Chalmers University of Technology

SEEX16

# Planning Report

*Students:*

Fagrell Peter
peterfag@chalmers.se

Kollberg Jens
jensko@chalmers.se

Rasmussen Elias
raelias@chalmers.se

Redmo Axelsson Erik
redmo@chalmers.se

Svensson Oskar
ossve@chalmers.se

Ybring Alexander
ybring@chalmers.se

*Supervisors:*

Bjerkeli Per
per.bjerkeli@chalmers.se

Toribio Maria Carmen
toribio@chalmers.se

8 februari 2023

# Table of Contents

# 1 Title

Detecting potential observations of dust in protoplanetary disk winds.

# 2 Background

Atacama Large Millimeter/submillimeter Array (ALMA) is an observatory located at 5000m above the ocean in the north of Chile [1]. The 66 antennas work together to capture high resolutions images of everything from our own solar system, to star formation and distant galaxies. These images are then stored in the public ALMA-archive, which today holds over 60 000 unique observations [2]. It is often of interest to search the archives for one specific type of observation. One example of this is the search for dust in outflows from protoplanetary disks.

The process of star formation begins with clouds of dust and gas that collapses together into a protostar forming a protoplanetary disk around it [3]. These differ from each other in mass, size, age, brightness, distance from earth, orientation to earth etc. As the gas and dust falls into the center, outflows from the disk and protostar is formed. Material leaves the system through jets and winds that drags it along from the surrounding cloud which forms outflows. There is still much uncertainty regarding a variety of aspects about these outflows, in particular how the launching takes place and if dust can be lifted from the disks into the outflows. Only a few images of these dust extensions in the wind has been found. That is why there is an interest in developing a tool to easier search the archive for such features. Currently, there exists no effective way to find protoplanetary objects of interest other than manually sorting through vast databases. This approach is not feasible due to factors such as cost, personnel and physical constraints. Therefore an automated solution would be desirable. A group began this work in 2022 [3] and created a search-tree for the archives which will be used in this project.

This project and program will be useful mainly for scientist wanting to find specific observations. It can also be of interest for anyone in regard to how machine learning generally can be used to search for data where it would be too time-consuming to search manually.

# 3 Aim

The project strives to produce a program whose main task is to distinguish specific observations taken by the ALMA observatory. More exact, the sought for observations shall contain images of protoplanetary disks where the disks show a deviation from Gaussianity. If successful, these images will later be used by scientists to investigate the presence of dust in parts of the disks' outflow called winds or jets. The program shall in turn increase the efficiency regarding research on how solar systems and planets are formed. Moreover, the program shall be flexible enough to easily be adjusted for other similar science projects.

# 4 Tasks

The main task of the project is to develop software that can classify images from the ALMA archive of protoplanetary disks where dust might be carried out by its winds. The intention is to construct a supervised convolutional neural network (CNN) that when fully trained can satisfy the main task. Other approaches such as using an unsupervised CNN might be considered if the former approach does not show good results. To abstract the project and ease understanding, this main task has been divided into sub tasks.

Critical for the project's success is to have the capability of effectively sorting out data with possible objects of interest from the ALMA archive. What an object of interest is will change depending on the current task at hand.

Once the data has been extracted, it needs to be formatted to fit into a uniform data set. With the same resolution and size, images can be used in a training set to establish what should and should not be sought for.

Another important objective of the project is to address the issue of limited training data. A large training data set will need to be created with as few biases as possible. There are only a few known observations of dust emission through winds from protoplanetary disks, which results in a lack of positive class images for the training set.

If an appropriate result and standard is reached, the software and images of interest should be presented, made publicly available and free to use. With this achieved, the studies of protoplanetary disks and protostars could advance with less manual labor.

# 5 Limitations

This project will mainly focus on machine learning and image analysis. The project will not focus as much on the astronomy aspect as the group has no background in this field of study and also because the aforementioned focus points are more important for developing the model.

Only 2D fits files from the ALMA archive will be used in this project. These files are produced by collapsing 3D images in the frequency domain. This results in loss of data as the size of an image goes down from gigabytes to megabytes. The hardware used in this project is not powerful enough to handle the size of the 3D images, thus 2D images will be used.

Because of the hardware constraints, only a subset of the ALMA archive will be used to train and test the model. The subset's size will be maximized with our constraints in mind.

