Active Learning application in a dark matter search with ATLAS PanDA and iDDS

Rui Zhang, on behalf of the ATLAS Collaboration

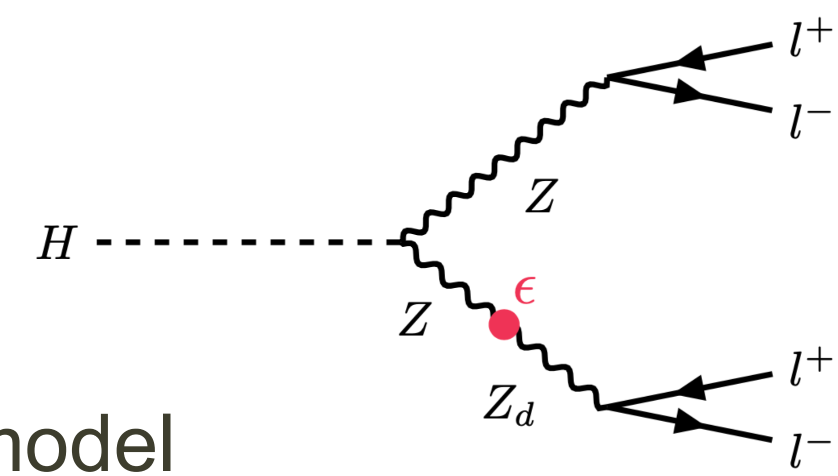
International Conference on New Frontiers in Physics (ICNFP 2023)

University of Wisconsin-Madison, Wisconsin

July 10—23, 2023

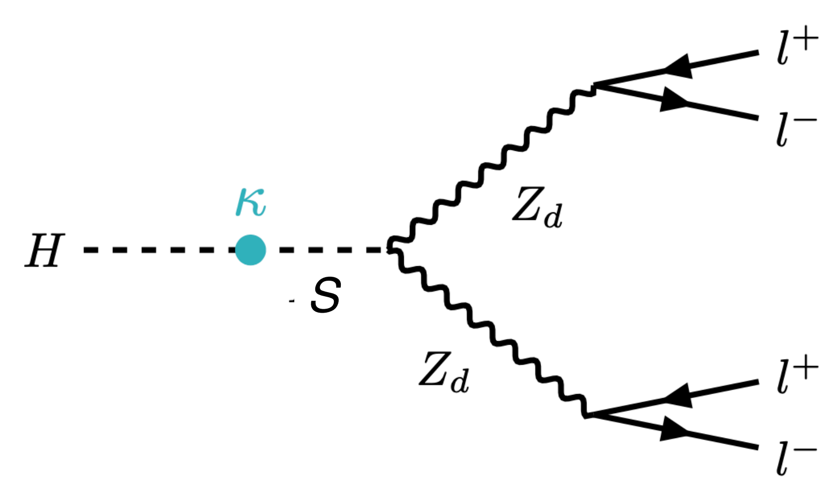
Beyond Standard Model physics parameters

- Hidden Abelian Higgs Model introduces a new Scalar S, new vector boson Z_d



(a) Hypercharge portal

- Five different parameters involved in the model
  - m_S - new scalar mass
  - m_{Z_d} - new vector boson mass
  - Γ_{Z_d} - decay width of the new vector boson
  - ϵ - mixing between Standard Model Z and Z_d
  - κ - mixing between Higgs Boson H and S

