Property Determinations of Open Clusters using Hierarchical Bayesian Modelling and Neural Networks

H. Westwood, ID: 1556764  
H. Leung, ID:  
27/03/2020

**Abstract:**

**Context:** Open clusters are very useful for age and metallicity estimations, which makes that excellent for tracing the evolution of galaxies. Current methods of estimating open cluster properties are time consuming and heavily rely on literature values. We put forward that a more efficient method is by using a framework consisting a neural network inside an HBM.   
**Aims:** To determine accurate properties of open clusters, using a neural network to convert fundamentals and observables, which can be used by a hierarchical Bayesian model to simultaneously constrain the fundamental variables and give population wide mean values of these properties.   
**Methods:** We trained a deep feedforward neural network to convert fundamentals: age, mass, [Fe/H], initial helium and mixing length to observables: effective temperature, stellar radius and Δν using a grid of stellar evolution tracks calculated using MESA (B. Paxton et al. (2010) [MESA]). This was then incorporated into an HBM and property estimations were made using this framework given suitable priors on the fundamentals and observational data for different clusters, taken from the literature.   
Results:  
Conclusion:

Estimated Word Count:   
Key words: neural network, HBM, Bayesian, M67

Contents:

Nomenclature:

**Introduction:**  
Open clusters are one of the most important types of objects when it comes tracing the evolutionary history of our galaxy, which is still very much an open topic of research as making accurate property determinations can be quite challenging but is crucial to figuring out the mechanisms behind galactic evolution. Open clusters are loosely gravitationally bound groups of typically a few hundred coeval stars, meaning all the stars formed at approximately the same time in a relatively small region of a galaxy where there was a particular chemical composition. The result of this being that open cluster members can be treated as having the same age and [Fe/H] (hereafter metallicity). This feature is hugely beneficial for property determinations over single star measurements as by measuring open cluster members the mean age and metallicity of these effectively converges to the true values (see \cite{Hippel\_2005} and \cite{salaris\_2004} for more details). The most common method of doing this is by isochrone fitting, where an isochrone is a plot on a Herzsprung-Russell which describes stars of the same age, and typically also keeps the chemical composition constant. Isochrone fitting requires sampling ages and metallicities and converting those fundamentals parameters to observables (effective temperature, luminosity etc.) such that they can be compared to the observed data for the open cluster. Depending on the method the isochrones are evaluated to determine how well each isochrone fits and the best isochrone’s age and metallicity are taken to be true for the cluster. However, certain choices for the method of sampling and parameter conversion tool, can cause the fitting process to be incredibly time consuming. There are a variety of sampling methods from fitting by-eye \citep{brandt \_2015} to automated fitting processes using software \citep(perren\_2015) and Bayesian analysis (\cite{hills\_2015}, \cite{jeffery\_2016}, \cite{jørgensen\_2005}). Which compounds upon the time for the conversion process from fundamentals to observables which in most studies to date has been done using stellar evolution models to generate the isochrones via frameworks like MESA (\cite{ball\_2017}, \cite{ball\_2018}, \cite{MESA}) and DSED (\cite{jeffery\_2016}, \cite{DSED}). These require generating a stellar evolutionary track for each sampled age and metallicity which can take hours per track generated, although some studies (see \cite{jørgensen\_2005}) skirt this by using a pre-existing set of tracks and interpolating between them but raises different problems. There are other open cluster dating methods like detached eclipsing binaries (\cite{brewer\_2016} and \cite{bavarsad\_2016}) and white dwarf cooling (\cite{kalirai\_2001}, \cite{bedin\_2015}) but these outside the scope of this study.

The purpose of this study is to improve upon the isochrone fitting methods stated above by avoiding lengthy parameter conversion and track interpolation by training a neural network on MESA tracks to convert fundamental to observable parameters (SEE NEURAL NETWORK SECTION FOR MORE DETAILS). The sampling is to be done using a hierarchical Bayesian model (HBM) where literature values inform the sampling such that minimal time is wasted sampling in parameter space unlikely to yield the true values. This method should allow for sampling of open clusters to be done much faster with the added benefit that the HBM allows for simultaneous constraint of a large number of variables and is able to estimate the spread of those variables, which haven’t been measured before for open clusters (SEE THE HBM SECTION FOR FURTHER DETAILS).

**Target selection and data collection:**  
We selected 6 open clusters (NGC 2682 also known as M67, NGC 6791, NGC 6819, Ruprecht 147, NGC 752, NGC 188) which have a large range of ages and metallicities that have been relatively well studied, see Table 1 for details on the results of previous studies.

To select stars to study from these open clusters we began by finding membership studies (calculations of the probability that a star belongs to a particular open cluster), from which we chose the membership study of M67 done by \cite{gao\_2018} and for the other 5 clusters we used the work of \cite{cantat-gaudin\_2018}.

What follows are the steps and measures used to discard stars and gather data to calculate luminosities.   
- These 2 membership papers contained GAIA DR2 IDs, which allowed us to query GAIA DR2 (\cite{GAIA\_mission}, \cite{GAIA\_DR2}) to get GAIA G-band apparent magnitudes and distance estimates for the stars calculated by \cite{bailer-jones\_2018}. We then discarded stars without G-band measurements.  
- Effective temperatures, Teff, are crucial for this study so we wanted to get accurate and consistent Teff measurements whenever possible. Some of the most accurate Teff determinations have been made using the infrared spectroscopy of the Apache Point Observatory Galactic Evolution Experiment (APOGEE) \citep{APOGEE}, which we queried through SDSS DR12. Unfortunately, APOGEE didn’t have effective temperatures for all the stars we were considering, so for the remaining stars we queried the SIMBAD database and discarded stars without a Teff from APOGEE or SIMBAD.   
- We then selected the greatest number of stars from the OCs that gave a mean membership uncertainty of 1%.   
- For stars that didn’t have APOGEE Teffs but multiple literature Teffs, we counted anomalies as stars whose Teffs differed from the median literature value by greater than 90k, which were removed. The Teff of these stars were then selected by the paper which provided the largest number of Teffs for that OC of the papers that provided Teffs for each star.   
- The absolute GAIA G-band magnitude was calculated as follows:

$ M\_G = m\_G + 5 - \log\_{10}(r) - A\_G $ eq.1

where r is the median of the distance estimates for the cluster and $A\_G$ is the line-of-sight extinction ignoring the effects of reddening, calculated by:

$ A\_G = R(G)E’ $ eq.2

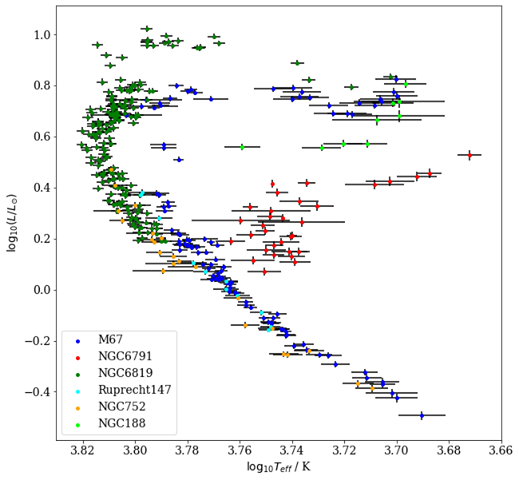
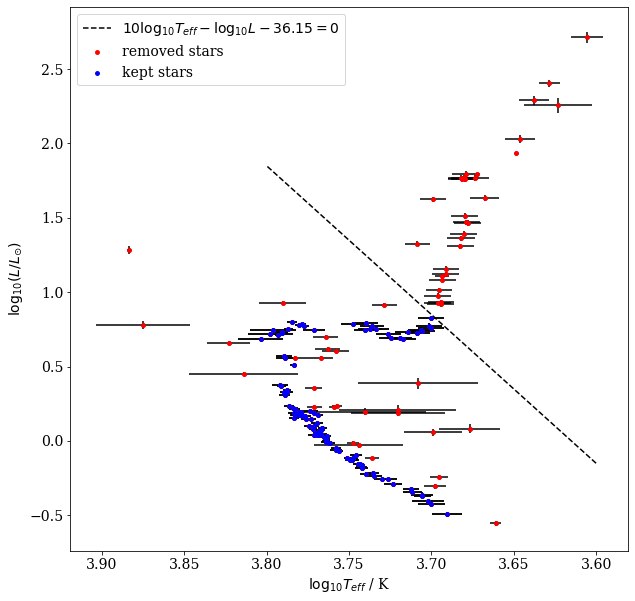
Where R(G) is the GAIA G-band extinction coefficient (2.294) determined by \cite{sanders\_2018} and E’ is the dust reddening taken from the best fit values in the bayestar2019 dust map. The GAIA G-band bolometric correction, $BC\_G$, was taken from a model dependent only Teff, by \cite{andrae\_2018}. From which we could get the bolometric magnitude:

$ M\_Bol = BC\_G + M\_G $ eq.3

And then we find the luminosity (in units of solar luminosities) by:

$ L = 10^{-0.4(M\_{Bol} - M\_{Bol, \odot})} $ eq.4

Where $M\_{Bol, \odot}$ is taken to be 4.75 \citep{MBOLSOL} We then plotted the HR-diagram and removed blue stragglers, binaries and other stars we thought might interfere with the HBM sampling. We removed the blue stragglers because they cannot be fitted to an isochrone as their properties evolve in an atypical way. Binaries are removed because their luminosities shift them far away from the line an isochrone would be fitted to and are removed like the other anomalies to assist with HBM sampling. A further cut that we made is removing the RGB stars which was done due to issues with neural network training (the reason for this is elaborated on further in the Neural Network section). The open clusters after star removal are shown in Figure 1. We also checked that after star removal the mean uncertainty per star for each cluster was still below 1% (which should be a table in the appendix)

   
**Figure 1**: HR diagram of the selected open clusters after removing blue stragglers, binaries, anomalies and the RGB (left). HR diagram of M67 with indication of removed stars (right).

As shown in Figure 1, NGC 752 and Ruprecht 147 don’t have sub giant branches with the selected stars, which is likely because they are the 2 youngest selected open clusters as shown in Table 1. These clusters should prove valuable in testing our methodology because without a sub giant branch fitting the isochrone to the clusters should be more difficult as the shape of the isochrone in this region is unclear.

**Table 1**: Literature search on the open clusters in this study. NC is not calculated i.e. the value was assumed for use in isochrone fitting.

|  |  |  |  |
| --- | --- | --- | --- |
| cluster | Age estimation, Gyr | Metallicity, [Fe/H] | Helium abundance |
| NGC 2682 (M67) | 3.64 \citep{Bossini}  3.45 ± 1.13 \citep{Netopil\_2016} 4.05 ± 0.05 \citep{jørgensen\_2005} | 0.03 \citep{Netopil\_2016} | 0.248 \citep{viani\_2017}  0.28 \citep{Dinescu\_1995} NC |
| NGC 6791 | 8.45 \citep{Bossini}  8.2 ± 0.3 \citep{mckeever\_2019} | +0.42 \citep{Netopil\_2016} | 0.297 ± 0.003 \citep{ mckeever\_2019}  0.30 ± 0.01 \citep{brogaard\_2012} NC |
| NGC 6819 | 2.00 \citep{Bossini}  2.11 ± 0.44 \citep{Netopil\_2016} | +0.09 \citep{Netopil\_2016} | 0.28 \cite{miglio\_2011} NC |
| Ruprecht 147 | 2.5 ± 0.25 \citep{curtis\_2013}  3 \citep{curtis\_2013}  3.25 \citep{curtis\_2013} | +0.08 \citep{bragaglia\_2018} | - |
| NGC 752 | 1.48 \citep{Bossini}  1.69 ± 0.66 \citep{Netopil\_2016} | -0.03 \citep{Netopil\_2016} | 0.26 \cite{Dinescu\_1995} NC |
| NGC 188 | 7.53 \citep{Bossini}  5.78 ± 0.03 \citep{hills\_2015}  6.45 ± 0.04 \citep{hills\_2015} | +0.11 \citep{Netopil\_2016}  +0.125 ± 0.003 \citep{hills\_2015} | 0.28 \cite{Dinescu\_1995} NC |

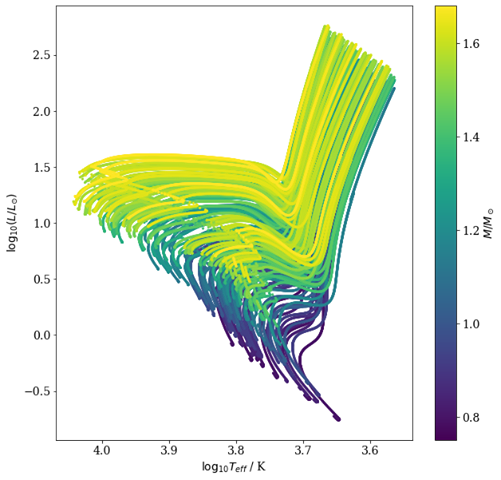
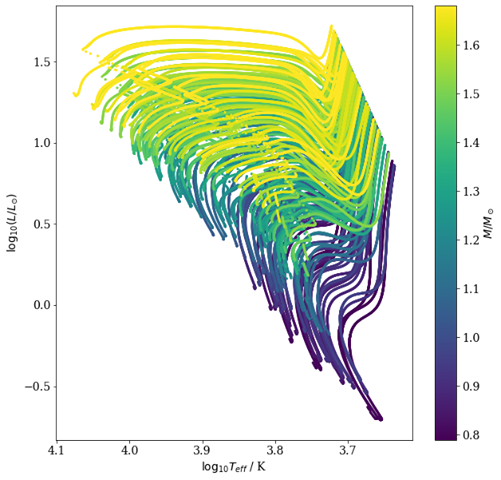
One of the features of the selected open clusters is that all the clusters are believed to be older than 1 Gyr. This is due to us wanting a large spread of stars along the isochrone to be informative for the HBM, but for clusters younger than 1 Gyr for stars to be sufficiently evolved must have high masses which exceed the mass range in the training grid shown in Table 2. It is also possible for younger clusters to be harder to observe due to large amounts of dust not having been used for star formation in the initial molecular cloud, which haven’t been dispersed through cosmic feedback processes which require post He-core burning stars to achieve.