Some images will be thrown out if two points of interest in an image are too close together to be cropped into our standard dimensions. This should be a rare occurrence but it means the model might miss some protoplanetary disks with dust in the winds.

# 6 Method

In this project, a CNN will be created to find protoplanetary disks with extensions. The CNN should be able to filter out said objects from the extensive ALMA archive.

The first approach will be to collect data, train a supervised CNN and apply it on the ALMA archive. The network will be trained on recognizing images containing dust being carried out of the system in the directions of outflows from protoplanetary disks. If the project was a success or not with this approach will depend on the results from the CNN.

Another approach to this project could be to use an unsupervised CNN to classify the archive. It thrives on classifying large quantities of unlabeled data [4]. Here, success would depend on a class of protoplanetary disks with extensions being distinguished.

## 6.1 Data acquisition

Data for the project will be collected from the ALMA archive through the tool ALminer, which is a Python based package made to efficiently query, analyze and visualize the ALMA science archive. It also allows direct download of data for further image processing [5].

## 6.2 Supervised learning

At an early stage, certain known observations of dust being carried out will be used for the positive class to create the training set for the CNN. These will be manually downloaded to then be augmented for a large enough training data set. The same will most likely be done for a negative class, even though it might create extensive bias. Later on, a smaller portion of the archive could be extracted to train on, with a part of it saved for testing afterward.

For the augmentation, techniques like rotation and mirroring of images will be used. On top of this, the amount of noise in the images will be altered. With this approach, one image could be made into thousands to prevent a lack of data and overfitting [6].

## 6.3 Unsupervised learning

Unlike the supervised CNN that will be trained for a specific goal, the unsupervised CNN model will supply unbiased results due to the fact that it does not have any prior knowledge of the expected outcome. Since it groups objects by similarity, without the need of labels, it could discover patterns or structure in the data that later can be used to identify underlying relationships between the variables. Thus, generating results undiscovered by the first approach, i.e. supervised learning.

## 6.4 Evaluation

To evaluate the project's outcome, the results produced by supervised as well as unsupervised learning will be reviewed. Even though the underlying aim is to find dust being carried out of a protoplanetary disk by wind, whether this occurs or not is beyond the scope of this project. As long as an effective program is constructed, with the capability of filtering through the ALMA archive and extracting relevant images, this project should be deemed successful. With produced material to study, the claim that dust leaving such systems by wind could affect the planetary creation has the potential to be asserted. At the very least, the program should streamline the process of extracting relevant data from the ALMA archive using some degree of machine learning. Anything else it might achieve is a bonus.

# 7 Social and ethical aspects

Even though societal and ethical aspects are not a primary concern for this project, there are some points worth acknowledging.

One important factor is to be aware of the limitations of the model and to interpret the results accordingly. Machine learning models can suffer from biases that are introduced during the training process. For example, if the data used to train the model is biased towards certain types of objects or observations, the model will also reflect this bias in its identifications. This could lead to incorrect identifications or the overlooking of important objects. Transparency about any limitations or possible limitations of the model is desirable so that there are as few misconceptions about the results as possible.

Another point worth noting is the fact that a machine learning model will perform the classification as opposed to an astronomer. Astronomers have a deep understanding of the field and the physical processes involved, which enables them to make informed decisions about object classification. Machine learning models, on the other hand, rely solely on the data and patterns they are trained on, and may lack the contextual understanding that an astronomer would have. Additionally, the results from the model can be difficult to interpret and understand, which can make it challenging to determine why certain decisions are made. This can make it nontrivial to identify and correct errors in the system.

Furthermore, this project aims to adhere to the FAIR Guiding Principles for the scientific community [7]. These principles, which stand for Findable, Accessible, Interoperable, and Reusable, promote the reusability and transparency of the project. However, given the complex nature of scientific data and the practical limitations of the project, it may not always be possible to adhere strictly to the FAIR principles. In such cases, the principles are balanced against the practical considerations and limitations, while still maintaining the highest standards of data management and responsibility.

