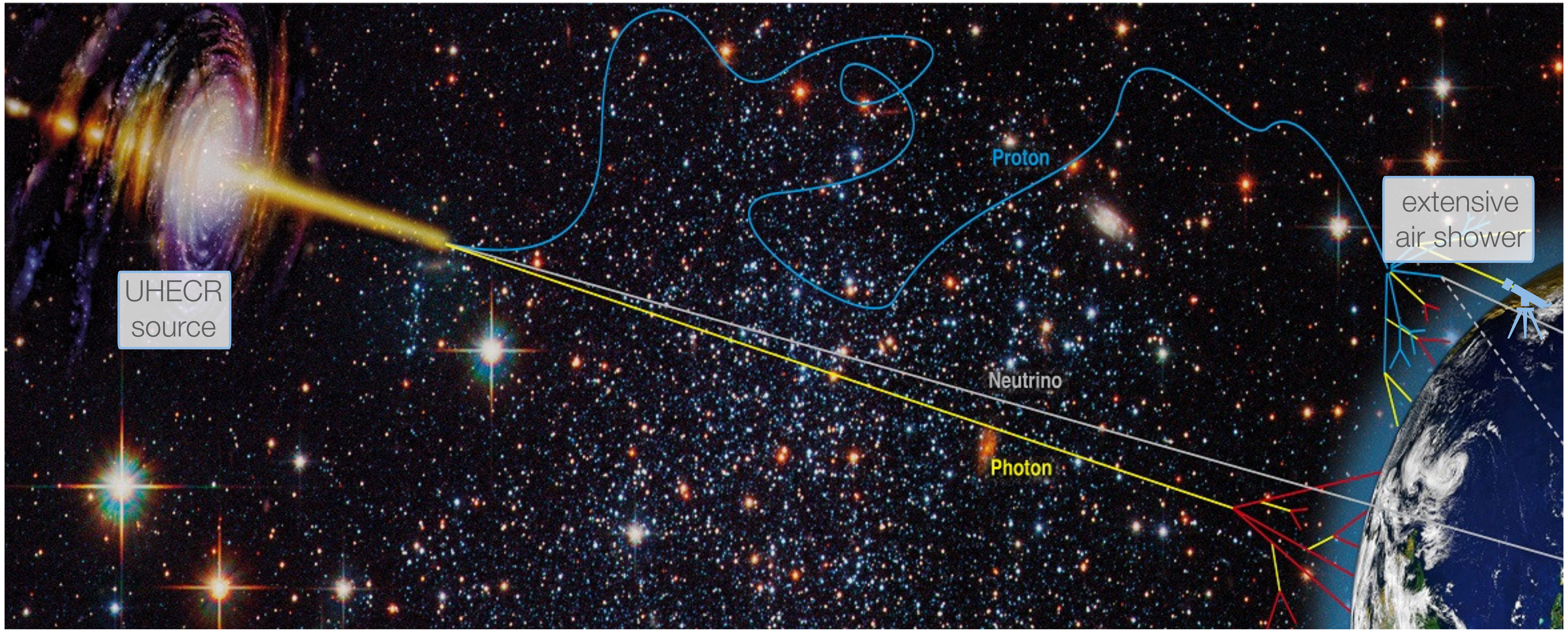
SPONSORED BY THE

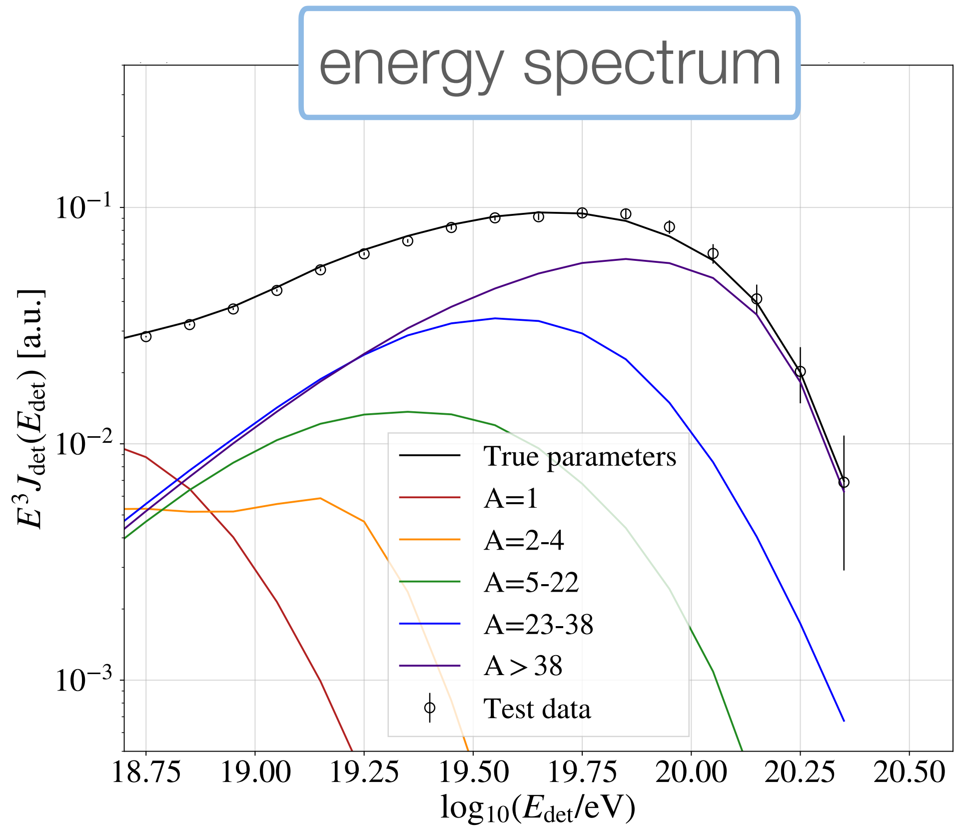
Federal Ministry
of Education
and Research

IDT-UM Collaboration Meeting

Parameter Estimation
by
conditional Invertible Neural Networks



https://www.uni-potsdam.de/typo3temp/assets/_processed_/b/c/csm_kosm_Strahlung2_b7bead59a7.jpg



Teresa Bister, Martin Erdmann, *Josina Schulte*

Introduction

11.05.21

Problem

Often, physical parameters cannot be measured directly

forward process

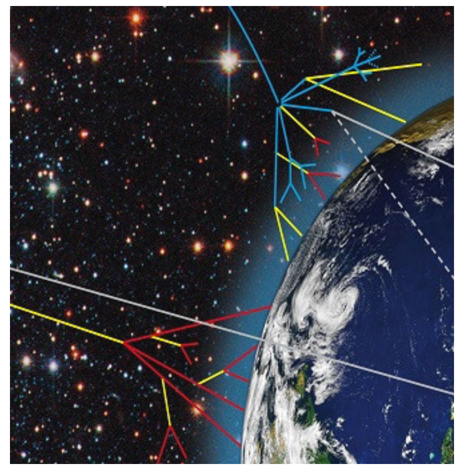
often with information loss

physical parameters θ

observables y

inverse process

analytically intractable

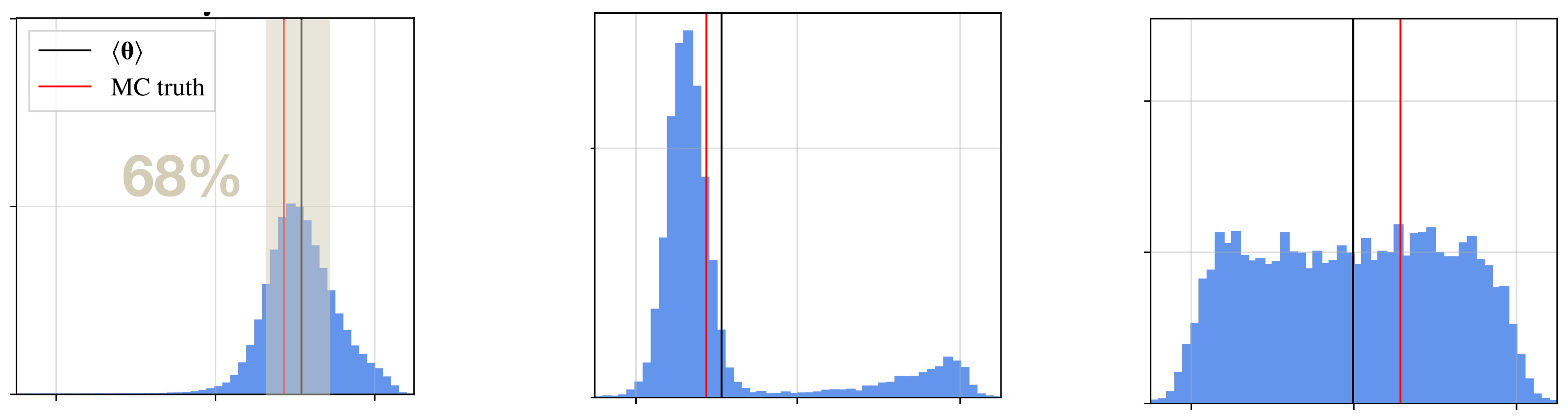


Goal

Reconstruct complete *posterior distribution* of physical parameters, conditioned on an observation

i. quantify uncertainty
ii. reveal multi-modal distributions
iii. identify degenerate & unrecoverable parameters

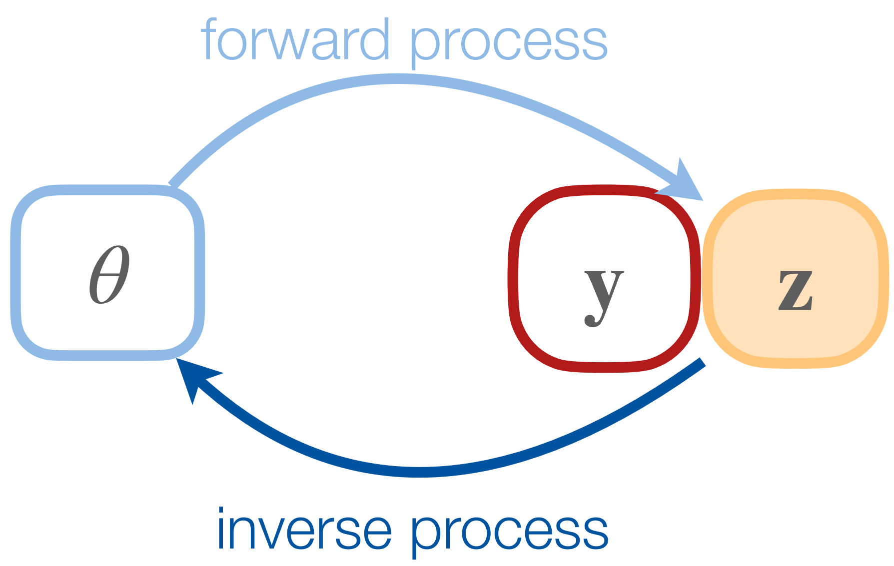
i. ii. iii.



Overview: Invertible Neural Network

Idea

invertible NN learns mapping between physical parameters θ & observables y

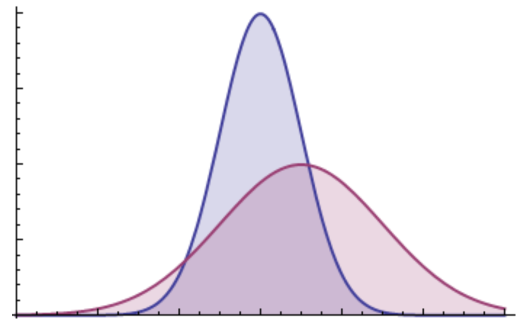


- training process: known forward model
- inverse pass: posterior distributions

very fast evaluation of many data sets

Approach

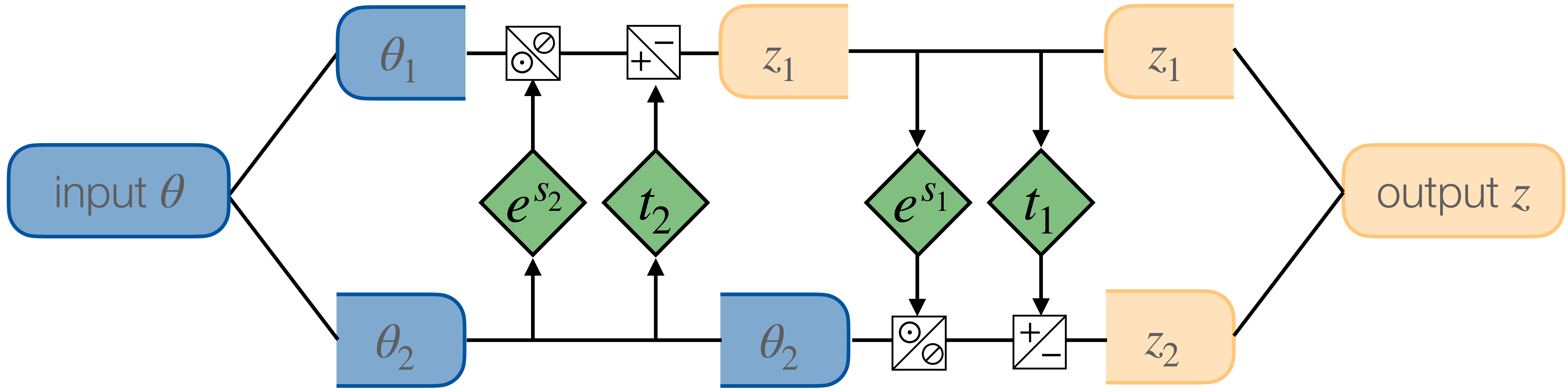
- add **latent variables z** to capture information otherwise lost
- **loss function**
  - minimizes difference: true underlying posterior distribution & cINN posterior distribution
  - ensures convergence to correct posterior distributions
- enforce normal distribution for latents z



Basic idea: Reversible blocks

reversible block

- split input vector θ in 2 halves: $\theta = [\theta_1, \theta_2]$
- transform by affine function: output $\mathbf{z} = [\mathbf{z_1}, \mathbf{z}_2]$
  coefficients $\exp(s_i)$ & t_i, $i \in \{1,2\}$
  (mappings s_i, t_i arbitrarily complicated, not invertible)
  $$\mathbf{z}_1 = \theta_1 \odot \exp(s_2(\theta_2)) + t_2(\theta_2)$$
  $$\mathbf{z}_2 = \theta_2 \odot \exp(s_1(\mathbf{z}_1)) + t_1(\mathbf{z}_1)$$
- Invertibility
  element-wise multiplication
  $$\theta_2 = (\mathbf{z}_2 - t_1(\mathbf{z}_1)) \odot \exp(-s_1(\mathbf{z}_1))$$
  $$\theta_1 = (\mathbf{z}_1 - t_2(\theta_2)) \odot \exp(-s_2(\theta_2))$$



based on *real-valued non-volume preserving* (**Real NVP**) architecture (*arXiv:1605.08803*) & **GLOW** (*arXiv:1807.03039*)

