# Learning dictionaries of New Physics with sparse local kernels

**Gaia Grosso**[1,2,3]*
gaia.grosso@cern.ch

**Katya Govorkova**[1]
ekaterina.govorkova@cern.ch

**Philip Harris**[1,2]
pcharris@mit.edu

**Eric Moreno**[1,2]
emoreno@mit.edu

**Ryan Raikman**[1]
rraikman@mit.edu

[1]MIT Laboratory for Nuclear Science, Cambridge, MA
[2]NSF AI Institute for Artificial Intelligence and Fundamental Interactions
[3]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA

## Abstract

Statistical anomaly detection empowered by AI is a subject of growing interest in high-energy physics and astrophysics. The unsupervised nature of the anomaly detection task combined with the highly complex nature of the LHC and astrophysical data give rise to a set of yet unaddressed challenges for AI. A particular challenge is the design of AI model architectures that are highly expressive, interpretable and incorporates physics knowledge. Under the assumption that the anomalous effects are mild perturbations of the nominal data distribution, *sparse* models represent an ideal family of functionals to learn interpretable models of the anomalies. In this work we propose a sparse model based on Gaussian kernels to construct a local representation of an anomaly score in semi supervised problems. Inspired by dictionary learning techniques we optimise the kernels' location over the input data, triggering a competition mechanism that induces the model's attention towards anomaly-enriched regions. We demonstrate the effectiveness using one-dimensional proof-of-concept numerical experiments and an application to gravitational wave anomaly detection.

## 1 Introduction

Novelty detection refers to the problem of recognising patterns in data that do not conform with any previously defined model of data generation. Within physics, generative models are often at the source of understanding experimental observations and discoveries. These models are theorised by a rigorous set of laws computing the probability of the observed data. The existence of unexpected features, outside predictions, within the data would imply either the generative model is incorrect, or there is a new phenomena that needs to be added to explain the data. Determining the statistical evidence for deviations of the data from the theoretical expectations (e.g. a “reference" model) is therefore crucial to achieving new scientific discoveries.

In absence of a signal model, an approach can be constructed that utilizes machine learning to train a binary classifier $f_{\mathbf{w}}$ to discriminate the data $\mathcal{D}$ from a reference distributed sample $\mathcal{R}$, and thus retrieve an estimate of the log-density-ratio between the generative models originating the two samples $f_{\mathbf{w}}(x) \approx \log\left[\frac{n(x|\mathcal{D})}{n(x|\mathcal{R})}\right]$. The log-density-ratio can then be used to design suitable two sample

[*] Corresponding Author

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

tests, such as the Neyman-Pearson test [1]. While this solution offers a valid framework to overcome the lack of theory assumptions for the data model, it still suffers from the expressivity bias induced by the choice of machine learning methods used in the binary classification task, a very relevant matter when the two datasets are almost identical, and there is a need to extract an interpretable understanding of detected features. A good trade-off between expressivity and interpretability can be found by imposing meaningful forms of regularisation, which induce physics motivated assumptions about the nature of the anomalies in the test design. In this work we propose a candidate family of functionals that embodies simple but powerful assumptions on the nature of the new physical processes arising in fundamental physics data. We build a family of functionals for anomaly detection as a *sparse* linear combination of Gaussian kernels with learnable coefficients and locations, that naturally fits the assumption of rare perturbations of the nominal data distribution.

## 2 Model design

New physical processes are expected to be rare, both in time and in space, producing a mild perturbation on top of the known physics models. The modification induced by the perturbation is unknown; it can be a localised over-density pattern or broad distortion. The function $f_{\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}$ modelling the log-density-ratio of the $d$-dimensional data input space should therefore return a null value for nearly all of the data, with the exception of the few regions where the anomalous signal events are present. It is then natural to think of *sparsity* as a meaningful form of regularisation to isolate the most anomalous region. At the same time, the lack of information about the location of the anomaly requires learning dynamics to be able to efficiently explore the phase space.

To address these points we propose a sparse linear combination of kernels $k_\theta$ (SparKer)

$$f_{\boldsymbol{a},\boldsymbol{\theta}}(x) = \sum_{i=1}^{\mathrm{M}} a_i \, k^i_{\boldsymbol{\theta}}(x),$$

with $a_i \in \mathbb{R}\ \forall i$ and sparsity enforced by choosing $\mathrm{M}$ to be much smaller than the number of data points. The $i$-th kernel is defined as a $d$-dimensional Gaussian distribution $k^i_{\mu_i,\sigma}$, with kernel's location $\mu_i = (\mu^1_i, \dots, \mu^d_i)$ and diagonal covariance matrix $\Sigma = \sigma \mathbb{I}_{d\times d}$.
Thus, the resulting model is characterised by three sets of parameters: the mixture coefficients $\boldsymbol{a} = [a_1, \dots, a_{\mathrm{M}}]$; the kernels' locations $\mu = [\mu_1, \dots, \mu_{\mathrm{M}}]$, and the kernels' width $\sigma$.

**Training task.** We solve the binary classification task by minimising the Neyman-Pearson (NP) loss function introduced in [1] (NPLM algorithm): $L_{\mathrm{NP}}[f_{\boldsymbol{a},\boldsymbol{\theta}}] = w \sum\limits_{x\in\mathcal{R}} (e^{f_{\boldsymbol{a},\boldsymbol{\theta}}} - 1) - \sum\limits_{x\in\mathcal{D}} f_{\boldsymbol{a},\boldsymbol{\theta}}$.
The weight parameter $w = \frac{\mathrm{N}(\mathcal{D}|\mathrm{H}_0)}{\mathrm{N}_{\mathcal{R}}}$ is used to handle unbalanced datasets, while preserving sensitivity to normalisation effect, as thoroughly explained in [1].

To solve this problem, we alternate a sparse coding step, in which the $\boldsymbol{a}$ coefficients are fitted while the kernel locations $\mu$ are fixed, and a dictionary learning stage, in which the $\boldsymbol{a}$ coefficients are fixed and the locations $\boldsymbol{\mu}$ are trained. We keep $\sigma$ constant and regard it as a hyperparameter of the model.

