Deep Learning to Reconstruct Gas Skymaps for Dark Matter Detection

Alexander Shmakov[*]
Department of Computer Science
University of California Irvine
ashmakov@uci.edu

Christopher M. Karwin[*]
Department of Physics & Astronomy
Clemson University
ckarwin@clemson.edu

Mohammadamin Tavakoli[*]
Department of Computer Science
University of California Irvine
mohamadt@uci.edu

Simona Murgia
Department of Physics & Astronomy
University of California Irvine
smurgia@uci.edu

Pierre Baldi
Department of Computer Science
University of California Irvine
pfbaldi@ics.uci.edu

Abstract

Fermi–LAT 's measurement of excess gamma-rays emanating from the Galactic center has sparked debate as to the source of this unexpected signal. Dark matter annihilation has emerged as a potential explanation, but confirming this hypothesis requires a comprehensive understanding of conventional gamma-ray sources. We evaluate and compare several machine learning approaches to accurately reconstruct ^{13}CO interstellar gas concentration maps, an important indicator for the primary source of gamma-rays. We apply recent advancements in machine learning to estimate these skymaps via deep neural networks and Gaussian processes. This first attempt at employing image reconstruction techniques for modeling the Milky Way gamma-ray background present an important step towards eliminating known gamma sources and uncovering the nature of the *Fermi*–LAT excess.

1 Introduction

Experimental data indicates that dark matter constitutes the majority of the mass in the Universe; however, its nature is not yet understood. Extensions to the Standard Model of particle physics predict the existence of weakly interacting particles that can annihilate, or decay, into gamma-rays. This electromagnetic debris is a potential messenger of dark matter annihilation, and, if detected, would provide valuable insight into the nature of dark matter. However, this search is limited by our incomplete understanding of gamma-ray emissions from conventional astrophysical processes.

[*]These authors contributed equally.

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

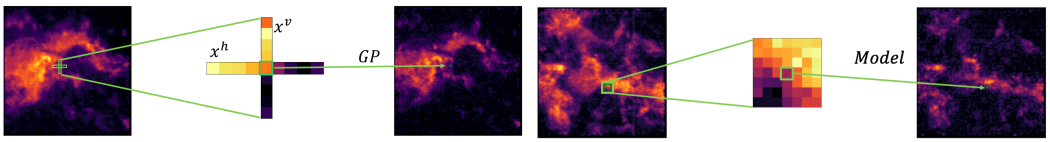

Figure 1: Left: An example of a vertical and a horizontal patches of CO map as the inputs of the Gaussian Process. Right: An example of a patch from the CO map use by convolutional neural network to predict the center of the corresponding patch in a ^{13}CO map.

The *Fermi* Large Area Telescope (*Fermi*–LAT) is a gammay-ray telescope in low-earth orbit performing a survey of gamma emissions in the universe. An excess in the *Fermi*–LAT gamma-ray observations of the Galactic center was first claimed by Goodenough and Hooper [17, 20], and, since then, numerous other analyses have confirmed its presence ([2, 1, 12, 15, 5, 4]). A striking feature of this emission is that its spatial morphology is consistent with annihilating dark matter. However, due to our incomplete understanding of conventional gamma-ray sources, the excess cannot be precisely characterized. Alternative explanations for the observations have been proposed, with the leading one attributing the signal to a collection of pulsars ([2, 1, 23]). The *Fermi*–LAT gamma-ray excess remains a debated topic. [22, 13].

One crucial direction into settling this debate is to better characterize the contribution of other gamma emitters. An accurate baseline would highlight any measured excess that cannot be attributed to known sources. The brightest contribution to the gamma-ray emission in the direction of the Galactic center originates from high-energy cosmic-rays interacting with the interstellar medium in the Galaxy, including hydrogen-rich interstellar gas. We refer to this emission as Galactic diffuse emission [3]. Models for this emission exist, however, they are limited by our incomplete understanding of the gas's structure. In particular, the data used to construct these models lack the resolution to capture highly structured components of the Galactic diffuse emission, and therefore could confound our ability to differentiate a dark matter signal from other, more structured components in the Galaxy [5].

