A&A, 683, A246 (2024)
https://doi.org/10.1051/0004-6361/202347758

Astronomy & Astrophysics

A deep-learning approach to the 3D reconstruction of dust density and temperature in star-forming regions

Victor F. Ksoll[1], Stefan Reissl[1], Ralf S. Klessen[1,2], Ian W. Stephens[3,4], Rowan J. Smith[9,10], Juan D. Soler[5], Alessio Traficante[5], Philipp Girichidis[1], Leonardo Testi[6,7], Patrick Hennebelle[8], and Sergio Molinari[5]

[1] Universität Heidelberg, Zentrum für Astronomie, Institut für Theoretische Astrophysik, Albert-Ueberle-Str. 2, 69120 Heidelberg, Germany
e-mail: v.ksoll@uni-heidelberg.de
[2] Universität Heidelberg, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany
[3] Department of Earth, Environment, and Physics, Worcester State University, Worcester, MA 01602, USA
[4] Center for Astrophysics, Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA
[5] Istituto di Astrofisica e Planetologia Spaziali (IAPS), INAF, Via Fosso del Cavaliere 100, 00133 Roma, Italy
[6] Alma Mater Studiorum Università di Bologna, Dipartimento di Fisica e Astronomia (DIFA), Via Gobetti 93/2, 40129 Bologna, Italy
[7] INAF – Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, 50125 Firenze, Italy
[8] Université Paris-Cité, Université Paris-Saclay, CEA, CNRS, AIM, 91191 Gif-sur-Yvette, France
[9] School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews, KY16 9SS, UK
[10] Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Received 18 August 2023 / Accepted 26 January 2024

ABSTRACT

Aims. We introduce a new deep-learning approach for the reconstruction of 3D dust density and temperature distributions from multi-wavelength dust emission observations on the scale of individual star-forming cloud cores (<0.2 pc).
Methods. We constructed a training data set by processing cloud cores from the Cloud Factory simulations with the POLARIS radiative transfer code to produce synthetic dust emission observations at 23 wavelengths between 12 and 1300 µm. We simplified the task by reconstructing the cloud structure along individual lines of sight (LoSs) and trained a conditional invertible neural network (cINN) for this purpose. The cINN belongs to the group of normalising flow methods and it is able to predict full posterior distributions for the target dust properties. We tested different cINN setups, ranging from a scenario that includes all 23 wavelengths down to a more realistically limited case with observations at only seven wavelengths. We evaluated the predictive performance of these models on synthetic test data.
Results. We report an excellent reconstruction performance for the 23-wavelength cINN model, achieving median absolute relative errors of about 1.8% in $\log(n/\mathrm{m}^{-3})$ and 1% in $\log(T_{\mathrm{dust}}/\mathrm{K})$, respectively. We identify trends towards an overestimation at the low end of the density range and towards an underestimation at the high end of both the density and temperature values, which may be related to a bias in the training data. After limiting our coverage to a combination of only seven wavelengths, we still find a satisfactory performance with average absolute relative errors of about 2.8% and 1.7% in $\log(n/\mathrm{m}^{-3})$ and $\log(T_{\mathrm{dust}}/\mathrm{K})$.
Conclusions. This proof-of-concept study shows that the cINN-based approach for 3D reconstruction of dust density and temperature is very promising and it is even compatible with a more realistically constrained wavelength coverage.

Key words. methods: statistical – stars: formation – dust, extinction

1. Introduction

A fundamental limitation of astronomical observations is that they only give access to the two-dimensional (2D) projection of cosmic structures onto the plane of the sky. Consequently, it is a central theme of modern astronomical and astrophysical research to resolve this degeneracy and try to reconstruct the underlying three-dimensional (3D) structures. When measuring lines, for example, we can take spectra at equally spaced positions within the area of interest and build a 3D cube of position-position-velocity (i.e. line-of-sight velocity) information. This PPV data is often used as an approximation for the intrinsic 3D position-position-position (PPP) structure of the emitting region (for further discussions, see e.g. Ballesteros-Paredes & Mac Low 2002; Beaumont et al. 2013). However, what to do when we are relying on continuum radiation, such as the thermal emission from interstellar dust grains (e.g. Molinari et al. 2010; Planck Collaboration XI 2014), is less clear. To address this challenge, we present a new deep learning approach for the reconstruction of 3D morphological information and employ multi-band observations in the wavelength regime from 12 to 1300 µm to estimate the 3D spatial distribution of dust density and temperature of star-forming cloud cores.

Thermal emission from interstellar dust grains is the dominant source of radiation across the sky at mid- and far-infrared wavelengths (see Hill et al. 2018, and references therein). This is the result of dust grains distributed throughout the interstellar medium (ISM), which are heated primarily by starlight and cool through thermal radiation (Tielens 2010; Draine 2011; Klessen & Glover 2016). The dust grains are heated to temperatures between

roughly 20 and 200 K depending on the spectrum and intensity of the interstellar radiation field and the size and optical properties of the grains (see Galliano et al. 2018, for a recent review).

Dust emission remains optically thin at relatively high column densities (see, e.g. Planck Collaboration XIX 2011), hence, it provides a crucial observable to study regions of the universe that are not accessible at visible wavelengths. Observations of the dust thermal emission have been used to characterise the evolution of the universe through the study of the cosmic infrared background (CIB, see Hauser & Dwek 2001). Radiation at far infrared wavelengths registered by the *Herschel* satellite has allowed for the reconstruction of star formation activity across the Milky Way disk (Molinari et al. 2016; Elia et al. 2021) and within nearby star-forming regions (André et al. 2010). Polarised dust thermal emission observed by the *Planck* satellite has provided the opportunity to infer the first whole-sky map of the projected Galactic magnetic field (see, e.g. Planck Collaboration XXXV 2016).

Interstellar dust also plays many critical roles in galactic evolution. It is a catalyser for the formation of molecular hydrogen (H_2), while sequestering select elements in solid grains (see, e.g. Gould & Salpeter 1963 or Jones & Ysard 2019). The interaction between dust grains and ultraviolet (UV) starlight releases electrons that can be the dominant source of heating for interstellar gas (Wolfire et al. 2003; Glover & Mac Low 2011). Dust grains also transfer the radiation pressure from starlight to the gas and couple it to the interstellar magnetic field through collisions (see, e.g. Draine 2003 for a review; or Reissl et al. 2018, 2023 for a microphysical model). Thus, reconstructing the distribution of the dust is crucial for understanding the physical conditions of the ISM.

Existing attempts to reconstruct the 3D distribution of dust on galactic scales often take into account distance information from stars in combination with individual extinction measurements (or the combination of *Gaia* with auxiliary data; see e.g. Lallement et al. 2018, 2019, 2022; Leike et al. 2020, 2022; or Zhang et al. 2023). A similar approach has also been adopted for assessing the 3D structure of individual molecular clouds (e.g. Rezaei Kh. et al. 2017, 2020; Zucker et al. 2021; Rezaei Kh. & Kainulainen 2022) or for building a realistic model of the matter distribution in the solar neighbourhood (e.g. Zucker et al. 2022, 2023). On the smaller scales of individual star-forming clumps, there have also been forward-modeling approaches introduced for the 3D dust distribution based on the combination with line data (see, e.g. Liseau et al. 2015) or from multi-frequency dust emission data by fitting overlapping Gaussian ellipsoids (Steinacker et al. 2005). Previous attempts to use machine learning to invert the radiative transfer problem have also been reported (e.g. Garcia-Cuesta et al. 2009).

In this study, we introduce a novel deep learning approach for the 3D reconstruction task, which employs a conditional invertible neural network (Ardizzone et al. 2019a,b). The latter belongs to the group of methods based on normalising flows (e.g. Kobyzev et al. 2021) and has the advantage of giving access to the full posterior distribution function. For this reason, the cINN architecture is particularly well suited for solving degenerate inverse problems and has been successfully applied to a range of subjects in astronomy. These include: the characterisation of stellar properties from photometry (Ksoll et al. 2020) or spectra (Kang et al. 2023b), prediction of exoplanet properties (Haldemann et al. 2023), analysis of emission lines in HII regions (Kang et al. 2022, 2023a), cosmic ray origin studies (Bister et al. 2022), and the reconstruction of galaxy assembly histories from numerical simulations (Eisert et al. 2023). Due to the lack of an observational ground truth sample, we have trained the cINN with data taken from numerical models of the turbulent multi-phase ISM, based on the Cloud Factory suite of simulations introduced by Smith et al. (2020), which we post-processed using detailed radiative transfer calculations (employing POLARIS, see Reissl et al. 2016, 2019) to bring them closer to the observational domain.

This paper is structured as follows. In Sect. 2, we outline the construction of the training data for our method from synthetic dust cloud simulations, including our setup for radiative transfer. Section 3 provides a summary of the invertible neural network approach, specifications of the inverse problem, implementation details, and our analysis methods. In Sect. 4, we present the evaluation of our trained models on synthetic test data and discuss the predictive performance of our approach. Lastly, Sect. 5 summarises our main results.

2. Training data

The main goal of this study is to reconstruct 3D dust distributions from the observed dust emission for sites of star formation on the scale of individual cloud cores. As we want to tackle this task with a supervised deep-learning approach, we therefore require a training data set consisting of 3D dust distributions with their properties and corresponding dust emission observations. As such a data set does not yet exist for real observations, we have turned to simulations to build a suitable database for training.

2.1. Simulation data

As a basis for our training data set we chose the AREPO-based (an adaptive Voronoi mesh hydrosolver, see Springel 2010), galactic-scale ISM hydrodynamics simulation suite Cloud Factory, introduced by Smith et al. (2020) and Izquierdo et al. (2021). The Cloud Factory self-consistently follows the formation of dense gas and molecular hydrogen with an average molecular weight of $\mu_g = 2.4$ in a Milky Way-like galactic gas disc at radii $4\,\mathrm{kpc} < r < 12\,\mathrm{kpc}$, including the effects of galactic scale forces, gas chemistry and cooling, and supernova feedback. The time-dependent chemical evolution is modelled as in Smith et al. (2014) using the hydrogen chemistry of Glover & Mac Low (2007) and the simplified CO treatment of Nelson & Langer (1997). It includes gas self-shielding from a UV field equivalent to that seen in the solar neighbourhood and cosmic-ray ionisation at the local rate as well.

The Cloud Factory employs a series of nested zooms with a base mass resolution of $1000\,M_\odot$ smoothly increased to $10\,M_\odot$ within a co-rotating box of size 3 kpc. In the co-rotating box individual cloud complexes are then selected and their resolution further increased to $0.25\,M_\odot$, which is equivalent to a spatial resolution better than 0.1 pc in gas with number densities higher than $10^9\,\mathrm{m}^{-3}$. By including the galactic scale forces that form the clouds, the Cloud Factory suite reproduces the turbulent gas motions on multiple scales as observed in the ISM. However, the current version of the suite does not include magnetic fields or other forms of stellar feedback, such as stellar winds, jets, or photoionisation.

For our analysis, we used Complex C and D, as shown in Fig. 5 of Smith et al. (2020), and only included gas at the highest resolution ($0.25\,M_\odot$ or four cells per local jeans length, whichever is higher). These molecular cloud complexes were

formed in regions that had previously experienced supernova feedback and, therefore, they already contained a well developed turbulent energy cascade. They are filamentary in structure, and extend for more than 100 pc along their longest axis. The density probability density function of the entire cloud complex peaks at number densities of around 10^8 m^{-3}, but extends beyond number densities of 10^{10} m^{-3}, which marks the point when sink particles may form (see Tress et al. 2020). Sink particles represent regions of star formation and at this resolution, they correspond to single star systems that may actually be multiples. While sinks may form above densities of 10^{10} m^{-3}, they will only be created if the gas passes energy checks to ensure it is bound and converging, so in practice the simulations include densities up to 10^{10}–10^{12} m^{-3}. Similarly, neighbouring gas cells can have material accreted by the sinks, but only when the gas becomes bound to the sink.

A high spatial resolution allows us to select the compact cloud cores on the scale of individual star-forming clumps that we want to analyse in this study. As a starting point, we extracted compact pre-stellar cloud cores (i.e. dense, cloud-like structures that have not formed a sink and are smaller than 0.2 pc) for this purpose from the Cloud Factory. However, we did not just want to model these early phases of star formation, but also more evolved star forming regions, particularly those affected by nearby stars. As the star formation prescription in the Cloud Factory forms sink particles on cluster scales (rather than individual stars), we chose to modify the raw simulation data in a post processing step by manually adding a star to model the types of evolved star-forming regions that we want to consider in this study. We guided these modifications on the example of a well-observed real-world counterpart of this type of star-forming cores, such as the nearby (d = 120 pc, Loinard et al. 2008), compact (0.1 pc), star-forming core ρ Oph A (Loren et al. 1990) in the Ophiuchus star-forming region.

We began our training set construction by cutting out cubes centered on high-density, core-like gas aggregations from the Cloud Factory in the complexes C and D, so that each cube contains a mass between 5 to 40 $M_\odot$ and a substructure with a gas number density of at least 10^{10} to 10^{11} m^{-3} (i.e. the lower density end of low-mass star-forming regions). In practise, the candidate positions for these cubes are determined by applying a density threshold to column density maps generated for three different (perpendicular) viewing angles of the complexes. We note that beyond these criteria, the cubes were randomly selected and do not represent any specific real-world counterpart star-forming core. For simplicity, we matched the Cloud Factory data to a regular grid. Initially, these cubes are selected on a $64 \times 64 \times 64$ pixel resolution, corresponding to a physical cube size of 0.4 pc. In total, we prepared a sample set of 11 036 individual cubes from The Cloud Factory in this first step. The initial large cube size serves primarily to avoid edge artefacts that can occur in the radiative transfer simulation when synthesising the dust temperature in post-processing for the later synthetic dust emission observations. For the final training data, we actually cropped out the inner $32 \times 32 \times 32$ pixel cubes, corresponding to a total 0.2 pc edge length. This resolution and size were chosen to keep the problem simple for this proof-of-concept and to reduce the overall training data size to facilitate the data handling during the training phase of our approach.

2.2. Synthetic images

To produce synthetic dust emission observations from the selected cloud model cubes, we employed the Monte Carlo (MC) radiative transfer (RT) code POLARIS[1] (Reissl et al. 2016, 2019). POLARIS calculates a dust temperature based on a given 3D density distribution and a specific dust composition, assuming an instantaneous temperature correction and a thermal equilibrium between the dust and its surroundings (for details we refer to Lucy 1999 and Bjorkman & Wood 2001). For the subsequent RT dust heating and emission simulations, we assumed a dust mass to gas mass ratio of $\delta_{\rm gd} = 1\%$ and a material composition of 37.5% graphite and 62.5% (astro)silicate for the grains typical for the ISM. The applied grain sizes are $a \in [5\,{\rm nm}, 250\,{\rm nm}]$ and the number of grains, N, follows a power-law $N(a) \propto a^{-3.5}$ (see e.g. Mathis 1977; Li & Draine 2001, for further details). We emphasise that the selected dust parameters in combination with the average molecular weight of the gas, $\mu_{\rm g}$, define an exact conversion factor from gas to dust density. Because of this, we use the gas density as a measure of the dust density in the following, without explicitly performing the conversion to remain consistent with the Cloud Factory.

We began by preparing the models for two distinct RT setups. In the first one, we considered the dust clouds to be only subject to the diffuse interstellar radiation field (ISRF). Here, we used the parameterisation of Mathis et al. (1983) for the spectral energy distribution with an intensity of $G_0 = 3$, which is typical for star-forming cores (Liseau et al. 2015). In the second scenario, we also added a single star inside the cube in addition to the background ISRF. We used the parameters of a typical B4-type star ($R = 4.33\,R_\odot$, $T_{\rm eff} = 16\,000$ K) in our MC dust heating simulation. We note that the dust MC heating by POLARIS in this step does not modify the ionisation state of the gas or redistribute the gas by means of radiatiave feedback. In each individual cube, we simply placed the B4-analogue star inside the inner $32 \times 32 \times 32$ pixels, selecting a point of low gas density. This procedure roughly emulates the fact that the feedback of such a star would likely clear out its immediate surroundings.

We generated synthetic, monochromatic dust emission observations with a 32×32 pixel resolution (matching the resolution of the underlying dust distribution) at 23 wavelengths between 12 and 1300 µm, matching the central wavelengths of bands available at various observational facilities (see Table A.1 for a full list). We note that we only generate synthetic observations for one viewing angle for each cube. In principle, it would be possible to include multiple viewing angles to increase the size of the training dataset, but given that our simulation suite provides a sufficiently large dataset for training from different physical regions, this is not necessary here. Nevertheless, after training, we also conducted a performance test of our algorithm on an individual region observed from different viewing angles (Appendix B.1), which confirms that the reconstructed 3D structure is nearly identical.

The choice for monochromatic emission observations is again made for simplicity, as modelling the full instrument responses of the considered bands is quite complex and beyond the scope of this proof of concept. Nevertheless, we wanted to select wavelengths that are actually accessible with current observational facilities; thus, we employed the corresponding central wavelengths. Henceforth, we refer to the different wavelengths by the names of the respective instrument bands in the following. We note that we did not consider wavelengths shorter than 12 µm because the influence of scattered light becomes

[1] POLARIS website: https://portia.astrophysik.uni-kiel.de/polaris/

non-negligible in this regime, adding extra complexity. At the current stage of our development, we have not considered instrumental effects related to the point spread functions (PSFs) of the various telescopes or observational noise. Thus, we treat our synthetic observations as fully resolved at all wavelengths and uncertainty free. Properly modelling these effects is not trivial either, particularly with respect to interferometric observations with ALMA, where simulations with a dedicated processing tool such as CASA[2] (CASATeam et al. 2022) would be necessary. Thus, we reserve a proper treatment of these effects for a follow-up work. Still, we want to note that accounting for uncertainties is well within the capabilities of the invertible neural network architecture used in this work, as demonstrated by Kang et al. (2023a).

We initially generated our synthetic dust emission observations assuming a distance of 3.703×10^{18} m (120 pc). To build a more generally applicable approach, we then rescaled the synthetic fluxes following:

$$\hat{f} = f \times \frac{d^2}{d_{\rm ref}^2}, \tag{1}$$

where d denotes the actual distance and $d_{\rm ref}$ is the reference distance (which the flux is scaled to) to determine a distance independent absolute flux measure. The choice of $d_{\rm ref}$ is arbitrary and since we operate on the logarithm of the fluxes in the following (see Sect. 3.3.1) only represents a linear offset to all fluxes, which will not notably affect the training outcome of our neural network approach. For simplicity, we set $d_{\rm ref} = 1$ m, so that the offset is zero in logarithmic space.

Having a complement of observations at 23 different wavelengths for a single real cloud core is typically unrealistic. Given the complexity of the 3D reconstruction task, we started out with this unrealistically large wavelength coverage to emulate a perfect information scenario and determined the best predictive performance our approach could achieve for this proof of concept. In addition, we also investigated a second, more realistically limited scenario, where we considered synthetic observations at the central wavelengths of the following bands: WISE 22 µm, SOFIA 89 µm and 154 µm, *Herschel* PACS 100 µm and 160 µm, *Herschel* SPIRE 350 µm, and LABOCA 870 µm (see Table A.1). This particular combination of bands is inspired by real observational data, which is, for instance, available for the ρ Oph A star forming cloud (Liseau et al. 2015; Santos et al. 2019). We emphasise here that this particular wavelength selection does not necessarily preserve the most information for the given inverse problem and that there may be a much more optimal subset of seven wavelengths among our total of 23 to maximise the reconstructive performance of the approach outlined below. Determining this combination is beyond the scope of this proof of concept and we reserve this to a dedicated follow-up study.

3. Reconstruction approach

To solve the inverse problem of recovering the 3D dust temperature and dust density distribution from the observed dust emission maps, we employed a supervised deep learning approach called an invertible neural network (INN). In the following, we provide a short summary of this methodology and outline our specific setup for the 3D dust reconstruction task.

[2] CASA website: https://casa.nrao.edu/

3.1. The conditional invertible neural network

The INN (Ardizzone et al. 2019a) belongs to the greater family of normalising flows (NFs, Tabak & Vanden-Eijnden 2010; Tabak & Turner 2013; Dinh et al. 2015; Rezende & Mohamed 2015; Kobyzev et al. 2021). More specifically, these are deep learning approaches that model complex distributions through sequences of invertible transformations of simpler known probability distributions (see also Kobyzev et al. 2021, for a review). Among the NF methods, the INN stands as a neural network (NN) architecture that is particularly well suited for solving degenerate inverse problems. Introducing a set of latent variables $\mathbf{z}$ to encode the information loss in the forward mapping $\mathbf{x} \rightarrow \mathbf{y}$ from the physical parameters $\mathbf{x}$ to a set of observables $\mathbf{y}$, which renders the inverse problem $\mathbf{y} \rightarrow \mathbf{x}$ degenerate, the INN can estimate full posterior distributions $p(\mathbf{x}|\mathbf{y})$ for the target parameters. This allows this method to both highlight and in some cases even break degeneracies in solving the inverse problem.

In this study, we employ an INN architecture called conditional invertible neural network (cINN, Ardizzone et al. 2019b). During training, this method learns a mapping of the physical parameters $\mathbf{x}$ to the latent variables $\mathbf{z}$, conditioned on the observables $\mathbf{y}$, that is the forward mapping denoted as:

$$\mathbf{z} = f(\mathbf{x}; \mathbf{c} = \mathbf{y}). \tag{2}$$

In doing so, the cINN encodes all variance of the physical parameters that is not explained by the corresponding observables in the latent variables, while the training process explicitly maintains a prescribed prior distribution $P(\mathbf{z})$ for the latent variables. At prediction time, the cINN can then query this encoded variance by drawing samples from the known prior distribution $P(\mathbf{z})$ of the latent variables and once it has been conditioned on a new query observation $\mathbf{y}'$, it can make use of its fully invertible architecture to generate corresponding samples of the posterior distribution $p(\mathbf{x}|\mathbf{y}')$ following:

$$p(\mathbf{x}|\mathbf{y}') \sim g(\mathbf{z}; \mathbf{c} = \mathbf{y}') \;\; \text{with} \;\; \mathbf{z} \propto P(\mathbf{z}), \tag{3}$$

where $g(\cdot; \mathbf{c}) = f^{-1}(\cdot; \mathbf{c})$ denotes the inverse of the forward mapping, f, for fixed condition, $\mathbf{c}$. For simplicity, $P(\mathbf{z})$ is usually prescribed to be a multivariate normal distribution with zero mean and unit covariance. The dimension of the latent space $\dim(\mathbf{z})$ is per construction equal to the dimension of the target parameter space $\dim(\mathbf{x})$. On the other hand, as the observations are treated as a condition their dimension can become arbitrarily large. In fact, the architecture of the cINN allows for the introduction of a feature extraction network, trained in tandem with the cINN itself, to transform the input observations into a more useful (learned) representation (Ardizzone et al. 2019b).

The invertibility of the cINN is achieved by employing so called conditional affine coupling blocks (Dinh et al. 2017; Ardizzone et al. 2019b). After splitting their input vector $\mathbf{u}$ into two halves $\mathbf{u}_1$ and $\mathbf{u}_2$, these coupling blocks perform two complementary affine transformations:

$$\begin{aligned} \mathbf{v}_1 &= \mathbf{u}_1 \odot \exp\left(s_2(\mathbf{u}_2, \mathbf{c})\right) \oplus t_2(\mathbf{u}_2, \mathbf{c}), \\ \mathbf{v}_2 &= \mathbf{u}_2 \odot \exp\left(s_1(\mathbf{v}_1, \mathbf{c})\right) \oplus t_1(\mathbf{v}_1, \mathbf{c}), \end{aligned} \tag{4}$$

to compute the halves, $\mathbf{v}_1$ and $\mathbf{v}_2$, of the output vector, $\mathbf{v}$, where $\odot$ and $\oplus$ denote elementwise multiplication and addition, respectively. Here, s_i and t_i represent arbitrarily complex transformations of the concatenation of $\mathbf{u}_i/\mathbf{v}_i$ and the conditioning

input **c**. To run the network in reverse, Eq. (4) is then trivially inverted given the output vector $\mathbf{v} = (\mathbf{v_1}, \mathbf{v_2})$ following:

$$\begin{aligned}\mathbf{u}_2 &= (\mathbf{v}_2 \ominus t_1(\mathbf{v}_1, \mathbf{c})) \odot \exp(-s_1(\mathbf{v}_1, \mathbf{c})),\\ \mathbf{u}_1 &= (\mathbf{v}_1 \ominus t_2(\mathbf{u}_2, \mathbf{c})) \odot \exp(-s_2(\mathbf{u}_2, \mathbf{c})),\end{aligned} \tag{5}$$

where $\ominus$ denotes elementwise subtraction. As the transformations s_i and t_i are always evaluated in the same direction in both the forward, Eq. (4) and backward pass, Eq. (5), of the coupling block, it is not necessary to choose them to be invertible themselves. In fact, s_i and t_i do not even need to be prescribed, but can be learned instead during the training of the cINN by representing them with small sub-networks; for instance, a fully connected neural network (Ardizzone et al. 2019a,b). The specific setup of the cINN and the coupling block architecture used in this work is described in Sect. 3.3.

3.2. Single LoS reformulation

The full inverse problem of the 3D reconstruction task consists of predicting the $2 \times N \times N \times N$ hypercube $\mathbf{X}$ of dust densities and temperatures from the $K \times N \times N$ cube, $\mathbf{Y}$, containing the corresponding observed dust emission in the K different wavelengths. Evidently, even for the small resolution of $N = 32$ that we selected, this is a very high-dimensional problem. To simplify our approach and mitigate the difficulties of high dimensionality, we therefore decided to reformulate the inverse problem. In particular, we reduced the 3D reconstruction problem to a matter of individual LoSs under the main assumption that the emission measured in any given pixel is independent of its neighbouring pixels. This independence assumption is especially valid for the perfectly resolved observation scenario that we consider here, but it should also hold (at least to first order) for real observations unless the PSF of the instrument is significantly larger than the pixel size, where smearing could become an issue. With that, we aim to use the vector of the K measured emission fluxes $(e_{\lambda_1}, \ldots, e_{\lambda_K})$ of a given pixel to recover the corresponding dust density $(n_1, \ldots, n_N)$ and temperature $(T_{\mathrm{dust},1}, \ldots, T_{\mathrm{dust},N})$ vectors. To avoid having to train two networks, we further combined the line-of-sight (LoS) dust densities and temperatures into a single vector, formulating the inverse problem as:

$$\begin{aligned}\mathbb{R}^K &\rightarrow \mathbb{R}^{2N}\\ (e_{\lambda_1}, \ldots, e_{\lambda_K}) &\rightarrow (n_1, \ldots, n_N, T_{\mathrm{dust},1}, \ldots, T_{\mathrm{dust},N}),\end{aligned} \tag{6}$$

for a given LoS.

