TIDMAD: Time Series Dataset for Discovering Dark Matter with AI Denoising

J. T. Fry[1*] jtfry@mit.edu
Xinyi Hope Fu[1] hopefu@mit.edu
Zhenghao Fu[1] fuzh@mit.edu
Kaliroe M. W. Pappas[1] kaliroe@mit.edu
Lindley Winslow[1] lwinslow@mit.edu
Aobo Li[2*] aol002@ucsd.edu

[1]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Halıcıoğlu Data Science Institute, Department of Physics, UC San Diego, La Jolla, CA 92093, USA
[*] Corresponding Authors

Abstract

Dark matter makes up approximately 85% of total matter in our universe, yet it has never been directly observed in any laboratory on Earth. The origin of dark matter is one of the most important questions in contemporary physics, and a convincing detection of dark matter would be a Nobel-Prize-level breakthrough in fundamental science. The ABRACADABRA experiment was specifically designed to search for dark matter. Although it has not yet made a discovery, ABRACADABRA has produced several dark matter search results widely endorsed by the physics community. The experiment generates ultra-long time-series data at a rate of 10 million samples per second, where the dark matter signal would manifest itself as a sinusoidal oscillation mode within the ultra-long time series. In this paper, we present the TIDMAD — a comprehensive data release from the ABRACADABRA experiment including three key components: an ultra-long time series dataset divided into training, validation, and science subsets; a carefully-designed denoising score for direct model benchmarking; and a complete analysis framework which produces a physics community-standard dark matter search result suitable for publication as a physics paper. This data release enables core AI algorithms to extract the dark matter signal and produce real physics results thereby advancing fundamental science. The data downloading and associated analysis scripts are available at `https://github.com/jessicafry/TIDMAD`.

1 Introduction

The quest to uncover the nature of dark matter is one of the biggest challenges in contemporary physics. Several key observations in astrophysics and cosmology have confirmed the existence of dark matter, which constitutes approximately 85% of all mass in the universe [2, 3, 26, 31, 32]. However, dark matter has never been detected by any detector on Earth. Because the composition of dark matter is unknown, theoretical physicists propose various dark matter candidates — hypothetical particles that can be characterized by their physical parameters. Experimental physicists then design experiments to search for these candidates. A convincing detection of any dark matter candidate would be a Nobel-Prize-level breakthrough in fundamental science, but even if nothing is detected, the null results still play a significant role in advancing our understanding of physics by setting limits within the physical parameter space. This means that a particular experiment has eliminated the

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

existence of a dark matter candidate within these limits and does not have sufficient sensitivity to test outside these limits. Reciprocally, these limits are used by theorists to propose better dark matter candidates, thereby improving our understanding of this mysterious constituent of our universe.

Attributable to its extremely rare interactions with normal matter, the signal of dark matter is often submerged in a sea of noise from various sources internal and external to the experimental apparatus. Machine learning (ML) offers a promising means to reduce this noise. Advancements in denoising techniques using ML algorithms have the potential to significantly improve dark matter analyses [27]. These techniques enable the detection of weaker dark matter signals, or in the case of no observation, the setting of stronger limits. In other words, improvements in data denoising directly enhance the scientific reach of dark matter experiments. In this paper, we present an ultra-long time series dataset produced by a real dark matter detector: ABRACADABRA (A Broadband/Resonant Approach to Cosmic Axion Detection with an Amplifying B-field Ring Apparatus, abbrev. ABRA-10cm). ABRA-10cm is the world leading sub-μeV dark matter experiment that pioneered the quantum enabled lumped element dark matter detection technique [21, 22, 28]. We operated the ABRA-10cm detector in February 2024 to obtain a special time series dataset for these studies: TIDMAD (TIme series dataset for discovering Dark Matter with Ai Denoising). These data are partitioned into three parts: (1) training data, (2) validation data, and (3) science data.

The training data include time series data where a dark matter-like signal is injected by hardware. If dark matter enters ABRA-10cm, it will manifest itself as a sinusoidal oscillation mode within the time series; therefore, the injected signals are also sinusoidal oscillations within a range of specific frequencies. While the signal shape is known, the dark matter signal amplitude and frequency are unknown parameters in dark matter searches. Both the detected (noisy) time series and the injected (ground truth, clean) time series are provided with one-to-one temporal correspondence. This allows the training of machine learning algorithms to denoise the detector data and recover the injected signal. The validation data is used to produce Benchmark 1: Denoising Score, see Section 4.1. Algorithms that effectively dampen the detector noise while amplifying the injected signal will achieve a better denoising score. The science data is collected without the injected signal with an extended duration to produce Benchmark 2: Dark Matter Limits. The limit generation procedure is detailed in Section 4.2. The scientific data are titled to reflect their use in producing real, community-standard physics results that are suitable for presentation in scientific journals. Several traditional and deep learning denoising algorithms are presented in Section 3 and Appendix C, where the resulting denoised data is benchmarked against the raw, un-denoised detector data.

1.1 Axion dark matter and ABRACADABRA

In recent years, the axion has emerged as one of the leading dark matter candidates as a result of its theoretical elegance. Axions interact with normal matter via electromagnetism, which can be characterized by a physics parameter $g_{a\gamma\gamma}$. Arising from its small mass $m_a < 1eV$ (10^{-6} times smaller than electron), axions act as a classical field oscillating at a frequency $f_a = m_a/2\pi$. Astrophysical measurements determine that the Earth exists in a bath of dark matter with a known local density of ρ_{DM} [12].

The latest advancements in quantum detector technology have facilitated new avenues to search for the axion. ABRA-10cm is one of the novel detectors designed to search for axions leveraging these advancements in quantum technologies [22]. ABRA-10cm capitalizes on the fact that we are immersed in a bath of a feebly electromagnetically-interacting, oscillating dark matter field to detect this elusive particle. Specifically, in the presence of a static magnetic field $\mathbf{B}_0$, the axion, henceforth referred to as dark matter, induces an oscillating magnetic field $\mathbf{B}_a$. Thus, to detect dark matter ABRA-10cm provides a strong magnetic field $B_0 = 1T$ and uses a superconducting pickup loop to observe the oscillating dark matter signal. Read out by a superconducting quantum interference device (SQUID), the pickup loop detects the dark matter signal as a time-oscillating current given by

$$\mathbf{J}_{eff} = g_{a\gamma\gamma}\sqrt{2\rho_{DM}}\mathbf{B}_0 cos(m_a t) \tag{1}$$

where the two parameters that define the theory, the coupling $g_{a\gamma\gamma}$ and mass m_a, appear as the relative strength of the signal and oscillation frequency respectively [21]. The total signal power expected in our detector is given by

$$A \equiv \langle|\Phi_a|^2\rangle = g^2_{a\gamma\gamma}\rho_{DM}\mathcal{G}^2 V^2 B^2_{max} \tag{2}$$

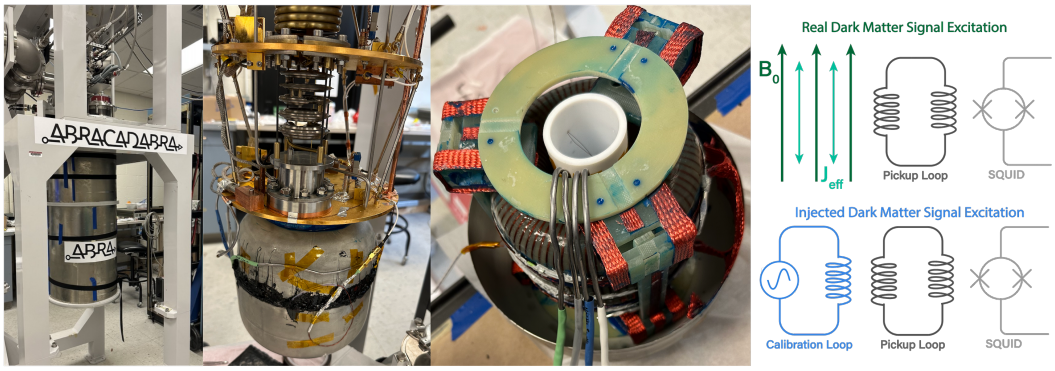

Figure 1: Left to right: ABRA-10cm dilution fridge with outer vacuum cans on; Coldest stage of ABRA-10cm fridge above shielded 1T superconducting torroidal magnet; Interior of ABRA-10cm magnet including pickup and calibration loop wires in the center of the magnet; Effective circuit diagram for both dark matter and injected signals.

where $\mathcal{G} \approx 0.0217$ is a geometric coupling, $V \approx 890\text{cm}^3$ is the magnetic field volume, and $B_{max} \approx 1$ T is the maximum static magnetic field [21].

The frequency of this oscillating signal is a model parameter meaning if we knew the dark matter mass, this frequency would be set. However, theoretical models point to a range of possible dark matter masses, not a singular value; this signal frequency range covers more than seven orders of magnitude making the data denoising task extremely challenging.

1.2 TIDMAD construction

In the classical analysis, we use a calibration procedure to determine the end-to-end response of our system for different signal frequencies. As shown in Figure 1, the ABRA-10cm detector contains a toroidal magnet equipped with both a pickup loop and a calibration loop. During calibration, we first inject a fake dark matter signal into the calibration loop at a specific frequency. This generates a sine wave with a known amplitude and frequency, creating a dark matter-like flux with our pickup loop. Finally, this flux is detected by the SQUID sensor for detector calibration.

The dark matter signal injected into the calibration loop by the signal generator follows the form prescribed by axion theory, as shown in Equation 1 and derived in Appendix A. We specifically choose to inject sine waves with frequencies from 1.1 kHz to 4.9 MHz, corresponding to axion masses $m_a = [0.005, 17]$ neV, to target the mass range that our experimental hardware is designed to detect. The injected signals were all set to an amplitude of 50 mV to ensure a reasonable signal-to-noise ratio. A total of 309 different frequencies were sequentially stepped through, from 1.1 kHz to 4.9 MHz, simulating 309 distinct axion masses in our detector hardware.

