PAPER • OPEN ACCESS

Extracting the gamma-ray source-count distribution below the Fermi-LAT detection limit with deep learning

To cite this article: A. Amerio *et al* JCAP09(2023)029

View the article online for updates and enhancements.

Extracting the gamma-ray source-count distribution below the Fermi-LAT detection limit with deep learning

A. Amerio,[a,b,*] A. Cuoco[b,c] and N. Fornengo[b,c]

[a]Instituto de Física Corpuscular (IFIC), University of Valencia and CSIC,
  Calle Catedrático José Beltrán 2, 46980 Paterna, Spain
[b]Department of Physics, University of Torino,
  Via P. Giuria 1, 10125 Torino, Italy
[c]Istituto Nazionale di Fisica Nucleare, Sezione di Torino,
  Via P. Giuria 1, 10125 Torino, Italy

E-mail: aurelio.amerio@ific.uv.es, alessandro.cuoco@unito.it,
  nicolao.fornengo@unito.it

Abstract. We reconstruct the extra-galactic gamma-ray source-count distribution, or dN/dS, of resolved and unresolved sources by adopting machine learning techniques. Specifically, we train a convolutional neural network on synthetic 2-dimensional sky-maps, which are built by varying parameters of underlying source-counts models and incorporate the *Fermi*-LAT instrumental response functions. The trained neural network is then applied to the *Fermi*-LAT data, from which we estimate the source count distribution down to flux levels a factor of 50 below the *Fermi*-LAT threshold. We perform our analysis using 14 years of data collected in the $(1, 10)$ GeV energy range. The results we obtain show a source count distribution which, in the resolved regime, is in excellent agreement with the one derived from cataloged sources, and then extends as $dN/dS \sim S^{-2}$ in the unresolved regime, down to fluxes of $5 \cdot 10^{-12}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The neural network architecture and the devised methodology have the flexibility to enable future analyses to study the energy dependence of the source-count distribution.

---

*Corresponding author.

Contents

1 Introduction

A substantial portion of the gamma-ray radiation that we observe from Earth is produced inside our own galaxy, mostly along the galactic plane. At high Galactic latitudes, however, the gamma-ray emission is mostly of cosmological origin. This extragalactic background radiation (EGB) is the sum of the emission from all the extragalactic gamma-ray sources [1, 2]. Most of the EGB is originated by various types of astrophysical sources which are seen from our viewpoint as point sources. A relevant observable is therefore their differential source-count distribution dN/dS which counts the number of sources at a given integral source flux S (for gamma-ray energies in a specific interval). For bright sources, which are observationally identified and therefore cataloged, the source-count distribution is directly measured by counting the objects in the catalog, at least above the threshold for which the catalog detection efficiency is equal to 100%. Below this flux threshold, where the detection efficiency is less than 100%, dedicated Monte Carlo simulations are required to accurately determine

the efficiency and use it as a correction to reconstruct the true underlying dN/dS [3–6]. For all those sources which are too faint to be resolved, their cumulative distribution of photons in the sky defines an almost isotropic cosmic field, conventionally called the isotropic diffuse gamma-ray background (IGRB) or, more precisely, the unresolved gamma-ray background (UGRB). Even though individual sources below the detector flux-threshold cannot be individually seen, it has been shown [7–10] nevertheless, that it is possible to infer their source-count distribution even in this regime, looking at the collective effects of these unresolved sources. This technique, called pixel-count distribution (or 1-point PDF) and pioneered for gamma-rays in [7], has been improved in [8–10] by employing a pixel-dependent approach, in order to fully explore all the available information and to incorporate the morphological variation of the gamma-ray emission components. Ref. [8] used the first 6 years of *Fermi*-LAT [11] data to measure the dN/dS for photons in the energy range (1,10) GeV down to fluxes about one order of magnitude below the *Fermi*-LAT detection threshold. In [9, 10] the same technique was used to extend the measurement of the dN/dS to several energy bands between 1 and 171 GeV, thus providing information on the energy dependence of the source-count distribution. This technique has then been used to study the contribution of individual source classes of emitters, in particular blazars [12–15]. The same methodology has also been used to characterize the properties of the unresolved sources in the Galactic Center region in relation to a possible signal from dark matter annihilation [16–25]. A further methodology called Compound Poisson Generator has also been developed [26], in order to handle biases possibly present in the 1-point PDF method.

In this paper we update the measurement of the dN/dS below the detection threshold to the increased statistics offered by 14 years of *Fermi*-LAT data. However, differently from refs. [8, 9], we adopt here a method based on machine learning techniques to obtain the dN/dS below the detection threshold. A similar methodology has also been used recently to investigate gamma-ray unresolved sources close to the Galactic center [27–29]. We train a convolutional neural network (CNN) on synthetic gamma-ray maps, built from a wide variety of source-count distributions and then apply the trained CNN to the 14-year *Fermi*-LAT map for photon energies in the (1,10) GeV band. We show that the CNN is able to reconstruct the dN/dS, thus obtaining an updated result which is fully compatible with the one obtained in [8]. The methodology presented here is also meant to be a proof of principle for the adoption of a CNN to the reconstruction of the source-count distribution of the extragalactic sky, with the future aim of properly investigating additional features of the dN/dS, like energy correlations and the presence of additional components which could be traced to dark matter. An advantage of the CNN method is that it avoids the need to calculate complicated and numerically demanding likelihoods, as in the 1-point PDF method. Furthermore, we will describe an improved version of the treatment of a CNN on a spherical domain which will further optimize the computational aspect.

Data selection and *Fermi*-LAT map generation is discussed in section 2, while section 3 discusses in detail how synthetic maps for the CNN training and validation are constructed. Section 4 discusses the neural network architecture that we use and its implementation, including validation and error estimation. Section 5 discusses the analysis of the *Fermi*-LAT map with the trained CNN and presents the ensuing results for the dN/dS. Section 6 gives our conclusions.

Healpix order	6, 7
Weeks	$9-745$
Emin	$1\,\mathrm{GeV}$
Emax	$10\,\mathrm{GeV}$
Instrument Response Functions (IRFs)	`P8R3_SOURCEVETO_V3`
EVCLASS	2048 (Source Veto)
EVTYPE	1 (Front)
ZMAX	90

Table 1. Fermi Tools settings used for the 14-year data set analysis.

2 Data selection

We process the *Fermi*-LAT data through the Fermi Tools suite [30] to produce a full-sky map of photon counts for gamma-ray energies in the (1–10) GeV range. We choose this range as a suitable balance between high statistics (which occurs at lower energies) and good angular resolution (which is better the higher the energy) of the detector, as well as being able to confront our results with those of ref. [8].

We consider the first 14 years of data collected by the *Fermi*-LAT (specifically, from week 9 to week 745 of operation), and Pass 8 event selection [31, 32]. We adopt the `P8R3_SOURCEVETO_V3` instrument response functions (IRFs), event class (EVCLASS) 2048 (`SOURCEVETO`) and event type (EVTYPE) 1 (`FRONT`). We used standard quality selection criteria, i.e. `DATA QUAL==1` and `LAT CONFIG==1`. The atmospheric gamma rays from the Earth limb emission is removed by adopting a cut on the maximum zenith angle ZMAX of 90 degrees. `FRONT` events refer to gamma-rays detected by the first layers of the tracking module of the *Fermi*-LAT detector, and are optimised in order to have a better angular resolution, which is a crucial requirement for our analysis. The `SOURCEVETO` class of events achieves good suppression of the charged cosmic-ray (CR) background while still retaining a large event statistics. All those settings are summarised in table 1. Finally, the maps are constructed with equal-area pixels by adopting the Healpix [33] pixelisation scheme of order $n = 6$ ($N_{\mathrm{side}} = 64$ resolution, which corresponds to $0.92°$ pixel side) although for testing we will also use maps with order $n = 7$ ($N_{\mathrm{side}} = 128$). The photon count map, in counts per pixel, for the 14 year data set is shown in figure 1. This is the map to which we will apply the CNN to extract the source-count distribution of gamma-ray sources.

Using the *Fermi*-LAT tools we further extract the *Fermi*-LAT exposure map and the point-spread-function (PSF) for photon energies in (1, 10) GeV, which will be needed to construct the synthetic maps used to train and validate the CNN. Some care is required, since the PSF is a rapidly varying function of energy. Following previous works [8, 34], we build the (1, 10) GeV mean PSF by averaging the energy-dependent PSF in the (1, 10) GeV energy range with a $E^{-2.4}$ weight, corresponding to the approximate energy spectrum of the high Galactic latitude gamma-ray sky. The average exposure map in the (1, 10) GeV energy bin and the averaged PSF profile are shown in figure 2 and in figure 3, respectively. The *Fermi*-LAT exposure slightly changes with energy, although its variation inside our energy bin is small.

A further dataset we will use is the catalog of resolved sources. In particular, we will employ the recently released Data Release 3 of the fourth *Fermi*-LAT catalog of gamma-rays sources (4FGL-DR3) [35], which increments the previous release to 12 years of data
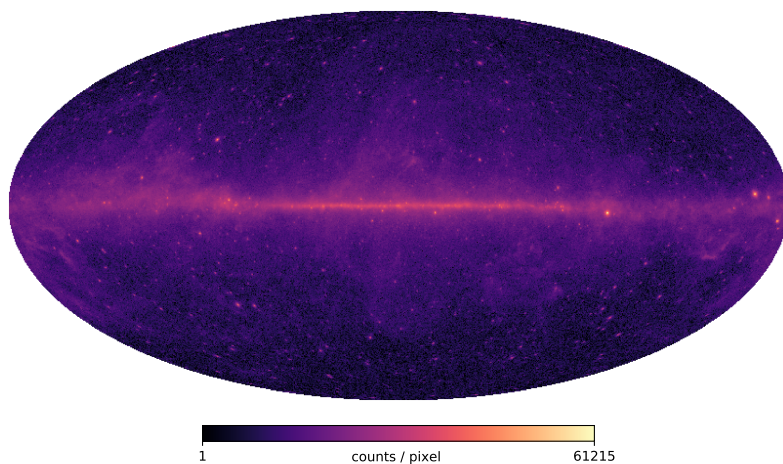
Figure 1. *Fermi*-LAT photon-counts map in units of counts per pixel in the 1–10 GeV energy range for the 14-years dataset, at the Healpix resolution $N_{\rm side} = 128$ ($n = 7$).
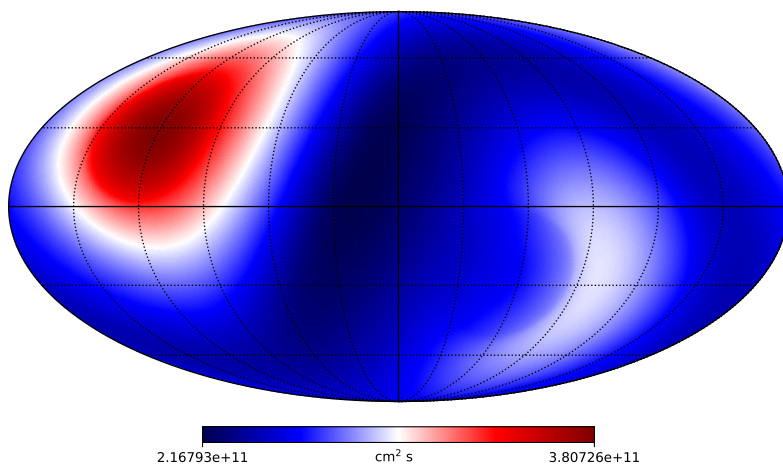

Figure 2. *Fermi*-LAT mean exposure map in the 1–10 GeV energy range for the 14-years data set, in units of $\rm cm^2\,s$, at the Healpix resolution $N_{\rm side} = 128$ ($n = 7$).

taking and contains 6658 resolved sources. From this catalog, we obtain the (1–10) GeV source-count distribution in the resolved regime, shown in figure 4. The detailed procedure with which we obtain the dN/dS from the list of sources is described in appendix B of [8]. Here and through the rest of the manuscript S will always refer to an integral flux in the range 1–10 GeV. The blue points refer to the regime in which the catalog is fully efficient in detecting sources, while the gray points denote sources for which the *Fermi*-LAT sensitivity to point sources starts degrading. Our aim is to infer the behaviour of the source count in the unresolved regime, i.e., for source fluxes below $S_{\rm th} \sim 2 \cdot 10^{-10}\,\rm cm^{-2}\,s^{-1}$ by adopting a CNN technique. We stress that in the following the dN/dS of resolved sources will be used only for comparison with the result of the CNN analysis, which are completely independent from it. Thus, the fact that we use 14 years of data while the catalog is based on 12 years does not have an impact on the analysis.
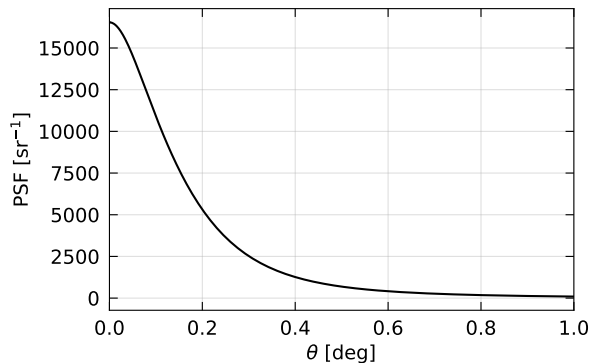
Figure 3. Average *Fermi*-LAT point spread function (PSF) for energies in $(1, 10)$ GeV.

3 Synthetic map generation

In order to instruct the CNN to extract the dN/dS from the *Fermi*-LAT photon-count map, we need to produce synthetic analogs of the map, constructed from a wide selection of dN/dS and for the same characteristics (detector exposure and PSF, known physical foregrounds) of the data we will be using. These maps will then be partly used to train a CNN and partly used to validate the CNN, to determine its performance and to estimate the uncertainty on the reconstructed dN/dS. Finally, we will apply the trained CNN to the *Fermi*-LAT data.

The gamma-ray maps are modeled as the sum of three contributions:

- A population of gamma-ray point sources. Since our analysis will be restricted to the high galactic latitude region (in order to minimize in the analysis the impact of the galactic foreground emission, as discussed in the following sections), these sources will mostly be of extra-galactic origin and can therefore be assumed to be distributed homogeneously across the sky. The high galactic latitude sky also contains a population of millisecond pulsars which can also be approximately assumed to be isotropically distributed [36], and thus are ideally included in our phenomenological dN/dS. In terms of photon flux, the sources are extracted, for each map realization, from the input dN/dS distributions, modeled as in section 3.1.

- Diffuse gamma-ray emission from the Milky Way. Galactic emission is the dominant contribution at low galactic latitudes, with a declining intensity at high latitudes. We model the galactic emission as discussed in section 3.2, by adopting a template model normalized to the *Fermi*-LAT data. Stability of the results against foreground modeling will be tested through the adoption of different templates.

- An isotropic background, which integrates all source emissions which are too faint to be processed by the CNN and possible truly diffused emission mechanisms, such as gamma-rays from cosmological cascades from ultra high energy cosmic rays or a possible contribution from dark matter annihilation or decay, as well as a residual instrumental CR background contamination. This term is modeled as a constant.

A flux map $\mathcal{M}$ will therefore be obtained as the sum of three contributions:

$$\mathcal{M} = \mathcal{S} + A_{\text{gal}} \cdot \mathcal{G} + F_{\text{iso}} \tag{3.1}$$

where $\mathcal{S}$ is the map due to the underlying distribution of sources (obtained from a dN/dS model as described in the next subsection), $\mathcal{G}$ is a template map for the foreground emission and F_{iso} is a constant that denotes the isotropic component discussed above. The constant A_{gal} is a normalization constant for the galactic foreground template.

The count map $\mathcal{N}$ is then obtained by multiplying the flux map by the *Fermi*-LAT exposure map $\mathcal{E}$, depicted in figure 2, and by taking into account the steradian-to-pixel conversion factor. Specifically, the counts in pixel a are obtained as:

$$\mathcal{N}_a = \mathcal{M}_a \cdot \mathcal{E}_a \cdot \frac{4\pi}{N_{\mathrm{pix}}} \tag{3.2}$$

where the number of pixels in the Healpix scheme is $N_{\mathrm{pix}} = 12 \cdot N_{\mathrm{side}}^2$ with $N_{\mathrm{side}} = 2^n$ with n the pixeling order.

Since the exposure slightly changes with energy, the above procedure could be improved by performing a convolution in energy of the photon flux with the exposure, instead of directly multiplying the two quantities as performed in eq. (3.2). However, the gradient in energy of the exposure in the $(1, 10)\,\mathrm{GeV}$ energy range is small, and the two procedures would provide very similar results. We, thus, use the simpler procedure outlined above. Eq. (3.2) also implies that we do not adopt energy dispersion. In the energy range 1–$10\,\mathrm{GeV}$ the energy dispersion is about 10% [31] and may induce normalization changes of similar size. Thus, as explained in the next section, we will re-derive the galactic foreground normalization with a dedicated fit in order to avoid possible biases in the normalization.

