Data Challenge 1 README

October 2022

General introduction

Welcome to the first COSI Data Challenge! This is the first of many COSI Data Challenges to be released on a yearly basis in preparation for the launch of the COSI Small Explorer mission (Tomsick et al. 2019) in 2027. The main goals of the COSI Data Challenges are to facilitate the development of the COSI data pipeline and analysis tools, and to provide resources to the astrophysics community to become familiar with COSI data. This first COSI Data Challenge was funded through NASA's Astrophysics Research and Analysis (APRA) for the release of high-level analysis tools for the COSI Balloon instrument (Kierans et al. 2017) , and thus the COSI Balloon model and flight data will be the focus this year. Future COSI Data Challenges will be released for the SMEX mission with increasingly more sophisticated tools and a larger range of astrophysical models and simulated sources each year. By the time we're ready to fly COSI, we will have simulated all of the main science objectives, developed the tools required to analyze each case, and have educated a broader community to perform the analysis.

Download and Install

Already have COSItools installed:

If you already have the COSItools installed on your computer, type *cosi* to navigate to the COSItools directory and activate the cosi python environment. Clone the cosi-data-challenge-1 repository.

New to COSItools:

Head to the feature/initialsetup branch of the cosi-setup Git repository and follow the readme guide: https://github.com/cositools/cosi-setup/tree/feature/initialsetup.

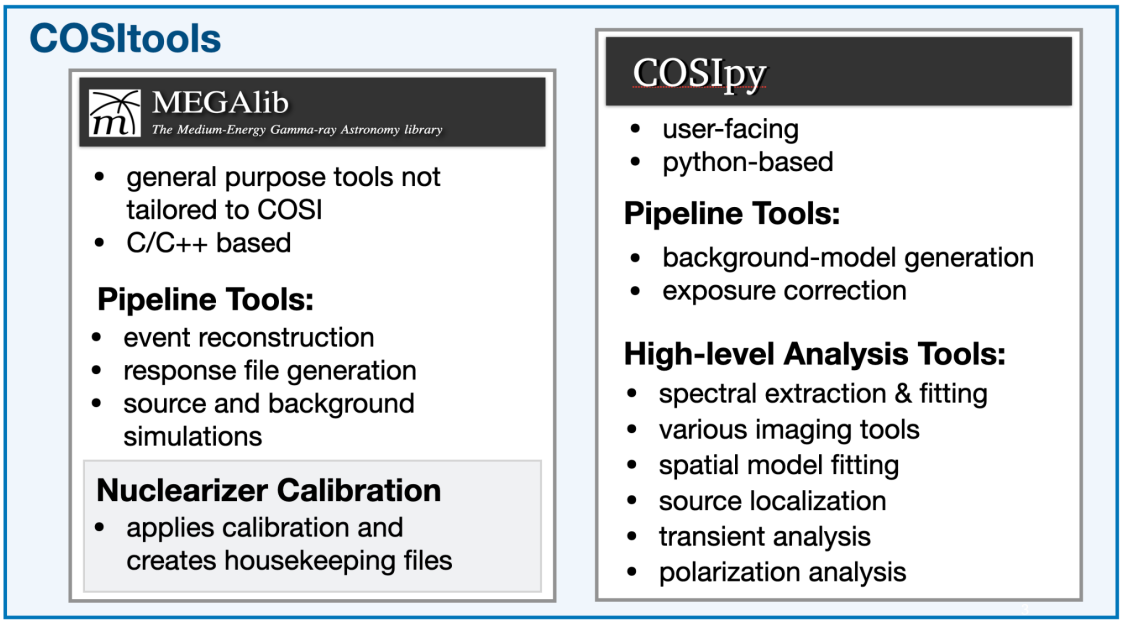
Using LFS to get the data:

Some of the data products are large, and we are using the Git Large File Server. You will need to first install git-lfs:
https://docs.github.com/en/repositories/working-with-files/managing-large-files/installing-git-large-file-storage

After Git LFS has been successfully installed, navigate to the cosi-data-challenge-1 directory and *git lfs pull* to download all of the data products. If this step is not performed, the file names will appear in the data_products directory, but they will only be placeholders and the size is only a few hundred kB. The total size of the data products directory should be 4.6 GB.

Getting Started

The COSI pipeline tools, COSItools, are divided into two programs (see figure below): MEGAlib and *cosipy*. MEGAlib (https://megalibtoolkit.com and https://github.com/zoglauer/megalib) is the Medium Energy Gamma-ray Astronomy Library, which is a general purpose tool that is state-of-the-art for MeV telescopes. MEGAlib performs the data calibration, event identification, reconstruction, as well as detailed source simulations. There are some high-level analysis tools in MEGAlib, but most of the COSI science analysis will be performed in *cosipy*. *cosipy* is the user-facing, python-based, high-level analysis tool for COSI data.



This Data Challenge will serve to introduce the community to *cosipy* and general Compton telescope analysis. We have prepared Jupyter Notebooks to walk the user through the analysis which are provided under spectral-fit and imaging; however, we suggest reading through the below description before attempting the notebooks.

cosipy was first developed by Thomas Siegert in 2019 to perform 511 keV image analysis from the 2016 COSI balloon flight (Siegert et al. 2020). Since then, it has been used for point source imaging and spectral extraction (e.g. Zoglauer et al. 2021), aluminum-26 spectral fitting (Beechert et al. 2022) and Al-26 imaging (in prep?), all using data from the COSI Balloon 2016 flight data. These analyses and the current Data Challenge use what we refer to as "*cosipy-classic*." The team is currently working on improved response handling and streamline tools built from the bottom up, and the new and improved *cosipy* will be the focus of next years' Data Challenge! With that in mind, there are still known issues and limitations with *cosipy-classic* that we will call out throughout this work.

