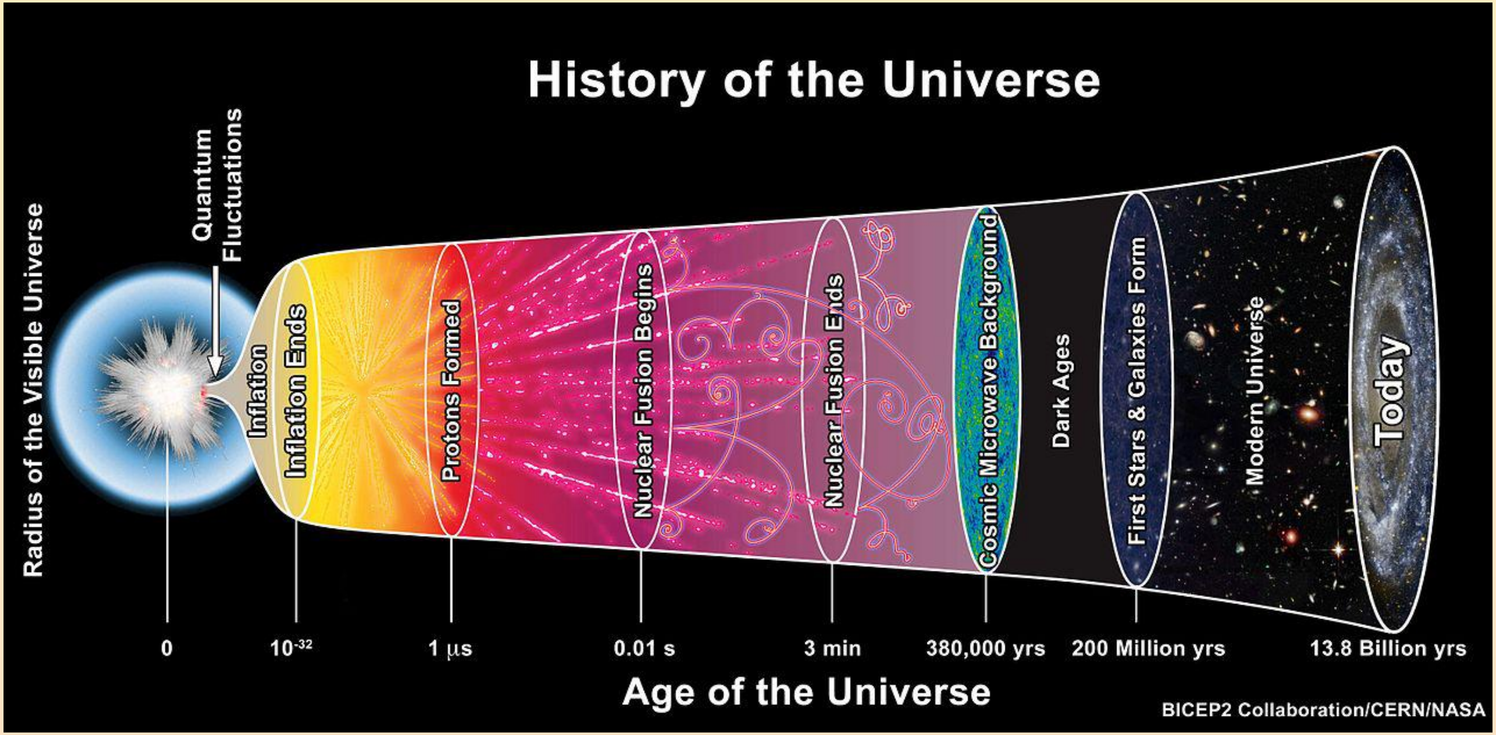
Optimizing Experiment Design with Machine Learning

STAMPS Seminar at Carnegie Mellon University

Dr. Pietro Vischia
pietro.vischia@cern.ch
@pietrovischia

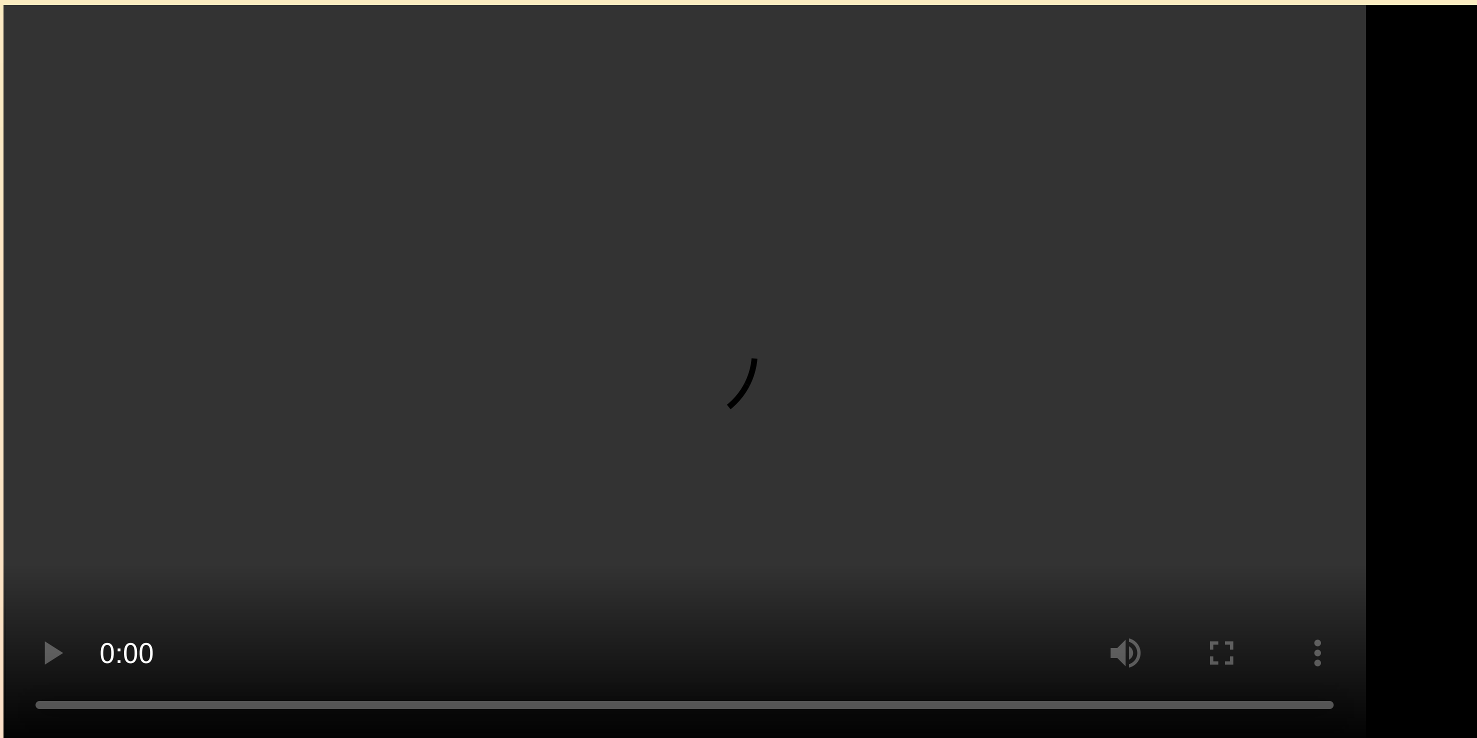
We Try to Understand the Universe



—
Image from CERN

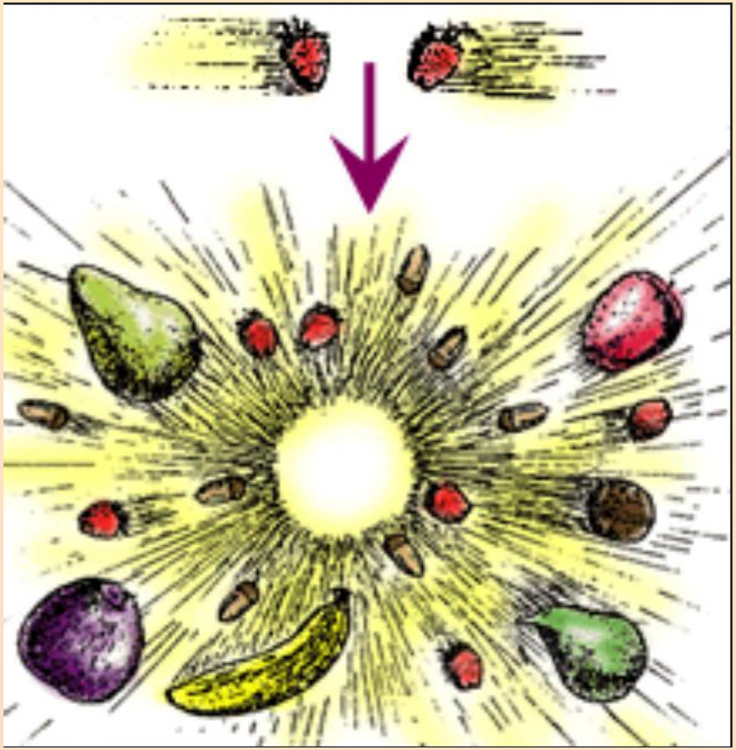
Collide Protons...