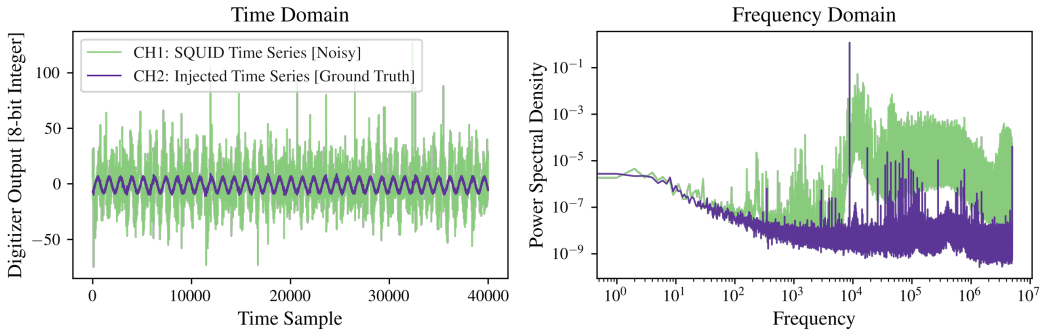

Figure 2: 10-millisecond snapshot of the time series in TIDMAD training dataset compared to the power spectral density of the same data snapshot.

The TIDMAD dataset presented in this work is inspired by this calibration procedure. The ABRA-10cm detector hardware enables us to simultaneously record two types of ultra-long time series: the one injected into the calibration loop (referred to as the "injected time series") and the one detected by the SQUID sensor coupled to the pickup loop (referred to as the "SQUID time series"). As shown in Figure 2, the injected time series exhibit a clear sinusoidal oscillatory signal, which can be considered the ground truth. Meanwhile, the SQUID time series contains the same ground truth submerged within a sea of detector noises. While the ground truth time series is simple, the detector readout contains a complex spectrum of noise spanning a large frequency range, which cannot be modeled using any simple approach [11]. The two time series are exactly aligned at every time step. This defines the signal recovery task: a model could be applied to the SQUID time series to reproduce the injected signal in the injected time series. A model trained on this task will be efficient in rejecting noise of different kinds while retaining the dark-matter-like signal within the detector. We then collected a science dataset where no fake dark matter signal is injected. The trained denoising model can then be applied to the SQUID time series of the science dataset. If a sinusoidal signal is found after denoising, it could potentially be a real dark matter particle entering the detector.

2 Dataset description

The data presented in this paper were acquired using the ABRACADABRA detector. The overall schematics of the data is shown in Figure 3. All data are saved as a series at 10 MS/s (Megasample per second), where each sample is a 8-bit integer ranging from -128 to 127. These integers can be converted into a physics units of mV (millivolts) with a scaling factor of $40/128$. The procured datasets are stored at Open Science Data Federation (OSDF) [6, 18, 37] in `.hdf5` format and can be accessed via the `download_data.py` script in the github repository provided.

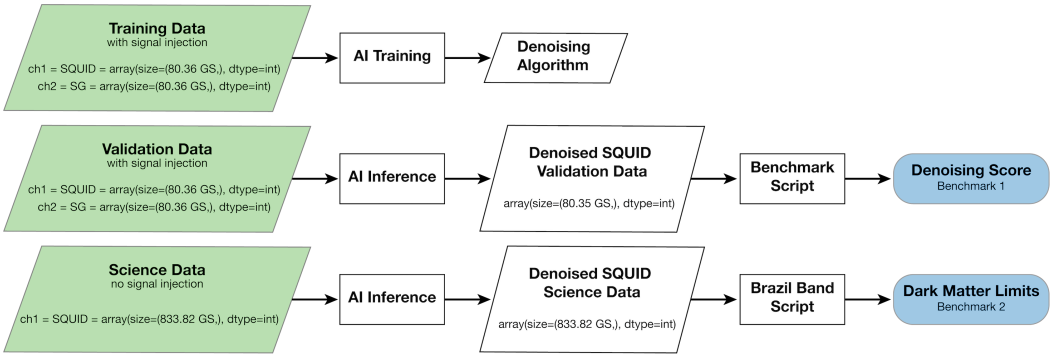

Figure 3: TIDMAD data flow explaining how data splits (green) result in benchmarks (blue). Rectangles correspond to provided scripts. Left to right, top to bottom the scripts are `train.py`, `inference.py`, `benchmark.py`, `process_science_data.py`, `brazilband.py`

Training dataset: The training dataset contains 80.23 Gigasamples of time series data, corresponding to roughly two hours of data collection. For training purposes, we injected a fake dark matter signal into the hardware as discussed in Section 1.2. The injected signal scans through dark matter frequencies from 1100 Hz to 5 MHz at two different amplitudes: 50 mV (standard) and 10 mV (weak). Only the standard injection was considered in the rest of this paper. However, the weak injection data is also available to download by using the additional `-w` flag with the `download_data.py` script, providing a more challenging scenario for signal recovery. All training data are partitioned into 20 files. Each file contains two channels: the injected time series is saved in the second channel (CH2), while the SQUID time series is saved in the first channel (CH1). As discussed in Section 1.2, the training task is to recover the CH2 ground truth time series using the CH1 noisy time series as input.

Validation dataset: The validation dataset, consisting of 80.23 Gigasamples, has the same format as the training dataset. The only difference is that the validation dataset was independently collected at a different time using the same detector apparatus, making it an out-of-sample dataset with slightly altered noise conditions. After training, users can perform model inference by running the `inference.py` script to denoise the CH1 time series while preserving the injected signal. The

denoised SQUID time series in CH1 and the injected time series in CH2 are then processed through the `benchmark.py` script. To determine the efficacy of the model's denoising, a benchmarking score called 'denoising score' is calculated, which will be discussed in detail in Section 4.1.

Science dataset: The science dataset comprises 833.82 Gigasamples of time series data collected over a 24-hour period. This data is distributed across 208 `.hdf5` files. Unlike the training and validation datasets, there are no injected signals, meaning that only the CH1 time series is saved per file. The inference task for ML models is to denoise the SQUID time series in CH1. The denoised science data is then analyzed to obtain a dark matter limit, which will be discussed in Section 4.2.

3 Experiments

We benchmarked eight different denoising algorithms including three traditional algorithms and five deep learning models. The traditional algorithm can be directly applied to the validation dataset using `inference.py`, while deep learning models need to be trained first on the training dataset with `train.py`. The eight algorithms are listed below:

- **Moving average**: a simple moving average with a window size of 19, implemented using the numpy convolve function.
- **Savitzky-Golay filter**: with a window size of 19 and a polynomial order of 11.
- **Fourier Averaging (baseline)**: this operation was originally adopted on ABRA-10cm to compress its data [21, 28]. For every 10-second time series segment, a Fourier transform is applied to each one-second interval, and the resulting 10 Fourier spectra are averaged to produce a single spectrum representing the 10-second segment. This method is considered the baseline approach, representing the current status quo of ABRA-10cm.
- **FC net**: an autoencoder architecture designed for transforming input data. This model consists of an encoder and a decoder. The encoder encodes the input data into a low-dimensional representation, while the decoder reconstructs the original data from this encoded representation. Both the encoder and decoder are composed of multiple fully-connected layers and activation layers. FC-Net outputs a single floating point number at each time step, and the training is conducted by minimizing the mean square error between this floating point number and corresponding ground truth time series at every sample.
- **WaveNet**: a deep neural network originally developed to generates high-quality raw audio waveforms by autoregressively predicting each sample using dilated causal convolutions [33]. The model was adapted with residual blocks containing exponentially increasing dilation rates, gated activations, and skip connections to maintain the same input-output shape for TIDMAD denoising task.
- **PU net**: a deep learning architecture based on the UNet architecture [25]. U-Net uses convolution layers as encoder and deconvolution layers as decoder, with contracting paths established between each pair of convolutional and deconvolution layers at the same level. This allows information at different encoding levels to flow to the decoding part. Positional encoding layers are introduced at all encoder layers to enhance the model's ability to understand positions in the time series. Since every sample of the ground truth time series has to be 8-bit integers ranging from -128 to 127, we require the model to output a 256-class classification decision at every time step, where each class corresponds to one possible output value. This effectively redefines the denoising task into a semantic segmentation task.
- **Transformer**: the transformer utilizes a self-attention mechanism to capture long-distance dependencies in sequences [34]. After processing by the multi-layer Transformer encoder, the model effectively extracts features and represents the input sequence. Finally, the encoded sequence is mapped to the output dimension through a linear layer for the same 256-class classification decision as PU-Net. Positional encoding is also added before the time series is fed into Transformer layers.
- **RNN Sequence to Sequence Model**: a neural architecture with separate encoder and decoder RNNs where the encoder processes the entire input sequence to produce a context representation, and the decoder autoregressively generates the output sequence [33]. Both

encoder and decoder adopts a LSTM architecture, and the decoder output is the same 256-class classification decision as PU-Net. PU-Net, Transformer, WaveNet, and RNN Seq2Seq are all trained using Focal Loss to handle class-imbalanced segmentation labels [17].

The benchmarking results of these models are discussed in Section 4. A hyperparameter search study is conducted on the Moving Average and Savitzky-Golay filter. The best-performing window size and polynomial order are chosen, with the rest of the study detailed in Appendix E. There are two additional constraints for the deep learning models. First, because of memory constraints, we segment the ultra-long SQUID and injected time series into smaller segments before feeding them into each model. The exact segment sizes are outlined in Table 1. Secondly, due to the broad frequency spectrum in our input data, we implemented frequency splitting for all models except WaveNet, training multiple specialized versions of each model to handle distinct frequency ranges. During benchmarking, we observed that a single WaveNet model is efficient to handle all frequencies. Both limitations and additional details of the deep learning models are discussed in Appendix B.

4 Evaluation metrics

We developed two benchmarking criteria to evaluate the performance of denoising algorithms. Benchmark 1: Denoising Score provides a quantitative measure of denoising performance based on the signal-to-noise ratio. This score is designed to be linear with respect to the noise level and equal to one when no denoising is applied. While Benchmark 1 offers a quick, straightforward assessment of model performance, it lacks direct relevance to fundamental science. To bridge this gap, we developed Benchmark 2: Dark Matter Limit, which directly links AI algorithms to community-standard physics result by automating the entire dark matter analysis on the science dataset. This benchmark allows AI algorithms to directly improve the physics reach of dark matter detectors.