The Simulated Data

For the first Data Challenge, we wanted to give the users a basic look at COSI data analysis, so we chose 3 straightforward examples based on the COSI science goals:

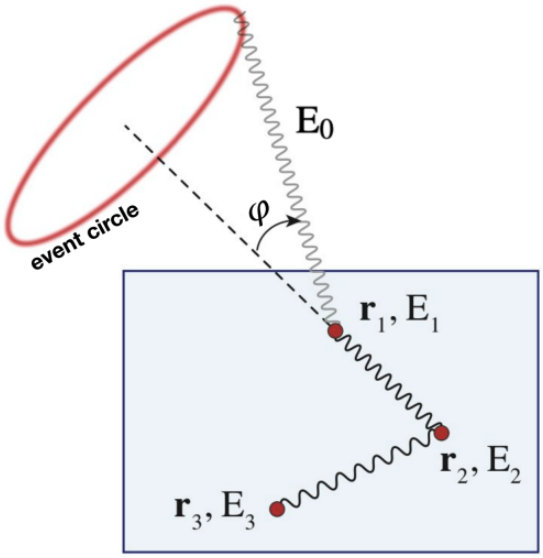
- Extracting the spectra from the Crab nebula, and other bright point sources
- Imaging bright point sources, such as the Crab and Cyg-X1
- Imaging diffuse emission from 511 keV and the ^{26}Al 1.8 MeV gamma-ray line

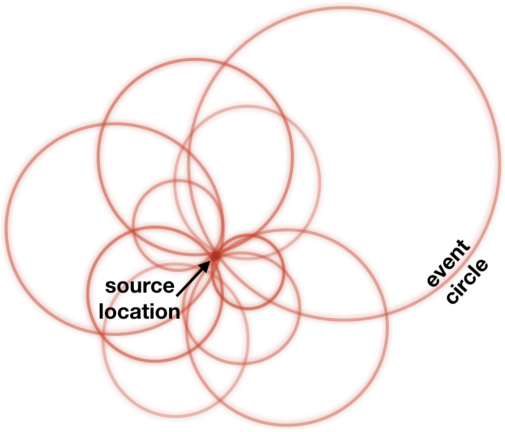
For each of these examples, we have provided a detailed description of the simulated sources and data products in the data_products directory. Each of the sources was simulated at x10 the astrophysical flux since the balloon flight had limited observation time, and because there were multiple detector failures during the balloon flight which reduced the effective area significantly.

General Compton Telescope Analysis Procedure

COSI is a Compton telescope, and analysis of MeV data is challenging due to the high-backgrounds and complicated instrument response. Thus, we will provide a basic description of Compton telescope data analysis here as an introduction for new users before diving into the Data Challenge analysis. For those who are new to Compton telescopes, we encourage you to read the review: Kierans, Takahashi & Kanbach 2022.

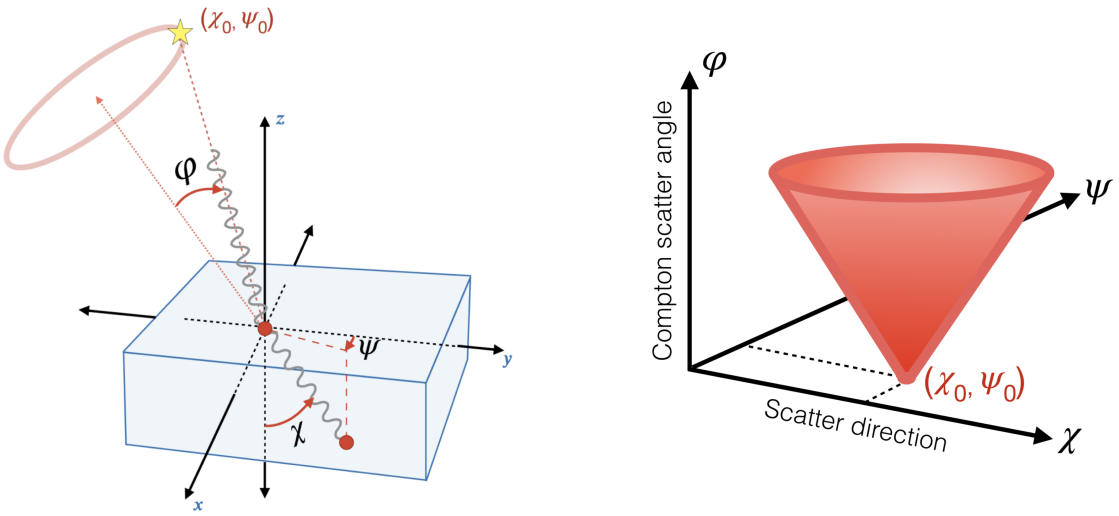
Compton telescopes provide a technique for single-photon detection, and each photon is measured as a number of energy deposits in the detector volume, shown in blue in the Figure below. These interactions must first be sequenced, and then reconstructed to determine the original direction of the photon, which is constrained to a circle on the sky defined with opening angle equal to the Compton scattering φ angle of the first interaction. The classic schematic for Compton telescopes is shown in the figure below, where this example shows a photon that results in 2 Compton scattering interactions, and finally a photoelectric absorption in interaction 3, fully containing the energy of the photon in the active detector volume. The event sequencing and reconstruction all occurs in the MEGAlib tool, and we will be starting our analysis with events that are already defined by their total energy deposit in the detector E_0, and the Compton scattering angle.





The above figure is what is traditionally shown for Compton telescopes, where an image of the source distribution can be realized by finding the overlap of event circles from multiple source photons, and then performing deconvolution techniques can recover a point-source image. This List-mode imaging approach (Zoglauer 2005) is implemented in MEGAlib, and can be used for strong point source, for example in laboratory measurements. However, this assumes a simplified detector response and cannot be used for the most sensitive analysis.

