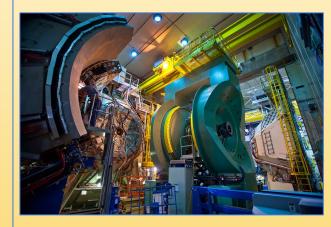
Current status of the Data and Analysis Preservation effort in the PHENIX experiment

Maxim Potekhin

Nuclear and Particle Physics Software Group







May 3, 2023

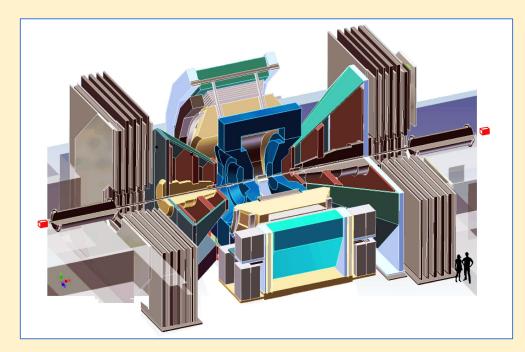
Overview

- **PHENIX** in a nutshell
- Data and Analysis Preservation (DAP) challenges in PHENIX and technical solutions identified, leveraged or developed to meet its goals
- Description of an effort currently in progress, to preserve the direct photon analysis based on the PHENIX data



PHENIX in a nutshell

- "Pioneering High Energy Nuclear Interaction eXperiment"
- A complex general purpose detector optimized for high rates
- Search for QGP and study of its properties, spin physics, electromagnetic probes, dimuons
- Complex analyses



• Please see the "PHENIX Collaboration Community" on Zenodo, the CERN-based digital repository: <u>https://zenodo.org/communities/phenixcollaboration/</u>

PHENIX today

- Data taking finished in 2016 with approx. 24PB of raw data accumulated
- Active analysis work underway (average ~10 articles a year in recent years)
 - About 260 published papers + *many* conference contributions, a total of >1200 entries on InspireHEP
 - Approx. 160 PhD theses
- Key active analyses
- Heavy flavor in central and forward rapidity
- Low p_T photons and thermal dileptons
- $\circ \qquad \text{High } p_{T} \text{ direct photon and} \\ \text{photon-jets} \\$

PIRE	stare v photo		
			(
	Literature Authors Jobs Seminars Conferences More		
Date of paper	202 medas 🖻 olici al	tator may	Most Recent V
	Elliptic flow measurement of J/ψ in PHENIX Run14 Au+Au at $\sqrt{s_{NN}}=200~{\rm GeV}$ Press calculated: - La blow (velocità U) for the uniformatic, (Ser 10, 2020) entre contract (Ser 10,	R reference search	0 o citations
Number of authors	Nuclear modification of hard scattering processes in small systems at PHENIX PREMIX collements - Netfork Remote at Use 13,782() = https://doi.org/10.1011		12
Single author 15 10 authors or lens 22	D) pat 🔄 dae 🔃 daen	🕅 reference search	⊕ 0 citations
Exclude RPP Diclude Revew of Particle Physics 202	J/ψ and ψ(25) Production in Small Systems with PHENIX PHENX Coldonator - Rima Carly Los Annual for the substantion, (bits 17, 2021) =/mir 2212.0000 (bits)-0-41 D of [= 100 - 00 - 00 - 00	S reference search	
Document Type	Exploring Hadron Spectra in Small Collision Systems at PHENIX	bå reference search	U Cossons
2 article 202 published (0) 233 conference paper 2	PHENEX Collaboration - V.A. Bedrahov (B. Petershung Pelytechnic Intel.) (Edgs 9, 3020) Published in: Mossow Unix/Phys.Bull. 77 (2020) 2, 176-177 - Contribution III: 2010 Lamonator Carefornize on Elementary Publick Physics, 176-177		
refere 2	අ වගා 🗄 මස 🖹 සභා Measurement of g-meson production in Cu + Au collisions at හැක = 200 GeV and U + U collisions at හැක = 193 GeV	R reference search	O citations //S
Author Hiddi Hamapali 232 Tako Sakauzh 232	невызанник и унивали росской на са и ла колнона в и у ₅₀ − 200 бек вис 0 + 0 больона в и у ₅₀ − 150 бек накак совоком са до дала и колнон чена са дала и опакона в и у ₅₀ − 200 бек вис 0 + 0 больона в и у ₅₀ − 150 бек пакане нероне са до дала и колнон чена са дала в опакона в опакона в и у ₅₀ − 200 бек вис 0 + 0 больона и у ₅₀ − 150 бек	🖾 reference search	D 1 citation
Thomas Kenneh Hemmick 239 Gabor David 239 Moley Chu 239	Improving constraints on gluon spin-momentum correlations in transversely polarized protons via midrapidity open-heavy-flavor electrons in $p^2 + p$ collisions at $\sqrt{s} = 200$	GeV	-
Edward P. Kistenev 239 Paul W. Stankus 230	D part E daw E damaens E dawn	🕅 reference search	3 citations
John Guy Lajole 229 Barbara V. Jacak 229 John S. Haggerry 229	The ϕ meson production from small to large systems of ion collisions at $\sqrt{S_{NN}} = 200$ and 133 GeV restrict calculations: usual Materials (i) Promoting Paragenties multi et al (ϕ 20, 202) Parallel (ϕ) (ϕ) (ϕ) (202)		97
Show 50 more	e ² DOI E data 🕄 data	🕅 reference search	C citations
Subject	Charm- and Bottom-Quark Production in Au+Au Collisions at $\sqrt{s_{_{NN}}}$ = 200 GeV		15
Experimen-Nucl 203 Experimen-HEP 09	PFUERCORRENTON - U.A. Activaya (Georgia Bane U., Marris) et al. (Mer 31, 2022) e Prinz 2020 1709 (push-rd)		



