Denoising Diffusion Probabilistic Models to Predict the Density of Molecular Clouds

(article starts on next page)

Denoising Diffusion Probabilistic Models to Predict the Density of Molecular Clouds

Duo Xu[1] , Jonathan C. Tan[1,2] , Chia-Jung Hsu[2], and Ye Zhu[3,4]
[1] Department of Astronomy, University of Virginia, Charlottesville, VA 22904-4235, USA; xuduo117@virginia.edu
[2] Department of Space, Earth & Environment, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden
[3] Department of Computer Science, Princeton University, Princeton, NJ 08544, USA
[4] Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616, USA
Received 2023 January 29; revised 2023 March 21; accepted 2023 April 4; published 2023 June 19

Abstract

We introduce the state-of-the-art deep-learning denoising diffusion probabilistic model as a method to infer the volume or number density of giant molecular clouds (GMCs) from projected mass surface density maps. We adopt magnetohydrodynamic simulations with different global magnetic field strengths and large-scale dynamics, i.e., noncolliding and colliding GMCs. We train a diffusion model on both mass surface density maps and their corresponding mass-weighted number density maps from different viewing angles for all the simulations. We compare the diffusion model performance with a more traditional empirical two-component and three-component power-law fitting method and with a more traditional neural network machine-learning approach. We conclude that the diffusion model achieves an order-of-magnitude improvement on the accuracy of predicting number density compared to that by other methods. We apply the diffusion method to some example astronomical column density maps of Taurus and the infrared dark clouds G28.37+0.07 and G35.39-0.33 to produce maps of their mean volume densities.

Unified Astronomy Thesaurus concepts: Interstellar medium (847); Astrostatistics (1882); Astrostatistics techniques (1886); Molecular clouds (1072); Magnetohydrodynamics (1964); Convolutional neural networks (1938)

1. Introduction

Giant molecular clouds (GMCs) are one of the most important components of the interstellar medium (ISM) in galaxies. The ISM is the material that fills the space between stars, consisting of gas and dust, as well as cosmic rays and magnetic fields (Spitzer 1978). The ISM plays a critical role in the life cycle of galaxies, regulating the rate of star formation (e.g., Leroy et al. 2008; Tacconi et al. 2020). GMCs are particularly important because they contain the majority of the dense gas in the ISM, which is required for the formation of stars (e.g., Shu et al. 1987; Saintonge & Catinella 2022). The physical conditions within GMCs are highly complex and dynamic, with variations in space and time of density, temperature, velocity, and magnetic field strength and direction. These conditions can lead to a wide range of phenomena, including the formation of protostars and star clusters (e.g., McKee & Ostriker 2007; Heyer & Dame 2015; Krumholz et al. 2019) and the formation of complex organic molecules (e.g., Herbst & van Dishoeck 2009; Jørgensen et al. 2020). Consequently, investigating the physical and chemical conditions of GMCs is a crucial step toward understanding the complex physical processes that occur within the Milky Way, as well as the properties and evolution of galaxies throughout the universe.

Among all the physical quantities of GMCs, the density (ρ), i.e., mass per unit volume,[5] is one of the most fundamental properties that relates to various physical quantity estimations,

such as the freefall time, the magnetic field strength (Davis 1951; Chandrasekhar & Fermi 1953; Beck 2015), and chemical reaction rates (Tielens & Hagen 1982; Wakelam et al. 2010; Gong et al. 2017). The freefall time depends on density as $t_{\rm ff} \propto \rho^{-1/2}$. Considering measurement of magnetic fields, one method commonly used is the Davis−Chandrasekhar −Fermi (DCF) method (Davis 1951; Chandrasekhar & Fermi 1953; Beck 2015). This estimates the plane-of-sky (POS) component of the magnetic field using polarized thermal dust emission (Rao et al. 1998; Planck Collaboration et al. 2016). The DCF method is based on the assumption that the magnetic field in the ISM is in a state of equipartition with the turbulent kinetic energy of the gas. This means that the magnetic field strength is proportional to the square root of the gas density and the turbulent velocity dispersion of the gas. Thus, in the DCF method, a good estimation of the gas density is required to obtain an accurate estimation of the magnetic field strength. Similarly, the gas density is a crucial factor affecting the rates of astrochemical reactions (Wakelam et al. 2010; Gong et al. 2017). Thus, a precise estimation of the gas density within GMCs is crucial for accurate prediction of molecular abundances and a better understanding of the chemical evolution in GMCs.

However, it is difficult to quantify the number density of GMCs from observations. The traditional approach of estimating the number density of GMCs is based on observations of column density and certain assumptions on the geometry of the clouds, for example, a cylindrical geometry for filamentary structures or spherical geometry for dense cores (André et al. 2014). Bisbas et al. (2021, 2023) proposed an empirical power law to convert the observed column density to the mean number density of GMCs based on the MHD simulations from Wu et al. (2017), which works decently but with noticeable scatter. Another method to constrain the number density of GMCs is utilizing density "probes," such as cyanoacetylene

---

[5] We will also use the number density of H nuclei as a metric of density. Under the assumption of an abundance of one He nucleus for every 10 H nuclei in interstellar gas, we have a mass per H nucleus of $\mu_{\rm H} = 1.4\,m_{\rm H} = 2.34 \times 10^{-24}$ g. Thus, $n_{\rm H} = 1\ {\rm cm}^{-3}$ is equivalent to $\rho = 2.34 \times 10^{-24}\ {\rm g\ cm}^{-3}$.

(HC$_3$N; Avery et al. 1982; Schloerb et al. 1983; Li & Goldsmith 2012). The relative intensity of different transitions of HC$_3$N is sensitive to the number density of the cloud, which makes it possible to constrain the mean number density directly by observing multiple transitions of HC$_3$N. Li & Goldsmith (2012) successfully observed $J = 2 - 1$ and 10–9 transitions of HC$_3$N in the Taurus B213 filament and constrained the number density of H$_2$ molecules, $n_{\rm H2} \sim (1.8 \pm 0.7) \times 10^4 \, {\rm cm}^{-3}$. Note that $n_{\rm H} = 2n_{\rm H2}$ under the assumption that all H is in the form of H$_2$. Unfortunately, the line ratio of HC$_3$N can only probe the number density at a relatively narrow range, between $\sim 10^4$ and $10^6 \, {\rm cm}^{-3}$ (Li & Goldsmith 2012), which limits its ability to infer the number density of the full range of structures that exist in GMCs. Consequently, a novel method to infer the number density of GMCs under a variety of physical conditions with high precision is in great demand. Machine learning makes it possible to learn from both the morphology of clouds and their column density to infer the mean number density rather than using a simple average power-law conversion.

Machine learning has gained great popularity among astronomers. For example, convolutional neural networks (CNNs) have been successfully applied to a series of tasks, including galaxy classification (Domínguez Sánchez et al. 2018; Becker et al. 2021); identification of structures like protostellar outflows, stellar-wind-driven bubbles, and Galactic cirrus filaments (Xu et al. 2020a, 2020b, 2022a; Smirnov et al. 2023); and inferring physical quantities based on observations, such as protostellar outflow inclination angles, magnetic field directions, stellar masses, exoplanet masses, galactic redshifts, and galactic star formation rates (Liu et al. 2020; Xu et al. 2022b; Zhang et al. 2022; Bisigello et al. 2023; Xu et al. 2023). Furthermore, CNNs have also been utilized to mitigate the impact of noise in astronomical observations (Gheller & Vazza 2022; Bartlett et al. 2023). For instance, Gheller & Vazza (2022) employed CNNs to remove noise and artifacts of radio interferometric images, while Bartlett et al. (2023) employed CNNs to diminish the impact of noise on various observation targets and preserve the morphology of galaxies. Meanwhile, generative adversarial networks (GANs) have been applied to a variety of tasks (Hemmati et al. 2022; Sweere et al. 2022). Sweere et al. (2022) applied GANs to generate super-resolution and denoised images from the XMM-Newton telescope. Hemmati et al. (2022) utilized GANs to effectively deblend galaxies from Hubble Space Telescope observations. More recently, denoising diffusion probabilistic models (DDPMs) have demonstrated their proficiency and robustness in image generation (Sohl-Dickstein et al. 2015; Ho et al. 2020). They have been particularly successful in creating realistic mock images that imitate observations of galaxies (Smith et al. 2022), making them well-suited for the prediction tasks in astronomy.

In this paper, we introduce the deep-learning method DDPMs to infer the number density of GMCs from column/surface density maps. We describe the diffusion model and how we generate the training set from MHD simulations in Section 2. Here we also introduce a CNN-based machine-learning approach, Convolutional Approach to Structure Identification-2D (CASI-2D), to infer the number density from column density. In Section 3, we evaluate our diffusion model in predicting the number density and compare with other approaches. We also apply our diffusion model to real

observations in Section 3. We summarize our results and conclusions in Section 4.

2. Data and Method

2.1. Magnetohydrodynamics Simulations

We carry out ideal MHD simulations based on the setup of Wu et al. (2020) and Hsu et al. (2023), which were conducted using the MUSCL-Dedner method and HLLD Riemann solver in the adaptive mesh refinement (AMR) code Enzo (Dedner et al. 2002; Wang & Abel 2009; Bryan et al. 2014). The simulations include self-gravity, magnetic fields, and heating/cooling based on a photodissociation model, which assumes a far-UV radiation field of $G_0 = 4$ Habings with attenuation from the $n_{\rm H} - A_V$ relation introduced in Wu et al. (2015) and cosmic-ray ionization rate $\zeta = 10^{-16} \, {\rm s}^{-1}$. Initially, two clouds with a radius of 20 pc are initialized in a 128 pc^3 domain resolved by 256^3 cells. The clouds have initial density $n_{\rm H} = 83 \, {\rm cm}^{-3}$, temperature $T = 15$ K, and solenoidal turbulent velocity field with $v_k^2 \propto k^{-4}$, $2 \leqslant k \leqslant 20$. The gas outside the clouds has a 10 times lower density $n_{\rm H} = 8.3 \, {\rm cm}^{-3}$ and 10 times higher temperature 150 K to balance the pressure. Note that while the GMCs have an initial temperature of 15 K, soon a multiphase temperature structure is established, e.g., with typical temperatures of $\sim$10–20 K at high densities ($n_{\rm H} \gtrsim 10^3 \, {\rm cm}^{-3}$), $\sim$40 K at intermediate densities ($n_{\rm H} \sim 10^2 \, {\rm cm}^{-3}$), and $\sim$1000 K at low densities ($n_{\rm H} \lesssim 10 \, {\rm cm}^{-3}$; see Figure 2 of Hsu et al. 2023).