Within the cINN framework, the LoS emission vectors, $\mathbf{y} = (e_{\lambda_1}, \ldots, e_{\lambda_K})$, correspond to the conditioning input and the combined vector of dust densities and temperatures, while $\mathbf{x} = (n_1, \ldots, n_N, T_{\mathrm{dust},1}, \ldots, T_{\mathrm{dust},N})$ denotes the target parameters. Consequently, the latent space introduced by the cINN has a dimension equal to that of $\mathbf{x}$, that is: $2N$.

At prediction time, a given query $K \times N \times N$ cube of emission maps is first decomposed into the N^2 line of sight emission vectors of length K. For each of these LoSs i (with $i \in \{1, \ldots, N^2\}$), the corresponding emission vector, $\mathbf{y}_i$, is then processed by the trained cINN, generating S samples of the full $2N$-dimensional joint posterior distribution $p(\mathbf{x}_i|\mathbf{y}_i)$ by sampling the latent space according to the known Gaussian prior distribution $P(z)$. Afterwards, we reassemble these N^2 LoS prediction results of size $2N \times S$ into the $2 \times N \times N \times N \times S$ hypercube, $\mathbf{X}_{\mathrm{samp}}$, of the density and temperature posterior samples.

It is worth noting that with this LoS decomposition approach, our method is not limited to the 32×32 pixel resolution in the plane of sky, so that larger dust emission maps can be processed as long as they match the physical resolution of 6.25×10^{-3} pc per pixel of the training data. With regard to depth, however, the presented approach is always limited to the 32 pixel depth corresponding to a physical size of 0.2 pc that the cINN is trained on. A possible avenue to create a more depth flexible extension of the method presented here could be to train the cINN on data with varying per pixel resolution and providing the physical resolution as an additional conditioning input. Because this would require a substantially larger training data set and notably increases the complexity of the inverse problem (perhaps even beyond the point of feasibility), we reserve this experiment to our follow-up studies and focus on the fixed depth scenario here. In any case, with the presented approach it is always possible to tailor the training data towards the characteristic sizes of the objects that are to be analysed and train a correspondingly specialised model, as we have done here for the example of very compact, star-forming cores.

Final training data set

Following the prescription of the reformulated inverse problem, we decomposed the $2 \times 11\,036$ training cubes into their respective LoSs, netting a total of $2 \times 11036 \times 32 \times 32 = 22\,601\,728$ vectors. Figure A.1 shows the corresponding effective prior distributions prescribed by the training data for dust density and temperature across all pixels of these lines of sights, as well as a correlation diagram indicating the coverage in the density-temperature space. In particular, we have covered a total density range from 3.3×10^6 to 2.2×10^{13} m^{-3}, although most of the data is concentrated between 10^8 and 10^{12} m^{-3}. The effective prior distribution for dust temperature ranges from 6.3 to 240 K, but is fairly skewed towards the 13 to 24 K interval, so that there are comparatively a lot fewer training pixels above a temperature of 32 K. This is a direct consequence of the fact that such high dust temperatures only occur in the relatively few pixels in the vicinity of a star. Given that half of our training cubes do not contain a star, the per-pixel dust temperature prior distribution is naturally biased towards this intermediate dust temperature regime because there are simply much more pixels that are either only subject to the ISRF to begin with or far enough away from the star to avoid being heated to very high temperatures. A corresponding diagram of the prior distributions of the measured fluxes at the 23 considered wavelengths across all pixels is provided in Fig. A.2. We emphasise that as a data-driven approach, the cINN is mostly limited to the parameter space covered by the training data. Although the cINN does exhibit some capability for extrapolation beyond the limits of the learned parameter space, there is in general no guarantee for a (physically) sound prediction outcome for inputs and targets that fall outside the described ranges.

We further split this data set randomly into a training (80% of the data) and test set (20% of the data). The latter serves as held-out data that is not seen during training of the cINN to later evaluate the convergence and performance of the model. While the split is in general randomly chosen, we make sure that the held-out test set contains a subset of 100 complete cubes. For this subset, we selected the same 50 cubes twice: once subject solely to the ISRF and once in the ISRF + star configuration. The aim is to evaluate how much the radiation setup affects the prediction outcome for the dust density and temperature. While we verified the model convergence on the greater test data set, the reported performance and all diagrams presented in Sect. 4 are based on this subset of 100 coherent cubes. Although this set of

$100 \times 32 \times 32 = 102\,400$ LoSs only represents 0.5% of the total data, it has been selected as a representative subset of the test data set in order to keep the memory requirements at a manageable level. For instance, storing the predicted posterior samples for these 100 cubes as an uncompressed csv table following the setup outlined further below already requires ~ 360 GB of memory.

3.3. Implementation details

We employed the Python deep learning module PYTORCH (Paszke et al. 2017) and the dedicated Framework for Easily Invertible Architectures (FrEIA[3], Ardizzone et al. 2019a,b) package to implement the cINN approach. For the affine coupling blocks, we employed the Generative Flow (GLOW; Kingma & Dhariwal 2018) configuration, in which the transformations s_1, t_1 and s_2, t_2 were jointly estimated by one sub-network each, which reduces the number of sub-networks in each coupling block from four to two. As sub-networks we utilised simple, fully connected networks with three layers of size 1024 and the rectified linear unit (ReLU) activation function. As in Ardizzone et al. (2019b), we also introduced a clamping procedure in the affine transformations in Eqs. (4) and (5) to the argument, s, of the exponential functions of the form:

$$s_{\text{clamp}} = \frac{2\alpha}{\pi} \arctan\left(\frac{s}{\alpha}\right), \tag{7}$$

with $\alpha = 1.9$. This procedure avoids instabilities arising from exploding magnitudes of $\exp(s)$. Furthermore, we alternated the affine coupling layers with random permutation layers, which randomly (but in a fixed and thus invertible manner) permute the output vector between each coupling layer to better intermix the information between the two streams u_1 and u_2. Our final network architecture (as determined via hyperparameter optimisation) is made up of nine coupling blocks in total. We also employed a simple feature extraction network, consisting of a three-layer (with 512, 512, and 256 nodes, respectively) fully connected network with ReLU activation functions, trained jointly with the cINN, to process the input observations.

3.3.1. Additional data preprocessing

Prior to training, we converted both the dust density and temperature to logarithmic space. This serves to prevent issues during training that can occur when the target parameters have a large dynamic range. This is particularly notable in the case of the dust density, which covers almost seven orders of magnitude. In addition, this implicitly ensures that the predicted dust densities and temperatures are always strictly positive. Afterwards, we performed two linear scaling operations on the training data. Each element x_i of the target parameters, $\mathbf{x}$, was rescaled by subtraction of its mean (over the entire training set) and then by division by its standard deviation, so that the resulting distribution of the rescaled $\hat{x}_i$ has zero mean and unit standard deviation. For the observables, we applied a matrix whitening procedure (Hyvärinen & Oja 2000) to the $M \times K$ matrix of training observations, $\mathbf{Y}$, where M is the number of training examples and K is the dimension of a single observation, $\mathbf{y}$, such that the rescaled observable matrix, $\hat{\mathbf{Y}}$, has a unit covariance matrix. Given they are linear transformations, these scaling operations are easily inverted to convert the cINN output back to the true target parameter space. The coefficients of these scaling operations were determined on the training data and at the prediction time applied in the same fashion to the new query input.

3.3.2. Training setup and sampling strategy

We trained our cINN approach via minimisation of the maximum likelihood loss, $\mathcal{L}$, as described in Ardizzone et al. (2019b), namely:

$$\mathcal{L} = \mathbb{E}_i \left(\frac{\|f(\mathbf{x}_i; \mathbf{c}_i, \Theta)\|_2^2}{2} - J_i \right), \tag{8}$$

where $J_i = \det(\partial f / \partial \mathbf{x}|_{\mathbf{x}=\mathbf{x}_i})$ denotes the determinant of the Jacobian matrix evaluated at training instance $\mathbf{x}_i$ and Θ represents the network weights. During training, the network weights, Θ, that minimise the loss function, $\mathcal{L}$, are determined using a standard stochastic gradient descent approach. This means that after making an initial random guess for the weights, they are iteratively updated in the direction of the gradient $\nabla_\Theta \mathcal{L}$ based on randomly drawn subsets (batches) of the training data until a convergence is reached. In particular, we employ the adaptive learning rate, momentum-based Adam (adaptive moment, Kingma & Dhariwal 2018) optimiser for this purpose (with $\beta_1 = \beta_2 = 0.8$). Here, we start with an initial learning rate (for Adam this is a scaling factor for the adaptive step size in the weight updates along the loss gradient) of $l_{\text{init}} = 9.642 \times 10^{-5}$ and then we reduced it by a factor of $\gamma = 0.831$ every 11 epochs. In total, our models were trained for 250 epochs, using a batch size of 512 and processing 4096 batches per epoch. We also employed an L2 weight regularisation with $\lambda = 6.093 \times 10^{-5}$. This setup was determined via hyperparameter optimisation, using the Hyperband algorithm (Li et al. 2018), a procedure that combines a random grid search approach with adaptive resource allocation and an early stopping criterion. Hyperband provides an efficient framework to test a large number of (randomly generated) hyperparameter configurations that finds a balance between running the training in full only for configurations that appear promising early (i.e. converge fast), while also allowing for some slower converging models that might reach a better final result. For more details on the logistics of Hyperband we refer to Li et al. (2018). Training a single network with the final setup described above takes about 19 h using GPU acceleration on a NVIDIA RTX 2080Ti graphics card.

At the prediction time, we then generated $S = 4096$ posterior samples for each new query LoS. This number of samples is chosen as a compromise between storage requirements and sample density, although experiments with even larger sample numbers have actually not shown a notable difference in the predicted posterior distributions, so this did not seem necessary within the framework of our analysis. A trained cINN can generate this amount of samples for 1024 LoSs (that is a single cube) in about 28 s (on a NVIDIA RTX 2080Ti), making the inference of the posterior distributions of the dust properties very efficient.

3.4. Making point estimates

To better compare the cINN predictions to the ground truth hypercubes, $\mathbf{X}$, in our synthetic test set, we computed a point estimate $\hat{\mathbf{X}}$ from the hypercube of posterior distribution samples, $\mathbf{X}_{\text{samp}}$, returned by the cINN. The most straightforward approach for this is to derive the maximum a posteriori (MAP) prediction values for the dust density and temperature in every pixel of the 3D cube, which consists of determining the most likely

[3] Available at https://github.com/vislearn/FrEIA

value of the target parameters from the corresponding posterior distribution. In the following, we describe how we tested two methods for computing the MAP estimate given the predicted posterior distribution from the cINN.

In the first approach, we treated the posterior distributions for density and temperature of each pixel individually, marginalising over all other pixels along the LoS, for which the cINN generated samples of the joint posterior distribution. From the corresponding set of posterior samples for each pixel, we identified the MAP estimate for density and temperature by employing a kernel density estimate (KDE) to first explicitly derive the probability density curve of the posterior and then find the peak of this curve. In practise, we used a Gaussian kernel function, determining the kernel bandwidth automatically with Silverman's rule of thumb (Silverman 1986) and evaluating it on an evenly spaced grid of 1024 points (between the minimum and maximum value of the posterior samples) to determine the MAP point estimates.

The cINN does not actually generate samples from the posterior distributions of dust density and temperature of each individual pixel but, rather, from the full joint posterior for density and temperature for all pixels along a given LoS. Therefore, to be completely correct, the MAP has to be determined as the most probable combination of values in the full 64-dimensional space that the cINN constructs the posterior samples in. To find the maximum of the probability density in this very high-dimensional space and then compare it to the marginalised MAP estimate, we employed the MeanShift algorithm (Fukunaga & Hostetler 1975; Comaniciu & Meer 2002). It is a gradient ascent approach whereby, given a set of N samples, the modes of the underlying density distribution can be found. MeanShift is an iterative procedure, in which the center of a kernel window is continuously moved into the direction of the maximum increase in density until convergence is reached. Given a kernel function $K(\mathbf{x})$ (e.g. a Gaussian kernel) and an initial position for the center $\mathbf{x}$ of the kernel window, the algorithm computes the so-called mean shift:

$$\mathbf{m}(\mathbf{x}) = \frac{\sum_{i=1}^{N} K\left(\mathbf{x} - \mathbf{x_i}\right) \mathbf{x}_i}{\sum_{i=1}^{N} K\left(\mathbf{x} - \mathbf{x_i}\right)} - \mathbf{x}, \tag{9}$$

which is the difference between the kernel weighted mean and the center of the kernel window. As demonstrated by Comaniciu & Meer (2002), this vector is proportional to the estimate of the density gradient estimate obtained with the same kernel; thus, it always points in the direction of maximum increase in density. Iteratively translating the kernel window in direction of the mean shift will therefore find a (local) maximum for the underlying density distribution (Comaniciu & Meer 2002). To find all modes of the distribution (and ideally the global maximum) this approach is then repeated for other initial kernel positions, scoring the identified peaks by their corresponding (kernel) density estimate. In a post-processing step, any spurious mode detections (such as plateaus in the distribution) or very close-by modes can then be further pruned (see for example Comaniciu & Meer 2002, for further details).

In practise, we employed the scikit-learn (Pedregosa et al. 2011) Python implementation of the MeanShift algorithm, which uses a flat kernel:

$$K(\mathbf{x}) = \begin{cases} 1 \text{ if } \|\mathbf{x}\| \leq \lambda \\ 0 \text{ if } \|\mathbf{x}\| > \lambda, \end{cases} \tag{10}$$

where λ denotes the bandwidth. To speed up computation, this implementation provides a binned seeding strategy, where the initial guesses for the kernel starting position are selected on a discretised grid instead of testing all of the individual sample points. The coarseness of this grid is determined by the bandwidth selected for the kernel. The automatic bandwidth selection that comes with this implementation (based on a nearest neighbour distance estimation) has, however, proven not to be robust enough for our very high-dimensional parameter space and would often select bandwidths that are too small for the kernel windows to find any data points inside of them (when used in combination with the binned-seeding approach). Since the computation time becomes prohibitively large without the binned seeding, we adopted a simple bandwidth selection procedure where we iteratively doubled an initial bandwidth guess of 32 until a bandwidth is found, with which the MeanShift algorithm converges. In practise, this simple approach leads to the selection of a bandwidth of 64 or 128 in most cases.

3.5. Spatial consistency

As we outline above, the cINN approach predicts the posterior distributions for density and temperature for a single LoS jointly. Consequently, the prediction preserves the consistency of the predicted posteriors along the LoS. Perpendicular to the LoS, however, we have (by construction) no such spatial consistency guarantee. Figure 1 provides an example of this behaviour, highlighting the gradual shift of the posteriors along the LoS, whereas perpendicular to it, they are not necessarily consistent. As a consequence, the MAP or MeanShift point estimates can often exhibit sharp discontinuities in the predicted densities and temperatures. This can be seen, for example, in the MAP estimates in the left panels of Fig. 1. As these discontinuities and sharp jumps are rather unphysical, we experimented with two approaches in order to mitigate the spatial consistency issue.

3.5.1. MNPCP point estimator

Our first approach consists of introducing a third, alternative point estimator that enforces a degree of spatial consistency perpendicular to the LoSs, which we refer to as the median neighbour pixel combined posterior (MNPCP) in the following. Figure 2 outlines the steps of the MNPCP approach. Looping over all pixels in the 3D cube of generated posterior samples, $\mathbf{X}_{\text{samp}}$, we first collected the samples for the current pixel and its 26 neighbouring pixels. We then determined the n and T_{dust} point estimates for the current pixel as the weighted median of this combined set of posterior samples. Here, each sample has been weighted according to the distance of the pixel to the query pixel using the city-block distance metric (Manhattan distance). For edge cases, we accumulate only samples from the existing neighbour pixels, meaning that no form of padding was applied. Taking, for example, a corner pixel, this means that samples from only the seven neighbour pixels are accumulated, as compared to the 26 neighbours available for an interior pixel.

3.5.2. Neighbour LoS reformulation

Aside from introducing an alternative point estimator to combat the spatial consistency issue, we also investigated whether a different reformulation of the inverse problem may improve the situation, in comparison to our primary formulation (introduced in Sect. 3.2). We refer to cINNs trained on the primary reformulation as a single LoS cINN (SLoS-cINN) in the following, whereas models for the alternative formulation outlined below shall be denoted as a neighbour LoS cINN (NLoS-cINN). To directly compare the SLoS and NLoS approaches, we tested both of them with all three introduced point estimators, namely: MAP, MeanShift, and MNPCP.

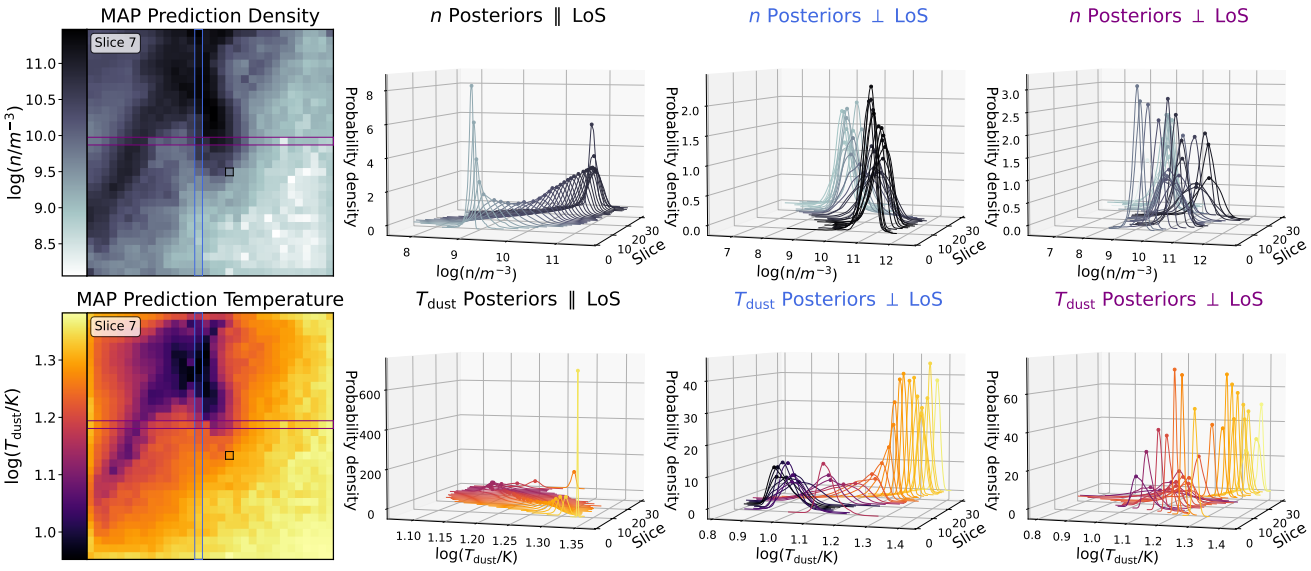

Fig. 1. Comparison of the predicted posterior distributions along vs. perpendicular to the LoS (into the plane). The left column shows an example slice with the MAP estimates for dust density (top) and dust temperature (bottom). The other three columns show the posterior distributions of dust density (top) and temperature (bottom) for the lines indicated in the left panels in black, blue and purple, respectively. Here, the black square denotes a LoS going into the plane of the image, whereas the blue and purple lines are perpendicular to the LoS along the x and y axis, respectively.

The NLoS reformulation aims at improving the spatial consistency perpendicular to the LoS by adding information of the neighbouring LoSs to the observables. Instead of taking only the vector of fluxes corresponding to the pixel of a given LoS, we go on to also consider the observed dust emission in the eight neighbouring pixels, so that the inverse problem becomes:

$$\begin{aligned} \mathbb{R}^{9K} &\rightarrow \mathbb{R}^{2N} \\ (\mathbf{y}_1, \ldots, \mathbf{y}_9) &\rightarrow (n_1, \ldots, n_N, T_{\mathrm{dust},1}, \ldots, T_{\mathrm{dust},N}), \end{aligned} \tag{11}$$

where $\mathbf{y}_i = (e_{i,\lambda_1}, \ldots, e_{i,\lambda_K})$ denote the nine emission flux vectors corresponding to the pixel ($\mathbf{y}_5$) and its neighbourhood. This reformulation has one immediate drawback, however, in that we lose some of the available training data. As we now require every LoS to have eight neighbours, we can no longer consider the edge cases in our training cubes, reducing the total amount of LoSs available for training data to $2 \times 11036 \times 30 \times 30 = 19\,864\,800$. Another disadvantage is a notably increased memory requirement when storing the training data as a simple csv-table, since we increased the size of the observables vector by a factor of 9. In our case, the table size increases from 55 to 138 GB, even though the latter set contains 2 736 928 fewer LoSs. Nevertheless, this is the most straightforward approach to providing the cINN with information on the vicinity of a given query pixel.

3.6. Performance evaluation

To quantify the overall performance on the held-out test set of 100 coherent cubes, we computed two metrics for the three different point estimation approaches as an average over all $N_{\mathrm{test}} = 100 \times 32 \times 32 \times 32 = 3\,276\,800$ test pixels. The first one is the normalised root mean squared error (NRMSE), defined as:

$$\mathrm{NRMSE} = \frac{1}{\Delta x_{\mathrm{TS}}} \sqrt{\frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \left(x_{\mathrm{i,pred}} - x_{\mathrm{i,true}}\right)^2}, \tag{12}$$

where $x_{\mathrm{i,true}}$ and $x_{\mathrm{i,pred}}$ refer to the ground truth and point estimate prediction of target parameter x for pixel i, and $\Delta x_{\mathrm{TS}} = \max(x_{\mathrm{TS}}) - \min(x_{\mathrm{TS}})$ denotes the range of target parameter, x, in the training data (6.82 and 1.58 for $\log(n/\mathrm{m}^{-3})$ and $\log(T_{\mathrm{dust}}/\mathrm{K})$, respectively). The second metric that we computed is the median $|\bar{e}_{\mathrm{rel}}|$ (and 25% and 75% quantiles) of the absolute relative error $|e_{\mathrm{i,rel}}|$, defined as:

$$|e_{\mathrm{i,rel}}| = \left| \frac{x_{\mathrm{i,pred}} - x_{\mathrm{i,true}}}{x_{\mathrm{i,true}}} \right|, \tag{13}$$

for pixel i.

4. Results

In this section, we outline the evaluation results regarding the predictive performance of our trained cINN models on the held-out 100 test cubes. Table 1 provides a summary of the NRMSE and absolute relative errors achieved by our three different cINN setups with the three different point estimation approaches. In addition, it also shows a breakdown of the results between the two radiative transfer configurations, that is, ISRF-only and ISRF + star. In the following, we first discuss the influence of the point estimator choice on the prediction results on the example of the SLoS-cINN that accounts for all 23 wavelengths (Sect. 4.1). We then compare the outcomes of the NLoS approach to the SLoS setup (Sect. 4.2) and present an analysis of the SLoS performance for the more realistically limited wavelength coverage experiment (Sect. 4.3). We conclude with a comparison of our approach with a classical SED fit to determine column densities (Sect. 4.4), followed by discussions on the physical feasibility of the approach (Sect. 4.5) and on the application of our setup to real observational data (Sect. 4.6).