From the count map $\mathcal{N}$, which contains the number of photon counts in each pixel as determined by the model, we finally produce a Poisson realization $\mathcal{N}_\mathcal{P}$, by extracting in each pixel a photon count from a Poisson distribution with mean equal to the photon number of the corresponding pixel in $\mathcal{N}$. $\mathcal{N}_\mathcal{P}$ constitutes our synthetic map realization, which mimics a count map of an experiment with the *Fermi*-LAT specifications, where the underling model is the one of eq. (3.1).

In the next sections we discuss in more detail how the components of eq. (3.1) are modeled.

3.1 Differential source-count distribution

Following [8], the differential source-count distribution dN/dS is modeled as a multi-break power-law (MBPL). A MBPL with N_b breaks located at fluxes S_j $(j = 1, 2, \ldots, N_b)$ is defined as:

$$\frac{dN}{dS} = A_S \times \begin{cases} \left(\frac{S}{S_0}\right)^{-n_1}, & S > S_1 \\[2mm] \left(\frac{S_1}{S_0}\right)^{-n_1+n_2}\left(\frac{S}{S_0}\right)^{-n_2}, & S_2 < S \leqslant S_1 \\ \vdots & \\ \vdots & \\ \left(\frac{S_1}{S_0}\right)^{-n_1+n_2}\left(\frac{S_2}{S_0}\right)^{-n_2+n_3}\cdots\left(\frac{S}{S_0}\right)^{-n_{N_b+1}}, & S \leqslant S_{N_b} \end{cases} \tag{3.3}$$

where S_0 is a reference flux value which we choose to be $S_0 = 3 \cdot 10^{-8}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, and n_j are the power-law indices. S denotes the source flux in the energy interval $(1, 10)\,\mathrm{GeV}$. The dN/dS distribution is normalized with an overall factor A_S.

Parameter	Prior	Range
A_S	log-flat	$[1, 15]$
$F_{\rm iso}$	log-flat	$[0.5 \cdot 10^{-7}, 7.0 \cdot 10^{-7}]$
S_1	log-flat	$[3 \cdot 10^{-9}, 5.0 \cdot 10^{-8}]$
S_2	log-flat	$[5 \cdot 10^{-11}, 3.0 \cdot 10^{-9}]$
S_3	log-flat	$[5 \cdot 10^{-12}, 5 \cdot 10^{-11}]$
n_1	flat	$[2, 4]$
n_2	flat	$[1.80, 2.15]$
n_3	flat	$[1.5, 2.5]$
n_4	flat	$[1.0, 3.0]$

Table 2. Ranges of the parameters and their sampling priors adopted in the map generation with $N_b = 3$ breaks. The normalization A_S is in units of $10^7\,{\rm cm^2\,s\,sr^{-1}}$. $F_{\rm iso}$ is in units of ${\rm cm^{-2}\,s^{-1}\,sr^{-1}}$. The breaks S_1 S_2 and S_3 are given in units of ${\rm cm^{-1}\,s^{-1}}$. All other quantities are dimensionless.

For definiteness, in this work we adopt 3 breaks and 4 slopes. The ranges in which the parameters are varied to produce different source-count distributions is reported in table 2. The ranges have been chosen in order to intercept the expected features of the dN/dS and are broad enough to allow for a good variability in the modeling of the source counts, in order to expose the CNN to a wide set of options. In fact, the prior ranges of the three break positions are chosen relative to the *Fermi*-LAT threshold for resolving point sources, $S_{\rm th} \sim 2 \cdot 10^{-10}\,{\rm cm^{-2}\,s^{-1}}$, such that S_1, S_2, S_3 are respectively larger, across and below $S_{\rm th}$. The flux interval over which we will try to reconstruct the dN/dS is set as $[5 \cdot 10^{-12}, 1 \cdot 10^{-7}]\,{\rm cm^{-2}\,s^{-1}}$. The upper limit of this interval is set by the brightest sources in the catalog, while the lower limit is set at about 1.6 orders of magnitude below the *Fermi*-LAT threshold for resolving sources. Sources fainter than the lower limit of this interval are assumed to just contribute to the isotropic component. This lower limit, indeed, is also close to the theoretical sensitivity given by the flux of a point source contributing exactly one photon, which we calculate as $F_{\rm sens} = 1/\mathcal{E}_{\rm av} = 3.74 \cdot 10^{-12}\,{\rm cm^{-2}\,s^{-1}}$ where $\mathcal{E}_{\rm av}$ is the average exposure in the region $|b| > 30°$ where we perform the analysis.

The prior ranges for the power-law slopes are anchored to the results from the 4FGL catalog in the resolved regime, and allow for progressively wider variability moving toward the unresolved regime. Since catalog data exhibit a source count compatible with S^{-2}, except for large fluxes, where the slope increases, we adopt priors for n_1 and n_2 around these behaviours. For n_3 and n_4, we progressively enlarge the prior ranges, in order to allow for more variability especially in the sub-threshold (unresolved) regime we are interested in. An example of a dN/dS generated from this procedure is shown in the right panel of figure 4, which also illustrates the notation.

The maps $\mathcal{S}$ are generated according to the following procedure:

1. A dN/dS is selected by random sampling of its parameters in the intervals and with the priors defined in table 2;

2. The number N_i of sources in the ith flux bin is calculated as $N_i = 4\pi (S_{i,\max} - S_{i,\min}) \, dN/dS|_i$ where $dN/dS|_i$ is the value at the center of the ith bin. The flux bins are determined by subdividing the interval of interest $[5 \cdot 10^{-12}, 1 \cdot 10^{-7}]\,{\rm cm^{-2}\,s^{-1}}$ in 10^3 log-intervals per decade;
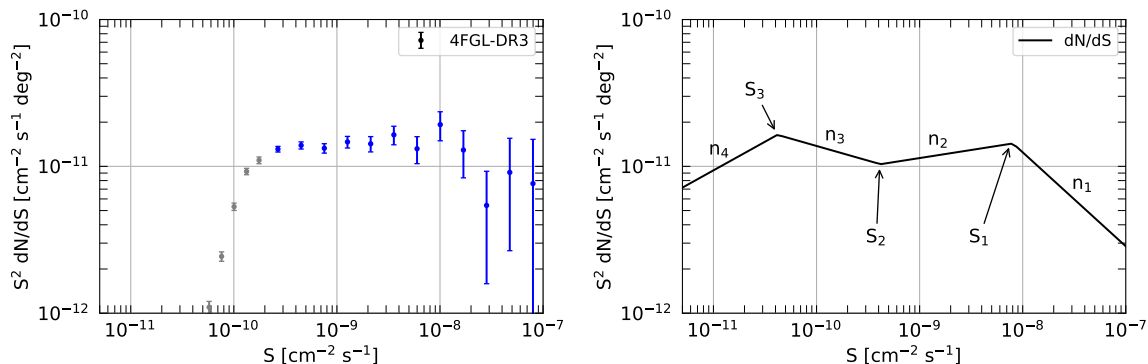
Figure 4. Left: the 1–10 GeV dN/dS for resolved *Fermi*-LAT sources, obtained from the 4FGL-DR3 [35] catalog. Right: an example of dN/dS with 3 breaks S_j ($j = 1, 3$) and four slopes, where n_i ($i = 1, 4$) is the inclination of the slope.

3. A Poisson extraction on N_i determines the number of sources $N_{P,i}$ in each flux interval;

4. The sources are randomly positioned on the sphere by uniformly sampling $\cos\theta$ in $[-1, 1]$ and φ in $[0, 2\pi]$, θ and φ being the spherical coordinates. The selected angular position is then converted into a pixel position in the Healpix pixeling scheme with the `ang2pix` routine. An example of a point-source map is shown in the upper panel of figure 5;

5. Finally, convolution with the *Fermi*-LAT PSF shown in figure 3 is applied to each point source. Since sources are assumed to be point-like, we simply replace each point source with an extended circular object with a radial profile equal to the PSF, and whose total flux is equal to the one of the original source. Since the image formation process is linear in the intensity, in case of overlapping sources (either because of their spatial position, or because of the PSF smearing) the resulting flux is additive. An example of flux map with the PSF applied is shown in the lower panel of figure 5.

In order to avoid possible biases in the generated maps, due to the resolution of the angular binning, the maps generated with the procedure described above have been produced at a resolution higher than the one adopted for the analysis. We have therefore built our maps with $N_{\text{side}} = 128$. These maps have then been downgraded to $N_{\text{side}} = 64$ for the analysis.

3.2 Galactic foreground and isotropic background

The galactic foreground is implemented through the adoption of the *Fermi*-LAT template model `gll_iem_v07`.[1] The map of the integrated flux in the $(1, 10)$ GeV energy range is shown in figure 6. To check consistency and stability of our results, we will also adopt a different version of the galactic template, namely model `gll_iem_v05`.[1] The extragalactic isotropic background instead is modeled as a constant flux F_{iso}, which captures all the unresolved emission which is too faint to be associated to the emission modeled through the dN/dS (i.e., which refers to fluxes $S < 5 \cdot 10^{-12} \, \text{cm}^{-2} \, \text{s}^{-1}$, below our region of interest).

As shown in figure 6, the galactic foreground dominates at low latitudes, although it extends to high latitudes with a smaller flux. For these reasons, low latitudes are unsuitable
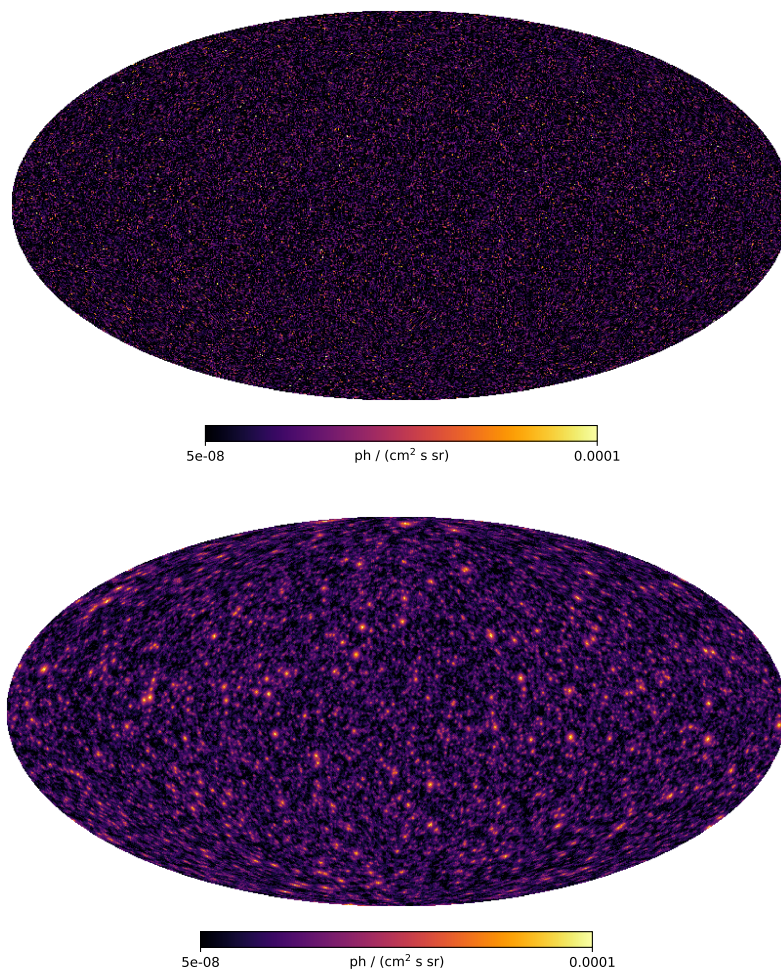
---

[1]

Figure 5. Simulated flux map of point sources extracted from a specific source-count distribution, before (upper panel) and after (lower panel) the PSF is applied, in units of photons/(cm^2 s sr). The maps are shown at Healpix resolution $N_{\rm side} = 128$ ($n = 7$).

for the study of the extra-galactic background. Our baseline analysis will therefore apply a mask to the maps to cover the dominant part of the galactic emission, for which we adopt a latitude cut $|b| < 30°$. In order to check the stability of the results against the foreground treatment, in section 5 we will show the results obtained with two additional latitude cuts: $|b| < 40°$ and $|b| < 50°$.

Given the various approximations we use, instead of using the nominal normalization of the galactic foreground models, we prefer to use a data-driven approach and normalize the galactic templates `gll_iem_v07` and `gll_iem_v05` directly to the observed *Fermi*-LAT map. This is realized by performing a maximum-likelihood Poisson fit, using as likelihood:

$$\mathcal{L} = \prod_{a=1}^{N_{\rm pix}} \frac{\lambda_a^{k_a} e^{-\lambda_a}}{k_a!} \tag{3.4}$$

where k_a is the number of photons in pixel a of the *Fermi*-LAT count map and λ_a the corresponding quantity from the model count-map, given by $\lambda_a = (A_{\rm gal}\,\mathcal{G}_a + F'_{\rm iso}) \cdot \mathcal{E}_a \cdot \frac{4\pi}{N_{\rm pix}}$,
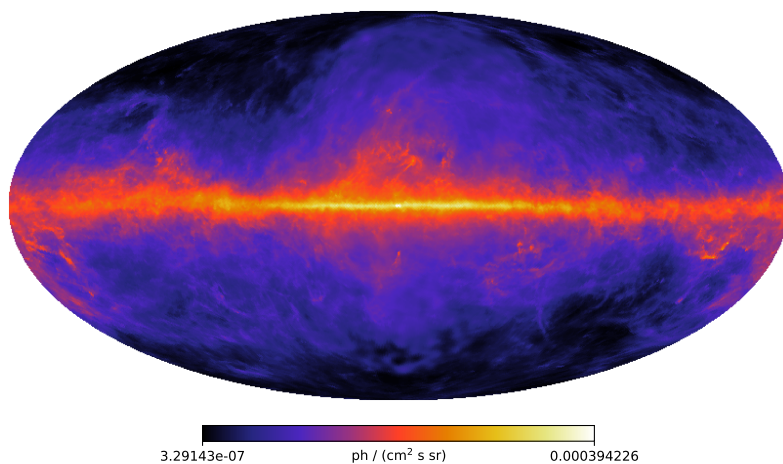
Figure 6. `gll_iem_v07` galactic foreground model [37], integrated in the $(1, 10)$ GeV energy bin, in units of photons/$(\mathrm{cm}^2\,\mathrm{s}\,\mathrm{sr})$, at the Healpix resolution $N_{\mathrm{side}} = 128$ $(n = 7)$.

Foreground template	Latitude cut	A_{gal}	F'_{iso}
`gll_iem_v07`	$\|b\| < 30°$	0.888 ± 0.005	$(4.91 \pm 0.04) \cdot 10^{-7}$
`gll_iem_v07`	$\|b\| < 40°$	0.874 ± 0.008	$(4.90 \pm 0.06) \cdot 10^{-7}$
`gll_iem_v07`	$\|b\| < 50°$	0.838 ± 0.013	$(5.05 \pm 0.08) \cdot 10^{-7}$
`gll_iem_v05`	$\|b\| < 30°$	1.030 ± 0.006	$(2.57 \pm 0.06) \cdot 10^{-7}$

Table 3. Values of the A_{gal} and F'_{iso} parameters obtained from a fit to the *Fermi*-LAT map, for different galactic foreground template models and latitude cuts. F'_{iso} is in units of $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}\,\mathrm{sr}^{-1}$.

where A_{gal} and F'_{iso} are the two fit parameters. In performing the fit we mask the sky with the low-latitude cut and with a $2°$ circular mask around each of the sources in the 4FGL catalog. We added a prime on the F'_{iso} derived from the maximum-likelihood fit above to explicitly distinguish it from the F_{iso} introduced earlier. Confronting with eq. (3.1), the F'_{iso} term denotes the sum of the purely isotropic contribution F_{iso} and the contribution of sources, $\mathcal{S}$ term, below the 4FGL detection threshold.

The results of the fit are shown in table 3, for three latitude cuts for the `gll_iem_v07` template and for $|b| < 30°$ for the alternative `gll_iem_v05` model. We see that both the normalisation parameter A_{gal} and F'_{iso} are consistent among themselves for a fixed galactic foreground model, implying consistency with respect to the latitude cut. The model `gll_iem_v05` instead requires a larger normalization, with an ensuing reduced value for F'_{iso} by a factor of 2. This reflects a certain amount of degeneracy which is present at high galactic latitude between the foreground emission and the isotropic emission, allowed by the uncertainties in the foreground model. The normalization of the `v07` model different from 1 is related to the fact that we do not include energy dispersion. The `v05` model, instead, has been created originally without including the effect of energy dispersion,[2] and indeed we obtain a normalization compatible with 1.
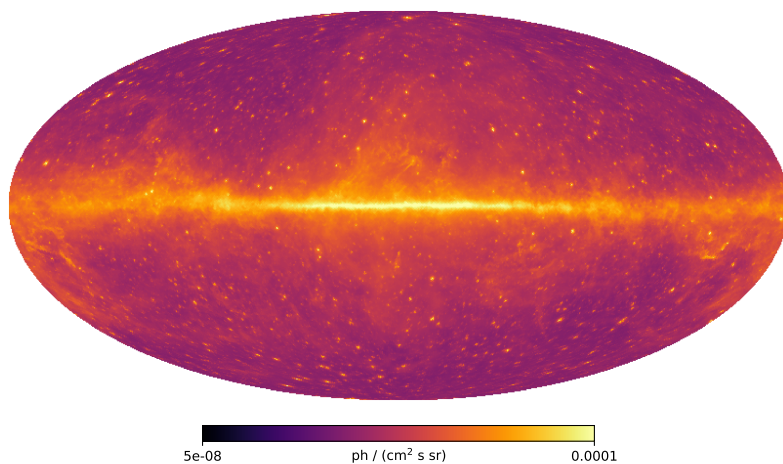
---

[2]https://fermi.gsfc.nasa.gov/ssc/data/access/lat/BackgroundModels.html.

