1 Background and Introduction

SuperCDMS is a direct-detection dark matter search and is one of a suite of experiments aiming to answer a top-priority question of the DOE, "What is dark matter?"

Evidence for non-luminous mass, "dark matter" abounds at the galactic and extra-galactic scale [1, 2, 3]. Theories of a dark-matter particle consistent with the observed standard model allow masses ranging from the μeV scale to the TeV scale [4]. Direct-detection dark matter experiments have put stringent limits on weakly-interacting dark matter in the GeV mass range; as the increased sensitivity has resulted in no reproducible detection, interest in the lower mass range has increased [5]. The SuperCDMS project is designed for world-leading sensitivity in the low mass range; this requires excellent data quality as well as novel hardware design.

The SuperCDMS-SNOLAB experiment expects a sensitivity of 10^{-41} cm^2 for 0.5 GeV dark matter for high-voltage silicon detectors and 10^{-40} cm^2 for 0.5 GeV dark matter for high-voltage germanium detectors, based on key performance parameters of detector resolution and noise [6]. However, these projections assume 80% post-analysis livetime. Since the expected dark matter signal increases at low energy, analyzers must push down their analysis as close to the noise wall as possible. This makes the low-threshold scientific results extremely sensitive to a changing noise environment. To achieve 80% livetime, we must support rapid science analysis of incoming data and develop new tools which can allow rapid analysis and alert analysis teams to changes in the noise environment within days. Big-data techniques—such as Generalized Adversarial Networks for creating salting data and change detection of time-series data—are ideally suited for such problems.

We request support for developing a method to automatically generate "salt" for blinded dark-matter data analysis. PI Roberts (physics) and Co-PI Banaei-Kashani (computer science) propose to develop new analysis methods that will support improved data quality.

2 Project Objectives

The objective of the proposed work is to ensure data quality sufficient to deliver the SuperCDMS-SNOLAB science goals listed in Table 1. To achieve this, funding for the following is requested:

1. Development of an efficient salting method with Generative Adversarial Networks (GANs) to allow analysis of full datasets — increasing the likelihood of identifying low-rate data-quality issues
2. Test the effectiveness of this salt for measuring bias in a low-threshold dark-matter analysis.

This work will support the SuperCDMS science goals, particularly low-mass dark matter analyses and modulation analyses.

Previous experience has taught the collaboration that science analysis is what reveals data-quality issues that affect the science reach of our data. PI Roberts has therefore focused on building the infrastructure needed for rapid science analysis: analyzers must be able to perform first-pass analyses within two weeks. This represents an incredible speed-up relative to the prior SuperCDMS experiment, where first-pass analyses often took between two to six months and sometimes even longer.

To achieve this goal, PI Roberts has led efforts to (1) build a maintainable end-user analysis environment with minimal onboarding time, (2) create data-analysis tutorials showing best-practice use of the system, (3) develop critical data catalog features that allow easy access to analysis datasets, (4) develop analysis libraries and methods that use reasonable amounts of computer memory, (5) improve end-user documentation and reporting across systems, and (6) pioneer the use of "containers," a technology that reduces the installation burden for developing software and allows the creation of test suites for core software.

However, rapid analysis will occur on only a tenth of the data — data division is the current plan for blinding analysis because it is rapid. From a data quality perspective, this makes it harder to identify low-frequency data quality issues. These data quality issues have a direct impact on the science reach of low-threshold searches and annual modulation searches; a blinding method that is rapid and allows analyzing the full dataset would be useful to the broader dark matter community.

PI Roberts requests funding for research focused on efficiently creating "salt": simulated signal that is added to a data set to estimate acceptance and rejection bias. Salting is appealing because it allows analyzers to look at the full set of data. However, it is slow. Several dark matter collaborations have explored data salting as a blinding method but it is too slow for a first analysis: SuperCDMS salting took over a year on a well-understood data set. An efficient, accurate salting method would provide a valuable blinding method to the entire dark matter community. PI Roberts has extensive experience with SuperCDMS data and analysis, and for the past year has collaborated with Associate Professor Farnoush Banaei-Kashani, an expert in machine learning who has applied data science methods to solve a wide variety of applied problems. Together they have created a Generative Adversarial Network that creates fake detector signals, a first step towards automated creation of salting data. PI Roberts requests funds to continue this collaborative work with particular application to SuperCDMS data.

3 Proposed Research

Previous experience has taught the collaboration that science analysis is what reveals data-quality issues that affect the science reach of our data. PI Roberts has therefore focused on building the infrastructure needed for rapid science analysis: analyzers must be able to perform first-pass analyses within two weeks. This represents an incredible speed-up relative to the prior SuperCDMS experiment, where first-pass analyses often took between two to six months and sometimes even longer.

Many aspects of the infrastructure needed for rapid science analysis are currently in testing or are nearing completion. However, the plan for rapid blinding is currently data division, which will only allow analyzers to look at a tenth of the data. This will make identification of low-frequency data quality issues difficult.

