Long Time Series Data Release from Broadband Axion Dark Matter Experiment

J. T. Fry
Department of Physics
Massachusetts Institute of Technology
Cambridge, MA 02139
jtfry@mit.edu

Aobo Li
Halıcıoğlu Data Science Institute
Department of Physics
University of California San Diego
La Jolla, CA 92093
liaobo77@ucsd.edu

Lindley Winslow
Department of Physics
Massachusetts Institute of Technology
Cambridge, MA 02139
lwinslow@mit.edu

Xinyi Hope Fu
Department of Physics
Massachusetts Institute of Technology
Cambridge, MA 02139
hopefu@mit.edu

Kaliroe M. W. Pappas
Department of Physics
Massachusetts Institute of Technology
Cambridge, MA 02139
kaliroe@mit.edu

Abstract

Axions are a promising dark matter candidate, yet their feeble interactions with visible matter pose a considerable challenge in detecting them. The ABRACADABRA-10cm experiment was meticulously built to generate long time series data within which axion signals could hide. Currently, the axion analysis of this dataset is only conducted in the frequency domain, omitting the valuable phase information embedded in the raw time series. In this public data release, we present a labeled dataset comprised of time series data collected from the ABRACADABRA detector, complete with axion-mimicking hardware signal injections. This dataset paper sets the stage for critical challenges faced in ABRACADABRA data analysis, including peak finding, denoising, and time series reconstructions. The success of machine learning algorithms in tackling these challenges will boost the experimental sensitivity to the enigmatic axion.

1 Introduction

Understanding the particle nature of dark matter is one of the most important questions faced by physics today. Despite compelling cosmological observations mandating that dark matter constitutes a significant 24% of the universe's energy content, no direct terrestrial laboratory detection has been achieved to date [1]. One proposed particle that can both serve as a dark matter candidate and simultaneously resolve the strong-CP problem of quantum chromodymanics is the axion [2, 3]. In the presence of a strong background magnetic field, axions interact with virtual photons via the axion-photon coupling to create real magnetic field that oscillates at a frequency corresponding to the axion mass. The **A** **B**roadband/**R**esonant **A**pproach to **C**osmic **A**xion **D**etection with an **A**mplifying **B**-field **R**ing **A**pparatus (ABRACADABRA) experiment searches for this axion-induced field through

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

a superconducting pickup structure at the center of a toroidal magnet. Subsequently, the signal is read out from the pickup structure with a SQUID current sensor [4, 5, 6].

Given this setup, the axion signal will appear as a narrow peak in the frequency spectrum at a frequency defined by the axion mass [7]. This extremely weak signal dwells above numerous sources of background noise, thus achieving lower noise levels is a persistent challenge for the ABRACADABRA experiment. The data acquisition system (DAQ) continuously reads out voltage of the SQUID as a function of time, i.e. time series data. In the traditional ABRACADABRA axion analysis, due to the large file size, we perform an online Fourier transform to convert the time series data into the frequency domain data. This process of online averaging over the real components of the Fourier transform will inevitably lose information. Raw time series data with phase information retention is necessary for other physics analyses which require us to detect time dependant changes in dark matter density as the earth rotates within a non-standard dark matter halo.

End-to-end machine learning algorithms have the potential to extract information from both time and frequency domains, thereby surpassing the performance of traditional analysis. In this work, we present a carefully-procured dataset with injected hardware signal mimicking axion behavior. Since the signal is injected at defined frequencies, the ground truth of this dataset is exactly known. Designing and benchmarking machine learning algorithms on this dataset will address critical challenges faced by ABRACADABRA and other axion dark matter experiments.

2 Dataset Description

In this work, we present a labeled dataset produced by the ABRACADABRA detector during calibration signal injection. Axion mimetic signal is injected into the system with a superconducting wire loop placed along the central axis of the toroidal magnet. Using a signal generator (SG), we drive a sinusoidal AC signal with varying frequency and amplitude into the calibration loop. This produces and axion-like oscillating field in the detector. By scanning across different frequencies (1100 Hz to 5 MHz) at various amplitudes (50 mV and 100 mV), we calibrate the end-to-end gain of our system in the axion mass range of interest. In doing so, we relate the detector response to the science quantity of interest — the axion to photon coupling.

