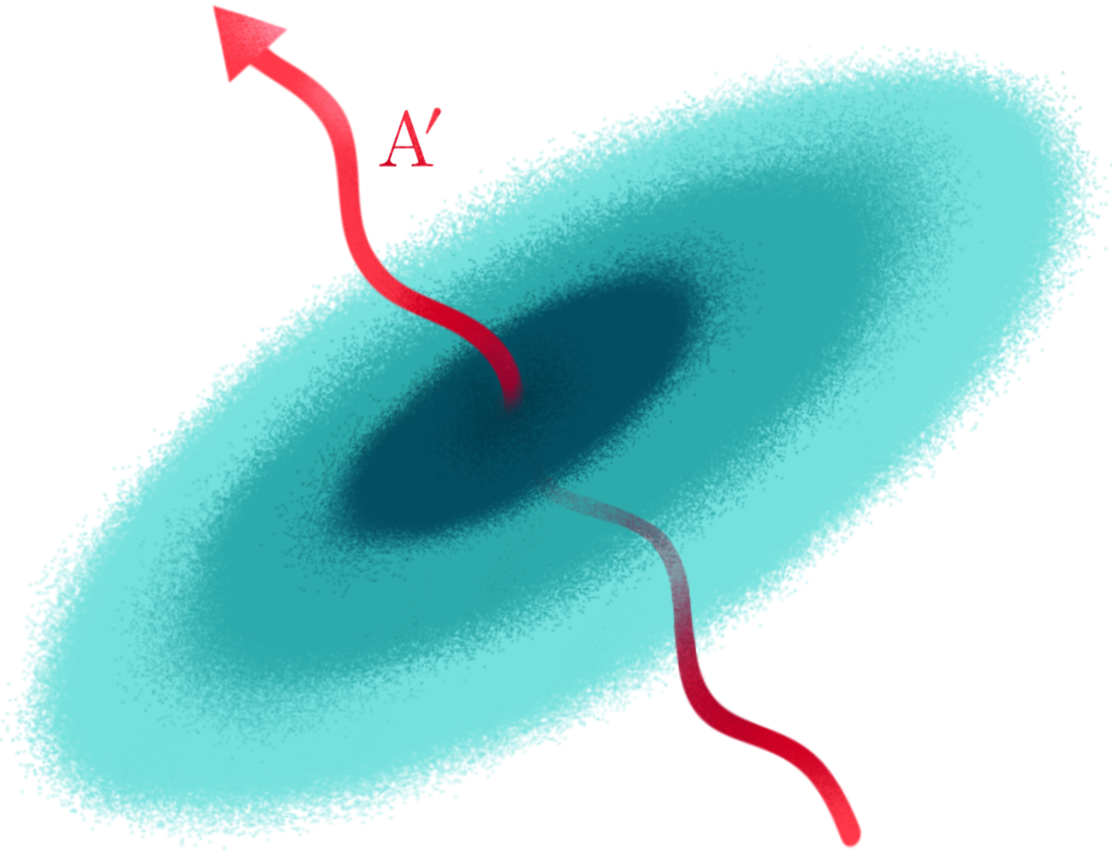
CHALMERS
UNIVERSITY OF TECHNOLOGY




# A Bayesian Framework for Dark Photon Searches at LDMX

Parameter Inference and Interaction Model Comparison

Master's thesis in Physics (MPPHS)

ADAM BERGER

DEPARTMENT OF PHYSICS AND ASTRONOMY

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

# A Bayesian Framework for Dark Photon Searches at LDMX

Parameter Inference and Interaction Model Comparison

Adam Berger

Department of Physics and Astronomy
*Division of Subatomic, High-Energy, and Plasma Physics*
Theoretical Subatomic Physics
Chalmers University of Technology
Gothenburg, Sweden 2026

A Bayesian Framework for Dark Photon Searches at LDMX:
Parameter Inference and Interaction Model Comparison

Adam Berger



Supervisors: Prof. Riccardo Catena, Taylor Gray
Examiner: Prof. Riccardo Catena

Master's Thesis 2026
Department of Physics and Astronomy
Division of Subatomic, High-Energy, and Plasma Physics
Theoretical Subatomic Physics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone: +46 31 772 1000

Cover: Designed by Vivien Dumm.

Typeset in LaTeX
Printed by Chalmers Digital Printing
Gothenburg, Sweden 2026

A Bayesian Framework for Dark Photon Searches at LDMX:
Parameter Inference and Interaction Model Comparison

Adam Berger
Department of Physics and Astronomy
Chalmers University of Technology

## Abstract

Despite decades of experimental searches, dark matter, a hypothetical form of matter with highly suppressed electromagnetic interactions, has yet to be detected. The absence of any detection in the 10 GeV–TeV mass range has motivated searches in lower mass ranges, reaching sub-GeV levels. Due to the Lee-Weinberg bound, a thermal-relic dark matter particle at this mass scale requires a new force mediator to bring the annihilation cross-section into agreement with the observed relic density. This force mediator is dubbed the dark photon, and several experiments designed to produce and detect it have been built or are currently under construction.

The main objective of this thesis is to develop a Bayesian framework for inferring the coupling strength and mass of the dark photon using data from the upcoming Light Dark Matter eXperiment (LDMX), a fixed target experiment where dark photons are expected to be produced in electron-tungsten collisions. The framework is validated on `MadGraph5`-simulated data sets to quantify LDMX's (Phase II) detection capabilities and parameter-inference power. This task is accomplished using dynamic nested sampling, designed for estimating posteriors and calculating marginal likelihoods. The secondary objective is to investigate whether the degeneracy between the dark photon mass and interaction model can be resolved through a combined analysis of the recoil-electron's transverse momentum $|p_T|$ and total energy $E$. The degeneracy is explored by comparing Bayes factors computed from the marginal likelihoods of the different models. The framework recovers the dark photon mass and coupling accurately for benchmarks corresponding to thermal relic targets for complex scalar dark matter with $\mathcal{R} \equiv m_{A'}/m_\chi = 2.5$, while identifying benchmarks with $\mathcal{R} = 2.2$, producing fewer expected signal events, as being below detection threshold. Using two-dimensional ($E$, $|p_T|$) kinematics, the Bayes factor comparison breaks the mass-model degeneracy between the interaction model groups ([KM, C, A] and [M, E]), with maximum Bayes factors of $\ln K = 591$. One-dimensional analysis of $E$ and $|p_T|$ alone still separates the model groups, though less decisively, with maximum Bayes factors of $\ln K \approx 42$ for $E$ alone and $\ln K \approx 22$ for $|p_T|$ alone. The introduction of a relic-target prior on $g_f$ further separates KM from [C, A], at the cost of additional assumptions.

**Keywords:** Bayesian Statistics, Parameter Inference, Dark Matter, Dark Photon, LDMX.

## Acknowledgements

First, I would like to thank both my supervisors, Prof. Riccardo Catena and Taylor Gray, for allowing me to contribute to their work and for their helpful guidance during this time. Secondly, I want to thank my family for always being supportive of my choices and for giving me the confidence to reach this far in my studies. Lastly, I would like to acknowledge my friends at Chalmers, with whom I have shared many laughs and rewarding discussions over the years. May we stay friends for a long time.

Adam Berger, Gothenburg, June 2026

# Table of contents


**List of figures** xi

**List of tables** xiii

**1 Introduction** 1
- 1.1 Background . . . 1
- 1.2 Bayesian Inference for Dark Photon Detection . . . 1
- 1.3 Limitations . . . 2
- 1.4 Thesis Structure . . . 3

**2 Dark Matter Theory** 5
- 2.1 Evidence for Dark Matter . . . 5
  - 2.1.1 Rotation Curves . . . 5
  - 2.1.2 Bullet Clusters . . . 5
  - 2.1.3 Cosmic Microwave Background Radiation . . . 6
  - 2.1.4 Structure Formation . . . 6
- 2.2 From WIMPs to Sub-GeV . . . 6
- 2.3 The Dark Photon . . . 7
- 2.4 LDMX . . . 8
  - 2.4.1 Dark Bremsstrahlung . . . 8
  - 2.4.2 Data Processing . . . 10
  - 2.4.3 Estimation of the Background . . . 12
- 2.5 Benchmarking on Relic Targets . . . 12

**3 Bayesian Inference & Sampling** 15
- 3.1 Bayesian Statistics in Physics . . . 15
  - 3.1.1 Parameter Inference . . . 16
  - 3.1.2 Credible Intervals . . . 16
- 3.2 Sampling the Posterior . . . 16
  - 3.2.1 Markov-Chain Monte-Carlo (MCMC) Sampling . . . 17
  - 3.2.2 Nested Sampling (NS) . . . 17
  - 3.2.3 Dynamic Nested Sampling (DNS) . . . 20
- 3.3 Likelihood Formulation . . . 21
- 3.4 Nuisance Parameters & Selecting Priors . . . 22
- 3.5 Model Comparison using Bayes Factors . . . 24

**4 Results** 27
- 4.1 Parameter Inference on KM Benchmarks . . . 27
  - 4.1.1 Posterior Recovery with DNS . . . 27
  - 4.1.2 MCMC vs DNS Cross-Validation . . . 31
- 4.2 Dark Photon Model Comparison . . . 34
  - 4.2.1 Noise Floor Estimation . . . 34

4.2.2 Bayes Factors . . . . . . . . . . 35
4.2.3 Bias Estimation under Model Misspecification . . . . . . . . . . 38
4.3 Relic Target Prior . . . . . . . . . . 39

**5 Discussion & Conclusion** **41**
5.1 Discussion . . . . . . . . . . 41
5.2 Future Considerations . . . . . . . . . . 41
5.3 Conclusion . . . . . . . . . . 42

**References** **43**

**A Benchmark Distributions** **A-1**

**B Tables** **B-1**

## List of Abbreviations

| | |
|---|---|
| **CMB** | Cosmic Microwave Background radiation |
| **WIMP** | Weakly Interacting Massive Particle |
| **LDMX** | Light Dark Matter eXperiment |
| **MCMC** | Markov-Chain Monte-Carlo |
| **DNS** | Dynamic Nested Sampling |
| **EOT** | Electrons on Target |
| **NS** | Nested Sampling |
| **SM** | Standard Model |
| **KM** | Kinetic Mixing |
| **DP** | Dark Photon |
| **DM** | Dark Matter |

# List of figures


2.1 Experimental setup for Phase II of LDMX. An 8 GeV electron beam hits a tungsten target, gaining momentum in the transverse direction and producing a dark photon through dark bremsstrahlung. The recoil electron's total energy and transverse momentum are recorded in an ECal, while any hadrons from non-signature events are detected by an HCal. The dark photon escapes the experiment undetected. . . . . . . . . . . . . . . . . . . . . 9
2.2 The dark bremsstrahlung process at LDMX: $e^-N \rightarrow e^-NA'$. An electron interacts with a tungsten nucleus, gains momentum in the transverse direction, and emits a dark photon. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
2.3 Kinematic distribution of the recoil electrons for $m_{A'} = 0.8$ GeV and $\epsilon = 0.5 \cdot 10^{-2}$ using the KM model, with background scaling BG = 100 **(left)**, and a Poisson realization of the same distribution **(right)**. The signal is displayed in red, and the background in blue. . . . . . . . . . . . . . . . . . . 13
2.4 Relic target curves for the KM model corresponding to $\mathcal{R} = 2.5$ and $\mathcal{R} = 2.2$. The colored dots mark the four benchmarks used to validate the Bayesian framework. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

3.1 Example of a nested sampling procedure. The live points start out uniformly across the entire prior. As the algorithm progresses, the live points hone in on the posterior peak. The red ellipse traces out the "slice" of the posterior defined by the live point with the smallest likelihood value. . . . . . . . . . 18
3.2 Example of Nested Sampling (NS) using a Gaussian likelihood and a five-dimensional prior. As the sampling progresses, $\ln X$ becomes increasingly negative. The dead points (maroon) are spread out roughly uniformly along $\ln X$, leading to equal resolution at the bulk of the posterior mass as in the empty prior landscape. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
3.3 Example of three Poisson probability distributions with $\lambda = 2, 6, 12$. . . . . 22
3.4 Prior distributions on the three nuisance parameters: log-normal on $\theta_S$ with fixed $\sigma_S = 0.1$ **(left)**; log-normal on $\theta_B$ shown with a representative $\sigma_B = 0.5$ **(center)**, where $\sigma_B$ itself is sampled from its hyperprior, a half-normal on $\sigma_B$ with scale 0.5 **(right)**. . . . . . . . . . . . . . . . . . . . . . . . . 25

4.1 Posterior distributions over the five parameters $(m_{A'}, g_f, \theta_S, \theta_B, \sigma_B)$ recovered for Benchmark 1 (a) and Benchmark 2 (b) with the Kinetic Mixing model using Dynamic Nested Sampling (DNS). For each benchmark, the left corner plot shows the posteriors when the background has been scaled by a factor of 1, and the right by a factor of 100. The dashed lines encase the 68% credible interval and median prediction of the five marginalized distributions corresponding to each parameter. The two-dimensional distributions are combinations of each of the five marginal posteriors, and their shaded regions encompass 39.3% and 86.5% of the posterior mass. . . . . . . . . . . . . . 29

4.2 Posterior distributions over the five parameters $(m_{A'}, g_f, \theta_S, \theta_B, \sigma_B)$ recovered for Benchmark 3 (a) and Benchmark 4 (b) with the Kinetic Mixing model using DNS. For each benchmark, the left corner plot shows the posteriors when the background has been scaled by a factor of 1, and the right by a factor of 100. The dashed lines encase the 68% credible interval and median prediction of the five marginalized distributions corresponding to each parameter. The two-dimensional distributions are combinations of each of the five marginal posteriors, and their shaded regions encompass 39.3% and 86.5% of the posterior mass. . . . . . . . . . . . . . . . . . . . . . 30
4.3 Comparison between MCMC sampling and DNS sampling on the same median Poisson realization out of 10 with BG = 1. The two recovered two-dimensional marginal posteriors are plotted on top of each other for comparison. Each subplot shows which benchmark was used in the generation of the data, and the orange triangle marks the true benchmark values. The color gradient on the DNS posterior and red outline for Markov-Chain Monte-Carlo (MCMC) mark the two-dimensional 68% and 95% credible intervals. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33
4.4 Bayes factors recovered with Dynamic Nested Sampling (DNS) by comparing marginal likelihoods between the five interaction models using both $E$ and $|p_T|$ kinematics. The y-axis shows what model was used in the generation of the data, while the x-axis shows what model was used in the likelihood while sampling. The numbers reported are $\ln K$, where $K = \frac{Z_{\text{true}}}{Z_{\text{hypothesis}}}$. A larger $\ln K$ means higher confidence that corresponding hypothesis model is ruled out. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
4.5 Same as Figure 4.4 using only total energy $E$ kinematics. . . . . . . . . . 37
4.6 Same as Figure 4.4 using only transverse momentum $|p_T|$ kinematics. . . . . 37
4.7 Bias estimation under model misspecification for the two-dimensional Bayes factor analysis. The y-axis shows what model was used in the generation of the data, while the x-axis shows what model was used in the likelihood while sampling. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
4.8 Same as Figure 4.4 but using a relic target prior on $g_f$ with $\mathcal{R} = 3$, allowing for separation between KM and [C, A]. . . . . . . . . . . . . . . . . . . . . . 39

A.1 Kinematic distributions of benchmark 1 (right), with a Poisson instance of said distribution (left). Background scaling BG = 1. . . . . . . . . . . . . . A-1
A.2 Kinematic distributions of benchmark 2 (right), with a Poisson instance of said distribution (left). Background scaling BG = 1. . . . . . . . . . . . . . A-1
A.3 Kinematic distributions of benchmark 3 (right), with a Poisson instance of said distribution (left). Background scaling BG = 1. . . . . . . . . . . . . . A-2
A.4 Kinematic distributions of benchmark 4 (right), with a Poisson instance of said distribution (left). Background scaling BG = 1. . . . . . . . . . . . . . A-2

# List of tables


4.1 The four KM relic target benchmarks that were used to test the Bayesian framework. Color coded to match to posterior figures. . . . . . . . . . . . . 27

4.2 Recovered $(\hat{m}_{A'}, \hat{g}_f)$ posteriors for the four KM benchmarks using DNS. Values are the median across $n = 10$ independent Poisson realizations of the per-realization posterior median, with 68% credible intervals given by the median across realizations of the 16th and 84th percentile bounds. The corner plots in Figures 4.1 and 4.2 show the representative realization (closest to the median across runs in the $(m_{A'}, g_f)$ plane) and may differ slightly from the values reported here. Rows with a cut value include the $E < 3.16$ GeV cut (Section 2.4.2). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28

4.3 MCMC sampler diagnostics across the four KM benchmarks, with background scaling BG = 1. Values are means/medians over $n = 10$ Poisson realizations. Acceptance fraction $\langle a \rangle$ should lie in $[0.2, 0.5]$ for healthy stretch-move sampling [12]. The autocorrelation time $\tau_{\max}$ is the maximum across the 5 sampled parameters; $\mathrm{ESS}_{\min}$ is the corresponding minimum effective sample size. Chain length $N = 10^4$ steps per walker, with $n_w = 50$ walkers. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

4.4 Cross-method validation: number of realizations (out of $n = 10$) in which the recovered 68% credible interval contains the true parameter value. Results from both DNS and MCMC are presented at background levels BG = 1 and BG = 100. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32

4.5 Relic target coupling values $g_f$ used as the generative-model couplings in the Bayes factor analysis, evaluated at the benchmark mass $m_{A'} = 0.1$ GeV. $N_{\mathrm{exp}}$ is the total number of expected signal events for the benchmark. . . . 34

4.6 Noise floor estimate for the Bayes factor calculations. Two independent `MadGraph5` batches of the KM signal ($\mathrm{KM}_{\mathrm{old}}$ and $\mathrm{KM}_{\mathrm{new}}$) were generated under identical physical parameters, with event counts scaled to match the KM-, M- and C-model relic-target benchmarks. For each Poisson realization, signal-plus-background data were drawn from both batches at a shared Poisson seed, and the log evidence of each dataset was computed under a common $\mathrm{KM}_{\mathrm{old}}$ likelihood. The reported quantity $|\Delta \ln Z| = |\ln Z_{\mathrm{new}} - \ln Z_{\mathrm{old}}|$ measures the shift in evidence induced by `MadGraph5` stochasticity. The minimum and maximum values characterize the signed scatter around zero. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

B.1 Recovered $(\hat{m}_{A'}, \hat{g}_f)$ posteriors for the four KM benchmarks using DNS at $200 \times 200$ resolution, with background scaling BG = 1. Values are the median across $n = 10$ independent Poisson realizations of the per-realization posterior median, with 68% credible intervals given by the median across realizations of the 16th and 84th percentile bounds. This higher-resolution analysis is presented as a cross-check of the primary $50 \times 50$ results (Table 4.2); recovered values are broadly consistent, and apparent differences for the low-signal benchmarks (BM3, BM4) lie within the realization-to-realization scatter expected at these low event counts and cannot be attributed to binning resolution alone. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . B-1
B.2 Recovered $(m_{A'}, g_f)$ posteriors for the four KM benchmarks, sampled with MCMC. Values quoted are the median and 68% credible interval over $n = 10$ independent Poisson realizations of each benchmark. Background scaling BG = 1. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . B-1

# 1. Introduction

## 1.1 Background

The frontiers of scientific discovery are continually being pushed forward. There are numerous examples in which a new type of astronomical object, planet, or entire field of physics is identified through inference of what *should* exist given our current understanding of the universe, even though we have not yet seen or proved it. One such example was in 1783, when the English clergyman and scientist John Michell proposed an object so massive that even light could not escape it. We first saw such an object in 2019, when the Event Horizon Telescope captured the first ever image of a black hole (M87). Another such example is the discovery of the planet Neptune through the irregular movements of Uranus. It was argued that a planet could be pulling on Uranus gravitationally, and this was confirmed to be the case when Johann Galle observed it telescopically in 1846 [7].

Today, one of the largest mysteries in physics is Dark Matter (DM), a proposed form of matter whose only confirmed interaction with regular matter is gravitational. Due to its nature, it is invisible to any known measurement technique, but its existence can still be inferred due to its effects on the dynamics of celestial bodies and its crucial role in the formation of the early universe. One of the most widely accepted theories is that the DM particles exist as an extension of the Standard Model (SM) particles, contained in a so-called dark sector, with its own set of particles and forces. While direct interactions between the dark sector and SM particles have not yet been observed, a wide range of theoretical models predict small but non-zero couplings, most notably through the theoretical existence of a mediator particle, the Dark Photon (DP) [11], which is what this work will focus on.

The Light Dark Matter eXperiment (LDMX) [31], a missing-momentum experiment currently under construction, is designed for detection of sub-GeV DM and DPs. Recent work [9] has shown that within the experimental setup of LDMX, where dark bremsstrahlung is the production method for DPs, there exists a degeneracy between different DP models and the mass of the DP. Specifically, the kinematic distributions of the recoil electrons, which constitute the signature at LDMX, depend both on the mass of the DP and what type of interaction is dominant between it and the SM fermions. This results in practically identical distributions for certain combinations of interaction model and DP mass. The authors of [9] suggest this degeneracy can be broken by combined analysis of both the final-state energy and the transverse momentum of the recoil electrons. Breaking of the degeneracy is essential for understanding the signal at LDMX: not only to be able to tell if DPs exist, but also to be able to distinguish which underlying interaction between it and the SM fermions is responsible for its production.