The initial magnetic field is oriented at an angle of $60°$ with respect to the collisional axis, and its strength varies from 10 to 30 to $50 \, \mu$G in different considered cases. Four additional levels of refinement are allowed to resolve the local Jeans length with eight cells. For each magnetic field, we select two different types of setup of the large-scale dynamics of the GMCs, noncolliding and colliding GMCs. In the colliding cases, the clouds have a relative velocity of $10 \, {\rm km \, s}^{-1}$ and are offset by $0.5R_{\rm GMC}$. The simulations do not include star formation or feedback, so they represent the structures that develop in the early phases of collapse up to the onset of star formation. The simulations are run for 5 Myr. We take two evolutionary stages, 3 and 4 Myr, for analysis from each run.

To enhance the diversity of the data set, we generate column density maps and their corresponding line-of-sight (LOS) mass-weighted number density across different scales by adopting different AMR levels with different physical resolutions. The image size in pixels is 128×128, with multiple physical scales, including 32, 16, 8, and 4 pc. In total, we have 7179 images in the data set, in which 70% are used for the training set and the remaining 30% are a test set. Figure 1 shows the correlation between the LOS mass-weighted number density and the column density for different simulations. Although the column density range is not the same for different simulations, it is obvious that the relation between the mass-weighted number density and column density is similar for all the different simulations.

2.2. Machine-learning Approaches

2.2.1. Denoising Diffusion Probabilistic Models

DDPMs (Sohl-Dickstein et al. 2015; Ho et al. 2020), hereafter called diffusion models for short, are the state-of-the-art generative method in deep-learning and computer vision research field in recent years. Specifically, diffusion models have been successfully
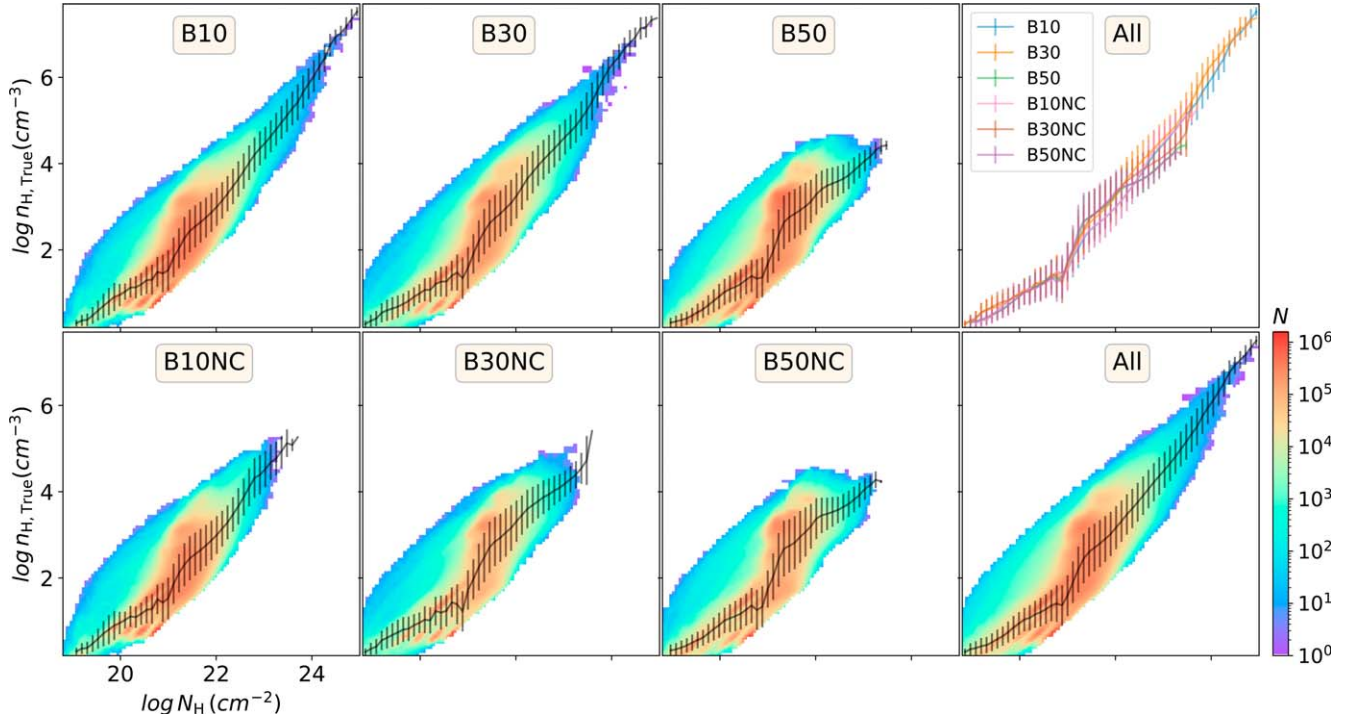
Figure 1. Relation between the LOS mass-weighted number density and the column density for different simulations. The rainbow color background indicates the 2D histogram of the distribution between number density and column density for each simulation. The lines with error bars represent the mean and their standard deviation of each column density bin.

applied in synthesizing various high-fidelity data, including images, video, and audio (Rombach et al. 2022; Singer et al. 2022; Zhu et al. 2023a; Zhu et al. 2023b).

The core formulation of diffusion models is inspired by nonequilibrium thermodynamics (Sohl-Dickstein et al. 2015), which models a stochastic Markov chain of T steps in two directions. The forward direction q, also known as the diffusion process, gradually adds stochastic Gaussian noises to a data sample x_0 following

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I), \quad (1)$$

where $\{\beta_t\}_t^T = 1$ are prescheduled variances. The other direction, often referred to as the reverse direction or generative process p, denoises a noisy sample x_T from a standard Gaussian distribution to a data sample x_0 as

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_t^\theta(x_t)\right) + \sigma_t z_t, \quad (2)$$

where ϵ_t^θ is the learnable noise predictor, $z_t \sim \mathcal{N}(0, I)$, and the variance of the reverse process σ_t^2 is set to be $\sigma^2 = \beta$. The actual neural network training process aims to learn the noise predictor to optimize the variational lower bound on negative log likelihood:

$$\mathcal{L} = \mathbb{E}_q\left[-\log p(x_T) - \sum_{t \geqslant 1}\log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})}\right]. \quad (3)$$

There are several reasons why we adopt the diffusion model as the tool to infer the number density of GMCs. First, the diffusion models have demonstrated great potential and impressive performance in learning the data distribution in

the generative field as the current mainstream method. Moreover, unlike the previous deep-learning methods such as GANs (Goodfellow et al. 2014), diffusion models are well known for the benefits of stable training, robust performance, and better interpretability and traceability via the rigorous mathematical formulations as a Markov chain. Additionally, originating from the natural thermodynamics problem, the diffusion models simulate a random walk process in the data space, which shares the intrinsic alignment and consistency with most scientific problems.

To adapt the diffusion models in our context, we deploy a diffusion model with a similar parameter setup to that in Ho et al. (2020). Specifically, the diffusion model has in total $T = 1000$ steps and is optimized using the variational loss in Equation (3), which makes valid the assumption that the reverse direction converges to the Gaussian stochastic diffusion process. In particular, we provide x_c as an additional input condition to make the prediction follow each individual observational sample. We train the diffusion model for 400 epochs and evaluate the performance on a sample in the test set. Figure 2 shows an example of the reverse process on our test data, where Gaussian noise is gradually converted to our target after 1000 time steps.

2.2.2. CASI-2D

In this subsection, we introduce the CASI-2D model to predict the number density of GMCs from the column density. We adopt the same CNN architecture, CASI-2D, from Van Oort et al. (2019). CASI-2D is an autoencoder with both residual networks (He et al. 2016) and a "U-net" (Ronneberger et al. 2015). CASI-2D has two major components, the encoder part and the decoder part. The encoder part extracts the features
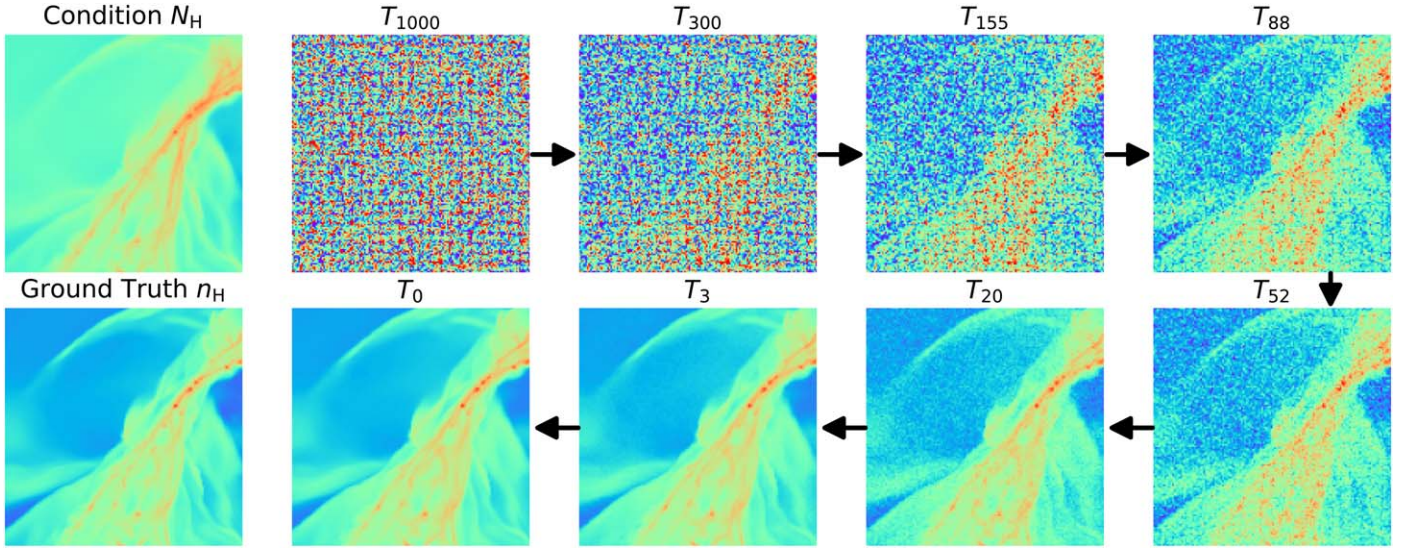
Figure 2. Demonstration of the diffusion process (reverse) on a sample in the test set.

from the input data and maps them into a lower-dimensional space, called the latent space, and then the decoder part takes this compressed representation from the latent space and attempts to reconstruct the target data. During the training process, the autoencoder is fed both input data, i.e., the column density map, and the target data, i.e., the number density map. It then learns to map the column density data to the number density data in a way that minimizes the difference between the reconstructed output and the target. We adopt the same training hyperparameters as Xu et al. (2020a, 2020b).