This data is produced by our digitizer which saves the time series at rate of 10 MS/s. The two channels of the digitizer are connected to the SQUID output (ch1) and the SG directly (ch2). Every 13 seconds, the SG changes its injection frequency. For each signal injection, we save long time series with 130 MS, injected SG frequency, and injected SG signal amplitude. Within each time series file there are 20 such signal injections with data from either ch1 or ch2. The graphical representation of the data stored in the time series file can be seen in Figure 1. The raw output from the digitizer is stored as 8 bit integers. This integer can be converted into units of mV with a scaling factor of $40/128$.

Additionally, the traditional analysis relies on power spectral densities (PSD). To procure this dataset, we first perform fast Fourier transform over each second of data then calculate the power in each frequency bin normalized to a single hertz bandwidth, thus creating a PSD. Since each signal injection lasts 13 seconds, 13 PSDs are created for each injection. We discard the first 3 PSDs to limit cross talk with the previous injection, then average the last 10 PSDs to create the averaged PSD for each injection. The averaged PSD data is stored in two files, one corresponding to each channel input. A representation of this data can be see in Figure 1.

	Time Series Data	Averaged PSD Data
File Name	`abra_TS_{SQUID/SG}_{00-30}.h5`	`abra_PSD_{SQUID/SG}.h5`
No. Signals per File	20	618
No. Data Points per Signal	130M	50M
HDF5 File Size	3.1 GB/1.5 GB	24 GB
Hardware Input	SQUID/SG	SQUID/SG
Injected Frequencies (Hz)	[1100, 1200, ... , 4.8M, 4.9M]	
Injected Amplitudes (mV)	[50, 100]	
DOI	`https://dataplanet.ucsd.edu:443/citation?persistentId=perma:83.ucsddata/GSDXLP`	

Table 1: Summary of critical information about this data release.

This dataset will be stored at the San Diego Supercomputer Center (SDSC, `https://www.sdsc.edu`). The dataset will be published using the DataPlanet software framework, an initiative by the

