EUROPEAN AI FOR FUNDAMENTAL PHYSICS CONFERENCE
EuCAIFCon 2025

Tests for model misspecification in simulation-based inference

Noemi Anau Montel

17 June 2025

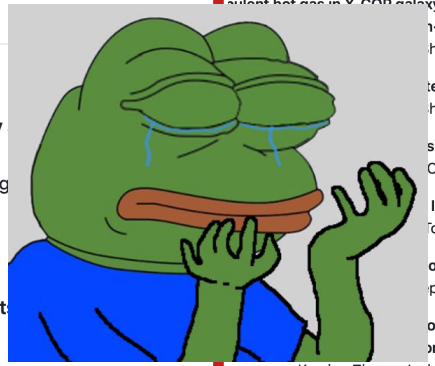
MAX PLANCK INSTITUTE
FOR ASTROPHYSICS

Simulation-based inference

Application to Real Data

Applications of neural simulation-based inference beyond synthetic data.

- **A robust neural determination of the source-count distribution of the Fermi-LAT sky** [arXiv]
  Christopher Eckner, Noemi Anau Montel, Florian List, Francesca Calore, Christoph Weniger
- **SimBIG : A Forward Modeling Approach To Analyzing Galaxy Clustering** [arXiv]
  ChangHoon Hahn et al
- **Mental speed is high until age 60 as revealed by analysis of over a million participants**
  Krause, Stefan T. Radev, Andreas Voss
- **A neural simulation-based inference approach for characterizing the Galactic Center γ-ray excess** [arXiv]
  Siddharth Mishra-Sharma, Kyle Cranmer
- **Towards constraining warm dark matter with stellar streams through neural simulation-based inference** *(Preliminary)* [arXiv]
  Joeri Hermans, Nilanjan Banik, Christoph Weniger, Gianfranco Bertone, Gilles Louppe
- **OutbreakFlow: Model-based Bayesian inference of disease outbreak dynamics with invertible neural networks and its application to the COVID-19 pandemics in Germany** [arXiv]
  Stefan T. Radev, Frederik Graw, Simiao Chen, Nico T. Mutters, Vanessa M. Eichel, Till Bärnighausen, Ullrich Köthe
- **Likelihood-free inference with neural compression of DES SV weak lensing map statistics** [arXiv]
  Niall Jeffrey, Justin Alsing, François Lanusse



Curated list of works: https://github.com/smsharma/awesome-neural-sbi

Model misspecification

Model

$x, \theta \sim p(x \mid \theta) p(\theta)$

Data

x_{obs}

Does my model fit the data?
If not, where and how does it fail?

Outline

Model misspecification *diagnostics*

1. Inference robustness checks
2. Posterior predictive validation
3. Contrastive model diagnostics

Classification from TASI lectures on structured reasoning for SBI, Christoph Weniger [to appear!!]

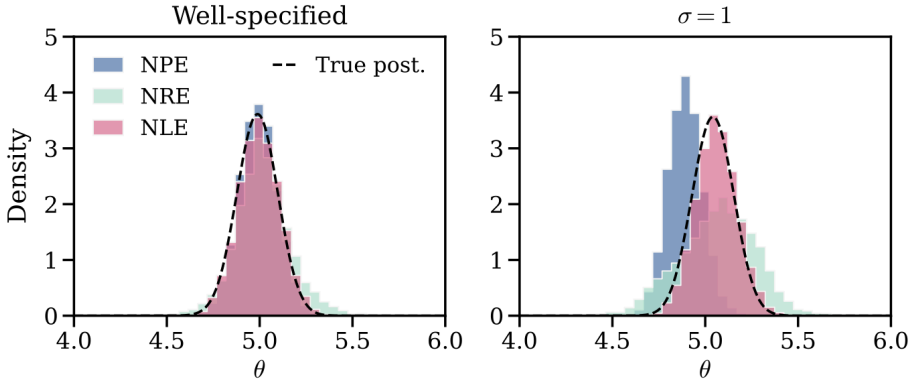
Structured test batteries from augmented simulators

NAM, J. Alvey, C. Weniger (PRD, 2025) [arXiv: 2412.15100]

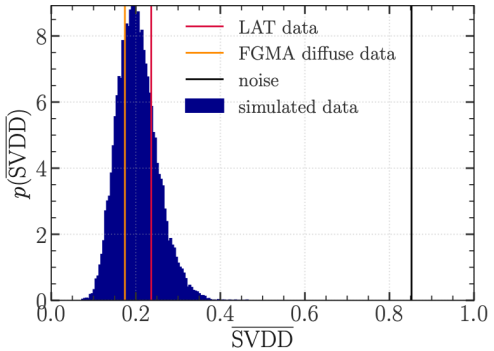
1. Inference robustness checks

Is our inference stable under small, structured changes to the data or the inference pipeline?

- Masking part of the data
- Changing summary network architecture
- Altering training algorithm, optimizers, seeds etc.



Cannon+22 [arXiv:2209.01845]



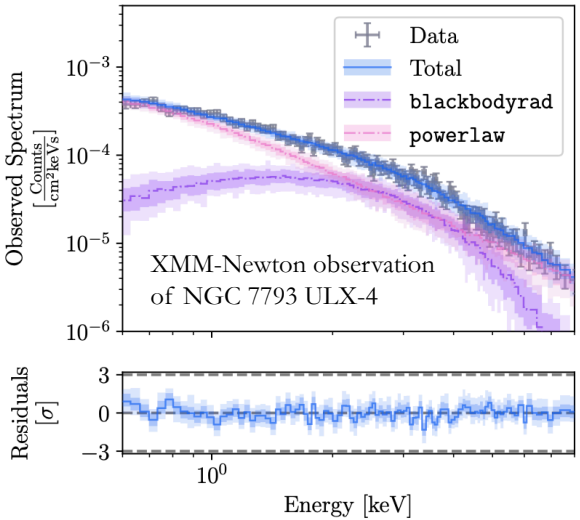
C. Eckner, NAM+25 [arXiv:2505.02906]

Check Christopher Eckner's talk and poster!