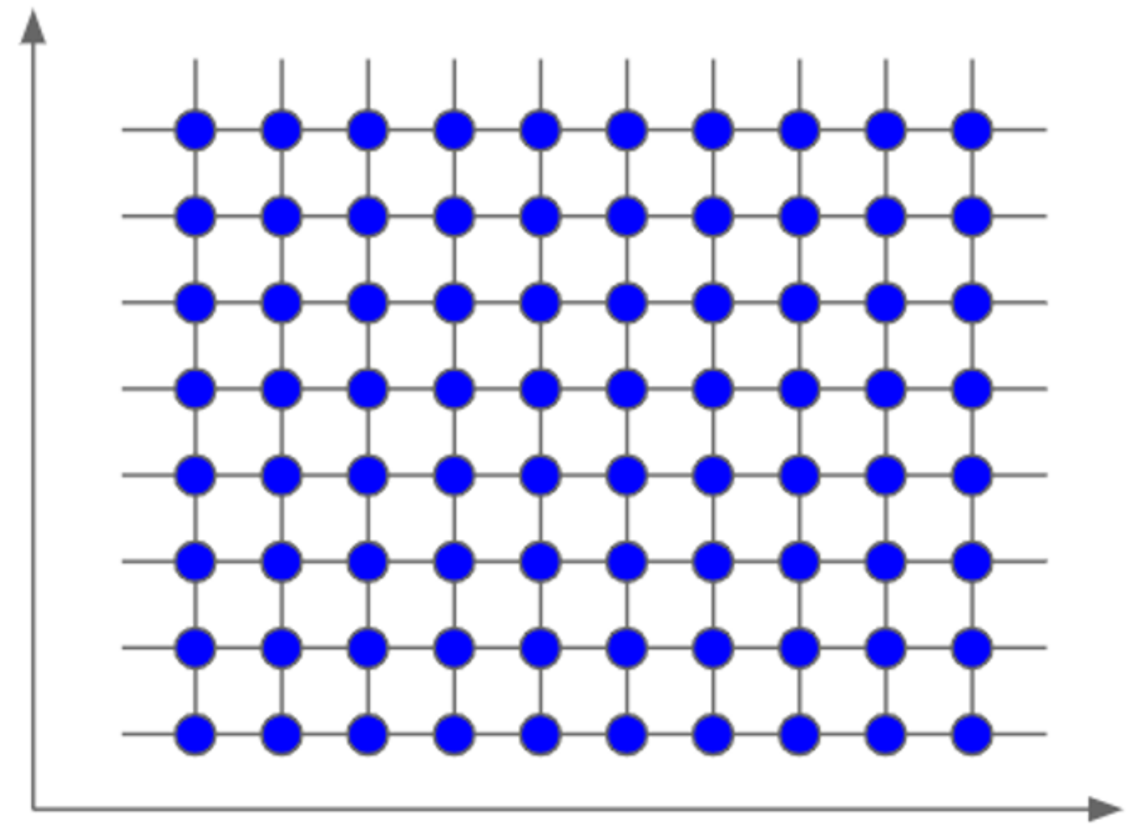


(b) Higgs portal

- ATLAS publication JHEP 03 (2022) 041

Search in high-dimensional parameter space

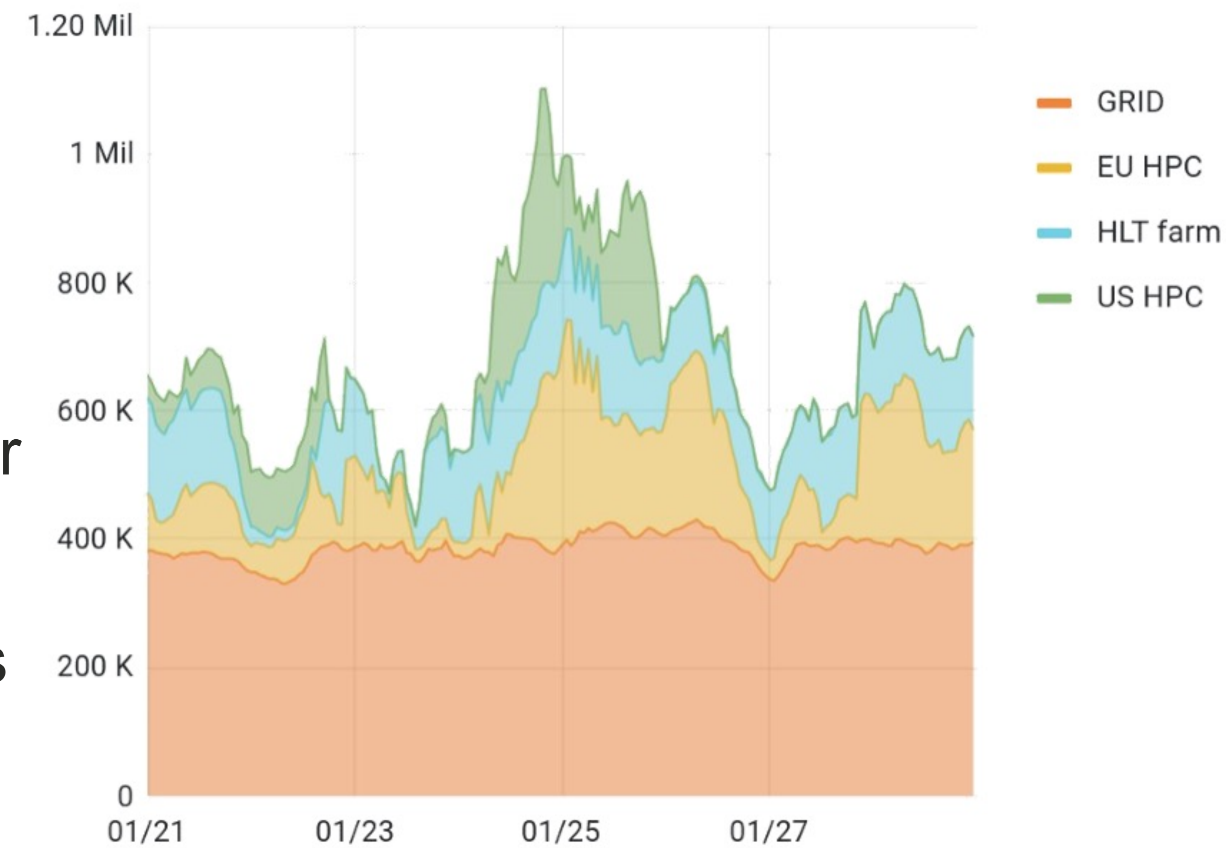
- Process for calculation of limits in phase space conceptional simple, e.g., do a “grid search”
  - Select boundaries in $m_{Z_d}, m_{H_d}, \Gamma_{Z_d}, \epsilon, \kappa$
  - Generate grid of signal samples
  - Calculate cross section limits for each sample
  - Bonus: with theory predictions one can infer contours between excluded and not excluded regions of parameter space



