

Predicting dust emission using galactic 21 cm data

Guangyu Zhang,^{a,b} Chi-Ting Chiang,^b Chris Sheehy,^b Anže Slosar^b and Jian Wang^a

^aState Key Laboratory of Particle Detection and Electronics, University of Science and Technology of China, Hefei 230026, China

^bPhysics Department, Brookhaven National Laboratory, Upton, NY 11973, USA

E-mail: zguangyu@mail.ustc.edu.cn, ctchiang@bnl.gov, csheehy@bnl.gov, anze@bnl.gov, wangjian@ustc.edu.cn

Abstract. Understanding large-angular-scale galactic foregrounds is crucial for future CMB experiments aiming to detect B -mode polarization from primordial gravitational waves. Traditionally, the dust component has been separated using its different frequency dependence. However, using non-CMB observations has potential to increase fidelity and decrease the reconstruction noise. In this exploratory paper we investigate the capability of galactic 21 cm observations to predict the dust foreground in intensity. We train a neural network to predict the dust foreground as measured by the Planck Satellite from the full velocity data-cube of galactic 21 cm emission as measured by the HI4PI survey. We demonstrate that information in the velocity structure clearly improves the predictive power over both a simple integrated emission model and a simple linear model. The improvement is significant at arc-minute scales but more modest at degree scales. This proof of principle on temperature data indicates that it might also be possible to improve foreground polarization templates from the same input data.

Contents

1	Introduction	1
2	Data and method	3
2.1	Integrated intensity model	5
2.2	Linear combination model	5
2.3	Deep neural network	5
3	Results	7
4	Conclusions	10
A	Optimal weights for the linear model	11
B	Effect of λ on the correlation between the predicted and target maps	13
C	Result for east and west galactic hemisphere division	13

1 Introduction

Modeling and understanding the foregrounds contaminating Cosmic Microwave Background (CMB) maps is crucial for deriving robust cosmological constraints from them. So far, the two main approaches employed in reducing the impact of unwanted foregrounds on the CMB maps rely on i) minimizing the problem by choosing the cleanest parts of the sky and ii) projecting the unwanted foreground components by the different spectral indices of various components in the microwave sky to build maps of foregrounds.

As we are approaching the forthcoming generation of the CMB experiments, exploring other possible techniques can be beneficial. This is particularly important for the future measurements that aim to constrain the presence of tensor modes in the primordial fluctuations, both for the ground-based observations focusing on small areas of the sky (such as Simons Observatory [1] and CMB-S4 [2]) and the satellite missions covering the full sky (such as PICO [3], and LiteBIRD [4]). Many of the future experiments will be incredibly deep, with sensitivities below $1 \mu\text{K arcmin}$, an over two order of magnitude increase over the depth of Planck Satellite maps [5]. Therefore, a foreground that might have been completely negligible at the Planck sensitivity level might suddenly be crucial. If these future observations will be performed from the ground, the situation is even more difficult as the range of frequency bands suitable for observation is considerably smaller than in the space. Finally, the polarization measurements have inherently more degrees of freedom as both the polarization intensity as well as the orientation can, in principle, change with frequency, an effect known as decorrelation [6, 7]. Therefore it would be helpful to develop additional methods for understanding foregrounds.

The most important foregrounds at frequencies of interest from the ground are the synchrotron and dust. The former has a strongly falling spectrum and becomes very small at frequencies beyond ~ 70 GHz. The latter has a rising spectrum and starts to dominate at frequencies higher than ~ 150 GHz. While the CMB fluctuations peak at around 220 GHz,

the global foreground minimum is at lower frequencies of 70 – 100 GHz [8]. In the CMB data analysis, the raw temperature maps of the sky are first cleaned using one of the component separation techniques (see e.g. Ref. [9] and reference there-in). The resulting CMB maps often contain potentially significant levels of residual foreground contamination. If a template for any such contamination is known, it can be exactly marginalized in the estimation of the power spectrum, a technique called *template marginalization*. In optimal quadratic estimation approaches, this is achieved by assigning infinite variance to the linear combinations of pixels corresponding to the template (see e.g. Ref. [10]). In pseudo- C_ℓ approaches the same effect can be achieved by fitting and subtracting the template and then correcting for the small bias this introduces in the power spectrum measurement (see e.g. Ref. [11]). In both cases, the more faithful the template map is, the better the results, but even with imperfect templates, the biases in estimated power spectra can be reduced to perhaps satisfactory levels. In particular, it is also possible to marginalize over several templates, which in effect models the true foreground contamination as some linear combination of those templates. As the number of templates increases, the statistical power is being lost, but template marginalization cannot introduce a bias, *unless* template correlates with the CMB signal. Therefore, it is useful to consider methods that predict foregrounds based on data that cannot be correlated with the true CMB signal. This lead us to consider the galactic 21 cm data.

The 21 cm line is the hyperfine splitting of the ground level of neutral hydrogen which arises due to alignment (or not) of the spins of electron and proton. Galactic 21 cm emission traces the neutral hydrogen in our own galaxy. Since various components of the milky way (stars, dust, neutral and ionized hydrogen, etc.) trace each other, it is reasonable to expect that 21 cm measurements might be useful for predicting the dust foreground contaminant. This is particularly attractive since the 21 cm maps are measured at frequencies which are orders of magnitude away from frequencies relevant for observations of CMB fluctuations. We can therefore be absolutely sure that these maps are blind to any structure in the CMB fluctuations. Note that this is not true when one attempts to generate foreground maps relying purely on the different spectral indices.

We are not the first ones to consider this idea. Several papers have discussed how dust is traced by the integrated intensity in the 21 cm observations. Already in 1955, Ref. [12] has noticed the association of dust and galactic 21 cm line. More recently, Ref. [13] has shown that the integrated intensity in galactic 21 cm data correlates with the dust polarization angle. However, the 21 cm dataset is considerably richer: the line is well resolved by modern surveys giving the full velocity structure of the hydrogen along the line-of-sight – at each spatial position one can measure the full profile of the velocity line. It is entirely plausible that the full information about the velocity structure along the line-of-sight encodes information about the physical environment in which the neutral hydrogen exists and can therefore be useful in inferring the dust polarization angle [14–16]. In this paper we explore this possibility in a way that does not assume any concrete physical model but uses existing data to see if dust can be predicted from many maps corresponding to velocity slices in 21 cm measurement using modern machine-learning techniques. As a first exploratory work we perform this using intensity alone, i.e. predicting intensity of dust emission from the full velocity cube of 21 cm data. The main result is that there indeed seems to be more information the velocity slices beyond that present in the integrated intensity map and which cannot be modeled by a naive linear model of the type which is traditionally utilized. This proof of concept anticipates future work involving polarization data, but it could also be useful itself for inferring the true cosmic infrared background.

