PROCEEDINGS OF SCIENCE

Approximating the COSI Telescope Response with Neural Networks

Pascal Janowski,[a,*] Savitri Gallego,[a] Uwe Oberlack,[a] Jan Peter Lommler,[a] Israel Martinez-Castellanos,[b,c] John Tomsick,[d] Andreas Zoglauer[d] and Steven E. Boggs[e] on behalf of the COSI collaboration

[a] *Institut für Physik & Exzellenzcluster PRISMA+, Johannes Gutenberg-Universität Mainz, 55099 Mainz, Germany*

[b] *Department of Astronomy, University of Maryland, College Park, MD 20742, USA*

[c] *NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, USA*

[d] *Space Sciences Laboratory, UC Berkeley, 7 Gauss Way, University of California, Berkeley, CA 94720, USA*

[e] *Department of Astronomy & Astrophysics, UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA*

E-mail: pjanowsk@students.uni-mainz.de, sgallego@uni-mainz.de, oberlack@uni-mainz.de, lommler@uni-mainz.de, imc@umd.edu, jtomsick@berkeley.edu, zoglauer@berkeley.edu, seboggs@physics.ucsd.edu

The Compton Spectrometer and Imager (COSI), a gamma-ray telescope set to launch in 2027 as a NASA Small Explorer satellite mission, is a compact Compton telescope consisting of a cross-strip germanium detector array. Owing to its wide field-of-view and excellent energy resolution, COSI is set to achieve an unprecedented angular resolution and line sensitivity among Compton telescopes in the 0.2-5 MeV energy band.

This requires a precise characterization of the instrument response, which is challenging to achieve with a traditional multi-dimensional binned response.

We therefore pursue a novel approximation scheme, consisting of a conditional autoregressive spline flow neural network and a spherical harmonics expansion, working in a dedicated relative coordinate space, to obtain a detailed model of the response function.

Early applications in unbinned maximum likelihood analysis, including RL-deconvolution and spectral fitting, underscore the broad utility of our method. This continuous approach has the potential to overcome the inherent limitations of conventional discretized models, effectively bridging the gap between COSI's innovative design and the practical challenges of gamma-ray data analysis.

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

*Speaker

PoS(ICRC2025)698

1. Introduction

The Compton Spectrometer and Imager (COSI) is a NASA SMEX mission, scheduled to launch in 2027, and designed to probe the MeV gap in the $0.2 - 5$ MeV range, a region in which sensitivity is markedly worse than in adjacent energy bands. This limitation is due to the high instrumental background as well as the inherent challenges of COSI's main method to measure gamma rays, Compton scattering.

Its primary science goals are to uncover the origin of galactic positrons, reveal galactic element formation, gain insight into extreme environments with polarization, and probe the physics of multimessenger events. All of these require analysis tools, which are currently in development and will become part of the dedicated Python library `cosipy` [1]. For all of them, it is crucial to have a precise model of the instrument response, the key to predicting the measurement of an incoming photon in the Compton Dataspace (CDS; see [2]). Current methods rely on high-dimensional response matrices, which face exponentially growing size limitations going to smaller resolutions, and thus would require close to super-computer level resources to fully utilize COSI's energy and angular resolution.

This work presents an alternative approximation scheme for the response function based on neural networks, whose details are outlined in section 2. With section 3 covering the method's ability to reach the desired accuracy, as well as its performance for broad usage, section 4 provides an overview of tailored applications for the response approximation, which will serve as the most important benchmark. Section 5 concludes with an outlook on our future work, further extending the response and analysis pipeline for the `cosipy` framework.

2. Response Approximation

The classical CDS of detected events after reconstruction is characterised by the measured energy E_m, the Compton scattering angle ϕ and the scattering direction in local detector coordinates $\psi\chi$. The expected distribution in the CDS caused by a given flux model F, which produces photons with energy E_i from local direction $\nu\lambda$, can be calculated by its convolution with the response function R

$$\frac{\mathrm{d}N}{\mathrm{d}t\mathrm{d}E_m\mathrm{d}\phi\mathrm{d}\psi\chi} = \int \underbrace{R\left(E_m, \phi, \psi\chi; \nu\lambda, E_i\right)}_{A_{\mathrm{eff}}(\nu\lambda, E_i)\, P(E_m, \phi, \psi\chi | \nu\lambda, E_i)} F\left(\nu\lambda_{\mathrm{gal}}(\nu\lambda, t), E_i\right) \mathrm{d}\nu\lambda\, \mathrm{d}E_i. \tag{1}$$

R is a product of the effective area A_{eff}, which depends only on the initial parameters, and a conditional pdf P that includes the physical principles of a Compton telescope, detector effects, and the specific geometric characteristics of the instrument.

