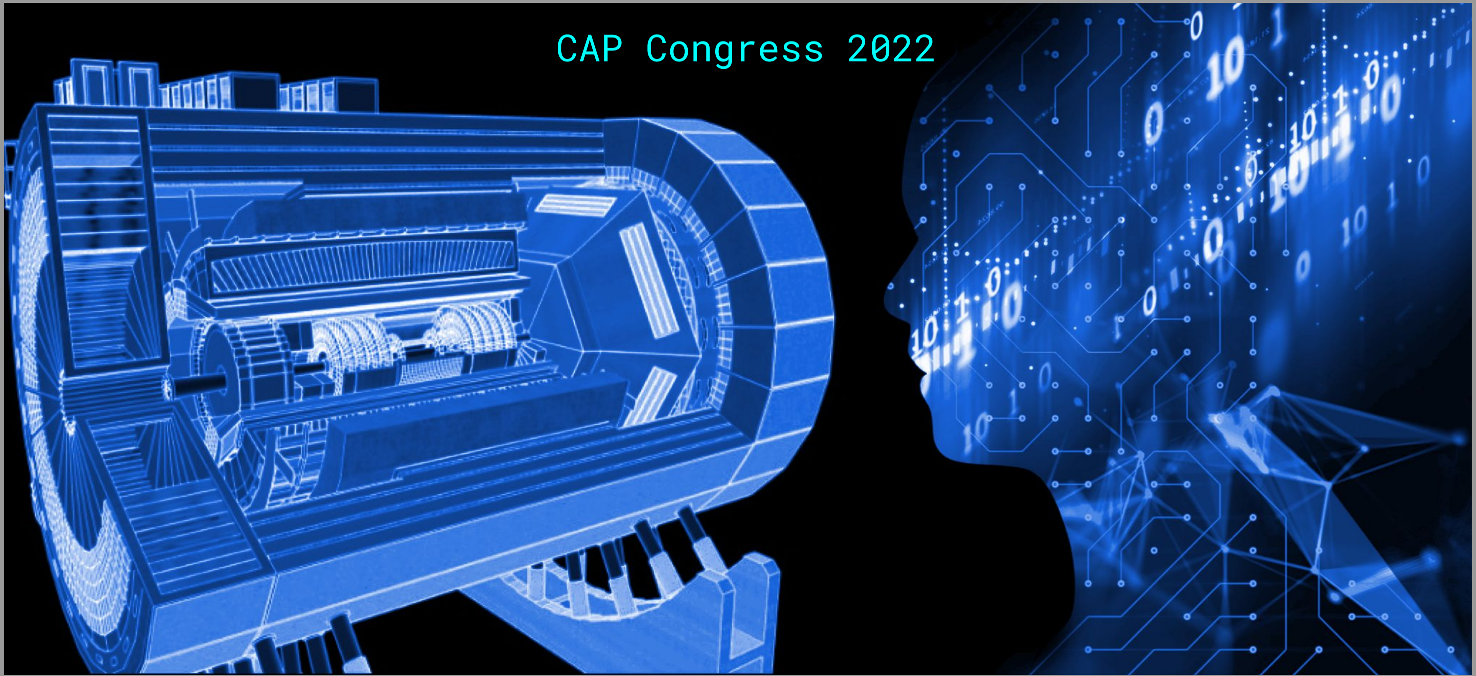
AI-assisted design of the EIC Detector



Cristiano Fanelli

AI for Design

It is a relatively new but active area of research.
Many applications in, e.g., industrial material, molecular and drug design.

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

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]

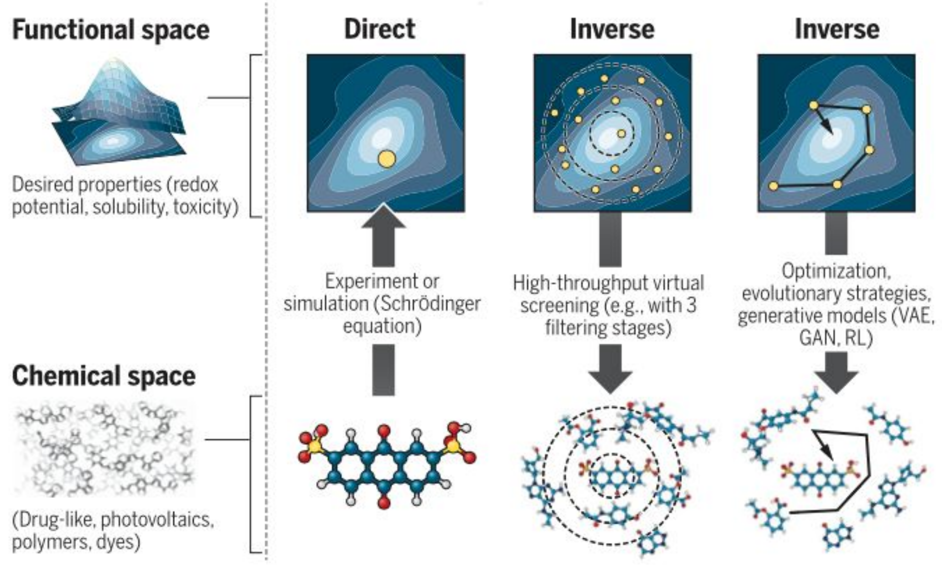


Fig. 2. Schematic of the different approaches toward molecular design. Inverse design starts from desired properties and ends in chemical space, unlike the direct approach that leads from chemical space to the properties.

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

AI for Detector Design

- When it comes to designing detectors with AI this is an area at its “infancy”.
- Typically full detector design is studied once the subsystem prototypes are ready (phase constraints from the full detector or outer layers are taken into consideration).
- Need to use advanced simulations which are computationally expensive (Geant).
- Many parameters (and multiple objective functions): curse of dimensionality [1].
- Entails establishing a procedural body of instructions [2].
- The choice of a suitable algorithm is a challenge itself (no free lunch theorem [3]) and always requires some degree of customization.
- Non-differentiable terms.

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

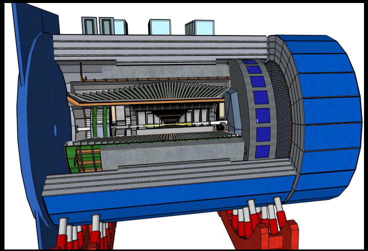
What follows based on a series of lectures on Detector Design with AI at the AI4NP Winter School

[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

AI-assisted Workflow

[See invited talk at IAEA Technical Meeting on AI]

IAEA
International Atomic Energy Agency



- AI can assist in designing more efficiently detectors (performance, costs).
- It helps steering the design (and eventually fine-tune it).
- It can capture hidden correlations among design parameters.

A.I.
gathers observations and suggests new points

customization

Design parameters

AI

Injection of Physics Events

Detector Simulation

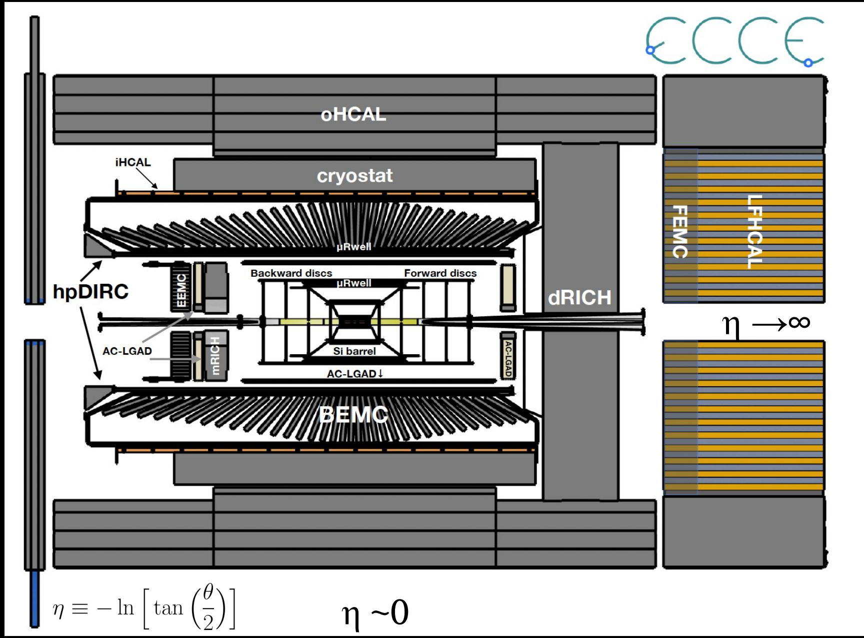
compute intensive (Geant4)

Analysis of High-level reconstruction of events

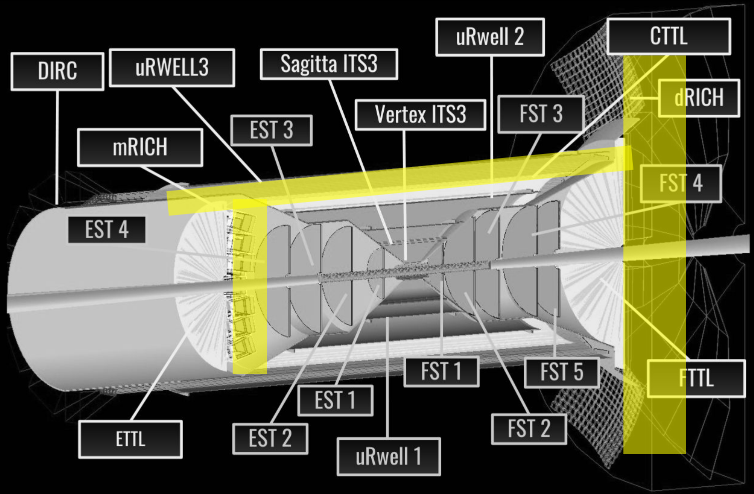
(AI/ML can also speed-up the simulation/reconstruction stack; cf. Amdahl's law)

The EIC Detector

We have a reference (ECCE) detector.
Possible updates are currently being investigated (detector-1).



