

# Tests for model misspecification in simulation-based inference: from local distortions to global model checks

Noemi Anau Montel,<sup>1,\*</sup> James Alvey,<sup>2,3,†</sup> and Christoph Weniger<sup>4,‡</sup>

<sup>1</sup>*Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85748 Garching, Germany*

<sup>2</sup>*Kavli Institute for Cosmology Cambridge, Madingley Road, Cambridge CB3 0HA, United Kingdom*

<sup>3</sup>*Institute of Astronomy, University of Cambridge,*

*Madingley Road, Cambridge CB3 0HA, United Kingdom*

<sup>4</sup>*GRAPPA Institute, Institute for Theoretical Physics Amsterdam,*

*University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands*

Model misspecification analysis strategies, such as anomaly detection, model validation, and model comparison are a key component of scientific model development. Over the last few years, there has been a rapid rise in the use of simulation-based inference (SBI) techniques for Bayesian parameter estimation, applied to increasingly complex forward models. To move towards fully simulation-based analysis pipelines, however, there is an urgent need for a comprehensive simulation-based framework for model misspecification analysis. In this work, we provide a solid and flexible foundation for a wide range of model discrepancy analysis tasks, using *distortion-driven model misspecification tests*. From a theoretical perspective, we introduce the statistical framework built around performing many hypothesis tests for distortions of the simulation model. We also make explicit analytic connections to classical techniques: anomaly detection, model validation, and goodness-of-fit residual analysis. Furthermore, we introduce an efficient self-calibrating training algorithm that is useful for practitioners. We demonstrate the performance of the framework in multiple scenarios, making the connection to classical results where they are valid. Finally, we show how to conduct such a distortion-driven model misspecification test for real gravitational wave data, specifically on the event GW150914.

## I. INTRODUCTION

The primary goal of the physical sciences is to refine analytic or computational models that form the backbone of our understanding of physical phenomena. Key steps in this process are designing, fitting, and validating different models against data. There are a variety of existing strategies to approach this. Powerful tools include Bayesian evidence estimation conditioned on observations for model comparison [1], or goodness-of-fit and hypothesis tests for establishing the validity of a model across the observed data and making discoveries [2, 3].

In recent years, simulation-based inference (SBI), also known as implicit-likelihood inference, has emerged as an important tool for inference when simulations from implicitly-defined models are available [4]. This approach is especially useful for managing the increasing dimensionality of datasets and the growing complexity of scientific models, where often a full probabilistic description is difficult or even impossible to define. It is also useful in the regime where classical methods are computationally demanding, but efficient simulations remain feasible.

In SBI, the modeling complexity is shifted from having to define a likelihood function to having to program a simulator, which may be easier than constructing an analytical probabilistic description. In principle, this allows one to include and account for more effects in the

modeling than in traditional methods. However, as in any statistical inference framework, the question of model misspecification — how to detect it, how it affects parameter reconstruction, and how to account for it — remains. In the SBI setting, model misspecification occurs when the data-generating function (the simulator) is a poor representation of the physical processes being analyzed. Modeling choices in the simulator are thus extremely important and the degree of simulator complexity requires a careful balance between over-fitting and under-fitting the data.

The majority of existing SBI applications have so far focused on uncertainty quantification in parameter inference tasks [e.g. 5–13].<sup>1</sup> Recently though, there has been a surge of interest in the development of SBI algorithms for other cornerstones of classical statistics: for example, model comparison through Bayesian evidence computation [14–16], and frequentist hypothesis testing [17–19]. In particular, calibrated binary classifiers have been shown to be equivalent to classical likelihood-ratio test statistics [17]. It has also been shown that it is possible to build confidence intervals with good frequentist properties [18], and obtain test statistics equivalent to profile likelihood ratios [19] in SBI settings. A recent example of this interest is the simulation-based cosmological analysis of KiDS-1000 data, where a classical goodness-of-fit measure for Bayesian settings has been adapted to inspect SBI posteriors [20].

\* noemiam@mpa-garching.mpg.de

† jbg2@cam.ac.uk

‡ c.weniger@uva.nl

<sup>1</sup> An extensive list of SBI applications can be found here: <https://github.com/smsharma/awesome-neural-sbi>.

The effects of model misspecification in SBI and possible mitigation strategies have also received some attention in the recent literature. Generally, a misspecification of the model is expected to lead to wrong inference results. In likelihood-based inference, a misspecified likelihood function is expected to always lead to the same incorrect results [21]. In contrast, it has been shown that the failure modes of SBI depend on the adopted method; this means that a consistency check between methods can be used as a diagnostic check [22]. Other diagnostic checks are based on testing whether the learned data summaries lead to outliers in latent space when applied to real-world data [23]. Lastly, in order to reduce the effect of model misspecification on inference results, various approaches to use a small set of (labeled or unlabeled) real-world data examples to make data summaries resilient against misspecification have been explored [24–27]. In this work, we will limit ourselves to developing diagnostic strategies for *detecting model misspecification*, which may serve as the basis for identifying and eventually correcting any insufficiency in a given simulation model.

With this context in mind, we propose a novel SBI framework for a wide variety of *locally interpretable* and *globally significant* model misspecification tests for simulated hypotheses. Our versatile and flexible framework is based on *model augmentation*, fully embracing the core aspect of SBI: if you can simulate it, you can test for it. In general, the proposed framework allows one to perform, in a practical and comprehensive way, analyses such as anomaly detection and model validation. It also allows us to assess their significance, and perform residual analyses. In limiting cases, we demonstrate the close connection of our approach to classical matched filtering and  $\chi^2$ -goodness-of-fit tests. Furthermore, we propose an efficient self-calibrating training algorithm that converges to look for distortions that are just plausible given the observational noise. We demonstrate the performance of the framework in an instructive scenario, and show a proof-of-concept application of the framework to real gravitational waves data.

The rest of the paper is organized as follows. In Section II, we motivate and describe our framework based on high-volume hypothesis testing via data augmentation. Section III displays an instructive example, highlighting its direct link to traditional techniques and analytic results. We present an application to gravitational waves data in Section IV. In Section V, we discuss possible improvements and the relevant limitations of our approach. Finally, we present some outlook and our conclusions in Section VI.

**Code:** The code to reproduce the examples in this work can be found at [NoemiAM/mist](https://github.com/NoemiAM/mist).

## II. MISSPECIFICATION TESTING IN SIMULATION-BASED INFERENCE

In general, we are interested in testing and searching for aspects of the data that might not be fully accounted for by our base model. These deviating features can arise in the form of, e.g., distortions in individual data bins, correlated distortions, excesses in specific Fourier modes, or additional model components, depending on the type of data at hand. Here, we describe a general SBI framework to *simultaneously* test for many types of deviations with respect to the base model in the data and discuss its connections to classical testing frameworks.

The proposed framework is summarized in Figure 1. It is based on a high-volume (i.e. large numbers of individual tests) hypothesis testing algorithm rooted in SBI (Section II.1), that broadly recovers classical techniques in anomaly detection and model validation in limiting cases. Given the large number of tests performed, it is necessary to account for correlated trials when assessing the overall significance. This is discussed in Section II.3. Furthermore, we make explicit connections between the proposed framework and classical ones in Section II.4. Finally, we discuss briefly training strategies in Section II.5.

### II.1. High-volume hypothesis testing

Our starting point for carrying out model misspecification testing using many different types of distortion (what we call *high-volume*) is the classical hypothesis testing framework [e.g. 28]. In a nutshell, hypothesis testing aims to assess a null hypothesis,  $H_0$ , by determining whether it can be rejected in favor of an alternative hypothesis,  $H_1$ , given observational data  $\mathbf{x}_{\text{obs}}$ . In practice, this involves defining a test statistic  $t(\mathbf{x})$ , a single real-valued summary, ideally designed to maximize the ability to distinguish between  $H_0$  and  $H_1$ . Evaluated on the observed data, the test statistic yields  $t_{\text{obs}} = t(\mathbf{x}_{\text{obs}})$ , which is then compared with its distribution under  $H_0$ , denoted as  $p(t|H_0)$ . The key quantity of interest in this process is the p-value, defined as the probability of observing a value of  $t$  at least as extreme as the observed one,  $t_{\text{obs}}$ , assuming the null hypothesis is true,  $p_{\text{obs}} = p(t > t_{\text{obs}}) = \int_{t_{\text{obs}}}^{\infty} p(t|H_0)dt$ .

In an SBI setting, the logical first step is to define the null hypothesis in terms of the base simulator,  $p_{\text{sim}}(\mathbf{x})$ , that implicitly defines the likelihood of the base model.<sup>2</sup> The null hypothesis is thus defined as

$$H_0 : \mathbf{x} \sim p_{\text{sim}}(\mathbf{x}) \equiv p(\mathbf{x}|H_0). \quad (1)$$

<sup>2</sup> In general, the typical setup will be parametric simulators in the form of  $p_{\text{sim}}(\mathbf{x}) = \int d\Theta p(\mathbf{x}|\Theta)p(\Theta)$ , where  $p(\mathbf{x}|\Theta)$  is the likelihood of the data given some model parameters  $\Theta$  with prior  $p(\Theta)$ . One can also choose to test simulators with constrained proposal distribution through active learning [29–31]. More generally,  $\mathbf{x} \sim p_{\text{sim}}(\mathbf{x})$  can also be samples from generative models.

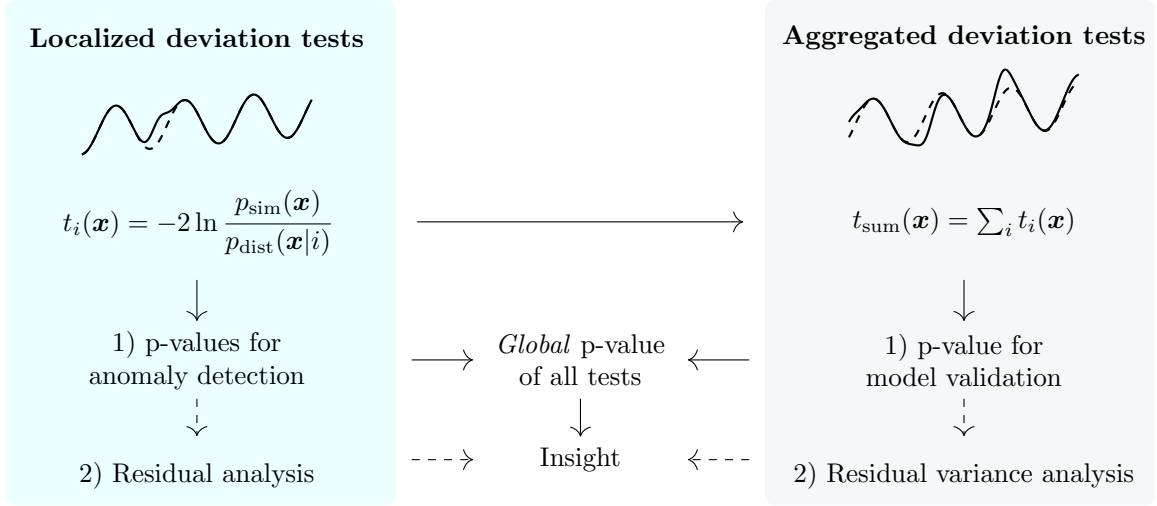


FIG. 1. Summary illustration of the presented framework for tests of model misspecification in SBI (see Section II for details). **Left panel:** An ensemble of localized test statistics is learned by neural networks (see Appendix A for details); they are typically more sensitive towards isolated distortions and in some limits can be the basis for anomaly detection. Their individual significance can be quantified with Monte-Carlo estimates, and in specific training scenarios (see Appendix C for details) one can visualize the distortions to the model in data space through residuals. **Right panel:** Aggregated test statistics can be constructed given any subset of localized test statistics; they are sensitive towards the cumulative evidence of multiple distortions and in some limits can be the basis for model validation tests. Their individual significance can be quantified with Monte-Carlo estimates, and in specific training scenarios one can perform a residual variance analysis. **Central panel:** We can estimate the overall global significance of all the performed tests, accounting for their correlation.

