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[1, 6, †]

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

(Dated: September 5, 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 from a population of relatively bright, unresolved point sources. We leverage recent developments in the field of simulation-based inference, in particular conditional density estimation with normalizing flows, in order to characterize the contribution of unresolved point sources to the GCE. Compared to traditional techniques based on the statistical distribution of photon counts, our method generically attributes a smaller fraction of the total GCE flux to an unresolved point source population in all cases considered, and we obtain a lower bound of $32.5^{+9.1}_{-18.7}$% on such a contribution in our baseline analysis.

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 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 [15, 26, 27] rather than a distribution expected due to DM annihilation, although these conclusions can depend on the 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]. 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 [34–36]. 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 high dimensionality of γ-ray data has traditionally necessitated a description of the photon map in terms of hand-crafted summary quantities *e.g.*, the probability distribution of photon counts [30, 37] or a wavelet decomposition of the photon map [31, 38, 39], in order to enable computationally tractable analyses. While effective, this reduced description necessarily involves loss of information compared to that contained in the original γ-ray map. On the other hand, recent developments in machine learning have enabled analysis techniques that can extract more information from high-dimensional datasets, leveraging more of the information contained in the forward model. Machine learning methods have recently shown promise for analyzing γ-ray data [40] and in particular for understanding the nature of the *Fermi* GCE [41, 42].

In this paper, we leverage recent developments in the

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

field of simulation-based inference (SBI, also referred to as likelihood-free inference; see, *e.g.*, Ref. [43] 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 [44, 45] in order 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 neural network architecture in order to automatically extract summary statistics from γ-ray maps optimized for the downstream task of estimating the distribution of parameters characterizing the contribution of modeled components to the GCE.

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 analysis 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 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 the datasets used. After a brief summary of traditional likelihood-based methods, 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 graph neural networks.

A. Datasets and the forward model

Datasets and region of interest: We use the datasets and spatial templates from Ref. [46] (packaged with Ref. [47]) to create the simulated maps of *Fermi* data in the Galactic Center region. The data and templates used correspond to 413 weeks of *Fermi*-LAT Pass 8 data collected between August 4, 2008 and July 7, 2016. The top quarter of photons by quality of PSF reconstruction in the energy range 2–20 GeV and event class `ULTRACLEANVETO` are used. The recommended quality cuts are applied, corresponding to zenith angle less than 90°, `LAT_CONFIG` = 1, and `DATA_QUAL` > 0.1.[1] The maps are spatially binned using `HEALPix` [48] with `nside`=128. This dataset has been previously used in the literature for traditional analyses based on explicit likelihoods [32, 33, 36] as well as machine learning-based [41] analyses for characterizing the GCE. All templates are normalized, per-pixel, within a region defined by $r < 30°$.

The inner region of the Galactic plane is masked at $|b| < 2°$, and a radial cut $r < 25°$ defines the region of interest (ROI). Even though the GCE is spatially confined to the inner 10–15° of the Galactic Center, 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. We mask resolved PSs from the 3FGL catalog at a radius of 0.8°, approximately corresponding to 99% PSF containment for photons in the data type employed [49].

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 emission accounting for, *e.g.*, uniform emission from unresolved extragalactic sources, *(iii)* emission from resolved PSs included in the *Fermi* 3FGL catalog [49], and *(iv)* lobe-like emission associated with the *Fermi* bubbles [50]. Finally, *(v)* Poissonian DM-like emission is modeled using a line-of-sight integral of the (squared) generalized Navarro-Frenk-White (NFW) [51, 52] 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, 53]. Here, r is the radial distance from the Galactic Center, $r_{\rm s} = 20\,$kpc is the Milky Way scale radius, and we take $R_{\odot} = 8.2\,$kpc as the distance to the Galactic Center [54, 55]. Templates for components *(ii)*–*(iv)* are obtained from Ref. [46].

The Galactic foreground component accounts for emission due to cosmic rays interacting with interstellar gas and radiation. 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 exhibit modulation on small scales. Normalizations of the gas-tracing components, subscripted 'brem/π^0', and the ICS-tracing component of the diffuse Galactic emission, 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], which

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

was based on the same *Fermi* dataset employed here and where 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 other Galactic diffuse models in Sec. IV B.

The total diffuse emission is modeled as a Poisson realization of a linear combination of the diffuse templates T_i, with their corresponding normalizations A_i regarded as parameters of the forward model; $p(x|\{A_i\}) = \mathrm{Pois}\left(\sum_i A_i T_i\right)$. Prior ranges on these parameters are specified in the left column of Tab. II.

PS emission forward model: Each PS population is completely specified by its spatial distribution, described by a spatial template, as well as the expected number of PSs and distribution of photon counts contributed by each PS. Two separate populations are considered: *(i)* PSs spatially correlated with the GCE, modeled using the generalized NFW profile in Eq. (1) with $\gamma = 1.2$, and *(ii)* those correlated with the Galactic disk, modeled as a doubly-exponential profile motivated by studies of the spatial distribution of Galactic millisecond pulsar populations [56, 57],

$$\rho_{\rm Disk}(R,z) \propto \exp\left(-\frac{R}{R_{\rm d}}\right)\exp\left(-\frac{|z|}{z_{\rm s}}\right) \qquad (2)$$

where R and z are the radial and vertical Galactic cylindrical coordinates and the disk scale height is set to $z_{\rm s} = 0.3\,{\rm kpc}$ and radius $R_{\rm d} = 5\,{\rm kpc}$ in the baseline scenario.

Photon counts from a generated PS population are put down on a map according to the *Fermi* PSF at 2 GeV, modeled as a King function, using the algorithm implemented in the code package `NPTFit-Sim` [58]. The SCD $\mathrm{d}N/\mathrm{d}S$ of each PS population, describing the differential number of sources emitting S photons in expectation, is modeled as a doubly-broken power law,

$$\frac{\mathrm{d}N}{\mathrm{d}S} = A_{\rm PS}\begin{cases}\left(\frac{S}{S_{\rm b,1}}\right)^{-n_1}, & S \geq S_{\rm b,1}\\ \left(\frac{S}{S_{\rm b,1}}\right)^{-n_2}, & S_{\rm b,1} > S \geq S_{\rm b,2}\\ \left(\frac{S_{\rm b,2}}{S_{\rm b,1}}\right)^{-n_2}\left(\frac{S}{S_{\rm b,2}}\right)^{-n_3}, & S_{\rm b,2} > S\end{cases} \qquad (3)$$

specified by the breaks $\{S_{\rm b,1}, S_{\rm b,2}\}$, slopes $\{n_1, n_2, n_3\}$, and an overall normalization $A_{\rm PS}$. We note that these parameters specify the spatially-averaged properties of the PS population—variation due to non-uniform exposure of the LAT instrument is accounted for in putting down simulated photon counts.

The final maps are obtained by combining a pixel-wise Poisson draw of the summed astrophysical templates with the simulated PS maps.

Since PS-like and Poissonian components of the model are exactly degenerate in the limit of each PS contributing $\lesssim 1$ photon counts in expectation (see Refs. [33, 59] for a detailed discussion of this degeneracy, in particular in the context of traditional likelihood-based methods), we impose priors on the *expected counts contributed per pixel* for these components. The motivation behind this is to place the two components on the same 'footing' and partially mitigate biases caused by an induced prior preferring one model over the other, which can prevent the expression of the natural degeneracy between the two components. For the PS-like components, this is in practice done by placing a uniform prior on $\langle S^{\rm PS}\rangle = \int \mathrm{d}S\,\langle \mathrm{d}N/\mathrm{d}S\rangle_{\rm pix}$, where $\langle \mathrm{d}N/\mathrm{d}S\rangle$ is the mean source-count distribution per pixel of the respective PS-like component.

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. (3), 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 counts contributed by the respective components in the ROI.

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 distribution 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 given in Tab. I. In order to improve sample efficiency, priors on the overall normalizations of each component are motivated by posteriors obtained from a Poissonian template fit to the real *Fermi* data.

B. Methods based on simplified likelihoods

Of central interest in Bayesian inference is the probability distribution of a set of parameters of interest θ given some data x—the posterior distribution $p(\theta \mid x)$. Bayes' theorem can be used to obtain the posterior as

$p(\theta \mid x) = p(\theta)\, p(x \mid \theta)/\mathcal{Z}$, where $p(x \mid \theta)$ is the likelihood and $\mathcal{Z} \equiv p(x)$ is the Bayesian evidence. 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.

The present problem is no exception. In their simplest incarnation, traditional template fitting methods model the counts map in the region of interest x as a Poisson realization of a linear combination of spatial templates T_i, $p(x|\{A_i\}) = \mathrm{Pois}\left(\sum_i A_i T_i\right)$, where the normalizations A_i of the respective templates are parameters of interest and there are no additional latent variables. Inference on the parameters of interest can easily be performed within a frequentist or Bayesian framework.

