

This is the accepted manuscript made available via CHORUS. The article has been published as:

Fundamental statistical limitations of future dark matter direct detection experiments

Charlotte Strege, Roberto Trotta, Gianfranco Bertone, Annika H. G. Peter, and Pat Scott

Phys. Rev. D **86**, 023507 — Published 3 July 2012

DOI: [10.1103/PhysRevD.86.023507](https://doi.org/10.1103/PhysRevD.86.023507)

Fundamental statistical limitations of future dark matter direct detection experiments

Charlotte Strege,¹ Roberto Trotta,^{1,2} Gianfranco Bertone,³ Annika H. G. Peter,⁴ and Pat Scott⁵

¹*Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK*

²*Kavli Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106-4030, USA*

³*Institute for Theoretical Physics, University of Amsterdam,*

Science Park 904, Postbus 94485, 1090 GL Amsterdam, The Netherlands

⁴*Department of Physics and Astronomy, University of California, Irvine, California 92697-4575, USA*

⁵*Department of Physics, McGill University, 3600 rue University, Montréal, QC, H3A 2T8, Canada*

We discuss irreducible statistical limitations of future ton-scale dark matter direct detection experiments. We focus in particular on the coverage of confidence intervals, which quantifies the reliability of the statistical method used to reconstruct the dark matter parameters, and the bias of the reconstructed parameters. We study 36 benchmark dark matter models within the reach of upcoming ton-scale experiments. We find that approximate confidence intervals from a profile-likelihood analysis exactly cover or over-cover the true values of the WIMP parameters, and hence are conservative. We evaluate the probability that unavoidable statistical fluctuations in the data might lead to a biased reconstruction of the dark matter parameters, or large uncertainties on the reconstructed parameter values. We show that this probability can be surprisingly large, even for benchmark models leading to a large event rate of order a hundred counts. We find that combining data sets from two different targets leads to improved coverage properties, as well as a substantial reduction of statistical bias and uncertainty on the dark matter parameters.

I. INTRODUCTION

Among the large number of possible dark matter candidates [1–4], weakly-interacting massive particles (WIMP) [5] are by far the most widely studied. WIMPs naturally arise from popular extensions of the Standard Model of particle physics (e.g., the lightest neutralino in supersymmetry [6, 7] and the B^1 in theories with universal extra dimensions [8–10]), and they naturally achieve the appropriate cosmological relic density through thermal freeze-out in the early Universe.

Several experiments are currently searching for these particles by looking for signals of WIMPs scattering on atomic nuclei in large underground detectors, and many others are planned for the next decade (see e.g. Ref. [1] and the discussion in Ref. [11]). Although the DAMA/LIBRA [12] and CoGeNT [13] collaborations have reported a modulation of the measured event rate that has been tentatively interpreted in terms of WIMPs (e.g. [14]), and the CRESST-II collaboration has found a large excess of events in the acceptance region where a WIMP signal would be expected [15], these results can hardly be reconciled with null searches from experiments such as XENON100 [16, 17], CDMS-II [18, 19], EDELWEISS-II [20] and ZEPLIN-III [21]. The controversy will hopefully be resolved by next-generation direct detection experiments, where larger rates and better statistics could lead to an incontrovertible discovery of dark matter.

If a WIMP-nucleon scattering signal is detected, the event rate and the shape of the measured spectrum of recoil energies can be used to determine the properties of the dark-matter particle, most importantly its mass and scattering cross-section. The constraining power of present and upcoming experiments has been thoroughly

discussed in the literature [11, 22–25]. Here, we present *irreducible statistical limitations* of future dark matter direct detection experiments.

We focus on two different issues: first, we explore the concept of *coverage* of confidence intervals, which quantifies the reliability of the statistical method adopted to reconstruct the WIMP parameters. We investigate the coverage of one-dimensional confidence intervals, constructed using an approximate method that relies on the assumption that profile likelihood ratios are chi-square distributed, based on Wilks’ theorem [26]. This approximate method of constructing confidence intervals is commonly used for frequentist data analysis in the literature in lieu of more complex methods (e.g. Feldman and Cousins [27]), which provide exact coverage by construction. The coverage of parameter reconstructions has been previously discussed in the context of direct detection [28] and collider identification [29] of supersymmetric models.

Second, we consider how well one can expect to reconstruct the WIMP properties from future direct-detection data, given the statistical fluctuations that will inevitably impact the observed energy spectrum. We perform parameter reconstructions on thousands of simulated data sets to estimate the *average* uncertainty and bias in the reconstructions of several different WIMP benchmark models. Additionally, we provide an estimate of the number of *outliers* in the parameter reconstructions. We show that for several different benchmark models that lead to small average uncertainties in the parameter reconstruction, a non-negligible percentage of all reconstructions results in a much larger uncertainty, as a result of statistical fluctuations that impact on each individual data set. Considering the number of outliers for different WIMP benchmark models is of crucial importance, since in practice there will be a unique realisation of each

experiment, and the constraints derived from a particular realisation can be very different from the outcome for the “average experiment”, as illustrated below. Finally, we investigate how the average uncertainty in the WIMP mass can be decreased by increasing the exposure of direct detection experiments, for several different benchmark points in WIMP parameter space.

The complementarity between direct detection experiments using different target materials, and the possibility of obtaining tighter constraints on the WIMP parameters when combining data from more than one experiment, have recently been emphasized in Ref. [11, 24, 30, 31]. Here we compare the coverage, uncertainty and bias of reconstructed parameters for various benchmark points, based either on mock data sets from a single xenon experiment, or a combined analysis of mock data from a xenon experiment and a germanium experiment.

Throughout our analysis we assume that the background event rate is negligible, and ignore uncertainties in the nuclear physics of elastic scattering and the local WIMP distribution function. We expect that the coverage, precision and bias of our reconstructions will degrade if the backgrounds are non-negligible and astrophysical uncertainties are fully taken into account. Given this optimistic set-up, we present here a set of *irreducible limitations* on WIMP parameter reconstruction from future direct-detection experiments, arising from fundamental statistical fluctuations driven by the Poisson nature of the event rate.

The paper is organized as follows: in Sec. II we introduce the formalism of direct dark matter detection and discuss the expected performance of upcoming experiments. In Sec. III we present our parameter reconstruction method and introduce the statistical quantities we use to quantify the performance of our reconstruction procedure. We present our results in Sec. IV and our conclusions in Sec. V.

II. DIRECT DARK MATTER DETECTION

A. Theoretical formalism

Dark matter direct detection experiments aim to detect signals of WIMPs scattering on target nuclei. The nuclear recoil spectrum for a WIMP of mass m_χ and a target nucleus of mass m_N has the form

$$\frac{dR}{dE_R}(E_R) = \frac{\rho_0}{m_\chi m_N} \int_{v > v_{\min}} d^3\vec{v} \frac{d\sigma}{dE_R} v f(\vec{v} + \vec{v}_E) \quad . \quad (1)$$

Here dR/dE_R has units of events per unit energy per unit time per unit target material mass, ρ_0 is the local dark matter density, σ is the WIMP-nucleus scattering cross-section and E_R is the WIMP-induced recoil energy of the nucleus. Neglecting gravitational focusing of WIMPs as they flow into the potential well of the Solar System, $f(\vec{v})$ is the normalized local WIMP velocity distribution func-

tion in the rest frame of the Galaxy, \vec{v}_E is the Earth’s velocity in this frame and \vec{v} is the velocity of the WIMPs in the rest frame of the Earth (which is also the WIMP-nucleon relative velocity, as to a good approximation the nucleons are at rest in the Earth frame). In this paper we focus on elastic WIMP-nucleus interactions. For elastic scattering the minimum velocity v_{\min} required for a WIMP of mass m_χ to be able to induce a nuclear recoil of energy E_R is

$$v_{\min} = \sqrt{\frac{m_N E_R}{2\mu_N^2}} \quad , \quad (2)$$

where $\mu_N = m_\chi m_N / (m_\chi + m_N)$ is the WIMP-nucleus reduced mass.

The differential scattering cross-section $d\sigma/dE_R$ includes different types of WIMP-nucleus interactions. We will assume that all events result from spin-independent WIMP-nucleus scattering and neglect all other types of interactions. In this case the differential scattering cross-section is given by

$$\frac{d\sigma}{dE_R} = \frac{m_N}{2v^2\mu_N^2} \sigma_N^{SI} \mathcal{F}^2(E_R) \quad , \quad (3)$$

where $\mathcal{F}(E_R)$ is the spin-independent nuclear form factor, which accounts for the finite extent and composite nature of the atomic nucleus, and σ_N^{SI} is the spin-independent (SI) zero-momentum WIMP-nucleus cross-section. This cross-section can be written in terms of the mass number of the nucleon A , its atomic number Z , the WIMP-proton coupling f_p , and the WIMP-neutron coupling f_n .

$$\sigma_N^{SI} = \frac{4}{\pi} \mu_N^2 (Z f_p + (A - Z) f_n)^2 \quad . \quad (4)$$

In the following we will assume that the WIMP-proton and WIMP-neutron couplings are very similar $f_p \sim f_n$ (as appropriate in most supersymmetric setups [32], but see also Refs.[33–36] for alternative scenarios), so that the WIMP-nucleus cross-section simplifies to $\sigma_N^{SI} = 4\mu_N^2 A^2 f_p^2 / \pi$. In analogy to this expression we define the WIMP-proton cross-section $\sigma_p^{SI} = 4\mu_p^2 f_p^2 / \pi$, with $\mu_p = m_\chi m_p / (m_\chi + m_p)$ the WIMP-proton reduced mass. The differential scattering cross-section can then be rewritten as

$$\frac{d\sigma}{dE_R} = \frac{m_N}{2v^2\mu_p^2} A^2 \sigma_p^{SI} \mathcal{F}^2(E_R) \quad . \quad (5)$$

In this analysis we use the Helm form factor [37]

$$\mathcal{F}(E_R) = 3 \frac{\sin(qr) - (qr)\cos(qr)}{(qr)^3} e^{-(qs)^2/2} \quad , \quad (6)$$

where $q = \sqrt{2m_N E_R}$ is the momentum transferred in the recoil, $s = 0.9$ fm, $r = \sqrt{c^2 + 7\pi^2 a^2/3 - 5s^2}$, $a = 0.52$ fm and $c = (1/23A^{1/3} - 0.6)$ fm. Using Eq. (5) the nuclear recoil spectrum can be rewritten as

$$\frac{dR}{dE_R}(E_R) = \frac{\rho_0 \sigma_p^{SI} A^2 \mathcal{F}^2(E_R)}{2\mu_p^2 m_\chi} \int_{v > v_{\min}} d^3\vec{v} \frac{f(\vec{v} + \vec{v}_E)}{v} \quad . \quad (7)$$

The quantities of interest are the WIMP mass m_χ and the spin-independent WIMP-proton cross-section σ_p^{SI} . The choice of target material enters the analysis via the mass number A and the form factor $\mathcal{F}(E_R)$, and through v_{\min} . Note for $m_\chi \gg m_N$, $v_{\min} \rightarrow \sqrt{E_R/2m_N}$, and hence the recoil spectrum depends on m_χ and σ_p^{SI} only via the degenerate combination $\sigma_p^{SI}/(\mu_p^2 m_\chi)$, which has a strong impact on the performance of the reconstruction of the WIMP properties, as we will see in the following sections.

