Classifying Galaxies Using a Novel non-NN Deep Learning Model

R. Brnich, A. Ghadimi, M. Marsee
PH582 Group i

Abstract

The need for deep learning in the astronomical community has become increasingly apparent, as advances in data acquisition now require complex algorithms and enormous computing power. Data sets such as the Sloan Digital Sky Survey (SDSS), which alone has catalogued over ⅓ of the sky, are growing in size so quickly that automated tools are now necessary for mining through the overwhelming abundance of information (Sánchez et al. 2018). As classification is a shared goal of Machine Learning and Astrophysics, we need to look at the most recent advancements in the world of deep learning of the past two years. Our primary goal is to construct a novel algorithm that can aptly classify galaxies into three known types (spiral, elliptical, and irregular) based on their observed and derived photometric properties. Our self-training classifier utilizes a semi-supervised learning technique and is based upon the DeepForest model, proposed and built by Zhou and Feng (2019). and our data source is the ALFALFA-SDSS Galaxy Catalog, which contains basic optical and derived properties of ~ 30,000 galaxies. We found that our classifier performed with 83% accuracy and an F1-score of 78%. The model's performance supports the notion that it is possible to a) classify galaxies without having to rely on the standard methods of backpropagation and differentiation and b) using only observed/derived properties and no images. DeepForest has opened the doors to allowing new frameworks for classification in the astronomical community and beyond.

Introduction

DeepForest

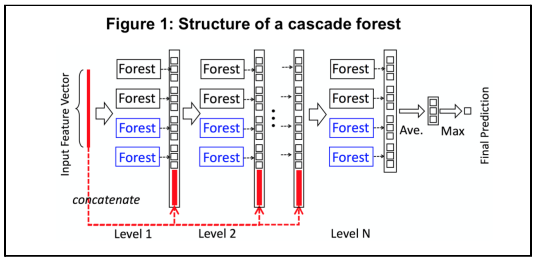


Figure 1: Structure of a cascade forest

In the search of what constituted modern machine learning architectures for galaxy classification, it had seemed that much of what we had already covered in lecture had been implemented widely in some form. Convolutional Neural Networks (CNNs) were fairly common, due to their clear usefulness in image analysis, in addition to the use of RandomForest, Boosting (e.g., ADA, XGB, etc.), Bagging, SVM, KNN and Stacking. Looking for more recently-developed classifiers in hopes of finding one not previously utilized, we found a new adaptation of Random Forest. This novel framework is referred to Deep Forest (DF) and was proposed by Zhou and Feng in 2017. It provides a

gcForest	**99.26%**
LeNet-5	99.05%
Deep Belief Net	98.75% [23]
SVM (rbf kernel)	98.60%
Random Forest	96.80%

Figure 2: gcForest's performance on the MNIST dataset

framework for deep learning that avoids the need for a differentiable module and back-propagation. DF is an ensemble of decision trees based on a cascading random forest structure and utilizes multi-grained scanning with a sliding window (gcForest). It allows for representation learning while executing layer-by-layer processing. The model complexity is data-dependent, meaning it works well even on small data sets. The general cascade structure is shown in Figure 1 (Zhou and Feng 2019), while the overall architecture of DF21, the latest implementation of DeepForest, is presented in Figure 3. In Figure 2, the model's performance on the MNIST dataset is shown alongside the performance of other commonly used classifiers.