2.1 Relative Coordinates

This complexity can be mitigated by expressing the response in a more suitable coordinate system. By reparameterizing in coordinates relative to the incoming photon, rather than the detector, we reduce the amount of physical detail that must be explicitly modeled

$$E_m,\, \psi\chi \;\rightarrow\; \epsilon_m = \frac{E_m - E_i}{E_i},\; \theta = \phi_{\mathrm{geo}} - \phi,\; \zeta. \tag{2}$$

ϵ_m defines the measured energy relative to the initial energy, thereby positioning the most dominant spectral feature, the photopeak ($E_m = E_i$), at $\epsilon_m = 0$. θ is the angular resolution measure (ARM), the difference between the geometric scattering angle $\phi_{\text{geo}\measuredangle}(\psi\chi, \nu\lambda)$ and the kinematic scattering angle ϕ, which follows from the Compton kinematic equation. ζ is the photon-frame azimuthal Compton scattering angle.
It should be noted that these coordinates depend on the initial photon parameters and therefore must be converted back to classical CDS coordinates used for analysis. Nevertheless, this internal parameterization creates a probability distribution landscape that is easier to interpolate, thereby improving the learnability of our neural network, which we cover in the next section.

2.2 Response PDF

While it might seem more intuitive to train a neural network to learn the response function as a whole, modeling P separately allows us to utilize its inherent normalization, which is the concept behind normalizing flow networks [3]. These networks start with a simple base distribution $B(\mathbf{u})$, which is in most cases a Gaussian. Via bijective and easily invertible transformations $\mathbf{x} = \mathbf{f}_\theta(\mathbf{u})$, with some trainable parameters θ, we obtain the target pdf

$$P(\mathbf{x}) = B\left(\mathbf{f}^{-1}(\mathbf{x})\right)\left|\det\left(\frac{\partial \mathbf{f}^{-1}}{\partial \mathbf{x}}\right)\right|. \tag{3}$$

Note that we used the basic change of variables principle with the Jacobian determinant of the inverse transformation, thereby maintaining inherent normalization. In practice, multiple such transformations are combined to model very complex distributions.
There are many different types of normalizing flow networks, which differ in the way they couple the components of $\mathbf{x}$, define the transformations $\mathbf{x} = \mathbf{f}_\theta$, and incorporate the conditioning, which in this case are $\nu\lambda$ and E_i. Our implementation is based on the recently proposed "Conditional-Autoregressive-Rational-Quadratic Spline" architecture [3], which is part of the `normflows` library [4] extending `PyTorch` [5].
When training the neural network, we minimize the loss, which is simply the negative log-likelihood,

$$-\frac{1}{N}\sum_{k=1}^{N}\log P([\mathbf{x}]_k) = -\frac{1}{N}\left[\log B([\mathbf{u}]_k) - \sum^{\text{layer}}\log|\det \mathbf{J}|\right]. \tag{4}$$

Here, $[\mathbf{x}]_k$ are reconstructed and selected Compton events, and $\mathbf{J}$ is the Jacobian. All events are obtained by simulating a sphere around the COSI telescope with a constant flux from 10 keV to 10 MeV. The coordinates are normalized to $[0, 1]$, while the initial energy is both scaled and sampled evenly in log-space to more accurately distribute the corresponding changes of the response. The azimuthal angle ν is periodic and can be encoded using sine and cosine, resulting in a total of four context and four latent variables. Optimization is performed with RAdam [6] in combination with a cosine learning rate decay.