4.1 Benchmark 1: denoising score

The denoising score is a modified signal-to-noise ratio (SNR) of the denoised CH1 SQUID time series. It is calculated over the validation dataset by first segmenting both the injected and SQUID time series into one-second segments. Each second of the time series is transformed into a power spectral density (PSD) using a squared fast Fourier transform. This frequency domain data records signal power at each frequency – $PSD(\nu)$. Since the injected dark matter signal is a clean sinusodial oscillation, it should appear as a single-bin peak (ν_0) in the PSD, while noise in SQUID time series is distributed across all frequency bins. The location of ν_0 is identified by the PSD of the injected time series (ground truth) as the largest single bin peak relative to its nearest neighbors.

$$\nu_0 = \underset{\nu}{\operatorname{argmax}}\Big(PSD_{\text{Injected}}(\nu) - (PSD_{\text{Injected}}(\nu - df) + PSD_{\text{Injected}}(\nu + df))\Big) \tag{3}$$

where df is the sampling frequency of 10^{-7} Hz. Once ν_0 is identified in the injected time series, the signal region is defined by selecting $n_{sig} = \pm 1$ bins around the signal frequency to account for spectral leakage. Similarly, the noise region is defined by selecting $n_{bkgd} = \pm 50$ bins outside of the signal region. By taking the ratio of the PSD in the signal region to that in the noise region, we acquire the SNR for each one-second segment PSD.

$$SNR_i = \left(\frac{P_{sig}}{P_{noise}}\right)_i = \sum_{\nu=\nu_0-\nu_{sig}}^{\nu_0+\nu_{sig}} PSD_i(\nu) \Bigg/ \sum_{\nu=\nu_0-\nu_{bkg}}^{\nu_0+\nu_{bkg}} PSD_i(\nu) \tag{4}$$

Multiplying by the sampling frequency (df) turns bin range ($n_{sig,bkg}$) to frequency range ($\nu_{sig,bkg}$).

The hardware setup includes a bandpass filter between the pickups and the digitizer, resulting in a frequency dependence for the signal magnitude in both the injected and the SQUID time series. To account for this, we first calculate the normalized injected SNR:

$$(SNR'_{\text{Injected}})_i = \frac{(SNR_{\text{Injected}})_i}{\max(SNR_{\text{Injected}})} \tag{5}$$

The SQUID SNR then gets multiplied to the corresponding, normalized SQUID SNR in the same one-second segments, and then summed over all one-second segments to produce Λ defined below:

$$\Lambda = \left(\frac{1}{n}\sum_{i=0}^{n}(SNR_{\text{SQUID}})_i \times (SNR'_{\text{Injected}})_i\right) \tag{6}$$

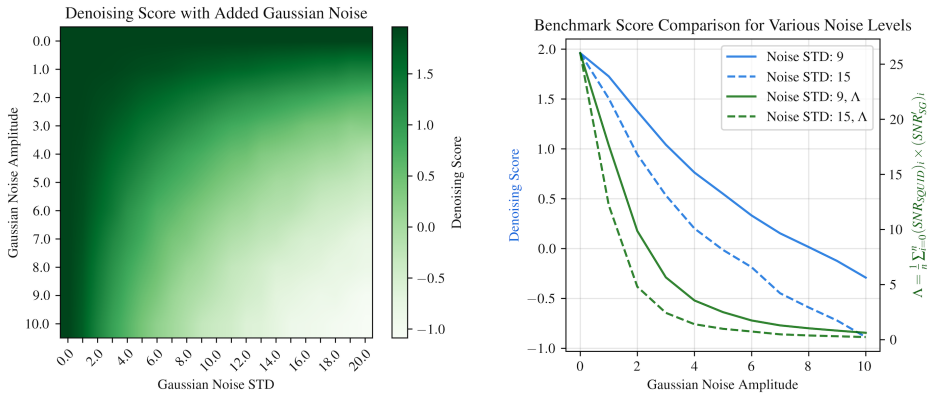

Figure 4: Left: The color bar represents the denoising score for 20s of raw data with added Gaussian noise showing that noisier data results in a lower score. Right: Denoising score and Λ for 20s of raw data with added Gaussian noise of variable noise amplitude. The exponential behavior of Λ can clearly be seen in contrast to the smooth linearity of the denoising score.

We examine the validity of Λ as a measure of denoising efficiency through a controlled study involving added Gaussian noise. The ABRA-10cm detector is subject to a variety of independent noise sources both internal and external, and both stable and stochastic in time. Together, these real noise sources produce a distinctly non-Gaussian noise distribution, as demonstrated in Figure 2. Although the detector noise itself is non-Gaussian, we introduce synthetic Gaussian noise to the real time series data to probe how Λ behaves under systematically varied and well-characterized noise conditions. This approach allows the denoising score to be linearized and validated in a controlled setting, even if not precisely replicating the complex, full statistical structure of the detector noise. In this study, Gaussian noise is added to the SQUID time series and Λ values are computed for a range of added noise amplitudes and standard deviations. As shown in Figure 4 (right), Λ exhibits an exponential decay trend with increased noise amplitude. To establish a linear correlation between the denoising score and noise, we apply a logarithmic transformation to Λ to obtain the final denoising score:

$$\text{Denoising Score} = log_{5.27}\Lambda \tag{7}$$

The base of the logarithm is chosen to be 5.27 so that the denoising score equals 1 for the raw SQUID time series over the validation dataset (i.e., when no denoising algorithm is applied). We further examined this denoising score over a range of imposed Gaussian noise amplitude and STD, and observed a smooth linear response as shown in Figure 4 (left).

This denoising score is implemented in the provided script, `benchmark.py`, which takes as input the denoised SQUID time series of the validation dataset produced by model inference. The script is designed for parallelization and takes about 30 minutes to run on an 8-core CPU node. To further reduce the time required for calculating the denoising score, we defined this second-by-second scan as the Fine Score and introduced a new Coarse Score Denoising Score. The Coarse Score is a tenfold downsample of the full Fine Scan thereby providing a fast benchmarking score that users can leverage to get a rough estimate of model performance in 10% the computational time.

Table 1 shows the denoising scores for all algorithms discussed in Section 3. The case with no denoising is shown in the first row, with its fine denoising score calibrated to 1. The deep learning algorithms repeatedly outperform all of the traditional denoising methods. Based on the results, all traditional algorithms decrease the denoising score because time domain averaging erases high-frequency signals in the region of interest. Meanwhile, all deep learning algorithms efficiently boost the denoising score. We observed that in all cases, the coarse denoising score is slightly higher than the fine denoising score. Surprisingly, we observed that the FC Net model achieved the best performance with a denoising score of 6.43.

The model training and inference is conducted upon the SDSC Expanse cluster equipped with V100 GPUs. For benchmarking tasks, model training and inference over the denoising score data takes less than O(100 GPU hours). The actual value varies slightly with model complexity and data size.

Table 1: Fine and coarse denoising score for raw data, traditional algorithms, and trained ML models. FS means frequency splitting, or training multiple versions of the same model to handle different frequency ranges. The detail of segment size and FS is discussed in Section B.

Algorithms	Segment Size	FS	Parameters	Fine Score	Coarse Score
None				1.00	1.10
Fourier Averaging	1×10^8	–	10-fold Average	0.24	0.26
Moving Average	1×10^6	–	window = 100	0.86	0.95
SG Filter	1×10^6	–	window = 19, order = 11	0.95	1.04
FC Net	4×10^4	Yes	See Appendix B	6.43	6.55
PU Net	4×10^4	Yes	See Appendix B	3.69	3.84
Transformer	2×10^4	Yes	See Appendix B	3.95	4.18
WaveNet	4×10^4	No	See Appendix B	4.99	5.16
RNN Seq2Seq	4×10^4	Yes	See Appendix B	3.38	3.79

4.2 Benchmark 2: dark matter limit

The second benchmark empowers algorithm creators with the capability to conduct a community-standard dark matter search using the science dataset. This benchmark bridges the gap between ML and particle physics; ML developers can translate algorithmic performance directly to improved particle physics experimental reach. By abstracting the full physics analysis chain, the ML community can translate their denoising results to physics results without understanding the underlying physics, all while maintaining full physics rigor.

The two physics parameters for the dark matter candidate in this paper are the dark matter mass (m_a) and the dark matter to electromagnetic coupling ($g_{a\gamma\gamma}$). Null results from different dark matter experiments place limits within this parameter space expressed by the shaded regions in Figure 5. In the physics community, a better dark matter limit is represented by pushing towards lower values of $g_{a\gamma\gamma}$ at different m_a. For dark matter physicists, this dark matter exclusion limit is the community standard for benchmarking different detector performances against each other on the metric of experimental sensitivity to theoretical dark matter candidates. We provide a comprehensive tool necessary for performing the statistical analysis to produce dark matter limits in Figure 5. The limit-setting procedure is repeated for 11.1 million independent m_a (f_a) from 0.4 neV (100 kHz) to 8 neV (2 MHz). The dark matter limit at each m_a is obtained using a frequentist log-likelihood ratio test statistic (TS). The details of this analysis can be found in Appendix C. However, by simply inputting the provided science data, denoised with an algorithm of choice, into the `brazilband.py` script, ML developers can automatically produce Figure 5 to measure the improvements to physics experimental reach enabled by their denoising algorithm.

This analysis is performed twice on the SQUID time series of the science dataset: once without any denoising algorithm which produces the ABRA-TIDMAD Raw limit, and once with FC Net, the top-performing denoising algorithm, which produces the ABRA-TIDMAD Denoised limit. These limits can be directly compared to the previous world-leading ABRA-10cm Run 3 limit, limits obtained by other dark matter experiments, as well as theoretical predictions [28]. While the ABRA-TIDMAD Denoised limit does not outperform the ABRA-10cm Run 3 limit because of hardware and time constraints, it is evident that denoising algorithms significantly improved the dark matter limit by 1-2 orders of magnitude across different m_a. Although the size of the ABRA-TIDMAD science dataset is only 1% of the ABRA-10cm Run 3 science dataset, the AI denoising algorithm boosted the ABRA-TIDMAD limit to nearly the same level as ABRA-10cm Run 3 and even surpassed the ABRA-10cm Run 3 limits at small m_a.

