PROCEEDINGS OF SCIENCE

# Real-time Likelihood Map Generation to Localize Short-duration Gamma-ray Transients

PoS(ICRC2025)587

**Jeremy Buhler**[a,*] **and Marion Sudvarg**[b]

[a] *Department of Computer Science and Engineering, Washington University*
*One Brookings Drive, St. Louis, MO, USA*

[b] *Department of Physics, Washington University*
*One Brookings Drive, St. Louis, MO, USA*

*E-mail:* jbuhler@wustl.edu, msudvarg@wustl.edu

High-energy transient astrophysical phenomena, such as supernovae and binary neutron star mergers, benefit from a multi-wavelength investigation in which a space- or balloon-based omnidirectional telescope detects and localizes early high-energy emissions (such as a gamma-ray burst), then alerts a narrow-field follow-up instrument to observe the source. The high-energy telescope must provide a map that assigns to each sky location a likelihood that the source appears there. To issue prompt alerts despite limits on communication bandwidth and latency, it is desirable to compute this map aboard the high-energy telescope, but doing so requires rapid response while computing under stringent size, weight, and power constraints.

This work describes a real-time likelihood mapping implementation for Compton telescopes that is suitable for on-board computation. We use an adaptive multi-resolution approach and exploit parallelism and data reduction opportunities to achieve sub-second construction of high-resolution maps (HEALPix $N_{\text{side}}$=64) using a detailed instrument response matrix on a low-power ($< 10$ W) embedded computing platform. We validate the speed and accuracy of our mapping approach on simulated high-energy transients from the third COSI Data Challenge.

39th International Cosmic Ray Conference (ICRC2025)
15–24 July 2025
Geneva, Switzerland

ICRC 2025
The Astroparticle Physics Conference
Geneva July 15-24, 2025

[*] Speaker

## 1. Introduction

Investigations of high-energy transient astrophysical phenomena, such as binary neutron star mergers and supernovae, can benefit from coordination between instruments with different imaging modalities. For gamma-ray bursts (GRBs), a high-energy telescope with a wide field of view (FoV) detects prompt emissions from a source and then communicates its location to a narrow-FoV partner instrument, such as an optical telescope, for follow-up observation. In this cooperative setting, the high-energy telescope must provide source localization that is both *prompt* and *informative* in order to maximize the chance that the follow-up instrument can observe as much of the transient event's light curve as possible. The standard of informative localization is a *likelihood map*, like that shown in Figure 1, that rates the likelihood that the source appears at each possible location in the sky. Prompt map generation is important not only for sources that fade quickly (within 1–2 seconds) but also for longer-lived sources; for the latter, the high-energy telescope can provide a series of increasingly accurate maps as it accumulates observations, so that partner instruments can begin to move toward a source right away and can refine their search strategy as the map improves.

Existing Compton telescopes, such as Fermi GBM and the planned COSI mission, send information about detected gamma rays to the ground, where terrestrial computation can map a transient's location within a few seconds [1]. But this arrangement imposes constraints that increase the time until a partner instrument can respond to an observed transient. Limited bandwidth to ground restricts the number of observations available for prompt construction of a source's likelihood map, no matter how bright it is; moreover, moving data to the ground incurs latency before the map can be computed. Instruments in orbits far from earth, such as the Advanced Particle-astrophysics Telescope (APT) [2] proposed for deployment at the Sun-Earth $L_2$ Lagrange orbit, may need to coordinate with space-based partner instruments, but limited bandwidth and light-speed delays would add tens of seconds to the time needed to produce a map with terrestrial computation. These issues can be alleviated by computing the map aboard the high-energy telescope itself.

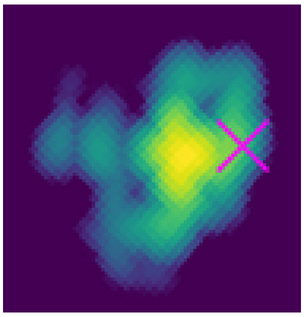

**Figure 1:** Detail of likelihood map for GRB GBM140329295 at HEALPix $N_{\text{side}}$=64 resolution, showing 90% containment region. Each HEALPix pixel covers ~0.84 deg$^2$. '×' denotes true source location.