The question remains whether this method is useful at all. If one needs the traditional component separation in order to train the neural network on one patch of the sky, then one might expect that this correlation will never be useful as long as traditional component separation is available. While we do not yet give a definite method for doing that, the general point remains that if more information is available, it can only be beneficial. For example, one could imagine a fitting procedure that uses multi-frequency observations of the CMB sky, together with galactic 21 cm maps to simultaneously solve for individual component maps *and* the way these individual maps map to the data vector: spectral indices and their variations, but also the nonlinear and somewhat non-local relation between microwave and 21 cm dust emission. It might also be possible, for example, to train on a small patch of extremely deep data microwave data and use this training where only deep 21 cm maps are available (in general, these are signal dominated everywhere). The main point of this paper is to show the existence of extra information.

The paper is structured as follows. In section 2 we present the data used in this work, the models employed and general caveats of the procedure. In section 3 we show the result and the improvements that can be made in dust predictions over what the integrated intensity predicts and finally conclude in section 4. In appendix A, we demonstrate the analytical calculation for the optimal weights of the linear model. In appendix B we study the effect of λ values on the correlation between the predicted and target maps. In appendix C we present the results for an additional sky division for constructing the training and test datasets.

2 Data and method

The main goal of this work is to investigate the capability of machine-learning methods to predict the dust intensity from the galactic 21 cm measurements. For the target dust map we use the publicly available `COMMANDER` dust component map¹ with the resolution of `Nside=512` produced by the Planck Satellite team using component separation [9]. For the 21 cm data we take the full-sky measurements of the 21 cm emission from our galaxy as measured by the HI4PI survey [17]. HI4PI survey was constructed from the Effelsberg-Bonn HI Survey [18] and the Parkes Galactic All-Sky Survey [19]. The survey has an angular resolution of 16.2 arcmin and a sensitivity of 43 mK, surpassing the legacy HI dataset from the Leiden/Argentine/Bonn Survey [20].

Because the standard neural network libraries are optimized to work with data that are sampled on a Cartesian grid but not on the spherical coordinates, we use `HEALPix`² [21] to gnomonically project 21 cm and dust all-sky maps to square cutouts. In particular, we divide the data into the north galactic hemisphere used for training the model and the south galactic hemisphere used for measuring the effectiveness of the model. To ensure no overlap between the training and test data sets, we do not use projection centers with declination below 17.65 degrees.

In detail, we process the data as follows:

- We start with full resolution maps that are pixelized on an `Nside=1024` `HEALPix` map.

¹<https://pla.esac.esa.int>

²<https://healpix.sourceforge.net>

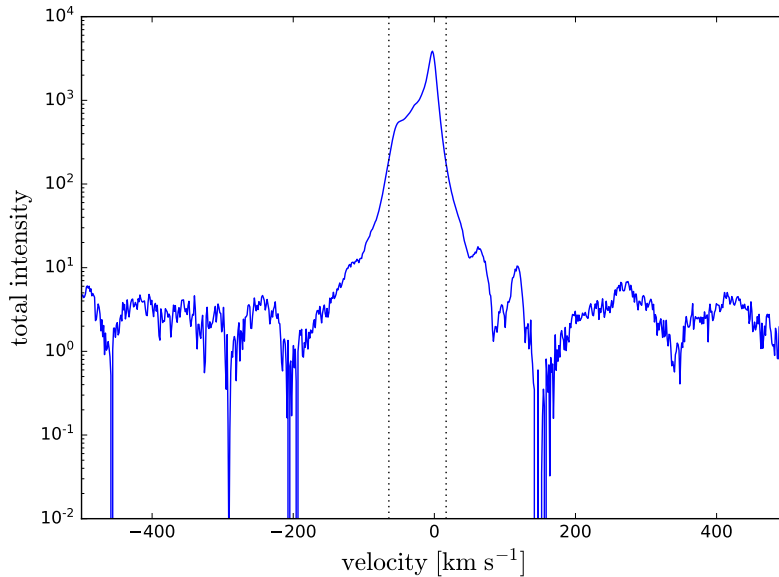


Figure 1. Total intensity of 10000 random pointings (excluding areas covered by the mask) for maps of all velocity slices. The vertical dotted lines show the range of -64.5 km s^{-1} and 17 km s^{-1} , which contains most of the 21 cm signal.

- For the 21 cm galactic data, we use data in the velocity range between -64.5 km s^{-1} and 17 km s^{-1} because it contains most of the 21 cm signal (see figure 1). This leaves us with 64 individual maps I_ν .³
- We smooth data by applying a butterworth filter in ℓ domain with resolution $\ell < 460$. This is to prevent aliasing of small-scale power into large-scale power, but we do note that is a lossy process that could be sacrificing real information.
- We apply the publicly available LR72 mask used in Ref. [7] on the dust and 21 cm maps for an effective sky fraction of 72%.
- We choose random cutout centers and rotations in the training and test datasets and project the input data onto a 64×64 Cartesian grid with the pixel size of 23.4 arcmin using a gnomonic projection routine in the HEALPix package. Such cutouts have a side length of 24.96 degrees, corresponding to a fundamental mode expressed in units of standard spherical harmonic ℓ of $\ell_F = 14.4$. To avoid regions largely covered by the mask, we discard the cutouts if the mask covers more than 70% of the projections. Since we work with projected flat-space maps of a small area, the spherical harmonic decomposition can be trivially obtained using a two-dimension fast Fourier transform [22]. We prepare the data so that the declination of the cutout centers follows $\cos(\text{DEC})$ distribution to avoid oversampling the high-declination sky.

The result of this process is 50000 sets of maps from the north galactic hemisphere for the training data and 1000 sets of maps from the south galactic hemisphere for the test data.

³To further test the choice of the velocity range, we also train the neural network with data in the range of $\pm 120 \text{ km s}^{-1}$, but we do not find a significant difference in the predicted correlation coefficients on all scales.

Each set contains a target dust map T and 64 21 cm maps I_i applied with the same cutout and mask. Since the cutout centers and rotations are randomly chosen, the 50000 maps are not completely independent and some of them can overlap by chance. However, this should not be an issue for training because the overlapping pixels will not appear at identical locations. If the neural network just memorizes the relation between dust and 21 cm data for certain pixels, it will not be able to generalize when these pixels show up in different locations for different cutouts, and we do not observe such a problem.