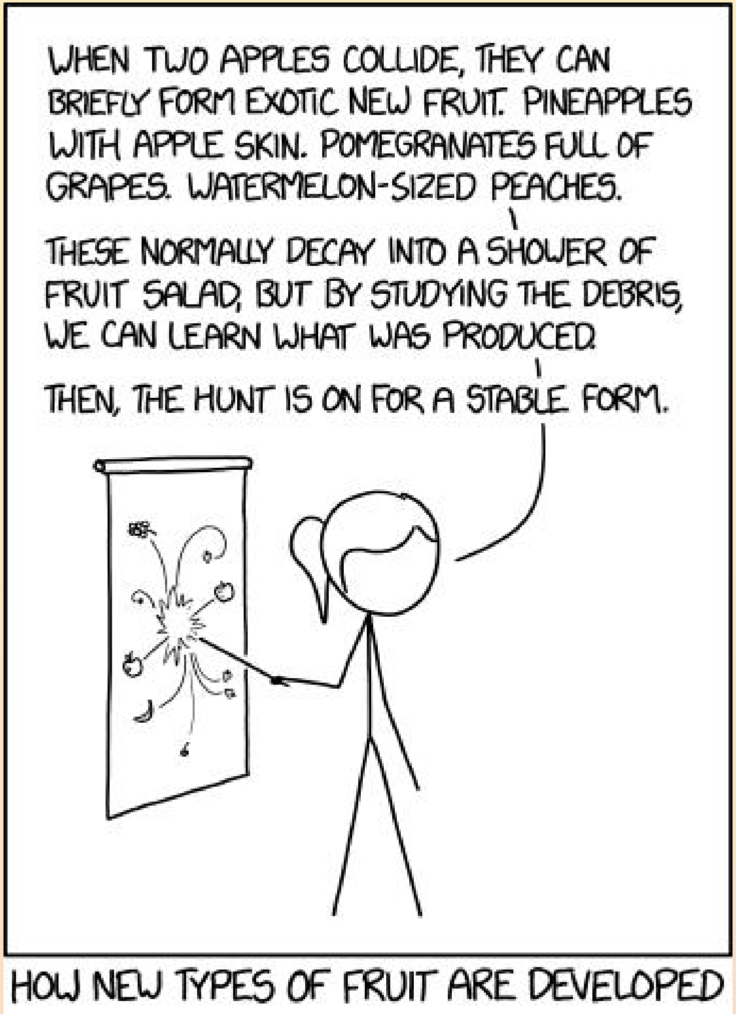
- The kinetic energy of two 88k tons aircraft carriers, each at 10km/h
- Packed into a transverse section of 16 micron



—
Video from YouTube

To See What Happens

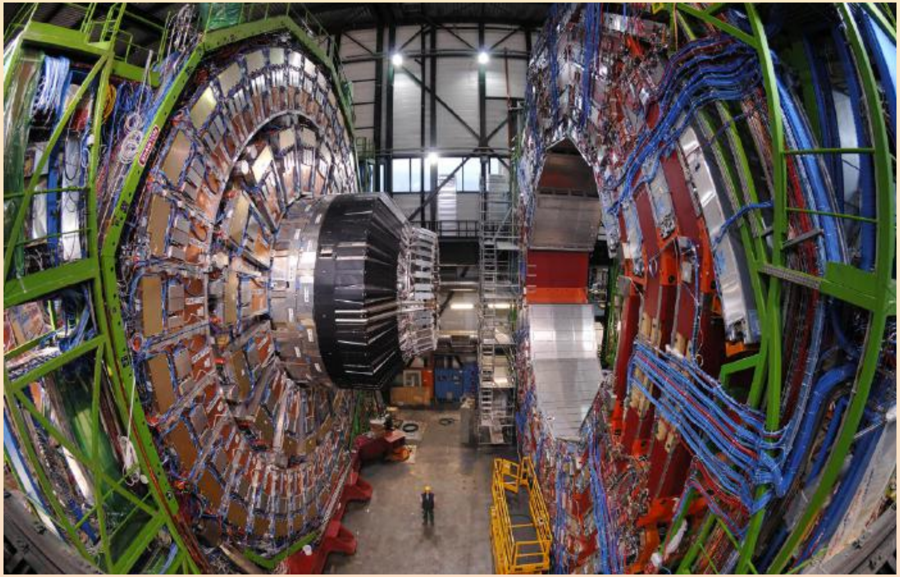




—
Images from ParticleAdventure.org and XKCD.com

Complex Experiments

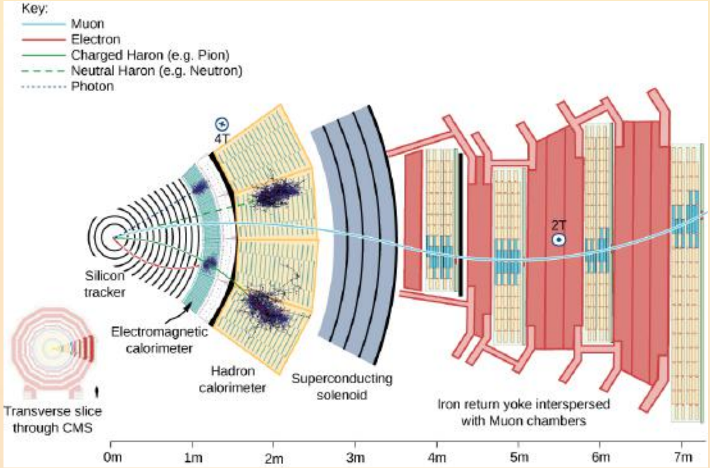
- Costly full simulation
- Interrelation of many parameters
- Large number of optimizable subdetectors
- Complexity prevents from optimizing targeting final goals



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 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."

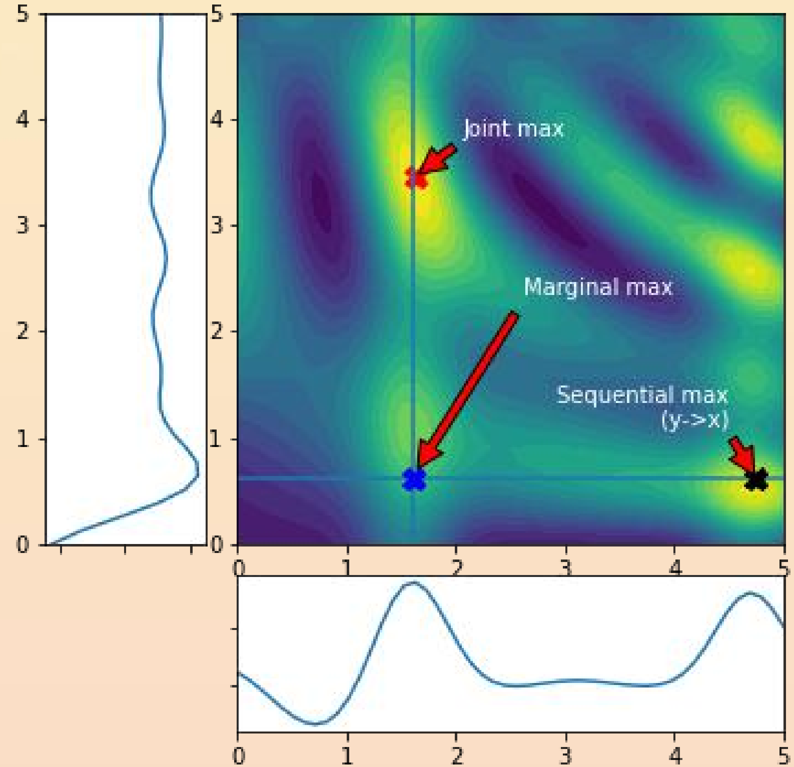
—
Slide by Lukas Heinrich at QCHS 2022. Referenced paper is doi:10.2514/6.2006-7242

Joint optimization: why?

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

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

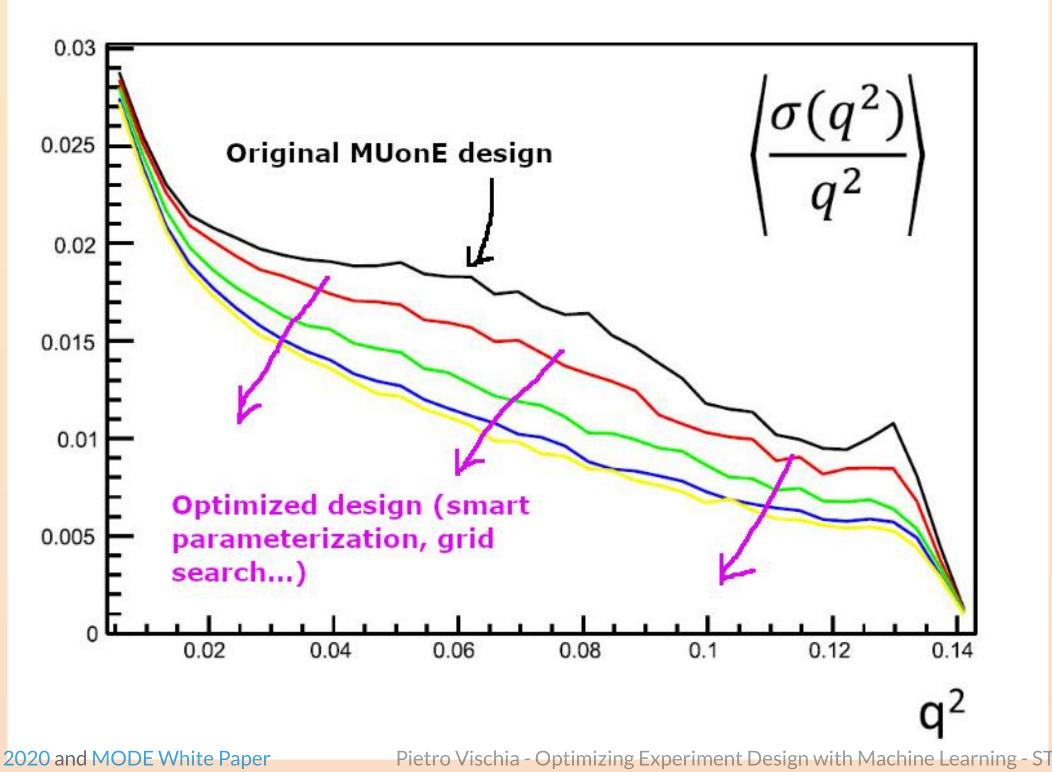
$$argmin_{x,y}\Big(\mathcal{L}(x,y)\Big)_y \neq argmin_y\Big(\mathcal{L}(x,y)\Big)$$



