AID(2)E:
AI-Assisted Detector Design at EIC

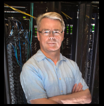
DE-FOA-0002785

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

Cristiano Fanelli on behalf of AID(2)E

Brookhaven National Laboratory
The Catholic University of America
Duke University
Jefferson Lab
William & Mary
Chartered 1693

Team Members



Torre Wenaus, PhD — coPI

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

Brookhaven National Laboratory

Meifeng Lin, PhD

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

Brookhaven National Laboratory

Tianle Wang, PhD

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

Brookhaven National Laboratory

Wen Guan, PhD

Expertise: distributed computing, workflow management, ML workflows

Brookhaven National Laboratory

Gabor Galgoczi, PhD — consulting

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

Brookhaven National Laboratory

Kolja Kauder, PhD — consulting

Expertise: Physics, Simulation

Brookhaven National Laboratory

Alex Jentsch, PhD — consulting

Expertise:

Brookhaven National Laboratory

Tanja Horn, PhD — coPI

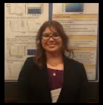
Expertise: medium energy nuclear physics, EIC, 3D hadron Imaging, calorimetry

The Catholic University of America

Anselm Vossen, PhD — coPI

Expertise: Physics, PID and calorimetry for the EIC

Duke University



Cynthia Nunez, PhD

Expertise: Postdoc: physics, simulations

Duke University

Connor Pecar, MS

Expertise: physics, PID and detector simulations for the EIC

Duke University

Markus Diefenthaler, PhD — coPI

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

Jefferson Lab

Makoto Asai, PhD — consulting

Expertise: Detector Simulations, Geant4

Jefferson Lab

Cristiano Fanelli, PhD — PI

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

William & Mary

Karthik Suresh PhD

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

William & Mary

Hemalata Nayak, MS

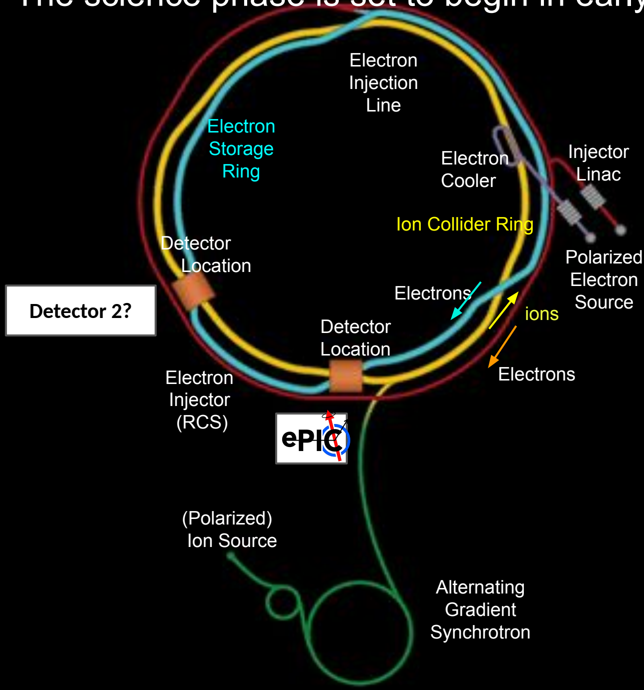
Expertise: Physics, Data Science, Optimization, Reconstruction

William & Mary

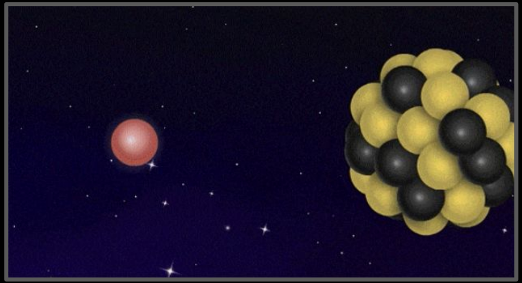
Electron Ion Collider

3 fundamental questions
(Nuclear Science Advisory Committee)

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 early 2030.



polarized electron - polarized protons/ions



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

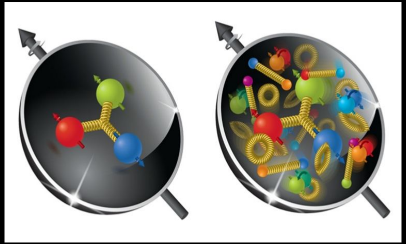
High luminosity up to 10^{34} $cm^{-2}s^{-1}$, a factor ~100-1000 times HERA

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

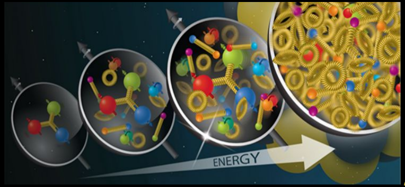
AI/ML will play a major role in optimizing this complex operation

How does the mass of the nucleon arise?

How does the spin of the nucleon arise?



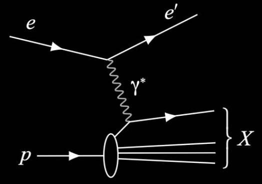
What are the emergent properties of dense systems of gluons?



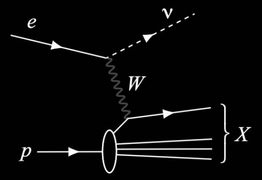
A Glimpse into EIC Physics

EIC Yellow Report. NIMA 1026 (2022): 122447
HERA Coll. , JHEP 1001:109(2010)

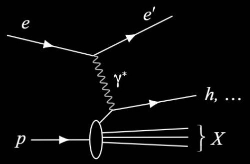
Neutral current inclusive DIS



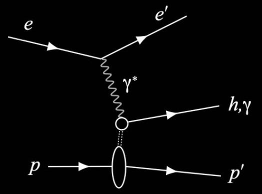
Charged current inclusive DIS



Semi-inclusive DIS



Exclusive DIS



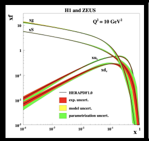
Detector requirements and design are tailored to optimize physics reach, guided by the EIC Yellow Report:

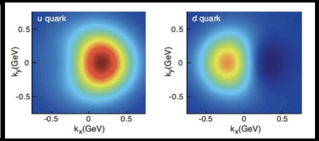
- Mass and Tomography
- Spin and Flavor Structure of the Nucleons and Nuclei
- Internal Landscape of Nuclei
- QCD at Extreme Parton Densities - Saturation
- etc

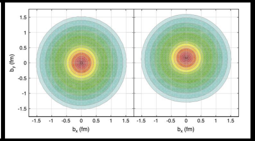
Important synergies with HL-LHC science program:

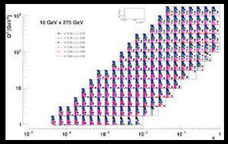
- Precision QCD studies with proton & nuclear targets $\boldsymbol{\alpha_s}$, quarkonia, quark exotica, jet physics in e-p collisions, …
- Precision electroweak and BSM physics Weak mixing angle, LFV, …
- etc

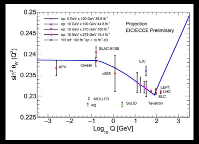
For more info in the RAG-based EIC Chatbot







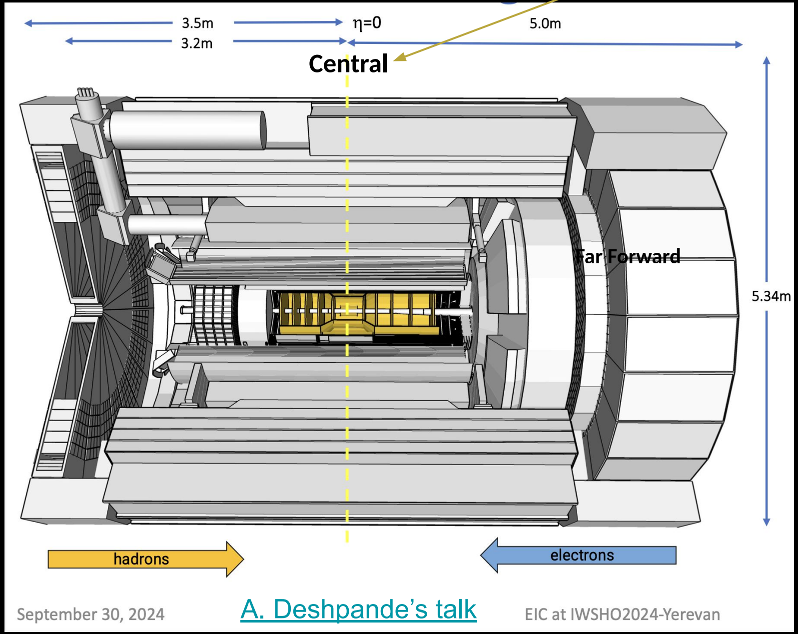




ePIC Detector

As of now, 171 institutions, 24 countries and 500+ participants

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



Tracking

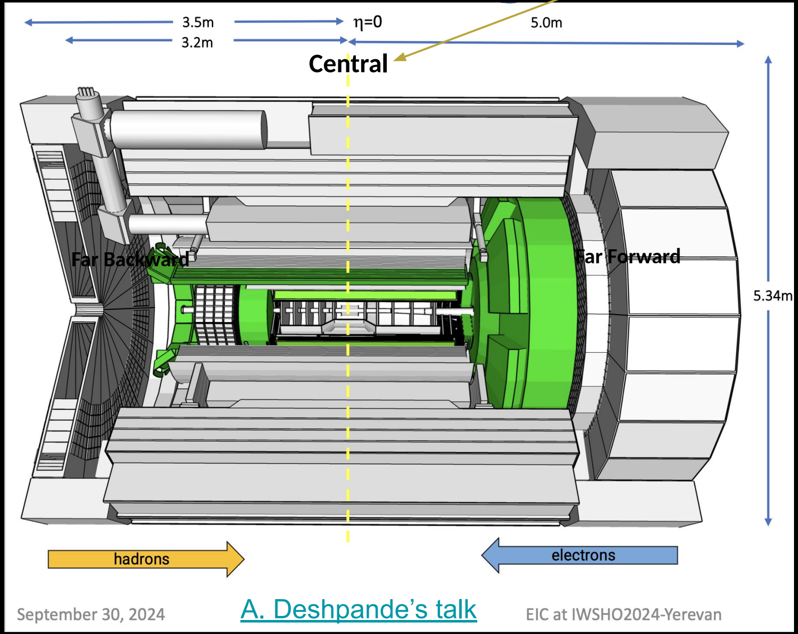
- New 1.7T solenoid
- Si MAPS Tracker
- MPGDs (μRWELL/μMegas)

