An empirical Bayesian prior for indirect dark matter searches in uncertain backgrounds

Alex Geringer-Sameth*
Astrophysics Group, Physics Department, Blackett Lab, Imperial College London, Prince Consort Road, London SW7 2AZ, UK [RT :Add LLNL!]

Andre Scaffidi†
Theoretical and Scientific Data Science, Scuola Internazionale Superiore di Studi Avanzati (SISSA), via Bonomea 265, 34136 Trieste, Italy

Roberto Trotta‡
Theoretical and Scientific Data Science, Scuola Internazionale Superiore di Studi Avanzati (SISSA), via Bonomea 265, 34136 Trieste, Italy
Astrophysics Group, Physics Department, Blackett Lab, Imperial College London, Prince Consort Road, London SW7 2AZ, UK
INFN – National Institute for Nuclear Physics, Via Valerio 2, 34127 Trieste, Italy
Italian Research Center on High Performance Computing, Big Data and Quantum Computing

(Dated: February 17, 2025)

Dark matter searches in gamma-ray data aim to reject the null-hypothesis that the observed data are generated from a background-only model. However, test statistics used for the hypothesis test are inevitably based on an incomplete understanding of the background. Even if false alarm probabilities are correctly calibrated later, the sensitivity to the signal of interest suffers. We propose a new test that aims to restore sensitivity using an informative prior – which we call *the empirical prior* – derived from sampling background-only regions around the target location. Our frequentist test statistic is built from the Bayes factor between a background-only model and a source signal model, with the empirical prior providing the distribution of the background parameters. We demonstrate from simulations that our new test statistic has superior power (at fixed false detection rate) compared with the traditional likelihood ratio, which is the default statistic adopted in the literature. The empirical prior can be thought of as a correction that pushes an initial uninformed prior closer to the unknown "true" one. Finally, we forecast the median detection significance to an annihilating dark matter signal and demonstrate a factor of ~ 4 improvement over the standard likelihood-based approach. While we present this approach in the context of dark matter searches, the empirical prior and the methodology presented in this work are general and can be conducted on any real data for which a sample of background regions is available.

I. INTRODUCTION

The detection of astrophysical sources in an imprecisely known background is a general problem in astronomy. The most common application is the generation of source catalogs from observations over some region of the sky [e.g. 1, and citations therein]. In this case the statistical questions are at the population level, e.g., characterizing an algorithm's completeness, depth, and bias. This paper focuses on a second application, where source detection is targeted at a single astronomical object of interest. The detection or non-detection of that individual source is the crucial question and the statistical considerations have a different flavor. One aims to quantify the false positive detection probability as accurately as possible while also pushing the sensitivity (or statistical power) as high as possible. The uncertain nature of the background impacts both of these tasks.

Traditionally, targeted source detection has been tackled with variations of the so-called "on/off" method: a measurement of the background in the vicinity of the target of interest is made, which is then subtracted off the on-source measurement to obtain an estimate of the source's flux [e.g. 2–7]. Model-fitting based approaches are also common. Fits to an astronomical scene are performed both with and without a model component corresponding to the target of interest and the difference in fit quality is a measure of detection significance. It is also possible to marginalize over model parameters for a Bayesian approach [AGS: Roberto add a cite here].

In the context of gamma-ray astronomy, which is the primary motivator of this work, the majority of astrophysical source detection methods assume a simple background model in the vicinity of the target of interest, namely, that photons are generated by a Poisson process over an underlying smooth intensity composed of extended diffuse processes, plus a set of discrete, known sources [8–13]. The brightest point-like sources are easily detectable and can be included as known, parameterized components in the background model. Diffuse and isotropic processes can also be included as templates

* geringersame1@llnl.gov
† ascaffid@sissa.it
‡ rtrotta@sissa.it

with fixed morphology but with amplitude and spectrum controlled by free parameters, which are usually fitted over a larger region of the sky [14–21]. This work focuses on addressing the effect of *unknown* contributors to the background — those which cannot be identified and then parameterized and included as model components — and designing a test statistic that improves sensitivity for dark matter searches under such conditions.

The Fermi Large Area Telescope (Fermi-LAT) [22] has provided extensive data for analyzing gamma-ray emissions from both dwarf galaxies and the Galactic Center (GC), two important classes of dark matter targets. Recent studies, including an 11-year analysis of 27 dwarf spheroidal galaxies, as well as global analyses by multiple gamma-ray experiments (Fermi-LAT, HAWC, H.E.S.S., MAGIC and VERITAS) indicate no significant dark matter signals, although new stringent upper limits on dark matter annihilation cross-sections have been established [23–28]. In contrast, investigations of the GC gamma-ray excess have revealed a multitude of differing interpretations ranging from dark matter (DM) to astrophysical [29–32], or even mis-modelling [33].

In the context of dark matter searches for dwarf galaxy gamma-ray emission, a simple background model is generally not an adequate representation of the actual sky. This leads to both decreased sensitivity for a dark matter signal and to higher false positive rates for discovery of emission from the target region of interest. Several works have pointed out that false positive source detections may commonly arise due to small-scale mis-modeling of what are assumed to be spatially-diffuse background components [34–38]. Additionally, most conventional searches are implicitly optimized to distinguish a new faint source from a smooth diffuse background, not a background teeming with unknown faint interlopers. Neglecting these interlopers effectively reduces the possible discovery significance of a new source.

A variety of studies have aimed to address these limitations. The "likelihood stacking" method introduced by Paliya *et al.* [39] presents a method in which the likelihood is the product of many regions of interest (ROIs) each centered at the location of a blazar. Each ROI is parameterized in terms of its source and background, using the same set of source parameters for each. The background parameters may differ among the ROIs and are all profiled out, leaving a (profile) likelihood specified in terms of the source parameters only. This procedure is repeated by fitting different areas of the sky with no apparent sources, but doesn't account for the *distribution* of background parameters across background regions.

Song *et al.* [40] focus on pulsars, and follow a similar procedure with the addition of likelihood functions at locations of control field test sources. The likelihood-stacking method used in the sources involves summing the test statistic (TS) values of all the sources of interest and comparing that sum to the cumulative TS of a population of control field test sources to quantify the significance of the stack. This method is more sensitive than simple count stacks and allows for the investigation of the spectral properties of the pulsar population by performing a TS stack in spectral parameter space.

Calore *et al.* [41] present a purely data-driven approach to derive constraints on dark matter properties. Their procedure involves a distribution, $P(D)$, somewhat akin to our empirical prior introduced in Sec. II, which is informed by the observed photon counts from dwarf galaxies. This distribution is learned; it is crafted from the photon counts in adjacent pixels, allowing for a more nuanced analysis of the gamma-ray background. However, it is important to note that the model primarily focuses on the aggregation of photon counts across different energy bins for each dwarf galaxy, without explicitly incorporating the spatial distribution of counts within each galaxy's region.

The method of Refs. [42, 43] introduces a procedure that is similar to our proposal, albeit framed as a likelihood-based method. Every background location is independently modelled as a diffuse background plus a dark matter point source with fixed spectrum but floating amplitude. The profile likelihood is computed for each background region as a function of the dark matter amplitude alone. The resulting collection of profile likelihood functions are then averaged and profiled.

In this paper, we present a new methodology that improves on the traditional "on/off" method for source searches in the presence of diffuse photon backgrounds that are both mismodeled and contain additional unknown components. The key components of our inference framework are (i) the empirical learning of the diffuse emission in the vicinity of a ROI, which is the result of a superposition of multiple physical and instrumental processes, and (ii) the inclusion of such as an informative "empirical prior" in the calculation of a Bayes factor, which we use as a frequentist test statistic. By introducing an informative prior for the background in the proximity of the ROI, we obtain higher sensitivity to dark matter annihilation signals that are improbable under the prior. Then, by adopting the Bayes factor between the background-only model and a model including dark matter as a frequentist test statistic, we connect with the usual hypothesis rejection framework.

This paper is organised as follows: In Sec. II we introduce the methodology to derive an empirical prior that encapsulates knowledge of the unknown background region around a target of interest and encodes it as a predictive model to be used in a later hypothesis test in the target ROI. We also demonstrate in this section how our method is augmented by treating the Bayes factor as a frequentist test statistic. In Sec. III we compare the performance of the new test statistic based on the empirical prior against that of the standard likelihood-ratio test statistic, exploring sensitivity in the presence of ambiguity between DM annihilation versus emission from an astrophysical point source behind a dwarf galaxy. We then conclude and discuss future applications of the method in Sec. IV.

II. METHODOLOGY

A. Motivation

In any source detection framework, frequentist or Bayesian, a model of the background is essential. In a frequentist search, knowledge of the background enters in two ways: in designing the test statistic T and in obtaining its sampling distribution under the null (i.e., background-only) hypothesis $\mathcal{H}_0$. There is an underappreciated distinction between these two uses of the background-model. Even if the test statistic is derived using an ill-suited background model, as long as its sampling distribution under an accurate $\mathcal{H}_0$ is found, one obtains a test that properly controls the false detection rate and returns a well-calibrated p value.

It is thus understandable that much work has been focused on ensuring that the null distribution of T (i.e, the sampling distribution of T under $\mathcal{H}_0$) correctly captures the behaviour of the true background distribution (for dark matter searches in dwarf galaxies, see, e.g., Refs. [44, 45] and many subsequent works).

However, relatively less attention has been paid to how the background enters into the design of the test statistic. The price to be paid for a failure in this area is suboptimal sensitivity to the signal of interest (even if the test procedure calculates the correct false detection rate). Without improving the background model used in the test statistic itself, the reach of searches for interesting and/or weak signals will be fundamentally limited. This paper aims to address this problem.

B. Source detection and hypothesis testing

In the frequentist context, source detection is formulated as a hypothesis test, in which the null hypothesis, $\mathcal{H}_0$, assumes the data in the ROI are generated by the same process that produces the background in its vicinity (i.e, the background-only model) and the alternative model, $\mathcal{H}_1$, assumes a background plus signal model (e.g., background plus dark matter annihilation). One then derives a test statistic $T(D)$ that is a function of the data, D, (e.g., a likelihood ratio, a particular summary statistic, or other cleverly-crafted function of the data). The test statistic should be designed so that extreme values of T reflect the presence of the sought-for signal. Evidence for a new source in the observed data $D_{\rm obs}$ is quantified by the p value: $p = P\left(T(D) > T_{\rm obs} \mid \mathcal{H}_0\right)$, where $T_{\rm obs} \equiv T(D_{\rm obs})$ is the value of the test statistic for the observed data[1]. The p value is thus the probability of the observed data that is as "signal-like" (or even more so) under the hypothesis that there is no new source present. Small values of p are thus often interpreted as evidence for a source discovery. In other words, the p value is often interpreted as a measure of detection significance. Caution, however, must be taken in the subtle interpretation of a p value: it indicates how incompatible the data are with the specified null model, but is not the probability that the null model itself is true, nor is it the probability that the exact observed data were produced by random chance [46, 47].

To characterise the sensitivity of an experiment for detecting a source of a particular specification (in terms of the parameterisation of $\mathcal{H}_1$, e.g., the source strength) it is helpful to consider its *median significance* [48]. This is the median, $p_{\rm med}$, of the distribution of p values in the case that the specified source signal is present. It is calculated by finding the median of T in repeated pseudo-experiments[2] with the specified injected signal and then computing the p value for the median T. In essence, the median sensitivity quantitatively represents the typical detection significance the experiment will achieve for a given signal model: half the time the signal will produce a higher significance than $p_{\rm med}$ and half the time at a lower significance.

For a Bayesian, source detection is not typically framed in terms of rejecting a null hypothesis or model, but rather it is viewed as a model comparison problem. Either the posterior odds or the Bayes factor comparing the two competing models (background-only vs background + source, e.g., dark matter annihilation) is evaluated [e.g., by comparing the Bayes factor against the Jeffreys' scale, 49]. In contrast to a frequentist hypothesis test, the two models are treated symmetrically and either may be favoured by the Bayes factor (or odds). Indeed, the in-built principle of Occam's razor penalises model with unnecessary free parameters and thus generally favours the simpler model [50, 51]. [AGS: maybe this frequentist vs bayesian comparison can be put in the discussion. Like the main message of the paper says "here's how you can do a better frequentist search" and then in the discussion we can say "by the way, since the test is based so heavily on Bayesian concepts it would be pretty easy to adapt it into a fully Bayesian model selection problem, get the evidence ratio, use the Jeffreys scale etc". Then in this paragraph we can just keep the one or two Bayesian sentences and say "we discuss this more later in Sec.XX".]. [DvD: A full comparison does seem a bit out of scope for the paper, but could definitely go in the discussion.] [RT :Note: WALDO paper also uses the BF; include discussion]

The Bayes Factor is defined as the ratio of the evidences (or marginal likelihoods) of the two models under

[1] We assume a one-sided hypothesis test, where larger values of T are less likely under $\mathcal{H}_0$ and thus are more "extreme".

[2] This calcuation of $p_{\rm med}$ assumes a one-sided test. In the more general setting $p_{\rm med}$ is computed as the median of the p values computed from each pseudo-experiment.

comparison, [49],

$$BF = \frac{P\left(D \mid \mathcal{H}_1\right)}{P\left(D \mid \mathcal{H}_0\right)}. \tag{1}$$

For even prior odds (i.e., $P\left(\mathcal{H}_1\right)/P\left(\mathcal{H}_0\right) = 1$), the posterior odds are equal to the Bayes factor.

The suggestion to use a Bayes factor as a significance criterion, that is, to consider its distribution under the null, goes back (insofar as we are aware) to [52]. Ever since, the statistics literature has debated the merits of using Bayes factors as test statistics, sometimes in the context of a Bayesian/Frequentist compromise [53–55]. Since in the evidence the parameters for each hypothesis are averaged over, using the Bayes factor as test statistics means that the test is between two simple hypotheses (i.e., neither has any free parameters). Therefore, the Neyman-Pearson lemma applies, just as in the case of the standard likelihood ratio test (LRT) for simple hypotheses. The key difference is that in the LRT the significance level of the test, α, corresponds to the type I error rate (i.e., to the probability, over repeated draws of the data, of incorrectly rejecting the null when it is in fact true). By contrast, when adopting the Bayes factor as a test statistics the significance α controls the type-I error rate *averaged* over the prior distribution for the parameters of the null hypothesis [56, 57], which need not correspond to the Frequentist type-I error rate. By introducing an Empirical Prior (see below) and using the Bayes factor as test statistics, our aim is to control the *average* type-I error rate of the test: i.e., to the extent that our EP is a representative distribution of background across the sky, our test controls the average error rate for randomly sampled regions of the sky.

[RT :revise and maybe move this to discussion later]

C. The empirical prior

Suppose the background model can be parameterized in terms of a set of parameters $\boldsymbol{\phi}$. In order to specify a prior distribution on $\boldsymbol{\phi}$, we consider a population of $n_{\rm bg}$ subregions, A_i, for $i = 1, \ldots n_{\rm bg}$ each with their own background parameter $\boldsymbol{\phi}_i$, where the subregions together partition a larger region surrounding the ROI. We assume that the background parameters for the ROI is a random sample drawn from the distribution of background parameters for the subregions, defined later in Sec. III A.

The fundamental assumption underlying the empirical prior is that by sampling the spatial region around the target ROI we obtain a representative distribution of *all* background processes that occur in the direction of the target. These processes, even if individually unidentifiable in one sample, can be captured by the mixture model represented by the empirical prior. In particular, we wish empirical prior to describe the presence of unresolved point sources to arbitrarily low fluxes, which would not appear as detectable sources in a conventional analysis.