—
Illustration (c) P. Vischia, book in preparation

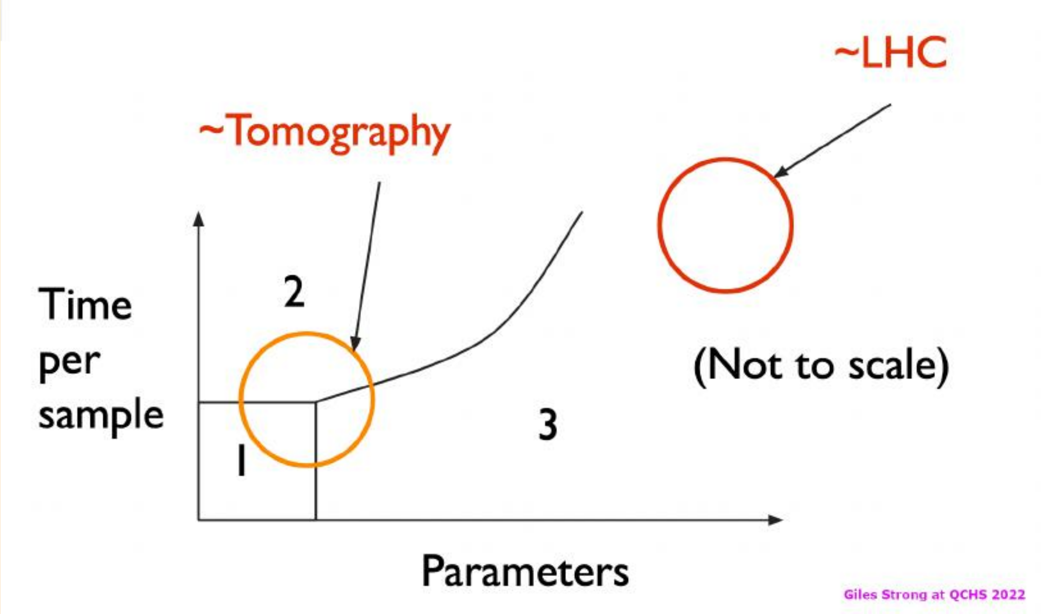
Large gains to be had

- MUonE: proposed 150 GeV muon beam experiment to be built at CERN
  - Measure precisely the q^2 differential cross section in electron-muon scattering
  - 40 tracking stations and a calorimeter
- Dramatic improvement in the resolution on q^2 even from a simple grid search



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

Different challenges require different methods



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

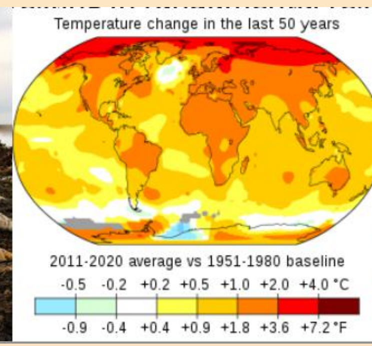
A moral imperative

Optimize...

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

...within constraints

- Unprecedented global challenges
- Society less receptive to fundamental research



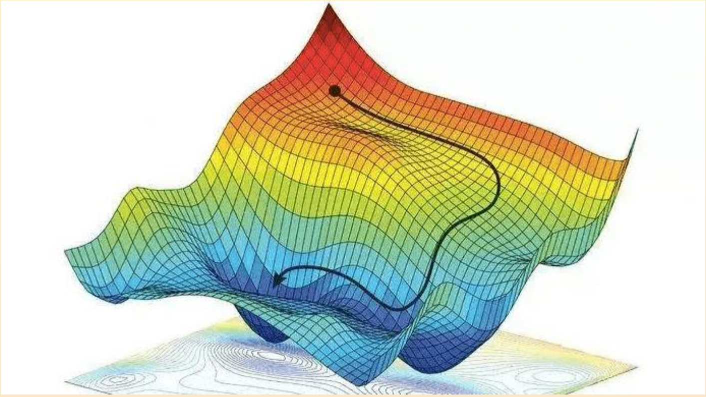
Maximum extraction of scientific value from the available resources

—

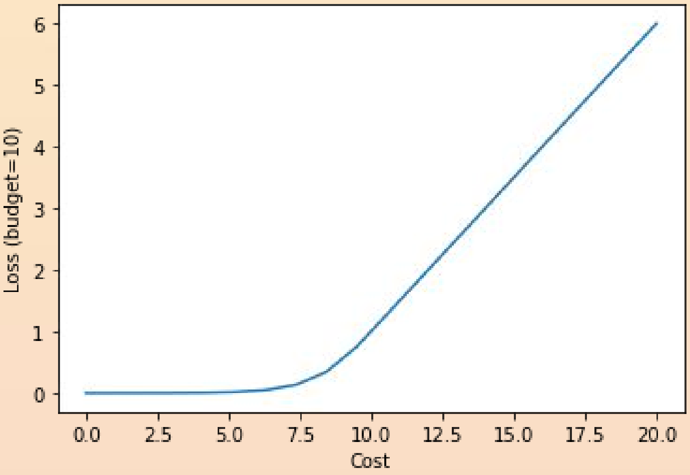
Image and concept by T. Dorigo, 2020 update of the European Strategy for Particle Physics (EUSUPP)

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 the MODE White Paper

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

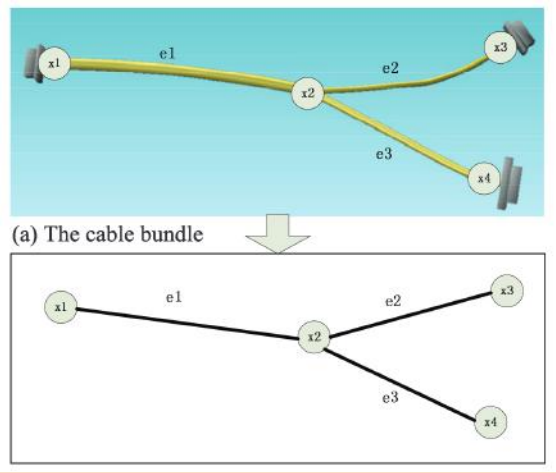
- Material availability (influenced e.g. also by wars) is also a concern, nowadays

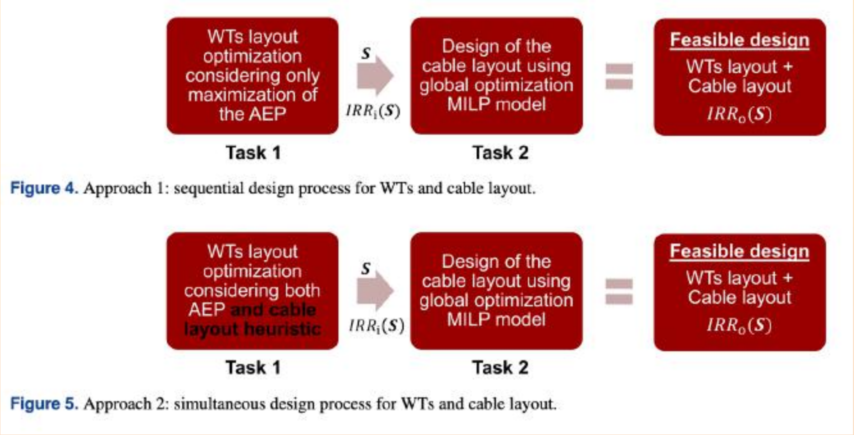
Figure 6: Road transport of a structure for the ATLAS air toroids. Photo reproduced from Ref. [181].

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

Maybe we can optimize cable layout

- Easily description as trees or graphs
- Alhough 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