3. Results

3.1. Comparison between Different Approaches

In this subsection, we adopt several different approaches to convert column density to mass-weighted number density on the LOS. Note that when converting from n_H and N_H to a total mass per unit volume, ρ, or a mass per unit area (mass surface density, Σ), as mentioned above, one must also account for He (assuming $n_{He} = 0.1 n_H$, so that the mass per H is $\mu_H = 1.4 m_H = 2.34 \times 10^{-24}$ g. Note also that observational studies of molecular clouds sometimes report results for n_{H2}, since H_2 is the main collision partner. We will convert these estimates to n_H assuming that all H is in the form of H_2 so that $n_H = 2 n_{H2}$.

We start with power-law fitting on the relation between the LOS mass-weighted number density of H nuclei (n_H) and the column density of H nuclei (N_H) for all simulations in Figure 1. We adopt two-component and three-component power laws to fit the $n_H - N_H$ relation, i.e., $n_H = f(N_H)$. Meanwhile, we conduct the "inverted" fitting, i.e., adopting two-component and three-component power laws to fit the $N_H - n_H$ relation $N_H = f(n_H)$, and then derive the inverse function $n_H = f^{-1}(N_H)$. We summarize the fitting results in Table 1. We then follow the power-law fitting results to convert the column density to mass-weighted number density. We show the fitting results in Figure 3. In addition, we present the result from machine-learning approaches, including CASI-2D and the diffusion model, in Figure 3.

Considering the results, it is obvious that there is significant dispersion between the true mass-weighted number density, $n_{H,True}$, and the predicted number density, $n_{H,Pred}$, that is

Table 1
Power-law Fitting Results[a]

Label	Components	Break Points	Power Indices
2PL	2	20.82	0.61/1.47
3PL	3	20.73/22.69	0.61/1.41/1.49
2PL-i	2	20.19	0.67/1.42
3PL-i	3	21.35/21.79	0.79/4.06/1.26

Note.
[a] Label, number of power-law components, break points of power-law fittings in log scale, and power indices of each power-law component.

converted by the power-law relation and that is predicted by CASI-2D. The predicted number density by the diffusion model has a much smaller dispersion around the true density. We present the dispersion values between $n_{H,True}$ and $n_{H,Pred}$ in Figure 3 for the various cases. The predicted number density by the diffusion model is an order of magnitude better than that by all the other approaches. The 2D distribution between the actual number density and the predicted number density is shown in the second row of Figure 3. The power-law conversion approach leads to a nonuniform distribution across densities, with a systematic underestimation at low density and a systematic overestimation at moderate density.

We proceed to measure the error distribution across the range of densities in the third row of Figure 3. It is observed that, except for the diffusion model, all other approaches have a nonuniform distribution across densities. In the last row of Figure 3, we provide an overview of the error distribution for all methods, i.e., listing the average offset (μ_b), internal standard deviation (σ_i), and total standard deviation ($\sigma_t = \sqrt{\mu_b^2 + \sigma_i^2}$). It can be seen that the diffusion model has the lowest offset bias and the smallest variation compared to the other methods. For the diffusion model the mean offset is at a level of 10% and the total standard deviation, which indicates the total deviation from the true value, is at the level of 38%.

To better visualize the performance between different approaches, we show a sample image from the test set and apply different approaches to obtain the predicted number density in Figure 4. Although the column density map is similar to the true
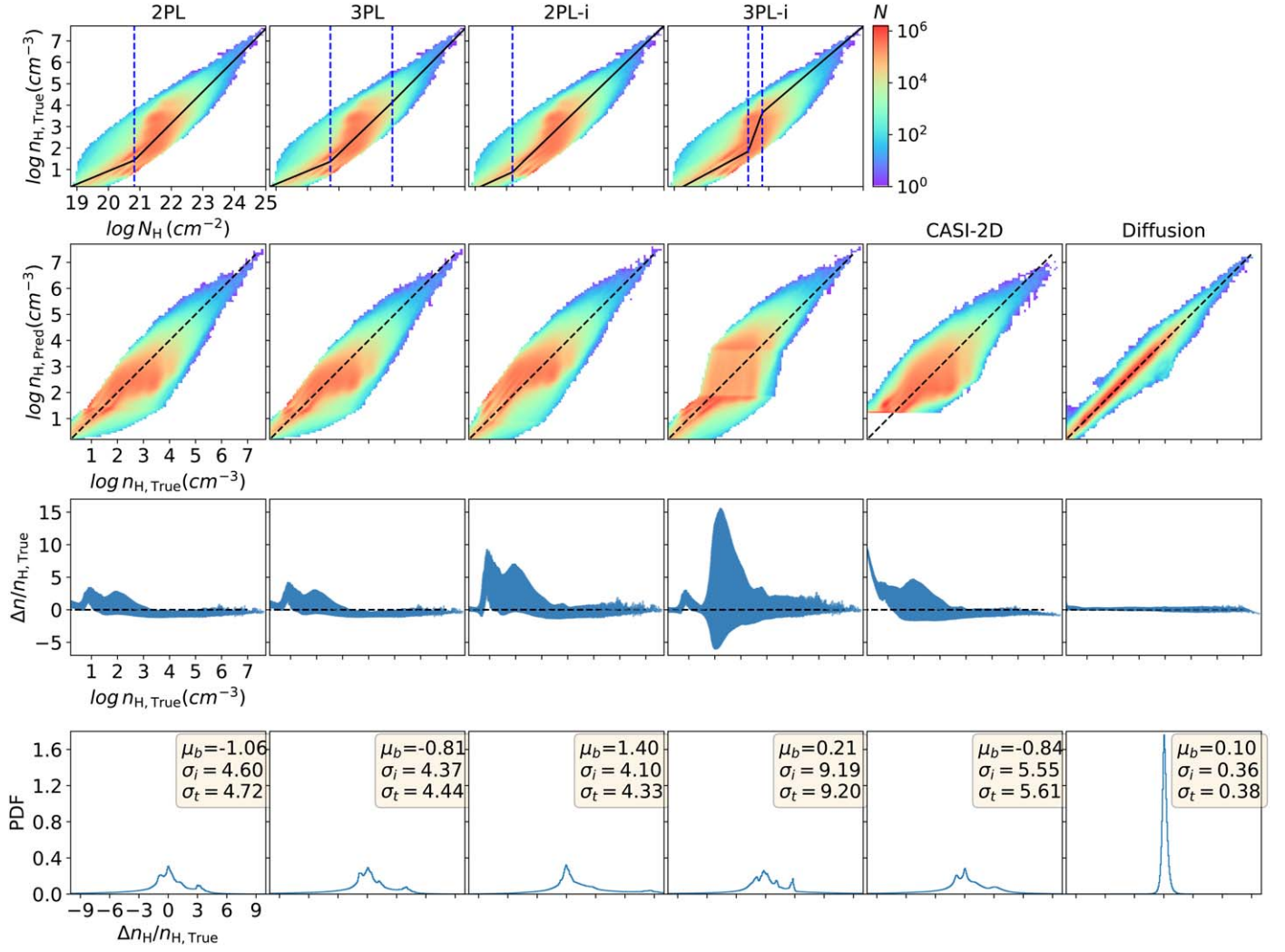
Figure 3. Summary of the performance of different approaches to convert the gas column density to number density on all the data samples.
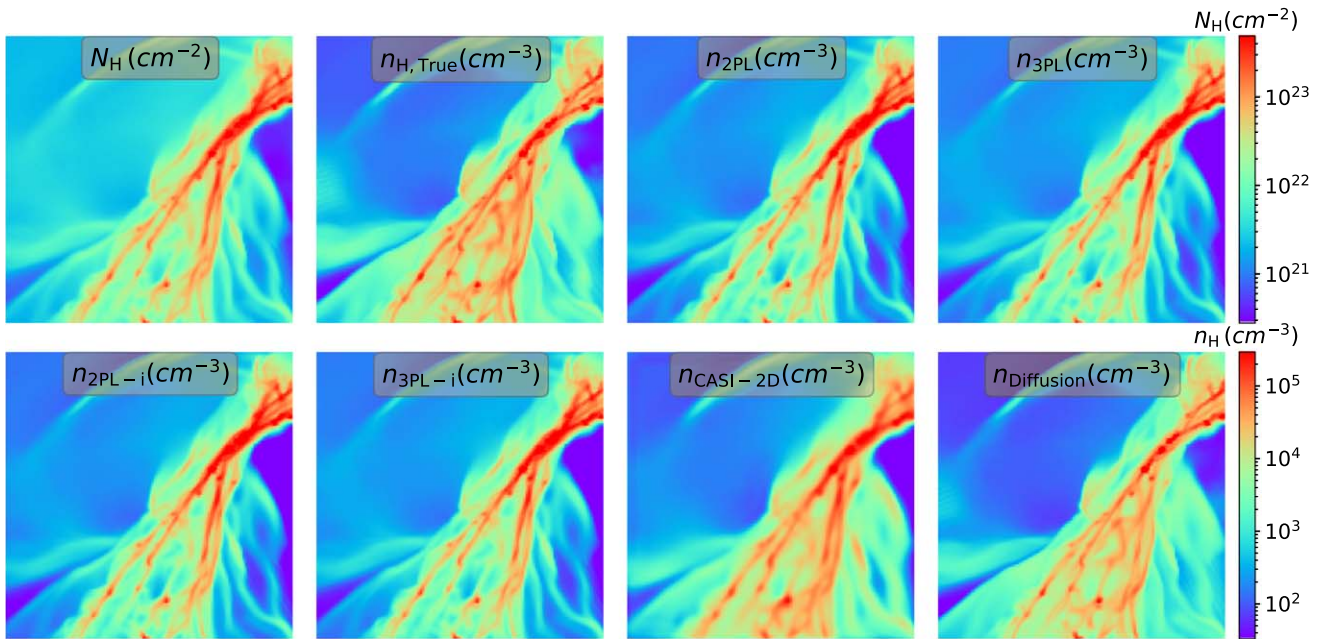