Therefore, the proposed work focuses on developing a method to rapidly create "salt" for science analysis so that analyzers can look at the full set of data in the initial analysis phase. We propose two parallel efforts to realize this goal:

- Train Generative Adversarial Networks (GANs) to create salt that obeys analysis-specific physics constraints.
- Re-analyze an existing dataset with this salt added to asses its effectiveness for blinded data analysis

The proposed research has the potential to create a new analysis method for the dark matter community that supports timely data quality monitoring. The SuperCDMS experiment has a data quality monitoring system in place and based on previous experience we expect this system to be highly effective at identifying hardware failures, network failures, and dramatic changes in the system. This system is not sufficient to realize the full science impact of the SuperCDMS-SNOLAB experiment as detailed below, making the proposed research high-priority for the dark matter field.

3.1 Intellectual Merit: SuperCDMS-SNOLAB Science Goals

The science goals for SuperCDMS-SNOLAB Operations are listed in Table 1. Each of these science goals is a search for a rare process and—within the timeframe defined by SuperCDMS-SNOLAB operations—is limited either by background ("background limited") or by exposure time ("exposure limited"). Ensuring good livetime is most important for exposure-limited searches. Our flagship low-mass dark matter search (SG-1) will be exposure limited for approximately two years and its ultimate sensitivity to low-mass dark

matter may be set by the noise level we are able to maintain over that time. Modulation searches (SG-6, SG-8) have their own data quality needs: experiment stability. Any periodic or low-frequency changes in the experiment weaken the sensitivity of modulation analyses.

Table 1: Science goals for SuperCDMS SNOLAB.

Primary Science Goals
SG-1 Search for DM with masses $< 10\,\mathrm{GeV/c^2}$ using complementary targets (Ge and Si) and complementary techniques (iZIP and HV) to understand residual backgrounds
SG-2 Design for the possibility of future upgrades that would further increase the low-mass sensitivity of the experiment to the level where solar neutrinos are detected
Secondary Science Goals
SG-3 Search for non-SI dark matter interactions within the EFT framework
SG-4 Observe coherent neutrino scattering of ^{8}B solar neutrinos
SG-6 Search for axions produced in the sun and the galaxy
SG-7 Search for lightly ionizing particles (LIPS)
SG-8 Search for annual modulation characteristic of a galactic dark matter "wind"
SG-9 Search for sub-GeV mass dark matter via electron scattering

(SG-1) Low-mass dark matter sensitivity SuperCDMS-SNOLAB aims for a sensitivity of 10^{-41} cm^2 at dark-matter masses of 0.5 GeV with its high-voltage silicon detectors. Because dark matter is more likely to deposit small amounts of energy in a detector, analysis in the low-mass dark matter region pushes as close to the noise wall as possible. Identifying any increases in the environmental noise promptly is therefore critical to allow for mitigation.

(SG-6, SG-8) Modulation searches Axions (SG-6) are expected to produce a daily modulation signal and an annual modulation signal (SG-8) is an expected feature of the current cosmological model of dark matter [7, 8]. DAMA has observed a possible annual modulation signal [9, 10] that other experiments have failed to reproduce [11, 12, 13]. The SuperCDMS-SNOLAB installation will be one of a set of the lowest-background dark matter experiments ever built and will have the potential to produce a strong annual modulation analysis.

Due to the fact there are many possible sources of variation, annual modulation analysis places stringent requirements on the data quality system. The SuperCDMS-SNOLAB change to a selective-detector readout, while necessary for data size considerations, may increase uncertainty in the trigger efficiency. Any modulation affects a modulation analysis, and modulations in-phase with expected signal will directly weaken the analysis sensitivity.

3.2 Challenges

Ensuring high data quality at SuperCDMS-SNOLAB will require shortening the time between taking the data and getting full analysis feedback on the science reach of that data within weeks rather than months.

Most of the solution for rapid science analysis is supported through NSF and DOE grants that support the SuperCDMS-SNOLAB experiment. The proposed effort to automate the creation of salting data for an efficient method of dark-matter analysis blinding is R&D that benefits the SuperCDMS collaboration and

has the potential to benefit the broader dark-matter field. The specific hypotheses posed by this work are listed in Table 2.

3.3 Hypotheses

PI Roberts seeks funding to efficiently create salting data with Generative Adversarial Networks (GANs), with the ultimate goal of allowing analyzers to look at the full data set (plus the "salt") in the blinded phase of analysis. In addition to providing a practical method of blinding to the dark matter community, this will also reduce the time needed to identify low-frequency data quality issues.

Table 2: Hypotheses of the proposed work.

Primary Hypothesis
H-1 Can Generative Adversarial Networks create effective "salt" for blinded dark-matter analysis?