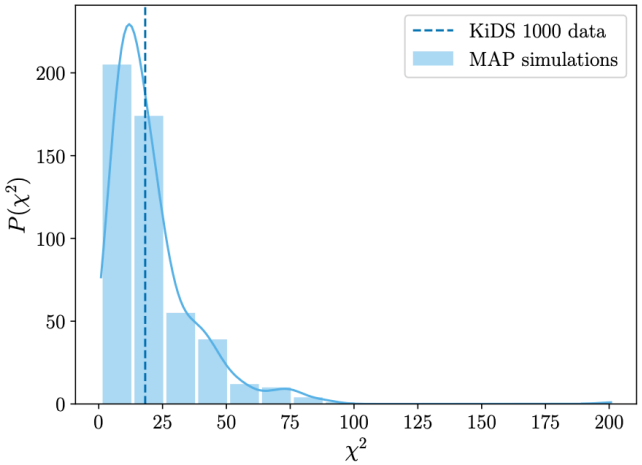
2. Posterior predictive validation

Does our model produce data that look like what we observed?

$$p(x_{\text{new}} \mid x_{\text{obs}}) = \int p(x_{\text{new}} \mid \theta)\, p(\theta \mid x_{\text{obs}})\, d\theta$$



Dupourqué+25 [arXiv:2506.05911]



KiDS-SBI 2024 analysis [arXiv:2404.15402]

3. Contrastive model diagnostics

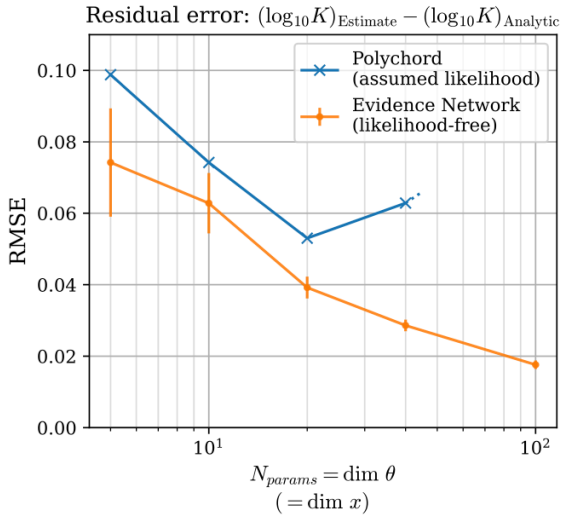
Which model describe our data better?

$$K = \frac{p(x_{obs}\,|\,H_1)}{p(x_{obs}\,|\,H_0)} \quad \text{with} \quad p(x_{obs}\,|\,H_j) = \int p(x_{obs}\,|\,\theta, H_j) p_j(\theta)\,\mathrm{d}\theta$$

SBI Bayesian model comparison via Evidence Networks:

Train a discriminative classifier via a special loss to distinguish between simulations from each model.

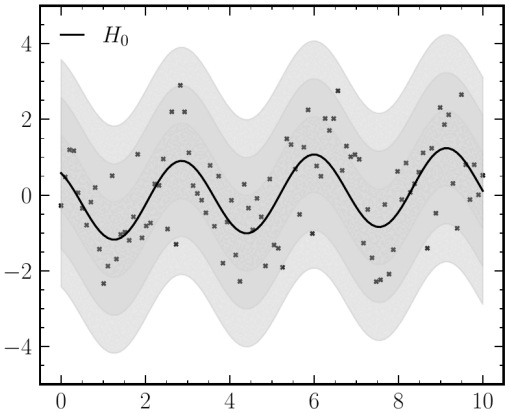
Jeffrey+23 [arXiv:2305.11241]



→ How to extend this contrastive diagnostic reasoning to a structured framework for multiple tests?

Structured test batteries from augmented simulators

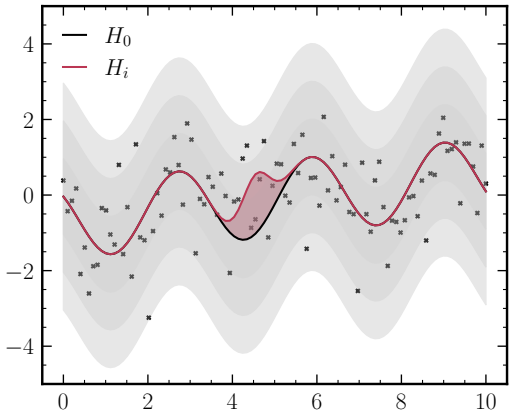
$$x^{(0)} \sim p(x \mid H_0)$$



NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

Structured test batteries from augmented simulators

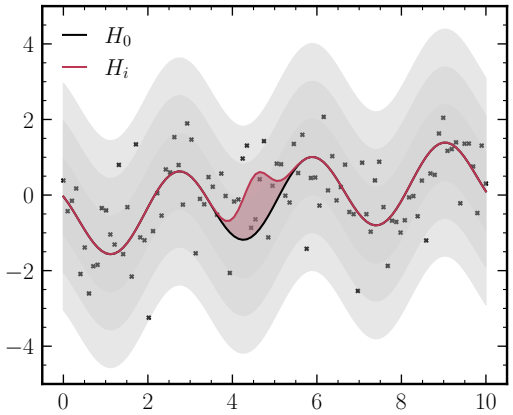
$$\begin{cases} x^{(0)} \sim p(x \mid H_0) \\ x^{(i)} \sim p(x \mid H_i) \end{cases}$$



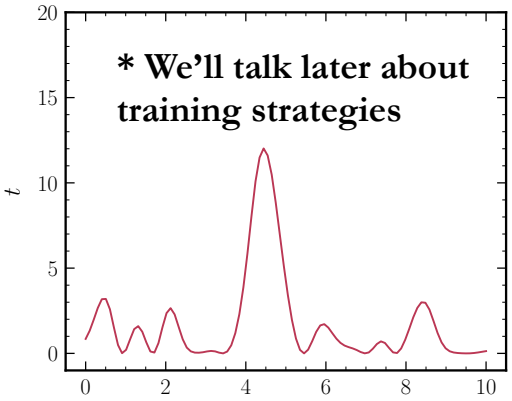
1. Define structured alternatives.

