Self-supervised Synthetic Pretraining for Inference of Stellar Mass Embedded in Dense Gas

Keiya Hirashima*
Center for Interdisciplinary Theoretical and Mathematical Sciences (iTHEMS)
RIKEN
Wako, Japan
keiya.hirashima@riken.jp

Shingo Nozaki
Department of Earth and Planetary Sciences
Kyushu University
Fukuoka, Japan
nozaki.shingo.307@s.kyushu-u.ac.jp

Naoto Harada
Department of Astronomy
University of Tokyo
Tokyo, Japan
hrdnot-3997@g.ecc.u-tokyo.ac.jp

Abstract

Stellar mass is a fundamental quantity that determines the properties and evolution of stars. However, estimating stellar masses in star-forming regions is challenging because young stars are obscured by dense gas and the regions are highly inhomogeneous, making spherical dynamical estimates unreliable. Supervised machine learning could link such complex structures to stellar mass, but it requires large, high-quality labeled datasets from high-resolution magneto-hydrodynamical (MHD) simulations, which are computationally expensive. We address this by pretraining a vision transformer on one million synthetic fractal images using the self-supervised framework DINOv2, and then applying the frozen model to limited high-resolution MHD simulations. Our results demonstrate that synthetic pretraining improves frozen-feature regression stellar mass predictions, with the pretrained model performing slightly better than a supervised model trained on the same limited simulations. Principal component analysis of the extracted features further reveals semantically meaningful structures, suggesting that the model enables unsupervised segmentation of star-forming regions without the need for labeled data or fine-tuning.

1 Introduction

Stellar mass is a fundamental stellar property that governs luminosity, lifetime, stellar evolutionary tracks, and stellar nucleosynthesis, which produce the chemical elements essential for the origin of our solar system and life. In astronomy, the Initial Mass Function (IMF) describes the stellar mass distribution [1]. Observations suggest that the IMF exhibits a remarkably similar shape across diverse environments, while the physical mechanisms governing this distribution remain unclear [2]. To uncover the origin of the IMF, it is essential to determine the masses of young, still-forming stars (protostars and pre-main-sequence stars) through observations. However, determining the masses of young stars from direct observations remains highly challenging. They are deeply embedded within their natal molecular clouds, making them invisible with optical light. Furthermore, their luminosity

*RIKEN Special Postdoctoral Researcher

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Machine Learning and the Physical Sciences.

originates primarily from gas accretion rather than stellar radiation, making mass estimates difficult using methods commonly applied to main-sequence stars.

Predicting stellar masses from their environments is a challenging task. The gas is highly inhomogeneous, making analytic models unreliable, while capturing the physics requires three-dimensional (3D) simulations, which are too expensive to produce in large numbers [3]. A promising approach is to combine high-resolution simulations with deep learning. In our work, three-dimensional (3D) magneto-hydrodynamical (MHD) simulations are employed to capture the physics of star formation and track stellar mass growth [4]. We then leverage two-dimensional (2D) maps of gas projected from these 3D high-resolution MHD simulations to develop deep learning models for predicting stellar masses. Since large amounts of high-resolution simulations or observational data with labels are rarely available, we propose a framework that combines self-supervised pretraining and downstream tasks. Models are trained on numerous synthetic images to learn robust visual representations, while limited high-resolution MHD simulations are reserved for zero-shot and frozen-feature evaluation on downstream tasks.

2 Related work

Self-supervised Learning for Astrophysics Self-supervised learning has become a powerful approach for extracting image representations without labels and could help mitigate challenges from limited labeled data. Early methods like MoCoV3 [5] and MAE [6] improved vision transformer (ViT) pretraining but often required supervised fine-tuning. In contrast, DINOv2 [7–9] captures semantic structures by enforcing consistency across multiple views, transfers to downstream tasks without fine-tuning backbones, and requires only lightweight classifiers or k-nearest neighbor evaluation. In astrophysics, self-supervised learning has enabled galaxy classification with sparse labels [10, 11], inference of galaxy properties by combining simulations and observations [12], and mitigation of observational biases through metadata [13], with DINOv2 recently applied to galaxy images [14].

