Use of Neural Network Machine Learning Models in Calculating Peculiar Velocity and Galactic Distances

by

Kaiden Elam

Submitted in Partial Fulfillment of the
Requirements for the Degree

Bachelor of Science

Supervised by
Dr. Rick Watkins

Department of Physics

Willamette University
College of Arts & Sciences
Salem, Oregon

2024

Presentations and publications:

"Using Machine Learning to Calculate Galactic Distances," SSRD Thesis Presentation. April 17th, 2024.

"Using Neural Networks to Measure the Distances to Galaxies," Fall Semester Thesis Progress Reports. November 30th, 2023.

"Using Machine Learning to Calculate the Distances to Galaxies," ATEP Thesis Proposal Presentations. May 2nd, 2023.

"Using Machine Learning to Find Galactic Distances," SCRP Symposium 2022. September 16th, 2022.

"PIXL Calibration & PIQUANT Documentation," JPL Summer Internship Program Final Student Presentations. July 26th, 2021.

Acknowledgments

My success in this project was made possible by having worked on an earlier version of it as part of SCRP 2022.

A tremendous thank-you to Rick Watkins, my advisor on this project.

Thanks to the Willamette physics department overall, for a fantastic four years, and the physics class of 2024 specifically. We did it! Through a global pandemic, wildfire seasons, and on one memorable occasion an ice storm.

To my AP Physics teacher, Mr. Gehring–thanks for helping me take the first steps that led me here, and showing me physics is something I was capable of.

And of course my family. Thanks for the support along the way.

Abstracts

Technical Abstract

This project used neural network models to calculate the peculiar velocity and distance values of spiral galaxies, with models constructed based on the baryonic Tully-Fisher Relation (bTFR). The novel approach of this project is that it used measured redshifts as target values to predict cosmological redshift values, and used these redshifts to recover peculiar velocities. Since peculiar velocity is not correlated with any properties of the galaxy, the peculiar velocity contribution to the target redshift value was averaged out by the model. Cosmicflows-4 (CF4) data and a set of simulated data based on CF4 were used to test the models. A correlation was found between predicted and reference (simulated or bTFR) peculiar velocities for both datasets, indicating neural networks can be used to calculate peculiar velocities and distance values.

General Abstract

This project used machine learning models to calculate the distances to spiral galaxies. The models were based on the baryonic Tulley-Fisher Relation, a current distance method for spiral galaxies. The Cosmicflows-4 (CF4) dataset was used in this project, as well as simulated data based on CF4 values. CF4 simulated data were used to construct a method based solely on distance measurements, as well as a novel method that did not require reliable distance comparison values and as such could be used for the real data. This second method provided usable results for both simulated and CF4 data, showing that galactic distance values can be predicted using machine learning models.

Table of Contents

List of Figures

1 Introduction

One of the thorniest problems in cosmology is finding an accurate method that can be used to determine galactic distances. Standing on the surface of the Earth, it is very difficult to tell how far away anything is; some objects are brighter and some are dimmer, but that has very little real correlation to distance. An object can be very far away but very bright, and thus easier to see in the sky than a close dim object. Gaining a sense of depth perception on the universe is an important part of being able to understand space, and the main way to do this is by finding distances to the lights in the sky.

There are a large number of existing distance methods that make up the cosmological distance ladder, including the standard candle measurements from the Tip of the Red Giant Branch and Cepheid variable stars. These two methods produce very reliable distances, but are only applicable to a small number of nearby galaxies. Thus, methods that work at larger distances must be found. Other distance methods in popular use include the Fundamental Plane calculation, which can be used on elliptical galaxies, and the baryonic Tully-Fisher relation (bTFR) for spiral galaxies. Of particular interest to this project is the bTFR.

This project uses neural networks, a type of machine learning algorithm, to calculate distances to galaxies. Rather than using a complex empirical relationship, such as the bTFR, the goal is to produce an algorithm that provides high-accuracy distance values when given information about a galaxy. By providing a set of values about a given galaxy and a target value, a neural net will determine relationships between the input values that produce an output as close to the target value as possible. Both distance and redshift values were used as targets in this project, with the redshift model providing a method that did not require known comparison values to find accurate distance and peculiar velocity data.

Both real Cosmicflows-4 (CF4) data (available online through the Extragalactic Distance Database [1]) and simulated data based on the Cosmicflows-4 dataset are used in this project. The simulated data is provided by Kourkchi et al., who have been working with the baryonic Tully-Fisher Relation and the CF-4 dataset to find galactic distances [2]. The simulated data includes known distance val-

ues, which means it is can be used as a calibration dataset for models produced. Simulated data can be used to create a neural net model that is as accurate as possible, which can then be applied to the real CF4 data.

While much ink has been spilt on the bTFR as a distance method since it was discovered in 2000, the use of machine learning in cosmology is not nearly as well investigated. Machine learning has been shown to be a useful tool, and this project aims to show it can be used to produce distance values that are more accurate than those reported in Cosmicflows-4.

Background necessary to understand this project will be provided, as well as a discussion of the CF4 dataset. The processes undertaken will be discussed, both using the simulated data to determine a viable approach, and then applying that approach to the CF4 dataset. Finally, the results of both these investigations will be discussed, including how the resulting CF4 machine learning model can be used as a distance indicator for spiral galaxies.

2 Background

2.1 Cosmological Distance Measurements

2.1.1 The Magnitude System

The magnitude of an object is a measure of its brightness. An object has two types of magnitude: the absolute magnitude M and the apparent magnitude m.

The apparent magnitude of an object is how bright it appears to be from some observation point, typically the Earth (or a telescope in Earth orbit) [3]. This quantity depends on distance, as light diffuses in all directions between the object being observed and the observer. Apparent magnitude is calculated using photometry and flux values. An object's apparent magnitude can be calculated in specific photometric passbands, or sections of the photometric spectrum. A passband only allows certain wavelengths through. Photometric passbands used to measure magnitudes can be in the ultraviolet, visible, or infrared parts of the photometric spectrum. Apparent magnitude can be calculated with the following equation:

$$m = -2.5 \log \left(\frac{F_x}{F_{x,0}} \right), \tag{2.1}$$

where F_x is the observed radiant flux and $F_{x,0}$ the reference flux, both in the passband being studied [3].

The absolute magnitude of an object is its intrinsic brightness, which is a measure of the luminosity of the object in question [4]. Absolute magnitude is defined specifically as the apparent magnitude as viewed from 10 parsecs away. It can be calculated in terms of distance d and apparent magnitude m.

The distance modulus μ of a galaxy is the difference between its apparent magnitude (m) and its absolute magnitude (M) [4]:

$$\mu = m - M. \tag{2.2}$$

Distance modulus is related to distance d through the following equation:

$$\mu = 5 \log \left(\frac{d}{10\text{pc}} \right). \tag{2.3}$$

Distance moduli is one of the main ways used to report galactic distances.

2.1.2 The Cosmological Distance Ladder

Measuring the distance to cosmological objects is a difficult task, so there are many methods that can be used. These methods are accurate (or usable) at different distances, which allows us to construct what is known as the cosmological distance ladder: a list of what distance methods can be used based on how far away an object is, and a way to determine how accurate that measurement will be. Each rung of the ladder is used to calibrate the next one.

A general rule is that the closer the object is, the more accurate the measurement will be. However, the accuracy also depends on the method that is used to obtain that measurement.

Objects with luminosities that can be determined by making distance-independent measurements are called standard candles. These objects are an important part of cosmological distance calculations, since they can act as calibration factors for other distance methods. If the absolute magnitude M of an object is known, that value can be used to find the distance modulus μ by using Equation 2.2.

The first rung of the cosmological distance ladder is parallax, the apparent shift of a star in the sky as the Earth orbits the Sun [5]. By measuring the angle created between a reference star that does not appear to move and the star being observed at two different times, the distance to the star can be calculated using basic trigonometry. This is shown in Figure 2.1.

One of the most common and reliable examples of standard candles are a class of stars called cepheid variable stars. These stars are fairly rare, but have a luminosity that oscillates over time and is very predictable. The period of oscillation is correlated with the absolute magnitude M, as can be seen in Figure 2.2, and as such the period can be used to find the luminosity of the star [6].

Since the absolute luminosity of a cepheid is easy to determine from its period, they provide excellent distance indicators that can be used as calibration values for other distance methods. Cepheids can only be identified in nearby galaxies, but when a galaxy that is being measured with a separate distance method does contain a cepheid, the distance measurements from each method can be compared and the non-cephid method can be calibrated using the cepheid value.

