

AID2E: AI-Assisted Detector Design at EIC

<https://aid2e.github.io/aid2e>



DE-FOA-0002875

BNL, T. Wenaus
CUA, T. Horn
Duke, A. Vossen
JLab, M. Diefenthaler
W&M, CF (PI)

Cristiano Fanelli

PI Exchange Meeting - Cristiano Fanelli - Nov 19-20, 2025



The AID2E Team





Team



Torre Wenaus, PhD

Expertise: Nuclear and particle physics software, Distributed Computing, Simulations



coPI



Meifeng Lin, PhD

Expertise: Physics, High performance computing, numerical simulations, Computational physics, exascale



Tianle Wang, PhD

Expertise: Postdoc - Physics, High Performance Computing, Workflow Management, Machine Learning



Wen Guan, PhD

Expertise: distributed computing, workflow management, ML workflows



Amit Bashyal, PhD

Expertise: AI/ML, optimization algorithms, scientific infrastructures



Fang-Ying Tsai, PhD

Expertise: AI/ML, HPC, and large-scale data processing



Gabor Galgoczi, PhD

Expertise: Physics, Data Science, MOO, Bayesian, Detectors



consulting



Kolja Kauder, PhD

Expertise: Physics, Simulation



consulting



Alex Jentsch, PhD

Expertise: Far Forward physics, EIC



consulting



Tanja Horn, PhD

Expertise: medium energy nuclear physics, EIC, 3D hadron Imaging, calorimetry



coPI



Baptiste Fraisse, PhD

Expertise: medium energy nuclear physics, simulations, calorimetry



Anselm Vossen, PhD

Expertise: Physics, PID and calorimetry for the EIC



coPI



Cynthia Nunez, PhD

Expertise: physics, simulations



Connor Pecar, MS

Expertise: physics, PID and detector simulations for the EIC



Markus Diefenthaler, PhD

Expertise: ePIC Software & Computing Coordinator, EIC Science, Simulations



coPI



Derek Anderson, PhD

Expertise: EIC Science, Reconstruction, Simulations, calorimetry, jets

Jefferson Lab



Makoto Asai, PhD

Expertise: Detector Simulations, Geant4



consulting



Cristiano Fanelli, PhD

Expertise: Data Science, Physics, MOO, Bayesian, Detectors, Artificial Intelligence, Computing



PI



Karthik Suresh, PhD

Expertise: Postdoc - Data Science, Physics, MOO, Bayesian, Detectors, Distributed Computing



Hemalata Nayak, MS

Expertise: Physics, Data Science, Optimization, Reconstruction





Electron Ion Collider



Electron Ion Collider

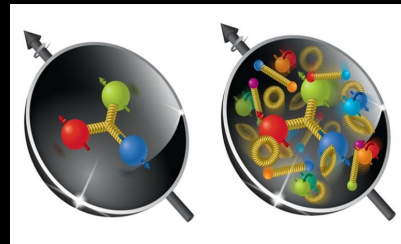


A US-led and international effort to build a precision machine to study the “glue” that binds us all. This will put the US at the frontier of nuclear physics research for the next 30 years. The science phase is set to begin in the 2030s.

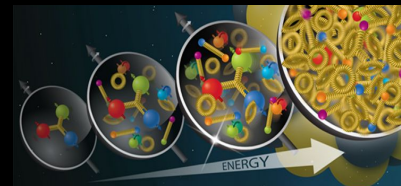
How does the **mass of the nucleon** arise?



How does the **spin of the nucleon** arise?



What are the **emergent properties of dense systems of gluons**?



polarized electron - polarized protons/ions

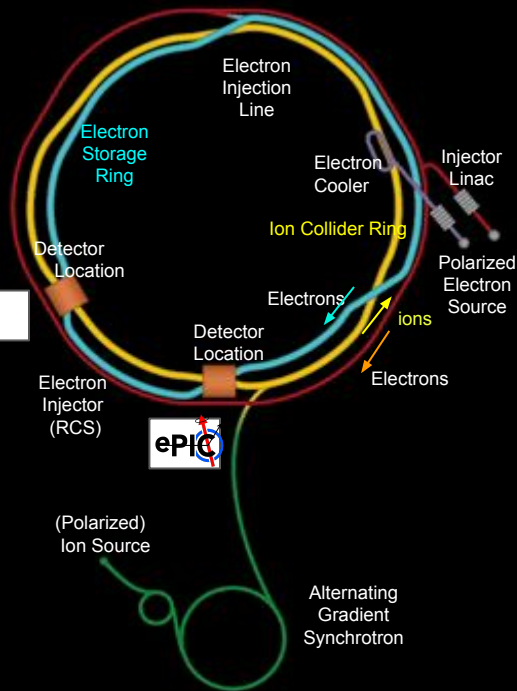


CoM energy $\sqrt{s_{e-p}} \sim (20-140) \text{ GeV}$

High luminosity up to $10^{34} \text{ cm}^{-2}\text{s}^{-1}$,
a factor $\sim 100-1000$ times HERA

Possibility of second detector in
addition to EIC Project Detector / ePIC.

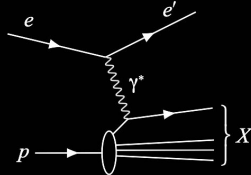
AI/ML will play a major role in
optimizing these complex operations
and accelerating the pace of discovery
by enhancing scientific precision



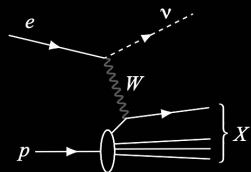
A Glimpse into EIC Physics



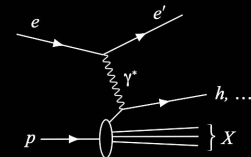
Neutral current
inclusive DIS



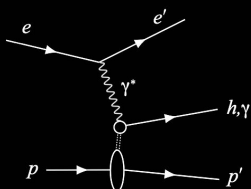
Charged current
inclusive DIS



Semi-inclusive DIS



Exclusive DIS



The broad scientific program drives the detector requirements
(cf. [EIC Yellow Report, NIMA 1026 \(2022\): 122447](#)):

- Mass and Tomography
- Spin and Flavor Structure of the Nucleons and Nuclei
- Internal Landscape of Nuclei
- QCD at Extreme Parton Densities - Saturation
- Important synergies with HL-LHC science program:
 - Precision QCD studies with proton & nuclear targets α_s , quarkonia, quark exotica, jet physics in e-p collisions, ...
 - Precision electroweak and BSM physics Weak mixing angle, LFV,
- Etc

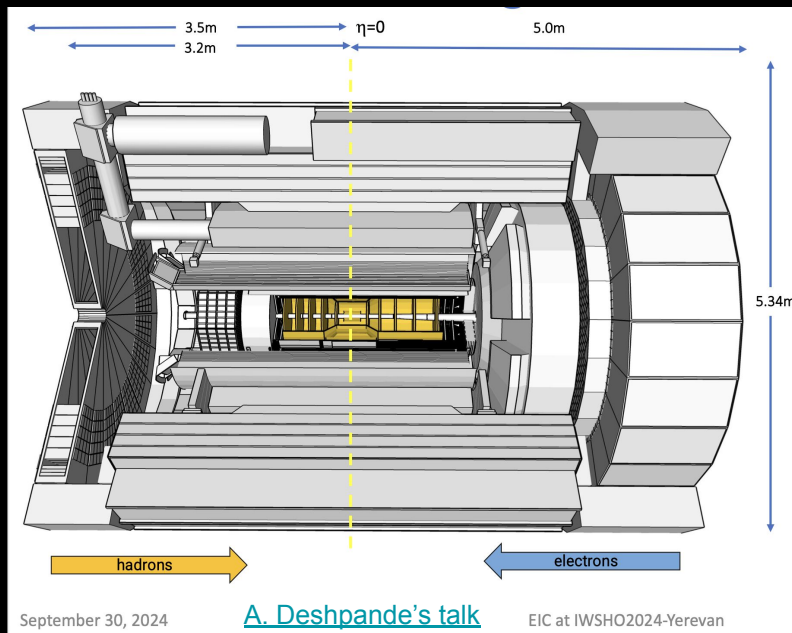
The optimization of the detector(s) must take the various physics objectives into account...

ePIC Detector



As of now, 180+ institutions, 25 countries and 1000+ collaborators

ePIC stands out as a highly **Integrated Detector** encompassing Central, Far-Forward, and Far-Backward regions, all crucial to access the EIC physics.



Tracking

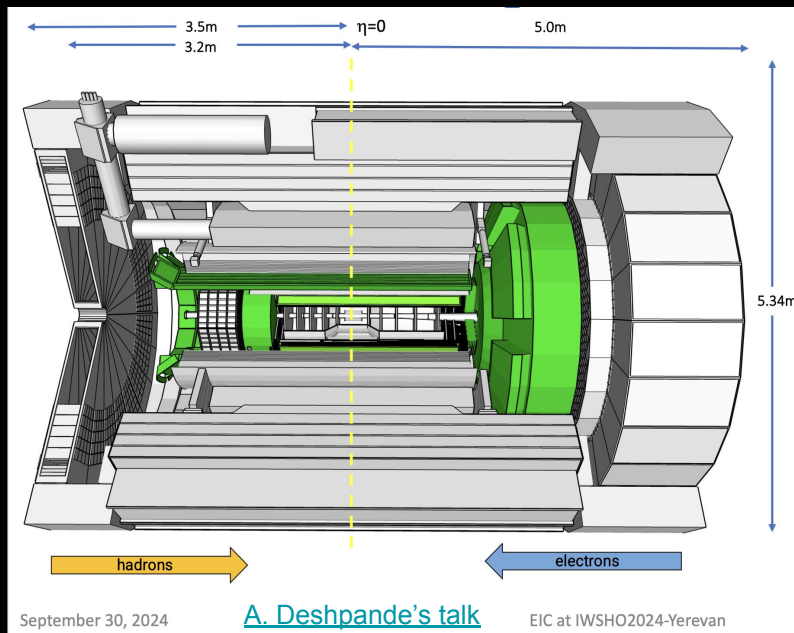
- New 1.7T solenoid
- Si MAPS Tracker
- MPGDs (μ RWELL/ μ Megas)

ePIC Detector



As of now, 180+ institutions, 25 countries and 1000+ collaborators

ePIC stands out as a highly **Integrated Detector** encompassing Central, Far-Forward, and Far-Backward regions, all crucial to access the EIC physics.



September 30, 2024

[A. Deshpande's talk](#)

EIC at IWSHO2024-Yerevan



Tracking

- New 1.7T solenoid
- Si MAPS Tracker
- MPGDs (μ RWELL/ μ Megas)

PID

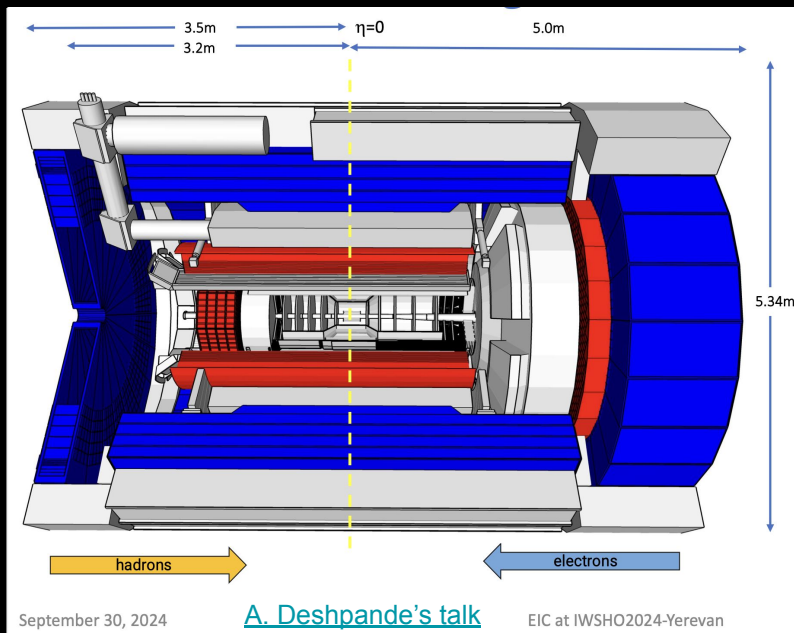
- hpDIRC
- pfRICH
- dRICH
- AC-LGAD (~ 30 ps TOF)

ePIC Detector



As of now, 180+ institutions, 25 countries and 1000+ collaborators

ePIC stands out as a highly **Integrated Detector** encompassing Central, Far-Forward, and Far-Backward regions, all crucial to access the EIC physics.