Pretraining with Synthetic Data Supervised deep learning has achieved remarkable success by training models on large labeled datasets. An alternative line of research investigates the use of synthetic data generated with mathematical equations to achieve competitive performance across various downstream tasks. Such data can be produced inexpensively and in large quantities, without demanding experiments, observations, or extensive computational resources, and without raising ethical or privacy concerns. A pioneering study [15] showed that supervised pretraining on fractal images alone can reach competitive accuracy on natural images, in some cases even surpassing models pretrained on `ImageNet-22k` [16]. This approach has since been extended to supervised learning with ViTs [16–19] and to self-supervised learning with convolutional neural networks [20, 21].

3 Methodology

3.1 Data Generation

Synthetic Images for Pretraining We extend the Flame algorithm [22] to generate our datasets of fractal images. With randomly sampled parameters $\theta_i = (a_i, b_i, c_i, d_i, e_i, f_i)$ for rotation and shifting fed to a translation w, coordinates are sampled through an iterated function system (IFS) [23],

$$w(\boldsymbol{x}; \theta_i) = \begin{pmatrix} a_i & b_i \\ c_i & d_i \end{pmatrix} \boldsymbol{x} + \begin{pmatrix} e_i \\ f_i \end{pmatrix}, \tag{1}$$

where $\boldsymbol{x}$ is a coordinate. At each sampling step, one of the non-linear variations (e.g., spherical and bubble) of the original Flame algorithm [22] is probabilistically applied, yielding the next sampled point $\boldsymbol{x}_{i+1} = w(\boldsymbol{x}_i; \theta_i)$. Those sampled points are interpolated to 336×336 images. We draw candidate frames with one million points, accept only those that cover $\geq 80\%$ of the image plane, and thus build a dataset of 1M images. Examples are shown in Fig. 1 (1). The dataset generation was executed with a throughput of 2.67 TFLOPS per image and 2.67 EFLOPS in total.

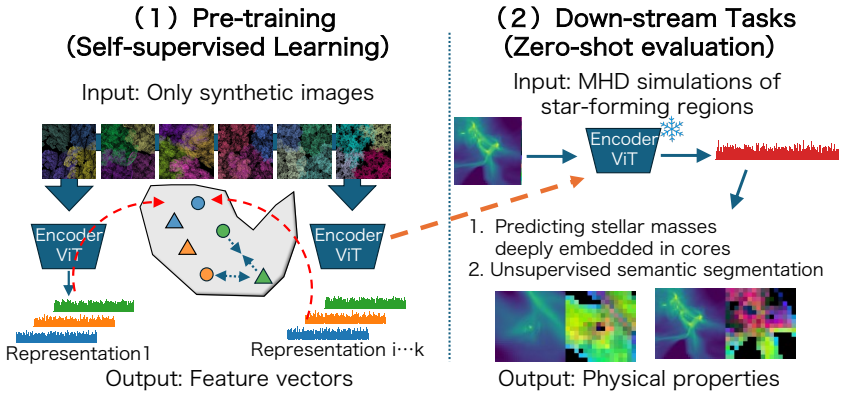


Figure 1: Overview of our model. *Left*: self-supervised pretraining with synthetic fractal images using DINOv2 to extract feature vectors. *Right*: zero-shot evaluation on simulation with the frozen encoder, applied to stellar mass prediction (k-NN) and semantic segmentation (PCA-based colors).

Simulations of star-forming regions We performed 3D MHD simulations with `SFUMATO`, an adaptive mesh refinement code [4, 24–26], in a cubic box of 4 parsec[2] per side containing 3000 $M_\odot$[3] of gas with an initial uniform proton density of 1365 cm^{-3} and a magnetic field of 10 μG along the z-axis. To follow the long-term evolution, we use the sink particle method [25], in which unstable dense clumps are replaced by sink particles that accrete gas within a fixed radius (5.0×10^{-4} pc). The accreted mass is taken as the protostellar mass, allowing us to trace protostellar growth. The finest spatial resolution is $\Delta x \sim 3 \times 10^{-3}$ parsec, sufficient to resolve the Jeans length with more than five cells, given an initial velocity field with a Mach number of 10. Our dataset consists of 32k snapshots of 0.5 pc regions centered on protostars, from which we construct 64×64 maps of column density, mean line-of-sight velocity, and its velocity dispersion along the x, y, and z axes, each paired with the elapsed time since protostar formation and the stellar mass shown in Fig. 1 (2). The simulation was executed with a throughput of 2540 TFLOPS per snapshot and 81.2 EFLOPS in total.

