DENOISING DIFFUSION PROBABILISTIC MODELS TO PREDICT THE NUMBER DENSITY OF MOLECULAR CLOUDS IN ASTRONOMY

Anonymous authors
Paper under double-blind review

ABSTRACT

Denoising Diffusion Probabilistic Models (DDPMs) have become the mainstream generative approach in the Machine Learning and Computer Vision area, achieving state-of-the-art performance in synthesizing high-quality images, videos, and audio. In this work, we bring the DDPMs out of the data generation tasks, but to a new scientific application field in astronomy for inferring the volume or number density of giant molecular clouds (GMCs) from projected mass surface density maps. Specifically, we adopt magnetohydrodynamic (MHD) simulations with different global magnetic field strengths and large-scale dynamics, *i.e.*, noncolliding and colliding GMCs. We train a DDPM on both mass surface density maps and their corresponding mass-weighted number density maps from different viewing angles for all the simulations. We compare our 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 (CASI-2D). Experiments show that DDPMs achieve an order of magnitude improvement in the accuracy of predicting number density compared to that by other methods, demonstrating the promising potential of applying DDPMs in astrophysics.

1 INTRODUCTION

Giant molecular clouds (GMCs) are the birthplace of stars (Shu et al., 1987). Investigating the physical and chemical conditions of GMCs is a crucial step to understand the evolution of the interstellar medium (ISM) and the formation of stars (McKee & Ostriker, 2007; Heyer & Dame, 2015). Among all the physical quantities of GMCs, the number density or volume density is one of the most fundamental properties that relates to various physical quantity estimations, such as the free-fall time, the magnetic field strength (Chandrasekhar & Fermi, 1953) and chemical reaction rates (Tielens & Hagen, 1982). 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 (HC_3N, Avery et al., 1982; Schloerb et al., 1983; Li & Goldsmith, 2012). The relative intensity of different transitions of HC_3N 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_3N. Li & Goldsmith (2012) successfully observed $J = 2-1$ and $10-9$ transitions of HC_3N in Taurus B213 filament and constrained the molecular gas number density to be around 1.8×10^4 cm^{-3}. Unfortunately, the line ratio of HC_3N can only probe the number density at a relatively narrow range, between 10^4 and 10^6 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 the cloud 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 have been successfully applied to identify structures (Xu et al., 2020a;b; 2022a) and infer physical quantities, such as protostellar outflow inclination angles and magnetic field directions (Xu et al., 2022b; 2023). More recently, Denoising Diffusion Probabilistic Models (DDPMs) Sohl-Dickstein et al. (2015); Ho et al. (2020), originated from the natural thermodynamics problem, simulate a random walk process in the data space, which shares the intrinsic alignment and consistency with most scientific problems. For example, the observed emission or absorption maps of GMCs that derive the column density map are part of the results of the random walk of photons in the ISM, *i.e.*, the scattering process. Inferring the raw mass-weighted number density distribution based on the observed column density inherits the basic concepts of diffusion models. In addition, DDPMs have demonstrated their proficiency and robustness in image and audio generation (Sohl-Dickstein et al., 2015; Ho et al., 2020; Zhu et al., 2023; Rombach et al., 2022), which are suitable for the prediction task in astronomy. Therefore, our primary objective in this work is to demonstrate the great potential for applying DDPMs in astronomy studies.

2 METHOD

The core formulation of diffusion models is inspired by non-equilibrium 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:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}), \tag{1}$$

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

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_t^\theta(\boldsymbol{x}_t)) + \sigma_t z_t, \tag{2}$$

where ϵ_t^θ is the learnable noise predictor, $z_t \sim \mathcal{N}(0, \mathbf{I})$, 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:

$$\mathcal{L} = \mathbb{E}_q[-\log p(\boldsymbol{x}_T) - \sum_{t\geq 1} \log \frac{p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)}{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})}]. \tag{3}$$

To adapt the diffusion models in our context, we deploy a diffusion model with a similar parameter setup as 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 $\boldsymbol{x}_c$ 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.

3 EXPERIMENTS

Data. We carry out ideal MHD simulations based on the set-up of Wu et al. (2020), which were conducted using the adaptive mesh refinement (AMR) code Enzo (Bryan et al., 2014) with a top grid resolved by 256^3 and 4 additional levels of refinement, achieving a resolution of 0.03125 pc[1] within a 128 pc^3 domain. The simulations include heating/cooling based on a photo-dissociation model with FUV radiation field of $G_0 = 4$ Habings and cosmic ray ionization rate $\zeta = 10^{-16}\,\mathrm{s}^{-1}$, self-gravity, magnetic fields and supersonic turbulence. The simulations do not include star formation or feedback, so represent the structures that develop in the early phases of collapse up to the onset of star formation. We select two different types of setup of the large-scale dynamics of the GMCs, noncolliding and colliding GMCs. For each setup, we implement three different magnetic field strengths, 10, 30, and 50 μG. In the colliding cases, the clouds have relative velocities of 10 km s^{-1},