**MESA stellar model**  
To generate a grid of stellar tracks to train a neural network on we chose to use Modules for Experiments in Stellar Astrophysics (MESA), which uses a specific model of stellar physics in order to evolve stars from some initial conditions and give their properties at a range of points across their lifetime (see \cite{MESA}, \cite{MESA2} and \cite{MESA3}). As our understanding of stellar evolution isn’t perfect, our stellar models aren’t perfect and as such there are simplifications must be made. An example is that the grid used in this study doesn’t account for stellar rotation which has knock on effects on temperature, luminosity and asteroseismic values (among others). However, studies like this assists in refining models by testing them.

**Table 2**: MESA tracks initial input parameter ranges

|  |  |  |
| --- | --- | --- |
| Input Parameter | range | increments |
| Age / Gyr | 0.04 - 32 | varied |
| M\_{init} / M\_{\odot} | 0.8 - 1.68 | 0.04 |
| [Fe/H]\_{init} | -0.8 - 0.6 | 0.2 |
| Y\_{init} | 0.24 - 0.32 | 0.02 |
| αMLT | 1.7 - 2.3 | 0.2 |

A more detailed description of the construction of the MESA grid can be found in \cite{li\_2018}, the only difference being in the initial input parameter ranges, shown here in Table 2.

   
Figure 2: HR diagram of 200 tracks from the training grid, with RGB (left) and without RGB (right). The reason why the right plot still has part of the lower RGB is because of neural network training and will be covered further in the neural network section.

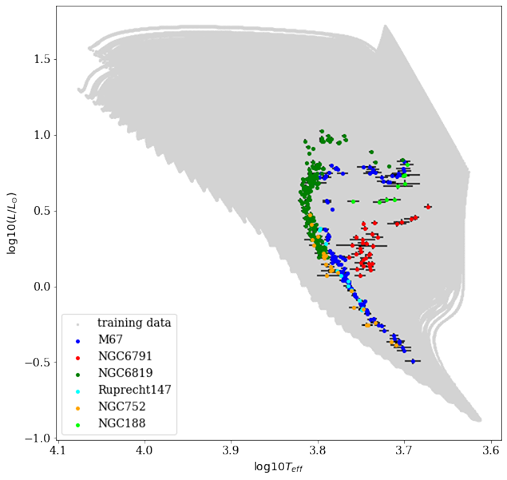


Figure 3: HR diagram of the selected open clusters, shown on top of the training data.

Figure 3 helps demonstrate that the selected open cluster’s true parameters can be found within the region covered by the training grid. Although it is possible that combinations of input parameters could still produce an isochrone that lies within the shaded region of Figure 3 but which is not covered by the input parameters of the grid, however parameter space covered by the training grid does cover metallicity estimates for the clusters from the literature shown in Table 1. As a result we feel confident that the training grid covers a large enough of age, metallicity and initial helium (initial helium never really goes above 0.32).

**Working with Neural Networks:  
 Machine Learning and Neural Networks:**  
Machine learning is the term used to describe a broad spectrum of statistical tools used to identify, reproduce or learn features of different kinds of trends. There are many different types of machine learning methods because not all methods are suitable for the limitless number of machine learning applications. For instance, in this study we will focussed on supervised machine learning where we present a trend that we want the neural network learn features of. This is in the form of the MESA grid where we have a set of input parameters we know map onto the output parameters as shown by the same trends in the MESA tracks generated from stellar fundamentals in Figure 2 having the same evolutionary shapes as the observational data shown in Figure 1. Within supervised machine learning there are the classification and regression subsets which both have a place in astronomy with classification methods using both supervised and unsupervised methods being useful for morphological classification (for a discussion of supervised methods see \cite{cheng\_2020}). Regression is quite different to this we want to reproduce the trends, as opposed to classification which learns the trend such that it can sort say images of galaxies into morphological groups. With regression methods, trends are learned and can be reproduced to predict points that the machine learning algorithm has never handled before but having learned the trends it knows how it should be handled.

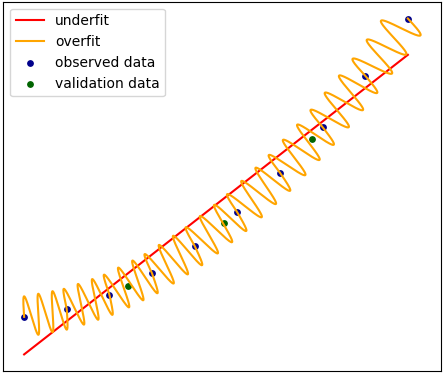
There have been a number of studies applying regression machine learning methods to best fitting stellar properties, like that done by \cite{hendriks\_2019} who used a deep neural network trained on a grid calculated using MESA to model asteroseismic properties and achieved consistent results to previous studies but using much less time to do so in a way that can be generalised to thousands of stars. These results are very similar to those we intend for this study but instead generalizable for many open clusters, which gives us confidence that at least our work with neural networks especially that with asteroseismic properties can be done successfully to at least the accuracy of previous works. There are also studies on just stellar fundamentals of main sequence stars using machine learning like that of \cite{bellinger\_2016} who use a random forest method that also uses MESA and again found that using machine learning they could achieve results consistent with previous work but using much less time. (add more examples of ML in literature?)

The machine learning of this study is focussed solely on using a deep, fully connected, feedforward neural network to construct what that is, I need to explain what an artificial neural network is. Artificial neural networks, commonly just referred to as neural networks are a set of machine learning methods which consist of a set of layers within which are some number of nodes also known as neurons hence neural network which have connections between subsequent layers. The simplest neural network requires an input layer consisting of a number nodes equal to the dimensionality of the input, similarly for an output layer and 1 layer in between the input and output known as a hidden layer which for this example may only have 1 node which would be connected to all the input and output nodes. The number of nodes and hidden layers and the way they are connected describes the “architecture” of the neural network. The significance of the size of the architecture is in that a larger architecture allows the neural network to reproduce more complex functions as nodes act almost like terms of a polynomial and so more nodes allows a higher order polynomial to be reproduced. Nodes are able to do this because for all non-input layer nodes, the value of each node is calculated using the values of the nodes in the preceding layer it is connected to using $\text{Output} = f(\sum\limits\_{i}^{N} W\_{i}I\_{i} + B)$. For simplicity this shall be for a fully connected neural network where each node in a non-input layer has connections to every node in the preceding layer. N is the number of nodes in the previous layer the current node is connected to, $ I\_{i} $ is the output of the ith node in the preceding layer, $ W\_{i} $ is the weight of the connection between the ith node of the preceding layer and current node, which describes the significance of the ith node relative to the other n-1 inputs to the current node and B is an additional input to the current node known as the “bias” which acts as a vertical translation variable. The function f which acts over the top of the weights, node inputs and bias is the activation function which in this case will be a non-linear activation function which works to penalize certain node outputs, typically negative node outputs are handled by non-linear activation functions by decreasing the weight of the connection of the current node to those in the succeeding layer which due to the reduction in the weight of the current means that it will be less informative to nodes of the next layer. Having explained somewhat of how a neural network functions I can now explicitly describe a deep, fully connected, feedforward, neural network as an artificial neural network, which uses deep learning meaning the neural network has more than one hidden layer, is fully connected where all nodes in a non-input layers have connections to all nodes of the preceding layer and it’s feedforward which is where the information flow from the input to the output.

Having covered the basic construction of a neural network the next stage is to explain how they are trained. Here I shall cover how to train neural networks using grid modelling: the grid being an array of possible input parameter values and the corresponding output parameters that you would expect/want the neural network to predict given the inputs, contains the data with which the neural network is able to learn and ideally improve to some desired accuracy by altering the weights and biases of its nodes to reproduce the trends in the data and enable it to make predictions of points not found in the training grid. The way in which the neural network makes decisions on how to improve its configuration by changing weights and biases is done on incremental steps of training called “epochs” and during an epoch the neural network will do backpropagation which allows it to determine the “complete error gradient vector”, $ \frac{\partial e}{\partial w} $ where $e$ is the error of the neural network known as the “loss function”, and $w$ is the set of neural network weights. The change to each weight can then be found from the negative of $ \frac{\partial e}{\partial w} $ from:

$ \Delta w = -\mu \frac{\partial e}{\partial w} $ eq.5 \citep{Bailer\_Jones\_2002}

Where $\mu$ describes the size of the weight update, typically referred to as the “learning rate”. A similar method of backpropagation is used to update the biases (for a more detailed description of backpropagation see \cite{NNbook}, \cite{Bailer\_Jones\_2002} and the references therein).

**Underfitting and Overfitting**  
By training for some number of epochs there are 2 extreme conditions where training the neural network won’t improve and may in fact become worse, both of which are related to the size of the neural network architecture (among other factors like regularization) and are demonstrated in Figure 4.  
  
Figure 4: Graphical showcase of neural network overfitting and underfitting.

For underfitting to occur, the neural network’s flexibility needs to be insufficient to fit the true trend of the data. This can be due to the architecture of the neural network being too small and so it doesn’t have enough nodes to reproduce the trends of the data. For example in Figure 4 a quadratic function is being fit by a linear polynomial which is obviously insufficient to properly fit the trend. At the other extreme is overfitting, where a neural network is given too much flexibility by having too large of an architecture, which is also characterized by the neural network losing it’s generalizability meaning the neural network is only trying to fit the points of the training grid and not predicting the overall trend of the data. Overfitting would go unnoticed if one was only concerned with the values of the loss function, as the loss would be very low because the grid points are the only points evaluated with the loss function but as shown by the green points in Figure 4 which follow the same trend as the observed data, the neural network does a terrible job at fitting to these points. Which is why when training having validation data is very useful. Validation data are points that are in the grid of training points which the neural network is not allowed to be trained on, which by finding loss for the validation points (the validation loss), it can be seen whether the neural network is overfitting if the validation loss is larger than the standard loss.

There are a few ways to minimize the effects of overfitting, first off is reducing the architecture to simply reduce the flexibility of the neural network, the second is increasing the number of training points which isn’t always feasible but does prevent the neural network fitting such complex functions when the points it needs to fit to are closer together. The third method is regularization which is a term added onto the loss function that penalizes large weights, thus reducing the size of the weights and consequently causing the neural network to not be able to generate complex models, preventing overfitting. It is possible to overregularize which would then move the neural network into the underfitting regime (a description of the other methods of minimizing overfitting can be found in \cite{prevent\_overfitting}).

**Training Stage 1: Main Sequence, Subgiant and RGB**

The initial process of training began by taking a subset of the training grid, such that we could figure out the approximate setup, (architecture and regularization wise) without having to waste time by training on the full grid that contains over 2.6 million points. The training stage 1 grid specs are given in Table 3 and whose HR diagram is shown in the left panel of Figure 2.

**Table 3**: Training Stage 1 setup. “-” shows that those parameter’s range is dependent on the input parameters.

|  |  |  |
| --- | --- | --- |
| Parameter | Form of data used by neural network | Training grid |
| Age / Gyr | Log\_10{Age} | 1 - 12 |
| M\_{init} / M\_{\odot} | Log\_10{M} | 0.8 - 1.68 |
| [Fe/H] | Log\_10{[Fe/H]} | -0.2, 0, 0.2 |
| Y\_{init} | Log\_10{Y\_{init}} | 0.26, 0.28, 0.30 |
| α\_{MLT} | Log\_10{α\_{MLT}} | 1.7 - 2.3 |
| Luminosity / L\_{\odot} | Log\_10{R} | - |
| Teff / K | Log\_10{Teff/5000} | - |
| \Delta \nu / \mu Hz | Log\_10{\Delta \nu} | - |

To elaborate a few aspects of Table 3, neural networks like values they’re handling to have a small dynamical range close to 0. To combat this the neural network is given the log of all the parameters, which immediately solves that for everything except luminosity and the effective temperature, which we solve by dividing Teff by 5000 which reduces the log temperatures below 1 and to solve the luminosity dynamical range we instead used radius as one of the neural network outputs, calculated using the luminosity and effective temperature using the Stefan-Boltzmann law. This means that we were attempting to train a neural network to convert inputted age, mass, [Fe/H], initial helium and mixing length to Teff, radius and $\Delta \nu$, using roughly solar metallicity and helium on MS and RGB, stellar tracks.