In the model described in Sec. II A on the other hand, the presence of a PS population where no PS can be individually localized or characterized introduces a large number of unobserved latent variables, specifically the position of and counts emitted by each PS. Ignoring the contribution from Poissonian templates for the moment and considering only a single PS population, the likelihood for the map x in the region of interest is formally given by

$$p(x \mid \theta_{\rm PS}) = \sum_n \int \mathrm{d}^n z\, p\left(n \mid \theta\right)\, p(z \mid \theta_{\rm PS})\, p(x|z), \quad (4)$$

where $\theta_{\rm PS}$ are the parameters of interest that characterize the spatial and counts distribution of sources parameterized, *e.g.*, by a broken power law as in Eq. (3). 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, and traditional likelihood-based methods aim to simplify the problem setting in order to enable its evaluation in a practical setting.

The 1-point PDF (probability distribution function) framework, first introduced in the context of γ-ray analyses in Ref. [60] and extended in Refs. [30, 37] under the name of non-Poissonian template fitting (NPTF), considers a simplification to Eq. (4) in terms of 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, localizing them within a pixel and modulating their expected number by the modeled spatial template (*e.g.*, GCE-correlated or disk-correlated in our case). 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 detailed discussed as well as numerical implementations of the method in Refs. [30, 47].

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

$$p(x_i^p \mid \theta_{\rm PS}) = \int \mathrm{d}S_i\, p(S_i \mid \theta_{\rm PS})\, p(x_i^p|S_i), \quad (5)$$

where S_i is the expected counts from the PS following some prior probability parameterized by $\theta_{\rm PS}$, in this case following a broken power law as in Eq. (3) with parameters $\theta_{\rm PS} = \{A_{\rm PS}, n_1, n_2, n_3, S_{\rm b,1}, S_{\rm b,2}\}$, and $p(x_i^p|S_i)$ is the distribution of actual counts given latent S_i, assumed to be given by a Poisson distribution. 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:

$$\begin{aligned} p(x^p \,|\theta_{\rm PS}) = &\sum_n p\left(n \mid \theta_{\rm PS}\right) \sum_{n_j} \delta\left(\sum_j n_j j - x^p\right) \\ &\times \delta\left(\sum_j n_j - n\right) \frac{n!}{\prod_j n_j} \prod_{j=1}^{n} p(x_i^p = j \mid \theta_{\rm PS})^{n_j}, \end{aligned} \quad (6)$$

where n_j is the number of PSs contributing j counts. The distribution of the number of PSs is usually assumed to follow a Poisson distribution on the mean expected number of PSs in the ROI $\lambda = \int \mathrm{d}S \left\langle \mathrm{d}N/\mathrm{d}S \right\rangle$ *i.e.*, $p(n \mid \theta_{\rm PS}) = \mathrm{Pois}(n \mid \lambda)$. In this case, the sum over n can be eliminated and the distribution of observed counts is given by

$$\begin{aligned} p(x^p \,|\theta_{\rm PS}) = &\sum_{n_j} \delta\left(\sum_j n_j j - x^p\right) \\ &\times \prod_j \mathrm{Pois}\left(n_j \mid \lambda\, p(x_i^p = j \mid \theta_{\rm PS})\right). \end{aligned} \quad (7)$$

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, $p(x \mid \theta_{\rm PS}) = \prod_p p(x^p \mid \theta_{\rm PS})$. We emphasize that we have only provided a brief overview of the method here, with further analytic simplifications, extensions to include 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. [47]. We note that including a finite point spread function renders the per-pixel likelihood only approximately correct, since this introduces correlations across pixels over on 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. Further generalizations of the method that can account for more extreme variations in the instrumental point-spread function and exposure, which are necessary for application to *e.g.* X-ray data, were introduced and studied in in Ref. [59].

In the same vein as NPTF, probabilistic cataloging [61, 62] is a complementary method for characterizing the sub-threshold contribution of PS populations that has found application in γ-ray analyses [63]. This technique keeps the latent variables in Eq. (4) *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 run the NPTF algorithm on *Fermi* data as a comparison point to previous studies employing the method. We use the NPTF likelihood implemented in `NPTFit` [47] and construct the posterior distribution over the parameters of interest described in Sec. II A using nested sampling implemented `dynesty` [64]. The static variant of nested sampling 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 $\Delta \log \mathcal{Z} < 0.01$.

1-point PDF-based techniques, in particular NPTF, have shown enormous success in characterizing γ-ray PS populations below the *Fermi* detection threshold, both in relation to the GCE [30, 32, 34, 35] and more generally *e.g.*, for characterizing the contribution of extragalactic PSs at high latitudes [65] and for searching for a DM annihilation signal from Galactic subhalos [66]. 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 [35, 36]. In particular, Refs. [30, 33, 36] 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. [34, 35] further showed and described analytically how mismodeling, in particular an unmodeled North-South asymmetry in a DM signal, could lead to the inference of spurious PSs in NPTF analyses of the GCE, a scenario that is seen to be preferred in real *Fermi* data.

The fact that NPTF analyses rely on a per-pixel likelihood can make them especially susceptible to the effects of mismodeling—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 mismodeling can mimic the effect of a PS population *through the statistics of their PDF*, ignoring any spatial correlations that could have an additional regularizing effect in the face of mismodeling. In the rest of this section, we will describe the components of our machine learning-based method that is able to leverage pixel-to-pixel spatial correlations, with the overall aim of extracting more information from γ-ray maps.

[SM: Mention the ignoring of spatial correlations earlier, maybe right before pcat]

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. In this setting, a model is 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 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) [67]—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 distance 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 methods have leveraged advancements in machine learning, in particular the ability of neural networks to extract useful features from high-dimensional data and to flexibly approximate functions and distributions, in order to address these issues, enabling new ways of performing inference on complex models defined through simulations; see Ref. [43] for a review of recent developments.

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 $\hat{p}_\phi(\theta \mid s)$, parameterized by ϕ, conditioned on summaries $s = s(x)$ from simulated samples x. This class of simulation-based inference techniques, known as conditional density estimation [68], 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 $\theta \sim p(\theta)$.

Normalizing flows: In this paper we employ normalizing flows [44, 45], 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 $\theta = f(z)$, from a simple base distribution $\pi(z)$ to the target distribution. Suppressing the conditional dependence on s for the moment for simplicity, we have

$$\hat{p}(\theta) = \pi(z) \left| \det \left(\frac{\partial z}{\partial \theta} \right) \right| = \pi(f^{-1}(\theta)) \left| \det J_{f^{-1}}(\theta) \right| \quad (8)$$

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 a differentiable and invertible transformation with a differentiable inverse. This renders the Jacobian and inverse in Eq. (8) 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 usually defined by a neural network and the base distribution $\pi(z)$ is chosen to be a standard Gaussian $z \sim \mathcal{N}(0, \mathbb{1})$, 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*—it is easy to sample from the base distribution and then run the forward transformation, obtaining a representative distribution of samples drawn according to $\theta \sim \hat{p}(\theta|x)$.

A number of methods have been proposed for defining the flow transformation, *e.g.*, based on affine transformations [69–72], spline-based transformations [73, 74], and continuous-time transformations [75]. We refer to Ref. [44] for a recent review of normalizing flows, including details of practical implementation as well as an overview of proposed methods.

Masked autoregressive flows for (conditional) density estimation: In this paper we use Masked Autoregressive Flows (MAFs) [69] 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 z to θ is expressed as

$$\theta_i = z_i \cdot \exp \alpha_i + \mu_i \quad (9)$$

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 using the recipe introduced in Ref. [76]. The inverse transformation is easily identified from Eq. (9). This allows for an analytically tractable Jacobian determinant, for an N-dimensional distribution given by

$$\left| \det J_{f^{-1}}(\theta) \right| = \exp \left(- \sum_{i=1}^{N} \alpha_i \right) \quad (10)$$

and a forward pass through the flow according to Eq. (9). Multiple transformations f_j can be composed together in order to model more expressive posteriors,

$$\hat{p}(\theta) = \pi \left(f^{-1}(\theta) \right) \prod_{j=1}^{K} \left| \det J_{f_j^{-1}}(z_{j-1}) \right| . \quad (11)$$

The log-probability of the posterior can then be computed using Eq. (10):

$$\log \hat{p}(\theta) = \log \left[\pi \left(f^{-1}(\theta) \right) \right] - \sum_{j=1}^{K} \sum_{i=1}^{N} \alpha_i^j , \quad (12)$$

which acts as the optimization objective during training. Here, we use 8 MAF transformations, each made up of a 2-layer neural network with 128 hidden units with tanh activations and masking used to enforce the autoregressive property. 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. (9) 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 (graph) 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) decomposition of the photon counts map, an 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 *e.g.*, estimating the posterior distributions of the parameters associated with the forward model.

Graph construction and network architecture: The `DeepSphere` architecture [77–79], with a configuration similar to that 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 as 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, 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. [77], 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 symmetric positive semi-definite, the graph Laplacian can be decomposed as $L = U\Lambda U^T$, 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 $h(L)x = Uh(\Lambda)U^T x$ [80].