[1] 1 pc=3.086×10^{16} m

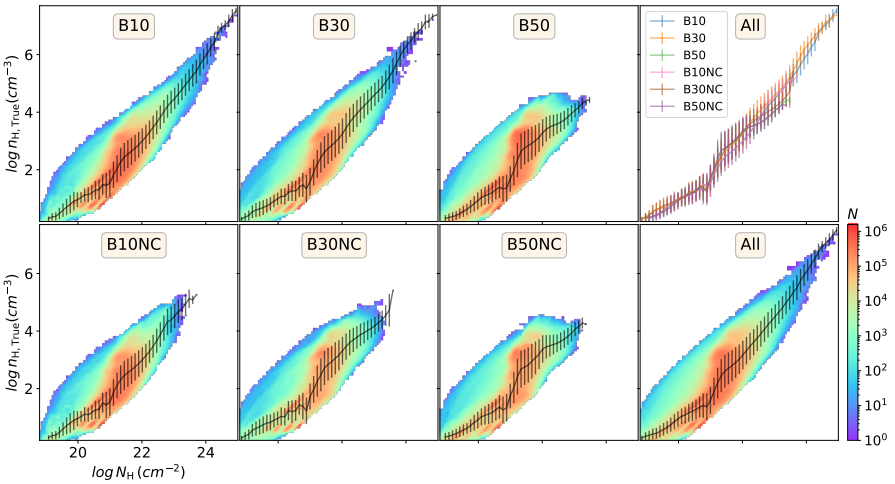

Figure 1: Relation between the LOS mass-weighted number density and the column density for different simulations. The rainbow color background indicates the 2D histogram of the distribution between number density and column density for each simulation. The lines with errorbars represent the mean and their standard deviation of each column density bin.

Table 1: Power-Law Fitting Results. Label, number of power-law components, break points of power-law fittings in log scale, and power indices of each power-law component.

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

and are offset by $0.5R_{\rm GMC}$. The GMCs have an initial temperature of 15 K, but soon establish a multiphase temperature structure. The simulations are run for 5 Myr. We take two evolutionary stages, 3 and 4 Myr, for each setup 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, 4 and 2 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.

Results. We adopt several different approaches to convert column density to mass-weighted number density on the LOS. We start with power-law fitting on the relation between the LOS mass-weighted number density of H nuclei ($n_{\rm H}$) and the column density of H nuclei ($N_{\rm H}$) for all simulations in Figure 1. We adopt two-components and three components power law to fit the $n_{\rm H} - N_{\rm H}$ relation, i.e., $n_{\rm H} = f(N_{\rm H})$. Meanwhile, we conduct the "inverted" fitting, i.e., adopting two-component and three component power laws to fit the $N_{\rm H} - n_{\rm H}$ relation $N_{\rm H} = f(n_{\rm H})$ and then derive the inverse function $n_{\rm H} = f^{-1}(N_{\rm H})$. 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 2. In addition, we present the result from machine learning approaches, including CASI-2D (Van Oort et al., 2019) and diffusion model, in Figure 2. It is obvious that there is significant dispersion between the true mass-weighted number density $n_{\rm H,True}$ and the predicted number density $n_{\rm H,Pred}$ that is converted by power-law and that predicted by CASI-2D. The predicted number density by the diffusion model has a much smaller dispersion around the true density. We list the dispersion between $n_{\rm H,True}$ and $n_{\rm H,Pred}$ in Figure 2. The predicted number density by

