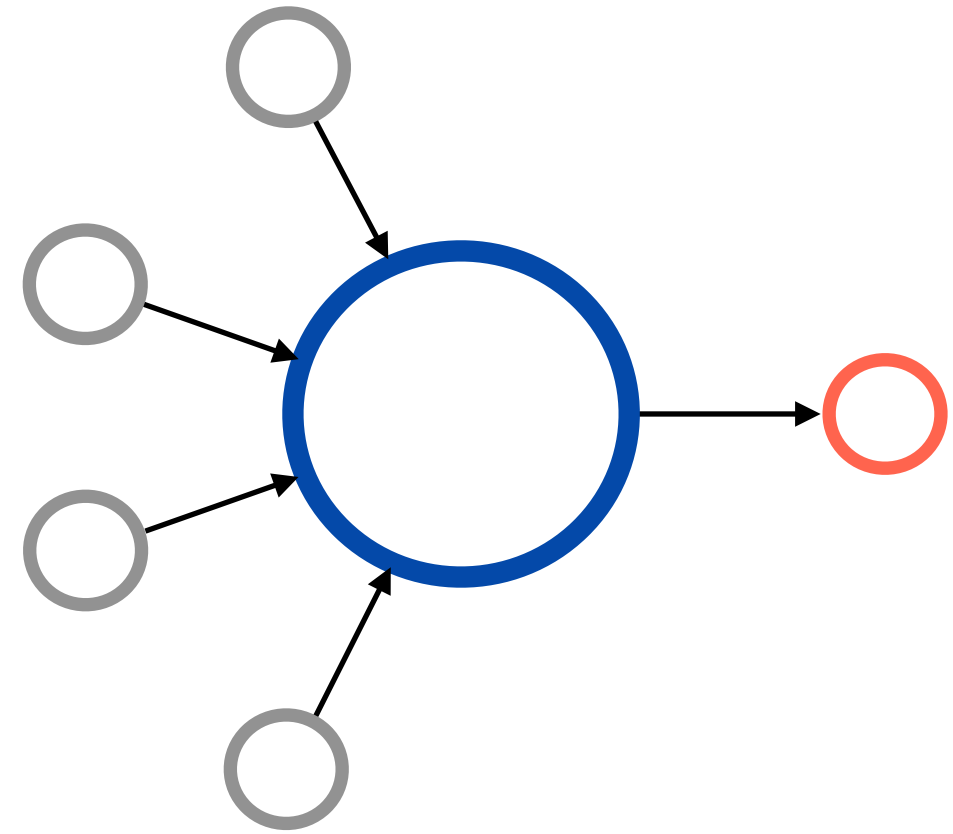
Hierarchical Neural SBI over Event Ensembles



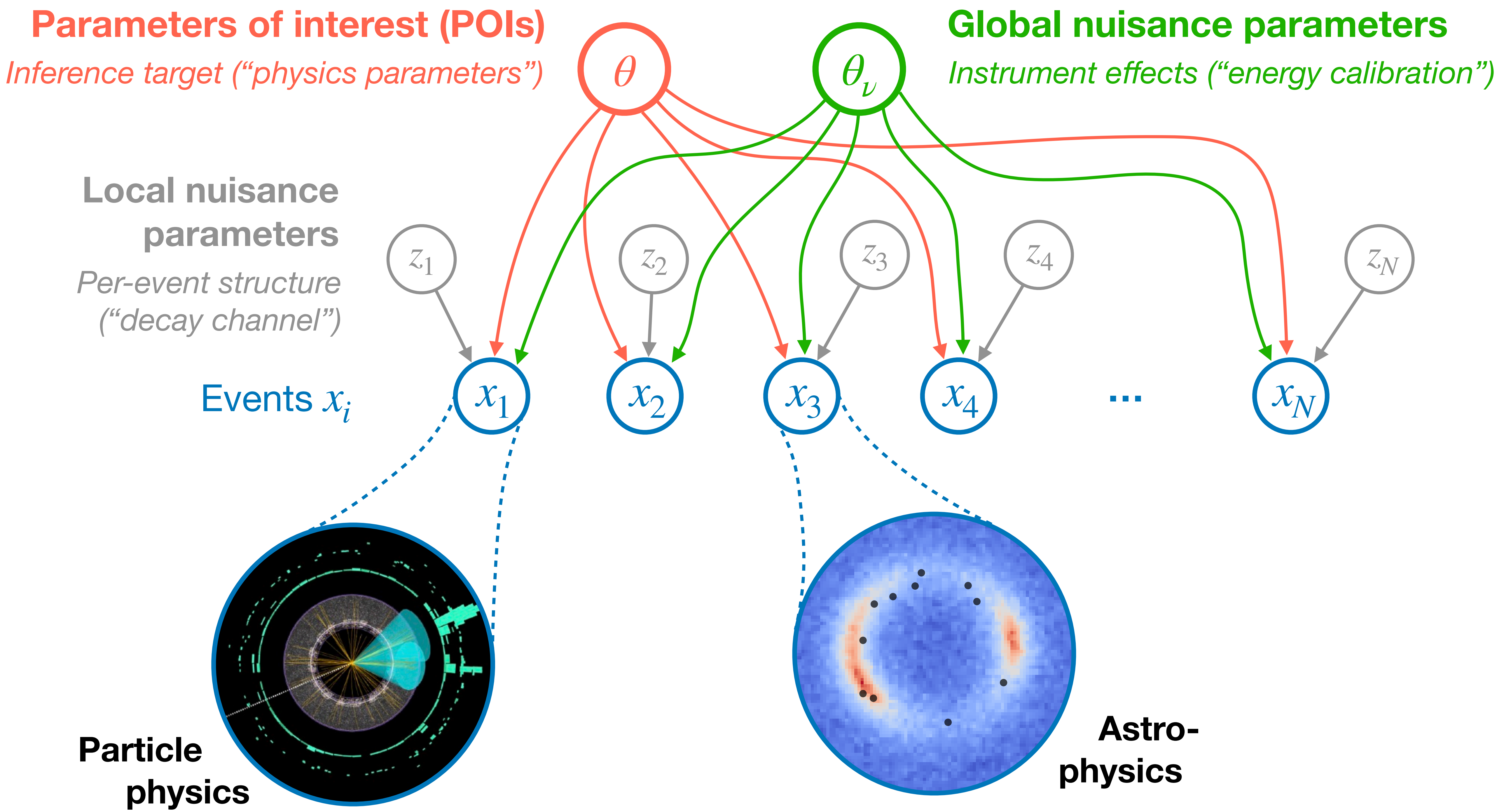
L. Heinrich (TU Munich), S. Mishra-Sharma (MIT),
C. Pollard (Warwick), P. Windischhofer (Chicago)

AI and the Uncertainty Challenge in Fundamental Physics, Paris, 28.11.2023

THE UNIVERSITY OF WARWICK
THE UNIVERSITY OF CHICAGO

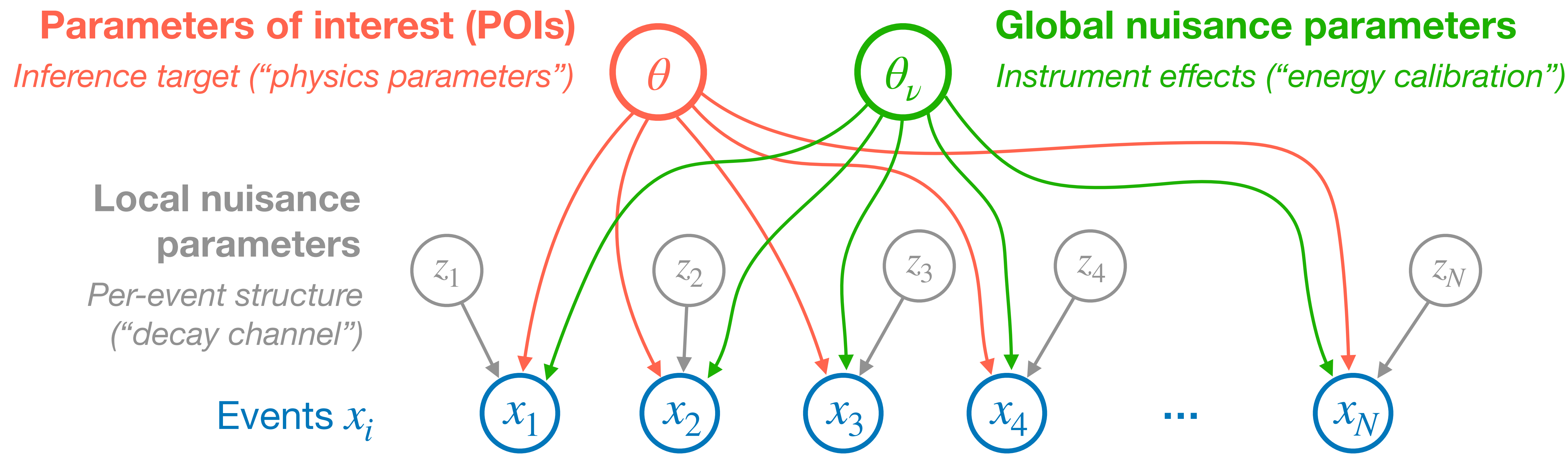
Scientific data analysis in a nutshell

Scientific data sets often have a hierarchical structure



Scientific data analysis in a nutshell

Scientific data sets often have a hierarchical structure



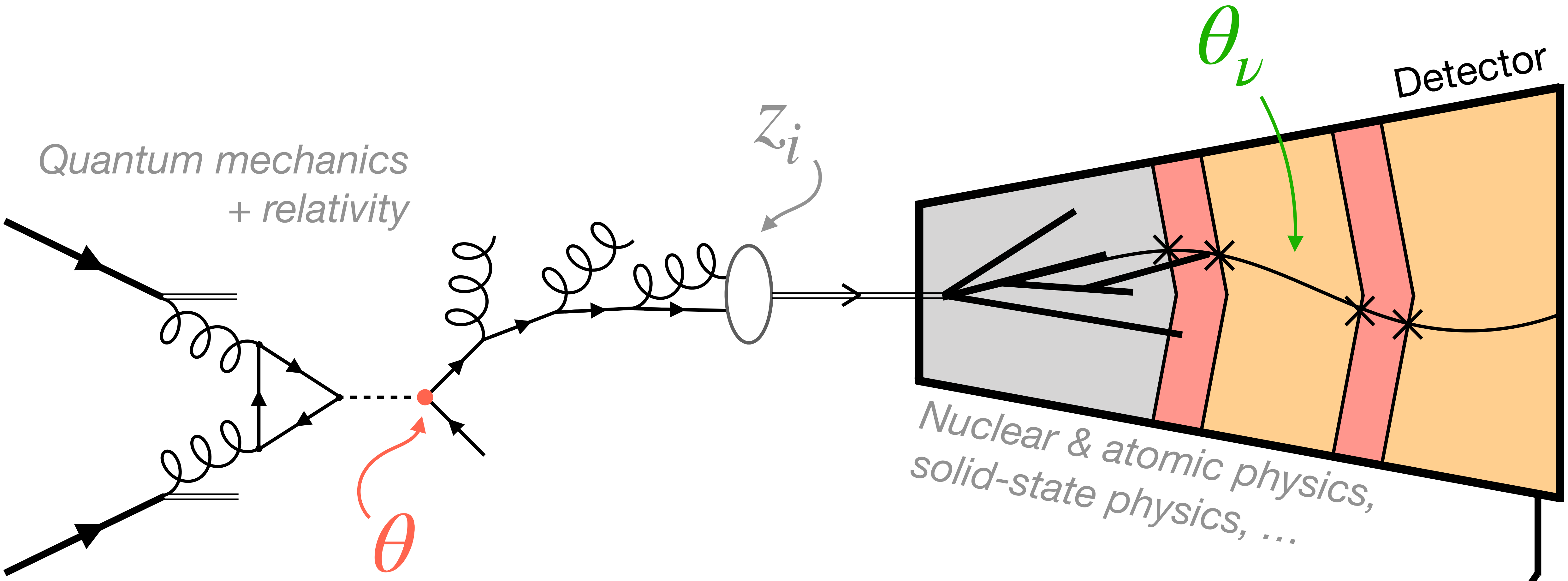
Dataset-wide likelihood:

$$p(\{x\}|\{z\}, \theta, \theta_\nu) = \sum_{N=0}^{\infty} p(N|\theta) \prod_{i=1}^{N} p(x_i|z_i, \theta, \theta_\nu)$$

Dataset cardinality
(e.g. Poisson rate)

Events are conditional IID

Simulation-driven science

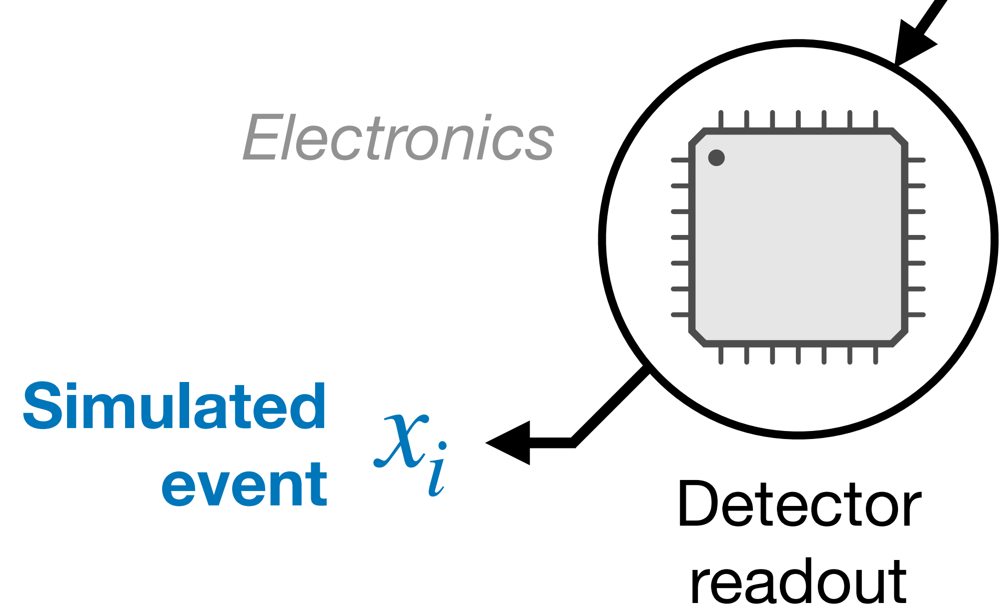


Particle physics:

