1 Background and Introduction

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. This work will support the SuperCDMS science goals, particularly low-mass dark matter analyses and modulation analyses, both of which are particularly sensitive to any change in the noise environment.

The SuperCDMS-SNOLAB experiment expects a sensitivity of 10^{-41} cm^2 using 0.5 GeV dark matter for high-voltage silicon detectors and 10^{-40} cm^2 using 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 an energy-independent efficiency of 80% after both livetime and analysis cuts. This is an aggressive data-quality goal; total efficiency of SuperCDMS-Soudan analyses is typically 50% and decreases strongly at low energies. Avoiding this decrease in efficiency at low energies is critical to maximizing science reach for dark matter searches: since the expected dark matter signal at all masses increases at low energy, analyzers must push down their analysis as close to the noise wall as possible. And if the efficiency is too low at lower energies, experiments can lose any sensitivity to lower-mass candidates. This makes the low-threshold scientific results extremely sensitive to a changing noise environment. To achieve 80% efficiency, we must support rapid science analysis of incoming data . However, we also need to blind the data to produce science results that we believe in. With current methods, data quality needs and trusting results are in tension: the need to look at data within days means that any blinding method must be nearly automatic. Data division, where only a portion of the data is inspected during analysis, is by far the fastest method and the one SuperCDMS has chosen. This method meets the speed requirement but also limits the set of data analyzers can look at, making it difficult to identify intermittent data issues. Salting, a method that injects fake signal events into a datastream, allows analyzers to inspect the full data set—but our current methods for generating “salt” requires full data analysis first and is therefore not able to allow full inspection during data quality analysis.

Machine-learning techniques—such as Generalized Adversarial Networks for creating salting data —are ideally suited for rapidly generating fake signals. Developing an ML-based method for data blinding has the potential to allow analyzers to inspect the full set of data while still performing blinded analysis, supporting the strongest data quality efforts while still preserving trust in our science results.

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 rapid inspection of the full dataset while preserving blinding. PI Roberts is the appointed data-quality coordinator for SuperCDMS operations and is the elected Software Working Group Chair. She has led the collaboration’s effort to reduce barriers to science analysis; this has been critical to her current collaboration with Co-PI Banaei-Kashani. Co-PI Banaei-Kashani specializes in solving domain-specific problems with Big Data methods and has worked with PI Roberts for over a year on prototype software, gaining familiarity with the data and analysis methods of the dark matter field. We request support for two students, one with ML experience and another with physics experience, to verify and further develop this ML-supported analysis method.

2 Project Objectives

The objective of the proposed work is (1) to develop a method that reliably and rapidly creates “salt” for blinded dark-matter analyses and (2) to ensure data quality sufficient to deliver the SuperCDMS-SNOLAB science goals described in Section **??**. To achieve this, funding for the following is requested:

1. Develop an efficient salting method with Generative Adversarial Networks (GANs) to allow analysis of full datasets
   (a) develop experiment-specific tests for the similarity of generated data to experimental data
   (b) develop a GAN model that incorporates the physics constraints of expected dark-matter signals
2. Test the effectiveness of this salt for blinding a low-threshold dark-matter analysis.

2.1 Why does this work need to happen now?

Previous experience has taught the collaboration that science analysis reveals data-quality issues that affect the science reach of our data. The SuperCDMS-SNOLAB experiment is currently testing sets of detectors and the first science data run is anticipated in early 2023. Data quality will soon become the primary determinant of the science reach of this incredible, multi-physics facility.

Part of what’s needed for data quality is in place - automated monitoring tools and an infrastructure that allows analyzers to perform first-pass analysis within two weeks of recording data. This prototype for this infrastructure is in place and is being stress-tested and improved during detector testing.

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. The proposed work provides a method for looking at the full dataset during first-pass analysis.

3 Proposed Research

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 SuperCDMS-SNOLAB experiment is a “direct-detection dark matter search”. It is a highly radiopure experiment with energy sensitivity down to the eV scale and is optimized for low-mass dark matter candidates. Because we know so little about dark matter, there are many science goals for the experiment, each looking for a different dark matter signature.

In general, each science goal is limited either by background (“background limited”) or by exposure time (“exposure limited”). Below we discuss two science goals of the SuperCDMS-SNOLAB experiment and explain why the proposed work is essential for maximizing their sensitivity.

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. This is the flagship SuperCDMS-SNOLAB analysis and it will be exposure limited for approximately two years; ensuring good efficiency is most important for exposure-limited searches.