3.2 Model Implementation

We employ a ViT-L/16 encoder within the DINOv2 framework for self-supervised pretraining and zero-shot or frozen-feature evaluation. The encoder is pretrained on 1M synthetic fractal images at a resolution of 336 for 100 epochs with a batch size of 1024 and a patch size of 16. For comparison, we implement a ResNet-18 [27] baseline trained in a fully supervised manner. Both models use a cosine-annealed learning rate schedule with a maximum of 0.04, including 10 warm-up epochs followed by 90 epochs of cosine decay. The ResNet-18 is trained with an L_2 regression loss. Simulation data are preprocessed by applying a logarithmic transformation to stellar mass and column density, and min–max normalization to mean line-of-sight velocity and its dispersion.

3.3 Experiments

Self-supervised Pretraining with Synthetic Data and k-NN Regression The pretrained ViT-L/16 encoder is applied to 32k snapshots from star-formation simulations at a resolution of 64 to obtain 1024-dimensional feature vectors. Principal component analysis (PCA) is fitted on the training split and applied to all features while preserving the full dimensionality of 1024 (PCA whitening). The transformed features are then evaluated with a distance-weighted k-nearest neighbors (k-NN) regressor ($k = 5$) to predict the logarithm of stellar mass, using 24k training and 8k test samples. Prediction performance is assessed in terms of root-mean-square error (RMSE) and the coefficient of determination (R^2).

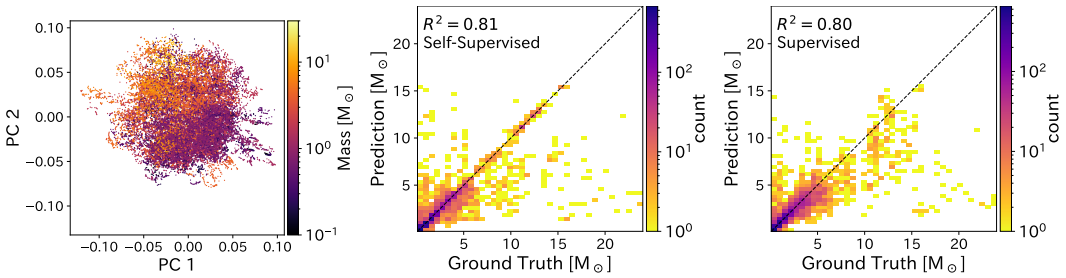
Zero-shot Unsupervised Feature Visualization To examine the semantic structure of the learned representations, feature vectors from the pretrained ViT-L/16 encoder are projected with PCA and

[2] 1 parsec $\sim 3.08 \times 10^{13}$ km ~ 3.26 light years

[3] 1 $M_\odot$ (solar mass) is equal to the mass of the Sun.

the first three components are mapped to the RGB color space. Although pretraining is performed with a patch size of 16, for visualization we linearly upsample 4×4 input patches to 16×16 prior to encoding, in order to maintain consistency with the token granularity of the model.

4 Results



(a) (b) (c)

Figure 2: Frozen-feature regression of stellar masses. **(a)** PCA projection of feature vectors from DINOv2 colored by stellar mass. **(b)** True versus predicted stellar masses using DINOv2 representations with k-NN regression. **(c)**True versus predicted stellar masses from a supervised ResNet-18 baseline.

We evaluate the ViT-L/16 encoder pretrained with fractal images on zero-shot and frozen-feature tasks, keeping the model parameters frozen.

Frozen-feature Regression on Stellar Masses Fig. 2a shows a scatter plot of the first and second PCA components of feature vectors from column density, mean line-of-sight velocity, and its velocity dispersion maps. The distribution exhibits a weak trend with stellar masses, which are otherwise difficult to infer from 2D density and velocity information alone. To evaluate predictive performance, we use all PCA components with a k-NN regressor trained on the training set and tested on the validation set. Fig. 2b and Fig. 2c compare stellar mass predictions from our method and a supervised ResNet-18 baseline, respectively. Both approaches follow the ground truth trend up to ~ 6 $M_\odot$, where more than 10^2 training samples are available. Beyond this, with fewer than 10 samples, ResNet-18 tends to underestimate stellar masses, while DINOv2 captures many true values in the range 6–15 $M_\odot$. At higher masses, neither model performs reliably due to data scarcity. Table 1 shows that PCA features slightly improve scores over raw feature vectors and that pretraining with fractal images markedly improves performance compared to the result of random initialization with DINOv2.