The third component that enters the recoil rate is the local astrophysical DM distribution, most importantly the local density ρ_0 and the WIMP velocity distribution $f(\vec{u})$. In this analysis we will model local astrophysics using the standard halo model. This model consists of an isothermal, spherically symmetric galactic WIMP distribution. In this model, WIMP velocities follow a non-rotating isotropic Maxwellian distribution in a Galactocentric frame with a one-dimensional velocity dispersion $v_0/\sqrt{2}$, where v_0 is the speed of the Local Standard of Rest. WIMPs traveling at very high velocities will escape the gravitational attraction of the galaxy and will therefore not be present in the halo. This is taken into account by truncating the velocity distribution at some escape velocity v_{esc} , leading to a WIMP velocity distribution function

$$f(\vec{v} + \vec{v}_E) = \begin{cases} \frac{N^{-1}}{v_0^3 \pi^{3/2}} e^{-(\vec{v} + \vec{v}_E)^2/v_0^2}, & \text{for } |\vec{v} + \vec{v}_E| < v_{esc} \\ 0 & \text{otherwise} \end{cases}, \quad (8)$$

with $N = \text{erf}(v_{esc}/v_0) - 2\pi^{-1/2}(v_{esc}/v_0)e^{-(v_{esc}/v_0)^2}$ a normalization factor which ensures that $\int d^3\vec{u} f(\vec{u}) = 1$. The velocity of the Earth with respect to the rest frame of the galaxy is given by the sum of the local circular velocity \vec{v}_0 , the Sun's peculiar velocity \vec{v}_{pec} and the Earth's velocity relative to the Sun \vec{v}_{orb}

$$\vec{v}_E = \vec{v}_0 + \vec{v}_{pec} + \vec{v}_{orb}. \quad (9)$$

The contribution of both $|\vec{v}_{pec}| \sim 10$ km/s and $\vec{v}_{orb} \sim 30$ km/s to \vec{v}_E is small compared to the contribution of $\vec{v}_0 \sim 200 - 300$ km/s. As we consider neither directional signatures nor the annual modulation of the nuclear recoil spectrum in this study, the latter two terms in Eq. (9) can be neglected and $\vec{v}_E \simeq \vec{v}_0$.

It is well known that there is a sizeable uncertainty on the astrophysical parameters ρ_0, v_0, v_{esc} and $f(\vec{u})$. Additionally, the standard halo model can only be considered a first approximation to a much more complicated halo profile [38–41]. In order to achieve a correct reconstruction of the WIMP parameters from experiment, it is of vital importance to take into account these uncertainties [23–25]. The aim of this paper is to investigate the coverage properties and the quality of the reconstruction for different WIMP benchmark models and identify any unavoidable statistical effects. In order to do so we will assume an ideal case, fixing all of the astrophysical parameters to their fiducial values and neglecting their uncertainties. The fiducial values we use are

$\rho_0 = 0.4$ GeV/cm³, $v_0 = 230$ km/s and $v_{esc} = 544$ km/s. We will investigate coverage properties of a more general framework that includes astrophysical uncertainties in the WIMP distribution function in a future work.

The total number of recoil events N_R can be found by weighting the nuclear recoil rate in Eq. (7) by the event acceptance $\epsilon(E_R)$, and integrating from some threshold energy E_{thr} to some maximum energy E_{\max} . Assuming that the acceptance is not energy-dependent, $\epsilon(E_R)$ simply falls out of the integral, and becomes a mean effective exposure ϵ_{eff} (which is the product of the detector mass and exposure time). N_R is then given by

$$N_R = \epsilon_{\text{eff}} \int_{E_{thr}}^{E_{\max}} dE_R \frac{dR}{dE_R}. \quad (10)$$

For our coverage study, we select a number of WIMP benchmark models, with benchmark mass and cross-section ranges $m_\chi = [25, 250]$ GeV and $\sigma_p^{SI} = [10^{-8}, 10^{-10}]$ pb. For each benchmark point the analysis is based on 10^3 mock data sets.

B. Future direct detection experiments

In order to assess the performance of the reconstruction of WIMP properties from next-generation direct detection data, we will use ton-scale, low-background versions of two current detectors. We will systematically investigate the constraints that data sets from these experiments can place on the WIMP properties for different benchmark models.

The most stringent constraints on WIMP properties are currently provided by the XENON100 collaboration [17]. The recently published 90% C.L. exclusion curve has a minimum cross-section of $\sigma_p^{SI} = 7.0 \times 10^{-9}$ pb at a WIMP mass $m_\chi = 50$ GeV [17]. These constraints will be improved further once data from the proposed XENON1T experiment becomes available in 2015 [42]. Additionally, the DARWIN project¹ is working towards a multi-ton scale noble liquid experiment which is expected to start running in 2017 and will probe spin-independent cross-sections down to 10^{-12} pb. [43] A second promising WIMP detection strategy is based on cryogenic detectors operating at very low temperatures, most notably the current CDMS-II germanium experiment [18]. The SuperCDMS and GEODM cryogenic germanium experiments aim to upgrade this experiment to the ton scale within the next decade [44]. A second planned experiment using cryogenic detectors operating at mK temperatures is EURECA². This experiment is pushing for a target mass of 1 ton and will probe cross-sections down to 10^{-10} pb.

¹ <http://darwin.physik.uzh.ch>

² <http://www.eureca.ox.ac.uk>

In this study we will use a ton-scale experiment with a liquid natural Xe target with average atomic mass 131 g/mol, and a ton-scale Ge experiment with atomic mass 73 g/mol. The characteristics of these detectors are chosen to reflect projects that can realistically be built within the next 5 - 10 years; they are given in Table I. Although large liquid argon experiments are also currently under construction, we choose not to include simulated argon data in this study, because previous studies have shown that germanium and xenon provide tighter constraints on the WIMP parameters and halo velocity distribution [11].

For both the xenon and the germanium experiment we assume a threshold energy of $E_{thr} = 10$ keV and only consider recoil energies below 100 keV. This is a reasonable cut-off, given the exponential decay of the WIMP-nucleus recoil spectrum with energy. Studies have shown that resolving the exponential decay at high energies is important for improving parameter reconstruction [25]. For both experiments we assume a total cut efficiency of $\eta_{cut} = 80\%$. Following Ref. [11], for the Xe experiment we take a fiducial detector mass of 5 tons and one year of operation. We assume that a percentage $A_{NR} = 50\%$ of all nuclear recoils in the fiducial region are accepted, so that, after inclusion of the overall cut efficiency, the effective exposure is $\epsilon_{eff} = 2.00$ ton \times year. For the germanium experiment we adopt a fiducial detector mass of 1 ton and an exposure of three years. Taking into account the percentage of events that survive the selection cuts η_{cut} and the nuclear recoil acceptance for germanium $A_{NR} = 90\%$ the effective exposure is $\epsilon_{eff} = 2.16$ ton \times years.

Several sources of background can induce additional recoil events in direct detection experiments, such as cosmic rays, or radioactive contaminations. Future detectors will apply a variety of advanced techniques in order to achieve extreme radio-purity and self-shielding of the detector, minimisation of cosmic ray events and precise determination of charge-to-light and charge-to-phonon ratios, in order to limit the background to < 1 event per effective exposure. Given these prospects in the following we assume that backgrounds are negligible.

We do not include the energy resolution of the detectors, as for both target materials including energy resolution smearing has a negligible impact on the recoil rate, except possibly near threshold. The scenario considered here is therefore somewhat idealised, which means that the statistical uncertainties we identify are unavoidable, inherent to the WIMP benchmark point and target exposure, rather than a reflection of systematic uncertainties in detector response, backgrounds or modelling of the dark matter halo.

III. STATISTICAL METHODOLOGY

A. Mock data generation

The data set for a direct dark matter experiment consists of the total number of observed events \hat{N}_R and the spectrum of recoil energies $\{\hat{E}_R^i\}$, with $i = 1, \dots, \hat{N}_R$. The likelihood function $\mathcal{L}(\theta)$ for the WIMP parameters $\theta = \{m_\chi, \sigma_p^{SI}\}$ is given by the Poisson probability of observing \hat{N}_R events, multiplied by the probabilities of each event of energy E_R^i having been drawn from the predicted probability distribution of event energies $P(E_R|\theta)$

$$\mathcal{L}(\theta) = \frac{N_R(\theta)^{\hat{N}_R}}{\hat{N}_R!} \exp[-N_R(\theta)] \prod_{i=1}^{\hat{N}_R} P(\hat{E}_R^i|\theta) \quad (11)$$

Notice that in the above we have replaced the (latent, unobserved) true recoil energy E_R^i by the observed value \hat{E}_R^i , thus assuming that energy resolution of the detectors is negligible, as outlined in the previous section. $N_R(\theta)$ can be computed from Eq. (10), using the experimental characteristics in Table I. The distribution $P(\hat{E}_R, \theta)$ is no more than the normalized recoil spectrum

$$P(\hat{E}_R, \theta) = \frac{dR/dE_R(\hat{E}_R, \theta)}{\int_{E_{min}}^{E_{max}} dE'_R dR/dE'_R(E'_R, \theta)} \quad (12)$$

where the rate $dR/dE_R(E_R, \theta)$ is given in Eq. (7). Note that the efficiency parameter ϵ_{eff} drops out in the one-event likelihood because we assume that this function is independent of recoil energy. For both the Xe and the Ge target the integration limits are $E_{min} = 10$ keV and $E_{max} = 100$ keV. As explained in the previous section no background events are included in \hat{N}_R , as we assume the background to be negligible. The so-called unbinned likelihood function in Eq. (11) has been employed by both the XENON and the CDMS collaborations [45, 46]. The likelihood function for the combined data set of our two toy experiments is given by the product of the individual likelihood functions, each found from Eq. (11).

The mock data sets for the experiments are generated as follows. First, the measured total number of counts \hat{N}_R is drawn from a Poisson distribution with mean equal to the benchmark number of counts N_R . Then, values for the measured recoil energies $\{\hat{E}_R^i\}$, $i = 1, \dots, \hat{N}_R$ are drawn from the differential event rate $dR/dE_R(E_R)$, given in Eq. (7), for the benchmark value of the parameters.

B. Parameter reconstruction technique

We employ Bayesian methods to scan over the parameter space and reconstruct the WIMP properties, see [47] for further details. The cornerstone of Bayesian param-

Target	E_{thr} [keV]	ϵ [ton \times year]	A_{NR}	ϵ_{eff} [ton \times yr]	# Background events
Xe	10.0	5.00	0.5	2.00	< 1
Ge	10.0	3.00	0.9	2.16	< 1

TABLE I. Primary characteristics of future ton-scale dark matter direct detection experiments using xenon and germanium as target materials. For further details see section IIB.

ter inference is Bayes’ theorem

$$p(\theta|d) = \frac{\mathcal{L}(\theta)p(\theta)}{p(d)}, \quad (13)$$

where $p(\theta|d)$ is the posterior probability density function (pdf), $\mathcal{L}(\theta)$ is the likelihood function and $p(\theta)$ is the prior distribution on the parameters. The evidence is given by $p(d)$, which in the context of parameter inference acts as a normalisation constant and will not be of interest in the following. There are two possible ways of looking at parameter inference: either in the Bayesian context (where the posterior pdf is the relevant quantity) or in the frequentist framework (where the likelihood function or a related test statistic is considered). In this work, we will use Bayesian Markov Chain Monte Carlo (MCMC) techniques to obtain samples from the posterior pdf of Eq. (13), but we will also use these samples to map the likelihood function in the parameter space of interest, here the WIMP mass and the WIMP-proton spin-independent scattering cross-section, $\theta = \{m_\chi, \sigma_p^{SI}\}$. In order to sample from the posterior distribution on these parameters, we have to specify their prior pdf $p(\theta)$. Without assuming a specific underlying WIMP model there are no a priori constraints on m_χ and σ_p^{SI} . Therefore, we choose uniform priors on the log of both the WIMP mass and cross-section, reflecting ignorance on their order of magnitude. The mass prior range is fixed to $1 \leq \log_{10}(m_\chi/\text{GeV}) \leq 3$. The range of the cross-section prior is chosen to span two orders of magnitude around the benchmark cross-section. We extend this range where required, to avoid regions of high posterior probability density touching the prior boundary.