When solving the sparse coding problem, we add a L2 regularization term on the coefficients $\boldsymbol{a}$ that makes the problem convex and enforces the solution to be smooth. The final loss function thus assumes the following form $L = L_{\mathrm{NP}} + \lambda_{\boldsymbol{a}} L_{\boldsymbol{a}}$, with the L2 regularisation term defined as $L_{\boldsymbol{a}} = \frac{1}{\mathrm{M}} \sum_{i=1}^{\mathrm{M}} a_i^2$, and $\lambda_{\boldsymbol{a}}$ being the coefficient regulating the relative contribution of the two terms.

During the dictionary learning step, we introduce a regularisation term based on a measure of the entropy of the kernels' locations in the data support. The entropy term biases the learning task towards solutions uniformly spacing the kernel locations over the space, as opposed to a localized aggregation of kernels around the same anomalous points. The loss function for the dictionary learning step assumes the following form $L = L_{\mathrm{NP}} - \lambda_H L_H$, where $L_H$ is the entropy loss defined as $L_H = -\sum_{j=1}^{\mathrm{M}} p_\mu(\mu_j) \log p_\mu(\mu_j)$, with $\lambda_H$ a coefficient regulating the relative contribution of the two terms. We choose to approximate the density distribution of the kernels as: $p_\mu(x) = \frac{1}{\mathrm{M}} \sum_{i=1}^{\mathrm{M}} k^i_{\boldsymbol{\theta}}(x)$.

The regularisation coefficients $\lambda_a$ and $\lambda_H$ constitute two additional hyperparameters of the machine learning task. It should be noted that the assumption of signal model independence forbids the optimisation of the regularisation coefficients on a specific hypothetical signal pattern. The tuning of both $\lambda_a$ and $\lambda_H$ must therefore follow considerations that are only based on their impact on the training task in presence of signal-free input data. Following the heuristic approach already used in Ref. [4], we base the choice of the regularisation coefficients on their effect on the statistical behaviour of the NP test statistic distribution under the null hypothesis. Configurations that return an empirical distribution compatible with a $\chi^2$ are considered good candidates (see Appendix B). In general, small values of the coefficients are preferable to keep the training stable and the regularisation bias small.

**Learning dynamics.** To solve the binary classification task we update the model parameters following the gradient of the loss. In absence of regularisations, the gradient of the NP loss, for a generic parameter $\theta$, is $\nabla_\theta L_{\rm NP}[f] = w \sum\limits_{x\in\mathcal{R}} e^{f(x)}\nabla_\theta f(x) - \sum\limits_{x\in\mathcal{D}} \nabla_\theta f(x)$. The loss gradient directly depends on the model's gradient, having an opposite sign when computed over the two samples.

The gradient of the SparKer model with respect to the $j$-the coefficient is given by $\nabla_{a_j} f(x) = k_j(x)$. This returns a kernel centered around the location of the $j$-th kernel; the more localized around $j$, the stronger the gradient, thus favouring local reconstruction.

The gradient of SparKer with respect to the $j$-th kernel location is $\nabla_{\mu_j} f(x) = a_j k_j(x)\frac{(x-\mu_j)}{\sigma^2}$. For positive $a_j$ the the $j$-th kernel is attracted by data points $x \in \mathcal{D}$ and repelled by reference points $x \in \mathcal{R}$, hence emphasizing over-densities as regions of interest. Conversely, for negative $a_j$ the kernels are repelled by elements of $\mathcal{D}$ and attracted by elements of $\mathcal{R}$, thus finding under-densities within the learning process. More generally, the loss presents a competition of the two training samples to attract kernel's attention to the most critical over- and under-densities. However, the dynamics is only active locally, as the gradient strength depends on the coupling of the kernel through euclidean a distance.

**Neyman-Pearson test.** As a final figure of merit, we adopt the Neyman-Pearson test as defined in [1]: $t(\mathcal{D}) = 2\max\limits_{\boldsymbol{\theta}} \sum\limits_{x\in\mathcal{D}} \log\frac{\mathcal{L}(\mathcal{D}|{\rm H}_\theta)}{\mathcal{L}(\mathcal{D}|{\rm H}_0)}$. In our setup, it can be straightforward computed from the value of the NP loss at the end of the training: $t(\mathcal{D}) = -2\min_{\boldsymbol{\theta}} L_{\rm NP}(\boldsymbol{\theta})$.

The NP test has been implemented using a mixture of static Gaussian kernels in [2]. In the numerical experiments reported in the following, we will refer to the latter as Falkon-NPLM model, whereas we will refer to the application of SparKer to the computation of the NP test as Sparker-NPLM model.

## 3 One-dimensional toy experiments

**Dataset.** The first benchmark we consider was introduced in Ref. [1]. It comprises a simple univariate setup, where the density distribution under the reference model is an exponentially falling distribution mimicking typical energy or momentum spectra of LHC data. To test the model ability to detect features of various narrowness and locations, we consider Gaussian signals with variable location $x_S$ and width $\sigma_S$, and a signal manifesting as an excess in the tail, as defined in [1]. We set the total number of expected experimental observations in absence of signal at ${\rm N}({\rm H}_0) = 2000$ events and the reference sample $\mathcal{R}$ is taken 100 times larger, ${\rm N}_{\mathcal{R}} = 200\,000$. The details of the signal injected are given in Table 1 of Appendix A.

**Model setup.** We consider three models with ${\rm M} = 10$, 50 and 1000 respectively. The L2 regularisation on the coefficients, $\lambda_a$, is set a $10^{-5}$ for the first two models and at $10^{-6}$ for the third one. The entropy regularisation coefficient, $\lambda_H$, is set at 0.001 for the first two models and at 0.01 for the third. The model is trained for 500 epochs, and each epoch alternates 100 sparse coding steps and 100 dictionary learning steps. We use Adam optimiser with learning rate $10^{-3}$. The width of the Gaussian kernels, $\sigma$, is set to the 90% quantile of the pair-wise euclidean distance between reference distributed data points. This choice has been made to align with the heuristic proposed in [2] for the Falkon-NPLM model, and thus facilitate the comparison. Other values of $\sigma$ are equally viable, and further optimizations are possible by aggregating the p-values for different $\sigma$ choices [3]. To highlight the gain introduced by allowing the kernels' locations to be learnable, we compare our study to Falkon-NPLM. For a fair comparison, we choose for Falkon-NPLM the same values of M and $\lambda_{\boldsymbol{a}}$ as in SparKer. The code to reproduce the experiments is available on GitHubGitHub . For a single training of the model, we used a GPU with 5 GB memory. The training time significantly depends on M. For our choices of M, the run of a single experiment takes between few minutes and one hour.

