21.10.2025

Searching for Primordial Black Holes with LSST

& The Impact of False Positives on Microlensing Detection

Benedict Crossey
Institute of Particle Physics Phenomenology
Durham University

In collaboration with Djuna Croon, Miguel Crispim Romão & Daniel Godines

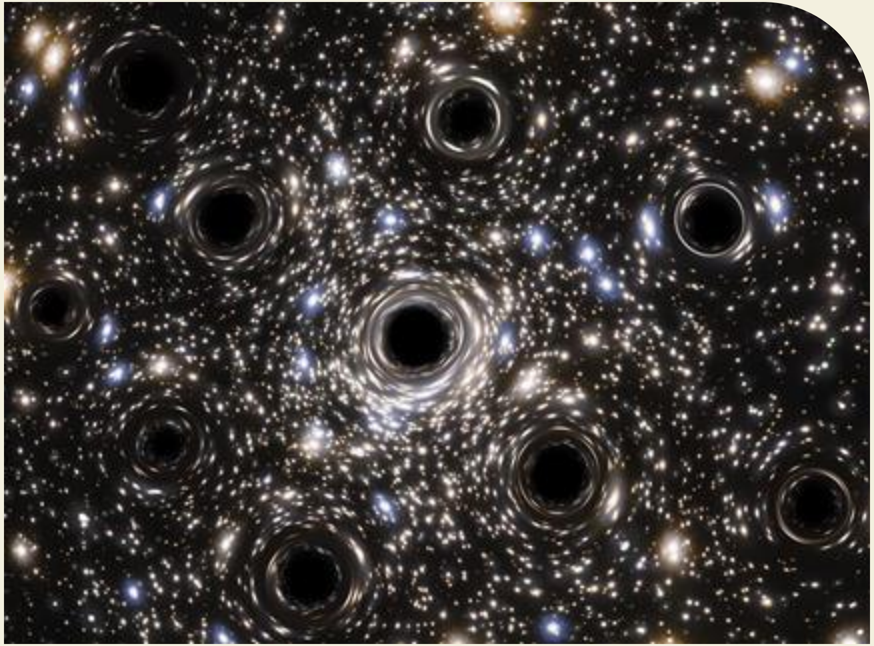
JCAP10 (2025) 066

arXiv:2506.20709

Credit: RubinObs/NOIRLab/SLAC/NSF/DOE/AURA

Introduction

- The Legacy Survey of Space and Time (LSST) is a full-sky survey which launched this summer at the Vera C. Rubin Observatory in Chile
- It will run for 10 years and observe ~10^9 stars in our galaxy
- Can be used to search for gravitational microlensing events
- Microlensing events can be used to identify – or put bounds on – primordial black holes (PBHs)