4.1. Choice of the point estimator and influence of the radiation configuration

Figure 3 shows a qualitative comparison of the point estimates for dust density and temperature to the ground truth for the MAP, MeanShift, and MNPCP estimators. In particular, we show the

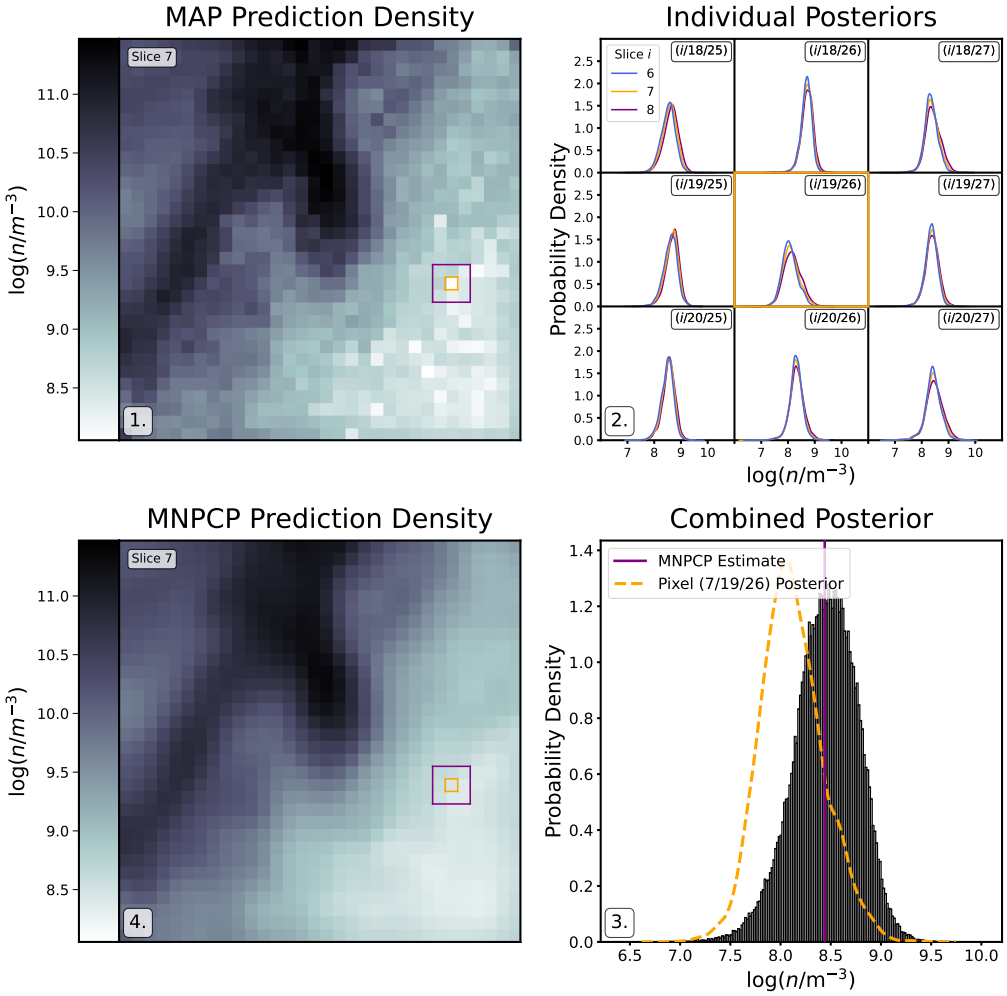

Fig. 2. Schematic outline of the median neighbour pixel combined posterior approach for point estimations based on the example of dust density. The procedure follows the panels from top left to bottom left in clockwise order. The top-left panel shows the MAP density estimates for a single slice of a dust cube (perpendicular to the LoS), highlighting the discontinuities that can occur in the MAP estimate. The query pixel, for which the MNPCP estimate will be computed, and its neighbourhood are indicated in orange and purple, respectively. The top-right panel shows the predicted posterior distributions for the dust density for the query pixel with index 7/11/8 (center subpanel indicated by the orange outline) and all 26 neighbouring pixels. The colours of the curves indicate the slice, *i*, that they come from, whereas the indices in the top right corner denote the pixel index perpendicular to the LoS. The bottom-right panel shows a histogram of the all posterior samples accumulated from the query pixel and its neighbours. Here, the orange dotted curve indicates the original (marginalised 1D) posterior distribution of the query pixel and the purple line marks the MNPCP estimate, which is the city-block distance weighted median of all posterior samples (the distance in pixels when allowing only right angle moves, no diagonals). Finally, the bottom-left panel presents the MNPCP dust density estimates of the slice shown in the top left with the orange and purple boxes indicating again the query pixel and its neighbourhood.

outcome for a single slice (perpendicular to the LoS) of one example cube of the test set. Here, the first four rows show the prediction results based on the SLoS-cINN, distinguishing the ISRF-only scenario (rows 1 and 2) and the ISRF + star radiation configuration (rows 3 and 4). In the ISRF scenario, we can see that all three point estimators provide a very decent reconstruction result for both dust density and temperature. However, the discontinuities in the MAP prediction (as previously discussed in Sect. 3) are quite notable and give the prediction outcome a noisy character. The MeanShift result in the third column also suffers from this effect, albeit to a slightly lesser degree. Given that both of these estimators have no spatial consistency guarantee perpendicular to the LoS, this is of course an expected result. Contrary to that, we can see that our MNPCP approach (fourth column) provides a much more consistent and smoothed prediction result than the other two estimators. Nevertheless, this can come at the expense of losing some of the finer, high-density features of the dust distribution. A full comparison of the prediction with the SLoS-cINN for all slices of this example cube in the ISRF-only configuration is given in Fig. B.1. We also refer to Fig. B.2 for a 3D visualisation of the prediction results in terms of isodensity surfaces.

The predictions for the same cube with the alternative radiation setup (rows 3 and 4, Fig. 3) indicate that the inclusion of a star affects not only the prediction of the dust temperature (as expected given the additional heating from the star), but the dust density estimates as well. Although the overall reconstruction of both dust density and temperature remain quite good in this example, it is obvious that the reconstructed dust density has lost accuracy in comparison to the cINN prediction for the same cube subject to only the ISRF. While the overall larger scale structures are still recovered well, a lot of the finer details of the dust distribution are lost compared to the prediction in the ISRF-only scenario. The inclusion of a star inside of the cube thus appears

Table 1. Summary of the predictive performance for our three different cINN setups and the three different point estimation methods.

				cINN		
Cube selection	Point estimator	Measure	Parameter	Single LoS	Neighbour LoS	Single LoS (7 wavelengths)
All cubes	MAP	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0707	0.0608	0.0981
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0300	0.0260	0.0458
		$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$1.85^{+1.99}_{-1.07}$	$1.52^{+1.62}_{-0.89}$	$2.61^{+3.13}_{-1.50}$
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	$0.87^{+1.39}_{-0.56}$	$0.75^{+1.15}_{-0.48}$	$1.47^{+2.49}_{-0.94}$
	MeanShift	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0640	0.0555	0.0855
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0251	0.0221	0.0372
		$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$2.01^{+1.87}_{-1.15}$	$1.63^{+1.57}_{-0.93}$	$3.10^{+2.52}_{-1.68}$
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	$1.01^{+1.39}_{-0.64}$	$0.84^{+1.19}_{-0.53}$	$1.97^{+2.04}_{-1.14}$
	MNPCP	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0620	0.0536	0.0838
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0245	0.0213	0.0359
		$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$1.84^{+1.78}_{-1.05}$	$1.54^{+1.48}_{-0.88}$	$2.81^{+2.47}_{-1.54}$
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	$0.96^{+1.26}_{-0.58}$	$0.82^{+1.07}_{-0.49}$	$1.72^{+1.89}_{-0.99}$
ISRF-only	MAP	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0510	0.0441	0.0917
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0633	0.0550	0.1135
		$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$1.41^{+1.45}_{-0.83}$	$1.18^{+1.20}_{-0.69}$	$2.25^{2.82}_{-1.29}$
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	$0.61^{+1.03}_{-0.39}$	$0.53^{+0.89}_{-0.34}$	$1.33^{+2.54}_{-0.87}$
	MeanShift	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0480	0.0409	0.0804
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0522	0.0455	0.0906
		$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$1.53^{+1.42}_{-0.88}$	$1.24^{+1.18}_{-0.72}$	$2.86^{+2.40}_{-1.56}$
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	$0.70^{+1.12}_{-0.45}$	$0.58^{+0.95}_{-0.37}$	$1.90^{+2.08}_{-1.12}$
	MNPCP	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0456	0.0395	0.0772
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0506	0.0439	0.0859
		$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$1.40^{+1.33}_{-0.80}$	$1.19^{+1.12}_{-0.69}$	$2.51^{+2.26}_{-1.38}$
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	$0.70^{+1.02}_{-0.42}$	$0.60^{+0.88}_{-0.36}$	$1.64^{+1.93}_{-0.97}$
ISRF + star	MAP	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0859	0.0739	0.1041
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0387	0.0336	0.0538
		$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$2.45^{+2.55}_{-1.38}$	$1.99^{+2.05}_{-1.12}$	$3.01^{+3.34}_{-1.71}$
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	$1.20^{+1.67}_{-0.74}$	$1.02^{+1.32}_{-0.62}$	$1.61^{+2.43}_{-0.98}$
	MeanShift	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0767	0.0670	0.0903
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0327	0.0289	0.0441
		$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$2.66^{+2.19}_{-1.44}$	$2.15^{+1.89}_{-1.18}$	$3.35^{+2.61}_{-1.78}$
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	$1.38^{+1.50}_{-0.80}$	$1.14^{+1.29}_{-0.67}$	$2.05^{+2.01}_{-1.15}$
	MNPCP	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0748	0.0648	0.0899
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0320	0.0279	0.0431
		$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$2.45^{+2.13}_{-1.34}$	$1.99^{+1.80}_{-1.10}$	$3.13^{+2.64}_{-1.69}$
			$\log(T_{\mathrm{dust}}/\mathrm{K})$	$1.27^{+1.39}_{-0.73}$	$1.08^{+1.16}_{-0.61}$	$1.79^{+1.85}_{-1.01}$

Notes. Listed are respectively the NRMSE and the median absolute relative error $|\bar{e}_{\mathrm{rel}}|$ (along with the 25% and 75% quantiles) for the dust density and temperature evaluated across all pixels of the set of test cubes indicated in the first column.

to add not only complexity to the prediction of the dust temperature, but also renders the recovery of the density more difficult. For a complete comparison of all slices of the example cube in the ISRF + star radiation configuration, we refer to Fig. B.4. The corresponding 3D isodensity surface visualisation is provided in Fig. B.5. To provide additional insights into the difference between the predictions on the ISRF-only and ISRF+star cube, Fig. B.3 presents an extension to Fig. 3, where some examples of the predicted posterior distributions are shown. Here, we find that the density posterior distributions become notably wider in the RT scenario that includes the star, which results in flat, plateau-like or even multi-peaked distributions for some pixels. In the latter cases, the peaks of the distributions may then no longer coincide with the ground truth (although the ground truth

is always part of the distribution); for instance, the MAP estimator returns a suboptimal result in comparison to the prediction for the same cube in the ISRF-only scenario. There are two possible explanations for the broadening of the density posterior distributions in the second RT scenario. The first is that this an intrinsic degeneracy of the problem. While the cubes share the same density distribution between the two RT setups, the resulting dust temperatures naturally differ. It is not unreasonable to believe that recovering the density can become more or less ambiguous depending on the temperature given that opacity depends on temperature. It is also worth noting that the temperature posterior distributions do not suffer the same broadening effect and appear similarly well constrained in both RT configurations. The second possible explanation is a suboptimal convergence of the network related to the sampling of the training data, which is discussed in more detail further below.

To quantify the overall performance of the SLoS-cINN in combination with the three different point estimators, we present a direct one-to-one comparison of the predicted dust densities and temperatures to the respective ground truth values, and the corresponding relative residuals for all 100×32^3 pixels in the test set in Fig. 4. As indicated by Table 1 and Fig. 4, the overall predictive performance of the SLoS-cINN in the 23-wavelength configuration is quite excellent across all three point estimation approaches. Here, we achieved NRMSEs between 0.06 to 0.07 in $\log(n/\mathrm{m}^{-3})$ and between 0.025 to 0.03 in $\log(T_{\mathrm{dust}}/\mathrm{K})$, corresponding to median absolute relative residuals, $|\bar{e}_{\mathrm{rel}}|$, in the ranges of 1.8 to 2.0% and 0.9 to 1.0%, respectively. Although there is a notable dispersion around a perfect one-to-one correlation between the estimated densities and temperatures and the corresponding ground truth, the binned median curve (and binned 25 and 75% quantile curves) in the relative residual diagrams indicates that a majority of pixels is indeed close to a perfect recovery for most of the range covered in density and temperature. Nevertheless, the relative residuals can reach up to 40 to 50% for a small number of individual pixels in terms of both density and temperature.

The binned median relative residual curves also highlight some systematic trends in the point estimation outcome. For the dust density, we find a notable trend towards overestimation of the density for pixels with a true density below $\log(n/\mathrm{m}^{-3}) = 8$. There is also a tendency (although to a lesser extent) for underestimations at the high density end, starting at $\log(n/\mathrm{m}^{-3}) = 11$. For the dust temperature, we also observe a systematic underestimation at temperatures higher than 100 K, and for the MeanShift and MNPCP point estimates a slight tendency for overestimation below 10 K. It appears that the recovery of dust density and temperature in terms of the point estimation tends to struggle more overall towards the extreme ends of the respective parameter range. Part of this difficulty can likely be attributed to the relative complexity of these more extreme environments, but there might be a more direct issue with our training data that could explain this decrease in the predictive performance at the edges of the parameter space. Figure A.2 shows the prior distributions of dust density and temperature across all pixels in our training set. Comparing the thresholds at which the binned median relative error starts to show systematic offsets in Fig. 4, that is, $\log(n/\mathrm{m}^{-3}) < 8$, $\log(n/\mathrm{m}^{-3}) > 11$ and $\log(T_{\mathrm{dust}}/K) > 2$, to the training set priors, we can see that there are comparatively a lot fewer pixels in these parameter ranges in the training data. This (relative) lack of training examples within these parameter ranges could lead to a suboptimal convergence of the cINN, so that it does not achieve the same robustness at the edges of the parameter space as it does within the intervals where a lot of training data is available. Given that the prior distributions of the dust densities and temperatures in our training data are (in part) dictated by the underlying dust cloud simulations, achieving a more even sampling across the parameter spaces is not a trivial matter and at this stage, this is beyond the scope of this proof of concept. It is also worth noting that very high temperature regions ($T_{\mathrm{dust}} > 100\,\mathrm{K}$) are both rare in reality and likely affected by strong feedback, being either part of an HII region or related to strong outflow activity. This adds further complexity to these extreme environments, which is also currently not accounted for in the Cloud Factory as noted in Sect. 2. We plan to investigate these effects and the sampling strategy further in future optimisation of our training set generation.

In comparing the three point estimators more in detail, we find that both MeanShift and MNPCP are less prone to large outlier values, as evident by the smaller dispersion around the one-to-one correlation in Fig. 4 and the lower NRMSE in Table 1. At the same time, it is the MAP estimator that returns the overall best $|\bar{e}_{\mathrm{rel}}|$ with 1.85% for $\log(n/\mathrm{m}^{-3})$ and 0.87% for $\log(T_{\mathrm{dust}}/\mathrm{K})$. Although the MNPCP estimator has a nominally better result for $|\bar{e}_{\mathrm{rel}}|$ with 1.84% for $\log(n/\mathrm{m}^{-3})$, it incurs a notably larger error in $\log(T_{\mathrm{dust}}/\mathrm{K})$ with 0.96%. The latter small performance decrease of the MNPCP approach in terms of $|\bar{e}_{\mathrm{rel}}|$ is likely a result of the effective smoothing that this estimator performs. As Fig. 4 shows, this enhances the underestimation tendencies at the high temperature end (also for high density but to a lesser degree) and, thus, the average error. For the MeanShift, which performs the worst in terms of $|\bar{e}_{\mathrm{rel}}|$, this might be a consequence of a suboptimally chosen bandwidth from our simplified bandwidth selection procedure (as described in Sect. 3). If, for instance, the selected bandwidth is too large, the MeanShift kernel will overly smooth the density distribution and likely miss narrow peaks. This results in a comparable (over-) smoothing effect to the MNPCP, as evidenced by the similar behaviour of the two methods in Fig. 4.

Figure 5 provides a breakdown of Fig. 4 for the best and worst case prediction outcomes, that is the five cubes with the best and five cubes with the worst NRMSE in the MAP point estimate. Averaged over the five best cubes $|\bar{e}_{\mathrm{rel}}|$ goes down to about 1% and 0.5% in $\log(n/\mathrm{m}^{-3})$ and $\log(T_{\mathrm{dust}}/\mathrm{K})$, respectively. In the worst cases on the other hand, $|\bar{e}_{\mathrm{rel}}|$ reaches up to 4.4% and 1.7% in $\log(n/\mathrm{m}^{-3})$ and $\log(T_{\mathrm{dust}}/\mathrm{K})$, respectively. What is interesting to note here is that the five best cubes are all only subject to the ISRF, whereas the five worst ones are all in the ISRF + star radiation configuration. This reaffirms our earlier assessment that the presence of a star in the cube notably complicates the problem. We further quantified this by breaking down the overall performance on the test set between the two radiation setups in Table 1. As we can see, $|\bar{e}_{\mathrm{rel}}|$ increases by almost a factor of two for the cubes with ISRF + star setup in comparison to the ISRF-only configuration cubes regardless of the choice of the point estimator. We also want to emphasise that the observed performance is not dependent on the selected viewing angle of the cubes. We have confirmed in a test limited to the 50 ISRF-only cubes that the cINN returns a similarly excellent reconstructive performance, when the cubes are observed from different directions. Thus, our choice to only generate synthetic observations from one direction in the training data has not introduced a bias in the form of a preferred viewing angle (for more details, see Appendix B.1).

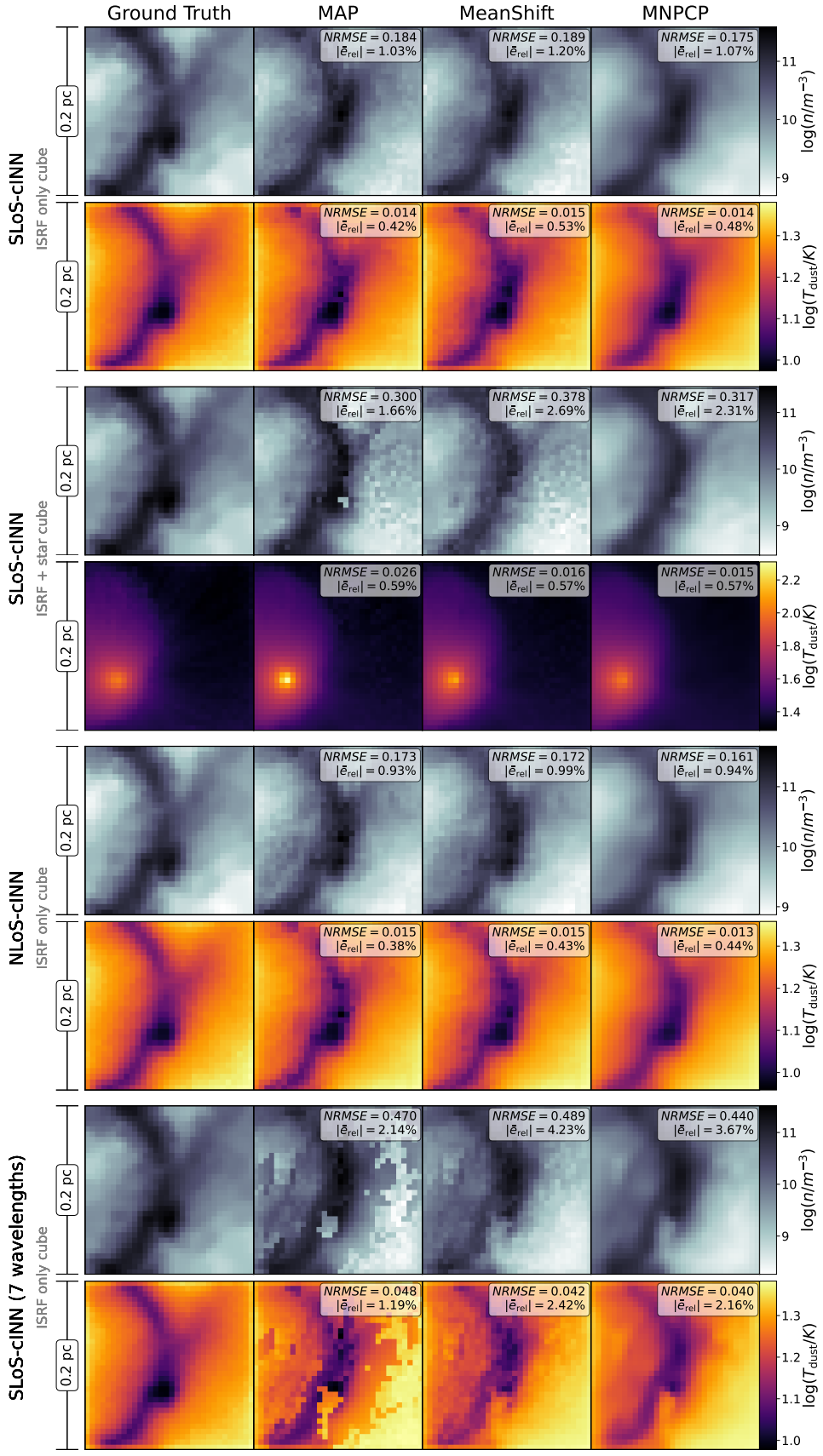

Fig. 3. Predicted dust densities and temperatures for a single example cube slice perpendicular to the LoS. A comparison between the three point estimation methods to the ground truth is shown in the left column. The top two rows give the 23-wavelength SLoS-cINN result for the ISRF-only scenario, while rows 3 and 4 display the counterpart for the ISRF + star case. Rows 5 and 6 present the NLoS-based outcome for the ISRF-only case, and the last two rows show the corresponding seven wavelengths SLoS-cINN prediction. The listed NRMSE and median absolute relative errors are averages over this slice only and not the entire cube.

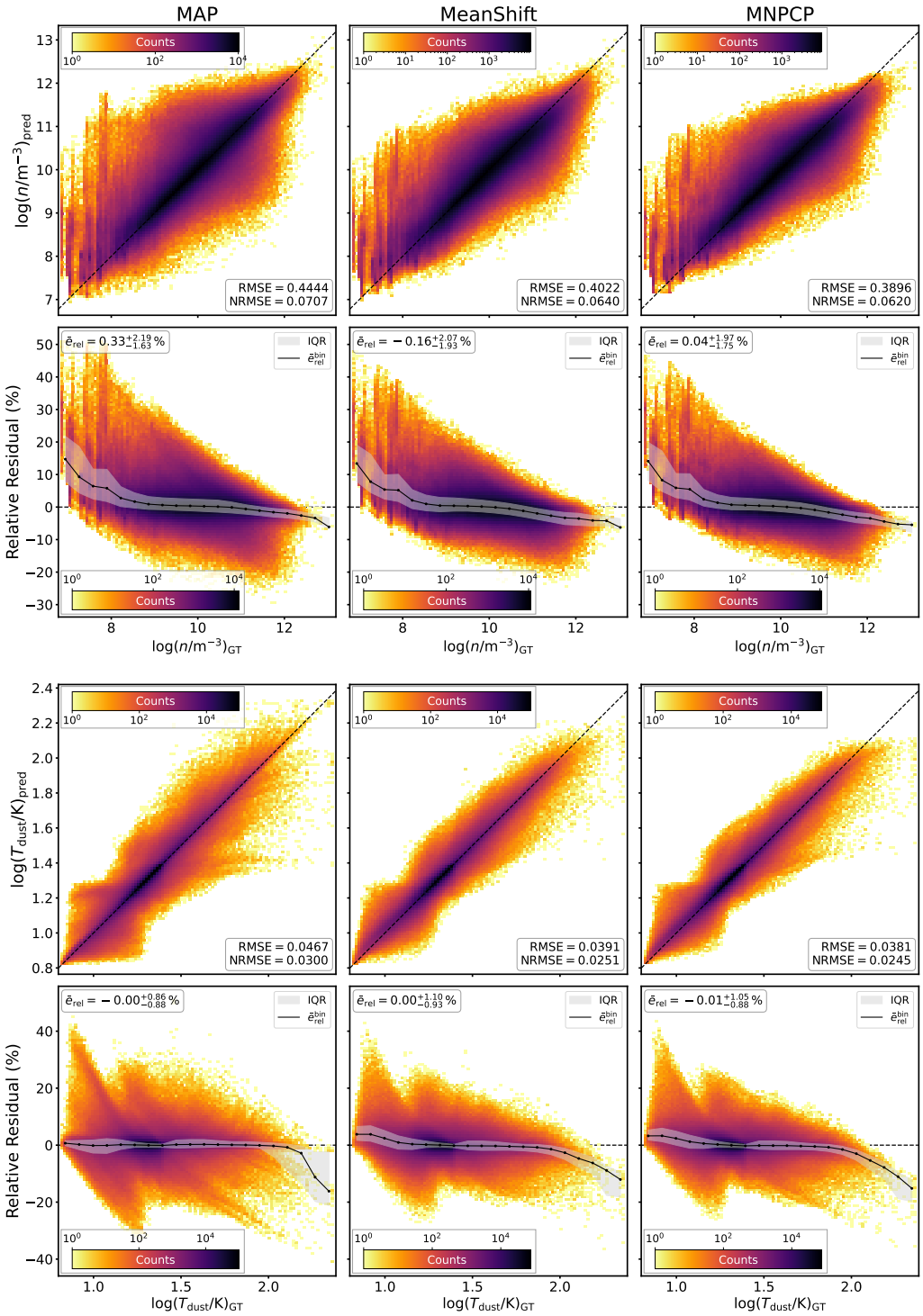

Fig. 4. Performance breakdown of the SLoS-cINN using 23 wavelengths. 2D histograms comparing the cINN predictions for dust density (top two rows) and temperature (bottom two rows) to the ground truth across all pixels of the test set data are given, distinguishing the results of the three point estimation procedures: MAP, MeanShift, and MNPCP. Rows 1 and 3 present the direct one-to-one correlation of the predicted parameters to the ground truth, whereas rows 2 and 4 provide the corresponding relative residuals. In the latter panels, the black curve and grey shaded area indicates a binned median relative residual along with the interquantile range between the 25% and 75% quantile of these bins.