With regards to a few choices made when training which in this study was done using keras \citep{keras}, the neural networks were trained almost entirely using the Nesterov Adam (Nadam) optimizer, which improves on the standard stochastic gradient decent (SGD) optimizer where the weight updates comes from moving in a direction of decreasing loss (see eq.5), which Nadam improves upon by having Nesterov momentum which acts similarly to normal momentum whereby it aids the weight updates moving along the same decrease in gradient direction which prevents the neural network bypass becoming stuck in a local minima. Except Nesterov momentum allows the gradient update to the weights helps correct the movement through parameter space due to the momentum movement because the movement due to momentum may not be in the same direction to the gradient update which always points in the correct direction. We found SGD to be useful in allowing the loss to slope down significantly after training for a period of time with Nadam but only when Nadam had been trained for an insufficient amount of time, although in the cases where SGD was used after Nadam we found the noise on the loss would disappear almost completely. For further details on Nadam see \cite{Nadam} and \cite{Nadam2}. For our choice of activation we selected elu, after initially testing relu we found elu performed much better, but found little difference for the dynamical range of our neural network inputs and outputs when we compared elu to swish. We calculated the training loss using the mean absolute error (MAE) instead of the mean squared error (MSE) because the training data has no noise on it as it’s calculated from models, which means the weight penalization should be proportional to the loss as per MAE, rather than MSE which doesn’t penalise when the loss is close to 0 and would be useful if using noisy training data, as it effectively makes allowances for the noise. We also chose L2 regularization when regularizing for similar reasons because we don’t need to account for outliers as the training data doesn’t contain any as it’s model generated, L2 regularization also performs better than say L2 regularization when all of the outputs are functions of the input. An additional strength of L2 is that there is a singular solution, which allows some reliability in the regularization before when training multiple neural networks but making very small tuning changes to certain training parameters. We also achieved some reliability in the neural network behaviour by randomizing the order of the data rows always with the same seed, which allowed training on batches to be consistent and allowed batches to be more representative of the features of the data and also importantly meant that the validation fraction (30%) was also more representative of the data.

We developed a training schedule following the method suggested by \cite{lrbatch}, where they compare a number of different training schedules and found that when increasing the batch size and increasing the learning rate during the early stages of training, you can achieve a very high learning rate with a small number of updates (to the weights), which makes the training process faster than a number of other methods. Batch size is very useful in neural network training because appropriate choices during the early stages of training allows for a larger rate of improvement because by giving batches roughly representative of the shape of the data, the neural network is able to reproduce trends faster by each weight update being more significant.

It’s recommended by \cite{lrbatch} that when increasing the learning rate and batch size to increase them proportionally and from testing we determined that the viable spread in learning at the start of training was 0.0001 to 0.001 and a fairly representative batch size was of 1000 points. The rough training outline is shown in Table 4 and the neural network hyperparameters that were kept the same during this stage of training are shown in Table 5.

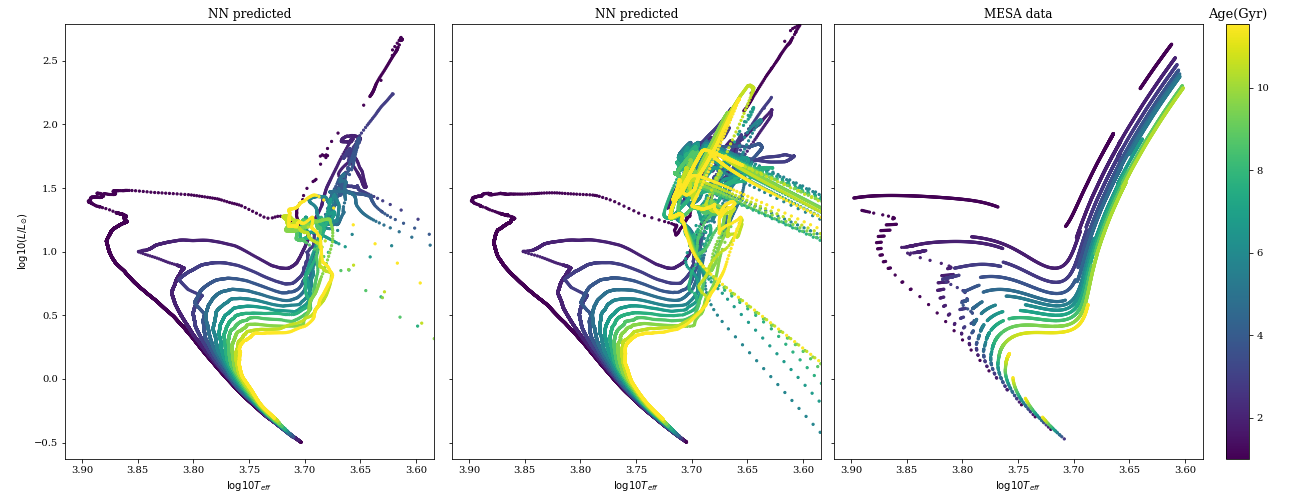
**Table 4**: Training schedule outline

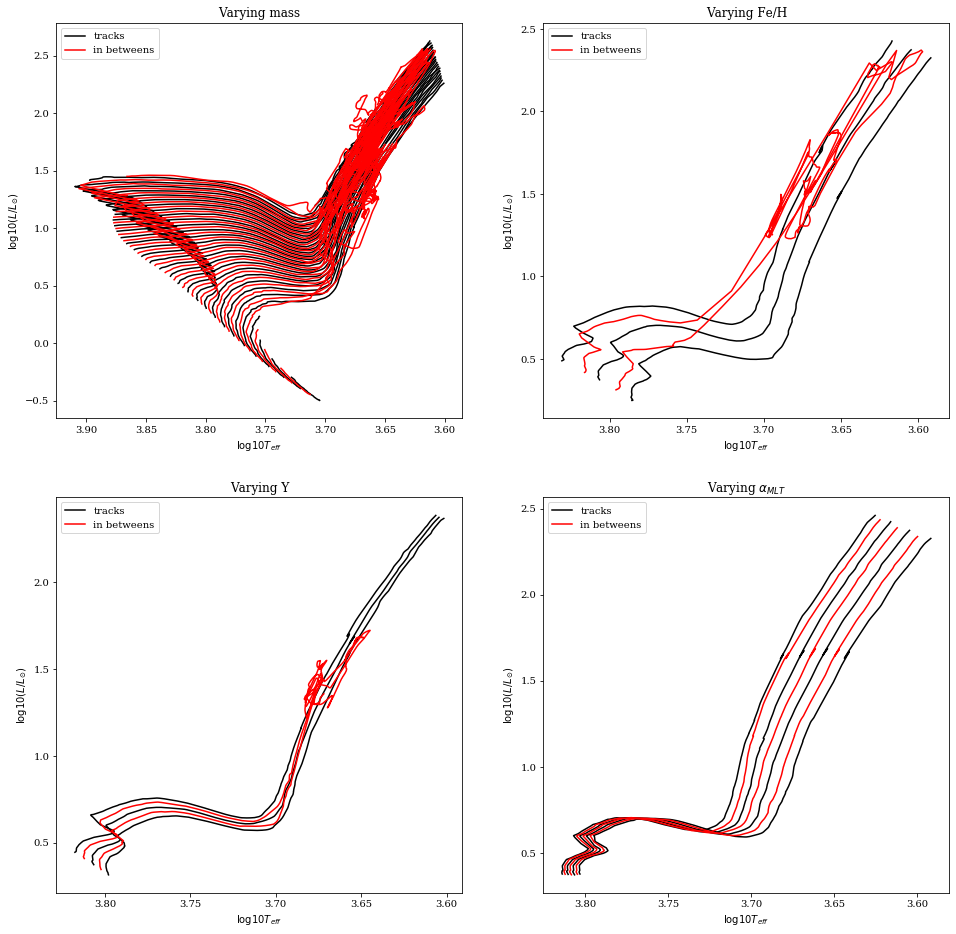
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Step 1 | Step 2 | Step 3 | Step 4 | Remaining steps |
| Learning rate | 0.0001 | 0.0002 | 0.0005 | 0.001 | Use previous step’s decayed learning rate |
| Batch size | 1000 | 2000 | 5000 | 1000 | Batch size increased in a 1x10x, 2x10x, 5 x10x, 1x10x+1 sequence until the max batch size reached |
| Epochs | 1000 | 1000 | 1000 | 3000 | 1000 until max batch size achieved and then trained until the loss is considered to have plateaued by eye. |

**Table 5**: Neural network hyperparameters kept constant during training stage 1. The hyperparameters of the Nadam optimizer are kept as the keras default.

|  |  |
| --- | --- |
| Hyperparameter | value |
| Weight initialization | Set seed as 53 |
| Optimizer | Nadam |
| Activation function | Elu |
| regression loss function | MAE |
| Batch normalization | Only on the input layer |
| Regularization | None |
| Dropout | None |

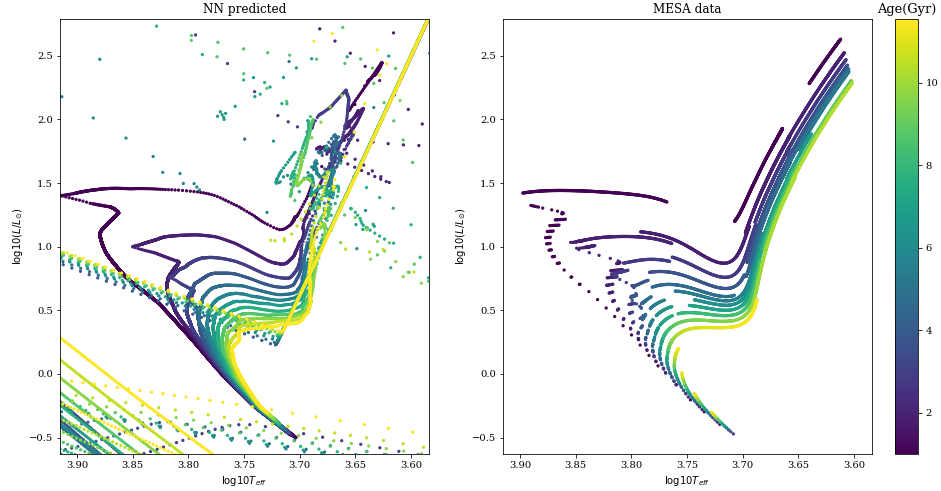
We began by trying to tune the size of the architecture, which we found with batch normalization on the input layer required a larger architecture then without, despite which the neural networks would still train faster. The first significant result was with an architecture of 8 layers of 256 nodes followed by 2 layers of 512 nodes, which for simplification hereinafter architecture will be written in the form layers x nodes i.e. for the above architecture 8x256,2x512. It was significant as it was the first neural network where we successfully got the MAE loss below 1x10-3 although there were a couple of noticeable unwanted traits about this neural network shown in figures 5 and 6.

  
**Figure 5**: Comparison of isochrones from a 8x256,2x512 architecture neural network. The left panel shows the isochrone after using the training schedule in **Table 4** to a batch size of 200,000 (approximately 1/6 of the max batch size). The middle panel shows the neural network isochrone after the full training schedule to max batch. The right panel shows the isochrone generated from the points in the training grid.

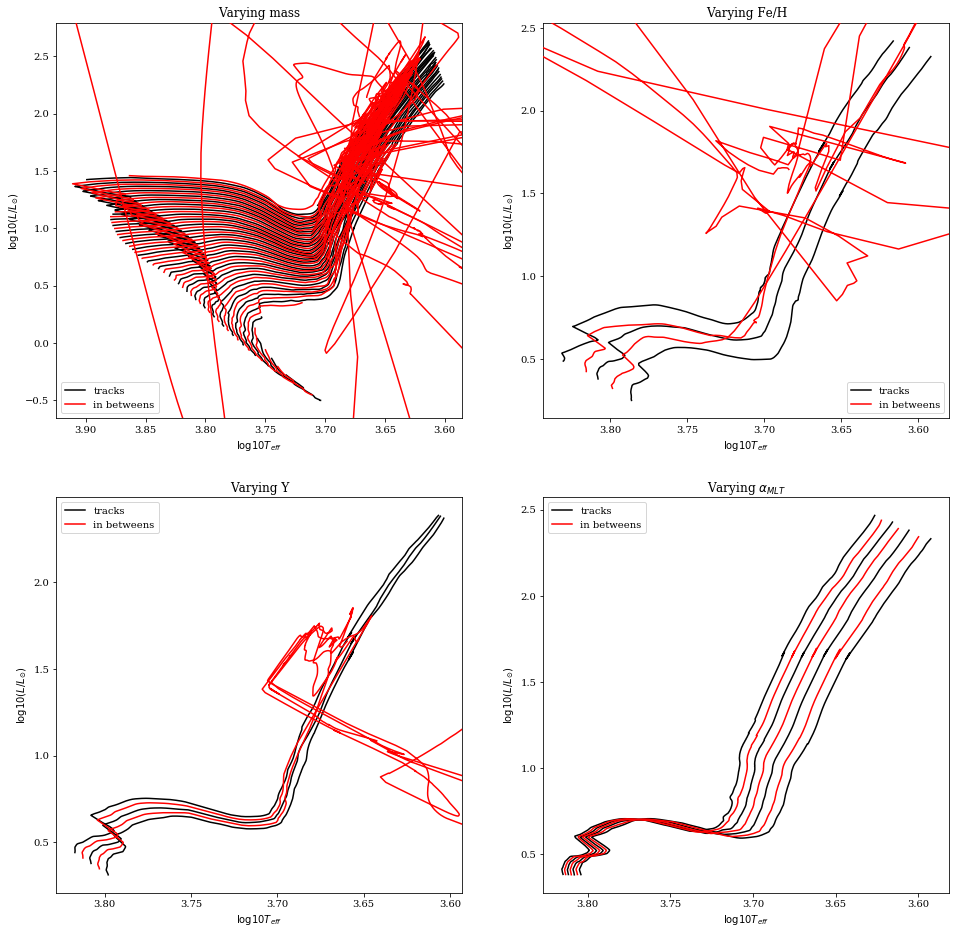
  
**Figure 6**: tracks from a 8x256,2x512 architecture neural network for different input parameters.

Figures 5 and 6 show the neural network is able to do a decent job at reproducing the main sequence except for the in-between [Fe/H] tracks in the top right panel of figure 6. By comparing the left and middle panels form figure 5 we see a behaviour that we prevalent throughout the course of every training we undertook in stage 1, where the more the neural network improved its approximation of the main sequence it would overfit more in the RGB. Figure 6 also gives a better visualization of where the overfitting is more significant, which is for the tracks of [Fe/H] and mass variation, especially in the RGB, though there is still overfitting in the main sequence particularly in both the in grid and off grid tracks for the higher mass tracks.