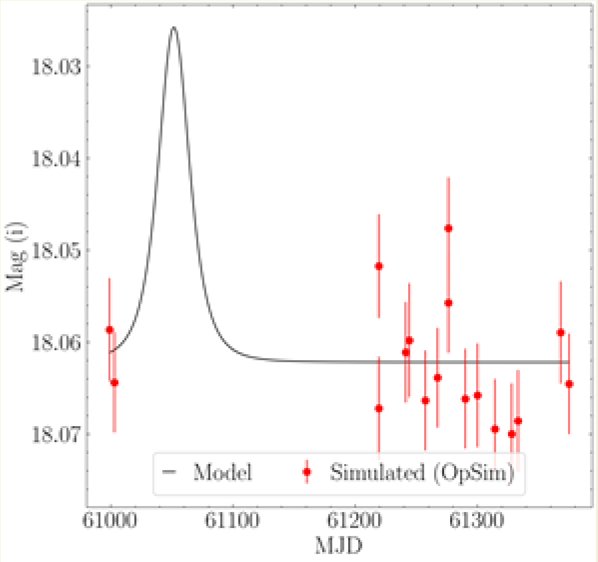
Credit: ESA/Hubble, N. Bartmann

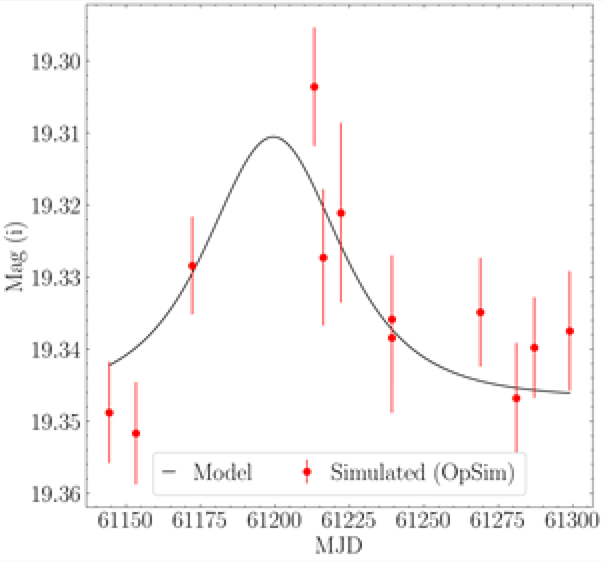
Identifying Microlensing in LSST Data

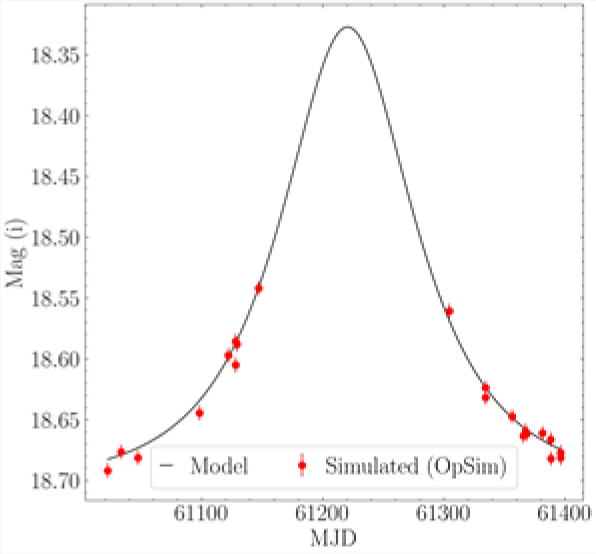
- Simulated ~10^6 Constant and point-like Microlensing (ML) light curves using the Operations Simulator (OpSim) codebase from Rubin
- Due to LSST cadence, can only expect to see ~800 observations in the bulge over the full 10 years

Identifying Microlensing in LSST Data

- Simulated ~10^6 Constant and point-like Microlensing (ML) light curves using the Operations Simulator (OpSim) codebase from Rubin
- Due to LSST cadence, can only expect to see ~800 observations in the bulge over the full 10 years

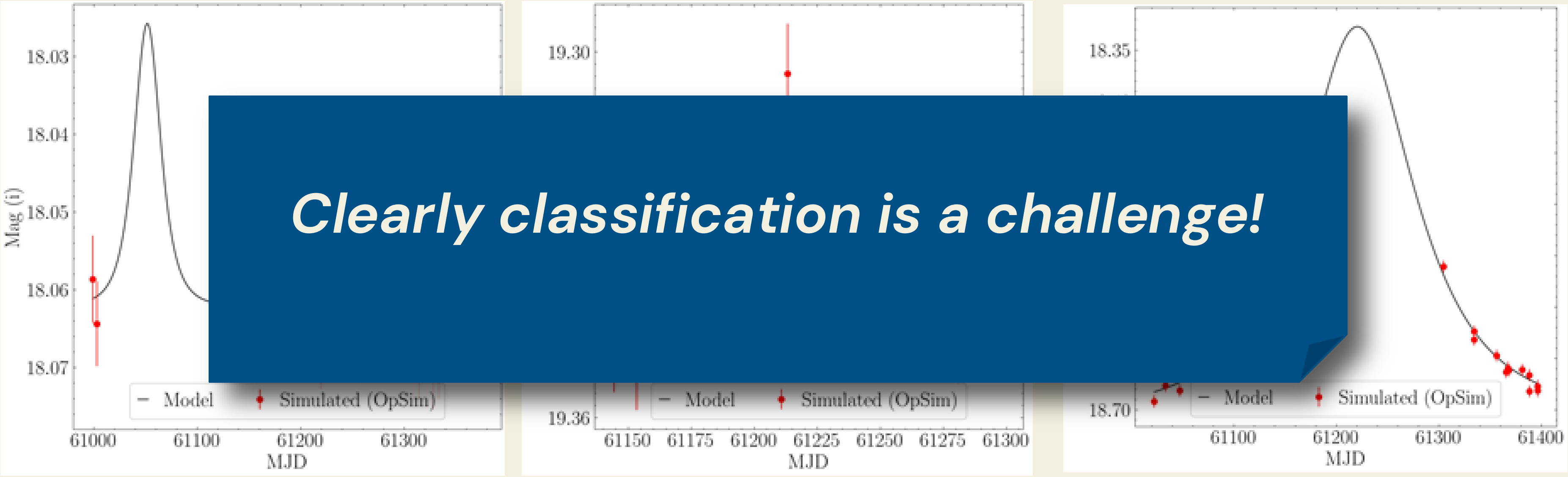






Identifying Microlensing in LSST Data

- Simulated ~10^6 Constant and point-like Microlensing (ML) light curves using the Operations Simulator (OpSim) codebase from Rubin
- Due to LSST cadence, can only expect to see ~800 observations in the bulge over the full 10 years



BDT and BIC Ratio as Classifiers

In order to distinguish between the two light curve classes (Constant and ML), we trained a boosted decision tree (BDT) on a subset of the data:

- 0 = Constant Class
- 1 = ML Class

BDT and BIC Ratio as Classifiers

In order to distinguish between the two light curve classes (Constant and ML), we trained a boosted decision tree (BDT) on a subset of the data:

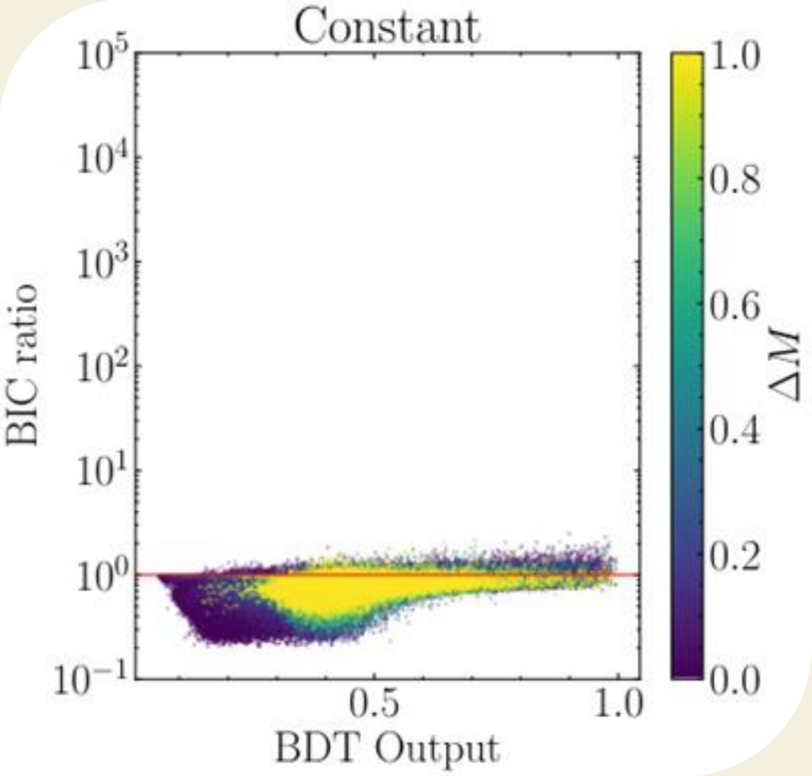
- 0 = Constant Class
- 1 = ML Class
- Bayesian Information Criterion (BIC):

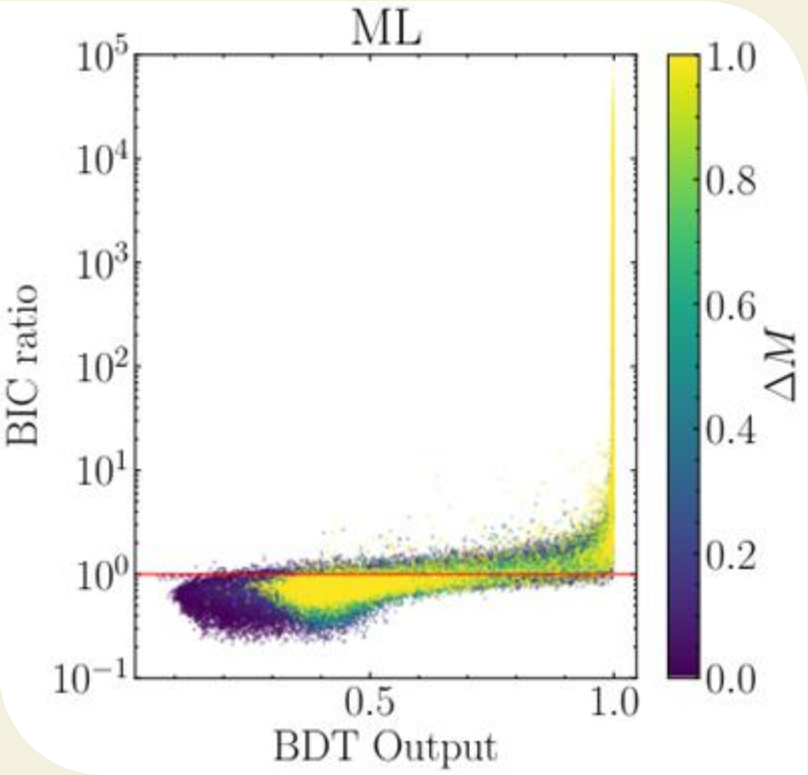
$$\mathrm{BIC}(\mathcal{M}) \;=\; k \ln n \;-\; 2 \ln \hat{L} = k \ln n + \chi^2$$

- Can compare the Constant and ML BICs for a given lightcurve using BIC ratio:

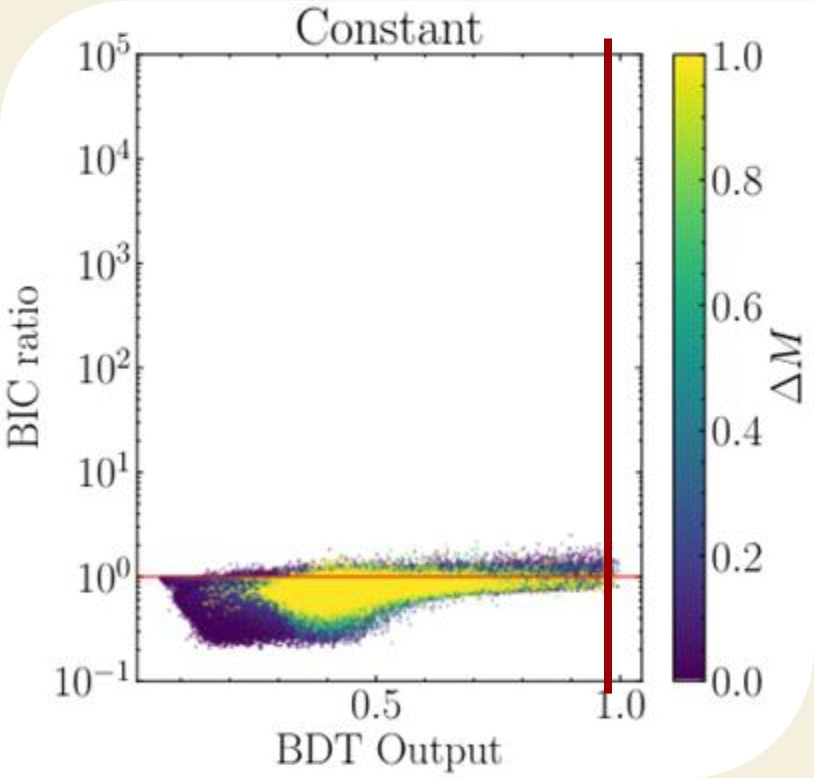
$$\text{BIC ratio} = \frac{\mathrm{BIC}(\mathcal{M}_{\mathrm{Const}})}{\mathrm{BIC}(\mathcal{M}_{\mathrm{ML}})} = \frac{\chi^2_{\mathrm{Const}} + \ln n}{\chi^2_{\mathrm{ML}} + 5 \ln n}$$