September 30, 2024

[A. Deshpande's talk](#)

EIC at IWSHO2024-Yerevan



Tracking

- New 1.7T solenoid
- Si MAPS Tracker
- MPGDs (μ RWELL/ μ Megas)

PID

- hpDIRC
- pfRICH
- dRICH
- AC-LGAD (~ 30 ps TOF)

Calorimetry

- Imaging Barrel EMCal
- PbWO₄ EMCal in backward direction
- Finely segmented EMCal +HCal in forward direction
- Outer HCal (sPHENIX re-use)
- Backwards HCal (tail-catcher)

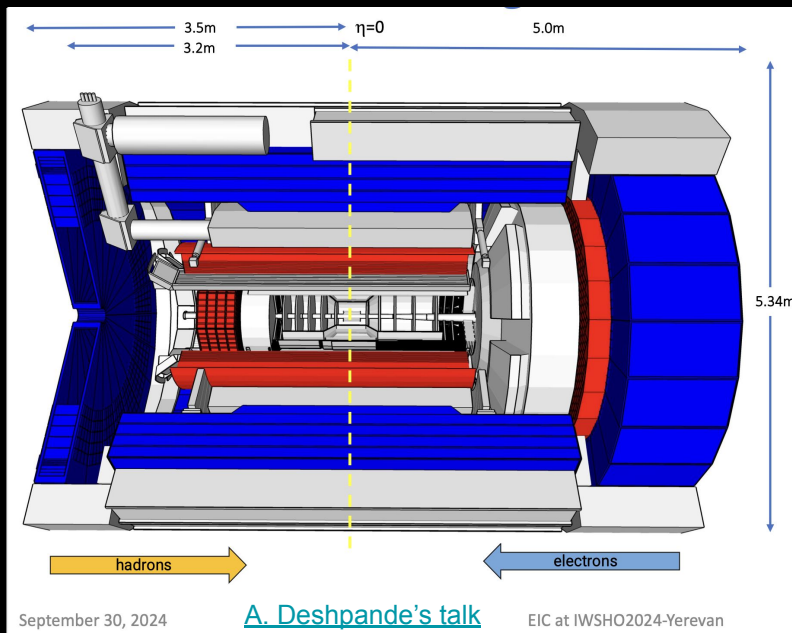
ePIC Detector



As of now, 180+ institutions, 25 countries and 1000+ collaborators

ePIC stands out as a highly **Integrated Detector** encompassing Central, Far-Forward, and Far-Backward regions, all crucial to access the EIC physics.

ePIC extends across -35m to $+35\text{m}$.



September 30, 2024

[A. Deshpande's talk](#)

EIC at IWSHO2024-Yerevan



Tracking

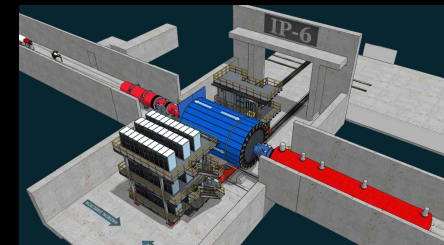
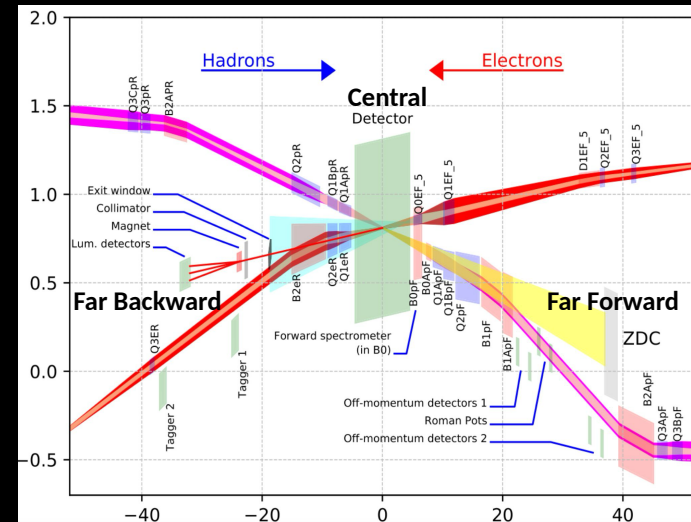
- New 1.7T solenoid
- Si MAPS Tracker
- MPGDs ($\mu\text{RWELL}/\mu\text{Megas}$)

PID

- hpDIRC
- pfRICH
- dRICH
- AC-LGAD ($\sim 30\text{ps TOF}$)

Calorimetry

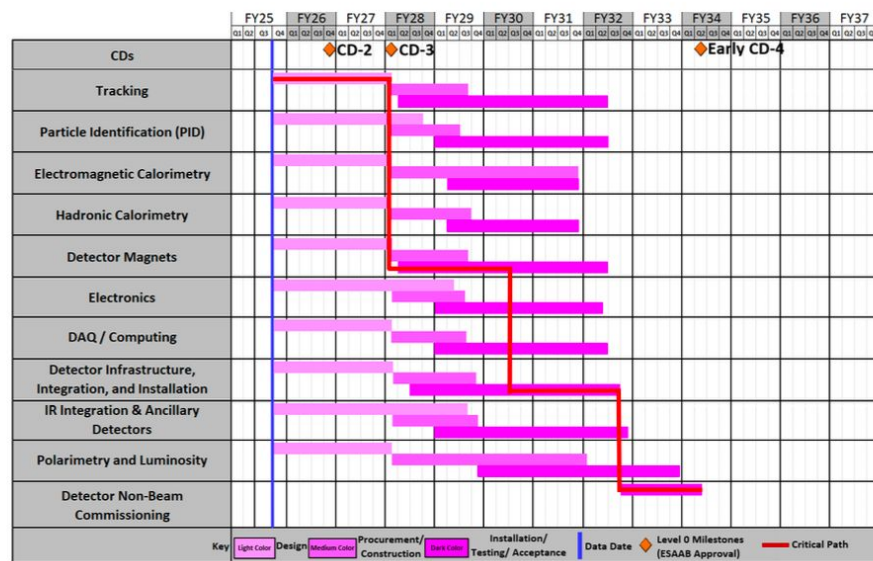
- Imaging Barrel EMCal
- PbWO₄ EMCal in backward direction
- Finely segmented EMCal +HCal in forward direction
- Outer HCal (SPHENIX re-use)
- Backwards HCal (tail-catcher)



The tightly integrated design of the ePIC detector leads to unique and complex optimization challenges...

<https://www.bnl.gov/eic/epic.php>

Project detector timeline



Collaboration working hard towards pre-TDR (for CD-2, ~60% design readiness)

ePIC Streaming Computing Model publication in preparation

— AID2E project is well aligned with the EIC schedule —



Detector Design



Detector Design: Considerations



- 1) **Traditional approach to Detector Design is limited:** Each sub-detector is optimized independently, often with a single objective within fixed global detector constraints.
 - This yields **suboptimal global performance**. Changing one sub-detector can affect the **entire detector**, so the full system should be optimized holistically.
- 2) **Accurate simulations (i.e., Geant4-based) are essential**
 - However they are slow and compute-intensive.
 - Fast simulations (ML-based/differentiable) only partly address the detector-design problem for several reasons:
 - They require large Geant4 training datasets
 - They can lack fidelity in critical phase-space regions
 - They may struggle to reproduce simultaneously and reliably multiple objectives
 - All sub-detectors must be fast-simulated, as overall speedup is ultimately limited by Amdahl's law
 - Even with fast simulations, high-dimensional optimization can dominate the runtime, making simulation speed less relevant.
- 3) **Large simulation campaigns* are necessary**, often using containerized workflows and distributed computing (e.g., NIM-A 1047 (2023): 167859).

- *Current simulation campaigns produce up to O(50) TB / month ([D. Kalinkin, Oct 2025](#))
- Towards a quantitative computing model ([M. Diefenthaler, Nov 2025](#))
- Simulating 5M charged particles for the tracker and PID system requires at least 15k CPU core-hours
- This cost increases substantially when additional particle types are included to design other sub-detectors

(Our Approach) **AID2E allows to optimize both (i) large-scale detector design and (ii) resources usage (“co-design”)**

- Enables holistic detector optimization using accurate full simulations (Geant4-based)
- Accelerate design process by reducing the number of design points explored during optimization

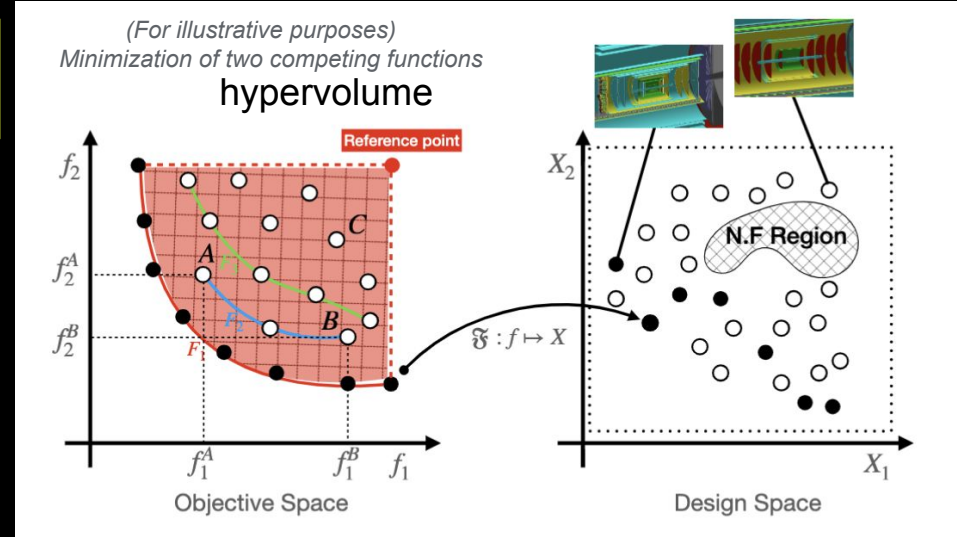
Multi-Objective Optimization



Multi-Objective Optimization

MOO needed to optimize a system of sub-detectors

- 3 Types of Objectives
 - **Intrinsic detector performance** (resolutions, efficiencies) for each sub-detector — Tracking, calorimetry, PID — noisy
 - **Physics-performance** — Multiple physics channels, equally important in the EIC physics program
 - **Costs** (e.g., material costs, provided a reliable parametrization)
- Objectives can be competing with each other
 - E.g. Better detector response come with higher costs; better resolutions may imply lower efficiencies; etc.



A generic MOO problem can be formulated as

$$\begin{aligned} \min \quad & f_m(\mathbf{x}) \quad m = 1, \dots, M, \\ \text{s.t.} \quad & g_j(\mathbf{x}) \leq 0, \quad j = 1, \dots, J, \\ & h_k(\mathbf{x}) = 0, \quad k = 1, \dots, K, \\ & x_i^L \leq x_i \leq x_i^U, \quad i = 1, \dots, N. \end{aligned}$$

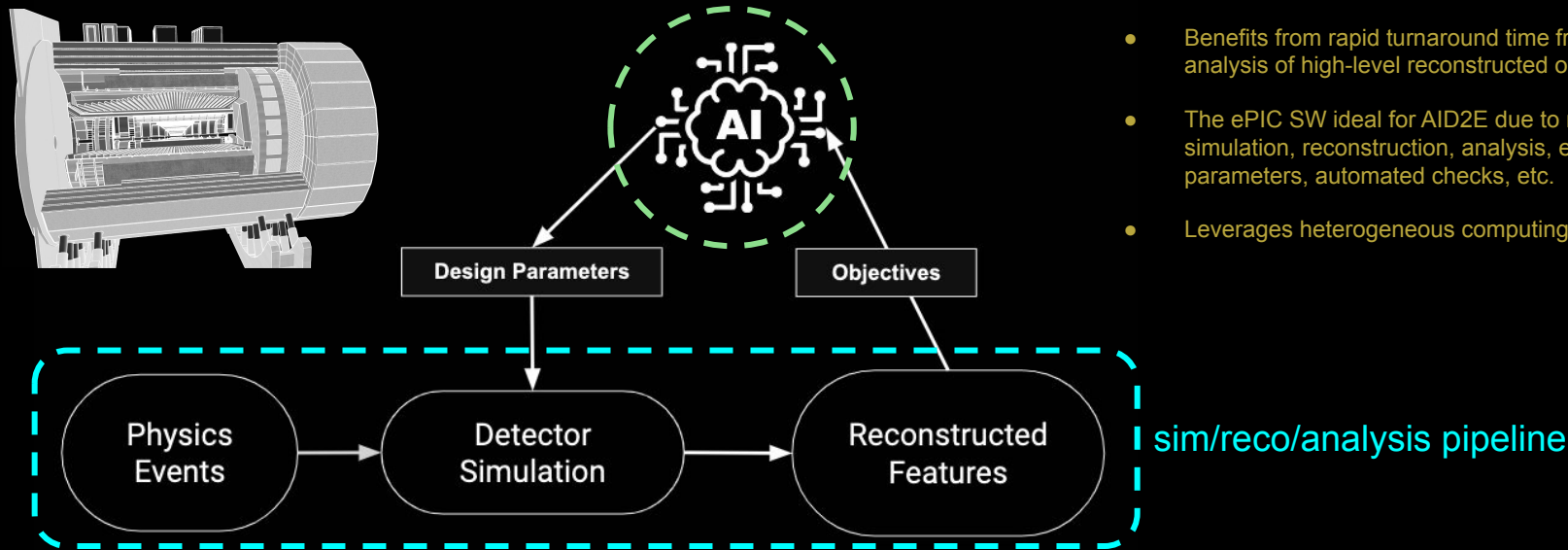
objectives

constraints

AI-Assisted Detector Design at EIC



- EIC is an end-to-end AI facility, AID2E contributes to applying AI to detector design and optimization.
- It is a prime example of the role AI can play in automation and scientific workflows.
- The AI-assisted design embraces all the main steps of the sim/reco/analysis pipeline.



