# 3D-PDR Orion dataset and NeuralPDR: Neural Differential Equations for Photodissociation Regions

**Gijs Vermariën,*** **Rahul Ravichandran & Serena Viti**

Leiden Observatory
Leiden University
PO Box 9513, 2300 RA Leiden, The Netherlands

**Thomas G. Bisbas**

Research Center for
Astronomical Computing
Zhejiang Lab
Hangzhou 311100, China

## Abstract

We present a novel dataset of simulations of the photodissociation region (PDR) in the Orion Bar and provide benchmarks of emulators for the dataset. Numerical models of PDRs are computationally expensive since the modeling of these changing regions requires resolving the thermal balance and chemical composition along a line-of-sight into an interstellar cloud. This often makes it a bottleneck for 3D simulations of these regions. In this work, we provide a dataset of 8192 models with different initial conditions simulated with 3D-PDR. We then benchmark different architectures, focusing on Augmented Neural Ordinary Differential Equation (ANODE) based models [2]. Obtaining fast and robust emulators that can be included as preconditioners of classical codes or full emulators into 3D simulations of PDRs.

## 1 Introduction

At the edges of interstellar clouds there exist PDR regions that are dominated by UV photon chemistry. Computational models of these regions are expensive, since both the physical conditions and chemical composition of these regions change as we move deeper into the cloud. In order to simulate this accurately, we need an iterative method that solves for heating, cooling and chemistry, creating a visual extinction based problem. The visual extinction is a measure for the decrease in radiation as we move into an astronomical object, and is related to the amount of hydrogen along a line of sight [Güver and Özel, 2009]. By solving the differential equations describing these processes along a line of sight, we obtain the temperatures and abundances. Having to solve along lines of sight makes comprehensive 3D simulations of these regions computationally expensive, raising the need for surrogates that provide approximations of the chemistry, freeing up budget for hydrodynamics and radiative transport.

In this article we provide a dataset inspired by the Orion Bar, a region within the Orion nebulae, that is currently being thoroughly investigated with the James Webb Space Telescope [Andree-Labsch et al., 2017, Habart et al., 2023, Peeters et al., 2024]. The dataset is thus derived from typical conditions in the Orion Bar, with radiation fields $10^1 \leq G_{\rm UV}(\rm Draine) \leq 10^4$, number densities $10^2 \leq n_{\rm H}(\rm cm^{-3}) \leq 10^7$ and lastly the cosmic ray ionisations $10^{-17} \leq \zeta(\rm s^{-1}) \leq 10^{-15}$. We use these parameters as initial condition for the simulation with 3D-PDR, the resulting 8192 models together we call the 3D-PDR Orion dataset[3].[Vermariën and Viti, 2024].

*vermarien@strw.leidenuniv.nl

[2]These models can be found at `https://github.com/uclchem/neuralpdr`.

[3]The 3D-PDR Orion dataset is available at `https://doi.org/10.5281/zenodo.13711174`

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

So far several machine learning based methods have been proposed to tackle problems in astrochemistry. Emulators have been trained to predict equilibrium chemistry [de Mijolla et al., 2019], time series based chemistry [Grassi et al., 2021, Holdship et al., 2021, Tang and Turk, 2022, Branca and Pallottini, 2022, Sulzer and Buck, 2023, Branca and Pallottini, 2024] and position based chemistry [Maes et al., 2024]. But most of these have limited accuracy, only generalize to a small part of the physical parameter space or fail to include the interaction between temperature and chemistry.

We then focus on benchmarking surrogate models (emulator) that work for large physical parameter spaces and are effective at emulating a selection of molecules that have complex formation pathways, without having to include the entire chemical network. To this end, we use Augmented Neural Ordinary equations (ANODE), where the augmented part of the NODE are auxiliary features that relate to the physical conditions of 3D-PDR models.

## 2 Methods

### 2.1 Augmented Neural Ordinary Differential Equations with auxiliary parameters

The concept of the Neural Ordinary Differential Equations [Chen et al., 2019, Kidger, 2022] is that instead of defining the right hand side (RHS) of an ordinary differential equation with expert chemistry and physics knowledge, we instead define an approximator $\tilde{f}$. In this case we use a neural network, that uses data to generate a nonlinear RHS that can approximate the series. This provides an integral of the form:

$$y(x_2) = y(x_1) + \int_{x_1}^{x_2} \tilde{f}(x, y, e)\mathrm{d}x \tag{1}$$

where $x$ is the independent variable, $y$ the dependent variables and $e$ auxiliary parameters. The addition of auxiliary parameters $e$, allows us to train a model that generalizes over many different physical models with different physical parameters. The usage of parameters to find more expressive NeuralODEs has been coined as augmented ODEs [Dupont et al., 2019] and parameterized ODEs [Lee and Parish], in this article we employ the term "auxiliary parameters" distinguish them from the physical parameters of the dataset and prevent confusion. Consequently, the acronym of the employed architecture, ANODE, still fits.

These neural differential equations can be combined with encoder and decoder models [Kramer, 1991], allowing one to construct a lower dimensional latent ODE, which can be solved at, typically, a lower cost [Grathwohl et al., 2018, Rubanova et al.]. This latent ODE can be defined by a small dummy chemical network [Grassi et al., 2021], constant terms [Sulzer and Buck, 2023] or a tensor expression akin to a larger chemical network [Maes et al., 2024].

### 2.2 The 3D-PDR Orion dataset

The 3D-PDR code [Bisbas et al., 2012] is a flexible code that can simulate photodissociation regions in both 1D and 3D. For the purpose of this paper, we generated a dataset of 1D models, each with a different initial condition of external radiation field $G_{\mathrm{UV},0}$, density $n_{\mathrm{H},0}$ and cosmic ray ionisation $\zeta_0$, inspired by the Orion cloud. We call this physical parameter space $P \in \mathbb{R}^3$. From this physical parameter space, we generate a total of 8192 models using Sobol sampling; these models are then divided into a training, validation and test set with a $0.70$, $0.15$ and $0.15$ split respectively.

For each sample in $p \in P$, we obtain one series consisting of 215 relative abundances $x_i(A_V) = n_i(A_V)/n_{\mathrm{H},0}$ as a function of 300 monotonously increasing visual extinctions ($A_V$), illustrated in appendix A. This gives us a series $X_p \in \mathbb{R}^{300\times215}$ for each point in the physical parameter space, $p \in P$, where each vector of abundances is linked to one visual "depth". Besides the abundances of the molecules, 3D-PDR provides us with outputs such as the extinction $A_V$, gas temperature $T_{\mathrm{gas}}(A_V)$, dust temperature $T_{\mathrm{dust}}$ $(A_V)$, $G_{\mathrm{UV}}$, $n_{\mathrm{H},0}$ and $\zeta_0$, giving the auxiliary parameters $E_p \in \mathbb{R}^{300\times6}$, with the last two parameters constants. We combine these aforementioned series into the dataset by picking the abundances of: $e^-$, H, $H_2$, O, CO, $H_2O$, CH, $O_2$, CN, $HCO^+$, NH, HCN, $C_2$, HCO, $H_2CO$, $CO^+$, CS and $CH_3OH$, adding the auxiliary parameters $E_p$ and two of the constant physical parameters, namely $n_{\mathrm{H},0}$and $\zeta_0$. This results in the complete dataset, with 8192 series of $D \in \mathbb{R}^{300\times25}$.

In order to prevent extremely small values from creating a too large dynamic range in log-space, we add a minor offset to each of the features: $\epsilon = 10^{-20}$ for the abundances $X_p$ and the auxiliary

parameters $E_p$, and $\epsilon = 10^{-10}$ for the visual extinction $A_V$. Subsequently, we apply a $\log_{10}$ transformation. Lastly, we standardize the data using the mean $\tilde{\mu}$ and standard deviation $\tilde{\sigma}$ for each of the log transformed features individually. The data transformation combined is:

$$D_i' = \frac{\log_{10}(D_i + \epsilon_i) - \tilde{\mu}}{\tilde{\sigma}} \tag{2}$$

### 2.3 On training neural differential equations

In this paper, we investigate two different types of architectures. Firstly the we explore the direct application of ANODE on the data, mapping from $D_i$ directly to the next visual extinction $D_{i+1}$, hereafter called the `evolve (e)` model. The second architecture uses an additional encoder and decoder, adapted from the encoder-decoder with bottleneck and latent ODE model from Sulzer and Buck [2023], which applies the same mapping but now with the ANODE in the latent space, hereafter called the `encoder-evolve-decode (eed)` model. To the latter, we also add the auxiliary parameters $E_p$ into the encoder with a latent bottleneck of size 5 (`eed-a`), adding them in the latent space giving it size 5+4 (`eed-b`) and adding them into the encoder but enlarging the bottleneck size to 9 (`eed-c`). All these models can be found at `https://github.com/uclchem/neuralpdr`.

In order to effectively train the evolve models, we use several methods. Firstly we use weight decay [Loshchilov and Hutter, 2017] to penalize large weights in the model, since they can lead to expensive to evaluate and instable differential equations. Additionally we initialize the weights for the ANODE with a truncated normal distribution and scale it with a factor $\sqrt{b/n}$ [He et al., 2015] with $b$ a hyperparameter and $n$ the width of the inputs into the layer. This ensures the RHS of the differential starts out small and improves training. Additionally, we use a learning rate scheduler, namely a cosine delay schedule with a linear warmup phase [Loshchilov and Hutter, 2016].We also propose to introduce the series in fractions, first training on the first part, then adding the second parts and so forth. This is combined with the learning rate scheduler, restarting the cosine delay schedule each time after introducing a new fraction.

In summary for the direct models, we choose, after hyperparameter tuning, a neural network of 592 neurons wide, 3 layers deep, a softplus activation function between all layers, a peak learning rate of $1.1 \times 10^{-3}$, a weight scale $b = 1.3$ with no truncation between -10 and 10, a weight decay of $1.2 \times 10^{-4}$, a batch size of 48 and the dataset split into three equal fractions, with each fraction being expanded after 50 epochs, with the full dataset being trained on for an extra 50 epochs, resulting in 200 epochs in total. We then perform an ablation study to compare the complete model with all these features, model (`e-a`), to a model without the weight decay (`e-b`), a model without the learning rate schedule (`e-c`), a model with normal weight initialization (`e-d`), a model without introducing the fractions of training data (`e-e`), and lastly a model with none of the above (`e-f`).

## 3 Experiments and results

We find that both the encoder-evolve-decoder and the evolve models can emulate 3D-PDR in a satisfactory fashion. We firstly investigate the MSE over all test samples per visual extinction index in figure 1. This figure shows that the first four evolve models (`e-a` through `e-d`) perform best in general, with every single ablation not performing significantly worse, (`e-f`) with all features ablated, shows a significantly worse performance. It is not clear which of the improvements to the training process we proposed is exactly responsible for the improvement in performance, but altogether they result in a well trained model. The family of `eed` models does not perform as well as the `evolve` models for most features, which could be attributed to the small bottleneck size. Only for the $A_V$ , the `eed` models do better, which is counter-intuitive since the $A_V$ prediction is reused as an input feature for the next evolution of the model, but apparently so, the inaccurate $A_V$ predictions are sufficient.

We show the predictions for a selection molecules, the temperatures and the auxiliary features inferred with the best model, (`e-a`) in figure 2, with all individual series in appendix C. The model shows good agreement for most features, with the molecules at high abundances hard to discern from the original data. Especially the high abundance, and therefore easily observable by telescope, molecules show a good agreement between the data and the predictions. This is also well within the underlying uncertainties of the underlying parameters as specified in the chemical databases [Millar et al., 2024]. The molecules at low abundances have many fluctuations in the data, the neuralODE however fits a

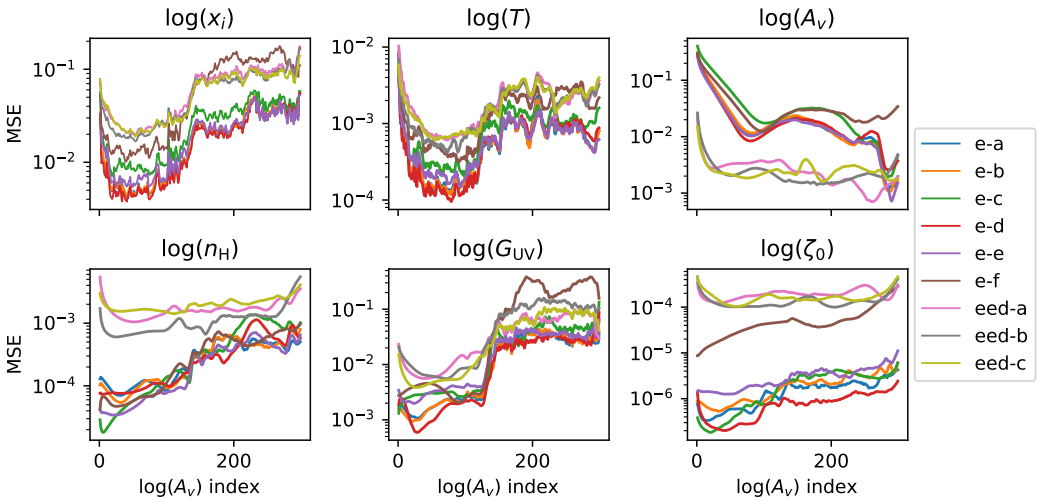


Figure 1: Comparison between different NeuralPDR configurations, the y-axis is the index of the visual extinction. The loss of the `a,b,c` and `e evolve` models largely overlap.

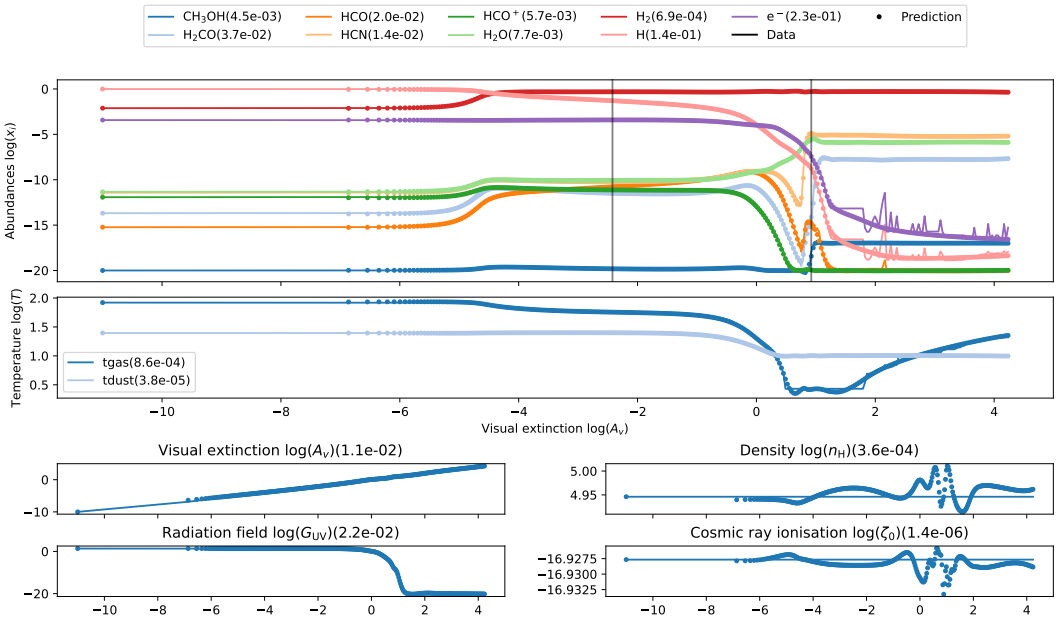


Figure 2: The result for one sample as inferred by the (`e-a`) emulator as a function of the log-$A_V$. The MSE in the feature space for this sample is listed in parenthesis behind each feature.

smooth function to it, providing a desirable smooth interpolation in this regime. This is preferred to overfitting the jumps in the data since they are artifacts of the 3D-PDR solver, caused by the iterative nature of solving the temperature and chemistry balance. The temperature is fitted well, with the gas temperature having a small deviation at the point where the trend in temperature changes rapidly. Both the changing and constant auxiliary features are reproduced well, with the constant features having the lowest errors of all.

The original dataset of 8192 samples was generated on an Intel(R) Core(TM) i9-13900 CPU system with 16 cores, taking 432 hours. The training happens on a system with a NVIDIA RTX 2080, taking 2.5 hours for 200 epochs. Inference then takes 6 seconds for 1228 samples including loading, compiling and saving the data. This critical speed up is exactly what we need for simulating PDRs.

## 4 Conclusions

In this paper we introduced the 3D-PDR Orion dataset, which provides 8192 models that capture the chemistry of the Orion Bar. The emulators that we train on this dataset can be used for higher spatial and temporal resolution simulations, enabling us to resolve PDR regions in more detail. We show that the benchmark ANODE architectures can provide high fidelity predictions at a fraction of the computational cost, especially with the addition of the weight decay, learning rate scheduler, small weight initialization and weight decay. With these emulators, we enable fast computation of the chemistry of PDR regions, either by replacing the classical code altogether, or by using the emulator as a preconditioner for the classical code. In future work, the dataset will be expanded upon to contain actual line traces of 3D simulations of different regions, with variable density, visual extinction and cosmic ray attenuation profiles along the line of sight [Gaches et al., 2019].

## Acknowledgments and Disclosure of Funding

G.V., R.R. and S.V. acknowledge support from the European Research Council (ERC) Advanced grant MOPPEX 833460. T.G.B. acknowledges support from the Leading Innovation and Entrepreneurship Team of Zhejiang Province of China (Grant No. 2023R01008). The authors declare not competing interests.

The ANODEs were implemented using `diffrax`[Kidger, 2022] and `jax`[Bradbury et al., 2018]. Plots were made using `matplotlib` [Hunter, 2007]. The dataset was serialized into its final format using `h5py` [Collette, 2013].

## References

S. Andree-Labsch, V. Ossenkopf-Okada, and M. Röllig. Modelling clumpy photon-dominated regions in 3D. Understanding the Orion Bar stratification. *Astronomy and Astrophysics*, 598:A2, February 2017. ISSN 0004-6361. doi: 10.1051/0004-6361/201424287.

T. G. Bisbas, T. A. Bell, S. Viti, J. Yates, and M. J. Barlow. 3D-PDR: A new three-dimensional astrochemistry code for treating Photodissociation Regions. *Monthly Notices of the Royal Astronomical Society*, 427(3): 2100–2118, December 2012. ISSN 00358711, 13652966. doi: 10.1111/j.1365-2966.2012.22077.x.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: Composable transformations of Python+NumPy programs, 2018.

L Branca and A Pallottini. Neural networks: Solving the chemistry of the interstellar medium. *Monthly Notices of the Royal Astronomical Society*, 518(4):5718–5733, December 2022. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stac3512.

Lorenzo Branca and Andrea Pallottini. Emulating the interstellar medium chemistry with neural operators, February 2024.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations, December 2019.

Andrew Collette. *Python and HDF5*. O'Reilly, 2013.

D. de Mijolla, S. Viti, J. Holdship, I. Manolopoulou, and J. Yates. Incorporating astrochemistry into molecular line modelling via emulation. *Astronomy and Astrophysics*, 630:A117, October 2019. ISSN 0004-6361. doi: 10.1051/0004-6361/201935973.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented Neural ODEs, October 2019.

Brandt A. L. Gaches, Stella S. R. Offner, and Thomas G. Bisbas. The Astrochemical Impact of Cosmic Rays in Protoclusters. I. Molecular Cloud Chemistry. *The Astrophysical Journal*, 878:105, June 2019. ISSN 0004-637X. doi: 10.3847/1538-4357/ab20c7.

T. Grassi, F. Nauman, J. P. Ramsey, S. Bovino, G. Picogna, and B. Ercolano. Reducing the complexity of chemical networks via interpretable autoencoders, October 2021.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models, October 2018.

Tolga Güver and Feryal Özel. The relation between optical extinction and hydrogen column density in the Galaxy. *Monthly Notices of the Royal Astronomical Society*, 400(4):2050–2053, December 2009. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2009.15598.x.

Emilie Habart, Romane Le Gal, Carlos Alvarez, Els Peeters, Olivier Berné, Mark G. Wolfire, Javier R. Goicoechea, Thiébaut Schirmer, Emeric Bron, and Markus Röllig. High-angular-resolution NIR view of the Orion Bar revealed by Keck/NIRC2. *Astronomy & Astrophysics*, 673:A149, May 2023. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202244034.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. https://arxiv.org/abs/1512.03385v1, December 2015.

J. Holdship, S. Viti, T. J. Haworth, and J. D. Ilee. Chemulator: Fast, accurate thermochemistry for dynamical models through emulation. *Astronomy & Astrophysics*, 653:A76, September 2021. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202140357.

J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Patrick Kidger. On Neural Differential Equations, February 2022.

Mark A Kramer. Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *AIChE Journal*, 37(2), 1991.

Kookjin Lee and Eric Parish. Parameterized Neural Ordinary Differential Equations: Applications to Computational Physics Problems.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. https://arxiv.org/abs/1608.03983v5, August 2016.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. https://arxiv.org/abs/1711.05101v3, November 2017.

S. Maes, F. De Ceuster, M. Van de Sande, and L. Decin. MACE: A Machine learning Approach to Chemistry Emulation, May 2024.

T. J. Millar, C. Walsh, M. Van De Sande, and A. J. Markwick. The UMIST Database for Astrochemistry 2022. *Astronomy & Astrophysics*, 682:A109, February 2024. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202346908.

Els Peeters, Emilie Habart, Olivier Berné, Ameek Sidhu, Ryan Chown, Dries Van De Putte, Boris Trahin, Ilane Schroetter, Amélie Canin, Felipe Alarcón, Bethany Schefter, Baria Khan, Sofia Pasquini, Alexander G. G. M. Tielens, Mark G. Wolfire, Emmanuel Dartois, Javier R. Goicoechea, Alexandros Maragkoudakis, Takashi Onaka, Marc W. Pound, Sílvia Vicente, Alain Abergel, Edwin A. Bergin, Jeronimo Bernard-Salas, Christiaan Boersma, Emeric Bron, Jan Cami, Sara Cuadrado, Daniel Dicken, Meriem Elyajouri, Asunción Fuente, Karl D. Gordon, Lina Issa, Christine Joblin, Olga Kannavou, Ozan Lacinbala, David Languignon, Romane Le Gal, Raphael Meshaka, Yoko Okada, Massimo Robberto, Markus Röllig, Thiébaut Schirmer, Benoit Tabone, Marion Zannese, Isabel Aleman, Louis Allamandola, Rebecca Auchettl, Giuseppe Antonio Baratta, Salma Bejaoui, Partha P. Bera, John H. Black, Francois Boulanger, Jordy Bouwman, Bernhard Brandl, Philippe Brechignac, Sandra Brünken, Mridusmita Buragohain, Andrew Burkhardt, Alessandra Candian, Stéphanie Cazaux, Jose Cernicharo, Marin Chabot, Shubhadip Chakraborty, Jason Champion, Sean W. J. Colgan, Ilsa R. Cooke, Audrey Coutens, Nick L. J. Cox, Karine Demyk, Jennifer Donovan Meyer, Sacha Foschino, Pedro García-Lario, Maryvonne Gerin, Carl A. Gottlieb, Pierre Guillard, Antoine Gusdorf, Patrick Hartigan, Jinhua He, Eric Herbst, Liv Hornekaer, Cornelia Jäger, Eduardo Janot-Pacheco, Michael Kaufman, Sarah Kendrew, Maria S. Kirsanova, Pamela Klaassen, Sun Kwok, Álvaro Labiano, Thomas S.-Y. Lai, Timothy J. Lee, Bertrand Lefloch, Franck Le Petit, Aigen Li, Hendrik Linz, Cameron J. Mackie, Suzanne C. Madden, Joëlle Mascetti, Brett A. McGuire, Pablo Merino, Elisabetta R. Micelotta, Karl Misselt, Jon A. Morse, Giacomo Mulas, Naslim Neelamkodan, Ryou Ohsawa, Roberta Paladini, Maria Elisabetta Palumbo, Amit Pathak, Yvonne J. Pendleton, Annemieke Petrignani, Thomas Pino, Elena Puga, Naseem Rangwala, Mathias Rapacioli, Alessandra Ricca, Julia Roman-Duval, Joseph Roser, Evelyne Roueff, Gaël Rouillé, Farid Salama, Dinalva A. Sales, Karin Sandstrom, Peter Sarre, Ella Sciamma-O'Brien, Kris Sellgren, Sachindev S. Shenoy, David Teyssier, Richard D. Thomas, Aditya Togi, Laurent Verstraete, Adolf N. Witt, Alwyn Wootten, Nathalie Ysard, Henning Zettergren, Yong Zhang, Ziwei E. Zhang, and Junfeng Zhen. PDRs4All - III. JWST's NIR spectroscopic view of the Orion Bar. *Astronomy & Astrophysics*, 685:A74, May 2024. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202348244.

Yulia Rubanova, Ricky T Q Chen, and David Duvenaud. Latent ODEs for Irregularly-Sampled Time Series.

Immanuel Sulzer and Tobias Buck. Speeding up astrochemical reaction networks with autoencoders and neural ODEs, December 2023.

Kwok Sun Tang and Matthew Turk. Reduced Order Model for Chemical Kinetics: A case study with Primordial Chemical Network, July 2022.

Gijs Vermariën and Serena Viti. 3DPDR Orion inspired dataset, September 2024.

## A The relation between $A_V$ index and $\log(A_V)$ for different models

We show here the connection between the index of the visual extinction and its logs values.

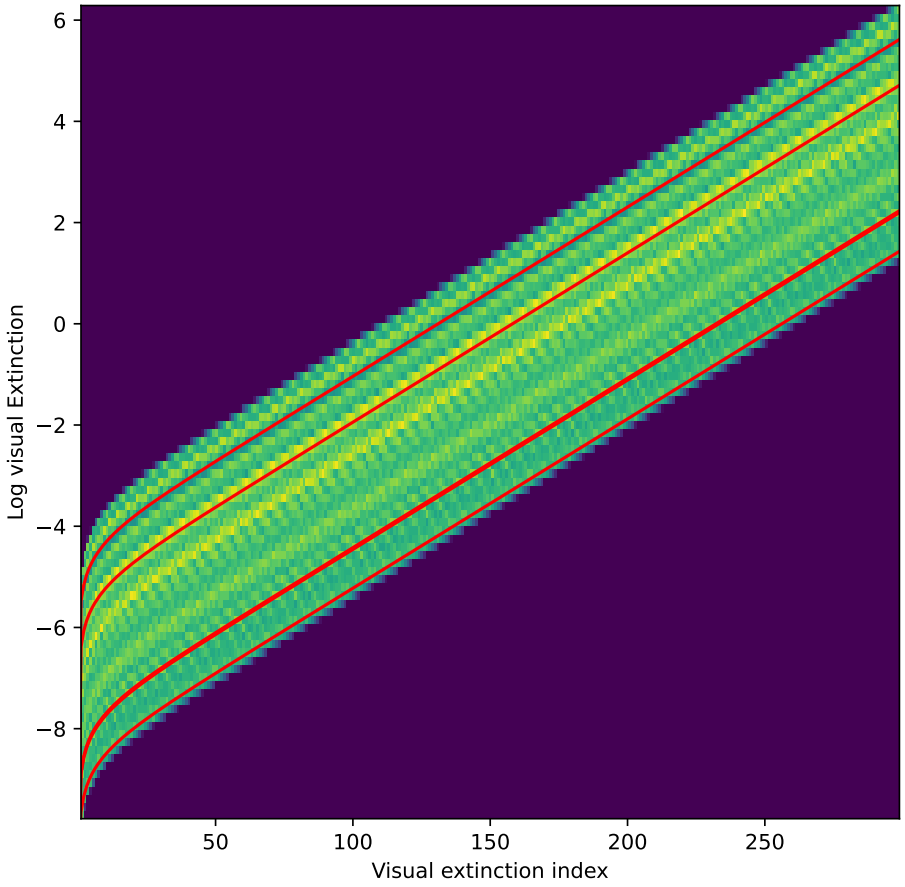


Figure 3: The distribution of visual extinction index versus the actual log values of the index. The 5 red lines are 5 individual samples, with the 2d histogram the distribution.

## B Loss function

We provide here the validation loss as it is computed during the training process, which is directly on the standardized training output.