In this study, we refine modeling of the highly structured component of the Galactic diffuse emission, with a focus on the emission related to interstellar molecular hydrogen (H_2). Directly measuring the concentration of H_2 in the Galaxy is difficult since it does not emit at characteristic frequencies. Instead, we use other molecules found in conjunction with H_2 as proxies for its distribution. Typically the H_2 concentration is modeled using CO measurements via radio telescopes. The concentrations of these molecules are related [25], and conversion factors have been determined to infer the H_2 concentration from CO measurements. However, CO becomes optically thick in high density cores of interstellar clouds [28], reducing the trace accuracy in these regions. Instead, we can employ the rarer isotopologues of CO , ^{13}CO and C^{18}O , which remain optically thin at higher densities. Knowing the concentration of CO and its isotopologues is therefore vital for accurately modeling the concentration of H_2 clouds; their contribution to Galactic diffuse emission; and, ultimately, to resolve the nature of the excess gamma-rays observed by *Fermi*–LAT in the Galactic center.

While promising, this approach is not straightforward. Primarily, ^{13}CO and C^{18}O emit at different spectral frequencies and with much lower brightness than CO [14]. This introduces added complexity to their measurement, and few surveys have included the isotopologues. Recently, the MOPRA CO Survey [10] has provided high resolution measurements of all three isotopologues in the vicinity of the Galactic center. This data provides an opportunity to use recent advancements in deep learning to model the isotopologue concentrations given CO measurements. We employ the MOPRA data as training examples for learning such a model. We fit classical Gaussian processes as well as deep (convolutional) neural networks, inspired by their success in computer vision, as well as their many other applications in physics [7, 9, 8, 27, 16, 6, 24]. In this paper, we present the first attempt at modeling the concentration of the ^{13}CO , the more ubiquitous and cleanly observable isotopologue, given the baseline CO concentration measured by MOPRA.

2 Methods

2.1 Data

Our dataset is generated from the MOPRA molecular gas survey [10, 11]. The complete survey contains 50 $1^\circ \times 1^\circ$ images along the Galactic plane measuring the brightness of the three major isotopologues of carbon monoxide: CO , $^{13}\mathrm{CO}$, and $\mathrm{C}^{18}\mathrm{O}$. The data covers Galactic longitudes between 300° and 350° and Galactic latitudes between $\pm 0.5^\circ$ ($^\circ$ = degrees). The brightness of the gas is measured as a function of gas velocity. From this, the column density for each is obtained by integrating over a given velocity range and multiplying by a physically-determined conversion factor. In particular, we divide the gas into 17 velocity bins, corresponding to Galactocentric radii. Assuming that the Galaxy is in uniform circular motion, the velocity of the gas traces its distance from the center of the Galaxy.

In order to simplify the regression problem and enlarge our effective dataset, we train a model that maps small $D_1 \times D_2$ sections (patches) of the sky to their respective center points (Figure 1). The validity of this simplification requires the gas column density to be locally correlated. That is, we assume it is unlikely that distant regions of the galaxy could significantly affect each others' concentrations. We evaluate this assumption in Section 3.1. This simplified formulation greatly reduces model size and raises the number of training examples from just 50 images to over 50 million patches when $D_1 = D_2 = 7$.

The original survey data contains systemic noise in the form of zero-valued vertical and horizontal artifacts. We apply a $\sigma = 1$ pixel Gaussian filter across both source and target images, which removes the large artifacts. Additionally, to ensure the highest quality data in the rarer isotopologues ($^{13}\mathrm{CO}$ and $\mathrm{C}^{18}\mathrm{O}$), we only keep pixels for which the brightness temperature exceeds the 3σ noise level. Otherwise the pixel value is set to zero. We split the survey data into 80% training and 20% testing subsets. However, this separation is not randomly selected. Instead, we construct the subsets so they contain two mutually exclusive regions of the sky. We maintain this separation so we may examine if the model overfits to a specific region of the Galactic center, which would impair its generalizability to other surveys. We examine this claim in Section 3.2.

