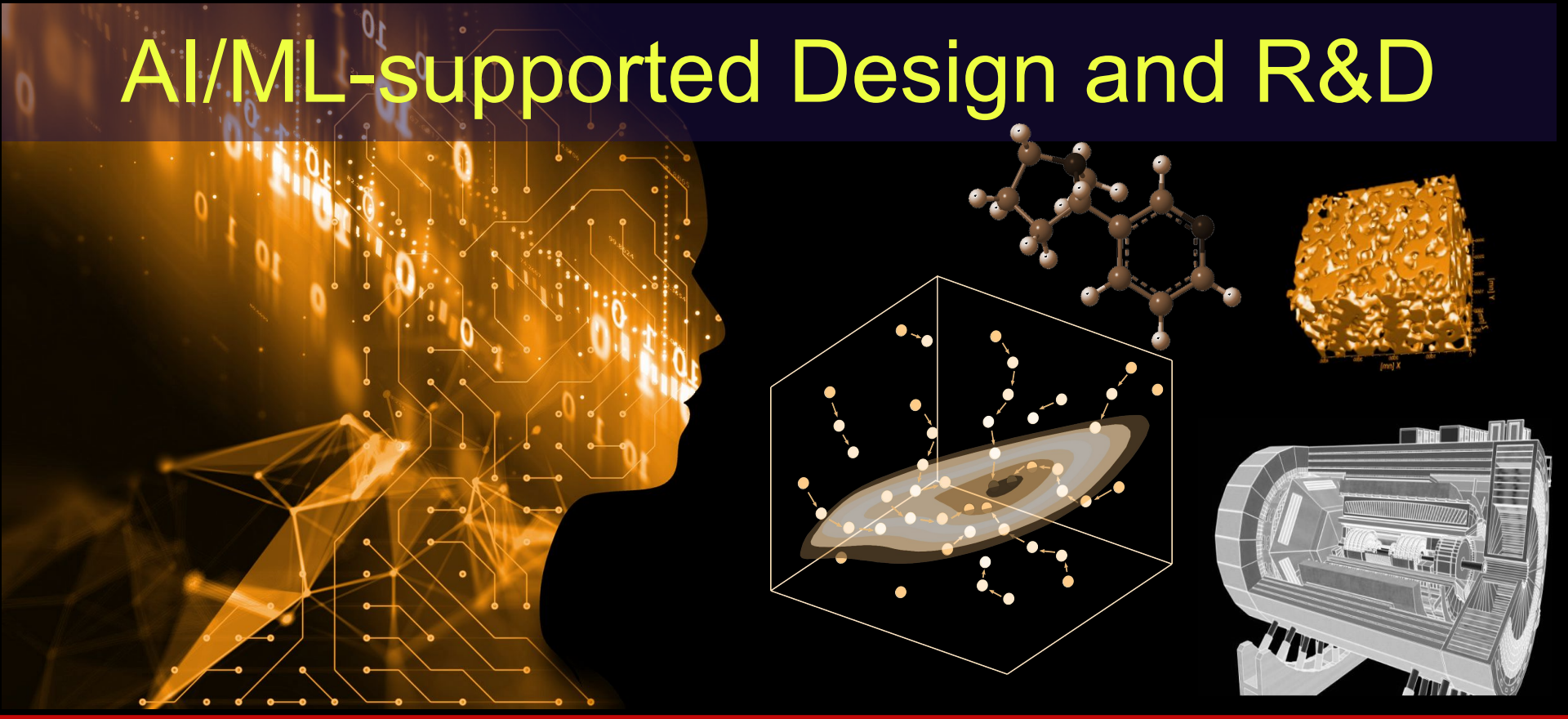
AI/ML-supported Design and R&D



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;[60,161] design of modular metamaterials[162]
Reinforcement learning	Maximize cumulative awards with agents reacting to the environments.	Deriving microstructure-based traction-[illegible]ion 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]

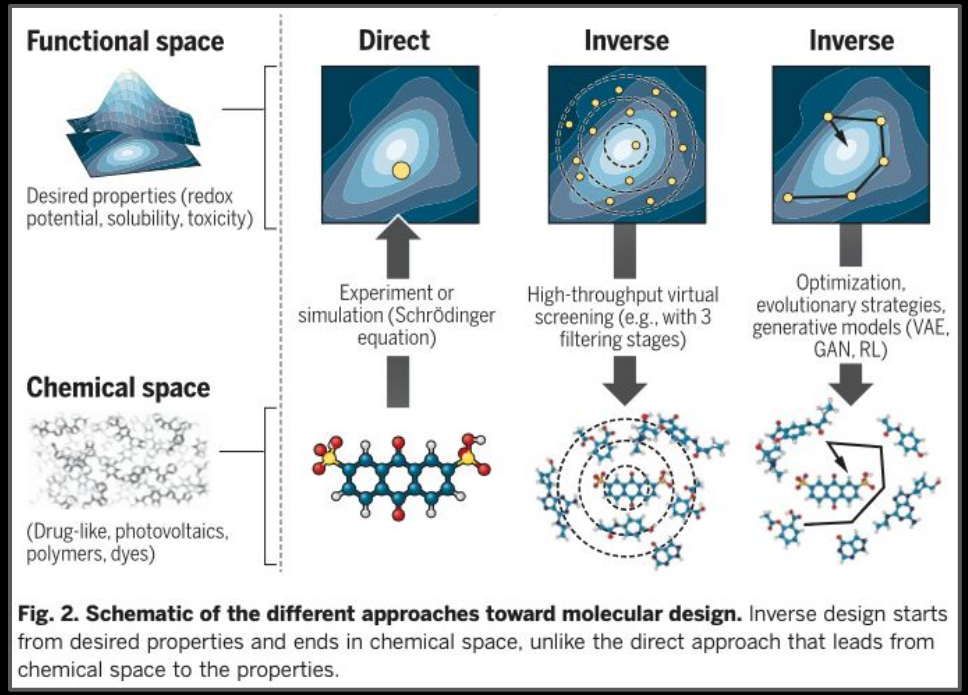


Fig. 2. Schematic of the different approaches toward molecular design. Inverse design starts from desired properties and ends in chemical space, unlike the direct approach that leads from chemical space to the properties.

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

AI for Experimental Design in NP/HEP

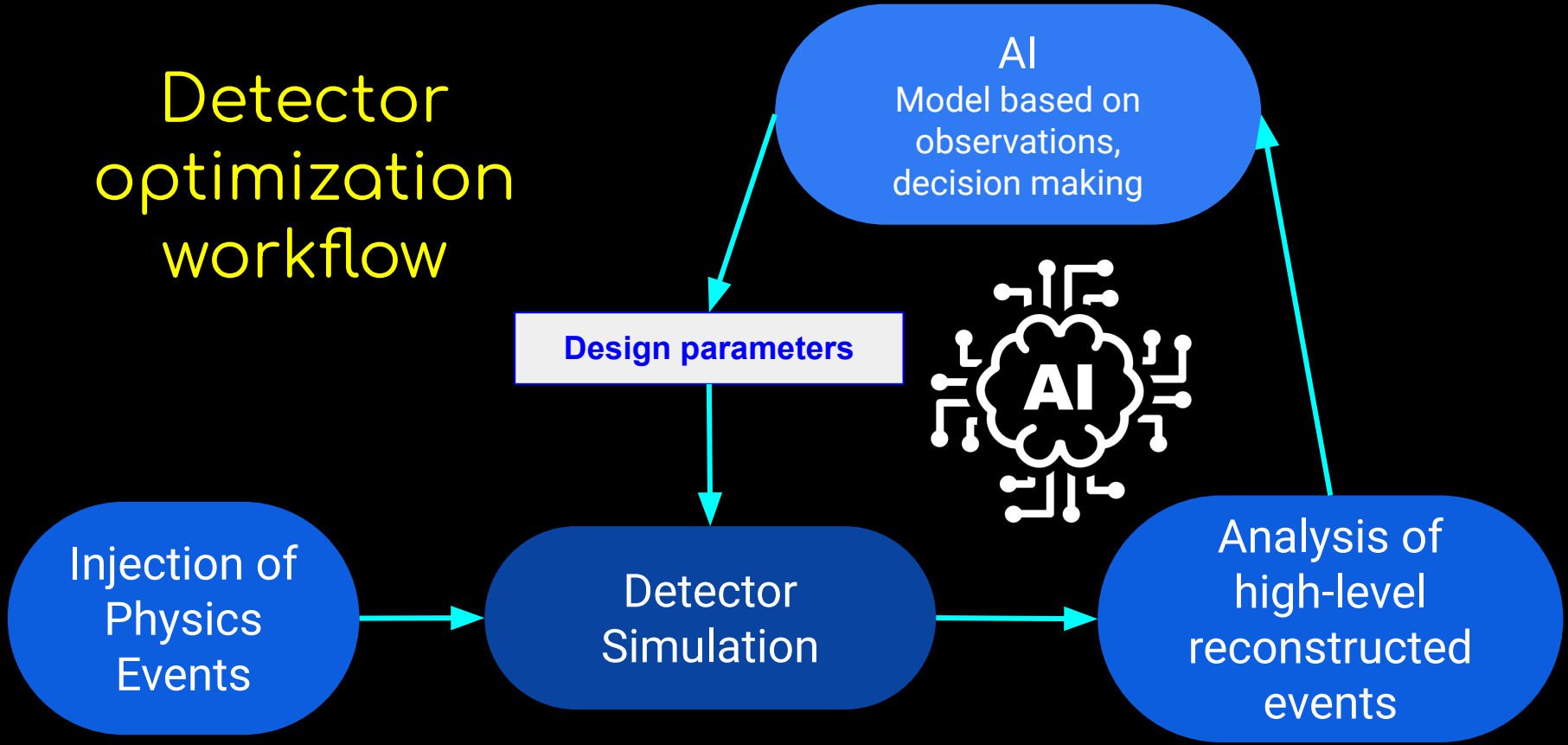
- When it comes to designing detectors and accelerators with AI this is an area at its “infancy”. What follows uses “detector” as example but applies to both detector and accelerator.
- Typically full detector design is studied once the subsystem prototypes are ready (phase constraints from the full 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