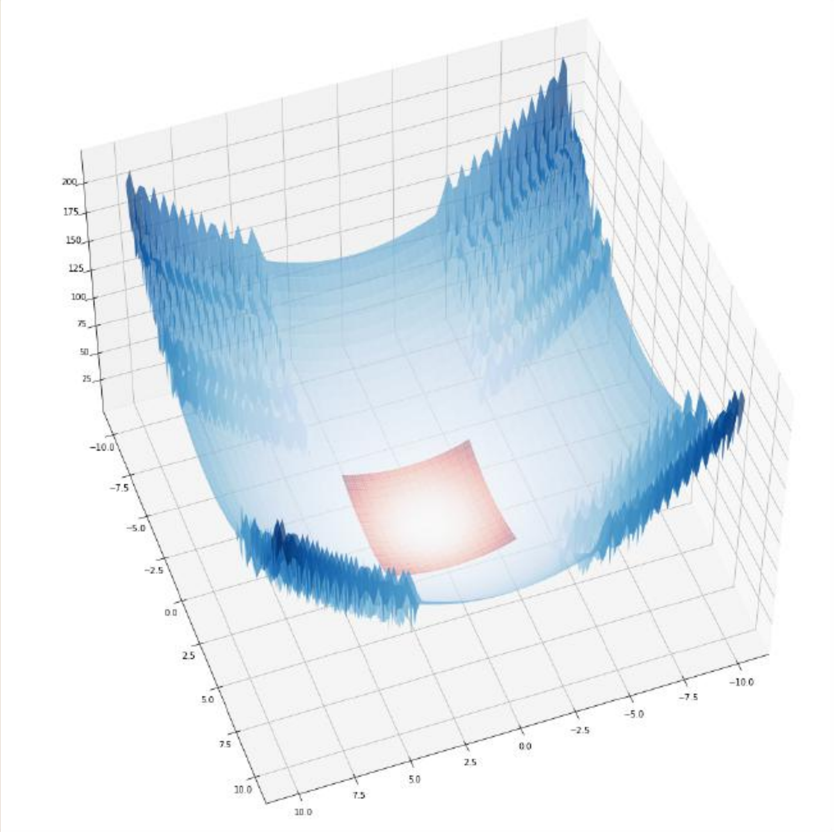


—

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

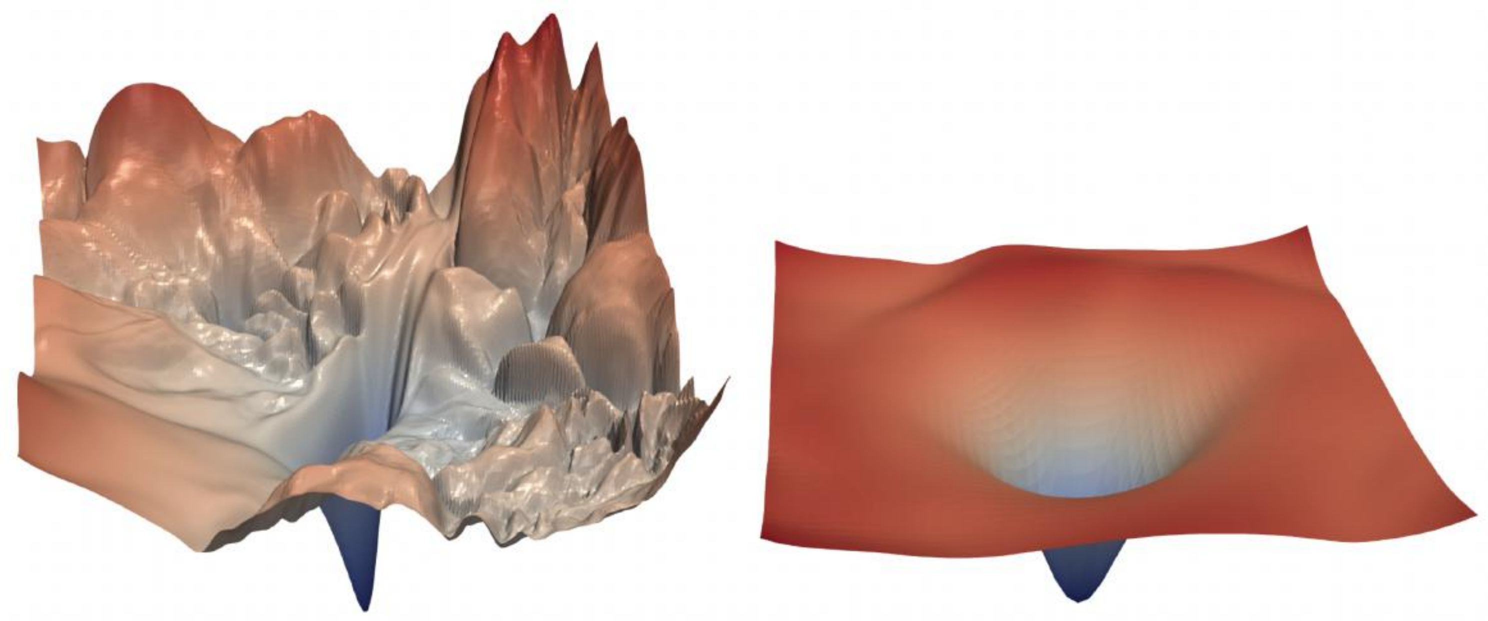
- Cannot parameterize everything
- The optimal solution: unrealistic
- Provide feasible solutions near optimality
- The physicist will fine tune



—
Illustration (c) P. Vischia, book in preparation

How far from optimality?

- Can we define in a general way an acceptable increase in loss?
  - Tradeoff performance/cost



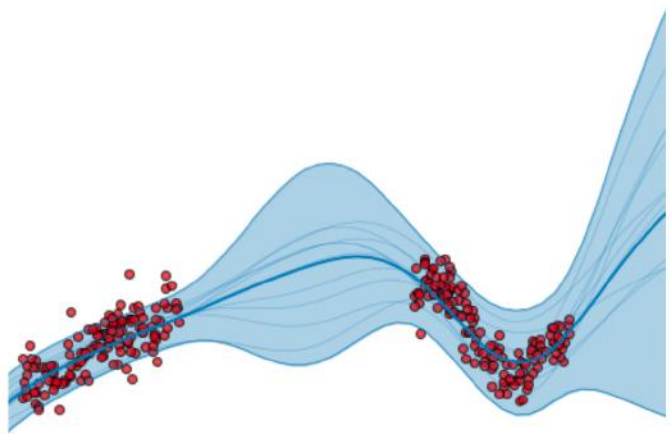
—

Image from Li et al (taken as pictorial representation out of its original context)

Maybe we should marginalize?

What is Bayesian learning?

- The key distinguishing property of a Bayesian approach is **marginalization** instead of optimization.
- Rather than use a single setting of parameters **w**, use all settings weighted by their posterior probabilities in a *Bayesian model average*.



Andrew Wilson at Hammers&Nails 2022

7 / 85

How to optimize an experiment

- We detailed our idea in the MODE White Paper
  - 109-page document drafting the way forward, joint with computer scientists from proton Computed Tomography
  - under revision for Reviews in Physics

Contents

1 Introduction 5

2 The State of the Art in Design Optimization and Differentiable Programming 9

3 Problem Description and Possible Solution 18

4 Example Use Cases 33

5 System Architecture and Requirements 80

6 Conclusions 83

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/

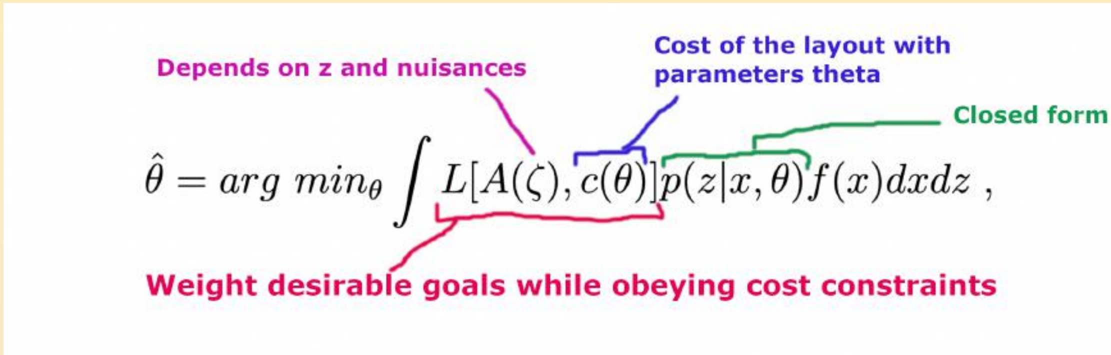
s-det] 22 Mar 2022

Ingredients

- 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

—
From our white paper

Optimization recipe



- For example, to identify smuggled material in a container

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

—
From our white paper

Differentiable happiness

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

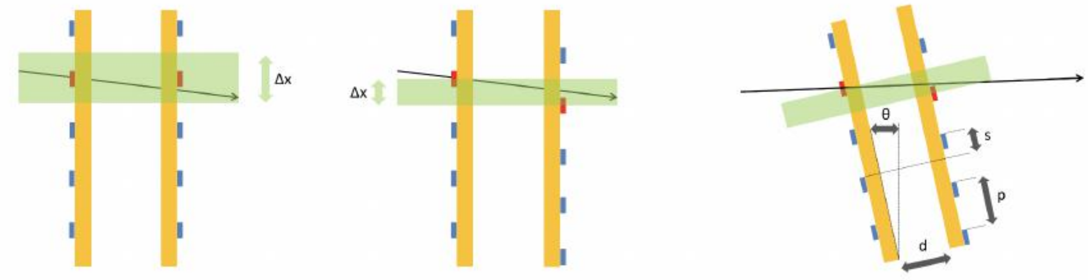


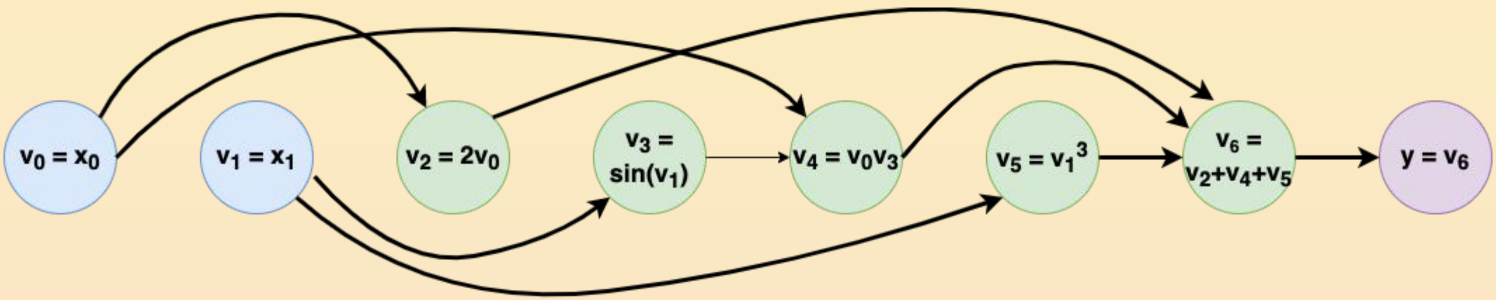
Figure 4: Left and center: a double-sided silicon strip sensor produces twice smaller resolution Δx on single-strip hit position for an orthogonally incident particle if strips on the two sides are staggered by half the strip pitch. Right: the four parameters affecting single-strip hit position resolution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