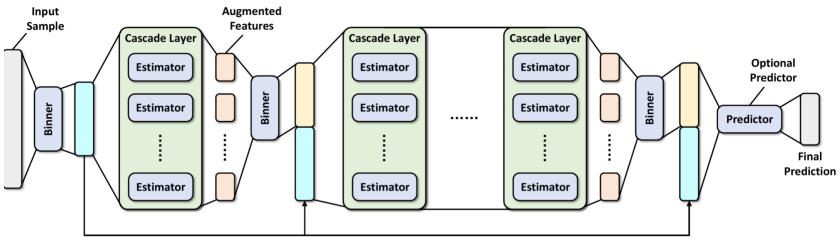


Figure 3: Architecture of DF21, the latest implementation of DeepForest

Source: deep-forest.readthedocs.io

Implementing this newly developed classifier will not only allow us to create a novel approach to galaxy classification, but will also act as a test of the classifier. Current deep model architecture is based upon a neural network (NN) structure, which uses backpropagation training on several layers of parameterized, differentiable, nonlinear models. Although powerful, NNs are notoriously difficult to train and require the architecture to be predetermined. Moreover, the backpropagation training technique uses a gradient descent algorithm that can very quickly diverge and interfere with the network's ability to learn patterns (Géron 2019).

We have therefore implemented a non-NN deep model, through an ensemble of decision trees, that is independent of backpropagation and is non-differentiable. The model follows a cascade-like structure, allowing its complexity to be automatically determined based on the data rather than manually specified before training. This cascade structure also guarantees the use of representation learning, for which its layer-by-layer processing is crucial for deep learning's success.

Ultimately, using a DeepForest architecture allows one to forgo the various issues found in the application of neural networks. While many in the past have used DNNs or other various traditional classifiers (or ensembles of them) to construct a system to classify galaxies, none have yet been able to implement a deep architecture while avoiding the drawbacks of neural networks. With the recent advent of DeepForest, one is able to almost have the best of both worlds. Given its improved performance on top of other classifiers on datasets such as MNIST, extending its application to galaxy classification looks promising.

Dataset

The ALFALFA-SDSS catalog used in our project does not include labeled instances, i.e., neither galaxy images nor features with labels of their corresponding shapes were given. It contains optical and derived properties of the galaxies as given in Tables 1 and 2 (Durbala et al. 2020; see Appendix A). Table 1 presents the ALFALFA Extragalactic Source Catalog and includes basic optical properties of galaxies that are cross-matched with SDSS. Table 2 contains the derived properties, calculated using various methods described in the paper.

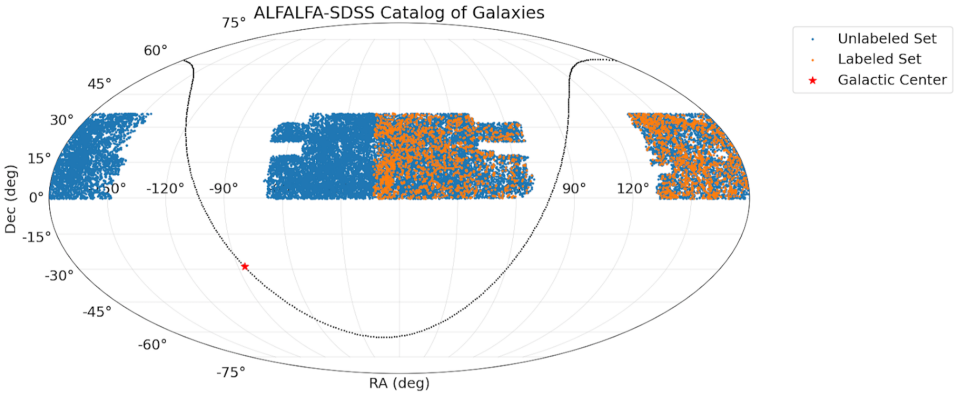


Figure 4: Distribution of Galaxies

We manually labeled ~10% (the first 3000 galaxies) of the catalog , using SIMBAD and VizieR, and used a semi-supervised model with the basic optical properties and the derived properties given in the catalog (Tables 1 and 2 of Appendix A). Figure 4 shows the distribution of the galaxies in Celestial coordinates. The catalog only consists of galaxies in the northern sky up to declination of ~40° due to the location of the Arecibo Observatory. The orange markers indicate the labeled set.

