Estimating Probability Densities of Tabular Data using a Transformer Model combined with Denoising Diffusion

Henry W. Leung [1 2]   **Jo Bovy** [1 2]   **Joshua S. Speagle** [1 2]

Abstract

Transformers are often the go-to architecture to build foundation models that ingest a large amount of training data. But these models do not estimate the probability density distribution when trained on regression problems, yet obtaining full probabilistic outputs is crucial to many fields of science, where the probability distribution of the answer can be non-Gaussian and multimodal. In this work, we demonstrate that training a probabilistic model using a denoising diffusion head on top of the Transformer provides reasonable probability density estimation even for high-dimensional inputs. The combined Transformer+Denoising Diffusion model allows conditioning the output probability density on arbitrary combinations of inputs and it is thus a highly flexible density function emulator of all possible input/output combinations. We illustrate our Transformer+Denoising Diffusion model by training it on a large dataset of astronomical observations and measured labels of stars within our Galaxy and we apply it to a variety of inference tasks to show that the model can infer labels accurately with reasonable distributions.

1. Introduction

Many scientific fields are becoming increasingly data-driven and use machine learning for a large variety of analysis and synthesis tasks. For example, in the field of astronomy, many areas benefit from large-scale surveys across multiple areas such as spectroscopy, photometry, and time-domain observations. Good examples of this are the *Gaia* survey,

which mostly produces tabular data, the Sloan Digital Sky Survey (SDSS; Blanton et al. 2017; Kollmeier et al. 2017) spectroscopic surveys and The Dark Energy Camera Legacy Survey (DECaLS; Dey et al. 2019) imaging surveys, and the upcoming Vera C. Rubin Observatory (LSST; Ivezić et al. 2019) which will produce multi-band images and time-series data for billions of objects. In order to build a machine learning model for astronomy that is useful across a wide range of use cases, such a model needs to be trained at scale across all of these datasets.

Foundation models are an active area of research in a wide range of areas, but for our purposes applications in scientific areas are most interesting. There they are showing great promise for revolutionizing science as they are able to leverage commonalities between different scientific areas (e.g., McCabe et al. 2023) and because they obey similar scaling laws as found in natural language processing, allowing these models to outperform traditional methods when trained with large amounts of data (Walmsley et al., 2024).

Transformers are the currently most powerful and flexible architecture for building foundation models, especially in natural language because the attention mechanism used in Transformers can capture contextual dependencies within a sequence of data and can therefore predict the missing/next words. But in the astronomical data context where the learning task is regression, they only predict scalars for each feature with perhaps a scalar predictive uncertainty to represent the width of a Gaussian uncertainty distribution. While the Gaussian assumption is often a good assumption in science due to the central limit theorem, it can be untrue/uninformative especially when using foundation models performing a wide range of tasks on labels whose distributions can be non-Gaussian and multi-modal.

To infer probability densities for the output instead of scalars, we need to incorporate generative models on top of the Transformers, such that the model can learn the underlying distribution of the training data abd can generate plausible samples. In our model, we want the model to generate plausible samples based on conditions from Transformers. Possible choice of generative models for our application include Normalizing Flows (NF; Jimenez Rezende & Mohamed 2015) or a Denoising Diffusion Probabilistic Model

[1]David A. Dunlap Department of Astronomy and Astrophysics, University of Toronto, 50 St. George Street, Toronto, Ontario, M5S 3H4, Canada [2]Dunlap Institute for Astronomy and Astrophysics, University of Toronto, 50 St. George Street, Toronto, Ontario, M5S 3H4, Canada. Correspondence to: Henry W. Leung <henrysky.leung@utoronto.ca>, Jo Bovy <bovy@utoronto.ca>, Joshua S. Speagle <j.speagle@utoronto.ca>.
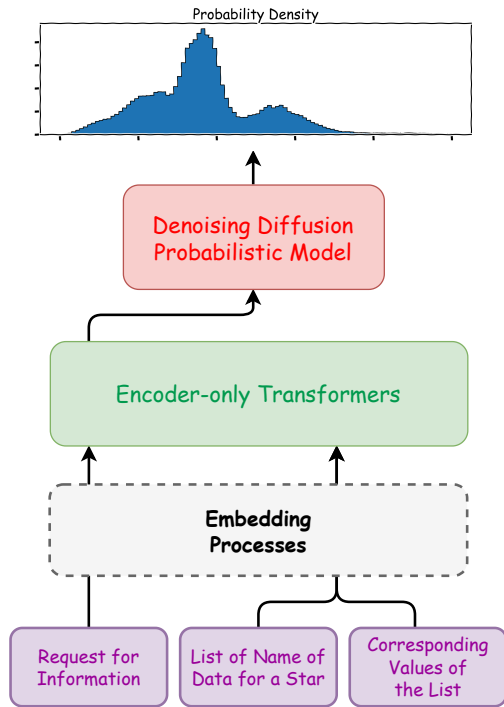
Figure 1. High level model architecture of our Transformer+Denoising Diffusion foundation model. The goal of the model is to estimate the probability density of an output based on a list of input scientific data and what unknown data is being requested. The role of the Denoising Diffusion Probabilistic Model (DDPM) is to turn the hidden state at the first position from the Transformer to a probability density distribution.

Figure 2. Probability density of different surface temperature T_{eff} (leftmost), surface gravity $\log g$ (middle left), metallicity [M/H] (middle right) and G-band luminosity (rightmost) of stars based on no conditions. The blue solid lines show the probability density from our model while the orange colored histogram shows the probability density from the training set. This figure demonstrates that our model learns the training set distribution on various labels when there is no condition.