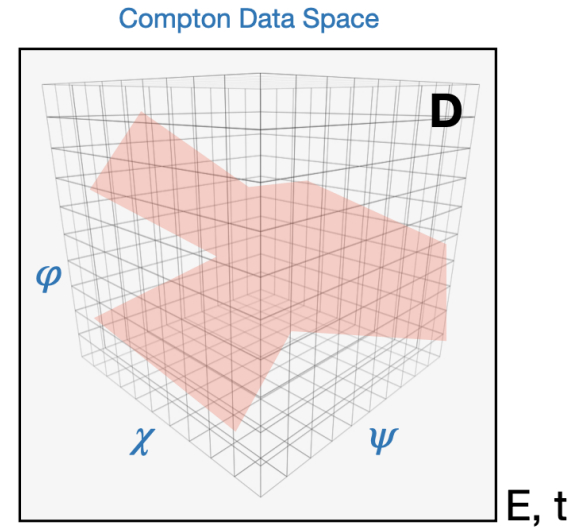
When performing analysis of astrophysical sources with high backgrounds, a more sophisticated description of the data space is required. This data space, along with the fundamentals of modern Compton telescope analysis techniques, was pioneered for the COMPTEL mission (Schönfelder et al. 1993 & Diehl et al. 1992) and is sometimes referred to as the COMPTEL Data Space, or simply the Compton Data Space (CDS). In the CDS, each event is defined by a minimum of 5 parameters. Three parameters describe the geometry of each event: the Compton scatter angle φ of the first interaction, and the polar and azimuthal angles of the scattered photon direction (χ, ψ). These angles as defined in instrument coordinates are shown in the figure below. The other two parameters in the CDS are the total photon energy (E_0) and the event time (t), but these are not always explicitly written.



The three scattering angles (φ, χ, ψ) make up 3 orthogonal axes of the CDS. As photons from a point source at location (χ_0, ψ_0) scatter in the detector, the CDS will be populated in the shape of a cone with apex at the source location, as shown in the figure on the right. This is the point spread function of a Compton telescope. The opening angle of the CDS cone is 90º since the Compton scatter angle is equal (within measurement error) to the deviation of the scattered photon direction. An extended source will appear as a broadened cone. The more familiar Angular Resolution Measure (ARM) for Compton telescopes is a 1-dimensional projection of the width of the CDS cone walls, representing the angular resolution.

All analysis with cosipy starts with reconstructed events defined in a photon list (MEGAlib's .tra files). The first step of any analysis is to bin the data into the CDS. For this Data Challenge, we are using 6º bin sizes for each of the scattering angles. This is because the angular resolution of the COSI Balloon instrument is ~6º at best, and smaller bins would be more computationally demanding. We also are using 10 energy bins for the continuum analyses, and 1 hour time bins. For the narrow-line sources, such as 511 keV or ^{26}Al, we use only 1 energy bin centered on the gamma-ray line of interest.

As a visual representation of the data (D) in the CDS, see the figure below. The three axes are the 3 scatter angles, and each bin contains the number of events, or counts, with that (φ, χ, ψ), shown with the red color fill (this is just a random distribution and not representative of what real data looks like in the CDS). The CDS is filled for each energy and time bin, represented by the subscript E,t in the figure. In cosipy-classic, χ and ψ are binned into in 1145 FISBEL bins, where FISBEL is MEGAlib's spherical axis binnings that has approximately equal solid angle for each pixel.



6º bins (30x φ bins, 1145x $\chi\psi$ FISBEL bins), 10x energy bins, 1hr time bins

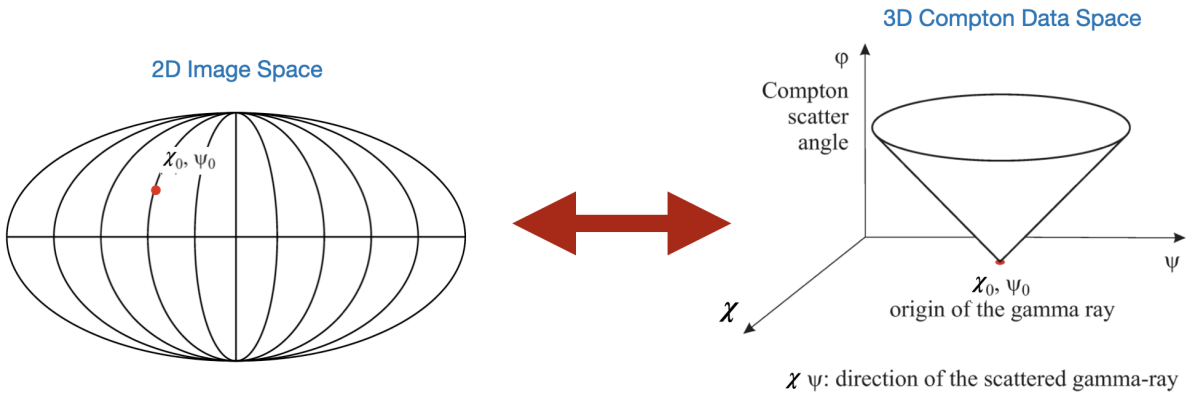
Now that we have a better understanding of the Compton telescope data space, and have our data binned in the CDS, we're ready to understand how to perform spectral analysis and imaging with COSI. There are 3 key pieces needed:

1) Response Matrix
2) Sky Model
3) Background Model

We will describe these components here and explain how they are used for general fitting procedures with COSI data. When running through the analysis notebooks pay attention to when these are initialized.