Figure 4. Comparison between different approaches to converting gas column density to number density on a sample in the test set.
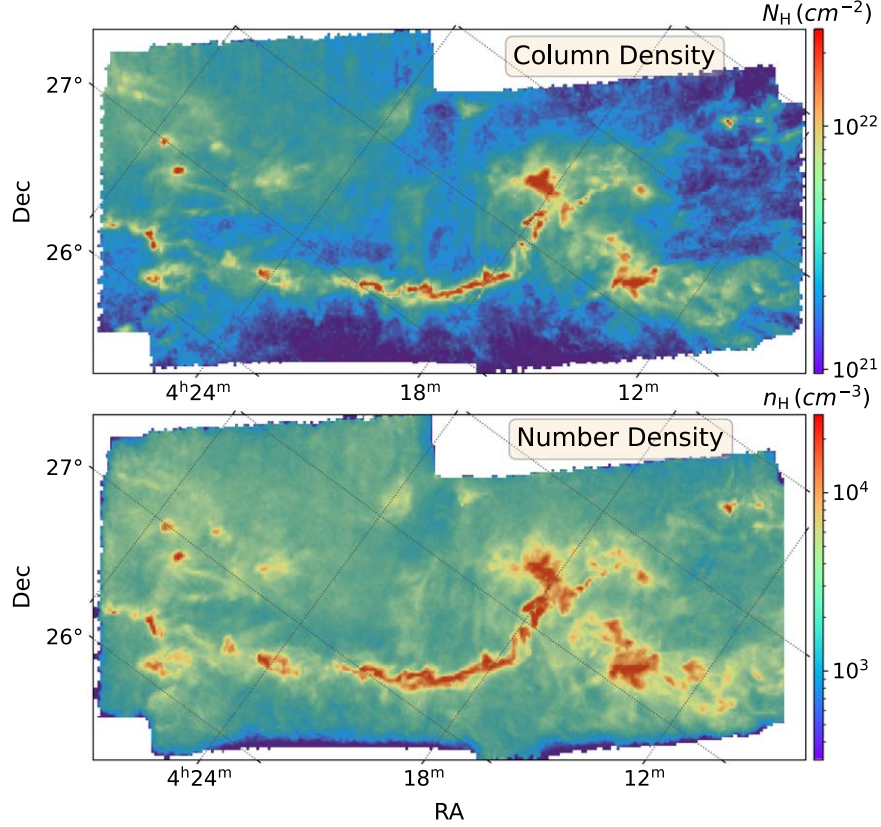
Figure 5. Herschel column density map of Taurus B213 (top) and the diffusion-model-predicted corresponding gas number density (bottom).

number density map, there is noticeable difference at relatively high column densities, where more structures are highlighted in the true number density map but appear faint in the column density map. The power-law conversion approach cannot reconstruct the structures that have relatively high number densities. CASI-2D is able to regenerate some of the high-density structures, but due to the intrinsic convolution manipulation, the number density map predicted by CASI-2D is smoother compared to the true number density map. The number density map predicted by the diffusion model is able to re-create the actual structures across a wide density range.

The performance of the proposed diffusion model exhibits an improvement of one order of magnitude compared to CASI-2D, which is a CNN similar to variational autoencoders (VAEs; Kingma & Welling 2014). In the context of machine learning, we discuss the reasons for the significant improvement achieved by our proposed diffusion model as follows. The DDPMs are formulated based on the Markov stochastic process and model a random walk in the data space, which aligns with most existing physics problems and is consistent with intrinsic properties of the natural world. In contrast, CNNs and VAEs were originally designed for image classification and generation tasks in computer vision and lack explicit connections to the physical world. For instance, the observed structure of GMCs is likely to be shaped, at least in part, by turbulent motions that involve compressions in a series of quasi-random directions. Then, the overall mass surface density map is constructed by summing a series of quasi-independent patches of volume density along the LOS. Thus, inferring the raw mass-weighted number density distribution based on the observed column density inherits the basic concepts of diffusion models. Furthermore, with respect to information

loss, DDPMs maintain the same data dimensionality throughout the entire denoising (i.e., prediction) process, thus better preserving the information conveyed by the original data. In contrast, CNNs and VAEs involve dimension reduction and information compression during training, resulting in inevitable information loss for the prediction objective. In terms of traceability and interpretability, we employ predefined Gaussian transition kernels to introduce and remove noise at each diffusion step in DDPMs. This provides us with superior traceability and interpretability for the data transition compared to CNNs and VAEs, whose traceability relies on a relatively vague gradient descent optimization direction.

3.2. Test on Herschel Observations of Taurus B213

In this subsection, we apply our diffusion model to a Herschel-derived column density map of Taurus B213. The Herschel column density map of Taurus B213 is obtained from Palmeirim et al. (2013), which has a resolution of $18\overset{''}{.}2$. The column density map is derived based on an optically thin graybody assumption with the function $I_\nu = B_\nu(T_d)\kappa_\nu\Sigma$, where I_ν is the observed surface brightness at frequency ν and κ_ν is the dust opacity per unit mass. Four data points of spectral energy distributions (SEDs) are included in the fitting, including 160, 250, 350 and 500 μm.

Taurus B213 is one of the closest star-forming filamentary structures located in the Taurus molecular cloud (Li & Goldsmith 2012; Palmeirim et al. 2013). Palmeirim et al. (2013) adopted a power-law radial density profile to fit the column density profile of the B213 filament and found the central density of the filament to be $n_{H.c} = 7.5 \times 10^4$ cm^{-3}. Li & Goldsmith (2012) analyzed multiple transitions of HC$_3$N in the B213
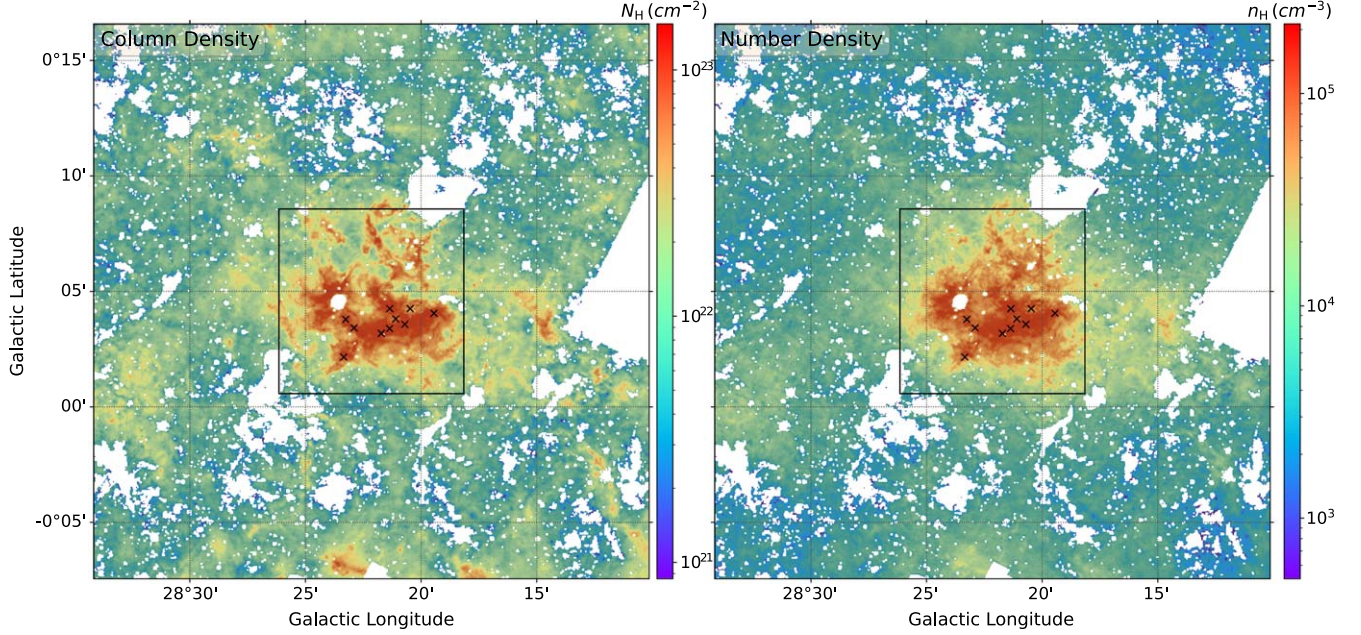
Figure 6. MIR-extinction-derived column density map of IRDC G28.37+0.07, i.e., Cloud C (left; from Kainulainen & Tan 2013), and the diffusion-model-predicted corresponding gas number density (right). Crosses indicate the locations that have number density estimates by Entekhabi et al. (2022).

filament and constrained the number density to be $n_H \sim 3.6 \times 10^4$ cm^{-3}. We apply our diffusion model to predict the LOS mass-weighted number density and show the predicted map in Figure 5. The peak number density of the B213 filament predicted by the diffusion model is $n_{H,peak} = 4.7 \times 10^4$ cm^{-3}. Note that the diffusion model predicts the LOS mass-weighted number density but not the actual central peak density of the filament. We adopt the power-law fitting results from Palmeirim et al. (2013) and calculate the LOS mass-weighted number density of the center of the B213 filament, which is $n_H = 4.0 \times 10^4$ cm^{-3}. Consequently, our diffusion model prediction is highly consistent with the estimation from other methods.

The above example examines the diffusion model's ability to accurately predict number density on real data that were not included in our training set. We further evaluate the diffusion model's performance on a new simulation in Appendix A, as well as its performance on data with varying physical resolutions in Appendix B. Our findings demonstrate that the diffusion model performs reasonably well in predicting number density on previously unseen data. As a comparison, we also apply CASI-2D to the Taurus data and report our results in Appendix C. Our findings suggest that the predictions of CASI-2D exhibit greater blurring, with a smoother density peak at a lower value. This observation implies that CASI-2D may not be as effective when dealing with unseen data.

3.3. Test on Extinction Maps of IRDCs

In this subsection, we apply our diffusion model to dust extinction maps of two infrared dark clouds (IRDCs), G28.37+00.07 (Cloud C) and G035.39-00.33 (Cloud H). IRDCs are cold, dense regions that are considered the birthplace of massive stars and star clusters. It is of great significance to study the physical and chemical conditions of IRDCs. Particularly, the number density of IRDCs is one of the most fundamental quantities.