(DDPM; Ho et al. 2020). Here we chose DDPM because of its flexibility of inferring a wide range of distributions with a single model. DDPM takes samples from the training data distribution, incrementally adding noise in order to corrupt the data, and the role of DDPM is to gradually de-noise the data. A population of new samples can be generated from de-noising random noise (often normal) with the trained DDPM.

The usage of Transformer and DDPM together is not a new idea and has been applied in various ways and various tasks. For example, in natural language settings like Kapur et al. (2024) for diagram generation using a Transformer with DDPM head, or using the Transformer architecture within DDPM like Gloeckler et al. (2024) for simulation-based inference and Peebles & Xie (2022) for natural language image generation.

In this work, we extend the work of Leung & Bovy (2024) (thereafter LB24) on a scientific foundation model by incorporating a DDPM head on top of a Transformer trained in a regression setting (as opposed to logistic regression used in natural language), trained with a small set of astronomical data of stars in our Galaxy. We demonstrate that the model
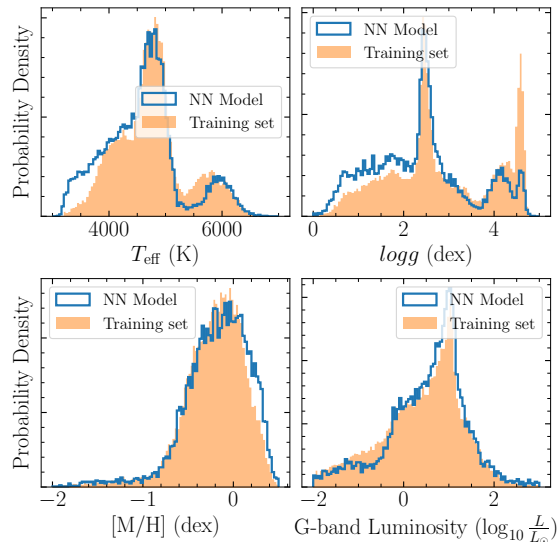
can provide reasonable probability density estimation when given arbitrary combinations of inputs and outputs, hence enhancing the information output. Our technique should be immediately applicable to other foundation models especially in science and with model with higher dimensional output.

2. Motivation

Neural network models trained on discriminative tasks in general do not provide probability density estimation for the outputs, because the learning objective is simply minimizing the difference between the model prediction and the ground truth. There is no mechanism in the model to estimate the probability distribution.

As demonstrated in LB24, it is possible to train a model where one can give observations and known properties of a star, and then request a quantity that you do not know about that star. For example, one can give a spectrum of a star and ask the model for the iron abundance, which is an easy, unambiguous task in astronomy. But when the model is asked to do an ambiguous task (i.e. there are multiple ground truths) like inferring the surface gravity $\log g$ from surface temperature T_{eff} (one can refer to the right panel of Figure 6 for the distribution of stars in this space), the model predicts a mean surface gravity with large
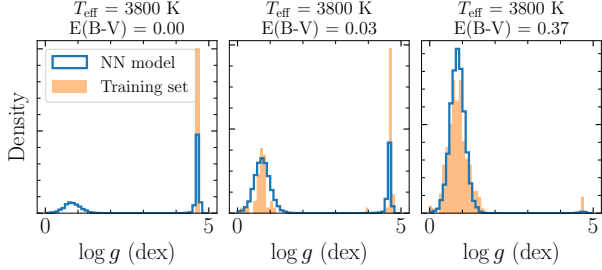
Figure 3. Probability density of surface gravity $\log g$ given different surface temperature T_{eff} and reddening $E(B-V)$ (hence extinction) of stars. All panels show the probability density from the model (solid blue lines) and the training set (orange filled area) based on the condition provided in the panel's title. As reddening increases, a higher proportion of stars are expected to be intrinsically bright giant stars rather than intrinsically dim dwarf stars. This figure demonstrates that our model learns the correct output distribution of various labels when there are only a few conditions leading to an ambiguous answer not captured by traditional Transformer-only models.

uncertainty. While this is correct when assuming Gaussian uncertainty, because there can be a wide range of possible surface gravities, the output is uninformative since the true probability distribution of surface gravity is composed of two widely-separated narrow peaks. This is why it is crucial to have a model that can output non-trivial probabilities while keeping the flexibility of the Transformer.

3. Model Implementation

We adopt a similar architecture, embedding process, and training procedure as in LB24, but modify their approach in a few aspects. Instead of an encoder-decoder Transformer, here we use an encoder-only architecture similar to Google's BERT (Devlin et al., 2018) to reduce of the number of attention mechanisms needed to be performed to reduce the computational cost. Moreover to alleviate the issue of having a limited context size during inference, the encoder-only Transformer architecture used here allows the context size during inference to be much longer than the one we have trained on, because we do not require positional-encoding here. The model accepts a list of data points that consist of the name of the data as well as their magnitude (e.g., $[T_{\text{eff}}, 7000\,\text{K}]$); the first position in the input sequence is always the name of the data being requested without the magnitude.

We use the same "non-linear embedding" process as described in LB24, which embeds data of kind x through

$$y_x = f(w_x \cdot M_x) + w_{b,x} \qquad (1)$$

where y_x is the final vectorized data for a particular kind of data x that will be fed into the encoder-only Transformer.
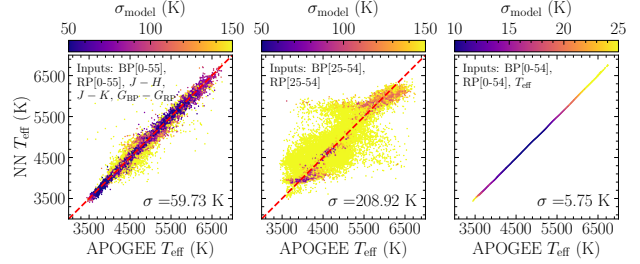

Figure 4. Inference of surface temperature T_{eff} from different combinations of input data from the testing set. NN T_{eff} is the median of the output probability distribution with the robust standard deviation (i.e., $1.4826 \times$ median absolute deviation) of the distribution as the output uncertainty represented by colors. In the left panel, the input contains the whole Gaia XP spectra with colors (113 data point in total which is almost double the 64 context size used during training) and the T_{eff} is accurate with reasonable uncertainty. In the middle panel, the input contains the most uninformative part of the Gaia XP spectra and the model prediction of T_{eff} also suffers with large uncertainty as expected. In the right panel, the same input as the middle panel is given, but in addition T_{eff} is mixed into the inputs, in this case the model predicts almost perfect T_{eff} with very low uncertainty.