Basic idea: Reversible blocks

reversible block

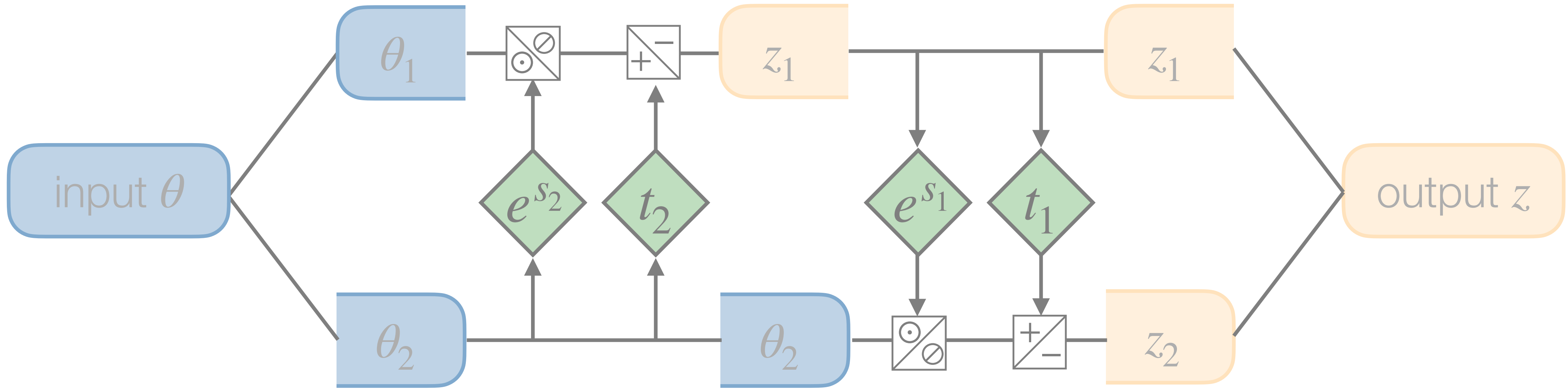
- split input vector θ in 2 halves: $\theta = [\theta_1, \theta_2]$
- transform by affine function: output $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2]$
  coefficients $\exp(s_i)$ & t_i $i \in \{1,2\}$
  (mappings s_i, t_i arbitrarily complicated, not invertible)

$$\mathbf{z}_1 = \theta_1 \odot \exp(s_2(\theta_2)) + t_2(\theta_2)$$
$$\mathbf{z}_2 = \theta_2 \odot \exp(s_1(\mathbf{z}_1)) + t_1(\mathbf{z}_1)$$

- Invertibility

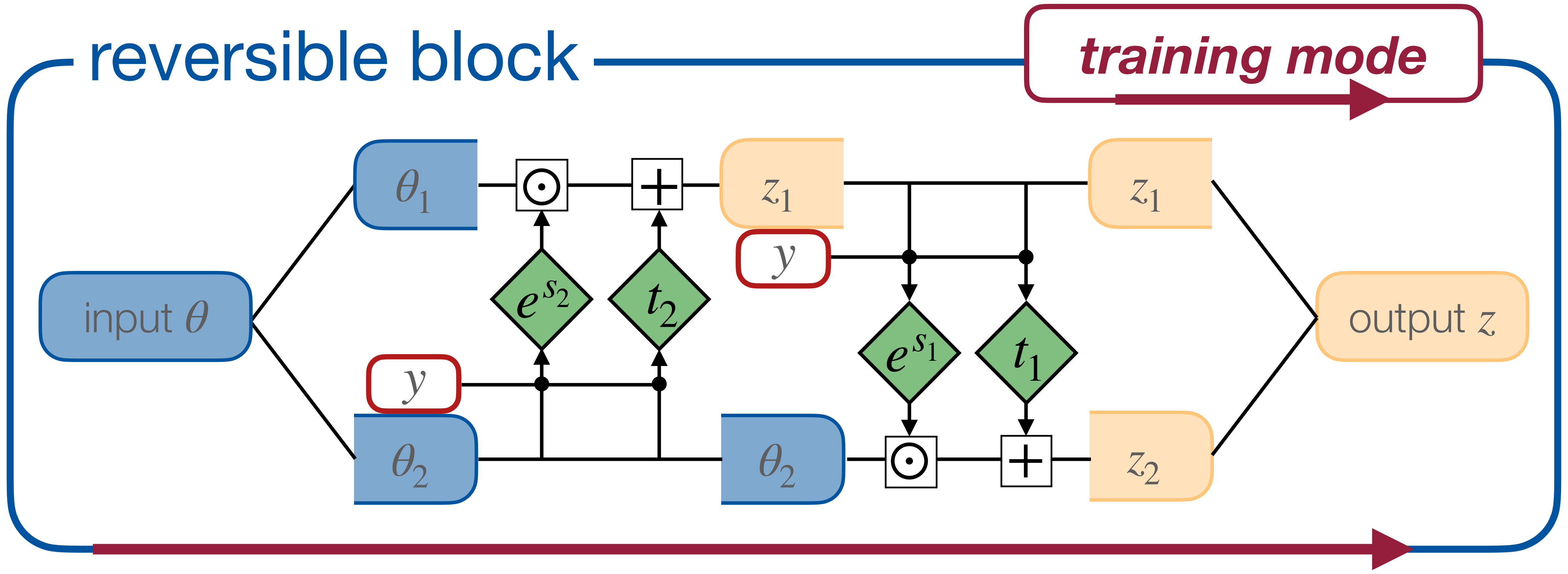
element-wise multiplication

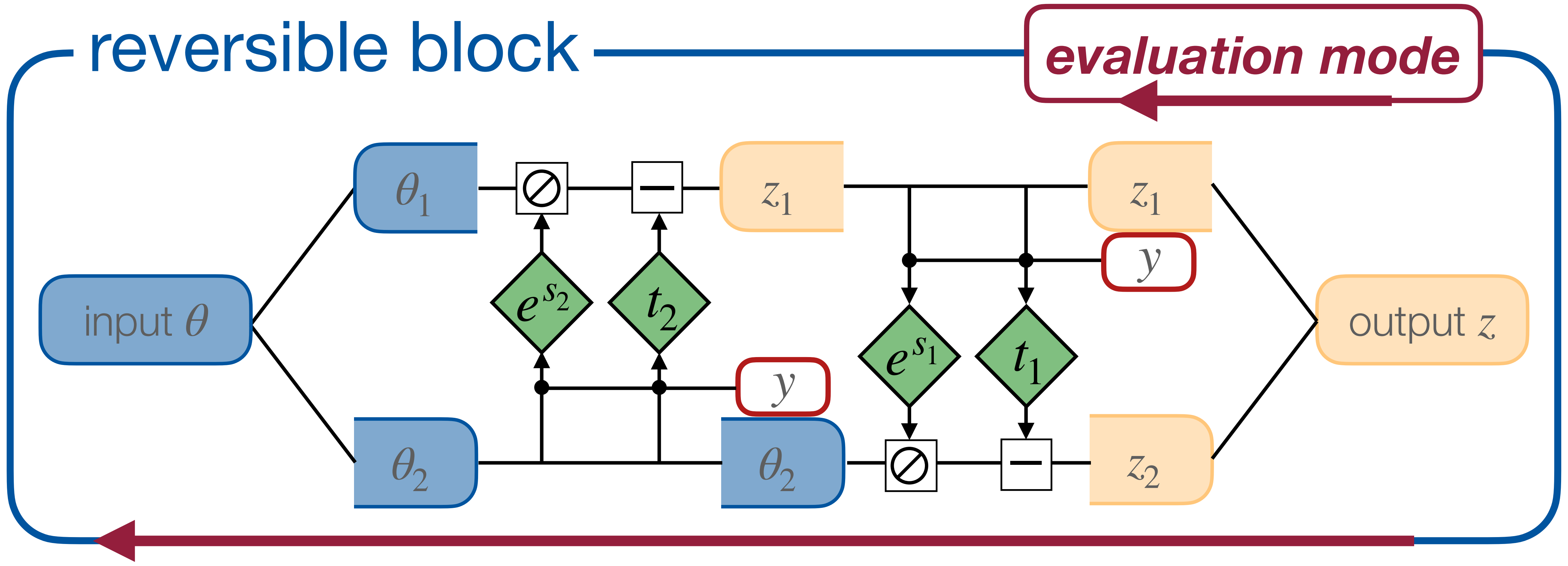
$$\theta_2 = (\mathbf{z}_2 - t_1(\mathbf{z}_1)) \odot \exp(-s_1(\mathbf{z}_1))$$
$$\theta_1 = (\mathbf{z}_1 - t_2(\theta_2)) \odot \exp(-s_2(\theta_2))$$



based on *real-valued non-volume preserving* (**Real NVP**) architecture (arXiv:1605.08803) & **GLOW** (arXiv:1807.03039)

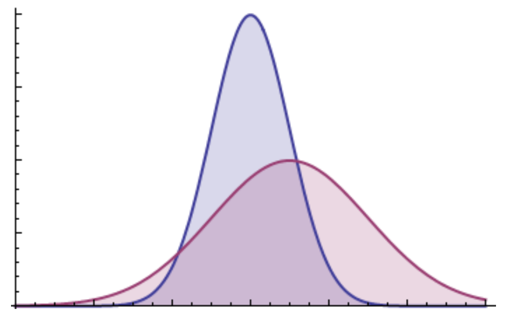
Using the observation y as *condition*





Suitable loss function

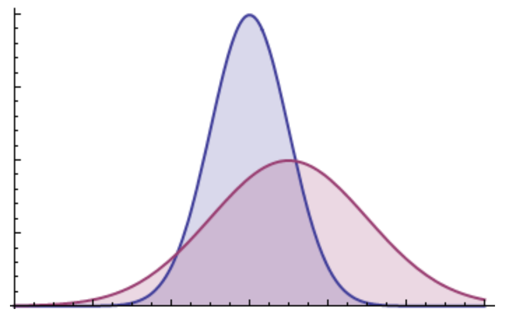
Kullback-Leibler divergence: *true posterior* $p(\theta\,|\,\mathbf{y})$ & *cINN posterior* $p_\phi(\theta\,|\,\mathbf{y})$



- $\mathscr{L} = \mathbb{KL}\big(p(\theta\,|\,\mathbf{y}) \,\|\, p_\phi(\theta\,|\,\mathbf{y})\big) = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\big(\log p(\theta\,|\,\mathbf{y}) - \log p_\phi(\theta\,|\,\mathbf{y})\big) = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\big(-\log p_\phi(\theta\,|\,\mathbf{y})\big)$

Suitable loss function

Kullback-Leibler divergence: *true posterior* $p(\theta\,|\,\mathbf{y})$ & *cINN posterior* $p_\phi(\theta\,|\,\mathbf{y})$



- $\mathscr{L} = \mathbb{KL}\big(p(\theta\,|\,\mathbf{y}) \,\|\, p_\phi(\theta\,|\,\mathbf{y})\big) = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\big(\log p(\theta\,|\,\mathbf{y}) - \log p_\phi(\theta\,|\,\mathbf{y})\big) = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\big(-\log p_\phi(\theta\,|\,\mathbf{y})\big)$

probability conservation: $p_\phi(\theta\,|\,\mathbf{y})\,\mathrm{d}\theta = p(z)\,\mathrm{d}z$

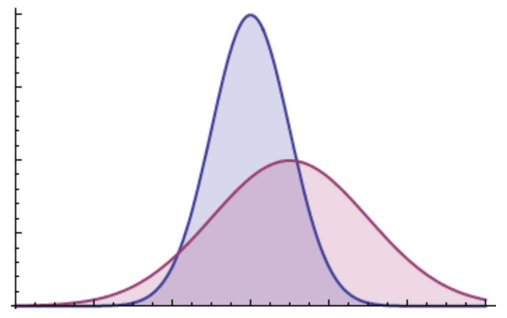
- $\mathscr{L} = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\big(-\log p_\phi(\theta\,|\,\mathbf{y})\big) = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\Big(\log\big(p(\mathbf{z})\cdot\big|\det\left(\frac{\partial\mathbf{z}}{\partial\theta}\right)\big|\big)\Big) = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\Big(\log\big(p(\mathbf{z})\big) + \log\big(\big|\det\left(\frac{\partial\mathbf{z}}{\partial\theta}\right)\big|\big)\Big)$

Using Gaussian distribution for latents z & Jacobian of reversible blocks = triangular matrix, where $\mathbf{z} = f(\theta)$

- $\mathscr{L} = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\big(\frac{1}{2}\,|f(\theta)|^2 + \sum s(\theta)\big)$

Suitable loss function

Kullback-Leibler divergence: *true posterior* $p(\theta\,|\,\mathbf{y})$ & *cINN posterior* $\boxed{p_\phi(\theta\,|\,\mathbf{y})}$



- $\mathcal{L} = \mathbb{KL}\big(p(\theta\,|\,\mathbf{y}) \,\|\, p_\phi(\theta\,|\,\mathbf{y})\big) = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\big(\log p(\theta\,|\,\mathbf{y}) - \log p_\phi(\theta\,|\,\mathbf{y})\big) = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\big(-\log p_\phi(\theta\,|\,\mathbf{y})\big)$

probability conservation: $\boxed{p_\phi(\theta\,|\,\mathbf{y})\,\mathrm{d}\theta} = \boxed{p(z)\,\mathrm{d}z}$

- $\mathcal{L} = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\big(-\log p_\phi(\theta\,|\,\mathbf{y})\big) = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\Big(\log\Big(p(\mathbf{z})\cdot\Big|\det\left(\frac{\partial\,\mathbf{z}}{\partial\theta}\right)\Big|\Big)\Big) = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\Big(\log\big(p(\mathbf{z})\big) + \log\Big(\Big|\det\left(\frac{\partial\,\mathbf{z}}{\partial\theta}\right)\Big|\Big)\Big)$