## 1.2 Bayesian Inference for Dark Photon Detection

This work develops a statistical framework for parameter inference and model discrimination of DPs at LDMX. Bayesian statistics is used to infer both the mass $m_{A'}$ and coupling strength $\epsilon$ of the DP within the Kinetic Mixing (KM) interaction model, using simulated total energy and transverse momentum data of the dark bremsstrahlung experiment at LDMX. The sensitivity of the models developed is benchmarked against $(m_{A'}, \epsilon)$-values that would be responsible for the DM abundance observed in the universe today, so-called relic targets. The framework is also evaluated against two different background scenarios, one with a normalized background and one where the background is scaled by a factor of 100. In addition to inference and sensitivity estimation for data generated by the KM model, four other types of interactions are investigated; Magnetic Dipole (M), Electric Dipole (E), Charge Radius (C) and Anapole Moment (A). Each of these interaction types is treated as a separate hypothesis, and Bayes factors are used to assess which is best supported by the data. The model comparison is performed under the combined total energy $E$ and transverse momentum $|p_T|$ distributions suggested for breaking the degeneracy between DP mass and interaction model in [9]. Further analysis using a thermal relic target prior on each model's coupling strength $g_f$ is performed in an attempt to separate KM from A and C.

The framework recovers the dark photon mass and coupling within the 68% credible interval for benchmarks corresponding to thermal relic targets with $\mathcal{R} \equiv m_{A'}/m_\chi = 2.5$ for background scaling BG = 1. For benchmarks corresponding to $\mathcal{R} = 2.2$, the framework does not detect a signal for most Poisson realizations of the data. This is due to the low number of expected events in this parameter space region for Phase II of LDMX, leading to wide credible intervals and with the background scaling of 100, a drowned-out, undetectable signal. A cross-validation between Dynamic Nested Sampling (DNS) and Markov-Chain Monte-Carlo (MCMC) shows that the two sampling methods largely agree, validating the posterior estimation of the DNS method. Calculated Bayes factors between interaction models show clear separation between the interaction model groups ([KM, C, A] and [M, E]) with a log-uniform prior on $g_f$. The groups produce nearly identical kinematic distributions for a given DP mass, differing only in the expected number of events, which can be accounted for by the framework by varying the coupling strength $g_f$. One-dimensional analysis of $E$ or $|p_T|$ alone still separates the model groups, though less decisively. Adding a relic target prior on $g_f$ further separates KM from A and C.

This thesis builds on the work of [9] and was developed in coordination with the research group's ongoing efforts. To keep the Bayesian framework compatible with their pipeline, the physical assumptions, `MadGraph5`-simulated data, data processing, and the Poisson counting likelihood are adopted from their current setup. The contribution of this work is the Bayesian part constructed on their foundation: parameter inference with DNS using signal and background nuisance parameters under a hierarchical prior, model discrimination via Bayes factors computed from marginal likelihoods of the five models, and a relic-target prior on $g_f$ that breaks the degeneracy within one of the model groups.

## 1.3 Limitations

In this work, the mass constraint of the DP is limited to the sub-GeV scale. Specifically, the prior on the DP mass is contained in the region $[10^{-3}, 1]$ GeV. The coupling strength for the five interaction models is also limited to a uniform prior on a log-scale in the range $[10^{-20}, 1]$, which means we are uninformative about the order of magnitude of these parameters. Towards the end of the analysis, a relic target prior on $g_f$ for complex scalar DM with $\mathcal{R} = 3$ and $\alpha_D = 0.5$ is added in order to further separate the interaction models.

The choice of relic target is made due to data being readily available for this particular combination of parameters. This work will, in accordance with [9], limit itself to the investigation of these aforementioned five different DP interaction models. The data is generated using simulation software, since LDMX has not run yet. This brings with it limitations on what can be studied, and the data is carefully constructed and modified to account for constraints in the experimental setup.

This work limits itself to studying Phase II at LDMX, which involves certain fixed experimental parameters that are detailed in Chapter 2. Bayesian framework testing is performed using four benchmarks selected for two different DP to DM mass-ratios, and taken from pre-calculated relic targets for a scalar DM model with fine structure constant $\alpha_D = 0.5$. The resolution of the mock signal data set simulated using `MadGraph5` is limited to 50 x 50 ($E$, $p_T$)-bins, but is also tested using 200 x 200 bins. Due to computational limits, this work computes 10 Poisson realizations per KM benchmark and 5 Poisson realizations per Bayes factor.

## 1.4 Thesis Structure

Chapter 2 introduces basic DM- and DP-theory relevant to the scope of this work. The chapter also includes detailed description of the experimental setup of Phase II at LDMX, including relevant details for the handling of the simulated data that will be used for analysis. Chapter 3 gives an in-depth guide of the Bayesian analysis and sampling methods that are used in this work. Chapter 4 presents the results of the analysis on relic-target benchmarks and model comparison. Chapter 5 summarizes and discusses the findings of this work.

# 2. Dark Matter Theory

This chapter introduces DM and DP theory relevant to this work. Section 2.1 presents some of the most convincing pieces of evidence for the existence of DM to date. Section 2.2 introduces the logic and motivation behind switching the search from higher mass-scales to sub-GeV levels, and why that requires the DP as a force mediator, introduced in section 2.3. Section 2.4 describes the experimental setup at LDMX, while explaining the different steps that have been taken in this work to reproduce the data that will come from the experiment. Finally, section 2.5 explains and presents the four benchmarks that are used to test the framework on simulated data.

## 2.1 Evidence for Dark Matter

Over the years there have been many cases of scientific discoveries that can only be explained by some sort of invisible, non-interacting particle that only affects objects around it gravitationally. Here, a few of the most notable ones are mentioned, to motivate the core search of a DM particle and its mediator, the DP.

### 2.1.1 Rotation Curves

Rotation curves are a measure of the orbital velocity of objects within a galaxy as a function of their distance from the galactic center. It was believed for a long time that objects further out from the rotational center of the galaxy would have a lower orbital velocity, as is the case e.g. for the planets orbiting the Sun in our solar system. However, in 1970, a study by Vera Rubin and Kent Ford [26] presented much more accurate observations of M31, measuring out up to 110 arcminutes from the galaxy's center. Later in the 1970s, more such studies followed, and it was observed that the rotation curves of galaxies seemed to level out with larger radii, rather than decrease. It was suggested that the explanation for these flat rotation curves was due to an unknown additional mass present in the outer edges of the galaxies. This missing mass, large structures encompassing galaxies at their outer edge, is today believed to be massive DM halos [7].

### 2.1.2 Bullet Clusters

Another strong piece of evidence for the existence of DM is the Bullet Cluster [10]. Two galaxy clusters set on a collision course have been observed through X-ray data, mapping the distribution of the baryonic mass inside the clusters, consisting of hot, dense plasma. When the clusters collided, the plasma experienced a drag force which caused it to slow down. The stars and galaxies, observed through optical wavelengths, simply passed through each other as the distances between them are too great for them to collide. However, in addition to X-ray observations, the gravitational lensing effect of the clusters was also measured, which takes into account the entire mass of the system. It was observed that, while the baryonic mass was slowed down, there was a significant displacement between the baryonic mass center and the total mass center of the clusters after the collision. The vast majority of the total mass of the two clusters simply passed through each other, indicating that the clusters are mostly made up of invisible non-interacting non-baryonic matter.

### 2.1.3 Cosmic Microwave Background Radiation

Recent data (2018) from the Planck telescope reveals details about the very early period of the universe through the Cosmic Microwave Background radiation (CMB) [1]. Before recombination, about 380 000 years after the Big Bang, the universe was so hot and dense that baryonic matter and photons were coupled tightly in a hot plasma. This plasma oscillated due to the competing forces of gravity and radiation pressure. Due to the non-interactive nature of DM, it did not experience the radiation pressure from the photons, and was allowed to create potential wells of gravity. When the universe cooled, the photons were released from their coupling with the baryonic matter, making their way through the universe, which we observe today as the CMB. When the radiation was released, it took with it a snapshot of the universe at recombination, and in it the gravitational imprints caused by DM can be measured. The relative highs and lows of these wells can be measured and directly related to the matter content of the universe. Following the Planck telescope data, the cold DM density has been constrained to $\Omega_c h^2 = 0.1200 \pm 0.0012$ and the baryon density to $\Omega_b h^2 = 0.02237 \pm 0.00015$. This confirms that the total non-baryonic DM density constitutes a much larger portion of the energy-mass content in the universe than baryonic matter.

### 2.1.4 Structure Formation

The final piece of evidence for the existence of DM presented here is that of the unexplainable structure formations of the universe. The CMB shows a remarkably smooth universe at recombination, with density fluctuations of only 1 part in $10^5$. Yet today, the universe is highly structured with galaxies, clusters, and the cosmic web at large. For the universe to have created such complex structures only due to gravity in a relatively short time period after recombination, something else had to have created gravitational structures already. Simulations of the formation of the universe show that the structures that can be seen today can form only with the inclusion of cold DM at recombination [8, 30].

## 2.2 From WIMPs to Sub-GeV

Historically, DM searches have been limited to the mass window of the Weakly Interacting Massive Particle (WIMP) which typically ranges from 10 GeV to TeV in scale. WIMPs are a proposed type of DM that can interact with SM particles through a weak-scale force. The reason WIMPs have been the strongest contender for the DM particle for so long is due to their weak-scale mass and weak-scale annihilation cross-section naturally freezing out[1] in the early universe with a relic density approximately matching that of the observed $\Omega_c h^2 \approx 0.12$ [7, 8].

Despite this match, decades of searching for the WIMP in the mass-scales where it should be found have yielded no detections. The most sensitive search to date, LZ, has found no evidence of WIMPs in the lowest region searched yet, the $3-9$ GeV regime, even up to sensitivities where solar neutrinos become an irreducible background [2]. This motivates

[1]meaning the DM annihilation rate falls below the rate at which the universe is expanding, leading to fewer interactions and effectively making the DM content of the universe constant.

the search for DM below the traditional 10 GeV scale, down to the sub-GeV scale. Due to the Lee-Weinberg bound, the calculated relic density would not match the observed value for a DM particle that only annihilates through a weak SM interaction at sub-GeV scale, as the annihilation cross-section would be too small [20]. In order to increase the annihilation rate of a proposed sub-GeV DM particle, a new mediator particle must exist, the DP [9], which is introduced in the next section.

## 2.3 The Dark Photon

Sub-GeV DM models require a mediator for interactions between SM particles and particles that belong to the dark sector. This mediator can be included in the underlying physics in a number of ways, some more complex than others. For this work, we adapt the standard approach of [9] in which a massive mediator is introduced as an extension of the SM through a new U(1) gauge group, $\mathrm{U}(1)_{\mathrm{D}}$. The new gauge group carries with it a corresponding gauge boson $A'$, the DP.

Since it is not known what types of interactions occur between $A'$ and the SM particles, other than that they are weak, *portals* are introduced as small, potentially observable interactions. The vector portal causes interaction between the $A'$ gauge boson and the SM gauge bosons corresponding to the groups $\mathrm{U}(1)_Y$ or $\mathrm{U}(1)_{\mathrm{em}}$ through kinetic mixing. The interaction occurs with $\mathrm{U}(1)_Y$ for energy scales above the electroweak symmetry-breaking scale, and with $\mathrm{U}(1)_{\mathrm{em}}$ below it. At sub-GeV levels, the interaction occurs between $A'$ and the SM photon of the $\mathrm{U}(1)_{\mathrm{em}}$ group, which is why we only consider that interaction in this work. We can write down a general form of the Lagrangian considering the Abelian gauge bosons corresponding to the $\mathrm{U}(1)_{\mathrm{em}}$ group and a dark $\mathrm{U}(1)_{\mathrm{D}}$ gauge group;

$$\mathscr{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} - \frac{1}{4}F_{D\mu\nu}F_D^{\mu\nu} - \frac{\epsilon}{2}F_{\mu\nu}F_D^{\mu\nu}, \tag{2.1}$$

where $\epsilon$ is the kinetic mixing parameter and $F_{\mu\nu}$ is the electromagnetic field strength tensor. The third term is responsible for the interaction between the bosons, and after diagonalization, the dark photon gains a coupling to the fermionic electromagnetic current $J^\mu = \sum_f Q_f \bar{f}\gamma^\mu f$ as

$$\mathscr{L}_{\mathrm{KM}} = -e\epsilon J^\mu A'_\mu, \tag{2.2}$$

where $e$ is the elementary charge. Therefore, the kinetic mixing induces an effective coupling strength $-e\epsilon Q_f$ between $A'$ and each charged SM fermion [11].

The dark sector could include more complex physics than just kinetic mixing. Higher-dimensional operators can be constructed which, if their signature is detected at a low-energy experiment like LDMX, could indicate additional UV physics beyond the minimal kinetic-mixing scenario [24]. In this thesis, the kinetic mixing scenario is extended by including four additional interactions, coming from higher-order dark electromagnetic moment operators. They are treated independently within an effective-field-theory framework, since LDMX cannot resolve the UV physics that would generate them [9]. Although UV physics could be responsible for generating combinations of these operators, this work considers each as a separate operator hypothesis to compare models using Bayes factors.

The higher-order interaction models are Magnetic Dipole (M), Electric Dipole (E), Charge Radius (C) and Anapole Moment (A). The full interaction Lagrangian, considering all five models following [9], takes the form

$$\begin{aligned}\mathscr{L}_{A'f} = & -e\epsilon Q_f\,\bar{f}\gamma^\mu f\,A'_\mu - \mu_f\,\partial_\nu(\bar{f}\sigma^{\mu\nu}f)\,A'_\mu - id_f\,\partial_\nu(\bar{f}\sigma^{\mu\nu}\gamma^5 f)\,A'_\mu \\ & - b_f\left[\partial^\nu\partial_\nu(\bar{f}\gamma^\mu f) - \partial^\mu\partial_\nu(\bar{f}\gamma^\nu f)\right]A'_\mu \\ & + a_f\left[\partial^\nu\partial_\nu(\bar{f}\gamma^\mu\gamma^5 f) - \partial^\mu\partial_\nu(\bar{f}\gamma^\nu\gamma^5 f)\right]A'_\mu,\end{aligned} \tag{2.3}$$

where, with respect to $A'$, $\mu_f$ and $d_f$ are the magnetic and electric dipole moments of the SM fermion $f$, $b_f$ the charge radius coupling and $a_f$ the anapole coupling. The tensor $\sigma^{\mu\nu}$ is the antisymmetric product of Dirac matrices and $\gamma^5$ the chirality operator. The KM coupling strength $\epsilon$ is dimensionless, while the other coupling strengths are dimensionful with $[\mu_f,\, d_f]$ having dimension $[\text{mass}]^{-1}$ and $[b_f,\, a_f]$ dimension $[\text{mass}]^{-2}$. This is consistent with the idea of their suppression at higher mass scales. It is convenient to define each of the coupling strengths under a common parameter:

$$g_f \equiv \begin{cases} \epsilon & \text{Kinetic Mixing} \\ \mu_f/e & \text{Magnetic Dipole} \\ d_f/e & \text{Electric Dipole} \\ b_f/e & \text{Charge Radius} \\ a_f/e & \text{Anapole Moment} \end{cases}. \tag{2.4}$$

This parametrization, adopted from [9], allows for comparison between the five interactions, with the elementary charge $e$ factored out of the dipole, charge radius and anapole couplings.

## 2.4 LDMX

LDMX is an upcoming accelerator experiment designed for both direct detection of DM particles and mediator particles, such as the DP, at sub-GeV scale [31]. It is designed to use missing-momentum of a fixed-target electron beam fired at a heavy material to detect signatures of DM or DP production. In order to do this, it must precisely measure the initial and final states of the incoming and outgoing particles.

The incoming electrons are counted using a trigger scintillator system, and their momentum and trajectory precisely identified using a tagging tracker. After the electrons have hit the target material, their altered trajectory and momentum are measured using a recoil tracker. The final state particles then enter a set of calorimeters, which measure their energy. The calorimeters act as a veto system, rejecting background events and pairing particles to signal events (see Figure 2.1) [3, 31].

### 2.4.1 Dark Bremsstrahlung

In order to detect a massive DP, one needs to consider its interactions with the SM particles. The part of the Lagrangian coupling the DP Gauge boson $A'$ to the electromagnetic current $J^\mu$ is structured in the same way as the regular photon interaction with the electromagnetic current, as shown in equation (2.2). This gives rise to the same types of production mechanisms as with the regular photon, which can be created in detection experiments. In this work, we will consider one of these processes in particular, namely what is known as *dark bremsstrahlung* (see Figure 2.2). The interaction gets its name from the analogous counterpart of the SM photon, Bremsstrahlung, in which a charged particle emits a photon

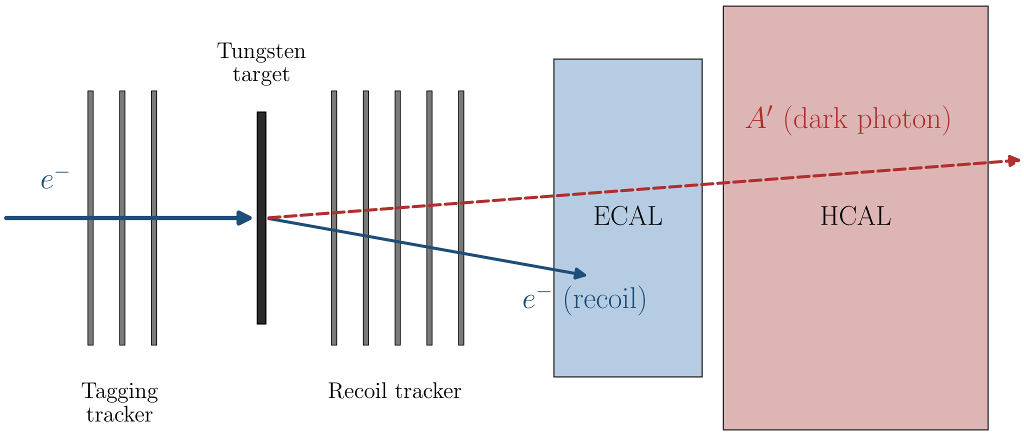



**Figure 2.1:** Experimental setup for Phase II of LDMX. An 8 GeV electron beam hits a tungsten target, gaining momentum in the transverse direction and producing a dark photon through dark bremsstrahlung. The recoil electron's total energy and transverse momentum are recorded in an ECal, while any hadrons from non-signature events are detected by an HCal. The dark photon escapes the experiment undetected.

when being decelerated [33]; for example by interaction with another particle. In the context of producing a DP, a scenario can be created where a high-energy electron beam is fired at a heavy material. The electrons scatter off the target nuclei, producing a DP that, being invisible to detectors, carries away the missing energy and momentum [11]. In order to separate dark bremsstrahlung processes from regular bremsstrahlung processes or any other type of interaction, the experiment uses various detectors. Both electromagnetic (ECal) and hadronic calorimeters (HCal) are used to detect any photons or hadrons that could be created in processes other than dark bremsstrahlung. A signal event is one where the recoil electrons have missing momentum and energy and no SM photons or hadrons were detected, suggesting that an undetectable DP was the cause of the electron's missing momentum and energy. Throughout this work, it is assumed that the DP predominantly decays invisibly into DM particles, requiring that the mass of the DP is at least twice that of the scalar DM particle ($m_{A'} > 2m_\chi$). Under this assumption, the DP escapes the experiment undetected [9].

The signature of dark bremsstrahlung at LDMX is therefore a recoil electron with substantially reduced energy and significant transverse momentum, accompanied by an absence of any explaining detections in the calorimeters. By measuring the recoil electron's total energy $E$ and transverse momentum $|p_T|$, one can infer the kinematics of the DP that was produced through energy-momentum conservation. This scenario is exactly what is to be created and studied at LDMX (Phase II), where an 8 GeV electron beam is fired at a thin plate of tungsten ($^{184}W$) [31]. Tungsten is chosen primarily for its high atomic number $Z$, which enhances the dark bremsstrahlung production rate through the $Z^2$ scaling of the cross section [31]. While the analytic treatment of the cross section assumes a static nucleus, the simulated signal samples used in this work include the full kinematics of the

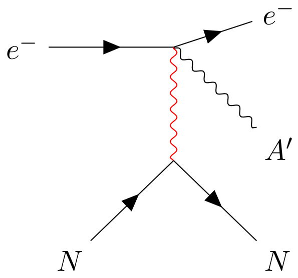



**Figure 2.2:** The dark bremsstrahlung process at LDMX: $e^-N \to e^-NA'$. An electron interacts with a tungsten nucleus, gains momentum in the transverse direction, and emits a dark photon.

recoiling tungsten nucleus, following [9]. The experiment will be conducted in different phases, where Phase II, which is considered in this work, is to be run for an extended period, resulting in approximately $10^{16}$ Electrons on Target (EOT).

