# Efficient Search for New Physics using Active Learning in the ATLAS Experiment with RECAST

*LHC Reinterpretation Forum 2022*

**Irina Espejo**
**12nd December 2022**

# The team

Kyle Cranmer
(UW Madison,
former NYU)

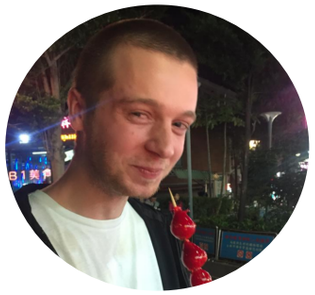
Lukas Heinrich
(TMU, former NYU)

Patrick Rieck
(NYU)

Zubair Bhatti
(NYU)

Phillip Gadow
(DESY)

Eleni Skorda
(Lund University)



Janik von Ahnen
(DESY)

Ruth Pöttgen
(Lund University)

# 4 years ago, same venue

## How it started

**Levelset Estimation by Bayesian Optimization**

K Cranmer, **L Heinrich**, G Louppe
4th Reinterpretation Workshop

May 15th 2018

NEW YORK UNIVERSITY /lukasheinrich

**Searching for BSM:**

fundamentally about finding the boundary between models that are consistent with the data and those that are not.
1D: intervals
2D: contours
ND: (hyper-)surfaces

Boundary is usually defined by iso-surfaces of a test statistic (e.g. CLs) at certain values.

**Problem:**

assessing models is computationally expensive.

**This Talk:**

how to find excursion sets / iso-surfaces of generic R^n functions in an efficient way.

## How it's going

ATL-PHYS-045-2022

ATLAS EXPERIMENT

**ATLAS PUB Note**
ATL-PHYS-PUB-2022-045
November 3, 2022

CERN

**Active Learning reinterpretation of an ATLAS Dark Matter search constraining a model of a dark Higgs boson decaying to two *b*-quarks**

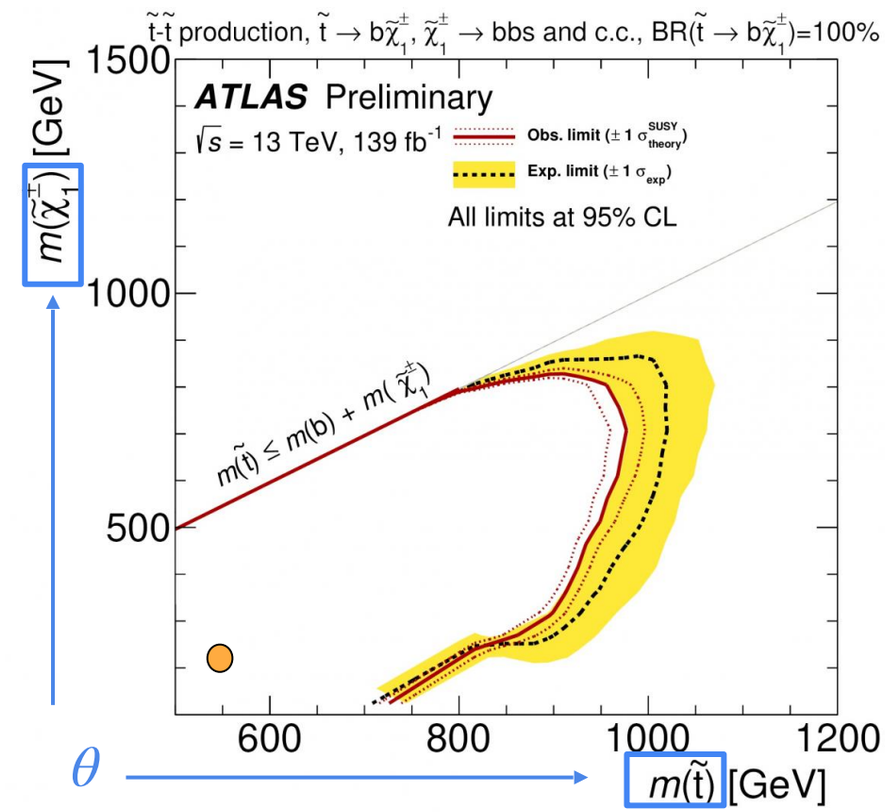
The ATLAS Collaboration

# Overview

Congrats! You have completed your analysis and after hard work, it's time to produce an **Exclusion Contour**

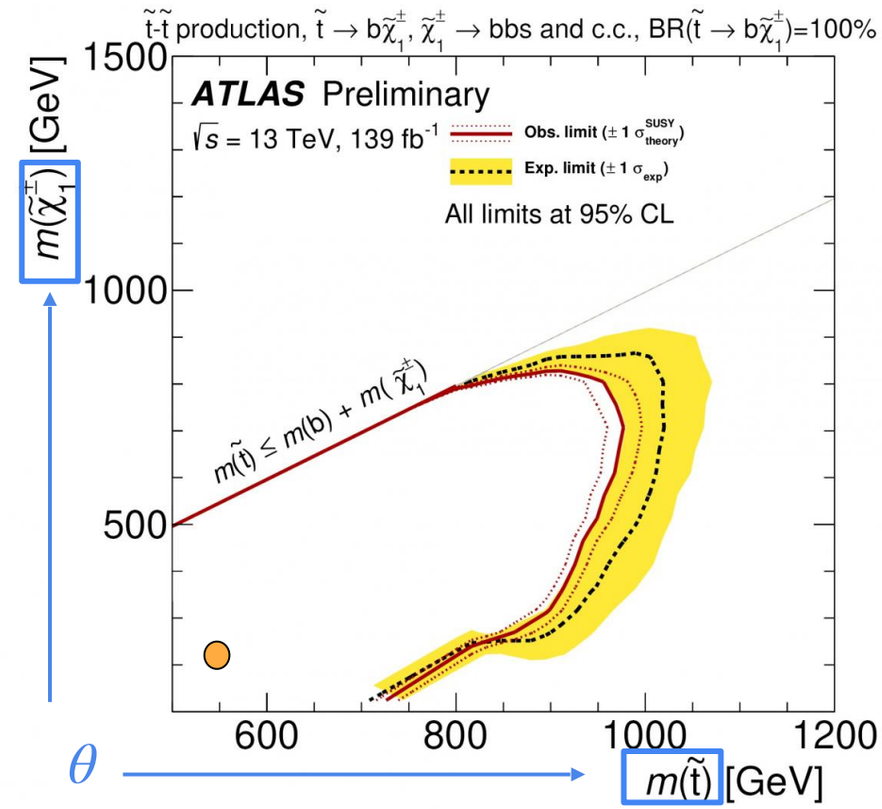
- *Reduction to 1D-2D* **because evaluating official MC Samples is** computationally expensive

- New Physics models are multidimensional, how to set limits over rest of parameters?



# Overview

Congrats! You have completed your analysis and after hard work, it's time to produce an **Exclusion Contour**

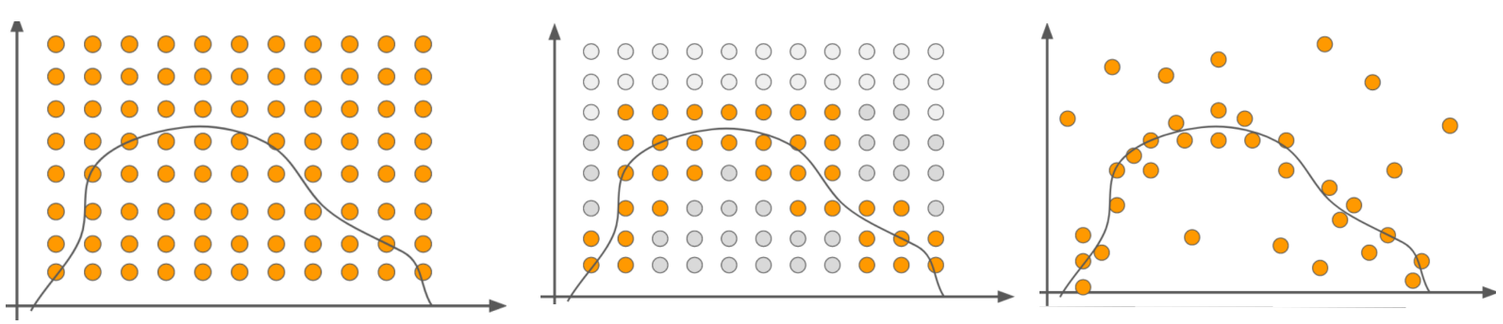



- *Reduction to 1D-2D* because evaluating official MC Samples is computationally expensive

- New Physics models are multidimensional, how to set limits over rest of parameters?

***Traditional grid approach scales exponentially with dimension, need a more efficient approach***

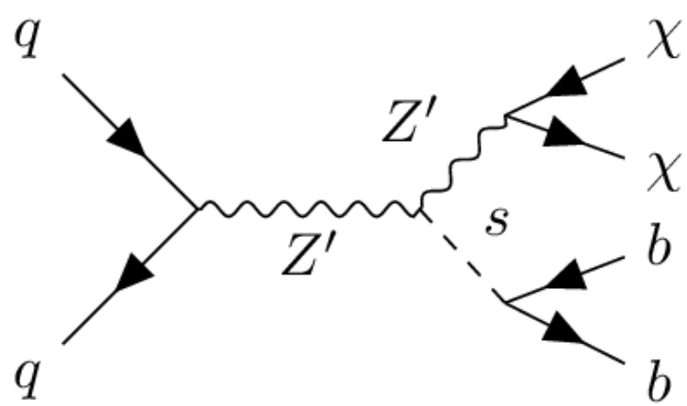
L. Heinrich et al. ACAT 2019



Current ATLAS approach | Better queries | **Active Learning method**

# ATLAS Mono-H(bb) search

- Reinterpretation of the ATLAS **search for Dark Matter in association with a Higgs boson decaying into b-quarks**




**ATLAS CONF Note**

ATLAS-CONF-2018-039

25th July 2018

**Search for Dark Matter Produced in Association with a Higgs Boson decaying to $b\bar{b}$ at $\sqrt{s}$ = 13 TeV with the ATLAS Detector using 79.8 fb$^{-1}$ of proton-proton collision data**

The ATLAS Collaboration

**ATLAS PUB Note**

ATL-PHYS-PUB-2019-032

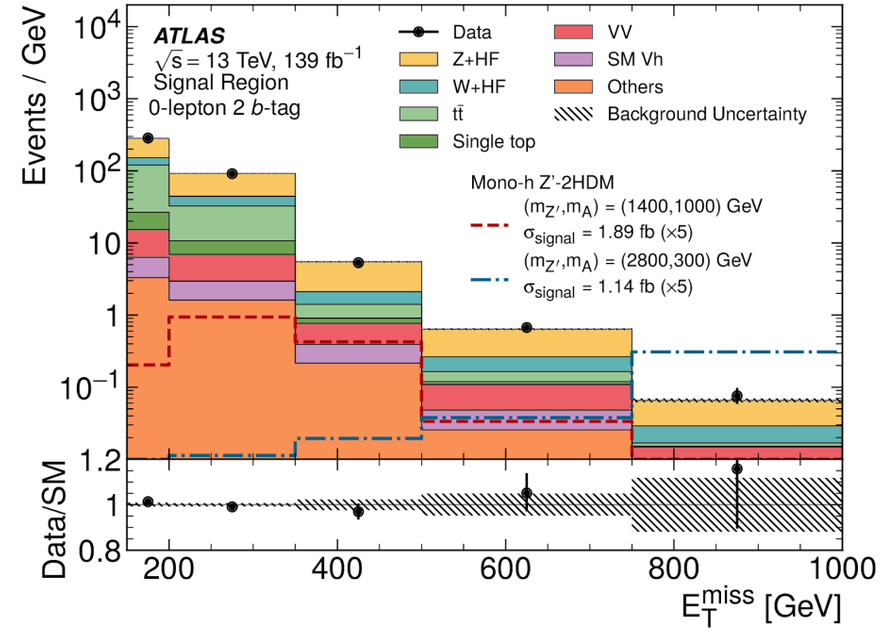
11th August 2019

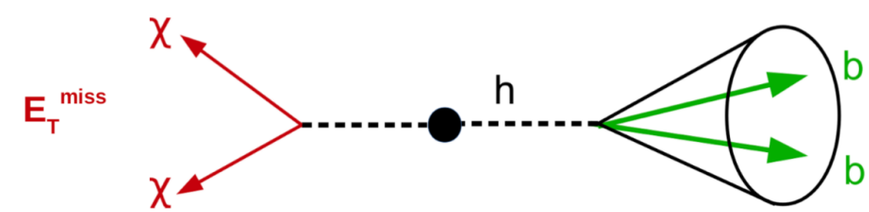
**RECAST framework reinterpretation of an ATLAS Dark Matter Search constraining a model of a dark Higgs boson decaying to two *b*-quarks**