Using Gaussian distribution for latents z & Jacobian of reversible blocks = triangular matrix, where $\mathbf{z} = f(\theta)$

- $\mathcal{L} = \mathbb{E}_{\theta\sim p(\theta|\mathbf{y})}\Big(\frac{1}{2}\,|f(\theta)|^2 + \sum s(\theta)\Big)$

reversible block

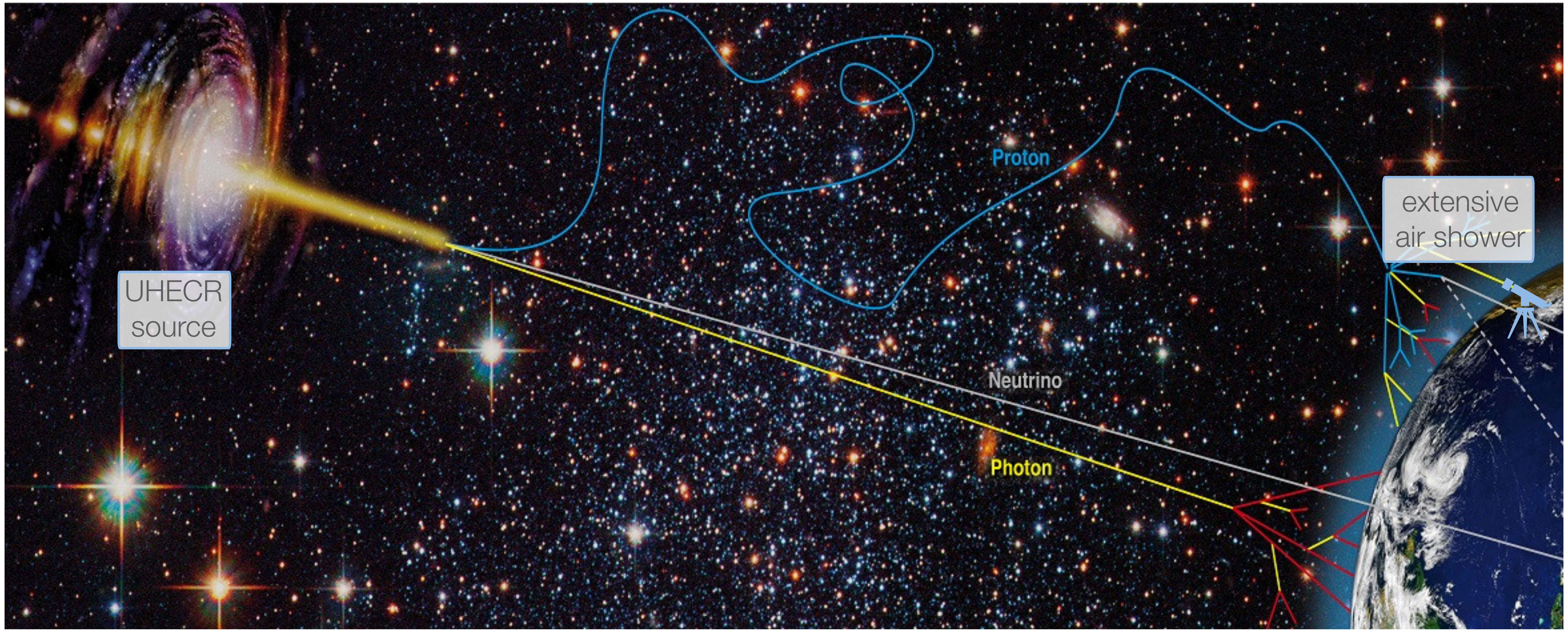
Jacobian

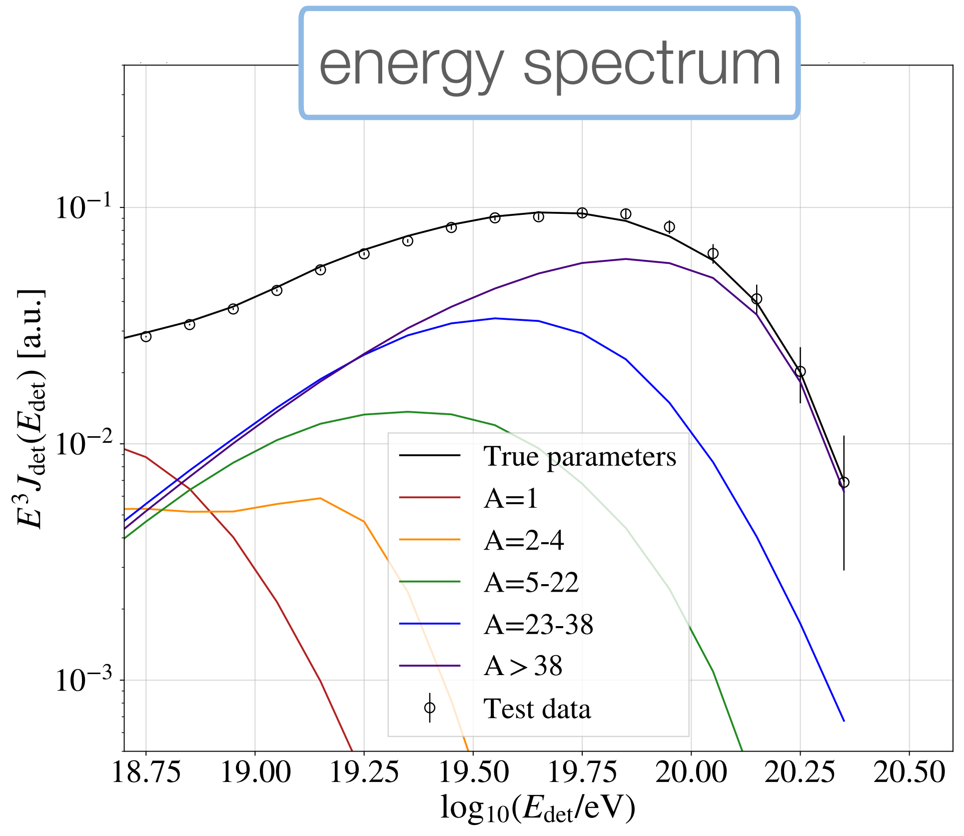
- example: $f_1(\theta) = \begin{cases} \mathbf{z}_1 = \theta_1 \odot \exp(s_2(\theta_2)) + t_2(\theta_2) \\ \theta_2 = \theta_2 \end{cases}$, $\dfrac{\partial f_1(\theta)}{\partial\theta} = \begin{pmatrix} \frac{\partial \mathbf{z}_1}{\partial\theta_1} & \frac{\partial \mathbf{z}_1}{\partial\theta_2} \\ \frac{\partial \theta_2}{\partial\theta_1} & \frac{\partial \theta_2}{\partial\theta_2} \end{pmatrix} = \begin{pmatrix} \mathrm{diag}(\exp\left(s_2(\theta_2)\right)) & \frac{\partial \mathbf{z}_1}{\partial\theta_2} \\ 0 & \mathbb{1} \end{pmatrix}$

(equivalently for $\mathbf{z}_2$ with $f_2(\theta)$)

- $\left|\det\dfrac{\partial\mathbf{z}}{\partial\theta}\right| = \dfrac{\partial f_1(\theta)}{\partial\theta}\cdot\dfrac{\partial f_2(\theta)}{\partial\theta} = \prod \exp(s_2(\theta_2))\cdot\exp(s_1(\mathbf{z}_1)) = \exp\Big(\sum s_2(\theta_2) + \sum s_1(\mathbf{z}_1)\Big)$

Example application on an astrophysical scenario



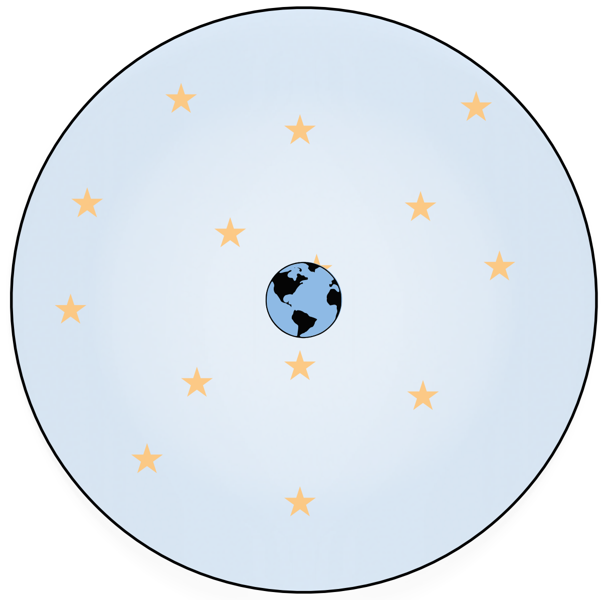


Example Application: Astrophysical setup

Idea

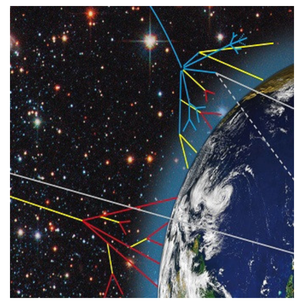
Ultra-high energy cosmic rays (UHECRs) emitted by homogeneously distributed sources

$E_{max} = R_{cut} \cdot Z$



- 5 different injected nuclei: **power-law** energy spectrum with maximum rigidity: $J_{inj}(E) \propto E^{-\gamma} \cdot f_{cut}(E, Z \cdot R_{cut})$

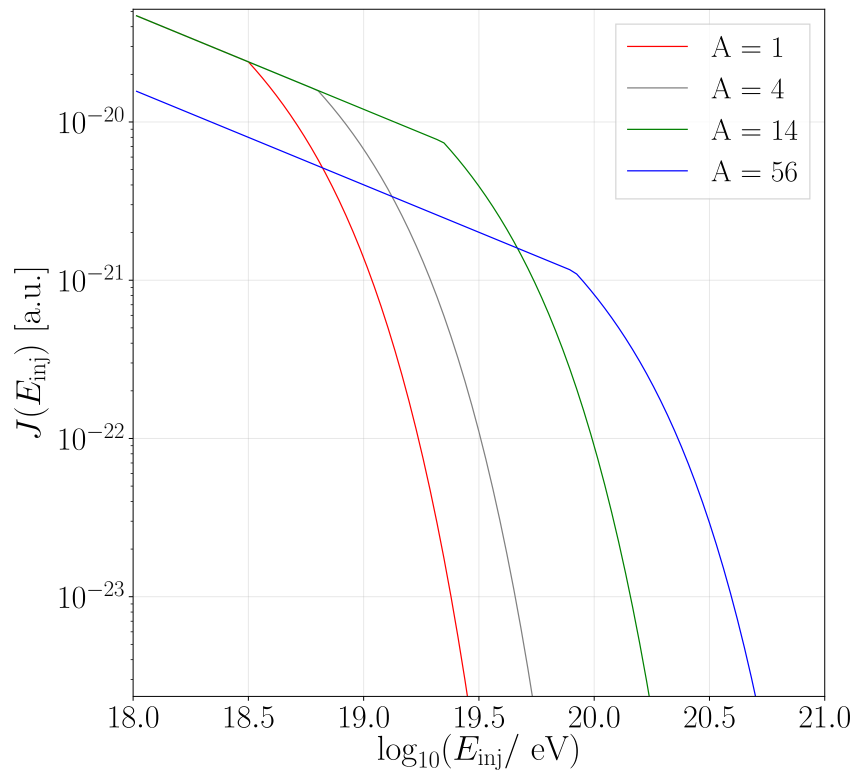
Detected energy spectrum contains only **implicit information**, no explicit inverse function can be formulated.



