Deciphering the Structure of Molecular Clouds with Neural Networks

Aryan Jain

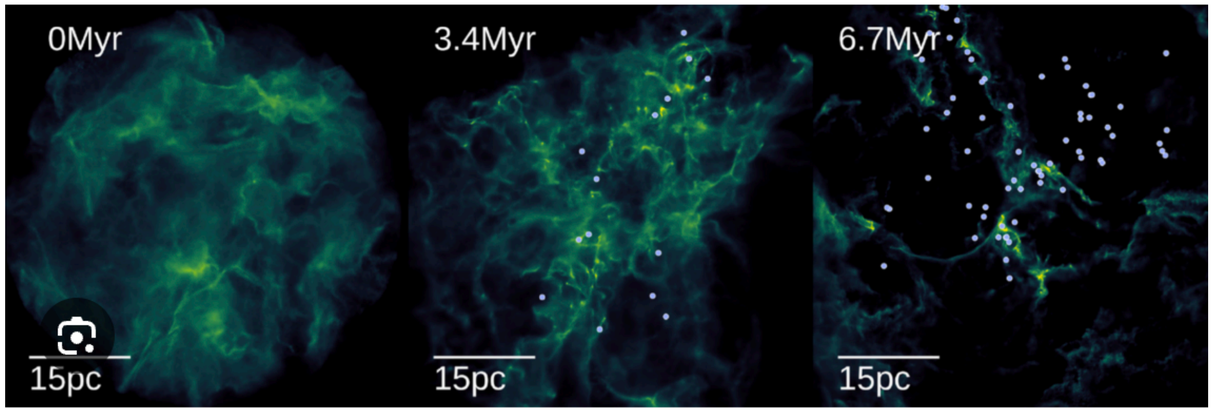
Supervisors: Marta Reina Campos , Shivan Khullar

What is the ISM, what are GMCs ?

The space between stars is not that empty !! It contains dust and gases

The ISM is a region of high interest for many researchers due to its turbulent environment

GMCs are prime regions of chaotic activity between molecular gases



Caption

Gravitational instability

Radiative Cooling

Self-gravity

Fragmentation of Gas

What is our Focus ?

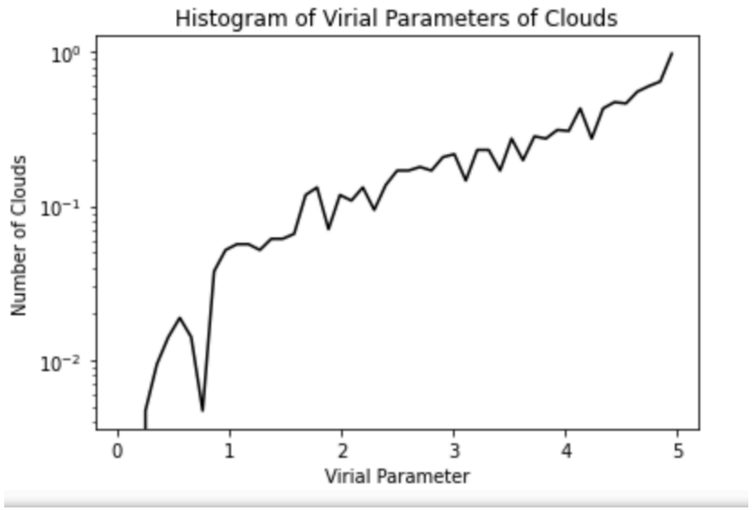
Enhancing GMC Observational Studies with Machine Learning

- **Objective:** Leverage machine learning models trained on galaxy simulations to support and extend the capabilities of observational studies.
- **Clustering Techniques:** Utilise algorithms such as Cloudphinder to identify and analyze cloud clusters in Milky Way-like galaxy simulations.
- **Predictive Modelling:** Train regression models on simulation-derived data to predict parameters for observational data, enabling insights into areas where direct observations are limited.

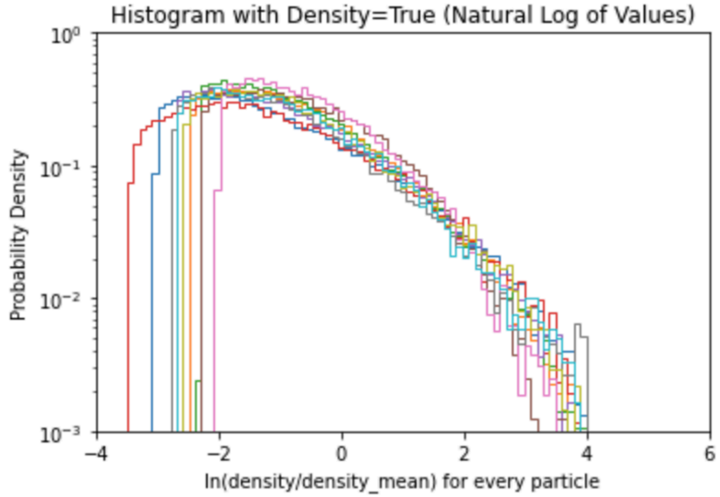
Studying molecular clouds in the FIRE-2 simulations

We use galaxy simulation catalogues of galaxies similar to the milky way galaxy

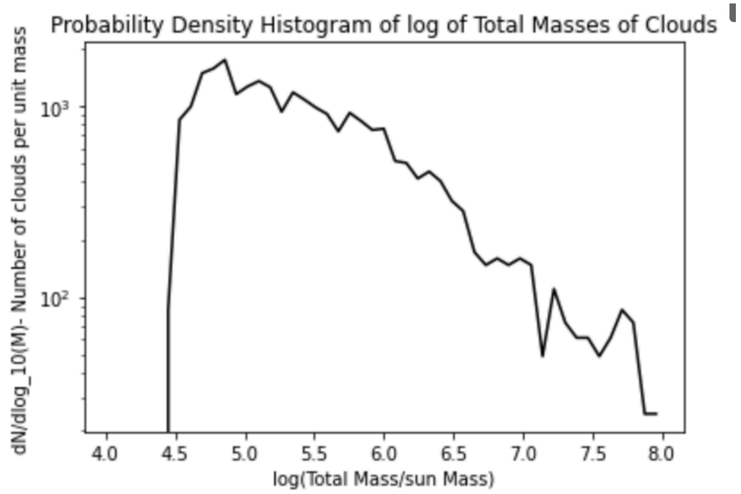
The algorithm clusters particles based on density peak plots and cuts-off at a lower bound



The virial parameter tells us the ratio between the kinetic and potential energy



The density plot tells us that majority of the particles have density lower than the mean density, from observations this plot follows a log- normal distribution