The function f is a typical activation function used in neural networks, w_x are the embedding weights, M_x is the magnitude of the data, and $w_{b,x}$ is a bias weight; all of these are particular to the kind of data x.

As shown in Figure 1, the outputs from the Transformer passed to the DDPM head are from the hidden state at the position of the request token (similar to [CLS] in Google's BERT classification token); the rest of the hidden states are not used. The DDPM head de-noises samples from a unit Gaussian distribution where the denoising process is controlled by the hidden state at the position of the request token. During inference, the probability distribution is generated by denoising samples from an unit normal distribution based on the output from the Transformer.

4. Datasets and Training

To illustrate the model, we use the same training and testing data set as in LB24. The data sets are small enough for quick proof-of-concept purposes but big enough for a meaningful foundation model to be fit, and the result can be directly compared to their work. The data set consists of Gaia DR3 photometry and XP spectra (De Angeli et al., 2023; Montegriffo et al., 2023), 2MASS photometry (Skrutskie et al., 2006), and stellar properties derived from high-resolution, high signal-to-noise-ratio near-infrared APOGEE DR17 (Abdurro'uf & et al., 2022) spectra for $\approx 100,000$ stars. The majority of these stars are sub-giants and red giants. Further details of these data are given in Appendix B.
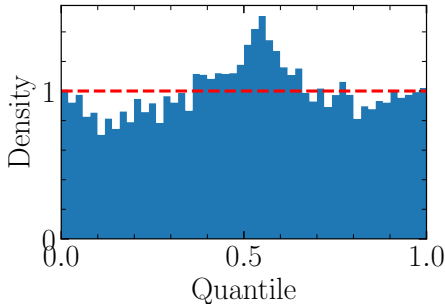
Figure 5. The distribution of the quantile a which the ground truth is found in the probability density distribution inferred from our model. Similar to the training procedure, we randomly select a subset of data as input and a random label as output for each star. The blue colored histogram show the quantile distribution while the red dotted line represents the uniform distribution, which the quantiles should follow if the model probability distributions were exactly correct. The quantile distribution closely follows the red dotted line with a slight over-concentration around the 50% quantile, indicating that we are slightly overestimating the uncertainty in the outputs.

In the data set, each star has a row of available observations in which missing data (represented by `NaN`) are replaced by a special padding token (the same idea used to mask empty spaces in sentences), which is masked automatically in the attention layer. Similar to LB24, we pick a random set of data for each star as input and a random one as output, which may or may not be included in the input already even for stars in the same batch during training. Hyper-parameters used in training are described in Section 5.

5. Training Hyper-parameters

The model used for training consists of total 3.7 million trainable parameters which is slightly less than half of the 8.8 million parameters used in LB2024. This is also many order of magnitudes less than those used in commercial models which have billions of parameters because this is a proof-of-concept. There are 118 of type of data and each embedding we have used is 192-dimensional. The encoder-only Transformer consists of 2 multi-head attention layers where each layers has 24 attention heads such that each head corresponds to 8-dimension in the embedding.

For the DDPM head, there are 1.8 million trainable parameters which is about half of the total 3.7 million parameters for the whole model. There are 120 diffusion steps, that is we generate a new sample of data by denoising 120 times from pure Gaussian noise. A sigmoid noise schedule β is used, which is defined as follow

$$\beta_t = \beta_{\text{start}} + (\beta_{\text{end}} - \beta_{\text{start}}) \cdot \frac{1}{1 + e^{-t}} \qquad (2)$$

where β_{start} and β_{end} are 0.00001 and 0.1, respectively and t are uniformly sampled from -6 to 6. The `AdamW` optimizer is used to train the whole model with an initial learning rate of 0.001 with a `CosineAnnealingWarmRestarts` learning rate scheduler with final learning rate of 1×10^{-10}; each cycle spans 1024 epochs for 10240 epochs in total during training. The training process took place with a single NVIDIA A100SXM4 for about 20 hours.

6. Results

To validate our model, we perform variety of tests in different scenarios. First, we test the model's ability to infer probability density distributions when no condition is provided, which essentially checks if the model can recover the training set distribution (Figure 2). The second test is to see if the model can recover probability density distributions when there are conditions (Figure 3). The third test is to check if the median and standard deviation of the distribution are accurate and provide reasonable uncertainty estimation (Figure 4). The final test checks how good the output probability densities are, by checking that the ground truth is found at uniformly distributed quantiles of the model probability densities (Figure 5).

Overall, our model works reasonably well when we use the probability density distribution to estimate population of stars (see Figure 2 and 3). Our model also works reasonably well when one simply uses the mean as the final prediction and the standard derivation as the output uncertainty as shown in Figure 4, similar to Figure 4 in LB24. The prediction is accurate and the estimated uncertainties are reasonable, with predictions far from the ground truth also being those with large uncertainty. The model also perform inference on arbitrary combinations of input and outputs; Figure 5 shows the quantile of the ground truth within the model predicted probability distribution. The uniformity in this quantile distribution demonstrates that the model provides a reasonable uncertainty in general.