2.3 Effective Area

The effective area is a rather smooth distribution in $\nu\lambda$, which changes continuously with E_i. This motivates its approximation with a spherical harmonics expansion $Y_{lm}(\nu\lambda)$ and energy-dependent coefficients $a_{lm}(E_i)$,

$$A_{\text{eff}}(\nu\lambda, E_i) = \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{+l} a_{lm}(E_i) Y_{lm}(\nu\lambda). \tag{5}$$

For a given (l, m), the coefficients, which are calculated using `healpy` [7, 8], are modeled as polynomials up to a given order. The real and imaginary parts are fitted separately.

3. Accuracy & Performance

Since the response is high-resolution and high-dimensional, and any data collected to verify it would only sample the space very sparsely, there is no direct way to assess the accuracy of the neural network prediction via residuals, as can be done for the effective area expansion. Therefore, we can only compare projections into one dimension, examine its ability to reproduce expected physical lines in slices, and determine its fitting performance in the applications discussed in section 4.

3.1 Response PDF

Figure 1 shows a comparison of the neural network pdf and the classical reference projected onto ϵ_m and ϕ for the most important gamma-ray lines at 511 keV and 1809 keV. Both demonstrate the ability of the approximation to reproduce peaks and the continuum landscape. However, its accuracy declines slightly when faced with discontinuous or highly context-dependent features.

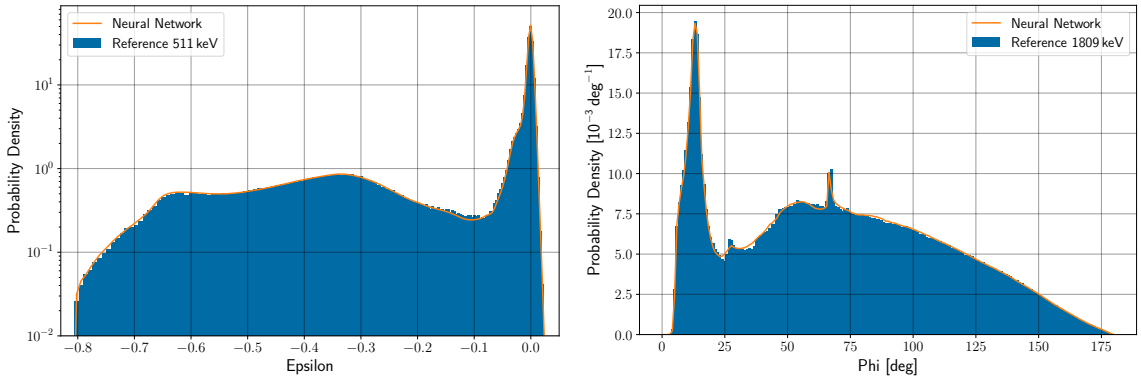

Figure 1: Projections of the response pdf, comparing the simulated, binned reference and the neural-network inference. The reference was accumulated over all incidence angles ν with scattering angle λ, $<, 25°$ to ensure sufficient statistics, whereas the neural network integration was carried out at $\nu = 45°$ and $\lambda = 10°$. The two panels show reference energies of (a) 511 keV projected onto ϵ_m (b) 1809 keV projected onto ϕ.

This is especially true for ζ, which encodes most of the detector's geometry information. Locations in the xy-plane are limited to the strip intersections that are used to detect interactions inside the germanium crystals (see [9]), making the pdf discontinuous and ζ highly dependent on the initial direction $\nu\lambda$. To address this, we apply a Gaussian blur to events before batching them for training. This prevents the neural network from learning artifacts. The required amount of blur depends on the amount of available data and the model size. For $\sigma = 5°$, the result is depicted in fig. 2.