Figure 7. The same as in figure 5, with the galactic foreground model normalized with $A_{\rm gal} = 0.886$ and an isotropic background component $F_{\rm iso} = 4 \cdot 10^{-7}\,{\rm cm}^{-2}\,{\rm s}^{-1}\,{\rm sr}^{-1}$. Units are photons/$({\rm cm}^2\,{\rm s}\,{\rm sr})$.

Throughout the rest of the analysis we will fix the normalization parameter $A_{\rm gal}$ to the value determined from the above fit. The values of $F'_{\rm iso}$ are, instead, used to define the range of the prior for $F_{\rm iso}$ in table 2 and *a posteriori* to check consistency of the results. One illustration of a flux map obtained as the sum of all components is shown in figure 7. Instead of fixing $A_{\rm gal}$, in principle, we could also leave it as a free parameter and instruct the network to predict its value. We found, however, that the above procedure constrains $A_{\rm gal}$ very precisely and it is thus preferable. Furthermore, we also found that fixing $A_{\rm gal}$ provides a better stability of the neural network.

Nonetheless, to better asses the impact of fixing the $A_{\rm gal}$ value, we also performed a further test using different values in the range $(0.78, 1.0)$ (i.e., a range centered on our fiducial value 0.888 ± 0.005, thus encompassing many standard deviations away from its central value.) We found that the dN/dS reconstructed by the CNN is largely unaffected by the employed value of $A_{\rm gal}$, while the specific $A_{\rm gal}$ choice mostly affect $F_{\rm iso}$, as expected because of the high galactic latitude degeneracy between the foreground emission and the isotropic emission. Example plots are shown in the appendix A.4.

3.3 Count maps

With a flux map $\mathcal{M}$ available, we can now produce a synthetic photon-count map $\mathcal{N}$ as described in eq. (3.2). One example is shown in figure 8. In order to generate maps compatible with a counting experiment like *Fermi*-LAT, we then add Poisson noise to the map performing a Poisson realization of the number of photons in each pixel of the map $\mathcal{N}$, and we dub this final map as $\mathcal{N}_{\mathcal{P}}$ count map. One example is shown in figure 9, which also shows the same map with the latitude cut applied: these are the type of maps that we use to train and validate the neural network. We have generated in total 1 million maps, 90% of which have been used for training and 10% for validation. The maps have been stored as a `TFRecord` dataset[3] for use with Tensorflow [38]. The maps have been generated in Healpix format with $N_{\rm side} = 128$ resolution (order $n = 7$) and then downsampled to $N_{\rm side} = 64$ (order $n = 6$) for the training and validation (this resolution corresponds to 49152 pixels).
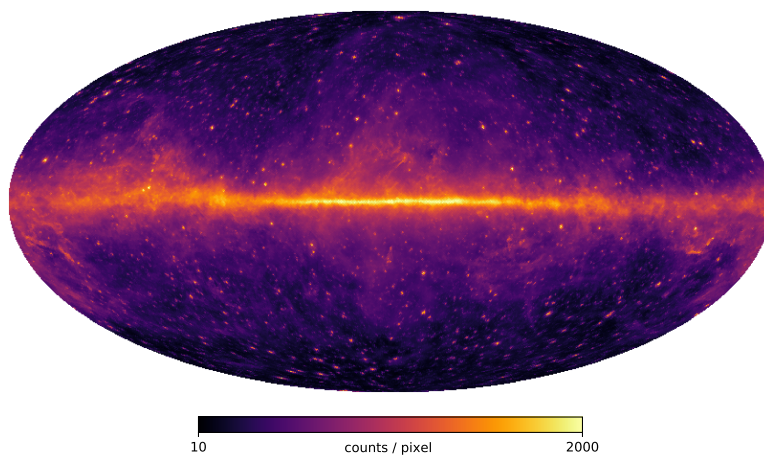

Figure 8. Simulated count map in units of counts per pixel, at the Healpix resolution $N_{\rm side} = 128$ ($n = 7$).
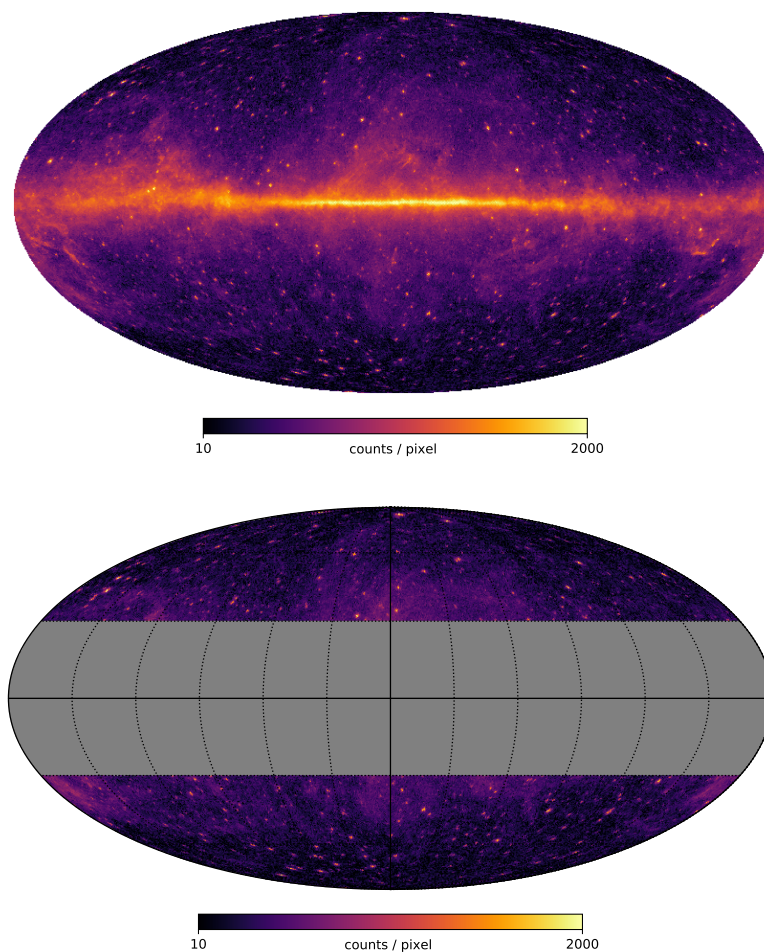

Figure 9. Simulated map in units of counts per pixel after adding Poisson noise from the synthetic model map of figure 8, at the same Healpix resolution $N_{\rm side} = 128$ ($n = 7$). The lower panel shows the map with a $|b| < 30°$ latitude cut.

The training of the CNN and the subsequent analysis of the *Fermi*-LAT data can be performed equivalently with maps in units of photon-count or in units of flux. Nonetheless, we will work with maps in units of flux, which are easily obtained from the $\mathcal{N}_\mathcal{P}$ count maps by using the exposure map and by inverting eq. (3.2). The reason to convert the $\mathcal{N}_\mathcal{P}$ count maps in units of flux is motivated by the fact that count maps contain a spurious large-scale angular dependence introduced by the exposure map (see figure 2) which might confuse the CNN, while this potential issue is avoided using flux maps. Finally, to further ease the learning process of the CNN we also apply a foreground subtraction procedure, i.e., to the above flux maps we subtract the original galactic template flux map. In this way the CNN can focus on learning the dN/dS, seeing the exposure and galactic foregrounds only as sources of noise, instead of trying to learn their detailed structure. These foreground-subtracted flux maps are the maps we use to train the CNN.

4   Neural network architecture and training

In this section we describe the architecture of the CNN we adopt and how data are handled. The input of the CNN are synthetic foreground-subtracted flux maps, and the output vector $\hat{y}_i = (dN/dS|_j, F_{\rm iso})$ contains the reconstructed $dN/dS|_j$ in 20 flux bins in $[5 \cdot 10^{-12}, 1 \cdot 10^{-7}]\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, plus the reconstructed value of $F_{\rm iso}$. The index i, thus, has values $i = 1, \ldots, 21$. A possible alternative would be to choose as output the breaks and the slopes of the dN/dS, i.e., the same parameters used to model the input dN/dS. Nonetheless, we found a better stability of the CNN using as output directly the dN/dS function as discretized above, and we will thus use this as output.

As further described below, we will use an approach in which we will let the CNN determine also the errors on the output parameters. The output vector will thus be effectively doubled. Beside the $\hat{y}_i$, we thus have also $\hat{\sigma}_i$ the output vector of the errors.

Since we are dealing with sky maps, we can approach the problem from the point of view of image analysis. However, the image here is in spherical projection. We therefore need to adopt a method able to deal with this situation.

4.1   Spherical neural networks

In the last years, the machine learning community has realised the importance of developing neural network architectures on complex topologies. One of the most interesting non-trivial domains is that of a sphere, since it cannot be mapped on a plane without introducing distortion. This is the case we are confronting here, since we are dealing with sky maps. There are many efforts to implement CNN architectures that can do convolutions on the sphere [39–43], and while there have been many proposals, there is not yet agreement on a standard algorithm.

In dealing with information extraction from spherical images, one has to cope with a set of issues: i) the algorithm needs to properly deal with the topology of the sphere without distorting the image or losing information; ii) the algorithm has to be computationally efficient; iii) it would be desirable to be able to use the same neural network architectures which already work for flat images, which have been proven to be very efficient and effective.

In fact, a full spherical architecture is usually slower than a flat architecture when applied to images of the same size, and it often happens that it is not possible to adopt pre-existing models without re-implementing them *ad-hoc*. On the other hand, the adoption of a flat-image architecture either introduces spherical distortions, due to the mapping of a

spherical image on a plane, or may loose some of the large scale information contained in the image.

In this work, we implemented a fast and reliable spherical architecture on the Healpix pixelisation of the sphere, taking inspiration from the idea proposed in [43] for the icosahedron. This method proved to be reliable when the information we need to retrieve from the maps is a small-scale effect, like in our case the distribution of sources, which are point-like (although smeared by the PSF) and isotropically distributed.

The method is based on the fact that, in the Healpix pixelisation scheme, each pixel has equal area and contains the information of the underlying field (in our case, the photon counts) averaged over that pixel. As discussed above, we train our neural network on maps of order $n = 6$ ($N_{side} = 64$), which contain 49152 pixels. Considering instead the Healpix sphere with base pixelisation of order $n = 0$ ($N_{side} = 1$), we can subdivide the map into 12 equal-area patches, each of which containing 4096 pixels. Each patch is an independent realisation of the underlying isotropic field and can be mapped into a 2D flat image. We call this parsing algorithm `map2patch`. Notice that this is not a re-sampling of the map: it's just a subdivision of the spherical image into 12 patches, in a way that preserves the area and content of each pixel. This allows us to use standard convolutional network techniques that work efficiently on flat images, while preserving the whole available statistics of the full map.

Convolution is performed on each patch separately. In order to be able to extract information from each patch in parallel, we define 3D convolutions, where the first two dimensions are the spatial dimensions, and the third dimension specifies the patch. In order to force the neural network to learn from each individual sky patch, we employ convolution filters of size $(N, N, 1)$, where N is the dimension of the filter. By doing so, we effectively apply 2D filters to the whole sky. Note that by employing such filter size, the output of each convolution will still have 12 separate images on the third axis, associated to each sky patch, while the number of channels will vary according to the number of filters employed. This technique will thus naturally scale for images with different color channels, such as RGB images, or spherical images for which several channels are available, as is the case where multiple energy bands are considered at once.

We stress that with our method we partially loose large-scale information and information stored at the border of the patches. Thus our algorithm would not be suitable in the case one is interested in recovering large-scale features. In our case, however, we are interested in exploring information which is statistically isotropic and/or lies at scales much smaller then the size of a patch. In this case our architecture is certainly viable and extremely efficient. Also, since we build our patch on the pixels of the Healpix pixelisation scheme, there is no spherical distortion in our images, unlike other architectures which employ the euclidean projection of the sphere.

This approach is extremely fast, since it leverages on optimised 3D convolutions. Furthermore the sphere-to-patch operation is performed asynchronously by the CPU while the accelerator is training the model, allowing for virtually no bottlenecks in the processing pipeline.

In order to cross-check our method, we also implemented the full-sky truly-spherical convolutions of [42], obtaining very similar results. In appendix A.5, we report some examples of the results and checks performed with a fully spherical convolution model. The two methods are therefore interchangeable for the type of information we want extract from the *Fermi*-LAT maps: however, our implementation allowed us to gain more than a factor of 10 in time of execution of the whole pipeline of training and validation. This convergence of

results with the fully spherical method of [42] gives us confidence on the reliability of our model, and the speed gain sets the basis for a future efficient extension of our analysis to several energy bins in the full *Fermi*-LAT energy range, in order to determine also the energy dependence of the dN/dS of unresolved sources.

4.2 Data pre-processing

The maps are constructed with the procedure outlined in section 3. As explained more in detail there, even though we apply a latitude cut to the maps, in order to further reduce the impact of the galactic foreground (and as it is customary in analyses of the extragalactic gamma-ray background), we subtract the galactic foreground (normalized by the A_{gal} constant) from the input maps. This will also facilitate to some extent the training of the CNN, whose goal is to reconstruct the source-count distribution. Photon-counts maps are converted in units of flux, by using the exposure map. In order to reduce the variability of the inputs seen by the CNN and stabilize its behaviour, we take the logarithm of the flux values in the pixels and then map them in $[-1, 1]$ range. Finally, each map is parsed into 12 patches using the `map2patch` algorithm discussed in the previous sections.

Given the approximate S^{-2} behavior of the dN/dS, for the output vector, we reconstruct $y_i = S^2 dN/dS|_i \times 10^{11} \, \mathrm{cm^2 \, s \, deg^2}$ ($i = 1, 20$) and $y_{21} = F_{\mathrm{iso}} \cdot 10^7 \, \mathrm{cm^2 \, s \, sr}$, in order to work with (pure) numbers of order unity. We adopt this strategy in order to avoid that the mean square error could favour the optimisation of any flux bin over another, making the deviations more even across the range of S.

4.3 Network architecture

In recent years, we have seen the rise of many convolutional neural network models for computer vision tasks [44–46]. Modern convolutional neural network architectures can have a very large numbers of internal parameters. This is due to the complex links between each neural unit, which allows for these models to be highly expressive and effectively behave as universal approximators for many kinds of tasks [47]. Unfortunately this comes with a downside: these models are prone to fitting too well the training dataset, which would lead to poorer performance on new data. In the field of Machine Learning, this problem is known as "overfitting" [48]. In order to prevent overfitting, a possible solution is to train the model on a big enough dataset. This will render impossible for the NN to learn exactly the representation of the training set, as well as showing a richer feature space, which will result in better performance on any additional sample for which we may want to get a prediction, like the real case scenario with the Fermi photon counts map. Another solution is to apply regularisation layers which force the NN to learn a more general representation of the data, by promoting several metrics. Some of the most common regularisation techniques include L1 and L2 regularisation, batch normalisation, and dropout [49–52].

We choose as our convolutional neural network architecture the EfficientNet V2M model [53]. This model has been proven to be faster to train then other models for computer vision tasks, while retaining the same expressivity and predictive power. Furthermore, its low memory footprint helps in training with more data at the same time (larger batch sizes), which in turn reduces the risk of overfitting. In this work, following the original implementation of EfficientNet, we will adopt batch normalisation layers after each convolution.

We additionally employ a fairly new trainable implementation of dropout, called concrete dropout [54], before each convolution layer. Dropout [49] has been proven to be an

efficient way to reduce overfitting in a model, by randomly turning off neurons or connections inside a model. Unfortunately, the dropout probability for each dropout layer has to be either hand picked or optimised through grid search, which is practically unfeasible for very large models. If the dropout probability is too small, the regularisation effect will not be very significant, while on the other hand too high a dropout probability will result in slower training and worse overall performance (underfitting). We therefore adopt the implementation of dropout as a concrete Bernoulli distribution [55], which allows the self-learning of the optimal dropout probability, effectively saving a conspicuous amount of time while training and ensuring that the model will be less prone to overfitting.