In Sect. 3, we specifically introduce the MNPCP approach, because the MAP and MeanShift estimators per construction of our inverse problem do not have a spatial consistency guarantee perpendicular to the LoS (as demonstrated in Fig. 3). To quantify whether the MNPCP approach improves upon this situation (beyond the qualitative comparison in Fig. 3), we computed the median difference in density and temperature for neighbouring pixels following:

$$\Delta_{\parallel \mathrm{LoS}} = \mathrm{median}\left\{x_{i+1,j,k} - x_{i,j,k}\right\} \begin{cases} \forall i \in \{1, \ldots, N-1\} \\ \forall j,k \in \{1, \ldots, N\} \end{cases} \tag{14}$$

and perpendicular to the LoS:

$$\Delta_{\perp \mathrm{LoS}} = \mathrm{median}\left\{\Delta_j, \Delta_k\right\}, \tag{15}$$

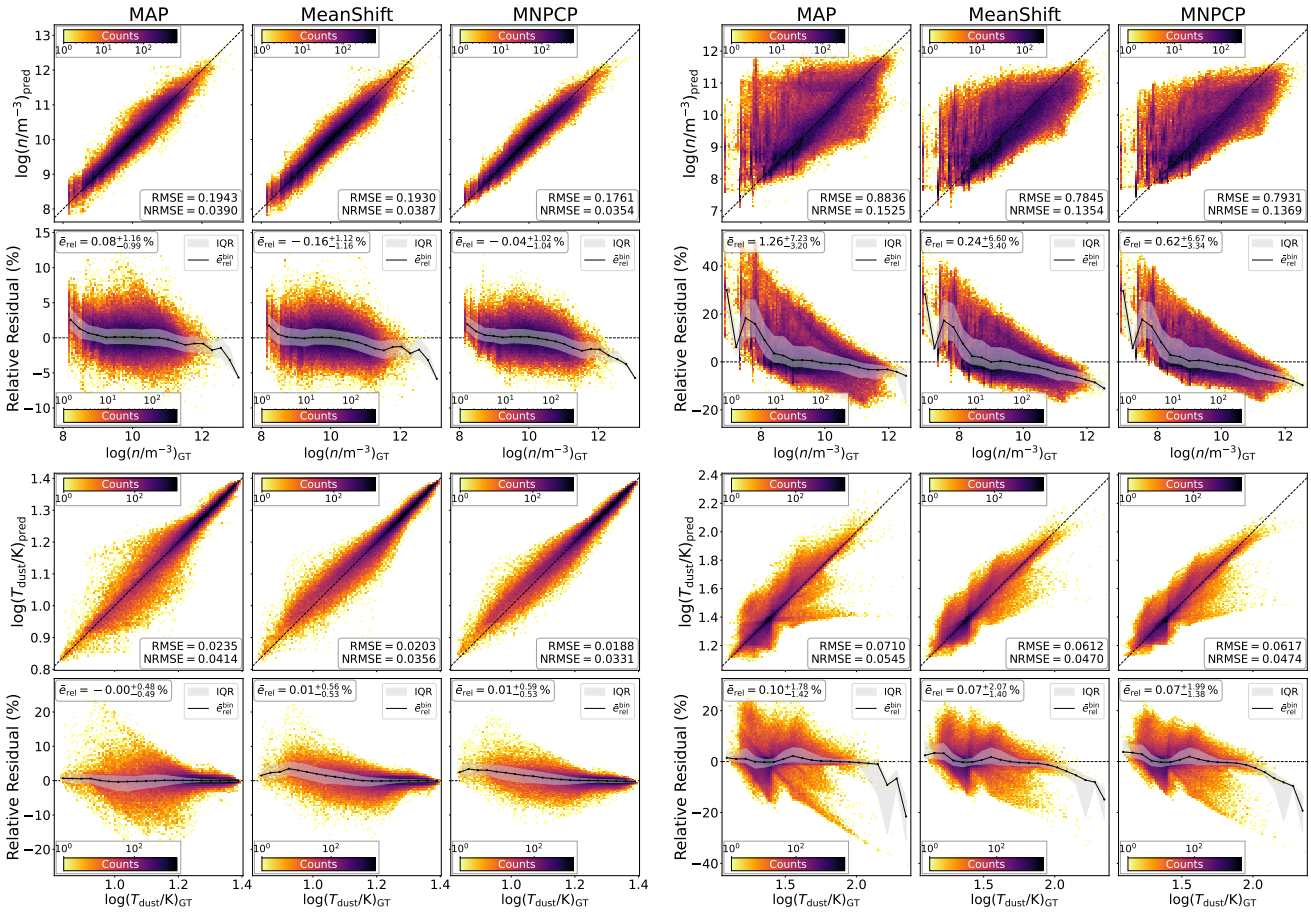

Fig. 5. Breakdown of the predictive performance of the SLoS-cINN for the best and worst cases. Analogously to Fig. 4, we show the 2D histograms for the one-to-one comparison of the prediction results (rows one and three) and their respective residuals (rows 2 and 4). The left three columns present the five best reconstructed cubes, whereas the five worst reconstructed ones are shown in the three right columns, respectively.

where

$$\Delta_j = \left\{ x_{i,j+1,k} - x_{i,j,k} \right\} \begin{cases} \forall j \in \{1, \ldots, N-1\} \\ \forall i, k \in \{1, \ldots, N\} \end{cases}$$

$$\Delta_k = \left\{ x_{i,j,k+1} - x_{i,j,k} \right\} \begin{cases} \forall k \in \{1, \ldots, N-1\} \\ \forall i, j \in \{1, \ldots, N\}, \end{cases}$$

for our three different point estimators. We then compared these results to the respective values obtained from the ground truth. The results of this analysis on the test set are summarised in Table 2. As expected, there is no preferred direction in the ground truth, with values of about 0.075 in $\log(n/\mathrm{m}^{-3})$ and 0.01 in $\log(T_\mathrm{dust}/\mathrm{K})$ for both $\Delta_{\parallel\mathrm{LoS}}$ and $\Delta_{\perp\mathrm{LoS}}$. In the MAP and MeanShift prediction results, on the other hand, we find $\Delta_{\perp\mathrm{LoS}}$ to be about twice as large as $\Delta_{\parallel\mathrm{LoS}}$ on average, confirming again the spatial consistency issue. In contrast, the MNPCP estimator offers a much more balanced result, achieving about even $\Delta_{\perp\mathrm{LoS}}$ and $\Delta_{\parallel\mathrm{LoS}}$ for $\log(T_\mathrm{dust}/\mathrm{K})$, and at least reducing $\Delta_{\perp\mathrm{LoS}}$ to about $1.5\Delta_{\parallel\mathrm{LoS}}$ for $\log(n/\mathrm{m}^{-3})$. Yet even with that outcome, the MNPCP estimate does not quite achieve the balance of the ground truth results. It is also interesting to note that all three point estimators return solutions where $\Delta_{\parallel\mathrm{LoS}}$ is notably smaller than in the ground truth. This indicates that the cINN prediction tends to return a smoother transition along the LoS than the ground truth. This is likely a result of the fact that the cINN returns a smooth continuous output, whereas the ground truth is limited by the coarseness of the simulation resolution.

In summary, we find an overall very satisfactory performance of the SLoS-cINN in the 23-wavelength configuration, providing a fairly robust recovery of dust density and temperature for most of the tested parameter range. We do note, however, a systematic decrease in performance towards the lower and upper limits of the trained range in terms of density and temperature, which can potentially be traced back to a relative lack of examples in these regimes in the training data. We also identified a dependence of the performance on the radiation setup of the test cubes, where the ones also hosting a star in addition to the ISRF appear more difficult to reconstruct. Regarding the choice of the point estimator, there is no clear winner in terms of the NRMSE and $|\bar{e}_\mathrm{rel}|$ performance indicators. Nevertheless, we believe that the MNPCP approach appears as the most reasonable solution because it can provide smooth reconstruction solutions both along and perpendicular to the LoS. One caveat to keep in mind is the fact that the MNPCP point estimator does amplify the systematic error tendencies at the lower and upper limits of the density and temperature ranges due to its inherent smoothing effect.

4.2. Taking the neighbouring LoSs into account

Rows 5 and 6 of Fig. 3 provide the qualitative comparison between the prediction outcomes of the three point estimation approaches for the NLoS-cINN, which is the model trained for the alternative formulation of the inverse problem described in

Table 2. Overview of the median difference, Δ, in dust density and temperature between neighbouring pixels along and perpendicular to the LoS for the three different point estimators and two inverse problem setups.

Inverse problem setup	Measure	$\log(n/\mathrm{m}^{-3})$		$\log(T_{\mathrm{dust}}/\mathrm{K})$	
		$\Delta_{\parallel\mathrm{LoS}}$	$\Delta_{\perp\mathrm{LoS}}$	$\Delta_{\parallel\mathrm{LoS}}$	$\Delta_{\perp\mathrm{LoS}}$
Single LoS (23 wavelengths)	Ground Truth	$0.075^{+0.086}_{-0.050}$	$0.073^{+0.082}_{-0.048}$	$0.010^{+0.010}_{-0.006}$	$0.010^{+0.009}_{-0.006}$
	MAP	$0.040^{+0.049}_{-0.024}$	$0.106^{+0.099}_{-0.059}$	$0.005^{+0.008}_{-0.003}$	$0.011^{+0.012}_{-0.006}$
	MeanShift	$0.050^{+0.043}_{-0.029}$	$0.098^{+0.084}_{-0.054}$	$0.007^{+0.007}_{-0.004}$	$0.011^{+0.010}_{-0.006}$
	MNPCP	$0.045^{+0.045}_{-0.027}$	$0.064^{+0.055}_{-0.035}$	$0.007^{+0.007}_{-0.004}$	$0.008^{+0.008}_{-0.005}$
Single LoS (7 wavelengths)	MAP	$0.032^{+0.037}_{-0.018}$	$0.107^{+0.106}_{-0.060}$	$0.003^{+0.006}_{-0.002}$	$0.012^{+0.014}_{-0.007}$
	MeanShift	$0.041^{+0.036}_{-0.025}$	$0.087^{+0.078}_{-0.048}$	$0.006^{+0.007}_{-0.004}$	$0.011^{+0.011}_{-0.006}$
	MNPCP	$0.038^{+0.041}_{-0.024}$	$0.063^{+0.056}_{-0.035}$	$0.006^{+0.007}_{-0.004}$	$0.008^{+0.008}_{-0.005}$
Neighbour LoS	Ground Truth	$0.077^{+0.085}_{-0.051}$	$0.077^{+0.083}_{-0.050}$	$0.011^{+0.010}_{-0.006}$	$0.010^{+0.009}_{-0.005}$
	MAP	$0.045^{+0.055}_{-0.027}$	$0.103^{+0.095}_{-0.057}$	$0.006^{+0.008}_{-0.004}$	$0.010^{+0.010}_{-0.006}$
	MeanShift	$0.056^{+0.048}_{-0.033}$	$0.095^{+0.082}_{-0.052}$	$0.008^{+0.008}_{-0.005}$	$0.010^{+0.009}_{-0.005}$
	MNPCP	$0.051^{+0.049}_{-0.030}$	$0.066^{+0.058}_{-0.037}$	$0.008^{+0.008}_{-0.008}$	$0.007^{+0.007}_{-0.004}$

Notes. The ground truth values differ slightly between the two setups, as the NLoS approach does not contain the edge LoSs of each cube.

Sect. 3.5. For direct compatibility the shown slice and example cube are the same as for the SLoS-cINN outcomes in Fig. 3, except that the 124 border pixels, which lack the required number of neighbouring LoSs for the prediction with the NLoS-cINN, are missing. At first glance the MAP and MeanShift results appear less noisy, which is subject to fewer strong discontinuities, with the NLoS-cINN in comparison to the SLoS-cINN outcome (first two rows of Fig. 3). In addition, the NLoS-cINN based reconstructions seem overall to be slightly more faithful to the ground truth in terms of the recovered details.

Looking at the spatial discontinuities perpendicular to the LoSs in the MAP and MeanShift estimates (see Table 2), it appears that the NLoS-cINN suffers on average from the same issue as the SLoS-based prediction results. $\Delta_{\perp\mathrm{LoS}}$ is still twice as large as $\Delta_{\parallel\mathrm{LoS}}$ for both density and temperature. Again, only the MNPCP approach achieves a more balanced result, but not quite at the level of the ground truth. Thus, accounting for the fluxes in the neighbouring LoSs does not appear to lead to a significant improvement of the spatial consistency perpendicular to the LoS in the prediction of dust density and temperature. It is worth noting, however, that this observation may only hold in the fully resolved dust emission map scenario that we have posed in this study, which renders neighbouring LoSs effectively independent. In real observations, however, where the PSF of a given instrument is larger than a single pixel, neighbouring pixels in the dust emission maps may become correlated. In the latter case, it is possible that the NLoS-cINN may perform better with regards to the spatial consistency. We will conduct a corresponding test in our subsequent work, once we have established a proper treatment of the instrument-related resolution effects during training.

Looking at the overall performance of the NLoS-cINN in comparison to the SLoS-cINN (see Table 1 and Fig. 6), we do find a general improvement. In particular, the NRMSE goes down to values as low as 0.054 and 0.0213 for $\log(n/\mathrm{m}^{-3})$ and $\log(T_{\mathrm{dust}}/\mathrm{K})$ with the MNPCP estimator, compared to the SLoS-cINN results of 0.062 and 0.0245. $|\bar{e}_{\mathrm{rel}}|$ with, for instance, the MNPCP estimator improves to 1.54% and 0.82% opposed to the SLoS-based performance of 1.84% and 0.96%, respectively. It is possible that this performance improvement is only a data selection effect, since the NLoS and SLoS test sets are not fully identical, with the former missing the edge LoSs of every cube. To test this hypothesis, we recomputed the SLoS-cINN performance, limited to the LoSs of the NLoS test set. This experiment reveals that the observed average performance improvement of the NLoS-cINN is real, as the SLoS-cINN performs even slightly worse on this limited test set, returning for instance $|\bar{e}_{\mathrm{rel}}|$ values of 1.86% and 0.98% for $\log(n/\mathrm{m}^{-3})$ and $\log(T_{\mathrm{dust}}/\mathrm{K})$ with the MNPCP estimator, respectively.

In summary, accounting for the fluxes of the neighbouring LoSs in the input has not achieved its initial goal of improving the spatial consistency of the predicted dust densities and temperatures perpendicular to the LoS. It has, however, demonstrated a slight improvement of the predictive performance of the model, indicating that knowledge of the fluxes in the neighbouring LoSs provides additional constraints for the prediction of the dust properties. However, this comes at a cost of flexibility, as all query LoSs now also require observations of the eight adjacent LoSs in this approach. All in all, the NLoS-cINN does not increase the spatial consistency and despite a slight performance improvement does not appear to be a markedly superior approach.

4.3. Realistic wavelength coverage test

As our final experiment, we trained and tested an SLoS-cINN for a more realistically limited wavelength coverage, corresponding to the central wavelengths of the following seven bands: WISE 22 μm, SOFIA 89 μm and 154 μm, *Herschel* PACS 100 μm and 160 μm, *Herschel* SPIRE 350 μm, and LABOCA 870 μm. The last two rows in Fig. 3 provide the qualitative example of the dust density and temperature prediction results for a single cube slice in comparison to the ground truth (see also Fig. A.3 for the corresponding input emission maps at the seven selected wavelengths, as indicated by the highlighted panels). It is immediately evident that the large reduction in wavelength coverage

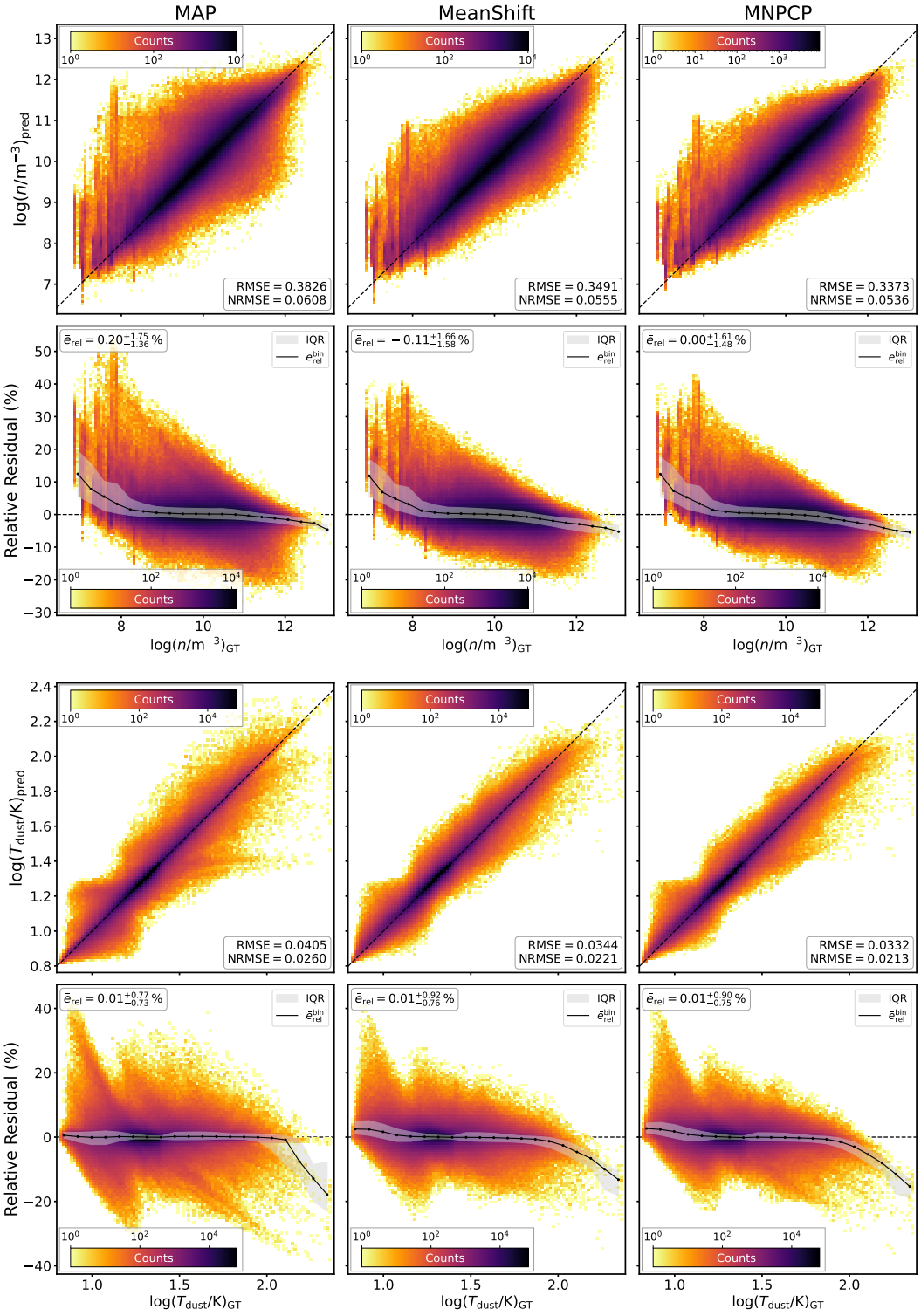

Fig. 6. Prediction performance summary for the NLoS-cINN on the held-out test data. 2D histograms of the direct one-to-one comparison of prediction results to the ground truth (rows one and three) and the respective residuals (rows 2 and 4) are shown for the three different MAP estimators (analogously to Fig. 4).

leads to a notably decreased quality in the reconstruction. While larger scale and more diffuse features of the density and temperature distributions are still being recovered very well, this is no longer true for narrow details at high density and low temperature, in particular, which tend to be only partially reconstructed. This is even more apparent in the full prediction summary of this cube in Fig. B.8 (and the corresponding isodensity surface diagram in Fig. B.9). In addition, the MAP point estimator appears to produce a lot more and larger spatial discrepancies in the predicted dust densities and temperatures perpendicular to the LoS. Interestingly, where the MeanShift algorithm appeared to show similar behaviour to the MAP estimator in the previous

SLoS-cINN setup in terms of spatial inconsistencies, it seems to provide more consistent prediction results here. Looking at $\Delta_{\perp\mathrm{LoS}}$ and $\Delta_{\parallel\mathrm{LoS}}$ in Table 2, this can be quantitatively confirmed at least for the dust density as well, with $\Delta_{\perp\mathrm{LoS}}$ being only about two times greater than $\Delta_{\parallel\mathrm{LoS}}$ in the MeanShift result, compared to $\Delta_{\perp\mathrm{LoS}} \approx 3\Delta_{\parallel\mathrm{LoS}}$ in the MAP outcome. With this outcome, it seems that a determination of the most likely solution in the joint target parameter space in this setup is more robust in terms of the spatial consistency of the dust density than the marginalisation approach in the MAP estimator.

Looking at the overall performance on the test set in Table 1, the limitation to only seven wavelengths in this setup increases the NRMSE to 0.084–0.098 and 0.036–0.046 for $\log(n/\mathrm{m}^{-3})$ and $\log(T_{\mathrm{dust}}/\mathrm{K})$, respectively. The corresponding $|\bar{e}_{\mathrm{rel}}|$ values rise to 2.6–3.1% and 1.5–2.0%. Although this is a notable performance decrease with $|\bar{e}_{\mathrm{rel}}|$ increasing by a factor of 1.5–2 in comparison to the 23-wavelength setup, considering that this cINN has 16 wavelengths fewer to work with, this is still a quite satisfactory performance. This indicates that the 3D reconstruction approach can be feasible even for more limited observational coverage.

Investigating the prediction performance in more detail in Fig. 7 (which provides a breakdown, analogously to Figs. 4 and 6), we find that the change in wavelength coverage also influences the systematic tendencies of the prediction results. Where for instance the binned median relative residual curve for the dust density in the 23-wavelength SLoS-cINN outcome exhibits a plateau between 10^8 and $10^{11}\,\mathrm{m}^{-3}$ and only really leans into systematic behaviour below and above this range, the systematic offsets are amplified in the seven wavelength prediction outcomes. In particular for the MeanShift and MNPCP point estimates we now find a pivot point at around $10^{10}\,\mathrm{m}^{-3}$, below which the density tends to be overestimated and above which it is systematically underestimated. We find a similar amplification of the earlier observed systematic behaviour in the MeanShift and MNPCP estimates of the dust temperature, where there is now a pivot point between over- and underestimation at about 19 K. Nevertheless, for most pixels, the error remains comparatively small, as indicated by $|\bar{e}_{\mathrm{rel}}|$, and only few individual pixels will manage to reach the maximum relative residual of about 60%. Part of this systematic behaviour can likely be attributed to the relative lack of examples in the training data towards the lower and upper limits of density and temperature, as we discuss in Sect. 4.1. However, given that this cINN also exhibits systematic offsets within the parameter ranges, where a lot of training data is available, this also points towards the increased complexity of the recovery task with much less observational information; in particular in the intermediate density and temperature ranges. Evidence for the intrinsic increase in complexity can be found in the shape of the predicted posterior distributions, a few examples of which are shown in Fig. B.3. Compared to the full 23-wavelength SLoS-cINN, the predicted posteriors of the seven wavelength limited model appear notably broader and often exhibit double or multi-peaked distributions for both density and dust temperature even in the ISRF-only RT configuration. This shows that the cINN has both a harder time to constrain the prediction and finds more degeneracy in the more limited problem.

In summary, although the 7-wavelength SLoS-cINN exhibits an (expected) reduction in predictive performance compared to the 23-wavelength counterpart, this experiment proves that the 3D reconstruction of the dust distributions is quite feasible even when subject to more realistic observational constraints. We also want to emphasise again that our wavelength selection does not necessarily preserve the most information for the given inverse problem. A different, more optimal selection of wavelengths might retain more of the predictive performance of the full 23-wavelength cINN. We plan to look further into both the optimal choice of wavelength combination and relative importance of each wavelength (for the reconstruction) for realistic applications in our continued development of this 3D reconstruction approach.

4.4. Comparison with classical SED fit

The dust emission is often modelled by a modified black body (MBB) in order to construct a column density map and the temperature distribution from multi-wavelength observations. As a final test, we compared the predictive performance of our cINN method to the results of a classical MBB fit. For the purposes of this test, we employed the full 23 wavelengths of coverage and focussed on the simpler RT scenario, comparing SED fit and cINN approach on the 50 cubes that are only subject to the ISRF. The dust emission, I_ν, may be approximated as an MBB by:

$$I_\nu = B_\nu\left(T_{\mathrm{dust}}\right)\left(1-\mathrm{e}^{-\tau_\nu}\right) \approx B_\nu\left(T_{\mathrm{dust}}\right)\tau_\nu = \mu_{\mathrm{g}} m_{\mathrm{H}} N_{\mathrm{H}} \delta_{\mathrm{gd}} \kappa_\nu B_\nu\left(T_{\mathrm{dust}}\right), \tag{16}$$

where τ_ν is the optical depth, m_{H} is the hydrogen mass, and the black body $B_\nu\left(T_{\mathrm{dust}}\right)$ spectrum is modified by the dust opacity:

$$\kappa_\nu = 2\left(\frac{\nu}{\nu_0}\right)^\beta \ \mathrm{cm}^2\ \mathrm{g}^{-1}\,. \tag{17}$$

For the opacity, we take a characteristic frequency of $\nu_0 = 600$ GHz and a spectral index of $\beta = 2$. Here, the dust temperature T_{dust} and the gas column density N_{H} are the fit parameters to be determined by a least-squares fit.