—

From our white paper

Automatic differentiation

$$z(x, y) = 2x + x\,sin(y) + y^3$$



Forward mode

- To the extreme, $f : \mathbb{R} \rightarrow \mathbb{R}^m$
- Evaluates $(\frac{\partial f_1}{\partial x}, \ldots, \frac{\partial f_m}{\partial x})$

Reverse mode

- To the extreme, $f : \mathbb{R}^n \rightarrow \mathbb{R}$
- Evaluate $\nabla f(\mathbf{x})(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})$

- Computational cost of calculating $\mathbf{J}_f(\mathbf{x})$ for $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ in $\mathbb{R}^n \times \mathbb{R}^m$

$\mathcal{O}(n\ \mathrm{time}(f))$ $\qquad$ $\mathcal{O}(m\ \mathrm{time}(f))$

When the likelihood is intractable

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

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

- $F(\cdot)$ 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
  - Descend to the minimum of approximated loss by following surrogate gradient

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

—

From our white paper

Advantages of surrogates

- 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 $F(\cdot)$

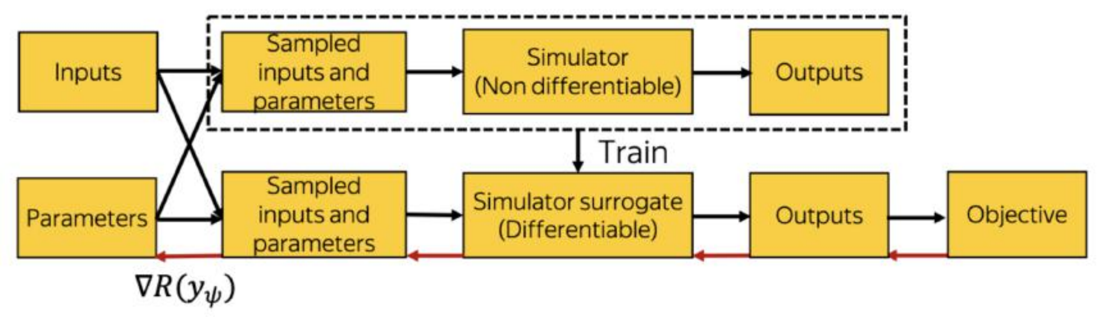


Figure 1: Simulation and surrogate training. *Black:* forward propagation. *Red:* error backpropagation.

—
Image from Shirobokov,..., Kagan, ... et al.

Vast set of use cases

- Already exploring
    - Muon tomography
    - LHCb and CMS calorimetry
    - SWGO placement/geometry of tanks
    - LEGEND optimization

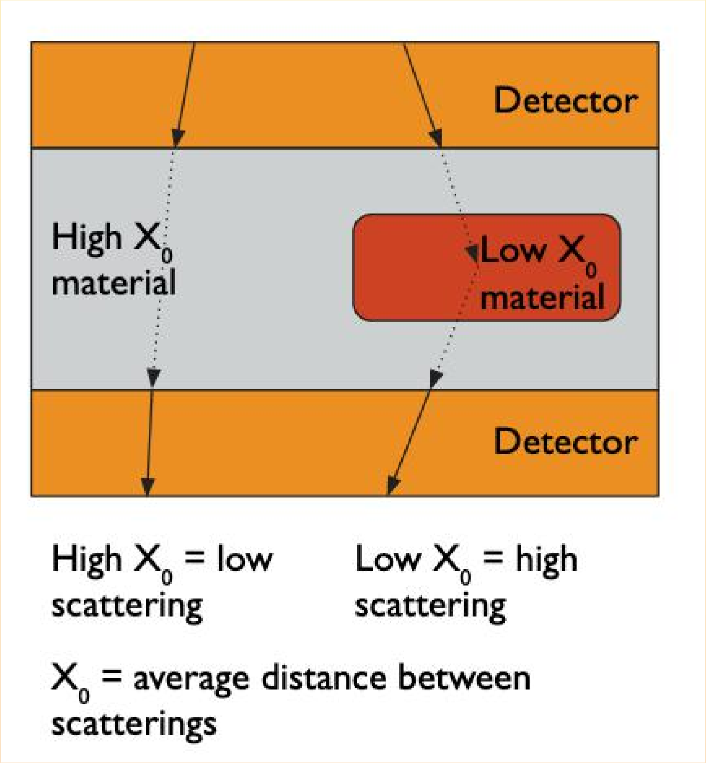
- 4 Example Use Cases
    - 4.1 Experiments at Accelerators
        - 4.1.1 Particle Accelerator Design and Control
        - 4.1.2 Calorimeter Optimization
        - 4.1.3 Hybrid Calorimeter for a Future Particle Collider
        - 4.1.4 Electromagnetic Calorimeter of a Muon Collider Experiment
        - 4.1.5 Optimization of the MUonE Detector
        - 4.1.6 Searches for Milli-charged Particles
    - 4.2 Astro-particle Physics and Neutrino Experiments
        - 4.2.1 High-Energy Gamma-Ray Astronomy
        - 4.2.2 Interferometric Gravitational-Wave Detectors
        - 4.2.3 Radio Detection of High-Energy Neutrinos
    - 4.3 Cosmic-Ray Muon Imaging
        - 4.3.1 Figures of Merit
        - 4.3.2 Parameters of the Optimization Task
        - 4.3.3 TomOpt: Differential Muon Tomography optimization
        - 4.3.4 Industrial Applications
        - 4.3.5 Portable Modular Detectors for Flexible Muography
    - 4.4 Proton Computed Tomography
    - 4.5 Low-Energy Particle Physics
    - 4.6 Error Analysis of Monte Carlo Data in Lattice QCD

—

From our white paper

Muon Tomography

- Want to infer properties (e.g. 3D map of elemental composition) of unknown volume
  - Shipping container, archeological site, nuclear waste dump, industrial machinery, etc.
- Muons from cosmic rays traverse us all the time
  - On average, 1 muon per cm^2 per minute
  - Change in kinematics provides handle for inference on X_0

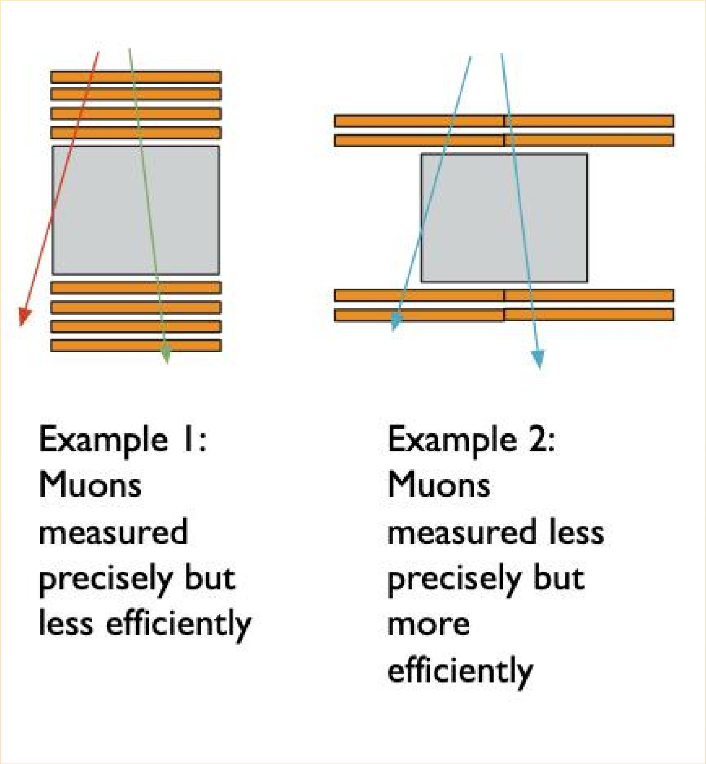


—
Figure from the TomOpt project

Domain knowledge is not enough

- Domain knowledge typically provides heuristics based on proxy objectives

- Will likely have a budget
  - Money, heat, power, positioning of detectors, imaging time...
- Will likely have varying purposes
  - Today want to spot uranium, tomorrow e.g. drugs

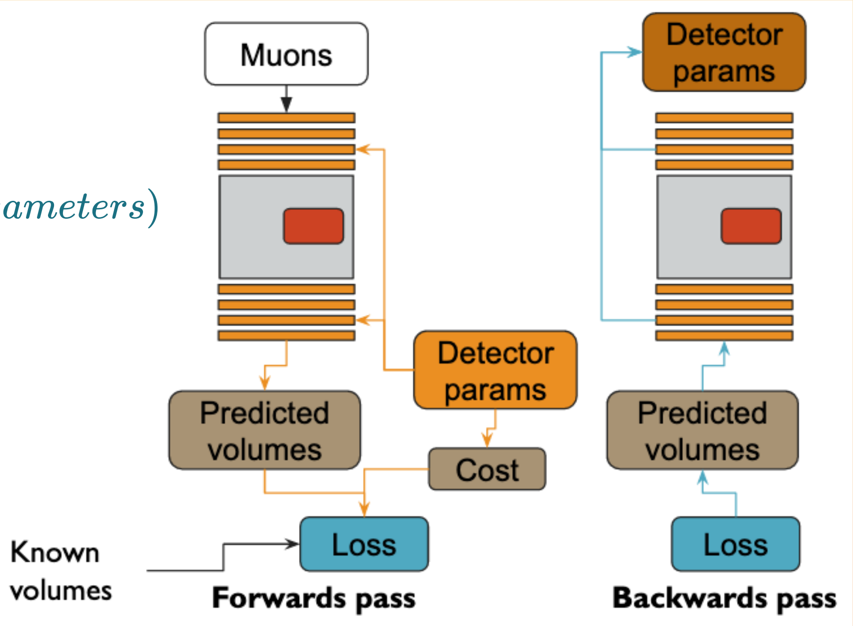


—
Figure from the TomOpt project

TomOpt

- Differential optimization of muon-tomography detectors (ongoing project)
    - Giles C. Strong, Tommaso Dorigo, Andrea Giammanco, Pietro Vischia, Jan Kieseler, Maxime Lagrange, Mariam Safieldin, Federico Nardi, Anna Bordignon, Haitham Zaraket, Max Lamparth, Federica Fanzago, Oleg Savchenko, Nitesh Sharma
    - Modular design in python, autodiff via PyTorch