We combatted the overfitting by halving the number of nodes per layer to an architecture of 8x128,2x256 the results of which are shown in figure 7 and 8.

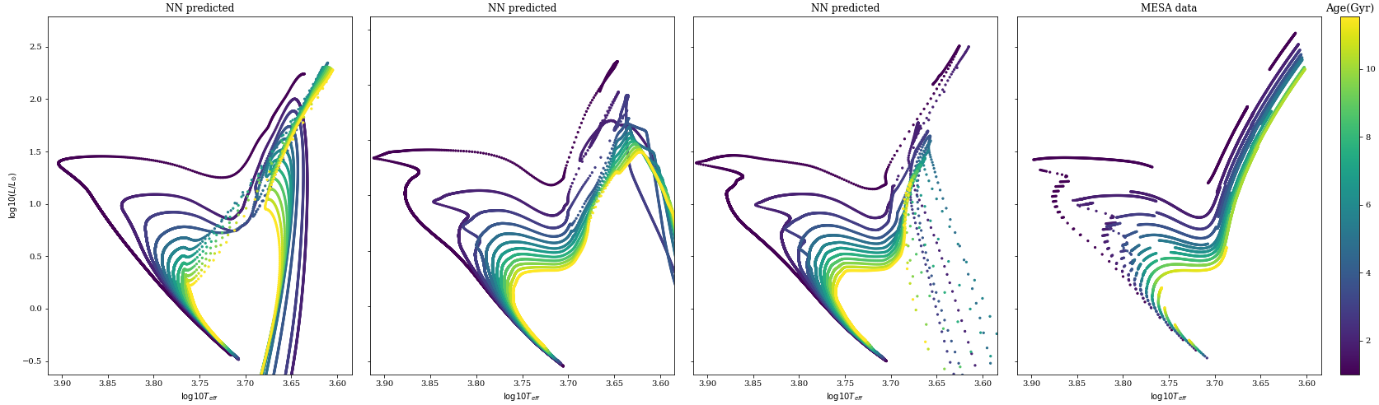


**Figure 7**: Isochrones after full training schedule using 8x128,2x256 architecture.

  
**Figure 8**: tracks from a 8x128,2x256 architecture neural network for different input parameters.

Reducing the architecture allowed the neural network to do a better job at reproducing the main sequence as shown by the hook shape of the lower ages tracks in figure 7 and the higher age tracks not overlapping close to the end of the sub giant branch. There is always a smaller amount of overfitting on the tracks themselves as they tend to be smoother. Figure 8 does look worse than figure 6 in terms of overfitting but in fact each track has less flexibility put into it which is why in figure 8 the tracks tend to move in straight lines unlike in figure 6 where there is much more flexibility which is why the tracks are much more compressed spiralling around in the RGB and thus has much worse overfitting. The only downside is with less overfitting came a slightly reduced loss of roughly 1x10-3.

Next, we employed L2 regularization to reduce overfitting, the results of which are shown in figure 9 and 10.

  
Figure 9: Comparison of isochrones with different regularizations for a neural network architecture of 8x128,2x256. The left most panel is with a regularization of 1x10-4, the middle left panel is with 1x10-6, the right middle panel is 1x10-7 and the right most panel is plotted from the training grid points.

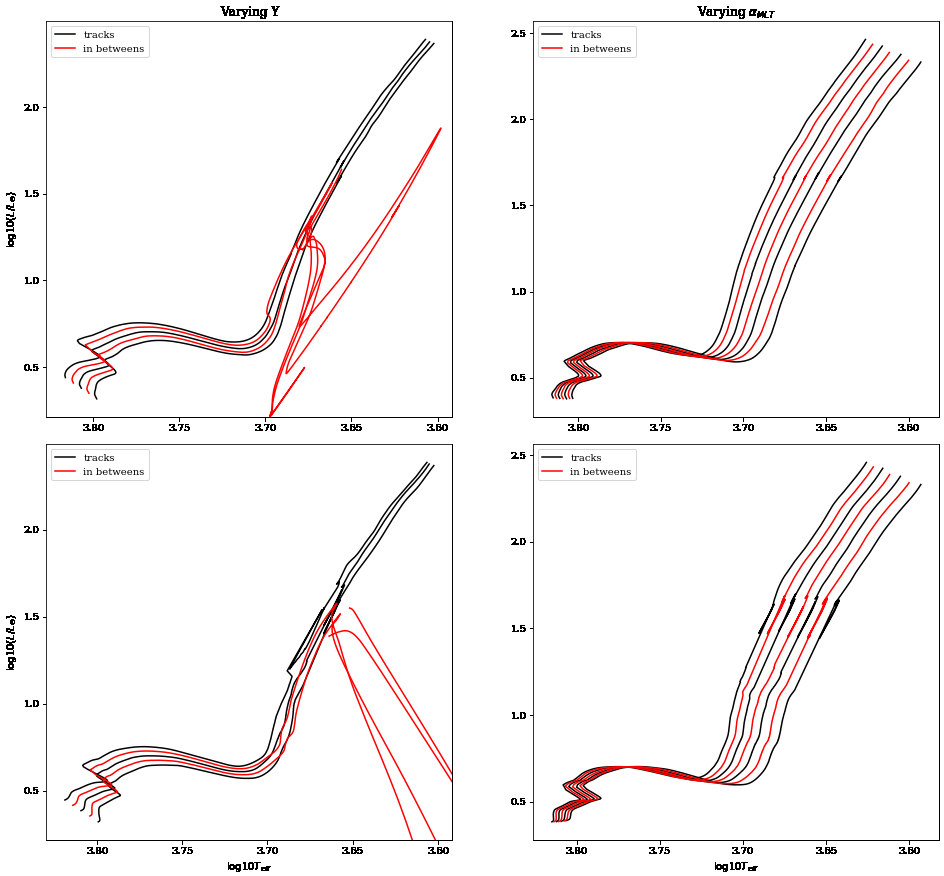
  
**Figure 10**: Comparison of the mass and MLT tracks between 1x10-6 (top) and 1x10-7 (bottoms) regularizations. (the mass and [Fe/H] graphs have been omitted as there wasn’t a clear difference between the 2 regularizations)

Figure 9 demonstrates that 1x10-4 is overregularized as the entire subgiant branch has been smoothed out for the lower age isochrones. Distinguishing between 1x10-6 and 1x10-7 regularization is harder to see in figure 9 but when looking at figure 10 there is an apparent difference between the 2 when looking at the main sequence, as 1x10-6 is significantly less “wiggly” demonstrating that it is overfitting less 1x10-7. Both regularizations still give poor behaviour in the RGB but 1x10-6 is superior in both on and off grid tracks and achieved the highest loss of all neural networks shown to this point with 7x10-4 loss (in MAE).

**Training Stage 2: Main sequence and Subgiant**  
Using a variety of methods we were unable to satisfactorily reduce the amount of overfitting in the RGB whilst maintaining a loss of 1x10-4, so we instead focussed just on main sequence and subgiant, which required truncating the MESA grid shown in Table 2, by keeping points which satisfied: $10log\_{10}(Teff)-log\_{10}(L)-35.5>0$, an inequality selected by eye which keeps the lower part of RGB so neural networks still reproduce the end of the subgiant branch correctly. The resulting grid’s HR diagram is shown in the right panel of figure 2. The best performing neural network trained with this grid, which will be referred to the MS+SG NN is described the tables in Appendix B.

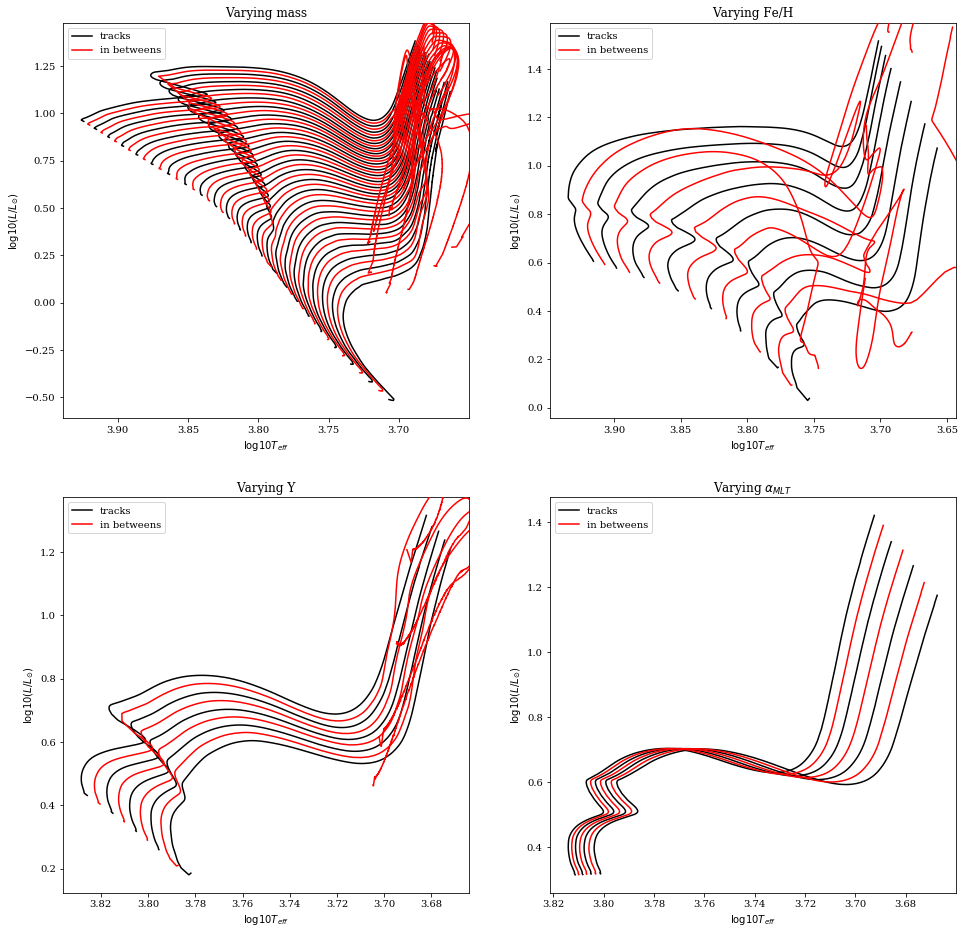
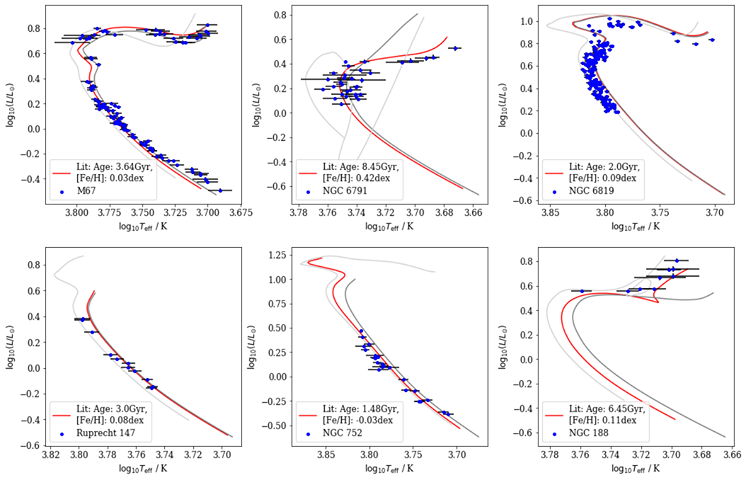
  
**Figure 10**: tracks of MS+SG NN for different input parameters.

Figure 10 shows that MS+SG NN performs well in most of the in-between regions of the input parameters except for in the low RGB as expected and the higher [Fe/H] tracks which we believe is because the training grid doesn’t have fine enough [Fe/H] track spacings for a parameter with which a small change to the input makes a large change to the output. For αMLT an incremental spacing of 0.2 is acceptable because the changes in αMLT are less impactful than for [Fe/H] in the output. Fortunately, of the open clusters we studied MS+SG NN only had trouble forming isochrones for NGC 6791 as shown in figure 11, due to NGC 6791’s high metallicity.

  
**Figure 11**: Literature isochrone with closest in-between [Fe/H] step on either side of the literature [Fe/H] (grey line is higher in-between [Fe/H] and light grey is lower in-betweeen [Fe/H]), for each of the open clusters we’re studying (see Table 1 for literature references).

We trained a second neural network in an attempt to improve isochrone fitting in the higher metallicity regions such that we could fit NGC 6791, the resulting neural network’s setup and training is described in Appendix C (this neural network will be referred to as NGC 6791 NN) and the in-between tracks plots are shown in figure 12.

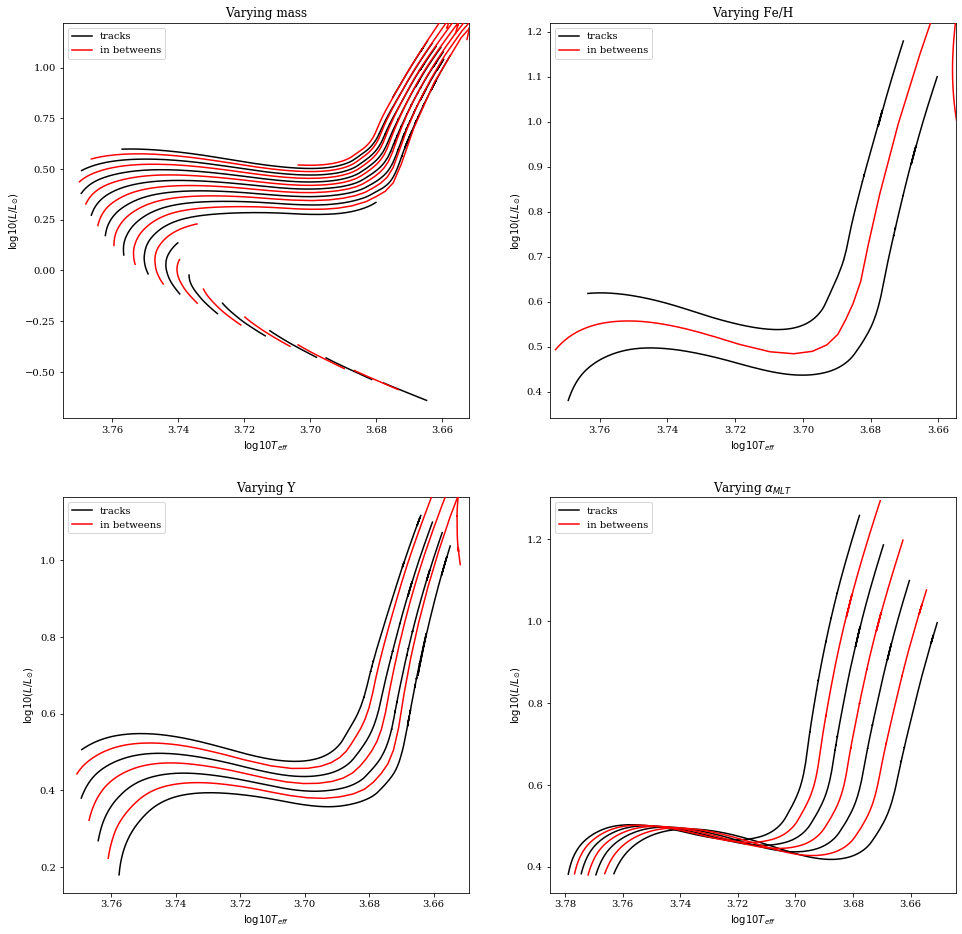
  
**Figure 12**: tracks of NGC 6791 NN for different input parameters.

Figure 12 shows that for a reduced metallicity range we were able to almost completely get rid of the overfitting issues that plagued the other neural networks, especially when looking at the metallicity tracks, and the lower RGB behaves correctly as well.

Having trained 2 neural networks to a sufficient level to reproduce the behaviour of age, mass, metallicity, helium and MLT in the approximate regions of the open clusters we wanted to study we felt these would work well when used to convert between fundamentals and observables in a hierarchical Bayesian model.

**Hierarchical Bayesian Modelling:**

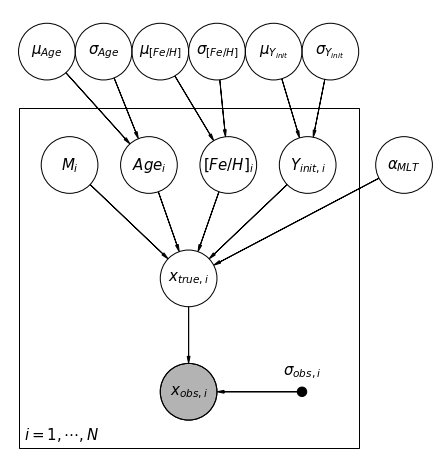
**HBMs an Introduction:**

Hierarchical Bayesian models are multileveled statistical models which are used to approximate the posterior distribution of some parameters, given some data. The strengths of HBMs come from that they are able to simultaneously constrain the parameters of multiple objects, by grouping together the data for all the objects which improves the constraint \citep{si\_2018} and by choosing appropriate samplers they can be fast. Additionally, the multileveled nature of HBMs allow you to use prior knowledge of the overarching features of a population to help constrain the individual members of that population and further constrain the properties of the population as a whole. This is done using parameter priors which are informed by population wide hyperpriors which for a parameter describes the belief in its mean (\mu), possible range which the mean might be within (\mu\_{range}), spread of that parameter through the population (\sigma), and the range of the spread (\sigma\_{range}). The hyperpriors are also given distributions to describe the likelihood in our believes in where the true population wide parameters lie. The mean of parameter has a distribution described by \mu and \mu\_{range} which is the standard deviation for normal distributions, similarly for the spread of a parameter being described by \sigma and \sigma\_{range}. Through sampling these 4 variables (\mu, \mu\_{range}, \sigma, \sigma\_{range}) for each parameter are constrained and are shifted and influence the prior distributions described by \mu and \sigma of each parameter. The priors generate pseudo populations which using a model enable population guesses from the prior to be compared to observational data and iteratively constrain and construct the posterior of parameter for each object and for the population wide hyperpriors.

HBMs have been used a number of times in astrophysics like \cite{jørgensen\_2005} and \cite{hippel\_2014} who found that the results of HBMs compared to standard isochrone fitting were at least equally as accurate and sometimes much better \citep{jørgensen\_2005} and that they produced much more information \citep{hippel\_2014}. HBMs have further been used to fit initial-final mass relations \citep{si\_2018}, age determinations of NGC 188 \citep{hills\_2015} and M67 \citep{jørgensen\_2005}. Further discussion of the uses of Bayesian modelling of stellar populations can be found in \cite{HBMbook} and \cite{HBMpaper}.

**HBM implementation:**

We generated a 4 level HBM, whose structure is described in figure 12, with the top level containing the 7 fundamental hyperpriors describing the overarching properties of the open cluster: the mean and spread of the age, [Fe/H] and Y\_{init}, with the 7th hyperparameter being the mean \alpha\_{MLT} which is taken to be constant for all stars in the cluster and is why \alpha\_{MLT} has no prior level as there is no star-to-star variation. The 2nd level is the prior level, describing the star by star fundamental properties, mass is within the prior level because the mass is star dependent not open cluster dependent. The prior level also contains priors on age, [Fe/H] and Y\_{init}. The 3rd level describes the “true” observables being the noiseless output of the neural network coming from the predictions from the fundamental priors. The 4th level is where the posterior is calculated by comparing “true” observables with added Gaussian noise, to the observed data using a Gaussian likelihood to give a probability of the HBM’s guesses constructed from the previous 3 levels.

   
**Figure 13**: Probabilistic graphical model of HBM setup.

To influence how the HBM samples around the hyperpriors and priors we set distributions on them. The mean hyperpriors and the mass prior are described by (1.1,1.1) Beta-functions, because the HBM needs a smooth function (which is why we don’t use a top-hat function) which still has boundaries of 0 probabilities because otherwise the HBM will sample parameter space the neural networks were not train in which causes convergence issues. The spread hyperpriors are chosen to be log-normal distributions because we know the spread of the age non-zero but is likely to be nearer 0 than far away from 0. Then the non-mass priors are normal distributions because the prior is informed by the mean and spread hyperpriors and the distribution of the stellar parameters should be cantered on the mean of the cluster. As the spread of the parameters should be independent of the mean properties of the open clusters, they initial hyperpriors are chosen to the same for each cluster as shown in Table 6.

**Table 6**: Parameters chosen to construct the lognormal of the spread hyperpriors used for each of the open clusters.

|  |  |  |
| --- | --- | --- |
| Hyperparameter | \mu | \sigma |
| \sigma\_{age} | log\_{10}(0.15) | 0.4 |
| \sigma\_{[Fe/H]} | log\_{10}(0.05) | 0.5 |
| \sigma\_{Y\_{init}} | log\_{10}(0.01) | 0.5 |

**Table 7**: Ranges adopted for the mean hyperpriors and for the mass prior for each of the studied open clusters. The \alpha\_{MLT} is not given as the full range used in the neural network training was always used (1.7 - 2.3).

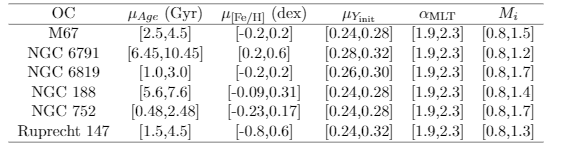


Table 7 shows the final mean hyperpriors and mass prior used for the HBMs of each of the studies clusters. These values were established by using literature values shown in Table 1 to run an initial set of HBMs to see if the ranges given were appropriate. If the resulting posteriors of any of the parameters for a cluster peaked out the sampling boundary, the boundary was translated to have the peak of that posterior at the centre of the sampling region, unless doing so would cause sampling outside of trained region of the neural network in which case the ranges were translated to centralize the posterior in the sampling region as much as possible without the sampling region extending beyond the trained neural network region.

3859 tracks in total  
2659783 datapoints

Acknowledgements:  
Guy Davies for their guidance, Tanda Li for permitting us to use their MESA grid.

This work has made use of data from the European Space Agency (ESA) mission {\it Gaia} (\url{https://www.cosmos.esa.int/gaia}), processed by the {\it Gaia} Data Processing and Analysis Consortium (DPAC, \url{https://www.cosmos.esa.int/web/gaia/dpac/consortium}). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the {\it Gaia} Multilateral Agreement.