2.2 Gaussian Process Regression

A Gaussian process is a stochastic model where any finite collection of observed data are jointly Gaussian with mean μ and covariance Σ. Using a finite set of observed data, we can obtain the closed-form predictive distribution over unobserved data by maximizing the multivariate Gaussian likelihood through kernel parameters ω [26]. Due to their capability to perform regression, the smoothness of their samples, and their ability to capture both local and universal correlations between data points, we employ Gaussian processes ($\mathcal{GP}$) to predict $^{13}\mathrm{CO}$ concentration.

For statistical stability, the mean of a $\mathcal{GP}$ is set to zero. A zero-mean $\mathcal{GP}$ is uniquely defined by its kernel κ. We employ a *Matern* kernel which is described below. In addition to being infinitely differentiable, leading to smooth samples in the predictions, the Matern kernel is a *stationary* kernel, meaning the pair-wise covariance of two points is only depending on their relative position. The length scale, l, controls the distance over which the $\mathcal{GP}$ interpolates between points. Due to the significant differences of pixel intensity over a small degree of sky maps, the ability to control the length of interpolation is a crucial property of the kernel function. To account for the inherent noise in MOPRA observations, a diagonal matrix $\sigma^2\mathcal{I}$ is added to model the variance throughout the observed data. The complete Gaussian process kernel consists of the Matern kernel, additional white noise, and the variance estimate.

$$\text{Matern}(x_1, x_2) = \nu_0\alpha(1+\sqrt{3}r)exp(-\sqrt{3}r);\ r = \frac{||x_1 - x_2||_2}{l} \quad ,\text{White}(x_1, x_2) = \nu_1 + \nu_2\delta_{x_1,x_2}$$

We train the $\mathcal{GP}$ on pairs of horizontal and vertical patches with a size of 9 pixels from CO map as the inputs. For each of these intersecting patches $x^h_{[i-4:i+4,j]}$ and $x^v_{[i,j-4,j+4]}$ we train separate $\mathcal{GP}$s predicting $y^h_{[i,j]}$ and $y^v_{[i,j]}$ which are the intensity of the pixel $[i, j]$ on the $^{13}\mathrm{CO}$ map. Then the final prediction for the intensity of the pixel $o_{[i,j]}$ on the $^{13}\mathrm{CO}$ map would be the mean of two separate

predictions.

$$o_{[i,j]} = \frac{y^h_{[i,j]} + y^v_{[i,j]}}{2}$$

During training, the vertical and horizontal stride size (the distance between two adjacent patches) is set to be 8. After all the prepossessing steps throughout the velocity bins, we extract the total number of 7840 horizontal and vertical pairs of patches per velocity bins. We train an independent $\mathcal{GP}$ for each velocity bin.

2.3 Patch Convolutional Neural Networks (CNNs)

We use K learned convolution filters of size $D_1 \times D_2$, operating on an entire patch, and independently for each velocity bin. We do not apply any padding the patches, and evaluate the CNN once for each patch. This results in a single latent vector with dimension K for each patch. We also include a parametric ReLU non-linearity, $PReLU(x) = \max(x, \alpha x)$, where α is a learnable parameter. This allows the network to learn a larger class of functions. Additionally, we normalized the latent vectors based on a moving average of their batch statistics [21], and we apply randomized drop-out to prevent overfitting. These latent vectors are fed through several fully-connected layers, each with their own $PReLU$ non-linearity, batch-normalization, and dropout. The resulting vector is fed through a final fully-connected layer which produces the scalar ^{13}CO concentration estimate.