Other distance methods that work further out in the ladder include measurements of Tip of the Red Giant branch stars, and (as will be used in this project)

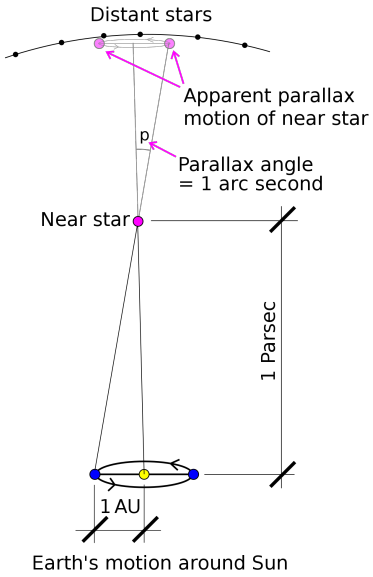

Figure 2.1: A diagram showing the parallax angle for a star 1 parsec from the Sun. This angle measurement is 1 arcsecond, and the basis of the parsec unit. [5]

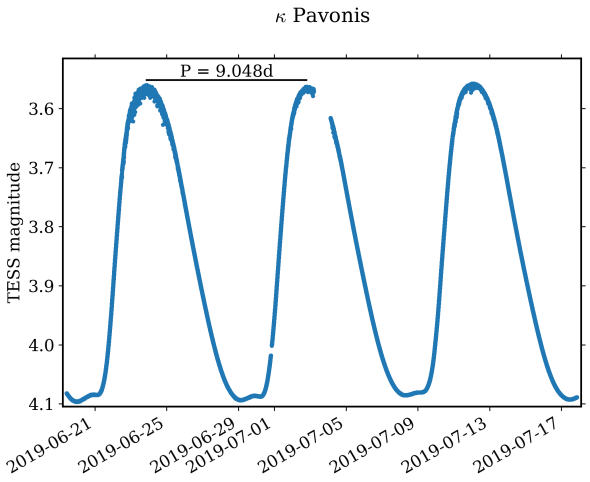

Figure 2.2: The light curve of κ Pavonis, a type II cepheid. Data to produce this graph were recorded by NASA's Transiting Exoplanet Survey Satellite (TESS). Notice the oscillation of the magnitude over a period of time. [7]

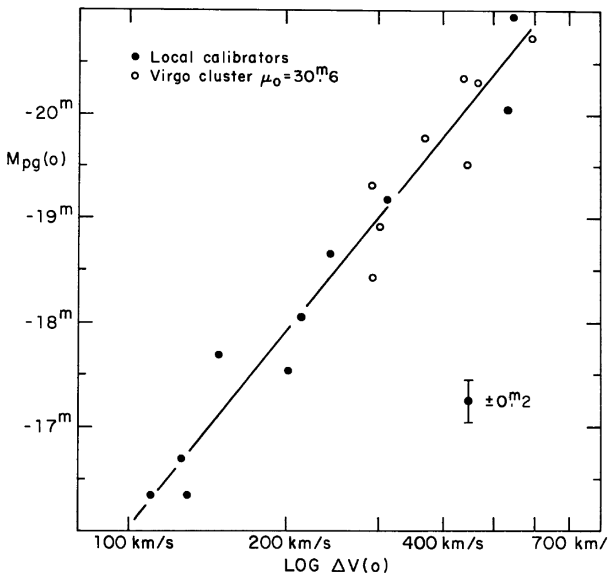

Figure 2.3: Figure 5(a) from [8], showing the Tully-Fisher absolute magnitude (here $M_{pg(o)}$) to rotation velocity ($\log \Delta V_{(o)}$) relation of galaxies from Virgo cluster galaxies and several others with known distances. The best visual fit of the relation is shown.

the baryonic Tully-Fisher relation. Accurate distance values can be used in the peculiar velocity calculation, in order to calculate how a galaxy is moving in space. This calculation combines redshift distance information with an independently-calculated distance measurement to produce a peculiar velocity value, which describes how a galaxy is moving in space independent of the expansion of the universe (cosmological redshift). The total redshift cz is a combination of cosmological redshift r and Doppler shift, which relies on peculiar velocity v:

$$cz = H_0 r + v. \tag{2.4}$$

2.1.3 The Tully-Fisher Relation

The Tully-Fisher relation (TFR), as first published in 1976, is a correlation between the H1 profile width of a spiral galaxy and its absolute magnitude [8] (see Figure 2.3). H1 profile width is a measure of a galaxy's rotation velocity: by observing a galaxy from the side it can be seen that one side of the galaxy is redshifted and the other side is blueshifted as that galaxy rotates. The red- and blueshift of the galaxy widens the H1 line, which can be used to calculate a value for rotation velocity.

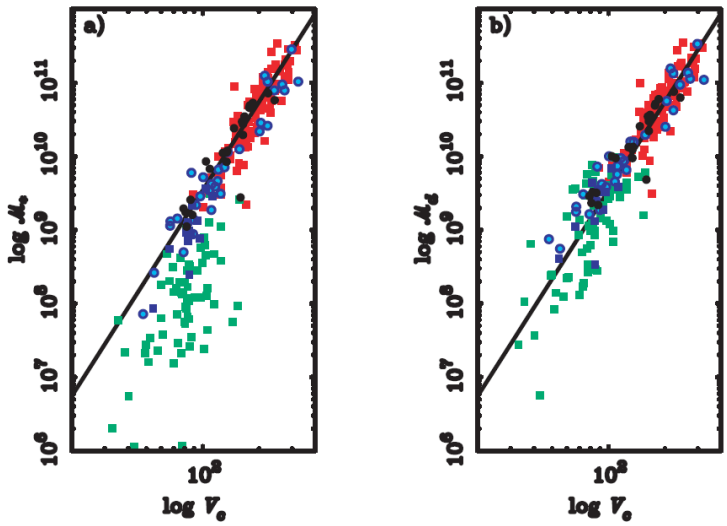

Figure 2.4: Figure 1 from [9]. Graph *a.*) plots stellar mass against rotation velocity, as in the standard Tully-Fisher relation, while graph *b.*) plots baryonic mass against rotation velocity–the baryonic Tully-Fisher relation. Notice the many green points in figure *a.*) that are off the linear fit are well accounted for in figure *b.*).

Given the rotation velocity of a galaxy, the TF relation can be used to determine a value for the galaxy's absolute magnitude M. The apparent magnitude of a galaxy is determined by observing the galaxy, and thus is an easy quantity to find. These two values can be used in Equation 2.2 to determine the distance to the galaxy in question, which is how distance is calculated using the Tully-Fisher relation.

2.1.4 The Baryonic Tully-Fisher Relation

The baryonic Tully-Fisher relation (bTFR) is an extension of the Tully-Fisher relation, designed to be more accurate for galaxies with noticeable gas mass contributions. The smaller a galaxy is, the more of an effect the gas mass has on its overall mass, and the less accurate the TFR gets. The TFR using the absolute magnitude of a given galaxy means it is only accounting for the stellar mass, since absolute magnitude is based on the luminosity of the star, which results in the breakdown for smaller galaxies. The baryonic mass accounts for the total mass of the galaxy, and as such produces a far more accurate relation for galaxies with non-negligible gas masses. Gas mass becomes non-negligible in galaxies of about 10^9 solar masses [10]. This is illustrated in Figure 2.4.

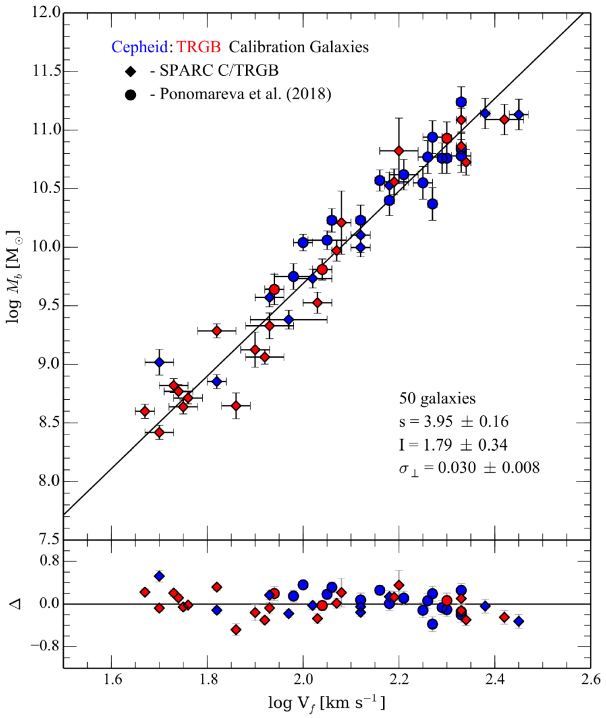

Figure 2.5: Figure 1 from [10], showing the bTFR diagram for TRGB (red) and cepheid (blue) galaxies. The linear fit is shown with parameters, as well as the residuals of the relation on the lower plot. Methods for calculating V_f and M_b are described in the original source.