### 2.4.2 Data Processing

The simulated signal samples, denoted $S$, are generated using the simulation software `MadGraph5_aMC@NLO`[2][5, 22], a matrix-element generator that automatically constructs Feynman diagrams from a given Lagrangian and computes the corresponding tree-level cross-section by integrating the squared matrix element $|M|^2$ over the final state phase-space. In the software, a number of parameters for the experiment can be used as input, including the DP mass $m_{A'}$, the number of final events and the Lagrangian of the corresponding DP model that is to be simulated. For each value of $m_{A'}$, `MadGraph5` outputs a cross-section $\sigma(m_{A'})$, along with an associated integration uncertainty. In addition, the software outputs the final state in the form of four-momenta, which are then projected onto the total energy $E$ and transverse momentum $|p_T|$ of the recoil electrons.

For each model, 100 simulations are run using `MadGraph5` corresponding to $m_{A'}$-values in the range $0.001-1$ GeV (log-scale) to generate a total of $10^6$ signal events per simulation. To prepare the data for statistical analysis, the ($E$, $|p_T|$) of the recoil electrons corresponding to a cross-section $\sigma(m_{A'})$ are binned for two different resolutions, 50x50 and 200x200, where 50x50 is the main one used throughout this work. Choosing the lower resolution comes down to managing computational cost while keeping the grid dense enough to resolve shape differences. A comparison between the two resolutions is performed and reported in Appendix B. The ranges of the distributions are $E \in [0, 8]$ GeV, as 8 GeV is the energy of the incoming electron beam, and for $|p_T| \in [0, 2]$ GeV, as the expected yield of events above 2 GeV is considered negligible. This means that each bin $i$ corresponds to a range of $E$ and $|p_T|$ where a detection has happened. If a count is registered in a bin, it means a detected recoil electron had both total energy $E$ and transverse momentum $|p_T|$ within the range of that bin.

There are several modifications applied to the signal and background distributions in order to make them similar to the actual data that will be produced at LDMX. As adopted

[2]The generation of the data was performed by Taylor Gray. This section describes the inputs and outputs of `MadGraph5` relevant for the upcoming analysis.

in [3][3], two primary kinematic cuts have been implemented due to the limitations of the underlying physics of the experiment and the experimental equipment itself. The first is a lower bound on the possible total energy of the detected recoil electrons, at $E = 0.05$ GeV. This cut is due to the detector's inability to track charged particles with energies below this value [31]. The second cut is a requirement on the recoil electrons, keeping them within a $\pm 40°$ angle of the electron beam axis. This angle-constraint manifests as a condition on the ($E$, $|p_T|$) values themselves, namely $|p_T| < E \cdot \sin 40°$. These two kinematic cuts are applied for all data sets. As an additional part of the analysis, a third cut of $E < 3.16$ GeV is implemented as a test on the framework's capabilities on handling a lesser signal, while also attempting to reduce the background in the kinematic distributions. This specific cut is reported in LDMX collaboration reports for the trigger requirement [3, 4].

To account for the uncertainty of the detectors at LDMX, a smearing is applied to the bins after the cuts have been applied. A real detector is not exact in the measurement of $E$ and $|p_T|$: a detection in one bin could have been the result of a recoil electron with kinematics corresponding to a neighboring bin. To model this, the bin contents are spread out according to a Gaussian distribution centered at the corresponding bin, applied separately for $E$ and $|p_T|$. The width of the Gaussian decides how much blurring of the bin occurs. For energy, the width is calculated as

$$\sigma_E(E) = \sqrt{s^2 E + c^2 E^2 + n^2}, \tag{2.5}$$

with $s = 0.2$ GeV$^{1/2}$, $c = 0.03$ and $n = 0$, taken from Table 3.10 in [3]. For $|p_T|$, the width is taken from the resolution values in Figure 3.26 of [3]. The resulting distribution gets washed out, modelling the uncertainty of the signal inherent to the detectors at LDMX.

The resulting kinematic distribution is normalized with the total number of simulated events and then scaled with the proper luminosity and coupling strength factor using the experimental parameters and the cross-sections calculated by `MadGraph5`. All simulation is performed with $g_f = 1$, which allows the distributions to be adapted to any coupling strength by normalizing them to the proper coupling strength scaling. The total number of expected dark photon events is given by the cross-section and the experimental parameters:

$$N_{A'}(m_{A'}, g_f) = \sigma_{A'}(m_{A'}, g_f = 1)\, \underbrace{\frac{T X_0 N_e N_0}{A}}_{\text{luminosity}}\, g_f^2, \tag{2.6}$$

where $\sigma_{A'}$ is given in cm$^2$, $T = 0.1$ is the number of radiation lengths, $X_0 = 6.76$ g/cm$^2$ the radiation length of the target, $N_e = 10^{16}$ the number of EOT, $A = 183.84$ g/mol the atomic mass of tungsten and $N_0$ is Avogadro's number. The scaling of the coupling constant as a normalization step allows the framework to adopt it as a variable parameter in the sampling process.

The values of each bin generally take on decimal values. As a final step before analysis, the mock data of a distribution with selected $m_{A'}$ and coupling strength $g_f$ is sampled through a random Poisson sampler. The reason for this is that the experiment at LDMX will detect an integer number of recoil electrons with energy and transverse momentum corresponding to each bin. A Poisson sampling of the distribution ensures integer counts in the bins, the drawback of which is the introduction of random Poisson noise. The Poisson

[3]and confirmed through discussion with Lund University.

realization of one distribution could look slightly different from another realization with the same simulated distribution as its origin. The framework mitigates this noise by running analysis over several Poisson realizations of the same mock data distribution.

### 2.4.3 Estimation of the Background

Although the veto-system at LDMX described in section 2.4.1 is designed to reject non-signal events, no missing-momentum experiment is entirely free of background processes. Among the background processes considered in [31], three illustrate the main mechanisms by which a recoil electron's kinematics can mimic those of a dark bremsstrahlung event:

1. *Electrons that did not react with the target* end up as high-energy showers in the calorimeters, which occasionally lose energy through electro-nuclear or photo-nuclear interactions in the ECal. This is the largest cause for background events at LDMX.

2. *Incident low-energy particles*, impurities in the electron beam with energies below the intended beam energy, could make their way to the calorimeters and mimic a signal event.

3. *Hard bremsstrahlung* occurs in the target and produces ordinary photons, which can be missed as background events by the detectors due to low measurement accuracy in energy.

Due to the lack of access to a combined two-dimensional background, it is modelled by combining two different sources. First, the background for the recoil electron's energy $E$ is taken as a digitization of Figure 10 in [31], where the energy axis is scaled by 2 to approximate Phase II's 8 GeV beam. Second, the background in $|p_T|$ is approximated as the $|p_T|$ distribution for the simulated `MadGraph5` data for kinetic mixing with $m_{A'} = 0.01$ GeV, used as a stand-in in the absence of a $|p_T|$ background. The two backgrounds are combined to form a two-dimensional distribution through an outer product, treating them as uncorrelated. The same cuts and smearing as described in section 2.4.2 are applied and then the background template is normalized. It can then be rescaled to a desired number, and in this work a background scaling of BG = 1 and BG = 100 is considered in the analysis. This scaling, together with the nuisance parameter $\theta_B$ introduced in section 3.4, sets the absolute background normalization during inference, while the template itself only controls its shape in the $(E, |p_T|)$ plane.

See Figure 2.3 for an example of a binned $(E, |p_T|)$-distribution of signal and background simulated with $m_{A'} = 0.8$ GeV and $\epsilon = 0.5 \cdot 10^{-2}$ using the kinetic mixing model. The background scaling is set to BG = 100. The distribution is displayed with the two primary kinematic cuts applied (excluding $E < 3.16$ GeV), and post-cut smearing. The full simulated kinematics is displayed to the left, while a Poisson instance of the same distribution is displayed to the right.

## 2.5 Benchmarking on Relic Targets

In this work, the approach of [9] is adopted in which a complex scalar DM $\chi$ charged under a dark $U(1)_D$ gauge group is considered, with dark fine-structure constant $\alpha_D = 0.5$. As mentioned above, for the invisible decay of the DP into DM particles the DP/DM mass

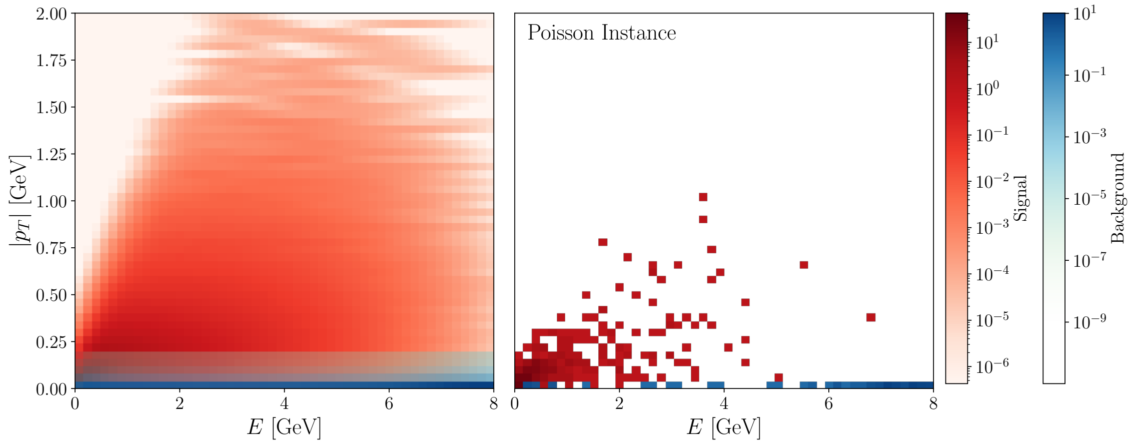



**Figure 2.3:** Kinematic distribution of the recoil electrons for $m_{A'} = 0.8$ GeV and $\epsilon = 0.5 \cdot 10^{-2}$ using the KM model, with background scaling BG = 100 **(left)**, and a Poisson realization of the same distribution **(right)**. The signal is displayed in red, and the background in blue.

ratio has to obey

$$\mathcal{R} \equiv m_{A'}/m_\chi > 2. \tag{2.7}$$

For a given $\mathcal{R}$, $\alpha_D$ and interaction model, requiring the thermal annihilation cross-section to agree with the observed relic density $\Omega_{\text{DM}}h^2 \approx 0.12$ traces a curve in the $(m_{A'}, g_f)$ plane, called a relic target curve. Points along this curve represent mass and coupling strength combinations we most expect to be realized given the assumptions made, as they would agree with observed data.

The Bayesian framework is benchmarked on four $(m_{A'}, g_f)$ combinations under the assumed kinetic mixing model. These benchmarks are extracted from digitization of Figure 6 in [6], and rounded to two significant figures. Two benchmarks from the $\mathcal{R} = 2.5$ and $\mathcal{R} = 2.2$ relic curves each are used, corresponding to $m_{A'} = 0.1$ and 0.01 GeV. The four benchmarks taken from their corresponding relic curves can be seen in Figure 2.4. In previous work [9], the $\mathcal{R} = 3$ scenario has been considered. Recent studies have excluded complex scalar DM with $\mathcal{R} \gtrsim 3$ [19], motivating the search at lower values, where the expected signal is correspondingly smaller. Values extracted from a curve with higher $\mathcal{R}$ yield a larger number of signal events, for a given $m_{A'}$. Therefore, exploring two different values of $\mathcal{R}$ in the range $\mathcal{R} \in [2, 3]$ probes the detection sensitivity of LDMX.

For each benchmark, mock data distributions are created to mimic the data that would be produced at LDMX for DP model with the corresponding parameters. This mock data is then compared to other signal data through Bayesian methods, detailed in the following chapter.

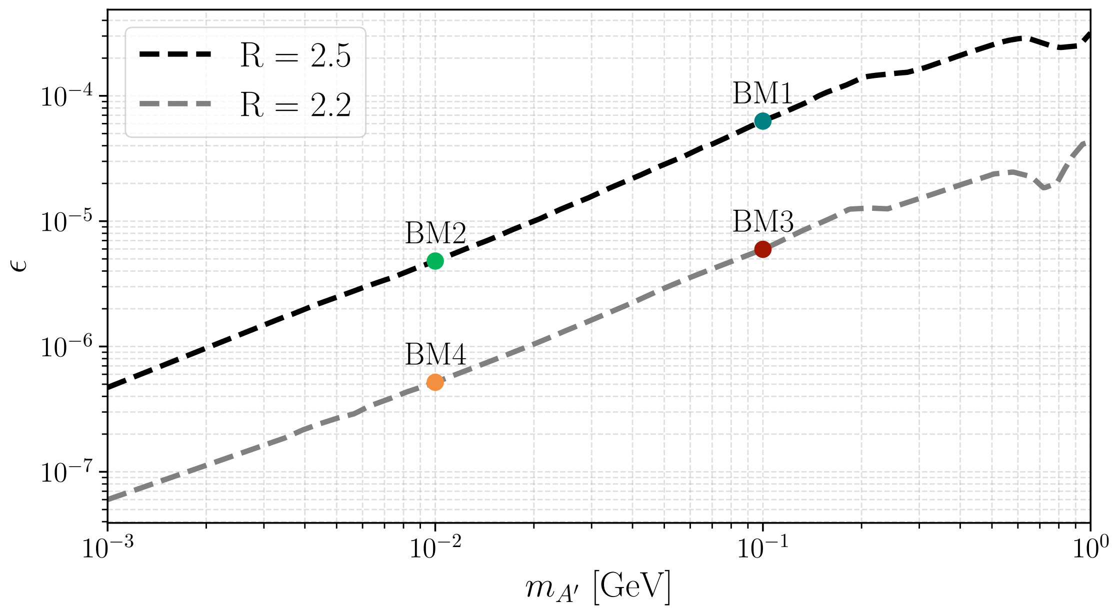



**Figure 2.4:** Relic target curves for the KM model corresponding to $\mathcal{R} = 2.5$ and $\mathcal{R} = 2.2$. The colored dots mark the four benchmarks used to validate the Bayesian framework.

# 3. Bayesian Inference & Sampling

This chapter goes over some basic concepts in Bayesian statistics, as well as the sampling methods and specific choices made in this work. Section 3.1 goes over the basic ideas of Bayes' theorem and how it is used from a physics perspective. In section 3.2, the various sampling methods used in this work are introduced, including some motivation on why we might be interested in sampling the posterior. Later, sections 3.3 and 3.4 cover the choice of likelihood and priors, explaining the model in terms of its nuisance parameters and why they are important. Lastly, section 3.5 introduces Bayes factors and how they are computed in this work, as well as how they are used in DP interaction model comparison.

## 3.1 Bayesian Statistics in Physics

When parameters are to be inferred in a scientific context, we are interested in the *posterior distribution* $p(\boldsymbol{\theta} \,|\, \mathcal{D}, \mathcal{M})$. The posterior is the probability distribution of a set of parameters $\boldsymbol{\theta}$, given a dataset $\mathcal{D}$ and our model $\mathcal{M}$ that we have engineered to try to explain the physical nature of our experiment. In reality, all models are wrong, and the goal of Bayesian statistics is to introduce uncertainties about our model, to allow observed data to update our prior beliefs. The posterior distribution is related to the *likelihood* and the *prior* through Bayes' Theorem:

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}, \mathcal{M}) = \frac{\mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\theta}, \mathcal{M})\pi(\boldsymbol{\theta})}{p(\mathcal{D} \,|\, \mathcal{M})}. \tag{3.1}$$

The likelihood $\mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\theta}, \mathcal{M})$ is a distribution quantifying the probability of observing the data given that our set of parameters $\boldsymbol{\theta}$ and our model $\mathcal{M}$, are correct. If the parameters agree with the data, the likelihood is large, and vice versa. The prior $\pi(\boldsymbol{\theta})$ is the probability density of the set of parameters $\boldsymbol{\theta}$, which is decided *a priori*. In general, the prior is model dependent. For this work, all models considered share the same prior for the majority of the analysis. The denominator $p(\mathcal{D} \,|\, \mathcal{M})$ is the *marginal likelihood* or *Bayesian evidence*. It is defined in equation (3.6) and is often difficult to define in a closed form for convenient calculation [34].

Bayes' theorem tells us how we should update our beliefs when presented with new data. The prior contains our initial beliefs about what the parameters of our model should be, while the likelihood encodes the magnitude at which the data updates that initial belief. Consider a scenario where we are entirely certain one of the parameters should be $\theta_1 = 2.0$, i.e. the prior is a delta function $\delta(\theta - 2.0)$. There would be no amount of data that could sway the power of such a prior, since the probability of $\theta_1$ being anything other than 2.0 is zero. For this reason, we must admit to some sort of uncertainty about our prior beliefs, and make it possible for the posterior to consider many alternatives in the probability of $\boldsymbol{\theta}$.

### 3.1.1 Parameter Inference

The posterior in equation (3.1) is joint over all parameters of the model. In many cases, only a few of the parameters are of interest, or the model contains *nuisance parameters*, which are covered in section 3.4. In order to extract the posterior over only a select number of parameters, the remaining ones have to be *marginalized* over. In other words, they have to be integrated out, projecting the multi-dimensional posterior onto a single axis, the axis of the parameter of interest e.g. $m_{A'}$:

$$p(m_{A'}\,|\,\mathcal{D},\mathcal{M}) = \int p(m_{A'},\boldsymbol{\theta}^*\,|\,\mathcal{D},\mathcal{M})d\boldsymbol{\theta}, \tag{3.2}$$

where $\boldsymbol{\theta}^*$ are all other parameters except $m_{A'}$. This marginalization propagates the uncertainties of the other parameters onto the marginal posterior of $m_{A'}$ [14].

Once the posteriors of the parameters of interest are recovered, they can be interpreted in several ways. This is often done through a point estimate, a number that somehow represents a parameter prediction or another quantity of the distribution. This could for example be the mean of the distribution, or the mode, which gives the single most probable parameter value. In this work, we use the median prediction. Choosing the median of the distribution ensures that 50% of the posterior mass is located above and 50% is located below the prediction value. This is common when using samplers, as the resolution around the peak can often be low, leading to uncertain maximum probability estimations [14].

### 3.1.2 Credible Intervals

In addition to the median prediction, this work uses 68% credible intervals (CI) for its uncertainty estimations. A $100 \cdot (1-\alpha)\%$ credible interval is any interval $[a, b]$ such that

$$\int_a^b p(\theta\,|\,\mathcal{D},\mathcal{M})d\theta = 1-\alpha, \tag{3.3}$$

where $p(\theta\,|\,\mathcal{D},\mathcal{M})$ is the marginal posterior for the parameter $\theta$. There are many possible intervals that follow this definition. In this work, the 16th and 84th percentiles of the marginal distributions are used for the 68% CI, placing 16% of the distribution in each tail of the posterior. For a Gaussian distribution, the Bayesian credible interval and the frequentist *confidence interval* coincide, and a 68% CI corresponds to the $\pm1\sigma$ bounds [32].

## 3.2 Sampling the Posterior

There are many cases where one might have a likelihood and prior distribution, but has no analytical way of calculating the posterior distribution. One of the more difficult parts of getting an estimation of the posterior is the denominator in equation (3.1), the so-called marginal likelihood. There are many sampling algorithms that handle this problem though, such as MCMC sampling (section 3.2.1), where the marginal likelihood is simply never calculated, the consequence of which is an un-normalized sampled posterior. Another example is Nested Sampling (NS) (section 3.2.2), which is specifically designed to approximate the marginal likelihood, at the cost of less efficiency in parameter inference. The sampling method used in this work is DNS (section 3.2.3), a modified NS algorithm that allocates live points to areas of higher posterior mass, combining the marginal likelihood calculation strategy of NS with high efficiency in parameter inference.

### 3.2.1 Markov-Chain Monte-Carlo (MCMC) Sampling

A Markov chain is a particular type of stochastic process defined by the Markov property

$$\mathbb{P}_{X_n|X_0,X_1,\ldots,X_{n-1}}(i_n|i_0,i_1,\ldots,i_{n-1}) = \mathbb{P}_{X_n|X_{n-1}}(i_n|i_{n-1}), \tag{3.4}$$

where $X_n$ are random variables and $i$ denotes an iteration in the process [21]. The condition describes a scenario where the state $n$ only depends on the state before it $n-1$. In essence, the probability of the state $n$ is only dependent on the previous state, and not any of the states that come before it. This concept is the core of MCMC, a category of sampling algorithms that, when implemented correctly, sample the posterior by proposing new steps that depend on the previous one. The Markov chain eventually reaches an equilibrium state where all subsequent samples follow the same distribution as the posterior.