This patch-based CNN may also be viewed as a single large CNN with kernel size $D_1 \times D_2$ which is applied to the entire image and returns another image. We can equate the window of this convolution layer scanning across the image to applying a small convolution layer to individual patches of size $D_1 \times D_2$. We examine the CNN in the patch interpretation so that we may directly compare its mechanism to the Gaussian process.

To effectively learn the widely varying dataset, the network is trained using either weighted mean absolute error (WMAE) or Poisson log-likelihood. Since the target ^{13}CO maps are rather sparse, we re-weight conventional absolute error so the network prioritizes correctly predicting bright regions over dim regions with a method akin to quantile regression. These weights are calculated from target maps that were smoothed with a Gaussian kernel and re-scaled to be in $[0, 1]$. We emphasize high concentrations regions by scaling their loss weight by up to $\sigma = 30$ times more than low concentration regions, and we progressively reduce this scaling, γ_t, to 1.0 throughout training in order to prevent bias. For input patches I, target values T, model f, and re-scaled epoch $t \in [0, 1]$, the losses are described as:

$$\gamma_t = \left(\sigma - \frac{\sigma}{1 + \exp\left(-10t + 5\right)}\right) (\text{Rescale}(T_i) + 1)$$

$$\mathcal{L}_{\text{WMAE}} = \frac{1}{N}\sum_{i=1}^{N} \gamma_t |(f(I_i) - T_i|$$

$$\mathcal{L}_{\text{Poisson}} = \frac{1}{N}\sum_{i=1}^{N} f(I_i) - T_i \log f(I_i)$$

We tune the network hyperparameters using the SHERPA hyperparameter optimization library [18, 19]. We tested 2000 network variations, using Gaussian Process optimization to suggest values for the learning rate, filter count, window size, and maximum scaling coefficient. We evaluated our model during hyperparameter optimization on a uniformly sampled 20% of the training data which we label as the validation dataset. Our final model has the following parameterization: 128 convolution filters with a window size of 7×7 and two fully-connected hidden layers. We evaluate the final network by training it for 200 iterations using the LAMB gradient descent optimizer [29], with learning rate of 10^{-3}, and a batch size of 8192 patches. This training takes approximately 2 hours when performing CNN calculations on four NVidia Titan X GPUs.

3 Results

We evaluate our models using several metrics. These metrics are computed independently for each source and target image plane in our testing dataset S_n and T_n. We measure the Pearson correlation

coefficient, the scaled absolute error, and a hot-spot scaled absolute error. See Figure 4 for full heatmaps across the testing dataset. We also evaluate the median values for these metrics on each model in Table 1.

When measuring relative error, conventional percent error proved difficult to interpret because our data can take on a large range of values and contains many near-zero values in sparse regions. Instead, to avoid direct division, we use a globally normalized variant which we call scaled absolute error. Here, we compare the absolute error with the maximal element of each image plane, providing a stable estimate of relative error. For the hot-spot absolute error, we normalize each velocity bin and select only locations in the image that are at least one standard deviation above the velocity bin mean. This serves as a simple marking of high-activity regions. Measuring scaled absolute error in these hot-spots informs us if the model is biased towards low or high-value regions. We notice that all models slightly degrade in performance on these hot-spot.

We notice that all of the models performed similarly on the dataset. The CNNs were best at broad prediction across the entire dataset, scoring the lowest and most consistent absolute error. However, the Matern GP performed best when predicting the bright hot-spots in the image. Since $^{13}\mathrm{CO}$ contribution for H_2 tracing is primarily in dense, high concentration regions, the effect of this disparity must be further evaluated an analysis of these models' ability to generate H_2 concentrations and, ultimately, gamma-ray emission maps. It is also important to evaluate if gamma-ray estimation requires absolute accuracy in hot-spots, or if the simulated emission are not greatly effected by absolute concentration.

3.1 Locality

During data preprocessing, we assumed that CO column densities are locally correlated. We examine this assumption by comparing the models' performance at various window sizes, independently optimizing hyperparameters at each window size. A performance comparison is presented in Figure 2. We notice that increasing the window size has a negligible effect on scaled absolute error, implying that additional long-distance information is not necessary for accurate prediction. In fact, larger window sizes decrease performance, likely because they reduce training sample count and increase the number of model parameters.