5 Limitations and applications

Hardware and datataking period: As illustrated in Figure 5, the baseline models fail to surpass the results of ABRA-10cm Run 3. This is attributed to hardware limitations and changes since the last

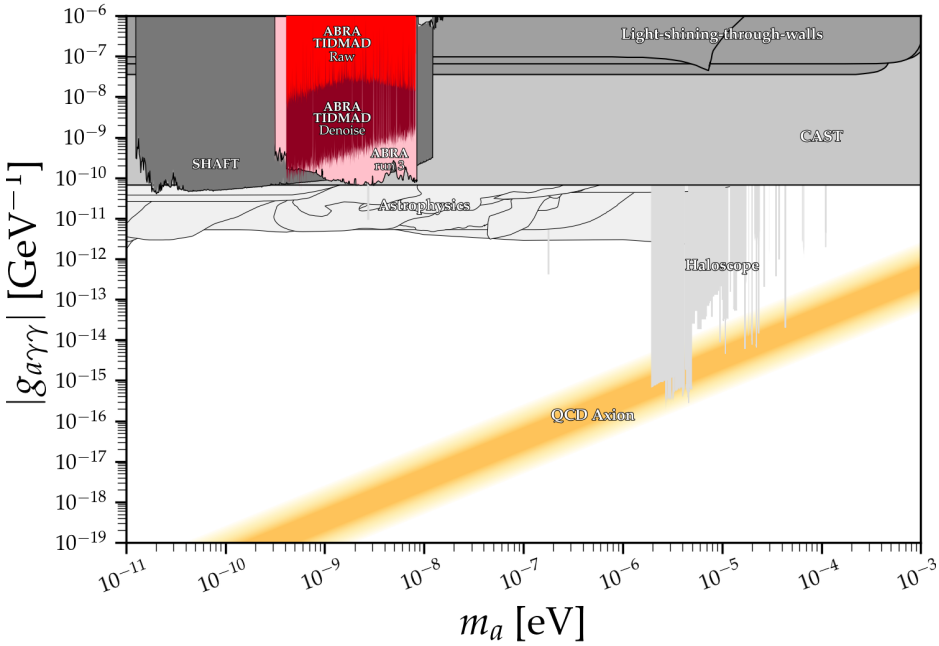

Figure 5: Plotted alongside the present, state-of-the-art axion dark matter limits (grey) are the 2σ exclusion limits for ABRA-10cm Run 3 (pink) [28], ABRA-TIDMAD Raw limit (red), and ABRA-TIDMAD Denoised limit from the trained FC net (maroon). The ABRA-TIDMAD result does not surpass ABRA-10cm Run 3 due to reduced geometric coupling and shorter data taking duration in the TIDMAD configuration. Shaded regions correspond to pairings of dark matter model parameters (m_a, $g_{a\gamma\gamma}$) that are ruled out by the specified experiment, and the bright yellow region indicates theoretical predictions (See Appendix D for details). The plotting script is modified from the publicly available `AxionLimits` repository [19].

data run, including (1) replacing the dark matter pickup cylinder with a pickup loop, consequently reducing the geometric coupling to the dark matter signal, and (2) reducing the data taking period to 24 hours from three months. The decision to implement these changes was driven by the aim to enable ABRA-10cm to simultaneously search for dark matter and gravitational waves, thereby enhancing the scientific scope of the experiment. The shortened data taking duration was necessitated by operational constraints of the dilution refrigerator. Because the signal-to-noise ratio scales as the fourth root of integration time, we can increase ABRA-10cm's sensitivity to dark matter by increasing the data taking period [20]. Another more efficacious way to increase this signal-to-noise ratio is improving our denoising with ML; a doubling of our noise reduction represents a *16x* speed up our data taking time revealing the out-sized return on investment in denoising techniques when compared to increased detector run time.

Null result vs. potential discovery: In Section 4.2, we discussed how to set a dark matter limit using the provided analysis scripts. This script assumes a null result as no 5σ dark matter candidates were identified in this region of parameter space by ABRA-10cm Run 3. Therefore, we assume a null result for this much shorter (24 hr) data taking. This assumption enables us to establish upper limits on the coupling parameter $g_{a\gamma\gamma}$ for every mass point. As shown in Figure 5, the ABRA-TIDMAD Raw limit without denoising covers a smaller region than ABRA-10cm Run 3 [28].

However, with the denoising algorithm applied to the 24-hour science data, ABRA-TIDMAD could potentially reach beyond the ABRA-10cm Run 3 region, where a discovery of dark matter is possible. In this paper, we focused on increasing the experimental sensitivity and setting exclusion limits. A straightforward modification to the interpretation of the TS would unlock the discovery potential of this analysis framework. Future efforts will focus on employing a more extensive dataset and implementing dark matter discovery analysis code. Given that discovering dark matter would be a Nobel Prize-level breakthrough, it is crucial to not only claim discovery but also to convince the scientific community of its validity. If TIDMAD users find any anomalous signals in the

science dataset, please contact the authors for further investigation and understanding of systematic uncertainties.

Generalizability: Axion dark matter is a specific subset of wave-like dark matter candidates, making the techniques developed in this paper broadly applicable to a wide range of wave-like dark matter experiments. Other axion dark matter experiments, including but not limited to ADMX [8], HAYSTAC [7], and DMRadio [9], also produces long time series data and search for similar peaks in frequency domain; any AI algorithm developed upon TIDMAD can be easily adapted and applied to these experiments. Because the denoising and benchmarking methods are detector-agnostic, the authors are presently collaborating with several other axion dark matter experiments beyond ABRA-10cm on implementing the denoising procedure presented in this paper to various experimental set ups. This work lays the groundwork for such generalization, and we hope the paper will encourage similar releases from other experimental collaborations. Furthermore, time series denoising algorithms are crucial for extracting wave-like signals in various areas of physics. In astrophysics, gravitational wave searches involve detecting chirp signals with durations on the order of seconds [1, 23], often buried within detector noise. In nuclear physics, denoising can enhance the efficiency of HPGe detectors [5] and bolometer detectors [35]. Advancements in denoising can be deployed across a suite of these physics experiments.

While the data released in this paper was tailored to our specific problem statement and benchmarks in physics, these ultra-long time series datasets have the potential to benefit a wide range of applications beyond physics. Similar to TIDMAD, many other scientific domains involves time series datasets exhibiting relatively uniform frequency characteristics, with the primary analytical task focused on extracting signals from these time series. Examples include pulsar timing from radio observatory data (astronomy)[16], detecting seismic arrivals above background noise (geology)[36], identifying sea surface and near-surface temperature anomalies (climate science)[30], and recognizing atrial fibrillation among noises and other rhythms in short-term ECG recordings (health science)[10]. If a foundation model were to be developed for general time series analysis in science, our frequency-rich, detector-generated, long time series data could provide a uniquely abundant source of spectral complexity.

6 Conclusions and other works

We present TIDMAD, the first dataset and benchmark designed to yield a community-standard dark matter search result. TIDMAD includes all necessary inputs and processing to train time series denoising algorithms and produce a science-level dark matter limit. Through a series of experiments, we developed five ultra-long time series deep learning algorithms, benchmarked their ability to recover hardware-injected signals, and set dark matter limits. Clear performance improvements were demonstrated on both benchmarks. Our future work will focus on enhancing the denoising algorithm to achieve better dark matter limits, expanding to other nuclear and particle experiments, and embedding these algorithms onto FPGA chips for real-time denoising during data taking.

The aim of this data release is to enable the ML community to use TIDMAD to develop algorithms tailored for data with highly coherent embedded signals. This development would not only extend the experimental reach of dark matter searches, leading to improved dark matter limits, but also allow the AI/ML community to make direct scientific advancements. This transparency aims to foster greater collaboration between the ML and particle physics communities, benefiting both fields.

Acknowledgements

J. T. Fry is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2141064. A. Li's work is supported by his startup grant. This material is based upon work supported by the National Science Foundation under Grant No. 2110720. We would like to thank San Diego Supercomputer Center (SDSC) and Open Science Data Federation (OSDF) for the data storage space and open access infrastructure. OSDF is funded under National Science Foundation under Grant No. 2030508, 2331480, and 2112167. We'd like to thank Frank Wuerthwein for many insightful discussions and offering the OSDF storage space, as well as Fabio Andrijauskas for the technical support. The computation of this project was conducted on SDSC Expanse cluster as well as MIT's subMIT cluster. We would also like to thank Joshua Foster for his invaluable contributions to the axion analysis.

References

[1] B. P. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, F. Acernese, K. Ackley, and C. et al Adams. Observation of gravitational waves from a binary black hole merger. *Phys. Rev. Lett.*, 116:061102, Feb 2016. doi: 10.1103/PhysRevLett.116.061102. URL `https://link.aps.org/doi/10.1103/PhysRevLett.116.061102`.

[2] C. B. Adams and et. al. Axion dark matter, 2023.

[3] N. Aghanim and et. al. Planck2018 results: I. overview and the cosmological legacy ofplanck. *Astronomy; Astrophysics*, 641:A1, September 2020. ISSN 1432-0746. doi: 10.1051/0004-6361/201833880. URL `http://dx.doi.org/10.1051/0004-6361/201833880`.

[4] Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Jos van der Velde, Steffen Vogler, and Carole-Jean Wu. Croissant: A metadata format for ml-ready datasets. In *Association for Computing Machinery*, DEEM '24, page 1–6, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706110. doi: 10.1145/3650203.3663326. URL `https://doi.org/10.1145/3650203.3663326`.

[5] Mark R. Anderson, Vasundhara Basu, Ryan D. Martin, Charlotte Z. Reed, Noah J. Rowe, Mehdi Shafiee, and Tianai Ye. Performance of a convolutional autoencoder designed to remove electronic noise from p-type point contact germanium detector signals. *Eur. Phys. J. C*, 82(12):1084, 2022. doi: 10.1140/epjc/s10052-022-11000-w.

[6] Fabio Andrijauskas, Derek Weitzel, and Frank Wuerthwein. Open science data federation - operation and monitoring. In *Practice and Experience in Advanced Research Computing 2024: Human Powered Computing*, PEARC '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704192. doi: 10.1145/3626203.3670557. URL `https://doi.org/10.1145/3626203.3670557`.

[7] Kelly M Backes, Daniel A Palken, S Al Kenany, Benjamin M Brubaker, SB Cahn, A Droster, Gene C Hilton, Sumita Ghosh, H Jackson, Steve K Lamoreaux, et al. A quantum enhanced search for dark matter axions. *Nature*, 590(7845):238–242, 2021.

[8] T. Braine, R. Cervantes, N. Crisosto, N. Du, S. Kimes, L. J. Rosenberg, G. Rybka, and J. et al. Yang. Extended search for the invisible axion with the axion dark matter experiment. *Phys. Rev. Lett.*, 124:101303, Mar 2020. doi: 10.1103/PhysRevLett.124.101303. URL `https://link.aps.org/doi/10.1103/PhysRevLett.124.101303`.

[9] L. Brouwer, S. Chaudhuri, H.-M. Cho, J. Corbin, W. Craddock, C. S. Dawson, A. Droster, J. W. Foster, J. T. Fry, P. W. Graham, R. Henning, K. D. Irwin, F. Kadribasic, Y. Kahn, A. Keller, R. Kolevatov, S. Kuenstner, A. F. Leder, D. Li, J. L. Ouellet, K. M. W. Pappas, A. Phipps, N. M. Rapidis, B. R. Safdi, C. P. Salemi, M. Simanovskaia, J. Singh, E. C. van Assendelft, K. van Bibber, K. Wells, L. Winslow, W. J. Wisniewski, and B. A. Young. Projected sensitivity of dmradio-m^3: A search for the qcd axion below 1 μeV. *Phys. Rev. D*, 106:103008, Nov 2022. doi: 10.1103/PhysRevD.106.103008. URL `https://link.aps.org/doi/10.1103/PhysRevD.106.103008`.