- Problem:
  - Generating Monte Carlo samples is computationally expensive

Monte Carlo Simulation

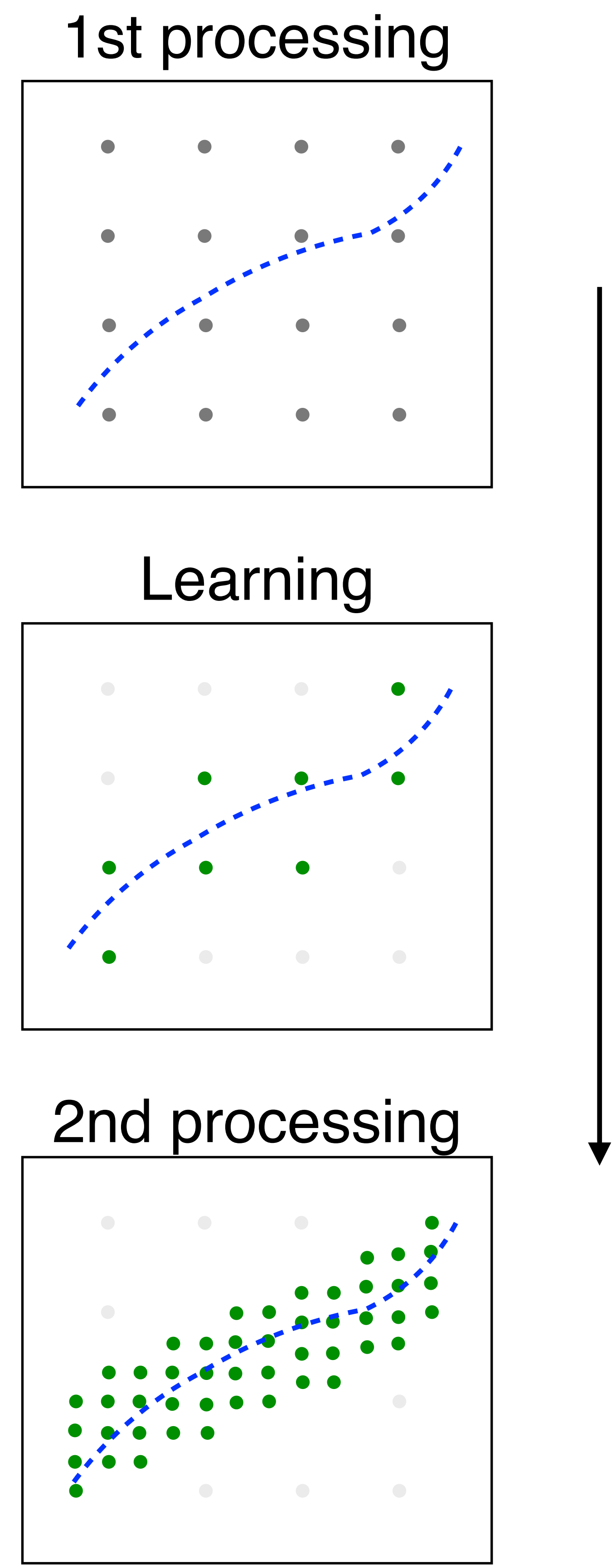
- Monte Carlo (MC) Event simulation is very computationally intensive
  - ~15h calculation time for 100 simulated events on a single CPU core
  - Might need around 10^5 events or more per signal samples
  - Using a grid of 10 values in two dimensions would require 10^7 events, about ~2k CPU months
  - Estimate is for fast calorimeter simulation, full calorimeter simulation requires even more computational resources



CPU resources available to the ATLAS experiment in late January 2023

Active Learning

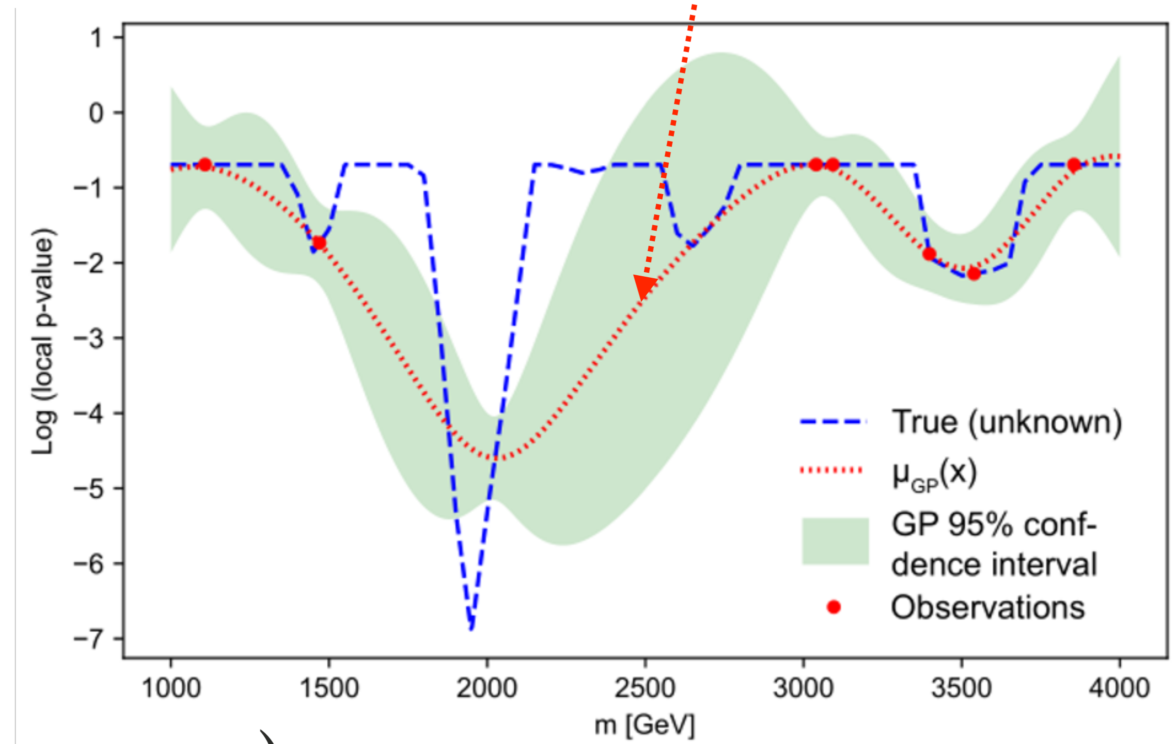
- An iterative process to collect new labelled data for optimisation tasks
- We developed a system to work around the computational constraints by using an Active Learning approach that includes
  - **Surrogate model** that approximates the function mapping physics parameters onto exclusion limits
    - $m_{Z_d}, m_{H_d}, \Gamma_{Z_d}, \epsilon, \kappa \to \sigma(m_{Z_d}, m_{H_d}, \Gamma_{Z_d}, \epsilon, \kappa)$
  - An **acquisition function** that determines which points in the physics parameter space to explore based on the surrogate model