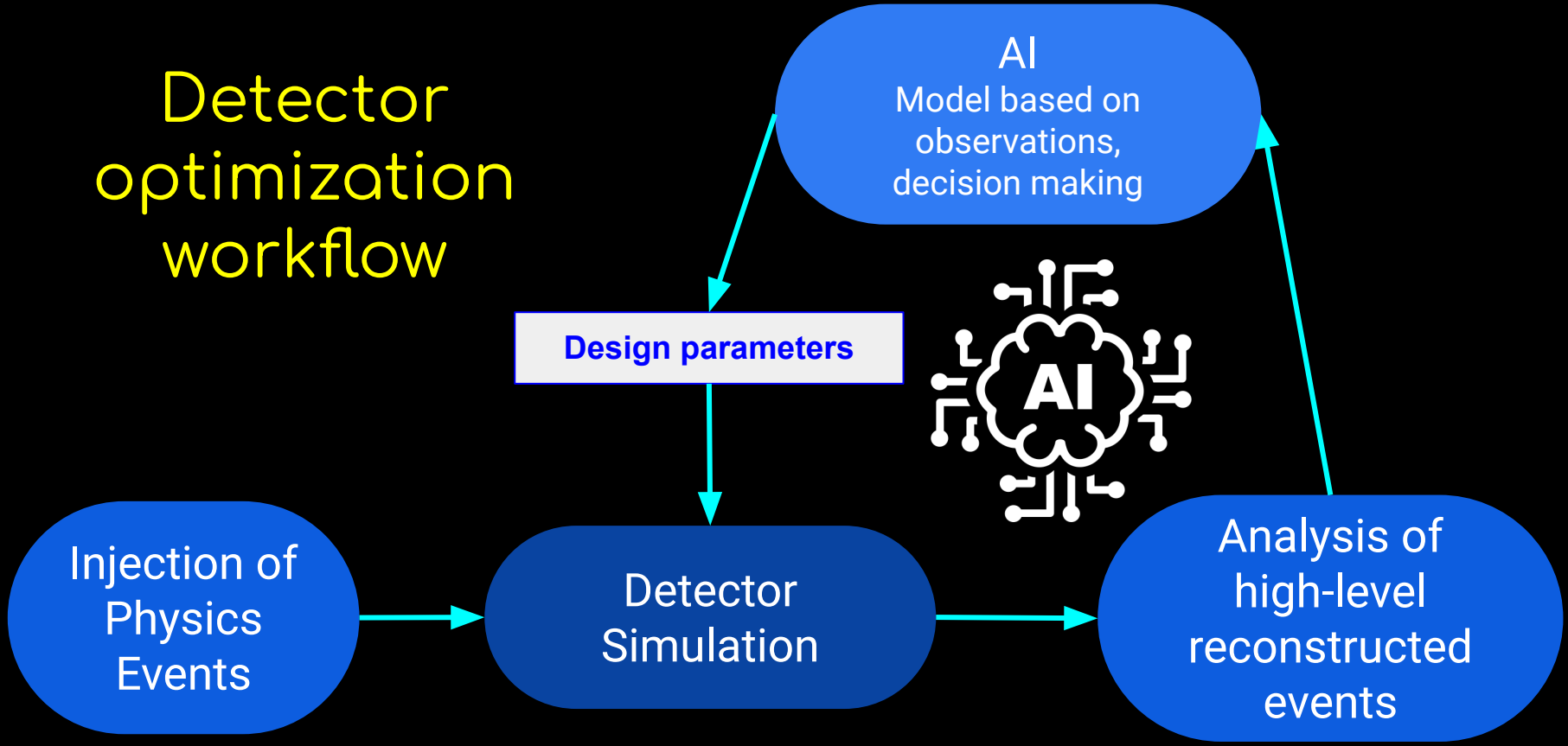
What follows largely based on a series of lectures on Detector Design with AI at the AI4NP Winter School

[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

Detector optimization workflow



Detector optimization workflow



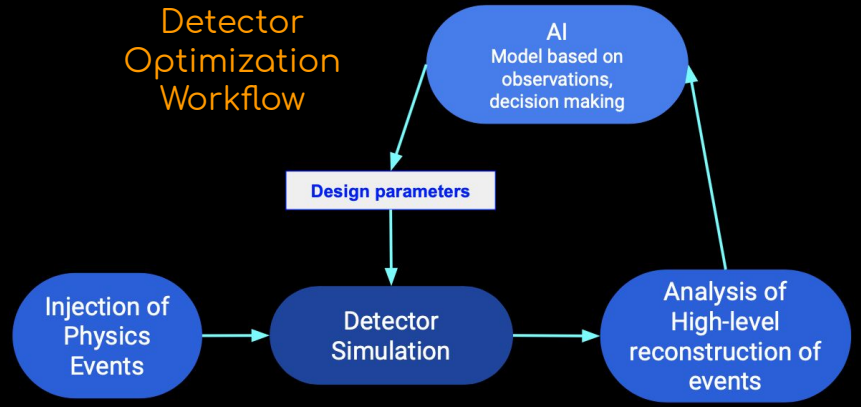
See Session on Simulations this afternoon

See Session on Reconstruction & Analysis on Wed, Sep 8

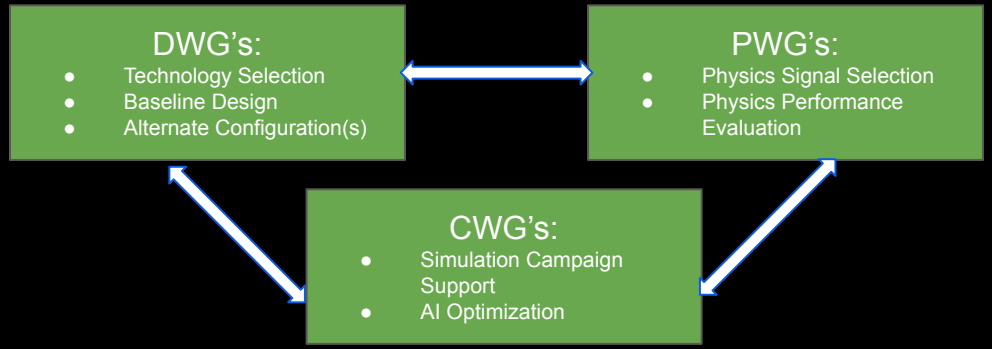
Why Design with AI now?

Optimization does not mean necessarily “fine-tuning”