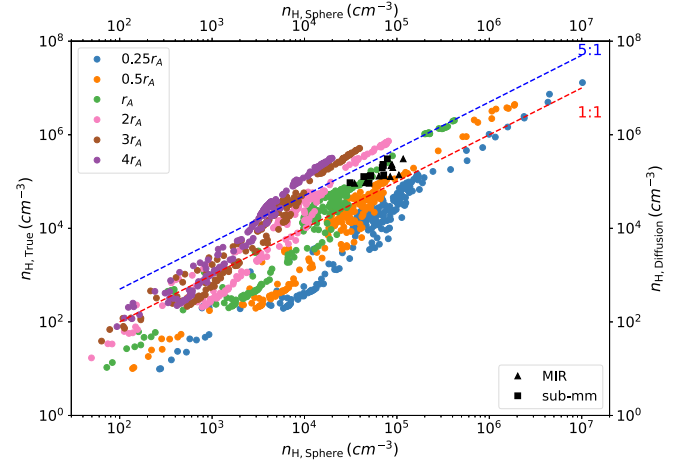We adopt the mid-IR (MIR) extinction (MIREX) map derived column density maps from Kainulainen & Tan (2013),

Figure 7. Comparison between the true mass-weighted LOS number density and that estimated by assuming a sphere with a certain radius on a simulation sample. Different colors indicate different radii of apertures, where r_A is 0.39 pc. Black symbols represent the comparison between the number density estimates by Entekhabi et al. (2022) and those by the diffusion model in IRDC G28.37+0.07, i.e., Cloud C. Squares and triangles represent different number density estimates in Entekhabi et al. (2022) based on two different data sets, MIR extinction from Spitzer and submillimeter emission from Herschel, respectively. The red line indicates the 1–1 line. The blue line indicates the 5–1 line.

which are derived from a 2″-resolution 8 μm Spitzer−IRAC extinction map (Butler & Tan 2012), along with a lower-resolution, larger-scale near-IR extinction correction. The relation of column density N_H to visual extinction A_V is $N_H = 1.9 \times 10^{21}$ cm$^{-2}(A_V/$mag$)$. We first apply the diffusion model to the column density map of IRDC G28.37+0.07, i.e., Cloud C, and show the predicted number density in Figure 6. The number density of Cloud C varies from 10^3 to 10^6 cm^{-3}. Entekhabi et al. (2022) applied spherical geometry with a radius of 0.39 pc to derive the number density from the column density map of Cloud C at 10 locations, as indicated in Figure 6. We compare the number density calculated by
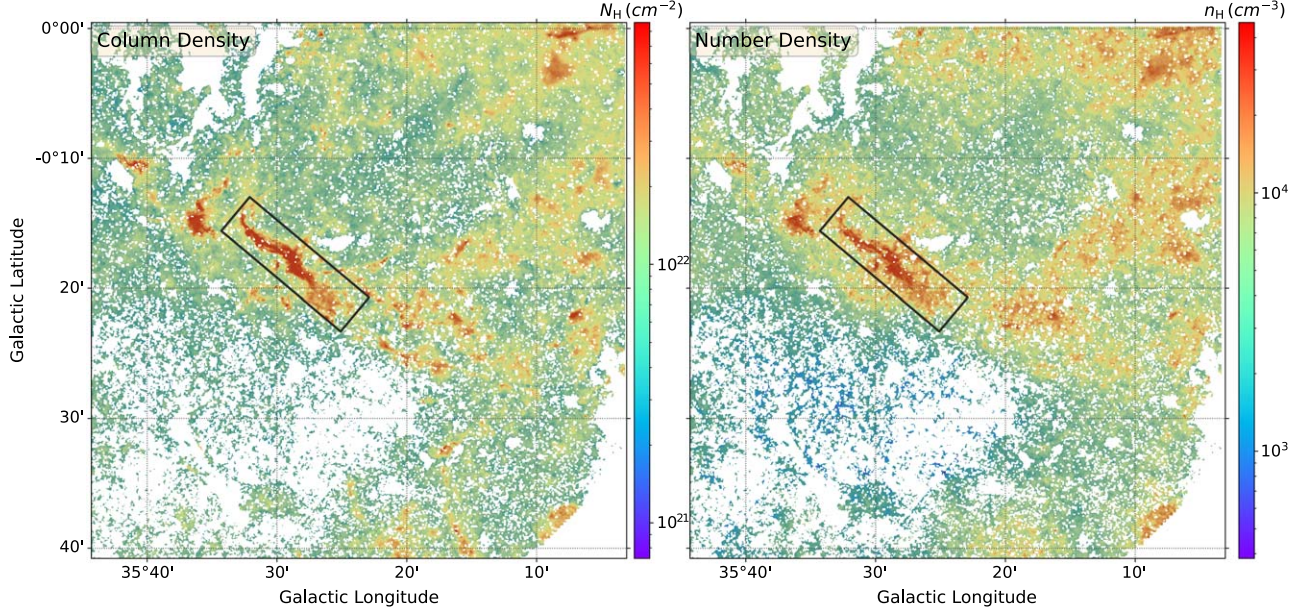
Figure 8. MIR-extinction-derived column density map of IRDC G35.39-0.33, i.e., Cloud H (left; from Kainulainen & Tan 2013), and the diffusion-model-predicted corresponding gas number density (right). The black box indicates the location of the filamentary structure studied in Jiménez-Serra et al. (2014).
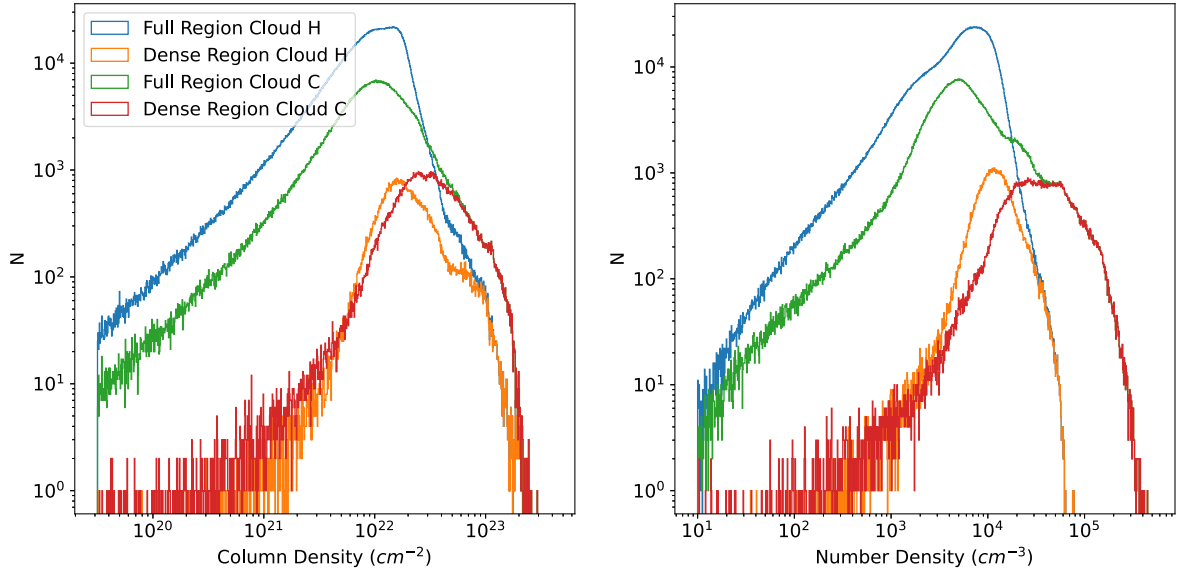

Figure 9. Distribution of gas column density and the mass-weighted number density predicted by the diffusion model for both Cloud C and Cloud H. The study of the entire Cloud C and Cloud H regions is represented by the blue and green lines, respectively. The red line corresponds to the study focused on the central dense region of Cloud C as illustrated by the black box in Figure 6. The orange line indicates the study conducted on the filamentary structure, located within the black box in Figure 8.

Entekhabi et al. (2022) to that by the diffusion model in Figure 7. It is likely that the previous estimation of the number density in Cloud C is systematically underestimated by a factor of 2, which may have a slight impact on the chemical modeling of Cloud C performed by Entekhabi et al. (2022). To evaluate the impact of the radius of the aperture in spherical geometry approximation, we apply different radii of apertures to calculate the mean number density in each circle and compare that with the true number density calculated from simulations. Figure 7 also shows the impact of aperture radius in density estimation. As the aperture increases, the mean density calculated with a circle-to-sphere approximation is likely underestimated in dense regions. When the aperture is small, the density

calculated with a circle-to-sphere approximation is similar to the true density, which indicates that the dense cores likely have a similar scale to this aperture radius. It is worth noting that this scale might only be an upper limit owing to the finite resolution of the simulation. At the low-density region, we can see a significant overestimation of density using a circle-to-sphere approximation with a small aperture. This is consistent with intuition, where the diffuse region has a much larger scale on the LOS than 0.39 pc. Consequently, we conclude that the density estimation with a circle-to-sphere approximation is sensitive to the choice of aperture radius.

Next, we apply the diffusion model to the column density map of IRDC G35.39-0.33, i.e., Cloud H, which is a highly filamentary

cloud. The predicted number density is shown in Figure 8. The number density of Cloud H varies from 10^3 to 10^5 cm^{-3}. Jiménez-Serra et al. (2014) adopted multitransition ^{13}CO and C^{18}O lines to estimate H$_2$ number density of the filamentary structure in Cloud H. Based on the large velocity gradient approximation of the non-LTE radiative transfer, Jiménez-Serra et al. (2014) constrained the number density to be $n_{\rm H,avg} = 8.3 \times 10^3$ to 1.2×10^4 cm^{-3}. The diffusion model predicts a mean number density of the filamentary structure of $n_{\rm H,avg,diffusion} = 1.3 \times 10^4$ cm^{-3}. The diffusion-model-predicted mean number density of the entire Cloud H is $n_{\rm H,avg,diffusion} = 6.1 \times 10^3$ cm^{-3}. The predicted number density by the diffusion model is thus consistent with that inferred by an independent non-LTE radiative transfer approach in Jiménez-Serra et al. (2014).

It is important to note that our objective is not to reproduce the measurements of individual points obtained by other methods. Rather, our aim is to establish the diffusion model as a useful tool for determining the number density of GMCs over a wide range of values. Such a tool can be useful for various analyses, including statistical analysis of turbulence. An example of comparing the column density distribution with the number density distribution in Cloud C and Cloud H is shown in Figure 9. The dissimilarity between the column density distribution and the number density distribution is evident in the high-density regions. This result may help astronomers in investigating the statistical properties of turbulence more thoroughly in the future. It is noteworthy that the quantity referred to as the number density does not reflect the true density; rather, it represents the mass-weighted number density along the LOS.

4. Conclusions

We have trained the deep-learning method diffusion probabilistic models to predict the LOS mass-weighted number density of GMCs from column density maps. We have tested the diffusion model performance on synthetic test samples and real observational data. Our main findings are as follows:

1. The diffusion model is able to predict the LOS mass-weighted number density with higher accuracy (an order of magnitude better) than that from traditional power-law fitting and that from CASI-2D.