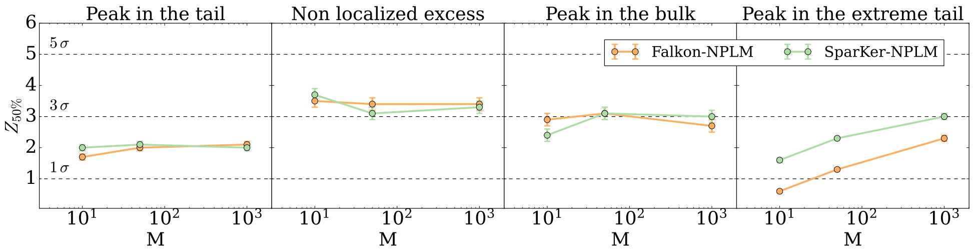


Figure 1: Sensitivity reach of the SparKer (green line) and the state-of-the-art (orange line) models with increasingly large number of kernels (M). We report the median $Z$-score calculated over 1000 experiments as a function of the model number of kernels M.

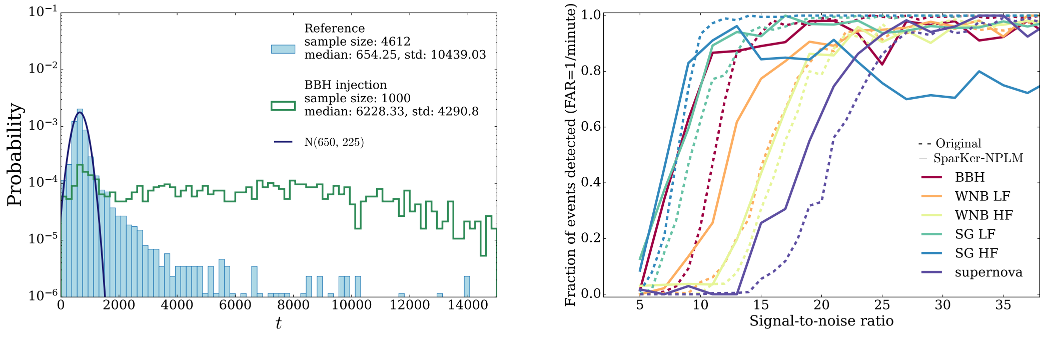
**Impact of sparsity and dictionary learning on detection performances.** We study the impact of an increasingly aggressive sparsity constraint on the anomaly detection performances of SparKer-NPLM, and we compare it with Falkon-NPLM. Results are summarised in Figure 1. For signal benchmarks that are visible in high density regions of the data, SparKer-NPLM maintains similar performances to Falkon-NPLM at all values of $\mathrm{M}$. For signal patterns that are progressively further from the bulk of the data distribution SparKer tends to outperform Falkon-NPLM. More importantly, we observe a smaller degradation with reduced number of kernels when compared with Falkon-NPL model. We conclude that the degrees of freedom are better directed toward the region of the input space where the anomaly is located.

## 4 Gravitational waves agnostic detection

We apply SparKer to the detection of anomalous gravitational wave (GW) sources. The data from the LIGO detector [5], available at [7], are time series which represent the strain caused by passing gravitational waves. In the absence of the GW signal, the strain consists of the noise coming from various sources.

**Dataset.** In this paper, we start from the anomaly detection method introduced in [6]. The semi-supervised approach GWAK (Gravitational Wave Anomalous Knowledge) builds latent space from the reconstruction loss of five autoencoders applied to each of the two LIGO detectors, yielding 10 dimensions. Each of the autoencoders is trained on a specific signal or background type: the main background is represented as white noise and taken from real data, with loud glitches removed; large transient backgrounds, known as glitches, are identified and trains separately; signals are represented by binary black holes (BBHs) and a low and high frequency generic sine-gaussian signal shape. Two additional axes consisting of the frequency correlation and Pearson correlation between the two detectors are added, yielding 12-dimensions. The 12 values are combined to determine an anomaly search region. In previous results, a simple linear combination was used [6]; this is not the most efficient way to select anomaly-enriched regions since it assumes linearity. Here, we use the "GWAK" values as input features for SparKer-NPLM to determine a more rigorous anomaly metric.

**Model setup.** We set $\mathrm{M} = 20$, $\lambda_a = 10^{-2}$ and $\lambda_H = 10^{-1}$. We train for 2000 epochs, alternating 10 sparse coding and 10 dictionary learning iterations. We use Adam optimiser with a learning rate $4 \cdot 10^{-2}$ (the code is available on `GitHub`). We consider six signal benchmarks: binary black hole (BBH), low frequency (LF) and high frequency (HF) generic sine-gaussian (SG) signal shape, low and high frequency white noise burst (WNB) and supernova. Three of them (BBH, SG HF and SG LF) are included in the training of GWAK autoencoders; whereas the remaining three are previously unseen. We use as data sample a collection of 2402 sequential observations of the GWAK algorithm as defined in [6], on data with signal injection from one of the considered signal benchmarks, or without any signal to calibrate the test. To build the reference sample we use $\mathrm{N}_{\mathcal{R}} = 24020$ GWAK observations on background only events, corresponding to 10 times more events than what constitutes a data sample[2]. Since signals are relatively sparse, a background reference dataset can be obtained from the actual datasets by simply shifting the time of the two datasets such that the correlation present when signals appear is broken.



(a)

(b)

Figure 2: (a) SparKer-NPLM test statistic distribution for background streams (light blue) and streams injected with BBH signal (green). (b) Trigger efficiency after glitches vetoing at false alarm rate of 1 per minute for SparKer-NPLM (solid line) and original GWAK (dashed line) for various signals.

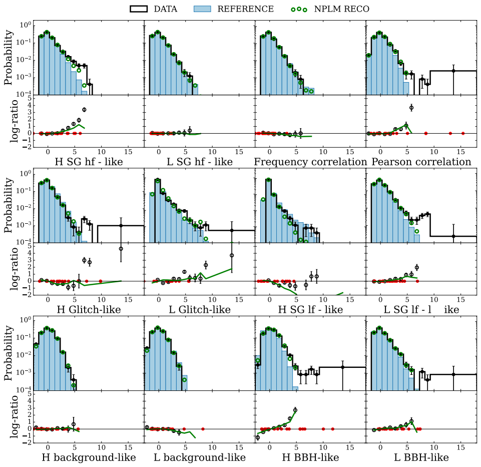


Figure 3: Top panels: the distributions in the absence of signals (light blue histogram) and in the presence of a binary black hole signal (black histogram) over the twelve input features marginals. Bottom panels: learned SparKer model (green line) and kernels' location (red dots).