To compare the cINN outcome with the SED fit results we compute the column number density, N, by integrating the density along each LoS for both the ground truth and SLoS-cINN prediction outcome, that is:

$$N = \int_0^L n\ \mathrm{d}l = \sum_{i=1}^{32} n_i\ \Delta l, \tag{18}$$

where L indicates the depth of our test cubes of 0.2 pc and $\Delta l = L/32$ is our cube spatial resolution. As a proxy for the temperature determined by the SED fit, we can compute a density-weighted average temperature $\bar{T}_{\mathrm{dust}}$ following:

$$\bar{T}_{\mathrm{dust}} = \frac{\sum_{i=1}^{32} n_i T_{\mathrm{dust},i}}{\sum_{i=1}^{32} n_i}. \tag{19}$$

Figure 8 provides an example comparison of the column number densities and density weighted average temperatures between the outcome of the SED fit, the cINN prediction and the ground truth for one example cube that is only subject to the ISRF. As this example shows, both the SED fit and cINN approach manage to reproduce the column number density map for this cube quite accurately with overall absolute relative errors below 1%. However, the result based on the MAP cINN estimates returns notably smaller residual errors than the SED fit. The maps based on the MeanShift and MNPCP point estimators on the other hand are about on par with the SED fit, if not slightly better. For MeanShift this is likely explained by the on average higher absolute relative error of the method itself in comparison to the MAP estimator (as we describe in Sect. 4.1). The MNPCP estimator on the

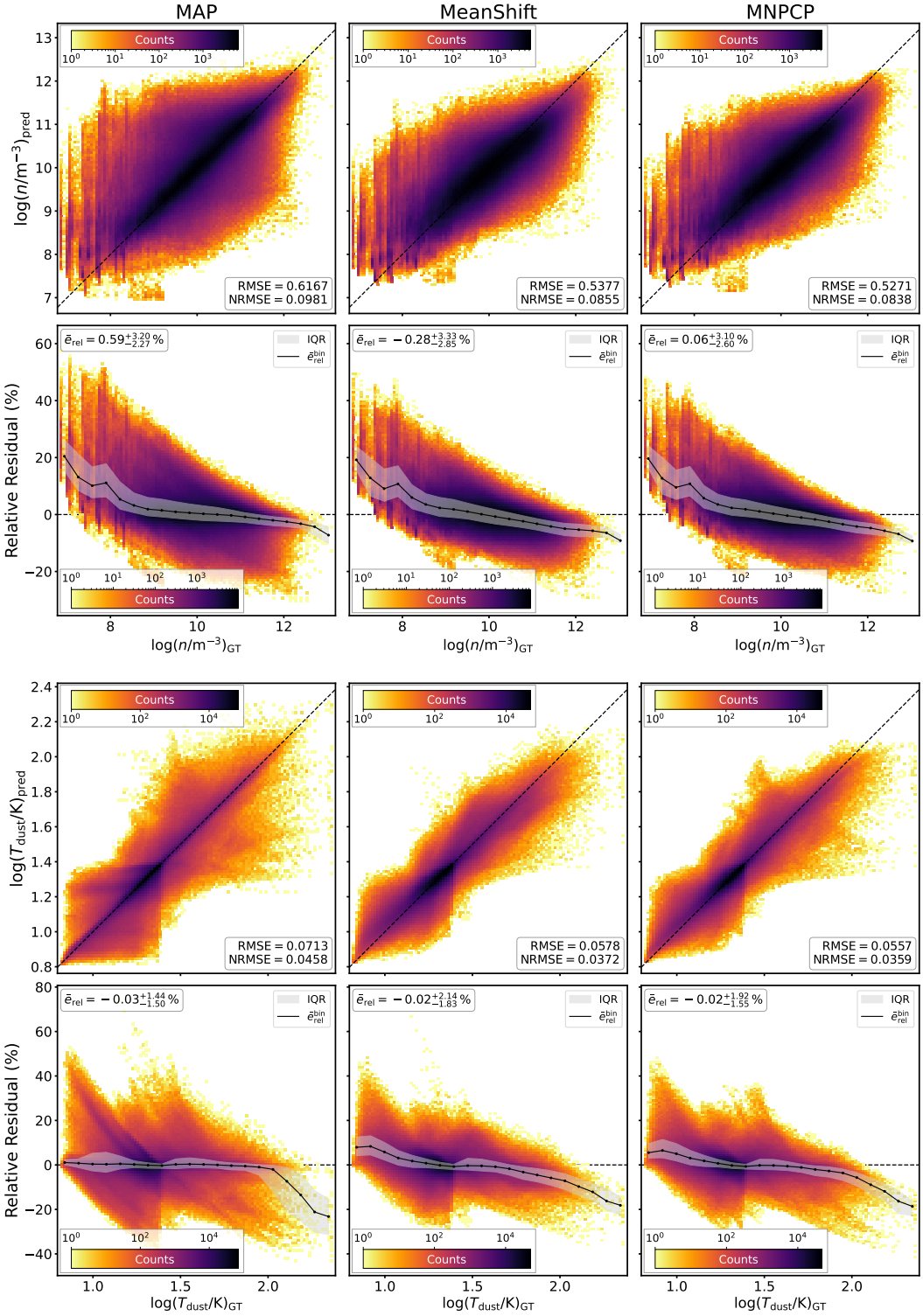

Fig. 7. Summary of the predictive performance on the held-out test data for the SLoS-cINN using only seven wavelenghts as input. Shown are 2D histograms of the direct one-to-one comparison of prediction results to the ground truth (rows one and three) and the respective residuals (rows 2 and 4) for the three different MAP estimators (analogously to Fig. 4).

other hand performs worse here because of its inherent smoothing action perpendicular to the LoS. As such small amounts of material may get mixed between adjacent LoSs, increasing the error on the column density. Looking now at the density averaged temperature $\bar{T}_{\rm dust}$, we find excellent results for the cINN based estimates, whereas the SED fit outcome notably underestimates the temperature by up to 10% or more. However, it is important to note that the density-weighted average temperature is only an approximation of the temperature that an SED fit would return, so a discrepancy is to be expected. Lastly, Fig. 9 provides the performance comparison between the SED fit and cINN approach across all pixels of the 50 considered test cubes. This figure confirms that the cINN based estimates of both column number density and density weighted average temperature are overall

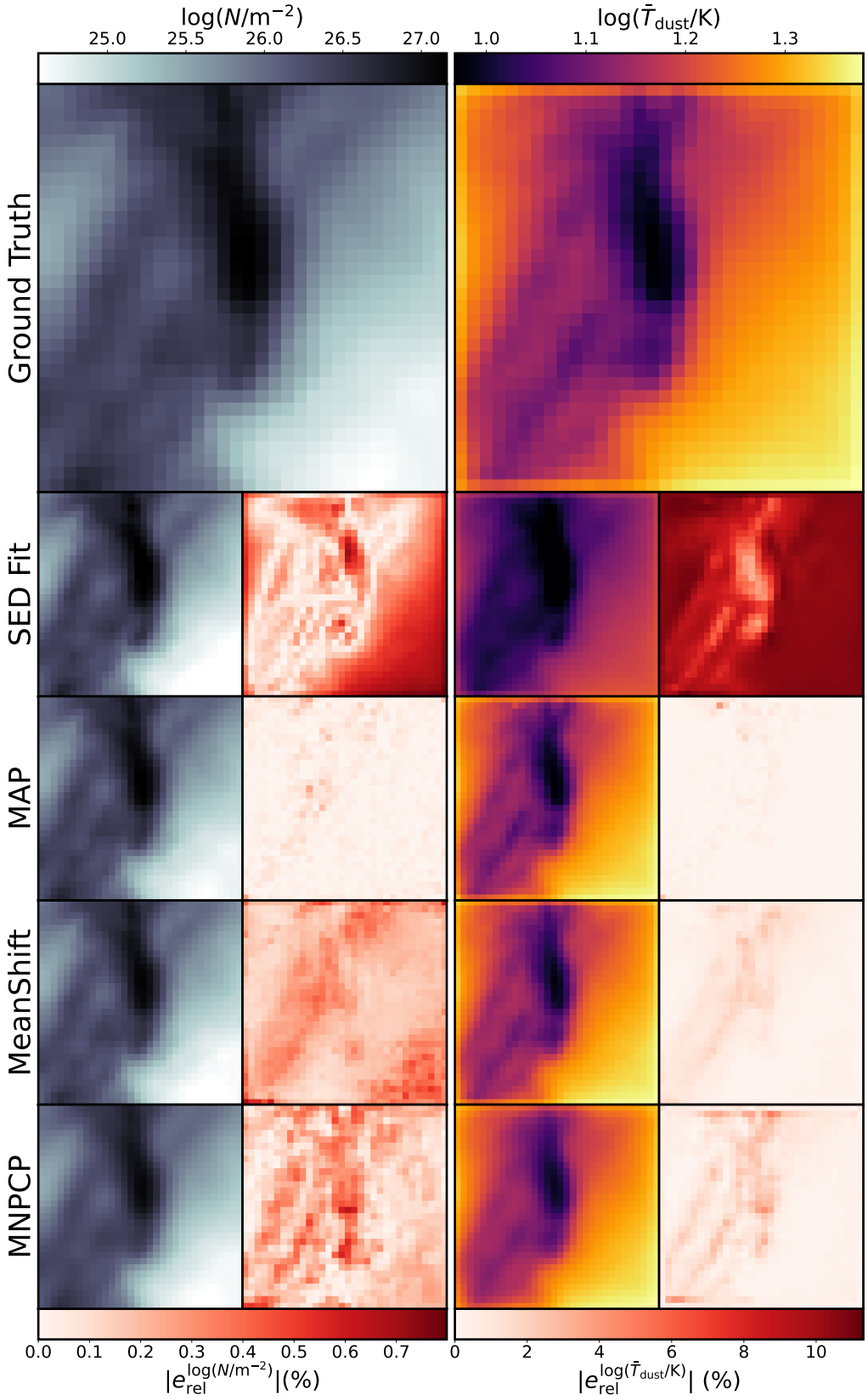

Fig. 8. Comparison of estimated column density and average temperature maps for one example cube in the ISRF-only RT configuration between a classical SED fit and our cINN approach. The large panels on the top show the ground truth for the column density (left) and density weighted average temperature (right), respectively. The smaller panels below each ground truth panel present a map of the estimates from the respective method on the left and a map of the absolute relative error $|e_{\rm rel}|$ on the right. The example cube as in Figs. 3, B.1, and B.2 is shown.

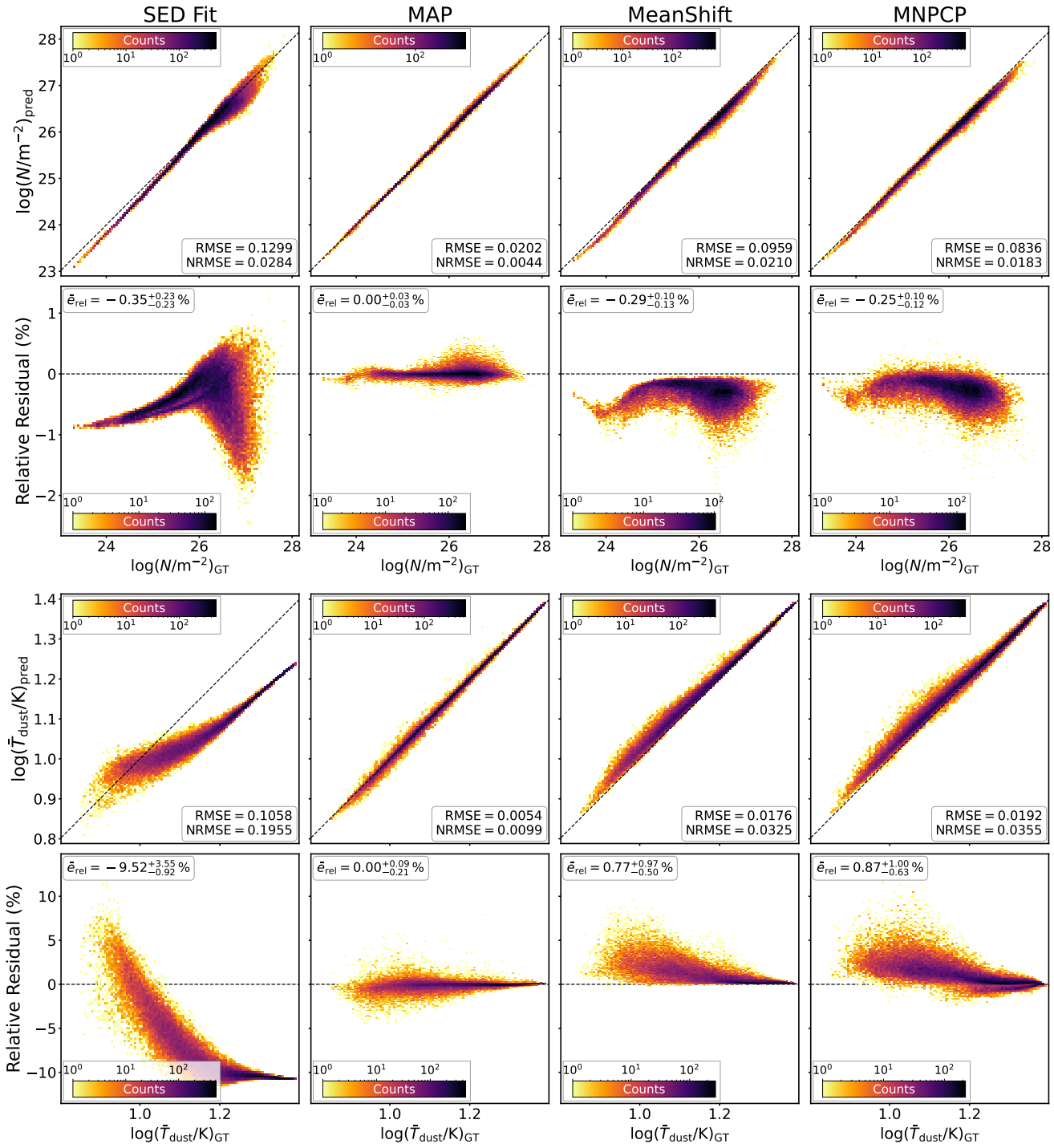

Fig. 9. Comparison of the predictive performance for column density and density averaged temperature between a classical SED fitting technique and our cINN approach across all pixels of the 50 test cubes that are only subject to the ISRF. The top two rows show the results for the estimated column density, whereas the bottom two present the density averaged dust temperature, $\bar{T}_{\rm dust}$. In each group, the upper panels provide a direct one-to-one comparison of the predicted values to the respective ground truth, whereas the lower panels show the corresponding relative residuals.

quite excellent. In addition, the cINN approach outperforms the SED fit in all cases, although the margin is comparatively small in the column number density for the MeanShift-based outcome. We also note a slight tendency of the MeanShift and MNPCP derived results towards overestimation of the density averaged temperature. We emphasise that the MBB SED fit may only provide a 2D estimate of the underlying density and temperature distributions, whereas the advantage of the cINN approach is the reconstruction of the full 3D information along the LoS. We also note that while this test indicates the MAP point estimate as the best suited choice to recover the column density, for the full 3D reconstruction, we still recommend the MNPCP estimate because of its higher degree of spatial consistency in the 3D structure.

4.5. On the physical feasibility of the 3D reconstruction

The physical reason why our approach of reconstructing the underlying 3D density and temperature structure works so well is that the dust opacity strongly depends on the energy of the incident electromagnetic wave. In addition, there is a dependence on the chemical composition and the grain size distribution, but we keep those fixed here to the standard Milky Way values (Sect. 2.2). For further details we refer to Tielens (2010) and Draine (2011). Overall, this means that different wavelengths trace different optical depths within the cloud. As a consequence, the observed radiation is accumulated in different ways along the LoS and the resulting SED encodes critical information about the density and temperature structure of the cloud, as also discussed in the Sect. 4.4. This process is highly degenerate, as varying cloud configurations can lead to very similar SEDs; thus, solving the inverse problem of reconstructing the cloud properties from the SED is a challenging task. This is where the INN architecture can play out its full potential since it belongs to the group of normalising flow methods. It offers direct access to the full posterior distribution function and enables us to approach this challenge in a statistically reliable way.

4.6. Application to real data

An application to real observational data of cloud cores is at this stage not currently feasible with our setup. The main reason for this is our treatment of the synthetic dust emission observations. The choice to only use monochromatic dust emission and to not include instrument-related effects, such as a consideration of the PSF or noise renders our synthetic dust emission observations notably different to what the actual observatories would measure. This results in a gap between the synthetic observable space learned by our cINN models and the actual observable space spanned by real measurements, so that prediction attempts are bound to fail (for an example see Appendix B.2). For an application to real data, the generation of the synthetic dust observations thus needs to be further refined by accounting for instance for the actual band width of real instruments and ensuring a proper sampling and modelling of their respective PSFs. We plan to address these challenges in our subsequent development of our method.

Apart from the current limitations of our synthetic dust emission observations, we also identified another area that may require improvements towards an application to real data. As we describe in Sect. 2, our training data generation is aimed at creating simplified, synthetic analogues to objects such as ρ Oph A. However, we suspect that our simplified prescription for adding a star to the cubes, that is, randomly placing it in a region of low dust density to emulate stellar feedback, is not likely to produce truly realistic results. In this approach, the dust clouds are merely illuminated by the star, rather than being mechanically affected by the stellar feedback. Consequently, we do not actually model clouds that are actively shaped by their nearby stars, as is (in particular) the case for cores such as that of ρ Oph A. For that reason, we plan to move to more sophisticated MHD simulations, where stellar feedback is properly accounted for before we apply our method to real observations.

5. Summary and conclusions

In this paper, we present a proof-of-concept deep-learning approach to reconstructing interstellar dust distributions in 3D from observed dust emission maps, focussing on the sub-parsec length scales of individual star-forming clumps. To train this approach, we modelled simplified analogues of the ρ Oph A star-forming clump, employing the Cloud Factory dust cloud simulations by Smith et al. (2020), Izquierdo et al. (2021). For the basis of our training database, we selected 11.065 cubes from the Cloud Factory, centered on clump-like dust aggregations, with a side length of 0.2 pc and $32 \times 32 \times 32$ pixel resolution. We simulated the corresponding dust emission observations with the POLARIS (Reissl et al. 2016, 2019) radiative transfer code at 23 different wavelengths between 12 µm and 1300 µm, matching the central wavelengths of the observational instruments of WISE, MSX, *Spitzer*, SOFIA, *Herschel*, ALMA, APEX, and CSO. For the radiative transfer, we considered two irradiation scenarios. In the first, the dust is only subject to the interstellar radiation field (ISRF) with an amplitude that matches conditions in nearby star-forming cores. In the second, we randomly inserted one B4-type star in a low density area inside the cube in addition to the ISRF in order to emulate cloud cores that are subject to the radiation of nearby stars. This procedure yields a final training dataset of 22.130 mock dust clumps, including their dust density and temperature distributions on a 3D grid and the corresponding simulated dust emission fluxes in 23 wavelengths.

To reduce the complexity and dimensionality of the inverse problem posed by the 3D reconstruction task, we broke the problem down to individual LoS under the assumption that (to first order) the emission measured in a given pixel only depends on the material along the corresponding LoS. Specifically, we aimed to recover the 32 dust densities and temperatures along a given LoS from the measured fluxes at different wavelengths in the corresponding pixel of the dust emission map. For this purpose, we trained a conditional invertible neural network (cINN), which is a deep-learning approach that can efficiently estimate full posterior distributions for the target parameters conditioned on the observations. To recover point estimates from the posterior distributions predicted by the cINN (and compare these to the ground truth), we tested three different methods. The first uses a 1D kernel density estimate to determine the maximum a posteriori (MAP) prediction (that is the most likely value) for the dust density and temperature from the marginalised 1D posterior distributions of the individual pixels. The second method employs the MeanShift algorithm to determine the most likely solutions as the point with the highest (probability) density in the full 64-dimensional space of the predicted posterior distributions. As both the MAP and MeanShift estimators have no spatial consistency guarantee perpendicular to the LoS, as per our reduction of the reconstruction task, we introduced a third estimator to rectify this circumstance. This method, dubbed median neighbour pixel combined posterior (MNPCP), determines the point estimates for a given pixel by accumulating the posterior samples of all neighbouring pixels and computing the median of dust density and temperature on this collection.

In total, we trained and tested cINNs for three different formulations of the reconstruction task. The first consists of a perfect information scenario, where we have access to the dust emission maps at all 23 wavelengths. In the second, we extended the network input to also account for the observed fluxes in the neighbouring pixels of the query LoS to evaluate whether this improves the spatial consistency of the prediction outcome. Lastly, we considered a more realistically limited observational scenario, where observations are only available in a combination of seven wavelengths (matching the coverage of real observational data that is for instance available for the star-forming core ρ Oph A). We then evaluated the performance of our trained cINN models on a synthetic test set that consists of 50 mock

dust clouds in both irradiation scenarios. Our main findings are summarised in the following.

The model trained for the 23-wavelength scenario achieves an excellent overall predictive performance, returning median absolute relative errors $|\bar{e}_{\rm rel}|$ of 1.84 to 2.01% in $\log(n/{\rm m}^{-3})$ and 0.87 to 1.01% in $\log(T_{\rm dust}/{\rm K})$, depending on the choice of the point estimator. Averaged over the five best reconstructed cloud cores, $|\bar{e}_{\rm rel}|$ even goes as low as 1% and 0.5% in $\log(n/{\rm m}^{-3})$ and $\log(T_{\rm dust}/{\rm K})$, respectively. In general, we find that the reconstructive performance is better for clouds that are only subject to the ISRF, indicating that the presence of a star in the vicinity of a mock dust cloud adds a notable level of complexity to the reconstruction task. Breaking the predictive performance down to the individual pixels, we also identified some systematic behaviours. Although the dust densities and temperatures were recovered well overall, we found trends for overestimating the density in the low density regime ($n < 10^8\,{\rm m}^{-3}$) and of slight underestimations for very-high-density regions ($n > 10^{11}\,{\rm m}^{-3}$). A similar tendency for underestimation also appears in the dust temperature point estimates for $T_{\rm dust} > 100$ K. We identified a bias in the training data as a potential explanation for this behaviour, as the training database contains significantly fewer pixels in these parameter ranges. It should be noted, however, that (in particular) the $T_{\rm dust} > 100$ K regime represents more extreme conditions that are more typically found in HII regions or in regions of strong outflow activity; however, these are neither accounted for in the Cloud Factory simulations nor the focus of our analysis. This incomplete modelling of the high temperature regime may also tie into the comparatively worse reconstructive performance.

The comparison of the three point estimators reveals no clear favourite in terms of the quantitative performance measures. Although the MeanShift and MNPCP estimators achieve a slightly better NRMSE than the MAP and the MAP estimator provides the smallest $|\bar{e}_{\rm rel}|$, the difference between the three methods is only on the order of 10^{-3} in NRMSE and 0.1% in $|\bar{e}_{\rm rel}|$. The MAP and MeanShift estimates do suffer from the aforementioned spatial inconsistencies perpendicular to the LoS (as per construction). Although the MeanShift results appear to be slightly less affected by this, we find that determining an optimal bandwidth for MeanShift in the 64-dimensional target parameter space can be tricky and may lead to oversmoothing. The MNPCP approach provides much better results in term of spatial consistency, but also exhibits oversmoothing behaviour, which amplifies the systematic trends of the method at the edges of the parameter ranges. Ultimately, we would generally recommend the MNPCP solution (despite its flaws), as it quite effectively avoids the more severe and unphysical spatial inconsistencies perpendicular to the LoS that the MAP and MeanShift approaches can hardly avoid (by construction).

For the cINN model that also accounts for the emission in the neighbouring LoSs, we found an overall slight increase in the predictive performance, but no significant improvement in the spatial consistency of the prediction outcome. This experiment indicates that accounting for the measured fluxes in the neighbouring pixels may help with constraining the dust density and temperatures overall, but cannot counteract the inherent spatial consistency bias of the LoS approach. In addition, this setup is less flexible as it requires a query LoS to have measurements for all eight neighbouring pixels, which excludes (for instance) edge pixels. Although the slight performance improvement of this approach is certainly desirable, it is overall not large enough to clearly distinguish this setup as a superior method in comparison to the single LoS approach.

Limiting the input to a more realistic coverage of seven wavelengths – in our case corresponding to the central wavelengths of the WISE 22 μm, SOFIA 89 μm and 154 μm, *Herschel* PACS 100 μm and 160 μm, *Herschel* SPIRE 350 μm, and LABOCA 870 μm bands – naturally leads to a decrease in predictive performance. Nevertheless, the cINN overall still achieves a satisfactory performance with $|\bar{e}_{\rm rel}|$ on the order of 2.6 to 3.1% for $\log(n/{\rm m}^{-3})$ and 1.5 to 2.0% in $\log(T_{\rm dust}/{\rm K})$. The most notable change in the behaviour of the seven wavelength cINN is an amplification of the average systematic errors in the predicted point estimates, leading to more difficulties overall in reconstructing the dust distributions in very dense structures or extremely diffuse regions. It is worth emphasising, however, that the tested selection of wavelengths was inspired by available data for ρ Oph A and not curated to maximise the network performance. A more optimal wavelength configuration is certainly likely to achieve even better reconstructive power. Lastly, we find that an application of our approach to real observational data is not yet feasible at this stage, as our simplified simulation setup is not able to correctly model all the nuances of real dust emission observations. We conclude that a more in-depth treatment of the synthetic dust emission observations and improvements to our simulation basis (to e.g. properly model the interaction between stars and the dust through stellar feedback) may be required.

In summary, we have shown that a cINN-based approach for the 3D reconstruction of dust distributions produces quite excellent results in a perfect information scenario on synthetic data and still retains a satisfactory performance, even for more realistically limited observational coverage. For future applications to real data, however, the method still requires a further refinement of the simulation setup for the training data to resolve the mismatch between real and synthetic observations, with some adjustment to the training set sampling to reduce the potential bias, and an analysis of the most informative band combinations to maximise performance in realistic coverage scenarios. Once these points have been resolved over the course of our follow-up studies and we can demonstrate a successful application to real observational data, we plan to make this approach available to the community in the form of an open-source tool. Access to a work-in-progress code may be provided upon reasonable request beforehand.

Acknowledgements. The team in Heidelberg acknowledges funding from the European Research Council via the ERC Synergy Grant "ECOGAL" (project ID 855130), from the German Excellence Strategy via the Heidelberg Cluster of Excellence (EXC 2181 – 390900948) "STRUCTURES", and from the German Ministry for Economic Affairs and Climate Action in project "MAINN" (funding ID 50OO2206). They also thank for computing resources provided by *The Länd* and DFG through grant INST 35/1134-1 FUGG and for data storage at SDS@hd through grant INST 35/1314-1 FUGG. RJS gratefully acknowledges an STFC Ernest Rutherford fellowship (grant ST/N00485X/1).

