Bayesian decomposition of the Galactic multi-frequency sky using probabilistic autoencoders*

Sara Milosevic1,2, Philipp Frank1,2, Reimar H. Leike1,2, Ancla Müller3, and Torsten A. Enßlin1,2

1 Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, 80539 München, Germany
e-mail: sara.mia.milosevic@gmail.com
2 Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85748 Garching, Germany
3 Ruhr-Universität Bochum, Universitätstr. 150, 44801 Bochum, Germany

Received 14 September 2020 / Accepted 8 March 2021

ABSTRACT

Context. All-sky observations show both Galactic and non-Galactic diffuse emission, for example from interstellar matter or the cosmic microwave background (CMB). The decomposition of the emission into different underlying radiative components is an important signal reconstruction problem.

Aims. We aim to reconstruct radiative all-sky components using spectral data, without incorporating knowledge about physical or spatial correlations.

Methods. We built a self-instructing algorithm based on variational autoencoders following three steps: (1) We stated a forward model describing how the data set was generated from a smaller set of features, (2) we used Bayes’ theorem to derive a posterior probability distribution, and (3) we used variational inference and statistical independence of the features to approximate the posterior. From this, we derived a loss function and optimized it with neural networks. The resulting algorithm contains a quadratic error norm with a self-adaptive variance estimate to minimize the number of hyperparameters. We trained our algorithm on independent pixel vectors, each vector representing the spectral information of the same pixel in 35 Galactic all-sky maps ranging from the radio to the γ-ray regime.

Results. The algorithm calculates a compressed representation of the input data. We find the feature maps derived in the algorithm’s latent space show spatial structures that can be associated with all-sky representations of known astrophysical components. Our resulting feature maps encode (1) the dense interstellar medium (ISM), (2) the hot and dilute regions of the ISM, and (3) the CMB, without being informed about these components a priori.

Conclusions. We conclude that Bayesian signal reconstruction with independent Gaussian latent space statistics is sufficient to reconstruct the dense and the dilute ISM, as well as the CMB, from spectral correlations only. The computational approximation of the posterior can be performed efficiently using variational inference and neural networks, making them a suitable approach to probabilistic data analysis.

Key words. methods: data analysis – methods: statistical – techniques: image processing – Galaxy: general – ISM: structure

1. Introduction

The interstellar medium (ISM) is a key element of the Milky Way and subject to both astrophysical and cosmological studies. It consists of localized components such as molecules and interstellar dust in cold clouds, atomic and ionized hydrogen, and hot plasma in the Galactic halo, as well as components which pervade the entire ISM, such as cosmic rays and magnetic fields. Our present knowledge about the existence of these components is based on the fact that they all contribute to the interstellar radiation field (e.g., Draine 2011). Each component, or the interplay of several components, generates radiation of a specific spectrum and can be reconstructed by component separation algorithms.

There are components showing very characteristic emission lines such as CO or neutral (HI) and ionized (HII) atomic hydrogen, which permits their Galactic distribution to be determined very precisely. Other components, however, can generate radiation distributed over completely opposite areas of the electromagnetic spectrum. For example, the interaction of cosmic rays and magnetic fields generates synchrotron radiation and can be observed in the radio regime, as in the 408 MHz radio map (Haslam et al. 1982), while the interaction of cosmic rays with interstellar matter imprints in the γ-ray regime due to hadron-nucleon collisions producing pions (e.g., Mannheim & Schlickeiser 1994). Another example is hot ionized plasma, which radiates in the X-ray regime when generated by supernovae, whereas hot molecules ionized by collisions emit in the UV regime (e.g., Ferriere 2001). In this specific case, both the UV and soft X-ray photons are again absorbed by interstellar dust, which prevents observations in regions of high dust density such as the Galactic plane. This interplay shows the high complexity of radiative extinction, which needs to be taken into account when reconstructing single emission components.

The cosmic microwave background (CMB), for example, is not affected by dust extinction, but it is superimposed by Galactic foregrounds. Cosmological studies aim to extract the CMB radiation from multiple frequency channels by identifying and systematically removing these Galactic foregrounds. In frequencies below 100 GHz, Galactic synchrotron and free-free emission contaminate the CMB; whereas, above 100 GHz, thermal dust emission and the cosmic infrared background...
dominate (Gold et al. 2011; Planck Collaboration X 2016). To distinguish between different sources of emission, members of the Planck Collaboration IV (2020) developed several component separation algorithms, for example the COMMANDER (Eriksen et al. 2008), SEVEM (Leach et al. 2008), or SMICA code (Delabrouille et al. 2003; Cardoso et al. 2008). The results obtained from those algorithms contain, among others, all-sky maps of the CMB; synchrotron; free-free; thermal and spinning dust; and CO line emission (Planck Collaboration X 2016). The approaches are, however, based on cosmological, astrophysical, and instrumental parameters, and they require preprocessed spectral templates or explicit knowledge about physical correlations. The COMMANDER code, for example, describes the data at a given frequency as the linear sum of several signal components and a noise term. The signal terms contain a number of astrophysical spectra describing zodiacal light, thermal and spinning dust emission, or synchrotron radiation, which depend on physical parameters, such as the dust temperature, and which are processed by an instrument response function that depends on frequency calibration factors and instrumental bandpass integration effects (Planck Collaboration X 2016). To verify component separation results in a manner independent of such extensive modeling, approaches allowing automated component identification have been increasingly pursued in recent years (Longo et al. 2019).

A broad range of machine learning algorithms was applied to cosmological and astrophysical problems (Fluke & Jacobs 2020). For example, Beaumont et al. (2011) employed the Support Vector Machine (SVM) algorithm to classify structures in the ISM. The algorithm was able to identify a supernovae remnant behind a molecular cloud based on a sample of manually classified data. Ucci et al. (2018a) examined the ISM composition of dwarf galaxies by processing the available spectral information in a machine learning code. The so-called GAME algorithm was trained on a large library of synthetic data (Ucci et al. 2018b) and recovered ISM properties such as metallicity, gas densities, and their respective correlations on the basis of spectral emission lines. By training a convolutional neural network on synthetic spectra, Murray & Peek (2019) were able to decompose the thermal phases of neutral hydrogen HI. This selection of algorithms represents supervised learning approaches, which require labeled or preclassified data in order to be trained. We investigate to what extent unsupervised approaches can be applied to Galactic observations. Our study is based on the work of Müller et al. (2018). The authors applied Gaussian mixture models (GMMs) to Galactic all-sky data in order to augment pixels with missing measurement data by learning spectral pixel-wise relations. They verified the prestated hadronic and leptonic component of the γ-ray sky (Selig et al. 2015) and presented a higher resolved hadronic component as well as a completion of nonexistent information. In their work, the authors had hoped to correlate the main component maps of the GMM with the different phases of the ISM, which however was not the case (see Appendix F in Müller et al. 2018). The idea of classifying astrophysical data in an unsupervised manner motivated the approach in the present study to explore other latent variable models, such as autoencoders, which are able to encode relevant representations of the data in their latent spaces. Autoencoders belong to a class of unsupervised machine learning approaches called representation learning (Hinton 1990; Bengio et al. 2013; Goodfellow et al. 2016). Representation learning methods are constructed to extract the underlying causes that generated the observed data, in other words they calculate the most informative representation of the observation (Bengio et al. 2013; Goodfellow et al. 2016). Autoencoders reduce the dimension of the input data to the so-called underlying, explanatory factors of variation in their latent space (Hinton & Salakhutdinov 2006; Bengio et al. 2013; Goodfellow et al. 2016). They are trained to regenerate their initial input as accurate as possible from their lower-dimensional latent space, meaning a generative process is modeled. We translated this functionality to the task of astrophysical component separation: Since astrophysical data display the superposition of several radiative processes, we modeled this superposition by the generative part of an autoencoder and extracted the underlying radiative processes as “features” in the latent space. In contrast to component separation algorithms similar to the COMMANDER code, we did not predefine the components to be separated, but rather constructed the algorithm to learn a lower-dimensional representation of the data and analyzed the resulting features after training. We stated a similar forward model compared to the COMMANDER code, but by the use of neural networks, we built our algorithm to learn the generative signal interactions from the data. This unsupervised and generic approach to classify astrophysical data could be used as a preprocessing step for subsequent analyses. Especially, considering the large quantities of astrophysical data produced every day by surveys such as the Sloan Digital Sky Survey, an effective preprocessing strategy needs to be developed to be able to analyze the vast amount of data effectively (Kremer et al. 2017; Reis et al. 2019).

The learning process of autoencoders, and in general of neural networks, is guided by so-called loss functions. In this study, we developed a specific loss function for spectral component separation. We used a Bayesian framework to formulate the signal reconstruction problem, and approximated the resulting posterior distribution using variational inference. Autoencoders consisting of a probabilistic inference and generation process are called variational autoencoders (VAE; Kingma & Welling 2013). This approach was, for example, used in the context of the Sloan Digital Sky Survey, as Portillo et al. (2020) successfully applied VAEs to the SDSS data. The authors efficiently reduced the dimension of 1000 input pixels to six latent components, while the VAE was able to outperform principal component analysis considering the spectral dimensionality reduction. Their latent space separates types of galaxies or detects outliers, making it a very useful preprocessing step for large astrophysical data. However, the authors claim that the uncertainty quantification could be improved and suggest to include the pixel-level noise as a separate feature to improve latent variable representations. In contrast to this study, we fully derived the loss function describing the component separation process that, compared to a standard VAE, contains additional terms guiding the learning process. These additional terms describe the influence of an adaptive model noise introduced by our forward model and weights introduced by our specific generative process. The Bayesian framework further allows to calculate uncertainties and to define the significance of each result.