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

ePIC Detector

As of now, 171 institutions, 24 countries and 500+ participants

ePIC stands out as an **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 (μRWELL/μMegas)

PID

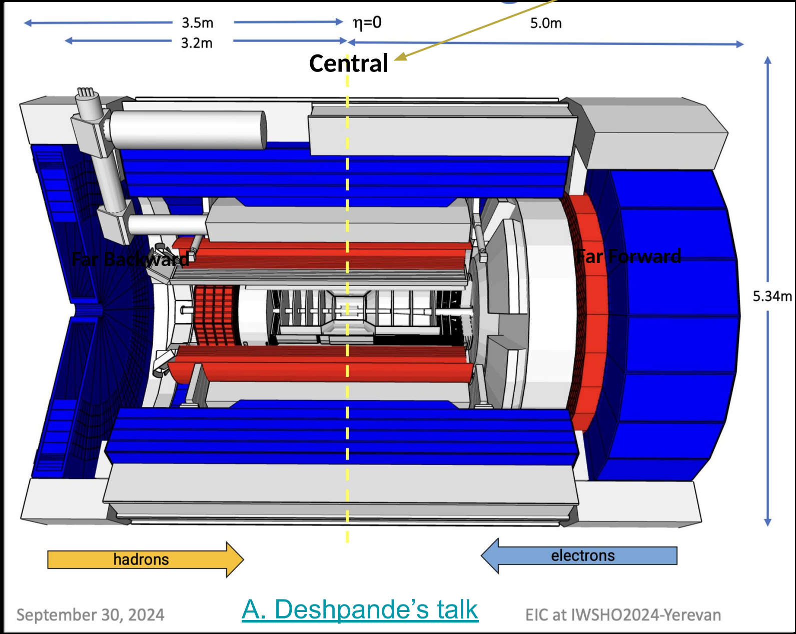
- hpDIRC
- pfRICH
- dRICH
- AC-LGAD (~30ps TOF)

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

ePIC Detector

As of now, 171 institutions, 24 countries and 500+ participants

ePIC stands out as an **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 (µRWELL/µMegas)

PID

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

Calorimetry

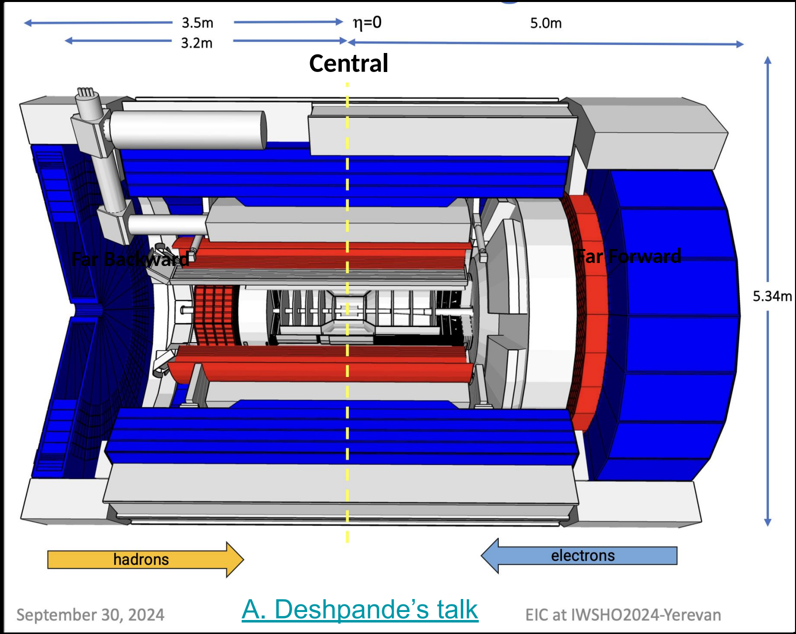
- Imaging Barrel EMCal
- PbWO4 EMCal in backward direction
- Finely segmented EMCal +HCal in forward direction
- Outer HCal (sPHENIX re-use)
- Backwards HCal (tail-catcher)

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

ePIC Detector

As of now, 171 institutions, 24 countries and 500+ participants

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

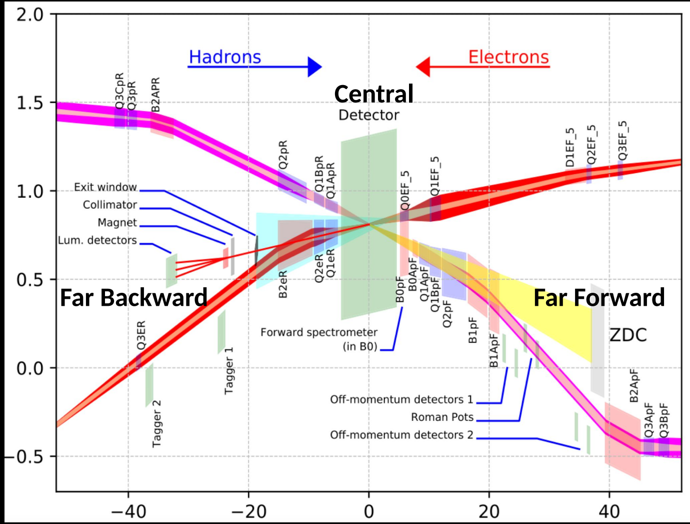


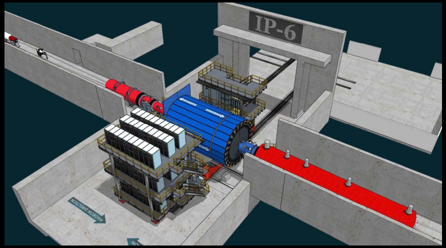
Tracking
- New 1.7T solenoid
- Si MAPS Tracker
- MPGDs (μRWELL/μMegas)

PID
- hpDIRC
- pfRICH
- dRICH
- AC-LGAD (~30ps 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 extends across -35m to +35m.





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

Traditional Approach to Detector Design

- Sub-detector systems are optimized individually, using single-objective criteria per sub-detector, within the constraints of the overall detector design. This approach often leads to suboptimal solutions.
- Each combination of sub-detector choices creates a new overall detector design. Accurate and reliable full simulation pipelines are required to reduce possible bias when exploring new designs. Fast simulations may become unreliable in regions that have not been previously explored or validated.
- Large simulation campaigns are required, often leveraging containerized software and distributed computing (e.g., NIM-A: 1047 (2023): 167859):
  - Each "design point" (a new detector configuration) potentially needs a new simulation campaign.
  - Exploring multiple design points demands significantly more simulations, increasing computational costs and complexity.

Reconstruction and Simulation Times	Times based on current software on modern cores
Reconstruction event processing time with background **[s]**	2
Reconstruction algorithmic speedup factor 10yrs out	1.5
Simulation event processing time with background **[s]**	15
Full simu speedup factor 10yrs out	1.5
Combined time with background, with speedup **[s]**	11

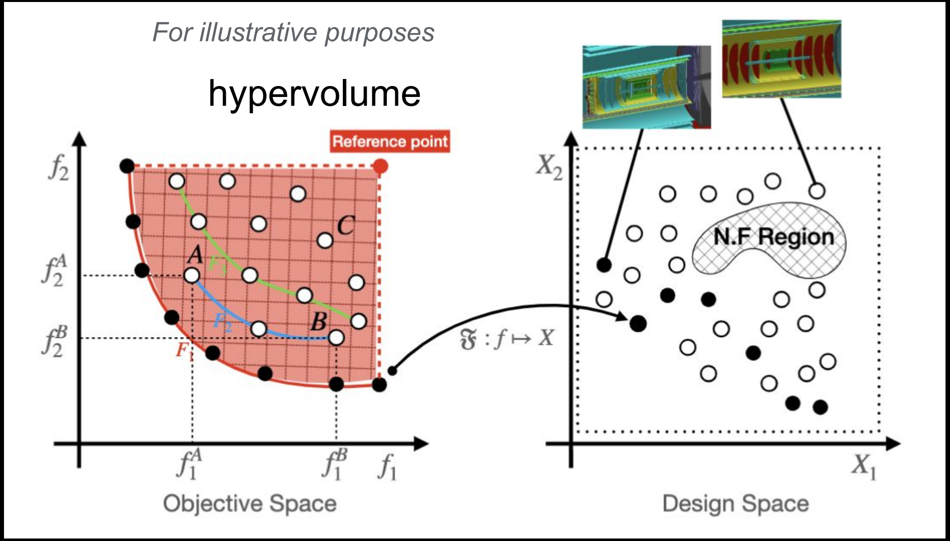
- Current simulation campaigns produce up to 15-20 TB / month (T. Britton, Oct 2024)
- Towards a quantitative computing model (M. Diefenthaler, Sep 2024)

Simulating 5M charged particles for the tracker and PID system would require at least 15k CPU core hours. This requirement can grow significantly when accounting for additional particle types or extending the scope to include neutrals to design other sub-detector systems.