Surrogate Model

- Surrogate Model: Gaussian Process
  - Non-parametric model, yielding probability distribution over possible functions that fit a set of points.
- Assumes output $f(\vec{x})$ is a random variable for each input $\vec{x}$ and $p(f(x_1), f(x_2), \ldots)$ is a multivariate gaussian:
  - $p(f(\vec{x})) = N(f(\vec{x}) \,|\, \vec{\mu}, \mathbf{K})$
  - mean vector $\vec{\mu}$
  - covariance matrix $\mathbf{K}_{ij} = k(x_i, x_j)$
  - kernel $k(x_i, x_j)$, many options, common choice:

Surrogate model (red) to describe truth (blue) within uncertainty

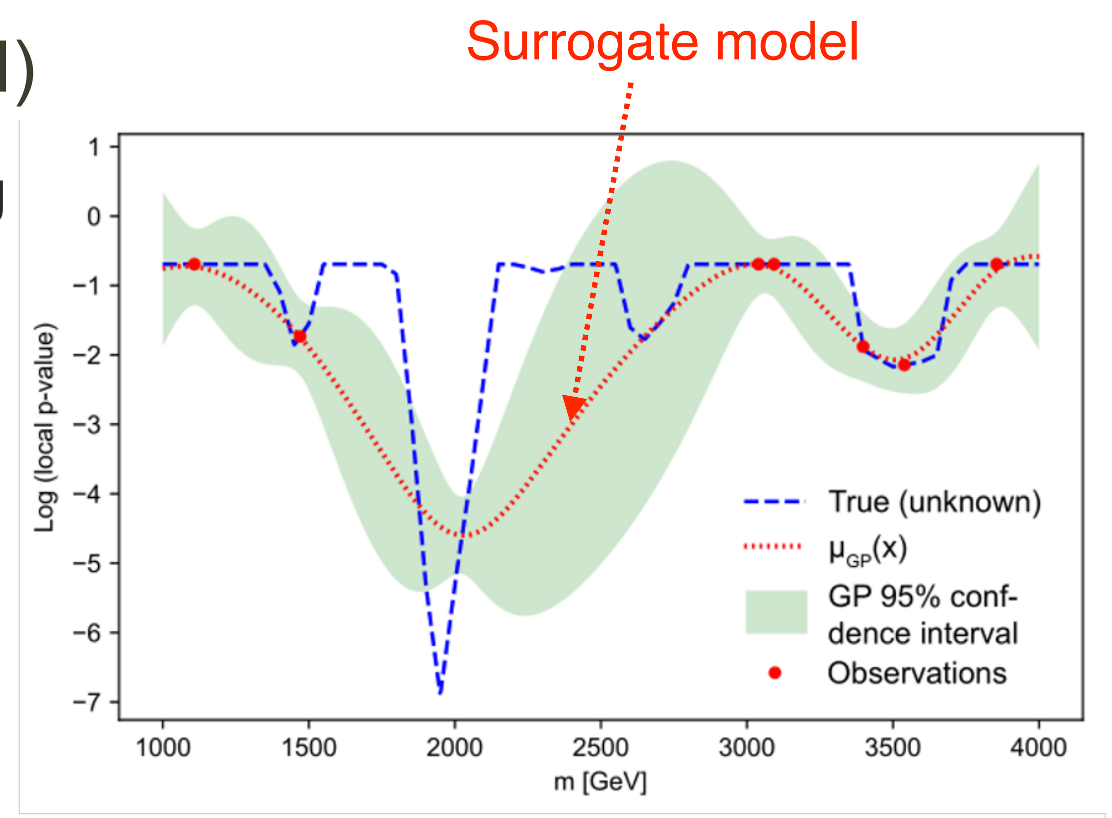


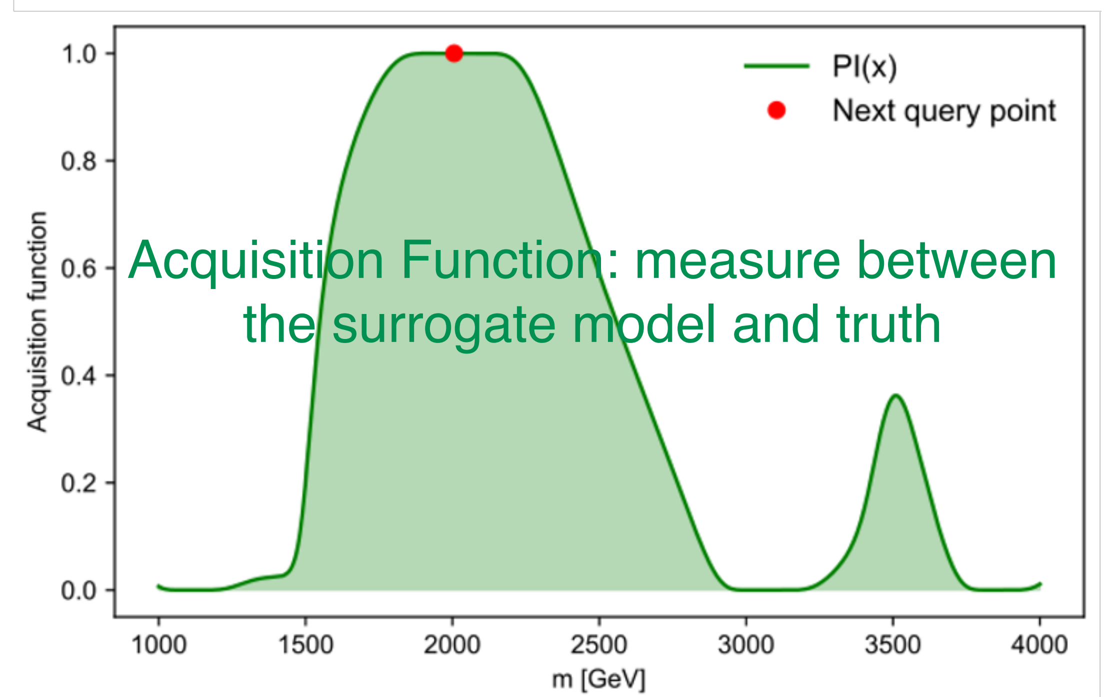