The ATLAS Collaboration

# ATLAS Mono-H(bb) search

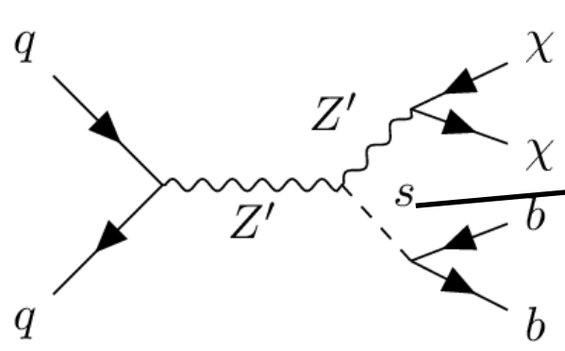
- Reinterpretation of the ATLAS **search for Dark Matter in association with a Higgs boson decaying into b-quarks**
  - Full Run 2 dataset
  - Numerous signal and control regions
  - *Sophisticated reconstruction techniques*: large-Radius jets with b-tagged subjects as Higgs boson decay candidates



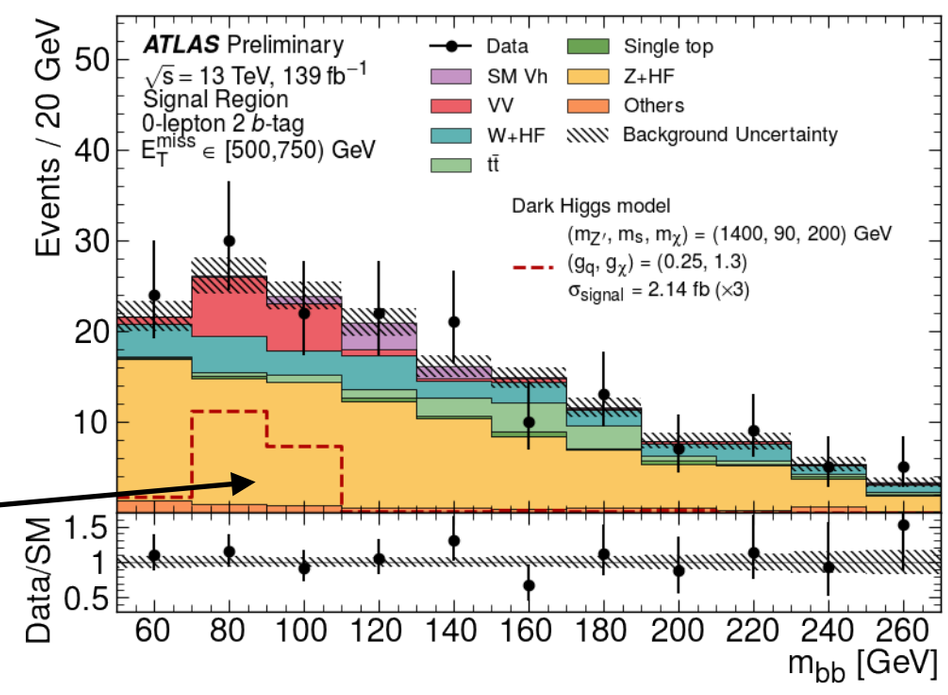




Non-trivial new physics search

4 parameters $\boldsymbol{\theta} = (m_{Z'}, m_s, m_\chi, g_\chi)$




Dark Higgs boson resonance



# Method

Cranmer et al. 2010

Heinrich et al. 2018

- Abstraction of full analysis pipeline, evaluated thanks to recast

$$f : \mathbb{R}^4 \to \mathbb{R}^4$$

$$\boldsymbol{\theta} \to \boldsymbol{y} := (\log \mu^{UL}_{exp}(\boldsymbol{\theta}),\ \log \mu^{UL}_{+1\sigma}(\boldsymbol{\theta}),\ \log \mu^{UL}_{-1\sigma}(\boldsymbol{\theta}),\ \log \mu^{UL}_{obs}(\boldsymbol{\theta}))$$

Contours of interest $\boldsymbol{y}(\boldsymbol{\theta}) = 0$ corresponds to $\mu^{UL}_{exp}(\boldsymbol{\theta}) = 1$ meaning $\sigma(\boldsymbol{\theta}) \approx \sigma_{BSM}(\boldsymbol{\theta})$

- Find the 4 contours with minimum num. queries to $f$ $\longrightarrow$ **Active Learning**

# Method

Cranmer et al. 2010

ATL-PHYS-PUB-2019-932

- Abstraction of full analysis pipeline, evaluated thanks to recast

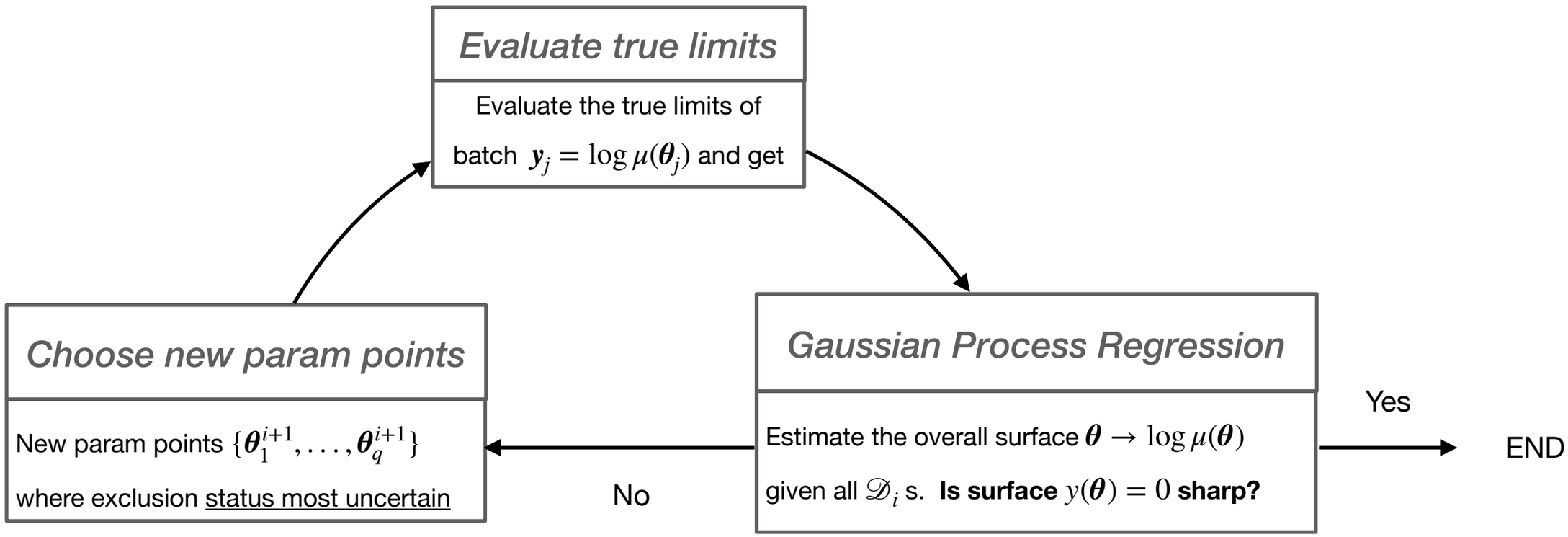
$$f : \mathbb{R}^d \to \mathbb{R}^4$$

$$\boldsymbol{\theta} \to \boldsymbol{y} := (\log \mu^{UL}_{exp}(\boldsymbol{\theta}),\ \log \mu^{UL}_{+1\sigma}(\boldsymbol{\theta}),\ \log \mu^{UL}_{-1\sigma}(\boldsymbol{\theta}),\ \log \mu^{UL}_{obs}(\boldsymbol{\theta}))$$

Contours of interest $\boldsymbol{y}(\boldsymbol{\theta}) = 0$ corresponds to $\mu^{UL}_{exp}(\boldsymbol{\theta}) = 1$ meaning $\sigma(\boldsymbol{\theta}) \approx \sigma_{BSM}(\boldsymbol{\theta})$

- Find the 4 contours with minimum num. queries to $f$ $\longrightarrow$ **Active Learning**

**For** each new batch of BSM parameter points $\{\boldsymbol{\theta}^i_1, \ldots, \boldsymbol{\theta}^i_q\}$ **do**:



# Method

Cranmer et al. 2010

ATL-PHYS-PUB-2019-932

- Abstraction of full analysis pipeline, evaluated thanks to recast

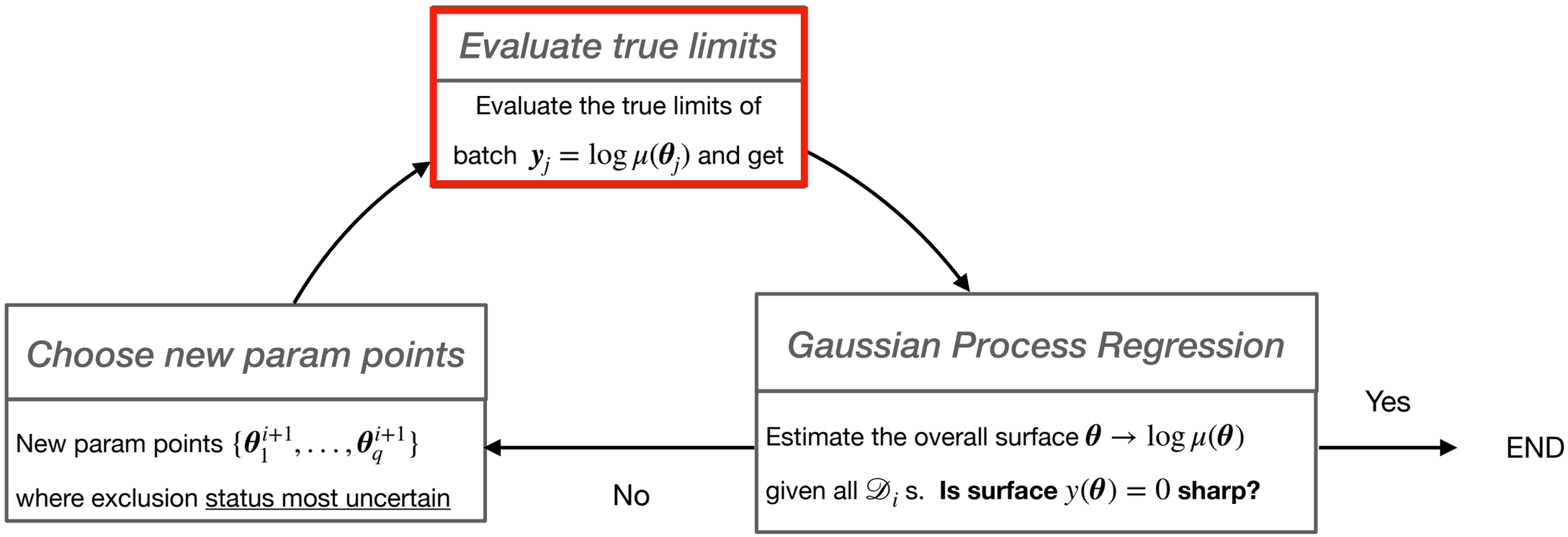
$$f : \mathbb{R}^d \to \mathbb{R}^4$$

$$\boldsymbol{\theta} \to \boldsymbol{y} := (\log \mu^{UL}_{exp}(\boldsymbol{\theta}),\ \log \mu^{UL}_{+1\sigma}(\boldsymbol{\theta}),\ \log \mu^{UL}_{-1\sigma}(\boldsymbol{\theta}),\ \log \mu^{UL}_{obs}(\boldsymbol{\theta}))$$

