Design of a machine learning model for the characterisation of young planets from dust morphologies in discs

Alessandro Ruzza, n° di matricola: 931750

During the last decades, breakthroughs in interferometric observations and the advent of new telescopes, such as the Atacama Large Millimiter Array, led to the first high resolution observations of protoplanetary discs, the planets' birthplace. These images have to be interpreted to characterise the observed discs and the embedded planets. Currently used approaches present some limitations which keep the problem open to improvements and new solutions. In this thesis we propose the adoption of machine learning methods for this purpose. More specifically we focused on neural networks, a class of algorithms that can learn underlying patterns in the available complete data and later apply the acquired model to analyse partial data predicting the unknown features. We thus collated synthetic images obtained from numerical simulations implementing a dataset aimed at this purpose, which was made publicly available along with a python package and some scripts developed for its usage and mantainance. We also provided a proof of concept in support of the new approach proposed.

Protoplanetary discs, usually observed at hundreds of parsec in the star-forming regions, are rotating discs of gas and dust revolving around a central star. They are a late product of the star's accretion disc evolution when most of its mass has accreted onto the central star and most of the energy released by the gravitional collapse that started the process has been exhausted through the disc emission which is now mainly due to the irradiation it receives from the star. In this enviroment planet formation takes place, profoundly shaping the distribution of the disc matter with the formation of peculiar substructures that exist in a wide variety of morphologies. For example, it has been shown that the gravitational interaction with a planet can originate gaps, i.e. annular regions of depletion (eg. Bryden et al. 1999; Dipierro et al. 2016).

In this thesis we focused on observations of the dust thermal emission in the millimetric region of the spectrum. The images obtained consist in spatially resolved maps of the flux density at specific wavelengths. From them, all the other properties of the disc and embedded planets have to be inferred to obtain a complete characterisation. The models describing the dynamics and optical properties of these objects depends upon numerous parameters, such as the gas and dust total masses, their ratio, viscosity and so on. These studies allow the development and test of theories regarding protoplanetary discs and planet formation offering a statistically valid pool of data beyond, in the latter case, the sole information obtainable from the solar system.

Among the different variables that could be inferred we chose to focus, as an interesting example, on the masses of gap opening planets. The methods presented can, however, be applied, with the appropriate changes, to different problems. The width of gaps has been linked through some empirical formulae to the masses of the planets responsible for their origin (Lodato et al., 2019; Kanagawa et al., 2015). Both these and more generally all the empirical relations proposed to determine specific disc features, are usually simple polynomial or power laws, which allow a fast computation of unknown properties and can be easily explained with theoretical arguments. On the other end, they lack in accuracy and precision requiring the use of more advanced tools. Numerical simulations are today's response to these needs. Physical models that describe the dynamics, particle interaction and radiative emission at a small scale are applied to simulate the overall disc evolution, and synthetize images analogous to the observations that would be obtained if the system was real. These results can thus be directly compared with the real data to check the assumptions made to start the simulation and select the parameters which provide the best fit.

The main downside of this approach resides in the computational complexity of the simulations which could take hours to complete. A typical disc characterisation requires at least hundreds of them to properly explore the hyperspace of likely values for the unknown disc features. This method is thus unfeasible for large sets of images which could be obtained, for example, in future surveys.

Neural netoworks could combine the convenience of analytical expressions with the accuracy and ability to work directly on the observed data of the numerical method. These types of algorithms are modelled on the functioning of a biological brain implementing elementary units, i.e. artificial neurons, connected in a specific structure (Mehta et al., 2019). In this thesis we used a feedforward neural network proposing a future improvement with convolutional layers. Hiding the details, a neural network is just a black box which takes an input and returns a related output with the ability to learn, in a previous step called training process, the correct relation from data. We propose, for example, to use neural networks to analyse the entire image of a protoplanetary disc presenting a gap, which will be provided in input, and return the mass of the planet that could have carved that gap, which will be the output of the network.