Motivations for DAP – data and analysis preservation

- In general, the goal of DAP is to have a reproducible analysis capability over a long period of time so as to retain the value of the data and effort
- However, it also brings benefits on the shorter time scale
 - Reliable reproducibility is a necessary component of any analysis, and DAP tools are well suited for that
 - It aids the capability to perform a modified or new analysis within the same framework
 - Onboarding new researchers is facilitated by knowledge management inherent in DAP, as well as good software management practices that it engenders
- The policy factor: funding agencies increasingly require both new and existing experiments to develop and implement plans for Open Data, which in practice implies not just preserving the data but also the tools and documentation necessary to access it.



Challenges

• Web infrastructure

 Without constant effort the web resources become obsolete, fragmented and the underlying technology (e.g. legacy versions of PHP, databases) can become difficult to maintain, especially with limited resources. Network security presents its own set of issues.

• Knowledge management

- As people move on to other projects, continuity of know-how becomes an issue. Software, detector and other documentation must remain accessible and useful in the long term.
- In the past, research document management was done using in-house solutions which share common problems with the legacy web infrastructure

• Software

OS versions, compilers, tools and components are not constant in the long term. Analysis
preservation implies that the exact software configuration and services are captured and
remain operational





Solutions

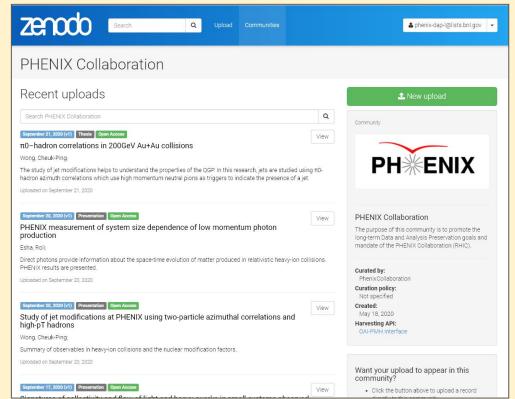
- Web infrastructure (aging, fragmented, hard to maintain)
 - Redesign of the PHENIX website with long-term maintenance in mind, utilizing a popular static website generation technology "Jekyll": phenix.bnl.gov
 - Consolidation, curation and reformatting of materials from previous web resources
 - Vastly improved security due to static nature of the site (important)
- Knowledge management (hard-to-discover, custom information resources)
 - Leveraging modern repositories and web portals to host research materials, replacing previous in-house solutions
 - A "Controlled Vocabulary" of keywords to make materials discoverable, e.g. conferences with PHENIX participation, physics topics/terms, detector components, software technologies – complements sophisticated Zenodo search capabilities
- Software (evolving platforms, compilers, components)
 - Containers (Docker+Singularity): capture of the software environment, utilization in REANA



Zenodo@CERN – the PHENIX page

https://zenodo.org/communities/phenixcollaboration

- A modern digital repository
- >600 PHENIX items, uploads ongoing, including presentations from >100 conferences and 160 PhD theses
- Branded, curated, findable, with DOIs
- Well-suited for long-term preservation
 - ...and also current activity: theses, analysis tutorials, presentations etc
 - Can be used to store data in almost any format and aggregation
- Excellent search capability
 - Keywords (including "controlled")
 - Elastic search





PHENIX keywords on the website ("vocabulary")

