ASTROCENT

WMLQ 2026
08.09.2026

Automated Data/MC Tuning via Phase-Space Optimization

Piotr Kalaczyński
Medical Physics & Radiation Detection Group, Astrocent

Nicolaus Copernicus Astronomical Center
Polish Academy of Sciences

AGH University of Krakow

CEAI
Center of Excellence in Artificial Intelligence

Funded by the European Union | European Funds for Smart Economy | Republic of Poland | Co-funded by the European Union

Motivation

A common problem: data ≠ MC, we need quality cuts

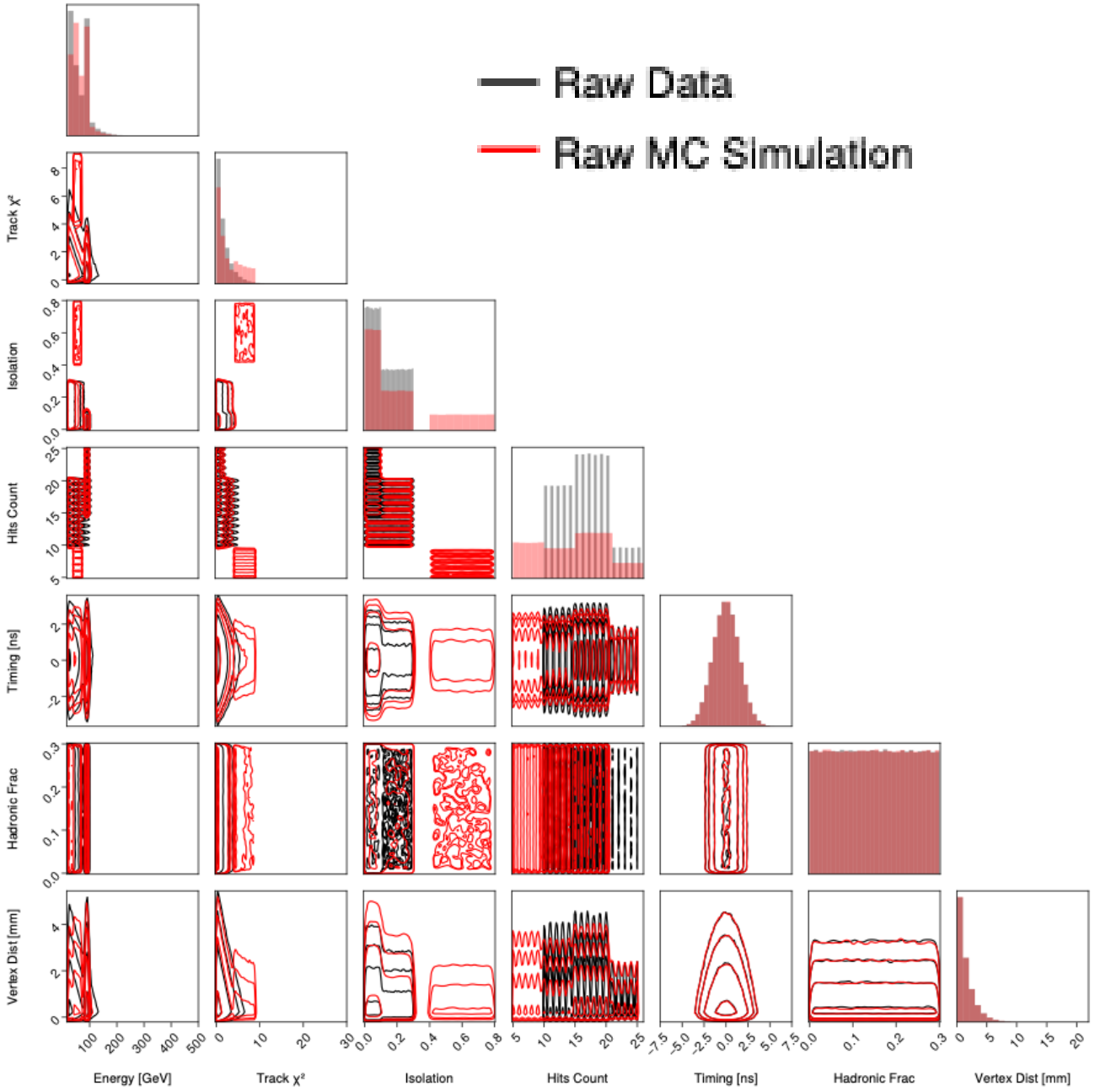
- The usual way:
  - manual adjustment
  - visual inspection
  - Slow
  - Biased
- What I suggest:
  - Automated optimization
  - Analogous to tuning ML model hyperparameters
  - But our target is a metric describing the agreement between 2 distributions (χ^2, …)
  - Less human bias
  - Faster
  - Better results