Radial Basis Function kernel $k(x_i, x_j) = \exp\left(-\frac{\left(x_i - x_j\right)^2}{2l^2}\right)$

Acquisition Function

- One example: **Probability of improvement** (PI)
  - Assume we want to find the minimum of an underlying true function, via optimising surrogate model
  - $f(x^*)$ is current minimum, then choose next value to sample for surrogate model, such that PI(x) = Prob$(f(x) - f(x^*) > 0)$ is maximal
  - PI(x) calculable as f(x) is a gaussian distribution with mean μ(x), and standard deviation σ(x)

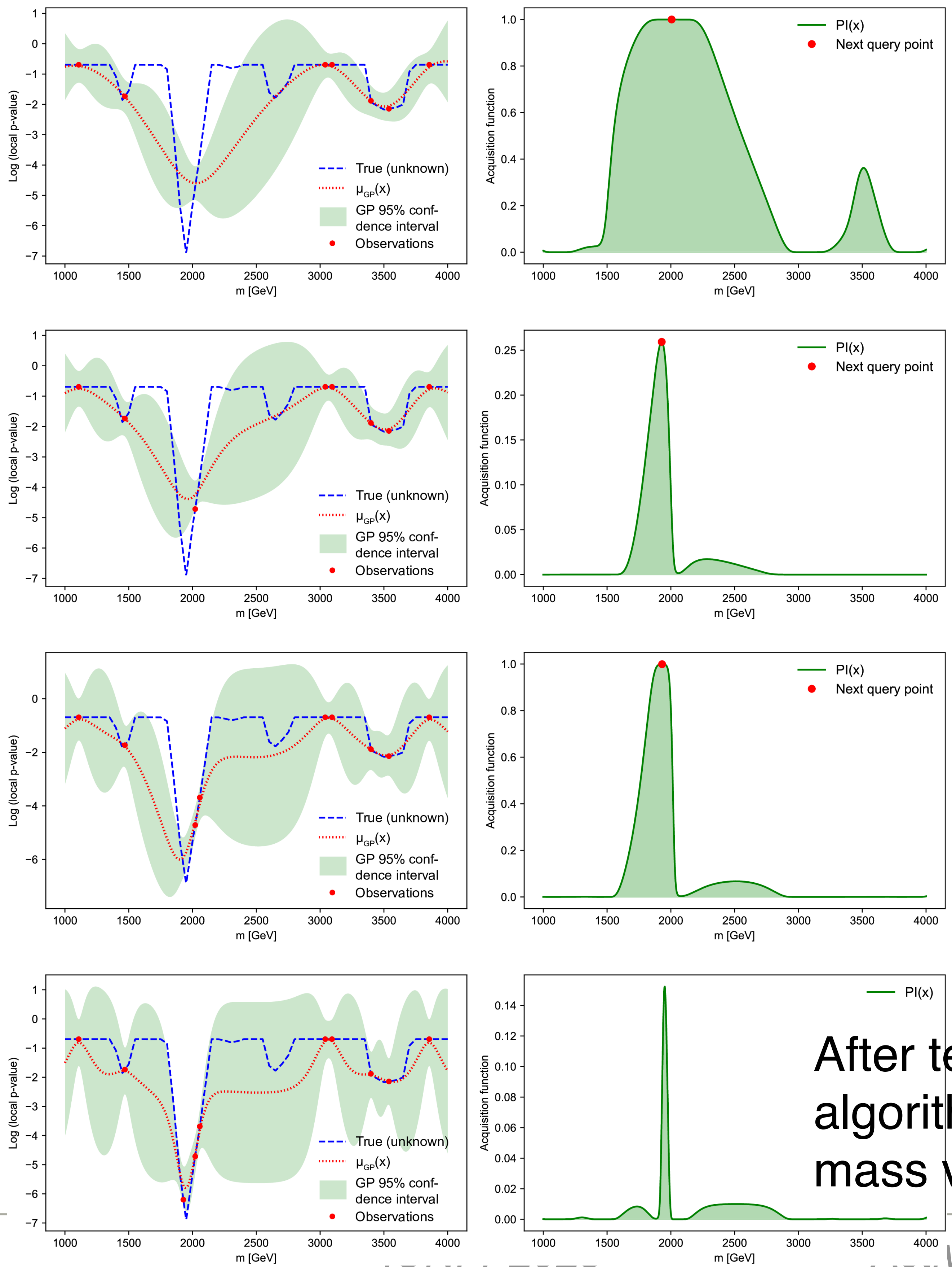
$$\text{PI}(x) = \Phi\left(\frac{f(x^*) - \mu(x)}{\sigma(x)}\right)$$, Φ is Gaussian cumulative distribution function