2. We applied the diffusion model to predict the LOS mass-weighted number density of the Taurus B213 filament from the Herschel-derived column density map. The predicted result is consistent with the estimation by the density probe HC$_3$N and by the cylindrical geometry fitting.

3. We applied the diffusion model to predict the LOS mass-weighted number density of two IRDCs from MIR-extinction-map-derived column density maps. The predicted result is consistent with the estimation by non-LTE radiative transfer modeling. However, application of a local circle-to-sphere conversion appears to underestimate the density by a factor of about 2, potentially due to the small-scale dense substructure within the volume.

4. The diffusion model, along with all the other methods of predicting density, encounters difficulties when there are differences between the properties of observational data and those of the training data, such as variations in physical resolutions or dissimilarities in column density distributions. Thus, it is important for more precise predictions for the model to be trained with data that most

accurately represent the observed system that is being investigated.

Appendix A
Model Evaluation on Different Simulations

In this appendix, we apply the power-law fitting result, CASI-2D, and the diffusion model to a new simulation that is conducted with a different code and with different physical conditions. We follow the same simulation setup in Xu et al. (2023). We conduct ideal MHD simulations with ORION2 (Li et al. 2021) to model turbulent clouds with periodic boundary conditions and without self-gravity. The simulation box is $5 \times 5 \times 5$ pc^3. The magnetic field is initialized along the z-direction. The gas is assumed to be an isothermal ideal gas with an initial temperature of 10 K. The three-dimensional Mach number is 10.5, which places the simulated cloud on the line width–size relation, $\sigma_{\rm 1D} = 0.72 R_{\rm pc}^{0.5}$ km s^{-1} (McKee & Ostriker 2007). The calculations use a base grid of 256^3 without AMR. Simulations are performed with a virial parameter of $\alpha_{\rm vir} = 5\sigma_v^2 R/(GM) = 2$. We adopt two different mass-to-flux ratios $\mu_\Phi = M_{\rm gas}/M_\Phi = 2\pi G^{1/2} M_{\rm gas}/(BL^2)$, $\mu_\Phi = 1$ and 16, which yield Alfvén mach numbers of 0.87 and 14, respectively. We show the correlation between the LOS mass-weighted
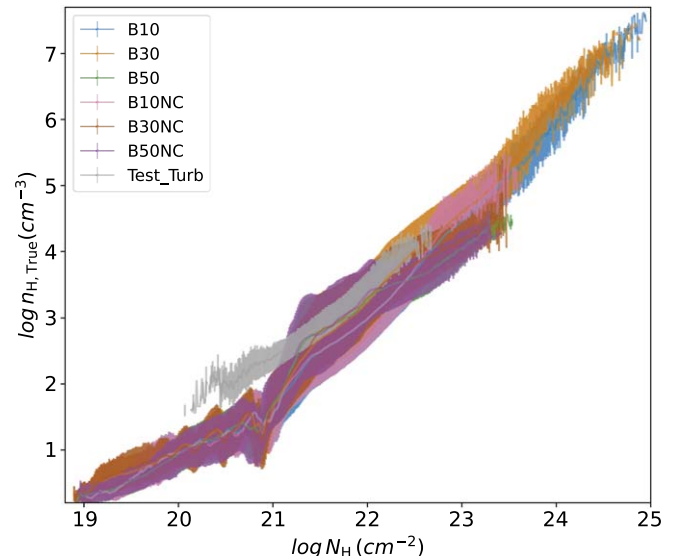

Figure 10. Relation between the LOS mass-weighted number density and the column density for different simulations. The lines with error bars represent the mean and their standard deviation of each column density bin.
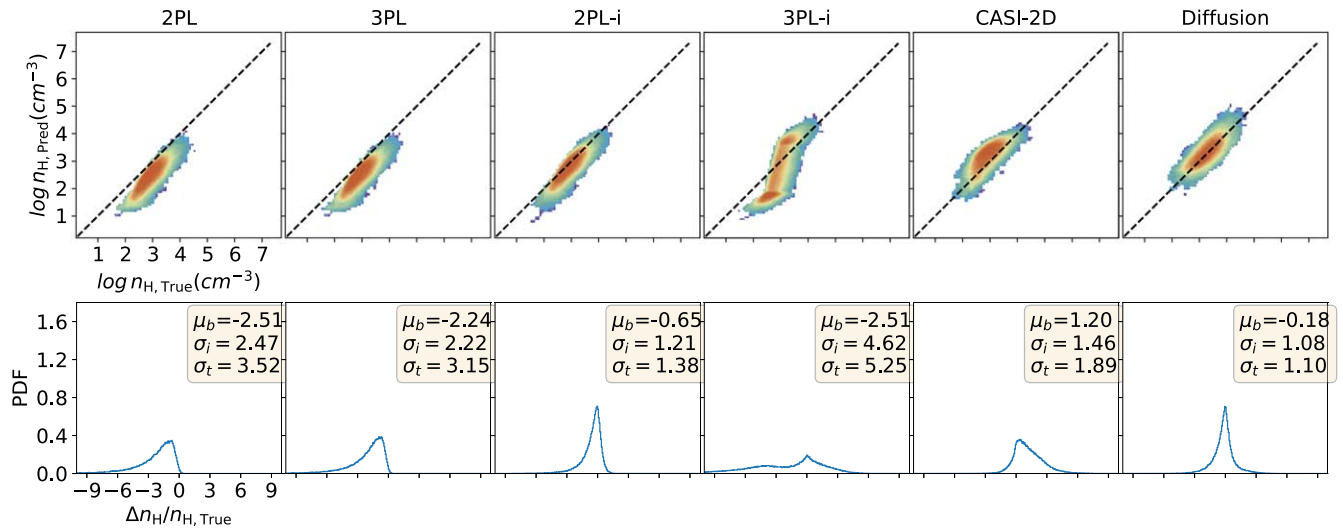
Figure 11. Summary of the performance of different approaches to convert the gas column density to number density on ORION2 turbulent simulations.

number density and the column density for these turbulent simulations in Figure 10. We also show the correlation between the LOS mass-weighted number density and the column density for all Enzo simulations, i.e., colliding and noncolliding multiphase GMCs with self-gravity, in Figure 10 for comparison. The correlation observed in the ORION2 pure turbulent simulations, which lack self-gravity, differs slightly from the correlation seen in the Enzo simulations that incorporate self-gravity and heating/cooling processes. The range of density variation in the ORION2 simulations is narrower compared to that in the Enzo simulations. However, the overall trend of both correlations is in close accordance.

We then apply the same power-law fitting results from Section 3.1 and two machine-learning approaches to predict the gas number density from their column density. We summarize our results in Figure 11. It is worth noting that all the approaches are derived or trained on the Enzo simulations. This ORION2 turbulent simulation is completely new to these approaches. One can clearly observe that all four power-law conversion approaches are systematically offset from the true value. Similarly, CASI-2D systematically overestimates the gas number density. On the other hand, the diffusion model is relatively robust in predicting the gas number density with only a small offset. In addition, the error distribution from the diffusion model predictions is more symmetric compared to the other approaches. This highlights the capability of the diffusion model in predicting gas number density on a previously unseen data set. However, this finding also highlights a potential limitation when applying the diffusion model to observational data sets. Specifically, if the properties of the observational data differ significantly from those of the training data, retraining the model may be necessary to ensure more accurate predictions.

Appendix B
Model Evaluation on Different Physical Scales

In this appendix, we evaluate the performance of the diffusion model on observational data at various scales. It is important to point out that the training data cover a particular range of physical resolutions, which spans from 0.315 to 0.25 pc pixel^{-1}. For nearby GMCs, such as Taurus, most column density maps obtained by

infrared telescopes, like Spitzer and Herschel, have a much higher physical resolution. As a result, to use the diffusion model on such data, we need to downsample it to a lower resolution that is within the range covered by our training set. The column density map of Taurus B213 obtained from Palmeirim et al. (2013) using Herschel has a high physical resolution of 0.002 pc pixel^{-1} when adopting a distance of 140 pc for Taurus, as discussed in Section 3.2. However, to use this map with our diffusion model, it needs to be downsampled by a factor of at least 15 to match the physical resolution range covered by our training data.

Here we assess the performance of the diffusion model on various physical scales using a subregion in Taurus, as illustrated in Figure 12. The diffusion model likely underestimates the number density by a factor of a few at higher resolutions that are beyond the scope of our training set. To obtain a more accurate assessment, we focus on the same locations observed by Li et al. (2012) and calculate the average number density in this region across different physical scales. Figure 13 presents the mean density prediction for the same region as Li et al. (2012) at different physical scales. To achieve the highest resolution within the training set, a minimum downsample factor of 15.4 is required. When the physical resolution is higher than the range of our training set, the diffusion model appears to underestimate the number density by a factor of a few, as compared to the measurements in Li et al. (2012). However, when we downsample the column density to a physical resolution of 0.04 and 0.05 pc pixel^{-1}, which corresponds to a downsampling factor of 20 and 24, respectively, the model predictions become relatively stable and consistent with the measurements. It is worth noting that we are unable to downsample the Taurus column density map by a larger factor owing to the restriction imposed by the diffusion model, which requires an image size of 128×128.

