A neural simulation-based inference approach for characterizing the Galactic Center γ-ray excess

Siddharth Mishra-Sharma[1, 2, 3, 4, 5, *] and Kyle Cranmer[5, 6, †]

[1] *Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2] *The NSF AI Institute for Artificial Intelligence and Fundamental Interactions*
[3] *Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[4] *Department of Physics, Harvard University, Cambridge, MA 02138, USA*
[5] *Center for Cosmology and Particle Physics, Department of Physics, New York University, New York, NY 10003, USA*
[6] *Center for Data Science, New York University, 60 Fifth Ave, New York, NY 10011, USA*

(Dated: October 13, 2021)

The nature of the *Fermi* γ-ray Galactic Center Excess (GCE) has remained a persistent mystery for over a decade. Although the excess is broadly compatible with emission expected due to dark matter annihilation, an explanation in terms of a population of unresolved astrophysical point sources *e.g.*, millisecond pulsars, remains viable. The effort to uncover the origin of the GCE is hampered in particular by an incomplete understanding of diffuse emission of Galactic origin. This can lead to spurious features that make it difficult to robustly differentiate smooth emission, as expected for a dark matter origin, from more "clumpy" emission expected for a population of relatively bright, unresolved point sources. We use recent advancements in the field of simulation-based inference, in particular density estimation techniques using normalizing flows, in order to characterize the contribution of modeled components, including unresolved point source populations, to the GCE. Compared to traditional techniques based on the statistical distribution of photon counts, our machine learning-based method is able to utilize more of the information contained in a given model of the Galactic Center emission, and in particular can perform posterior parameter estimation while accounting for pixel-to-pixel spatial correlations in the γ-ray map. This makes the method demonstrably more resilient to certain forms of model misspecification. On application to *Fermi* data, the method generically attributes a smaller fraction of the GCE flux to unresolved point sources when compared to traditional approaches. We nevertheless infer such a contribution to make up a non-negligible fraction of the GCE across all analysis variations considered, with at least 38^{+9}_{-19}% of the excess attributed to unresolved points sources in our baseline analysis.

CONTENTS

I. INTRODUCTION

Dark matter (DM) represents one of the major unsolved problems in particle physics and cosmology today. The traditional Weakly-Interacting Massive Particle (WIMP) paradigm envisions production of dark matter in the early Universe through freeze-out of dark sector particles weakly coupled to the Standard Model (SM) sector. In this scenario, one of the most promising avenues of detecting a dark matter signal is through an observation of excess γ-ray photons at $\sim$ GeV energies from

* sm8383@nyu.edu; ORCID: 0000-0001-9088-7845
† kyle.cranmer@nyu.edu; ORCID: 0000-0002-5769-7094

DM-rich regions of the sky produced through the cascade of SM particles resulting from DM self-annihilation.

The *Fermi* γ-ray Galactic Center Excess (GCE), first identified over a decade ago using data from the *Fermi* Large Area Telescope (LAT) [1], is an excess of photons in the Galactic Center with properties—such as energy spectrum and spatial morphology—broadly compatible with expectation due to annihilating DM [2–16]. The nature of the GCE remains contentious however, with competing explanations in terms of a population of unresolved astrophysical point sources (PSs), in particular millisecond pulsars (MSPs), remaining viable [9, 17–25]. Analyses of the morphology of the excess have shown it to prefer a spatial distribution correlated with baryonic structures in the Galactic Center region rather than a distribution expected due to DM annihilation [15, 26, 27], although these conclusions can depend on details of the modeling [28, 29]. Studies leveraging the statistical distribution of photon counts in the Galactic Center have shown the γ-ray data to prefer a point source origin of the excess [30–33], a conclusion corroborated using wavelet-based techniques [31]. Recent studies have, however, pointed out the potential of unknown systematics, such as the poorly understood morphology of the diffuse foreground emission and the existence of unmodeled point source populations, to affect the conclusions of these analyses [34]. Ref. [32] showed that many of these issues can be ameliorated through the use of better diffuse foreground models, as well as by augmenting existing models with additional degrees of freedom.

The complexity associated with analyzing high-dimensional γ-ray maps—typically binned spatially using a pixelization scheme—has motivated the use of approximate likelihoods based on *e.g.*, the statistics of photon counts in individual pixels [30, 35, 36] or scale decomposition of the photon map using wavelet techniques [31, 37–39], in order to enable computationally tractable analyses. Under certain assumptions, using such approximations can capture all of the information contained in a given spatial model of the γ-ray data. This is the case, *e.g.*, for a likelihood based on the expected probability distribution of photon counts factorized across pixels when pixel-to-pixel correlations can be assumed to be negligible. When such correlations are present, however, the use of such approximations necessarily involves loss of information compared to that contained in the original γ-ray map.

Recent developments in machine learning have enabled analysis techniques that can extract more information from high-dimensional datasets, and can therefore be used to leverage more of the information contained in models of γ-ray emission. Machine learning methods have recently shown promise for analyzing γ-ray data [40] and specifically for understanding the nature of the *Fermi* GCE [41–43]. In particular, Ref. [41] used a method based on Bayesian neural networks in order to infer the flux fractions associated with various modeled components in the Galactic Center region, finding the GCE to be predominantly smooth in contrast to prior analyses depending the statistics of photon counts. Ref. [42] extended this framework, using a novel non-parametric approach [44] to extract the characteristics of the PS population associated with the GCE, finding a non-negligible portion of the emission to be attributable to a dim PS population. We will show the results of our analysis on *Fermi* data to be qualitatively consistent with those obtained in that work.

In this paper, we present a complementary approach that leverages recent developments in the field of simulation-based inference (SBI, also referred to as likelihood-free inference; see, *e.g.*, Ref. [45] for a recent review) in order to weigh in on the nature of the GCE. In particular, we use conditional density estimation techniques based on normalizing flows [46, 47] to characterize the contributions of various modeled components, including "clumpy" PS-like and "smooth" DM-like emission spatially tracing the GCE, to the γ-ray photon sky at $\sim$ GeV energies in the Galactic Center region. Rather than using hand-crafted summary statistics, we employ a graph-based spherical convolutional neural network architecture (previously utilized in Refs. [41, 42]) in order to extract summaries from γ-ray maps optimized for the downstream task of estimating the distribution of parameters characterizing the contribution of modeled components to the GCE. Unlike traditional approaches based on the statistics of photon counts, this approach allows us to capture more of the information contained in a model of the Galactic Center emission, and in particular implicitly uses the distribution of pixel-to-pixel correlations as an additional discriminating handle. As we will show, this makes our method more resilient to certain systematic uncertainties compared to these approaches. A schematic illustration of our method is presented in Fig. 1.

This paper is organized as follows. In Sec. II we describe our forward model and analysis framework based on neural simulation-based inference. In Sec. III we validate our pipeline on mock observations of the *Fermi* GCE. Section IV presents an application of the method to *Fermi* γ-ray data, including systematic variations on the analysis. In Sec. V we study the susceptibility of the analysis to known mismodeling of the signal and background templates. We conclude in Sec. VI.

II. METHODOLOGY

We begin by describing the various ingredients of our forward model and datasets used. After a brief summary of established methods based on explicit likelihoods, we detail our analysis methodology going over, in turn, the general principles behind simulation-based inference, posterior estimation using normalizing flows, and learning representative summary statistics from high-dimensional γ-ray maps with neural networks.

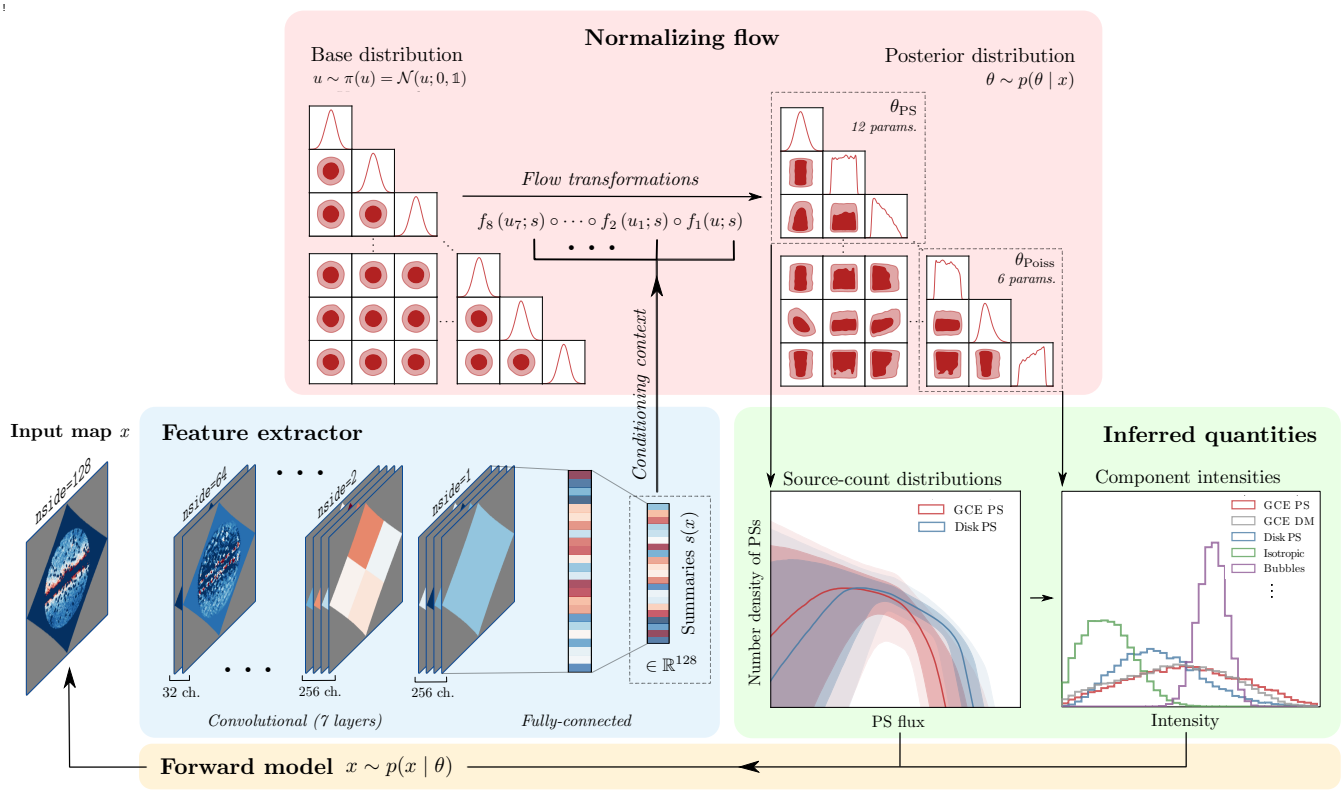

FIG. 1. A schematic overview of the inference framework used in this work. A normalizing flow is used to model posterior distribution of the parameters of interest characterizing the contribution of point source populations as well as diffuse ("smooth") components to the γ-ray data. The flow transformation from the base distribution to the posterior is conditioned on learned summaries of the γ-ray map extracted using a convolutional neural network. The normalizing flow and feature-extractor neural networks are trained simultaneously using maps simulated from the forward model. Once trained, samples from the flow can be generated conditioned on a new dataset of interest in order to obtain an estimate of the corresponding parameter posteriors, which can be used to infer physical quantities of interest such as source-count distributions of modeled PS populations as well as fluxes associated with the diffuse components. See Sec. II for a detailed description of the analysis pipeline.

A. Datasets and the forward model

Datasets and region of interest: We use the datasets and spatial templates from Refs. [48, 49] to create simulated maps of *Fermi*-LAT data in the Galactic Center region. The templates and data used correspond to 413 weeks of *Fermi*-LAT Pass 8 data taken between August 4, 2008 and July 7, 2016. The top quartile of photons as graded by quality of PSF reconstruction in the energy range 2–20 GeV and event class `ULTRACLEANVETO` are used. The conventional quality cuts are applied: zenith angle less than 90°, `LAT_CONFIG`==1, and `DATA_QUAL`==1.[1] The maps are binned spatially using the `HEALPix` [50] pixelization scheme with resolution parameter `nside`=128, roughly corresponding to pixel area $\sim 0.5\,\mathrm{deg}^2$. This dataset has been previously used in the literature for analyses based on explicit likelihoods [32–34] as well as machine learning-based analyses [41] for characterizing the GCE. All templates are normalized, per-pixel, within a region defined by $r < 30°$.

The inner region of the Galactic plane, where the observed emission is especially difficult to model, is masked at $|b| < 2°$, and a radial cut $r < 25°$ defines the region of interest (ROI) for our analysis. Even though the GCE is spatially confined to the inner 10–15° of the Galactic Center [10, 11], using a larger ROI improves the ability to constrain other spatially extended templates and helps mitigate spatial degeneracies that would otherwise crop up in a smaller ROI. On the other hand, using a ROI that is too large can exacerbate the effects of misspecified spatial templates [51]. We mask resolved PSs from the 3FGL catalog [52] at a radius of 0.8°, approximately corresponding to 99% PSF containment for photons in the data type employed [52].

Diffuse emission forward model: The simulated data maps are a combination of diffuse (alternatively referred to as smooth or Poissonian) and PS contributions. The smooth contributions include *(i)* the Galactic diffuse foreground emission, *(ii)* spatially isotropic emis-

[1] https://fermi.gsfc.nasa.gov/ssc/data/analysis/documentation/Cicerone/Cicerone_Data_Exploration/Data_preparation.html

sion accounting for, *e.g.*, uniform emission from unresolved sources of extragalactic origin, *(iii)* emission from resolved PSs included in the *Fermi* 3FGL catalog [52], and *(iv)* lobe-like emission associated with the *Fermi* bubbles [53]. Finally, *(v)* Poissonian DM-like emission is modeled using a line-of-sight integral of the (squared) generalized Navarro-Frenk-White (NFW) [54, 55] profile,

$$\rho_{\rm gNFW}(r) \propto \frac{1}{(r/r_{\rm s})^{\gamma}\,(1+r/r_{\rm s})^{3-\gamma}} \tag{1}$$

with inner slope $\gamma = 1.2$ motivated by previous GCE analyses [8, 10, 56]. Here, r is the radial distance from the Galactic Center, $r_{\rm s} = 20\,{\rm kpc}$ is the Milky Way scale radius, and we take $R_{\odot} = 8.2\,{\rm kpc}$ as the distance to the Galactic Center [57, 58]. Templates for components *(ii)*–*(iv)* are obtained from Ref. [49].

The Galactic foreground component accounts for γ rays produced due to cosmic rays interacting with interstellar gas and radiation, which makes up the majority of the observed emission in the Galactic Center region. In particular, bremsstrahlung emission from cosmic-ray electrons scattering off of gas as well as photons produced as a result of the decay of pions produced through cosmic ray protons scattering elastically with the gas both trace the Galactic gas distribution, modulated by the incoming cosmic ray density. These components exhibit structure on smaller angular scales. Additionally, inverse Compton (up-)scattering (ICS) of the interstellar radiation field by cosmic ray electrons produces an important component of the γ-ray Galactic diffuse emission which spatially traces the Galactic charge carrier density and does not show modulation on small scales. Normalizations of the gas-tracing components, subscripted 'brem/π^0', and the ICS-tracing component, subscripted 'ICS', are included separately in our forward model. Templates for these two components are described in our baseline configuration by Model O, introduced in Ref. [32]. There, it was found to be better fit, as quantified by the likelihood of describing the data up to Poisson noise, to the counts map in the Galactic Center region compared to diffuse foreground templates previously employed in GCE analyses. We explore the effect of variations on the assumed Galactic diffuse model in Sec. IV C.

The diffuse emission templates have been pre-smoothed with the *Fermi* point-spread function (PSF) at 2 GeV for the dataset employed, modeled as a pair of King functions.[2] The total diffuse emission in a given pixel p, x^p, is modeled as a Poisson realization of a linear combination of the diffuse templates T_i^p, where i indexes the individual templates, with their corresponding normalizations A_i regarded as parameters of the forward model; $x^p \sim {\rm Pois}\,(x^p \mid \sum_i A_i T_i^p)$.

[2] https://fermi.gsfc.nasa.gov/ssc/data/analysis/documentation/Cicerone/Cicerone_LAT_IRFs/IRF_PSF.html

PS emission forward model: Assuming the locations of individual PSs are not known *a-priori*, the statistics of multiple PS populations can be completely specified through *(i)* their spatial distribution, described by templates T^p discretized over pixels p, *(ii)* the distribution of expected photon counts S contributed by each PS, $p(S)$, and *(iii)* the distribution of the number of PSs for each population. Additionally, the modeled instrumental point-spread function quantifies the spatial distribution of photon counts sourced by an individual PS around its location due to the finite angular resolution of the LAT instrument.

Here, we parameterize the distributions of photon counts S contributed by each PS through a doubly-broken power law,

$$p(S \mid \theta_{\rm PS}) \propto \begin{cases} \left(\frac{S}{S_{\rm b,1}}\right)^{-n_1}, & S \geq S_{\rm b,1} \\ \left(\frac{S}{S_{\rm b,1}}\right)^{-n_2}, & S_{\rm b,1} > S \geq S_{\rm b,2} \\ \left(\frac{S_{\rm b,2}}{S_{\rm b,1}}\right)^{-n_2}\left(\frac{S}{S_{\rm b,2}}\right)^{-n_3}, & S_{\rm b,2} > S \end{cases} \tag{2}$$

specified by the break locations $\{S_{\rm b,1}, S_{\rm b,2}\}$, spectral indices (slopes) $\{n_1, n_2, n_3\}$, and appropriately normalized to unity. Higher subscript indices correspond to dimmer parts of the source-count distribution. Together, we denote these parameters by $\theta_{\rm PS}$.