We consider various predictive models that produce \tilde{T} given input maps I_i . Our success is characterized by the cross-correlation coefficient given by

$$r(\ell) = \frac{C_{T\tilde{T}}(\ell)}{\sqrt{C_{TT}(\ell)C_{\tilde{T}\tilde{T}}(\ell)}}. \quad (2.1)$$

We compute the cross-correlation coefficients on individual Fourier modes first and then bin the data into ℓ with bin size ℓ_F . When compressed into a single number, we average the cross-correlation coefficients of individual Fourier modes over two bins: $\ell_F - 100$ and $100 - 200$, denoted respectively as r_{50} and r_{150} . Note that r is invariant under the total amplitude rescaling of \tilde{T} and hence we ignore any overall normalization factor. A cross-correlation coefficient of 1 corresponds to perfect prediction $\tilde{T} = T$ on a given scale, while 0 a map that is completely uncorrelated. In principle $r < 0$ is possible, but these values correspond to a simple sign change.

We have attempted several models to predict \tilde{T} . In order of increasing complexity, they are described below:

2.1 Integrated intensity model

Our baseline and simplest model is that the integrated 21 cm intensity traces the dust emission. That is, $\tilde{T} = \sum_i I_i$. Other models will be judged by whether they offer (or not) improvement over this simplest model.

2.2 Linear combination model

In the linear combination model, we model the target dust map as a simple linear combination of input 21 cm maps with a scale-independent weight in Fourier space

$$\tilde{T} = \sum_i w_i I_i, \quad (2.2)$$

where the index i runs from 1 to 64 for different velocity slices. Note that the integrated intensity model is a special case of the linear combination model with $w_i = 1$. Appendix A outlines the procedure for optimizing the linear weights given the training data. In principle, the linear weights can be “scale dependent”, i.e. a distinct set of linear weights for each angular scale ℓ . However, our numerical experiment has shown that scale-dependent weights are prone to overfitting. Namely, the results on *test* data get worse as the number of weights increases, indicating an overfitting problem. Therefore, in this paper we restrict ourselves to scale-independent weights.

2.3 Deep neural network

Our most advanced predictive model uses a deep neural network [23]. Deep neural networks are a subset of machine learning algorithms used for solving a wide range of problems, such as

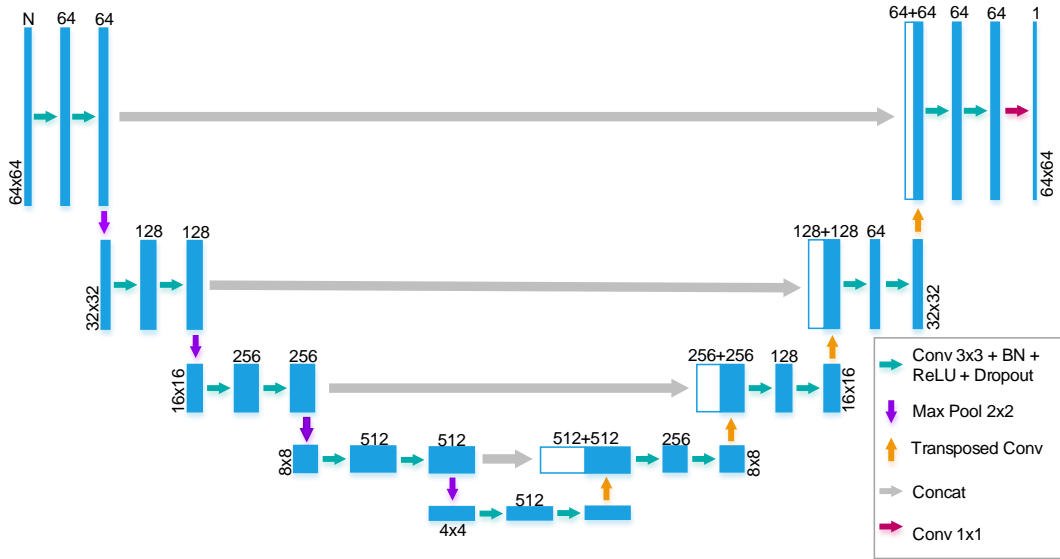


Figure 2. The architecture of the deep neural network, U-Net, adopted in this paper. The green right arrow contains a convolution layer with a 3×3 kernel, stride 1 and same padding, a batch normalization layer, a rectified linear unit for activation, and a dropout layer with dropout rate of 0.3; the purple down arrow represents a maximum pooling of 2×2 kernel and stride 2; the gray right arrow represents concatenation of the current and the previous downward part outputs; the yellow up arrow represents the transposed convolution with 2×2 kernel and stride 2.

image recognition, machine translation and image generation. A deep neural network which consists of convolution layers is also called a convolutional neural network (CNN) [24, 25]. A convolution layer is some set of $N \times N$ kernel applied to the input. The kernel parameters are adjusted by the back propagation algorithm during training. It can effectively extract useful features from an image.

Our neural network structure is based on the U-Net model [26]. U-Net is an image-to-image network, which takes images as input and also outputs images. U-Net was designed to perform image segmentation tasks. The task in this project is similar to image segmentation, i.e. both generating pixel-to-pixel maps. Since this model has no fully connected layer, it can fit different input image sizes using the same model.

Figure 2 shows the U-Net architecture with the 64×64 input images adopted in the paper. The U-Net has a downward part and an upward part. The downward part consists of four sets of double convolution operations and a maximum pooling (purple down arrow) with a 2×2 kernel and stride 2. Each convolution operation (green right arrow) contains a convolution layer with a 3×3 kernel, stride 1 and same padding, a batch normalization layer [27], a rectified linear unit for activation, and a dropout layer with dropout rate of 0.3 to regularize the network [28]. The upward part concatenates the current and previous downward part outputs (gray right arrow) and then uses transposed convolution [29] with a 2×2 kernel and stride 2 (yellow up arrow) to recover the original image size. The concatenating operation keeps the pixel information along with learned features. The output of the U-Net is images of 64×64 pixels, so they can be directly compared with the target dust maps.