- Benefits from rapid turnaround time from simulations to analysis of high-level reconstructed observables
- The ePIC SW ideal for AID2E due to modularity of simulation, reconstruction, analysis, easy access to design parameters, automated checks, etc.
- Leverages heterogeneous computing

AID2E provides a framework for the holistic optimization of sub-detector systems

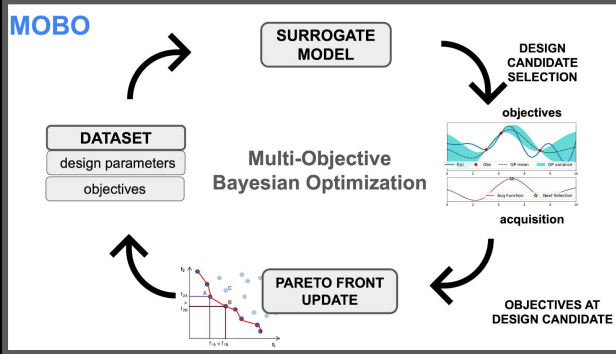
A complex problem with (i) **multiple design parameters**, driven by (ii) **multiple objectives** subject to (iii) **constraints**



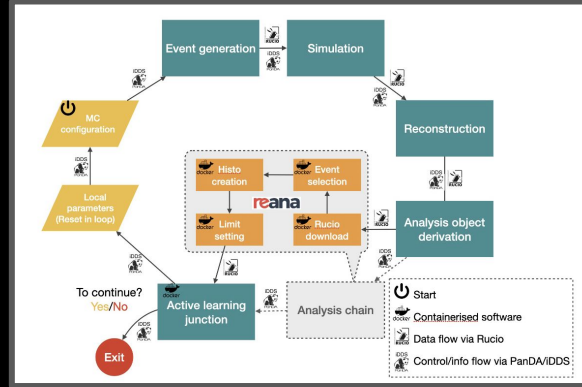
Multi-Objective Optimization Engines

Distribution and Workload Management

Data Science & Analytics

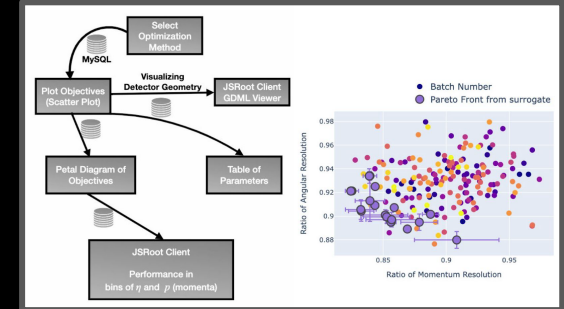


(i) Contributing to advance state of the art MOO complexity (e.g., Multi-Objective Bayesian Optimization) to accommodate a large number of objectives. AID2E supports also other methods (e.g., MOGA) and explores usage of physics-inspired approaches

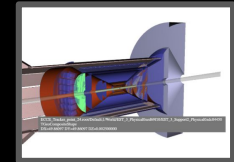


(ii) We leverage cutting-edge workload management systems capable of operating at massive data and handle complex workflows

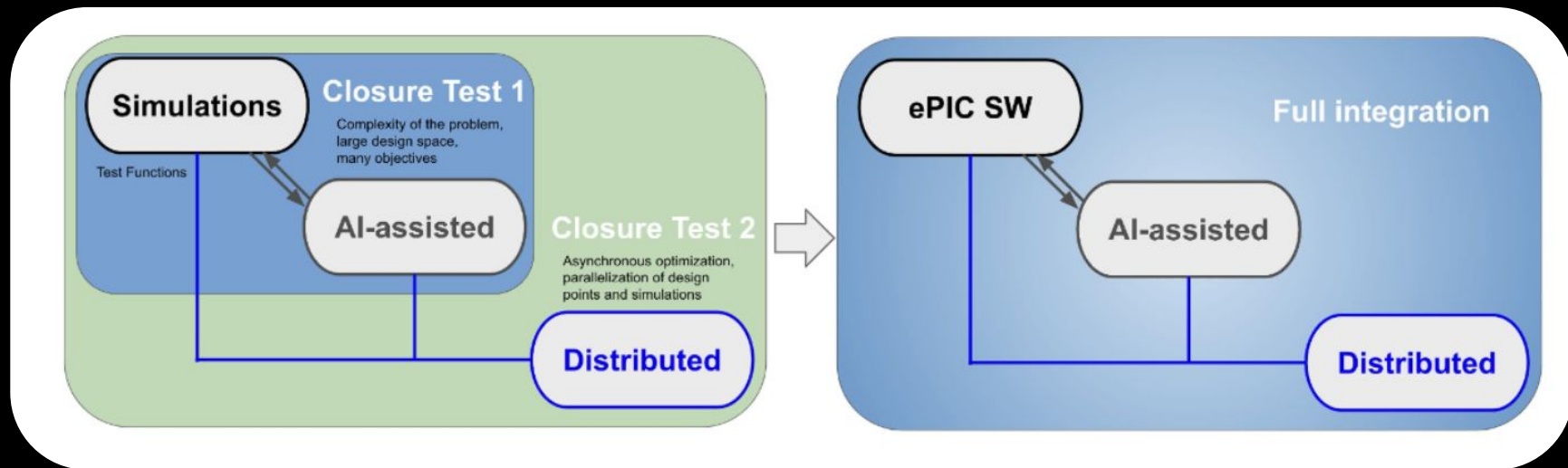
CF, Z. Papandreou, K. Suresh, et al. NIMA: 1047 (2023): 167748.
CF JINST 17.04 (2022): C04038.



(iii) Development of suite of data science tools for interactive navigation of Pareto front (multi-dim design with multiple objectives). Point are determined with uncertainties.



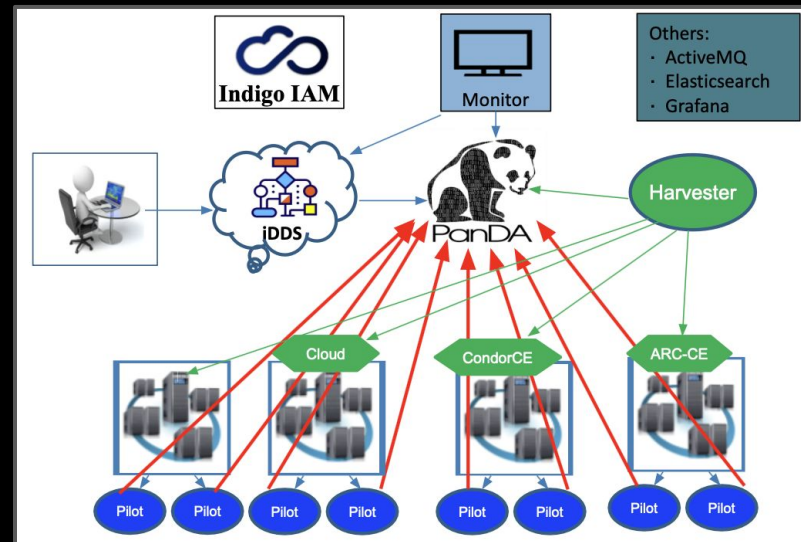
Closure Tests + Full Integration



(Rationale)

Goals:

- **Enhance Workflow Management for Design Optimization:** Adapt [PanDA/iDDS](#) AI/ML services to support a Function-as-a-Task workflow management for design optimization with MOO
- **Ensure System Scalability and Robustness:** Stress-testing scalability, robustness across distributed resources
- **Assessing Consistency:** Compare results against the closure test to evaluate consistency.



PanDA (Production and Distributed Analysis system):

- Distributed Workload Management
 - General interface for users, one authentication for all sites
 - Integrate different resource providers (Grid, Cloud, k8s, HPC and so on), hide the diversities from users, large scale

iDDS (intelligent Data Delivery Service):

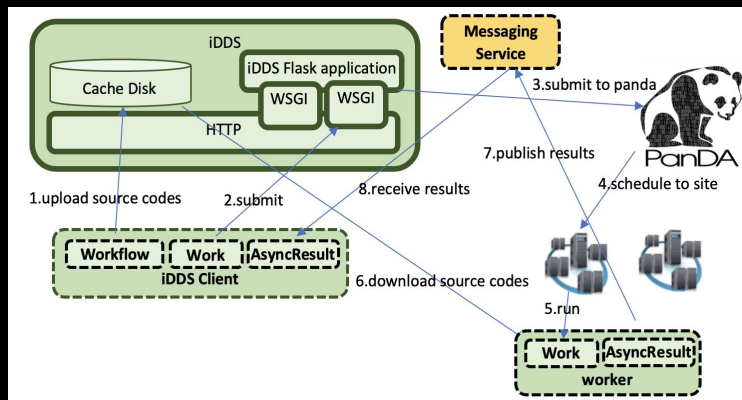
- Workflow Management Orchestration

CHEP2023 Talk: T. Maeno, et al. Utilizing Distributed Heterogeneous Computing with PanDA in ATLAS

CHEP2023 Talk: W.Christian, et al. Distributed ML with PanDA and iDDS in ATLAS
iDDS: Intelligent Distributed Dispatch and Scheduling for Workflow Orchestration.
arXiv:2510.02930, Oct 2025 (submitted for publication)

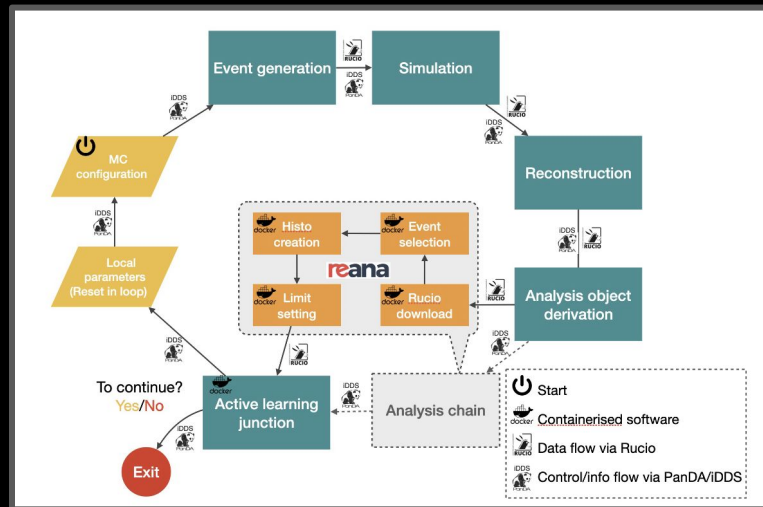
PanDA/iDDS supported complex workflow managements; different use cases in production:

- Fine-grained Data Carousel for **LHC ATLAS**
- DAG management for **Rubin Observatory** to sequence data processing
- Distributed HyperParameter Optimization (HPO)
- Monte Carlo Toy based Confidence Limits
- Active Learning assisted technique to boost the parameter search in New Physics search space