We begin by covering the spatial region around the target ROI with $n_{\rm bg}$ (overlapping) sub-regions A_i, $i = 1, \ldots, n_{\rm bg}$, as illustrated in Fig. 1. In each sub-region i, the likelihood is given by $P(\boldsymbol{D}_i^{\rm bg} \mid \boldsymbol{\phi})$, the probability of observing the sub-region's data $\boldsymbol{D}_i^{\rm bg}$ under a background model specified by $\boldsymbol{\phi}$. We adopt a binned Poisson likelihood function throughout our analysis as usual in Fermi-LAT analyses [58, 59]:

$$P(\boldsymbol{D}_i^{\rm bg} \mid \boldsymbol{\phi}) = \prod_{i,j} \frac{\lambda_{i,j}^{n_{i,j}}}{n_{i,j}!} \mathrm{e}^{-\lambda_{i,j}} \ , \tag{2}$$

where $\lambda_{i,j}$ is the expected number of counts in bin j for sub-region A_i, specified in Sec.III A, with $n_{i,j}$ the observed number of counts in that bin; we allow j to index both spatial and energy bins. We define a prior, $\pi_{\rm ur}(\boldsymbol{\phi})$, on the background parameters ϕ, in Sec.III A, which encapsulates our prior belief about the background parameters $\boldsymbol{\phi}$ at a *randomly selected spatial location*[3]. One may think of $\pi_{\rm ur}$ as the initial, uninformed state of belief about the distribution of background parameters. As we will discuss later, this initial prior can be rather non-informative, but for our demonstration we choose it to be informed by the population of spectra of identified point-sources observed and presented in the Fermi-LAT collaboration source catalogs [60, 61].

We seek a distribution over the parameters $\boldsymbol{\phi}$ that will describe the background processes at work in a randomly selected sub-region. If we denote the complete data from all the sub-regions (excluding the target ROI) as $\boldsymbol{D}^{\rm bg} = \{\boldsymbol{D}_1^{\rm bg}, \ldots, \boldsymbol{D}_{n_{\rm bg}}^{\rm bg}\}$, then:

$$\begin{aligned} P(\boldsymbol{\phi} \mid \boldsymbol{D}^{\rm bg}) &= \sum_{i=1}^{n_{\rm bg}} P(\boldsymbol{\phi}, A_i \mid \boldsymbol{D}^{\rm bg}) \\ &= \sum_{i=1}^{n_{\rm bg}} P(\boldsymbol{\phi} \mid A_i, \boldsymbol{D}^{\rm bg}) P(A_i \mid \boldsymbol{D}^{\rm bg}) \\ &= \frac{1}{n_{\rm bg}} \sum_{i=1}^{n_{\rm bg}} P(\boldsymbol{\phi} \mid \boldsymbol{D}_i^{\rm bg}) \\ &= \frac{1}{n_{\rm bg}} \sum_{i=1}^{n_{\rm bg}} \frac{P(\boldsymbol{D}_i^{\rm bg} \mid \boldsymbol{\phi}) \pi_{\rm ur}(\boldsymbol{\phi})}{Z_i^{\rm ur}} \ . \end{aligned} \tag{3}$$

The A_i in the first line denotes which sub-region was selected. In the second-to-last line we used the fact that $P(\boldsymbol{\phi}|A_i, \boldsymbol{D}^{\rm bg}) = P(\boldsymbol{\phi} \mid \boldsymbol{D}_i^{\rm bg})$, i.e., conditioning on region i means that $\boldsymbol{\phi}$ is informed only by the data $\boldsymbol{D}_i^{\rm bg}$ from that region. We also assumed that that $P(A_i \mid \boldsymbol{D}^{\rm bg}) = P(A_i) = 1/n_{\rm bg}$, since each background region is selected at random from the $n_{\rm bg}$ available, independently of $\boldsymbol{D}^{\rm bg}$. The last line includes the Bayesian evidence, defined by

$$Z_i^{\rm ur} = \int_{\boldsymbol{\phi}} P(\boldsymbol{D}_i^{\rm bg} \mid \boldsymbol{\phi}) \pi_{\rm ur}(\boldsymbol{\phi}) \mathrm{d}\boldsymbol{\phi}. \tag{4}$$

[3] The choice of subscript for this prior, which we call "the ur-prior", come from the German word "ur", for "primitive".

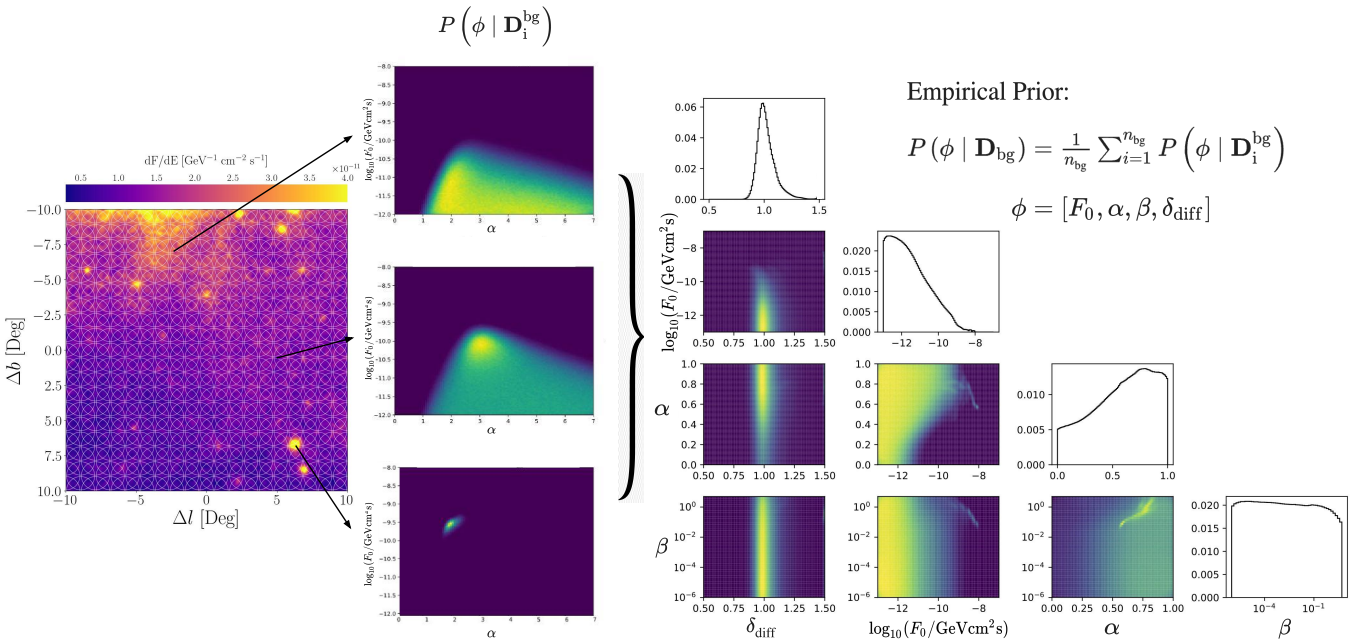

FIG. 1. Visual demonstration of the derivation of the empirical prior, obtained from Fermi data in a $20^\circ \times 20^\circ$ ROI centered on the location of the dwarf galaxy Carina II at RA = 114.11° and Dec = -58.0° (axes are relative latitude and longitude $(\Delta b, \Delta l)$ in a Cartesian projection centered at this RA and Dec). From each sub-region A_i (white circles in the left panel) we obtain a posterior density over the background model parameters $\phi = [F_0, \alpha, \beta, \delta_{\rm diff}]$ (middle panels, for three different sub-regions). We then average these posteriors over the different sub-regions to form the empirical prior (triangle plot on the right).

The quantity $P(\boldsymbol{\phi} \mid \boldsymbol{D}^{\rm bg})$ on the left-hand side of Eq. (3) is thus the posterior distribution for the background model's parameters averaged over the background regions. This object serves as a "predictive posterior distribution" as commonly used in Bayesian updating [62, 63]. Another way of looking at it is as an equal-weight mixture of the posterior distributions for each of the background region, noting how the probability density inside the sum in simply the posterior probability for $\boldsymbol{\phi}$ given each sub-region's data. A diagrammatic demonstration of the derivation of this distribution is illustrated in Fig. 1.

We denote the value of the background parameters for the target ROI by $\boldsymbol{\phi}^\star$, emphasizing once more that the target ROI's data are *not* used in Eq. (3). We now reinterpret the posterior distribution $P(\boldsymbol{\phi} \mid \boldsymbol{D}^{\rm bg})$ as a prior for the background processes at work in the direction of the target, i.e.

$$\pi_{\rm ep}(\boldsymbol{\phi}^\star) \equiv P(\boldsymbol{\phi}^\star \mid \boldsymbol{D}^{\rm bg}). \tag{5}$$

We call this distribution the *empirical prior*: it contains information about the abundance and properties of all components of the background in the vicinity of the target, including unresolved faint point sources of all types. Using the empirical prior as the background model for our target of interest should correctly include possible interloper point sources and residual modeling uncertainties of the diffuse and isotropic background components, as we discuss in Sec. III A. It also properly handles the fact that knowledge of the background comes from limited data $\boldsymbol{D}_i^{\rm bg}$ in each sub-region, as encoded by the individual sub-region posteriors.

D. The Bayes factor as a test statistic

The empirical prior serves to define a test statistic designed to distinguish between two models for the ROI data: the first, that the data are a realisation from the same background processes at work in the surrounding sub-regions (including unknown faint contributors and variations from fixed templates); the second, that the ROI data include an additional signal of interest (in our case, dark matter annihilation).

Let $\boldsymbol{\theta}_s$ denote the set of parameters that describe the signal of interest (e.g., the dark matter particle mass, annihilation cross-section, and density profile of the dwarf galaxy halo) and $\pi(\boldsymbol{\theta}_s)$ be a prior on these parameters. The data from the ROI is $\mathbf{D}^\star$. We propose to use as test statistic the evidence ratio (i.e., the Bayes factor) between the background-only null model and a back-

ground+signal alternative. In both null and alternative, the background-only model includes the empirical prior over the background parameters $\boldsymbol{\phi}$. The test statistic $T_{\rm BF}$ is thus:

$$T_{\rm BF}(\mathbf{D}^\star) = \frac{P\left(\mathbf{D}^\star \mid \mathcal{H}_{\rm ep+s}\right)}{P\left(\mathbf{D}^\star \mid \mathcal{H}_{\rm ep}\right)} = \frac{Z_{\rm ep+s}(\mathbf{D}^\star)}{Z_{\rm ep}(\mathbf{D}^\star)}, \tag{6}$$

where

$$Z_{\rm ep}(\mathbf{D}^\star) = \int d\boldsymbol{\phi}\, P(\mathbf{D}^\star \mid \boldsymbol{\phi}) \pi_{\rm ep}(\boldsymbol{\phi}), \tag{7}$$

is the evidence for the background-only model, and

$$Z_{\rm ep+s}(\mathbf{D}^\star) = \int d\boldsymbol{\phi} \int d\boldsymbol{\theta}_s\, P(\mathbf{D}^\star \mid \boldsymbol{\phi}, \boldsymbol{\theta}_s) \pi_{\rm ep}(\boldsymbol{\phi}) \pi(\boldsymbol{\theta}_s) \tag{8}$$

is the evidence for the model that includes a signal contribution.

It is useful to contrast our approach with conventional source detection based on a maximum profile likelihood ratio test statistic, $\lambda(\mathbf{D}^\star)$, defined as [64, 65],

$$\lambda(\mathbf{D}^\star) = 2 \ln \frac{\sup\limits_{\theta_b, \boldsymbol{\theta}_s} P\left(\mathbf{D}^\star \mid \theta_b, \boldsymbol{\theta}_s\right)}{\sup\limits_{\theta_b} P\left(\mathbf{D}^\star \mid \theta_b, \boldsymbol{\theta}_s = 0\right)}, \tag{9}$$

where the parameters are fit via optimisation rather than marginalised over as in Eq. (6). Another important distinction is in the background model used in Eq. 9. Conventional background models (denoted by θ_b above) do not include a distribution over the unknown background processes that can only be quantified by empirically sampling the sky, but only fairly rigidly parameterized contributions from physics-based templates. Such a fixed parametric form of the likelihood can translate into systematic errors when evaluating the p-value for the presence of a signal [66–69]. The empirical prior, as introduced in Sec. II C, alleviates such mis-specification by incorporating in an agnostic way the influence of all background sources in the region around the target.

Previous work has sought to alleviate this problem in the gamma-ray setting, while retaining the maximum likelihood approach. Refs. [37, 70] included potential faint sources associated with identified radio galaxies, though without a term analogous to the empirical prior that quantifies the abundance and properties of such sources. The method of Linden [42], discussed in the introduction, is closer to what we propose here. In that work the background model at every point includes an additional contribution from dark matter annihilation with a free amplitude informed by a sampling procedure, and employs a maximum likelihood test statistic. A key distinction of our method is that we instead incorporate the information contained in the empirical prior by using it in the computation of a Bayes factor test statistics.

III. APPLICATION TO DARK MATTER SEARCHES IN DWARFS

In this section we demonstrate the benefits of the empirical prior by carrying out a sensitivity analysis for a dark matter annihilation search in dwarf galaxies. We investigate two scenarios: a dark matter signal emitted from the target (in this case, a simulated Carina II dwarf) and emission from an ordinary astrophysical power law source that is coincidentally located at the position of the target. We compare the performance of what we call Hypothesis Testing with the Empirical Prior (HTEP), which incorporates the informative empirical prior $\pi_{\rm emp}$, with a conventional test based on the profile likelihood ratio assuming the standard isotropic+diffuse Poisson background model (and no empirical prior information). The two scenarios (i.e., a dark matter source and power law interloper at the location of the targer) are chosen to have similar values of the likelihood ratio test statistic, which means that conventional likelihood ratio tests cannot distinguish between the two, which limits the maximum significance that a conventional test could return for a dark matter signal. We then show that using the empirical prior with the Bayes factor test statistic we can distinguish between the two scenarios, which enables the discovery of fainter dark matter signals.

A. Background model for the test statistic

In the calculation of the test statistic $T_{\rm BF}$, the background in sub-region A_i ($i = 1, \ldots, n_{\rm bg}$) is described by a model that is the sum of the following components:

1. A fixed template corresponding to the isotropic gamma-ray background $T_{\rm iso}$, including instrumental effects and misidentified cosmic ray contamination [71]. The isotropic background was determined by the Fermi Collaboration via a full-sky fit [72]. The uncertainties on the energy spectrum amount to no more than 1.3% for the most important energies below about 30 GeV and less than about 9% in the remaining energy range we consider up to 500 GeV. For this reason, we do not introduce nuisance parameters for the isotropic background and instead fix its contribution to the value given by the provided template[4].

2. Diffuse galactic emission, with a parameter we denote $\delta_{\rm diff}$ that scales the amplitude relative to the baseline Fermi model $T_{\rm diff}$. The contribution of Galactic diffuse emission is captured by the Fermi Collaboration's galactic interstellar emission

[4] `http://fermi.gsfc.nasa.gov/ssc/data/access/lat/BackgroundModels.html`; `iso_P8R3_SOURCE_V3_v1.txt`.

model[5], created from a full-sky fit to physically-motivated templates derived with the `GALPROP` cosmic ray propagation code [73, 74]. The uncertainties in this model are not easily quantified and so the energy-independent normalization factor $\delta_{\rm diff}$ allows for local deviations from unit scaling[6].