Experiment ⊛ ▼ Results ⊘ ▼ Detectors Physics (97 items)	Experi	ment ⊛ ▼ Results 🖉 ▼ Detectors 🗘 ▼	Offline Software 🕞 🔻	Each keyword is a link to a functioning Zenodo query
Keyword 3he+au	Description Helium3-on-gold collisions	Conferences (97 items)		
anisotropy	Anisotropy	Keyword	Description	
asymmetry	Asymmetry	aum16	RHIC & AGS Annual	Users Meeting (2016)
au+au	Gold-on-gold collisions	aum17	RHIC & AGS Annual	Users Meeting (2017)
azimuthal	Azimuthal	aum18	RHIC & AGS Annual	Users Meeting (2018)
b-meson	B meson	aum19	RHIC & AGS Annual	Users Meeting (2019)
backward-rapidity	The backward kinematic region	aum20	RHIC & AGS Annual	Users Meeting (2020)
binary scaling	Binary scaling	aum21	RHIC & AGS Annual	Users Meeting (2021)
bose-einstein	Bose-Einstein statistics	aum22	RHIC & AGS Annual	Users Meeting (2021)
bose-enisten	Particles containing the b-guark	charm21 10		orkshop on CHARM Physics
	Centrality characteristic of the c	cipanp18	Conf. on the Intersec	tions of Particle And Nuclear Physics (2018)
centrality	Color Glass Condensate (type o	cipanp22	Conf. on the Intersec	tions of Particle And Nuclear Physics (2022)
cgc		cpod17		set of Deconfinement (2017)
charm	Particles containing the c-quark			set of deconfinement (2018)
charmonium	Meson containing a c-quark an	dis17	Deep Inelastic Scatte	ring (2017)
cnm effects	Cold Nuclear Matter effects	dis18	Deep Inelastic Scatte	ring (2018)
correlations	Various types of correlations	dis19	Deep Inelastic Scatte	
cronin effect	Cronin effect	dis21	Deep Inelastic Scatte	
cross section	Cross section (as it applies to so	dis22	Deep Inelastic Scatte	ring (2022)
cu+au	Copper-on-gold collisions	dnp19	DNP (2019)	
cu+cu	Copper-on-copper collisions	dnp20	DNP (2020)	
cumulant	Cumulant	epshep17	EPS HEP 2017	
d+au	Deutron-on-gold collisions	eunpc22		lysics Conference 2022
d-meson	D meson	fwph21		d physics and QCD (2021)
dca	Distance of Closest Approach	ghp17		APS Topical Group on Hadronic Physics (2017)
dielectron	A pair of electrons	ghp19		APS Topical Group on Hadronic Physics (2019)
dilepton	A pair of leptons	hfwinc22		ction in Nuclear Collisions (2022)
dimuon	A pair of muons produced in a	hp16	Hard Probes 2016	
direct photon	Direct photons produced in nuc	hp18	Hard Probes 2018	
drell-yan	Drell-Yan type of process	hp20	Hard Probes 2020	
electron	Electron	hptlhc19		e RHIC/LHC era (2019)
		hq18	Hot Quarks 2018	