The PS component of the simulated *Fermi* map is created as follows, practically implemented using the code package `NPTFit-Sim` [59]. The total number of PSs to be simulated is drawn as $n \sim {\rm Pois}(n \mid n_{\rm pix}\lambda)$, where $n_{\rm pix}$ is the number of pixels in the ROI and λ is the mean number of PSs per pixel. The sample of PS angular positions $\{r_n\}$ is drawn from a PDF constructed by linearly interpolating the relevant pixel-wise spatial template T^p; $\{r_n\} \sim p(r) \propto T(r)$. The expected number of photons emitted by each PS, indexed by i, is drawn by first sampling from the mean PDF of expected photon counts in Eq. (2), $S \sim p\,(S \mid \theta_{\rm PS})$, and scaling this as $S_i = S\epsilon(r_i)/\langle\epsilon\rangle$ to account for variations in the *Fermi* exposure at the sampled PS positions, $\epsilon(r_i)$, over the mean exposure $\langle\epsilon\rangle$ in the ROI. The actual sample of photon counts emitted by the simulated PSs, $\{x_n\}$, is taken to be a Poisson realization of this expectation; $x_i \sim {\rm Pois}\,(x_i \mid S_i)$. Given the angular positions of and photon counts emitted by PSs $\{r_n, x_n\}$, the radial coordinates of photons relative to the positions of PSs are drawn following the modeled *Fermi* PSF, with the azimuthal coordinates sampled uniformly assuming a spherically-symmetric PSF. This procedure is repeated for each PS population, and the final simulated PS map is constructed by binning the sampled photon positions within the ROI according to the pixelization scheme used. In practice, in order to avoid computational costs associated with simulating a large number of low-flux PSs, the dim component of the PS population below a specified threshold is partially accounted for in the DM-like component, as described in detail towards the end of this

subsection.

In the NPTF literature, modeled PS populations are often compactly described through the so-called source-count distribution (SCD) $\mathrm{d}^2N/\mathrm{d}S\mathrm{d}\Omega$, which quantifies the differential number density of sources per unit angular area emitting S photons in expectation. The source-count distribution jointly describes the distribution of photon counts from individual PSs $p(S \mid \theta_{\rm PS})$ and their mean per-pixel abundance λ, and is related to these as

$$\frac{\mathrm{d}^2N}{\mathrm{d}S\mathrm{d}\Omega} = \lambda\, p(S \mid \theta_{\rm PS})/\Omega_{\rm pix} \tag{3}$$

where the the pixel area $\Omega_{\rm pix}$ is used to convert the per-pixel source count to per-area, rendering it agnostic to pixel size. We will present our results in terms of the source fluxes ($\mathrm{d}^2N/\mathrm{d}F\mathrm{d}\Omega$) rather than expected counts ($\mathrm{d}^2N/\mathrm{d}S\mathrm{d}\Omega$), with the conversion $S = \langle\epsilon\rangle F$ where $\langle\epsilon\rangle$ is the mean exposure in the region considered. In the analysis ROI used here, the mean exposure is $\langle\epsilon\rangle \simeq 7 \times 10^{10}\,\mathrm{cm}^2\,\mathrm{s}$. For brevity, we will denote the distribution as $\mathrm{d}N/\mathrm{d}F$, leaving the per-area normalization implicit.

In this paper, we consider two independent PS populations: *(i)* those spatially correlated with the GCE, modeled the same as the Poissonian counterpart using a line of sight integral of the (squared) generalized NFW profile in Eq. (1) with $\gamma = 1.2$, and *(ii)* those spatially correlated with the Galactic disk, modeled by a doubly-exponential profile motivated by studies of the spatial distribution of Galactic millisecond pulsar populations [60, 61],

$$\rho_{\rm Disk}(R, z) \propto \exp\left(-\frac{R}{R_{\rm d}}\right) \exp\left(-\frac{|z|}{z_{\rm s}}\right) \tag{4}$$

where R and z are the radial and vertical Galactic cylindrical coordinates, and the disk scale height and radius are set to $z_{\rm s} = 0.3\,\mathrm{kpc}$ and $R_{\rm d} = 5\,\mathrm{kpc}$ respectively in the baseline scenario. The final maps are obtained by combining the diffuse and PS emission components of the forward model.

Prior specification: We use uniform priors for the normalization factors of the Poissonian templates. For the PS components, we use uniform priors on the parameters that characterize the broken power-law distribution of photon counts within the intervals defined below. The break associated with the brighter end of the SCD, $S_{\rm b,1} \in [5, 40]$ photons, reflects a 'turn-on' associated with the source luminosity function, above which sources are either individually resolved or not inferred to exist. This turn-on is further enforced by specifying a highest slope $n_1 \in [10, 20]$ that is steeply rising with decreasing S. The middle slope, $n_2 \in [1.1, 1.99]$, is associated with the physical luminosity function of the source population, typically expected to be in this specified range for a Galactic pulsar population [25].

Emission from a PS population is nearly degenerate but still statistically distinguishable from that following a Poisson distribution when associated with sources emitting $\sim$ 0.1–1 counts in expectation [42]; in practice, however, residual effects of model misspecification and degeneracies between multiple PS populations can make characterizing the source-count distribution in this low-photon regime challenging [33]. The dimmer break, $S_{\rm b,2} \in [0.1, 4.99]$ photons, therefore specifies a regime where we do not attempt to explicitly characterize the PS population. This is enforced by allowing for a lowest slope $n_3 \in [-10, 1.99]$ that is steeply falling with decreasing S, encouraging the SCD to turn off in this regime. This gives preference to the smooth component in absorbing flux close to and below the single photon regime, and our analysis therefore conservatively aims to estimate a *lower bound* on the contribution of PS emission to the GCE by primarily considering the relatively bright regime of the source-count distribution. In order to quantify the effect of the prior in the low-photon regime, we also explore an alternative specification where the lower range of the upper break prior is brought down to a single photon, $S_{\rm b,1} \in [1, 30]$ photons, giving the PS component more overlap closer to the degeneracy regime and thus allowing it to account for more of the dim emission. In App. A, we show how the prior choices map onto the source-count distribution for the baseline and alternative configurations.

The overall abundance of PSs associated with a modeled population is specified as follows. Rather than sampling the expected number of PSs per pixel λ with a uniform prior, we instead uniformly sample a related parameter $\langle S^{\rm PS}\rangle = \int \mathrm{d}S\, S\, \lambda\, p(S \mid \theta_{\rm PS})$, the expected number of photon counts contributed by the PS population per pixel. Similarly, for the Poissonian GCE component, the template normalization $A_{\rm GCE}$ is reparameterized through a constant multiplicative factor into the mean per-pixel expected counts $\langle S^{\rm Poiss}_{\rm GCE}\rangle$. This is done in order to place the flux distribution of the PS-like component $\langle S^{\gtrsim 1\,\rm ph}\rangle$ on the same 'footing' as that associate with smooth emission $\langle S^{\lesssim 1\,\rm ph}\rangle$. Since a uniform prior on λ would not correspond to a uniform prior on $\langle S^{\rm PS}\rangle$, these reparameterizations *a-priori* distribute photons approximately uniformly among the regimes $\langle S^{\lesssim 1\,\rm ph}\rangle$ and $\langle S^{\gtrsim 1\,\rm ph}\rangle$. We note here the possibility of using other prior prescription proposed in the literature, *e.g.* in Ref. [62] where, in addition to enforcing an equivalence between dim PSs and smooth emission (rather than enforcing a distinction between relatively-bright PSs and smooth emission as done here), the SCD slopes are specified in terms of the angles between adjacent parts of the broken power law and the break positions are specified as a fraction relative to the brightest break.

The forward model is thus specified by a total of 18 parameters—6 for the overall normalizations of the Poissonian templates $\{\langle S^{\rm Poiss}_{\rm GCE}\rangle, A_{\rm brem/\pi^0}, A_{\rm ICS}, A_{\rm iso}, A_{\rm bub}, A_{\rm 3FGL}\}$, and 6×2 parameters modeling the source-count distributions associated with GCE-correlated and disk-correlated PS populations $\{\langle S^{\rm PS}\rangle, n_1, n_2, n_3, S_{\rm b,1}, S_{\rm b,2}\}$. The priors used in the forward model are summarized in Tab. I. In order to improve sample efficiency, the priors

Poissonian		PS-like (GCE and disk)	
Parameter	**Prior range**	**Parameter**	**Prior range**
$\langle S^{\rm Poiss}_{\rm GCE}\rangle$	[0, 2.5] ph	$\langle S^{\rm PS}\rangle$	[0, 2.5] ph
$A_{\rm brem/\pi^0}$	[6, 12]	n_1	[10, 20]
$A_{\rm ICS}$	[1, 6]	n_2	[1.1, 1.99]
$A_{\rm iso}$	[0, 1.5]	n_3	[-10, 1.99]
$A_{\rm bub}$	[0, 1.5]	$S_{\rm b,1}$	[5, 40] ph
$A_{\rm 3FGL}$	[0, 1.5]	$S_{\rm b,2}$	[0.1, 4.99] ph

TABLE I. Parameter priors used for the components of the forward model described in Sec. II A. All priors are uniform within the ranges specified. Priors on the Poissonian components, corresponding to overall normalization, are shown in the left table column, while those of the GCE- and disk-correlated PS components, parameterized according to Eq. (2), are shown in the right table column. The overall normalizations of the Poissonian GCE and PS-like components are parameterized through the mean number of photon counts contributed by the respective components in the ROI.

are motivated by posteriors obtained from a Poissonian template fit to the real *Fermi* data.

B. Inference with likelihoods based on simplified data representations

Before discussing the methodology used in this paper in detail, we will provide a brief overview of an established class of techniques—Non-Poissonian template fitting—that have been successfully deployed in order to characterize the contribution of PSs to the GCE. We will focus on a schematic description of the method without delving into details of the implementation, aiming to highlight the elements that introduce approximations and where our ML-based approach differs.

A central object in statistical inference is the likelihood $p(x \mid \theta)$, which quantifies the probability of an observation x given parameters of interest θ. In the simplest incarnation of astrophysical template-fitting methods dealing with counts data, the likelihood of the map x in the region of interest is computed as a pixel-wise product of Poisson likelihoods with mean given by a linear combination of spatial templates T_i^p, $p(x \mid \theta) = \prod_p \mathrm{Pois}\,(x^p \mid \sum_i A_i T_i^p)$, where normalizations A_i of the respective spatial templates are the parameters of interest. This captures the diffuse part of the forward model described in Sec. II A, and inference here can easily be performed within a frequentist or Bayesian framework.

In practice, unobserved latent variables z are often involved in the data-generation process, and computing the likelihood involves marginalizing over the latent space, $p(x \mid \theta) = \int \mathrm{d}z\, p(x \mid \theta, z)$. In typical problems of interest, the high dimensionality of the latent space often means that this integral is intractable, necessitating simplifications in statistical treatment as well as theoretical modeling. For the forward model in Sec. II A, the presence of PS populations introduces a large number of latent variables, specifically the position of and counts emitted by each PS. Ignoring the contribution from diffuse components for the moment and considering only a single isotropically-distributed PS population, the likelihood for the map x in the region of interest is given by

$$p(x \mid \lambda, \theta_{\rm PS}) = \sum_{n=0}^{\infty} \int \mathrm{d}^n z\, p\,(n \mid \lambda)\; p(z \mid \theta_{\rm PS})\, p(x|z), \quad (5)$$

where $\theta_{\rm PS}$ parameterize the distribution of photon counts from individual PSs. n is the total number of PSs in the ROI, with the sum running over all possible number of PSs. This high-dimensional integral is, for all practical purposes, computationally intractable. The presence of a finite instrumental PSF introduces additional latent processes, decoupling the positions of the photons and PSs. Given these difficulties, a simplification of the problem setting is typically required to make further progress.

The 1-point PDF (probability distribution function) framework, first introduced in the context of γ-ray analyses in Ref. [35] and extended to allow for non-trivial spatial PS distributions in Refs. [30, 36] under the name of non-Poissonian template fitting (NPTF), considers a simplification of the problem by computing the pixel-wise likelihood assuming each pixel to be statistically independent (*1-point* then referring to values over individual, independent spatial positions in the sky). This significantly reduces the latent space dimensionality by eliminating the positions of individual PSs as latent variables. Since non-Poissonian template fitting has been widely used in analyses of the GCE, we briefly outline the basic philosophy behind this method, pointing the interested reader to a more detailed discussion as well as numerical implementations in Refs. [30, 49].

Since emission from each PS can be regarded as independent conditioned on $\theta_{\rm PS}$, the probability of a given PS, indexed i, emitting x_i^p photons in a pixel p is given by

$$p(x_i^p \mid \theta_{\rm PS}) = \int \mathrm{d}S_i\, p(S_i \mid \theta_{\rm PS})\, p(x_i^p|S_i), \quad (6)$$

where S_i are the expected photon counts from the PS following some probability distribution parameterized by $\theta_{\rm PS}$, in this case following a doubly-broken power law with parameters $\theta_{\rm PS} = \{n_1, n_2, n_3, S_{\rm b,1}, S_{\rm b,2}\}$, and $p(x_i^p|S_i)$ is the distribution of actual counts given latent S_i, assumed to follow a Poisson distribution on S_i. The probability of having a total of x_p counts in a pixel from multiple PSs is then described by a multinomial distribution, subject to the constraint that the total number

of counts be equal to the observed counts:

$$p(x^p\,|\lambda,\theta_{\rm PS}) = \sum_{n=0}^{\infty} p\left(n \mid \lambda\right) \sum_{n_j} \delta\left(\sum_j n_j j - x^p\right) \times \delta\left(\sum_j n_j - n\right) \frac{n!}{\prod_j n_j} \prod_{j=1}^{n} p(x_i^p = j \mid \theta_{\rm PS})^{n_j}, \tag{7}$$

where n_j is the number of PSs contributing j counts. The distribution of the number of PSs in a pixel is usually assumed to follow a Poisson distribution on the mean expected number of PSs λ *i.e.*, $p(n \mid \lambda) = {\rm Pois}(n \mid \lambda)$. In this case, the sum over n can be eliminated and the distribution of observed counts is given by

$$p(x^p\,|\lambda,\theta_{\rm PS}) = \sum_{n_j=0}^{\infty} \delta\left(\sum_j n_j j - x^p\right) \times \prod_j {\rm Pois}\left(n_j \mid \lambda\, p(x_i^p = j \mid \theta_{\rm PS})\right). \tag{8}$$

where $p(x_i^p = j \mid \theta_{\rm PS})$ is given by Eq. (6). While not immediately obvious from this expression, eliminating the positions of individual PSs as latent parameters as well as the sum over the possible number of PSs n renders the per-pixel likelihood tractable, and the total data likelihood can then be computed as a product over pixels, $p(x \mid \lambda, \theta_{\rm PS}) = \prod_p p(x^p \mid \lambda, \theta_{\rm PS})$.

We emphasize that we have only provided a brief overview of the NPTF method here, with further analytic simplifications, extensions to approximately incorporate the effect of non-trivial instrumental point-spread function and exposure, as well as a numerical recipe for evaluating the likelihood described in detail in Ref. [49]. We note that including the effect of a finite point-spread function in the NPTF framework renders the per-pixel likelihood only approximately correct, since this introduces correlations across pixels over the scale of the PSF size. Previous studies have shown this approximation to be accurate enough for the present problem when using a pixel size of the order of the PSF size itself [33]. Further generalizations of the method that can account for more extreme variations in the instrumental point-spread function and exposure without resorting to an approximate treatment—necessary for application to *e.g.* X-ray data—were introduced and studied in Ref. [62].

Probabilistic cataloging [63, 64] is another method that has been proposed for characterizing the sub-threshold contribution of PS populations in counts data and found application in γ-ray analyses [65]. This technique keeps the latent variables in Eq. (5) *i.e.*, the positions and expected fluxes of individual PSs, as parameters of interest, and uses trans-dimensional sampling techniques to obtain the distribution over possible catalogs of unresolved PS populations. For computational reasons, probabilistic cataloging techniques generally require a strong assumption on the nature of the putative PS population and can thus produce highly prior-dependent results.

In this paper, we show results on *Fermi* data using the NPTF algorithm in order to establish a comparison point to previous GCE studies employing the method. We perform these analyses within a Bayesian framework, obtaining an approximation to the posterior distribution $p(\theta \mid x) = p(\theta)\, p(x \mid \theta)/\mathcal{Z}$, where $\mathcal{Z} \equiv p(x)$ the Bayesian evidence. We use the NPTF likelihood implemented in `NPTFit` [49] and obtain representative posterior samples over the parameters of interest described in Sec. II A using nested sampling [66, 67] implemented in `dynesty` [68]. The static variant of the nested sampling algorithm is run in its default configuration with 1000 live points, stopping when the estimated contribution of the remaining posterior volume to the log-evidence falls below $\Delta \log \mathcal{Z} < 0.1$. Although it's possible to correct for non-uniform exposure within the NPTF framework by considering independent sub-regions with different exposure values, given the fairly uniform *Fermi* exposure in the Galactic Center region we use the mean exposure in our NPTF benchmarks for simplicity.

1-point PDF-based techniques, and in particular NPTF, have been widely applied for characterizing γ-ray PS populations below the *Fermi* detection threshold, both in relation to the GCE [30, 32, 69–71] and more generally *e.g.*, for characterizing the contribution of extragalactic PSs at high latitudes [72–74] and for searching for a DM annihilation signal from Galactic subhalos [75]. It has recently been pointed out, however, that signal and foreground mismodeling associated in particular with the emission in the Galactic Center region can hamper the ability to accurately characterize the contribution of PSs to the GCE [34, 69]. In particular, Refs. [30, 33, 34] pointed out that spurious residuals associated with foreground mismodeling can lead to the mischaracterization of a purely DM signal as a population of PSs. Ref. [32] recently showed that many of the issues associated with the expression of such effects in *Fermi* data could be mitigated through the use of better Galactic foreground models along with affording them more degrees of freedom on large angular scales. Refs. [69, 70] further showed and described analytically how mismodeling, in particular an unmodeled asymmetry in a DM signal, could lead to the spurious inference of PSs in NPTF analyses of the GCE.