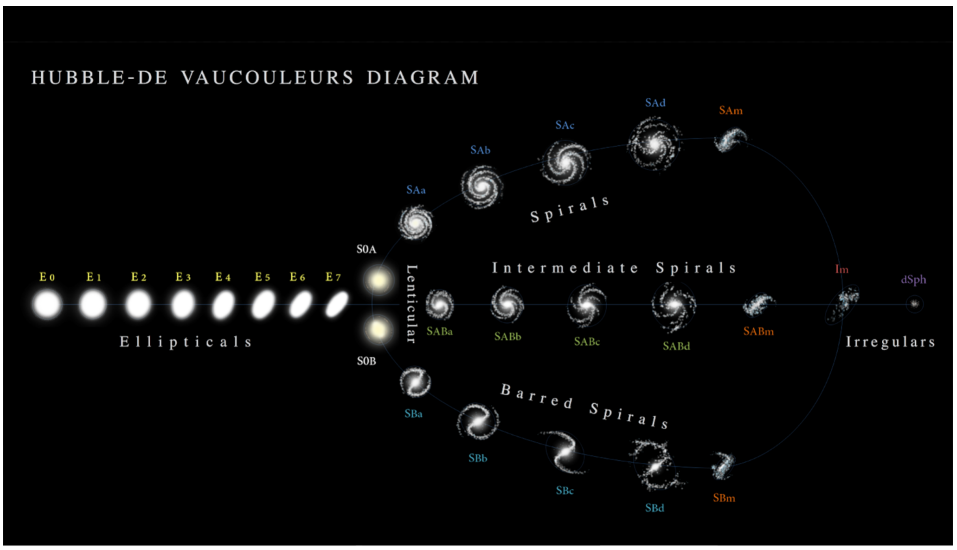


Figure 5: Hubble-De Vaucouleurs Diagram (Source: Wikipedia)

One of the most detailed systems of morphology classification is the Hubble - de Vaucouleurs method (Fig. 5). The galaxies were initially labeled using this system. After a preliminary training, it was apparent that this detailed system has a poor performance

when classifying galaxies without the use of images. The simplified labels are given in the histogram below (Fig. 6). Label S therefore includes both Spiral and Bar Spiral galaxies; IRR includes Sm/Sdm, SBm/Sdm, Im, S-IRR; LENT includes S0/S0a, SB0/SB0a; Dwarf contains dG dSph, dS, dSB, and, dI; and FAINT are those that did not have a label and could not be visually classified.

Each type has corresponding physical properties, which can also be used for further classification and categorization. The overall color (i.e., temperature) of a galaxy is one major indicator of its general type. Spiral galaxies generally have higher star-forming regions with younger and bluer stars, while elliptical galaxies tend to contain older, redder stars. The axial ratio is another property that can be used to classify galaxy types. It is defined as the ratio of the minor axis (b) to the major axis (a) of an ellipse, with eccentricity defined as e = 1 - (b/a).

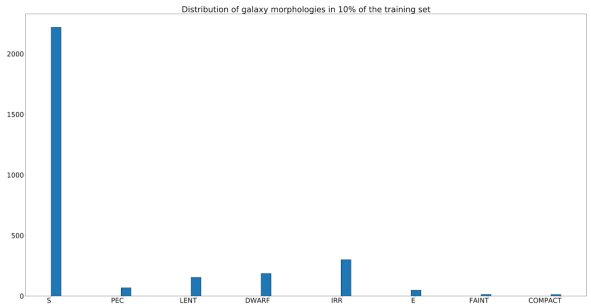


Figure 6: Distribution of Galaxy Morphologies

The Star Formation Rates (SFR) given in the catalog can help us determine the type of the galaxy in addition to the color, i.e. a galaxy with a higher SFR + bluer color is more likely to be a spiral galaxy than elliptical.