Even though the model presented here generates only one dimensional distributions, the model is capable of providing multi-dimensional distributions by using the flexibility of the Transformer to generate a single dimension at a time sequentially. An example is shown in Figure 6 where the model distribution in the three-dimensional space of surface temperature$-$surface gravity$-$metallicity. We first ask the model to generate the metallicity $[\text{M/H}]$ distribution, then use the $[\text{M/H}]$ samples to generate surface gravity samples, and use pairs of samples to generate surface temperature samples. This procedure allows one to generate probability distributions for *any* combination of output quantities, thus
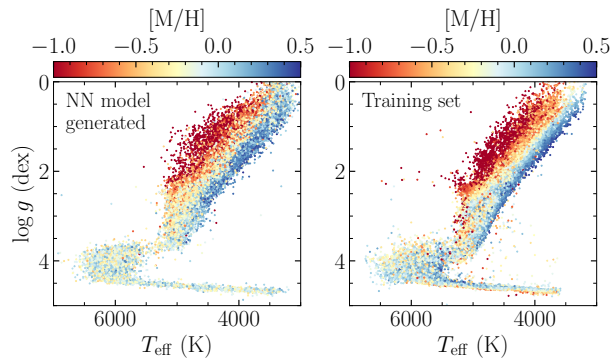
Figure 6. Surface temperature $T_{\rm eff}$ and surface gravity $\log g$ diagram (a.k.a. Kiel Diagram) reconstructed from with our model on the left with the ground-truth distribution on the right. Because our model only output one-dimensional distributions, the diagram was generated by first requesting the distribution of metallicity, then ask for the distribution of $T_{\rm eff}$ for each metallicity sample and finally ask for the distribution of $\log g$ for each pair of $T_{\rm eff}$ and metallicity samples. Our model's reconstruction represents the training set distribution well.

solving the combinatoric problem of learning large numbers of high-dimensional probability distributions.

7. Conclusion

In this work, we have trained an encoder-only Transformer with a denoising diffusion probabilistic model head on top with a small set of astronomical dataset of stars in our Galaxy. We demonstrated that the model can infer probability densities for the outputs reasonably well when the model is asked to recover the training set probability density distribution of various labels or when asking the model for conditional probability density distributions, making this model a very flexible density function emulator. The prediction and uncertainty of the distributions from the model are accurate and reasonable, making the model useful when the user only wants a scalar prediction and a scalar uncertainty instead of a whole distribution. The approach we demonstrate here can be used in other scientific foundation models in regression settings to make the model's outputs more informative.

Software and Data

Python is used in this work and all packages used here are open-source software include `pytorch`, `numpy`, `matplotlib`, `pandas`, `h5py`, `tqdm`, `astropy`, `mwdust`, `GaiaXPy`, `gaiadr3_zeropoint`, `MyGaiaDB` & `mwdust`. All astronomical data used in this work are public from the Sloan Digital Sky Survey and European Space Agency Gaia mission.

The code for this work is available at `https://github.com/henrysky/stars_foundation_diffusion`.

Acknowledgements

Impact Statement

The work presented here can be applied to many other field of science. This paper presents work whose goal is to advance Foundation model in scientific fields. There are minimal societal consequences and ethical concerns of our work.

References

Abdurro'uf and et al. The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data. ApJS, 259(2):

35, April 2022. doi: 10.3847/1538-4365/ac4414.