Contours of interest $\boldsymbol{y}(\boldsymbol{\theta}) = 0$ corresponds to $\mu^{UL}_{exp}(\boldsymbol{\theta}) = 1$ meaning $\sigma(\boldsymbol{\theta}) \approx \sigma_{BSM}(\boldsymbol{\theta})$

- Find the 4 contours with minimum num. queries to $f$ $\longrightarrow$ **Active Learning**

**For** each new batch of BSM parameter points $\{\boldsymbol{\theta}^i_1, \ldots, \boldsymbol{\theta}^i_q\}$ **do**:



# Evaluate param points

## *on MC Production System*

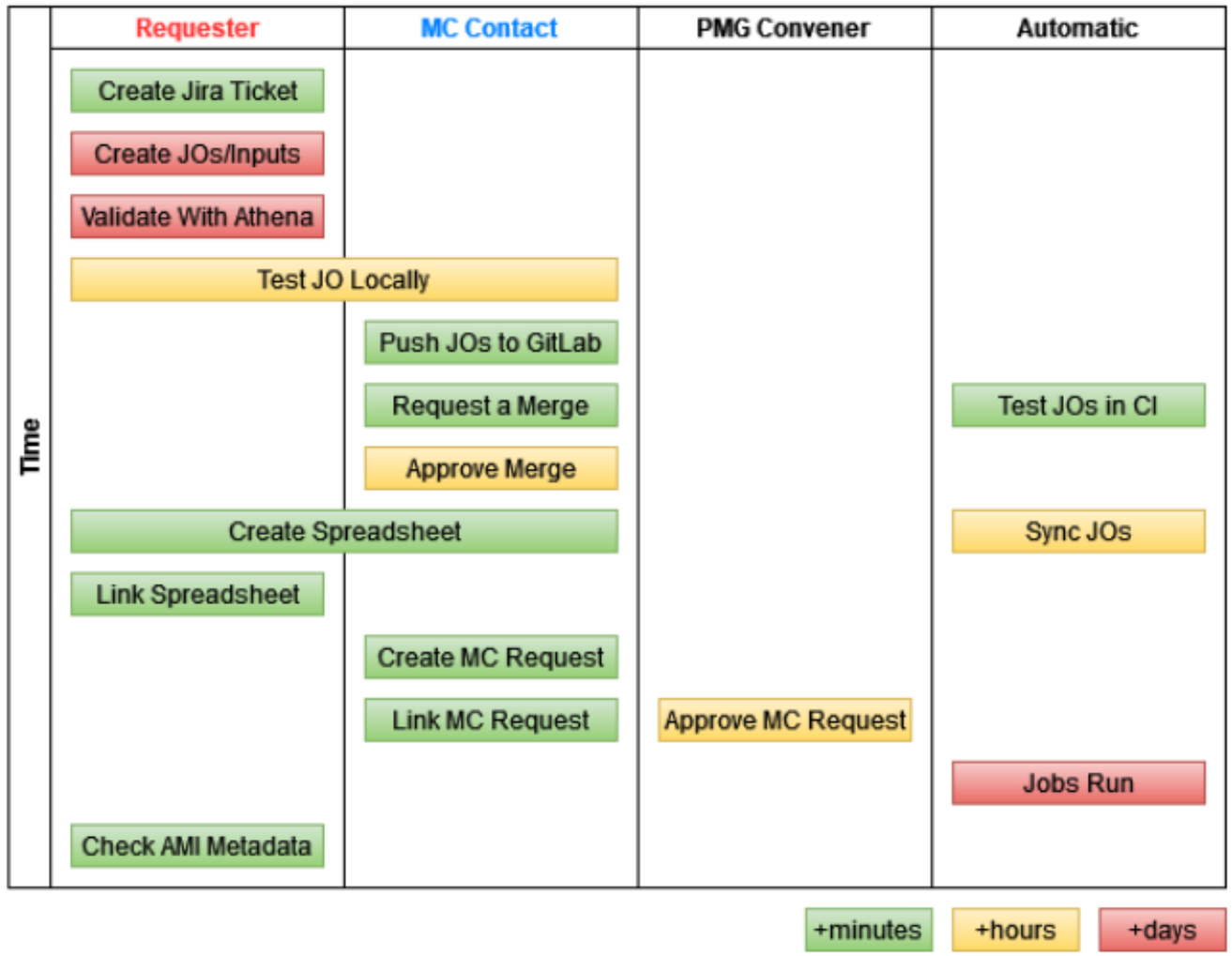
Latency is ~ 10 days.

Latency improving since the start of project thanks to

Zubair Bhatti (NYU).

<u>Ideally fully automatized</u> but:

- Approval needs different people at ATLAS
- Pipeline has many steps with legacy software/ protocols
- Involves active monitoring of jobs.
- Methodology needs to be crystal clear and as simple as possible for approval




Source: Zubair Bhatti

# Method

Cranmer et al. 2010

ATL-PHYS-PUB-2019-932

- Abstraction of full analysis pipeline, evaluated thanks to recast

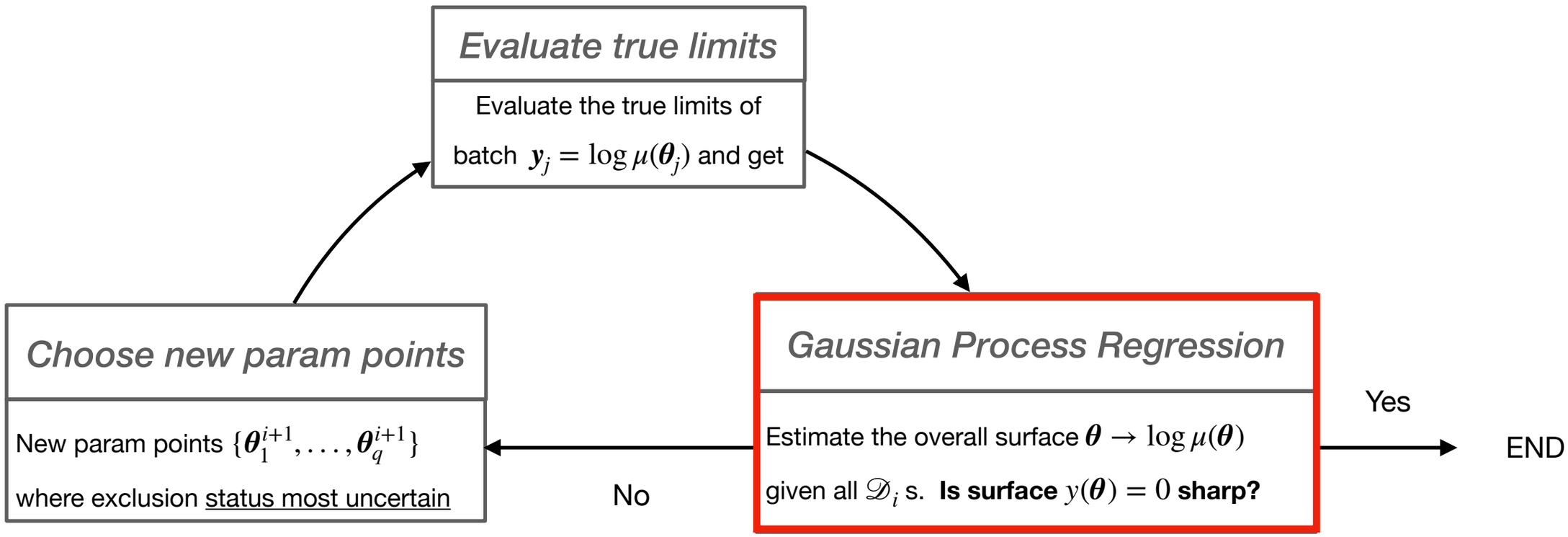
$$f : \mathbb{R}^d \to \mathbb{R}^4$$

$$\boldsymbol{\theta} \to \boldsymbol{y} := (\log \mu^{UL}_{exp}(\boldsymbol{\theta}),\ \log \mu^{UL}_{+1\sigma}(\boldsymbol{\theta}),\ \log \mu^{UL}_{-1\sigma}(\boldsymbol{\theta}),\ \log \mu^{UL}_{obs}(\boldsymbol{\theta}))$$

Contours of interest $\boldsymbol{y}(\boldsymbol{\theta}) = 0$ corresponds to $\mu^{UL}_{exp}(\boldsymbol{\theta}) = 1$ meaning $\sigma(\boldsymbol{\theta}) \approx \sigma_{BSM}(\boldsymbol{\theta})$

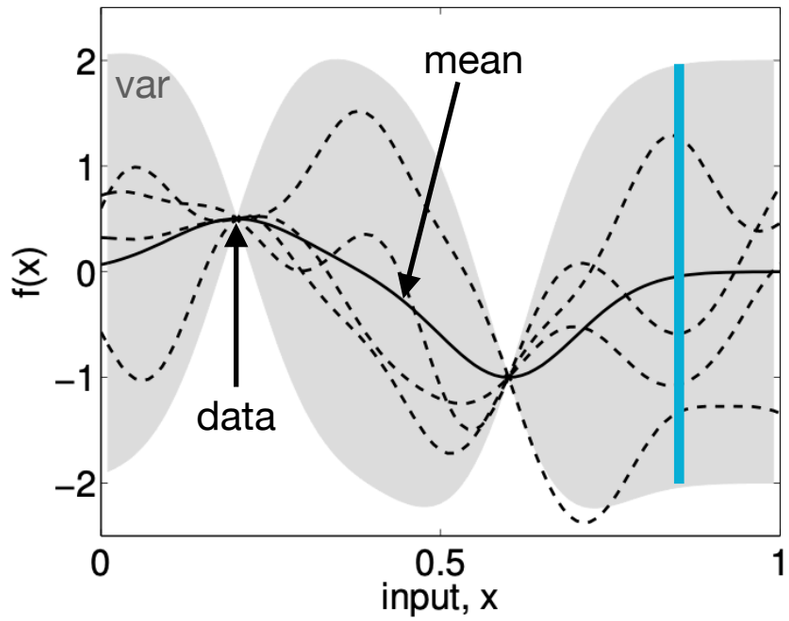
- Find the 4 contours with minimum num. queries to $f$ $\longrightarrow$ **Active Learning**