3. To account for the presence of faint interloping sources along the line of sight, but unrelated to the dark matter signal of interest, we add an additional point source at the center of the ROI whose spectrum is described by a total flux $F_{\rm tot}$, spectral index α, and spectral curvature β, according to the "log-parabola" form,

$$\frac{dF(E)}{dE} = F_0 \left(\frac{E}{E_0}\right)^{-\alpha-\beta\log\left(\frac{E}{E_0}\right)} . \tag{10}$$

The reference energy $E_0 = 1$ GeV is fixed and F_0 is a function of $F_{\rm tot}$, α, and β, obtained from the relation $F_{\rm tot} = \int_{0.5\ \rm GeV}^{500\ \rm GeV} dF/dE\, dE$. Allowing for curvature in the spectrum is important here because this can differentiate a dark matter signal (expected to have a highly curved spectrum) from astrophysical gamma-ray sources, whose spectra are nearly always much flatter ($\beta \sim 0$), see, e.g., Ref. [75].

The overall background model for each region is thus the sum of the templates from the isotropic and diffuse contributions, plus the additional point source:

$$\frac{{\rm d}\Phi_{\rm bkg}}{{\rm d}E{\rm d}\Omega} = T_{\rm iso} + \delta_{\rm diff} T_{\rm diff} + \frac{dF(E)}{dE}, \tag{11}$$

where ${\rm d}\Phi_{\rm bkg}/{\rm d}E{\rm d}\Omega$ is total intensity of the background model (photons per energy per solid angle final) and the final. The set of background model parameters is thus $\boldsymbol{\phi} = (F_{\rm tot}, \alpha, \beta, \delta_{\rm diff})$.

For the methodological demonstration in this paper, we construct the empirical prior using ~ 12.5 years of Fermi-LAT observations[7] of a $20^\circ \times 20^\circ$ region centered on the location of the nearby dwarf galaxy Carina II, located at RA = 114.11° and Dec = -58.0° [76, 77]. The gamma-ray events are binned spatially into 0.05° pixels and into 30 log-spaced energy bins between 0.5 and 500 GeV. We form $n_{\rm bg} = 10^4$ circular sub-regions, each of radius 1°, centred on a 100×100 equally-spaced grid of locations between ±8° of Carina II and oriented along galactic latitude and longitude. The binned events within each sub-region form the data sets $\boldsymbol{D}_i^{\rm bg}$. Note that the sub-regions are highly overlapping, which is appropriate given our definition of the empirical prior: the background processes in the target direction will be taken to be those from a single randomly selected location in the vicinity of the target.

The Carina II dwarf [78] was chosen as it a nearby, potentially attractive dark matter target. Upon its discovery in 2018 we noticed a slight gamma-ray excess with a spectrum quite similar to that of the Reticulum II dwarf [79], consistent with dark matter annihilation (see also Ref. [80]). However, its location fairly close to the Galactic plane and the abundance of similarly significant nearby locations makes a dark matter interpretation highly fragile. The present framework is an attempt to go deeper. Rather than being forced to dismiss a potentially valuable dark matter target due to the possibilities of interlopers or diffuse background mis-modeling, we aim to show that it is possible to restore sensitivity through a more careful probabilistic characterisation of the observed gamma-ray background.

To compute the empirical prior distribution we need to set the ur-prior $\pi_{\rm ur}$. In principle one could make an uninformative choice and let the empirical prior update the initial uncertainty in the parameters to a more informed distribution. We propose using a more informative ur-prior that reflects at least some prior knowledge of the properties of faint point sources. We chose $\pi_{\rm ur}(\boldsymbol{\phi}^\star)$ in such a way that it represents the spectral diversity of the Fermi point source catalogue [60, 61]:

$$\pi_{\rm ur}(\boldsymbol{\phi}^\star) = \begin{cases} \delta_{\rm diff} \sim \text{Uniform}\,[0.5, 1.5]\,, \\ \log_{10}(F_0/\text{GeVcm}^2\text{s}) \sim \text{Uniform}\,[-13, -7]\,, \\ \log_{10}\beta \sim \text{Uniform}\,[-6, , \log_{10}(5)]\,, \\ \alpha|\beta \sim \text{Uniform}\,[-2\beta\log(500), 4 - 2\beta\log(2)]\,. \end{cases} \tag{12}$$

The above choice was not fit to the gamma-ray data to be analyzed ($\boldsymbol{D}^{\rm bg}$) in any way. Rather, it was designed "by eye" so that sample spectra drawn from the prior cover a range of realistic-looking spectra without being too strongly weighted toward any particular shape. In particular, the conditional prior on α given β is such that the peak of the spectral energy distribution $\log E_{\rm peak}$ [75] is roughly uniform between 0.5 and 500 GeV.

There are likely improvements in the ur-prior that would result in a further increase of sensitivity. For now, we make two comments. First, we note that a "bad" choice of ur-prior will not lead to incorrect conclusions (i.e. erroneous detection p values): it would simply result in a test statistic that is less sensitive to a dark matter signal than it could be. Second, the ur-prior's

---

[5] `gll_iem_v07.fits`.

[6] An analogous approach was taken in previous studies of the background [e.g.,49]. The introduction of such dwarf-dependent scaling factors is important since the empirically derived background surrounding the dSphs has been shown to deviate from the Fermi Galactic interstellar emission model in ways beyond what is expected from Poisson fluctuations [15, 31, 50]. We note that using an energy-independent rescaling factor for the background may not fully account for an additional effect due to the background residual's spectral shape varying from location to location.

[7] We use Pass 8 SOURCE events prepared by the Fermi Science Tools (`http://fermi.gsfc.nasa.gov/ssc/`) with the parameters `zmax=90` and `filter="(DATA_QUAL>0) && (LAT_CONFIG==1)"`. The $20^\circ \times 20^\circ$ binned counts map is obtained with the `gtbin` tool in the `CAR` projection aligned with galactic longitude and latitude.

effect on the test statistic is reduced because it is combined with the likelihood $\boldsymbol{D}_i^{\rm bg}$ before being used in the empirical prior. It is not quite a hierarchical Bayesian framework because the ur-prior enters in each sub-region separately, rather than as a single hyperprior shared by all likelihoods at the sub-region level. The refinement of the ur-prior and whether a true hyperprior can be used instead are interesting topics for future work.

With the choice of ur-prior as in Eq. (12), the resulting empirical prior is shown in the right-most panel of Fig. 1. The Z_i^{ur} denominators in the empirical prior from Eqn. 4 were calculated using `MultiNest` [81] with 1000 live points and a tolerance of 0.1. We observe interesting structure in the marginal distribution for F_0 and α, corresponding to the extensive population of point sources in the ROI. The majority of the empirical prior's density is concentrated at low F_0 values, where it captures the presence of faint, below-threshold sources in the ROI. The 1D marginal posterior for the diffuse component scaling parameter $\delta_{\rm diff}$ remains highly peaked at a value of 1, showing that the diffuse component normalisation is in good agreement with the ROI. The posterior for the curvature parameter β is unconstrained, showing no preference for curved spectra in the population of point sources.

B. Dark matter signal model

The differential photon flux from dark matter annihilation has the well-known form

$$\frac{{\rm d}F_{\rm DM}}{{\rm d}E} = J\frac{\langle\sigma v\rangle}{8\pi m_\chi^2}\frac{dN_\gamma}{dE}, \tag{13}$$

where dN_γ/dE is the spectrum of gamma rays emitted from a single pair annihilation event, $\langle\sigma v\rangle$ is the velocity-weighted dark matter annihilation cross-section, m_χ the dark matter mass. We assume annihilation into $\tau^+\tau^-$ lepton pairs, and use the dN_γ/dE spectra compiled by Cirelli *et al.* [82] with electroweak corrections [83].

We assume the spatial morphology of dark matter annihilation is narrower than the telescope PSF so that the emission appears point-like. In this case, a scalar J is sufficient to quantify the macroscopic distribution of dark matter within the dwarf galaxy halo [84, 85]. Since the annihilation cross-section and J are perfectly degenerate, we may fix $\langle\sigma v\rangle = 3\times10^{-26}$ cm^2/s, corresponding to a canonical WIMP dark matter particle with relic abundance matching observations. Therefore, the dark matter model parameters reduce to $\boldsymbol{\theta}_{\rm DM} = (m_\chi, J)$. We set the prior on the dark matter model parameters as

$$\pi(\boldsymbol{\theta}_{\rm DM}) = \begin{cases} \log_{10}(m_\chi/{\rm GeV}) \sim {\rm Uniform}\left[\log_{10}(6), 4\right], \\ \log_{10}(J/{\rm GeV}^2{\rm cm}^{-5}) \sim {\rm Uniform}\left[17, 22\right]. \end{cases} \tag{14}$$

The dark matter mass range is chosen to span the orders of magnitudes relevant for WIMP candidates able to annihilate into a $\tau^+\,\tau^-$ final state, with an upper limit chosen to match the preferred weak scale SUSY scenarios, of order 10 TeV. The range of values for J is informed by what has been observed for nearby dwarfs, spanning from $\log_{10} J \sim 17$ to 21, as inferred from dynamical modelling, e.g., [84–88].

C. Dark matter annihilation vs. an astrophysical interloper

1. *Hypothesis testing with the empirical prior (HTEP)*

As motivated in Secs. II B and II D, we adopt as frequentist test statistic the Bayes factor between a model for the gamma-ray emission at the target that includes both dark matter and a background model vs a null background-only model, where in both cases the background model is informed by the empirical prior on possible point source contributions. The explicit form of the test statistic given by Eq. (6) (with s=$\{{\rm dm}\}$) is then:

$$\begin{aligned} T_{\rm BF}(\mathbf{D}^\star) &\equiv \frac{Z_{\rm dm+ep}}{Z_{\rm ep}} \\ &= \frac{\int d\boldsymbol{\theta}_{\rm DM}\int d\boldsymbol{\phi}^\star\; p(\mathbf{D}^\star \mid \boldsymbol{\theta}_{\rm DM}, \boldsymbol{\phi}^\star)\,\pi(\boldsymbol{\theta}_{\rm DM})\,\pi_{\rm ep}(\boldsymbol{\phi}^\star)}{\int d\boldsymbol{\phi}^\star\; p(\mathbf{D}^\star \mid \boldsymbol{\phi}^\star)\,\pi_{\rm ep}(\boldsymbol{\phi}^\star)}, \end{aligned} \tag{15}$$

where $Z_{\rm dm+ep}$ is the Bayesian evidence for the target data under a model that includes both dark matter (dm) and the background model with the empirical prior (ep), and $Z_{\rm ep}$ is the Bayesian evidence under a model with only the empirical prior. The prior $\pi(\boldsymbol{\theta}_{\rm DM})$ is given in Eq. (14), while the background components parameters' empirical prior $\pi_{\rm ep}(\boldsymbol{\phi}^\star)$ is described in Sec. III A. The quantity $Z_{\rm ep}$ represents the evidence for the background model alone, with the empirical prior applied to its parameters. The likelihood functions in Eq. (15) take the Poisson form of Eq. (2).

The simulated data at the location of the target, $\mathbf{D}^\star$, is generated from the Fermi diffuse and isotropic background templates as described in Sec. III A, with the addition of a source of either DM annihilation as parameterised by the model in Sec. III B or a point source with spectrum given by Eqn. (10), which we call an "interloper", i.e., a point source that might be mistaken for DM at the target.

In order to compute the p-value for this test statistic, we need to establish its asymptotic distribution, which requires computing $T_{\rm BF}$ over $N \gg 1$ realisations of simulated target data[8]. The Bayes factor from Eqn. (15) is calculated for each simulated dataset by using the nested sampling implementation `Multinest`[81, 89, 90] to evaluate the model evidences. We use 1000 live points and

[8] Note that we keep the empirical prior fixed under replica of the target data, i.e., the empirical prior (obtained from $n_{\rm bg}$ subregions around the ROI in question) is computed only once.

a tolerance of 0.1. We observed a good compromise between stability of the evidence evaluation and computation time with this choice of hyperparameters.

2. *Maximum profile likelihood ratio test*

In order to benchmark our novel HTEP approach, we compare with the standard maximum profiled likelihood ratio test. In the standard scenario, which does not use information from the EP, one would perform a hypothesis test comparing a background model encompassing a fixed isotropic component and a diffuse component with a free normalization, with a model in which an additional dark matter source was present. The profile likelihood test statistics given in Eq. (9), becomes in this case:

$$\lambda_{\rm ML} = 2\log\left(\frac{\hat{L}_{\rm dm+bkg}}{\hat{L}_{\rm bkg}}\right) = 2\log\frac{\sup_{\boldsymbol{\theta}_{\rm DM},\delta_{\rm diff}} P(\mathbf{D}|\boldsymbol{\theta}_{\rm DM},\delta_{\rm diff})}{\sup_{\delta_{\rm diff}} P(\mathbf{D}|\boldsymbol{\theta}_{\rm DM}=0,\delta_{\rm diff})}. \quad (16)$$

The likelihood $\hat{L}_{\rm dm+bkg}$ is again Poisson in nature and it is maximized over $\delta_{\rm diff}$ and the DM model parameters introduced in Sec. III B, while $\hat{L}_{\rm bkg}$ is maximized over the galactic diffuse amplitude parameter $\delta_{\rm diff}$ with the dark matter signal is fixed to zero. The minimisation of the log-likelihood ratio in Eqn. (16) is carried out using the 'optimize' module from `scipy` [91]. Given the low-dimensional nature of the optimisation we observed stable and consistent behaviour in the result.

D. Asymptotic distribution of test statistics and Asimov data

In order to perform a hypothesis test for the presence of a DM source at the target location, we need to determine the asymptotic distributions of both $T_{\rm BF}$ and $\lambda_{\rm ML}$ under data generated from the null hypothesis.

Our fundamental assumption is that the EP captures the nature of the background distribution in absence of a DM source. Therefore, under the background-only null hypothesis, we simulate data at the target location by sampling parameter values from the 4-dimensional empirical prior and, for each sample, building the predicted gamma-ray flux from Eq. (11). We then generate simulated data from the Poisson likelihood and use Eqs. (15) and (16) to compute the value of $T_{\rm BF}$ and likelihood ratio test statistics, respectively, and, over repeated sampling, generate their respective distribution under the null hypothesis. We notice that the likelihood ratio test features a misspecified background model, with only a single free parameter (the diffuse template normalization), while the background-only simulated data additionally include a contribution from undetected point sources, weighted by the EP (which is not included in the traditional profile likelihood analysis). By contrast, the BF contains the information from the EP, and therefore includes a correctly specified background model. This crucial difference will later play out in the increased sensitivity of the BF test statistics, as shown below.

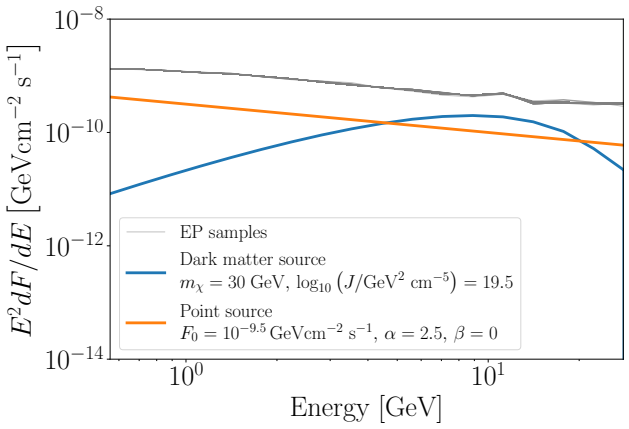

FIG. 2. Differential fluxes of the DM and point source benchmarks used to test the relative discerning power of the E-Bayt and profile likelihood based statistics presented and discussed in Sec. III E. The grey lines represent the spectra in the $20^o \times 20^o$ ROI centred on Carina II pseudo-data samples drawn from the empirical prior [A: Not sure if these EP samples make sense.]. [RT :This figure seems wrong: how can the PS be confused with the DM?]