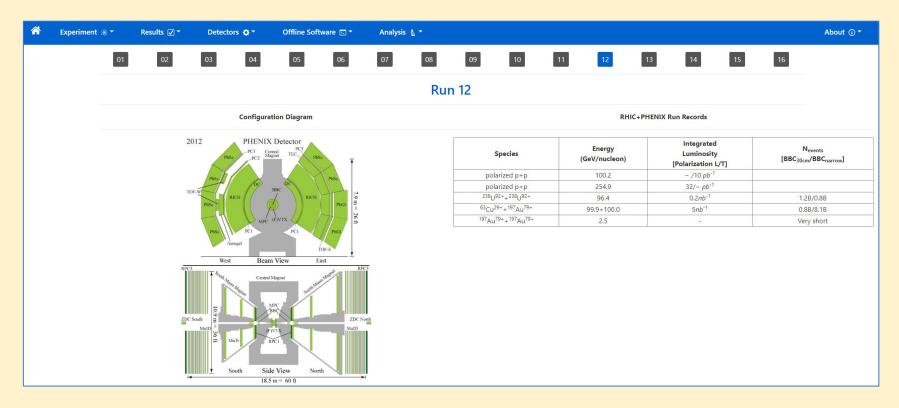


Example of a subsystem page on the PHENIX Website

							Electromagneti		
Write	-ups								
1000	OI 10.5281/zenodo.3		Calorimeter (EMCal) – Detec		C C C C C C C C C C C C C C C C C C C		Calorimeter		
0.00	OI 10.5281/zenodo.3 OI 10.5281/zenodo.3			e PHENIX Electromagnetic Ca	lorimeter (G.David)				
			nggers (A.bazilevsky)						
These	S								
-	OI 10.5281/zenodo.3				lisions at √s _{NN} =62. <mark>4</mark> GeV and √s	_{NN} =39 GeV beam energi	es (Vladimir Khachatryan)		
• 0	OI 10.5281/zenodo.3	85870 Inclusive jet production	in proton-proton and copper	-gold collisions at √sNN = 20	0 GeV (Arbin Timilsina)				
Public	ations								
• PI	HENIX Calorimeter	(NIM A 499, 2003, doi.ora/10.	1016/S0168-9002(02)01954-X	()					
		5		 PHENIX Calorimeter (NIM A 499, 2003, doi.org/10.1016/S0168-9002(02)01954-X) 					
High Energy Beam Test of the PHENIX Lead-Scintillator EM Calorimeter High Energy Beam Test of the PHENIX Lead-Scintillator EM Calorimeter									
Prese	Igh Energy Beam 1 ntations 01 10.5281/zenodo.4		or EM Calorimeter High Ene agnetic Calorimeter (Gabor D		(Lead-Scintillator EM Calorimet	er			
Prese	ntations OI 10.5281/zenodo.4		agnetic Calorimeter (Gabor D	avid)	(Lead-Scintillator EM Calorimet	er			
Preser • •	ntations OI 10.5281/zenodo.4	07113 PHENIX Focus: Electrom	agnetic Calorimeter (Gabor D	avid)	(Lead-Scintillator EM Calorimet	er			
Preser • •	ntations 01 10.5281/zenodo.4 ples and Acces	O7113 PHENIX Focus: Electrom	agnetic Calorimeter (Gabor D k Node (used for charg	avid) Jed particle analyses)	(Lead-Scintillator EM Calorimet	er			
Preser • © Variak	ntations 01 10.5281/zenodo.4 ples and Acces Name	oria PHENIX Focus: Electrom ors under PHCentralTrac Description	agnetic Calorimeter (Gabor D k Node (used for charg	avid) ged particle analyses) MC (cm)	(Lead-Scintillator EM Calorimet	er			
Preser • C Variak Type float	ntations of 10.5281/zenodo.4 ples and Acces Name get_pemcx	PHENIX Focus: Electrom ors under PHCentralTrac Description x-component of the projectic	agnetic Calorimeter (Gabor D k Node (used for charg in of the cgl track onto the EM in of the cgl track onto the EM	avid) ged particle analyses) MC (cm) MC (cm)	(Lead-Scintillator EM Calorimet	er			
Variak float	ntations or 10.5281/zenodo.4 ples and Acces Name get_pemcx get_pemcy get_pemcz	PHENIX Focus: Electrom ors under PHCentralTrac Description x-component of the projectic y-component of the projectic	agnetic Calorimeter (Gabor D k Node (used for charg in of the cgl track onto the EM in of the cgl track onto the EM in of the cgl track onto the EM	avid) ged particle analyses) MC (cm) MC (cm) MC (cm)	(Lead-Scintillator EM Calorimet	er			
Variak Variak Type float float	Diagonal and Access Name get_pemcx get_pemcz get_plemcz get_plemc	PHENIX Focus: Electrom ors under PHCentralTrac Description x-component of the projectic y-component of the projectic z-component of the projectic path Length following particle	agnetic Calorimeter (Gabor D k Node (used for charg on of the cgl track onto the EM on of the cgl track onto the EM on of the cgl track onto the EM on of the cgl track onto the EM	avid) ged particle analyses) MC (cm) MC (cm) MC (cm) MC (cm)			that the former is more useful for calculating		
Preset • D Variak Type float float float	Diagonal and Access Name get_pemcx get_pemcz get_pemcz get_pemc	PHENIX Focus: Electrom ors under PHCentralTrac Description x-component of the projectic y-component of the projectic z-component of the projectic path Length following particle time of the EMC hit. This time	agnetic Calorimeter (Gabor D k Node (used for charg an of the cgl track onto the EM an of the cgl track onto the EM an of the cgl track onto the EM in of the cgl track onto the EM trajectory from vertex to EM thas been back-corrected inF	Jed particle analyses) MC (cm) MC (cm) MC (cm) MC (cm) MC PHCentralTracks to be the phy			that the former is more useful for calculating		
Preset • • • Variak Type float float float	I 10.5281/zenodo.4	OPTI3 PHENIX Focus: Electrom OPTIS UNDER PHCentralTrac Description x-component of the projectic y-component of the projectic z-component of the projectic path Length following particle time of the EMC hit. This time properties of a charged track.	agnetic Calorimeter (Gabor D k Node (used for charg on of the cgl track onto the EN on of the cgl track onto the EN on of the cgl track onto the EN of the cgl track onto the EN trajectory from vertex to EN has been back-corrected inF en the track model projectior	avid) Jed particle analyses) MC (cm) MC (cm) MC (cm) HC PHCentralTracks to be the phy n and the hit in emc			that the former is more useful for calculating		
Preset • • • Variak Type float float float float	I 10.5281/zenodo.4	OPTI3 PHENIX Focus: Electrom Ors under PHCentralTrac Description x-component of the projectic y-component of the projectic z-component of the projectic path Length following particl time of the EMC hit. This time properties of a charged track. difference in phi (rads) between	agnetic Calorimeter (Gabor D k Node (used for charg on of the cgl track onto the EN on of the cgl track onto the EN on of the cgl track onto the EN the trajectory from vertex to EN the track model projection the track model projection a	avid) Jed particle analyses) MC (cm) MC (cm) MC (cm) HCentralTracks to be the phy and the hit in emc and the hit in emc			that the former is more useful for calculating		