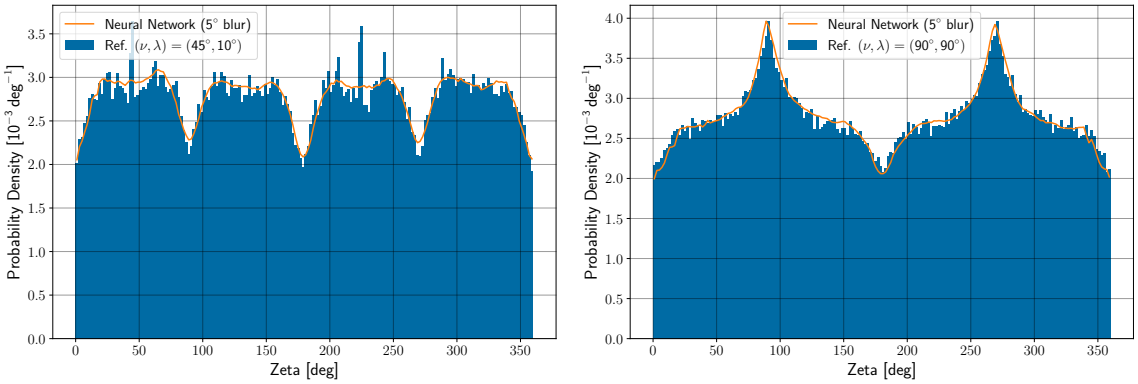

Figure 2: Projections of the response pdf onto ζ, comparing the simulated, binned reference and the neural-network inference, both at 1809 keV. The network was trained on ζ data blurred by a Gaussian with $\sigma = 5°$. The reference was accumulated over $\pm 10°$ around (a) $(\nu\lambda) = (45°, 10°)$ (b) $(\nu\lambda) = (90°, 90°)$.

With these being projections, physical lines allow for a deeper understanding of the actually learned physical information. The most dominant lines present in the response are the photopeak ($E_i = E_m$), the single-escape peak ($E_i = E_m + 511\,\text{keV}$), the annihilation peak ($E_m = 511\,\text{keV}$), and two incomplete absorption peaks (more details in appendix A). The latter exhibit complex relationships between different CDS variables, which makes them another valuable indicator for the obtained accuracy. Additionally, threshold and saturation effects of the germanium strips can be observed, as well as cuts applied to two-site events due to hit-ordering ambiguities. Figure 3 shows two response slices for given source and CDS configurations. The different lines we predict are also marked.

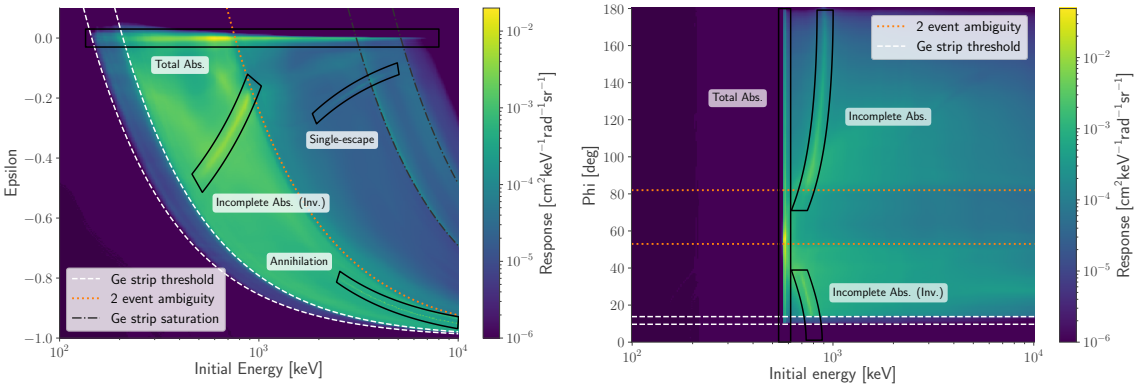

Figure 3: Network response slices with key features annotated. (a) E_i and ϵ_m. (b) E_i and ϕ.

3.2 Effective Area

As an example, fig. 4 shows the approximated effective area and the residuals normalized by the reference's statistical uncertainty at 511 keV. Residuals remain minor, well within statistical errors, and the median relative deviation is only 1 %, so efforts are focused on enhancing the pdf approximation.