Multi-Objective Optimization

MOO is 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

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

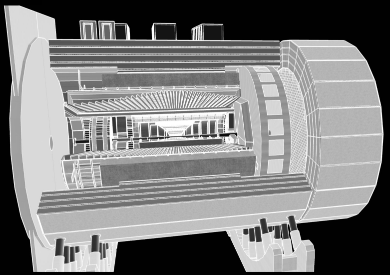
objectives

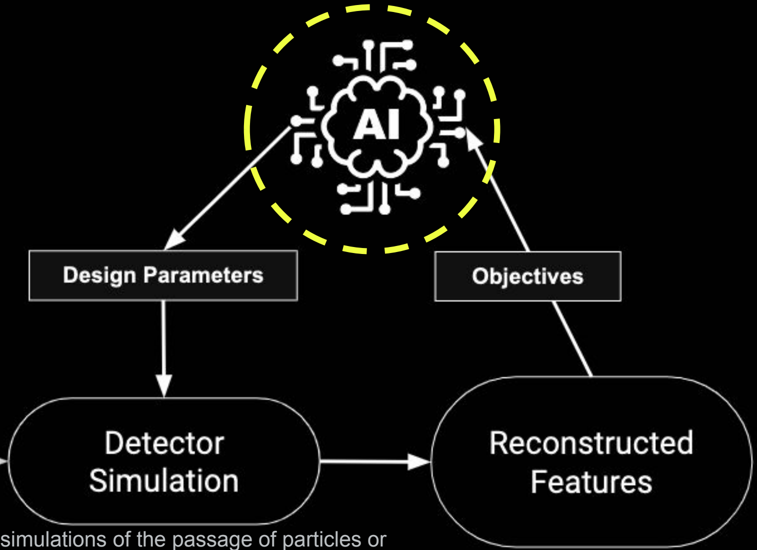
constraints

ranges

AI-Assisted Detector Design at EIC

The AI-assisted design embraces all the main steps of the sim/reco/analysis pipeline…





Accurate simulations of the passage of particles or radiation through matter

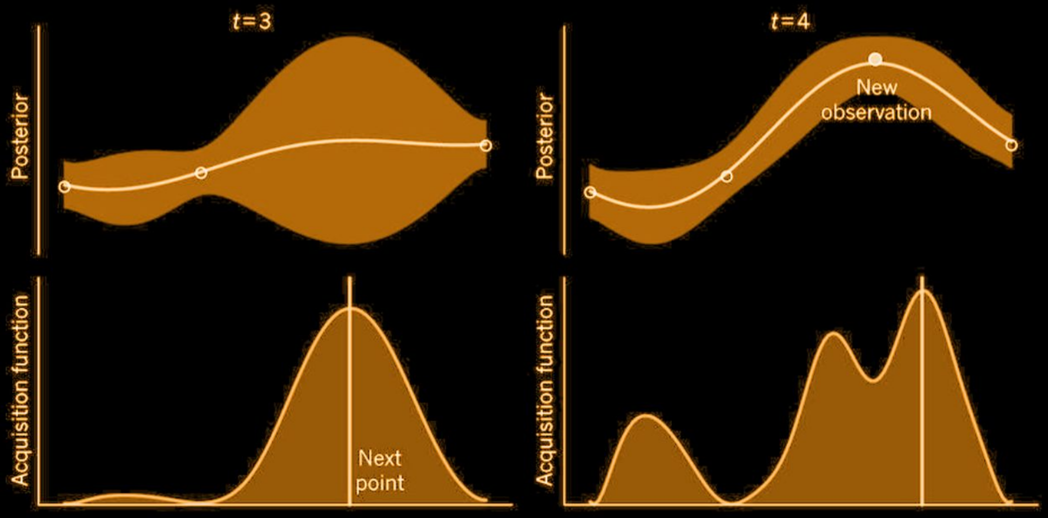
- Benefits from rapid turnaround time from simulations to analysis of high-level reconstructed observables
- The ePIC SW stack offers multiple features that facilitate AI-assisted design (e.g., modularity of simulation, reconstruction, analysis, easy access to design parameters, automated checks, etc.)
- Leverages heterogeneous computing
- AI-assisted design started being used since proto-collaboration phase (NIM-A 1047 (2023): 167748)

Provide a framework for an holistic optimization of the sub-detector system
A complex problem with (i) multiple design parameters, driven by (ii) multiple objectives (e.g., detector response, physics-driven, costs) subject to (iii) constraints

Those at EIC can be the first large-scale experiments ever realized with the assistance of AI

Bayesian Optimization in a nutshell

- BO is a sequential strategy developed for global optimization.
- After gathering evaluations we builds a posterior distribution used to construct an **acquisition function**.
- This cheap function determines what is **next query point**.

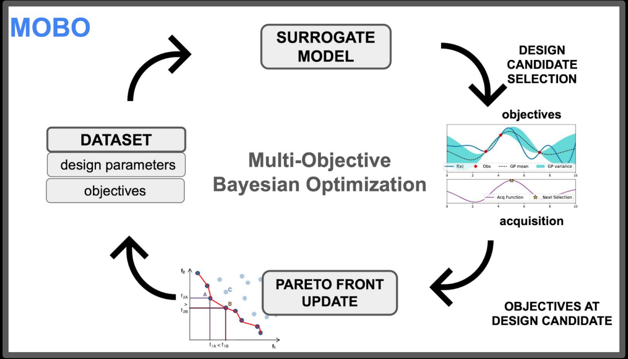


1. Select a Sample by Optimizing the Acquisition Function.
2. Evaluate the Sample With the Objective Function.
3. Update the Data and, in turn, the Surrogate Function.
4. Go To 1.

Contributions

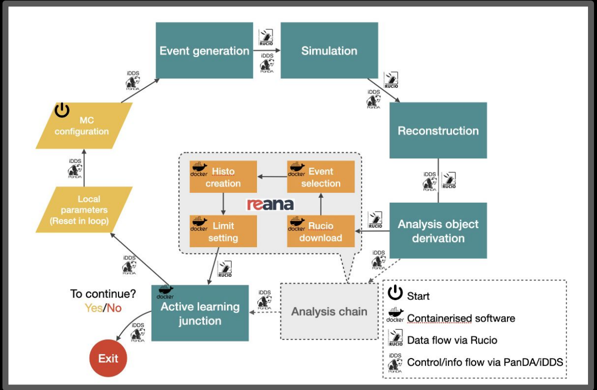
The AID2E Collaboration, "AI-assisted detector design for the EIC", 2024 JINST 19 C07001

Multi-Objective Optimization



(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

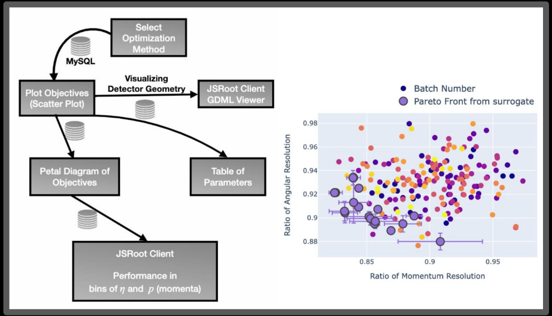
Distribution and Workload Management



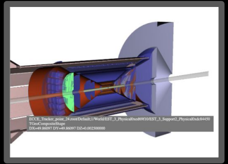
(ii) Will leverage cutting-edge workload management systems capable of operating at massive data and handle complex workflows

https://github.com/aid2e

Human-in-the loop: interactive Pareto navigation



(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.