The baryonic Tully-Fisher Relation accounts for the galaxies with low gas mass:solar mass ratios by correlating baryonic mass (rather than absolute magnitude) to rotation velocity, as is shown in Figure 2.5. The baryonic mass M_b of a galaxy is its stellar mass M_* plus its gas mass M_g, or $M_b = M_* + M_g$. M_* and M_g can be calculated using distance-dependant equations.

To find distance using this method, the baryonic mass of the galaxy must be calculated using the baryonic Tully-Fisher relation, as shown in Figure 2.5. The baryonic mass value produced by the bTFR becomes a reference value. Since baryonic mass can also be calculated as a distance-dependant value, distance curves can be produced over a range of test distances. Baryonic mass is a combination of the stellar and gas mass of the galaxy, both of which are distance-dependant values. The gas mass of a galaxy can be calculated using distance and the flux in the H1 line. The stellar mass of a galaxy can be calculated from distance and a set of luminosity measurements conducted in several passbands, along with several reference values such as the mass of our sun. Once produced, these distance curves can be added to produce a baryonic mass curve and compared to the reference

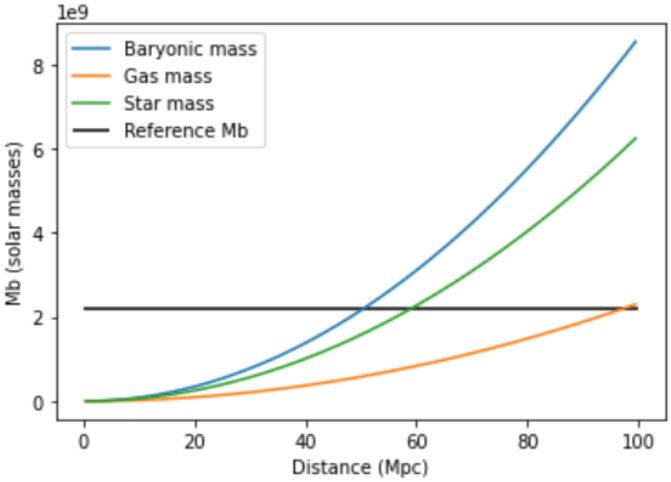

Figure 2.6: The method via which distance can be calculated using the baryonic Tully-Fisher method. Baryonic mass as determined using the bTFR is the reference M_b line at around 2 solar masses. Baryonic mass M_b can be calculated over a range of test distances by adding distance-dependant values for M_g and M_*, generating distance curves that are then compared to the reference value. The gas mass M_g and stellar mass M_* curves are also shown.

baryonic mass produced by the bTFR, and a distance measurement is obtained by finding the intersection of the mass curves with the reference value. This process is illustrated in Figure 2.6.

2.2 Neural Networks

A neural network, or neural net, is a type of machine learning algorithm. At the most basic level, it acts as a generalization of a linear fit. The neural net is composed of layers, which are in turn composed of nodes.

As can be seen in Figure 2.7, a neural net takes a certain number of inputs and combines them to get an output. A target value is also provided. Each layer takes as inputs the outputs of the previous layer. Layers are composed of nodes, which do the real work of providing a fit. Each node combines the values from every node on the layer just before its own, assigning them weights and producing a value. A node with i inputs will be assigned a value V_{node} given by:

$$V_{node} = \sum_{n=1}^{i} W_n I_n, \tag{2.5}$$

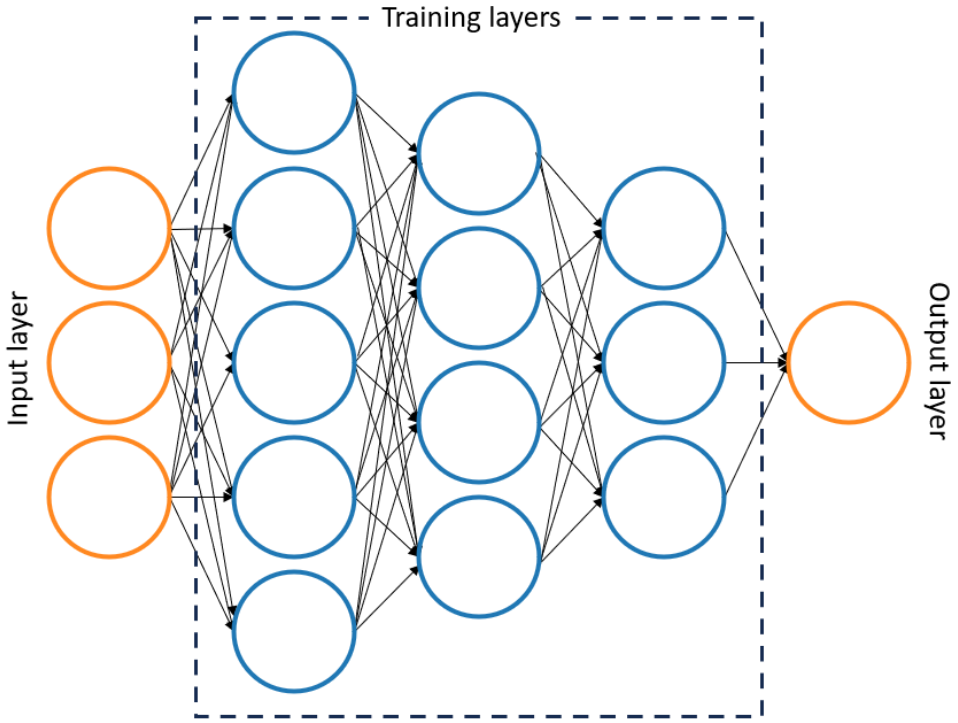

Figure 2.7: A visual representation of a single epoch of a neural network. Each node in a layer combines and assigns weights to the values from the previous layer, and then passes on its combined value if it is close to the target value (as determined by the node's activation function). This process repeats for all the training layers, which then combine to an output layer.

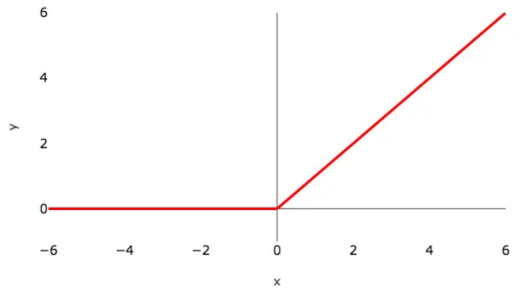

Figure 2.8: A graph of the Rectified Linear Unit (ReLU) activation function. Values calculated for each node are fed through this function, and nodes with values over zero produce outputs. On this plot, x is the value of the node, and y is the value passed to the next node. [11]

where each W_n is the weight assigned to an input value I_n. This weight assignment is how neural nets can be considered generalizations of linear fits. The node's activation function determines in what form the combined value is passed on to the next layer as an input [11]. The activation function of a node is an important part of the model, since it is what allows a neural net to produce a nonlinear model. The activation function used in this project, ReLU (Rectified Linear Unit), is illustrated in Figure 2.8. Depending on the weighted sum value assigned to the node, the function "activates" and calculates what is passed on to the next node based on the node's value. In this case, any values above zero are sent out. Once one layer has completed its calculations and produced outputs, the next layer of the model will in turn assign weights to the previous layer's nodes in an attempt to gain accuracy. Once all the layers have been processed, a final value is produced. A neural net is iterative in that it is run multiple times for a given number of epochs, and the weights that survive this process are the ones that produce a final value that is as close to the reference value as possible. This process is known as training a model.

A neural net tends to be more accurate than a standard linear fit, since the way the variables are combined is much more adaptable: they can model nonlinear relationships. They don't often produce equations that can be extracted, but the models are easy to recreate given a set of parameters. Once a model is trained, it can take a set of similarly-shaped inputs to the training data and predict them to produce an output value based on how the model was trained.

3 Data

3.1 Cosmicflows-4

The Cosmicflows-4 (CF4) dataset is a collection of galactic distance estimates for 55,877 galaxies in 38,065 groups. Of these galaxies, 12,412 have distance data provided by the Tully-Fisher or baryonic Tully-Fisher relations [12]. These 12,412 galaxies are the ones that are of interest in this project.