Because the likelihood function is unimodal and well-behaved, and the parameter space is of low dimensionality ($D = 2$), we can efficiently sample the posterior pdf using MCMC methods and use the ensuing samples to map out the likelihood function in a quasi-frequentist sense (see [48] for a detailed study of profile likelihood evaluation using Bayesian techniques in the context of supersymmetric models). To this end, we use a Metropolis-Hastings algorithm [49, 50] to generate a “chain” of samples from the posterior pdf. As our proposal distribution we take a two-dimensional Gaussian centred on the previous point in the chain; its covariance matrix is chosen according to earlier test runs. For some of the benchmark points we consider, the shape of the posterior distribution can vary strongly because of statistical fluctuations in the data realisation. In these cases, to achieve an efficient and complete sampling of the posterior we adopt a mixture strategy MCMC: our proposal distribution is a mixture

of two different two-dimensional Gaussians, whose covariance matrices are chosen (from earlier test runs) to match the two very different shapes of the posterior distribution that can arise from the same benchmark model due to statistical fluctuations in the data (“good” reconstructions and “bad” reconstructions, to be defined more precisely below). Every third proposal of the MCMC is not drawn from this Gaussian mixture, but instead is taken in a random direction, with a step size tuned to achieve an acceptable efficiency, in order to protect against under-exploration of the tails of the posterior.

Each Markov chain contains a minimum number $N = 3 \times 10^5$ samples; this ensures high enough statistics for a successful coverage investigation. Some benchmark models lead to a very spread-out posterior distribution. In these cases we further increased the number of points in the chains, up to a maximum of $N = 5 \times 10^5$ points. We discarded the initial 10^4 samples of each chain (the so-called “burn-in”). We checked that this is sufficient to ensure that the resulting distribution is independent of the starting point of the MCMC and that the results of our analysis are stable when the length of the chains is doubled. Finally, we tested our MCMC method on toy models with known analytic posterior distributions, in order to verify its suitability and numerical stability.

C. Coverage

There are two ways of reporting inferences: $x\%$ credible intervals (Bayesian) contain a fraction x of the posterior probability; they express the posterior degree of belief about the value of the parameter considered after the data and any prior information have been taken into account. An $x\%$ confidence interval (Frequentist) is built from the likelihood function alone, and, ideally, it ought to contain (“cover”) the true value of the parameter $x\%$ of the time, when repeatedly applied to mock data generated from those true parameter values. This requirement leads to the concept of “coverage”. Coverage is an inherently frequentist concept, and it is not necessarily of concern to Bayesian statistics, although reliable behaviour of Bayesian credible intervals under repeated sampling is arguably also a desirable property. In the following, we will mainly focus on evaluating the coverage and other statistical properties of (frequentist) confidence intervals, for the reasons outlined below.

The profile likelihood test statistic for a point X in some N -dimensional subspace Θ_N of the full M -dimensional parameter space Θ_M (i.e. $X \in \Theta_N \subset \Theta_M$),

is

$$\lambda(X) = -2 \ln \left(\frac{\mathcal{L}[X, \hat{\Theta}_{M-N}(X)]}{\mathcal{L}_{\max}} \right) . \quad (14)$$

Here \mathcal{L}_{\max} is the unconditional maximum likelihood i.e. the global maximum likelihood value across the entire M -dimensional parameter space. $\mathcal{L}[X, \hat{\Theta}_{M-N}(X)]$ is the conditional maximum likelihood for the given point X . The subspace Θ_{M-N} refers to the section of Θ_M that is not spanned by Θ_N . $\hat{\Theta}_{M-N}(X)$ is the conditional maximum likelihood estimate of the values of the parameters in Θ_{M-N} for X , i.e. the specific combination of the other $M - N$ parameters that maximises the likelihood for the chosen X in Θ_N . Confidence intervals with exact coverage can always be constructed by Monte Carlo evaluation of the distribution of $\lambda(X)$, as described in Ref. [27], but in practice this may be a complicated and time-consuming procedure.

Wilks' theorem [26] shows that under certain regularity conditions, Eq. (14) converges asymptotically to a chi-square distribution with N degrees of freedom. Assuming Wilks' theorem holds, it is simple to define confidence intervals using the profile likelihood function and standard lookup tables for the chi-square distribution. However, in practice there is no guarantee that such confidence intervals will have the desired coverage properties, especially in cases where the likelihood function is strongly non-Gaussian, which leads to a lack of convergence of the test statistic to its asymptotic behaviour. Under-coverage (over-coverage) of a confidence interval means that the interval is too short (too large). While over-coverage is unnecessarily conservative, under-coverage can be a particularly severe problem, as the true value of the parameters will lie outside the stated interval a larger fraction of the time than its stated confidence level implies.

In the following analysis we discuss the coverage of Wilks-based 1D confidence intervals for the WIMP mass and spin-independent cross-section. The profile likelihood is constructed by binning the 2D parameter space ($\{m_\chi, \sigma_p^{SI}\}$), and determining the test statistics (14) in each bin. We then use Wilks' theorem to find the confidence level of interest. We used 750 bins in each direction of parameter space, choosing the bin size so that they covered the whole range spanned by the samples. We found that a significantly larger number of bins leads to large numerical noise, while a smaller number gives too coarse a likelihood mapping and hence artificial over-coverage (as tested on Gaussian toy models, for which the coverage is exact).

D. Performance of parameter reconstruction

In addition to determining how well the Wilks-based confidence levels cover the benchmark models, we are interested in estimating how well one may expect to constrain WIMP properties from future direct detection data

sets, *including realisation noise*. An important indicator is the uncertainty in the reconstructed parameters. In order to quantify this, we consider the expected fractional uncertainty (e.f.u.) along a direction in parameter space. The fractional uncertainty (f.u.) is defined as the fractional length of the 68% confidence interval relative to the benchmark parameter value θ_{true} :

$$\text{f.u.} = \frac{\theta_{\max}^{68\%} - \theta_{\min}^{68\%}}{\theta_{\text{true}}} . \quad (15)$$

The e.f.u. is the average of this quantity over 100 reconstructions. However, even a benchmark model with a small *average* f.u. may contain a sizeable number of reconstructions with a large parameter uncertainty. Therefore, in addition to the e.f.u. we also count the number of ‘bad’ reconstructions in 100 reconstructions. A bad case is defined as a reconstruction with an $\text{f.u.} > 0.75$, in which case only very limited constraints can be placed on the parameter in question (m_χ or σ_{SI}) from the data.

The f.u. is somewhat similar to the statistical quantity known as effect size [51, 52], which for the case of σ_{SI} is

$$d \equiv \frac{(\hat{\sigma}_{\text{SI}} - \sigma_{\text{SI, null}})}{SD} . \quad (16)$$

Here $\hat{\sigma}_{\text{SI}}$ and SD are the mean and standard deviation, respectively, of a series of repeated measurements of σ_{SI} . In our case, an equivalent role to $\hat{\sigma}_{\text{SI}}$ and SD are played by the best-fit reconstructed value of σ_{SI} , and half the width of the corresponding 68% CI. This is because these quantities are good estimators for, respectively, the true value of σ_{SI} and the standard deviation of $\hat{\sigma}_{\text{SI}}$, the observed best-fit value. The quantity $\sigma_{\text{SI, null}}$ refers to the value of σ_{SI} under the null hypothesis, i.e. the default situation against which the effect is being sought. In our case, the null hypothesis is simply that there is no WIMP signal, so $\sigma_{\text{SI}} = 0$. Therefore, in the limit of zero bias, where the best-fit value of σ_{SI} is exactly equal to the benchmark value, e.f.u. is approximately equivalent to $2d^{-1}$. The case of WIMP mass is less straightforward, as m_χ is undefined under the null hypothesis.

One of the basic properties of statistical inference is that the power of a statistical test (its ability to avoid excluding a true hypothesis that differs from the null hypothesis) increases with d [52, 53]. This is simply the statement that larger effects can be detected more easily. We can therefore see that the e.f.u. not only relates to the precision with which the WIMP mass can be reconstructed, but also gives some idea of the statistical power for detection of a WIMP with this mass. That is, a smaller e.f.u. indicates that a model can be detected more easily, so we expect the e.f.u. to roughly track the sensitivity of an experiment across the WIMP parameter space.

We can further investigate the performance of the statistical reconstruction by explicitly considering the bias³

³ Another useful quantity is the so-called “mean squared error”

for the parameters m_χ and σ_p^{SI} . The statistical bias for a parameter θ is the expectation value of the difference between the best fit value $\hat{\theta}_{\text{bf}}$ resulting from the reconstruction and the true value θ_{true} , i.e.

$$\text{bias} = \langle \hat{\theta}_{\text{bf}} - \theta_{\text{true}} \rangle. \quad (17)$$

As for the e.f.u., the expectation is taken by averaging the observed bias over 100 reconstructions. In the following we focus on the e.f.u. and bias of the reconstructed WIMP mass, as the performance of the reconstruction is expected to typically be poorer in the mass than the cross-section direction, due to the impact of statistical fluctuations on the observed recoil spectrum.

IV. RESULTS

A. The impact of statistical fluctuations on the reconstruction

We investigate the performance of the reconstruction of WIMP properties for six benchmark masses $m_\chi = \{25, 35, 50, 70, 100, 250\}$ GeV, and six spin-independent WIMP-proton cross-sections $\sigma_p^{SI} = \{1.00 \times 10^{-8}, 3.98 \times 10^{-9}, 1.58 \times 10^{-9}, 6.31 \times 10^{-10}, 2.51 \times 10^{-10}, 1.00 \times 10^{-10}\}$ pb, thus 36 benchmark models in total. The number of dark matter recoil events above threshold for our Xe experiment (see section II B) for these benchmark points is in the range $10 \lesssim N_R \lesssim 4000$. As we focus on the case of a significant detection in a future experiment, we do not investigate the statistical properties of benchmark points in the very low counts regime, where $N_R < 10$, as it is hard to constrain much of anything with fewer than ~ 10 events.

Before we present results for our coverage study and the quantitative description of the performance of parameter estimation, we show examples of good and poor reconstructions of WIMP parameters based on the mock data sets of a specific benchmark point. These examples illustrate points that will be important in our coverage and performance studies.

Two examples of the reconstruction using Xe data are shown in Fig. 1 for a benchmark model with WIMP mass $m_\chi = 50$ GeV and spin-independent WIMP-proton cross-section $\sigma_p^{SI} = 2.51 \times 10^{-10}$ pb. This is an example of a benchmark point for which the performance of the reconstruction can vary strongly with the mock data. We show on the left of Fig. 1 an example of a “good” reconstruction (i.e., well constrained likelihood in the $m_\chi - \sigma_p^{SI}$ plane), and on the right of Fig. 1 an example of a “bad” reconstruction (leading to an essentially unconstrained

likelihood). For both cases we show the 68.3% and 95.4% likelihood contours (top) and the energy spectrum of the mock events (bottom), compared with the theoretical spectrum of the benchmark model (shown in black).