The fact that NPTF analyses rely on a simplified per-pixel likelihood can make them especially susceptible to the effects of model misspecification (alternatively referred to as mismodeling)—systematic departures of the forward model from the true data-generating process. This can be intuited from the fact that, assuming a corresponding permutation of template pixel labels, the NPTF likelihood is invariant to a permutation of pixels within the analysis ROI. This means that residuals associated with a misspecified background model can mimic the effect of PSs through the distribution of their photon counts, disregarding the specific spatial structure associated with a PS population. The full likelihood sketched out in Eq. (5) and implicitly defined by the forward model

described in Sec. II A contains significantly more spatial structure than is encoded in the distribution of photon counts, and in particular accounts for the distribution of pixel-to-pixel correlations in the γ-ray map; see also Ref. [42] for an extended discussion on this point. In the rest of this section, we will describe the building blocks of our machine learning-based method that, in contrast to NPTF, aims to estimate the likelihood implicitly associated with the γ-ray forward model, leveraging pixel-to-pixel spatial correlations with the overall aim of more robustly characterizing the PS contribution to the GCE.

C. Simulation-based inference

Simulation-based inference (SBI) refers to a class of methods for performing inference when the data-generating process does not have a tractable likelihood. This is the case for the model described in Sec. II A, where the likelihood in Eq. (5) cannot be used explicitly for practical purposes without further simplifications. The model is then defined through a simulator as a probabilistic program, often knows as a forward model. Samples x from the simulator then *implicitly* define a likelihood, $\{x\} \sim p(x \mid \theta)$. In the simplest existing realizations of SBI, simulated samples $\{x\}$ can be compared to a given dataset of interest x', with the approximate posterior defined by parameter values whose corresponding samples most closely resemble x' according to some distance metric. Such methods—usually grouped under the umbrella of Approximate Bayesian Computation (ABC) [76]—are not uncommon in astrophysics and cosmology. Nevertheless, they suffer from several downsides. The curse of dimensionality usually necessitates reduction of data to representative, hand-crafted, lower-dimensional summary statistics $s(x)$, resulting in loss of information. A notion of distance in the lower-dimensional summaries domain as well as a tolerance threshold, $||s(x) - s(x')|| < \epsilon$, is necessary to trade off between precision and sample efficiency, leading to inexact inference. Additionally, the ABC analysis must be performed anew for each new target dataset.

Recent advances in machine learning, particularly the proliferation of neural network architectures suited to a variety of data structures and the development of algorithms that can efficiently approximate functions and distributions in high dimensions, have galvanized the field of simulation-based inference, substantially increasing its domain of applicability; see Ref. [45] for a review of recent developments. In the following subsections, we will describe the specific SBI methods employed in this work for parameter estimation on the forward model described in Sec. II A.

D. Conditional density estimation with normalizing flows

We approximate the joint posterior $p(\theta \mid x)$ over the parameters of interest θ through a distribution $\hat{p}_\phi(\theta \mid s)$ conditioned on summaries $s = s(x)$ from simulated samples $\{x\}$, parameterized by ϕ and modeled by a neural network. This class of simulation-based inference techniques, known as conditional density estimation [77, 78], directly models the posterior distribution given a set of samples $\{x\} \sim p(x \mid \theta)$ produced from the forward model, where parameters θ are sampled according to some prior proposal distribution $\{\theta\} \sim p(\theta)$. We note that, given the absence of explicit labels associated with the sampled parameters of interest, estimating the probability density is an example of an unsupervised learning problem.

Normalizing flows: In this paper we employ normalizing flows [46, 47], a class of models that provide an efficient way of constructing flexible and expressive high-dimensional probability distributions. Normalizing flows model the (conditional) distribution over the parameters of interest $\hat{p}_\phi(\theta \mid s)$ as a series of transformations, denoted by f such that $\theta = f(u)$, from a simple base distribution $\pi(u)$ to the target distribution. Suppressing the conditional dependence on s for the moment for simplicity, we have

$$\hat{p}(\theta) = \pi(u) \left| \det \left(\frac{\partial u}{\partial \theta} \right) \right| = \pi(f^{-1}(\theta)) \left| \det J_{f^{-1}}(\theta) \right| \quad (9)$$

where $\det J_{f^{-1}}$ is the Jacobian of the inverse transformation f^{-1}.

The defining characteristic of transformations in flow-based models is that they be diffeomorphic *i.e.*, f be differentiable and invertible with a differentiable inverse. This renders the Jacobian and inverse in Eq. (9) computable, allowing for the evaluation of the probability density of the target distribution $\hat{p}(\theta)$ at a given parameter point θ once the transformation is defined. In practice, the transformation f (or f^{-1}) is chosen such that $\det J$ can be efficiently computed and is usually defined by a neural network, and the base distribution $\pi(u)$ is chosen to be a standard Gaussian $u \sim \mathcal{N}(u; 0, \mathbb{1})$, which we follow here.

A crucial property of diffeomorphic transformation such as those that define normalizing flows is that multiple transformations can be chained together through composition. Given two transformations f_1 and f_2, their composition will also be differentiable and invertible: $\det J_{f_1 \circ f_2}(\theta) = \det J_{f_2}(f_1(\theta)) \det J_{f_1}(\theta)$ and $(f_2 \circ f_1)^{-1} = f_1^{-1} \circ f_2^{-1}$. This can be used to define more expressive probability distributions by chaining together several flow transformation. 'Flow' thus refers to the trajectory through which parameters in the simple base distribution are transformed into the target parameter space, and 'normalizing' refers to the inverse transformation

into the base distribution. Flow-based models are *generative*—given a new dataset x', it is easy to sample from the base distribution and then run the forward transformation conditioned on x', obtaining a set of parameter samples representative of the posterior distribution, $\{\theta\} \sim \hat{p}(\theta|x')$.

A number of methods have been proposed for defining the flow transformation *e.g.*, based on affine transformations [79–82], spline-based transformations [83, 84], and continuous-time transformations [85]. We refer to Ref. [46] for a recent review of normalizing flows, including details of practical implementations as well as an overview of proposed methods.

Masked autoregressive flows for (conditional) density estimation: In this paper we use Masked Autoregressive Flows (MAFs) [79] to define the flow transformation. Autoregressive models can be used to learn a complex joint probability density $p(\theta)$ as a product of one-dimensional conditional densities where each θ_i depends only on the previous $\theta_{1:i-1}$ in the parameter sequence: $p(\theta) = \prod_i p(\theta_i \mid \theta_{1:i-1})$. The MAF is built using blocks of affine transformations subject to the autoregressive constraint; for a single block, the affine transformation from u to θ is expressed as

$$\theta_i = u_i \cdot \exp \alpha_i + \mu_i \tag{10}$$

where $\mu_i = g_{\mu_i}(\theta_{1:i-1}; s)$ and $\alpha_i = g_{\alpha_i}(\theta_{1:i-1}; s)$ are scaling and shift factors modeled by neural networks and additionally parameterized by summaries s from the forward model. The autoregressive property is enforced by masking out connections between network layers using the recipe introduced in Ref. [86]. The inverse transformation is easily identified from Eq. (10). This allows for an analytically tractable Jacobian determinant, for an N-dimensional distribution given by

$$\left|\det J_{f^{-1}}(\theta)\right| = \exp\left(-\sum_{i=1}^{N} \alpha_i\right) \tag{11}$$

and a forward pass through the flow according to Eq. (10). Multiple transformations f_j can be composed together in order to model more expressive posteriors,

$$\hat{p}(\theta \mid s) = \pi\left(f^{-1}(\theta)\right) \prod_{j=1}^{K} \left|\det J_{f_j^{-1}}(u_{j-1})\right| \tag{12}$$

where we have reinstated the conditional dependence on data summaries s, keeping it implicit in the transformations on the right hand side. The log-probability of the posterior can then be computed using Eq. (11):

$$\log \hat{p}(\theta \mid s) = \log\left[\pi\left(f^{-1}(\theta)\right)\right] - \sum_{j=1}^{K}\sum_{i=1}^{N} \alpha_i^j, \tag{13}$$

which acts as the optimization objective during training. Here, we use 8 MAF transformations, each made up of a 2-layer masked neural network with 128 hidden units and tanh activations. The ordering of parameters in the autoregressive sequence is randomly permuted between successive transformations in order to reduce dependence on the specific ordering of input variables. Each transformation is conditioned on summaries $s(x)$ extracted from the γ-ray maps x (described in the next section below) by including these as additional inputs into the transformation block *i.e.*, the scaling and shift factors in Eq. (10) can be expressed as $\mu_i = g_{\mu_i}(\theta_{1:i-1}; s(x))$ and $\alpha_i = g_{\alpha_i}(\theta_{1:i-1}; s(x))$.

E. Learning summary statistics with neural networks

The curse of dimensionality makes it computationally inefficient to condition the density estimation task on the raw dataset x *i.e.*, the γ-ray pixel counts map in the region of interest (ROI). Representative summaries $s = s_\varphi(x)$ of the data can therefore be used in order to enable a tractable analysis, where φ parameterizes the data-to-summary transformation. Although many choices for data summaries are possible—*e.g.*, a Principal Component Analysis (PCA) or angular power spectrum decomposition of the photon counts map, or simply a histogram of the photon counts—in this paper, we use a neural network to automatically learn low-dimensional summaries that are optimized for the specific downstream task at hand of estimating the posterior distributions of the parameters associated with the forward model.

Graph construction and network architecture: The `DeepSphere` architecture [87–89], with a configuration similar to and inspired by that employed in Ref. [41], is used to extract representative summaries from γ-ray maps and is briefly outlined here. `DeepSphere` is a graph-based spherical convolutional neural network (CNN) architecture tailored to data sampled on a sphere, and in particular is able to leverage the hierarchical structure of data in the `HEALPix` representation. This makes it well-suited for our purposes.

The `HEALPix` sphere can be represented in terms of a weighted undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ where $\mathcal{V}$ is the set of $N_{\rm pix} = |\mathcal{V}|$ vertices, $\mathcal{E}$ is the set of edges connecting pixels, and A is the weighted adjacency matrix. Each pixel i is represented by a vertex $v_i \in \mathcal{V}$ and is connected to the 8 (or 7, depending on the pixel) vertices v_j which represent the neighboring pixels j of pixel i, forming edges $(v_i, v_j) \in \mathcal{E}$. The weights of the adjacency matrix over neighboring pixels (i, j) are given by $A_{ij} = \exp\left(-\|r_i - r_j\|_2^2/\rho^2\right)$ where r_i specifies the 3-dimensional coordinates of pixel i. The kernel widths ρ at a given `HEALPix` resolution are obtained from Ref. [87], which used empirical measures of rotational equivariance in order to optimize for this hyperparameter.

We use the combinatorial graph Laplacian, defined as

$L = D - A$, where D is the diagonal degree matrix, and which can be used to define a Fourier basis on a graph. By construction being symmetric and positive semi-definite, the graph Laplacian can be decomposed as [90] $L = U\Lambda U^T$, where U is an orthonormal eigenvector matrix and Λ is a diagonal eigenvalue matrix. The Laplacian eigenvectors then define the graph Fourier basis, with the Fourier transform $\tilde{x}$ of a signal x on a graph being its projection $\tilde{x} = U^T x$. Given a convolutional kernel h, graph convolutions can be efficiently performed in the Fourier basis as $h(L)x = Uh(\Lambda)U^T x$ [90].

The isotropic `DeepSphere` convolutional kernel h is defined as a linear combination of Chebychev polynomials, $h(L) = \sum_{k=0}^{K} c_k T_k(L)$ where T_k are the order-k Chebyshev polynomials and c_k are the $K+1$ filter coefficients which are the trainable parameters to be learned during model optimization. The graph filtering operation can then be expressed as

$$h(L)x = U\left(\sum_{k=0}^{K} c_k T_k(\Lambda)\right)U^T x = \sum_{k=0}^{K} c_k T_k(L)x. \quad (14)$$

We set $K = 5$, having checked that larger values do not quantitatively affect the results of the analysis. $T_k(\Lambda)$ acts on the diagonal eigenvalue matrix, $T_k(\Lambda_{ii}) = T_k(\Lambda)_{ii}$.

Following Refs. [41, 88], the feature extraction architecture is built out of graph convolutional layers which involve progressively coarsening the pixel representation of the γ-ray maps while increasing the number of filter channels at each step. The input map corresponds to the 16,384 pixels at `HEALPix` resolution `nside`=128 in the nested pixel ordering within the single pixel corresponding to `nside`=1 covering the Galactic Center region, with the masked pixels set to zero. Each graph convolution operation is followed by a batch normalization, a ReLU nonlinearity, and a max pooling operation which downsamples the representation by a factor of 4 into the next coarser `HEALPix` resolution, starting with input maps at `nside`=128 until a single pixel channel at `nside`=1 remains after the final convolutional layer. All together, 7 layers of this kind are employed. The number of filter channels is doubled at each convolutional layer until a maximum of 256.

The output of the final convolutional layer is augmented with 2 additional auxiliary variables—the log-mean and log-standard deviation of the γ-ray map within the region of interest—and passed, via a ReLU nonlinearity, through a fully-connected layer with 1024 hidden units outputting a desired number of summary features, which we take as 128 in our baseline configuration. Pixels outside of the ROI as well as masked PSs are set to zero in the input maps. All input maps are standardized to zero mean and unit variance across the training sample.

Using a convolutional neural network-based feature extractor, we implicitly use an approximation to the full data likelihood in Eq. (5) associated with our forward model of emission in the Galactic Center region. The method is thus able to capture pixel-to-pixel correlations in the γ-ray map, mitigating some of the limitations of approximate likelihood-based methods described in Sec. II B.

Optimization, training, and evaluation: The optimization objective in Eq. (13), $\log \hat{p}_\phi(\theta \mid s_\varphi(x))$, is used to train the graph convolutional and normalizing flow neural networks simultaneously, optimizing their respective parameters $\{\varphi, \phi\}$. 10^6 samples are generated using the prior proposal distribution of parameters given in Tab. I, and models are optimized with batch size 256 using the `AdamW` [91, 92] optimizer with initial learning rate 10^{-3} and weight decay 10^{-5}, using cosine annealing to decay the learning rate across epochs. Training proceeds for up to 30 epochs with early stopping if the validation loss, evaluated on 15% of samples held out, has not improved after 8 epochs.

After training, given a new dataset of either real or simulated *Fermi* data in our ROI, the posterior is obtained by drawing samples from the flow within the prior distribution using rejection sampling, conditioning each flow transformation on summaries extracted by the convolutional neural network with the new dataset as input. The model is *amortized*, which means that after the upfront cost of training the neural network, the required number of posterior samples corresponding to a new dataset can be obtained on a few-second timescale. This makes it efficient to validate the performance of a trained model using mock data, which we do in the following section before applying the method to *Fermi* data.

III. TESTS ON SIMULATED DATA

We begin by validating our pipeline on simulated *Fermi* data. We create simulated datasets with the parameters of interest in the forward model fixed to posterior medians obtained in a fit of the baseline model to real *Fermi* data, and test the ability of our model to infer the presence of either DM-like or PS-like signals on top of the modeled astrophysical background.