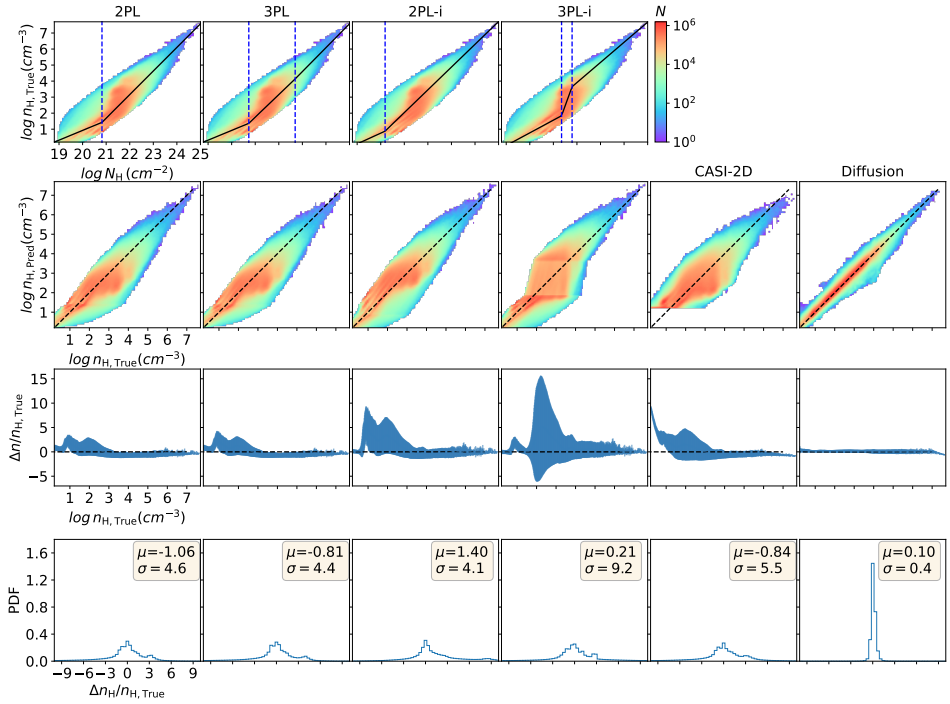

Figure 2: Summary of the performance of different approaches to convert the gas column density to number density on all the data samples.

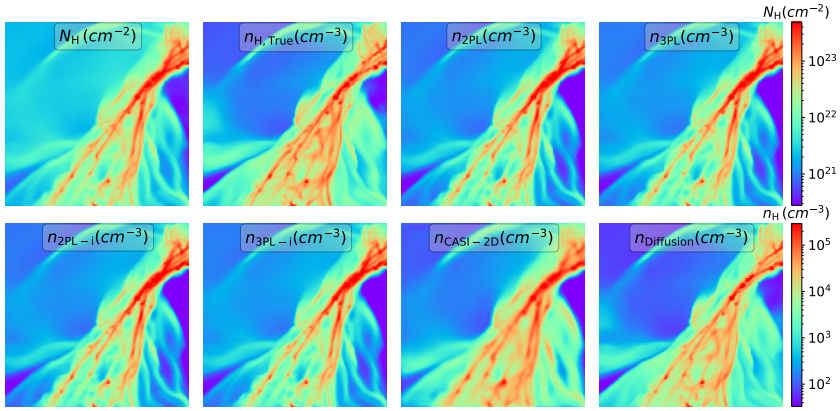

Figure 3: Comparison between different approaches to converting gas column density to number density on a sample in the test set.

the diffusion model is an order of magnitude better than that by all the other approaches. We show a sample image from the test set and apply different approaches to obtain the predicted number density in Figure 3. Although the column density map is similar to the true 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 number density map predicted by the diffusion model is able to recreate the actual structures across a wide density range.

4 Conclusion

In this work, we deploy the DDPMs to predict the LOS mass-weighted number density of GMCs from column density maps in astronomy. We achieve large improvements in the performance for both synthetic test samples and real observational data, demonstrating the effectiveness and great potential for astronomical applications using diffusion models.

REFERENCES

P. André, J. Di Francesco, D. Ward-Thompson, S. I. Inutsuka, R. E. Pudritz, and J. E. Pineda. From Filamentary Networks to Dense Cores in Molecular Clouds: Toward a New Paradigm for Star Formation. In Henrik Beuther, Ralf S. Klessen, Cornelis P. Dullemond, and Thomas Henning (eds.), *Protostars and Planets VI*, pp. 27–51, January 2014. doi: 10.2458/azu_uapress_9780816531240-ch002.

L. W. Avery, J. M. MacLeod, and N. W. Broten. A model of Taurus Molecular Cloud 1 based on HC3N observations. *The Astrophysical Journal*, 254:116–125, March 1982. doi: 10.1086/159713.

Thomas G. Bisbas, Jonathan C. Tan, and Kei E. I. Tanaka. Photodissociation region diagnostics across galactic environments. *Monthly Notices of the Royal Astronomical Society*, 502(2):2701–2732, April 2021. doi: 10.1093/mnras/stab121.

Thomas G. Bisbas, Ewine F. van Dishoeck, Chia-Yu Hu, and Andreas Schruba. PDFCHEM: A new fast method to determine ISM properties and infer environmental parameters using probability distributions. *Monthly Notices of the Royal Astronomical Society*, 519(1):729–753, February 2023. doi: 10.1093/mnras/stac3487.