For the first example (left) both the 68.3% and the 95.4% confidence level spans a small range of masses and the benchmark point is well reconstructed. The distribution of the observed energies agrees well with the true benchmark rate. In contrast, the second example (right) leads to confidence levels that spread over a large mass range; at 95.4% confidence only a lower limit on the WIMP mass can be inferred (note that the 95.4% contour does not close, but is cut off at the upper mass prior limit $m_\chi = 1000$ GeV). The benchmark point is badly reconstructed mostly because of the presence of a relatively large number of high-energy counts at $E > 40$ keV. Events with these energies are an unlikely realisation of the benchmark WIMP spectrum, but can appear in the data due to statistical fluctuations. Poisson noise has flattened the observed energy spectrum relative to the predicted energy spectrum. The confidence intervals show “runaway” behaviour towards high mass because a flat energy spectrum is indicative of high masses, and the energy spectra for $m_\chi \gg m_N$ are nearly identical. As an example, the theoretical spectrum for a WIMP model with $m_\chi = 250$ GeV, $\sigma_p^{SI} = 6.31 \times 10^{-10}$ pb is shown in red in the bottom right panel. Clearly this model is a better fit to the simulated events than the benchmark model.

Note that this benchmark model leads to a large number of events ($N_R \sim 100$), so that one would naively expect that statistical fluctuations in the realised spectrum ought to have a minor impact. This is clearly not the case, as the bad reconstruction in the right panels of Fig. 1 shows that even with ~ 100 events, the parameter reconstruction can be poor. Even though we show in the rest of this section that this benchmark is relatively well-behaved—the coverage is exact for most intervals, the e.f.u. and bias are low, and the expected number of large-f.u. outliers is fairly small—there is a non-negligible probability that particular realisations of data sets for this benchmark lead to catastrophically poor WIMP parameter reconstructions.

B. Results from the coverage analysis

In order to investigate the coverage results for the 1D 68.3% and 95.4% confidence intervals for m_χ and σ_p^{SI} , for both Xe data and a combination of Xe+Ge data, we generate 1000 mock data sets for each of the 36 benchmark models, as outlined in section III. The 1D 68.3% (1σ) and 95.4% (2σ) confidence levels are constructed using Wilks’ theorem and we count how often the true value of the WIMP mass and cross-section are found within the stated CL. We further subdivide the 1000 reconstructions into 10 subsets, of 100 reconstructions each, and we compute the coverage for each subset. We take the standard

(MSE) for the parameters, given by the sum of the bias squared and the variance. We have found that the MSE behaves qualitatively similarly to the e.f.u., so we do not discuss it separately.

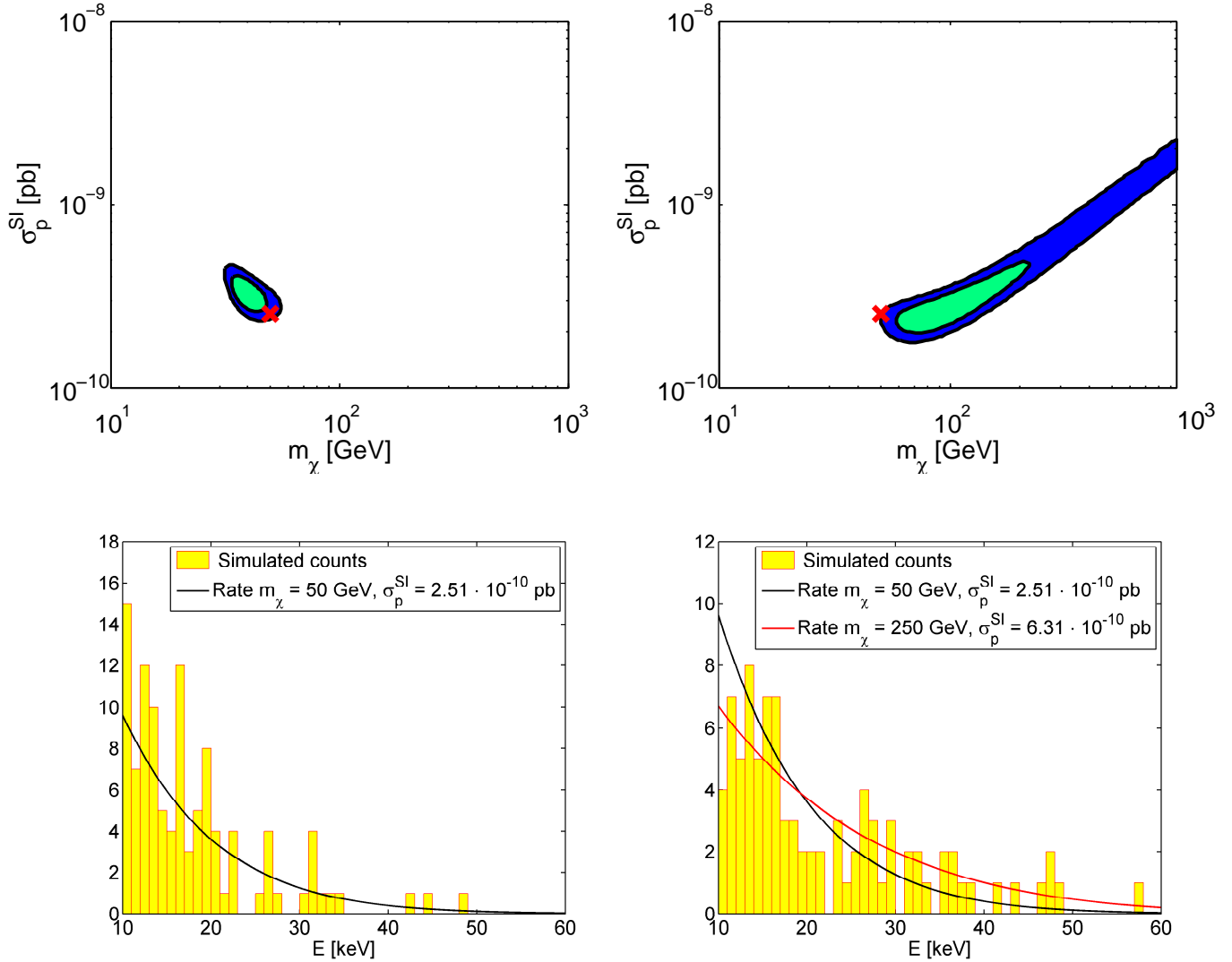


FIG. 1. The left (right) panels show examples for a good (bad) reconstruction of the WIMP benchmark model with true values $m_\chi = 50$ GeV, $\sigma_p^{SI} = 2.51 \times 10^{-10}$. The difference is exclusively in statistical fluctuations in the simulated data. Top panels: 68.3% and 95.4% confidence levels in the $m_\chi - \sigma_p^{SI}$ plane; the red cross shows the true value. Bottom panel: energy spectrum of the mock data (yellow histogram - recall that we use an unbinned likelihood function, the counts are binned for a better visualization), true rate $dR/dE(E)$ (black) and for the “bad” reconstruction an example of a rate (red) with a higher likelihood than the true rate.

error of these ten values to estimate the statistical error of our coverage analysis, encompassing the uncertainty coming from finite numerical samples of the likelihood and the finite number of reconstructions. Although this statistical error on the coverage value varies mildly across benchmark points, it is sufficient for our purposes to use its average over all benchmark points. This leads to an estimated 1σ error of 4.5% for the 68.3% intervals, and of 1.9% for the 95.4% intervals.

We start by discussing the 1D 68.3% and 95.4% confidence intervals for m_χ , shown in the top and bottom panels of Fig. 2, respectively. On the left-hand side we

show the coverage results obtained for a Xe target, on the right-hand side we show results for the combined data set Xe+Ge. From the above estimate of the error on the coverage, we define the coverage to be “exact” if it lies in the range (63.8, 72.8)% and (93.5, 97.3)% for the 68.3% and 95.4% contours, respectively. Benchmark points showing “exact” coverage within errors are displayed in green. Coverage values $> 72.8\%$ ($> 97.3\%$) correspond to over-coverage and are shown in red. Coverage values $< 63.8\%$ ($< 93.5\%$) correspond to under-coverage. However, none of the benchmark points studied here leads to under-coverage of any of the confidence

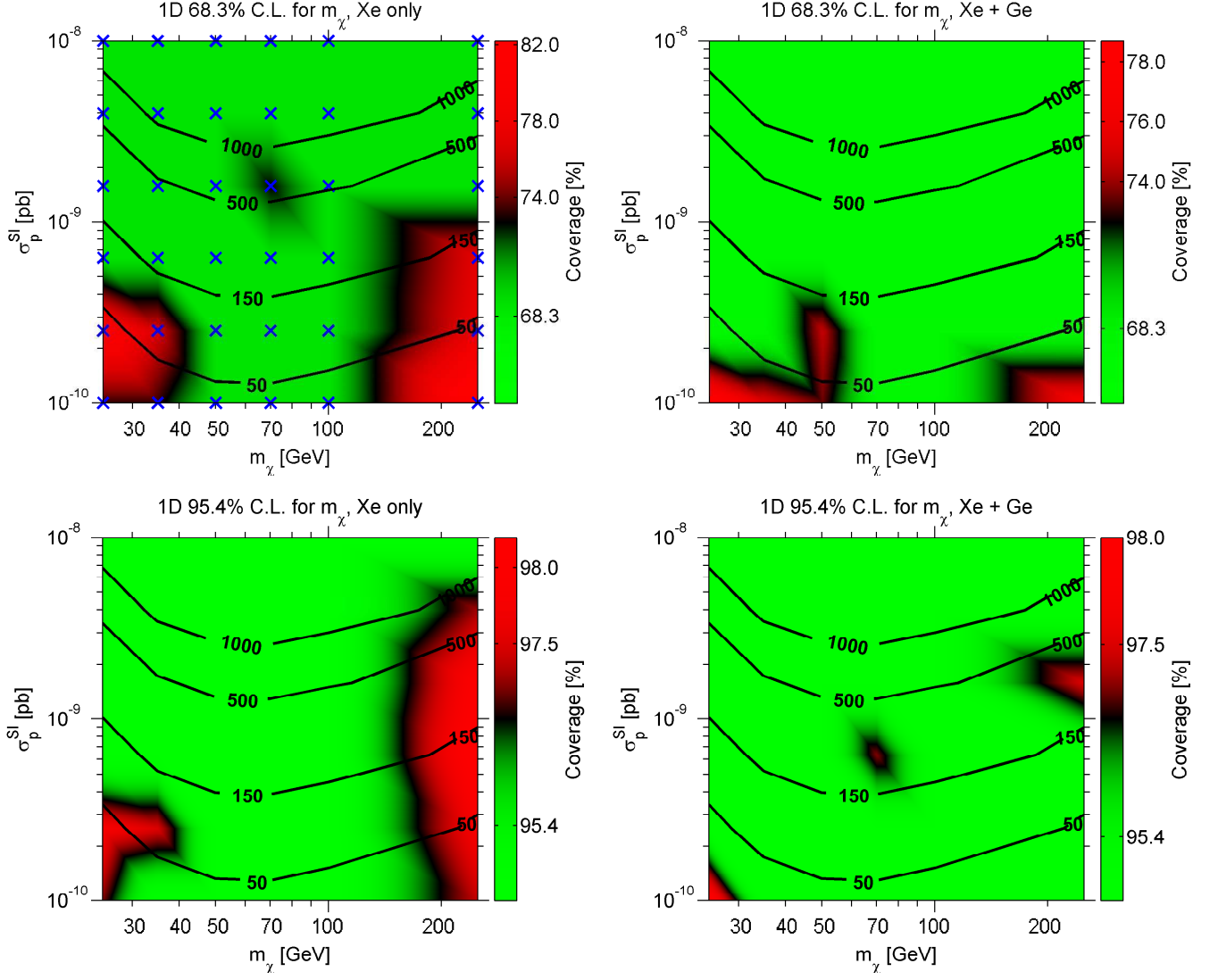


FIG. 2. Coverage results for the 1D 68.3% (top) and 95.4% (bottom) confidence interval for the WIMP mass in the $m_\chi - \sigma_p^{SI}$ plane, for simulated Xe target (left) and for a combination of Xe+Ge (right). Green (red) regions show “exact” coverage (over-coverage), as defined in the text. Black regions correspond to a transition from exact coverage to over-coverage. No under-coverage is observed. Isocontours of the expected number of counts in the Xe experiment are given in black. In the upper left plot, the benchmark points studied are indicated by blue crosses. The ‘flares’ pattern seen in some points are an artefact of the interpolation scheme used to generate the plots.

intervals. Benchmark points at the upper boundary of exact coverage or the lower boundary of over-coverage are displayed in black. For reference, isocontours of the expected number of counts N_R in a Xe experiment are also shown.

