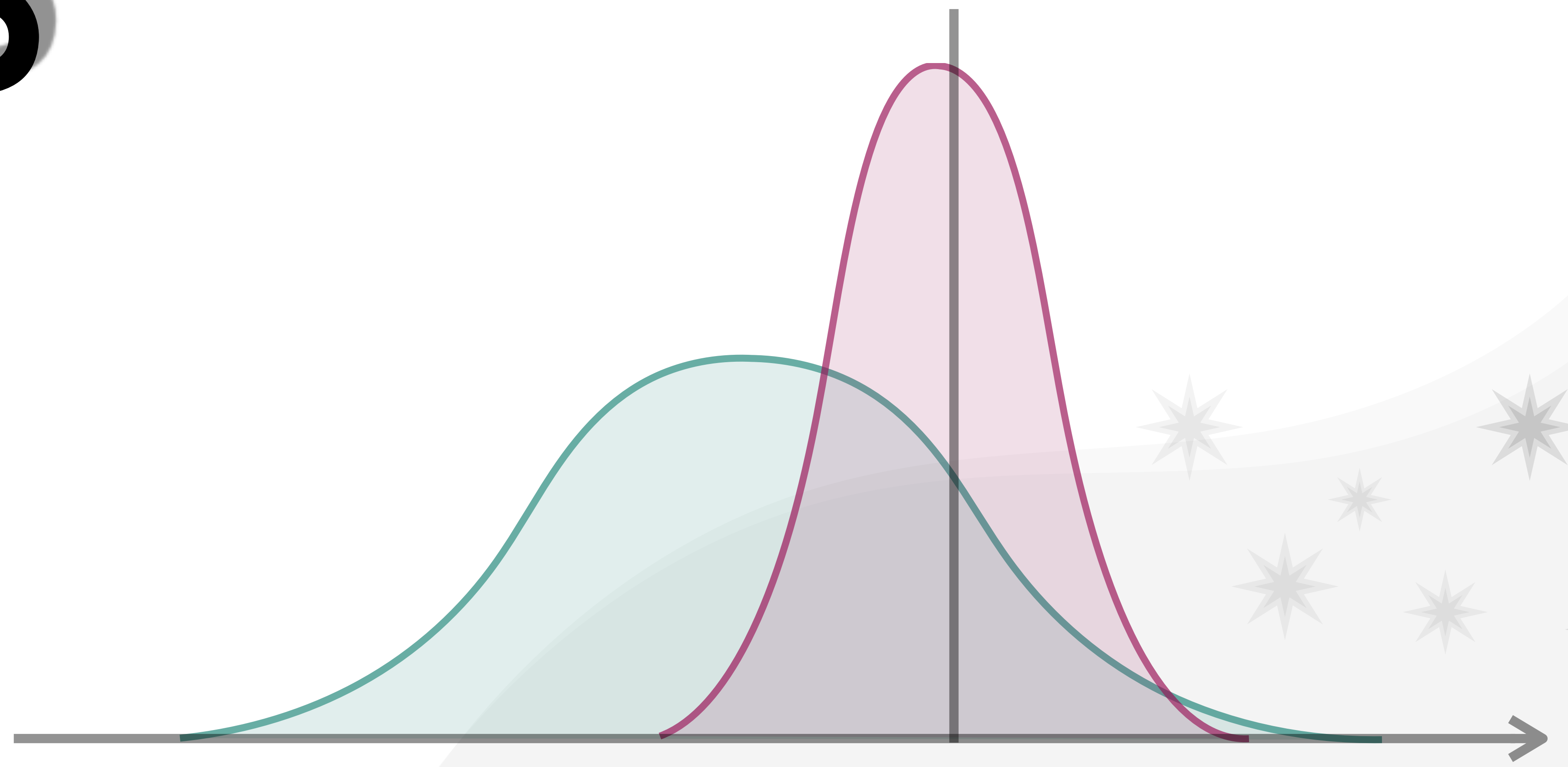
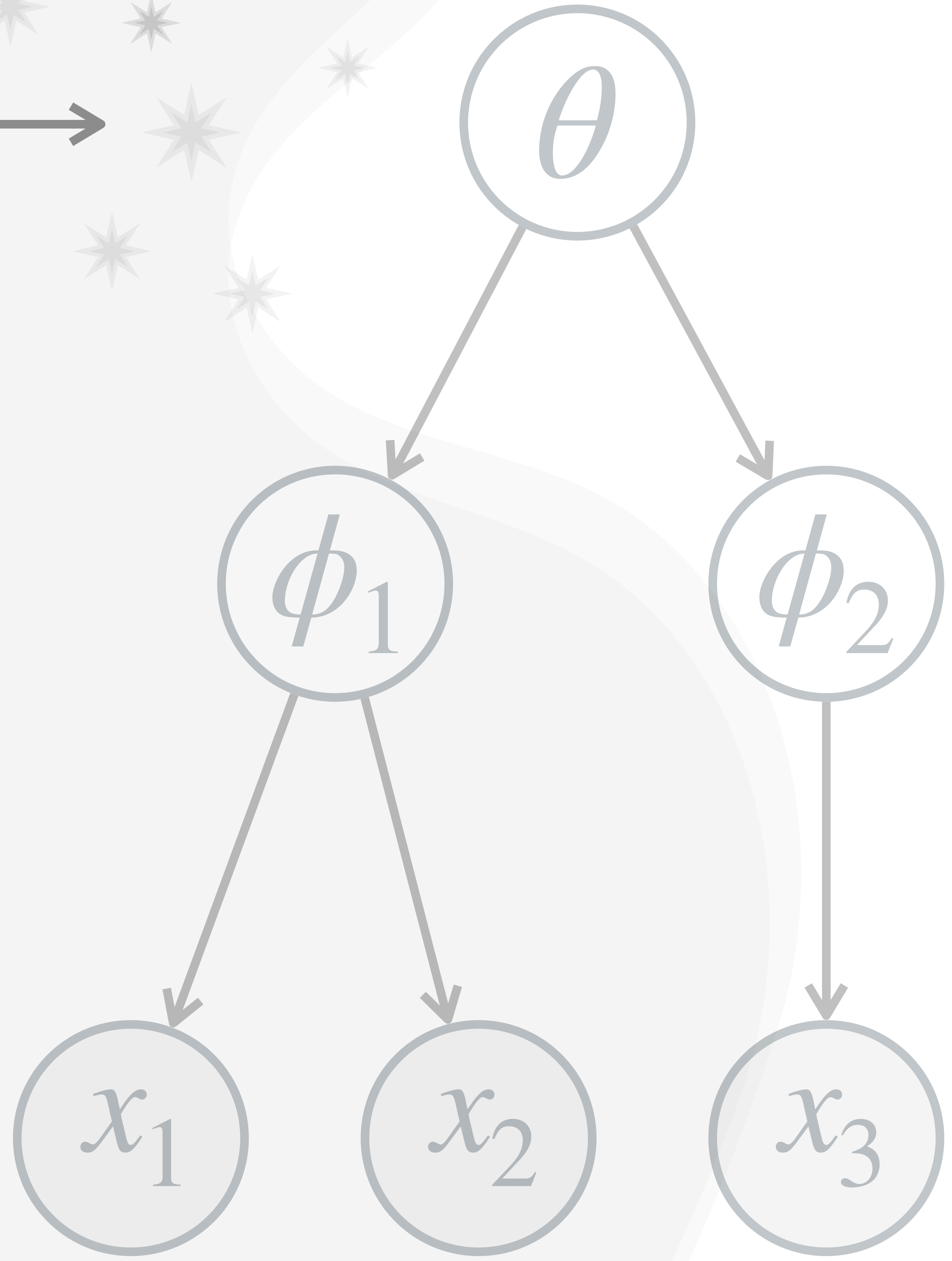
BAYESIAN APPROACHES IN ASTROPHYSICS





Francesca Capel

Max Planck Institute for Physics, Munich, Germany

Systematic Effects and Nuisance Parameters in Particle Physics Data Analyses

Banff International Research Station, 23rd-28th April 2023

ODSL ORIGINS Data Science Lab

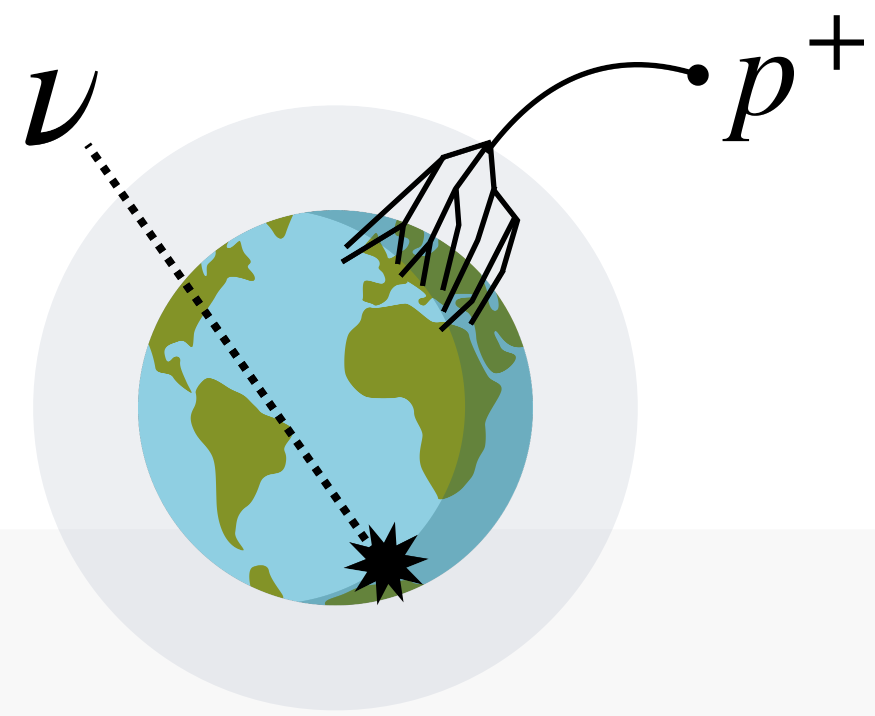
SFB 1258 Neutrinos Dark Matter Messengers

MAX-PLANCK-GESELLSCHAFT

Max-Planck-Institut für Physik (Werner-Heisenberg-Institut)

INTRODUCTION

My background:



Cosmic ray and neutrino astrophysics

Detector development & data analysis

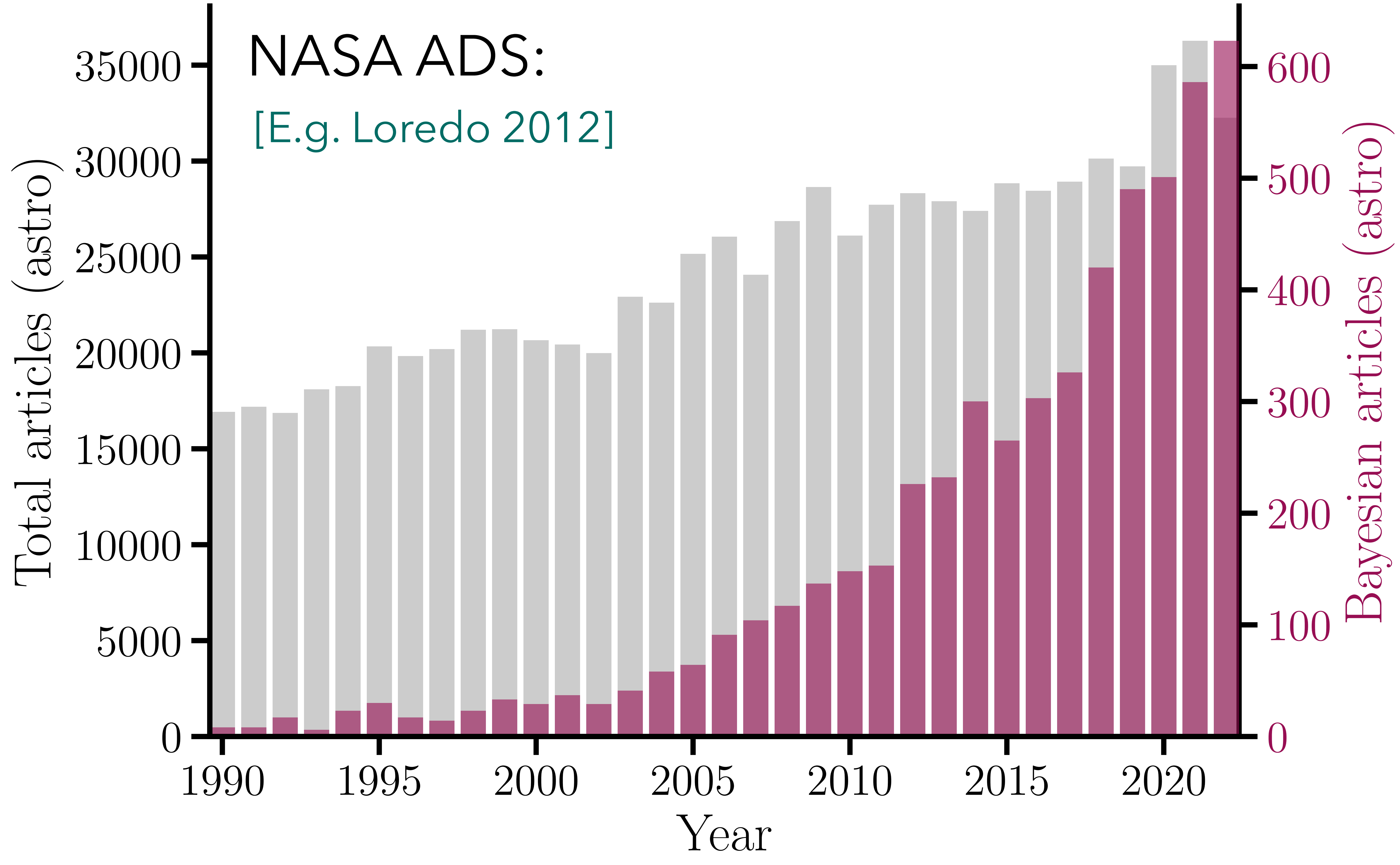
Identifying astrophysical sources

Astroparticle: More frequentist than Bayesian

INTRODUCTION

Bayesian approaches widely used in astrophysics and cosmology

[Trotta *Contemp. Phys* 2008; Hobson+ 2010; Loredo 2012; Schafer *Annu. Rev. Stat. Appl.* 2015; Feigelson+ *Annu. Rev. Stat. Appl.* 2021]



Driving factors:

- Observational - lack of controlled repeated experiments
- Computational developments
- Increasing complexity
- Lots of data - systematics become increasingly important

OUTLINE

1 Bayesian view on systematic uncertainties

2 General strategies for handling systematic uncertainties

3 Summary & outlook

Examples from astrophysics and cosmology

BAYESIAN SYSTEMATICS

Nuisance parameters are just parameters

Marginalisation summarises the result:

[Dawid 1980; Sinervo, *PHYSTAT* 2003; Sivia+Skilling 2006; Heinrich+Lyons *Annu. Rev. Nucl. Part. Sci.* 2007; Gelman+ BDA 2013]

Nuisance

Prior quantifies uncertainty

Interesting

$$\pi(\theta, \phi \,|\, x) = \frac{\pi(x \,|\, \theta, \phi)\pi(\theta, \phi)}{\int d\theta \int d\phi \;\pi(x \,|\, \theta, \phi)\pi(\theta, \phi)}$$

$$\pi(\theta \,|\, x) = \int d\phi \;\pi(\theta, \phi \,|\, x)$$

Credible regions,
limits etc.

[Talks by Cousins & Davison]

Comparison with frequentist

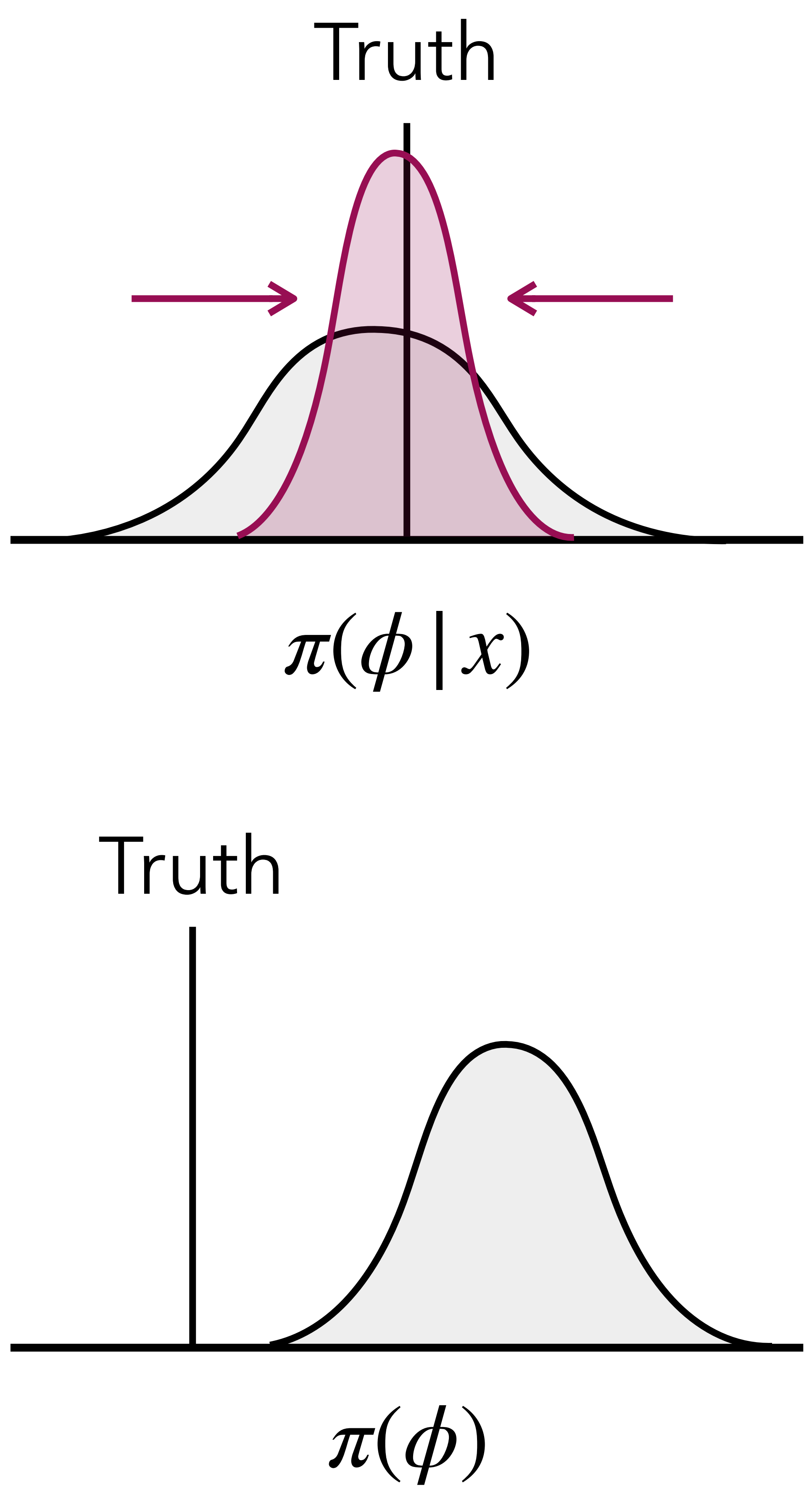
- Clear rationale
- No "special treatment" of nuisance parameters
- Priors necessary
- Coverage not guaranteed