https://ai4eicdetopt.pythonanywhere.com/

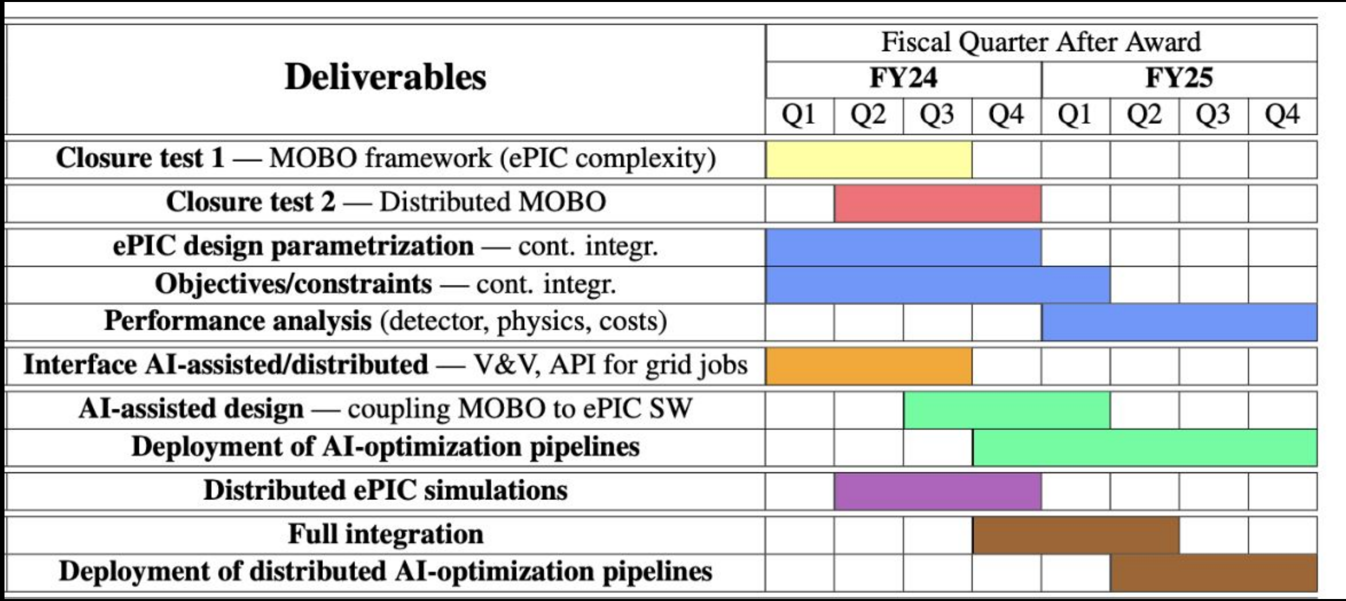
https://wandb.ai/scheduler/AID2E-Closure-1

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

Benefits of AID2E

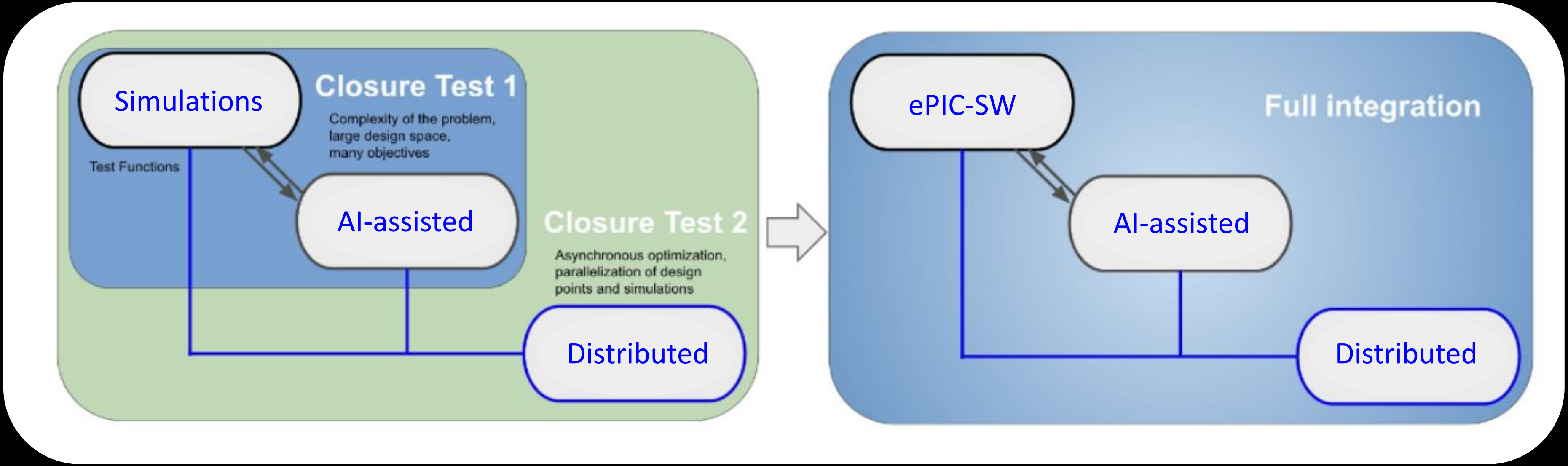
- Integrating holistic, multi-objective optimization into detector design marks a significant paradigm shift, with AI-assisted methods poised to profoundly impact large-scale NP projects such as at the EIC.
  - AI provides quantifiable insights into complex design and objective spaces, enabling a comprehensive evaluation of various tradeoff design solutions.
- A fractional improvement in the objectives translates to a more efficient use of beam time which will make up a majority of the cost of the EIC over its lifetime.
- Examining solutions on the Pareto front of EIC detectors at different values of the budget can have great cost benefits.
- Implementing AI-assisted, multi-objective optimization accelerates the design process, quantifies trade-offs between design points, and produces designs that optimize both performance and cost. This approach will also be valuable during construction, accommodating new constraints as they arise.
- Possibility of extending this framework to other computational intensive tasks such as calibrations and alignment of detectors.

Deliverables and Staffing
(estimated at start of project)



- All deliverables for FY24 met.
- In the following I will provide more details on the accomplished work.

AID2E Closure Tests and Workflow

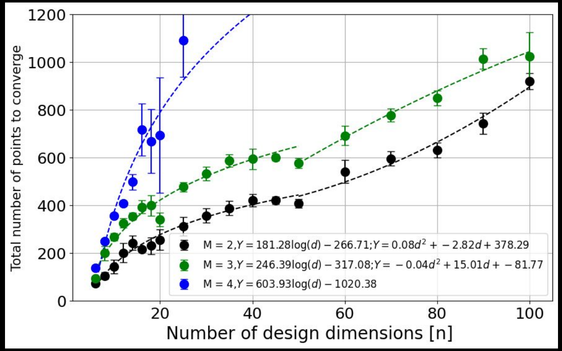


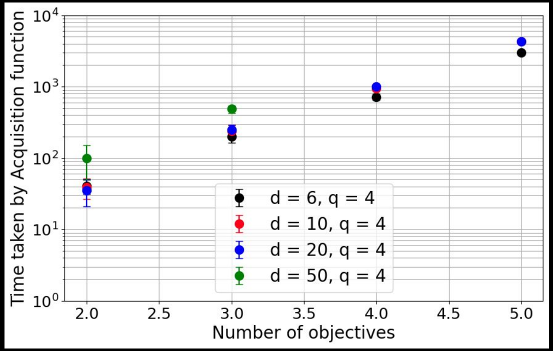
(Rationale)

✅ Closure Test 1

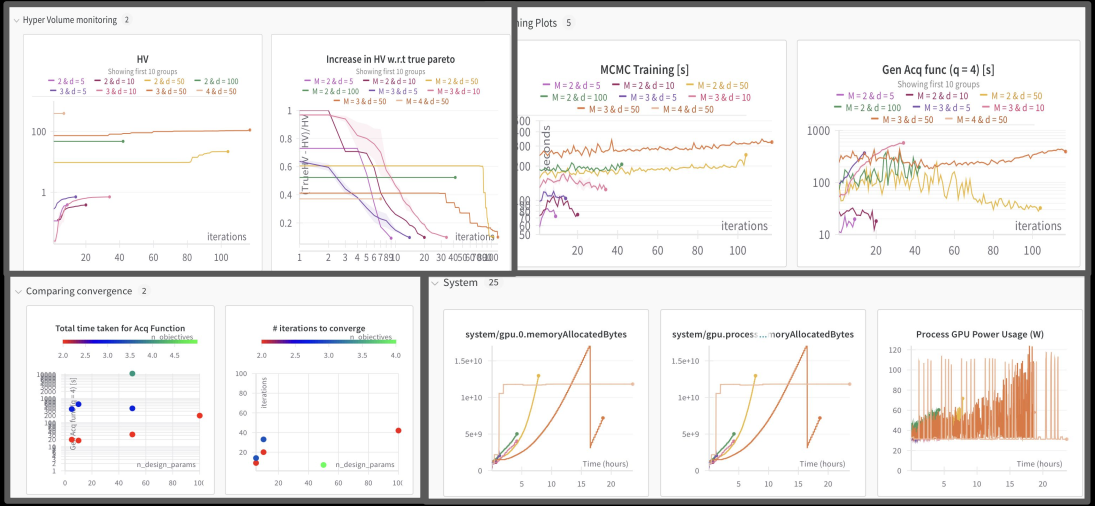
Goals:

- Utilize test functions to evaluate proximity to known Pareto front
  - Accuracy of optimization, convergence properties, compute resources
- Characterization of Complexity
  - Stress-testing for problems with increased complexity





[W&B dashboard for monitoring]



Shown here:

- Test function: [DTLZ]
- Technique: [MOBO]

✅ Closure Test 1: MOBO Complexity

n: number of design points
d: design dimensionality (each point)
M: objectives

Gaussian Process
$O(n^3)$

- Surrogate model.
- SAAS[1] priors have been proven to be successful up to 388 design dimensions
- Assumes several design variables has increased importance compared to others
- Computational expensive as iteration increases
- Benefit from GPU hardware acceleration

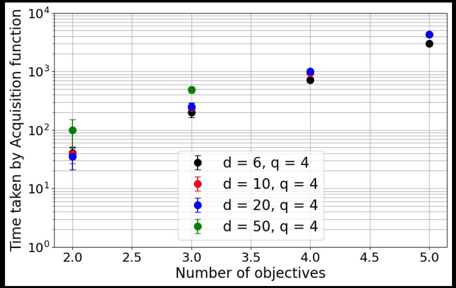
Bayesian Sampling from posteriors
NUTS – $O(Md^{5/4})$ [NUTS]

- Sample L points from the posterior distribution
- HMC is a popular algorithm, NUTS is a variant
- Mainly depends on the number of objectives and design space dimensions
- Has minimal dependence on iteration.
- GPU acceleration through JAX backend.