Structured test batteries from augmented simulators

$$\begin{cases} x^{(0)} \sim p(x \mid H_0) \\ x^{(i)} \sim p(x \mid H_i) \end{cases} \longrightarrow t_i(x) = -2 \log \frac{p(x \mid H_0)}{p(x \mid H_i)}$$



1. Define structured alternatives.

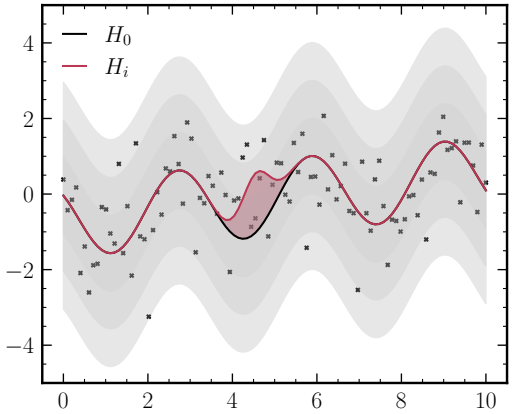


2. Define log-likelihood ratio test statistics for the alternatives.

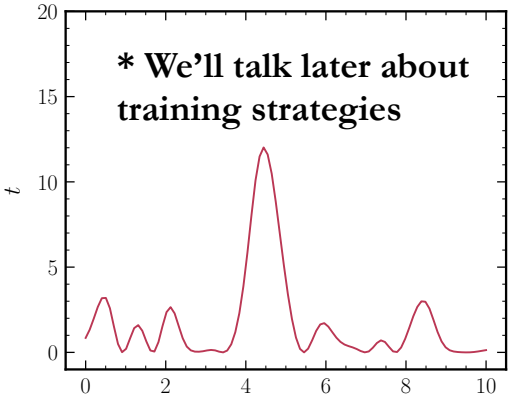
NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

Structured test batteries from augmented simulators

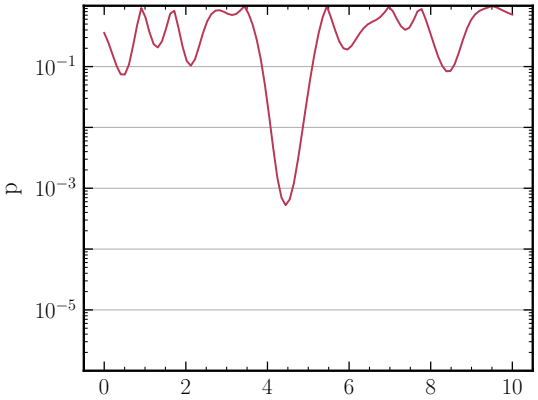
$$\begin{cases} x^{(0)} \sim p(x \mid H_0) \\ x^{(i)} \sim p(x \mid H_i) \end{cases} \longrightarrow t_i(x) = -2 \log \frac{p(x \mid H_0)}{p(x \mid H_i)} \longrightarrow \mathrm{p}_i(x) = \mathbb{E}_{x \sim p(x \mid H_0)}[\mathbb{I}(t(x) > t(x_{obs}))]$$



1. Define structured alternatives.



2. Define log-likelihood ratio test statistics for the alternatives.

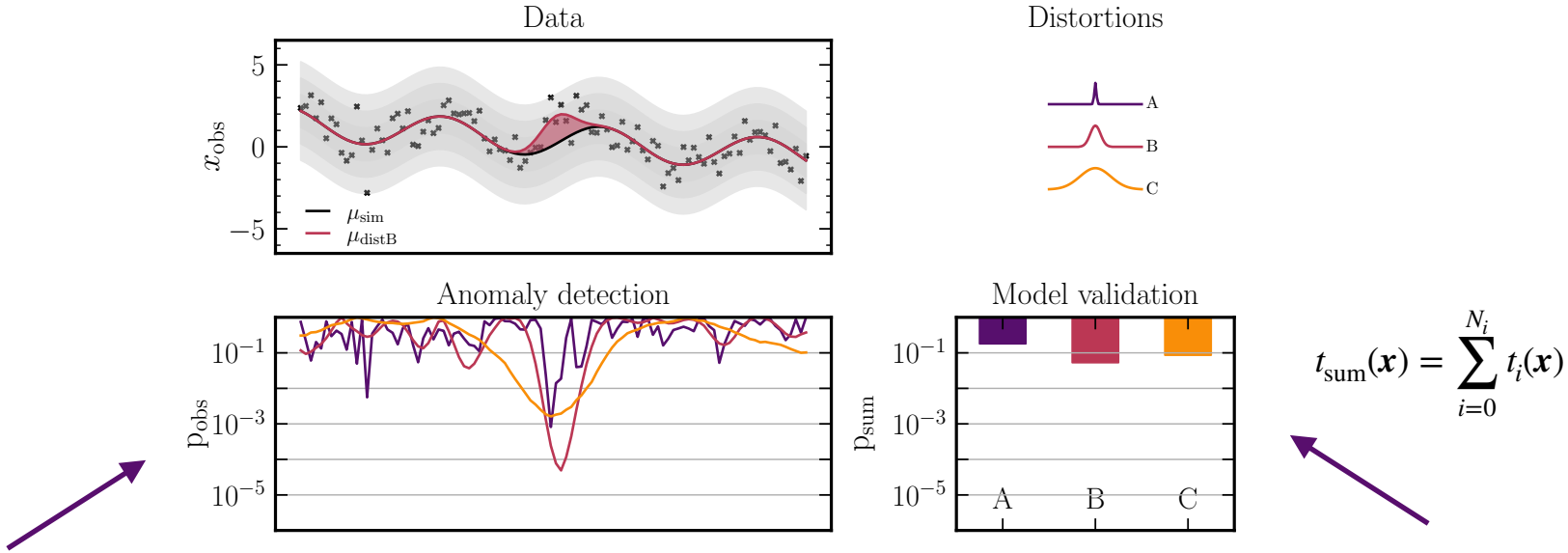


3. From test statistics to p-values.

NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

Structured test batteries from augmented simulators:

Localized and aggregated tests



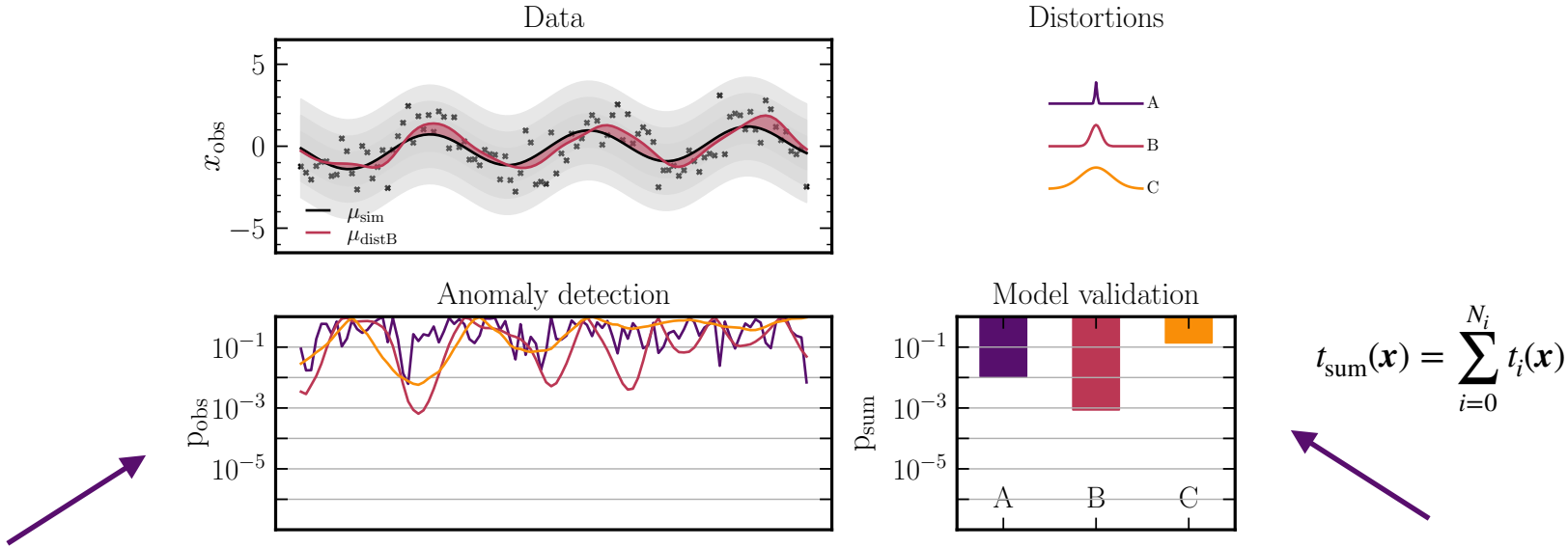
Localized test statistics are more sensitive towards single isolated distortions, and, in some limits, lead to matched filter and anomaly localization "bump-hunt" type of analyses.

Aggregated test statistics provides complementary information about the statistical significance of favoring the alternatives H_i over the baseline model H_0, and, in some limits, lead to model validation statistics.

NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

Structured test batteries from augmented simulators:

Localized and aggregated tests



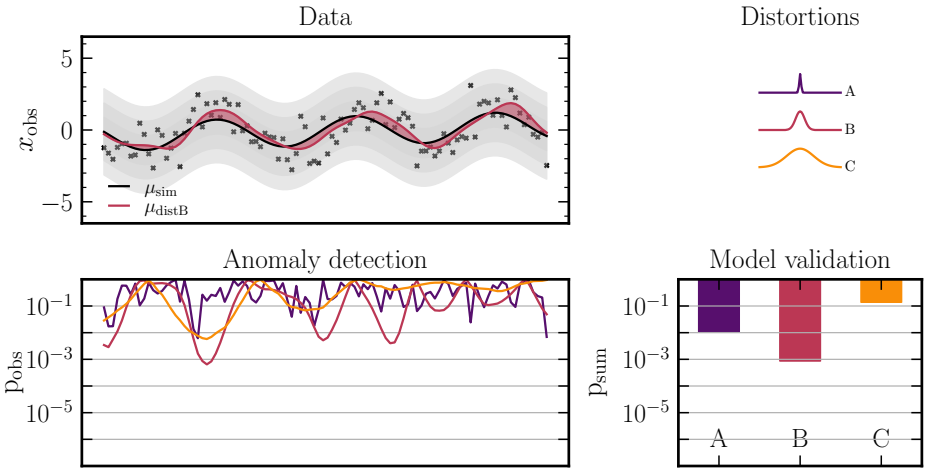
Localized test statistics are more sensitive towards single isolated distortions, and, in some limits, lead to matched filter and anomaly localization "bump-hunt" type of analyses.

Aggregated test statistics provides complementary information about the statistical significance of favoring the alternatives H_i over the baseline model H_0, and, in some limits, lead to model validation statistics.

NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

Structured test batteries from augmented simulators:

Global significance



Because many hypotheses (localized and aggregated) are tested in parallel, multiple testing corrections are required. A common approach is to compute a **global p-value** based on the most extreme individual result.

Global p-value $p_{\mathrm{glob}} = 7.57 \times 10^{-3}$

NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

Structured test batteries from augmented simulators:

Connection to classical testing frameworks

- **Localized** test statistics for **additive distortions** are closely related to matched filters and signal-to-noise ratio (SNR) statistics.

$$H_i : \tilde{\boldsymbol{x}} = \boldsymbol{x} + \epsilon \cdot \boldsymbol{n}^{(i)} \quad \text{with} \quad \epsilon \sim \mathcal{U}(-b, b)$$

Assuming a Gaussian likelihood function for the base model, in the large sample limit, and for scenarios where the maximum-likelihood estimator is not significantly correlated with the distortion, the test statistic for a given distortion is directly related to the signal-to-noise ratio (SNR) of that distortion in the data