To test for deviations in the data that are not fully described by the base model/simulator, we construct an *ensemble* of  $N_{\text{alt}}$  alternative data-generating functions, i.e. alternative simulation models. These alternative simulation models are defined by suitably augmenting the base simulator model with distinct *stochastic distortions*  $i$  to the data,

$$\tilde{\mathbf{x}} \sim p_{\text{dist}}(\tilde{\mathbf{x}}|i) : \tilde{\mathbf{x}} \sim p_{\text{dist}}(\tilde{\mathbf{x}}|\mathbf{x}, i) \quad \text{with } \mathbf{x} \sim p_{\text{sim}}(\mathbf{x}) \quad (2)$$

The stochastic distortions model any deviating feature of the data we want to test for. Each of these possible distortions defines an alternative hypothesis

$$H_i : \mathbf{x}_i \sim p_{\text{dist}}(\mathbf{x}|i) \equiv p(\mathbf{x}|H_i) \quad \text{with } i = 1, \dots, N_{\text{alt}}, \quad (3)$$

against which we test our base model  $H_0$ . In general, the index  $i$  that characterizes the distortion does not have to be discrete, i.e. one could parameterize the distortions in a continuous way.<sup>3</sup> In this work, however, we use the discrete indexing.

As for the test statistic, we consider the widely used log-likelihood ratio test statistic<sup>4</sup>

$$t_i(\mathbf{x}) \equiv -2 \ln \frac{p(\mathbf{x}|H_0)}{p(\mathbf{x}|H_i)} = -2 \ln \frac{p_{\text{sim}}(\mathbf{x})}{p_{\text{dist}}(\mathbf{x}|i)}. \quad (4)$$

In the applications below, we will approximate this ensemble of  $N_{\text{alt}}$  test statistics  $t_i$  via neural networks. The learned neural test statistics allow us to test for multiple types of deviations in the data simultaneously. We refer the reader to Section II.5 for more details regarding specific training strategies.

It is important to emphasize that, within SBI, ensemble hypothesis testing for model distortions is achieved *operationally* through data augmentation. In simple terms, different hypothesis  $H_i$ , each describing the data distribution via a different likelihood functions  $p(\mathbf{x}|H_i)$ , manifest in SBI through distinct data generation processes — simulation models with diverse stochastic distortions,  $p_{\text{dist}}(\mathbf{x}|i)$  — which implicitly encode the alternative likelihoods. The flexibility of neural networks enables high-volume hypothesis testing for all these distortions.

<sup>3</sup> For example, particle physics “bump-hunt” searches [3] are one common scenario in which a continuous labeling is natural.

<sup>4</sup> We know from the Neyman-Pearson lemma [2] that the log-likelihood ratio provides the optimal test statistic to maximally distinguish between  $H_0$  and  $H_1$  in case of simple hypotheses (i.e. hypotheses that fully specify the probability distribution of the data). For composite hypotheses (i.e. hypotheses where some of the distribution parameters are not specified), the log-likelihood ratio test is usually generalized by optimizing the likelihood over the parameter spaces of both the null and alternative hypotheses [32]. We note that in most of the use-cases this framework can be applied to, one usually does not have a uniformly most powerful test.

## II.2. Localized and aggregated tests

We refer to each test statistic  $t_i$  as *localized*, not in the spatial sense, but because it is tied to a specific, single, narrowly defined distortion scenario in the data. These 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.

For any subset of alternative hypotheses  $H_i$ , we can also consider the sum over all localized test statistics,

$$t_{\text{sum}}(\mathbf{x}) = \sum_{i=0}^{N_{\text{alt}}} t_i(\mathbf{x}) , \quad (5)$$

which is expected to be particularly useful if there is a general tendency to prefer alternative hypothesis  $H_i$  over  $H_0$ . This *aggregated* test statistic provides *complementary* information about the statistical significance of favoring the alternatives  $H_i$  over the baseline model  $H_0$ . It is thus sensitive towards the cumulative evidence of multiple distortions being present in the data. The precise meaning of  $t_{\text{sum}}(\mathbf{x})$  depends on the specific choice of the subset of alternative hypotheses  $H_i$ , allowing flexibility in designing custom tests that target particular classes of deviations (for a short discussion see Section V). For example, we will see that under certain conditions this aggregated test statistic asymptotically follows a  $\chi^2$  distribution. The connection of the framework to classical tests will be discussed in Section II.4, with more details provided in Appendix B.

## II.3. Individual and global significance estimates

A sufficiently large test statistic for a given discrepancy (whether a localized one  $t_i$  or an aggregated one  $t_{\text{sum}}$ ) hints that the corresponding alternative hypothesis is preferred over the baseline assumption  $H_0$ . As is widely done, to correctly quantify the statistical significance of a discrepancy, one can use a Monte Carlo estimate of the corresponding p-value.<sup>5</sup> This estimate requires a sufficiently large number of samples from the base model, but can be done in parallel for all alternative hypotheses.

The above construction provides information about the *individual* significance of each specific alternative (e.g. a localized anomaly or a cumulative effect of many distortions), without accounting for the fact that multiple hypotheses (e.g. the presence of a specific localized

distortion or the possibility that there are many subtle correlated discrepancies not being modeled) are being tested. Therefore, we need to account for the fact that, when testing a large number of alternative hypotheses, the probability of observing a large test statistic purely by chance increases. This is analogous to the “look-elsewhere effect” in particle physics [33], where searching over many possible signal locations increases the chance of a statistical fluctuation appearing significant.

To estimate the significance of the individual discrepancies from the null hypothesis  $H_0$  at a more *global* level, accounting for the large number of performed tests and their possible correlations, it is necessary to estimate a global p-value for the overall, i.e. trials-corrected, significance of the minimum observed p-value from all tests. In different words, the global p-value estimates the probability of observing the most extreme test outcome across any of the hypotheses being tested.

Operationally, this global p-value can be obtained as follows. First, for  $N_{\text{mc}}$  Monte Carlo samples, we compute p-values (see Footnote 5) for all of the  $N_t$  individual tests (either localized or aggregated) of interest. Each of these Monte Carlo samples generates  $N_t$  p-values, one for each test. While any single p-value is uniformly distributed under the null hypothesis, it is important to note that they exhibit correlations among themselves. To account for this, from each of the  $N_{\text{mc}}$  samples, we extract the minimum p-value across the  $N_t$  tests, yielding  $N_{\text{mc}}$  such minima. Now, we want to ask the following question: given an observation on which we have performed  $N_t$  tests, what is the probability of having observed the most significant deviation (i.e. the one with the test yielding the minimum p-value)? We can use the  $N_{\text{mc}}$  minimum p-values to answer this question by comparing the distribution of these minimum p-values with the minimum p-value of the observation. This allows us to properly assess the global significance of the observed minimum p-value, the global p-value.

## II.4. Connection to classical testing frameworks

Under a few key assumptions, the above framework for misspecification testing in SBI via high-volume hypothesis testing can be connected to classical testing methods. For a detailed derivation of the following results we refer the reader to Appendix B.

Given a base simulator from which one can sample  $\mathbf{x} \sim p(\mathbf{x}|H_0)$ , let us consider simple *stochastic additive non-Gaussian distortions* with specific noise directions  $\mathbf{n}^{(i)}$  in data space

$$H_i : \tilde{\mathbf{x}} = \mathbf{x} + \epsilon \cdot \mathbf{n}^{(i)} \quad \text{with} \quad \epsilon \sim \mathcal{U}(-b, b) . \quad (8)$$

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  $\mathbf{n}^{(i)}$ , we derive the central result that **the test statistic for a given distortion**

<sup>5</sup> Operationally, this is easily computed as

$$p(\mathbf{x}_{\text{obs}}) = \mathbb{E}_{\mathbf{x} \sim p_{\text{sim}}(\mathbf{x})} [\Theta(t(\mathbf{x}) - t(\mathbf{x}_{\text{obs}}))] , \quad (6)$$

where  $\Theta[\cdot]$  is the Heaviside step function; or equivalently

$$p(\mathbf{x}_{\text{obs}}) = \mathbb{E}_{\mathbf{x} \sim p_{\text{sim}}(\mathbf{x})} [\mathbb{I}(t(\mathbf{x}) > t(\mathbf{x}_{\text{obs}}))] , \quad (7)$$

where  $\mathbb{I}[\cdot]$  is an indicator function, which is unity if the condition is true and zero otherwise.



$\mathbf{n}^{(i)}$  is directly related to the signal-to-noise ratio (SNR) of that distortion in the data

$$t_i(\mathbf{x}) \simeq \text{SNR}_i^2(\mathbf{x}) + C \quad (9)$$

where  $C$  is a constant that depends on the prior distribution of the distortion amplitude  $\epsilon$  and is defined in Equation B20, and the SNR is

$$\text{SNR}_i(\mathbf{x}) = \frac{\epsilon^*(\mathbf{x})}{\sigma}, \quad (10)$$

where  $\epsilon^*(\mathbf{x})$  is the maximum likelihood estimator (MLE) of  $\epsilon$  (Equation B15), and  $\sigma^2$  the variance of the MLE (Equation B16).

This is directly analogous to the matched filtering technique used in signal detection, where the data is correlated with a set of template signals to find the one with the highest SNR. Performing high-volume hypothesis testing in this limiting case and under the above assumptions, allows one to obtain results simultaneously for many types of distortions  $\mathbf{n}^{(i)}$ .

Let us now restrict the generality of the distortions such that each distortion corresponds to a deviation along one of the *standard basis* vectors in data space, i.e.  $\mathbf{n}^{(i)} = \mathbf{e}_i$ , where  $\mathbf{e}_i$  is the unit vector in the  $i$ th data space dimension. Furthermore, let us consider a diagonal noise covariance matrix. Under the same assumptions as before, we find that the sum of the test statistics over all alternative hypotheses  $H_i$  for the stochastic additive distortions under consideration, i.e. the aggregated test statistic from Equation 5, is related to the Pearson's  $\chi^2$  test statistic (see derivation for Equation B22):

$$t_{\text{sum}}(\mathbf{x}) = \chi^2(\mathbf{x}) + \text{const}. \quad (11)$$

Thus, the sum test statistic captures the cumulative effect of small deviations across multiple dimensions, reflecting an overall mismatch between the data and the model, and providing a goodness-of-fit measure.

### II.5. Training strategies

In this work, we adopt two different training strategies to estimate the localized test statistics  $t_i$ . The first one is very general and does not rely on any prior assumptions, whereas the second one is motivated by the connection of our framework to classical testing frameworks as discussed in the previous section.

**BCE** - The first training strategy employs discriminative classifiers to approximate likelihood ratio statistics using the binary cross entropy (BCE) loss [17]. Thus, we refer to it as the BCE training strategy. For more details we refer the reader to Appendix A.

**SNR** - The second training strategy stems from the results discussed in Section II.4, where it was shown that under certain assumptions the test statistic for a given

distortion  $\mathbf{n}^{(i)}$  is directly related to the SNR of that distortion in the data (Equation 9). Hence, these test statistics can be equivalently trained by minimizing a Gaussian negative log-likelihood loss for the MLE of the matched filter  $\epsilon^*(\mathbf{x})$  and its variance  $\sigma^2$ . Thus, we refer to it as the SNR training strategy. For more details we refer the reader to Appendix C.

A direct comparison of the training strategies for the example presented in Section III can be found in Appendix D. We also apply both training strategies to the gravitational waves example in Section IV.