Metallicity is another feature used for galaxy classification, and is generally dependent upon stellar mass. The Mass-Metallicity Relation (Gao et al. 2018) indicates that the metallicity of a galaxy increases with its mass. One can also determine the color, shape, and metallicity of a galaxy by using the Stellar Mass in conjunction with SFR (Durbala et. al 2020).

As discussed previously, determining the distribution of high/low metallicity galaxies is a complicated task. It is postulated that high-redshift galaxies are metal-poor as not many stars had undergone stellar nucleosynthesis in the early Universe. However, metal-poor galaxies have also been found in the local Universe. Initially we planned to use the galaxy distance information from the catalog to find if there are any correlations between metallicity and distance. However, pur catalog consists of galaxies up to a redshift of 0.06, Therefore determining a correlation requires additional catalogs with high-redshift galaxies.

Methods

We first proposed that unsupervised learning would be beneficial to our goals because we are seeking to classify the galaxies and discover patterns using unlabeled photometric data. A clustering analysis approach would give us a way to partition the data and find feature relations that may have gone unnoticed due to the sheer size of

our data set. The DeepForest model is particularly advantageous for unsupervised learning, as clustering algorithms don't require labels: "The employment of completely-random tree forests not only helps enhance diversity, but also provides an opportunity to exploit unlabeled data" (Zhou and Feng 2019).

However, after trying several techniques and adding feature selection, it was nearly impossible to partition our dataset into any clear groups (see Fig. 7). We concluded that another learning method would be necessary to classify these galaxies. Because our data set is so robust, individually labeling every instance would be painstakingly time-consuming. However, labeling some portion of the set gives us more insight for our classifier's predictions than random, unlabeled guessing would. This gives us another avenue to explore for classification: semi-supervised learning. In this approach, we train our model on a subset of labeled data, i.e., 3,000 instances in the catalog. Then this classification model will be used to predict class labels on the rest of the data. The predicted labels with the highest probability of being correct are then referred to 'pseudo-labels,' and are combined with the original labeled training data to re-train the classifier. This will increase our usable dataset from 3000 back to over 30,000!

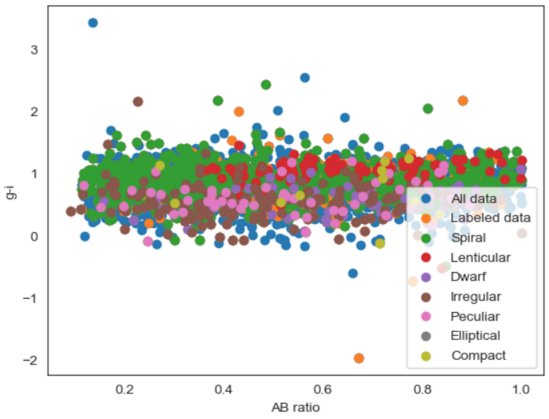
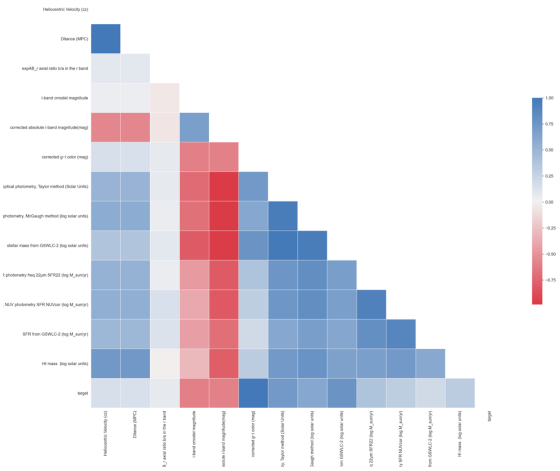


Figure 7: Correlation plot of color and axes ratio of the different morphological types. No clear evidence of clustering is observed.