Tracker System + PID



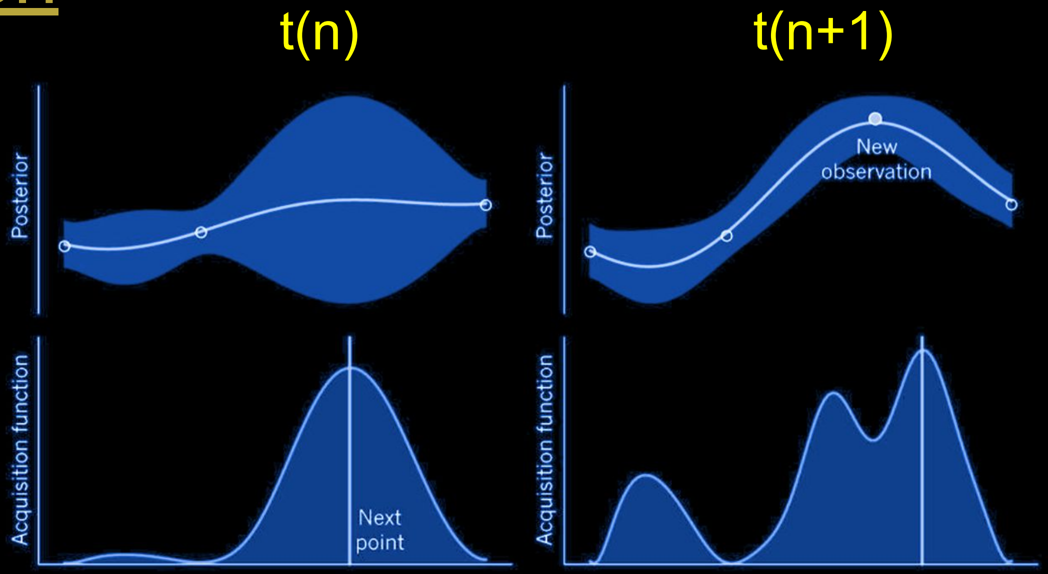
- The tracking system reconstructs charged particle tracks. It combines different technologies.
- Imaging Cherenkov detectors are the backbone of PID in EIC. Compute intensive to simulate / reconstruct.
- In this presentation: detector design, simulation/reconstruction with AI/ML for EIC

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.**



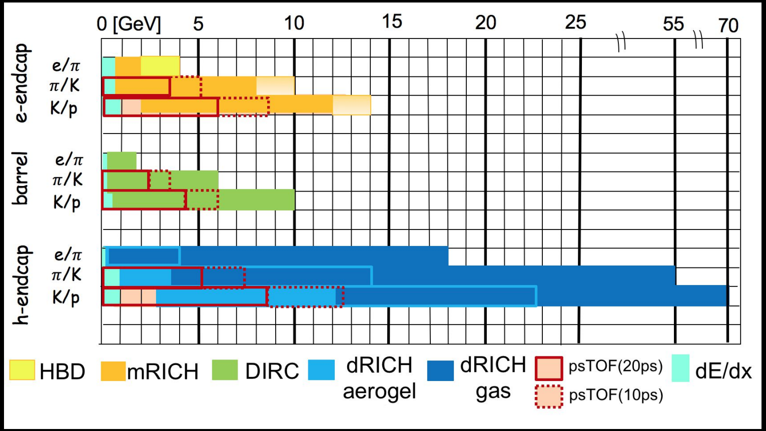
1. Select a Sample by Optimizing the Acquisition Function.
2. Evaluate the Sample With the Objective Function.
3. Update the Data and, in turn, the Surrogate Function.
4. Go To 1.

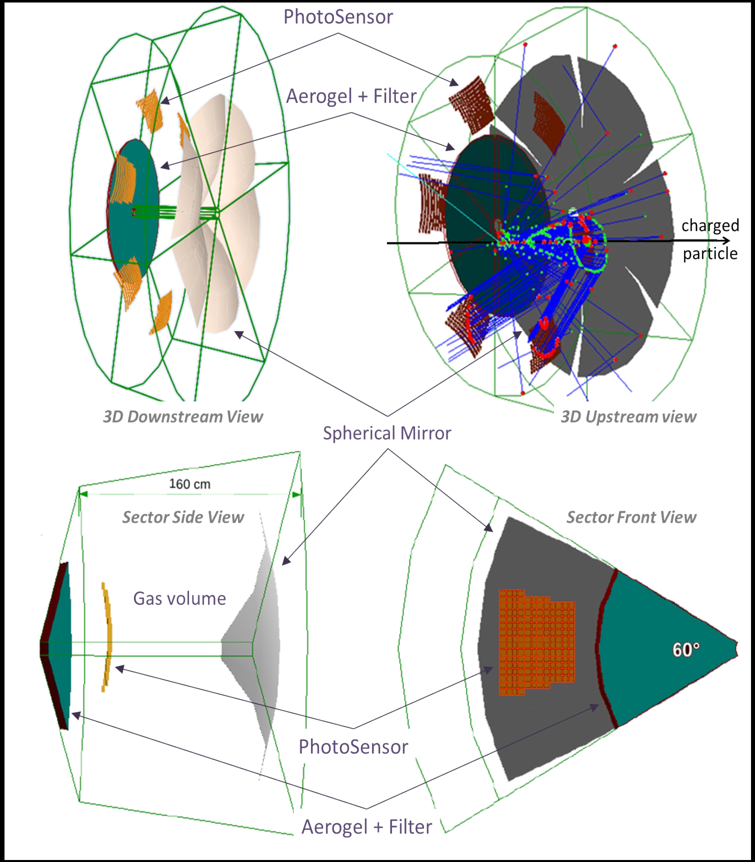
http://krasserm.github.io/2018/03/21/bayesian-optimization/
http://krasserm.github.io/2018/03/19/gaussian-processes/

Dual RICH: case study

E. Cisbani, A. Del Dotto, CF*, M. Williams et al.
"AI-optimized detector design for the future Electron-Ion Collider: the dual-radiator RICH case."
JINST 15.05 (2020): P05009.

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

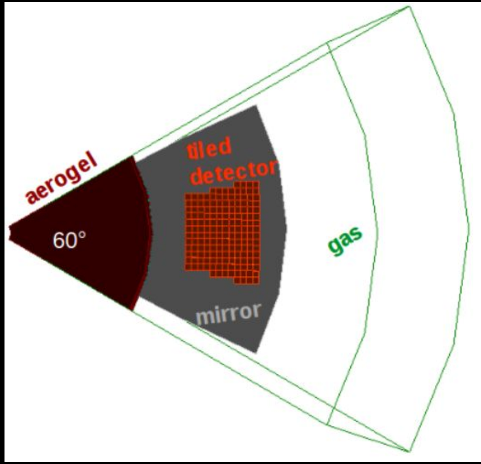
- 6 Identical open sectors (petals)
- Optical sensor elements: 8500 cm^2/sector, 3 mm pixel
- Large focusing mirror

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.

Convergence Criteria

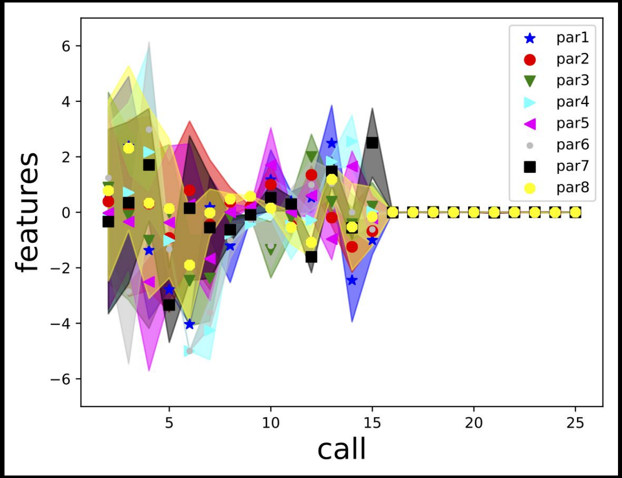
- Can in general be applied in the design space, in the objective space, or looking at the behavior of the acquisition function.

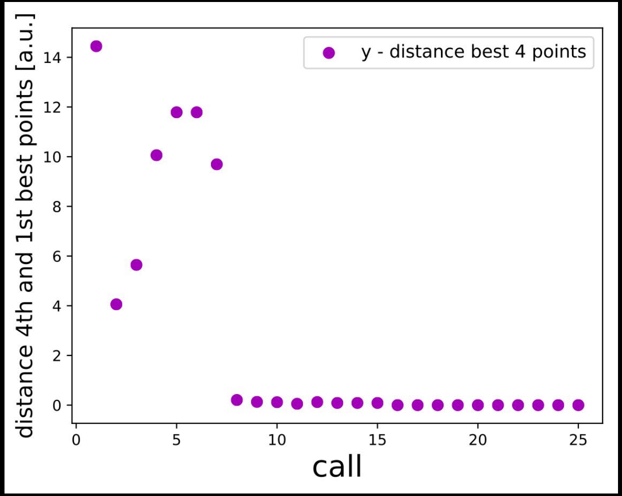
- We defined a set of conditions to ensure convergence:

  - These correspond to the logic AND of booleans on each feature and on the variation of the figure of merit.

  - They are built on standardized Z and Fisher statistics.

- Pre-processing of data required to remove outliers.

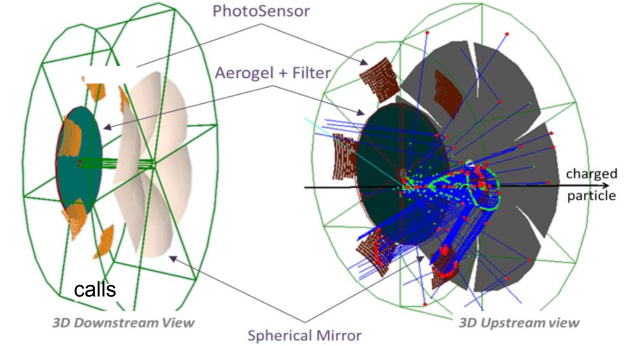


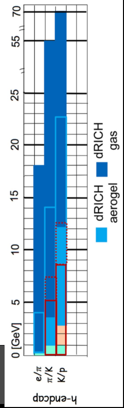


Dual RICH: ante proposal

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

- Two radiators with different refractive indices for continuous momentum coverage.
- Simulation of detector and processes is compute-intensive
- Legacy design from INFN (EICUG2017).