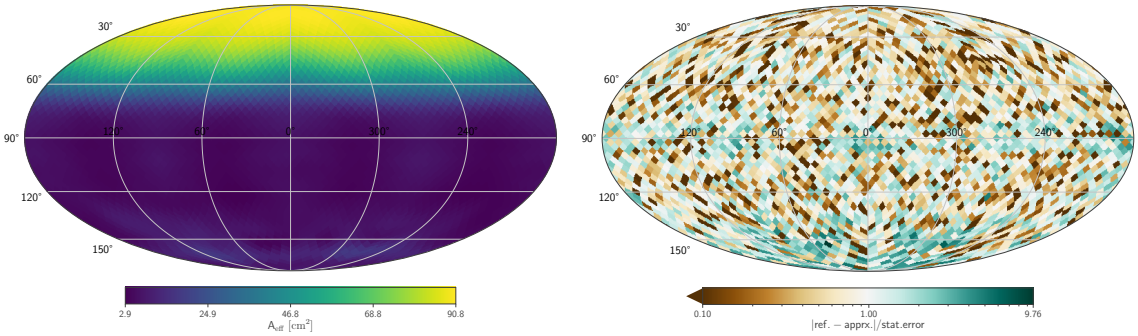

Figure 4: (a) Approximated effective area at 511 keV using a spherical-harmonic expansion up to $l_{\max} = 24$. (b) Residuals, defined as the absolute difference between approximation and reference, normalized by the reference's statistical uncertainty. Both in local detector coordinates $\nu\lambda$.

3.3 Implementation & Inference

The full response approximation is implemented in `PyTorch` [5], utilizing all available GPUs to achieve optimal inference time. The effective area requires calculating both the polynomial coefficients and the spherical harmonics. The former can be evaluated using Horner's method (see appendix B), which is automatically parallelized. For the latter, we employ the `spherical` library [10] with `torch` backend, which performs fast harmonic approximations on the GPU. We achieve about 4.8 million evaluations per second on the following hardware:

3× NVIDIA 2080 Ti (12 GB) GPUs, Intel Xeon Gold 5218 CPU, 256 GB RAM

For the pdf, we distribute the load across all GPUs and reach approximately 220,000 evaluations per second with 53 million trainable parameters. The conversion between relative coordinates and classical CDS is fully vectorized and adds no significant delay. In total, this means it is not feasible to perform inference of the neural network on the fly, for example repeatedly calculating an updated likelihood. Instead, the response for specific events is evaluated once beforehand and can then be reused throughout the actual analysis.

4. Application

The approximation marks a shift from using a binned response matrix—which would require slow, inaccurate integration to recover bins—to directly modeling the underlying response function. This enables an unbinned Poisson maximum-likelihood estimation framework, eliminating artifacts that arise from the coarse binning forced by our data sparsity

$$-\log \mathcal{L} = N - \sum_{\text{events}} \log \frac{\mathrm{d}N}{\mathrm{d}t\mathrm{d}E_m\mathrm{d}\phi\mathrm{d}\psi\chi}. \tag{6}$$

N is the total expected number of events. It was successfully applied to spectral analyses of point sources such as the Crab nebula and gamma-ray bursts (GRBs), as well as RL image deconvolution. In the process, we developed additional techniques to approximate the background distribution, which is also part of N and $\mathrm{d}N$, using a similar neural network architecture. Furthermore, we optimized the numerical integration required for likelihood calculations, which must be performed for

each individual event, significantly reducing both the required amount of RAM and precomputation time. A detailed discussion of these background modeling and optimization techniques, together with the associated fitting results, is beyond the scope of this presentation and will be addressed in a separate publication.

5. Outlook

The current version of the response approximation and its application pipeline is based on the COSI Data Challenge 3 (DC3) [11], a yearly release combining increasingly realistic simulated data and analysis tools to prepare for the eventual launch of the COSI telescope. This means that the response approximation will need to adapt continuously to changes in detector and geometry effects, as well as to different choices of event selection criteria. In this process, the effects of additional training data, increases in model size, and varying degrees of ζ blur on performance, especially in the analysis pipeline, need to be explored. Additionally, COSI is designed to analyze the polarization of gamma radiation, which will introduce additional conditioning variables into the neural network pdf.
In parallel, it is necessary to adapt the `cosipy` framework to the new unbinned analysis pipeline based on our response approximation and to develop further techniques to achieve smooth and high-accuracy likelihood computation while managing memory and computational demands. The goal is to support likelihood sampling approaches, such as MCMC, for robust error estimation, while simultaneously lowering hardware requirements to broaden accessibility.