While rapid, detailed source likelihood mapping has been demonstrated for, e.g., LIGO [3] using terrestrial computation, prior work has not yet constructed such maps for gamma-ray transients in real time under the stringent size, weight, and power constraints of on-board computation in space. To address this gap, we have developed a real-time, low-resource implementation of likelihood mapping for Compton telescopes. It uses a detailed instrument response model to construct high-resolution maps (HEALPix $N_{\text{side}}$=64, sub-degree scale) in under one second on an embedded processor using less than 10 W of power. Using simulated data from the third COSI Data Challenge [4], we demonstrate that, given COSI's current detector model, our mapping algorithm can localize gamma-ray transients to within a few degrees using as few as 150 Compton events from the source, even in the presence of background radiation. Our results suggest that real-time, on-board source mapping is feasible for cooperative imaging of astrophysical transients.

## 2. Problem Background

A high-energy telescope continuously observes signals from incoming particles. In the absence of transient phenomena, these particles arise from diffuse *background radiation*. When a transient occurs, the telescope observes additional gamma rays from a point source in direction $\vec{s}$ for a length of time $\Delta t$. The goal is to localize the source, that is, to infer the direction $\vec{s}$. Because this inference cannot be made with certainty, we instead construct a *likelihood map* that assigns to each possible $\vec{s}$ a (log-)likelihood ratio between two hypotheses: the null hypothesis $H_B$ that observations arise from the background alone, and the alternative hypothesis $H_s$ that they arise from both the background and a source at $\vec{s}$. The source's brightness and its *energy spectrum*, i.e., the distribution of its photon's energies, must be inferred as part of estimating the map.

Incoming gamma-ray photons interact with the detector by Compton scattering, producing *Compton rings* that constrain the source direction. Other types of background particles can also produce detector signals that are interpreted (incorrectly) as Compton events, giving rise to spurious rings. A Compton ring has a center vector $\vec{c}$ and an angular radius $\phi$ and is paired with a measurement $E_m$ of the incident photon's energy. These parameters $(E_m, \phi, \vec{c})$ define the *Compton data space* (CDS) of possible rings. Each observation used to calculate the map is a point in the CDS together with the time at which it is detected; let $D$ be the set of all such observations for one transient.

**Mapping Approach.** Our mapping algorithm broadly follows that of the `cosipy` library [5], which combines the maximum-likelihood approach of [6] to point-source localization with discretized sky-map construction in the spirit of [7]. In what follows, we assume that the burst's energy spectrum and the (constant) rate $\rho_B$ at which Compton rings arise from background radiation are known; we will address these limitations in the next section.

The probability models for source and background Compton ring generation are respectively an *instrument response* $R(\vec{s}, E_i)$ and a *background model* $B$. Each model discretizes the CDS into voxels and, for each CDS voxel, gives the fraction of observed rings from the process expected to lie within that voxel. Discretization of source directions $\vec{s}$ and center vectors $\vec{c}$ in the CDS uses the HEALPix [8] equal-area tiling of the sphere. $R$ and $B$ are derived empirically from simulations that capture biases in observed ring parameters due to the telescope's geometry (e.g., whether shielding blocks particles from some directions), detector electronics, and analysis software, as well as biases inherent to the physics of Compton scattering. Given an energy spectrum for the source, we reduce $R$ to its spectrum-weighted average $\overline{R}(\vec{s})$ over $R$'s $E_i$ dimension, since a set of events sampled from the spectrum will exhibit aggregate CDS voxel frequencies that follow $\overline{R}$.

For the null hypothesis $H_B$ that $D$ arises from the background alone, we assume that rings in voxel $v$ of the CDS are produced by a Poisson process of intensity $\lambda_B[v] = B[v] \cdot \rho_B \Delta t$. For the alternate hypothesis $H_s$ that $D$ arises partly from a source in direction $\vec{s}$, we assume that rings in voxel $v$ are produced by a Poisson process of intensity $\lambda_s[v] = \overline{R}(\vec{s})[v] \cdot \rho_s \Delta t + \lambda_B[v]$, where $\rho_s$ is the intensity of the source. Let $n_D[v]$ be the number of observed rings for voxel $v$, and let $\Pr_{pois(\lambda)}[k]$ be the probability that a Poisson random variable with mean $\lambda$ is equal to $k$. Then the desired likelihood ratio is given by

$$\frac{L(H_s \mid D)}{L(H_B \mid D)} = \prod_v \frac{\Pr_{pois(\lambda_s[v])}[n_D[v]]}{\Pr_{pois(\lambda_B[v])}[n_D[v]]}.$$

We do not know $\rho_s$ *a priori*, so for each source direction $\vec{s}$, we choose $\rho_s$ to maximize $L(H_s \mid D)$. We follow `cosipy`'s iterative likelihood maximization approach based on Newton's method.