Goal

Constrain **7 astrophysical source parameters**

- spectral index γ, maximum rigidity R_{cut}, 5 mass fractions $a(\mathrm{A}_i)$

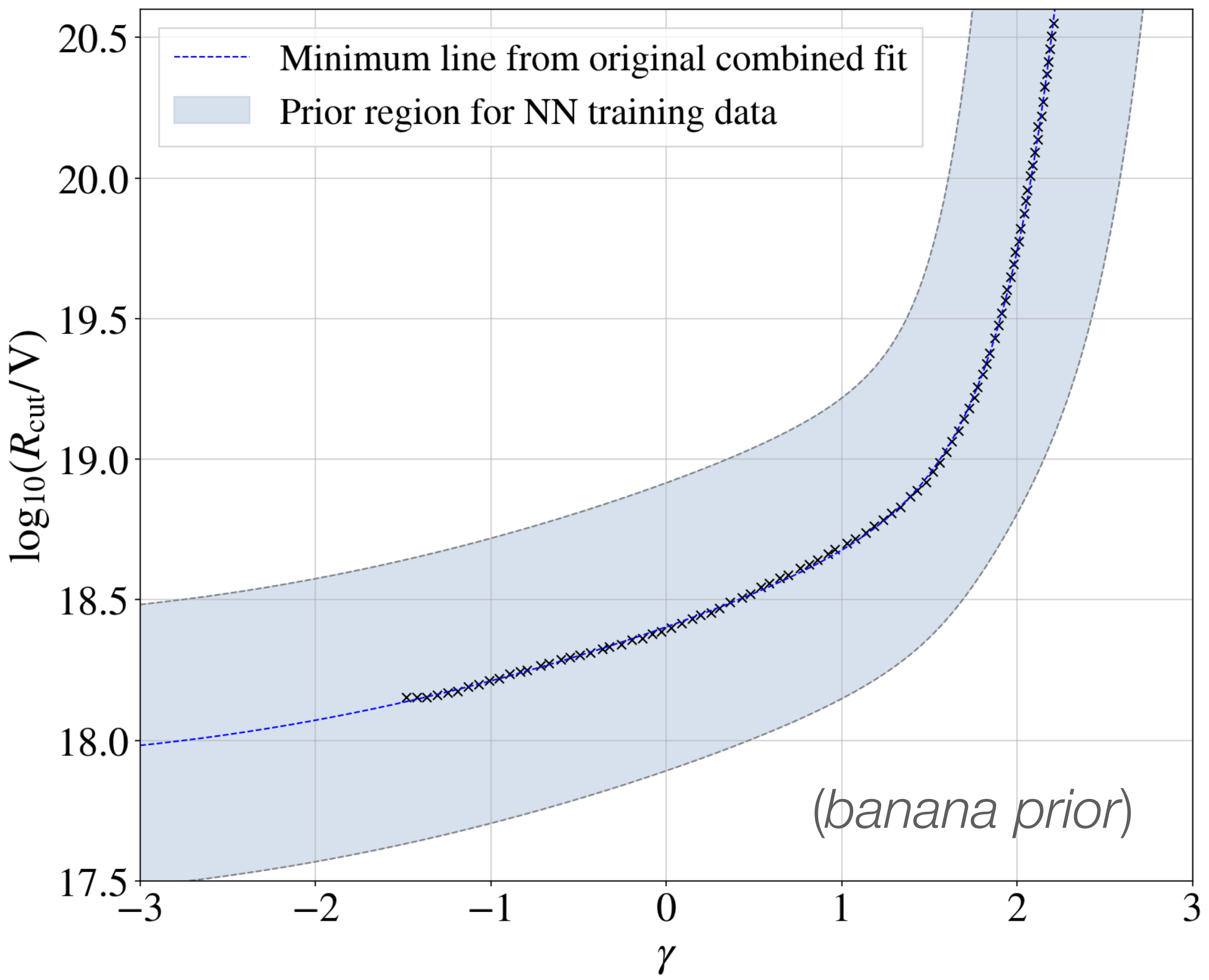


Training data

Source

Prior *(on source parameters)* for **training data set** *(size: 100.000)*

- spectral parameters γ, R_{cut} from prior analyses

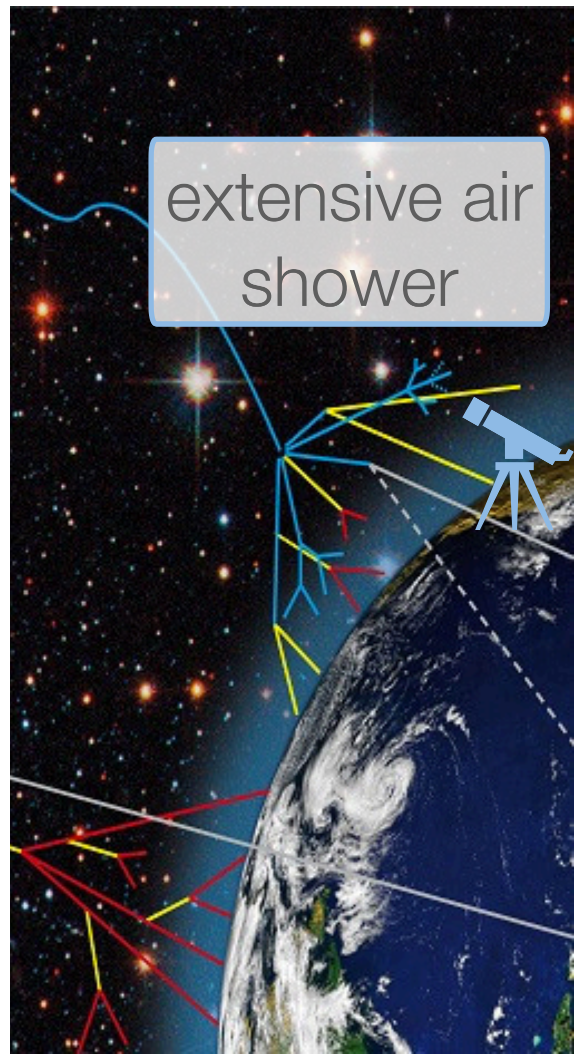


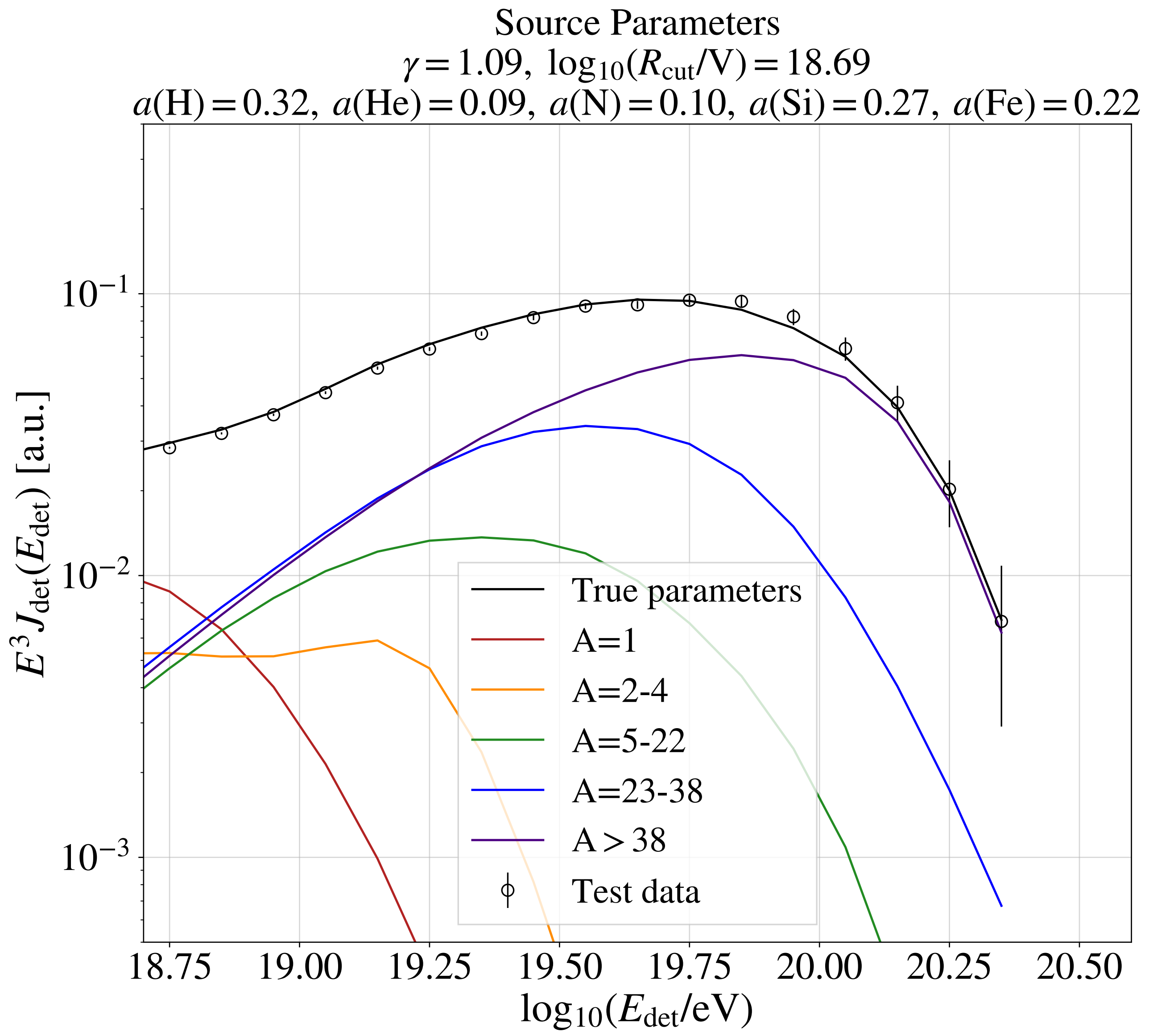
- 5 elements (H, He, N, Si, Fe): unit simplex
  → 7 free model parameters

Observation

Only energy spectrum as observation
binned spectrum multiplied with E^3

obtained from CRPropa 3 simulations



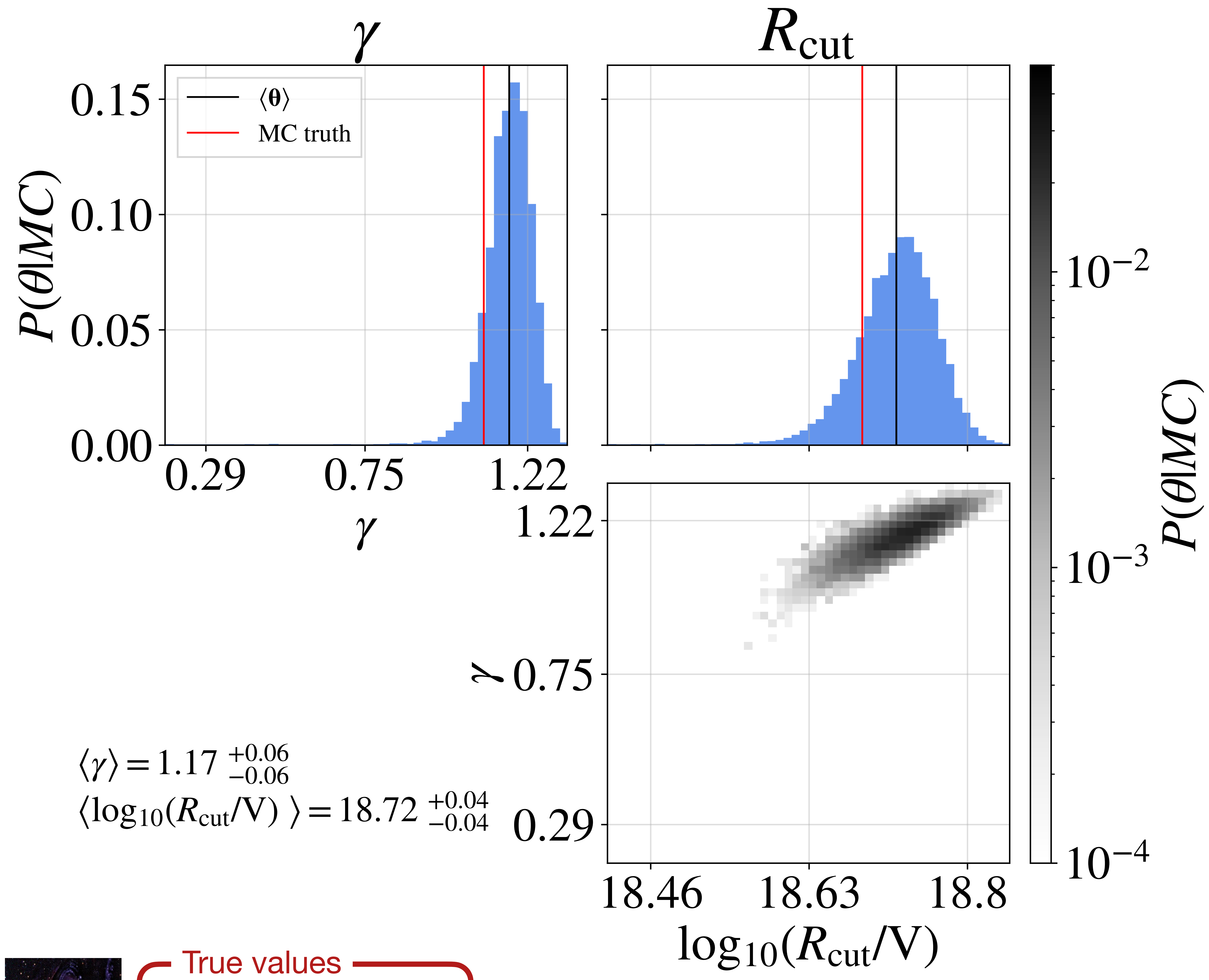


Example observation

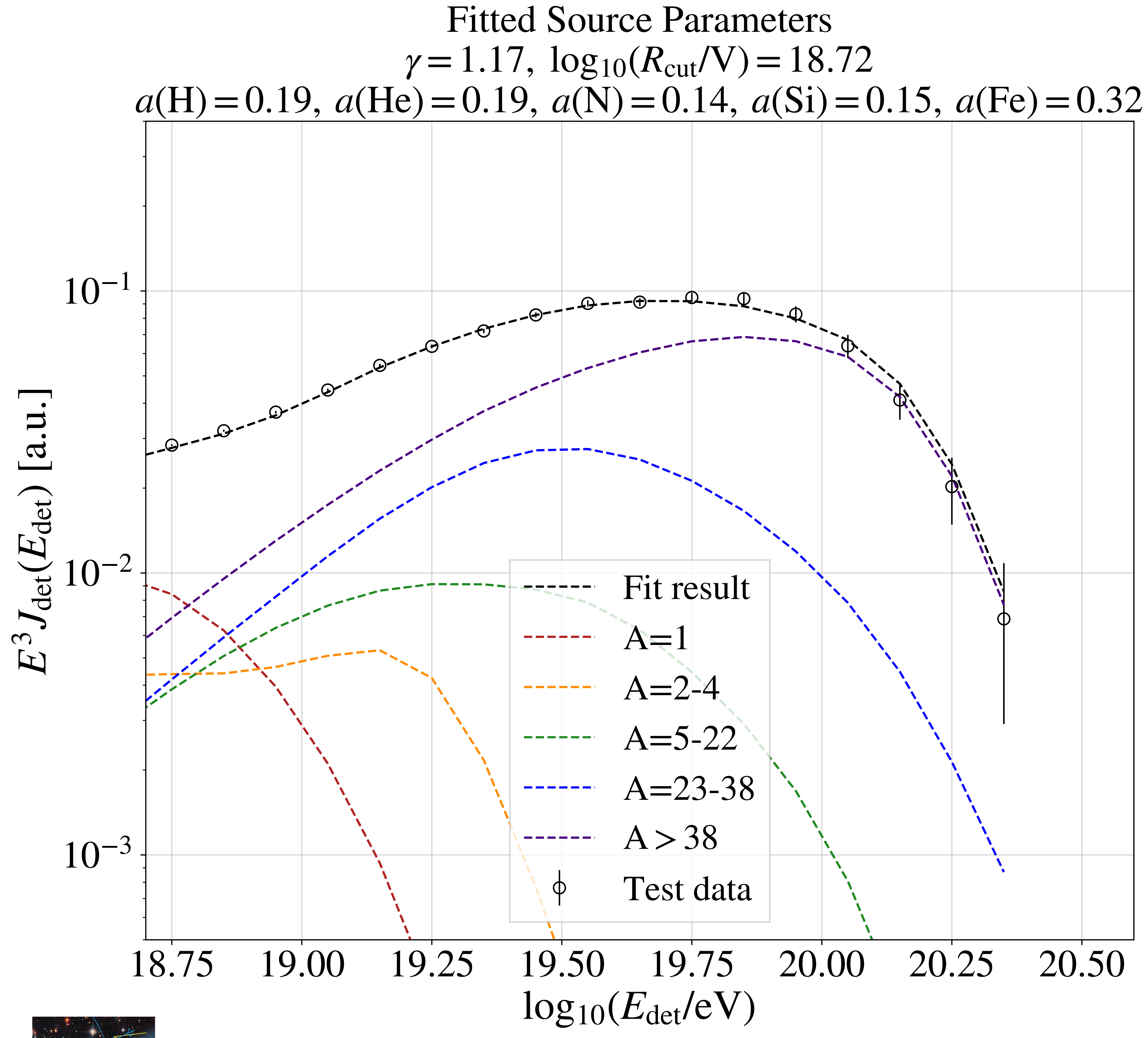
(binned between $10^{18.7}\,\mathrm{eV} - 10^{20.4}\,\mathrm{eV}$ *with bin width* $\log_{10}(E_{det}/\,\mathrm{eV}) = 0.1$*)*