Schema of how a workflow executes a function at remote distributed resources

Bayesian optimisation based active learning with Panda/iDDS

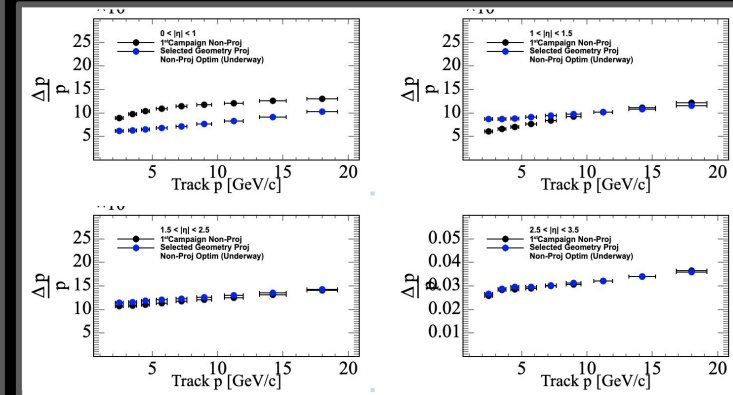
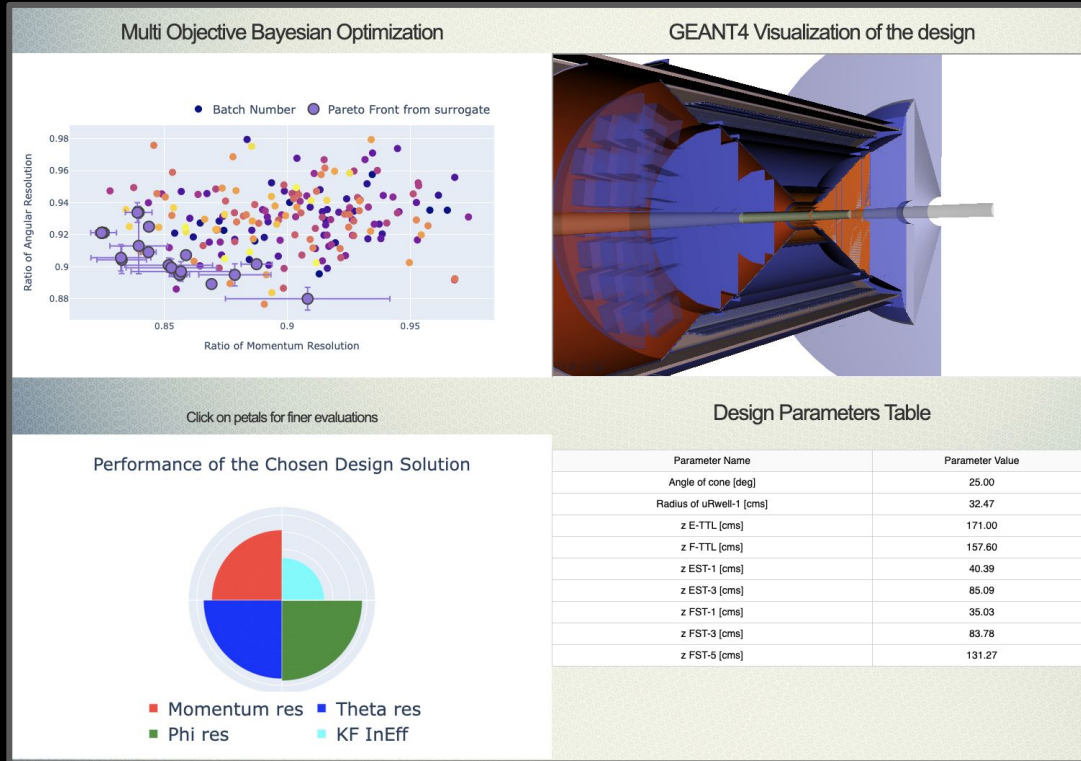


See in what follows **Closure Test 2** (convergence on Pareto fronts) and **full integration results** (using ePIC SW stack)

Interactive Pareto Exploration



C.Fanelli et al, NIM A, 2023, 167748



A graphic featuring the letters 'AI' in a large, black, serif font. The background is a blue-tinted image of a human head profile in silhouette, filled with a circuit board pattern and binary code (0s and 1s).

AI

Deliverables, Budget and Timeline

A graphic featuring the letters 'DE' in a large, black, serif font. The background is a blue-tinted image of a modern building with a glass facade, reflecting the sky and surrounding environment.

DE²

Gantt Chart & Budget



Deliverables	Fiscal Quarter After Award							
	FY24				FY25			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Closure test 1 — MOBO framework (ePIC complexity)								
Closure test 2 — Distributed MOBO								
ePIC design parametrization — cont. integr.				NCE				
Objectives/constraints — cont. integr.				NCE				
Performance analysis (detector, physics, costs)						NCE		
Interface AI-assisted/distributed — V&V, API for grid jobs								
AI-assisted design — coupling MOBO to ePIC SW								
Deployment of AI-optimization pipelines							NCE	
Distributed ePIC simulations				NCE				
Full integration					NCE			
Deployment of distributed AI-optimization pipelines							NCE	

All	Year 1	Year 2	Total
a) Funds allocated (\$k)	622	777	1,399
b) Actual costs to date (\$k)	440	582	1,022
Funds allocated (\$k)	Year 1	Year 2	Total
BNL	190	200	390
CUA	46	46	92
Duke	90	92	182
JLab	110	194	304
W&M	186	245	431
Actual costs to date (\$k)	Year 1	Year 2	Total
BNL	173	192	365
CUA	0	14	14
Duke	90	48	138
JLab	0	87	87
W&M	177	241	418

- All FY24 deliverables have been largely completed, with only minor activities pending due to delays associated with hiring processes.
- A substantial portion of the FY25 deliverables has also been achieved.
- Remaining tasks have been extended under no-cost extensions (NCEs) across all participating institutions to accommodate administrative and onboarding delays experienced over the past two years.
 - William & Mary (W&M) and Duke University: extended through February 2026
 - Brookhaven National Laboratory (BNL): extended through FY26
 - Catholic University of America (CUA) and Jefferson Lab (JLab): extended through approximately January 2027
- Further details on the completed and ongoing work are provided in the following.



(Recent) Milestones



AxScheduler: Translates Ax trials into executable design jobs — defining parameters, code, and job structure for automated execution

Job / MultiStepJob: Unified abstraction for single and DAG-based multi-step workflows.

Runners: Provide a unified interface to submit, monitor, and manage Job execution across heterogeneous computing backends.

JoblibRunner

Lightweight local runner for parallel function evaluations using Joblib

SlurmRunner

Distributes and manages Jobs across SLURM nodes on local HPC systems

PanDAiDDSRunner

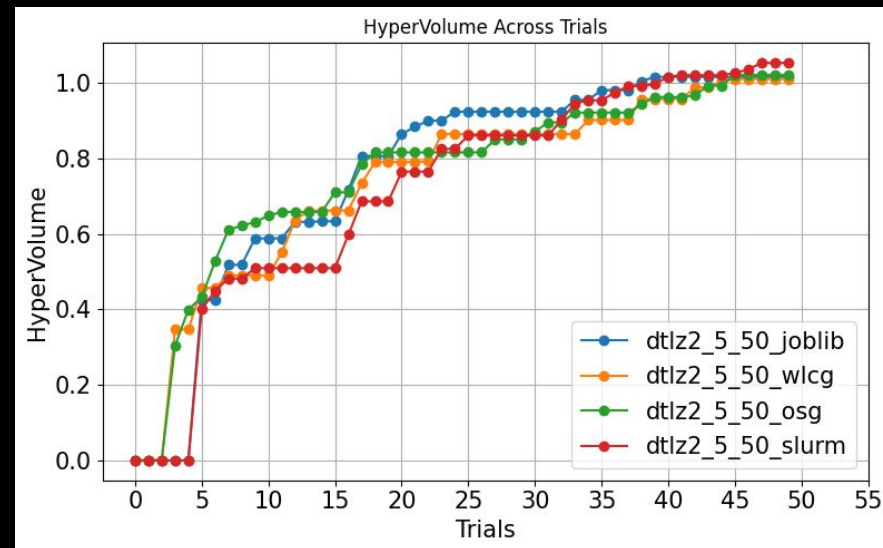
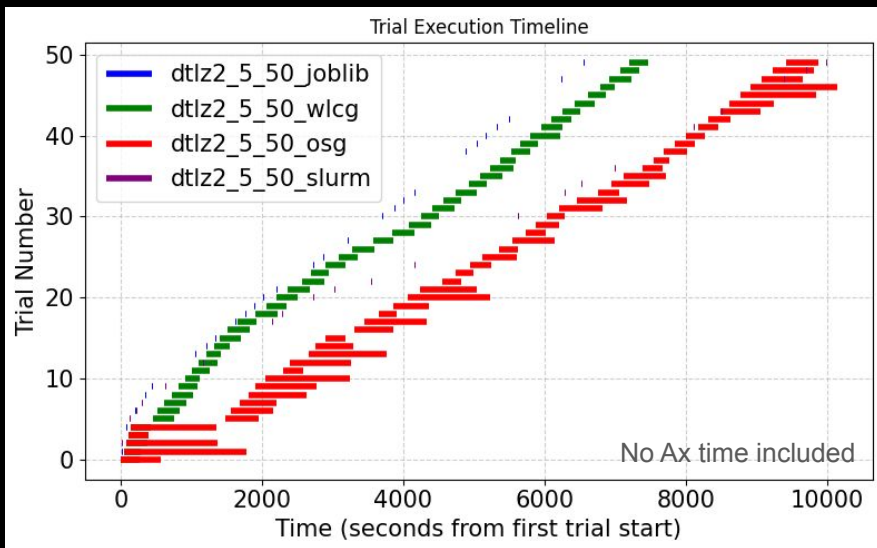
Tar-packages scripts and dispatches them to distributed PanDA iDDS infrastructure for remote execution.

~ increasing queue time, hence recommended for increasingly complex workflows

Stress Tests Across Job Schedulers



- Joblib, Slurm, PanDA (OSG and WLC)
Studies with DTLZ benchmarks (known Pareto front)



- The width of the line represents the total execution time for the trial. Joblib has the least time, Followed by Slurm and then WLCG and OSG (across US grids).
- All the runners perform similar optimization (plot on right)

MultiStep job capability



- A **MultiStepJob** breaks a large task into smaller connected steps. Each step can have its own runner (i.e., SLURM, joblib, PanDA), and steps can run sequentially or in parallel, based on dependencies.
- Key advantages:
 - **Reusable**: Steps can be shared across workflows
 - **Transparent**: Each step tracked individually
 - **Scalable**: Parallel step execution within a job*
- An example Workflow execution for a MultiStep job is shown below. The steps can have dependencies, represented as arrows, across jobs.

More info at this [link](#)

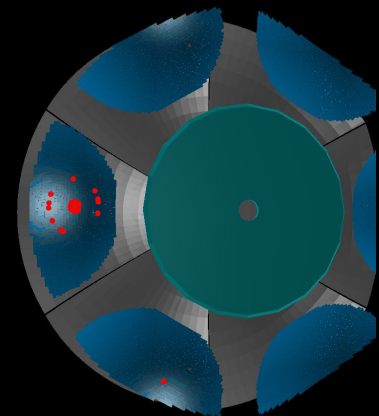
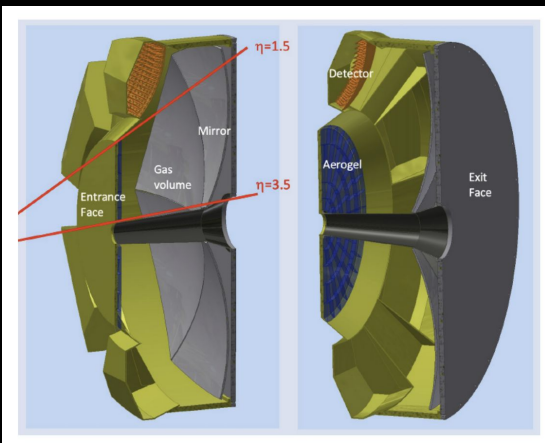
Step 1: Prepare Input
JobLib Runner