## III. INSTRUCTIVE EXAMPLE

In this section we show the results of our unified framework for an instructive example: a white noise time series that can be thought of as a toy example for the analysis of e.g. gravitational waves or light curves. The data is defined across 100 evenly spaced bins, on a grid  $\mathbf{y} = (y_1, \dots, y_{100})$  from  $y_1 = -10$  to  $y_{100} = 10$ . Our baseline model consists of unit variance, uncorrelated Gaussian noise in each bin, along with a deterministic signal component that is defined by a sinusoidal function of the grid  $\mathbf{y}$ . Specifically, the sinusoidal function has parameters  $\Theta$  that define an adjustable phase, a linear trend, and an offset. Together, these define a mean  $\mu_j(\Theta) = \sin(y_j + 0.5\Theta_0) + 0.1\Theta_1 y_j + 0.5\Theta_2$ , where  $j$  is the grid index, with  $\Theta$  sampled from  $p(\Theta) = \mathcal{U}(-1, 1)^3$ . The baseline model is thus defined by  $p_{\text{sim}}(\mathbf{x}) = \int d\Theta \mathcal{N}(\mathbf{x}; \mu(\Theta), \Sigma = 1) p(\Theta)$ .

We consider a set of three different stochastic additive distortions with three different correlation scales — dubbed distortions A (spanning 5 bins), B (spanning 21 bins), and C (spanning 61 bins) respectively. We model these correlated distortions with convolutions with kernel sizes corresponding to their correlation scales. By convention, we fix the maximum of the kernels to one. Their position is sampled across the whole data length and their amplitude is sampled from  $\epsilon \sim \mathcal{U}(-b, b)$ , as in Equation 8. In Section III.1 we show how to adaptively learn the strength of the distortions parameter  $b$  during training.

During training, the data is affected by a single distortion at a time and we train in parallel a different network for each different correlation scale. This is to better showcase their differences at inference time, but in principle a single network would suffice for the framework. Here, we adopt the SNR strategy, introduced in Section II.5, that allows one to directly estimate the SNR (linked to the test statistic as in Equation 9) by learning the matched filter estimates  $\epsilon_\phi(\mathbf{x})$  and their variances  $\sigma_\phi^2$  (as defined in Equations C3 and C4). A comparison of the SNR and BCE training strategies with the analytical expectation for this example and further details are given in Appendix D.

In Figures 2 and 3 we show a summary of our results for

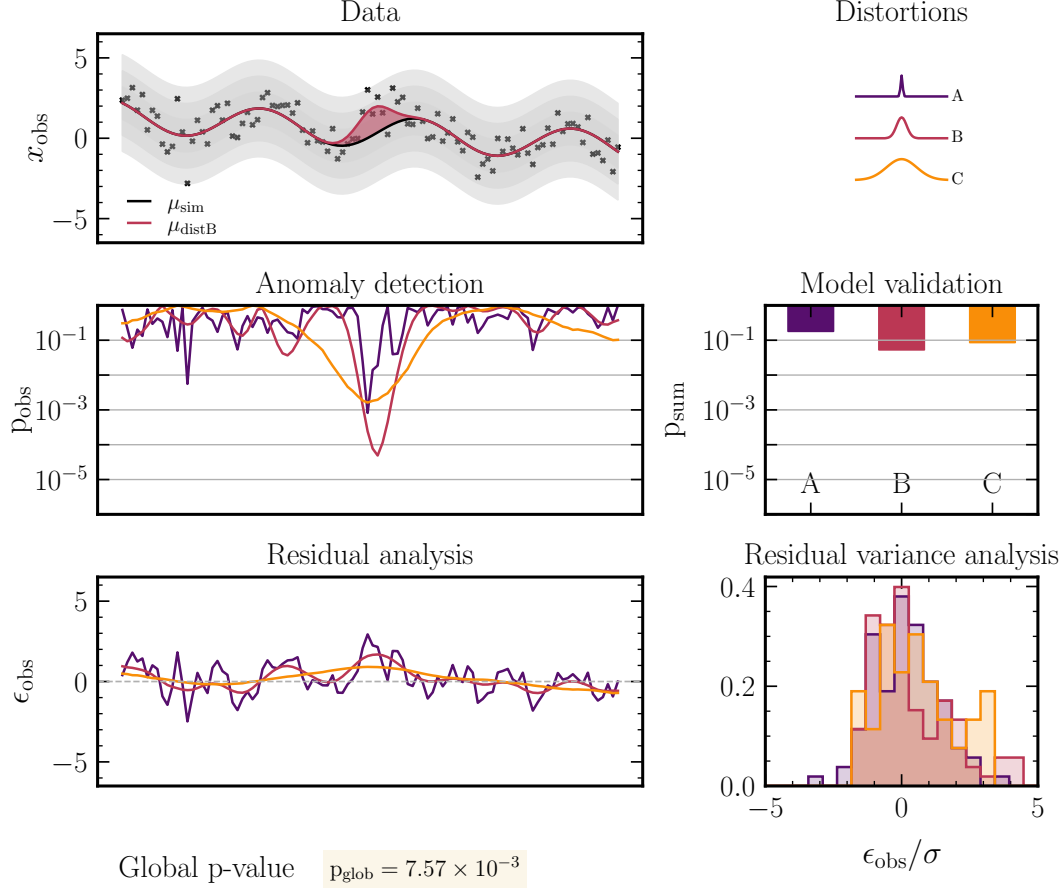


FIG. 2. Comprehensive summary of the framework results for the instructive example presented in Section III. The panels show the results by following the structure of our framework, as represented in the summary graphic Figure 1. The **upper-left panel** depicts scattered data points  $\mathbf{x}_{\text{obs}}$ , the baseline signal  $\mu_{\text{sim}}$ , and the signal distorted by an additive stochastic distortion of type B,  $\mu_{\text{distB}}$ , as described in Section III. The gray bands highlights the 1-, 2-, and 3- $\sigma$  regions of the baseline Gaussian noise. The **upper-right panel** visualizes the three types of deviations with different correlation scale under investigation. The results from different networks is color-coded based on the type of distortion they were trained on in the following panels. The **center-left panel** showcases the significance of localized distortions, the **center-right panel** the significance of the aggregated distortions, and the **bottom text** the global significance. Finally, the **bottom panels** the strength of the distortions in data space and their variance. These latter results are achievable only through the SNR training strategy (see Section II.5).

a single distortion of type B and for multiple distortions of type B respectively. In the upper left panel of both figures, we show the data points  $\mathbf{x}_{\text{obs}}$ , the baseline signal  $\mu_{\text{sim}}$ , and the signal distorted by an additive stochastic distortion of type B,  $\mu_{\text{distB}}$ . The gray shaded areas show the 1-, 2-, and 3-sigma Gaussian noise regions. The rest of the panels highlight the range of results obtained using our framework, and follow the same logic as represented in the summary, Figure 1. These can be described in more detail as follows:

*Anomaly detection (central left panel):* This panel shows the anomaly detection significance of localized distortions for the three correlation-scales, shown in the legend (upper right panel). In Figure 2, we see that there is a clear minimum p-value for the correct correlation scale of the individual distortion. Indeed, we see that the most powerful individual test corresponds to the correct correla-

tion scale at the right point in the data, aligning with matched filtering expectations. In Figure 3, where there are multiple distortions, we see that we are able to capture their relative significance, and again identify the correct correlation scale.

*Residual analysis (bottom left panel):* The residual analysis shows where the distortion is located in data space and how large it is in the same units as the input data. This is enforced by the convention that the maximum element of the kernel of the convolution for the correlated distortion is one. This is shown for the three correlation-scales. It is most useful if there is a localized excess. The behaviour as a residual is perhaps most obvious for the smallest correlation scale, where we see that it tracks upwards and downwards noise fluctuations on top of the signal.

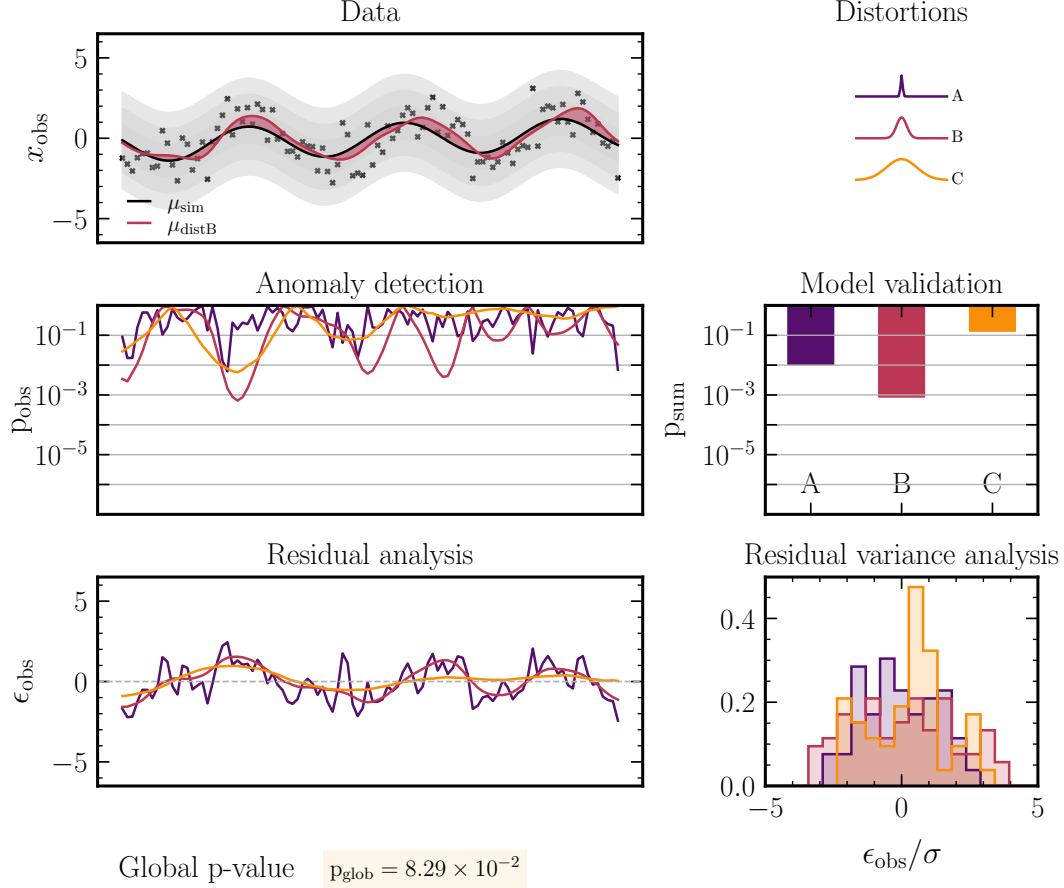


FIG. 3. Same as Figure 2, but applied to data distorted by many small distortions.

*Model validation (central right panel):* The aggregated p-value is useful to check if the model gives a good overall description of the data. It is especially useful when there is no clear single distortion, but many across the data, contributing to the signal, as in Figure 3. We can also use this to identify the most likely subset of distortions in our data (e.g. here the ‘B’ distortion is correctly singled out in Figures 2 and 3 by the analysis).

*Residual variance analysis (bottom left panel):* The residual variance analysis helps to visualize the model validation. In the presence of Gaussian noise it should be standard normally distributed, but the further it is from normality (under the assumptions in Section II), the greater the indication for misspecification.

*Global p-value (bottom text):* This is the overall p-value that accounts for the number of tests performed (see Section II.3). The p-values from the localized and aggregated tests in the second rows are individual p-values for specific tests. Here, we choose to treat the aggregated and localized tests on equal footing by combining them into a single global p-value, which accounts for all 303 tests under consideration (3 aggregated tests and 300 localized tests). Of course, one could opt to derive global

p-values separately for the localized and aggregated tests if desired. This choice can be adapted to the specific testing framework or interpretational needs.