The `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 filering operation can then be expressed as

$$h(L)x = U\left(\sum_{k=0}^{K} c_k T_k(\Lambda)\right)U^T x = \sum_{k=0}^{K} c_k T_k(L)x. \quad (13)$$

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, 78], 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 representation 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 the 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 neural network-based feature extractor, we implicitly use an approximation to the full data likelihood in Eq. (4) 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 likelihood-based methods described in Sec. II B.

Optimization, training, and evaluation: The optimization objective in Eq. (12), $\log p_\phi(\theta \mid s_\varphi(x))$, is used to train the graph-based and normalizing flow neural networks simultaneously, optimizing the respective parameter sets $\{\varphi, \phi\}$. 10^6 samples are generated using the prior proposal distribution of parameters given in Tab. I, and models are optimized with batch size 128 using the `AdamW` [81, 82] 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 held-out samples, 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 20,000 samples from the flow within the prior distribution using rejection sampling, conditioning each flow transformation on summaries extracted by the graph-based 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.

III. TESTS ON SIMULATED DATA

We begin by validating our pipeline on simulated *Fermi* data. We create simulated datasets by drawing parameter values from ranges motivated by a fit of the model to real *Fermi* data in our baseline ROI (described in Sec. IV), 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 1 shows results of the analysis pipeline conditioning the trained model on five simulated realizations of maps where the GCE consists of purely DM-like emission. The left column shows the median as well as middle-68/95% containment of the point-wise posterior on the source-count distributions of GCE- and disk-correlated point source emission in red and blue, respectively. 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 perfectly degenerate) and the approximate 1-σ threshold for detecting individual sources (below which the degeneracy is often empirically observed in practice [32, 33]). The middle column shows the posteriors on various modeled emission components, excluding 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 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 attributed to unresolved PSs. Some PS-like emission is inferred in most cases as well however, due to a combination of degeneracy with both disk-correlated PSs as well as DM-like flux. The overall flux of all modeled components is seen to be consistent with their true underlying values.

Figure 2 shows the corresponding results for simulated data containing PS-like emission correlated with the GCE. Here, simulations were produced such that the highest break of the GCE-correlated PS SCD was contained between 5 and 20 expected photon counts. We see that PS-like emission is successfully inferred in each case, while at the same time exemplifying a 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. Some degeneracy between GCE- and disk-correlated PSs is seen, although the true SCDs are seen to lie within the 95% containment interval of the inferred point-wise SCD posteriors in each case.

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 pipeline 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 obtained for the various analysis configurations explored, for both the SBI and NPTF pipelines, 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 at which the GCE source count distribution peaks, fraction of the overall emission attributed to disk-correlated PSs, and flux at which the disk source count distribution peaks.

The results of the NPTF analysis are shown in the bottom panel of Fig. 3. 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. The top panel of Fig. 3 shows results using the neural simulation-based analysis pipeline introduced in this paper. Although posteriors for the astrophysical background templates are seen to be broadly consistent with those inferred in the NPTF anlaysis, the preference for PSs is reduced in this case, with $32.5^{+9.1}_{-18.7}$% of the GCE emission being PS-like. We also note that the inferred GCE-correlated SCD peaks at values lower than those inferred from previous NPTF analyses, which have generally found the bulk of expected emission from PSs to lie just below the 3FGL PS detection threshold [30] at $\sim$ 2–3 $\times\, 10^{-10}\,\mathrm{ph\,cm^{-2}\,s^{-1}}$. In this baseline configuration, the SCD peak is constrained to be $2.1^{+1.1}_{-2.0} \times 10^{-11}\,\mathrm{ph\,cm^{-2}\,s^{-1}}$. We note that that the position of the SCD peak is related to the flux at which the largest number of PSs are expected to lie, rather than a statement about which fluxes the majority of the emission comes from.

A. Signal injection test on data

A crucial self-consistency test is the ability of the analysis to recover an artificial signal injected onto the real γ-ray data. As shown in Ref. [36], 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 the

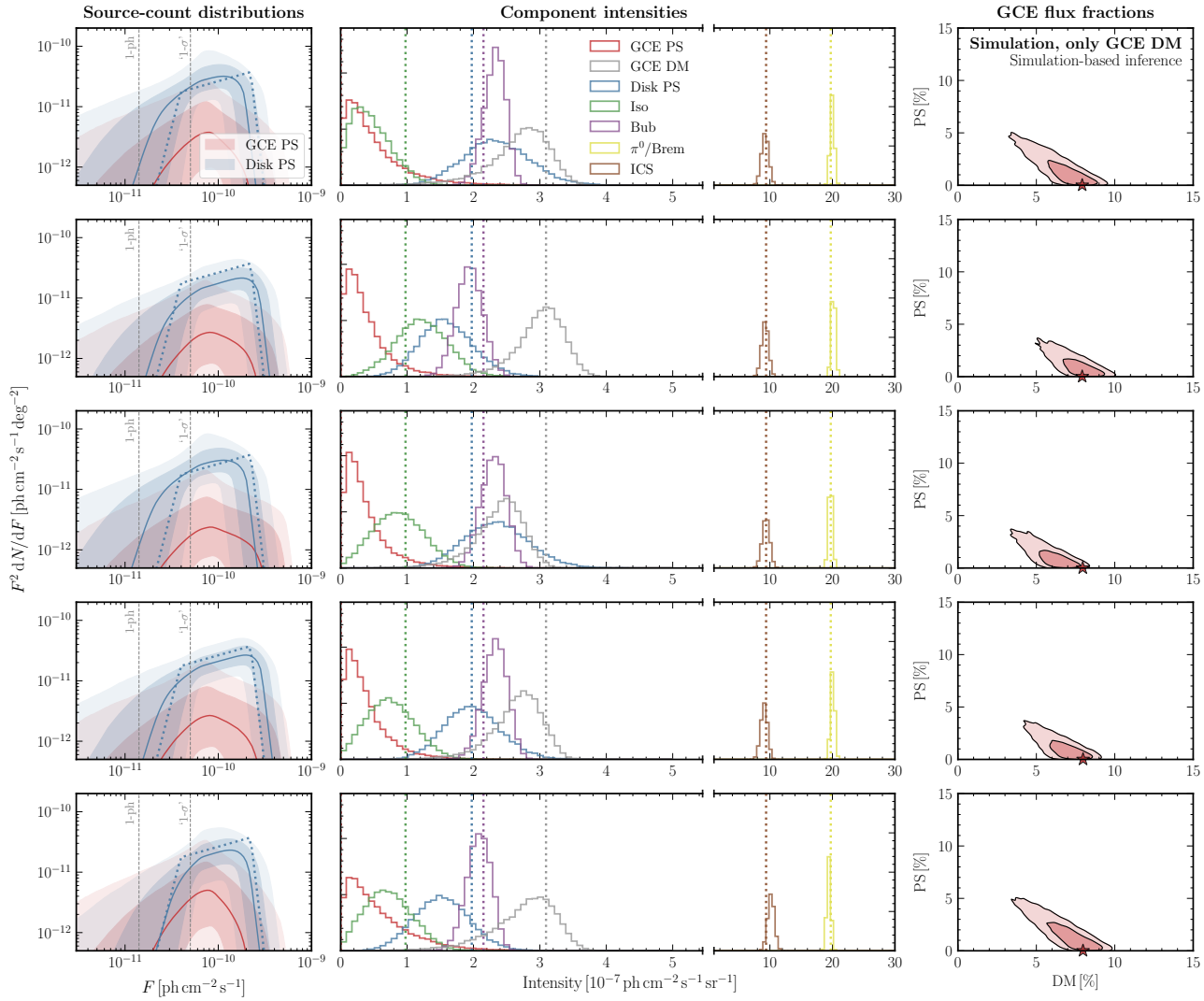

FIG. 1. Results of the analysis pipeline on simulated *Fermi* data where the GCE consists of purely DM-like emission, with different rows corresponding to five different simulated realizations. The left column shows the inferred source-count distribution posteriors for GCE-correlated (red) and disk-correlated (blue) PS. Dashed vertical lines corresponding to the flux associated with 1 expected photon count per source and the approximate 1-σ threshold for detecting individual sources are given for reference. Solid lines correspond to the inferred point-wise median, with the lighter and darker bands representing the point-wise middle-68% and 95% posterior containment respectively. The middle panel shows the posteriors for the Poissonian templates. The right panel 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.