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

1 Define design parametrization and space

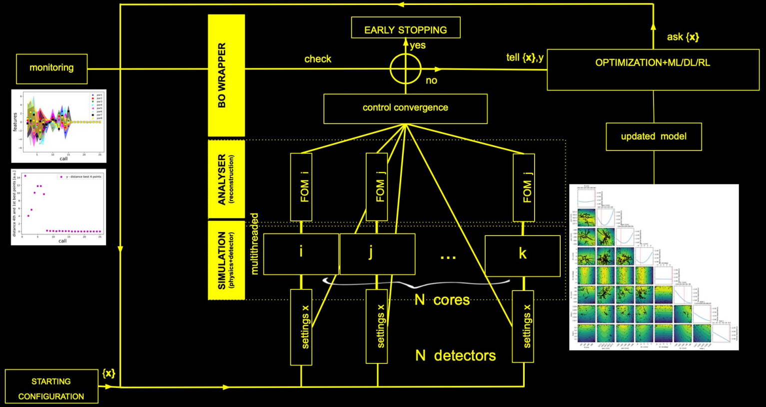
parameter	**description**	**range** [units]	**tolerance** [units]
R	mirror radius	[290,300] [cm]	100 [μm]
pos r	radial position of mirror center	[125,140] [cm]	100 [μm]
pos l	longitudinal position of mirror center	[-305,-295] [cm]	100 [μm]
tiles x	shift along x of tiles center	[-5,5] [cm]	100 [μm]
tiles 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]

2 Come up with a smart objective; study / characterize properties (noise, stats needed etc): simulation + reconstruction

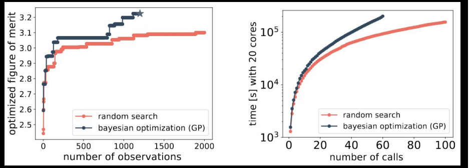
$$N\sigma = \frac{|\langle\theta_K\rangle - \langle\theta_\pi\rangle|\sqrt{N_\gamma}}{\sigma_\theta^{1p.e.}}$$

$$h = 2\cdot\left[\frac{1}{(N\sigma)|_1} + \frac{1}{(N\sigma)|_2}\right]^{-1}$$

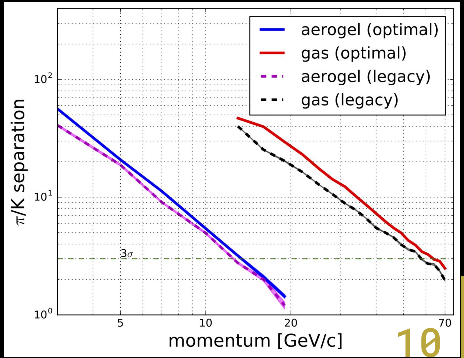
3 Optimization framework (embed convergence criteria)

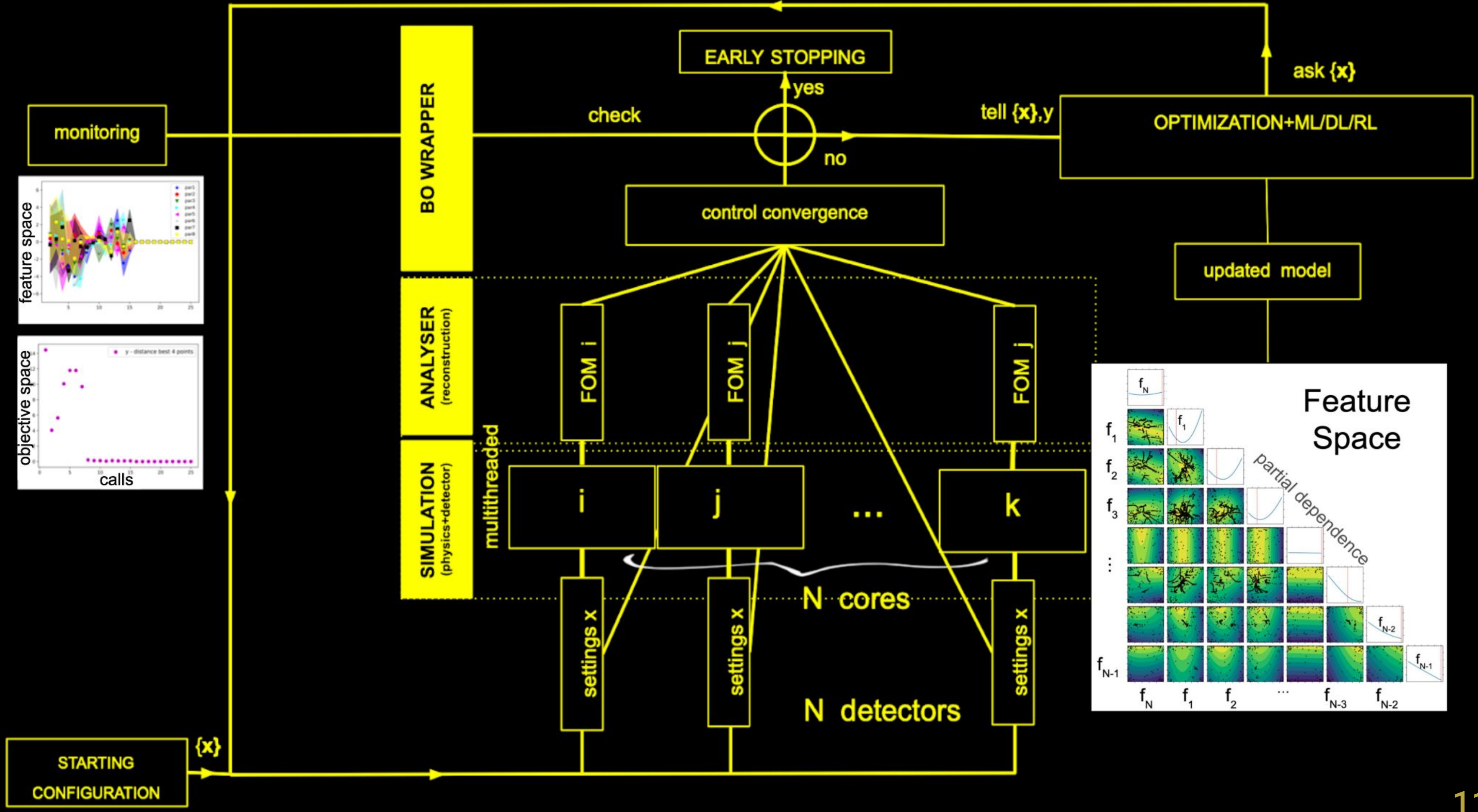


4 Analysis + Validation



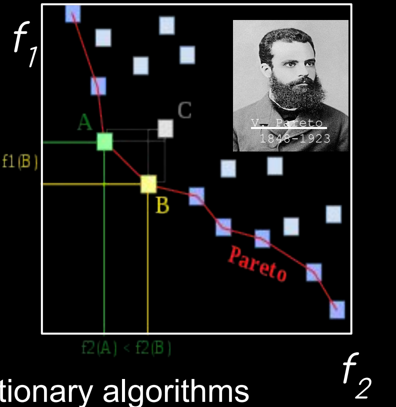
principled vs random





EARLY STOPPING
yes
check
no
tell {x},y
ask {x}
monitoring
BO WRAPPER
OPTIMIZATION+ML/DL/RL
control convergence
updated model
feature space
objective space
calls
ANALYSER (reconstruction)
SIMULATION (physics+detector)
multithreaded
FOM i
FOM j
FOM j
i
j
...
k
N cores
settings x
settings x
settings x
N detectors
Feature Space
partial dependence
STARTING CONFIGURATION
{x}

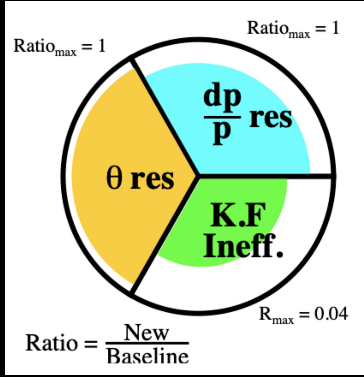
Multi-Objective Optimization



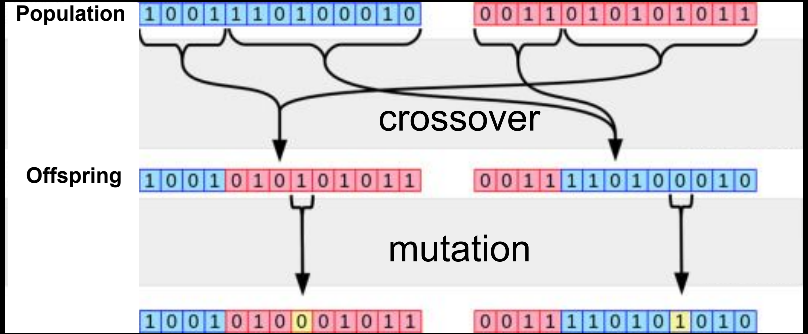
- The problem becomes challenging when the objectives are of conflict to each other, that is, the optimal solution of an objective function is different from that of the other.
- In solving such problems, with or without constraints, they give rise to a trade-off optimal solutions, popularly known as 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
- During the proposal we used both evolutionary (1) and bayesian approaches (2). I will describe now (1). For implementation details see talk by K. Suresh.

