UNIVERSITY OF CALIFORNIA
SANTA CRUZ

NEURAL PROBABILISTIC MODELING FOR ASTROPHYSICS AND GALAXY EVOLUTION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PHYSICS

by

David M. Reiman

September 2020

The Dissertation of David M. Reiman
is approved:

Professor Joel R. Primack, Chair

Professor J. Xavier Prochaska

Professor David Koo

Dr. Doug Hellinger

Quentin Williams
Interim Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Neural Probabilistic Modeling for Astrophysics and Galaxy Evolution

by

David M. Reiman

Statistical modeling in modern astrophysics and cosmology frequently involves simplified analytic models which fail to capture the underlying complexity of the problem at hand. And though traditional machine learning methods have proved useful, they scale poorly with dataset size and will struggle to make optimal use of the vast quantities of data soon to be produced by near-future surveys. Recently, neural networks have provided a powerful new foothold for probabilistic modeling in the presence of large data volumes by acting as expressive universal function approximators. In this thesis, I consider two applications of neural networks: (i) I use a convolutional neural network with an adversarial regularizing loss to deblend superimposed galaxy images. I focus on two-component blends and show that a model trained with a combination of supervised pixel-wise and regularizing adversarial losses provides high fidelity deblended images. And (ii) I use normalizing flows, a neural density estimation technique, to model the distribution over intrinsic quasar continua near Lyman-α given the redward spectrum. I then constrain the timeline of the Epoch of Reionization by measuring the neutral fraction of hydrogen in the spectrum of two $z > 7$ quasars. I also describe ongoing work on recovering the local density field in the vicinity of distant galaxies with attention-based graph neural networks and likelihood free inference for astrophysical simulator

models with intractable likelihoods.

To my parents...

Mark and Lori,

To my sisters...

Anna and Stephanie,

To my brothers...

Jake and Adam.

Part I

Introduction

Chapter 1

Foreword

Astronomy is the most ancient of the natural sciences, and though the tools and methods astronomers use to learn about the cosmos have improved radically since Galileo, they have perhaps never improved as rapidly as now. The fields of astronomy and cosmology are presently overflowing with data, largely due to improvements in technology such as the charge-coupled device and the wide-field digital camera.

In the recent past, images for astronomical surveys were captured on thin glass plates covered in photo-sensitive silver salts. For example, the Palomar Observatory Sky Survey photographed the night sky using color-sensitive (red or blue) 14-inch photographic plates which allowed for a rudimentary measurement of galaxy color. Similarly, the spectra (intensity of light as a function of wavelength) of celestial bodies was recorded by shining diffracted light onto photographic plates.

Since then, the advent of the charge-coupled device (CCD) has overtaken the use of photographic plates, though some observatories have only adopted the CCD as

recently as the 1990s. Among the first astronomical images recorded via charge-coupled devices were low resolution captures of the lunar surface (1974), Uranus (1976), and Saturn (1976). The original CCD image of the moon was 100 pixels square (or 0.01 megapixels). In comparison, the Hubble Space Telescope is presently equipped with Wide Field Camera 3, a fourth-generation Hubble instrument with a 16 megapixel CCD. With its capability to create mosaic images of the cosmos, Hubble has produced images eclipsing 1.5 gigapixels in size (for example, of our galactic neighbor Andromeda). Near-future ground-based instruments such as the Vera C. Rubin Observatory boast 6.2 gigapixels cameras. Moreover, the telescope enjoys a 3.5-degree field of view—over 75 times larger than that of Hubble's Wide Field Camera 3.

With such advancements in imaging technology came rapid influxes of data. In fact, modern astronomical surveys capture nightly data volumes larger than entire surveys of a decade ago [56], rendering real-time analysis of the data all but impossible. Where astronomers previously identified celestial objects by visual inspection of the photographic plates, modern astronomers rely on computational methods to parse vast quantities of data in short order. Rich with data, astronomers look to the burgeoning field of machine learning for its scalable algorithms and impressive automated pattern recognition capabilities.

Machine learning, a term coined by Arthur Samuel of IBM in 1959, is a hybrid field combining applied statistics and computer science with the intent to discover and learn the underlying structure in data without explicitly programmed instructions. Classical machine learning algorithms are principally inductive (as are all statistical

models): they attempt to learn, for example, the fundamental mapping between a set of predictor variables (referred to as *features*) and response variables (often termed *labels*) or to infer the often complex and high dimensional distribution from which a body of data arises given only discrete examples. As an elementary but familiar example, linear regression attempts to learn the linear function which maps a set of predictor variables to a single response variable by fitting such a function to a collection of examples. It is important to note that each statistical model encodes an inductive bias: the set of assumptions which the model relies on to generalize to unseen data. In the case of linear regression for example, this inductive bias is the assumption of linearity.

At present, the most prominent variation of machine learning concerns the study and application of *artificial neural networks*. Neural networks rose to prominence after their decisive victory in the annual ImageNet image classification competition of 2012 [57]. In this field of *deep learning*, statistical models loosely based on the animal brain learn increasingly non-linear representations of input data by passing features through sequential layers of artificial neurons. These representation spaces typically make downstream tasks such as classification or regression trivial for example by contorting what would be a highly non-linear classification boundary in the data space to a simply linear boundary in the representation space.

The expressivity of such a network, a measure of the number of linear regions in functions represented by the network, has been shown to increase exponentially with network depth (number of layers) but only polynomially with layer width [72]. This realization led researchers to build deeper and deeper neural networks.

Modern neural networks are highly parameterized statistical models often containing millions to billions of free parameters. They are commonly equipped with useful architectural innovations which expedite the learning process or encode prior knowledge for the problem at hand (for instance by introducing an advantageous inductive bias). At present, a number of neural network varieties have emerged as powerful state-of-the-art models for various research areas (typically for distinct data paradigms) such as convolutional neural networks [58] for images and videos, and transformers [104] for tasks involving sentences and other sequences of discrete random variables.

Many open problems in astronomy and astrophysics are largely translatable to these machine learning subfields. To give an example, the same convolutional neural networks which exhibit a super-human ability to classify natural images can automate the morphological classification of galaxy images [19]. Indeed, neural networks enjoy many properties crucial to astronomers: (a) neural networks can implicitly detect complex relationships among predictor variables and between predictor and response variables—relationships that are often infeasible to identify or express a priori, (b) model performance scales well with data volume, a result of the substantial number of free parameters (so-called *model capacity*), and (c) after training, inference speeds (for small or mid-sized models) are on the order of milliseconds with modern hardware.