Step 2: Run parallel
Simulations
PanDARunner

Step 3: Analyze result
SlurmRunner

Step 4: Aggregate →
JSON output
JobLib Runner

AID2E in Action: dRICH example



C. Pecar (Duke), C. Nunez (Duke)

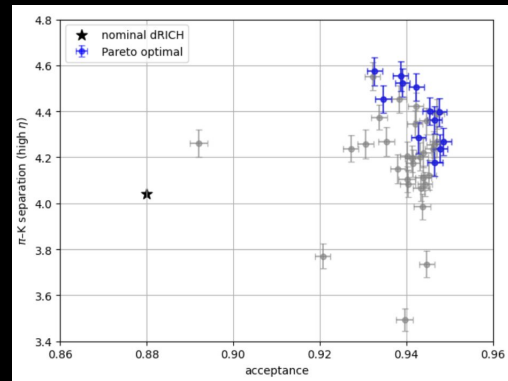
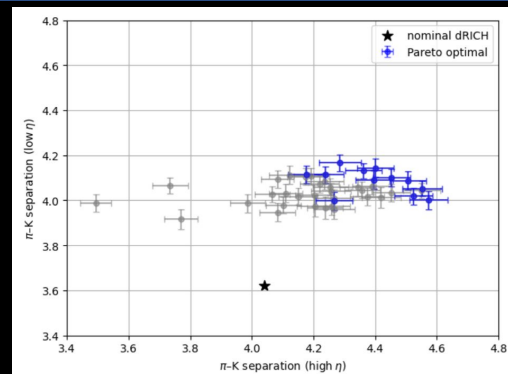
- **Particle ID in $1.5 < \eta < 3.5$**
 - Aerogel radiator $n \approx 1.02$
 - C2F6 gas radiator $n \approx 1.0008$
 - Constrained length to 1.2 m
- **Performance goals:**
 - π -K separation in different regions
 - $p = 15$ GeV, $1.5 < \eta < 2.5$
 - $p = 45$ GeV, $2.5 < \eta < 3.5$
 - Photon acceptance ($N_{ph} > 5$)

parameters 6 pars: 1 mirror + sensor box + snout length

objectives

$$N_{MAE}^{\pi-K} = \frac{\text{PID separation} \times \text{photon acceptance}}{\langle |\theta_{Ch}^{\pi} - \theta_{Ch}^K| \sqrt{N_{photons}} \rangle}$$

penalty: optical reconstruction error

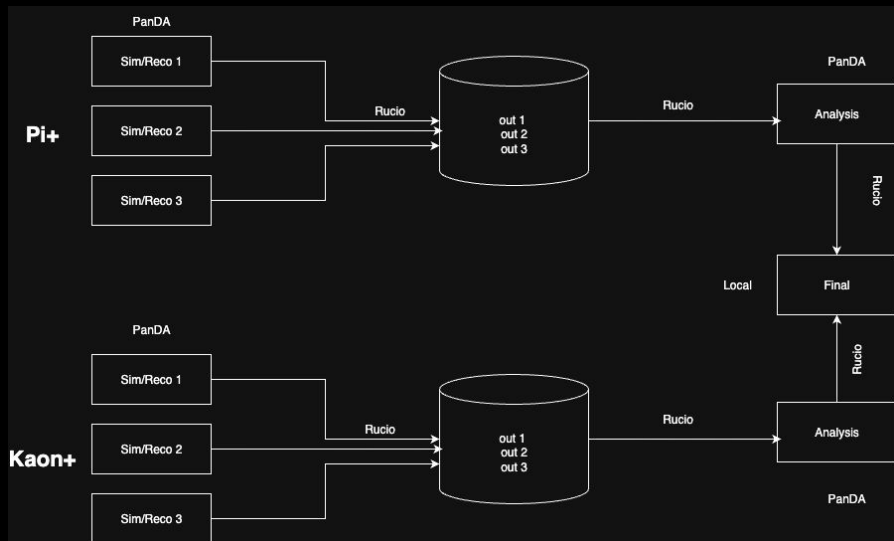


Presented @
ePIC dRICH sim WG

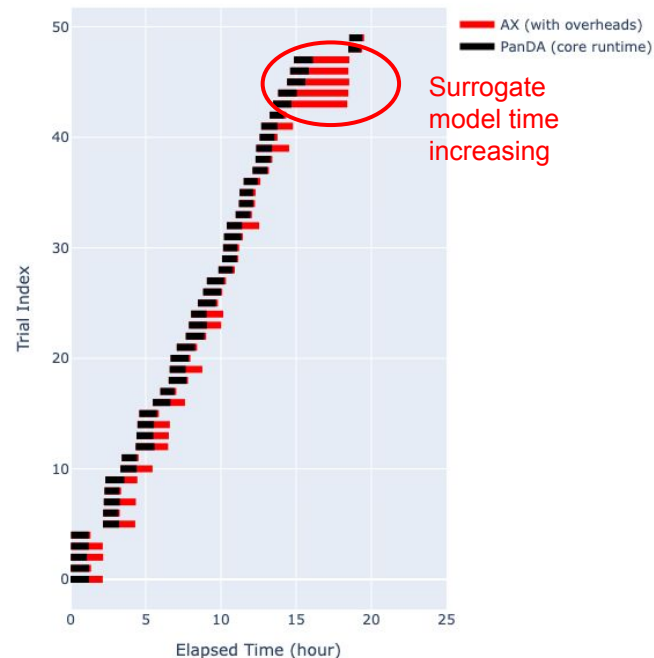
AID2E in Action: dRICH example

dRICH MOBO workflow works as a multistep job:

The optimization of parameters depend on pion and kaon performance across different momentum ranges in each trial.

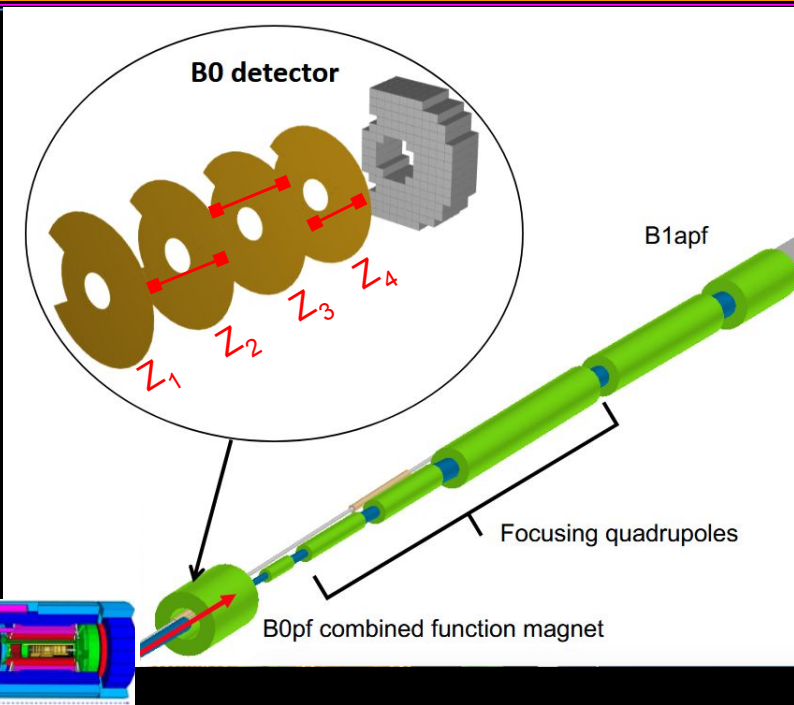


Trial Duration vs Elapsed Time

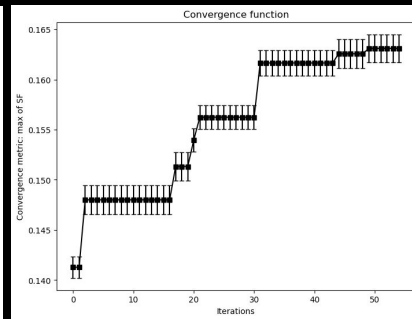
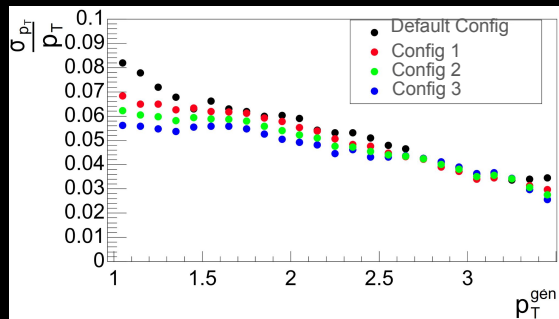


- **Red color** shows the overhead for each ax-trial (model's fit time + gen time)
- **Black color** gives information on cost of overheads when job is idle or in queue. PanDA core time is the time spent on queue and design evaluation time (Sim+Reco+Ana) for each design point

AID2E in Action: Far-Forward B0 example



- Pars: $z_1, z_2, z_3, z_4, c_{n,m}$ (active zones decomposition Zernike poly)
- Inputs: protons, or Λ from Sullivan process @ 5x41, 10x100, 18x275 GeV
- Objective: P or P_T resolution in B0
Acquisition function: logNEI; Initialization: $5xN_{\text{pars}}$ for final calculations; $1xN_{\text{pars}}$ for batch size; Stopping criteria: 100 iterations



Ongoing discussion within ePIC on inclusion of Physics Objectives:

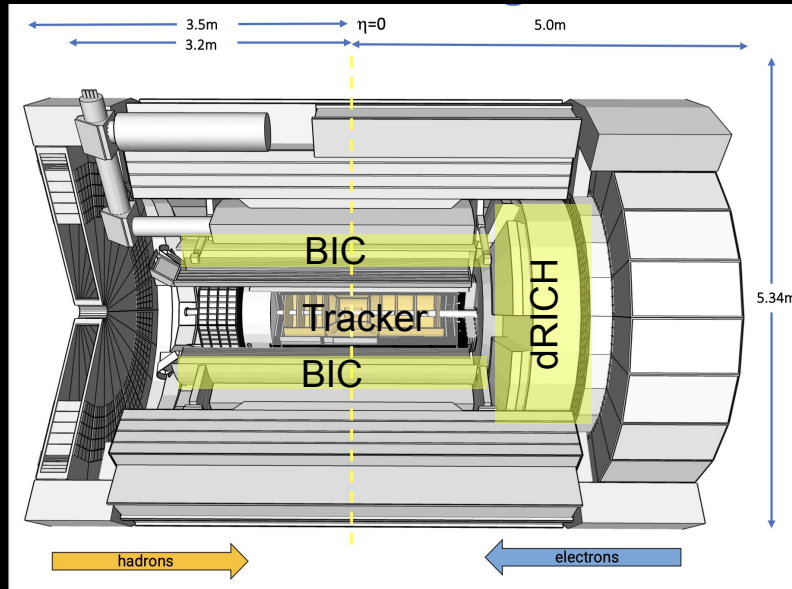
- Λ -reconstruction using multiple ePIC sub-detectors (ZDC, B0, EndCap, LFHcal) — more reconstructed events.
- Ongoing validation: energy correction, vertex resolution, and integration with holistic e^- reconstruction.

Towards Holistic Optimization

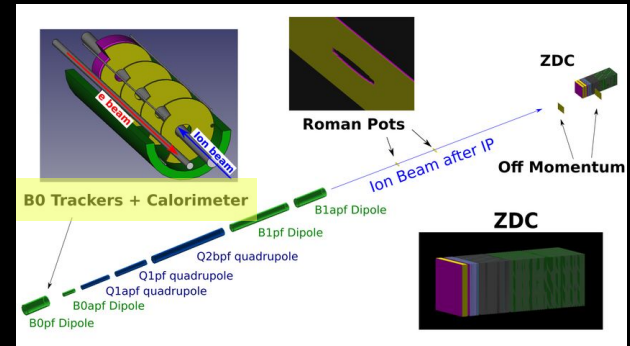


- (Ongoing) Stress-test performance combining Tracker + dRICH + BIC (from central detector) + Far-Forward B0 $\sim O(50)$ pars + 4-5 objectives \rightarrow paper in preparation
- Integrating different simulations (particle guns or physics simulators) through ePIC SW stack

Central



+ Far-Forward





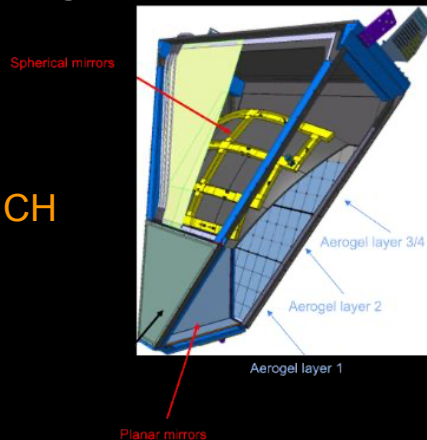
Spin-Off & Other Activities



Spin-Off: CLAS12 RICH Alignment



- Multiple mirror geometry
- 'Black-box' simulations make conventional alignment difficult
 - AID2E used for first successful alignment of CLAS12 RICH with 480 parameters using TurBO
- Significant improvement in performance