The most common implementation of MCMC-sampling is the Metropolis-Hastings algorithm [16]. Applied to the problem of parameter inference, where we are interested in sampling an unknown posterior from which we can only evaluate the numerator in equation (3.1), the algorithm starts at state $\boldsymbol{\theta}_n$ (a point in parameter space). A new state $\boldsymbol{\theta}'$ is drawn from a proposal distribution $Q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}_n)$. In theory, $Q$ does not need to look anything like the posterior we are interested in sampling. Labeling the product of the likelihood and prior as $p^*(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$, the proposed step is accepted with probability

$$\alpha = \min\left(1, \frac{p^*(\boldsymbol{\theta}' \,|\, \mathcal{D},\mathcal{M})}{p^*(\boldsymbol{\theta}_n \,|\, \mathcal{D},\mathcal{M})} \frac{Q(\boldsymbol{\theta}_n \,|\, \boldsymbol{\theta}')}{Q(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}_n)}\right). \tag{3.5}$$

If the step is accepted, we set $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}'$, otherwise, $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n$. If $Q$ is symmetric, the fraction in (3.5) becomes a fraction between $p^*$ for the two points. In theory, the algorithm will draw the steps taken by the algorithm closer to regions of the posterior with high probability. Since each step depends on the last, it can take a considerable number of steps to get enough independent samples from the distribution. Fortunately, one often only needs a small number of independent samples to approximate the distribution being sampled [21]. After an appropriate number of steps, and after proper convergence of the Markov Chain, the samples themselves will be distributed as the posterior, with the drawback of not being properly normalized.

This work makes use of the MCMC sampler package `emcee` [12], an ensemble sampler utilizing the affine-invariant stretch move [15]. A number of walkers $n_w$ take steps in parallel, where the next proposed step is defined by the current position of the other walkers. This makes the step size adaptive, adjusting the sampler to posterior shapes that a fixed step size would have difficulty navigating.

### 3.2.2 Nested Sampling (NS)

The main objective with NS [27] is to calculate the Bayesian evidence. With MCMC sampling, the overall normalization factor in equation (3.1) is not needed. Instead, MCMC sampling estimates the shape of the posterior, so that a mean/median/mode value and variance could be determined for the parameters we are trying to infer. In order to properly compare how well different models fit our data using Bayes factors (see section 3.5) this normalization factor is needed. It takes the form of an integral over the entire prior volume $\Omega_{\boldsymbol{\theta}}$:

$$Z = \int_{\Omega_{\boldsymbol{\theta}}} \mathcal{L}(\mathcal{D} \,|\, \boldsymbol{\theta},\mathcal{M})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{3.6}$$

This is in general a very difficult integral to calculate, spanning as many dimensions as there are parameters. Nested sampling tackles this problem by reshaping the prior landscape into a volume by the use of a prior transform. Each point in a $d$-dimensional unit cube $u \in [0, 1]^d$ gets mapped to the prior space via a cumulative distribution function, and the integral in (3.6) becomes a one-dimensional integral

$$Z = \int_0^1 \mathcal{L}(X(\lambda))dX(\lambda), \tag{3.7}$$

where $X(\lambda)$ is the prior volume contained in the region where the likelihood exceeds the threshold $\lambda$ [27];

$$X(\lambda) = \int_{\mathcal{L}(\boldsymbol{\theta})>\lambda} \pi(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{3.8}$$

As $\lambda$ increases, the volume $X$ shrinks from $X = 1$, where the entire prior lies above $\lambda$, to $X = 0$, where $\lambda$ is the maximum likelihood. To estimate this integral, the sampler places a number of *live points* uniformly spaced inside the prior volume. The algorithm then evaluates which of these points has the lowest likelihood and discards it, labeling it as a *dead point*. A new point is sampled inside the now smaller prior volume, defined by the reach of all the current live points (see Figure 3.1). The algorithm keeps doing this until the prior volume is as small as a certain convergence criterion allows. Each iteration, the "area" that the current live points encompass is multiplied by the most recent dead point, adding a "slice" to the total calculation of $Z$ [27].

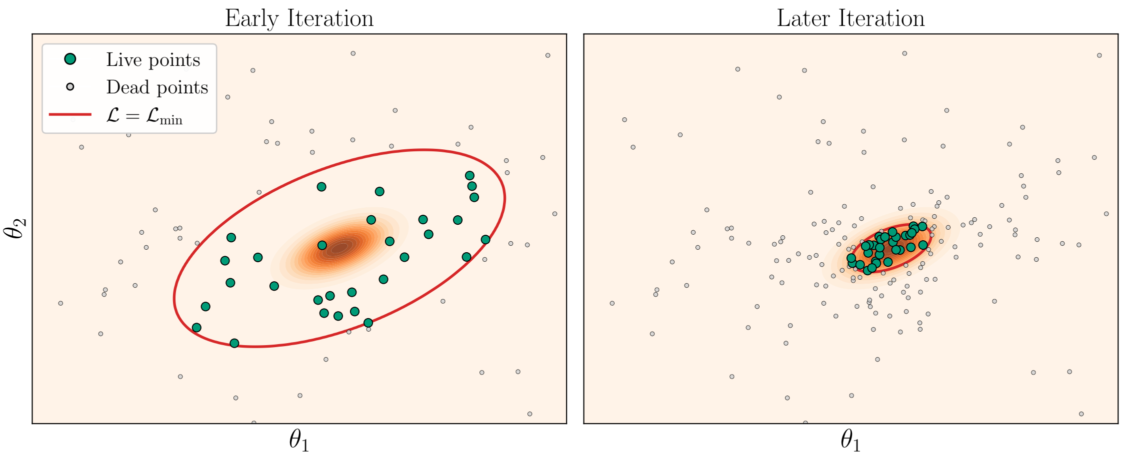



**Figure 3.1:** Example of a nested sampling procedure. The live points start out uniformly across the entire prior. As the algorithm progresses, the live points hone in on the posterior peak. The red ellipse traces out the "slice" of the posterior defined by the live point with the smallest likelihood value.

The nested sampling algorithm proceeds as follows:

1. Draw $N$ live points from the prior $\pi(\boldsymbol{\theta})$.

2. Identify the live point with the lowest likelihood, $\mathcal{L}_i$, and record it as a dead point with an associated prior volume $X_i$ at iteration $i$.

3. Replace the dead point with a new sample drawn from the prior, subject to the constraint $\mathcal{L}(\boldsymbol{\theta}) > \mathcal{L}_i$.

4. Update the Bayesian evidence estimate $Z \approx \sum_i \mathcal{L}_i w_i$, where $w_i$ is the prior-volume weight associated with dead point $i$, and repeat from step 2.

5. Terminate when the contribution of the remaining live points to $Z$ falls below a chosen tolerance $\Delta \log Z$.

The weights $w_i$ in step 4 are calculated as the difference between the prior volume $X$ at step $i$ and the volume $X$ at the previous step $i-1$:

$$w_i = X_{i-1} - X_i. \tag{3.9}$$

The prior volume at a given iteration $X_i$ cannot be calculated directly. If that was the case then there would be no need for the NS algorithm. Instead, certain statistical tricks are used to estimate $X_i$. We know that for a given iteration $i-1$, we have $N$ live points uniformly distributed over the prior volume $X_{i-1}$. By definition, the volumes associated with each of those live points are themselves uniformly distributed on $[0, X_{i-1}]$. We can therefore define an associated relative prior volume $u_k = X(\mathcal{L}_k)/X_{i-1}$ which is uniform on $[0,1]$, and since the lowest likelihood point is discarded at each iteration of NS, that point has the largest corresponding prior volume $X_k$, and also largest corresponding $u_k$. We can therefore equate the shrinkage ratio $t_i = X_i/X_{i-1}$ to the maximum of $N$ independent uniform random variables

$$t_i = X_i/X_{i-1} = \max(u_1, u_2, ..., u_N), \tag{3.10}$$

which is a standard order statistic distributed as a beta distribution $\text{Beta}(N,1)$, with expected log-shrinkage $\mathbb{E}[\ln t_i] = -\frac{1}{N}$. Since the prior volume contracts multiplicatively at each iteration, $X_i = X_{i-1}t_i$, the log-volume after $i$ iterations is the sum of $i$ independent log-shrinkages

$$\ln X_i = \sum_{j=1}^{i} \ln t_j, \tag{3.11}$$

with expected value $\mathbb{E}[\ln X_i] = -i/N$. Approximating $X_i$ by its expected log-value gives

$$X_i \approx \exp(-i/N). \tag{3.12}$$

The weights in equation (3.9) can then be approximately computed, and equation (3.7) becomes the finite sum $Z \approx \sum_i \mathcal{L}_i w_i$ over the dead points [27].

Selecting a new live point in step 3 is not trivial. Since the probability of finding a point across the whole prior volume $X_0$ inside $X_i$ is $X_i/X_0$, rejection sampling becomes inefficient as $X_i$ shrinks. `dynesty` [28, 29], which is a Python package that is used in this work for NS and DNS, has many different sampling methods that can be chosen from. In this work, slice sampling is chosen.

A practical side-effect of nested sampling is that in the estimation of the Bayesian evidence, samples of the posterior are produced as well. The dead points are assigned importance weights

$$p_i = \mathcal{L}_i w_i / Z, \tag{3.13}$$

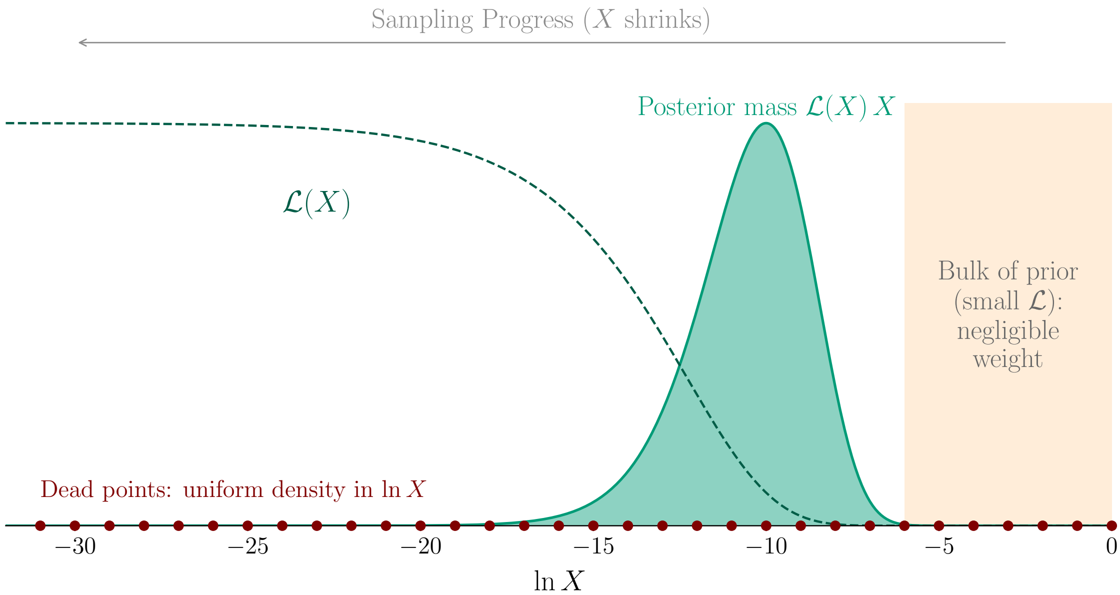



**Figure 3.2:** Example of NS using a Gaussian likelihood and a five-dimensional prior. As the sampling progresses, $\ln X$ becomes increasingly negative. The dead points (maroon) are spread out roughly uniformly along $\ln X$, leading to equal resolution at the bulk of the posterior mass as in the empty prior landscape.

which appropriately assigns each point to the posterior. The problem is that the sampling of the dead points themselves does not differentiate for areas of high posterior mass (see Figure 3.2). In other words, with a fixed number of live points $N$, the posterior is sampled just as densely in areas of low likelihood as it is in areas of high likelihood. One could increase $N$ to get a higher-resolution sampling on the whole posterior, but in the process areas of non-interest also get sampled more, which is computationally inefficient. For estimation of the Bayesian evidence, this is fine, but for detailed posterior mapping, an allocation of sampling resources to areas where the posterior mass $\mathcal{L}(X)X$ is large is preferred [17].

### 3.2.3 Dynamic Nested Sampling (DNS)

With MCMC, we could reliably sample the posterior, but had no way of computing the marginal likelihood in equation (3.6). NS solved the calculation of the Bayesian evidence, but is generally less efficient for posterior recovery. With DNS, regular NS is improved with regard to posterior estimation, by first performing regular NS to calculate the marginal likelihood, and then allocating live points to areas of high interest (using information gained from the first exploration), increasing sampling resolution around areas of high posterior mass [17].

Unlike NS, which uses a fixed number of live points $N$ throughout the entire sampling procedure, DNS lets the number of live points $n_i$ vary between the initial run and subsequent allocation runs. A relative weight $G \in [0, 1]$ is assigned to the DNS algorithm depending on how much one wants it to focus on sampling the posterior, as opposed to calculating the marginal likelihood. $G = 0$ corresponds to optimizing for evidence calculation while $G = 1$ means better posterior estimation. Each dead point after the initial regular NS run is assigned an importance

$$I(G, i) = (1 - G)\frac{I_Z(i)}{\sum_j I_Z(j)} + G\frac{I_{\boldsymbol{\theta}}(i)}{\sum_j I_{\boldsymbol{\theta}}(j)}, \tag{3.14}$$

where $I_Z$, the importance that dead point $i$ holds for evidence calculation, is defined as

$$I_Z(i) \propto \frac{\mathrm{E}[Z_{\geq i}]}{n_i}, \quad Z_{\geq i} \equiv \sum_{k \geq i} \mathcal{L}_k w_k(\mathbf{t}), \tag{3.15}$$

and $I_{\boldsymbol{\theta}}$, the importance corresponding to parameter inference, is defined as

$$I_{\boldsymbol{\theta}}(i) \propto \mathcal{L}_i \mathrm{E}[w_i(\mathbf{t})]. \tag{3.16}$$

After assigning importance, a new sampling run is begun. This time, the sampler runs from likelihood point $\mathcal{L}_{j-1}$ to $\mathcal{L}_{k+1}$ where $j$ and $k$ are the first and last point in the collection of dead points that have importance larger than some predetermined fraction $f$. In this way, the algorithm picks out a region of "high importance" depending on the choice of $G$ and $f$, where the new sampling will take place [17]. Any number of re-runs of this kind can be made in succession, and therefore the accuracy of the sampling method increases with time, until a certain stopping criterion is met.

In this work, `dynesty`'s [28] implementation of DNS is used with an initial baseline run of $N = 500$ live points, an importance weight of $G = 0.8$ and the sampling method slice sampling [23]. This combination prioritizes posterior resolution while still producing accurate evidence estimates. This is appropriate for the task of this work, inferring parameters and calculating Bayes factors. The algorithm is terminated when the default stopping criterion for `dynesty` is met, which is when the contribution of the live points falls below the tolerance $\Delta \log Z = 0.01$.

## 3.3 Likelihood Formulation

Choosing the likelihood can be one of the most difficult tasks in designing a Bayesian framework. The answer often lies in what type of experiment one is trying to model. In this work, we are modeling recoil electrons falling into ($E$, $|p_T|$) bins independently, with a small probability and large number of independent trials ($10^{16}$ EOT). LDMX is therefore what is called a counting experiment, and the count in each bin is Poisson-distributed:

$$P(X = d) = \frac{\lambda^d e^{-\lambda}}{d!}, \tag{3.17}$$

where $d$ is the number of observed events and $\lambda$ the expected number of occurrences [34]. An example of the Poisson distribution for three different $\lambda$'s can be found in Figure 3.3. We can model the likelihood of $d_i$ number of events landing in a specific bin $i$ as the Poisson distribution where we input the number of events we expect in bin $i$ as $\lambda_i$:

$$\mathcal{L}(d_i \mid \lambda_i) = \frac{\lambda_i^{d_i} e^{-\lambda_i}}{d_i!}. \tag{3.18}$$

Since the bins themselves are independent, the combined likelihood over all bins is simply the product over all the individual likelihoods. Labeling the combined data of each bin as $\vec{d} = (d_1, d_2, ..., d_{n_{\text{bins}}})$, the combined expected number of events in each bin as $\vec{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_{n_{\text{bins}}})$, and treating each bin as independent, we get

$$\mathcal{L}(\vec{d} \mid \vec{\lambda}) = \prod_{i=1}^{n_{\text{bins}}} \frac{\lambda_i^{d_i} e^{-\lambda_i}}{d_i!}, \tag{3.19}$$

as the combined likelihood of all the bins. Because the product of many small probabilities tends to become too small to work with, samplers work with the log-likelihood, turning the product into a sum, which is more stable to compute:

$$\ln \mathcal{L}(\vec{d} \mid \vec{\lambda}) = \sum_{i=1}^{n_{\text{bins}}} (d_i \ln \lambda_i - \lambda_i - \ln d_i!). \tag{3.20}$$

This is exactly the likelihood that is used in the sampling done in this work. $\vec{\lambda}$, the combined means of each Poisson-bin, is the simulated signal that has been generated using `MadGraph5`. When it is compared with the data $\vec{d}$, it evaluates the likelihood of the data matching the expected number of counts in each bin as equation (3.20).

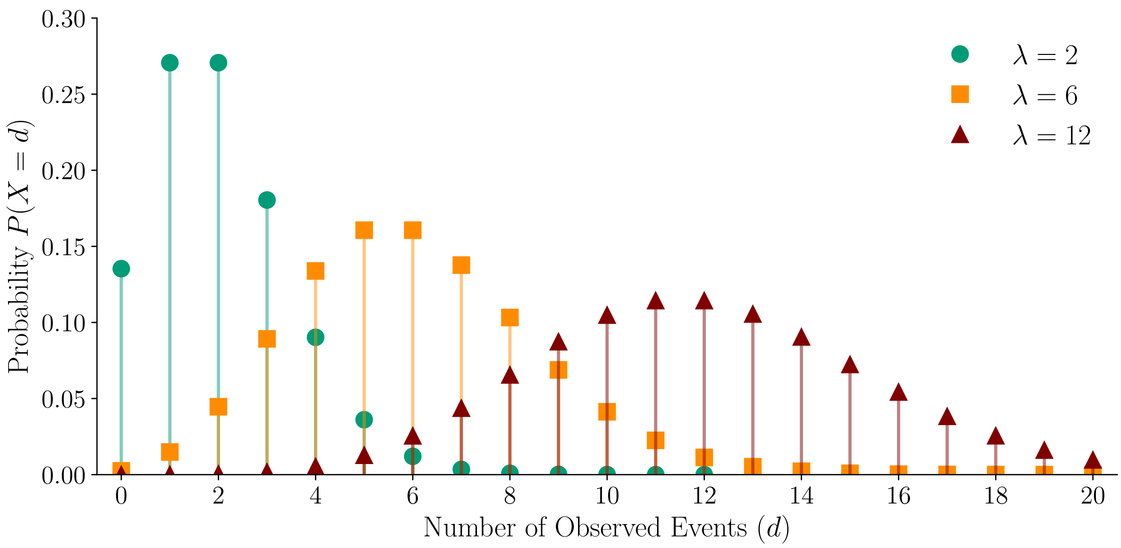



**Figure 3.3:** Example of three Poisson probability distributions with $\lambda = 2, 6, 12$.

Since the $\ln d_i!$ term does not depend on $\vec{\lambda}$, it is simply a normalization term, and can safely be dropped when sampling the posterior with the intention of parameter estimation (where we only care about shape). It can be advantageous to do so since $d_i!$ can become very large for certain data sets, causing problems with computation. However, when calculating the Bayesian evidences, this term would matter since it contributes to the overall normalization of the likelihood. Since this work only uses Bayesian evidence to calculate relative differences between identical data sets using Bayes factors (which are not affected and neither are their uncertainties), this problem is avoided, but it is good to keep it in mind [28].

## 3.4 Nuisance Parameters & Selecting Priors

For the two parameters of interest, we choose an uninformative log-uniform prior with bounds $m_{A'} \in [10^{-3}, 1]$ GeV and $g_f \in [10^{-20}, 1]$. This choice is consistent across all five interaction models, ensuring the prior volume factor cancels in Bayes factor calculation (outlined in section 3.2.3). The choice of upper bound on the $g_f$ prior is well motivated for the Kinetic Mixing model, as $g_f = \epsilon$ is a dimensionless scaling parameter defined on $[0, 1]$. However, the other models' coupling strengths are dimensionful, and could take on values larger than 1. In this work, the prior is shared across all models, ensuring the prior dependence in the marginal likelihood cancels for the calculation of Bayes factors. An analysis was performed after the calculation of the Bayes factors in section 4.2.2, Figure 4.4, to check if the posterior is clipped at the upper boundary of $g_f$. The estimated 99.9th percentile of the recovered coupling never exceeds $\log_{10}(\hat{g}_{f,99.9}) = -1.69$, leaving over 1.6 orders of magnitude of headroom to the upper prior bound. This confirms that the upper bound on the prior for $g_f$ is justified in this analysis. For future work where higher event counts or different model assumptions might push posteriors closer to the upper bound, the prior on $g_f$ should be reconsidered due to the dimensional nature of the coupling for non-KM interactions.

In order to quantify uncertainty about the likelihood, new parameters can be introduced in the sampling process. Often called *nuisance parameters*, these are not of interest when trying to make a prediction of the underlying physics. Instead, the parameters are introduced with a corresponding prior and often, a corresponding uncertainty to allow assumptions of the underlying physics to be malleable during the sampling process. In this work there are several nuisance parameters that could be introduced, each corresponding to an uncertainty one might have about the experimental setup. For example, the number of EOT might be measured by the instruments in the experiment, but keeping track of individual electrons is inherently subject to measurement errors. One could model this uncertainty by adding a prior on the number of EOT $N_e$. A common choice is a normal distribution with a mean around the measured value and a variance corresponding to the uncertainty of the detectors. This gives the sampler room to explore other values for $N_e$ around that measured mean, and therefore allows it to marginalize over the uncertainty, shifting the posterior distribution if the data fits better to a different value for $N_e$. The nuisance parameters can then be marginalized over, propagating their uncertainties onto the marginal posterior distribution of the parameters of interest.