For the Xe-only case, we find that most benchmark points lead to exact coverage of the 1D 68.3% and 95.4% contours. For the 68.3% interval there is a region observed at high cross-sections and intermediate WIMP masses that borders on over-coverage; this is most likely the result of a statistical fluctuation. For both the 68.3% interval and the 95.4% interval, two regions leading to significant over-coverage can be identified, one at large

$m_\chi = 250$ GeV, and another at small $m_\chi = 25, 35$ GeV; both regions correspond to a small σ_p^{SI} . The over-coverage observed in the first region is a result of the high-mass degeneracy (for $m_\chi \gg m_N$, dR/dE_R depends only on $\sigma_p^{SI}/(\mu_p^2 m_\chi)$; refer to Sec. II A). The importance of this effect decreases with increasing cross-section because the slope of the energy spectrum is better resolved with more events, and hence is more sensitive to slight changes in v_{\min} . The high-mass degeneracy leads to a 1D profile likelihood that can no longer be well approximated by a Gaussian, such that the test statistic $\lambda(m_\chi)$ defined in Eq. (14) starts to deviate from a chi-square distribution. The difference between the histogram of

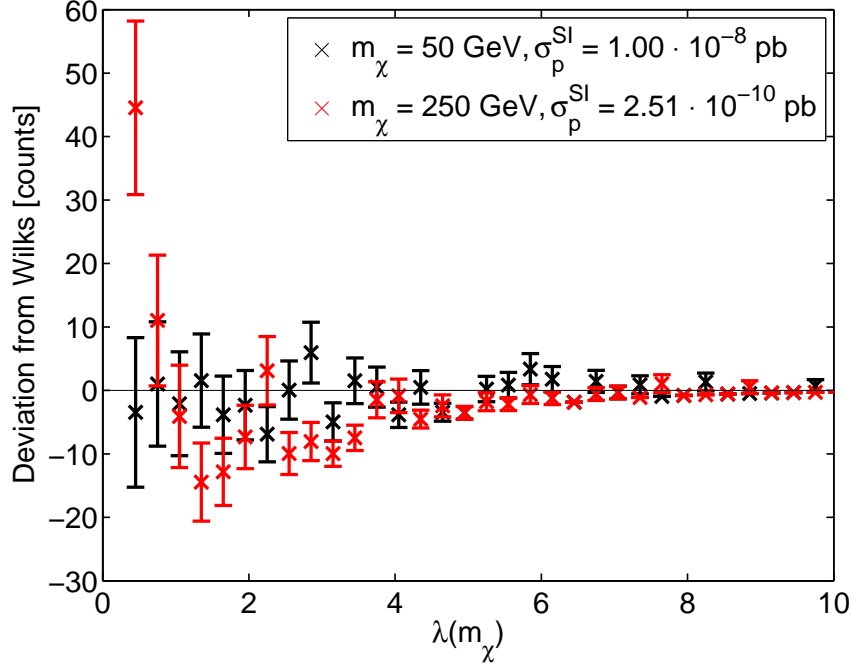


FIG. 3. Difference between the histogram of the profile likelihood test statistic $\lambda(m_\chi)$ from mock data sets and the value of the chi-square distribution with 1 degree of freedom (as predicted by Wilks' theorem) at the centre of each bin, as a function of $\lambda(m_\chi)$, for two different WIMP benchmark points.. This difference quantifies the deviation from Wilks' theorem for these two benchmark points. For each benchmark point, 10^3 realisations of mock data sets have been used to construct this histogram. Errorbars assume Poisson count statistics.

$\lambda(m_\chi)$ values from the mock data and the chi-square distribution with 1 degree of freedom (as predicted by Wilks' theorem) is shown in Fig. 3 for a high-mass benchmark point suffering from over-coverage ($m_\chi = 250$ GeV, $\sigma_p^{SI} = 2.51 \cdot 10^{-10}$ pb; see left-hand side of Fig. 2). For comparison, we also show the same quantity for a benchmark point where the agreement with the predicted chi-square distribution is much better ($m_\chi = 50$ GeV, $\sigma_p^{SI} = 10^{-8}$ pb), and whose coverage is exact to within errors. In contrast, for the high-mass point we observe significant discrepancies in the test statistics $\lambda(m_\chi)$ for values $\lesssim 4$, which explains why over-coverage is observed for this benchmark point.

The over-coverage observed at small m_χ and σ_p^{SI} is a result of the low number of counts for this benchmark model. Due to the low statistics in the region of parameter space the 1D profile likelihood is no longer well approximated by a Gaussian, hence the asymptotic behaviour of Wilks' theorem is less accurate. The deviation from Wilks' for these benchmark points is qualitatively similar to the red curve in Fig. 3, albeit less extreme.

Coverage improves when the Ge data are added to the analysis, as can be seen in the right panels of Fig. 2. Exact coverage is obtained in most of the parameter space. An exception is observed at $m_\chi = 70$ GeV,

$\sigma_p^{SI} = 6.31 \times 10^{-10}$ pb for the 95.4% plot, where slight over-coverage is found. Because neighbouring benchmark points are exactly covered, we interpret this as a statistical fluctuation. Both regions of over-coverage identified in the Xe-only case shrink significantly when adding Ge data to the analysis. For both the 68.3% and the 95.4% interval the over-coverage at large m_χ is almost completely eliminated, except at small σ_p^{SI} (for the 68.3% interval), for which the total number of expected events is $\mathcal{O}(10)$. For higher σ_p^{SI} , over-coverage of high-mass benchmark models is reduced since the likelihood is tighter for a combined analysis of Xe+Ge. The remaining over-coverage of the 95.4% interval at $m_\chi = 250$ GeV, $\sigma_p^{SI} = 1.58 \times 10^{-9}$ pb corresponds to a value of 97.5%, which is just above the border of exact coverage at 97.3%. However, at lower masses, especially for the 68.3% contour, over-coverage at very low cross-sections $\sigma_p^{SI} \approx 10^{-10}$ pb is not removed. In general, we find that the possibility of over-coverage remains as long as WIMP parameters are poorly constrained, which occurs most frequently for benchmark points which imply a low expected number of events. Both problems are resolved to some extent with the addition of data sets from a second experiment.

We display the results of our coverage analysis for

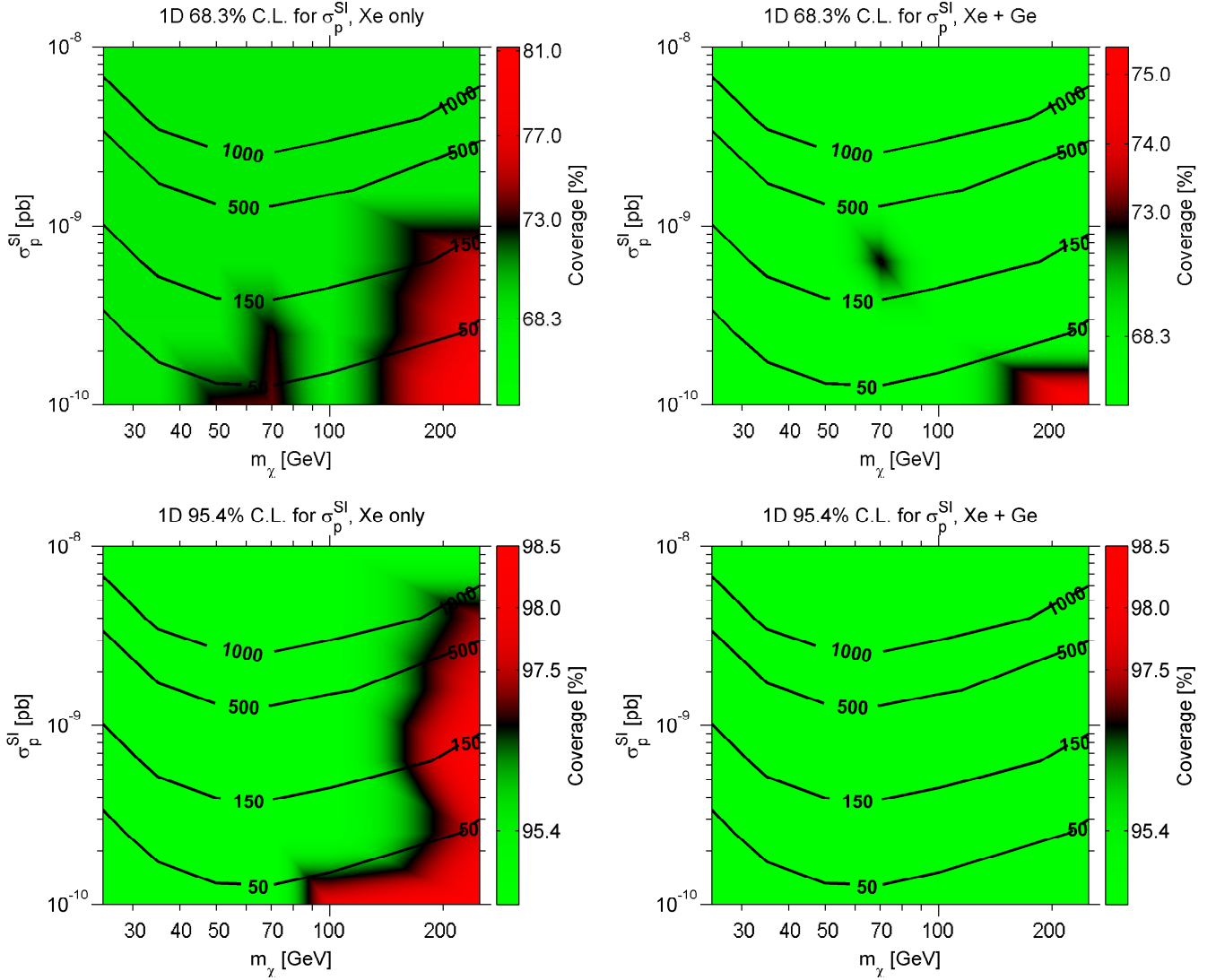


FIG. 4. As in Fig. 2, but for the 1D confidence intervals for σ_p^{SI} . A significant improvement in the coverage when combining Xe+Ge is apparent.

the 1D 68.3% and 95.4% confidence intervals for σ_p^{SI} in Fig. 4. The left-hand plot shows the results for a Xe target, the right-hand plot shows the results for combined Xe+Ge data. In the case in which we consider the Xe data alone, most of the parameter space corresponds to exact coverage, but for both the 1σ and the 2σ intervals a large region at high masses $m_\chi = 250$ GeV is over-covered. For the 95.4% interval this region is spread over almost the entire cross-section range, and extends to $m_\chi = 100$ GeV at low cross-sections. For the 68.3% interval a small region of over-coverage is found at intermediate WIMP masses $m_\chi = 50, 70$ GeV and low σ_p^{SI} . For the 95.4% contour the corresponding benchmark points systematically show a coverage percentage at least 1% above the exact value of 95.4%.