- 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

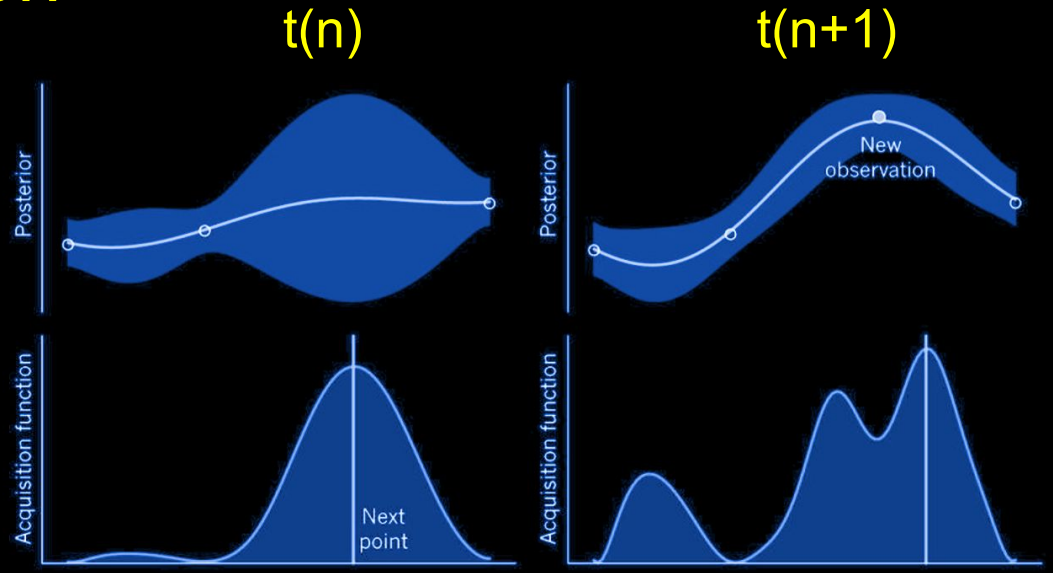


AI promotes Interaction among Working Groups



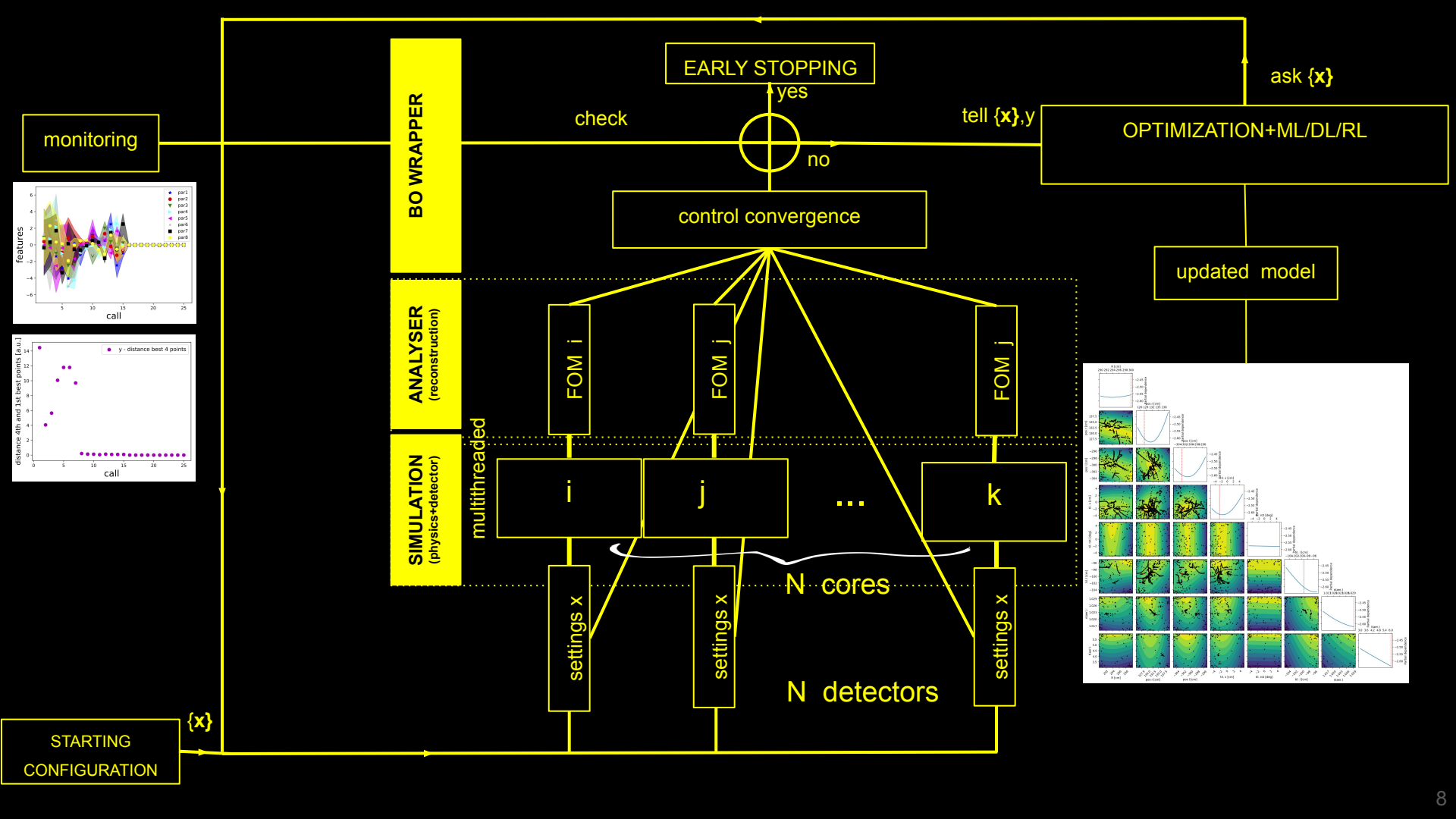
Bayesian Optimization

- 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**.



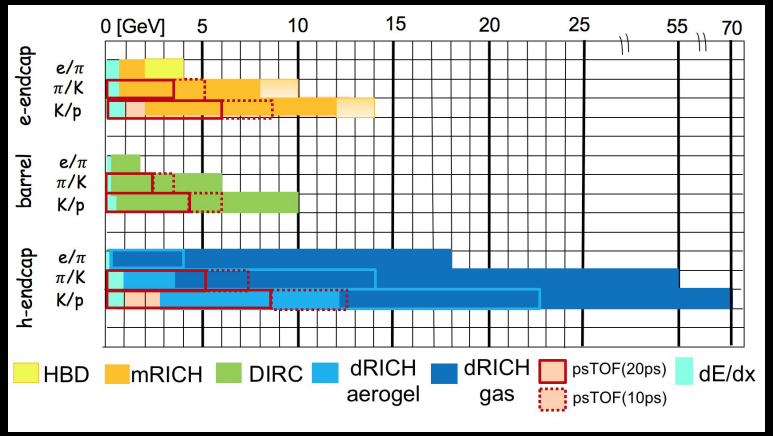
1. Select a Sample by Optimizing the Acquisition Function.
2. Evaluate the Sample With the Objective Function.
3. Update the Data and, in turn, the Surrogate Function.
4. Go To 1.

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


monitoring
BO WRAPPER
EARLY STOPPING
yes
check
tell {x},y
ask {x}
OPTIMIZATION+ML/DL/RL
no
control convergence
updated model
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
{x}
STARTING
CONFIGURATION
features
call
distance 4th and 1st best points [a.u.]
y - distance best 4 points
call

Dual RICH: case study

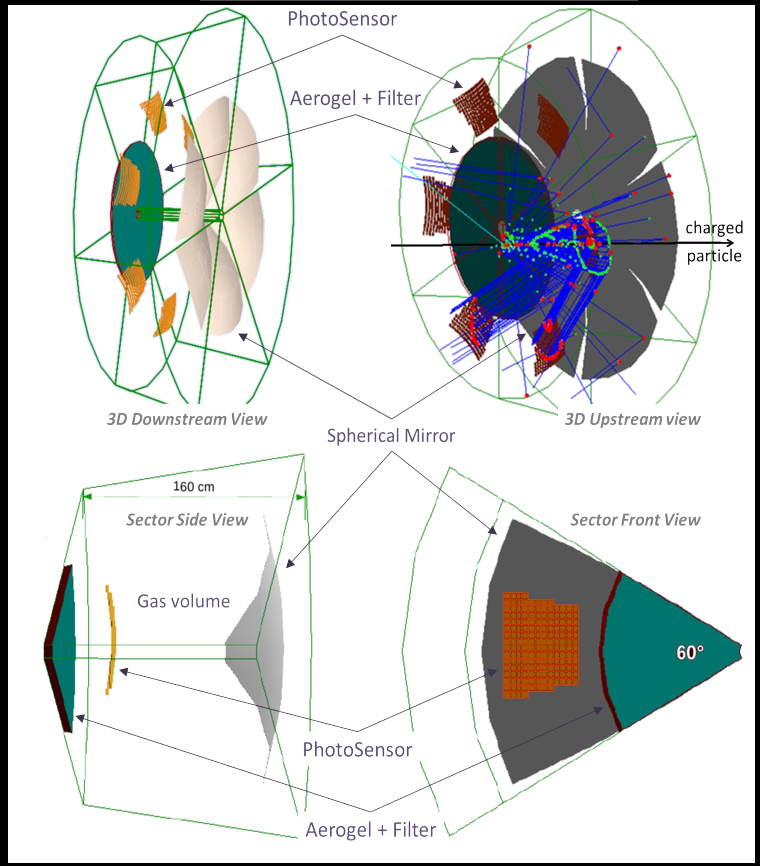
E. Cisbani, A. Del Dotto, CF*, M. Williams et al.
"AI-optimized detector design for the future Electron-Ion Collider: the dual-radiator RICH case."
Journal of Instrumentation 15.05 (2020): P05009.



- 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

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

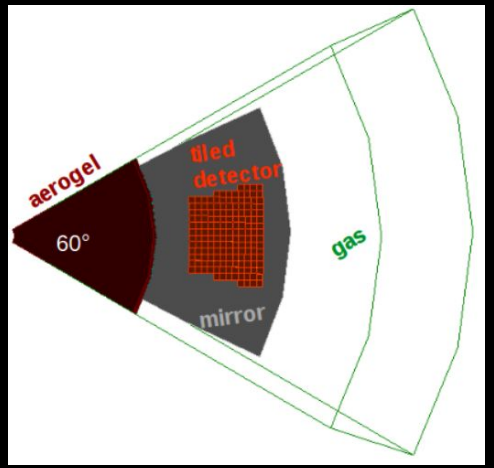


Construction Constraints

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

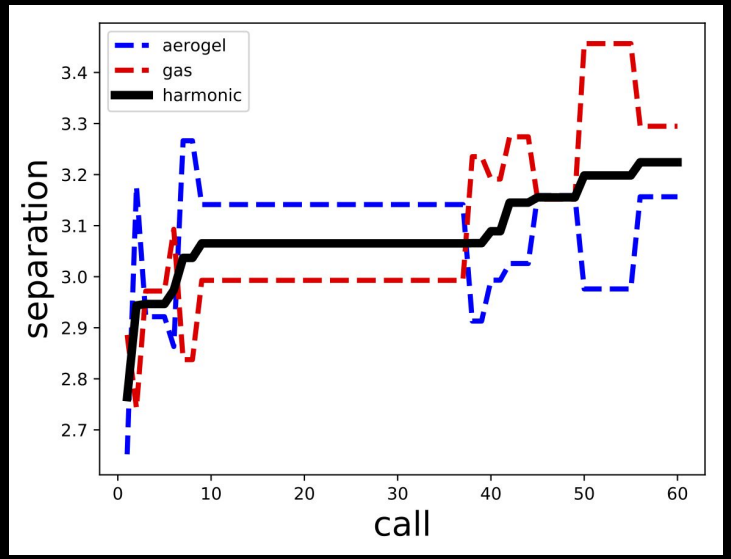
parameter	description	range [units]	tolerance [units]
R	mirror radius	[290,300] [cm]	100 [μm]
pos r	radial position of mirror center	[125,140] [cm]	100 [μm]
pos l	longitudinal position of mirror center	[-305,-295] [cm]	100 [μm]
tiles x	shift along x of tiles center	[-5,5] [cm]	100 [μm]
tiles y	shift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