Run summary pages (rebuilt on the new site)





PHENIX HEPData presence

HEPData		🛛 About 🕞 Submissic	on Helţ				
		Q phenix Reset search Advanced					
1 Max results - ↓ IF Sort b	oy + I ^z Reverse order	Showing 10 of 79 results					
Date	Reset	* * 1 2 * *					
autiliate	dda -	Nonprompt direct-photon production in Au+Au collisions at $\sqrt{s_{_{FW}}}=200$ GeV					
0	O 2023	The PHENIX collaboration Acharya, U.A.; Adare, A.; Aidala, C.; et al.					
Collaboration	Reset	36 authors from 71 institutions, 26 pages, 30 figures, 4 tabels, 2014 data. Physical Review C. Plain text data tables for the points plotted in figures for this and previous phenix publications are (or will be) publicly availabl http://www.phenix.bnl.gov/papers.html, 2022	le at				
* PHENIX	79	Inspire Record 2061074 % DOI 10.17182/hepdata.129292					
Subject_areas		The measurement of the direct photon spectrum from Au+Au collisions at $\sqrt{s_{35}} = 200$ GeV is presented by the <u>photons</u> collaboration using the external-photon conversion technique for DN=-39% central collisions in a transverse-momentum (pr) range of 0.8–10 GeV photons, above prompt photon production from hard-scattering processes, is observed for pr < 6 GeV(c. Nonprompt direct photons are measured by subtracting the prompt component, which is estimated as N _{mit} scaled direct photons from p + p collisions at 200 GeV					
nucl-ex	65	9 data tables match query					
hep-ex	39	Figure 12 Direct photon R_p every 20% contrality					
		Figure 13 Direct photon R ₂ every 10% centrality.					
CM Energies (GeV)		Figure 14 Invariant yield of direct photons, every 10% centrality					
$0.0 \leq \sqrt{s} \leq 1.0$		More					
1.0 ≤ √s < 2.0							
2.0 ≤ √s < 5.0 5.0 ≤ √s < 10.0							
5.0 ≤ √s < 10.0		Dilepton mass spectra in p+p collisions at sqrt(s)= 200 GeV and the contribution from open charm					
10.0 2 13 1 200.0		The PHENIX collaboration Adare, A.; Afanasiev, S.; Aidala, C.; et al.					
	Next 5 Show All	Phys.Lett.B 670 (2009) 313-320, 2009.					
Authors		Binspire Record 778611 € DOI 10.17182/hepdata.73669					
Q Search authors		The <u>come</u> operiment has measured the electron-positron pair mass spectrum from 0 to GeV/c ² in pr p collisions at sqr(b):200 CF. The contributions from light mecon-decays to e ⁺ e ⁺ -pairs have been determined based on measurements of hadron pairs. They account for not bear set of the electron positron decays of based based based based on measurements of hadron pairs. They account for the entry level electron decays of based b					
Fokin, S.L.	79	III 2 data tables match query					
Morrison, D.P.	79	Table 1 Differential charm cross section at mid rapidity An additional +39.5 microbarm error, due to the validity of the model used to extrapolate the data, is not included The contribution from beauty estimated to be 3.7 microbarm, has been su	ubtrac				
van Hecke, H.W.	79	branching ratio used was 9.5 + 1.0%.					
Akiba, Y.	78	Table 2 Total charm cross section An additional systematic error of +- 200 microbarm, due to the validity of the model used to extrapolate the data, is not included. To obtain the total charm cross section, the differential charm cross section and the total charm cross section	is been				
Bazilevsky, A.	78	whole rapidity range, using a HVQMNR rapidity distribution with aCTEQSM PDF.					
Brooks, M.L.	78						

HEPData submissions mandated for all new publications in PHENIX

We are also revisiting older publications, using GitHub to develop materials and coordinate teamwork (with PRs etc)

Within the last 3 years, the number of published PHENIX HEPData entries increased from 23 to 80 – special thanks to the team at UTK

The PHENIX OpenData entry

opendata _{CERN}	Search	٩	Help	About 👻
	PRENAX * Include on-demand datasets Filter by type Derived Include Include	Sort by: Display Best match asc. Display Found 1 result.]	
ALICE ALAS		© CERN. 2014–2021 · Terms of Use · Privacy Policy · Help ·		CERN