Figure 2 shows results of the analysis conditioning the trained baseline model on five simulated maps where the GCE consists of purely DM-like emission, drawing 20,000 representative samples in each case. The left column shows the median (solid lines) as well as middle-68/95% containment (dark/light shaded regions) of the posteriors on the source-count distributions $F^2 \mathrm{d}N/\mathrm{d}F$ of GCE-correlated (red) and disk-correlated (blue) PS emission, evaluated point-wise in flux F. The dashed grey vertical lines correspond to the flux associated with a single expected photon count per source (below which Poissonian and PS-like emission is expected to be nearly degenerate) and the approximate 1-σ threshold for detecting individual sources (below which the degeneracy is often observed in practice [32, 33]). The middle column shows the posteriors on various modeled emission components, excluding

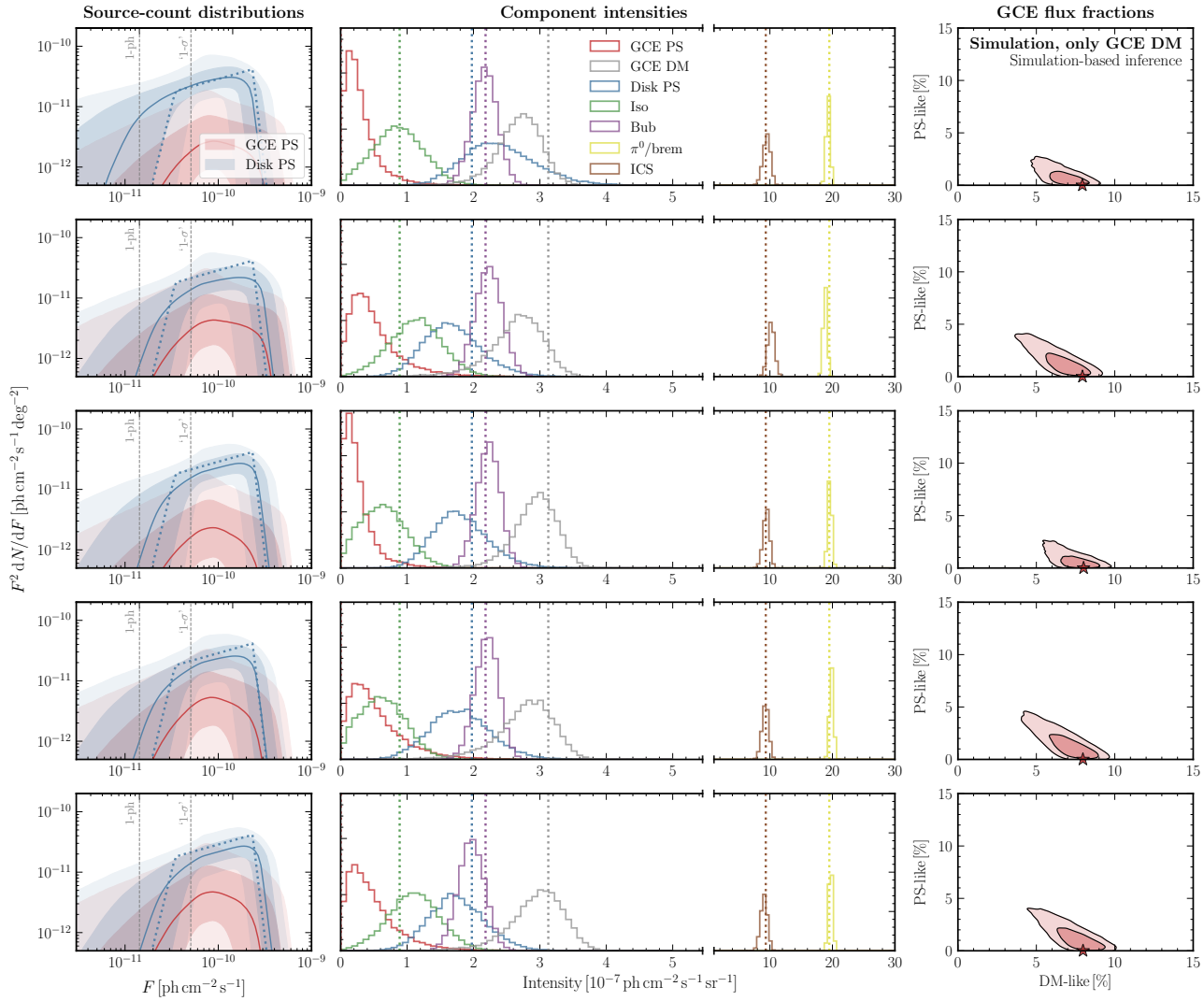

FIG. 2. Results of the analysis on simulated *Fermi* data where the GCE consists of purely DM-like emission, with different rows corresponding to five different simulated realizations. The left column shows the inferred source-count distribution posteriors for GCE-correlated (red) and disk-correlated (blue) PSs. Dashed vertical lines corresponding to the flux associated with 1 expected photon per source and the approximate 1-σ threshold for detecting individual sources are shown for reference. Solid lines correspond to the inferred posterior median, and the lighter and darker bands represent the middle-68% and 95% posterior containments respectively, evaluated point-wise in flux F. The middle column shows the posteriors for the Poissonian templates. The right column shows the joins posterior on the flux fractions of DM-like and PS-like emission. The dotted lines (in the left two columns) and the stars (in the right column) correspond to the true simulated quantities. DM-like emission is successfully inferred in each case, with the other parameter posteriors corresponding faithfully to the true simulated values.

emission from resolved 3FGL PSs as the posterior in that case is largely unconstrained owing to the fact that resolved PSs are masked out in the analysis. The right column shows the joint posterior on the fraction of DM- and PS-like emission in proportion to the total inferred flux in the ROI. The true underlying parameter values from which the data was generated are represented by dotted lines in the left and middle columns, and by star markers in the right column. We see that, in all cases shown, the pipeline successfully recovers the presence of DM-like emission, with little flux—$\lesssim 10\%$ of the total inferred GCE emission in all cases—attributed to PSs.

Figure 3 shows the corresponding results for simulated data containing PS-like emission correlated with the GCE. We see that PS-like emission is successfully inferred in each case, while at the same time exemplifying some degeneracy with the Poissonian component. Furthermore, as seen in the left column, the method is able to characterize the contribution of the two modeled PS components through the inferred source-count distribution. The inferred posteriors for the contribution of the DM-like component are seen to be compatible with zero. The overall flux of all modeled components, both PS and diffuse, is seen to be consistent with the true values used

for the simulations in both sets of tests.

IV. RESULTS ON *FERMI* DATA

We finally apply our neural simulation-based inference pipeline to the real *Fermi* dataset. As a point of comparison, we also run the NPTF method described in Sec. II B on the data using the same spatial templates and prior assumptions as those used in the corresponding SBI analyses. A summary of the results for different analysis configurations, obtained by re-training the model using different assumptions about spatial templates or parameter priors, is shown in Table II, including the fraction of overall emission attributed to the GCE, fraction of the GCE attributed to PS-like emission, flux corresponding to the highest break in the GCE broken power-law source count, fraction of the overall emission attributed to disk-correlated PSs, and flux corresponding to the highest break in the disk-correlated broken power-law source count. Medians as well as middle-68% ranges on the respective posteriors are presented. In these analyses, we draw a larger number 50,000 of samples from the trained flow in order to reduce sample variance when quoting summary quantities of the inferred posteriors.

A. Baseline analysis on *Fermi* data

Figure 4 shows posterior distributions for the baseline analysis on *Fermi* data, with the top panel showing results for the SBI analysis and bottom panel corresponding to the NPTF analysis. Consistent with previous studies using a similar configuration, a significant fraction of the GCE—$55.0^{+8.8}_{-22.9}$%—is attributed to PS-like emission within the NPTF framework. For the SBI analysis, although posteriors for the astrophysical background templates are seen to be broadly consistent with those inferred in the NPTF analysis, the preference for PSs is reduced, with $37.9^{+8.9}_{-19.2}$% of the GCE emission being PS-like. We also note that, in both cases, the inferred GCE-correlated source-count distribution sits at lower values than those inferred in previous NPTF analyses, which have generally found the bulk of PSs to lie just below the 3FGL PS detection threshold at $\sim$ 2–3 $\times$ $10^{-10}\,\mathrm{ph\,cm^{-2}\,s^{-1}}$ [30]. Given our doubly-broken power-law parameterization, the actual peak of the SCD is not physically meaningful—it can be driven simply by the position of the lowest break which marks a soft boundary between PS-like and smooth (but still possibly PS-driven) emission for accounting purposes. Instead, we use the upper break as a proxy for where the brightest unresolved sources are inferred to lie. In this baseline configuration, the SCD upper break is constrained to be $1.3^{+0.3}_{-0.4} \times 10^{-10}\,\mathrm{ph\,cm^{-2}\,s^{-1}}$, corresponding to $\sim$ 8—10 photons. This 'dimming' of the inferred SCD compared to previous NPTF analyses was also observes in Ref. [42], where it was found to be largely driven by the use of more up-to-date diffuse models. We note that even though the SBI analysis prefers a smaller GCE fraction in point sources, there is significant posterior overlap in the inferred joint PS and DM flux fraction posteriors between the SBI and NPTF analyses.

B. Signal injection test on *Fermi* data

A crucial test of self-consistency is the ability of the method to recover an artificial signal injected onto the real γ-ray data. As shown in Ref. [34], initial applications of the NPTF to the GCE would generally fail this closure test, with implications for characterizing the nature of PSs in the Galactic Center explored in Refs. [32, 33]. In particular, it was shown that this test can help diagnose underlying issues associated with mismodeling of the diffuse foreground emission, which have the potential to bias the characterization of PS populations. Recent NPTF analyses using improved descriptions of foreground modeling [32] show consistent behavior under thist closure test. We perform a version of this test within our framework, testing the ability of our method to recover different mock signals injected onto the real *Fermi* data.

Figure 5 shows the results of this test, with the different rows corresponding to different signal configurations—purely DM, bright PSs, medium-bright PSs, and dim PSs. Bright, medium-bright, and dim PS configurations are taken have a maximum PS flux (given by the highest break in Eq. (2)) at 20, 10, and 5 photon counts respectively, with other parameters set to median values inferred on real *Fermi* data, except the lower break, which was set to 2 photon counts. The leftmost columns show the baseline analysis on *Fermi* data, with subsequent columns showing signals of progressively larger sizes injected onto the data, up to approximately the size of the original GCE signal. The dotted horizontal and vertical lines show the total emissions including the injected signal and the median fluxes for the PS and DM components of the GCE inferred without any additional injected signal, respectively.

The additional injected signal is seen to be reconstructed correctly within the inferred 95% confidence interval in all four cases. For the DM signal (top row), the brighter inferred DM signals tend to be slightly overestimated. The injected PS signals (rows 2–4) are correctly reconstructed in all cases, with the dimmer PS signals showing a more prominent flat direction with Poissonian emission, as expected.

C. Systematic variations on the analysis

We test the robustness of our results by exploring several systematic variations on the baseline analysis, using alternative descriptions for the diffuse foreground emission template, the spatial distribution of disk-correlated sources, and prior configuration. Here, the

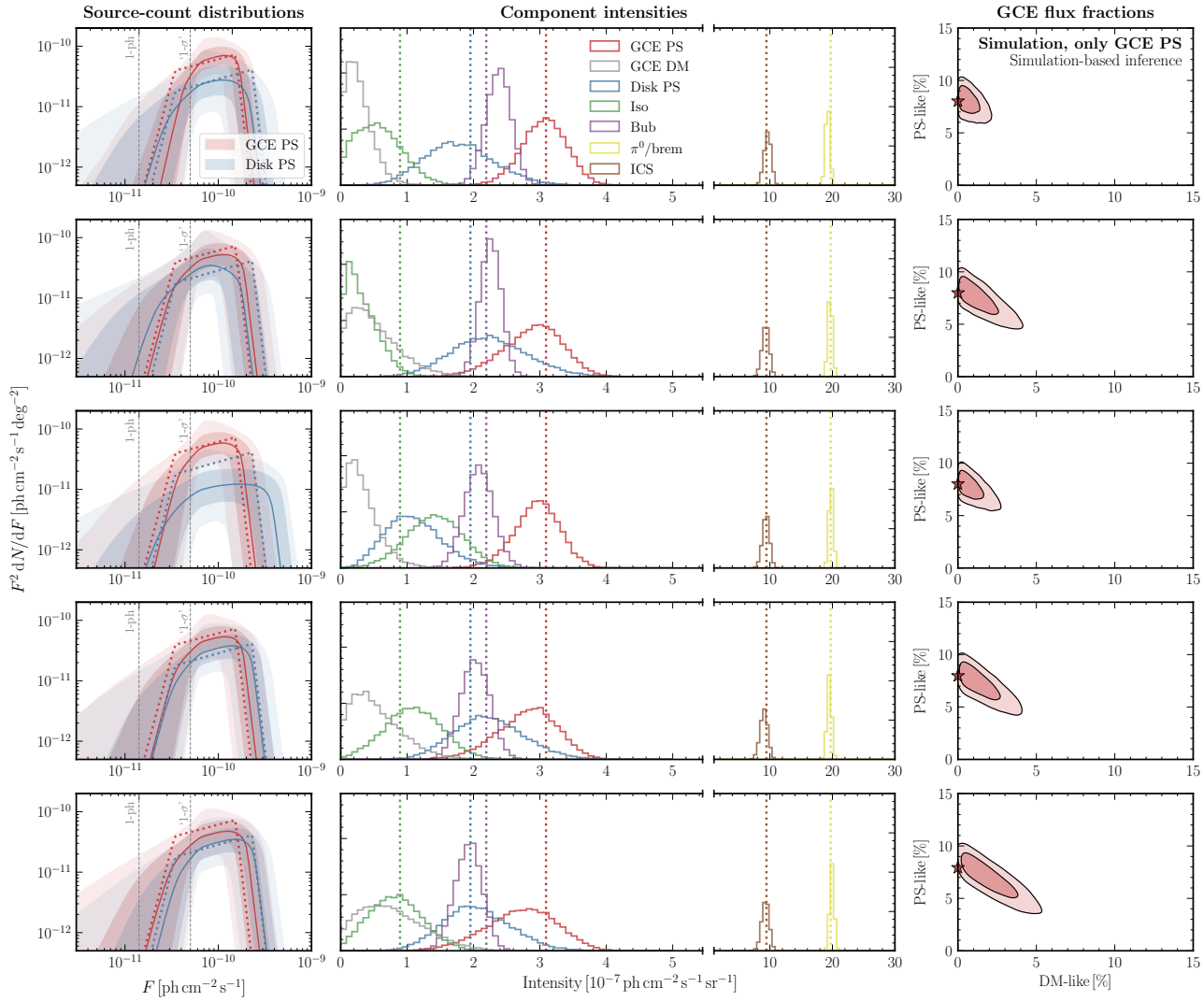

FIG. 3. Same as Fig. 2, but for five simulated realization of *Fermi* data where the GCE consists of predominantly PS-like emission. PS-like emission is inferred in each case, with the other posteriors corresponding faithfully to their true simulated quantities. The GCE-correlated source-count distribution is also seen to be successfully recovered in the left panel. We note that, as detailed towards the end of Sec. II A, PS flux below ~ 5 photons is partially accounted for by the smooth DM-like component, which is responsible for the sharp turn-off in the modeled as well as inferred GCE-correlated SCD with decreasing flux.

neural network is re-trained on a new set of simulations obtained using the alternative forward model before applying it to *Fermi* data. Results of these analysis variations are summarized in Tab. II.

Variation on the diffuse foreground model: In addition to diffuse Model O considered in the baseline analysis, we consider the alternative Models A and F from Ref. [11] to model the diffuse foreground emission, again including separate templates for gas-correlated emission and inverse Compton scattering. While shown to be a worse fit to the present dataset [32], these models have been previously used in the GCE literature [32, 69, 70] and provide a useful comparison point.

Results for these variations are shown in Figs. 6 and 7, respectively. In each case, results using the SBI pipeline are shown in the top row, with corresponding results using the NPTF pipeline in the bottom row. A somewhat larger fraction of the GCE, $47.2^{+10.5}_{-24.6}\%$, is attributed to PSs when using diffuse Model A (Fig. 6) compared to the baseline analysis using Model O. The corresponding NPTF analysis finds a still larger fraction of $74.9^{+6.6}_{-22.5}\%$. Using Model F, $62.5^{+10.1}_{-26.9}\%$ of the GCE is attributed to PSs, with qualitatively similar results found by the NPTF analysis. The total emission absorbed by the GCE in this case is about $\sim 60\%$ of that found in the baseline scenario. This is consistent with the results of Ref. [32], which found that the total GCE flux could vary by up to a factor of ~ 2 between analyses using different diffuse models.

Configuration	Method	GCE / Total	GCE PS / GCE	$F_{\rm b,1}^{\rm GCE}$	Disk PS / Total	$F_{\rm b,1}^{\rm Disk}$	Posteriors
-	-	[%]	[%]	[10^{-10} ph cm^{-2} s^{-1}]	[%]	[10^{-10} ph cm^{-2} s^{-1}]	-
Baseline	SBI	$7.8^{+0.2}_{-0.6}$	$37.9^{+8.9}_{-19.2}$	$1.3^{+0.3}_{-0.4}$	$5.0^{+0.5}_{-1.1}$	$2.2^{+0.2}_{-0.5}$	Figure 4
	NPTF	$7.7^{+0.2}_{-0.6}$	$55.0^{+8.8}_{-22.9}$	$1.1^{+0.1}_{-0.2}$	$5.4^{+0.5}_{-1.1}$	$2.0^{+0.2}_{-0.5}$	
Dif. Model A	SBI	$6.3^{+0.2}_{-0.6}$	$47.2^{+10.5}_{-24.6}$	$1.3^{+0.3}_{-0.4}$	$4.9^{+0.5}_{-1.2}$	$2.5^{+0.2}_{-0.5}$	Figure 6
	NPTF	$6.7^{+0.2}_{-0.6}$	$74.9^{+6.6}_{-22.5}$	$1.1^{+0.1}_{-0.2}$	$5.1^{+0.5}_{-1.3}$	$2.2^{+0.2}_{-0.5}$	
Dif. Model F	SBI	$4.7^{+0.2}_{-0.6}$	$62.5^{+10.1}_{-26.9}$	$1.5^{+0.3}_{-0.5}$	$5.2^{+0.4}_{-1.1}$	$2.5^{+0.2}_{-0.5}$	Figure 7
	NPTF	$5.2^{+0.2}_{-0.5}$	$67.5^{+8.6}_{-26.7}$	$1.1^{+0.2}_{-0.3}$	$6.4^{+0.5}_{-1.1}$	$2.0^{+0.2}_{-0.4}$	
Thick disk	SBI	$7.9^{+0.3}_{-0.6}$	$51.9^{+8.5}_{-22.2}$	$1.4^{+0.2}_{-0.4}$	$3.2^{+0.5}_{-1.2}$	$2.5^{+0.3}_{-0.6}$	Figure 8
	NPTF	$8.2^{+0.3}_{-0.7}$	$75.0^{+7.1}_{-22.6}$	$1.1^{+0.1}_{-0.2}$	$2.3^{+0.7}_{-1.1}$	$3.1^{+0.6}_{-1.2}$	
Alt. priors	SBI	$7.7^{+0.2}_{-0.6}$	$54.2^{+11.9}_{-27.4}$	$0.9^{+0.2}_{-0.4}$	$5.9^{+0.5}_{-1.1}$	$2.4^{+0.2}_{-0.4}$	Figure 12
	NPTF	$7.9^{+0.2}_{-0.6}$	$77.7^{+6.5}_{-21.4}$	$0.9^{+0.1}_{-0.3}$	$5.9^{+0.5}_{-1.1}$	$2.3^{+0.2}_{-0.4}$	

TABLE II. Inferred values for the inferred GCE flux as a fraction of the total flux, the GCE PS-like flux as a fraction of the total GCE flux, the position of the upper source count flux break $F_{\rm b,1}$ for the GCE and disk PS components, and the disk flux as a fraction of the total flux. For the baseline configuration as well as the various systematic variations explored, the median along with the 16th and 84th posterior percentile values are shown for the simulation-based inference (SBI) and NPTF analyses.