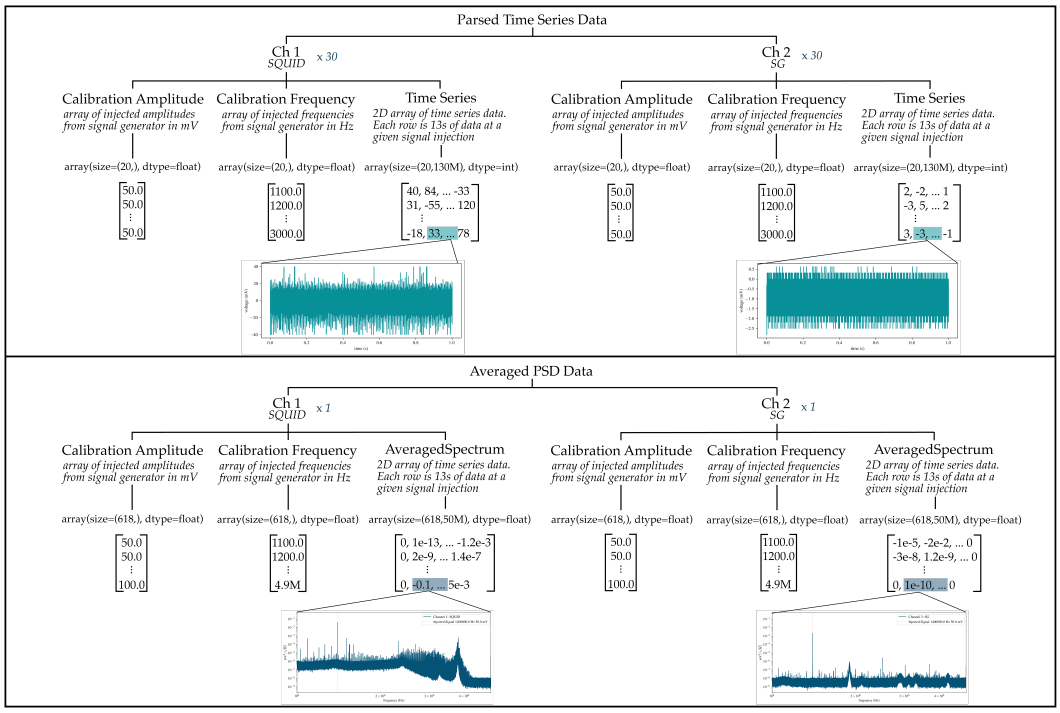

Figure 1: The top (bottom) trees graphically depict the data structure of the time series (power spectral density) data in this public data release. The time series arrays are in raw digitizer units (see Section 2 for scaling) while the PSD data is in units of power $[mV^2/\sqrt{Hz}]$.

UCSD Halıcıoğlu Data Science Institute (HDSI). To access the data, users should visit the DataPlanet website at `https://dataplanet.ucsd.edu`. In the list of dataverses, there is a dedicated Physics dataverse under which users will find the "ABRACADABRA Experiment" dataset. Alternatively, users could directly access the dataset page via the following link `https://dataplanet.ucsd.edu:443/citation?persistentId=perma:83.ucsddata/GSDXLP`.

Within the dataset page, users will find a downloading script `download_ABRACADABRA_data.py`. The entire dataset can be downloaded by running this script with the following command line:

```
python download_ABRACADABRA_data.py [destination_directory]
```

If no `[destination_directory]` is specified, the dataset will be downloaded into the current working directory. Please note that the overall dataset size is roughly 1 Tb.

Alternatively, individual files listed in Table 2 can be downloaded via the `wget` command:

```
wget https://osg-sunnyvale-stashcache.t2.ucsd.edu:8443/ucsd/physics/
    ABRACADABRA/[file_name]
```

`[file_name]` should be replaced with individual file the use would like to download. For a full list of released file names, please refer to Appendix A.

3 Traditional Analysis

In this section, we present the traditional analysis methods and results currently adopted by the ABRACADABRA experiment. The analysis was performed upon the average PSD dataset described in Section 2. The goal of the traditional calibration analysis is to determine the end-to-end gain of the ABRACADABRA apparatus over the entire frequency range. This calibration allows us to ascertain one of our principal physics quantities of interest, the maximum sensitivity to the axion to two photon coupling $g_{a\gamma\gamma}$.

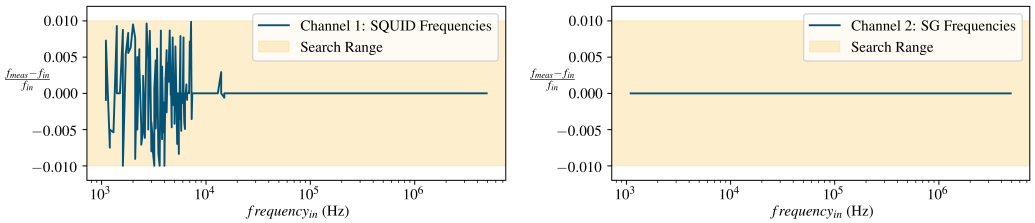

Figure 2: To the left (right) is the difference between the measured frequency and the injected frequency from the SQUID (SG) data. Highlighted is the frequency range in which the traditional algorithm searched.

The key step in the calibration analysis is a peak finding algorithm to identify an injected signal. Within a small window centered around the known injected signal frequency, a peak is defined as the frequency bin (f_{meas}) whose power is a factor of three larger than the bins ten steps above and ten steps below. While this algorithm is good at finding high frequency calibration peaks (Figure 2), the peak finding algorithm could be improved for low frequencies in which the signal is attenuated through a high pass filter making the power significantly smaller. Furthermore, this calibration analysis relies on truncating the PSD data to a narrow frequency bandwidth of interest. For the axion science data analysis, the axion mass, corresponding to signal frequency, is unknown; truncating the data around a narrow frequency bandwidth is therefore impossible. A much more computationally intensive technique is therefore used in the axion science data analysis, making calibration with this technique infeasible [4, 5, 8].