Greg L. Bryan, Michael L. Norman, Brian W. O'Shea, Tom Abel, John H. Wise, Matthew J. Turk, Daniel R. Reynolds, David C. Collins, Peng Wang, Samuel W. Skillman, Britton Smith, Robert P. Harkness, James Bordner, Ji-hoon Kim, Michael Kuhlen, Hao Xu, Nathan Goldbaum, Cameron Hummels, Alexei G. Kritsuk, Elizabeth Tasker, Stephen Skory, Christine M. Simpson, Oliver Hahn, Jeffrey S. Oishi, Geoffrey C. So, Fen Zhao, Renyue Cen, Yuan Li, and Enzo Collaboration. ENZO: An Adaptive Mesh Refinement Code for Astrophysics. *The Astrophysical Journal Supplement Series*, 211(2):19, April 2014. doi: 10.1088/0067-0049/211/2/19.

S. Chandrasekhar and E. Fermi. Magnetic Fields in Spiral Arms. *The Astrophysical Journal*, 118:113, July 1953. doi: 10.1086/145731.

Mark Heyer and T. M. Dame. Molecular Clouds in the Milky Way. *Annual Review of Astronomy and Astrophysics*, 53:583–629, August 2015. doi: 10.1146/annurev-astro-082214-122324.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Di Li and Paul F. Goldsmith. Is the Taurus B213 Region a True Filament?: Observations of Multiple Cyanoacetylene Transitions. *The Astrophysical Journal*, 756(1):12, September 2012. doi: 10.1088/0004-637X/756/1/12.

Christopher F. McKee and Eve C. Ostriker. Theory of Star Formation. *Annual Review of Astronomy and Astrophysics*, 45(1):565–687, September 2007. doi: 10.1146/annurev.astro.45.051806.110602.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

F. P. Schloerb, R. L. Snell, and J. S. Young. Structure of dense molecular gas in TMC 1 from observations of three transitions of HC3N. *The Astrophysical Journal*, 267:163–173, April 1983. doi: 10.1086/160854.

Frank H. Shu, Fred C. Adams, and Susana Lizano. Star formation in molecular clouds: observation and theory. *Annual Review of Astronomy and Astrophysics*, 25:23–81, January 1987. doi: 10.1146/annurev.aa.25.090187.000323.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

A. G. G. M. Tielens and W. Hagen. Model calculations of the molecular composition of interstellar grain mantles. *Astronomy and Astrophysics*, 114(2):245–260, October 1982.

Colin M. Van Oort, Duo Xu, Stella S. R. Offner, and Robert A. Gutermuth. CASI: A Convolutional Neural Network Approach for Shell Identification. *The Astrophysical Journal*, 880(2):83, August 2019. doi: 10.3847/1538-4357/ab275e.

Benjamin Wu, Jonathan C. Tan, Fumitaka Nakamura, Sven Van Loo, Duncan Christie, and David Collins. GMC Collisions as Triggers of Star Formation. II. 3D Turbulent, Magnetized Simulations. *The Astrophysical Journal*, 835(2):137, February 2017. doi: 10.3847/1538-4357/835/2/137.

Benjamin Wu, Jonathan C. Tan, Duncan Christie, and Fumitaka Nakamura. GMC Collisions as Triggers of Star Formation. VII. The Effect of Magnetic Field Strength on Star Formation. *The Astrophysical Journal*, 891(2):168, March 2020. doi: 10.3847/1538-4357/ab77b5.

Duo Xu, Stella S. R. Offner, Robert Gutermuth, and Colin Van Oort. Application of Convolutional Neural Networks to Identify Stellar Feedback Bubbles in CO Emission. *The Astrophysical Journal*, 890(1):64, February 2020a. doi: 10.3847/1538-4357/ab6607.

Duo Xu, Stella S. R. Offner, Robert Gutermuth, and Colin Van Oort. Application of Convolutional Neural Networks to Identify Protostellar Outflows in CO Emission. *The Astrophysical Journal*, 905(2):172, December 2020b. doi: 10.3847/1538-4357/abc7bf.

Duo Xu, Stella S. R. Offner, Robert Gutermuth, Shuo Kong, and Hector G. Arce. A Census of Protostellar Outflows in Nearby Molecular Clouds. *The Astrophysical Journal*, 926(1):19, February 2022a. doi: 10.3847/1538-4357/ac39a0.

Duo Xu, Stella S. R. Offner, Robert Gutermuth, and Jonathan C. Tan. A Census of Outflow to Magnetic Field Orientations in Nearby Molecular Clouds. *The Astrophysical Journal*, 941(1):81, December 2022b. doi: 10.3847/1538-4357/aca153.

Duo Xu, Chi-Yan Law, and Jonathan C. Tan. Application of Convolutional Neural Networks to Predict Magnetic Fields Directions in Turbulent Clouds. *The Astrophysical Journal*, 942(2):95, January 2023.

Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal and conditional generation. In *International Conference on Learning Representations*, 2023.