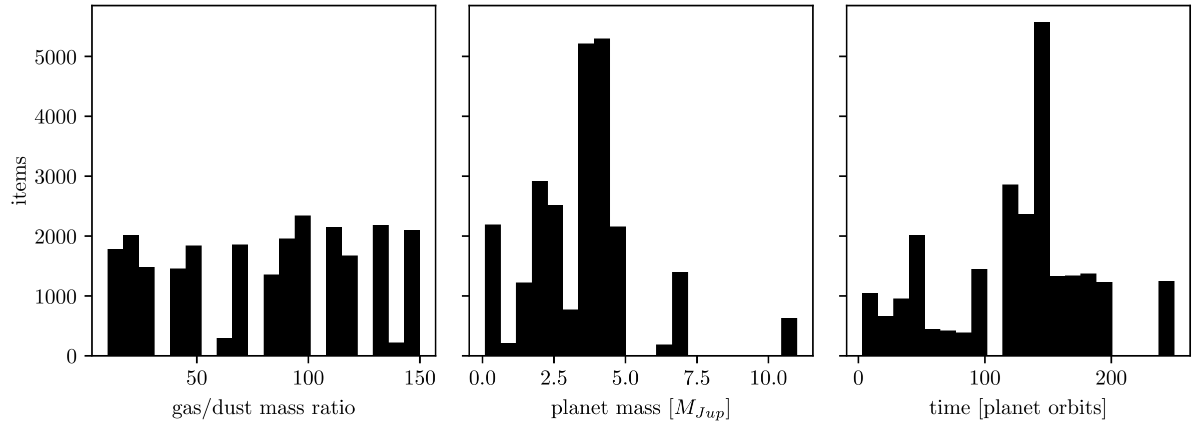

Figure 1: Histograms showing the distributions of values, among the dataset, for the three parameters more significantly explored. From left to right, these parameters are: the gas to dust mass ratio, the mass of the embedded planet and the number of orbits completed by the planet during the hydrodynamical simulation before saving the snapshot. The total number of items is 24720.

The training process requires a set of data where the expected value for each observation is known. The main objective of this thesis was the design of this dataset in both its content and interface. We collated fictious images of protoplanetary discs obtained through numerical hydrodynamical and then radiative simulations, with an additional step aimed at reproducing the limited resolution of the telescopes, which consisted in the convolution of the images with Gaussian beams of different dimensions. Since one of the downsides of the numerical approach is its high computational cost, running all the simulations from the beginning would partly defeat the purpose of proposing this new method. We thus designed the dataset on a collaborative basis, granting open access to this resource, and allowing people working in the field to contribute providing the results of any numerical simulations that they could have performed in the context of their research. In the version deployed with this thesis, the results of the first step, and for some of them also of the second step, were kindly shared with us by two researchers. We then run the remaining steps of the simulations.

In total, we collated 24720 different images significantly exploring the three parameters in figure 1 (gas to dust mass ratio, planet mass and time of the snapshot) where the distribution of the explored values is shown. Additionally, the dataset contains images depicting discs from three main inclinations (0°, 30° and 60°) and convolved with Gaussian beams of ten different sizes, in order to reproduce the current best resolutions achievable with existing observatories and to explore the upper and lower limits. The data of each image is stored in a separate file in the `FITS` format whose header stores all the parameters characterising the disc, which are known from being used to perform the simulations. They are then indexed by two files. One of them contains only the basic information needed to uniquely identify each item while the other one stores all the information saved in the files' headers.

This structure was conceived in the perspective of making the dataset publicly available online. In this scenario the user would have first to download the index to obtain the remote location of each image to retrieve the entire database or, alternatively, download also the second index file that contains all the disc's parameters, to apply a first selection and download only the desired subset of items.

In addition to the storage structure, we also developed a python package containing 3 submodules. The first one implements functions that can be used to download, partly or completely, the dataset checking the integrity of the retrieved files. The second submodule handles the index files providing functions for reading their data returning a python dictionary and for adding new entries with an automated process for large collections of `FITS` files. The last one contains functions aimed at assembling the dataset and preprocessing each image obtained from the radiative simulation, which is done retrieving the parameters stored in different files, convolving the image with the gaussian beams and, eventually, generating the file where all these data are stored. Some python scripts were also developed to interface these functions from the command line. Both the database and the code developed are published in two GitHub repositiories accessible at this link: `https://github.com/dust-busters/`.

Finally, we presented a proof of concept implementing a feedforward neural network, using TensorFlow libraries, to infer the planet's mass as previously explained. Note that all the images in the dataset depict discs with an embedded planet that opened a gap. We trained the model selecting, from the entire database, a subset of images all showing the disc at the same inclination and all convolved with a Gaussian beam of the same size, chosen in accord to the modern best observations. This selection was done both to reduce the computational complexity