We define the loss function of the neural network to be

$$L(T, \tilde{T}) = \lambda [(1 - r_{50}) + (1 - r_{150})] + \frac{1}{N_{\text{pixel}}} \sum_{a \in \text{pixel}} |T_a - \tilde{T}_a|, \quad (2.3)$$

where T and \tilde{T} are respectively the true and predicted maps and λ is a hyper-parameter controlling the relative contributions from the two terms of the loss function. The first term is the mean cross-correlation coefficients (defined in eq. 2.1) at angular scale bins $\ell = \ell_F - 100$ and $\ell = 100 - 200$. Since $|r(\ell)| \leq 1$, minimization of $L(T, \tilde{T})$ leads to $r_{50} \rightarrow 1$ and $r_{150} \rightarrow 1$, hence a higher correlation between T and \tilde{T} . The second term of the loss function is the mean pixel-by-pixel difference between T and \tilde{T} . The pixel difference is quantified by the L1 loss instead of the more popular L2 loss so that the model will not be too sensitive to the outliers and bright spots, which occur when the point sources in the dust and 21 cm maps are not masked out. Though our primary goal is to make the predicted maps be as highly correlated with the target maps as possible, i.e. minimization of the first term, we include the pixel difference in the loss function because the cross-correlation coefficient is invariant under a total amplitude rescaling of \tilde{T} . This helps break the degeneracy of the loss function for the neural network and the training is more stable. In appendix B we study the effect of λ values on the correlation between the predicted and target maps. We find that for $\lambda \geq 30$, the changes of the correlation is not significant on all scales. In this paper we set $\lambda = 30$ so that the two loss terms have similar amplitudes (observed during training).

We train the model for 40 epochs, with a batch size of 64 so each epoch has 782 iterations. The learning rate is set to 10^{-3} in the beginning of the training, and is reduced by a factor of 10 for every 10 epochs.

3 Results

Figure 3 shows a set of example cutout maps from the test dataset. The top left, top right, bottom left, and bottom right panels show respectively the target dust map, the integrated intensity model map, the linear combination model map, and the U-Net predicted map. The darker dots are the results of the mask. Since the normalization of the integrated intensity map is different from the other maps, for a better comparison we normalize the maps using their maximum pixel values. It is immediately striking how much better the neural network prediction is compared to naive summing over all frequencies. However, we note that visual comparison can often be very deceiving – a human eye is trained to pick individual features, which are dominated by high spatial frequency features. For scientific interest we are most interested in recovery of large-scale smooth components – these are the scales most relevant for the cosmological measurement such as tensor modes where we believe this method could be most useful (once adapted for polarization).

To better compare the performance between various models, we compute the cross-correlation coefficients between the target dust maps and maps predicted by different models. The left and right panels of figure 4 show a realization-by-realization comparison from the test set of r_{50} and r_{150} , respectively. The x -axis displays the cross-correlation coefficients of the integrated intensity model, which we use as a reference to judge whether the other models provide improvement. The y -axis displays the cross-correlation coefficients of the other models: the red circles are for the linear combination model and the blue triangles are for the U-Net model. We first notice that the linear combination model with the optimal weights computed from the training set provides similar performance as the integrated intensity model

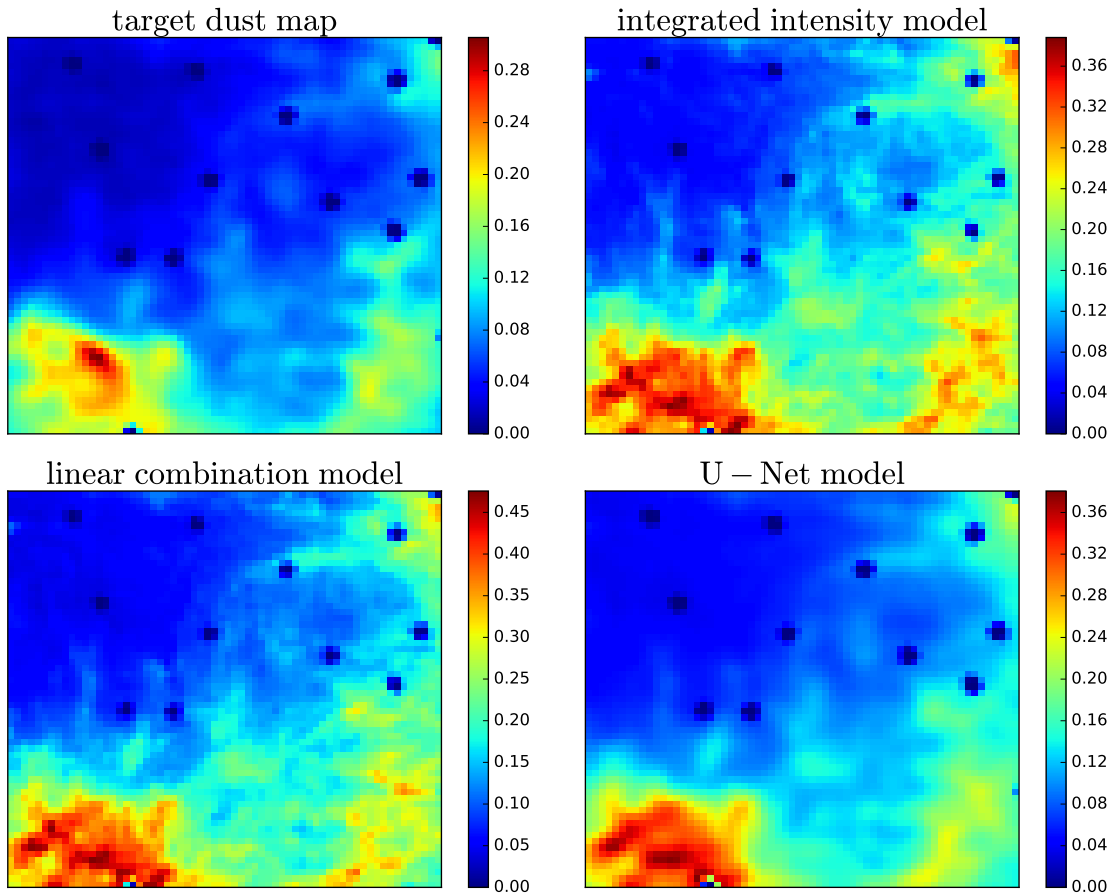


Figure 3. Example maps of one cutout from the test set. (Top left) target dust map; (top right) integrated intensity model map; (bottom left) linear combination model map; (bottom right) U-Net model map. The darker dots are the results of the mask. Since the normalization of the integrated intensity map is different from the other maps, for a better comparison we normalize the maps using their maximum pixel values. The ability of U-Net model to improve the integrated intensity map on general morphological features in the map is striking.

for both r_{50} and r_{150} on the test set. This indicates that a linear transformation of the 21 cm data is insufficient to outperform the integrated intensity model. The U-Net model provides a nonlinear transformation of the input 21 cm data. We find that on both large ($\ell_F \leq \ell < 100$) and small ($100 \leq \ell < 200$) scales the U-Net model outperforms the integrated intensity model for most of the realizations, and the improvement is more apparent on small scale. Specifically, out of 1000 realizations, U-Net outperforms the integrated intensity model in 764 and 889 realizations for r_{50} and r_{150} , respectively. While the improvement is modest, it demonstrates that the neural network indeed learns non-trivial features and transformation of the map compared to the linear combination model.