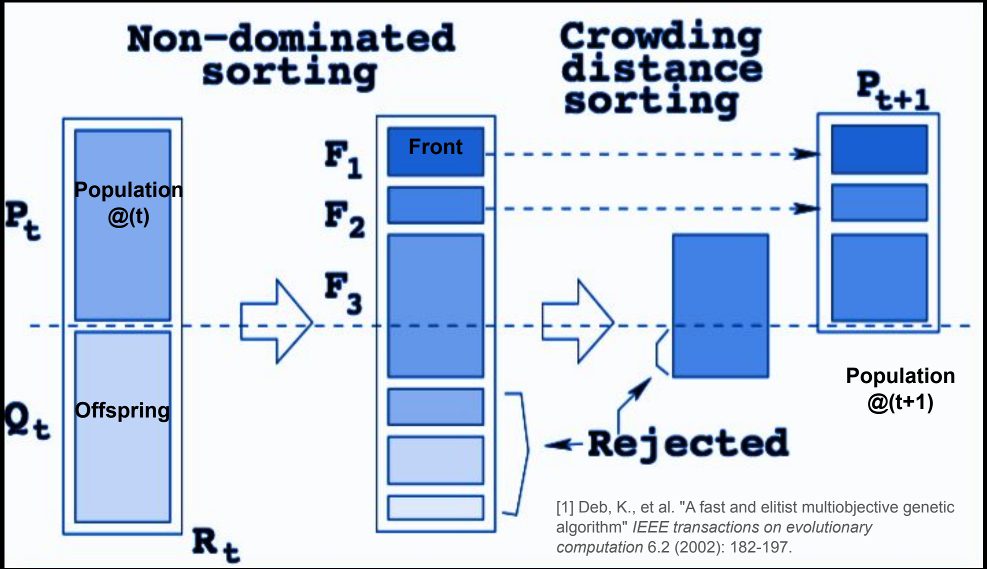
The ECCE Tracker Design Optimization considered simultaneously:

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



Elitist Non-Dominated Sorting Genetic

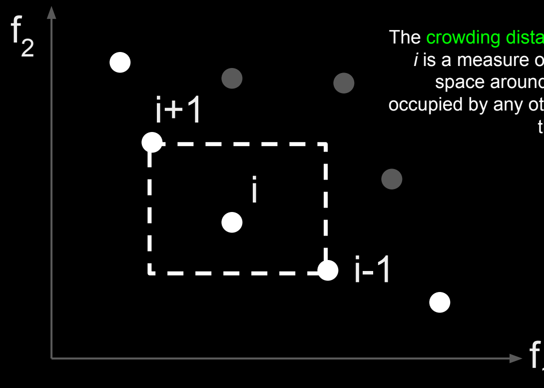




This is one of the most popular approach (>35k citations on google scholar), characterized by:

- 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.



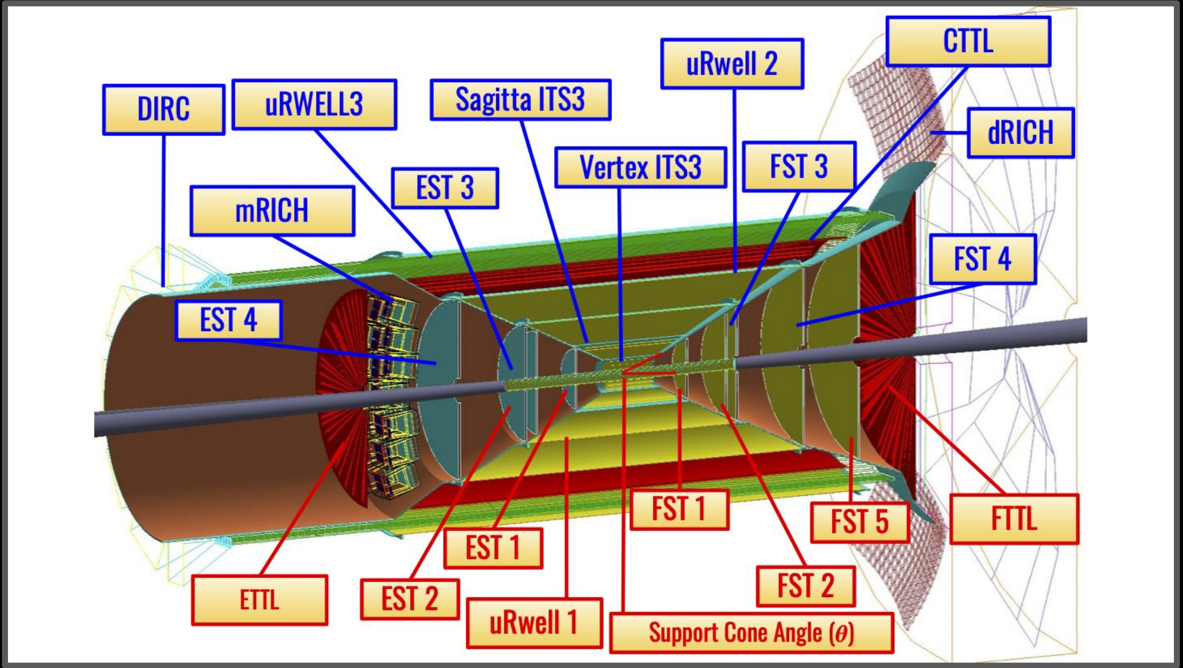
The crowding distance d_i of point *i* is a measure of the objective space around *i* which is not occupied by any other solution in the population.

The EIC Detector Tracker

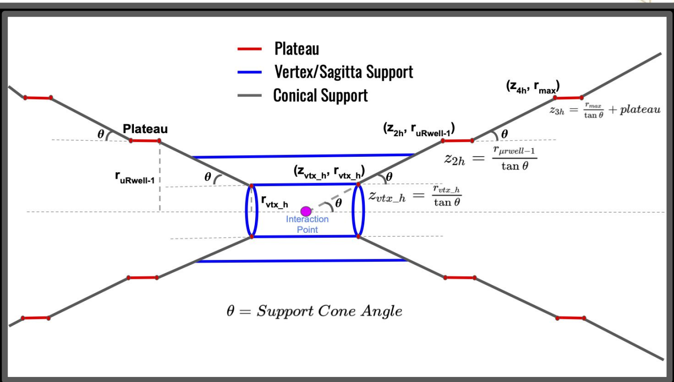
CF, K. Suresh, Z. Papandreou et al (ECCE)

AI-assisted Optimization of the ECCE Tracking System at the Electron Ion Collider

arXiv:2205.09185



Parametrization



see talk by K. Suresh

Non-Projective VS Projective, actually…

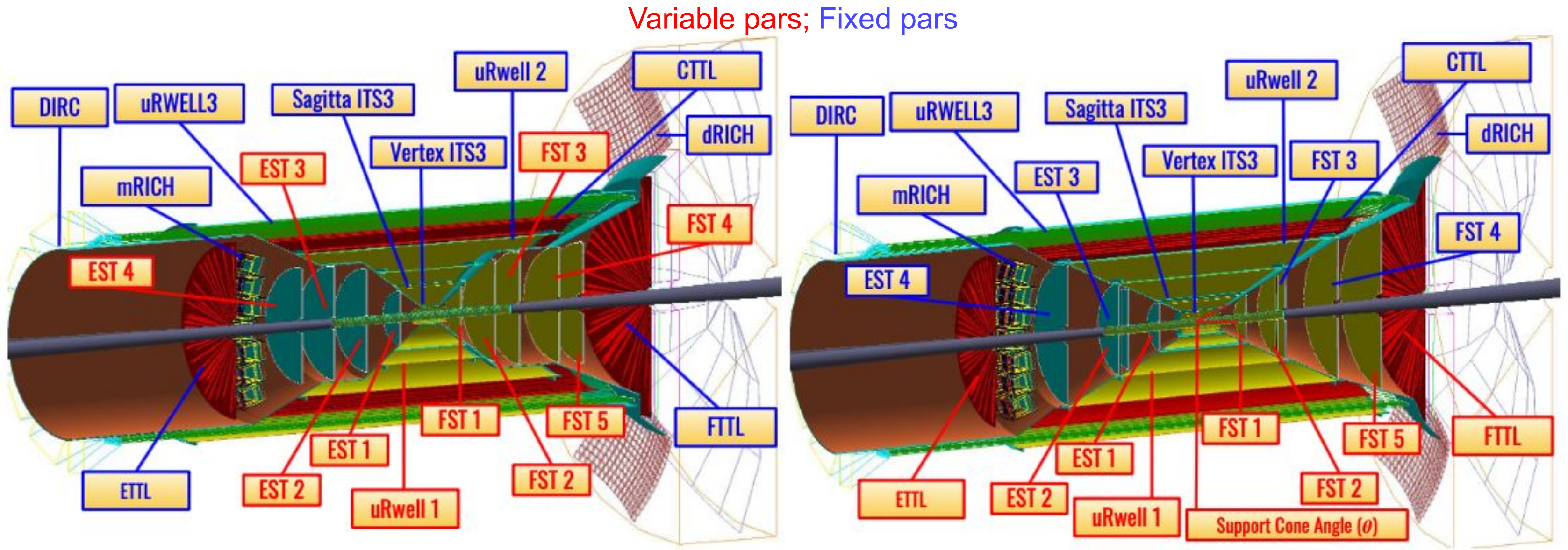
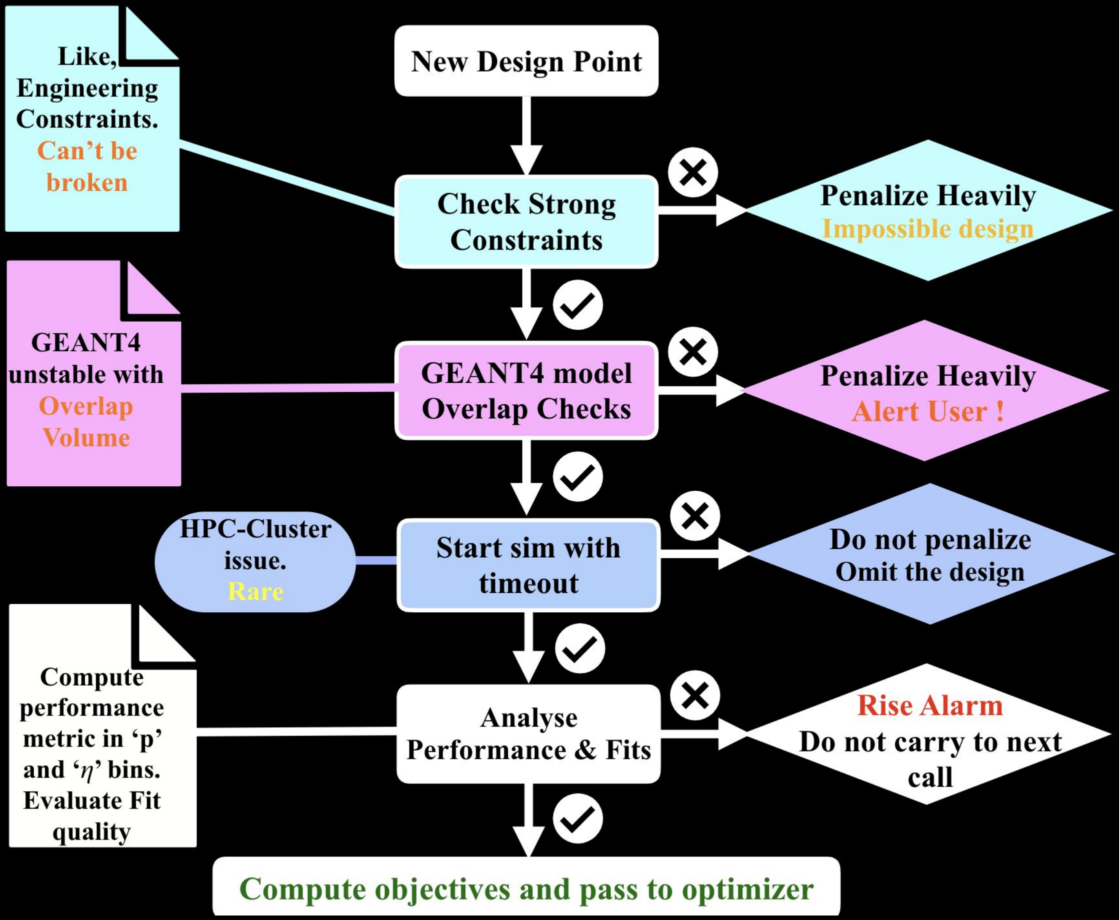