We emphasize once again that this approach relies on the empirical prior being a correct description of additional components (beyond the isotropic and diffuse templates) being present in the true background – consistent with the fundamental assumption that, whilst not a perfect representation of the 'true' background model, the empirical prior is an effective description of the background variability. We also note that our setup is idealized, in that the simulated data come from precisely the same model as is included in the BF background – which is unlikely to hold exactly true for real data.

Next, we need to determine the distribution of $T_{\rm BF}$ and of the likelihood ratio statistics under data generated from the alternative model. To this end, we simulate datasets $\mathbf{D}^\star_{\rm DM}$ and $\mathbf{D}^\star_{PS}$ including the presence of annihilating DM or of a point source interloper at the target location. This target location is located at a fixed relative galactic latitude and longitude within the $20°\times20°$ region in which the empirical prior was constructed, specifically, at 5 deg longitude from the centre $\Delta b, \Delta l = [0°, 5°]$. Similarly to the procedure above, we simulate each data set by adding the additional spectral component (DM or PS, with parameters fixed to benchmark values given below) to the background spectrum computed from the EP, and subsequently drawing Poisson counts.

The resulting distributions of $T_{\rm BF}$ and $\lambda_{\rm ML}$ constitute the alternative hypothesis for the test, $\mathcal{H}_1$.

In order to conduct a sensitivity study, it is necessary to simulate a large number of pseudo-data realizations under various dark matter model parameters, $\boldsymbol{\theta}_{\rm DM}$. For each realization, we calculate the detection significance using both the $T_{\rm BF}$ test statistic and the likelihood ratio test statistics. Sensitivity can be quantified by estimating statistical power, i.e., the fraction of realizations rejecting the null hypothesis at a (conventional) 3σ level, or by finding the median detection significance ($p_{\rm med}$), i.e., the median p-value obtained from pseudo-data [RT :but we only do the latter, no? If so, remove the former]. This is how we carry out the comparison in section III E for one single DM vs interloper scenario, but it would be computationally demanding to extend such an approach to many values of $\boldsymbol{\theta}_{\rm DM}$.

In order to avoid the computational burden of extensive simulations, for the sensitivity analysis in Sec. III F, we exploit the idea of the so-called "Asimov data set" for the signal hypothesis, a concept introduced in Ref. [92]. The Asimov data set is a single realization of the data, in which the counts in each bin are equal to the expected counts the alternative model. As shown in **??**, this dataset behaves like a "median data set", allowing us to estimate the asymptotic $p_{\rm med}$ using a single realization.

E. Sensitivity comparison for three benchmarks

We now compare the sensitivity (defined as median p-value) the profile likelihood method and our BF test statistics with simulated target data, dt, being produced from either a DM annihilation signal or a point source interloper. Specifically, we simulate $\mathbf{D}^\star$ under three scenarios:

1. Dark matter annihilation into $\tau^+\tau^-$ with $m_\chi = 30\,\text{GeV}$ and $\log_{10} J = 19.5$ with fixed velocity averaged annihilation cross-section of $\langle\sigma v\rangle = 3\times10^{-26}$ cm^{3}s^{-1}.

2. A power law point source interloper with parameters $\phi^* = \{\log_{10}(F_0) = -9.5, \alpha = 2.5, \beta = 0, \delta_0 = 1\}$.

3. No signal: Data realisations from samples of the EP that represent the null hypothesis.

The differential flux spectra for the three above components are shown in Fig. 2 [RT :this figure seems wrong]. We present the results of this benchmark analysis in Fig. 3, Fig. 4 and Fig. 5, while a sensitivity study over the full range of DM parameters is discussed in Sec. III F.

We show the distribution of the test statistics $T_{\rm BF}$ and $\lambda_{\rm ML}$ under mock data realisations for the three aforementioned scenarios in the left panel of Fig. 3, where we also plot the Asimov data for the DM and point source benchmarks. This first result demonstrates that the mode of the distribution of $T_{\rm BF}$ (vertical axis) is displaced to larger values for the case of a DM signal (purple), compared to the case when the signal is due to a point source interloper (orange). In the latter case, the distribution of $T_{\rm BF}$ is consistent with that under the null hypothesis, which is expected given that the empirical prior was constructed to represent the existence of possible point sources at the target location [RT :add null hypothesis marginal to the right axis]. By contrast, the power of the likelihood ratio test statistics to distinguish between the two signal benchmarks is almost zero, as reflected by the high degree of overlap in the distribution of $\lambda_{\rm ML}$ for DM and point source samples along the horizontal axis. The empirical Bayes factors' superior ability to distinguish a DM signal from a point source interloper is additionally shown by the separation between Asimov data points along the vertical axis, but not along the horizontal axis.

The right panel of Fig. 3 shows the distribution of the p-value for rejecting the null hypothesis (that the data are generated from the EP, see Sec.II B) for each pseudo-data sample $\mathbf{D}^\star_{\rm DM}$ and $\mathbf{D}^\star_{PS}$. We observe that 90% of pseudo-data containing a DM signal (purple points) are found below the black line, indicating higher power of the empirical Bayesian method than the profile likelihood. Considering data generated by an interloper source (orange points), $\sim 66\%$ of cases are found above the black line, meaning that the empirical Bayes factor fails to reject the null hypothesis more often than the likelihood ratio test. The bold cross markers again show the Asimov data sets for the two scenarios with error bars denoting the nested sampling uncertainty. The distance of each Asimov data point from the diagonal line shows the increase or decrease in median significance we expect from using the empirical Bayes factor vs. the likelihood ratio.

Another representation of this result in shown in Fig. 4, where each panel gives the survival function $S_T(t) = P(T > t)$, i.e., the fraction of the pseudo-data sets with a test statistic T larger than a given value t, for data generated under each of the three scenarios. In the top panel, $T = \lambda_{\rm ML}$, the survival functions for the DM and point source interloper pseudo-data are very similar, meaning that the power (type-II error rate) is approximately the same for these two cases. However, the bottom panel, where $T = T_{\rm BF}$, shows that pseudo-data generated by dark matter annihilation (purple) leads to larger test statistic values (i.e., higher detection significance) than data coming from a point source interloper (orange). A horizontal dotted line [RT :missing!] indicates the median probability, and the point where it crosses the survival function gives the median test statistic – to be compared with the vertical lines, an approximation to the same quantity obtained from a single Asimov data set.

Finally, we conclude this first demonstration with Fig. 5 where we show the fraction of data sets for which the null hypothesis is rejected with a significance level of p_t (Neyman-Pearson testing). In other words, for a given p_t, the figure shows the probability of rejecting the

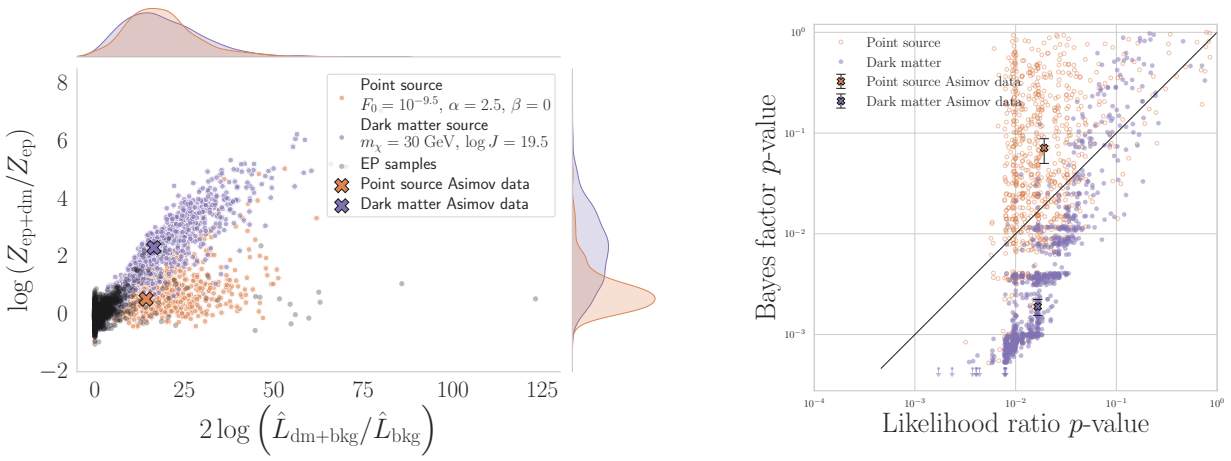

FIG. 3. **Left:** Comparison of the two test statistics $\lambda_{\rm ML}$ (conventional profile maximum likelihood ratio) and $T_{\rm BF}$ (Bayes factor with empirical prior) [RT :update axis labels to match], for simulated data generated under three different scenarios: from the empirical prior background model (black), from a power law point source interloper (orange), and from dark matter annihilation (purple), with benchmarks parameters described in Sec. III E. The thick crosses denote the Asimov data for the two different signal scenarios (see Sec. III D). One-dimensional marginal densities are plotted for each test statistic along the axes. **Right:** Each point corresponds to a simulated data set under either the DM scenario (purple) or the interloper point source scenario (orange). The horizontal axis shows the p-value for rejecting the background-only hypothesis using the conventional profile maximum likelihood ratio test statistic, while the vertical axis is the p-value from the Bayes factor with empirical prior. [RT :make axis labels bigger]

null hypothesis with significance p_t when the data are generated either from dark matter (purple) or an interloper point source (orange). The Asimov p-value is the median p-value for each this signal model: for example, if the target data is generated with a DM source and we use the likelihood ratio test statistic (top panel, purple curve), the Asimov data indicate that there is a 50% probability to reject the null hypothesis at the $p_t = 0.02$ significance. The actual distribution obtained numerically crosses the 50% horizontal probability line almost exactly at the same value. Comparing the point source and dark matter curves in the top panel, we see that they are very close, which reflects once again the inability to distinguish between these two alternatives using the maximum profile likelihood ratio test statistics (for the point source case, there is a small but inconsequential difference between the estimated mean significance via Asimov data and the actual value obtained numerically).

The bottom panel shows the same plot but using $T_{\rm BF}$ as test statistics: now, the median significance for data generated from a point source interloper scenario decreases to about $p \sim 0.07$, while the median significance from dark matter increases to $p \sim 0.002$ (an order of magnitude reduction in p value for the same data). This demonstrates that $T_{\rm BF}$ is less sensitive to interlopers and more sensitive to a DM signal, compared to a traditional likelihood-based test that does not incorporate prior information from the distribution of point sources gleaned from the ROI.

F. Dark matter signal sensitivity forecast

We now present a sensitivity study comparing the empirical Bayes factor sensitivity for a DM signal to that obtained from the conventional maximum profile likelihood ratio test. As above, we conduct this sensitivity study at a fixed reference annihilation cross-section for the DM annihilating to purely $\tau^+\tau^-$ final states as described in Sec. III B. As motivated above, we do not simulated repeated pseudo-data for each $\boldsymbol{\theta}_{\rm DM}$ parameter value, but instead evaluate both $T_{\rm BF}$ and $\lambda_{\rm ML}$ at the Asimov data for a given DM signal injection with m_χ and J factors, and thus obtain an approximate median sensitivity (validated by the example above). This is done on a 20×25 log-spaced grid $m_\chi \in [1, 10^4]$ and uniformly on $\log_{10} J \in [18, 21]$.

As mentioned in Sec.III B we, conduct the sensitivity study in the m_χ-J plane (for a fixed $\langle\sigma v\rangle = 3 \times 10^{-26}$ cm^2/s) in order to investigate the performance of the empirical prior as a function of the J-factor. Of course, given that $\langle\sigma v\rangle$ and J enter linearly into the annihilation flux spectrum in Eqn. (13), our results can be re-scaled for different values of the annihilation cross-section.

The results of this sensitivity study are given in Fig. 6. The colour scale in the first two panels shows the median detection significance $p_{\rm med}$ to reject the null hypothesis calculated for the two test statistics, with the black and red contours corresponding to a median detection sig-

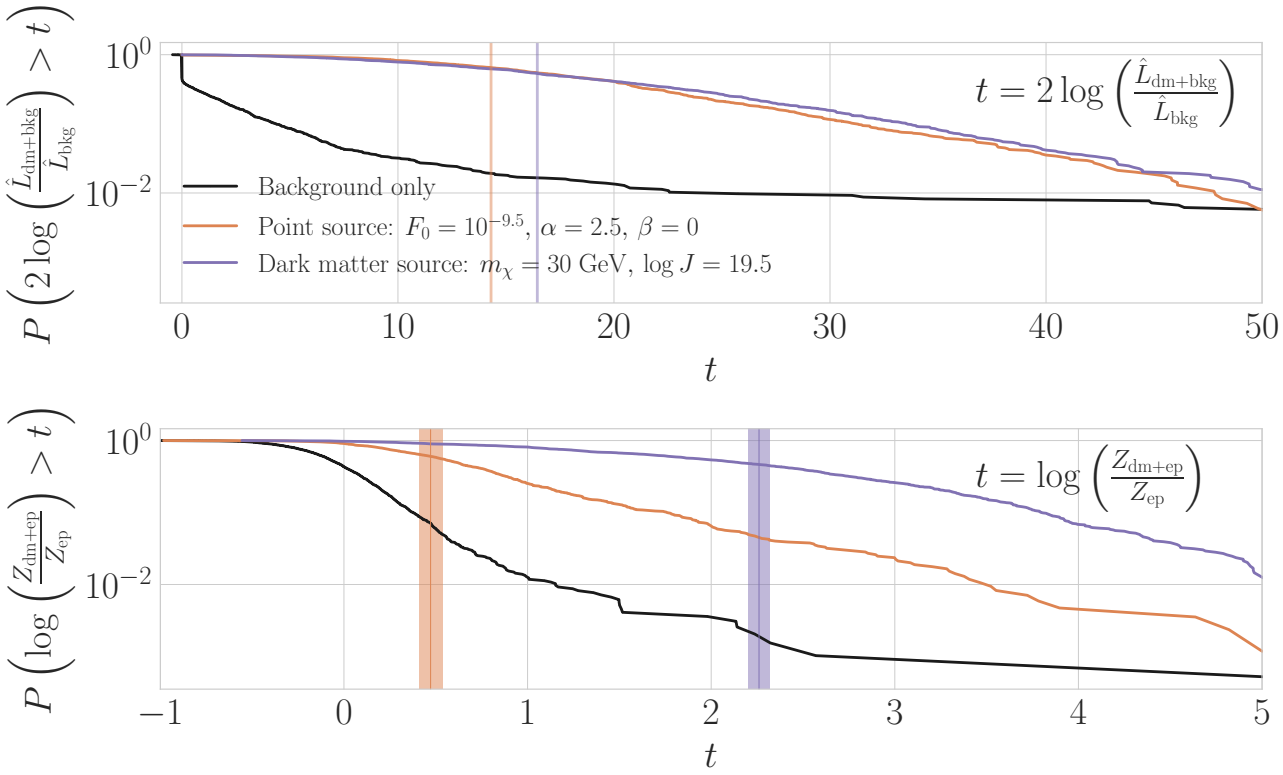

FIG. 4. Survival function, i.e the proportion of the simulated data sets with test statistic larger than a given value t for the profile maximum likelihood ratio test (top) and Bayes factor with empirical prior (bottom). The black line is for 'background only' samples drawn from the empirical prior. The vertical lines denote the value of the test statistics corresponding to the median as obtained from Asimov data, with coloured bands in the bottom panel denoting nested sampling uncertainty in Z. [RT :horizontal line missing; update test statistics notation to be consistent]

nificance[9] of 3σ: each contour marks the boundary between a median detection significance greater and smaller than 3σ. This means that, for dark matter models with J value above the contour, there is a greater than 50% probability of making a discovery at significance of 3σ. All regions that correspond to a median significance $> 5\sigma$ are coloured yellow. This choice of cutoff is in line with the common standard in the high-energy physics community, that considers 5σ the threshold for the claim of a 'discovery' [92–95]. More practically, however, the lack of statistics in the tails of the null-hypothesis distributions of the test statistics make formally calculating p-values at such extreme Asimov data values potentially inaccurate. However, past a certain threshold the exact value of the p-value no longer matters.