Example scenario: cINN posteriors & prediction

cINN posteriors of source parameters



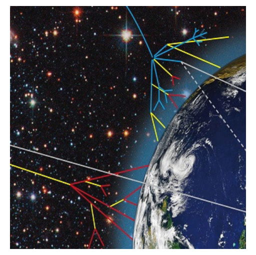
Observables: bin content of energy spectrum



True values

$\gamma = 1.09$

$\log_{10}(R_{\mathrm{cut}}/\mathrm{V}) = 18.69$



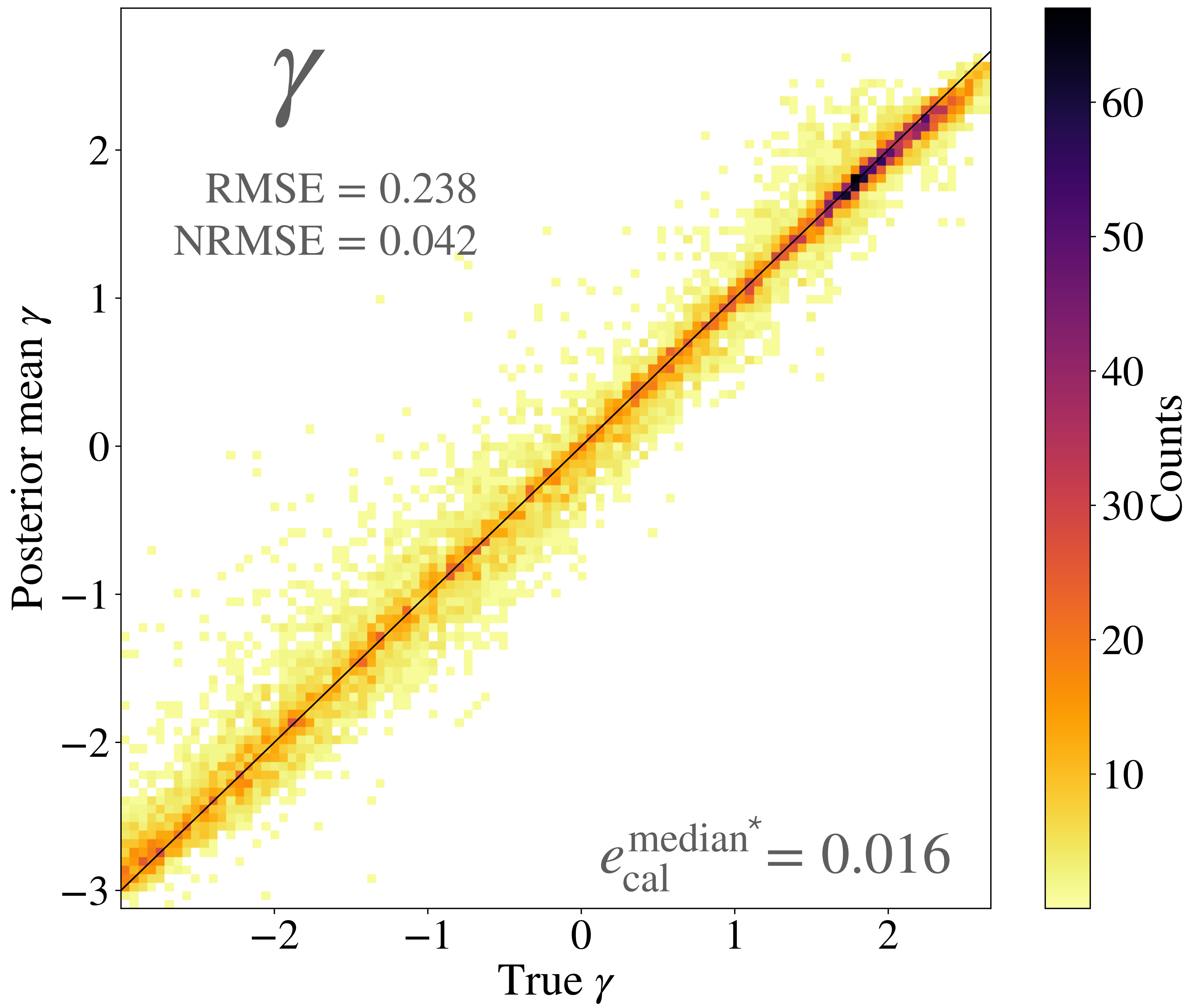
→ cINN recovers source parameters & correlations between them

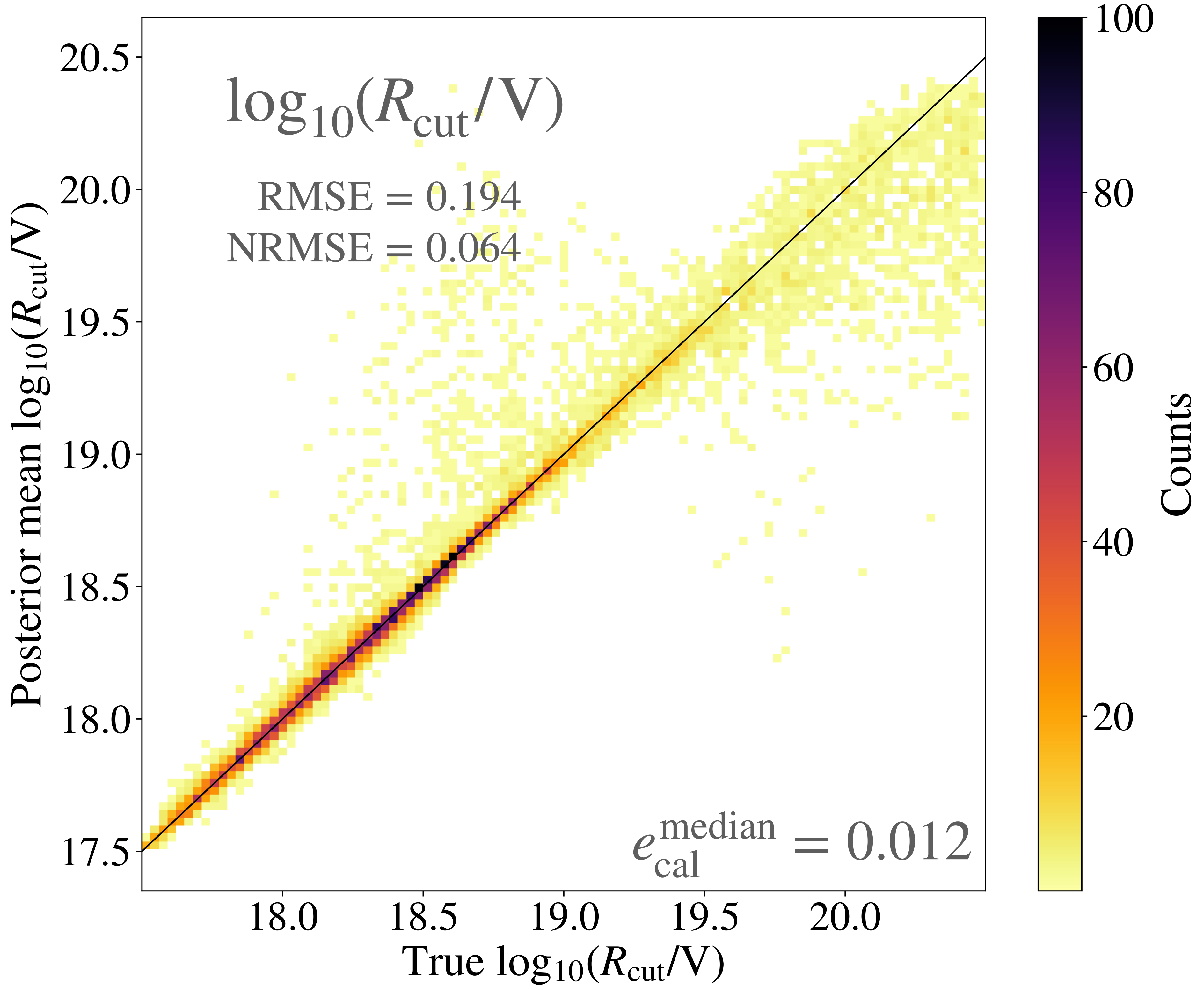
cINN summary statistics

**median over absolute values in range of confidence intervals $(0.01, 0.99)$ in 0.01 steps*

Look at whole test data set *(size 10.000)*

- Histogram **posterior mean vs true simulation value**





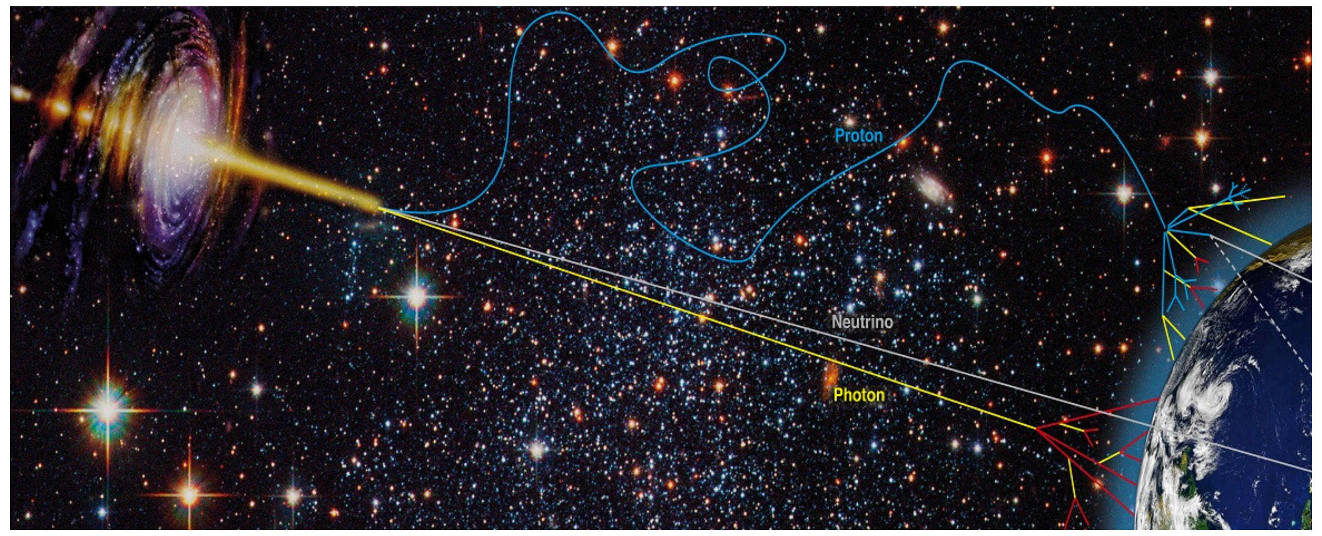
- **Calibration error**: $e_{\mathrm{cal}} = q_{\mathrm{inliers}} - q$ for estimation of correctness of the widths of the posterior distributions
  - confidence interval q & fraction of observations $q_{\mathrm{inliers}} = \frac{N_{\mathrm{inliers}}}{N}$ with true value in q-confidence interval

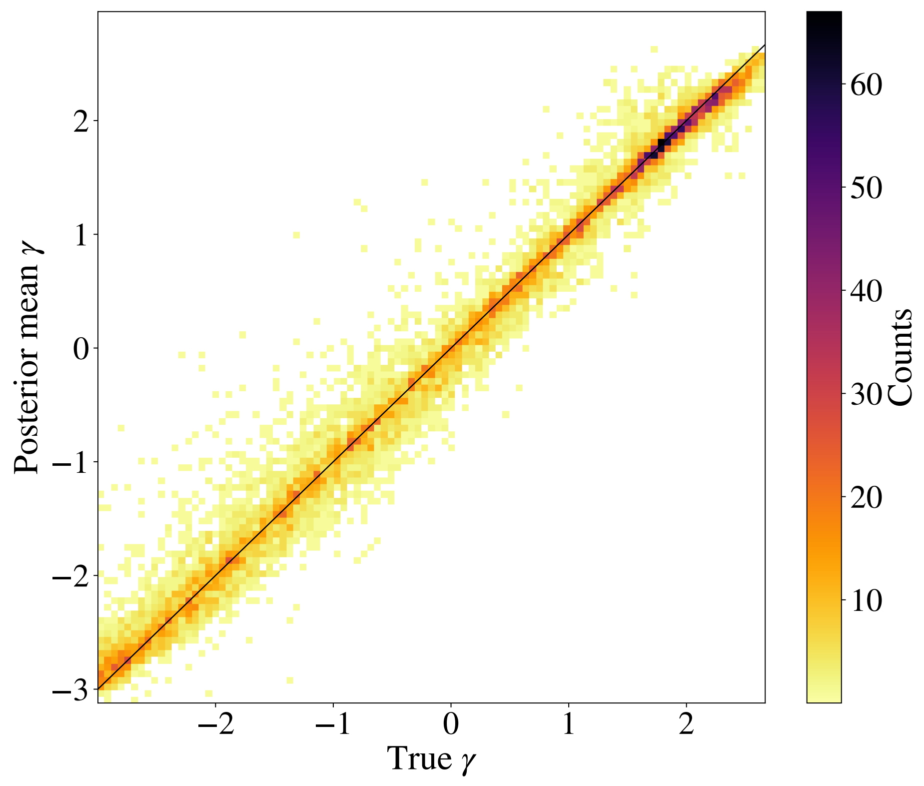
→ **Reconstruction quality of cINN good for many test data sets**