Figure 5: **Tracking and PID system in the non-projective (left) and the ongoing R&D projective (right) designs:** the two figures show the different geometry and parametrization of the ECCE non-projective design (left) and of the ongoing R&D projective design to optimize the support structure (right). Labels in red indicate the sub-detector systems that were optimized, while the labels in blue are the sub-detector systems that were kept fixed due to geometrical constraint. The non-projective geometry (left) is a result of an optimization on the inner tracker layers (labeled in red) while keeping the support structure fixed, The angle made by the support structure to the IP is fixed at about 36.5°. The projective geometry (right) is the result of an ongoing project R&D to reduce the impact of readout and services on tracking resolution.

Constraints, Overlaps, & Other

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

sub-detector	constraint	description
EST/FST disks	$min\left\{\sum_i^{disks}\left\|\frac{R^l_{out}-R^l_{in}}{d}-\left\lfloor\frac{R^l_{out}-R^l_{in}}{d}\right\rfloor\right\|\right\}$	**soft constraint**: sum of residuals in sensor coverage for disks; sensor dimensions: d = 17.8 (30.0) mm
EST/FST disks	$z_{n+1} - z_n >= 10.0$ cm	**strong constraint**: minimum distance between 2 consecutive disks
sagitta layers	$min\left\{\left\|\frac{2\pi r_{sagitta}}{w}-\left\lfloor\frac{2\pi r_{sagitta}}{w}\right\rfloor\right\|\right\}$	**soft constraint**: residual in sensor coverage for every layer; sensor strip width: w = 17.8 mm
μRWELL	$r_{n+1} - r_n >= 5.0$ cm	**strong constraint**: minimum distance between μRwell barrel layers

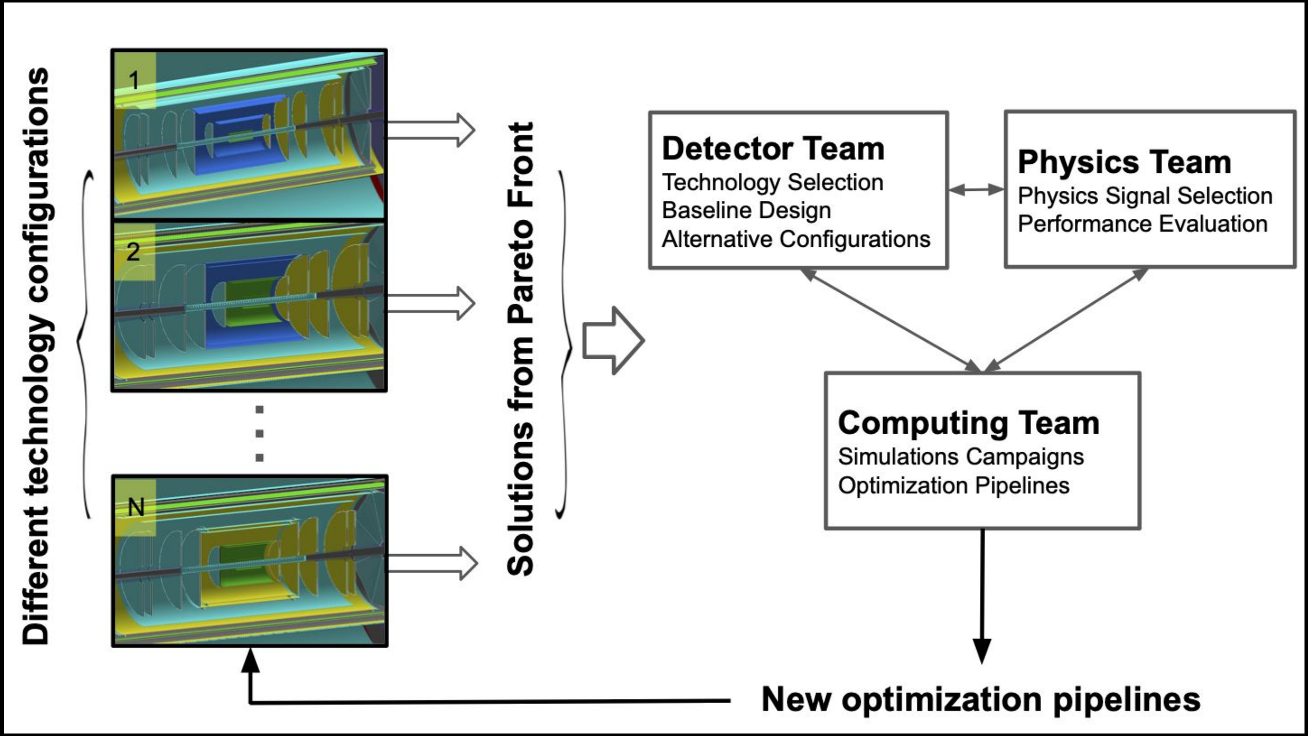


Integration during EIC Detector Proposal

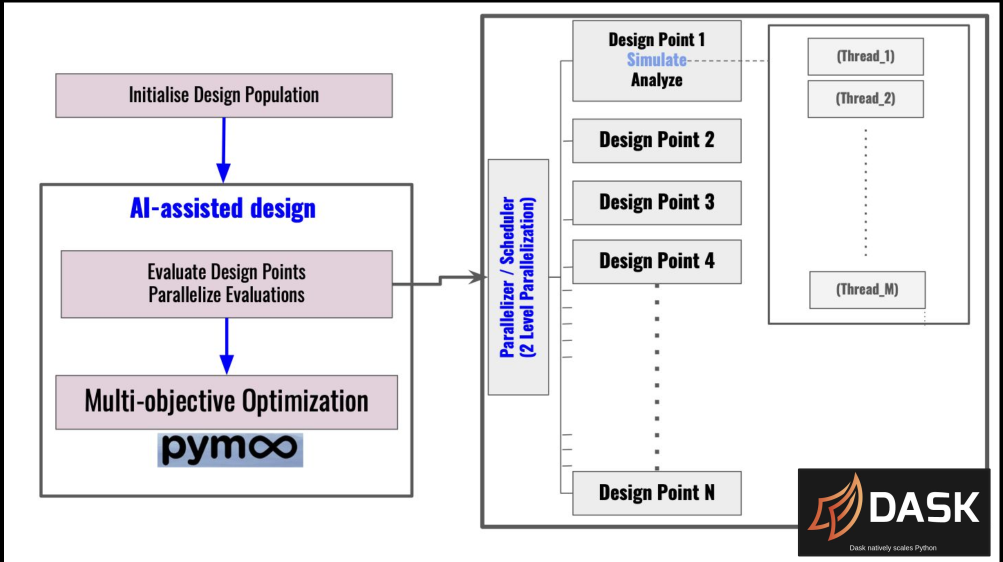
“Optimization” does not mean necessarily “fine-tuning”

Light/smart optimization pipelines ran during the “explorative” phase of the detector proposal