2. Data and methods

In the following, we describe how we modeled a loss function describing the problem of spectral component separation. The loss function guided the learning process of our algorithm and was based on statistical assumptions rather than knowledge about physical correlations. In Sect. 2.2, we state a generative model describing how 35 Galactic all-sky maps were generated by a smaller number of signals. We combined the generative model with statistical prior knowledge on the signals
and applied Bayes’ theorem to calculate an analytic solution to the signal reconstruction problem. Using variational inference, we derived a minimization objective approximating the signal reconstruction, which we optimized with neural networks. The approximated signals are called “features” in this study, following machine learning terminology. The architecture of our algorithm was based on autoencoders, which are trained to reconstruct their input data from a lower-dimensional, latent representation. After training, this latent representation contained our resulting features. In Sect. 2.3 we explain how the derived minimization objective was translated to a variational autoencoder architecture.

2.1. Data

Our goal was to determine a compressed spectral representation of the Galactic sky. We built an unsupervised algorithm to learn this representation from data itself rather than defining specific components to be reconstructed. To give the algorithm a broad range of data to learn from, we chose a data set consisting of 39 Galactic all-sky maps compiled by Müller et al. (2018), covering frequencies from the radio to γ-ray regime.

To generate the data set, the authors assembled information from all-sky surveys (Müller et al. 2018, and references therein) with at least 80% spatial coverage in HEALPix format (Gorski et al. 2005), and used the highest resolution map for each frequency. The UV regime is not part of their data set, since the resolution is not determined in the beginning, but we aimed to associate it with posteriori with relevant physical quantities. We assumed that each $d_i \in D$ was generated from a source vector $s_i \in \mathbb{R}^l$ with $l \leq k$, and the collection of these vectors is expressed by an indexed source set $S = \{s_1, \ldots, s_p\}$. This relation is described in a data model

$$D = \tilde{D} + N_{\text{model}},$$

where we defined the observed data $D$ to be composed of $\tilde{D} := (d_1, \ldots, d_p)$, which is the output of a generative process $G : \mathbb{R}^l \to \mathbb{R}^k$, and the model noise $N_{\text{model}}$. The generative process $G(\cdot)$ mapping the variables of $S$ pixel by pixel to $\tilde{D}$, that is $\tilde{D} = G(S)$, was unknown at this point.

2.2. Model design

Determining a lower-dimensional representation of the spectral information in our data set is an inverse problem: We were looking for unknown quantities, our so-called features, and procedures applied to them that could have contained the full information in $D$. To describe this problem, we used generative modeling to define a data model with corresponding parameters $\Theta$. We solved the inverse problem with the posterior probability distribution $P(\Theta | D)$ of the model parameters $\Theta$, which we specify in the next paragraph. Using Bayes’ theorem, this posterior can be expressed by the data likelihood and the prior distribution up to a $\Theta$-independent factor, reading $P(\Theta | D) \propto P(D | \Theta) \times P(\Theta)$. We performed a pixel-based analysis of image data by assuming magnitude pixel vectors to be independent. This enabled us to factorize the data likelihood and prior distributions. In the following, we describe how we calculated the respective distributions per pixel and how we combined all derivations to one solution for the entire data set $D$.

We started by defining a generative process leading from an abstract source $S$ to the observations in the data set $D$. $S$ was not determined in the beginning, but we aimed to associate $S$ with posteriori with relevant physical quantities. We assumed that each $d_i \in D$ was generated from a source vector $s_i \in \mathbb{R}^l$ with $l \leq k$, and the collection of these vectors is expressed by an indexed source set $S = \{s_1, \ldots, s_p\}$. This relation is described in a data model

$$D = \tilde{D} + N_{\text{model}},$$

where we defined the observed data $D$ to be composed of $\tilde{D} := (d_1, \ldots, d_p)$, which is the output of a generative process $G : \mathbb{R}^l \to \mathbb{R}^k$, and the model noise $N_{\text{model}}$. The generative process $G(\cdot)$ mapping the variables of $S$ pixel by pixel to $\tilde{D}$, that is $\tilde{D} = G(S)$, was unknown at this point.

2.2.1. Prior distributions

We factorized the prior distribution of the set of source vectors $S$ as $P(S) = \prod_{i=1}^{p} P(s_i)$. The prior on $P(s_i)$ can be arbitrarily complex, but without loss of generality we stated a transformation $s_i = T(z_i)$ of the source vectors to an indexed set of latent vectors $z_i \in \mathbb{R}^l$, providing a standardized prior distribution $P(z_i) = G(z_i, \mathbb{1})$. The coordinate transformation into the eigenspace of the prior (e.g., Devroye 1986; Kingma & Welling 2013; Rezende et al. 2014; Titsias & Lázaro-Gredilla 2014) enabled us to absorb all complex and unknown structures of the source space in the transformation $T(\cdot)$. This transformation results in a unit Gaussian prior distribution. Using this definition, we rewrote the generative process as a parametrized function of the latent variables $\tilde{d}_i = G(s_i) = G(T(z_i)) =: f_\theta(z_i)$, with $\theta$ denoting the parameters of the transformed generative forward model. We did not model the generative process explicitly, but used a neural network to learn the function $f_\theta$ from the data.

We further defined the model noise $N_{\text{model}} = (n_1, \ldots, n_p)$ as an indexed set of $p$ noise vectors $n_i \in \mathbb{R}^k$, and assumed the pixel-wise noise was independent and identically distributed with a Gaussian distribution of zero mean and noise covariance $N \in \mathbb{R}^{k \times l}$, meaning $P(N_{\text{model}}) = \prod_{i=1}^{p} P(n_i) = \prod_{i=1}^{p} G(n_i, N)$. This noise covariance induces a metric in data space, which is between two magnitude pixel vectors $d_i$ and $\tilde{d}_i$. $N$ thus indicates how accurately the data vector $d_i$ can be reconstructed using the latent variables $z_i$. Since this accuracy was unknown, we introduced $N$ as an inference parameter. We then performed a transformation to the prior eigenspace of the form $N = f_\theta(\xi_N)$ with the latent noise parameter $\xi_N \in \mathbb{R}$, being distributed as $P(\xi_N) = G(\xi_N, 1)$ with the transformation

---

1 A single pixel vector reads $d_i = (d_i^{(1)}, d_i^{(2)}, \ldots, d_i^{(l)})$, where $d_i^{(j)}$ is the magnitude flux value of the $j$th sky map at pixel $i$. 
with the indexed latent space set \( \mathbf{\theta} \) distribution on these forward model parameters, we assumed a uniform prior described in Sect. 2.2. It consists of the latent vectors \( \mathbf{z} \) parameter vector of the generative forward model.

parameter \( \mathbf{\psi} \). For our specific application, we chose a log-normal mapping

\[
N = t_{\mathbf{\psi}}(\mathbf{\xi}_N) = \exp(\mathbf{\mu}_N + \mathbf{\sigma}_N \mathbf{\xi}_N),
\]

with mean \( \mathbf{\mu}_N \) and standard deviation \( \mathbf{\sigma}_N \). In other words, this is a diagonal noise covariance matrix \( N \) with transformation parameters \( \mathbf{\psi} = (\mathbf{\mu}_N, \mathbf{\sigma}_N)^T \) and a single global noise parameter \( \mathbf{\xi}_N \), which we aimed to learn.

Combining the definitions of the noise and forward model, we rewrote the data model from Eq. (1) as a pixel-wise expression:

\[
d_i = f_{\mathbf{\psi}}(z_i) + n_i.
\]

At this point, we defined the data model parameter vector \( \mathbf{\Theta} \) described in Sect. 2.2. It consists of the latent vectors \( \mathbf{z} \), the latent noise parameter \( \mathbf{\xi}_N \), which indirectly defines the model noise \( \mathbf{n} \), and the parameters of the generative model \( \mathbf{\theta} \). The parameters \( \mathbf{\theta} \) are the parameters of the neural network used to approximate the generative process, and they are specified in further detail in Sect. 2.3. Since we did not have prior knowledge on these forward model parameters, we assumed a uniform prior distribution on \( \mathbf{\theta} \) (see Table 1). This resulted in \( \mathbf{\Theta} = (\mathbf{Z}, \mathbf{\theta}, \mathbf{\xi}_N) \) with the indexed latent space set \( Z = (z_1, \ldots, z_p) \).

### 2.2.2. Data likelihood

We included our forward model in the pixel-wise likelihood

\[
P(D | \mathbf{\Theta}) = \prod_{i=1}^{P} P(d_i | Z, \mathbf{\theta}, \mathbf{\xi}_N)
\]

by marginalizing over the model noise

\[
P(d_i | Z, \mathbf{\theta}, \mathbf{\xi}_N) = \int d\mathbf{n} P(d_i | \mathbf{n}, Z, \mathbf{\theta}, \mathbf{\xi}_N) = \int d\mathbf{n} \delta(d_i - f_{\mathbf{\psi}}(\mathbf{z}_i) - \mathbf{n}_i) G(n_i, \mathbf{\psi}(\mathbf{\xi}_N)) = G(d_i - f_{\mathbf{\psi}}(\mathbf{z}_i), \mathbf{\psi}(\mathbf{\xi}_N)),
\]

where we assumed the noise \( \mathbf{n} \) to be Gaussian distributed and a priori independent of \( Z \) and \( \mathbf{\theta} \). A Gaussian distributed likelihood states that we want the squared deviation of all reconstructed brightness magnitudes compared to the original data to be small. A way to generalize this requirement in the case of more complicated and higher dimensional data distributions is a GAN-VAE (Larsen et al. 2016), where the likelihood is represented by a neural network as well.