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 4 shows the results of this test, with the different rows corresponding to different signal configuration—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. (3)) at 20, 10, and 5 photon counts respectively, with other parameters the same as those inferred on real *Fermi* data. The leftmost columns shows 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 expected total emission on top of 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 brightest tested DM signal is seen to partially reconstruct as PS-like, which could be attributed to the larger mag-

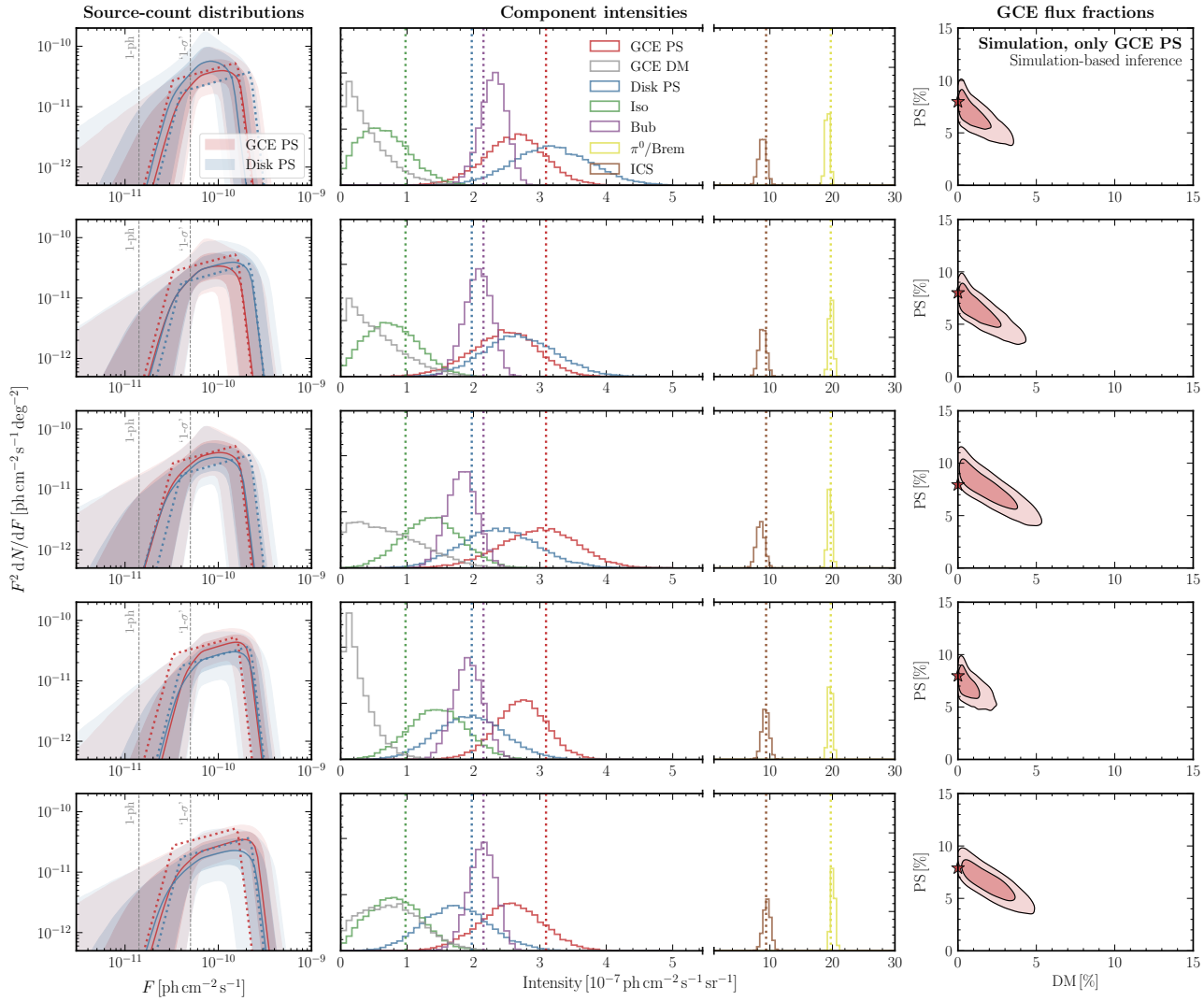

FIG. 2. Same as Fig. 1, but for five simulated realization of *Fermi* data where the GCE consists of predominantly PS-like emission. PS-like emission is inferred in each case, with the other posteriors corresponding faithfully to their true simulated quantities. The GCE-correlated source-count distribution is also seen to be successfully recovered in the left panel.

nitude of Poisson fluctuations in this case mimicking the effect of an unresolved PS population. 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.

B. 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 and the spatial distribution of disk-correlated sources. Results of these 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 formally a worse fit to the present dataset [32], these models have been previously used in the GCE literature [32, 34, 35] and provide a useful comparison point.

Results for these variations are shown in Figs. 5 and 6, 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. Qualitatively similar results are obtained when using Model A (Fig. 5) compared to the baseline analysis in Fig. 3 using Model O, with $34.7^{+9.9}_{-19.8}$% of the GCE flux attributed to PS, roughly half the fraction found by the NPTF analysis. Using Model F on the other hand,

Configuration	Method	GCE fraction	GCE PS fraction	$F^{\rm GCE}_{\rm peak}$	Disk PS fraction	$F^{\rm Disk}_{\rm peak}$	Posteriors
-	-	-	-	[ph cm^{-2} s^{-1}]	-	[ph cm^{-2} s^{-1}]	-
Baseline	SBI	$7.7^{+0.2}_{-0.6}$%	$32.5^{+9.1}_{-18.7}$%	$2.3^{1.2}_{2.2}$	$4.9^{+0.5}_{-1.1}$%	$3.3^{1.1}_{2.7}$	Figure 3
	NPTF	$7.7^{+0.2}_{-0.6}$%	$55.0^{+8.8}_{-22.9}$%	$2.6^{+0.9}_{-1.5}$	$5.4^{+0.5}_{-1.1}$%	$3.7^{+0.9}_{-1.9}$	
Diffuse Model A	SBI	$6.3^{+0.2}_{-0.6}$%	$34.7^{+9.9}_{-19.8}$%	$2.5^{1.3}_{2.3}$	$5.8^{+0.4}_{-1.0}$%	$3.7^{0.9}_{2.9}$	Figure 5
	NPTF	$6.7^{+0.2}_{-0.6}$%	$74.9^{+6.6}_{-22.5}$%	$3.0^{+1.0}_{-1.8}$	$5.1^{+0.5}_{-1.3}$%	$4.3^{+0.6}_{-2.2}$	
Diffuse Model F	SBI	$4.7^{+0.3}_{-0.6}$%	$61.8^{+10.1}_{-27.0}$%	$3.0^{1.0}_{2.6}$	$5.3^{+0.4}_{-1.1}$%	$4.3^{1.0}_{2.4}$	Figure 6
	NPTF	$5.2^{+0.2}_{-0.5}$%	$67.5^{+8.6}_{-26.7}$%	$3.0^{1.0}_{2.0}$	$6.4^{+0.5}_{-1.1}$%	$5.0^{0.4}_{1.9}$	
Thick disk	SBI	$7.9^{+0.3}_{-0.6}$%	$51.5^{+9.2}_{-22.2}$%	$2.6^{1.1}_{2.4}$	$3.2^{+0.5}_{-1.2}$%	$3.5^{1.1}_{2.8}$	Figure 7
	NPTF	$8.2^{+0.3}_{-0.7}$%	$75.0^{+7.1}_{-22.6}$%	$3.0^{+1.0}_{-1.8}$	$2.3^{+0.7}_{-1.1}$%	$3.0^{+1.0}_{-2.0}$	

TABLE II. 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. (3), 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 counts contributed by the respective components in the ROI.

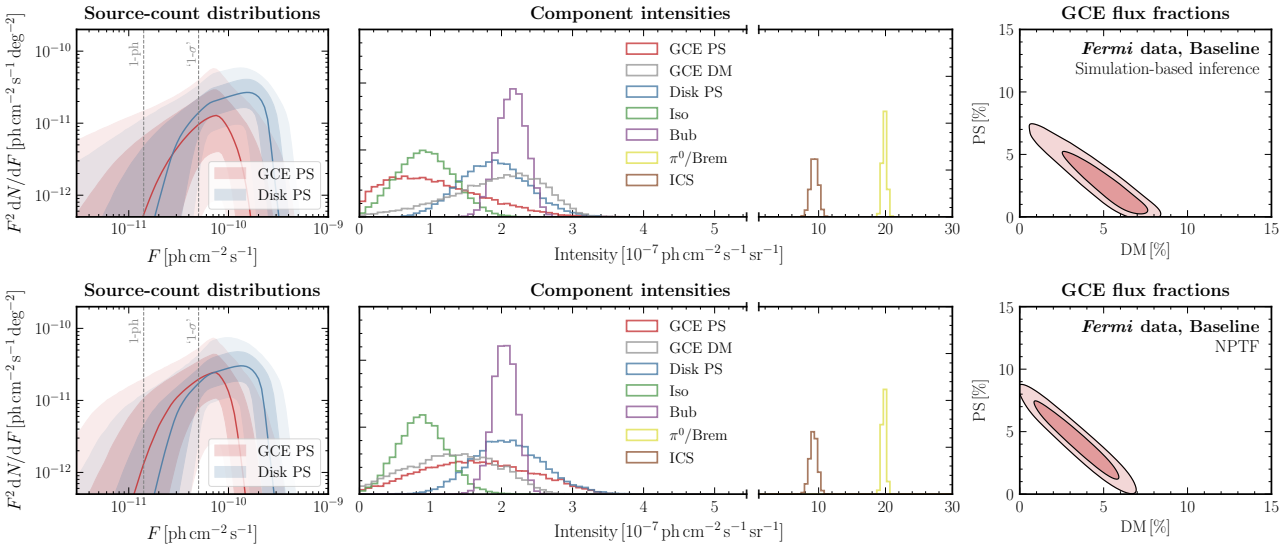

FIG. 3. Results of the baseline analysis on real *Fermi* data. *(Top row)* Analysis using neural simulation-based inference with normalizing flows, and *(bottom row)* using the 1-point PDF likelihood implemented in the non-Poissonian template fitting (NPTF) framework. While moderate preference for a 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).