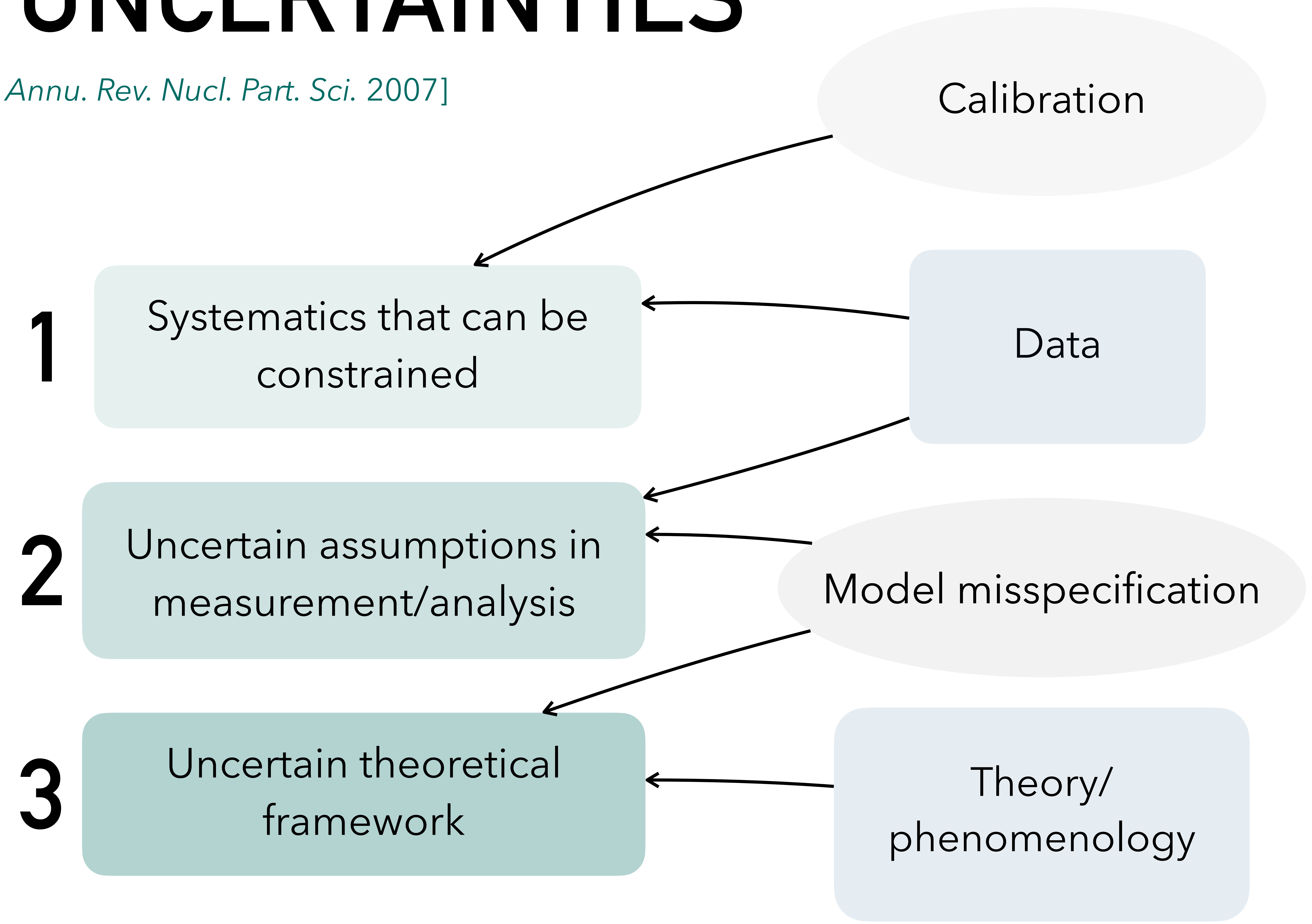
[Loredo *SCMA* 1992; Conrad *Astropart. Phys.* 2015]

SYSTEMATIC UNCERTAINTIES

[Sinervo *PHYSTAT* 2003; Heinrich+Lyons *Annu. Rev. Nucl. Part. Sci.* 2007]

[Talks by Wardle & Algeri]





SYSTEMATIC UNCERTAINTIES

1 | Systematics that can be constrained | E.g. Calibration of detector, characterisation of the effective area, ancillary measurements

"On/off" signal measurement for low count data

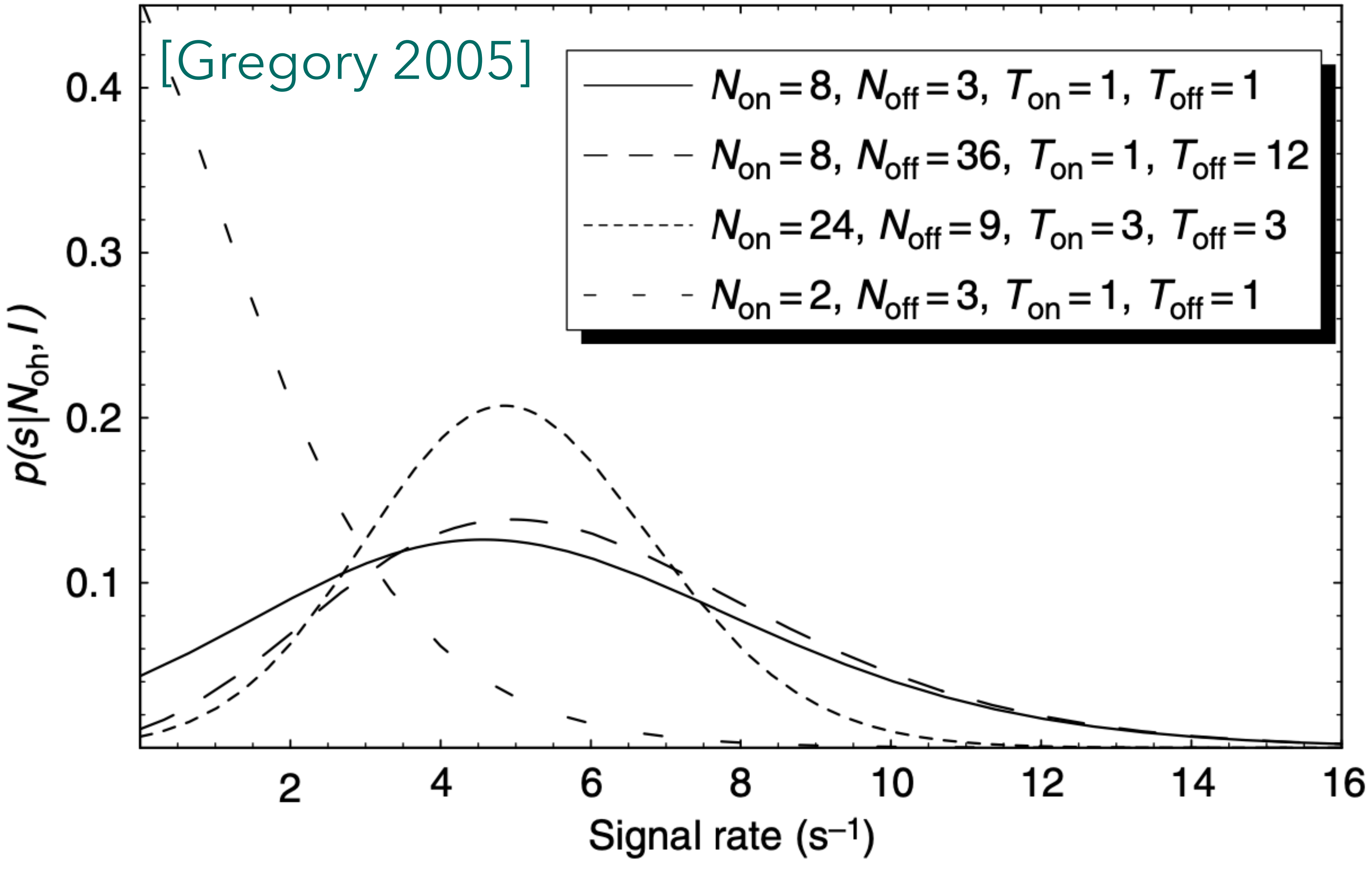
[Helene *NIM* 1983+1984; Prosper *NIMA* 1985; Prosper *PRD* 1988; Loredo *SCMA* 1992; Gregory 2005; Gillessen+Harney *A&A*. 2005; Knoetig *ApJ* 2014; Casadei *ApJ* 2015; Nosek+Nosková *NIMA* 2016]

[Talk by Cousins]

Astrophysics application: **Gamma-ray data**

- Off: Point away from source and measure b
- On: Point at source and measure s + b
- Marginalise to study source signal
- Quantify significance

$$\pi(s \,|\, N_{\rm on}) = \int_0^{b_{\rm max}} {\rm d}b \;\pi(s, b \,|\, N_{\rm on})$$



SYSTEMATIC UNCERTAINTIES

2 Uncertain assumptions in measurement/analysis

Not well-constrained by observations, assumptions and simplifications necessary

Magnetic field modelling in the search for ultra-high-energy cosmic ray (UHECR) sources

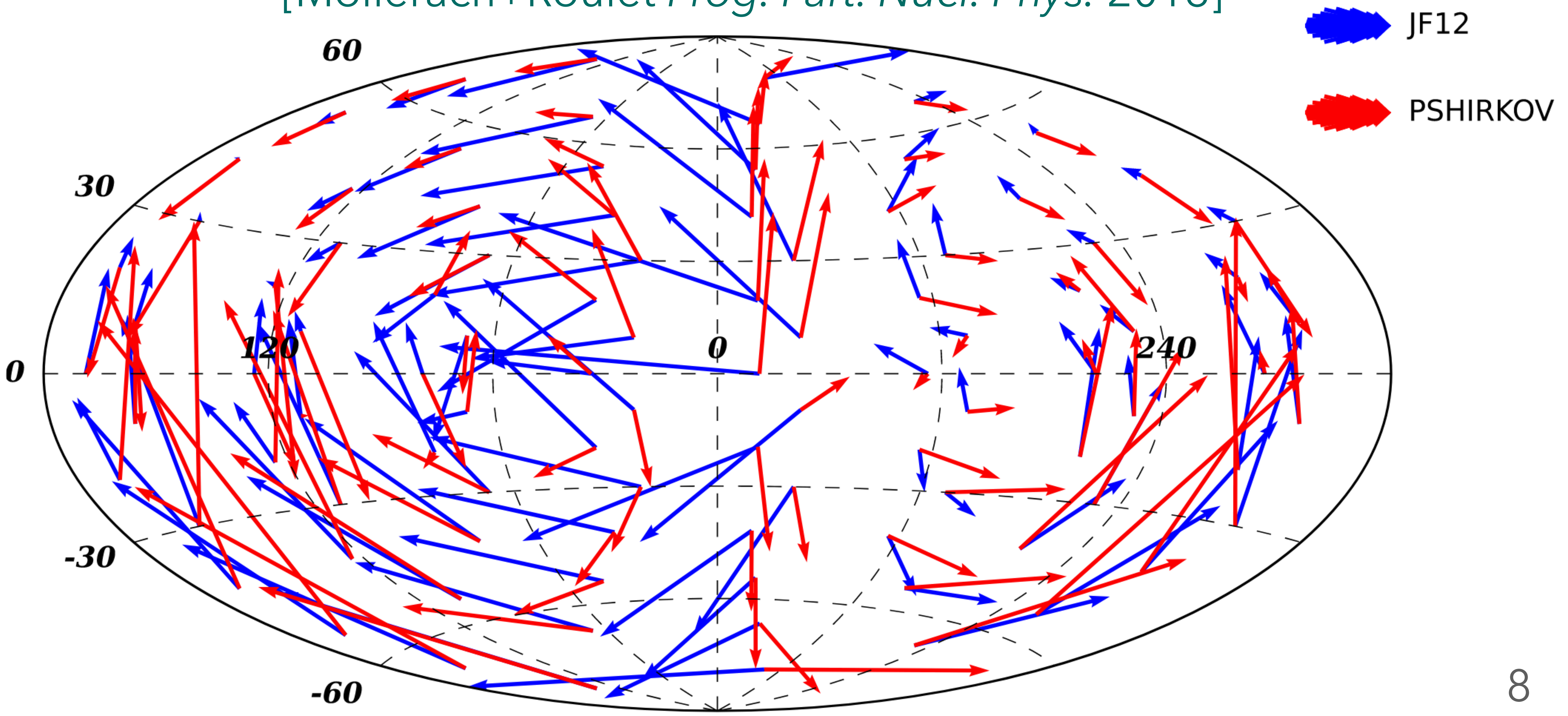
[Pschirkov+ *ApJ* 2011; Janson+Farrar *ApJ* 2012; Unger+Farrar *UHECR* 2018; Batista+ *Front. Astron. Space Sci*. 2019; Jaffe *Galaxies* 2019; Watanabe+ *UHECR* 2022]