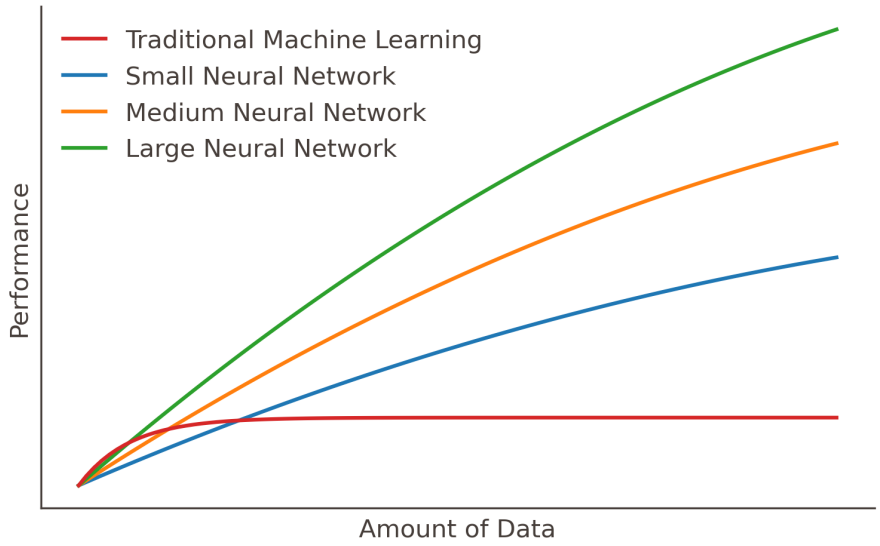


Figure 1.1: A depiction of deep learning performance versus data volume compared with traditional machine learning approaches.

Apart from purely scientific motivations, the natural sciences provide fertile ground for designing and implementing statistical models. This can be described in a number of points (originally and most eloquently put by my advisor Joel Primack): (a) scientific data is plentiful, (b) scientific data is widely available to the public (albeit often after an initial period of sequestration), and (c) we presume Nature to obey universal truths which should be revealable in scientific data.

In this thesis, I describe my pursuit to reveal a number of these truths. I will work at the interface of physics (principally astronomy) and machine learning to do so. In the introduction, I provide motivation for the use of neural networks in probabilistic modeling for astrophysics. Subsequently, I provide additional context for two manuscripts I've written over the course of my degree. I then discuss ongoing

work and conclude with a final discussion of my research and an overview of the many promising paths forward in the field of astrostatistics.

Chapter 2

Motivation

In the field of astronomy, probabilistic modeling is commonly employed as a means to describe the generative process of a collection of data via some underlying physics. Indeed, our observations must comply with many unbreakable known laws of physics (and perhaps some unknown or hidden laws) and therefore the space of possible models, though still large, is constrained.

However, traditional modeling techniques face two major problems. Astronomers most often employ parametric models wherein a fixed family of functions is specified and its optimal member is chosen based on the probability it assigns to a finite dataset via the likelihood function. Though these families of functions may be infinite sets, their ability to describe the process at hand is often limited by the expressivity of the family itself. If the functional family is misspecified, the optimal member will still fail to capture the underlying complexity of the data-generating process (the *modeling problem*). Moreover, in many domains including astronomy, likelihood functions are commonly

assumed to be Gaussian when the data is in truth highly non-Gaussian (the *likelihood problem*) for example by containing many modes, being highly skewed, or possessing very long tails.

Artificial neural networks, large computational graphs which act as expressive universal function approximators, solve both the modeling problem and the likelihood problem alike by replacing the model or the likelihood with a flexible, data-driven approximation. In doing so, their nascent use in astronomy and astrophysics has provided a powerful new foothold for probabilistic modeling in the presence of large data volumes.

In this thesis, I use neural networks both as flexible models of astrophysics as well as stand-ins for intractable likelihood functions and other conditional probability density functions. I focus on two projects I have led: (i) deblending galaxy image superpositions using convolutional neural networks with an adversarial regularizing loss, and (ii) predicting the conditional probability over plausible intrinsic quasar emission near Lyman-α given the redward spectrum.

In Part II, I describe my work using a generative adversarial network (GAN) regularized convolutional neural network to deblend galaxy superpositions on the sky. In a GAN, two neural networks engage in a zero-sum game wherein one neural network (the generator) attempts to produce samples similar to samples from the true data distribution. A second neural network (the discriminator) acts as a binary classifier and is trained to distinguish true samples from samples produced by the generator. The networks train in this manner until the generator can produce images indistinguishable from the true images—this training pushes generated images to manifold of natural

images to ensure realistic generations. I focus on two-component blends and use a dataset of artificial blends made of low-redshift Sloan Digital Sky Survey (SDSS) [4] composite images of the g, r, and i bands. I show that a pixel-wise loss combined with the GAN objective provide a strong foundation for deblending: the learned discriminator ensures that generated deblended images match the distribution over true galaxy images and the pixel-wise loss ensures that the features of the true underlying components in the blend are recovered.

Then, in Part III, I employ normalizing flows to predict the intrinsic continua near Lyman-α in high-redshift quasars to measure the damping wing of neutral hydrogen in the early Universe and thereby constrain the timeline of reionization. The normalizing flow acts as a learned bijection between the data space and an isotropic multivariate Gaussian space of equal dimensionality. The network sequentially contorts samples from one space into samples from the other. By doing so, inference becomes trivial: data are cast into the Gaussian space where density evaluation is trivial. To sample, we draw from the Gaussian space and cast the samples into the data space via the inverse transformation. I frame the problem as a conditional density estimation task and approximate the full distribution over intrinsic continua near Lyman-α given the redward spectrum by exploiting correlations in the quasar emission features. Our model is fully-probabilistic and allows calculation of confidence intervals via simple Monte Carlo sampling.

I then describe ongoing work which may be of interest to the reader in Part IV. This includes recovering the local density field in the vicinity of distant galaxies with

graph neural networks and using normalizing flows for likelihood-free inference for astrophysical simulator models with intractable likelihoods.

And finally, in Part V, I conclude my dissertation with a discussion of my work and a note on the promising future of machine learning applications to astrophysics and cosmology.

Part II

Deblending galaxy superpositions with branched generative adversarial networks

Chapter 3

Manuscript

3.1 Abstract

Near-future large galaxy surveys will encounter blended galaxy images at a fraction of up to 50% in the densest regions of the universe. Current deblending techniques may segment the foreground galaxy while leaving missing pixel intensities in the background galaxy flux. The problem is compounded by the diffuse nature of galaxies in their outer regions, making segmentation significantly more difficult than in traditional object segmentation applications. We propose a novel branched generative adversarial network (GAN) to deblend overlapping galaxies, where the two branches produce images of the two deblended galaxies. We show that generative models are a powerful engine for deblending given their innate ability to infill missing pixel values occluded by the superposition. We maintain high peak signal-to-noise ratio and structural similarity scores with respect to ground truth images upon deblending. Our model also predicts

near-instantaneously, making it a natural choice for the immense quantities of data soon to be created by large surveys such as LSST, Euclid and WFIRST.

3.2 Introduction

Large galaxy surveys necessarily encounter galactic blends due to projection effects and galactic interactions. Often, the objective of astronomical research involves measuring the properties of isolated celestial bodies. The methods used therein frequently make strong assumptions about the isolation of the object. In such cases, galactic blends can be severe enough to warrant discarding the images. Current surveys such as DES (Dark Energy Survey) [2], KiDs (Kilo-Degree Survey) [18] and HSC (Hyper Suprime-Cam) [71] as well as future surveys such as LSST (Large Synoptic Survey Telescope) [62], Euclid [5] and WFIRST (Wide-Field Infrared Survey Telescope) [31] will generate immense quantities of imaging data within the next decade. Thus, a robust and high-throughput solution to galaxy image deblending is vital for drawing maximal value from near-future surveys.