Blanton, M. R., Bershady, M. A., Abolfathi, B., Albareti, F. D., Allende Prieto, C., Almeida, A., Alonso-García, J., Anders, F., Anderson, S. F., Andrews, B., Aquino-Ortíz, E., Aragón-Salamanca, A., Argudo-Fernández, M., Armengaud, E., Aubourg, E., Avila-Reese, V., Badenes, C., Bailey, S., Barger, K. A., Barrera-Ballesteros, J., Bartosz, C., Bates, D., Baumgarten, F., Bautista, J., Beaton, R., Beers, T. C., Belfiore, F., Bender, C. F., Berlind, A. A., Bernardi, M., Beutler, F., Bird, J. C., Bizyaev, D., Blanc, G. A., Blomqvist, M., Bolton, A. S., Boquien, M., Borissova, J., van den Bosch, R., Bovy, J., Brandt, W. N., Brinkmann, J., Brownstein, J. R., Bundy, K., Burgasser, A. J., Burtin, E., Busca, N. G., Cappellari, M., Delgado Carigi, M. L., Carlberg, J. K., Carnero Rosell, A., Carrera, R., Chanover, N. J., Cherinka, B., Cheung, E., Gómez Maqueo Chew, Y., Chiappini, C., Choi, P. D., Chojnowski, D., Chuang, C.-H., Chung, H., Cirolini, R. F., Clerc, N., Cohen, R. E., Comparat, J., da Costa, L., Cousinou, M.-C., Covey, K., Crane, J. D., Croft, R. A. C., Cruz-Gonzalez, I., Garrido Cuadra, D., Cunha, K., Damke, G. J., Darling, J., Davies, R., Dawson, K., de la Macorra, A., Dell'Agli, F., De Lee, N., Delubac, T., Di Mille, F., Diamond-Stanic, A., Cano-Díaz, M., Donor, J., Downes, J. J., Drory, N., du Mas des Bourboux, H., Duckworth, C. J., Dwelly, T., Dyer, J., Ebelke, G., Eigenbrot, A. D., Eisenstein, D. J., Emsellem, E., Eracleous, M., Escoffier, S., Evans, M. L., Fan, X., Fernández-Alvar, E., Fernandez-Trincado, J. G., Feuillet, D. K., Finoguenov, A., Fleming, S. W., Font-Ribera, A., Fredrickson, A., Freischlad, G., Frinchaboy, P. M., Fuentes, C. E., Galbany, L., Garcia-Dias, R., García-Hernández, D. A., Gaulme, P., Geisler, D., Gelfand, J. D., Gil-Marín, H., Gillespie, B. A., Goddard, D., Gonzalez-Perez, V., Grabowski, K., Green, P. J., Grier, C. J., Gunn, J. E., Guo, H., Guy, J., Hagen, A., Hahn, C., Hall, M., Harding, P., Hasselquist, S., Hawley, S. L., Hearty, F., Gonzalez Hernández, J. I., Ho, S., Hogg, D. W., Holley-Bockelmann, K., Holtzman, J. A., Holzer, P. H., Huehnerhoff, J., Hutchinson, T. A., Hwang, H. S., Ibarra-Medel, H. J., da Silva Ilha, G., Ivans, I. I., Ivory, K., Jackson, K., Jensen, T. W., Johnson, J. A., Jones, A., Jönsson, H., Jullo, E., Kamble, V., Kinemuchi, K., Kirkby, D., Kitaura, F.-S., Klaene, M., Knapp, G. R., Kneib, J.-P., Kollmeier, J. A., Lacerna, I., Lane, R. R., Lang, D., Law, D. R., Lazarz, D., Lee, Y., Le Goff, J.-M., Liang, F.-H., Li, C., Li, H., Lian, J., Lima, M., Lin, L., Lin, Y.-T., Bertran de Lis, S., Liu, C., de Icaza Lizaola, M. A. C., Long, D., Lucatello, S., Lundgren, B., MacDonald, N. K., Deconto Machado, A., MacLeod, C. L., Mahadevan, S., Geimba Maia, M. A., Maiolino, R., Majewski, S. R., Malanushenko, E., Malanushenko, V., Manchado, A., Mao, S., Maraston, C., Marques-Chaves, R., Masseron, T., Masters, K. L., McBride, C. K., Mc-Dermid, R. M., McGrath, B., McGreer, I. D., Medina Peña, N., Melendez, M., Merloni, A., Merrifield, M. R., Meszaros, S., Meza, A., Minchev, I., Minniti, D., Miyaji, T., More, S., Mulchaey, J., Müller-Sánchez, F., Muna, D., Munoz, R. R., Myers, A. D., Nair, P., Nandra, K., Correa do Nascimento, J., Negrete, A., Ness, M., Newman, J. A., Nichol, R. C., Nidever, D. L., Nitschelm, C., Ntelis, P., O'Connell, J. E., Oelkers, R. J., Oravetz, A., Oravetz, D., Pace, Z., Padilla, N., Palanque-Delabrouille, N., Alonso Palicio, P., Pan, K., Parejko, J. K., Parikh, T., Pâris, I., Park, C., Patten, A. Y., Peirani, S., Pellejero-Ibanez, M., Penny, S., Percival, W. J., Perez-Fournon, I., Petitjean, P., Pieri, M. M., Pinsonneault, M., Pisani, A., Poleski, R., Prada, F., Prakash, A., Queiroz, A. B. d. A., Raddick, M. J., Raichoor, A., Barboza Rembold, S., Richstein, H., Riffel, R. A., Riffel, R., Rix, H.-W., Robin, A. C., Rockosi, C. M., Rodríguez-Torres, S., Roman-Lopes, A., Román-Zúñiga, C., Rosado, M., Ross, A. J., Rossi, G., Ruan, J., Ruggeri, R., Rykoff, E. S., Salazar-Albornoz, S., Salvato, M., Sánchez, A. G., Aguado, D. S., Sánchez-Gallego, J. R., Santana, F. A., Santiago, B. X., Sayres, C., Schiavon, R. P., da Silva Schimoia, J., Schlafly, E. F., Schlegel, D. J., Schneider, D. P., Schultheis, M., Schuster, W. J., Schwope, A., Seo, H.-J., Shao, Z., Shen, S., Shetrone, M., Shull, M., Simon, J. D., Skinner, D., Skrutskie, M. F., Slosar, A., Smith, V. V., Sobeck, J. S., Sobreira, F., Somers, G., Souto, D., Stark, D. V., Stassun, K., Stauffer, F., Steinmetz, M., Storchi-Bergmann, T., Streblyanska, A., Stringfellow, G. S., Suárez, G., Sun, J., Suzuki, N., Szigeti, L., Taghizadeh-Popp, M., Tang, B., Tao, C., Tayar, J., Tembe, M., Teske, J., Thakar, A. R., Thomas, D., Thompson, B. A., Tinker, J. L., Tissera, P., Tojeiro, R., Hernandez Toledo, H., de la Torre, S., Tremonti, C., Troup, N. W., Valenzuela, O., Martinez Valpuesta, I., Vargas-González, J., Vargas-Magaña, M., Vazquez, J. A., Villanova, S., Vivek, M., Vogt, N., Wake, D., Walterbos, R., Wang, Y., Weaver, B. A., Weijmans, A.-M., Weinberg, D. H., Westfall, K. B., Whelan, D. G., Wild, V., Wilson, J., Wood-Vasey, W. M., Wylezalek, D., Xiao, T., Yan, R., Yang, M., Ybarra, J. E., Yèche, C., Zakamska, N., Zamora, O., Zarrouk, P., Zasowski, G., Zhang, K., Zhao, G.-B., Zheng, Z., Zheng, Z., Zhou, X., Zhou, Z.-M., Zhu, G. B., Zoccali, M., and Zou, H. Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. AJ, 154(1):28, July 2017. doi: 10.3847/1538-3881/aa7567.

Bovy, J., Rix, H.-W., Green, G. M., Schlafly, E. F., and Finkbeiner, D. P. On Galactic Density Modeling in the Presence of Dust Extinction. ApJ, 818(2):130, February 2016. doi: 10.3847/0004-637X/818/2/130.

Carrasco, J. M., Weiler, M., Jordi, C., Fabricius, C., De Angeli, F., Evans, D. W., van Leeuwen, F., Riello, M., and Montegriffo, P. Internal calibration of Gaia BP/RP

low-resolution spectra. A&A, 652:A86, August 2021. doi: 10.1051/0004-6361/202141249.

De Angeli, F., Weiler, M., Montegriffo, P., Evans, D. W., Riello, M., Andrae, R., Carrasco, J. M., Busso, G., Burgess, P. W., Cacciari, C., Davidson, M., Harrison, D. L., Hodgkin, S. T., Jordi, C., Osborne, P. J., Pancino, E., Altavilla, G., Barstow, M. A., Bailer-Jones, C. A. L., Bellazzini, M., Brown, A. G. A., Castellani, M., Cowell, S., Delchambre, L., De Luise, F., Diener, C., Fabricius, C., Fouesneau, M., Frémat, Y., Gilmore, G., Giuffrida, G., Hambly, N. C., Hidalgo, S., Holland, G., Kostrzewa-Rutkowska, Z., van Leeuwen, F., Lobel, A., Marinoni, S., Miller, N., Pagani, C., Palaversa, L., Piersimoni, A. M., Pulone, L., Ragaini, S., Rainer, M., Richards, P. J., Rixon, G. T., Ruz-Mieres, D., Sanna, N., Sarro, L. M., Rowell, N., Sordo, R., Walton, N. A., and Yoldas, A. Gaia Data Release 3. Processing and validation of BP/RP low-resolution spectral data. A&A, 674:A2, June 2023. doi: 10.1051/0004-6361/202243680.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, art. arXiv:1810.04805, October 2018. doi: 10.48550/arXiv.1810.04805.

Dey, A., Schlegel, D. J., Lang, D., Blum, R., Burleigh, K., Fan, X., Findlay, J. R., Finkbeiner, D., Herrera, D., Juneau, S., Landriau, M., Levi, M., McGreer, I., Meisner, A., Myers, A. D., Moustakas, J., Nugent, P., Patej, A., Schlafly, E. F., Walker, A. R., Valdes, F., Weaver, B. A., Yèche, C., Zou, H., Zhou, X., Abareshi, B., Abbott, T. M. C., Abolfathi, B., Aguilera, C., Alam, S., Allen, L., Alvarez, A., Annis, J., Ansarinejad, B., Aubert, M., Beechert, J., Bell, E. F., BenZvi, S. Y., Beutler, F., Bielby, R. M., Bolton, A. S., Briceño, C., Buckley-Geer, E. J., Butler, K., Calamida, A., Carlberg, R. G., Carter, P., Casas, R., Castander, F. J., Choi, Y., Comparat, J., Cukanovaite, E., Delubac, T., DeVries, K., Dey, S., Dhungana, G., Dickinson, M., Ding, Z., Donaldson, J. B., Duan, Y., Duckworth, C. J., Eftekharzadeh, S., Eisenstein, D. J., Etourneau, T., Fagrelius, P. A., Farihi, J., Fitzpatrick, M., Font-Ribera, A., Fulmer, L., Gänsicke, B. T., Gaztanaga, E., George, K., Gerdes, D. W., Gontcho, S. G. A., Gorgoni, C., Green, G., Guy, J., Harmer, D., Hernandez, M., Honscheid, K., Huang, L. W., James, D. J., Jannuzi, B. T., Jiang, L., Joyce, R., Karcher, A., Karkar, S., Kehoe, R., Kneib, J.-P., Kueter-Young, A., Lan, T.-W., Lauer, T. R., Le Guillou, L., Le Van Suu, A., Lee, J. H., Lesser, M., Perreault Levasseur, L., Li, T. S., Mann, J. L., Marshall, R., Martínez-Vázquez, C. E., Martini, P., du Mas des Bourboux, H., McManus, S., Meier, T. G., Ménard, B., Metcalfe, N., Muñoz-Gutiérrez, A., Najita, J., Napier, K., Narayan, G., Newman, J. A., Nie, J., Nord, B., Norman, D. J., Olsen, K. A. G., Paat, A., Palanque-Delabrouille, N., Peng, X., Poppett, C. L., Poremba, M. R., Prakash, A., Rabinowitz, D., Raichoor, A., Rezaie, M., Robertson, A. N., Roe, N. A., Ross, A. J., Ross, N. P., Rudnick, G., Safonova, S., Saha, A., Sánchez, F. J., Savary, E., Schweiker, H., Scott, A., Seo, H.-J., Shan, H., Silva, D. R., Slepian, Z., Soto, C., Sprayberry, D., Staten, R., Stillman, C. M., Stupak, R. J., Summers, D. L., Sien Tie, S., Tirado, H., Vargas-Magaña, M., Vivas, A. K., Wechsler, R. H., Williams, D., Yang, J., Yang, Q., Yapici, T., Zaritsky, D., Zenteno, A., Zhang, K., Zhang, T., Zhou, R., and Zhou, Z. Overview of the DESI Legacy Imaging Surveys. AJ, 157(5):168, May 2019. doi: 10.3847/1538-3881/ab089d.

García Pérez, A. E., Allende Prieto, C., Holtzman, J. A., Shetrone, M., Mészáros, S., Bizyaev, D., Carrera, R., Cunha, K., García-Hernández, D. A., Johnson, J. A., Majewski, S. R., Nidever, D. L., Schiavon, R. P., Shane, N., Smith, V. V., Sobeck, J., Troup, N., Zamora, O., Weinberg, D. H., Bovy, J., Eisenstein, D. J., Feuillet, D., Frinchaboy, P. M., Hayden, M. R., Hearty, F. R., Nguyen, D. C., O'Connell, R. W., Pinsonneault, M. H., Wilson, J. C., and Zasowski, G. ASPCAP: The APOGEE Stellar Parameter and Chemical Abundances Pipeline. AJ, 151 (6):144, June 2016. doi: 10.3847/0004-6256/151/6/144.

Gloeckler, M., Deistler, M., Weilbach, C., Wood, F., and Macke, J. H. All-in-one simulation-based inference. *arXiv e-prints*, art. arXiv:2404.09636, April 2024. doi: 10.48550/arXiv.2404.09636.

Ho, J., Jain, A., and Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv e-prints*, art. arXiv:2006.11239, June 2020. doi: 10.48550/arXiv.2006.11239.

Ivezić, Ž. et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. ApJ, 873(2):111, March 2019. doi: 10.3847/1538-4357/ab042c.

Jimenez Rezende, D. and Mohamed, S. Variational Inference with Normalizing Flows. *arXiv e-prints*, art. arXiv:1505.05770, May 2015. doi: 10.48550/arXiv.1505.05770.

Kapur, S., Jenner, E., and Russell, S. Diffusion On Syntax Trees For Program Synthesis. *arXiv e-prints*, art. arXiv:2405.20519, May 2024. doi: 10.48550/arXiv.2405.20519.

Kollmeier, J. A., Zasowski, G., Rix, H.-W., Johns, M., Anderson, S. F., Drory, N., Johnson, J. A., Pogge, R. W., Bird, J. C., Blanc, G. A., Brownstein, J. R., Crane, J. D., De Lee, N. M., Klaene, M. A., Kreckel, K., MacDonald, N., Merloni, A., Ness, M. K., O'Brien, T., Sanchez-Gallego, J. R., Sayres, C. C., Shen, Y., Thakar, A. R., Tkachenko, A., Aerts, C., Blanton, M. R., Eisenstein,

D. J., Holtzman, J. A., Maoz, D., Nandra, K., Rockosi, C., Weinberg, D. H., Bovy, J., Casey, A. R., Chaname, J., Clerc, N., Conroy, C., Eracleous, M., Gänsicke, B. T., Hekker, S., Horne, K., Kauffmann, J., McQuinn, K. B. W., Pellegrini, E. W., Schinnerer, E., Schlafly, E. F., Schwope, A. D., Seibert, M., Teske, J. K., and van Saders, J. L. SDSS-V: Pioneering Panoptic Spectroscopy. *arXiv e-prints*, art. arXiv:1711.03234, November 2017. doi: 10.48550/arXiv.1711.03234.

Leung, H. W. and Bovy, J. Towards an astronomical foundation model for stars with a transformer-based model. MNRAS, 527(1):1494–1520, January 2024. doi: 10.1093/mnras/stad3015.

McCabe, M., Régaldo-Saint Blancard, B., Holden Parker, L., Ohana, R., Cranmer, M., Bietti, A., Eickenberg, M., Golkar, S., Krawezik, G., Lanusse, F., Pettee, M., Tesileanu, T., Cho, K., and Ho, S. Multiple Physics Pretraining for Physical Surrogate Models. *arXiv e-prints*, art. arXiv:2310.02994, October 2023. doi: 10.48550/arXiv.2310.02994.

Montegriffo, P., De Angeli, F., Andrae, R., Riello, M., Pancino, E., Sanna, N., Bellazzini, M., Evans, D. W., Carrasco, J. M., Sordo, R., Busso, G., Cacciari, C., Jordi, C., van Leeuwen, F., Vallenari, A., Altavilla, G., Barstow, M. A., Brown, A. G. A., Burgess, P. W., Castellani, M., Cowell, S., Davidson, M., De Luise, F., Delchambre, L., Diener, C., Fabricius, C., Frémat, Y., Fouesneau, M., Gilmore, G., Giuffrida, G., Hambly, N. C., Harrison, D. L., Hidalgo, S., Hodgkin, S. T., Holland, G., Marinoni, S., Osborne, P. J., Pagani, C., Palaversa, L., Piersimoni, A. M., Pulone, L., Ragaini, S., Rainer, M., Richards, P. J., Rowell, N., Ruz-Mieres, D., Sarro, L. M., Walton, N. A., and Yoldas, A. Gaia Data Release 3. External calibration of BP/RP low-resolution spectroscopic data. A&A, 674: A3, June 2023. doi: 10.1051/0004-6361/202243880.

Pace, R. K. and Barry, R. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, May 1997. URL http://www.sciencedirect.com/science/article/B6V1D-3WNMX3B-B/1/4baafcb4328934d470158b0233c44102.

Peebles, W. and Xie, S. Scalable Diffusion Models with Transformers. *arXiv e-prints*, art. arXiv:2212.09748, December 2022. doi: 10.48550/arXiv.2212.09748.

Skrutskie, M. F., Cutri, R. M., Stiening, R., Weinberg, M. D., Schneider, S., Carpenter, J. M., Beichman, C., Capps, R., Chester, T., Elias, J., Huchra, J., Liebert, J., Lonsdale, C., Monet, D. G., Price, S., Seitzer, P., Jarrett, T., Kirkpatrick, J. D., Gizis, J. E., Howard, E., Evans, T., Fowler, J., Fullmer, L., Hurt, R., Light, R., Kopan, E. L., Marsh,

K. A., McCallon, H. L., Tam, R., Van Dyk, S., and Wheelock, S. The Two Micron All Sky Survey (2MASS). AJ, 131(2):1163–1183, February 2006. doi: 10.1086/498708.