The full calibration analysis description and list of hardware can be found in Appendix B.

4 Scientific Challenges

Machine learning algorithms can be used to resolve critical challenges presented in Section 3 as well as enabling new avenues of physics search in ABRACADABRA. The traditional calibration analysis requires the injected signal's frequency as an input. When searching for axions, the dataset will contain no injected calibration signal, therefore the exact frequency of the axion peak will be unavailable. Moreover, since the axion frequency is correlated with axion mass — an important free parameter in the KSVZ/DFSZ theory, determination of this frequency is pivotal for making discovery claims and validating theories. Machine learning algorithms possess the capability to harness insights from both the time and frequency domains, enabling the detection of axion peaks without relying on prior information of the signal frequency. These algorithms can be trained using labeled calibration data generated through the signal injection process and subsequently applied to physics datasets without such injection. Furthermore, a valuable improvement in the calibration analysis lies in preserving the complete frequency information, obviating the need for frequency domain cuts to detect injected signals.

Machine learning algorithms could also be used to denoise the long time series dataset. The major source of noise in ch1 data comes from background electromagnetic fields created externally or internally when conductive components oscillate. Given that the ch2 data on the DAQ represents the true injected signal, it is free of such oscillation-induced noise. Training machine learning algorithms to reproduce the long time series in ch2 compared from ch1 input can shed light on persistent noise sources exclusively within the detector signal. This denoising model could then serve as a filter for scientific data where no signals are intentionally injected, effectively reducing noise interference. Furthermore, this denoising task could be executed in either the frequency domain or the time series data, providing flexibility in noise reduction approaches.

Lastly, machine learning algorithms could enable ABRACADABRA to search for physics in previously unavailable regime. This data release represents only one hour of data collection, yet requires an enormous amount of storage ($\sim$1 Tb). Furthermore, storing and analyzing such raw time series data for a multiyear physics run would result in an impractical $\sim$ 40 PB of data. Therefore ABRACADABRA only saves frequency domain data in traditional analysis. The recent trend of time-frequency analysis in machine learning community could be used to overcome this disadvantage.

The raw time series data could be converted to a time-frequency series format, where the 13-second raw time series is converted to 1-second raw time series (time domain) portion followed by 12-second average PSD (frequency domain) portion. This saving format allows us to suppress dataset size by at least a factor of 10. Because the transition between the frequency domain data and time series data collection is uninterrupted, time-frequency analysis algorithms could use the raw time series portion to reconstruct the phase information within the average PSD portion. Reconstructing phase information in ABRACADABRA will open up new avenues for new physics search such as non-standard halo dark matter models or axion interferrometry [9, 10].

5 Broader Impacts

Limitations One limitation of this work is the disproportionate calibration signal amplitude. For our traditional calibration peak finding algorithm to work, injected signals must be large. However, an expected axion signal amplitude in the detector would be orders of magnitude smaller due to dark matter's feeble interactions with the standard model. In the future, we would like to expand this dataset to include smaller injected signal amplitudes. Secondarily, our benchmark peak finding analysis comes from our calibration pipeline which truncates frequency information. For future benchmarking, we will run our science data analysis peak finding algorithm, which does not assume knowledge of signal frequency, on the dataset presented in this work. The third limitation is the parsing algorithm used to label the dataset. Due to hardware response delays, each injected frequency duration is not precisely 13 seconds. This causes drift in the parsing and labeling algorithm such that the first seconds of high injected frequency data are contaminated with signal from the frequency immediately prior. While solvable with the current dataset by omitting the first 3 seconds of every time series chunk, we plan on improving the parsing and labeling algorithm for future works.

