod22	Critical Point and Onset of Deconfinement 2022
dis17	Deep Inelastic Scattering (2017)
dis18	Deep Inelastic Scattering (2018)
dis19	Deep Inelastic Scattering (2019)
dis21	Deep Inelastic Scattering (2021)
dis22	Deep Inelastic Scattering (2022)
dis23	Deep Inelastic Scattering (2023)
dis24	Deep Inelastic Scattering (2024)
dnp19	DNP (2019)

The website: physics keywords, as functional links

Physics (98 items)

Keyword	Description
3he+au	Helium3-on-gold collisions
anisotropy	Anisotropy
asymmetry	Asymmetry
au+au	Gold-on-gold collisions
azimuthal	Azimuthal
b-fraction	fraction of b-quarks
b-meson	B meson
backward-rapidity	The backward kinematic region
binary scaling	Binary scaling
bose-einstein	Bose-Einstein statistics
bottom	Particles containing the b-quark (bottom)
centrality	Centrality characteristic of the collision
cgc	Color Glass Condensate (type of matter)
charm	Particles containing the c-quark (charm)
charmonium	Meson containing a c-quark and its antiparticle
cnm effects	Cold Nuclear Matter effects
correlations	Various types of correlations
cronin effect	Cronin effect
cross section	Cross section (as it applies to scattering)
cu+au	Copper-on-gold collisions
cu+cu	Copper-on-copper collisions
cumulant	Cumulant
d+au	Deutron-on-gold collisions
d-meson	D meson
dca	Distance of Closest Approach
dielectron	A pair of electrons
dilepton	A pair of leptons
dimuon	A pair of muons produced in a collision
direct photon	Direct photons produced in nuclear collisions
drell-yan	Drell-Yan type of process

More complex (multi-keyword) queries can be constructed directly on the Zenodo website if necessary.