3.4 H-1: Using Generative Adversarial Networks for creating fake data

PI Roberts has collaborated with Co-PI Farnoush Banaei-Kashani for the past year to generate "salt" data using Generative Adversarial Networks (GANs).

"Salting" adds simulated signal to a data set. This allows analyzers to look at the full data set during a blinded analysis, making it ideal for dark matter analysis where full investigation of the data is often necessary to fully understand the detector response. However, creating realistic simulated signal currently takes years. GANs specialize in learning to generate new data that is indistinguishable from the training set and have the potential to create realistic simulated signal within weeks. Using GANs for data salting could provide a powerful and much-needed analysis tool to SuperCDMS and the broader dark matter community.

A GAN consists of a Generator and a Discriminator. The Generator generates fake samples of data (in our case sensor data from cryogenic crystals) and tries to fool the Discriminator. The Discriminator, on the other hand, tries to distinguish between the real and fake samples. This competition repeats many times when the GAN is being "trained." With each repetition the Generator and Discriminator get better and better in their respective jobs: generating fake data samples (the Generator) and discriminating fake data samples from real data samples (the Discriminator). As a result, we train the Generator to maximize the probability of the Discriminator in making a mistake. The Discriminator, on the other hand, is based on a model that estimates the probability that the sample that it got is received from the real data set and not from the Generator. Once training is complete, the Generator is isolated from the rest of the network and used for efficient and accurate fake data generation, in our case the "salt".

We have adapted TimeGAN [14], a specialized type of GAN, to generate the time-series data that SuperCDMS-SNOLAB sensors record. Figure 1 shows example input data — real data recorded in the SuperCDMS Soudan experiment — and the generated data.

Two tasks are required before these results will be ready for testing in a prior analysis: (1) While our results look promising by eye, we need to check that the noise characteristics of the generated salt are indistinguishable from real data on SuperCDMS data-quality plots. (2) Our current method of fake-data generation does not include physics constraints like the expected energy spectrum due to a dark-matter interaction.

Developing SuperCDMS-specific quality measures of generated data is straightforward. This proposal requests funding for a physics-focused Master's student for this work because we will use previous physics analyses as a guide.

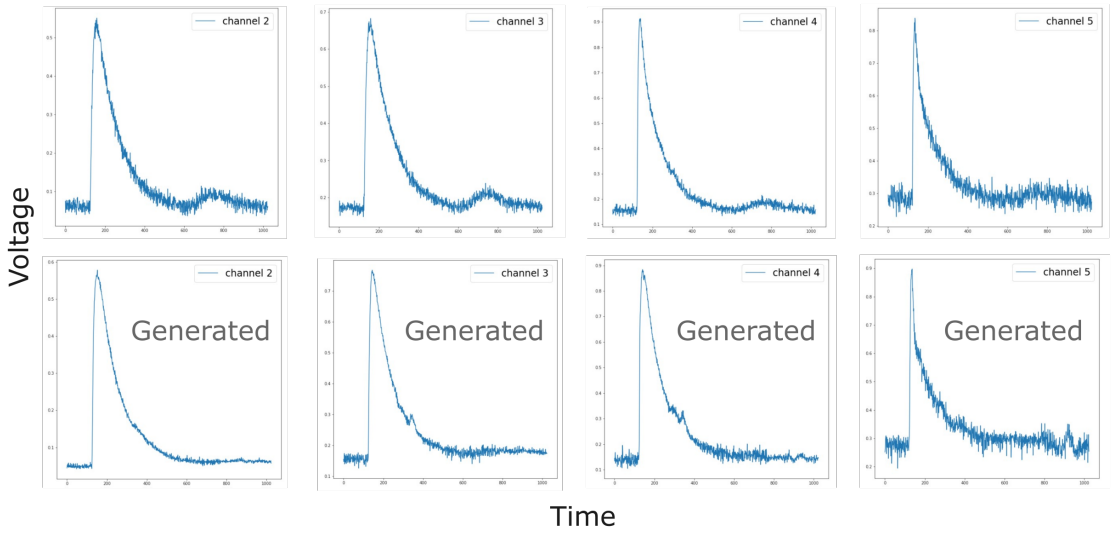

Figure 1: Input data (top) and data generated by TimeGAN (bottom). The plots and TimeGAN training are Selvakumar Jayaraman's work, under the guidance of PI Roberts and Farnoush Banaei-Kashani. The generated data looks promising by eye; SuperCDMS-specific quality measures are needed, as well as physical constraints on the size of the signal.

To impose physics constraints on the generated data, we intend to exploit physics-guided machine learning methodologies [15]. Because physics-based GAN models benefit from both the effectiveness of GAN and physics knowledge, this solution is expected to outperform our existing model in terms of accuracy and efficiency.

3.4.1 H-1: Novel aspects of salting creation: Creating fake, time-series data with constraints has not been used by the dark matter field as a salt-generation method.