High-fidelity simulations encode collective domain knowledge of the past decades

→ *Sample from* $p(\{x\}|\{z\}, \theta, \theta_\nu)$

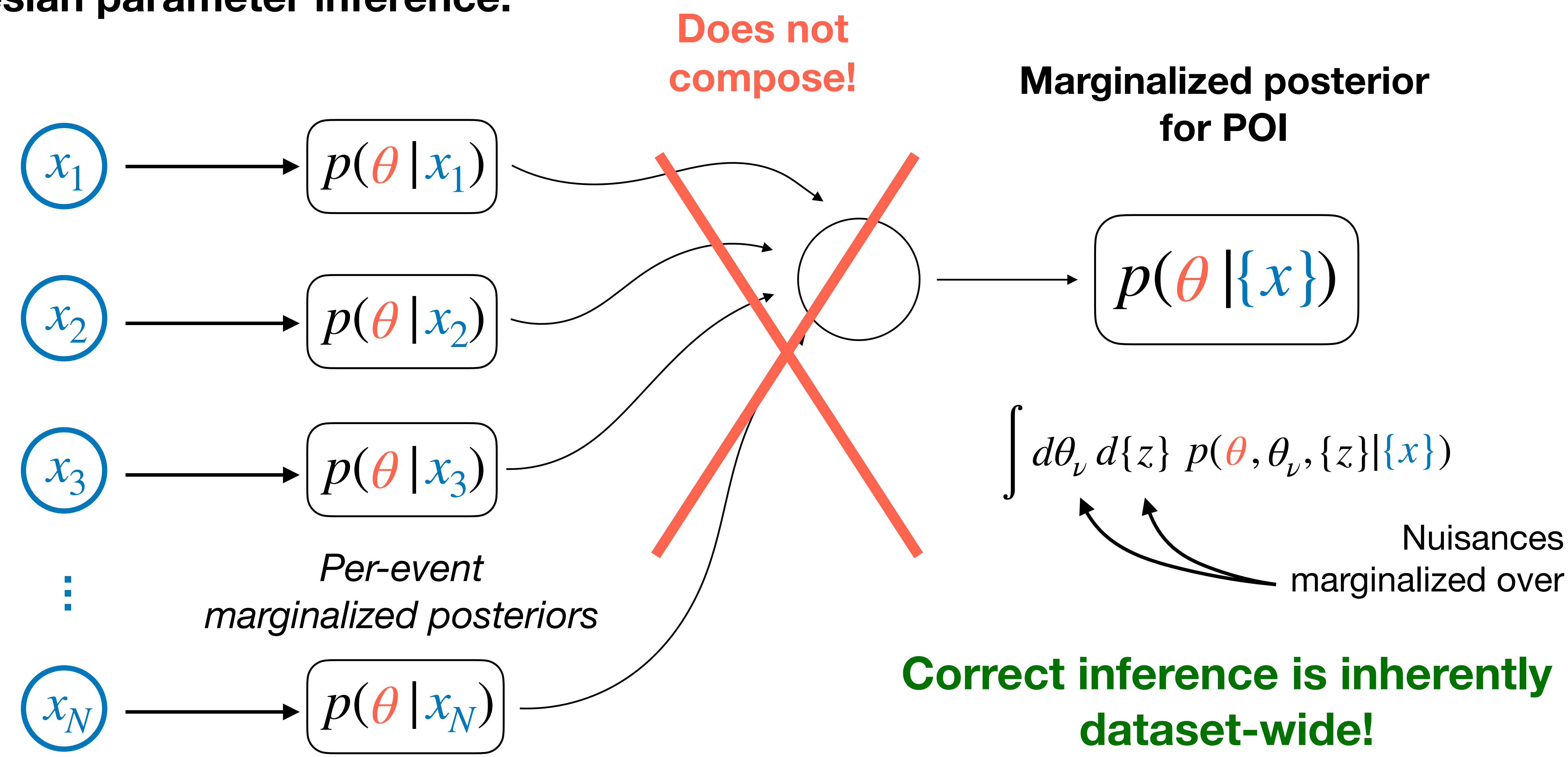
Simulation-based inference shines!



Parameter inference

$$p(\{x\}|\{z\},\theta,\theta_\nu) = \sum_{N=0}^{\infty} p(N|\theta) \prod_{i=1}^{N} p(x_i|z_i,\theta,\theta_\nu)$$

Bayesian parameter inference:



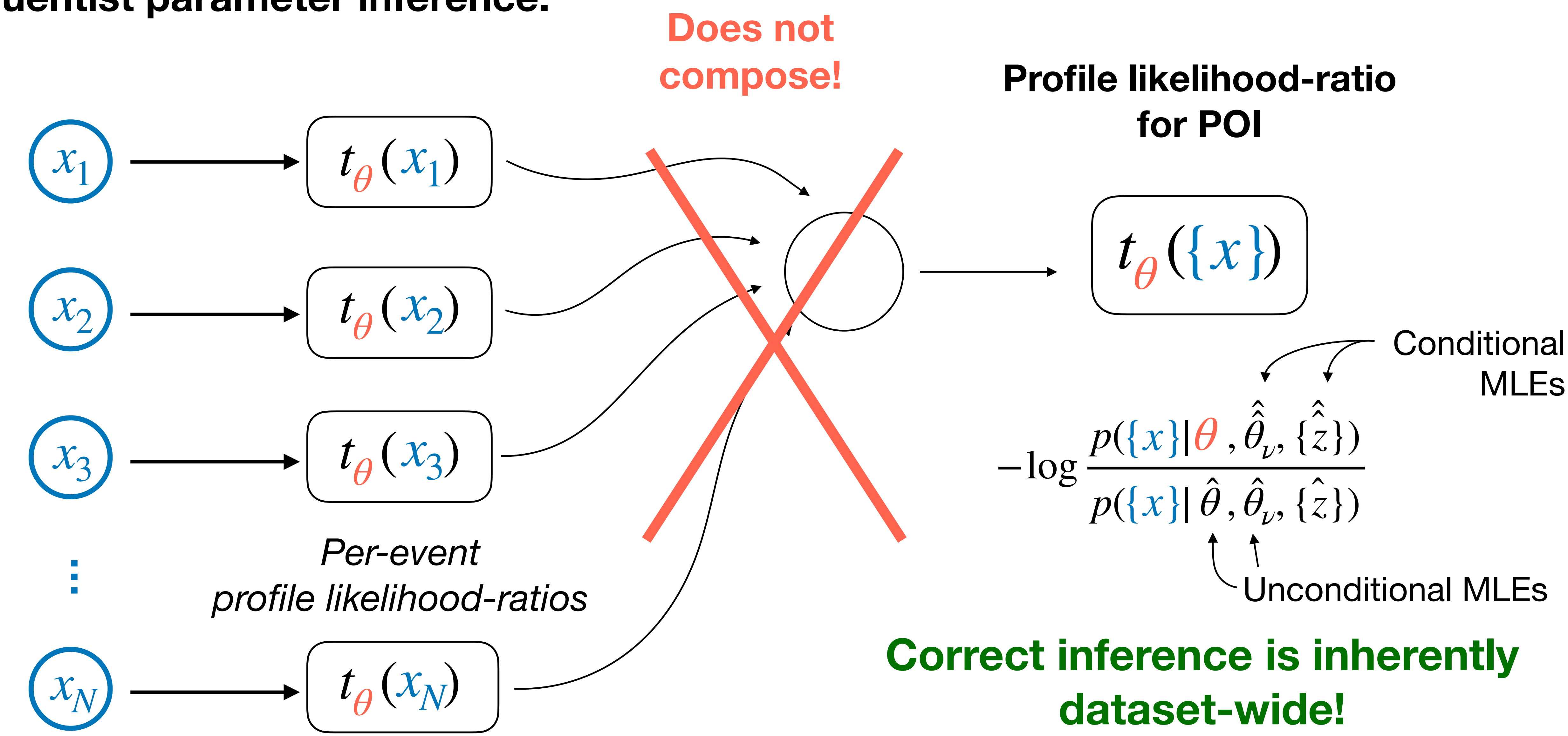
Correct inference is inherently dataset-wide!

(In the presence of global nuisance parameters)

Parameter inference

$$p(\{x\}|\{z\},\theta,\theta_\nu) = \sum_{N=0}^{\infty} p(N|\theta) \prod_{i=1}^{N} p(x_i|z_i,\theta,\theta_\nu)$$

Frequentist parameter inference:



Correct inference is inherently dataset-wide!

(In the presence of global nuisance parameters)

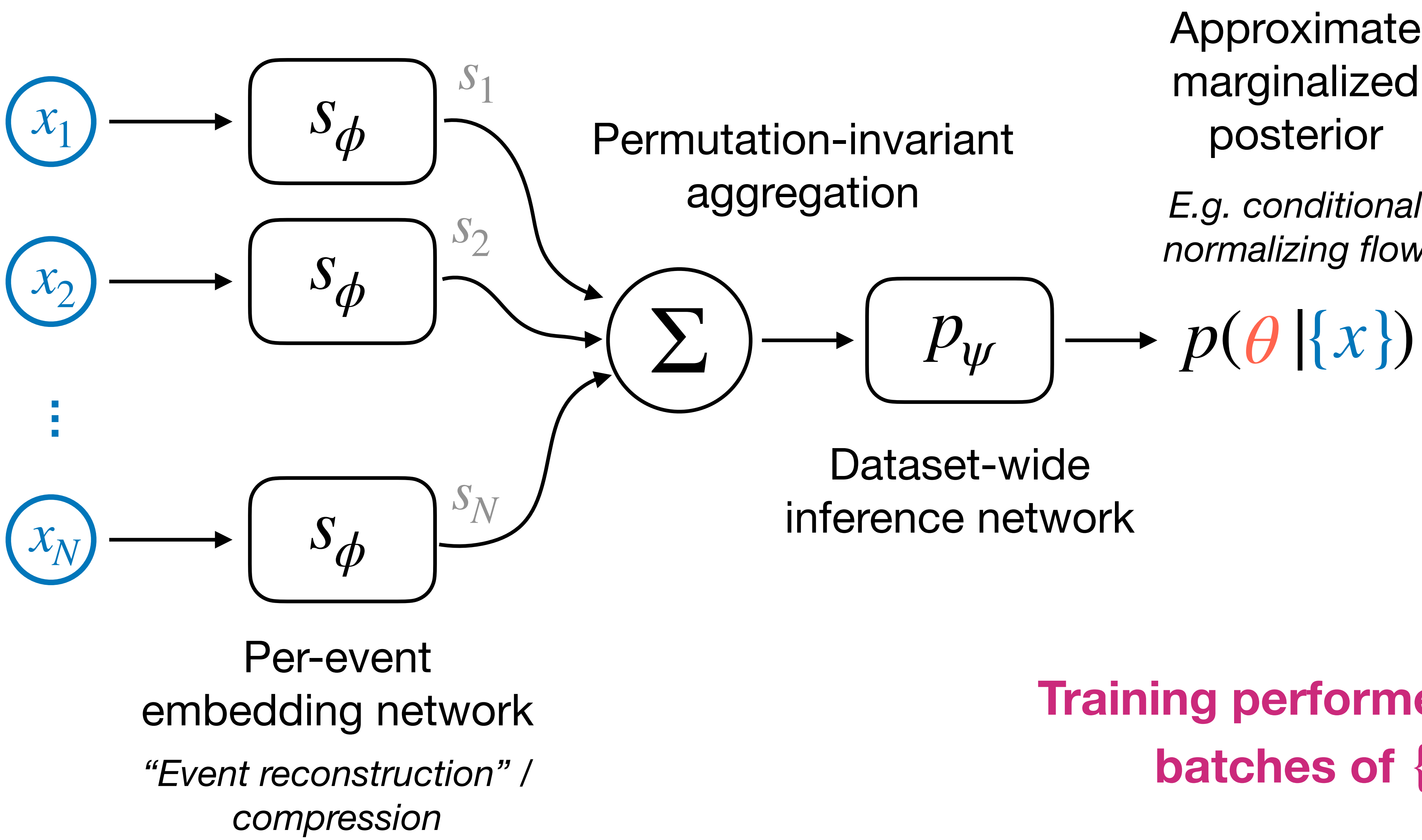
Our question:

Can SBI teach us anything about dataset-wide parameter inference?