The right panel compares the performance of the two test statistics by demonstrating that the 3σ contour calculated with $T_{\rm BF}$ is lowered by a factor of ~ 4 in J: the use of the empirical prior, coupled with the Bayes factor test statistics, therefore leads to enhanced sensitivity to weaker dark matter signals – a consequence of the fact that $T_{\rm BF}$ incorporates information (neglected in the conventional profile likelihood) about the data-generating model coming from the EP. We additionally observe a more rapid increase of sensitivity from $\sim 3\sigma$ in the empirical Bayesian result, as opposed to the likelihood ratio which more slowly transitions from $\sim 2\sigma$ to $\sim 5\sigma$.

[RT :more caveats here?]

IV. CONCLUSION

In this paper, we introduced a novel, hybrid Frequentist-Bayesian methodology for dark matter searches in gamma-ray data that addresses the limitations of conventional statistical testing, particularly in the presence of significant background model uncertainties. Our approach employs an empirical prior derived from sampling nearby background-only regions, which effectively incorporates the characteristics of unresolved faint sources emission components into the hypothesis testing framework, which are usually neglected when only

[9] For the sake of connecting with usual practice, the p-values have been converted into "sigma units" of a Gaussian distribution by using the CDP Φ_N of a standard normal, i.e. detection significance $= \Phi_N^{-1}(1 - p_{\rm med})$.

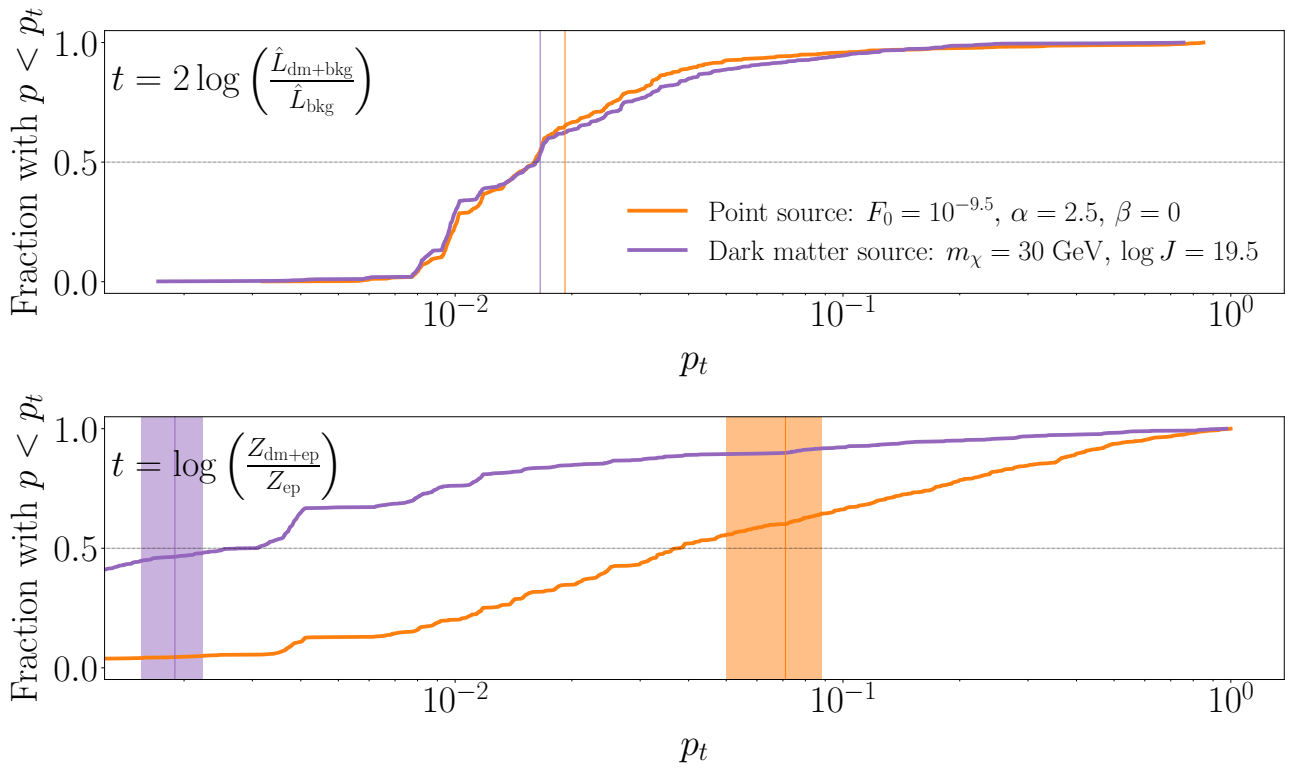

FIG. 5. As in Fig. 5, but displaying the proportion of pseudo-data sets that lead to a rejection of the null with a Neyman-Pearson significance of p_t. The vertical lines represent the Asimov data realisations with the bands on the bottom panel indicating sampling uncertainty in Z.

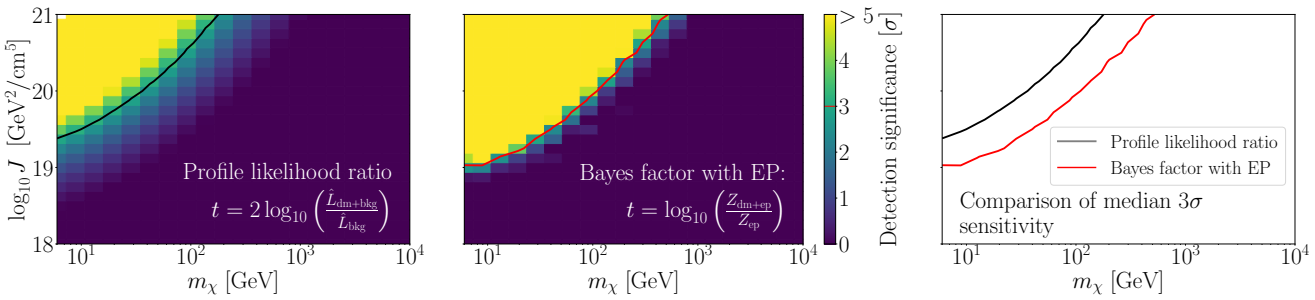

FIG. 6. Median detection significance (in units of Gaussian σ) as a function of dark matter mass and J value, for a fixed thermal relic cross section, for the conventional maximum profile likelihood ratio test statistic (left) and the empirical Bayes factor test statistic (middle). The far right panel demonstrates the increase in sensitivity by comparing the respective 3σ median sensitivity contours. [RT :Make lines a bit thicker in left and central panel; change right panel to show the improvement factor instead (linear scale on the y-axis, and plot only 1 curve); make sure to add y-axis ticks to this panel]

considering a diffuse background and isotropic contribution.

As a first test study, we demonstrated the efficacy of our methodology by comparing the performance of the new test statistic, based on a Bayes factor that utilizes the empirical prior, with the traditional maximum profile likelihood ratio statistic. Through simulations, we showed that our method exhibits superior power in detecting dark matter signals while simultaneously reducing the false positive rate associated with point source interlopers. The empirical prior, constructed from actual `Fermi` data, proved to be a robust tool to describe the underlying background processes, enhancing the sensitivity of our detection framework to annihilating DM.

It is important to note that a misspecification of the background model has two potential and very different

impacts on our results. While an imperfect background model in the test statistic $T_{\rm BF}$ can lead to reduced sensitivity, an incorrect background model used to derive its sampling distribution would result in inaccurate p-values. To ensure that the p-value is correct, the null distribution should be obtained by sampling random regions of the sky, computing the test statistic $T_{\rm BF}$ in each region, and generating a histogram. Instead, in the current analysis, the null sampling distribution is generated by assuming the empirical prior model to be correct. Ideally, a larger region of the sky (e.g., a 40×20 degrees region) should be used, with half the region dedicated to building the empirical prior and the other half to determining the sampling distribution of the test statistic. We leave the exploration of this more refined method to future work.

We then conducted a sensitivity study spanning a range of dark matter masses and J-factors, which highlighted the significant improvements offered by our empirical Bayesian method, which resolves the problem of model misspecification that arises when using a likelihood ratio test solely including template-based isotropic and diffuse components. We found that the empirical Bayes factor test statistic consistently achieves higher detection significances, allowing for the identification of weaker dark matter signals compared to conventional methods. This is particularly relevant for searches in regions with high uncertainties in the background model, such as the galactic center or dwarf spheroidal galaxies.

The methodology presented here is not only applicable to simulated data but is also readily extendable to real gamma-ray observations. The flexibility of the empirical prior makes it a valuable option for various astrophysical contexts where background characterization poses a challenge. Whilst beyond the scope of this methods paper, we propose future work involving a real analysis of Carina I and Carina II, where the spatial (approximately half a degree) separation of these objects make an explicit likelihood based determination of the background processes difficult to model. Constructing an empirical prior in this case could prove useful in conservatively incorporating an approximation of all background components from these two faint dwarfs.

While promising, our empirical prior method has potential shortcomings. Its effectiveness is highly dependent on the quality and representativeness of the sub-areas A_i used as the corpus. If these sub-areas do not accurately reflect the broader background processes, the prior may be biased. This issue is related to the 'honest corpus' discussed in Ref. [96].

The method also assumes that background processes in the sub-areas are identical to those in the target region, which may not hold if there are spatial or temporal variations. Additionally, the fixed empirical prior might not adapt well to new or unforeseen background variations, and may be sensitive to variability in the sampled sub-areas. These limitations suggest the need for enhancing corpus representativeness, increasing the number of sub-areas sampled, incorporating adaptive elements into the prior, and continuously validating it against new data. Additionally, it should be noted that the availability of an empirical prior does not preclude using more specific and/or auxiliary prior information where available.

In conclusion, our approach offers a promising avenue for enhancing the robustness and sensitivity of dark matter searches in gamma-ray data. By leveraging the empirical prior, we can more accurately discern potential dark matter signals amidst complex and uncertain backgrounds.

Appendix A: The meaning of the empirical prior

This section builds up a qualitative intuition for the role of the empirical prior. We begin with a brief discussion regarding model misspecification. That is, the imperfect modeling of representative stochastic nature of the data. We expect the empirical prior to be most useful when it is close to the "true" distribution representing the actual distribution of the background. While it is true that, in general, both the prior and the likelihood are approximations, misspecification is more likely to lie with the prior than the likelihood. This is because the likelihood, in the case of gamma-ray data, is a straightforward convolution of a Poisson process – which we are nearly sure is the correct distribution – with an instrument response. In order to explore misspecification in the context of the methodology presented in this paper, we consider the expected value of the empirical prior:

$$\mathrm{E}_{\boldsymbol{D}^{\rm bg}}\left[P(\phi^* \mid \boldsymbol{D}^{\rm bg})\right] = \sum_{\boldsymbol{D}^{\rm bg}} P(\phi^* \mid \boldsymbol{D}^{\rm bg}) \cdot P(\boldsymbol{D}^{\rm bg}) \,, \tag{A1}$$

that is, the empirical prior we would obtain if we had an infinite number of background samples $\boldsymbol{D}_i^{\rm bg}$. The expectation with respect to the sample background data $\boldsymbol{D}^{\rm bg}$ is [RT :Why is this different from the above equation? Simply replaced $P(\boldsymbol{D}^{\rm bg})$ with $\pi_{\rm tr}$, no?],

$$\mathrm{E}_{\boldsymbol{D}^{\rm bg}} = \sum_{\boldsymbol{D}^{\rm bg}} P(\phi^* \mid \boldsymbol{D}^{\rm bg}) \int d\phi' P_{\rm tr}(\boldsymbol{D}^{\rm bg} \mid \phi') \pi_{\rm tr}(\phi') \tag{A2}$$

$$= \sum_{\boldsymbol{D}^{\rm bg}} P(\phi^* \mid \boldsymbol{D}^{\rm bg}) \cdot Z_{\rm tr}(\boldsymbol{D}^{\rm bg}) \,, \tag{A3}$$

where $\pi_{\rm tr}$ is the true distribution of background model parameters over the sky (using a fully complete parameterization ϕ^* of the background model [RT :what if this is not true? This is also part of what constitutes model misspecification...]), and $P_{\rm tr}$ represents the true likelihood connecting the true background in a region with the data that would be observed. In other words, the sum in the expectation is over all possible states of the background ϕ^* and over all possible data sets that could arise from each state. Note that $P(\boldsymbol{D}^{\rm bg})$, the inner integral over ϕ',

is just the evidence for the true background model, that we can denote $Z_{\rm tr}(\boldsymbol{D}^{\rm bg})$.

Following on, and now explicitly inserting the data index i over which the expectation is summed, the expected empirical prior based on our approximate model is then evaluated by substituting the explicit expression for the empirical prior from Eqn. (3):

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{D}^{\rm bg}}\left[P(\phi^* \mid \boldsymbol{D}_i^{\rm bg})\right] &= \sum_{\boldsymbol{D}^{\rm bg}} P(\phi^* \mid \boldsymbol{D}_i^{\rm bg}) Z_{\rm tr}(\boldsymbol{D}^{\rm bg}) \\
&= \sum_{\boldsymbol{D}_i^{\rm bg}} \left(\frac{P(\boldsymbol{D}_i^{\rm bg} \mid \phi^*)\pi_{\rm ur}(\phi^*)}{Z_i(\boldsymbol{D}_i^{\rm bg})}\right) Z_{\rm tr}(\boldsymbol{D}^{\rm bg}) \\
&= \mathrm{E}_{P(\boldsymbol{D}^{\rm bg}|\boldsymbol{\phi}^\star)}\left[\frac{Z_{\rm tr}(\boldsymbol{D}^{\rm bg})}{Z(\boldsymbol{D}^{\rm bg})} \,\middle|\, \phi^*\right] \pi_{\rm ur}(\phi^*).
\end{aligned} \tag{A4}$$

The expectation in the last line is with respect to our adopted likelihood function $P(\boldsymbol{D}^{\rm bg} \mid \boldsymbol{\phi}^\star)$. Therefore, the expected empirical prior is our ur-prior $\pi_{\rm ur}(\boldsymbol{\phi}^\star)$ scaled by the expected evidence ratio between the true model and our model for a dataset generated under our model with fixed parameters $\boldsymbol{\phi}^\star$ [RT :But in principle the true model could have a different set of parameters, as well, no?].