The over-coverage at large σ_p^{SI} is a result of the high-mass degeneracy, analogously to what has been explained

above for the mass. The over-coverage at intermediate WIMP masses can be explained using Fig. 1. Good reconstructions yield one dimensional profile likelihood functions that are approximately Gaussian, and thus lead to exact coverage. For bad reconstructions, the likelihood is spread over a larger range and thus the statement that σ_p^{SI} is over-covered for intermediate WIMP masses is a statement about the ratio of good to bad parameter fits. Due to low statistics resulting from the low number of counts the 1D profile likelihood function can no longer be well approximated by a chi-square distribution, Wilks' theorem becomes less accurate and over-coverage is observed. On the other hand, the over-coverage around 50 GeV WIMPs is not very significant, being close in magnitude to the numerical uncertainty of our coverage values, and therefore could be interpreted as a statistical fluke.

As with the WIMP mass, coverage improves with the

m_χ [GeV]	σ_p^{SI} [pb]	N_R	Coverage [%]			
			1D 68.3% m_χ	1D 95.4% m_χ	1D 68.3% σ_p^{SI}	1D 95.4% σ_p^{SI}
35	10^{-10}	29	73.3 (75.4)	96.1 (96.3)	69.2 (68.7)	96.9 (95.5)
50	10^{-10}	38	68.3 (73.5)	95.7 (96.3)	73.3 (71.2)	96.9 (96.8)
100	1.58×10^{-9}	527	70.3 (69.2)	96.0 (95.3)	68.9 (68.4)	94.9 (95.6)
250	10^{-8}	1671	68.0 (66.7)	95.9 (94.9)	69.2 (67.6)	95.7 (95.2)

TABLE II. Results of the coverage analysis of the 1D confidence intervals for four selected benchmark points. Results for the Xe data alone are given, as well as for the combined analysis of Xe+Ge (in parenthesis).

addition of data from a Ge target (right plots in Fig. 4). For the 68.3% contour the over-covered region at intermediate $m_\chi = 50, 70$ GeV vanishes completely and is now exactly covered (apart from what can again be interpreted as a statistically non-significant fluctuation around 70 GeV, which appears as a ‘flare’ pattern in the figure). The over-covered region at high WIMP masses $m_\chi = 250$ GeV shrinks significantly, but is difficult to eliminate at low cross-sections $\sigma_p^{SI} = 10^{-10}$ pb, as discussed above. The improvement in the coverage is even greater for the 2σ contour. For a combined analysis of data from Xe+Ge the over-coverage observed for the Xe target completely vanishes; the entire parameter space is exactly covered. The coverage results for a selected subset of benchmark points are shown in Table II.

Overall, our coverage analysis concludes that the approximate confidence intervals for the studied benchmark points either cover exactly or over-cover the true values of the parameters – i.e., they are conservative. The two most important effects at play are the large mass degeneracy, and strong statistical fluctuations that are important even for a relatively large numbers of expected counts (~ 100). We have shown that addition of data from a second target such as Ge leads to significant improvement on both fronts. We point out that the observed over-coverage can in principle be remedied using methods such as Feldman-Cousins to build confidence intervals with guaranteed exact coverage.

We have also investigated coverage properties of the credible intervals obtained from the Bayesian posterior. For well-reconstructed benchmark points, credible intervals are numerically identical to confidence intervals, since we have taken flat priors on our WIMP parameters of interest, so their coverage properties are the same. However, for badly reconstructed points (i.e., lying on the high-mass degeneracy line) the posterior is cut off at large masses and cross-sections by the prior range. This means that the ensuing 1D marginal posterior and thus also the credible intervals become a function of the prior range adopted for the mass and cross section, which is clearly unsatisfactory (this effect has also been pointed out in another context by Ref. [54]). As a consequence, the coverage of Bayesian credible intervals exhibits broadly the same trends as highlighted above for the frequentist intervals, but also shows a tendency towards under-coverage in some regions. As those results are however sensitive to the choice of prior range, we do not present

coverage results for Bayesian credible intervals in this work – a thorough exploration of this issue would require a study of how such properties change as a function of the prior ranges chosen. We emphasize however that the prior ranges have no impact on our results for the frequentist confidence intervals.

C. Accuracy and precision of parameter reconstruction

We now consider the question of the accuracy and precision of the parameter reconstruction. We start by investigating the expected fractional uncertainty (e.f.u.) for m_χ , introduced in section IIID. The e.f.u. quantifies the average fractional standard deviation of the reconstructed WIMP mass value and thus is a measure of the precision of the reconstruction. We show the e.f.u. in the $m_\chi - \sigma_p^{SI}$ plane in Fig. 5 (notice that the upper limit of the colorbar is set to e.f.u. = 1.5 for display purposes, but this limit is surpassed in many cases). Isocontours of the expected number of counts in a Xe target are shown in black. Isocontours of the number of “bad” cases (i.e., with an f.u. > 0.75) are shown in white. Considering the number of “bad” cases is very important, since this number quantifies the probability that, for a given WIMP benchmark point (that may lead to a reasonably small *average* uncertainty on m_χ), the experiment results in a data set that leaves the WIMP mass essentially unconstrained.

High-mass benchmark points lead to a likelihood function with a long tail in the $m_\chi - \sigma_p^{SI}$ plane, and thus are expected to have a very high e.f.u.. We are most interested in the region where the transition from good to poor performance takes place.

We will first discuss the e.f.u. results from Xe data only. As a general pattern, the larger m_χ and the smaller σ_p^{SI} , the larger the e.f.u. value for the benchmark point. We will discuss the e.f.u. results at high ($\sigma_p^{SI} = 10^{-8}$ pb), intermediate ($\sigma_p^{SI} = 10^{-9}$ pb) and low ($\sigma_p^{SI} = 10^{-10}$ pb) cross-sections.

At high ($\sigma_p^{SI} = 10^{-8}$ pb) cross-sections, most benchmark masses lead to a small e.f.u., and thus a small uncertainty in the reconstructed WIMP mass. The e.f.u. does not exceed 0.15 for $m_\chi \leq 100$ GeV and is significantly smaller for small $m_\chi = 25, 35$ GeV (e.f.u. = 0.03). The fraction of bad reconstructions is < 1%. However,

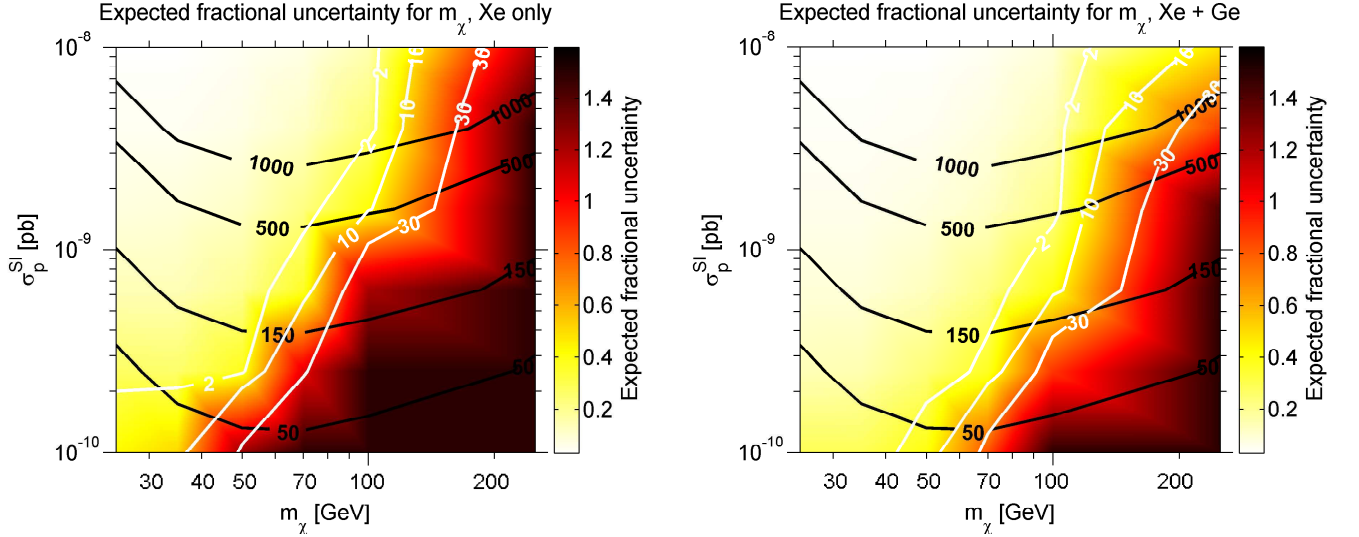


FIG. 5. Expected fractional uncertainty (e.f.u.) for the WIMP mass in the $m_\chi - \sigma_p^{SI}$ plane, for a Xe (Xe+Ge) target in the left (right) plot, quantifying the precision of the mass reconstruction (low e.f.u. corresponding to better precision). Isocontours of the expected number of counts in the Xe experiment are given in black; isocontours of the percentage of “bad” reconstruction (f.u. > 0.75) are shown in white.

even for this large cross-section and the resulting large number of events, $N_R = 1671$, the high-mass benchmark point $m_\chi = 250$ GeV leads to an e.f.u. > 1.00 . Such a large e.f.u. means that the WIMP mass is left essentially unconstrained by the data, and the confidence levels inhabit the region of degeneracy at high masses and cross-sections.

For intermediate benchmark cross-sections ($\sigma_p^{SI} = 10^{-9}$ pb), the overall precision is quite good. For benchmark masses $m_\chi \leq 70$ GeV the e.f.u. is < 0.30 and the WIMP mass is well constrained. This is also reflected in the number of bad reconstructions: for $m_\chi \leq 50$ GeV this number is $< 1\%$; for $m_\chi = 70$ GeV only a couple of bad cases occur for 100 reconstructions. At higher m_χ the e.f.u. increases rapidly. For example, at $m_\chi = 100$ GeV the e.f.u. increases from 0.41 to 1.21 when decreasing the cross-section from $\sigma_p^{SI} = 1.58 \times 10^{-9}$ (corresponding to $N = 527$ events) to $\sigma_p^{SI} = 6.31 \times 10^{-10}$ (corresponding to $N = 210$ events). Therefore, at $\sigma_p^{SI} = 10^{-9}$ this benchmark point lies on the borderline between good and bad performance of the reconstruction. At cross-sections $\sigma_p^{SI} \leq 10^{-9}$ and high WIMP masses ($m_\chi \geq 100$ GeV), the e.f.u. is systematically > 0.75 (sometimes $\gg 0.75$), meaning that the WIMP mass becomes essentially unconstrained in 20% or more of the reconstructions. This is to be expected, due to the $m_\chi - \sigma_p^{SI}$ degeneracy that occurs at high masses. However, it is interesting to see how pronounced this effect is even at a relatively small mass ($m_\chi \approx 100$ GeV).

The situation deteriorates significantly for $\sigma_p^{SI} = 10^{-10}$ pb, leading to a small number of counts [$\mathcal{O}(10)$] for all m_χ . This is reflected in the e.f.u., which is of order ~ 0.50 for small $m_\chi = 25, 35$ GeV. This corresponds

to weak constraints on the WIMP mass, and leads to an average uncertainty of more than 100% for $m_\chi \geq 50$ GeV. Similarly, while for small WIMP masses just above 5% of all reconstructions are bad, this number is significantly higher for high-mass WIMP models. Even for an intermediate $m_\chi = 50$ GeV, $\sim 30\%$ of reconstructions are bad. We emphasize once more that this is due to statistical fluctuations in the realization of the energy spectrum, and therefore an unavoidable effect.