3.2 Generalizability

In order to evaluate if our models overfit to a specific region of the sky, we examine if model performance worsens further away from the training regions. When creating the training-testing data split, we ensured that the two datasets represented mutually exclusive regions. We trained on Galactic longitudes between 300° and 340°, and we tested between 340° to 350°. We examine the relationship between scaled absolute error and distance from the training region in Figure 3. We do not see a significant trend as the model predicts further away from the training region, implying good generalization in the Galactic center.

4 Discussion and Broader Impact

In this paper, we demonstrate the applicability of machine learning techniques to reconstructing the distribution of $^{13}\mathrm{CO}$, and validate it with available data. This is the first, crucial step to establishing the validity of this methodology which hinges on machine learning reliably reconstructing this emission for regions of the sky where observations are not available. This work benefits the broader astrophysical community by allowing researchers to evaluate theories in regions of the sky where expensive, accurate survey data is not yet available. The next step towards unraveling the *Fermi*–LAT excess to apply this technique to the analysis of the gamma-ray emission data from the inner Milky Way, which requires a full analysis of the *Fermi*–LAT data itself, including modeling of all components of Galactic diffuse emission.

5 Acknowledgements

The work of AS, MT, and PB is in part supported by grants NSF NRT 1633631 and ARO 76649-CS to PB. We would also like to thank Yuzo Kanomata for his computational support.

References

[1] Kevork Abazajian, Nicolas Canac, Shunsaku Horiuchi, and Manoj Kaplinghat. Astrophysical and dark matter interpretations of extended gamma ray emission from the galactic center. *Physical Review D*, 90, 02 2014.

[2] Kevork N Abazajian. The consistency of fermi-LAT observations of the galactic center with a millisecond pulsar population in the central stellar cluster. *Journal of Cosmology and Astroparticle Physics*, 2011(03):010–010, mar 2011.

[3] Markus Ackermann, Marco Ajello, WB Atwood, Luca Baldini, Jean Ballet, Guido Barbiellini, D Bastieri, K Bechtol, R Bellazzini, B Berenji, et al. Fermi-lat observations of the diffuse γ-ray emission: implications for cosmic rays and the interstellar medium. *The Astrophysical Journal*, 750(1):3, 2012.

[4] Prateek Agrawal, Brian Batell, Patrick J Fox, and Roni Harnik. Wimps at the galactic center. *Journal of Cosmology and Astroparticle Physics*, 2015(05):011, 2015.

[5] M Ajello, A Albert, WB Atwood, G Barbiellini, D Bastieri, K Bechtol, Ronaldo Bellazzini, E Bissaldi, RD Blandford, ED Bloom, et al. Fermi-lat observations of high-energy γ-ray emission toward the galactic center. *The Astrophysical Journal*, 819(1):44, 2016.

[6] P. Baldi. *Deep Learning in Science: Theory, Algorithms, and Applications*. Cambridge University Press, Cambridge, UK, 2021. In press.

[7] P Baldi, P Sadowski, and D Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5, 2014.

[8] Pierre Baldi, Kevin Bauer, Clara Eng, Peter Sadowski, and Daniel Whiteson. Jet substructure classification in high-energy physics with deep neural networks. *Physical Review D*, 93(9):094034, 2016.

[9] Pierre Baldi, Kyle Cranmer, Taylor Faucett, Peter Sadowski, and Daniel Whiteson. Parameterized neural networks for high-energy physics. *The European Physical Journal C*, 76(5):235, 2016.

[10] Catherine Braiding, Graeme F Wong, Nigel I Maxted, Donatella Romano, Michael G Burton, Rebecca Blackwell, MD Filipović, MSR Freeman, B Indermuehle, J Lau, et al. The mopra southern galactic plane co survey data release 3. *Publications of the Astronomical Society of Australia*, 35, 2018.