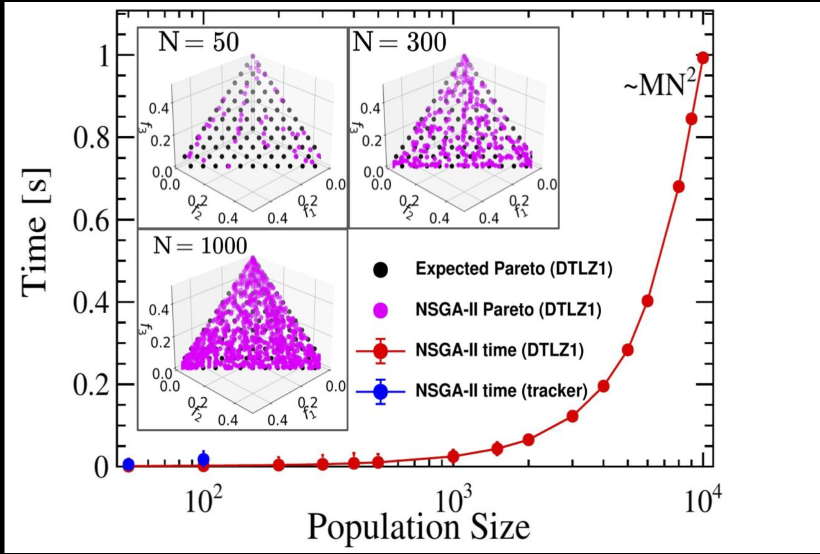
- We want to use these algorithms to: (1) steer the design and suggest parameters that a “manual”/brute-force optimization will likely miss to identify; (2) further optimize some particular detector technology (see d-RICH paper, e.g., optics properties)
- AI allows to capture hidden correlations among the design parameters.
- All “steps” (physics, detector) involved in the AI optimization, strong interplay between working groups



Computational Resources



time taken by GA + sorting



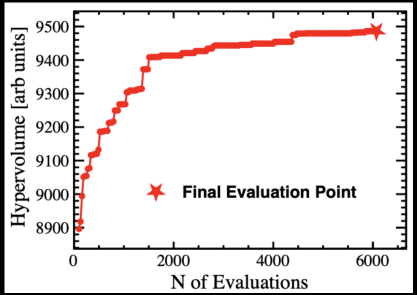
description	symbol	value
population size	N	100
# objectives	M	3
offspring	O	30
design size	D	11 (9)
# calls (tot. budget)	–	200
# cores	–	same as offspring
# charged π tracks	N_{trk}	120k
# bins in η	N_η	5
# bins in p	N_p	10

- Used a test problem DTLZ1
- Verified scaling following MN^2 and convergence to true front
- ~1s/call with 10^4 size!
- For 11 variables and 3 objectives needs ~ 10000 evaluations to converge

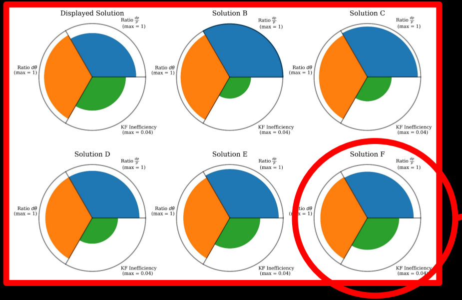
  ~10k CPUhours / pipeline

"Navigate" Pareto Front

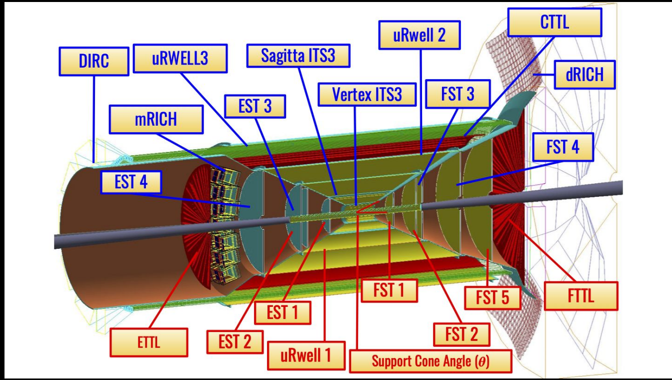
1 Can take a snapshot any time during evaluation



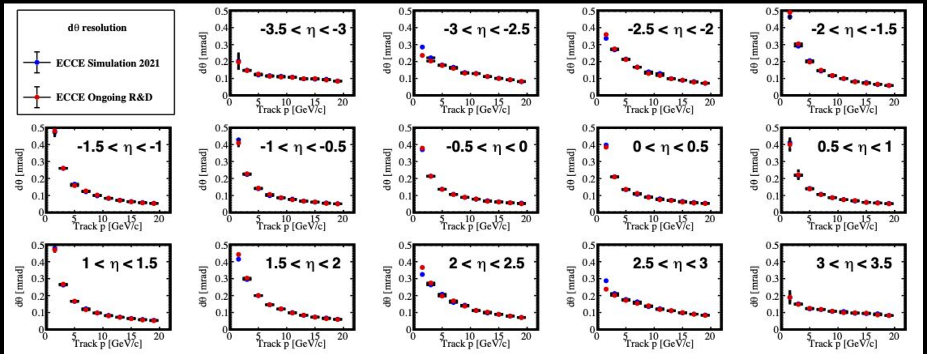
2 Updated Pareto Front at time t

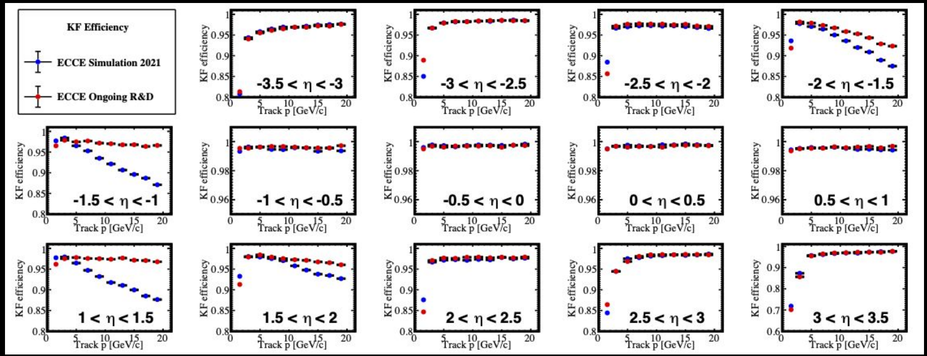


3 At each point in the Pareto front corresponds a design



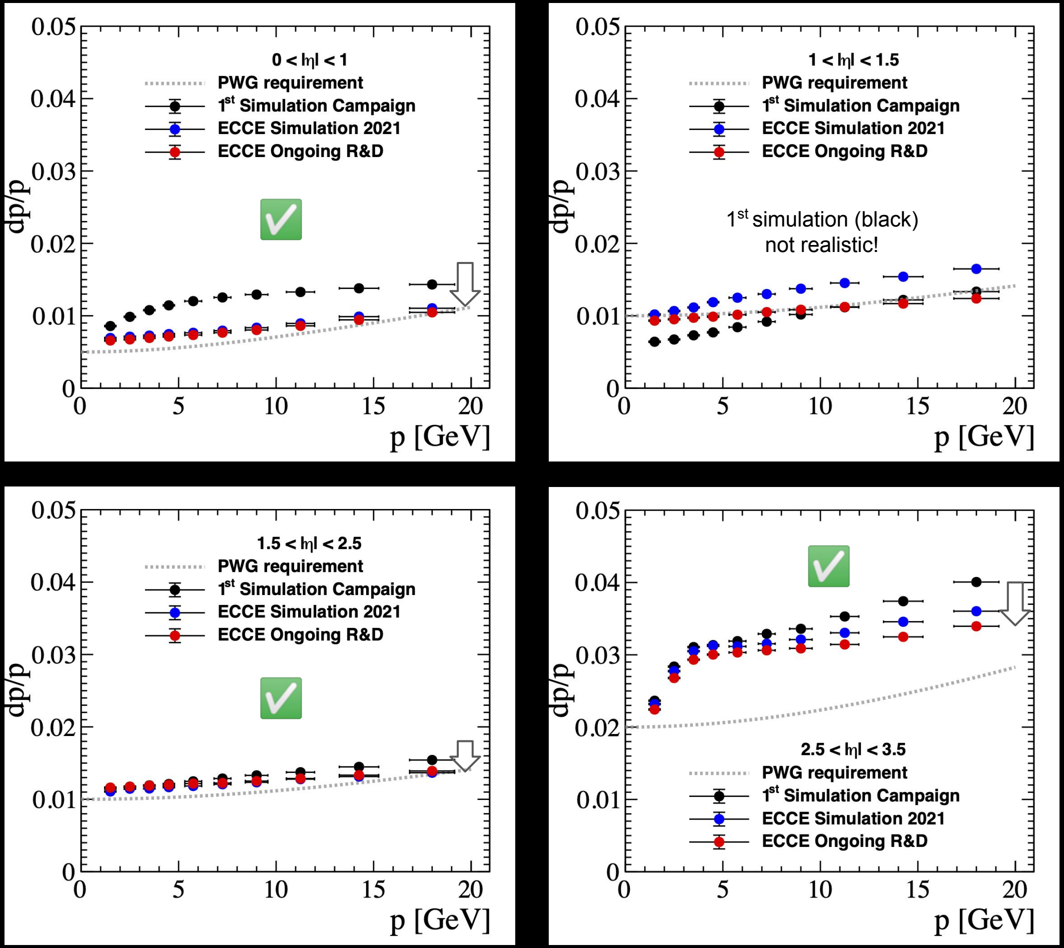
4 Analysis of Objectives (momentum resolution, angular resolution, KF efficiency)