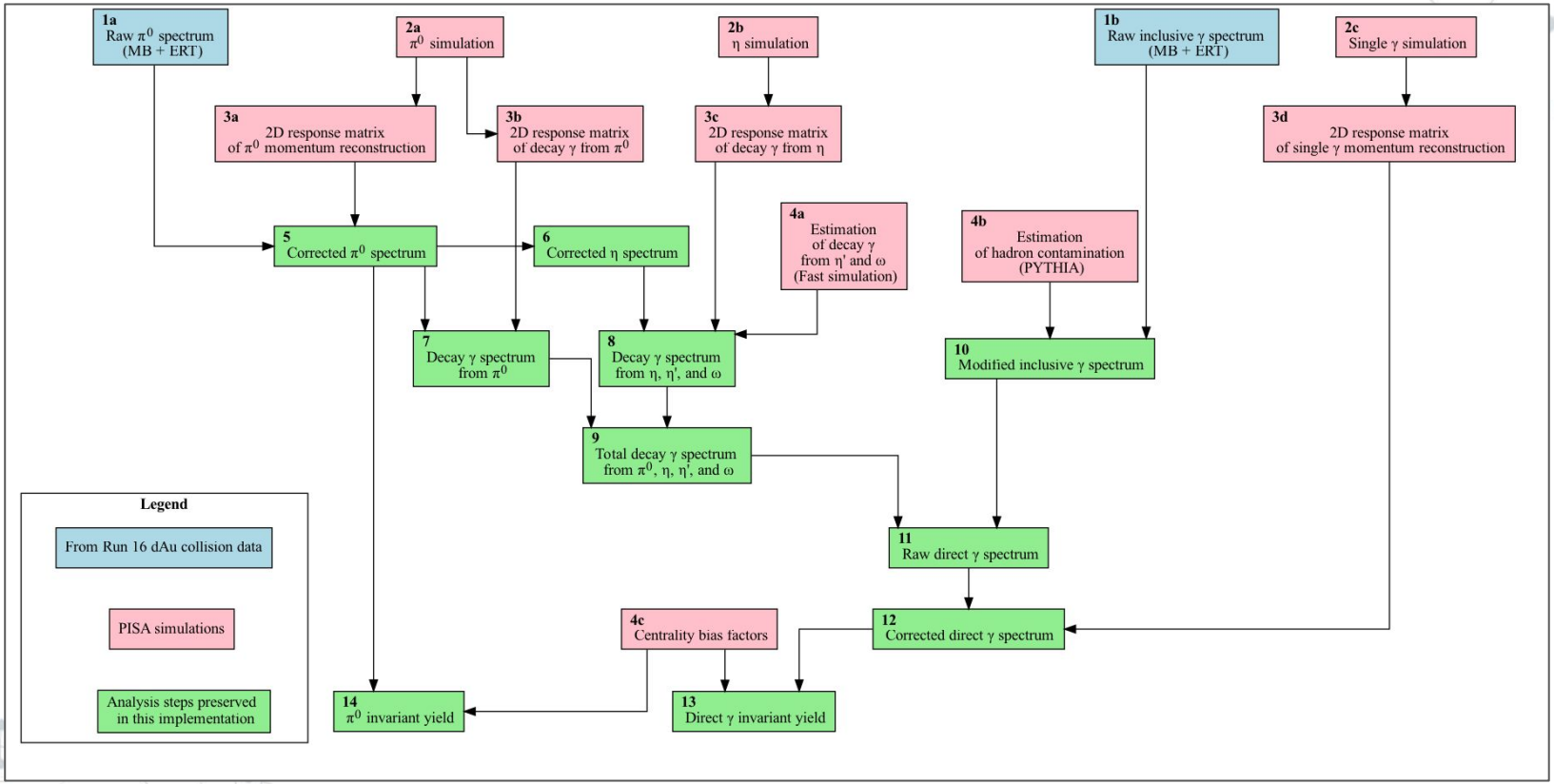
Experience in Analysis Preservation

- To state the obvious: analysis preservation is only possible with full commitment of people involved in that particular analysis, and provided a complete and accurate flowchart of the analysis has been created (the exact format is unimportant)
- We recently performed a complete preservation of an important analysis (nuclear modification factor in d+Au collisions). One more analysis (dimuon decay of J/psi) is in the pipeline.

An example: direct γ and π^0 analysis

- The main motivation for this study was an initially puzzling result regarding the nuclear modification factor in peripheral d+Au collisions. The background is explained in a PHENIX note: https://doi.org/10.5281/zenodo.8169171
- Ultimately it was determined that the key to understanding this result is to use the correct technique of estimating the centrality of collisions, making use of the Electromagnetic Calorimeter data. This has substantial scientific importance, and hence was chosen for preservation.
- This analysis featured a well documented flowchart and other crucial documentation which made this effort possible.

Direct γ and π^0 analysis: the flowchart



Direct γ and π^0 analysis: documentation (web)

https://www.phenix.bnl.gov/analysis/dAuPi0Photon.html

Direct γ in d+Au collisions

The measurement of γ and π^0 yields in d+Au interactions is important for studying the formation of quark-gluon plasma (QGP) in heavy ion collisions.

One way to measure QGP formation is by observing jet suppression using the nuclear modification factor R_{AB}, which compares the yield of a particle (in this case, the π^0) observed in AB is the same as that observed in p+p. If R_{AB} is less than one, then the yield in AB is suppressed, and if it is greater than one, then it is enhanced.

For a more detailed explanation that includes the motivation and physics background, please refer to this write-up: DOI 10.5281/zenodo.8169171.

- Direct γ in d+Au collisions
  - The Analysis Outline
    - General Analysis Workflow Diagram
    - Source Code
    - Input Data
    - Calibration Dependencies
    - Running the Analysis in Containers
      - Singularity
      - Docker
      - Building the Image
    - Running the Analysis with REANA
    - Confirming the Results
    - Analysis Steps
      - 1a. Raw π^0 spectrum (MB + ERT)
      - 2a. π^0 simulation
      - 3a. 2D response matrix of π^0 momentum reconstruction
      - 5. Corrected π^0 spectrum
      - 6. Corrected η spectrum
      - 7. Decay γ spectrum from π^0
      - 8. Decay γ spectrum from η, η′, and Ω
      - 9. Total decay γ spectrum from π^0, η, η′, and Ω
      - 10. Modified inclusive γ spectrum
      - 11. Raw direct γ spectrum
      - 12. Corrected direct γ spectrum
      - 13. Direct γ invariant yield
      - 14. π^0 invariant yield