- UHECRs are **deflected** by magnetic fields as they propagate from extragalactic sources
- **Galactic magnetic field** has a **lensing** effect that is model dependent

IMAGINE: Joint fit of all constraints

[Boulanger+ *JCAP* 2018]

[Mollerach+Roulet *Prog. Part. Nucl. Phys.* 2018]



SYSTEMATIC UNCERTAINTIES

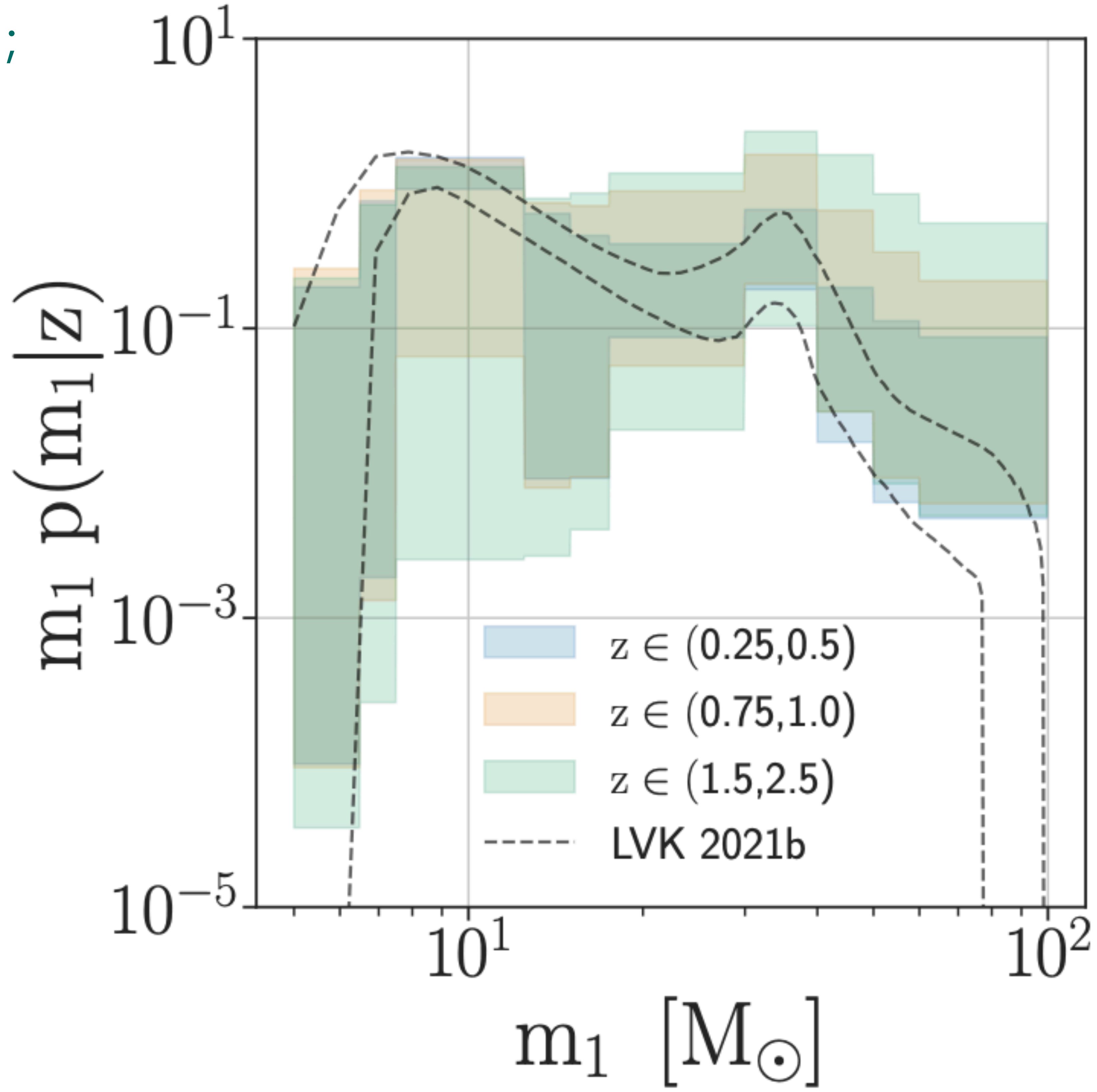
3 Uncertain theoretical framework

Empirical functional forms, choice of priors

Inference of binary black hole population properties from gravitational wave data

[Mandel+ *MNRAS* 2016; Abbott+ *ApJL* 2021; Tiwari+ *Class. Quantum Grav.* 2021; Callister+ *ApJL* 2022; Callister+Farr 2023; Ray+ 2023]

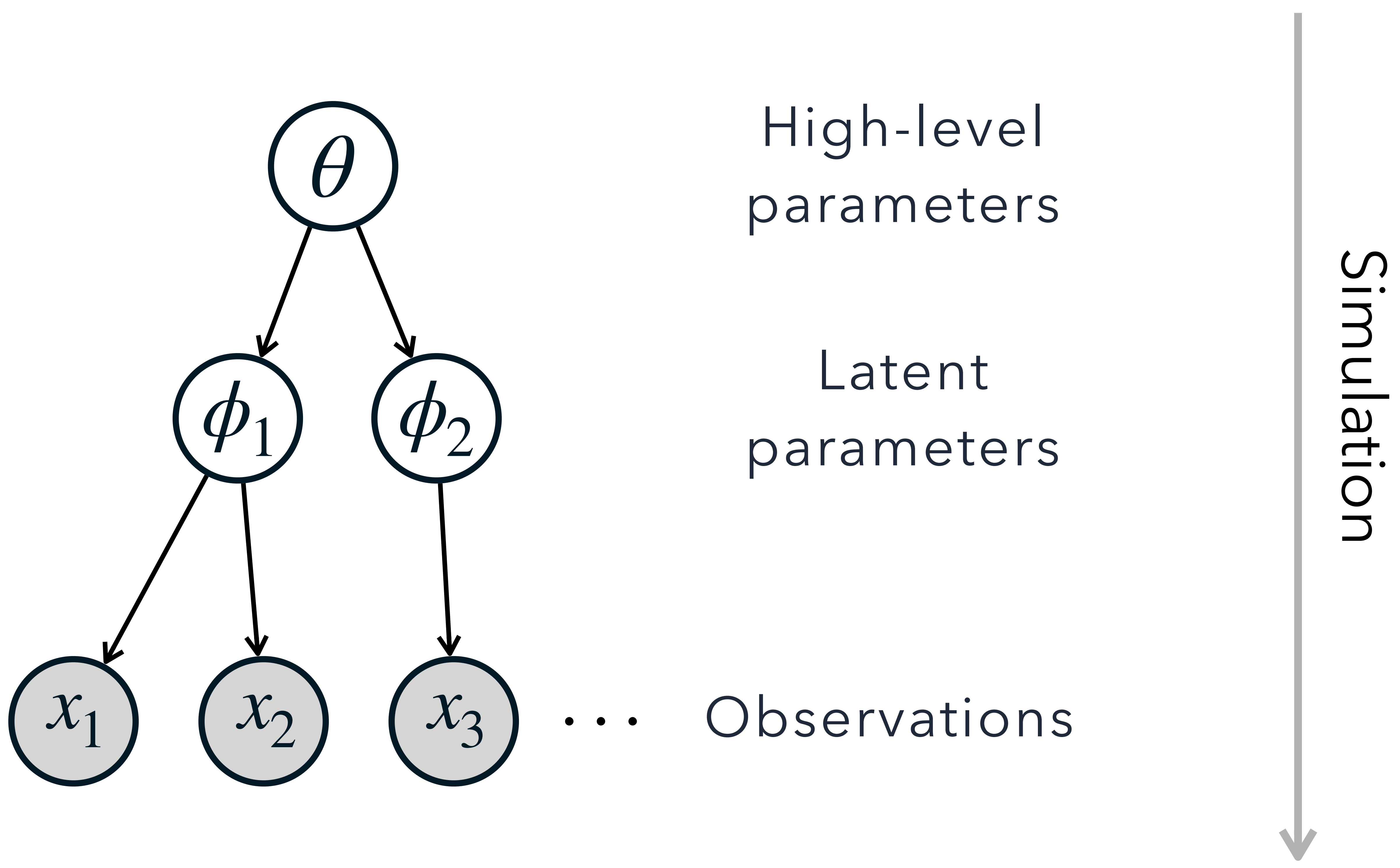
- Mass distribution and evolution informs **physics of black hole formation**
- Usual: power law + Gaussian
- **Alternative: Non-parametric**
- E.g. Binned **Gaussian process** to capture correlation between mass and redshift distributions [Ray+ 2023]



BAYESIAN STRATEGIES

Forward or generative modelling

- Identify known and expected systematics
- Map out complexity
- Contains all information needed to derive the likelihood function
- Prior predictive checks to evaluate prior choices
- Testbench to evaluate possible simplifications or assumptions