- It's a start...
- OpenData: a point of synthesis for software, data and documentation, capable of handling complex analysis cases and making substantial amounts of data accessible
- Contents of this particular package:
 - Derived data (Ntuples)
 - ROOT macros
 - Detailed instructions (PDF)
- Subject area:
 - Analyses based on the EMcal data

Capturing the Software Environment



- We use images to capture a few PHENIX SW environments, as a solution for the changing OS landscape, software components and dependencies
- Images created for PHENIX range from simple ones, capturing legacy ROOT and compilers, to the complete software stack as it is installed on the facility (large!)
- NB. interoperability between Docker and Singularity, i.e. containers can run in batch
- In PHENIX, we are using GitHub to manage Dockerfiles, Docker Hub for image delivery and also a private Docker registry at BNL *to provision software to REANA*
- Full images are stored in a private registry and accessible at the facility



REANA

- The core goal of REANA is reproducible analysis
- The workflow description syntax in REANA is a clear improvement, compared to a free form assembly of shell and other scripts, as it establishes a structured approach to analysis description (YAML)
- The learning curve in REANA is particularly easy for linear workflows; general DAGs are (a lot more) complex
- Individual software images can be set for steps in workflows more flexibility
- The PHENIX team is currently focusing on one specific REANA analysis direct photon production in *d*+*Au* collisions

reana Home Examples Get Started Documentation News Roadmap Contact Blog reana Reproducible research data analysis platform Flexible Scalable Reusable Free Support for remote Free Software, MIT Run many Containerise once. compute clouds. licence. Made with computational reuse elsewhere. workflow engines. e at CFRN. Cloud-native.







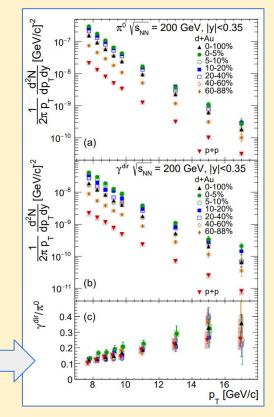


The study – use case

- PHENIX is a relativistic heavy-ion experiment and its main focus is determination of the properties of the Quark-Gluon Plasma (QGP).
- One of the most important signatures of formation of QGP was the jet suppression in heavy ion collisions, quantified by the so-called nuclear modification factor: R_{AB} (p_T) = Y_{AB}(p_T)/(N_{coll}×Y_{pp}(p_T)), where
 - \circ $\,$ A, B denote the two colliding nuclei
 - \circ Y_{AB} is the inv. yield measured in A+B collisions, and Y_{pp} is the yield measure in pp collisions
 - N_{coll} is the number of binary nucleon-nucleon collisions; its estimate is typically model-dependent (e.g. using the Glauber model)
- The main motivation for this study was to determine whether there can be a bias in centrality estimates in "small systems" such as *d*+*Au* collisions, resulting in hard-to-explain behavior of the nuclear modification factors. For background and details please see <u>https://arxiv.org/abs/2303.12899</u>

The measured yields of γ^{direct} and π^{0} , and their ratio: $\gamma^{\text{direct}}/\pi^{0}$

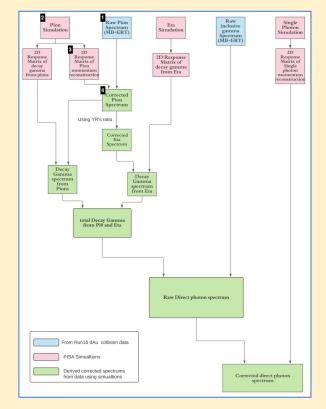
- d+Au collisions
- https://arxiv.org/abs/2303.12899
- On the bottom chart: note the similar ratio for most collisions except most central ones
- An important result that warrants analysis preservation
- The Electromagnetic Calorimeter was the principal detector used in this analysis





Workflow diagrams - a low tech but effective tool

- On the right the actual flowchart used in adopting the direct photon/EMCal analysis in PHENIX, to REANA
- Complementary to a good verbal description, published at, and cross referenced with the textual description of the analysis on the PHENIX website
- Diagrams like this one are being considered as a potential requirement for future analysis notes in PHENIX, as an effective and relatively low-cost policy
- In combination with REANA, enhances knowledge sharing and transfer within and between working groups





The documentation page for this analysis (excerpt)