Acquisition function
qNEHVI – $O(Md(n+i)^M)$ [2]

- Captures HV improvement
- A "cheaper" function to evaluate as a proxy for the black box function
- Scales nonlinearly with iteration, total points explored, design space and objective space.
- Partially benefitted by GPU acceleration.

Complexity Studies



- Benefitting from GPU acceleration
- With sufficient parallelization, if possible, the time associated to the MOBO part at some point becomes dominant (bottom plot shown at 15th iteration with number of points between ~70-160)

q: batch size

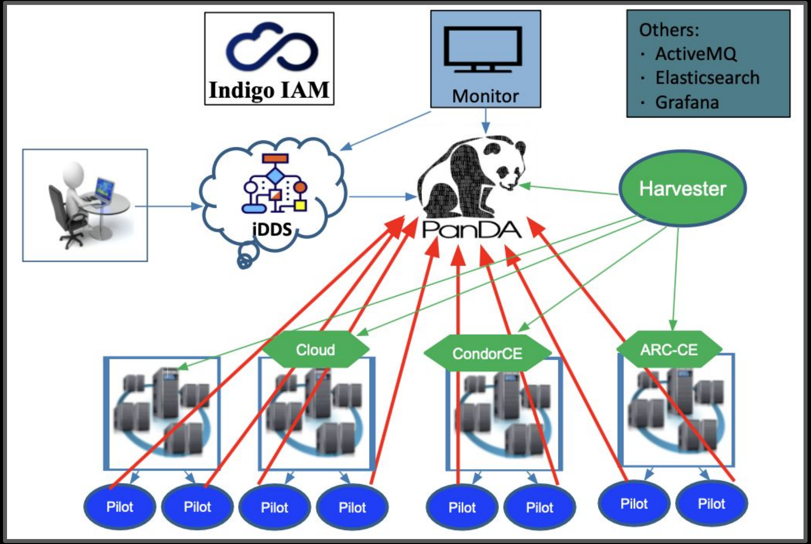
✅ Closure Test 2: PaNDA/iDDS

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 Machine Learning with PanDA and iDDS in ATLAS

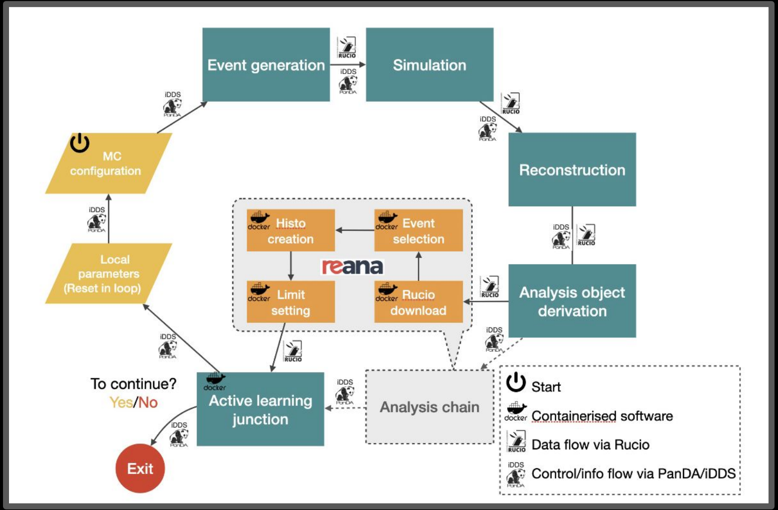
PanDA: Production and Distributed Analysis System. *Comput Softw Big Sci* **8**, 4 (2024)

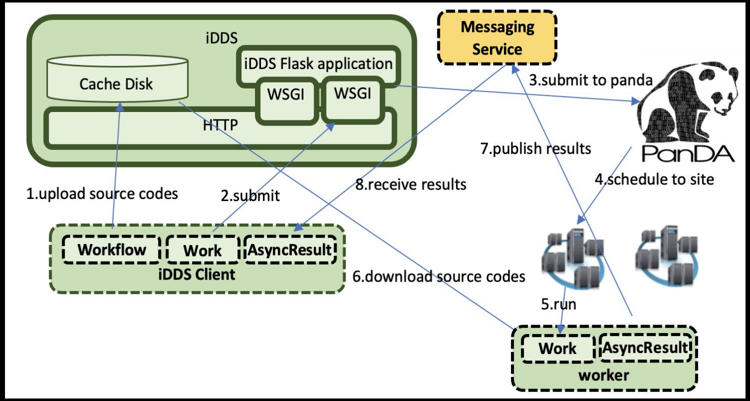
✅ Closure Test 2: PaNDA/iDDS

Bayesian optimisation based active learning with Panda/iDDS

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





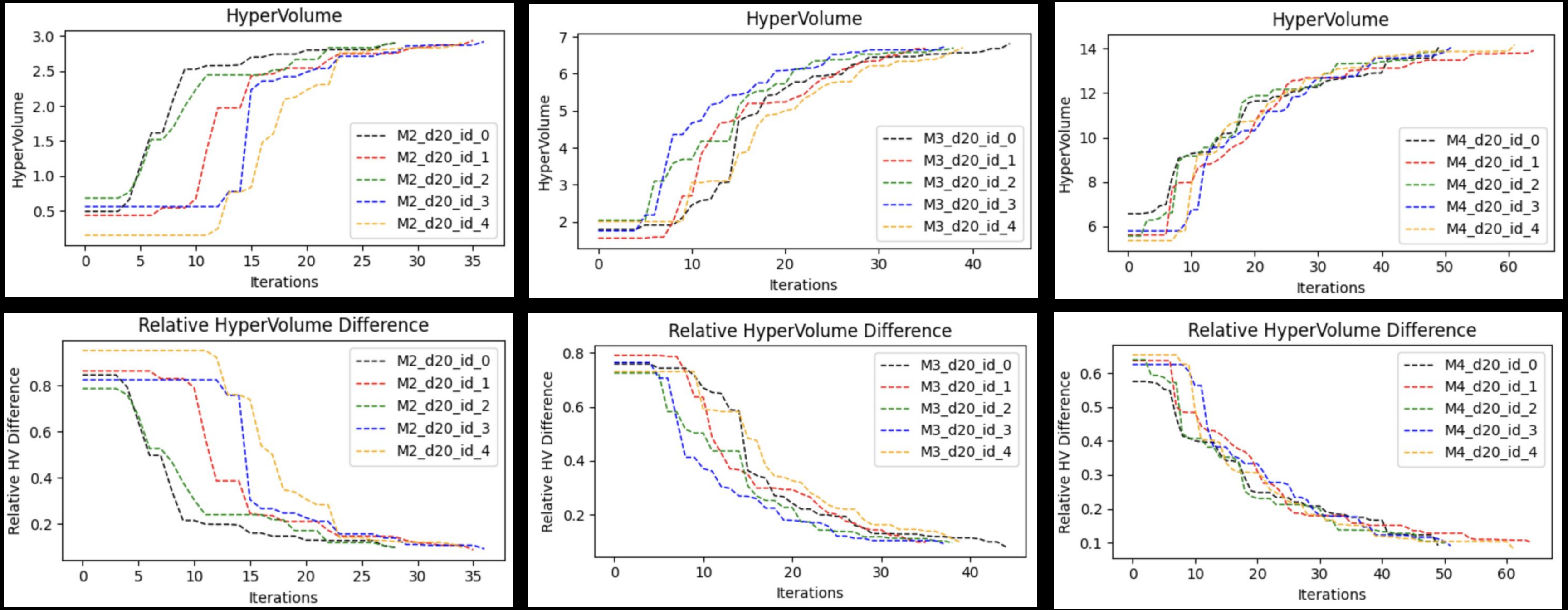
Schema of how a workflow executes a function at remote distributed resources

Closure Test 2:

obtaining convergence on Pareto fronts using test functions and distributed computing