github.com/pkalaczynski/data_MC_autotune

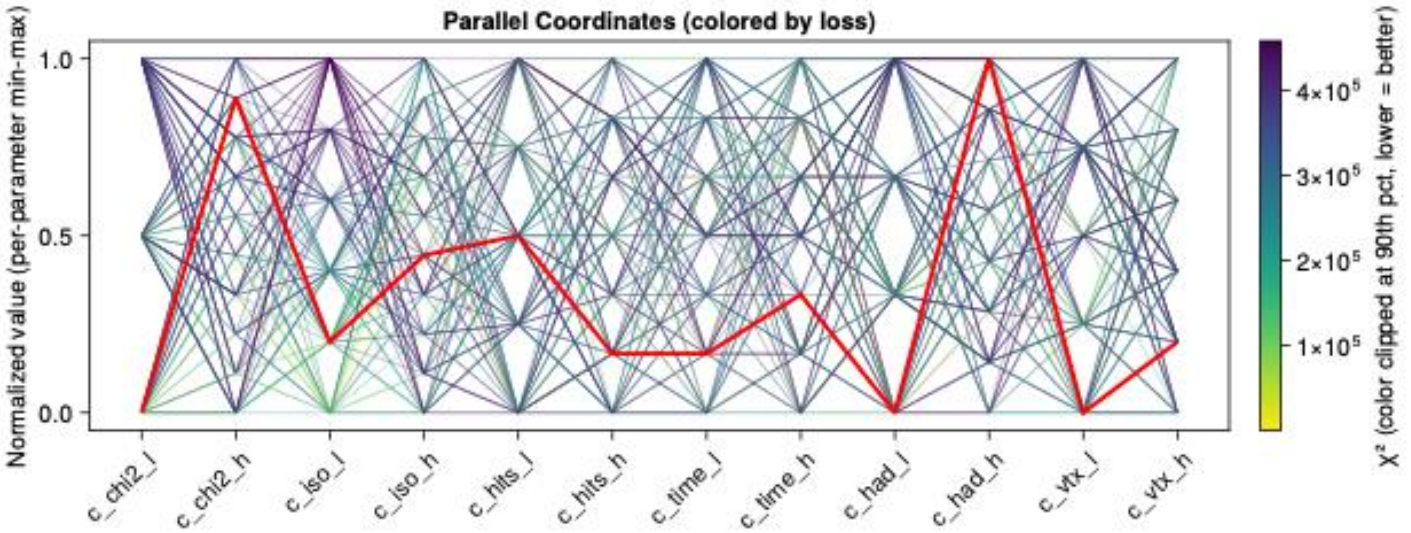
Synthetic MC and data:

- LHC-inspired („fake Z-boson” measurement)
- With intentionally introduced mismatch
- Some fake parameters:
  - Track χ^2
  - Isolation
  - Hits Count
  - Timing [ns]
  - Hadronic Fraction
  - Vertex Distribution [mm]
- Observable: Energy [GeV]
  - Range: 20 – 120 GeV
  - Signal peak: 82 – 98GeV
- Statistics:
  - Data: 600k events
  - MC: 800k events