This plot shows us that most of the molecular clouds have a map of about 10^5 solar masses

The project aims at aiding observational studies of GMCs by using supervised learning models. We use 3 major machine learning models:

- Neural Network
- Decision Trees
- Random Forest

There models are used to predict the following quantities:

1. 3-d volume density of clouds in (M_sun/pc^3)
2. Shape of the PDF of cloud particle densities
3. 3-d dispersion velocity

Model Input Data

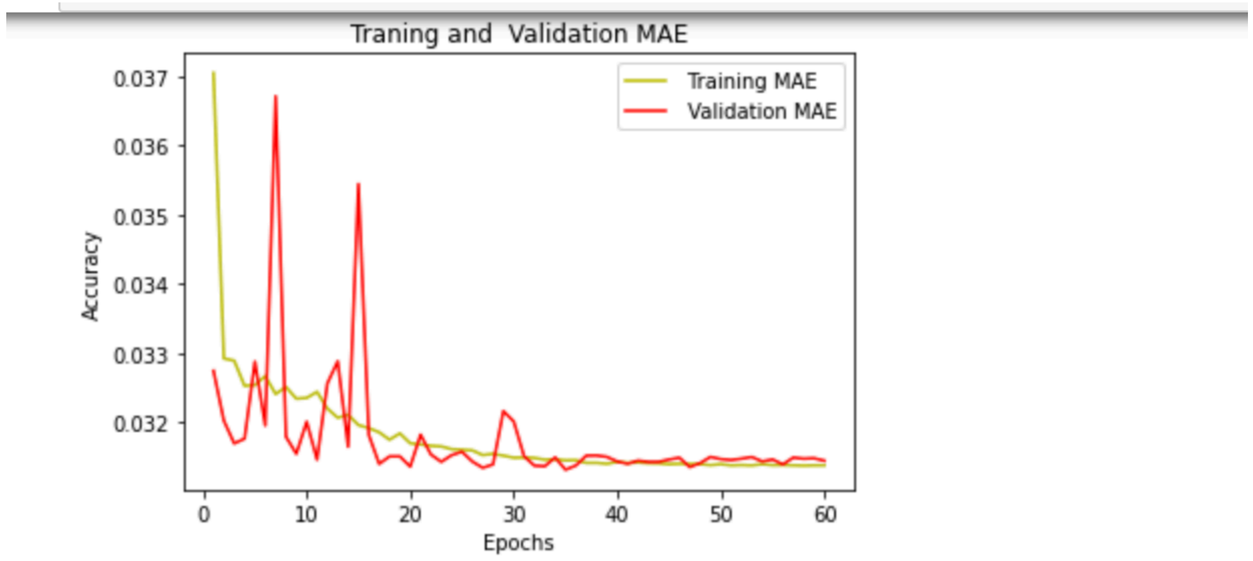
The regression models we use take observationally available data as its input and learns a function the output physical quantity

We use the following quantities as inputs to our models:

1. Galacto-centric radius in pc
2. Surface density (2-d) cloud density in M_sun/pc^2
3. Dispersion velocity in the z direction in km/s

3-D Density Model results

We used Neural Networks and Random Forest to train a model that predicts 3-d volume density of GMCs




```
139/139 [==============================] - 0s 932us/step
MAPE scaled( data both log scaled and normalized ): 10.378721360166487
MAPE linear : 68.15267814636762
```

Figure 1 : mean absolute percentage error and training loss of Neural Network model for predicting 3-d density

```
MAPE: 10.902430714709514
139/139 [==============================] - 0s 929us/step
MAPE scaled( data both log scaled and normalized ): 10.378721360166487
MAPE linear : 68.15267814636762
```