4.4 Bayesian error estimation and cost function

In order to correctly estimate the error on a prediction, it is necessary to correctly estimate both the systematic error (sometimes called epistemic error [56]) and aleatory error components of a measurement. One way to estimate such components together is through a Bayesian approach, and in particular by defining a Bayesian neural network capable of estimating the posterior distribution of the desired observables. It has been shown [57] that a NN with arbitrary depth and non-linearities, with dropout applied before every weight layer, is mathematically equivalent to an approximation of a Bayesian model.

In a subsequent analysis [58] it has been shown how to estimate the first and second momenta (mean and variance) of the posterior distribution, under the assumption of a Gaussian likelihood of unknown variance. In particular [58] shows how to define a neural network capable of determining the appropriate variances of such a multivariate Gaussian. In our case, training a neural network with such an approach is advantageous, as the neural network will learn to weight less the component of the measurement at low fluxes, which we expect to be more uncertain, which helps in obtaining stable measurements for the test dataset.

Following [58] we define the heteroscedastic Gaussian negative log-likelihood:

$$NLL = \sum_{i=1}^{N} \left[\frac{(y_i^{\text{true}} - y_i^{\text{pred}})^2}{2\sigma_i^2} + \frac{1}{2}\log(\sigma_i^2) + \frac{1}{2}\log(2\pi) \right] \tag{4.1}$$

where we compare the true dN/dS and F_{iso} (transformed as discussed in section 4.2) used to generate the simulated map with the prediction of the neural network, and thus $N = 20 + 1$. For the training of the neural network, we omit the last term of eq. (4.1) which is an irrelevant constant. Regarding the errors, since the first term in eq. (4.1) is a decreasing function of the σ_i while the second is increasing, this ensures that there will be a minimum which will provide the optimal estimate for the σ_i themselves.

In principle, the above cost function can easily be generalized to account for the covariance among the different bins and F_{iso}. In practice, however, this implies estimating also the off-diagonal terms of a 21×21 symmetric matrix. Such matrix has to be inferred autonomously from data without any supervision, making the problem very challenging to optimise in this form. We thus refrain from attempting estimating the covariance. For CNN applications where the covariance between a set of parameters has been estimated, see e.g. refs. [27, 29].

4.5 Training and validation

We train our model using a Tensor Processing Unit (TPU) v3, kindly provided by the Google[TM] TPU Research Cloud. We adopt the Adam optimizer. We employ a cyclical
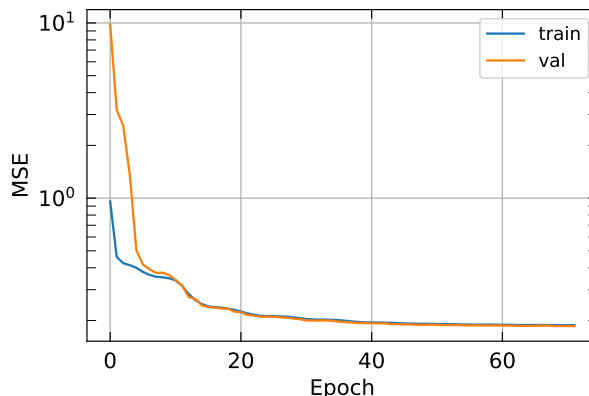
Figure 10. Epoch evolution of the MSE for the training (blue line) and validation (orange line) sets.

triangular learning rate [59] to avoid falling into local minima of the parameter space. We adopt a batch size equal to $128 \times 8 = 1024$ to fully exploit the computational capabilities of the TPU. We train the model for 72 epochs, for a total of 9 triangular learning rate cycles, with learning rate ranging from 10^{-6} to 10^{-3}. As already anticipated, we have generated 1 million maps, 90% of which have been employed in the training and 10% for validation.

To test the convergence of the model we evaluate the mean square error (MSE) defined as $MSE = \sum_{i=1}^{N} (y_i^{\mathrm{true}} - y_i^{\mathrm{pred}})^2$. In figure 10 we plot the MSE as a function of the training epoch for the training and validation dataset. We can see that, starting from epoch 10, the validation and training lines are close to each other, which means that our model is not overfitting the training dataset. We observe that the MSE reaches a plateau at around epoch 60, meaning that further training of our model would not lead to improvement in performance.

Once the CNN is fully trained, we extensively check the convergence of the reconstructed dN/dS and F_{iso} against their inputs, by using the maps of the validation set. Figure 11 shows a few illustrative examples of dN/dS, randomly generated according to the priors and parameters intervals listed in table 2, together with their reconstructions. The dashed black lines show the input dN/dS and the black solid lines denote the corresponding distributions reconstructed by the CNN. The input and reconstructed values of F_{iso} are also quoted. The colored bands refer to the 1σ and 2σ confidence intervals as estimated by the CNN. The CNN recovers a dN/dS which is compatible with the input. The error bands are large for high fluxes, due to the fact that the number of sources at those values of S is low and therefore a large statistical uncertainty is expected. The error band becomes large again at low fluxes, indicating that the CNN reaches a confusion limit: even though here the number of sources is very large, their faintness makes progressively harder to identify their number and a confusion between faint sources and the average isotropic component F_{iso} arises, the lower the valus of S becomes. This same behaviour also occurs in the 1point-PDF analysis of [8]. Concerning the value of F_{iso}, the reconstruction is very good, compatible with the input and with a small uncertainty.

4.6 Frequentist error estimation and bias

Besides the Bayesian error automatically provided as output by the CNN, we also estimate the error in an alternative frequentist-like way, in order to provide a cross-check. This second
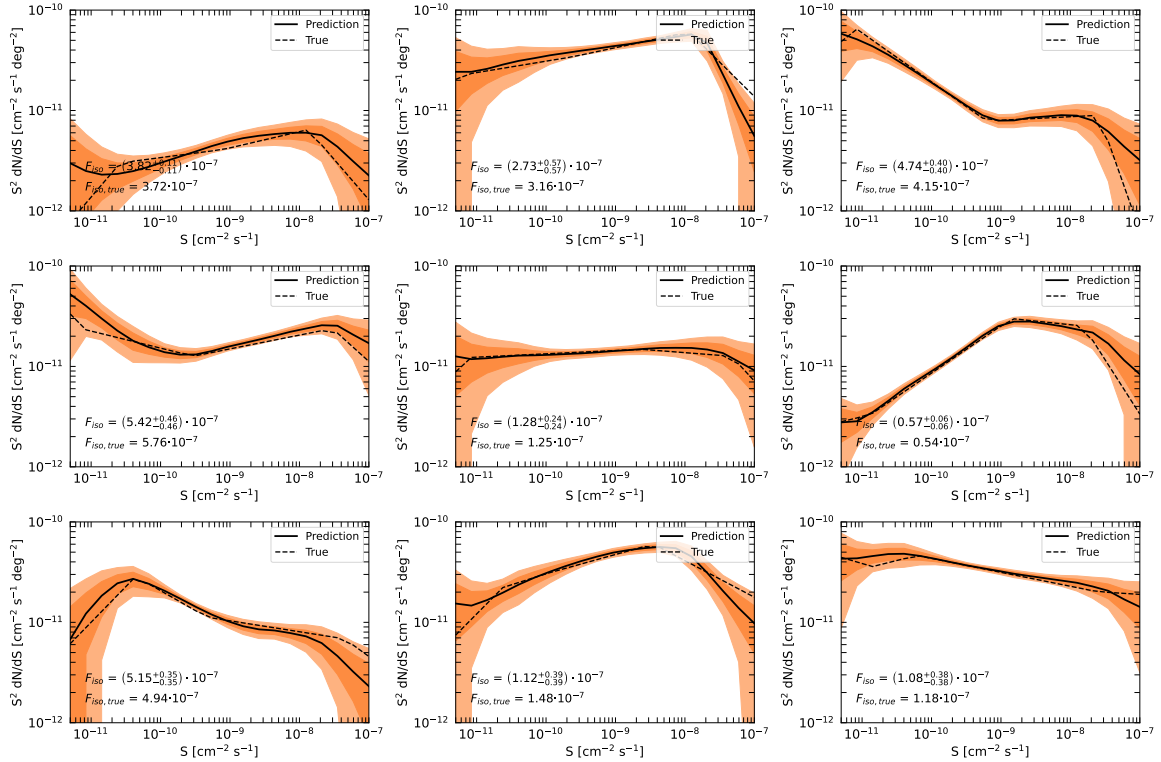
Figure 11. Some example dN/dS from the validation dataset. Each panel shows the input dN/dS and $F_{\rm iso}$ and the corresponding quantities reconstructed by the CNN. The colored bands indicate 1–2 σ uncertainty regions. See text for more details.

method gives as a further advantage also the possibility to estimate the amount of bias on the reconstructed dN/dS. To this aim, we have built, for each of the 20 bins in flux and for $F_{\rm iso}$, the histogram of the deviations between the true and reconstructed dN/dS and $F_{\rm iso}$. The results for 14 values of S across its interval of interest, and for $F_{\rm iso}$, are shown in figure 12 (the remaining histograms are similar in content and are not shown here for economy of space). The plot shows the distribution of models which exhibit a specific relative deviation (expressed as percentage of the true value). By using these distributions of the deviations (normalized to unity), we derived the 1σ and 2σ confidence level intervals on the reconstructed dN/dS and $F_{\rm iso}$. As anticipated in section 4.5, at low and large fluxes the distributions are wide, implying a poorer ability to reconstruct the underlying physical models, while in the flux interval $(1 \cdot 10^{-11}, 2 \cdot 10^{-8})\,{\rm cm}^{-2}\,{\rm s}^{-1}$ the 1σ error is below 30%, with an error below 20% for S in $(1 \cdot 10^{-10}, 1 \cdot 10^{-8})\,{\rm cm}^{-2}\,{\rm s}^{-1}$. The 1σ error on $F_{\rm iso}$ is also of the order of 20%. The last panel of figure 12 shows the histogram for the reconstructed value of the total photon flux: this is a sanity check and shows that the CNN correctly reproduces the normalization of the maps in terms of total counting rate. The uncertainty on the total flux is 5%. The 1σ (non symmetric) intervals for all the 20 flux bins are also reported in figure 13. A more direct comparison of the Bayesian and frequentist errors is provided in the results section.

The same histograms of figure 12 allow us also to check whether the reconstructed values are biased. We define the bias as the deviation between the median of the distribution and the input value. The values of the bias are shown, for each of the 20 values of flux, in figure 13 as squares. We see that for all fluxes, the bias is very close to zero, which implies
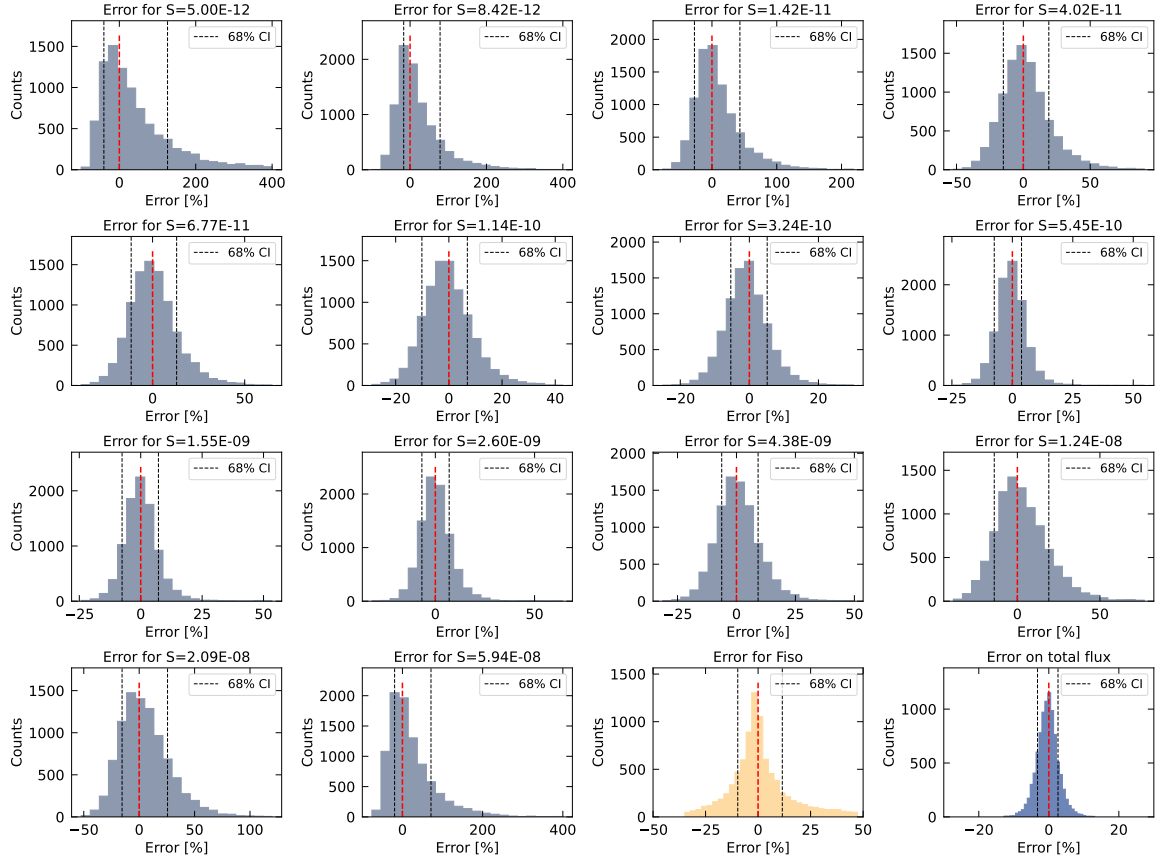
Figure 12. Histograms of the percentage deviations between the true and reconstructed dN/dS for 14 values of S across its interval of interest, and for F_{iso}. The bottom right panel shows the same, but for a derived quantity, namely the total flux. The dotted vertical black lines indicate the 68% confidence region.
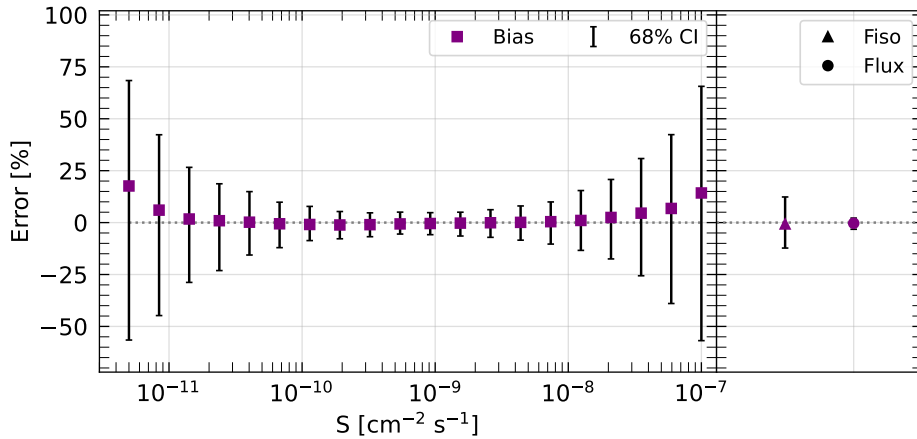

Figure 13. 68% error for the reconstructed dN/dS as function of S, for F_{iso} and for the total flux. Also shown is the bias given by the median of the histogram for each S bin (squares).
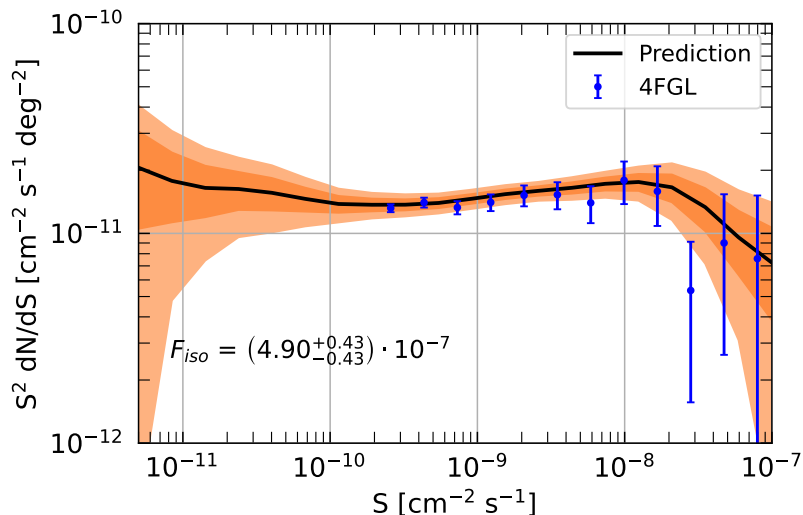
Figure 14. Reconstructed dN/dS and F_{iso} and their Bayesian errors, when the trained CNN is applied to the *Fermi*-LAT map. This is the main baseline result of the analysis.

that the reconstructed values not only are determined with good precision (since the error bar is small), but also without a bias toward larger or lower values. The only exception is a positive bias for very low fluxes, where the CNN meets its confusion limit, and for very large fluxes, where the low source-counts statistics produces a small overestimate of the underlying dN/dS. In both cases, however, the bias is small compared to the overall error.