Our approach

Use deep set for dataset-wide SBI

Varying cardinality, local + global nuisance parameters



Training performed on batches of $\{x\}, \theta$

Example: varying cardinality

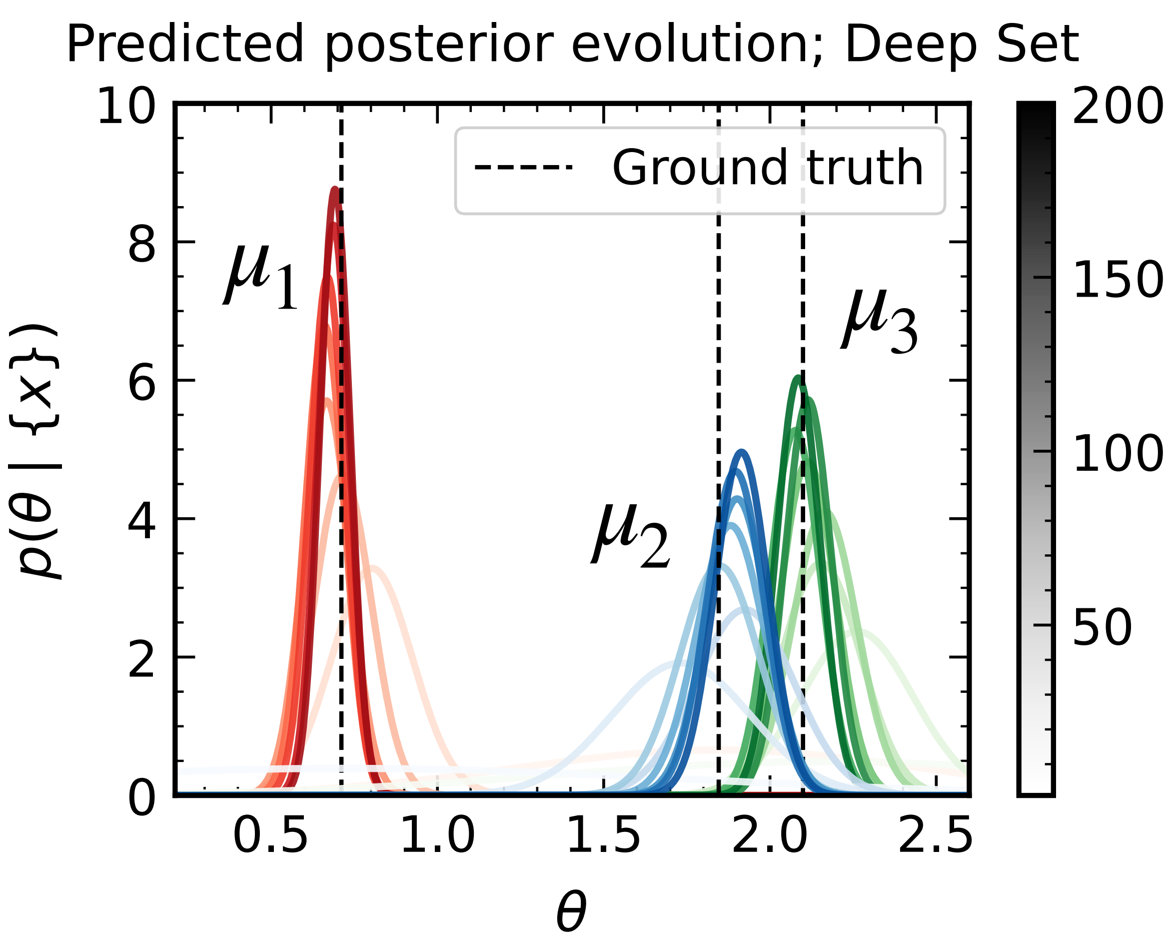
Infer mean vector from events x_i drawn from 3-dim normal distribution

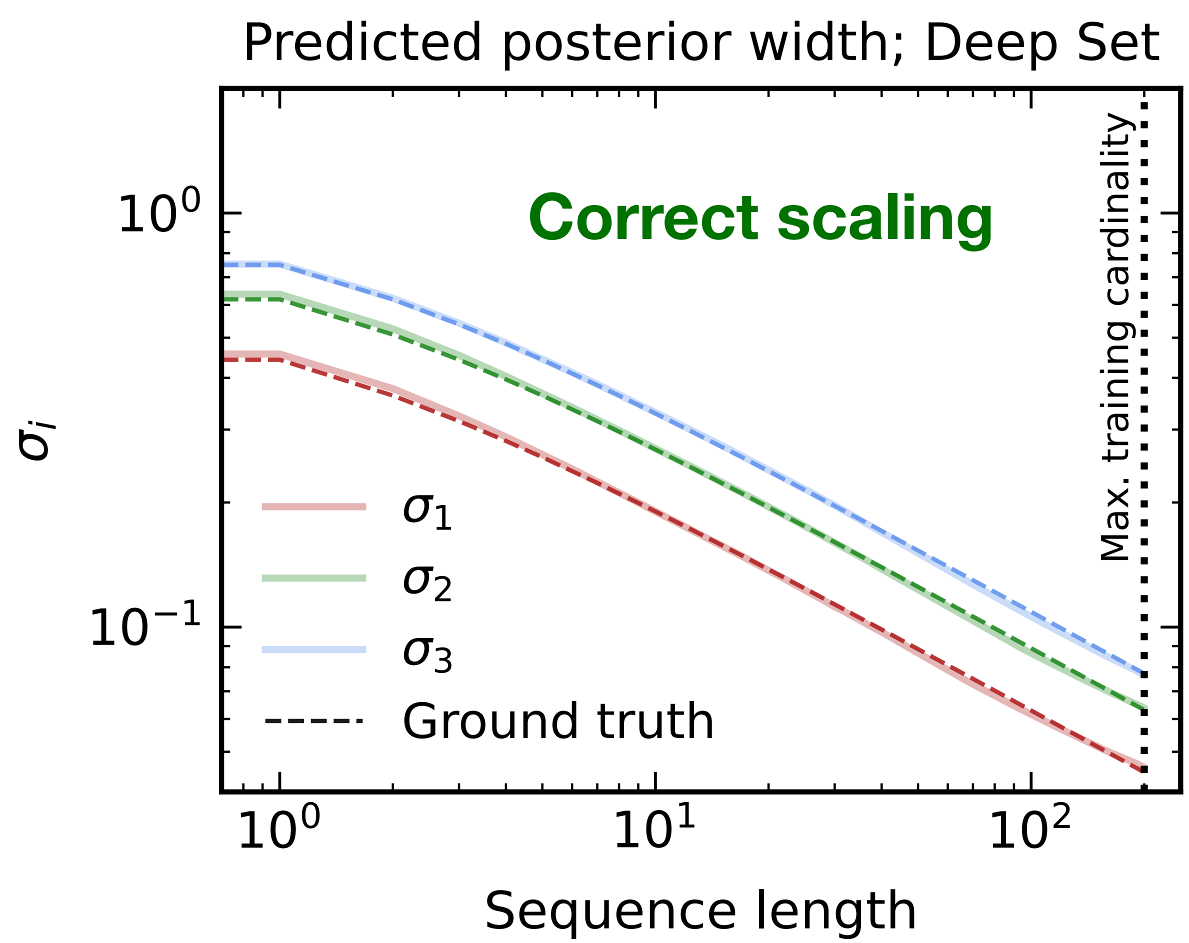
$$x_i \sim \mathrm{No}(\mu_{\mathrm{true}}, \Sigma_{\mathrm{true}}) \qquad \theta = \{\mu_1, \mu_2, \mu_3\}$$

Unknown mean vector

(Diagonal) covariance matrix assumed known

Inferred parameters of interest

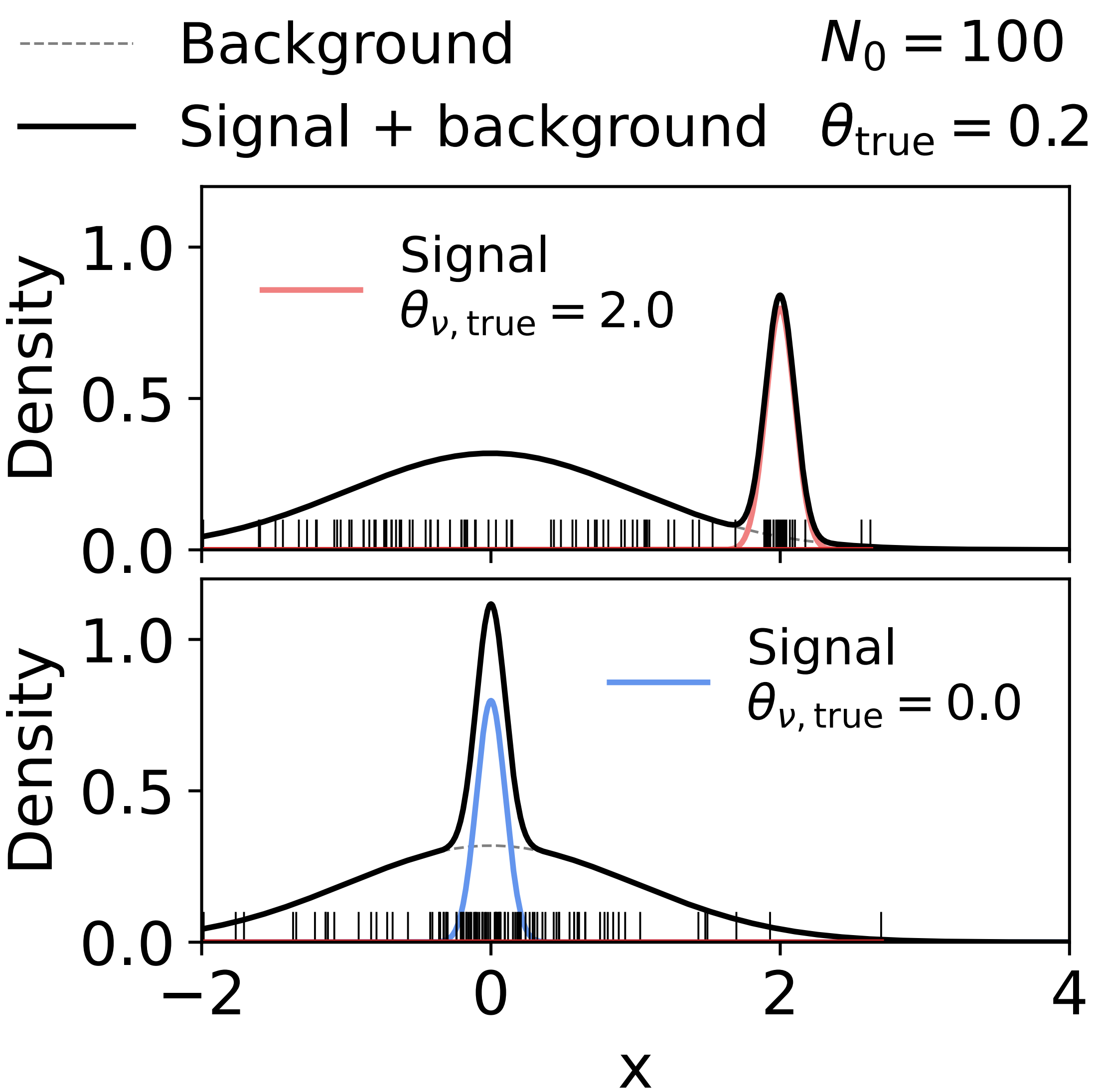




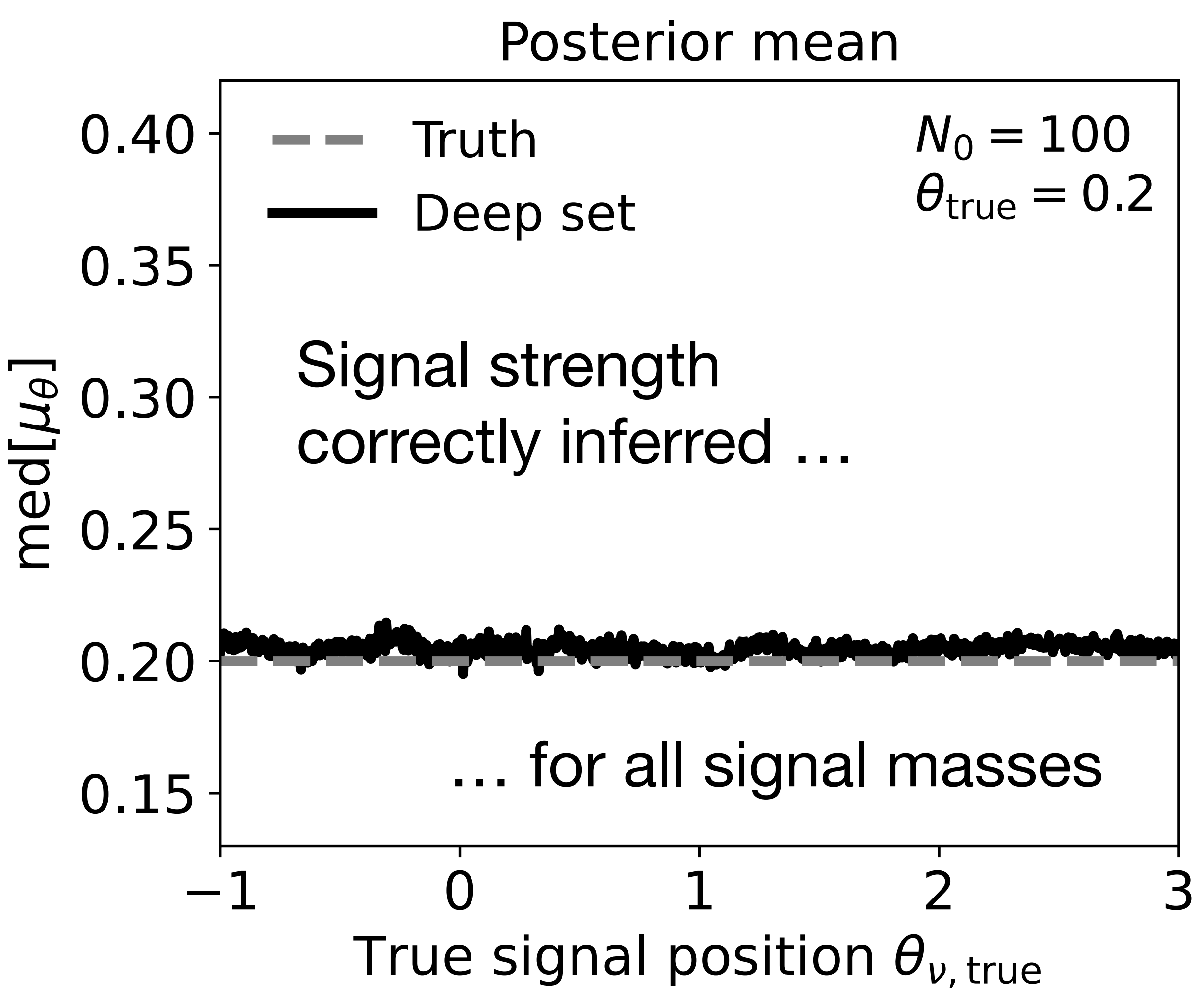
Example: “bump hunt”