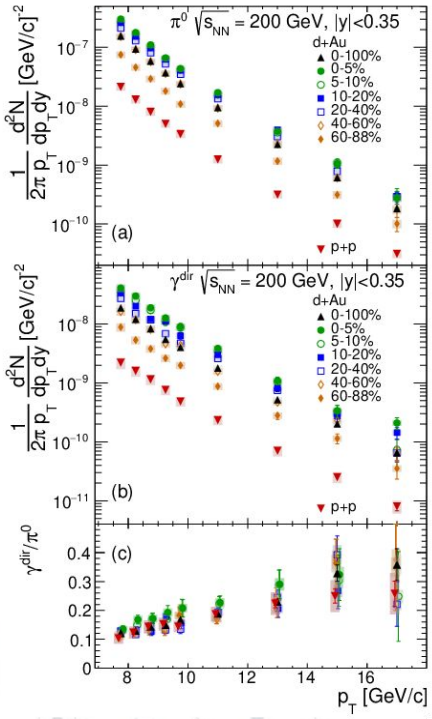
A detailed, step-by-step description of the analysis procedure, with references to the tagged blocks in the flowchart.

To DAG or not to DAG?

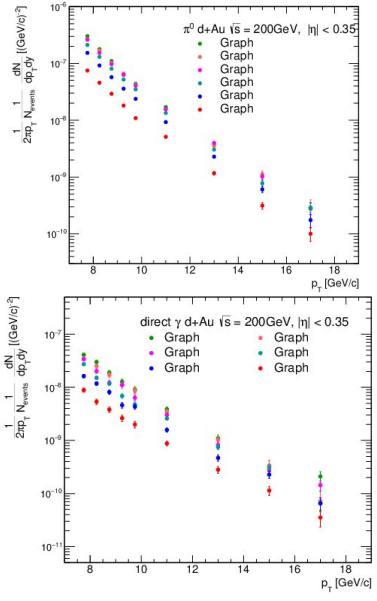
- Encoding the complete analysis in a Directed Acyclic Graph is an attractive concept, but in this case it creates more problems than it solves
  - Error recovery becomes more difficult
  - It is potentially harder to modify individual steps of the analysis and its overall logic, should this become necessary
  - Handling intermediate data may require extra planning and effort (cf. disk space etc)
  - DAG-capable workflow description languages are not very easy to master
- For these reasons, it was decided to keep the analysis as a set of self-contained and well documented blocks and execute them separately on REANA (and other platforms) according to the logic in the flowchart.

Validation of the procedure

Published



REANA run



A test run of the preserved analysis by a person (a computer science student who is not a PHENIX member), based solely on documentation, produced the correct published result.

First publication from RHIC on the topic of the preserved analysis:

https://arxiv.org/abs/2408.12072

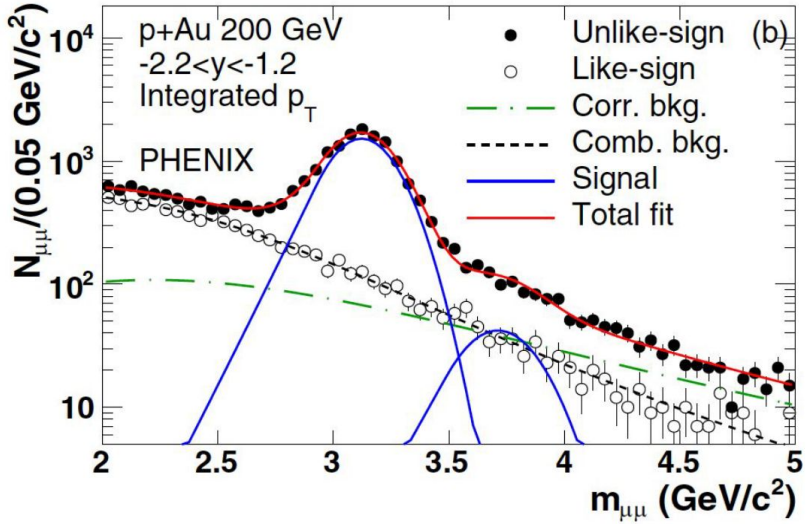
“Preservation of the Direct Photon and Neutral Meson Analysis in the PHENIX Experiment at RHIC”