First, notice that if our background model was complete and correct, so that $\pi_{\rm ur}(\boldsymbol{\phi}^\star) = \pi_{\rm tr}(\boldsymbol{\phi}^\star)$ and $P(\boldsymbol{D}^{\rm bg} \mid \boldsymbol{\phi}^\star) = P_{\rm tr}(\boldsymbol{D}^{\rm bg} \mid \boldsymbol{\phi}^\star)$, then the expected empirical prior would be the true background distribution $\pi_{\rm tr}(\boldsymbol{\phi}^\star)$. This situation is useless of course — if we already knew the true distribution of parameters over the true background model we wouldn't need to try to construct the empirical prior in the first place. [RT :yes, but again this assumes that ϕ^* is the same set for both models, which is not necessarily true and a strong assuption]

However, Eq. (A4) suggests that the empirical prior should be at least *closer* to the truth than our initial prior $\pi_{\rm ur}(\boldsymbol{\phi}^\star)$. If $\pi_{\rm ur}(\boldsymbol{\phi}^\star)$ is "too large" (more bias toward the model [RT :what does that mean?]) at a particular value of $\boldsymbol{\phi}^\star$, then the evidence ratio for a data set generated by our model $P(\boldsymbol{D}^{\rm bg} \mid \boldsymbol{\phi}^\star)$ will favor model rather than the true model, [RT :thus reducing the density at that value?].

Likewise, if $\pi_{\rm ur}(\boldsymbol{\phi}^\star)$ is "too small" (less bias toward the model [RT :unclear]) at a particular $\boldsymbol{\phi}^\star$, then we expect the evidence ratio to favor the true background model, for data generated using $\boldsymbol{\phi}^\star$. In this case the expected empirical prior will be larger than $\pi_{\rm ur}(\boldsymbol{\phi}^\star)$. So as a result, in both situations the empirical prior is "closer to what it should be", in some qualitative sense.

We can assert that the empirical prior $\pi_{\rm ep}$ can be thought of as a 'fuzzy' approximation of some underlying 'true' distributions which encapsulates the underlying background distribution in a given set of background regions. The background regions for which the empirical prior is derived can therefore be analogously thought of as the 'corpus' of studies proposed in Refs. [96, 97] that serves to create a slightly more uninformative [RT :why un-informative?], but certainly not diffuse prior.

Consider again Eqn. (A4). If each of the sampled background areas i were observed for an infinite amount of time, in virtue of the central limit theorem the likelihood $P(\boldsymbol{D}_i^{\rm bg} \mid \phi^*)$ would overwhelm the prior $\pi_{\rm ur}(\boldsymbol{\phi}^\star)$ and tend toward a delta function at the 'true' values of $\boldsymbol{\phi}^\star$ for that small region of sky[10]. With enough sampled regions and long enough exposures the empirical prior would simply be the true background model. With finite data, each delta function is smeared into a distribution which represents our realistic understanding of the background in region i. The empirical prior is then a mixture of these distributions, and is similar to a kernel density estimate of the true $\boldsymbol{\phi}^\star$ distribution. Longer exposures will lower the "bandwidth" of the kernel, i.e. reduce the uncertainty on $\boldsymbol{\phi}^\star$ for individual background samples.

The mixture in Eq. (A4) is also limited by the finite number of background samples. This finite n means that the empirical prior will not incorporate sufficiently rare processes. The first limitation is addressed by longer exposure, the second by a larger sky area[11] over which to sample the background. Given an existing data set, the empirical prior represents our best understanding of the properties of the background.

[10] The background processes in each small sky region are governed by a random process. I.e. there is some probability that there exist particular faint sources along any given line of sight and that the diffuse processes in the Milky Way have some amplitude. This randomness can be quantified by the "true" distribution of parameters $\boldsymbol{\phi}^\star$, $\pi(\boldsymbol{\phi}^\star)$. This distribution could in principle be constructed from commonly studied quantities like the source count distribution dN/dS (density of sources per flux level) as well as the distribution of spectral properties for different classes of sources, but we are not concerned with that here, as we adopt a more agnostic approach via the assertion of the empirical prior.

[11] Although this needs to be tensioned against the requirement to limit the ROI to a 'representative' region around the target, lest it deviates from what can be expected at the target.

[1] M. Masias, J. Freixenet, X. Lladó, and M. Peracaula, A review of source detection approaches in astronomical images, **422**, 1674 (2012).

[2] T. P. Li and Y. Q. Ma, Analysis methods for results in gamma-ray astronomy., Astrophys. J. **272**, 317 (1983).

[3] L. Szentmiklósi, Z. Révay, T. Belgya, A. Simonits, and Z. Kis, Combining prompt gamma activation analysis

and off-line counting, Journal of Radioanalytical and Nuclear Chemistry **278**, 657 (2008).
[4] B. Huber, C. Farnier, A. Manalaysay, U. Straumann, and R. Walter, A stacking method to study the gamma-ray emission of source samples based on the co-adding of fermi-lat count maps, Astronomy and Astrophysics **547** (2012).
[5] C. Consortium, *Science with the Cherenkov Telescope Array* (World Scientific Publishing, 2019).
[6] C. Deil, R. Zanin, J. Lefaucheur, C. Boisson, *et al.*, Data formats for gamma-ray astronomy – version 0.3, in *Analysis Methods for Gamma-Ray Astronomy* (SpringerLink, 2022).
[7] E. Fenimore and T. Cannon, Coded mask instruments for gamma-ray astronomy, in *Analysis Methods for Gamma-Ray Astronomy* (SpringerLink, 1978).
[8] M. D. Mauro, The origin of the fermi-lat γ-ray background (2016), arXiv:1601.04323 [astro-ph.HE].
[9] J. Schmitt, J. Starck, J. Casandjian, J. Fadili, and I. Grenier, Poisson denoising on the sphere: application to the fermi gamma ray space telescope, Astronomy and Astrophysics **517**, 10.1051/0004-6361/200913822 (2010).
[10] F. Guglielmetti, R. Fischer, and V. Dose, Bayesian mixture models for poisson astronomical images, in *Springer Series in Astrostatistics*, Vol. 902 (2012) pp. 197–202.
[11] J. Tucker, Evaluation of two methods for incorporating a systematic uncertainty into a test of the background-only hypothesis for a poisson process, in *Proceedings of the Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology* (2008) pp. 40–43.
[12] D. Hearn, Consistent analysis of gamma-ray astronomy experiments, Nuclear Instruments and Methods **70**, 200 (1969).
[13] W. Wheaton, A. L. Dunklee, A. S. Jacobsen, J. Ling, W. Mahoney, and R. Radocinski, Multiparameter linear least-squares fitting to poisson data one count at a time, The Astrophysical Journal **438**, 322 (1995).
[14] Rowell, G. P., A new template background estimate for source searching in tev -ray astronomy, AA **410**, 389 (2003).
[15] A. Cuoco, A. Sellerholm, J. Conrad, and S. Hannestad, Anisotropies in the diffuse gamma-ray background from dark matter with fermi lat: a closer look, Monthly Notices of the Royal Astronomical Society **414**, 2040 (2010).
[16] M. Di Mauro and F. Donato, Composition of the fermi-lat isotropic gamma-ray background intensity: Emission from extragalactic point sources and dark matter annihilations, Physical Review D **91**, 123001 (2015).
[17] B. Hensley, V. Pavlidou, and J. Siegal-Gaskins, Novel techniques for decomposing diffuse backgrounds, Monthly Notices of the Royal Astronomical Society **433**, 591 (2012).
[18] M. Fornasa, J. Zavala, M. Sánchez-Conde, F. Prada, M. Vogelsberger, and C. Frenk, Dark matter implications of the fermi-lat measurement of anisotropies in the diffuse gamma-ray background: status report, Nuclear Instruments Methods in Physics Research Section A-accelerators Spectrometers Detectors and Associated Equipment **692**, 132 (2011).
[19] A. Abdo, M. Ackermann, M. Ajello, W. Atwood, L. Baldini, J. Ballet, G. Barbiellini, D. Bastieri, B. Baughman, K. Bechtol, *et al.*, Spectrum of the isotropic diffuse gamma-ray emission derived from first-year fermi large area telescope data, Physical review letters **104 10**, 101101 (2010).
[20] B. Zhou, Y.-F. Liang, X. Huang, X. Li, Y. Fan, L. Feng, and J. Chang, Gev excess in the milky way: The role of diffuse galactic gamma-ray emission templates, Physical Review D **91**, 123010 (2014).
[21] T. F. C. A. Abdo, M. Ackermann, M. Ajello, L. Baldini, J. Ballet, G. Barbiellini, D. Bastieri, K. Bechtol, R. Bellazzini, B. Berenji, *et al.*, Constraints on cosmological dark matter annihilation from the fermi-lat isotropic diffuse gamma-ray measurement, Journal of Cosmology and Astroparticle Physics **2010**, 014.
[22] W. B. Atwood, A. A. Abdo, M. Ackermann, W. Althouse, B. Anderson, M. Axelsson, L. Baldini, J. Ballet, D. L. Band, G. Barbiellini, J. Bartelt, D. Bastieri, B. M. Baughman, K. Bechtol, D. Bédérède, F. Bellardi, R. Bellazzini, B. Berenji, G. F. Bignami, D. Bisello, E. Bissaldi, R. D. Blandford, E. D. Bloom, J. R. Bogart, E. Bonamente, J. Bonnell, A. W. Borgland, A. Bouvier, J. Bregeon, A. Brez, M. Brigida, P. Bruel, T. H. Burnett, G. Busetto, G. A. Caliandro, R. A. Cameron, P. A. Caraveo, S. Carius, P. Carlson, J. M. Casandjian, E. Cavazzuti, M. Ceccanti, C. Cecchi, E. Charles, A. Chekhtman, C. C. Cheung, J. Chiang, R. Chipaux, A. N. Cillis, S. Ciprini, R. Claus, J. Cohen-Tanugi, S. Condamoor, J. Conrad, R. Corbet, L. Corucci, L. Costamante, S. Cutini, D. S. Davis, D. Decotigny, M. DeKlotz, C. D. Dermer, A. de Angelis, S. W. Digel, E. do Couto e Silva, P. S. Drell, R. Dubois, D. Dumora, Y. Edmonds, D. Fabiani, C. Farnier, C. Favuzzi, D. L. Flath, P. Fleury, W. B. Focke, S. Funk, P. Fusco, F. Gargano, D. Gasparrini, N. Gehrels, F. X. Gentit, S. Germani, B. Giebels, N. Giglietto, P. Giommi, F. Giordano, T. Glanzman, G. Godfrey, I. A. Grenier, M. H. Grondin, J. E. Grove, L. Guillemot, S. Guiriec, G. Haller, A. K. Harding, P. A. Hart, E. Hays, S. E. Healey, M. Hirayama, L. Hjalmarsdotter, R. Horn, R. E. Hughes, G. Jóhannesson, G. Johansson, A. S. Johnson, R. P. Johnson, T. J. Johnson, W. N. Johnson, T. Kamae, H. Katagiri, J. Kataoka, A. Kavelaars, N. Kawai, H. Kelly, M. Kerr, W. Klamra, J. Knödlseder, M. L. Kocian, N. Komin, F. Kuehn, M. Kuss, D. Landriu, L. Latronico, B. Lee, S. H. Lee, M. Lemoine-Goumard, A. M. Lionetto, F. Longo, F. Loparco, B. Lott, M. N. Lovellette, P. Lubrano, G. M. Madejski, A. Makeev, B. Marangelli, M. M. Massai, M. N. Mazziotta, J. E. McEnery, N. Menon, C. Meurer, P. F. Michelson, M. Minuti, N. Mirizzi, W. Mitthumsiri, T. Mizuno, A. A. Moiseev, C. Monte, M. E. Monzani, E. Moretti, A. Morselli, I. V. Moskalenko, S. Murgia, T. Nakamori, S. Nishino, P. L. Nolan, J. P. Norris, E. Nuss, M. Ohno, T. Ohsugi, N. Omodei, E. Orlando, J. F. Ormes, A. Paccagnella, D. Paneque, J. H. Panetta, D. Parent, M. Pearce, M. Pepe, A. Perazzo, M. Pesce-Rollins, P. Picozza, L. Pieri, M. Pinchera, F. Piron, T. A. Porter, L. Poupard, S. Rainò, R. Rando, E. Rapposelli, M. Razzano, A. Reimer, O. Reimer, T. Reposeur, L. C. Reyes, S. Ritz, L. S. Rochester, A. Y. Rodriguez, R. W. Romani, M. Roth, J. J. Russell, F. Ryde, S. Sabatini, H. F. W. Sadrozinski, D. Sanchez, A. Sander, L. Sapozhnikov, P. M. S. Parkinson, J. D. Scargle, T. L. Schalk, G. Scolieri, C. Sgrò, G. H. Share, M. Shaw, T. Shimokawabe, C. Shrader, A. Sierpowska-Bartosik, E. J. Siskind, D. A. Smith, P. D. Smith, G. Spandre, P. Spinelli, J. L. Starck, T. E. Stephens, M. S. Strickman, A. W. Strong, D. J. Suson, H. Tajima, H. Taka-

hashi, T. Takahashi, T. Tanaka, A. Tenze, S. Tether, J. B. Thayer, J. G. Thayer, D. J. Thompson, L. Tibaldo, O. Tibolla, D. F. Torres, G. Tosti, A. Tramacere, M. Turri, T. L. Usher, N. Vilchez, V. Vitale, P. Wang, K. Watters, B. L. Winer, K. S. Wood, T. Ylinen, and M. Ziegler, The Large Area Telescope on the Fermi Gamma-Ray Space Telescope Mission, Astrophys. J. **697**, 1071 (2009), arXiv:0902.1089 [astro-ph.IM].

[23] S. Hoof, A. Geringer-Sameth, and R. Trotta, A Global Analysis of Dark Matter Signals from 27 Dwarf Spheroidal Galaxies using 11 Years of Fermi-LAT Observations, JCAP **02**, 012, arXiv:1812.06986 [astro-ph.CO].

[24] A. McDaniel, M. Ajello, C. M. Karwin, M. Di Mauro, A. Drlica-Wagner, and M. A. Sánchez-Conde, Legacy analysis of dark matter annihilation from the Milky Way dwarf spheroidal galaxies with 14 years of Fermi-LAT data, Phys. Rev. D **109**, 063024 (2024), arXiv:2311.04982 [astro-ph.HE].

[25] D. Kerszberg (Fermi-LAT, HAWC, H.E.S.S., MAGIC, VERITAS), Search for dark matter annihilation with a combined analysis of dwarf spheroidal galaxies from Fermi-LAT, HAWC, H.E.S.S., MAGIC and VERITAS, PoS **ICRC2023**, 1426 (2023).

[26] M. Di Mauro, M. Stref, and F. Calore, Investigating the effect of Milky Way dwarf spheroidal galaxies extension on dark matter searches with Fermi-LAT data, Phys. Rev. D **106**, 123032 (2022), arXiv:2212.06850 [astro-ph.HE].

[27] S. Li, Y.-F. Liang, and Y.-Z. Fan, Search for gamma-ray emission from the 12 nearby dwarf spheroidal galaxies with 12 years of Fermi-LAT data, Phys. Rev. D **104**, 083037 (2021), arXiv:2110.01157 [astro-ph.HE].

[28] V. Gammaldi, J. Pérez-Romero, J. Coronado-Blázquez, M. Di Mauro, E. Karukes, M. A. Sánchez-Conde, and P. Salucci. (Fermi-LAT), Dark Matter search in dwarf irregular galaxies with the Fermi Large Area Telescope, PoS **ICRC2021**, 509 (2021), arXiv:2109.11291 [astro-ph.CO].

[29] M. Ackermann *et al.* (Fermi-LAT), The Fermi Galactic Center GeV Excess and Implications for Dark Matter, Astrophys. J. **840**, 43 (2017), arXiv:1704.03910 [astro-ph.HE].