- Inference chain as differentiable pipeline
    - Can compute $p(\Delta output|\Delta detector\ parameters)$
- Task as loss function
    - Including target (e.g. prediction uncertainty), costs, constraints
- Backpropagate and optimize as usual
    - Gradient descent



—
Figure from the TomOpt project

Muon Generation

- Formulas by 2015 and 2016 models
- Account for Earth's curvature
- Code handles many muons at once (batch)

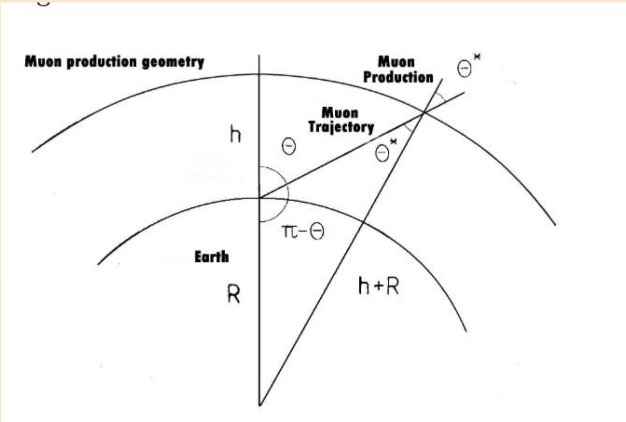
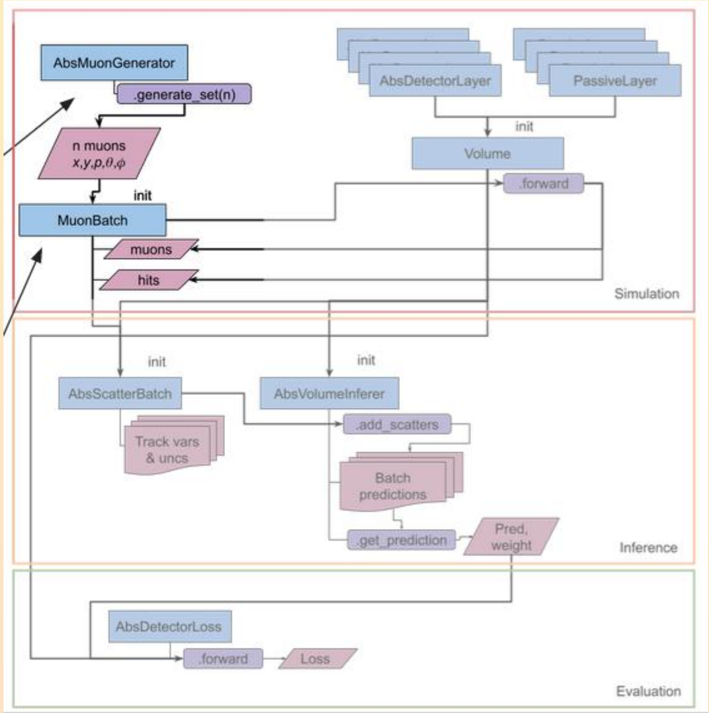


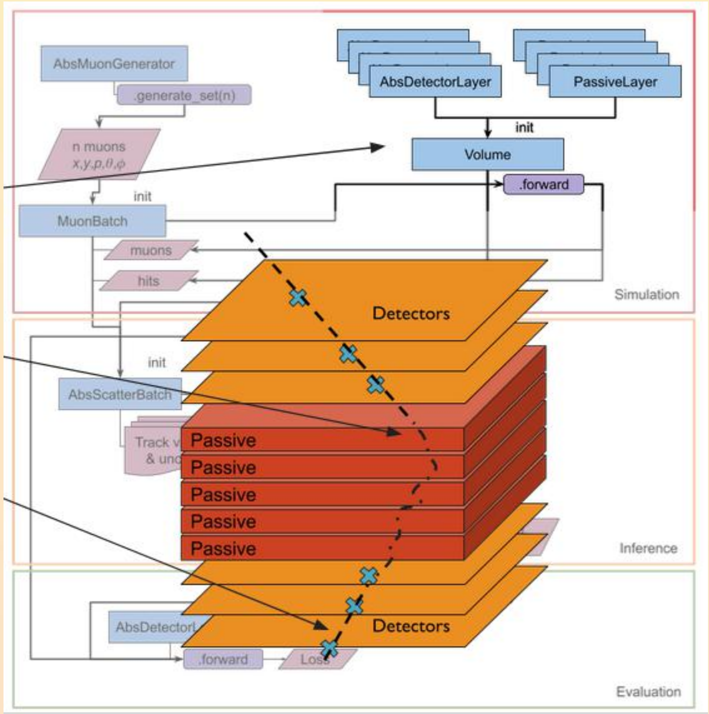
Fig. 1. The relation of the observed zenith angle of muons, θ^*, to the zenith angle at the muon production point in the atmosphere, θ. R is the radius of the Earth. Adopted from [3][4]



—
Left: Guan et al, Right: G.C.Strong

Volume Specification

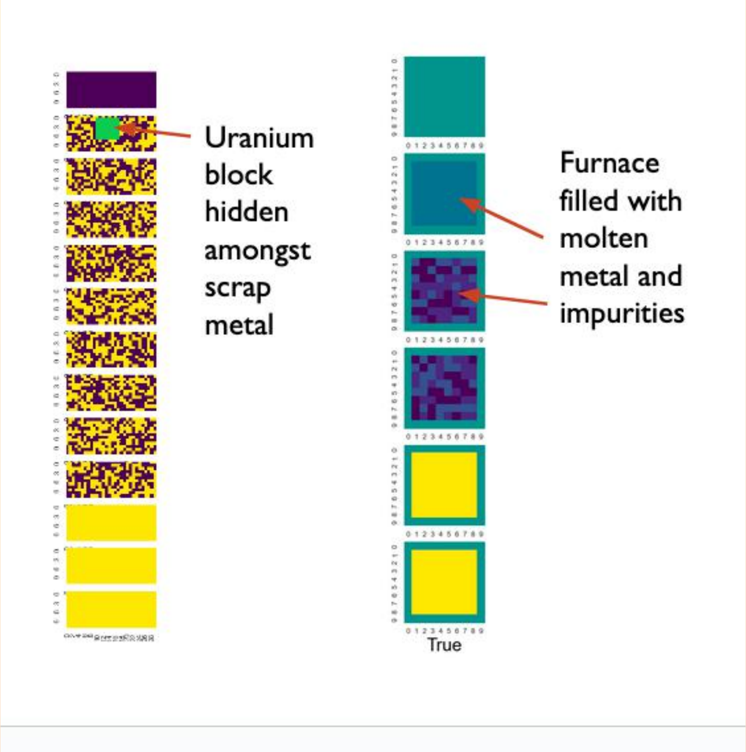
- Volume made up of stacked layers in z
- Passive layers scatter muon
  - PDG and GEANT models both available
  - Voxelized passive layers (x, y)
- Active layers record muon hits
  - Parameterized efficiency and resolution (cost, physics constraints)



—
Figure by G.C.Strong

Monte Carlo Truth

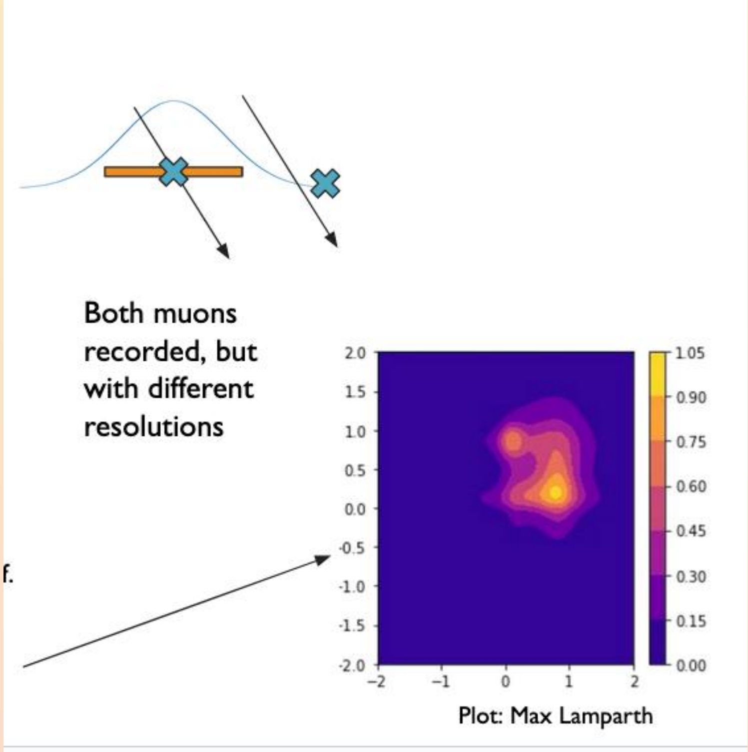
- Per each scenario, can build voxelized random volumes
  - Each voxel can be a different material
  - tomoNext: material mixture per voxel