References

André, P., Men'shchikov, A., Bontemps, S., et al. 2010, A&A, 518, A102
Ardizzone, L., Kruse, J., Rother, C., & Köthe, U. 2019a, Analyzing Inverse Problems with Invertible Neural Networks, in International Conference on Learning Representations
Ardizzone, L., Lüth, C., Kruse, J., Rother, C., & Köthe, U. 2019b, CoRR, 1907.02392
Ballesteros-Paredes, J., & Mac Low, M.-M. 2002, ApJ, 570, 734
Beaumont, C. N., Offner, S. S. R., Shetty, R., Glover, S. C. O., & Goodman, A. A. 2013, ApJ, 777, 173
Bister, T., Erdmann, M., Köthe, U., & Schulte, J. 2022, Eur. Phys. J. C, 82, 171
Bjorkman, J. E., & Wood, K. 2001, ApJ, 554, 615
CASATeam, Bean, B., Bhatnagar, S., et al. 2022, PASP, 134, 114501

Comaniciu, D., & Meer, P. 2002, IEEE Trans. Pattern Anal. Mach. Intell., 24, 603
Cortes, P. C., Remijan, A., Hales, A., et al. 2022, ALMA Technical Handbook, ALMA Doc. 9.3, ver. 1.0
Dinh, L., Krueger, D., & Bengio, Y. 2015, in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Workshop Track Proceedings, eds. Y. Bengio, & Y. LeCun
Dinh, L., Sohl-Dickstein, J., & Bengio, S. 2017, Density estimation using Real NVP, in International Conference on Learning Representations
Dole, H., Lagache, G., & Puget, J.-L. 2003, ApJ, 585, 617
Draine, B. 2003, ARA&A, 41, 241
Draine, B. T. 2011, Physics of the Interstellar and Intergalactic Medium (Princeton: Princeton University Press)
Egan, M. P., Price, S. D., Moshir, M. M., Cohen, M., & Tedesco, E. 1999, The Midcourse Space Experiment Point Source Catalog Version 1.2 Explanatory Guide, Technical Report, AD-A381933; AFRL-VS-TR-1999-1522
Eisert, L., Pillepich, A., Nelson, D., et al. 2023, MNRAS, 519, 2199
Elia, D., Merello, M., Molinari, S., et al. 2021, MNRAS, 504, 2742
Exter, K. 2017, Quick-Start Guide to HERSCHEL–PACS The Photometer, HERSCHEL-HSC-DOC-2151, version 1.0
Fukunaga, K., & Hostetler, L. 1975, IEEE Trans. Information Theory, 21, 32
Galliano, F., Galametz, M., & Jones, A. P. 2018, ARA&A, 56, 673
Garcia-Cuesta, E., de la Torre, F., & de Castro, A. J. 2009, Machine Learning Approaches for the Inversion of the Radiative Transfer Equation, eds. S.-I. Ao, B. Rieger, & S.-S. Chen (Dordrecht: Springer Netherlands), 319
Glover, S. C. O., & Mac Low, M. 2007, ApJS, 169, 239
Glover, S. C. O., & Mac Low, M. M. 2011, MNRAS, 412, 337
Gould, R. J., & Salpeter, E. E. 1963, ApJ, 138, 393
Haldemann, J., Ksoll, V., Walter, D., et al. 2023, A&A, 672, A180
Harper, D. A., Runyan, M. C., Dowell, C. D., et al. 2018, J. Astron. Instrum., 7, 1840008
Hauser, M. G., & Dwek, E. 2001, ARA&A, 39, 249
Hill, R., Masui, K. W., & Scott, D. 2018, Appl. Spectrosc., 72, 663
Hyvärinen, A., & Oja, E. 2000, Neural Networks, 13, 411
Izquierdo, A. F., Smith, R. J., Glover, S. C. O., et al. 2021, MNRAS, 500, 5268
Jones, A. P., & Ysard, N. 2019, A&A, 627, A38
Kang, D. E., Pellegrini, E. W., Ardizzone, L., et al. 2022, MNRAS, 512, 617
Kang, D. E., Klessen, R. S., Ksoll, V. F., et al. 2023a, MNRAS, 520, 4981
Kang, D. E., Ksoll, V. F., Itrich, D., et al. 2023b, A&A, 674, A175
Kingma, D. P., & Dhariwal, P. 2018, in Advances in Neural Information Processing Systems, eds. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, (Curran Associates, Inc.), 31
Klessen, R. S., & Glover, S. C. O. 2016, in Saas-Fee Advanced Course, eds. Y. Revaz, P. Jablonka, R. Teyssier, & L. Mayer, 43, 85
Kobyzev, I., Prince, S. J., & Brubaker, M. A. 2021, IEEE Trans. Pattern Anal. Mach. Intell., 43, 3964
Ksoll, V. F., Ardizzone, L., Klessen, R., et al. 2020, MNRAS, 499, 5447
Lallement, R., Capitanio, L., Ruiz-Dern, L., et al. 2018, A&A, 616, A132
Lallement, R., Babusiaux, C., Vergely, J. L., et al. 2019, A&A, 625, A135
Lallement, R., Vergely, J. L., Babusiaux, C., & Cox, N. L. J. 2022, A&A, 661, A147
Leike, R. H., Glatzle, M., & Enßlin, T. A. 2020, A&A, 639, A138
Leike, R. H., Edenhofer, G., Knollmüller, J., et al. 2022, ArXiv e-prints [arXiv:2204.11715]
Li, A., & Draine, B. T. 2001, ApJ, 554, 778
Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. 2018, J. Mach. Learn. Res., 18, 1
Liseau, R., Larsson, B., Lunttila, T., et al. 2015, A&A, 578, A131
Loinard, L., Torres, R. M., Mioduszewski, A. J., & Rodríguez, L. F. 2008, ApJ, 675, L29
Loren, R. B., Wootten, A., & Wilking, B. A. 1990, ApJ, 365, 269
Lucy, L. B. 1999, A&A, 344, 282
Mathis, J. S., Rumpl, W., & Nordsieck, K. H. 1977, ApJ, 217, 425
Mathis, J. S., Mezger, P. G., & Panagia, N. 1983, A&A, 128, 212
Molinari, S., Swinyard, B., Bally, J., et al. 2010, A&A, 518, A100
Molinari, S., Schisano, E., Elia, D., et al. 2016, A&A, 591, A149
Nelson, R. P., & Langer, W. D. 1997, ApJ, 482, 796
Paszke, A., Gross, S., Chintala, S., et al. 2017, Automatic Differentiation in PyTorch, in NIPS Autodiff Workshop
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Planck Collaboration XIX. 2011, A&A, 536, A19
Planck Collaboration XI. 2014, A&A, 571, A11
Planck Collaboration XXXV. 2016, A&A, 586, A138
Reissl, S., Wolf, S., & Brauer, R. 2016, A&A, 593, A87
Reissl, S., Klessen, R. S., Mac Low, M.-M., & Pellegrini, E. W. 2018, A&A, 611, A70
Reissl, S., Brauer, R., Klessen, R. S., & Pellegrini, E. W. 2019, ApJ, 885, 15
Reissl, S., Meehan, P., & Klessen, R. S. 2023, A&A, 674, A47
Rezaei Kh., S., & Kainulainen, J. 2022, ApJ, 930, L22
Rezaei Kh., S., Bailer-Jones, C. A. L., Hanson, R. J., & Fouesneau, M. 2017, A&A, 598, A125
Rezaei Kh., S., Bailer-Jones, C. A. L., Soler, J. D., & Zari, E. 2020, A&A, 643, A151
Rezende, D., & Mohamed, S. 2015, in Proceedings of Machine Learning Research, Proceedings of the 32nd International Conference on Machine Learning, eds. F. Bach, & D. Blei (Lille, France: PMLR), 37, 1530
Santos, F. P., Chuss, D. T., Dowell, C. D., et al. 2019, ApJ, 882, 113
Silverman, B. W. 1986, Density Estimation for Statistics and Data Analysis (Chapman and Hall)
Smith, R. J., Glover, S. C. O., Clark, P. C., Klessen, R. S., & Springel, V. 2014, MNRAS, 441, 1628
Smith, R. J., Treß, R. G., Sormani, M. C., et al. 2020, MNRAS, 492, 1594
Springel, V. 2010, MNRAS, 401, 791
Steinacker, J., Bacmann, A., Henning, T., Klessen, R., & Stickel, M. 2005, A&A, 434, 167
Tabak, E., & Vanden-Eijnden, E. 2010, Commun. Math. Sci., 8, 217
Tabak, E., & Turner, C. 2013, Commun. Pure Appl. Math., 66, 145
Tielens, A. G. G. M. 2010, The Physics and Chemistry of the Interstellar Medium
Tress, R. G., Smith, R. J., Sormani, M. C., et al. 2020, MNRAS, 492, 2973
Valtchanov, I. 2018, The Spectral And Photometric Imaging Receiver (SPIRE) Handbook, HERSCHEL-HSC-DOC-0798, version 3.2
Wolfire, M. G., McKee, C. F., Hollenbach, D., & Tielens, A. G. G. M. 2003, ApJ, 587, 278
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Zhang, X., Green, G. M., & Rix, H.-W. 2023, MNRAS, 524, 1855
Zucker, C., Goodman, A., Alves, J., et al. 2021, ApJ, 919, 35
Zucker, C., Goodman, A. A., Alves, J., et al. 2022, Nature, 601, 334
Zucker, C., Alves, J., Goodman, A., Meingast, S., & Galli, P. 2023, in Protostars and Planets VII, eds. S. Inutsuka, Y. Aikawa, T. Muto, K. Tomida, & M. Tamura, ASP Conf. Ser., 534, 43

Appendix A: Training set generation

This appendix provides additional material with regard to the construction of our training data. Table A.1 provides a summary of the different bands we consider in our generation of the synthetic dust emission maps in Section 3. Listed are the instrument name, band ID, the central wavelength we assume for the band, and the reference indicating from where we have extracted the respective information. Figures A.1 and A.2 provide histograms for the effective prior distributions for the dust density and temperature, and the simulated fluxes in all wavelengths that the cINN sees during training, summarised over all the pixels in our training data set. We emphasise that these effective priors are an outcome of our training data selection, depicting the actual distributions of the respective parameters in the final training data set, and not a prescription after which the training data is selected. Lastly, Figure A.3 shows an example of the emission maps generated by POLARIS at the 23 different wavelengths that we consider for the example cube used in Figures 3 (top row), B.1, B.2, and B.3 (top row).

Table A.1: Overview of filter bands and instruments considered in this study.

Instrument	Band	$\lambda(\mu m)$	Reference
WISE	3	12	Wright et al. (2010)
WISE	4	22	Wright et al. (2010)
MSX/SPIRIT-III	2	12.13	Egan et al. (1999)
MSX/SPIRIT-III	3	14.65	Egan et al. (1999)
MSX/SPIRIT-III	4	21.34	Egan et al. (1999)
Spitzer	MIPS	24	Table 1, Dole et al. (2003)
SOFIA / HAWC+	A	53	Harper et al. (2018)
SOFIA / HAWC+	B	63	Harper et al. (2018)
SOFIA / HAWC+	C	89	Harper et al. (2018)
SOFIA / HAWC+	D	154	Harper et al. (2018)
SOFIA / HAWC+	E	214	Harper et al. (2018)
Herschel	PACS 1	70	PACS Photometer Quickstart Guide, Exter (2017)
Herschel	PACS 2	100	PACS Photometer Quickstart Guide, Exter (2017)
Herschel	PACS 3	160	PACS Photometer Quickstart Guide, Exter (2017)
Herschel	SPIRE 1	250	Table 5.2, SPIRE Handbook, Valtchanov (2018)
Herschel	SPIRE 2	350	Table 5.2, SPIRE Handbook, Valtchanov (2018)
Herschel	SPIRE 3	500	Table 5.2, SPIRE Handbook, Valtchanov (2018)
ALMA	Band 10	350	Table 7.1, Cortes et al. (2022)
			Table 7.1, Cortes et al. (2022)
ALMA	Band 9	460	Table 7.1, Cortes et al. (2022)
			Table 7.1, Cortes et al. (2022)
ALMA	Band 8	690	Table 7.1, Cortes et al. (2022)
			Table 7.1, Cortes et al. (2022)
ALMA	Band 7	945	Table 7.1, Cortes et al. (2022)
			Table 7.1, Cortes et al. (2022)
ALMA	Band 6	1300	Table 7.1, Cortes et al. (2022)
			Table 7.1, Cortes et al. (2022)
APEX	LABOCA	870	APEX LABOCA website
			(https://www.apex-telescope.org/ns/laboca-calibration/)
CSO	Bolocam	1100	Bolocam website
			(http://www.cso.caltech.edu/bolocam/ProposerInfo.html)

Notes. For each instrument and band, we list the central wavelength λ and the literature reference for the quoted values. We note that we treat ALMA Band 10 and Herschel SPIRE 2 as one band in our analysis, because they share the same central wavelength and we are not considering any instrument related effects. Consequently, there is no difference in the corresponding synthetic dust emission maps between SPIRE 2 and ALMA Band 10 in our analysis.

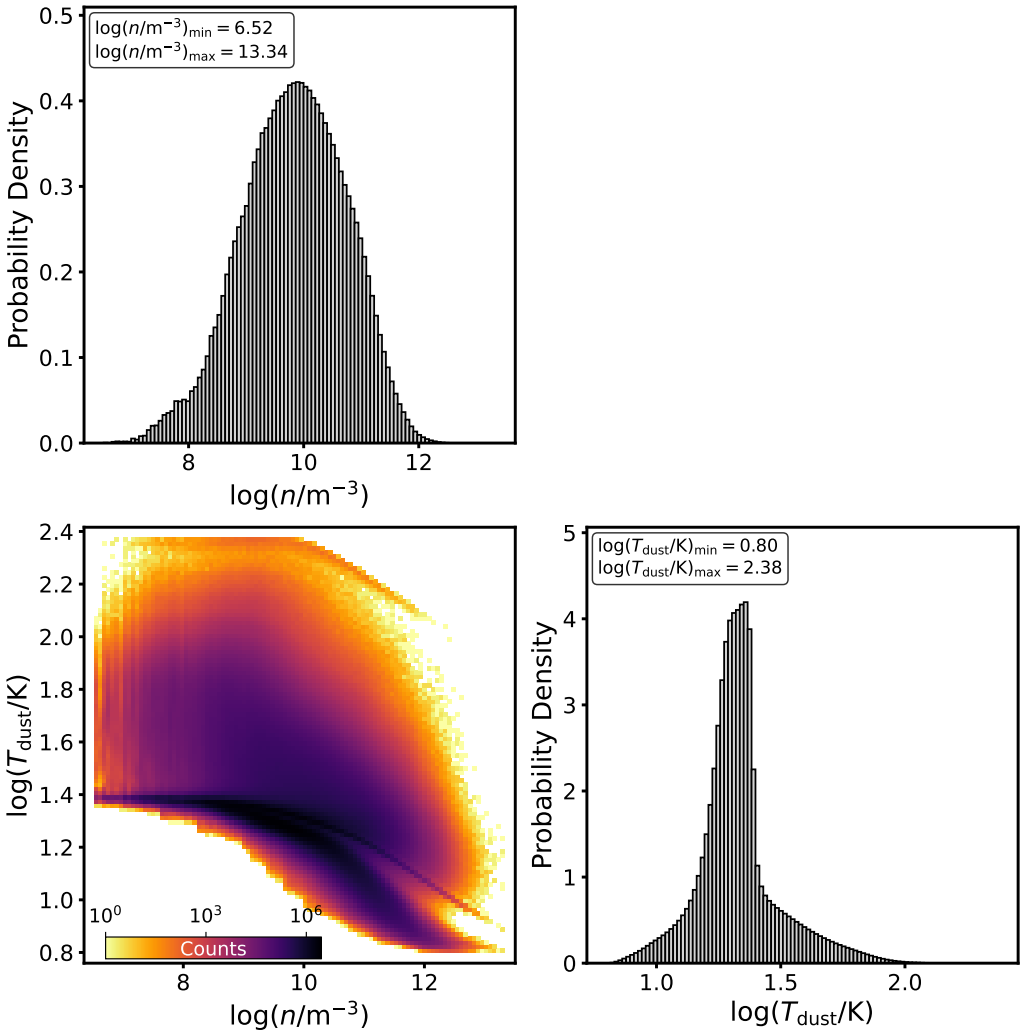

Fig. A.1: Histograms of the prior distributions and correlation of the target parameters in the training data over all pixels. The top-left and bottom-right panels show the 1D distributions of the number density and dust temperature, respectively. In both panels, the boxes at the top left provide the minimum and maximum of the respective parameter. The bottom left panel presents a 2D histogram of the effective prior distribution in the combined density-temperature space.

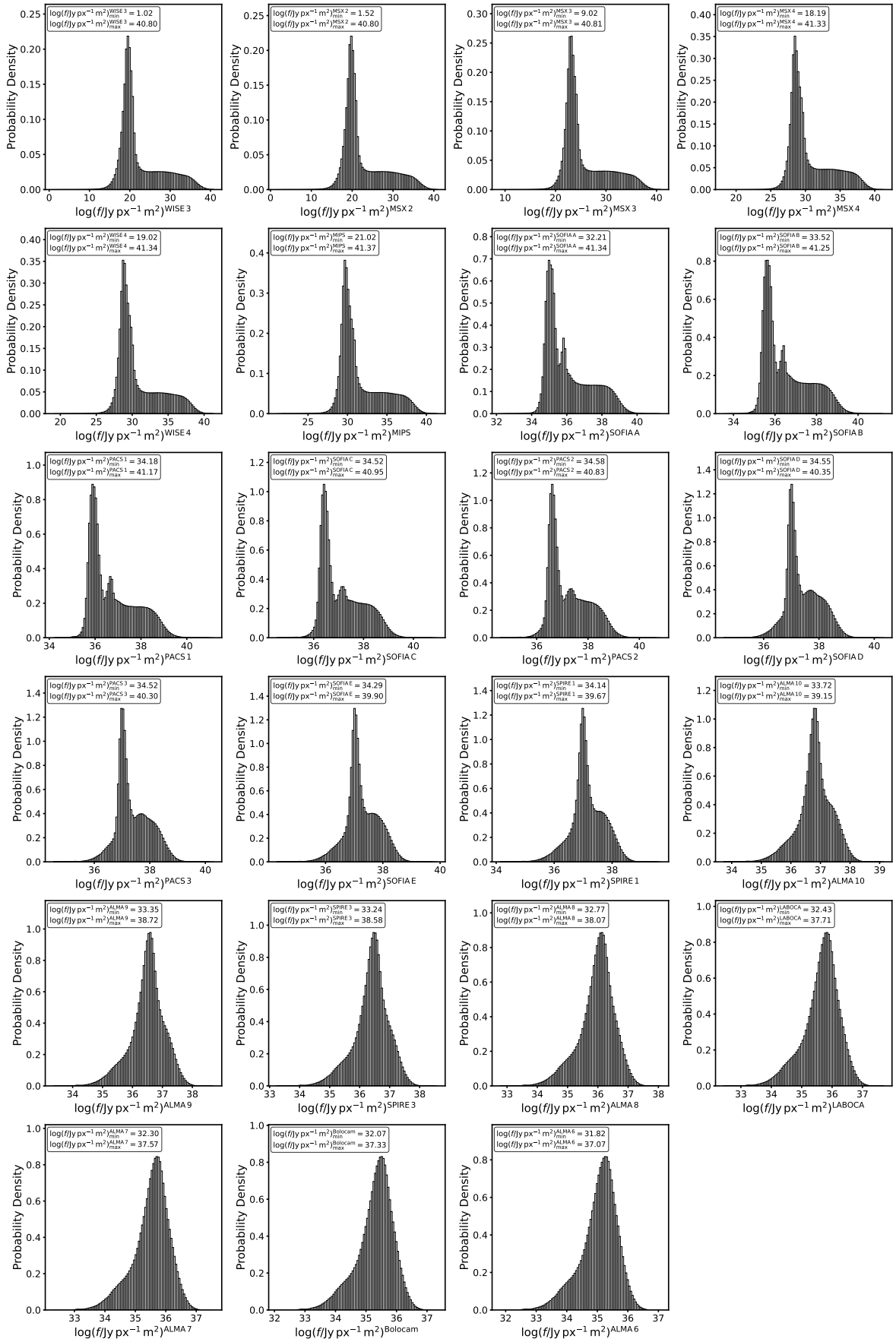

Fig. A.2: Histograms of the prior distributions for the observables in the training data over all cubes and pixels

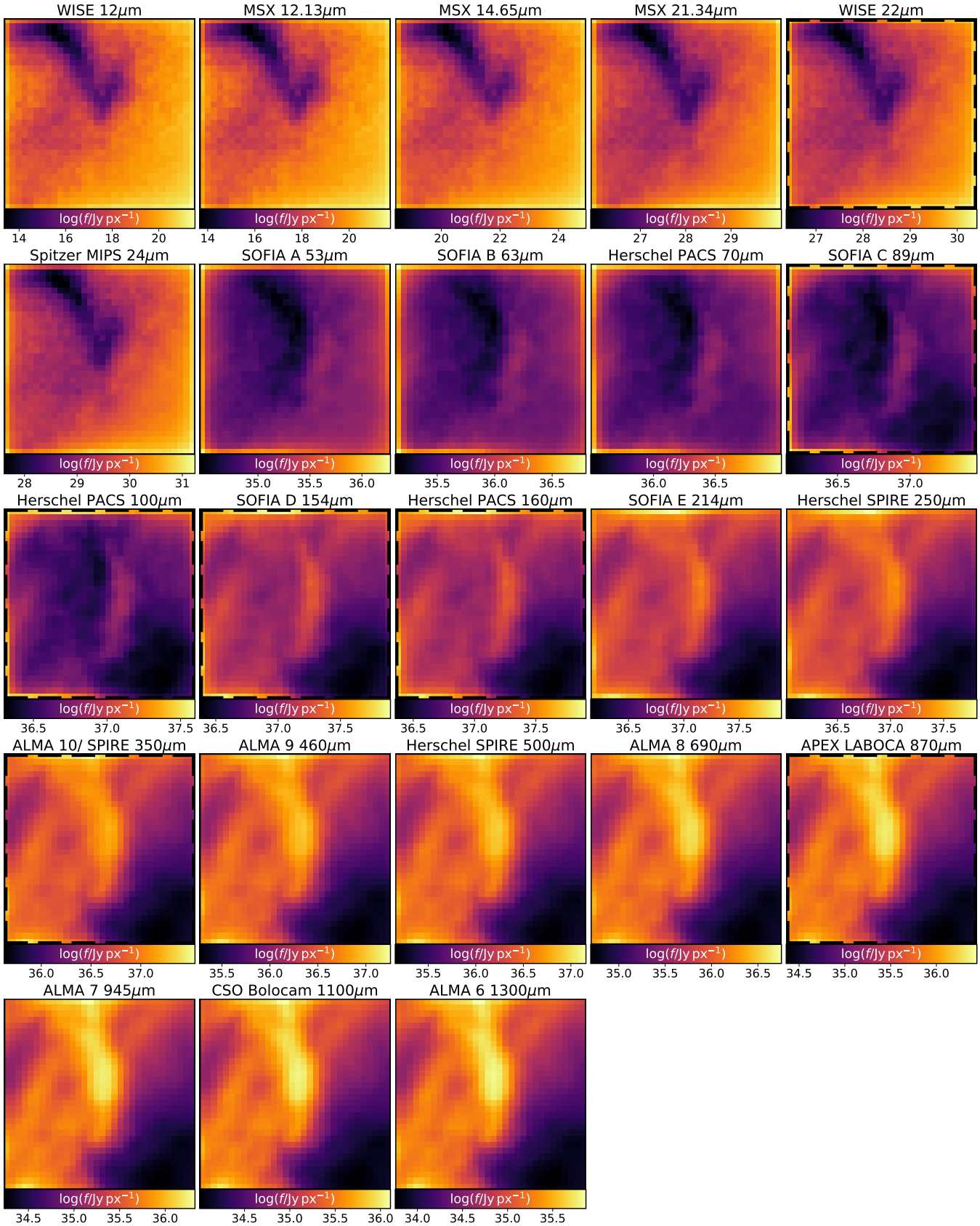

Fig. A.3: Example of the synthetic dust emission maps at all 23 considered wavelengths that serve as the input to our cINN approach. These correspond to bands of specific instruments (as labelled on the top of each panel), but do not account for PSF-related resolution effects of the respective telescopes. The shown example corresponds to the example cube in the ISRF-only configuration that is also the subject of Figures 3, B.1, and B.2. The panels outlined with the black dashed lines correspond to the seven wavelengths considered in our more limited experiment in Section 4.3. We emphasise that the presented flux maps are corrected for distance.

Appendix B: Additional material for the performance analysis

This appendix provides complementary diagrams for the performance analysis of our cINN approach outlined in Section 4. Figure B.1 provides an extended comparison of the point estimation results to the ground truth for the example cube shown in the first to rows of Figure 3, depicting all 32 slices of the cube. As an alternative visualisation of this comparison, Figure B.2 shows 3D isodensity surface diagrams for densities of 10^{10} (grey) and 10^{11} m^{-3} (red) for two different rotation angles of the example cube, comparing again the results of our three point estimators to the ground truth. As discussed in Section 4.1, this diagram illustrates the very good recovery of the 3D dust structure at intermediate densities, whereas some finer, high-density structures may be subject to underestimation and are, thus, lost in this visualisation. As an extension to Figure 3, Figure B.3 provides

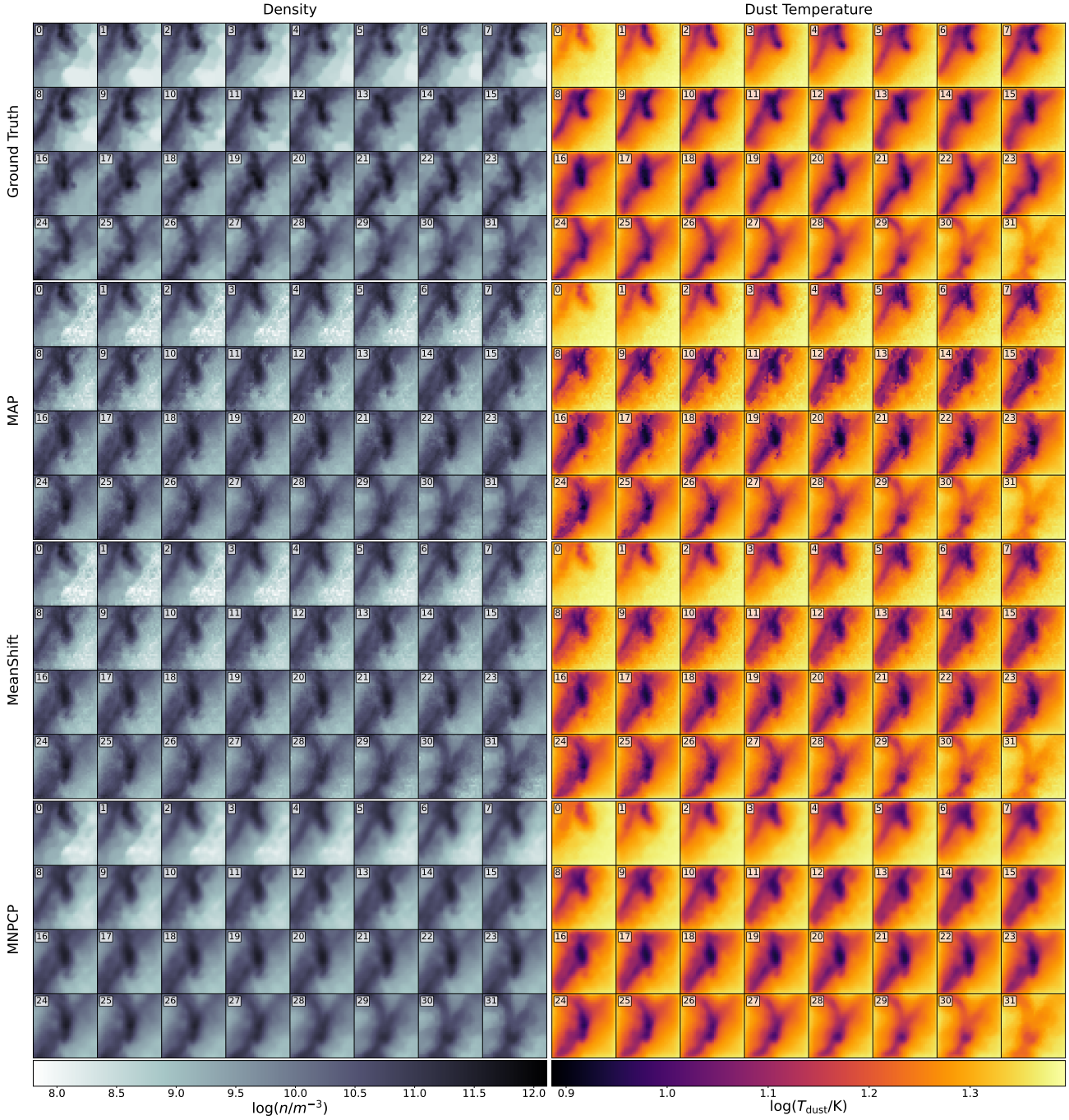

Fig. B.1: Comparison of the point estimate prediction results to the ground truth for all slices of one example cube that is only subject to the ISRF. In each panel, the subpanels show from top left to bottom right the cube slices going along the LoS from front to back. The left and right columns show the dust density and dust temperature respectively. From top to bottom, the rows indicate the ground truth and the MAP, MeanShift, and MNPCP estimates based on the outcome of the single LoS cINN using all 23 wavelengths. This diagram shows the full cube of the single slices shown in the first two rows of Figure 3.