[10] Gari D Clifford, Chengyu Liu, Benjamin Moody, H Lehman Li-wei, Ikaro Silva, Qiao Li, AE Johnson, and Roger G Mark. Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[11] ABRACADABRA Collaboration. Systematic approach to understanding noise sources in abracadabra-10 cm. *To be submitted to JINST*, 2025.

[12] Pablo F de Salas and A Widmark. Dark matter local density determination: recent observations and future prospects. *Reports on Progress in Physics*, 84(10):104901, oct 2021. doi: 10.1088/1361-6633/ac24e7. URL `https://dx.doi.org/10.1088/1361-6633/ac24e7`.

[13] Michael Dine and Willy Fischler. The not-so-harmless axion. *Physics Letters B*, 120(1):137–141, 1983. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(83)90639-1. URL `https://www.sciencedirect.com/science/article/pii/0370269383906391`.

[14] Michael Dine, Willy Fischler, and Mark Srednicki. A simple solution to the strong cp problem with a harmless axion. *Physics Letters B*, 104(3):199–202, 1981. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(81)90590-6. URL `https://www.sciencedirect.com/science/article/pii/0370269381905906`.

[15] Joshua W. Foster, Nicholas L. Rodd, and Benjamin R. Safdi. Revealing the dark matter halo with axion direct detection. *Phys. Rev. D*, 97:123006, Jun 2018. doi: 10.1103/PhysRevD.97.123006. URL https://link.aps.org/doi/10.1103/PhysRevD.97.123006.

[16] George Hobbs, R. Edwards, and R. Manchester. Tempo2, a new pulsar timing package. 1. overview. *Mon. Not. Roy. Astron. Soc.*, 369:655–672, 2006. doi: 10.1111/j.1365-2966.2006.10302.x.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. URL https://openaccess.thecvf.com/content_ICCV_2017/papers/Lin_Focal_Loss_for_ICCV_2017_paper.pdf.

[18] Sydney Montiel, Alexsandra Guadarrama, and Fabio Andrijauskas. Using the open science data federation for data distribution: Big bear solar observatory use case. In *Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration*, PEARC '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713989. doi: 10.1145/3708035.3736032. URL https://doi.org/10.1145/3708035.3736032.

[19] Ciaran O'Hare. cajohare/axionlimits: Axionlimits, 2025. URL https://doi.org/10.5281/zenodo.3932430.

[20] Jonathan Ouellet and Zachary Bogorad. Solutions to axion electrodynamics in various geometries. *Physical Review D*, 99(5), March 2019. ISSN 2470-0029. doi: 10.1103/physrevd.99.055010. URL http://dx.doi.org/10.1103/PhysRevD.99.055010.

[21] Jonathan L. Ouellet and et. al. First results from abracadabra-10 cm: A search for sub-μeV axion dark matter. *Phys. Rev. Lett.*, 122:121802, Mar 2019. doi: 10.1103/PhysRevLett.122.121802. URL https://link.aps.org/doi/10.1103/PhysRevLett.122.121802.

[22] Jonathan L. Ouellet and et. al. Design and implementation of the abracadabra-10 cm axion dark matter search. *Phys. Rev. D*, 99:052012, Mar 2019. doi: 10.1103/PhysRevD.99.052012. URL https://link.aps.org/doi/10.1103/PhysRevD.99.052012.

[23] Kaliroë M. W. Pappas, Jessica T. Fry, Sabrina Cheng, Arianna Colón Cesaní, Jonathan L. Ouellet, Chiara P. Salemi, Inoela Vital, Lindley Winslow, Valerie Domcke, Sung Mook Lee, Joshua W. Foster, Reyco Henning, Yonatan Kahn, Nicholas L. Rodd, and Benjamin R. Safdi. High-frequency gravitational wave search with abracadabra-10 cm, 2025. URL https://arxiv.org/abs/2505.02821.

[24] John Preskill, Mark B. Wise, and Frank Wilczek. Cosmology of the invisible axion. *Physics Letters B*, 120(1):127–132, 1983. ISSN 0370-2693. doi: https://doi.org/10.1016/0370-2693(83)90637-8. URL https://www.sciencedirect.com/science/article/pii/0370269383906378.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv*, 2015.

[26] Vera C. Rubin and Jr. Ford, W. Kent. Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions. *APJ*, 159:379, February 1970. doi: 10.1086/150317.

[27] Muhammed Saleem, Alec Gunny, Chia-Jui Chou, Li-Cheng Yang, Shu-Wei Yeh, Andy Chen, Ryan Magee, William Benoit, Tri Nguyen, Pinchen Fan, Deep Chatterjee, Ethan Marx, Eric Moreno, Rafia Omer, Ryan Raikman, Dylan Rankin, Ritwik Sharma, Michael Coughlin, Philip Harris, and Erik Katsavounidis. Demonstration of machine learning-assisted real-time noise regression in gravitational wave detectors, 06 2023.

[28] Chiara P. Salemi and et. al. Search for low-mass axion dark matter with abracadabra-10 cm. *Phys. Rev. Lett.*, 127:081801, Aug 2021. doi: 10.1103/PhysRevLett.127.081801. URL https://link.aps.org/doi/10.1103/PhysRevLett.127.081801.

[29] M.A. Shifman, A.I. Vainshtein, and V.I. Zakharov. Can confinement ensure natural cp invariance of strong interactions? *Nuclear Physics B*, 166(3):493–506, 1980. ISSN 0550-3213. doi: https://doi.org/10.1016/0550-3213(80)90209-6. URL https://www.sciencedirect.com/science/article/pii/0550321380902096.

[30] Thomas M. Smith, Richard W. Reynolds, Thomas C. Peterson, and Jay Lawrimore. Improvements to noaa's historical merged land–ocean surface temperature analysis (1880–2006). *Journal of Climate*, 21(10): 2283 – 2296, 2008. doi: 10.1175/2007JCLI2100.1. URL https://journals.ametsoc.org/view/journals/clim/21/10/2007jcli2100.1.xml.

[31] Max Tegmark and et. al. Cosmological parameters from sdss and wmap. *Physical Review D*, 69(10), May 2004. ISSN 1550-2368. doi: 10.1103/physrevd.69.103501. URL http://dx.doi.org/10.1103/PhysRevD.69.103501.

[32] J. Anthony Tyson, Greg P. Kochanski, and Ian P. Dell'Antonio. Detailed mass map of cl 0024+1654 from strong lensing. *The Astrophysical Journal*, 498(2):L107–L110, May 1998. ISSN 0004-637X. doi: 10.1086/311314. URL http://dx.doi.org/10.1086/311314.

[33] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[35] Kenneth J. Vetter et al. Improving the performance of cryogenic calorimeters with nonlinear multivariate noise cancellation algorithms. *Eur. Phys. J. C*, 84(3):243, 2024. doi: 10.1140/epjc/s10052-024-12595-y.

[36] Spahr C. Webb. 19 - seismic noise on land and on the sea floor. In William H.K. Lee, Hiroo Kanamori, Paul C. Jennings, and Carl Kisslinger, editors, *International Handbook of Earthquake and Engineering Seismology, Part A*, volume 81 of *International Geophysics*, pages 305–318. Academic Press, 2002. doi: https://doi.org/10.1016/S0074-6142(02)80222-4. URL https://www.sciencedirect.com/science/article/pii/S0074614202802224.

[37] Derek Weitzel, Brian Bockelman, Duncan A. Brown, Peter Couvares, Frank Würthwein, and Edgar Fajardo Hernandez. Data access for ligo on the osg. In *Practice and Experience in Advanced Research Computing 2017: Sustainability, Success and Impact*, PEARC '17, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450352727. doi: 10.1145/3093338.3093363. URL https://doi.org/10.1145/3093338.3093363.

A Dark Matter Signal and Signal Injection

Axion dark matter appears in our detector as a time oscillating current with two free model parameters $(m_a, g_{a\gamma\gamma})$ described by Equation 1. This signal can be fully derived from axion theory and is extremely well-defined. While a full derivation of this equation is presented in Reference [20], schematically, this theoretical prediction for the dark matter signal in our detection comes from four phenomena – the omnipresent axion field, axion's interactions with electromagnetism, the specific geometry of our detector, and the velocity distribution of the dark matter.

Under certain conditions, axions are created in the early universe via the misalignment mechanism [13, 24]. When axions are created they produce a time oscillating omnipresent axion field with an oscillation frequency equal to m_a and phase coherence. If axions are dark matter then the abundance of axions and the strength of this axion field are set by our astrophysical observations of dark matter density [12].

The axion field can interact with other physical forces including electromagnetism. If an axion field collides with an electromagnetic field, some of the axions will be transformed into photons. These photons produce a secondary electromagnetic field – an axion-induced electromagnetic field. In the geometry of our detector, the axion-induced electromagnetic field is read out with a pickup wire which samples the field like a typical radio antenna. This means the time oscillating axion field turns into a time oscillating current in our detector (i.e. our dark matter signal).

If earth were stationary, sitting in a uniform bath of dark matter, then the dark matter signal frequency would be exactly the axion field oscillation frequency (i.e. the axion mass) and completely coherent. However, the earth is moving within the Milky Way galaxy and thus the axion field gains a velocity with respect to earth at about $v_{DM} \sim 220km/s$. Doppler shifting spreads this frequency such that it is distribution around the original field frequency $\Delta f = v_{DM}^2 f$ [20]. This frequency distribution is six orders of magnitude smaller than the signal frequency, therefore can be treated a coherent, single frequency sine-wave to good approximation. For signal injection, this single frequency approximation is used while for the dark matter analysis, we model the full frequency distribution.

Thus, we have established that the dark matter signal in the ABRA-10cm detector is approximately sine-wave current. The frequency of this oscillating signal is a model parameter meaning if we knew the axion mass, this frequency would be set. However, theoretical models point to a range of possible axion masses, not a singular value. Ideally we would create a dark matter detector that could search the entire range of valid dark matter masses, but experimental constraints such detector size, configuration, and readout electronics preclude this possibility. Instead, experiments must be tailored to search for smaller areas of the axion mass parameter space with ABRA-10cm the detector being specifically designed to target $m_a = 0.4 - 10$ neV [21].

The second free signal parameter is the sine wave amplitude which is proportional to $g_{a\gamma\gamma}$, the strength of the axion's interactions with electromagnetism. Theoretical calculations constrain this parameter to $g_{a\gamma\gamma} = Cm_a$ where $C = [-0.39, 0.22]$ depending on the theory [14, 29]. Theoretically motivated axion couplings can be seen in Figure 5 as the gold band in the $m_a, g_{a\gamma\gamma}$ parameter space. Experimentally, the goal is to detect ever smaller signal amplitudes to reach lower values of $g_{a\gamma\gamma}$.