### III.1. Self-calibrating distortions algorithm

An important consideration in setting up this framework is determining the variance of the distortion amplitude, as defined by the parameter  $b$  (see Equation 8). Ideally, with sufficiently flexible networks and infinite training data, the test results would be largely invariant to the specific variations of  $b$ , as long as it is not too small. In practice, however, choosing distortions that are only slightly larger than the natural statistical baseline model variations results in more efficient training.

In simple cases one can make an educated guess about  $b$  or even calculate it analytically, but it is a straightforward extension of the SNR training strategy (Appendix C) to make it adaptive so as to converge only to plausible distortion amplitudes. In other words, it converges to distortions that are significant enough to be detectable, but not so significant that they are clearly ruled out. One way to make this quantitative is to define

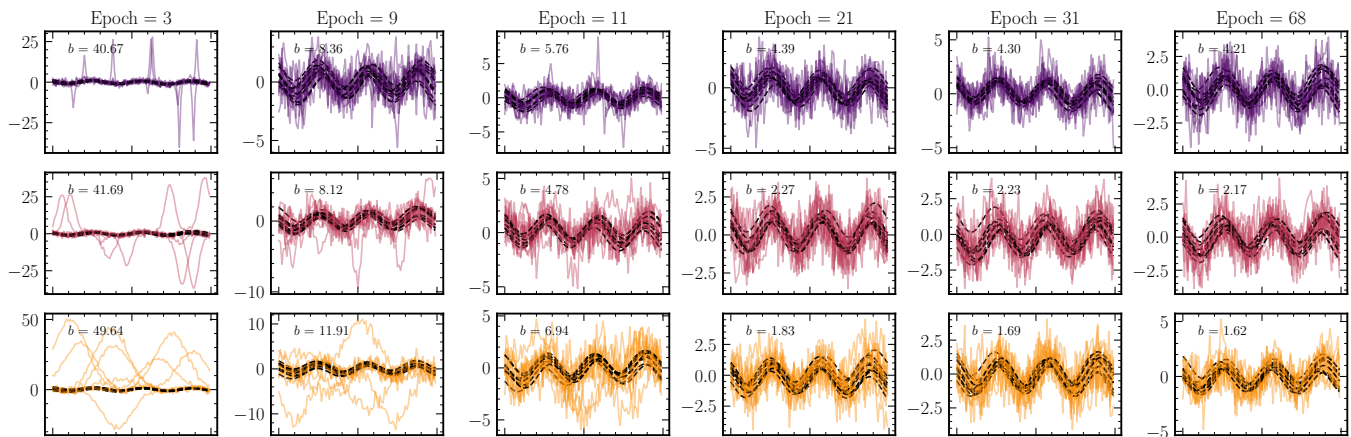


FIG. 4. Illustration of the adaptive training of distortion amplitudes in our framework. The figure shows how the generated distortions (color-coded to match the legend in Figure 2) dynamically adjust to envelop the deterministic part of the baseline model (black dashed lines) during training. The baseline model noise is not shown in the plot for clarity purposes. By adaptively tuning the distortion amplitude parameter  $b$  based on the learned variance  $\sigma^2$  and a desired maximum signal-to-noise ratio  $\text{SNR}_{\text{max}}$ , the distortions remain plausible.

a maximum SNR for a given distortion,  $\text{SNR}_{\text{max}}$ . This is realized by the maximum value of  $\epsilon$ , i.e.  $b$ . Writing this in terms of the variance  $\sigma$ , we see that:

$$b = \text{SNR}_{\text{max}} \sigma = \frac{\text{SNR}_{\text{max}}}{\sqrt{(\mathbf{n}^{(i)})^T \Sigma^{-1} \mathbf{n}^{(i)}}}. \quad (12)$$

Since  $\sigma^2$  is a learned parameter of the model in the SNR training strategy, we can anchor  $b$  to its learned value on the fly by choosing a desired  $\text{SNR}_{\text{max}}$  (for our example we set  $\text{SNR}_{\text{max}} = 5$ ), and generate training data with the more suitable distortion amplitude as the network is learning.

Figure 4 shows this in practice, highlighting how the distortions correctly end up enveloping the model.<sup>6</sup> This adaptive approach ensures that the alternative hypotheses explored during high-volume hypothesis testing are both challenging and realistic, enhancing the sensitivity of our method to subtle model discrepancies.

#### IV. APPLICATION TO GRAVITATIONAL WAVES

In this section, we move beyond the toy problem that we have investigated in detail above and use our framework to analyse real data. In particular, we take the example of gravitational waves (GWs), as detected by the LIGO-Virgo-Kagra collaboration. Specifically, we will focus on the first detection by the LIGO detectors [34–36],

<sup>6</sup> For simplicity, since in this example the variance  $\sigma^2$  is pretty much constant across distortions of the same correlation scale, we estimate a single  $b$  for all localized distortions of the same correlation scale, anchoring it to the mean value of  $\sigma^2$ .

GW150914, and illustrate how to perform an additional post-analysis quality check with our framework. Beyond this, our framework could be used to test other various aspects of GW data analysis pipelines. For example, it could be used to test for waveform systematics after carrying out Bayesian inference (see e.g. [37–39] for discussions on systematics in the LIGO context), compare and contrast different models for detector noise in the presence/absence of astrophysical signals [40, 41], or identify and flag detector artifacts or glitches (see e.g. [42]) that may result in a mismodelling or biasing of the parameter estimation. However, we postpone a full follow up of this application to a future work.

##### IV.1. Bayesian inference step

The starting point of our analysis is the circumstance where we have fitted some model to data, and are then looking to evaluate whether there is any evidence for mismodelling. In the GW context, we carry out a likelihood-based Bayesian inference analysis of the first LIGO event, GW150914 [34–36]. From a technical point of view, this involves (a) defining an analysis window and accessing the data from the GWOSC<sup>7</sup> [43], (b) estimating the power-spectral density  $S_n(f)$  (PSD) around the event, (c) choosing a signal model to fit to the data using an appropriate likelihood and sampler.

In this work, we use the `jimgw` library [44], built on top of the `ripple` waveform package [45] to analyse GW150914, with the priors and detector setups exactly as described in Ref. [44]. As far as (a) is concerned,

<sup>7</sup> Gravitational Wave Open Science Center



we analyse the data from both the Hanford (H1) and Livingston (L1) detectors, with a 4 s detection window centred on a GPS trigger time of 1126259462.4, sampled at 4096 Hz. For (b), we estimate the PSD using a data segment of length 16 s starting 32 s before the beginning of the detection window. Finally, for (c), we use the `IMRPhenomD` waveform model [46, 47], as implemented in the `ripple` codebase [45]. The result of this is a set of posterior samples  $\Theta \sim p(\Theta|h_{H_1}, h_{L_1})$  over the gravitational wave model parameters  $\Theta$ <sup>8</sup> given the detector data  $h_{H_1}, h_{L_1}$  for Hanford and Livingston, respectively. These (noise-free) posterior-predictive samples are visualised (after the data processing steps described below) in the top panel of Figure 5 alongside the real (processed) data for the Hanford detector.

## IV.2. Data processing setup

For this application, we use our model misspecification framework to check for deviations from the posterior-predictive distribution obtained in the inference step. We do this for whitened time domain data  $d_w(t)$  in the Hanford detector as a concrete example.

To generate simulated data in the time domain under our null hypothesis model  $H_0$ , we use the following procedure. First, we take a posterior sample  $\Theta \sim p(\Theta|h_{H_1}, h_{L_1})$  and generate the frequency domain signal  $\tilde{h}_{H_1}(f; \Theta)$  using the `IMRPhenomD` waveform model. In addition, we generate noise in the frequency domain  $\tilde{n}(f)$  by sampling from the PSD estimated during the inference step. Then, before taking the inverse Fast Fourier Transform (FFT), we normalize the frequency domain strain by the square-root of the PSD and filter the data using a bandpass filter between 20 Hz and 1024 Hz<sup>9</sup>, with additional notches of a width 0.1 Hz removed at 60, 120, and 180 Hz. For visualization purposes, we also downsample the resulting time domain data by a factor 8 and consider a time window of 0.2 s around the trigger time. To summarise in the language of Section II, our baseline model in this context is given by  $p_{\text{sim}}(d_w(t)) = \int d\Theta p(d_w(t)|\Theta, S_n)p(\Theta|h_{H_1}, h_{L_1})$ . It is worth stressing that using posterior samples here is somewhat for illustrative purposes, since it is expected that this will slightly overestimate distortions that are degenerate with the effect of model parameter changes. In general, parameters should be drawn from the full prior or truncated versions thereof [29, 31].

To analyse the real data from the Hanford detector for GW150914, there is one additional step that we must carry out. In particular, before taking the FFT of the raw

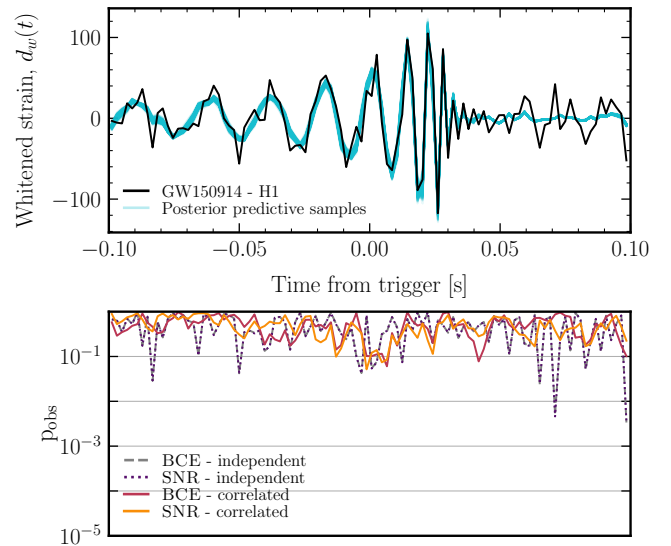


FIG. 5. **Top panel:** GW150914 data for the Hanford detector and posterior-predictive distribution samples from the Bayesian inference step as described in Section IV.1 and processed as described in Section IV.2. **Bottom panel:** Results of our framework for GW150914 from the Hanford detector. We test for an independent bin-wise distortion and a correlated one, using both our training strategies, dubbed BCE and SNR respectively (Section II.5). 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.

time domain data to move to the frequency domain where the whitening and filtering is carried out, we must apply a window function. Here, we apply the same Tukey window function as in the original Bayesian inference analysis (see the `jingw` code [44]).

## IV.3. Results

For the purpose of this example, our baseline model is the one described in the previous section. As deviations from the base simulator, we consider two types of stochastic additive distortions: an independent, bin-wise distortion for each processed time step (totally 102 bins), and a correlated distortion spanning eleven processed time steps generated through a convolution in the same way as the one for the example in Section III.

For both classes of distortions, we tested both our training strategies, BCE and SNR (Section II.5). The results are presented in Figure 5. As expected, no significant anomaly is present in the modeling of GW150914, with global p-values for all the types of analyses of around a few tenths. We do not report a plot for the aggregated test statistic for model validation since it also shows similarly a very small significance.

We have also tested the results against different processing steps of our data. In particular, we also tried

<sup>8</sup> Here,  $\Theta$  consists of all the standard intrinsic (as relevant to the `IMRPhenomD` waveform model [46, 47]) and extrinsic parameters describing the properties and location of the GW source.

<sup>9</sup> Note that this is the same frequency range used during the inference step [44].

using unwhitened data, with a 40-400 Hz bandpass filter, or downsampling by a factor of 2 instead of 8. In the former case, this introduces the non-trivial feature that the noise in the time domain is correlated. Nonetheless, in all of these additional cases the results were consistent with those shown in Figure 5.