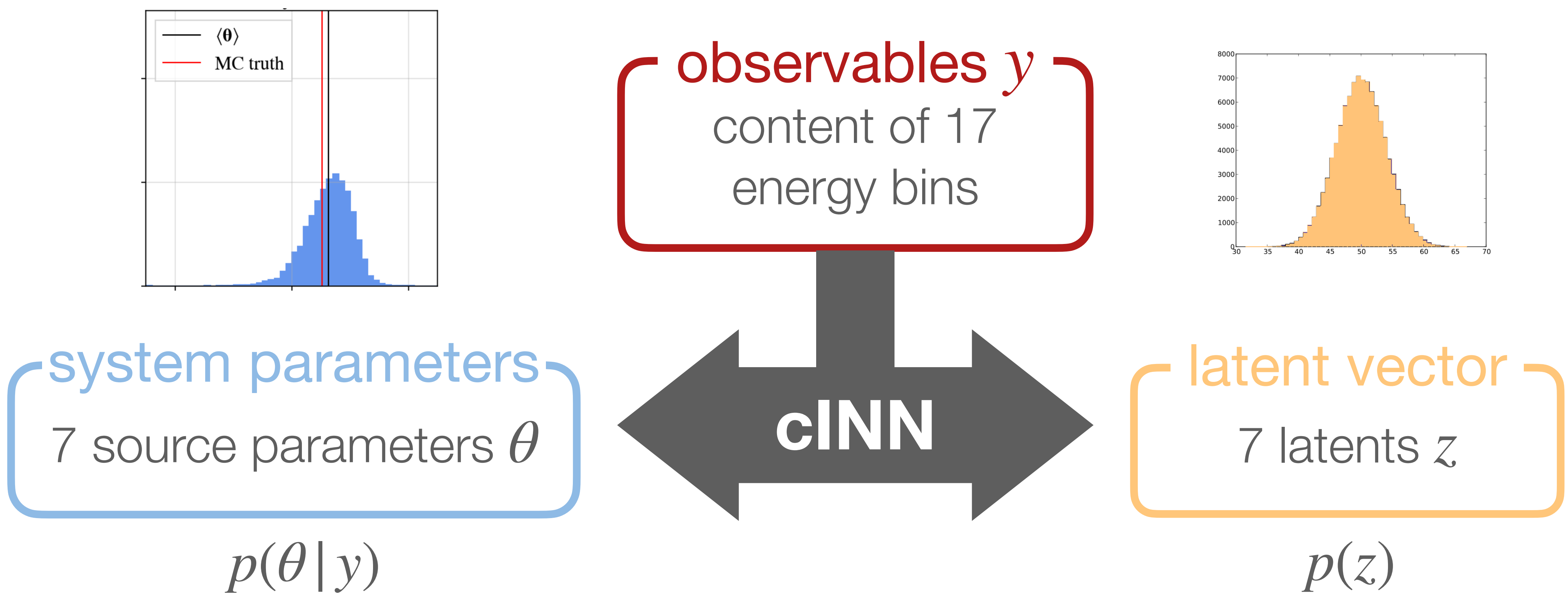
Summary

cINN

- built with **reversible blocks** & trained with suitable loss function
- enables **creation of posterior distributions**
  - estimate of actual value & corresponding uncertainty
  - correlations between parameters
- provides **fast evaluation** of many test data sets
  - training (GPU): ~1h, evaluation: ~s

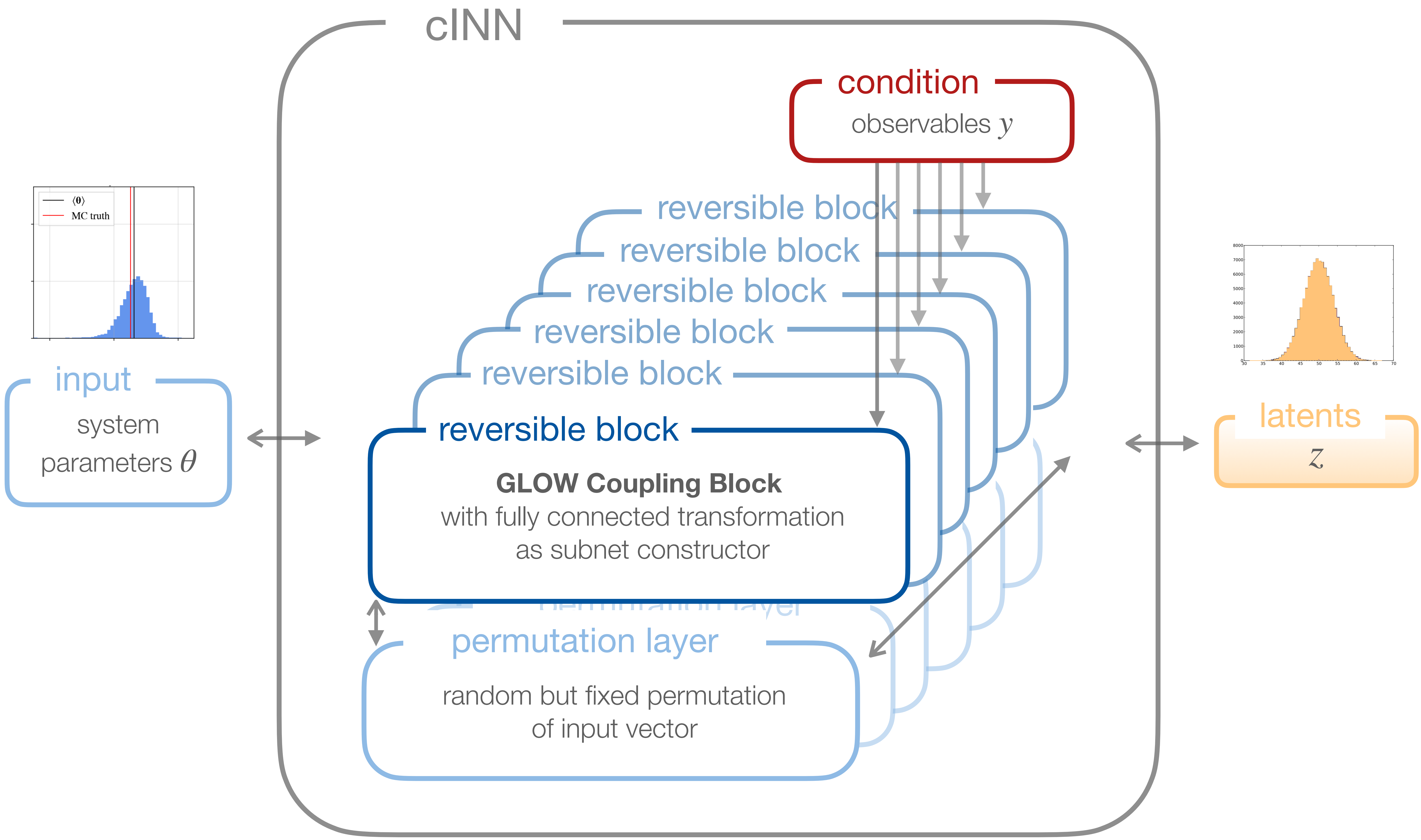






Backup

Conditional Invertible Neural Network structure



Network setup

Network settings

Reversible blocks: 6

GLOW subnetwork structure

fully connected transformation
as subnet constructor (width 256)

Training settings

Initial learning rate: 10^{-3}

Learning rate decay: cosine

Final learning rate: 10^{-5}

Adam optimizer betas: $(0.9,\ 0.999)$

Batch size: 1000

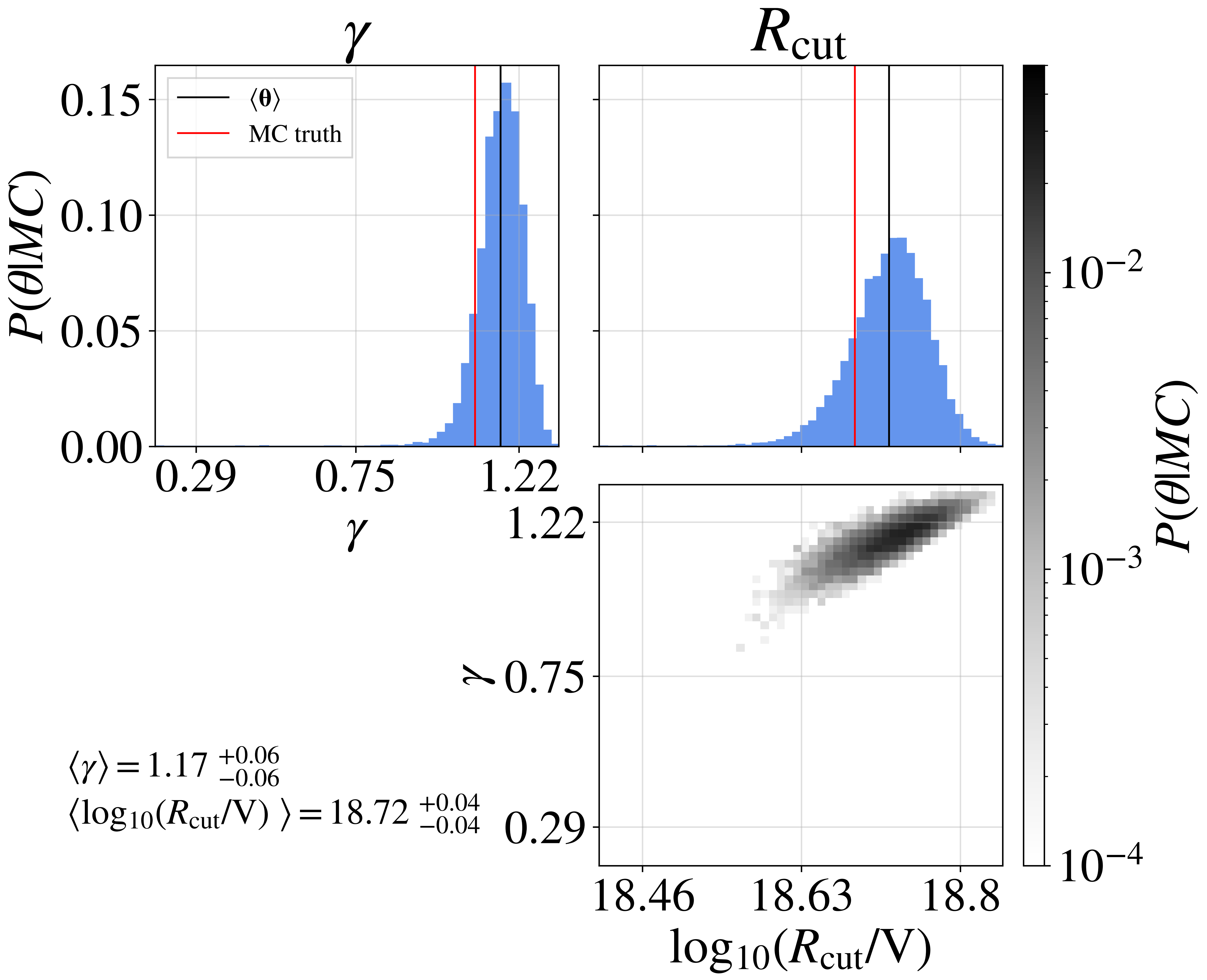
Epochs: 500

subnetworks

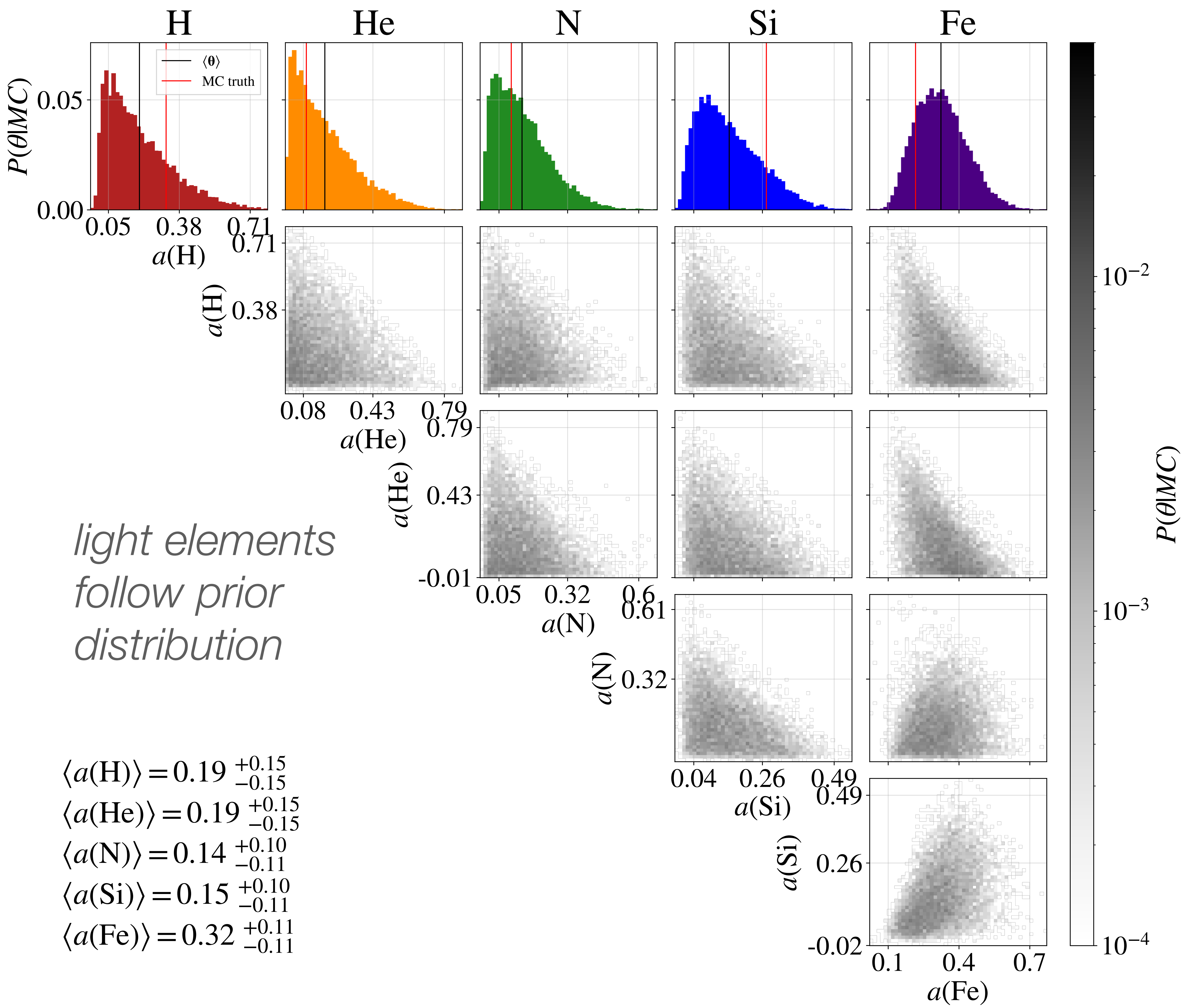
- **fully connected** transformation with internal size (here: 256)
- 3 nodes/ layers with **ReLU activation** of dropout (here: $p_{\text{dropout}} = 0.0$) of **linear transformation**
- last layer: linear transformation only