As expected, the e.f.u. improves considerably with the addition of data from a Ge target. For fixed cross section, the 30% bad reconstruction isocontour shifts to higher mass values by $\sim 50\%$ with respect to the reconstruction with Xe data alone. Because the e.f.u. is correlated with the percentage of poor reconstructions, we also see that it decreases dramatically at fixed WIMP parameters (often by $> 50\%$) with the inclusion of the Ge data.

Fig. 6 shows the value of the e.f.u. as a function of the exposure ϵ for a WIMP with cross-section $\sigma_p^{SI} = 10^{-9}$ pb and for three different benchmark masses. Solid lines correspond to the e.f.u. from a Xe target only, dashed lines show results for combining data from a Xe and a Ge experiment. For the Xe only case, for massive WIMPs ($m_\chi = 250$ GeV), the expected fractional uncertainty is always greater than unity, as a consequence of the degeneracy. For intermediate ($m_\chi = 50$ GeV) and small mass WIMPs ($m_\chi = 25$ GeV), the e.f.u. drops sharply with increasing exposure. In particular, it is still of order $\sim 30 - 40\%$ for an exposure of 1 ton \times year, and it is reduced to less than 10% for a Xe experiment with exposure 10 ton \times year. When combining Xe + Ge data the situation improves for all benchmark masses. For massive WIMPs ($m_\chi = 250$ GeV) an e.f.u. smaller than unity can be achieved for a Xe experiment with expo-

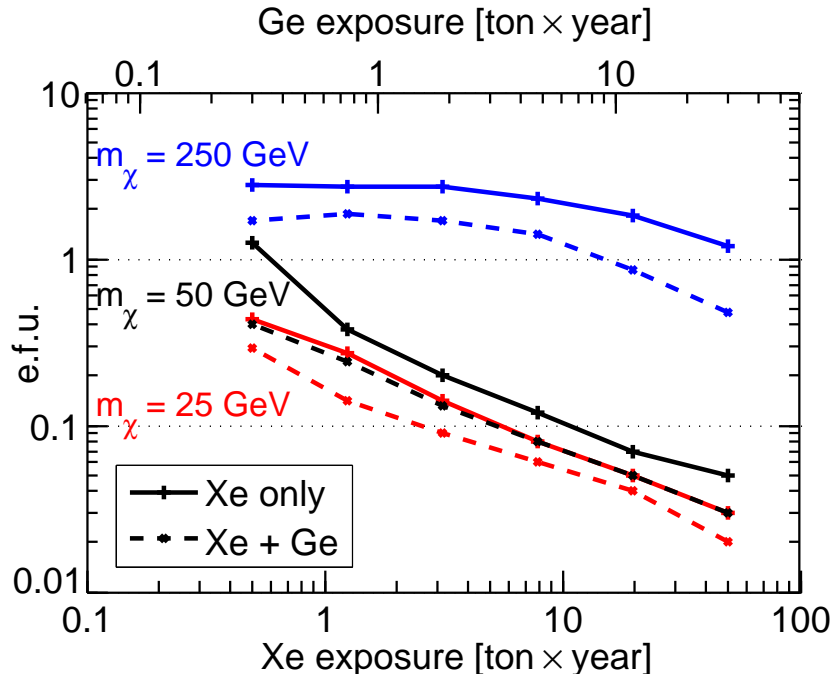


FIG. 6. Expected fractional uncertainty (e.f.u.) on the WIMP mass as a function of exposure for a Xenon experiment (bottom axis) and a Germanium experiment (top axis) required to achieve this e.f.u. for a WIMP with cross-section $\sigma_p^{SI} = 10^{-9}$, for three different benchmark masses $m_\chi = 25$ GeV (red), $m_\chi = 50$ GeV (black) and $m_\chi = 250$ GeV (blue). Solid lines correspond to e.f.u. results for Xe only, dashed lines correspond to e.f.u. results for a Xe + Ge target.

sure ~ 20 ton \times year and a Ge experiment with exposure ~ 10 ton \times year. For larger exposures the e.f.u. further decreases. For both intermediate ($m_\chi = 50$ GeV) and small ($m_\chi = 25$ GeV) WIMP masses the e.f.u. for Xe + Ge is significantly smaller than in the Xe only case. The e.f.u. strongly decreases as the exposures of the Xe and Ge targets are increased. In particular, for an intermediate (low) mass WIMP an expected fractional uncertainty of less than 10% can be achieved for a 3 (1.5) ton \times year exposure for Ge and a 5 (3) ton \times year exposure for Xe. These trends are qualitatively consistent with those found by Refs. [55, 56].

However, we caution that the e.f.u. will be higher in reality for a fixed exposure and benchmark point, because of astrophysical and nuclear-physics uncertainties.

The fractional mass bias in the $m_\chi - \sigma_p^{SI}$ plane for a Xe target (Xe and Ge target) is displayed on the left (right) of Fig. 7. Almost no negative bias in the mass is observed. If a bias exists, it typically goes in the direction of a larger m_χ than the true value, as a consequence of the high mass-cross section degeneracy. In fact, the distribution of reconstructions that reach up onto the degeneracy curve explains the features of Fig. 7. In comparing Figs. 5 and 7, we find that the curve for e.f.u. = 0.8 corresponds closely to the curve of bias = 0.2. When a large fraction of reconstructions are bad, both the e.f.u. and bias increase because the high mass-cross section curve

becomes populated with high-likelihood fits. The extension of the confidence levels to this region of the parameter space means that the best-fit mass is typically higher than the true mass, so that both the uncertainty in the mass and its bias become large.

The performance of the statistical reconstruction (as quantified by the e.f.u., the number of bad cases and the fractional bias in the WIMP mass) is summarised for four benchmark points in Table III.

D. Comparison with other coverage studies

We have focused on reconstructing phenomenological WIMP-related variables (mass, spin-independent cross section) rather than theoretical parameters in specific theories for WIMP physics. Perhaps not surprisingly, our results differ from recent studies of the coverage properties of parameters of specific supersymmetric models from particle-physics experiments, including direct-detection data [28, 29]. Ref. [29] found that supersymmetric parameters were consistently over-covered when attempting to reconstruct the ‘SU3’ benchmark point with mock ATLAS data on sparticle masses and mass splittings. In contrast, consistent (and sometimes drastic) under-coverage was observed [28] for two different benchmark points reconstructed using mock ton-scale di-

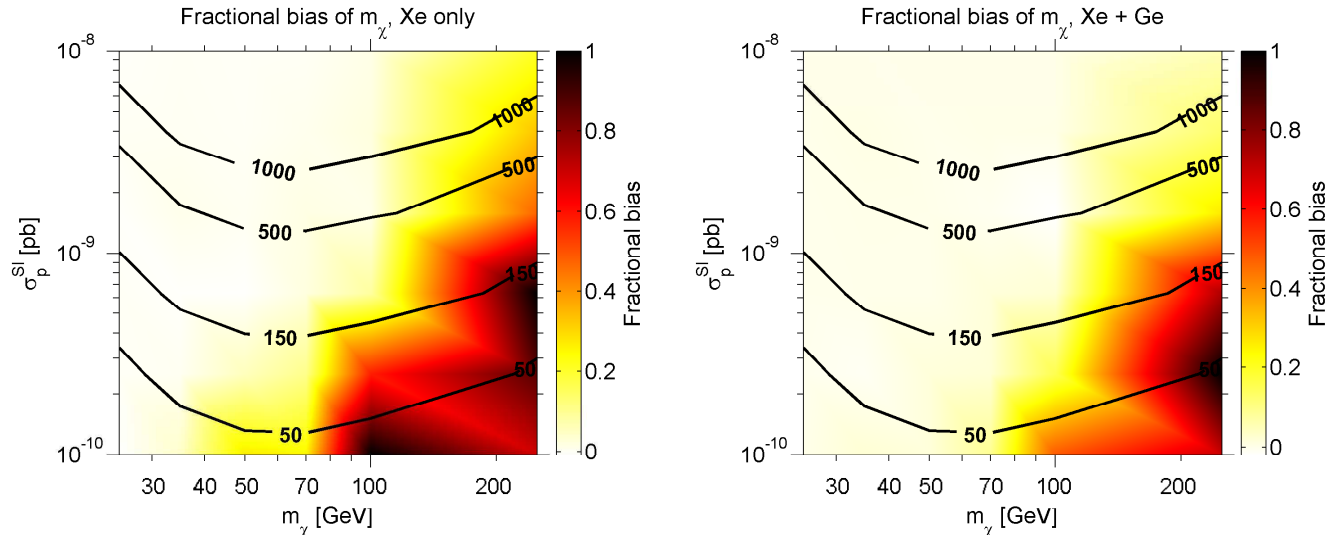


FIG. 7. As in Fig.5, but for the fractional bias of the WIMP mass, i.e. the bias of the WIMP mass relative to the benchmark mass (notice that almost no negative bias is observed).

m_χ [GeV]	σ_p^{SI} [pb]	N_R	e.f.u.	# bad cases	fractional bias for m_χ
35	10^{-10}	29	0.51 (0.29)	7 (0)	0.042 (0.023)
50	10^{-10}	38	1.24 (0.40)	32 (4)	0.272 (0.017)
100	1.58×10^{-9}	527	0.41 (0.22)	9 (0)	0.014 (-0.020)
250	10^{-8}	1671	1.20 (0.48)	51 (13)	0.205 (0.052)

TABLE III. Summary of the performance of the statistical reconstruction four selected WIMP benchmark models. The benchmark (true) mass and cross-section and the corresponding number of counts for the Xe experiment are shown. We give the expected fractional uncertainty, the number of “bad” (f.u. > 0.75) cases and the fractional bias in m_χ for the Xe data alone and for the combined analysis of Xe+Ge (in parenthesis).

rect detection data.

Here, we observed exact coverage in a large portion of the phenomenological parameter space we investigated. Unlike in supersymmetric analyses, the parameter space considered here does not include complicated theoretical boundaries where the likelihood function is not defined. Substantial over-coverage is therefore not expected in our results for cases with reasonable statistics (i.e. where Wilks’ Theorem does not break down simply due to low-number statistics). Furthermore, the relationship between parameters of interest (here, WIMP mass and cross-section) and observables (i.e., counts) is far simpler here than when one works with fundamental supersymmetric parameters (which are connected to observables via complex, non-linear Renormalization Group Equations that make the likelihood function highly non-Gaussian in the parameters). Therefore, sampling issues that might plague supersymmetric parameter spaces and lead to under-coverage are not observed in our setup.

Taking the results of all three studies together, we conclude that coverage properties are good when the scanning is done over a set of parameters that have a simple mapping to the observables (as was seen in [29]). As the observables on which a (typically approximately Gaus-

sian) likelihood function is defined become a highly complicated function (i.e. via highly non-linear transformations) of the parameters of interest, the coverage becomes less exact, and a detailed numerical investigation is required to establish the coverage properties. The upshot of this for dark matter searches is that simple model-independent analyses using phenomenological particle-physics parameters for WIMPs can generally be expected to have good coverage, but the mapping onto specific model spaces will typically not retain this property.

V. CONCLUSIONS

We have studied the statistical properties of approximate confidence intervals on WIMP parameters, using mock data from future ton-scale direct detection experiments. We have focused in particular on the effect of unavoidable statistical fluctuations in the data. Contrary to what has been observed in GUT-scale SUSY parameterisations, we see that coverage for phenomenological WIMP parameters (mass, cross-section) is generally quite good. We have observed a small amount of over-coverage for certain benchmark points, i.e. the con-

structed confidence intervals are conservative. We have traced this over-coverage back to either statistical fluctuations, which become most important for benchmark points leading to a low expected number of counts, or to the degeneracy between the WIMP mass and cross-section, that occurs at large WIMP masses in the likelihood function. In both cases the profile likelihood is not well approximated by a Gaussian, such that Wilks' theorem no longer accurately described the behaviour of the test statistics $\lambda(m_\chi)$ and $\lambda(\sigma_{SI})$. This problem is much less severe than in the SUSY case; in general, it appears that the less complicated and nonlinear a function the likelihood is of the underlying parameter space, the better the coverage properties. Finally, we remind the reader that coverage issues can in principle be resolved altogether by constructing intervals that have exact coverage, e.g. by using the Feldman-Cousins method.