Note that the example application that we have demonstrated here is only one option as far as GW data analysis is concerned, although the application to real data is an important step. As discussed at the start of this section, it would be interesting to take this further and study either a broader class of events, or look to calibrate detector performance and waveform models.

## V. DISCUSSION

This work focuses on detecting a wide variety of model discrepancies through high-volume hypothesis testing. However, we do not address how to adjust inference algorithms to accommodate these discrepancies if they are identified, nor do we examine the robustness of estimators in the presence of systematic errors. We acknowledge several recent efforts aimed at the first issue — specifically, adapting inference algorithms trained on one simulator for application to another. These approaches involve selecting subsets of observables that remain relatively consistent across different simulators [48, 49], employing domain adaptation techniques during the training phase [50, 51], or correcting minor inaccuracies in the simulations through calibration methods [52]. Regarding the second issue, the robustness of SBI techniques, such as neural posterior estimation and neural ratio estimation, to distributional shifts was recently investigated by Ref. [53].

With this context in mind, the main advantage of high-volume hypothesis testing for many test statistics  $t_i$ , localized in the space of distortions, is that any mismodelling can be systematically identified and visualized, as e.g. in Figure 2. An aggregated test statistic on the other hand can only reveal that some part of the data is mismodeled, but it cannot tell us exactly how. By directly targeting localized test statistics  $t_i$ , our method opens up a plethora of hypothesis tests for the same trained network, since the way we combine them into aggregated tests is entirely flexible. In Section II, we highlighted the summed test statistic given by  $t_{\text{sum}} = \sum_i t_i$ . However, in principle many other options are open. For example, we could define aggregated tests for specific subsets of the distortions (e.g., focusing on distortions at a particular correlation scale, or on the left half of the data space etc.), or construct complex statistics like a “double-excess” test statistic for the probability of two large excesses anywhere in the data. Such a construction, which would be analytically challenging, is now trivial to implement and to compute the significance for, by summing only the largest and second-largest localized test values.

On the other hand, this flexibility does come with com-

putational costs, as the dimensionality of the network output scales with the number of alternative hypotheses (see e.g. Equations A4 and C3). Hence, when looking for a wider variety of distortions, it could be preferable to directly target global test statistics, or test statistics at the level of informative lower dimensional summaries of the data.

One of the main benefits of our framework is that the choice of sampling distributions for the alternative hypotheses (or equivalently the form of the model augmentation  $\mathbf{n}$ ) can be tailored to the questions of interest. Whilst, in principle, any choice will lead to a concrete set of test statistics, there are likely to be trade-offs between complexity, specificity, and functional form of the contrastive distribution. Consequently, the implications of these choices on the statistical power of the test to detect significant features will vary on a case by case basis. Regardless, the adaptive algorithm described in Section III.1 can be employed to set the variations of the strength of the distortions, so as they are plausible. Furthermore, we note that this model augmentation need not occur in data space directly. For example, in gravitational wave analyses, the augmentation could be performed at the signal level in frequency space, offering additional flexibility in defining meaningful alternative hypotheses.

As a last point of discussion, we expand upon our findings regarding the main upsides and downsides of the two training strategies (Section II.5). We note that the followings are not exhaustive results, but intuitions from our experiments that can serve as useful starting point for further investigation. By construction, the SNR training strategy provides clear and interpretable results, allowing one to visually inspect the amplitude of deviations in specific distortion directions through  $\epsilon$  (see e.g. the bottom left panel of Figure 2). Furthermore, it allows one to adaptively learn during training the most appropriate amplitude for the distortions of interest (Section III.1 and Figure 4). On the other hand, the BCE training strategy does not depend on any prior assumption. As such, we expect it to better perform in general scenarios. Indeed, we have found that it seems to give a slightly better performance when describing excesses and localizing distortions, as shown, e.g., in Figure 7.

## VI. CONCLUSION

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 simulation-based inference context. By leveraging classical concepts, it provides a flexible and comprehensive approach to simultaneously perform many hypothesis tests and quantify their statistical significance (Section II). In Section III, we demonstrated the application of the framework to a toy example, before applying it to real data in the gravitational wave context in Sec-

tion IV. The main conclusions of this work are then as follows:

*Detection of model misspecification:* Our framework uses high-volume hypothesis testing to detect model misspecification. This framing allows us to unify the ideas of anomaly detection (localized test statistics) and model validation (aggregated tests), as described in Section II. In addition, due to its simulation-based nature, it also is extremely flexible as far as the classes and types of distortion that can be searched for.

*Efficient:* As described in Section II, one way to think about our framework is as a collection of hypothesis tests comparing the various distorted data models to a base simulator. Via the training strategies described in Section II.5, 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. Furthermore, for the SNR training strategy, we have developed an adaptive algorithm that can be used to calibrate the scale of distortions searched for in the data (Section III.1).

*Principled:* Although the method is simulation-based, it is firmly rooted in classical statistical principles. Indeed,

we discussed at length the connections to classical hypothesis testing in Section II. For example, we showed how our framework reduces to the classical concepts of matched filtering and  $\chi^2$  goodness-of-fit tests under certain conditions. This adds an additional level of interpretability to the results derived in this work.

To conclude, we argued at the start of this work that there was a crucial need for model misspecification frameworks in the context of SBI. The presented work may serve as a step towards that direction. This will allow us to move beyond the parameter estimation regime, and encourage more end-to-end, simulation-based analysis pipelines for real-world data settings across astrophysics, particle physics, and cosmology.

## ACKNOWLEDGEMENTS

The work of NAM, CW, and JA was supported by a project that has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (Grant agreement No. 864035 – UnDark). JA is supported by a fellowship from the Kavli Foundation.

- 
- [1] J. Skilling, *Bayesian Analysis* **1**, 833 (2006).
  - [2] J. Neyman and E. S. Pearson, *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **231**, 289 (1933).
  - [3] T. A. Collaboration, *Physics Letters B* **716**, 1 (2012).
  - [4] K. Cranmer, J. Brehmer, and G. Louppe, *Proceedings of the National Academy of Sciences* **117**, 30055 (2020), arXiv:1911.01429 [cs, stat].
  - [5] S. Mishra-Sharma and K. Cranmer, *Physical Review D* **105**, 063017 (2022), arXiv:2110.06931 [astro-ph.HE].
  - [6] N. Anau Montel, A. Coogan, C. Correa, K. Karchev, and C. Weniger, *Monthly Notices of the Royal Astronomical Society* **518**, 2746 (2022), arXiv:2205.09126 [astro-ph.CO].
  - [7] B. Tucci and F. Schmidt, *Journal of Cosmology and Astroparticle Physics* **05**, 063 (2024), arXiv:2310.03741 [astro-ph.CO].
  - [8] G. F. Abellán, G. Cañas Herrera, M. Martinelli, O. Savchenko, D. Sciotti, and C. Weniger, “Fast likelihood-free inference in the LSS Stage IV era,” (2024), arXiv:2403.14750 [astro-ph.CO].
  - [9] J. Alsing, T. Charnock, S. Feeney, and B. Wandelt, *Monthly Notices of the Royal Astronomical Society* **488**, 4440 (2019), arXiv:1903.00007 [astro-ph.CO].
  - [10] C. Modi, S. Pandey, M. Ho, C. Hahn, B. R.-S. Blencard, and B. Wandelt, “Sensitivity Analysis of Simulation-Based Inference for Galaxy Clustering,” (2023), arXiv:2309.15071 [astro-ph.CO].
  - [11] K. Karchev and R. Trotta, (2024), arXiv:2409.03837 [astro-ph.CO].
  - [12] U. Bhardwaj, J. Alvey, B. K. Miller, S. Nissanke, and C. Weniger, *Physical Review D* **108**, 042004 (2023), arXiv:2304.02035 [gr-qc].
  - [13] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, *Phys. Rev. Lett.* **127**, 241103 (2021), arXiv:2106.12594 [gr-qc].
  - [14] N. Jeffrey and B. D. Wandelt, *Machine Learning: Science and Technology* **5**, 015008 (2024), arXiv:2305.11241 [astro-ph, stat].
  - [15] T. Gessey-Jones and W. J. Handley, (2023), arXiv:2309.06942 [astro-ph.IM].
  - [16] A. S. Mancini, M. M. Docherty, M. A. Price, and J. D. McEwen, *RAS Techniques and Instruments* **2**, 710 (2023), arXiv:2207.04037 [astro-ph, physics:physics].
  - [17] K. Cranmer, J. Pavez, and G. Louppe, “Approximating likelihood ratios with calibrated discriminative classifiers,” (2016), arXiv:1506.02169 [stat.AP].
  - [18] N. Dalmaso, R. Izbicki, and A. B. Lee, “Confidence sets and hypothesis testing in a likelihood-free inference setting,” (2020), arXiv:2002.10399 [stat.ME].
  - [19] L. Heinrich, “Learning Optimal Test Statistics in the Presence of Nuisance Parameters,” (2022), arXiv:2203.13079 [physics, stat].
  - [20] M. von Wietersheim-Kramsta, K. Lin, N. Tessore, B. Joachimi, A. Loureiro, R. Reischke, and A. H. Wright, “KiDS-SBI: Simulation-Based Inference Analysis of KiDS-1000 Cosmic Shear,” (2024).
  - [21] H. White, *Econometrica* **50**, 1 (1982), publisher: [Wiley, Econometric Society].
  - [22] P. Cannon, D. Ward, and S. M. Schmon, *Investigating the Impact of Model Misspecification in Neural*

