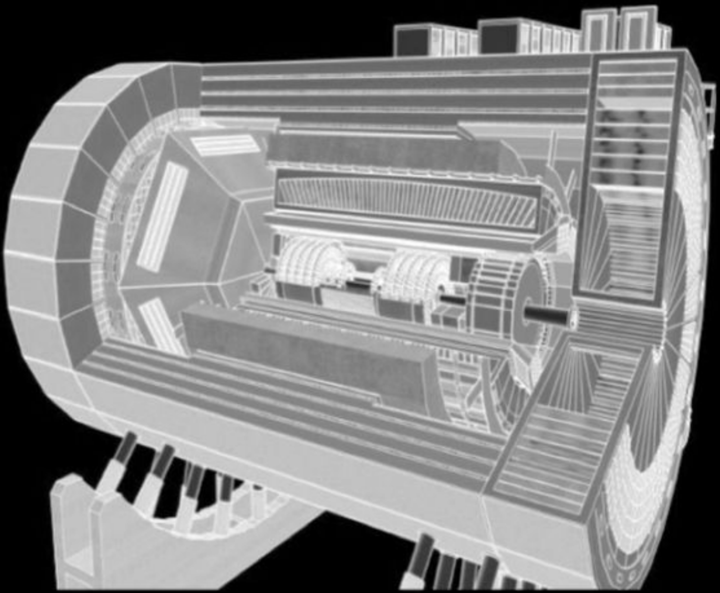
AI/ML for Detector Optimization



BNL, HEP SW Foundation, JLab
Software & Computing Round Table
November 9 2021

Cristiano Fanelli
cfanelli@mit.edu

AI for Design

Relatively new but active area of research.

Many applications in, e.g., industrial material, molecular and drug design.

Guo, Kai, et al. *Materials Horizons* 8.4 (2021): 1153-1172.

Table 1 Popular ML methods in design of mechanical materials

ML method	Characteristics	Example applications in mechanical materials design
Linear regression; polynomial regression	Model the linear or polynomial relationship between input and output variables	Modulus[112] or strength[123] prediction
Support vector machine; SVR	Separate high-dimensional data space with one or a set of hyperplanes	Strength[123] or hardness[125] prediction; structural topology optimization[159]
Random forest	Construct multiple decision trees for classification or prediction	Modulus[112] or toughness[130] prediction
Feedforward neural network (FFNN); MLP	Connect nodes (neurons) with information flowing in one direction	Prediction of modulus,[97,112] strength,[93] toughness[130] or hardness;[97] prediction of hyperelastic or plastic behaviors;[143,145] identification of collision load conditions;[147] design of spinodoid metamaterials[163]
CNNs	Capture features at different hierarchical levels by calculating convolutions; operate on pixel-based or voxel-based data	Prediction of strain fields[104,105] or elastic properties[102,103] of high-contrast composites, modulus of unidirectional composites,[136] stress fields in cantilevered structures,[137] or yield strength of additive-manufactured metals;[121] prediction of fatigue crack propagation in polycrystalline alloys;[140] prediction of crystal plasticity;[120] design of tessellate composites;[107–109] design of stretchable graphene kirigami;[155] structural topology optimization[156–158]
Recurrent neural network (RNN); LSTM; GRU	Connect nodes (neurons) forming a directed graph with history information stored in hidden states; operate on sequential data	Prediction of fracture patterns in crystalline solids;[114] prediction of plastic behaviors in heterogeneous materials;[142,144] multi-scale modeling of porous media[173]
Generative adversarial networks (GANs)	Train two opponent neural networks to generate and discriminate separately until the two networks reach equilibrium; generate new data according to the distribution of training set	Prediction of modulus distribution by solving inverse elasticity problems;[138] prediction of strain or stress fields in composites;[139] composite design;[164] structural topology optimization;[165–167] architected materials design[115]
Gaussian process regression (GPR); Bayesian learning	Treat parameters as random variables and calculate the probability distribution of these variables; quantify the uncertainty of model predictions	Modulus[122] or strength[123,124] prediction; design of supercompressible and recoverable metamaterials[110]
Active learning	Interacts with a user on the fly for labeling new data; augment training data with post-hoc experiments or simulations	Strength prediction[124]
Genetic or evolutionary algorithms	Mimic evolutionary rules for optimizing objective function	Hardness prediction;[126] designs of active materials;[160,161] design of modular metamaterials[162]
Reinforcement learning	Maximize cumulative awards with agents reacting to the environments.	Deriving microstructure-based traction-separation laws[174]
Graph neural networks (GNNs)	Operate on non-Euclidean data structures; applicable tasks include link prediction, node classification and graph classification	Hardness prediction;[127] architected materials design[168]

Z. Zhou et al., *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019

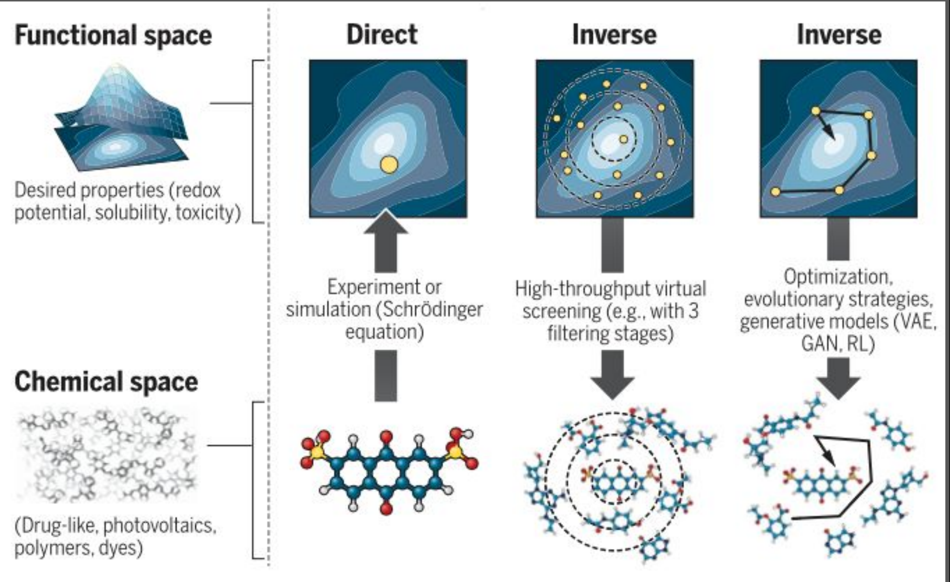


Fig. 2. Schematic of the different approaches toward molecular design. Inverse design starts from desired properties and ends in chemical space, unlike the direct approach that leads from chemical space to the properties.

B. Sanchez-Lengeling, A. Aspuru-Guzik. *Science* 361.6400 (2018): 360-365.

Usage of ML/DL typically requires large datasets (design points)…

AI for Experimental Design in NP/HEP

[1] Bellman, Richard. *Dynamic programming*. Vol. 295. RAND CORP SANTA MONICA CA, 1956.
[2] CF et al. *JINST* 15.05 (2020): P05009.
[3] Wolpert, D.H., Macready, W.G., 1997. Trans. Evol. Comp 1, 67–82

- Detector design needs advanced simulations which are computationally expensive (Geant). Deal with non-differentiable terms, noisy functions. We want to reduce computing budget.
- In general the full detector design is studied once the subsystem prototypes are ready (phase constraints from the full detector or outer layers are taken into consideration). This is what is typically done without AI.
- When it comes to designing detectors with AI this is an area at its “infancy”.
- Complex detector designs typically entail:
  - Many parameters (and multiple objective functions): curse of dimensionality [1].
  - Incorporating a complex body of instructions [2].
  - Some degree of customization: the choice of a suitable algorithm is a challenge itself (no free lunch theorem [3]).

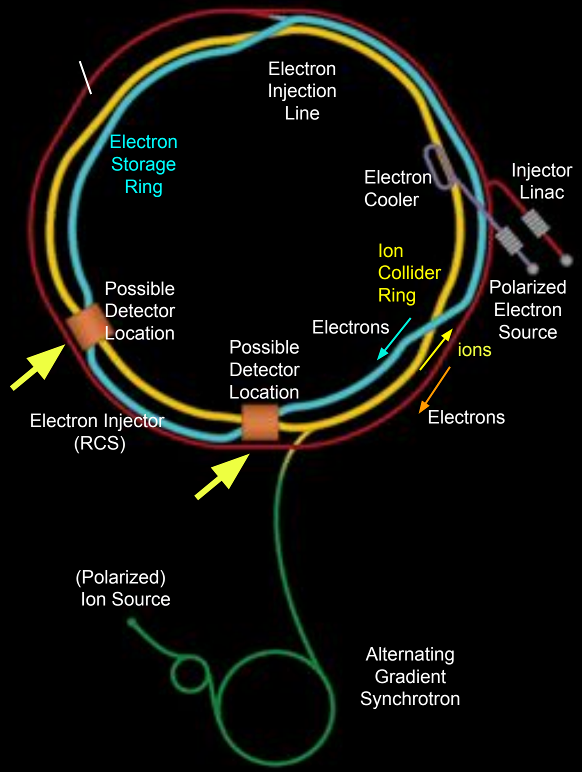
AI offers SOTA solutions to solve complex optimization problems in an efficient way.

Opportunity for new experiments in their design and R&D phase

Electron-Ion Collider

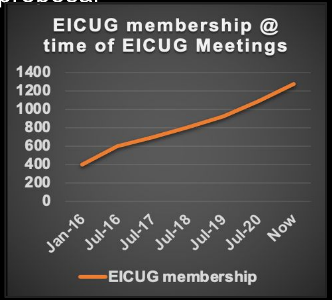
arXiv:1212.1701, arXiv:2103.05419

Will be built at Brookhaven National Laboratory in ~2030.
Use existing infrastructure of RHIC.

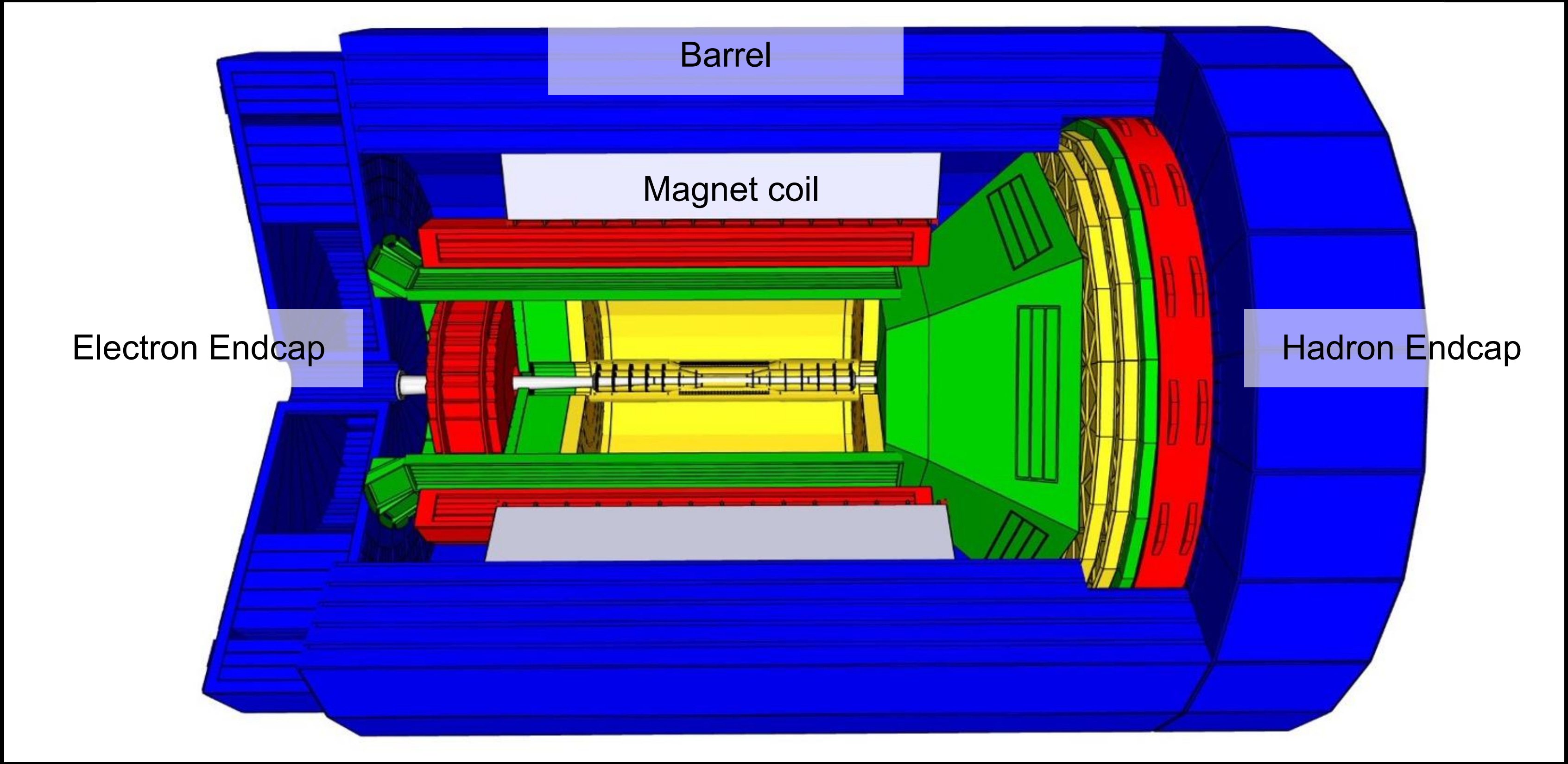


A machine for delving deeper than ever before into the building blocks of matter

- Physics Goals (findings from NAS Committee):
  - 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?
- The Machine will be capable to perform
  - High luminosity measurements **(10^{33} cm^{-2} s^{-1} - 10^{34} cm^{-2} s^{-1})**
  - Flexible center-of-mass energy range. $\sqrt{s} = \sqrt{4E_eE_p}$
  - Deliver highly polarised electron and proton/ light ion beams
  - Almost 4π hermetic detector
- Worldwide interest in the EIC
- Three proto-collaborations (ATHENA, CORE, ECCE) working for the detector proposal

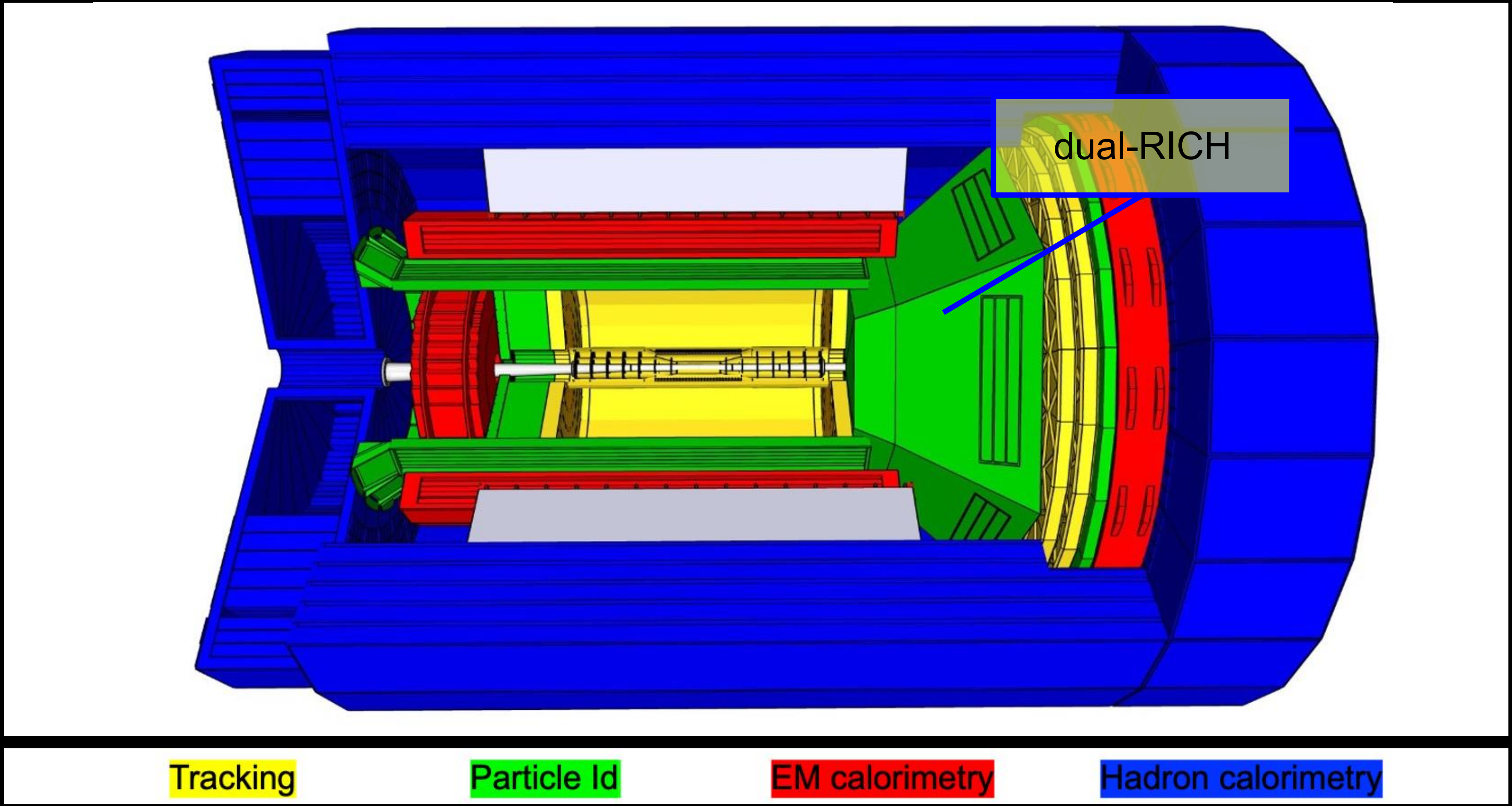


Overview of an example EIC Central Detector

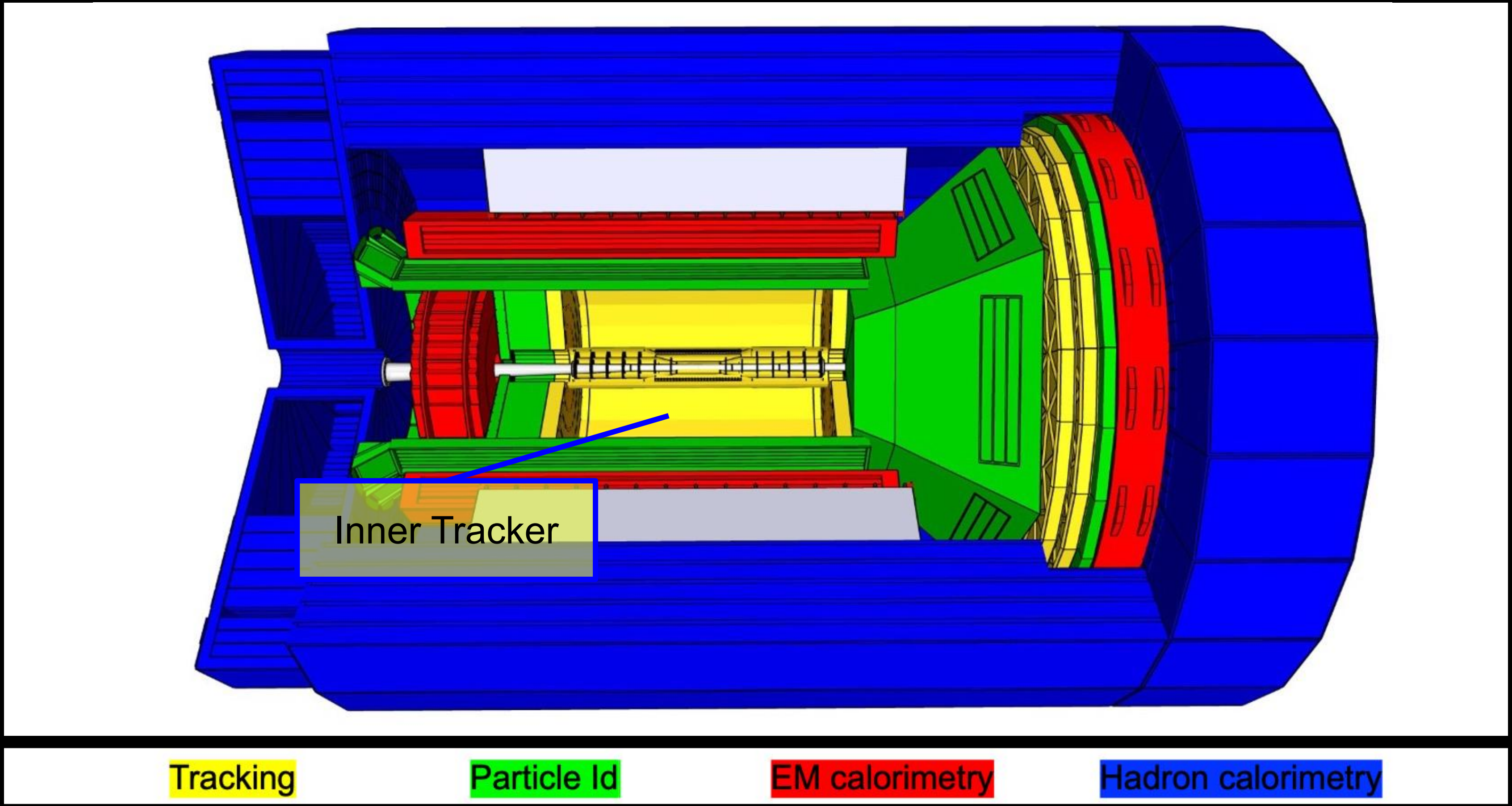


Tracking Particle Id EM calorimetry Hadron calorimetry

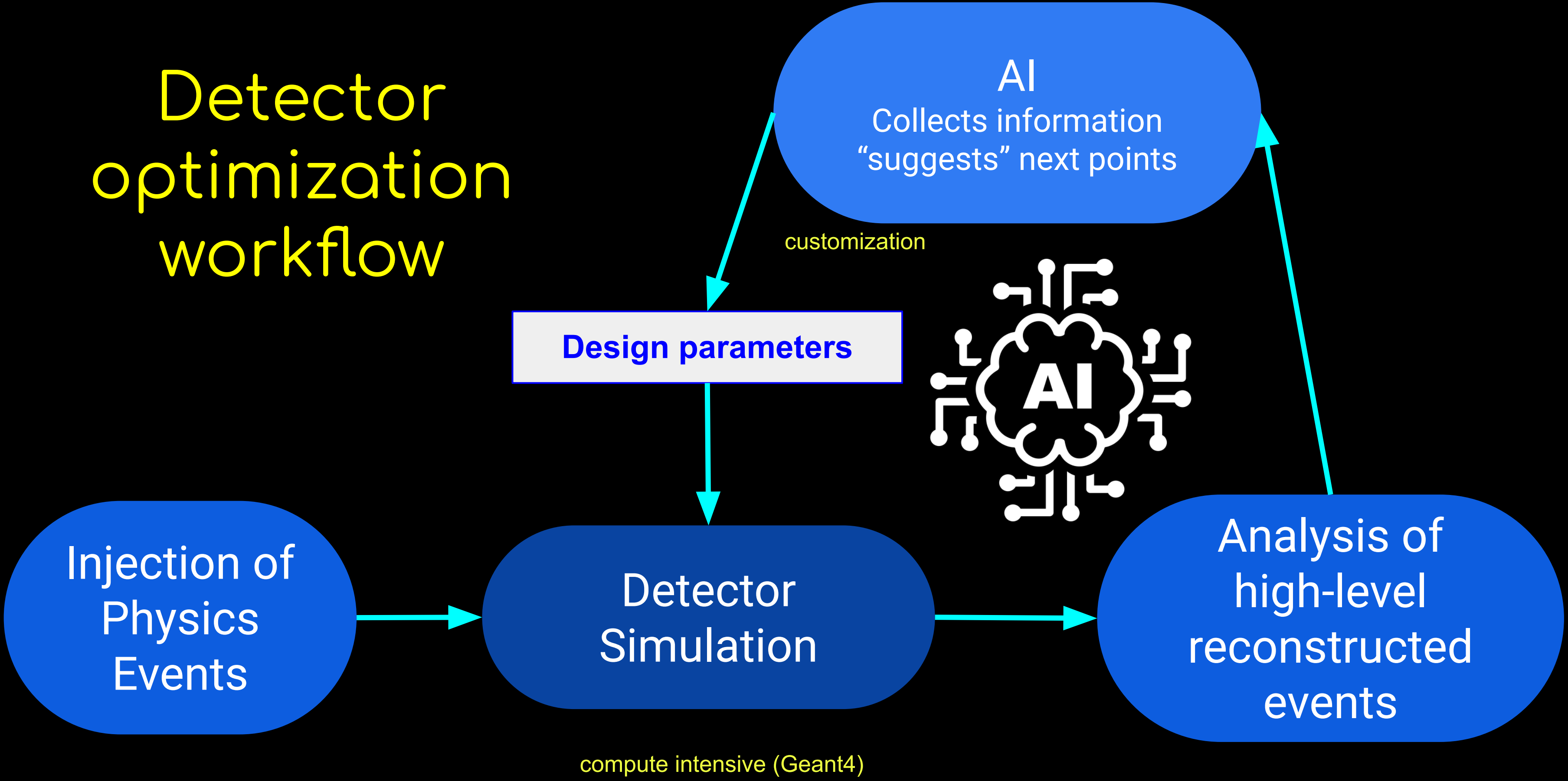
In what follows: 1st example



In what follows: 2nd example



Detector optimization workflow



Why Design with AI?

Optimization does not mean necessarily “fine-tuning”

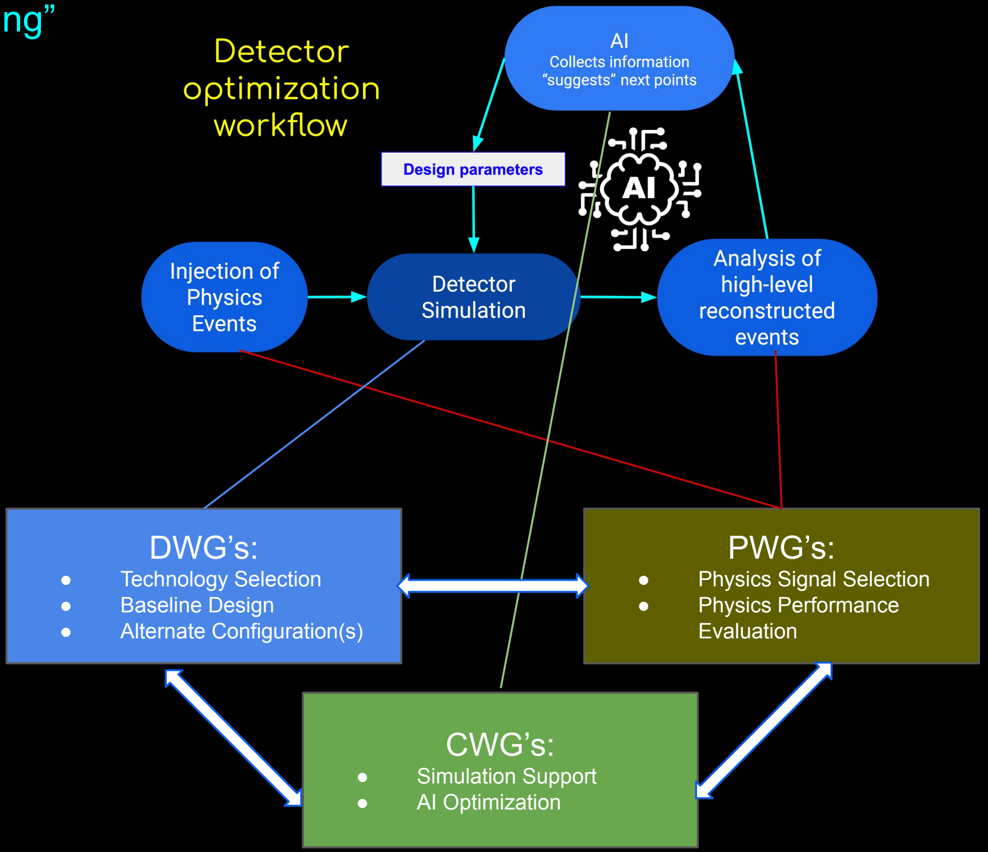
- We want to use these algorithms to:

  (1) steer the design and suggest parameters that a “manual” optimization will likely miss to identify;

  (2) further optimize

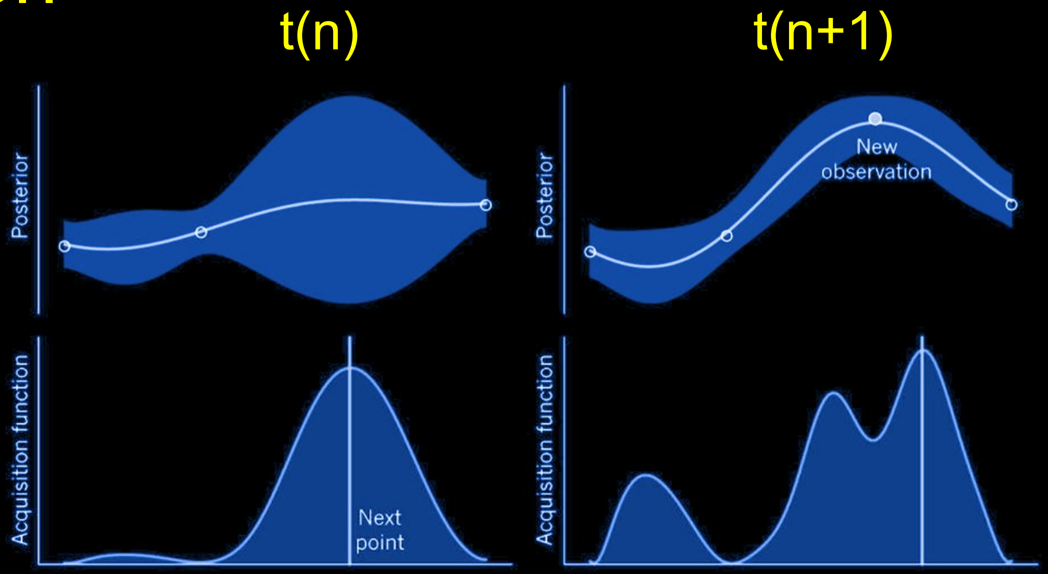
- AI allows to capture hidden correlations among the design parameters.
- All “steps” (physics, detector) involved in the AI optimization, strong interplay between working groups

AI promotes interaction among Working Groups



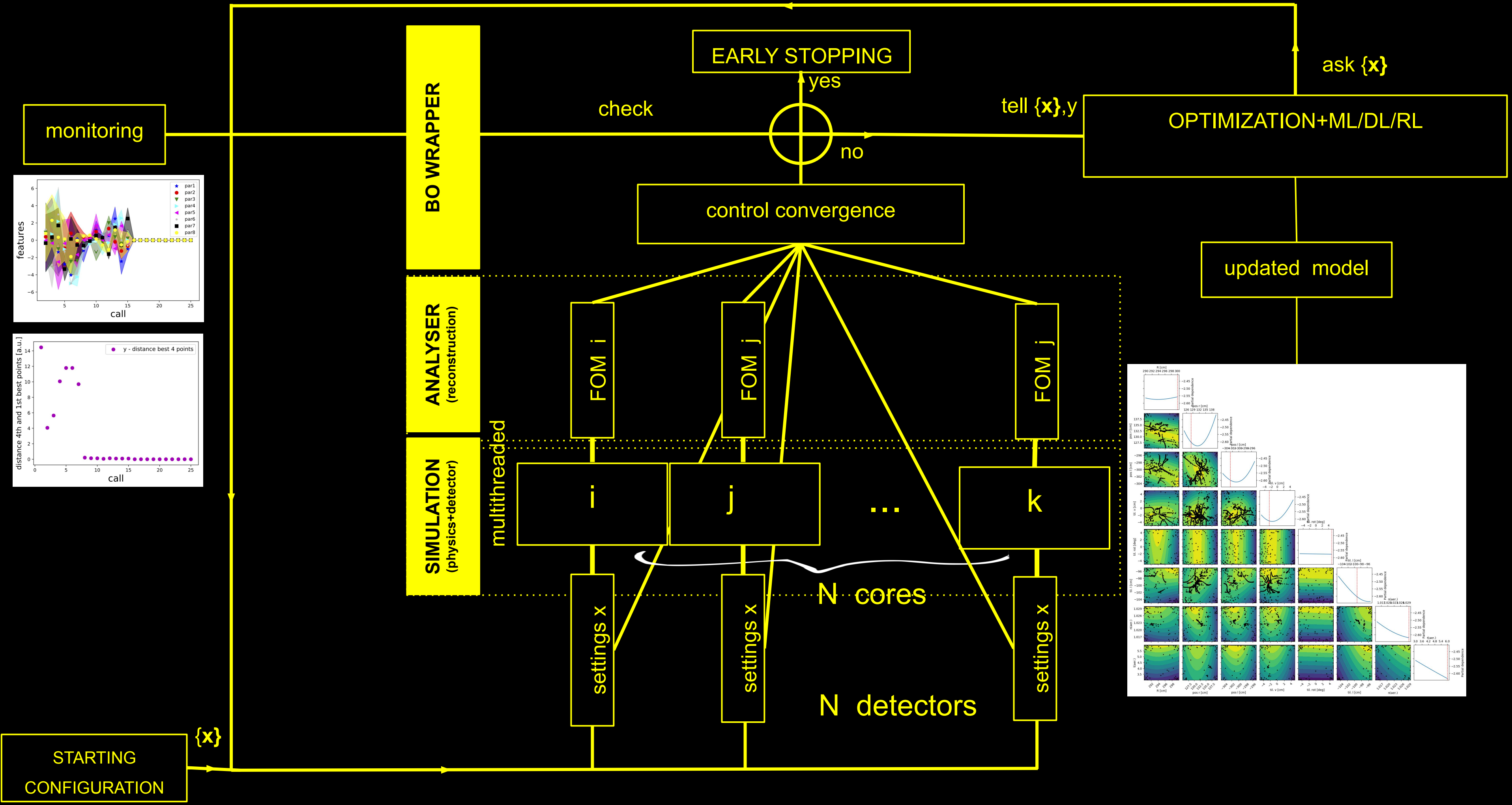
Bayesian Optimization

- 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.

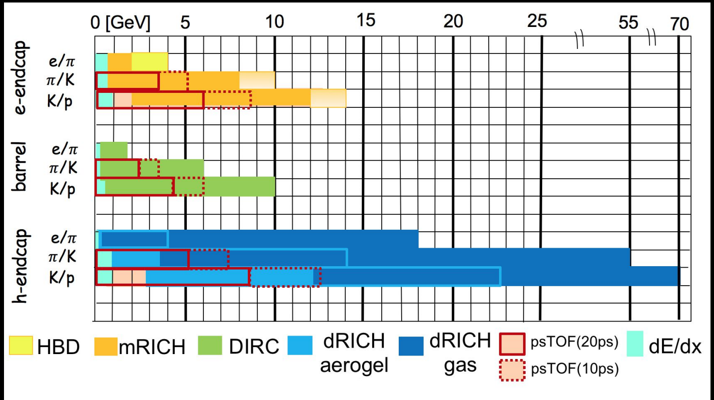
http://krasserm.github.io/2018/03/21/bayesian-optimization/
http://krasserm.github.io/2018/03/19/gaussian-processes/


EARLY STOPPING
yes
check
tell {x},y
ask {x}
OPTIMIZATION+ML/DL/RL
monitoring
BO WRAPPER
no
control convergence
updated model
ANALYSER
(reconstruction)
FOM i
FOM j
FOM j
SIMULATION
(physics+detector)
multithreaded
i
j
...
k
N cores
settings x
settings x
settings x
N detectors
{x}
STARTING
CONFIGURATION
features
call
par1
par2
par3
par4
par5
par6
par7
par8
y - distance best 4 points
distance 4th and 1st best points [a.u.]
call

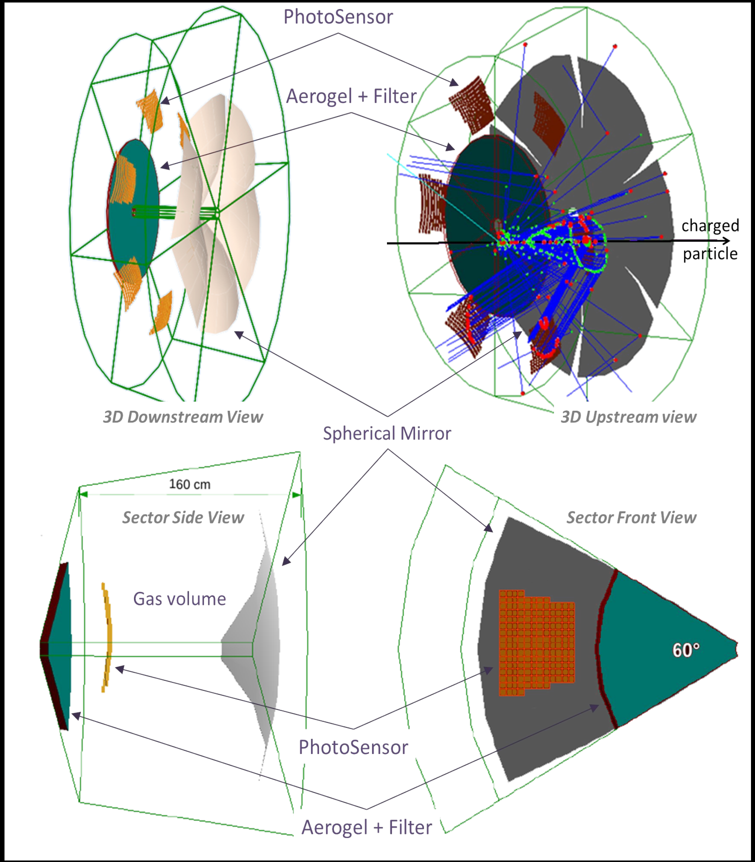
Dual RICH Example

E. Cisbani, A. Del Dotto, CF*, M. Williams et al.
JINST 15.05 (2020): P05009.

AI-optimized detector design for the future Electron-Ion Collider:
the dual-radiator RICH case.



aerogel (4 cm, n(400 nm): 1.02)
+ 3 mm acrylic filter
+ gas (1.6 m, n(C_2F_6): 1.0008)



- Continuous momentum coverage.
- Simple geometry and optics, cost effective.
- Legacy design from INFN, see EICUG2017

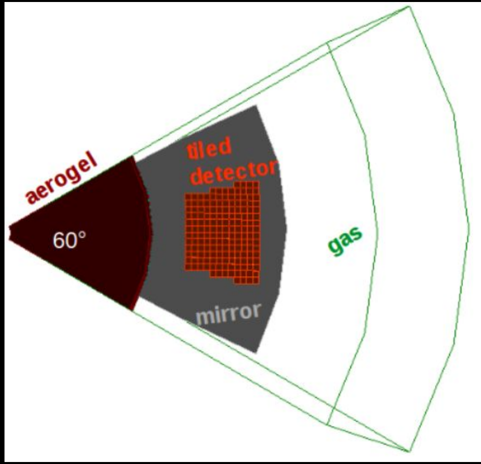
- 6 Identical open sectors (petals)
- Optical sensor elements:
  8500 cm^2/sector, 3 mm pixel
- Large focusing mirror

Construction Constraints

The idea is that we have a bunch of parameters to optimize that characterize the detector design.
We know from previous studies their ranges and the construction tolerances.

Variations below these values are irrelevant

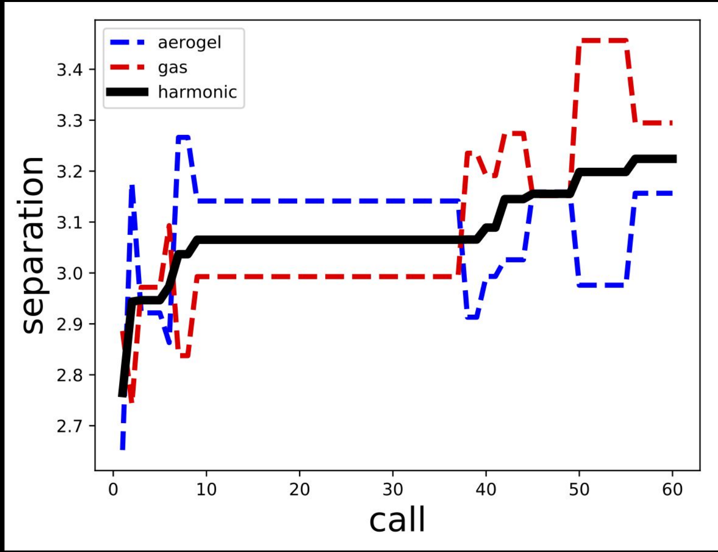
parameter	**description**	**range** [units]	**tolerance** [units]
R	mirror radius	[290,300] [cm]	100 [μm]
pos r	radial position of mirror center	[125,140] [cm]	100 [μm]
pos l	longitudinal position of mirror center	[-305,-295] [cm]	100 [μm]
tiles x	shift along x of tiles center	[-5,5] [cm]	100 [μm]
tiles y	shift along y of tiles center	[-5,5] [cm]	100 [μm]
tiles z	shift along z of tiles center	[-105,-95] [cm]	100 [μm]
$n_{aerogel}$	aerogel refractive index	[1.015,1.030]	0.2%
$t_{aerogel}$	aerogel thickness	[3.0,6.0] [cm]	1 [mm]

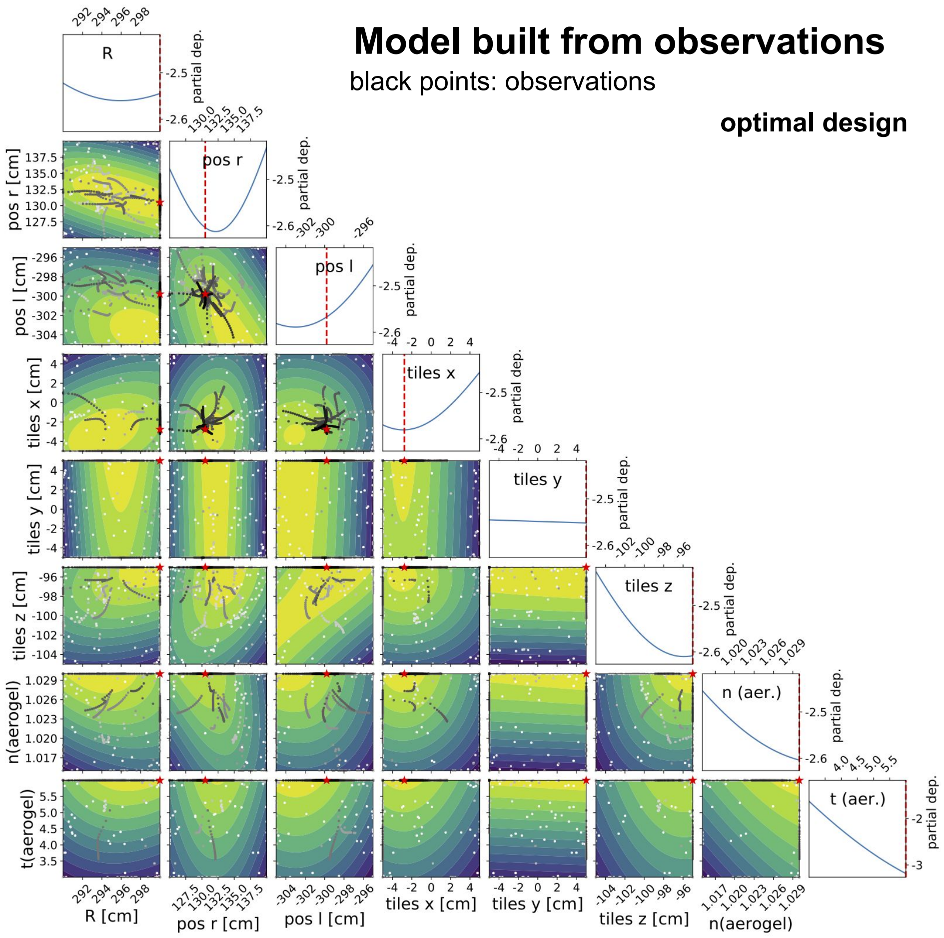


Ranges depend mainly on mechanical constraints and optics requirements.
These requirements can change in the next future based on inputs from prototyping.

The Model and the Optimized FoM

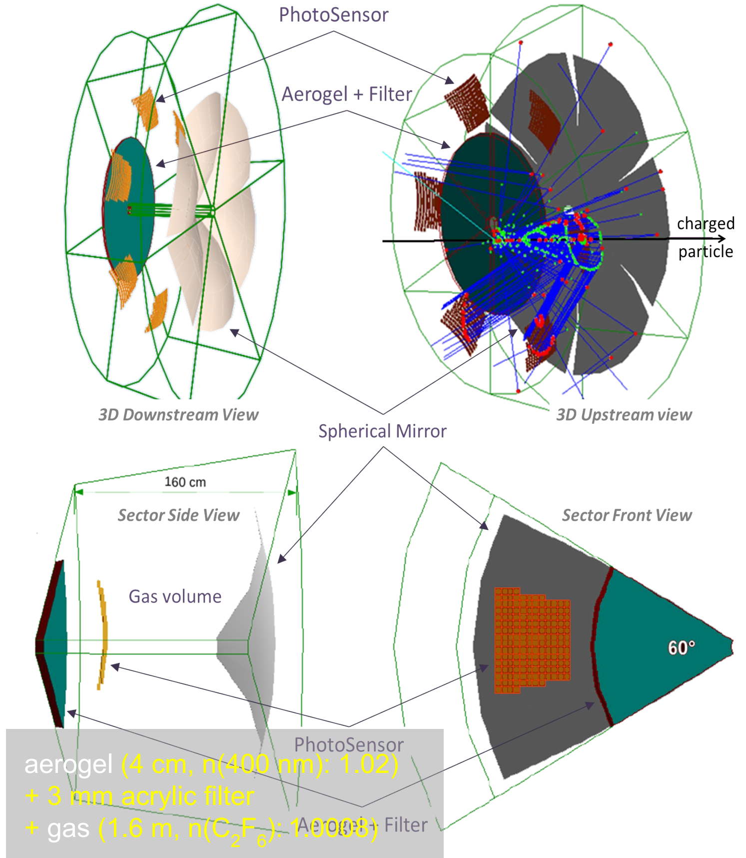
$$N\sigma = \frac{||\langle\theta_K\rangle - \langle\theta_\pi\rangle||\sqrt{N_\gamma}}{\sigma_\theta^{1p.e.}}$$

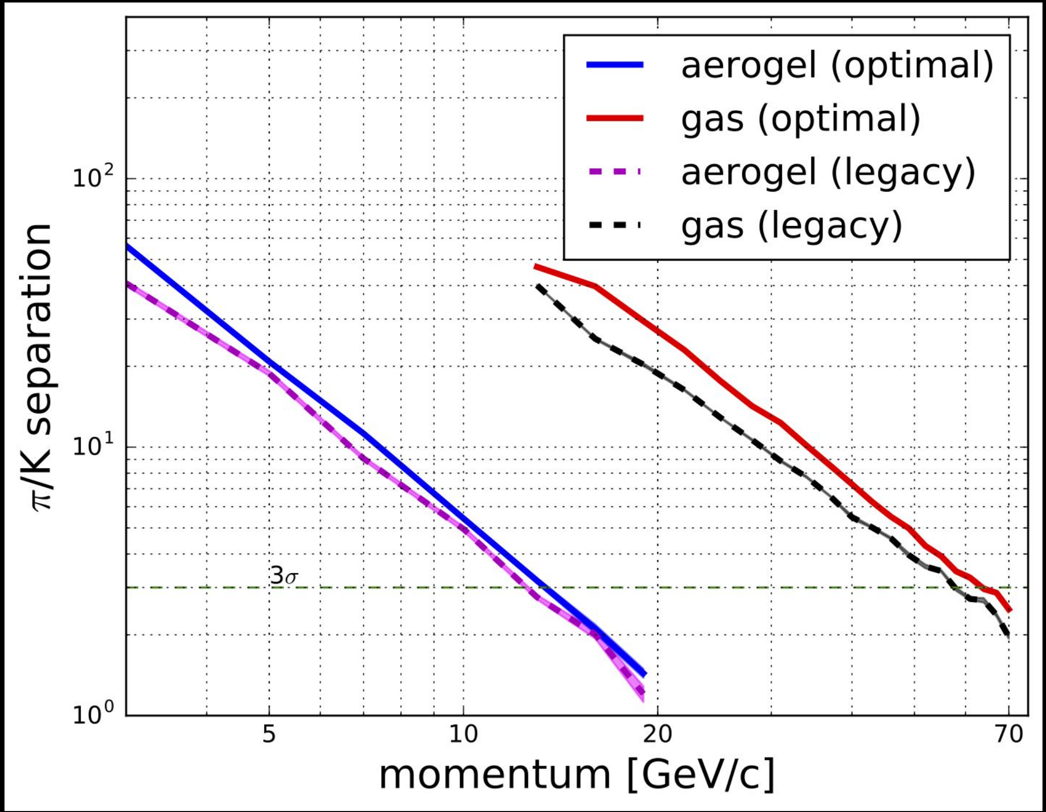




AI-Optimized dRICH

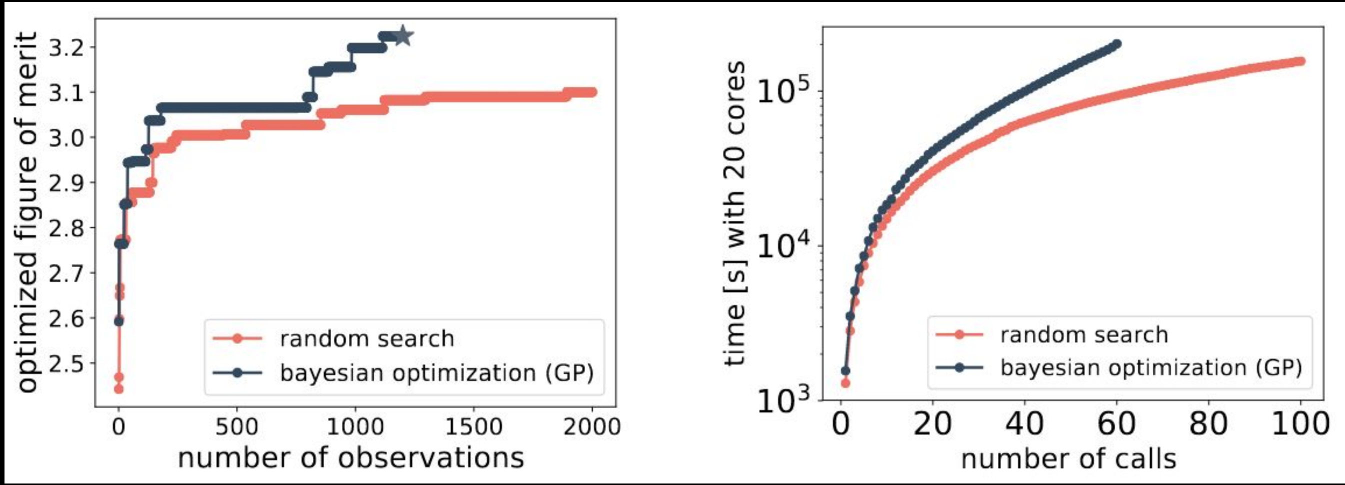
E. Cisbani, A. Del Dotto, CF*, M. Williams et al.
JINST 15.05 (2020): P05009.





- Statistically significant Improvement in both parts.
- In particular in the gas region where the 5σ threshold shifted from 43 to 50 GeV/c and the 3σ one extended up to
- Notice that before this study we did not know “how well” the legacy design was performing.

Comparison with Random Search



Each call:
400 tracks generated/core
20 cores

1 design point ~ 10 mins/CPU

Budget: 100 calls

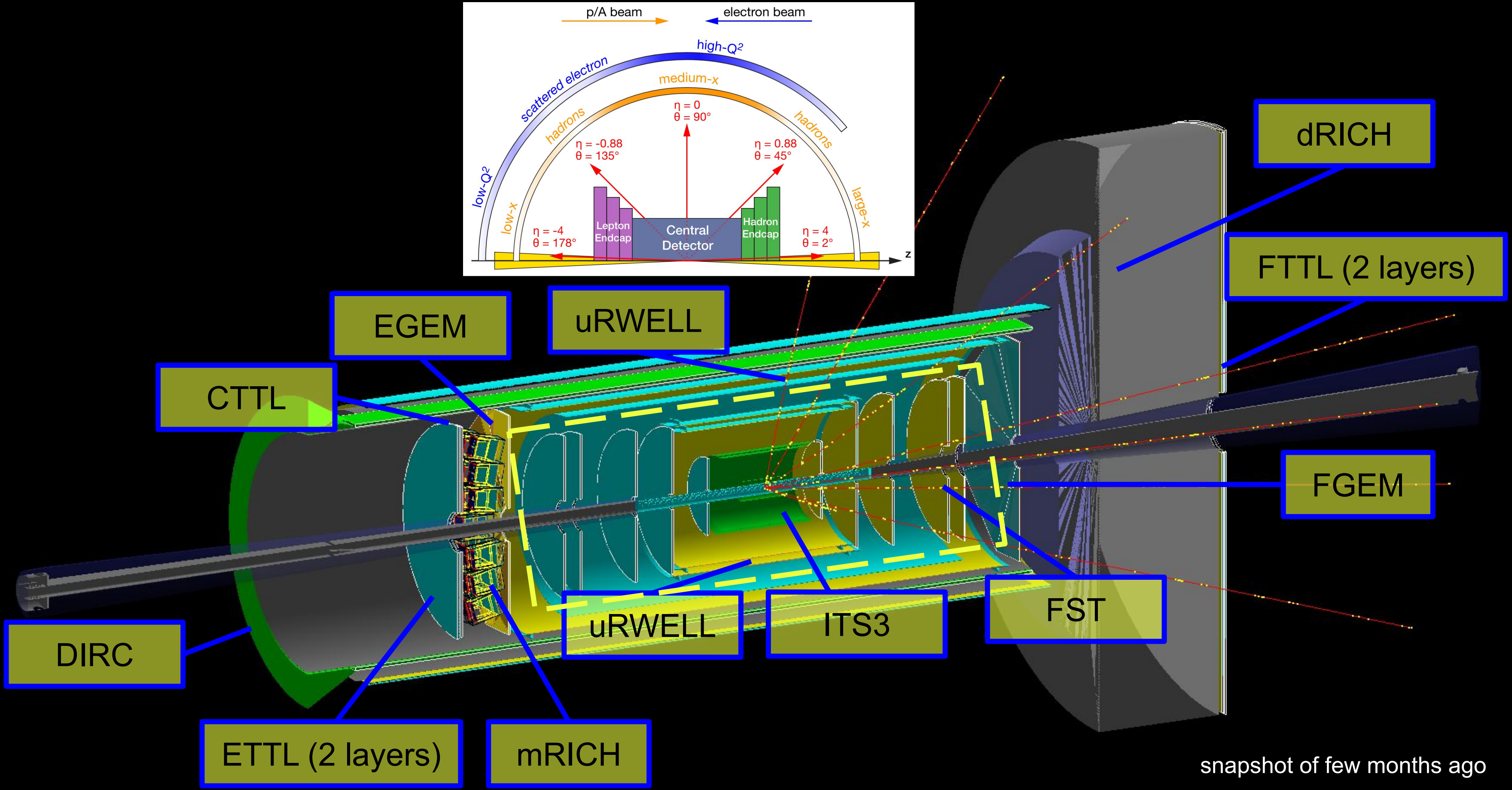
- BO with GP scales cubically with number of observations.

- Bayesian optimization methods are more promising because they offer principled approaches to weighting the importance of each dimension.

- For this 8D problem - even with 50 cores, RS looks unfeasible due to the curse of dimensionality.
  - Recall that the probability of finding the target with RS is $1-(1-v/V)^T$, where T is trials, v/V is the volume of target relative to the unit hypercube

Bergstra, Bengio, “Random search for hyper-parameter optimization”, J. Mach. Learn. Res.13 (Feb) (2012) 281–305.

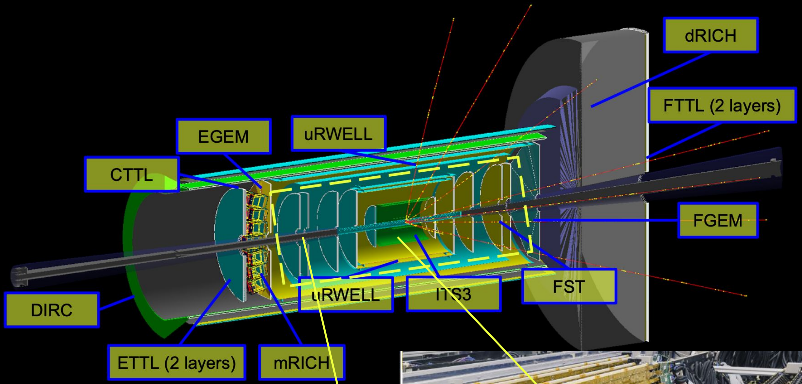
ECCE Tracker Example



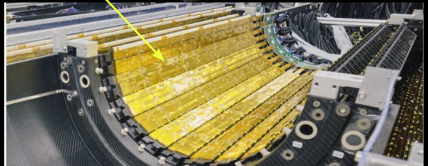
snapshot of few months ago

ECCE Inner Tracker

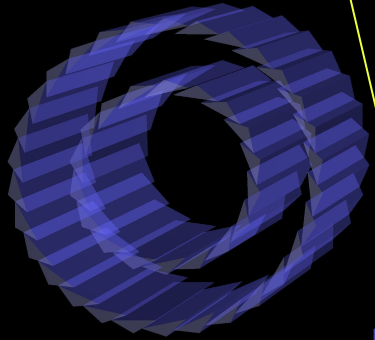
[arXiv:2102.08337]

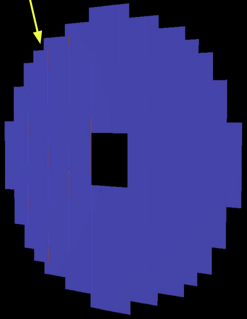


- Geometric parameters have significant impact in the performance of the tracker
- Effective parameterization of the detector design can reduce dimensionality
- Encode different geometric and structural constraints
  - ITS3 constrained (fixed strip length)
  - Mechanical constraints due to support structures
- The performance can be characterised by multiple figures of merit, a.k.a. “objectives” (e.g., resolution, reconstruction efficiency for the tracks). Analyses with 3 and 4 objectives…

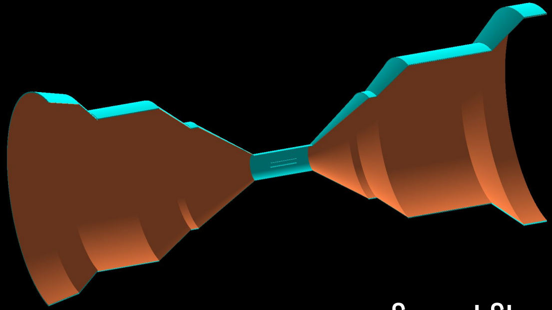


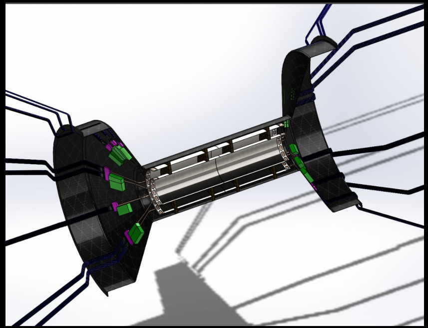
ALICE Si Vertex tracker





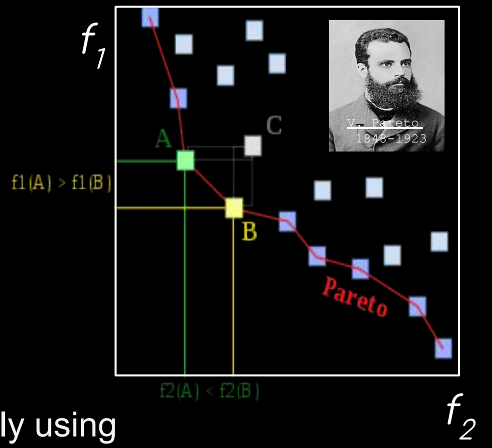
Fine grained implementation





Support Structures

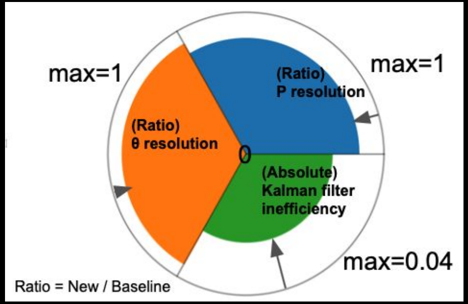
Multi-Objective Optimization



- The problem becomes challenging when the objectives are of conflict to each other, that is, the optimal solution of an objective function is different from that of the other.
- In solving such problems, with or without constraints, they give rise to a trade-off optimal solutions, popularly known as Pareto-optimal solutions.
- Due to the multiplicity in solutions, these problems were proposed to be solved suitably using evolutionary algorithms which use a population approach in its search procedure.
- MO-based solutions are helping to reveal important hidden knowledge about a problem – a matter which is difficult to achieve otherwise.

The ECCE Inner Tracker Design Optimization considers simultaneously:

- momentum resolution
- angular resolution
- Kalman filter efficiency
- (pointing resolution)
- Mechanical constraints



Frameworks

- Notice that MOO with dynamic/evolutionary algorithms (see, e.g., [1-3]) are probably the most utilized approaches, followed by more recent developments on multi-objective bayesian optimization (see, e.g., [4-7]). Using them has the advantage of having an entire community developing those tools.

  https://github.com/topics/multi-objective-optimization

- Agent-based approaches to MOO are also possible (see, e.g., [8]), but won't be discussed here.
- Remarkably these approaches can accommodate mechanical and geometrical constraints during the optimization process.

[1] J. J. Durillo and A. J. Nebro, "jMetal: A Java framework for multi-objective optimization," Advances in Engineering Software, vol. 42, no. 10, pp. 760–771, 2011.

[2] F.-A. Fortin, F.-M. De Rainville, M.-A. G. Gardner, M. Parizeau, and C. Gagné, "DEAP: Evolutionary algorithms made easy," The Journal of Machine Learning Research, vol. 13, no. 1, pp. 2171–2175, 2012.

[3] J. Blank and K. Deb, "pymoo: Multi-objective Optimization in Python," IEEE Access, vol. 8, pp. 89497–89509, 2020

[4] M. Laumanns and J. Ocenasek, "Bayesian optimization algorithms for multi-objective optimization," in International Conference on Parallel Problem Solving from Nature, pp. 298–307, Springer, 2002.

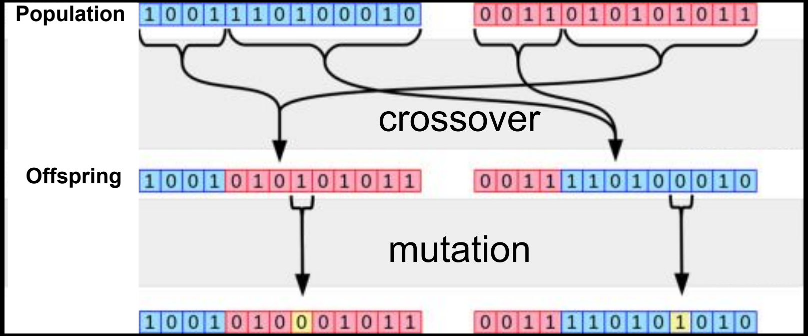
[5] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, "Botorch: Programmable bayesian optimization in pytorch," arXiv preprint arXiv:1910.06403, 2019.

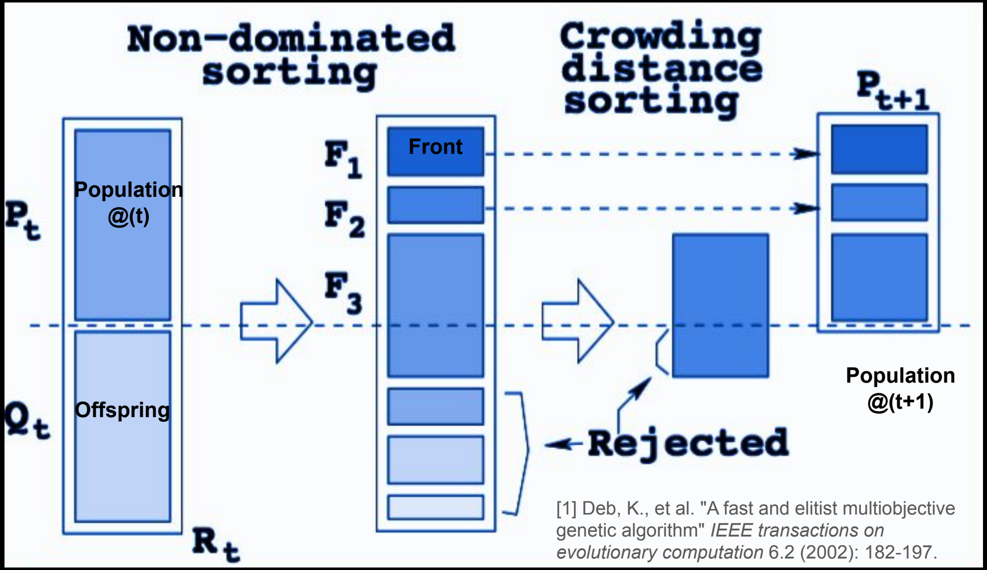
[6] P. P. Galuzio, E. H. de Vasconcelos Segundo, L. dos Santos Coelho, and V. C. Mariani, "MOBOpt—multi-objective Bayesian optimization," SoftwareX, vol. 12, p. 100520, 2020.

[7] A. Mathern, O. S. Steinholtz, A. Sjöberg, M. Önnheim, K. Ek, R. Rempling, E. Gustavsson, and M. Jirstrand, "Multi-objective constrained Bayesian optimization for structural design," Structural and Multidisciplinary Optimization, pp. 1–13, 2020.

[8] R. Yang, X. Sun, and K. Narasimhan, "A generalized algorithm for multi-objective reinforcement learning and policy adaptation," in Advances in Neural Information Processing Systems, pp. 14636–14647, 2019

Elitist Non-Dominated Sorting Genetic

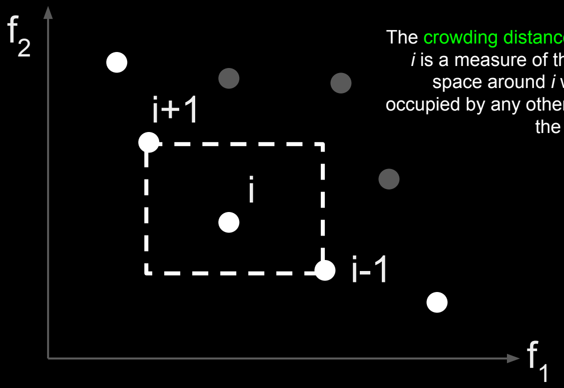




This is one of the most popular approach (>35k citations on google scholar), characterized by:

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

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.



The crowding distance d_i of point i is a measure of the objective space around i which is not occupied by any other solution in the population.

Under the hood…

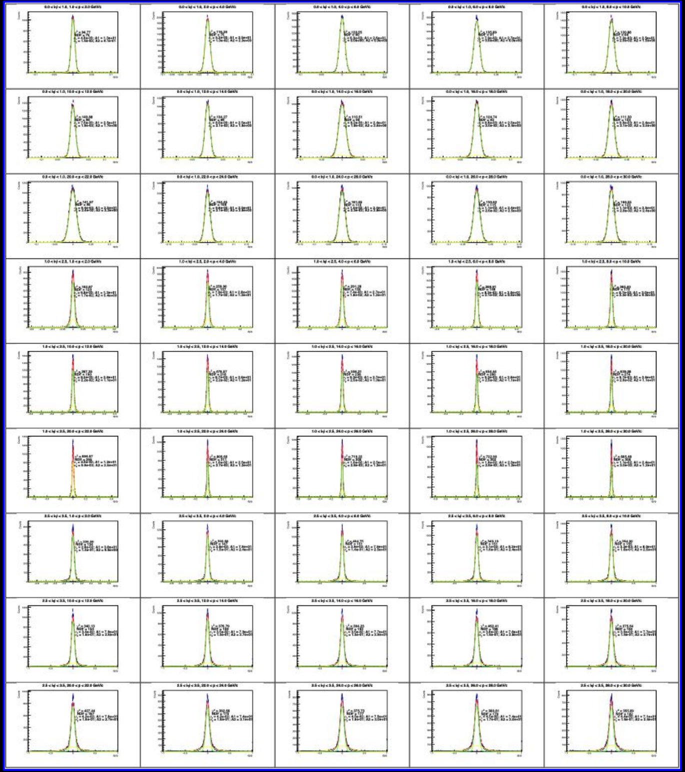
1) Start with the definitions:

$$\frac{\Delta P}{P} = \frac{P_{reco} - P_{true}}{P_{true}}$$

$$\Delta\Theta = \Theta_{reco} - \Theta_{true}$$

Fine-grained analysis in phase-space

2) Extract resolutions in bins of the phase-space (P,η). Do this for baseline and for each new design. For each bin we can calculate ratios.



3) We use global weighted quantities for the objectives representing the resolutions. Weights are obtained propagating uncertainties from the fits.

$$\underset{\text{Ratio wrt reference}}{R(f)} = \frac{1}{N_\eta} \sum_\eta \left(\frac{\sum_p \overset{\text{from fits}}{w_{p,\eta}} \cdot R(f)_{p,\eta}}{\sum_p w_{p,\eta}} \right)$$

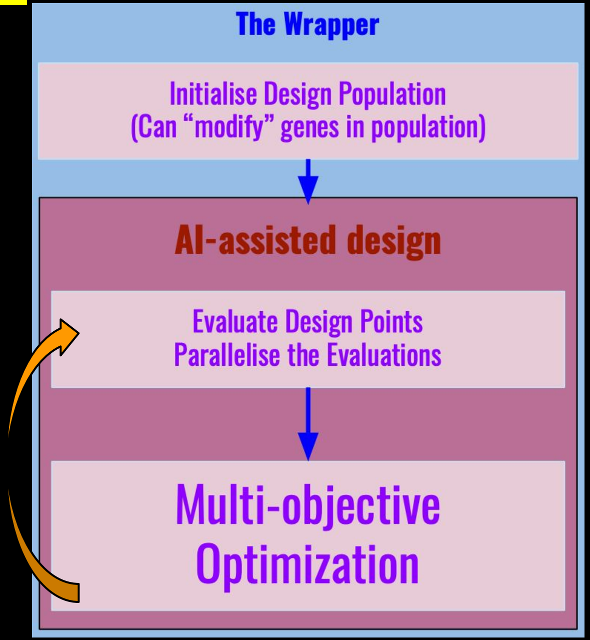
✅ momentum resolution
✅ angular resolution

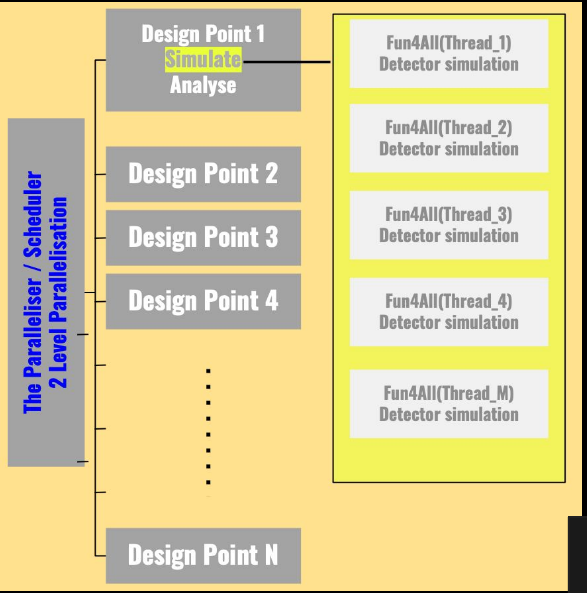
4) We directly calculate the global KF inefficiency as:

$$R(\mathrm{KF}) = \frac{\mathrm{N(bad\ \ tracks)}}{\mathrm{N(tot.\ \ tracks)}}$$

✅ Kalman filter (in)efficiency

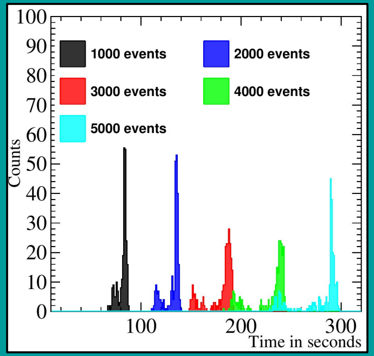
Software Stack





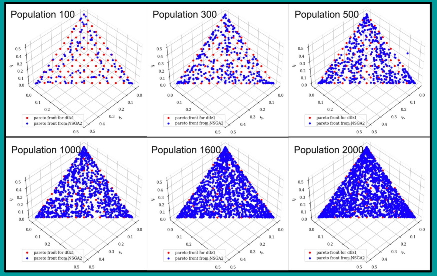
Simulation Time

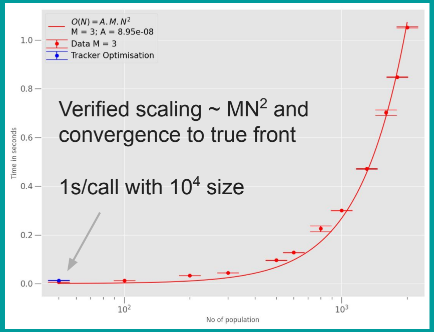
Simulating 80000 in total for each evaluation, 1 evaluation takes <=80 mins



Dask natively scales Python

GA + sorting Time





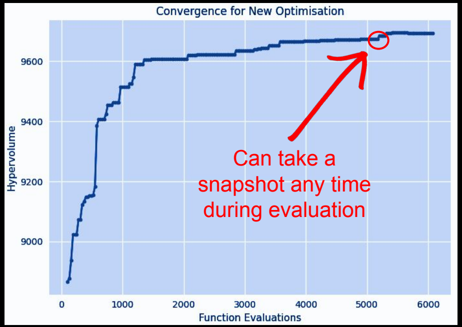
Used a test problem DTLZ1

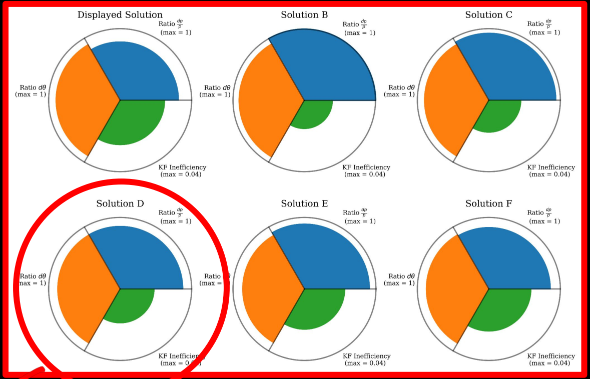
- 2-level parallelization
  [Launch multiple design points in parallel; each point is parallelized]
- With 11 variables and 3 objectives ~10k CPUhours

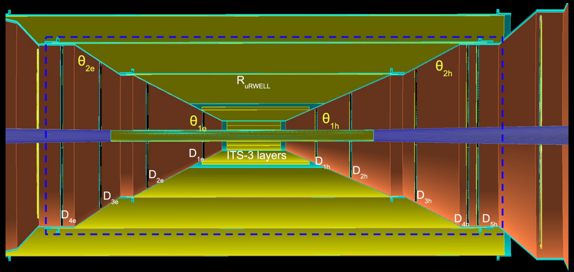
Optimal Design Solutions

This is (already) an unprecedented attempt in detector design for complexity!

- ≥ 11 parameters
- 3 (4) objectives
- ≥ 5 mechanical constraints
- Population size 100
- Offspring distr. over ≥ 30 cores
- 80000 tracks / design point
- ~1h / design point



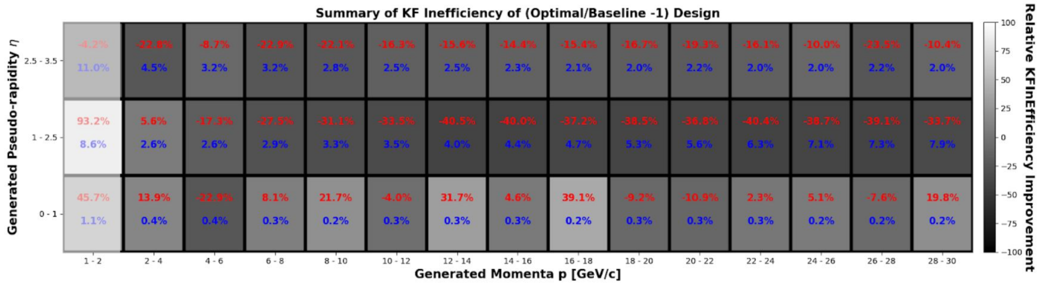




Select one solution (e.g., D) from the Pareto front and evaluate performance

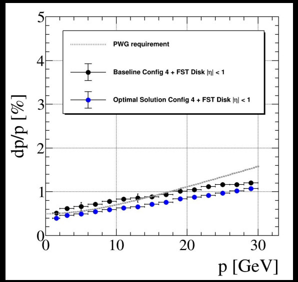
KF inefficiency

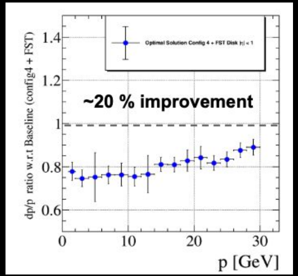
- Optimal/baseline -1
- Baseline Ineff



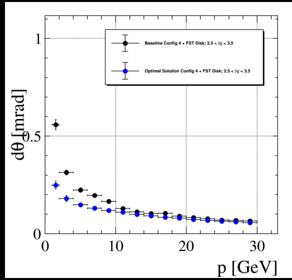
Generated Pseudo-rapidity η	1 - 2	2 - 4	4 - 6	6 - 8	8 - 10	10 - 12	12 - 14	14 - 16	16 - 18	18 - 20	20 - 22	22 - 24	24 - 26	26 - 28	28 - 30
2.5 - 3.5	-4.2% / 11.0%	-22.8% / 4.5%	-8.7% / 3.2%	-22.9% / 3.2%	-22.1% / 2.8%	-16.3% / 2.5%	-15.6% / 2.5%	-14.4% / 2.3%	-15.4% / 2.1%	-16.7% / 2.0%	-19.3% / 2.2%	-16.1% / 2.0%	-10.0% / 2.0%	-23.5% / 2.2%	-10.4% / 2.0%
1 - 2.5	93.2% / 8.6%	5.6% / 2.6%	-17.3% / 2.6%	-27.5% / 2.9%	-31.1% / 3.3%	-33.5% / 3.5%	-40.5% / 4.0%	-40.0% / 4.4%	-37.2% / 4.7%	-38.5% / 5.3%	-36.8% / 5.6%	-40.4% / 6.3%	-38.7% / 7.1%	-39.1% / 7.3%	-33.7% / 7.9%
0 - 1	45.7% / 1.1%	13.9% / 0.4%	-22.9% / 0.4%	8.1% / 0.3%	21.7% / 0.2%	-4.0% / 0.3%	31.7% / 0.3%	4.6% / 0.3%	39.1% / 0.2%	-9.2% / 0.3%	-10.9% / 0.3%	2.3% / 0.3%	5.1% / 0.2%	-7.6% / 0.2%	19.8% / 0.2%

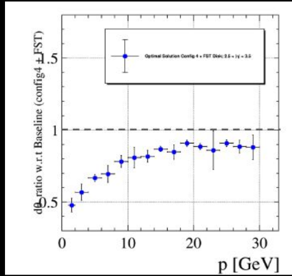
dP/P Resolution





dθ Resolution





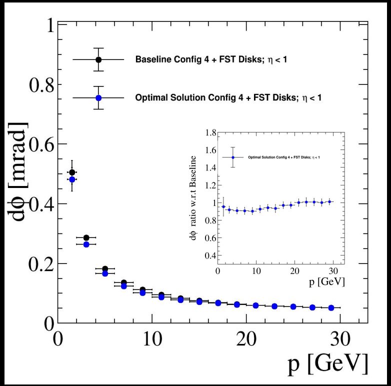
Shown for |η|<1. Performance is studied in the entire acceptance (other bins)

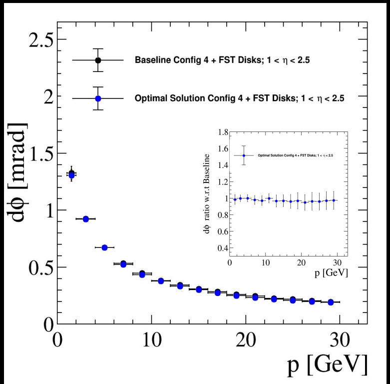
Validation and decision making

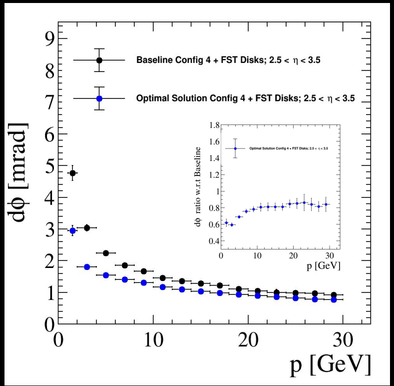
Validation is done by looking at other figures of merit characterizing the detector performance that have not been directly used in the optimization process

The decision making process done after optimization.
For each design solution in the Pareto Front one can study the corresponding detector performance.

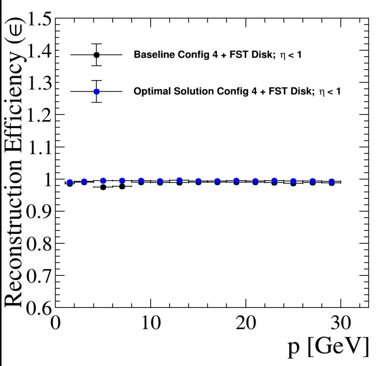
dφ Resolution

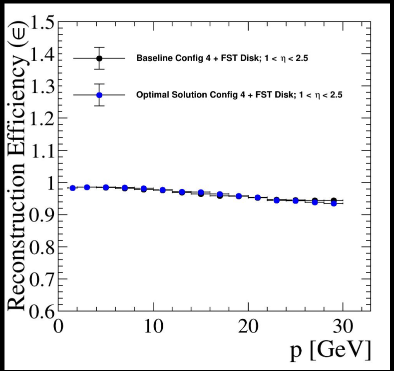


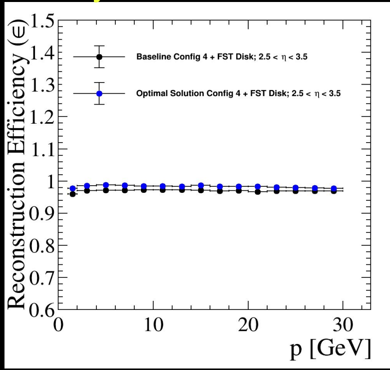




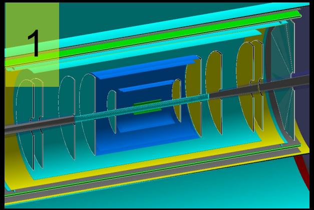
Reconstruction Efficiency



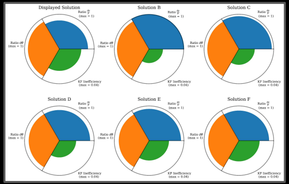


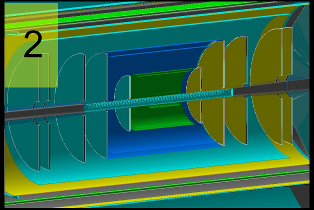


Different Technologies/Multiple Pipelines

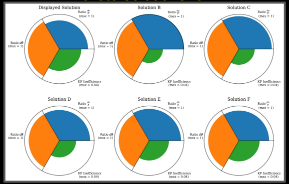


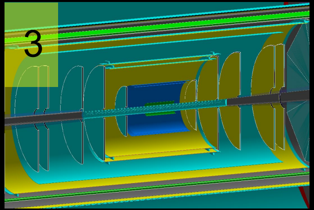
Pareto solutions for 1



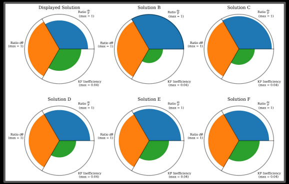


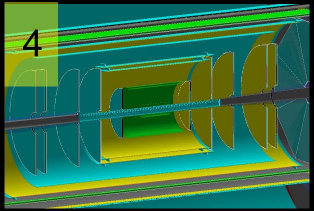
Pareto solutions for 2



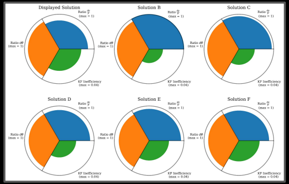


Pareto solutions for 3





Pareto solutions for 4



E.g. Inner Tracker Barrel (+ disks in the h-endcap and e-endcap)

- Configuration 1: 2-vtx (ITS3) + 2-sagitta (ITS2) + 2-outer layer (ITS2)
- Configuration 2: 2-vtx (ITS3) + 2-sagitta (ITS3) + 2-outer layer (ITS2)
- Configuration 3: 2-vtx (ITS3) + 2-sagitta (ITS2) + 2-outer layer (uRwell)
- Configuration 4: 2-vtx (ITS3) + 2-sagitta (ITS3) + 2-outer layer (uRwell)

DWG's:

- Technology Selection
- Baseline Design
- Alternative Configuration(s)

PWG's:

- Physics Signal Selection
- Physics Performance Evaluation

CWG's:

- Simulation Campaign Support
- AI Optimization

updated configurations with any additional requirements

New optimization pipelines

Novel Aerogel Material **aefib**

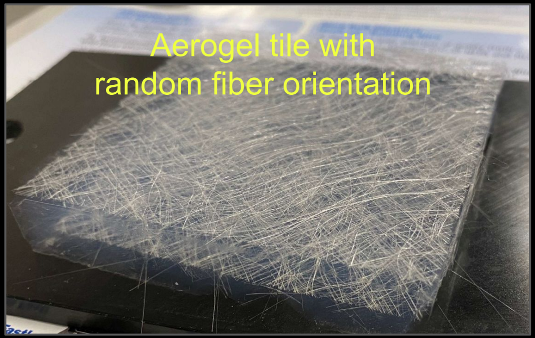
The team: V. Berdnikov, J. Crafts, E. Cisbani, CF, T. Horn, R. Trotta

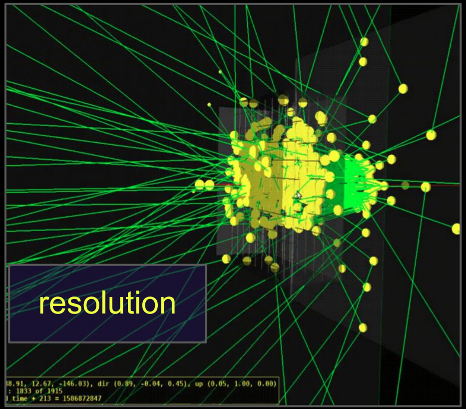
- Aerogels with low refractive indices are very fragile tiles break during production and handling, and their installation in detectors.
- To improve the mechanical strength of aerogels, Scintilex developed a reinforcement strategy. The general concept consists of introducing fibers into the aerogel that increase mechanical strength, but do not affect the optical properties of the aerogel.
- Paper in preparation.

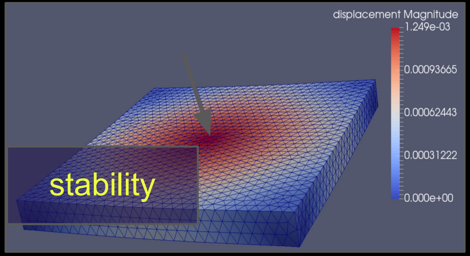
Software Stack

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

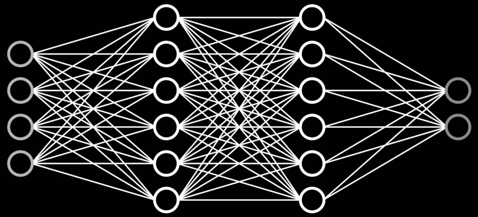


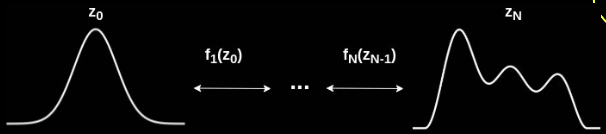


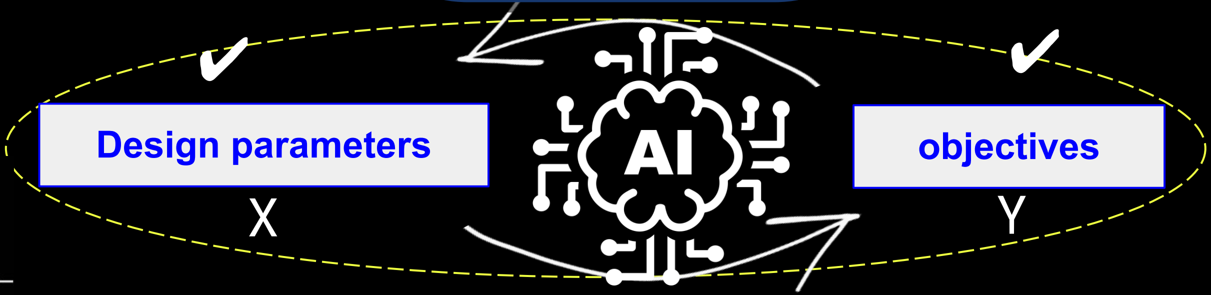


…with large datasets

A.I.
gathers observations and suggests new points







Injection of Physics Events

Detector Simulation

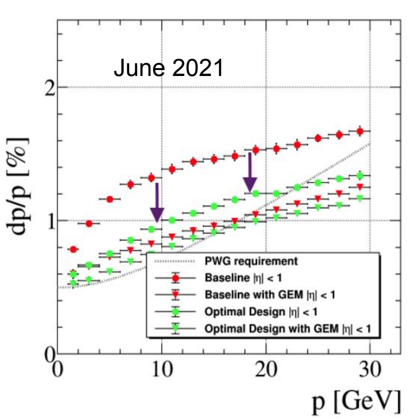
Analysis of High-level reconstruction of events

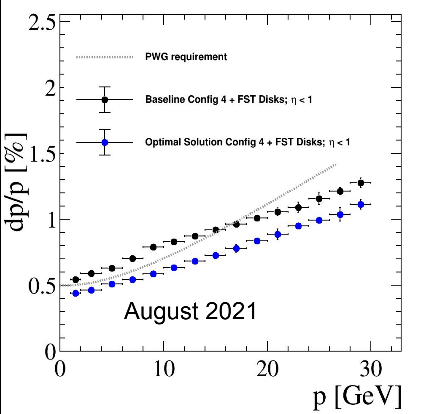
(Considerations)

Summary

[1] Liu, Xin, et al. IEEE Trans Parallel Distrib Syst 32.4 (2020): 975-987.
[2] R. Stevens, et al. *AI for Science*. No. ANL-20/17. ANL, IL (US), 2020.

- EIC can be one of the first experiments to systematically leverage on AI during the R&D and Design phase.
- ECCE created an AI WG to lead these efforts with an unprecedented attempt in detector design (multidimensional design and multiple objectives).
- None ever accomplished a multi-dimensional / multi-objective optimization of many sub-detectors combined together within the global design.
  - costs can be included provided reliable parametrization
  - speed-up bottlenecks (sim/reco steps)
- Larger populations of design points can improve accuracy of the Pareto front. A recent trend in MOO is distributed optimizations and implementation on supercomputers [1].
- AI can assist the development of a detector during the design phase. This can be extended to other instrumentation designs in the industry (e.g., medical imaging)





One of the conclusions from the DOE Town Halls on AI for Science on 2019 [2] was that *"AI techniques that can optimize the design of complex, large-scale experiments have the potential to revolutionize the way experimental nuclear physics is currently done".*

Spares