To better examine the detailed performance of different models, in figure 5 we plot the mean cross-correlation coefficient between the target dust maps and different models as a function of ℓ for training (left panel) and test (right panel) sets. The red solid, green dashed, and blue dot-dashed lines show the results for integrated intensity, linear combination, and U-Net models, respectively. The bands show the errors on the means. For the test set,

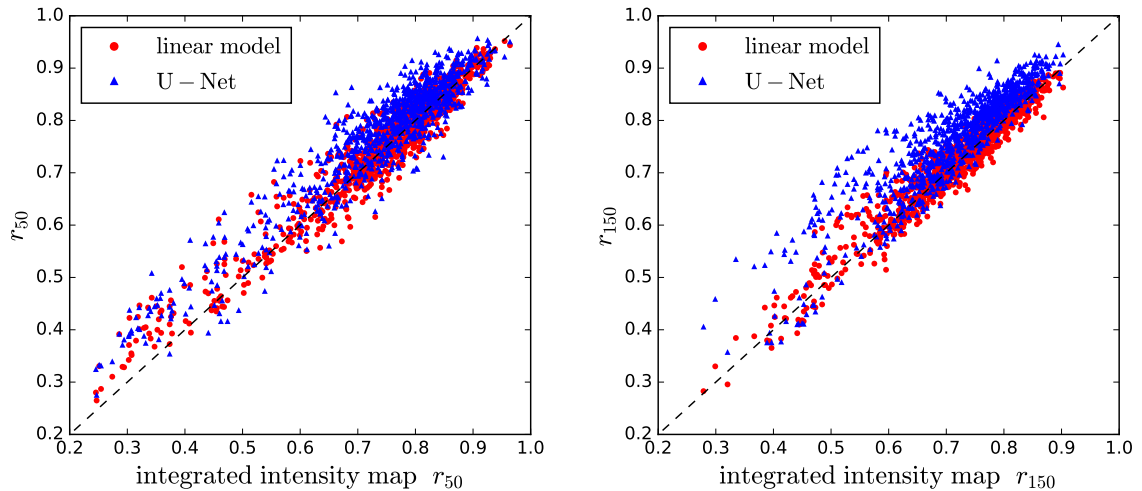


Figure 4. Realization-by-realization comparison of r_{50} (left) and r_{150} (right) for the test set. The x -axis displays the cross-correlation coefficients between the target dust maps and the integrated intensity model maps, which we use as a reference to judge whether the other models provide improvement. The y -axis displays the cross-correlation coefficients between the target dust maps and the other models: the red circle is for the linear combination model and the blue triangle is for the U-Net model. Points above the dashed line indicate improvement, while points below correspond to deterioration with respect to the integrated intensity prediction. Out of 1000 realizations, U-Net outperforms the integrated intensity model in 764 and 889 realizations for r_{50} and r_{150} , respectively.

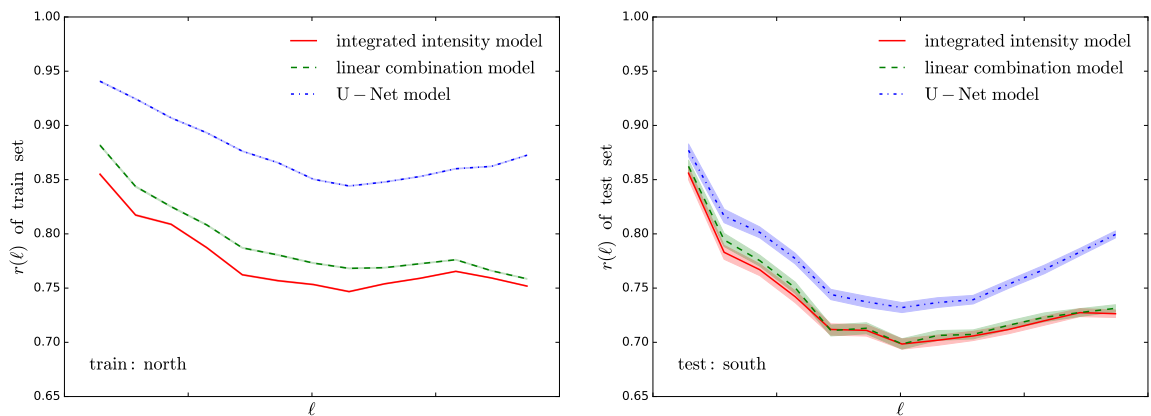


Figure 5. Mean cross-correlation coefficient between the target dust maps and different models as a function of ℓ for training (left) and test (right) sets. The red solid, green dashed, and blue dot-dashed lines show the results for integrated intensity, linear combination, and U-Net models, respectively. The bands show the errors on the means.

the conclusion is the same as figure 4. Namely, the linear combination model has a similar performance as the integrated intensity model on all scales, whereas the U-Net model on average has a better performance on all scales. Between training and test sets, we find that the linear and U-Net models perform better for the training than the test set, and the difference is larger for the U-Net model. This is a common outcome for training a neural network and indicates modest overfitting. In particular, the performance of the U-Net model worsens more

strongly on large than on small scales. This is probably due to the fact that there are fewer large-scale modes and so it is easier to overfit on large scales. Another interesting finding is that the integrated intensity model performs worse for the test than the training set. We note that there are no parameters to be fitted for the integrated intensity model, so there should be no overfitting issue. In appendix C, we investigate this issue further by dividing the sky into east and west galactic hemispheres, and train and test on east and west hemispheres, respectively. However, we do not find significant difference of the integrated intensity model. Hence, we conclude that the difference of the integrated intensity model performance between training (north) and test (south) sets shown in figure 5 is likely a statistical fluctuation.

4 Conclusions

This paper describes the principle and the feasibility of predicting the microwave dust signal using the 21 cm hydrogen line signal. We compare three models to do this work, an integrated intensity model, a linear combination model, and a neural network model. The neural network model shows the best performance on average but has a larger scatter than the linear combination model.

More precisely, the neural network offers a modest improvement on all scales, but the improvement is more apparent on small ($100 \leq \ell < 200$) than on large scales ($\ell_F \leq \ell < 100$). One possible explanation is that there are fewer modes on large scales, hence it is more likely to overfit. Nevertheless, based on the fact that the linear combination model on the full velocity slices of 21 cm data cannot outperform the neural network, we conclude that the neural network is indeed picking non-trivial features and transformation of the maps.