- Simulation-based Inference*, Tech. Rep. arXiv:2209.01845 (arXiv, 2022) arXiv:2209.01845 [cs, stat] type: article.
- [23] M. Schmitt, P.-C. Bürkner, U. Köthe, and S. T. Radev, “Detecting Model Misspecification in Amortized Bayesian Inference with Neural Networks: An Extended Investigation,” (2024), arXiv:2406.03154.
  - [24] D. Huang, A. Bharti, A. Souza, L. Acerbi, and S. Kaski, *Advances in Neural Information Processing Systems* **36**, 7289 (2023).
  - [25] A. Wehenkel, J. L. Gamella, O. Sener, J. Behrmann, G. Sapiro, M. Cuturi, and J.-H. Jacobsen, “Addressing Misspecification in Simulation-based Inference through Data-driven Calibration,” (2024).
  - [26] D. Ward, P. Cannon, M. Beaumont, M. Fasiolo, and S. Schmon, *Advances in Neural Information Processing Systems* **35**, 33845 (2022).
  - [27] C. Dellaporta, J. Knoblauch, T. Damoulas, and F.-X. Briol, in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics* (PMLR, 2022) pp. 943–970, ISSN: 2640-3498.
  - [28] K. Cranmer, “Practical statistics for the LHC,” (2015), arXiv:1503.07622 [physics.data-an].
  - [29] B. K. Miller, A. Cole, P. Forré, G. Louppe, and C. Weniger, “Truncated Marginal Neural Ratio Estimation,” (2021).
  - [30] G. Papamakarios, D. C. Sterratt, and I. Murray, “Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows,” (2019), arXiv:1805.07226 [stat.ML].
  - [31] N. Anau Montel, J. Alvey, and C. Weniger, “Scalable inference with Autoregressive Neural Ratio Estimation,” (2023), arXiv:2308.08597 [astro-ph.IM].
  - [32] S. S. Wilks, *The Annals of Mathematical Statistics* **9**, 60 (1938).
  - [33] E. Gross and O. Vitells, *The European Physical Journal C* **70**, 525–530 (2010).
  - [34] B. P. Abbott *et al.* (LIGO Scientific, Virgo), *Phys. Rev. Lett.* **116**, 061102 (2016), arXiv:1602.03837 [gr-qc].
  - [35] B. P. Abbott *et al.* (LIGO Scientific, Virgo), *Phys. Rev. Lett.* **116**, 241102 (2016), arXiv:1602.03840 [gr-qc].
  - [36] B. P. Abbott *et al.* (LIGO Scientific, Virgo), *Phys. Rev. D* **93**, 122003 (2016), arXiv:1602.03839 [gr-qc].
  - [37] B. P. Abbott *et al.* (LIGO Scientific, Virgo), *Class. Quant. Grav.* **34**, 104002 (2017), arXiv:1611.07531 [gr-qc].
  - [38] R. Gamba, M. Breschi, S. Bernuzzi, M. Agathos, and A. Nagar, *Phys. Rev. D* **103**, 124015 (2021), arXiv:2009.08467 [gr-qc].
  - [39] K. K. H. Lam, K. W. K. Wong, and T. D. P. Edwards, *Phys. Rev. D* **109**, 124009 (2024), arXiv:2306.17245 [gr-qc].
  - [40] B. P. Abbott *et al.* (LIGO Scientific, Virgo), *Phys. Rev. Lett.* **116**, 131103 (2016), arXiv:1602.03838 [gr-qc].
  - [41] R. Legin, M. Isi, K. W. K. Wong, Y. Hezaveh, and L. Perreault-Levasseur, (2024), arXiv:2410.19956 [astro-ph.IM].
  - [42] S. B. Coughlin *et al.*, *Phys. Rev. D* **99**, 082002 (2019), arXiv:1903.04058 [astro-ph.IM].
  - [43] R. Abbott *et al.* (LIGO Scientific, Virgo), *SoftwareX* **13**, 100658 (2021), arXiv:1912.11716 [gr-qc].
  - [44] K. W. K. Wong, M. Isi, and T. D. P. Edwards, *Astrophys. J.* **958**, 129 (2023), arXiv:2302.05333 [astro-ph.IM].
  - [45] T. D. P. Edwards, K. W. K. Wong, K. K. H. Lam, A. Coogan, D. Foreman-Mackey, M. Isi, and A. Zimmerman, *Phys. Rev. D* **110**, 064028 (2024), arXiv:2302.05329 [astro-ph.IM].
  - [46] S. Husa, S. Khan, M. Hannam, M. Pürrer, F. Ohme, X. Jiménez Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044006 (2016), arXiv:1508.07250 [gr-qc].
  - [47] S. Khan, S. Husa, M. Hannam, F. Ohme, M. Pürrer, X. Jiménez Forteza, and A. Bohé, *Phys. Rev. D* **93**, 044007 (2016), arXiv:1508.07253 [gr-qc].
  - [48] N. Echeverri, F. Villaescusa-Navarro, C. Chawak, Y. Ni, C. Hahn, E. Hernandez-Martinez, R. Teyssier, D. Angles-Alcazar, K. Dolag, and T. Castro, “Cosmology with one galaxy? – the astrid model and robustness,” (2023), arXiv:2304.06084 [astro-ph.CO].
  - [49] N. S. M. de Santi, H. Shao, F. Villaescusa-Navarro, L. R. Abramo, R. Teyssier, P. Villanueva-Domingo, Y. Ni, D. Anglés-Alcázar, S. Genel, E. Hernández-Martínez, U. P. Steinwandel, C. C. Lovell, K. Dolag, T. Castro, and M. Vogelsberger, *The Astrophysical Journal* **952**, 69 (2023).
  - [50] P. Swierc, M. Tamargo-Arizmendi, A. Čiprijanović, and B. D. Nord, “Domain-adaptive neural posterior estimation for strong gravitational lens analysis,” (2024), arXiv:2410.16347 [astro-ph.IM].
  - [51] J.-Y. Lee, J.-h. Kim, M. Jung, B. K. Oh, Y. Jo, S. Park, J. Lee, Y.-S. Ting, and H. S. Hwang, *The Astrophysical Journal* **975**, 38 (2024).
  - [52] H. Jia, “Cosmological analysis with calibrated neural quantile estimation and approximate simulators,” (2024), arXiv:2411.14748 [astro-ph.CO].
  - [53] A. Filipp, Y. Hezaveh, and L. Perreault-Levasseur, “Robustness of neural ratio and posterior estimators to distributional shifts for population-level dark matter analysis in strong gravitational lensing,” (2024), arXiv:2411.05905 [astro-ph.CO].
  - [54] A. Mao, M. Mohri, and Y. Zhong, “Cross-entropy loss functions: Theoretical analysis and applications,” (2023), arXiv:2304.07288 [cs.LG].



## APPENDICES

Appendix A describes a general training strategy for our model misspecification testing framework. An alternative training strategy, that stems from the connection of the framework to classical testing (Appendix B), is presented in Appendix C. Furthermore, a comparison of the two strategies to each other and the analytical counterpart is presented in Appendix D.

### Appendix A: Classifier-based training strategy

As discussed in Section II, we want to approximate an ensemble of  $N_{\text{alt}}$  test statistics  $t_i$  (Equation 4) via neural networks. As originally proposed in Ref. [17], discriminative classifiers can be used to approximate the generalized likelihood ratio statistic when only a generative model for the data is available. The classifiers  $f_{i,\phi}$  can be optimized through gradient descent using the standard binary cross-entropy loss [54] as the optimization objective,

$$\mathcal{L}_{\text{BCE}}^{(i)}[f_{i,\phi}(\mathbf{x})] = \mathbb{E}_{\mathbf{x} \sim p_{\text{sim}}(\mathbf{x})} [-\ln \sigma(f_{i,\phi}(\mathbf{x}))] + \mathbb{E}_{\mathbf{x} \sim p_{\text{dist}}(\mathbf{x}|i)} [-\ln \sigma(1 - f_{i,\phi}(\mathbf{x}))], \quad (\text{A1})$$

although other classes of loss functions could also be employed [e.g. 14].

After optimisation, each classifier estimates the likelihood ratio

$$f_{i,\phi}(\mathbf{x}) \approx \ln \frac{p_{\text{dist}}(\mathbf{x}|i)}{p_{\text{sim}}(\mathbf{x})}, \quad (\text{A2})$$

and we can define the localized test statistics in terms of the its output via

$$\hat{t}_i(\mathbf{x}) = 2f_{i,\phi}(\mathbf{x}) \simeq -2 \ln \frac{p_{\text{sim}}(\mathbf{x})}{p_{\text{dist}}(\mathbf{x}|i)}. \quad (\text{A3})$$

In the interest of efficiency, rather than training individual classifiers for each possible localized distortion  $i$ , we typically train single multi-output networks of the form

$$\mathbf{f}_\phi(\mathbf{x}) : \mathcal{D} \rightarrow \mathbb{R}^{N_{\text{alt}}}, \quad (\text{A4})$$

where  $\mathcal{D}$  refers to the data space of  $\mathbf{x}$ . The total loss is given as a sum over the individual losses,

$$\mathcal{L}_{\text{BCE}}[\mathbf{f}_\phi(\mathbf{x})] = \sum_{i=1}^{N_{\text{alt}}} \mathcal{L}_{\text{BCE}}^{(i)}[f_{i,\phi}(\mathbf{x})], \quad (\text{A5})$$

where the sum runs over all hypotheses  $H_i$ .

### Appendix B: Connection to classical testing frameworks

In this appendix, we derive the connection between the proposed distortion-driven model misspecification test-

ing, presented in Section II, and classical testing frameworks. Throughout this derivation, a number of assumptions (highlighted in bold) will be made in order to align our general framework with more specific traditional testing methods.

To begin, we note that traditional tests often use *profiled likelihoods*, whereas our method considers *likelihoods marginalized over model parameters*. Thus, in order to establish the connection between our method and classical testing frameworks, we first compute the relationship between marginal and profile likelihoods for  $H_0$  and  $H_i$ .

**Marginal null hypothesis.** Given a model likelihood  $p(\mathbf{x}|\Theta)$ , our base hypothesis is defined by *marginalizing* over its parameters  $\Theta$

$$p(\mathbf{x}|H_0) = \int d\Theta p(\Theta)p(\mathbf{x}|\Theta) \quad (\text{B1})$$

where  $p(\Theta)$  is the prior distribution over the model parameters. In contrast, the *profile* likelihood  $p(\mathbf{x}|\Theta_{\mathbf{x}}^*)$ , is the value of the likelihood at its maximum-likelihood estimator (MLE)

$$\Theta_{\mathbf{x}}^* = \arg \max_{\Theta} p(\mathbf{x}|\Theta). \quad (\text{B2})$$

To connect marginal and profile likelihoods, we assume the **large sample limit**, where it is possible to approximate the likelihood function as Gaussian in  $\Theta$ ,<sup>10</sup> centered around the MLE  $\Theta_{\mathbf{x}}^*$  and with covariance  $\Sigma_{\Theta_{\mathbf{x}}^*}$

$$p(\mathbf{x}|\Theta) \propto \mathcal{N}(\Theta; \Theta_{\mathbf{x}}^*, \Sigma_{\Theta_{\mathbf{x}}^*}). \quad (\text{B3})$$

Under the above assumption, we can connect marginal and profile likelihoods as follows

$$\begin{aligned} p(\mathbf{x}|H_0) &= p(\mathbf{x}|\Theta_{\mathbf{x}}^*) \int d\Theta p(\Theta) \frac{p(\mathbf{x}|\Theta)}{p(\mathbf{x}|\Theta_{\mathbf{x}}^*)} \\ &\stackrel{\text{B3}}{\simeq} p(\mathbf{x}|\Theta_{\mathbf{x}}^*) \times \int_{\Theta \sim \mathcal{N}(\Theta_{\mathbf{x}}^*, \Sigma_{\Theta_{\mathbf{x}}^*})} p(\Theta) \sqrt{(2\pi)^d \det \Sigma_{\Theta_{\mathbf{x}}^*}}. \end{aligned} \quad (\text{B4})$$

**Marginal alternative hypothesis.** Let us now consider, as alternative hypotheses  $H_i$ , the case where the distortions are in the form of simple **stochastic additive non-Gaussian distortions** in specific noise directions  $\mathbf{n}^{(i)}$ ,

$$H_i : \tilde{\mathbf{x}} = \mathbf{x} + \epsilon \cdot \mathbf{n}^{(i)} \quad \text{with} \quad \epsilon \sim \mathcal{U}(-b, b), \quad (\text{B5})$$

where  $\mathbf{x} \sim p(\mathbf{x}|H_0)$ . Note that, although  $\epsilon$  is a random variable, the noise directions  $\mathbf{n}^{(i)}$  are considered to be *fixed*. The bounds  $b$  for the prior of  $\epsilon$  are chosen large enough so that  $H_i$  is significantly different from  $H_0$ , while specific choices will be discussed below in Section B.1.

<sup>10</sup> Note that this does not require that  $p(\mathbf{x}|\Theta)$  is Gaussian in  $\mathbf{x}$ .