A. Incomplete Absorption

The response features two additional resonances which are characteristic for the underlying Compton effect of the COSI telescope. The energies E_2 and $E_1 = E_m - E_2$, being the two deposited energies of a photon inside the detector, can be determined with

$$E_2 = \frac{511\,\mathrm{keV}}{1 + \frac{511\,\mathrm{keV}}{E_m} - \cos(\phi)}. \tag{7}$$

Assuming that the geometric scattering angle ϕ_{geo} would have been obtained without the missing energy $E_\Delta = E_i - E_m$ the incomplete absorption feature is defined by

$$\phi_{\mathrm{geo}} = \arccos\left(1 - \frac{511\,\mathrm{keV}}{E_2 + E_\Delta} + \frac{511\,\mathrm{keV}}{E_1 + E_2 + E_\Delta}\right). \tag{8}$$

A second absorption peak is created by events, whose event order was wrongly reconstructed, meaning the scattering direction defined by $\psi\chi$ is inverted and E_1,E_2 in eq. (8) are switched.

B. Horner's Method

A polynomial $p(x) = c_0 + c_1 x + c_2 x^2 + \ldots + c_n x^n$ can be calculated efficiently in a recursive manner, reducing the required number of multiplications from $2n - 1$ for all exponents to n

$$p(x) = c_0 + x\left(c_1 + x\left(\ldots + x\left(c_{n-1} + c_n x\right)\ldots\right)\right). \tag{9}$$

Acknowledgments

The Compton Spectrometer and Imager is a NASA Explorer project led by the University of California, Berkeley with funding from NASA under contract 80GSFC21C0059. Resources supporting this work were provided by the NASA High-End Computing (HEC) Program through the NASA Advanced Supercomputing (NAS) Division at Ames Research Center.

SG and JL acknowledge support by DLR grant 50OO2218. Resources supporting this work were provided by National High Performance Computing (NHR) South-West at Johannes Gutenberg University Mainz.

References

[1] I. Martinez, *The cosipy library: COSI's high-level analysis software*, *PoS (ICRC2023)* **444** (2023) 858 [doi:10.22323/1.444.0858]

[2] C. Kierans, T. Takahashi and G. Kanbach, *Compton Telescopes for Gamma-Ray Astrophysics*, *Handbook of X-ray and Gamma-ray Astrophysics*, Springer Nature Singapore (2022) 1–72, ISBN 9789811645440 [doi:10.1007/978-981-16-4544-0_46-1]

[3] C. Durkan, A. Bekasov, I. Murray and G. Papamakarios, *Neural Spline Flows*, *Advances in Neural Information Processing Systems* **32** (2019), Curran Associates, Inc.

[4] V. Stimper *et al.*, *normflows: A PyTorch Package for Normalizing Flows*, *J. Open Source Software* **8**(86) (2023) 5361 [doi:10.21105/joss.05361]

[5] J. Ansel, E. Yang, H. He *et al.*, *PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation*, *ASPLOS* **2** (2024) 929 [doi:10.1145/3620665.3640366]

[6] L. Liu *et al.*, *On the Variance of the Adaptive Learning Rate and Beyond*, *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, (2020)

[7] A. Zonca *et al.*, *healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in Python*, *J. Open Source Softw.* **4**(35) (2019) 1298 [doi:10.21105/joss.01298]

[8] K. M. Górski *et al.*, *HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere*, *Astrophys. J.* **622** (2005) 759 [astro-ph/0409513]

[9] J. A. Tomsick *et al.*, *The Compton Spectrometer and Imager*, *PoS (ICRC2023)* **444** (2023) 745 [doi:10.22323/1.444.0745]

[10] F. Bigi, G. Fraux, N. J. Browning and M. Ceriotti, *Fast evaluation of spherical harmonics with sphericart*, *J. Chem. Phys.* **159** (2023) 064802 [doi:10.1063/5.0156307]

[11] The Compton Spectrometer and Imager (COSI) Collaboration, *COSI Data Challenges*, Zenodo (2025) [doi:10.5281/zenodo.15126188]