**Results.** Figure 2a shows the distribution of the SparKer-NPLM test statistic in the absence of signal (light blue histogram) and with BBH signal injection (green histogram). The bulk of the distribution in the null hypothesis follows a normal distribution. Outliers highlight the existence of anomalous events due to glitches, which are noise transients present in data. To overcome this problem, in the original GWAK implementation glitches are vetoed using a heuristic over a larger time window where glitch identification becomes easier. The implementation of the veto in the NPLM-SparKer is left to future work, for now we will neglect these outliers and consider the asymptotic gaussian distribution as a good approximation of the test statistic after glitches removal. Figure 2b compares the trigger efficiency of SparKer-NPLM (solid lines) and the original GWAK implementation (dashed lines) for a False alarm rate (FAR) of 1 event per minute as a function of the signal-to-noise ratio. Different colours represent different signal benchmarks. Sparker-NPLM shows improved trigger efficiency over the original implementation.

Moreover, SparKer-NPLM is able to identify signal-enriched tails in the sample, providing useful insight for the interpretation of the anomaly. An example of BBH signal identification and reconstruction with SparKer is given in Figure 3, and additional plots are reported in Appendix C.

## 5 Conclusions, limitations and outlook

We have shown successful applications of SparKer on a 1-dimensional toy model, and a 12-dimensional gravitational wave search, showing the potential for SparKer to act as an effective anomaly detection algorithm, while producing interpretable results. Future extensions of this work involve incorporating a way to handle systematic failures like glitches in the model following [8], along with execution time and model design optimisation.

[2] This guarantees reasonable modelling of the background expectation for the 12D input space.

## Acknowledgments

This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). The authors acknowledge support from the National Science Foundation with grant numbers OAC-2117997 (A3D3) and CSSI-1931469. Computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

## References

[1] D'Agnolo, R.T., Wulzer, A. (2018). Learning new physics from a machine. Physical Review D.

[2] Letizia, M., Losapio, G., Rando, M. et al. Learning new physics efficiently with nonparametric methods. Eur. Phys. J. C 82, 879 (2022).

[3] G. Grosso and M. Letizia, Multiple testing for signal-agnostic searches of new physics with machine learning. arXiv:2408.12296 [hep-ph].

[4] d'Agnolo, R.T., Grosso, G., Pierini, M., Wulzer, A. and Zanetti, M., 2021. Learning multivariate new physics. The European Physical Journal C, 81(1), pp.1-21.

[5] The LIGO Scientific Collaboration and J Aasi et al. Advanced ligo. Classical and Quantum Gravity, 32(7):074001, mar 2015.

[6] GWAK: gravitational-wave anomalous knowledge with recurrent autoencoders. Ryan Raikman et al 2024 Mach. Learn.: Sci. Technol. 5 025020

[7] R. Abbott et al. (LIGO Scientific Collaboration, Virgo Collaboration and KAGRA Collaboration), "Open data from the third observing run of LIGO, Virgo, KAGRA and GEO", ApJS 267 29 (2023)

[8] d'Agnolo, R.T., Grosso, G., Pierini, M., Wulzer, A. and Zanetti, M., 2021. Learning new physics from an imperfect machine. The European Physical Journal C, 82(3), pp. 275

## A One-dimensional toy model: dataset details

The first benchmark we consider was introduced in [1]. It comprises a simple univariate setup. The density distribution under the null hypothesis is defined as

$$n(x|\mathrm{H}_\mathbf{0}) = \mathrm{N}(\mathcal{D}|\mathrm{H}_\mathbf{0})\, e^{-x}, \tag{1}$$

where $\mathrm{N}(\mathrm{H}_\mathbf{0})$ denotes the total number of expected events in the dataset. To test the model ability to detect features of various narrowness and in different locations, we consider Gaussian signals of the form

$$n(x|\mathrm{G}_{\bar{x},\sigma_{\mathrm{NP}}}) = \mathrm{N}(\mathrm{S})\frac{1}{\sqrt{2\pi}\sigma_S}\ \exp\left[-\frac{(x-x_S)^2}{2\sigma_S^2}\right], \tag{2}$$

and a signal manifesting as an excess in the tail defined by the analytic model

$$n(x|\mathrm{E}) = \frac{\mathrm{N(S)}}{2}x^2e^{-x}\,. \tag{3}$$

The data distribution under the alternative hypotheses $\mathrm{H}_1$ has therefore the following form

$$n(x|\mathrm{H}_1) = \ n(x|\mathrm{H}_\mathbf{0}) + n(x|\mathrm{S})\,, \tag{4}$$

with $\mathrm{S} = \{\mathrm{G}_{\mu,\sigma}, \mathrm{E}\}$. We set the total number of expected experimental observations in absence of signal at $\mathrm{N}(\mathrm{H}_\mathbf{0}) = 2000$ events and the reference sample $\mathcal{R}$ is taken 100 times larger, $\mathrm{N}_\mathcal{R} = 200\,000$. The details of the signal injected are given in Table 1. A graphic representation of the four signal benchmarks is given in Figure 4.

| | | $x_S$ | $\sigma_S$ | $\mathrm{N(S)}/\mathrm{N}(\mathrm{H}_\mathbf{0})$ | $Z_{id}$ |
|---|---|---|---|---|---|
| **Gaussian peaks** | bulk | 1.6 | 0.16 | $4.5 \cdot 10^{-2}$ | 4.9 |
| | tail | 6.4 | 0.16 | $5 \cdot 10^{-3}$ | 4.0 |
| | extreme tail | 9 | 0.16 | $2.5 \cdot 10^{-3}$ | 12.5 |
| **Tail excess** | excess | | | $4.5 \cdot 0^{-2}$ | 4.5 |

Table 1: Summary of the 1D signal benchmarks. $Z_{id}$ represents the ideal Z-score computed as the Neyman-Pearson log-likelihood-ratio test with known true alternative hypothesis.