We have also attempted to train a neural network with moment maps (i.e. maps of the type $\sum \nu^m I_\nu$) rather than full velocity cubes, but this do not result in any improvements over the integrated intensity map and hence we do not present those results. The same is true using the integrated intensity map as the input to the neural network, indicating that the resulting map is not just a sufficiently nonlinear transformation of the integrated intensity map. However, it is possible that a different network architecture coupled with preprocessing of the velocity cube could result in improvements to the presented method.

The next step would be to apply this method to the polarization. Polarization of dust is result of an alignment of dust particles with the magnetic field. The distribution of the magnetic fields is complex but again related to the morphology of matter distribution in our galaxy. It has been shown that the integrated intensity maps processed with a rolling Hough transform [30] correlate significantly with the dust polarization angle [13, 16]. However, the use of Hough transform in that work was purely phenomenological. Moreover, the full velocity cube can separate distinct streams of neutral hydrogen that appear blended in projection and potential contain more information. Hence, a natural extension of Refs. [13, 16] would be to replace the integrated intensity maps with the full velocity cube and the Hough transform with a neural network in order to *learn* the necessary transformation. We leave this for future work.

Acknowledgments

Results in this paper were obtained using the Institutional Cluster in Scientific Data and Computing Center at the Brookhaven National Laboratory under proposal 304400. Some of the results in this paper have been derived using the HEALPix package.

A Optimal weights for the linear model

Generally, the scale-dependent linear weights in Fourier space can be written as

$$\tilde{T}_\ell = \sum_i w_{\ell,i} I_{\ell,i}. \quad (\text{A.1})$$

To solve the weights $w_{\ell,i}$, we minimize

$$\chi_\ell^2 = \sum_{n=1}^N \left(T_\ell^n - \tilde{T}_\ell^n \right)^2 = \sum_{n=1}^N \left(T_\ell^n - \sum_i w_{\ell,i} I_{\ell,i}^n \right)^2, \quad (\text{A.2})$$

where the superscript n denotes the training set index and N is the number of training set. Requiring the derivative on χ_ℓ^2 with respect to $w_{\ell,j}$ to be zero, we obtain

$$-2 \sum_{n=1}^N I_{\ell,j}^n \left(T_\ell^n - \sum_i w_{\ell,i} I_{\ell,i}^n \right) = 0, \quad (\text{A.3})$$

which then leads to

$$\sum_{n=1}^N \sum_i w_{\ell,i} I_{\ell,i}^n I_{\ell,j}^n = \sum_{n=1}^N T_\ell^n I_{\ell,j}^n. \quad (\text{A.4})$$

The above equation is a set of linear equations, and the solutions can be found by performing one matrix multiplication. Specifically, if we define the mean power spectra

$$\bar{C}_{I_i I_j}(\ell) = \frac{1}{N} \sum_{n=1}^N I_{\ell,i}^n I_{\ell,j}^n, \quad \bar{C}_{T I_j}(\ell) = \frac{1}{N} \sum_{n=1}^N T_\ell^n I_{\ell,j}^n, \quad (\text{A.5})$$

we have

$$\vec{w}_\ell = [\bar{\mathbf{C}}_{I_i I_j}(\ell)]^{-1} \vec{\bar{C}}_{T I}(\ell), \quad (\text{A.6})$$

where \vec{w}_ℓ and $\vec{\bar{C}}_{T I}$ are vectors of 64 components and $\bar{\mathbf{C}}_{I_i I_j}$ is a 64×64 matrix.

Mathematically, one can show that the solution minimizing the difference between T and \tilde{T} (equation (A.6)) also maximize the correlation coefficient of the mean power spectrum

$$\bar{r}(\ell) = \frac{\bar{C}_{T \tilde{T}}(\ell)}{\sqrt{\bar{C}_{T T}(\ell) \bar{C}_{\tilde{T} \tilde{T}}(\ell)}}, \quad (\text{A.7})$$

where

$$\begin{aligned} \bar{C}_{T T}(\ell) &= \frac{1}{N} \sum_{n=1}^N T_\ell^n T_\ell^n, \\ \bar{C}_{T \tilde{T}}(\ell) &= \frac{1}{N} \sum_{n=1}^N T_\ell^n \tilde{T}_\ell^n = \frac{1}{N} \sum_{n=1}^N T_\ell^n \sum_i w_{\ell,i} I_{\ell,i}^n = \sum_i w_{\ell,i} \bar{C}_{T I_i}(\ell), \\ \bar{C}_{\tilde{T} \tilde{T}}(\ell) &= \frac{1}{N} \sum_{n=1}^N \tilde{T}_\ell^n \tilde{T}_\ell^n = \frac{1}{N} \sum_{n=1}^N \sum_{ij} w_{\ell,i} w_{\ell,j} I_{\ell,i}^n I_{\ell,j}^n = \sum_{ij} w_{\ell,i} w_{\ell,j} \bar{C}_{I_i I_j}(\ell). \end{aligned} \quad (\text{A.8})$$

Specifically, taking the derivative on $\bar{r}(\ell)$ with respect to $w_{\ell,k}$, we obtain

$$\frac{\partial \bar{r}(\ell)}{\partial w_{\ell,k}} = \frac{\frac{\partial \bar{C}_{T\tilde{T}}(\ell)}{\partial w_{\ell,k}} \bar{C}_{\tilde{T}\tilde{T}}(\ell) - \frac{1}{2} \frac{\partial \bar{C}_{\tilde{T}\tilde{T}}(\ell)}{\partial w_{\ell,k}} \bar{C}_{T\tilde{T}}(\ell)}{\sqrt{\bar{C}_{TT}(\ell) \bar{C}_{\tilde{T}\tilde{T}}^3(\ell)}}, \quad (\text{A.9})$$

where

$$\begin{aligned} \frac{\partial \bar{C}_{T\tilde{T}}(\ell)}{\partial w_{\ell,k}} &= \frac{1}{N} \sum_{n=1}^N T_\ell^n I_{\ell,k}^n = \bar{C}_{TI_k}, \\ \frac{\partial \bar{C}_{\tilde{T}\tilde{T}}(\ell)}{\partial w_{\ell,k}} &= \frac{1}{N} \sum_{n=1}^N 2 \sum_i w_{\ell,i} I_{\ell,i}^n I_{\ell,k}^n = 2 \sum_i w_{\ell,i} \bar{C}_{I_i I_k}. \end{aligned} \quad (\text{A.10})$$

Combining the above results, the numerator in equation (A.9) becomes

$$\sum_{ij} w_{\ell,i} w_{\ell,j} [\bar{C}_{TI_k} \bar{C}_{I_i I_j}(\ell) - \bar{C}_{I_i I_k} \bar{C}_{TI_j}(\ell)]. \quad (\text{A.11})$$