Example scenario: cINN posteriors with elemental fractions

Posteriors from cINN



$\langle\gamma\rangle = 1.17\,^{+0.06}_{-0.06}$
$\langle\log_{10}(R_{\mathrm{cut}}/\mathrm{V})\rangle = 18.72\,^{+0.04}_{-0.04}$

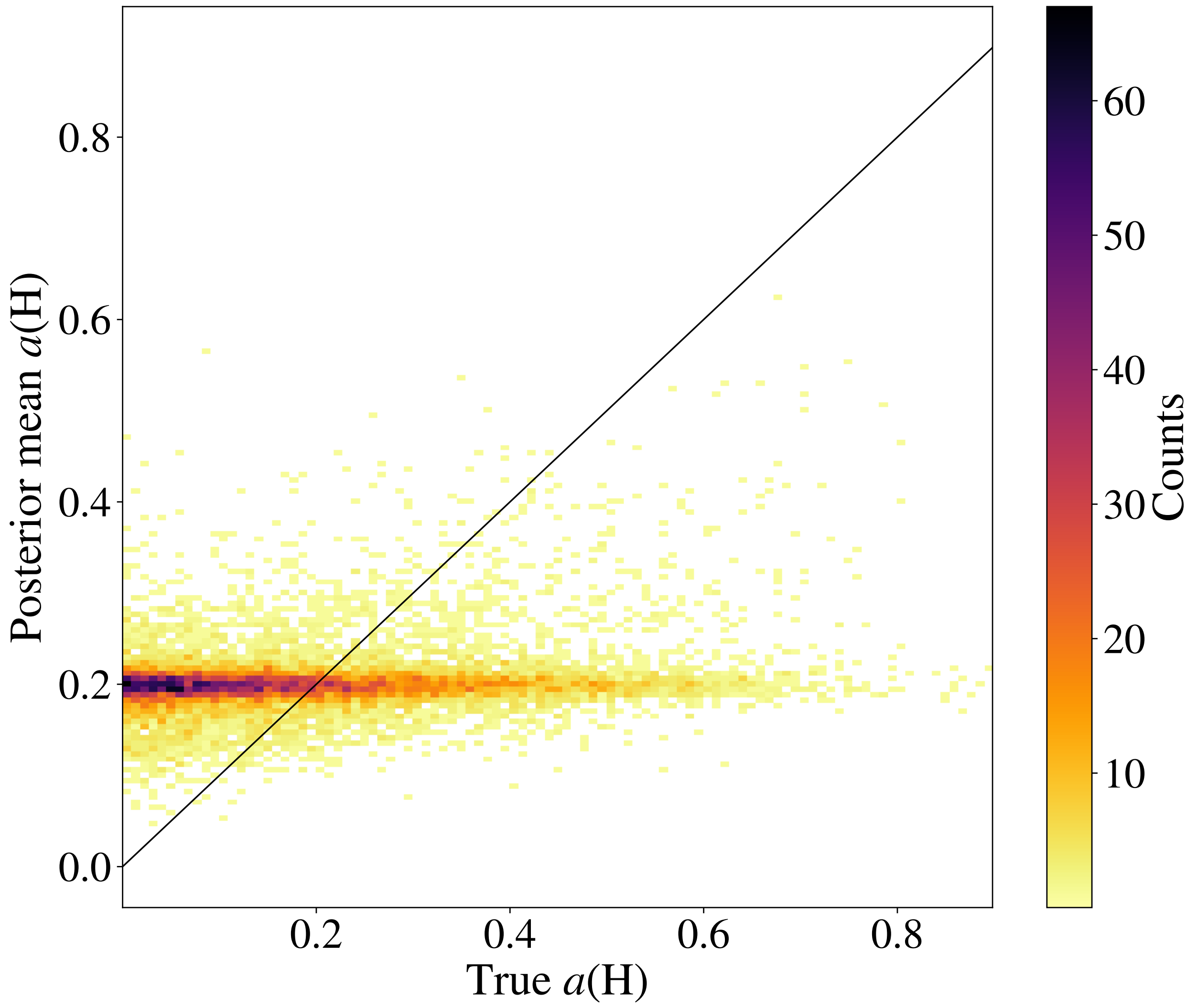


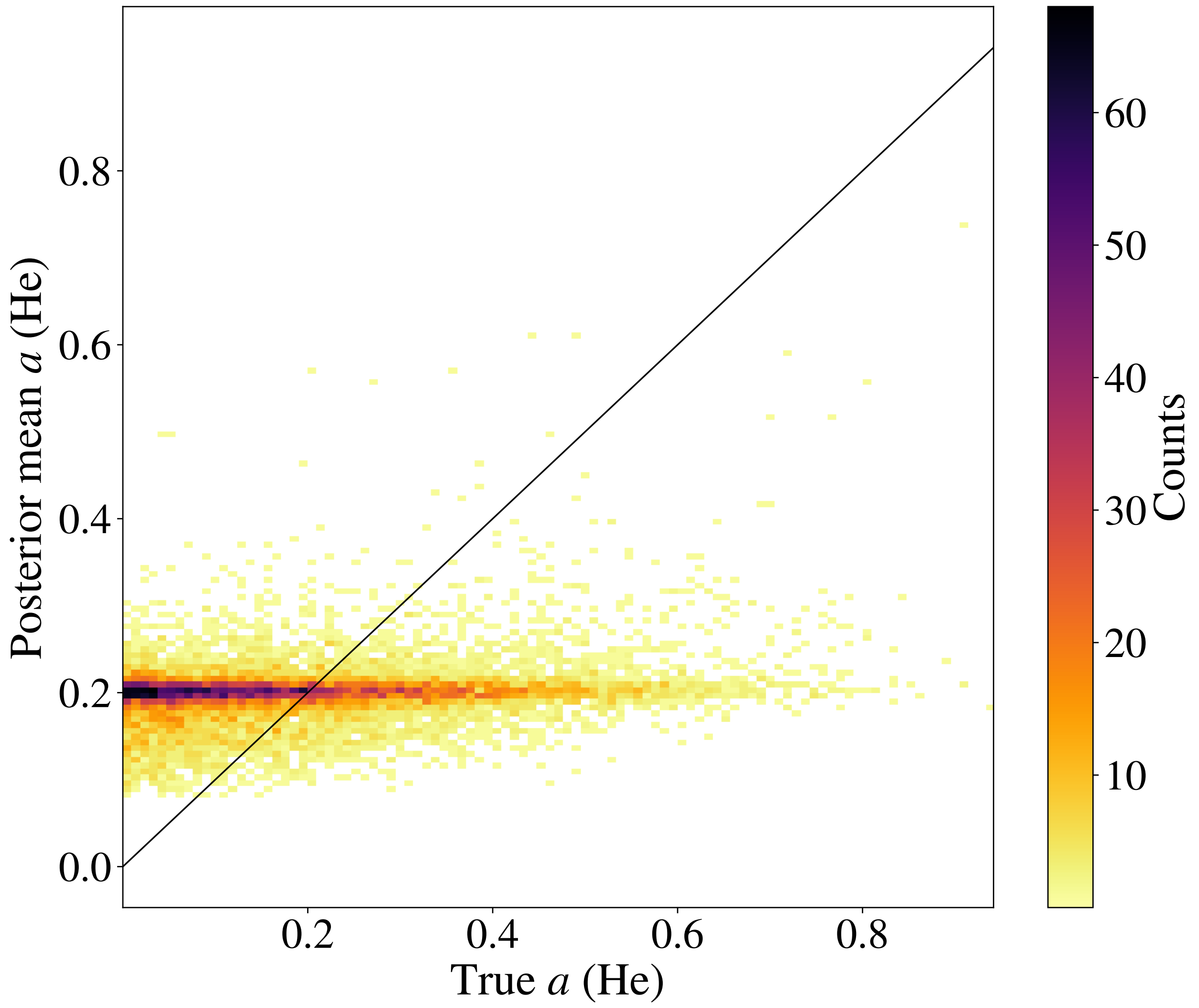
light elements follow prior distribution

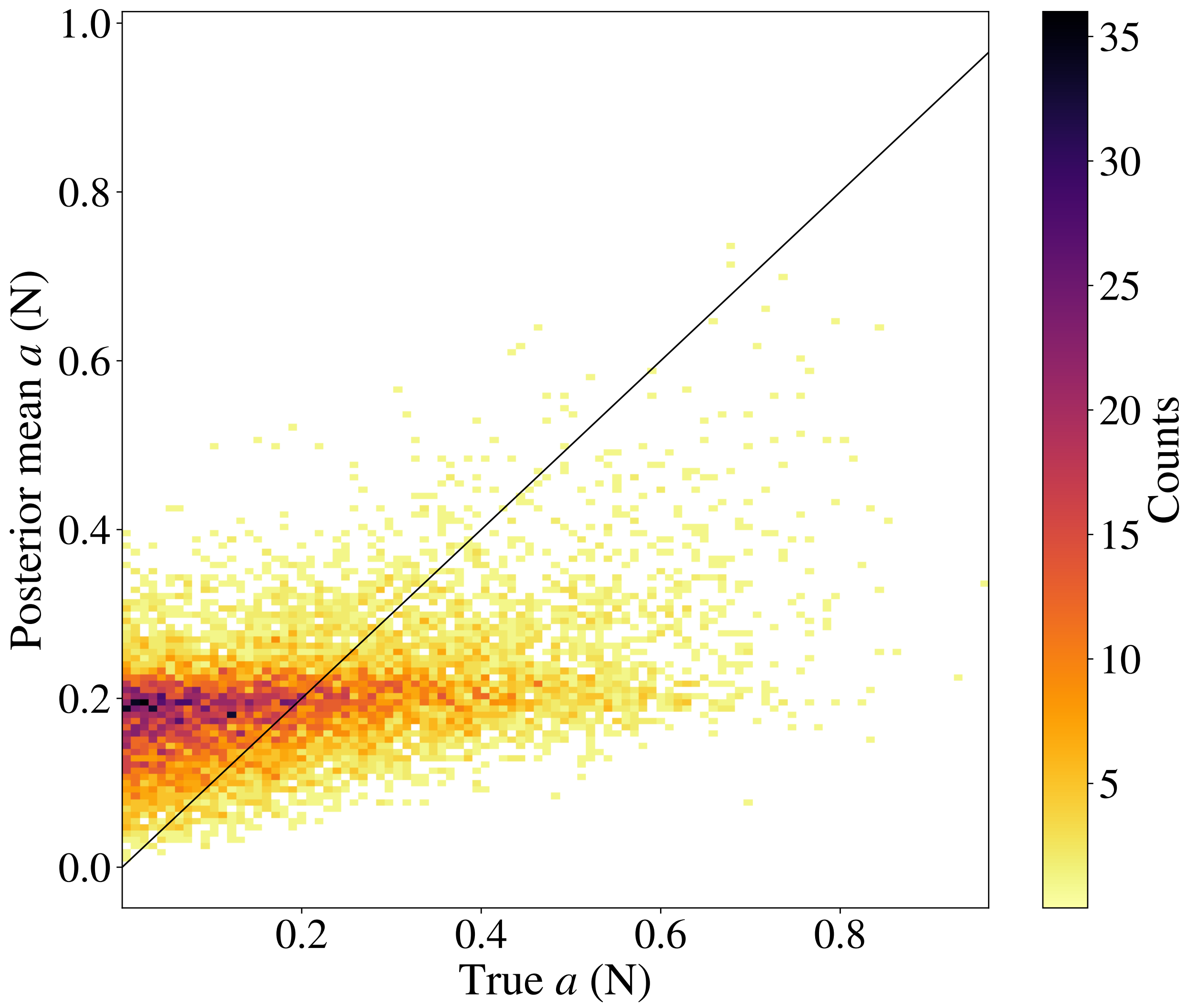
$\langle a(\mathrm{H})\rangle = 0.19\,^{+0.15}_{-0.15}$
$\langle a(\mathrm{He})\rangle = 0.19\,^{+0.15}_{-0.15}$
$\langle a(\mathrm{N})\rangle = 0.14\,^{+0.10}_{-0.11}$
$\langle a(\mathrm{Si})\rangle = 0.15\,^{+0.10}_{-0.11}$
$\langle a(\mathrm{Fe})\rangle = 0.32\,^{+0.11}_{-0.11}$