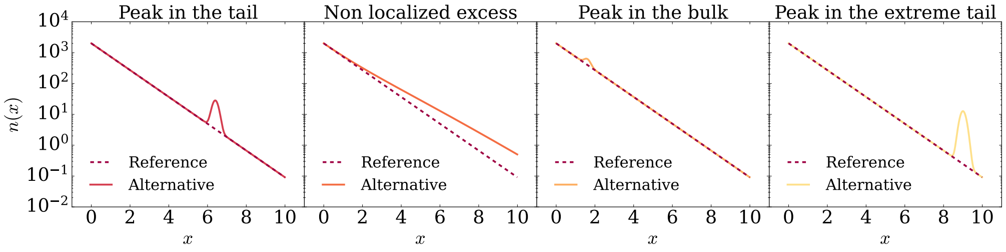


Figure 4: EXPO-1D signal benchmarks representation.

## B Statistical properties of the NP test statistic using the SparKer model.

In the contect of the numerical experiments presented in Section 3, we study the distribution of the NP test statistic in absence of signal injection and we observe a good compatibility with a $\chi^2$ with arbitrary number of degrees of freedom. The same properties were observed for the kernel methods used in [2], and guarantee a good behaviour of the NP test, allowing the extraction of a reliable approximated p-value from the asymptotic $\chi^2$ model. For completeness, we report in Figure 5 the empirical distribution of the NP test statistic for different values of M.

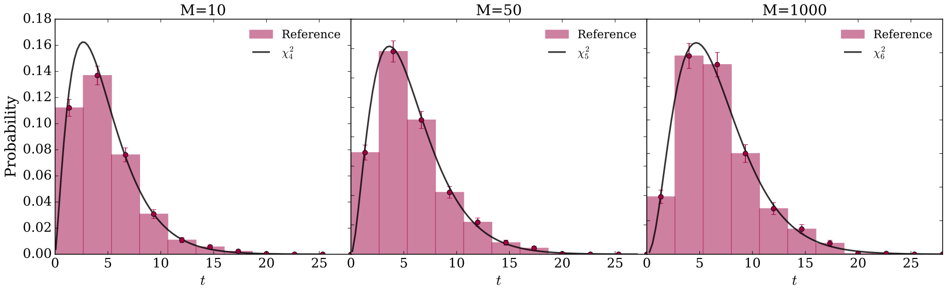


Figure 5: Test statistic distribution under the $\mathrm{H}_{\mathbf{0}}$ hypothesis for different values of M. The distributions are all well compatible with a $\chi^2$ distribution with number of degrees of freedom to be fitted.

## C Gravitational waves detection

In this appendix we report additional plots concerning the gravitational waves detection use case presented in Section 4. Figures 6, 7, 8, 9, 10, 11 show the distribution of the learned kernel's centroids $\mu_i$ over the marginals of the 12D GWAK space. The black historgrams represent the typical learned kernels' locations for signal-free toys, while the different coloured histograms across figures represent the distribution for toys with various signal injections. The locations of kernels in presence of signal injection are generally shifted towards the tails of some marginals. This is in line with the expectation that data stream containing signal events are scored higher by the GWAK features used as input. We highlight with a filled histogram the kernels' locations that are associated to the top 10% positively weighted kernels. Those generally fall in the tail of the distributions suggesting that the nature of the anomaly in the tail is an overdensity, as expected.

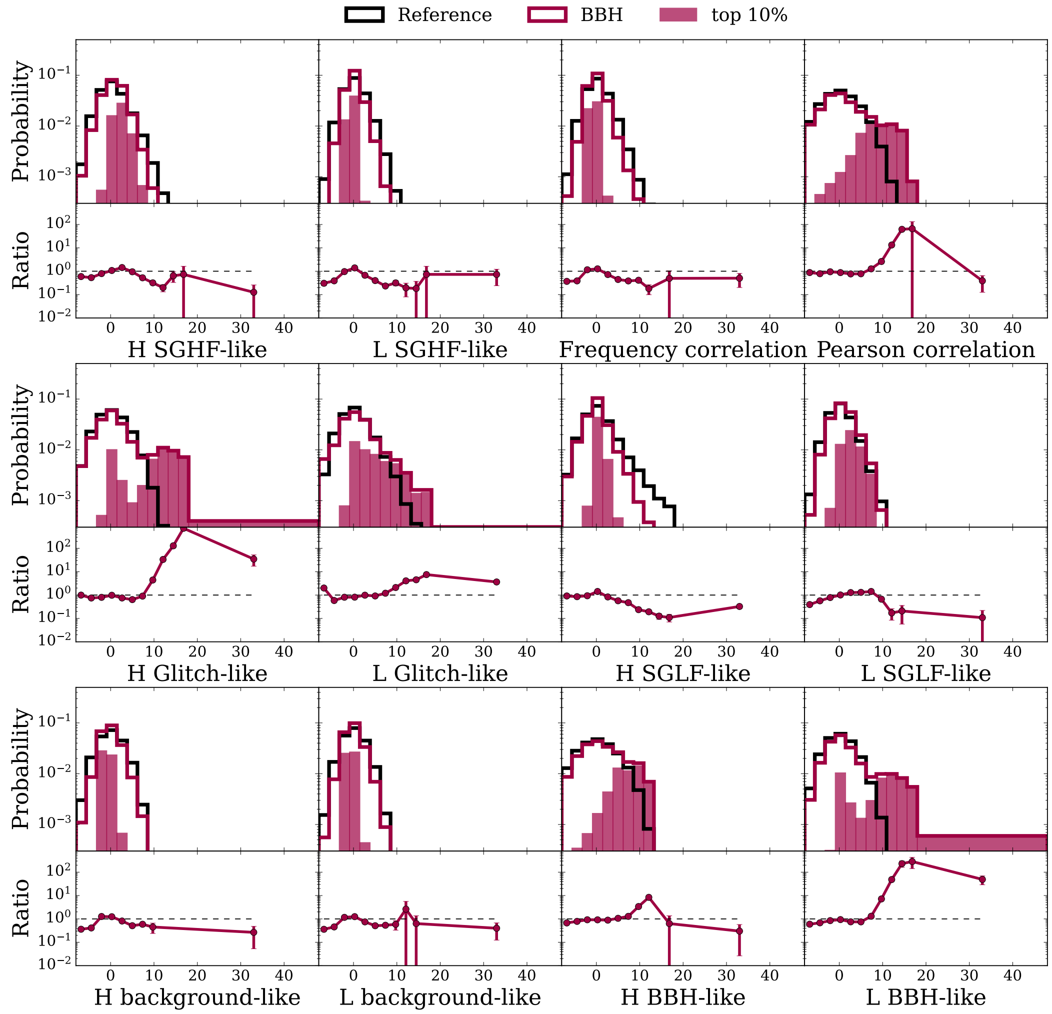


Figure 6: Learned kernels' location marginal distributions over the 12 input features. The distribution in the absence of signals (black) and in presence of BBH signals (dark red) are compared. The top 10% positively weighted kernels are reported in the filled histogram.