CF4 is a useful dataset because it provides a far larger number of galaxies than any previous Cosmicflows dataset. Cosmicflows datasets have come a long way since the first release in 2008, which provided values for 1791 galaxies. In particular, CF4 provides TFR samples that balance out the samples provided by the CF3 dataset, many of which are concentrated in the galactic south. Data from the Arecibo Legacy Fast Arecibo (ALFA) L-band Feed Array, supplemented by the Green Bank and Parkes telescopes, provided a large number of TFR and bTFR distances in the galactic north. Overall, the CF4 dataset provides bTFR samples across the sky, as can be seen in Figure 3.1. The Cosmicflows-4 dataset is described in detail in Reference [12] and is available online in the Extragalactic Distance Database [1].

CF4 data will be used in this project as input data for the neural nets produced. Relevant CF4 values used for models are as follows: rotation linewidth $\log W^i_{mx}$; H1 line flux $F21$; and apparent magnitude values in the g, r, i, z, w1, and w2 photometric passbands; as well as the redshift measured in the reference frame provided by the cosmic background radiation ($v3k$), and of course distance moduli values.

3.2 CF4 Simulated Data

An existing study [2] has used the baryonic Tully-Fisher relation to estimate distances for a subset of CF4 galaxies, which will be used as comparison values

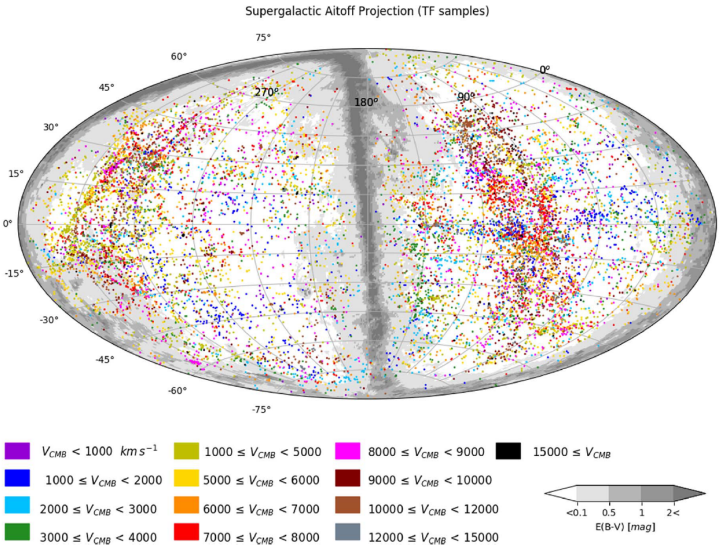

Figure 3.1: A projection in galactic coordinates of the TFR galaxies in CF4. The colors are related to the velocity of each galaxy. Extinction levels where the Milk Way blocks our view of the sky is shown in gray. Note how the galaxies shown span the range of the sky: the TFR galaxies form a good representative sample of the sky. Figure 2 from Reference [12].

for the outputs of the neural net models used in this project. The authors of this study have also produced a set of simulated data that mimics the CF4 dataset. This simulated data will be used to test methods and ensure they are viable, before they are applied to the real CF4 data.

The values in the simulated data are: Catalogue of Principle Galaxies (PGC) value, distance modulus, error in distance modulus, radial velocity (redshift) in CMB frame, a linewidth value related to rotation ($\log W^i_{mx}$), H1 line flux ($F21$), and the fully-corrected WISE w1 band apparent magnitude value. The values provided that were of use in this project were the linewidth value related to rotation ($\log W^i_{mx}$), H1 line flux ($F21$), and the fully-corrected WISE w1 band apparent magnitude value; as well as both distance modulus and CMB redshift. The simulated data set does not include peculiar velocity information, so for methods involving peculiar velocity comparison data those values had to be generated. These generated peculiar velocity values could be added to the simulated CMB redshift values, producing typical redshift values that include both peculiar velocity and the CMB redshift.

4 Methods

4.1 Distance Modulus Method

The first step of this project was to ensure that neural networks could be used to calculate distances, specifically distance moduli. The simulated data was used for this approach. The simulated data have three values that can be used as inputs to the model: linewidth related to rotation ($\log W^i_{mx}$), H1 line flux ($F21$), and fully-corrected WISE w1 band value. It includes as well the CMB redshift of each galaxy and distance modulus values. The main trial-and-error process of this section was producing a neural network that would combine those three input values and return an accurate distance modulus. A distance modulus calculated from the redshift distance (based on Equation 2.2) was used as the target value for the neural net training:

$$DM_{cz} = 5 \log \left(\frac{cz}{75} \right) + 25. \tag{4.1}$$

There was some noise in the simulated data from the peculiar velocities included in the redshift distance, which defined the expected results of the model. A successful model was one that produced distance modulus data with a standard deviation from the simulated distance moduli within the noise value–in this case about 0.25.

The approach that provided the best results in this situation was to average the results of running three similar neural networks. The basic model used in this averaging process had an input layer of three values (the number of inputs available in the simulated data), and training layers of five nodes, four nodes, three nodes, two nodes, with a final output later of one node to flatten the data to a single distance modulus value, as can be seen in Figure 4.1. The other two neural networks in the average had the four-node layer and the three-node layer, respectively, removed from the model to ensure variation in the model outputs. Running three separate models produced better results than running the same model three times.

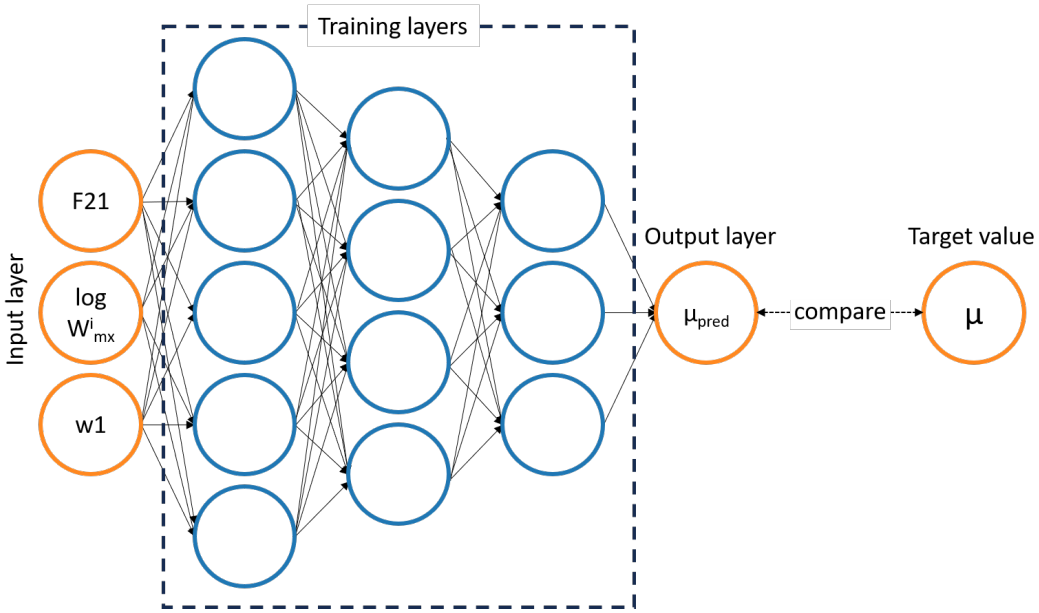

Figure 4.1: A visual representation of the neural network model that was used to predict distance modulus data using the simulated CF4 dataset. The three input values in the dataset are inputs to the model, which then has three training layers, of five, four, and three nodes respectively. Finally, the last training layer is flattened to a single distance modulus output value, which is compared to the distance modulus provided in the simulated dataset.

As mentioned, the main way that model results using the simulated data were analysed was by comparing the outputs of the model to the known distance moduli of the simulated data. This took the form of graphing these values, predicted data vs. simulated data. The residuals were also graphed against the simulated distance moduli, and the standard deviation of the residuals was found to be useable as an average uncertainty value. The standard deviation was compared to the noise in the simulated data.

Every model would inevitably have some outliers, which were always obvious on the graphs, particularly the residuals graph. By limiting the residuals by eye, the outliers could be excluded to produce a more accurate model. By keeping track of how many values are in the original dataset and how many values are in the limited dataset, it can be assured only a minimal number of outliers were excluded.

4.2 Redshift and Peculiar Velocity Method

Once the distance modulus analysis proved that neural networks could be used to predict distance data, the next step was to find a method that did not require known comparison values. Typically, the value being predicted by a neural net has a set of data where that value is known, which can be used as a target to train the model. However, non-simulated distance data is not accurate enough to use as target values for training a neural network. While the simulated data provides known distance values, and thus can be used to test other ways of using a neural net to predict distance, a method had to be found that does not require known training values. This led to the method that was the focus of this project: using redshift target data to predict the cosmological redshift of a galaxy, and then calculating its peculiar velocity. The neural network predicts the cosmological redshift of a galaxy, a quantity that does not include peculiar velocity information, which can then be subtracted from the original full redshift value of a galaxy (cosmological redshift + peculiar velocity) to return a peculiar velocity value. The simulated data were used to produce a method that achieved this goal, a method which, once it worked, was applied to the CF4 data.