### 2.2.3. Approximating the posterior distribution

Combining the prior distributions listed in Table 1 and the data likelihood in Eq. (4), the posterior distribution \( P(\mathbf{\Theta} | D) \) reads:

\[
P(Z, \mathbf{\theta}, \mathbf{\xi}_N | D) \propto \prod_{i=1}^{p} P(d_i | Z, \mathbf{\theta}, \mathbf{\xi}_N) \prod_{i=1}^{p} P(z_i) P(\mathbf{\theta}) P(\mathbf{\xi}_N)
\]

\[
\propto \prod_{i=1}^{p} \left( G(d_i - f_{\mathbf{\psi}}(z_i), \psi(\mathbf{\xi}_N)) \right) \times G(\mathbf{\xi}_N, 1),
\]

where we used \( P(\mathbf{\theta}) = \text{const} \). To overcome analytic limitations, we used variational inference (e.g., Blei et al. 2017) to approximate the posterior distribution \( P(\mathbf{\Theta} | D) \) with an easier to calculate distribution \( Q_{\mathbf{\Theta}}(\mathbf{\Theta} | D) \) with variational parameters \( \Phi \). We defined a suitable approximate distribution \( Q_{\mathbf{\Theta}}(\mathbf{\Theta} | D) \) and used the Kullback-Leibler Divergence\(^2\) as a measure to evaluate the dissimilarity of \( P(\mathbf{\Theta} | D) \) and \( Q_{\mathbf{\Theta}}(\mathbf{\Theta} | D) \).

Assuming that \( Z, \mathbf{\theta} \) and \( \mathbf{\xi}_N \) are a posteriori independent, the approximate distribution \( Q_{\mathbf{\Phi}} \) was written as

\[
Q_{\mathbf{\Phi}}(Z, \mathbf{\theta}, \mathbf{\xi}_N | D) = Q_{\mathbf{\Phi}}(Z | D) Q_{\mathbf{\Phi}}(\mathbf{\theta} | D) Q_{\mathbf{\Phi}}(\mathbf{\xi}_N | D).
\]

For \( Q_{\mathbf{\Phi}}(Z | D) = \prod_{i=1}^{p} Q_{\mathbf{\Phi}}(z_i | d_i) \) we chose a pixel-wise independent Gaussian distribution \( \mathcal{G}(\mathbf{z}_i - \mathbf{\mu}_i, \mathbf{\Sigma}_i) \) based on the maximum entropy principle (Jaynes 1982). The principle states if only the mean \( \mathbf{\mu} \) and covariance \( \Sigma \) of data are known, then the knowledge contained in that data set can be best expressed by a Gaussian distribution with exactly this mean \( \mathbf{\mu} \) and covariance \( \Sigma \) (e.g., Enßlin 2019). Later, we discuss that by construction \( \mathbf{\mu}_i \) and \( \Sigma_i \) are functions of the input data, making the choice of a Gaussian distribution for \( Q_{\mathbf{\Phi}}(Z | D) \) valid. For \( Q_{\mathbf{\Phi}}(\mathbf{\theta} | D) \) and \( Q_{\mathbf{\Phi}}(\mathbf{\xi}_N | D) \) we chose maximum a posteriori solutions. In practice, we evaluated the approximation at the variational parameter values \( \hat{\mathbf{\theta}}, \hat{\mathbf{\xi}}_N \) expressed by \( Q_{\mathbf{\Phi}}(\mathbf{\theta} | D) = \delta(\hat{\mathbf{\theta}} - \mathbf{\theta}) \) and \( Q_{\mathbf{\Phi}}(\mathbf{\xi}_N | D) = \delta(\hat{\mathbf{\xi}}_N - \hat{\mathbf{\xi}}_N) \).

To approximate \( P(\mathbf{\Theta} | D) \) with \( Q_{\mathbf{\Phi}}(\mathbf{\Theta} | D) \), we used the Kullback-Leibler Divergence

\[
D_{KL} \left[ Q_{\mathbf{\Phi}}(Z, \mathbf{\theta}, \mathbf{\xi}_N | D) || P(Z, \mathbf{\theta}, \mathbf{\xi}_N | D) \right] =
\]

\[
\int dZ d\mathbf{\theta} d\mathbf{\xi}_N \ln \left( \frac{Q_{\mathbf{\Phi}}(Z, \mathbf{\theta}, \mathbf{\xi}_N | D)}{P(Z, \mathbf{\theta}, \mathbf{\xi}_N | D)} \right).
\]

Inserting the derived expressions from the previous sections and using Monte Carlo Methods to approximate integrals with finite sums, results in

\[
D_{KL}[Q_{\mathbf{\Phi}}(\cdot) || P(\cdot)] = \frac{1}{2} \sum_{i=1}^{P} \left[ -\ln(\Sigma_i) - l(1 + \ln(2\pi)) + \frac{1}{p} \hat{\mathbf{\xi}}^2 - \right.\]

\[
+ \ln \left( \frac{1}{Q_{\mathbf{\Phi}}(\mathbf{\xi}_N | D)} \right) \right] + \ln \left( \frac{1}{Q_{\mathbf{\Phi}}(\mathbf{\xi}_N | D)} \right)
\]

\[
+ \left. \ln(\Sigma_i + \mathbf{\mu}_i\mathbf{\mu}_i^T) + \ln \left( \frac{1}{Q_{\mathbf{\Phi}}(\mathbf{\xi}_N | D)} \right) \right] + \mathcal{H}_0,
\]

\[\text{KL-Divergence} \]
where we absorbed all constant terms into $\mathcal{H}_0$. The full calculations with all intermediate steps are presented in Appendix B.

2.3. NEAT-VAE

Our goal was to infer a lower-dimensional representation of the data, which in our case was expressed by the latent source space $Z$. We achieved this by minimizing Eq. (8), which describes the spurious amount of artificial information introduced by the approximation of $P$ with $Q_{\theta}$. In the following, we explain how this objective function was translated into the framework of a latent variable model called variational autoencoder (Kingma & Welling 2013).

A basic autoencoder is constructed to compute a low-dimensional, latent representation of higher-dimensional input data. This is achieved by training the autoencoder to reconstruct the original input as accurate as possible from the reduced, latent representation (e.g., Rumelhart et al. 1985; Hinton & Salakhutdinov 2006; Goodfellow et al. 2016). In this context, training describes the minimization of an objective or loss function, for example the mean squared error between input data and the reconstructed output data. The dimensionality reduction of the input to the latent space occurs in the so-called encoder, the latent space itself is called bottleneck layer, and in the decoder, the latent space gets translated back to data space. Variational autoencoders (VAEs) offer a probabilistic framework to jointly optimize latent variable (or generative) models and inference models (Kingma & Welling 2019). Equation (8) was transformed to such a VAE framework (see Fig. 1) as follows:

- **Generative Process:** Neural networks are generalized function approximators. Since the exact form of the data generating process $f_{\theta}(z_i) = d_i$ was unknown, we used a neural network with input $z_i$ and output $d_i$ to approximate $f_{\theta}$, and we used back propagation to optimize the parameters $\theta$ of the network, which corresponded to the generative decoder.

- **Variational Inference:** The inference of the latent space variables was approximated by $Q_{\theta}(z_i \mid d_i) = Q(d_i \mid \mu_i, \Sigma)$. We built a function delivering posterior samples following this variational distribution through a neural network with input $d_i$ and output $z_i$. We let the pixel-wise mean $\mu_i \in \mathbb{R}^t$ and covariance $\Sigma \in \mathbb{R}^{t \times t}$ be determined by a parametrized function of the input data $e_{\phi}(d_i) = (\mu_i, \log \Sigma)$, where $\phi$ contains the parameters of the function $e_{\phi}$ and the matrix logarithm of $\Sigma$ is calculated. This parametrization ensured the variance to be positive and the calculation to be numerically stable, since the logarithmic function maps the small values of the variance to a larger space. By inverse transform sampling, we defined the posterior latent space variables as $z_i = \mu_i + \exp \left( \frac{1}{2} \log (\Sigma) \right) \cdot \epsilon_i = \mu_i + \Sigma^{1/2} \epsilon_i$, with an auxiliary variable $\epsilon_i$, and $P(\epsilon_i) = \mathcal{N}(\epsilon_i, 1)$. In practice, we approximated the function $e_{\phi}(d_i)$ by the variational inference encoder. We took $\sqrt{\Sigma}$ to be diagonal and described it by its diagonal vector $\mathcal{N}(\sqrt{\Sigma})$, allowing us to calculate $\mu_i$ and $\sqrt{\Sigma}$ as two distinct outputs of the encoder network.

- **Independent Representation:** Based on the input data, the encoder network delivers latent space variables $z_i$ with mean $\mu_i$ and covariance vector $\Sigma_i = \text{diag} \left( \mathcal{N}(\Sigma) \right)$. Using this definition we found the optimal independent approximation to the posterior $P(\theta \mid D)$. This lead to a disentanglement of the input information, meaning each dimension of $z_i$, which is equivalent to a hidden neuron in the bottleneck layer, encoded (approximately) mutually independent features of the data.

The minimization objective in Eq. (8) contains the loss function of a classic VAE with three modifications: (1) We included the size of the latent space $l$ and its corresponding weight (see Appendix B for the derivation) to be able to compare different latent space sizes with each other, (2) in addition to the network weights, we aimed to optimize the noise covariance and thus the latent noise variable $\xi_i$. The prior on this latent noise added an extra term proportional to $\xi_i^{-2}$ to the objective function, and (3) by including noise in the data model, the likelihood contains a factor $1/(l \xi_i^2)$ and contributed an additional term $\ln(n \xi_i^2)$ from the normalization. Since we expanded the VAE framework by the adaption to noise, we named our algorithm NEAT-VAE (NoisE AdapTing Variational AutoEncoder).