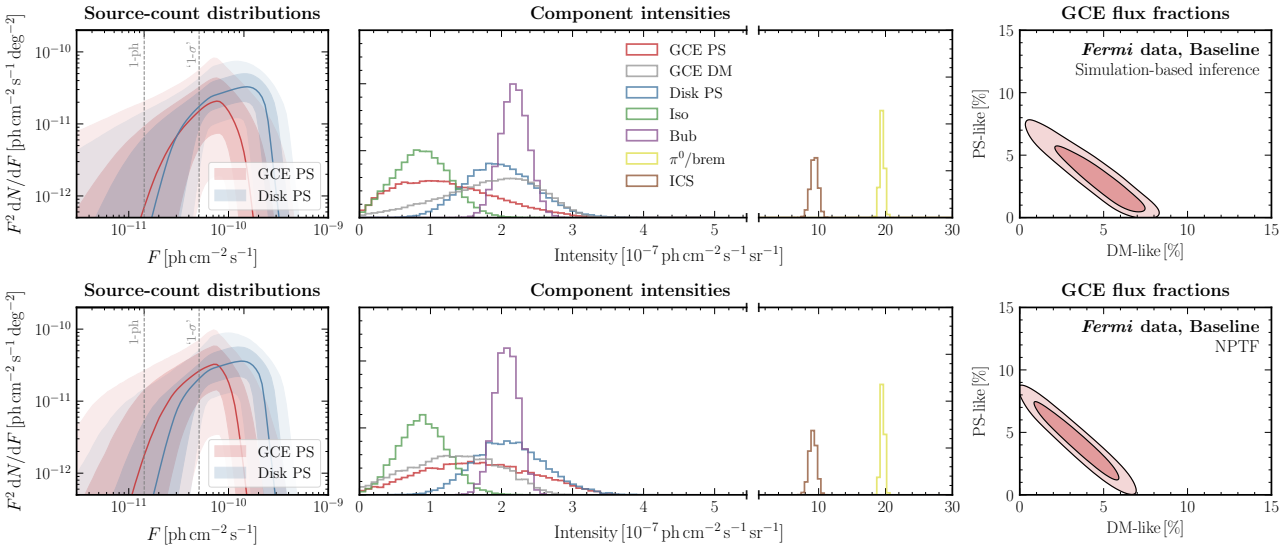

FIG. 4. Results of the baseline analysis on real *Fermi* data. (Top row) Analysis using neural simulation-based inference with normalizing flows, and (bottom row) using the 1-point PDF likelihood implemented in the non-Poissonian template fitting (NPTF) framework. While moderate preference for a PS-like origin of the GCE is seen in the case of the NPTF analysis (bottom), the simulation-based inference analysis attributes a smaller fraction of the GCE to PS-like emission (top).

Variation on the disk template: The baseline scenario considered a disk-correlated PS population with a spatial distribution given by Eq. (4), setting the scale height scale height $z_{\rm s} = 0.3\,$kpc corresponding to the 'thin-disk' scenario. Given uncertainties in the spatial distribution of the point source population (in particular, that of millisecond pulsars) associated with the Galactic disk, a 'thick-disk' spatial distribution has been employed in the literature as an alternative model [30, 32, 34], where the scale height is typically set to $z_{\rm s} = 1\,$kpc.

Results using a thick-disk template for the disk-correlated PS population are shown in Fig. 8. For the SBI analysis, a larger fraction $51.9^{+8.5}_{-22.2}$% of the GCE flux is attributed to a PS population in this case compared to the baseline scenario, with the GCE flux

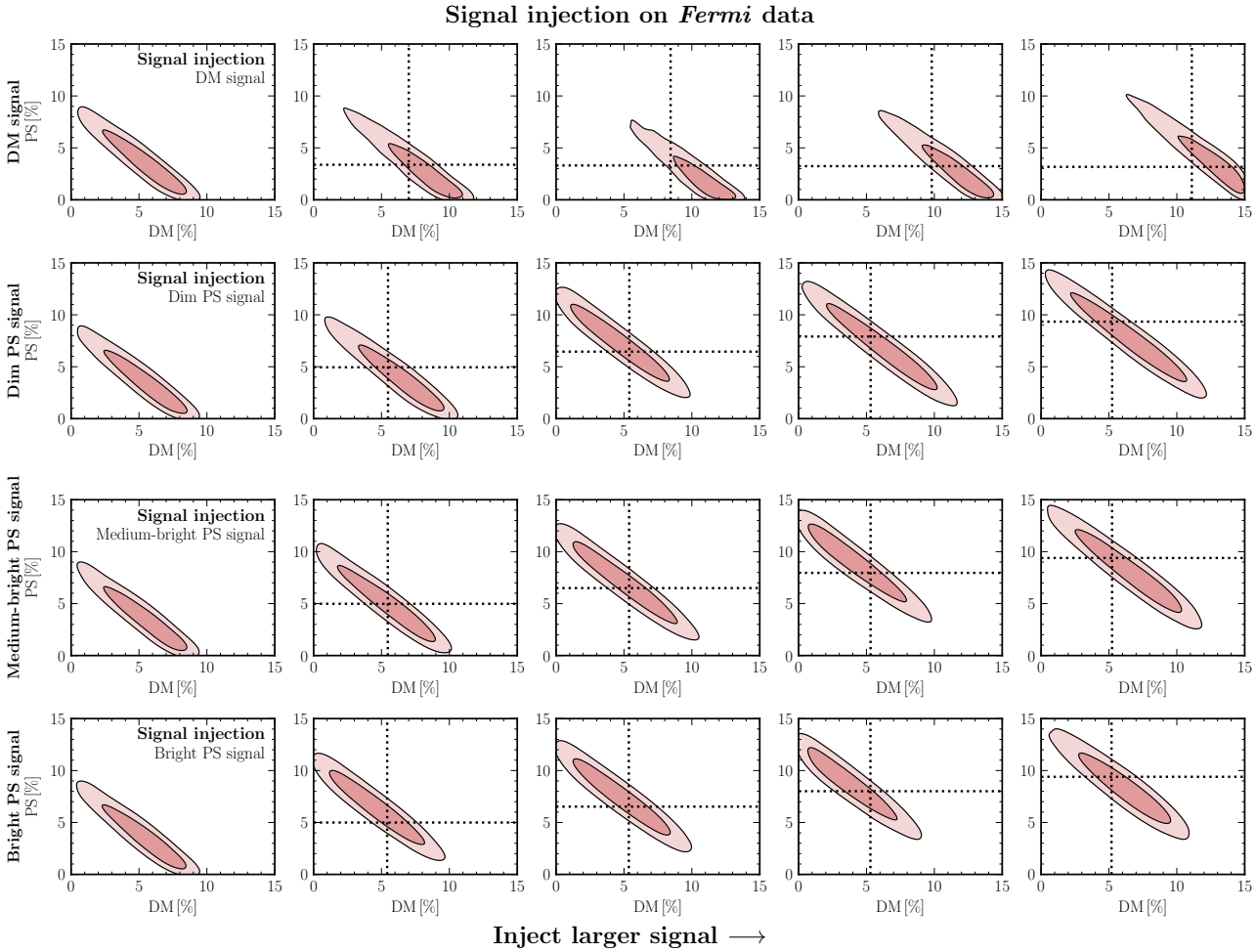

FIG. 5. Joint posterior for the flux fraction of PS-like and DM-like emission when an artificial DM signal is injected onto the real *Fermi* data. The different rows correspond to different signal types, from top to bottom, purely DM, dim PSs (maximum of 5 expected counts per PS), moderately-bright PSs (maximum of 10 expected counts per PS), and bright PSs (maximum of 20 expected counts per PS). The leftmost panels shows the baseline analysis on *Fermi* data, with subsequent panels showing results with progressively larger signals injected onto the data. The dotted lines show the expected total emissions including the injected signal and the median fluxes. The additional injected DM and PS signals are seen to correctly reconstructed within the respective posterior bounds in all cases.

itself being slightly larger. Once again, the NPTF analysis estimates a higher relative fraction $75.0^{+7.1}_{-22.6}\%$ of the GCE in point sources. The emission attributed to disk-correlated PSs is reduced in this case compared to the baseline scenario, possibly indicating a redistribution of PS-like emission between the GCE- and disk-correlated components.

Alternative prior specification: In the baseline analysis, we have chosen to enforce a soft distinction between relatively-bright PSs emitting $\gtrsim 5$ photons in expectation, and a combination of dimmer PSs and smooth emission following Poisson statistics taken together. This is done by placing a prior on the source-count slope below this chosen counts threshold that encourages a steeply-falling distribution with decreasing PS flux, allowing for a conservative interpretation of our results as a lower bound on the amount of PS emission. We also explore an alternative configuration where the lower break on the SCD is allowed to go down to a single photon, giving the PS component more overlap with the dim emission, and thus accounting for more emission in the PS-like component. The results of this analysis on data are summarized in the last row of Tab. II.

Reassuringly, the total flux attributed to the GCE is consistent between the alternative and baseline prior choices. As expected, allowing the lower SCD break to go down to smaller expected counts increases the fraction of the GCE flux attributable to PS-like emission, for the SBI case increasing the median fraction by $\sim 16\%$ relative to the baseline case. The NPTF analysis sees a larger increase in PS flux, with the median increasing by $\sim 23\%$. The fact that the NPTF analysis is relatively more sensitive to the details of modeling close to the single-photon limit is not surprising—since mismodeling is most likely to affect this dimmer regime in the source-

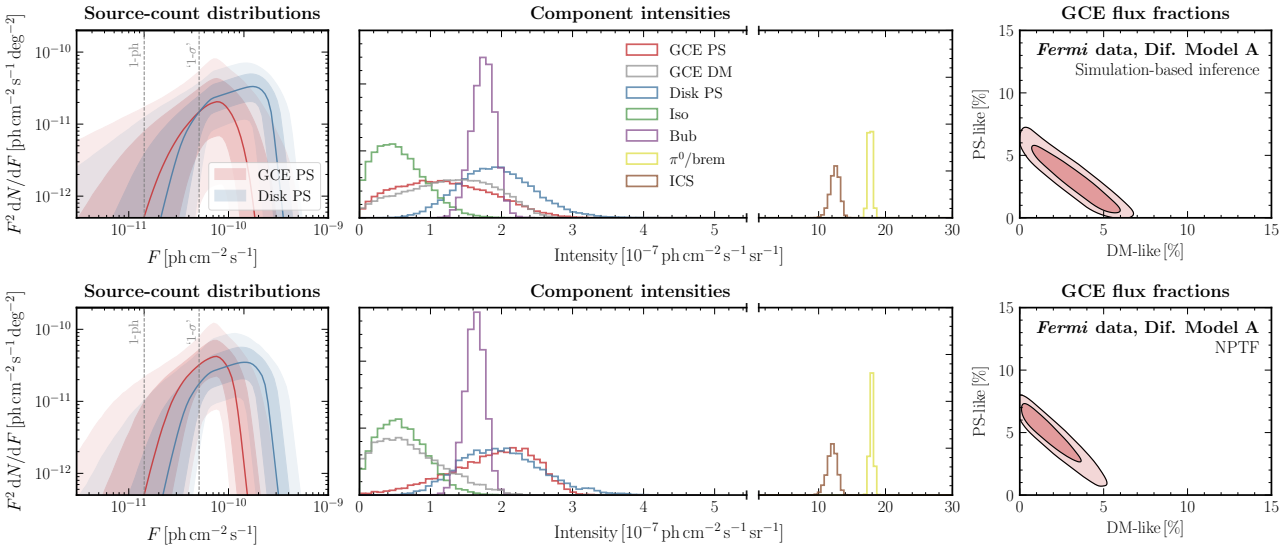

FIG. 6. Same as Fig. 4, but with the diffuse foreground emission modeled using the alternative Model A.

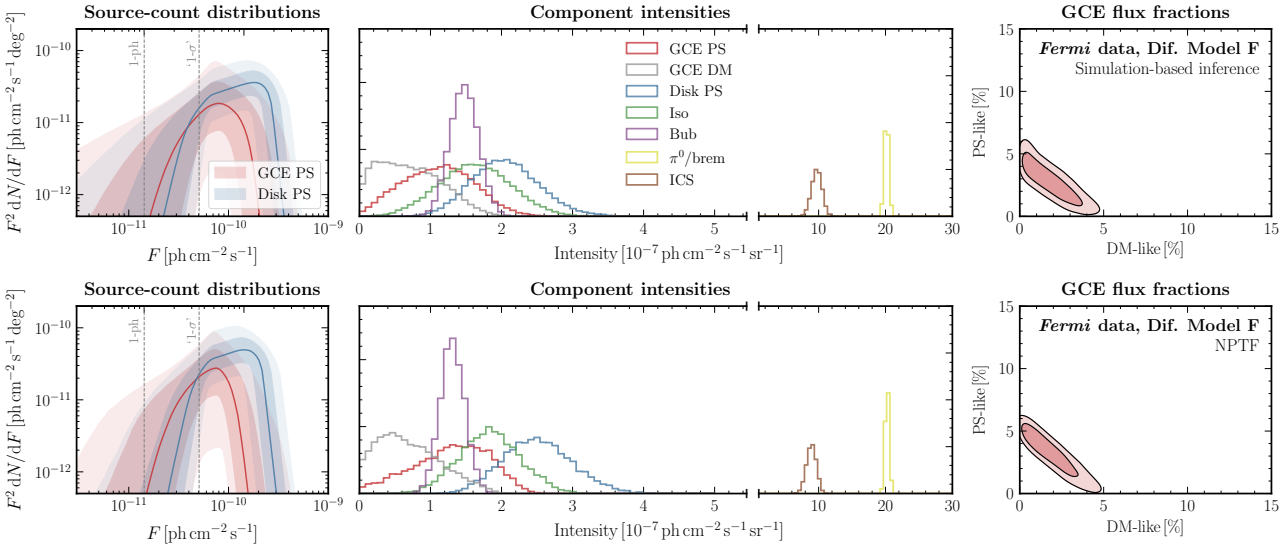

FIG. 7. Same as Fig. 4, but with the diffuse foreground emission modeled using the alternative Model F.

count distribution, this is also where the two methods can be expected to diverge more significantly. The position of the upper flux break, quantifying the inferred fluxes of the brightest sources in the PS population, is slightly reduced to $0.9^{+0.2}_{-0.4}\times10^{-10}$ ph cm^{-2} s^{-1}, corresponding to 5–7 photons. The posterior distributions for these cases, as well as the prior distribution on the source-count distribution corresponding to the two prior choices, are shown in App. A. There, we also check that the inferred emission below 5 photons is consistently redistributed between the PS-like and DM-like components when using the two different prior configurations.

V. SUSCEPTIBILITY TO MODEL MISSPECIFICATION

Given the complex astrophysical environment in the Galactic Center, a key challenge in γ-ray analyses of the GCE is that associated with effects of mismodeled signal and background templates. As explored in detail in Refs. [30, 32, 33, 69, 70] within the NPTF framework, mismodeling can hamper the characterization of an Inner Galaxy PS population and, if sufficiently severe, can result in the attribution of mismodeled residuals to a spurious PS population when the underlying emission is actually smooth in nature.

In this section we assess the susceptibility of our simulation-based inference pipeline to several known sources of mismodeling. We do so by creating mock data

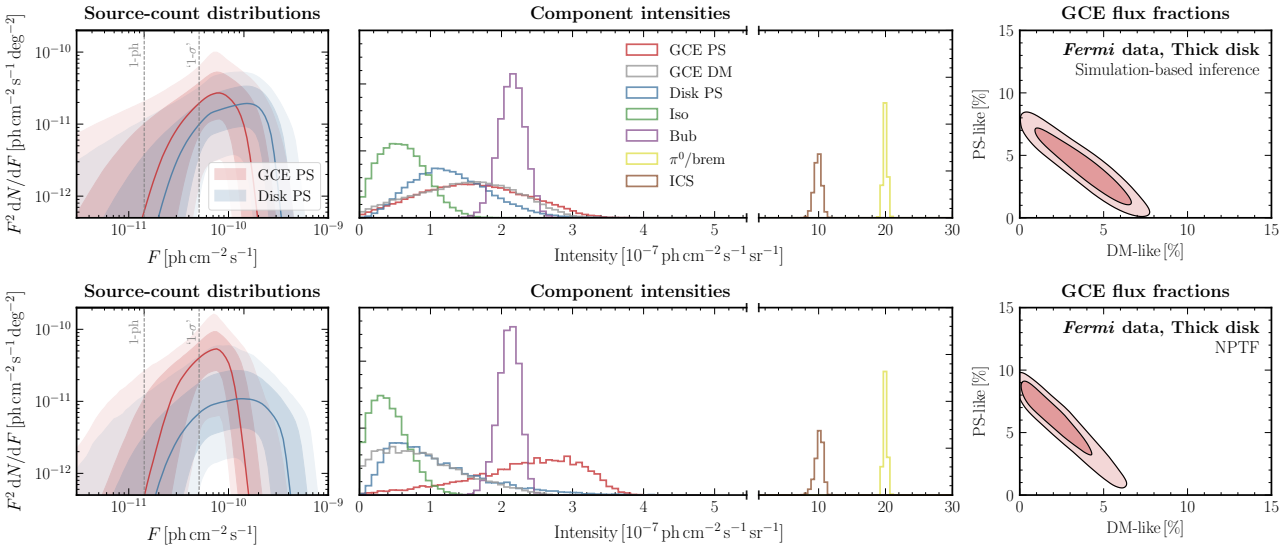

FIG. 8. Same as Fig. 4, but with the spatial distribution of disk-correlated PSs modeled using a thick-disk template (scale factor $z_s = 1\,\mathrm{kpc}$ in Eq. (4)) rather than the default thin-disk template ($z_s = 0.3\,\mathrm{kpc}$).

with a smooth GCE signal and a background model that was perturbed compared to the model that was used to train the SBI pipeline, and analyzing it with our baseline neural network *i.e.*, the one trained on the forward model described in Sec. II A and used in the baseline analysis on data in Sec. IV. The ability of our method to correctly characterize the injected signal is then indicative of the level of robustness that can be expected in the real data under corresponding circumstances. Results for the various tests performed are shown in Fig. 9, and will be described below. In each case, we show posteriors obtained by combining 50,000 samples from analyses of 10 different mock datasets (thinned by a corresponding factor of 10) in order to characterize the 'average' mismodeling associated with a given configuration. The first row of Fig. 9 shows the aggregate analysis without mismodeling *i.e.*, conditioned on mock data created with the same forward model as that used for training the neural posterior estimator, as a point of comparison.