[30] R. Bartels, S. Krishnamurthy, and C. Weniger, Strong support for the millisecond pulsar origin of the Galactic center GeV excess, Phys. Rev. Lett. **116**, 051102 (2016), arXiv:1506.05104 [astro-ph.HE].

[31] J. Petrović, P. D. Serpico, and G. Zaharijaš, Galactic Center gamma-ray "excess" from an active past of the Galactic Centre?, JCAP **10**, 052, arXiv:1405.7928 [astro-ph.HE].

[32] M. Di Mauro, Characteristics of the Galactic Center excess measured with 11 years of *Fermi*-LAT data, Phys. Rev. D **103**, 063029 (2021), arXiv:2101.04694 [astro-ph.HE].

[33] R. K. Leane and T. R. Slatyer, The enigmatic Galactic Center excess: Spurious point sources and signal mismodeling, Phys. Rev. D **102**, 063019 (2020), arXiv:2002.12371 [astro-ph.HE].

[34] P. Bruel, Extending the event-weighted pulsation search to very faint gamma-ray sources, Astronomy Astrophysics 10.1051/0004-6361/201834555 (2018).

[35] M. Ackermann, A. Albert, B. Anderson, W. B. Atwood, L. Baldini, G. Barbiellini, D. Bastieri, K. Bechtol, R. Bellazzini, E. Bissaldi, R. D. Blandford, E. D. Bloom, R. Bonino, E. Bottacini, T. J. Brandt, J. Bregeon, P. Bruel, R. Buehler, G. A. Caliandro, R. A. Cameron, R. Caputo, M. Caragiulo, P. A. Caraveo, C. Cecchi, E. Charles, A. Chekhtman, J. Chiang, G. Chiaro, S. Ciprini, R. Claus, J. Cohen-Tanugi, J. Conrad, A. Cuoco, S. Cutini, F. D'Ammando, A. de Angelis, F. de Palma, R. Desiante, S. W. Digel, L. Di Venere, P. S. Drell, A. Drlica-Wagner, R. Essig, C. Favuzzi, S. J. Fegan, E. C. Ferrara, W. B. Focke, A. Franckowiak, Y. Fukazawa, S. Funk, P. Fusco, F. Gargano, D. Gasparrini, N. Giglietto, F. Giordano, M. Giroletti, T. Glanzman, G. Godfrey, G. A. Gomez-Vargas, I. A. Grenier, S. Guiriec, M. Gustafsson, E. Hays, J. W. Hewitt, D. Horan, T. Jogler, G. Jóhannesson, M. Kuss, S. Larsson, L. Latronico, J. Li, L. Li, M. Llena Garde, F. Longo, F. Loparco, P. Lubrano, D. Malyshev, M. Mayer, M. N. Mazziotta, J. E. McEnery, M. Meyer, P. F. Michelson, T. Mizuno, A. A. Moiseev, M. E. Monzani, A. Morselli, S. Murgia, E. Nuss, T. Ohsugi, M. Orienti, E. Orlando, J. F. Ormes, D. Paneque, J. S. Perkins, M. Pesce-Rollins, F. Piron, G. Pivato, T. A. Porter, S. Rainò, R. Rando, M. Razzano, A. Reimer, O. Reimer, S. Ritz, M. Sánchez-Conde, A. Schulz, N. Sehgal, C. Sgrò, E. J. Siskind, F. Spada, G. Spandre, P. Spinelli, L. Strigari, H. Tajima, H. Takahashi, J. B. Thayer, L. Tibaldo, D. F. Torres, E. Troja, G. Vianello, M. Werner, B. L. Winer, K. S. Wood, M. Wood, G. Zaharijas, S. Zimmer, and Fermi-LAT Collaboration, Searching for Dark Matter Annihilation from Milky Way Dwarf Spheroidal Galaxies with Six Years of Fermi Large Area Telescope Data, Phys. Rev. Lett. **115**, 231301 (2015), arXiv:1503.02641 [astro-ph.HE].

[36] A. Geringer-Sameth, S. M. Koushiappas, and M. G. Walker, Comprehensive search for dark matter annihilation in dwarf galaxies, Phys. Rev. D **91**, 083535 (2015), arXiv:1410.2242 [astro-ph.CO].

[37] E. Carlson, D. Hooper, and T. Linden, Improving the sensitivity of gamma-ray telescopes to dark matter annihilation in dwarf spheroidal galaxies, Phys. Rev. D **91**, 061302 (2015), arXiv:1409.1572 [astro-ph.HE].

[38] T. Daylan, S. Portillo, and D. Finkbeiner, Inference of unresolved point sources at high galactic latitudes using probabilistic catalogs, The Astrophysical Journal **839**, 10.3847/1538-4357/aa679e (2016).

[39] V. S. Paliya, A. Domínguez, M. Ajello, A. Franckowiak, and D. Hartmann, Fermi-lat stacking analysis technique: An application to extreme blazars and prospects for their cta detection, The Astrophysical Journal Letters **882**, L3 (2019).

[40] Y. Song, T. A. D. Paglione, J. Tan, C. Lee-Georgescu, and D. Herrera, A stacking survey of gamma-ray pulsars, Monthly Notices of the Royal Astronomical Society **524**, 5854–5868 (2023).

[41] F. Calore, P. D. Serpico, and B. Zaldivar, Dark matter constraints from dwarf galaxies: a data-driven analysis, JCAP **10**, 029, arXiv:1803.05508 [astro-ph.HE].

[42] T. Linden, Robust method for treating astrophysical mismodeling in dark matter annihilation searches of dwarf spheroidal galaxies, Physical Review D **101**, 10.1103/physrevd.101.043017 (2020).

[43] T. Linden, Star-forming galaxies significantly contribute to the isotropic gamma-ray background, Physical Review D **96**, 10.1103/physrevd.96.083001 (2017).

[44] A. Geringer-Sameth and S. M. Koushiappas, Exclusion of

Canonical Weakly Interacting Massive Particles by Joint Analysis of Milky Way Dwarf Galaxies with Data from the Fermi Gamma-Ray Space Telescope, Phys. Rev. Lett. **107**, 241303 (2011).

[45] M. Ackermann, A. Albert, B. Anderson, L. Baldini, J. Ballet, G. Barbiellini, D. Bastieri, K. Bechtol, R. Bellazzini, E. Bissaldi, E. D. Bloom, E. Bonamente, A. Bouvier, T. J. Brandt, J. Bregeon, M. Brigida, P. Bruel, R. Buehler, S. Buson, G. A. Caliandro, R. A. Cameron, M. Caragiulo, P. A. Caraveo, C. Cecchi, E. Charles, A. Chekhtman, J. Chiang, S. Ciprini, R. Claus, J. Cohen-Tanugi, J. Conrad, F. D'Ammando, A. de Angelis, C. D. Dermer, S. W. Digel, E. do Couto e Silva, P. S. Drell, A. Drlica-Wagner, R. Essig, C. Favuzzi, E. C. Ferrara, A. Franckowiak, Y. Fukazawa, S. Funk, P. Fusco, F. Gargano, D. Gasparrini, N. Giglietto, M. Giroletti, G. Godfrey, G. A. Gomez-Vargas, I. A. Grenier, S. Guiriec, M. Gustafsson, M. Hayashida, E. Hays, J. Hewitt, R. E. Hughes, T. Jogler, T. Kamae, J. Knödlseder, D. Kocevski, M. Kuss, S. Larsson, L. Latronico, M. Llena Garde, F. Longo, F. Loparco, M. N. Lovellette, P. Lubrano, G. Martinez, M. Mayer, M. N. Mazziotta, P. F. Michelson, W. Mitthumsiri, T. Mizuno, A. A. Moiseev, M. E. Monzani, A. Morselli, I. V. Moskalenko, S. Murgia, R. Nemmen, E. Nuss, T. Ohsugi, E. Orlando, J. F. Ormes, J. S. Perkins, F. Piron, G. Pivato, T. A. Porter, S. Rainò, R. Rando, M. Razzano, S. Razzaque, A. Reimer, O. Reimer, S. Ritz, M. Sánchez-Conde, N. Sehgal, C. Sgrò, E. J. Siskind, P. Spinelli, L. Strigari, D. J. Suson, H. Tajima, H. Takahashi, J. B. Thayer, L. Tibaldo, M. Tinivella, D. F. Torres, Y. Uchiyama, T. L. Usher, J. Vandenbroucke, G. Vianello, V. Vitale, M. Werner, B. L. Winer, K. S. Wood, M. Wood, G. Zaharijas, S. Zimmer, and Fermi-LAT Collaboration, Dark matter constraints from observations of 25 MilkyÂ Way satellite galaxies with the Fermi Large Area Telescope, Phys. Rev. D **89**, 042001 (2014), arXiv:1310.0828 [astro-ph.HE].

[46] R. L. Wasserstein and N. A. Lazar, Asa statement on statistical significance and p-values, The American Statistician **70**, 129 (2016).

[47] E. Dunn, Weighing the evidence: On the use and misuse of p-values, The Dunn Lab (2016).

[48] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, The European Physical Journal C **71**, 10.1140/epjc/s10052-011-1554-0 (2011).

[49] H. Jeffreys, *Theory of Probability*, 3rd ed. (Oxford University Press, 1939).

[50] R. E. Kass and A. E. Raftery, Bayes factors, Journal of the American Statistical Association **90**, 773 (1995).

[51] M. D. Lee and E. J. Wagenmakers, *Bayesian Cognitive Modeling: A Practical Course* (Cambridge University Press, 2014).

[52] I. J. Good, Saddle-point methods for the multinomial distribution, The Annals of Mathematical Statistics **28**, 861 (1957).

[53] I. J. Good, The bayes/non-bayes compromise: A brief review, Journal of the American Statistical Association **87**, 597 (1992).

[54] J. O. Berger, B. Boukai, and Y. Wang, Unified frequentist and bayesian testing of a precise hypothesis, Statistical Science **12**, 133 (1997).

[55] J. O. Berger, Could fisher, jeffreys and neyman have agreed on testing?, Statistical Science **18**, 1 (2003).

[56] J. Zhang, Bayesian (mean) most powerful tests, Australian & New Zealand Journal of Statistics **59**, 43 (2017).

[57] A. Fowlie, Neymanâ-pearson lemma for bayes factors, Communications in Statistics - Theory and Methods **52**, 5379 (2023).

[58] W. Atwood, A. Albert, L. Baldini, M. Tinivella, J. Bregeon, M. Pesce-Rollins, C. Sgrò, P. Bruel, E. Charles, A. Drlica-Wagner, A. Franckowiak, T. Jogler, L. Rochester, T. Usher, M. Wood, J. Cohen-Tanugi, and S. Zimmer, Pass 8: Toward the full realization of the fermi-lat scientific potential (2013), arXiv:1303.3514 [astro-ph.IM].

[59] P. Bruel, A new method to perform data-model comparison in fermi-lat analysis, Astronomy amp; Astrophysics **656**, A81 (2021).

[60] M. Ajello *et al.* (Fermi-LAT), 3FHL: The Third Catalog of Hard Fermi-LAT Sources, Astrophys. J. Suppl. **232**, 18 (2017), arXiv:1702.00664 [astro-ph.HE].

[61] F.-L. Collaboration, Fermi large area telescope fourth source catalog data release 4 (4fgl-dr4), arXiv: Instrumentation and Methods for Astrophysics (2024).

[62] N. Hjort, F. A. Dahl, and G. H. Steinbakk, Post-processing posterior predictive p values, Journal of the American Statistical Association **101**, 1157 (2006).

[63] S. Sinharay and H. Stern, Posterior predictive model checking in hierarchical models, Journal of Statistical Planning and Inference **111**, 209 (2003).

[64] R. Protassov, D. V. van Dyk, A. Connors, V. Kashyap, and A. Siemiginowska, Statistics, handle with care: Detecting multiple model components with the likelihood ratio test, The Astrophysical Journal **571**, 545 (2002).

[65] W. F. Tompkins, Applications of likelihood analysis in gamma-ray astrophysics, arXiv: Astrophysics (2002).

[66] J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer, Likelihood-free inference with an improved cross-entropy estimator, ArXiv (2018).

[67] V. M. H. Ong, D. J. Nott, M. N. Tran, S. A. Sisson, and C. C. Drovandi, Likelihood-free inference in high dimensions with synthetic likelihood, Computational Statistics Data Analysis **128**, 271 (2018).

[68] S. Self and K. Liang, Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, Journal of the American Statistical Association **82**, 605 (1987).

[69] K. Wachter, R. Leach, and E. Kellogg, Parameter estimation in x-ray astronomy using maximum likelihood, The Astrophysical Journal **230**, 274 (1979).

[70] D. Hooper and T. Linden, On The gamma-ray emission from Reticulum II and other dwarf galaxies, **2015**, 016 (2015), arXiv:1503.06209 [astro-ph.HE].

[71] A. A. Abdo, M. Ackermann, M. Ajello, W. B. Atwood, L. Baldini, J. Ballet, G. Barbiellini, D. Bastieri, B. M. Baughman, K. Bechtol, R. Bellazzini, B. Berenji, R. D. Blandford, E. D. Bloom, E. Bonamente, A. W. Borgland, J. Bregeon, A. Brez, M. Brigida, P. Bruel, T. H. Burnett, S. Buson, G. A. Caliandro, R. A. Cameron, P. A. Caraveo, J. M. Casandjian, E. Cavazzuti, C. Cecchi, Çelik, E. Charles, A. Chekhtman, C. C. Cheung, J. Chiang, S. Ciprini, R. Claus, J. Cohen-Tanugi, L. R. Cominsky, J. Conrad, S. Cutini, C. D. Dermer, A. de Angelis, F. de Palma, S. W. Digel, G. Di Bernardo, E. d. C. e Silva, P. S. Drell, A. Drlica-Wagner, R. Dubois,

D. Dumora, C. Farnier, C. Favuzzi, S. J. Fegan, W. B. Focke, P. Fortin, M. Frailis, Y. Fukazawa, S. Funk, P. Fusco, D. Gaggero, F. Gargano, D. Gasparrini, N. Gehrels, S. Germani, B. Giebels, N. Giglietto, P. Giommi, F. Giordano, T. Glanzman, G. Godfrey, I. A. Grenier, M.-H. Grondin, J. E. Grove, L. Guillemot, S. Guiriec, M. Gustafsson, Y. Hanabata, A. K. Harding, M. Hayashida, R. E. Hughes, R. Itoh, M. S. Jackson, G. Jóhannesson, A. S. Johnson, R. P. Johnson, T. J. Johnson, W. N. Johnson, T. Kamae, H. Katagiri, J. Kataoka, N. Kawai, M. Kerr, J. Knödlseder, M. L. Kocian, F. Kuehn, M. Kuss, J. Lande, L. Latronico, M. Lemoine-Goumard, F. Longo, F. Loparco, B. Lott, M. N. Lovellette, P. Lubrano, G. M. Madejski, A. Makeev, M. N. Mazziotta, W. McConville, J. E. McEnery, C. Meurer, P. F. Michelson, W. Mitthumsiri, T. Mizuno, A. A. Moiseev, C. Monte, M. E. Monzani, A. Morselli, I. V. Moskalenko, S. Murgia, P. L. Nolan, J. P. Norris, E. Nuss, T. Ohsugi, N. Omodei, E. Orlando, J. F. Ormes, D. Paneque, J. H. Panetta, D. Parent, V. Pelassa, M. Pepe, M. Pesce-Rollins, F. Piron, T. A. Porter, S. Rainò, R. Rando, M. Razzano, A. Reimer, O. Reimer, T. Reposeur, S. Ritz, L. S. Rochester, A. Y. Rodriguez, M. Roth, F. Ryde, H. F.-W. Sadrozinski, D. Sanchez, A. Sander, P. M. S. Parkinson, J. D. Scargle, A. Sellerholm, C. Sgrò, M. S. Shaw, E. J. Siskind, D. A. Smith, P. D. Smith, G. Spandre, P. Spinelli, J.-L. Starck, M. S. Strickman, A. W. Strong, D. J. Suson, H. Tajima, H. Takahashi, T. Takahashi, T. Tanaka, J. B. Thayer, J. G. Thayer, D. J. Thompson, L. Tibaldo, D. F. Torres, G. Tosti, A. Tramacere, Y. Uchiyama, T. L. Usher, V. Vasileiou, N. Vilchez, V. Vitale, A. P. Waite, P. Wang, B. L. Winer, K. S. Wood, T. Ylinen, and M. Ziegler, Spectrum of the isotropic diffuse gamma-ray emission derived from first-year fermi large area telescope data, Physical Review Letters **104**, 10.1103/physrevlett.104.101101 (2010).