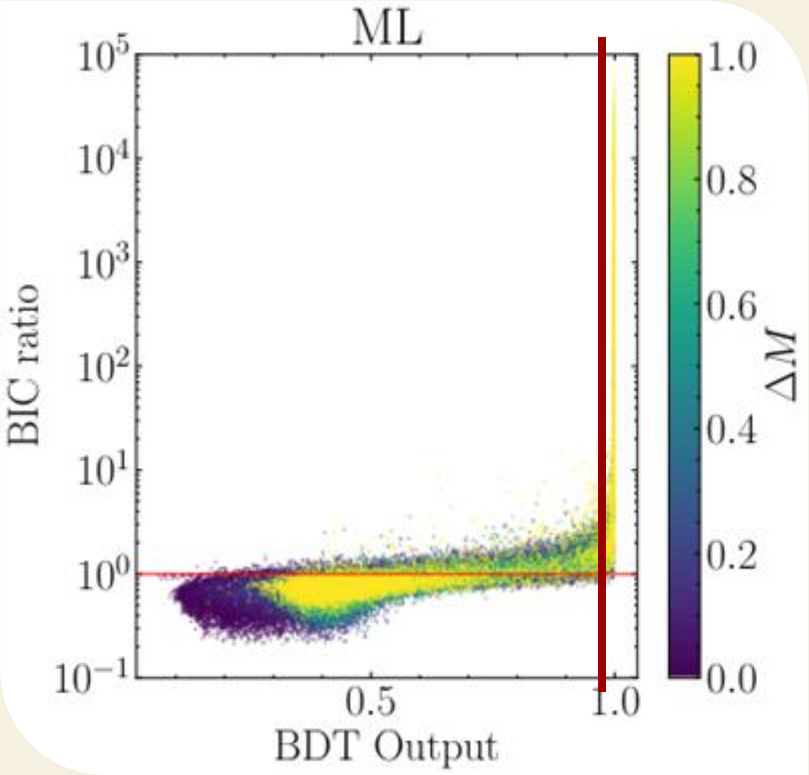
BDT and BIC Ratio as Classifiers



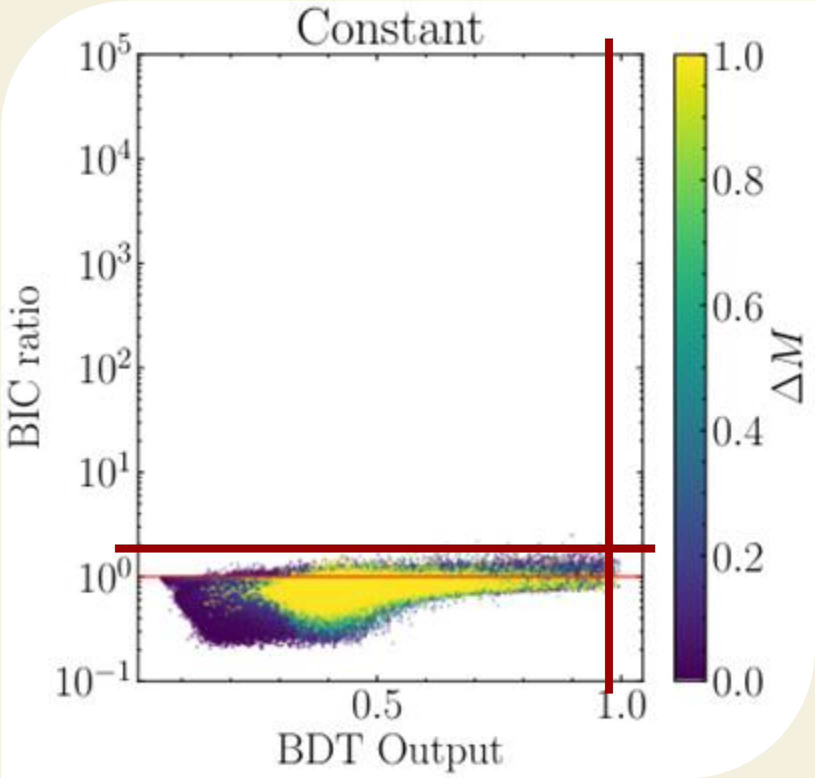


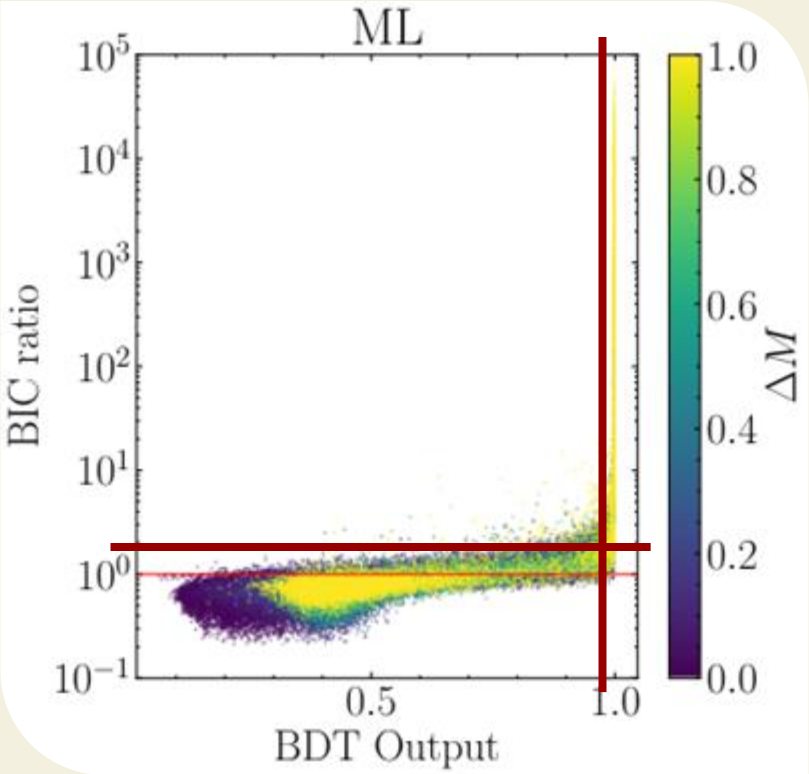
BDT and BIC Ratio as Classifiers





BDT and BIC Ratio as Classifiers





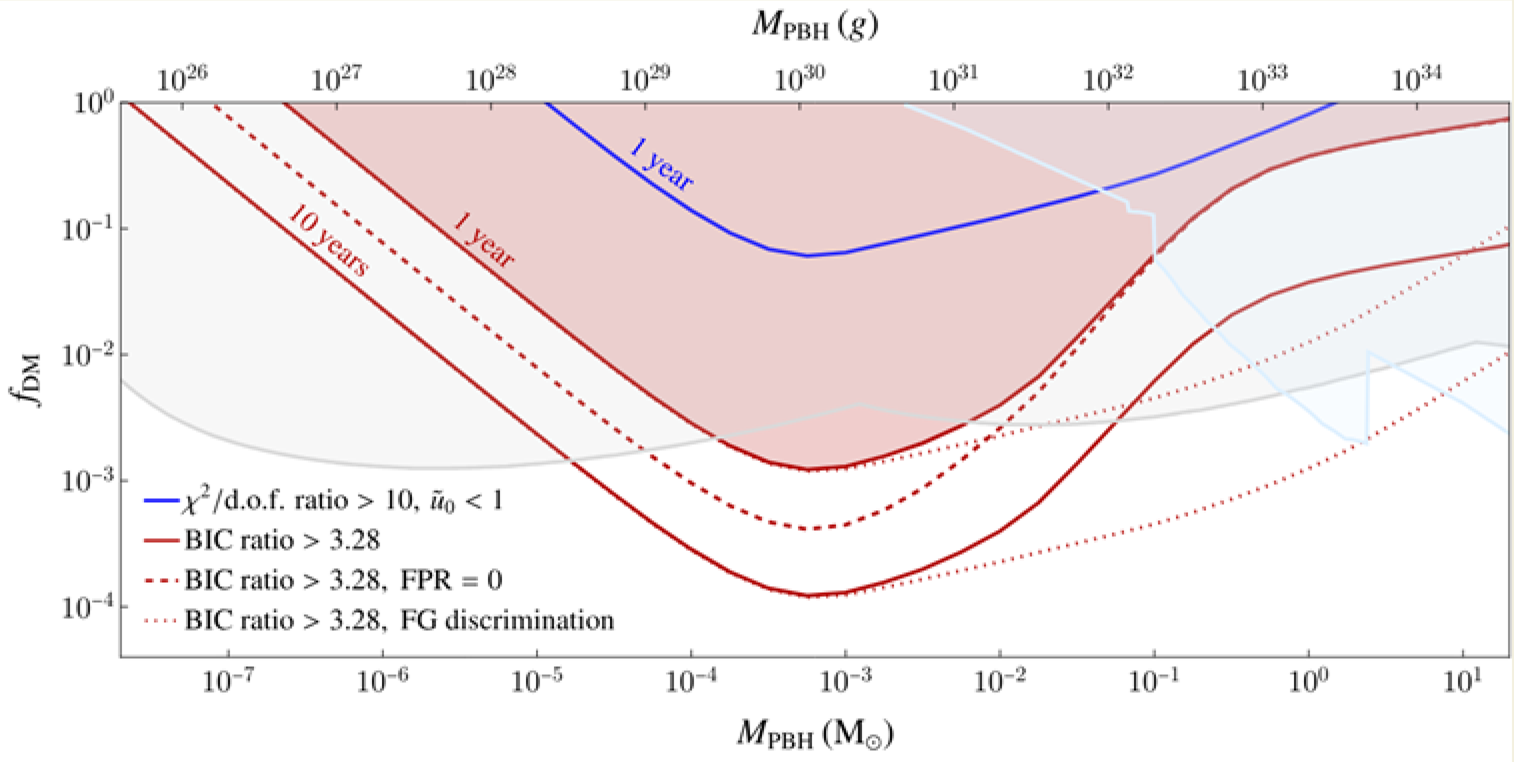
The Importance of False Positives

- Expect some amount of false positives to contaminate the data:
    - False Positive Rate (FPR) = N_{FP} / (N_{FP} + N_{TP})
- With ~10^6 light curves simulated, cannot predict an FPR of lower than ~10^{-6} per star per year (without extrapolation)
- Bad news for LSST – ~10^9 stars visited, so with this FPR we would expect ~10^3 false positives – completely drowning out the expected foreground (from our modelling ~10^2 foreground events expected per year)
- ***Therefore it is absolutely crucial to effectively control the false positive rate if we want to set novel constraints on PBHs using Rubin.***

Extrapolating to Lower FPR

cut	FPR	$\bar{\epsilon}$
BIC ratio > 3.28	10^{-7}	0.38
BDT > 0.999	10^{-7}	0.34
χ^2/d.o.f. ratio > 10	3.5×10^{-4}	0.30
χ^2/d.o.f. ratio > 10, $\tilde{u}_0 < 1$	1.1×10^{-4}	0.20

Projected Constraints on PBHs



Thank you for listening!

Any questions?

Backup Slides

Tail Fits

- Pareto Distribution:

$$f(x, b) = b/x^b$$

- Johnson S_B:

$$f(x, a, b) = (b/(x(1-x)))\phi(a + b\log(x/(1-x)))$$

where ϕ is the normal distribution probability distribution function.

Analytic Efficiency Function

$$\epsilon = \left(\left(\frac{t_{\mathrm{E}}}{t_0} \right)^{-1/t_r} + 1 \right)^{-1}$$

Calculating Constraints

$$\kappa = 2 \sum_{i=1}^{N_{\mathrm{bins}}} \left[N_i^{\mathrm{FG,eff}} - N_i^{\mathrm{SIG}} + N_i^{\mathrm{SIG}} \ln \frac{N_i^{\mathrm{SIG}}}{N_i^{\mathrm{FG,eff}}} \right]$$

- 90% CL found by locating (M_{PBH}, f_{DM}) for which $\boldsymbol{\kappa}$ = 4.61

