MODE

Optimize experiment design using differentiable programming

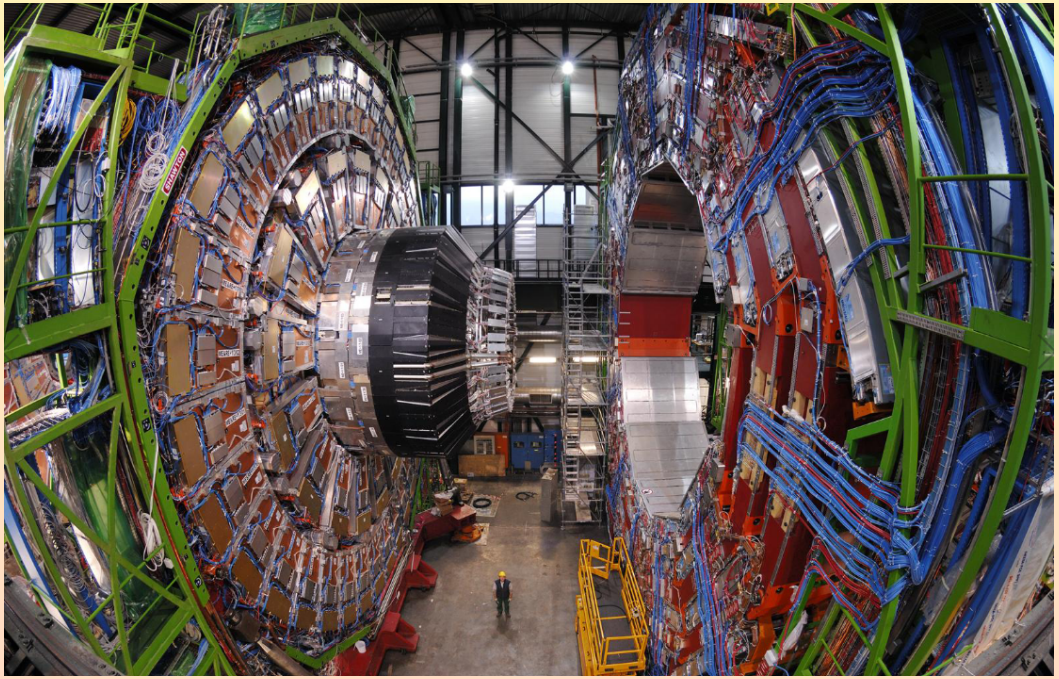
Pietro Vischia[1]
@pietrovischia
for MODE

[1] CP3 — IRMP, Université catholique de Louvain

HSF Detector Simulation Working Group, 2022.11.28

Complex experiments

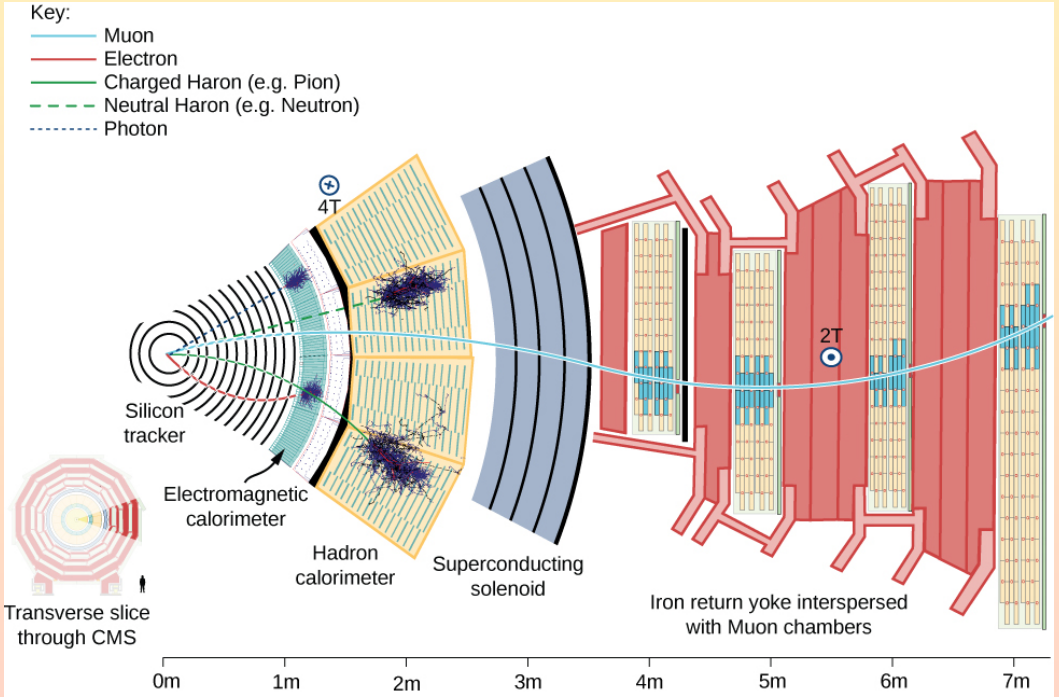
- Costly full simulation
- Interrelation of many parameters
- Large number of optimizable subdetectors
- Complexity prevents from optimizing targeting final goals



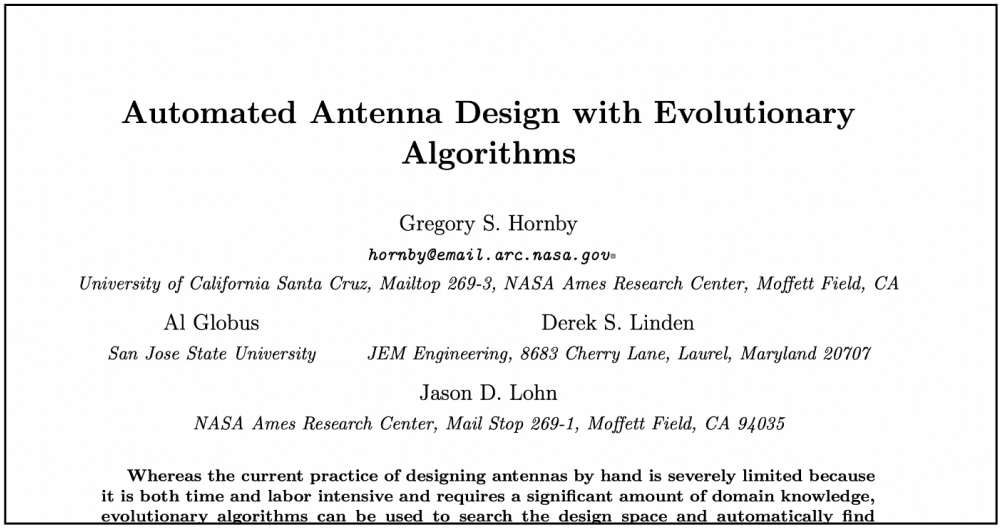
Robustness is not optimization

- 50+ years old detector design concepts served us well but may now be assisted by AI

- *Track first, destroy later*
- Redundancy in the detection systems
- Symmetrical layouts
  - No guarantee of optimality
- Subdetector-specific figures of merit



Optimal design requires huge domain expertise



Automated Antenna Design with Evolutionary Algorithms

Gregory S. Hornby
hornby@email.arc.nasa.gov
University of California Santa Cruz, Mailtop 269-3, NASA Ames Research Center, Moffett Field, CA

Al Globus
San Jose State University

Derek S. Linden
JEM Engineering, 8683 Cherry Lane, Laurel, Maryland 20707

Jason D. Lohn
NASA Ames Research Center, Mail Stop 269-1, Moffett Field, CA 94035

Whereas the current practice of designing antennas by hand is severely limited because it is both time and labor intensive and requires a significant amount of domain knowledge, evolutionary algorithms can be used to search the design space and automatically find

"The current practice of designing and optimizing antennas by hand is limited in its ability to develop new and better antenna designs because it requires significant domain expertise and is both time and labor intensive."

Slide by Lukas Heinrich at QCHS 2022. Referenced paper is http://dx.doi.org/10.2514/6.2006-7242

Why do we need a joint optimization?