Narrow “signal” with unknown mass on top of broad “background”

Find posterior on signal fraction, marginalized over signal mass (global nuisance parameter)



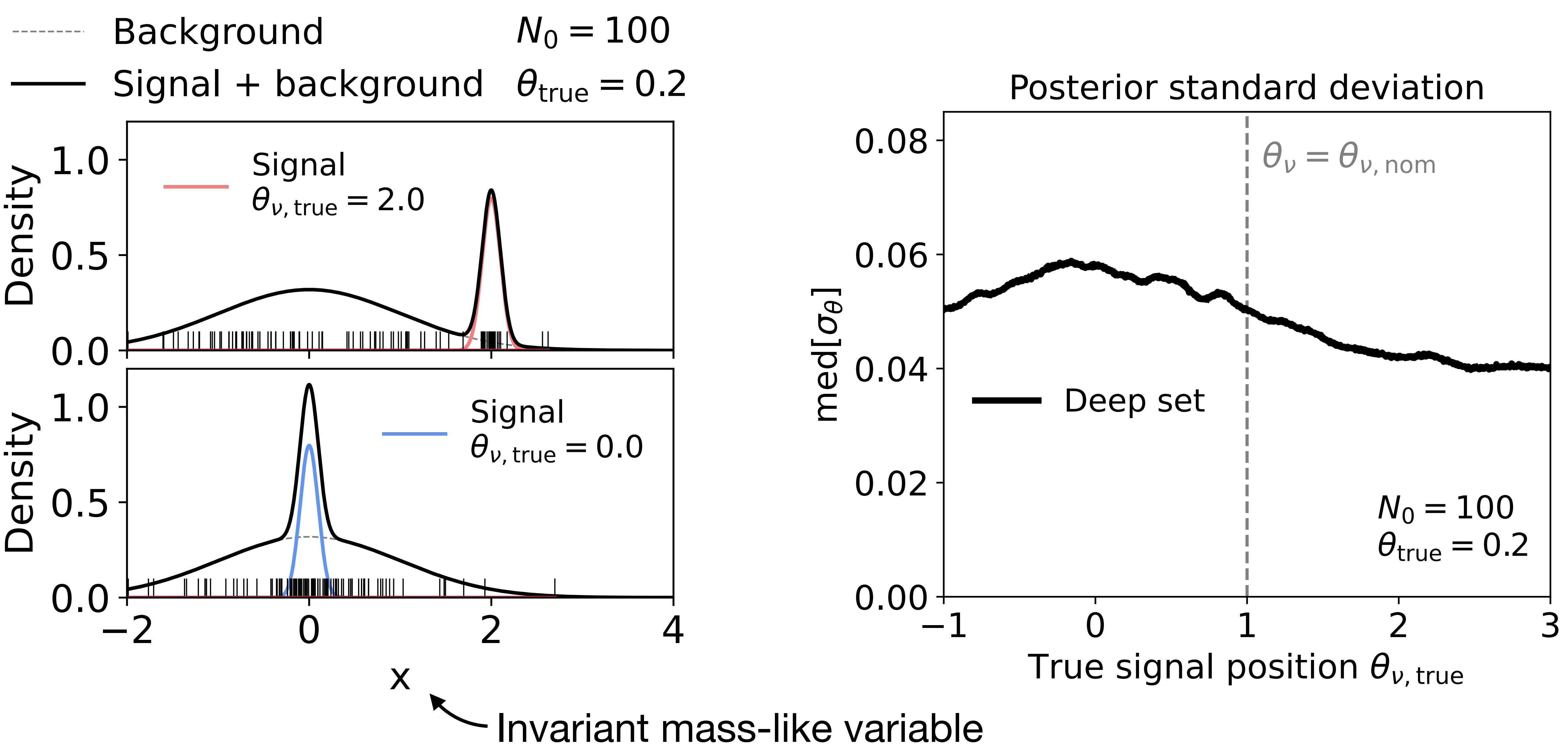
Invariant mass-like variable



Example: “bump hunt”

Narrow “signal” with unknown mass on top of broad “background”

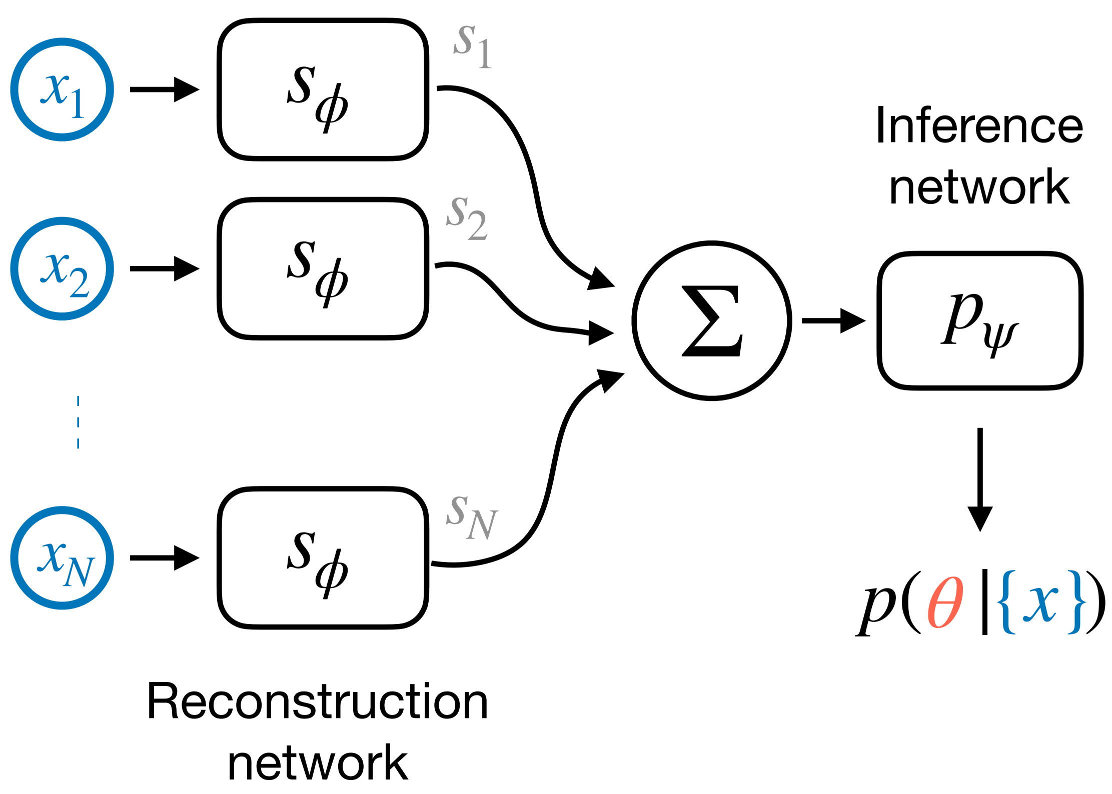
Find posterior on signal fraction, marginalized over signal mass
(global nuisance parameter)

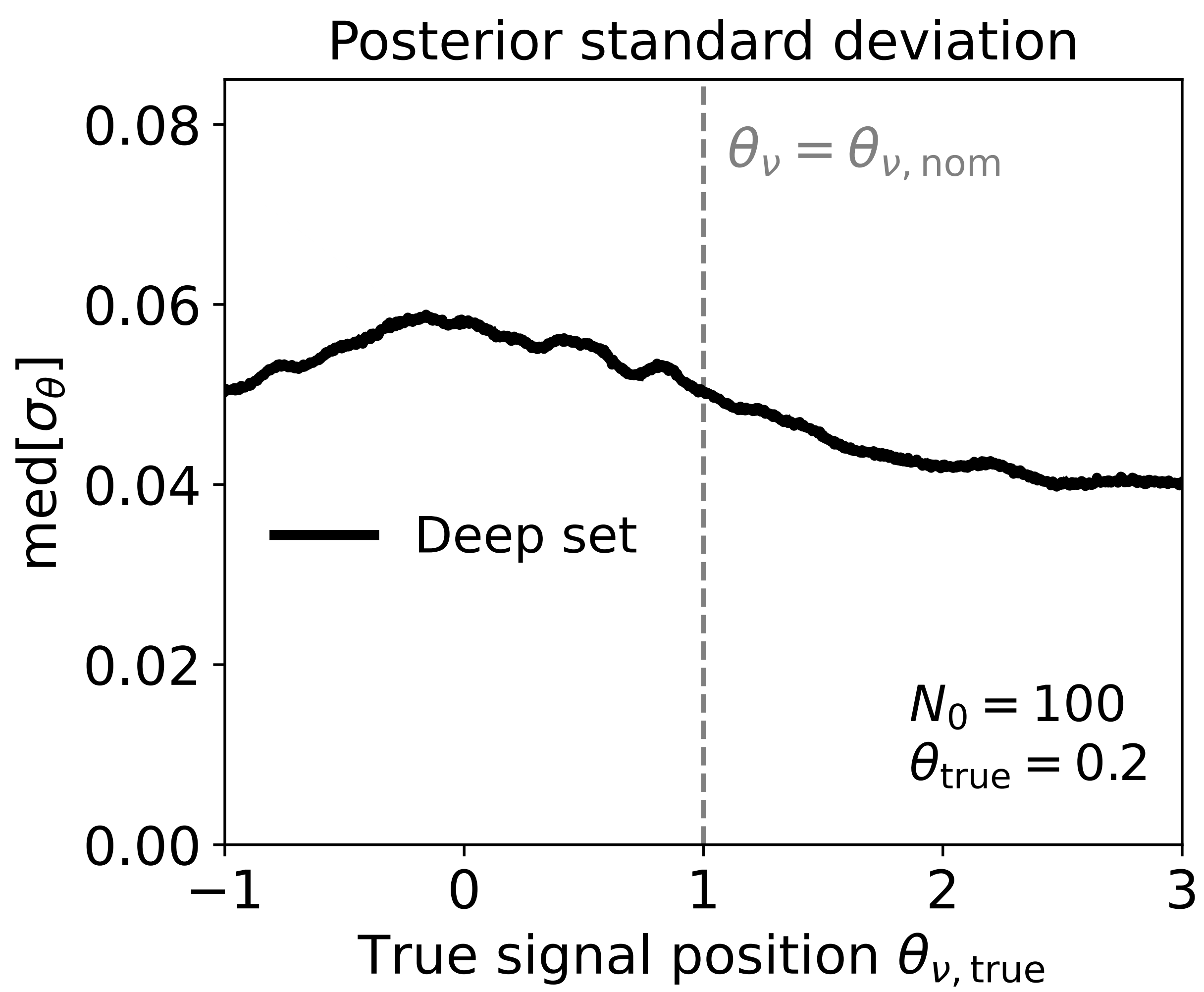


Example: “bump hunt”

Narrow “signal” with unknown mass on top of broad “background”

Find posterior on signal fraction, marginalized over signal mass (global nuisance parameter)