*	Experiment 🕸 🔻	Results 🖌 🕶	Detectors 🔅 🔻	Offline Software 🕞 🔻	Analysis 🛓 🏲
		0EmbedFiles.C bedFiles t p()			
	In the RE	ANA script, this is used	as follows: cat pi@run.sc	ript root -b. Note that a PHE	NIX-specific ROOT library libTHmul.so is loaded in the beginning, as this is necessary for proper operation of the macro.
	Please re	fer to the relevant folde	r in the PHENIX GitHub re	epository for access to the actual	material.
	This is th	e driver script P10EmbedF	iles.csh. Note that symb	polic links are created to feed suc	ccessive files from a holding folder, to the ROOT macro.
	sourc: foread l: c: c: ri ri ri end	n/tcsh e ./setup_env.csh ch i (`seq 0 1 \$1`) n -s gpfs/mrt/gpfs07t s -l pi@_dAu%B.root s -l pi@_dAu%B.root t pi@run.script roo v EmbedPi0dAu.root Emb m pi@_dAu%B.root cf embedPi0dAu.tar Emb	t -b edPi0dAu_\$i.root	/phnxreco/emcal/Pi0/test/simPi	10_\$1.root p10_dAuM8.root
	Processin	ng of input files takes pla	ace sequentially and in th	is case takes a significant amoun	t of time compared to other steps, i.e. a feew hours.
					ng easier. Upon retrieval the data need to be merged using the utility had@henix which is done in Block 3 (see below). Upon completion of this he next step is launched. An example of the cownload command, assuming the workflow was named "embed":
	reana	-client download -w em	bed embedPi0dAu.tar		
	2D Res	ponse Matrix of Pi	on Momentum Reco	onstruction (Block 3)	
	The origi	nal macro generationRM	_Pi0.cc was cleaned up (including removal of interactive	graphics) and renamed generationRM_Pi0.C.
	Tar file co	ontaining multiple ROOT	files (see Block 2 descrip	otion above) is uploaded as inpu	t for this step. Abbreviated contents of driver script look as follows:
	source	n/tcsh e ./setup_env.csh henix EmbedPi0dAu.root -l -b -q 'generationRM			
	The macr	ro generates the file Emb	edPi0dAu.root by mergin	g inputs via haddPhenix and prod	duces Pion_RM. root. The complete description is in generationRM_Pi0.yam1, which resides with all subsidiary scripts in the folder generationRM.
	The work	flow description is as fo	llows:		



Current status of this study

- The results are public at https://arxiv.org/abs/2303.12899
- The initial version of this analysis was ported to REANA in 2021, it included a set of the most important analysis blocks shown above, but not quite complete
- It is clearly beneficial to expand the part of the analysis converted for reuse and preservation this is the focus of the current work
- The analysis itself has progressed and improved in the past two year, so this must be reflected in the REANA materials
- This round of updates is expected to be completed in Summer of 2023

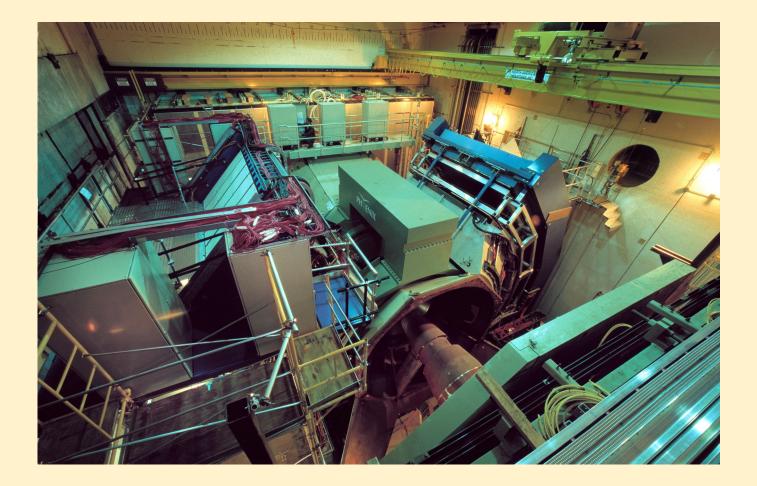


Challenges, Lessons and Plans

- In a legacy experiment
 - Too many components and dependencies may have accumulated over time, to have an unencumbered build on top of a "clean" OS
 - Services (like databases) and conditions data can also be tightly coupled to a particular facility
 - All these factors impede making the software images public
 - Analysis code is rarely documented in sufficient detail, to enable straightforward preservation
- Leveraging existing, community-supported platform allows to make substantial progress in the area of analysis preservation even with limited resources, the support from the facility is instrumental
- "Plan and execute early" is the best advice that new experiments can take
- PHENIX plans to continue its vigorous work of HEPData materials preparation, research document management in Zenodo and analysis preservation on REANA