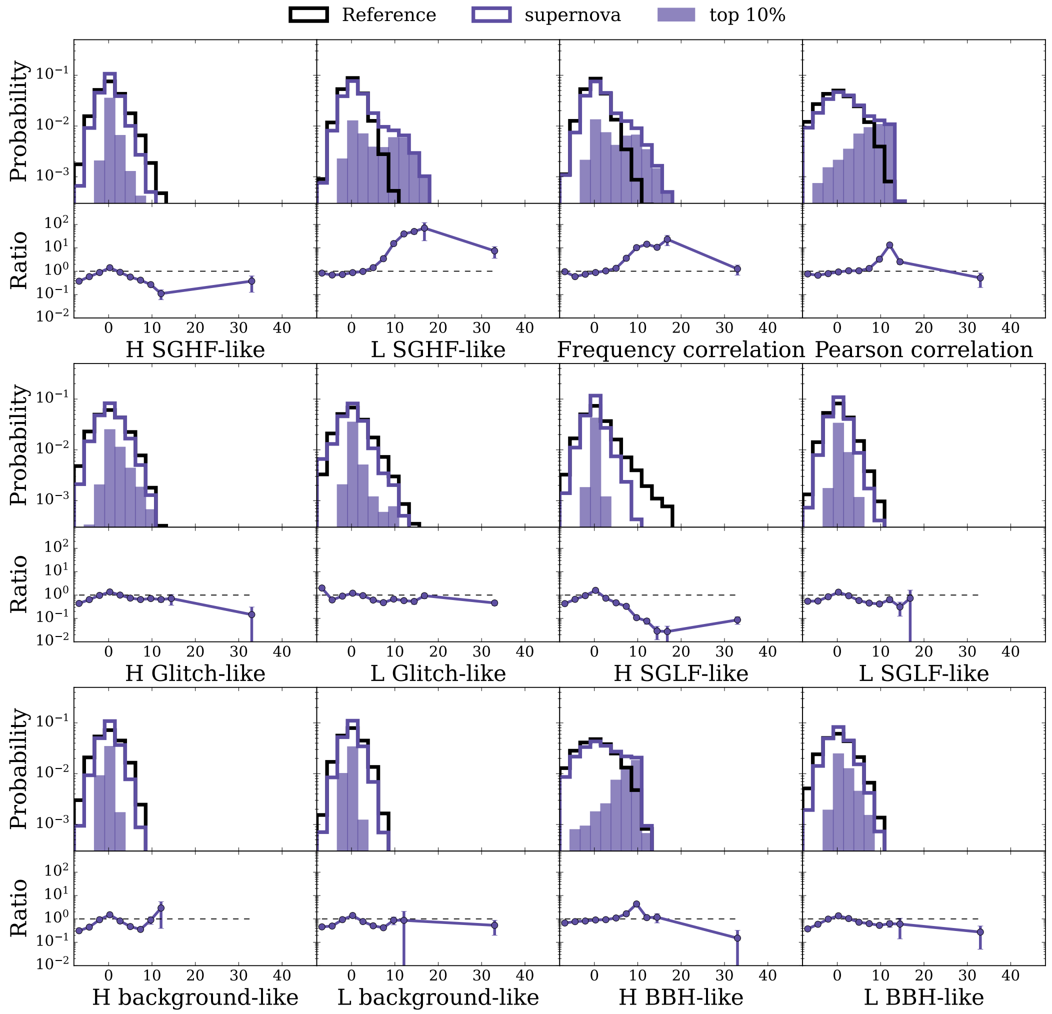


Figure 7: Learned kernels' location marginal distributions over the 12 input features. The distribution in the absence of signals (black) and in presence of supernova signals (dark blue) are compared. The top 10% positively weighted kernels are reported in the filled histogram.