Test of diffuse mismodeling using an alternative diffuse emission template: We create mock data using diffuse Model A, and analyze it using our baseline analysis pipeline trained with Model O. The aggregated results over 10 different maps are shown in the second row of Fig. 9. We see that even though some of the other diffuse component posteriors are shifted relative to their true values, the DM-like emission is faithfully recovered, and no additional PS-like flux is inferred.

A data-driven test of large-scale mismodeling: We construct a data-driven model of foreground mismodeling on large spatial scales (specifically, well above the scale of the instrumental PSF) and assess the ability of our method to recover a smooth DM-like signal in this case. Following Ref. [93], we perform a Poissonian template analysis on the *Fermi* dataset x, modulating the diffuse model template $T_{\rm dif}$, which describes the bremsstrahlung and neutral pion decay components of diffuse Model O, by an (exponentiated) Gaussian process (GP) f:

$$x \sim \mathrm{Pois}\left(\sum_{i\neq \mathrm{dif}} A_i T_i + \exp\left(f\right) A_{\rm dif} T_{\rm dif}\right). \qquad (15)$$

The other Poissonian templates T_i, including a GCE DM template and the inverse Compton component of the diffuse foreground model, are treated as before using an overall normalization factor A_i. $f \sim \mathcal{N}(m, K)$ is the GP component with prior mean m set to zero, and the covariance K described using the Matérn kernel with smoothness parameter $\nu = 5/2$. We refer to Ref. [93] for further details of the analysis, as well a validation of the GP-augmented template fitting pipeline on simulated data.

Five random samples from the Gaussian process describing multiplicative mismodeling relative to the real *Fermi* data when using our baseline diffuse Model O are shown in Fig. 10. The largest mismodeling by magnitude in this case is inferred to be concentrated in the southern regions of the baseline ROI. We note that, when analyzing the real *Fermi* data, the recovered GCE flux tends to be lower by up to 40% when using the GP-modulated diffuse model compared to that obtain in a Poissonian fit without the GP, with the missing emission absorbed by the GP-modulated template. This is indicative of the fact that a component of the centrally concentrated emission could be better described by the modulated template rather than the generalized NFW template modeling DM annihilation. We leave a detailed study of implications of this fact for the morphology of the excess to future work. When creating simulated data containing DM-like emission in association with this modulated template, the

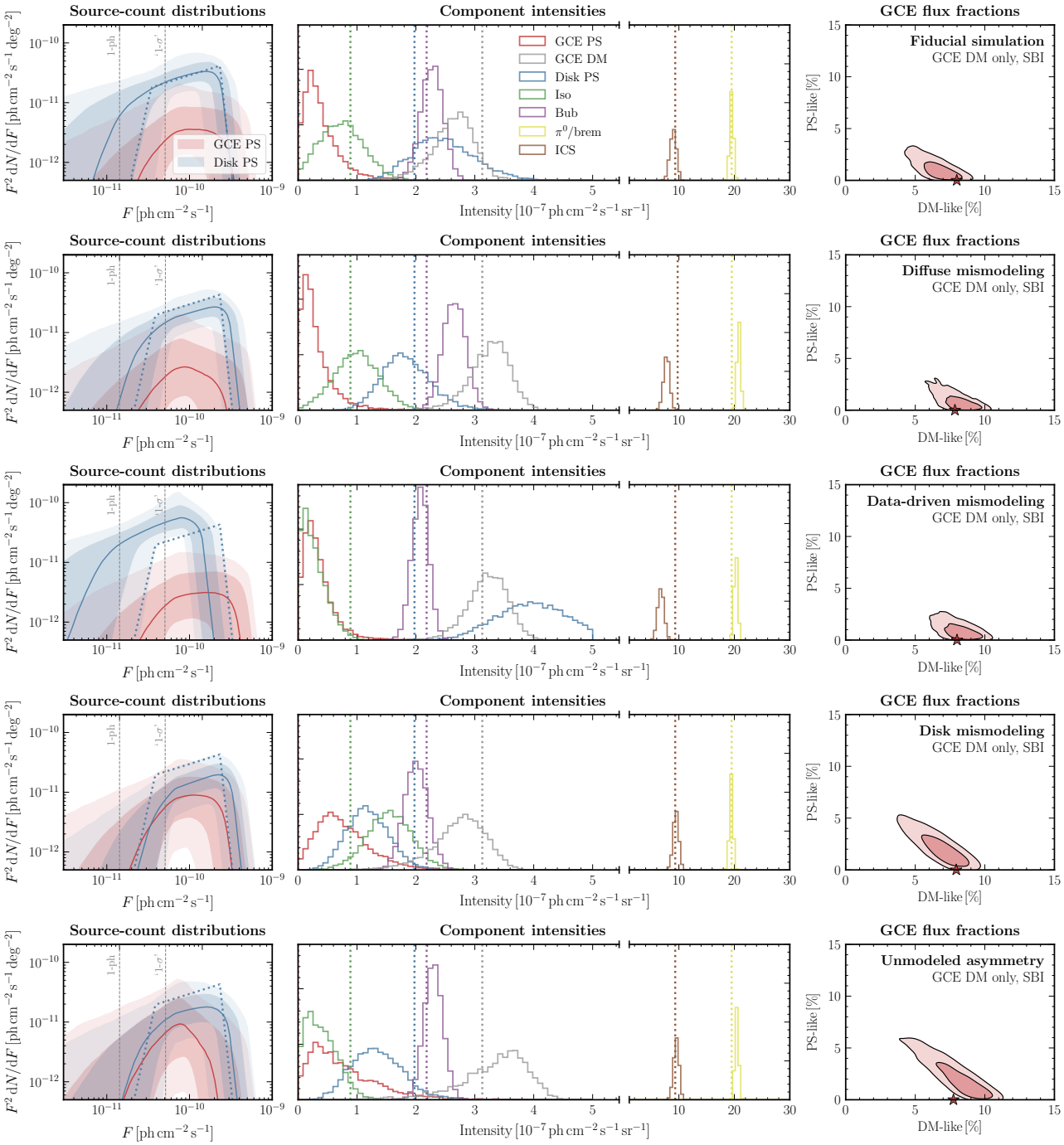

FIG. 9. Effect of mismodeling on a smooth GCE within our analysis framework. Each row shows aggregate posteriors collected over 10 simulated samples; row-wise from top to bottom: *(i)* No mismodeling; simulated data is constructed with the same templates as those used in the forward model for training. *(ii)* Mock data created with diffuse Model A, showing a possible effect of diffuse mismodeling. *(iii)* Mock data where the diffuse template, described by Model O, is modulated by draws from a Gaussian process modeling large-scale mismodeling inferred from the real *Fermi* data. *(iv)* Mock data where the thick-disk template is used in lieu of the thin-disk template. *(v)* Mock data where the GCE signal in the Northern hemisphere is twice as large as that in the Southern hemisphere. While some PS-like emission is inferred, it is consistent with zero in all cases, and evidence for a smooth GCE is robust.

fraction of DM-like flux in the simulation was correspondingly reduced by 40%.

In order to test the effect of such mismodeling on recovery of a DM signal we modulate the bremsstrahlung and neutral pion decay-tracing components of Model O using samples drawn from the inferred Gaussian process. These simulated samples are used as mock data that are then analyzed with our baseline model, where the unmodulated Model O was used to create training samples. The results of this test are shown in the third row of Fig. 9. It can be seen that while large-scale mismodeling can distort the total flux attributed to individual modeled components, in particular causing the disk-correlated PS emission to be significantly overestimated, preference for a smooth origin of the signal remains robust.

Effect of mismodeling the disk spatial template: We replace the thin-disk template, described by a scale height $z_s = 0.3\,\mathrm{kpc}$ in Eq. (4), with a thick-disk template with $z_s = 1\,\mathrm{kpc}$ in the simulated data. Results of then analyzing 10 mock maps using the thin-disk template used in the baseline configuration are shown in the fourth row of Fig. 9. We see that the disk-correlated PS emission is underestimated, and a small amount of GCE PS-like emission is inferred while the marginal DM-like posterior is not significantly affected. This could be indicative of a reshuffling of the emission between disk- and GCE-correlated components.

Effect of an unmodeled asymmetry in the signal: Besides mismodeling associated with astrophysical background templates, another concern is that associated with mismodeling of the signal emission itself. In particular, as pointed out in Refs. [69, 70], a North-South asymmetry in a putative dark matter signal, if unaccounted for, could lead to spurious inference of a PS population associated with the purely smooth, asymmetric signal in the NPTF framework. Refs. [69, 70] found preference for such a scenario in real *Fermi* data, with the GCE signal in the Northern hemisphere a factor of ~ 2 larger than that in the Southern hemisphere when the GCE template in the two regions is floated separately in a ROI defined by $r < 10^\circ$. In this case, for certain diffuse models, no preference for a PS-like GCE was found in contrast to the case when a single template was used to model the GCE.

We test the impact of a North-South-asymmetric dark matter signal within our framework by running our baseline pipeline on simulated datasets where the dark matter-like signal in the Northern hemisphere of the ROI is 2 times larger than that in the Southern hemisphere, mimicking the preference in real data found in Refs. [69, 70]. The result of this test over 10 such simulated realizations is shown in the last row of Fig. 9. We see that even with the presence of a substantially asymmetric DM-like signal we retain preference for a predominantly smooth GCE. While some additional PS-like emission is inferred, the effect is small compared to that exhibited within the NPTF framework in analogous tests [69, 70]. We attribute this to the fact that the `DeepSphere`-based convolutional neural network feature extractor can account for pixel-to-pixel correlations in the γ-ray counts map, and can thus be sensitive to *local* PS-like structures. In contrast, the 1-point PDF-based NPTF framework, being agnostic to the ordering of the pixels, can notice spurious PS-like structures in the distribution of 'residuals' associated with an asymmetric signal when analyzed with a symmetric template. As done in Ref. [32], we emphasize that the presence of a substantial asymmetry in the GCE signal, if not attributed to diffuse mismodeling, would point towards astrophysical explanations of the GCE since a true dark matter signal would not be expected to be significantly asymmetric.

In all cases tested, while posteriors for certain templates can show systematic biases, preference for a smooth origin of the GCE remains robust and the fraction of inferred PS-like emission is compatible with zero. Finally, it is also interesting to similarly consider the effect of mismodeling on a PS-like GCE signal. We perform a subset of the tests described above on simulated GCE PS signals in App. B, showing successful recovery of an overwhelmingly PS-like GCE in the face of mismodeling.

VI. DISCUSSION AND CONCLUSIONS

In this paper, we have leveraged recent advances in neural simulation-based inference in order to jointly characterize a putative DM-like signal and PS-like population associated with the observed *Fermi* Galactic Center Excess. Consistent with Ref. [41] which used a Bayesian neural network and first leveraged a `DeepSphere`-based feature extraction architecture for analyzing γ-ray data in the Galactic Center region, our analysis based on conditional posterior density estimation with normalizing flows finds a reduced contribution associated with a potential population of unresolved PSs to the GCE compared to previous analyses based on the photon statistics of the γ-ray map. In particular, depending on the analysis configuration, we find a median value of ~ 40–60% as the fraction of GCE emission that can be attributed to a PS population, with the brightest unresolved sources inferred to be at somewhat smaller fluxes $\sim 10^{-10}\,\mathrm{ph\,cm^{-2}\,s^{-1}}$ compared to values found in previous analyses based on the non-Poissonian template fitting (NPTF) framework [30]. The NPTF analyses performed in this work find a similarly dimmer source-count distribution, in all cases however attributing a larger fraction ~ 50–80% of the GCE to a PS population as compared to the corresponding SBI analyses. Even though the SBI analyses presented here generically attributed a smaller fraction of the GCE flux to PSs, we note that there is significant overlap within posterior uncertainties between results returned by the two methods, as can be seen from

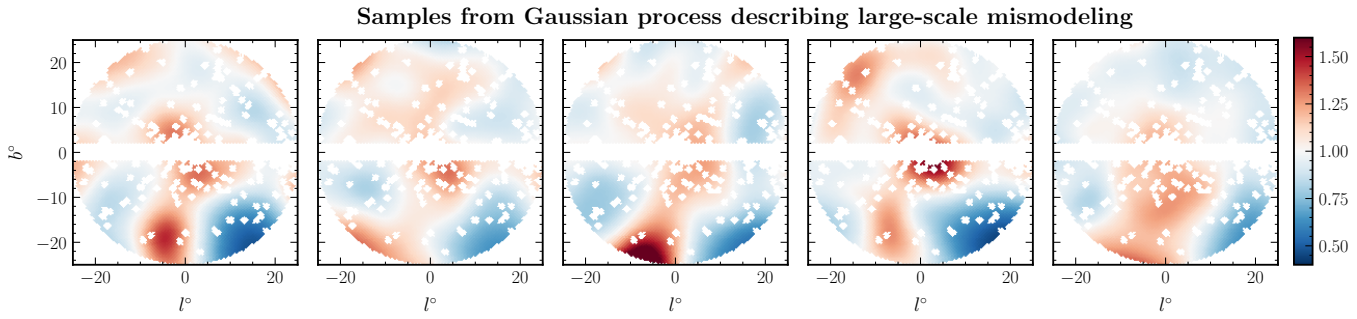

FIG. 10. Five random samples from the Gaussian process description of large-scale multiplicative mismodeling associated with the gas-correlated component of diffuse foreground Model O when applied to the real *Fermi* data.

Tab. II.

The results of this paper are broadly consistent with and complementary to those obtained in Ref. [42], which used a `DeepSphere`-based architecture which was, in contrast to our parametric approach, combined with a novel neural network-based non-parameteric approach to infer the counts distributions associated to PS populations using histograms with modeled uncertainties [44]. Their approach does not explicitly distinguish between Poissonian and PS-like components, treating emission associated with the inferred counts PDF below some threshold as effectively Poissonian. While this makes a direct comparison to the results of their analysis challenging, the overall conclusions regarding the fraction of emission that can be attributed to PSs and the characteristics of the GCE and disk source-count distributions are qualitatively similar between the two studies. In particular, both papers find a dimmer GCE-correlated source-count distribution, with a smaller lower bound $\gtrsim \mathcal{O}(40\%)$ on the fraction of the total GCE emission associated to PSs compared to previous studies based on the NPTF.

Our qualitative conclusions are robust to the systematic variations we have explored, including different models for the diffuse foreground and spatial distribution of disk-correlated PS emission. We used a novel Gaussian process-based method to construct a data-driven model of large-scale spatial mismodeling, finding our method to be resilient to such effects when it comes to inferring the presence of DM-like emission. As in any Galactic Center γ-ray analysis, we caution of the potential of unknown systematics, such as mismodeling on the scale of the size of the *Fermi*-LAT point-spread function, to bias the results and conclusions of our analysis. Although machine learning-based analyses can utilize more of the information encoded in the forward model, and in particular in the present case can take advantage of pixel-to-pixel correlations, this can also make them more susceptible to specific modeled features compared to traditional techniques based on data reduction to hand-crafted data summaries. We leave a more detailed investigation of the potential impact of these effects on our analysis to future work.

Several improvements to the framework presented here are possible. Although we have used a dataset restricted to the top quartile of photons by quality of PSF reconstruction, as shown in Ref. [69] the use of a larger data sample can provide improved sensitivity to a PS population while acting as a consistency wo the with results obtained on the smaller sample. The inclusion of energy-binning information in the analysis can be implemented in a straightforward manner by splitting up the data and template maps into individual energy bins and feeding these as separate channels in the graph-convolutional feature extraction network. The use of more complex feature extraction architectures can additionally improve the robustness of our results. While we have considered a simulated-based inference framework based on posterior density estimation with normalizing flows, alternative frameworks based on likelihood-ratio estimation [94–100] or flow-based likelihood estimation [101, 102] can provide complementary ways to characterize the γ-ray PS population in the Galactic Center. Additionally, the use of sequential active-learning methods [102] and methods that make use of additional latent information from the simulator [94–96, 103, 104] can significantly improve the simulation sample efficiency and allow for extensions to more complex forward models, which can be important in particular for an energy-binned analysis and if including additional degrees of freedom for the astrophysical background models.