The J/ψ dimuon decay study

- This study is our next candidate for analysis preservation
- PHENIX Publication: Phys. Rev. C 102, 014902
- Scientific significance
  - Heavy quarkonia suppression is one of the hallmark signs of quark-gluon plasma formation
  - Detecting signs of final state suppression due to quark-gluon plasma formation in small systems was done by comparing the nuclear modification measurements at forward and backward rapidity in three different systems: p+Al, p+Au, and ^{3}He+Au
  - The analysis shows that nuclear absorption and gluon anti-shadowing describe the PHENIX data very well at backward rapidity and the suppression is likely not due to quark gluon plasma formation

The J/ψ dimuon decay study: novelty

- The study featured an improved fitting method of the *background estimation*, necessary for extraction of the physics result



The J/ψ dimuon study: status of analysis preservation

- This is work in progress, and a substantial amount of detailed information has been already collected in the form of a write-up, first created in February 2024
- In addition, the general workflow of the dimuon analysis has been documented in the form of a PHENIX tutorial: https://doi.org/10.5281/zenodo.4011095
- We are moving towards a REANA implementation based on this information

The RHIC DAP Initiative

- The RHIC program is entering the final stage of data taking
- A long-term strategy and platform for data and analysis preservation become increasingly important
- In consultations with STAR Collaboration, a DAP initiative has been created by the Scientific Data and Computing Center (SDCC) personnel in the form of the Program Development (PD) proposal, led by E.Lancon
- See next slide for details

PD Request: Description of Deliverables (credit: E.Lancon)

- Comprehensive DAP Plan:
  - Ensures integrity, accessibility, and longevity of RHIC data and analysis post-RHIC operation.
- Dedicated Research Portal:
  - Offers a comprehensive overview of RHIC data and analyses.
  - Facilitates easy data access for future researchers.
- AI Chatbot:
  - Powered by large language models (LLMs) trained on domain specific data. Handles extensive information, provides code snippets, facilitates data access, and answers complex questions standard tools would not be able to.
  - Acts as an intelligent assistant for RHIC data analysis and technical information, simplifying data access and complex queries.
  - Valuable for re-analyzing data and answering domain-specific questions when human experts are unavailable.
- REANA (Reproducible research data analysis) Platform Implementation:
  - Preserves RHIC analysis software environment and workflow descriptions.
- Comprehensive Documentation:
  - Outlines data formats, analysis pipelines, and the scientific rationale behind specific analyses.
- Continuous Improvement:
  - DAP strategy will evolve through collaboration with the RHIC user community.

Wish list

- The RHIC experiments and their Analysis Preservation effort would benefit greatly from an equivalent of the Open Data portal (as it exists at CERN) deployed and available to the RHIC community
- Open Data is an efficient way to integrate heterogeneous material (data, documentation, software) in an accessible package of any level of complexity
- Open Data is based on Invenio, and Invenio has been supported at BNL for the past few years. Can the Open Data web application be ported to a different site outside of CERN?
- A key component of DAP in general and of platforms such as Open Data is the mass storage, which in case of Open Data is mostly EOS
- There is currently no EOS equivalent in the US in terms of scale and accessibility

Summary

- PHENIX has made substantial progress in its Data and Analysis Preservation effort overy the past five years
- In the past two years, most of the publication-related data has been committed to the HEPData portal and this work is ongoing
- Zenodo remains one of the principal PHENIX DAP components with 700 items committed, including indexed presentations from 132 conferences
- REANA-based analysis preservation effort resulted in one analysis preserved in substantial detail and validated against the actual publication plots, and another important analysis is in the pipeline
- There is a recent RHIC DAP initiative taking shape