This work employs a hierarchical model, which means the priors of one or several of the nuisance parameters will themselves depend on other nuisance parameters [14]. In this work, three nuisance parameters are used in the inference of $m_{A'}$ and $g_f$; the third is hierarchical and acts as a hyperparameter on the prior of the second. Firstly, we introduce two multiplicative scaling parameters on the signal and background, denoted as $\theta_S$ and $\theta_B$ respectively. For an $E$ and $|p_T|$ bin $i$, we have

$$\lambda_i = S_i \theta_S + B_i \theta_B, \tag{3.21}$$

where $\lambda_i$ is the expected event count, $S_i$ is the contribution from the simulated signal and $B_i$ is the contribution from the background. The primary function of the scaling parameter on the signal, $\theta_S$, is to account for the uncertainty inherent in the production of the simulated data using `MadGraph5`. This uncertainty is estimated to lie around 10%[1], meaning that for a certain count in a specific bin, the expected error is 10% of that count. To simulate this error, the prior distribution for $\theta_S$ is set to a log-normal distribution with median 1 and shape parameter $\sigma = 0.1$. Because the median of a log-normal distribution is

[1]from observation of the errors calculated within `MadGraph5`.

given by $e^{\mu}$, we set $\mu = 0$ in order to enforce equal probability mass on each side of $\theta_S = 1$:

$$\pi(\theta_S) = \ln\mathcal{N}(\mu = 0,\, \sigma_s^2 = 0.01) = \frac{1}{\theta_S \cdot 0.1\sqrt{2\pi}} \exp\left(-\frac{(\ln\theta_S)^2}{0.02}\right). \tag{3.22}$$

This prior lets the sampler explore a region around the median $\theta_S = 1$ to account for the error in the simulation software. Furthermore, the background scale factor $\theta_B$ is also chosen to have a log-normal prior with a median at 1. This ensures that a negative value for $\theta_B$ is not allowed, which is especially important since we will make the standard deviation (or shape parameter) of this log-normal our third and final nuisance parameter. The log-normal prior for $\theta_B$ takes the form

$$\pi(\theta_B) = \ln\mathcal{N}(\mu = 0, \sigma_B^2) = \frac{1}{\theta_B\sigma_B\sqrt{2\pi}} \exp\left(-\frac{(\ln\theta_B)^2}{2\sigma_B^2}\right). \tag{3.23}$$

Since $\theta_S$ and $\theta_B$ are multiplicative scaling parameters, log-normals are a natural choice as they span the positive real numbers and are symmetric in log space. Finally, since we are unsure about the uncertainty of $\theta_B$, the hyperparameter $\sigma_B$ is also given a prior. We know that an uncertainty cannot be negative, and that it is probably relatively small in comparison to the order of magnitude of the parameter it is assigned to. It is therefore motivated to choose a half-normal for $\pi(\sigma_B)$, i.e. a normal distribution centered at 0 but cut off at the peak, allowing only positive numbers. This ensures a maximum prior density at 0, yet the prior is limited to the positive real numbers. Distributions of this kind, such as the half-normal or the heavy-tailed half-Cauchy, have been advocated over inverse-gamma priors for hierarchical variance parameters [13]. The choice of standard deviation for $\pi(\sigma_B)$ is relatively arbitrary, but it is chosen in this work to be $\sigma_{\sigma_B} = 0.5$. This provides a weakly informative constraint in order not to let $\theta_B$ vary too much during the sampling procedure. The equation for the half-normal prior on $\sigma_B$ is

$$\pi(\sigma_B) = \mathcal{N}^+(\mu = 0, \sigma_{\sigma_B}^2 = 0.25) = \sqrt{\frac{2}{\pi\sigma_{\sigma_B}^2}} \exp\left(-\frac{\sigma_B^2}{2\sigma_{\sigma_B}^2}\right) = \sqrt{\frac{8}{\pi}} \exp\left(-2\sigma_B^2\right), \quad \sigma_B \geq 0. \tag{3.24}$$

Generally, priors with hard boundaries like the half-normal can be problematic in sampling algorithms like MCMC. Walkers can get stuck at the boundaries, since half of the proposed steps are likely to be rejected. The hard boundary is handled by the prior transform function necessary for DNS, as it maps the unit cube $u \in [0,1]^d$ to the whole parameter space. The three priors on the nuisance parameters $\theta_S$, $\theta_B$ and the hyperparameter $\sigma_B$ can be seen plotted in Figure 3.4.

## 3.5 Model Comparison using Bayes Factors

One conceivable approach to model selection is the introduction of the theoretical model itself as another parameter with a likelihood function capable of transcending models. Given a data set, one could then let the sampler explore this new additional parameter and it would give a posterior distribution of what model fits the data best. However, this would require the definition of a complex likelihood that takes into account different simulated data sets and somehow compares them within the sampling process at each iteration and/or walker step. Instead, this work employs a more computationally simple method for model comparison, given that one has means of calculating the marginal likelihood: Bayes factors.

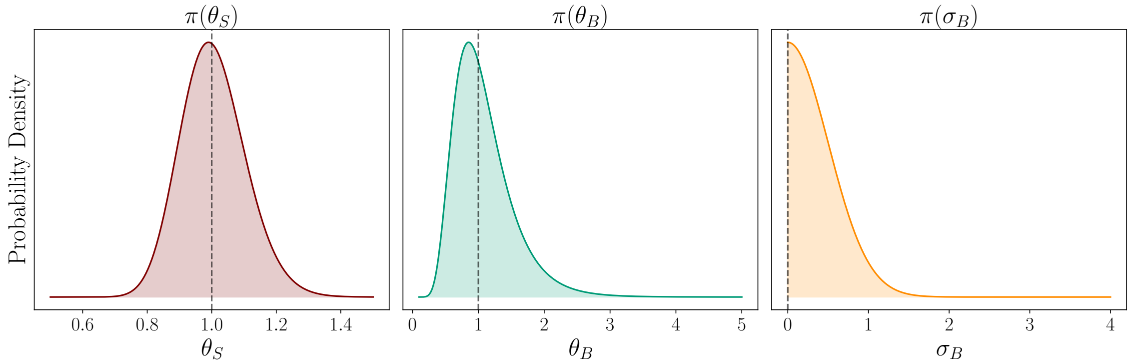



**Figure 3.4:** Prior distributions on the three nuisance parameters: log-normal on $\theta_S$ with fixed $\sigma_S = 0.1$ **(left)**; log-normal on $\theta_B$ shown with a representative $\sigma_B = 0.5$ **(center)**, where $\sigma_B$ itself is sampled from its hyperprior, a half-normal on $\sigma_B$ with scale 0.5 **(right)**.

The Bayes factor $K_{\mathrm{AB}}$ for the comparison of model A and model B is defined as the ratio of the two Bayesian evidences:

$$K_{\mathrm{AB}} = \frac{Z_{\mathrm{A}}}{Z_{\mathrm{B}}} = \frac{\int_{\Omega_{\boldsymbol{\theta}}} \mathcal{L}(\mathcal{D} \mid \boldsymbol{\theta}, \mathrm{A}) \pi(\boldsymbol{\theta} \mid \mathrm{A}) d\boldsymbol{\theta}}{\int_{\Omega_{\boldsymbol{\theta}}} \mathcal{L}(\mathcal{D} \mid \boldsymbol{\theta}, \mathrm{B}) \pi(\boldsymbol{\theta} \mid \mathrm{B}) d\boldsymbol{\theta}}, \tag{3.25}$$

where in this work, the prior $\pi(\boldsymbol{\theta})$ is for the primary analysis not model dependent, canceling in the calculation of the Bayes factor. A Bayes factor $K_{\mathrm{AB}} > 1$, or $\ln K_{\mathrm{AB}} > 0$ indicates that the marginal likelihood, defined in equation (3.6), is greater for model A than model B, and therefore that model A is more strongly supported by the data. By this definition, this work labels model A as the *true model* and model B as the *hypothesis model*. Mock data will be generated using the true model, after which Bayes factors are calculated by comparing the recovered marginal likelihoods for each hypothesis model with the true model by sampling the posterior using DNS.

The marginal likelihood measures the average predictive success of the model, as it integrates the likelihood over all possible parameter combinations. Given this definition, the Bayes factor inherently brings with it an Occam's razor penalty for complex models. If the likelihood peak is large, due to a good fit to the data through a model with many parameters, those extra parameters also result in a larger dimensional parameter space. The average is therefore diluted in poorly fitted regions, and the penalty is only overcome if the fit is good enough. In this work, all five models share identical parameters and priors, so the Occam penalty has no effect; the Bayes factors instead reflect relative goodness-of-fit alone [34].

To perform the model comparison for this work, we simply modify our likelihood to depend on model selection. The Poisson log-likelihood in equation (3.20) will have a different mean $\vec{\lambda}_{\mathcal{M}}$ depending on which DP interaction model is assumed to be the correct one. In this way, a "hypothesis model" $\mathcal{M}$ can be selected for the likelihood before sampling the posterior using DNS. The sampler tries to fit the parameters to the data and simultaneously calculates the Bayesian evidence $Z_{\mathcal{M}}$. This is repeated for all five interaction models, yielding a marginal likelihood for each. These Bayesian evidences can then be compared through the Bayes factor, which gives an indication of which model is more favored by the data.

# 4. Results

This chapter contains the results of the various tests performed on the Bayesian framework. Section 4.1 contains the DNS sampling of the posterior for four relic-target KM benchmarks. A few different variations of the benchmarks are reported. The benchmarks were run with both background scaling BG = 1 and BG = 100, with and without the $E < 3.16$ GeV cut. This is to show the effects of a significant number of background events at LDMX, and determine what kind of effect it will have on the inference of the model parameters. A cross-validation between MCMC and DNS is made in section 4.1.2, in order to confirm the framework's consistency across samplers. The model comparison using Bayes factors is presented in section 4.2. Here, a noise floor estimation is included, along with Bayes factors calculated from running the sampler with a model-specific likelihood on data generated from each interaction model. There is also a bias estimation showing how much the sampler drifts in its median prediction if the wrong model is used in the likelihood. Section 4.3 shows that introducing a relic target prior on $g_f$ breaks degeneracy between KM and [C, A].

## 4.1 Parameter Inference on KM Benchmarks

The four benchmarks on which the Bayesian framework has been tested are presented and color-coded in Table 4.1. The benchmarks can be grouped into two groups: [BM 1, BM 2], which are taken from relic target calculations assuming a complex scalar DM model, a dark-sector-fine-structure constant $\alpha_D = 0.5$ and a DP to DM mass-ratio of $R = m_{A'}/m_\chi = 2.5$. The second group, [BM 3, BM 4], assumes $R = 2.2$. These values were extracted by digitizing Figure 6 in [6]. For all benchmarks, the true values of the nuisance parameters $\theta_S$ and $\theta_B$ are set each to 1.0.

**Table 4.1:** The four KM relic target benchmarks that were used to test the Bayesian framework. Color coded to match to posterior figures.

| Parameters | BM 1 | BM 2 | BM 3 | BM 4 |
|---|---|---|---|---|
| $m_{A'}$ [GeV] | 0.1 | 0.01 | 0.1 | 0.01 |
| $g_f$ | $6.3 \cdot 10^{-5}$ | $4.8 \cdot 10^{-6}$ | $6.0 \cdot 10^{-6}$ | $5.2 \cdot 10^{-7}$ |
| $\mathcal{R}$ | 2.5 | 2.5 | 2.2 | 2.2 |

### 4.1.1 Posterior Recovery with DNS

The posteriors shown in Figures 4.1 and 4.2 are of a representative realization, chosen as the run which produced the median prediction $(\hat{m}_{A'}, \hat{g}_f)$ closest to the median of 10 Poisson realizations of the KM data. This is done for both background scaling BG = 1 and BG = 100. The corner plots contain the marginal posteriors of all five parameters, including the nuisance parameters $\theta_S$, $\theta_B$ and the hyperparameter $\sigma_B$. As is apparent, the fitting to the data becomes progressively less stable for each benchmark. This is mainly due to the fact that the coupling strength $g_f$ acts as a scaling parameter on the expected number of events, $N \propto g_f^2$. Since the coupling strength drops approximately two orders of magnitude between BM1 and BM4, and BM4 has an order of magnitude lower $m_{A'}$[1], $N$ drops by a factor of $\approx 70$ to only $3-4$ total events in a given Poisson realization (see Appendix A). This is simply too little information for the sampler to go on, which leads to wide posteriors on $m_{A'}$, spanning the whole prior range. This effect is even more noticeable when the background is scaled by 100. In Figure 4.2, it can be seen that for benchmark 3 and 4, the large background of BG = 100 disrupts the signal. The posterior has difficulty reading the shape and scaling in the signal data, apparent by the mostly uniform, uninformative shape on both $m_{A'}$ and $g_f$. The modes of the posteriors sometimes still end up around the truth, indicating that a large variety of parameters agree with the data, but values close to the truth are still preferred.

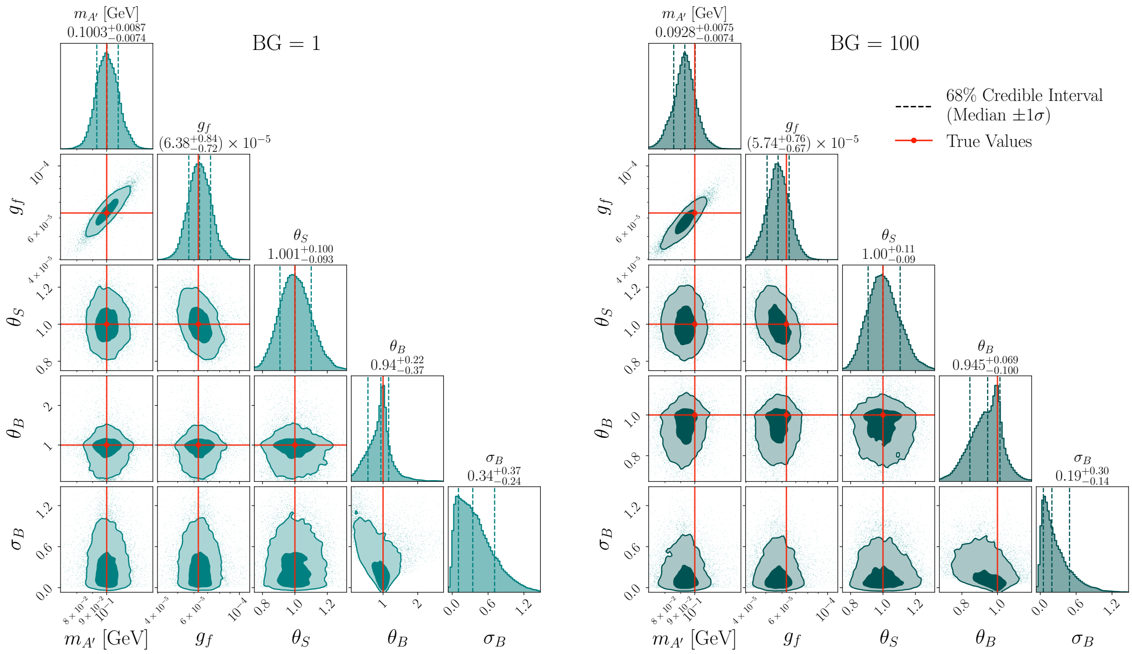
The combined statistics for the 10 Poisson realizations are shown in Table 4.2. Here, the values presented are all the medians of the median predictions across the 10 Poisson realizations. The +/- values on these predictions represent the medians of the 16th and 84th percentile bounds on the posteriors. The values from the analysis with the $E < 3.16$ GeV cut are also included.

**Table 4.2:** Recovered $(\hat{m}_{A'}, \hat{g}_f)$ posteriors for the four KM benchmarks using DNS. Values are the median across $n = 10$ independent Poisson realizations of the per-realization posterior median, with 68% credible intervals given by the median across realizations of the 16th and 84th percentile bounds. The corner plots in Figures 4.1 and 4.2 show the representative realization (closest to the median across runs in the $(m_{A'}, g_f)$ plane) and may differ slightly from the values reported here. Rows with a cut value include the $E < 3.16$ GeV cut (Section 2.4.2).

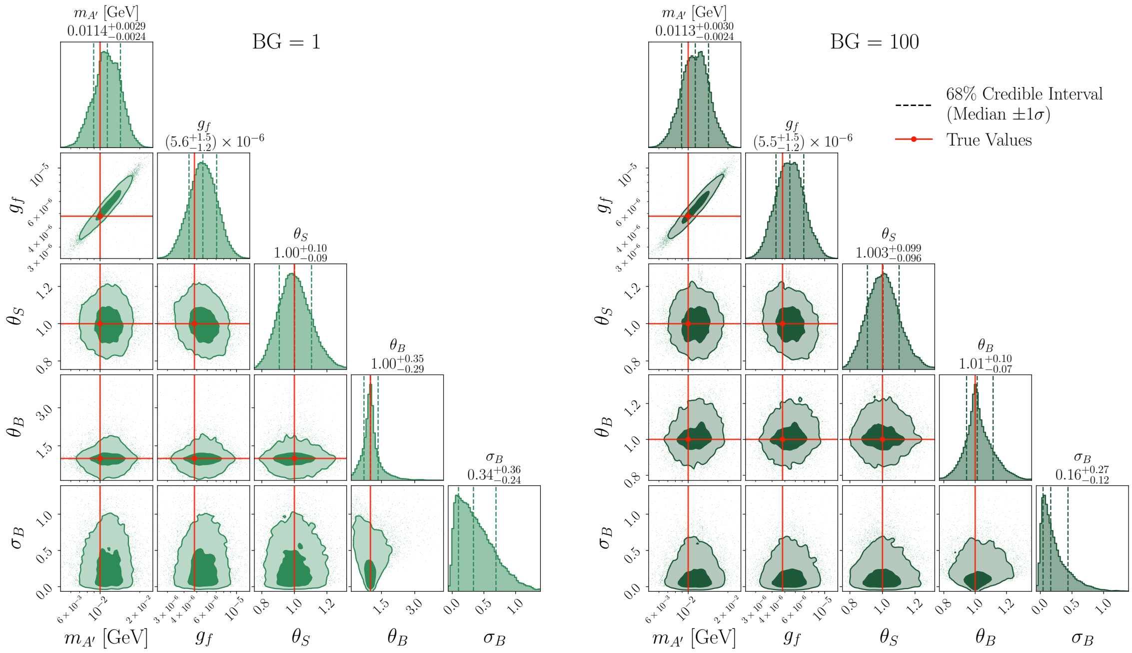
| BM | $m_{A'}^{\text{true}}$ [GeV] | $g_f^{\text{true}}$ | BG | Cut [GeV] | $\hat{m}_{A'}$ [GeV] | $\hat{g}_f$ |
|---|---|---|---|---|---|---|
| 1 | 0.10 | $6.3 \times 10^{-5}$ | 1 | — | $0.099^{+0.009}_{-0.007}$ | $(6.3^{+0.8}_{-0.8}) \times 10^{-5}$ |
| | | | 100 | — | $0.096^{+0.005}_{-0.007}$ | $(6.1^{+0.6}_{-0.7}) \times 10^{-5}$ |
| | | | 1 | 3.16 | $0.096^{+0.009}_{-0.009}$ | $(6.1^{+0.9}_{-0.8}) \times 10^{-5}$ |
| | | | 100 | 3.16 | $0.097^{+0.008}_{-0.008}$ | $(6.1^{+0.8}_{-0.8}) \times 10^{-5}$ |
| 2 | 0.01 | $4.8 \times 10^{-6}$ | 1 | — | $0.0114^{+0.0029}_{-0.0024}$ | $(5.6^{+1.4}_{-1.2}) \times 10^{-6}$ |
| | | | 100 | — | $0.0108^{+0.0029}_{-0.0023}$ | $(5.2^{+1.5}_{-1.1}) \times 10^{-6}$ |
| | | | 1 | 3.16 | $0.012^{+0.004}_{-0.003}$ | $(5.7^{+2.0}_{-1.4}) \times 10^{-6}$ |
| | | | 100 | 3.16 | $0.011^{+0.004}_{-0.003}$ | $(5.2^{+1.9}_{-1.3}) \times 10^{-6}$ |
| 3† | 0.10 | $6.0 \times 10^{-6}$ | 1 | — | $0.040^{+0.285}_{-0.035}$ | $(0.45^{+5.64}_{-0.36}) \times 10^{-6}$ |
| | | | 100 | — | $0.058^{+0.298}_{-0.052}$ | - |
| | | | 1 | 3.16 | $0.038^{+0.282}_{-0.034}$ | $(0.38^{+4.80}_{-0.29}) \times 10^{-6}$ |
| | | | 100 | 3.16 | $0.059^{+0.291}_{-0.054}$ | - |
| 4† | 0.01 | $5.2 \times 10^{-7}$ | 1 | — | $0.020^{+0.098}_{-0.017}$ | $(0.48^{+2.47}_{-0.40}) \times 10^{-6}$ |
| | | | 100 | — | $0.032^{+0.311}_{-0.029}$ | - |
| | | | 1 | 3.16 | $0.017^{+0.092}_{-0.014}$ | $(0.40^{+2.57}_{-0.33}) \times 10^{-6}$ |
| | | | 100 | 3.16 | $0.034^{+0.316}_{-0.031}$ | - |

† The posteriors for both $m_{A'}$ and $g_f$ span a large fraction of the prior range. For most of these prior-dominated cases the median prediction and 68% CI cannot be attributed to parameter recovery. See Figure 4.2 for the representative posteriors.