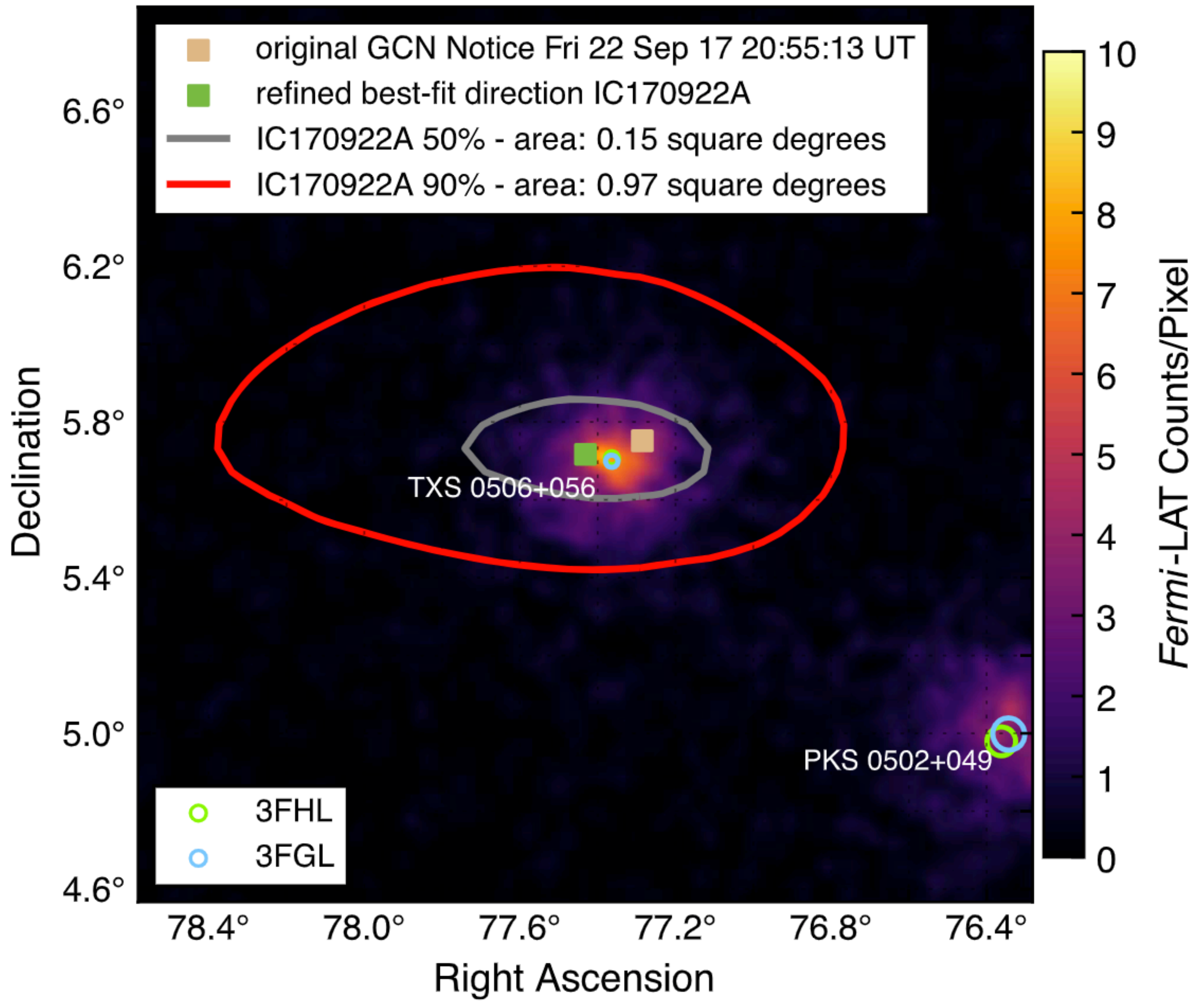


BAYESIAN STRATEGIES

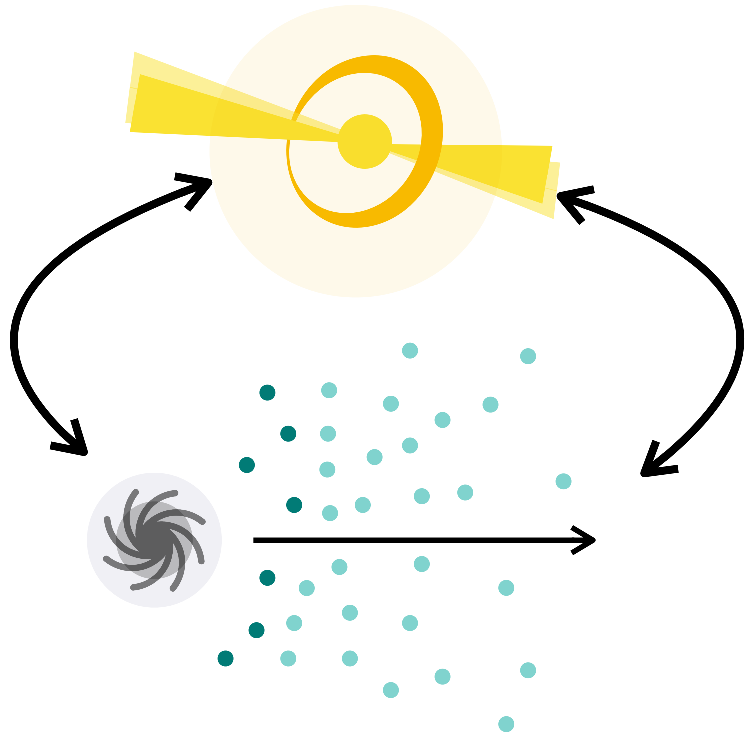
Forward or generative modelling

Example: Evaluating a possible blazar-neutrino connection from a population perspective

Individual association



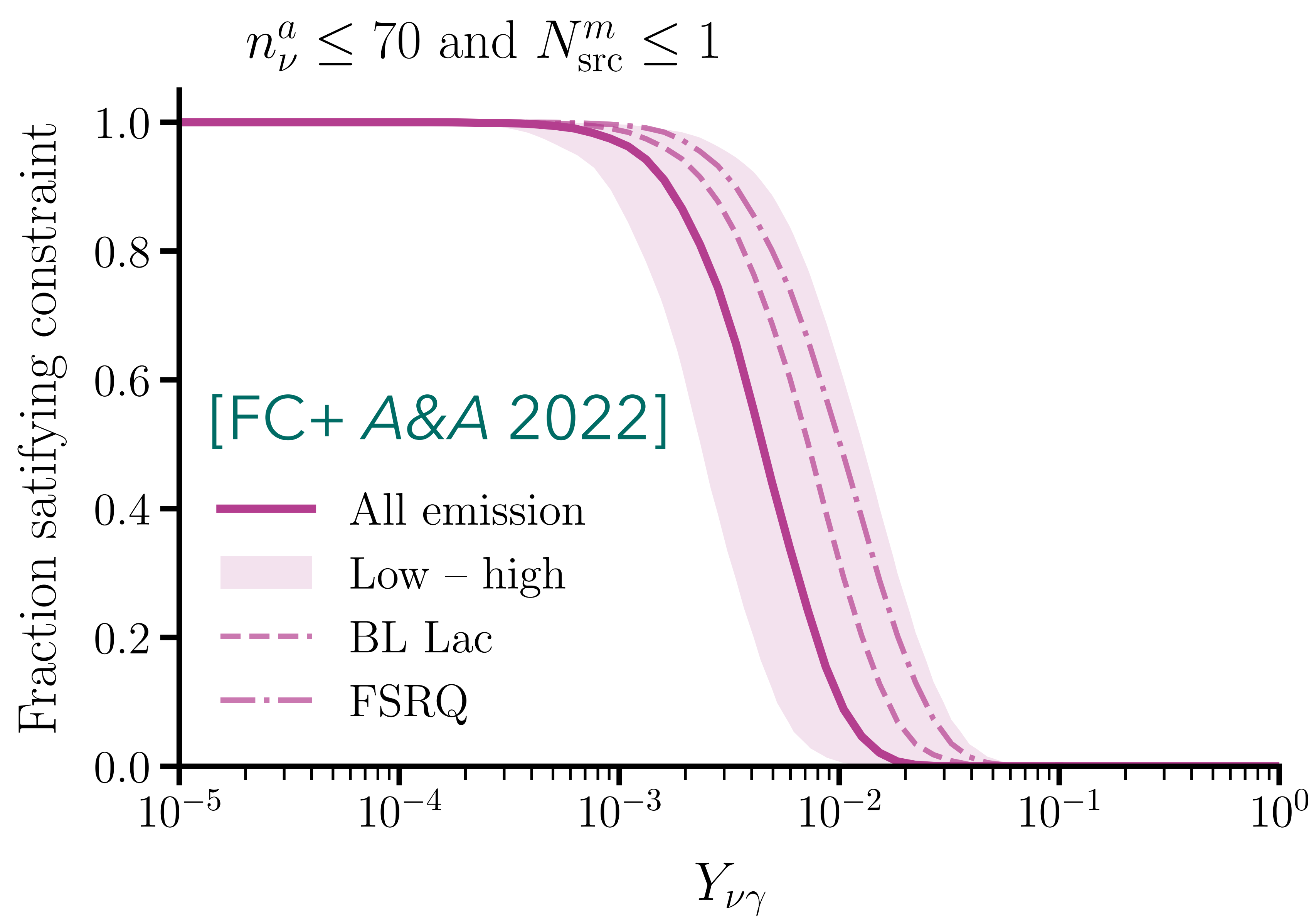
[Aartsen+ *Science* 2018]



Key assumption for ~3σ result: Proportionality between gamma-ray and neutrino fluxes

$$F^{\nu} = Y_{\nu\gamma} F^{\gamma}$$

What does this imply for the population?

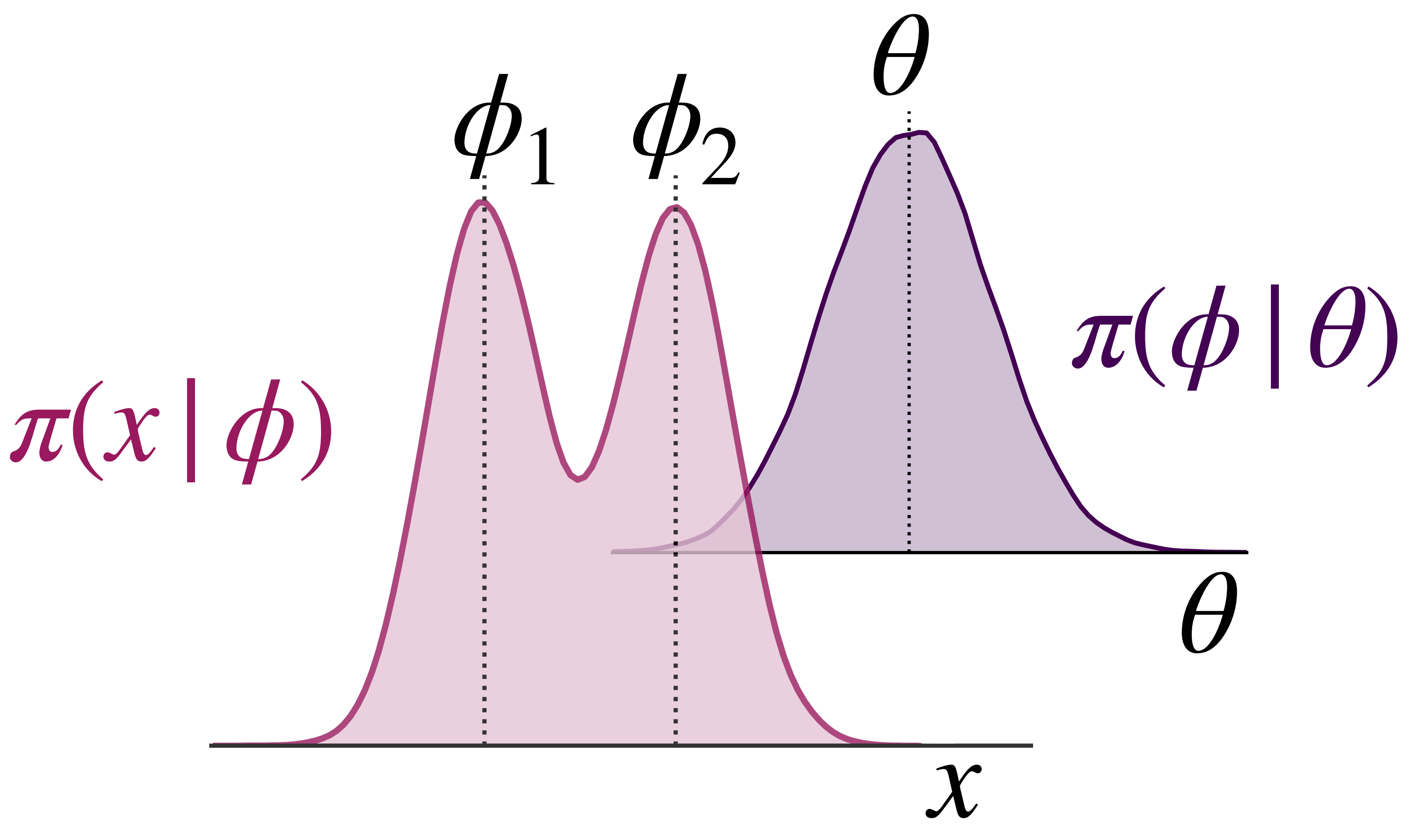


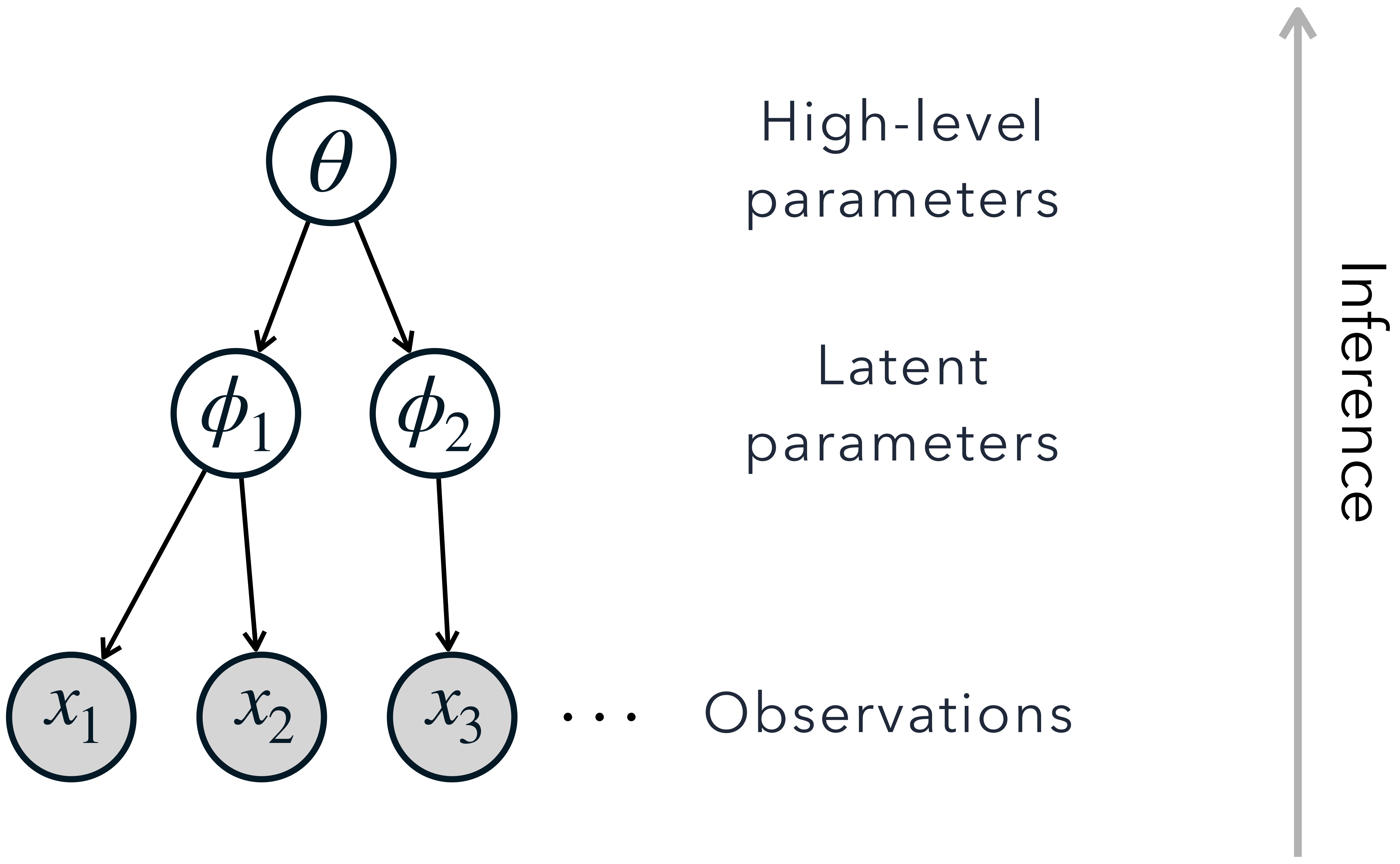
BAYESIAN STRATEGIES

Hierarchical modelling

Organise parameters into a hierarchy that describes the forward model

$$\mathscr{L}(x, \theta) = \pi(x \,|\, \phi)\ \pi(\phi \,|\, \theta) \ldots$$





Goal: "Good enough" model considering available constraints

BAYESIAN STRATEGIES

Hierarchical modelling

Widely applicable to a broad range of problems

[See Loredo *ACNA* 2013]

Fitting lines and correlations

[Kelly *ApJ* 2007; Andreon+Hurn *Stat. Data Min.* 2012]

Supernova cosmology and H_0 estimation

[Shariff+ *ApJ* 2016; Feeney+ *MNRAS* 2018; Hinton+ *ApJ* 2019]

Gravitational wave astronomy

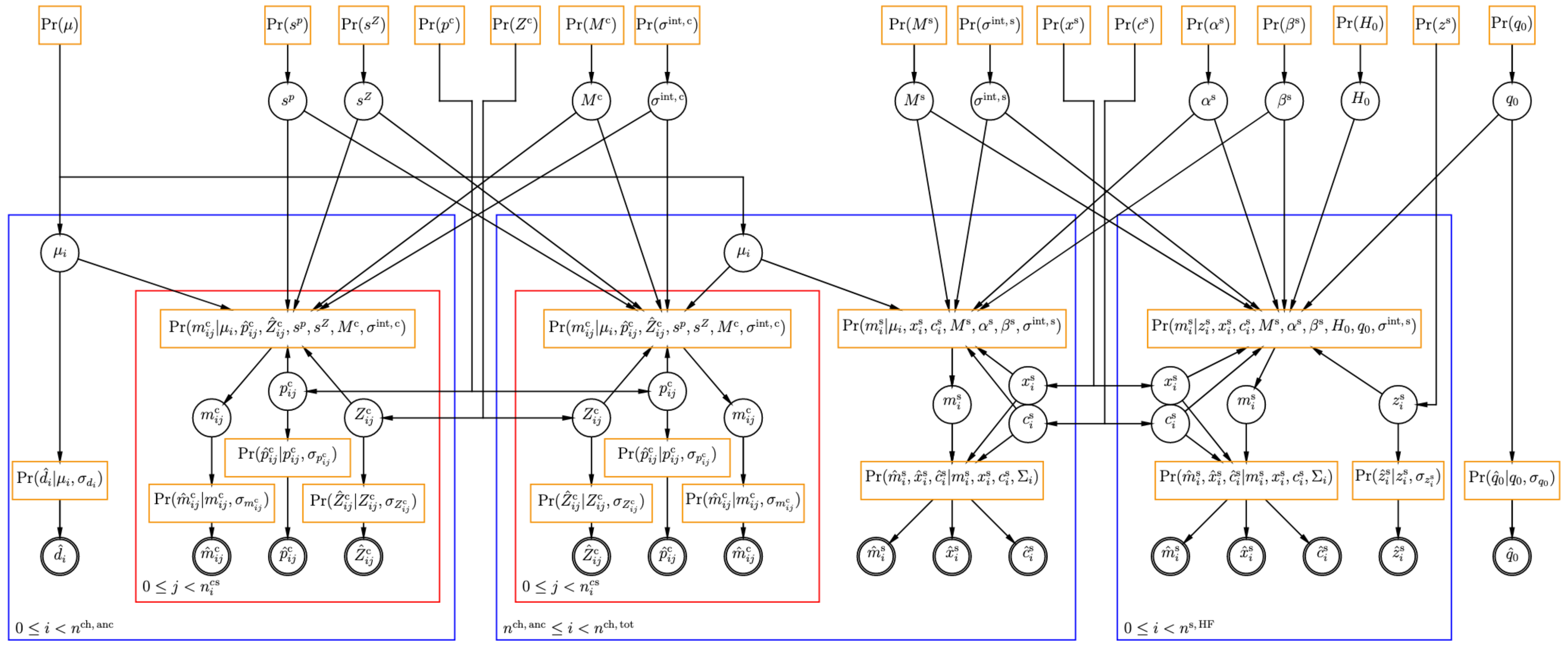
[Thrane+Talbot *PASA* 2019 + refs therein]

Photometric redshifts

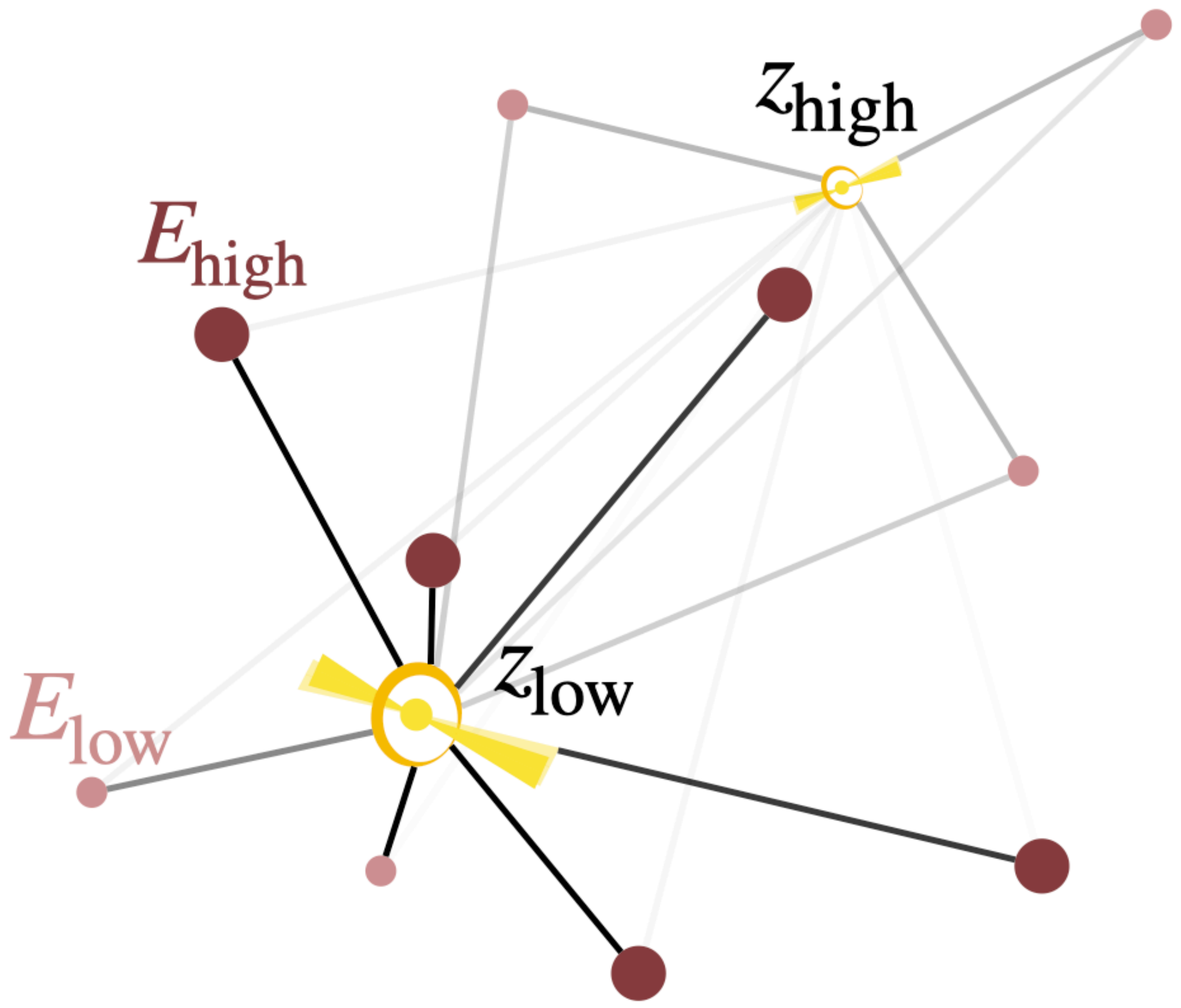
[Malz *PRD* 2021; Newman+Gruen *Annu. Rev. Astron. Astrophys.* 2022; Malz+Hogg *ApJ* 2022; Leistedt+ *ApJS* 2023]

Astrophysical association and cross-matching

[Soiaporn+ *Ann. Appl. Stat.* 2012; Budavári+Loredo *Annu. Rev. Stat. Appl.* 2015; FC+Mortlock *MNRAS* 2019]



[Feeney+ MNRAS 2018]



BAYESIAN STRATEGIES

Inference in high-dimensional parameter spaces

Markov chain Monte Carlo

Algorithm to numerically approximate **high-dimensional integrals** (e.g. expectation values, variances of parameters)

Exact convergence in the limit of infinite samples

$$\int_{\Theta} \mathrm{d}\theta \, f(\theta)p(\theta) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} f(\theta_n)$$

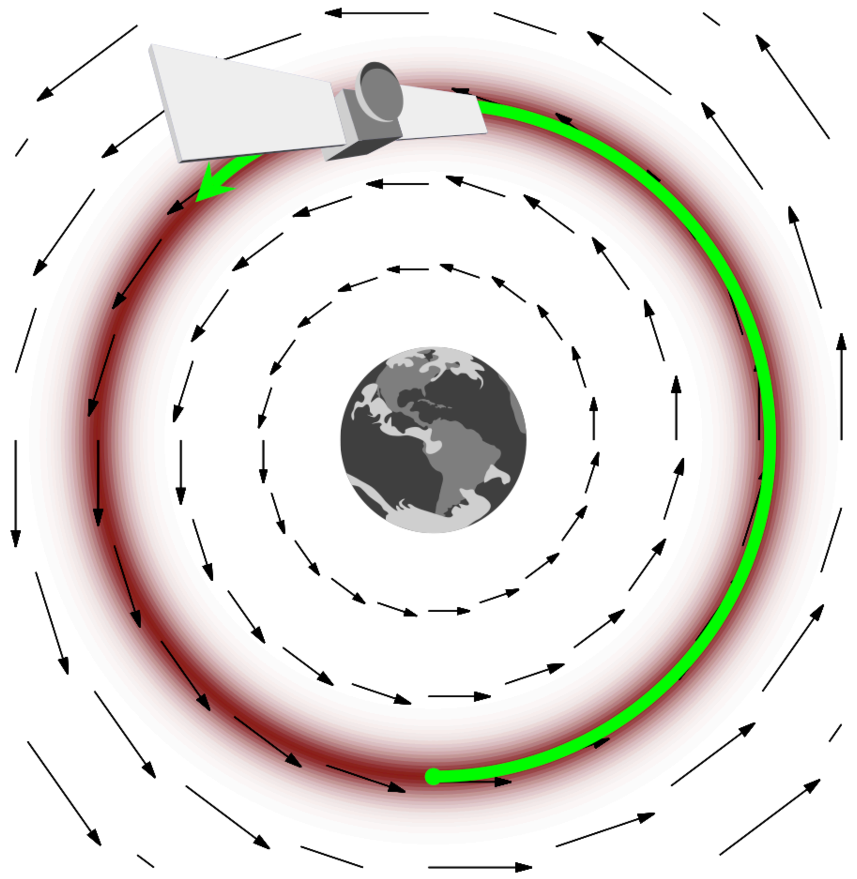
Large numbers of free parameters possible

Hamiltonian Monte Carlo

A type of Markov chain Monte Carlo that uses **Hamiltonian dynamics** to move **efficiently** through high-dimensional parameter spaces

$$\theta \longrightarrow (\theta, p) \qquad \frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{\partial H}{\partial p}$$

$$H(\theta, p) \equiv \log P(\theta, p) \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial H}{\partial \theta}$$



[Betancourt 2017]

Uncertainty quantification for free

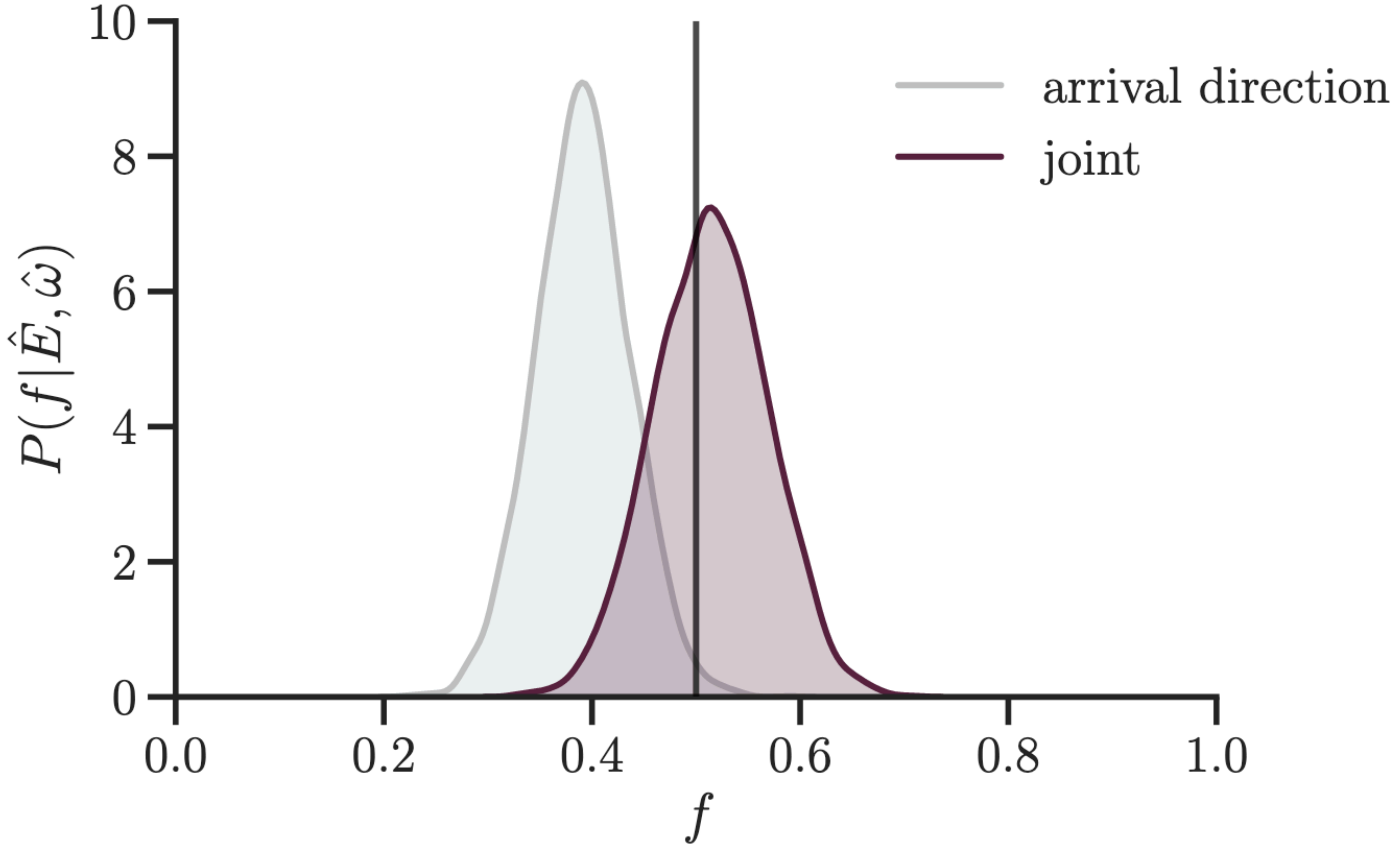
BAYESIAN STRATEGIES

Model checking: Fitting simulated data

- Consistency check, test for model expansion
- Analysis of fraction (f) of UHECRs from set of sources
- Data: detected energies (E) and arrival directions (ω)

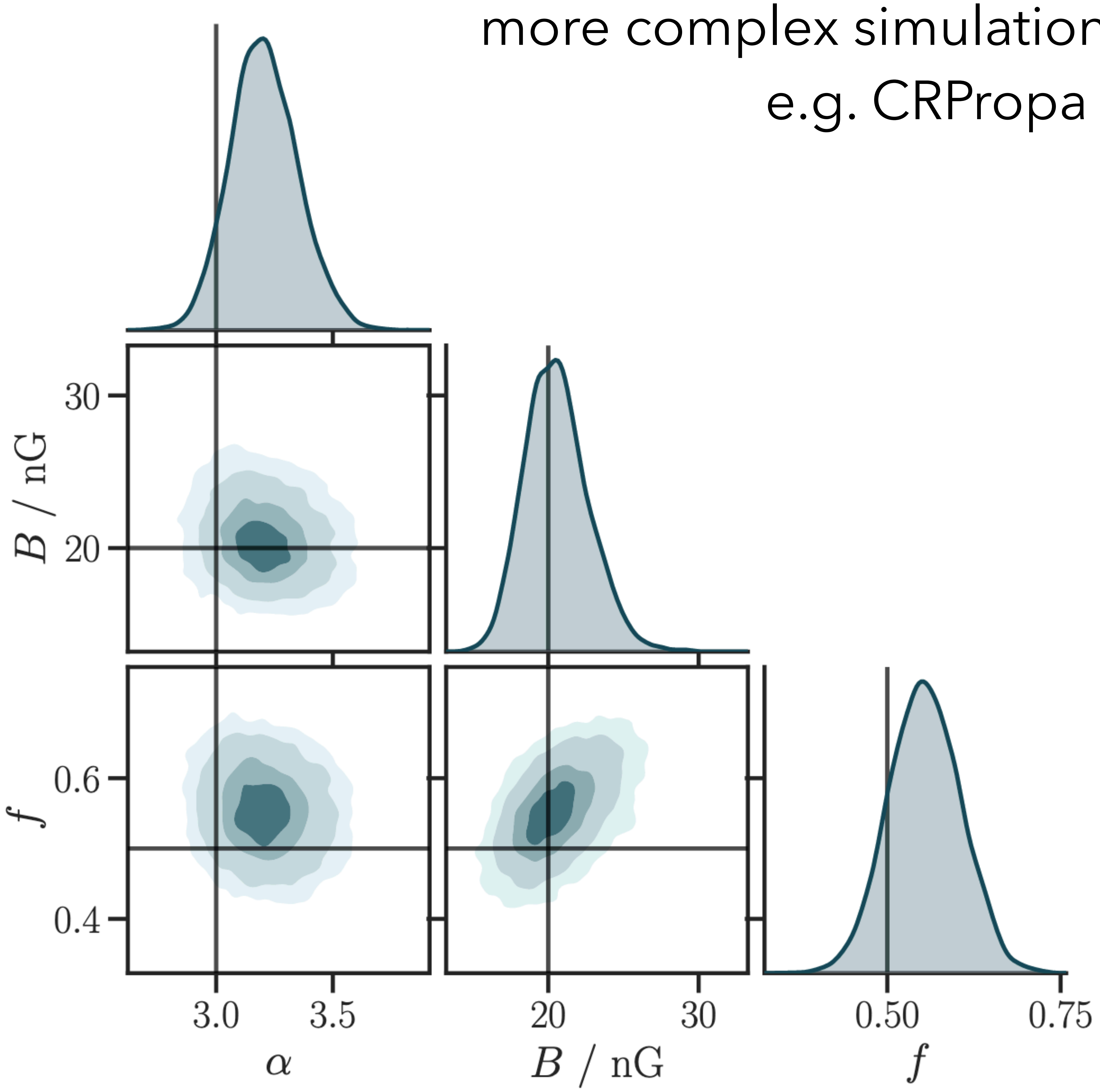
Examples from [FC+Mortlock MNRAS 2019]

Truth



Bias if energies left out of the fit

Key parameters recovered when fitting more complex simulations e.g. CRPropa 3



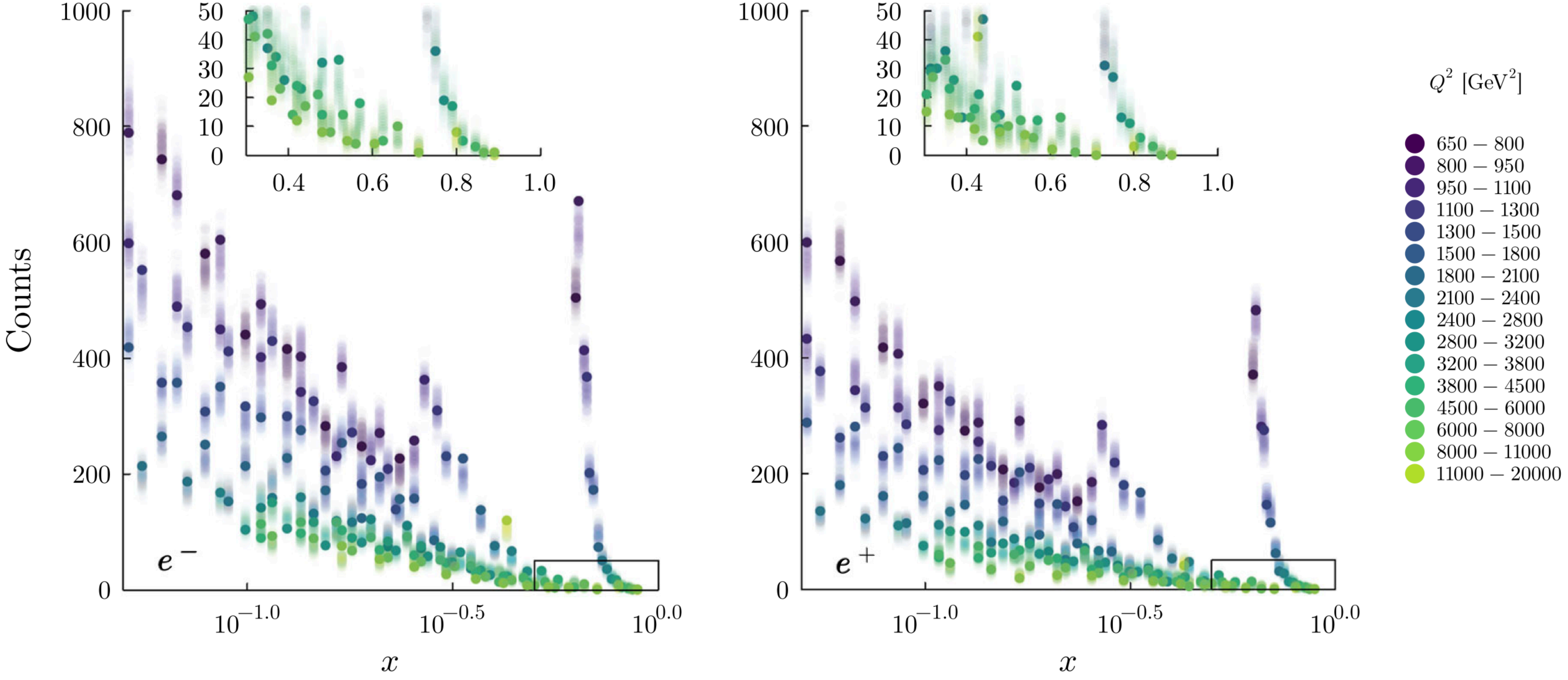
BAYESIAN STRATEGIES

Model checking: Posterior predictive checks

Example from [Aggarwal+ PRL 2023]

- Goodness of fit, test for model expansion
- Parton density extraction from high-x ZEUS data
- Data: Counts in x–Q² bins

$$\pi(n_{\mathrm{rep}} \,|\, n) = \int \mathrm{d}\Theta \;\pi(n_{\mathrm{rep}} \,|\, \Theta)\pi(\Theta \,|\, n)$$



BAYESIAN STRATEGIES

Model averaging

[Hoeting Statist. Sci. 1999; Parkinson+Liddle *Stat. Data Min.* 2012, Vehtari+Ojanen *Stat. Surv.* 2012]

- Discrete models
- All possible models known (M-closed)
- Model uncertainty quantifiable
- Non-trivial to evaluate evidence via MCMC
- Nested sampling is widely used

[Buchner *Stat. Surv.* 2023]

$$\pi(\theta|x) = \frac{\sum_k \pi(\theta|x, \mathcal{M}_k)\pi(\mathcal{M}_k|x)}{\sum_k \pi(\mathcal{M}_k|x)}$$

Model prior

Evidence

Is the average model useful? [Talk by Wardle]

PRACTICAL ASPECTS

More parameters ⇒ increasingly computationally expensive

Implementation of strategies can be challenging in different scenarios
E.g. Visualisation of predictive checks, likelihood difficult/impossible to write down
[Talks by Lee & Shen]

Data may not be available in a way that allows for complete modelling
[Loredo *ACNA* 2013]

Overfitting possible if using the same data for model building and inference
[Betancourt 2020]

Limited protection against "Unknown unknowns"
[Hatfield 2022]

SUMMARY & OUTLOOK

Bayesian methods are **conceptually more straightforward**, but can still be challenging to implement (not just priors)

It is **worth it** to make the most of our data!

Some thoughts:

- Availability of reusable data and reproducible methods is key
- Complementarity with frequentist approaches - why not do both?