[11] Michael G Burton, Catherine Braiding, Christian Glueck, Paul Goldsmith, Jarryd Hawkes, David J Hollenbach, Craig Kulesa, Christopher L Martin, Jorge L Pineda, Gavin Rowell, et al. The mopra southern galactic plane co survey. *Publications of the Astronomical Society of Australia*, 30, 2013.

[12] Francesca Calore, Ilias Cholis, and Christoph Weniger. Background model systematics for the Fermi GeV excess. *arXiv preprint arXiv:1409.0042*, 2014.

[13] Laura J Chang, Siddharth Mishra-Sharma, Mariangela Lisanti, Malte Buschmann, Nicholas L Rodd, and Benjamin R Safdi. Characterizing the nature of the unresolved point sources in the galactic center: An assessment of systematic uncertainties. *Physical Review D*, 101(2):023014, 2020.

[14] James J. Condon and Scott M. Ransom. 7. spectral lines. *Essential Radio Astronomy*, page 233–276, 2016.

[15] Tansu Daylan, Douglas P Finkbeiner, Dan Hooper, Tim Linden, Stephen KN Portillo, Nicholas L Rodd, and Tracy R Slatyer. The characterization of the gamma-ray signal from the central milky way: a compelling case for annihilating dark matter. *arXiv preprint arXiv:1402.6703*, 2014.

[16] M. Fenton, A. Shmakov, T. Ho, S. Hsu, D. Whiteson, and P. Baldi. Permutationless many-jet event reconstruction with symmetry preserving attention networks. 2020. Submitted. Also arXiv:2010.09206.

[17] Lisa Goodenough and Dan Hooper. Possible evidence for dark matter annihilation in the inner milky way from the fermi gamma ray space telescope. 2009.

[18] Lars Hertel, Julian Collado, Peter Sadowski, and Pierre Baldi. Sherpa: Hyperparameter optimization for machine learning models. Machine Learning Open Source Software 2018: Sustainable communities. openreview.net, 2018.

[19] Lars Hertel, Julian Collado, Peter Sadowski, Jordan Ott, and Pierre Baldi. Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*, 2020. In press.

[20] Dan Hooper and Lisa Goodenough. Dark Matter Annihilation in The Galactic Center As Seen by the Fermi Gamma Ray Space Telescope. *Phys. Lett. B*, 697:412–428, 2011.

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org, 2015.

[22] Rebecca K Leane and Tracy R Slatyer. Dark matter strikes back at the galactic center. *arXiv preprint arXiv:1904.08430*, 2019.

[23] Samuel K. Lee, Mariangela Lisanti, Benjamin R. Safdi, Tracy R. Slatyer, and Wei Xue. Evidence for unresolved γ-ray point sources in the inner galaxy. *Phys. Rev. Lett.*, 116:051103, Feb 2016.

[24] Lingge Li, Nitish Nayak, Jianming Bian, and Pierre Baldi. Efficient neutrino oscillation parameter inference using gaussian processes. *Physical Review D*, 101(1):012001, 2020.

[25] Liszt, H. S., Pety, J., and Lucas, R. The co luminosity and co-h2 conversion factor of diffuse ism: does co emission trace dense molecular gas?*. *A&A*, 518:A45, 2010.

[26] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[27] Chase Shimmin, Peter Sadowski, Pierre Baldi, Edison Weik, Daniel Whiteson, Edward Goul, and Andreas Søgaard. Decorrelated jet substructure tagging using adversarial neural networks. *Physical Review D*, 96(7):074034, 2017.

[28] Thomas L. Wilson, Kristen Rohlfs, and Susanne Hüttemeister. *Tools of Radio Astronomy*. 2009.

[29] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes, 2019.