	ResNet-18		DINOv2 + k-NN (k=5)		
Methods	Random Init.	Pretrained	Random Init.	Pretrained	with PCA whitening
R^2 (↑)	-1.9	0.80	-0.58	0.80	**0.81**
RMSE (↓)	0.34	0.089	0.52	0.089	**0.088**

Table 1: R^2 and RMSE of frozen-feature regression on stellar mass using ResNet-18 and DINOv2.

Zero-shot Semantic Segmentation with PCA-based Colors Fig. 3 shows four examples (a–d), each containing four panels: column density $N_{\rm HI}$, mean line-of-sight velocity $v_{\rm los}$, its velocity dispersion σ_v, and a color map based on the first three PCA components of the 1024-dimensional feature vectors. Black areas correspond to either diffuse, low-density regions or regions of very high velocity dispersion, the latter likely marking sites of ongoing star formation. Yellow to yellow-green areas highlight regions of low velocity dispersion (Fig. 3a–3c). Magenta and dodgerblue indicate negative and positive line-of-sight velocities in regions of high velocity dispersion (Fig. 3b–3d), where gas may accrete onto dense cores and contribute to stellar growth. Notably, this semantic segmentation arises directly from the PCA projection of pretrained representations, without any labeled data or supervised fine-tuning.

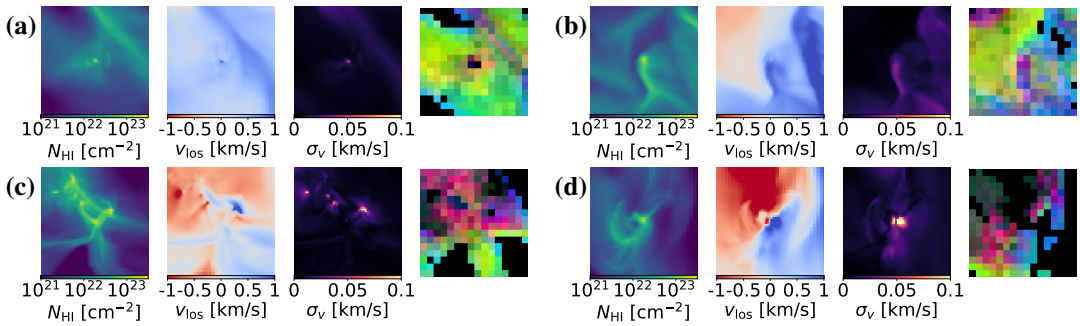


Figure 3: Snapshots from MHD simulations with visualizations of PCA components of feature vectors. Each panel shows four maps: column density $N_{\rm HI}$, mean line-of-sight velocity $v_{\rm los}$, its velocity dispersion σ_v, and a color map of the first three PCA components from image patches.

5 Summary and Discussion

Our results demonstrate that self-supervised synthetic pretraining can serve as a data-efficient alternative to supervised pipelines in high-resolution, yet data-limited, MHD simulations, with a ViT encoder achieving performance comparable to that of supervised learning. PCA-based visualization further revealed meaningful structures, such as dense cores and inflows, motivating the extension of this approach to stellar property prediction and its direct application to observational data.

The present approach indicates a potential to predict protostellar masses, and incorporating information from more extended gas and velocity fields may ultimately enable predictions of the final stellar mass and the IMF. The framework, however, still relies on labeled simulation data for training in order to apply it to observations. Meanwhile, PCA-based segmentation highlights that broad structural patterns can be identified without labels, though addressing observational noise—potentially by constructing datasets from the noise itself—will be crucial for robustness in practice.

Acknowledgments and Disclosure of Funding

The authors thank Takaharu Yaguchi, Kazuki Tokuda, Yoshito Shimajiri, and Kana Moriwaki for fruitful discussions. This work was initiated following discussions at the International Conference on Scientific Computing and Machine Learning 2025 and the 37th Rironkon (Community of Theoretical Astrophysics) annual symposium 2024. KH is supported by the Special Postdoctoral Researchers Program at RIKEN. SN is supported by JST SPRING, Grant Number JPMJSP2136. NH is supported by JSPS KAKENHI Grant Number JP25K17434. Numerical computations were carried out on Flatiron Institute's Rusty computing cluster, the Cray XC50 (Aterui II) at the Center for Computational Astrophysics of the National Astronomical Observatory of Japan, and the Yukawa-21 at the Yukawa Institute for Theoretical Physics (YITP), Kyoto University. We gratefully acknowledge the Flatiron Institute and the Scientific Computing Core for their support.