In this case, the likelihood for the  $H_i$  hypothesis can be expressed as a convolution of the likelihood under  $H_0$  via,

$$\begin{aligned} p(\mathbf{x}|H_i) &= \int d\epsilon p(\epsilon) p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | H_0) \\ &= \int d\epsilon p(\epsilon) \int d\Theta p(\Theta) p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta). \end{aligned} \quad (\text{B6})$$

Following the same steps as in the previous section to derive Equation B4 for  $p(\mathbf{x}|\Theta)$ , it is possible to connect the marginal and the profile likelihood of  $p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta)$  assuming the large sample limit. Here, to facilitate our ultimate goal, i.e. to compute the ratio between  $p(\mathbf{x}|H_0)$  and  $p(\mathbf{x}|H_i)$  for the test statistic, we consider the additional assumption that the **MLE of  $p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta)$  is not significantly correlated with the distortions  $\epsilon \mathbf{n}^{(i)}$** . In other words, we assume that the best-fit value for  $\Theta$  is not significantly affected by a single extra noise degree of freedom

$$\Theta_{\mathbf{x}}^* \simeq \Theta_{\mathbf{x} - \epsilon \mathbf{n}^{(i)}}^*. \quad (\text{B7})$$

With the above assumption, we can derive,

$$\begin{aligned} p(\mathbf{x}|H_i) &= \int d\epsilon p(\epsilon) \int d\Theta p(\Theta) p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta) \\ &\stackrel{\text{B7}}{=} \int d\epsilon p(\epsilon) p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta_{\mathbf{x}}^*) \\ &\quad \int d\Theta p(\Theta) \frac{p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta)}{p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta_{\mathbf{x}}^*)} \\ &\stackrel{\text{B3, B7}}{\simeq} \int d\epsilon p(\epsilon) p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta_{\mathbf{x}}^*) \\ &\quad \times \mathbb{E}_{\Theta \sim \mathcal{N}(\Theta_{\mathbf{x}}^*, \Sigma_{\Theta_{\mathbf{x}}^*})} p(\Theta) \sqrt{(2\pi)^d \det \Sigma_{\Theta_{\mathbf{x}}^*}}. \end{aligned} \quad (\text{B8})$$

We can further simplify the above expression by introducing the MLE for  $\epsilon$ ,  $\epsilon_{\mathbf{x}}^*$ , and the second derivative around the curvature,  $\sigma_{\epsilon_{\mathbf{x}}^*}^{-2}$ ,

$$\epsilon_{\mathbf{x}}^* \equiv \arg \max_{\epsilon} p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta_{\mathbf{x}}^*), \quad (\text{B9})$$

$$\sigma_{\epsilon_{\mathbf{x}}^*}^{-2} \equiv \partial_{\epsilon}^2 \ln p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta_{\mathbf{x}}^*)|_{\epsilon = \epsilon_{\mathbf{x}}^*}. \quad (\text{B10})$$

We then obtain

$$\begin{aligned} p(\mathbf{x}|H_i) &\simeq p(\mathbf{x} - \epsilon_{\mathbf{x}}^* \mathbf{n}^{(i)} | \Theta_{\mathbf{x}}^*) \\ &\quad \times \int d\epsilon p(\epsilon) \frac{p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta_{\mathbf{x}}^*)}{p(\mathbf{x} - \epsilon_{\mathbf{x}}^* \mathbf{n}^{(i)} | \Theta_{\mathbf{x}}^*)} \\ &\quad \times \mathbb{E}_{\Theta \sim \mathcal{N}(\Theta_{\mathbf{x}}^*, \Sigma_{\Theta_{\mathbf{x}}^*})} p(\Theta) \sqrt{(2\pi)^d \det \Sigma_{\Theta_{\mathbf{x}}^*}} \\ &\simeq p(\mathbf{x} - \epsilon_{\mathbf{x}}^* \mathbf{n}^{(i)} | \Theta_{\mathbf{x}}^*) p(\epsilon_{\mathbf{x}}^*) \sqrt{2\pi \sigma_{\epsilon_{\mathbf{x}}^*}^2} \\ &\quad \times \mathbb{E}_{\Theta \sim \mathcal{N}(\Theta_{\mathbf{x}}^*, \Sigma_{\Theta_{\mathbf{x}}^*})} p(\Theta) \sqrt{(2\pi)^d \det \Sigma_{\Theta_{\mathbf{x}}^*}}, \end{aligned} \quad (\text{B11})$$

where, in the second step, the integral over  $\epsilon$  is approximated by assuming that  $p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta_{\mathbf{x}}^*)$  is sharply peaked around  $\epsilon_{\mathbf{x}}^*$ , as expected in the large sample limit.

**Marginal test statistic.** We have now all the elements to compute the test statistic quantity of interest (Equation 4),

$$\begin{aligned} t_i(\mathbf{x}) &= -2 \ln \frac{p(\mathbf{x}|H_0)}{p(\mathbf{x}|H_i)} \\ &\stackrel{\text{B4, B11}}{\simeq} -2 \ln \frac{p(\mathbf{x} | \Theta_{\mathbf{x}}^*)}{p(\mathbf{x} - \epsilon_{\mathbf{x}}^* \mathbf{n}^{(i)} | \Theta_{\mathbf{x}}^*) p(\epsilon_{\mathbf{x}}^*) \sqrt{2\pi \sigma_{\epsilon_{\mathbf{x}}^*}^2}} \end{aligned} \quad (\text{B12})$$

To connect the above quantity to classical analysis frameworks, we consider their common assumption of a **Gaussian likelihood function**, and define

$$p(\mathbf{x} | \Theta) \equiv \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}(\Theta), \Sigma), \quad (\text{B13})$$

$$p(\mathbf{x} - \epsilon \mathbf{n}^{(i)} | \Theta) \equiv \mathcal{N}(\mathbf{x} - \epsilon \mathbf{n}^{(i)}; \boldsymbol{\mu}(\Theta), \Sigma) \quad (\text{B14})$$

where  $\boldsymbol{\mu}(\Theta)$  is the model prediction and  $\Sigma$  the likelihood covariance matrix.

Given the Gaussian assumption, we can straightforwardly compute the value of the MLE of  $\epsilon$  and its variance

$$\epsilon_{\mathbf{x}}^* \stackrel{\text{B9, B14}}{=} \frac{\Delta \mathbf{x}^T \Sigma^{-1} \mathbf{n}^{(i)}}{(\mathbf{n}^{(i)})^T \Sigma^{-1} \mathbf{n}^{(i)}} \quad (\text{B15})$$

$$\sigma_{\epsilon_{\mathbf{x}}^*}^{-2} \stackrel{\text{B10, B14}}{=} (\mathbf{n}^{(i)})^T \Sigma^{-1} \mathbf{n}^{(i)} \quad (\text{B16})$$

where we have defined the residual between the data and the maximum-likelihood model prediction,

$$\Delta \mathbf{x} \equiv \mathbf{x} - \boldsymbol{\mu}(\Theta_{\mathbf{x}}^*). \quad (\text{B17})$$

Finally, in the case of a Gaussian likelihood with non-Gaussian additive distortions, assuming the large sample limit, and that the model parameter MLE is not significantly affected by the distortions, we obtain that the marginal likelihood ratio can be written as

$$t_i(\mathbf{x}) = -2 \ln \frac{p(\mathbf{x}|H_0)}{p(\mathbf{x}|H_i)} \stackrel{11}{\simeq} \text{SNR}_i^2(\mathbf{x}) + C \quad (\text{B18})$$

where we have introduced the signal-to-noise ratio

<sup>11</sup> Expanding the computations

$$\begin{aligned} t_i &\stackrel{\text{B12, B13, B14, B17}}{\simeq} -(\Delta \mathbf{x} - \epsilon_{\mathbf{x}}^* \mathbf{n}^{(i)})^T \Sigma^{-1} (\Delta \mathbf{x} - \epsilon_{\mathbf{x}}^* \mathbf{n}^{(i)}) \\ &\quad + \Delta \mathbf{x}^T \Sigma^{-1} \Delta \mathbf{x} + 2 \ln p(\epsilon_{\mathbf{x}}^*) + \ln(2\pi \sigma_{\epsilon_{\mathbf{x}}^*}^2) \\ &\stackrel{\text{B15}}{=} \left( \frac{\Delta \mathbf{x}^T \Sigma^{-1} \mathbf{n}^{(i)}}{\sqrt{(\mathbf{n}^{(i)})^T \Sigma^{-1} \mathbf{n}^{(i)}}} \right)^2 + 2 \ln p(\epsilon_{\mathbf{x}}^*) + \ln(2\pi \sigma_{\epsilon_{\mathbf{x}}^*}^2) \\ &\stackrel{\text{B15, B16}}{=} \left( \frac{\epsilon_{\mathbf{x}}^*}{\sigma_{\epsilon_{\mathbf{x}}^*}} \right)^2 + 2 \ln p(\epsilon_{\mathbf{x}}^*) + \ln(2\pi \sigma_{\epsilon_{\mathbf{x}}^*}^2). \end{aligned}$$

(SNR) for template  $\mathbf{n}^{(i)}$ ,

$$\text{SNR}_i(\mathbf{x}) = \frac{\epsilon_{\mathbf{x}}^*}{\sigma_{\epsilon_{\mathbf{x}}^*}} \stackrel{\text{B15, B16}}{=} \frac{\Delta \mathbf{x}^T \Sigma^{-1} \mathbf{n}^{(i)}}{\sqrt{(\mathbf{n}^{(i)})^T \Sigma^{-1} \mathbf{n}^{(i)}}}, \quad (\text{B19})$$

and the constant

$$C = 2 \ln p(\epsilon_{\mathbf{x}}^*) + \ln(2\pi\sigma_{\epsilon_{\mathbf{x}}^*}^2). \quad (\text{B20})$$

To sum up, in the case of a general Gaussian likelihood function and general distortion components  $\mathbf{n}^{(i)}$ , the individual test statistics  $t_i(\mathbf{x})$  measure the strength of the evidence for the presence of the distortion  $\mathbf{n}^{(i)}$  in the data (Equation B18). This is directly analogous to the matched filtering technique used in signal detection, where the data is correlated with a set of template signals to find the one that maximizes the SNR.

Let us now restrict to the case where the distortions correspond to deviations along the standard basis vectors in data space. Specifically, we set the distortion directions to be the unit vectors in the  $i$ th dimension of the data space, i.e.  $\mathbf{n}^{(i)} \equiv \mathbf{e}_i$ . We find that

$$\text{SNR}_i^2(\mathbf{x}) = [\Delta \mathbf{x}^T \Sigma^{-1} \Delta \mathbf{x}]_i \stackrel{\text{B17}}{=} [(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})]_i. \quad (\text{B21})$$

Thus, considering all the possible alternatives of the standard basis and summing over them

$$t_{\text{sum}}(\mathbf{x}) = [(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})] + \text{const.} = \chi^2 + \text{const.} \quad (\text{B22})$$

where we have recovered the classical Pearson's  $\chi^2$  test up to a constant. We discuss in the next section when the conditions under which this constant is effectively independent of  $\mathbf{x}$ .

### B.1. Choice of prior boundaries

In our derivations, we have assumed a uniform prior for the distortion amplitude  $\epsilon$ , specifically  $p(\epsilon) \equiv \mathcal{U}(-b, b)$ . The choice of the boundaries  $b$ , hence of  $p(\epsilon)$ , affects the constant term  $C$  in the test statistic (see Equation B20).

One way to make this quantitative is to define a maximum SNR for a given distortion,  $\text{SNR}_{\text{max}}$ . This is realized by the maximum value of  $\epsilon$ , i.e.  $b$ . Writing this in terms of the variance  $\sigma_{\epsilon_{\mathbf{x}}^*}$ , we see that:

$$b = \text{SNR}_{\text{max}} \sigma_{\epsilon_{\mathbf{x}}^*} = \frac{\text{SNR}_{\text{max}}}{\sqrt{(\mathbf{n}^{(i)})^T \Sigma^{-1} \mathbf{n}^{(i)}}}. \quad (\text{B23})$$

Hence, as long as the prior boundaries  $b$  are chosen to be sufficiently wide, the constant  $C$  in the test statistic becomes effectively independent of  $\mathbf{x}$ .

Using the SNR training strategy, described in the following Appendix C, the prior boundaries  $b$  can be adaptively learned by the algorithm, as explained in Section III.1 and shown in Figure 4.