5 Results

After the training and validation processes described in the previous section, we have applied the fully trained CNN to the 14 years *Fermi*-LAT data. The result of our baseline analysis, which adopts the `gll_iem_v07` galactic foreground and a latitude cut of $|b| < 30°$, is shown in figure 14. The blue line is the reconstructed dN/dS and the error bands are the Bayesian errors obtained as described in section 4.4. The blue points show the dN/dS of the resolved sources, obtained from the 4FGL-DR3 catalog. We can see that in the resolved limit, the CNN reconstructs a dN/dS fully compatible with the one derived from the catalog, but then extends it to the unresolved regime with a behaviour compatible with $dN/dS \sim S^{-2}$ down to the smallest flux considered of $S = 5 \cdot 10^{-12}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. Below $S \sim 10^{-10}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ the dN/dS exhibits a slight turn-up, although the feature has a very low statistical significance.

The result obtained here is also compatible with the one obtained in [8], as shown in figure 15. The difference between the two analyses, apart from the technique adopted to extract the dN/dS, stands in the fact that here we update the data set to 14 years of *Fermi*-LAT data collection and to improved detector response functions, as compared to [8]. Another difference is that in [8] the dN/dS was reconstructed with a broken power-law, while here we are allowing for a more flexible functional dependence (we are using 20 points in the flux interval shown in figure 14). The result is nevertheless quite compatible with a power law behaviour with $dN/dS \sim S^{-2}$ for mid-values of S, $dN/dS \sim S^{-3}$ for $S > 10^{-8}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and a slight decrease in the power-law index for fluxes $S < 10^{-10}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$.

In figure 16 we show the measurement for the Fermi map employing the frequestist estimation of the error. It can be seen that the error is very similar to the Bayesian one,
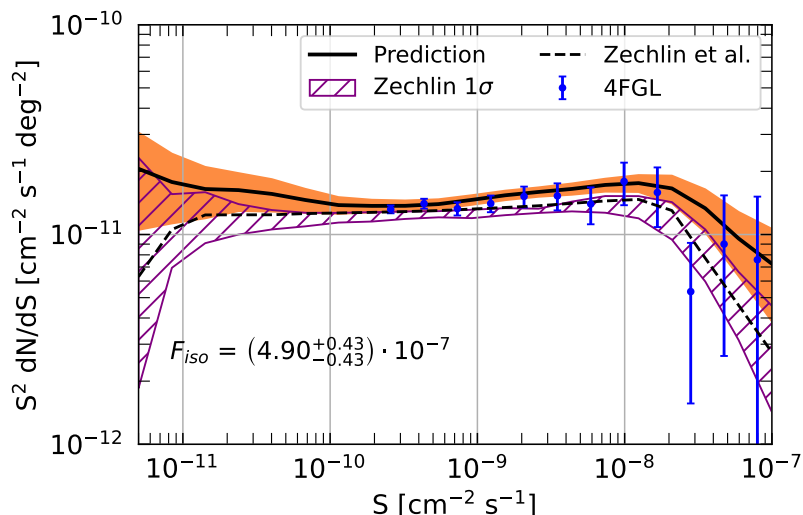
Figure 15. Same as figure 14, including a comparison with the result of ref. [8].
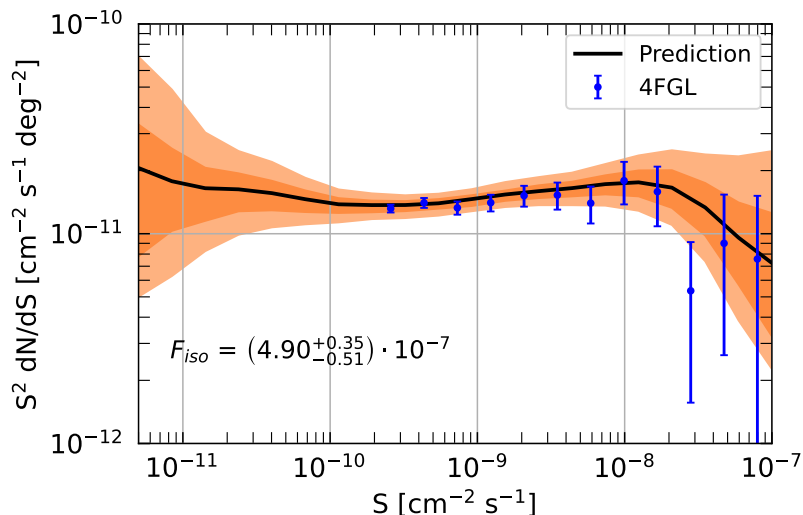

Figure 16. Same as in figure 14, but with errors estimated through a frequentist approach.

with some minimal difference at low fluxes. This provides confidence that the estimated error is robust.

The CNN provides also the value $F_{\text{iso}} = 4.90 \pm 0.43 \cdot 10^{-7} \, \text{cm}^{-2} \, \text{s}^{-1} \, \text{sr}^{-1}$, which can be compared with the value $F'_{\text{iso}} = 4.91 \pm 0.04 \cdot 10^{-7} \, \text{cm}^{-2} \, \text{s}^{-1} \, \text{sr}^{-1}$ reported in table 3. We remind that F'_{iso} is the sum of F_{iso} and of the contribution of sources below the 4FGL catalog threshold. Indeed, if we calculate the contribution of sources using our best fit dN/dS integrated in the range $[5 \cdot 10^{-12}, 1 \cdot 10^{-10}] \, \text{cm}^{-2} \, \text{s}^{-1}$, the upper limit being an approximate value for the catalog threshold, and we add F_{iso} we obtain a value of $(6.41 \pm 0.61) \cdot 10^{-7} \, \text{cm}^{-2} \, \text{s}^{-1} \, \text{sr}^{-1}$. This is in slight tension with F'_{iso}, although the two agree at the $\sim 2\sigma$ level.

In order to test the stability of the results, we repeated the analysis by using the same `gll_iem_v07` galactic foreground but with two different latitude cuts: $|b| < 40°$ and $|b| < 50°$. The positive consequence of a higher latitude cut is a lower residual foreground contamination.
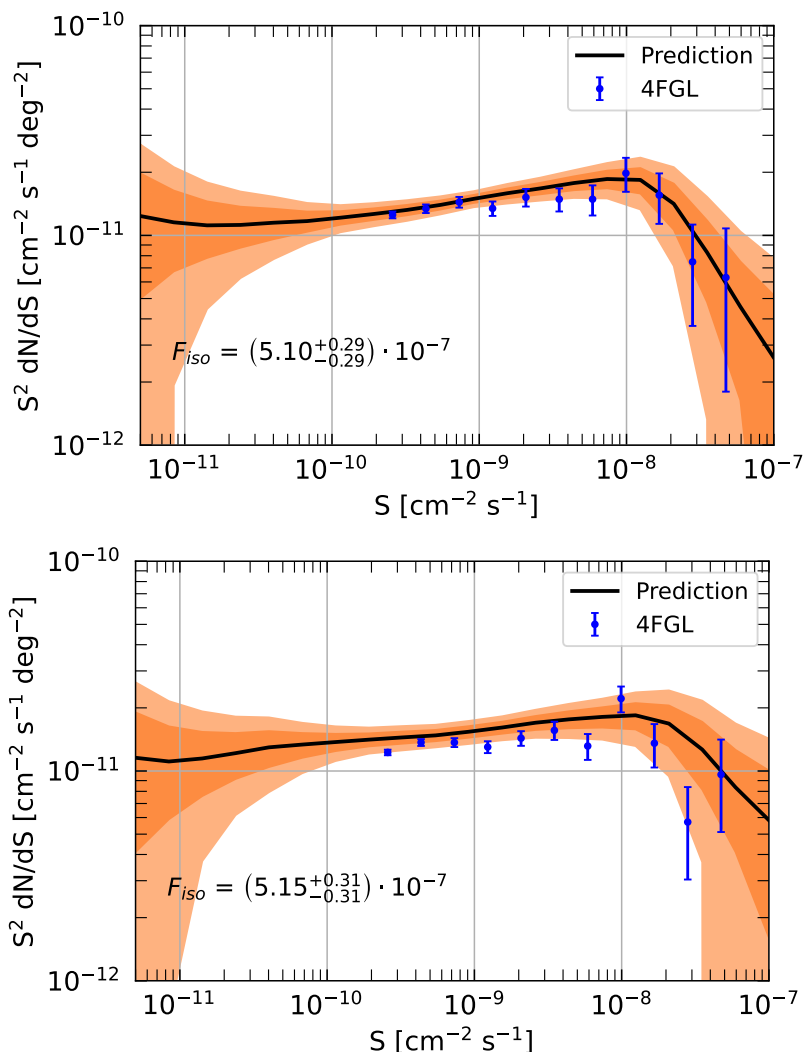
Figure 17. Same as figure 14 but for a latitude cut of 40° and 50°. The blue datapoints of the dN/dS of the 4FGL sources have been updated using only the source in the given region of interested.

However, the amount of available data is reduced. The CNN has been fully re-trained for both of these situations, and the ensuing results when applied to the *Fermi*-LAT data are shown in figure 17. We notice that the results, both in terms of behaviour and uncertainty estimate, are fully compatible with the results of the baseline analysis. The dN/dS exhibits a slight decrease toward lower fluxes as compared to the baseline case, although the size of this effect is not statistically significant. The values of $F_{\rm iso}$ are also fully compatible among them.

A second test of stability has been performed in order to check the impact of galactic foreground modeling. We have performed a new training of the CNN with the alternative template `gll_iem_v05`, and with the $|b| < 30°$ latitude cut. The result is shown in figure 18. The reconstructed dN/dS matches to a large degree the source-count distributions obtained with `gll_iem_v07`. This reassures us that, even though the galactic foreground modeling introduces a degree of uncertainty, nevertheless the results of the CNN for dN/dS are remarkably stable. The value of $F_{\rm iso}$ is smaller in this case, equal to $F_{\rm iso} = 1.82 \pm 0.15 \cdot 10^{-7}\,{\rm cm}^{-2}\,{\rm s}^{-1}\,{\rm sr}^{-1}$,
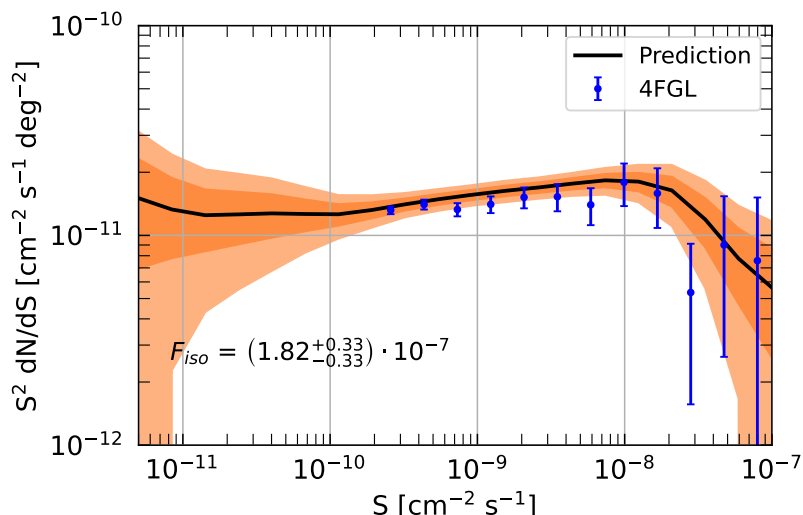
Figure 18. Same as figure 14 but using a CNN trained using the foreground model v05.

which can be compared with the value $F'_{\rm iso} = 2.57 \pm 0.06 \cdot 10^{-7}\,{\rm cm}^{-2}\,{\rm s}^{-1}\,{\rm sr}^{-1}$ reported in table 3. Adding to $F_{\rm iso}$ the integral of the contribution of sources in the interval in the range $[5 \cdot 10^{-12}, 1 \cdot 10^{-10}]\,{\rm cm}^{-2}\,{\rm s}^{-1}$ using the obtained dN/dS, we obtain a total flux of $(2.92 \pm 0.54) \cdot 10^{-7}\,{\rm cm}^{-2}\,{\rm s}^{-1}\,{\rm sr}^{-1}$, which is in very good agreement with $F'_{\rm iso}$.

In appendix A, we also discuss the results of further tests that we performed to cross-check the stability of the dN/dS derived in this section, and that confirm the robustness of the result.

6 Conclusions

In this paper we have studied the differential source-count distribution of gamma-ray sources below the *Fermi*-LAT threshold for source identification. We adopted a convolutional neural network methodology devised to deal with spherical sky-maps and able to extract information related to small scale fluctuations. The technique is based on the EfficientNet V2M model, applied to a patched version of the 2-dimensional spherical sky-map projection of the gamma-ray emission. The neural network has been trained on 900k synthetic maps and validated on 100k additional maps. These synthetic maps have been generated from the flux of random selections of source-counts distributions, to which galactic foreground models and an isotropic component have been added. After training and validation, the neural network has been applied to the *Fermi*-LAT sky-map.

We concentrated our analysis on the photon energy interval $(1, 10)\,{\rm GeV}$ and to 14 years of data. The main result is shown in figure 14: the reconstructed source count distribution exhibits a $dN/dS \sim S^{-2}$ behaviour over almost four orders of magnitude in flux in the range $[5 \cdot 10^{-12}, 1 \cdot 10^{-8}]\,{\rm cm}^{-2}\,{\rm s}^{-1}$, which includes the unresolved regime range $[5 \cdot 10^{-12}, 2 \cdot 10^{10}]$. In the regime where *Fermi*-LAT has sensitivity to individually resolve gamma-ray sources, figure 14 shows that the neural network fairly reproduces the observations, giving confidence on its reliability and strength. While in the unresolved regime the result is in agreement with previous studies performed with different methodologies. We have also further validated the result performing several test of stability which have confirmed that our findings is stable and robust.

The methodology presented here provides a proof of principle for the adoption of a CNN to the reconstruction of the source-count distribution of the extra-galactic sky. With respect to other methods to extract the source-count distribution in the unresolved regime, like the 1-point PDF one, the use of a CNN avoids the need to calculate complicated and numerically demanding likelihoods. Possible future applications includes the extension to multiple energy ranges and energy correlations, and the investigation of features in the dN/dS which might indicate the presence of exotic components like dark matter. A further aspect which would be interesting to investigate is to compare different approaches to machine learning on spherical domains. In particular, it might be interesting to employ a graph convolutional neural network as suggested in [41], in order to preserve spatial relations between distant pixels, and rotational invariance. Finally, recent developments in the field of simulation based inference, such as our case, suggest that a promising approach to estimate the complex posterior distribution of the objective parameters are the so-called neural likelihood ratio estimation techniques [60], and in particular Truncated Marginal Neural Ratio Estimation [61–63]. Since these techniques are tailored for simulation based problems, we hypothesise that it would be possible to achieve similar performance to our current neural network architecture with an even simpler structure and possibly less training samples.

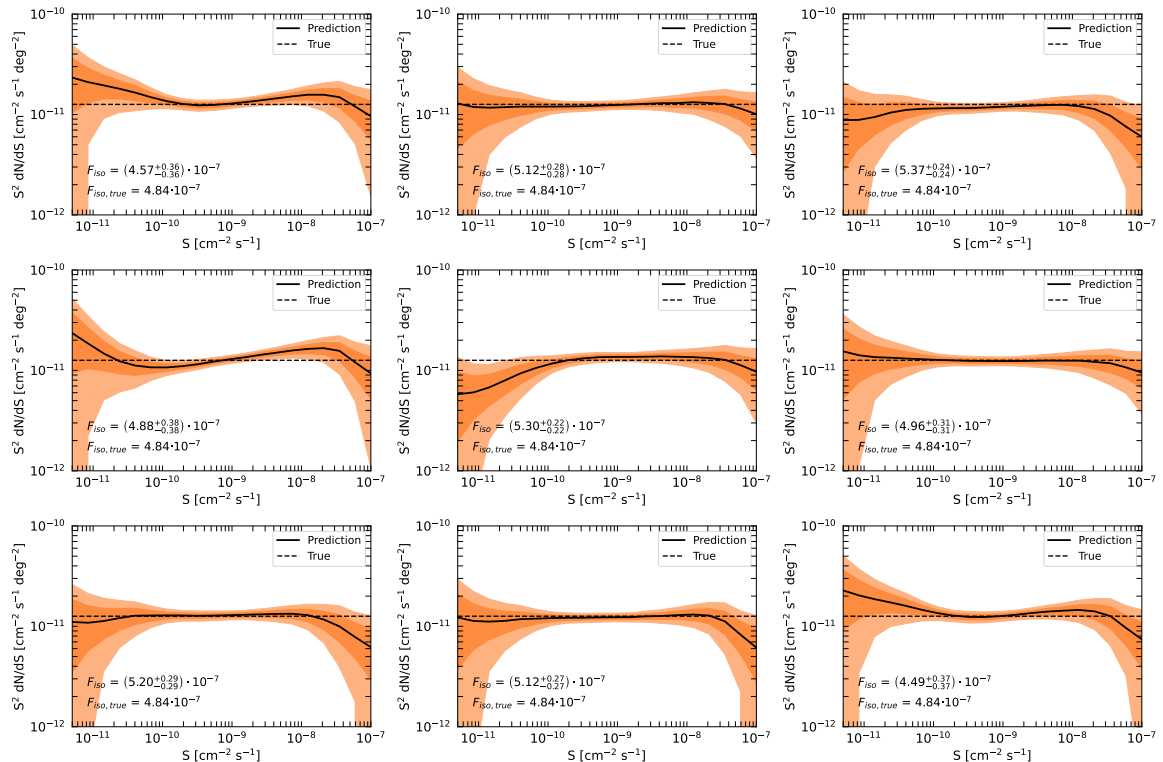## Acknowledgments

---

[4]https://github.com/nmik/Xgam.

Figure 19. Same as figure 11 but for 6 different random realization of the same flat $S^2 dN/dS$ and same $F_{\rm iso,true} = 4.84 \cdot 10^{-7} \, {\rm cm}^{-2} \, {\rm s}^{-1} \, {\rm sr}^{-1}$.

A Further tests

In this appendix we describe further tests that we have performed to cross-check the stability and robustness of the dN/dS derived in the main text.

A.1 Flat $S^2 \, dN/dS$

Figure 19 shows the stability of the CNN output for a specific repeated dN/dS input. In this case, the CNN, trained on the wide variability of models of table 2, is applied to a set of maps, all generated with a specific $dN/dS \sim S^{-2}$. This specific choice of input dN/dS is meant to test a case which is expected to be similar to the real *Fermi*-LAT case, as was found in [8] for mid-values of S. The plots show that the CNN is remarkably consistent in reproducing the correct behaviour. The drop at high fluxes is mostly due to the fact that the CNN has been instructed with maps that contain a dropping dN/dS at high fluxes (see table 2 — the reason for this choice of prior is that sources in the 4FGL catalog exhibit this drop in the source count). For mid-values of S, the reconstructed $S^2 dN/dS$ is quite flat, with a small error band. At low S, well below the *Fermi*-LAT threshold and when approaching the confusion limit, some mild deviation occurs, but the true dN/dS lies always within the uncertainties. These tests gives us confidence that the CNN should be able to faithfully reconstruct the source distribution of the *Fermi*-LAT sky, when applied to the real data in section 5.

As a further test, we show in figure 20 the case where we apply our method to maps which contain a galactic foreground model different from the one used in the training. The
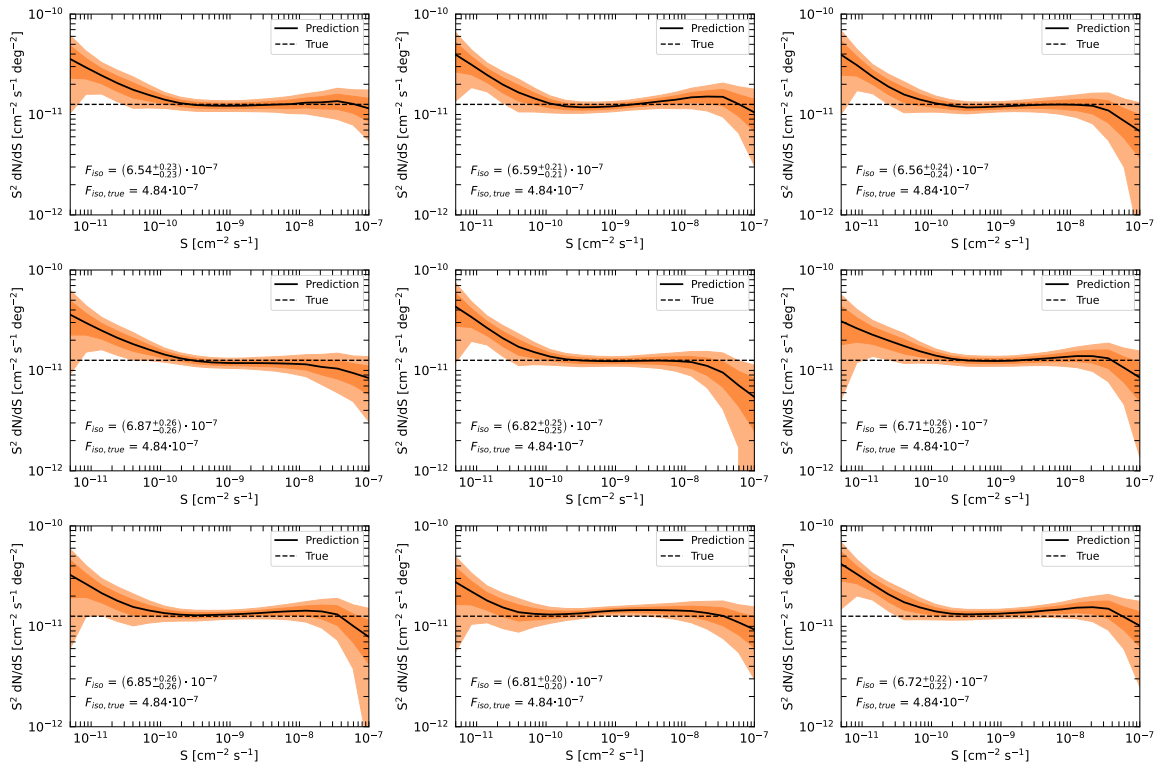
Figure 20. Same as figure 19 but for synthetic maps generated using version v05 of the foreground model.

CNN has been built (like in our baseline case) with maps generated with the galactic template model `gll_iem_v07`. The same model is used in the foreground subtraction performed in the analysis. On the contrary, for this test we have generated maps using the alternative `gll_iem_v05` foreground model, and let the above `gll_iem_v07`-trained CNN to analyze them. This has been done in order to verify the resilience of the method on the imperfect knowledge of the galactic foreground. Figure 20 shows that the reconstructed dN/dS has a level of agreement with the input model comparable to the analogous case of figure 19. Some larger deviation is present at very low fluxes, even though the reconstructed and input models are compatible within 1σ.

A.2 UltraCleanVeto selection

In this section we further test the robustness of the results to the *Fermi*-LAT data selection. In particular, while for the main results we used the `SOURCEVETO` event class selection, here we test the `ULTRACLEANVETO` selection. This event class has more stringent cuts with respct to the `SOURCEVETO` class, and it thus contains a lower residual charged cosmic-ray background, at the prize of a lower effective area (by about 15%). Except for the different data selection all the rest of the analysis is performed in the same identical way as for the `SOURCEVETO` case. The resulting dN/dS is shown in figure 21 and it can be seen that it is compatible with the main results, while, as expected, the value of $F_{\rm iso}$ is a bit lower.
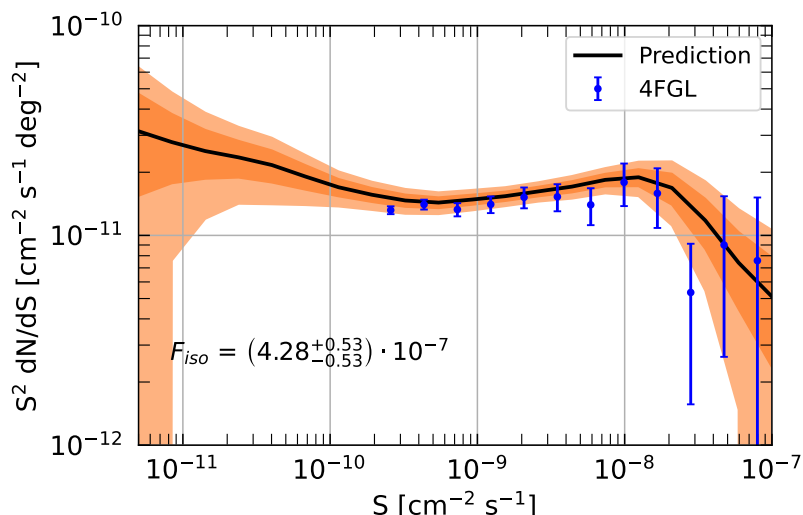
Figure 21. Same as figure 14 but for the `ULTRACLEANVETO` *Fermi*-LAT data selection.

A.3 Multipole analysis

Since foreground modeling can be an important source of systematic uncertainty, we investigate this potential issue performing a further test. As discussed above, the signal we try to extract from the gamma ray maps, is a small-scale effect, since it is due to a distribution of point sources. To a large degree, this point sources are isotropically distributed in the sky and contribute to the fluctuations of this isotropic field at small angular scales. On the contrary, the galactic foreground is quite diffuse, as can be seen in figure 6, and therefore contributes mainly to large-scale anisotropies. In order to further remove from the data the possible large-scale residual component of the galactic foreground, which might still be present after imperfect foreground removal performed with the use of the templates, we adopt a method based on the transformation of the flux to harmonic space and the removal from the maps of the low-multipoles (i.e., large angular scales) contribution. We stress that this procedure is applied only to the *Fermi*-LAT data map, and not to the synthetic maps. For the latter, the foreground subtracted is the same used to generate the maps, so foreground subtraction is, by definition, "perfect". The procedure follows these steps:

- We start with the foreground-subtracted flux map, to which we apply the $|b| < 30°$ latitude cut and a 2° mask around each of the bright sources of the 4FGL catalog (which would otherwise dominate the angular power spectrum). We first determine and remove the monopole ($l = 0$) and dipole ($l = 1$) components from the masked map (for this, we use the `remove_dipole` routine from the Healpy library). In this way, after the subtraction procedure outlined below, the monopole (i.e. the total flux) and dipole information of the *Fermi*-LAT map are retained.

- This map is then subject to harmonic space decomposition:

$$\mathcal{M}_{\text{masked}}(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{lm} Y_{lm}(\theta, \phi) \tag{A.1}$$

where θ and ϕ are angles on the sphere, $Y_{lm}(\theta, \phi)$ the spherical harmonics and the harmonic coefficients a_{lm} completely encode the same information present in the map

$\mathcal{M}_{\text{masked}}$ (notice that, even though in eq. (A.1) we formally decompose over all multipoles, the monopole and dipole amplitude are vanishing, since they have been removed from $\mathcal{M}_{\text{masked}}$);

- The harmonic coefficients a_{lm} are determined with the `map2alm` Healpy function. The ones that are retained for the foreground cleaning are those referring to multipole values $l < l_{\text{max}}$, where l_{max} will be varied and the dependence of the results with l_{max} will be studied (at this level, the monopole and dipole are not present, because of the above point);

- By using the a_{lm} corresponding to multipoles $l < l_{\text{max}}$, we use eq. (A.1) to construct a map $\mathcal{M}_{\text{residual}}$ which contains only the large-scale component. This is performed with `alm2map` Healpy function.

- We subtract the residual map from the original map, thus obtaining a map where the large scale residual components have been removed. The size of the scale is determined by $l < l_{\text{max}}$.

With this algorithm we obtain maps where the large scale residual features associated to imperfect galactic foreground subtraction are further removed. The results we obtain on the reconstructed dN/dS for the *Fermi*-LAT data (after this "multipole cleaning") are shown in figure 22, for different values of l_{max} and for the `gll_iem_v07` template. By comparing figure 22 with figure 14, which refers to the same foreground template and latitude cut, we can see that the results are extremely stable up to $l_{\text{max}} \sim 100$. This means that our baseline analysis is not affected by incomplete foreground subtraction, and can be considered reliable. The fall-off of the dN/dS when $l_{\text{max}} > 100$ is expected since $l_{\text{max}} = 100$ corresponds to an angular scale of the order of $2°$, which starts to be compatible with the size of (PSF-smeared) point sources. Therefore, a cleaning with $l_{\text{max}} > 100$ starts to remove point sources instead of galactic foregrounds.

The same analysis has been performed on a pipeline fully based on `gll_iem_v05`: the result is shown in figure 23, to be compared with figure 18. Also in this case, the results are extremely stable up to $l_{\text{max}} \sim 100$, and the same above considerations apply.

Finally, to check the strength of the method against the imperfect knowledge of the background, we performed the same analysis of multipole cleaning by adopting our baseline pipeline (based on `gll_iem_v07`) on maps generated with `gll_iem_v05` (similarly to what has been done for figure 20). The results are shown in figure 24, where again we see that the results are pretty consistent, up to $l \sim 100$. It can be seen, moreover, that the procedure of multipole cleaning helps in removing the, nonetheless small, positive bias which is present toward low fluxes in this case.

In conclusion, these tests give confidence on the reliability of the method and that the main result of figure 14 corresponds to a fair representation of the source-count distribution of the unresolved gamma-ray sources.
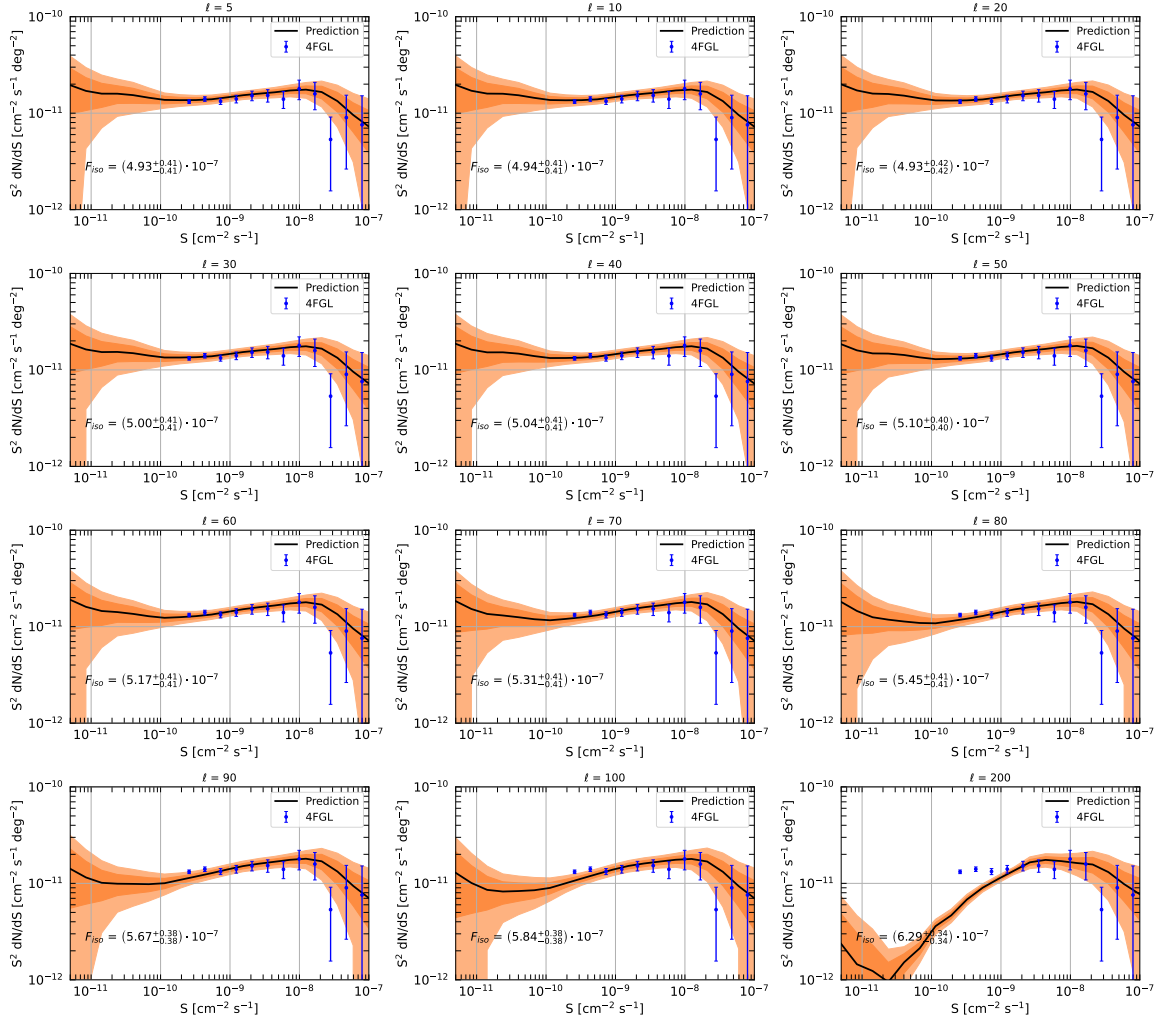
Figure 22. Same as figure 14 but where the *Fermi*-LAT map given as input to the CNN has been processed with the improved foreground cleaning procedure based on multipole decomposition described in the text. Each panel refer to a different maximal multipole $\ell_{\max}$ used for the cleaning.