References

[1] Salpeter, E. E. The Luminosity Function and Stellar Evolution. *ApJ*, 121:161, 1955.

[2] Offner, S. S. R., P. C. Clark, P. Hennebelle, et al. The Origin and Universality of the Stellar Initial Mass Function. In H. Beuther, R. S. Klessen, C. P. Dullemond, T. Henning, eds., *Protostars and Planets VI*, pages 53–75. 2014.

[3] Pelkonen, V. M., P. Padoan, T. Haugbølle, et al. From the CMF to the IMF: beyond the core-collapse model. *MNRAS*, 504(1):1219–1236, 2021.

[4] Nozaki, S., H. Fukushima, K. Tokuda, et al. Tracking Star-forming Cores as Mass Reservoirs in Clustered and Isolated Regions Using Numerical Passive Tracer Particles. *ApJ*, 980(1):101, 2025.

[5] Chen, X., S. Xie, K. He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649. 2021.

[6] He, K., X. Chen, S. Xie, et al. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988. 2022.

[7] Caron, M., H. Touvron, I. Misra, et al. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640. 2021.

[8] Zhou, J., C. Wei, H. Wang, et al. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022.

[9] Oquab, M., T. Darcet, T. Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification.

[10] Hayat, M. A., G. Stein, P. Harrington, et al. Self-supervised Representation Learning for Astronomical Images. *ApJ*, 911(2):L33, 2021.

[11] Desmons, A., S. Brough, F. Lanusse. Detecting galaxy tidal features using self-supervised representation learning. *MNRAS*, 531(4):4070–4084, 2024.

[12] Eisert, L., C. Bottrell, A. Pillepich, et al. ERGO-ML: comparing IllustrisTNG and HSC galaxy images via contrastive learning. *MNRAS*, 528(4):7411–7439, 2024.

[13] Rizhko, M., J. S. Bloom. AstroM^3: A Self-supervised Multimodal Model for Astronomy. *AJ*, 170(1):28, 2025.

[14] Parker, L., F. Lanusse, S. Golkar, et al. AstroCLIP: a cross-modal foundation model for galaxies. *MNRAS*, 531(4):4990–5011, 2024.

[15] Kataoka, H., K. Okayasu, A. Matsumoto, et al. Pre-training without natural images. In *Computer Vision – ACCV 2020: 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 – December 4, 2020, Revised Selected Papers, Part VI*, page 583–600. Springer-Verlag, Berlin, Heidelberg, 2020.

[16] Kataoka, H., R. Hayamizu, R. Yamada, et al. Replacing labeled real-image datasets with auto-generated contours. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21232–21241. 2022.

[17] Nakashima, K., H. Kataoka, A. Matsumoto, et al. Can vision transformers learn without natural images? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):1990–1998, 2022.

[18] Nakamura, R., H. Kataoka, S. Takashima, et al. Pre-training Vision Transformers with Very Limited Synthesized Images . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20303–20312. IEEE Computer Society, Los Alamitos, CA, USA, 2023.

[19] Nakamura, R., R. Tadokoro, R. Yamada, et al. Scaling backwards:minimal synthetic pre-training? In *Proceedings of the 2024 European Conference on Computer Vision (ECCV)*. 2024.

[20] Baradad Jurjo, M., J. Wulff, T. Wang, et al. Learning to see by looking at noise. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan, eds., *Advances in Neural Information Processing Systems*, vol. 34, pages 2556–2569. Curran Associates, Inc., 2021.

[21] Baradad, M., C.-F. R. Chen, J. Wulff, et al. Procedural image programs for representation learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc., Red Hook, NY, USA, 2022.

[22] Draves, S., E. Reckase. The fractal flame algorithm, 2008. Accessed 2025-08-15.

[23] Barnsley, M. *Fractals Everywhere*. Academic Press, San Diego, 1988.

[24] Matsumoto, T. Self-Gravitational Magnetohydrodynamics with Adaptive Mesh Refinement for Protostellar Collapse. *PASJ*, 59:905, 2007.

[25] Matsumoto, T., K. Dobashi, T. Shimoikura. Star Formation in Turbulent Molecular Clouds with Colliding Flow. *ApJ*, 801(2):77, 2015.

[26] Fukushima, H., H. Yajima. Radiation hydrodynamics simulations of massive star cluster formation in giant molecular clouds. *MNRAS*, 506(4):5512–5539, 2021.

[27] He, K., X. Zhang, S. Ren, et al. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. 2016.