Figures and Tables

Metric	Equation	Matern GP	WMAE CNN	Poisson CNN
Correlation Coefficient	$\frac{\sum_{n=1}^{N}(T_n - \bar{T})(O_n - \bar{O})}{\sqrt{\sum_{n=1}^{N}(T_n - \bar{T})^2}\sqrt{\sum_{n=1}^{N}(O_n - \bar{O})^2}}$	0.84	**0.87**	0.85
Scaled Absolute Error	$\frac{1}{N}\sum_{n=1}^{N}\frac{\lvert f(I_n) - T_n\rvert}{\max_n \lvert T_n\rvert}$	0.0137	**0.0135**	0.0136
Hot-Spot Absolute Error	$\frac{1}{N}\sum_{T_n > \Sigma}\frac{\lvert f(I_n) - T_n\rvert}{\max_n \lvert T_n\rvert}$	**0.0238**	0.0257	0.0263

Table 1: A list of metrics and the median values that each method achieved on the testing datasets. These values are from the best networks produced by the Sherpa hyperparameter optimization step.

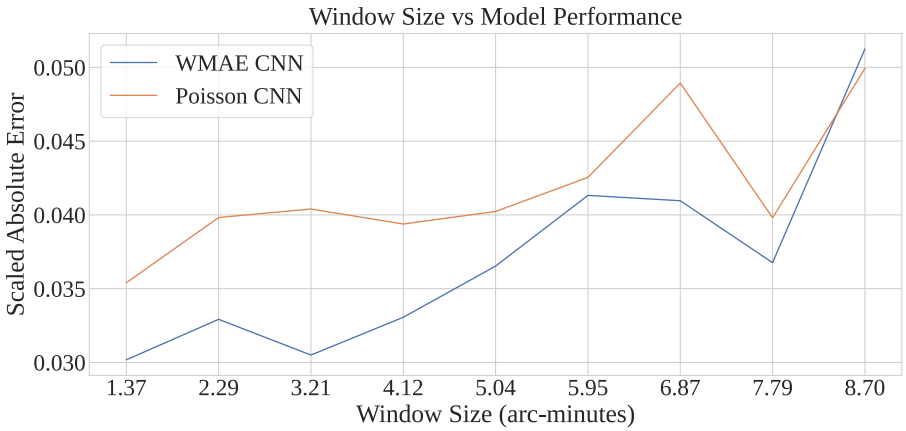

Figure 2: The mean scaled absolute error of models that were tested during hyperparameter optimization, grouped by their window size.

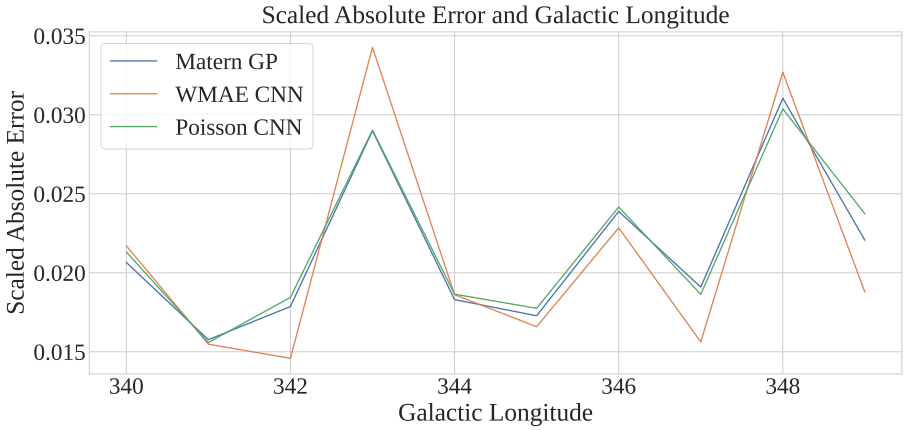

Figure 3: The mean scaled absolute error across all velocity bins, grouped by galactic longitude.