To implement the self-supervised learning, we used sklearn's SelfTrainingClassifier, which requires that the unlabeled portion of the data be assigned a label of -1. This created an issue as the 3,000 manually labeled had string labels. To remedy this problem, we used label encoding to transform the string labels into numbers. This way the self-training algorithm can differentiate between the labeled data (positive values) and the unlabeled data (those marked as -1). We can then train the model on a combined training set of the self-trained labels and some of the hand-labeled data. Then, we can finally evaluate the model on a test set consisting of the remaining hand-labeled data.

Figure 8: Heat map of feature correlations from Table 1 (Durbala et al. 2020)



Results

Because we are using a fairly large data set that includes several properties of the galaxies, it is difficult for us to determine which features to extract by solely relying on visual plots. Thus, we created a visual heat map that shows correlation coefficients for every parameter in the data set. Using the data from Table 1 (Appendix A), we have identified the features that are the most and least closely correlated with each other and with the target variable, G-I Color Index (Fig. 8). We took into account the results from this heat map along with the importance of the physical parameters described above, and decided on 4 parameters to use for classification: axial ratio, G-I color index, stellar mass, and star formation rate.

The primary metric we used to evaluate the performance of our system was its accuracy score. Taking into account all four features, DeepForest obtained an accuracy score of 0.834 on a test set of 500 samples. This is a noticeable improvement on a simultaneous RandomForest model we ran, which only obtained an accuracy score of 0.784. These results add to the validity of the value of DeepForest as a new and improved ensemble classifier.

One unfortunate issue we had to deal with after the selection of our features was data loss. No series of experimental detections are perfect, and that being the case, there were a number of entries whose axial ratio, color, stellar mass or star formation rate was unrecorded ("nan" in the dataset) and thus that particular entry would be unusable. A preliminary model with only axial ratio, color and stellar mass only reduced our usable data from 31,501 to 28,267 entries (~10% reduction). Including star formation rate for a total of 4 features, our dataset plummets to 4,523 total usable entries (~86% reduction); however its performance increases. The accuracy score for the 3-parameter model was 0.804, whereas including the star formation rate yielded an accuracy score of 0.834. While it is nice to have large datasets for training, it is also important to consider the quality of the being used, and how the feature selection reflects that. As they say (while not as drastic in our case), "Garbage in - garbage out".

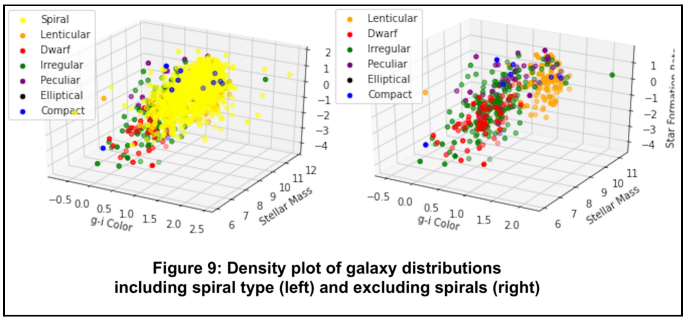


Figure 9: Density plot of galaxy distributions including spiral type (left) and excluding spirals (right)

```
print(classification_report(y_test,y_pred_DF))
```

	precision	recall	f1-score	support
COMPACT	0.00	0.00	0.00	1
DWARF	0.56	0.21	0.30	24
E	0.00	0.00	0.00	5
IRR	0.50	0.10	0.17	30
LENT	0.50	0.08	0.13	26
PEC	0.00	0.00	0.00	4
S	0.85	0.99	0.91	410
accuracy			0.83	500
macro avg	0.34	0.20	0.22	500
weighted avg	0.78	0.83	0.78	500

Figure 10: Performance evaluation

Being collected from a limited section of sky, the SDSS-ALFALFA data is dominated by spiral galaxies (Fig. 9). This gave us a limited number of alternative types of galaxies to train on. However

our model still was able to discern a level of difference between the different types, as opposed to just simply labeling everything as a spiral galaxy.