We have found that the statistical bias and expected fractional uncertainty of the reconstructed WIMP mass and cross-section are more serious problems, which cannot be resolved by employing a different method of constructing confidence intervals. The parameter reconstruction can be ruined by statistical fluctuations that flatten the observed energy recoil spectrum with respect to the true underlying model, leading to an essentially unconstrained likelihood function, so that only a lower limit can be placed on the WIMP mass and cross-section. This was found to be important even at intermediate WIMP masses and cross-sections. Therefore, even for benchmark models leading to a relative large expected number of counts ($\gtrsim \mathcal{O}(100)$), statistical fluctuations can result in a strong bias (i.e. low accuracy) and a low precision of the reconstruction of the WIMP parameters.

We have shown that a combination of data sets from two independent experiments with different target materials can significantly improve the coverage properties,

reduce the bias and increase the accuracy and precision of the reconstruction. Furthermore, we have shown that the precision of the reconstruction can be improved considerably if the exposure of the experiment(s) is increased.

Our investigation has assumed negligible backgrounds and fixed important sources of uncertainties, such as astrophysical quantities describing the local dark matter distribution. Our modelling of the experimental likelihood was correspondingly simplified. Therefore, the large bias and low precision of the reconstructed parameters discovered for a number of benchmark models is a fundamental result of statistical fluctuations in the realisation of the energy spectrum. We expect that including the energy resolution, non-negligible backgrounds and astrophysical uncertainties in the analysis would further degrade the performance of the reconstruction.

ACKNOWLEDGMENTS

We would like to thank Miguel Pato, Alastair Currie and Henrique Araujo for useful discussions. C.S. and R.T. thank the University of Zurich for hospitality. R.T. would like to thank Kavli Institute for Theoretical Physics at the UCSB for hospitality. C.S. would like to thank the McGill High Energy Theory Group for hospitality. C.S. is partially supported by a scholarship of the "Studienstiftung des deutschen Volkes". This research was supported in part by the National Science Foundation under Grant No. NSF PHY05-51164. A.H.G.P. is supported by a Gary McCue Fellowship through the Center for Cosmology at UC Irvine and by NASA Grant No. NNX09AD09G. P.S. is supported by the Lorne Trotter Chair in Astrophysics and an Institute for Particle Physics Theory Fellowship. G.B. is supported by a European Research Council Starting Grant, under grant agreement No. 277591.

-
- [1] G. Bertone, ed., *Particle dark matter: Observations, models and searches* (Cambridge University Press, 2010).
 - [2] L. Bergstrom, Rept. Prog. Phys. **63**, 793 (2000), arXiv:hep-ph/0002126 [hep-ph].
 - [3] C. Munoz, Int. J. Mod. Phys. A **19**, 3093 (2004), arXiv:hep-ph/0309346 [hep-ph].
 - [4] G. Bertone, D. Hooper, and J. Silk, Phys. Rept. **405**, 279 (2005), arXiv:hep-ph/0404175 [hep-ph].
 - [5] G. Steigman and M. S. Turner, Nuclear Physics B **253**, 375 (1985).
 - [6] K. Griest, Phys. Rev. D **38**, 2357 (1988).
 - [7] G. Jungman, M. Kamionkowski, and K. Griest, Phys. Rep. **267**, 195 (1996), arXiv:hep-ph/9506380.
 - [8] T. Appelquist, H.-C. Cheng, and B. A. Dobrescu, Phys. Rev. D **64**, 035002 (2001), arXiv:hep-ph/0012100.
 - [9] G. Servant and T. M. P. Tait, New J. Phys. **4**, 99 (2002), arXiv:hep-ph/0209262.
 - [10] H.-C. Cheng, J. L. Feng, and K. T. Matchev, Phys. Rev. Lett. **89**, 211301 (2002), arXiv:hep-ph/0207125.
 - [11] M. Pato, L. Baudis, G. Bertone, R. Ruiz de Austri, L. E. Strigari, *et al.*, Phys. Rev. D **83**, 083505 (2011), arXiv:1012.3458 [astro-ph.CO].
 - [12] R. Bernabei, P. Belli, F. Cappella, R. Cerulli, C. Dai, *et al.*, Eur. Phys. J. C **67**, 39 (2010), arXiv:1002.1028 [astro-ph.GA].
 - [13] C. Aalseth *et al.* (CoGeNT collaboration), Phys. Rev. Lett. **106**, 131301 (2011), arXiv:1002.4703 [astro-ph.CO].
 - [14] C. Arina, J. Hamann, R. Trotta, and Y. Y. Wong, (2011), arXiv:1111.3238 [hep-ph].
 - [15] G. Angloher, M. Bauer, I. Bavykina, A. Bento, C. Bucci, *et al.*, (2011), arXiv:1109.0702 [astro-ph.CO].
 - [16] E. Aprile *et al.* (XENON100 Collaboration), Phys. Rev. Lett. **105**, 131302 (2010), arXiv:1005.0380 [astro-ph.CO].
 - [17] E. Aprile *et al.* (XENON100), Phys. Rev. Lett. **107**, 131302 (2011), arXiv:1104.2549 [astro-ph.CO].

- [18] Z. Ahmed *et al.* (The CDMS-II Collaboration), Science **327**, 1619 (2010), arXiv:0912.3592 [astro-ph.CO].
- [19] Z. Ahmed *et al.* (CDMS-II Collaboration), Phys. Rev. Lett. **106**, 131302 (2011), arXiv:1011.2482 [astro-ph.CO].
- [20] E. Armengaud *et al.* (EDELWEISS Collaboration), Phys. Lett. **B702**, 329 (2011), arXiv:1103.4070 [astro-ph.CO].
- [21] D. Akimov, H. Araujo, E. Barnes, V. Belov, A. Bewick, *et al.*, (2011), arXiv:1110.4769 [astro-ph.CO].
- [22] A. M. Green, JCAP **0807**, 005 (2008), arXiv:0805.1704 [hep-ph].
- [23] L. E. Strigari and R. Trotta, JCAP **11**, 19 (2009), arXiv:0906.5361 [astro-ph.HE].
- [24] Y. Akrami, C. Savage, P. Scott, J. Conrad, and J. Edsjö, JCAP **4**, 12 (2011), arXiv:1011.4318 [astro-ph.CO].
- [25] A. H. G. Peter, Phys. Rev. D **83**, 125029 (2011), arXiv:1103.5145 [astro-ph.CO].
- [26] S. S. Wilks, The Annals of Mathematical Statistics **9**, pp. 60 (1938).
- [27] G. J. Feldman and R. D. Cousins, Phys. Rev. D **57**, 3873 (1998).
- [28] Y. Akrami, C. Savage, P. Scott, J. Conrad, and J. Edsjö, JCAP **1107**, 002 (2011), arXiv:1011.4297 [hep-ph].
- [29] M. Bridges, K. Cranmer, F. Feroz, M. Hobson, R. R. de Austri, *et al.*, JHEP **1103**, 012 (2011), arXiv:1011.4306 [hep-ph].
- [30] A. H. G. Peter, Phys. Rev. D **81**, 087301 (2010), arXiv:0910.4765.
- [31] S. D. McDermott, H.-B. Yu, and K. M. Zurek, (2011), arXiv:1110.4281 [hep-ph].
- [32] J. R. Ellis, K. A. Olive, and C. Savage, Phys. Rev. D **77**, 065026 (2008), arXiv:0801.3656 [hep-ph].
- [33] S. Chang, J. Liu, A. Pierce, N. Weiner, and I. Yavin, JCAP **1008**, 018 (2010), arXiv:1004.0697 [hep-ph].
- [34] J. L. Feng, J. Kumar, D. Marfatia, and D. Sanford, Phys. Lett. B **703**, 124 (2011), arXiv:1102.4331 [hep-ph].
- [35] Z. Kang, T. Li, T. Liu, C. Tong, and J. M. Yang, JCAP **1101**, 028 (2011), arXiv:1008.5243 [hep-ph].
- [36] M. Pato, JCAP **1110**, 035 (2011), arXiv:1106.0743 [astro-ph.CO].
- [37] J. D. Lewin and P. F. Smith, Astroparticle Physics **6**, 87 (1996).
- [38] J. I. Read, L. Mayer, A. M. Brooks, F. Governato, and G. Lake, Mon. Not. R. Astron. Soc. **397**, 44 (2009), arXiv:0902.0009.
- [39] M. Vogelsberger, A. Helmi, V. Springel, S. D. M. White, J. Wang, C. S. Frenk, A. Jenkins, A. Ludlow, and J. F. Navarro, Mon. Not. R. Astron. Soc. **395**, 797 (2009), arXiv:0812.0362.
- [40] M. Kuhlen, N. Weiner, J. Diemand, P. Madau, B. Moore, D. Potter, J. Stadel, and M. Zemp, JCAP **2**, 30 (2010), arXiv:0912.2358 [astro-ph.GA].
- [41] M. Lisanti, L. E. Strigari, J. G. Wacker, and R. H. Wechsler, Phys. Rev. D **83**, 023519 (2011), arXiv:1010.4300 [astro-ph.CO].
- [42] E. Aprile, Journal of Physics Conference Series **308**, 012010 (2011).
- [43] L. Baudis (for the DARWIN Consortium), (2012), arXiv:1201.2402 [astro-ph.IM].
- [44] E. Figueroa-Feliciano, in *American Institute of Physics Conference Series*, Vol. 1200, edited by G. Alverson, P. Nath, & B. Nelson (2010) pp. 959–962.
- [45] E. Aprile *et al.* (XENON100 Collaboration), Phys. Rev. D **84**, 052003 (2011), arXiv:1103.0303 [hep-ex].
- [46] Z. Ahmed *et al.* (CDMS Collaboration), Phys. Rev. Lett. **102**, 011301 (2009), arXiv:0802.3530 [astro-ph].
- [47] R. Trotta, Contemp. Phys. **49**, 71 (2008), arXiv:0803.4089 [astro-ph].
- [48] F. Feroz, K. Cranmer, M. Hobson, R. Ruiz de Austri, and R. Trotta, JHEP **1106**, 042 (2011), arXiv:1101.3296 [hep-ph].
- [49] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, The Journal of Chemical Physics **21**, 1087 (1953).
- [50] W. K. Hastings, Biometrika **57**, 97 (1970).
- [51] J. Cohen, *Statistical power analysis for the behavioral sciences* (L. Erlbaum Associates, 1988).
- [52] R. Bausell and Y. Li, *Power Analysis for Experimental Research: A Practical Guide for the Biological, Medical and Social Sciences* (Cambridge University Press, 2006).
- [53] K. Murphy and B. Myors, *Statistical power analysis: a simple and general model for traditional and modern hypothesis tests*, Inquiry and Pedagogy Across Diverse Contexts Series (L. Erlbaum Associates, 2004).
- [54] C. Arina, J. Hamann, and Y. Y. Wong, JCAP **1109**, 022 (2011), arXiv:1105.5121 [hep-ph].
- [55] A. M. Green, JCAP **7**, 5 (2008), arXiv:0805.1704.
- [56] M. Drees and C.-L. Shan, JCAP **6**, 12 (2008), arXiv:0803.4477.