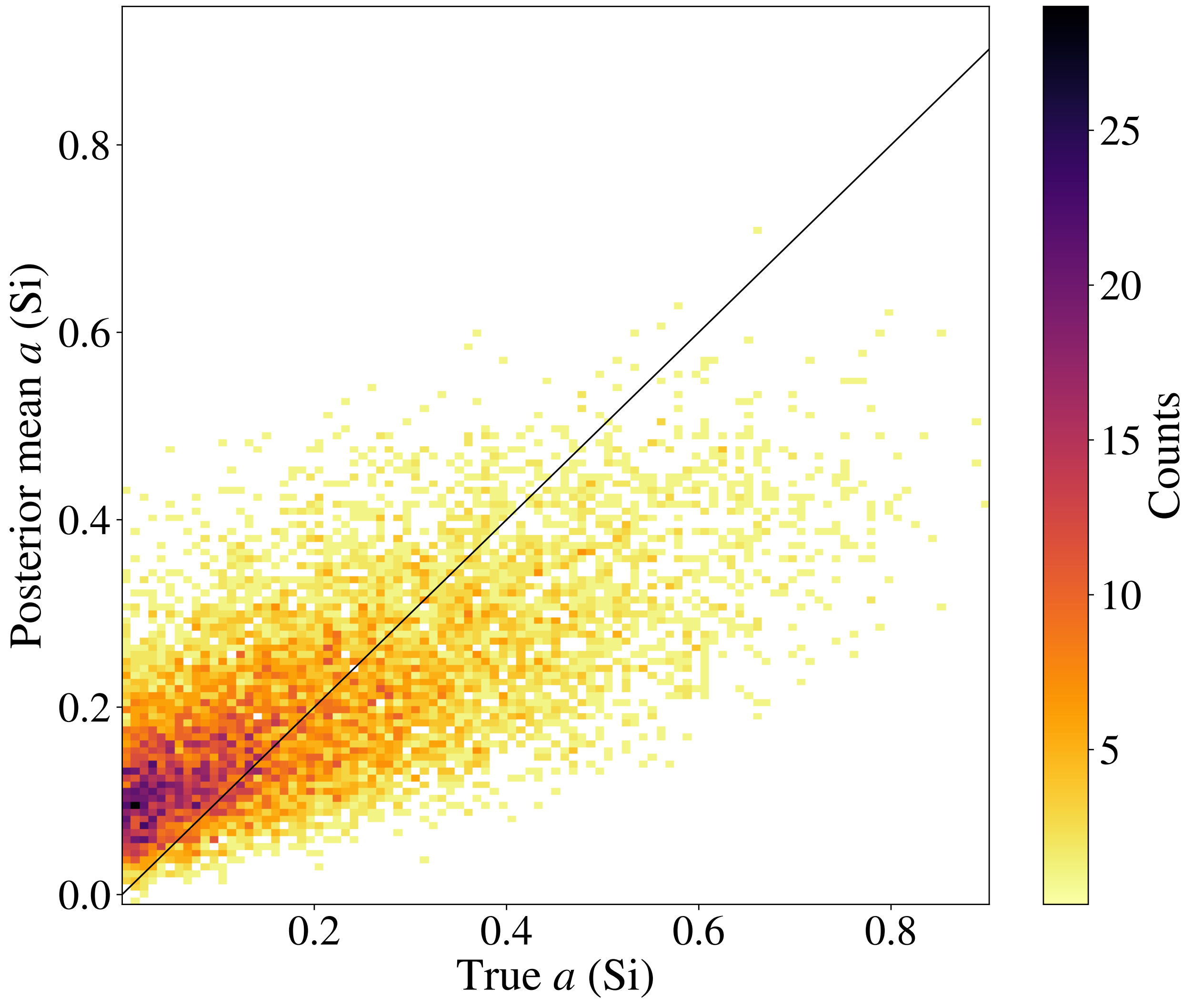
→ constraining power on elemental fractions at the sources only from energy spectrum not very high

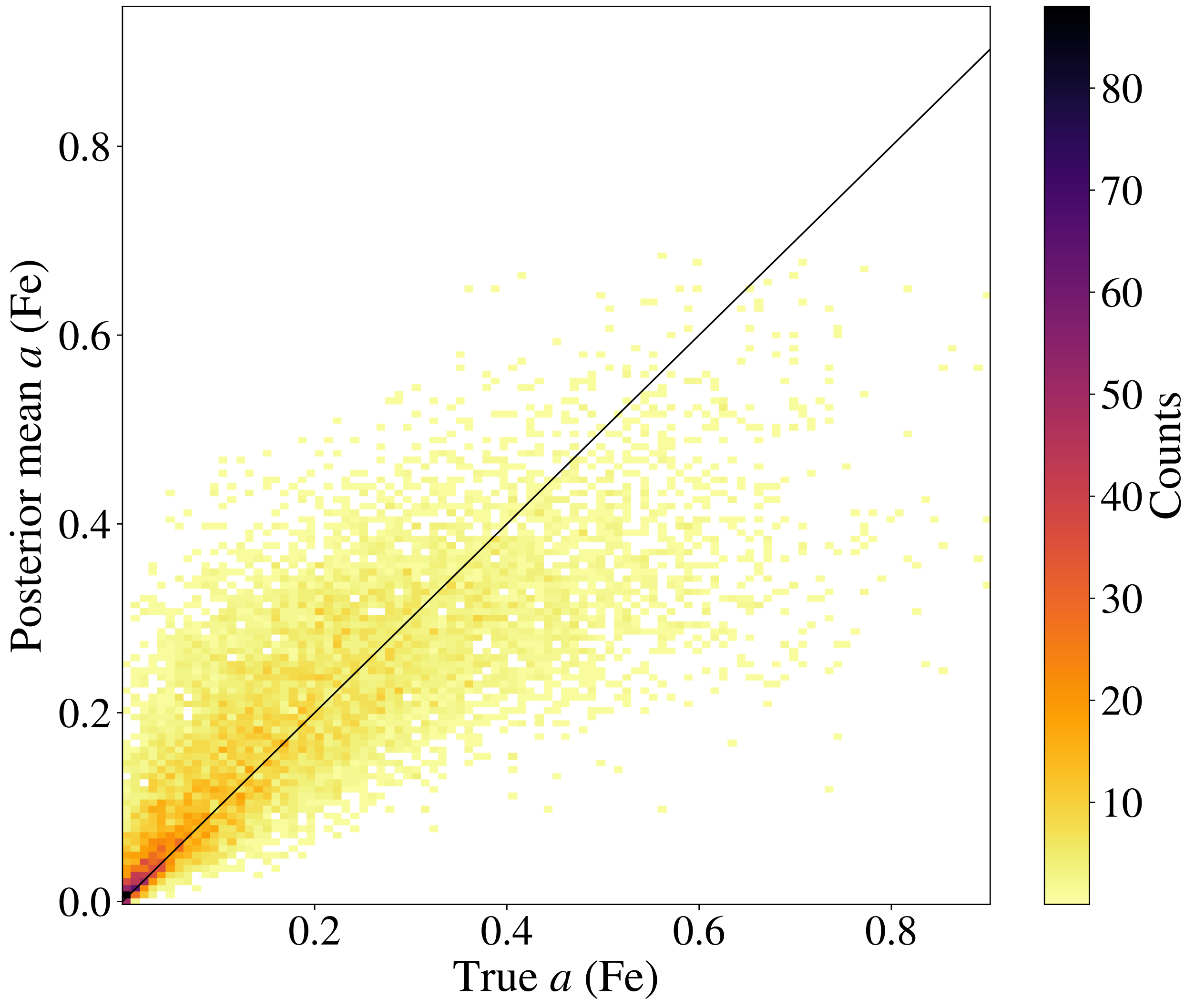
cINN summary statistics: elemental fractions











→ constraining power on fractions at the sources only from energy spectrum not very high

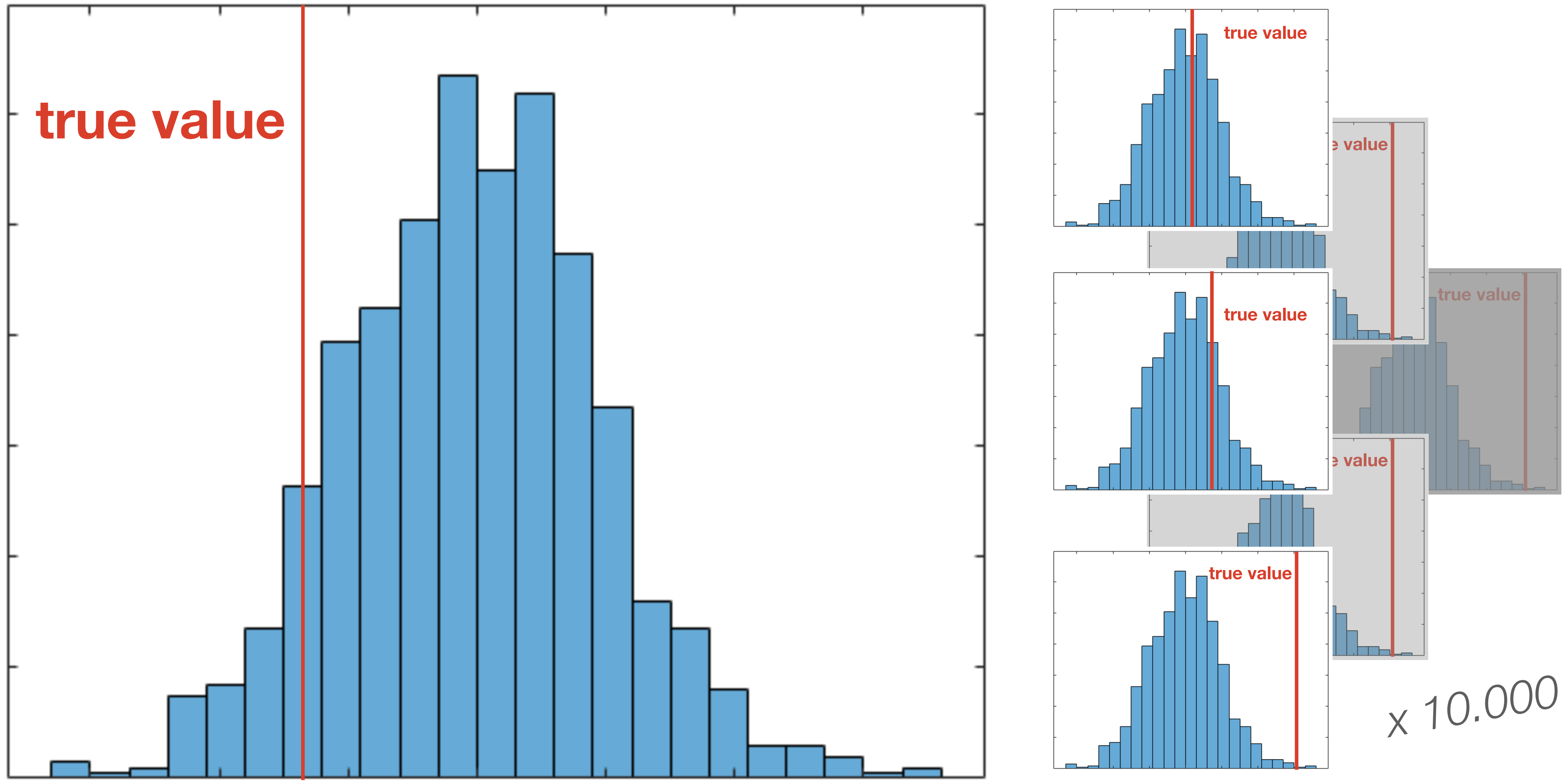
→ contributions of heavy elements more prominent in energy spectrum

Calibration error

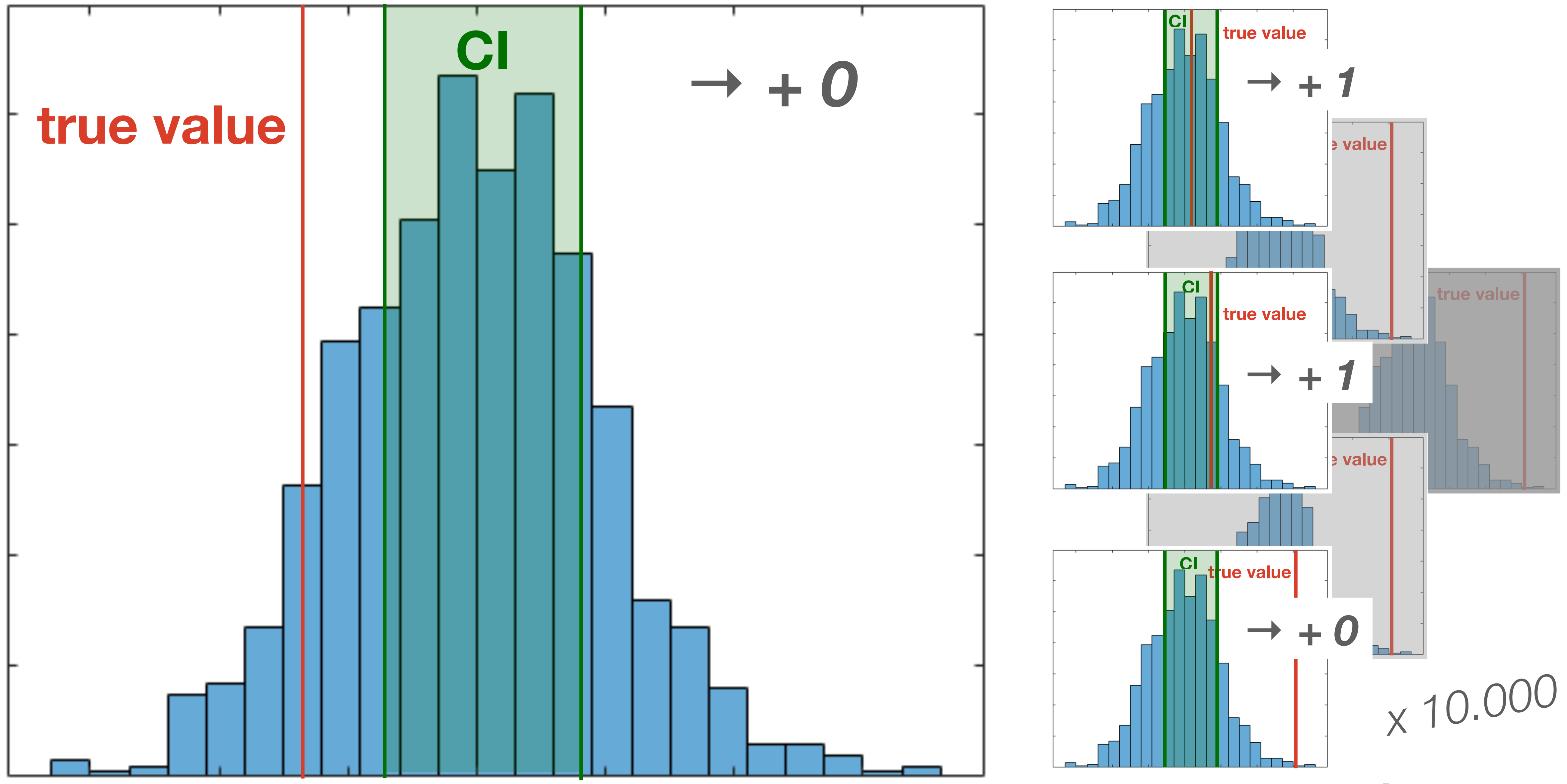
For estimation of correctness of the widths of the posterior distributions

- $e_{\text{cal}} = q_{\text{inliers}} - q$
  - confidence interval q & fraction of observations $q_{\text{inliers}} = \dfrac{N_{\text{inliers}}}{N}$ with true value in q-confidence interval
- $e_{\text{cal}}^{\text{med}}$: median over range of confidence intervals $(0.01, 0.99)$ in 0.01 steps

For each free parameter:
Evaluate posteriors for all test data sets (10.000)



Count for how many test sets the true value lies in the q-confidence interval



→ ***e.g. = 5.108/10.000***