$52.9^{+10.8}_{-26.1}$% of the emission is attributed to PSs, with a marginally larger amount found by the NPTF analysis. The total emission absorbed by the GCE in this case is about half 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 diffuse models.

Variation on the disk template: The baseline scenario considered a disk-correlated PS population with a spatial distribution given by Eq. (2), 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, 36], where the scale height is typically set to $z_{\rm s} = 1\,$kpc.

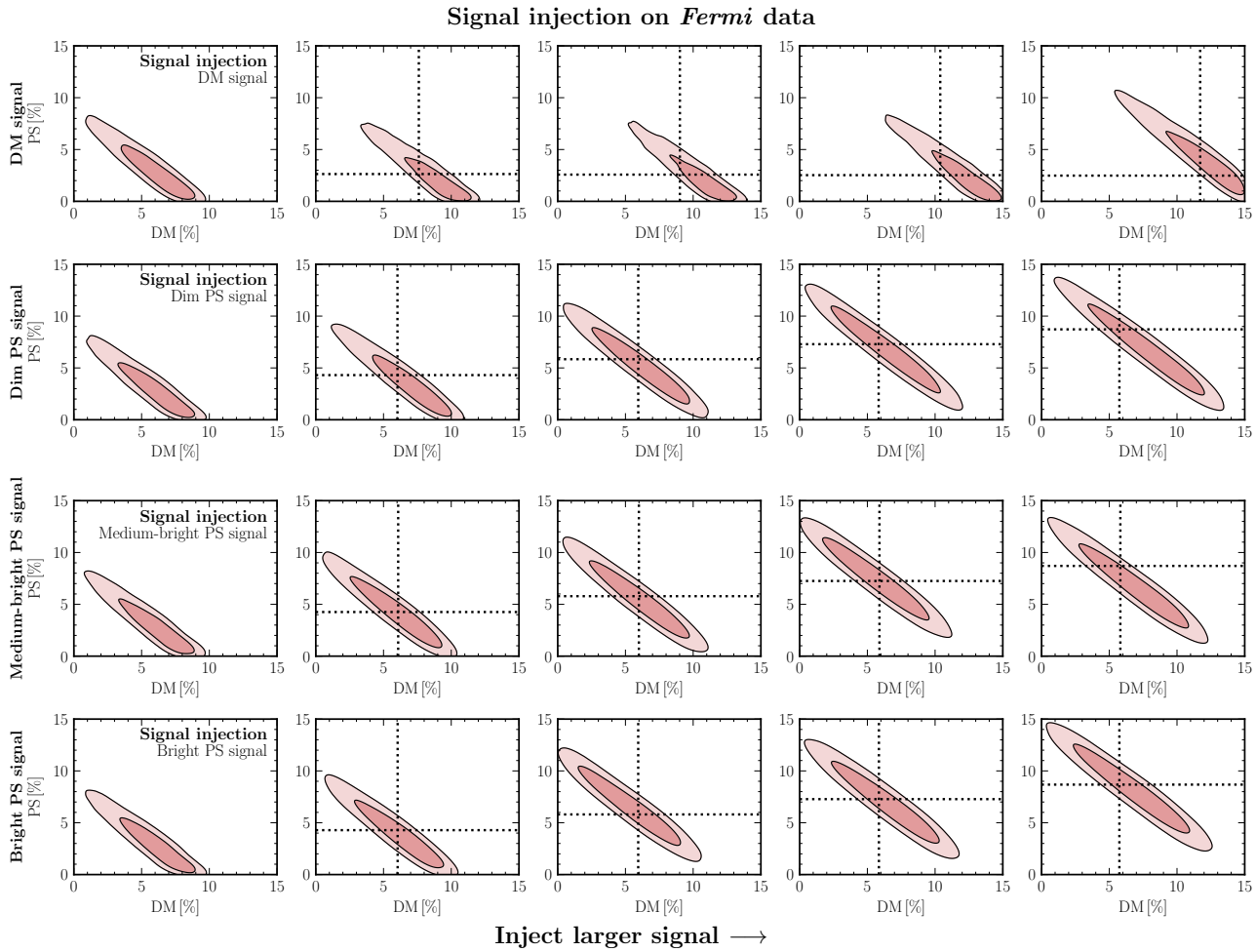

FIG. 4. Joint posterior for the flux fraction of PS-like and DM-like emission when an artificial DM signal is injected onto the real *Fermi* data. The different rows correspond to different signal types, from top to bottom, purely DM, dim PSs (maximum of 5 expected counts per PS), moderately-bright PSs (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 emission on top of the median initial inferred flux. The additional injected DM and PS signals are seen to correctly reconstructed within the respective posterior bounds in all cases.

Results using a thick-disk template for the disk-correlated PS population are shown in Fig. 7. For the SBI analysis, a larger fraction $51.5^{+9.2}_{-22.2}$% of the GCE flux is attributed to a PS population in this case compared to the baseline scenario, with the GCE flux itself being slightly larger. Once again, the NPTF analysis estimates a higher fraction $75.0^{+7.1}_{-22.6}$% of the GCE in point sources.

V. SUSCEPTIBILITY TO MISMODELING

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, 32–35] 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 this scenario. We do so by simulating mock data with a smooth GCE signal, additionally containing known mismodeling, and analyzing it with our baseline pipeline. The ability of our method to correctly characterize the injected signal is then indicative of the level of misattribution 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 by combining samples obtained by analyzing 10 different mock datasets in order to

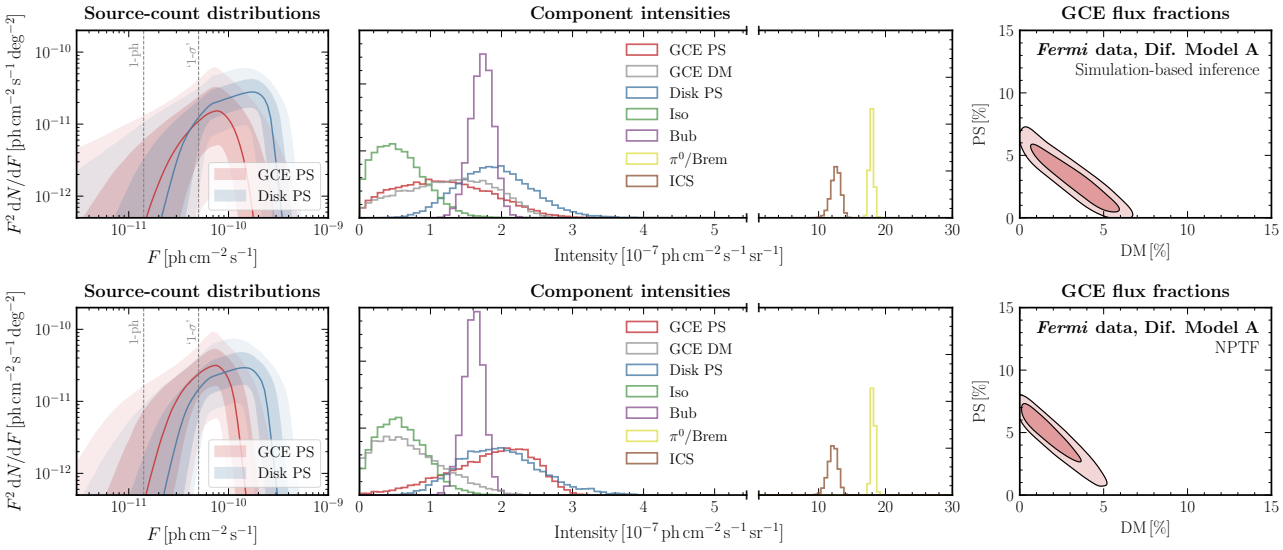

FIG. 5. Same as Fig. 3, but for a model where the diffuse foreground emission is modeled using the alternative Model A.