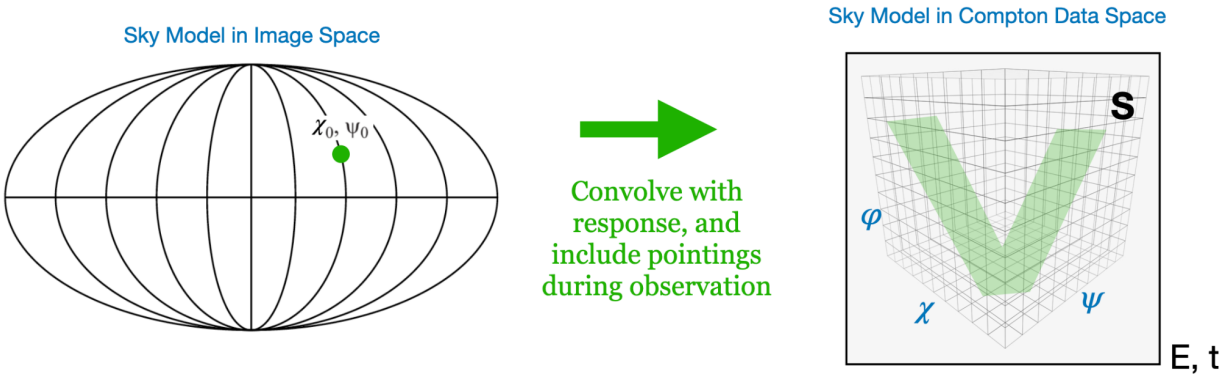
Response Matrix

The response matrix (R) for Compton telescopes represents the probability that a photon with energy E initiating from Galactic coordinates (l,b) interacts in the detector resulting in an event with measured energy E' and scattering angles (φ,χ,ψ). This is a matrix that is 2 dimensions larger (l,b) than the CDS and describes the transformation between image space and the CDS, taking into account the accurate response of the instrument. We build the response matrix through large MEGAlib simulations of an isotropic source.



Sky Model

For forward-folding analysis methods, we assume a source sky distribution, referred to as the sky model (S). The sky model in image space can be simply a point source or it can be a complicated diffuse model. This model is convolved with the instrument response matrix, and includes knowledge of the instrument aspect during observations, to determine the representation of the sky model in the CDS. For example, here we are showing this for a simple point source, and we regain the expected cone-shape in the CDS.

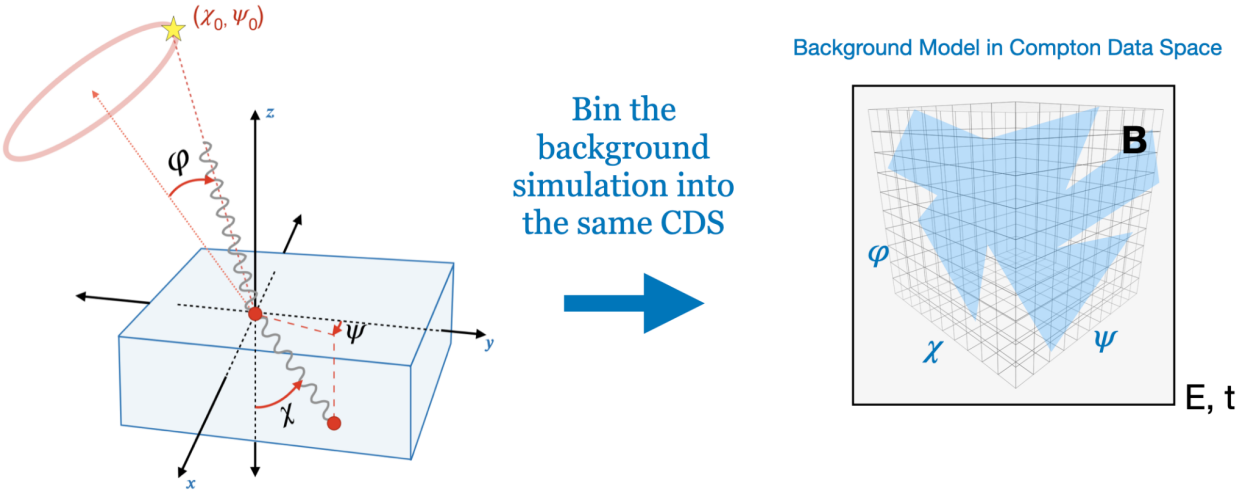


Background Model

We require an accurate estimate of the backgrounds during observations. This background model (B) can be achieved in a number of ways, for example by using the measured flight data from source-starved regions. Alternatively, one can perform full bottoms-up simulations of the gamma-ray background at balloon-flight altitudes, including activation. For this first Data Challenge, we are using this approach, and the simulation is further described in data_products; however, for future Data Challenges, we will be employing multiple background-model approaches.

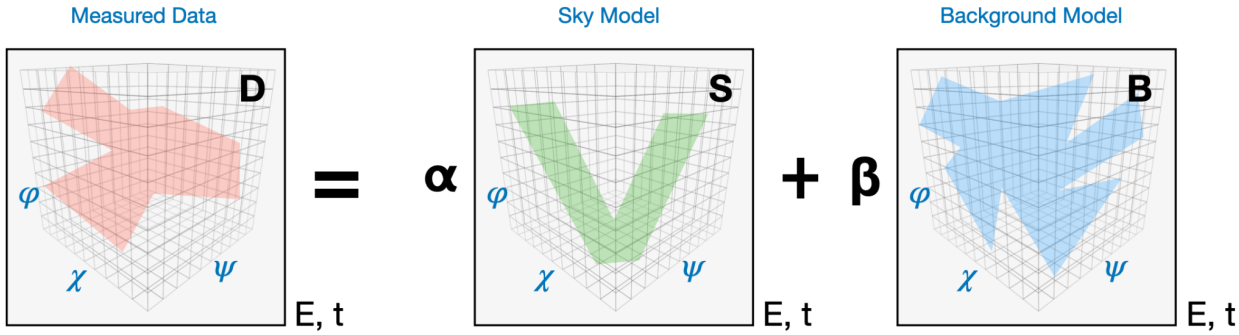
With the background model generated from simulations, then we first have to bin it in the same CDS that we have used for the data and the source model. We have already performed this

step for you, and have provided an .npz file, which is a zipped numpy array of the background simulation in the CDS.



Fitting General Principle

Finally, we have all of our components, and now we can perform our analysis. When model fitting, we generally fit for the amplitude of the source and background models that describe the data: D = **α**S + **β**B, shown schematically in the figure below, which is done for each energy bin and time bin independently. Through this procedure for spectral fitting and spatial model fitting, one can maximize the likelihood in the CDS.