To inject a fake signal into the hardware, we replicate the signal current, given by Equation 1, with a signal generator. The signal generator is connected to a calibration loop depicted in Figure 1 designed to mimic an axion field incident on the detector. We specifically choose to inject sine waves with frequencies from 1.1 kHz to 4.9 MHz, $m_a = [0.005, 17]$ neV, to contain the masses our experimental hardware was built to target. The injected fake signals we used all have amplitudes of 50 mV to achieve a reasonable signal-to-noise ratio. While injecting smaller fake signal amplitudes would effectively simulate dark matter candidates with smaller electromagnetic couplings, fake signals smaller than 50 mV are difficult to detect with traditional techniques. Though smaller couplings provide an interesting ML task, our denoising score benchmark is predicated on finding injected signals with traditional techniques and subsequently we did not use smaller fake signal amplitudes. However, we did take data injected with signals ranging from 1.1 kHz to 4.9 MHz at a signal amplitude of 10 mV. While this data was not used in our benchmark creation or model training, it is publicly available (see Datasheet) for an extra challenge.

To summarize, our signal injection scheme involves exciting hardware with a sine wave from a signal generator. We sequentially step through 309 different frequencies from 1.1 kHz to 4.9 MHz to

simulate 309 different axion masses in our detector hardware all with an amplitude of 50 mV so that the fake signal is visible above the detector noise floor.

B Details and limitations of deep learning model

There are two special treatments we took to train the five deep learning models:

Training segmentation: To feed the time series data into limited GPU memories, the training time series are segmented into 4 milliseconds. This imposes a fundamental lower limit on the frequencies that the models can detect. While this lower limit, approximately 250 Hz given the sampling frequency of 10 MS/s, is relatively small, it does establish a foundational lower threshold for frequency resolution in the model. The transformer model requires additional memory, therefore we have to further reduce the segment to 2 millisecond or 500 Hz. Both of these limits are well below the dark matter search range: 0.4 neV (100 kHz) to 8 neV (2 MHz).

Frequency splitting: Since the injected dark matter signal spans two orders of magnitude in frequency, the observed features in the injected time series significantly vary. During training, we noticed that a single deep learning model (with the exception of WaveNet) to denoise the entire dataset would fail to generalize across the different injected frequency ranges. To address this issue, we trained four deep learning models per architecture, each focusing on a specific frequency range: the first covering the low-frequency regime (training/validation files 0-3), the second covering the mid-low regime (training/validation files 4-9), the third covering the mid-high regime (training/validation files 10-14), and the fourth covering the high-frequency regime (training/validation files 15-19). During the benchmark 1 inference, we selected the input validation data corresponding to the frequency range for which each model was trained and averaged the results of the four models. For the benchmark 2 inference, we ran all four models on the science data and selected the highest-performing model, as represented in Figure 5.

The WaveNet architecture was the only deep learning model that generalized across the entire frequency range of validation data. All trained models are made available in `https://drive.google.com/drive/folders/16ORX1b2zo1_lOYYAcRBgddBuYImjOBxs?usp=share_link`.

The hyperparameters of the FC Net are listed below:

```
AE(
  (encoder): Sequential(
    (0): Linear(in_features=40000, out_features=4000, bias=True)
    (1): ReLU()
    (2): Linear(in_features=4000, out_features=400, bias=True)
    (3): ReLU()
    (4): Linear(in_features=400, out_features=40, bias=True)
  )
  (decoder): Sequential(
    (0): Linear(in_features=40, out_features=400, bias=True)
    (1): ReLU()
    (2): Linear(in_features=400, out_features=4000, bias=True)
    (3): ReLU()
    (4): Linear(in_features=4000, out_features=40000, bias=True)
  )
)
```

The output of FC Net at every time step is a single floating point number. An MSE loss is calculated between the floating point number and the ground truth value.

The PU Net model consists of four down layers and four up layers, with contracting paths between each pair of layers. The down layers include Max Pooling and two convolutional operations, while the up layers comprise Deconvolution and Convolution. Additionally, positional encoding is added after each down layer [34]. Lastly, the output is fed into a linear layer to produce 256-dimensional

vector at each time step. The detailed model hyperparameter could be found within the `network.py` script in `https://github.com/jessicafry/TIDMAD`.

The transformer model processes the time series data by using an Embedding layer to encode each input, converting 8-bit integers in the range of $(-128, 127)$ into a 32-dimensional vector. Positional encoding is then added to the embedded time series [34]. This augmented data is fed into a Transformer Encoder with two layers, each containing two heads, 128 hidden dimensions, and a 0.1 dropout rate. Finally, the output is passed through a linear layer to produce a 256-dimensional vector at each time step.

For both PU Net and Transformer, the output at each time step is a 256-dimensional vector, corresponding to 256 possible output classes. This can be considered as a time series semantic segmentation task where there are 256 possible classes to choose from. We adopted Focal Loss in Object Detection to address the class imbalance problem in semantic segmentation task [17].

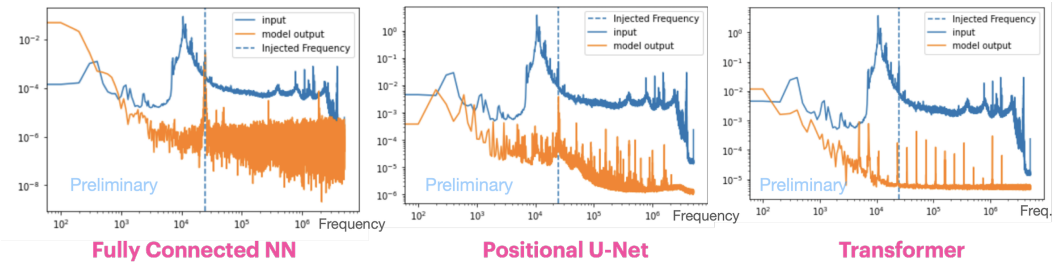

Figure 6: The denoising performance of FC Net (Left), PU Net (Middle), and Transformer (Right) at a single injected frequency. The plot is made by Fourier-transform the time series into frequency space.

The denoising performance of three models at a single injected frequency is illustrated in Figure 6. In this specific instance, the PU-Net model demonstrates superior denoising performance. However, when evaluated using the Denoising Score across all frequencies, the FC Net significantly outperforms the other two models by a large margin.

The WaveNet model implements the core dilated causal convolution architecture with 10 residual blocks using exponentially increasing dilation rates (1, 2, 4, 8, 16, ..., 512), kernel size of 12, and 32 residual channels with 64 gate channels for the gated activation mechanism (tanh × sigmoid). Each block processes input through dilated causal convolutions to ensure no future information leakage, applies gated activation, and produces both residual connections (for depth) and skip connections (accumulated across all blocks for the final output). The model uses 32 skip channels and projects the final accumulated skip connections through two 1D convolutions to produce the output shape of (batch, sequence_length, 256) while achieving a large receptive field of over 1000 timesteps through the dilated convolution hierarchy, making it suitable for modeling long-term dependencies in sequential data like audio waveforms.

The simple RNN Seq2Seq model follows the classical encoder-decoder architecture with vocabulary size of 256 and 256 output classes, using 128-dimensional embeddings and 256-dimensional LSTM hidden states across 2 layers with 0.1 dropout. The encoder processes the input sequence (batch, seq_len) through an embedding layer and LSTM to produce final hidden and cell states, which initialize the decoder LSTM that generates output logits of shape (batch, seq_len, num_classes) by processing the same input sequence (a simplified teacher-forcing approach for same-length input/output tasks). The model uses separate embedding and LSTM layers for encoder and decoder, accumulates the decoder's hidden states across all timesteps, and projects them through a linear layer to produce class distributions for each position, making it suitable for sequence-to-sequence tasks where input and output have identical lengths, such as sequence labeling or token-level classification with 256 possible output categories per position.

C Frequentist log-likelihood test statistics

The detailed analysis flow to produce the dark matter limit is depicted in Figure 7. The first step involves performing a fast Fourier transform on the time series data in 10-second segments to produce

power spectral densities (PSDs). These PSDs are then averaged across the full dataset to generate our average PSD, reflecting the power in the pickup loop as a function of frequency. Since one of the physics parameters, dark matter mass (m_a), is directly proportional to the frequency, the analysis script conducts 11.1 million independent searches for dark matter with varying mass points from 0.4 neV (100 kHz) to 8 neV (2 MHz).

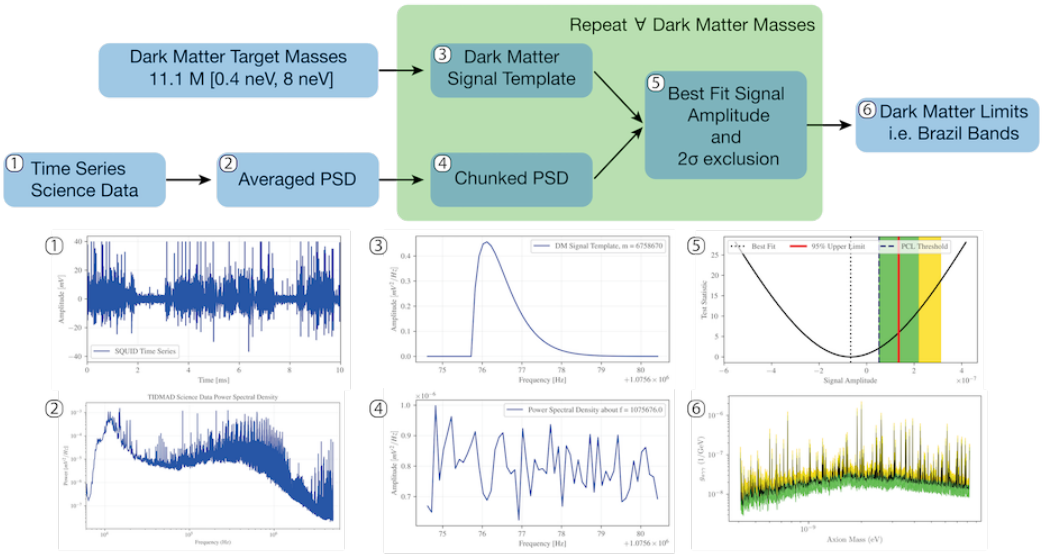

Figure 7: This represents the analysis flow for dark matter science data and the detection logic to build the brazil band limits from time series data (raw or denoised). Key types of data are plotted corresponding to their step in the analysis chain.