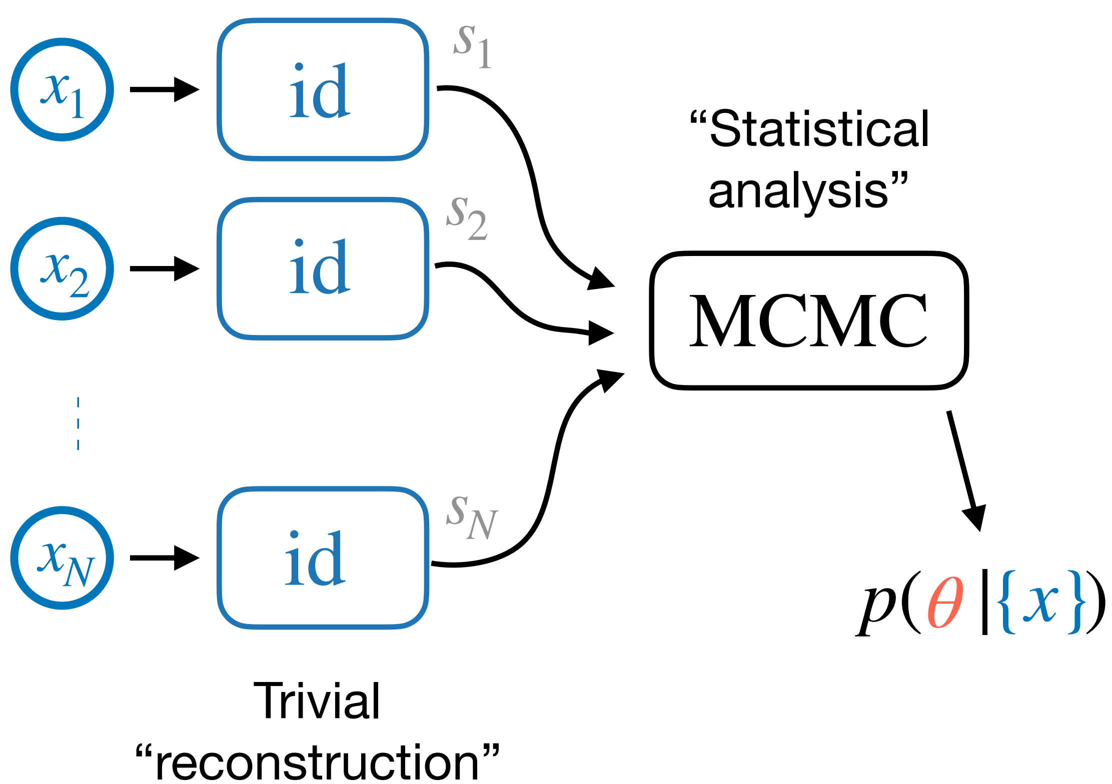


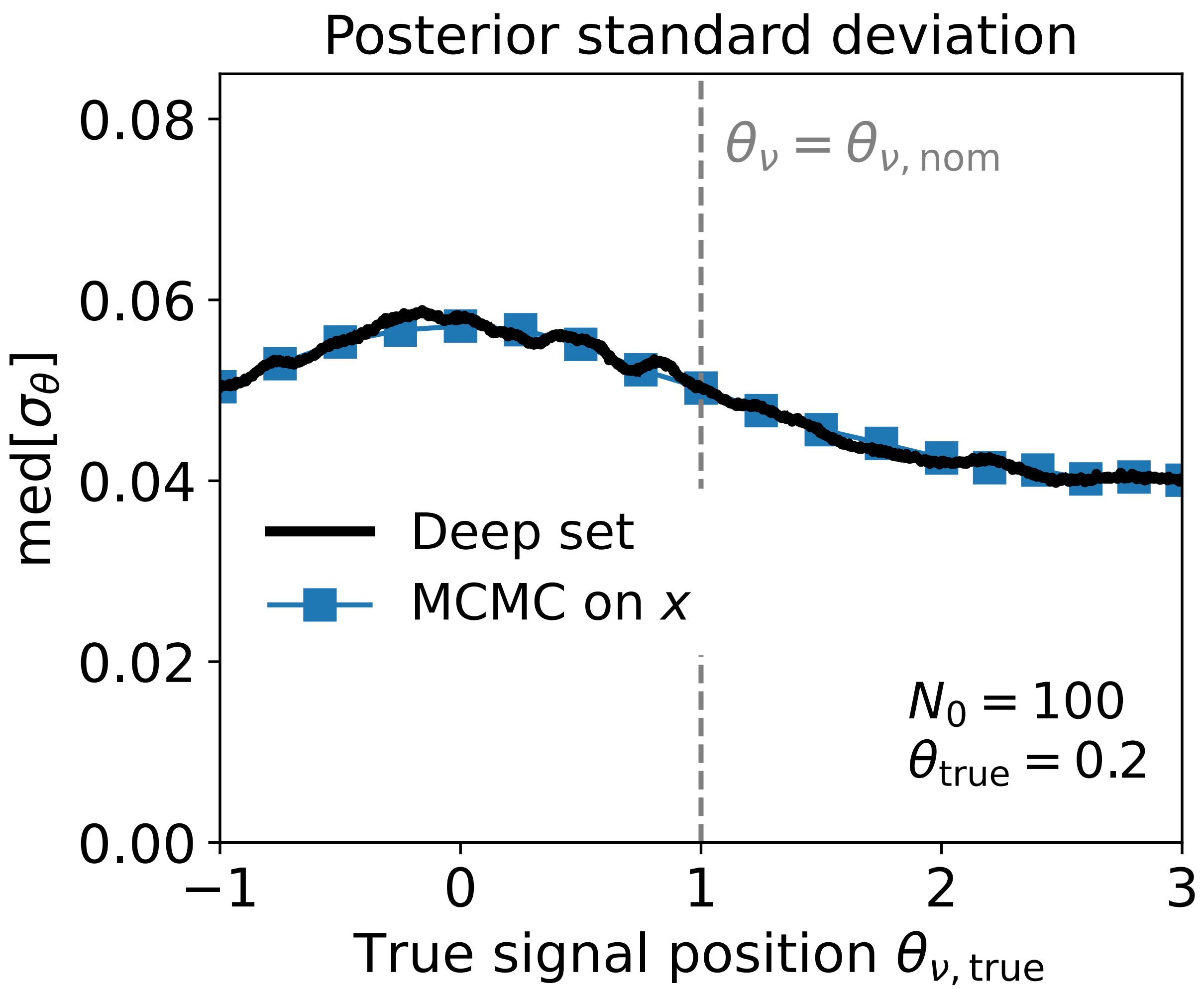
Example: “bump hunt”

Narrow “signal” with unknown mass on top of broad “background”

Find posterior on signal fraction, marginalized over signal mass (global nuisance parameter)

“Ground truth”: MCMC on full per-event information





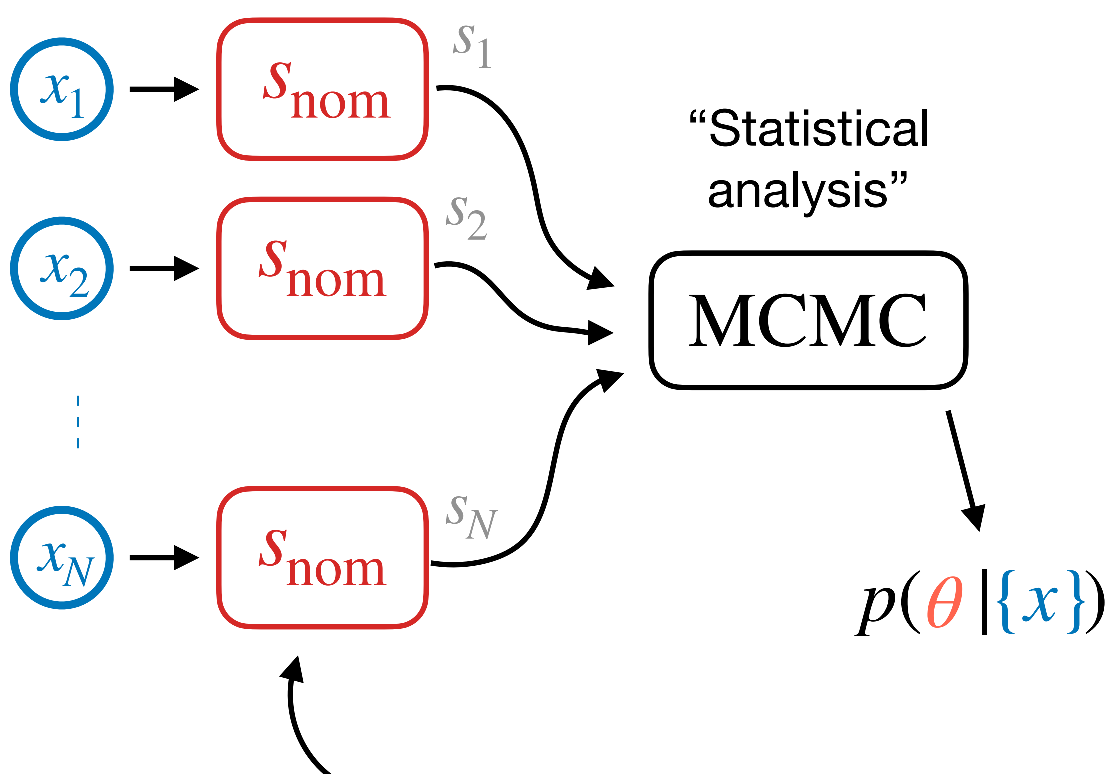
Deep set achieves optimal accuracy!

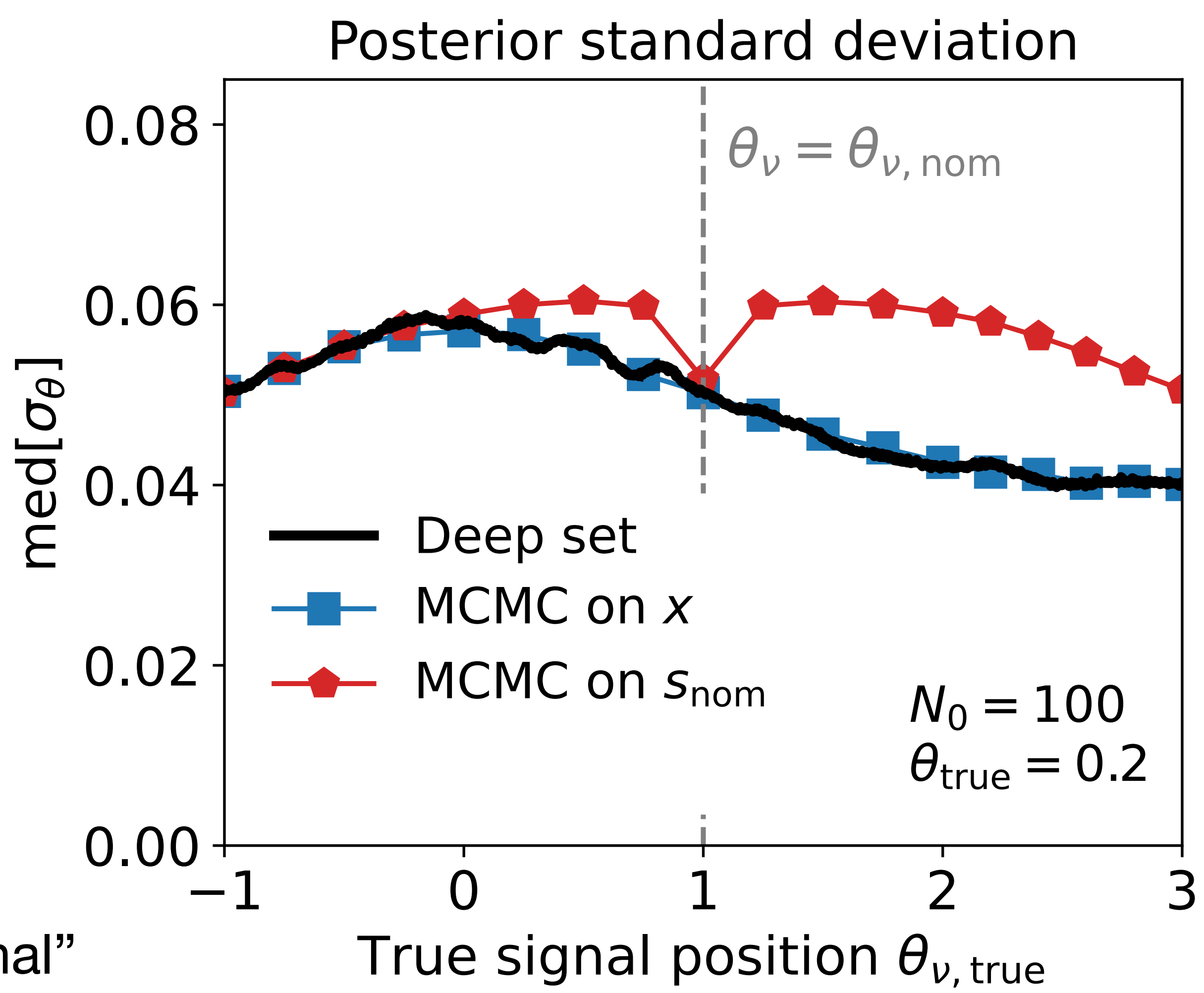
Example: "bump hunt"

Narrow "signal" with unknown mass on top of broad "background"

Find posterior on signal fraction, marginalized over signal mass (global nuisance parameter)

Common practice in HEP:
Use classifier output for inference



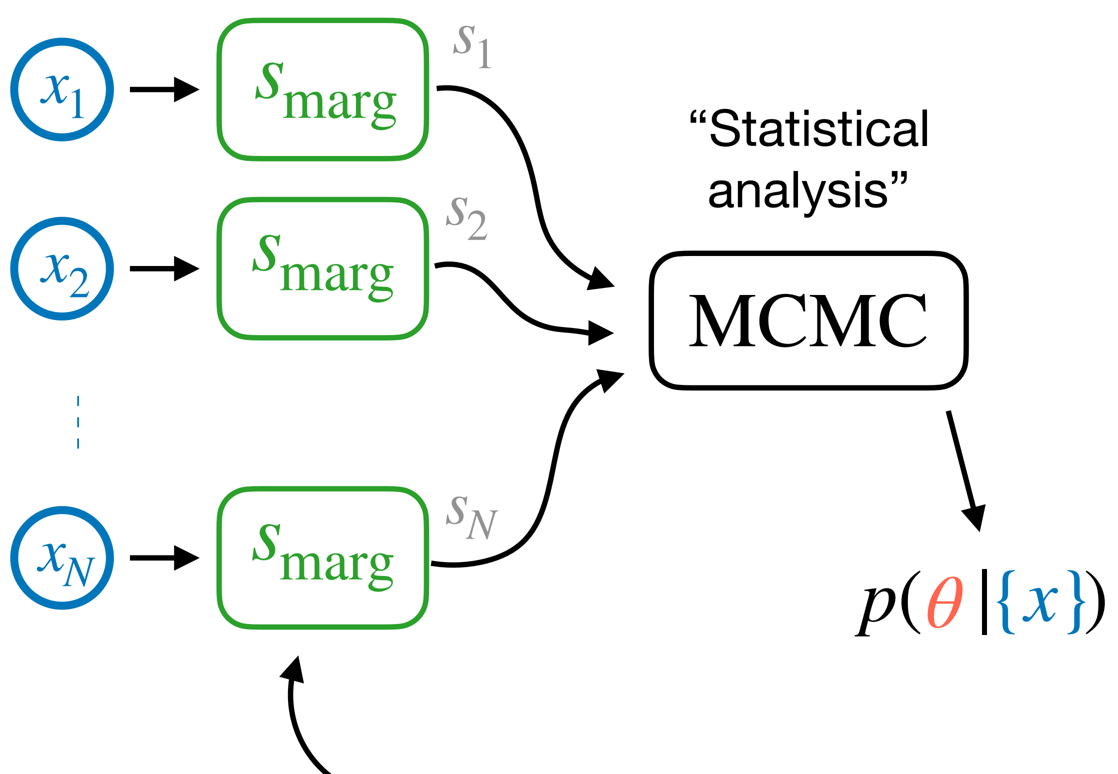


Example: "bump hunt"