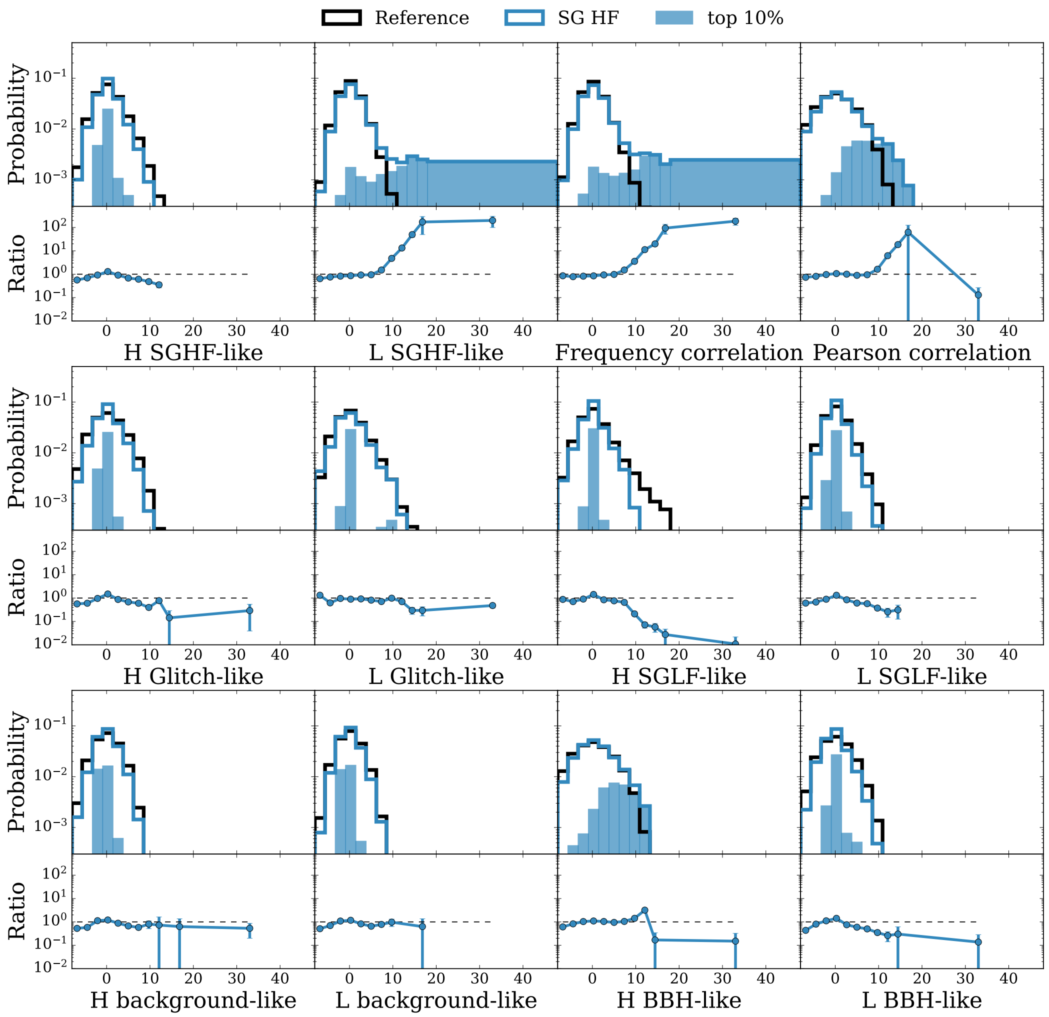


Figure 8: Learned kernels' location marginal distributions over the 12 input features. The distribution in the absence of signals (black) and in presence of high frequency sine Gaussian signals (blue) are compared. The top 10% positively weighted kernels are reported in the filled histogram.

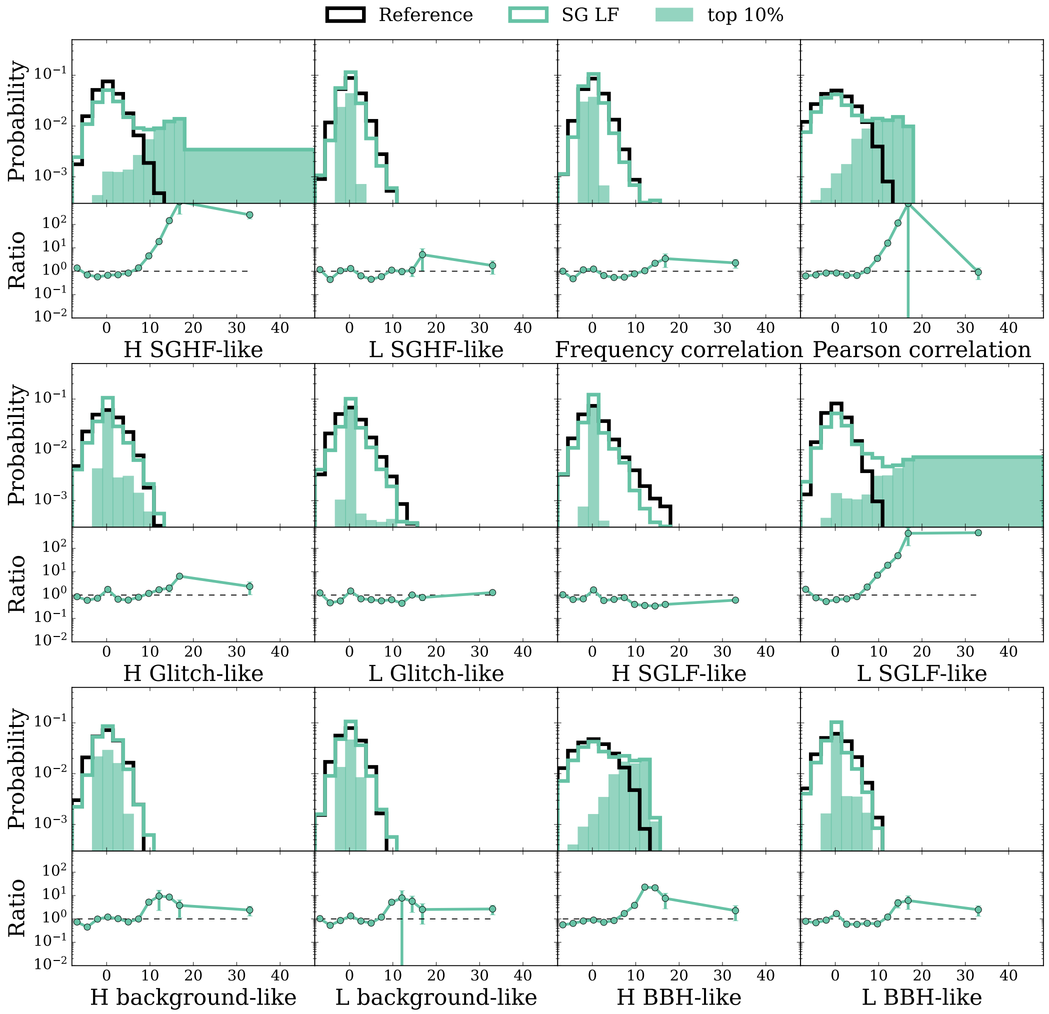


Figure 9: Learned kernels' location marginal distributions over the 12 input features. The distribution in the absence of signals (black) and in presence of low frequency sine Gaussian signals (seagreen) are compared. The top 10% positively weighted kernels are reported in the filled histogram.