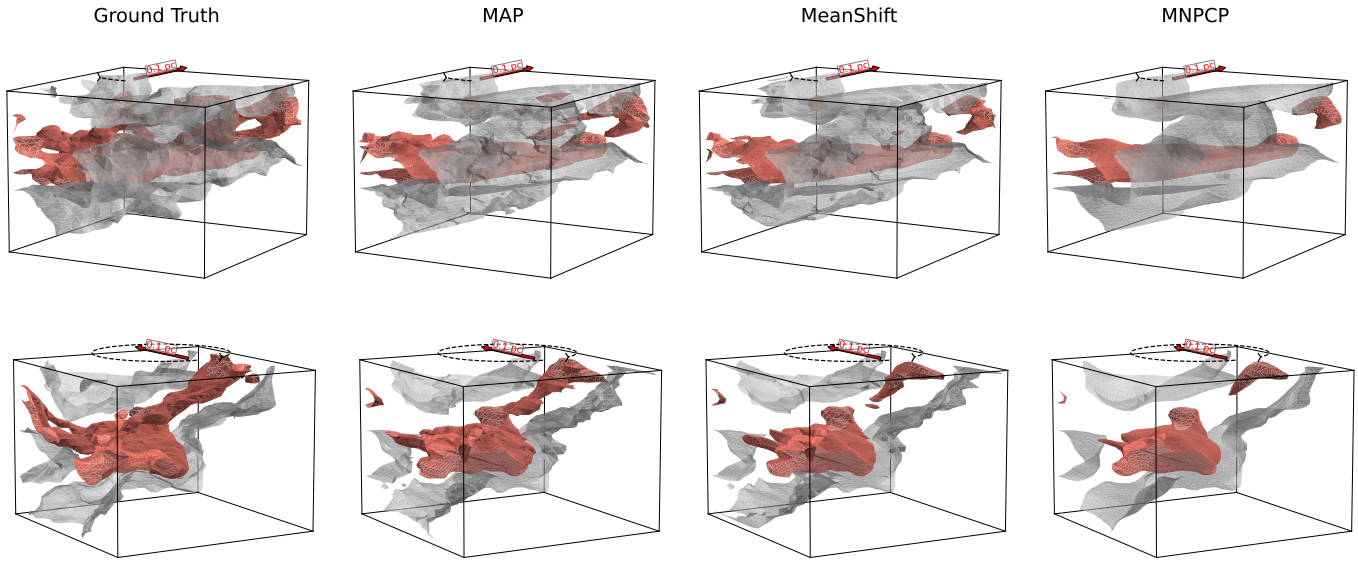

Fig. B.2: 3D isodensity surface diagrams for one example test cube in the ISRF-only radiation configuration, comparing the SLoS-cINN prediction results to the ground truth. Here the two rows show different rotation angles of the given cube. The grey surfaces indicate a density of 10^{10} m^{-3}, whereas red surfaces mark the 10^{11} m^{-3} density level.

Table B.1: Performance summary for the viewing angle comparison experiment.

			Viewing angle		
Point estimator	Measure	Parameter	Front face	Right face	Bottom face
MAP	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0510	0.0531	0.0517
		$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0633	0.0682	0.0648
	$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$1.41^{+1.45}_{-0.83}$	$1.53^{+1.58}_{-0.90}$	$1.38^{+1.39}_{-0.81}$
		$\log(T_{\mathrm{dust}}/\mathrm{K})$	$0.61^{+1.03}_{-0.39}$	$0.66^{+1.16}_{-0.43}$	$0.58^{+0.96}_{-0.37}$
MeanShift	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0480	0.0499	0.0490
		$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0522	0.0561	0.0455
	$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$1.53^{+1.42}_{-0.88}$	$1.62^{+1.52}_{-0.94}$	$1.48^{+1.38}_{-0.85}$
		$\log(T_{\mathrm{dust}}/\mathrm{K})$	$0.70^{+1.12}_{-0.45}$	$0.73^{+1.23}_{-0.47}$	$0.65^{+1.03}_{-0.41}$
MNPCP	NRMSE	$\log(n/\mathrm{m}^{-3})$	0.0456	0.0476	0.0455
		$\log(T_{\mathrm{dust}}/\mathrm{K})$	0.0506	0.0545	0.0501
	$\lvert\bar{e}_{\mathrm{rel}}\rvert$ (%)	$\log(n/\mathrm{m}^{-3})$	$1.40^{+1.33}_{-0.80}$	$1.50^{+1.44}_{-0.87}$	$1.37^{+1.30}_{-0.79}$
		$\log(T_{\mathrm{dust}}/\mathrm{K})$	$0.70^{+1.02}_{-0.42}$	$0.74^{+1.15}_{-0.45}$	$0.67^{+0.96}_{-0.40}$

Notes. NRMSE and the median absolute relative error $\lvert\bar{e}_{\mathrm{rel}}\rvert$ (along with the 25% and 75% quantiles) are given, respectively, for the dust density and temperature prediction evaluated across all pixels of the 50 test cubes subject to the ISRF-only case. All predictions are made with the 23-wavelength SLoS-cINN here (Section 4.1).

examples for the predicted posterior distributions in the different network and cube configurations analysed in Figure 3.

Analogously to Figures B.1 and B.2, Figures B.4 and B.5 show the respective results for the cube in the ISRF + star radiation configuration (rows 2&3 of Figure 3), Figures B.6 and B.7 present the outcome as derived from the NLoS-cINN (Section 4.2), and Figures B.8 and B.9 illustrate the predictions from the SLoS-cINN that is limited to the seven wavelengths configuration (Section 4.3).

Appendix B.1. Influence of the observation angle

In Section 2.2, we outline that during our training set generation we only simulate synthetic observations for the simulation cubes from one viewing angle. In particular, it our simulated observations are performed from the same direction for all training cubes. Although somewhat unlikely, this circumstance begs the question, whether the chosen viewing angle may have introduced a bias in our trained method, so the reconstruction from other viewing angles may differ. Naturally, the simulated synthetic observations depend on the viewing angle of the cloud as the contribution of different parcels of gas in the cloud to the observed dust emission depends on the 3D position after all. Therefore, it is to be expected that a reconstruction of the same cloud from two different viewing angles is not 100% identical. To test how the viewing angle may affect the reconstructive performance and rule out a direction bias in our method, we conducted an experiment where we rotated the 50 cubes that

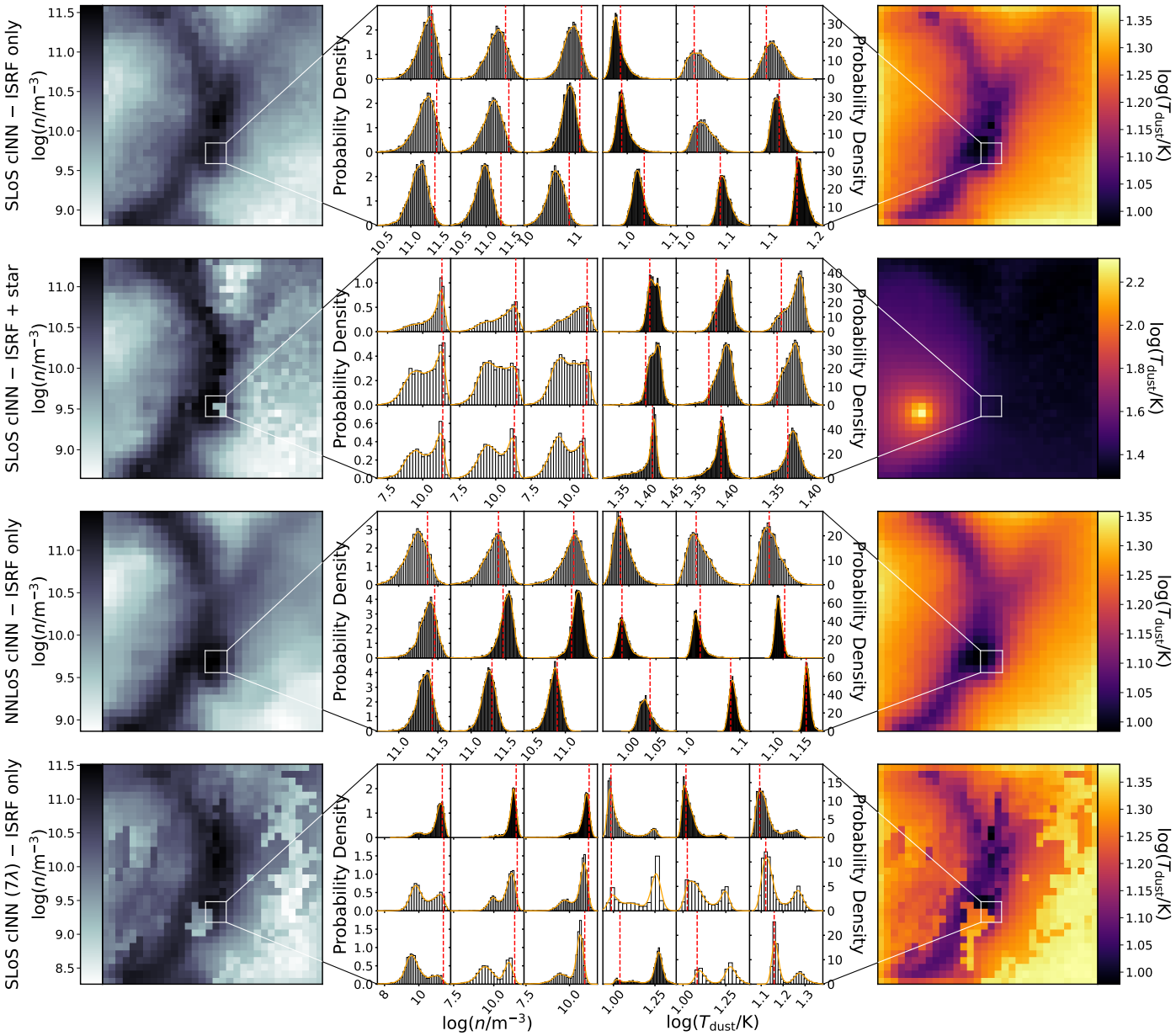

Fig. B.3: Comparison of the predicted posterior distributions for a few example pixels of the cubes shown in Figure 3. The panels in the leftmost and right-most column show the MAP prediction results for density and dust temperature for the same slice as in Figure 3, respectively. The middle two columns provide histograms of the predicted posterior distributions for the nine pixels indicated by the white square in the left-most and right-most columns. The orange curve represents the kernel density estimate used to determine the MAP values, whereas the red dashed line mark the respective ground truth value of each example pixel. We note that within each of the posterior histogram panels, the subpanels share the same x axis column-wise and the same y axis row-wise.

are only subject to the ISRF, resimulating the corresponding synthetic dust emission observations and then reconstructing again the 3D distribution of the dust with the 23-wavelength SLoS-cINN. Compared to the main viewing angle, for which we present the performance in Section 4.1, we tested two additional observation directions. In the first, the cubes are rotated by 90° clockwise around the vertical axis in the plane of the primary viewing angle; namely, we are now observing the right face of the cube instead of the front face. For the second test, we rotated the cube around the horizontal axis by 90° clockwise, so that we are now observing the bottom face of the cube.

Figure B.10 provides an example comparison of the reconstruction result for one slice (along the original LoS) of one example cube (the same as in Figure 3) from the three different viewing angles that we have tested. As we can see, the cINN recovers the bulk of the structure in this slice quite well, independently of the observation direction. Naturally, smaller details do differ a bit, but as mentioned before this is to be expected, both because the information encoded in the synthetic dust emission maps may not be identical and because the spatial consistency of the predicted posterior distributions depends on the viewing angles. Where the original viewing angles produces posteriors that are consistent along the LoS going into the plane of the

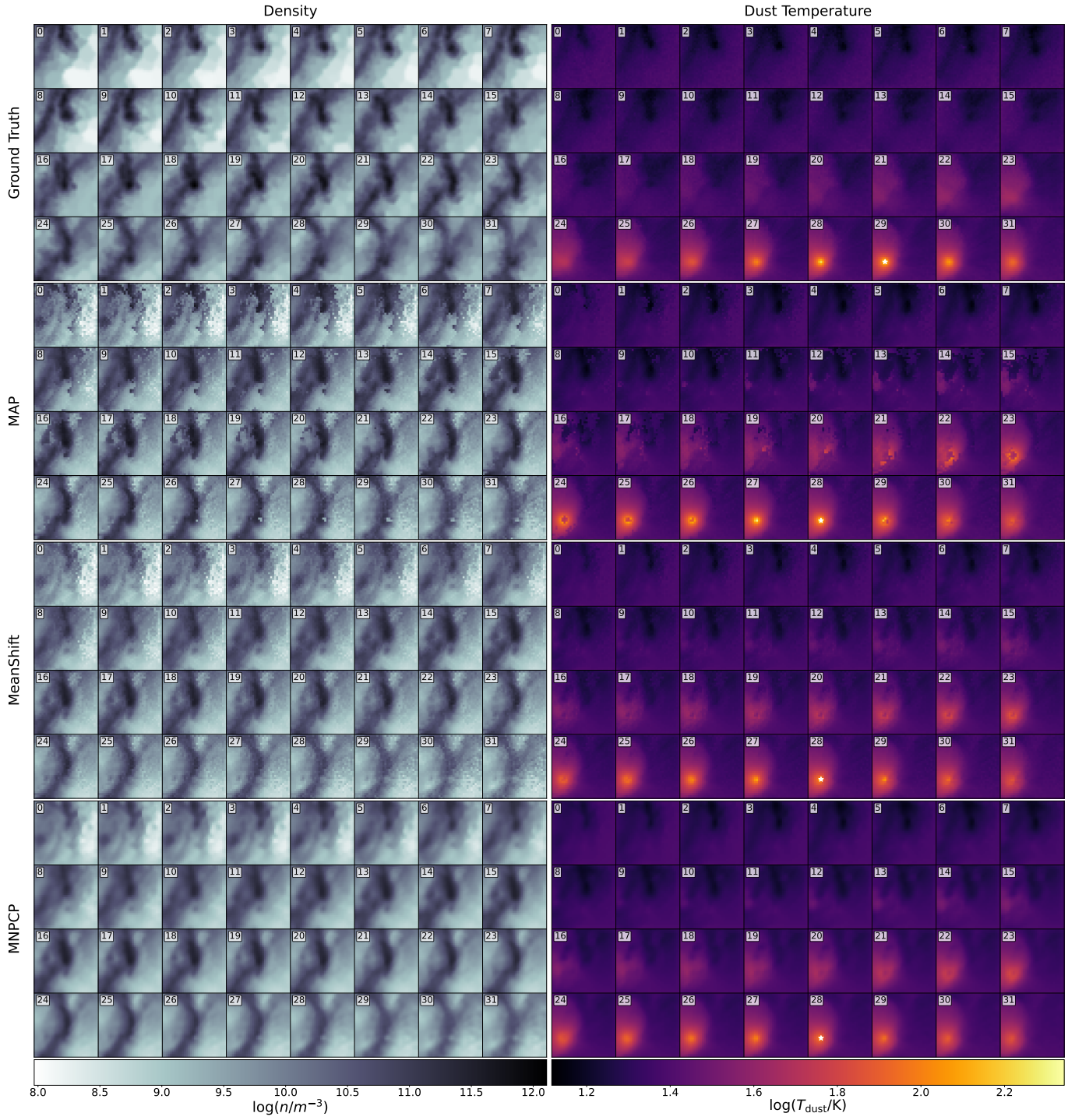

Fig. B.4: Comparison of the point estimate prediction results to the ground truth for all slices of one example cube that is subject to the ISRF and one B4 star. In each panel, the subpanels show (from top-left to bottom-right) the cube slices going along the LoS from front to back. The left and right column show the dust density and dust temperature respectively. From top to bottom, the rows indicate the ground truth and the MAP, MeanShift and MNPCP estimates, based on the outcome of the SLoS-cINN using all 23 wavelengths. This diagram shows the full cube of the single slice shown in rows 3 and 4 of Figure 3. The white star symbol in the right column indicates the approximate position of the star, defined as the hottest pixel in the cube for both the ground truth and the prediction results.

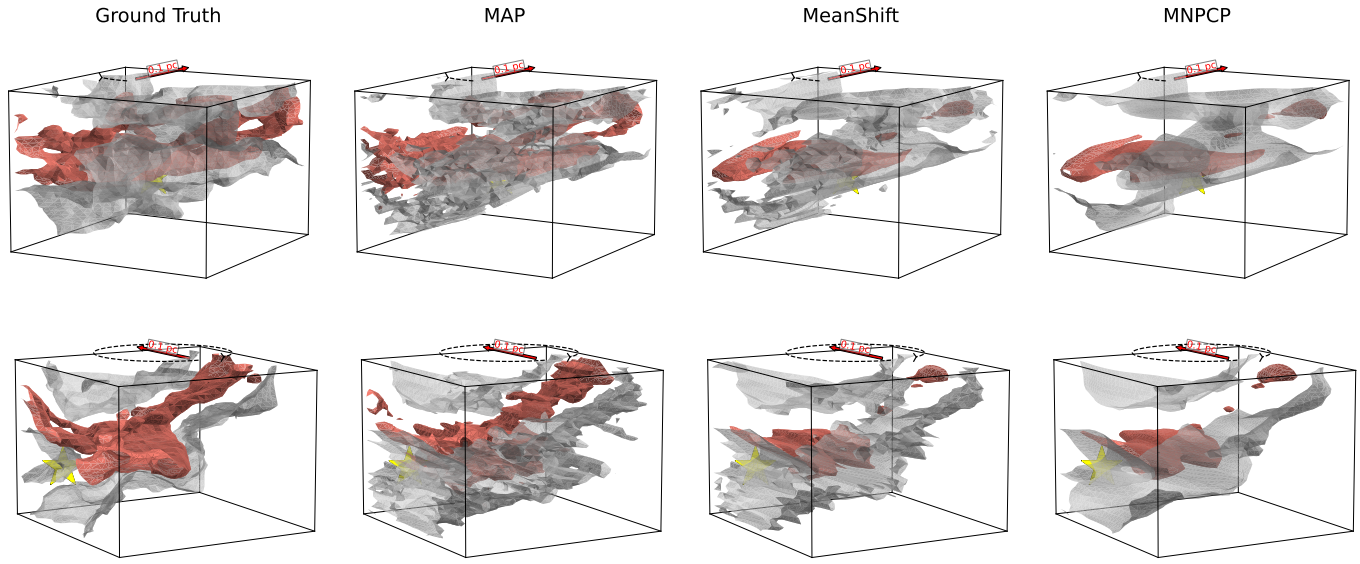

Fig. B.5: 3D isodensity surface diagrams for one example test cube in the ISRF + star radiation configuration, comparing the SLoS-cINN prediction results to the ground truth. Here the two rows show different rotation angles of the given cube. The grey surfaces indicate a density of 10^{10} m^{-3}, whereas red surfaces mark the 10^{11} m^{-3} density level. The yellow star indicates the position of the B4 anaologue placed inside the cube.

shown slice, the other observation directions provide consistency along the perpendicular LoSs now. As a consequence, features in the reconstruction may appear a little more elongated along the given LoS compared to the other viewing angles, as becomes apparent when comparing the two right columns in Figure B.10. Table B.1 provides a summary of the predictive performance of the 23-wavelength SLoS-cINN for the three different viewing angles across all 50 test cubes. As we can see, the performance is very similar, independently of the direction from where the cubes were observed. As such, we can conclude that the cINN has not been biased towards a prefered observational direction and can reconstruct the synthetic dust clouds quite reliably even from different viewing angles.

Appendix B.2. Influence of the instrument PSF

In Section 2.2, we are ignoring any instrument related effects in our generation of the synthetic dust emission observations used throughout the rest of this study. Our reasoning for this is that proper modelling of these effects is not trivial. Nevertheless, the latter will be a necessary step in the ongoing development of our approach in order to ultimately apply it to real observational data. In this appendix, we perform a simplified test to demonstrate how a cINN that is trained on fully resolved data behaves on data that is subject to resolution effects. For this purpose, we took the 50 test cubes subject only to the ISRF and modified their corresponding synthetic dust emission maps to simulate the influence of the PSFs of the respective telescope instruments. For simplicity, we approximated this effect by convolving the fully resolved synthetic dust emission maps with a Gaussian, where the standard deviations are matched to the full width half maxima (FWHM) of the PSFs. Notably, this includes the implicit assumption that all considered telescopes would have the same pixel size; in this case, matching that of our input simulated data.

For the convolution, we assumed a distance to the observed cubes, at which the PSFs of all considered instruments are sampled by at least two pixels in the dust emission maps (i.e. to fulfil the Nyquist sampling criterion). Given the excellent spatial resolution of ALMA, this distance would be quite large if we were to consider all 23 simulated wavelengths (namely larger than 8500 pc). Therefore, we conduct this test only on the seven wavelength subset used in Section 4.3, for which the PSF sampling criterion is fulfilled at a distance of $d = 397$ pc. Figure B.11 shows an example for how this PSF consideration affects the input dust emission maps (analogously to Figure A.3, we select the same example cube here). As we can see, if we require that all seven considered filters are Nyquist sampled, meaning that the selected distance is naturally dictated by the instrument with the best resolution out of the seven (in this case Herschel PACS at 100 μm), then there is already a notable loss in detail in the input dust emission maps for instruments with a worse resolution.

We went on to test how a cINN (namely the seven wavelength SLoS-cINN discussed in Section 4.3) that is trained on perfectly resolved data performs on input observations that are subject to these varying resolution effects at the 397 pc distance on the 50 ISRF-only test cubes. Figure B.12 presents an example MAP prediction result in comparison to the ground truth for the same cube shown in Figure B.1. As we can see, the cINN can no longer recover the dust densities and temperatures at all, a result that we find for all 50 tested cubes. This outcome is not surprising, however, as this resolution effect is not small, rendering data affected by it likely quite far outside of the observable domain that this cINN has learned, where it is bound to fail. While this particularly strong failure here might be exacerbated by the limited wavelength coverage or filter choice, this still confirms the necessity to account for instrumental effects already during training.

To identify at which point the prediction breaks down, we followed up with an additional test that further simplifies the consideration of the PSF effects. Instead of convolving the emission maps in each wavelength with the (approximate) PSF of the corresponding instrument filter, we applied the same PSF resolution to all wavelengths and investigated the beam size at which the prediction starts to deteriorate. For simplicity, we directly

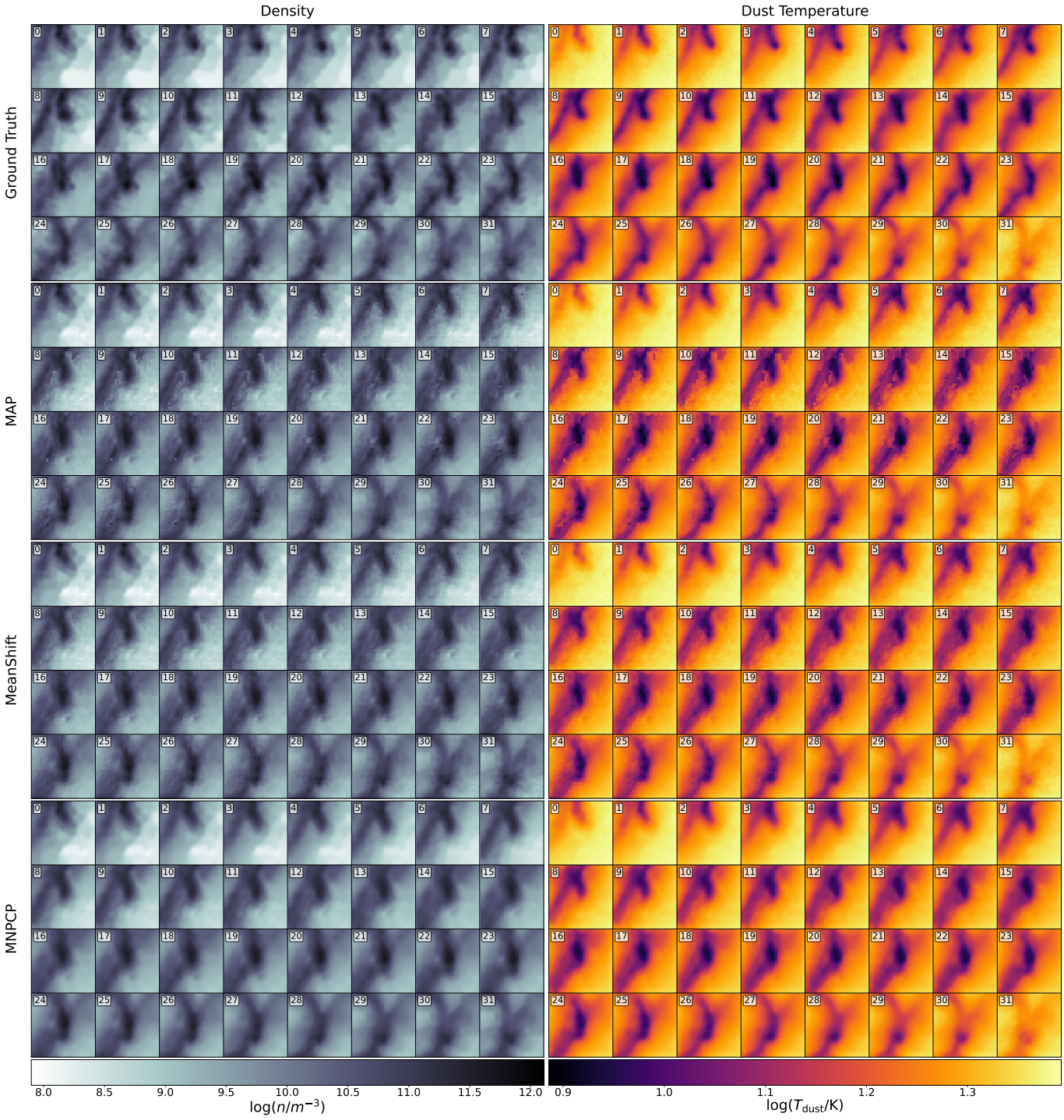

Fig. B.6: Comparison of the point estimate prediction results to the ground truth for all slices of one example cube that is only subject to the ISRF. In each panel, the subpanels show from top left to bottom right the cube slices going along the LoS from front to back. The left and right columns show the dust density and dust temperature, respectively. From top to bottom, the rows indicate the ground truth and the MAP, MeanShift, and MNPCP estimates based on the outcome of the NLoS-cINN using all 23 wavelengths (as opposed to the SLoS-cINN outcome shown in Fig. B.1). This diagram shows the full cube of the single slice shown in rows 5&6 of Figure 3.

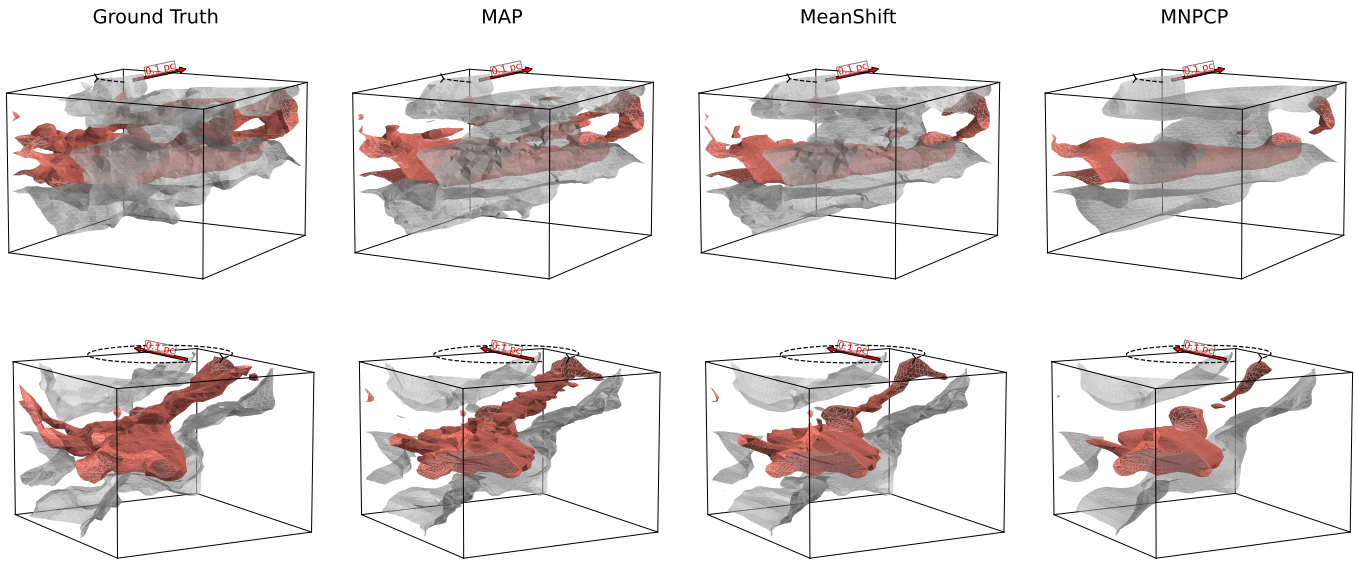

Fig. B.7: 3D isodensity surface diagrams for one example test cube in the ISRF-only radiation configuration, comparing the NLoS-cINN prediction results to the ground truth in the left column. Here, the two rows show different rotation angles of the given cube. The grey surfaces indicate a density of $10^{10}\,\mathrm{m}^{-3}$, whereas red marks the $10^{11}\,\mathrm{m}^{-3}$ density level. These data are analogous to the SLoS-cINN results in Figure B.2.

specified the FWHM of the Gaussian convolution kernel in pixels for this experiment. For reference, the tested FWHM values of 1.5, 2, 3, 5, and 10 pixels correspond to distances of 79.9, 106.5, 159.8, 266.4, and 532.7 pc, respectively, when using (for instance) the Herschel SPIRE 350 µm filter. Figure B.13 provides a summary of this test for the same example cube (in the ISRF-only configuration) used in Figure B.12, comparing the prediction results on the smoothed emission maps for varying FWHMs to the reference outcome on the original synthetic observations. As we can see, the cINN predictions are somewhat robust with respect to the smoothing of the input emission maps up to a FWHM of the Gaussian beam of 3 pixels, exhibiting only small changes in the NRMSEs for this example. However, the prediction outcome rapidly deteriorates for the larger tested FWHMs of 5 and 10 pixels. Given this outcome, the failure in the previous test might be explained by a combination of the complexity of varying resolution at each wavelength and some rather large PSFs at the tested distance for the central wavelengths of Herschel SPIRE 350µm and APEX LABOCA. In conclusion, this test shows that our approach can compensate a minor degree of unaccounted for resolution effects if they are the same at all wavelengths. Nevertheless, for realistic applications with variations between observational instruments, proper modelling of the instrumental effects within the training data will be necessary to build a truly robust reconstruction model.

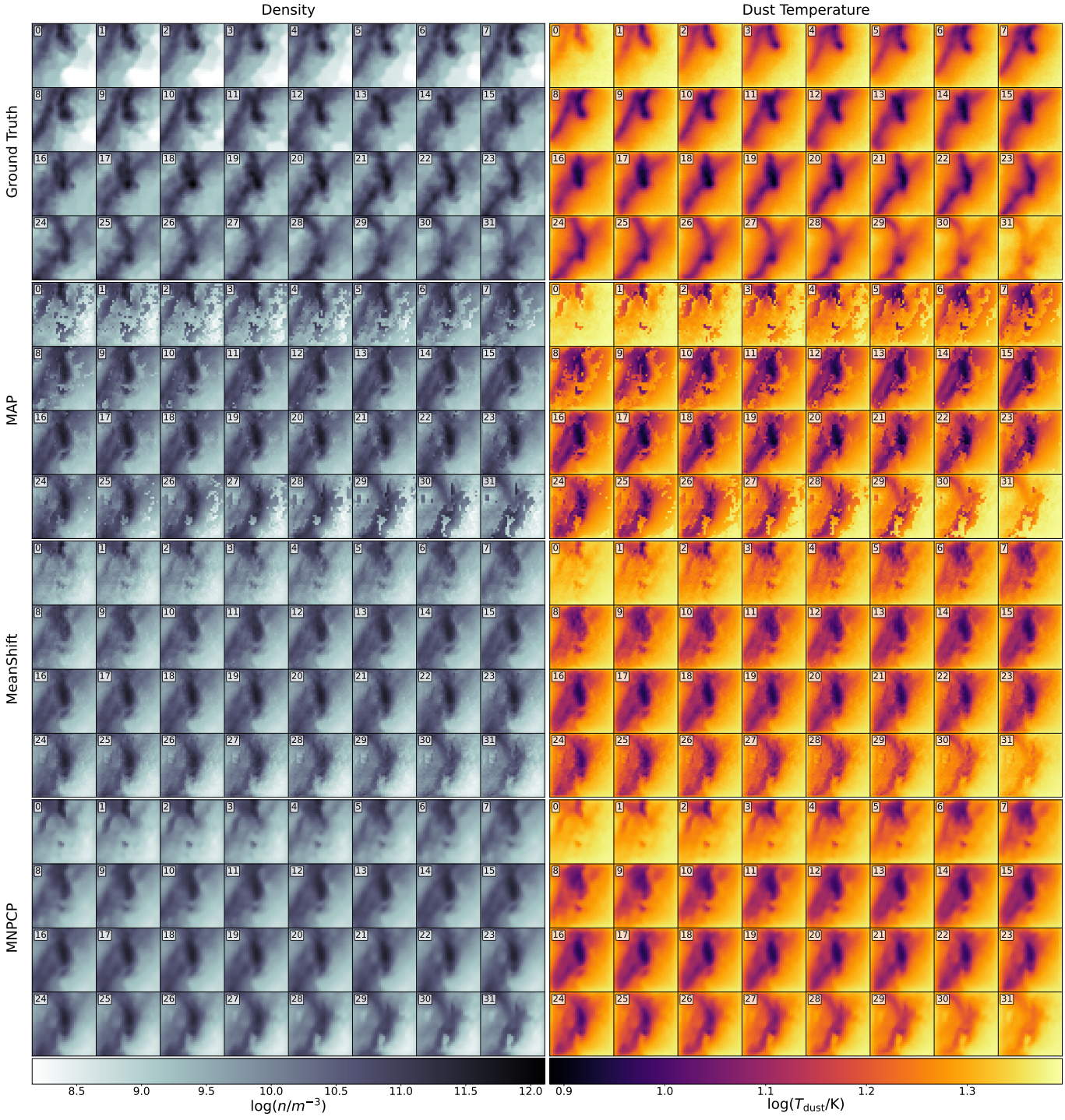

Fig. B.8: Comparison of the point estimate prediction results to the ground truth for all slices of one example cube that is only subject to the ISRF. In each panel, the subpanels show (from top-left to bottom-right) the cube slices going along the LoS from front to back. The left and right columns show the dust density and dust temperature respectively. From top to bottom the rows indicate the ground truth and the MAP, MeanShift, and MNPCP estimates based on the SLoS-cINN that uses only seven wavelengths. This diagram shows the full cube of the single slice shown in the last two rows of Figure 3.

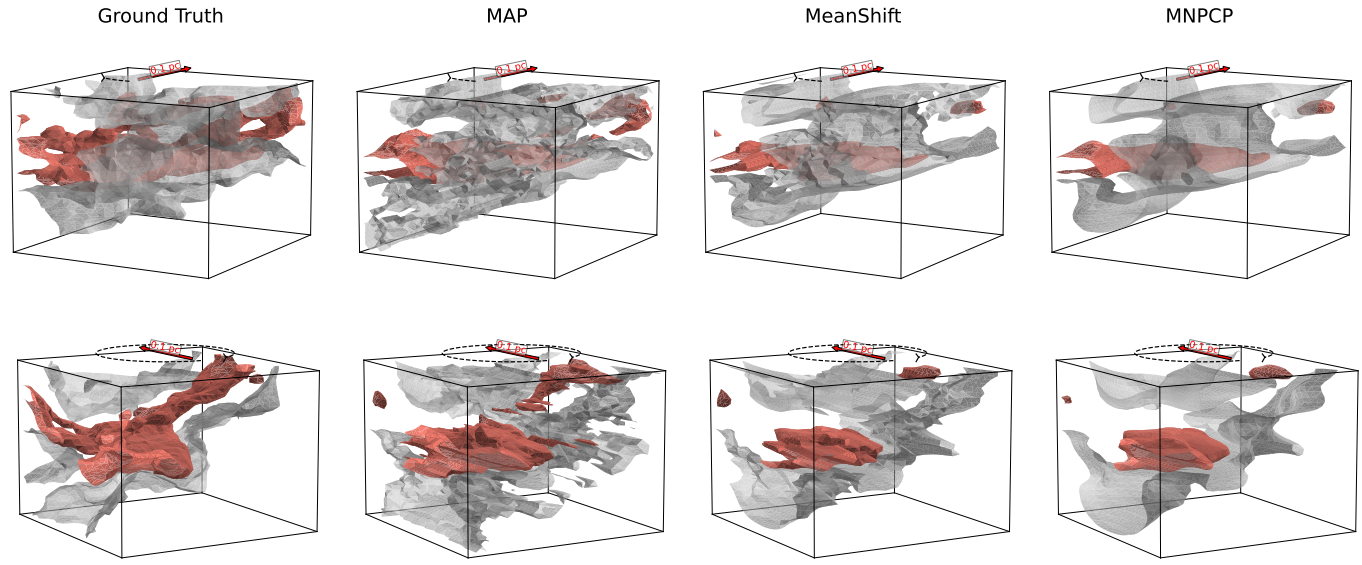

Fig. B.9: 3D isodensity surface diagrams for one example test cube in the ISRF-only radiation configuration, comparing the SLoS-cINN prediction results, based on only seven wavelengths, to the ground truth. Here, the two rows show different rotation angles of the given cube. The grey surfaces indicate a density of 10^{10} m^{-3}, whereas red surfaces mark the 10^{11} m^{-3} density level. These data are analogous to the 23-wavelength SLoS results shown in Figure B.2.

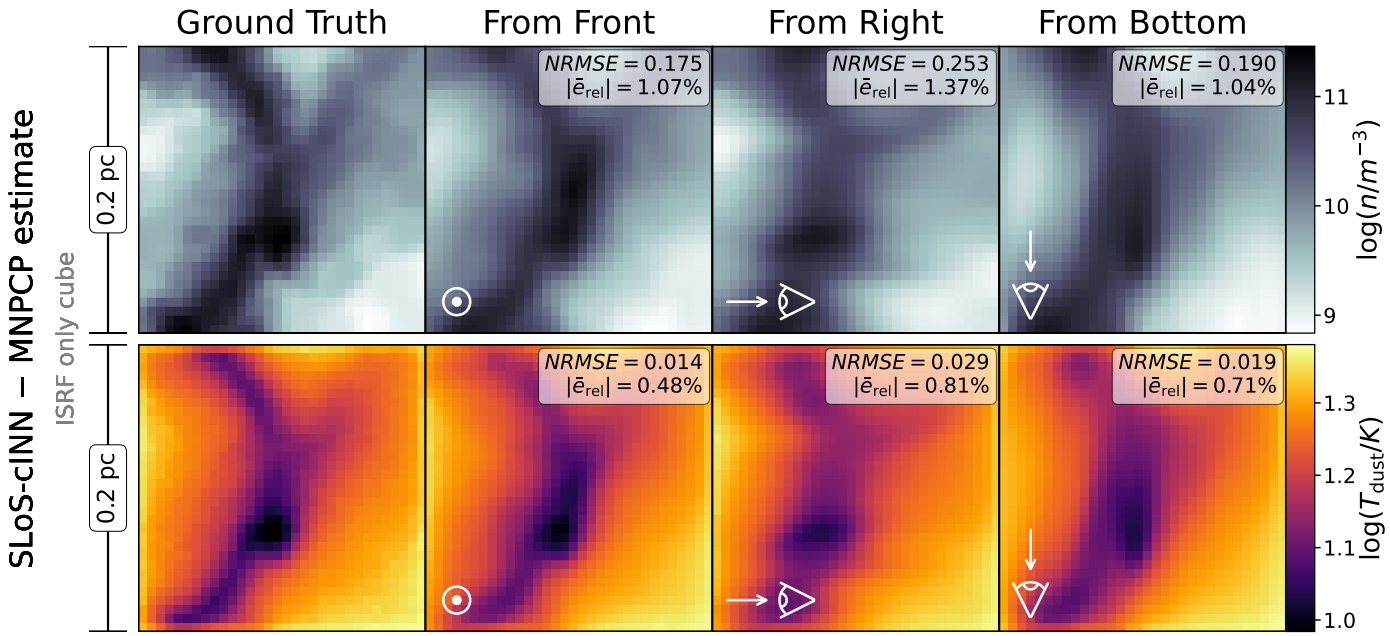

Fig. B.10: Comparison of MNPCP prediction results from different viewing angles to the ground truth for a single slice of an example cube subject only to the ISRF. Shown is the same example cube as in Figure 3 and all predictions are determined with the 23-wavelength SLoS-cINN. The presented slice is taken along the main viewing angle analysed in this work (Section 4). The results in the second column correspond to the nominal prediction outcome derived from the main viewing angle (i.e. the cube is observed from the front). The arrow and eye symbol in the third and fourth columns indicate the direction from which the cube is observed instead to generate the corresponding prediction result. The NRMSEs and median absolute relative errors listed in the three right-most panels are determined over the depicted slice only.

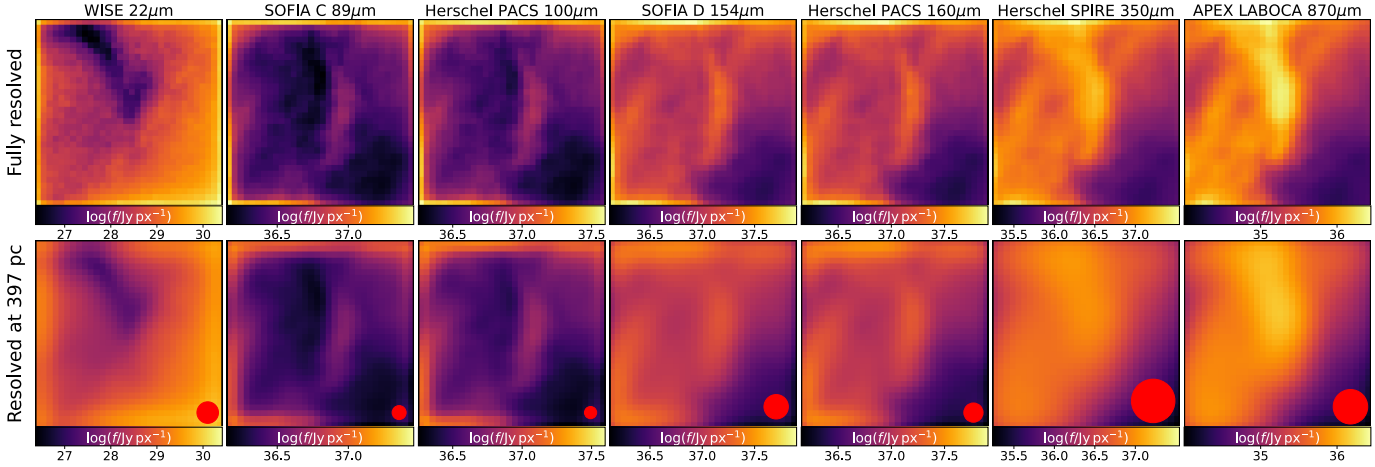

Fig. B.11: Comparison of dust emission maps between the perfect resolution scenario and a case that accounts for the PSFs of the respective instruments for the limited filter configuration used in Section 4.3. The red circle in the bottom row panels indicates the FWHM of the respective PSFs. We note that this example is based on the assumption that all considered telescopes share the same pixel size, matching our simulation resolution at a query distance of d = 397 pc.

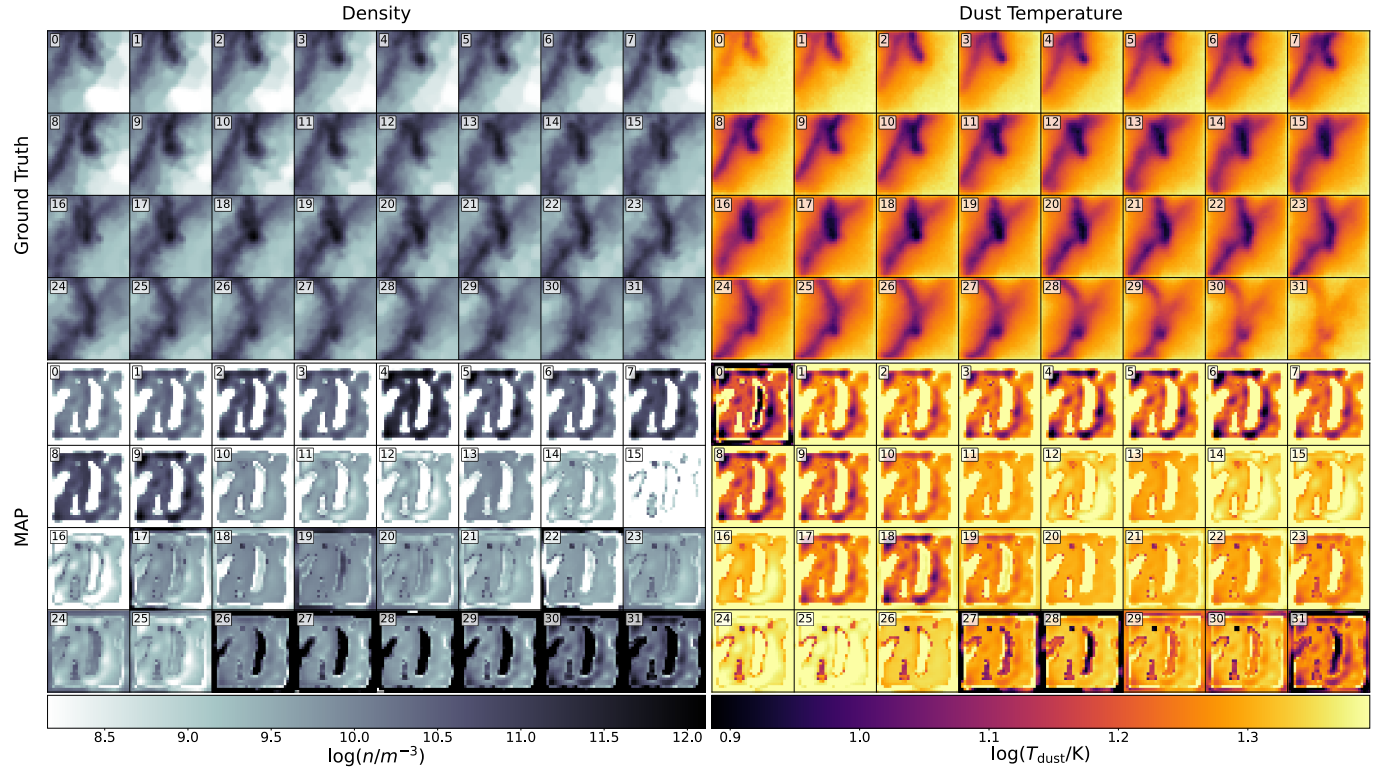

Fig. B.12: Comparison of the seven wavelength SLoS-cINN MAP prediction result to the ground truth for one example cube that is only subject to the ISRF. Same cube as in Figure B.8 is shown, with the difference being that here the input dust emission maps were first convolved with the PSFs of the respective instruments, corresponding to the input shown in the bottom row in Figure B.11.

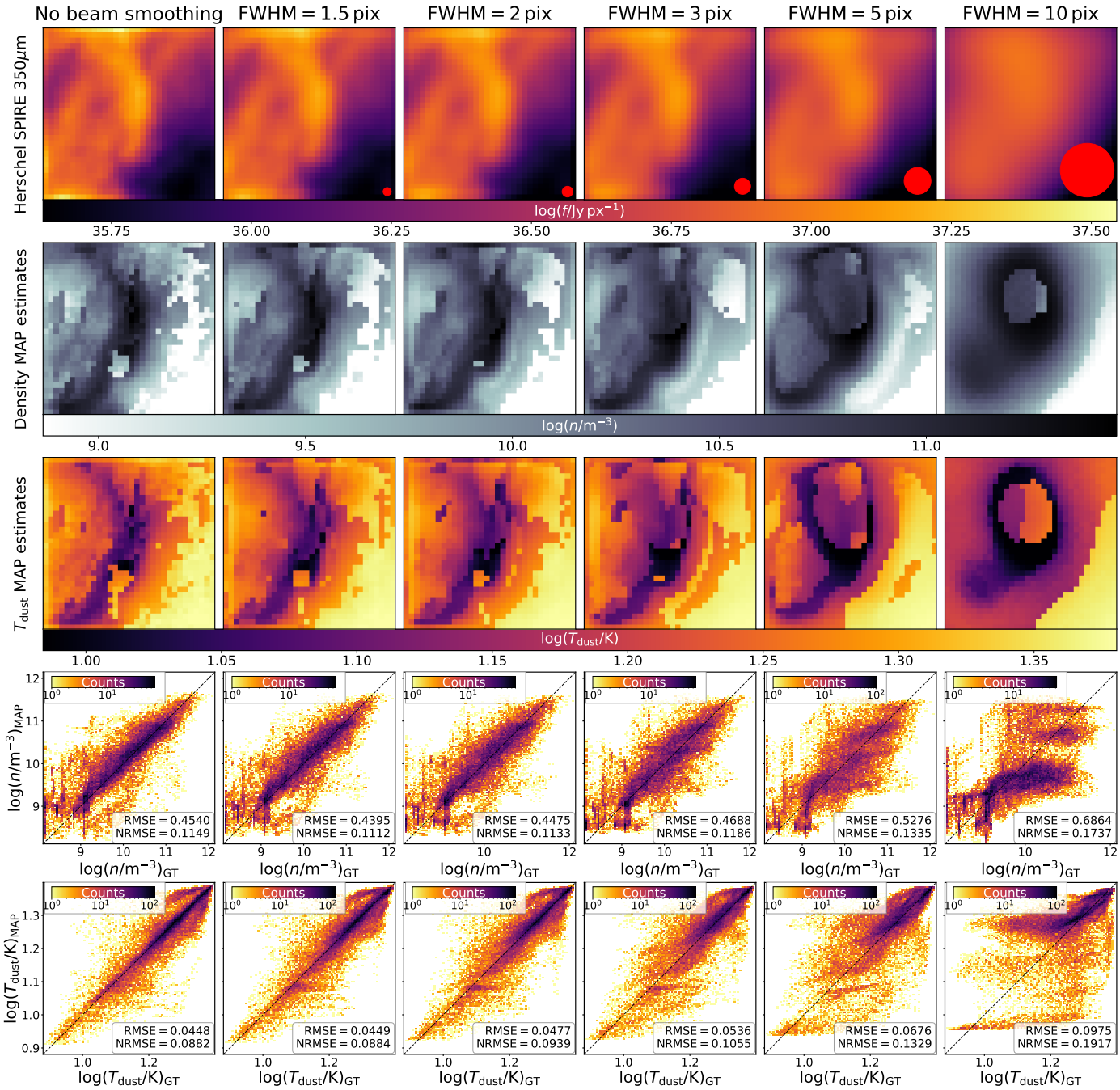

Fig. B.13: Comparison of prediction results with the seven wavelength SLoS-cINN for different amounts of smoothing applied to the input observations at all wavelengths equally. The top row shows the emission maps at the central wavelength of the Herschel SPIRE 350 μm filter as a an example to illustrate the effect of the convolution with a Gaussian beam. The red circle in each of these panels indicates the FWHM of the respective Gaussian kernel. The second and third row show the corresponding MAP estimates for dust density and temperature, respectively, for one example slice of the same cube analysed throughout the paper (i.e. the same slice as in Figures 3, B.3, B.10). The fourth and the fifth rows present a 2D histogram that directly compares the ground truth to the predicted density and temperature MAP estimates, respectively, for all pixels of this test cube, summarising the predictive performance.