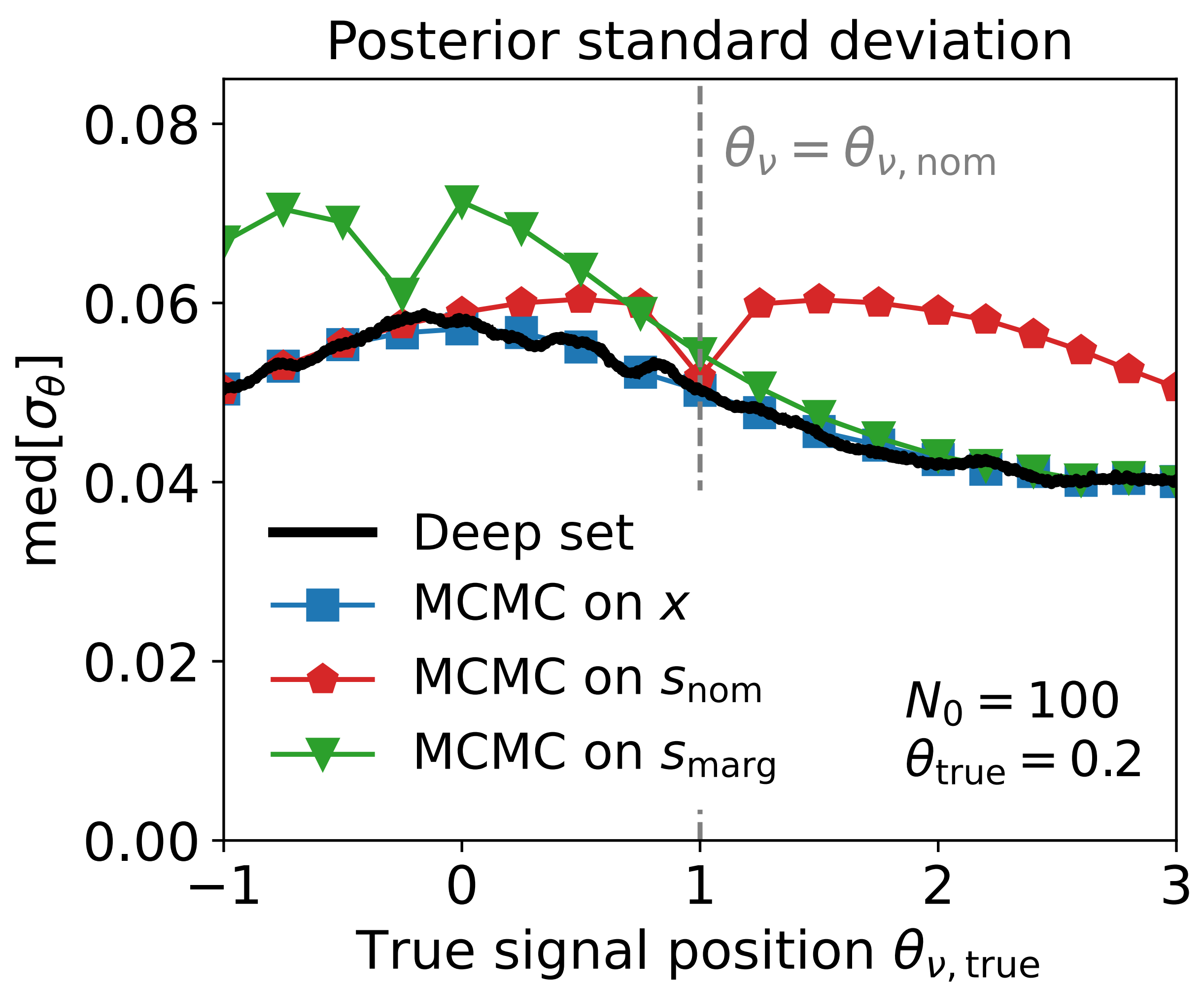
Narrow "signal" with unknown mass on top of broad "background"

Find posterior on signal fraction, marginalized over signal mass (global nuisance parameter)

Common practice in HEP:
Use classifier output for inference



Classifier trained to separate "marginalized" signal from background



Example: “bump hunt”

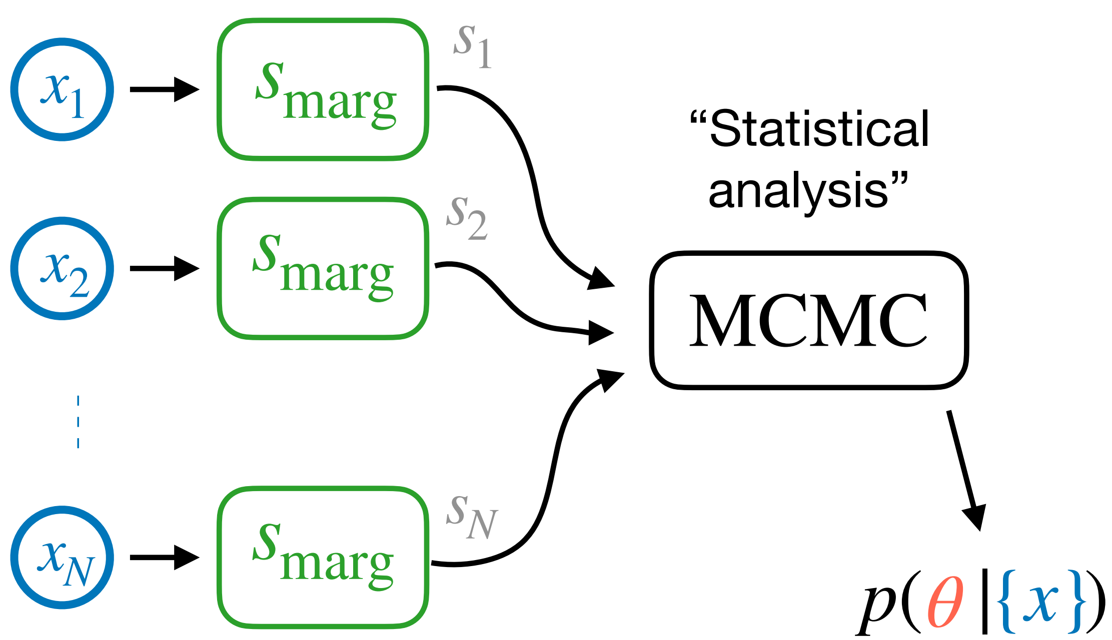
Narrow “signal” with unknown mass on top of broad “background”

marginalized over signal mass
e parameter)

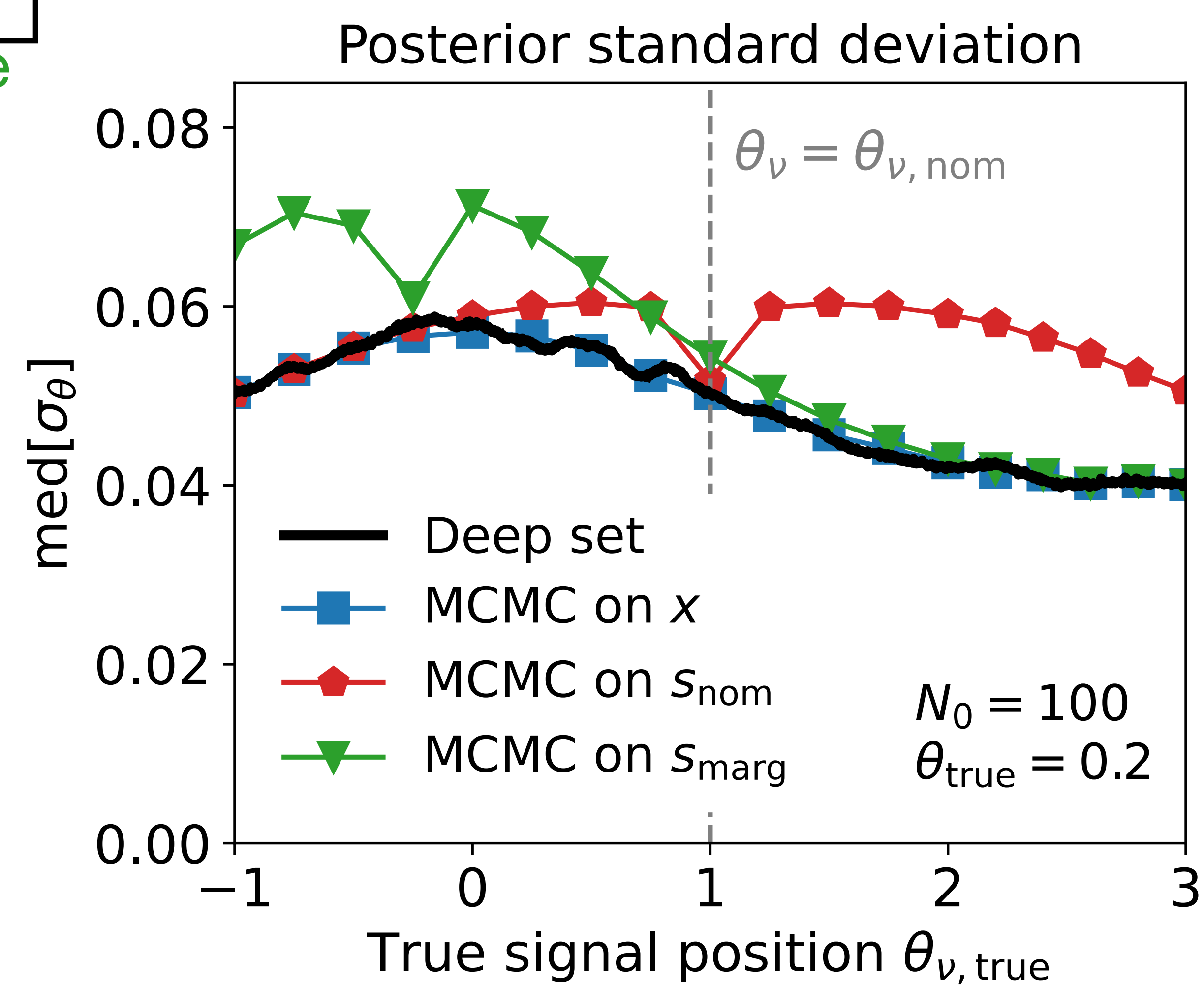
To avoid information loss, need classifier to be parameterized w.r.t. all nuisance parameters

(possibly intractably many!)

Use classifier output for inference



Classifier trained to separate “marginalized” signal from background



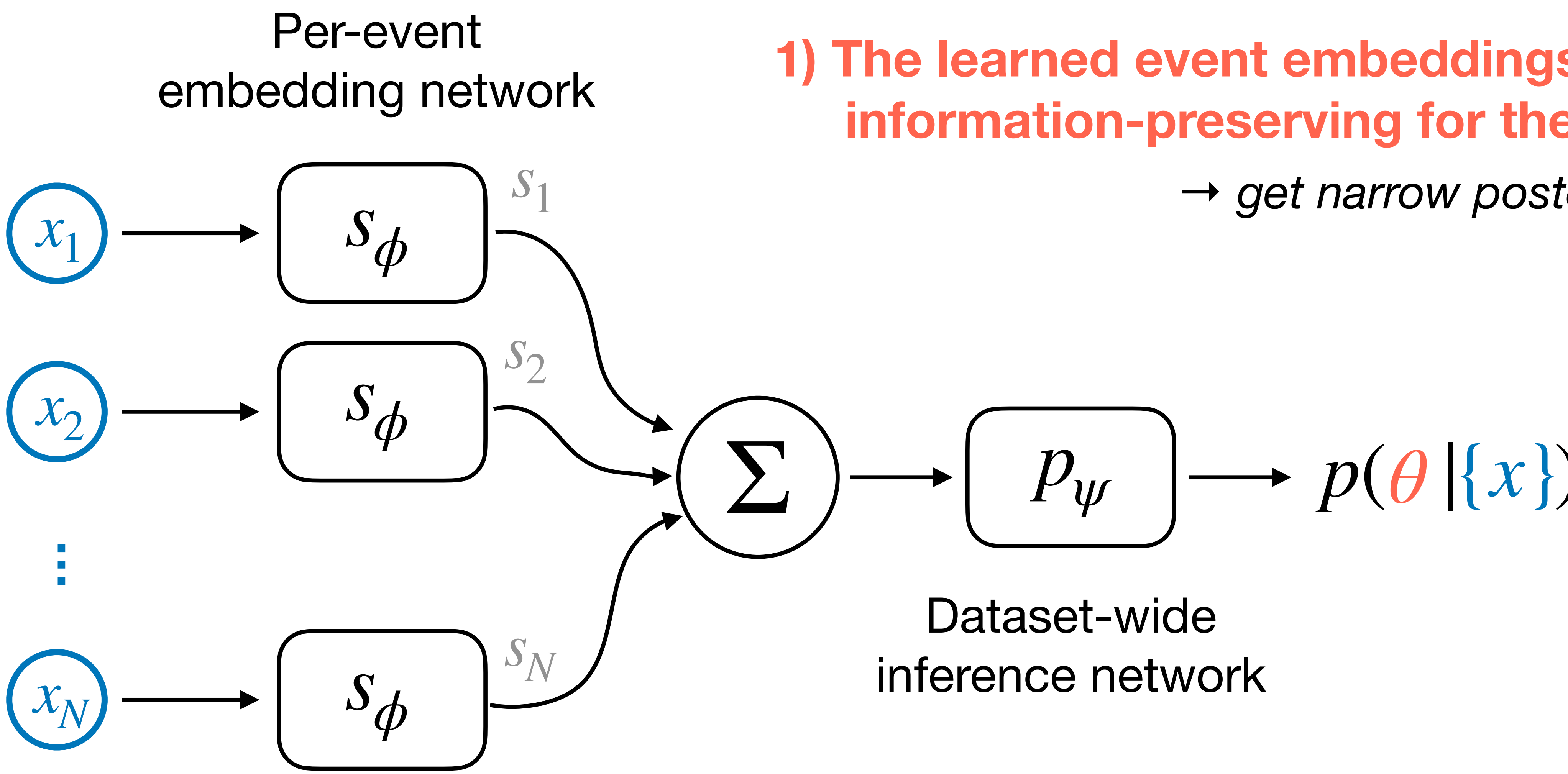
What is happening here?