While the source direction $\vec{s}$ is fixed in the galactic reference frame, the telescope's orientation changes with time, so that the source appears to move in its frame of reference (which is also the frame of the response $R$). For this reason, we use a record of the telescope's orientation over time (sampled once per second) to convert $\vec{s}$ to a (discretized) path $\pi = \{(\vec{s}_1, \tau_1), (\vec{s}_2, \tau_2), \ldots\}$ in its frame; the source appears at $\vec{s}_i$ in the telescope's frame for a duration $\tau_i$. We then replace the response $\overline{R}(\vec{s})$ in $\lambda_s[v]$ with $R^*(\vec{s}) = \sum_i \tau_i \overline{R}(\vec{s}_i)$. Averaging the response over the path is acceptable if orientation undergoes little change during a transient; future work will compute and sum separate likelihoods for events occurring in each segment of the path.

## 3. Real-Time Mapping Implementation

Our mapping computation seeks to construct high-resolution likelihood maps while limiting computational cost. This cost includes ① inferring missing parameters (background intensity, spectrum) needed by the likelihood model; ② computing the spectrum-averaged instrument response $\overline{R}$; and, for each candidate source direction $\vec{s}$, ③ converting $\vec{s}$ to a path in the instrument frame and ④ computing its likelihood ratio. Below, we describe our approach to ① and optimizations targeting ② and ④. We implemented map construction in Python for ease of prototyping and reuse of `cosipy` functions but minimized Python interpreter overhead by implementing performance-critical computations using Numpy vectorization and, in some cases, the Numba JIT compiler.

Throughout, we assume that the instrument response $R$ already resides in DRAM. We used a response published as part of the third COSI Data Challenge [4], which discretizes $\vec{s}$ and $\vec{c}$ into 768 bins (HEALPix $N_{\text{side}}$=8), $E_i$ and $E_m$ into 10 bins each, and $\phi$ into 30 bins, occupying ~6.6 GiB of memory. For purposes of mapping, the 3-dimensional array of CDS voxels can be flattened into a linear array of size $10 \times 30 \times 768$.

**Inference of Missing Parameters.** We estimate the background intensity $\rho_B$ during a transient as the average intensity in the prior 600 seconds. We reduce this interval if it overlaps a qualitative change in event rate, such as passage through the South Atlantic Anomaly. In actual deployment, a running estimate of $\rho_B$ would be maintained continually and so would not contribute to the time needed to generate a map when a transient occurs. We estimate the source's spectrum as a histogram of the measured energies $E_m$ of the events $D$, discretized using the same energy bins used for $E_i$ in $R$. This does not correct for the background's spectrum or for discrepancies between true $E_i$ and measured $E_m$. However, we find empirically that our maps are insensitive to small spectral changes.

**Size Reduction of Instrument Response.** Reducing $R$ to its spectrum average $\overline{R}$, extracting and summing slices of $\overline{R}$ for the path corresponding to $\vec{s}$, and computing the likelihood ratio for $\vec{s}$ all require time proportional to the number of voxels in $R$'s discretized CDS. Although a moderate-intensity transient may produce several thousand Compton events in $D$, the discretization used by $R$ in this work creates $\sim 2.3 \times 10^5$ CDS voxels. Hence, for such a transient, the majority of voxels $v$ in $R$ have $n_D[v] = 0$. The Poisson intensities for these "empty" voxels are not used in computing $L(H_s \mid D)$, except as part of the aggregate intensity $\sum_v \overline{R}(\vec{s})[v]$. Hence, after computing this aggregate once for each $\vec{s}$, we may eliminate empty voxels entirely from the computation.

To do so, we first compute a sparse list of voxels occupied by at least one event in $D$. We then discard all entries of $R$ and $B$ corresponding to CDS voxels not on this list. Recall that $B$, as well as each CDS-shaped subarray $R(\vec{s}, E_i)$, can be expressed as a linear array of voxels. Hence, we may convert the sparsified CDS back to a (smaller) dense linear array by consecutively renumbering the surviving voxels and applying the same renumbering to the voxel numbers for each discretized event in $D$. This transformation makes the problem size proportional to the number of non-empty voxels while still using dense rather than sparse array arithmetic for the mapping computation.