Using the transformations illustrated before, the final objective function only depends on the generative decoder parameters $\theta$, the variational encoder parameters $\phi$ and the latent

---

**Fig. 1.** Model design of our algorithm (NEAT-VAE). (1) In the forward model, we assumed the 35 Galactic all-sky maps $D$ to be generated by a smaller number of features $Z$ and some additive noise $N$. The generative process $f_{\theta}$, which we approximated by the decoder neural network, was learned by the algorithm. (2) We calculated the joint posterior distribution of the features $Z$, the network parameters $\theta$, and the noise parameter $\xi_i$ using statistical priors and Bayes’ theorem. (3) We approximated the posterior distribution using variational inference, the maximum entropy principle, and inverse transform sampling. Batches of spatially independent pixel vectors $d_i$ served as input data, where each vector contains the spectral information of the same pixel in 35 Galactic all-sky maps. The algorithm was constructed to infer a latent representation $z_i$ of the input data (encoder), and to regenerate its input $d_i$ as accurate as possible from the latent space (decoder). The minimization objective, or loss function, guiding the algorithm’s learning process is Eq. (8).
noise parameters \(\tilde{\xi}_N\). We implemented an autoencoder architecture (encoder network, bottleneck layer, decoder network) with Eq. (8) as the respective loss function in the PyTorch framework\(^3\). The framework enabled us to calculate derivatives of Eq. (8) with respect to \(\theta\), \(\phi\) and \(\tilde{\xi}_N\) using automated back propagation, and to minimize the loss with a built-in optimizer. The conducted experiments are described in Appendix C.

2.4. Evaluation metric

**Reconstruction accuracy.** To evaluate the ability of the NEAT-VAE to reconstruct its input, we defined the reconstruction accuracy by the mean squared error of input maps and reconstructed (output) maps by

\[
\text{MSE}(\mathbf{d}, \hat{\mathbf{d}}) = \frac{1}{p} \sum_{i=1}^{p} (d_i - \hat{d}_i)^2, \tag{9}
\]

with the number of pixels \(p\), the 35-dimensional data vectors \(d_i \in \mathbf{D}\) and the corresponding reconstructed vectors \(\hat{d_i} \in \hat{\mathbf{D}}\).

**Significance of results.** To investigate whether there is an order among the latent feature maps calculated by the NEAT-VAE with respect to information content, we evaluated the significance of each feature for the reconstruction of the input data by

\[
S_{\text{feature}} = \frac{1}{p} \sum_{i=1}^{p} \left( \frac{\tilde{z}_{\text{feature map}, i} - \bar{z}_{\text{feature map}}}{\sigma^2_{\text{feature}, i}} \right)^2, \tag{10}
\]

where \(\tilde{z}_{\text{feature map}, i}\) is the intensity value at the \(i\)th pixel, \(\bar{z}_{\text{feature map}}\) is the mean intensity value of a single feature map, and \(\sigma^2_{\text{feature}, i}\) denotes the posterior variance as calculated by the algorithm at the \(i\)th pixel. This means we calculated the feature map variance, which describes the fluctuations within the posterior mean map, weighted by the posterior feature variance averaged over all pixels \(p\). The significance expresses the ratio of the magnitude of fluctuations within a map compared to the uncertainties of the map. In this context, a high significance highlights the most relevant features for the reconstruction, while a feature significance below 1.0 corresponds to an insignificant feature, as the posterior uncertainty is larger compared to the posterior mean values of the map.

**Encoded Information.** We investigated which data information a feature map encodes by considering the generative process of the NEAT-VAE and using the back propagation algorithm: Since the amount of parameters in the latent space is much smaller compared to the data space, the autoencoder is forced to extract the essential information of the input data in order to generate it again. The generative process therefore offers, at least approximately, an explanation for the information flow within the autoencoder, and we visualized to what extend specific features are generating the individual Galactic all-sky maps using sensitivity analysis (e.g., Zurada et al. 1994). Here, we computed the gradients of the reconstructed output maps with respect to the feature maps using the back propagation algorithm. The values of the resulting decoder Jacobian are displayed in HEALPix pixelization in Appendix E.

\(^3\) https://pytorch.org/, publicly available at https://gitlab.mpdpf.mpg.de/msara/neat_vae

We used two other measures to quantify correlations between subsequent all-sky maps. One is the discrete case defined by

\[
\rho(X; Y) = \frac{\text{cov}(X, Y)}{\text{std}(X) \cdot \text{std}(Y)}, \tag{11}
\]

where \(\text{cov}(X, Y)\) is the covariance of two maps \(X\) and \(Y\), and \(\text{std}(\cdot)\) denotes the standard deviation of the respective maps. A value of \(-1.0\) corresponds to negative correlation, \(0.0\) means the values are not correlated, and \(1.0\) corresponds to positive correlation. The Pearson correlation coefficient only accounts for linear relations between two maps. To account for nonlinear relations as well, we used the mutual information

\[
I(X; Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \ln \left( \frac{P(x, y)}{P(x)P(y)} \right), \tag{12}
\]

which can be calculated from two-dimensional histograms of feature maps \(X\) and output maps \(Y\). For a given number of bins, we represented the joint probability distribution \(P(x, y)\) in an \((x, y)\)-shaped matrix by counts per bin, while the marginalized distributions \(P(x)\) and \(P(y)\) were obtained by summing over the respective \(y\) and \(x\)-axes of this matrix. In our interpretations, we used all above measures to evaluate the encoded information in the latent space (mutual information of feature and input maps), the generative process from the latent space (decoder Jacobian maps), and the linear correlations between maps (correlation plots and Pearson coefficient) in order to determine the physical content of the resulting features.

3. Results and discussion

We applied the NEAT-VAE to the described 35 Galactic all-sky maps with equal resolution. We trained the algorithm on magnitude pixel vectors \(\mathbf{d}\), independently, we specifically did not include spatial correlations. Each pixel vector contains the spectral information of the same pixel in all 35 Galactic all-sky maps. After training, the NEAT-VAE yields a posterior probability distribution of the reduced spectral information of the input data in its latent space. In the following, we show all-sky representations of the latent space variables \(z_i \in Z\), where the dimension of the latent space corresponds to the number of hidden neurons in the bottleneck layer and was varied in different experiments. We call the subsequent hidden neurons “features” and the resulting full-sky representations “feature maps”.

3.1. Dimensionality reduction

We first analyzed how the dimension of the latent space in the NEAT-VAE, that is the number of features, correlates with the reconstruction quality of the generative process (Fig. 2). Using Eq. (9), we observe that only three features are required to achieve an MSE below 0.1, which describes an average deviation of the reconstructed values compared to the input values of a natural logarithm-based flux magnitude. For small values, the absolute uncertainty on logarithmic scale equals the relative uncertainty on linear scale, meaning MSE = 0.1 corresponds to a relative uncertainty of \(\approx 10\%\). The MSE decreases for a further growing number of features, and stagnates around a value of 0.02, corresponding to a relative uncertainty of \(\approx 2\%\) at approximately ten features. We interpret this as an indication for a high redundancy of the information contained in the 35 Galactic all-sky maps, since increasing the number of latent features does not increase the quality of the reconstruction any further.
Fig. 2. Reconstruction mean squared error (MSE) and values of the minimized Kullback-Leibler Divergence (Loss) \(\Delta D_{KL}\) (\(D_{KL}\) in Eq. (8) except constant terms) depending on the dimension of the latent space (also called number of features, x-axis). The values were determined using the NEAT-VAE with a configuration of six layers, 30 hidden neurons in the encoder and decoder layers, noise transformation parameters \(\mu_N = -7\) and \(\sigma_N = 1\), learning rates of 0.005 for the network weights and 0.001 for \(\xi_N\), and a batchsize of 128. We did not track all normalization constants through the calculations, which leads to negative values for the loss.

3.2. Morphology of features

Based on the experiments with different latent space dimensions, we examined the spatial structures in the feature maps in more detail. We recognize spatially correlated structures of the input data in the feature maps. We assume this to be a meaningful result, since the autoencoder only computes spectral correlations among the magnitude flux values within one pixel vector \(d_i\), and is not informed about spatial structures. We also observe feature maps with the same morphology to occur in latent spaces of different dimensions. Averaged over all experiments we conducted (see Appendix C), we call the three most significant features A, B, and C.

For the configuration shown in Fig. 3, the features have significance values (Eq. (10)) \(S_{\text{feature}A} = 1.67 \times 10^2\), \(S_{\text{feature}B} = 1.41 \times 10^2\), and \(S_{\text{feature}C} = 4.68 \times 10^1\). The remaining features have significance values ranging from 8.75 to 3.82 \(\times 10^1\) and encode artifacts and other morphologies of the input data, as displayed in Appendix D. A similar behavior was also observed by Müller et al. (2018): The GMM components used in their study also encode artifacts of the input data when the number of components is increased. However, posterior samples of the latent space of our algorithm show that the NEAT-VAE does not always assign information to each and every feature: Starting from twelve features and adding further neurons to the latent space, the significance of the added features drops below 1.0. This means the posterior variance of a feature map is greater than the fluctuations within the map, and the resulting posterior samples show white noise statistics. On average, ten to eleven features are significant throughout our experiments with varying latent space size. We assume that from this point on, our algorithm has identified all mutually independent features of the input data. When further dimensions are added to the latent space, the algorithm “tunes out” those degrees of freedom by making them insignificant.

In the next sections, we focus on the three most significant features of the configuration with ten latent space features (since from this point on the reconstruction does not change significantly, see Fig. 2), and their physical interpretation. Visual inspections of the other features (Appendix D) indicate that the separation into independent components is not fully reached, possibly due to the finite amount of optimization. For example, morphological structures similar to the North Galactic Spur, the Fermi bubbles, and measurement artifacts imprint onto several features simultaneously.