[1] which actually contributes to a higher $N$, since the production cross-section grows as $m_{A'}$ decreases.

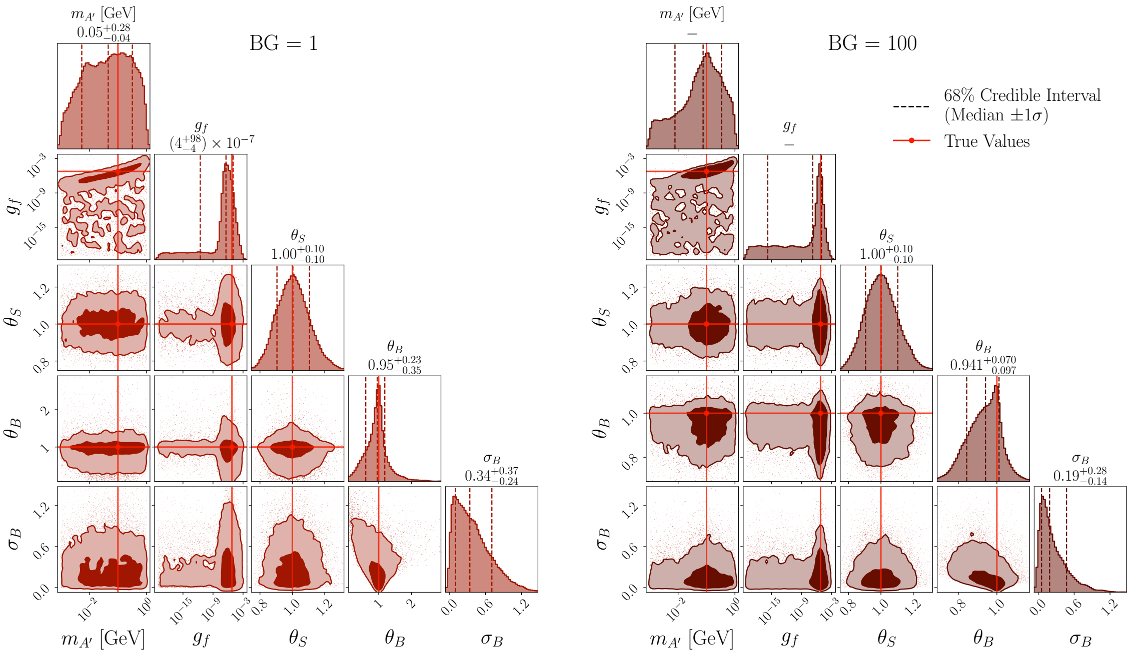



**(a)** BM 1

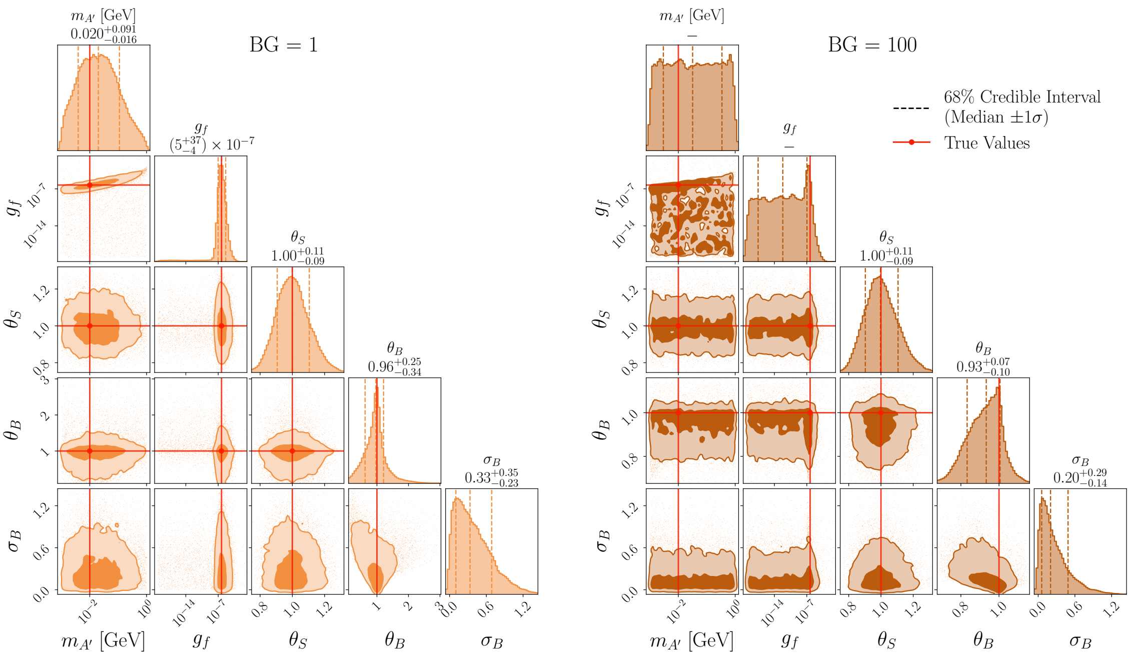



**(b)** BM 2

**Figure 4.1:** Posterior distributions over the five parameters $(m_{A'}, g_f, \theta_S, \theta_B, \sigma_B)$ recovered for Benchmark 1 (a) and Benchmark 2 (b) with the Kinetic Mixing model using DNS. For each benchmark, the left corner plot shows the posteriors when the background has been scaled by a factor of 1, and the right by a factor of 100. The dashed lines encase the 68% credible interval and median prediction of the five marginalized distributions corresponding to each parameter. The two-dimensional distributions are combinations of each of the five marginal posteriors, and their shaded regions encompass 39.3% and 86.5% of the posterior mass.



**(a)** BM 3



**(b)** BM 4

**Figure 4.2:** Posterior distributions over the five parameters $(m_{A'}, g_f, \theta_S, \theta_B, \sigma_B)$ recovered for Benchmark 3 (a) and Benchmark 4 (b) with the Kinetic Mixing model using DNS. For each benchmark, the left corner plot shows the posteriors when the background has been scaled by a factor of 1, and the right by a factor of 100. The dashed lines encase the 68% credible interval and median prediction of the five marginalized distributions corresponding to each parameter. The two-dimensional distributions are combinations of each of the five marginal posteriors, and their shaded regions encompass 39.3% and 86.5% of the posterior mass.

The $E < 3.16$ GeV cut leaves the BM1 and BM2 predictions largely unchanged, but pushes BM3 and BM4 further into non-detection. Most indicative of this is the BM3 and BM4, BG = 100 cases where the median 16th and 84th percentile are unconstrained, spanning several orders of magnitude from the median $\hat{g}_f$.

### 4.1.2 MCMC vs DNS Cross-Validation

In this section, the results for posterior recovery using MCMC on the four benchmarks are presented and cross-validated with DNS. The comparison is done using the same Poisson realization, to accurately determine the differences between the two sampling methods. Similarly to section 4.1.1, the benchmarks have been run with a total of 10 Poisson realizations each, where the median realization is selected based on the distance its $(\hat{m}_{A'}, \hat{g}_f)$ pair has from the median of all $(\hat{m}_{A'}, \hat{g}_f)$ pairs across the 10 realizations. The MCMC algorithm was run using the `emcee` package, with 5000 steps as burn-in period followed by $N_{\text{steps}} = 10^4$ sampling steps, using $n_w = 50$ walkers. The DNS posteriors are the same as in section 4.1.1, using $N = 500$ live points, $G = 0.8$ and slice sampling.

A comparison between the two-dimensional posteriors over $m_{A'}$ and $g_f$ recovered with both sampling methods is presented in Figure 4.3. We see that both samplers perform very similarly in parameter inference and posterior recovery, however MCMC fails to capture the entire tail of the posterior in the same way as DNS in benchmark 3. This could probably be improved by using a larger number of sampling steps in the MCMC run, leading to walkers eventually exploring the entire posterior. The fact that DNS and MCMC have nearly identical posterior recovery means they validate each other and the framework implementation well. For this reason, DNS is suitable as primary sampling method in this work, as it also calculates the Bayesian evidence.

The MCMC runs converged adequately, as indicated by the maximum autocorrelation times $\tau_{\max}$ in table 4.3. Here, the relevant diagnostics of the MCMC chain are presented for each benchmark. Mean acceptance fraction $\langle a \rangle$ slightly below 0.5 indicates that roughly half of the suggested steps were accepted. The minimum Effective Sample Size ($\text{ESS}_{\min}$) measures how many truly independent samples were gathered at a minimum for the five parameters.

**Table 4.3:** MCMC sampler diagnostics across the four KM benchmarks, with background scaling BG = 1. Values are means/medians over $n = 10$ Poisson realizations. Acceptance fraction $\langle a \rangle$ should lie in $[0.2, 0.5]$ for healthy stretch-move sampling [12]. The autocorrelation time $\tau_{\max}$ is the maximum across the 5 sampled parameters; $\text{ESS}_{\min}$ is the corresponding minimum effective sample size. Chain length $N = 10^4$ steps per walker, with $n_w = 50$ walkers.

| BM | $\langle a \rangle$ | median $\tau_{\max}$ | max $\tau_{\max}$ | min $\text{ESS}_{\min}$ | $N(\text{ESS} \geq 2000)$ |
|---|---|---|---|---|---|
| 1 | 0.46 | 132 | 281 | 1779 | 7/10 |
| 2 | 0.43 | 194 | 504 | 991 | 9/10 |
| 3 | 0.42 | 139 | 289 | 1728 | 9/10 |
| 4 | 0.42 | 182 | 353 | 1416 | 9/10 |

Table 4.4 shows how many of the 10 Poisson realizations recovered a median parameter prediction within a 68% CI for both sampling methods. This number does not indicate

detection, as a posterior spanning multiple orders of magnitude has a very wide 68% CI. See for example Figure 4.2b BG = 100, where a posterior spanning the prior of $m_{A'}$ still has the truth within its 68% CI.

**Table 4.4:** Cross-method validation: number of realizations (out of $n = 10$) in which the recovered 68% credible interval contains the true parameter value. Results from both DNS and MCMC are presented at background levels BG = 1 and BG = 100.

| BM | BG | Method | $m_{A'}$ | $g_f$ | $\theta_S$, $\theta_B$ |
|---|---|---|---|---|---|
| 1 | 1 | DNS | 7/10 | 7/10 | 10/10, 10/10 |
| | | MCMC | 7/10 | 9/10 | 10/10, 10/10 |
| | 100 | DNS | 10/10 | 2/10 | 10/10, 7/10 |
| | | MCMC | 10/10 | 2/10 | 10/10, 7/10 |
| 2 | 1 | DNS | 7/10 | 7/10 | 10/10, 10/10 |
| | | MCMC | 8/10 | 8/10 | 10/10, 10/10 |
| | 100 | DNS | 8/10 | 8/10 | 10/10, 8/10 |
| | | MCMC | 8/10 | 8/10 | 10/10, 9/10 |
| 3 | 1 | DNS | 7/10 | 3/10 | 10/10, 10/10 |
| | | MCMC | 7/10 | 3/10 | 10/10, 10/10 |
| | 100 | DNS | 9/10 | 5/10 | 10/10, 8/10 |
| | | MCMC | 9/10 | 5/10 | 10/10, 7/10 |
| 4 | 1 | DNS | 10/10 | 7/10 | 10/10, 10/10 |
| | | MCMC | 10/10 | 7/10 | 10/10, 10/10 |
| | 100 | DNS | 6/10 | 6/10 | 10/10, 8/10 |
| | | MCMC | 6/10 | 6/10 | 10/10, 8/10 |

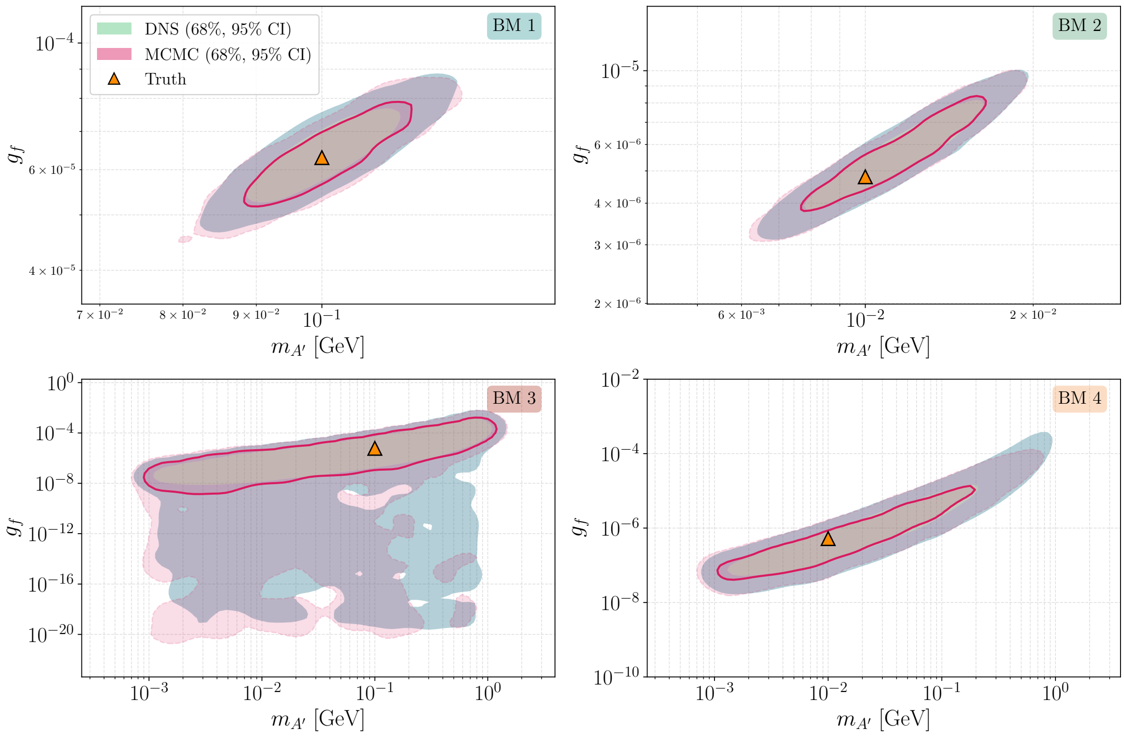



**Figure 4.3:** Comparison between MCMC sampling and DNS sampling on the same median Poisson realization out of 10 with BG = 1. The two recovered two-dimensional marginal posteriors are plotted on top of each other for comparison. Each subplot shows which benchmark was used in the generation of the data, and the orange triangle marks the true benchmark values. The color gradient on the DNS posterior and red outline for MCMC mark the two-dimensional 68% and 95% credible intervals.

## 4.2 Dark Photon Model Comparison

In addition to inferring the model parameters of a single model, the Bayesian framework is equipped to compare competing models. One might be interested to know if, given a data set, which of the five interaction models considered in this work (Kinetic Mixing (KM), Magnetic Dipole (M), Electric Dipole (E), Charge Radius (C) and Anapole Moment (A)), is most likely to be the one to have produced said data. This comparison is done using the Bayes factor, which is a ratio of evidences $K_{\mathrm{AB}} = \frac{Z_{\mathrm{A}}}{Z_{\mathrm{B}}}$, where A is the true model and B is the hypothesis model. The Bayes factor can be interpreted using Jeffreys' scale [18], where $|\ln K| > 5$ is by convention taken as decisive evidence that the hypothesis model can be rejected. For each pair of true model and hypothesis model, the Bayesian evidences are computed using DNS on $n = 5$ identical Poisson realizations on the data, such that data-dependent likelihood normalizations (see section 3.3) are canceled in the calculation of $\ln K$. The choice of lowering the number of Poisson realizations from $n = 10$ was made to save on computation time. The data is generated using each model under $m_{A'} = 0.1$ GeV and corresponding $g_f$ consistent with relic targets for the respective model, which are presented in table 4.5. The values are extracted in the same way as the relic targets in Figure 1 in [9], but for $R = 2.5$.

**Table 4.5:** Relic target coupling values $g_f$ used as the generative-model couplings in the Bayes factor analysis, evaluated at the benchmark mass $m_{A'} = 0.1$ GeV. $N_{\mathrm{exp}}$ is the total number of expected signal events for the benchmark.

| Model | $g_f$ | $N_{\mathrm{exp}}$ |
|---|---|---|
| KM | $6.3 \times 10^{-5}$ | 250 |
| M | $3.8 \times 10^{-3} \cdot e$ | 3900 |
| E | $3.8 \times 10^{-3} \cdot e$ | 3900 |
| C | $3 \times 10^{-2} \cdot e$ | 5300 |
| A | $3 \times 10^{-2} \cdot e$ | 5300 |

In section 4.2.1, there is a discussion about the relative noise floor in these calculations due to the inherent Monte Carlo stochasticity in the simulation of the data using `MadGraph5` [5], along with a description of how this noise floor was estimated. The Bayes factors calculated using combined $E$ and $|p_T|$ kinematics as well as $E$ and $|p_T|$ separately are presented and compared in section 4.2.2. There is also a bias estimation under model misspecification, showing how much the recovered median $\hat{m}_{A'}$ and $\hat{g}_f$ could differ from the true values if the analysis is performed using the wrong interaction model on LDMX data, presented in section 4.2.3.

### 4.2.1 Noise Floor Estimation

The signal kinematics used in both data generation and likelihood evaluation are Monte Carlo samples from `MadGraph5`, bringing with them an inherent uncertainty. Two independent runs on the same DP interaction model using the same benchmark therefore yield slightly different kinematic distributions. Given a comparison between two such batches, the Bayesian evidence difference $\Delta \ln Z$ will be non-zero even though they both describe the physics of the same model. This sets a noise floor below which Bayes factors cannot be attributed to a model difference.

To approach this problem, a simple comparison was made using a different batch of simulated KM data ($KM_{new}$). This data set was compared with the data used throughout this work, $KM_{old}$, which is simulated under the exact same parameters. Data were generated using both $KM_{new}$ and $KM_{old}$. The difference between the Bayesian evidences $\Delta \ln Z = \ln Z_{new} - \ln Z_{old}$ was computed across $n = 10$ Poisson realizations, with $KM_{old}$ in the role of the hypothesis model. The benchmarks create roughly the same total number of events in the bins for the pairs [M, E] and [C, A] respectively, but they differ a lot between those groups. This means the statistics are different, so the event counts in each of the data generations were scaled to match the M-model and C-model relic-target benchmarks, in order for the statistics to remain the same as in the Bayes factor analysis in section 4.2.2.

Table 4.6 reports median $\Delta \ln Z$, standard deviation $\sigma$, and min/max bounds across the $n = 10$ realizations for each calibration.

**Table 4.6:** Noise floor estimate for the Bayes factor calculations. Two independent `MadGraph5` batches of the KM signal ($KM_{old}$ and $KM_{new}$) were generated under identical physical parameters, with event counts scaled to match the KM-, M- and C-model relic-target benchmarks. For each Poisson realization, signal-plus-background data were drawn from both batches at a shared Poisson seed, and the log evidence of each dataset was computed under a common $KM_{old}$ likelihood. The reported quantity $|\Delta \ln Z| = |\ln Z_{new} - \ln Z_{old}|$ measures the shift in evidence induced by `MadGraph5` stochasticity. The minimum and maximum values characterize the signed scatter around zero.

| Calibration | median $\lvert\Delta \ln Z\rvert$ | $\sigma$ | $[\min, \max]$ |
|---|---|---|---|
| Scaled to KM-model event count | 2.4 | 9.5 | [−10.9, +25.4] |
| Scaled to M-model event count | 10.3 | 12.1 | [−3.9, +32.6] |
| Scaled to C-model event count | 9.2 | 13.0 | [−13.3, +23.7] |

This noise floor estimation is very rough due to the low sample size, and is meant to serve the discussion of the results in section 4.2.2, rather than give a concrete quantitative calculation of the stochastic error in the `MadGraph5` simulations.

### 4.2.2 Bayes Factors

Figure 4.4 shows the results of the Bayes factor calculations on relic target benchmarks for the five interaction models. We see a clear separation between two groups, namely [KM, C, A] and [M, E], which is in agreement with the work of Catena et al. [9]. The models within each group produce similar kinematic distributions for a given $m_{A'}$, where only the number of events, $N$, differs. It is therefore possible for the sampler to match a different model within the same group to the number of events $N$ in a given distribution simply by fitting to a different $g_f$, as $N \propto g_f^2$. Looking at the Bayes factors it is clear that the separation between these two groups is decisive, given Jeffreys' scale threshold of $|\ln K| > 5$. When comparing the two groups, Bayes factors reach all the way from 38 to 591, where the scale is assumed to be due to the large difference in statistics between benchmarks.