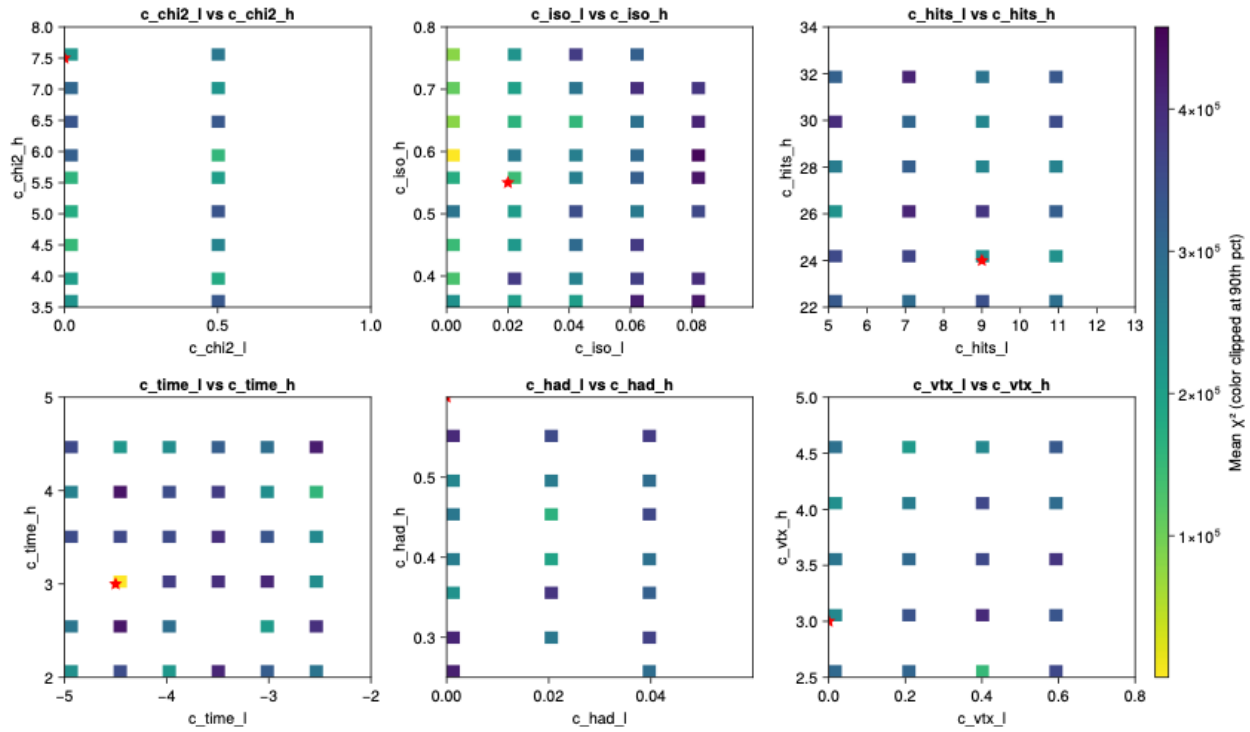


Our optimization problem:

- Objective minimize the loss function
- loss: $\chi^2_{\text{loss}} = \sum \frac{(w_{\text{data}} - w_{\text{MC}})^2}{w_{\text{data}} + w_{\text{MC}}}$
- Tool: HyperOpt.jl
- Method: Random Search