This method begins with the redshift values given in the simulated data, which are the cosmological redshift values and as such contain no peculiar velocity information. Peculiar velocities on a Gaussian with a standard distribution of $300km/s$ were added to these redshift values, and this value (cosmological redshift + Gaussian peculiar velocity) was used as target values for a neural net model. These values had to be scaled by dividing by the largest value in the list to ensure the neural network ran as smoothly as possible. This led to the results overall being normalized, and thus some scaling was required when producing final results.

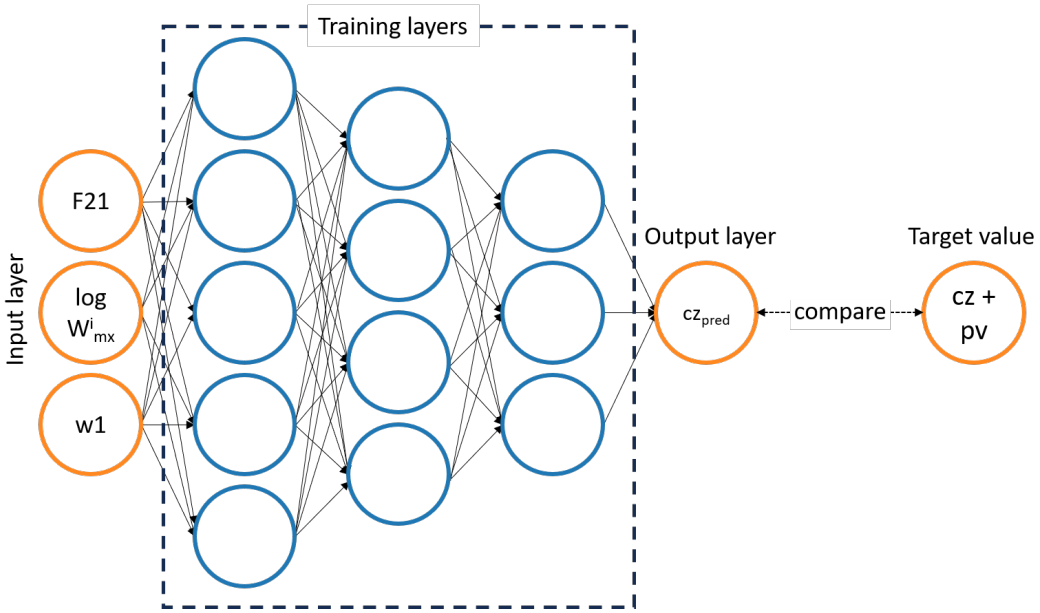

Figure 4.2: A visual represention of the neural network model that was used in the peculiar velocity analysis of the simulated data. The three input values provided in the simulated data result in an input layer of 3 nodes. The training layers had five nodes, four nodes, and three nodes respectively, which are flattened into a single output layer. This output predicts the cosmologicial redshift of a galaxy, and is compared to the redshift target value.

The inputs to this neural net remained the same as the distance modulus model, since the simulated data was still being used: linewidth related to rotation ($\log W^i_{mx}$), H1 line flux ($F21$), and fully-corrected WISE w1 band value. The output, however, was a prediction of the cosmological redshift, as the noise from the peculiar velocities was eliminated by the neural network. The shape of the neural network used in this method was consistent with that used in the distance modulus analysis, with an input layer of three values (the inputs available in the simulated data), and training layers of five nodes, four nodes, three nodes, and finally an output layer of one node. This model is visually represented in Figure 4.2. Neural nets were not averaged for this approach, since the goal was to provide proof of concept instead of completely accurate results.

Once cosmological redshift values had been predicted using the neural net, the accuracy of this calculation could be tested. The first step in this process was to create a histogram of the predicted redshift values and look at the shape of the resulting graph, in order to eyeball if the neural net is working as intended. The predicted redshifts can be plotted against the cosmological redshifts provided in the simulated data, to ensure there is a correlation and determine if the model is fitting the full range of values well.

The second part of this analysis works specifically with the peculiar velocities of the galaxies. By subtracting the predicted redshift values from the original redshift plus Gaussian peculiar velocity values, a peculiar velocity for each galaxy can be recovered. These two peculiar velocity values are then graphed against each other, in order to confirm there is a correlation and thus that the neural net is producing accurate values. The scaling factor used to ensure results from the neural net must be taken into account when constructing this graph.

4.2.1 Applying this Method to the Cosmicflows-4 Data

This method can also be used on the CF4 dataset. It requires a few tweaks, mainly in the way data is provided for the model to use, but also in the shape of the neural net. Since peculiar velocity information is included in the measured redshifts reported in the dataset, Gaussian velocities do not have to be added to redshift values. Instead, the target value used for the neural net was simply the scaled $v3k$ value from the CF4 data.

The values used as inputs for the simulated data (linewidth related to rotation ($\log W^i_{mx}$), H1 line flux ($F21$), and fully-corrected WISE w1 band value) were used originally, but there are far more values given in the CF4 data that can be used as inputs, so long as they do not contain redshift information. Several combinations of inputs were experimented with, including the values that make up the distance modulus calculation from the bTFR: rotation linewidth $\log W^i_{mx}$; H1 line flux $F21$; and the g, r, i, z, w1, and w2 photometric passband values.

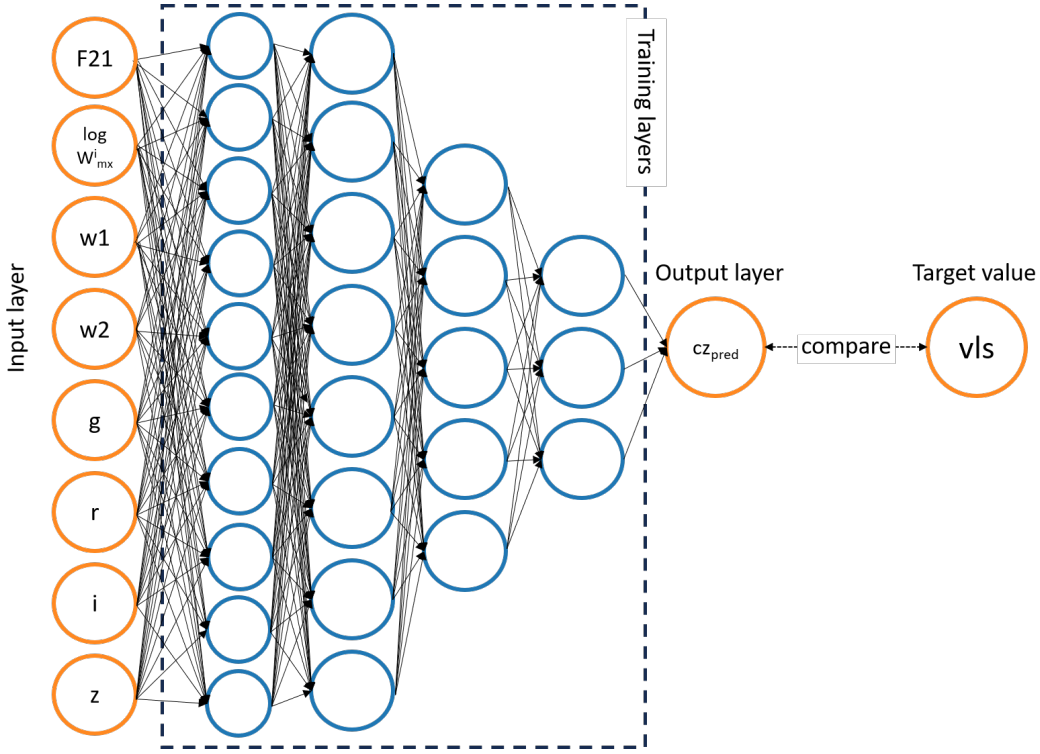

Figure 4.3: A visual represention of the neural network model that was used in the peculiar velocity analysis of the CF4 data. Eight input values that were found to produce good results means the model has an input layer of 8 nodes. The training layers had ten nodes, eight nodes, five nodes, and three nodes respectively, which are flattened into a single output layer. This output predicts the cosmologicial redshift of a galaxy, and is compared to the redshift target value.

This set of inputs proved the most effective, and is illustrated in Figure 4.3. The output of models remains the cosmological redshift, with the neural net having reduced the peculiar velocity "noise" to determine cosmological redshift.