|  |
| --- |
| Ted von Hippel: <https://arxiv.org/pdf/astro-ph/0509152.pdf> |
| M. Salaris et al. 2004: “The age of the oldest Open Clusters”  **Cluster age estimates in table 1**  <https://www-aanda-org.ezproxye.bham.ac.uk/articles/aa/full/2004/04/aah4736/aah4736.html>  The most reliable stellar ages are obtained for the star clusters belonging to the various populations, i.e., the globular clusters (GCs) in the halo, thick disk and bulge, and the open clusters (OCs) in the thin disk. The advantage of dating star clusters over individual stars – whose age determination relies entirely on the knowledge of individual metallicities, effective temperatures and gravities (or absolute magnitudes), which have to be fitted by the appropriate theoretical model – stems from the fact that star clusters are made of coeval objects, largely with the same initial chemical composition and located at the same distance, so that it is possible to use morphological parameters deduced from theoretical isochrones in order to derive their age. |
| Clio Bertelli Motta et al. 2016: “Observing the products of stellar evolution in the old open cluster M67 with APOGEE”  **Isochrone fitting M67**  <https://academic-oup-com.ezproxye.bham.ac.uk/mnras/article/466/2/2161/2687805> |
| D. Stello, et al. 2016: “THE K2 M67 STUDY: REVISITING OLD FRIENDS WITH K2 REVEALS OSCILLATING RED GIANTS IN THE OPEN CLUSTER M67”  **Isochrone fitting/grid based modelling**  <https://openresearch-repository.anu.edu.au/bitstream/1885/152166/2/01_Stello_THE_K2_M67_STUDY%253A_REVISITING_2016.pdf> |
| 1. F. Oliveira, et al. 2013: “Fitting isochrones to open cluster photometric data III. Estimating metallicities from UBV photometry”   **Isochrone fitting metallicity using the cross-entropy global optimization algorithm contains explanation**  “The metallicity is a critical parameter that affects the correct determination of fundamental characteristics of a stellar cluster and has important implications in Galactic and stellar evolution research. Fewer than 10% of the 2174 currently atalogued open clusters have their metallicity determined in the literature. In this work we present a method for estimating the metallicity of open clusters via non-subjective isochrone fitting using the cross-entropy global optimization algorithm applied to UBV photometric data.”  “Important questions that depend on metallicity, which is usually measured by the [Fe/H] ratio, are the determination of chemical abundance gradients (see Lépine et al. 2011, and references therein), determination of the rotational speed of the spiral pattern, and the co-rotation radius (Dias & Lépine 2005), and in the stellar context the empirical determination of the initial mass function, among many other fields of study.”  “Very schematically, the CE procedure provides a simple adaptive way of estimating the best-fit parameters. It involves an iterative procedure that follows the steps outlined below:  - random generation of the initial sample of fit parameters, respecting predefined criteria;  - selection of the best candidates based on calculated weighted likelihood values;  - generation of a random fit parameter sample derived from a new distribution based on the previous step;  - repeat until convergence or stopping criteria reached.” |
| Monteiro, H.; Dias, W. S.; Caetano, T. C. 2010: “Fitting isochrones to open cluster photometric data. A new global optimization tool”  **Isochrone fitting using cross-entropy global optimization algorithm ORIGINAL**  <https://ui.adsabs.harvard.edu/abs/2010A%26A...516A...2M/abstract> |
| H. Pöhnl and E. Paunzen 2010: “A statistical method to determine open cluster metallicities”  Method which differs from isochrone fitting as it doesn’t really on metallicity measurements.  “The study of open cluster metallicities helps to understand the local stellar formation and evolution throughout the Milky Way. Its metallicity gradient is an important tracer for the Galactic formation in a global sense. Because open clusters can be treated in a statistical way, the error of the cluster mean is minimized.”  <https://www-aanda-org.ezproxye.bham.ac.uk/articles/aa/pdf/2010/06/aa10855-08.pdf> |
| L. A. Balona, et al. (2013)  **Isochrone fitting**  “we determine the distance and age of NGC 6819 using several different methods. From isochrone fitting we find the age of the cluster to be about 2.5 Gyr” |
| L. N. Brewer, et al. (2016): “Determining the Age of the Kepler Open Cluster NGC 6819 With a New Triple System and Other Eclipsing Binary Stars”  **Ageing using Detached Eclipsing Binaries**  “using all measured eclipsing binary star masses and radii, we constrain the  age to 2.38 ± 0.05 ± 0.22 Gyr. The quoted uncertainties are estimates of measurement and systematic uncertainties (due to model physics differences and metal content), respectively”  “Measurements of the masses and radii of the component stars in detached eclipsing binaries (DEB) can be used to precisely determine the age of the stars if at least one of the eclipsing stars has begun to evolve away from the main sequence. The use of mass and radius (M – R)  measurements of eclipsing stars avoids or minimizes systematic uncertainties introduced by factors such as distance, interstellar reddening, and color-temperature conversions that can affect age measurements (Andersen 1991; Torres et al. 2010). When eclipsing binaries occur in star clusters, their utility increases dramatically because they place constraints on the age of all the stars in the cluster. Multiple DEBs in a cluster can provide M – R measurements for stars having a range in mass, and can constrain the age even more tightly”  “Kalirai et al. (2001) describes how the age of NGC 6819 can be found using the faint end of the white dwarf cooling sequence, and Bedin et al. (2015) present an age of 2.25 ± 0.20 Gyr using that technique.”  <https://arxiv.org/pdf/1601.04069.pdf> |
| Bavarsad E. A. et al. (2016) : “THE DETACHED ECLIPSING BINARY KV 29 AND THE AGE OF THE OPEN CLUSTER M11”  **More detailed description of how age determinations work with DEBs**  <https://iopscience.iop.org/article/10.3847/0004-637X/831/1/48> |
| Kalirai JS, et al. (2001)  “The CFHT Open Star Cluster Survey. II. Deep CCD photometry of the old open star cluster NGC 6819”  **Age determinations using white dwarf cooling** |
| Bedin L. R. et al. (2015)  “Hubble Space Telescope observations of the Kepler-field cluster NGC 6819 – I. The bottom of the white dwarf cooling sequence”  **Age determination of NGC 6819 via white dwarf cooling**  “2.25 ± 0.20 Gyr”  <https://academic-oup-com.ezproxye.bham.ac.uk/mnras/article/448/2/1779/1056101> |
| Lund M. N. et al. (2016): “Asteroseismology of the Hyades with K2: first detection of main-sequence solar-like oscillations in an open cluster”  **first ever detections of solar-like oscillations in main-sequence stars in an open cluster**  **used kepler** |
| Hekker, S. et al. (2011): “Asteroseismic inferences on red giants in open clusters NGC 6791, NGC 6819, and NGC 6811 using Kepler”  **Asteroseismology of OC red giants using Kepler** |
| EMBEDDED CLUSTERS IN MOLECULAR CLOUDS  <https://pdfs>.semanticscholar.org/dd06/a8a9143f9e168dc2dc30c8cc3a15cc10fb40.pdf  galactic clusters form in giant molecular clouds (GMCs) and during their formation and earliest stages of evolution are completely embedded in molecular gas and dust, and thus obscured from view. Given the constraints imposed by traditional techniques of optical astronomy, direct observation and study of young embedded clusters had been extremely difficult, if not impossible. However, during the past two decades the development of infrared astronomy and, more recently, infrared array detectors, has dramatically improved this situation. |
| Jeffery, E. J. et al. (2016): “A BAYESIAN ANALYSIS OF THE AGES OF FOUR OPEN CLUSTERS”  **Isochrone fitting using Bayesian MCMC and Dartmouth Stellar Evolution Database (Dotter**  **et al. 2008) to create tracks.**  “Star clusters have long been important tools for studying  stellar evolution, specifically because they play the pivotal role  in determining the ages of stars. The most commonly used  method for measuring the age of an open star cluster involves  fitting an isochrone to the cluster’s observed color–magnitude  diagram (CMD), specifically to the cluster’s main sequence  turn-off (MSTO). Generating and fitting isochrones to a cluster  CMD to determine its age also requires knowledge of the  cluster’s metallicity, distance, and reddening. Oftentimes,  finding a best fit of these three parameters (plus age) is a  subjective process, as some of these parameters are correlated  with each other. This difficulty is reflected in isochrones that  appear to fit the CMD equally well with various combinations  of cluster parameters (see, for example, Figure 2 of VandenBerg & Stetson 2004). Moreover, the fit of the MSTO can be  challenging and isochrones may give inconsistent results in  different filters, even when using the same cluster parameters  (see, for example, Figure 10 of Sarajedini et al. 1999).”  <https://iopscience.iop.org/article/10.3847/0004-637X/828/2/79/pdf> |
| Hills, S. et al. (2015)  **Bayesian isochrone fit of NGC 188**  <https://iopscience.iop.org/article/10.1088/0004-6256/149/3/94/pdf> |
| Brandt, T. D. et al. (2015): “Rotating Stellar Models Can Account for the Extended Main Sequence Turnoffs in Intermediate Age Clusters”  **Isochrone fitting “by-eye”**  <https://arxiv.org/abs/1504.04375> |
| Perren, G. I. et al. (2015)  Automated isochrone fitting  <https://arxiv.org/pdf/1412.2366.pdf> |
| Appourchaux et al. (2015) |
| Ball and Gizon, (2017)  MESA  Ball, W. H. et al. (2018)  MESA |
| jørgensen\_2005  Bayesian age determinations |
| cantat-gaudin\_2018  **membership paper with 1k clusters**  <https://www.aanda.org/articles/aa/full_html/2018/10/aa33476-18/aa33476-18.html> |
| gao\_2018  M67 membership found using random forest  <https://ui.adsabs.harvard.edu/abs/2018ApJ...869....9G/abstract> |
| sanders\_2018  **where we get the G-band extinction coefficient**  <https://arxiv.org/pdf/1806.02324.pdf> |
| green\_2019  **3D dustmap**  <https://iopscience.iop.org/article/10.3847/1538-4357/ab5362> |
| andrae\_2018  **Teff 🡪 bolometric correction model** |
| MBOLSOL |
| li\_2018  Tanda’s training grid setup using MESA  <https://arxiv.org/pdf/1810.13015.pdf> |
| cheng\_2020  **comparison of different supervised machine learning methods for morphology classification**  <https://arxiv.org/abs/1908.03610> |
|  |
|  |
|  |
| M67 age estimation papers: Sarajedini et al. 2009; Bellini et al. 2010a; Kharchenko et al. 2013  Some metallicity estimate papers: Chen et al. (2003); Magrini et al. (2009).  NGC 6819 age determinations: Kalirai et al. 2001; Basu et al. 2011  NGC 6819 metallicity determination: [Fe/H]= +0.09 ± 0.03; Bragaglia et al. 2001 |