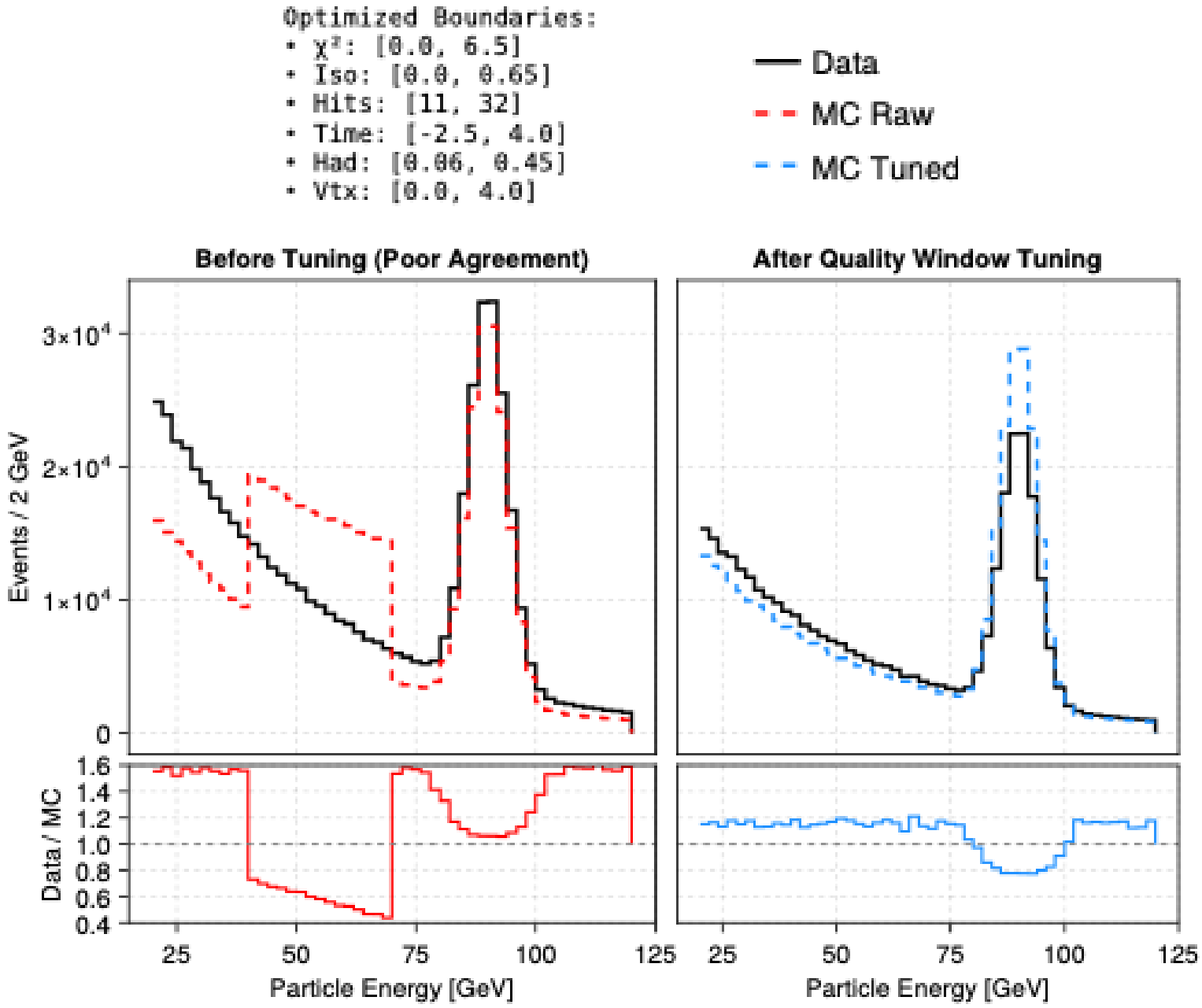


Optimization results

Some comments:

- The cut is ofc on track χ^2 , not χ^2_{loss} ☺
- We allow lower and upper bounds because why not
- Overall improvement
- Would be more interesting to compare against some human-made cuts though …

```
Optimized Boundaries:
* χ²: [0.0, 6.5]
* Iso: [0.0, 0.65]
* Hits: [11, 32]
* Time: [-2.5, 4.0]
* Had: [0.06, 0.45]
* Vtx: [0.0, 4.0]
```



Let's try real data!

Published April 3, 2025 | Version v1

Dataset Open

FAIR Universe - HiggsML Uncertainty Challenge Public Dataset

Bhimji, Wahid[1]; Calafiura, Paolo[1]; Chakkappai, Ragansu[2, 3]; Chang, Po-Wen[4]; Chou, Yuan-Tang[5]; Diefenbacher, Sascha[1]; Dudley, Jordan[1, 6]; Farrell, Steven[1]; Ghosh, Aishik[7, 1]; Guyon, Isabelle[8]; *et al.* Show all 19 authors

Process	Number Generated	LHC Events	Label
Higgs	52 040 227	1015	signal
Z Boson	160 383 358	1 002 395	background
Di-Boson	605 118	3 783	background
ttbar	7 070 398	44 192	background

Preselection Cuts

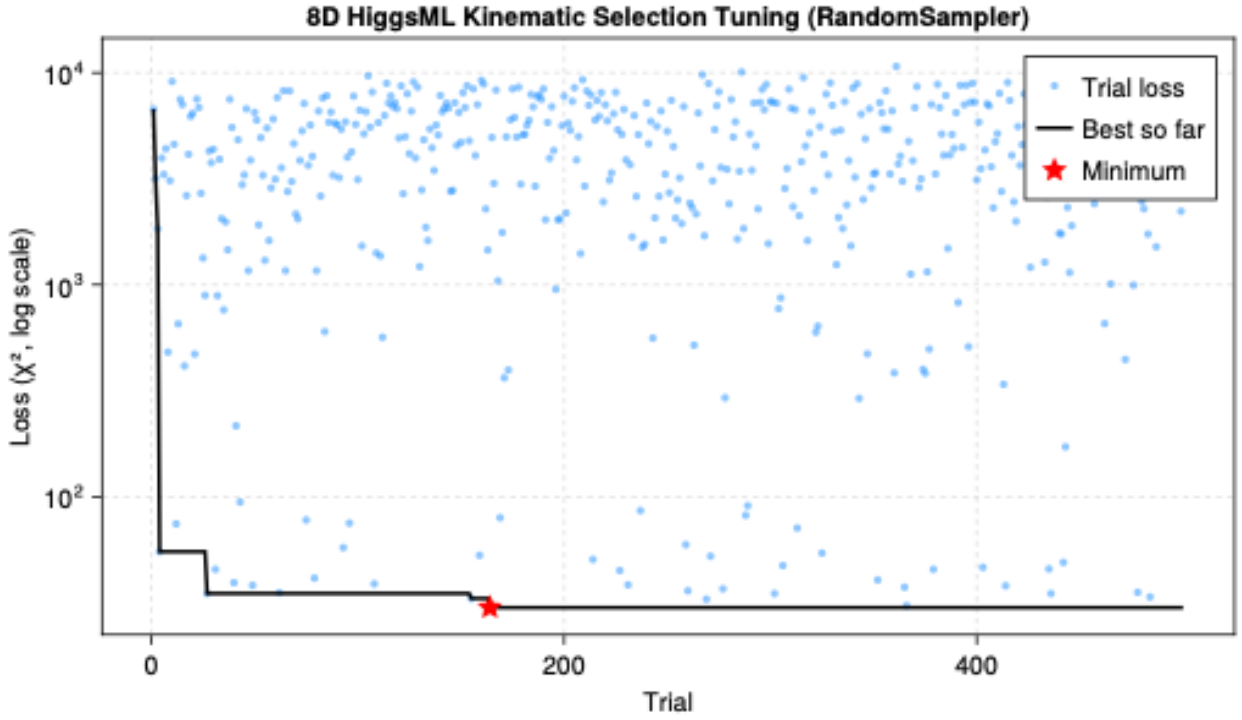
Criteria	Pre-selected cut	Post selection cut
Number of τhad	1	
Number of τlep	1	
pTτhad	> 20GeV	> 26GeV
pTτleppTτlep	> 20GeV	> 20GeV
pTleadingjet	> 20GeV	> 26GeV
pTsubleadingjet	> 20GeV	> 26GeV
Charge	Opposite Charges	

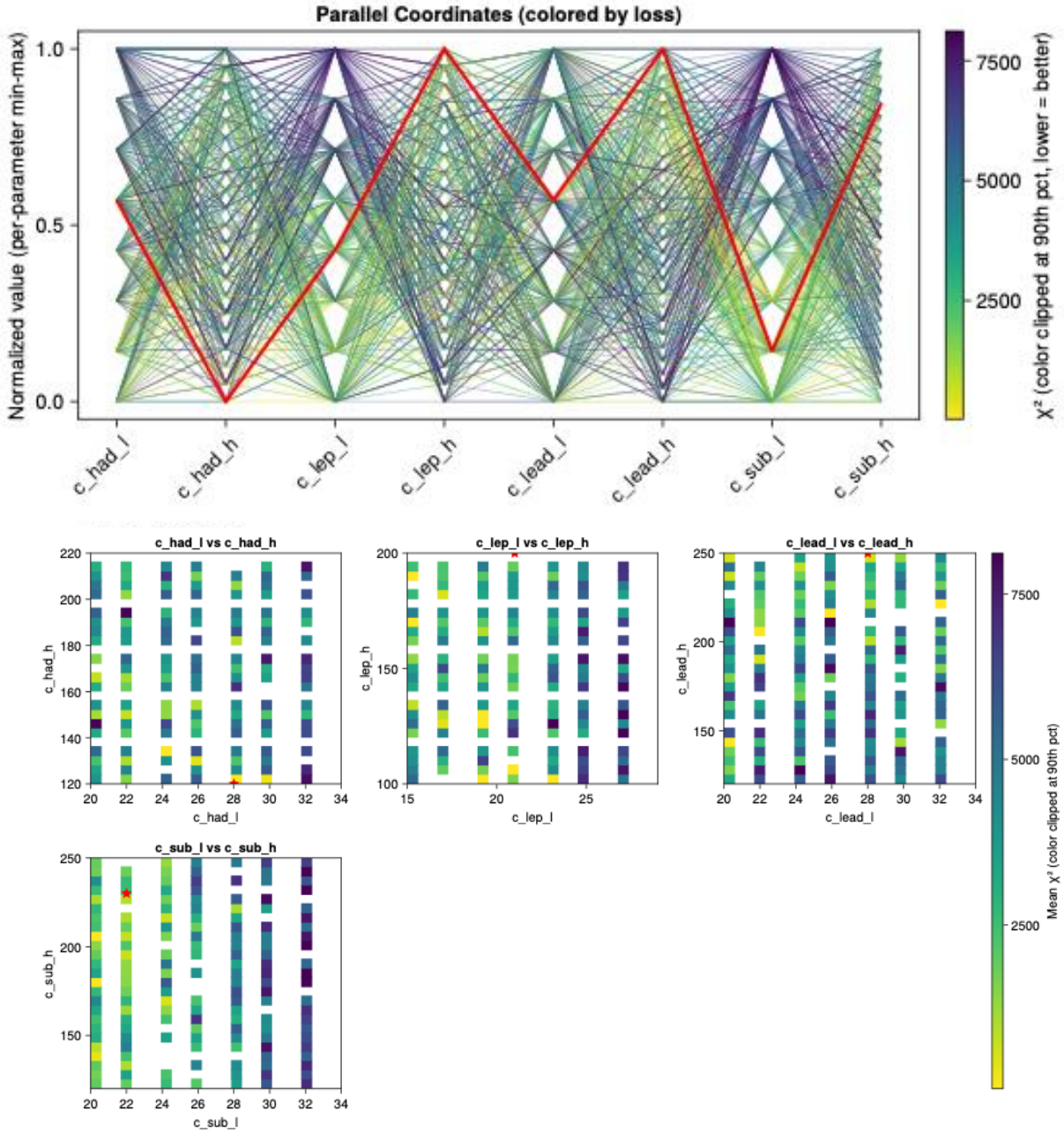
Some comments:

- I use a random subset of 100k events to have something a bit more manageable with modest resources
- The dataset was originally not intended for data/MC tuning but we don't mind
- MC is real, „data" obtained by reweighting
- What is important: it comes with a set of official cuts

Our optimization problem:

- Same as for synthetic data
- Just different parameters





Optimization results

Some comments:

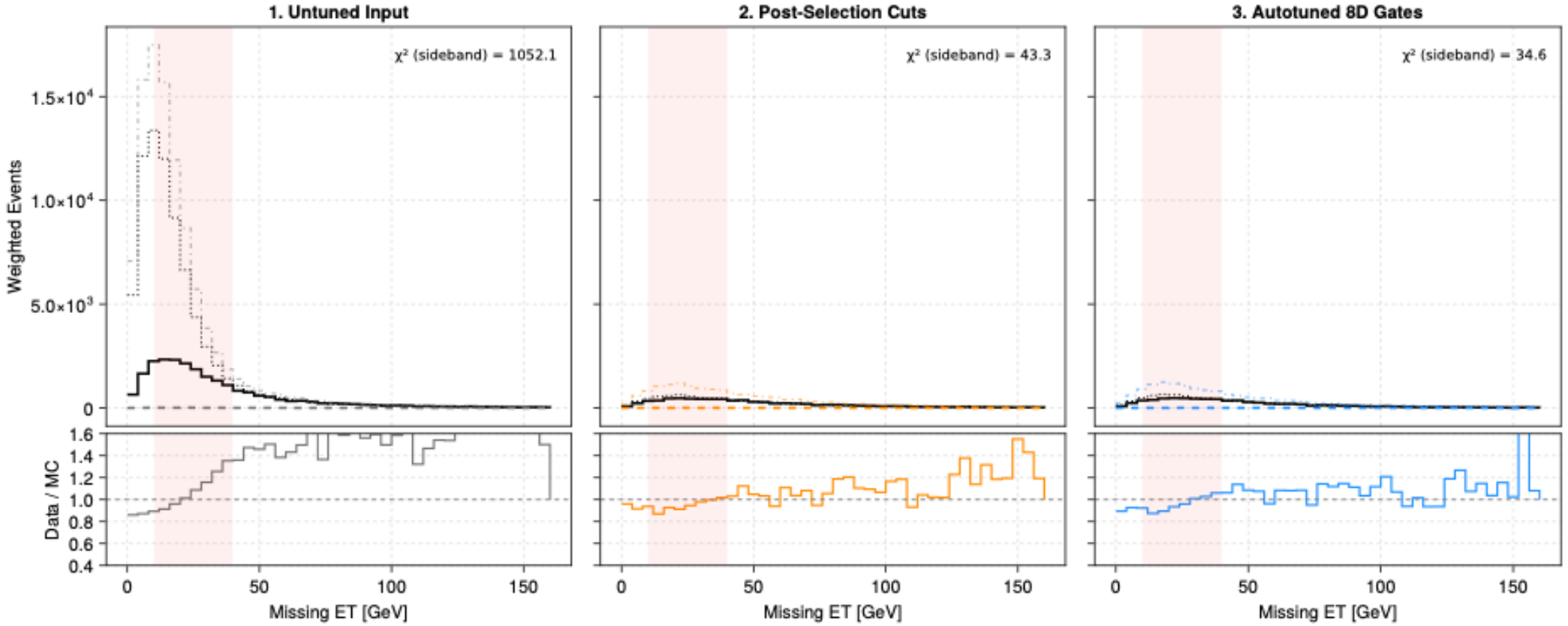
- Red band: excluded from the loss calculation
- Overall improvement
- The result is not really crazy compared to benchmark cuts

```
Benchmark Cuts:
• pT Had      : > 26.0 GeV
• pT Lep      : > 20.0 GeV
• pT Leading Jet: > 26.0 GeV
• pT Sublead Jet: > 26.0 GeV

Tuned Cuts:
• pT Had      : [28.0, 120.0]
• pT Lep      : [21.0, 200.0]
• pT Leading Jet: [28.0, 250.0]
• pT Sublead Jet: [22.0, 230.0]

Shaded band: sideband
reject window
[10.0, 48.0] GeV
```

Data (signal)
Data (background)
MC (signal)
MC (background)

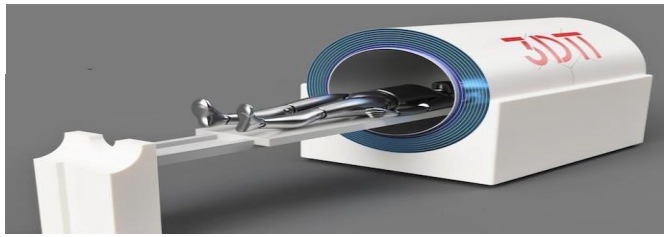


Summary:

- Automation in data/MC tuning can really make life easier and results nicer!
- Succesfull test on both synthetic and real LHC data
- Definitely not making use of the full potential of the machinery yet …
- You can play with it yourself: github.com/pkalaczynski/data_MC_autotune

Outlook:

- Try on other real data from:
  - KMNeT neutrino telescope (data & MC already available)
  - 3Dπ PET scanner developed @ Astrocent (for now only MC)
  - … your experiment(s)? ☺
- Try longer searches, more sophisticated search types, etc …



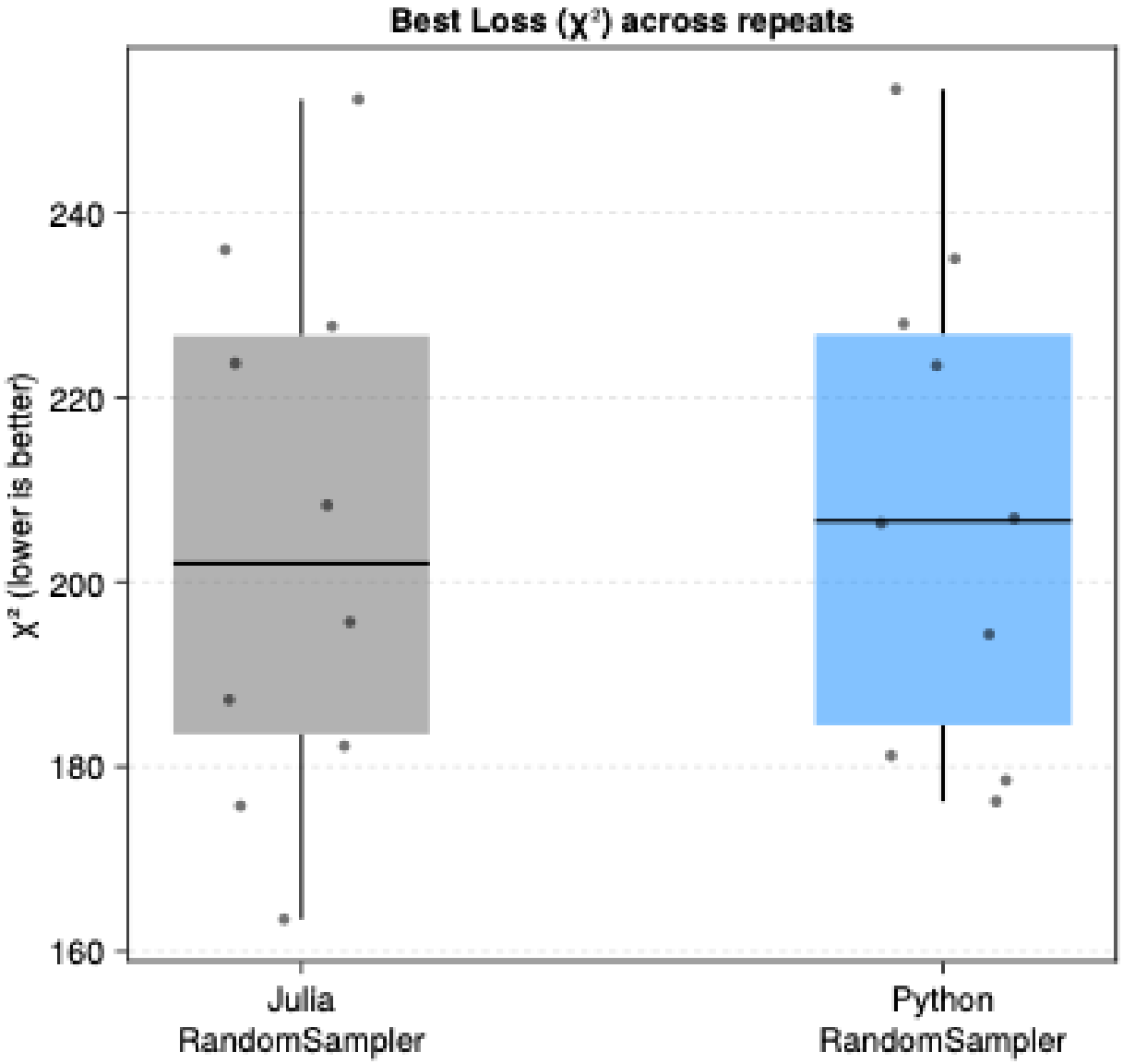
Thank you for your attention!