This issue is especially pressing given the details of LSST, for example. Beginning full operation in 2023, LSST is expected to retrieve 15 PB of image data over its scheduled 10-year operation with a limiting magnitude of $i \approx 27$. In the densest regions of the sky (100 galaxies per square arc-minute), up to half of all images are blended with center-to-center distance of 3 arc-seconds, with one-tenth of all galaxy images blended with center-to-center distance of 1 arc-second. Even considering the

most typical regions of sky (37 galaxies per square arc-minute), simulation estimates suggest around 20 percent of galaxies will be superimposed at 3 arc-second separation and up to 5 percent at 1 arc-second separation. [13]. At 1.5 PB per year and image sizes of approximately 6 GBs, conservative estimates predict the LSST capable of producing 200,000 wide-field images. This results in 1 Billion postage stamp galaxy images per year [62]. Improved handling of blended images can thus be expected to save up to 200 Million galaxy images from being discarded each year. 50 Million of these galaxy images are predicted to be blended with 1 arc-second center-to-center separation, posing an extremely difficult task for deblending algorithms. Moreover, detecting blends in the first place will pose an additional challenge for LSST (and other near-future surveys), especially for galaxies blended at 1 arc-second separation and a median seeing of $\sim$0.67 arc-seconds.

Pioneering work in deblending can be traced back to Jarvis & Tyson's "Faint Object Classification and Analysis System" [46] and Irwin's "Automatic Analysis of Crowded Fields" [44]. In Jarvis & Tyson's FOCAS algorithm, objects are initially identified and segmented by comparing the local flux density to that of the average sky density computed in a 5-by-5 pixel box filter. Blended sources are then separated by scanning the principle axis of the object centroid for a multi-modal intensity signal; if found, objects are deblended by drawing a boundary perpendicular to the principle axis at the point of minimum intensity. Meanwhile, in Irwin's analysis, a maximum likelihood parameter estimation scheme was used to segment stars near the center of globular clusters. The technique iteratively estimates the local sky background then di-

vides the field into images and analyzes each image for blends, estimates source position and shape, updates the background estimate and repeats. Irwin's method offered great progress in regions of high number density where the majority of images overlap even at high isophotes.

Since Jarvis & Tyson and Irwin, many wide-field image processing algorithms have been proposed including NEXT [6] and the widely used SEXTRACTOR [10]. NEXT was among the first wide-field image processing methods to utilize artificial neural networks, approaching the problem of detection as one of clustering then employing a modified version of the FOCAS algorithm to deblend superimposed objects. First, a neural network compresses windows of the input image into a dense vector, performing a non-linear principal component analysis. These representations are passed through a second network which classifies the central pixel in the window as belonging to an object or the background. Neighboring object pixels are grouped together and parameters such as the photometric barycenter and principal axis are computed for each contiguous cluster. Building upon the FOCAS algorithm, NEXT then searches for multiple peaks in the light distribution along the principal axis as well as five other axes rotated by up to π radians from the principal axis. When two peaks are found in the distribution, the objects are split along a line perpendicular to the axis joining the peaks. In this way, the NEXT deblending algorithm is an extension of FOCAS in which the assumption that the multi-peaked light distribution will occur along the principal axis is relaxed.

Modern approaches to deblending include gradient and interpolation-based deblending (GAIN) [107], morpho-spectral component analysis (MUSCADET) [48] and

constrained matrix factorization (SCARLET) [66]. In both MUSCADET and SCARLET, each astronomical scene is assumed to be the composition of two non-negative matrices which encode the spectral energy distribution and spatial shape information of a finite number of components which sum to represent the scene. Though the techniques share many similarities, the more recent approach of SCARLET can be viewed as a generalization of MUSCADET which allows any number of constraints to be placed on each source. Meanwhile, GAIN acts as a secondary deblender to repair flawed images by making use of image intensity gradient information and interpolating the missing flux from background sources after the foreground object is segmented. While GAIN, MUSCADET and SCARLET appear to be powerful deblending algorithms in their own right, here we present a method inspired by promising results in computer vision and deep learning.

Deep convolutional neural networks have proven highly effective at classifying images of galaxies [19]. Moreover, their use as generative models for galaxy images has yielded impressive results. In particular, generative adversarial networks (GANs) with deep convolutional generators are able to model realistic galaxies by learning an approximate mapping from a latent space to the distribution over galaxy images given large datasets such as those provided by Galaxy Zoo [29, 61].

Here, we introduce an algorithm which offers progress on both the task of deblending conspicuously blended galaxies as well as the challenge of fast, end-to-end inference of galaxy layers: a branched generative adversarial network. GANs offer an elegant solution to the problem of deblending by allowing us to combine the standard content loss function of a convolutional neural network (e.g. mean squared error) with

the adversarial loss signal of a second "discriminator" network which pushes solutions to the galaxy image manifold and thereby ensures that our deblending predictions appear to be galaxy-like in a probabilistic sense.

After training our deblending network, a forward pass of input to deblended outputs takes a trivial amount of time on a modern laptop. We focus on galaxy pairs which are known to be blended though future work may involve the identification of blends and inference of the number of galaxies involved. With this work, we intend to make usable hundreds of millions more galaxy images, while saving time and effort for astronomers to spend on other pertinent questions.

3.3 Methods

Generative adversarial networks (GANs) are a family of unsupervised learning models used to learn a map between two probability distributions [33, 32]. Often, these probability distributions live in high-dimensional space. For example, images are realizations of a high-dimensional joint probability distribution over the pixel intensities. Instead of positing a parametric family of functions and using maximum likelihood methods to determine an analytic form for these complex distributions, GANs use an artificial neural network to learn the map between a known (hereafter latent) probability distribution and the target data distribution. This map is called the generator. The generator is trained by a second network which learns to discriminate between samples from the target distribution and samples from the generator's approximate distribution.

This second network is called the discriminator. The generator and discriminator share a loss function in which the discriminator aims to maximize its ability to discriminate between true and generated samples while the generator aims to maximize its ability to generate samples indistinguishable from the true distribution samples and therefore "trick" the discriminator. This back-and-forth training can be viewed as a zero-sum game between neural networks where the goal is to reach Nash equilibrium [33]. It is important to note that the generator relies upon the discriminator's signal (and hence its accuracy) to adjust its parameters and increase the quality of its generated outputs. For this reason, it is crucial that both networks train in tandem with one network not overpowering the other.

In the problem of deblending images of galaxies, we treat the joint distribution over pixel intensities of the blended images as the latent distribution. Samples from the latent distribution are denoted I^{BL}. The generator's purpose is to map samples from this latent distribution to corresponding samples in the deblended galaxy image distribution. Samples from the target distribution are denoted I^{PB}, preblended galaxy images. Our goal is to train a generator using a training set of blended images (see Data section below), I^{BL}, and their corresponding original images, I^{PB}. The discriminator provides gradient information to the generator to ensure that the generated deblended images lie on the natural image manifold. The training of the generator is also guided with supervision in the form of a pixel-wise mean squared error (see Loss section below) which further coaches the generator to closely reproduce explicit ground truth galaxy images used to make an artificial blend. After training, the generator is capable of