We proceed to evaluate the performance of the diffusion model on Cloud C and Cloud H at various resolutions. The distances to Cloud C and Cloud H are 5 and 2.9 kpc, respectively, indicating that their physical resolutions are at the upper limit of the resolutions covered by our training set. However, it should be noted that the presence of masked areas in the column density maps of both clouds results in a
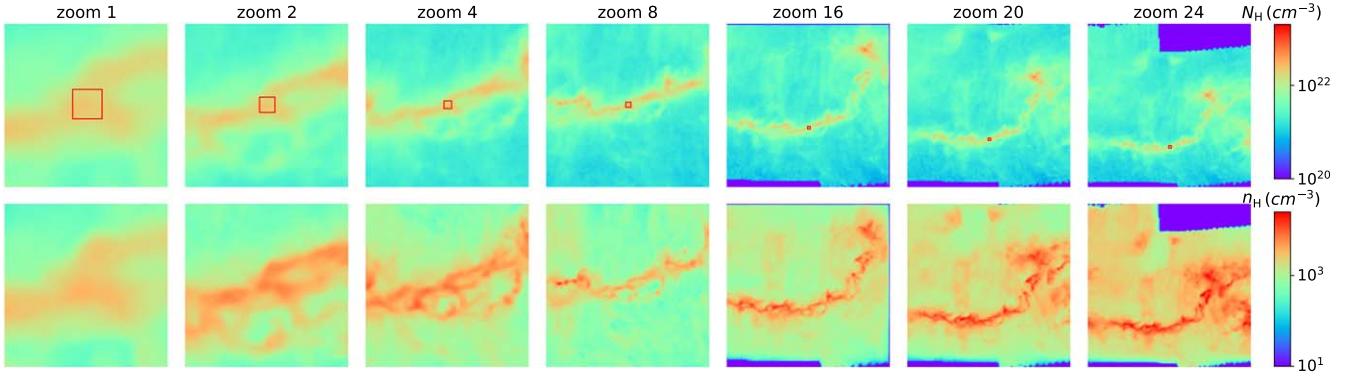
Figure 12. Diffusion model performance on different physical scales on a subregion in Taurus. The top row shows column density maps. The bottom row shows the number density maps predicted by the diffusion model. The rectangle indicates the location of the observation by Li et al. (2012).
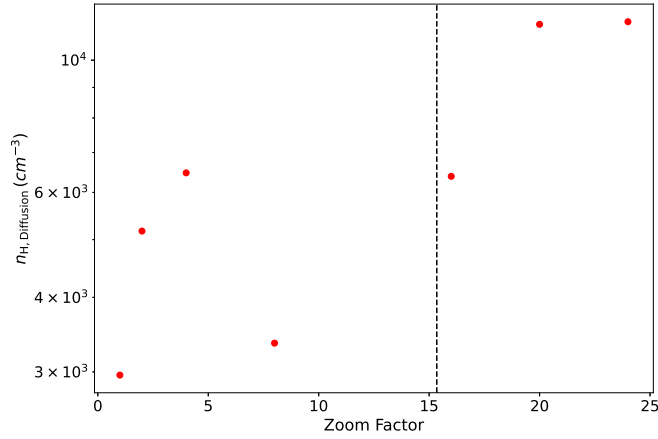

Figure 13. Predictions of the number density for the region examined by Li et al. (2012) at different physical scales. The minimum downsampling factor necessary to match the highest resolution in the training set is indicated by the vertical dashed line.

substantial decrease in the values of downsampled maps when padded with zeros. As a result, we only perform a limited evaluation at a few resolutions as a basic check on the method.

Figures 14 and 15 depict the performance of the diffusion model on various physical resolutions of Cloud C and Cloud H. It is evident that the predicted number density reduces by several factors as we downsample the data. To quantify this trend accurately, we take Cloud C as an example, and Figure 16 displays the diffusion model's prediction on various physical resolutions at locations examined by
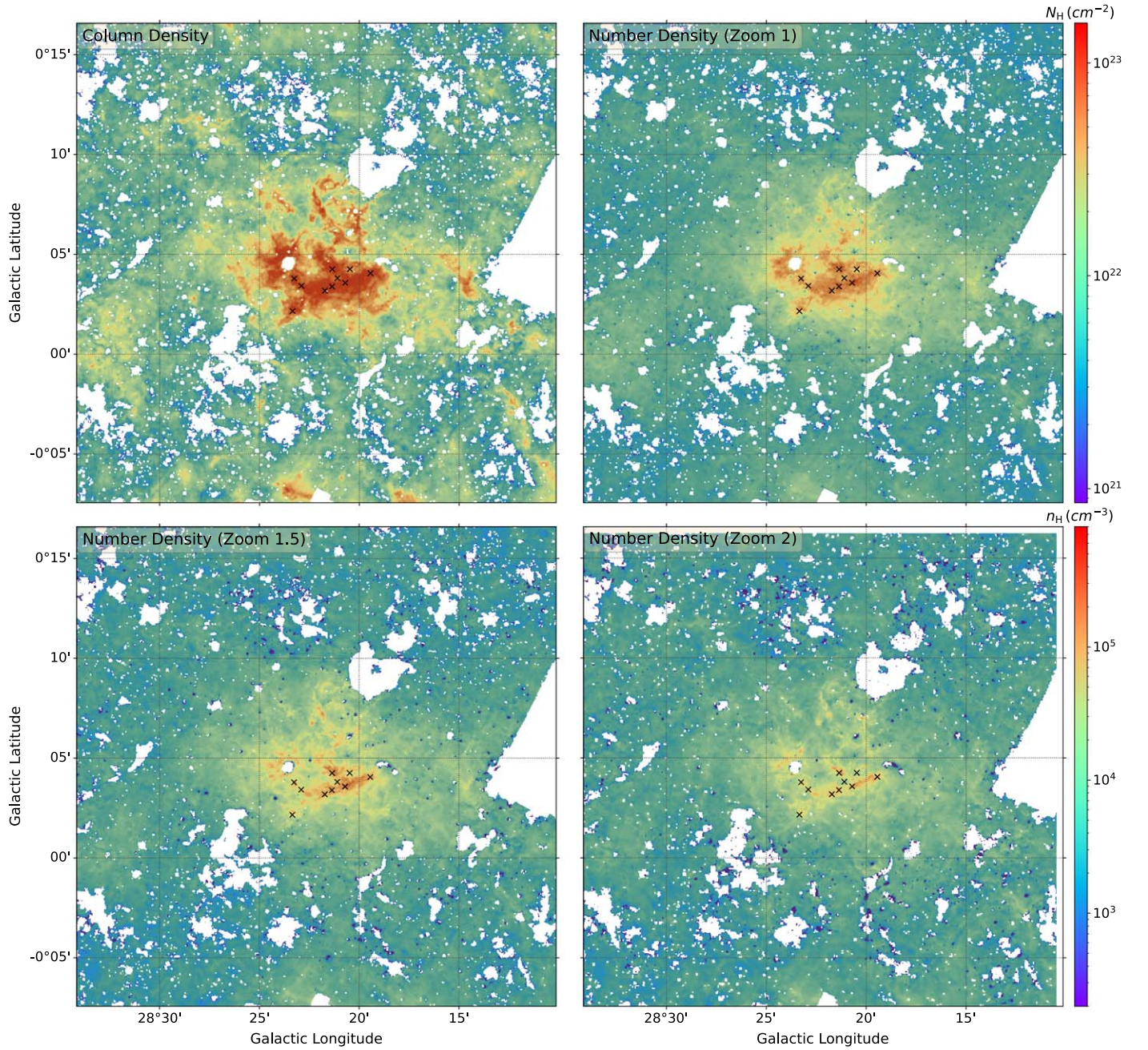
Figure 14. Diffusion model performance on different physical resolutions on Cloud C.
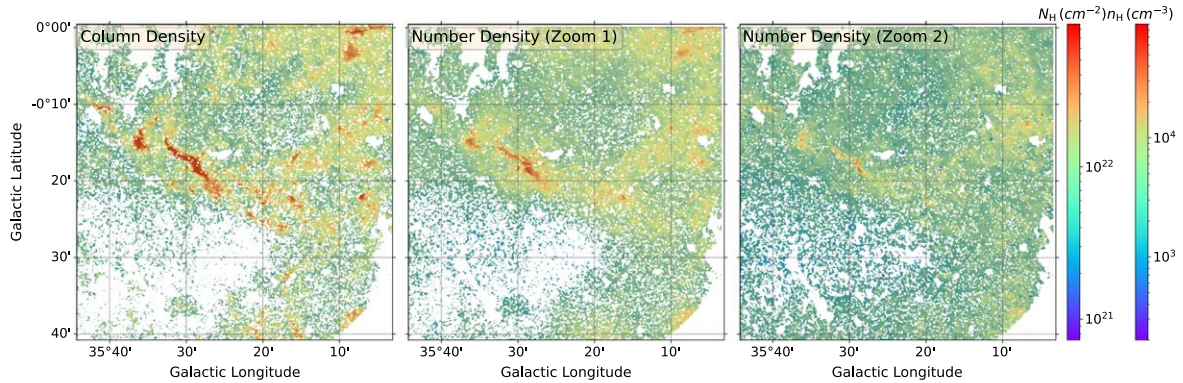

Figure 15. Diffusion model performance on different physical resolutions on Cloud H.
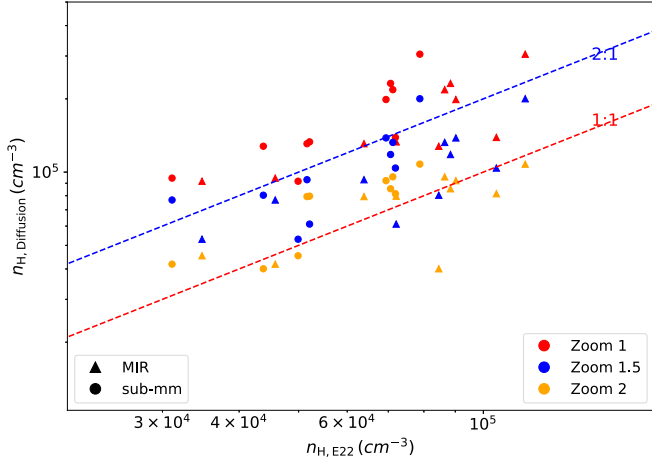
Figure 16. Comparison between the number density estimates by Entekhabi et al. (2022) and those by the diffusion model in Cloud C. Squares and triangles represent different number density estimates in Entekhabi et al. (2022) based on two different data sets, MIR extinction from Spitzer m and submillimeter emission from Herschel. The red line indicates the 1–1 line. The blue line indicates the 2–1 line. Different colors indicate predicted number density at different physical resolutions.

Entekhabi et al. (2022). The predicted number density decreases by a factor of two when we downsample the input column density by a factor of two.

In light of the differences in physical resolution that may exist between the observational data sets and the training data, it is important to exercise caution when applying the diffusion model to such data sets. In such cases, the diffusion model would need to be retrained with data that are representative of the relevant physical resolutions before being applied to actual observational data sets.