The errors reported in Figures 4.4 and 4.5 are median absolute deviations (MAD), across the five Poisson realizations of each cell. For a fixed cell, let $\ln K_i$ be the value in Poisson realization $i = 1, ..., 5$, and median($\ln K^*$) the median over those five. The MAD is calculated as the median across realizations of the absolute value of the difference

$$\mathrm{MAD} = b \cdot \mathrm{median}(|\ln K_i - \mathrm{median}(\ln K^*)|), \tag{4.1}$$

where $b = 1.4826$ in order to make MAD consistent with $1\sigma$ for Gaussian distributions [25], and median($\ln K^*$) is the median of all five values of $\ln K$ across the Poisson realizations. MADs are more resistant to outliers than standard deviation, which is a real risk with only five realizations. A large spread between the MAD values can be seen in the cells of Figures 4.4 and 4.5, which is explained by the low number of Poisson runs.

The Bayesian evidences carry with them an intrinsic error due to the stochastic nature of DNS. This uncertainty is reported by `dynesty` as $\sigma_{\ln Z}$, which propagates to the Bayes factors through their calculation $\ln K = \ln Z_{\text{true}} - \ln Z_{\text{hypothesis}}$. These errors are independent, so the sampler uncertainty on $\ln K$ can be calculated using the standard propagation of errors formula:

$$\sigma_{\ln K} = \sqrt{\sigma^2_{\ln Z_{\text{true}}} + \sigma^2_{\ln Z_{\text{hypothesis}}}}. \tag{4.2}$$

Across the model pairs and all five realizations, $\sigma_{\ln K}$ has a median of roughly 0.11 nats, which rises to 0.15 for higher statistics benchmarks like those using models M and E as the true model. Looking at Figure 4.4, these numbers are very low compared to the decisive Bayes factors ranging from ≈38-591 nats. The sampler precision is therefore negligible compared to the discriminating power of the relic target data, the Jeffreys' scale threshold of $|\ln K| > 5$, the calculated noise floor and the calculated MADs. Although the uncertainty is of the correct order of magnitude of the non-discriminating Bayes factors in Figure 4.4, these values are not taken as indicating any sort of information regarding model selection in this work.

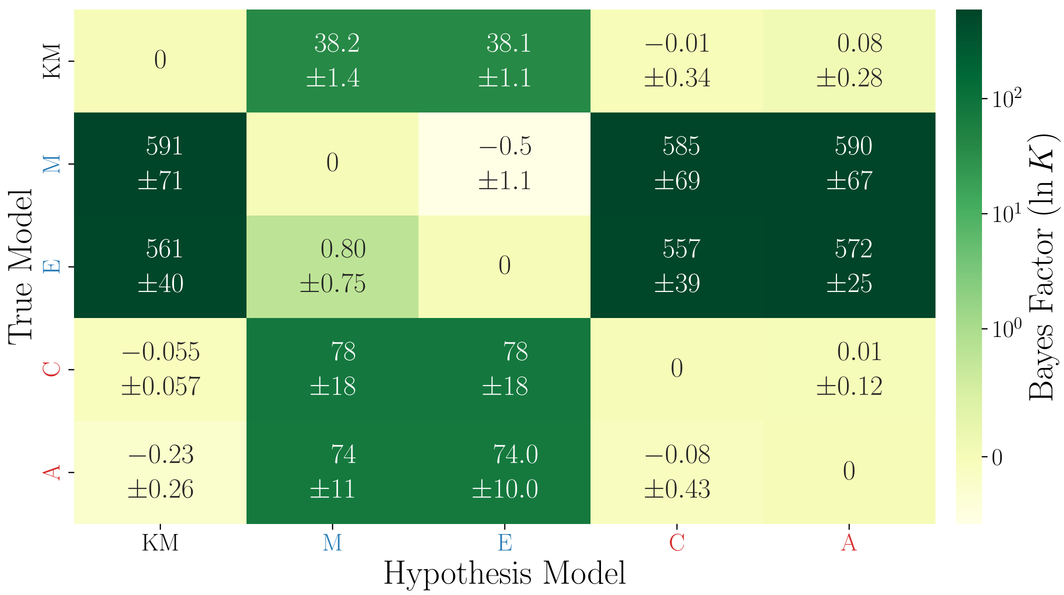



**Figure 4.4:** Bayes factors recovered with Dynamic Nested Sampling (DNS) by comparing marginal likelihoods between the five interaction models using both $E$ and $|p_T|$ kinematics. The y-axis shows what model was used in the generation of the data, while the x-axis shows what model was used in the likelihood while sampling. The numbers reported are $\ln K$, where $K = \frac{Z_{\text{true}}}{Z_{\text{hypothesis}}}$. A larger $\ln K$ means higher confidence that corresponding hypothesis model is ruled out.

The Bayes factor analysis was also run using only total energy $E$ or transverse momentum $|p_T|$ kinematics from the recoil electrons, reported in Figures 4.5 and 4.6. It is apparent that using only one-dimensional data in the distributions leads to lower statistics, indicated by lower overall Bayes factors. This is not surprising, as not only is one whole dimension of shape information lost, but there is also a decrease in the number of bins in the kinematic distribution from 50x50 = 2500 to only 50. The Bayes factors sometimes fall below the estimated two-dimensional noise floor, but no estimation was made for the one-dimensional case.

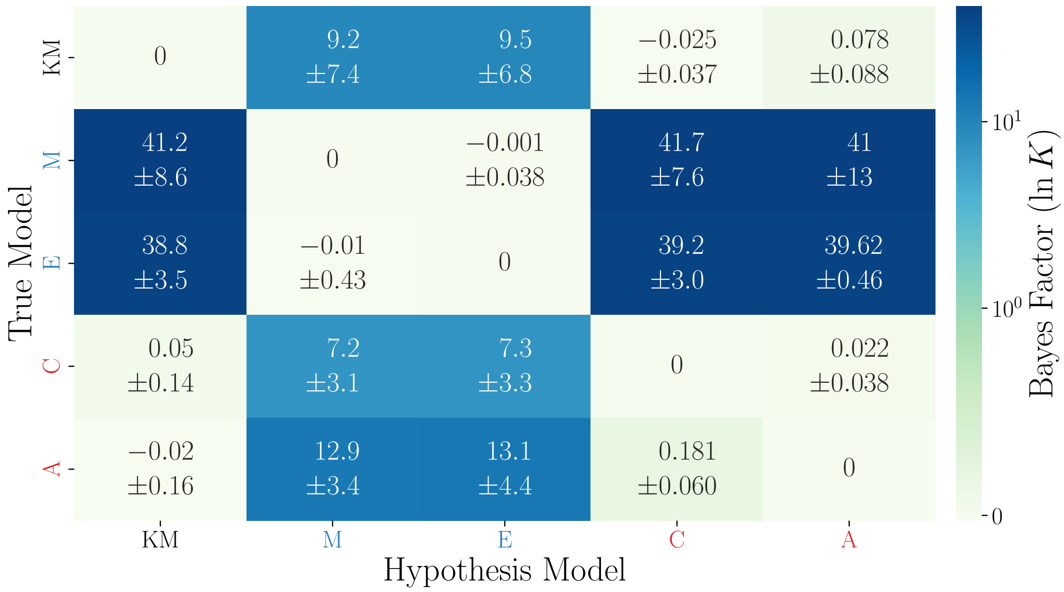



**Figure 4.5:** Same as Figure 4.4 using only total energy $E$ kinematics.

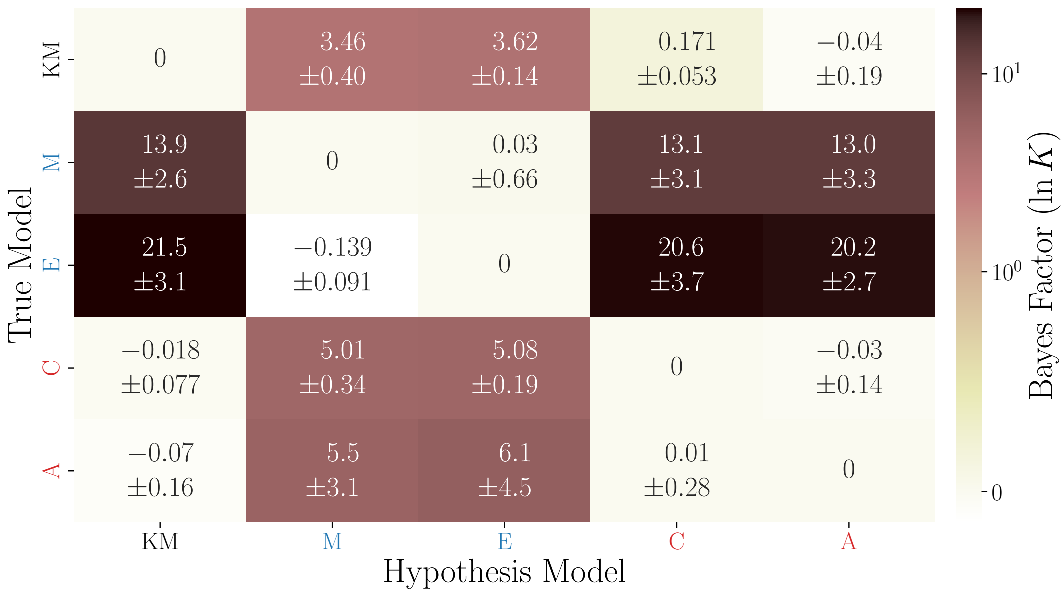



**Figure 4.6:** Same as Figure 4.4 using only transverse momentum $|p_T|$ kinematics.

### 4.2.3 Bias Estimation under Model Misspecification

A bias estimation was performed, showing the difference between the median DNS-predicted parameters ($\hat{m}_{A'}$, $\hat{g}_f$) and the true values, displayed in Figure 4.7. For each of the corresponding runs that are reported in Figure 4.4, the difference between the predicted parameters and the true values is reported as

$$\begin{aligned} \widetilde{m} &\equiv \log_{10}(\hat{m}_{A'}/\mathrm{GeV}) - \log_{10}(m_{A'}^{\mathrm{true}}/\mathrm{GeV}), \\ \widetilde{g}_f &\equiv \log_{10}\hat{g}_f - \log_{10} g_f^{\mathrm{true}}. \end{aligned} \tag{4.3}$$

Under this definition, e.g. $\widetilde{g}_f \approx -2$ means that the sampler has fitted a coupling strength two orders of magnitude lower than the true $g_f^{\mathrm{true}}$ to the data.

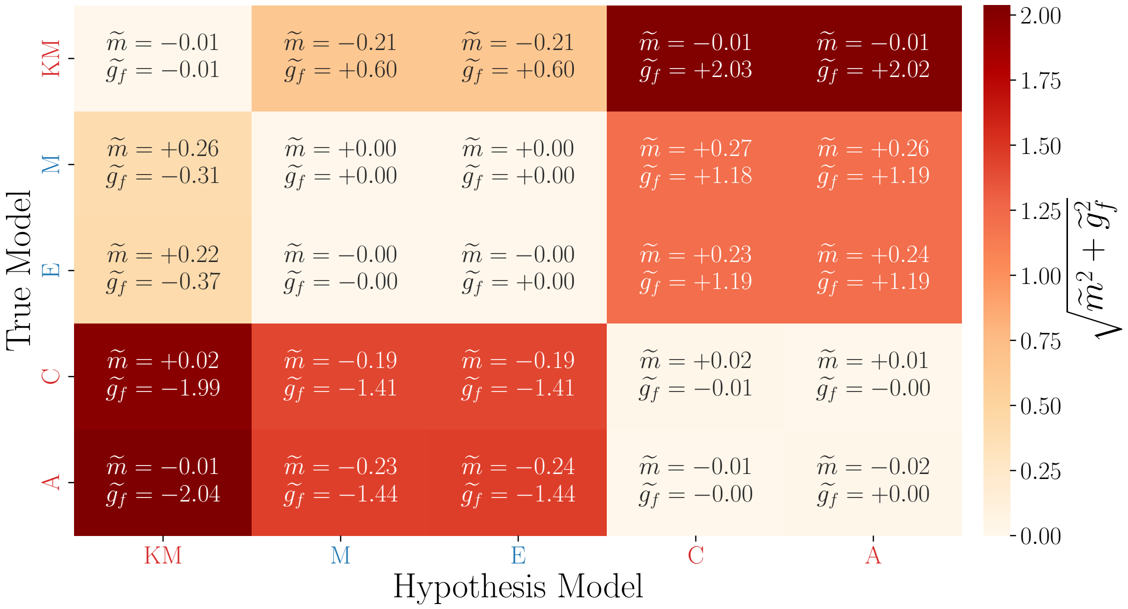



**Figure 4.7:** Bias estimation under model misspecification for the two-dimensional Bayes factor analysis. The y-axis shows what model was used in the generation of the data, while the x-axis shows what model was used in the likelihood while sampling.

As shown in Figure 4.7, the model agreement within groups that was apparent in Figure 4.4 between KM and [C, A] can be explained by the sampler fitting a coupling strength that is two orders of magnitude too high or low to the data. The shape of the distributions between the models is nearly identical, and the only difference is the total event counts $N$. The sampler can account for this difference in $N$ by fitting a different $g_f$ to account for this difference. Similarly, for hypothesis models where the Bayes factor is large, the shape of the kinematic distributions, governed by the mass of the DP, should be different between the true model and the hypothesis model. To match the shape of the hypothesis model, $\hat{m}_{A'}$ gets biased towards a different value.

The apparent asymmetry in the cells that represent cross-group predictions can be explained by the mismatch between the two models' kinematic shapes. When the sampler looks at data generated with e.g. the KM model and compares it to M-signal, the sampler cannot recreate the correct shape of the distribution at the true mass; median prediction $\hat{m}_{A'}$ gets shifted $\approx -0.21$ orders of magnitude away from the truth. In the reverse case, the $\hat{m}_{A'}$ gets biased by $\approx +0.26$ orders of magnitude. The reason $g_f$ gets biased asymmetrically ($\widetilde{g}_f \approx +0.60$ in the first scenario and $\widetilde{g}_f \approx -0.31$ in the second), is that the DP mass also affects the cross-section, and therefore the total event counts. Indeed we can see from the posteriors in section 4.1 that $m_{A'}$ and $g_f$ are correlated. The sampler then tries to account for this shift by fitting a different coupling strength, scaling the signal by the amount that matches the mass. Since the mass bias differs in magnitude between the two directions, so does the required coupling compensation. Figure 4.7 therefore shows the consequences for parameter prediction one could expect when using this Bayesian framework with the wrong model likelihood for the analysis of real data from LDMX.

## 4.3 Relic Target Prior

As a final proof-of-concept, a prior on the parameter $g_f$ is introduced based on the relic target curves in Figure 1 of [9]. Abandoning the log-uniform prior on $g_f$, every sampled point $m_{A'}^{\text{prop}}$ is assigned to a log-normal centered at the corresponding relic target $g_f$, with shape parameter $\sigma_g = 0.1$. This new prior is model-dependent, and introducing it has the effect of constraining $g_f$ to a prior belief based on an assumed $\mathcal{R} = 3$ and $\alpha_D = 0.5$, differing from the $\mathcal{R} = 2.5$ case of Figure 4.4. This leads to the different model priors no longer canceling in equation (3.25), which makes separation possible, and the Bayesian evidence is weakened if $\hat{g}_f$ strays too far from the relic target. The larger $\mathcal{R}$ brings with it slightly altered coupling strengths for all the models with $m_{A'} = 0.1$ GeV. The hope was that the bias reported in Figure 4.7 would be exploited in a way that allows for separation between the KM interaction model and [C, A]. The result is reported in Figure 4.8. We get a clear separation between KM and [C, A], where the relic target prior has penalized the fitting of a $g_f$ away from the relic target. Within each higher order group the separation remains weak, since the paired models' relic target curves are nearly identical at $m_{A'} = 0.1$ GeV: [C, A] shows essentially no separation, while [M, E] separates only marginally.

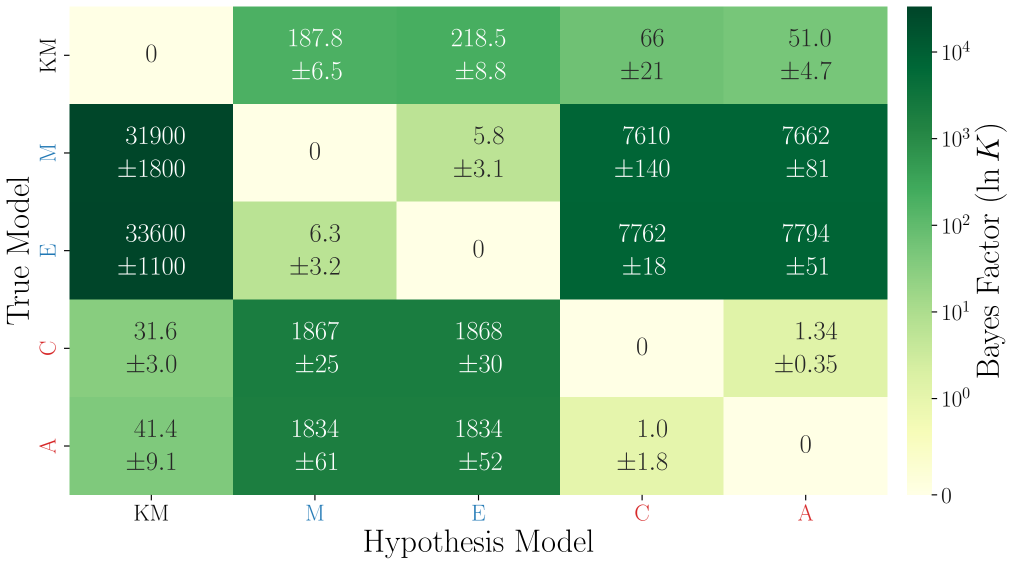



**Figure 4.8:** Same as Figure 4.4 but using a relic target prior on $g_f$ with $\mathcal{R} = 3$, allowing for separation between KM and [C, A].

# 5. Discussion & Conclusion

## 5.1 Discussion

The Bayesian framework recovers $(m_{A'}, g_f)$ within the 68% CI for the $\mathcal{R} = 2.5$ benchmarks (BM1, BM2) under BG = 1 for 7/10 Poisson realizations. At BG = 100 this still holds for BM2 (8/10), but BM1's coupling drops to 2/10. DNS and MCMC give the same number, so this is not a sampler problem. It likely comes from the large background, which pulls the correlated recovered mass and coupling slightly low and tightens their intervals, leaving the true coupling at the upper edge where it is often missed. Even so, the true $g_f$ still falls within the median of the 16th and 84th percentile as shown in Table 4.2, so the coupling is still recovered on average.

The benchmarks corresponding to $\mathcal{R} = 2.2$ behave differently. They sit below the detection threshold for the luminosity of LDMX Phase II, indicated by the prior-dominated marginal posteriors for BG = 1 and the near uninformative $m_{A'}$ posteriors for BG = 100. However, a posterior mode around the truth is still recovered for marginal posteriors over $g_f$, suggesting the likelihood provides more information about the overall scaling of the event counts than the shape of the distribution at these benchmarks. This mode sits on a broad, prior-dominated hill spanning several decades in $g_f$, resulting in a failure for the median and 68% CI to localize it. This is why, for BM3, the true $g_f$ only falls within the 68% CI for 3/10 Poisson realizations (see Table 4.4), and why this benchmark reads as being a non-detection.

The reason a mode survives at all comes down to the background. The marginal posteriors over the nuisance parameters are for all benchmarks centered around the true values with CIs in the expected ranges given their priors. For benchmarks corresponding to $\mathcal{R} = 2.2$, the total number of expected events is very low (dropping by $\approx 10^4$ between BM1 and BM4), so the signal is disrupted by a large background. The mode around $g_f^{\text{true}}$ is probably due to $\theta_B$ having low variance, allowing for a faint distinction of the signal. Indeed, we can see that the marginal posterior over $\sigma_B$ is bounded closer to zero for [BM3, BM4], BG = 100, indicating that the sampler was handed more background information to constrain $\theta_B$ tightly around the true value of $\theta_B^{\text{true}} = 1.0$, to separate the signal from the background.

## 5.2 Future Considerations

A few directions could make the analysis more robust or extend its capabilities. Due to computational limits, the number of Poisson realizations run as an attempt to mitigate Poisson noise has been kept conservative. Five Poisson realizations per Bayes factor lead to noisy MADs. Increasing the number of Poisson realizations would be advantageous to gain a more robust analysis with less noise. Section 4.2.1 shows the calculation of the expected noise floor between `MadGraph5` runs, which could contribute to skewed Bayes factors. This estimation is rough, with only one KM simulation run as a comparison. A more thorough analysis of the noise between `MadGraph5` runs is needed to investigate the effect it has on the Bayes factors.