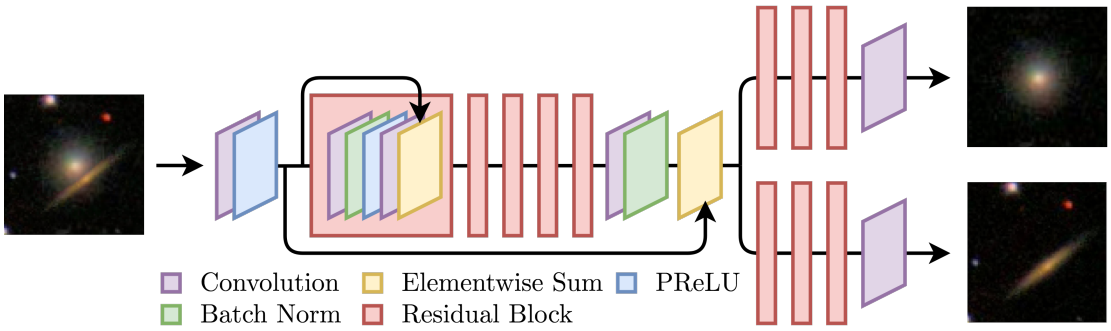


Figure 3.1: Branched residual network architecture for our deblender GAN. The number of residual blocks in the "root" and "branches" are integer-valued hyperparameters we denote M and N, respectively. Larger values of M and N generally produce higher quality results at the cost of training time. Here, we show (M, N) = (5, 3) as an example though the trained deblender herein used (M, N) = (10, 6). Although all residual blocks share a common structure and skip connection, here we only explicitly show the inner-workings of the first.

inferring the unseen independent layers of galaxy images with high fidelity (see Results section below).

3.3.1 Architecture

The architecture of our GAN is a modified version of the successful SRGAN (super resolution GAN) architecture [59]. Our generator is a very deep, branched residual network [39] consisting of a large number of residual blocks and skip-connections (see Figure 3.1). We reference the generator as G and denote the learnable parameters of the generator network as θ_G. Residual networks (ResNets) have performed

impressively on a variety of image detection and segmentation problems. The ResNet architecture alleviates the vanishing gradient problem by including a large number of skip-connections wherein gradient information can flow undegraded to shallow network layers, allowing for very deep network architectures. The building blocks of residual networks are residual blocks: a standard convolution with kernel size $k = [3, 3]$ followed by batch normalization [43], activation, another convolution, batch normalization and finally an elementwise sum with the original input of the block. We refer to the first M residual blocks as the "root" layers whereafter the network splits into two "branches" with N residual blocks each. For the results presented herein, we have chosen (M, N) = (10, 6). Larger values of (M, N) are generally preferable though training times can be prohibitively expensive for large values of either. For activation functions we've chosen parametric rectified linear units (PReLU), an adaptation of leaky ReLU wherein the slope parameter α becomes a learnable parameter [40].

The discriminator is a deep convolutional neural network consisting of many convolutional blocks (see Figure 3.2). A single convolutional block consists of a convolution with kernel size $k = [3, 3]$ and unit stride followed by batch normalization and an activation layer; we employ standard leaky ReLU activations throughout with $\alpha = 0.2$. We reference the discriminator as D and denote the learnable parameters of the discriminator as θ_D.

For network regularization, we rely on our batch normalization layers in place of common alternatives such as dropout. In essence, batch normalization layers regularize by multiplying the hidden units by a random value (one over the batch standard

deviation) and subtracting a random value (the batch mean). Since these values are different for each mini-batch, the layers must learn to be robust to added noise.

We used TensorFlow [1] to build and run our computational graph; training was executed on a single Nvidia GTX 1080 Ti Founder's Edition video card with 11 GB of VRAM and 3584 Nvidia CUDA cores.

3.3.2 Loss

The loss function of the generator is made of three components. As in the SRGAN prescription, these losses can be broken up into two classes: (i) the adversarial loss and (ii) the content loss.

3.3.2.1 Adversarial Loss

The adversarial loss is based on the discriminator's ability to differentiate the generator's samples from true data samples. In essence, this part of the loss ensures the generator's samples lie on the galaxy image manifold (i.e. they look convincingly like galaxies). We define p_{data} and p_z as the probability distributions over the pre-blended and blended image sets, respectively, wherein x is a single pre-blended image and z is a single blended image.

$$l_{ADV} = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{3.1}$$

$$\hat{\theta}_G = \arg\min_G l_{ADV}(D, G) \tag{3.2}$$

$$\hat{\theta}_D = \arg\max_D l_{ADV}(D, G) \tag{3.3}$$

In deblending use cases, this loss alone is insufficient since we aim to faithfully deblend an image into its exact components (not simply any two images that look like galaxy images). To ensure this, we employ content losses.

3.3.2.2 Content Loss

We compute the pixel-wise mean squared error (MSE) between the generator's sample (I^{DB}) and the corresponding ground truth images (I^{PB}). This is equivalent to maximizing the effective log-likelihood of the data (in this case pixel intensities) conditional upon the network parameters given a Gaussian likelihood function. This is justified by a combination of the central limit theorem and the observation that pixel intensities arise from a very large amount of uncertain physical factors. Thus, for images of pixel width W and pixel height H, the pixel-wise mean squared error is given by the following.

$$l_{MSE} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \left[I^{PB}_{x,y} - G_{\theta_G}(I^{BL})_{x,y} \right]^2 \tag{3.4}$$

The second component of the content loss uses deep feature maps of the pre-trained VGG19 19-layer convolutional neural network [96]. We chose the activations of the deepest convolutional layer (corresponding to the fourteenth overall layer

of VGG19). The ground truth and generator output images are passed through the VGG19 network and their corresponding feature maps at the last convolutional layer are extracted. We compute the pixel-wise mean squared error between these feature maps and include this in the total generator loss function. For feature maps of pixel width W and pixel height H, the VGG mean squared error loss is computed as follows.

$$l_{VGG} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \left[VGG_{14}(I^{PB})_{x,y} - VGG_{14}(G_{\theta_G}(I^{BL}))_{x,y} \right]^2 \tag{3.5}$$

In the original SRGAN implementation, the VGG feature maps were used as a content loss which resulted in perceptually satisfying images at the cost of exact similarity to the ground truth image. For our purposes, we employ the VGG loss as a "burn-in" loss function to break out of otherwise difficult to escape local minima. In the case of galaxy images, these local minima resulted in entirely black images for the first few hundred thousand batches. Using the VGG loss, we were able to escape the local minima in fewer than a thousand batches. To explain this improvement, note that entirely black images share much in common with images of galaxies in the dark surrounding sky, but those same flat dark images have no meaningful structural features. Because the VGG loss sets the target as deep layers of a CNN trained for image processing tasks, it effectively measures and penalizes the lack of pattern found in the predominately black images.

Following the SRGAN prescription, we scale the adversarial loss by 10^{-3} to approximately match its magnitude to the mean squared error content loss. We used a

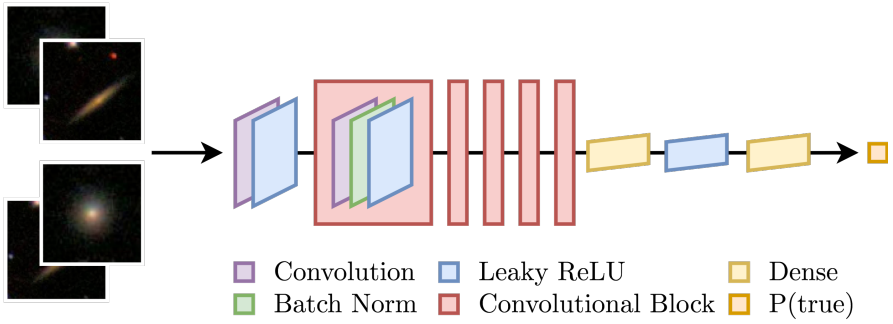


Figure 3.2: Deep convolutional network architecture for the deblender GAN discriminator. All convolutions use a kernel size of (3, 3) whereas strides alternate between (2, 2) and (1, 1). The number of convolutional blocks is a hyperparameter though it is constrained by the requirement that images maintain integer dimensions given that non-unity convolutional strides reduce this dimensionality by a factor of the stride assuming zero padding. Although all convolutional blocks share a common structure, here we only explicitly show the inner-workings of the first.

discriminator-to-generator training ratio of one: for each iteration, both the discriminator's parameters and the generator's parameters were updated once. It should be noted that we have to compute the MSE and adversarial losses twice—once for each image output of the generator. The total loss is the average of the losses computed for each output image.