✅ Closure Test 2: Results

Examples of optimization pipelines run using PanDA

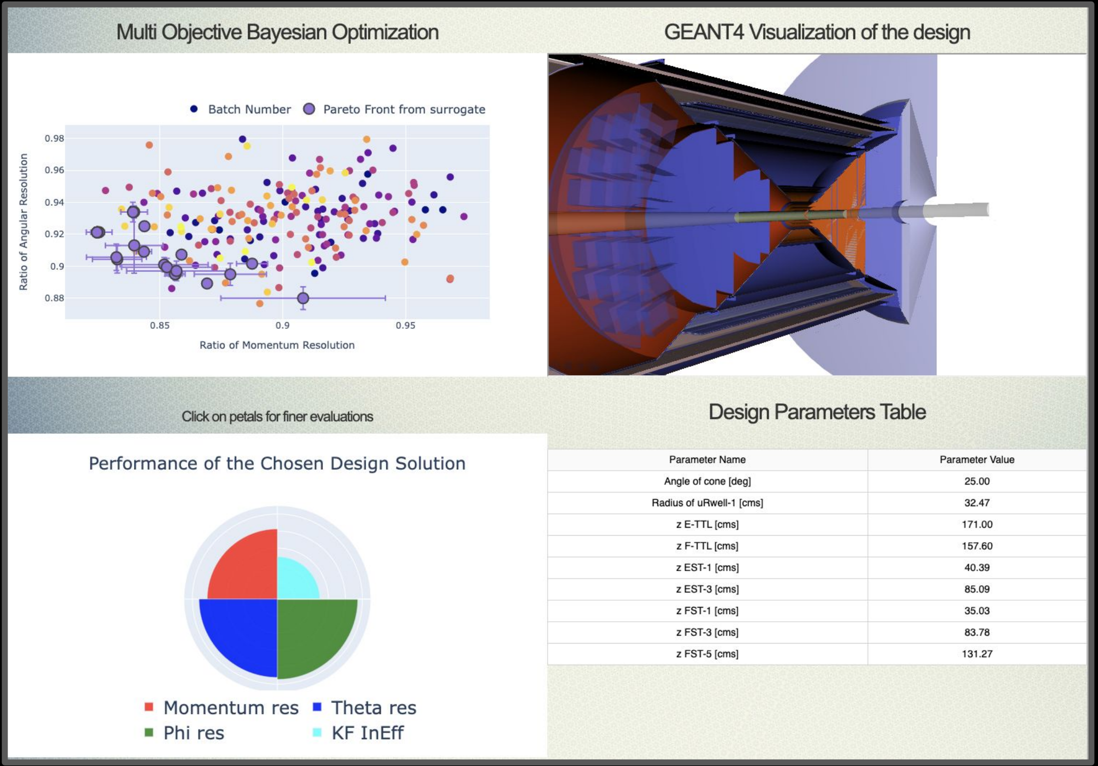


- Relative HyperVolume Difference = (HVol_pareto - HVol)/HVol_pareto
  When reaches the tolerance (0.1 here), stops the training
- HVol_pareto: HyperVolume of Pareto Front

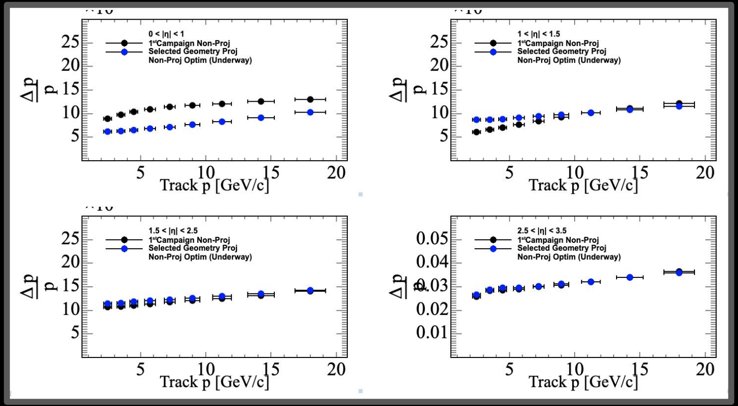
Recent tests covered:
- d=50 and M=3

✅ Interactive Pareto Front Navigation

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

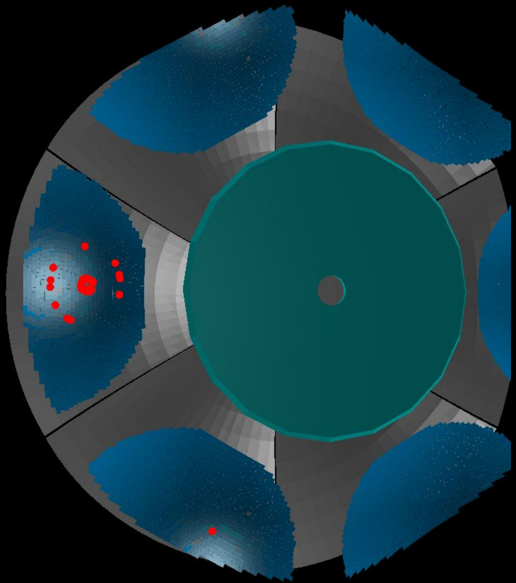


Parameter Name	Parameter Value
Angle of cone [deg]	25.00
Radius of uRwell-1 [cms]	32.47
z E-TTL [cms]	171.00
z F-TTL [cms]	157.60
z EST-1 [cms]	40.39
z EST-3 [cms]	85.09
z FST-1 [cms]	35.03
z FST-3 [cms]	83.78
z FST-5 [cms]	131.27



✅ Integration with ePIC SW: ePIC dRICH

Considering all the constraints as ePIC is in the process of finalizing engineering designs, we can select those sub-detectors that still have tunable parameters

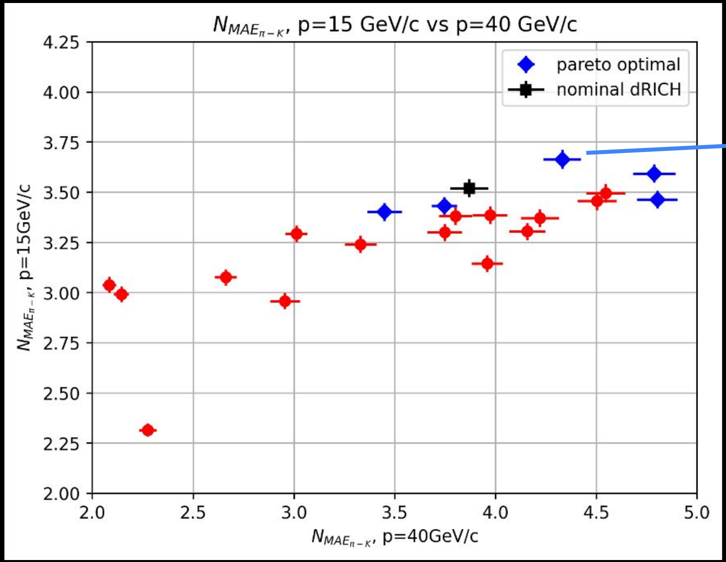


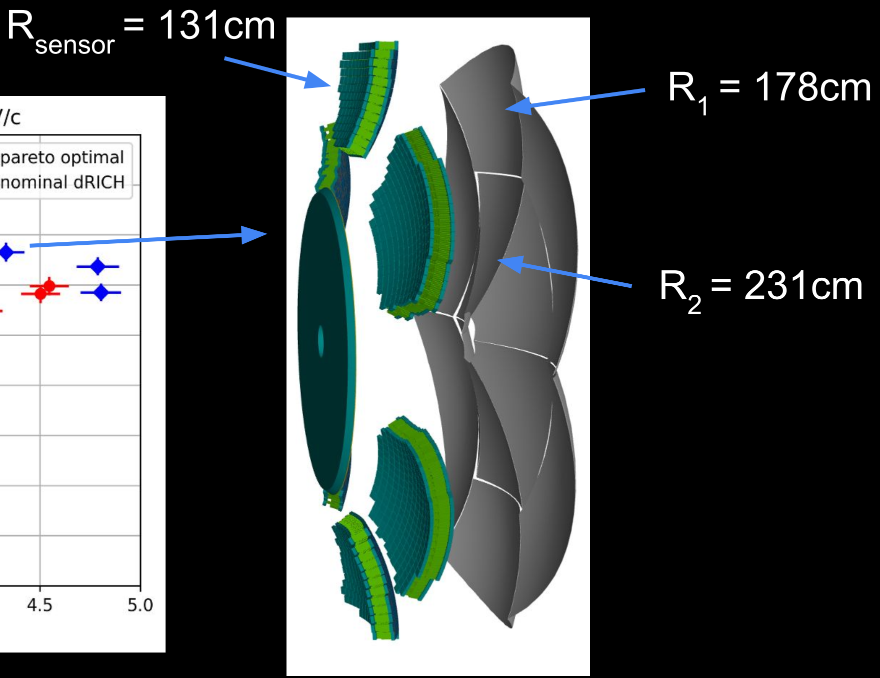
dual-RICH

- *dual-RICH: two radiators for wide momentum coverage (~ 3GeV/c - 60GeV/c), $1.5 < \eta < 3.5$*
- *Simultaneously focus all η regions, gas and aerogel rings*
- *Mirror, sensor placement and radii, gas, mirror material (lower cost material)...*
- *PID performance, costs, …*

E. Cisbani *et al* 2020 *JINST* **15** P05009

✅ Integration with ePIC SW: ePIC dRICH

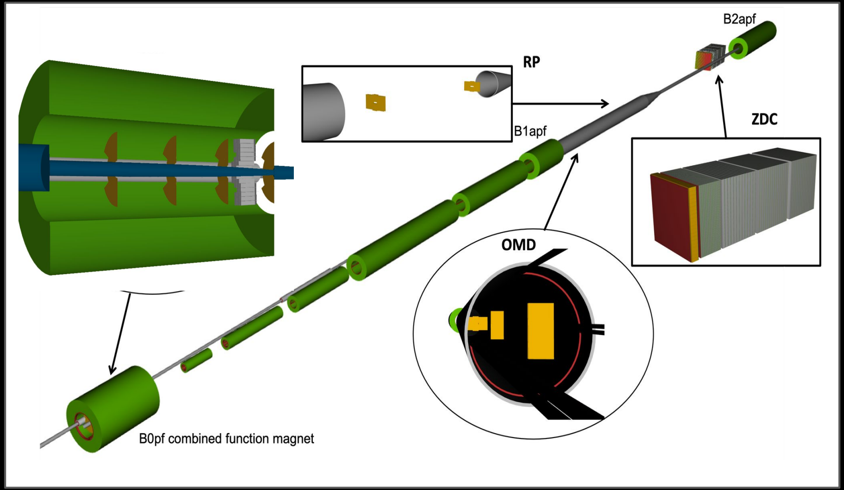




- Three-objective optimization: π-K separation at 15 GeV/c, 40 GeV/c, and fraction of tracks with > 5 photons
- Work ongoing with ePIC to refine two-mirror reconstruction algorithm, finalize optimization

✅ Integration with ePIC SW: Far-Forward

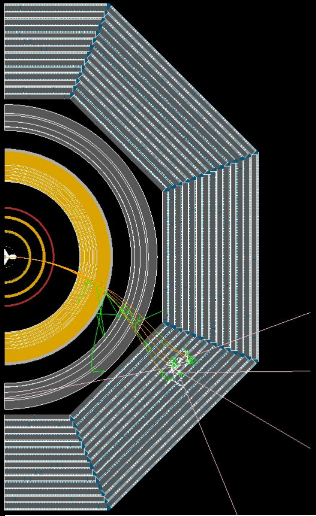
Far-Forward detectors

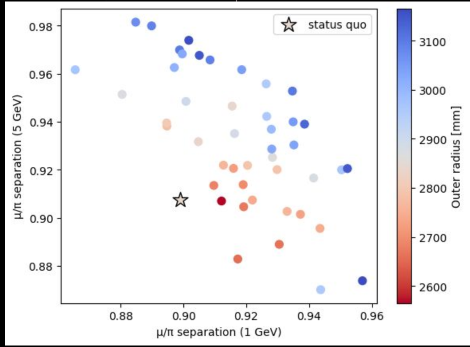


- *B0 subsystem*
  - *Tracker (AC-LGAD)*
  - *Crystal Calorimeter*
  - *Proton tagging critical for Forward Physics*
- *Magnetic Field is inhomogeneous & Mechanical constraints restrict detector real estate (entire length of B0 fixed)*
- *Tracking layers, ECAL crystals & tiling of crystals*
- *Objectives: Tracking resolution and detector acceptance.*

Ongoing discussion with the ePIC working group to consolidate optimization

Spin-off: Detector2 μId/HCAL





- Physics Motivation: Muon channels (J/Psi DDVCS), cost effective HCAL
- Iron/Scintillator sandwich integrated in flux return
  - Longitudinal segmentation for better h/μ ID, energy reconstruction with ML
  - R&D on fast scintillator (readout) ($\mathcal{O}(50ps)$) for ToF ongoing
  - Possible solution for endcap HCAL of ePIC
- **Pilot project:**
  - Optimize μID performance @1 & 5 GeV
  - Parameters: number of layers, thickness of passive iron layers
- **Activities to pursue in the future:**
  - Optimize μID and ToF/$\boldsymbol{\sigma}_E$ concurrently →competing requirements
    Active/passive detector ratio
  - Explore complex configurations (e.g. nonlinear layer thickness), parameters
  - Holistic optimization of magnet/detector geometry
    →explore physics impact and complementarity to project detector

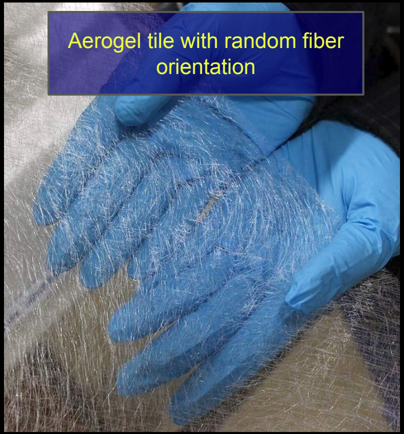
Spin-off: Material Design

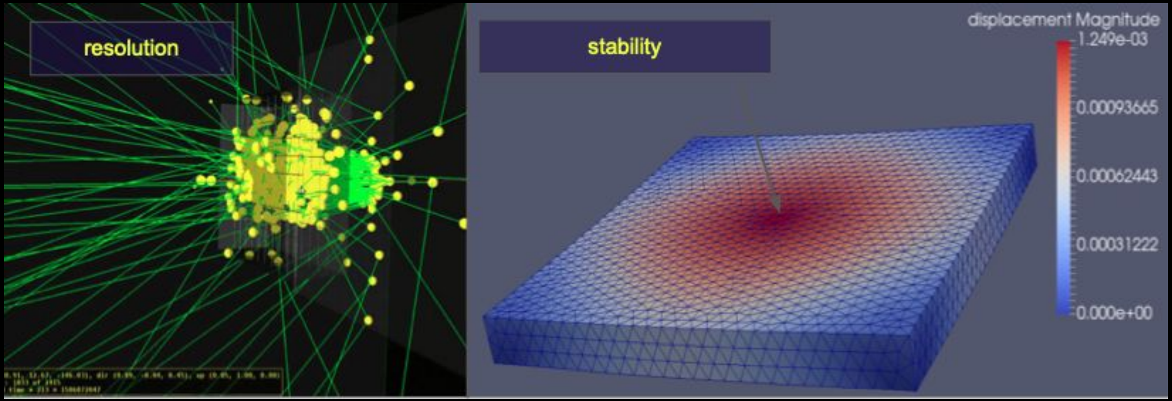
Reinforced novel aerogel material with fibers

Software Stack

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

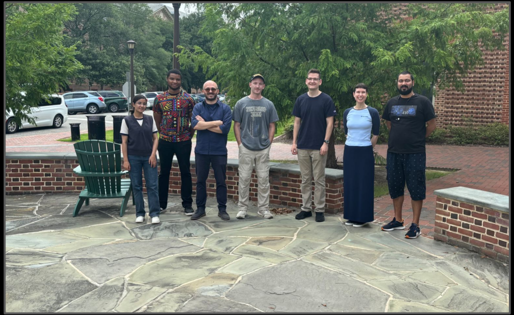


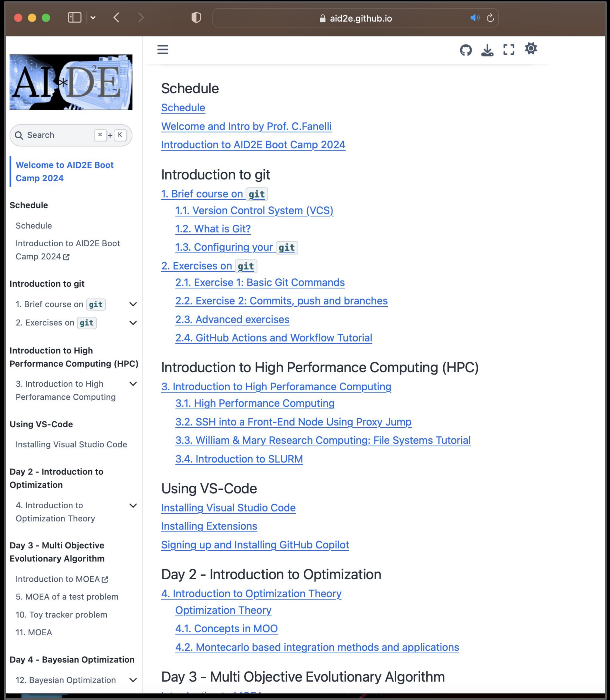


Publication in preparation

Documentation and Outreach

- GitBook and/or other knowledge sharing platforms will be part of the initiatives related to documentation and outreach
- Offering opportunities for experiential learning with easy access for beginners
  - 1 week summer bootcamp
  - Final Projects
- The first AID2E bootcamp took place in July 2024
  - https://aid2e.github.io/boot-camp-2024/intro.html





Conclusions

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

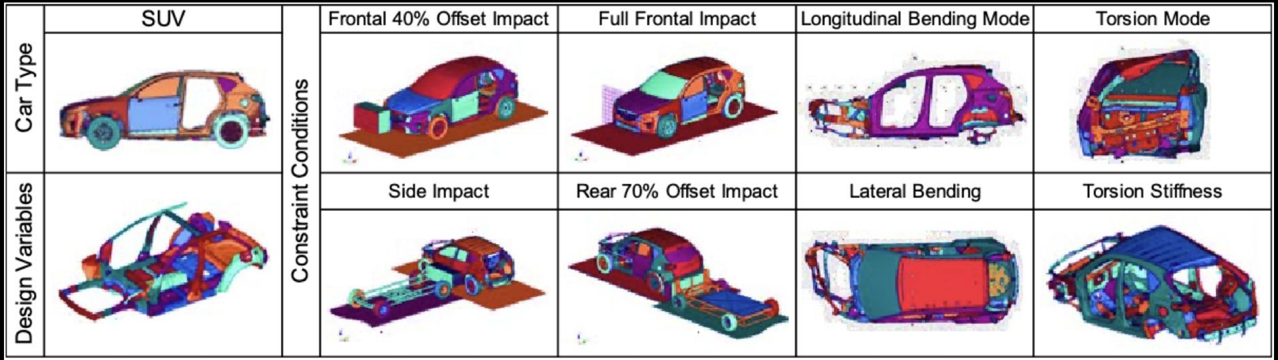
- We completed coupling of different MOO techniques to the ePIC SW, closure tests and other FY24 deliverables that demonstrated effective distribution and expected scalability.
- We are on track to deliver a framework that can optimize holistically a large-scale detector:
  - Shown ongoing activities with detector subsystems (dRICH, far-forward) in ePIC.
  - The ePIC design is in progress with CD-2 not before end of 2025.
  - Optimization studies of Detector II subsystems and magnet could provide valuable insights on complementarity with ePIC.
- In detector projects, most changes happen during the construction phases (e.g., changes in the available material or budget). AID2E will be an ideal tool to optimize design changes with objectives (e.g., reduce cost).
- This framework inherently offers broader impacts, as it can be adapted for use in various experiments and is suitable for a wide range of compute-intensive applications that necessitate MOO (e.g., calibrations, alignments, novel material design, etc)

Spares

Multi-Objective Optimization: Example

Vehicle Design Optimization:

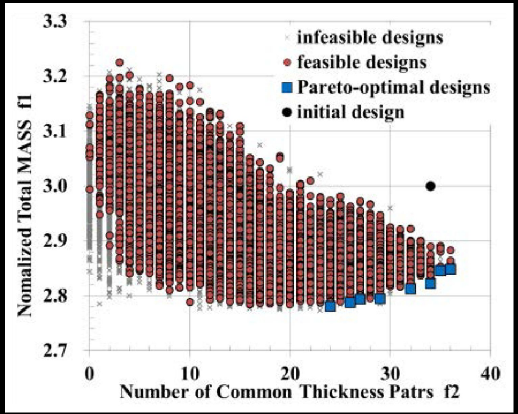
222 parameters + 2 objectives + 54 constraints



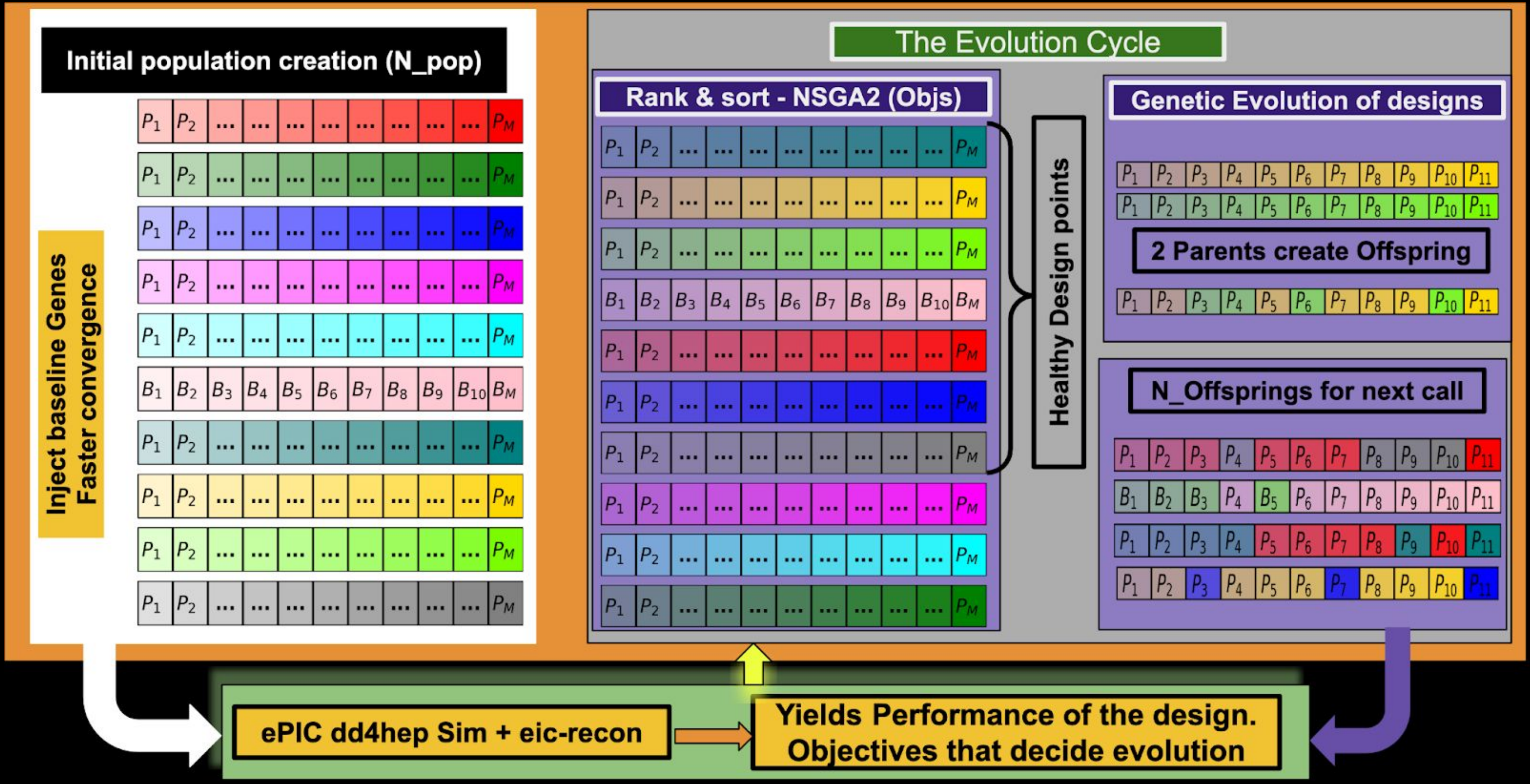
T. Kohira et al. Proposal of benchmark problem based on real-world car structure design optimization. GECCO 2018

Objectives:

- minimize total vehicle mass of three vehicles (Mazda 3, 6, CX-5)
- maximize number of parts shared across vehicles

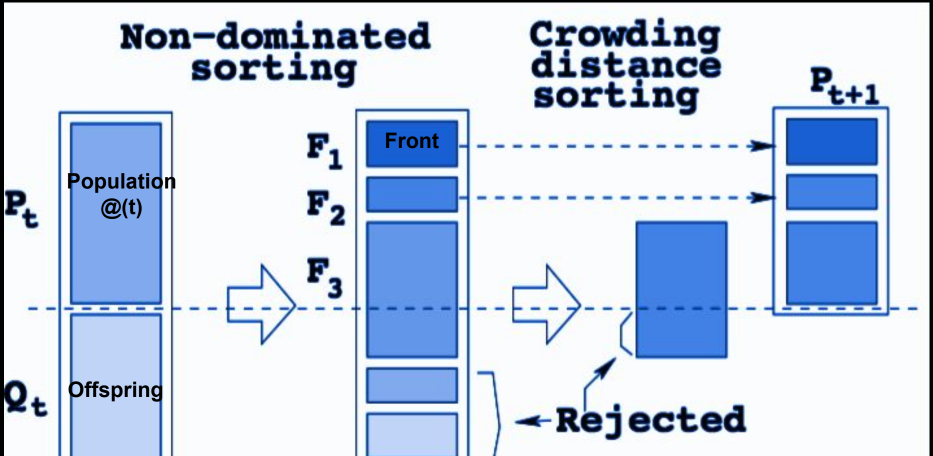


MOGA Pipeline

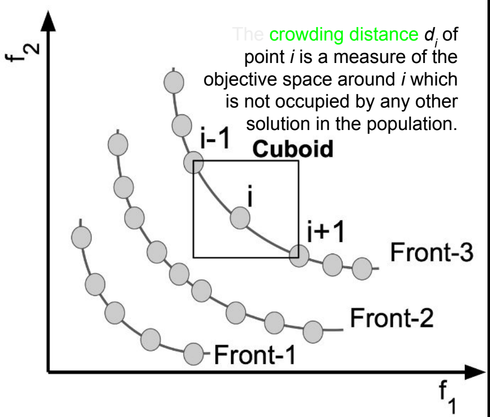


Non-Dominated Sorting Genetic Algorithm

The population R_t is classified in non-dominated fronts.
Not all fronts can be accommodated in the N slots of available in the new population P_{t+1}. We use **crowding distance** to keep those points in the last front that contribute to the highest diversity.

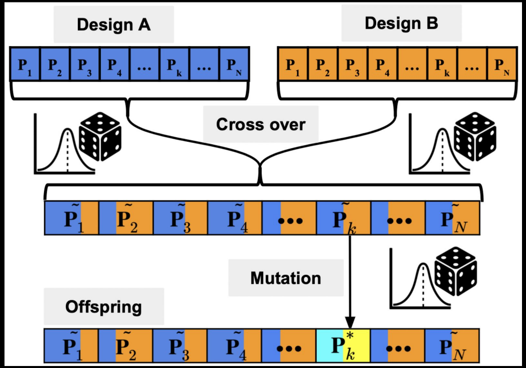


[1] Deb, K., et al. "A fast and elitist multiobjective genetic algorithm" *IEEE transactions on evolutionary computation* 6.2 (2002): 182-197



This is one of the most popular approach (>35k citations on google scholar), characterized by:

- Use of an elitist principle
- Explicit diversity preserving mechanism
- Emphasis in non-dominated solutions



This is to illustrate Binary Crossover