Since diffuse mismodeling is the largest source of uncertainty in any analysis that aims to characterize the GCE, we also note the possibility of using adversarial learning methods [105] or distance correlations [106] to account for systematic differences between the modeled and real *Fermi* data. Alternatively, generative modeling of the diffuse foreground either in a Gaussian process-based data-driven framework or using, *e.g.*, autoencoders trained on an ensemble of plausible diffuse models, can provide a principled way to account for the large latent space associated with diffuse emission modeling. Motivated by quantitative variations in our results on *Fermi* data when using different disk templates, self-consistently accounting for plausible variations in the spatial distribution of disk-correlated PSs can strengthen the results of our analysis when it comes to characterizing

the PS population in the Galactic Center. These extensions can lead to a more robust characterization of an unresolved PS population in the Galactic Center region associated with the GCE, and we leave their study to future work.

The code used to obtain the results in this paper as well as a pre-trained neural network model associated with the baseline analysis presented here is available at https://github.com/smsharma/fermi-gce-flows.

ACKNOWLEDGMENTS

We thank Johann Brehmer and Tracy Slatyer for helpful conversations. We are grateful to Florian List and Nick Rodd for carefully reading an earlier version of this paper and for their many helpful comments. SM would like to thank the Center for Computational Astrophysics at the Flatiron Institute for their hospitality while this work was being performed. This work was performed in part at the Aspen Center for Physics, which is supported by National Science Foundation grant PHY-1607611. The participation of SM at the Aspen Center for Physics was supported by the Simons Foundation. SM is supported by the NSF CAREER grant PHY-1554858, NSF grants PHY-1620727 and PHY-1915409, and the Simons Foundation. KC is partially supported by NSF awards ACI-1450310, OAC-1836650, and OAC-1841471, the NSF grant PHY-1505463, and the Moore-Sloan Data Science Environment at NYU. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics of U.S. Department of Energy under grant Contract Number DE-SC0012567. We thank the *Fermi*-LAT Collaboration for making publicly available the γ-ray data used in this work. This work made use of the NYU IT High Performance Computing resources, services, and staff expertise. This research has made use of NASA's Astrophysics Data System. This research made use of the `astropy` [107, 108], `dynesty` [68], `getdist` [109], `IPython` [110], `Jupyter` [111], `matplotlib` [112], `MLflow` [113], `nflows` [114], `NPTFit` [49], `NPTFit-Sim` [59], `NumPy` [115], `pandas` [116], `PyGSP` [117], `Pyro` [118], `PyTorch` [119], `PyTorch Geometric` [120], `PyTorch Lightning` [121], `seaborn` [122], `sbi` [123], `scikit-learn` [124], `SciPy` [125], and `tqdm` [126] software packages. We acknowledge the use of data products and templates from the code repository associated with Ref. [41].[3] We acknowledge use of the `DeepSphere` graph-convolutional layer from the code repository associated with Ref. [87].[4]

APPENDIX

Appendix A: Prior-predictive distributions and results for alternative priors

Figure 11 shows the prior distribution induced on the source-count distribution for the baseline PS model with upper SCD break priors uniform in the interval $S_{\mathrm{b},1} \in [5, 40]$ photons (left) and the alternative prior specification with $S_{\mathrm{b},1} \in [1, 30]$ photons (right). The latter prescription gives the PS-like component more overlap with emission just above 1-photon, since the slope below the second break encourages the SCD to steeply drop. It can be seen that both prior choices still allow for significant PS-like emission below their respective counts soft thresholds.

Figure 12 shows posterior distributions for the analysis using the alternative prior set. These results are summarized in the bottom row of Tab. II. As expected, both the SBI (top row) and NPTF (bottom row) analyses show a larger inferred PS flux compared to the analysis using the baseline prior choice. Reassuringly, the total flux absorbed by both GCE components taken together remains consistent between the analyses with different prior choices.

Finally, Fig. 13 shows a check of how the partitioning of flux between PS-like and DM-like components varies between the two prior choices. The excess dark matter flux (shown as inferred counts per-pixel $\langle S \rangle$) in the baseline prior configuration (topmost data point) is seen to be consistent with the cumulative excess flux below 5 photons in the alternative prior configuration compared to the baseline one (second data point from the top). When this excess flux is added to the total PS flux in the baseline configuration (middle data point), the combination (second data point from the bottom) is additionally seen to be consistent with the total PS flux in the alternative prior configuration (bottommost data point). We note that this test is merely heuristic—in particular, since the posteriors for baseline and alternative prior analyses are described by independent samples, the component counts were combined or subtracted assuming uncorrelated errors (computed as standard deviations over the respective posteriors), which is certain to not be the case. However, this test is indicative of the fact that the inferred flux below the threshold we set for accounting purposes is redistributed between the PS and DM components, as would be expected if the two analyses were self-consistent.

[3] https://github.com/FloList/GCE_NN

[4] https://github.com/deepsphere/deepsphere-pytorch

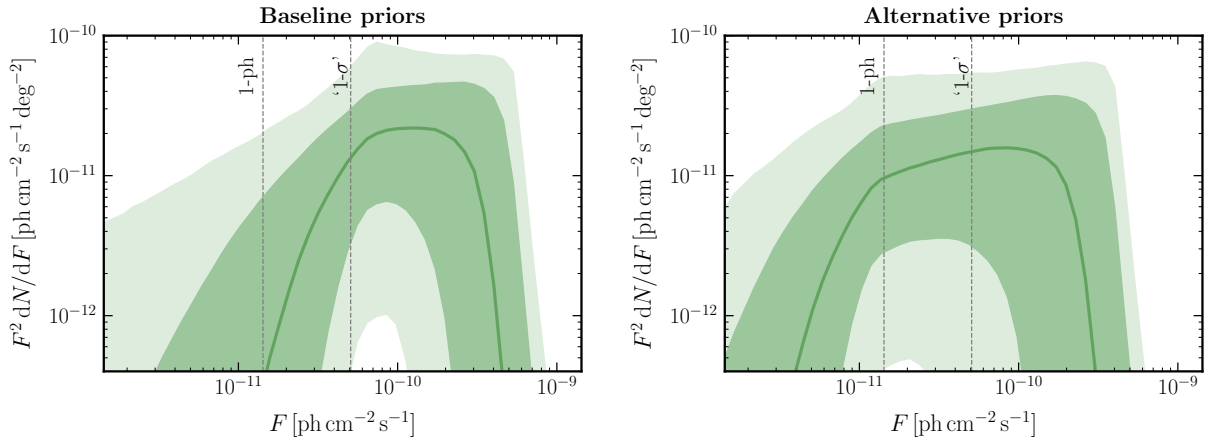

FIG. 11. Prior-predictive distribution on the source-count distribution for the baseline PS model priors (left) and alternative prior specification giving the PS component more overlap with emission close to the single-photon limit (right). The median (lines), middle-68% containment (darker bands), and middle-95% containment (lighter bands) regions are shown.

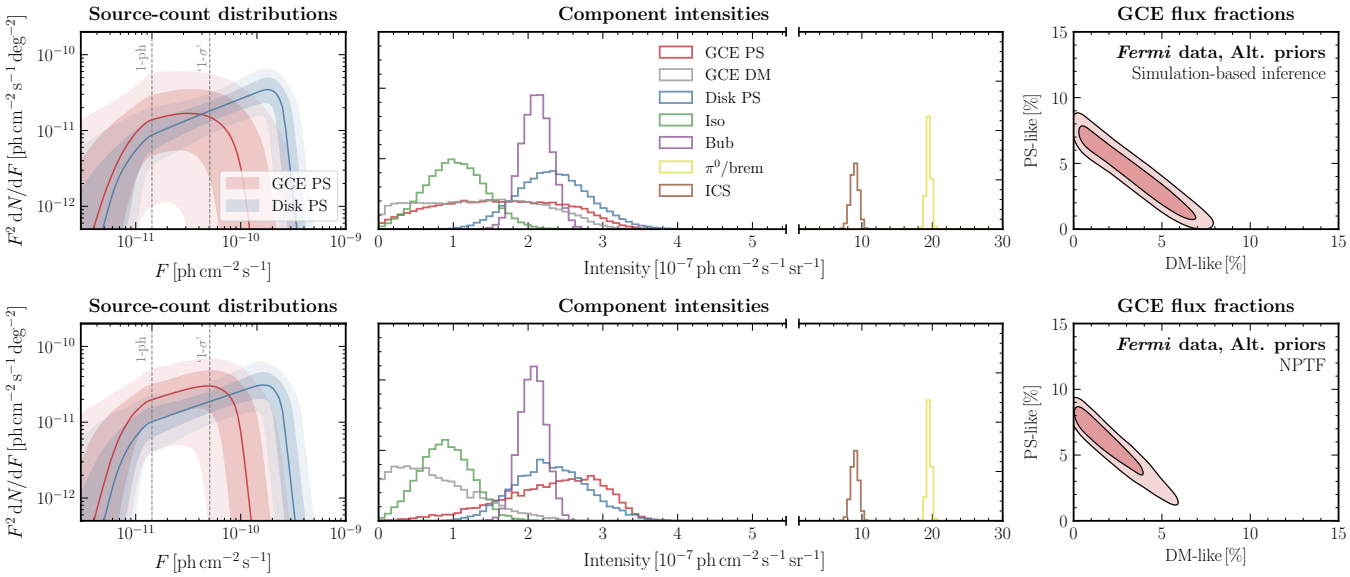

FIG. 12. Same as Fig. 4, but using the alternative prior specification for the PS model, with the break on the lower SCD break uniform within $S_{\mathrm{b},1} \in [1, 30]$ photons rather than $S_{\mathrm{b},1} \in [5, 40]$ photons. This configuration gives the PS model more overlap with the dim PS-like emission.

Appendix B: Mismodeling effects on a simulated GCE PS signal

Figure 14 shows the analog of Fig. 9 where we test the effect of mismodeling on a PS-like rather than smooth GCE. As in the test on simulated maps with a smooth GCE described in Sec. V, the data containing GCE-correlated PSs is created using a forward model that is different in a specific way from that used do train the neural network model: *(i)* No mismodeling; simulated data is constructed with the same templates as those in the forward model used for training the posterior estimator (top row). *(ii)* Mock data created with diffuse Model A, showing the effect of diffuse mismodeling (middle row). *(iii)* Mock data where the thick-disk template is used in lieu of the thin-disk template (bottom row).

In each case, the aggregate posterior described by 50,000 samples obtained over 10 simulations and then thinned by a factor of 10 is shown. Since the GP-modulated (smooth) diffuse template tends to absorb a substantial fraction of the GCE flux when applied to real *Fermi* data, a test using this modulated template was not performed here as it would not yield self-consistent results. It can be seen from the rightmost column of Fig. 14 that a substantially PS-like GCE is recovered in both cases tested, although a small fraction of flux is attributed to DM-like emission when mismodeling of the diffuse emission template is considered.

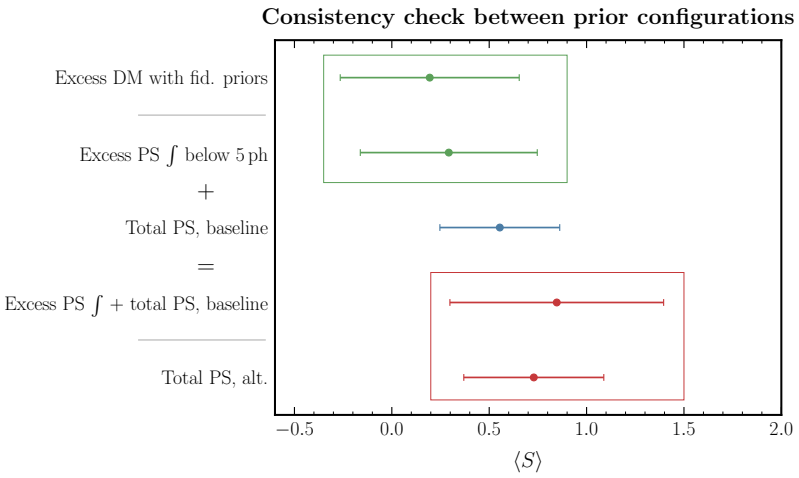

FIG. 13. A heuristic check of the distribution of PS-like flux below 5 photons in the analyses with baseline and alternative priors. The excess dark matter flux (shown as counts per pixel $\langle S \rangle$) in the baseline prior configuration (topmost data point) is seen to be consistent with the cumulative excess flux below 5 photons in the alternative prior configuration compared to the baseline one (second data point from the top). When this excess flux is added to the total PS flux in the baseline configuration (middle data point), the combination (second data point from the bottom) is additionally seen to be consistent with the total PS flux in the alternative prior configuration (bottommost data point).

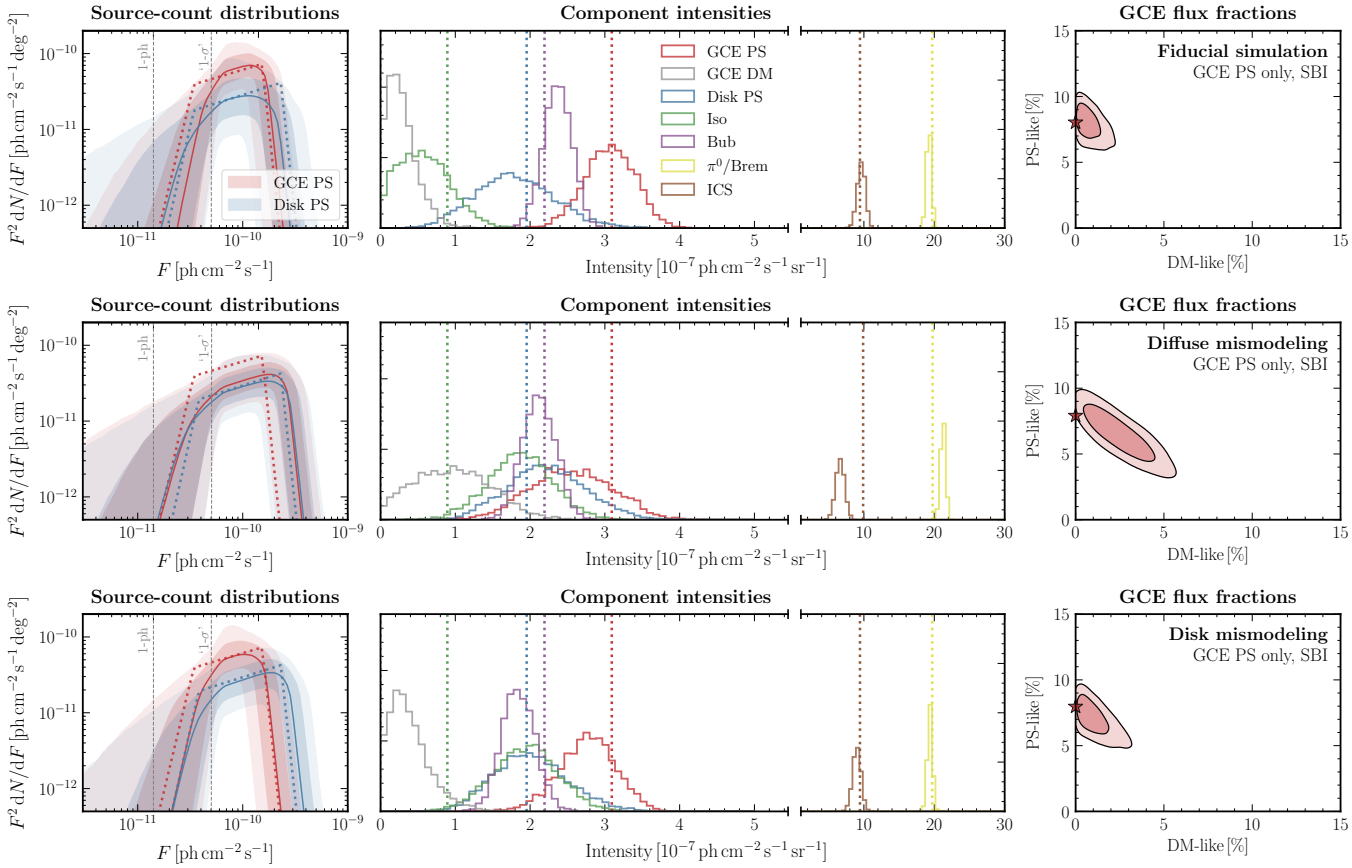

FIG. 14. Effect of mismodeling on a PS-like GCE within our analysis framework. Each row shows aggregate posteriors collected over 10 simulated samples; row-wise from top to bottom: *(i)* No mismodeling; simulated data is constructed with the same templates as those used in the forward model. *(ii)* Mock data created with diffuse foreground Model A, showing a possible effect of diffuse mismodeling. *(iii)* Mock data where the thick-disk template is used in lieu of the thin-disk template. A substantially PS-like GCE is inferred, although a subdominant fraction of DM-like flux is inferred as well when considering diffuse foreground mismodeling.