3.3.3 Data

As noted in [17], large near-future surveys such as LSST will encounter three classes of blends: *ambiguous* blends, *conspicuous* blends and *innocuous* blends. Here,

we focus primarily on the conspicuous (confirmed blends) and innocuous (near blends) classes. Future work may include classifying ambiguous blends.

We used 141,553 images available from Galaxy Zoo [61] via the Kaggle Galaxy Zoo classification challenge. These are galaxy images from the sixth data release [4] of the Sloan Digital Sky Survey (SDSS). SDSS is a survey covering nearly 26 percent of the sky and capturing photometric data in five filters, though the images used here were composite images of the g, r and i bands. Each image is of standard (i.e. non-extended) JPEG format and therefore has a bit depth of 8 bits per color channel. According to [61], the images were cropped such that the resolution was $0.024R_p$ arc-second per pixel with R_p the Petrosian radius of the system.

Each image is originally of dimension (H, W, C) = (424, 424, 3) where H, W and C are the height, width and channel dimensions, respectively. We first crop the center of these images to dimension (H, W, C) = (240, 240, 3) and then downsample them using a bicubic sharpener to (H, W, C) = (80, 80, 3). Leaving one image centered, we then perturb the second image by flipping it both horizontally or vertically according to a Bernoulli distribution with $p = 0.5$, rotating uniformly on the interval $\theta \in [0, 2\pi]$, displacing both horizontally or vertically uniformly on the interval $dx, dy \in [10, 50]$ pixels and finally scaling log-uniformly on the interval $r \in [1/e, \sqrt{e}]$ where r is the scale ratio. These images serve as our pre-blended ground truth images (I^{PB}); we scale their pixel values to the interval $I \in [-1, 1]$.

The images are then blended. We select two blending schemes for which we train and evaluate on independently: (1) blending by selecting the pixelwise max be-

tween the two images (2) blending by computing the pixelwise mean between the two images. In general, true galaxy blends include some combination of the two: predominantly mean blending with min/max blending where dust and gas occlude light from the background galaxy, for example.

Each blending scheme offers a unique set of challenges for our network. For example, in pixelwise max blending, all pixel information from the occluded galaxy is lost and the network must infer it from the surrounding image. On the other hand, pixelwise max blended images often display hard edges or lines that may make segmentation of the foreground galaxy substantially easier.

In pixelwise mean blending, each pixel contains information relevant to both components. In principle, this should make it easier to infer the properties of the occluded galaxy. However, the blended images are generally more diffuse and blurry than pixelwise max blending which makes segmentation of the foreground galaxy difficult.

Our image blending prescriptions were inspired in part by the catalog of true blended Galaxy Zoo images presented in [50]. We find that random samples from our datasets closely resemble blends from this catalog. We scale the blended image pixel values to the interval $I \in [0, 1]$ as recommended in the SRGAN paper. These blended images are realizations of the latent distribution of images (I^{BL}).

Note that it is a necessary requirement for our method that the image is centered upon one galaxy; this spatial information informs the independent behavior of each branch. During training, the network learns that one branch corresponds to the deblended central galaxy, while the other branch corresponds to the deblended off-center

galaxy. Without such a feature (which should always be possible in true blends with a simple image crop), the network may be confounded by which branch to assign to which galaxy, and for example create the same deblended galaxy image in each branch. By using a principled centering scheme, this problem is largely diminished. This scheme should be easily adaptable to more than two layers by deblending the center galaxy from all other galaxies, then sending the deblended output back into the generator.

3.3.4 Training

During training, the generator network is presented with a batch of blended inputs (I^{BL}); we use a batch size of 16 samples. The generator predicts the deblended components (I^{DB}) of each blended input. The discriminator is given both a batch of generated deblended images and their corresponding preblended ground truth images. The discriminator updates its parameters to maximize its discriminative ability between the two distributions. Mathematically speaking, minimization of the adversarial loss is equivalent to minimization of the Jenson-Shannon divergence between the generated distribution (I^{DB}) and the data distribution (I^{PB}) whereas minimization of the mean squared error term equates to minimization of an effective negative Gaussian likelihood over pixel intensities. In reality, the generator is trained to minimize the sum of these components by averaging the gradient signals from each. This is done via stochastic gradient descent on the network's high-dimensional error manifold defined by its compound loss function with respect to its learnable network parameters.

Both the generator and discriminator loss functions are optimized in high-

dimensional neural network weight space via the Adam optimizer (adaptive moment estimation) [51], a method for stochastic optimization which combines classical momentum with RMSProp. With momentum, Adam uses past gradients to influence current gradient-based steps, accelerating learning in valleys of the error manifold. Following the running-average style of RMSProp, it emphasizes the most recent gradients while decaying influence from much earlier ones, hence adapting better to the present loss surface curvature.

The steps down the error manifold are further scaled by a hyperparameter referred to as the learning rate (although the true rate is also naturally scaled by adaptive momentum). We choose an initial learning rate of 10^{-4}. After 100,000 update iterations (1.6 million individual images or five epochs), we decrease the learning rate by an order of magnitude to 10^{-5}. We then train for another 100,000 update iterations after which the network parameters are saved and can be used for inference on unseen images.

3.4 Results

To evaluate our model, we infer upon blended images withheld from the training dataset. For these blends, we have access to the to the true preblended images and therefore are able to quote relevant metrics for image comparison (see below). It should be noted that the galaxies which make up the testing set blended images have also been withheld in the creation of the training set. That is, the deblender GAN has never seen this particular blend nor the underlying individual

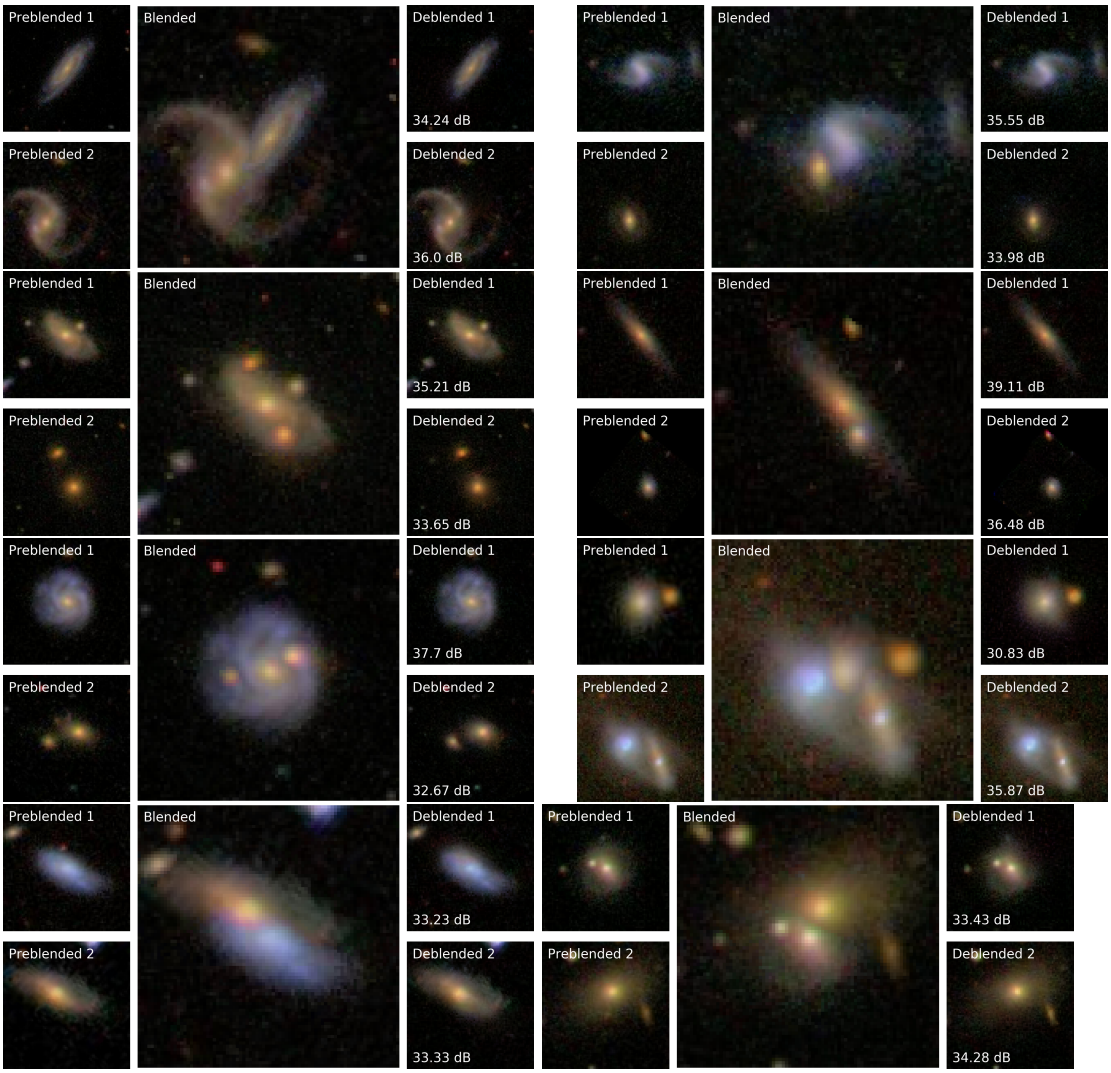


Figure 3.3: A selection of successful deblender GAN predictions on pixelwise maximum blended images. On the left side of each panel are two preblended Galaxy Zoo images (I^{PB}) which were superimposed to create the central blended input image. The trained generator's deblended predictions (I^{DB}) are on the right of each panel. Superimposed upon the deblended predictions are the associated PSNR scores between the deblended image and its related preblended image. A variety of galaxy morphologies are represented here—in each case, our deblender GAN successfully segments and deblends the foreground galaxy while imputing the most likely flux for the occluded pixels in the background galaxy.

galaxies during training. A selection of results are presented in Figure 3.3 and Figure 3.6 with a more extensive catalog of predictions in a GitHub repository available at https://github.com/davidreiman/deblender-gan-images.

We select two image comparison metrics for determining the quality of the deblended images in relation to the ground truth images: the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM).

The peak signal-to-noise ratio is often used as an evaluation metric for compression algorithms. It is a logarithmic measure of the similarity between two images, in our case the ground truth preblended image (I^{PB}) and the deblended re-creation (I^{DB}). The equation for PSNR is written in terms of the maximum pixel intensity (MAX) of the truth image and the mean squared error (MSE) between the test and ground truth images.

$$\text{PSNR} = 20 \log_{10}(\text{MAX}) - 10 \log_{10}(\text{MSE}) \tag{3.6}$$

The structural similarity index [105] is a method for evaluating the perceptual quality of an image in relation to an unaltered, true image. It was invented as an alternative to PSNR and MSE methods which measure error. Instead, SSIM measures the structural information in an image where structural information is understood in terms of pixel-wise correlations. SSIM scores are measured via sliding windows of user-defined width in which the means (μ_x, μ_y), variances (σ_x^2, σ_y^2) and covariance (σ_{xy}) of pixel intensities are computed in the same windowed region of each image being

compared, the window in one image denoted x and the other denoted y. Included are the constants c_1 and c_2 which stabilize the division to avoid undefined scores—here we use $c_1 = 0.01$ and $c_2 = 0.03$. A single SSIM score between two images is the mean of all windowed SSIM values across the extent of the images.

$$\text{SSIM} = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3.7}$$

We tested on 3,200 blends withheld from the training set. This corresponds to 6,400 individual galaxy images. We computed 6,400 PSNR and SSIM scores and quote their mean, median, minimum, maximum and variance in Table 3.1. Their full distributions are displayed in Figure 3.4 and Figure 3.5.

The deblender GAN performs impressively with respect to both metrics. For reference, values of PSNR above 30 dB are considered acceptable for lossy image compression. Considering that these images were completely reconstructed out of a blended image (rather than compressed) the scores set a high benchmark for deblending. It should be noted, however, that comparing these PSNR and SSIM scores directly to those of natural images is likely in error. Many galaxy images are dominated by a black background which can bias both the mean squared error and pixelwise correlations upward and thereby inflate the resulting PSNR and SSIM scores. Still, we quote our scores as a benchmark for comparison with future iterations of our own model and collation with alternative deblending algorithms working with a similar dataset and image format.

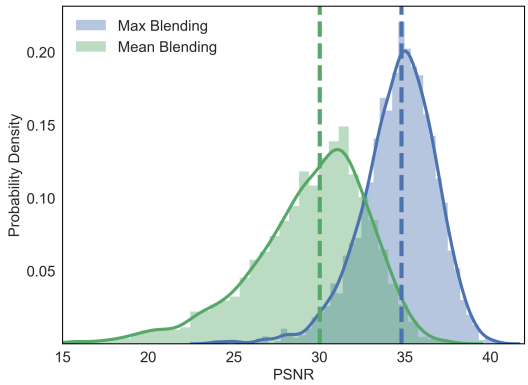


Figure 3.4: Distribution of peak signal-to-noise ratio (PSNR) scores between deblended images and pre-blended images for each blending scheme.

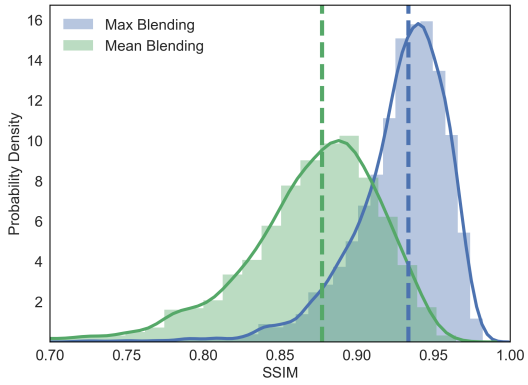


Figure 3.5: Distribution of structural similarity (SSIM) scores between deblended images and pre-blended images for each blending method.

