**Research Proposal**

**1. Proposed project title.**

Representation learning for neutral hydrogen mapping.

**2. Introduction and Background.**

Resolving the phase structure of neutral hydrogen (Hi) gas is crucial for understanding the life cycle of the interstellar medium (ISM). The fraction of cold neutral medium (F\_cnm) and opacity correction factor (R\_hi) are two important values that decide the phase structure of neutral hydrogen.

However, accurate measurements of R\_hi and F\_cnm are limited by the availability of background continuum sources for measuring Hi absorption. In traditional methods, optical depth information needed to be provided in order to get the R\_hi and F\_cnm. In most of the cases, optical depth information does not exist and the accurate prediction of R\_hi and F\_cnm is impossible in traditional methods.

**3. Problem statement.**

As accurate measurements of Hi temperature and density are limited by the availability of background continuum sources for measuring Hi absorption. As a result, the prediction of R\_hi and F\_cnm is impossible using traditional methods.

Due to the property of the spectra, applying deep convolutional neural networks is a solution for predicting the parameters that we care about without needing so much background information. Recently, a study has been put forward by Claire E. Murray, et al where they applied a 2-layer convolutional neural network to predict R\_hi and F\_cnm based on synthetic spectra data [1]. To be more specific, they first used synthetic spectra data to train the CNN model and then tested the model on 58 ground truth samples. However, there are still some problems in this field that need to be explored and solved.

1) The CNN model in the study is simple. More sophisticated design of CNN may leverage the prediction accuracy of R\_hi and F\_cnm.

2) The CNN model in the study tends to over-interpret sharp spectral edges, which results in an under-prediction of f\_cnm. This phenomenon can be considered as an overfitting problem and may be caused by the relatively simple CNN structure.

3) the value of R-hi and F\_cnm in an area usually has a closed relationship with the surrounding area’s R\_hi and F\_cnm values. The study calculated the values only based on the spectra of the area, which may cause accuracy reduction. We would like to explore the prediction of F\_cnm and R\_hi using spectra data and surrounding information together.

4) As CNN usually requires large training set and CNN model in the study is lacked of interpretability, one important aspect of the research is to make the model be interpretable to the astronomer.

In summary, there is a need for a better convolutional neural network architecture and data processing technique for the better prediction of R\_hi and F\_cnm in improving the prediction accuracy and reducing the overfitting problem.

**4. Objectives**

The objective of the research is to explore new convolutional neural network architectures, which can achieve higher regression accuracy in the R\_hi and F\_cnm prediction and solving the problem of overfitting in spectra data with sharp edges. Particularly, the study has the following sub-objectives:

1)To provide a review of R\_hi and F\_cnm calculation methods.

2) Perform regression (using these features) to determine R\_hi and F\_cnm with the absence of background information. To be more specific, we would like to explore convolutional neural network architectures or new mechanism for leveraging prediction accuracy and reducing overfitting problems.

3) Exploring 2D representation methods from 1D signals. To be more specific, the signal of spectra is a one-dimensional data. We would like to develop method that can represent 1D data with 2D representation and compare model fitting performance with two different data formats. One example is the spectrogram used in audio or ECG signal transformation.

4) As the noise in the universe is huge, which may affect the values in each channel if the spectra data. Extracting the feature and recognize the noise in an spectra vector is important. One potential solution is using self-supervised learning (comparative learning) method to train CNN and make the output as close as possible based on real vector and vector that contains gaussian noise.

The result of this study will be valuable to astronomy research as well as related data processing fields in developing better practice and tools for handling 1D data regression problems.

**5. Methodology**

The primary research method for this study is literature review, theoretical and mathematical analysis and comparative experiments. In the first stage of the research, we will first review different types of research in one dimensional data processing using convolutional neural networks and data processing techniques. As the research [1] contains the drawbacks of over-interpreting sharp spectral edges, which results in an under-prediction of f\_cnm. One direction is to design CNN that can focus on border area of the signal so as to leverage the model accuracy in Hi property prediction.

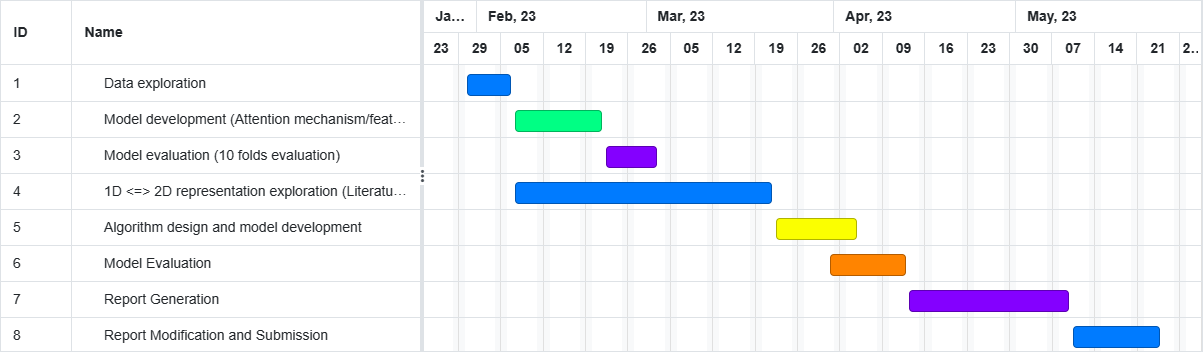
In the second stage of the research, we would like to focus on the 2D representation of 1D signal. Deep dive the representation method of transforming 1D signal into 2D dimensions (e.g. spectrogram or line plot) and put forward new methods of dimensional transformation. Then, we would like to further compare the model performance of 1D and 2D convolutional neural network. The final report will be based on both tasks.

**6. Raw data and resources.**

The research will mainly be based on synthetic spectra data with noise in the spectral (velocity) axis, which was put forward by E. Saury et al in 2014 (E. Saury, et, al 2014). The training set is a data cube with a size of (101, 512, 512) includes (512x512) spectra and their ground-truth cold atomic hydrogen gas fraction (FCNM) and opacity correction factor (R), where F\_cnm range: from ~0.0 to 1 and R\_hi range: >= 1 (E. Saury, et, al 2014). For model testing, we will use 58 HI spectra samples from [1] as our testing set. This 58 samples contain 30 spectra samples from 21-SPONGE (Murray et al. 2015, 2018b) survey and 28 samples from Millennium Survey (Heiles & Troland 2003a,b). Each testing sample is a 1 × 414 channels of spectra with corresponding R\_hi and F\_cnm. In order to prevent the uncertainty of the CNN, ten-fold validation will be used in our experiments.

Model training will be based on pytorch. Laptop equipped with GTX1080 will be used for CNN trainings.

|  |  |  |  |
| --- | --- | --- | --- |
| Data Files | Usage | Data Shape | description |
| [Tb\_n01\_pw02\_vs12\_512\_thick\_noise\_1.0\_K\_beam\_1.45\_pix\_.fits](https://anu365-my.sharepoint.com/personal/u1129269_anu_edu_au/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fu1129269%5Fanu%5Fedu%5Fau%2FDocuments%2FHI%2FDecomposition%2Fdata%2FTb%5Fn01%5Fpw02%5Fvs12%5F512%5Fthick%5Fnoise%5F1%2E0%5FK%5Fbeam%5F1%2E45%5Fpix%5F%2Efits&parent=%2Fpersonal%2Fu1129269%5Fanu%5Fedu%5Fau%2FDocuments%2FHI%2FDecomposition%2Fdata&ga=1) | Training (spectra) | 512 x 512 x 101 | Synthetic spectra data for training |
| [fcnm\_map\_n01\_pw02\_vs12\_Tcut\_500K\_512px.fits](https://anu365-my.sharepoint.com/personal/u1129269_anu_edu_au/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fu1129269%5Fanu%5Fedu%5Fau%2FDocuments%2FHI%2FDecomposition%2Fdata%2Ffcnm%5Fmap%5Fn01%5Fpw02%5Fvs12%5FTcut%5F500K%5F512px%2Efits&parent=%2Fpersonal%2Fu1129269%5Fanu%5Fedu%5Fau%2FDocuments%2FHI%2FDecomposition%2Fdata&ga=1) | Training (ground truth) | 512 x 512 | F\_cnm ground truth for training |
| [R\_map\_n01\_pw02\_vs12\_Tcut\_500K\_512px.fits](https://anu365-my.sharepoint.com/personal/u1129269_anu_edu_au/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fu1129269%5Fanu%5Fedu%5Fau%2FDocuments%2FHI%2FDecomposition%2Fdata%2FR%5Fmap%5Fn01%5Fpw02%5Fvs12%5FTcut%5F500K%5F512px%2Efits&parent=%2Fpersonal%2Fu1129269%5Fanu%5Fedu%5Fau%2FDocuments%2FHI%2FDecomposition%2Fdata&ga=1) | Training (Ground truth) | 512 x 512 | R\_hi ground truth for training |
| [training\_data\_kok14\_galfa.pickle -> X\_test](https://anu365-my.sharepoint.com/personal/u1129269_anu_edu_au/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fu1129269%5Fanu%5Fedu%5Fau%2FDocuments%2FHI%2FDecomposition%2Fdata%2Ftraining%5Fdata%5Fkok14%5Fgalfa%2Epickle&parent=%2Fpersonal%2Fu1129269%5Fanu%5Fedu%5Fau%2FDocuments%2FHI%2FDecomposition%2Fdata&ga=1) | Model Testing | 58 x 414 | spectra data for testing |
| [training\_data\_kok14\_galfa.pickle -> Y\_test](https://anu365-my.sharepoint.com/personal/u1129269_anu_edu_au/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fu1129269%5Fanu%5Fedu%5Fau%2FDocuments%2FHI%2FDecomposition%2Fdata%2Ftraining%5Fdata%5Fkok14%5Fgalfa%2Epickle&parent=%2Fpersonal%2Fu1129269%5Fanu%5Fedu%5Fau%2FDocuments%2FHI%2FDecomposition%2Fdata&ga=1) | Model Testing | 58 x 2 | F\_cnm and R\_hi ground truth for testing |

7. **Timeline**

**8. References.**

[1] Claire E. Murray, et, al, 2020, Extracting the cold neutral medium from HI emission with deep learning: Implications for Galactic.

[2] E. Saury, et, al, 2014, The structure of the thermally bistable and turbulent atomic gas in the local interstellar medium. arXiv:1301.3446v1.

[3] Heiles, C., & Troland, T. H. 2003a, ApJS, 145, 329—. 2003b, ApJ, 586, 1067.

[4] Murray, C. E., Stanimirovi ́c, S., Goss, W. M., et al. 2015, ApJ, 804, 89.