Appendix C
CASI-2D Performance on Observational Data

In this appendix, we provide an overview of the performance of CASI-2D on observational data. Figures 17 and 18 show the number density predictions of CASI-2D for GMCs. Upon comparison with the predictions of the diffusion model, it is evident that the CASI-2D predictions are more blurry, with a smoother density peak at a lower value.
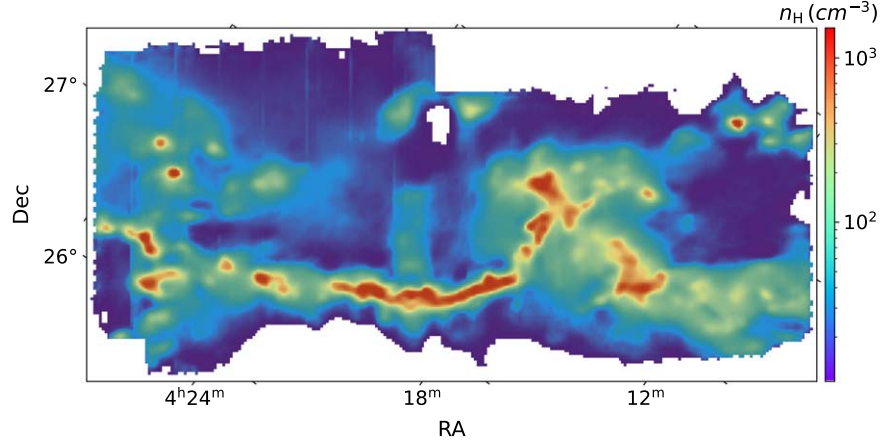

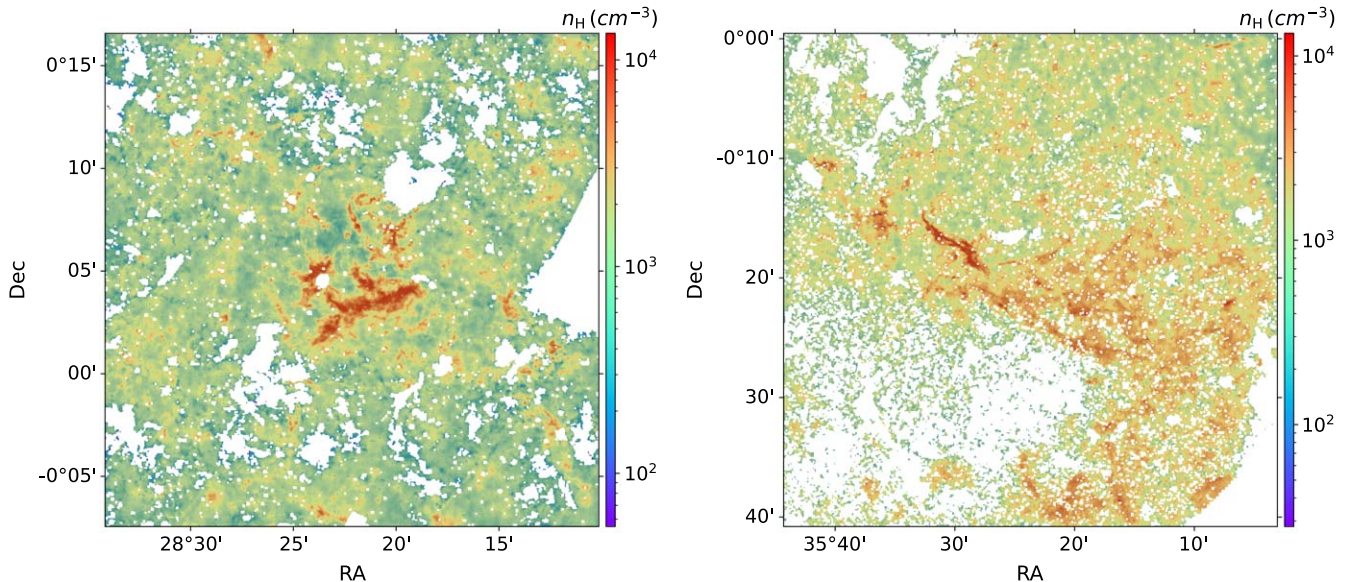**Figure 17.** Number density predicted by CASI-2D on Taurus B213.

Figure 18. Number density predicted by CASI-2D on Cloud C (left) and Cloud H (right).

13

ORCID iDs

Duo Xu ⓘ https://orcid.org/0000-0001-6216-8931
Jonathan C. Tan ⓘ https://orcid.org/0000-0002-3389-9142
Ye Zhu ⓘ https://orcid.org/0000-0003-0774-9375

References

Andre´, P., Di Francesco, J., Ward-Thompson, D., et al. 2014, in Protostars and Planets VI, ed. H. Beuther et al. (Tucson, AZ: Univ. Arizona Press), 27
Avery, L. W., MacLeod, J. M., & Broten, N. W. 1982, ApJ, 254, 116
Bartlett, O. J., Benoit, D. M., Pimbblet, K. A., Simmons, B., & Hunt, L. 2023, MNRAS, 521, 6318
Beck, R. 2015, A&ARv, 24, 4
Becker, B., Vaccari, M., Prescott, M., & Grobler, T. 2021, MNRAS, 503, 1828
Bisbas, T. G., Tan, J. C., & Tanaka, K. E. I. 2021, MNRAS, 502, 2701
Bisbas, T. G., van Dishoeck, E. F., Hu, C.-Y., & Schruba, A. 2023, MNRAS, 519, 729
Bisigello, L., Conselice, C. J., Baes, M., et al. 2023, MNRAS, 520, 3529
Bryan, G. L., Norman, M. L., O'Shea, B. W., et al. 2014, ApJS, 211, 19
Butler, M. J., & Tan, J. C. 2012, ApJ, 754, 5
Chandrasekhar, S., & Fermi, E. 1953, ApJ, 118, 113
Davis, L. 1951, PhRv, 81, 890
Dedner, A., Kemm, F., Kröner, D., et al. 2002, JCoPh, 175, 645
Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. 2018, MNRAS, 476, 3661
Entekhabi, N., Tan, J. C., Cosentino, G., et al. 2022, A&A, 662, A39
Gheller, C., & Vazza, F. 2022, MNRAS, 509, 990
Gong, M., Ostriker, E. C., & Wolfire, M. G. 2017, ApJ, 843, 38
Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. 2014, in Proc. of the 27th Int. Conf. on Neural Information Processing Systems, Vol. 2 (New York: ACM), 2672
He, K., Zhang, X., Ren, S., & Sun, J. 2016, in 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Piscataway, NJ: IEEE), 770
Hemmati, S., Huff, E., Nayyeri, H., et al. 2022, ApJ, 941, 141
Herbst, E., & van Dishoeck, E. F. 2009, ARA&A, 47, 427
Heyer, M., & Dame, T. M. 2015, ARA&A, 53, 583
Ho, J., Jain, A., & Abbeel, P. 2020, in Advances in Neural Information Processing Systems, Vol. 33, ed. H. Larochelle et al. (Red Hook, NY: Curran Associates, Inc.), 6840
Hsu, C.-J., Tan, J. C., Christie, D., Cheng, Y., & O'Neill, T. J. 2023, MNRAS, 522, 700
Jiménez-Serra, I., Caselli, P., Fontani, F., et al. 2014, MNRAS, 439, 1996
Jørgensen, J. K., Belloche, A., & Garrod, R. T. 2020, ARA&A, 58, 727
Kainulainen, J., & Tan, J. C. 2013, A&A, 549, A53
Kingma, D. P., & Welling, M. 2014, in Int. Conf. on Learning Representations
Krumholz, M. R., McKee, C. F., & Bland-Hawthorn, J. 2019, ARA&A, 57, 227
Leroy, A. K., Walter, F., Brinks, E., et al. 2008, AJ, 136, 2782
Li, D., & Goldsmith, P. F. 2012, ApJ, 756, 12
Li, P., Cunningham, A., Gaches, B., et al. 2021, JOSS, 6, 3771
Li, P. S., Martin, D. F., Klein, R. I., & McKee, C. F. 2012, ApJ, 745, 139
Liu, H., Xu, Y., Wang, J., et al. 2020, ApJ, 894, 70
McKee, C. F., & Ostriker, E. C. 2007, ARA&A, 45, 565
Palmeirim, P., André, P., Kirk, J., et al. 2013, A&A, 550, A38
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 586, A138
Rao, R., Crutcher, R. M., Plambeck, R. L., & Wright, M. C. H. 1998, ApJL, 502, L75
Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. 2022, in IEEE/CVF Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE), 10674
Ronneberger, O., Fischer, P., & Brox, T. 2015, in Int. Conf. on Medical Image Computing and Computer-assisted Intervention (Cham: Springer), 234
Saintonge, A., & Catinella, B. 2022, ARA&A, 60, 319
Schloerb, F. P., Snell, R. L., & Young, J. S. 1983, ApJ, 267, 163
Shu, F. H., Adams, F. C., & Lizano, S. 1987, ARA&A, 25, 23
Singer, U., Polyak, A., Hayes, T., et al. 2022, arXiv:2209.14792
Smirnov, A. A., Savchenko, S. S., Poliakov, D. M., et al. 2023, MNRAS, 519, 4735
Smith, M. J., Geach, J. E., Jackson, R. A., et al. 2022, MNRAS, 511, 1808
Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. 2015, in Proc. of the 32nd Int. Conf. on Machine Learning, Vol. 37, ed. F. Bach & D. Blei (PLMR Press), 2256
Spitzer, L. 1978, Physical Processes in the Interstellar Medium (New York: Wiley)
Sweere, S. F., Valtchanov, I., Lieu, M., et al. 2022, MNRAS, 517, 4054
Tacconi, L. J., Genzel, R., & Sternberg, A. 2020, ARA&A, 58, 157
Tielens, A. G. G. M., & Hagen, W. 1982, A&A, 114, 245
Van Oort, C. M., Xu, D., Offner, S. S. R., & Gutermuth, R. A. 2019, ApJ, 880, 83
Wakelam, V., Smith, I. W. M., Herbst, E., et al. 2010, SSRv, 156, 13
Wang, P., & Abel, T. 2009, ApJ, 696, 96
Wu, B., Tan, J. C., Christie, D., & Nakamura, F. 2020, ApJ, 891, 168
Wu, B., Tan, J. C., Nakamura, F., et al. 2017, ApJ, 835, 137
Wu, B., Van Loo, S., Tan, J. C., & Bruderer, S. 2015, ApJ, 811, 56
Xu, D., Law, C.-Y., & Tan, J. C. 2023, ApJ, 942, 95
Xu, D., Offner, S. S. R., Gutermuth, R., Kong, S., & Arce, H. G. 2022a, ApJ, 926, 19
Xu, D., Offner, S. S. R., Gutermuth, R., & Oort, C. V. 2020a, ApJ, 890, 64
Xu, D., Offner, S. S. R., Gutermuth, R., & Oort, C. V. 2020b, ApJ, 905, 172
Xu, D., Offner, S. S. R., Gutermuth, R., & Tan, J. C. 2022b, ApJ, 941, 81
Zhang, S., Zhu, Z., & Kang, M. 2022, MNRAS, 510, 4473
Zhu, Y., Wu, Y., Deng, Z., Russakovsky, O., & Yan, Y. 2023a, arXiv:2302.08357
Zhu, Y., Wu, Y., Olszewski, K., et al. 2023b, Int. Conf. on Learning Representations