$$t_i(\boldsymbol{x}) \simeq \mathrm{SNR}_i^2(\boldsymbol{x}) + \text{const.}$$

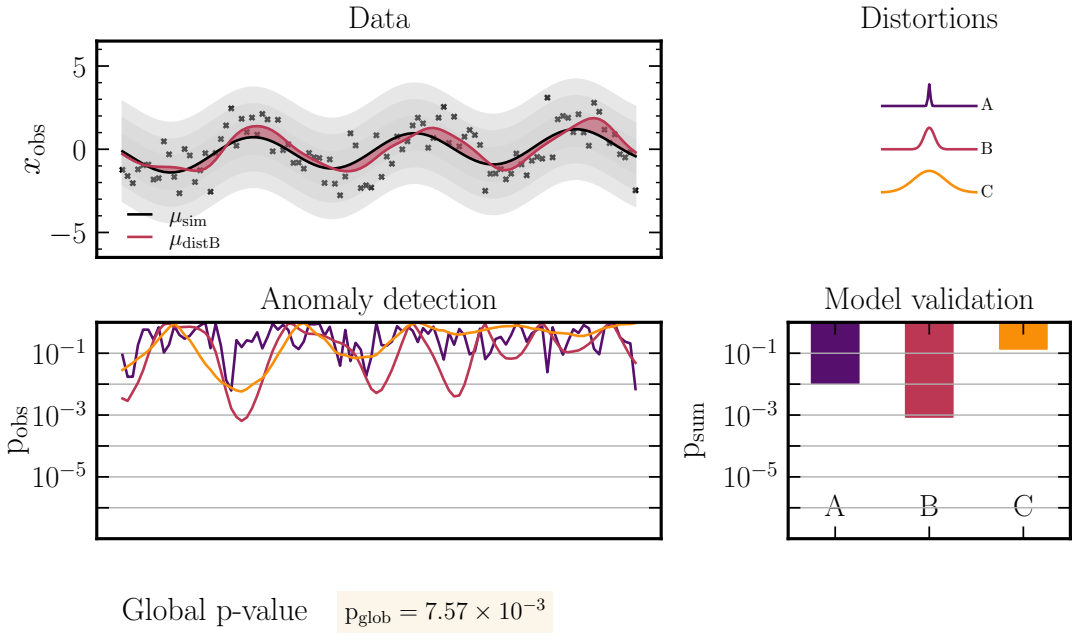
- **Aggregated** discrepancy scores reduce to χ^2 tests when distortions are orthogonal and noise is Gaussian.

$$t_{sum}(\boldsymbol{x}) \simeq \chi^2 + \text{const.}$$

- In this sense, simulator augmentation **extends** residual analysis and goodness-of-fit testing to the flexible, implicit-likelihood setting of SBI.

NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

What is the **training strategy** to build such structured test statistics batteries from augmented simulators?



NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

Structured test batteries from augmented simulators:

Training strategies

BCE: discriminative classifiers can be used to approximate the generalized likelihood ratio statistic.

$$\mathscr{L}^{(i)}_{\text{BCE}}\left[f_{i,\phi}(\boldsymbol{x})\right] = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x}|H_0)}\left[-\ln\sigma(f_{i,\phi}(\boldsymbol{x}))\right] + \mathbb{E}_{\boldsymbol{x}\sim p_(\boldsymbol{x}|H_i)}\left[-\ln\sigma(1-f_{i,\phi}(\boldsymbol{x}))\right] \qquad \text{with} \qquad t \approx 2f_{i,\phi}(\boldsymbol{x})$$

SNR: minimizing a Gaussian negative log-likelihood loss for the MLE of the matched filter $\epsilon(\boldsymbol{x})$ and its variance σ^2.

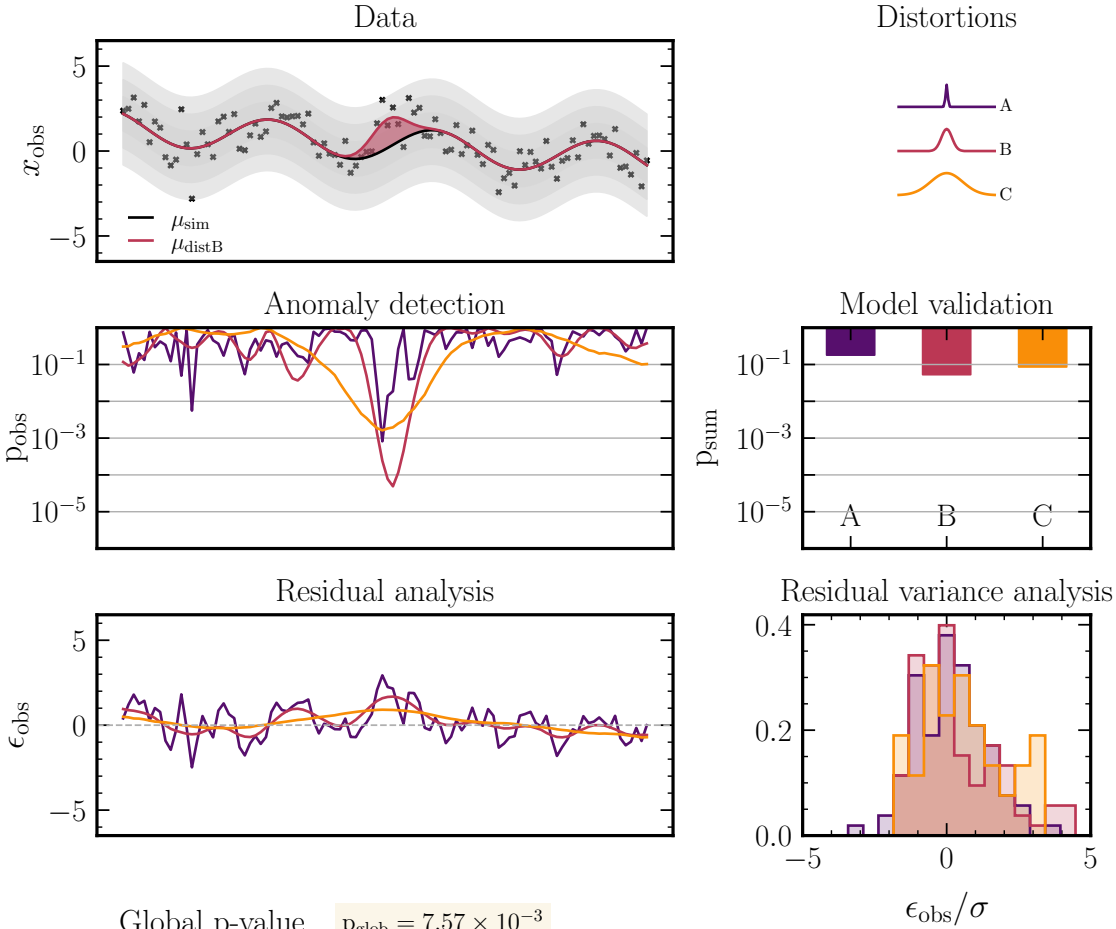
$$\mathscr{L}^{(i)}_{\text{SNR}}\left[\epsilon_{i,\phi}(\boldsymbol{x}), \sigma^2_{i,\phi}\right] = \mathbb{E}_{\boldsymbol{x},\epsilon\sim p(\boldsymbol{x}|H_i,\epsilon)p(\epsilon)}\left[\frac{(\epsilon_{i,\phi}(\boldsymbol{x})-\epsilon)^2}{\sigma^2_{i,\phi}} + \ln\sigma^2_{i,\phi}\right] \qquad \text{with} \qquad t_i(\boldsymbol{x}) \propto \frac{\epsilon_{i,\phi}(\boldsymbol{x})}{\sigma_{i,\phi}}$$