—
Figure by G.C.Strong

Make muon hits differentiable

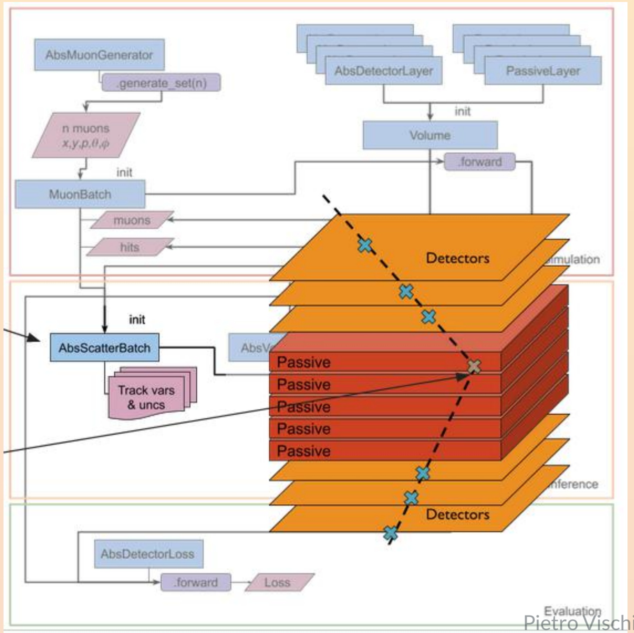
- Associate a distribution to resolution and efficiency
  - e.g. Gaussian centered on panel and width equal to panel span
  - p.d.f. of the muon position is now differentiable
- Further generalization: Gaussian Mixture models



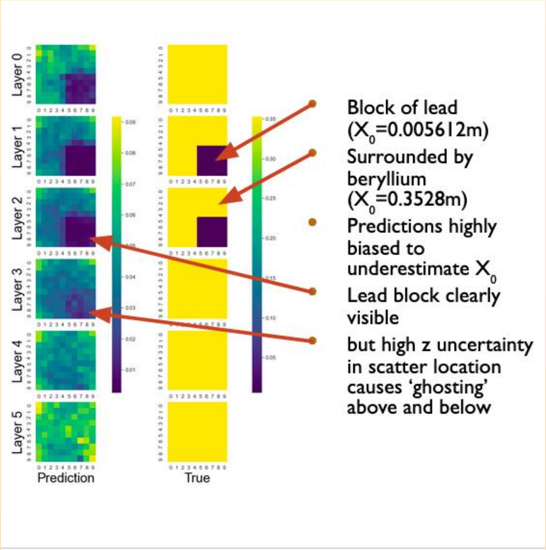
—
Figure by G.C.Strong

From hits to tracks

- Analytic maximum likelihood fit
  - considering uncertainty and efficiency of hits
  - fully differentiable w.r.t. detector parameters
- Provides track parameters and their uncertainties



- POCA (POint of Closest Approach)
  - assume one scattering in one point
  - invert model to compute X_0
  - average X_0 per voxel



—
Figure by G.C.Strong

Volume Inference with Expectation Maximization

- Iterative algorithm
  - estimate scatter density based on current estimate of the image
  - update estimate of the image based on the estimated scatter density

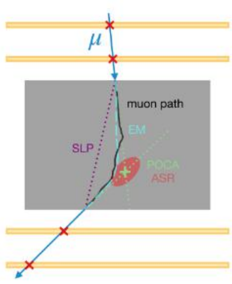
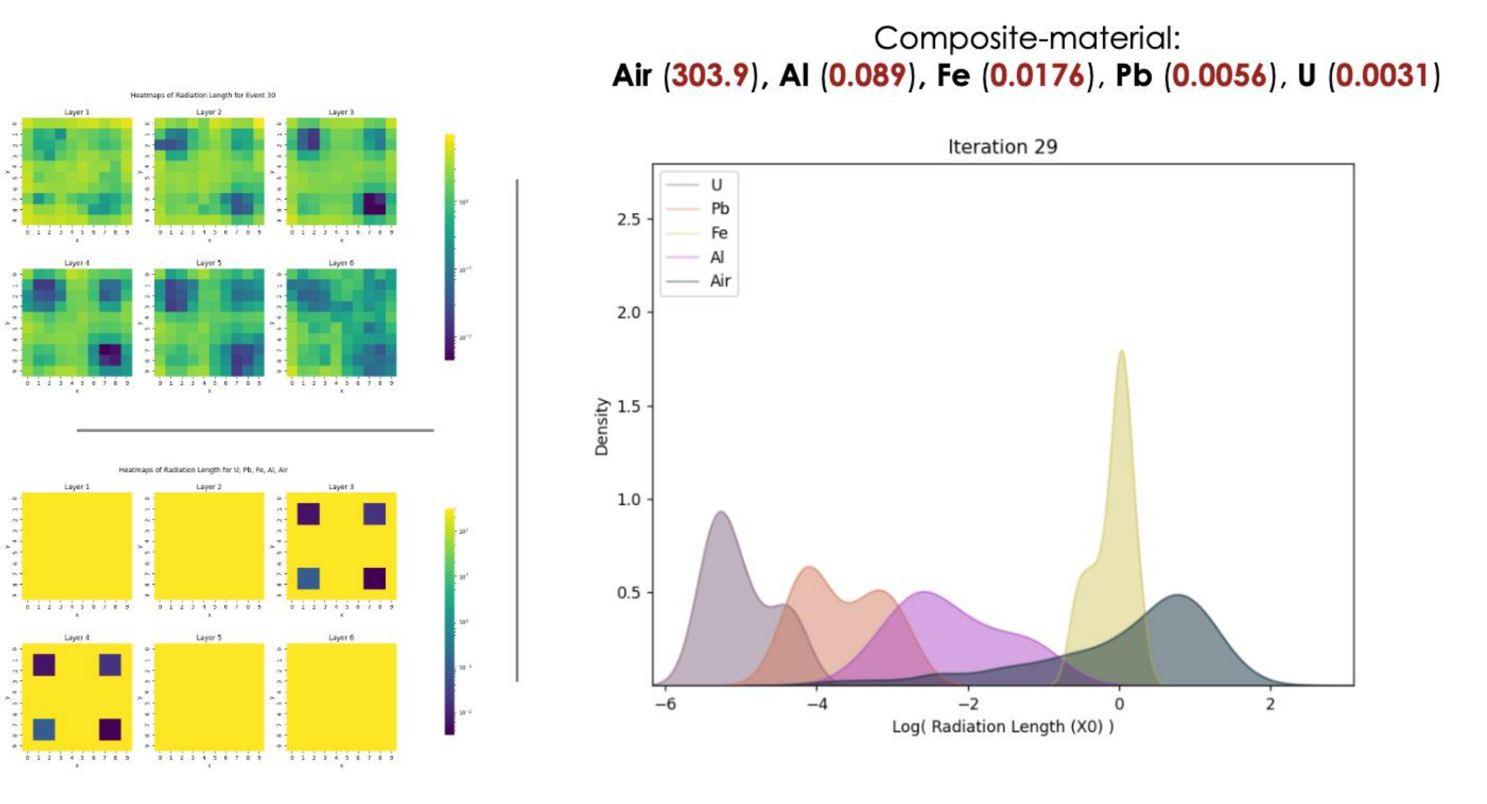


Figure 2: Illustration of commonly used algorithms for reconstructing images in muon scattering tomography. Barnes et al. 2023

1) Gather measurements of scattering and momentum for each muon i=1 to M: $(\Delta\theta_x, \Delta\theta_y, \Delta x, \Delta_y, p_r^2)_i$.
2) Estimate geometry of interaction of each muon with each voxel j=1 to N: $(L, T)_{ij}$.
3) For each muon voxel pair, compute the weight matrix: W_{ij} using (24).
4) Initialize the scattering density in each voxel with a guess : $\lambda_{j,old}$.
5) Do until (stopping criteria)
   a) For each muon, compute $\Sigma_{D_i}^{-1}$ using (29) and taking the inverse.
   b) For each muon voxel pair, compute the conditional expectation terms: S_{ij} using (43).
   c) Compute $\lambda_{j,new}$ using (38)
   d) Set $\lambda_{j,old} = \lambda_{j,new}$ for all voxels.
6) End do