**Multi-resolution Map Refinement.** We are interested mainly in the regions of a map most likely to contain the source. If these regions can be identified in a low-resolution map, then higher-resolution mapping can be limited to them. We therefore implemented an adaptive multi-resolution approach, similar to that of [9, 10], to reduce the cost of mapping. We first produce a map of the detector's entire field of view at low resolution (HEALPix $N_{\text{side}}$=4). We then compute the 90% containment region [11] for the source relative to the most likely map pixel. Any pixel that is not either in this region *or* immediately adjacent to it is discarded. We include adjacent pixels based on empirical observation that, even at containments $\gg$ 90%, the source's true location was often adjacent to the computed containment region rather than within it. Pixels that are not discarded are subdivided into their four nested subpixels at the next larger HEALPix $N_{\text{side}}$, and likelihoods are recomputed only for these subpixels. We iterate this process of subdivision up to the final map resolution.

**Parallel Computation.** Mapping can be parallelized across multiple CPUs. At a high level, likelihoods for individual source directions $\vec{s}$ can be computed independently. At a lower level, linear-algebraic operations and iteration over CDS voxels can be parallelized. Our implementation eschews high-level parallelism over $\vec{s}$ in favor of the lower level for two reasons. First, high-level parallelism in Python requires spawning multiple processes, which both adds overhead and complicates sharing common data such as the response $R$ between processes. Second, our multi-resolution approach greatly reduces the number of directions $\vec{s}$ to be computed, to the point that not enough parallelism is available to amortize multi-process overhead.

We exploit parallelism in mapping in two ways. First, trimming the response $R$ to just its non-empty voxels and averaging it over the spectrum is done in parallel for each value of $\vec{s}$. Second, for each $\vec{s}$, the log-likelihood ratio is computed in parallel over CDS voxels and then accumulated. In particular, we parallelize the step-size computation in the Newton's method iteration used to fit the intensity $\rho_s$. We parallelize these computations across threads using Numba's `prange` construct.

## 4. Validation

We tested our mapping implementation on simulated transients from the third COSI Data Challenge [4]. COSI is a low-earth orbit gamma-ray observatory to be launched in 2027. The Data Challenge includes an instrument response matrix $R$ discretized as described in Section 3, three months of Compton events (real and spurious) from simulated background radiation, and a set of simulated transients, including 11 GRBs and 6 magnetar giant flares (MGFs) with locations, light curves, and spectra taken from published sources. We adopted the discretized CDS defined by $R$ and used the full three-month background simulation to estimate the model $B$.

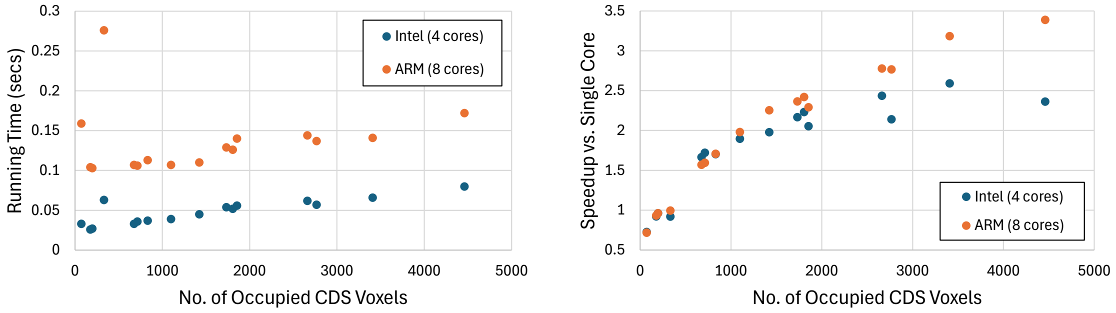

**Figure 2:** Running times and speedups of mapping at $N_{\text{side}}$=64 on 16 GRB and MGF transients.

**Computational Cost.** We assessed the cost of mapping the 17 simulated transients. For each, we chose a time interval that covered a large majority of its Compton events, starting with the first second in which the event rate significantly ($p < 0.1$) exceeded the background rate (when a detector might be expected to trigger) and continuing until the first second after peak intensity at which the event rate fell below this significance threshold. We measured the time to produce a map with $N_{\text{side}}$=64 resolution ($\sim 0.84$ deg$^2$/pixel), starting from the (un-discretized) event list $D$.