Plugging in equation (A.6), which is equivalent to equation (A.4) and

$$\sum_i w_{\ell,i} \bar{C}_{I_i I_j} = \bar{C}_{TI_j}, \quad (\text{A.12})$$

into the above equation, we obtain

$$\begin{aligned} &\sum_{ij} w_{\ell,i} w_{\ell,j} [\bar{C}_{TI_k}(\ell) \bar{C}_{I_i I_j}(\ell) - \bar{C}_{I_i I_k}(\ell) \bar{C}_{TI_j}(\ell)] \\ &= \sum_j w_{\ell,j} [\bar{C}_{TI_k}(\ell) \bar{C}_{TI_j}(\ell) - \bar{C}_{TI_k}(\ell) \bar{C}_{TI_j}(\ell)] = 0. \end{aligned} \quad (\text{A.13})$$

Since the solution that minimizes the difference between T and \tilde{T} also makes $\partial \bar{r}(\ell)/\partial w_{\ell,k} = 0$, it also maximizes the cross-correlation coefficient of the mean power spectrum. However, this does not guarantee that it maximizes the mean cross-correlation of the power spectra

$$\frac{1}{N} \sum_{n=1}^N \frac{C_{T\tilde{T}}^n(\ell)}{\sqrt{C_{TT}^n(\ell) C_{\tilde{T}\tilde{T}}^n(\ell)}}. \quad (\text{A.14})$$

The above calculation assumes that the linear weights are scale dependent, i.e. there is one set of weights for each ℓ bin. One can generalize the above derivation to the *scale-independent* weights as

$$\tilde{T} = \sum_i w_i I_i. \quad (\text{A.15})$$

In other words, there is only one set of weights for all ℓ bins. In practice we find that the scale-independent weights suffer less overfitting issue than the scale-dependent weights, hence in this paper we shall only present the results of the linear combination model using the scale-independent weights.

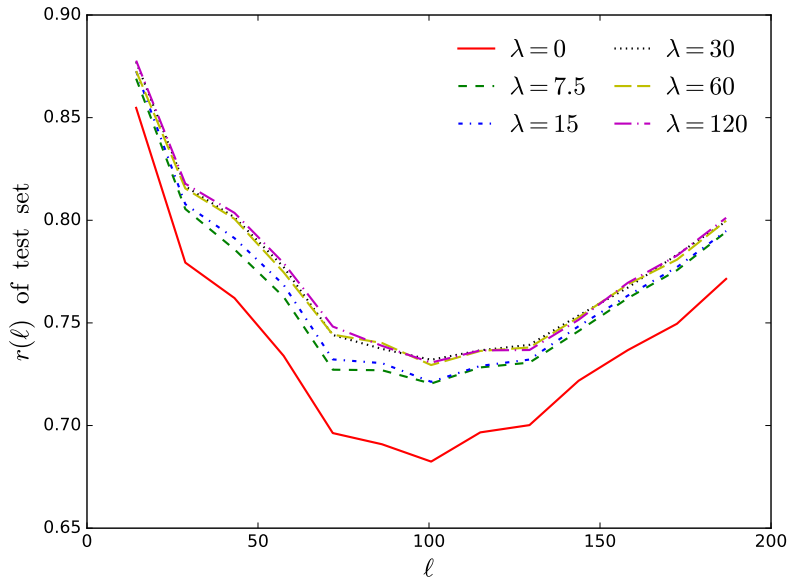


Figure 6. Mean cross-correlation coefficient between the target dust maps and different models as a function of ℓ for the test set for various λ values (shown by different colors and line styles).

B Effect of λ on the correlation between the predicted and target maps

The loss function for training the neural network is defined in eq. 2.3 and λ controls the relative contributions between the cross-correlation and the pixel difference. To study how the value of λ affects the results of the neural network, we train the U-Net model for different λ from 0 to 120. Figure 6 presents the mean cross-correlation coefficient between the target dust maps and different models as a function of ℓ for the test set for various λ values (shown by different colors and line styles). We find that the correlation is lower for smaller λ . In particular, for $\lambda = 0$ the correlation is significantly lower. This is likely because using only the pixel difference in the loss function puts too much weight on the small-scale information and the improvement on the large-scale correlation, which is our main scientific interest, is limited. As $\lambda \geq 30$, there is no significant change for the correlation on all scales. Thus, in this paper we set $\lambda = 30$ and the two loss terms have similar amplitudes.

C Result for east and west galactic hemisphere division

In figure 5, we find a significant difference of the integrated intensity model performance in training (north) and test (south) sets. To investigate this, we divide the sky into east and west galactic hemispheres, and the rest of the data preparation follows the procedures in section 2. To ensure no overlap for data in the two hemispheres, we add a buffer of 36 degrees, so at the equator the east and west hemisphere has RA ranges of $(198^\circ, 342^\circ)$ and $(18^\circ, 162^\circ)$. We then train the models using the data from the east hemisphere and access the model performance in the west hemisphere. Figure 7 shows the training and test results in left and right panel, respectively. First, we note that there is no significant difference of the integrated intensity model for training and test sets. Thus, the difference we see between north and south galactic hemispheres is likely a statistical fluctuation. In addition, we find similar conclusion as in

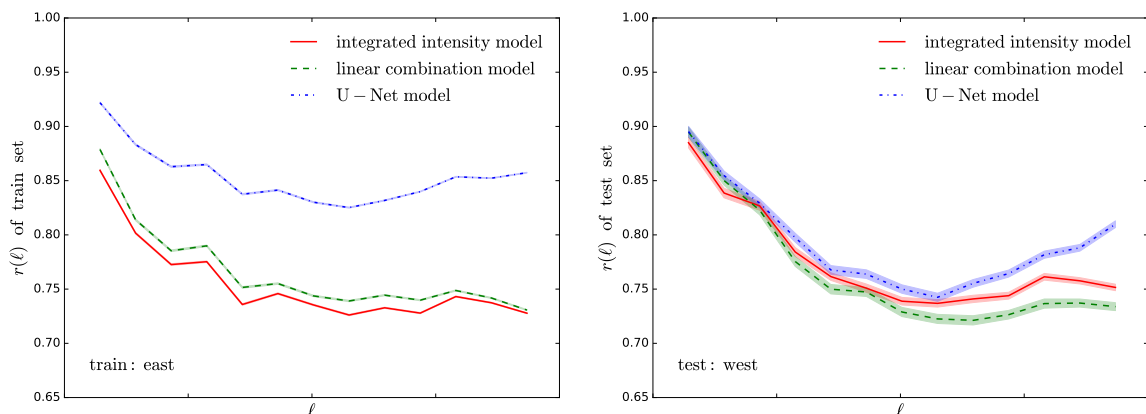


Figure 7. Same as figure 5, but the training and test sets are taken from east and west galactic hemispheres, respectively.

figure 5. Namely, for the test set, the linear model has a similar performance as the integrated intensity model, whereas the U-Net model outperforms. However, the improvement is more significant on small scales. Since a different sky division yields comparable results, we conclude that the training of the neural network does not suffer strong overfitting.