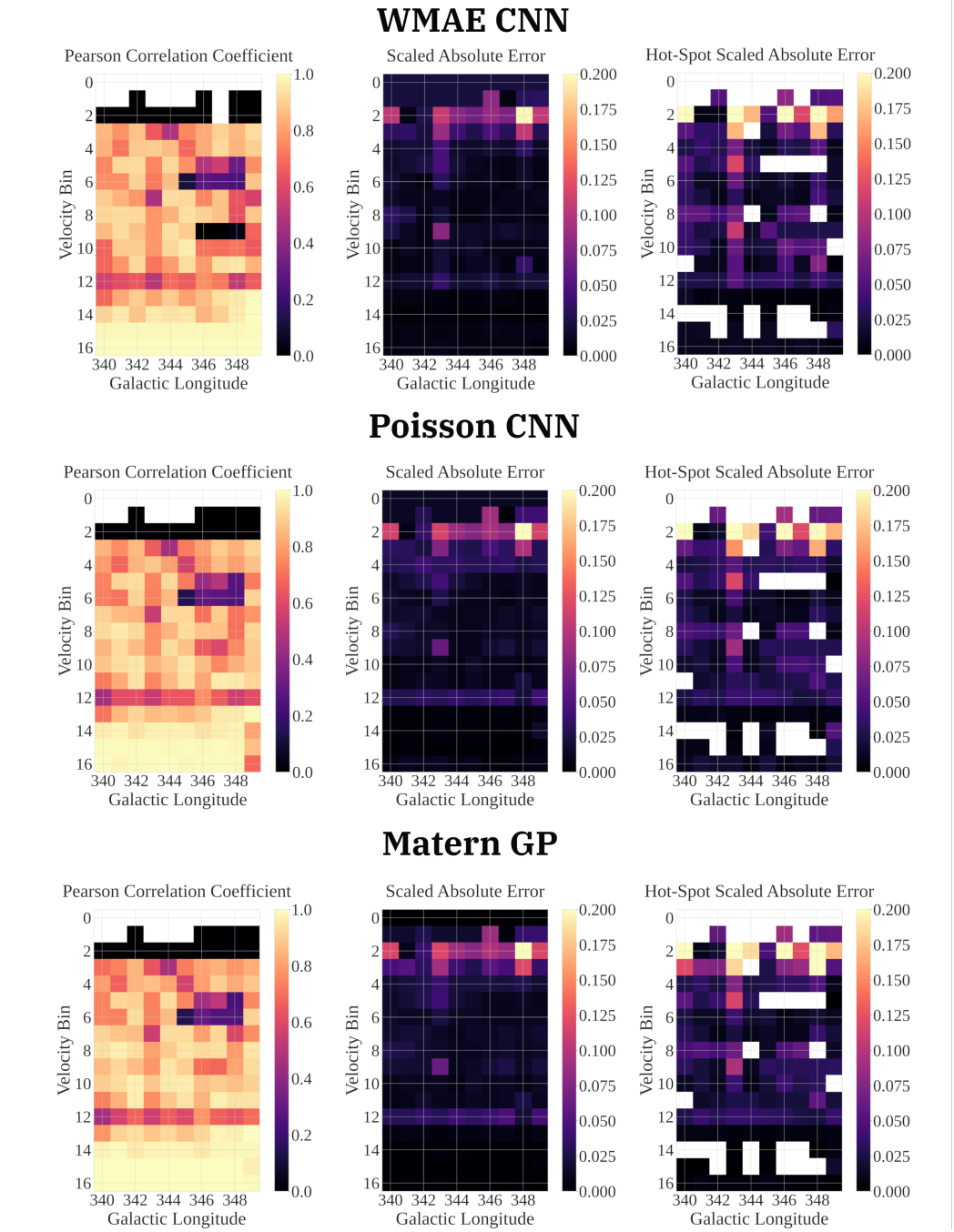

Figure 4: Heatmaps comparing metric values across every velocity and longitude bin for every model. White regions in the heatmaps indicate that the bin returned an invalid value for the metrics or there were no available hot-spots in the region.