The posterior mean values in Fig. 3 reflect the optimally compressed representation of the input data as calculated by our algorithm. We observe that the overall sign of the maps changes for different hyperparameter configurations. We do not observe the change of the sign to follow a specific rule or to depend on the sign of other features from the same hyperparameter configuration. Hence, we assume the overall sign not to have any physical interpretation and only to depend on the initialization of the network parameters. We also observe that the relation of positive and negative values within single maps is constant throughout different choices of hyperparameters, that is the overall feature or feature changes. Since these signs do not change the morphology of the maps, we chose the sign of the color code in the map display of Fig. 3 according to the input map which it most closely resembles. This corresponds to positive values for stronger emitting regions and negative values for weaker emitting regions.

3.3. Identifying the information encoded by feature A

Feature A, which is the most significant feature in 98% of the examined hyperparameter configurations (see Appendix C), is displayed in Fig. 3a. From a visual analysis, we recognize a positive color-coded Galactic plane and negative color-coded Galactic poles in the posterior mean. In the eastern part of the Galactic plane, we see a bright, circular structure in the Cygnus region. Further in the east, south of the Galactic plane, structures similar to the Perseus region occur. The circular shaped structure north of the Galactic center resembles the Ophiuchus region and southwestern of the Galactic center structures similar to the Small and Large Magellanic clouds can be recognized. In the western part of the Galactic plane, the bright structures match the Orion region. The posterior variance shows a high certainty in the region of the Galactic plane and the structures of the surrounding latitudes, while at the southern and northern Galactic poles, the uncertainty increases.

Interpretation. We find compelling evidence that feature A traces the dense and dusty parts of the ISM: Based on calculations of the mutual information with 512 bins, the top three data sets contributing to feature A are the AKARI far-infrared 140 \(\mu\)m input data with mutual information \(I = 1.91\), the IRIS infrared 100 \(\mu\)m input data with \(I = 1.84\), and the AKARI far-infrared 160 \(\mu\)m input data, also with \(I = 1.84\). This frequency band of the interstellar radiation field is dominated by the infrared emission of dust (e.g., Draine 2011, p.121). Feature A shares the least mutual information with the ROSAT X-ray 1.545 keV input data with \(I = 0.20\). Next, we analyzed the decoder Jacobian maps (see Appendix E), which display the gradients of the reconstructed Galactic all-sky maps with respect to latent space features in HEALPix format. We observe the following: The lower and mid-lattitudes of the \(\gamma\)-ray regime strongly depend on feature A, and going from higher to lower energies, the overall-dependence on feature A grows. Especially, the low-energy regime of the Fermi data is dominated by hadronic interactions of cosmic rays with the interstellar medium (ISM) and thus shows the gas distribution (Selig et al. 2015). In the X-ray regime, small to no dependence is observed; however, the
Fig. 3. Main results of the NEAT-V AE: (panel a) feature A, (panel b) feature B and (panel c) feature C are the most significant representations of the input data. Left panels: posterior mean, right panels: posterior variance. The NEAT-V AE was trained to reduce the spectral information of 35 Galactic all-sky maps to ten features in the latent space with the configuration described in Fig. 2. We show the three most important feature maps in descending order of significance according to Eq. (10). The colors in the posterior mean of feature C (left panel in (c)) were inverted for illustration purposes. The gray pixels correspond to missing values in the input data.

mid- and high latitudes of the soft X-ray data (0.212 keV and 0.197 keV) show negative gradients toward feature A. Here, negative gradients highlight an inverse-proportional relationship of feature A and the reconstructed X-ray maps, which is plausible by the physical context: Radiation from X-rays is extincted by cold interstellar gas (Ferriere 2001), thus the absence of X-ray emission reveals regions where interstellar matter is present. The Hα map at 656.3 nm, which displays emission due to hydrogen transitions occurring from the second excited state $n = 3$ to the first excited state $n = 2$ (e.g., Finkbeiner 2003), positively depends on feature A. This corresponds to neutral hydrogen preferably residing in the dense ISM, as traced by feature A. We observe a very strong dependence of the Galactic planes of the 12 and 25 µm infrared data on feature A, while with
Fig. 4. Correlation of feature A with published astrophysical components tracing interstellar matter. From top left to bottom right: (panel a) thermal dust component in logarithmic scaling (Planck Collaboration X 2016), (panel b) 2D histogram of the posterior mean intensity of feature A and the thermal dust component (mutual information $I = 1.72$, Pearson correlation coefficient $\rho = 0.98$), (panel c) hadronic component of the $\gamma$-ray spectrum in logarithmic scaling (Selig et al. 2015) with white pixels denoting missing values, (panel d) 2D histogram of the posterior mean intensity of feature A and the hadronic component ($I = 1.22$, $\rho = 0.90$). For the 2D histograms, the intensity ranges of the maps were divided into 256 equal bins and the number of intensity pairs per bin was displayed as counts on logarithmic scaling. Bright colors denote a high number of counts. Mutual information was calculated as described in Eq. (12) with 512 bins.

The positive dependencies of sky maps on feature A describe areas of our Galaxy with a high density of interstellar matter, while the negative gradients correlate with the extinction of emissions by interstellar matter. We assume positive gradients highlight pixels in the latent space used to generate the corresponding pixels in data space, while negative gradients mark an anti-correlation of feature and data pixels. On the basis of this specific combination of gradients and the mutual information, we infer that feature A encodes dense regions of the ISM.

Discussion. We can test our hypothesis by investigating the correlation of feature A with dust as a tracer for the ISM (Kennicutt & Evans 2012). Figure 4b shows the relationship of feature A and the thermal dust emission (Fig. 4a) as calculated by the Planck COMMANDER code (Eriksen et al. 2008; Planck Collaboration X 2016). The posterior mean of feature A has a strong, positive, and linear correlation ($\rho = 0.98$) with the thermal dust component. Both the COMMANDER and the NEAT-VAE algorithm perform a Bayesian, pixel-wise analysis based on a data model containing the linear sum of a signal function and noise, but with two main differences: First, we only employed statistical information in our forward model, while the COMMANDER code included detailed physical templates for the various emission processes contributing to the radio to far-infrared sky, as well as calibration and correction factors, and a prior for the CMB. Second, the COMMANDER algorithm has 11 million free parameters to tune (Planck Collaboration X 2016), while our model has only about 5000. The algorithms are not
directly comparable, since the COMMANDER code was especially developed to separate the Galactic foregrounds in order to reconstruct the CMB, while the NEAT-VAE only seeks to find an informative representation of the input data. But in this context, the NEAT-VAE summarized the input data into categories, one of which already contains dust emission. With this result, we assume it is possible to derive the dust component based on feature A with little computational effort.

We investigate another tracer for interstellar matter, namely the hadronic \( \gamma \)-ray component derived by Selig et al. (2015), shown in Fig. 4c. It represents the \( \gamma \)-ray emission due to the interaction of cosmic ray protons with interstellar matter and it is positively correlated \((\rho = 0.90)\) with feature A, see Fig. 4d. This correlation is reasonable, since Selig et al. (2015) composed the hadronic component of the low-energy \( \gamma \)-ray maps, on which feature A also depends. These results meet our initial aim of finding a reduced representation of the input data that combines redundant information in one feature, in this case tracers for dense regions in the ISM.

The high significance of feature A is likely to result from the choice of input data, since most of the maps in our data set \( D \) represent emission from the dense interstellar matter. We did not include the CO line emission data, but from the positive gradients of the Planck 100–353 GHz channels with respect to feature A, which contain the information of the CO line emission (Planck Collaboration X 2016), we assume that the Galactic plane of feature A most likely would generate the CO data. This would support our interpretation of feature A, since CO is a tracer for molecular interstellar gas (Scoville et al. 1987), and thus, for regions of high density.

### 3.4. Identifying the information encoded by feature B

Feature B, the second most significant feature in 98% of our experiments, is displayed in Fig. 3b. We see a negative color-coded equator which resembles the Galactic plane, with positive color-coded bulges north and south of the plane. Especially the northern bulge structure looks similar to the morphology of the north polar Spur. Due to this extinction, it is likely that regions of the ISM and is thus mostly absorbed in low and mid-latitudes (Ferriere 2001). Regarding the gradient maps in Appendix E, there is little to no dependence of the reconstructed \( \gamma \)-ray data on feature B. Starting from X-ray data at 1.545 keV, the positive gradients get stronger with decreasing energy, and especially the bulge-shaped area of the reconstructed X-ray data strongly depends on feature B. The soft X-ray regime around 0.25 keV, which coincides with hydrogen cavities (e.g., Sanders et al. 1977; Snowden et al. 1990), also shows a weak, positive dependence on feature B. The infrared and microwave regime again show little to no dependence, but we observe small negative gradients to be present whereas in the corresponding gradients maps of feature A, strong positive gradients occur. The reconstructed 1420 MHz and 408 MHz radio maps show positive dependence on feature B, especially in the bulge-shaped areas. This data set detects the synchrotron radiation generated by the interaction of cosmic ray electrons with magnetic fields in the ISM (Ginzburg & Syrovatskii 1965), and the intensity of this radiation depends on the density of relativistic particles and the magnetic field strength. The former one is predominantly found in the hot areas of the ISM if only for the much larger volume occupation of this phase within the Milky Way (Cox 2005). Thus, we assume feature B encodes the enhanced presence of such cosmic ray electrons. With exception of the low energy X-ray and radio synchrotron regime, we observe the data predominantly generated by feature A to have little to no dependence on feature B, which we interpret as feature B being complementary to dense regions of the ISM. In combination with the indications from the mutual information with the X-ray input data and positive gradients, we assume feature B encodes tracers for dilute and hot regions of the ISM.