In addition to the accuracy scores, we printed a classification report (Fig. 10) for our model's performance on the test set. Based on the weighted average, we see a fair amount of success, although the model does suffer a bit with non-spiral types due to their low sample count.

Conclusion

As previously stated, the Based on its performance in semi-supervised learning alongside our self-training classifier, we see that DeepForest is an apt tool for classifying galaxy morphology based on both observed and derived photometric properties. Here we have demonstrated both a non-image based approach to galaxy classification, and a successful implementation of a new and upcoming classifier.

Future Work

In the future it would be interesting to incorporate a larger dataset covering more of the observable sky. This could allow us to have both more samples of and better performance on making distinctions between spiral and non-spiral galaxies. Another thing to explore would be the implementation of more observed and derived features, in hopes of finding even stronger links between these properties and specific galaxy morphology.

Python Notebook on Colab

https://drive.google.com/file/d/14v1ei1MdlWbsrAqDXW_Z6nuyfw7lkQx1/view?usp=sharing

References

Adriana Durbala, et al. 2020, AJ, 160, 271

Blanton M. R., Kazin E., Muna D., Weaver B. A., Price-Whelan A., 2011, AJ, 142, 31

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. Sebastopol (CA): OReilly Media.

“Hubble de Vaucouleurs Galaxy Morphology Diagram.” *Wikimedia Commons.*

“Model Architecture.” *Model Architecture - Deep Forest (DF21) Documentation, http://deep-forest.readthedocs.io/en/latest/advanced_topics/architecture.html.*

Muñoz-Mateos, J. C., Sheth, K., Gil de Paz, A., et al. 2013, ApJ, 771, 59

NASA-Sloan Atlas. Retrieved from http://www.nsatlas.org.

Querejeta M., et al., 2015, ApJS, 219, 5

Sánchez, H. D., Huertas-Company, M., Bernardi, M., Tuccillo, D., & Fischer, J. L. (2018). Improving galaxy morphologies for SDSS with Deep Learning. *Monthly Notices of the Royal Astronomical Society, 476*(3), 3661-3676.

Sheth, K., Regan, M., Hinz, J. L., et al. 2010, PASP, 122, 1397

Zhi-Hua Zhou and Ji Feng. Deep Forest: Towards an Alternative to Deep Neural Networks. In IJCAI, pages 3553–3559, 2017.

Zhi-Hua Zhou and Ji Feng. Deep Forest. National Science Review, 6(1): 74–86, 2019.

Appendix A: Tables 1 and 2 of the ALFALFA-SDSS Catalog (Durbala et al. 2020)

Table 1
Basic Optical Properties of Cross-listed Objects in the ALFALFA-SDSS Catalog

AGC	Flag	SDSS objID	R.A. J2000	Decl. J2000	$V_{\rm helio}$ km s^{-1}	D Mpc	σ_D Mpc	Ext$_g$ mag	Ext$_i$ mag	expAB_r	$\sigma_{\exp AB_r}$	cmodel$_i$ mag	$\sigma_{\rm cmodel_i}$ mag
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	1	1237679455462228052	0.656670	16.65222	5839	82.8	2.2	0.11	0.06	0.77	0.01	15.07	0.01
3	1	1237679502171701451	0.692920	18.88583	7883	107.6	2.3	0.11	0.06	0.47	0.01	13.08	0.01
4	1	1237678660887445652	0.737080	4.20889	8621	118.0	2.3	0.09	0.05	0.64	0.01	13.88	0.01
6	1	1237679476933722181	0.790420	21.95972	6561	88.8	2.2	0.15	0.08	0.73	0.01	13.52	0.01
7	1	1237656496724639763	0.796670	15.96500	11223	155.2	2.2	0.15	0.08	0.37	0.01	13.27	0.01
8	1	1237652944786292742	0.811670	16.14556	1050	13.2	1.3	0.15	0.08	0.79	0.01	10.46	0.01
10	1	1237669682261983347	0.835420	8.61861	11941	165.4	2.1	0.24	0.13	0.89	0.01	13.41	0.01
11	1	1237680297818390670	0.839580	22.10250	4445	62.8	2.4	0.17	0.09	0.72	0.01	14.80	0.01
12	1	1237663234988769589	0.835000	29.79722	6980	95.0	2.4	0.19	0.10	0.59	0.01	14.26	0.01
13	1	1237663234452947057	0.871670	27.35139	7749	105.9	2.2	0.18	0.09	0.78	0.01	13.52	0.01

(This table is available in its entirety in machine-readable form.)