[1] W. B. Atwood *et al.* (Fermi-LAT), Astrophys. J. **697**, 1071 (2009), arXiv:0902.1089 [astro-ph.IM].
[2] L. Goodenough and D. Hooper, (2009), arXiv:0910.2998 [hep-ph].
[3] D. Hooper and L. Goodenough, Phys. Lett. B **697**, 412 (2011), arXiv:1010.2752 [hep-ph].
[4] A. Boyarsky, D. Malyshev, and O. Ruchayskiy, Phys. Lett. B **705**, 165 (2011), arXiv:1012.5839 [hep-ph].
[5] D. Hooper and T. Linden, Phys. Rev. D **84**, 123005 (2011), arXiv:1110.0006 [astro-ph.HE].
[6] K. N. Abazajian and M. Kaplinghat, Phys. Rev. D **86**, 083511 (2012), [Erratum: Phys.Rev.D 87, 129902 (2013)], arXiv:1207.6047 [astro-ph.HE].
[7] D. Hooper and T. R. Slatyer, Phys. Dark Univ. **2**, 118 (2013), arXiv:1302.6589 [astro-ph.HE].
[8] C. Gordon and O. Macias, Phys. Rev. D **88**, 083521 (2013), [Erratum: Phys.Rev.D 89, 049901 (2014)], arXiv:1306.5725 [astro-ph.HE].
[9] K. N. Abazajian, N. Canac, S. Horiuchi, and M. Kaplinghat, Phys. Rev. D **90**, 023526 (2014), arXiv:1402.4090 [astro-ph.HE].
[10] T. Daylan, D. P. Finkbeiner, D. Hooper, T. Linden, S. K. N. Portillo, N. L. Rodd, and T. R. Slatyer, Phys. Dark Univ. **12**, 1 (2016), arXiv:1402.6703 [astro-ph.HE].
[11] F. Calore, I. Cholis, and C. Weniger, JCAP **03**, 038 (2015), arXiv:1409.0042 [astro-ph.CO].
[12] K. N. Abazajian, N. Canac, S. Horiuchi, M. Kaplinghat, and A. Kwa, JCAP **07**, 013 (2015), arXiv:1410.6168 [astro-ph.HE].
[13] M. Ajello *et al.* (Fermi-LAT), Astrophys. J. **819**, 44 (2016), arXiv:1511.02938 [astro-ph.HE].
[14] T. Linden, N. L. Rodd, B. R. Safdi, and T. R. Slatyer, Phys. Rev. D **94**, 103013 (2016), arXiv:1604.01026 [astro-ph.HE].
[15] O. Macias, C. Gordon, R. M. Crocker, B. Coleman, D. Paterson, S. Horiuchi, and M. Pohl, Nature Astron. **2**, 387 (2018), arXiv:1611.06644 [astro-ph.HE].
[16] H. A. Clark, P. Scott, R. Trotta, and G. F. Lewis, JCAP **07**, 060 (2018), arXiv:1612.01539 [astro-ph.HE].
[17] K. N. Abazajian, JCAP **03**, 010 (2011), arXiv:1011.4275 [astro-ph.HE].
[18] D. Hooper, I. Cholis, T. Linden, J. Siegal-Gaskins, and T. Slatyer, Phys. Rev. D **88**, 083009 (2013), arXiv:1305.0830 [astro-ph.HE].
[19] F. Calore, M. Di Mauro, and F. Donato, Astrophys. J. **796**, 1 (2014), arXiv:1406.2706 [astro-ph.HE].
[20] I. Cholis, D. Hooper, and T. Linden, JCAP **06**, 043 (2015), arXiv:1407.5625 [astro-ph.HE].
[21] J. Petrović, P. D. Serpico, and G. Zaharijas, JCAP **02**, 023 (2015), arXiv:1411.2980 [astro-ph.HE].
[22] Q. Yuan and K. Ioka, Astrophys. J. **802**, 124 (2015), arXiv:1411.4363 [astro-ph.HE].
[23] T. D. Brandt and B. Kocsis, Astrophys. J. **812**, 15 (2015), arXiv:1507.05616 [astro-ph.HE].
[24] A. Gautam, R. M. Crocker, L. Ferrario, A. J. Ruiter, H. Ploeg, C. Gordon, and O. Macias, (2021), arXiv:2106.00222 [astro-ph.HE].
[25] H. Ploeg, C. Gordon, R. Crocker, and O. Macias, JCAP **12**, 035 (2020), arXiv:2008.10821 [astro-ph.HE].
[26] O. Macias, S. Horiuchi, M. Kaplinghat, C. Gordon, R. M. Crocker, and D. M. Nataf, JCAP **09**, 042 (2019), arXiv:1901.03822 [astro-ph.HE].
[27] R. Bartels, E. Storm, C. Weniger, and F. Calore, Nature Astron. **2**, 819 (2018), arXiv:1711.04778 [astro-ph.HE].
[28] M. Di Mauro, Phys. Rev. D **102**, 103013 (2020), arXiv:2010.02231 [astro-ph.HE].
[29] M. Di Mauro, Phys. Rev. D **103**, 063029 (2021), arXiv:2101.04694 [astro-ph.HE].
[30] S. K. Lee, M. Lisanti, B. R. Safdi, T. R. Slatyer, and W. Xue, Phys. Rev. Lett. **116**, 051103 (2016), arXiv:1506.05124 [astro-ph.HE].
[31] R. Bartels, S. Krishnamurthy, and C. Weniger, Phys. Rev. Lett. **116**, 051102 (2016), arXiv:1506.05104 [astro-ph.HE].
[32] M. Buschmann, N. L. Rodd, B. R. Safdi, L. J. Chang, S. Mishra-Sharma, M. Lisanti, and O. Macias, Phys. Rev. D **102**, 023023 (2020), arXiv:2002.12373 [astro-ph.HE].
[33] L. J. Chang, S. Mishra-Sharma, M. Lisanti, M. Buschmann, N. L. Rodd, and B. R. Safdi, Phys. Rev. D **101**, 023014 (2020), arXiv:1908.10874 [astro-ph.CO].
[34] R. K. Leane and T. R. Slatyer, Phys. Rev. Lett. **123**, 241101 (2019), arXiv:1904.08430 [astro-ph.HE].
[35] D. Malyshev and D. W. Hogg, Astrophys. J. **738**, 181 (2011), arXiv:1104.0010 [astro-ph.CO].
[36] S. K. Lee, M. Lisanti, and B. R. Safdi, JCAP **05**, 056 (2015), arXiv:1412.6099 [astro-ph.CO].
[37] B. Balaji, I. Cholis, P. J. Fox, and S. D. McDermott, Phys. Rev. D **98**, 043009 (2018), arXiv:1803.01952 [astro-ph.HE].
[38] S. D. McDermott, P. J. Fox, I. Cholis, and S. K. Lee, JCAP **07**, 045 (2016), arXiv:1512.00012 [astro-ph.HE].
[39] Y.-M. Zhong, S. D. McDermott, I. Cholis, and P. J. Fox, Phys. Rev. Lett. **124**, 231103 (2020), arXiv:1911.12369 [astro-ph.HE].
[40] S. Caron *et al.*, (2021), arXiv:2103.11068 [astro-ph.HE].
[41] F. List, N. L. Rodd, G. F. Lewis, and I. Bhat, Phys. Rev. Lett. **125**, 241102 (2020), arXiv:2006.12504 [astro-ph.HE].
[42] F. List, N. L. Rodd, and G. F. Lewis, (2021), arXiv:2107.09070 [astro-ph.HE].
[43] S. Caron, G. A. Gómez-Vargas, L. Hendriks, and R. Ruiz de Austri, JCAP **05**, 058 (2018), arXiv:1708.06706 [astro-ph.HE].
[44] F. List, in *Proceedings of the 38th International Conference on Machine Learning* (2021).
[45] K. Cranmer, J. Brehmer, and G. Louppe, Proc. Nat. Acad. Sci. **117**, 30055 (2020), arXiv:1911.01429 [stat.ML].
[46] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, (2019), arXiv:1912.02762 [cs.LG].
[47] D. J. Rezende and S. Mohamed, in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, JMLR Workshop and Conference Proceedings, Vol. 37, edited by F. R. Bach and D. M. Blei (2015) pp. 1530–1538.
[48] S. Mishra-Sharma, N. L. Rodd, and B. R. Safdi, "Supplementary material for NPTFit," (2016).
[49] S. Mishra-Sharma, N. L. Rodd, and B. R. Safdi, Astron.

J. **153**, 253 (2017), arXiv:1612.03173 [astro-ph.HE].
[50] K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelman, Astrophys. J. **622**, 759 (2005), arXiv:astro-ph/0409513.
[51] L. J. Chang, M. Lisanti, and S. Mishra-Sharma, Phys. Rev. D **98**, 123004 (2018), arXiv:1804.04132 [astro-ph.CO].
[52] F. Acero *et al.* (Fermi-LAT), Astrophys. J. Suppl. **218**, 23 (2015), arXiv:1501.02003 [astro-ph.HE].
[53] M. Su, T. R. Slatyer, and D. P. Finkbeiner, Astrophys. J. **724**, 1044 (2010), arXiv:1005.5480 [astro-ph.HE].
[54] J. F. Navarro, C. S. Frenk, and S. D. M. White, Astrophys. J. **462**, 563 (1996), arXiv:astro-ph/9508025 [astro-ph].
[55] J. F. Navarro, C. S. Frenk, and S. D. White, Astrophys.J. **490**, 493 (1997), arXiv:astro-ph/9611107 [astro-ph].
[56] B. Zhou, Y.-F. Liang, X. Huang, X. Li, Y.-Z. Fan, L. Feng, and J. Chang, Phys. Rev. D **91**, 123010 (2015), arXiv:1406.6948 [astro-ph.HE].
[57] J. Bovy, arXiv e-prints , arXiv:2012.02169 (2020), arXiv:2012.02169 [astro-ph.GA].
[58] Gravity Collaboration, Astron. Astrophys. **625**, L10 (2019), arXiv:1904.05721 [astro-ph.GA].
[59] N. L. Rodd and M. W. Toomey, *NPTFit-Sim* (2017).
[60] D. R. Lorimer *et al.*, Mon. Not. Roy. Astron. Soc. **372**, 777 (2006), arXiv:astro-ph/0607640 [astro-ph].
[61] R. T. Bartels, T. D. P. Edwards, and C. Weniger, Mon. Not. Roy. Astron. Soc. **481**, 3966 (2018), arXiv:1805.11097 [astro-ph.HE].
[62] G. H. Collin, N. L. Rodd, T. Erjavec, and K. Perez, (2021), arXiv:2104.04529 [astro-ph.IM].
[63] B. J. Brewer, D. Foreman-Mackey, and D. W. Hogg, Astron. J. **146**, 7 (2013), arXiv:1211.5805 [astro-ph.IM].
[64] R. Liu, J. D. McAuliffe, and J. Regier, (2021), arXiv:2102.02409 [astro-ph.IM].
[65] T. Daylan, S. K. N. Portillo, and D. P. Finkbeiner, Astrophys. J. **839**, 4 (2017), arXiv:1607.04637 [astro-ph.IM].
[66] F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt, (2013), arXiv:1306.2144 [astro-ph.IM].
[67] J. Skilling, Bayesian Anal. **1**, 833 (2006).
[68] J. S. Speagle, Monthly Notices of the Royal Astronomical Society **493**, 3132 (2020).
[69] R. K. Leane and T. R. Slatyer, Phys. Rev. D **102**, 063019 (2020), arXiv:2002.12371 [astro-ph.HE].
[70] R. K. Leane and T. R. Slatyer, Phys. Rev. Lett. **125**, 121105 (2020), arXiv:2002.12370 [astro-ph.HE].
[71] F. Calore, F. Donato, and S. Manconi, (2021), arXiv:2102.12497 [astro-ph.HE].
[72] M. Lisanti, S. Mishra-Sharma, L. Necib, and B. R. Safdi, Astrophys. J. **832**, 117 (2016), arXiv:1606.04101 [astro-ph.HE].
[73] H.-S. Zechlin, A. Cuoco, F. Donato, N. Fornengo, and M. Regis, Astrophys. J. Lett. **826**, L31 (2016), arXiv:1605.04256 [astro-ph.HE].
[74] H.-S. Zechlin, A. Cuoco, F. Donato, N. Fornengo, and A. Vittino, Astrophys. J. Suppl. **225**, 18 (2016), arXiv:1512.07190 [astro-ph.HE].
[75] J. J. Somalwar, L. J. Chang, S. Mishra-Sharma, and M. Lisanti, Astrophys. J. **906**, 57 (2021), arXiv:2009.00021 [astro-ph.CO].
[76] D. B. Rubin, The Annals of Statistics **12**, 1151 (1984).
[77] G. Papamakarios and I. Murray, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16 (Curran Associates Inc., Red Hook, NY, USA, 2016) pp. 1036–1044, arXiv:1605.06376 [stat.ML].
[78] K. Cranmer and G. Louppe, J. Brief Ideas (2016), 10.5281/zenodo.198541.
[79] G. Papamakarios, T. Pavlakou, and I. Murray, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (Curran Associates Inc., Red Hook, NY, USA, 2017) pp. 2335–2344.
[80] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16 (Curran Associates Inc., Red Hook, NY, USA, 2016) pp. 4743–4751.
[81] L. Dinh, J. Sohl-Dickstein, and S. Bengio, in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (2017).
[82] L. Dinh, D. Krueger, and Y. Bengio, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, edited by Y. Bengio and Y. LeCun (2015).
[83] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, edited by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett (2019) pp. 7509–7520.
[84] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, in *1st Workshop on Invertible Neural Networks and Normalizing Flows at ICML 2019* (2019) arXiv:1906.02145.
[85] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019* (2019) arXiv:1810.01367 [cs.LG].
[86] M. Germain, K. Gregor, I. Murray, and H. Larochelle, in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, JMLR Workshop and Conference Proceedings, Vol. 37, edited by F. R. Bach and D. M. Blei (2015) pp. 881–889.
[87] M. Defferrard, M. Milani, F. Gusset, and N. Perraudin, in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* (2020).
[88] N. Perraudin, M. Defferrard, T. Kacprzak, and R. Sgier, Astron. Comput. **27**, 130 (2019), arXiv:1810.12186 [astro-ph.CO].
[89] M. Defferrard, N. Perraudin, T. Kacprzak, and R. Sgier, in *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019) arXiv:1904.05146 [cs.LG].
[90] M. Defferrard, X. Bresson, and P. Vandergheynst, in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, edited by D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett (2016) pp. 3837–3845.
[91] D. P. Kingma and J. Ba, in *3rd International Conference*

on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, edited by Y. Bengio and Y. LeCun (2015).
[92] I. Loshchilov and F. Hutter, in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019* (2019).
[93] S. Mishra-Sharma and K. Cranmer, in *Machine Learning and the Physical Sciences Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS)* (2020) arXiv:2010.10450 [astro-ph.HE].
[94] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, Phys. Rev. D **98**, 052004 (2018), arXiv:1805.00020 [hep-ph].
[95] J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer, Proc. Nat. Acad. Sci. **117**, 5242 (2020), arXiv:1805.12244 [stat.ML].
[96] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, Phys. Rev. Lett. **121**, 111801 (2018), arXiv:1805.00013 [hep-ph].
[97] K. Cranmer, J. Pavez, and G. Louppe, (2015), arXiv:1506.02169 [stat.AP].
[98] J. Hermans, V. Begy, and G. Louppe, arXiv:1903.04057 [cs, stat] (2020), arXiv: 1903.04057.
[99] B. K. Miller, A. Cole, G. Louppe, and C. Weniger, (2020), arXiv:2011.13951 [astro-ph.IM].
[100] B. K. Miller, A. Cole, P. Forré, G. Louppe, and C. Weniger, (2021), 10.5281/zenodo.5043707, arXiv:2107.01214 [stat.ML].
[101] C. Winkler, D. Worrall, E. Hoogeboom, and M. Welling, (2019), arXiv:1912.00042 [cs.LG].
[102] G. Papamakarios, D. Sterratt, and I. Murray, in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 89, edited by K. Chaudhuri and M. Sugiyama (PMLR, 2019) pp. 837–848.
[103] J. Brehmer, F. Kling, I. Espejo, and K. Cranmer, Comput. Softw. Big Sci. **4**, 3 (2020), arXiv:1907.10621 [hep-ph].
[104] M. Stoye, J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer, in *Machine Learning and the Physical Sciences Workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS)* (2018) arXiv:1808.00973 [stat.ML].
[105] G. Louppe, M. Kagan, and K. Cranmer, (2016), arXiv:1611.01046 [stat.ML].
[106] G. Kasieczka and D. Shih, Phys. Rev. Lett. **125**, 122001 (2020), arXiv:2001.05310 [hep-ph].
[107] A. M. Price-Whelan *et al.*, Astron. J. **156**, 123 (2018), arXiv:1801.02634.
[108] T. P. Robitaille *et al.* (Astropy), Astron. Astrophys. **558**, A33 (2013), arXiv:1307.6212 [astro-ph.IM].
[109] A. Lewis, (2019), arXiv:1910.13970 [astro-ph.IM].
[110] F. Perez and B. E. Granger, Computing in Science and Engineering **9**, 21 (2007).
[111] T. Kluyver *et al.*, in *ELPUB* (2016).
[112] J. D. Hunter, Computing In Science & Engineering **9**, 90 (2007).
[113] A. Chen *et al.*, in *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning*, DEEM'20 (Association for Computing Machinery, New York, NY, USA, 2020).
[114] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, "nflows: normalizing flows in PyTorch," (2020).
[115] C. R. Harris *et al.*, Nature **585**, 357 (2020).
[116] W. McKinney, in *Proceedings of the 9th Python in Science Conference*, edited by S. van der Walt and J. Millman (2010) pp. 51 – 56.
[117] M. Defferrard, L. Martin, R. Pena, and N. Perraudin, "PyGSP: Graph Signal Processing in Python," (2017).
[118] E. Bingham *et al.*, J. Mach. Learn. Res. **20**, 28:1 (2019).
[119] A. Paszke *et al.*, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019) pp. 8024–8035.
[120] M. Fey and J. E. Lenssen, in *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019) arXiv:1903.02428 [cs.LG].
[121] W. Falcon *et al.*, "Pytorchlightning/pytorch-lightning: 0.7.6 release," (2020).
[122] M. Waskom *et al.*, "mwaskom/seaborn: v0.8.1 (september 2017)," (2017).
[123] A. Tejero-Cantero *et al.*, Journal of Open Source Software **5**, 2505 (2020).
[124] F. Pedregosa *et al.*, Journal of Machine Learning Research **12**, 2825 (2011).
[125] P. Virtanen *et al.*, Nature Methods (2020), https://doi.org/10.1038/s41592-019-0686-2.
[126] C. da Costa-Luis *et al.*, "tqdm: A fast, Extensible Progress Bar for Python and CLI," (2021).