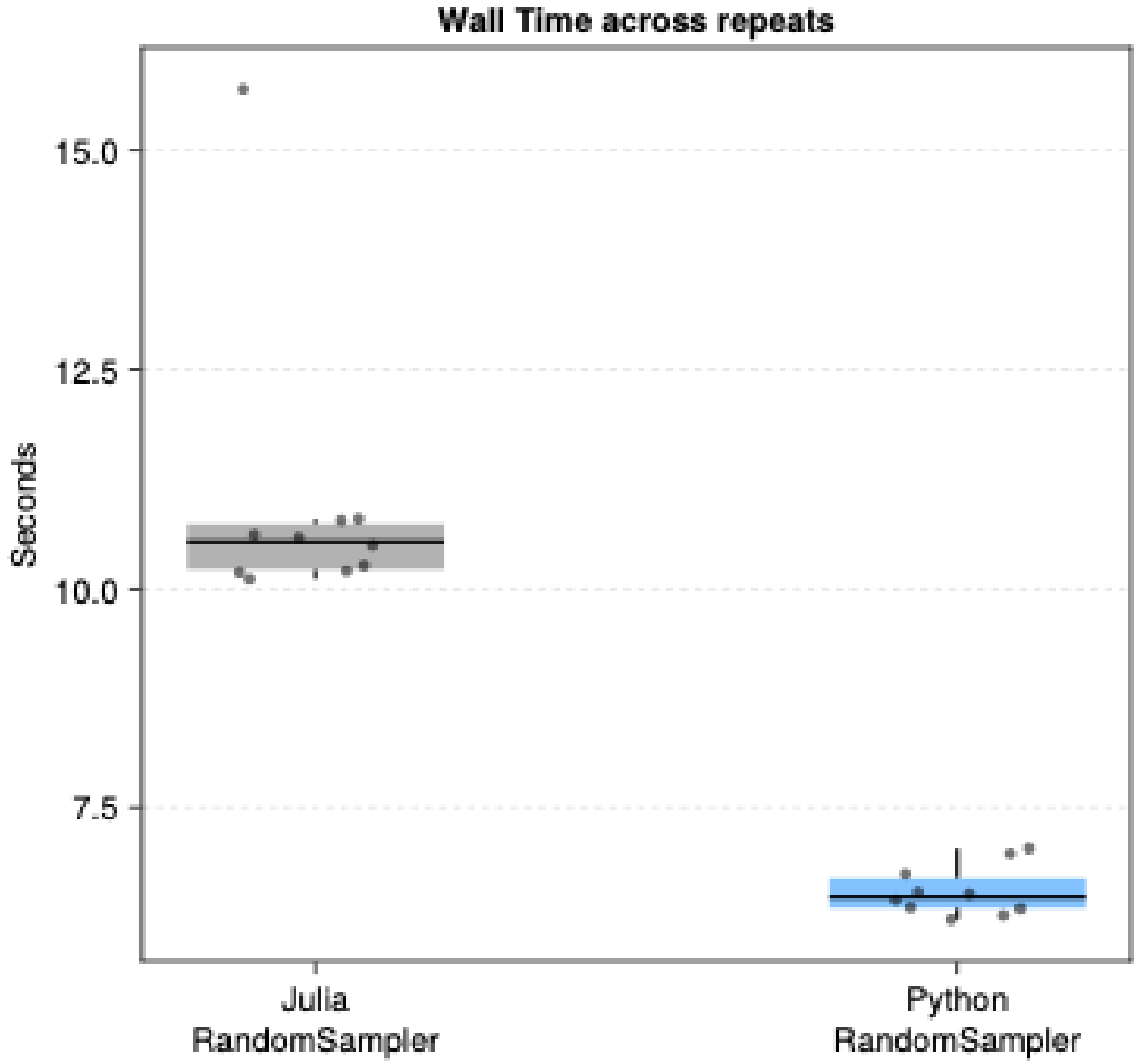
Funded by the European Union | European Funds for Smart Economy | Republic of Poland | Co-funded by the European Union

Backup

Performance comparison: HyperOpt.jl vs Optuna

Dataset: synthetic data & MC





Performance comparison: HyperOpt.jl vs Optuna

Dataset: HiggsML