Summary of the EM algorithm for muon tomography. Schultz et al. 2007

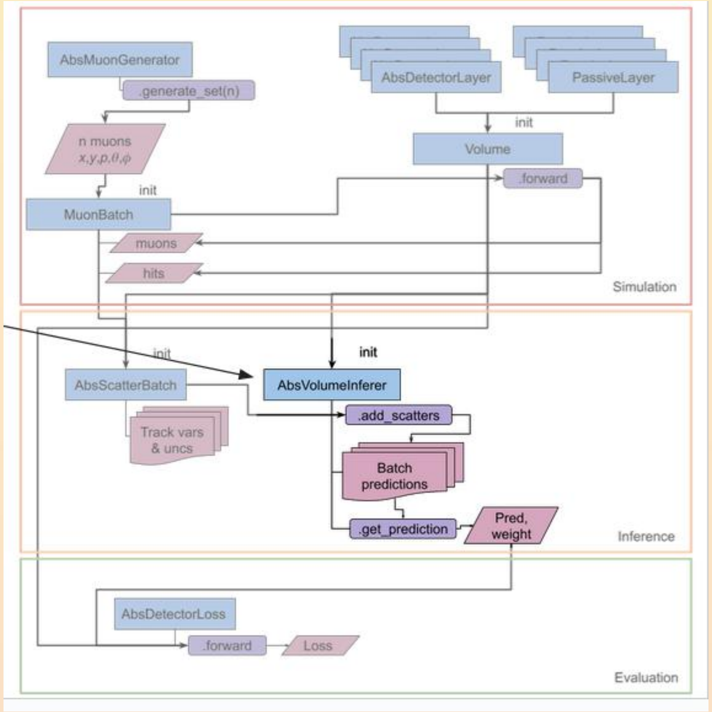
EM preliminary performance



Volume inference

- The main boundary is that inference algorithm must be differentiable

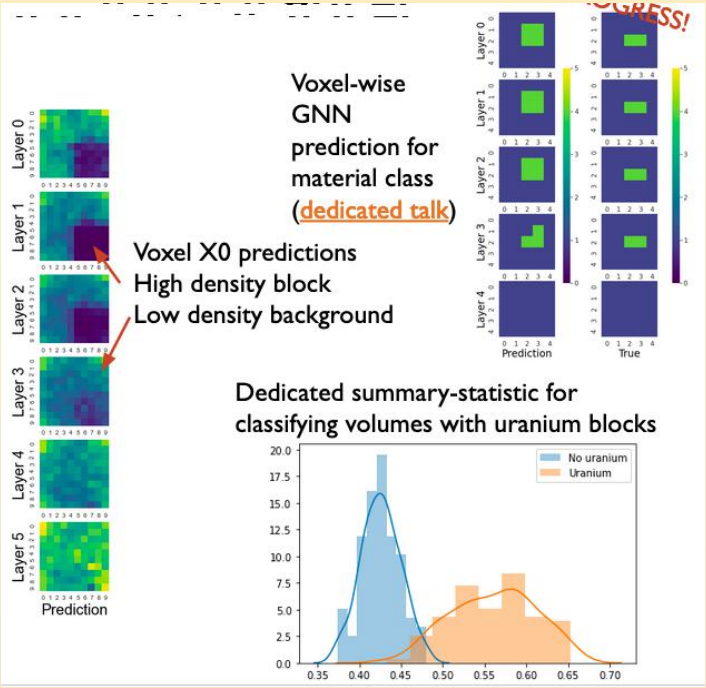
- Basic approach of inverting scatter model to compute X_0 is highly biased
- Maybe a task-specific summary statistic would work better



—
Figure by G.C.Strong

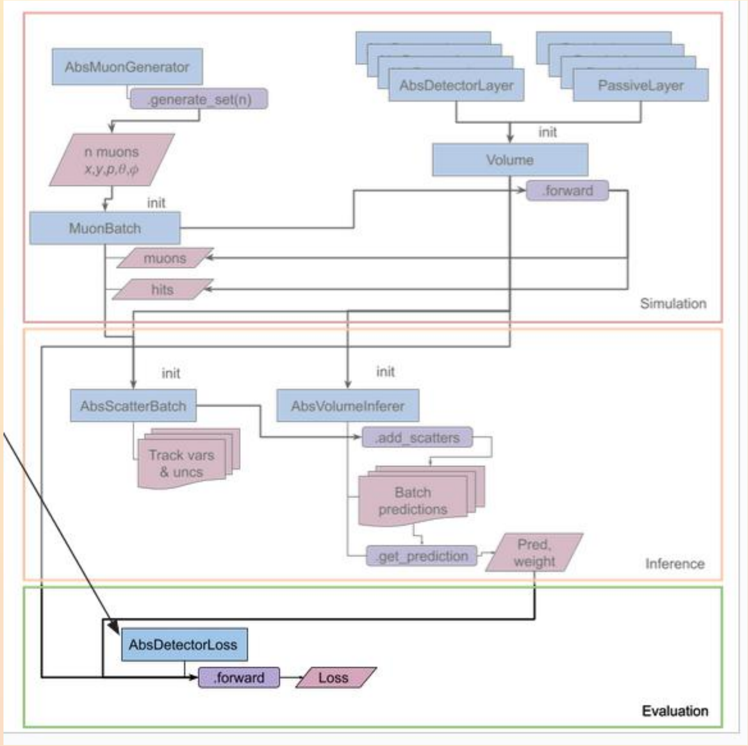
It's all about summaries

- Promising performance
- Muon track quantities differentiable
  - can compute uncertainties due to spatial resolution
  - useful for aggregating
- Summary can be learned
  - Graph neural network (see G.C.Strong's talk)

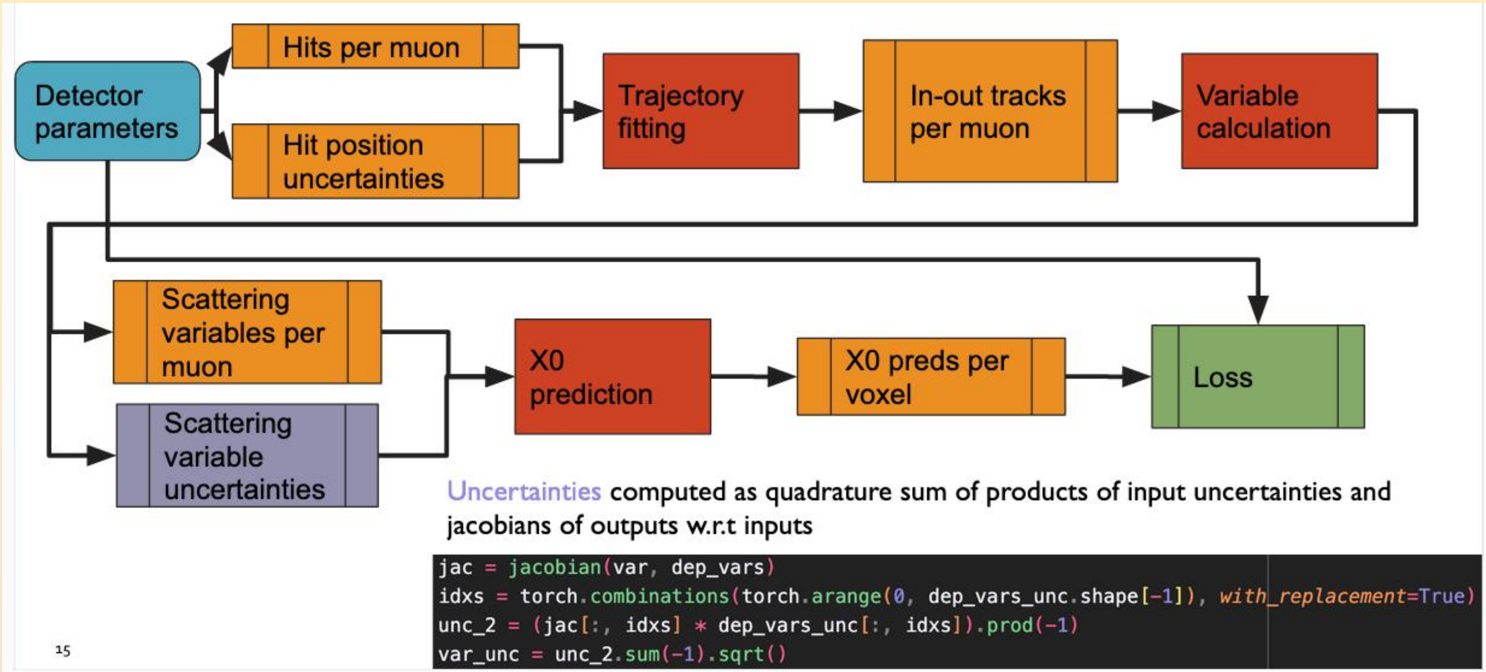


Optimization

- Regular gradient descent of a loss function
  - account for cost of the detector and other constraints
  - standard optimisers to update detector parameters



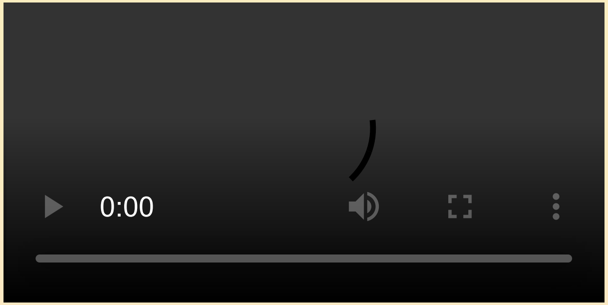
Pipeline

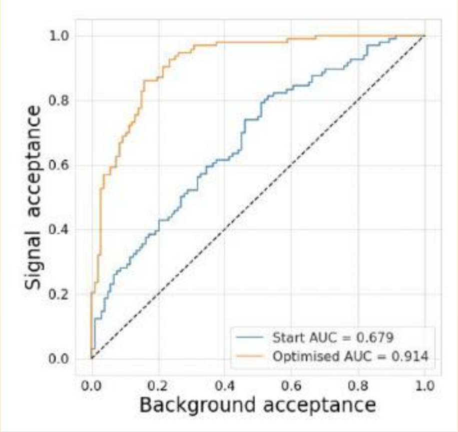


—
Figure by G.C.Strong

Encouraging results

- Inference via low-dim summaries
- E.g. identify uranium in container





—
Material from `TomOpt` draft paper

The MODE Collaboration

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

- Joint effort (created 11.2020) 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 Universitäët Kaiserslautern Mr. **Max Aehle**, Prof. **Nicolas Gauger**, Dr. **Lisa Kusch** At Technische Universitäët 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

Series of yearly workshop

- First installment in Louvain-la-Neuve (Belgium)
- Second installmentnt in Kolymbari (Greece)
  - 37 talks, 9 posters, one data challenge with prizes, recordings will be online soon
- You are all invited to the Third MODE Workshop, to be held in Princeton (USA)