References

- [1] SIMONS OBSERVATORY collaboration, J. Aguirre et al., *The Simons Observatory: Science goals and forecasts*, *JCAP* **1902** (2019) 056 [[1808.07445](#)].
- [2] K. N. Abazajian, P. Adshead, Z. Ahmed, S. W. Allen, D. Alonso, K. S. Arnold et al., *CMB-S4 Science Book, First Edition, arXiv e-prints* (2016) arXiv:1610.02743 [[1610.02743](#)].
- [3] S. Hanany, M. Alvarez, E. Artis, P. Ashton, J. Aumont, R. Aurlien et al., *PICO: Probe of Inflation and Cosmic Origins, arXiv e-prints* (2019) arXiv:1902.10541 [[1902.10541](#)].
- [4] M. Hazumi et al., *LiteBIRD: A Satellite for the Studies of B-Mode Polarization and Inflation from Cosmic Background Radiation Detection*, *J. Low. Temp. Phys.* **194** (2019) 443.
- [5] Planck Collaboration, Y. Akrami, F. Arroja, M. Ashdown, J. Aumont, C. Baccigalupi et al., *Planck 2018 results. I. Overview and the cosmological legacy of Planck, arXiv e-prints* (2018) arXiv:1807.06205 [[1807.06205](#)].
- [6] C. Sheehy and A. Slosar, *No evidence for dust B-mode decorrelation in Planck data*, *Phys. Rev. D* **97** (2018) 043522 [[1709.09729](#)].
- [7] Planck Collaboration, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini et al., *Planck 2018 results. XI. Polarized dust foregrounds, arXiv e-prints* (2018) arXiv:1801.04945 [[1801.04945](#)].
- [8] P. A. R. Ade, Y. Akiba, A. E. Anthony, K. Arnold, M. Atlas, D. Barron et al., *a Measurement of the Cosmic Microwave Background B-Mode Polarization with Polarbear*, *Publication of Korean Astronomical Society* **30** (2015) 625.
- [9] Planck Collaboration, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini et al., *Planck 2018 results. IV. Diffuse component separation, arXiv e-prints* (2018) arXiv:1807.06208 [[1807.06208](#)].
- [10] F. Elsner, B. Leistedt and H. V. Peiris, *Unbiased pseudo- C_ℓ power spectrum estimation with mode projection*, *MNRAS* **465** (2017) 1847 [[1609.03577](#)].

- [11] D. Alonso, J. Sanchez, A. Slosar and LSST Dark Energy Science Collaboration, *A unified pseudo- C_ℓ framework*, *MNRAS* **484** (2019) 4127 [[1809.09603](#)].
- [12] A. E. Lilley, *The Association of Gas and Dust from 21-CM Hydrogen Radio Observations.*, *ApJ* **121** (1955) 559.
- [13] S. E. Clark, J. C. Hill, J. E. G. Peek, M. E. Putman and B. L. Babler, *Neutral Hydrogen Structures Trace Dust Polarization Angle: Implications for Cosmic Microwave Background Foregrounds*, *Phys. Rev. Lett.* **115** (2015) 241302 [[1508.07005](#)].
- [14] S. E. Clark, *A New Probe of Line-of-sight Magnetic Field Tangling*, *ApJ* **857** (2018) L10 [[1802.00011](#)].
- [15] S. E. Clark, J. E. G. Peek and M.-A. Miville-Deschênes, *The Physical Nature of Neutral Hydrogen Intensity Structure*, *ApJ* **874** (2019) 171 [[1902.01409](#)].
- [16] S. E. Clark and B. S. Hensley, *Mapping the Magnetic Interstellar Medium in Three Dimensions Over the Full Sky with Neutral Hydrogen*, [1909.11673](#).
- [17] HI4PI Collaboration, N. Ben Bekhti, L. Flöer, R. Keller, J. Kerp, D. Lenz et al., *HI4PI: A full-sky H I survey based on EBHIS and GASS*, *A&A* **594** (2016) A116 [[1610.06175](#)].
- [18] B. Winkel, J. Kerp, L. Flöer, P. M. W. Kalberla, N. Ben Bekhti, R. Keller et al., *The Effelsberg-Bonn H I Survey: Milky Way gas. First data release*, *A&A* **585** (2016) A41 [[1512.05348](#)].
- [19] P. M. W. Kalberla and U. Haud, *GASS: The Parkes Galactic All-Sky Survey. Update: improved correction for instrumental effects and new data release*, *A&A* **578** (2015) A78 [[1505.01011](#)].
- [20] P. M. W. Kalberla, W. B. Burton, D. Hartmann, E. M. Arnal, E. Bajaja, R. Morras et al., *The Leiden/Argentine/Bonn (LAB) Survey of Galactic HI. Final data release of the combined LDS and IAR surveys with improved stray-radiation corrections*, *A&A* **440** (2005) 775 [[astro-ph/0504140](#)].
- [21] K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke et al., *HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere*, *ApJ* **622** (2005) 759 [[arXiv:astro-ph/0409513](#)].
- [22] F. Bernardeau, C. Pitrou and J.-P. Uzan, *CMB spectra and bispectra calculations: making the flat-sky approximation rigorous*, *J. Cosmology Astropart. Phys.* **2011** (2011) 015 [[1012.2652](#)].
- [23] Y. LeCun, Y. Bengio and G. Hinton, *Deep learning*, *Nature* **521** (2015) 436 EP .
- [24] M. D. Zeiler and R. Fergus, *Visualizing and understanding convolutional networks*, in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars, eds.), (Cham), pp. 818–833, Springer International Publishing, 2014.
- [25] V. Dumoulin and F. Visin, *A guide to convolution arithmetic for deep learning*, 2016.
- [26] O. Ronneberger, P. Fischer and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.
- [27] S. Ioffe and C. Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, pp. 448–456, 2015, <http://jmlr.org/proceedings/papers/v37/loff15.pdf>.
- [28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, *Dropout: A simple way to prevent neural networks from overfitting*, *Journal of Machine Learning Research* **15** (2014) 1929.
- [29] A. Odena, V. Dumoulin and C. Olah, *Deconvolution and checkerboard artifacts*, *Distill* (2016) .
- [30] P. Hough, “Method and Means for Recognizing Complex Patterns.” U.S. Patent 3.069.654, December, 1962.