[72] M. Ackermann, M. Ajello, A. Albert, W. B. Atwood, L. Baldini, J. Ballet, G. Barbiellini, D. Bastieri, K. Bechtol, R. Bellazzini, E. Bissaldi, R. D. Blandford, E. D. Bloom, E. Bottacini, T. J. Brandt, J. Bregeon, P. Bruel, R. Buehler, S. Buson, G. A. Caliandro, R. A. Cameron, M. Caragiulo, P. A. Caraveo, E. Cavazzuti, C. Cecchi, E. Charles, A. Chekhtman, J. Chiang, G. Chiaro, S. Ciprini, R. Claus, J. Cohen-Tanugi, J. Conrad, A. Cuoco, S. Cutini, F. D'Ammando, A. de Angelis, F. de Palma, C. D. Dermer, S. W. Digel, E. d. C. e. Silva, P. S. Drell, C. Favuzzi, E. C. Ferrara, W. B. Focke, A. Franckowiak, Y. Fukazawa, S. Funk, P. Fusco, F. Gargano, D. Gasparrini, S. Germani, N. Giglietto, P. Giommi, F. Giordano, M. Giroletti, G. Godfrey, G. A. Gomez-Vargas, I. A. Grenier, S. Guiriec, M. Gustafsson, D. Hadasch, K. Hayashi, E. Hays, J. W. Hewitt, P. Ippoliti, T. Jogler, G. Jóhannesson, A. S. Johnson, W. N. Johnson, T. Kamae, J. Kataoka, J. Knödlseder, M. Kuss, S. Larsson, L. Latronico, J. Li, L. Li, F. Longo, F. Loparco, B. Lott, M. N. Lovellette, P. Lubrano, G. M. Madejski, A. Manfreda, F. Massaro, M. Mayer, M. N. Mazziotta, J. E. McEnery, P. F. Michelson, W. Mitthumsiri, T. Mizuno, A. A. Moiseev, M. E. Monzani, A. Morselli, I. V. Moskalenko, S. Murgia, R. Nemmen, E. Nuss, T. Ohsugi, N. Omodei, E. Orlando, J. F. Ormes, D. Paneque, J. H. Panetta, J. S. Perkins, M. Pesce-Rollins, F. Piron, G. Pivato, T. A. Porter, S. Rainò, R. Rando, M. Razzano, S. Razzaque, A. Reimer, O. Reimer, T. Reposeur, S. Ritz, R. W. Romani, M. Sánchez-Conde, M. Schaal, A. Schulz, C. Sgrò, E. J. Siskind, G. Spandre, P. Spinelli, A. W. Strong, D. J. Suson, H. Takahashi, J. G. Thayer, J. B. Thayer, L. Tibaldo, M. Tinivella, D. F. Torres, G. Tosti, E. Troja, Y. Uchiyama, G. Vianello, M. Werner, B. L. Winer, K. S. Wood, M. Wood, G. Zaharijas, and S. Zimmer, The Spectrum of Isotropic Diffuse Gamma-Ray Emission between 100 MeV and 820 GeV, Astrophys. J. **799**, 86 (2015), arXiv:1410.3696 [astro-ph.HE].

[73] M. Ackermann, M. Ajello, W. B. Atwood, L. Baldini, J. Ballet, G. Barbiellini, D. Bastieri, K. Bechtol, R. Bellazzini, B. Berenji, R. D. Blandford, E. D. Bloom, E. Bonamente, A. W. Borgland, T. J. Brandt, J. Bregeon, M. Brigida, P. Bruel, R. Buehler, S. Buson, G. A. Caliandro, R. A. Cameron, P. A. Caraveo, E. Cavazzuti, C. Cecchi, E. Charles, A. Chekhtman, J. Chiang, S. Ciprini, R. Claus, J. Cohen-Tanugi, J. Conrad, S. Cutini, A. de Angelis, F. de Palma, C. D. Dermer, S. W. Digel, E. do Couto e Silva, P. S. Drell, A. Drlica-Wagner, L. Falletti, C. Favuzzi, S. J. Fegan, E. C. Ferrara, W. B. Focke, P. Fortin, Y. Fukazawa, S. Funk, P. Fusco, D. Gaggero, F. Gargano, S. Germani, N. Giglietto, F. Giordano, M. Giroletti, T. Glanzman, G. Godfrey, J. E. Grove, S. Guiriec, M. Gustafsson, D. Hadasch, Y. Hanabata, A. K. Harding, M. Hayashida, E. Hays, D. Horan, X. Hou, R. E. Hughes, G. Jóhannesson, A. S. Johnson, R. P. Johnson, T. Kamae, H. Katagiri, J. Kataoka, J. Knödlseder, M. Kuss, J. Lande, L. Latronico, S.-H. Lee, M. Lemoine-Goumard, F. Longo, F. Loparco, B. Lott, M. N. Lovellette, P. Lubrano, M. N. Mazziotta, J. E. McEnery, P. F. Michelson, W. Mitthumsiri, T. Mizuno, C. Monte, M. E. Monzani, A. Morselli, I. V. Moskalenko, S. Murgia, M. Naumann-Godo, J. P. Norris, E. Nuss, T. Ohsugi, A. Okumura, N. Omodei, E. Orlando, J. F. Ormes, D. Paneque, J. H. Panetta, D. Parent, M. Pesce-Rollins, M. Pierbattista, F. Piron, G. Pivato, T. A. Porter, S. Rainò, R. Rando, M. Razzano, S. Razzaque, A. Reimer, O. Reimer, H. F.-W. Sadrozinski, C. Sgrò, E. J. Siskind, G. Spandre, P. Spinelli, A. W. Strong, D. J. Suson, H. Takahashi, T. Tanaka, J. G. Thayer, J. B. Thayer, D. J. Thompson, L. Tibaldo, M. Tinivella, D. F. Torres, G. Tosti, E. Troja, T. L. Usher, J. Vandenbroucke, V. Vasileiou, G. Vianello, V. Vitale, A. P. Waite, P. Wang, B. L. Winer, K. S. Wood, M. Wood, Z. Yang, M. Ziegler, and S. Zimmer, Fermi-lat observations of the diffuse -ray emission: Implications for cosmic rays and the interstellar medium, The Astrophysical Journal **750**, 3 (2012).

[74] F. Acero, M. Ackermann, M. Ajello, A. Albert, L. Baldini, J. Ballet, G. Barbiellini, D. Bastieri, R. Bellazzini, E. Bissaldi, E. D. Bloom, R. Bonino, E. Bottacini, T. J. Brandt, J. Bregeon, P. Bruel, R. Buehler, S. Buson, G. A. Caliandro, R. A. Cameron, M. Caragiulo, P. A. Caraveo, J. M. Casandjian, E. Cavazzuti, C. Cecchi, E. Charles, A. Chekhtman, J. Chiang, G. Chiaro, S. Ciprini, R. Claus, J. Cohen-Tanugi, J. Conrad, A. Cuoco, S. Cutini, F. D'Ammando, A. d. Angelis, F. d. Palma, R. Desiante, S. W. Digel, L. D. Venere, P. S. Drell, C. Favuzzi, S. J. Fegan, E. C. Ferrara, W. B. Focke, A. Franckowiak, S. Funk, P. Fusco, F. Gargano, D. Gasparrini, N. Giglietto, F. Giordano, M. Giroletti, T. Glanzman, G. Godfrey, I. A. Grenier, S. Guiriec, D. Hadasch, A. K.

Harding, K. Hayashi, E. Hays, J. W. Hewitt, A. B. Hill, D. Horan, X. Hou, T. Jogler, G. Jóhannesson, T. Kamae, M. Kuss, D. Landriu, S. Larsson, L. Latronico, J. Li, L. Li, F. Longo, F. Loparco, M. N. Lovellette, P. Lubrano, S. Maldera, D. Malyshev, A. Manfreda, P. Martin, M. Mayer, M. N. Mazziotta, J. E. McEnery, P. F. Michelson, N. Mirabal, T. Mizuno, M. E. Monzani, A. Morselli, E. Nuss, T. Ohsugi, N. Omodei, M. Orienti, E. Orlando, J. F. Ormes, D. Paneque, M. Pesce-Rollins, F. Piron, G. Pivato, S. Rainò, R. Rando, M. Razzano, S. Razzaque, A. Reimer, O. Reimer, Q. Remy, N. Renault, M. Sánchez-Conde, M. Schaal, A. Schulz, C. Sgrò, E. J. Siskind, F. Spada, G. Spandre, P. Spinelli, A. W. Strong, D. J. Suson, H. Tajima, H. Takahashi, J. B. Thayer, D. J. Thompson, L. Tibaldo, M. Tinivella, D. F. Torres, G. Tosti, E. Troja, G. Vianello, M. Werner, K. S. Wood, M. Wood, G. Zaharijas, and S. Zimmer, Development of the model of galactic interstellar emission for standard point-source analysis of fermi large area telescope data, The Astrophysical Journal Supplement Series **223**, 26 (2016).

[75] A. Geringer-Sameth, S. M. Koushiappas, M. G. Walker, V. Bonnivard, C. Combet, and D. Maurin, Astrophysical explanations of suspected dark matter signals in dwarf galaxies, arXiv e-prints , arXiv:1807.08740 (2018), arXiv:1807.08740 [astro-ph.HE].

[76] T.-T. Ge, X.-N. Sun, R.-z. Yang, Y.-F. Liang, and E. Liang, Diffuse γ-ray emission around the massive star forming region of carina nebula complex, Monthly Notices of the Royal Astronomical Society 10.1093/mnras/stac2885 (2022).

[77] T. F.-L. Collaboration, Fermi large area telescope third source catalog, The Astrophysical Journal Supplement Series **218**, 23 (2015).

[78] G. Torrealba, V. Belokurov, S. E. Koposov, K. Bechtol, A. Drlica-Wagner, K. A. G. Olsen, A. K. Vivas, B. Yanny, P. Jethwa, A. R. Walker, T. S. Li, S. Allam, B. C. Conn, C. Gallart, R. A. Gruendl, D. J. James, M. D. Johnson, K. Kuehn, N. Kuropatkin, N. F. Martin, D. Martinez-Delgado, D. L. Nidever, N. E. D. Noël, J. D. Simon, G. S. Stringfellow, and D. L. Tucker, Discovery of two neighbouring satellites in the Carina constellation with MagLiteS, **475**, 5085 (2018), arXiv:1801.07279 [astro-ph.GA].

[79] A. Geringer-Sameth, M. G. Walker, S. M. Koushiappas, S. E. Koposov, V. Belokurov, G. Torrealba, and N. W. Evans, Indication of Gamma-Ray Emission from the Newly Discovered Dwarf Galaxy Reticulum II, Phys. Rev. Lett. **115**, 081101 (2015), arXiv:1503.02320 [astro-ph.HE].

[80] A. McDaniel, M. Ajello, C. M. Karwin, M. Di Mauro, A. Drlica-Wagner, and M. A. Sánchez-Conde, Legacy analysis of dark matter annihilation from the Milky Way dwarf spheroidal galaxies with 14 years of Fermi -LAT data, Phys. Rev. D **109**, 063024 (2024), arXiv:2311.04982 [astro-ph.HE].

[81] F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt, Importance nested sampling and the multinest algorithm, The Open Journal of Astrophysics **2**, 10.21105/astro.1306.2144 (2019).

[82] M. Cirelli, G. Corcella, A. Hektor, G. Hütsi, M. Kadastik, P. Panci, M. Raidal, F. Sala, and A. Strumia, PPPC 4 DM ID: a poor particle physicist cookbook for dark matter indirect detection, **2011**, 051 (2011), arXiv:1012.4515 [hep-ph].

[83] P. Ciafaloni, D. Comelli, A. Riotto, F. Sala, A. Strumia, and A. Urbano, Weak corrections are relevant for dark matter indirect detection, **2011**, 019 (2011), arXiv:1009.0224 [hep-ph].

[84] A. Chiappo, J. Cohen-Tanugi, J. Conrad, L. Strigari, B. Anderson, and M. Sánchez-Conde, Dwarf spheroidal j-factors without priors: A likelihood-based analysis for indirect dark matter searches, Monthly Notices of the Royal Astronomical Society **466**, 669 (2016).

[85] A. Pace and L. Strigari, Scaling relations for dark matter annihilation and decay profiles in dwarf spheroidal galaxies, Monthly Notices of the Royal Astronomical Society 10.1093/mnras/sty2839 (2018).

[86] A. Geringer-Sameth, S. M. Koushiappas, and M. Walker, Dwarf Galaxy Annihilation and Decay Emission Profiles for Dark Matter Experiments, Astrophys. J. **801**, 74 (2015), arXiv:1408.0002 [astro-ph.CO].

[87] K. Boddy, J. Kumar, L. Strigari, and M. Y. Wang, Sommerfeld-enhanced j-factors for dwarf spheroidal galaxies, Physical Review D **95**, 123008 (2017).

[88] S. Ando, A. Geringer-Sameth, N. Hiroshima, S. Hoof, R. Trotta, and M. G. Walker, Structure formation models weaken limits on WIMP dark matter from dwarf spheroidal galaxies, Phys. Rev. D **102**, 061302 (2020), arXiv:2002.11956 [astro-ph.CO].

[89] F. Feroz and M. P. Hobson, Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses: Multimodal nested sampling, Monthly Notices of the Royal Astronomical Society **384**, 449–463 (2008).

[90] F. Feroz, M. P. Hobson, and M. Bridges, Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics, Monthly Notices of the Royal Astronomical Society **398**, 1601–1614 (2009).

[91] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . Contributors, Scipy 1.0: Fundamental algorithms for scientific computing in python, Nature Methods **17**, 261 (2020).

[92] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, European Physical Journal C **71**, 1554 (2011).

[93] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, European Physical Journal C **71**, 1554 (2011).

[94] CERN, Why do physicists mention "five sigma" in their results?, CERN Courier (2023).

[95] L. Lyons, Discovering the significance of 5 sigma, arXiv preprint arXiv:1310.1284 (2013).

[96] E. van Zwet and A. Gelman, A proposal for informative default priors scaled by the standard error of estimates, The American Statistician **76**, 1 (2022), https://doi.org/10.1080/00031305.2021.1938225.

[97] E. van Zwet, S. Schwab, and S. Senn, The statistical properties of rcts and a proposal for shrinkage (2020), arXiv:2011.15004 [stat.ME].