Table 2
Derived Properties of Cross-listed Objects in the ALFALFA-SDSS Catalog

AGC	γ_g	γ_i	$M_{\rm icorr}$	$\sigma_{M_{\rm icorr}}$	$(g-i)_{\rm corr}$	$\sigma_{(g-i)_{\rm corr}}$	log $M_\star$ Taylor	$\sigma_{\log M_\star}$ Taylor	log $M_\star$ McGaugh	$\sigma_{\log M_\star}$ McGaugh	log $M_\star$ GSWLC	$\sigma_{\log M_\star}$ GSWLC	logSFR$_{22}$	$\sigma_{\rm logSFR_{22}}$	logSFR$_{\rm NUVcor}$	$\sigma_{\rm logSFR_{NUVcor}}$	logSFR GSWLC	$\sigma_{\rm logSFR}$ GSWLC	$M_{\rm H\,I}$	$\sigma_{M_{\rm H\,I}}$
	mag	mag	mag	mag	mag	mag	log($M_\odot$)		log($M_\odot$)		log($M_\odot$)		log($M_\odot$ yr^{-1})		log($M_\odot$ yr^{-1})		log($M_\odot$ yr^{-1})		log($M_\odot$)	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
1	0.73	0.39	−19.62	0.07	0.48	0.06	9.34	0.05	9.78	0.02	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	9.67	0.05
3	1.40	0.77	−22.39	0.11	1.06	0.14	10.85	0.11	10.89	0.02	⋯	⋯	−0.40	0.07	−0.22	0.03	⋯	⋯	10.28	0.05
4	1.25	0.68	−21.66	0.07	0.83	0.09	10.39	0.07	10.62	0.02	⋯	⋯	0.12	0.03	0.25	0.02	⋯	⋯	9.74	0.06
6	1.21	0.64	−21.38	0.07	0.77	0.07	10.24	0.05	11.23	0.02	⋯	⋯	1.50	0.02	1.40	0.03	⋯	⋯	9.29	0.07
7	1.59	0.86	−23.14	0.13	1.00	0.19	11.10	0.14	11.21	0.01	11.39	0.01	−0.06	0.05	−0.05	0.04	0.01	0.12	10.36	0.05
8	0.69	0.48	−20.26	0.22	1.64	0.05	10.41	0.09	9.32	0.09	⋯	⋯	−1.97	0.09	−1.56	0.09	⋯	⋯	8.96	0.10
10	1.70	0.87	−22.86	0.04	1.11	0.04	11.07	0.03	11.22	0.01	⋯	⋯	0.44	0.03	0.59	0.03	⋯	⋯	10.64	0.05
11	0.55	0.34	−19.33	0.10	0.65	0.07	9.33	0.06	9.58	0.03	⋯	⋯	−2.18	0.66	−0.90	0.04	⋯	⋯	9.25	0.06
12	1.01	0.56	−20.85	0.09	0.77	0.10	10.03	0.08	10.29	0.02	⋯	⋯	−0.15	0.03	⋯	⋯	⋯	⋯	9.79	0.05
13	1.27	0.70	−21.77	0.06	1.05	0.06	10.59	0.05	10.79	0.02	⋯	⋯	0.12	0.02	0.21	0.03	⋯	⋯	9.24	0.08

(This table is available in its entirety in machine-readable form.)