Future Work and Application Applying machine learning techniques to improve the peak finding, denoising, and time series reconstruction of axion dark matter data can further the experimental reach of a host of axion dark matter experiments. With this public data release, we open the door for ML analyses in the interesting field of axion dark matter experiments. In the future we plan to release the time-frequency series dataset as we described near the end of Section 4. By mining the time domain of axion data, increased sensitivity can be reached without extensive hardware upgrades. The techniques developed on this dataset are broadly applicable to future experiment such as DMRadio and Haystac [11, 12].

Social Impact While there are many long time series datasets (e.g. stock prices, weather data), much of this data is relatively flat in the frequency domain. Our data release and subsequent models built on frequency rich long time series data, could fill in this data gap in the growing field of ML. The main negative social impact is the energy consumption to the environment of taking axion dark matter data, as well as storing and analyzing ~ 1 TB of data.

6 Acknowledgements

J. T. Fry is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2141064. 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 for the data storage space and open access infrastructure, as well as Arun Kumar at Halicioglu Data Science Institute who led the development of DataPlanet.

References

[1] M. Markevitch et al. "Direct Constraints on the Dark Matter Self-Interaction Cross Section from the Merging Galaxy Cluster 1E 0657–56". In: *The Astrophysical Journal* 606.2 (May 2004), p. 819. DOI: 10.1086/383178. URL: https://dx.doi.org/10.1086/383178.

[2] R. D. Peccei and Helen R. Quinn. "CP Conservation in the Presence of Instantons". In: *Phys. Rev. Lett.* 38 (1977), pp. 1440–1443. DOI: 10.1103/PhysRevLett.38.1440.

[3] F. Wilczek. "Problem of Strong P and T Invariance in the Presence of Instantons". In: *Phys. Rev. Lett.* 40 (5 Jan. 1978), pp. 279–282. DOI: 10.1103/PhysRevLett.40.279. URL: https://link.aps.org/doi/10.1103/PhysRevLett.40.279.

[4] Chiara P. Salemi et al. “Search for Low-Mass Axion Dark Matter with ABRACADABRA-10 cm”. In: *Phys. Rev. Lett.* 127 (8 Aug. 2021), p. 081801. DOI: 10.1103/PhysRevLett.127.081801. URL: https://link.aps.org/doi/10.1103/PhysRevLett.127.081801.

[5] Jonathan L. Ouellet et al. “Design and implementation of the ABRACADABRA-10 cm axion dark matter search”. In: *Physical Review D* 99.5 (Mar. 2019). DOI: 10.1103/physrevd.99.052012. URL: https://doi.org/10.1103%2Fphysrevd.99.052012.

[6] Jonathan L. Ouellet et al. “First Results from ABRACADABRA-10 cm: A Search for Sub-μeV Axion Dark Matter”. In: *Phys. Rev. Lett.* 122 (12 Mar. 2019), p. 121802. DOI: 10.1103/PhysRevLett.122.121802. URL: https://link.aps.org/doi/10.1103/PhysRevLett.122.121802.

[7] L.F. Abbott and P. Sikivie. “A cosmological bound on the invisible axion”. In: *Physics Letters B* 120.1 (1983), pp. 133–136. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)90638-X. URL: https://www.sciencedirect.com/science/article/pii/037026938390638X.

[8] Joshua W. Foster, Nicholas L. Rodd, and Benjamin R. Safdi. “Revealing the dark matter halo with axion direct detection”. In: *Phys. Rev. D* 97 (12 June 2018), p. 123006. DOI: 10.1103/PhysRevD.97.123006. URL: https://link.aps.org/doi/10.1103/PhysRevD.97.123006.

[9] N. Wyn Evans, Ciaran A. J. O’Hare, and Christopher McCabe. *SHM^{++}: A Refinement of the Standard Halo Model for Dark Matter Searches in Light of the Gaia Sausage*. 2018. arXiv: 1810.11468 [astro-ph.GA].

[10] Joshua W. Foster et al. “Dark matter interferometry”. In: *Physical Review D* 103.7 (Apr. 2021). DOI: 10.1103/physrevd.103.076018. URL: https://doi.org/10.1103%2Fphysrevd.103.076018.