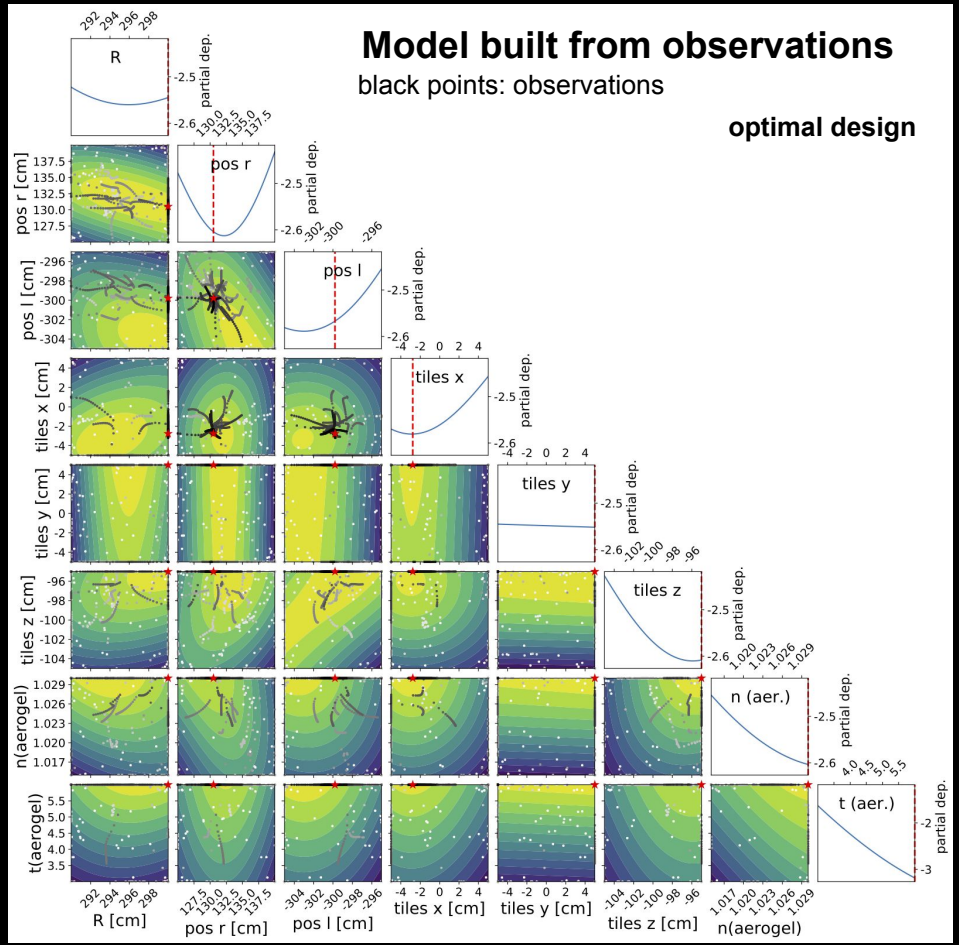


Ranges depend mainly on mechanical constraints and optics requirements.
These requirements can change in the next future based on inputs from prototyping.

The Model and the Optimized FoM

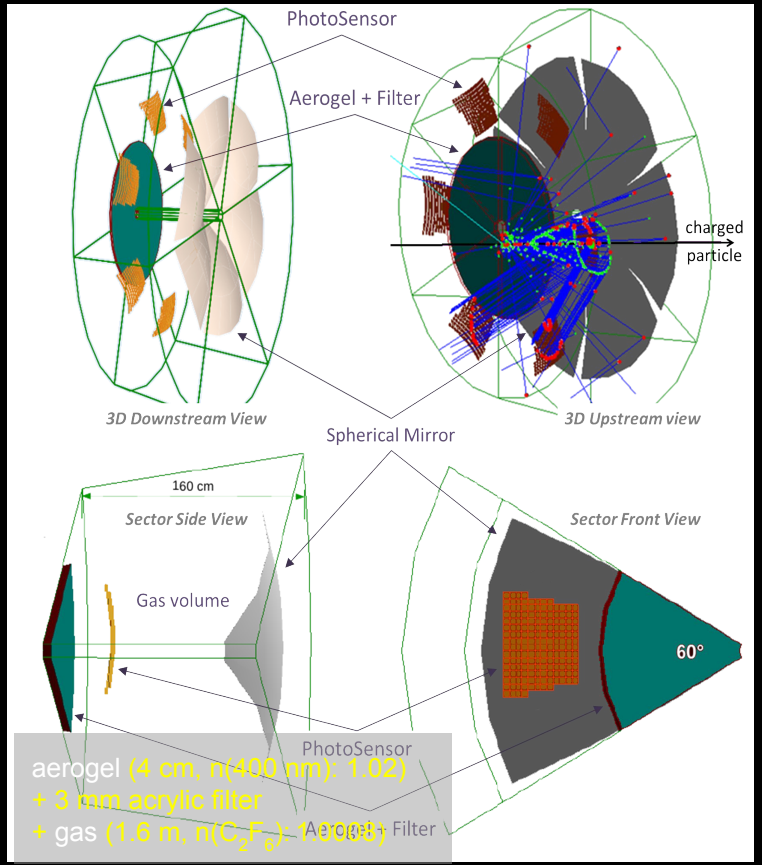
$$N\sigma = \frac{||\langle\theta_K\rangle - \langle\theta_\pi\rangle||\sqrt{N_\gamma}}{\sigma_\theta^{1p.e.}}$$

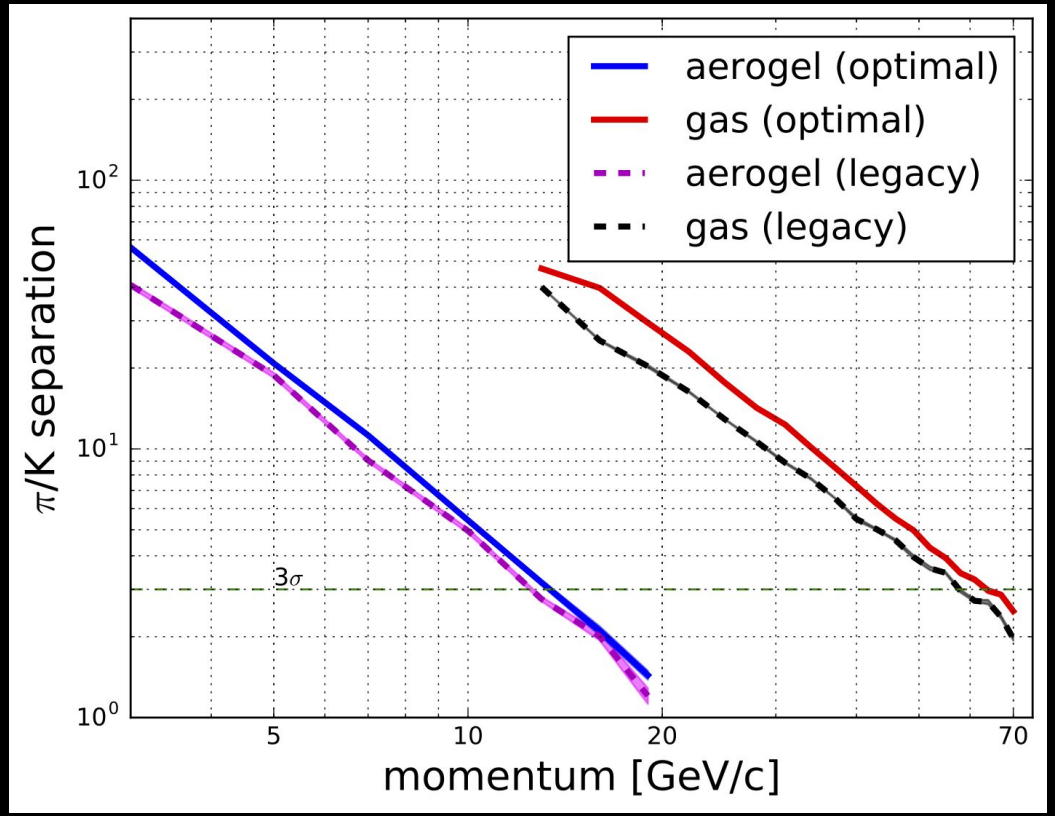




AI-Optimized dRICH

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

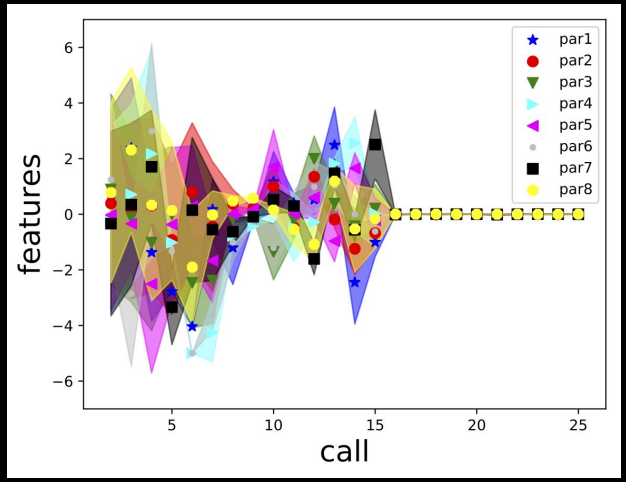


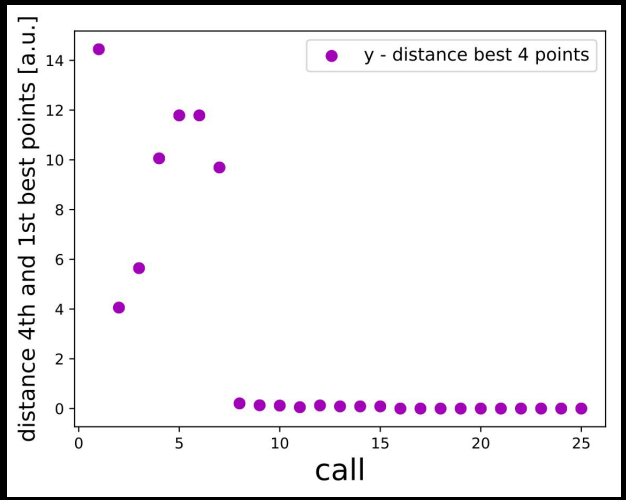


- Statistically significant Improvement in both parts.
- In particular in the gas region where the 5σ threshold shifted from 43 to 50 GeV/c and the 3σ one extended up to
- Notice that before this study we did not know “how well” the legacy design was performing.