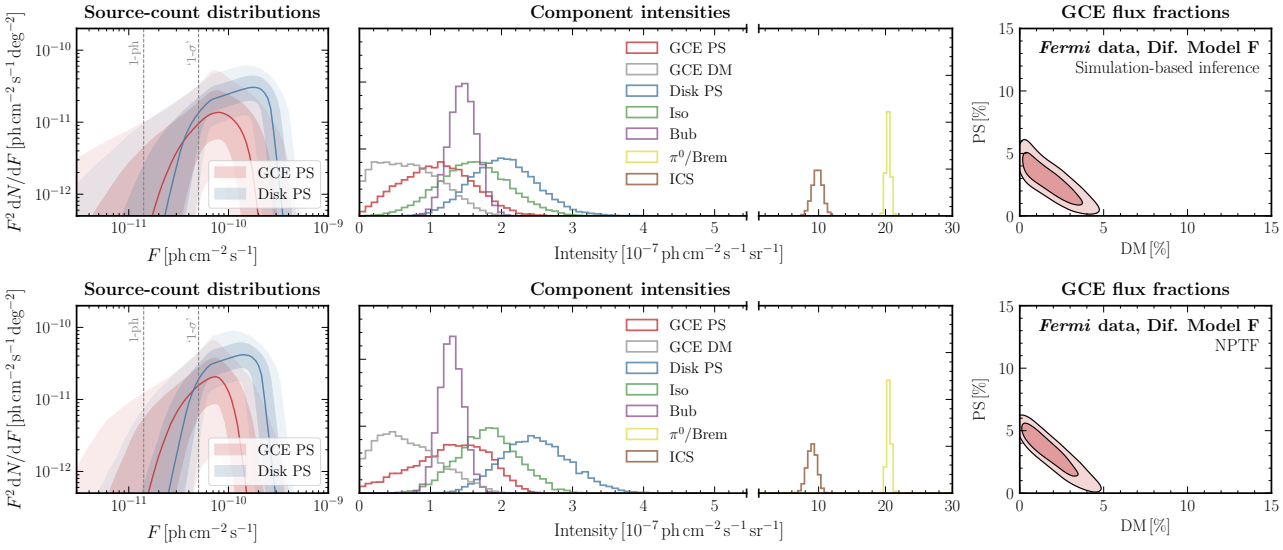

FIG. 6. Same as Fig. 3, but for a model where the diffuse foreground emission is modeled using the alternative Model F.

characterize the 'average' mismodeling associated with a given configuration. The first row of Fig. 9 shows the aggregate analysis without mismodeling *i.e.*, using mock data created with the same templates used in the analysis, as a point of comparison. In all cases tested, while posteriors for certain templates can show systematic biases, preference for a smooth GCE remains robust and the fraction of PS-like emission is compatible with zero (right-most column of Fig. 9).

Test of diffuse mismodeling using an alternative template: We create mock data using diffuse Model A, and analyze it using our baseline analysis pipeline with Model O. The aggregated results over 10 different maps are shown in the second row of Fig. 9. We see that while the marginalized DM posterior faithfully corresponds to the true underlying value, the PS template picks up small amount of residual flux.

A data-driven test of large-scale mismodeling: We construct a data-driven model of large-scale foreground mismodeling and assess the ability of our method to recover a smooth DM-like signal in this case. Following Ref. [83], 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

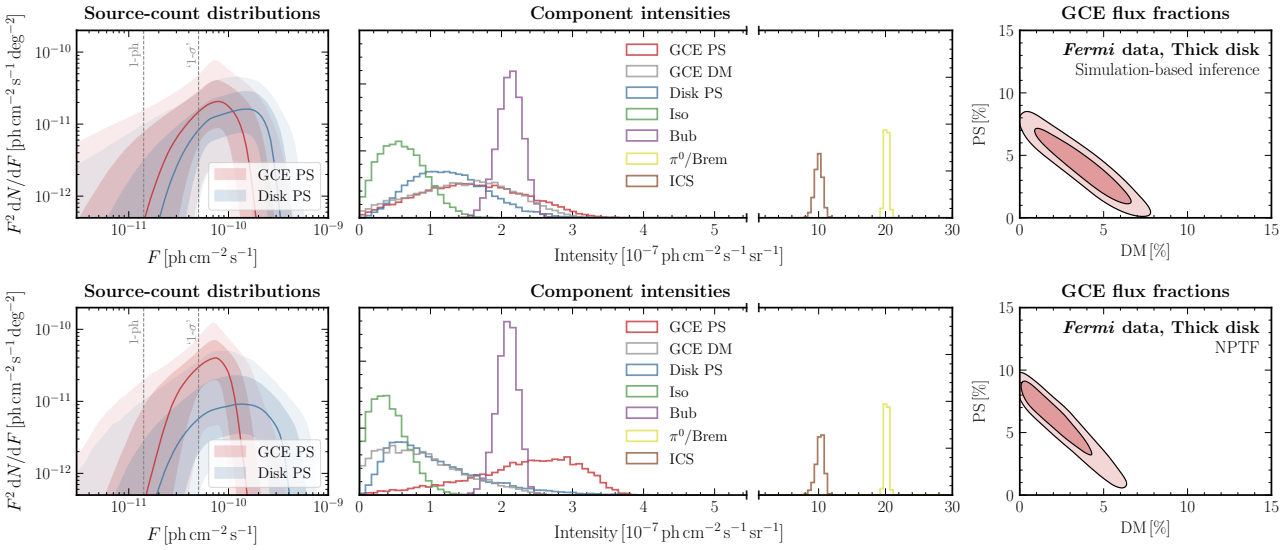

FIG. 7. Same as Fig. 3, but for a model where the spatial distribution of disk-correlated PSs is modeled using a thick-disk template (scale factor $z_s = 1\,\mathrm{kpc}$ in Eq. (2)) rather than the default thin-disk template ($z_s = 0.3\,\mathrm{kpc}$).

(exponentiated) Gaussian process (GP):

$$x \sim \mathrm{Pois}\left(\sum_{i\neq\mathrm{dif}} A_i T_i + \exp(f)\, A_{\mathrm{dif}} T_{\mathrm{dif}}\right). \qquad (14)$$

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 mean m set to zero, and the covariance K described using the Matérn kernel with smoothness parameter $\nu = 5/2$. We refer to Ref. [83] for further details of the analysis, as well a validation of the GP-augmented template fitting pipeline on simulated data.

Five fair 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. 8. The largest mismodeling by magnitude in this case is inferred to be concentrated in the southern regions of the baseline ROI. We note that 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, indicating 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.

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 then analyzed with our standard pipeline, using the unmodulated Model O to model the diffuse emission.

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, 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. (2), 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. While disk mismodeling can distort the inferred SCD of disk-correlated PSs away from the truth, the smooth GCE signal is seen to be successfully recovered in this case.

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. [34, 35], a North-South asymmetry in a putative dark matter signal, if unaccounted for, could lead to the inference of a spurious PS population associated with the purely smooth, asymmetric signal in the traditional NPTF framework. Refs. [34, 35] 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 smaller 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

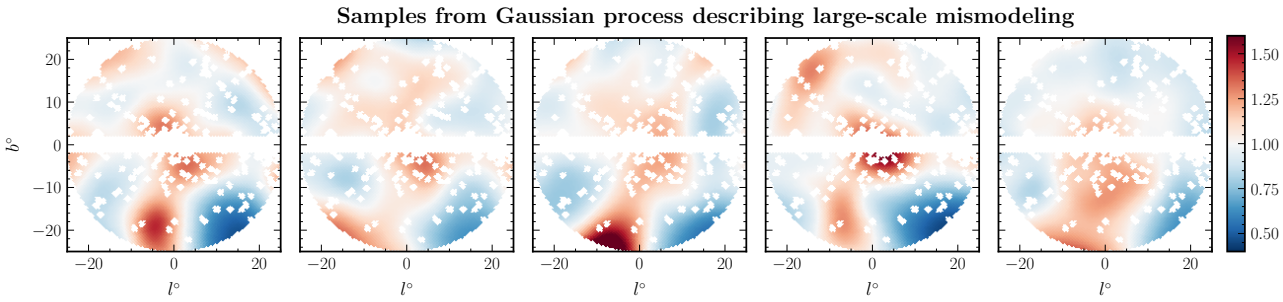

FIG. 8. Five fair samples from the Gaussian process description of large-scale multiplicative mismodeling associated with the gas-correlated component of diffuse foreground Model O when applied to the real *Fermi* data.

matter signal within our framework by running our 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. [34, 35]. The results of this test on 10 such simulated realizations is shown in the last row of Fig. 9. We see that the presence of a substantially asymmetric DM signal has only a marginal impact on the inferred posteriors, and does not lead to a spurious preference for a PS population as was found in the NPTF framework. We attribute this to the fact that the `DeepSphere`-based 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 in Ref. [32], we emphasize that the presence of an 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.