[11] L. Brouwer et al. “Projected sensitivity of mml:math xmlns:mml="http://www.w3.org/1998/Math/MathML" display="inline"mml:mrowmml:msupmml:mrowmml:mtextDMRadio-m/mml:mtext/mml:mrowmml:mrowmml:mn3/mml:mn/mml:mrow/mml:msup/mml:mrow/mml:math : A search for the QCD axion below mml:math xmlns:mml="http://www.w3.org/1998/Math/MathML" display="inline"mml:mrowmml:mn1/mml:mnmml:mtext /mml:mtextmml:mtext /mml:mtextmml:mi mathvariant="normal"/mml:mimml:mieV/mml:mi/mml:mrow/mml:math”. In: *Physical Review D* 106.10 (Nov. 2022). DOI: 10.1103/physrevd.106.103008. URL: https://doi.org/10.1103%2Fphysrevd.106.103008.

[12] HAYSTAC Collaboration et al. *New Results from HAYSTAC’s Phase II Operation with a Squeezed State Receiver*. 2023. arXiv: 2301.09721 [hep-ex].

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **The paper's contribution and scope is discussed in the Abstract, Section 4**
   (b) Did you describe the limitations of your work? **The paper's limitations are discussed in Section 4**
   (c) Did you discuss any potential negative societal impacts of your work? **The paper's societal impacts are discussed in Section 4**
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Our work conforms to the ethics review guidelines.**
2. If you are including theoretical results... **This paper does not contain any theoretical results.**
   (a) Did you state the full set of assumptions of all theoretical results?
   (b) Did you include complete proofs of all theoretical results?
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **Yes, we include the data necessary to reproduce the main experimental results. The code necessary for reproducing the benchmark result is proprietary and not included. However, the methodology of the code that produces the benchmark results is described in detail in Section 3.**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **The data and data splits are described in detail in Section 2**
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **Systematic errors are estimated further downstream in the classical benchmark analysis.**
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **This is reported in Section 4**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators?
   (b) Did you mention the license of the assets? **The dataset is self-taken at our local axion dark matter detector.**
   (c) Did you include any new assets either in the supplemental material or as a URL? **The dataset is provided in Section 2.**
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **The data come from an axion dark matter experiment. No human-related or human-created content is contained in the dataset.**
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **The data come from an axion dark matter experiment. No human-related content or identifiable information is contained in the dataset.**
5. If you used crowdsourcing or conducted research with human subjects... **All data was taken from a local axion dark matter detector. We did not use crowdsourcing or human subjects.**
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable?
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable?
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?

A Individual File List

Time Series Data	
abra_TS_SQUID_00.h5	abra_TS_SG_00.h5
abra_TS_SQUID_01.h5	abra_TS_SG_01.h5
abra_TS_SQUID_02.h5	abra_TS_SG_02.h5
abra_TS_SQUID_03.h5	abra_TS_SG_03.h5
abra_TS_SQUID_04.h5	abra_TS_SG_04.h5
abra_TS_SQUID_05.h5	abra_TS_SG_05.h5
abra_TS_SQUID_06.h5	abra_TS_SG_06.h5
abra_TS_SQUID_07.h5	abra_TS_SG_07.h5
abra_TS_SQUID_08.h5	abra_TS_SG_08.h5
abra_TS_SQUID_09.h5	abra_TS_SG_09.h5
abra_TS_SQUID_10.h5	abra_TS_SG_10.h5
abra_TS_SQUID_11.h5	abra_TS_SG_11.h5
abra_TS_SQUID_12.h5	abra_TS_SG_12.h5
abra_TS_SQUID_13.h5	abra_TS_SG_13.h5
abra_TS_SQUID_14.h5	abra_TS_SG_14.h5
abra_TS_SQUID_15.h5	abra_TS_SG_15.h5
abra_TS_SQUID_16.h5	abra_TS_SG_16.h5
abra_TS_SQUID_17.h5	abra_TS_SG_17.h5
abra_TS_SQUID_18.h5	abra_TS_SG_18.h5
abra_TS_SQUID_19.h5	abra_TS_SG_19.h5
abra_TS_SQUID_20.h5	abra_TS_SG_20.h5
abra_TS_SQUID_21.h5	abra_TS_SG_21.h5
abra_TS_SQUID_22.h5	abra_TS_SG_22.h5
abra_TS_SQUID_23.h5	abra_TS_SG_23.h5
abra_TS_SQUID_24.h5	abra_TS_SG_24.h5
abra_TS_SQUID_25.h5	abra_TS_SG_25.h5
abra_TS_SQUID_26.h5	abra_TS_SG_26.h5
abra_TS_SQUID_27.h5	abra_TS_SG_27.h5
abra_TS_SQUID_28.h5	abra_TS_SG_28.h5
abra_TS_SQUID_29.h5	abra_TS_SG_29.h5
abra_TS_SQUID_30.h5	abra_TS_SG_30.h5