**Discussion.** Figure 5, showing the correlation between Feature A and B, supports our interpretation: The features, in our case the dense and dilute regions of the ISM, are basically uncorrelated with a Pearson correlation coefficient \( \rho = 0.03 \). This relationship is reasonable since the features describe two distinct categories of the ISM: The data generated by feature A are based on emissions from cold and warm ionized gas (as observed in the 21 cm and H\(_{I}\) line emissions), dust, and interactions with interstellar matter (cosmic rays). The ionization of the warm gas occurs mainly due to photo-ionization by \( O \) and \( B \) stars, the hottest and most massive stars of the Milky Way. Due to their ratio of mass and luminosity, these stars have a short main sequence life cycle and are thus found near their initial birthplaces, the dense ISM (e.g., Blome et al. 1997, p.60). Feature B, however, encodes data which represent radiation of an even hotter medium, the hot ionized plasma, which is generated by supernova explosions (e.g., Kahn 1980). The radiative processes encoded by features A and B can thus be traced back to two fundamentally different origins, namely emission from interstellar matter and from stellar explosions.

One regime of the electromagnetic spectrum missing in our input data is the ultraviolet (UV) frequency band. In this regime, the emission of very hot gas which gets ionized by collisions can be observed, but UV radiation does not penetrate dense regions of the ISM and is thus mostly absorbed in low and mid-latitudes (Ferriere 2001). Due to this extinction, it is likely that
Fig. 6. Panel a: CMB as derived by the Planck Collaboration IV (2020), panels b–d: correlations of the posterior mean of feature C with (panel b) the CMB, mutual information $I = 0.88$ and Pearson correlation coefficient $\rho = 0.70$, (panel c) feature A, $I = 0.29$ and $\rho = 0.01$, and (panel d) feature B, $I = 0.20$ and $\rho = 0.01$. The histograms and mutual information were compiled as described in Fig. 4. Bright colors in the histograms denote a high number of counts.

UV radiation would be interpreted by the NEAT-VAE as redundant to the soft X-ray data, which shows similar dust absorption patterns. This would support our interpretation of feature B to encode the hot and dilute ISM.

3.5. Identifying the information encoded by feature C

The third most significant feature in 74\% of the investigated configurations, Feature C, is displayed in Fig. 3c. Here, we inverted the colors of the posterior mean to resemble the color-coding of the CMB, see Fig. 6a. The regions of the Galactic plane and Cygnus have strong, positive values, while most of the other structures fluctuate around zero. The uncertainty is highest in the Galactic plane, especially in the Galactic center. There are bulge-like shapes in the variance map north and south of the Galactic center denoting a medium level of uncertainty. Toward higher latitudes, the uncertainty is very low.

Interpretation. The mutual information of feature C is highest with the Planck data of 70, 100 and 143 GHz with values of $I = 1.01, 0.95, 0.82$, respectively, and the least mutual information is observed with the ROSAT X-ray 1.545 keV data with $I = 0.12$. According to the decoder Jacobian maps in Appendix E, feature C mostly generates the Planck 30–217 GHz channels, especially the 70 and 100 GHz data, in which the CMB can be observed (Planck Collaboration I 2020). All other reconstructed maps show little to no dependence on feature C, with exception of a weak, positive dependence of the soft ROSAT X-ray data that we address in Sect. 3.6. Based on the mutual information and the strong positive gradients in the microwave regime around 100 GHz, we assume feature C to encode the CMB.

Discussion. Considering our physical interpretation of features A and B, it is reasonable that the CMB emission is encoded in a separate feature, since the underlying physical processes in the Early Universe shaping the CMB fluctuations are independent of the processes in the ISM. To test our hypothesis we compare feature C to the CMB all-sky map derived by the Planck COMMANDER code in Fig. 6a. Feature C shows a positive, linear correlation ($\rho = 0.70$) with the CMB (Fig. 6b). A main difference between feature C and the CMB can be seen in the Galactic plane, which feature C encodes (in addition to the Cygnus region) with high intensities, but also with high uncertainty (see Fig. 3c). We assume the high intensities in the Galactic plane of feature C to originate from the model’s architecture: Since we do not take spatial correlations into account, the algorithm is constructed to learn spectral relations of each pixel vector independently. The Galactic planes of the input maps are dominated by high intensity values, while for other latitudes, more low-intensity values are present. We assume this generates two
distinct clusters of intensity ranges, leading to two different spectral relations learned by the algorithm. We can also observe this distinction in intensity by analyzing the contributions of pixels for data generation (see Appendix E). The structure of the Galactic plane is recognizable in most of the decoder Jacobian maps, even though the algorithm has no information about spatial correlations. This again indicates the relations detected in different intensity regimes to represent different processes in the NEAT-VAE algorithm, which however are attributable to internal computations rather than physical properties.

3.6. Nonphysical interpretation of the NEAT-VAE

We finally examine the decoder Jacobian maps which do not have a clear physical interpretation. For example, we observe feature A (the dense ISM) contributing to the 1420 and 408 MHz radio data, as well as feature C (the CMB) to supposedly generate the soft X-ray data.

The NEAT-VAE algorithm recombines all features of the latent space in a highly nonlinear way to generate the output maps, meaning the gradients in Appendix E cannot always be considered independently or in a linear way. Especially when the absolute gradient values of the reconstructed maps are large for more than one feature, a holistic analysis is required. In most of the displayed cases in Appendix E, there is one feature dominantly generating the reconstruction of input data (displayed by one gradient map (red and blue colors) per row showing stronger intensities than others). One exception, however, is the radio data, where we observe positive gradients with respect to both feature A and B. A speculative physical explanation is that feature A marks regions of enhanced magnetic field strength and feature B of enhanced relativistic electron densities, both quantities that in combination determine the synchrotron emission. However, we rather assume that these two features have no physical meaning for synchrotron emission, but the internal computations of our algorithm run most efficiently when latent values of features A and B are used for radio data generation. One always has to consider the fact that our algorithm has no information about physical relations and primarily optimizes a statistical function. Another example for such a behavior can be found in the soft X-ray data, where we observe strong negative gradients with respect to feature A and weak positive gradients toward both features B and C. This mixing of features in a partially contrasting manner indicates the nonlinearity of the function mapping from the latent space values to the reconstructed output data: We hypothesize some values might be used by the algorithm just to tune others out, since the decoder has to regenerate the data based on all available features. This discussion shows that the decoder Jacobians give a valid first overview of the strongest correlations, but have to be interpreted with care. Not each contribution to the gradient represents a physical causality, especially since all gradients are entangled and cannot be considered independently.

4. Conclusions

In summary, we derived a representation learning algorithm (NEAT-VAE) for spectral component separation on a pixel basis. The architecture of the NEAT-VAE was based on variational autoencoders and was constructed to learn a compressed representation of its input data. The learning process is guided by a minimization objective, which we designed to describe a component separation process following three steps: (1) We allowed the input data to be generated by a smaller set of features and some additive noise, where the generative process was automatically learned by our algorithm. (2) We calculated the posterior distribution of these features using statistical priors and Bayes’ theorem, and (3) approximated the posterior distribution using variational inference, the maximum entropy principle, and inverse transform sampling.

We ran our experiments on 35 Galactic all-sky maps ranging from radio to \( \gamma \)-radiation with a spatial resolution of \( N_{\text{side}} = 128 \). Our algorithm computed a compressed representation of the input data in spectral direction, without the knowledge of physical or spatial correlations. After training, the calculated latent space pixels were merged to all-sky maps. The most significant representation of the spectral information in our input data contains three separated astrophysical morphologies, namely emission patterns of the dense ISM, a superposition of the hot and dilute ISM, and the CMB. The association with physical quantities was verified in three ways by (1) correlation plots, (2) calculations of the Pearson correlation coefficient and (3) the mutual information with the dust and CMB component maps of the Planck Collaboration. Since the NEAT-VAE processes pixel vectors independently without the knowledge of their spatial correlations, we conclude that the morphologies of the dense and dilute ISM, as well as the CMB are sufficiently defined by their pixel-wise spectral correlations.

The minimization objective of the NEAT-VAE and the associated hyperparameter tuning mainly determine the performance of the algorithm. Since the mathematical derivation was performed in a Bayesian framework, the initial signal component separation problem was described in a probabilistic manner. This enabled us to track uncertainties and evaluate the significance of each feature map. Hyperparameter tuning was reduced by including parameters in the Bayesian inference process, as in our case the noise covariance of the model noise.

Based on our results, we conclude that spectral independent morphologies can be disentangled by representation learning algorithms, however a more detailed separation into single astrophysical components requires spatial correlations and physical knowledge. We propose our algorithm could be used as a preprocessing step in order to perform a component separation analysis more efficient. For example, the task of deriving the thermal dust emission is computationally very expensive. If only the data necessary for the separation of the dust component were used, such as our feature map encoding the dense regions of the ISM, the separation process might perform faster and more efficient compared to analyzing a larger data set with redundant and entangled information.

This work shows that Bayesian signal inference can be performed with neural networks in an unsupervised manner, and that variational autoencoders are a suitable architecture for problems consisting of generative modeling and inference tasks. It also shows that purely statistical approaches are not sufficient to replace physical analyses, but when customized to the specific problem, for example by incorporating statistical independence, they reveal meaningful relations in astrophysical data.

Acknowledgements. We acknowledge Jakob Knollmüller for helpful discussions and thank an anonymous reviewer for useful comments that improved the clarity and quality of the manuscript.