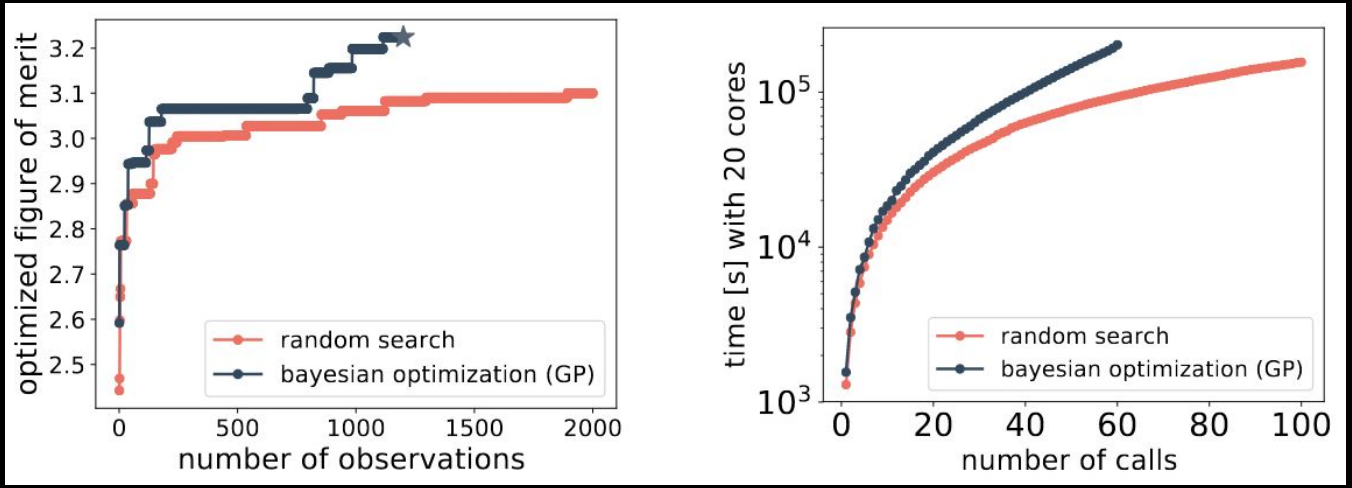
Convergence Criteria

- 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.





Comparison with Random Search



Each call:
400 tracks generated/core
20 cores

1 design point ~ 10 mins/CPU

Budget: 100 calls

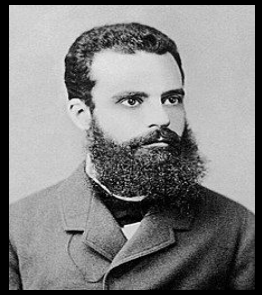
- BO with GP scales cubically with number of observations.
- Bayesian optimization methods are more promising because they offer principled approaches to weighting the importance of each dimension.
- For this 8D problem - even with 50 cores, RS looks unfeasible due to the curse of dimensionality.
  - 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.

Multiple Objectives!

[1] Deb, Kalyanmoy. "Multi-objective optimisation using evolutionary algorithms: an introduction." *Multi-objective evolutionary optimisation for product design and manufacturing*. Springer, London, 2011. 3-34.

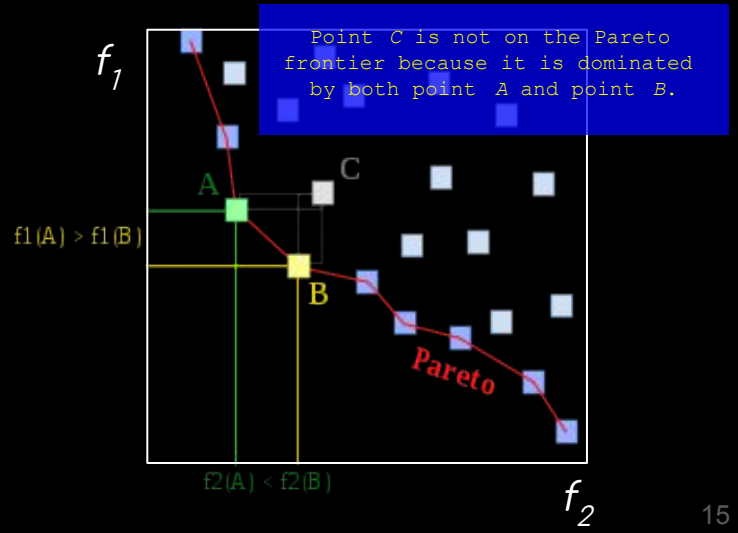
- The problem becomes challenging when the objectives are of conflict to each other, that is, the optimal solution of an objective function is different from that of the other.
- In solving such problems, with or without constraints, they give rise to a trade-off optimal solutions, popularly known as Pareto-optimal solutions.



V. Pareto, 1848-1923

- 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 [1].



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.

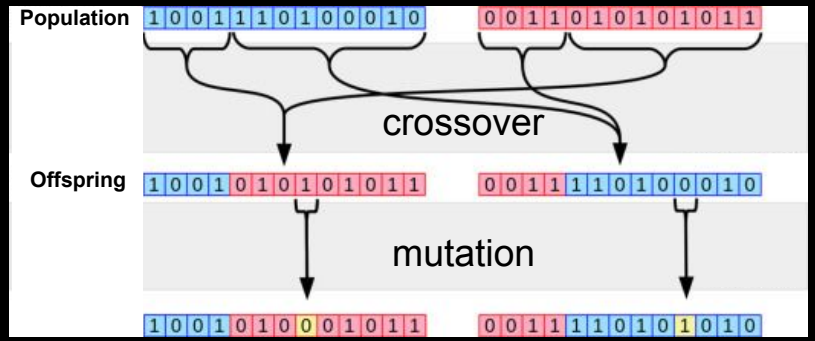
[5] M. Balandat, B. Karrer, D. R. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, "Botorch: Programmable bayesian optimization in pytorch," arXiv preprint arXiv:1910.06403, 2019.

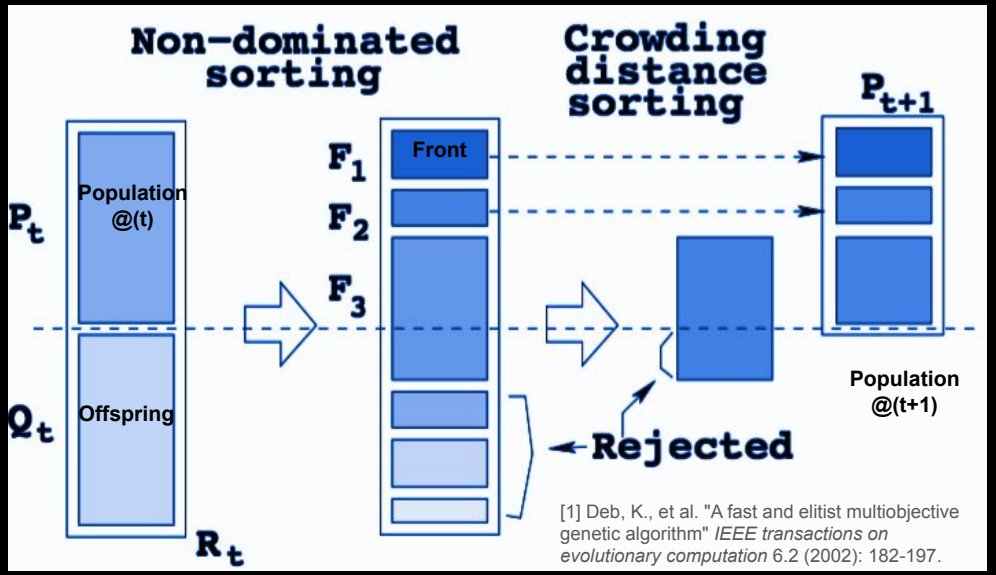
[6] P. P. Galuzio, E. H. de Vasconcelos Segundo, L. dos Santos Coelho, and V. C. Mariani, "MOBOpt—multi-objective Bayesian optimization," SoftwareX, vol. 12, p. 100520, 2020.

[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

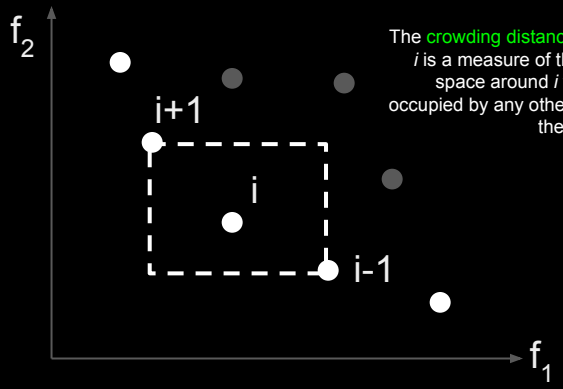




This is one of the most popular approach (>35k citations on google scholar), characterized by:

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

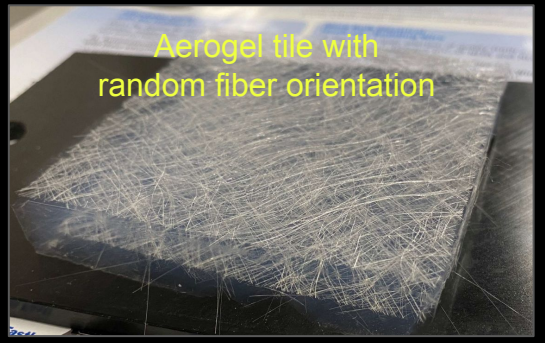
The population R_t is classified in non-dominated fronts. Not all fronts can be accommodated in the N slots of available in the new population P_{t+1}. We use crowding distance to keep those points in the last front that contribute to the highest diversity.



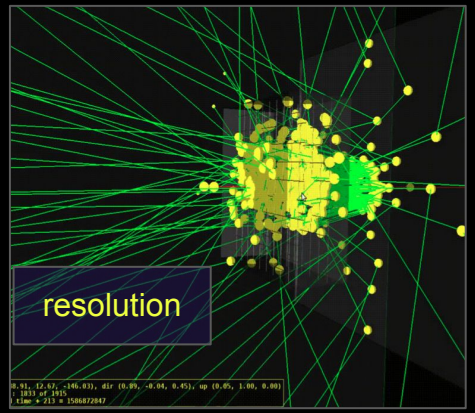
The crowding distance d_i of point *i* is a measure of the objective space around *i* which is not occupied by any other solution in the population.