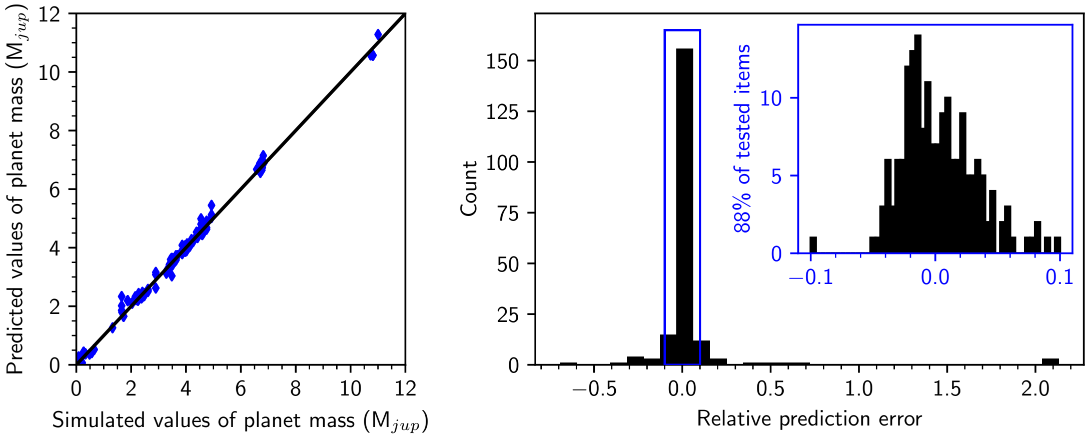

Figure 2: Left panel: correlation between the predicted planet's mass and the real value used to run the simulation for all the items in the test set. The black line draws the $M_{pred} = M_{sim}$ curve. Right panel: distribution of the relative prediction error $(M_{pred} - M_{sim})/M_{sim}$ computed for the test set. The blue box zooms in the -10% to 10% range which contains the 88% of all the items. This distribution exhibits mean and standard deviation of respectively 4% and 28%. The root mean squared error is $0.14 M_{jup}$.

required and to demonstrate how the database can be easily manipulated using the developed interface. The 1024 images selected in this way were divided, assigning respectively the 80% and 20% of the whole set to each group, among a training set used during the training process and a validation set. The elements of this last group were not used during the training process allowing to eventually test the obtained model on never seen images.

Figure 2 shows the final results obtained testing the trained model on the validation set. The histogram shows the distribution of the relative error which exhibit a mean value of 4% and a standard deviation of 28%. The planet's masses are thus slightly overestimated, and the errors are highly variable depending on the analysed image. The mean value of the absolute relative errors computed on the test set is 8% while the root mean square error is 0.14M$_{Jup}$. We considered this last metric the prediction uncertainty of our model.

We also tested the trained model on the real image of the DS Tau disc. Veronesi et al. (2020) showed using numerical simulations the presence of an embedded planet estimating a mass of $3.5 \pm 1 M_{Jup}$ which is in fair agreement with the mass predicted by our neural network of $4.12 \pm 0.14 M_{Jup}$. This estimate has been computed averaging the results obtained with three identical neural networks, which have been trained on the same set but shuffled in different ways. Notice the significant reduction of the uncertainty.

The results obtained are thus promising and successfully proved the feasibility of our approach. Future works will have to focus on the expansion of the dataset, encouraging contribution from the community, and on the study of more sophisticated models of neural networks exploring also different applications in the same context.

References

G. Bryden et al. Tidally induced gap formation in protostellar disks: Gap clearing and suppression of protoplanetary growth. *The Astrophysical Journal*, 514(1):344–367, mar 1999. doi: 10.1086/306917.

Giovanni Dipierro et al. Two mechanisms for dust gap opening in protoplanetary discs. *Monthly Notices of the Royal Astronomical Society: Letters*, 459(1):L1–L5, 03 2016. ISSN 1745-3925. doi: 10.1093/mnrasl/slw032.

Kazuhiro D. Kanagawa et al. MASS ESTIMATES OF a GIANT PLANET IN a PROTOPLANETARY DISK FROM THE GAP STRUCTURES. *The Astrophysical Journal*, 806(1):L15, Jun 2015. doi: 10.1088/2041-8205/806/1/l15.

Giuseppe Lodato et al. The newborn planet population emerging from ring-like structures in discs. *Monthly Notices of the Royal Astronomical Society*, 486(1):453–461, Mar 2019. ISSN 1365-2966. doi: 10.1093/mnras/stz913.

Pankaj Mehta et al. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 810:1–124, May 2019. ISSN 0370-1573. doi: 10.1016/j.physrep.2019.03.001.

Benedetta Veronesi et al. Is the gap in the DS Tau disc hiding a planet? *Monthly Notices of the Royal Astronomical Society*, 495(2):1913–1926, May 2020. ISSN 0035-8711. doi: 10.1093/mnras/staa1278.