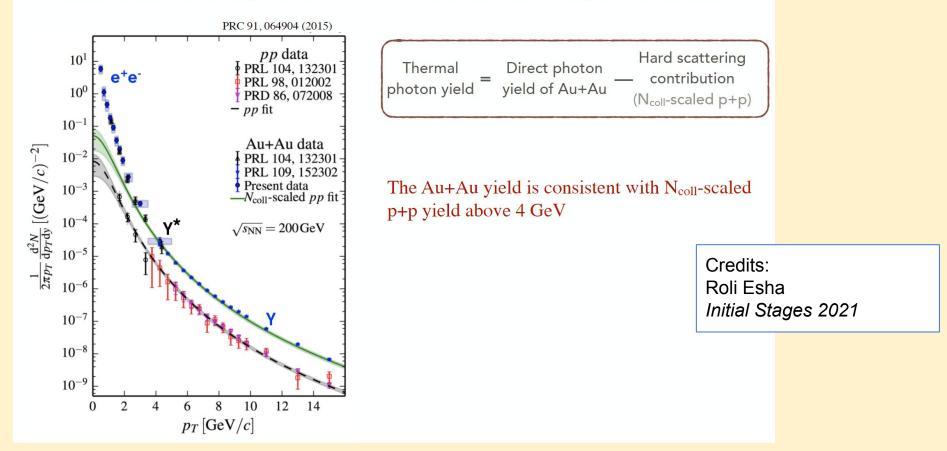
Backup slides







Direct photon yields for p+p collisions at 200 GeV are consistent with pQCD calculations





Knowledge Management

- Need to keep records of software provenance, dependencies, configuration, use etc the "know how"
- Software preservation ≠ Analysis preservation
- Keep track of "data artifacts" such as conditions-type data which may be produced for the purposes of a particular analysis and depend on details known mostly to the people involved in this analysis (misc. cuts, maps, lists, numerical constants in macros etc)
- There is a legacy solution which is a requirement to record such info in a dedicated section of the "Analysis Note" which must accompany every paper, but in reality its efficacy is variable and often insufficient
- Hard to provide continuity of know-how as people move on; knowledge dissipates



Benefits of DAP

DAP practices have the potential to enhance quality of the science output in the near term by helping ensure reproducibility and robustness of the results

DAP focus on knowledge management is conducive to efficient knowledge transfer within the collaboration and across projects (cf. onboarding graduate students)

Software management, packaging and containerization facilitates deployment

Modern digital repositories create efficient document management solutions on any time scale (cf. OpenData, Zenodo etc)





Lessons learned

- DAP: plan and start early
 - The effort will pay for itself by increasing overall productivity of the experiment
 - Will be hard or impossible to "catch up" later
- Avoid building in-house information systems, there are many tools available
 - State-of-the-art services such as GitHub, Zenodo, OpenData, HEPData, REANA, Rivet, Inspire (publication catalog) etc cover a vast majority of the experiments' needs
 - There must be no coupling to a particular MC/reco framework
- Containerization solves many of the challenges of capturing the software environment
 - Use it the right (portable) way, with services (DB) made accessible
- Create websites for the long haul (static site generation works well)
 - Avoid platforms that will require updates and maintenance in the long term e.g. Drupal
 - Any resource will become overgrown/obsolete in absence of editors
 - Avoid resource fragmentation



Tiers of Data Access (incl. HEPData and OpenData)

- Level 1: Data Products used in publications.
 - Such as data points and errors used in plots, in numeric format
 - o cf. the "HEPData" portal: https://www.hepdata.net/
- Level 2: Special Purpose Datasets (e.g. for education and outreach).
 - Select datasets + virtualized or otherwise portable analysis software + documentation
 - cf. the "OpenData" portal: <u>https://opendata.cern.ch/</u>
- Level 3: Reconstructed Open Data; may be released in future (e.g. based on policy)
 - Implies a more complex analysis environment than in Level 2
 - Requires adequate software and computing infrastructure to be properly used
- Level 4: Raw Data. Preserved, but not considered useful for release.



REANA – a few notes

REANA allows the user to record crucial components of analyses:

- The software environment (by reference to images and libraries, environment etc)
- The workflow(s), in one of the available YAML formats
- Data components to be staged in and staged out, any other auxiliary files that are need

The "workspace" paradigm (essentially a sandbox) enforces completeness of the description and provision of well-defined dependencies.

Also of note is a good CLI, a full Python API and Jupyter integration (e.g. one can open a notebook inside a workspace). The workspace is persistent (if needed).

A variety of computational back-ends is supported (even simultaneously – via hybrid pipelines)



DAP: Challenges and observations

If there is one lesson in this story it is the need to take a "holistic approach" – data without the software is often useless, as is software without build and verification systems and/or necessary additional data (alignment, calibration, magnetic field maps etc.) These are typically stored separately and involve distinct services that evolve on independent timescales and with lifetimes typically much shorter than the period for which the corresponding "data" needs to be preserved.

https://doi.org/10.5281/zenodo.2653526 "Software Preservation and Legacy issues at LEP" (J.Shiers)

No matter what preservation tools are developed that might enable reuse of software, analysis techniques, and data, if they are not conceived from the beginning as an integral part of the standard frameworks, retrofitting will be nearly impossible.

https://arxiv.org/abs/1810.01191 "HSF White Paper: Data and Software Preservation to Enable Reuse"