- Other examples include **upper confidence bound**, **expected improvement**
- See "Active Learning for Excursion Set Estimation" by Kyle Cranmer et al. for an entropy-based acquisition function

Iterative feedback and learning

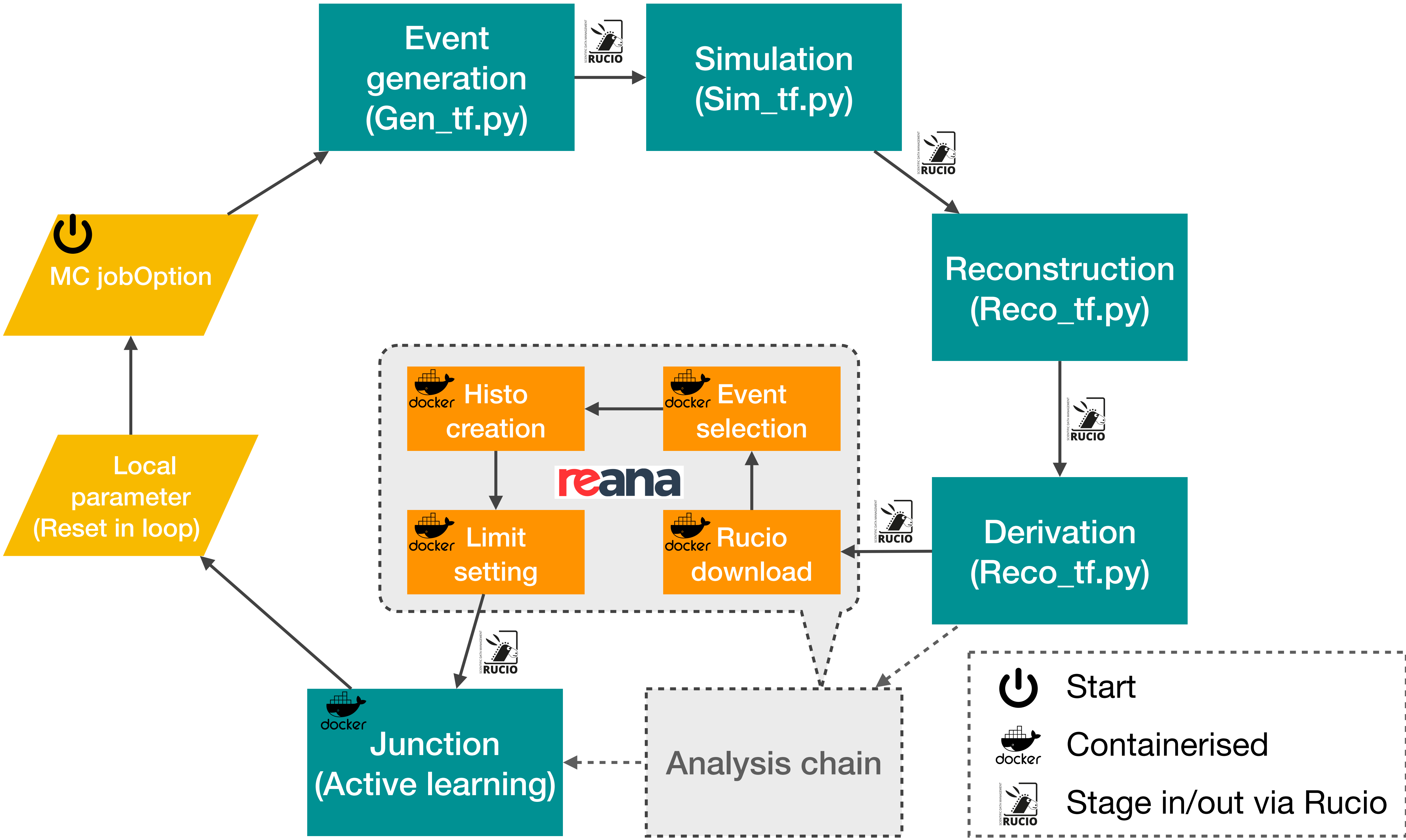


time

After testing around 10 points, the algorithm successfully finds the desired mass value within 2% of the true value.