Walmsley, M., Bowles, M., Scaife, A. M. M., Shingirai Makechemu, J., Gordon, A. J., Ferguson, A. M. N., Mann, R. G., Pearson, J., Popp, J. J., Bovy, J., Speagle, J., Dickinson, H., Fortson, L., Géron, T., Kruk, S., Lintott, C. J., Mantha, K., Mohan, D., O'Ryan, D., and Slijepevic, I. V. Scaling Laws for Galaxy Images. *arXiv e-prints*, art. arXiv:2404.02973, April 2024. doi: 10.48550/arXiv.2404.02973.

Probability Density of the California Housing Dataset learnt by Transformer+DDPM
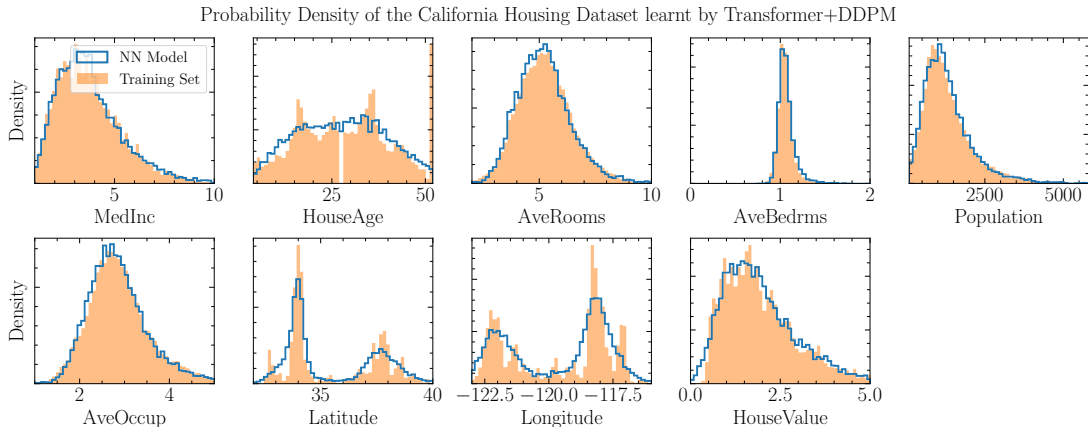
Figure 7. Probability density of the dataset based on no conditions (i.e., request values based only on padding). The orange filled areas are the density of the training set while the blue lines show the model prediction. In general, the predictions closely follow the training set distributions.

A. Demonstration with California Housing Data

To demonstrate that our method works in a simple setting with a standard dataset, we have trained our model on the much simpler California Housing dataset (Pace & Barry, 1997) with only 20,640 data points and 9 features included in `scikit-learn`. We treat this dataset as tabular data and follow the same training procedure as in the main text (i.e., randomly select a subset of features as input and one feature as output during training for each house in each batch). The model consists of 47,425 trainable parameters including the 16-dimensional embeddings where 21,377 of those parameters comes from the DDPM head.

Figure 7 demonstrates that the model successfully predicts the distribution of all 9 features in the training set and Figure 8 shows that the model also can provide conditional probability distributions. In both cases, the distributions look reasonable and follow the training set.

B. Astronomical Datasets

Gaia XP spectra are low-resolution optical to near-infrared ($R \sim 30$ to 100, 330 to $1050\,\mathrm{nm}$; Carrasco et al. 2021) spectra obtained from the Blue and Red Photometers (BP/RP) aboard the Gaia spacecraft. Unlike usual higher resolution stellar spectra, the Gaia XP spectra were released as 110 coefficients of an orthogonal basis function expansion, where lower-order coefficients explain large-scale features of the spectra (hence more information like T_eff) and higher-order coefficients explain small-scale features including the noise. We simply treat each coefficient as a kind of data. We normalize the XP coefficients by the Gaia G-band apparent flux.

Stellar parameter labels like the surface temperature T_eff, surface gravity $\log g$, and the overall metallicity [M/H] are from the SDSS APOGEE data release 17 (Blanton et al., 2017; Abdurro'uf & et al., 2022). APOGEE is a high-resolution ($R \sim 22,000$), high signal-to-noise (100 per pixel typically) spectroscopic survey in both northern and southern hemisphere in the near infrared H-band wavelength region of $1.5 - 1.7\mu m$. The stellar parameter labels are derived from APOGEE spectra using the APOGEE Stellar Parameter and Chemical Abundances Pipeline (García Pérez et al., 2016).

Photometry J, H, K band in infrared are from Two Micron All Sky Survey (Skrutskie et al., 2006) and G_BP, G_RP from Gaia spacecraft, $J - H$, $J - K$ and $G_\mathrm{BP} - G_\mathrm{RP}$ are simply the difference between these photometric bands. Reddening $E(B-V)$ are calculated with a three-dimensional extinction map `Combined19` (Bovy et al., 2016) based on three-dimensional position (i.e. two-dimensional coordinates on the sky as well as distance) of stars relative to Earth.
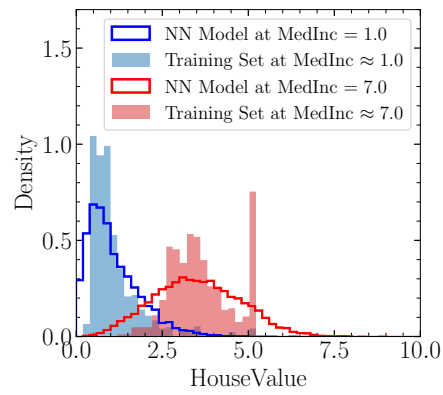
Figure 8. Conditional probability density of house value when providing the owner's median income. The colored areas are the density of the training set while the solid colored lines show the model prediction. People with higher median income tend to own a more expensive house as expected.