Duke

Spin-Off: hKLM compact μ Id/HCAL/ToF

Simultaneous optimization of μ Id and HCAL performance

- Very good μ Id and HCAL performance in simulation ($<40\%/\sqrt{E}$)
- Exploration of parameter space provides insights into tradeoffs (e.g. compensation for less material with greater longitudinal segmentation using ML reconstruction)
- Explored **more complex configurations** (pre-shower, radius dependent layer thickness)
- Future work will include PID in objectives and simultaneous optimization of reco algorithms

→ Another successful AID2E utilization:

Design and Expected Performance for an hKLM at the EIC, [ArXiv: 2511.08432 \[physics.ins-det\]](#)

(32)

Spin-Off: Material Design

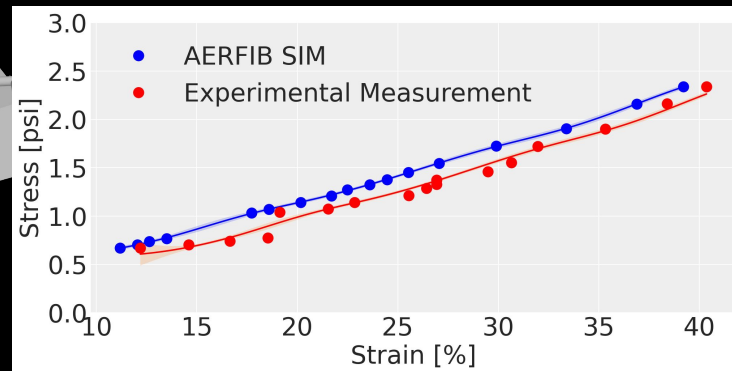
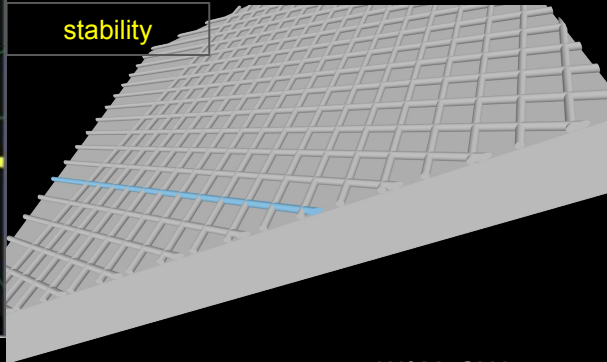
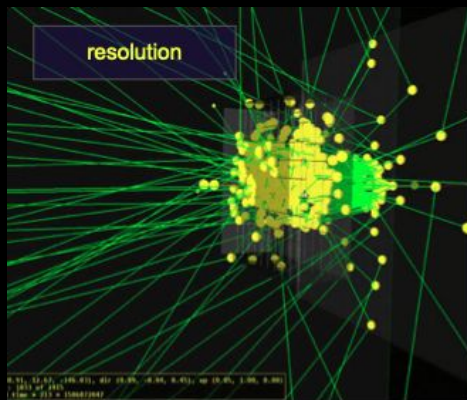
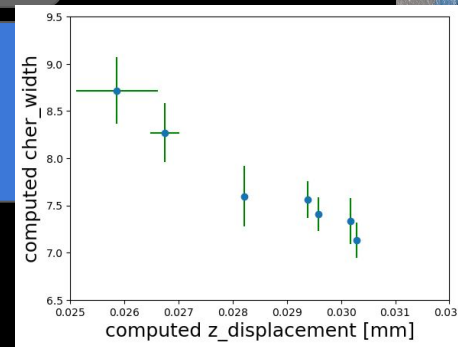


Reinforced novel aerogel material with fibers

Simple Ring Imaging CHerenkov Geant4 based simulation
Aerogel + Optical Fibers

Gmsh - define geometry and produce mesh
ElmerGrid - convert the gmsh mesh to elmer compatible mesh
ElmerSolver - do modeling (solve linear and nonlinear equation)
Paraview - visualize Elmer Solver and provide a python interface to automate

Aerogel tile with random fiber orientation



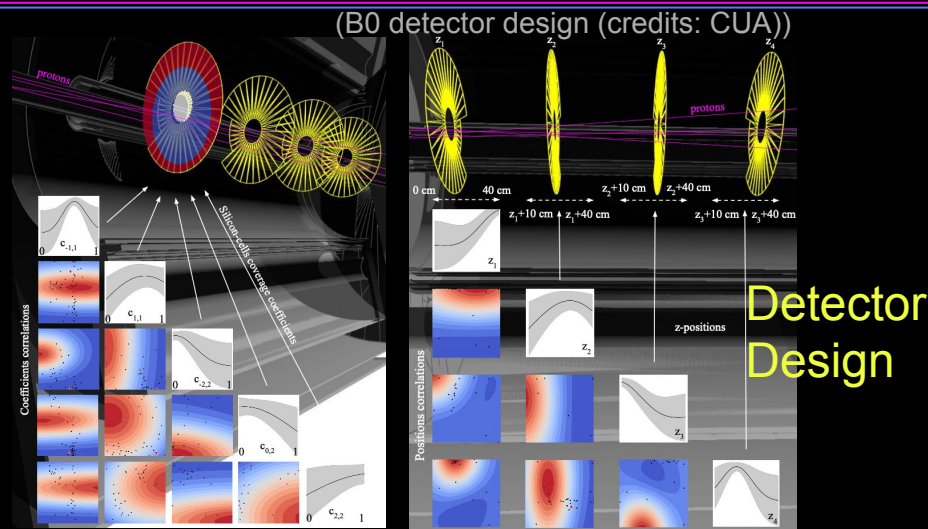
Publication in preparation

W&M, CUA

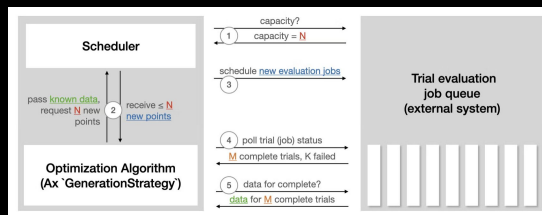
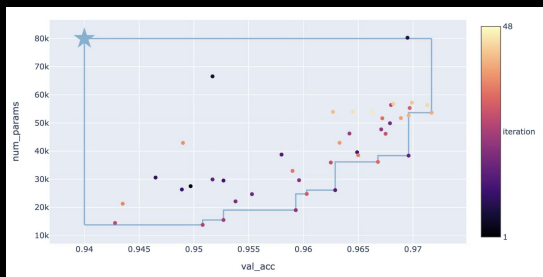
AID2E Broad Impact



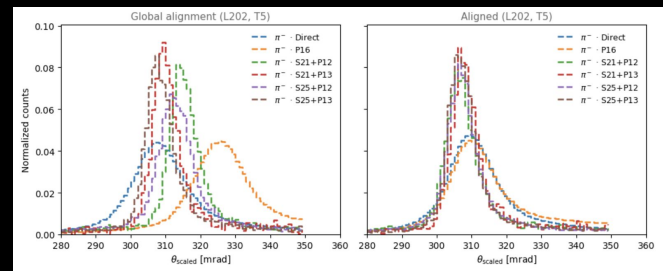
- **Detector Design**
- **Detector Optimization**: Alignment (cf. CLAS12 RICH using AID2E), Calibration
- **Neural Architecture Search** — NN performance vs complexity
- (Many) **Other Applications**



Neural Architecture Search



<https://pytorch.org/blog/effective-multi-objective-neural-architecture/>



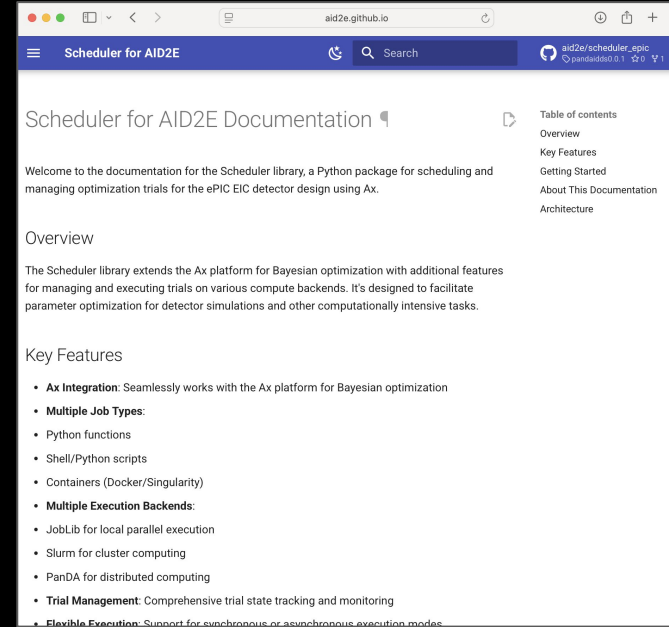
Detector Optimization

(CLAS12 RICH Alignment (credits: Duke))

Enhancing Usability & Flexibility



- Make AID2E easy to discover, adjust and utilize.
 - Implementing an intuitive interface for broad adoption
 - GitBook & other interactive knowledge sharing platforms are part of the initiatives related to **Documentation** and work dissemination
 - See, e.g.,: http://aid2e.github.io/scheduler_epic/
https://deepwiki.com/aid2e/scheduler_epic/4-multi-step-workflows



- **AID2E bootcamp @W&M** (1 week + final projects) — Opportunities for experiential learning with easy access for beginners
 - <https://aid2e.github.io/boot-camp-2024/>

A blue-tinted graphic featuring a silhouette of a human head in profile, facing right. Inside the head, there are glowing circuit lines and binary digits (0s and 1s). The letters 'AI' are prominently displayed in a large, black, serif font over the head.

AI

Project Extension

A blue-tinted graphic showing a stylized, futuristic cityscape or industrial structure with various levels and platforms. The letters 'DE' are written in a large, grey, serif font, with a superscript '2' (DE²) positioned above the 'E'.

DE²

Extension of AID2E



Extend AID2E framework (PanDA+MOO) by adding LM agent for autonomous optimization and system control with HITL

LM Agent



prompt

monitor/control

- Optimization Strategy
- Distribution/Orchestration
- Code Generation

```
for total_epochs in range(epochs):
    epoch_loss = 0
    total_graphs = 0
    not_train = 0
    for batch in train_loader:
        batch.to(device)
        optimizer.zero_grad()
        output = net(batch)
        loss = F.mse_loss(output, batch.y[:, target_idx].unsqueeze(1))
        loss.backward()
        epoch_loss += loss.item()
        total_graphs += batch.num_graphs
    optimizer.step()
    train_avg_loss = epoch_loss / total_graphs
```

Human in the loop



LM automates the workflow:
optimization pipelines, tasks
distribution, code generation

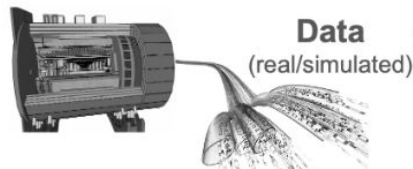
Human-in-the-loop control:
experts monitor, validate,
and guide the process

Multi-objective optimization:
with tools such as BoTorch
and pymoo

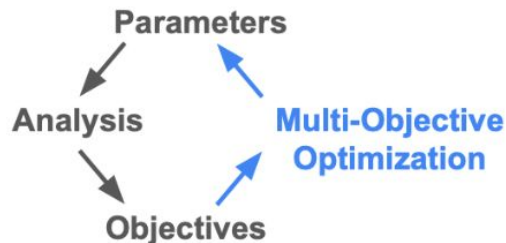
**Distributed workflow
management:** enabled
through PanDA



DE-FOA-0003458



Data
(real/simulated)



BoTorch
pymoo

What is Retrieval Augmented Generation (RAG)?

- Access up to date information without explicitly training of LLM.
- Reduce “Hallucination” of LLM.
- Grounding LLM to truth to increase reliability by providing citations.

(agent) <http://rag4eic.ds.wm.edu/>

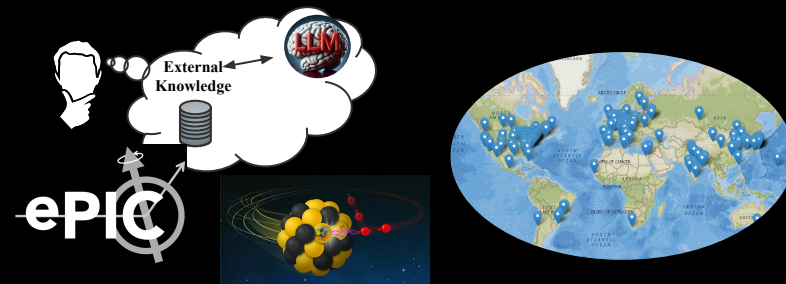
(source) <https://github.com/oi4eic/EIC-RAG-Project>

Why need RAG for Large Scale Physics Experiments?