At each mass point, the dark matter limit is obtained using a frequentist log-likelihood ratio test statistic (TS) [15]. Given the local velocity distribution and density of dark matter from astrophysical measurements, as well as our choice of dark matter mass, we create a dark matter signal template for each mass point. These templates are compared to a chunked frequency subset of the average PSD, constructed using a sliding window whose width scales as $\delta f/f \approx 5.5 \times 10^{-6}$. We use equation 1.1 and calibration data to produce the other physics parameter $g_{a\gamma\gamma}$ given detector geometry. By floating the template signal amplitude and allowing the mean background level of noise to vary within each sliding window, we fit the signal template to the data to construct a likelihood as a function of $g_{a\gamma\gamma}$. We then use the TS to determine the 95% one-sided upper limits on $g_{a\gamma\gamma}$ for every mass point [15, 28]. The resulting limits on $g_{a\gamma\gamma}$ as a function of m_a (black line) as well as the $1/2\sigma$ containment (green, yellow) can be seen on the dark matter sensitivity plot, i.e. "Brazil band" in Figure 7.

D Current Axion Limits

In Figure 5, the TIDMAD dark matter limits, denoised and raw, are presented alongside ABRA-10cm's previous Run 3 limits and various state-of-the-art axion dark matter experiments and observations. In gold, the theoretically motivated couplings for dark matter candidates are highlighted as discussed in Appendix A. In this appendix, a brief summary of the origins of these limits will be provided.

Light-shining-through-walls: In this class of experiments, a high-intensity laser is directed towards a solid barrier in a magnetic field. While conventional light cannot traverse the barrier, the interaction with the magnetic field may cause a fraction of the light to convert into axions. As a result of their weak interaction with matter, these axions could pass through the wall. On the opposite side, detectors are placed to identify any reconverted light, which would indicate the presence of axions. Light-shining-through-walls experiments must both create axions from photons and detect these axions by converting them back into photons, whereas ABRA-10cm only needs to detect axions, not create them.

Cavity Haloscopes: This class of experiments generate a strong magnetic field to stimulate axions to convert into microwave photons within a resonant cavity-enclosed space. The resonant cavity is finely tuned to amplify specific frequencies of electromagnetic radiation. Sensitive radio receivers then measure the power within the cavity to identify any potential photon signals indicative of axions. While the conversion mechanism is identical to ABRA-10cm, the resonant cavity only amplifies targeted frequencies. In contrast, ABRA-10cm's readout chain has broadband amplification of axion induced signals. ABRA-10cm, Cavity Haloscopes, and SHAFT are all examples of Haloscopes – experiments that search for dark matter axions in our galaxy's dark matter halo.

SHAFT: The Search for Halo Axions with Ferromagnetic Toroids (SHAFT) experiment is an axion haloscope with a broadband readout, similar to ABRA-10cm in both detection and readout mechanism. There are two main differences between SHAFT and ABRA-10cm (1) SHAFT uses toroidal magnets with *ferromagnetic* material in the core to convert the axions (2) SHAFT contains a pairs of stacked ferromagnetic toroids each of which has a separate pickup coil and SQUID readout.

Astrophysics: There are numerous astrophysical processes that would be altered if the axion exists. Broadly, this class of exclusions takes astrophysical observations, calculates how these processes would change if axions exits, and sets limits on possible axion couplings. These limits include the following astrophysical processes. **Stellar Cooling.** Axions produced in hot astrophysical plasma can transport energy out of stars. This transport of energy critically affects stellar lifetimes, thus observations of stellar energy-loss rates can set limits on axion's couplings to matter. **Photon Flux** Large photon fluxes from astrophysical objects like Supernova 1987A traverse the galaxy before being detected terrestrially. Within the galactic magnetic field, some of these supernova photons could be converted into axions. By observing the gamma-ray signals from such events, strong bounds on axions couplings to photons can be derived. **Black Hole Superradiance.** Light particles, like axions, affect the gravitational waves emitted by black holes through the superradiance mechanism in which axion fields extract energy and angular momentum from the black hole. Observations of stellar black hole spin measurements can therefore constrain allowable axion couplings.

CAST: The CERN Axion Solar Telescope (CAST) experiment is a prominent axion helioscope. In contrast to *halo*scopes which search for axions created in the early universe within the dark matter halo surrounding our galaxy, helioscopes search for axions created in our Sun's heliosphere. CAST uses a strong, movable superconducting magnet to convert axions produced in the core of our Sun into into X-ray photons when aligned with the Sun. CAST is equipped with highly sensitive X-ray detectors at both ends of the magnet, designed to capture these photons. By tracking the Sun and searching for excess X-rays that correlate with solar axions, CAST aims to detect solar axions.

E Hyperparameters of Moving Average and SG Filter

We conducted additional studies to understand the effects of different hyperparameters of Moving Average and Savitzky–Golay filter algorithms. The Moving Average method has one hyperparameter: WINDOW SIZE, while the Savitzky–Golay filter has two hyperparameters: WINDOW SIZE and POLYNOMIAL ORDER.

E.1 Experimental Setup

We designed 60 experimental trials to systematically understand these hyperparameters:

- **Moving Average (10 trials):** Window size randomly sampled from the range [10, 5000]
- **Savitzky–Golay filter (50 trials):** Window size randomly sampled as odd numbers from [11, 2001], and polynomial order randomly sampled from [2, 20]

For each trial, we executed the corresponding algorithm and calculated both fine and coarse denoising scores using the evaluation procedure described in Section 4. Each tials take about 6 hours on a 8-Core CPU machine with 64Gb memories.

Table 2: Fine and coarse denoising score for Moving Average and Savitzky–Golay filter under different parameters

Algorithms	Window Size	Polynomial Order	Fine Score	Coarse Score
None			1.00	1.10
Moving Average	19	-	0.86	0.95
Moving Average	99	-	0.56	0.69
Moving Average	100	-	0.52	0.64
Moving Average	622	-	-0.31	-0.47
Moving Average	974	-	-0.57	-0.47
Moving Average	1169	-	-0.61	-0.72
Moving Average	2632	-	-1.07	-0.97
Moving Average	2632	-	-1.07	-0.94
Moving Average	2664	-	-1.07	-0.94
Moving Average	3040	-	-1.23	-1.14
Moving Average	3810	-	-1.23	-1.14
SG Filter	19	11	0.95	1.04
SG Filter	171	5	0.58	0.69
SG Filter	945	2	-0.41	-0.31
SG Filter	405	5	0.30	0.42
SG Filter	1245	15	-2.77	-2.35

E.2 Result and Analysis

The results from the 10 Moving Average trials are presented in Table 2. The complete results from all 50 Savitzky–Golay filter trials are displayed in Figure 5, with a representative subset also included in Table 2.

Moving Average Performance: The denoising score consistently improves as the WINDOW SIZE decreases.

Savitzky–Golay Filter Performance: The algorithm achieves better performance with smaller window sizes and lower polynomial orders. Notably, the best denoising score was achieved in a trial with WINDOW SIZE 19 and POLYNOMIAL ORDER 11. This means the SG fitter fits an 11-dimensional polynomial to every 19 elements in the time series — representing an extreme overfitting scenario. In other words, SG filter in this trial attempts to memorize and preserve all features in the time series, resulting in a Fine Denoising Score of 0.94, which **asymptotically approach but still inferior to the baseline score of 1.0** (when no denoising algorithm is applied).

To validate if moving average has a similar asymptotic behavior, we conducted an additional Moving Average trial with window size 19. This trial, as expected produced the best fine denoising score at 0.85.

E.3 Conclusion and Key Finding

Critically, **none of the 60 trials surpassed the baseline denoising score of 1.0, where no denoising algorithm is applied**. This demonstrates that traditional algorithms are fundamentally inadequate for handling the complex noise characteristics within the TIDMAD dataset. At best, these methods can only asymptotically approximate the raw time series without providing meaningful denoising improvements. These findings underscore the necessity of deep learning-based approaches for effective denoising in the TIDMAD dataset, as traditional signal processing methods prove insufficient for this challenging task.

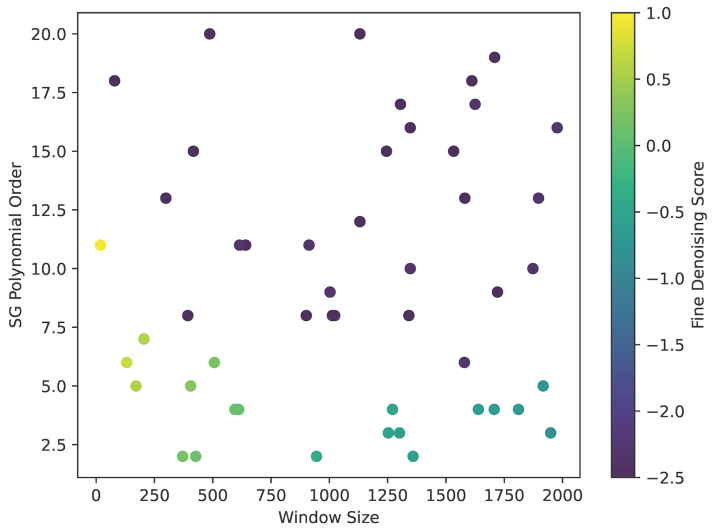

Figure 8: Result of all 50 trials of SG filter. The x axis is the window size, the y axis is the polynomial order of SG filter, while the color of each dot represents its fine denoising score.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made in the paper. The claims in the introduction, abstract, and main paper are the same; we introduce a new, segmented dataset from a physics detector, a denoising score for direct model benchmarking, and a complete dark matter analysis framework for science legible benchmarking.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Included in Section 5 is a discussion of the limitation of this work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors

should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: In Section 1 and Section A, we discuss the physics theory and assumptions for our result. Additionally, for the evaluation metric, we provide a set of assumptions for the benchmark construction in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We publicly provide all of the data and code necessary to reproduce all experimental results of the paper. Please use the link in the abstract to access all necessary materials for reproduction.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We publicly provide all of the data and code necessary to reproduce all experimental results of the paper. Please use the link in the abstract to access our data, benchmarking code, and analysis code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In Section 2, we describe the data training splits. In section 3, we describe the hyperparameters used.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In Section 4, the second benchmark – the dark matter limit – is accompanied by the relevant confidence intervals and statistical significance of the scientific result. Additionally, the noise assumptions are discussed with respect to the denoising score.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: As stated in the body of the paper, the model training and inference is conducted upon the SDSC Expanse cluster equipped with V100 GPUs. For benchmarking tasks, model training and inference over validation and science dataset takes less than O(100 GPU hours). The actual value varies slightly with model complexity and data size.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We do not have human subjects or any data related to human beings. *Safety:* the authors do not foresee how this data could be directly or indirectly used to harm people. *Security:* the research does not introduce security vulnerabilities. *Descrimination:* our technology can not be used to discriminate people. *Surveillance:* we do not use surveillance data. *Deception and Harassment:* this is physics data – it can not be used to facilitate deceptive interactions. *Environment:* our research does not promote negative impacts to the environment outside of the computational resources used for the research. *Human Rights:* no humans were used in this data collection or benchmark creation. *Bias and fairness:* our methods do not encode bias against certain people. Scientific bias is addressed in Section 5.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This is discussed in Section 5. This is no negative societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our data, models, and benchmarks relate to particle physics data. There is no risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators of the data are the authors of the paper. This data was produced by a particle physics detector operated by the authors.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The assets introduced in the paper are documented in the paper and provided with the publically available code repository.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: There was no crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: There were no human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: No LLM was used in the core methods of this research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