If we don't know what the source should look like, instead of providing a sky model, we can be agnostic and perform image deconvolution. The data is a combination of the response convolved with the source sky distribution, and the addition of the background:

$$D = R \times S + B$$

'Data': Counts per - Energy E - Time t - CDS $\phi\psi\chi$	**Response:** Relating photon with $Et\phi\psi\chi$ to sky coordinates (l,b)	**Sky:** The source image.	**Background:** Everything that we measure that is not from the source.
Counts	**Heavy simulations Not invertible**	**What we want to know**	**Unknown Dominates >99%**

Since the response is non-invertible, we must use iterative deconvolutions to arrive at the sky distribution. In cosipy-classic, we will introduce you to a modified Richardson-Lucy algorithm, which is a special case of the expectation maximization algorithm developed for COMPTEL's images of diffuse gamma-ray emission (Knödlseder et al. 1999). The Richardson-Lucy algorithm starts from an initial image, and iteratively compares the data in the CDS to the sky distribution convolved with the transpose of the response matrix R:

$$\hat{S}_{i+1} = \hat{S}_i \left(\frac{D}{\hat{S}_i \times R} \times R^{\dagger} \right)$$

This will evolve into the maximum likelihood solution. There are other image deconvolution techniques that will be used for COSI imaging, and those will be introduced in subsequent Data Challenges.

Next Steps

Now that you have a better feeling for the Compton telescope analysis process, at least in a general sense, you're ready to start on the analysis examples. First, we recommend you check out the detailed description of the simulations in the data_products directory. The easiest analysis is the spectral fitting, and we recommend you start there. The Richardson-Lucy imaging is computationally intensive, and still largely hard-coded in this release, and thus we recommend you work through this notebook second. And then finally, we have also included some of the COSI flight data, specifically for analysis of the Crab as seen during the 2016 flight. After you've run through the spectral fitting and image deconvolution with the simulated data, you should be ready to analyze this real flight data!

Simulations

For the first Data Challenge, we wanted to give the users a basic look at COSI data analysis, so we chose 3 straightforward examples based on the COSI science goals:

- Extracting the spectra from the Crab, Cen A, Cyg X-1, and Vela
- Imaging bright point sources, such as the Crab and Cyg X-1
- Imaging diffuse emission from 511 keV and the ^{26}Al 1.8 MeV gamma-ray line

For each of these examples, we have provided a detailed description of the simulated sources and data products here in the data_products directory. Each of the sources was simulated at x10 the astrophysical flux since the balloon flight had limited observation time, and because there were multiple detector failures during the balloon flight which reduced the effective area significantly.

The simulations were all performed in MEGAlib, with an accurate mass model of the COSI Balloon instrument. The COSIBalloon.9Detector.geo.setup model, which includes the detector failures, was used for all of the simulations. Each of the continuum simulations was performed for 100 keV – 10 MeV, and an energy range selection of <5 MeV was used in MEGAlib's mimrec event selection tool.

Data Products:

We have included many combinations of the source simulations and background to allow for flexibility and further testing with these files. All files are either the .npz zipped numpy array format for the CDS-binned response matrix or background model, or the MEGAlib photon list .tra.gz format. MEGAlib was used to perform all of these simulations, and the source models are described in detail below.

Within this directory, there are 2 background models and 3 response matrices.

- Scaled_Ling_BG_1x.npz: background model generated from C. Karwin's scaled 1x Ling background simulation
- Scaled_Ling_BG_3x.npz: background model generated from C. Karwin's scaled 3x Ling background simulation for more statistics
- Continuum_Response.npz: 6º response used for spectral analysis and imaging continuum sources
- 511keV_imaging_response.npz: 6º imaging response required for RL imaging of Galactic positron annihilation
- 1809keV_imaging_response.npz: 6º imaging response required for RL imaging of Galactic Al-26

There is the full-sky simulation with background and all of the sources combined:

- DC1_combined_10x.tra.gz: 4 point sources with 10x flux (Crab, Cyg X1, Cen A, Vela), 511 kev & Al-26 lines, 1x Ling background

There is each of the sources individually with the background:

- Point_sources_10x_BG.tra.gz: 4 point sources with 10x flux (Crab, Cyg X1, Cen A, Vela) and 1x Ling background
- Crab_BG_10x.tra.gz: Crab with 10x flux and 1x Ling background
- CenA_BG_10x.tra.gz: Cen A with 10x flux and 1x Ling background
- GC511_10xFlux_and_Ling.inc1.id1.extracted.tra.gz: 511 emission with 10x flux with Ling background
- DC1_Al26_10xFlux_and_Ling.inc1.id1.extracted.tra.gz: Al26 emission with 10x flux with Ling background

Each of the sources is also included without background:

- Point_sources_10x.tra.gz: 4 point sources with 10x flux (Crab, Cyg X1, Cen A, Vela)
- Crab_only_10x.tra.gz: Crab with 10x flux
- CygX1_only_10x.tra.gz: Cyg X1 with 10x flux
- CenA_only_10x.tra.gz: Cen A with 10x flux
- Vela_only_10x.tra.gz: Vela with 10x flux
- GC_511_10xFlux_only.inc1.id1.extracted.tra.gz: 511 emission with 10x flux
- Al26_10xFlux_Only.inc1.id1.extracted.tra.gz: Al26 emission with 10x flux

And finally, there is 1 background simulation (this is not required for any analysis, but is included here for posterity):

- Scaled_Ling_BG_1x.tra.gz: C. Karwin's scaled 1x Ling background simulation

All files have the same start and stop time, in unix time:

start time: 1463443400.0 s
stop time: 1467475400.0 s
total time: 4032000.0 s = 46.67 days

Which corresponds to May 17 2016 00:03:20 GMT to Jul 02 2016 16:03:20 GMT, covering the full COSI Balloon flight in 2016.