- EIC large scale experiment (e.g., EICUG ~1,500 users, ePIC 180 institutions)
- Regular updates to documents, Wiki etc; Tot document size ~ scale of experiment
- Newbies may take months to get to know the full experimental details

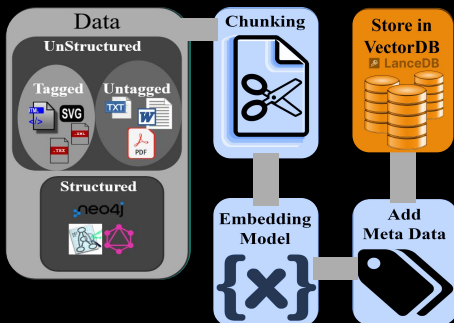
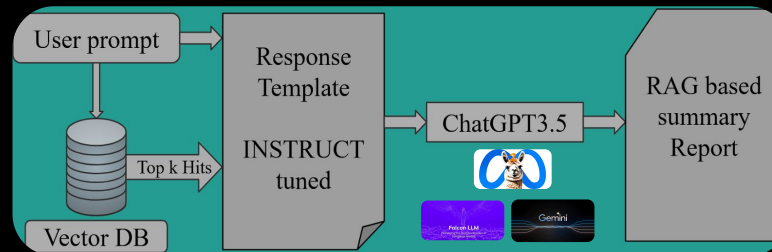
“Ingestion” of data

- Creation of vectorized knowledge base
- Every node below influence RAG performance
- 200+ recent arXiv papers on EIC (since 2021)



“Inference”

- Given a prompt compute similarity index to most similar vectors in VectorDB
- Use LLM to further narrow down and summarize the finding





Conclusions



Conclusions

<https://aid2e.github.io/aid2e>



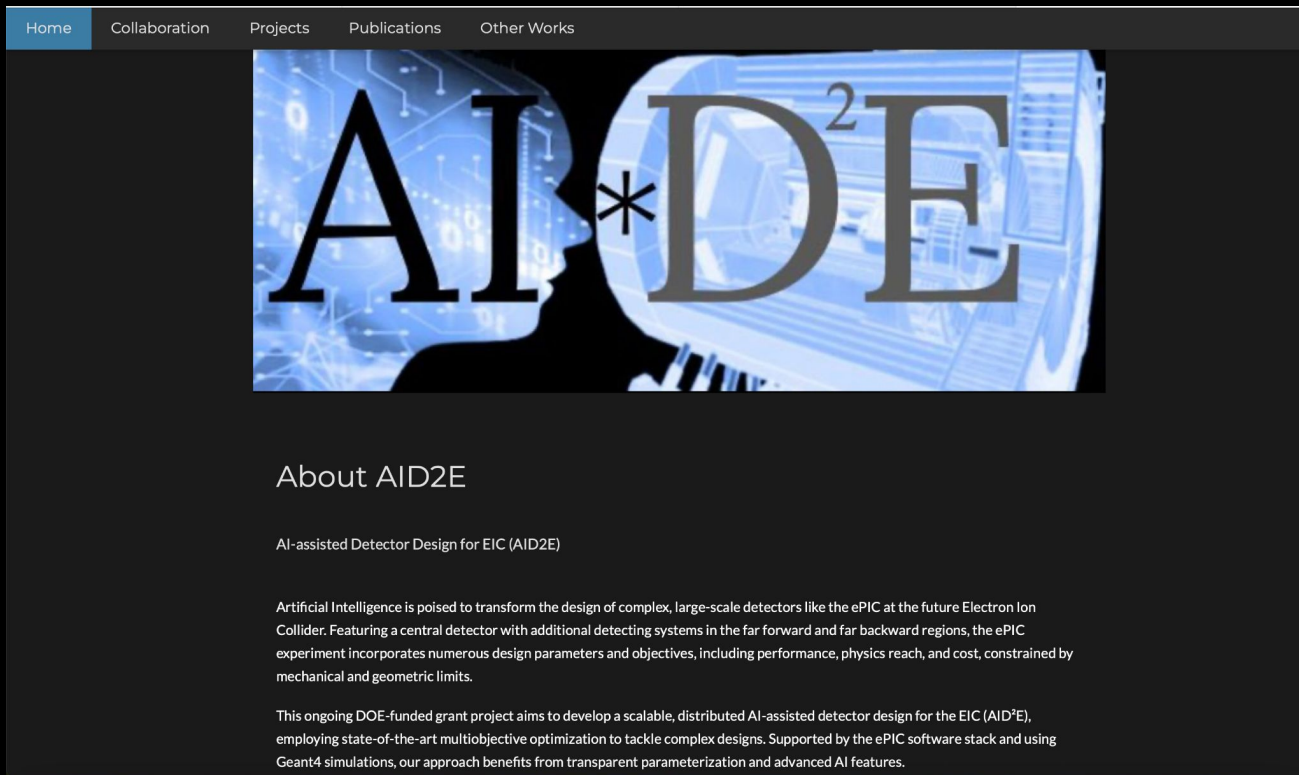
The EIC could feature the first large-scale experiments designed and optimized with the aid of Artificial Intelligence.

By applying AI to detector design and optimization, AID2E is a prime example of AI's role in automating scientific workflows.

- **AID2E integrates advanced workload management and multi-objective optimization using full ePIC software stack**
 - Completed FY24 deliverables (distribution, scalability, closure tests)
 - Several major FY25 milestones already achieved; on track through NCE
- **A new AID2E paper in preparation**
 - Will highlight scalability, compute performance, and integration with ePIC workflows
 - Draft targeted by end of the calendar year
- **Holistic detector optimization is now within reach**
 - Alignment with EIC project schedule
 - Demonstrated subsystem applications (dRICH, far-forward detectors)
 - Working with ePIC Collaboration to integrate AID(2)E into ePIC simulation/optimization campaigns
- **Towards a general framework for the community**
 - Applications include Detector-2@EIC (designed to complement ePIC), FCC concepts at CERN, and other large-scale detector/accelerator optimization problems
 - Adaptable to calibrations, alignments, materials R&D, and other compute-intensive tasks across experiments
 - Supports iterative (re)design, useful during different phases of an experiment (e.g., construction phase with budget and material constraints)
- **Ongoing priority is delivering a user-friendly, generalizable optimization framework**
 - Work is underway to make the tool accessible to non-experts by ensuring it is easy to discover, adjust, and tune

The timing of AID(2)E could not be better as the preTDR / preparation for CD-2 is underway...

<https://aid2e.github.io/aid2e>



The screenshot shows the AID2E website homepage. At the top is a navigation bar with links: Home, Collaboration, Projects, Publications, and Other Works. Below the navigation bar is a large hero image featuring the AID²E logo, which consists of a stylized head profile with circuitry and the text AID²E. Below the hero image is the section "About AID2E".

About AID2E

AI-assisted Detector Design for EIC (AID2E)

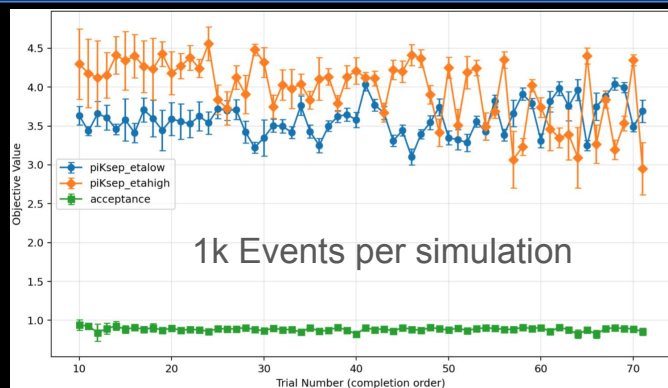
Artificial Intelligence is poised to transform the design of complex, large-scale detectors like the ePIC at the future Electron Ion Collider. Featuring a central detector with additional detecting systems in the far forward and far backward regions, the ePIC experiment incorporates numerous design parameters and objectives, including performance, physics reach, and cost, constrained by mechanical and geometric limits.

This ongoing DOE-funded grant project aims to develop a scalable, distributed AI-assisted detector design for the EIC (AID²E), employing state-of-the-art multiobjective optimization to tackle complex designs. Supported by the ePIC software stack and using Geant4 simulations, our approach benefits from transparent parameterization and advanced AI features.

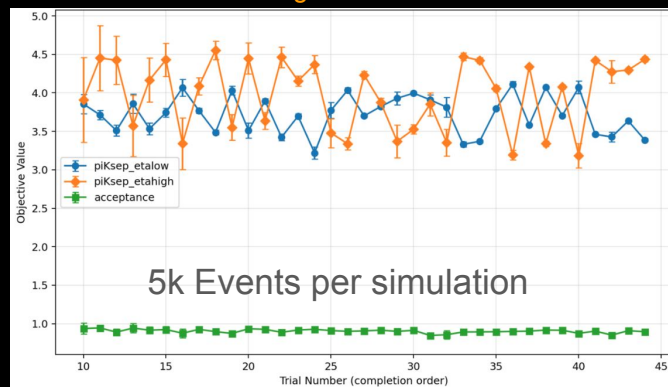
Backup



dRICH Performance Analysis on PanDA



1k & 5k: The mean of π^-/K @ low $\eta \approx 3.5-4.0$, π^-/K @ high $\eta \approx 4.0-4.5$, and acceptance ≈ 0.9 . - For 1k: there's **more scatter** and **wider uncertainty**, The optimizer's feedback (posterior update) is noisier, which can **slow convergence**



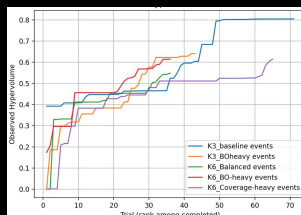
The Objective: Identify minimum stats needed for simulation to achieve statistical significant convergence in shortest number of iterations

- The 100-event run has the noisiest objectives (and therefore the slowest convergence)
- 5k gives cleaner signals early (fast initial HV gain) but appears to plateau.
- Since 1k and 5k means look similar once you ignore error bars, 1k is “good enough” as a default, and 5k is best used selectively to make early modeling cleaner (initial quality). e.g. if we have a smaller budget for #trials, we run 5k events.