**For** each new batch of BSM parameter points $\{\boldsymbol{\theta}^i_1, \ldots, \boldsymbol{\theta}^i_q\}$ **do**:



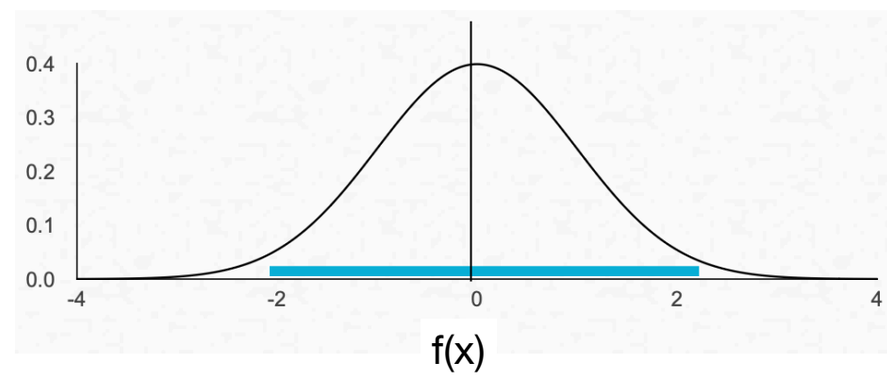
# Gaussian Process Regression

## *Basics*



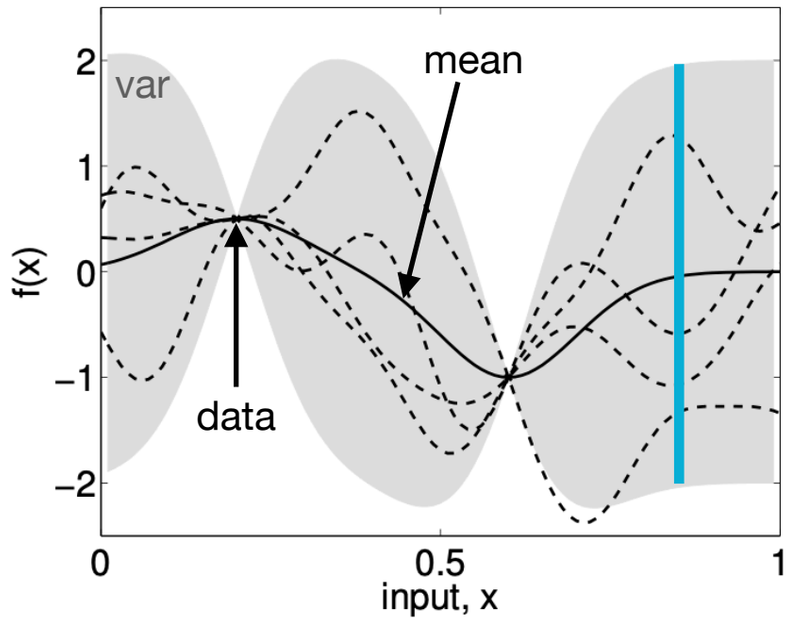

Source: Rasmussen & Williams

- GPR assumes data points have a **joint** Multivar Gaussian distr.
- Posterior distr. collapses on data points (no noise)
- For each point x : posterior ~$N(\text{mean}(x), \text{var}(x))$



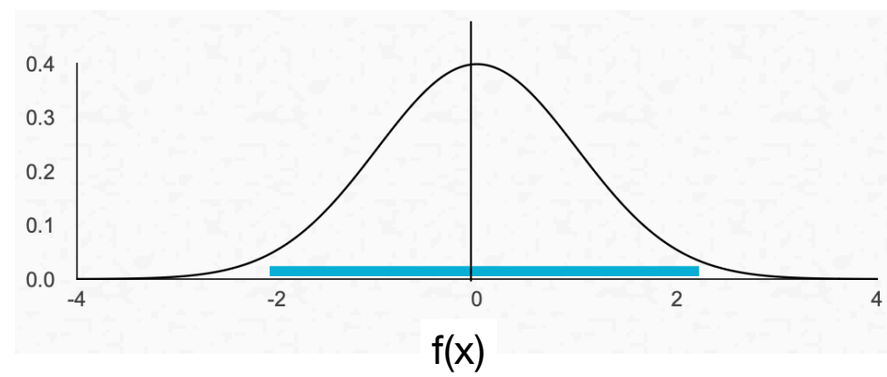
# Gaussian Process Regression

## *Basics*




Source: Rasmussen & Williams

- GPR assumes data points have a **joint** Multivar Gaussian distr.
- Posterior distr. collapses on data points (no noise)
- For each point x : posterior ~$N(\mathrm{mean}(x), \mathrm{var}(x))$



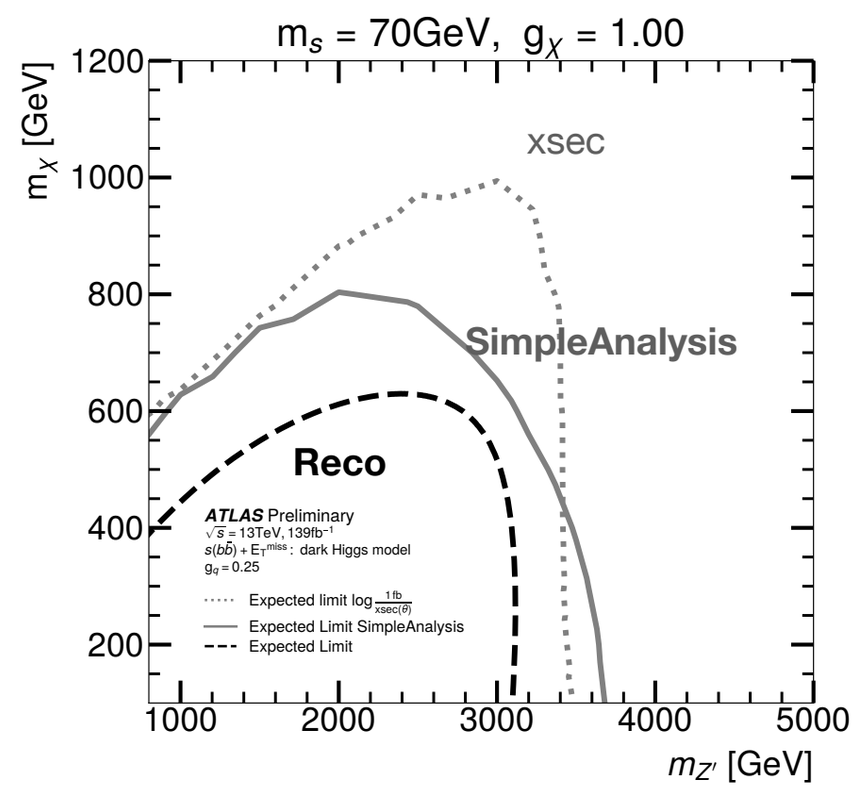

## *Multitask*

Task #1: regression SimpleAnalysis data
Task #2: regression reco-level data

**Few points reco-level but many with SimpleAnalysis**

*A vanilla GP with only reco-level data has no physics inductive bias and no info outside small 4D region*

*Leads to unreasonable fits*



# Warm start

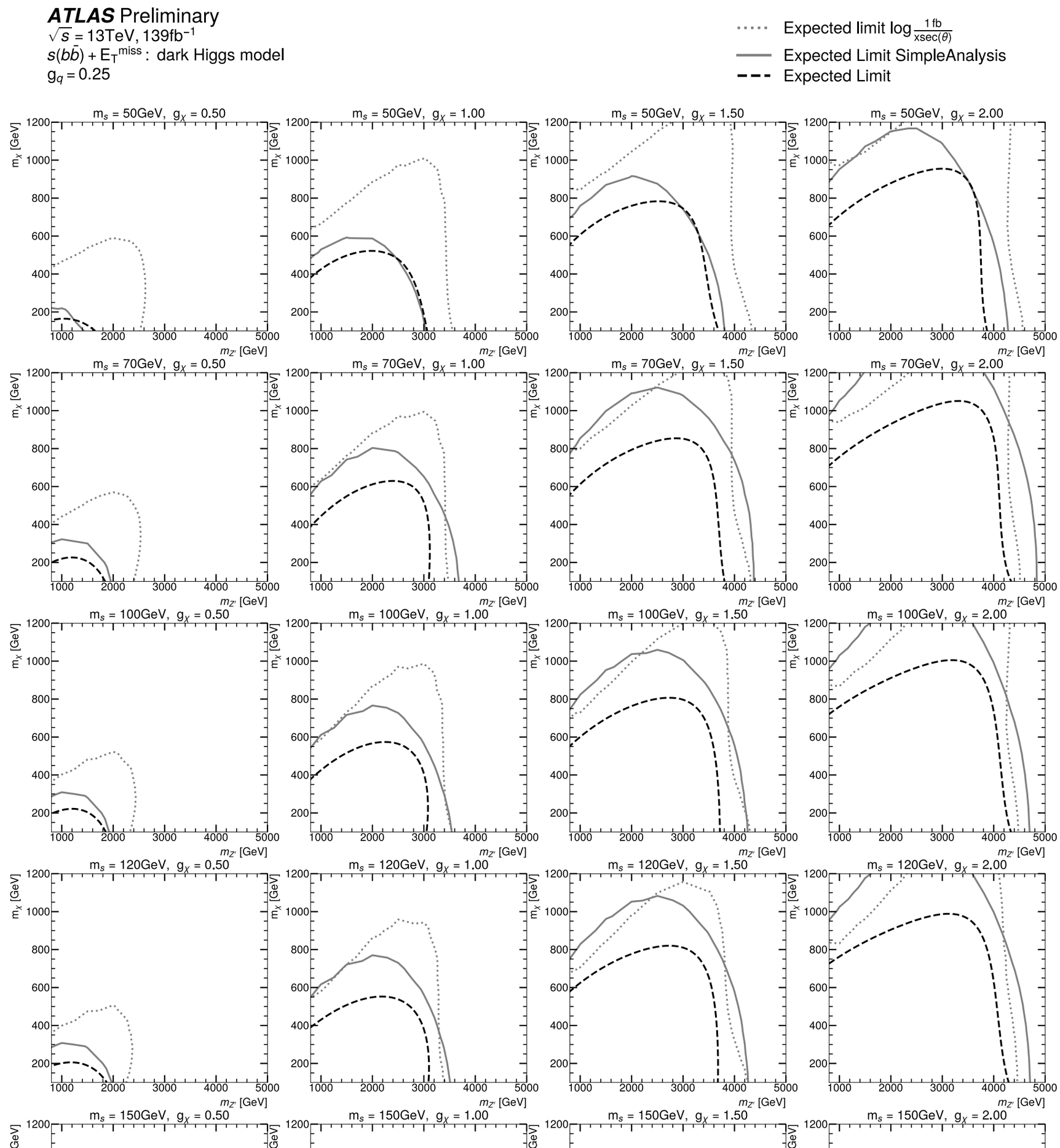
1. Populate param space with

SimpleAnalysis (low fidelity)

2. Query reco-level (high-fidelity)

Saves resources

Better than start from scratch




Irina Espejo iem244@nyu.edu

# Method

Cranmer et al. 2010

ATL-PHYS-PUB-2019-932

- Abstraction of full analysis pipeline, evaluated thanks to recast

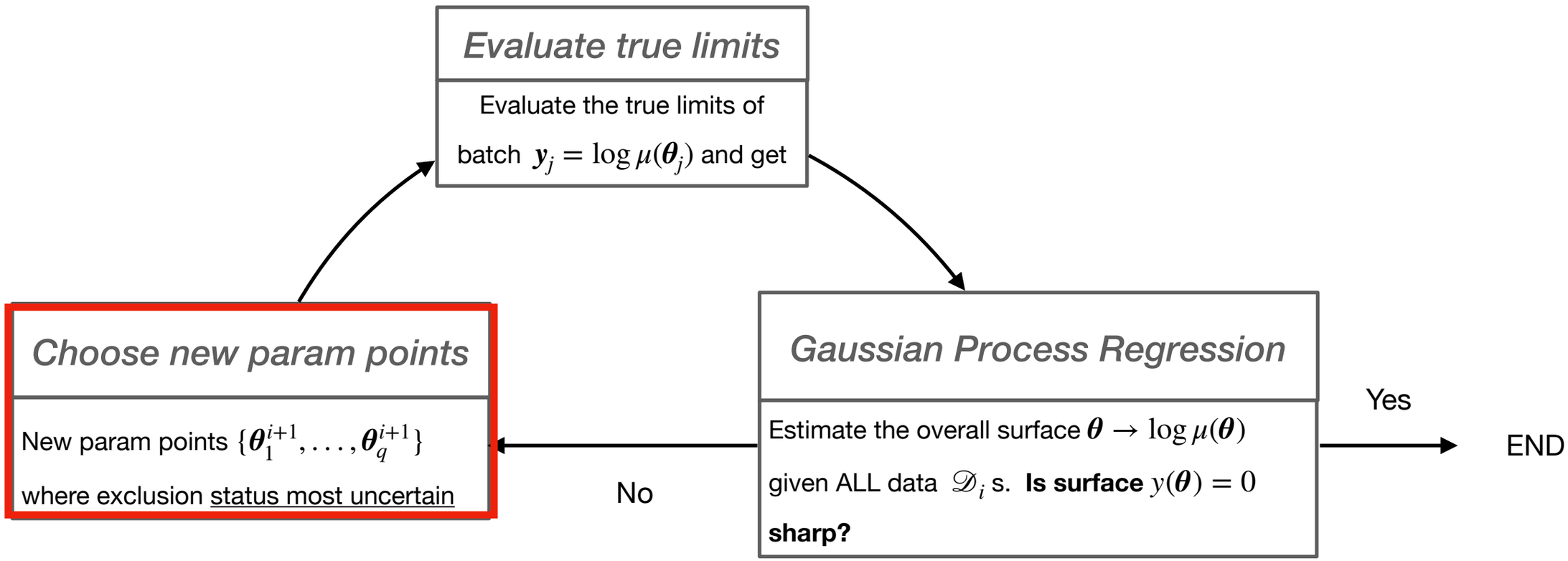
$$f : \mathbb{R}^d \to \mathbb{R}^4$$

$$\boldsymbol{\theta} \to \boldsymbol{y} := (\log \mu^{UL}_{exp}(\boldsymbol{\theta}),\ \log \mu^{UL}_{+1\sigma}(\boldsymbol{\theta}),\ \log \mu^{UL}_{-1\sigma}(\boldsymbol{\theta}),\ \log \mu^{UL}_{obs}(\boldsymbol{\theta}))$$

Contours of interest $\boldsymbol{y}(\boldsymbol{\theta}) = 0$ corresponds to $\mu^{UL}_{exp}(\boldsymbol{\theta}) = 1$ meaning $\sigma(\boldsymbol{\theta}) \approx \sigma_{BSM}(\boldsymbol{\theta})$

- Find the 4 contours with minimum num. queries to $f$ $\longrightarrow$ **Active Learning**

**For** each new batch of BSM parameter points $\{\boldsymbol{\theta}^i_1, \ldots, \boldsymbol{\theta}^i_q\}$ **do**:



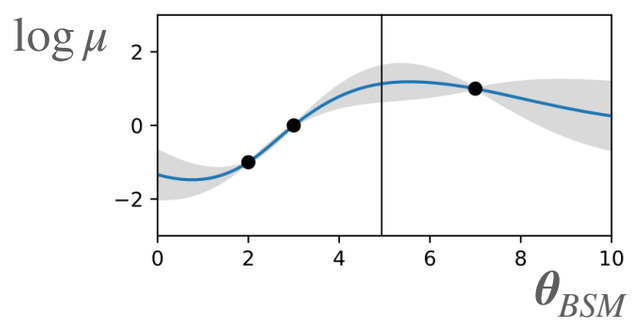
# Choosing new param points

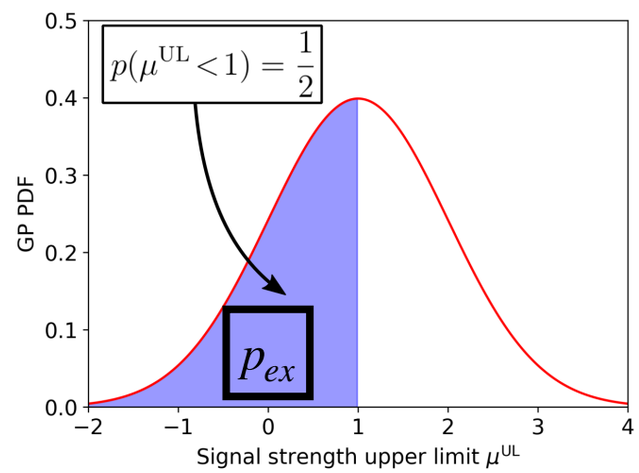
## *Entropy and Poisson-disc*

- **Exclusion Entropy** based on posterior

$$H(\boldsymbol{\theta}) = -\,p_{ex}\log p_{ex} - (1-p_{ex})\log(1-p_{ex})$$

Borrowed from Patrick Rieck



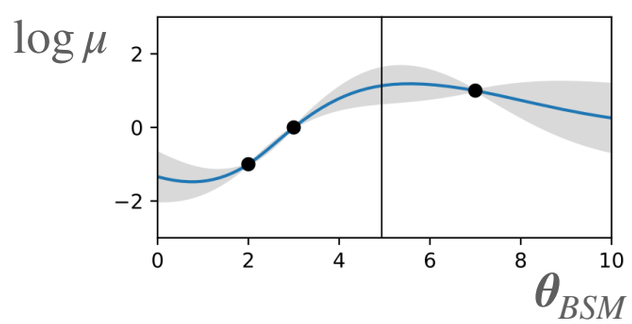
# Choosing new param points

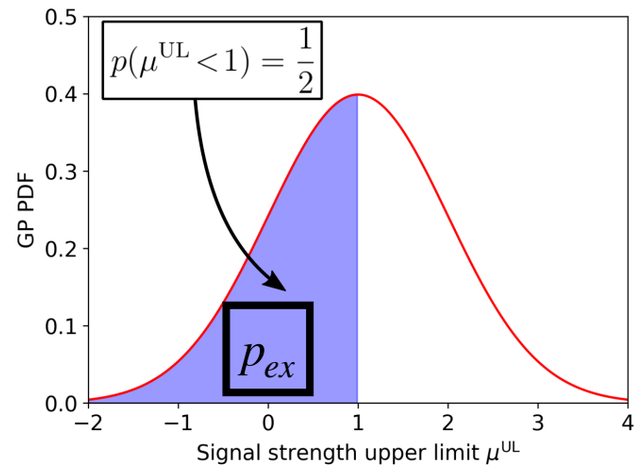
## *Entropy and Poisson-disc*

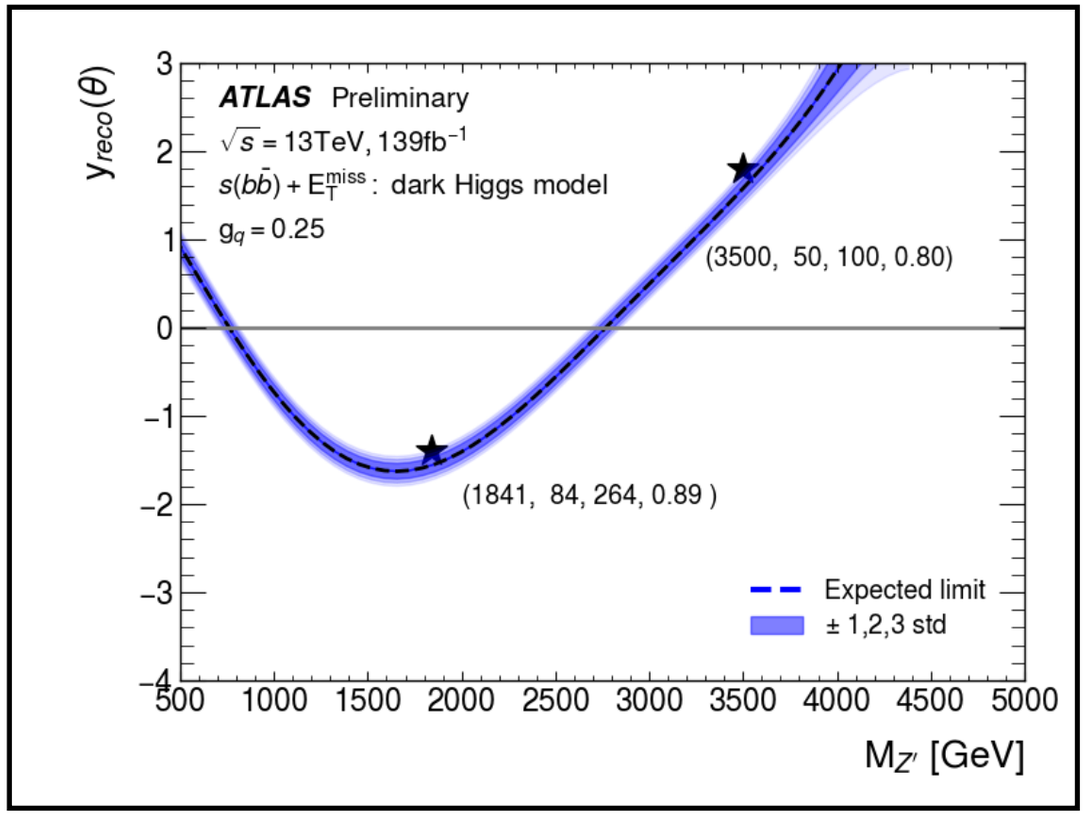
- **Exclusion Entropy** based on posterior

$$H(\boldsymbol{\theta}) = -\,p_{ex}\log p_{ex} - (1-p_{ex})\log(1-p_{ex})$$

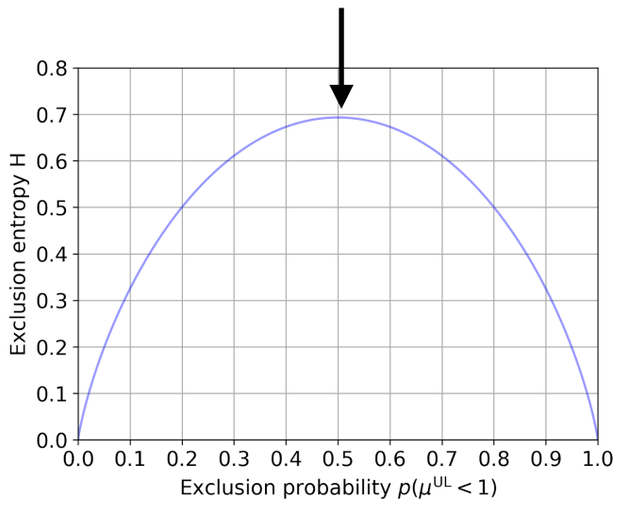
Borrowed from Patrick Rieck




- Maximum entropy achieved at $\frac{1}{2}$, maximum uncertainty wether exclude $\boldsymbol{\theta}$ or not

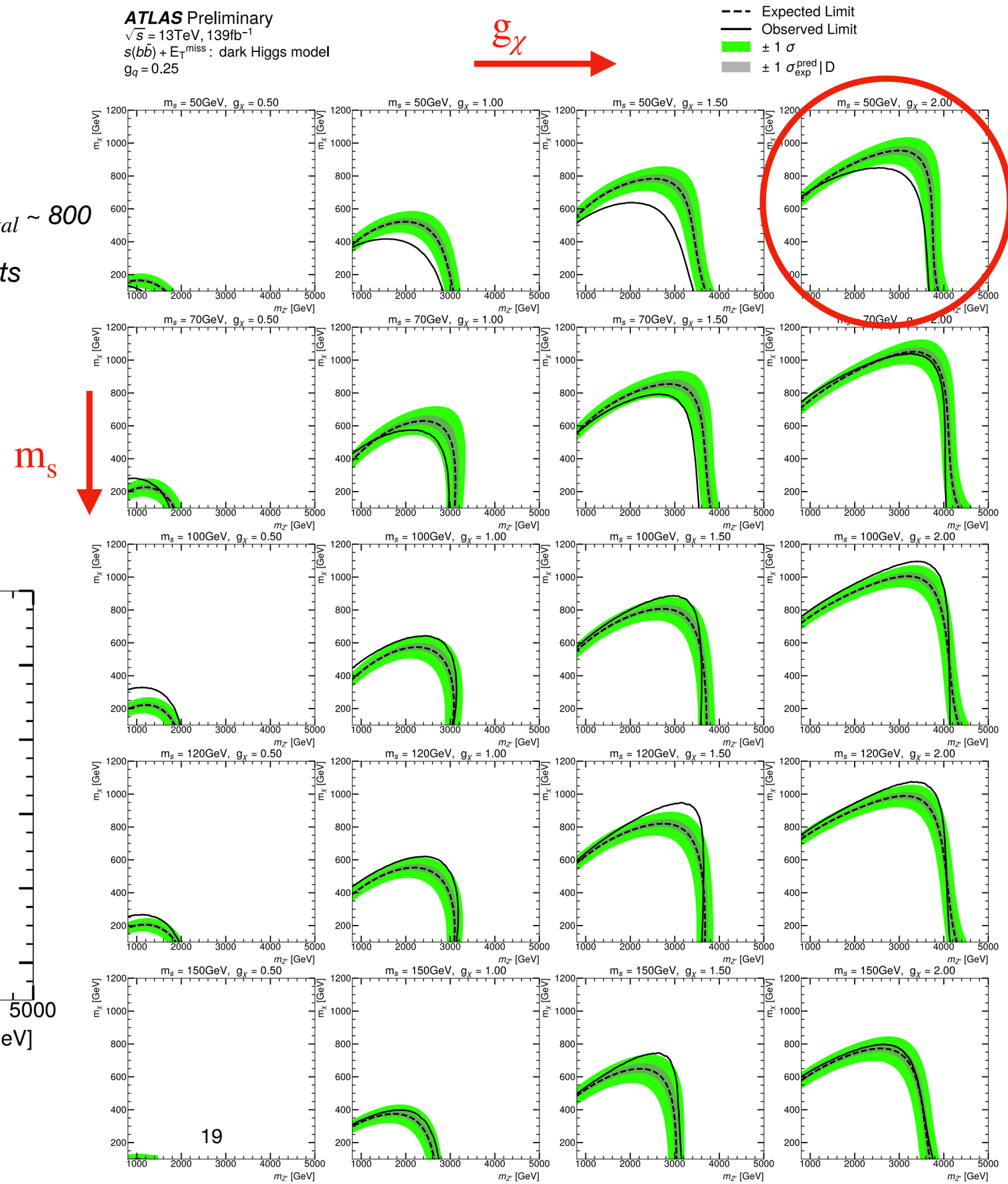


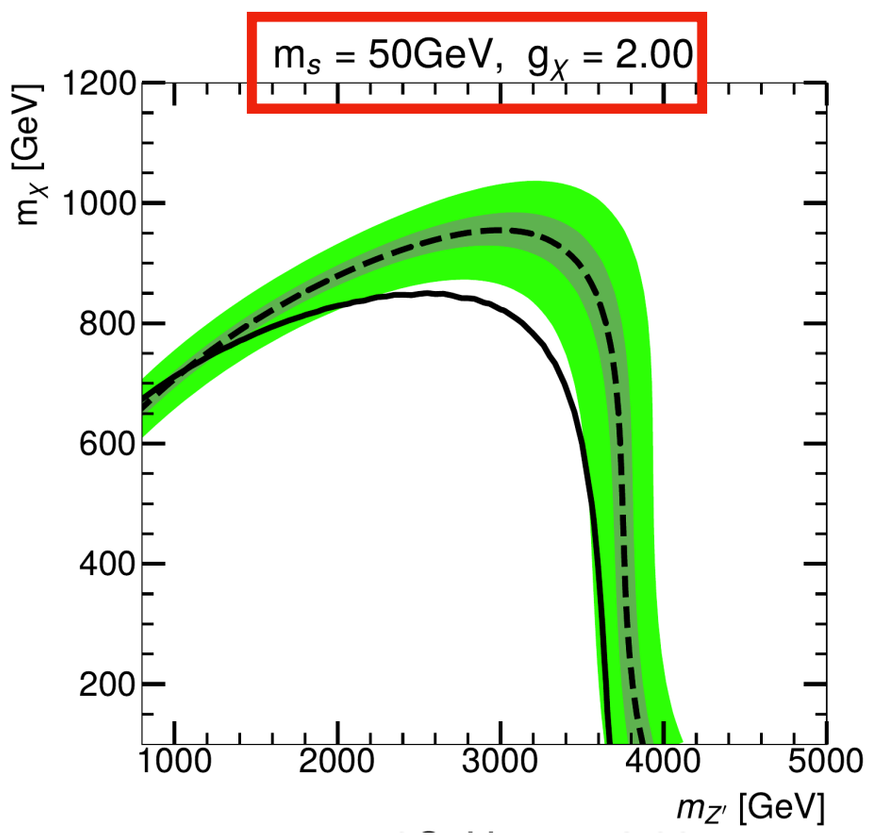
# Results

*Number of points evaluated high-res:* $\mathscr{D}_{reco,total}$ *~ 800*

*Active Learning rounds: 4* $\mathscr{D}_{SA}$ *~5000 points*

Recall $\boldsymbol{\theta} = (m_{Z'}, m_s, m_\chi, g_\chi)$



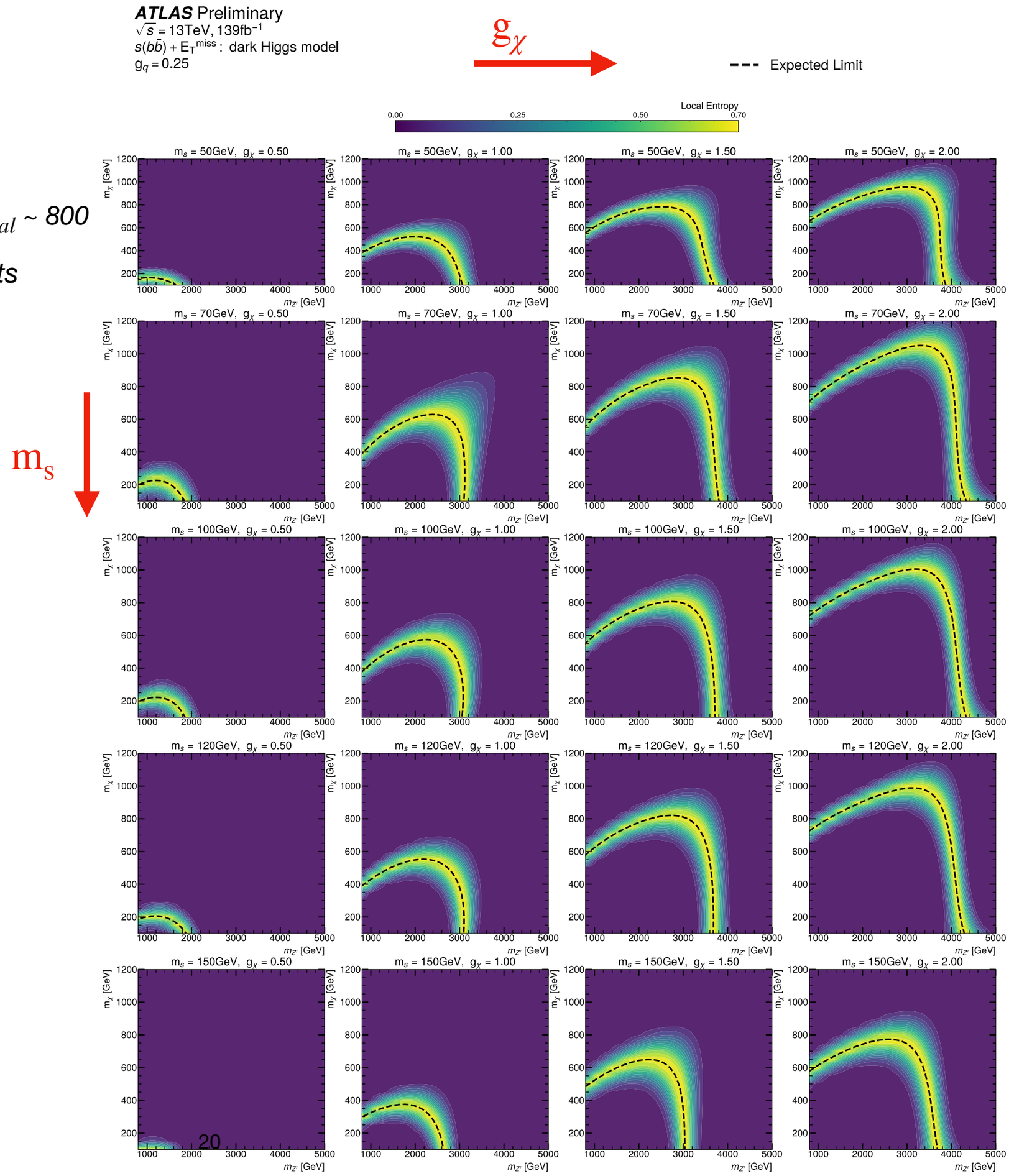

Irina Espejo iem244@nyu.edu

# Results

*Number of points evaluated high-res:* $\mathscr{D}_{reco,total}$ *~ 800*

*Active Learning rounds: 4* $\mathscr{D}_{SA}$ *~5000 points*

Recall $\boldsymbol{\theta} = (\mathrm{m}_{Z'}, \mathrm{m}_s, \mathrm{m}_\chi, \mathrm{g}_\chi)$



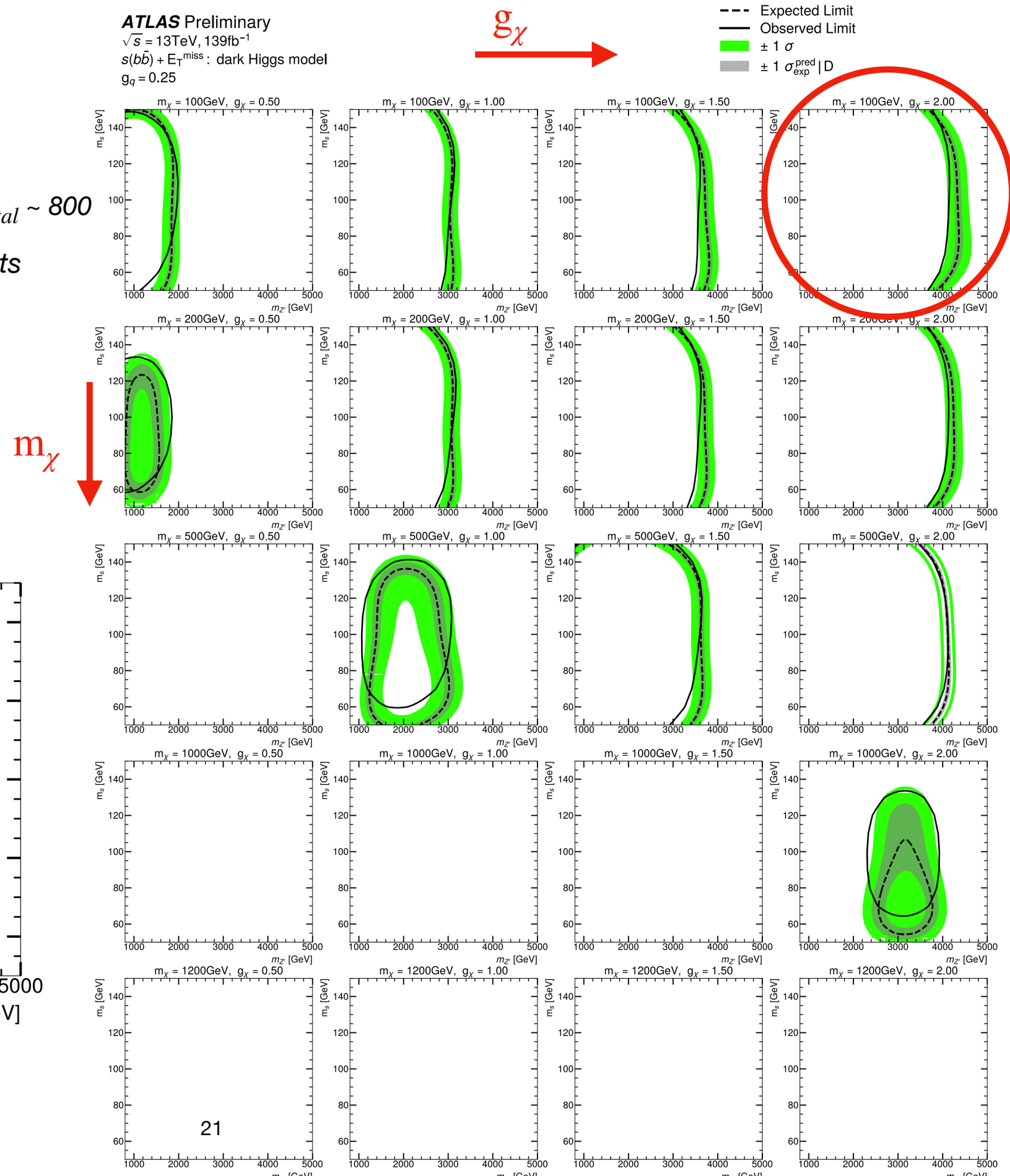

Irina Espejo iem244@nyu.edu

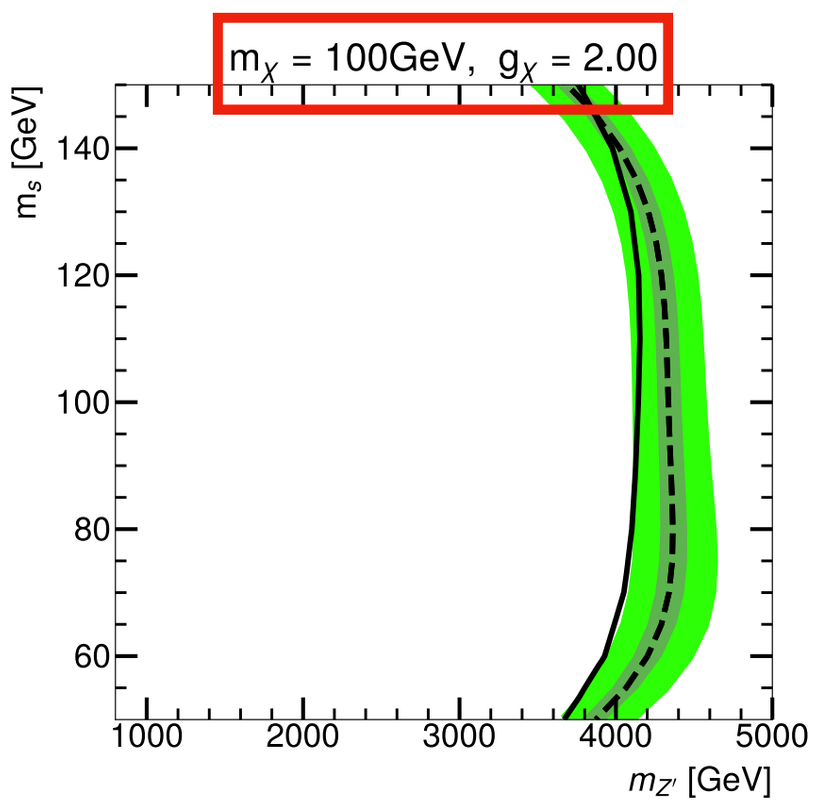
# Results

*Number of points evaluated high-res:* $\mathscr{D}_{reco,total}$ *~ 800*

*Active Learning rounds: 4* $\mathscr{D}_{SA}$ *~5000 points*

Recall $\boldsymbol{\theta} = (\mathrm{m}_{Z'}, \mathrm{m}_s, \mathrm{m}_\chi, \mathrm{g}_\chi)$



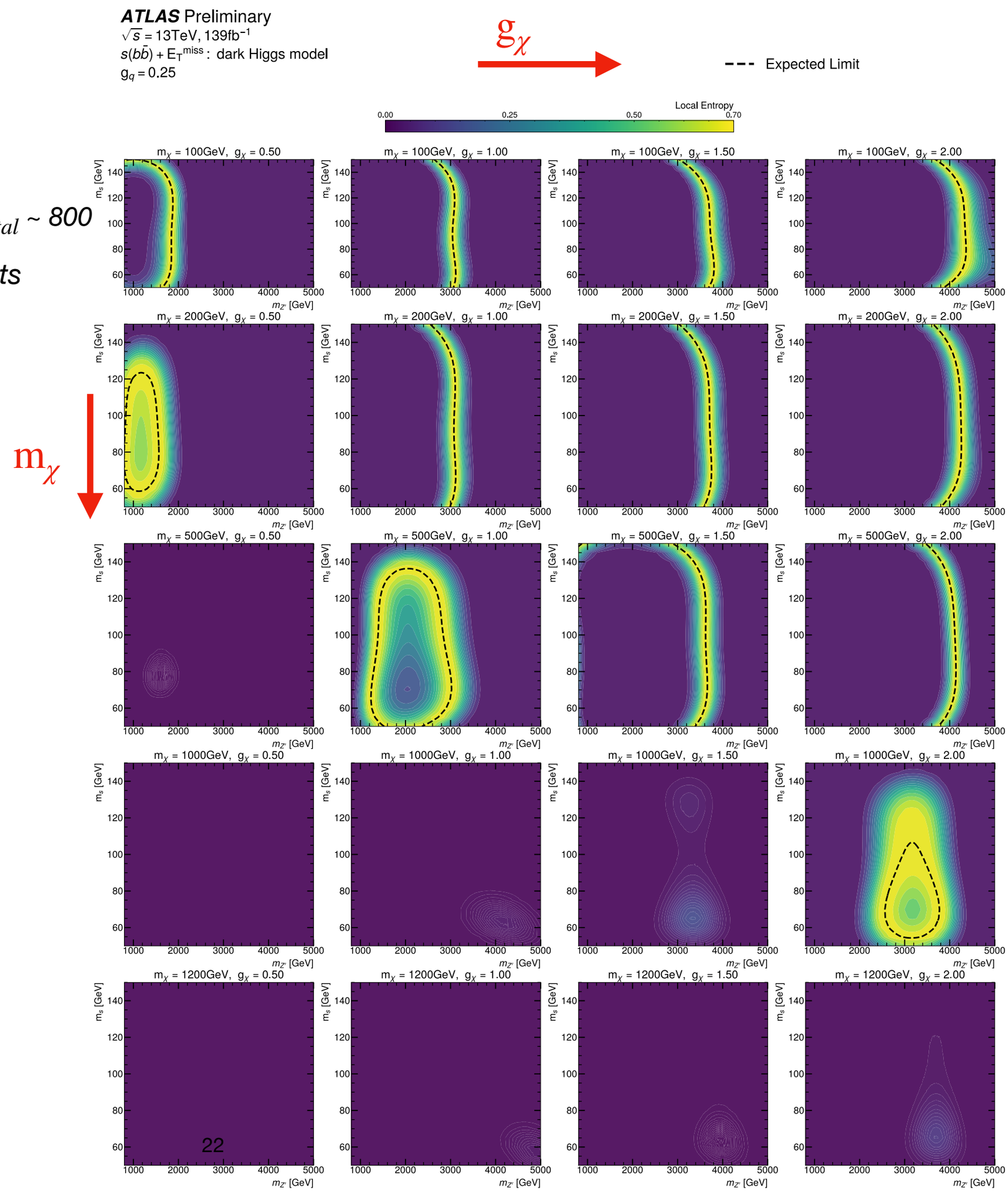

Irina Espejo iem244@nyu.edu

# Results

*Number of points evaluated high-res:* $\mathscr{D}_{reco,total}$ *~ 800*

*Active Learning rounds: 4* $\mathscr{D}_{SA}$ *~5000 points*

Recall $\boldsymbol{\theta} = (\mathrm{m}_{Z'}, \mathrm{m}_s, \mathrm{m}_\chi, \mathrm{g}_\chi)$

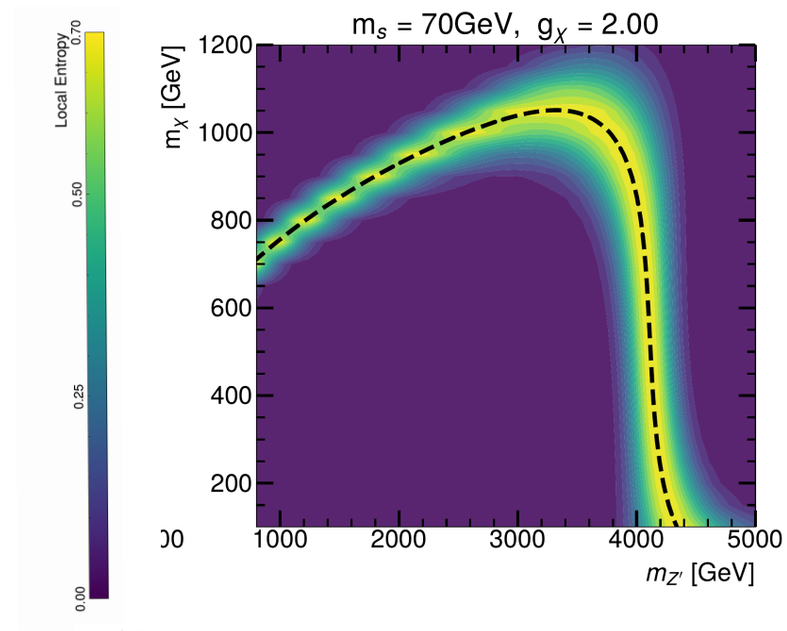



Irina Espejo iem244@nyu.edu

# Results

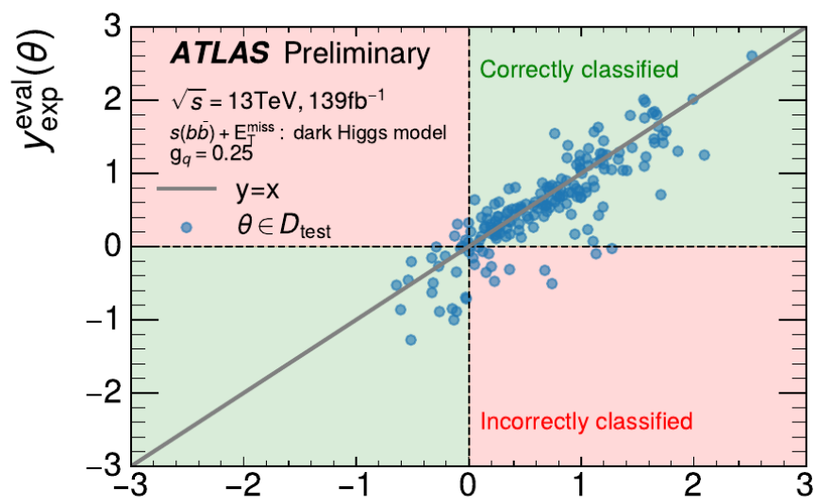
## *Convergence and evaluation metrics*

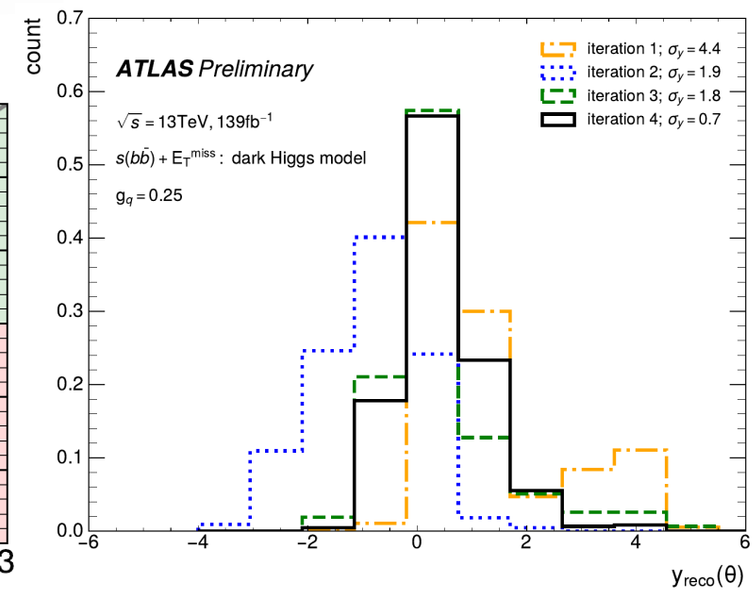
- *Sharp local entropy around expected contour*




- *True target distribution concentrates around 0*

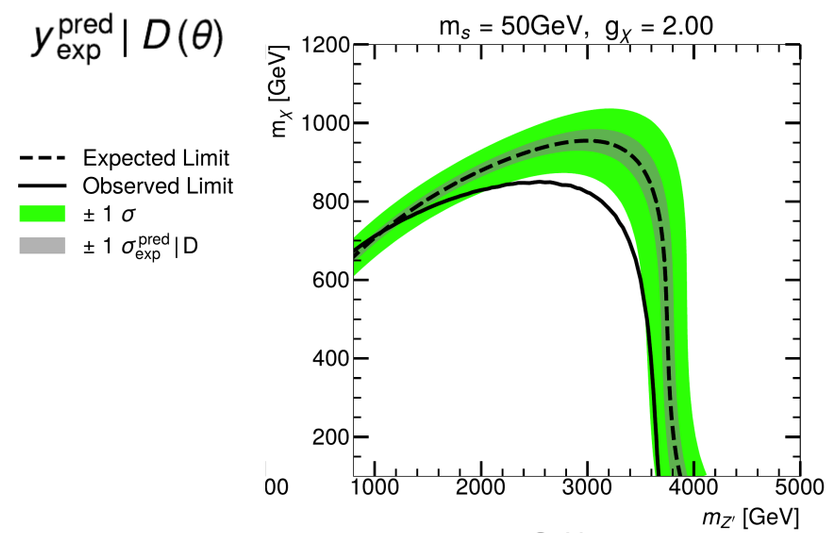
- *Classification confusion matrix test set!*




- *Posterior variance on expected*

  $y_{exp,reco}$ << *posterior mean* $y_{exp,reco} \pm y_{\pm 1\sigma,reco}$



# Take home message

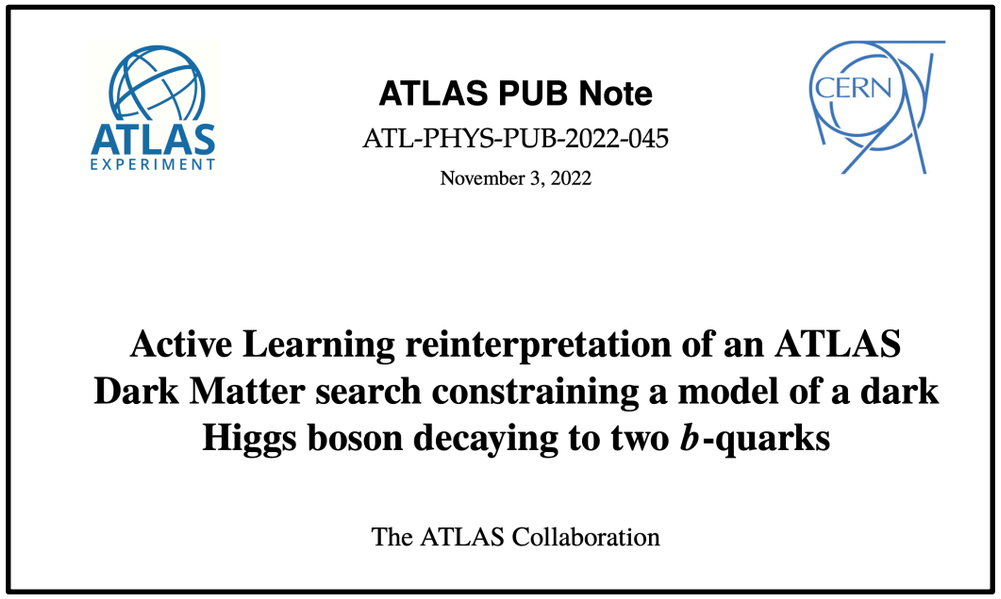
- Active Learning can scale the production of exclusion contours to > 2D in a full analysis.

- Warm start is helpful to populate param space before starting Active Learning.

- Cross-section data is close to SimpleAnalysis in some regions of param space.

  Cross-section data can be a good warm start, widely available.

ATL-PHYS-2022-045


ATLAS PUB Note
ATL-PHYS-PUB-2022-045
November 3, 2022

**Active Learning reinterpretation of an ATLAS Dark Matter search constraining a model of a dark Higgs boson decaying to two *b*-quarks**

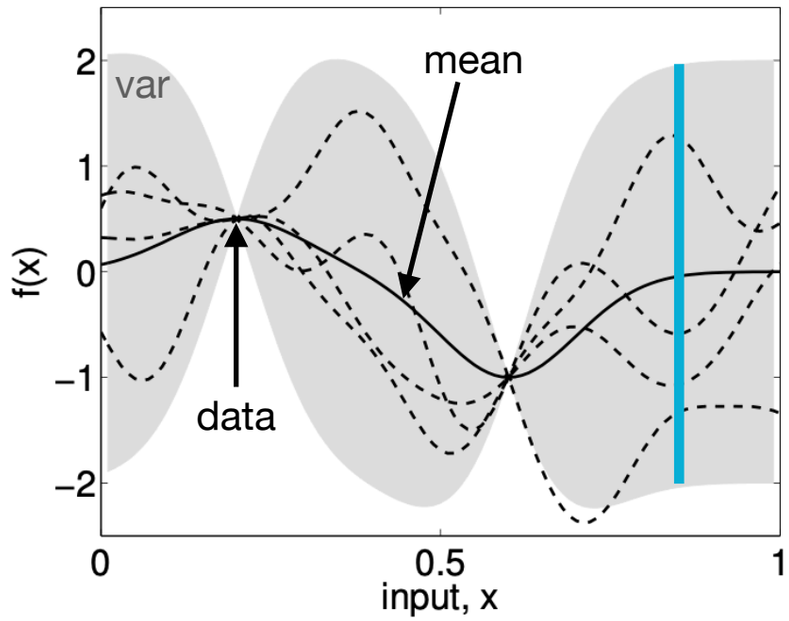
The ATLAS Collaboration

# Thank you!

# Q&A

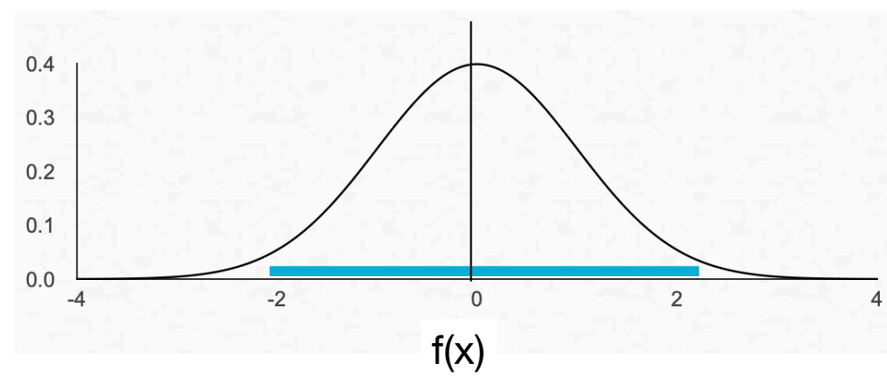
# Gaussian Process Regression

## *Basics*




Source: Rasmussen & Williams

- GPR assumes data points have a **joint** Multivar Gaussian distr.
- Posterior distr. collapses on data points (no noise)
- For each point x : posterior | data ~$\mathcal{N}(\mathrm{mean}(x), \mathrm{var}(x))$




## *Multitask*

Task #1: regression SimpleAnalysis data
Task #2: regression reco-level data

$$(y_{exp,\,SA}(\boldsymbol{\theta}), y_{exp,\,reco}(\boldsymbol{\theta})) \sim \mathcal{GP}(m(\boldsymbol{\theta}), k(\boldsymbol{\theta}, \boldsymbol{\theta}'))$$

$$m(\boldsymbol{\theta}) := \boldsymbol{w}^T\boldsymbol{\theta} + b \quad \text{with} \quad \boldsymbol{w} \in \mathbb{R}^4, b \in \mathbb{R}$$

$$k(\boldsymbol{\theta}, \boldsymbol{\theta}') = k_{task} \odot k_{RBF}(\boldsymbol{\theta}, \boldsymbol{\theta}') + \epsilon^2 \delta(\boldsymbol{\theta}, \boldsymbol{\theta}')$$

where $k_{task} = \begin{pmatrix} \Sigma_{rr} & \Sigma_{rs} \\ \Sigma_{rs} & \Sigma_{ss} \end{pmatrix}$ with $\Sigma_{rr}, \Sigma_{ss}, \Sigma_{rs} \in \mathbb{R}$

where $k_{RBF}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp\left(-\frac{||\boldsymbol{\theta} - \boldsymbol{\theta}'||^2}{2\,l^2}\right)$, with $l \in \mathbb{R}_+$

**Few points reco-level but many with SimpleAnalysis**

*A vanilla GP with only reco-level data has no physics inductive bias and no info outside small 4D region*

*Leads to unreasonable fits*

Bonilla et al. 2007
Implemented with **GPyTorch** (Gardner et al. 2018)
Hamelijnck et al. 2019