We tested two machine configurations. The first is expected to fly on the planned ADAPT balloon-borne gamma-ray telescope [12]; we used 4 of the performance cores on its Intel Core i7-13700TE CPU. The second machine is a small embedded system (NVidia Jetson Orin NX) with an 8-core ARM Cortex-A78AE CPU, which better matches the capacity of devices that could be deployed for orbital missions. It consumes less than 10 W when fully occupied with the mapping computation. (We did not use the Jetson's GPU because it did not offer a speed advantage over the ARM cores for the required linear algebraic operations.)

Figure 2 shows mapping times and multi-core speedups for 16 of the 17 transients vs. the number of *distinct CDS voxels* occupied by the Compton events (including background) for each. Mapping required under 0.3 s on the ARM and under 0.1 s on the Intel CPU in all cases, and the cost overall scaled linearly with the number of voxels to be processed. For all these transients, the number of occupied voxels is at most a few percent of the $> 200,000$ bins defined by the response, which makes our size reduction optimization effective. The figure does not include the bright GRB GBM090424592, which had over 100,000 Compton events occupying over 42,000 CDS bins; however, even this burst was mapped in 0.51 s on the ARM and 0.20 s on the Intel CPU, matching the observed linear trend. Scaling of the mapping computation with more CPU cores was limited for most transients by the small overall number of non-empty CDS voxels. For the bright GBM09042592, speedup reached 3.92× on the 8 cores of the ARM and 2.66× on 4 cores of the Intel, which is likely limited by non-parallelized parts of the computation.

Our multi-resolution approach was crucial for performance. A complete $N_{\text{side}}$=64 map requires calculating likelihoods for 49,152 sky pixels, which for the 17 DC3 transients averaged 2.8 s on Intel and 12.2 s on ARM. In contrast, our implementation computed fewer than 1500 pixels for all transients, and usually fewer than 500, while reproducing the 90% containment region of the map. Overall, our implementation generated high-resolution maps within a fraction of a second of receiving the input events for transients with a wide range of brightness, even on resource-constrained hardware.

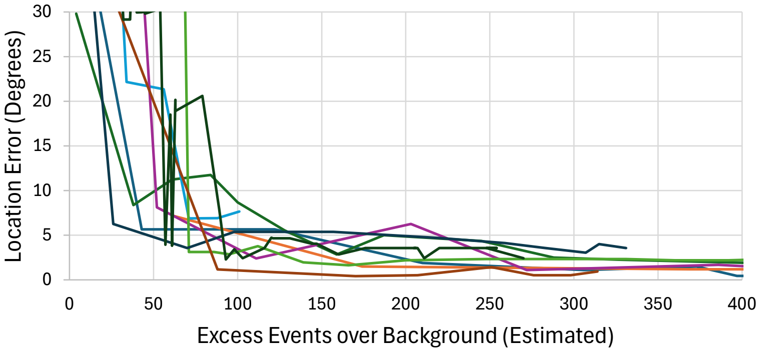

**Figure 3:** Accuracy as a function of excess events over background for ten long-duration transients.

**Localization Accuracy.** We next evaluated how accurately our likelihood maps localized each transient. We measured accuracy as the angular distance between the true source location and the center of the HEALPix pixel in the map with highest log-likelihood ratio. We emphasize that this experiment assesses how well our real-time mapping implementation performed with the given data; we make no claim about the resolution of the COSI instrument itself.

For these tests, we assumed the transient to start at the first one-second interval with significant excess events over background, as described above. We divided the 17 transients into two groups: *short* transients, with no significant excess after at most 2 (and usually after $< 1$) seconds, and *long* transients, which exhibited such excess at longer times (6–46 s) after the first significant second.

For short transients, which included all six MGFs and one GRB, our goal was to generate a map as soon as possible after the end of the burst. We stopped collecting Compton events for a transient after the first $^1/_{10}$ of a second in which the excess over background was not significant at $p < 0.1$. For these transients, the resulting maps had a mean accuracy of 2.28 degrees, with the maximum error being 3.99 degrees.

For long transients, real-time mapping must balance quick response with map accuracy. Earlier, less accurate maps can guide follow-up observations to the correct area of the sky, while later, more accurate maps refine the area to be searched. For each long transient, we therefore computed a new map each second after the first one during which we observed significant excess, using all events observed up to that point, and evaluated how these maps' accuracy improved as more events accumulated over time. We estimated the number of *excess* events seen due to the transient each second by subtracting the expected number of background events from the total event count so far.