Imaging Metrics	Mean	Median	Max	Min	Var
PSNR (dB)	34.61	34.83	40.45	21.34	4.88
SSIM	0.928	0.934	0.982	0.576	0.001

Table 3.1: Summary statistics of the PSNR and SSIM score distributions. Maximum PSNR and SSIM scores are comparable with high-quality compression algorithms (PSNR $\in [30, 50]$ dB).

3.5 Discussion

We've shown promising initial results for the application of a novel generative adversarial network architecture to the problem of galaxy deblending. Deep generative models are natural solutions to deblending for their ability to infill missing pixel values behind blends using information learned about the distribution of natural images provided during training. The discriminative loss ensures that all deblended images lie on the natural image manifold of galaxies. In addition, the branched structure allows our model to identify separate galaxies and segment them accordingly without human intervention or labeling; our approach only requires that one of the galaxies lie at the center of the image. Our branched GAN naturally handles the large incoming quantities of data from surveys like LSST and DES, deblending images near-instantaneously.

We've used the peak signal to noise ratio (PSNR) and structural similarity index (SSIM) as a metrics of quality in the segmentation. Though we've found no other applications of deep learning to the problem of deblending, we have quoted our PSNR

and SSIM scores here as a benchmark for future comparisons with our own branched GAN revisions and with the alternative deblending algorithms.

We encountered a variety of issues while training our branched deblender GAN. Most notable of these is the tendency to fall into local minima near the start of training. Since galaxy images are largely black space the GAN learns that it can initially trick the discriminator by making all black images. We broke out of the minima by including a "burn-in" period of a few thousand batches using only a mean squared error loss between the deep activations of the pre-trained VGG19 network. Another solution to this issue may be to pre-train the discriminator though there exists the possibility of mode collapse wherein the generator and discriminator performance are highly unequal and learning ceases.

There is great room for improvement in our branched GAN. Our blending schema was chosen to match the curated catalog of true blends in the Galaxy Zoo catalog presented in [50] though it certainly could be improved upon. In truth, galaxy blends consist of a combination of pixelwise sum (generally in more diffuse regions) and pixelwise max/min (generally where dense regions of the foreground galaxy obstructs the background galaxy flux). Pure pixelwise max blending tends to create unrealistic blended images where low flux foreground galaxies fall upon bright regions of the background galaxy—this generally leads to blends with artificially incomplete foreground galaxies which have been "cutoff" by the high flux pixels of the background images. Images of this type bias our estimates of both PSNR and SSIM scores. Indeed, pixelwise max blending is in some sense the worst case for deblending wherein precisely zero

information about the background galaxy is encoded in the pixel intensities in regions where the foreground galaxy eclipses it. On the other hand, a variety of generated blends exhibit artificially hard lines between overlapping galaxies which likely makes the galaxies easier to deblend. Moreover, the presence of unmasked background galaxies in the Galaxy Zoo images give rise to artificial penalties in the mean squared error loss function when they are assigned to the incorrect deblended galaxy image. We also note a slight misalignment of the RGB channels of the Galaxy Zoo images which gives rise to unnatural color gradients—a feature that our network learned to reproduce. We propose to address issues in our blending schema and issues in data creation in general in future work (see below).

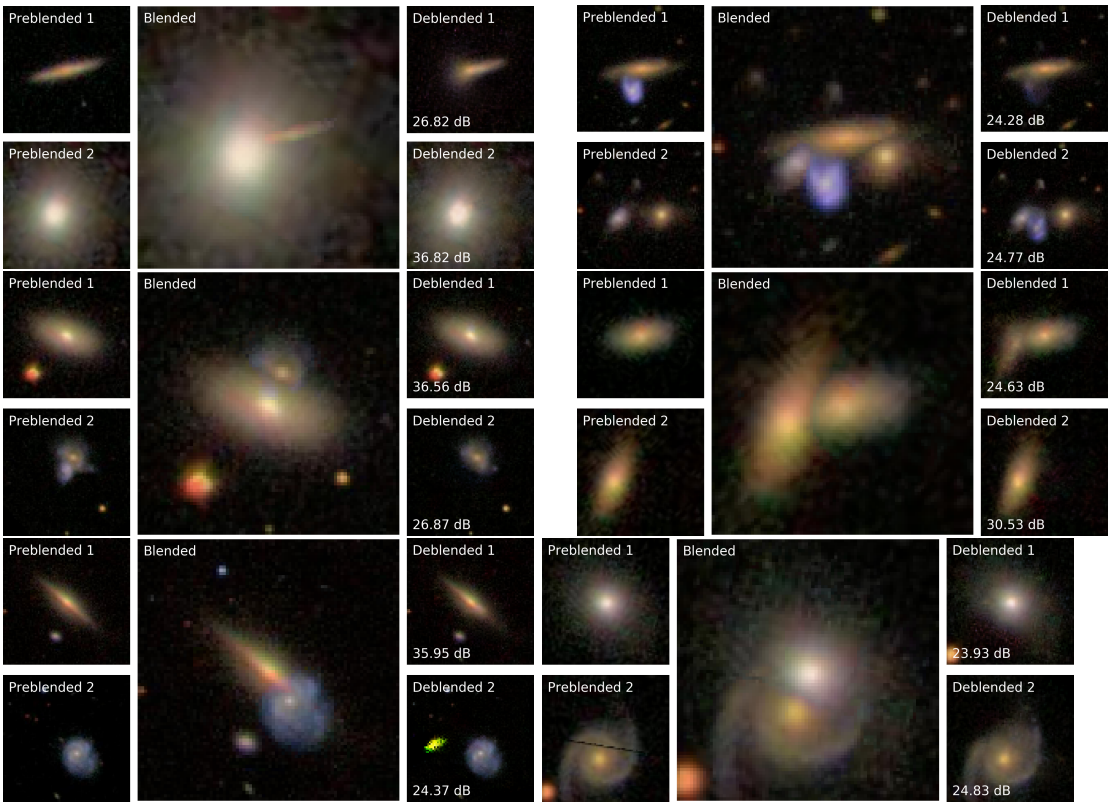


Figure 3.6: A selection of failed deblender GAN predictions. There exists a variety of failure modes in the deblender GAN predictions: (1) incomplete deblends wherein artifacts of one galaxy are left in the prediction for the other galaxy. (2) Incomplete galaxy generation due to our blending prescription. In selecting the pixelwise maximum value during the blending process, foreground galaxies that eclipse bright regions like the core of the background galaxy often are non-realistically "cutoff" by the overwhelming background flux. This leaves our network very little with which to predict from and often results in predictions that appear to be only a part of the associated preblended galaxy. (3) Associating background galaxies with the incorrect blended galaxy. This negatively biases the PSNR scores for otherwise acceptable predictions—a representative example is given in the last image of the above figure. (4) Deblending artifacts: in a handful of predictions, we notice random artifacts of unnatural color such as the green aberration in the second-to-last image above; a possible explanation being that a bright celestial object in the same relative position to the galaxy appears in neighboring galaxy images of the galaxy image manifold.

The deblender model presented herein would likely benefit from multiband input data spanning a larger range of the electromagnetic spectrum. Deep neural networks have proven powerful tools for estimating photometric redshifts from multiband galaxy images alone with no manual feature extraction [85]. The ability to learn approximate photometric redshift estimates from input images would likely strengthen the deblending performance of our network. To this end, future iterations of our deblender GAN will utilize multiband inputs.