VI. DISCUSSION AND CONCLUSIONS

In this paper, we have leveraged recent advances in neural simulation-based inference in order to characterize a putative point source population associated with the observed *Fermi* Galactic Center Excess. Consistent with Ref [41] which used Bayesian neural networks and first leveraged the `DeepSphere` graph-based 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 (non-Poissonian template fitting). In particular, depending on the analysis configuration, we find a median value of $\sim$ 30–50% as the fraction of GCE emission that can be attributed to a PS population, with the inferred source-count distribution peaking at smaller fluxes ~ 2–$3 \times 10^{-11}\,\mathrm{ph\,cm^{-2}\,s^{-1}}$ compared to values found in previous analyses based on the non-Poissonian template fitting (NPTF) framework [30], where the SCD is seen to peak just below the threshold for resolution of individual PSs, roughly $\sim 2 \times 10^{-10}\,\mathrm{ph\,cm^{-2}\,s^{-1}}$. The NPTF analyses performed in this work find a similarly dim source-count distribution, in all cases however attributing a larger fraction $\sim$ 50–75% of the GCE to a PS population as compared to the corresponding SBI analyses.

Our qualitative conclusions are robust to the systematic variations we have explored, including variations on the the diffuse foreground model and spatial modeling of disk-correlated PS emission. We used a novel Gaussian process-based method to construct a data-driven model of large-scale mismodeling, finding our method to be resilient to such mismodeling when it comes to differentiating between a smooth and PS-like GCE. As in any Galactic Center γ-ray analysis, given the poorly understood astrophysical emission in this region, 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.

Several improvements to the framework presented here are possible. The inclusion of energy-binning information in the analysis can be implemented 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 and flow 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 [84–90] or flow-based likelihood estimation [91, 92] can provide complementary ways to characterize the γ-ray PS population in the Galactic Center. Additionally, the use of sequential active-learning methods [92] and methods that make use of additional latent information from the simulator [84–86, 93, 94] can significantly improve the sample efficiency

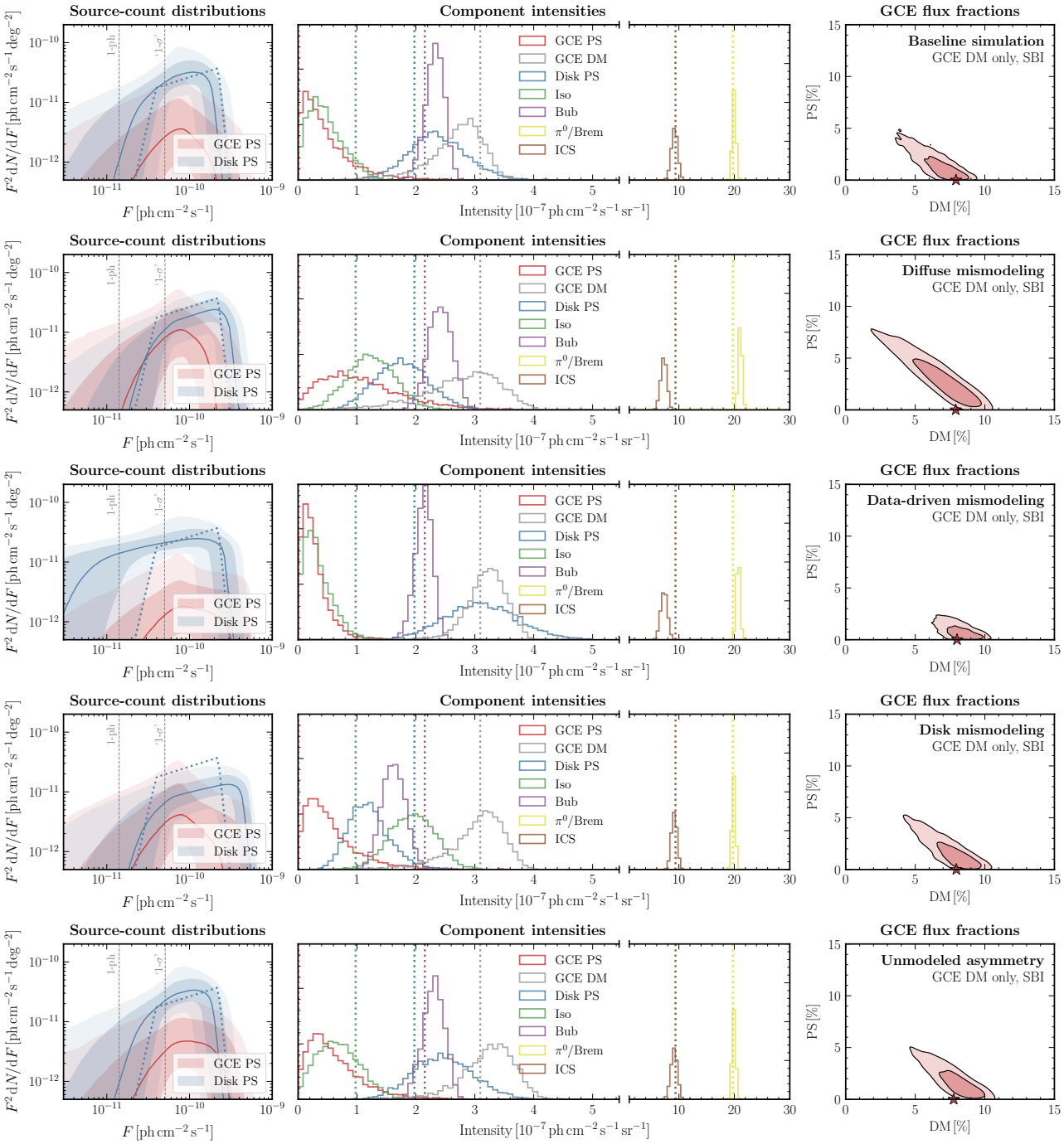

FIG. 9. Effect of mismodeling on a smooth GCE within our analysis framework. Each row shows aggregate posteriors collected over 10 simulated samples; row-wise from top to bottom: *(i)* No mismodeling; simulated data is constructed with the same templates as those used in the forward model. *(ii)* Mock data created with diffuse Model A, showing the 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.

of simulations and allow for extensions to more complex latent spaces, which will be important in particular for an energy-binned analysis and when 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 [95] 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. Similarly, 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 is available at `https://github.com/smsharma/fermi-flows`.

ACKNOWLEDGMENTS

We thank Johann Brehmer, Florian List, Nick Rodd, and Tracy Slatyer for helpful conversations. 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. 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. SM is supported by the NSF CAREER grant PHY-1554858, NSF grants PHY-1620727 and PHY-1915409, and the Simons Foundation. 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/`). 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` [96, 97], `dynesty` [64], `getdist` [98], `IPython` [99], Jupyter [100], `matplotlib` [101], `MLflow` [102], `nflows` [103], `NPTFit` [47], `NumPy` [104], `pandas` [105], `Pyro` [106], `PyTorch` [107], `PyTorch Geometric` [108], `PyTorch Lightning` [109], `seaborn` [110], `sbi` [111], `scikit-learn` [112], `SciPy` [113], and `tqdm` [114] software packages. We acknowledge the use of the code repository associated with Ref. [41], in particular the associated data products and templates.[2] We acknowledge the use of the code repository associated with Ref. [77], in particular the implementation of the `DeepSphere` graph convolutional kernel.[3]

[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].

[2] `https://github.com/FloList/GCE_NN`
[3] `https://github.com/deepsphere/deepsphere-pytorch`