A.4 Changing A_{gal}

In order to test the stability of the results against changes in the values of A_{gal}, we have provided, as input to our trained CNN, *Fermi*-LAT maps cleaned with different values of A_{gal}, ranging from 0.78 to 1.0 (our fiducial value is 0.888 ± 0.005, thus the chosen range encompass many standard deviations around its central value). The results are shown in figure 25 below. It can be seen that the dN/dS is considerably robust and stable to different values of A_{gal}, despite the large range explored. Instead, different values of A_{gal} seem to imply a significant variation in the reconstructed value of F_{iso}, as expected since at the large Galactic latitudes explored in our analysis ($|b| > 30°$) there is a certain amount of degeneracy between a purely isotropic emission and the Galactic foreground emission itself. However, this does not affect the reconstructed dN/dS.
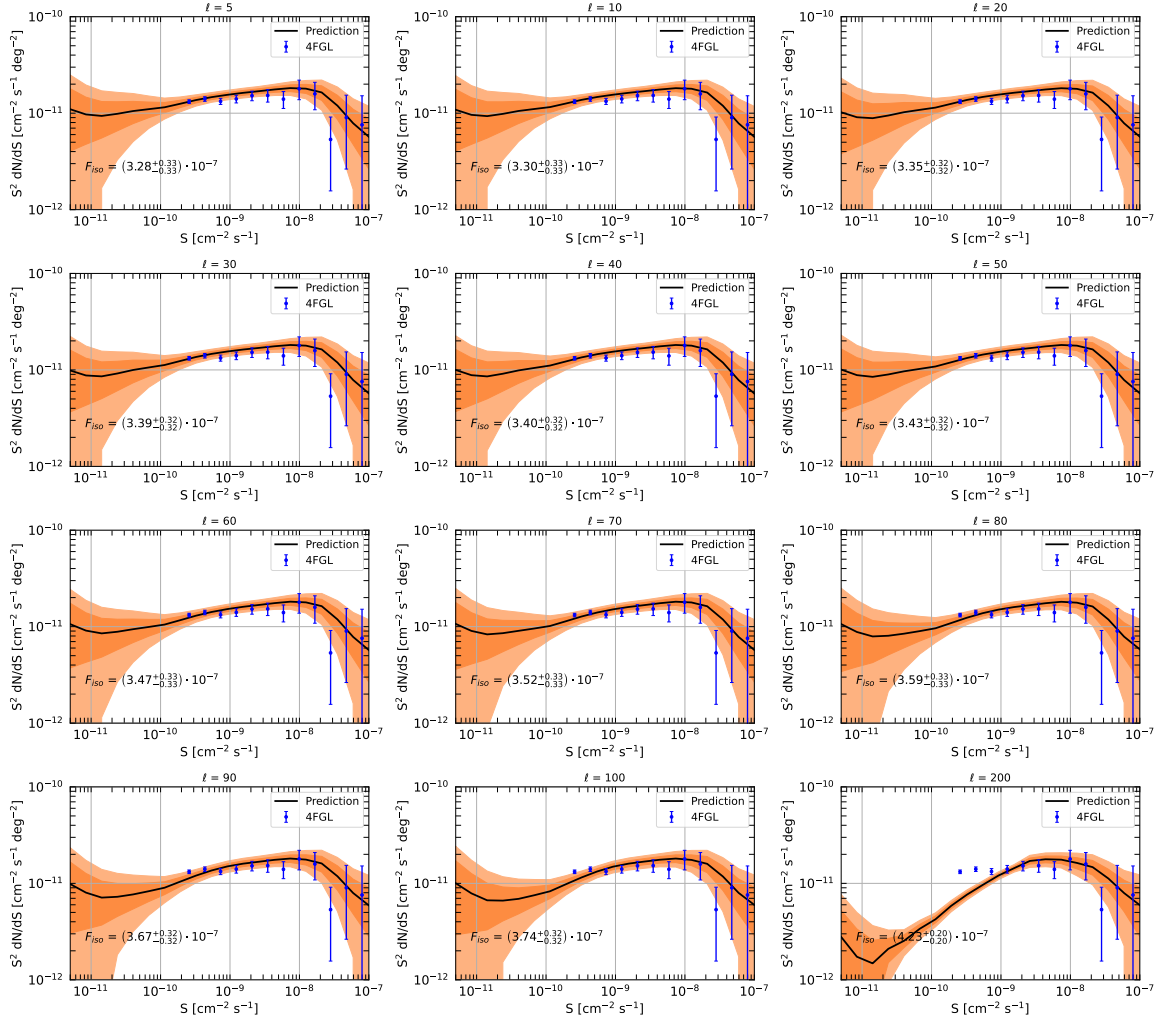
Figure 23. Same as figure 22 but using a CNN trained with the v05 version of the foreground model, as well as using as input a *Fermi*-LAT count map processed with the same foreground model.

A.5 Fully spherical CNN

In this section we show a comparison of the results obtained using a fully spherical CNN, instead of the `map2patch` algorithm used in the main analysis. In particular, we adopt here the NNHealpix [42] algorithm.

Figure 26 shows the reconstructed dN/dS, $F_{\rm iso}$ and their frequentist errors, using the spherical CNN applied to the *Fermi*-LAT map. The results are fully compatible with the ones obtained with the `map2patch` algorithm, both for the mean value and the errors, as can be seen by confronting with figure 16.

Figure 27 instead, shows some examples of random input dN/dS and $F_{\rm iso}$ from the validation dataset of the Spherical CNN, and the corresponding quantities and their errors reconstructed by the CNN. The spherical CNN has similar performances as the CNN that uses `map2patch`, as can be seen comparing with figure 11. The same occurs also in figure 28, which refers to a flat $S^2 dN/dS$, to be compared with figure 19 obtained with `map2patch`.
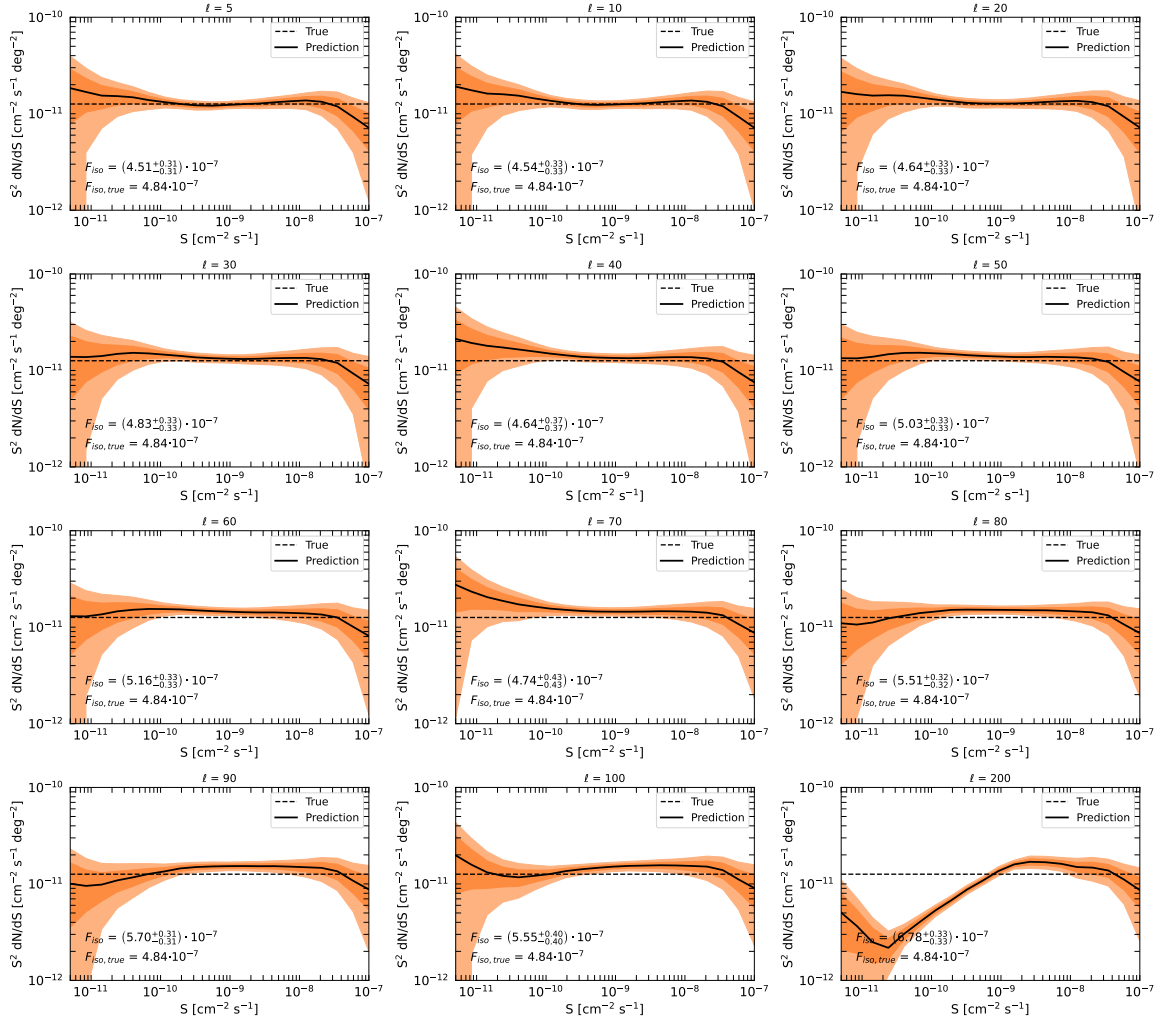
Figure 24. Same as figure 22 but using as input a synthetic map generated with a flat $S^2 dN/dS$ and with the v05 version of the foreground model.
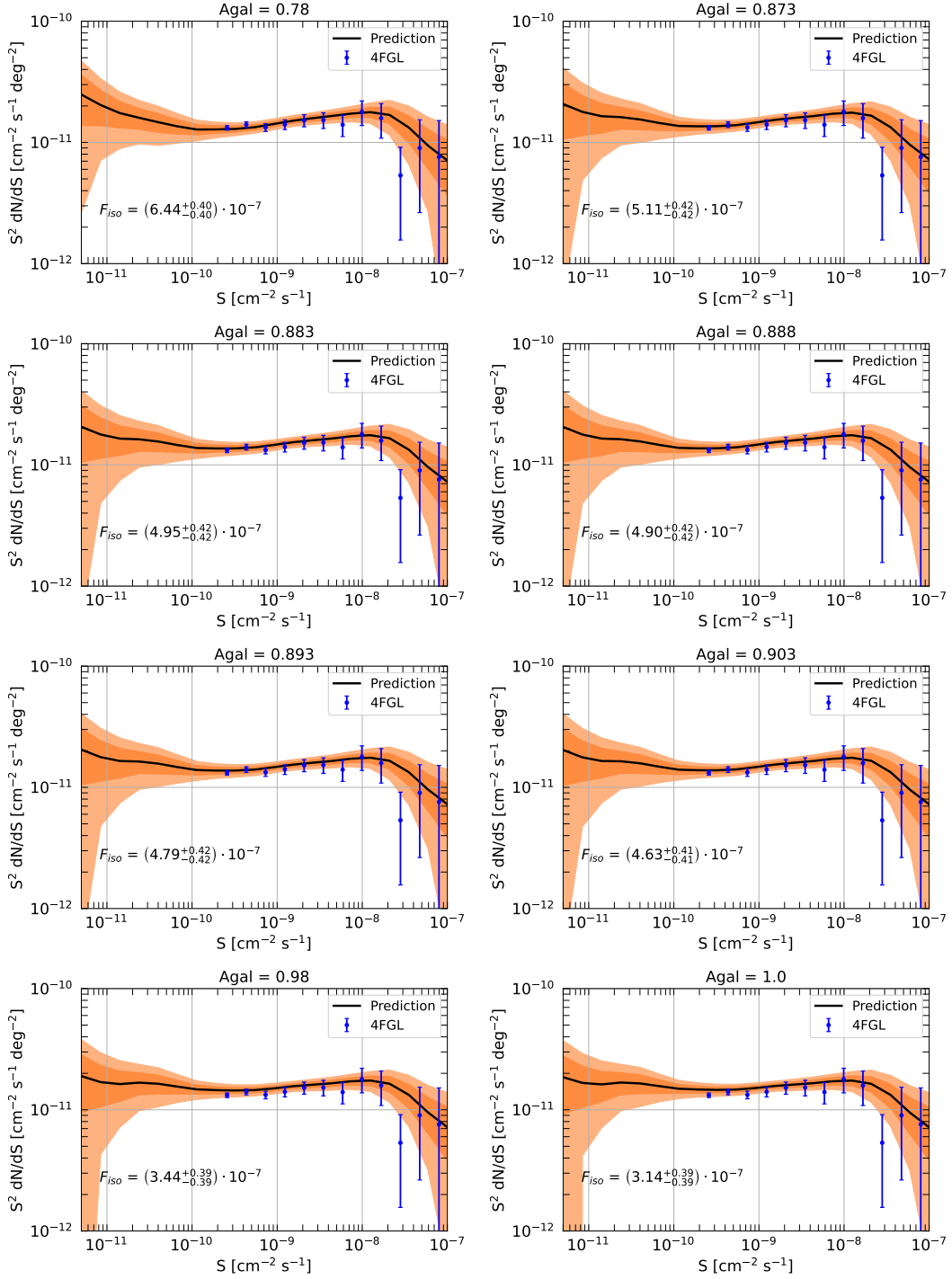
Figure 25. Reconstructed dN/dS and F_{iso} and their Bayesian errors, when the trained CNN is applied to the *Fermi*-LAT map using different values of A_{gal} as indicated in the panel labels.
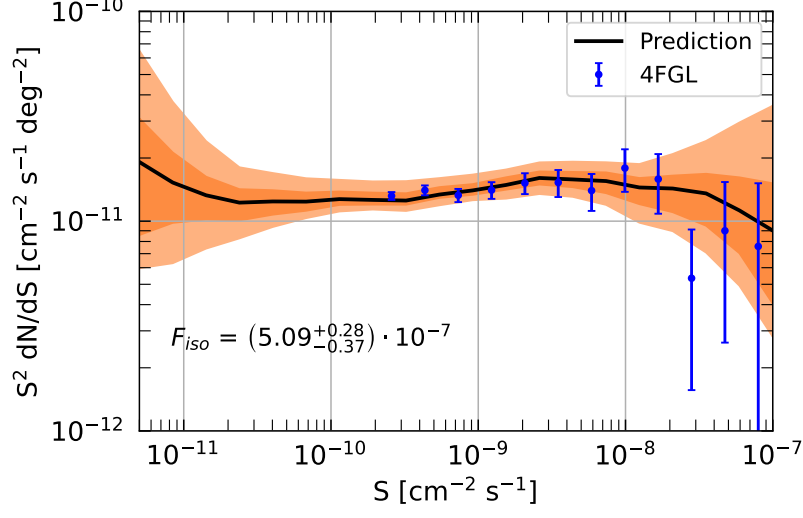
Figure 26. Reconstructed dN/dS and $F_{\rm iso}$ and their errors, using a fully spherical CNN, trained on simulated maps and then applied to the *Fermi*-LAT map.
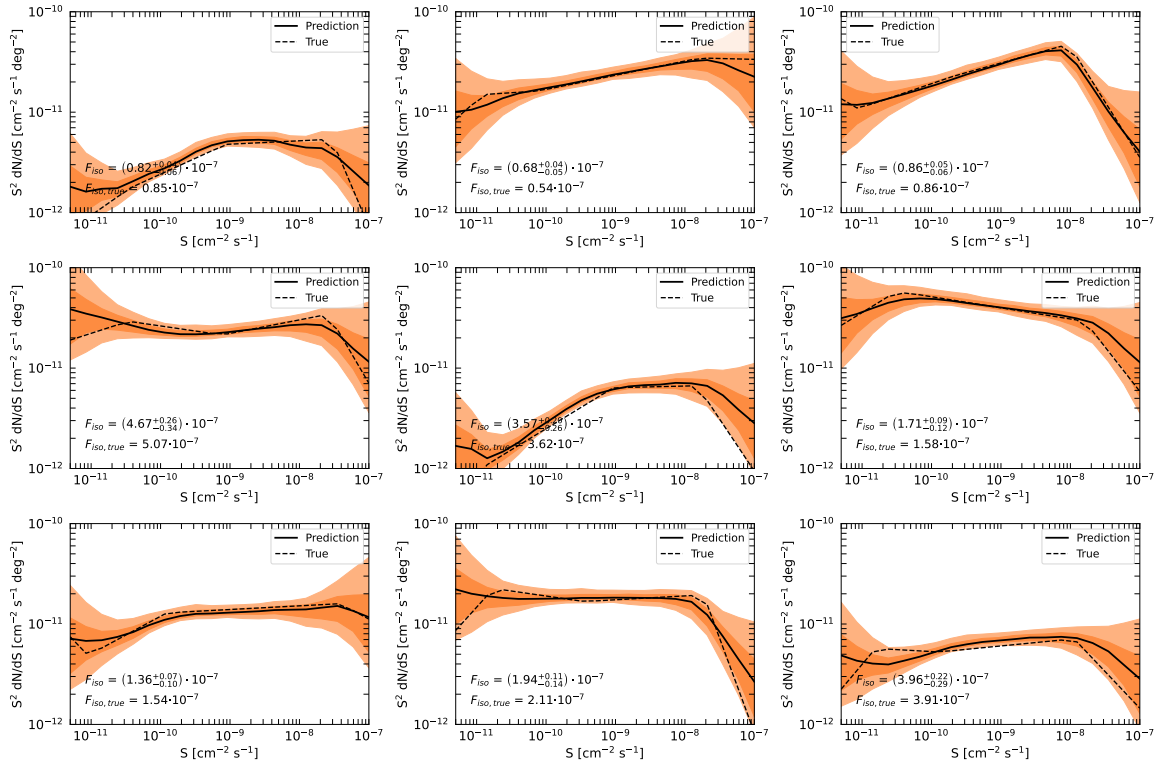