|  |
| --- |
| B. Paxton et al. (2010)  @article{MESA,  title={MODULES FOR EXPERIMENTS IN STELLAR ASTROPHYSICS (MESA)},  volume={192},  DOI={10.1088/0067-0049/192/1/3},  number={1},  journal={The Astrophysical Journal Supplement Series},  author={Paxton, Bill and Bildsten, Lars and Dotter, Aaron and Herwig, Falk and Lesaffre, Pierre and Timmes, Frank},  year={2010},  pages={3}} |
| @article{MESA2,  title={MODULES FOR EXPERIMENTS IN STELLAR ASTROPHYSICS (MESA): PLANETS, OSCILLATIONS, ROTATION, AND MASSIVE STARS},  volume={208},  DOI={10.1088/0067-0049/208/1/4},  number={1},  journal={The Astrophysical Journal Supplement Series},  author={Paxton, Bill and Cantiello, Matteo and Arras, Phil and Bildsten, Lars and Brown, Edward F. and Dotter, Aaron and Mankovich, Christopher and Montgomery, M. H. and Stello, Dennis and Timmes, F. X. et al.},  year={2013},  pages={4}} |
| @article{MESA3,  title={MODULES FOR EXPERIMENTS IN STELLAR ASTROPHYSICS (MESA): BINARIES, PULSATIONS, AND EXPLOSIONS},  volume={220},  DOI={10.1088/0067-0049/220/1/15},  number={1},  journal={The Astrophysical Journal Supplement Series},  author={Paxton, Bill and Marchant, Pablo and Schwab, Josiah and Bauer, Evan B. and Bildsten, Lars and Cantiello, Matteo and Dessart, Luc and Farmer, R. and Hu, H. and Langer, N. et al.},  year={2015},  pages={15}} |
| Dotter, A. et al. (2008)  @article{DSED,  title={The Dartmouth Stellar Evolution Database},  volume={178},  DOI={10.1086/589654},  number={1},  journal={The Astrophysical Journal Supplement Series},  author={Dotter, Aaron and Chaboyer, Brian and Jevremović, Darko and Kostov, Veselin and Baron, E. and Ferguson, Jason W.},  year={2008},  pages={89-101}} |
| D. Bossini et al. (2019)  @article{Bossini,  title={Age determination for 269 Gaia DR2 open clusters},  volume={623},  DOI={10.1051/0004-6361/201834693},  journal={Astronomy & Astrophysics},  author={Bossini, D. and Vallenari, A. and Bragaglia, A. and Cantat-Gaudin, T. and Sordo, R. and Balaguer-Núñez, L. and Jordi, C. and Moitinho, A. and Soubiran, C. and Casamiquela, L. et al.},  year={2019},  pages={A108}}  <https://arxiv.org/pdf/1901.04733.pdf> |
| M. Netopil et al. (2016)  @article{Netopil\_2016,  title={On the metallicity of open clusters},  volume={585},  DOI={10.1051/0004-6361/201526370},  journal={Astronomy & Astrophysics},  author={Netopil, M. and Paunzen, E. and Heiter, U. and Soubiran, C.},  year={2016},  pages={A150}}  <https://www.aanda.org/articles/aa/pdf/2016/01/aa26370-15.pdf> |
| Balona, L. A. et al. (2013)  @article{balona\_2013,  title={Kepler observations of the open cluster NGC 6819},  volume={430},  DOI={10.1093/mnras/stt148},  number={4},  journal={Monthly Notices of the Royal Astronomical Society},  author={Balona, L. A. and Medupe, T. and Abedigamba, O. P. and Ayane, G. and Keeley, L. and Matsididi, M. and Mekonnen, G. and Nhlapo, M. D. and Sithole, N.},  year={2013},  pages={3472-3482}} |
| L. N. Brewer, et al. (2016)  @article{brewer\_2016,  title={DETERMINING THE AGE OF THEKEPLEROPEN CLUSTER NGC 6819 WITH A NEW TRIPLE SYSTEM AND OTHER ECLIPSING BINARY STARS},  volume={151},  DOI={10.3847/0004-6256/151/3/66},  number={3},  journal={The Astronomical Journal},  author={Brewer, Lauren N. and Sandquist, Eric L. and Mathieu, Robert D. and Milliman, Katelyn and Geller, Aaron M. and Jeffries, Jr., Mark W. and Orosz, Jerome A. and Brogaard, Karsten and Platais, Imants and Bruntt, Hans et al.},  year={2016},  pages={66}} |
| Bavarsad E. A. et al. (2016)  @article{bavarsad\_2016,  title={THE DETACHED ECLIPSING BINARY KV 29 AND THE AGE OF THE OPEN CLUSTER M11},  volume={831},  DOI={10.3847/0004-637x/831/1/48},  number={1},  journal={The Astrophysical Journal},  author={Bavarsad, Ernest A. and Sandquist, Eric L. and Shetrone, Matthew D. and Orosz, Jerome A.},  year={2016},  pages={48}} |
| Kalirai JS, et al. (2001)  @article{kalirai\_2001,  title={The CFHT Open Star Cluster Survey. II. Deep CCD Photometry of the Old Open Star Cluster NGC 6819},  volume={122},  DOI={10.1086/321141},  number={1},  journal={The Astronomical Journal},  author={Kalirai, Jasonjot Singh and Richer, Harvey B. and Fahlman, Gregory G. and Cuillandre, Jean-Charles and Ventura, Paolo and D’Antona, Francesca and Bertin, Emmanuel and Marconi, Gianni and Durrell, Patrick R.},  year={2001},  pages={266-282}} |
| Bedin L. R. et al. (2015)  @article{bedin\_2015,  title={Hubble Space Telescope observations of the Kepler-field cluster NGC 6819 – I. The bottom of the white dwarf cooling sequence★},  volume={448},  DOI={10.1093/mnras/stv069},  number={2},  journal={Monthly Notices of the Royal Astronomical Society},  author={Bedin, L. R. and Salaris, M. and Anderson, J. and Cassisi, S. and Milone, A. P. and Piotto, G. and King, I. R. and Bergeron, P.},  year={2015},  pages={1779-1788}} |
| Lund M. N. et al. (2016)  @article{lund\_2016,  title={Asteroseismology of the Hyades with K2: first detection of main-sequence solar-like oscillations in an open cluster},  volume={463},  DOI={10.1093/mnras/stw2160},  number={3},  journal={Monthly Notices of the Royal Astronomical Society},  author={Lund, Mikkel N. and Basu, Sarbani and Silva Aguirre, Víctor and Chaplin, William J. and Serenelli, Aldo M. and García, Rafael A. and Latham, David W. and Casagrande, Luca and Bieryla, Allyson and Davies, Guy R. et al.},  year={2016},  pages={2600-2611}} |
| Hekker, S. et al. (2011)  @article{hekker\_2011,  title={Asteroseismic inferences on red giants in open clusters NGC 6791, NGC 6819, and NGC 6811 usingKepler},  volume={530},  DOI={10.1051/0004-6361/201016303},  journal={Astronomy & Astrophysics},  author={Hekker, S. and Basu, S. and Stello, D. and Kallinger, T. and Grundahl, F. and Mathur, S. and García, R. A. and Mosser, B. and Huber, D. and Bedding, T. R. et al.},  year={2011},  pages={A100}} |
| Von Hippel T. (2005)  @article{Hippel\_2005,  title={Galactic open clusters},  journal={Resolved Stellar Populations, ASP Conf.},  author={von Hippel, T.},  year={2005}} |
| Salaris, M. et al. (2004)  @article{salaris\_2004,  title={The age of the oldest Open Clusters},  volume={414},  DOI={10.1051/0004-6361:20031578},  number={1},  journal={Astronomy & Astrophysics},  author={Salaris, M. and Weiss, A. and Percival, S. M.},  year={2004},  pages={163-174}} |
| Jeffery, E. J. et al. (2016)  @article{effery\_2016,  title={A BAYESIAN ANALYSIS OF THE AGES OF FOUR OPEN CLUSTERS},  volume={828},  DOI={10.3847/0004-637x/828/2/79},  number={2},  journal={The Astrophysical Journal},  author={Jeffery, Elizabeth J. and Hippel, Ted von and Dyk, David A. van and Stenning, David C. and Robinson, Elliot and Stein, Nathan and Jefferys, William H.},  year={2016},  pages={79}} |
| Hills, S. et al. (2015)  @article{hills\_2015,  title={BAYESIAN INVESTIGATION OF ISOCHRONE CONSISTENCY USING THE OLD OPEN CLUSTER NGC 188},  volume={149},  DOI={10.1088/0004-6256/149/3/94},  number={3},  journal={The Astronomical Journal},  author={Hills, Shane and von Hippel, Ted and Courteau, Stéphane and Geller, Aaron M.},  year={2015},  pages={94}} |
| Brandt, T. D. et al. (2015)  @article{brandt \_2015,  title={ROTATING STELLAR MODELS CAN ACCOUNT FOR THE EXTENDED MAIN-SEQUENCE TURNOFFS IN INTERMEDIATE-AGE CLUSTERS},  volume={807},  DOI={10.1088/0004-637x/807/1/25},  number={1},  journal={The Astrophysical Journal},  author={Brandt, Timothy D. and Huang, Chelsea X.},  year={2015},  pages={25}} |
| Perren, G. I. et al. (2015)  @article{perren\_2015,  title={AsteCA: Automated Stellar Cluster Analysis},  volume={576},  DOI={10.1051/0004-6361/201424946},  journal={Astronomy & Astrophysics},  author={Perren, G. I. and Vázquez, R. A. and Piatti, A. E.},  year={2015},  pages={A6}} |
| Ball and Gizon, (2017)  @article{ball\_2017,  title={Surface-effect corrections for oscillation frequencies of evolved stars},  volume={600},  DOI={10.1051/0004-6361/201630260},  journal={Astronomy & Astrophysics},  author={Ball, W. H. and Gizon, L.},  year={2017},  pages={A128}} |
| Ball, W. H. et al. (2018)  @article{ball\_2018,  title={Surface effects on the red giant branch},  volume={478},  DOI={10.1093/mnras/sty1141},  number={4},  journal={Monthly Notices of the Royal Astronomical Society},  author={Ball, W H and Themeßl, N and Hekker, S},  year={2018},  pages={4697-4709}} |
| @article{jørgensen\_2005,  title={Determination of stellar ages from isochrones: Bayesian estimation versus isochrone fitting},  volume={436},  DOI={10.1051/0004-6361:20042185},  number={1},  journal={Astronomy & Astrophysics},  author={Jørgensen, B. R. and Lindegren, L.},  year={2005},  pages={127-143}} |
|  |
| @article{curtis\_2013,  title={RUPRECHT 147: THE OLDEST NEARBY OPEN CLUSTER AS A NEW BENCHMARK FOR STELLAR ASTROPHYSICS},  volume={145},  DOI={10.1088/0004-6256/145/5/134},  number={5},  journal={The Astronomical Journal},  author={Curtis, Jason L. and Wolfgang, Angie and Wright, Jason T. and Brewer, John M. and Johnson, John Asher},  year={2013},  pages={134}} |
| @article{bragaglia\_2018,  title={The chemical composition of the oldest nearby open cluster Ruprecht 147},  volume={619},  DOI={10.1051/0004-6361/201833888},  journal={Astronomy & Astrophysics},  author={Bragaglia, Angela and Fu, Xiaoting and Mucciarelli, Alessio and Andreuzzi, Gloria and Donati, Paolo},  year={2018},  pages={A176}} |
| @article{APOGEE,  title={The Apache Point Observatory Galactic Evolution Experiment (APOGEE)},  volume={154},  DOI={10.3847/1538-3881/aa784d},  number={3},  journal={The Astronomical Journal},  author={Majewski, Steven R. and Schiavon, Ricardo P. and Frinchaboy, Peter M. and Prieto, Carlos Allende and Barkhouser, Robert and Bizyaev, Dmitry and Blank, Basil and Brunner, Sophia and Burton, Adam and Carrera, Ricardo et al.},  year={2017},  pages={94}} |
| @article{GAIA\_mission,  title={The Gaia mission},  volume={595},  DOI={10.1051/0004-6361/201629272},  journal={Astronomy & Astrophysics},  author={Prusti, T. and de Bruijne, J. H. J. and Brown, A. G. A. and Vallenari, A. and Babusiaux, C. and Bailer-Jones, C. A. L. and Bastian, U. and Biermann, M. and Evans, D. W. and Eyer, L. et al.},  year={2016},  pages={A1}} |
| @article{GAIA\_DR2,  title={Gaia Data Release 2},  volume={616},  DOI={10.1051/0004-6361/201833051},  journal={Astronomy & Astrophysics},  author={Brown, A. G. A. and Vallenari, A. and Prusti, T. and de Bruijne, J. H. J. and Babusiaux, C. and Bailer-Jones, C. A. L. and Biermann, M. and Evans, D. W. and Eyer, L. and Jansen, F. et al.},  year={2018},  pages={A1}} |
| @article{bailer-jones\_2018,  title={Estimating Distance from Parallaxes. IV. Distances to 1.33 Billion Stars in Gaia Data Release 2},  volume={156},  DOI={10.3847/1538-3881/aacb21},  number={2},  journal={The Astronomical Journal},  author={Bailer-Jones, C. A. L. and Rybizki, J. and Fouesneau, M. and Mantelet, G. and Andrae, R.},  year={2018},  pages={58}} |
| @article{SIMBAD,  title={The SIMBAD astronomical database},  volume={143},  DOI={10.1051/aas:2000332},  number={1},  journal={Astronomy and Astrophysics Supplement Series},  author={Wenger, M. and Ochsenbein, F. and Egret, D. and Dubois, P. and Bonnarel, F. and Borde, S. and Genova, F. and Jasniewicz, G. and Laloë, S. and Lesteven, S. et al.},  year={2000},  pages={9-22}} |
| @article{sanders\_2018,  title={Isochrone ages for ∼3 million stars with the second Gaia data release},  volume={481},  DOI={10.1093/mnras/sty2490},  number={3},  journal={Monthly Notices of the Royal Astronomical Society},  author={Sanders, Jason L and Das, Payel},  year={2018},  pages={4093-4110}} |
| @article{green\_2019,  title={A 3D Dust Map Based on Gaia, Pan-STARRS 1, and 2MASS},  volume={887},  DOI={10.3847/1538-4357/ab5362},  number={1},  journal={The Astrophysical Journal},  author={Green, Gregory M. and Schlafly, Edward and Zucker, Catherine and Speagle, Joshua S. and Finkbeiner, Douglas},  year={2019},  pages={93}} |
| @article{andrae\_2018,  title={Gaia Data Release 2},  volume={616},  DOI={10.1051/0004-6361/201732516},  journal={Astronomy & Astrophysics},  author={Andrae, René and Fouesneau, Morgan and Creevey, Orlagh and Ordenovic, Christophe and Mary, Nicolas and Burlacu, Alexandru and Chaoul, Laurence and Jean-Antoine-Piccolo, Anne and Kordopatis, Georges and Korn, Andreas et al.},  year={2018},  pages={A8}} |
| @ARTICLE{MBOLSOL,  title = "{Book Review: Transactions of the IAU general assembly (23rd) / Kluwer, 1999}",  journal = {The Observatory},  year = 1999,  month = oct,  volume = {119},  number = {1152},  pages = {289}} |
|  |
| @article{li\_2018,  title={Asteroseismic modelling of the subgiant μ Herculis using SONG data: lifting the degeneracy between age and model input parameters},  volume={483},  DOI={10.1093/mnras/sty3000},  number={1},  journal={Monthly Notices of the Royal Astronomical Society},  author={Li, Tanda and Bedding, Timothy R and Kjeldsen, Hans and Stello, Dennis and Christensen-Dalsgaard, Jørgen and Deng, Licai},  year={2018},  pages={780-789}} |
| @article{cheng\_2020,  title={Optimizing automatic morphological classification of galaxies with machine learning and deep learning using Dark Energy Survey imaging},  volume={493},  DOI={10.1093/mnras/staa501},  number={3},  journal={Monthly Notices of the Royal Astronomical Society},  author={Cheng, Ting-Yun and Conselice, Christopher J and Aragón-Salamanca, Alfonso and Li, Nan and Bluck, Asa F L and Hartley, Will G and Annis, James and Brooks, David and Doel, Peter and García-Bellido, Juan et al.},  year={2020},  pages={4209-4228}} |
| @article{mckeever\_2019,  title={The Helium Abundance of NGC 6791 from Modeling of Stellar Oscillations},  volume={874},  DOI={10.3847/1538-4357/ab0c04},  number={2},  journal={The Astrophysical Journal},  author={McKeever, Jean M. and Basu, Sarbani and Corsaro, Enrico},  year={2019},  pages={180}} |
| @article{viani\_2017,  title={Isochrones of M67 with an Expanded Set of Parameters},  volume={160},  DOI={10.1051/epjconf/201716005005},  journal={EPJ Web of Conferences},  author={Viani, Lucas and Basu, Sarbani},  year={2017},  pages={05005}} |
| @article{brogaard\_2012,  title={Age and helium content of the open cluster NGC  6791 from multiple eclipsing binary members},  volume={543},  DOI={10.1051/0004-6361/201219196},  journal={Astronomy & Astrophysics},  author={Brogaard, K. and VandenBerg, D. A. and Bruntt, H. and Grundahl, F. and Frandsen, S. and Bedin, L. R. and Milone, A. P. and Dotter, A. and Feiden, G. A. and Stetson, P. B. et al.},  year={2012},  pages={A106}} |
| @article{miglio\_2011,  title={Asteroseismology of old open clusters with Kepler: direct estimate of the integrated red giant branch mass-loss in NGC 6791 and 6819},  volume={419},  DOI={10.1111/j.1365-2966.2011.19859.x},  number={3},  journal={Monthly Notices of the Royal Astronomical Society},  author={Miglio, A. and Brogaard, K. and Stello, D. and Chaplin, W. J. and D’Antona, F. and Montalbán, J. and Basu, S. and Bressan, A. and Grundahl, F. and Pinsonneault, M. et al.},  year={2011},  pages={2077-2088}} |
| @article{hendriks\_2019,  title={Deep Learning Applied to the Asteroseismic Modeling of Stars with Coherent Oscillation Modes},  volume={131},  DOI={10.1088/1538-3873/aaeeec},  number={1004},  journal={Publications of the Astronomical Society of the Pacific},  author={Hendriks, L. and Aerts, C.},  year={2019},  pages={108001}} |
| @article{bellinger\_2016,  title={FUNDAMENTAL PARAMETERS OF MAIN-SEQUENCE STARS IN AN INSTANT WITH MACHINE LEARNING},  volume={830},  DOI={10.3847/0004-637x/830/1/31},  number={1},  journal={The Astrophysical Journal},  author={Bellinger, Earl P. and Angelou, George C. and Hekker, Saskia and Basu, Sarbani and Ball, Warrick H. and Guggenberger, Elisabeth},  year={2016},  pages={31}} |
| @misc{NNbook,  added-at = {2019-01-15T22:46:49.000+0100},  author = {Nielsen, Michael A.},  biburl = {https://www.bibsonomy.org/bibtex/274383acee84241145ff4ffede9658206/slicside},  interhash = {04d527cadd39f888fc3babcad3343362},  intrahash = {74383acee84241145ff4ffede9658206},  keywords = {ba-2018-hahnrico},  publisher = {Determination Press},  timestamp = {2019-01-15T22:46:49.000+0100},  title = {Neural Networks and Deep Learning},  type = {misc},  url = {http://neuralnetworksanddeeplearning.com/},  year = 2018  } |
| @article{Bailer\_Jones\_2002,  author = {Bailer-Jones C. A. L., Gupta R., Singh H. P.},  title = "{An Introduction to Artificial Neural Networks}",  keywords = {Astrophysics},  booktitle = {Automated Data Analysis in Astronomy},  year = 2002,  editor = {{Gupta}, Ranjan and {Singh}, Harinder P. and {Bailer-Jones}, Coryn A.~L.},  month = jan,  pages = {51},  archivePrefix = {arXiv},  eprint = {astro-ph/0102224},  primaryClass = {astro-ph},  adsurl = {https://ui.adsabs.harvard.edu/abs/2002adaa.conf...51B},  adsnote = {Provided by the SAO/NASA Astrophysics Data System}  } |
| @ARTICLE{prevent\_overfitting,  author = {{Ghojogh}, Benyamin and {Crowley}, Mark},  title = "{The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial}",  journal = {arXiv e-prints},  keywords = {Statistics - Machine Learning, Computer Science - Machine Learning},  year = 2019,  month = may,  eid = {arXiv:1905.12787},  pages = {arXiv:1905.12787},  archivePrefix = {arXiv},  eprint = {1905.12787},  primaryClass = {stat.ML},  adsurl = {https://ui.adsabs.harvard.edu/abs/2019arXiv190512787G},  adsnote = {Provided by the SAO/NASA Astrophysics Data System}  } |
| @misc{keras,  title={Keras},  author={Chollet, Fran\c{c}ois and others},  year={2015},  howpublished={\url{https://keras.io}},  } |
| @article{Nadam,  title={Incorporating Nesterov Momentum into Adam},  conference={ICLR Workshop},  author={Timothy Dozat},  year={2016},  } |
| @article{Nadam2,  title={On the importance of initialization and momentum in deep learning},  conference={In Proceedings of the 30th International Conference on Machine Learning (ICML-13)},  author={I. Sutskever, J. Martens, G. Dahl, G. Hinton},  year={2013},  } |
| @ARTICLE{lrbatch,  author = {{Smith}, Samuel L. and {Kindermans}, Pieter-Jan and {Ying}, Chris and  {Le}, Quoc V.},  title = "{Don't Decay the Learning Rate, Increase the Batch Size}",  journal = {arXiv e-prints},  keywords = {Computer Science - Machine Learning, Computer Science - Computer Vision and Pattern Recognition, Computer Science - Distributed, Parallel, and Cluster Computing, Statistics - Machine Learning},  year = 2017,  month = nov,  eid = {arXiv:1711.00489},  pages = {arXiv:1711.00489},  archivePrefix = {arXiv},  eprint = {1711.00489},  primaryClass = {cs.LG},  adsurl = {https://ui.adsabs.harvard.edu/abs/2017arXiv171100489S},  adsnote = {Provided by the SAO/NASA Astrophysics Data System}  } |
| @ARTICLE{Dinescu\_1995,  author = {{Dinescu}, Dana I. and {Demarque}, Pierre and {Guenther}, D.~B. and  {Pinsonneault}, M.~H.},  title = "{The Ages of the Disk Clusters NGC 188, M67, and NGC 752, Using Improved Opacities and Cluster Membership Data}",  journal = {\aj},  keywords = {OPEN CLUSTERS AND ASSOCIATIONS: INDIVIDUAL: NGC 188, OPEN CLUSTERS AND ASSOCIATIONS: INDIVIDUAL: M 67, OPEN CLUSTERS AND ASSOCIATIONS: INDIVIDUAL: NGC 752},  year = 1995,  month = may,  volume = {109},  pages = {2090},  doi = {10.1086/117434},  adsurl = {https://ui.adsabs.harvard.edu/abs/1995AJ....109.2090D},  adsnote = {Provided by the SAO/NASA Astrophysics Data System}  } |
| @article{si\_2018,  title={Bayesian hierarchical modelling of initial–final mass relations acrossstar clusters},  volume={480},  DOI={10.1093/mnras/sty1913},  number={1},  journal={Monthly Notices of the Royal Astronomical Society},  author={Si, Shijing and van Dyk, David A and von Hippel, Ted and Robinson, Elliot and Jeffery, Elizabeth and Stenning, David C},  year={2018},  pages={1300-1321}} |
| @article{hippel\_2014,  title={The power of principled bayesian methods in the study of stellar evolution},  volume={65},  DOI={10.1051/eas/1465007},  journal={EAS Publications Series},  author={von Hippel, T. and van Dyk, D.A. and Stenning, D.C. and Robinson, E. and Jeffery, E. and Stein, N. and Jefferys, W.H. and O'Malley, E.},  year={2014},  pages={267-287}} |
| @INBOOK{HBMbook,  author = {{Loredo}, Thomas J. and {Hendry}, Martin A.},  title = "{Bayesian multilevel modelling of cosmological populations}",  booktitle = {Bayesian Methods in Cosmology},  year = 2010,  editor = {{Hobson}, Michael P. and {Jaffe}, Andrew H. and {Liddle}, Andrew R. and  {Mukeherjee}, Pia and {Parkinson}, David},  pages = {245},  adsurl = {https://ui.adsabs.harvard.edu/abs/2010bmic.book..245L},  adsnote = {Provided by the SAO/NASA Astrophysics Data System}  } |
| @ARTICLE{HBMpaper,  author = {{Loredo}, Thomas J. and {Hendry}, Martin A.},  title = "{Multilevel and hierarchical Bayesian modeling of cosmic populations}",  journal = {arXiv e-prints},  keywords = {Astrophysics - Instrumentation and Methods for Astrophysics, Statistics - Applications},  year = 2019,  month = nov,  eid = {arXiv:1911.12337},  pages = {arXiv:1911.12337},  archivePrefix = {arXiv},  eprint = {1911.12337},  primaryClass = {astro-ph.IM},  adsurl = {https://ui.adsabs.harvard.edu/abs/2019arXiv191112337L},  adsnote = {Provided by the SAO/NASA Astrophysics Data System}  } |
|  |
| @article{cantat-gaudin\_2018,  title={A Gaia DR2 view of the open cluster population in the Milky Way},  volume={618},  DOI={10.1051/0004-6361/201833476},  journal={Astronomy & Astrophysics},  author={Cantat-Gaudin, T. and Jordi, C. and Vallenari, A. and Bragaglia, A. and Balaguer-Núñez, L. and Soubiran, C. and Bossini, D. and Moitinho, A. and Castro-Ginard, A. and Krone-Martins, A. et al.},  year={2018},  pages={A93}} |
| @article{gao\_2018,  title={A Machine-learning-based Investigation of the Open Cluster M67},  volume={869},  DOI={10.3847/1538-4357/aae8dd},  number={1},  journal={The Astrophysical Journal},  author={Gao, Xinhua},  year={2018},  pages={9}} |
|  |