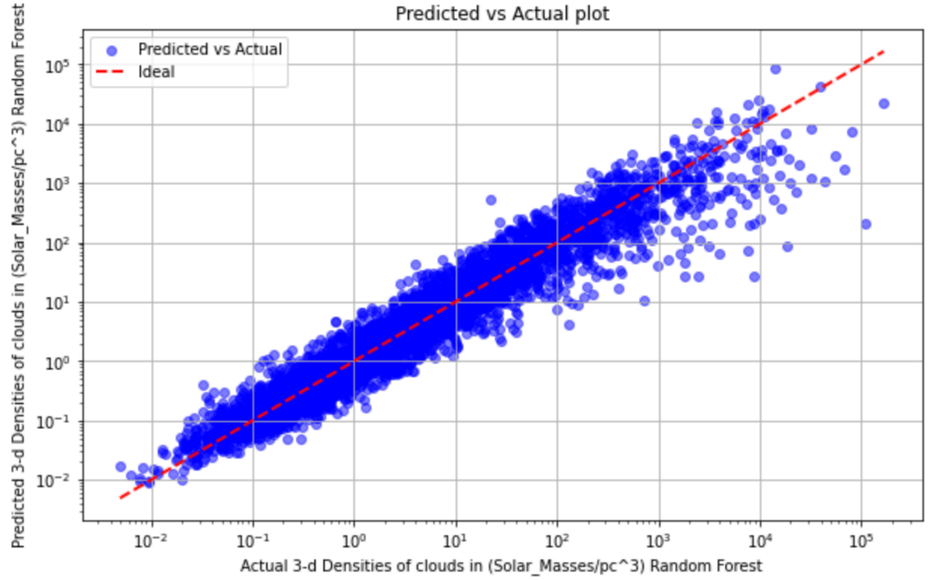
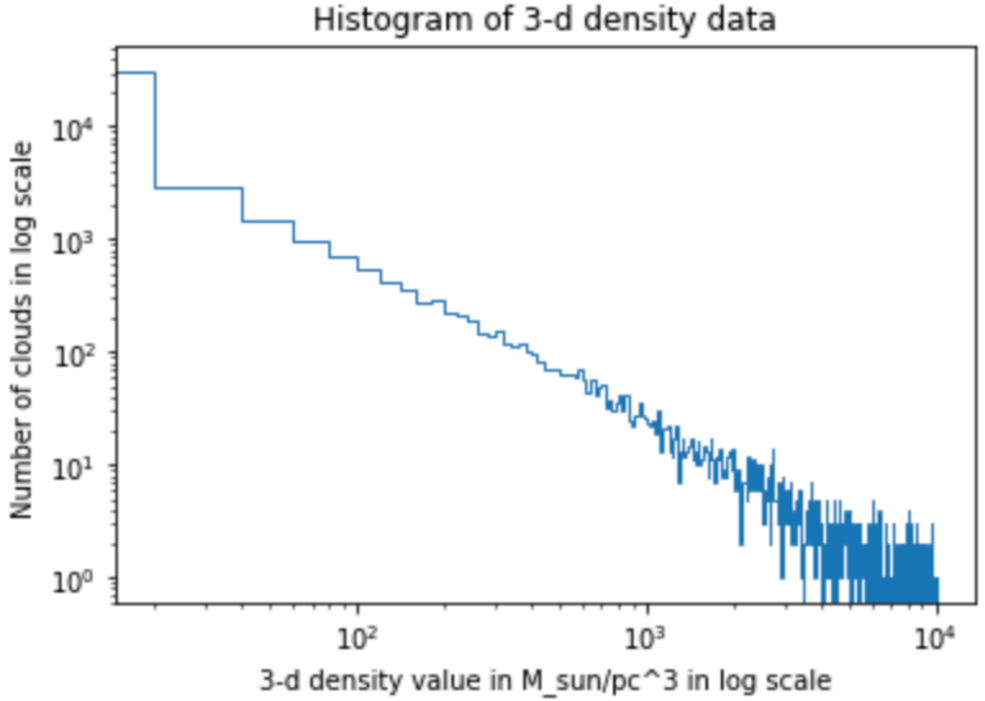
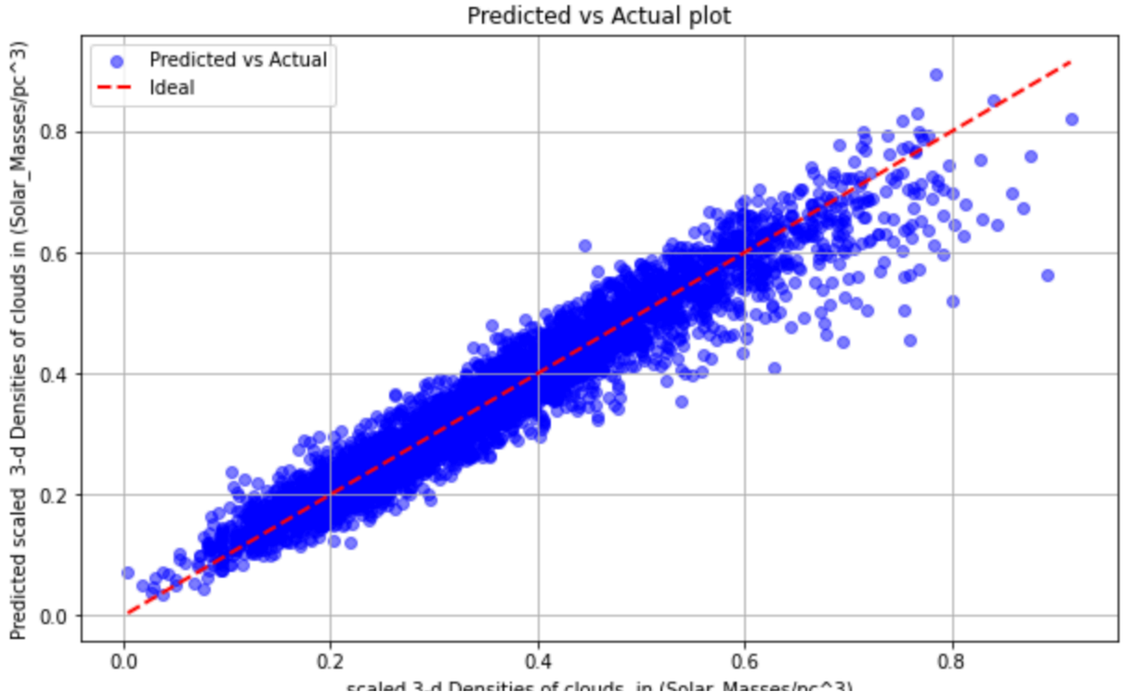


Figure 2: mean absolute percentage error of Random Forest model for predicting 3-d density

What does the 3-d Model tell us ?

1. Most of the 3-d densities are close to 1 with a median of 2.996 but mean of 430.26 meaning that the distribution of the 3-d densities are skewed to the left
2. The model predictions are very accurate for smaller values and have a log mean absolute error meaning the error increases with higher values of 3-d density.