A second limitation is the binning of the kinematics. An unbinned analysis of the data supplied at LDMX could be advantageous due to the high precision it would provide in capturing the recoil electron kinematics. Right now, the binned analysis inherently strips away some information from the data. While computationally efficient, the binned analysis partially loses detailed shape differences in the $(E, |p_T|)$-distribution which are important in both model comparison and parameter inference. However, one must also consider that while the data might show a precise point in $(E, |p_T|)$ space for a recoil electron, such a data point is still not without its uncertainty. Various errors can affect the data not only in the LDMX detectors, but also in the simulation of the signal. A higher resolution of 200x200 does show slightly narrower 68% CIs (Appendix B), indicating that higher resolution does reveal more information in the data.

A third consideration is in regard to the choice of prior for the coupling strength during model comparison. The benchmarks tested in this work fall far from the edges of the log-uniform $g_f \in [10^{-20}, 1]$ prior. If, however, the true $g_f$ is located close to one of these edges, the prior dependence in equation (3.25) could in theory affect the Bayes factors. In addition to this, the higher-order couplings could in theory take on values larger than 1. It is therefore recommended to increase the scope of the prior on $g_f$ for analysis of real LDMX data.

Finally, the relic target prior could be made less dependent on assumptions. The Bayes factors presented in section 4.3 show that the method of introducing a prior on $g_f$ based on the individual relic targets of the models works for separating KM from [C, A]. The downside to this approach is that one has to assume many things about the DP in order to fix a relic curve for each model. Solving this problem could include sampling both $\mathcal{R}$ and $\alpha_D$ as hyperparameters on the $g_f$ prior. This would need to incorporate some sort of Boltzmann solver into the framework, which can calculate relic target curves based on the hyperparameters. This is left for future work.

## 5.3 Conclusion

In conclusion, the framework performs well on benchmarks corresponding to $\mathcal{R} = 2.5$, while identifying $\mathcal{R} = 2.2$ as below detection threshold at LDMX. A larger background scaling has a negative effect on detection capabilities, leading to large Poisson fluctuations. While the $E < 3.16$ GeV cut removes a large portion of the background, it also removes information about the signal, weakening the detection capabilities on benchmarks where it is already low. The apparent degeneracy between the groups outlined in [9] is broken for two-dimensional kinematics, as the Bayesian framework handles separation between the interaction groups [KM, C, A] and [M, E]. Analysis using only $E$ or $|p_T|$ information still holds some discriminating power between groups, with some Bayes factors below the estimated two-dimensional noise floor. Combined, they produce essential shape differences that result in decisive Bayes factors. Separating KM from C and A within a group requires the model-dependent relic prior on $g_f$, which trades the shared prior cancellation for added assumptions.

# References


[1] N. Aghanim et al. "Planck 2018 results: VI. Cosmological parameters". In: *Astronomy & Astrophysics* 641 (2020), A6. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201833910. URL: http://dx.doi.org/10.1051/0004-6361/201833910.

[2] D. S. Akerib et al. *Searches for Light Dark Matter and Evidence of Coherent Elastic Neutrino-Nucleus Scattering of Solar Neutrinos with the LUX-ZEPLIN (LZ) Experiment*. 2025. arXiv: 2512.08065 [hep-ex]. URL: https://arxiv.org/abs/2512.08065.

[3] Torsten Åkesson et al. *LDMX – The Light Dark Matter eXperiment*. 2025. arXiv: 2508.11833 [hep-ex]. URL: https://arxiv.org/abs/2508.11833.

[4] Torsten Åkesson et al. "Sensitivity of an early dark matter search using the electromagnetic calorimeter as a target for the Light Dark Matter eXperiment". In: *Journal of High Energy Physics* 2025.12 (Dec. 2025). ISSN: 1029-8479. DOI: 10.1007/jhep12(2025)150. URL: http://dx.doi.org/10.1007/JHEP12(2025)150.

[5] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[6] Asher Berlin et al. "Dark matter, millicharges, axion and scalar particles, gauge bosons, and other new physics with LDMX". In: *Physical Review D* 99.7 (Apr. 2019). ISSN: 2470-0029. DOI: 10.1103/physrevd.99.075001. URL: http://dx.doi.org/10.1103/PhysRevD.99.075001.

[7] Gianfranco Bertone and Dan Hooper. "A History of Dark Matter". In: (2016). DOI: 10.48550/ARXIV.1605.04909. URL: https://arxiv.org/abs/1605.04909.

[8] Gianfranco Bertone, Dan Hooper, and Joseph Silk. "Particle dark matter: evidence, candidates and constraints". In: *Physics Reports* 405.5-6 (Jan. 2005), pp. 279–390. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2004.08.031. URL: http://dx.doi.org/10.1016/j.physrep.2004.08.031.

[9] Riccardo Catena, Taylor R. Gray, and Thomas Jerkvall. *Production of Dark Photons through Higher Electromagnetic Moments at LDMX: Simulations and Model Discrimination*. 2025. DOI: 10.48550/ARXIV.2502.13635. URL: https://arxiv.org/abs/2502.13635.

[10] Douglas Clowe et al. "A Direct Empirical Proof of the Existence of Dark Matter". In: *The Astrophysical Journal* 648.2 (Aug. 2006), pp. L109–L113. ISSN: 1538-4357. DOI: 10.1086/508162. URL: http://dx.doi.org/10.1086/508162.

[11] Marco Fabbrichesi, Emidio Gabrielli, and Gaia Lanfranchi. "The Dark Photon". In: (2020). DOI: 10.48550/ARXIV.2005.01515. URL: https://arxiv.org/abs/2005.01515.

[12] Daniel Foreman-Mackey et al. "emcee: The MCMC Hammer". In: *Publications of the Astronomical Society of the Pacific* 125.925 (Mar. 2013), pp. 306–312. ISSN: 1538-3873. DOI: 10.1086/670067. URL: http://dx.doi.org/10.1086/670067.

[13] Andrew Gelman. “Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)”. In: *Bayesian Analysis* 1.3 (2006). ISSN: 1936-0975. DOI: `10.1214/06-ba117a`. URL: `http://dx.doi.org/10.1214/06-BA117A`.

[14] Andrew Gelman et al. *Bayesian Data Analysis*. Chapman and Hall/CRC, Nov. 2013. ISBN: 9780429113079. DOI: `10.1201/b16018`. URL: `http://dx.doi.org/10.1201/b16018`.

[15] Jonathan Goodman and Jonathan Weare. “Ensemble samplers with affine invariance”. In: *Communications in Applied Mathematics and Computational Science* 5.1 (Jan. 2010), pp. 65–80. ISSN: 1559-3940. DOI: `10.2140/camcos.2010.5.65`. URL: `http://dx.doi.org/10.2140/camcos.2010.5.65`.

[16] W. K. Hastings. “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: `10.1093/biomet/57.1.97`.

[17] Edward Higson et al. “Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation”. In: *Statistics and Computing* 29.5 (Dec. 2018), pp. 891–913. ISSN: 1573-1375. DOI: `10.1007/s11222-018-9844-0`. URL: `http://dx.doi.org/10.1007/s11222-018-9844-0`.

[18] Robert E. Kass and Adrian E. Raftery. “Bayes Factors”. In: *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795. DOI: `10.1080/01621459.1995.10476572`. eprint: `https://doi.org/10.1080/01621459.1995.10476572`. URL: `https://doi.org/10.1080/01621459.1995.10476572`.

[19] Gordan Krnjaic. *Testing Thermal-Relic Dark Matter with a Dark Photon Mediator*. 2025. arXiv: `2505.04626 [hep-ph]`. URL: `https://arxiv.org/abs/2505.04626`.

[20] Benjamin W. Lee and Steven Weinberg. “Cosmological Lower Bound on Heavy-Neutrino Masses”. In: *Phys. Rev. Lett.* 39 (4 July 1977), pp. 165–168. DOI: `10.1103/PhysRevLett.39.165`. URL: `https://link.aps.org/doi/10.1103/PhysRevLett.39.165`.

[21] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. ISBN: 9780521642989. URL: `https://www.inference.org.uk/itprnn/book.pdf`.

[22] MadGraph5 Collaboration. *MadGraph5_aMC@NLO*. GitHub. 2026. URL: `https://github.com/madgraph5/` (visited on 04/28/2026).

[23] Radford M. Neal. *Slice Sampling*. 2000. arXiv: `physics/0009028 [physics.data-an]`. URL: `https://arxiv.org/abs/physics/0009028`.

[24] Thomas G. Rizzo. “Dark moments for the Standard Model?” In: *Journal of High Energy Physics* 2021.11 (Nov. 2021). ISSN: 1029-8479. DOI: `10.1007/jhep11(2021)035`. URL: `http://dx.doi.org/10.1007/JHEP11(2021)035`.

[25] Peter J. Rousseeuw and Christophe Croux. “Alternatives to the Median Absolute Deviation”. In: *Journal of the American Statistical Association* 88.424 (1993), pp. 1273–1283. DOI: `10.1080/01621459.1993.10476408`.

[26] Vera C. Rubin and Jr. Ford W. Kent. “Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions”. In: *The Astrophysical Journal* 159 (Feb. 1970), p. 379. ISSN: 1538-4357. DOI: `10.1086/150317`. URL: `http://dx.doi.org/10.1086/150317`.

[27] John Skilling. "Nested sampling for general Bayesian computation". In: *Bayesian Analysis* 1.4 (2006), pp. 833–859. DOI: 10.1214/06-BA127.

[28] Joshua S Speagle. "dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences". In: *Monthly Notices of the Royal Astronomical Society* 493.3 (Feb. 2020), pp. 3132–3158. ISSN: 1365-2966. DOI: 10.1093/mnras/staa278. URL: http://dx.doi.org/10.1093/mnras/staa278.

[29] Speagle, Joshua S. and Koposov, Sergey and Sergachev, Ilya and contributors. *Quickstart: Basic Post-Processing.* https://dynesty.readthedocs.io/en/v3.0.0/quickstart.html#basic-post-processing. Accessed: 2026-03-18. 2024.

[30] Volker Springel et al. "Simulations of the formation, evolution and clustering of galaxies and quasars". In: *Nature* 435.7042 (2005), pp. 629–636. ISSN: 1476-4687. DOI: 10.1038/nature03597. URL: http://dx.doi.org/10.1038/nature03597.

[31] Torsten Åkesson et al. "Light Dark Matter eXperiment (LDMX)". In: (Aug. 2018). arXiv: 1808.05219 [hep-ex]. URL: https://arxiv.org/abs/1808.05219.

[32] Roberto Trotta. "Bayes in the sky: Bayesian inference and model selection in cosmology". In: *Contemporary Physics* 49.2 (Mar. 2008), pp. 71–104. ISSN: 1366-5812. DOI: 10.1080/00107510802066753. URL: http://dx.doi.org/10.1080/00107510802066753.

[33] Yung-Su Tsai. "Pair Production and Bremsstrahlung of Charged Leptons". In: *Rev. Mod. Phys.* 46 (1974), pp. 815–851. DOI: 10.1103/RevModPhys.46.815. Erratum: Rev. Mod. Phys. **49**, 421 (1977).

[34] P A and Zyla et al. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 2020). Chapter 40: Statistics. ISSN: 2050-3911. DOI: 10.1093/ptep/ptaa104. URL: http://dx.doi.org/10.1093/ptep/ptaa104.

# A. Benchmark Distributions

Below are examples of the different distributions created with the four benchmarks that the framework is tested on, and corresponding Poisson realizations.

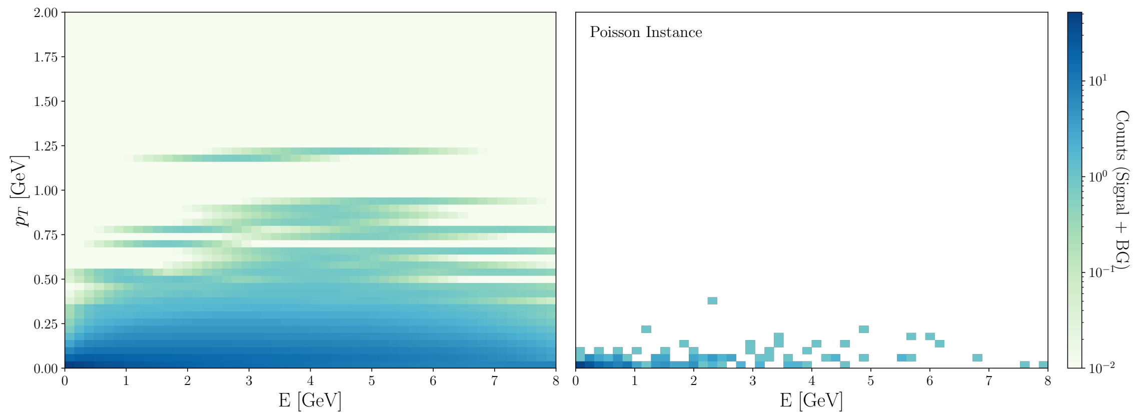



**Figure A.1:** Kinematic distributions of benchmark 1 (right), with a Poisson instance of said distribution (left). Background scaling BG = 1.

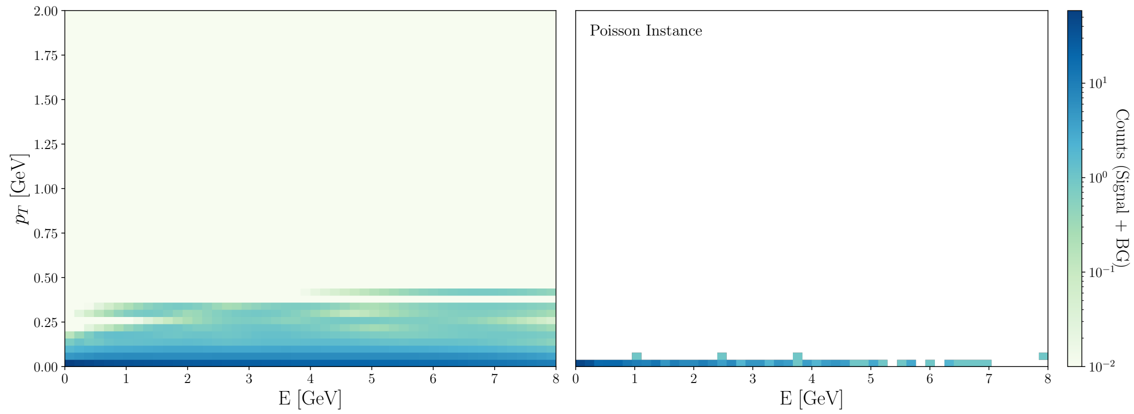



**Figure A.2:** Kinematic distributions of benchmark 2 (right), with a Poisson instance of said distribution (left). Background scaling BG = 1.

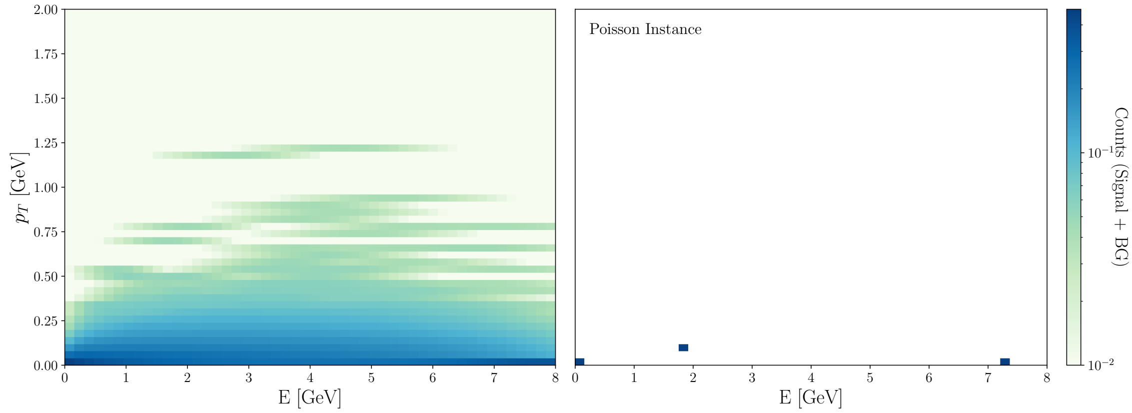



**Figure A.3:** Kinematic distributions of benchmark 3 (right), with a Poisson instance of said distribution (left). Background scaling BG = 1.

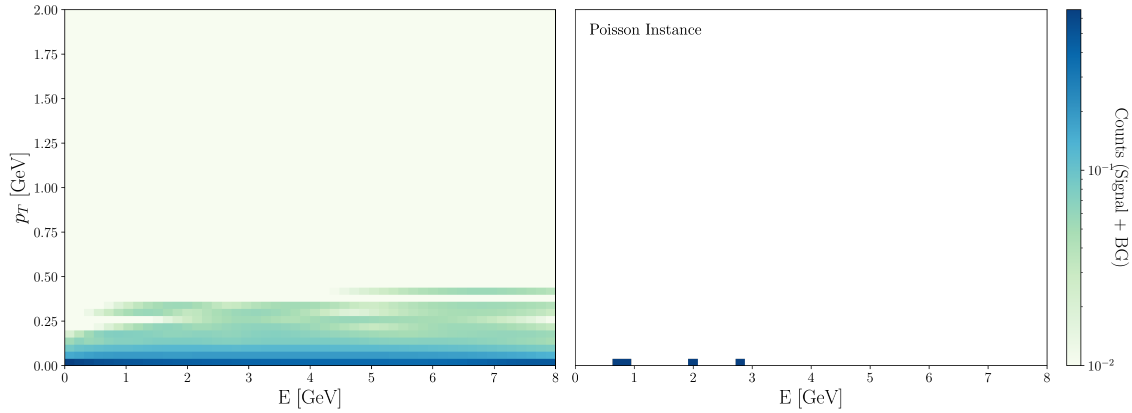



**Figure A.4:** Kinematic distributions of benchmark 4 (right), with a Poisson instance of said distribution (left). Background scaling BG = 1.

# B. Tables

Below, a few tables are listed that display the test of a 200 x 200 bin resolution (Table B.1) as well as the recovered parameters of the MCMC cross-validation (Table B.2).

**Table B.1:** Recovered $(\hat{m}_{A'}, \hat{g}_f)$ posteriors for the four KM benchmarks using DNS at $200 \times 200$ resolution, with background scaling BG = 1. Values are the median across $n = 10$ independent Poisson realizations of the per-realization posterior median, with 68% credible intervals given by the median across realizations of the 16th and 84th percentile bounds. This higher-resolution analysis is presented as a cross-check of the primary $50 \times 50$ results (Table 4.2); recovered values are broadly consistent, and apparent differences for the low-signal benchmarks (BM3, BM4) lie within the realization-to-realization scatter expected at these low event counts and cannot be attributed to binning resolution alone.

| BM | $m_{A'}^{\text{true}}$ [GeV] | $g_f^{\text{true}}$ | $\hat{m}_{A'}$ [GeV] | $\hat{g}_f$ |
|---|---|---|---|---|
| 1 | 0.10 | $6.3 \times 10^{-5}$ | $0.1009^{+0.0076}_{-0.0067}$ | $(6.24^{+0.72}_{-0.63}) \times 10^{-5}$ |
| 2 | 0.01 | $4.8 \times 10^{-6}$ | $0.0100^{+0.0010}_{-0.0010}$ | $(4.80^{+0.56}_{-0.51}) \times 10^{-6}$ |
| 3 | 0.10 | $6.0 \times 10^{-6}$ | $0.0639^{+0.0760}_{-0.0320}$ | $(3.09^{+5.70}_{-2.10}) \times 10^{-6}$ |
| 4 | 0.01 | $5.2 \times 10^{-7}$ | $0.0106^{+0.0300}_{-0.0067}$ | $(4.85^{+8.10}_{-3.60}) \times 10^{-7}$ |

Table B.2 contains the median predictions of $m_{A'}$ and $g_f$ and the median 16th and 84th percentiles for the posteriors.

**Table B.2:** Recovered $(m_{A'}, g_f)$ posteriors for the four KM benchmarks, sampled with MCMC. Values quoted are the median and 68% credible interval over $n = 10$ independent Poisson realizations of each benchmark. Background scaling BG = 1.

| BM | $m_{A'}^{\text{true}}$ [GeV] | $g_f^{\text{true}}$ | $\hat{m}_{A'}$ [GeV] | $\hat{g}_f$ |
|---|---|---|---|---|
| 1 | 0.10 | $6.3 \times 10^{-5}$ | $0.0984^{+0.0089}_{-0.0081}$ | $(6.33^{+0.85}_{-0.78}) \times 10^{-5}$ |
| 2 | 0.01 | $4.8 \times 10^{-6}$ | $0.0114^{+0.0035}_{-0.0025}$ | $(5.46^{+1.50}_{-1.30}) \times 10^{-6}$ |
| 3 | 0.10 | $6.0 \times 10^{-6}$ | $0.0431^{+0.2800}_{-0.0380}$ | $(4.68^{+56.00}_{-3.70}) \times 10^{-7}$ |
| 4 | 0.01 | $5.2 \times 10^{-7}$ | $0.0200^{+0.0970}_{-0.0160}$ | $(4.70^{+25.00}_{-3.90}) \times 10^{-7}$ |

DEPARTMENT OF PHYSICS AND ASTRONOMY
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY