

Connecting Optical Morphology, Environment, and H I Mass Fraction for Low-Redshift Galaxies Using Deep Learning

JOHN F. WU^{1,2}

¹Department of Physics & Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA

²Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA

ABSTRACT

A galaxy’s morphological features encode details about its gas content, star formation history, and feedback processes, which play important roles in regulating its growth and evolution. We use deep convolutional neural networks (CNNs) to learn a galaxy’s optical morphological information in order to estimate its neutral atomic hydrogen (H I) content directly from SDSS *gri* image cutouts. We are able to accurately predict a galaxy’s logarithmic H I mass fraction, $\mathcal{M} \equiv \log(M_{\text{HI}}/M_*)$, by training a CNN on galaxies in the ALFALFA 40% sample. Using pattern recognition (PR), we remove galaxies with unreliable \mathcal{M} estimates. We test CNN predictions on the ALFALFA 100%, xGASS, and NIBLES catalogs, and find that the CNN consistently outperforms previous estimators. The H I-morphology connection learned by the CNN appears to be constant in low- to intermediate-density galaxy environments, but it breaks down in the highest-density environments. We also use a visualization algorithm, Gradient-weighted Class Activation Maps (Grad-CAM), to determine which morphological features are associated with low or high gas content. These results demonstrate that CNNs are powerful tools for understanding the connections between optical morphology and other properties, as well as for probing other variables, in a quantitative and interpretable manner.

Keywords: Galaxies, Galaxy evolution, Galaxy processes, Galaxy environments, Astronomy data analysis, Astronomy data visualization

1. INTRODUCTION

Neutral atomic hydrogen (H I) is the dominant component of cool gas in the interstellar medium (ISM) for low-redshift galaxies (e.g., Saintonge et al. 2017). Although neutral gas is crucial for understanding how galaxies evolve and grow over cosmic timescales, H I is difficult to detect in extragalactic sources because of its weak 21-cm emission line. This observational challenge has been partially mitigated by large H I surveys such as the H I Parkes All Sky Survey (HIPASS; Barnes et al. 2001), the Arecibo Legacy Fast ALFA Survey (ALFALFA; Giovanelli et al. 2005), and the GALEX Arecibo SDSS Survey (GASS; Catinella et al. 2010), which have taken a census of the brightest H I sources in the local Universe. New radio telescopes such as MeerKAT, ASKAP (Australian Square Kilometre Array Pathfinder), and eventually the SKA will allow us to measure H I at much lower masses (M_{HI}) and at higher

redshifts; see, e.g., Looking at the Distant Universe with the MeerKAT Array (LADUMA; Blyth et al. 2016), MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE; Jarvis et al. 2016), Wide-field ASKAP L-Band Legacy All-sky Blind surveyY (WALLABY; Koribalski et al. 2020), and Deep Investigation of Neutral Gas Origins (DINGO).¹

Small and incomplete H I samples currently limit our ability to study gas properties in typical galaxies beyond $z \approx 0.05$. Since H I is so important to galaxy evolution but challenging to measure, astronomers have devised proxies for estimating galaxies’ gas content. For example, Kannappan (2004) proposed “photometric” gas fractions, leveraging the valuable connection between global gas content and optical properties. Zhang et al. (2009) tightened the relationship by accounting for *i*-band surface brightness in addition to *g* – *r* color. More complicated heuristics and machine learning models have also been used (e.g., Teimoorinia et al. 2017; Rafiefer-

¹ <https://dingo-survey.org/>

antsoa et al. 2018), although these estimators become more difficult to interpret as the number of parameters increases. Indeed, computer vision algorithms seem to perform spectacularly well at predicting galaxy properties directly from optical imaging (e.g., Dieleman et al. 2015; Huertas-Company et al. 2019; Morningstar et al. 2019; Pasquet et al. 2019; Wu & Boada 2019), but because these models often have millions of parameters, it can be difficult to understand what makes them so successful.

We train a deep convolutional neural network (CNN) to directly predict the logarithm of the H I mass fraction, $\mathcal{M} \equiv \log(M_{\text{HI}}/M_*)$, using three-band optical image cutouts from the Sloan Digital Sky Survey (SDSS). After demonstrating that our trained model can predict \mathcal{M} to within 0.23 dex for $\alpha.40$, we test the CNN on independent data sets and examine which factors result in poor predictions. We use the same CNN method to distinguish ALFALFA detections from non-detections, and estimate a galaxy’s likelihood of detection in an ALFALFA-like survey from its *gri* imaging. Using this pattern recognition system, we evaluate only the robust predictions on test data sets again and compare to results in the literature. We investigate how the relationship between optical imaging and H I content varies with galaxy environment. We also use the Grad-CAM algorithm to localize image features that the CNN associates with high or low gas mass fraction in order to visually interpret which morphological features are relevant to machine learning predictions; it essentially tells us which parts of the image the CNN is looking at in order to determine the gas mass fraction (see, e.g., Peek & Burkhardt 2019).

The paper is structured as follows. We describe the H I catalogs and optical imaging in Section 2, and explain some details of the CNNs in Section 3. In Section 4, we present our results showing that a CNN trained on ALFALFA can accurately recover \mathcal{M} , and report test results on independent data sets. In Section 5, we use pattern recognition to identify galaxies that are expected to be detected by an ALFALFA-like survey, and in Section 6 we compare our results to other \mathcal{M} estimators in the literature. In Section 7, we quantify the impact of environmental effects and study how the connection between H I content and optical morphology breaks down in the most overdense environments. In Section 8, we discuss and interpret the morphological features that a CNN “sees” in order to distinguish gas-rich systems from gas-poor galaxies. We discuss future prospects in Section 9, and report our conclusions in Section 10. In the Appendix, we present comparisons between CNNs and simpler ma-

chine learning models and tests of CNN performance when artificial sources are added to the input images. Throughout this paper, we assume a cosmology with $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$. All of the code used in our analysis is publicly available at <https://github.com/jwuphysics/HI-convnets>.

2. DATA

We make use of four H I data sets in our analysis: $\alpha.40$, $\alpha.100$, NIBLES, and the xGASS representative sample. Because these data sets have different selection criteria, they are useful for testing our CNN methods on galaxy samples with varying H I mass fraction distributions. In Figure 1, we show the cumulative distribution functions of \mathcal{M} , the logarithmic H I mass fraction. The catalogs’ stellar and H I properties are also summarized in Table 1. We describe the data sets and our selection criteria in greater detail below.

SDSS imaging—The ALFALFA, NIBLES, and xGASS data sets have publically available catalogs of optical counterparts. We obtain *gri* imaging from the SDSS DR14 (Abolfathi et al. 2018) SkyServer using the Image Cutout service² queried via a custom Python script. The conversion of *gri* imaging to RGB channels is a modified version of the Lupton et al. (2004) algorithm, as described on the SkyServer website.³ Downloaded JPG images have 224×224 pixels at the native SDSS pixel scale ($0.396'' \text{ pixel}^{-1}$), which corresponds to angular sizes of $1.48'' \times 1.48''$.

Stellar masses—In order to compute gas mass fractions, H I detections are crossmatched with galaxies in the SDSS DR7 MPA-JHU value-added catalog (Kauffmann et al. 2003; Brinchmann et al. 2004; Tremonti et al. 2004; Salim et al. 2007). All stellar mass estimates assume a Chabrier (2003) initial mass function. ALFALFA is crossmatched on the basis of PhotoObjID ($\alpha.40$) or a $1''$ search radius ($\alpha.100$). The xGASS and NIBLES catalogs already include crossmatched stellar masses. We keep only the galaxies with valid stellar mass estimates.

2.1. ALFALFA $\alpha.40$

The Arecibo Legacy Fast ALFA (ALFALFA) survey is a $z \leq 0.06$ blind search for H I at high Galactic latitudes (Giovanelli et al. 2005). The ALFALFA $\alpha.40$ catalog covers 40% (2800 deg^2) of the full survey area (Haynes et al. 2011); most of these detections (12,468 sources) lie within the SDSS footprint. Our sample in-

² <http://skyserver.sdss.org/dr14/en/help/docs/api.aspx>

³ <https://www.sdss.org/dr14/imaging/jpg-images-on-skyserver/>

Table 1. H I data sets

Data set ^a	N	$\log(M_{\text{HI}}/M_{\odot})$			$\log(M_{\star}/M_{\odot})$			$\mathcal{M}_{\text{true}}$		
		0.16	0.50	0.84	0.16	0.50	0.84	0.16	0.50	0.84
$\alpha.40\text{A}$	7128	9.21	9.71	10.06	8.71	9.56	10.41	-0.52	0.15	0.68
$\alpha.40\text{B}$	4644	9.21	9.68	10.02	8.73	9.41	10.07	-0.26	0.25	0.70
$\alpha.100$	6087	9.22	9.67	10.03	8.62	9.38	10.24	-0.37	0.28	0.75
NIBLES	899	8.54	9.18	9.73	8.52	9.62	10.68	-1.19	-0.50	0.29
xGASS	1179	8.68	9.27	9.84	9.52	10.30	10.95	-1.78 ^b	-1.11	-0.27

NOTE— For each data set, we show the number of galaxies (N) after crossmatching to SDSS, performing all cuts, and removing sources in common with other catalogs. We report 0.16, 0.50 (median), and 0.84 quantile values for the H I mass, stellar mass, and gas mass fraction.

^aWe have removed overlaps between the $\alpha.40\text{A}$, $\alpha.100$, NIBLES, and xGASS samples. $\alpha.40\text{B}$ is a subset of $\alpha.40\text{A}$.

^bThe xGASS catalog includes upper limits on $\mathcal{M}_{\text{true}}$ (see text for details).

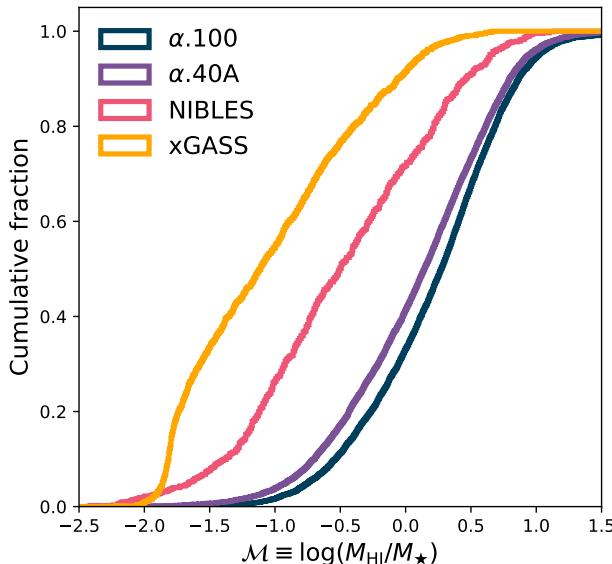


Figure 1. H I mass fraction cumulative distribution functions for the ALFALFA parent samples, xGASS representative sample, and NIBLES SDSS-crossmatched sample.

cludes rare, high-mass H I systems that are not necessarily representative of the probed cosmic volume. There also exists a nearly volume-limited ALFALFA subsample at $z \leq 0.05$, but we are interested in training our CNN with the larger data set. We select sources with `OCcode = 1` in order to retain $\alpha.40$ detections with SDSS counterparts, and we drop all sources with duplicate matches to DR7 identifiers. This cut reduces the number of H I sources to 11,739. We create the $\alpha.40\text{A}$ catalog, which contains 7,399 galaxies with valid stellar mass estimates, and the $\alpha.40\text{B}$ catalog, a subset of $\alpha.40\text{A}$ containing

4,797 galaxies with valid M_{\star} , SFR, gas metallicity, and spectroscopic redshift measurements.

2.2. ALFALFA $\alpha.100$

The $\alpha.100$ catalog contains 31,502 extragalactic sources across the full ALFALFA sky (Haynes et al. 2018). A large fraction of these sources do not overlap with the SDSS footprint, and a significant subset of $\alpha.100$ is already catalogued in $\alpha.40$. Since there is no public combined catalog for $\alpha.100$ with SDSS identifiers, we perform a custom crossmatching exercise. We crossmatch sources in the ALFALFA Spring sky to the SDSS MPA-JHU catalog using a $1''$ radius, exclude systems with other H I counterparts within a $1.9''$ radius (to avoid ALFALFA source blending), and exclude SDSS galaxies with neighboring optical sources within a $55''$ radius (to avoid fiber collisions). Since we will primarily be using $\alpha.100$ as an independent test data set, we also remove duplicates of $\alpha.40\text{A}$ and xGASS sources in $\alpha.100$ via another positional crossmatch (with search radius of $1''$). These selection criteria leave 6,087 galaxies in the $\alpha.100$ catalog. Among the H I catalogs, the $\alpha.100$ sample has the highest gas mass fractions (Figure 1).

2.3. NIBLES

The Nançay Interstellar Baryons Legacy Extragalactic Survey (NIBLES) catalog of H I detections contains 1,864 low-redshift galaxies with heterogeneous absolute z -band magnitudes (van Driel et al. 2016). The NIBLES sample is characterized by a wide range of stellar masses and intermediate values of \mathcal{M} relative to $\alpha.40$ and xGASS (Table 1). We remove systems with very low ($< 10^8 M_{\odot}$) or unavailable MPA-JHU stellar mass estimates. We visually inspect the SDSS image

cutouts and eliminate sources with no apparent optical counterpart near the center, those with only a point source in the center, and those significantly corrupted by imaging artifacts, leaving 941 galaxies. Finally, we also remove sources that overlap with ALFALFA and/or xGASS. There are 899 remaining galaxies in the cleaned NIBLES data set.

2.4. *xGASS representative sample*

ALFALFA detections tend to be the most H I-rich systems in the local Universe, and differ from the majority of galaxies found in optical surveys. We use the extended GALEX Arecibo SDSS Survey representative sample (xGASS; Catinella et al. 2018) in order to repeat our analysis on galaxies with more typical star formation and gas properties. xGASS consists of 1,179 galaxies with stellar masses between $9 \leq \log(M_*/M_\odot) \leq 11.5$ in the redshift range $0.01 \leq z \leq 0.05$. It is evident from Figure 1 that xGASS galaxies are the most gas-poor of our three samples, in part due to their relatively high stellar masses (Table 1). All xGASS systems have ancillary SDSS photometry and spectroscopy. The sample spans a range of galaxy morphologies, from passive ellipticals to starbursting mergers, and is complete down to $\mathcal{M} \approx -1.70$ for galaxies with $\log(M_*/M_\odot) \geq 9.7$. The most gas-poor members of the xGASS sample only have 5σ upper limits on M_{HI} available. However, we include them in our sample because the more massive systems have been observed to similar \mathcal{M} completeness, and the entire sample has a common gas mass limit $\log(M_{\text{HI}}/M_\odot) = 8$.

3. METHODOLOGY: DEEP NEURAL NETWORKS

We optimize deep CNNs in order to predict the H I mass fraction directly from three-band SDSS images, i.e., arrays of $3 \times 224 \times 224$ pixels. Because the goal is to estimate \mathcal{M} , a scalar quantity, we are optimizing the CNN to solve a regression task rather than a classification problem (although in later sections we will also use CNNs for classification). Training a neural network requires several steps, which can briefly described as follows. The CNN ingests a batch of images and outputs predictions ($\mathcal{M}_{\text{pred}}$) one batch at a time. These predictions are compared to their ground truth values (i.e., $\mathcal{M}_{\text{true}}$) via the *loss function*, which measures the level of discrepancy. CNN model parameters are then updated using an optimization algorithm that minimizes the loss. This process iterates until all samples in the training set have been used (signaling the end of an epoch), at which point the loss can be reported for the validation set, and the training loop repeated. The optimization details are very similar to the training routine described in Appendix A of Wu & Boada (2019).

We implement and optimize our deep convolutional neural network using `fastai` (Howard & Gugger 2020), which is built on PyTorch. All choices of CNN hyperparameters or tweaks have been empirically tuned in order to optimize training. Performance is primarily quantified by the root mean squared error (RMSE) metric, which also serves as our loss function:

$$\text{RMSE} \equiv \sqrt{\langle |\mathcal{M}_{\text{pred}} - \mathcal{M}_{\text{true}}|^2 \rangle}. \quad (1)$$

Another important metric of performance is the linear regression slope between $\mathcal{M}_{\text{true}}$ and $\mathcal{M}_{\text{pred}}$. A slope of unity indicates that there is no regression bias, and a shallower slope generally signifies that the CNN suffers from loss of predictive power (regression attenuation).⁴ Other works characterize the scatter using the standard deviation of the difference between predictions and truths, which is systematically lower than the RMSE if there is non-zero mean error (or offset).

We use the xresnet family of CNN architectures, which are enhanced versions of the original residual neural networks (He et al. 2015, 2018). Our 34-layer xresnets are further modified such that the usual Rectified Linear Unit (ReLU) activation functions are replaced with Mish (Misra 2019), and simple self-attention layers are added after convolutions in the residual blocks (Zhang et al. 2018). We train each model from scratch, as no pre-trained CNN with this architecture is available. In order to iteratively update the CNN’s weights, we use a combined Rectified Adam (Liu et al. 2019) and LookAhead (Zhang et al. 2019) optimizer. Weight decay with a coefficient of 0.01 is applied to all trainable layers except batch normalization layers (Goyal et al. 2017); note that we use true weight decay rather than the L2 norm (see Loshchilov & Hutter 2017 for details).

It is typical to evaluate deep learning models using a validation set drawn from the same distribution as the training set. If the model performs well on the training data but fails to perform well on the validation data, then it is a sign that the model suffers from overfitting. We randomly split the data by 80%/20% for training/validation sets, unless otherwise noted.

We train batches of 64 images at a time using a Nvidia P100 graphics processing unit (GPU). The learning rate is scheduled according to the “one-cycle” policy for 40 epochs (using the default hyperparameters set by `fastai`; Smith 2018), we set a maximum learning rate of 0.01. Dihedral group operations are randomly applied to images in order to augment the training set by a factor

⁴ A slope of zero and an RMSE equal to the inherent scatter can be achieved by always predicting the validation sample’s mean.

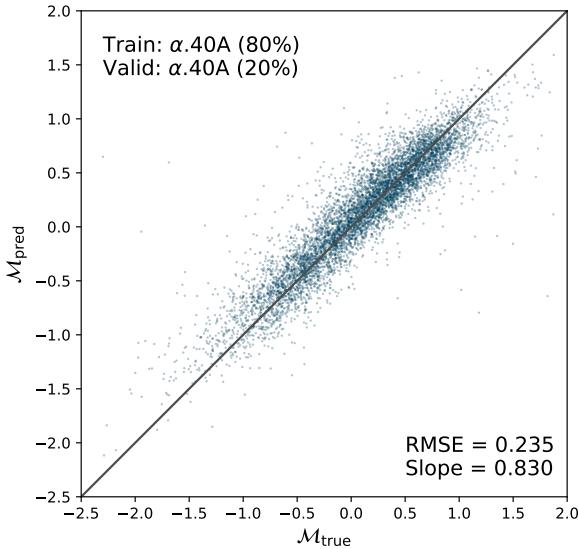


Figure 2. CNN-predicted logarithmic H I mass fraction ($\mathcal{M}_{\text{pred}}$) plotted against measured values ($\mathcal{M}_{\text{true}}$) for the $\alpha.40A$ validation subsample. The RMSE loss (in units of dex) and the linear regression slope are shown.

of eight and to force the CNN to learn such symmetries. The same transformations are applied during test-time augmentation to the validation data. We always report RMSE performance for the validation or test set.

We also consider simpler models for regressing \mathcal{M} based on the image data. These other regression models are tested in the Appendix A. Traditional statistical or classical machine learning methods are unable to represent morphological features using pixels as features, while CNNs are designed to encode image shapes, patterns, and textures over multiple scales in a translation- and rotation-invariant way. In Appendix B, we demonstrate that CNN predictions do not change significantly if the input images are injected with artificial point sources, which reveals that the CNN has learned a robust representation of galaxy morphology. Thus, we continue our analysis and discussion using the CNN results.

4. RESULTS

4.1. Training on $\alpha.40$

We train a deep CNN using 80% of the $\alpha.40A$ data set, and evaluate its performance using the remaining 20% validation data set. The optimized model can predict \mathcal{M} to within RMSE = 0.235 dex. In Figure 2, we show that the predicted and true values of H I mass fraction agree over two orders of magnitude in \mathcal{M} , and that the slope is close to unity. Our results demonstrate that there

exists a strong relationship between the H I content and the morphological information learned by a CNN from SDSS *gri* image cutouts.

We repeat the exercise using the smaller $\alpha.40B$ sample, and find very similar results (RMSE = 0.235 dex), even though $\alpha.40A$ is larger than $\alpha.40B$ by 54%. When we examine the standard deviation of \mathcal{M} for both subsamples, we find that $\alpha.40A$ has much larger scatter (0.68 dex) than $\alpha.40B$ (0.50 dex). The broader selection criteria for subsample A allows galaxies with poorer constraints on SFR or metallicity, which also means that galaxies with uncommon morphological or H I properties are included. Therefore, it is not surprising that the smaller $\alpha.40B$ subsample, which contains fewer outliers, produces similar results to the larger $\alpha.40A$ subsample.⁵

4.2. Dependence on galaxy properties

We find that a trained CNN can accurately recover \mathcal{M} from optical imaging for the $\alpha.40$ data set. No systematic biases are present, although incorrect predictions tend to be scattered toward the center of the \mathcal{M} distribution rather than toward the extrema. Generally, the CNN tends to overpredict gas mass fractions for low- \mathcal{M} galaxies.

We examine trends between $\Delta\mathcal{M} \equiv \mathcal{M}_{\text{pred}} - \mathcal{M}_{\text{true}}$ and other physical properties of galaxies. For example, it may be that the CNN tends to under- or overpredict \mathcal{M} based on some image feature that also correlates with other galaxy properties. However, we find that $\Delta\mathcal{M}$ does not vary systematically with any other property, and nor does the amount of scatter in $\Delta\mathcal{M}$. The only trend that we find is a negative correlation between $\Delta\mathcal{M}$ and $\mathcal{M}_{\text{true}}$, which is expected when the regression slope is less than unity. In Figure 3, we show trends between $\Delta\mathcal{M}$ and $\mathcal{M}_{\text{true}}$, redshift, stellar mass, SFR, and gas metallicity for the $\alpha.40B$ validation data set (959 H I sources). $\Delta\mathcal{M}$ also does not correlate with specific SFR, 4000 Å break strength, or the δ_5 environmental parameter (discussed in Section 7).

4.3. Testing on $\alpha.100$, NIBLES, and xGASS

We use the CNN trained on $\alpha.40A$ to estimate $\mathcal{M}_{\text{pred}}$ for galaxies in the $\alpha.100$, NIBLES, and xGASS test data sets. In the first set of entries in Table 2, we report our results for each data set. We list the RMSE, slope, σ , and mean offset in order to quantify performance, al-

⁵ Training a CNN on the $\alpha.40A$ sample leads to better validation performance than training on the full $\alpha.100$ sample (RMSE = 0.25 dex). This is likely because the $\alpha.40A$ catalog (Haynes et al. 2011) is more cleanly matched to SDSS than our custom cross-matching of $\alpha.100$ to SDSS sources (Haynes et al. 2018).

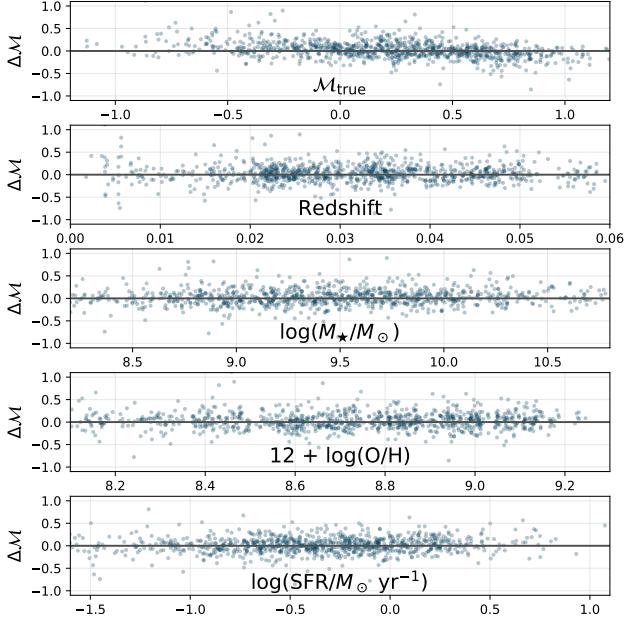


Figure 3. Comparisons between logarithmic H I mass fraction residuals ($\Delta\mathcal{M}$) and observed H I mass fraction, redshift, stellar mass, SFR, and gas metallicity. Only the $\alpha.40B$ validation data are shown. Aside from an anti-correlation in $\Delta\mathcal{M}$ vs M_{true} , which appears because the linear regression slope is shallower than unity, no correlation between residuals and other galaxy properties is observed.

though we primarily rely on the first two metrics. CNN predictions are also shown in Figure 4.

4.3.1. $\alpha.100$ results

By construction, our $\alpha.100$ test sample does not overlap with $\alpha.40$, so it can be used as an independent test for the trained CNN. In the top panel of Figure 4, we show a scatter plot of $\mathcal{M}_{\text{pred}}$ against $\mathcal{M}_{\text{true}}$ for $\alpha.100$. Overall, we find that the CNN recovers $\mathcal{M}_{\text{true}}$ accurately to within 0.30 dex. The performance is weaker than the $\alpha.40$ validation sample (RMSE = 0.235 dex), which implies that the CNN is unable to perfectly generalize to the $\alpha.100$ catalog. However, we note that our custom crossmatch of $\alpha.100$ to SDSS counterparts is different from the published $\alpha.40$ crossmatch to SDSS counterparts (Haynes et al. 2011), and that disparate selection effects may cause issues in generalization as well. One consequence of the imperfect crossmatching is that several sources apparently have implausibly high gas mass fractions. When 43 sources with $\mathcal{M}_{\text{true}} > 2$ are removed from the $\alpha.100$ sample, we find that the test performance is RMSE = 0.23 dex, which is in agreement with our $\alpha.40$ training sample. We will discuss the impacts of additional selection effects in Section 6.

4.3.2. NIBLES results

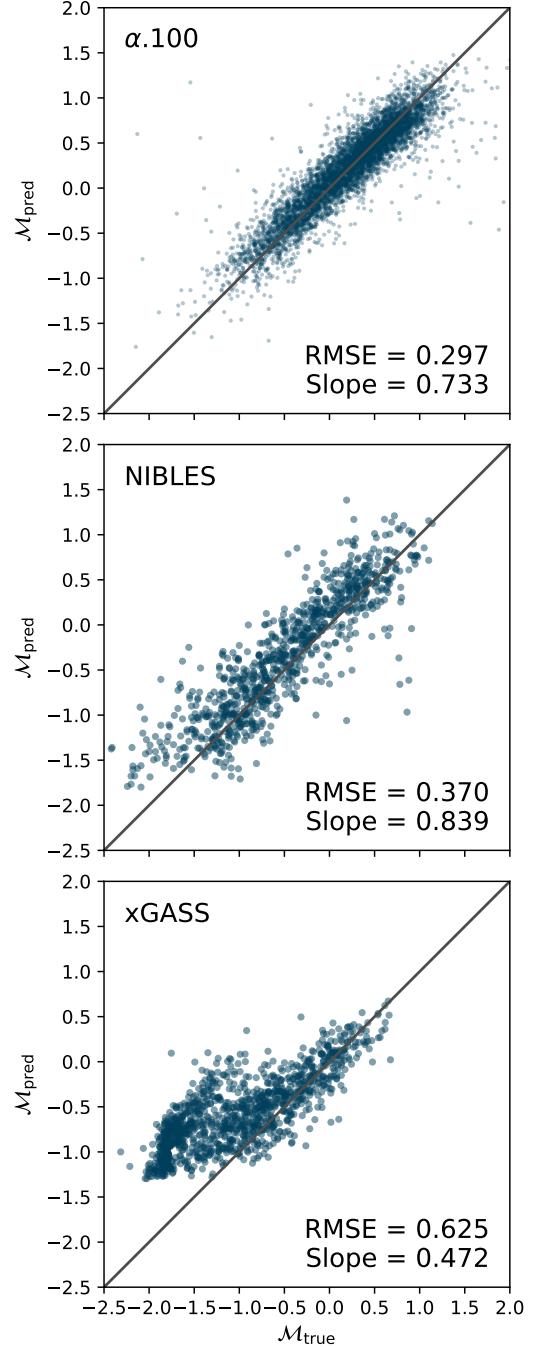


Figure 4. CNN-predicted logarithmic H I mass fraction ($\mathcal{M}_{\text{pred}}$) plotted against measured values ($\mathcal{M}_{\text{true}}$) for the $\alpha.100$, NIBLES, and xGASS test samples.

The NIBLES catalog presents another opportunity to test our CNN trained on $\alpha.40A$. While the two samples have comparable stellar mass distributions, NIBLES extends to significantly lower gas mass fractions than $\alpha.40$. We find that the CNN is generally able to recover $\mathcal{M}_{\text{true}}$ to within 0.37 dex (see center panel of

Figure 4). We note that $\mathcal{M}_{\text{pred}}$ estimates for NIBLES are systematically high by about 0.14 dex (Table 2). Our results are consistent with previous findings that ALFALFA H I fluxes are about 0.16 dex (45%) higher than NIBLES measurements for the overlapping sample (possibly due to a combination of flux calibration differences and multi-beam flux reconstruction; van Driel et al. 2016). If we rescale H I masses by the 0.16 dex systematic offset, then we find that the CNN’s predictions are in better agreement with NIBLES measurements (RMSE = 0.34 dex).

4.3.3. xGASS results

The xGASS representative galaxy sample contains a considerable number of massive elliptical galaxies with little gas content (see Table 1). For these gas-poor systems, the CNN tends to overpredict \mathcal{M} because it has been trained on $\alpha.40A$, which comprises mostly H I-rich, star-forming galaxies. We observe that the CNN does not output values below $\mathcal{M}_{\text{pred}} \approx -1.3$ for xGASS (although this does not appear to be a severe problem for NIBLES). As a result, our metrics indicate that the RMSE is large and the slope is too flat, as shown in the bottom panel of Figure 4. Therefore, the CNN is unable to generalize to *out-of-distribution* galaxies such as massive, gas-poor systems in xGASS.

5. PATTERN RECOGNITION FOR OUT-OF-DISTRIBUTION SAMPLES

5.1. Out-of-distribution samples

The trained CNN is capable of making accurate predictions where the training and test set distributions of $\mathcal{M}_{\text{true}}$ overlap. For example, over 95% of the $\alpha.40A$ training sample have $\mathcal{M}_{\text{true}} > -1$, and it is evident from Figure 4 that the CNN’s predictions for xGASS are more accurate for higher values of $\mathcal{M}_{\text{true}}$. However, for galaxies without measured H I masses, it is not known *a priori* whether a galaxy’s $\mathcal{M}_{\text{true}}$ is within the distribution of the training sample, and we must consider whether a galaxy is out-of-distribution in the input space (image data) rather than the target space ($\mathcal{M}_{\text{true}}$). A CNN (or any regression algorithm) is essentially a mapping between the input and output space, so it is feasible that the out-of-distribution $\mathcal{M}_{\text{true}}$ can be related to out-of-distribution galaxy images. For the rest of the paper, we will consider $\alpha.40A$ detections to be “in-distribution” and ALFALFA non-detections to be “out-of-distribution” for the trained CNN.

A neural network can be trained to recognize patterns in order to classify galaxies as in- or out-of-distribution (Hopfield 1987; Bishop 1995; Kinney et al. 1996). Pattern recognition (PR) algorithms have been used exten-

sively in astronomy (e.g., Zhu et al. 2014; Teimoorinia et al. 2017; Caldeira et al. 2019; Ćiprijanović et al. 2020), and they can supplement other machine learning estimators by classifying whether an input is representative of a training distribution. One such example is presented by Teimoorinia et al. (2017), where a shallow neural network is trained to distinguish ALFALFA H I detections from non-detections. Their method relies on 15 galaxy parameters derived from SDSS spectroscopy and photometry. However, one of the restrictions with this approach is that it necessitates spectroscopic observations and ancillary data in order to make predictions. Our method only requires an image of the galaxy. We proceed by training a CNN PR algorithm based on optical imaging in order to distinguish ALFALFA detections from non-detections.

5.2. Pattern recognition with a CNN

We obtain *gri* imaging for a sample of $z < 0.06$ SDSS galaxies that are located in the ALFALFA footprint but are undetected in the $\alpha.100$ catalog. To ensure a clean sample of ALFALFA non-detections, we impose the same selection cuts used for $\alpha.100$ (Section 2.2), and remove all galaxies with neighboring H I sources within an Arecibo beam radius. We randomly select 7,399 ALFALFA non-detections from this parent sample to enforce balanced classes with $\alpha.40A$ detections.

We train a CNN PR algorithm using exactly the same architecture as before, except that the final layer now predicts two outputs. After applying a sigmoid function, these two outputs represent the probabilities of detection (p_{CNN}) and non-detection ($1 - p_{\text{CNN}}$) in an ALFALFA-like survey. The $\alpha.40A$ sample serves as ground truths for the ALFALFA detection category, and the non-detection sample serves as ground truths for the non-detection category. We use the same optimization routine as before, except that cross-entropy loss is used as the optimization function since the objective now is binary classification. We also apply label smoothing with $\epsilon = 0.05$, so that optimized values of p_{CNN} gravitate toward 0.05 and 0.95 for non-detections and detections, respectively (see, e.g., Müller et al. 2019). We optimize the CNN using the same hyperparameters discussed in Section 3.

5.3. PR results

In the top and bottom panels of Figure 5, we show the distributions of p_{CNN} for the $\alpha.40A$ and ALFALFA non-detection samples, i.e., the training subsamples. Their p_{CNN} distributions are strongly peaked at ~ 0.95 and 0.05, respectively, indicating that the CNN robustly identifies $\alpha.40A$ detections solely using optical imaging. The trained PR is able to distinguish ALFALFA

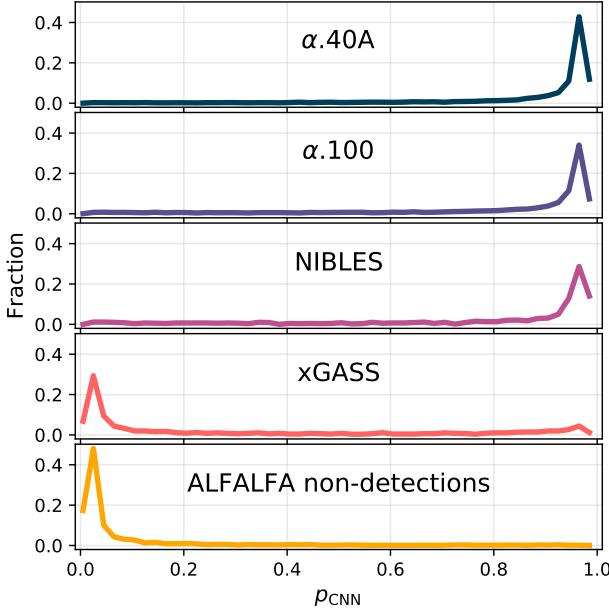


Figure 5. Pattern recognition probability (p_{CNN}) distributions for different galaxy samples. The $\alpha.40A$ and ALFALFA non-detections samples are used to train the CNN, while the $\alpha.100$, NIBLES, and xGASS are test galaxy samples.

detections from non-detections with 93% accuracy and AUC = 0.93 (area under the curve for the receiver operating characteristic).⁶

We examine the p_{CNN} distributions for the $\alpha.100$, NIBLES, and xGASS samples, which are also shown in Figure 5. As expected, the $\alpha.100$ galaxies are predominantly characterized by high values of p_{CNN} . Most galaxies in the NIBLES sample are also at $p_{\text{CNN}} \approx 0.95$ and would be detected by an ALFALFA-like survey. The xGASS representative galaxy sample, conversely, is characterized mostly by low values of p_{CNN} , although there is a small fraction with high p_{CNN} . Our results verify that the gas mass fraction estimates suffered from out-of-distribution error when evaluated on the xGASS sample, but did not encounter the same issue for the $\alpha.100$ and NIBLES samples.

5.4. Combining $\mathcal{M}_{\text{pred}}$ and p_{CNN}

We can use the CNN PR results to assess the reliability of $\mathcal{M}_{\text{pred}}$ on the test data. We consider various threshold values of p_{CNN} to separate detections from

⁶ The receiver operating characteristic (ROC) curve evaluates a model’s performance at all classification decision boundaries. Generally, the false positive rate is plotted against the true positive rate, such that the expected area under the curve (AUC) is 0.5 for a completely random model with balanced classes, and the AUC = 1 for a perfectly accurate and precise model.

non-detections. $p_{\text{CNN}} > 0.5$ represents a natural decision boundary, although in practice we may find that other values are better. For example, we have trained the CNN PR using an equal ratio of detected and undetected galaxies, but the ALFALFA survey only detects about 20% of typical galaxies in the low-redshift Universe (Catinella et al. 2010), which may imply that a higher value of p_{CNN} is desirable for rejecting false detections in typical massive galaxy samples.

In Table 2, we present CNN regression results for all H I data sets using several choices of PR decision boundary. For the ALFALFA data sets, we find that different cuts in p_{CNN} does not significantly impact regression performance (as characterized by, e.g., RMSE). This is expected because the ALFALFA samples should not contain out-of-distribution examples, and thus the PR cuts will not strongly affect the results.⁷ For NIBLES, performance modestly improves as the p_{CNN} threshold increases from zero (effectively the same as no cut) to 0.5, 0.8, and 0.9. We find that strict p_{CNN} cuts remove gas-poor (e.g., $\mathcal{M}_{\text{true}} < -1$) NIBLES systems for which $\mathcal{M}_{\text{pred}}$ is systematically overestimated. The most dramatic improvement is seen in xGASS, for which a PR cut of $p_{\text{CNN}} > 0.5$ removes nearly three quarters of the sample. Increasing the p_{CNN} threshold further refines the CNN performance; the RMSE, σ , and mean offset decrease, and the slope increases. For $p_{\text{CNN}} > 0.9$, the xGASS test set RMSE (0.24 dex) is comparable to that of the $\alpha.40A$ training set (0.23 dex).

A larger threshold for p_{CNN} enforces higher purity of ALFALFA-like galaxies. We recommend that $p_{\text{CNN}} > 0.9$ be used for typical massive galaxy samples similar to the xGASS representative sample. However, purity comes at the cost of completeness, and we note that different science goals may call for a different balance between complete versus clean samples. For optically selected samples, we find that a $p_{\text{CNN}} > 0.9$ threshold can robustly remove H I non-detections.

6. A COMPARISON OF \mathcal{M} ESTIMATORS

6.1. Colors and morphological parameters

Many works have studied the correlations between galaxy properties and their H I content. Kannappan (2004) finds that optical and near-infrared colors

⁷ There is a subtle effect with the $\alpha.100$ test set that causes the RMSE to slightly *increase* as we restrict p_{CNN} to higher values. Galaxies labeled with lower p_{CNN} generally have moderate H I properties (i.e., they are near the mode of the $\mathcal{M}_{\text{true}}$ distribution), whereas sources labeled with higher values of p_{CNN} might be extremely H I-rich (i.e., they have a broader distribution of $\mathcal{M}_{\text{true}}$). The end result is a < 0.01 dex increase as we shift the decision boundary from $p_{\text{CNN}} > 0.5$ to 0.9.

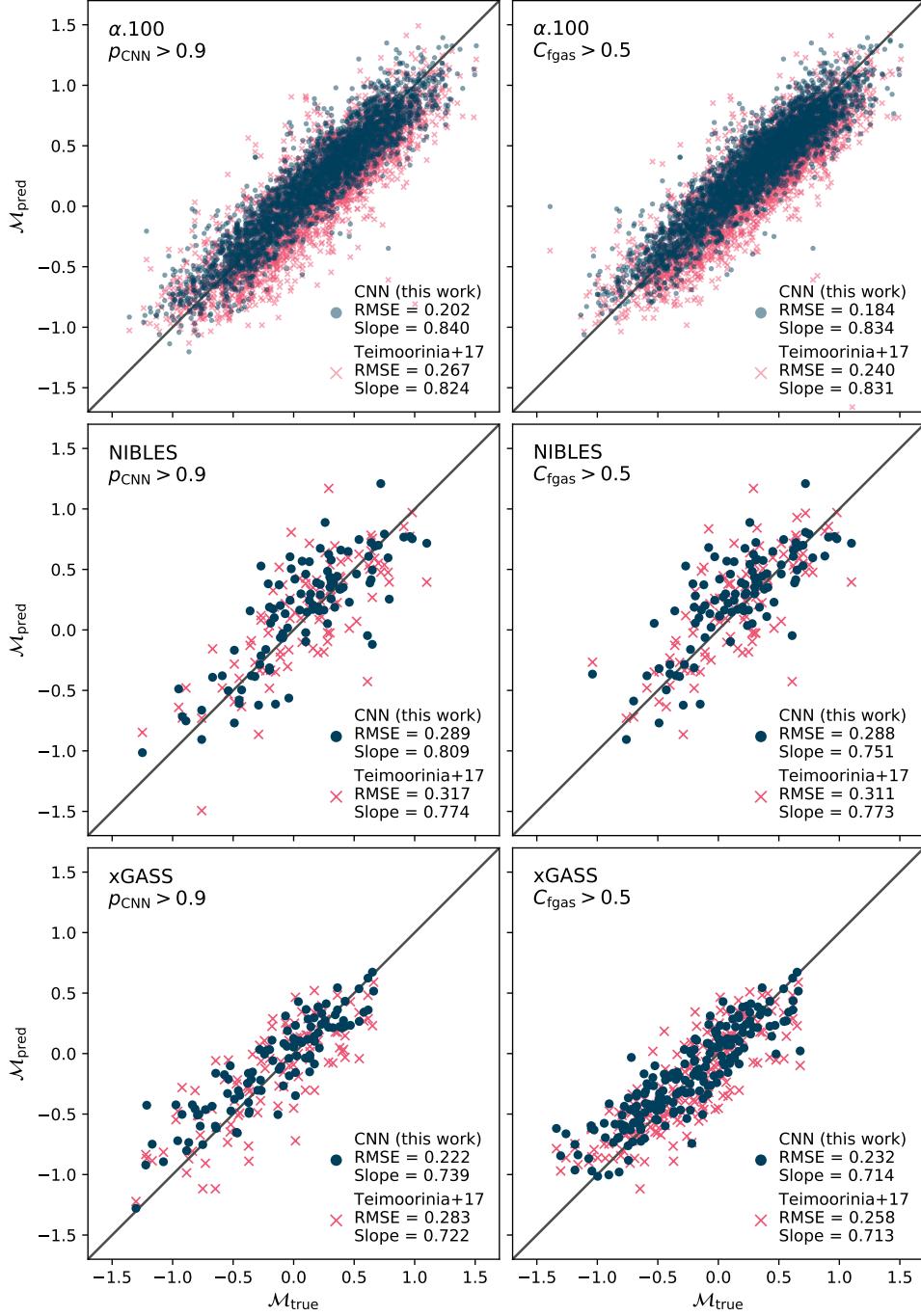


Figure 6. Comparison of M_{pred} versus M_{true} for our CNN trained on $\alpha.40A$ (blue circles) and the fully connected neural network presented (pink crosses; Teimoorinia et al. 2017). We show results for the $\alpha.100$ (top), NIBLES (middle), and xGASS (bottom) samples, using a selection of $p_{\text{CNN}} > 0.9$ (left) or $C_{\text{fgas}} > 0.5$ (right). Performance metrics are listed in each panel.

Table 2. Combined pattern recognition and $\mathcal{M}_{\text{pred}}$ results

PR cut	Data set	N	RMSE	Slope	σ	Offset
			(dex)		(dex)	
None	$\alpha.40A$	7128	0.2335	0.8427	0.2318	0.0285
	$\alpha.100$	6087	0.2975	0.7325	0.2974	-0.0076
	NIBLES	899	0.3705	0.8395	0.3416	0.1438
	xGASS	1179	0.6254	0.4724	0.4284	0.4558
$p_{\text{CNN}} > 0.5$	$\alpha.40A$	6674	0.2320	0.8429	0.2302	0.0291
	$\alpha.100$	5249	0.3043	0.7192	0.3042	-0.0086
	NIBLES	767	0.3566	0.8518	0.3238	0.1499
	xGASS	326	0.3237	0.6423	0.2962	0.1314
$p_{\text{CNN}} > 0.8$	$\alpha.40A$	6004	0.2334	0.8434	0.2316	0.0293
	$\alpha.100$	4454	0.3123	0.7084	0.3122	-0.0077
	NIBLES	663	0.3508	0.8586	0.3170	0.1506
	xGASS	217	0.2755	0.6882	0.2669	0.0706
$p_{\text{CNN}} > 0.9$	$\alpha.40A$	5319	0.2322	0.8468	0.2304	0.0290
	$\alpha.100$	3787	0.3107	0.7116	0.3107	-0.0085
	NIBLES	572	0.3435	0.8703	0.3116	0.1451
	xGASS	143	0.2449	0.7486	0.2365	0.0667

NOTE— The scatter can be quantified using the RMSE or standard deviation (σ) metrics, where lower is better. We also list the regression slope (closer to unity is better) and average offset (closer to zero is better). All data sets are independent from each other, such that N is sometimes smaller than the sample sizes listed in Table 1.

are connected to gas-to-stellar mass fraction, and reports a relationship between $u - K$ and \mathcal{M} with only $\sigma = 0.37$ dex scatter. [Zhang et al. \(2009\)](#) use i -band surface brightness, μ_i , in addition to $g - r$ color to predict \mathcal{M} . They find that the best-fit relation reduces the scatter to 0.31 dex. Other morphological parameters have been shown to tighten the relationship between galaxy colors and \mathcal{M} , such as stellar mass surface density and axial ratios, which reduce $\sigma \sim 0.30$ dex (e.g., [Catinella et al. 2010](#); [Huang et al. 2012](#); [Li et al. 2012](#); [Catinella et al. 2013](#); [Eckert et al. 2015](#)).

[Teimoorinia et al. \(2017\)](#) extend the [Zhang et al. \(2009\)](#) linear method by fitting a quadratic estimator to the inputs. They report a best fit of $\sigma = 0.22$ dex on the training data; however, this estimator is unable to generalize well to gas-poor samples such as GASS. Conversely, [Catinella et al. \(2010\)](#) predict \mathcal{M} for the GASS sample using UV-optical colors and stellar mass surface density, but their method begins to systematically fail in the gas-rich regime of very blue galaxies. The lack of generalizability suggests that more robust estimators are needed, and that pattern recognition should be used to exclude galaxies that yield poor predictions (e.g., [Catinella et al. 2013](#)).

6.2. Machine learning methods

Machine learning has recently become more popular for constructing powerful and flexible \mathcal{M} estimators. [Rafieferantsoa et al. \(2018\)](#) estimate gas mass frac-

tion for observed and simulated galaxy samples by using a variety of machine learning algorithms, such as random forests, gradient-boosted trees, and deep neural networks. Their algorithms can achieve $\text{RMSE} = 0.25 - 0.3$ dex when trained on real data (using photometric and environmental parameters as input features), but the estimators are unable to accurately predict \mathcal{M} when trained on simulated data (see also [Andrianomena et al. 2020](#)). We also train classical machine learning algorithms to model three-color *image* inputs, and demonstrate that they predict \mathcal{M} to $\text{RMSE} = 0.31$ dex for $\alpha.40A$ (Appendix A).

[Teimoorinia et al. \(2017\)](#) show that a fully connected neural network (FCNN) can be used to estimate \mathcal{M} to $\sigma = 0.22$ dex on independent test sets when trained using 15 galaxy properties. Although several of these input features are known to individually covary with \mathcal{M} , the non-linear combination of these properties processed by a shallow neural network is able to robustly outperform traditional regression methods (see [Ellison et al. 2016](#) for details on the network architecture). They find that $g - r$ color and μ_i are the most important parameters for regression (similar to previous results), and that a galaxy's bulge-to-total fraction also plays a role in governing \mathcal{M} . [Teimoorinia et al. \(2017\)](#) also present a pattern recognition method for rejecting galaxies that are not representative of the training sample and tend to be incorrectly predicted. The authors combine the PR probability and the epistemic uncertainty determined from an ensemble of neural network predictions, σ_{fitN} , to form C_{fgas} , which parameterizes the robustness of their \mathcal{M} prediction. [Teimoorinia et al. \(2017\)](#) have publically released \mathcal{M} and C_{fgas} estimates for a sample of over 500,000 SDSS galaxies.

6.3. Comparison to [Teimoorinia et al. \(2017\)](#)

The FCNN by [Teimoorinia et al. \(2017\)](#) predicts \mathcal{M} with remarkably low scatter. Because their method outperforms previous photometric gas fraction techniques (e.g., [Kannappan 2004](#); [Zhang et al. 2009](#)), their results serve as a valuable baseline for comparison to our work. In order to make a fair comparison with their results, we first crossmatch our data to their catalogs using a $1''$ radius. This process removes a non-trivial fraction of the test set before any PR cuts are made; after cross-matching, 67% of $\alpha.40A$, 87% of $\alpha.100$, 18% of NIBLES, and 86% of xGASS remain.

In Figure 7, we show $\mathcal{M}_{\text{pred}}$ versus $\mathcal{M}_{\text{true}}$ scatter plots comparing our predictions with [Teimoorinia et al. \(2017\)](#) results. Overall, we find that the CNN outperforms the FCNN according to terms of scatter and slope. Using a conservative threshold of $p_{\text{CNN}} > 0.9$, we observe that

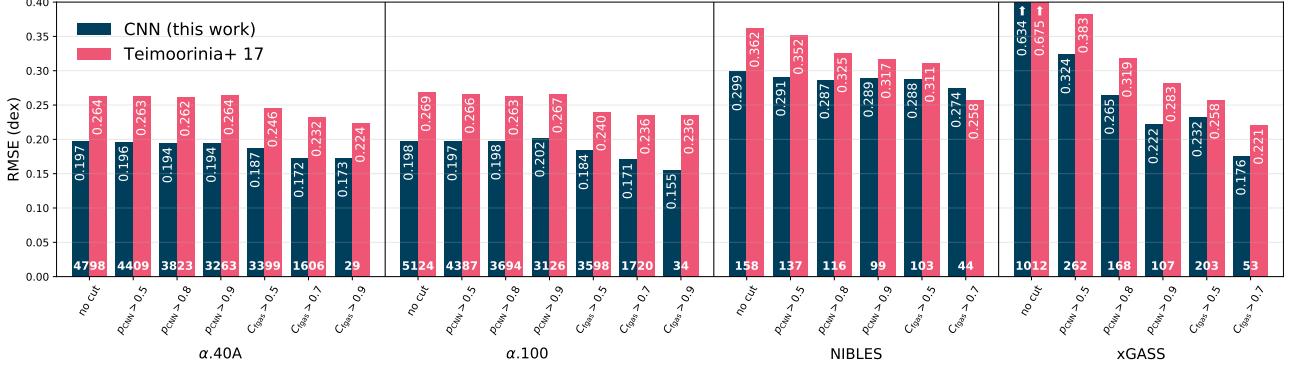


Figure 7. Comparison of results using our CNN method (blue) and a fully connected neural network (red; Teimoorinia et al. 2017) for ALFALFA, xGASS, and NIBLES galaxies crossmatched with the Teimoorinia et al. (2017) catalog. We show the RMSE for test subsamples selected using various choices of pattern recognition (C_{fgas} and p_{CNN}) or no cut at all; in every case the CNN recovers \mathcal{M} with lower scatter. The number of galaxies in each test set is shown in at the bottom of each bar plot. Detailed comparisons with additional metrics are provided in Table 3.

the CNN predicts \mathcal{M} to RMSE = 0.20 dex for $\alpha.100$, compared to 0.27 dex for the FCNN. If we instead apply $C_{\text{fgas}} > 0.5$ to both the CNN and Teimoorinia et al. (2017) results, we find RMSE = 0.18 and 0.24 dex, respectively, for $\alpha.100$. Both methods give comparable results for the crossmatched NIBLES catalog (RMSE = 0.29 dex for the CNN, and 0.31 dex for the FCNN). For xGASS, we find that the CNN predictions are robust using either $p_{\text{CNN}} > 0.9$ (RMSE = 0.22 dex) or $C_{\text{fgas}} > 0.5$ (RMSE = 0.23 dex), whereas the FCNN performs better using the latter (0.26 dex).

In Table 3, we provide detailed comparisons of our results with those published by Teimoorinia et al. (2017). We test different p_{CNN} and C_{fgas} thresholds in addition to the $C_{\text{fgas}} > 0.5$ threshold recommended by Teimoorinia et al. (2017). For all data sets, the CNN performs extremely well under the p_{CNN} and C_{fgas} selection criteria. However, it is difficult to compare the CNN and FCNN results for C_{fgas} thresholds higher than 0.5. For example, a $C_{\text{fgas}} > 0.9$ cut removes all NIBLES and xGASS galaxies and only leaves a few galaxies in alpha.40 (N = 29) and alpha.100 (34). Even with a more moderate cut ($C_{\text{fgas}} > 0.7$), the sample size is small for NIBLES (44) and xGASS (53). Therefore, our conclusions are based on the $C_{\text{fgas}} = 0.5$ decision boundary.

We primarily quantify our results using the RMSE scatter. However, previous works often report σ when discussing scatter (Zhang et al. 2009; Teimoorinia et al. 2017). When there is no mean offset between $\mathcal{M}_{\text{pred}}$ and $\mathcal{M}_{\text{true}}$, σ is comparable to the RMSE. However, the RMSE will more heavily penalize predictions when there is an offset, such as the 0.14 dex offset in H I mass between NIBLES and ALFALFA, or the > 0.20 dex systematic overpredictions for H I-poor galaxies in xGASS (prior to applying PR cuts). We recommend using σ

for comparison only when the mean offset is small compared to the scatter.⁸ Nevertheless, we find that the CNN consistently outperforms the FCNN according to both RMSE and σ .

It is intriguing that the CNN results are significantly improved after crossmatching with the Teimoorinia et al. (2017) catalog (i.e., comparing RMSE in Tables 2 and 3). Although we previously find RMSE = 0.30 dex for the $\alpha.100$ test set, we now report RMSE = 0.20 dex, in part because the sources with unrealistic $\mathcal{M}_{\text{true}}$ values identified in Section 4.3.1 have been removed. Substantial improvements are also evident for the other data sets. This discrepancy is most likely due to selection criteria introduced by Teimoorinia et al. (2017), which can remove galaxies with uncertain physical properties. A similar selection effect accounts for why $\alpha.40B$ has smaller intrinsic scatter in \mathcal{M} than $\alpha.40A$. Teimoorinia et al. (2017) restrict their analysis to galaxies with magnitudes in all SDSS bands, redshifts, sizes, morphological measurements, and environmental parameters. Galaxies with all 15 properties are characterized by higher signal-to-noise observations and are more likely to have secure H I and stellar mass measurements. The omitted galaxies tend to be redder and therefore more likely to contribute error and scatter. We conclude that the combination of selection criteria introduced during

⁸ It is worth noting that Teimoorinia et al. (2017) systematically predict lower \mathcal{M} than we do, which causes negative offsets in their ALFALFA predictions relative to $\mathcal{M}_{\text{true}}$. Equivalently, our NIBLES and xGASS predictions have positive offsets relative to $\mathcal{M}_{\text{true}}$. The systematic offset between our CNN and their FCNN predictions is likely due to their strict exclusion of galaxies with neighboring H I sources, which leads to a 0.14 dex difference in $\mathcal{M}_{\text{true}}$ between their clean and contaminated samples.

Table 3. Comparison of CNN and FCNN Results

PR cut	Data set	N	CNN (this work)				FCNN (Teimoorinia et al. 2017)			
			RMSE (dex)	Slope	σ (dex)	Offset	RMSE (dex)	Slope	σ (dex)	Offset
None	$\alpha.40A$	4798	0.1967	0.8524	0.1953	0.0238	0.2637	0.8317	0.2556	-0.0647
	$\alpha.100$	5124	0.1975	0.8462	0.1971	0.0125	0.2688	0.8243	0.2574	-0.0775
	NIBLES	158	0.2988	0.7890	0.2791	0.1090	0.3622	0.7194	0.3474	0.1060
	xGASS	1012	0.6338	0.4584	0.4258	0.4697	0.6751	0.4098	0.4793	0.4758
$p_{\text{CNN}} > 0.5$	$\alpha.40A$	4409	0.1958	0.8531	0.1942	0.0247	0.2626	0.8297	0.2548	-0.0636
	$\alpha.100$	4387	0.1968	0.8473	0.1963	0.0138	0.2660	0.8267	0.2550	-0.0758
	NIBLES	137	0.2909	0.7898	0.2739	0.1009	0.3518	0.7208	0.3435	0.0815
	xGASS	262	0.3235	0.6221	0.2953	0.1333	0.3829	0.5796	0.3649	0.1182
$p_{\text{CNN}} > 0.8$	$\alpha.40A$	3823	0.1938	0.8574	0.1922	0.0250	0.2617	0.8343	0.2530	-0.0672
	$\alpha.100$	3694	0.1980	0.8440	0.1974	0.0160	0.2634	0.8263	0.2521	-0.0765
	NIBLES	116	0.2869	0.7981	0.2760	0.0826	0.3252	0.7361	0.3226	0.0509
	xGASS	168	0.2645	0.6602	0.2568	0.0665	0.3185	0.6566	0.3182	0.0283
$p_{\text{CNN}} > 0.9$	$\alpha.40A$	3263	0.1944	0.8589	0.1927	0.0261	0.2638	0.8340	0.2552	-0.0670
	$\alpha.100$	3126	0.2021	0.8405	0.2015	0.0160	0.2665	0.8239	0.2548	-0.0783
	NIBLES	99	0.2895	0.8090	0.2822	0.0706	0.3167	0.7737	0.3174	0.0234
	xGASS	107	0.2216	0.7392	0.2098	0.0742	0.2826	0.7222	0.2835	0.0145
$C_{\text{fgas}} > 0.5$	$\alpha.40A$	3399	0.1869	0.8306	0.1863	0.0157	0.2460	0.8179	0.2347	-0.0738
	$\alpha.100$	3598	0.1838	0.8344	0.1837	0.0063	0.2404	0.8311	0.2243	-0.0867
	NIBLES	103	0.2876	0.7507	0.2692	0.1044	0.3113	0.7734	0.3091	0.0476
	xGASS	203	0.2317	0.7138	0.2189	0.0777	0.2577	0.7127	0.2581	0.0112
$C_{\text{fgas}} > 0.7$	$\alpha.40A$	1606	0.1725	0.8142	0.1725	0.0039	0.2321	0.7968	0.2083	-0.1025
	$\alpha.100$	1720	0.1706	0.8238	0.1707	-0.0009	0.2356	0.8318	0.2074	-0.1119
	NIBLES	44	0.2744	0.5946	0.2532	0.1124	0.2577	0.7070	0.2605	-0.0109
	xGASS	53	0.1758	0.7999	0.1589	0.1124	0.2208	0.7494	0.2134	-0.0640
$C_{\text{fgas}} > 0.9$	$\alpha.40A$	29	0.1730	0.6614	0.1735	-0.0290	0.2236	0.6599	0.1666	-0.1523
	$\alpha.100$	34	0.1555	0.7549	0.1524	-0.0405	0.2358	0.7454	0.1446	-0.1879

NOTE— We compare $\mathcal{M}_{\text{pred}}$ from this work and from the fully connected neural network trained by Teimoorinia et al. (2017) using all data sets and several choices of pattern recognition cuts. In order to facilitate an equal comparison, only galaxies with matches in the Teimoorinia et al. (2017) catalog are evaluated. Bolded values indicate superior performance for a given combination of PR cut, data set, and metric.

the comparison with Teimoorinia et al. (2017) explains the major improvement in our CNN predictions.

7. THE IMPACT OF ENVIRONMENT

The morphology and H I properties of a galaxy are strongly sensitive to its surrounding environment (e.g., Serra et al. 2012; Jones et al. 2016). If a CNN can accurately learn a connection between galaxy optical imaging and \mathcal{M} when trained only on systems in clustered environments but fails to accurately estimate H I content for a test sample of isolated galaxies (or vice versa), then it may be a sign that the distribution of galaxy morphologies has shifted. In essence, we wish to probe whether the H I-morphology relation learned by the CNN covaries with environment.

7.1. Galaxy overdensity

In order to quantitatively investigate the impacts of environment, we parameterize the environment using the projected galaxy density (e.g., Cooper et al. 2008):

$$\Sigma_5 = \frac{3}{\pi D_5^2}, \quad (2)$$

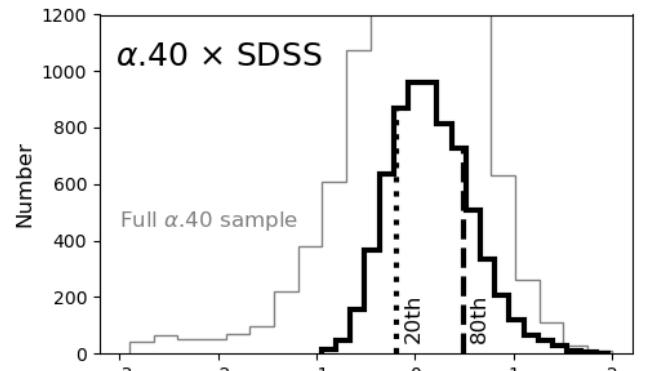


Figure 8. Thick black lines show histogram distributions of normalized galaxy overdensity for the $\alpha.40$ sample. Dotted and dashed vertical lines in both panels represent the 20th and 80th percentile values for δ_5 , respectively. We also show the parent $\alpha.40$ sample in light gray (prior to optical crossmatching with the SDSS catalog).

where D_5 is the projected physical distance to each galaxy's fifth-nearest neighbor (including its own optical counterpart). We use neighboring galaxies in the NASA-

Sloan Atlas (version 1.0.1; Blanton et al. 2011) within a velocity window of $\pm 1000 \text{ km s}^{-1}$ in order to compute Σ_5 for each H I source. We enforce a $D > 10 \text{ Mpc}$ distance cut in order to prevent contamination or biases from the Local Group. Following Cooper et al. (2008), we divide Σ_5 by its median over a sliding redshift boxcar window of size $\Delta z = 0.02$, which removes redshift dependence. The final result is a normalized overdensity parameter, $1 + \delta_5$. In Figure 8, we show the distribution of $\log(1 + \delta_5)$ for our $\alpha.40$ sample crossmatched with spectroscopically confirmed SDSS optical counterparts (we also show the full $\alpha.40$ sample in gray). It is apparent that the optical-H I crossmatching exercise removes $\alpha.40$ systems in the lowest-density environments. In Figure 9, we provide examples of SDSS image cutouts for ALFALFA galaxies at various environmental densities.

We select 80% of the higher- δ_5 galaxies for training and set aside the remaining 20% (with lower δ_5) for validation. In other words, we test whether a CNN trained on galaxies in higher-density environments can accurately predict the H I content of galaxies in lower-density environments. If this turns out to be the case, all else unchanged, then the environment does not significantly impact the connection between H I richness and optical imaging learned by the CNN. We also split the sample such that the 80% with lower δ_5 is used for training, and the 20% with higher δ_5 is used for validation. As a baseline comparison, we test the case where the training and validation set are randomly split (but trained in the same manner otherwise).⁹ We repeat tests five times for each training/validation split, and report the RMSE average and standard deviation in Table 4.

Our initial tests suggest that the galaxy H I-morphology connection *varies significantly* with environment. We find that a CNN trained only on galaxies in overdense environments and validated on systems in underdense environments performs better than the inverse. Surprisingly, the CNN validated on lower- δ_5 systems even outperforms the randomized baseline. However, this effect is fully explained by the validation scatter in \mathcal{M} when we select subsamples by a range in δ_5 . When we compare the CNN performance normalized by the inherent scatter of the validation subsample (the rightmost column in Table 4), it becomes apparent that the CNN validated on higher- δ_5 environments still performs

significantly worse than those validated on random or lower- δ_5 environments. We conclude that the CNN is able to generalize predictions in a way that yields good performance in underdense environments when exposed to galaxies in more overdense environments (relative to randomized validation subsamples), yet the opposite is not true.

These results can be interpreted as evidence that the H I-morphology connection is controlled by different physical mechanisms in the highest- δ_5 environments. Galaxies residing in clusters are subject to ram pressure stripping, tidal forces, galaxy-galaxy interactions, and other effects that can leave morphological imprints and also depress their gas content (e.g., Chung et al. 2009; Fabello et al. 2012). By training on subsamples of galaxies characterized by relatively lower-density surroundings, a CNN is unlikely to learn about the morphological cues associated with extreme physics of clustered environments, and therefore our experiment is able to distinguish between “typical” and “overdense” modes of the H I-morphology connection. It is also worth noting that these tests may not even capture the full extent of the environmental effects, as the $3.8'$ Arecibo beam may cause overestimation of H I mass or misidentification of optical counterparts in groups and clusters (e.g., Serra et al. 2015; Stevens et al. 2019). Such errors may artificially boost H I content and thereby ameliorate the CNN’s performance in high-density environments.

7.2. The overdensity transition regime

We observe a stark difference in CNN performance across different density regimes, but it is unlikely that there is a sharp transition in environmental effects. Gas mass fraction is known to depend on a satellite galaxy’s distance toward the center of its group or cluster host (e.g., Brown et al. 2017). “Pre-processing” in only moderately overdense environments can also depress galaxies’ gas masses (Odekon et al. 2016). We probe the gradual onset of environmental effects by repeating the analysis in Section 7.1 and reserving 20% of the galaxies for validation based on their δ_5 . We show the normalized RMSE as a function of the validation set δ_5 in Figure 10. For example, one of the validation data sets in $\alpha.40A$ comprises galaxies with δ_5 values in the $0.1 - 0.3$ quantile range, and the training set would consist of the remainder of the sample. The central validation δ_5 quantile is 0.2, corresponding to a value of $\log(1 + \delta_5) = -0.20$, and the normalized RMSE is approximately 0.40 ± 0.01 .

We find that the H I-morphology relation transitions to a different “mode” at high δ_5 . For low-density environments, a CNN is able to leverage the morphological information learned in intermediate- and high-density

⁹ For each environmental test run, a 34-layer xresnet is trained for 10 epochs using a learning rate of 0.03, batch size of 32, weight decay of 10^{-4} , and the validation subsample is evaluated using test-time augmentation. These hyperparameters have been chosen to best optimize the CNN in a smaller number of training epochs so that we can run multiple tests quickly.

Table 4. $\alpha.40A$ results split by environment

Training $N = 5922$	Validation $N = 1477$	Validation σ (dex)	Validation RMSE (dex)	Normalized RMSE
[0.2, 1.0)	[0, 0.2)	0.5241	0.2184 ± 0.0022	0.4167 ± 0.0042
[0, 0.8)	[0.8, 1.0)	0.6706	0.3269 ± 0.0066	0.4874 ± 0.0098
Random	Random	0.6036	0.2557 ± 0.0094	0.4237 ± 0.0156

NOTE— Comparison of CNN performance using different training/validation splits for $\alpha.40A$. The training and validation subsamples are either randomly selected or separated by δ_5 quantile according to an 80%/20% ratio in the given quantile range.

regimes and accurately predict the gas mass fraction directly from imaging. For high-density environments, a CNN is not able to generalize information learned from low- and intermediate-density regimes as well, and the normalized RMSE increases significantly. A physical explanation for this transition is the growing importance of ram pressure stripping, tidal forces, and other gas depletion effects in overdense environments. We determine that these effects become increasingly significant at 0.8 quantile in δ_5 for $\alpha.40A$, corresponding to a normalized overdensity of $\log(1 + \delta_5) \geq 0.5$; for lower values of δ_5 , the physics that govern this H I-morphology relation are constant.

8. INTERPRETING MORPHOLOGICAL FEATURES

Deep learning models often contain millions of trainable parameters, which makes them difficult to interpret compared to classical statistical or smaller machine learning models. For this reason, deep neural networks are often viewed as opaque systems that cannot be trusted. Indeed, there are many cases in which deep learning algorithms make high-confidence predictions while failing spectacularly; e.g., when a CNN is confronted with adversarial examples, out-of-distribution predictions, or domain adaptation problems (e.g., Amodei et al. 2016). In Section 4.3, we examined how a CNN trained on ALFALFA could not be used to make predictions for xGASS without pattern recognition, because the two galaxy populations have different distributions of physical properties. Alternatively, poor generalization across domains can occur if the training and test data sets are systematically different, e.g., if the training data comprises images of simulated galaxies while the test set solely comprises images of observed galaxies (e.g., Rafieferantsoa et al. 2018; Andrianomena et al. 2020). In Section 7, we exploited these failure modes in order to probe how galaxy environment impacts the learned H I-morphology relation. However, other unknown factors may cause CNNs

to perform poorly. Therefore, it is critical to investigate whether or not trained CNNs make sensible predictions in line with our physical intuition.

It is possible to decipher deep CNNs by examining which parts of an image are most relevant for making certain predictions. This method of localizing image features is generally known as *input attribution*, because output predictions can be directly attributed to pixels from the input images. Input attribution methods such as saliency maps and class activation maps (Simonyan et al. 2013; Zhou et al. 2016) are useful for identifying the image features that a trained CNN “looks at” in order to make its predictions. Other methods can also provide valuable insights, such as by visualizing the learned convolutional layers in optimzied CNNs (e.g., Zeiler & Fergus 2013).

8.1. Grad-CAM

To interpret our results, we make use of the Gradient-weighted Class Activation Map (Grad-CAM) visualization algorithm (Selvaraju et al. 2017). Grad-CAM is an input attribution tool that highlights the activated “neurons” in a trained CNN corresponding to the pixels in an input image that are most important for predicting a designated class. These visual explanations enable us to directly attribute output predictions to input morphological features represented as pixels. In order to use this algorithm, we reformulate our gas mass fraction regression problem to a binary classification problem.

We train a CNN to classify gas-rich and gas-poor galaxies in the $\alpha.40A$ sample. We define low- \mathcal{M} (gas-poor) and high- \mathcal{M} (gas-rich) as $\mathcal{M} < -0.5$ (1,327 objects) and $\mathcal{M} > 0.5$ (1,922 objects) respectively, so that the two classes are well-separated. According to this classification, some star-forming galaxies are labeled as “gas-poor,” so this scheme is only appropriate for the H I-rich ALFALFA sample.

We also use a simple CNN for visualization purposes. Our previous architecture (34-layer xresnet) contains many pooling layers that each decrease the resolution

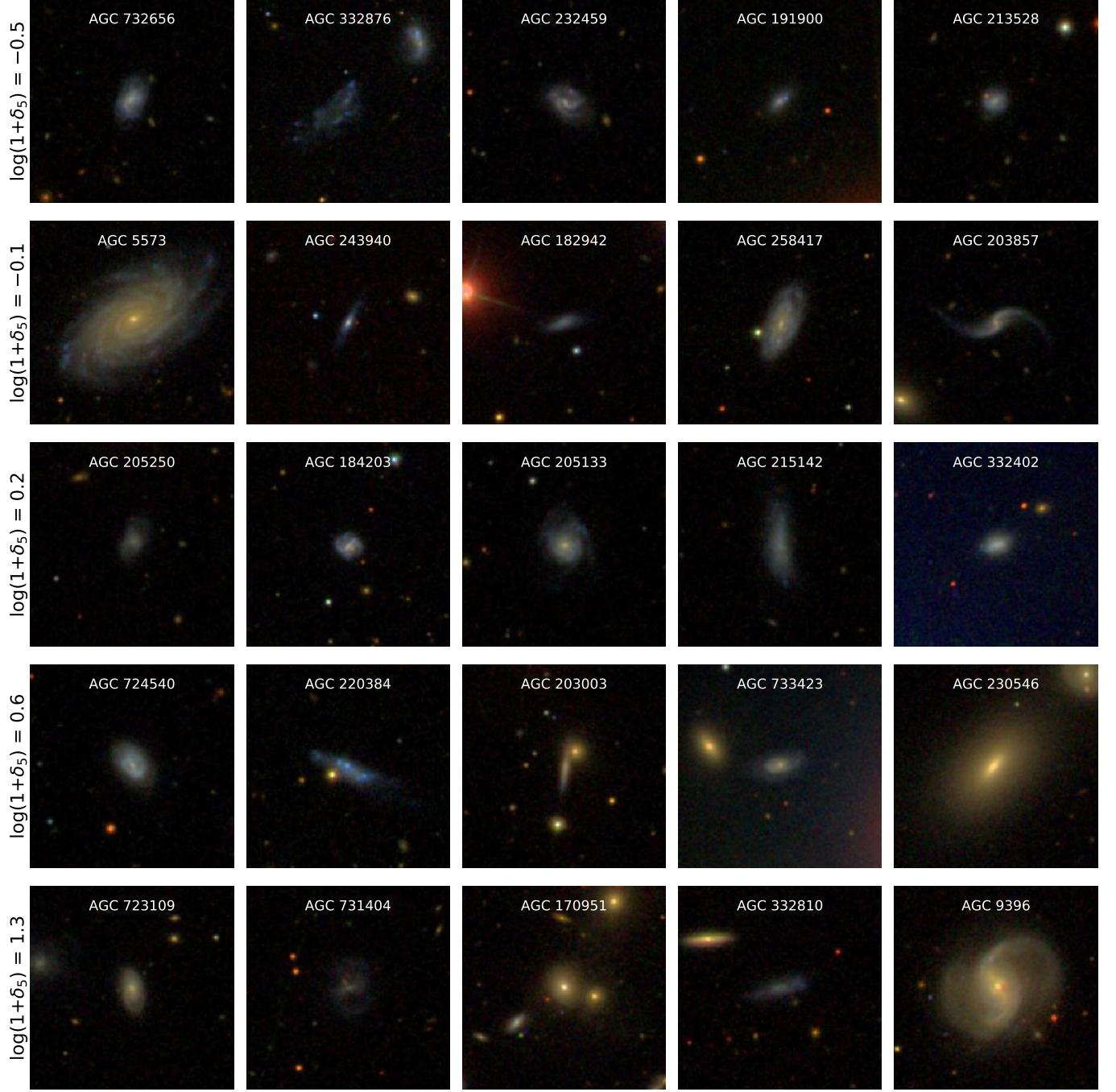


Figure 9. Example SDSS images of $\alpha.40A$ galaxy image cutouts in different environmental regimes. Each row depicts five galaxies with environmental overdensity closest to the value indicated on the left. The example galaxies range from below 10th percentile to over 90th percentile in $\log(1 + \delta_5)$.

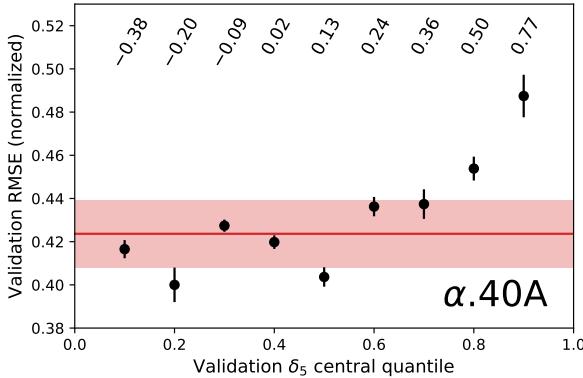


Figure 10. CNN validation performance across different environmental densities shown in black markers and error bars. The performance is the RMSE normalized by the inherent scatter in \mathcal{M} for the validation set; we show the mean and standard deviation for five tests. Each validation set is constructed from a 20% range in δ_5 , and the corresponding central $\log(1 + \delta_5)$ value is shown at the top. Validation results from randomly drawn subsamples are shown in red.

by a factor of two, such that the final Grad-CAM result is a 7×7 pixel feature map. Instead, we use a shallower CNN that consists of a basic CNN stem and two residual blocks containing two convolutional layers each (see, e.g., Howard & Gugger 2020). The final convolutional layer outputs a 56×56 pixel feature map. We use the same optimization methods as in Section 3, except that we train for only 10 epochs (at which point we reach convergence) and optimize using cross entropy loss with $\epsilon = 0.05$ label smoothing. The shallow network classifies $\alpha.40A$ galaxies by gas richness with 98% accuracy.

8.2. The most important morphological features

The trained CNN outputs probabilities for each predicted class for an input image. Grad-CAM can be used to highlight the most important morphological features used for making correct and incorrect predictions, and both sets of image features are valuable for understanding what the CNN has learned. The association between blue stellar populations at the edge of a galaxy’s star-forming disk and a high- \mathcal{M} classification, for example, strengthens our confidence in trained CNN. Examples of the galaxy image cutout, low- \mathcal{M} features, and high- \mathcal{M} features are shown in Figure 11. Below, we list the most commonly observed results.

1. H II regions, often indicated by bright, blue, compact features, and spiral arms, usually signify that a galaxy has high gas mass fraction.
2. Red central regions due to older stellar populations tend to be associated with low \mathcal{M} . However,

red stellar populations can sometimes be conflated with dust.

3. If the flocculent outer regions of a galaxy are blue, then the CNN tends to predict that it is gas-rich, but if the outer regions are populated with redder stars (e.g., panel d of Figure 11), then the galaxy is more often predicted to be gas-poor (e.g., Koopmann & Kenney 2004).
4. Nearby objects in the field of view, even ones that are clearly background or foreground objects, are often considered by the CNN. They may be highlighted as evidence for low or high \mathcal{M} depending on their relative color to the main system; however, their contributions to the overall prediction are usually subdominant. In Appendix B, we verify that artificial point sources are generally unimportant for the CNN’s decision-making process.

8.3. The value of single-band imaging

It is clear that the CNN relies on color information to identify gas-rich or gas-poor regions. Galaxy morphology is another useful, albeit subdominant, parameter for estimating gas mass fraction (e.g., Zhang et al. 2009; Eckert et al. 2015; Teimoorinia et al. 2017). However, morphological features often covary with color because galaxy structures are linked to their stellar populations. Therefore, we aim to identify the most crucial morphological features using monochromatic imaging, i.e., single-channel image cutouts with summed g , r , and i flux. We first verify that this is possible by repeating the regression analysis in Section 4 using single-band imaging for $\alpha.40A$, and recover \mathcal{M} to within $\text{RMSE} = 0.29$ dex. We also note that CNNs have had some success predicting other gas-phase metallicity from single-band imaging (see Section 5.3 of Wu & Boada 2019).

We train a CNN to classify gas-rich and gas-poor galaxies using monochromatic imaging. The data set, model, and optimization steps are otherwise the same as described in the previous section. We find that the shallow CNN is able to classify monochromatic $\alpha.40A$ galaxies to over 90% accuracy.

In Figure 11, we show Grad-CAM results on single-band imaging. By examining the highlighted activations on monochromatic images, we are able to discern the morphological indicators of gas richness. We find that the CNN inspects the outskirts of galaxies and highlights point source-like objects when identifying gas-rich features. Many of these compact features are star-forming knots in spiral arms or the extended disk, but the CNN also takes background sources into consideration (e.g.,

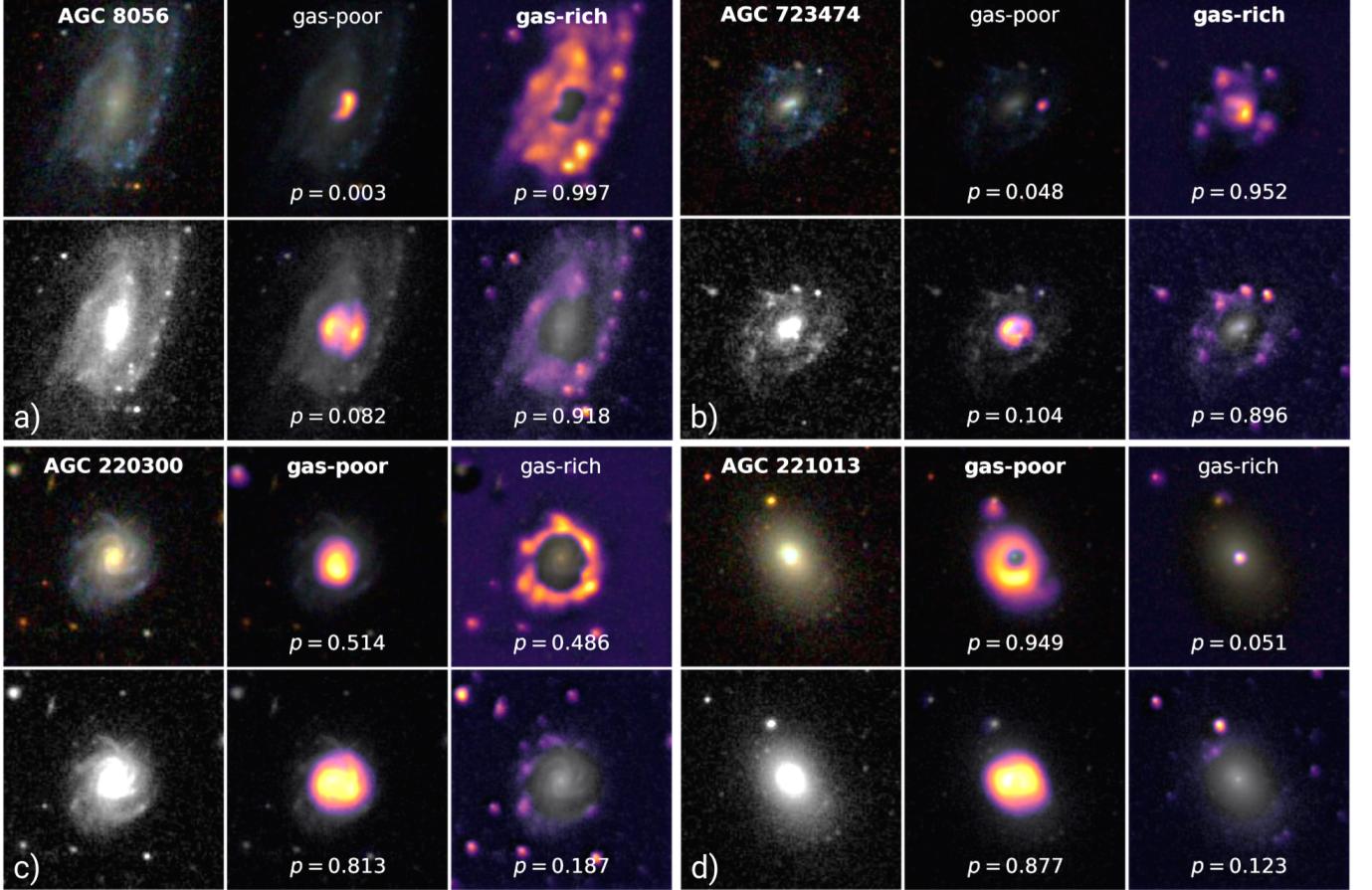


Figure 11. Grad-CAM heatmaps shown for SDSS images using trained CNNs. Each panel shows, from left to right, the SDSS *gri* image cutout, the heatmap of activations corresponding to gas-poor features, and the heatmap of activations corresponding to gas-rich features. Grad-CAM heatmaps are shown for for *gri* color input images (upper) and monochromatic input images (lower). Gas-poor/gas-rich labels are bolded for ground truth values, and CNN probabilities are provided for each class. The image contrast has been reduced for visualization purposes.

panel c). Grad-CAM also reveals that the CNN focuses on galaxy centers when identifying gas-poor features in monochromatic imaging. We surmise that it is relying the central surface brightness to recognize whether a galaxy is gas-poor; surprisingly, it is able to leverage this information even though no distance information is provided.

9. DEEPER IMAGING AND FUTURE SURVEYS

Future optical-wavelength surveys will offer deeper imaging data sets useful for characterizing the gas properties of galaxies. We obtain *grZ* imaging from the DESI Legacy Imaging Surveys DR8 (Legacy Survey; Dey et al. 2019) in order to compare with our previous results. Using the online interface¹⁰, we query 448×448 pixel JPG cutouts at the native $0.262'' \text{ pixel}^{-1}$ scale for both the $\alpha.40$ and xGASS samples (again using optical counter-

part coordinates for the former). Legacy Survey imaging is deeper than that of SDSS by about two magnitudes, and has higher angular resolution (although it remains seeing-limited). Deep optical imaging is particularly critical for identifying low-surface brightness features in galaxies with complex star formation histories or recent gas accretion (e.g., Duc et al. 2015; Geréb et al. 2016; Hagen et al. 2016).

Using Legacy Survey imaging, and the same training methodology as described in Section 3, we find similar results for $\alpha.40A$ as before. Our early results are promising and suggest that deeper optical imaging may be useful for improving \mathcal{M} predictions.¹¹ It is

¹¹ It is difficult to directly compare the two imaging data sets: Legacy Survey image cutouts have an expanded field-of-view ($1.96'$) compared to SDSS imaging ($1.48'$), and the use of *Z* rather than *i*-band imaging in the reddest channel may also affect CNN performance.

¹⁰ <https://legacysurvey.org/viewer>

also worth noting that the Legacy Survey DR8 imaging suffers from some imaging issues, such as pixel bleed, sky subtraction, and inconsistent zero-points in different bands, which may prevent the model from learning as much as it can. These effects must be remedied if we want to maximize scientific gains through the combination of deep learning and wide-field optical/near-infrared surveys (e.g., the Vera C. Rubin Observatory (VRO) Legacy Survey of Space and Time (LSST), [Ivezic et al. 2019](#); and the *Nancy Grace Roman Space Telescope* (*RST*, formerly *WFIRST*; [Spergel et al. 2015](#)).

Current H I surveys are mostly mass-limited, but SKA precursor surveys such as DINGO and LADUMA will be much more sensitive to gas-poor galaxy populations. These new surveys will allow us to construct data sets similar to the xGASS representative sample or the volume-limited RESOLVE survey (REsolved Spectroscopy Of a Local VolumE; e.g., [Stark et al. 2016](#)), except with orders of magnitude more detections at the same H I mass threshold. In the future, we may be able to take deep H I 21-cm line observations of some small patch of sky, and then use deep optical imaging in overlapping portions in order to generate \mathcal{M} predictions for galaxies across the entire optical survey area (e.g., [Domínguez Sánchez et al. 2019](#); [Khan et al. 2019](#)). The methods introduced in this paper may also allow us to probe the redshift evolution of the overdensity transition regime (Section 7) or evolution of the most relevant morphological features associated with gas richness over cosmic timescales (Section 8). These tantalizing prospects can be realized, but only if the co-evolving H I and stellar mass functions (e.g., [Lemonias et al. 2013](#)) and their effects on the priors baked into the trained CNN model are taken into account (e.g., by sampling according to a known distribution; [Buda et al. 2017](#)). Moreover, cosmic variance effects for deep H I surveys need to be considered (e.g., [Moster et al. 2011](#)). Finally, it is imperative to deploy robust pattern recognition algorithms in order to safeguard against out-of-distribution errors and gauge the reliability of machine learning predictions.

10. CONCLUSIONS

In this work, we have found that deep CNNs can predict a galaxy’s H I mass fraction (\mathcal{M}) solely from *gri* imaging to within RMSE = 0.23 dex for the $\alpha.40A$ sample, demonstrating that there is a strong connection between galaxy morphology and H I content. We have also trained a CNN for pattern recognition (PR), e.g., determining whether a galaxy is likely to be detected by an ALFALFA-like survey based on its optical imaging. The combined regression and PR results generalize well

to new test data sets, and our experiments indicate that PR threshold of $p_{\text{CNN}} > 0.9$ is best-suited for optically selected galaxy samples. We find that the CNN consistently outperforms previous machine learning methods on matched test data; using a $p_{\text{CNN}} > 0.9$ PR cut, we report RMSE = 0.20 dex scatter for $\alpha.100$, 0.29 dex scatter for NIBLES, and 0.22 dex scatter for xGASS. Our methodology can be augmented with deeper imaging or larger and more diverse galaxy samples. With the advent of next-generation H I 21-cm emission line surveys with the SKA precursor telescopes, and LSST and the *Roman Space Telescope* on the horizon, it will soon be possible to generate enormous CNN-predicted H I catalogs.

We are able to probe the environmental dependence of the H I-morphology relation by independently training and validating CNNs using subsamples stratified by galaxy overdensity (i.e., δ_5 , the normalized projected density). For high-density environments, a CNN trained on lower- δ_5 examples is unable to accurately estimate \mathcal{M} from optical imaging. However, if the validation set comprises galaxies in low- or intermediate-density environments, then a CNN has no trouble predicting \mathcal{M} . We propose that in the most overdense environments, $\log(1+\delta_5) \gtrsim 0.5$ for $\alpha.40A$, physical processes such as ram pressure stripping, tidal interactions, and other gas depletion effects are responsible for “breaking down” the H I-morphology relation observed in less dense environments.

We have also reformulated the problem of estimating \mathcal{M} as a binary classification task in order to better understand how CNNs are able to distinguish gas-poor from gas-rich systems. We use Gradient-weighted Class Activation Maps (Grad-CAM) to localize the optical features that are most important for predicting whether or not a galaxy is gas-rich. Bright star-forming regions and clumpy blue features usually imply high \mathcal{M} , while central red bulges and older stellar populations at large radii often indicate low \mathcal{M} . The CNN successfully distinguishes gas-rich and gas-poor galaxies with $> 90\%$ accuracy using single-band optical images, implying that it is able to identify purely morphological features for estimating gas content.

We have highlighted several ways that deep learning and computer vision can be useful for understanding galaxy evolution. Apart from predicting \mathcal{M} and the gas fraction’s reliability directly from optical imaging, CNNs can also be used to gauge the impact of covarying galaxy properties such as environmental overdensity. These methods are visually interpretable and provide key insights into the physical processes and stellar/ISM structures that are most closely connected to the H I properties in galaxies.

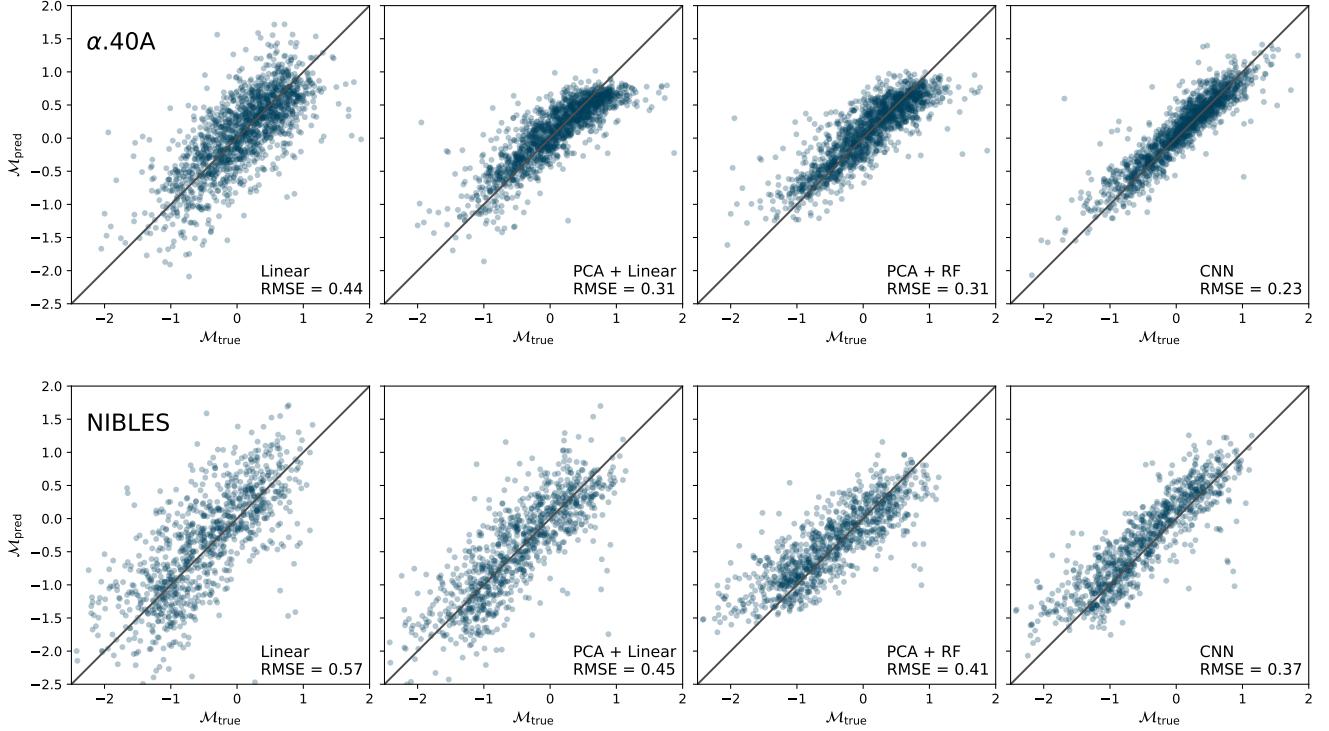


Figure 12. Comparison of different regression models for estimating \mathcal{M} using the ALFALFA $\alpha.40A$ validation sample (top) and the NIBLES test sample (bottom). Each panel shows $\mathcal{M}_{\text{pred}}$ plotted against $\mathcal{M}_{\text{true}}$ for a linear model trained on block-reduced images (left), a linear model trained on PCA-processed images (center-left), a random forest model trained on the PCA-processed images (center-right), and a CNN trained on the original images (right).

APPENDIX

A. COMPARING CNNS TO SIMPLER MODELS

Classical machine learning algorithms and statistical methods are not well-suited for operating on image data because they are not invariant to translation, scaling, or rotation. In other words, individual pixels may represent different galaxy features for different images, and any model that treats a pixel as a static feature will not perform well in computer vision problems. CNNs are able to encode optimized representations of galaxy features through convolution operations, which largely do not depend on the feature’s absolute location within an image. Moreover, these invariances can be learned via redundant convolutional filters by employing data augmentation (such as shifts, rotations, and crops; e.g., Dieleman et al. 2015).

Our data set consists of SDSS image cutouts centered on the optical sources of various H I catalog members. Although the galaxies are still observed at different position angles and inclinations, and have different physical scales because they span $0 \leq z \leq 0.06$, the galaxies’ central regions always occupy the center pixels. For this reason, it may be possible to use simple models, rather than a CNN, in order to estimate the H I mass fraction. We use the `sklearn` Python package to pre-process and fit our data.

The image data are represented as arrays with $3 \times 224 \times 224$ elements. Most simple regression models are ill-equipped to handle 150,528 inputs at a time, so we use either one of two methods to lower the dimensionality of the input data. The first method is to block-reduce each training image using a 7×7 kernel (also known as average binning or pooling), resulting in a $3 \times 32 \times 32$ -shaped input. Each block-reduced array is then flattened into a one-dimensional vector, so that the independent variables can be written as a $N \times 3,072$ matrix, where N is the number of galaxies in the training set. The second method is to use a principal components analysis (PCA) on the training set, where only the top

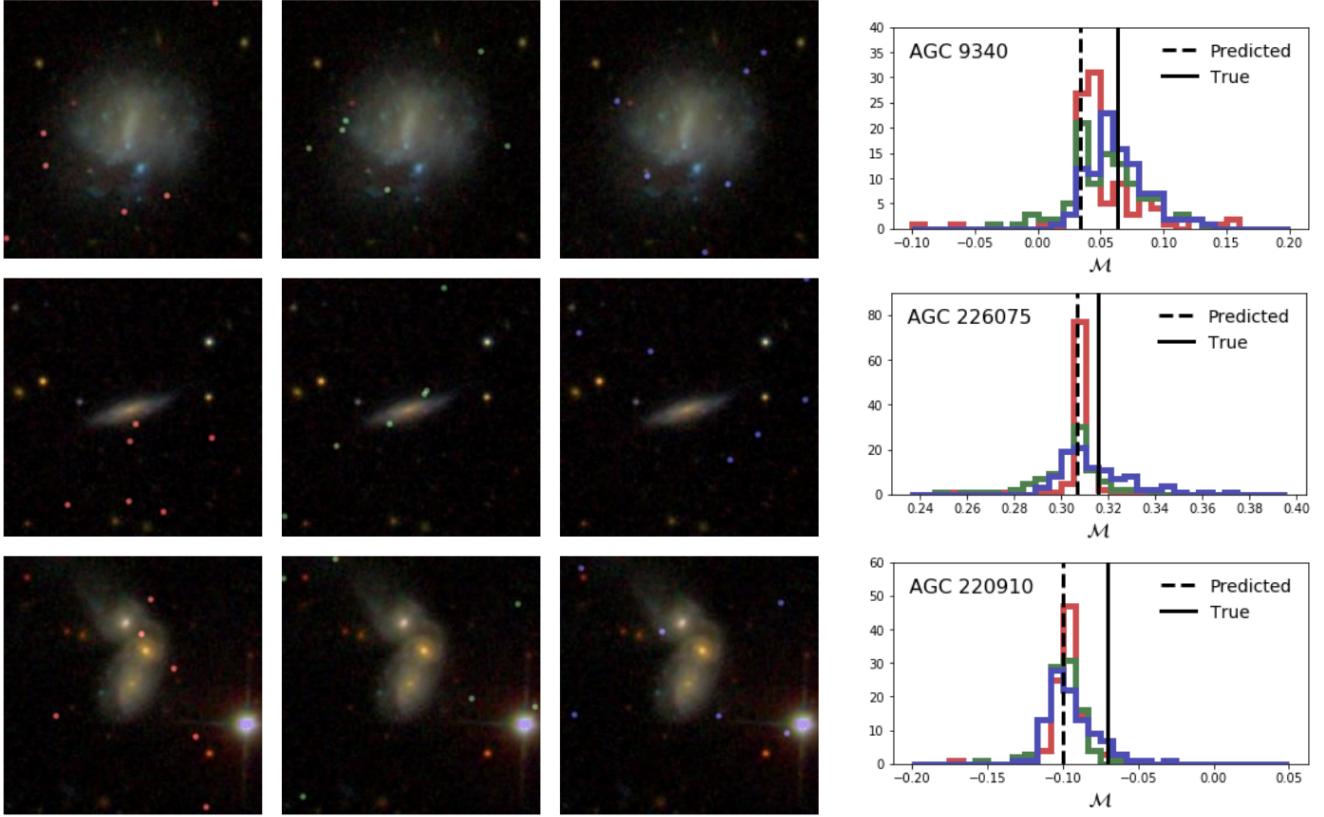


Figure 13. Examples of artificial point source injection for AGC 9340 (upper), AGC 226075 (center), and AGC 220910 (lower). For each row, starting from left to right, we show images with six artificial red sources, green sources, and blue sources. The right-most panel in each row compares the original $\mathcal{M}_{\text{pred}}$ (dashed vertical line) and $\mathcal{M}_{\text{true}}$ (solid vertical line) to histograms of 100 simulations for each injected source color (red, green, and blue).

16 components are flattened and saved. The PCA-processed independent variables can be written as a $16 \times 150,528$ matrix.

After reducing the dimensionality of the training inputs, we select one of two algorithms for statistical regression. The first algorithm is an ordinary least-squares regression to fit a low-order polynomial model. In practice we find that a linear model always outperforms a quadratic model, so only linear regression results are included in this discussion. The linear regression model requires 3,073 trainable parameters for block-reduced images, and 17 trainable parameters for PCA-processed images. The second algorithm is a random forest (RF), which bootstraps (i.e., samples with replacement) 100 random decision tree estimators. The RF regression model is optimized according to the mean squared error loss.

In Figure 12, we compare different models trained on 80% of $\alpha.40A$ and validated on the remaining 20% for $\alpha.40A$ (left) and for the entire NIBLES test set (right). All model predictions are impacted by the 0.16 dex systematic offset in H I mass between NIBLES and ALFALFA (van Driel et al. 2016). The linear models fit to block-reduced images do not perform well, as indicated by high scatter. PCA-processed data provide better results than block-reduced images, although overall performance is still modest. Ultimately, CNNs outperform all of the simpler models that we test in terms of slope and scatter.

B. ROBUSTNESS TO PERTURBATIONS

A concern with CNNs and deep learning algorithms is whether or not their predictions can be significantly swayed by image noise or other perturbations. For example, galaxy images with foreground stars or other faint background sources should not cause predictions to vary wildly (unless they are located in regions where the CNN ascribes high importance, which we can probe using Grad-CAM; see Section 8). To test our method’s performance when small changes are added to the images, we randomly add colored point sources (resembling artificial stars) to three representative galaxy images,

and allow the CNN to infer $\mathcal{M}_{\text{pred}}$. Specifically, we add six artificial sources (two-dimensional circular Gaussian profiles with a 5-pixel radius, all of which are red, green, or blue) to random locations in the original image. Figure 13 shows examples of the three galaxy images with injected artificial sources. 100 trials are run for each of the three colors, and we compare these perturbed predictions to the original estimate ($\mathcal{M}_{\text{pred}}$) and the ground truth ($\mathcal{M}_{\text{true}}$). We find the CNN is able to make generalized predictions that does not depend on the injected point sources; the typical scatter due to these injected sources is much smaller (< 0.05 dex) than the overall RMSE. Thus, we conclude that our trained CNNs are robust to perturbations such as artificial point sources.

REFERENCES

- Abolfathi, B., Aguado, D. S., Aguilar, G., et al. 2018, ApJS, 235, 42
- Amodei, D., Olah, C., Steinhardt, J., et al. 2016, arXiv e-prints, arXiv:1606.06565
- Andrianomena, S., Rafieferantsoa, M., & Davé, R. 2020, MNRAS, 492, 5743
- Barnes, D. G., Staveley-Smith, L., de Blok, W. J. G., et al. 2001, MNRAS, 322, 486
- Bishop, C. M. 1995, Neural Networks for Pattern Recognition (USA: Oxford University Press, Inc.)
- Blanton, M. R., Kazin, E., Muna, D., Weaver, B. A., & Price-Whelan, A. 2011, AJ, 142, 31
- Blyth, S., Baker, A. J., Holwerda, B., et al. 2016, in Proceedings of MeerKAT Science: On the Pathway to the SKA. 25-27 May, 4
- Brinchmann, J., Charlot, S., White, S. D. M., et al. 2004, MNRAS, 351, 1151
- Brown, T., Catinella, B., Cortese, L., et al. 2017, MNRAS, 466, 1275
- Buda, M., Maki, A., & Mazurowski, M. A. 2017, arXiv e-prints, arXiv:1710.05381
- Caldeira, J., Wu, W. L. K., Nord, B., et al. 2019, Astronomy and Computing, 28, 100307
- Catinella, B., Schiminovich, D., Kauffmann, G., et al. 2010, MNRAS, 403, 683
- Catinella, B., Schiminovich, D., Cortese, L., et al. 2013, MNRAS, 436, 34
- Catinella, B., Saintonge, A., Janowiecki, S., et al. 2018, MNRAS, 476, 875
- Chabrier, G. 2003, PASP, 115, 763
- Chung, A., van Gorkom, J. H., Kenney, J. D. P., Crowl, H., & Vollmer, B. 2009, AJ, 138, 1741
- Ćiprijanović, A., Snyder, G. F., Nord, B., & Peek, J. E. G. 2020, Astronomy and Computing, 32, 100390
- Cooper, M. C., Newman, J. A., Weiner, B. J., et al. 2008, MNRAS, 383, 1058
- Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168
- Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441
- Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al. 2019, MNRAS, 484, 93
- Duc, P.-A., Cuillandre, J.-C., Karabal, E., et al. 2015, MNRAS, 446, 120
- Eckert, K. D., Kannappan, S. J., Stark, D. V., et al. 2015, ApJ, 810, 166
- Ellison, S. L., Teimoorinia, H., Rosario, D. J., & Mendel, J. T. 2016, MNRAS, 455, 370
- Fabello, S., Kauffmann, G., Catinella, B., et al. 2012, MNRAS, 427, 2841
- Geréb, K., Catinella, B., Cortese, L., et al. 2016, MNRAS, 462, 382
- Giovanelli, R., Haynes, M. P., Kent, B. R., et al. 2005, AJ, 130, 2598
- Goyal, P., Dollár, P., Girshick, R., et al. 2017, arXiv e-prints, arXiv:1706.02677
- Hagen, L. M. Z., Seibert, M., Hagen, A., et al. 2016, ApJ, 826, 210
- Haynes, M. P., Giovanelli, R., Martin, A. M., et al. 2011, AJ, 142, 170
- Haynes, M. P., Giovanelli, R., Kent, B. R., et al. 2018, ApJ, 861, 49
- He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv e-prints, arXiv:1512.03385
- He, T., Zhang, Z., Zhang, H., et al. 2018, arXiv e-prints, arXiv:1812.01187
- Hopfield, J. J. 1987, Proceedings of the National Academy of Science, 84, 8429
- Howard, J., & Gugger, S. 2020, arXiv e-prints, arXiv:2002.04688
- Huang, S., Haynes, M. P., Giovanelli, R., & Brinchmann, J. 2012, ApJ, 756, 113
- Huertas-Company, M., Rodriguez-Gomez, V., Nelson, D., et al. 2019, MNRAS, 489, 1859
- Hunter, J. D. 2007, Computing in Science and Engineering, 9, 90
- Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111

- Jarvis, M., Taylor, R., Agudo, I., et al. 2016, in Proceedings of MeerKAT Science: On the Pathway to the SKA. 25–27 May, 6
- Jones, M. G., Papastergis, E., Haynes, M. P., & Giovanelli, R. 2016, MNRAS, 457, 4393
- Kannappan, S. J. 2004, ApJL, 611, L89
- Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, MNRAS, 341, 33
- Khan, A., Huerta, E. A., Wang, S., et al. 2019, Physics Letters B, 795, 248
- Kinney, A. L., Calzetti, D., Bohlin, R. C., et al. 1996, ApJ, 467, 38
- Koopmann, R. A., & Kenney, J. D. P. 2004, ApJ, 613, 866
- Koribalski, B. S., Staveley-Smith, L., Westmeier, T., et al. 2020, arXiv e-prints, arXiv:2002.07311
- Lemonias, J. J., Schiminovich, D., Catinella, B., Heckman, T. M., & Moran, S. M. 2013, ApJ, 776, 74
- Li, C., Kauffmann, G., Fu, J., et al. 2012, MNRAS, 424, 1471
- Liu, L., Jiang, H., He, P., et al. 2019, arXiv e-prints, arXiv:1908.03265
- Loshchilov, I., & Hutter, F. 2017, arXiv e-prints, arXiv:1711.05101
- Lupton, R., Blanton, M. R., Fekete, G., et al. 2004, PASP, 116, 133
- McKinney, W. 2010, in Proceedings of the 9th Python in Science Conference, ed. S. van der Walt & J. Millman, 51 – 56
- Misra, D. 2019, arXiv e-prints, arXiv:1908.08681
- Morningstar, W. R., Perreault Levasseur, L., Hezaveh, Y. D., et al. 2019, ApJ, 883, 14
- Moster, B. P., Somerville, R. S., Newman, J. A., & Rix, H.-W. 2011, ApJ, 731, 113
- Müller, R., Kornblith, S., & Hinton, G. 2019, arXiv e-prints, arXiv:1906.02629
- Odekon, M. C., Koopmann, R. A., Haynes, M. P., et al. 2016, ApJ, 824, 110
- Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26
- Paszke, A., Gross, S., Massa, F., et al. 2019, in Advances in Neural Information Processing Systems 32, ed. H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, & R. Garnett (Curran Associates, Inc.), 8024–8035
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825
- Peek, J. E. G., & Burkhart, B. 2019, ApJL, 882, L12
- Rafieferantsoa, M., Andrianomena, S., & Davé, R. 2018, MNRAS, 479, 4509
- Saintonge, A., Catinella, B., Tacconi, L. J., et al. 2017, ApJS, 233, 22
- Salim, S., Rich, R. M., Charlot, S., et al. 2007, ApJS, 173, 267
- Selvaraju, R. R., Cogswell, M., Das, A., et al. 2017, in 2017 IEEE International Conference on Computer Vision (ICCV), 618–626
- Serra, P., Oosterloo, T., Morganti, R., et al. 2012, MNRAS, 422, 1835
- Serra, P., Koribalski, B., Kilborn, V., et al. 2015, MNRAS, 452, 2680
- Simonyan, K., Vedaldi, A., & Zisserman, A. 2013, arXiv e-prints, arXiv:1312.6034
- Smith, L. N. 2018, arXiv e-prints, arXiv:1803.09820
- Spergel, D., Gehrels, N., Baltay, C., et al. 2015, arXiv e-prints, arXiv:1503.03757
- Stark, D. V., Kannappan, S. J., Eckert, K. D., et al. 2016, ApJ, 832, 126
- Stevens, A. R. H., Diemer, B., Lagos, C. d. P., et al. 2019, MNRAS, 483, 5334
- Teimoorinia, H., Ellison, S. L., & Patton, D. R. 2017, MNRAS, 464, 3796
- Tremonti, C. A., Heckman, T. M., Kauffmann, G., et al. 2004, ApJ, 613, 898
- van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science and Engineering, 13, 22
- van Driel, W., Butcher, Z., Schneider, S., et al. 2016, A&A, 595, A118
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2019, arXiv e-prints, arXiv:1907.10121
- Wu, J. F., & Boada, S. 2019, MNRAS, 484, 4683
- Zeiler, M. D., & Fergus, R. 2013, arXiv e-prints, arXiv:1311.2901
- Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. 2018, arXiv e-prints, arXiv:1805.08318
- Zhang, M. R., Lucas, J., Hinton, G., & Ba, J. 2019, arXiv e-prints, arXiv:1907.08610
- Zhang, W., Li, C., Kauffmann, G., et al. 2009, MNRAS, 397, 1243
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. 2016, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2921–2929
- Zhu, W. W., Berndsen, A., Madsen, E. C., et al. 2014, ApJ, 781, 117

Software: Numpy ([van der Walt et al. 2011](#)), scikit-learn ([Pedregosa et al. 2011](#)), Scipy ([Virtanen et al. 2019](#)), matplotlib ([Hunter 2007](#)), Pandas ([McKinney 2010](#)), Pytorch ([Paszke et al. 2019](#)), Fastai (<https://github.com/fastai/fastai>)

ACKNOWLEDGMENTS

The author would like to thank the anonymous referee for useful and detailed comments that have significantly improved this manuscript. The author would like to thank Josh Peek for suggesting the idea of using CNNs to probe galaxy environments and many other useful discussions. The author also thanks Luke Leisman and Mike Jones for helpful conversations regarding the ALFALFA data. The author acknowledges support from the National Science Foundation under grants NSF AST-1517908 and NSF AST-1616177, and also thanks the Pascal Institute for their hospitality (The Self-organised Star Formation Process program). This research was supported by the Munich Institute for Astro- and Particle Physics (MIAPP) which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy - EXC-2094 - 390783311. This work made use of Google Colab and Google Compute Engine.

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.