Figure 27. Some example dN/dS from the validation dataset of the spherical CNN. Each panel shows the input dN/dS and $F_{\rm iso}$ and the corresponding quantities reconstructed by the CNN. The colored areas indicate the 1σ and 2σ uncertainty bands.
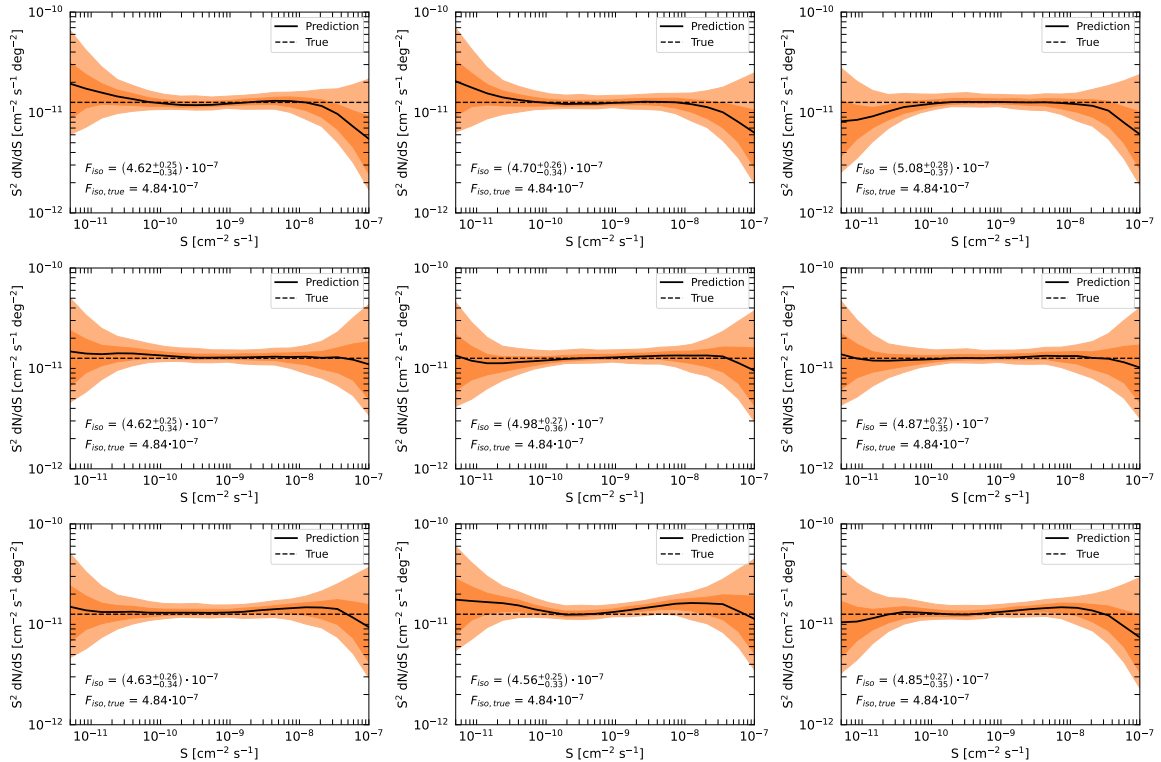
Figure 28. Same as figure 27 but for 6 different random realization of a flat $S^2 dN/dS$ and same F_{iso}.

References

[1] M. Fornasa and M.A. Sánchez-Conde, *The nature of the Diffuse Gamma-Ray Background*, *Phys. Rept.* **598** (2015) 1 [arXiv:1502.02866] [INSPIRE].

[2] FERMI-LAT collaboration, *The spectrum of isotropic diffuse gamma-ray emission between 100 MeV and 820 GeV*, *Astrophys. J.* **799** (2015) 86 [arXiv:1410.3696] [INSPIRE].

[3] FERMI-LAT collaboration, *The Fermi-LAT high-latitude Survey: Source Count Distributions and the Origin of the Extragalactic Diffuse Background*, *Astrophys. J.* **720** (2010) 435 [arXiv:1003.0895] [INSPIRE].

[4] FERMI-LAT collaboration, *Resolving the Extragalactic γ-Ray Background above 50 GeV with the Fermi Large Area Telescope*, *Phys. Rev. Lett.* **116** (2016) 151105 [arXiv:1511.00693] [INSPIRE].

[5] M. Ajello et al., *The Origin of the Extragalactic Gamma-Ray Background and Implications for Dark-Matter Annihilation*, *Astrophys. J. Lett.* **800** (2015) L27 [arXiv:1501.05301] [INSPIRE].

[6] L. Marcotulli, M. Di Mauro and M. Ajello, *Source-count Distribution of Gamma-Ray Blazars*, *Astrophys. J.* **896** (2020) 6 [arXiv:2006.04703] [INSPIRE].

[7] D. Malyshev and D.W. Hogg, *Statistics of gamma-ray point sources below the Fermi detection limit*, *Astrophys. J.* **738** (2011) 181 [arXiv:1104.0010] [INSPIRE].

[8] H.-S. Zechlin, A. Cuoco, F. Donato, N. Fornengo and A. Vittino, *Unveiling the Gamma-ray Source Count Distribution Below the Fermi Detection Limit with Photon Statistics*, *Astrophys. J. Suppl.* **225** (2016) 18 [arXiv:1512.07190] [INSPIRE].

[9] H.-S. Zechlin, A. Cuoco, F. Donato, N. Fornengo and M. Regis, *Statistical Measurement of the Gamma-ray Source-count Distribution as a Function of Energy*, *Astrophys. J. Lett.* **826** (2016) L31 [arXiv:1605.04256] [INSPIRE].

[10] M. Lisanti, S. Mishra-Sharma, L. Necib and B.R. Safdi, *Deciphering Contributions to the Extragalactic Gamma-Ray Background from 2 GeV to 2 TeV*, *Astrophys. J.* **832** (2016) 117 [arXiv:1606.04101] [INSPIRE].

[11] FERMI-LAT collaboration, *The Large Area Telescope on the Fermi Gamma-ray Space Telescope Mission*, *Astrophys. J.* **697** (2009) 1071 [arXiv:0902.1089] [INSPIRE].

[12] M. Korsmeier, E. Pinetti, M. Negro, M. Regis and N. Fornengo, *Flat-spectrum Radio Quasars and BL Lacs Dominate the Anisotropy of the Unresolved Gamma-Ray Background*, *Astrophys. J.* **933** (2022) 221 [arXiv:2201.02634] [INSPIRE].

[13] S. Manconi, M. Korsmeier, F. Donato, N. Fornengo, M. Regis and H. Zechlin, *Testing gamma-ray models of blazars in the extragalactic sky*, *Phys. Rev. D* **101** (2020) 103026 [arXiv:1912.01622] [INSPIRE].

[14] M. Di Mauro, S. Manconi, H.-S. Zechlin, M. Ajello, E. Charles and F. Donato, *Deriving the contribution of blazars to the Fermi-LAT Extragalactic γ-ray background at E > 10 GeV with efficiency corrections and photon statistics*, *Astrophys. J.* **856** (2018) 106 [arXiv:1711.03111] [INSPIRE].

[15] H.-S. Zechlin, S. Manconi and F. Donato, *Constraining Galactic dark matter with gamma-ray pixel counts statistics*, *Phys. Rev. D* **98** (2018) 083022 [arXiv:1710.01506] [INSPIRE].

[16] S. Mishra-Sharma, N.L. Rodd and B.R. Safdi, *NPTFit: A code package for Non-Poissonian Template Fitting*, *Astron. J.* **153** (2017) 253 [arXiv:1612.03173] [INSPIRE].

[17] L.G.C. Bariuan and T.R. Slatyer, *Characterizing the expected behavior of non-Poissonian template fitting*, *Phys. Rev. D* **107** (2023) 103014 [arXiv:2207.13097] [INSPIRE].

[18] S.K. Lee, M. Lisanti and B.R. Safdi, *Distinguishing Dark Matter from Unresolved Point Sources in the Inner Galaxy with Photon Statistics*, *JCAP* **05** (2015) 056 [arXiv:1412.6099] [INSPIRE].

[19] S.K. Lee, M. Lisanti, B.R. Safdi, T.R. Slatyer and W. Xue, *Evidence for Unresolved γ-Ray Point Sources in the Inner Galaxy*, *Phys. Rev. Lett.* **116** (2016) 051103 [arXiv:1506.05124] [INSPIRE].

[20] T. Linden, N.L. Rodd, B.R. Safdi and T.R. Slatyer, *High-energy tail of the Galactic Center gamma-ray excess*, *Phys. Rev. D* **94** (2016) 103013 [arXiv:1604.01026] [INSPIRE].

[21] R.K. Leane and T.R. Slatyer, *Revival of the Dark Matter Hypothesis for the Galactic Center Gamma-Ray Excess*, *Phys. Rev. Lett.* **123** (2019) 241101 [arXiv:1904.08430] [INSPIRE].

[22] L.J. Chang, S. Mishra-Sharma, M. Lisanti, M. Buschmann, N.L. Rodd and B.R. Safdi, *Characterizing the nature of the unresolved point sources in the Galactic Center: An assessment of systematic uncertainties*, *Phys. Rev. D* **101** (2020) 023014 [arXiv:1908.10874] [INSPIRE].

[23] M. Buschmann et al., *Foreground Mismodeling and the Point Source Explanation of the Fermi Galactic Center Excess*, *Phys. Rev. D* **102** (2020) 023023 [arXiv:2002.12373] [INSPIRE].

[24] R.K. Leane and T.R. Slatyer, *Spurious Point Source Signals in the Galactic Center Excess*, *Phys. Rev. Lett.* **125** (2020) 121105 [arXiv:2002.12370] [INSPIRE].

[25] R.K. Leane and T.R. Slatyer, *The enigmatic Galactic Center excess: Spurious point sources and signal mismodeling*, *Phys. Rev. D* **102** (2020) 063019 [arXiv:2002.12371] [INSPIRE].

[26] G.H. Collin, N.L. Rodd, T. Erjavec and K. Perez, *A Compound Poisson Generator Approach to Point-source Inference in Astrophysics*, *Astrophys. J. Suppl.* **260** (2022) 29 [arXiv:2104.04529] [INSPIRE].

[27] S. Mishra-Sharma and K. Cranmer, *Neural simulation-based inference approach for characterizing the Galactic Center γ-ray excess*, *Phys. Rev. D* **105** (2022) 063017 [arXiv:2110.06931] [INSPIRE].

[28] F. List, N.L. Rodd and G.F. Lewis, *Extracting the Galactic Center excess' source-count distribution with neural nets*, *Phys. Rev. D* **104** (2021) 123022 [arXiv:2107.09070] [INSPIRE].

[29] F. List, N.L. Rodd, G.F. Lewis and I. Bhat, *The GCE in a New Light: Disentangling the γ-ray Sky with Bayesian Graph Convolutional Neural Networks*, *Phys. Rev. Lett.* **125** (2020) 241102 [arXiv:2006.12504] [INSPIRE].

[30] FERMI-LAT collaboration, https://fermi.gsfc.nasa.gov/ssc/data/analysis/.

[31] FERMI-LAT collaboration, *Pass 8: Toward the Full Realization of the Fermi-LAT Scientific Potential*, arXiv:1303.3514] [INSPIRE].

[32] FERMI-LAT collaboration, *Fermi-LAT improved Pass~8 event selection*, in *8th International Fermi Symposium: Celebrating 10 Year of Fermi*, Baltimore, U.S.A. (2018) [arXiv:1810.11394] [INSPIRE].

[33] 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] [INSPIRE].

[34] M. Fornasa et al., *Angular power spectrum of the diffuse gamma-ray emission as measured by the Fermi Large Area Telescope and constraints on its dark matter interpretation*, *Phys. Rev. D* **94** (2016) 123005 [arXiv:1608.07289] [INSPIRE].

[35] FERMI-LAT collaboration, *Incremental Fermi Large Area Telescope Fourth Source Catalog*, *Astrophys. J. Suppl.* **260** (2022) 53 [arXiv:2201.11184] [INSPIRE].

[36] FERMI-LAT collaboration, *The Second Fermi Large Area Telescope Catalog of Gamma-ray Pulsars*, *Astrophys. J. Suppl.* **208** (2013) 17 [arXiv:1305.4385] [INSPIRE].

[37] FERMI-LAT collaboration, *Fermi Large Area Telescope Fourth Source Catalog*, *Astrophys. J. Suppl.* **247** (2020) 33 [arXiv:1902.10045] [INSPIRE].

[38] M. Abadi et al., *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*, https://www.tensorflow.org (2015).

[39] Y.-C. Su and K. Grauman, *Learning Spherical Convolution for Fast Features from 360° Imagery*, arXiv:1708.00919.

[40] T.S. Cohen, M. Weiler, B. Kicanaoglu and M. Welling, *Gauge Equivariant Convolutional Networks and the Icosahedral CNN*, arXiv:1902.04615.

[41] M. Defferrard, M. Milani, F. Gusset and N. Perraudin, *DeepSphere: a graph-based spherical CNN*, arXiv:2012.15000.

[42] N. Krachmalnicoff and M. Tomasi, *Convolutional Neural Networks on the HEALPix sphere: a pixel-based algorithm and its application to CMB data analysis*, *Astron. Astrophys.* **628** (2019) A129 [arXiv:1902.04083] [INSPIRE].

[43] M. Eder, T. Price, T. Vu, A. Bapat and J.-M. Frahm, *Mapped Convolutions*, arXiv:1906.11096.

[44] Z. Li, F. Liu, W. Yang, S. Peng and J. Zhou, *A survey of convolutional neural networks: Analysis, applications, and prospects*, *IEEE Trans. Neural Netw. Learn. Syst.* **33** (2022) 6999.

[45] J. Gu et al., *Recent advances in convolutional neural networks*, *Pattern Recognit.* **77** (2018) 354.

[46] A. Dhillon and G.K. Verma, *Convolutional neural network: a review of models, methodologies and applications to object detection*, *Prog. Artif. Intell.* **9** (2019) 85.

[47] K. Hornik, M. Stinchcombe and H. White, *Multilayer feedforward networks are universal approximators*, *Neural Netw.* **2** (1989) 359.

[48] X. Ying, *An overview of overfitting and its solutions*, *J. Phys. Conf. Ser.* **1168** (2019) 022022.

[49] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, J. Mach. Learn. Res. **15** (2014) 1929 [ɪɴSPIRE].

[50] S. Ioffe and C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, arXiv:1502.03167 [ɪɴSPIRE].

[51] T. van Laarhoven, *L2 Regularization versus Batch and Weight Normalization*, arXiv:1706.05350.

[52] C. Garbin, X. Zhu and O. Marques, *Dropout vs. batch normalization: an empirical study of their impact to deep learning*, *Multimed. Tools Appl.* **79** (2020) 12777.

[53] M. Tan and Q.V. Le, *EfficientNetV2: Smaller Models and Faster Training*, arXiv:2104.00298.

[54] Y. Gal, J. Hron and A. Kendall, *Concrete Dropout*, arXiv:1705.07832.

[55] C.J. Maddison, A. Mnih and Y. Whye Teh, *The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables*, arXiv:1611.00712.

[56] E. Hüllermeier and W. Waegeman, *Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods*, arXiv:1910.09457.

[57] Y. Gal and Z. Ghahramani, *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, arXiv:1506.02142.

[58] A. Kendall and Y. Gal, *What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?*, arXiv:1703.04977.

[59] L.N. Smith, *Cyclical Learning Rates for Training Neural Networks*, arXiv:1506.01186.

[60] C. Durkan, I. Murray and G. Papamakarios, *On Contrastive Learning for Likelihood-free Inference*, arXiv:2002.03712.

[61] B.K. Miller, A. Cole, P. Forré, G. Louppe and C. Weniger, *Truncated Marginal Neural Ratio Estimation*, in *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, online conference, Canada (2021) [arXiv:2107.01214] [ɪɴSPIRE].

[62] B.K. Miller, A. Cole, C. Weniger, F. Nattino, O. Ku and M.W. Grootes, *swyft: Truncated Marginal Neural Ratio Estimation in Python*, *J. Open Source Softw.* **7** (2022) 4205 [ɪɴSPIRE].

[63] N. Anau Montel and C. Weniger, *Detection is truncation: studying source populations with truncated marginal neural ratio estimation*, in *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, U.S.A. (2022) [arXiv:2211.04291] [ɪɴSPIRE].