Implementation

- Used pchain (PanDA) and iDDS
- loops automated no human intervention
- Seamless transition PanDA ⇄ REANA



Results - 1D

- First we have run some 1D results to verify if we can reproduce the results in the publication

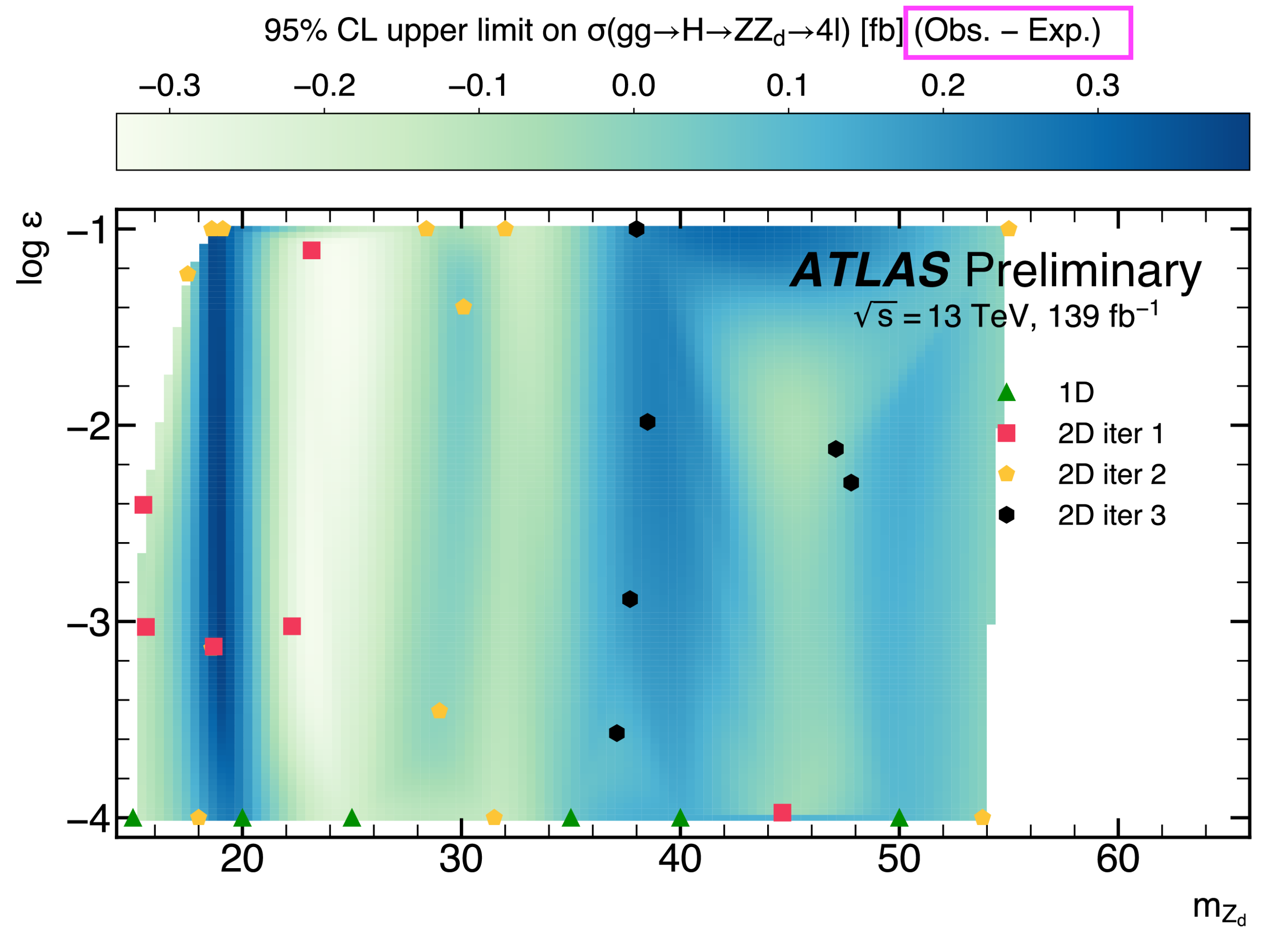
m_{Z_d}	HepData		This study		Ratio of HepData to this study	
[GeV]	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
15	0.34	0.48	0.32	0.48	1.04	1.00
20	0.74	0.53	0.73	0.51	1.03	1.04
25	0.37	0.54	0.37	0.55	0.99	0.99
30	0.56	0.51	0.57	0.52	0.99	0.98
40	0.39	0.25	0.39	0.26	1.00	0.97
50	0.30	0.21	0.30	0.20	1.01	1.03