The overall approach to models using CF4 data instead of simulated data also varied, mainly in how long they had to be run. Since values from CF4 data are less accurate overall than those using simulated data, the models had to be run for longer to ensure they converged. Averaging the results of three models was also returned to, in order to have as accurate a value as possible. Models in this case, since there were 8 input values, had an input layer of 8 nodes, and layers of 10, 8, 5, and 3 nodes before they were flattened to a single output value. This is illustrated in Figure 4.3. The first model in the average included all these layers, while in the second and third models the 8-node and 10-node layers were removed, respectively. As before, an average of three models was preferred to an average of one model run three times.

The redshifts of CF4 galaxies are well-known, since it is easy to measure redshift. The most straightforward way to determine if a model is producing values in the right ballpark is to plot its redshift results against the target CF4 redshift values and look for a correlation, as well as to plot the residuals. These two values should be close, but not entirely the same, if the peculiar velocity component is being correctly averaged out.

Once it has been determined the model is working as intended, the peculiar velocities of the CF4 galaxies can be calculated. By subtracting the calculated cosmological redshift values from the known redshift values, the peculiar velocity values for each galaxy can be recovered. By making a histogram of these calculated peculiar velocity values, it can be determined if they are vaguely Gaussian in shape, as well as their magnitude.

Finally, the peculiar velocity values calculated from the neural net can be compared to the peculiar velocity values calculated by the bTFR using CF4 data. Distance modulus is the quantity given in the dataset, but it can be converted into peculiar velocity via distance (d):

$$v = cz - H_0 d, \tag{4.2}$$

where v is the peculiar velocity of the galaxy, cz is the redshift value, and H_0 is assumed to be $75 km/s/Mpc$. The calculated peculiar velocity values from the neural net can be plotted against the peculiar velocities calculated using the bTFR. These values should be roughly correlated, since the neural net constructs a relation similar to the bTFR, but there is no way to know which calculation is more accurate.

5 Results

5.1 Distance Modulus Method

This project began with an investigation of neural networks used to calculate distance modulus values from the simulated data, as described in Section 4.1.

5.1.1 Simulated Data Noise Calculation

The first step in conducting an effective analysis of the simulated data was to calculate the included noise. It was known there was noise in the data, but not what value had been used.

To calculate the noise in the simulated data, distance modulus values were calculated using the non-distance values included in the data, by automating the process outlined in Section 2.1.4. Once distance modulus values had been calculated, they could be graphed against the distance modulus values provided in the simulated data and a standard deviation value found. This is illustrated in Figure 5.1. As can be seen in the figure, the standard deviation value calculated was 0.2523, which indicates the noise in the simulated data was about 0.25. This became a comparison value for distance modulus-based models using the simulated data.

5.1.2 Distance Modulus Models

Many neural networks were constructed, but the one that produced the most accurate results used the three inputs in the simulated data to predict the distance modulus of each galaxy, using as a training value the distance modulus calculated using the redshift. Predicted distance moduli were then compared to the distance moduli provided in the simulated data, and their standard deviation was compared to the noise value in the simulated data (for easy comparison, 0.25 was assumed) to

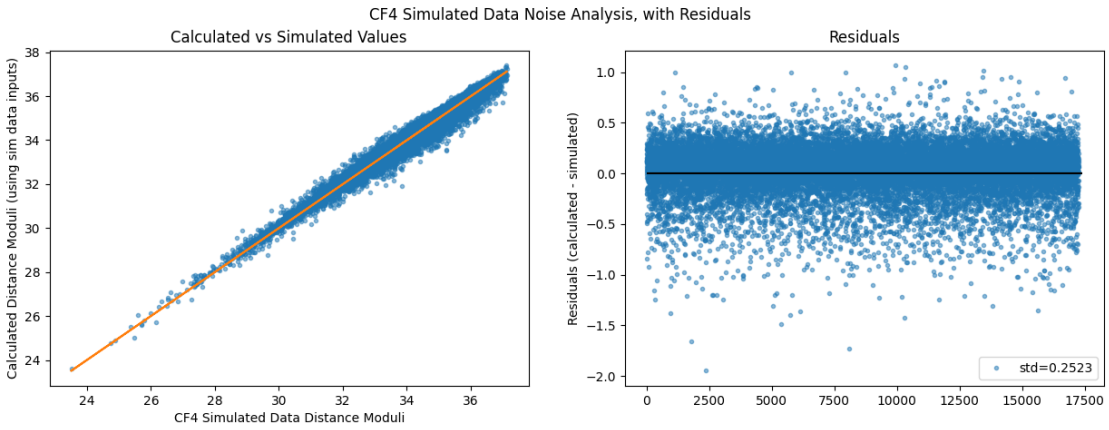

Figure 5.1: Calculated distance moduli using values from the simulated data are graphed against the distance moduli reported in the CF4 simulated data. Residuals are also shown, with a standard deviation of approximately 0.25.

determine the accuracy of the model. The model used for simulated data distance modulus predictions used the entire dataset as both training and test data.

Figure 5.2 illustrates the results of this method. The standard deviation with residuals limited to avoid outliers was calculated as 0.24, which is also comparable to the simulated data's noise value of 0.25. Since this approach allows for predictions of all galaxies in the dataset, rather than only a percentage of them as would be possible if a portion of the data was used as specifically training data and another portion for specifically testing data, training models on all of the data was the preferred approach in this project.

Residuals from the three models that made up the final calculation as shown in Figure 5.2 are provided in Figure 5.3.

5.2 Redshift and Peculiar Velocity Method

Once it was shown that neural networks could be used to predict distance moduli, the real work of this project could begin: finding a method that would predict distance without a known comparison value. This approach involved redshifts and peculiar velocities. By adding a set of Gaussian peculiar velocity values to the redshifts given in the simulated data, cosmological redshift values can be predicted using a neural network, as described in Section 4.2. Once these cosmological redshift values are calculated, the known redshift values from the simulated data can be subtracted, recovering the peculiar velocity values. The calculated peculiar velocity values are shown in Figure 5.4.

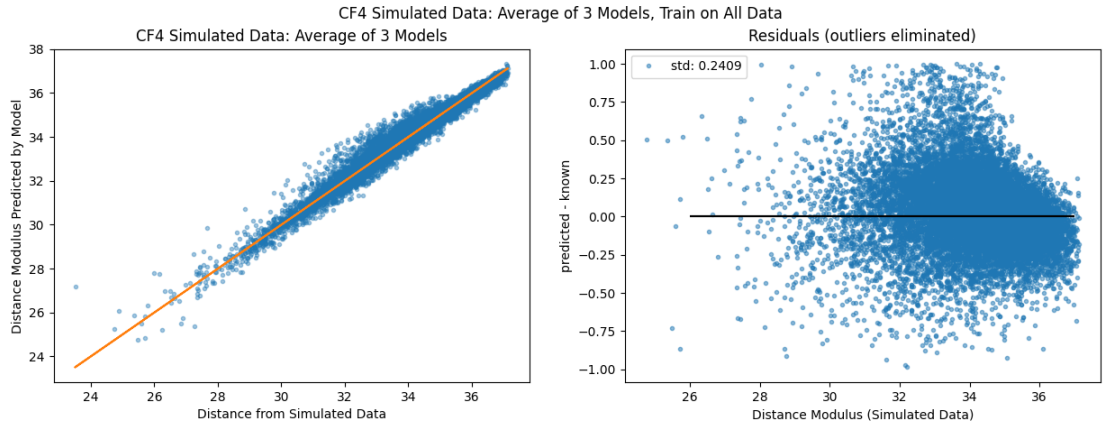

Figure 5.2: Distance moduli predicted by a neural net model using the whole dataset as both training and test data are plotted against the distance moduli reported in the CF4 simulated data. Residuals are also shown, plotted against the known distance moduli. The standard deviation, or uncertainty, in this model is 0.24.

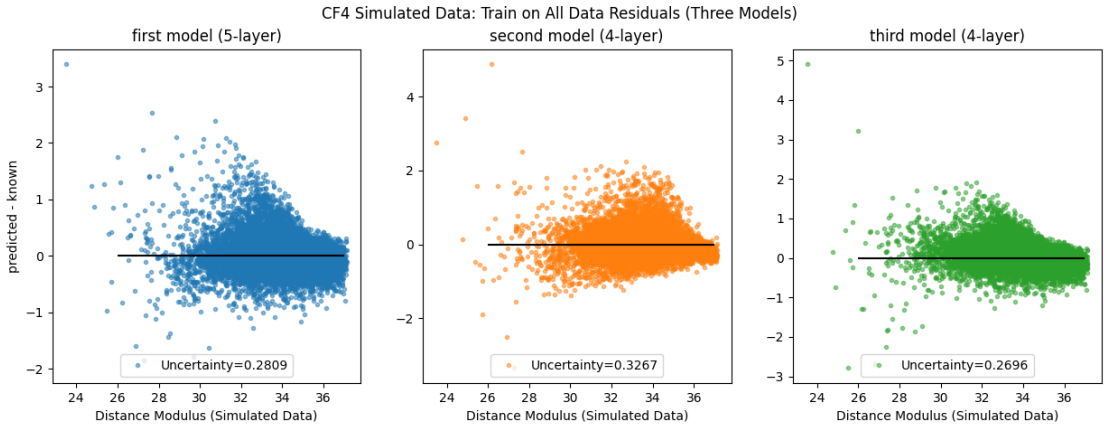

Figure 5.3: Residuals from the three models averaged to produce Figure 5.2 are plotted against the known distance moduli. Standard deviation values are shown.