Figure 3 shows how mapping accuracy improves as more events from the transient accumulate. After 150 excess events, accuracy is typically within 5 degrees and may continue to improve with additional events. Final accuracy by the end of the transient averaged 1.65 degrees for transients with at least 150 excess events. One dim GRB, GBM140329295, yielded only around 100 excess events overall and was therefore localized only to within 7.65 degrees. The fraction of excess vs. background events had little impact on accuracy, even when background events were five-fold more frequent than events from the transient.

## 5. Conclusion and Future Work

Cooperation between telescopes with different imaging modalities to study high-energy transients requires rapid, accurate localization. We have demonstrated high-resolution likelihood mapping for gamma-ray transients in a fraction of a second, using low-power computing resources in the range of what could be flown aboard an atmospheric or orbital mission. Our work suggests that future missions targeting these phenomena can utilize on-board computation to reduce alert latency, incorporate more observations into prompt analyses without the limits of communication to the ground, and automate coordination with partner instruments.

Because our current implementation can produce a map in 100–300 ms, future work will focus on ways to improve the accuracy of our maps while retaining sub-second latency. For the current simulated COSI data, our spectral and background estimates do not limit accuracy, which does not improve if these estimates are replaced with ground-truth values. More likely sources of systematic error in our method include not accounting for polarization of a transient's gamma rays, as well as the limited resolution of DC3's instrument response matrix. Addressing either of these limitations requires either a larger discretized response, which would challenge an onboard computer's DRAM capacity, or a different, more compact representation of the response. For example, machine learning could produce a functional model approximating a high-resolution response in much less space. To use such a model while retaining efficiency, we would employ the (currently unused) GPU capacity of our Jetson computing platform.

## References

[1] A. Goldstein, C. Fletcher, P. Veres, M.S. Briggs, W.H. Cleveland, M.H. Gibby et al., *Evaluation of automated Fermi GBM localizations of gamma-ray bursts*, The Astrophysical Journal **895** (2020) 40.

[2] J. Buckley, Adapt, S. Alnussirat, C. Altomare, R.G. Bose, D.L. Braun et al., *The Advanced Particle-astrophysics Telescope (APT) Project Status*, in *37th International Cosmic Ray Conference*, p. 655, Mar., 2022.

[3] L.P. Singer and L.R. Price, *Rapid bayesian position reconstruction for gravitational-wave transients*, Phys. Rev. D **93** (2016) 024013.

[4] Compton Spectrometer and Imager (COSI) Collaboration, *COSI Data Challenges*, Apr., 2025. 10.5281/zenodo.15126188.

[5] I. Martinez et al., *The cosipy library: COSI's high-level analysis software*, in *Proc. of 38th Int'l Cosmic Ray Conf.*, vol. 444, pp. 858:1–858:8, July, 2023.

[6] A. Pollock, G. Bignami, W. Hermsen, G. Kanbach, G. Lichti, J. Masnou et al., *Search for gamma-radiation from extragalactic objects using a likelihood method*, *Astronomy and Astrophysics* **94** (1981) 116.

[7] T. Herbert, *Estimating the sky map in gamma-ray astronomy with a compton telescope*, IEEE Transactions on Nuclear Science **38** (1991) 563.

[8] K.M. Gorski et al., *HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere*, *The Astrophysical Journal* **622** (2005) 759.

[9] P. Fernique, M. Allen, T. Boch, A. Oberto, F. Pineau, D. Durand et al., *Hierarchical progressive surveys: Multi-resolution HEALPix data structures for astronomical images, catalogues, and 3-dimensional data cubes*, *Astronomy & Astrophysics* **578** (2015) A114.

[10] L.P. Singer and L.R. Price, *Rapid Bayesian position reconstruction for gravitational-wave transients*, *Phys. Rev. D* **93** (2016) 024013.

[11] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, *Annals of Mathematical Statistics* **9** (1938) 60.

[12] W. Chen, J. Buckley et al., *Simulation of the instrument performance of the Antarctic Demonstrator for the Advanced Particle-astrophysics Telescope in the presence of the MeV background*, in *Proc. of 38th Int'l Cosmic Ray Conf.*, vol. 444, pp. 841:1–841:9, July, 2023.

## Acknowledgments

We wish to acknowledge our colleagues in the APT collaboration (https://adapt.physics.wustl.edu/) and to thank the COSI Science Team for assistance with `cosipy` and the DC3 data sets. This work was supported by NASA award 80NSSC21K1741.