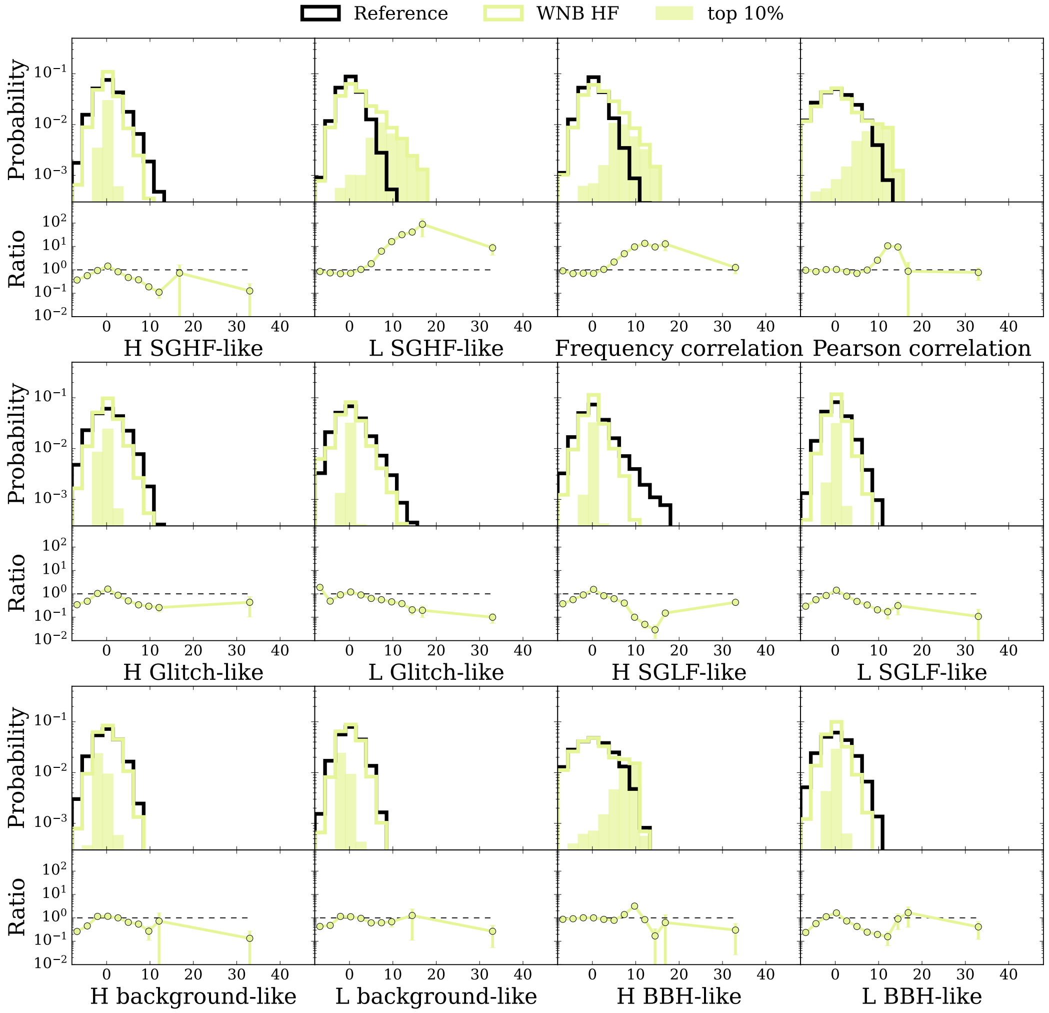


Figure 10: Learned kernels' location marginal distributions over the 12 input features. The distribution in the absence of signals (black) and in presence of high frequency white noise burst (lime green) are compared. The top 10% positively weighted kernels are reported in the filled histogram.

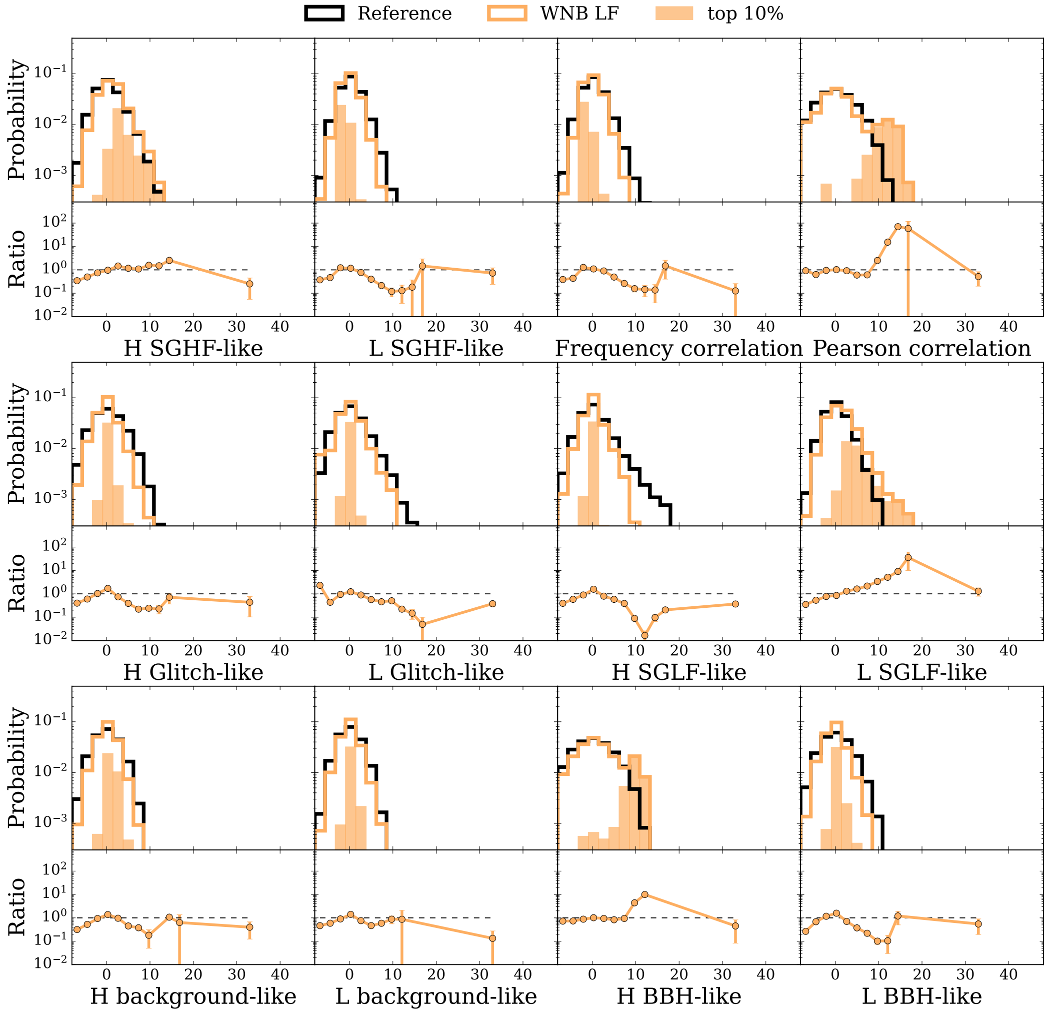


Figure 11: Learned kernels' location marginal distributions over the 12 input features. The distribution in the absence of signals (black) and in presence of low frequency white noise burst (orange) are compared. The top 10% positively weighted kernels are reported in the filled histogram.