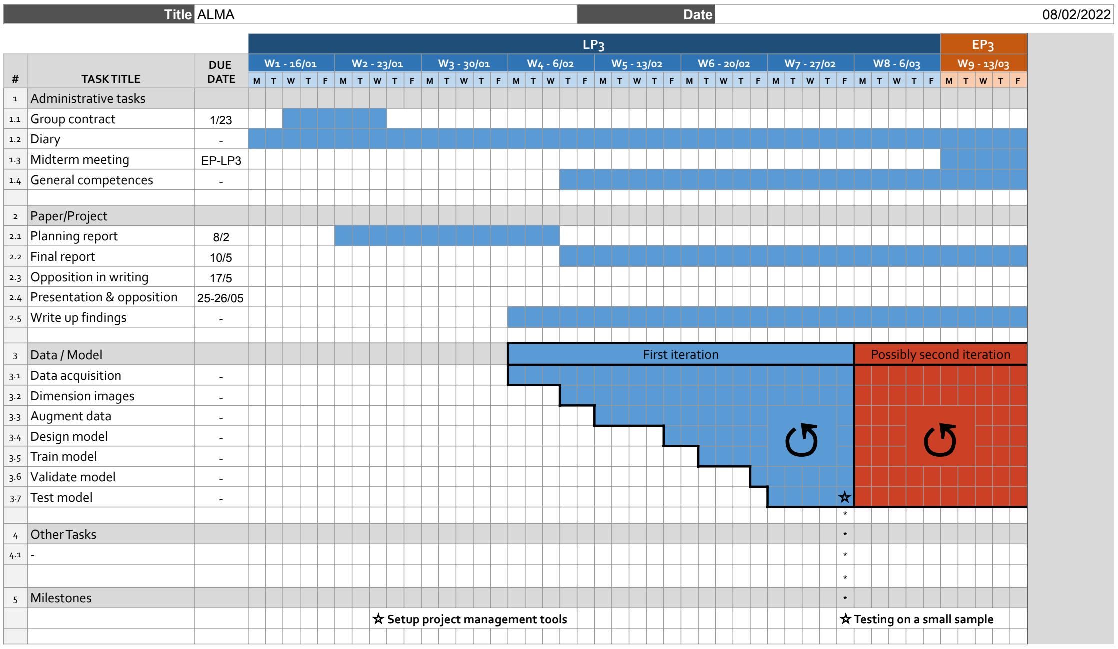
Another aspect worth taking into consideration is the risk with algorithms such as these influencing the choice of research objects for scientists. By creating an algorithm for the purpose explained in Aim, see section 3, the future research might be steered into a direction not present without such an algorithm. Certain errors made by the algorithm will also have an impact on researchers' choice of objects and studies.

# 8 Timetable

This chapter presents a Gantt chart that visually represents the schedule. The timeline outlines the major tasks and milestones required to complete the project, including anticipated due dates. The main motive of using a Gantt chart is to visualize the project timeline and task dependencies, as well as expected progress throughout the project.

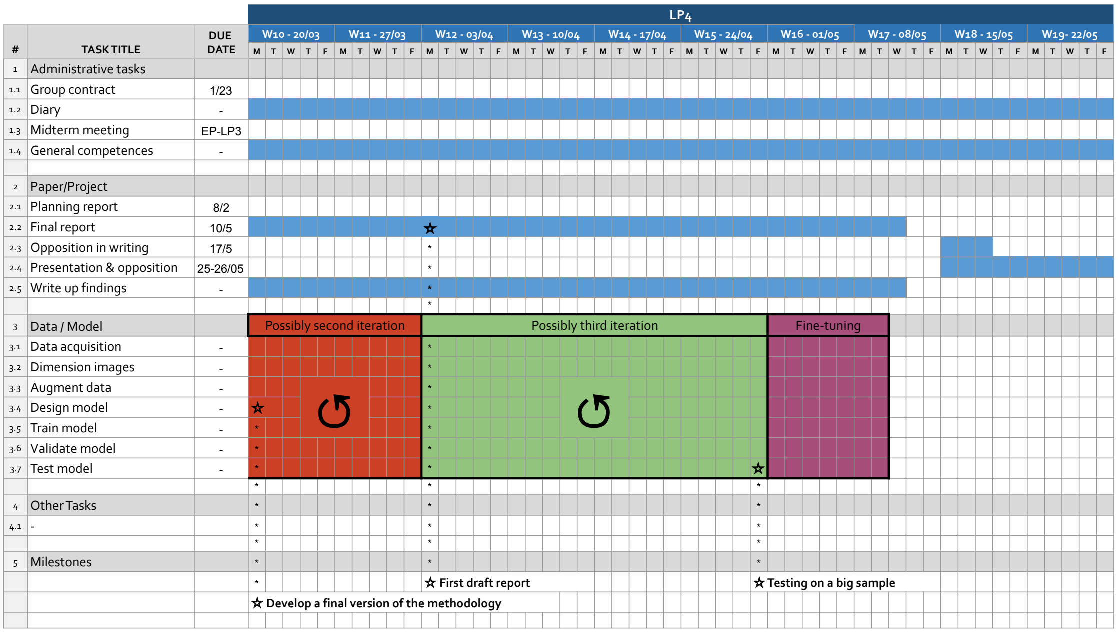
The idea, as one can tell from the Gantt chart shown in the appendix, see section 9, is to carry out the process involving the model iteratively. What this means is that the steps involved in each iteration build on the results of the previous one. These steps include data acquisition, dimensioning and augmentation of images, designing the model, training, validating, testing, and finally refinement of the model. By continuously improving the model through these iterations, the team can achieve the desired performance.

# 9 Appendix



| Title | ALMA | Date | 08/02/2022 |
|---|---|---|---|

| # | TASK TITLE | DUE DATE |
|---|---|---|
| 1 | Administrative tasks | |
| 1.1 | Group contract | 1/23 |
| 1.2 | Diary | - |
| 1.3 | Midterm meeting | EP-LP3 |
| 1.4 | General competences | - |
| 2 | Paper/Project | |
| 2.1 | Planning report | 8/2 |
| 2.2 | Final report | 10/5 |
| 2.3 | Opposition in writing | 17/5 |
| 2.4 | Presentation & opposition | 25-26/05 |
| 2.5 | Write up findings | - |
| 3 | Data / Model | |
| 3.1 | Data acquisition | - |
| 3.2 | Dimension images | - |
| 3.3 | Augment data | - |
| 3.4 | Design model | - |
| 3.5 | Train model | - |
| 3.6 | Validate model | - |
| 3.7 | Test model | - |
| 4 | Other Tasks | |
| 4.1 | - | |
| 5 | Milestones | |



| # | TASK TITLE | DUE DATE |
|---|---|---|
| 1 | Administrative tasks | |
| 1.1 | Group contract | 1/23 |
| 1.2 | Diary | - |
| 1.3 | Midterm meeting | EP-LP3 |
| 1.4 | General competences | - |
| 2 | Paper/Project | |
| 2.1 | Planning report | 8/2 |
| 2.2 | Final report | 10/5 |
| 2.3 | Opposition in writing | 17/5 |
| 2.4 | Presentation & opposition | 25-26/05 |
| 2.5 | Write up findings | - |
| 3 | Data / Model | |
| 3.1 | Data acquisition | - |
| 3.2 | Dimension images | - |
| 3.3 | Augment data | - |
| 3.4 | Design model | - |
| 3.5 | Train model | - |
| 3.6 | Validate model | - |
| 3.7 | Test model | - |
| 4 | Other Tasks | |
| 4.1 | - | |
| 5 | Milestones | |

# References

[1] ALMA Observatory. Atacama Large Millimeter/submillimeter Array; 2023. Available from: `https://www.almaobservatory.org/`.

[2] Atacama Large Millimeter/submillimeter Array. ALMA Science Archive;. Available from: `https://almascience.nrao.edu/aq/?result_view=observations`.

[3] Hjält M, Larsson C, Rosén A, Thim L, Thure T. Studie av molekylära utflöden från protoplanetära skivor En systematisk arkivstudie av observationer från ALMA-teleskopet Kandidatarbete inom Rymd-, geo-och miljövetenskap;. Available from: `www.chalmers.se`.

[4] Alex Ross. What Is Unsupervised Learning?; 2021. Available from: `https://unsupervised.com/resources/blogs/what-is-unsupervised-learning/`.

[5] Ahmadi A. ALminer: ALMA Archive Mining & Visualization Toolkit; 2023. Available from: `https://github.com/emerge-erc/ALminer`.

[6] What Is Image Augmentation; 2023. Available from: `https://albumentations.ai/docs/introduction/image_augmentation/`.

[7] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016 3;3.