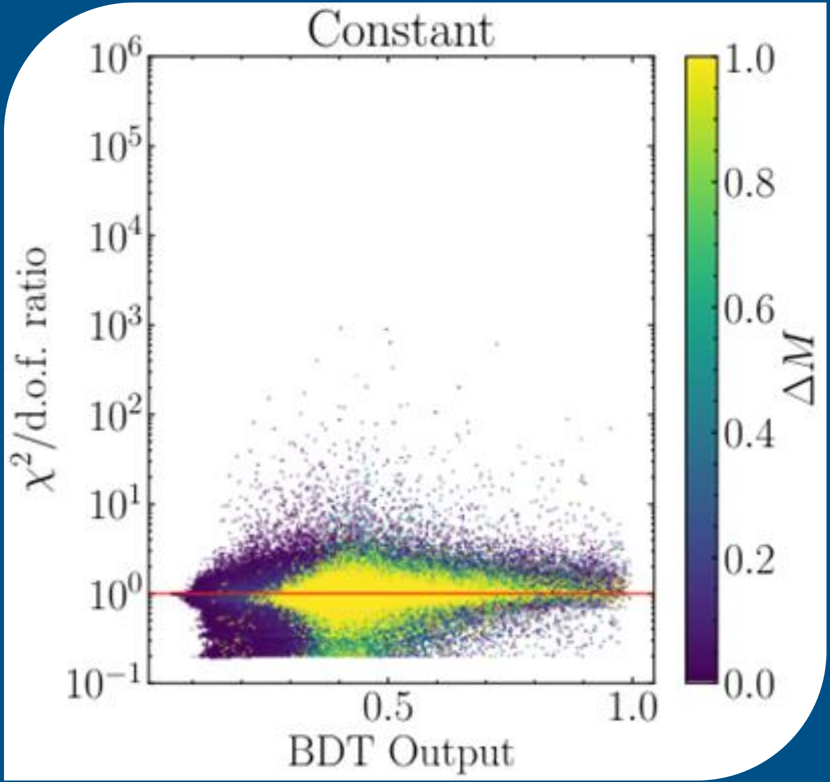
χ^2 Ratio

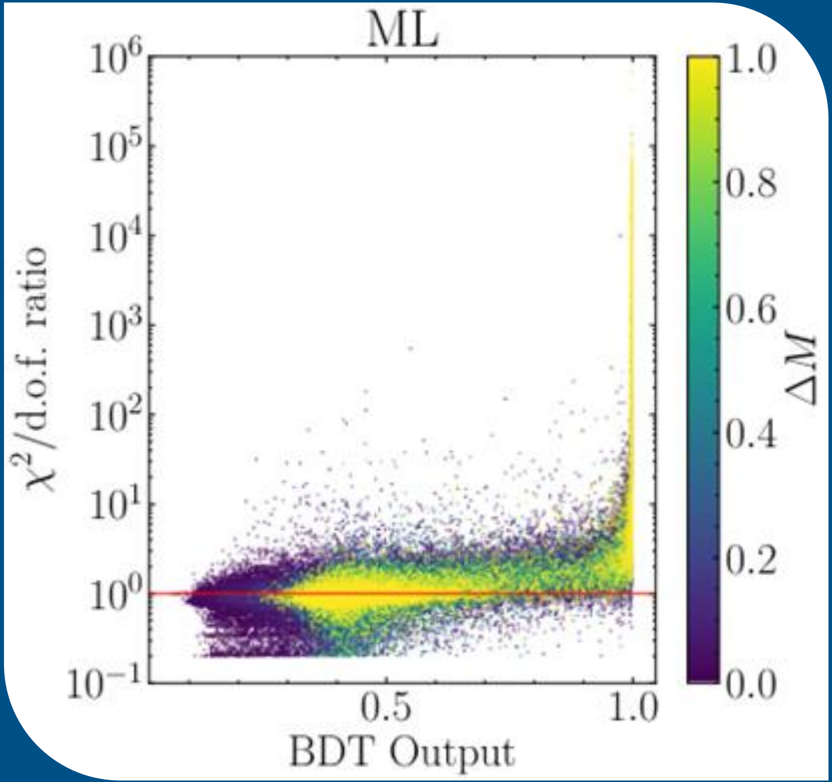
- Even for Constant class lightcurves, can often get low χ^2 values for ML fits by fitting to the noise
- Can compare the Constant and ML goodness-of-fits for a given lightcurve using the ratio of the reduced χ^2:

$$\chi^2/\text{d.o.f. ratio} \equiv \frac{\chi^2_{\text{Const}}/\nu_{\text{Const}}}{\chi^2_{\text{ML}}/\nu_{\text{ML}}}$$

  - $\nu_{\text{Const}} = n - 1$; $\nu_{\text{ML}} = n - 5$
  - Often get very large values for Constant lightcurves, suggesting ML model fits better than Constant model!