As described in the main cosi-data-challenge-1 README, these files are stored on Git's Large File Server, and one needs to install git-lfs to access them.

Source Models

Point Sources

There are four bright point sources that are included in these simulations: the Crab nebula, Cygnus X-1, Centaurus A, and Vela. The spectra and flux for each of these sources was determined from the literature. We have used 10x the flux values for these sources since the COSI Balloon flight had limited observation times, and the effective area was limited due to detector failures. The COSI SMEX mission is expected to be 50x more sensitive than the balloon-borne mission, so please keep that in mind when you see the detections in this Data Challenge.

Crab:

(l,b) = (184.56, -5.78)
Spectral shape: Band function from Jourdain et al. 2020
Flux = 0.48977 ph/cm^2/s between 100 keV and 10 MeV

Cen A:

(l,b) = (309.52, 19.42)
Spectral shape: SED from HESS+LAT collaboration 2018
Flux: 0.03609 ph/cm^2/s between 100 keV and 10 MeV

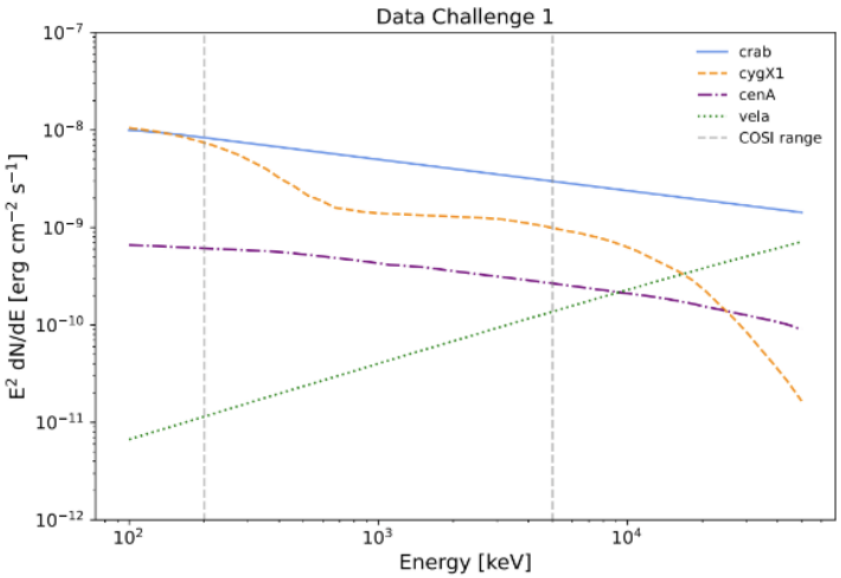
Cyg X1:

(l,b) = (71.33, 3.07)
Spectral shape: SED from Kantzas+21
Flux = 0.40644 ph/cm^2/s between 100 keV - 10 MeV

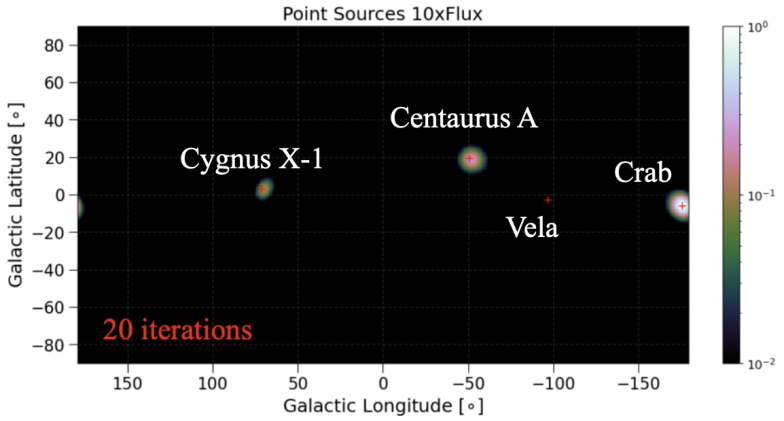
Vela:

(l,b) = (263.55, -2.79)
Spectral shape: SED from Mattana+11
Flux = 0.00120 ph/cm2/s between 100 keV - 10 MeV

The input spectrum for these point sources is shown below.



The simulations are run in MEGAlib's cosima tool, and then a list-mode image is created in mimrec to confirm the correct point source locations:

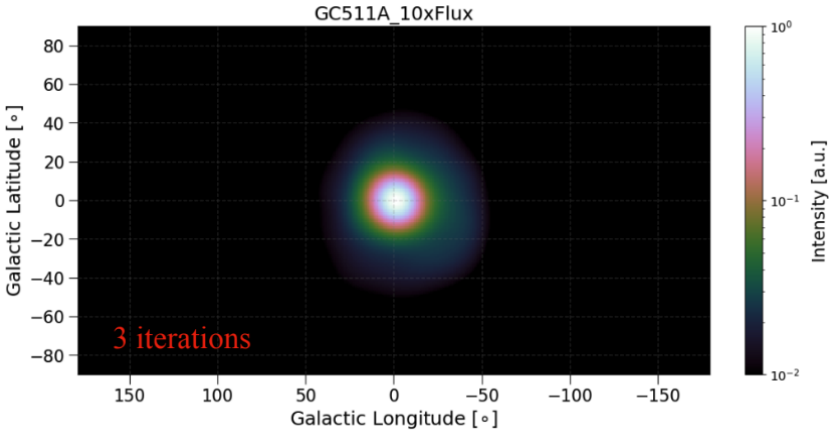


Positron Annihilation at 511 keV

The morphology of the 511 keV emission from positron annihilation is not well constrained. For this first Data Challenge, we have used the model defined in [Knödlseder et al. 2005], where the emission was fit with a 2-D asymmetric Gaussian spatial model with the following parameters:

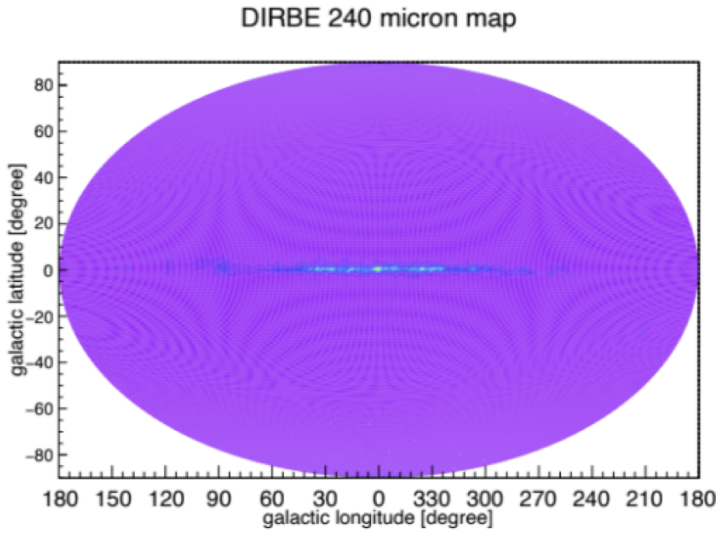
Quantity	Measured value
RMLR (DOF)	462.2 (5)
l_0	$-0.6° \pm 0.3°$
b_0	$+0.1° \pm 0.3°$
Δl (*FWHM*)	$8.1° \pm 0.9°$
Δb (*FWHM*)	$7.2° \pm 0.9°$
511 keV flux (10^{-3} ph cm^{-2} s^{-1})	1.09 ± 0.04

The emission was simulated as a 511 keV mono-energetic source, and the image from mimrec confirms the extended emission in the Galactic Center.

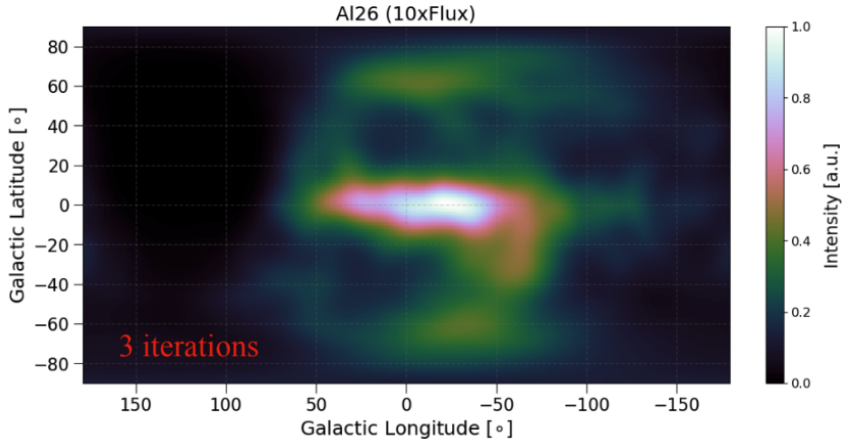


Aluminum-26 Decay at 1.8 MeV

The Diffuse Infrared Background Experiment (DIRBE) 240 um map has been shown to be a good tracer for the Al-26 emission, as measured by COMPTEL and INTEGRAL/SPI. We use that distribution as the spatial model for the Al-26 emission, and the inner Galaxy flux was normalized to 3.3×10^{-4} ph/cm^2/s (Diehl et al. 2006). This model is described further in Beechert et al. 2022.



The emission was simulated with a 1.8 MeV mono-energetic source, and the image from mimrec confirms the extended emission along the Galactic disk.



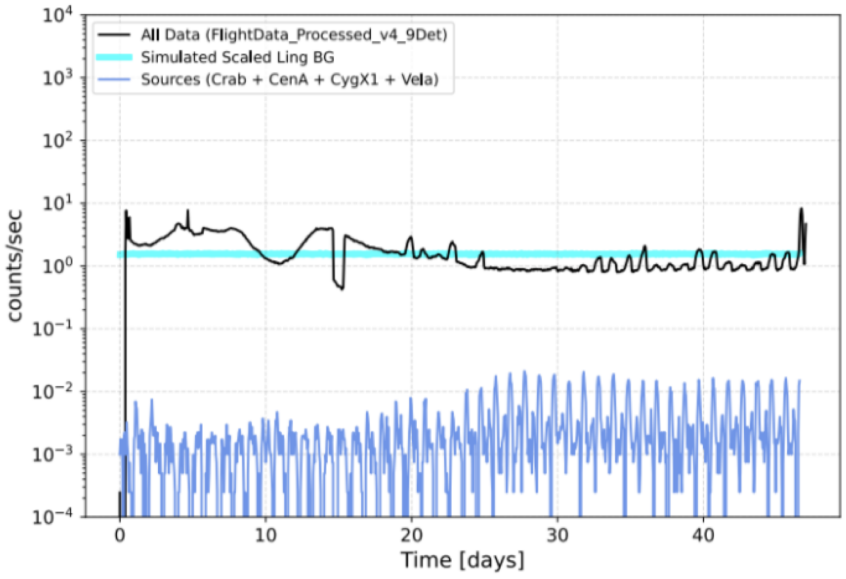
Note that the list-mode imaging method employed in MEGAlib is not optimized for diffuse sources, and the extra structure out of the Galactic plane is an imaging artifact.