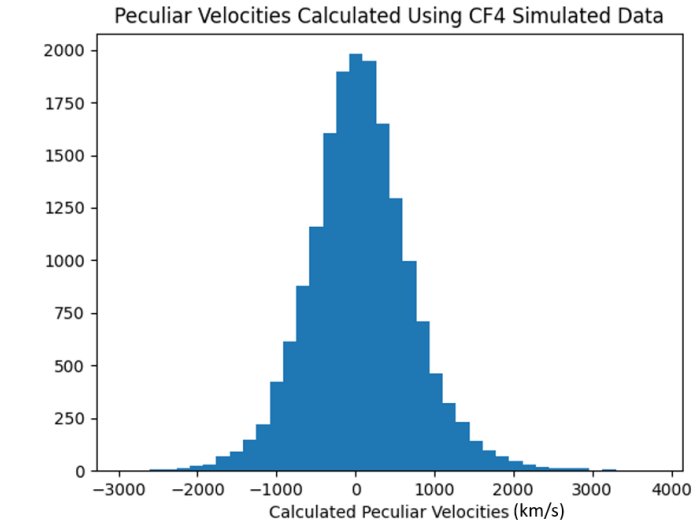

Figure 5.4: Peculiar velocity values calculated from the results of a neural net model are shown. They are Gaussian in shape.

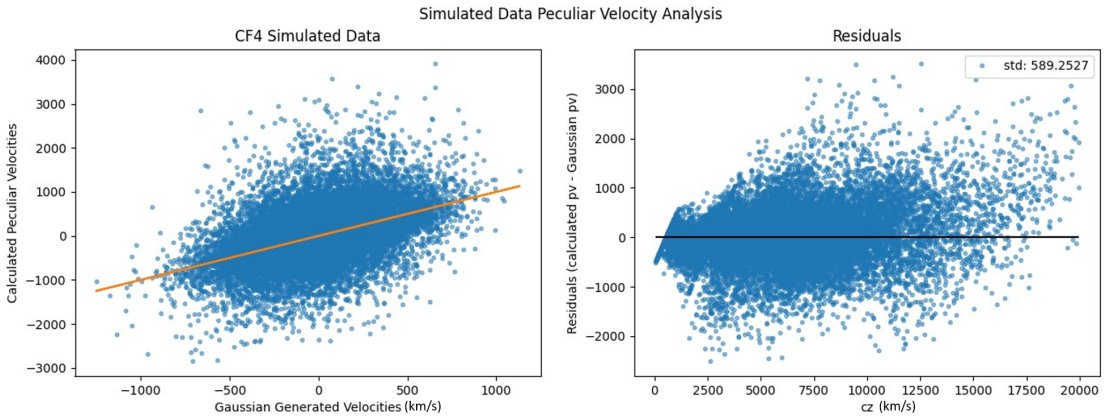

Figure 5.5: Peculiar velocity values calculated from the results of a neural net model are plotted against the Gaussian generated peculiar velocity values. The residuals are also shown, plotted against redshift values from the CF4 simulated data. The standard deviation, or uncertainty, of these values is $589.25 km/s$.

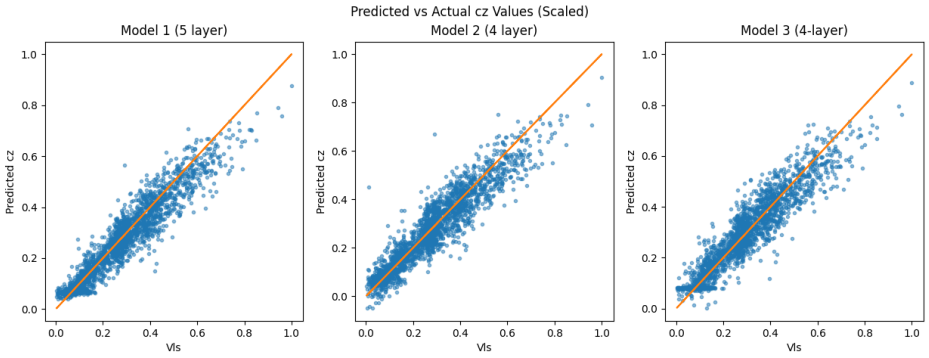

Figure 5.6: Redshift values predicted by three neural net models, plotted against the redshift values provided in CF4. Results are normalized against the largest CF4 redshift value for prediction accuracy.

Once calculated, peculiar velocity values can then be compared to the known Gaussian peculiar velocities, as shown in Figure 5.5. These values are correlated, which indicates this method is a viable way to calculate peculiar velocities. The neural net calculation has succeeded in recovering the Gaussian generated peculiar velocity values.

The residuals do not depend meaningfully on redshift, so standard deviation is indicative of uncertainty in this case. The standard deviation in this relation was calculated to be $589.25 km/s$.

5.2.1 Cosmicflows-4 Data

Once analysis of the simulated data was conducted and a viable distance method found, it could be applied to the Cosmicflows-4 dataset. Analysis of CF4 data was the final goal of this project. Results produced using CF4 data are actual distances to galaxies, as opposed to the proof of concept models produced using the simulated data.

The peculiar velocity analysis of the CF4 data produced similar results to the peculiar velocity analysis as applied to the simulated data, described in Section 5.2. However, since this analysis was the culmination of the project, more detail was put into the models: three models were produced with slight variations, all of which predicted redshift using inputs based on the bTFR calculation, and these results were averaged. The redshift results of these three models are shown in Figure 5.6.

Once redshift values have been predicted, peculiar velocity values for each galaxy can be calculated by subtracting the predicted redshift values from the

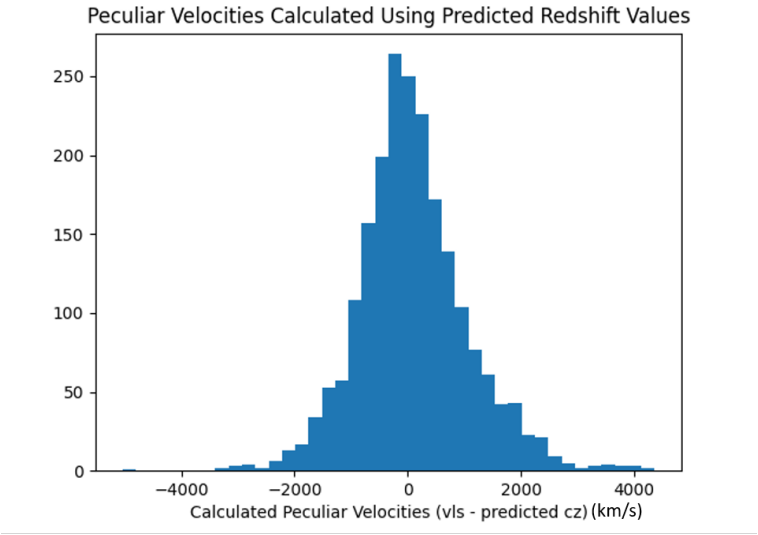

Figure 5.7: Histogram of peculiar velocities calculated using predicted redshift values. They are Gaussian in shape.

CF4 redshifts. These values were Gaussian, as was expected, which is shown in Figure 5.7.

The final analysis of the calculated peculiar velocity values is to compare them to bTFR peculiar velocities, as is discussed at the end of Section 4.2.1. The velocities calculated from predicted redshifts can be plotted against those calculated using bTFR distance moduli values, and a correlation shown. This plot is provided in Figure 5.8. Standard deviation is also a useful indicator of uncertainty in this case, and as such the uncertainty in this relation can be calculated as $568.44 km/s$, which is shown in Figure 5.9. This correlation indicates that this method of predicting peculiar velocities, and thus distance moduli, can be used as a distance indicator for spiral galaxies. The neural network constructs a similar relation to the bTFR, but there is no way to know which set of values is more accurate.