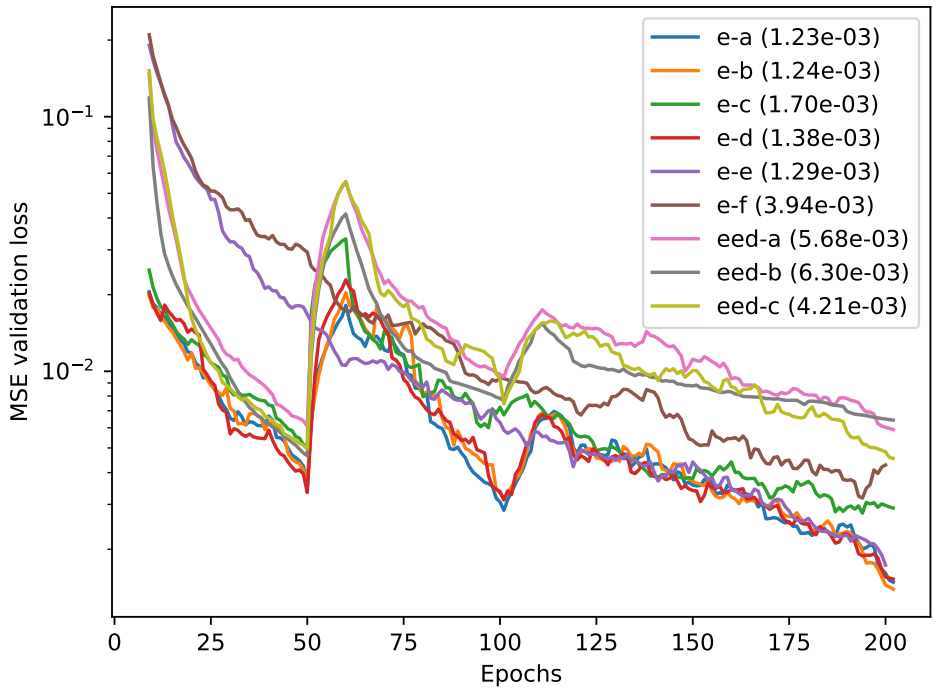


Figure 4: A comparison of the validation for the losses, with a moving average of 10 steps. The final loss is denoted in the legend behind each architecture.

## C MSE for each individual feature

The mean squared error in feature space for each of the series for the different models.

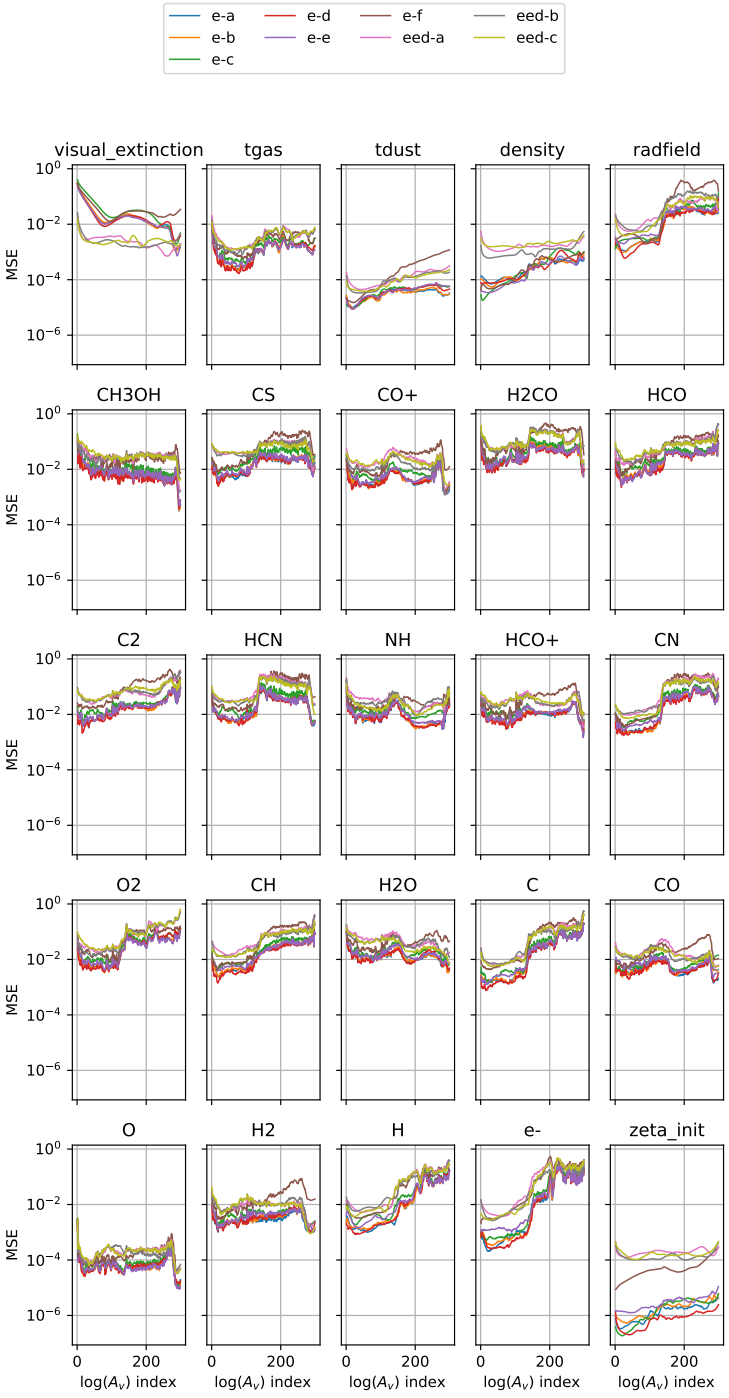


Figure 5: The MSEs for each individual molecule and all other features, for all different architectures.