3-d density Model on Observational Data

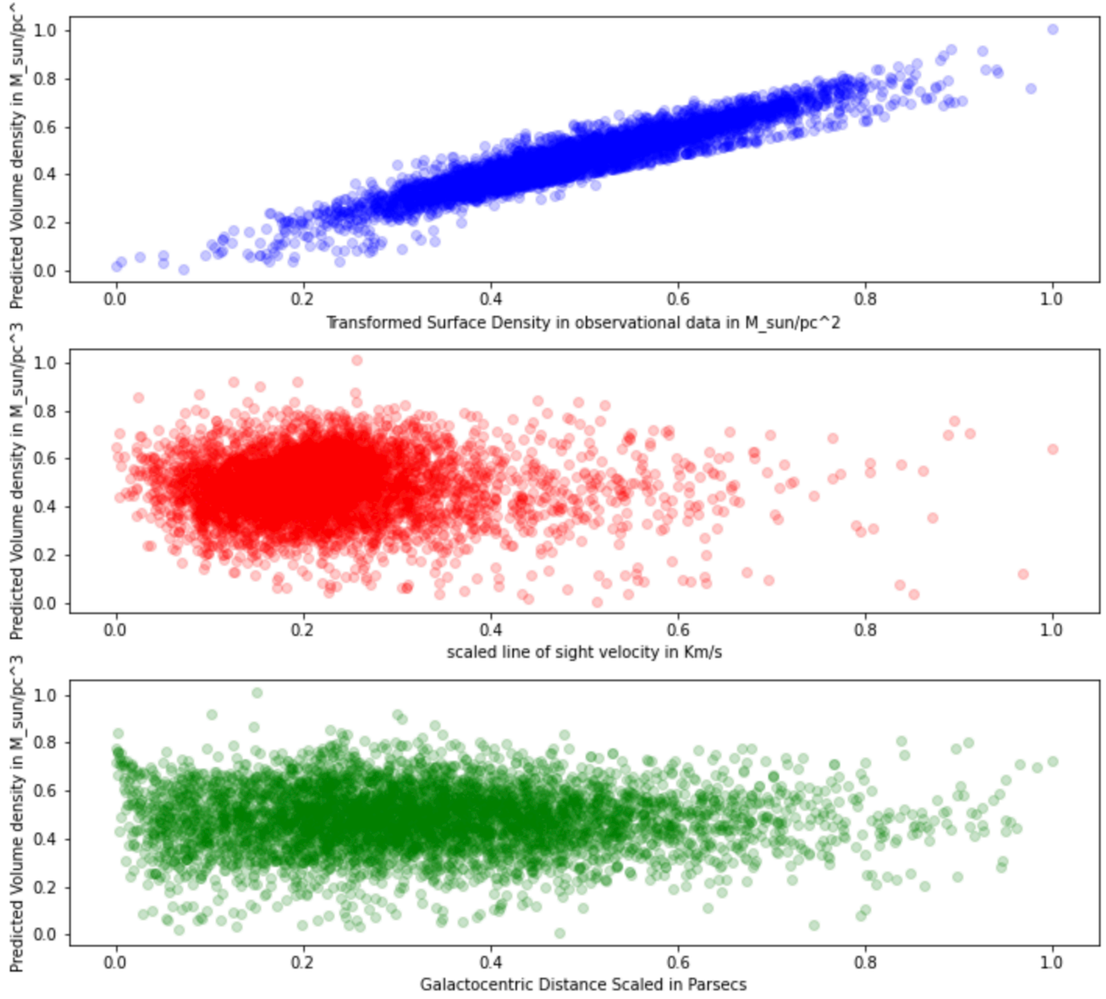


Figure 1: Observational Data Scatter Plots

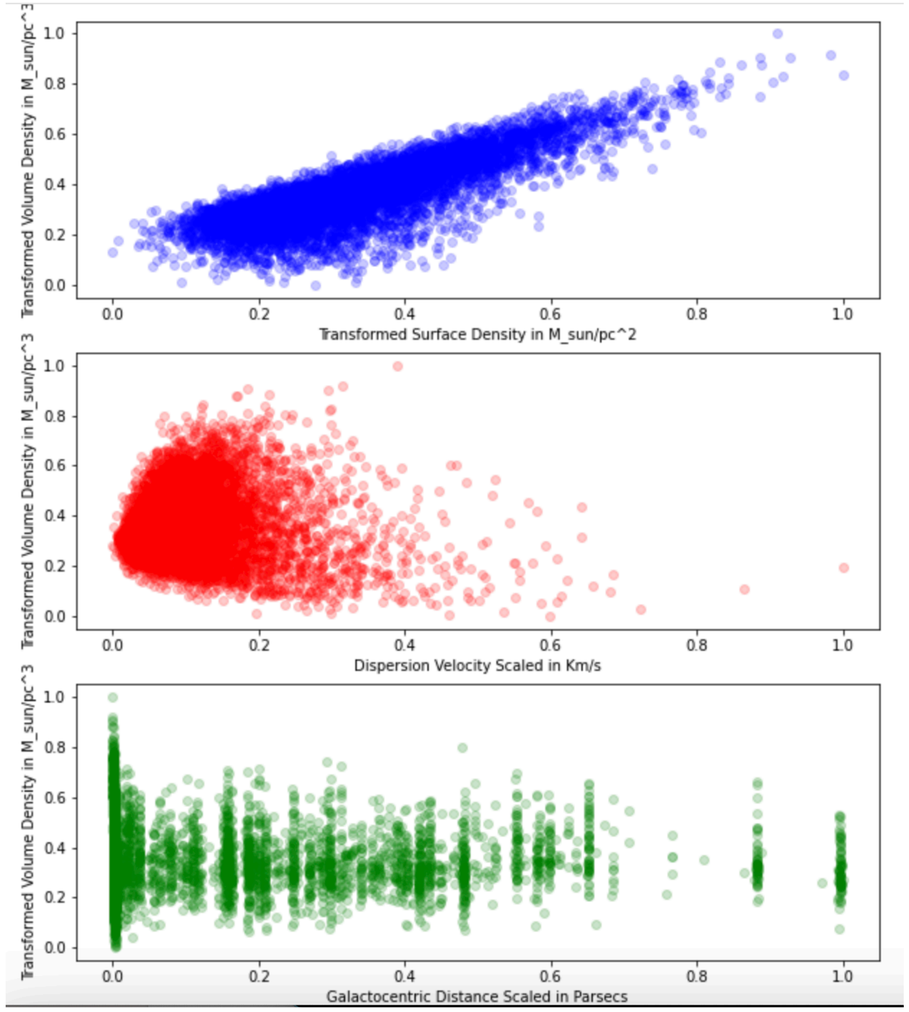


Figure 2 : Simulated Testing Data Scatter Plots

Cloud particle Density Model

The shape of the Cloud particle densities is defined by two parameters :

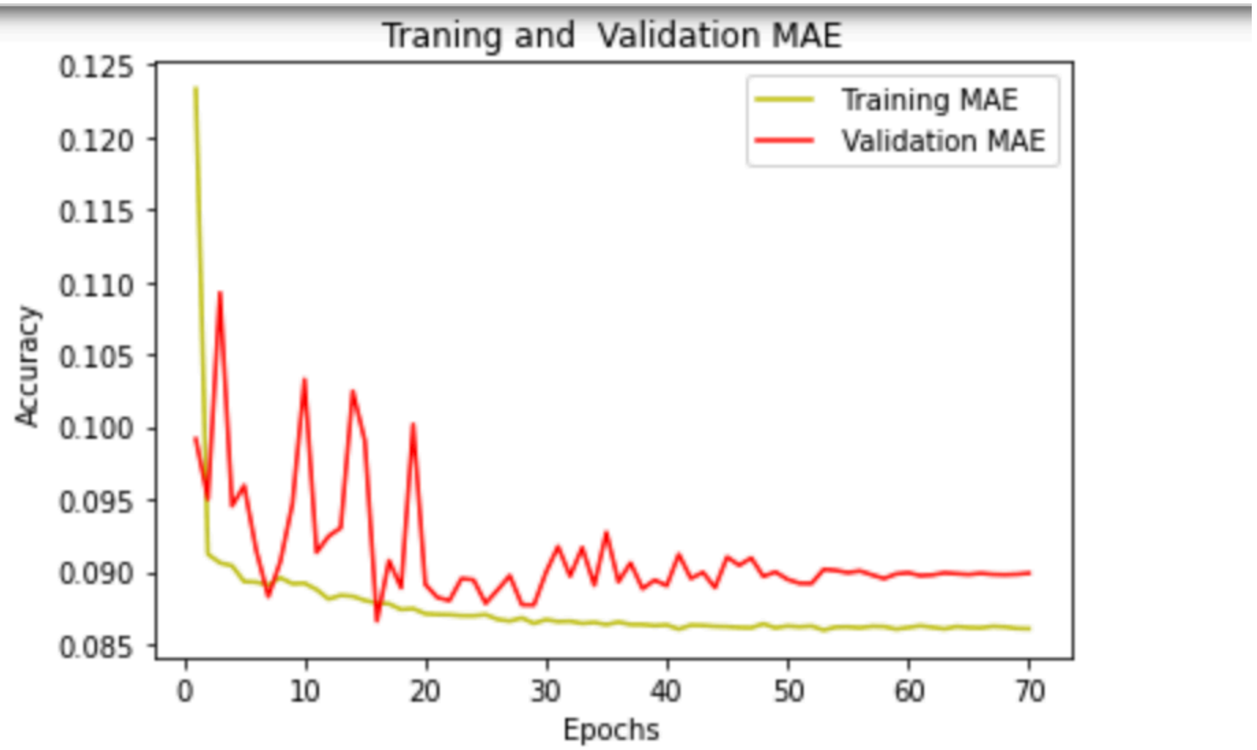
- S : The mean of the log difference of cloud particle density and mean cloud particle density in M_sun/ pc^3
- The dispersion of the cloud particle densities in the middle 68th percentile range.It is typically measured as the difference between the 84th and 16th percentiles of the logarithmic densities.

$$s_i = \log\left(\frac{\rho_i}{\bar{\rho}}\right)$$

$$\text{variance-like measure} = \log_{10}(P_{84}) - \log_{10}(P_{16})$$

$$\text{mean}_s = \frac{1}{N}\sum_{i=1}^{N} s_i$$

Variance Model Training Results




```
25/25 [==============================] - 0s 1ms/step
MAPE scaled( data only normalized): 12.036531203066474
MAPE linear : 59.96540748898165
```

Figure 1: Training Loss and Mean Absolute Percentage Error

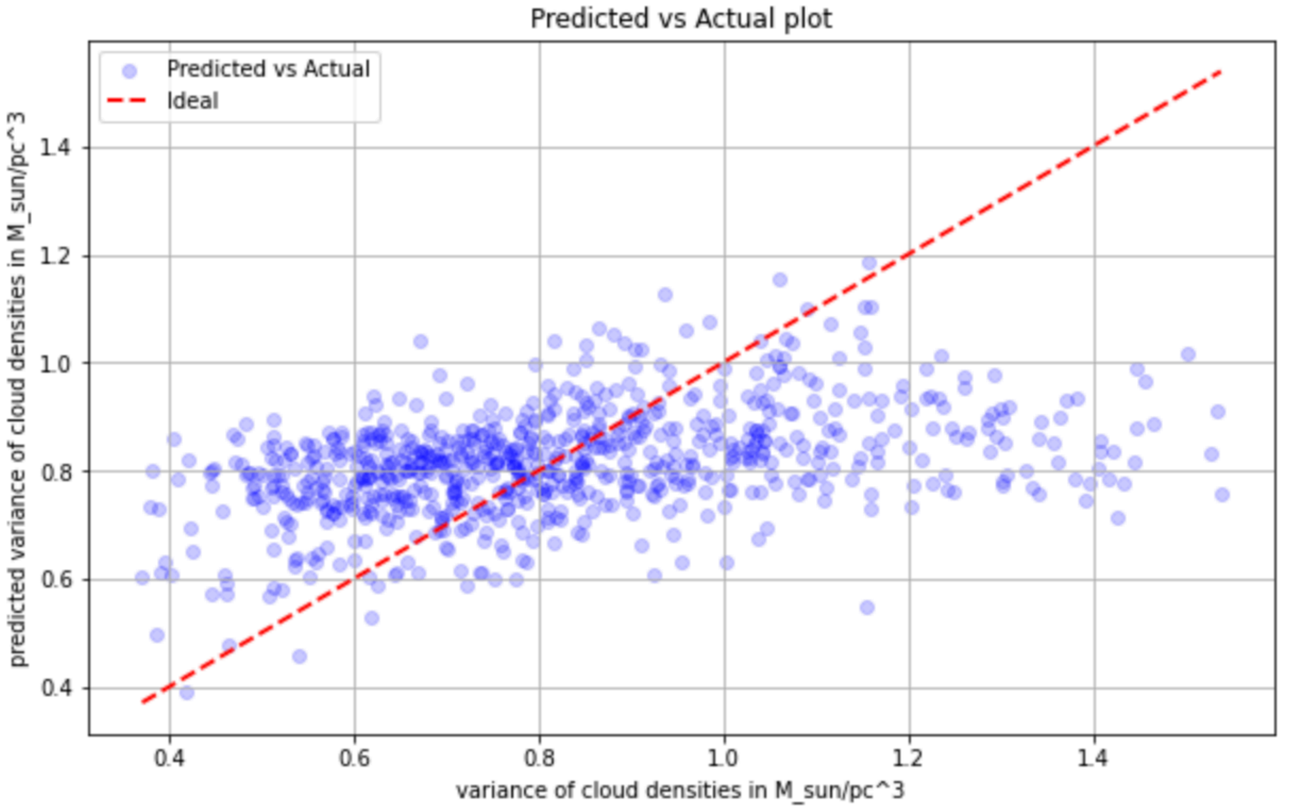
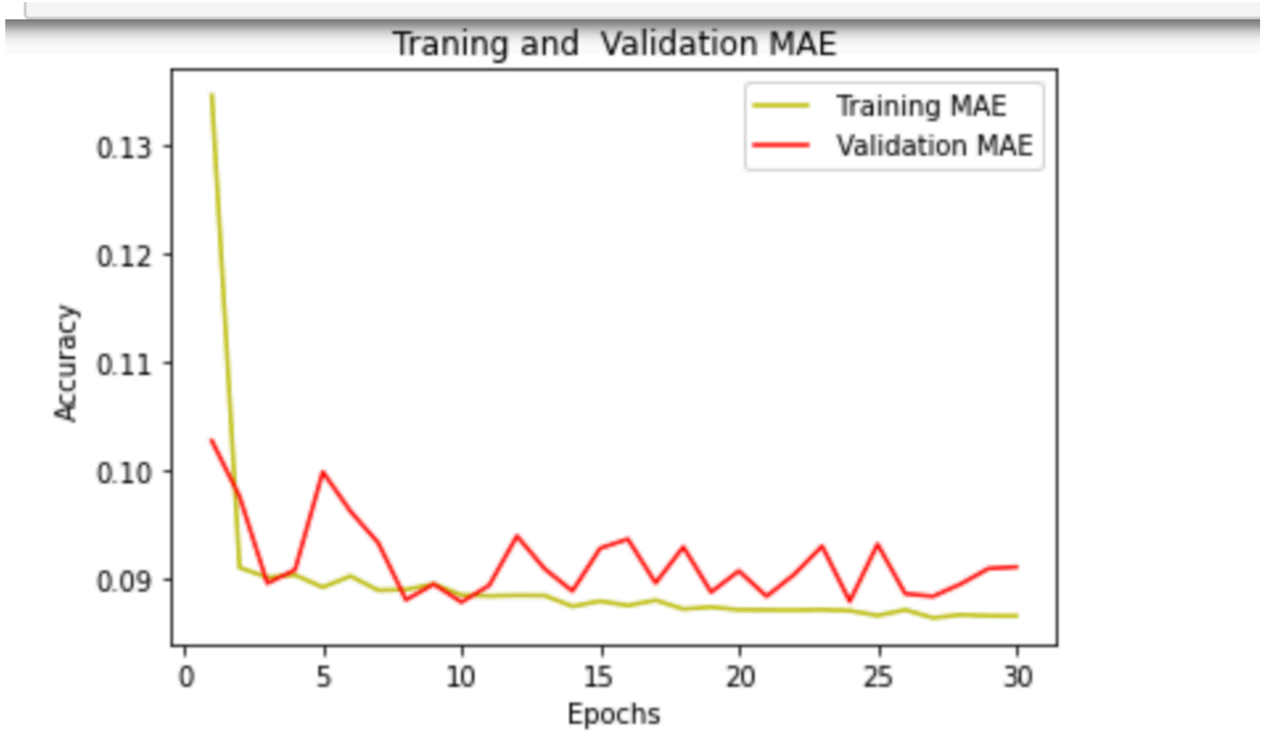


Figure 2: Graph of Predicted Vs Actual testing data in the original scale

Mean Model Results




```
25/25 [==============================] - 0s 1ms/step
MAPE scaled( data only normalized): 12.095640092704105
MAPE linear : 61.80807480553071
```

Figure 1: Training Loss and Mean Absolute Percentage Error

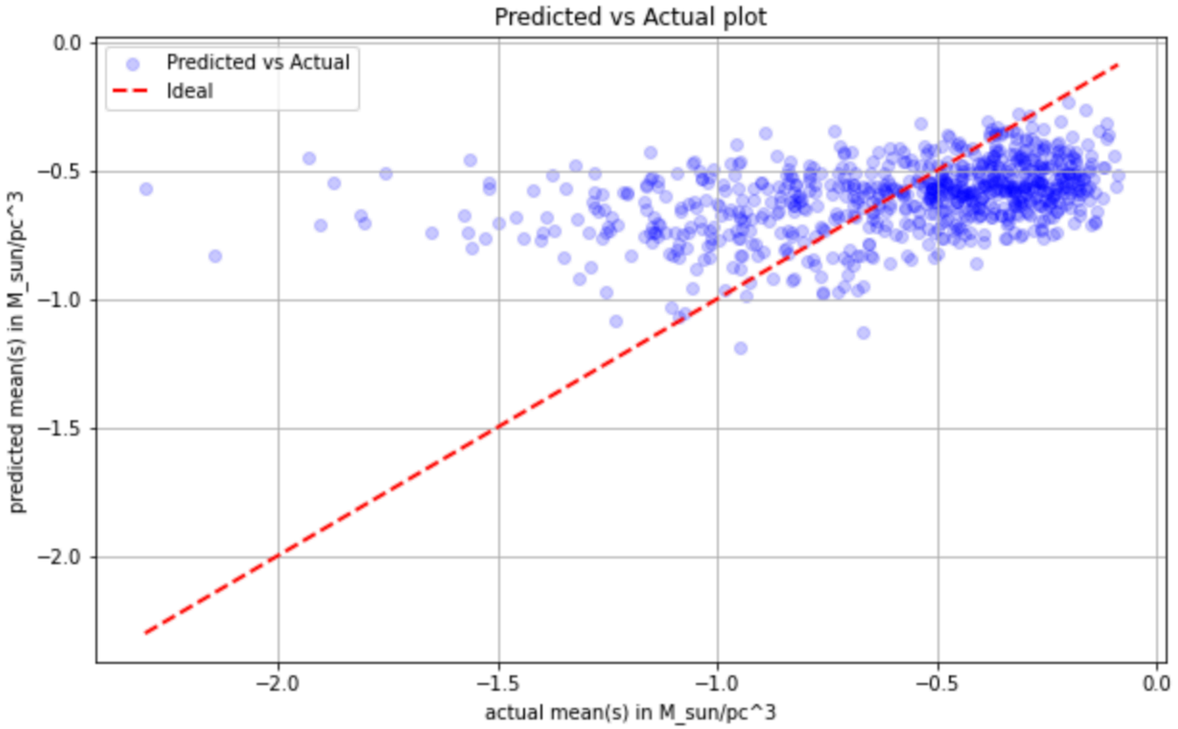


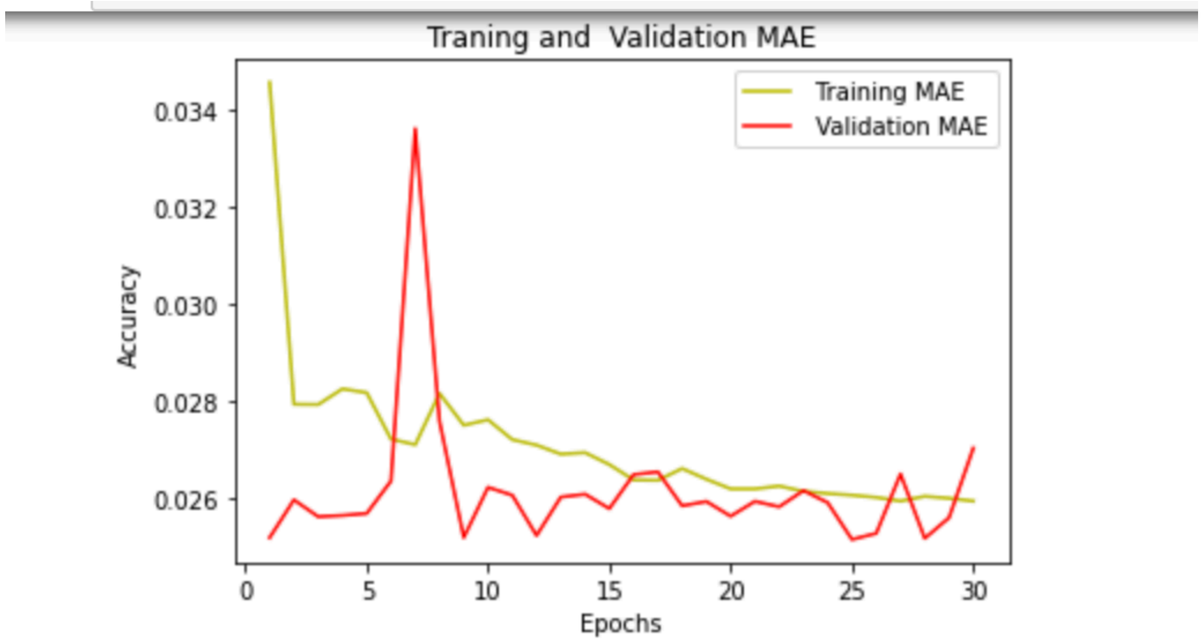
Figure 2: Graph of Predicted Vs Actual testing data in the original scale

Understanding the Suboptimal Performance of the Mean and Variance Model

The mean and variance models exhibit weaker performance compared to the 3D density model due to the following reasons:

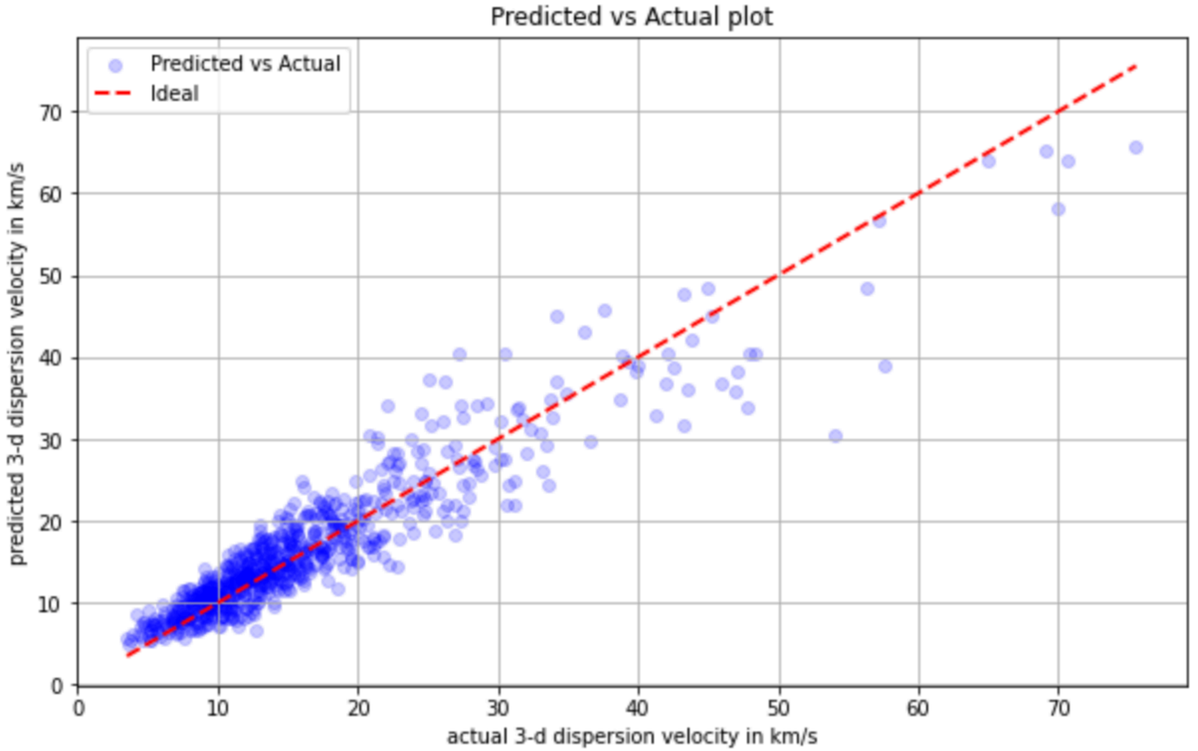
1. **Lack of Strong Correlation:** The mean and variance metrics do not show a strong correlation with the input quantities
2. **Dominance of 2D Density:** The network primarily learns the correlation between 3D density and 2D density, making the other two input quantities largely insignificant in the 3D density model, but this advantage does not translate to the mean and variance models

3-d velocity Dispersion Model



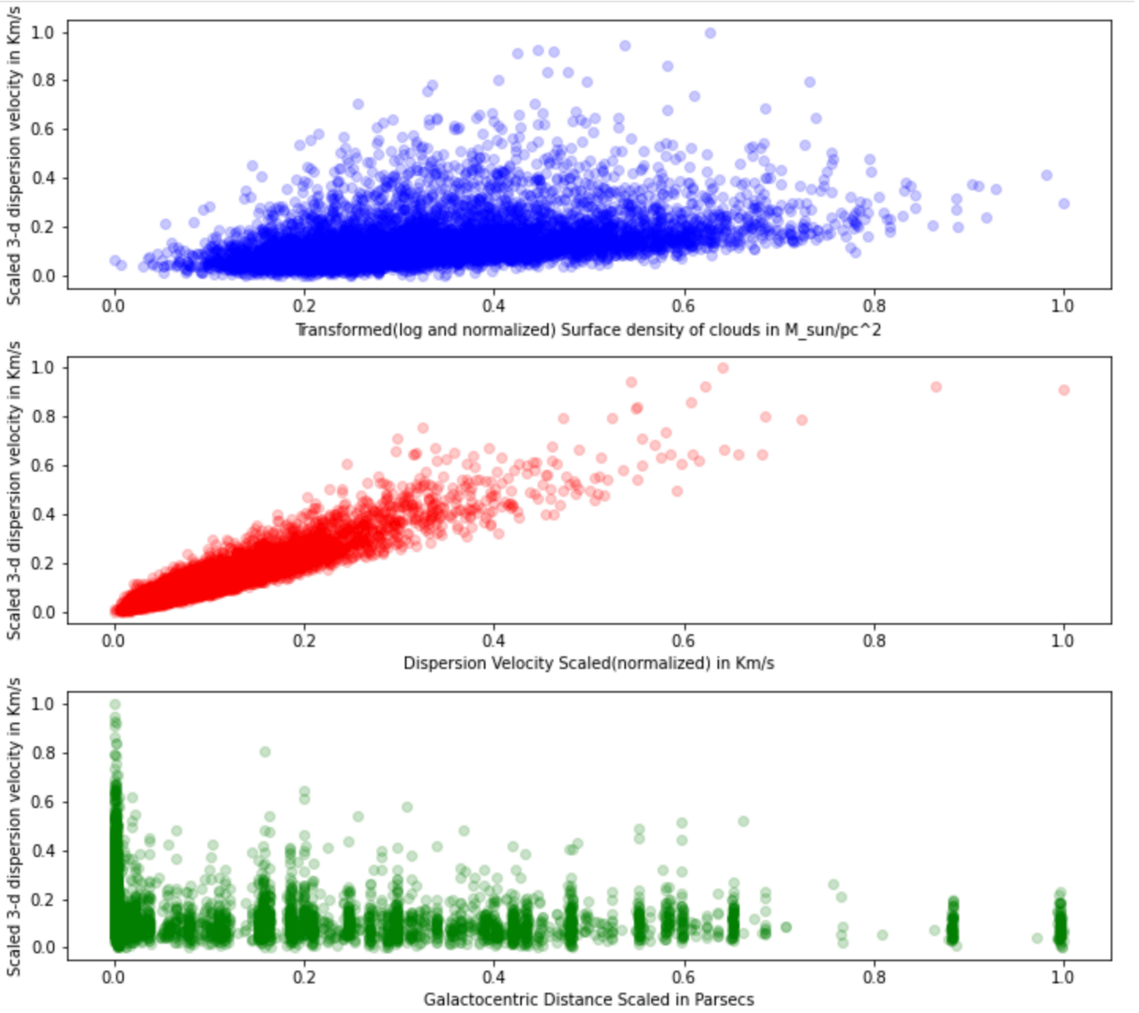

```
25/25 [==============================] - 0s 1ms/step
MAPE scaled( data only normalized): 29.030978228772092
MAPE linear : 16.80538770588763
```