Averaged PSD Data	
abra_PSD_SQUID.h5	abra_PSD_SG.h5

B Traditional Calibration

In the traditional ABRA analysis, the hardware signal injection is used to calibrate the detector system. By comparing the injected signal and the detected signal, the end-to-end gain of the system is determined.

The calibration analysis procedure is detailed below:

1. Each averaged PSD file is truncated, omitting frequencies outside of a 2000 Hz window centered at the known injected frequency.
2. Within this frequency window, a peak finding algorithm is used to identify the signal frequency range.
3. A cumulative distribution function is created over the entire frequency range to calculate the total power injected.
4. The signal power is then compared to the expected power given the injected signal amplitude for each entry in the file, resulting in an end-to-end gain, V_{DAQ}/V_{SG} in units of mV/mV for every injected signal frequency and amplitude.

The same analysis was repeated twice over the ch1 SQUID readout and ch2 SG readout to compare results. In ideal cases, the reproduced gain spectrum should remniscent of a flat line. As shown in Figure 3, the analysis result of SG readout represents the ground-truth information which is roughly flat, while the ch1 SQUID readout exhibits certain nontrivial shape. The shape in ch1 signal is a consequence of hardware noise as well

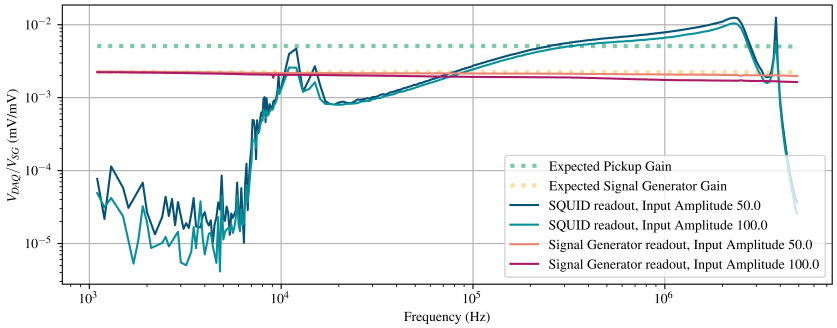

Figure 3: The gain shown here is the result of the traditional axion mimetic injection analysis. The gain is defined as the output voltage detected at the digitizer over a corresponding input voltage from the signal generator. This includes both expected and measured gains for the SQUID readout and the direct signal generator readout.

as uncertainties in the peak finding algorithms. Both of these issues can potentially be addressed by machine learning algorithms, as we discuss in Section 4.

The expected gain is calculated as a sequence of transfer functions taking into account the attenuators (50 dB for each channel), impedence mismatch, SQUID flux conversion, pickup to calibration loop mutual inductance, and splitters. The hardware filters for ch1 include a 10kHz high pass filter and a 5MHz low pass filter to decrease noise. Ch2 frequencies are unfiltered. Besides the pickup to calibration loop mutual inductance which is simulated with COMSOL multiphysics, the transfer factors are analytically calculated.