Deep set can learn <u>arbitrary</u> permutation-invariant functions

→ sufficiently expressive to aggressively amortize the "reconstruction" + inference task

1) The learned event embeddings are information-preserving for the POI

→ *get narrow posteriors*



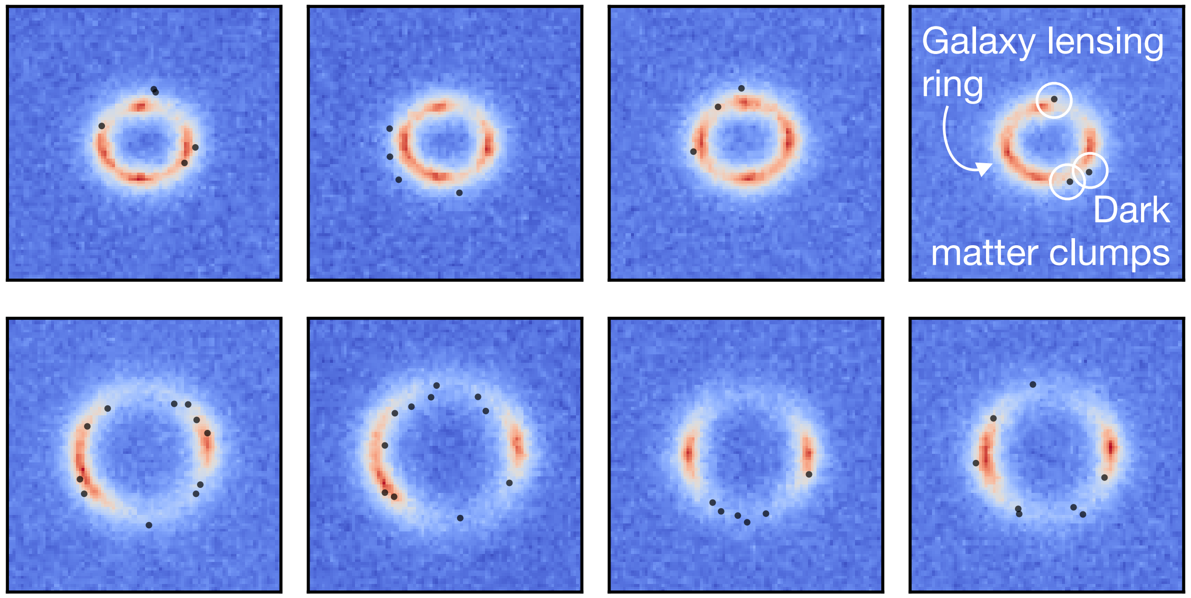
2) The embeddings are guaranteed to compose under addition!

→ *cheap to update posterior estimate with new data*

More complicated example: strong lensing

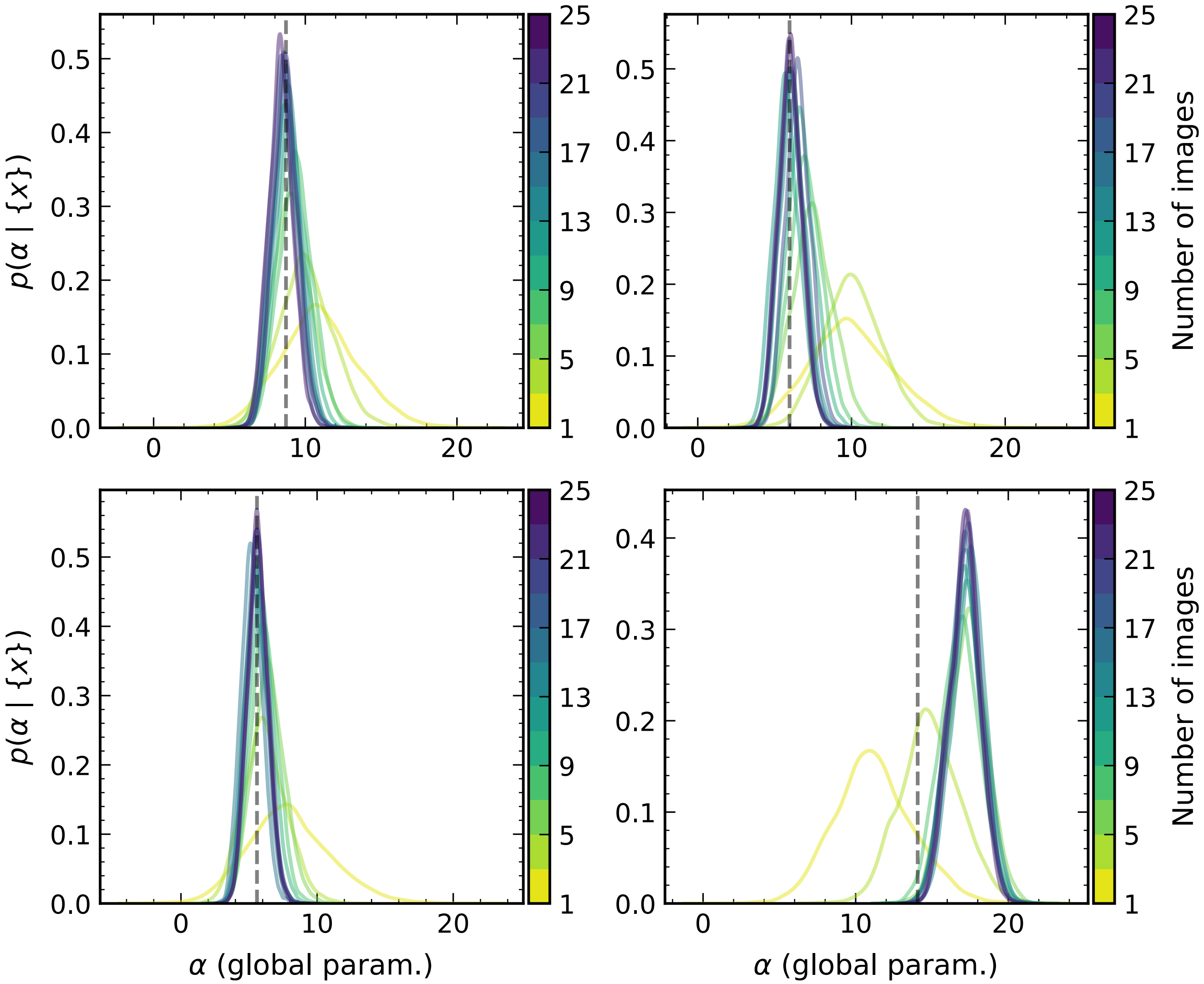
⟵ Different latent realizations $z_{sub} \sim p(z_{sub} \mid \theta_{glob}, \theta_{loc})$ ⟶

⟵ Vary θ_{glob} and θ_{loc} ⟶



Strong gravitational lensing + dark matter clumps

Recover true parameter values as $N \to \infty$

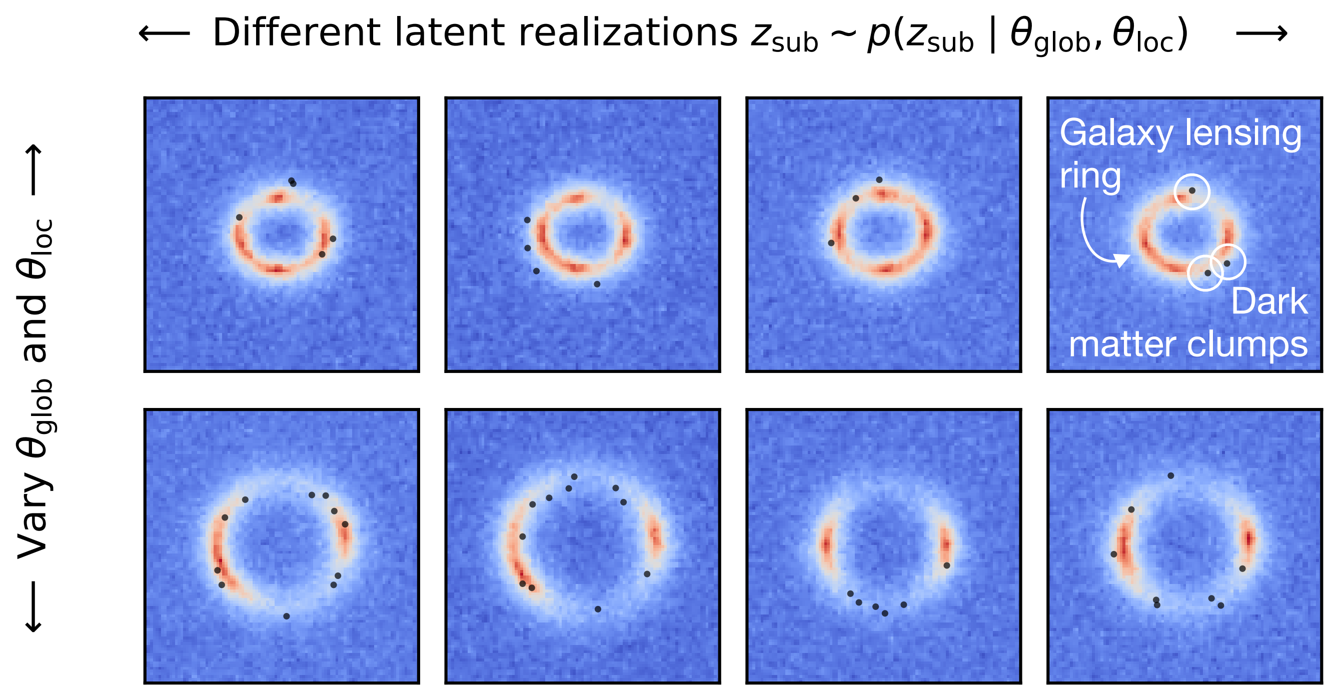


"Substructure fraction"

Global parameters: dark matter clump population parameters

Local parameters: per-image lensing & realization of dark matter clumps

No tractable likelihood!



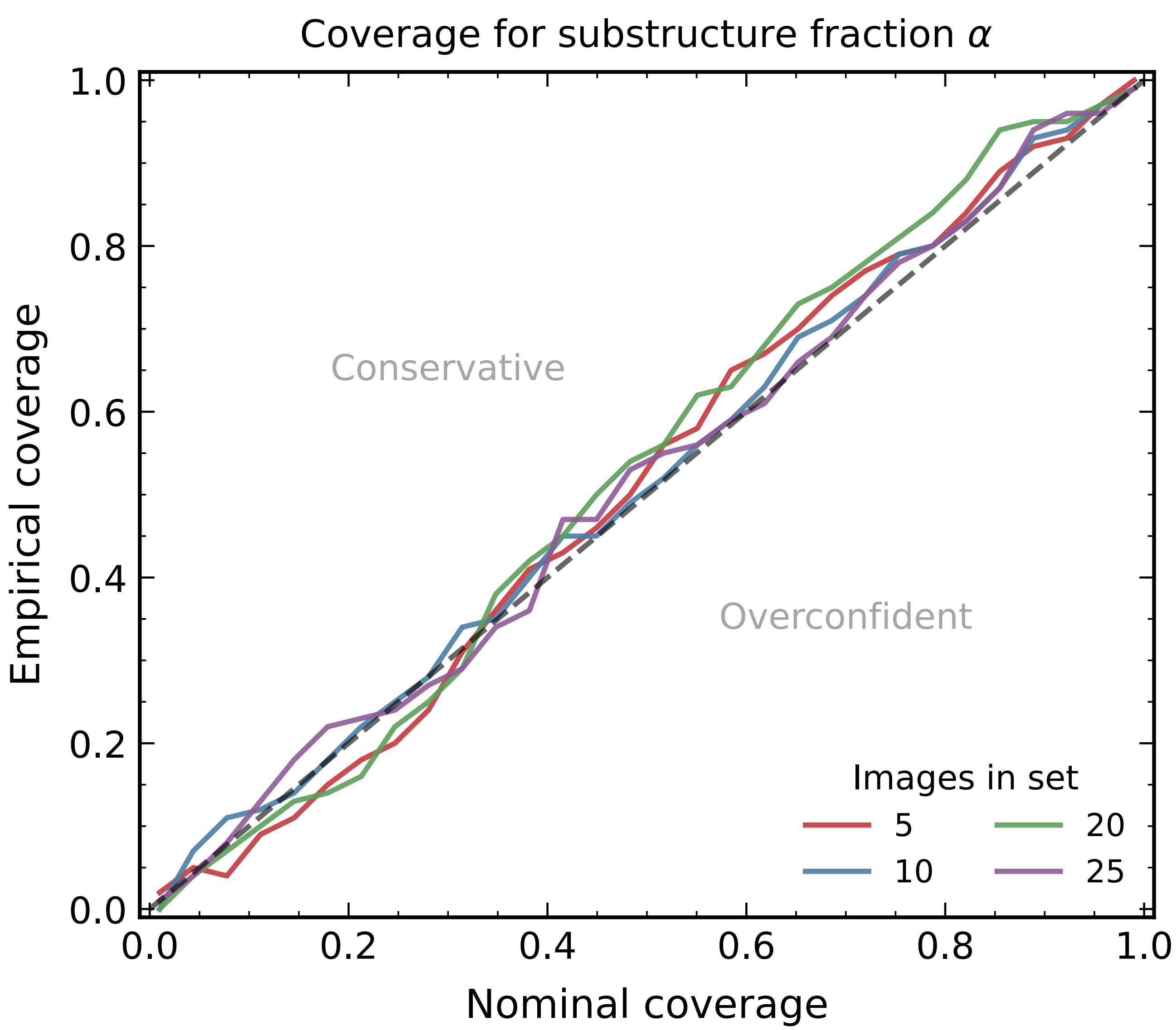
Strong gravitational lensing + dark matter clumps

Global parameters: dark matter clump population parameters

Local parameters: per-image lensing & realization of dark matter clumps

No tractable likelihood!