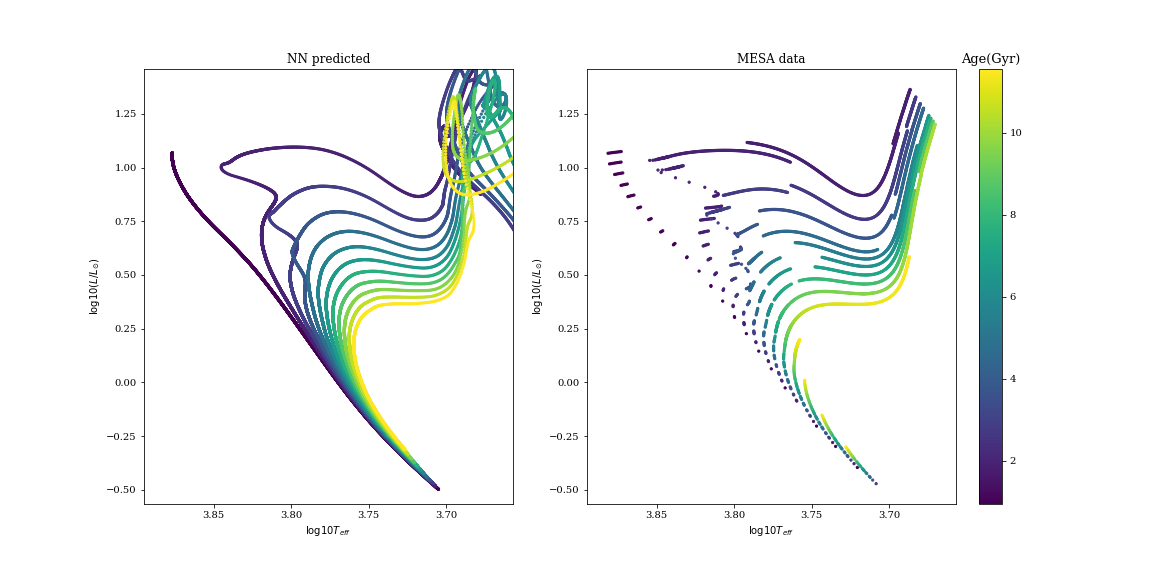
**Appendix A:**  
- observed data table

**Appendix B:**  
**Table B.1**: Training schedule of MS+SG NN. ‘prev’ = Use previous step’s decayed learning rate

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Training step | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Learning rate | 0.0001 | 0.0002 | 0.0005 | 0.001 | prev | prev | prev | prev | prev | prev |
| Batch size | 1x10^3 | 2x10^3 | 5x10^3 | 2x10^4 | 5x10^4 | 2x10^5 | 5x10^5 | 1x10^6 | 2x10^6 | Max\_batch |
| Epochs | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 48100 |

**Table B.2**: Neural network hyperparameters kept constant during MS+SG NN training. The hyperparameters of the Nadam optimizer are kept as the keras default.

|  |  |
| --- | --- |
| Hyperparameter | value |
| Architecture | 128x10 |
| Weight initialization | Set seed as 53 |
| Optimizer | Nadam |
| Activation function | Elu |
| regression loss function | MAE |
| Batch normalization | Only on the input layer |
| Regularization | L2, 1x10^{-6} |
| Dropout | None |

 **Figure B.1**: MS+SG NN predicted isochrones (left) and truncated MESA training grid (right).

**Appendix C**:  
A separate grid was used to train the NGC 6791 NN where the parameter ranges are shown in **Table C.1**.

**Table C.1**: NGC 6791 NN training grid.

|  |  |  |
| --- | --- | --- |
| Input Parameter | range | increments |
| Age / Gyr | 5 - 12 | varied |
| M\_{init} / M\_{\odot} | 0.8 - 1.52 | 0.04 |
| [Fe/H]\_{init} | 0 – 0.6 | 0.2 |
| Y\_{init} | 0.26 – 0.32 | 0.02 |
| αMLT | 1.7 - 2.3 | 0.2 |

**Table C.2**: Training schedule of NGC 6791 NN. ‘prev’ = Use previous step’s decayed learning rate

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Training step | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Learning rate | 0.0001 | 0.0002 | 0.0005 | 0.001 | prev | prev | prev | prev | prev | prev |
| Batch size | 1x10^2 | 2x10^2 | 5x10^2 | 1x10^3 | 2x10^3 | 5x10^3 | 2x10^4 | 5x10^4 | 2x10^5 | Max\_batch |
| Epochs | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 9000 |

The training grid for NGC 6791 has (approximately) 1/10 the number of points of the MS+SG NN grid which is why the batch sizes are 1/10 the size.

**Table C.3**: Neural network hyperparameters kept constant during NGC 6791 NN training. The hyperparameters of the Nadam optimizer are kept as the keras default.

|  |  |
| --- | --- |
| Hyperparameter | value |
| Architecture | 100x10 |
| Weight initialization | Set seed as 53 |
| Optimizer | Nadam |
| Activation function | Elu |
| regression loss function | MAE |
| Batch normalization | Only on the input layer |
| Regularization | L2, 1x10^{-5} |
| Dropout | None |

CONCERNS:  
- introduction doesn’t discuss papers but more references methods that those papers use SEE NEURAL NETWORK SECTION FOR ALTERNATE METHOD OF WRITING ABOUT PAPERS  
- make sure to properly subscript Teff and A\_G etc.  
- in acknowledgements do we have to acknowledge keras and MESA, GAIA, pymc3 etc.?  
- acknowledge SDSS and everything  
- is it okay to copy paste this whole GAIA acknowledgement chunk of text?  
- consider writing neural networks as NN  
- Cybenko  
- do appendices need titles?  
- **for project work 2 I suggest you make the Github repo public and submit the link to the repo.**

- limitations of using single G-band extinction coefficient  
- limit of Fe/H step fineness  
- harder to fit isochrone when you don’t have RGB stars  
- could cite Silva Aguirre et al. 2015 for an extended discussion on the various methods of dating stars