Novel Aerogel Material aefib

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



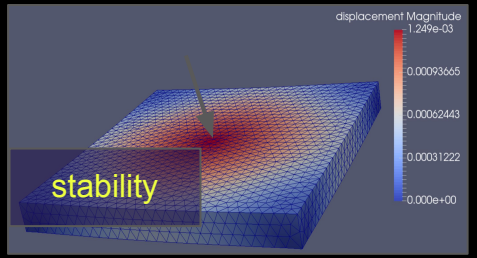
- Aerogels with low refractive indices are very fragile tiles break during production and handling, and their installation in detectors.
- To improve the mechanical strength of aerogels, Scintilex developed a reinforcement strategy. The general concept consists of introducing fibers into the aerogel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



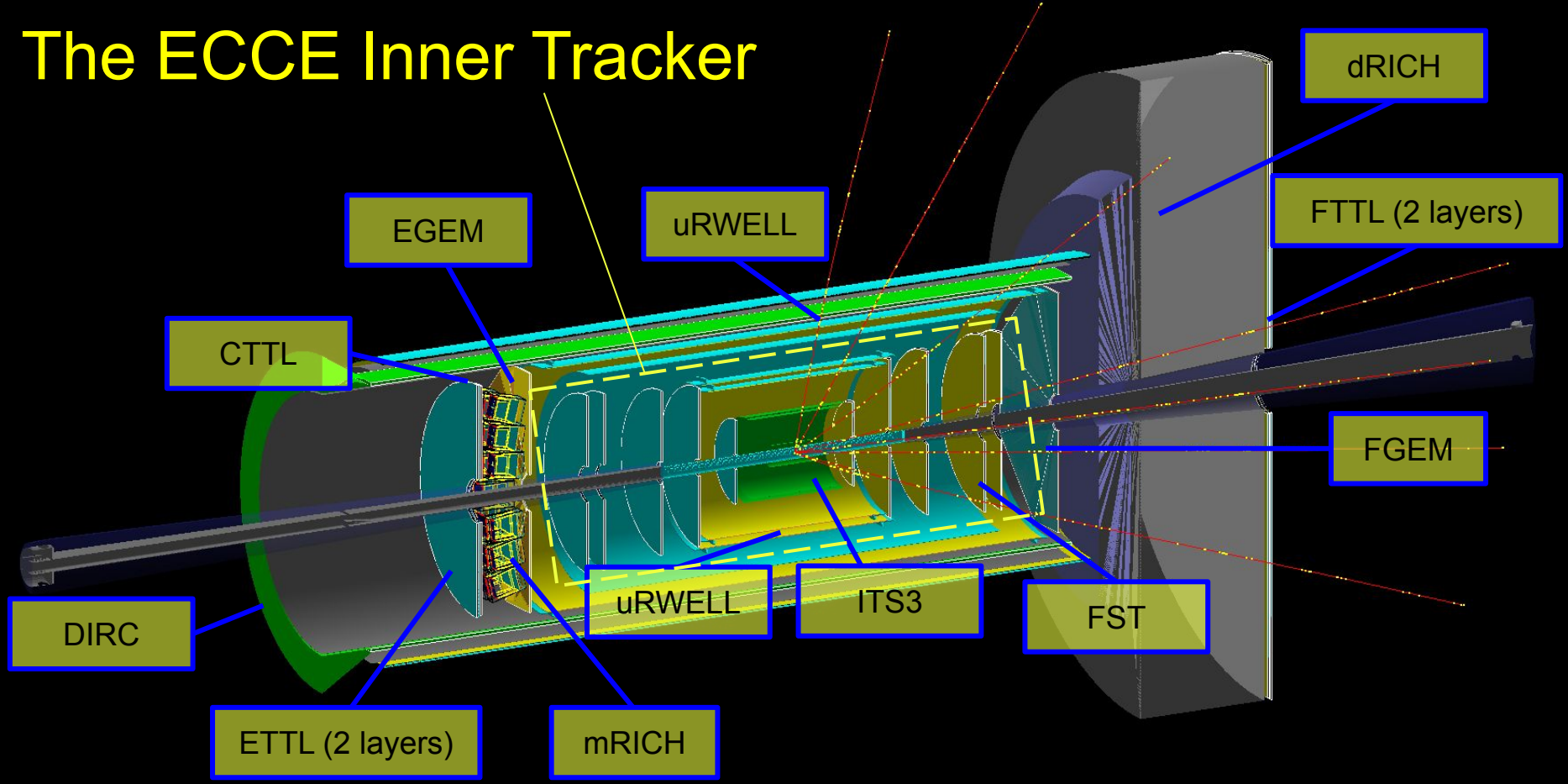
Software Stack

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



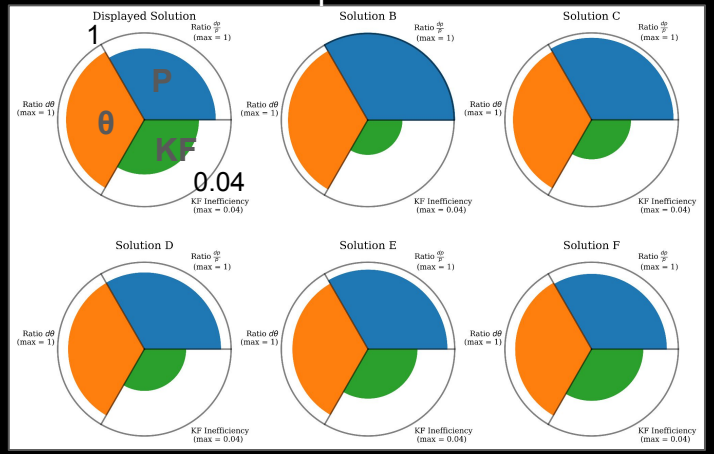
The ECCE Inner Tracker



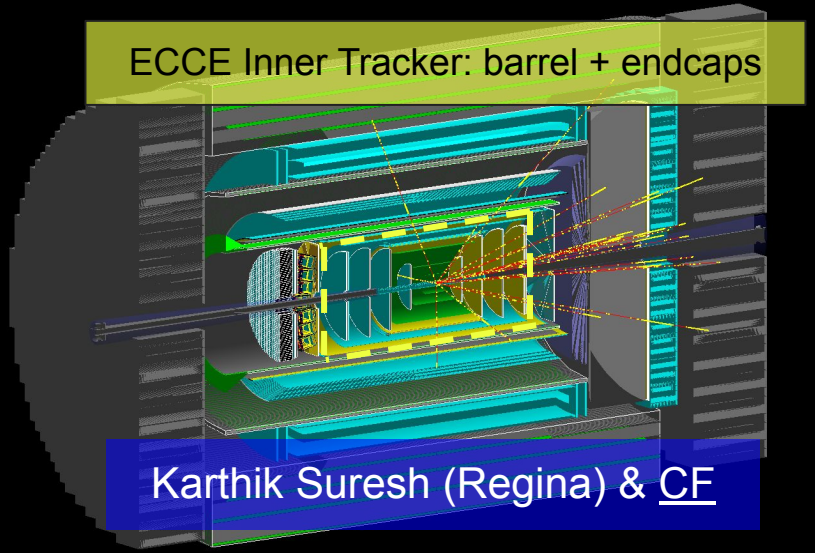
The ECCE Inner Tracker

- Design include simultaneously:
  - momentum resolution
  - angular resolution
  - Kalman filter efficiency
  - Mechanical constraints

- Pareto front: multiple candidate solutions

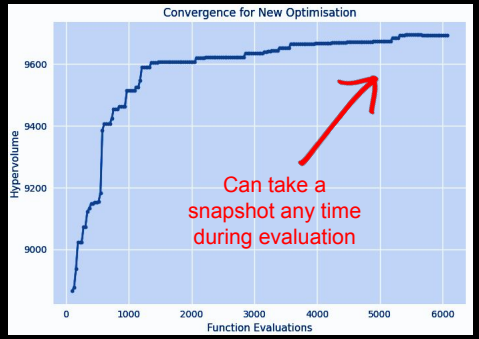


Ratios are with respect to a reference design
Each proposed design is consistent with an Aluminum support shell from the reference design



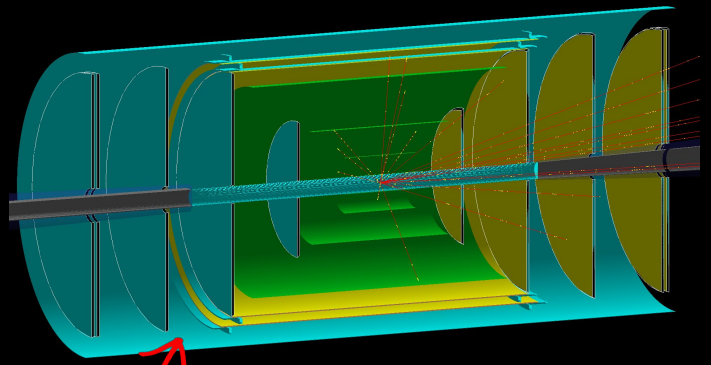
Karthik Suresh (Regina) & CF

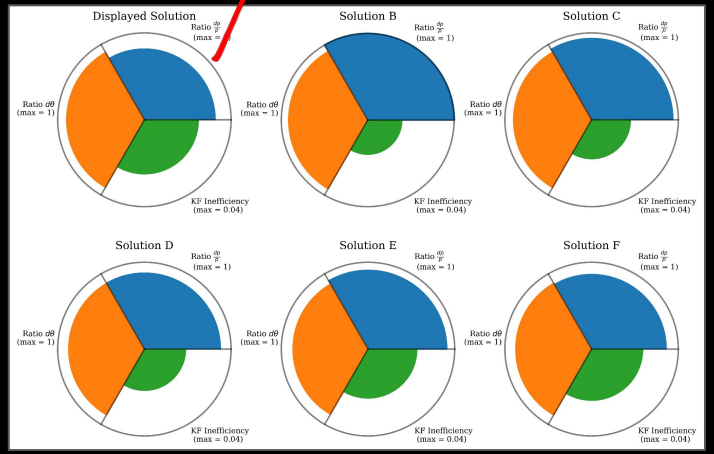
- >= 11 parameters
- 3 (4) objectives
- Population size 100
- Offspring distributed over >=30 cores
- 80000 tracks / design point
- ~1h / design point

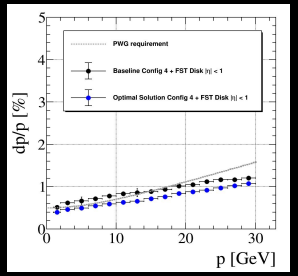


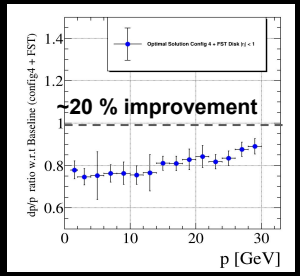
This is (already) an unprecedented attempt in detector design for complexity!

The ECCE Inner Tracker

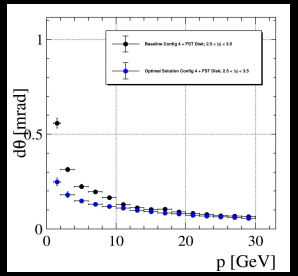


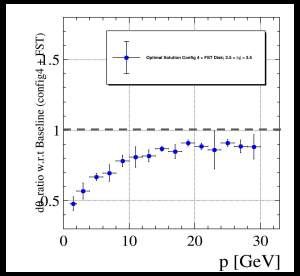






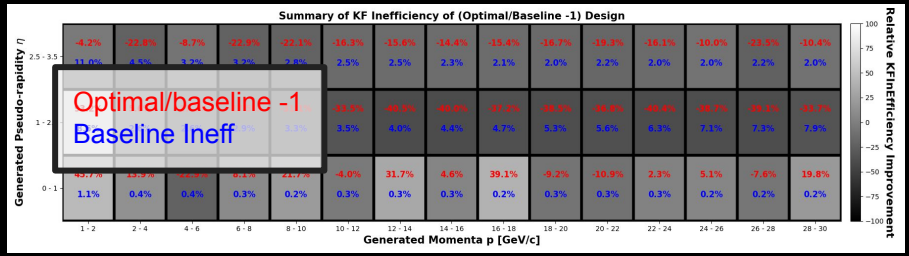
P Reso





θ Reso

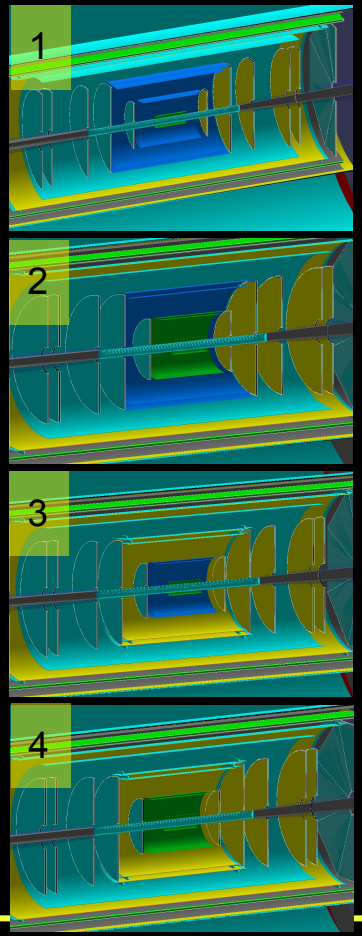
KF Ineff

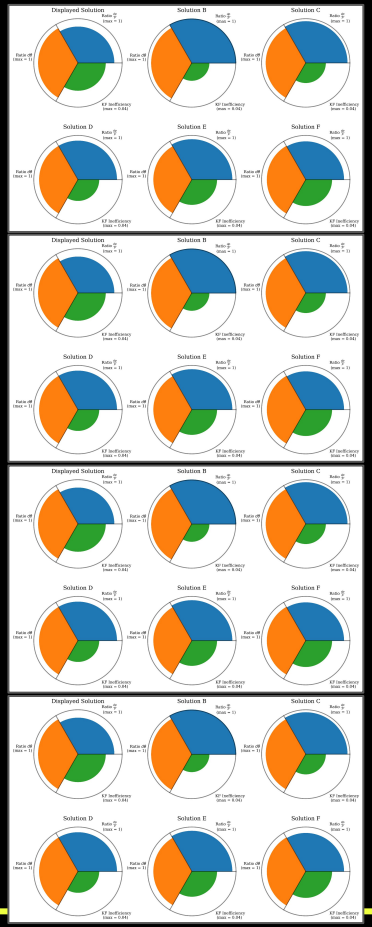


See talk by W. Phelps for more details

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

Multiple Pipelines: Example





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
- Alternate Configuration(s)

PWG's:
- Physics Signal Selection
- Physics Performance Evaluation

CWG's:
- Simulation Campaign Support
- AI Optimization

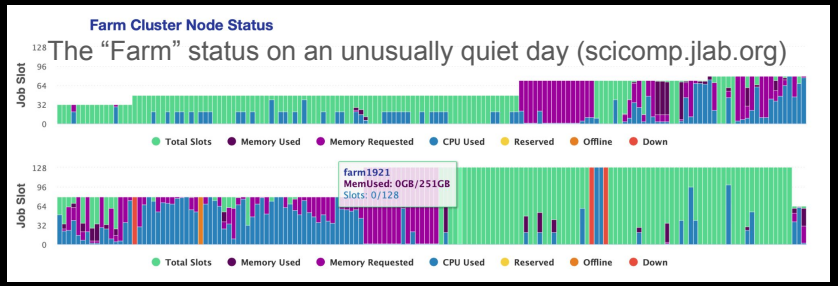
updated configurations with any additional requirements

New optimization pipelines

Resources

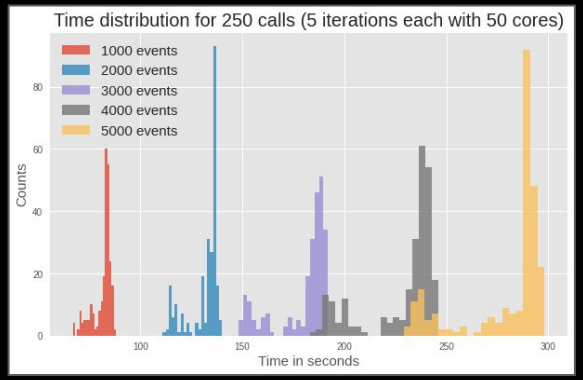
OUR PROBLEM: Inner Tracker
>= 11 parameters
3 objectives
Population size 100
Offspring distributed over >=30 cores

- running on scicomp @ JLab



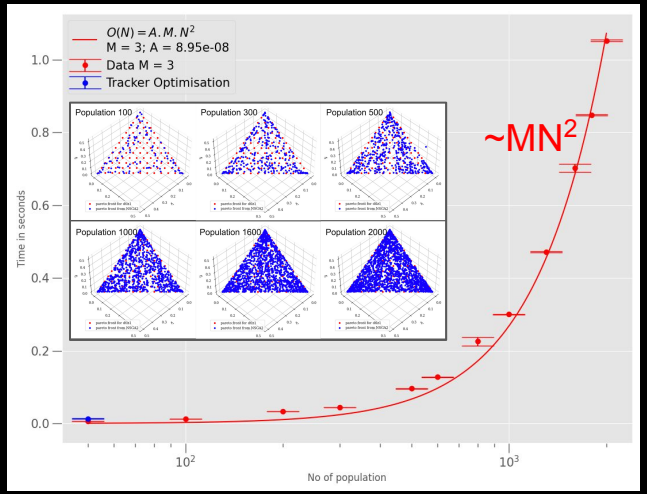
The scientific computing cluster has:
- 25k cores EIC Projects are allocated 10%
- 1PB for EIC use
- Batch use as well as interactive use supported with
  - Nodes with up to two 32 core AMD Epyc Processors (128 threads), 256GB Ram, 1TB SSD local storage
  - 3 Nodes with 4 Titan RTX Cards (24 GB Memory)
  - GPU nodes also available through jupyterhub.jlab.org

- Characterization of simulation times



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

- Characterization of time taken by GA + sorting



- 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

MOEA Parallelization

[1] M.T. Jensen, IEEE Trans. Evol. Comput., 2003

- Well known that NSGA-II increase in computational complexity as **O(MN^2)** [1].
- A recent trend in MOEA is distributed NSGA-II and implementation on supercomputers. This is useful when large populations are needed (*e.g.*, 10^5), due to complexity and/or to approximate the Pareto front with high accuracy.
- A custom optimized parallel NSGA-II called swNSGA-II has been designed for Sunway TaihuLight [2] supercomputer.

[2] Liu, Xin, et al. IEEE Trans Parallel Distrib Syst 32.4 (2020): 975-987.

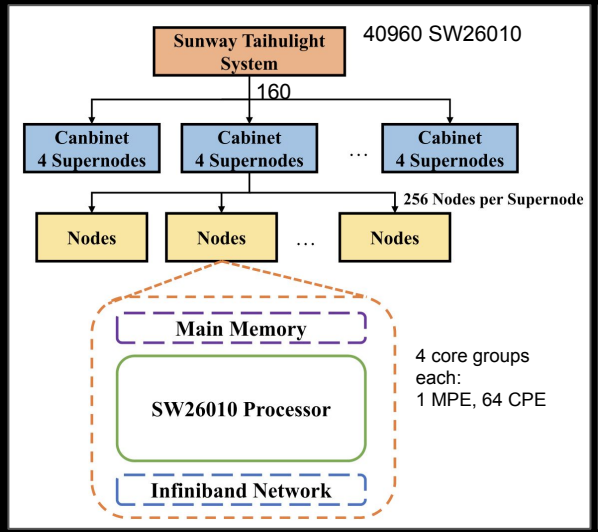


TABLE 3
The Running Time of swNSGA-II on Multiple Core Group(s)

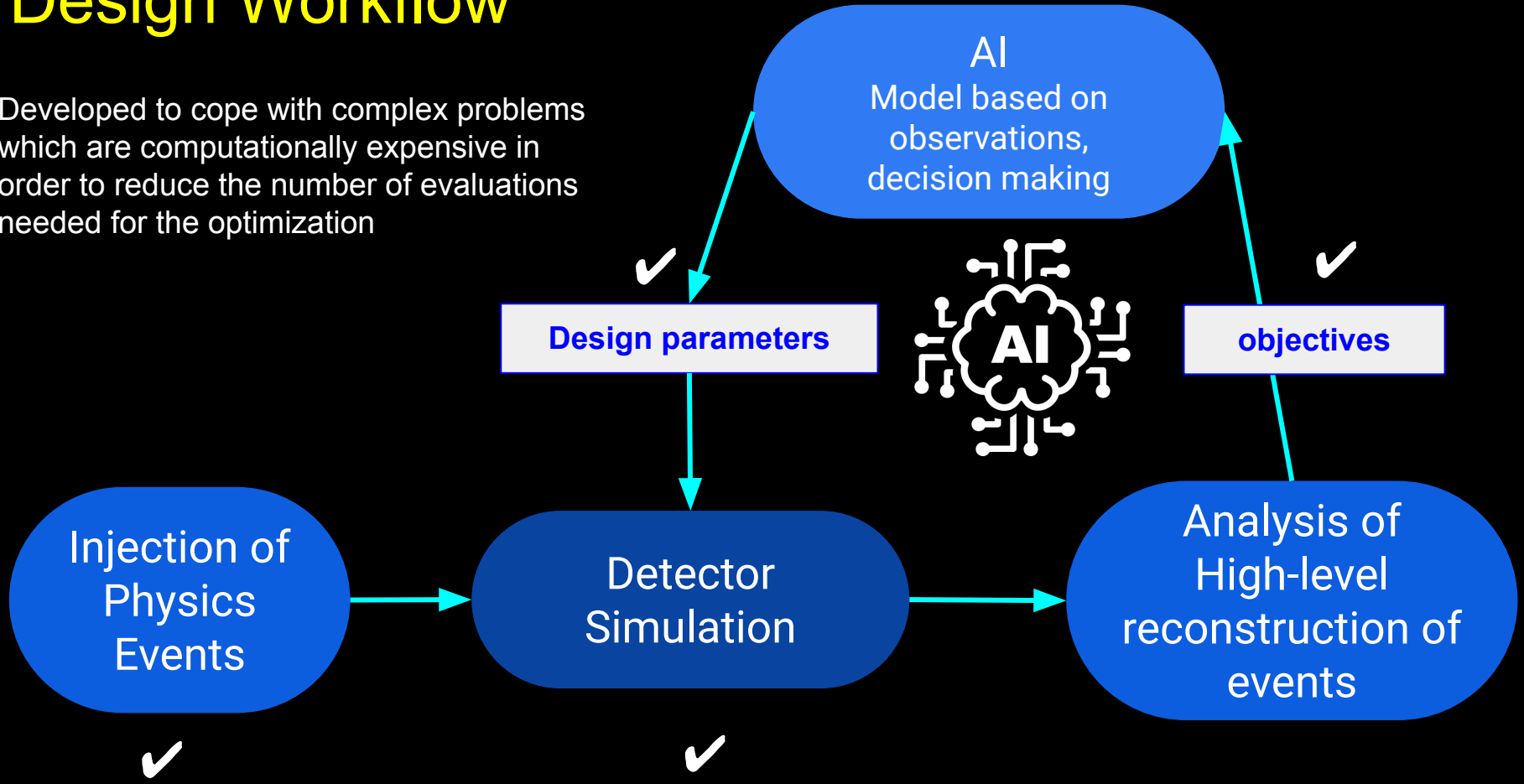
	CG(s)	Time in second(s)	Speedup
Path Planning			
NSGA-II	1*	4954.15	N/A
swNSGA-II	1	222.31	22.28
	2	51.06	24.19
	4	12.67	391.01
	8	3.53	1403.44
	16	1.15	4307.96
	32	0.34	14571.03
	64	0.19	26074.47
	100	0.12	41284.58
ZDT1			
NSGA-II	1*	3134.64	N/A
swNSGA-II	1	255.46	12.27
	2	54.77	57.23
	4	12.09	259.35
	8	2.78	1128.17
	16	0.70	4446.49
	32	0.24	13073.62
	64	0.07	45043.89
	100	0.05	62692.80

*MPE only.

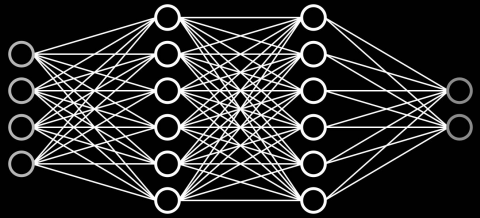
- swNSGA-II utilize process and thread level parallelism based on an improved island master-puppet model.
- Performance have been benchmarked against conventional NSGA-II with a speedup of ~$5 \cdot 10^4$ for standard optimization problems.
- Comparisons with GPU (GeForce GT 630) -based NSGA-II done using 1 core group only (64 CPE), obtaining a speedup of ~10 with large populations.

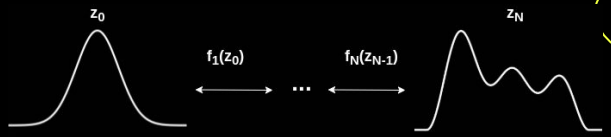
Design Workflow

Developed to cope with complex problems which are computationally expensive in order to reduce the number of evaluations needed for the optimization

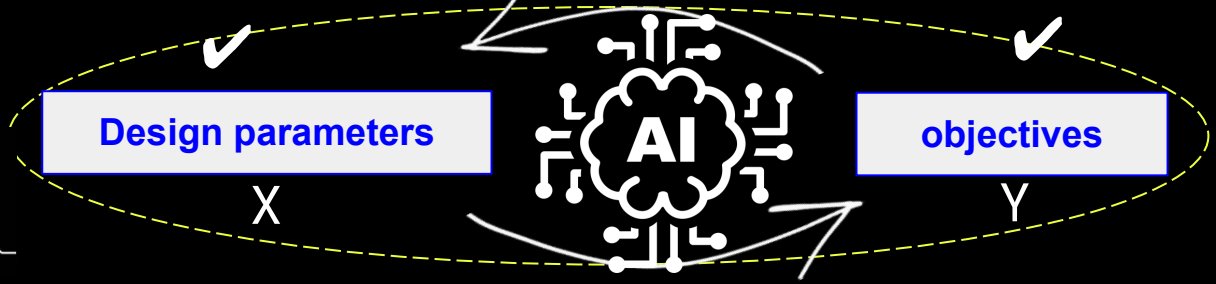


With large datasets...





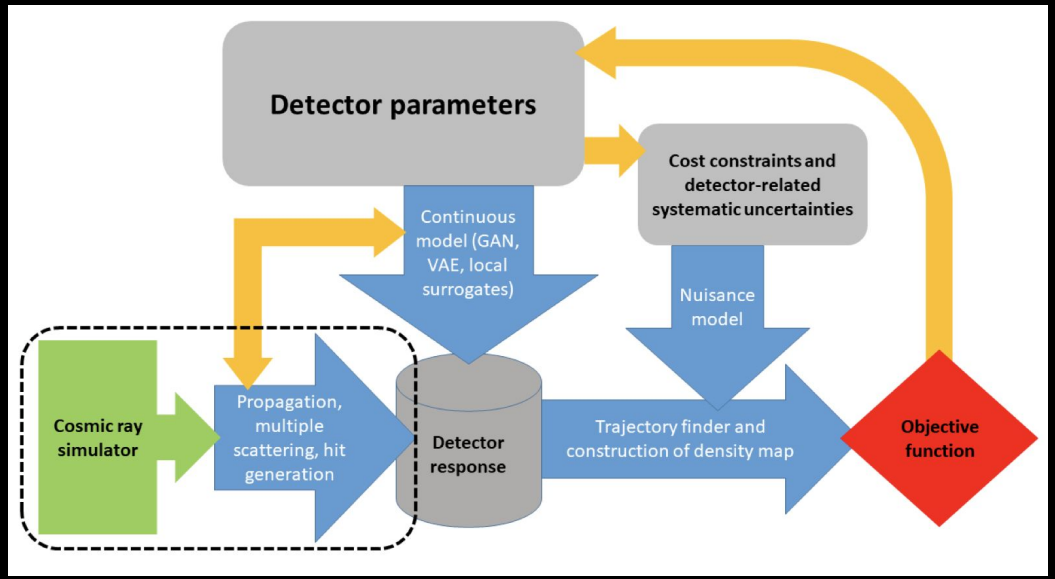
AI
Model based on observations, decision making



Injection of Physics Events → Detector Simulation → Analysis of High-level reconstruction of events

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

Summary

- EIC can be one of the first experiment to be designed with the support of AI.
- ECCE is leading these efforts with an unprecedented attempt in detector design (multidimensional design and objective spaces).
- None ever accomplished a multi-dimensional / multi-objective optimization of the global design, i.e., made by many sub-detectors combined together, that can be solved with AI
  - Costs can be explicitly included during the optimization provided a reliable parametrization)
  - An intrinsic overhead regards compute expensive simulations (+ reconstruction/analysis). How to speed up bottlenecks and overall these steps? See discussion in the Sessions on: Simulations, Reco & Analysis.
  - Larger populations of design points can be simulated to improve accuracy of the Pareto front in multidimensional spaces with AI-based accelerated optimizations.

Likely future detectors will be designed with the help of AI achieving optimal performance and cost reduction.
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*".