Background Radiation

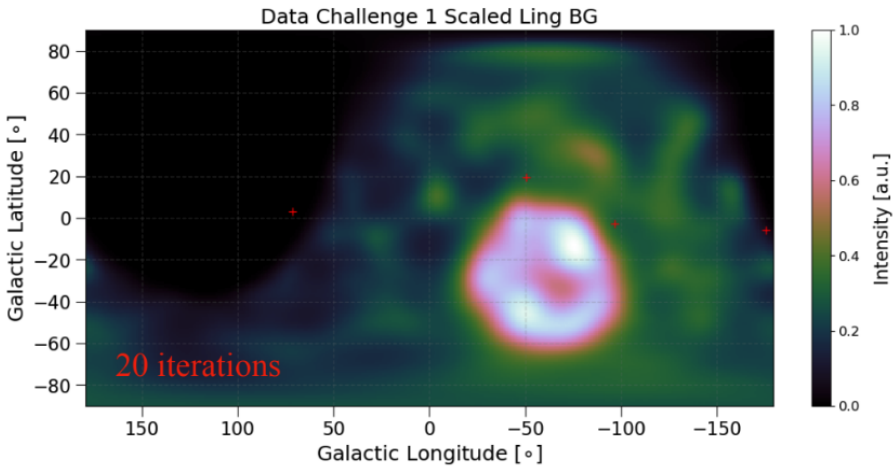
The background radiation model used for the Data Challenge is based on the semi-emperical model from Ling 1975, which has been used for all COSI balloon analyses. The model

describes the angular and atmospheric depth dependance of gamma-rays within 0.3 – 10 MeV and includes 3 primary gamma-ray sources: continuum emission (Bremsstrahlung from cosmic ray interactions in the atmosphere), 511 keV line component (from electron-positron annihilation in the atmosphere, and cosmic diffuse component (down-scattered gamma-rays from the extragalactic background and Galactic plane). The Ling model requires a description of the atmosphere's density, interaction depth, and mass absorption coefficients, which are obtained through the NRLMSISE-00 model. An altitude of 33.5 km was assumed for these models - altitude drops and changes in longitude and latitude were not taken into account for this simulation.

The amplitude of the Ling background was scaled so the total integrated background spectrum from simulation matched closely to what was measured during flight, as can been seen in the figure below. The flight data ("All Data" label) shows a time-variable background count rate that was influenced by the geomagnetic cutoff, and balloon altitude drops. The cyan line shows the scaled Ling background model. As a reference, the count rate from the 1x flux point sources are shown, and the signal is at most a few percent of the background count rate.

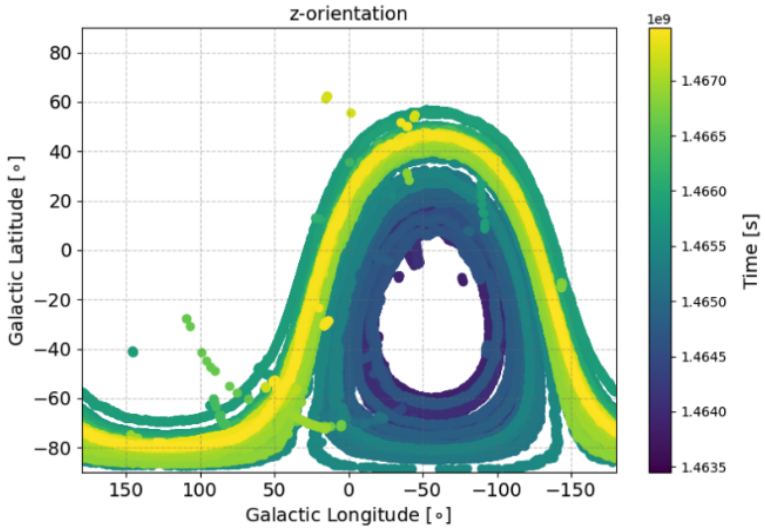


An image of the background simulation traces the exposure map, since the orientation of the COSI Balloon in Galactic coordinates was included in the simulation:

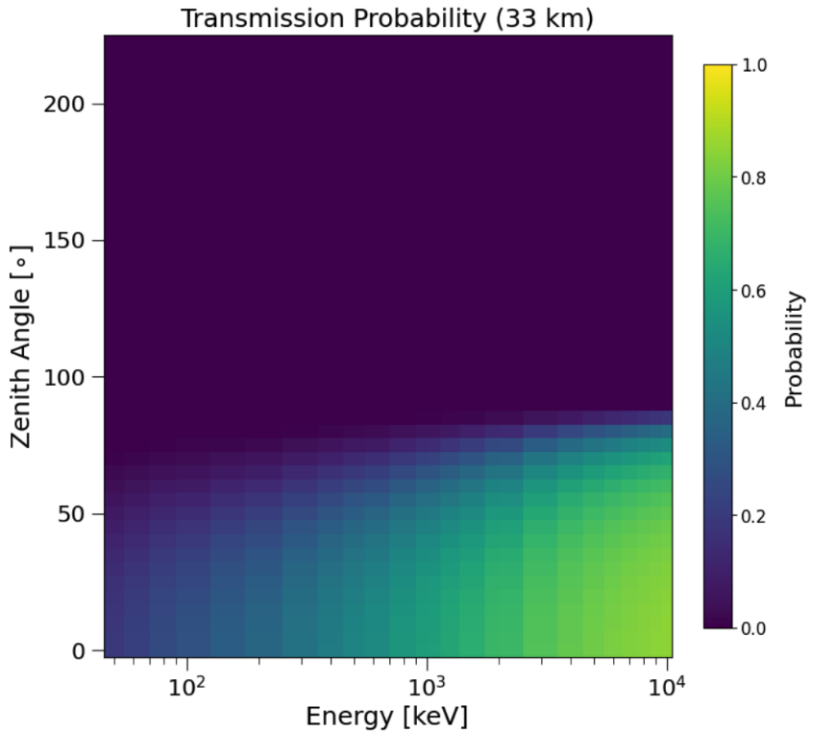


Flight path and transmission

The source simulations include the real flight aspect information so that the balloon path and source exposure time is accurate. This can be seen in the below plot showing the Galactic longitude and latitude of the zenith direction of the COSI balloon as a function of time.



Furthermore, the transmission probability of the source photons in the atmosphere are calculated for each instance of the simulation. The probability of transmission is taken at a constant altitude of 33 km, and is shown as a function of zenith angle and energy in the below figure.


Transmission Probability (33 km)
Zenith Angle [∘]
200
150
100
50
0
10^2
10^3
10^4
Energy [keV]
1.0
0.8
0.6
0.4
0.2
0.0
Probability