## Appendix C: SNR-based training strategy

We have seen in Section II.4 and in Appendix B that the test statistic for a given distortion  $\mathbf{n}^{(i)}$  is directly related to the SNR of that distortion in the data (Equation B18), where the SNR is given in Equation B19. Thus, these test statistics can be equivalently trained by minimizing a Gaussian negative log-likelihood loss for the MLE of the matched filter  $\epsilon_{\phi, i}(\mathbf{x})$  and its variance  $\sigma_{\phi, i}^2$  given a distortion  $i$

$$\begin{aligned} \mathcal{L}_{\text{SNR}}^{(i)}[\epsilon_{i, \phi}(\mathbf{x}), \sigma_{i, \phi}^2] = \\ \mathbb{E}_{\mathbf{x}, \epsilon \sim p_{\text{dist}}(\mathbf{x}|i, \epsilon)p(\epsilon)} \left[ \frac{(\epsilon_{i, \phi}(\mathbf{x}) - \epsilon)^2}{\sigma_{i, \phi}^2} + \ln \sigma_{i, \phi}^2 \right]. \end{aligned} \quad (\text{C1})$$

It is then straightforward to compute the SNR and hence the test statistics of interest from Equation B18 having the estimates  $\epsilon_{i, \phi}(\mathbf{x})$  and  $\sigma_{i, \phi}^2$

$$\hat{t}_i(\mathbf{x}) \propto \frac{\epsilon_{i, \phi}(\mathbf{x})}{\sigma_{i, \phi}}. \quad (\text{C2})$$

As discussed in Appendix A for the classifier-based training strategy, in the interest of efficiency, rather than training individual networks  $\epsilon_{i, \phi}$  and  $\sigma_{i, \phi}^2$  for each possible localized distortion  $i$ , we typically train single multi-output networks of the form

$$\epsilon_{\phi}(\mathbf{x}) : \mathcal{D} \rightarrow \mathbb{R}^{N_{\text{alt}}}, \quad (\text{C3})$$

$$\sigma_{\phi}^2 : \mathcal{D} \rightarrow \mathbb{R}^{N_{\text{alt}}}. \quad (\text{C4})$$

The total loss is given as a sum over the individual losses,

$$\mathcal{L}_{\text{SNR}}[\epsilon_{\phi}(\mathbf{x}), \sigma_{\phi}^2] = \sum_{i=1}^{N_{\text{alt}}} \mathcal{L}_{\text{SNR}}^{(i)}[\epsilon_{i, \phi}(\mathbf{x}), \sigma_{i, \phi}^2], \quad (\text{C5})$$

where the sum runs over all hypotheses  $H_i$ .

## Appendix D: Comparison of training strategies and analytical check

In this section we verify experimentally the connection to classical testing frameworks (Appendix B) of our method, and compare the two proposed training strategies (Appendix A and Appendix C). Importantly, we derive the analytical quantities for *profiled* likelihoods, while our estimates marginalize over background model variations.

For a first comparison, we further simplify the instructive example presented in Section III by considering only *uncorrelated bin-wise* additive stochastic distortions. In this simple scenario it is straightforward to compute the analytical (profiled) expectation. We show the comparison in Figure 6. There, we compare the analytical (profiled) expectations (dotted black lines) with the results obtained using our two training strategies — the classifier-based method from Appendix A (pink lines)

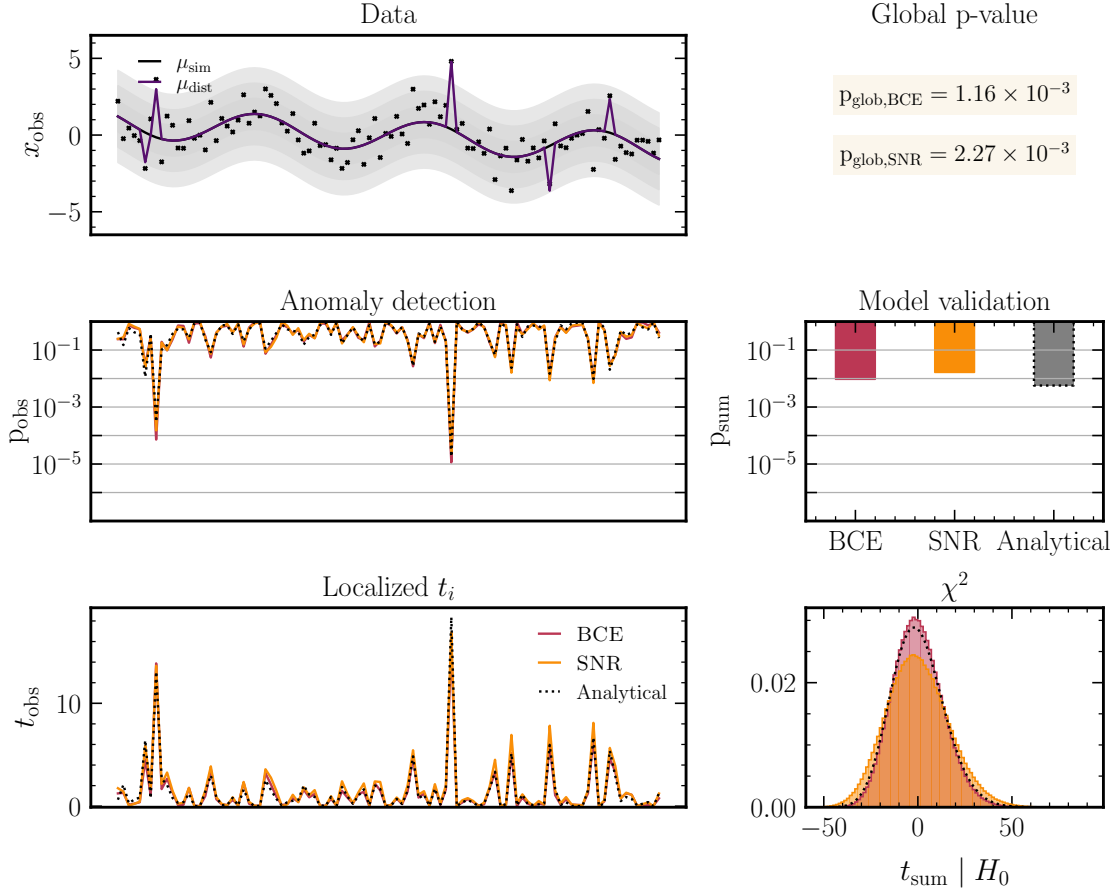


FIG. 6. Similar to Figures 2 and 3, but for uncorrelated bin-wise distortions. We show in dotted black the analytical (profiled) expectation, in pink the results obtained using the classifier-based (BCE) training strategy presented in Appendix A, and in orange the results obtained using the SNR-based (SNR) training strategy presented in Appendix C. The **upper-left, central-left, and central-right panels** are the same as in Figures 2 and 3. In the **lower-left panel** we show the localized test statistics estimated via Equation A3 and Equation C2 for the pink and orange lines respectively. Finally, in the **lower-right panel** we show the distribution of the aggregated test statistic  $t_{\text{sum}}$  (Equation 5) under the null hypothesis (i.e. applying the network to simulations  $\mathbf{x} \sim p_{\text{sim}}(\mathbf{x})$ ). As discussed in Section II and in Appendix B, for simple deviations along the standard basis vector in data space, the aggregated test statistics follows the classical Pearson’s  $\chi^2$  test.

and the SNR-based method from Appendix C (orange lines) — in the context of uncorrelated bin-wise additive stochastic distortions. The lower-left panel shows the localized test statistics estimated via Equations A3 (pink lines) and C2 (orange lines), where we see that both neural estimators closely match the analytical expectations. As a consequence, also the localized (central-left panel) and aggregated (central-right panel) significances match. Finally, the lower-right panel presents the distribution of the aggregated test statistic  $t_{\text{sum}}$  under the null hypothesis ( $\mathbf{x} \sim p_{\text{sim}}(\mathbf{x})$ ), confirming that it follows the classical Pearson’s  $\chi^2$  distribution as discussed in Section II and detailed in Appendix B.

We then consider the same setup as in the instructive example presented in Section III, with three correlated distortions of different correlation scales. We show the

comparison in Figure 7. There, we compare the analytical (profiled) expectations (dotted black lines) with the results obtained using our two training strategies. The second row shows the localized test statistics estimated via Equations A3 (dashed lines) and C2 (solid lines). In the last two rows, we can see that the two training strategies broadly agree, and in addition that the more the MLE prediction (shown in the first row) absorbs the distortion, the less the neural network-based and the analytical (profiled) prediction for the test statistic agree. We have checked that the analytical predictions and the neural network-based estimates do agree in the absence of a varying background. Thus we conclude that the mismatch is solely due to the MLE compensating the distortion, in case of the analytical profiled test statistic. Note that the neural network-based marginalized test statistics are, by construction, able to pick up the distortion accounting for the model variations.

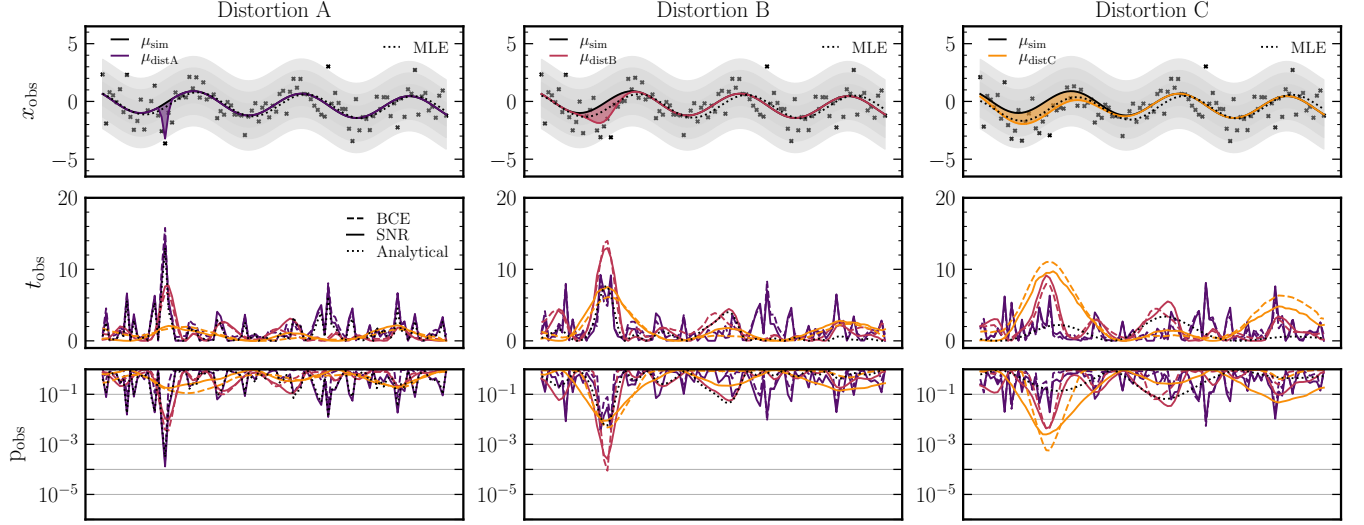


FIG. 7. A comparison of the training strategies and analytical expectation for correlated distortions. Each column is for data distorted by a deviation with different correlation scale, as labeled and color-coded in Figure 2. The **first row** shows the data distorted by the distortion, as in the first panel of Figure 2. We show with a dotted line the MLE prediction. The **second row** shows the test statistic estimated with the classifier-based training strategy through Equation A3 (dashed lines), with the SNR-based training strategy through Equation C2 (solid lines) or analytically (dotted black line). The **last row** shows the corresponding significance in terms of p-values. In the last two rows, we can see that the more the MLE prediction absorbs the distortion, the less the neural network-based and the analytical prediction for the test statistic agree. We note in this case that this is the expected behaviour and there is no reason that the two should agree when the MLE is significantly shifted.