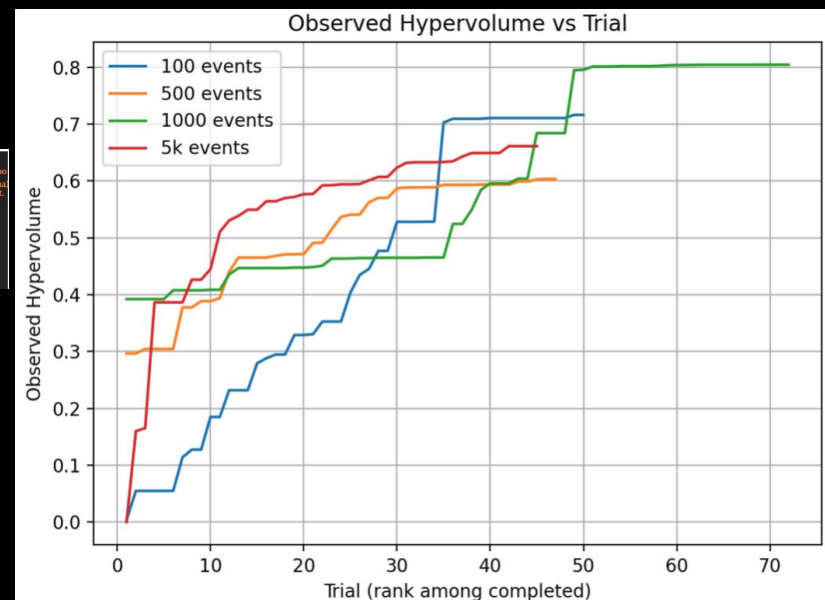
Concurrency = 3

```
Summary
Tasks total: 600
done: 600
is point combination tasks * go trials - 600
Each task runs 1000 events (sim reco ana)
Multithreaded is not supported in sim yet.

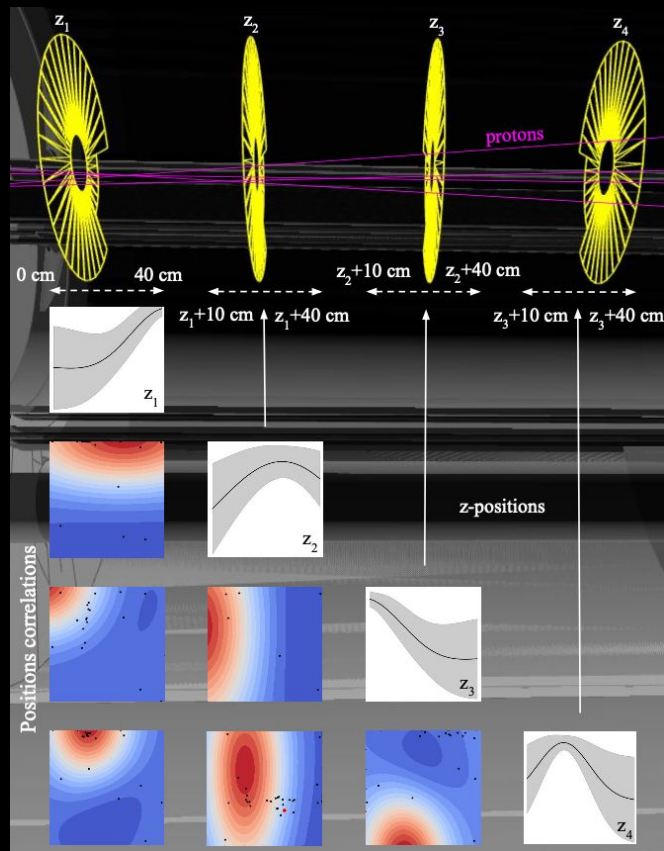
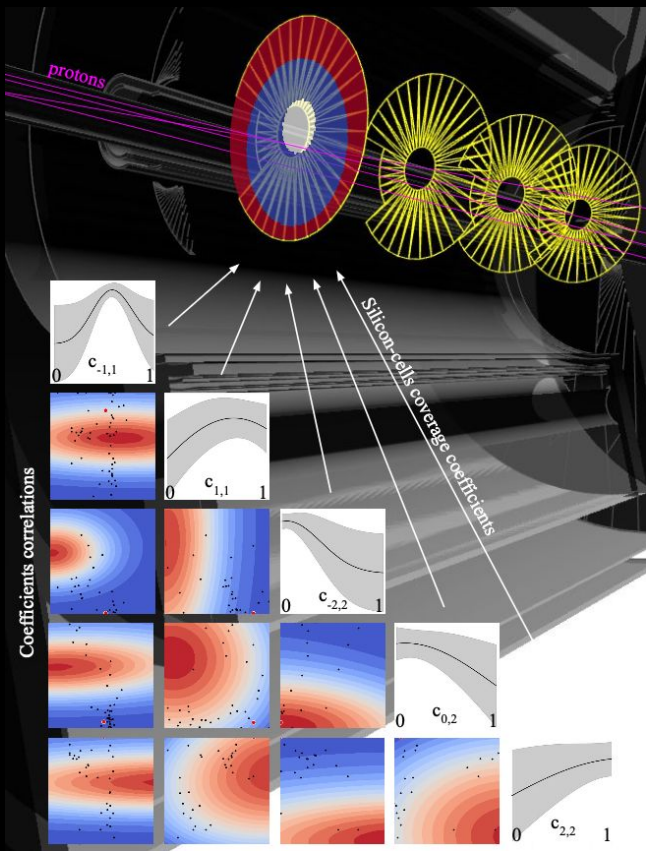
Wall time (sum): 325.88 h
Core-hours (sum): 325.88 core-h
Avg wall time per task: 0.54 h
Avg core-hours per task: 0.54 core-h
CPU hours (efficiency-weighted sum): 293.29 CPU-h
Avg CPU efficiency (weighted): 90.0%
```



Fang-Ying Tsai (BNL)



AID2E in Action: Far-Forward B0 example



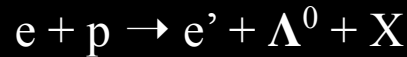
Acquisition function: $\log \text{NEI}$;

Initialization: $5 \times N_{\text{pars}}$

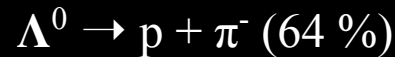
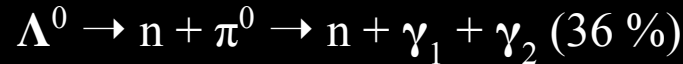
$1 \times N_{\text{pars}}$ for batch size;

Stopping criteria: 100 iterations

- Only one measurement of the **Kaon structure** in the 80's with **Drell-Yan** process [1]
- The future **ePIC** setup at **EIC** facility could bring new and broader data
- *E.g.* through the inclusive measurement of the 'strange' **Sullivan process** :



- The Kaon structure is inferred from the measurement of the Λ momentum
- The **Λ observables are challenging to measure**: unstable with **two decay channels**



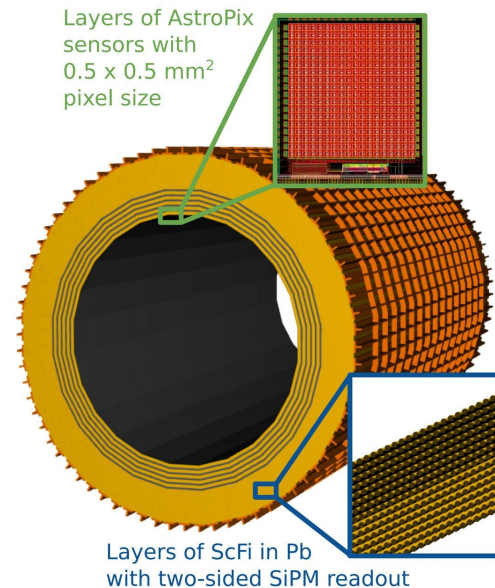
- It is crucial to optimize the ePIC setup to make this first-of-its-kind study possible
- The Sullivan process is a **severe test of holistic optimization** for AID2E project

- **Energy correction** required to finalize the reconstruction: **meeting CCWG Nov. 19th**
- **Validation** of the algorithm: vertex positions (must reproduce the Λ -decay exponential)
- **Charged decay channel** ($\Lambda^0 \rightarrow p + \pi^-$) under investigation (magnets-dependent)
- No-decay : possible detections of raw Λ s to be investigated
- **Crucial need of reconstructing the backward electron for a future holistic opti.**

- **Barrel-Imaging Calorimeter (BIC):** ePIC barrel EMCal (BEMC)
 - Primary purpose is detection of e^- , γ at mid-rapidity
- **Hybrid detector technology:** interleaved layers of Si sensors and Pb+Scintillating Fibers (ScFi)
 - Si layers → Precise position measurement
 - PbScFi layers → Precise energy measurement
- BIC design largely finalized → **valuable testbed for AID(2)E**
 - Both for software integration *and* methodology validation
- **Problem definition:** repeats optimization carried out during BEMC technology selection using AID(2)E
 - **What is the minimum number of Si layers needed to meet physics requirements?**

Problem details:

- 5 parameters
 - > Including/excluding 2nd through 6th Si layer
- Working toward 3 objectives
 - > e^- energy resolution, e^-/π^0 separation, and total cost
 - > All 3 used in prior optimization

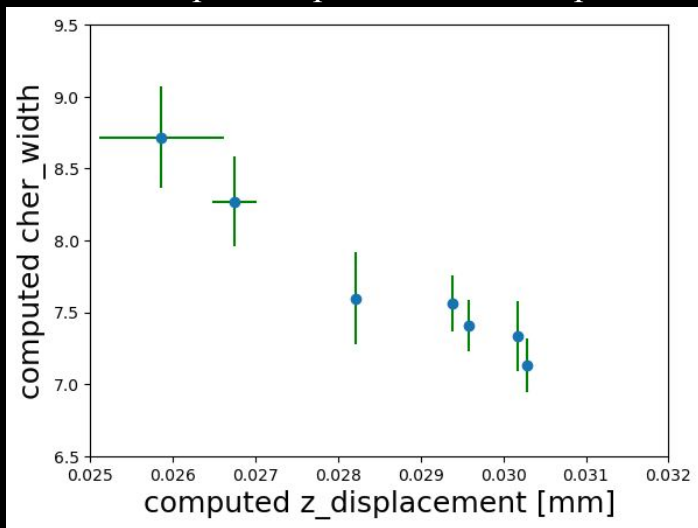


Design Parameters	Range	Multiplicative factor	Tolerance	Remarks
a_thick	46 – 60	1.0 [mm]	1 [mm]	Aerogel tile thickness [mm]
a_width	80 – 120	1.0 [mm]	1 [mm]	Aerogel tile width (height) [mm]
a_rind	1.01 – 1.04	1.0 [no unit]	0.002	Relation between rind and Young's modulus. ParameterType.FLOAT
f_rotx	-5 – 5	0.1 [deg]	0.1 deg ?	Rotations included in FEM simulations. But this implementation has to be revisited
f_roty	-5 – 5	0.1 [deg]	0.1 deg ?	
f_gap	5.0 – 30.0	1.0 [mm]	1 [mm]	Simulation are stable now.
f_diameter	10 – 40	0.005 [mm]	0.005 [mm]	Diameter of the fibres
f_pitch	10 – 30	0.5 [mm]	0.5 [mm]	Distance between fibers along x-y

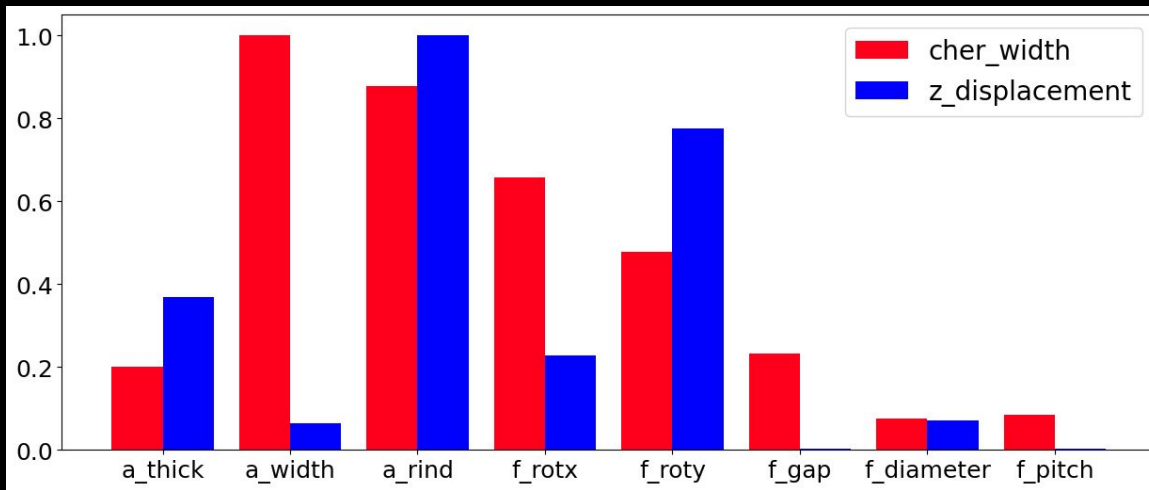
AerFib: What does MOBO find?



Predicted pareto optimal solution 7 points



The most important parameter at end of 60th iteration



comp_mean_cherwidths	comp_mean_z_displacements	comp_sems_cherwidths	comp_sems_z_displacements	a_thick	a_width	a_rind	f_rotx	f_roty	f_gap	f_diameter	f_pitch	pitch
14.26420	0.030282	0.187027	0.000399	58.512657	101.825535	1.040	-3.9	-1.1	17.1	0.147	8.224006	8.22
14.66766	0.030174	0.244024	0.001055	57.765580	103.233279	1.040	-3.5	-1.2	16.4	0.133	9.012985	9.01
14.81014	0.029580	0.179016	0.001717	58.962156	100.458672	1.040	-4.2	-1.1	17.5	0.159	7.573453	7.57
15.11720	0.029380	0.195284	0.001694	59.052033	100.480343	1.040	-4.2	-1.1	17.5	0.158	7.494971	7.49
15.19146	0.028218	0.321963	0.000870	58.235682	102.455963	1.040	-3.8	-1.1	16.8	0.143	8.474173	8.47
16.54180	0.026740	0.314514	0.013417	59.200601	118.177587	1.039	4.5	5.0	3.1	0.050	5.130616	5.13
17.43304	0.025860	0.352385	0.037940	59.266780	118.322944	1.039	4.5	5.0	2.9	0.050	5.120960	5.12