[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. **125**, 121105 (2020), arXiv:2002.12370 [astro-ph.HE].
[35] R. K. Leane and T. R. Slatyer, Phys. Rev. D **102**, 063019 (2020), arXiv:2002.12371 [astro-ph.HE].
[36] R. K. Leane and T. R. Slatyer, Phys. Rev. Lett. **123**, 241101 (2019), arXiv:1904.08430 [astro-ph.HE].
[37] S. K. Lee, M. Lisanti, and B. R. Safdi, JCAP **05**, 056 (2015), arXiv:1412.6099 [astro-ph.CO].
[38] B. Balaji, I. Cholis, P. J. Fox, and S. D. McDermott, Phys. Rev. D **98**, 043009 (2018), arXiv:1803.01952 [astro-ph.HE].
[39] S. D. McDermott, P. J. Fox, I. Cholis, and S. K. Lee, JCAP **07**, 045 (2016), arXiv:1512.00012 [astro-ph.HE].
[40] S. Caron, K. Dijkstra, C. Eckner, L. Hendriks, G. Jóhannesson, B. Panes, R. Ruiz De Austri, and G. Zaharijas, (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] S. Caron, G. A. Gómez-Vargas, L. Hendriks, and R. Ruiz de Austri, JCAP **05**, 058 (2018), arXiv:1708.06706 [astro-ph.HE].
[43] K. Cranmer, J. Brehmer, and G. Louppe, Proceedings of the National Academy of Sciences **117**, 30055 (2020).
[44] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, arXiv preprint arXiv:1912.02762 (2019).
[45] D. Rezende and S. Mohamed, in *International Conference on Machine Learning* (PMLR, 2015) pp. 1530–1538.
[46] S. Mishra-Sharma, N. L. Rodd, and B. R. Safdi, "Supplementary material for NPTFit," (2016).
[47] S. Mishra-Sharma, N. L. Rodd, and B. R. Safdi, Astron. J. **153**, 253 (2017), arXiv:1612.03173 [astro-ph.HE].
[48] 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.
[49] F. Acero *et al.* (Fermi-LAT), Astrophys. J. Suppl. **218**, 23 (2015), arXiv:1501.02003 [astro-ph.HE].
[50] M. Su, T. R. Slatyer, and D. P. Finkbeiner, Astrophys. J. **724**, 1044 (2010), arXiv:1005.5480 [astro-ph.HE].
[51] J. F. Navarro, C. S. Frenk, and S. D. M. White, Astrophys. J. **462**, 563 (1996), arXiv:astro-ph/9508025 [astro-ph].
[52] J. F. Navarro, C. S. Frenk, and S. D. White, Astrophys.J. **490**, 493 (1997), arXiv:astro-ph/9611107 [astro-ph].
[53] 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].
[54] J. Bovy, arXiv e-prints , arXiv:2012.02169 (2020), arXiv:2012.02169 [astro-ph.GA].
[55] Gravity Collaboration, Astron. Astrophys. **625**, L10 (2019), arXiv:1904.05721 [astro-ph.GA].
[56] D. R. Lorimer *et al.*, Mon. Not. Roy. Astron. Soc. **372**, 777 (2006), arXiv:astro-ph/0607640 [astro-ph].
[57] R. T. Bartels, T. D. P. Edwards, and C. Weniger, Mon. Not. Roy. Astron. Soc. **481**, 3966 (2018), arXiv:1805.11097 [astro-ph.HE].
[58] N. L. Rodd and M. W. Toomey, *NPTFit-Sim* (2017).
[59] G. H. Collin, N. L. Rodd, T. Erjavec, and K. Perez, (2021), arXiv:2104.04529 [astro-ph.IM].
[60] D. Malyshev and D. W. Hogg, Astrophys. J. **738**, 181 (2011), arXiv:1104.0010 [astro-ph.CO].
[61] B. J. Brewer, D. Foreman-Mackey, and D. W. Hogg, Astron. J. **146**, 7 (2013), arXiv:1211.5805 [astro-ph.IM].
[62] R. Liu, J. D. McAuliffe, and J. Regier, arXiv e-prints , arXiv:2102.02409 (2021), arXiv:2102.02409 [astro-ph.IM].
[63] T. Daylan, S. K. N. Portillo, and D. P. Finkbeiner, Astrophys. J. **839**, 4 (2017), arXiv:1607.04637 [astro-ph.IM].
[64] J. S. Speagle, Monthly Notices of the Royal Astronomical Society **493**, 3132 (2020).
[65] M. Lisanti, S. Mishra-Sharma, L. Necib, and B. R. Safdi, Astrophys. J. **832**, 117 (2016), arXiv:1606.04101 [astro-ph.HE].
[66] J. J. Somalwar, L. J. Chang, S. Mishra-Sharma, and M. Lisanti, Astrophys. J. **906**, 57 (2021), arXiv:2009.00021 [astro-ph.CO].
[67] D. B. Rubin, The Annals of Statistics **12**, 1151 (1984).
[68] G. Papamakarios and I. Murray, arXiv:1605.06376 [cs,

stat] (2018), arXiv: 1605.06376.
[69] 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.
[70] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, Advances in neural information processing systems **29**, 4743 (2016).
[71] L. Dinh, J. Sohl-Dickstein, and S. Bengio, arXiv preprint arXiv:1605.08803 (2016).
[72] L. Dinh, D. Krueger, and Y. Bengio, arXiv preprint arXiv:1410.8516 (2014).
[73] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, Advances in Neural Information Processing Systems **32**, 7511 (2019).
[74] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, arXiv preprint arXiv:1906.02145 (2019).
[75] W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, arXiv preprint arXiv:1810.01367 (2018).
[76] M. Germain, K. Gregor, I. Murray, and H. Larochelle, in *International Conference on Machine Learning* (PMLR, 2015) pp. 881–889.
[77] M. Defferrard, M. Milani, F. Gusset, and N. Perraudin, arXiv preprint arXiv:2012.15000 (2020).
[78] N. Perraudin, M. Defferrard, T. Kacprzak, and R. Sgier, Astron. Comput. **27**, 130 (2019), arXiv:1810.12186 [astro-ph.CO].
[79] M. Defferrard, N. Perraudin, T. Kacprzak, and R. Sgier, in *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019) arXiv:1904.05146.
[80] M. Defferrard, X. Bresson, and P. Vandergheynst, Advances in neural information processing systems **29**, 3844 (2016).
[81] D. P. Kingma and J. Ba, in *ICLR (Poster)* (2015).
[82] I. Loshchilov and F. Hutter, in *International Conference on Learning Representations* (2019).
[83] S. Mishra-Sharma and K. Cranmer, in *34th Conference on Neural Information Processing Systems* (2020) arXiv:2010.10450 [astro-ph.HE].
[84] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, Physical Review D **98**, 052004 (2018), arXiv: 1805.00020.
[85] J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer, Proceedings of the National Academy of Sciences **117**, 5242 (2020), arXiv: 1805.12244.
[86] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, Physical Review Letters **121**, 111801 (2018), arXiv: 1805.00013.
[87] K. Cranmer, J. Pavez, and G. Louppe, arXiv:1506.02169 [physics, stat] (2016), arXiv: 1506.02169.
[88] J. Hermans, V. Begy, and G. Louppe, arXiv:1903.04057 [cs, stat] (2020), arXiv: 1903.04057.
[89] B. K. Miller, A. Cole, G. Louppe, and C. Weniger, (2020), arXiv:2011.13951 [astro-ph.IM].
[90] B. K. Miller, A. Cole, P. Forré, G. Louppe, and C. Weniger, (2021), 10.5281/zenodo.5043707, arXiv:2107.01214 [stat.ML].
[91] C. Winkler, D. Worrall, E. Hoogeboom, and M. Welling, arXiv preprint arXiv:1912.00042 (2019).
[92] G. Papamakarios, D. Sterratt, and I. Murray, in *The 22nd International Conference on Artificial Intelligence and Statistics* (PMLR, 2019) pp. 837–848.
[93] J. Brehmer, F. Kling, I. Espejo, and K. Cranmer, Comput. Softw. Big Sci. **4**, 3 (2020), arXiv:1907.10621 [hep-ph].
[94] M. Stoye, J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer, (2018), arXiv:1808.00973 [stat.ML].
[95] G. Louppe, M. Kagan, and K. Cranmer, (2016), arXiv:1611.01046 [stat.ML].
[96] A. M. Price-Whelan *et al.*, Astron. J. **156**, 123 (2018), arXiv:1801.02634.
[97] T. P. Robitaille *et al.* (Astropy), Astron. Astrophys. **558**, A33 (2013), arXiv:1307.6212 [astro-ph.IM].
[98] A. Lewis, (2019), arXiv:1910.13970 [astro-ph.IM].
[99] F. Perez and B. E. Granger, Computing in Science and Engineering **9**, 21 (2007).
[100] T. Kluyver *et al.*, in *ELPUB* (2016).
[101] J. D. Hunter, Computing In Science & Engineering **9**, 90 (2007).
[102] A. Chen, A. Chow, A. Davidson, A. DCunha, A. Ghodsi, S. A. Hong, A. Konwinski, C. Mewald, S. Murching, T. Nykodym, *et al.*, in *Proceedings of the fourth international workshop on data management for end-to-end machine learning* (2020) pp. 1–4.
[103] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios, "nflows: normalizing flows in PyTorch," (2020).
[104] C. R. Harris *et al.*, Nature **585**, 357 (2020).
[105] W. McKinney, in *Proceedings of the 9th Python in Science Conference*, edited by S. van der Walt and J. Millman (2010) pp. 51 – 56.
[106] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman, Journal of Machine Learning Research (2018).
[107] 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.
[108] M. Fey and J. E. Lenssen, in *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019).
[109] W. Falcon *et al.*, "Pytorchlightning/pytorch-lightning: 0.7.6 release," (2020).
[110] M. Waskom *et al.*, "mwaskom/seaborn: v0.8.1 (september 2017)," (2017).
[111] A. Tejero-Cantero *et al.*, Journal of Open Source Software **5**, 2505 (2020).
[112] F. Pedregosa *et al.*, Journal of Machine Learning Research **12**, 2825 (2011).
[113] P. Virtanen *et al.*, Nature Methods (2020), https://doi.org/10.1038/s41592-019-0686-2.
[114] C. O. da Costa-Luis, JOSS **4**, 1277 (2019).