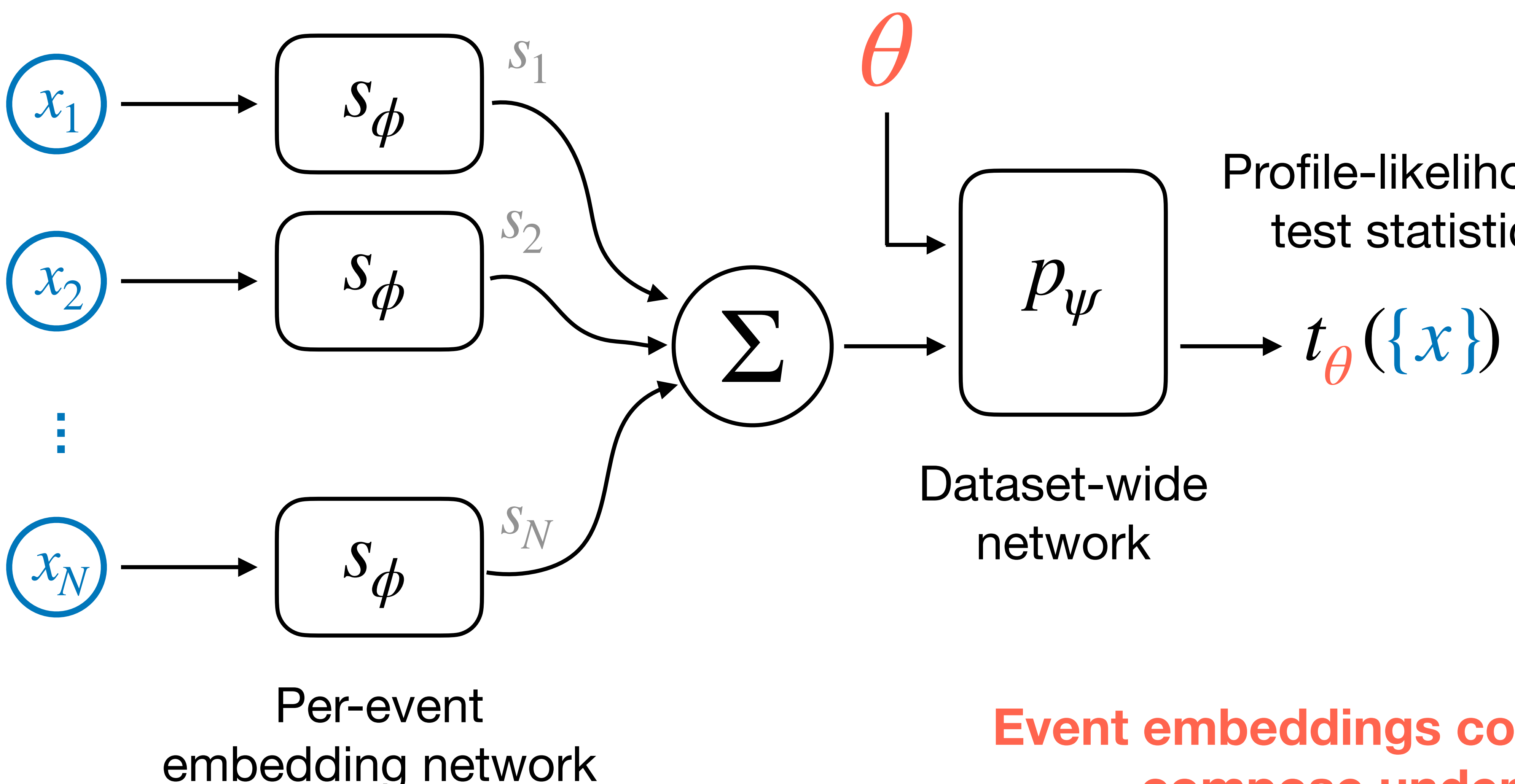
Uncertainty estimates reliable



Frequentist-style inference also works

Deep set can learn arbitrary permutation-invariant functions

→ sufficiently expressive to aggressively amortize the "reconstruction" + inference task



Event embeddings continue to compose under addition

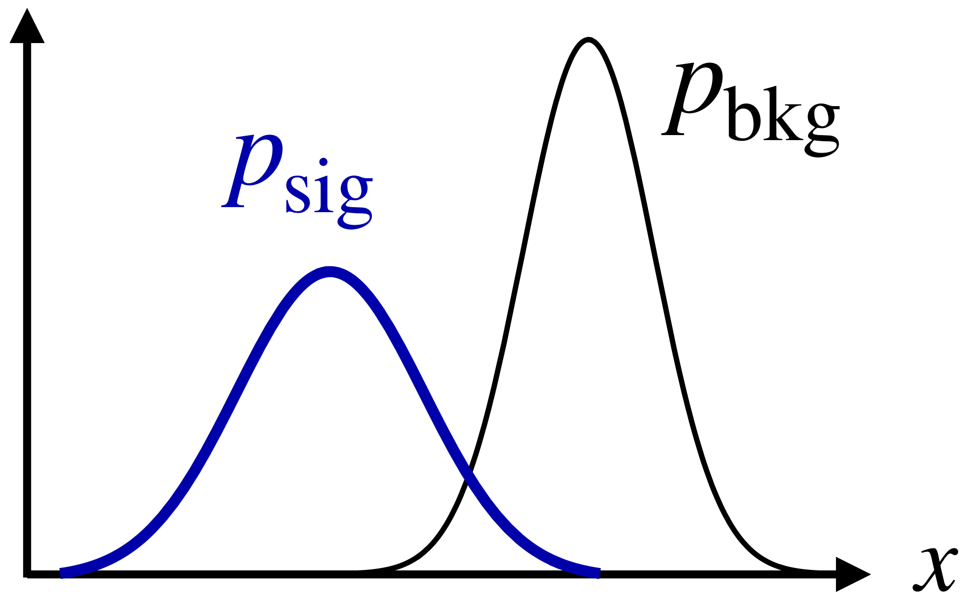
→ *cheap to update test statistic with new data*

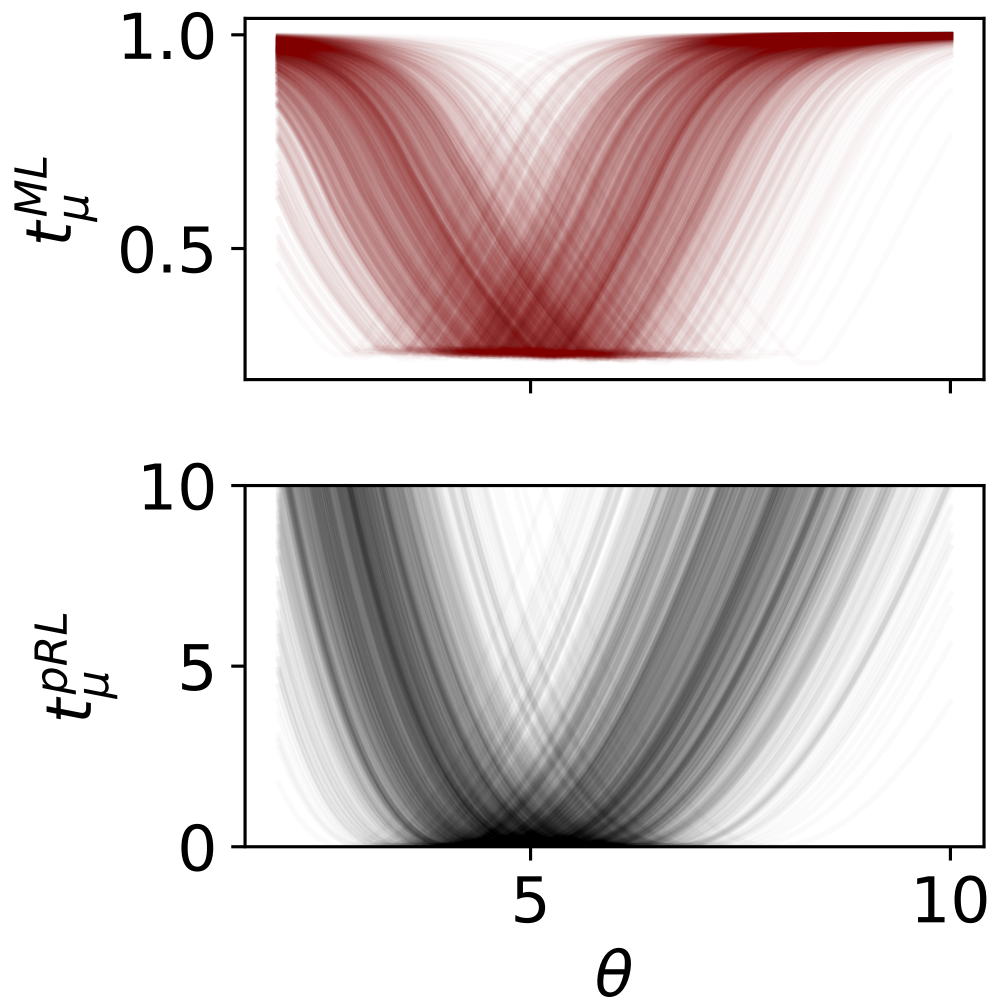
Mixture of Gaussians with different strengths

$$p(x) = \mu_{\text{sig}}\, p_{\text{sig}}(x) + \mu_{\text{bkg}}\, p_{\text{bkg}}(x)$$

θ — *Parameter of interest*

θ_ν — *Global nuisance parameter*

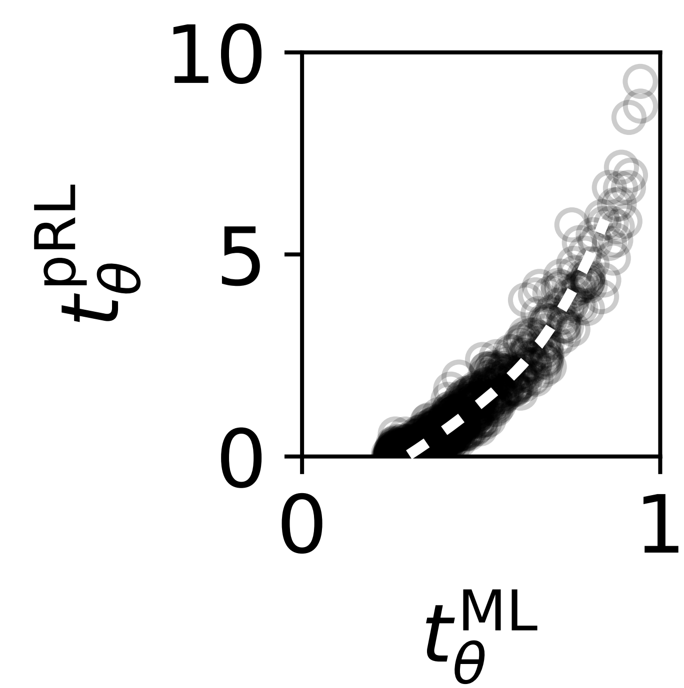




Learned test statistic

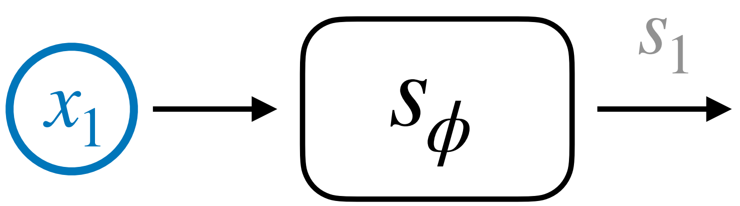
Bijective →

True profile-likelihood test statistic



Summary and outlook

Dataset-wide SBI works reliably in the presence of local and global nuisance parameters *(as others have also shown!)*



Learned event embedding is information-preserving for parameter of interest …

(Without requiring parameterization in terms of nuisance parameters!)

… composes under addition …

(Trivial to update inference on larger dataset!)

$$t_\theta(\{x\})$$

$$p(\theta \,|\, \{x\})$$

… and enables Bayesian and Frequentist amortized parameter inference.

But: requires training on *batches of datasets*

→ How to go beyond a few 10^3 events / dataset?

→ How to interpret event embedding beyond training cardinality?

→ Effects of deficiencies in simulation?

More information: [arXiv:2306.12584]