When distortions correspond to structured changes in space (e.g., image domains), the individual networks can be trained jointly using **shared neural architectures:** $\boldsymbol{f_\phi}(\boldsymbol{x}) : \mathscr{D} \to \mathbb{R}^{N_i}$ and $\boldsymbol{\epsilon_\phi}(\boldsymbol{x}) : \mathscr{D} \to \mathbb{R}^{N_i}$, $\boldsymbol{\sigma^2_\phi} : \mathscr{D} \to \mathbb{R}^{N_i}$

NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

Residual analysis

SNR training strategy allows to visualize where distortions are located in data space and how large they are.

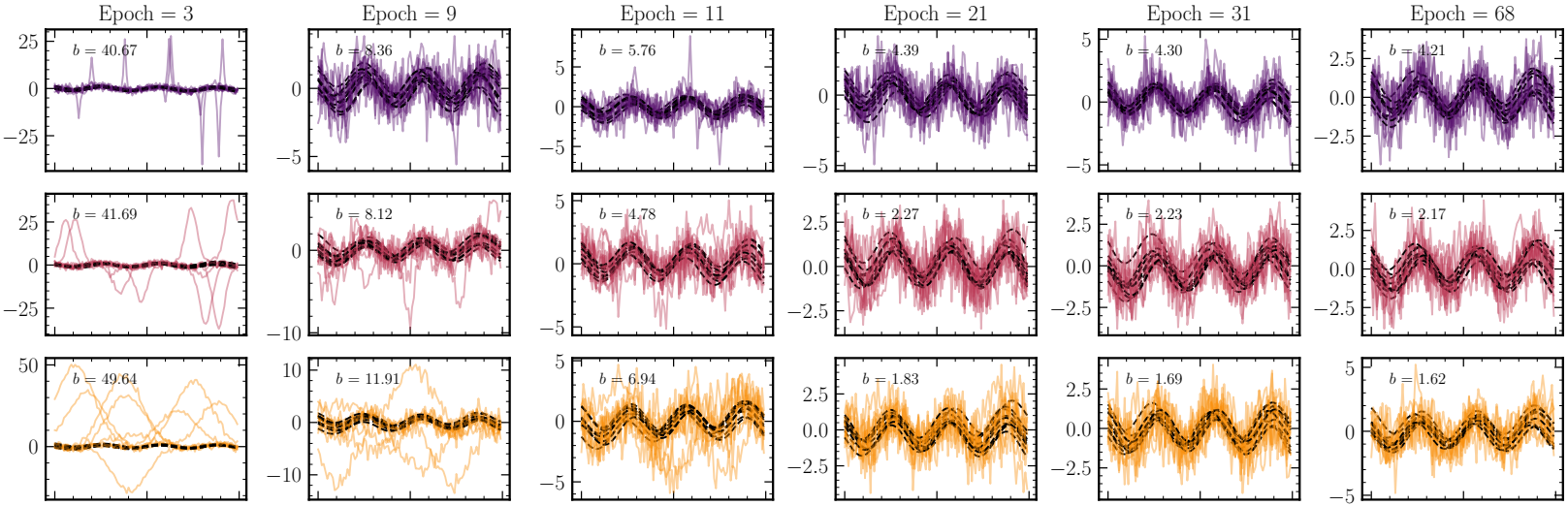


Global p-value $p_{\rm glob} = 7.57 \times 10^{-3}$

Adaptive learning of distortions amplitude

$$b = \mathrm{SNR}_{\max}\sigma = \frac{\mathrm{SNR}_{\max}}{\sqrt{(\boldsymbol{n}^{(i)})^T \Sigma^{-1} \boldsymbol{n}^{(i)}}}$$

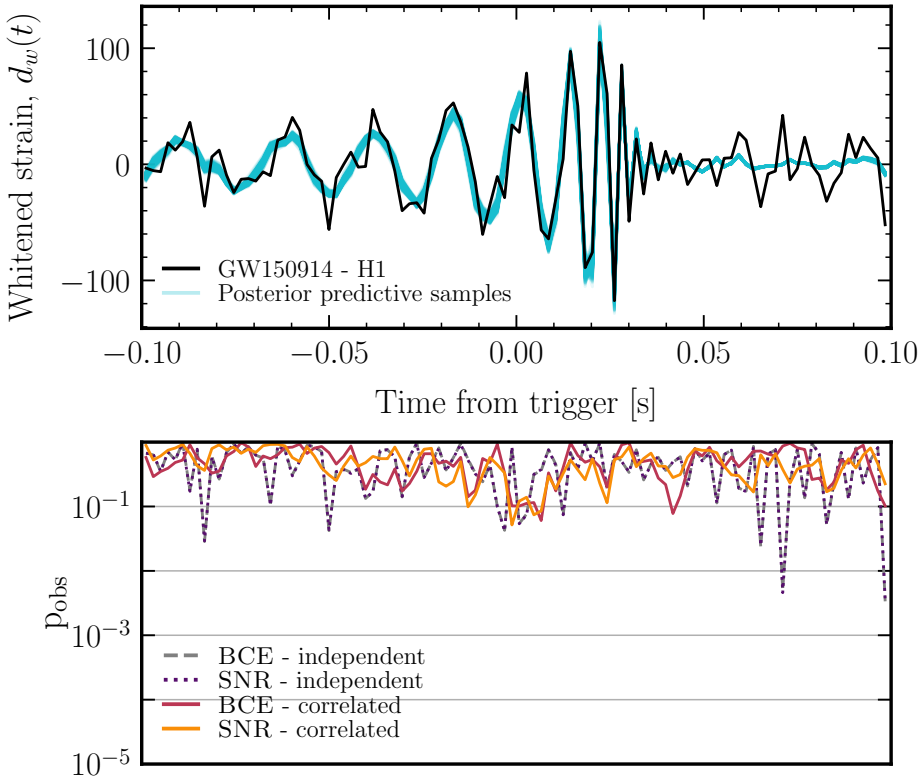
The algorithm converges to distortions that are significant enough to be detectable, but not so significant that they are clearly ruled out.



NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

An application to GW150914

- Fit model to data using `jimgw` [Wong+23 - arXiv:2302.05333].
- Construct structured test batteries of independent and correlated distortions to the model and test for misspecification.
- As expected, no significant anomaly is present in the modelling of GW150914, with global *p*-values for all the types of analyses of around a few tenths.



NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

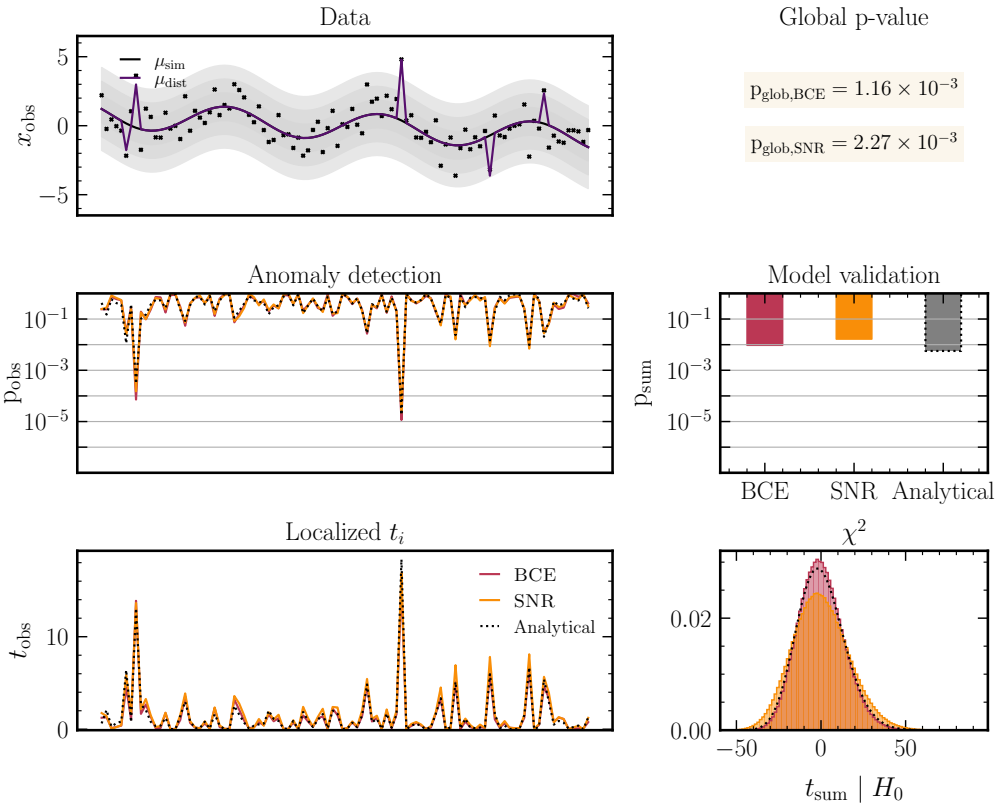
Summary

- **Model misspecification analysis** strategies are integral to advancing our understanding of physical phenomena.
- The framework presented here is designed to carry this out in a SBI context. By leveraging classical concepts, it provides a flexible and comprehensive approach to **simultaneously perform many hypothesis tests and quantify their statistical significance**.
- Via two **training strategies**, we can actually test (and Monte Carlo sample) all of these alternative hypotheses simultaneously. This makes the pipeline very **efficient** when looking to test for broad classes of mismodelling, while still maintaining the ability to carry out individual, targeted tests.
- The **SNR training strategy**, can be used to visualize model residuals and calibrate the scale of distortions searched for in the data.

Thanks!