References

Ackermann, M., Ajello, M., Albert, A., et al. 2012, ApJS, 203, 4
Atwood, W. B., Abdo, A. A., Ackermann, M., et al. 2009, ApJ, 697, 1071
Beaumont, C. N., Williams, J. P., & Goodman, A. A. 2011, ApJ, 741, 14
### Appendix A: Astrophysical data set

**Table A.1.** Overview of the data sets used in this study.

| Mission/survey       | Astrophysical contribution                                                                 | Frequency range | Maps | References                                      |
|----------------------|-------------------------------------------------------------------------------------------|-----------------|------|-------------------------------------------------|
| Haslam radio         | Synchrotron radiation from the interactions of cosmic ray electrons with magnetic fields in the ISM. | 408 MHz         | 1    | Haslam et al. (1982), Remazeilles et al. (2015) |
| Reich radio          |                                                                                           | 1420 MHz        | 1    | Reich (1982), Reich & Reich (1986), Reich et al. (2001) |
| HI line emission     | Emission from atomic hydrogen hyper fine level transitions. Proxy of the atomic hydrogen column density. | 21 cm           | 1    | HI4PI Collaboration (2016)                      |
| *Planck* microwave   | Low frequency range is dominated by synchrotron emission, high frequency range by thermal dust emission. Further emission originates from the CO-line, spinning dust (rotating dust grains), free-free (electron-ion collisions), and the CMB. | 30–857 GHz      | 9    | Planck Collaboration I (2016)                   |
| AKARI far-infrared   | Thermal dust emission.                                                                      | 90–160 µm       | 3    | Doi et al. (2015)                               |
| IRIS infrared        |                                                                                           | 72–100 µm       | 4    | Neugebauer et al. (1994), Miville-Deschênes & Laigle (2005) |
| Hα line emission     | Hydrogen transitions from the second to the first excited state. Proxy for the amount of warm to hot, partly ionized diffuse gas in the vicinity of the Sun (as self-absorbed). | 656.3 nm        | 1    | Denison et al. (1999), Gauystad et al. (2001), Madsen et al. (2001), Finkbeiner (2003) |
| ROSAT X-ray          | Emission from hot, largely ionized gas regions of the ISM with typically low density.       | 0.197–1.545 keV | 6    | Snowden et al. (1995, 1997), Freyberg (1998), Freyberg & Egger (1999) |
| *Fermi* γ-ray        | Low energy range is dominated by hadronic interactions of cosmic ray protons with the ISM, high energy range is dominated by leptonic interactions of cosmic ray electrons with the Galactic photon field (inverse-Compton scattering). | 0.85–217.22 GeV | 9    | Atwood et al. (2009), Ackermann et al. (2012), Selig et al. (2015) |

**Notes.** Data was preprocessed by Müller et al. (2018) as described in Sect. 2.1. Our data compilation **D** consists of 35 all-sky maps at different frequencies. The columns of the table specify (1) the mission or survey which provided the corresponding maps, (2) the astrophysical origin of the brightness in the maps, (3) the frequency or wavelength range, (4) the number of maps that belong to the specific mission or survey, and (5) the corresponding bibliographic reference.
Appendix B: Calculations

Here, we provide the analytic steps between Eqs. (7) and (8). For completeness, we start our calculations with the full posterior \( P(\theta \mid \mathbf{D}) = P(\mathbf{D} \mid \theta) \times P(\theta)/P(\mathbf{D}) \):

\[
D_{KL}(Q_{\phi(\cdot)} \mid \mid P(\cdot)) = \int dZ \, d\theta \, d\xi_N \, Q_{\phi}(Z, \theta, \xi_N \mid \mathbf{D}) \ln \left\{ \frac{Q_{\phi}(Z, \theta, \xi_N \mid \mathbf{D})}{P(Z, \theta, \xi_N \mid \mathbf{D})} \right\}
\]

\[
= \int \left( \prod_{i=1}^{p} dz_i \, Q_{\phi}(z_i \mid d_i) \right) \int d\theta \, Q_{\phi}(\theta \mid \mathbf{D}) \int d\xi_N \, Q_{\phi}(\xi_N \mid \mathbf{D})
\]

\[
\times \ln \left\{ \frac{\prod_{i=1}^{p} Q_{\phi}(z_i \mid d_i) \, Q_{\phi}(\theta \mid \mathbf{D}) \, Q_{\phi}(\xi_N \mid \mathbf{D})}{P(\theta) \, P(\xi_N) \, \prod_{i=1}^{p} G \left( d_i - f_\phi(z_i), t_\phi(\xi_N) \right) G(z_i, \mathbb{1})} \right\}
\]

\[
= \int \left( \prod_{i=1}^{p} dz_i \, Q_{\phi}(z_i \mid d_i) \right) \int d\theta \, Q_{\phi}(\theta \mid \mathbf{D}) \int d\xi_N \, Q_{\phi}(\xi_N \mid \mathbf{D})
\]

\[
\times \left\{ \sum_{i=1}^{p} \ln Q_{\phi}(z_i \mid d_i) + \ln Q_{\phi}(\theta \mid \mathbf{D}) + \ln Q_{\phi}(\xi_N \mid \mathbf{D}) - \ln P(\theta)
\]

\[
- \ln P(\xi_N) + \ln P(\mathbf{D}) - \sum_{i=1}^{p} \ln \left( G(z_i, \mathbb{1}) \, G \left( d_i - f_\phi(z_i), t_\phi(\xi_N) \right) \right) \right\}
\]

\[
= \int \left( \prod_{i=1}^{p} dz_i \, Q_{\phi}(z_i \mid d_i) \right) \left( \sum_{i=1}^{p} \ln Q_{\phi}(z_i \mid d_i) \right) \int d\theta \, Q_{\phi}(\theta \mid \mathbf{D}) \int d\xi_N \, Q_{\phi}(\xi_N \mid \mathbf{D})
\]

Inserting the definitions for \( Q_{\phi}(\theta \mid \mathbf{D}) = \delta(\theta - \hat{\theta}), Q_{\phi}(\xi_N \mid \mathbf{D}) = \delta(\xi_N - \hat{\xi}_N) \) and \( Q_{\phi}(z_i \mid d_i) = G(z_i - \mathbf{u}, \Sigma) \), we get
\[ D_{\text{KL}}[Q(\cdot) \parallel P(\cdot)] = \sum_{i=1}^{p} \left( \ln G(z_i - \mu_i, \Sigma_i) \right)_{\theta(z_i - \mu_i, \Sigma_i)} - \left( \ln G(\xi_N, 1) \right)_{\hat{\theta}(\xi_N)} - \sum_{i=1}^{p} \left( \ln G(z_i, 1) \right)_{\theta(z_i - \mu_i, \Sigma_i)} - \sum_{i=1}^{p} \left( \ln G(d_i - f_g(z_i), \eta \xi_N) \right)_{\theta(z_i - \mu_i, \Sigma_i)} + \mathcal{H}_0, \]  

where \( \mathcal{H}_0 \) collects all terms independent of the parameters. We renamed the subsequent terms of the Kullback-Leibler Divergence as follows and calculated each expression respectively:

- **T1** = \( \left( \ln G(z_i - \mu_i, \Sigma_i) \right)_{\theta(z_i - \mu_i, \Sigma_i)} \),
- **T2** = \( \left( \ln G(\xi_N, 1) \right)_{\hat{\theta}(\xi_N)} \),
- **T3** = \( \left( \ln G(z_i, 1) \right)_{\theta(z_i - \mu_i, \Sigma_i)} \),
- **T4** = \( \left( \ln G(d_i - f_g(z_i), \eta \xi_N) \right)_{\theta(z_i - \mu_i, \Sigma_i)} \).

**First energy term T1:**

\[
T_1 = \int dz_i G(z_i - \mu_i, \Sigma_i) \ln \left( \frac{\exp(-\frac{1}{2}(z_i - \mu_i)^T \Sigma_i^{-1} (z_i - \mu_i))}{\sqrt{2\pi} \Sigma_i} \right) 
= \int dz_i G(z_i - \mu_i, \Sigma_i) \left[ -\frac{1}{2} (z_i - \mu_i)^T \Sigma_i^{-1} (z_i - \mu_i) - \frac{1}{2} \ln (2\pi) \right].
\]

The trace of a scalar is the scalar itself, resulting in \((z_i - \mu_i)^T \Sigma_i^{-1} (z_i - \mu_i) = \text{tr} ((z_i - \mu_i)^T \Sigma_i^{-1} (z_i - \mu_i))\). Further, the trace is invariant under cyclic permutations: \(\text{tr} ((z_i - \mu_i)^T \Sigma_i^{-1} (z_i - \mu_i)) = \text{tr} (\Sigma_i^{-1} (z_i - \mu_i)(z_i - \mu_i)^T)\). Finally, the trace is a linear operator and therefore commutes with the expectation:

\[
= -\frac{1}{2} \text{tr} \left( \int dz_i G(z_i - \mu_i, \Sigma_i) \Sigma_i^{-1} (z_i - \mu_i)(z_i - \mu_i)^T \right) 
= -\frac{1}{2} \int dz_i G(z_i - \mu_i, \Sigma_i) \left[ \ln(\Sigma_i) + l \ln(2\pi) \right] 
= -\frac{1}{2} \left[ \text{tr}(\Sigma_i) + l \ln(2\pi) \right].
\]

In the upper calculation, we carried the term \(l(1 + \ln (2\pi))\) through our computations, since we changed the number of latent space features \(l\) in our different experiments.