TIDMAD: Supplemental Information

Contents

1 Datasheet **2**

1.1 Motivation 2

1.2 Composition 2

1.3 Collection process 2

1.4 Preprocessing/cleaning/labeling 3

1.5 Uses 3

1.6 Distribution 4

1.7 Maintenance 4

2 Dataset and code access **4**

3 Croissant metadata **5**

4 Author statement **5**

1 Datasheet

1.1 Motivation

1. **For what purpose was the dataset creates?** Our datasets were created to train and benchmark ultra-long time series denoising frameworks for the discovery of dark matter.
2. **Who created the dataset and on behalf of which entity?** This dataset is the direct output of the ABRACADABRA detector on behalf of the researchers on the author list. The ABRACADABRA detector was built by the ABRACADABRA collaboration including J. T. Fry and the authors of [28].
3. **Who funded the creation of the dataset?** This work was generously funded by the National Science Foundation under grant numbers NSF-PHY-1658693, NSF-PHY-1806440. J. T. Fry is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2141064

1.2 Composition

1. **What do the instances that comprise the dataset represent?** Each instance represents a voltage at a moment in time read out by our detector. For the SQUID data, this voltage comes from flux on the pickup loop of wire, converted to a voltage by the SQUID detector, read out by our digitizer. For the SG data, this voltage comes directly from a signal generator passed through a power splitter.
2. **How many instances are there in total?** There are 867,260,000,000 voltage instances total.
3. **Does the dataset contain all possible instances or is it a sample of instances from a larger set?** The voltage produced by the SQUID and the SG are continuous. The instances are sampled from this continuous voltage stream at a constant rate of 10MS/s.
4. **What data does each instance consist of?** The data each instance consists of is a raw 8-bit integer from our digitizer. To convert the raw 8-bit integer to a voltage, each bit must be scaled by the ADC voltage i.e. multiply by $40mV/128$.
5. **Is there a label or target associated with each instance?** Yes, for the calibration data, each instance of the SQUID data corresponds to a target which is the instance in the SG data.
6. **Is any information missing form individual instances?** No.
7. **Are relationships between individual instances made explicit?** The instances are related because they come from the same detector just sampled at a different moment in time.
8. **Are there recommended data splits?** Yes, please see Section 2.l
9. **Are there any errors, sources of noise, or redundancies in the dataset?** There are no redundancies. Yes, there are many sources of detector noise.
10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources?** The data is self-contained.
11. **Does the dataset contain data that may be considered confidential?** No
12. **Does the dataset containd ata that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** No.

1.3 Collection process

1. **How was the data associated with each instance acquired?** The data associated with each instance is acquired by the ABRACADABRA detector. Full details can be viewed in Sections 1.1 and 2.
2. **What mechanisms or procedures were used to collect the data?** Full details can be viewed in Sections 1.1 and 2 and reference [28]. The hardware necessary for producing said data include, but are not limited to, an Oxford dilution refrigerator, 1T superconducting magnet, two-stage Magnicon SQUID, superconducting pickup loop, superconducting calibration loop, signal generator, digitizer, and data acquisition computer.

Table 3: Summary of critical information about this data release. These are the data files used for training and benchmarking of the baseline algorithms provided.

	Training Data	Validation Data	Science Data
File Name	abra_training _00{00-19}.h5	abra_validation _00{00-19}.h5	abra_science _0{000-207}.h5
No. Data Points per File	2.01e9	2.01e9	4.01e9
HDF5 File Size	2.2 GB	2.2 GB	2.7 GB
ch1 Hardware Input	SQUID	SQUID	SQUID
ch2 Hardware Input	SG	SG	
Injected frequencies (Hz)	[1100, 1200, ... , 4.8M, 4.9M]		
Injected amplitudes (mV)	50		

Table 4: Summary of auxiliary files in this data release. These files provide an interesting challenge for the user, however were not used in the training or validation of the baseline models.

	Aux Training Data	Aux Validation Data
File Name	abra_training_00{20-39}.h5	abra_validation_00{20-39}.h5
No. Data Points per File	2.01e9	2.01e9
HDF5 File Size	2.2 GB	2.2 GB
ch1 Hardware Input	SQUID	SQUID
ch2 Hardware Input	SG	SG
Injected Frequencies (Hz)	[1100, 1200, ... , 4.8M, 4.9M]	
Injected Amplitudes (mV)	10	

3. **If the dataset is sampled from a larger set, what was the sampling strategy?** The voltage produced by the SQUID and the SG are continuous. The instances are sampled from this continuous voltage stream at a constant rate of 10MS/s. The sampling strategy is deteministic and regular.
4. **Who was involved in the data collection process and how were they compensated?** To run the ABRACADABRA experiment, one graduate student, J. T. Fry, was needed. This graduate student was paid via NSF fellowship.
5. **Over what timeframe was the data collected?** The data were collected from 2/21/24 - 2/23/24.
6. **Were any ethical review processes conducted?** No. These data do not involve humans.
7. **Does this dataset relate to people?** No.

1.4 Preprocessing/cleaning/labeling

1. **Was any preprocessing/cleaning/labeling of the data done?** No.

1.5 Uses

1. **Has the dataset been used for any tasks already?** No, this dataset has not been used for any tasks yet.
2. **Is there a repository that links to any or all papers or systems that use the dataset?** No. This dataset has yet to be used outside of this paper.
3. **What other tasks could the dataset be used for?** As discussed in Section 5, these data can generally be used for training time series algorithms. Due to its high coherence and extensive length, it is perfect for cross cutting applications.
4. **Is there anything about the composition of the dataset or the way it was collected and preprocessed that might impact future use?** No.
5. **Are there tasks for which the dataset should not be used?** No.

1.6 Distribution

1. **Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?** Yes, the dataset is open to the public.
2. **How will the dataset be distriuted?** The dataset is publically available to be downloaded from the Open Science Data Federation cache. For download instructions, please see Section 2.
3. **When will the dataset be distributed?** The dataset is presently available.
4. **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** No.
5. **Have any third parties imposed IP-based or other restriction on the data associated with the instances?** No.
6. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** No.

1.7 Maintenance

1. **Who will be supporting/hosting/maintaining the dataset?** The dataset is hosted at Open Science Data Federation (OSDF). The data storage at OSDF was offered to Dr. Aobo Li by the director of San Diego Supercomputer Center (SDSC). OSDF also provide distributed cache of the dataset across its global cache location. For more detail, please refer to OSDF Website*. Dr. Aobo Li will be maintaining the dataset.
2. **How can the owner/curator/mangager of the dataset be contacted?** Please email Dr. Aobo Li at (liaobo77@ucsd.edu).
3. **Is there an erratum?** No.
4. **Will the dataset be updated?** No, the dataset will not be updated.
5. **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?** This dataset does not relate to people.
6. **Will older versions of the dataset continue to be supported?** Yes, they will continue to be supported.
7. **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** No. Access to the ABRACADABRA detector is controlled.

2 Dataset and code access

The data downloading and associated analysis scripts are available at `https://github.com/jessicafry/TIDMAD`. All data were uploaded to Open Science Data Federation and cached using their distributed cache system. The TIDMAD dataset can be downloaded using the `download_data.py` script provided in this GitHub repository. This script runs without any external dependencies. This script downloads data by generating a series of wget commands and executing them in a bash environment. `download_data.py` has the following argument:

- `-output_dir -o`: Destination directory where the file will be downloaded, default: current working directory.
- `-cache -c`: Which OSDF cache location should be used to download data. Options include [NY/NorCal/SoCal/Director(default)]:
  - NY: New York
  - NorCal: Sunnyvale
  - SoCal: San Diego
  - Director: automatically find the fastest cache location based on user's location.
    - **WARNING**: Director cache is sometimes unstable. We recommend switching to a different cache if the download fails.

*`https://osg-htc.org/services/osdf.html`

- `-train_files -t`: Number of training files to download, must be an integer between 0 and 20, default 20.
- `-validation_files -v`: Number of validation files to download, must be an integer between 0 and 20, default 20.
- `-science_files -s`: Number of science files to download, must be an integer between 0 and 208, default 208.
- `-f, -force`: Directly proceed to download without showing the file size and asking the confirmation question.
- `-sk, -skip_downloaded`: Skip the file that already exists at `-output_dir`.
- `-w, -weak`: Download the weak signal version of training and validation files. In this version, the injected signal is 1/5 the amplitude of the normal version. This is a more challenging denoising task. Note that the normal version has a file range 0000-0019, while the weak version has a file range of 0020-0039.
- `-p, -print`: Print out all wget commands instead of actually executing the download commands.

In the same github repository, we also provided a `filelist.dat` file which contains line-by-line wget command to download the entire dataset.An example wget command is given here:

```
wget https://osdf-director.osg-htc.org/ucsd/physics/ABRACADABRA/ABRA_aires_
validation_data/abra_validation_0009.h5
```

We have also provided a link to download all trained models used to produce the results in the main manuscript. The models can be downloaded from `https://drive.google.com/drive/folders/16ORX1b2zo1_lOYYAcRBgddBuYImj0Bxs?usp=share_link`.

3 Croissant metadata

We created a croissant metadata file `TIDMAD_croissant.json` using protocal presented in [4].

4 Author statement

The authors of this paper all bear responsibility in the case of violation of rights. The information provided in the paper and supplementary material is truthful and accurate. The code from this paper is hosted, managed, and maintained by the paper author J. T. Fry at `https://github.com/jessicafry/TIDMAD`. The data from this paper is hosted, managed, and maintained by the paper author Dr. Aobo Li with download instructions in Section 2. The dataset is released under the Creative Commons Attribution (CC BY) license. The code is released under the GNU General Public License (GPL), version 3.