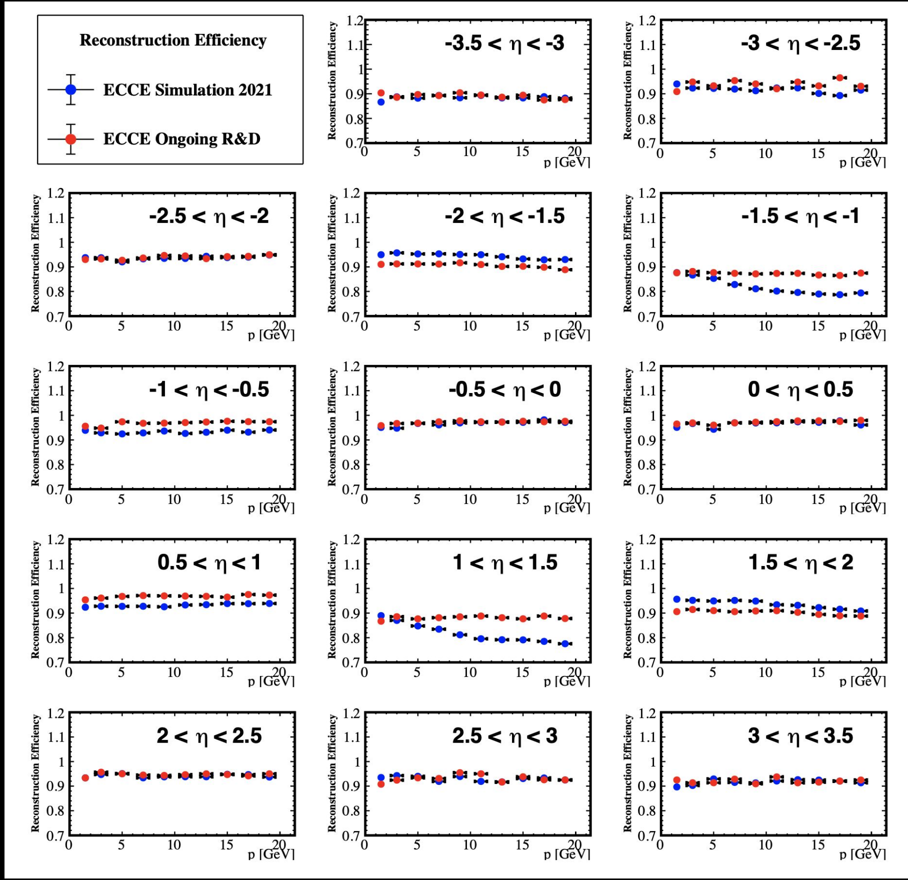
"Evolution"

- Black points represent the first simulation campaign, and a preliminary detector concept in phase-I optimization which did not have a developed support structure;
- Blue points represent the fully developed simulations for the final ECCE detector proposal concept; red points the ongoing R&D for the optimization of the support structure.
- Compared to black, there is an improvement in performance in all η bins with the exception of the transition region, an artifact that depends on the fact that black points do not include a realistic simulation of the material budget in the transition region!
- In the transition region, it can be also appreciated the improvement provided by the projective design



Validation

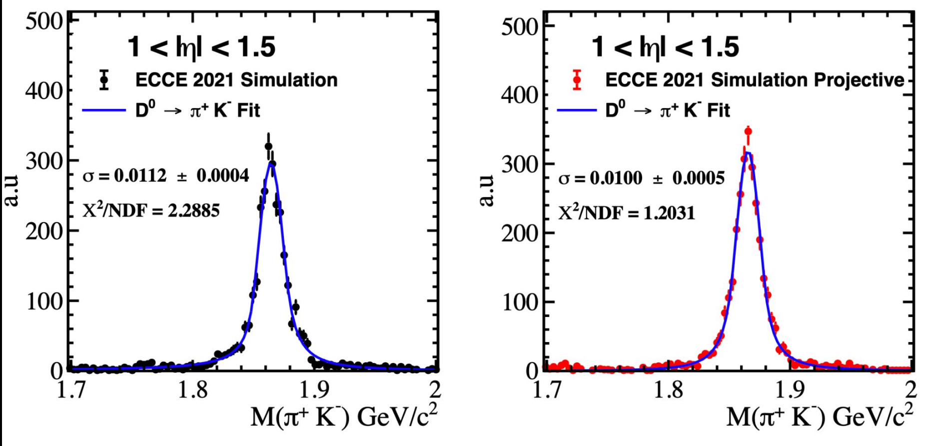
Reconstruction Efficiency



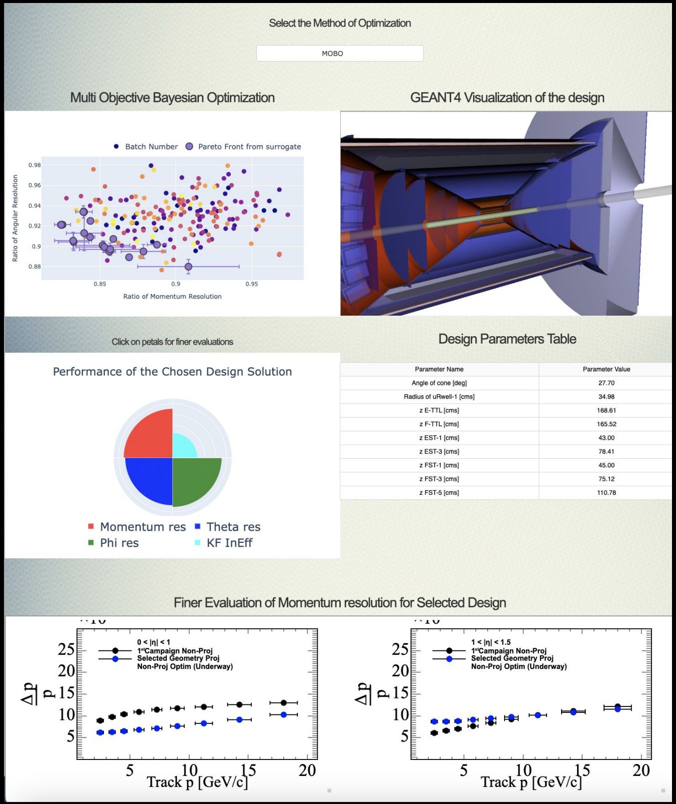
Performance evaluated after optimization process (both designs).

Notice red points are related to an ongoing project R&D with a projective support structure for the ECCE tracker.

D0 invariant mass from semi-inclusive deep inelastic scattering



Navigate interactively

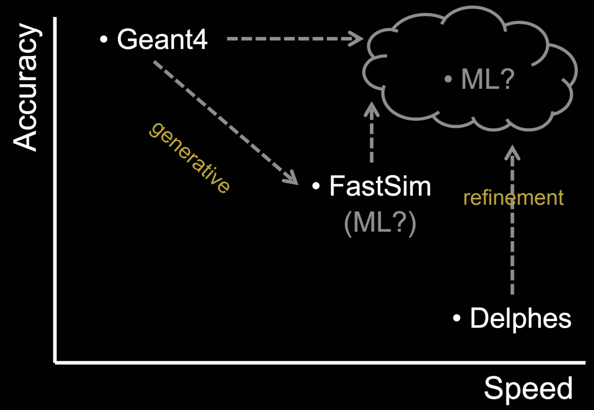


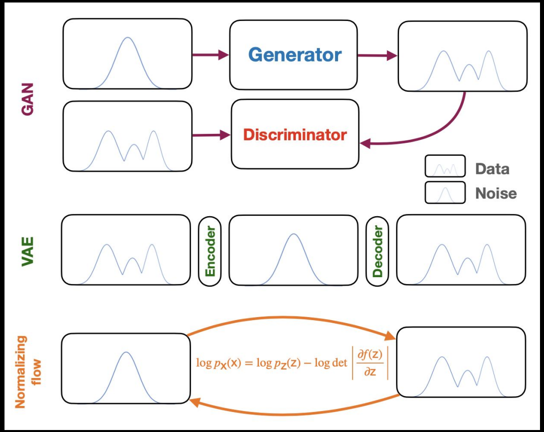
- Visualization of results from approximated Pareto front
- Exploration in a multiple objective space
- Facilitate study/comparison of tradeoff solutions
- Here MOBO is used using BoTorch/Ax (benefit from strong community support — Facebook)

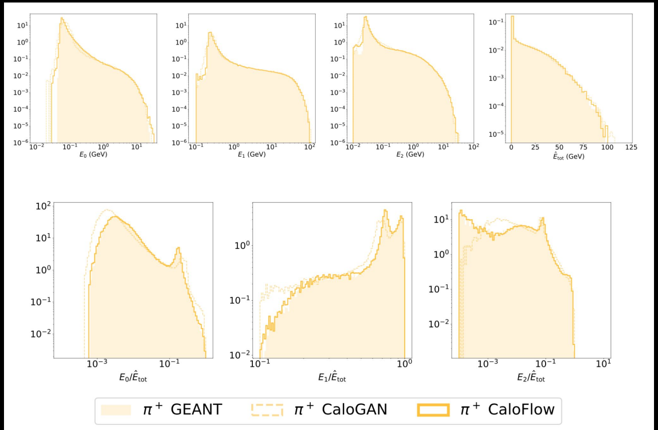
K. Suresh (U. of Regina) https://ai4eicdetopt.pythonanywhere.com

CF, Z. Papandreou, K. Suresh, *Designing EIC with the assistance of AI: strategies and perspectives* (in progress)