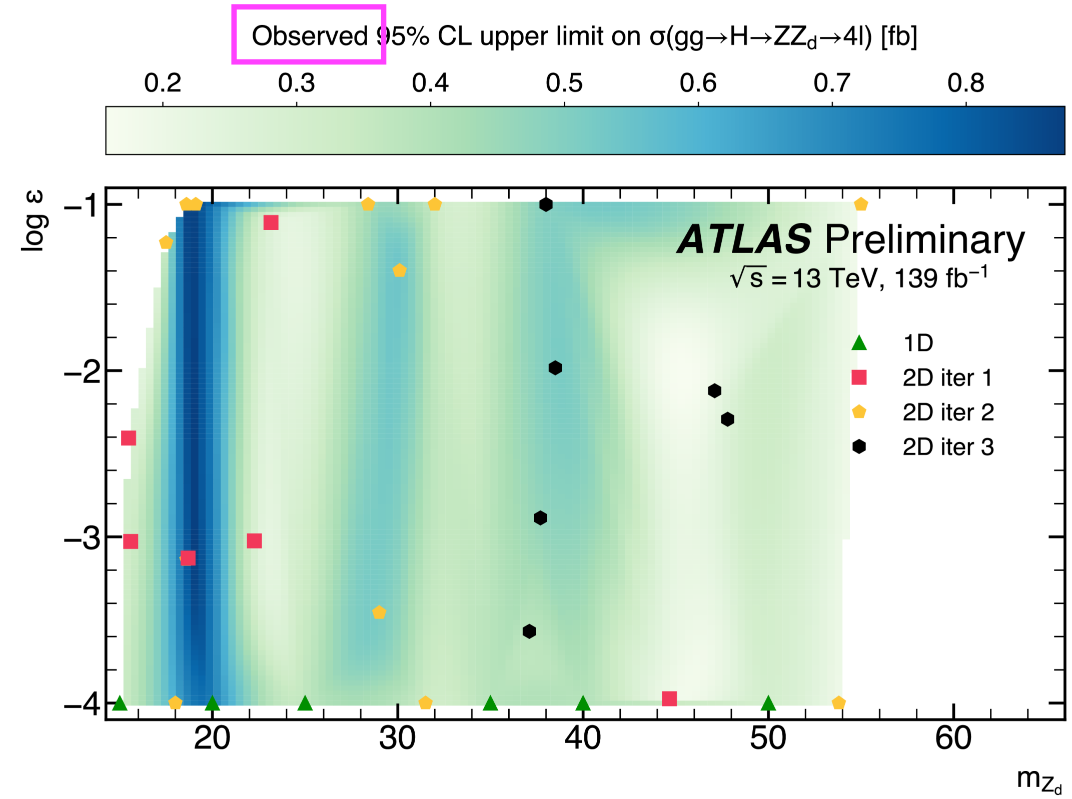
  - Agreement is within <4%

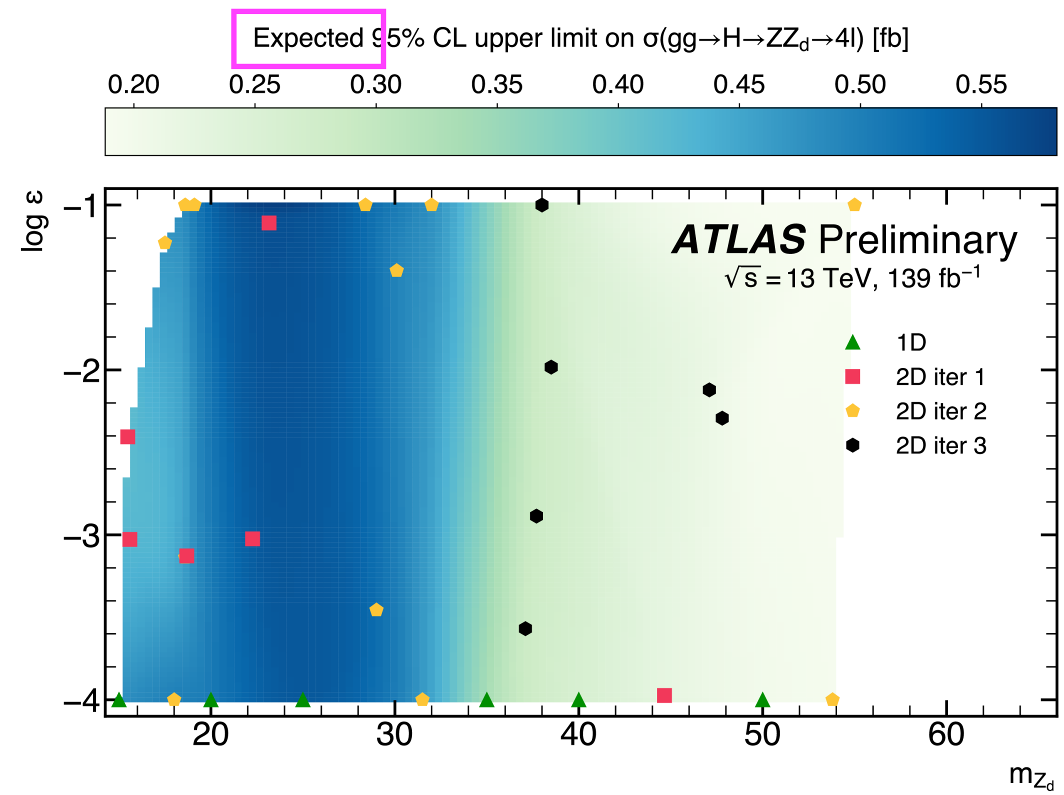
- Major differences:
  - Limits obtained with pseudo-data (publication) vs asymptotic Formulae (this study)
  - Different ATLAS software versions (signal MC are simulated in different versions)
  - Different NP pruning scheme

Results - 2D

- Parameter phase space bounded by
  - $m_{Z_d} = [15, 55], \epsilon = [0.0001, 0.1]$
- Goal is to looking for max(Obs-Exp) to find excess
- Run in 4 iterations with 30 points total







Summary

- Demonstrated active learning driven re-analysis for published dark sector analysis
  - Extended dark Z search channel from 1D to 2D parameter space in m_{Z_d} and ϵ
  - PubNote ATL-PHYS-PUB-2023-010
- Establishing a 2nd demonstrator based on heavy Higgs Boson search
  - 3D parameter space
  - Theory predictions to prioritise exclusion contours
- Many other applications and possibilities for this tool and workflow
- Another application of active learning using a different tool:
  - Active Learning reinterpretation of an ATLAS Dark Matter search constraining a model of a dark Higgs boson decaying to two b-quarks ATL-PHYS-PUB-2022-045

Thank you for your attention!