We can observe much better performance compared to Mean(s) and variance model

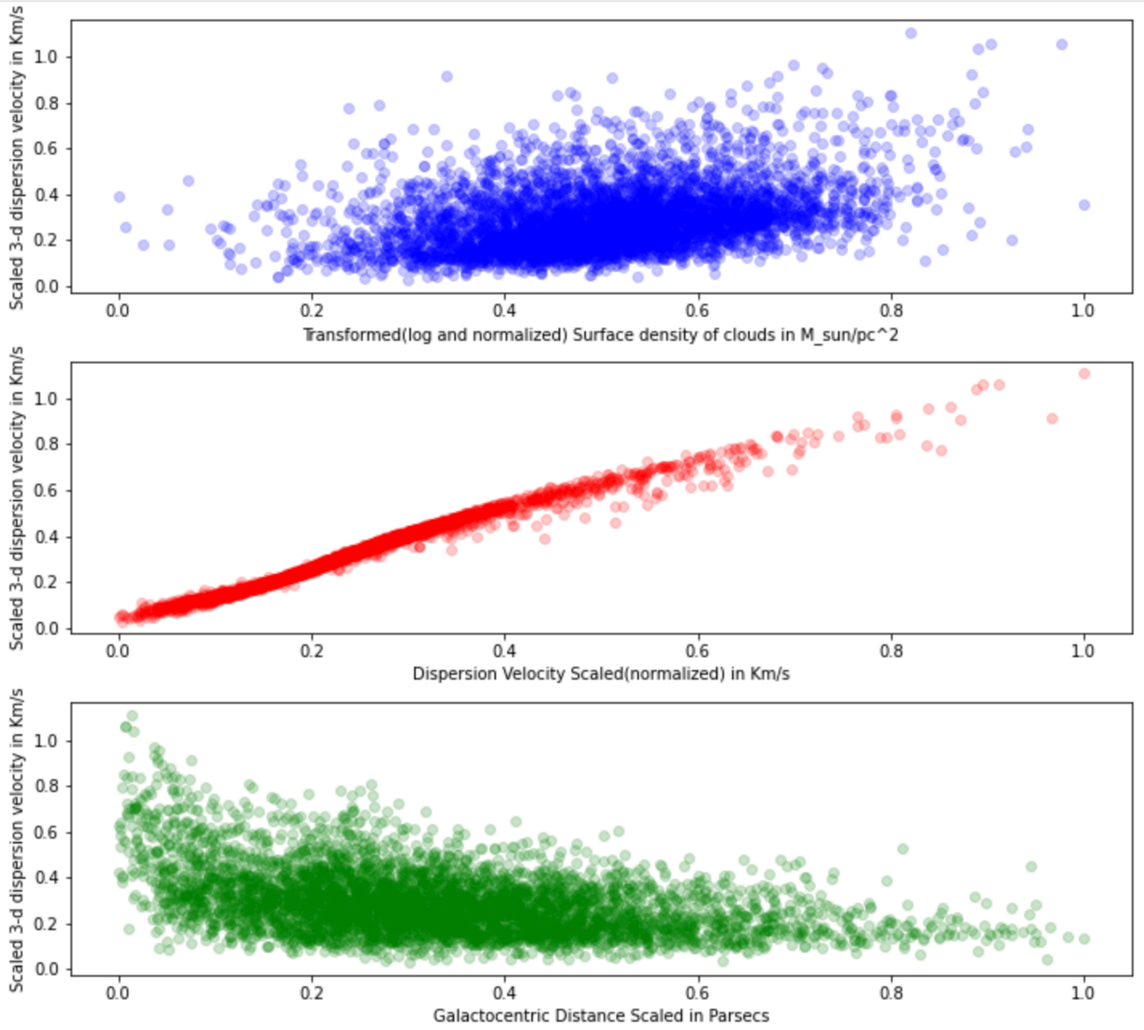


There is a positive correlation between the actual and predicted data indicating that the model learned the distribution of the 3-d dispersion velocities

Observational data on 3-d dispersion Velocity Model



Simulated Data Scatter Plots



Observational Data Scatter Plots

The 3-d dispersion velocity model has prediction distributions following the same trend as the simulated data

Conclusions

1. **Model Development:** Machine learning models were designed to predict unobservable quantities in GMCs, focusing on 3D volume density and 3D dispersion velocity.

2. **Training Performance:** These models excelled during training due to strong correlations between 3D density, 3D dispersion velocity, and input data.

3. **Observational Data Alignment:** The models successfully matched trends in observational data with their predictions, demonstrating effective fitting of input data to the predicted output distribution.

4. **Normalisation Challenge:** Future work should explore methods to renormalize data to its original scale for more accurate predictions.

5. **Potential Approach:** Implementing z-score normalisation could offer a more effective alternative to simple normalisation in preserving data integrity.

6.

Thank You