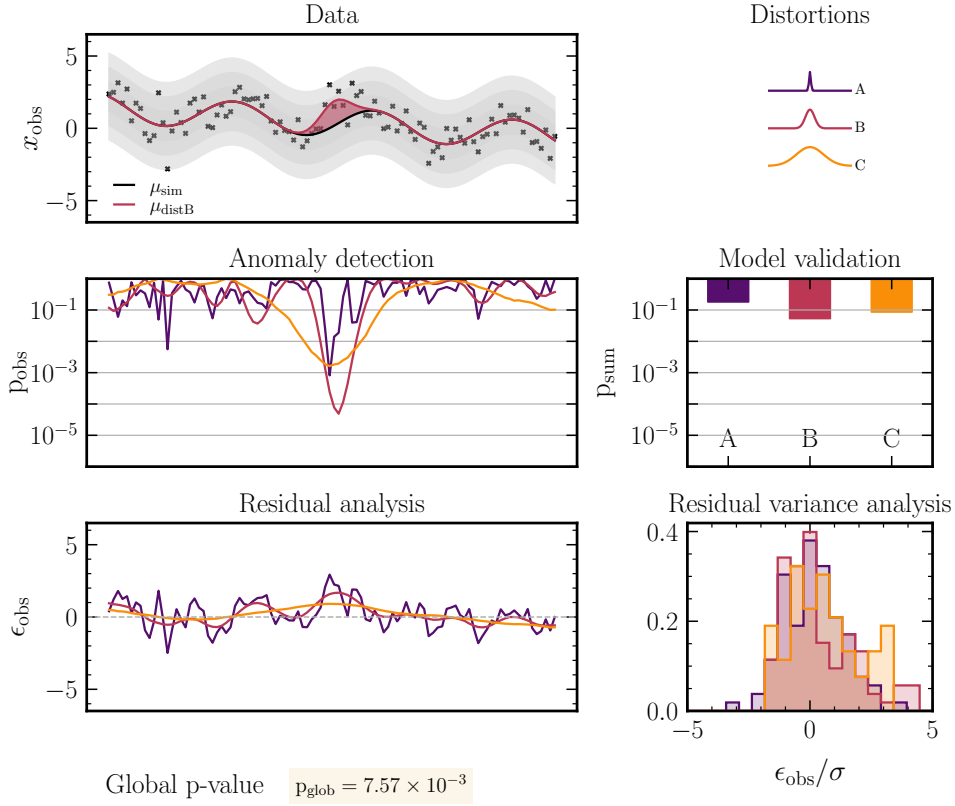
Backup slides

Framework summary: **analytic test**



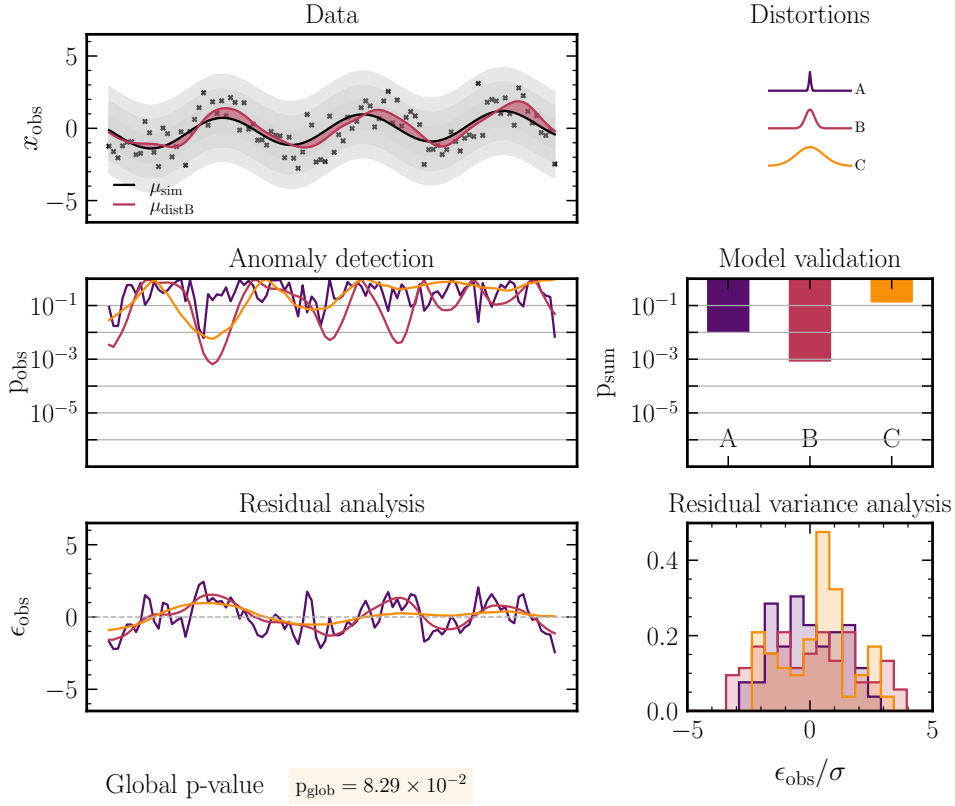
NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

Framework summary: **correlated distortions**



NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

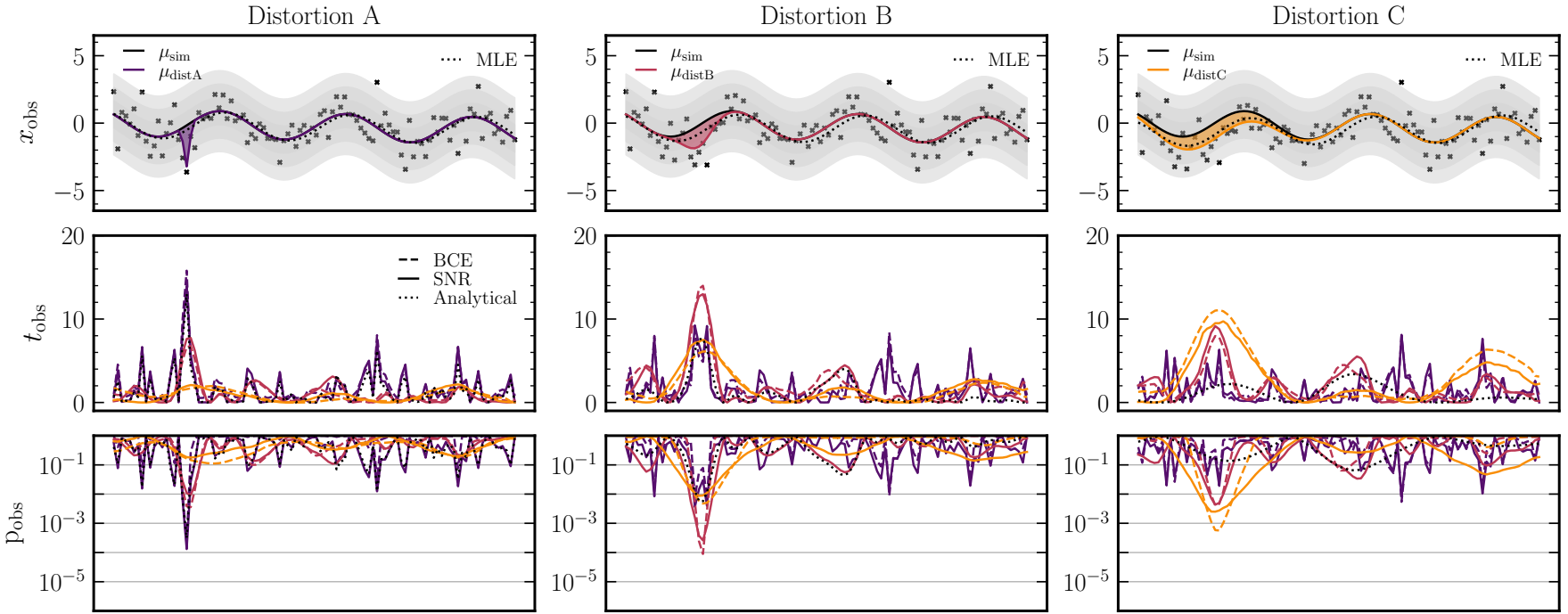
Framework summary: **multiple small correlated distortions**



NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]

The effect of **marginalization**

$$t(x) = -2\log\frac{\int p(x\,|\,\theta, H_0)p_0(\theta)\mathrm{d}\theta}{\int p(x\,|\,\theta, H_1)p_1(\theta)\mathrm{d}\theta}$$



NAM, J. Alvey, C. Weniger [arXiv: 2412.15100]