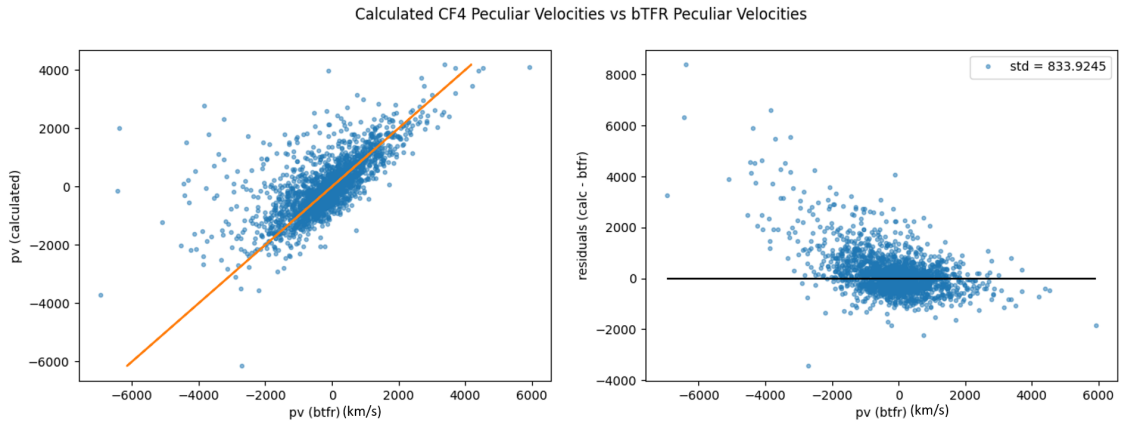

Figure 5.8: Peculiar velocities calculated using predicted redshift values plotted against peculiar velocities calculated from bTFR distance moduli. Residuals are also shown, as well as standard deviation, which can be treated as the uncertainty value. The uncertainty in this relation is $833.92 km/s$. The correlation of predicted and bTFR peculiar velocities proves neural nets can be used as a distance indicator for spiral galaxies.

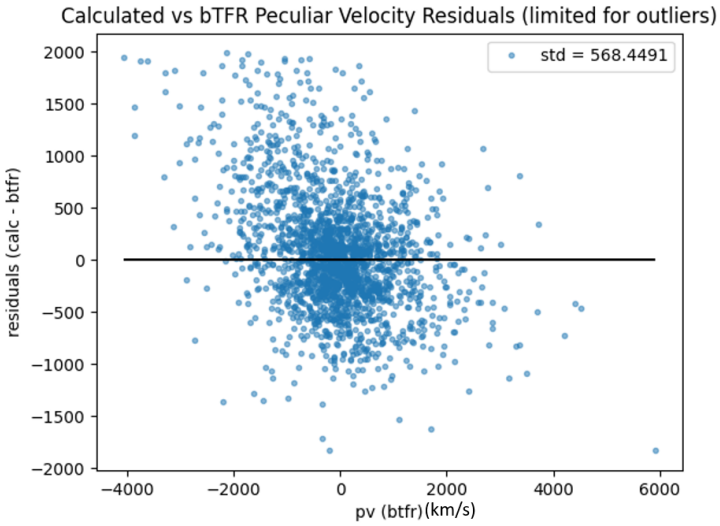

Figure 5.9: Residuals as plotted in Figure 5.8, limited to values between ± 2000. Of the 2109 values in the dataset, 2035 are included. The standard deviation is $568.44 km/s$.

6 Discussion

This project aimed to show that neural network models can be used to calculate peculiar velocties and galactic distances, which it did. The simulated data was used to construct models that acted as proof of concept, first for a purely distance modulus-based approach and more importantly for the peculiar velocity analysis. When this peculiar velocity analysis was applied to the CF4 data, it produced Gaussian peculiar velocity values that were correlated with the peculiar velocity values that can be calculated from bTFR data. This correlation suggests the neural net successfully filtered peculiar velocity values from the redshift, returning cosmological redshift that could be used to recover peculiar velocity.

The peculiar velocity analysis of CF4 data, and the correlation of calculated and bTFR peculiar velocity values specifically, as seen in Figure 5.8, shows that predicting distances using neural net models is possible.

The first step in this project was showing that neural nets could be used to predict distance modulus values, which is shown in Figure 5.2. The uncertainty in predicted values when they are compared to known distance modulus values is 0.2409, which is under the simulated data's noise value of 0.25. An uncertainty value below the noise indicates that the model is accurately predicting distance moduli values. The model for this approach uses the full dataset as both training and test data, which allows for predicted values across the whole dataset.

Once neural nets had shown they could be used to predict distance moduli accurately, a method had to be produced that allowed for distance prediction without known distance modulus values. This method took the form of redshift prediction, which could be analysed by looking at peculiar velocities. The proof of concept of this model using simulated data is shown in Figure 5.4: the Gaussian recovered peculiar velocities. This method is further proven in Figure 5.5, the correlation between calculated and known peculiar velocity values. This correlation has a standard deviation, or uncertainty, of $589.25 km/s$.

Finally, the peculiar velocity analysis could be applied to the CF4 dataset. These results are presented in Figure 5.8, the correlation between peculiar velocities calculated from predicted redshift values and peculiar velocities calcu-

lated from bTFR data. The uncertainty in this relation was calculated to be $568.45 km/s$, the calculation of which is shown in Figure 5.9. This correlation indicates that this method, of predicting redshift values and using them to calculate peculiar velocities, is a viable distance method. While the bTFR values are more commonly used, there is no way to tell which set of peculiar velocities is more accurate.

With a neural net-based approach shown to work for the CF4 data, using values based on the bTFR, it is possible to extend this idea to other similar distance methods. For example, the Fundamental Plane method used to calculate distances to elliptical galaxies contains a calculation that can be deconstructed to determine input values, like the bTFR calculation was. Fundamental Plane data is provided in the Sloane Digital Sky Survey (SDSS) dataset, making it an ideal next step to continue this project. Other similar distance calculations for other types of galaxies could also be investigated.

Bibliography

[1] L. R. e. a. E. Shaya, R. Brent Tully, "The Extragalactic Distance Database (EDD)." https://edd.ifa.hawaii.edu/dfirst.php?, 2023.

[2] E. Kourkchi, R. B. Tully, H. M. Courtois, A. Dupuy, and D. Guinet, "Cosmicflows-4: the baryonic Tully–Fisher relation providing 10 000 distances," *Monthly Notices of the Royal Astronomical Society*, vol. 511, no. 4, pp. 6160–6178, 2022.

[3] Wikipedia, "Apparent magnitude — Wikipedia, the free encyclopedia." http://en.wikipedia.org/w/index.php?title=Apparent%20magnitudeoldid=1222487485, 2024.

[4] Wikipedia, "Absolute magnitude — Wikipedia, the free encyclopedia." http://en.wikipedia.org/w/index.php?title=Absolute%20magnitudeoldid=1183688948, 2024.

[5] Wikipedia, "Stellar parallax — Wikipedia, the free encyclopedia." http://en.wikipedia.org/w/index.php?title=Stellar%20parallaxoldid=1217668122, 2024.

[6] W. L. Freedman, C. D. Wilson, and B. F. Madore, "New cepheid distances to nearby galaxies based on bvri ccd photometry. ii-the local group galaxy m33," *Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 372, May 10, 1991, p. 455-470.*, vol. 372, pp. 455–470, 1991.

[7] Wikipedia, "Cepheid variable — Wikipedia, the free encyclopedia." http://en.wikipedia.org/w/index.php?title=Cepheid%20variableoldid=1219934617, 2024.

[8] R. B. Tully and J. R. Fisher, "A new method of determining distances to galaxies," *Astronomy and Astrophysics*, vol. 54, pp. 661–673, 1977.

[9] S. S. McGaugh, J. M. Schombert, G. D. Bothun, and W. De Blok, "The baryonic Tully-Fisher Relation," *The Astrophysical Journal*, vol. 533, no. 2, p. L99, 2000.

[10] J. Schombert, S. McGaugh, and F. Lelli, "Using the Baryonic Tully–Fisher Relation to Measure Ho," *The Astronomical Journal*, vol. 160, no. 2, p. 71, 2020.

[11] V. Jain, "Everything you need to know about "Activation Functions" in Deep learning models." https://towardsdatascience.com/everything-you-need-to-know-about-activation-functions-in-deep-learning-models-84ba9f82c253, 2019.

[12] R. B. Tully, E. Kourkchi, H. M. Courtois, G. S. Anand, J. P. Blakeslee, D. Brout, T. de Jaeger, A. Dupuy, D. Guinet, C. Howlett, *et al.*, "Cosmicflows-4," *The Astrophysical Journal*, vol. 944, no. 1, p. 94, 2023.