**Second energy term T2:**

\[
T_2 = \int d\xi_N \ln \left( \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} \xi_N^2) \right) \delta(\xi_N - \xi_N) 
= -\frac{1}{2} \xi_N^2 - \frac{1}{2} \ln(2\pi) 
= -\frac{1}{2} \xi_N^2 + C_2.
\]
\[ T_3 = \int d z_i \mathcal{G}(z_i - \mu_i, \Sigma_i) \ln \left( \frac{\exp(-\frac{1}{2} z_i^T \Sigma_i^{-1} z_i)}{\sqrt{2\pi I}} \right) \]

\[ = -\frac{1}{2} \ln \left( \int d z_i \mathcal{G}(z_i - \mu_i, \Sigma_i) \frac{1}{2} z_i^T \Sigma_i^{-1} z_i + \int d z_i \mathcal{G}(z_i - \mu_i, \Sigma_i) \frac{1}{2} \ln(2\pi I) \right) = \text{const.} \]

\[ = -\frac{1}{2} \text{tr} \left( \int d z_i \mathcal{G}(z_i - \mu_i, \Sigma_i) z_i z_i^T \right) + C_3. \quad (B.4) \]

The expression \( z_i z_i^T \) can be rewritten as \( z_i z_i^T = (z_i - \mu)(z_i - \mu)^T + z_i \mu^T + \mu z_i^T \). We inserted this in the line above:

\[ = -\frac{1}{2} \text{tr} \left( \int d z_i \mathcal{G}(z_i - \mu_i, \Sigma_i) \right) \left[ (z_i - \mu_i)(z_i - \mu_i)^T - \mu \mu^T + z_i \mu^T + \mu z_i^T \right] + C_3 \]

\[ = -\frac{1}{2} \text{tr} \left( \Sigma_i + \mu \mu^T \right) + C_3. \quad (B.5) \]

**Fourth energy term T4:**

\[ T_4 = \int d z_i \mathcal{G}(z_i - \mu_i, \Sigma_i) \ln \mathcal{G}(d_i - f_\theta(z_i), I) \]

\[ = \int d z_i \mathcal{G}(z_i - \mu_i, \Sigma_i) \ln \left( \frac{\exp(-\frac{1}{2}(d_i - f_\theta(z_i))^T \phi^{-1}(\xi_N))}{\sqrt{2\pi I \phi(\xi_N)}} \right) \]

\[ = -\int d z_i \mathcal{G}(z_i - \mu_i, \Sigma_i) \left[ \frac{1}{2} (d_i - f_\theta(z_i))^T \phi^{-1}(\xi_N) (d_i - f_\theta(z_i)) + \frac{1}{2} \ln(2\pi \phi(\xi_N)) \right] = \text{const.} \]

\[ = -\frac{1}{2} \int d z_i \mathcal{G}(z_i - \mu_i, \Sigma_i) \left[ (d_i - f_\theta(z_i))^T \phi^{-1}(\xi_N) (d_i - f_\theta(z_i)) \right] \]

\[ = -\frac{1}{2} \text{tr} \left( \phi(\xi_N)^{-1} (d_i - f_\theta(z_i)) (d_i - f_\theta(z_i))^T \right) \]

\[ = -\frac{1}{2} \text{tr} \left( \ln \phi(\xi_N) \right) + C_4. \quad (B.6) \]

We cannot exactly evaluate the integral in this expression using analytic techniques. Fortunately, we can make use of Monte Carlo methods to approximate the expectation with a finite sum:

\[ \int d z_i \mathcal{G}(z_i - \mu_i, \Sigma_i) \text{tr} \left[ \phi(\xi_N)^{-1} (d_i - f_\theta(z_i)) (d_i - f_\theta(z_i))^T \right] \approx \frac{1}{M} \sum_{m=1}^{M} \text{tr} \left[ \phi(\xi_N)^{-1} (d_i - f_\theta(z_i^{(m)})) (d_i - f_\theta(z_i^{(m)}))^T \right]. \quad (B.7) \]

With this method we achieved a high accuracy estimator (even with a small number of samples) for the expectation value, as long as we draw samples \( z_i^{(m)} \) from the distribution \( \mathcal{G}(z_i - \mu_i, \Sigma_i) \) (Bishop 2006). This means the samples \( z_i \) are a function of \( \mu_i \) and \( \Sigma_i \) by \( z_i = \mu_i + \Sigma_i \epsilon_i \) with an auxiliary variable \( \epsilon_i \) and \( P(\epsilon_i) = \mathcal{G}(\epsilon_i, I) \) (see Sect. 2.3 for details). We sampled once for each pixel, that is \( M = 1 \). The full Kullback-Leibler Divergence \( (B.1) \) then reads:

\[ D_{\text{KL}}[\mathcal{Q}(\cdot) \parallel P(\cdot)] = \frac{1}{2} \sum_{i=1}^{P} \left[ -\text{tr} \left( \ln \Sigma_i \right) - I(1 + \ln(2\pi)) + \frac{1}{2} \Sigma_i \right] + \text{tr} \left( \Sigma_i + \mu_i \mu_i^T \right) + \text{tr} \left( \frac{1}{2} \phi(\xi_N)^{-1} (d_i - f_\theta(z_i)) (d_i - f_\theta(z_i))^T \right) \]

\[ + \text{tr} \left( \ln \phi(\xi_N) \right) + \mathcal{H}_0, \]

where we absorbed all constant terms in a redefined \( \mathcal{H}_0 \).

A100, page 17 of 26
Appendix C: Hyperparameters

Table C.1. Hyperparameters of NEAT-V AE.

| Hyperparameter          | Sampling sets                                                                 |
|-------------------------|-------------------------------------------------------------------------------|
| Layers                  | {6, 8, 10, 16, 24}                                                           |
| Hidden neurons          | {30, … , 37}                                                                 |
| Bottleneck neurons      | {10, … , 35}                                                                 |
| $\mu_N$                 | $\{−11, −9, 21, −8, −5, −1\}$                                               |
| $\sigma_N$              | Unif[1, 2]                                                                    |
| LR network weights      | {0.005, 0.001, 0.0005, 0.01}                                                 |
| LR $\xi_N$              | {0.0025, 0.0005, 0.0025, 0.005}                                              |
| Batch size              | {16, 64, 128, 256, 512}                                                      |

Notes. The number of layers, neurons per layer and bottleneck neurons determine the network architecture. $\mu_N$ and $\sigma_N$ are transformation parameters of the noise covariance matrix $N$. The optimization of learnable parameters is determined by the learning rates (LR) for network weights $\phi$, $\theta$ and the latent noise $\xi_N$, which are tuned to minimize the objective function in Eq. (B.8). When using mini-batching, the batch size determines how many data samples are used to compute the loss function before back propagation and model updating is performed.

Hyperparameters for the NEAT-VAE are (1) the number of network layers, (2) the number of hidden neurons per layer, (3) the number of neurons in the bottleneck layer, (4) the mean $\mu_N$ and (5) the standard deviation $\sigma_N$ of the log-normal model for the noise covariance transformation, the learning rates of (6) network weights and (7) noise parameter weights, and (8) the batch size. We examined 50 configurations of the NEAT-VAE, which consist of different arrangements of hyperparameters. The values for the hyperparameters were randomly chosen from limited intervals, which we specified for each hyperparameter based on prior experiments (see Table C.1). We implemented rectified linear unit (ReLU) functions as activation functions in each layer except the output layer and used the Adam optimizer (Kingma & Ba 2014) to update the network weights $\phi$ and $\theta$, and the latent noise value $\xi_N$. We trained each configuration for $(10^5 \times \text{batchsize})/p$ epochs with $p$ denoting the number of pixels. We built our NEAT-VAE as a descriptive rather than a predictive model, since we aim to learn the underlying, independent features generating one certain data set. For this reason, we did not split the data into training, validation and test sets, nor did we analyze overfitting. For reproducibility, we fixed the random seed to 123.
Appendix D: Other features

In the experiment discussed in Sect. 3, our NEAT-V AE algorithm mapped 35 Galactic all-sky maps to ten latent space features that encode the essential information required to reconstruct the input again. The latent space exhibits an order in significance of the features, which we measured by the ratio of mean fluctuations compared to feature posterior variance (see Eq. (10)). Features A, B, and C, which have the highest significance, are shown in Fig. 3. The remaining seven features (with their mean and variance in HEALPix representation) are displayed in the following in order of descending significance. By a visual analysis, we can recognize artifacts, for example, of the IRIS scanning scheme, in features D, F, G, and H. Feature H also seems to encode structures near the Galactic plane of the $H\alpha$ map, the mean of feature J encodes structures similar to the Fermi Bubbles in high energy $\gamma$-ray data. All-sky images of the 35 Galactic input data are displayed in the leftmost columns of Appendix E.

**Fig. D.1.** Full latent space representation with 10 neurons in descending order of significance. Significance values are: $S$(feature D) = 38.18, $S$(feature E) = 34.07, and $S$(feature F) = 13.77.
Fig. D.2. Least significant latent space feature maps. Significance values are: $S$(feature G) = 13.63, $S$(feature H) = 11.85, $S$(feature I) = 10.34, and $S$(feature J) = 8.75.
Appendix E: Decoder Jacobian maps

The following panels show the gradients of reconstructed Galactic all-sky maps (that is the output of our NEAT-VAE algorithm) with respect to the latent space features A, B, and C. The Galactic input data is displayed in the leftmost column, while the single features mark the top row. The resulting gradient values in each pixel are shown as a HEALPix map connecting a Galactic map and a feature map, with red colors indicating positive gradient values and blue colors denoting negative gradient values.

Fig. E.1. Decoder Jacobian maps. Derivatives of reconstructed Galactic all-sky maps with respect to latent feature maps A, B, and C (top panel).
Fig. E.1. continued.
Fig. E.1. continued.
Fig. E.1. continued.
Fig. E.1. continued.
Fig. E.1. continued.