χ^2 Ratio

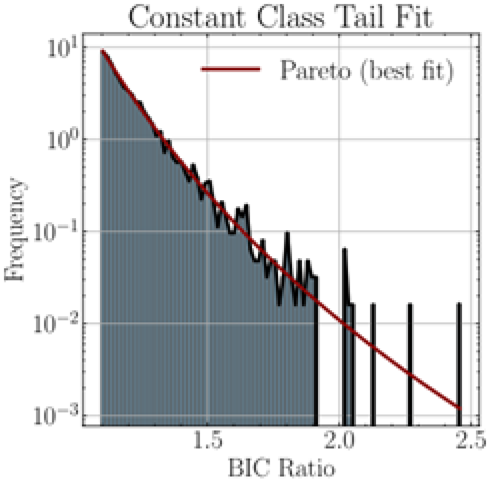

Constant
10^6
10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
χ^2/d.o.f. ratio
0.5
1.0
BDT Output
1.0
0.8
0.6
0.4
0.2
0.0
ΔM

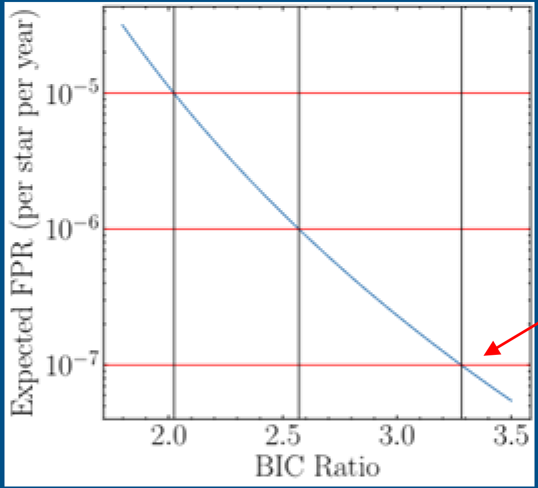

ML
10^6
10^5
10^4
10^3
10^2
10^1
10^0
10^{-1}
χ^2/d.o.f. ratio
0.5
1.0
BDT Output
1.0
0.8
0.6
0.4
0.2
0.0
ΔM

Extrapolating to Lower FPR

BIC Ratio

- To bring number of false positive events expected per year down to same level as foreground, need FPR ~ 10^{-7}
- Zooming into tail composed of top 1% of BIC ratios for Constant class, we find we can fit the tail well with a Pareto distribution using `distfit`:



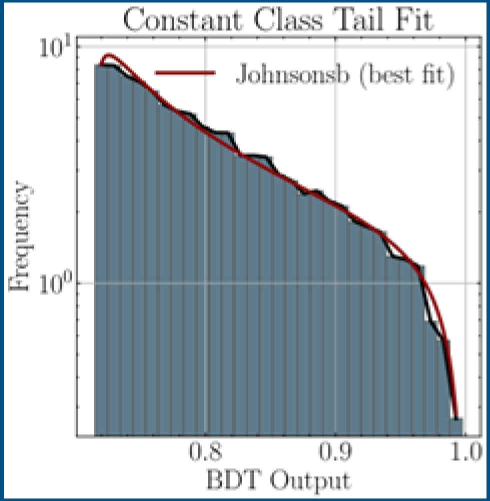


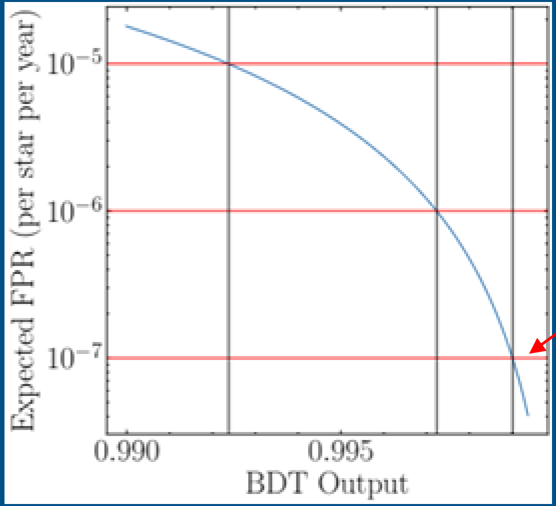
By extrapolation, a cut on the BIC ratio > 3.28 gives an expected FPR of 10^{-7}

Extrapolating to Lower FPR

BDT Output

- BDT output is bounded between 0 and 1, but can fit with a bounded distribution
- Zooming into tail composed of top 1% highest values of BDT, find the Johnson S_B to be a good fit using `distfit`:





A cut on BDT output > 0.999 gives an expected FPR of 10^{-7}

Effective ML Detection Efficiency

- Can define an effective efficiency per binned t_E as $\varepsilon = N_{>cuts}/N_{total}$
- We find an analytic function to fit the efficiency (solid line):

$$\epsilon = A\left(\frac{t_{\mathrm{E}}}{\mathrm{days}}\right)^{k} e^{\lambda(-t_{\mathrm{E}}/\mathrm{days})} + c$$

- This doesn't asymptote to 1 for large t_E as some other fits suggested in the literature do (dashed line)
- We use this efficiency and foreground estimates to calculate the number of expected point-like microlensing events and hence put constraints on PBHs