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. Our ultimate sensitivity to low-mass dark matter may be set by the noise level we are able to maintain over that time.

Identifying any increases in the environmental noise promptly is therefore critical to allow for mitigation.

Modulation searches Modulation searches have their own data quality needs: experiment stability. Any periodic or low-frequency changes in the experiment weaken the sensitivity of modulation analyses.

Axions are expected to produce a daily modulation signal and an annual modulation signal 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 Work Plan

Table 1: Task breakdown of the proposed work.

PI Roberts, MIS student	Co-PI Banaei-Kashani, CS student	End Date
Verify GAN using domain-specific measures	Introduce physics constraints into GAN model	Feb. 2024
Milestone: progress report at DANCE-ML meeting		Mar. 2024
Reproduce prior salting analysis w/GAN	Update GAN model and retrain as needed	Feb. 2025
Salting analysis for SuperCDMS-SNOLAB data	Package code for modularity, reusability; improve GAN model based on feedback	Feb. 2026
Milestone: initial code release and JOSS publication		Feb. 2026
Milestone: virtual workshop at DANCE-ML meeting		Mar. 2026
Provide labeled datasets, test automatic salt generation	Test change-detection algorithms for GAN retraining	Sep. 2026
Milestone: code release		Sep. 2026

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.

Hypotheses
H-1 Can Generative Adversarial Networks create effective "salt" for blinded dark-matter analysis?
H-1A Can Generative Adversarial Networks generate data that is indistinguishable from detector data?
H-1B Can Generative Adversarial Networks generate data that obeys the expected physics of dark-matter signal data?
H-2 Can Generative Adversarial Networks create effective "salt" for blinded dark-matter analysis during an experiment without any analyzer intervention?
H-2A Can Generative Adversarial Networks identify the need to retrain without human intervention?

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

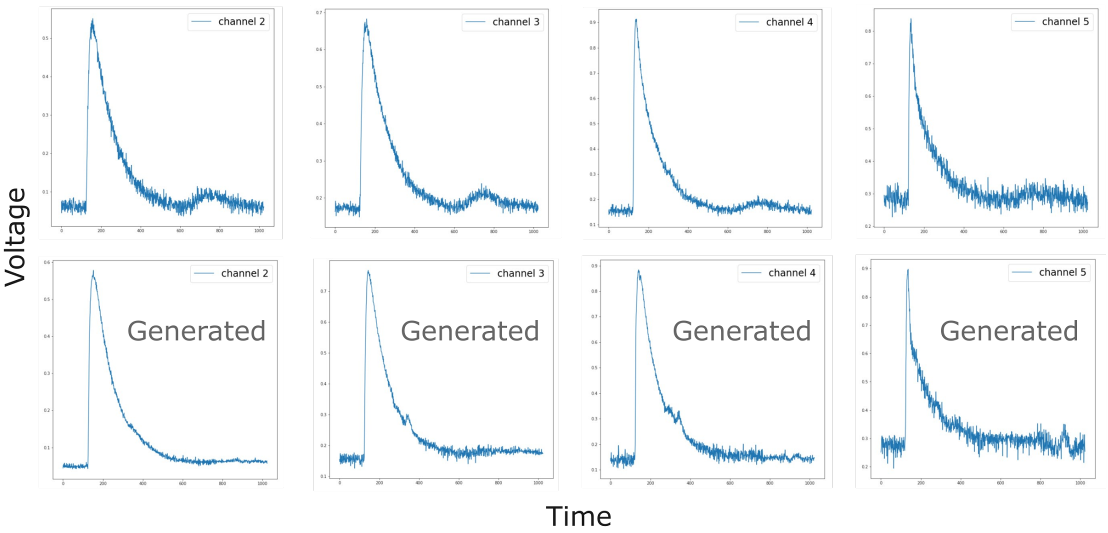

Figure 1: Input data (top) and data generated by TimeGAN (bottom). The plots and TimeGAN training are Selvakumar Jayaraman's work, under the guidance of PI Roberts and Farnoush Banaei-Kashani. The generated data looks promising by eye; SuperCDMS-specific quality measures are needed, as well as physical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

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 because we can use previous physics analyses as a guide. This proposal requests funding for a physics-focused Master's student for this work.

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.

To impose physics constraints on the generated data, we intend to exploit physics-guided machine learning methodologies [15]. What we've shown is a proof of principle. Because physics-based GAN models benefit from both the effectiveness of GAN and physics knowledge, this updated solution (code?) is expected to generate even more accurate fake data, faster.