- Joint optimization yields in general different solution than optimization of individual features
  - Both marginally and sequentially

$$argmin_{x,y}\Big(\mathcal{L}(x,y)\Big)_x \neq argmin_x\Big(\mathcal{L}(x,y)\Big)$$

$$argmin_{x,y}\Big(\mathcal{L}(x,y)\Big)_y \neq argmin_y\Big(\mathcal{L}(x,y)\Big)$$

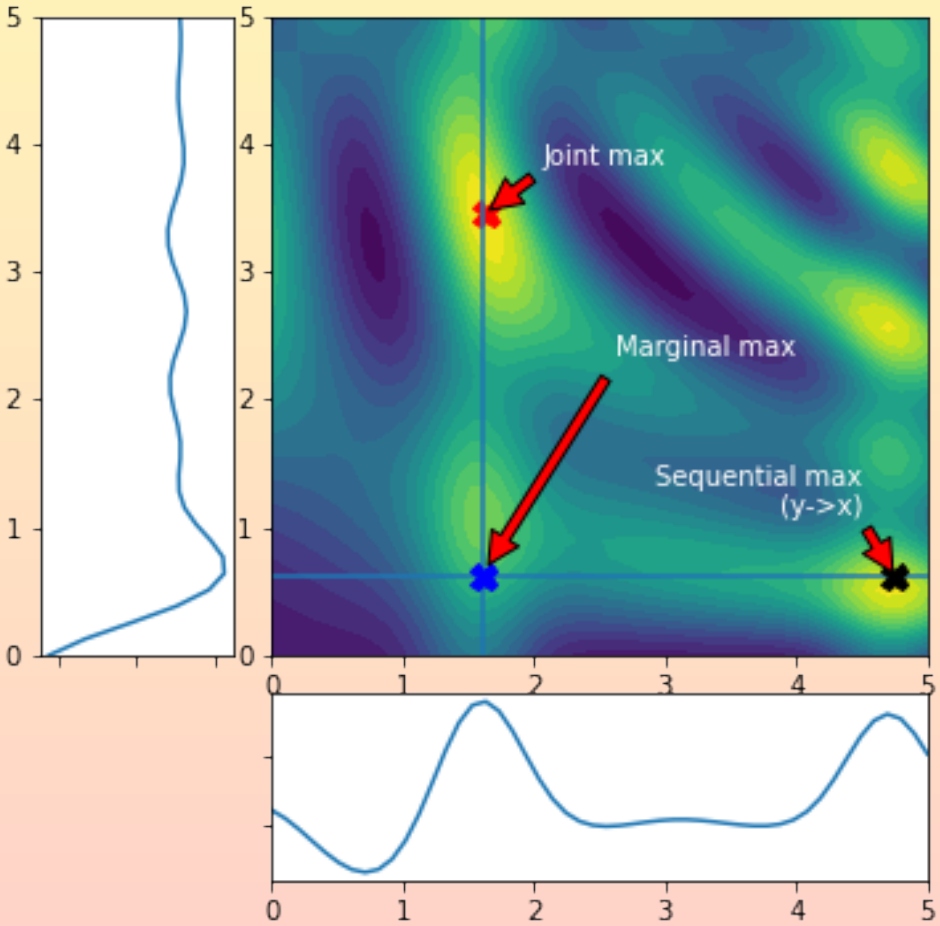


Illustration (c) P. Vischia, book in preparation

Large gains even with naïve techniques

- MUonE detector: proposed to be built at CERN to use a 150 GeV muon beam
  - Measure precisely the q^2 differential cross section in electron-muon scattering
  - This in turn constrains directly the leading source of systematic uncertainty in ~~the anomalous magnetic moment of the muon~~ a very important quantity, protagonist of recent contrasting results
  - 40 tracking stations and a calorimeter
- **Even a simple grid search yielded dramatic improvement** in the resolution on q^2

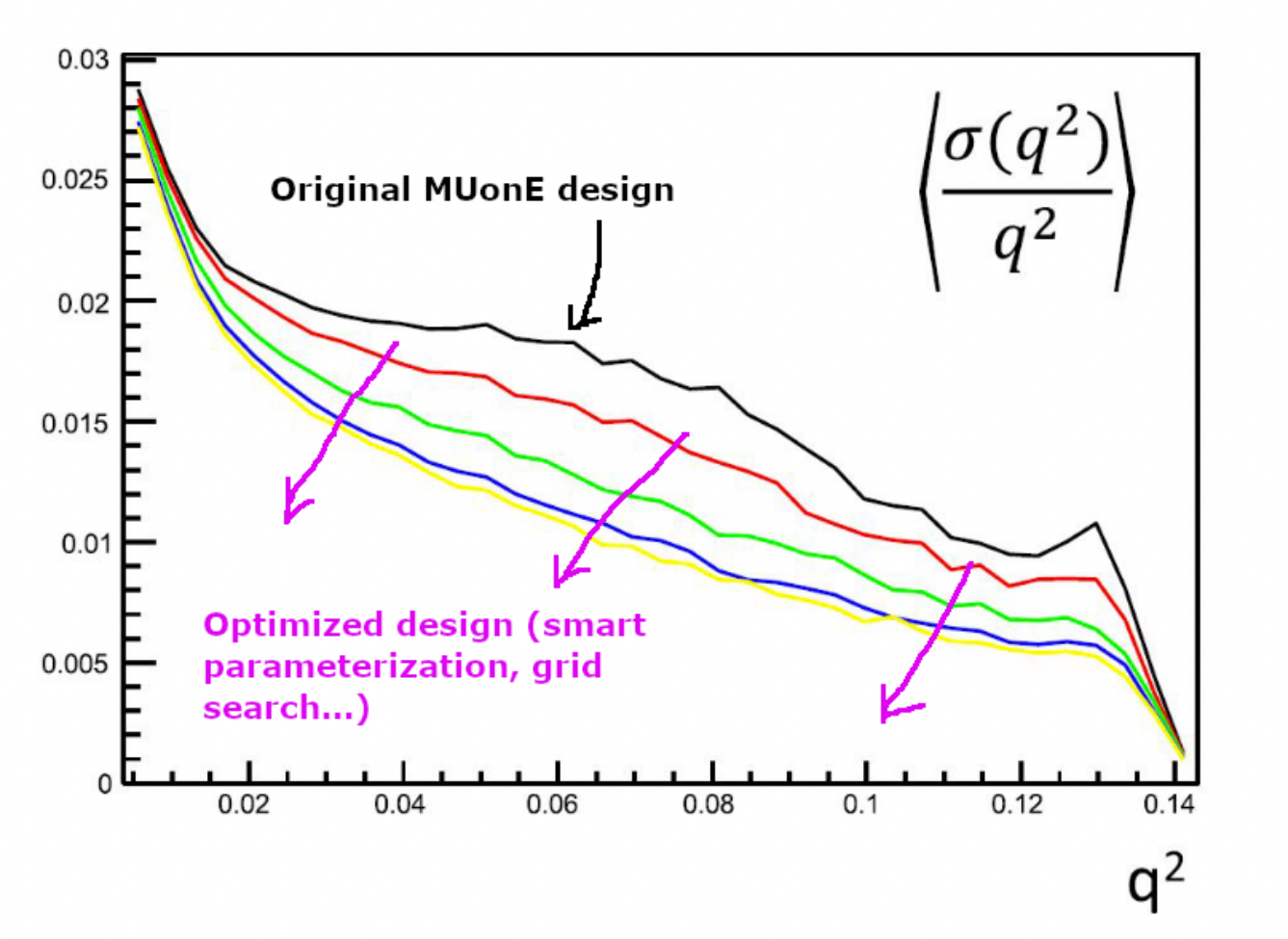
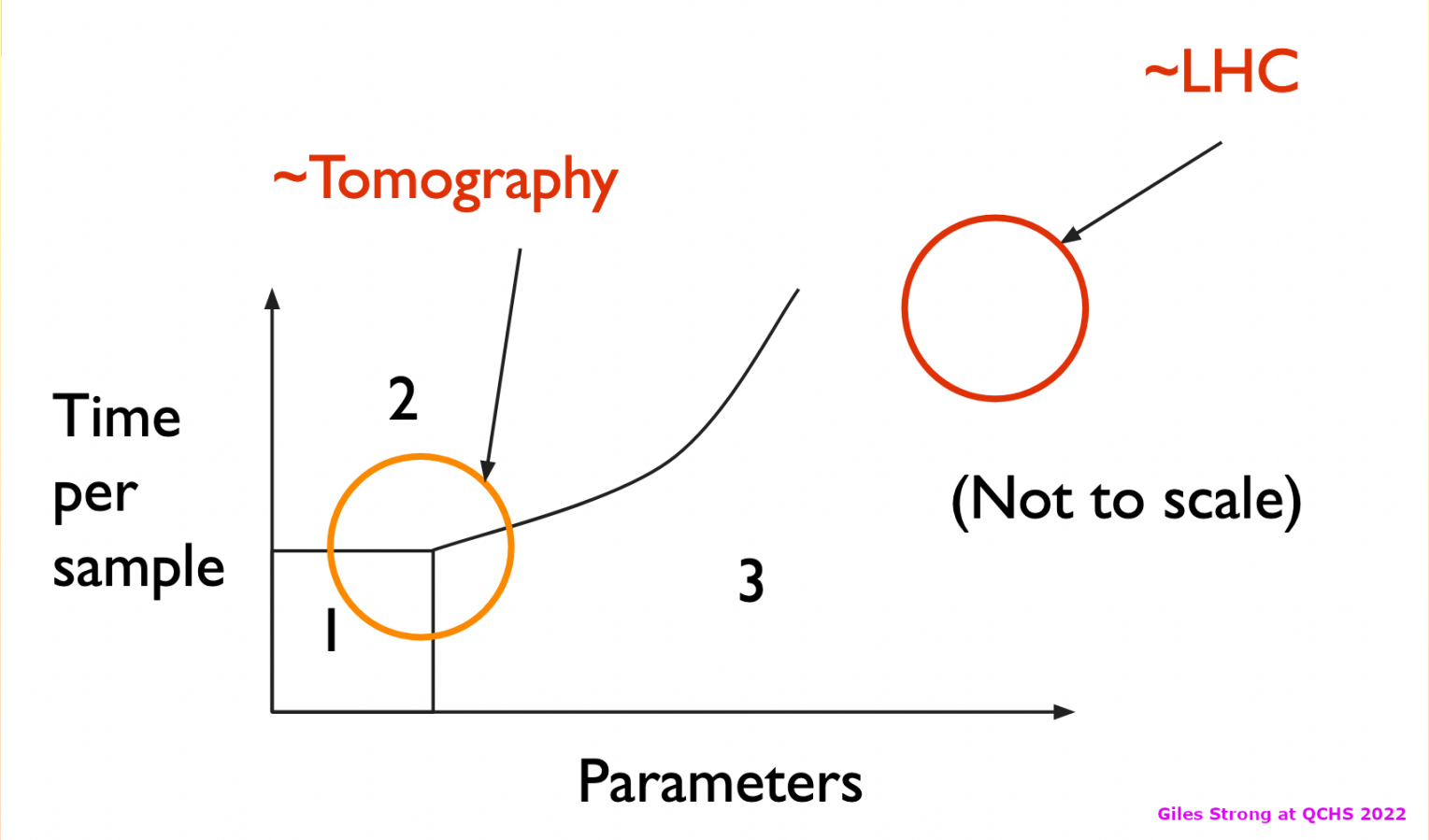


Image from Physics Open, 4:100022, 2020 and MODE White Paper 2203.13818

Different challenges require different methods



- **1:** Grid/random search
- **2:** Bayesian opt, simulated annealing, genetic algos, ...
- **3:** *Gradient-based optimization* (Newtonian, gradient descent, BFGS, ...)

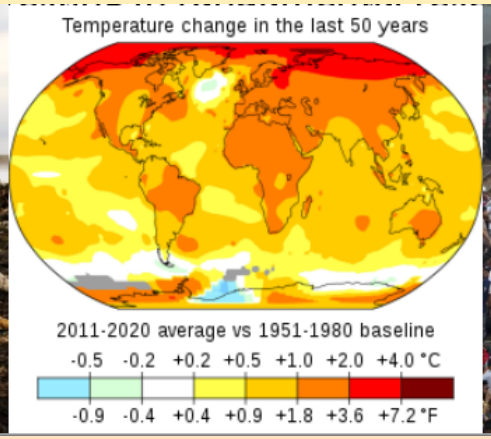
Full optimization is crucial when budget is constrained

Optimize...

- New large, long-term projects[1]
- Push technological skills to the limit

...within constraints

- Unprecedented global challenges
- Society less receptive to fundamental research



Maximum extraction of scientific value from the available resources is a moral imperative!

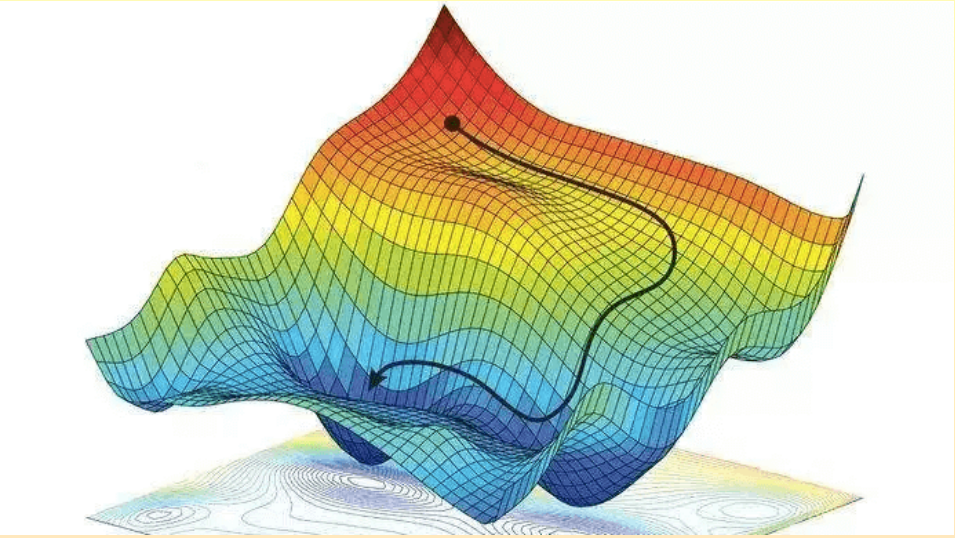
Slide by Tommaso Dorigo

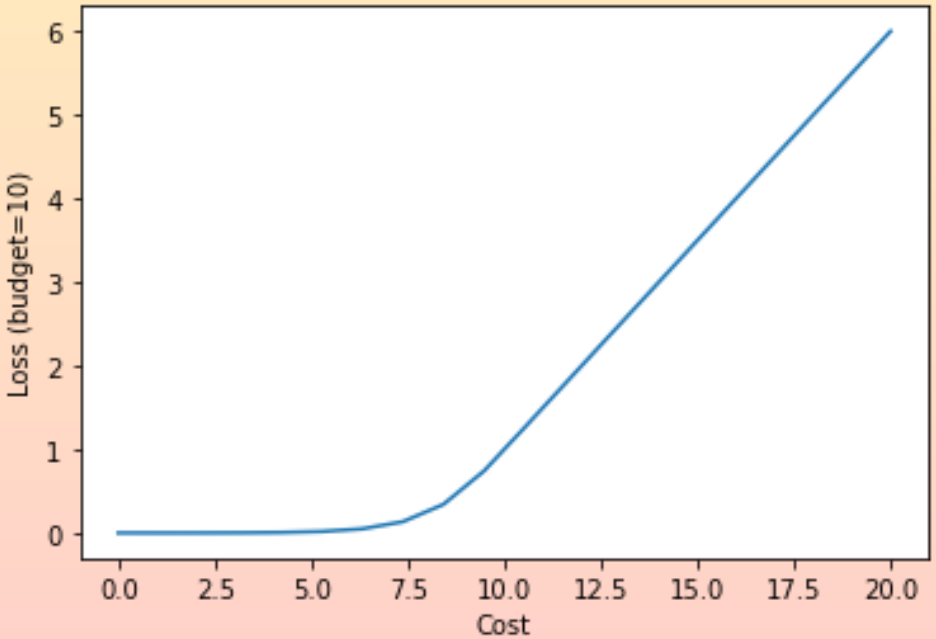
[1] 2020 update of the European Strategy for Particle Physics (EUSUPP)

Finite Budget: loss and constraints

- Optimization via gradient descent
  - Target-oriented loss functions
- Constraints inserted as penalization
  - Additional term to the loss

$$\mathcal{L} = \mathcal{L}(\text{physics output}) + \lambda\Big(\mathcal{L}(\text{cost})\Big)$$





Top illustration from easyai.tech, bottom one from MODE White Paper 2203.13818

Guarantee feasibility within constraints

- Monetary cost
- Case-specific technical constraints

$$\mathcal{L}_{\text{cost}} = c(\theta, \phi)$$

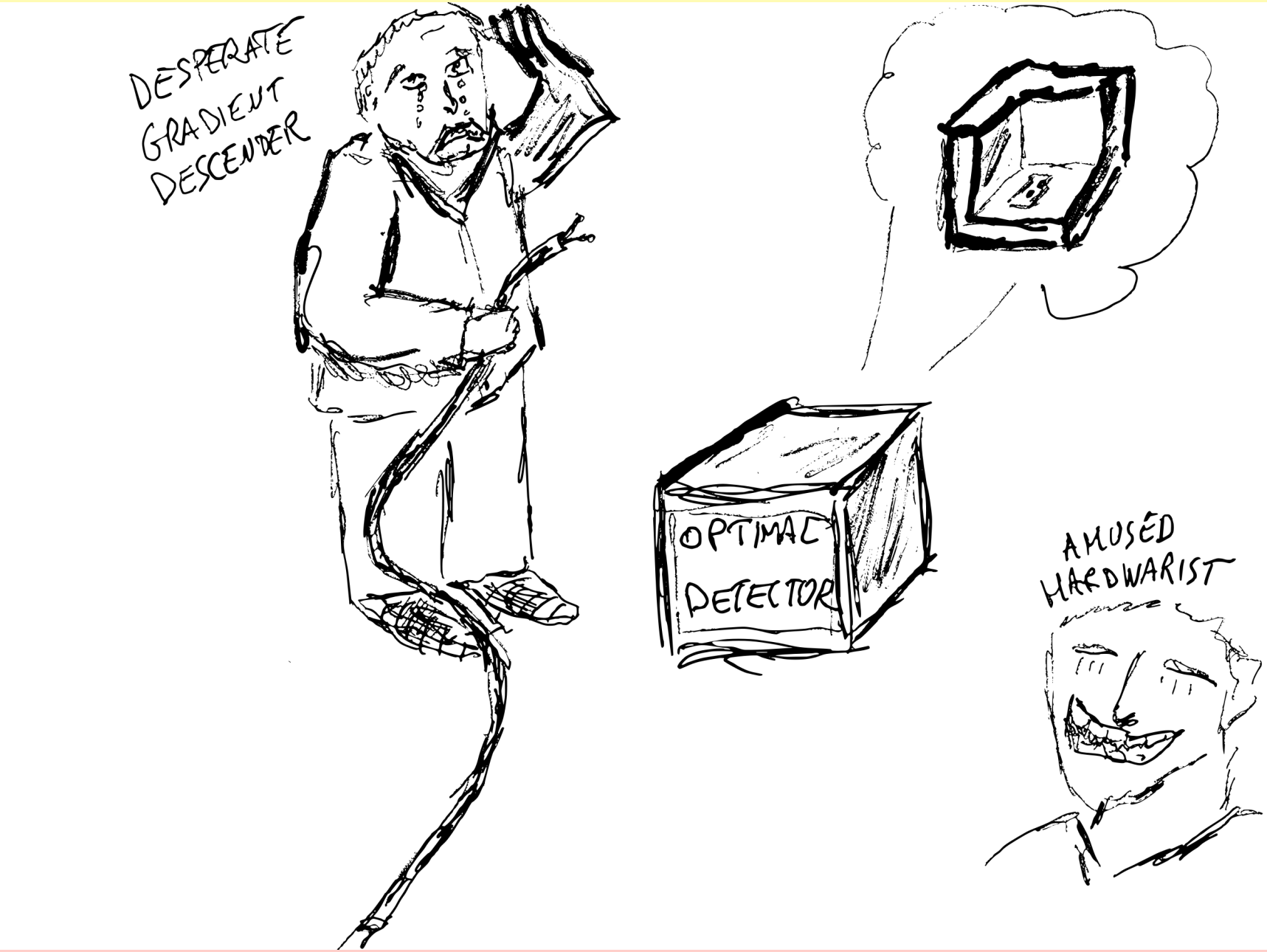
- θ: local, specific to the technology used (e.g. active components material)
- ϕ: global, describing overall detector conception (e.g. number, size, position of detector modules)
- Fixed costs can be added separately to the loss function

Optimization has practical consequences

- Material availability (influenced e.g. also by wars) is also a concern, nowadays

Figure 6: Road transport of a structure for the ATLAS air toroids. Photo reproduced from Ref. [181].

If you can't turn it on, it's not optimal



Maybe we can optimize cable layout

- Cable layout can be easily described w/ trees or graphs
- Cable layout cost function intrinsically discontinous and nonsmooth
- Mostly gradient-free tree searches
- Maybe further studies on the loss landscape can help in solving this in a differentiable way

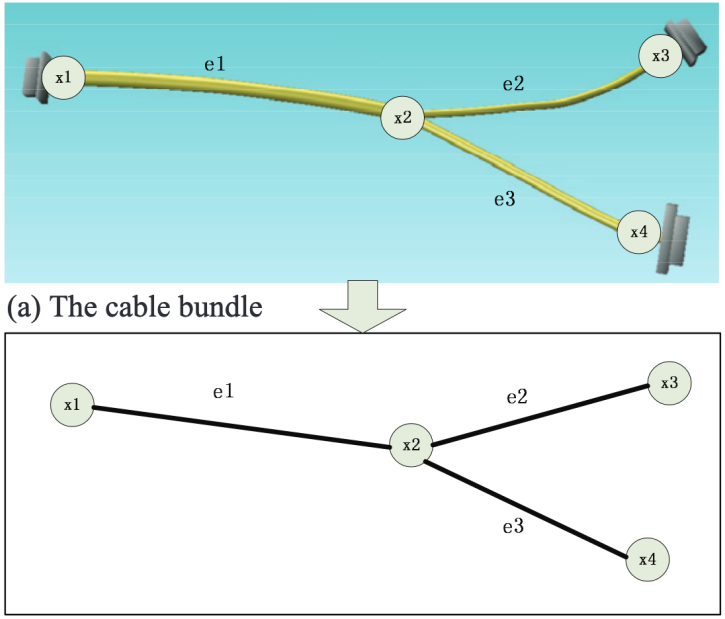


Figure 4. Approach 1: sequential design process for WTs and cable layout.

Figure 5. Approach 2: simultaneous design process for WTs and cable layout.

Left image: 10.1016/j.phpro.2012.05.297, right image: 10.5194/wes-7-925-2022

Assist the physicist with a landscape of solutions

- Cannot parameterize everything
- “**The** optimal solution” unrealistic
- Provide feasible solutions near optimality
- The physicist will fine tune

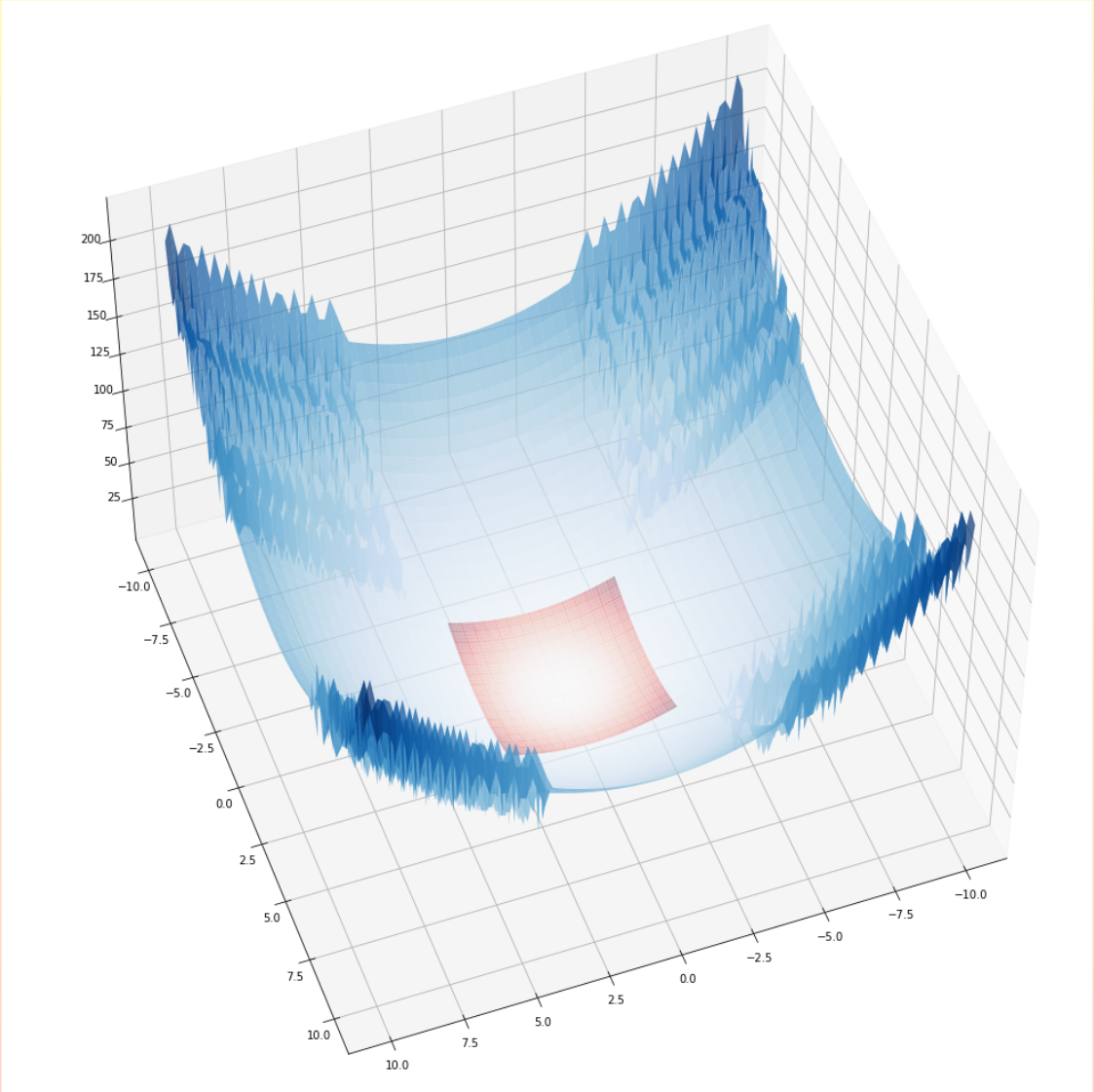


Illustration (c) P. Vischia, book in preparation

The MODE White Paper (2203.13818)

- 109-page document drafting the way forward
  - Featuring several computer scientists from https://sivert.info
    (proton Computed Tomography)
- Submitted to Reviews in Physics

Contents

Toward the End-to-End Optimization of Particle Physics Instruments with Differentiable Programming: a White Paper

Tommaso Dorigo[1,2], Andrea Giammanco*[1,3], Pietro Vischia[1,3] (editors) , Max Aehle[4], Mateusz Bawaj[5], Alexey Boldyrev[1,6], Pablo de Castro Manzano[1,2], Denis Derkach[1,6], Julien Donini[1,7], Auralee Edelen[8], Federica Fanzago[1,2], Nicolas R. Gauger[4], Christian Glaser[1,9], Atılım G. Baydin[1,10], Lukas Heinrich[1,11], Ralf Keidel[12], Jan Kieseler[1,13], Claudius Krause[1,14], Maxime Lagrange[1,3], Max Lamparth[1,11], Lukas Layer[1,2,15], Gernot Maier[16], Federico Nardi[1,2,17,7], Helge E. S. Pettersen[18], Alberto Ramos[19], Fedor Ratnikov[1,6], Dieter Röhrich[20], Roberto Ruiz de Austri[19], Pablo Martínez Ruiz del Árbol[1,21], Oleg Savchenko[2,3], Nathan Simpson[22], Giles C. Strong[1,2], Angela Taliercio[3], Mia Tosi[1,2,17], Andrey Ustyuzhanin[1,6], and Haitham Zaraket[1,23]

[1]MODE Collaboration, https://mode-collaboration.github.io/

Optimization of an experiment: ingredients (2203.13818)

- Multidimensional stochastic input variable $x \sim f(x)$ from simulator of physics process
  - Potentially dependent on latent variables

Optimization of an experiment: ingredients (2203.13818)

- Multidimensional stochastic input variable $x \sim f(x)$ from simulator of physics process
  - Potentially dependent on latent variables
- Sensor readouts $z \sim p(z|x, \theta)$

Optimization of an experiment: ingredients (2203.13818)

- Multidimensional stochastic input variable $x \sim f(x)$ from simulator of physics process
  - Potentially dependent on latent variables
- Sensor readouts $z \sim p(z|x, \theta)$
- High-level features $\zeta(\theta) = R[z, \theta, \nu(\theta)]$

Optimization of an experiment: ingredients (2203.13818)

- Multidimensional stochastic input variable $x \sim f(x)$ from simulator of physics process
  - Potentially dependent on latent variables
- Sensor readouts $z \sim p(z|x, \theta)$
- High-level features $\zeta(\theta) = R[z, \theta, \nu(\theta)]$
- Low-dimensional summary for inference, $s = A[\zeta(\theta)]$

Optimization of an experiment: ingredients (2203.13818)

- Multidimensional stochastic input variable $x \sim f(x)$ from simulator of physics process
  - Potentially dependent on latent variables
- Sensor readouts $z \sim p(z|x, \theta)$
- High-level features $\zeta(\theta) = R[z, \theta, \nu(\theta)]$
- Low-dimensional summary for inference, $s = A[\zeta(\theta)]$
- Optimization metric to find values of θ that optimize inference made with s

Optimization of an experiment: recipe (2203.13818)

Depends on z and nuisances

Cost of the layout with parameters theta

Closed form

$$\hat{\theta} = arg\ min_\theta \int L[A(\zeta), c(\theta)] p(z|x,\theta) f(x) dx dz \ ,$$

Weight desirable goals while obeying cost constraints

For example, to identify smuggled material in a container:

$$L = \left(1 + e^{k(c_\theta - c_0)}\right) \sum_Z \left[w_{imp}(Z) m_{50,\alpha}^{\text{concealed}}[s(Z)] \right]$$

If everything is differentiable, then we are happy (2203.13818)

- Domain knowledge crucial to parameterize systems in an optimal way (pun intended)

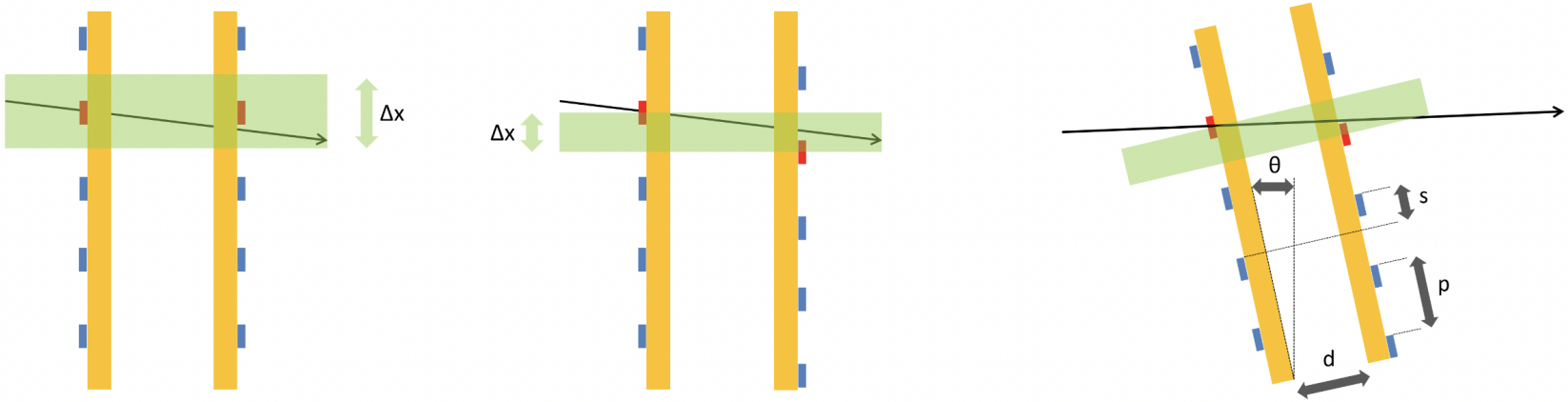


Figure 4: Left and center: a double-sided silicon strip sensor produces twice smaller resolution Δx on single-strip hit position for an orthogonally incident particle if strips on the two sides are staggered by half the strip pitch. Right: the four parameters affecting single-strip hit position resolution (tilt angle θ, strip pitch p, sensor distance d, staggering s).

Surrogates for intractable likelihoods (2203.13818)

- $p(\cdot)$ not in closed form
  - Sample $x_i \sim f(x)$
  - Then z_i distributed as emulator, $x_i \sim F(x_i, \theta)$

$$\hat{\theta}_{\text{approx}} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} L\Big[A(R(z_i)), c(\theta)\Big]$$

- $F(\cdot)$ nondifferentiable stochastic simulator
  - Replace with local surrogate $z = S(y, x, \theta)$, where y describes the stochastic variation of the approx distribution
  - Learn surrogate separately

$$\nabla_\theta \Big(\widehat{L(z)}\Big) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta L\Big[A\Big(R(S(y_i, x_i, \theta))\Big), c(\theta)\Big]$$

Why using surrogates (your favourite VAE, GAE, NF, whatever) is cool

- Subset of relatively simple class of functions (but they must be able to reproduce $F(\cdot)$ well)
- Learn by training (*hic sunt leones*), (but $N(\text{eval } F) \geq \mathcal{O}(dim(\theta))$)
- Automatically get AD out of the box even if original $F(\cdot)$ is not differentiable
- Evaluation of surrogate (for optimization) much faster than evaluation of $F(\cdot)$

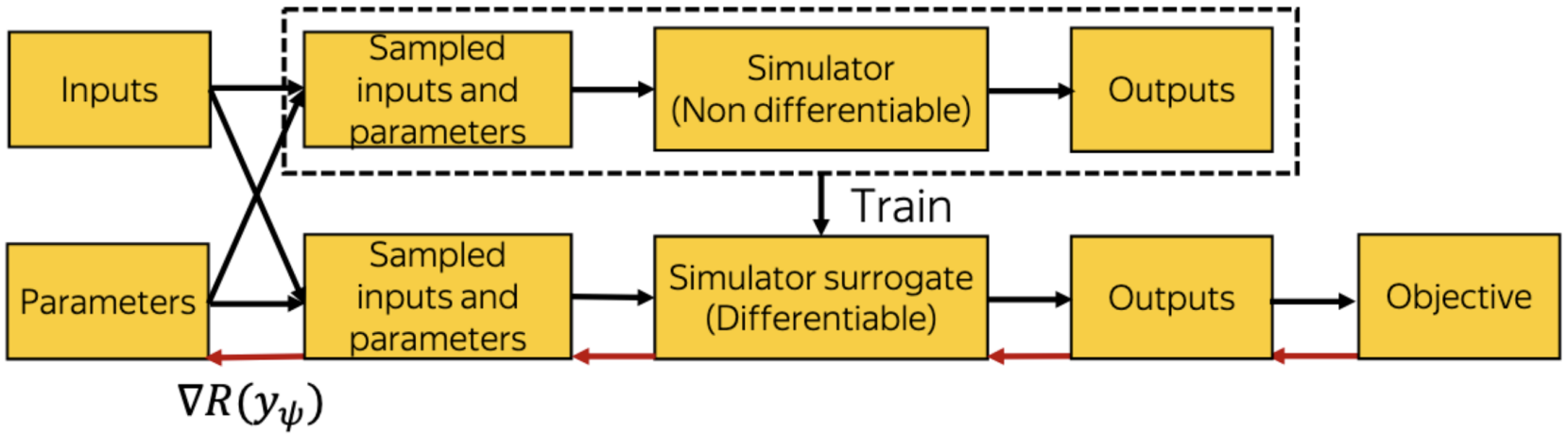


Figure 1: Simulation and surrogate training. *Black:* forward propagation. *Red:* error backpropagation.

Images from Shirobokov,..., Kagan, ... *et al.* 2002.04632

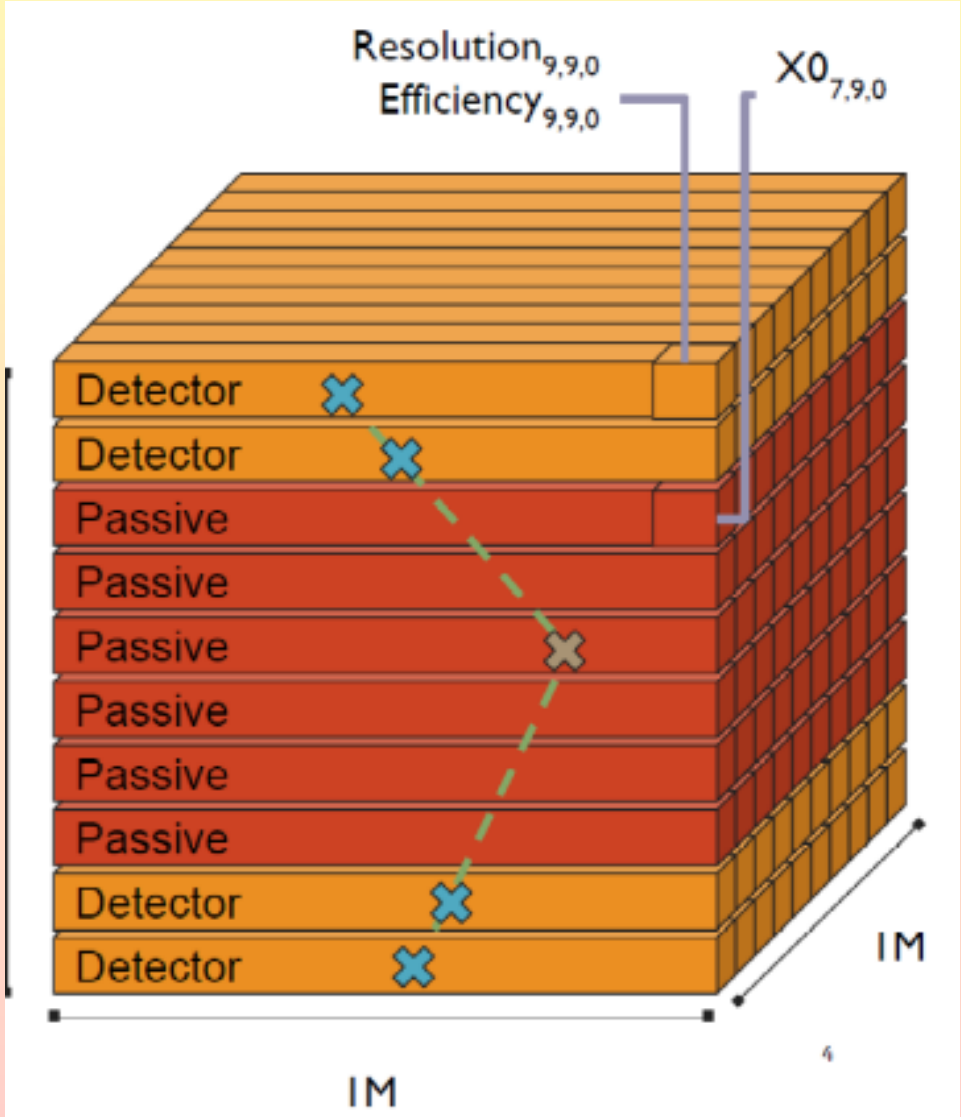
Vast and diverse set of use cases (2203.13818)

We are already exploring several use cases (2203.13818)

- Muon tomography (e.g. for nuclear safety)
- LHCb and CMS calorimeter optimization
- SWGO placement and geometry of tanks
- LEGEND optimization

Mature use case

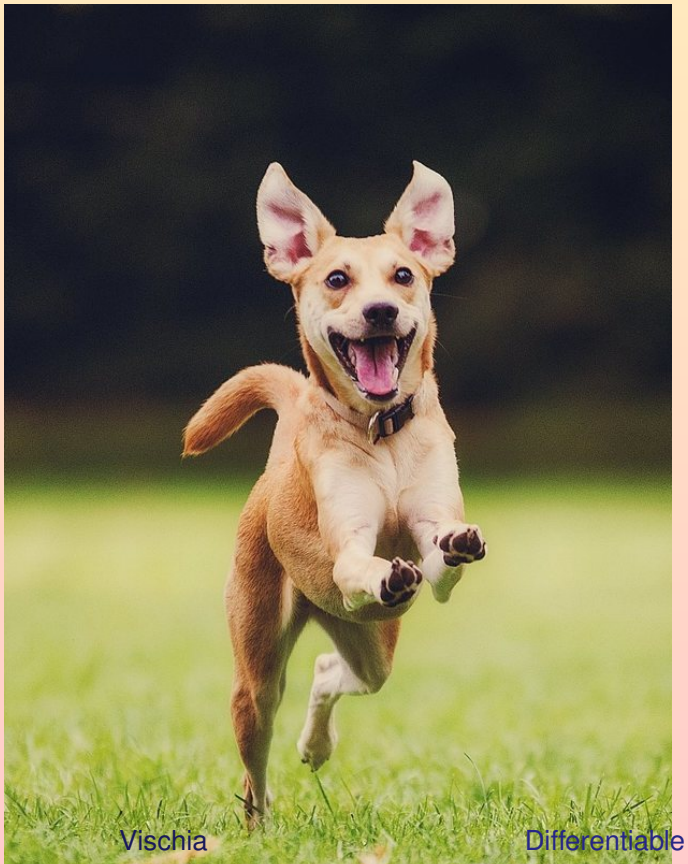
- Scan un unknown material using cosmic muons
- See Giles' presentation for methodology and early results

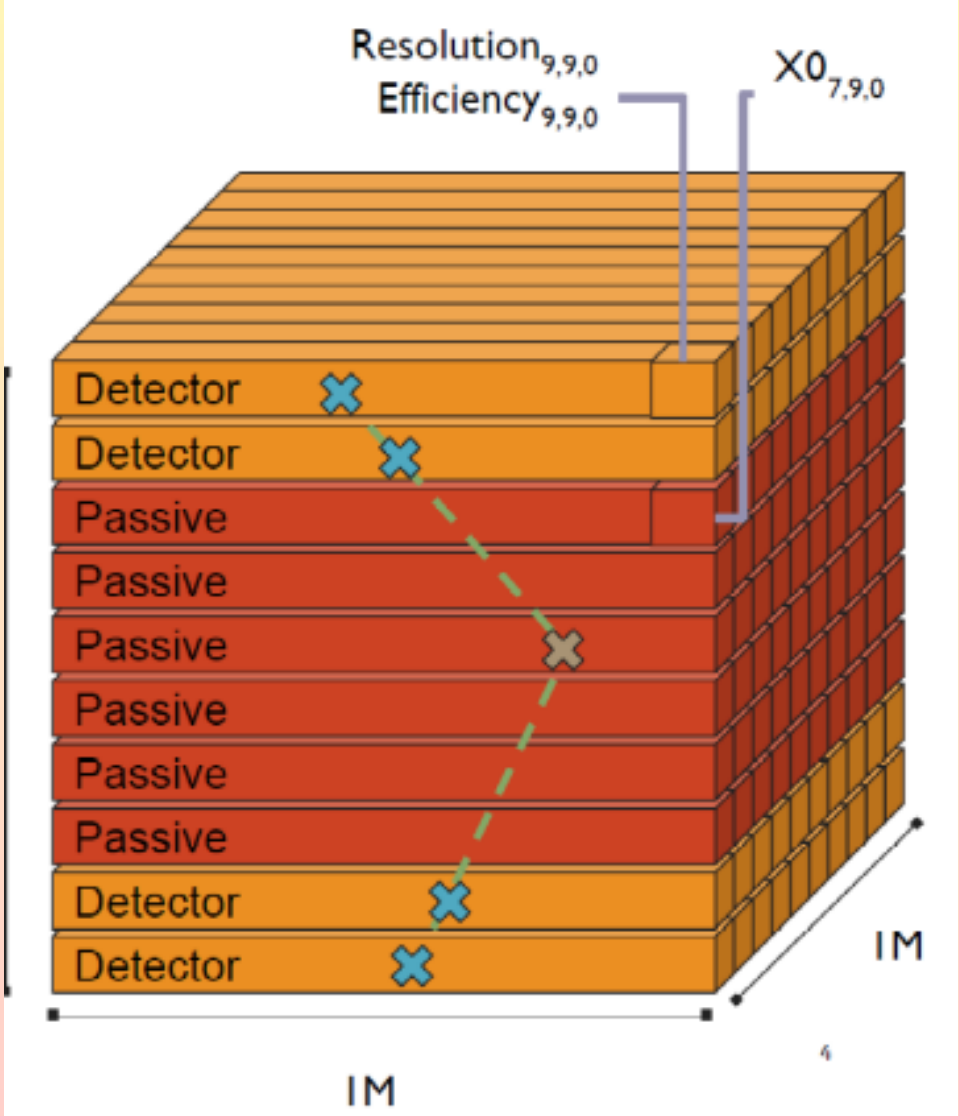


Illustrations from TomOpt draft paper

Detector Modelling and Nondifferentiable Elements

- Muons hit or miss
  - Model via resolution and efficiency distribution (e.g. Gaussian mixture model)
  - Now differentiable!
- Account for discrete number of layers as continuous “layer efficiency” allowed to kill a layer by dropping to zero
  - When nonzero, multiplier of the per-element efficiency
- Or resort to reinforcement learning?





Particle accelerators, some of the most complex instruments in existence

- Strong impact in medical and industrial applications
- Interconnected sub-systems
- Time-varying inputs and responses
- Thousands of adjustable variables
- Highly nonlinear responses

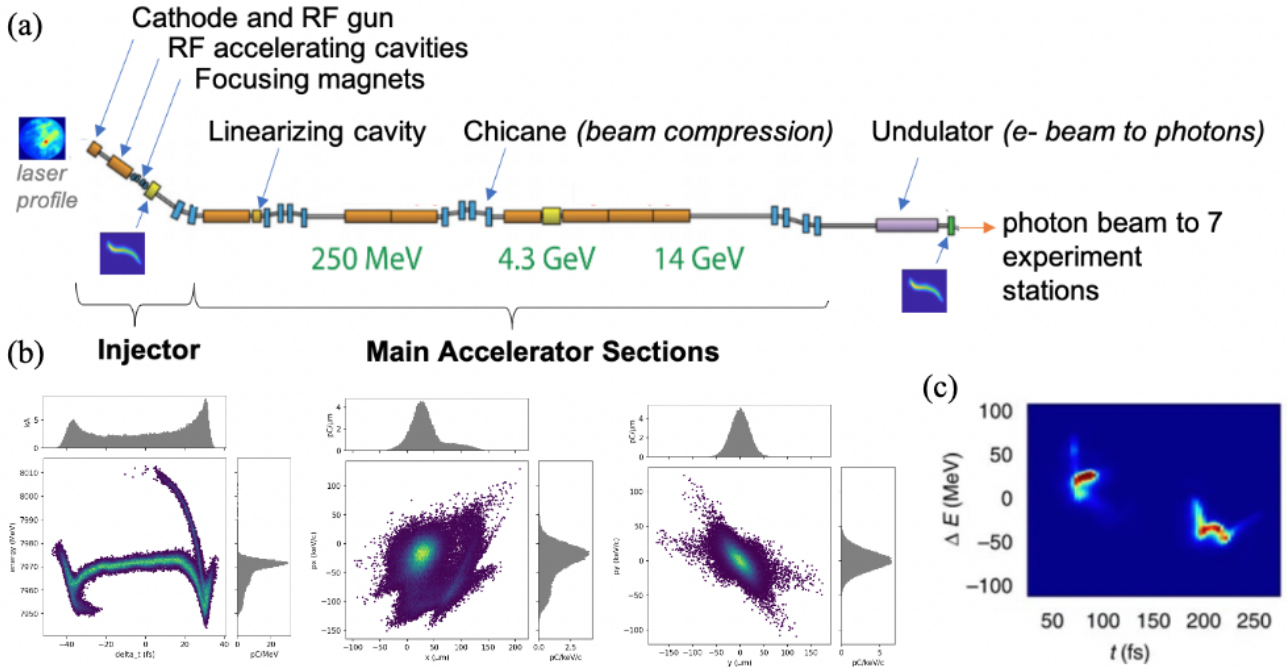


Figure 7: Example of a linear accelerator and major components (in this case the Linac Coherent Light Source) (a). Examples of 2D projections of 6D beam phase space, from physics simulations in `Bmad` [184] (b). Example of measured longitudinal (duration vs. energy) phase space in an operating mode with two electron bunches (c, reproduced from Ref. [185]).

Optimizing a particle accelerator

- **Initial design:** detailed physics simulations, and heuristic automated multi-objective optimization
  - Tradeoff speed/accuracy
  - Differentiable simulators are crucial
- **Online control:** often only a few parameters at a time, often sequential optimization
  - Modern ML-based optimization has been shown to significantly improve over standard techniques

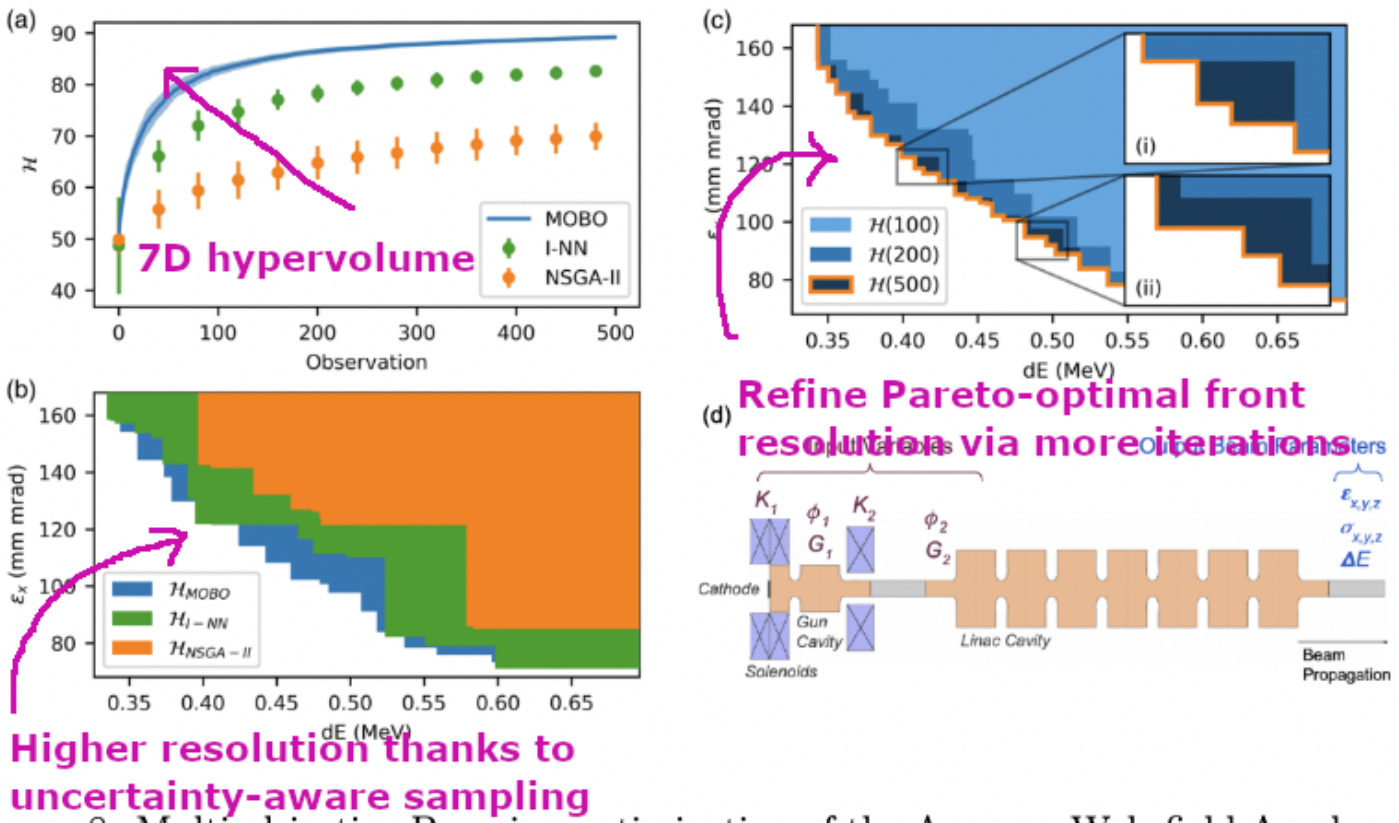


Figure 8: Multi-objective Bayesian optimization of the Argonne Wakefield Accelerator (reproduced from Ref. [189]). (a) Multi-objective Bayesian optimization is more

Beyond surrogates: MadJax (ACat proceedings 2203.00057)

- Nondifferentiable elements of the pipeline must be made differentiable
  - Surrogates
  - Software redesign
- Matrix Element computation from Madgraph made differentiable with JAX
- Embeddable in full optimization pipeline
- Future: “J(ax)EANT4”, “PYTHJ(ax)A”?

Example of interface

1. Generation:

```
generate p p > t t~, t > b udsc udscx , t~ > b~ udsc udscx
output madjax generated_ttbar
set auto_update 0
```

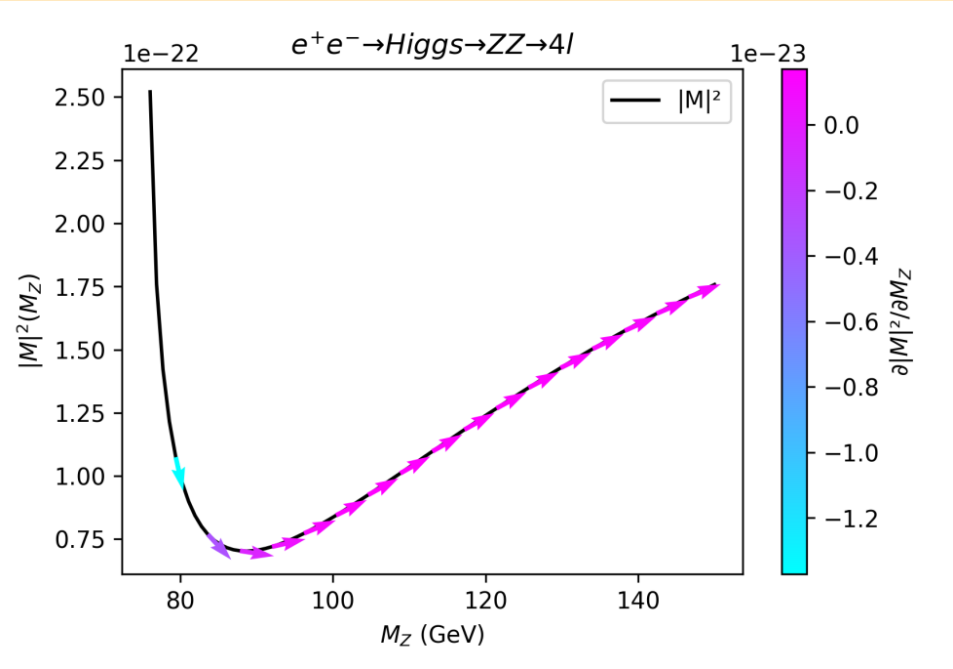
2. Evaluation:

```
import madjax
mj = madjax.MadJax('generated_ttbar')
E_cm = 14000 #GeV
process = 'Matrix_1_gg_ttx_t_budx_tx_bxdux'
matrix_element = mj.matrix_element(E_cm,process)

parameters = ('mass',6): 173.0 #set top mass
phasespace_coords = [0.1]*14 #14D phasespace

val, grad = matrix_element(parameters,phasespace_coords)
grad[('mass', 6)] #gradient wrt top mass
```

Example of results



The kind of pipeline we want to build (2203.13818)

- Support a variety of simulation packages and their required libraries
  - Make them differentiable (MadJax, GEANT4 ongoing work, etc), or use surrogates
- Support different scales of devices under optimization
  - Try to make it semi-transparent to the user
  - Opportunistic resource allocation
- API for the most common infrastructure (e.g. Kubernetes)

The optimization system should include the following main software components:

- structural storage (database) management;
- volume storage management;
- cloud compute management;
- simulation package connection and software environment configuration;
- user task management;
- optimization monitoring/benchmarking interface;
- black-box optimization runtime (Python, Julia, ...);
- differentiable optimization runtime (PyTorch [49], Tensorflow [50], JAX [88]).

The above is a non-exhaustive list, and is only meant to offer a view of the typical deployment needs of the system we imagine.

MODE: building a community

The MODE Collaboration (created 11.2020)

https://mode-collaboration.github.io/

- Joint effort of particle physicists, nuclear physicists, astrophysicists, and computer scientists

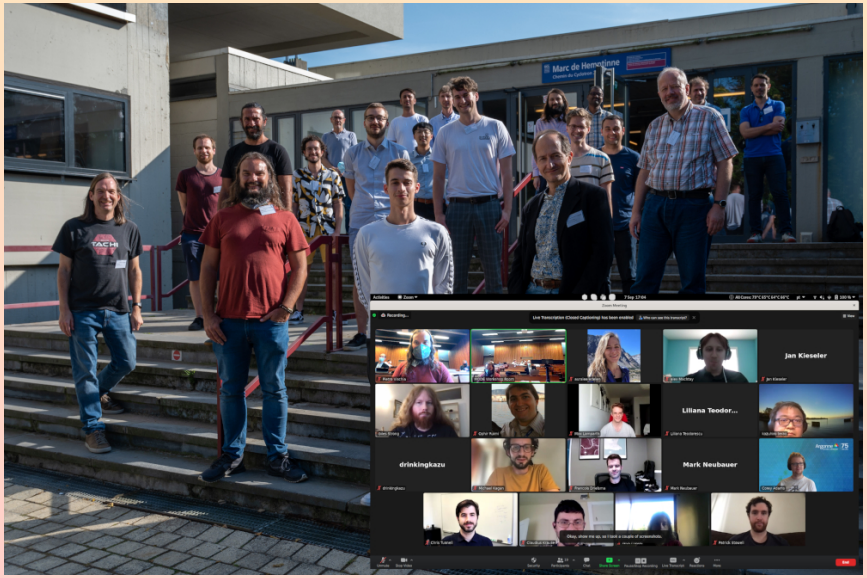
At INFN and Università of Padova Dr. **Tommaso Dorigo**, Dr. **Pablo De Castro Manzano**, Dr. **Federica Fanzago**, Dr. **Lukas Layer**, Dr. **Giles Strong**, Dr. **Mia Tosi**, and Dr. **Hevjin Yarar**
At Université catholique de Louvain Dr. **Andrea Giammanco**, Prof. **Christophe Delaere**, Mr. **Maxime Lagrange**, and Dr. **Pietro Vischia**
At Université Clermont Auvergne, Prof. **Julien Donini**, and Mr. **Federico Nardi** (joint with Universitá di Padova)
At the Higher School of Economics of Moscow, Prof. **Andrey Ustyuzhanin**, Dr. **Alexey Boldyrev**, Dr. **Denis Derkach**, and Dr. **Fedor Ratnikov**
At the Instituto de Física de Cantabria, Dr. **Pablo Martínez Ruíz del Árbol**
At CERN, Dr. **Jan Kieseler**, Dr. **Sofia Vallecrosia**
At University of Oxford Dr. **Atilim Gunes Baydin**
At New York University Prof. **Kyle Cranmer**
At Université de Liège Prof. **Gilles Louppe**
At GSI/FAIR Dr. **Anastasios Belias**
At Rutgers University Dr. **Claudius Krause**
At Uppsala Universitet Prof. **Christian Glaser**
At TU-München, Prof. **Lukas Heinrich** and Mr. **Max Lamparth**
At Durham University Dr. **Patrick Stowell**
At Lebanese University Prof. **Haitham Zaraket**
At Technische Universitàët Kaiserslautern Mr. **Max Aehle**, Prof. **Nicolas Gauger**, Dr. **Lisa Kusch** At Technische Universitàët Worms Prof. **Ralf Keidel**
At Princeton University Prof. **Peter Elmer**
At University of Washington Prof. **Gordon Watts**
At SLAC Dr. **Ryan Roussel**

The Scientific Coordinator of the MODE Collaboration is Dr. **Tommaso Dorigo**, INFN-Sezione di Padova
The Steering Board of the MODE Collaboration includes:

- Prof. **Julien Donini**, UCA
- Dr. **Tommaso Dorigo**, INFN-PD
- Dr. **Andrea Giammanco**, UCLouvain
- Dr. **Fedor Ratnikov**, HSE
- Dr. **Pietro Vischia**, UCLouvain

The MODE Workshop Series on AutoDiff for Experimental Design

- First installment: 09.2021, UCLouvain
  - Sponsored by JENAS and IRIS-HEP
- 105 participants
  - $\sim 30\%$ in person despite pandemic!
- ~ 30 talks
- Six communities
  - Muon tomography
  - Collider physics
  - Nuclear physics
  - AstroHEP
  - Neutrino physics
  - Computer science



Second installment: Kolympari (Crete) 12–16 September 2022

- 87 participants, **50 in person!**, several computer scientists (Google, TUKaiserslautern, HSWorms, UniRouen)
- Full program at https://indico.cern.ch/event/1145124/
- Keynote speakers: **Max Sagebaum** (TUKaiserslautern), **Adam Paszke** (Google Brain)
- **37 talks, 9 posters**, one data challenge

MODE: the future of experimental design

Maximum extraction of scientific value

Challenge current design concepts

Assist with a landscape of solutions

MODE

Modular pipelines powered by autodiff

Make generators differentiable where possible

Create and guide a multidisciplinary community