ML-“accelerated” Simulations





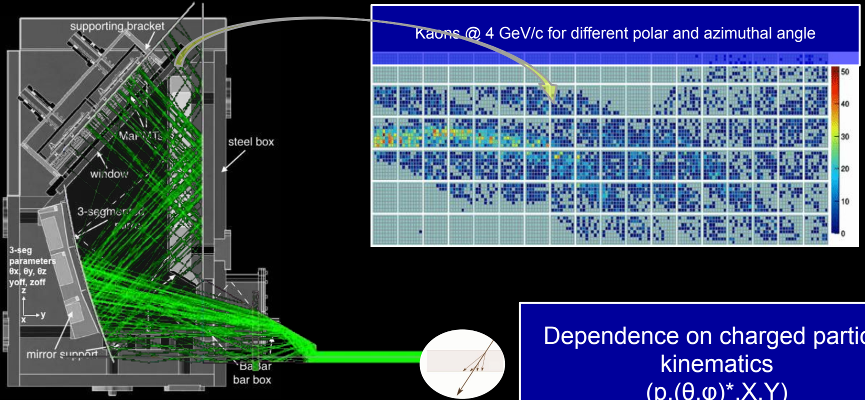


C. Krause, D. Shih, CaloFlow, arXiv:2106.05285

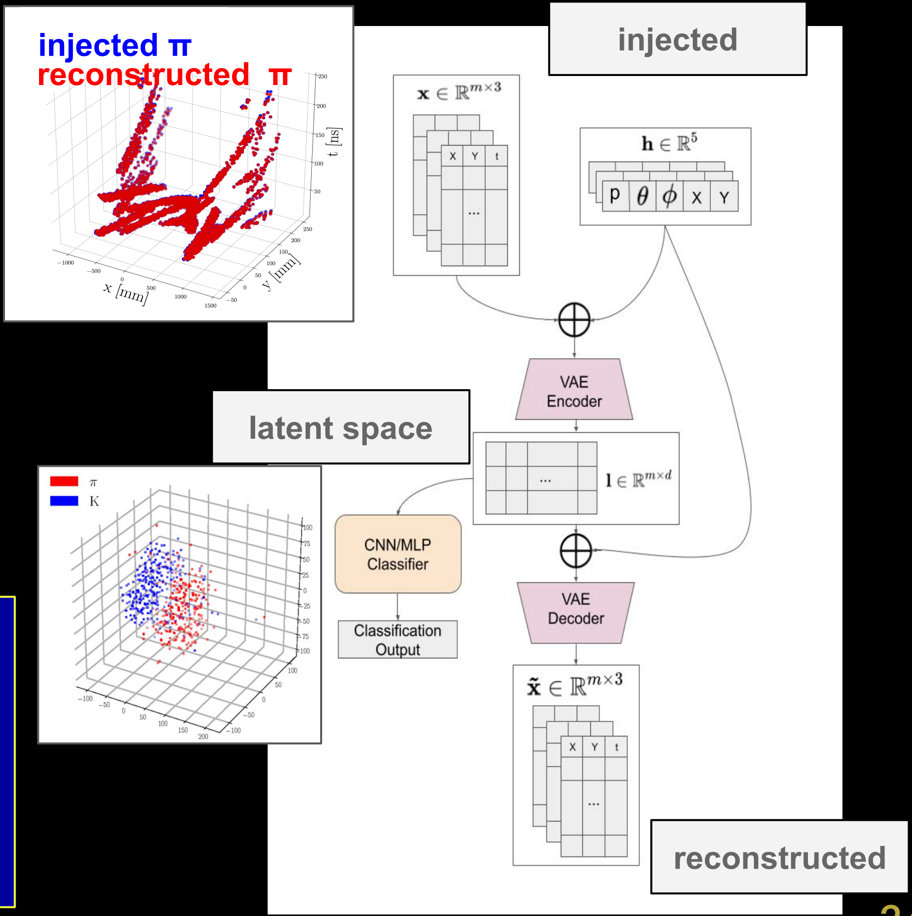
- Computational demands for simulation of current and next generation HEP experiments inspired investigation of surrogates using deep generative models (GAN, VAE, NF based) to decrease simulation time while maintaining fidelity — “real” and “fake” harder to distinguish with NF
- Complex detectors require many fully simulated events as a dataset for the ML architecture
- Notice that a new detector design requires a new dataset…

A. Adelmann et al., New directions for surrogate models and differentiable programming for High Energy Physics detector simulation arXiv:2203.08806v1 and references therein

ML-“accelerated” Sim + Reco



Dependence on charged particle kinematics
(p,(θ,φ)*,X,Y)

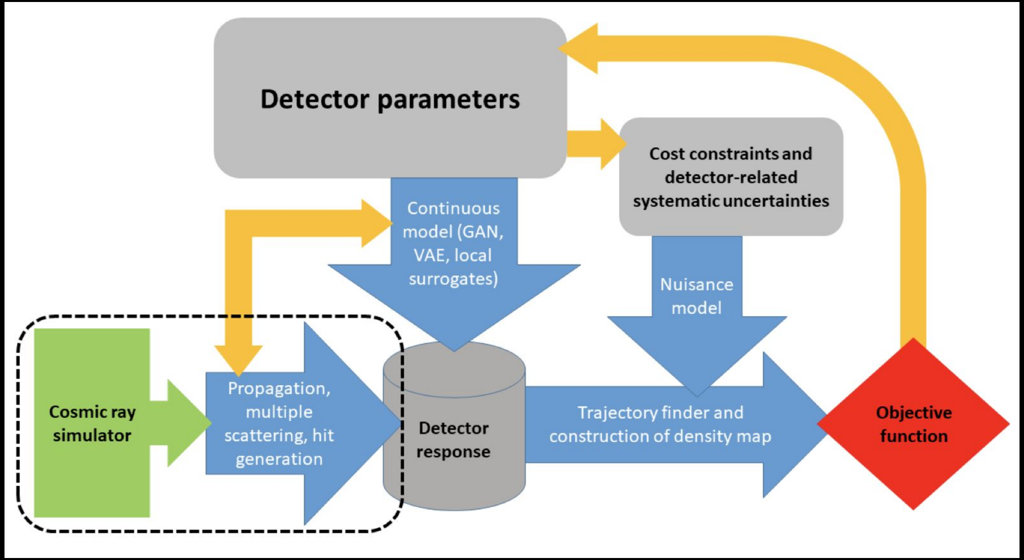


CF and J. Pomponi, DeepRICH, MLST (IOP) 1.1 (2020): 015010

- It is fast* and provides accurate reconstruction
  - 99% FastDIRC; 1us (GPU) vs 1ms (CPU) / particle
- Can be extended to multiple particle types
- Can be generalized to fast simulation
- Can utilize (x,y,t) patterns if time is measured
- Can deal with different topologies and detectors
- Deeply learns the detector response (real data can be injected)

ML Optimized Design of Experiments – MODE

- Detectors design with AI is gaining a lot of interest.
- MODE is a recently formed collaboration of physicists and computer scientists who target the use of differentiable programming in design optimization of detectors for particle physics applications
- Ambitious project: develop a modular, customizable, and scalable, fully differentiable pipeline for the end-to-end optimization of articulated objective functions that model in full the true goals of experimental particle physics endeavours, to ensure optimal detector performance, analysis potential, and cost-effectiveness.



Conceptual layout of an optimization pipeline for a muon radiography apparatus.

An end to end optimization requires modeling of simulations. Requires collect reference data to train the surrogate models ML implementations.

A. G. Baydin et al. Nuclear Physics News 31.1 (Mar 30, 2021): 25-28.

AI4EIC

First Workshop on September 2021 – [JINST proceedings]

Formation of EICUG AI WG (a.k.a. AI4EIC) early 2022 https://eicug.github.io/

AI4EIC website https://eic.ai

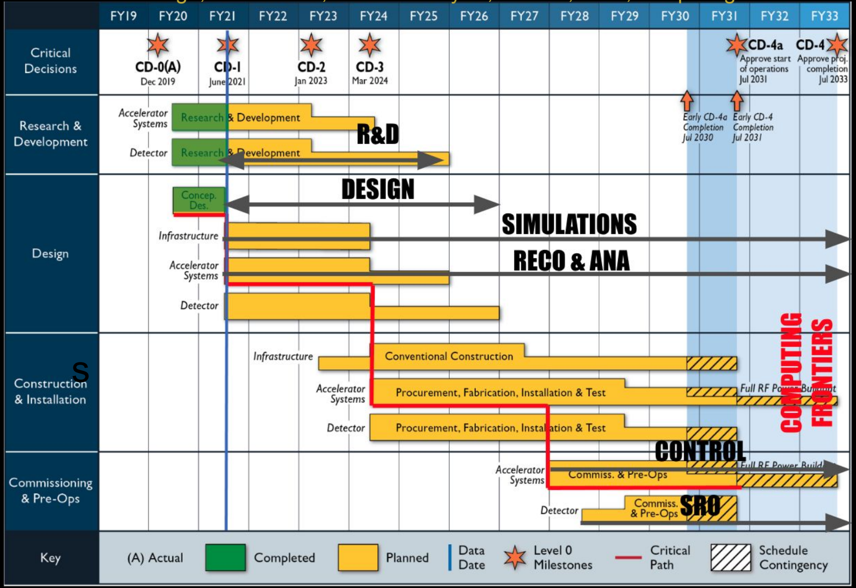
Next meeting: topic-oriented on UQ

Next workshop on October 10-14 2022 at W&M

https://eic.ai/workshop

AI4EIC Workshops structured to reflect the EIC schedule

6 sessions: Design, Simulations, Reco & Analysis, Control, SRO, computing frontiers



AI Community @ EIC

- AI4EIC Workshops
- Tutorials
- Schools
- Jamboree
- Hackathons
- Kaggle Challenges
- Outreach

It may develop “sub-WG” groups

(From the AI4EIC Workshop):

- AI for EIC Design*
- AI for EIC (Fast) Simulations
- AI for EIC Data Reco & Analysis
- AI for EIC Control*: automated workflows; data quality monitoring; anomaly detection
- AI for EIC Streaming Readout
- AI for EIC Computing frontiers
- AI for EIC theory; phenomenology;

(From our meetings):

- Additional areas

AI4EIC survey form: https://forms.gle/6LADKTGaX7DeTVE46

Results preview: https://indico.bnl.gov/event/15636/

*Schedule presented @ AI4EIC in 2021

Summary

One of the conclusions from the DOE Town Halls on AI for Science on 2019 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"*.

- AI can assist the design and R&D of complex experimental systems by providing more efficient design (considering multiple objectives) and optimizing the computing budget needed to achieve that.
- EIC is one of the first experiments to be designed with the support of AI (already since 2020 with dRICH design and during detector proposal for the tracker — See K. Suresh talk).
    - Roughly 1M CPU-core hours/year are anticipated for these studies (which will be extended to include PID detectors, e.g., the dRICH) for detector-1.
- Cherenkov detectors are the backbone of PID at EIC. Need for fast simulations and fast reconstruction/pattern recognition; generally, AI/ML for SRO-related activities. Pivotal for EIC.

None ever accomplished a multi-dimensional / multi-objective optimization of the global design

Costs can be explicitly included during the optimization provided a reliable parametrization)

An intrinsic overhead regards compute expensive simulations + reconstruction/analysis.

Larger populations to improve accuracy of the Pareto front

Just few AI/ML applications for Cherenkov, particularly utilizing low-level features. A lot to be done!

Possibility to leverage advancements in ML implemented on heterogeneous computing architectures.

Spares