Moreover, our model is restricted to inference on solely nearby (low-redshift) galaxies. In deep learning, we assume that examples from the training and testing sets are drawn from the same distribution. For our deblender, this limits feasible test images to those composed of galaxies from the same redshift and mass bin as the training images. It is commonly understood that galaxy morphology evolves with redshift and therefore galaxy images at high-redshift will look largely different than nearby galaxies [3], i.e. they belong to a different image manifold. Though the galaxy distributions are distinct, the distribution of pixel intensities in the images share a great deal of similarities such as dark backgrounds populated by luminous, diffuse objects. This observation makes it an area of interest apt for transfer learning which is a topic we will explore in future work.

Also, in future work we will apply our branched deblender GAN to data generated from GalSim [93] in which we can robustly blend any number of galaxies and apply a variety of PSFs. With this, we hope to generalize our results to galaxy images from a wide variety of potential surveys and thereby salvage an increasing number of blended

images which would otherwise be unused in the study of galaxies and galaxy evolution. Moreover, we plan to implement a variable number of GAN branches, allowing any blended object to be deblended after estimating the number of objects involved in the blend. An ultimate goal is the full deep learning pipeline from source identification to blending classification and deblending.

Acknowledgements

The authors would like to thank Shawfeng Dong for offering his compute resources to train and test our model. We would also like to thank Joel Primack and Doug Hellinger for their helpful feedback on a draft of this paper. In addition, we thank David Koo, Sandra Faber, Tesla Jeltema, and Spencer Everett for many insightful discussions and recommendations.

The authors would also like to acknowledge the Sloan Digital Sky Survey for providing the data used herein.

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/.

Finally, we'd like to thank the Galaxy Zoo team for their efforts in providing a neatly curated dataset of SDSS galaxy images to the astronomy community.

Part III

Fully probabilistic quasar continua predictions near Lyman-α with conditional neural spline flows

Chapter 4

Manuscript

4.1 Abstract

Measurement of the red damping wing of neutral hydrogen in quasar spectra provides a probe of the epoch of reionization in the early Universe. Such quantification requires precise and unbiased estimates of the intrinsic continua near Lyman-α (Lyα), a challenging task given the highly variable Lyα emission profiles of quasars. Here, we introduce a fully probabilistic approach to intrinsic continua prediction. We frame the problem as a conditional density estimation task and explicitly model the distribution over plausible blue-side continua ($1190\,\text{Å} \leq \lambda_{\text{rest}} < 1290\,\text{Å}$) conditional on the red-side spectrum ($1290\,\text{Å} \leq \lambda_{\text{rest}} < 2900\,\text{Å}$) using *normalizing flows*. Our approach achieves state-of-the-art precision and accuracy, allows for sampling one thousand plausible continua in less than a tenth of a second, and can natively provide confidence intervals on the blue-side continua via Monte Carlo sampling. We measure the damping wing

effect in two $z > 7$ quasars and estimate the volume-averaged neutral fraction of hydrogen from each, finding $\bar{x}_{\rm HI} = 0.304 \pm 0.042$ for ULAS J1120+0641 ($z = 7.09$) and $\bar{x}_{\rm HI} = 0.384 \pm 0.133$ for ULAS J1342+0928 ($z = 7.54$).

4.2 Introduction

Prior to the emergence of the first luminous sources, the intergalactic medium (IGM) was filled with a dense gas of neutral hydrogen. During the epoch of reionization, nascent stars, galaxies and quasars ionized the surrounding IGM with ultraviolet (UV) radiation. Reionization induced a patchy topology upon the universe wherein ionization bubbles grew about each source until the ionization regions merged and percolated, and residual neutral hydrogen was left primarily in the deep potential wells of dark matter halos. Recent measurements of the cosmic microwave background suggest that reionization of the IGM occurred at $z \sim 8$ [90], but there remains uncertainty about its precise timing and nature—constraining the history of reionization is a goal of modern cosmology.

A largely neutral IGM leaves marked signatures in source spectra. At redshifts beyond $z \approx 6$, the spectra of sources sufficiently luminous to be observed with modern telescopes exhibit a near-complete suppression of flux blueward of Lyα—an effect known as the Gunn-Peterson trough [37]. In comparison, at lower redshifts where the residual neutral hydrogen has gathered in the potential wells of dark matter halos and other large-scale structures, we observe discrete absorption features (the Lyα forest) in source

spectra that trace out the matter field of the universe [70, 41]. These latter observations confirm the presence of a highly ionized IGM.

These considerations suggest that one could probe the epoch of reionization by searching for the presence of the Gunn-Peterson trough in a sequence of luminous sources of increasing redshift. However, the presence of the Gunn-Peterson trough alone is insufficient to prove that the emitting source is surrounded by a neutral IGM [69]—due to the considerable optical depth of the neutral IGM, even a modest residual neutral fraction of hydrogen in a largely reionized IGM will suppress the transmittance of flux blueward of Lyα to near-zero.

Instead, unambiguous evidence of a neutral IGM may be provided by measurement of the *red damping wing* of the Gunn-Peterson trough [69], which is only identifiable for very large column densities of neutral hydrogen such as those at or prior to reionization. This broad absorption feature extends redward of Lyα out to a rest-frame wavelength of $\lambda_{\rm rest} \approx 1260\,\text{Å}$ in the quasar's rest-frame. By measuring the damping wing in a series of high-redshift sources, one can place constraints on the timeline of the early universe IGM phase transition from neutral to reionized. Thus far, the only sources luminous enough to permit measurement of the damping wing are quasars, although there remains optimism that gamma-ray burst afterglows may offer an additional avenue [100].

Measurement of the damping wing is complicated by a variety of factors, notably: (i) the possibility of a potent absorption system along the quasar line-of-sight whose proximity alters the profile of the wing, and (ii) the uncertainty in our estimation

of the intrinsic quasar flux (termed the "continuum") whose profile allows us to measure the damping wing width. As noted by [69], continua prediction is especially difficult when the emitting source is a quasar since "quasars have strong, broad Lyα emission lines with profiles that are highly variable." Thus, a model which predicts intrinsic quasar continua should be expressive and ideally provide calibrated uncertainties with its predictions.

Another complicating factor is estimating the extent of the quasar near-zone. In the proximity of powerful UV sources such as quasars, the increased photo-ionization rate of neutral hydrogen leads to a diminishing number of Lyα absorbers near the redshift of the source. This consequence is known as the line-of-sight proximity effect [7], and its associated spectral feature is denoted the *ionization near-zone*, or simply the proximity zone. Our model of the damping wing relies on our knowledge of the blueward edge of the quasar proximity zone, and though we have heuristics to locate the edge of the near-zone, our uncertainty in the true location affects our estimates of the damping wing strength.

Intrinsic continua prediction has been studied widely and approached in a variety of manners. Previous approaches generally predict the blue-side continua ($1190\,\text{Å} \leq \lambda_{\text{rest}} < 1290\,\text{Å}$) using information encoded in the red-side spectrum ($\lambda_{\text{rest}} \geq 1290\,\text{Å}$). This information arises from correlations in the emission features of the blue and red-side spectra [11]. Such models are typically trained on the spectra of quasars at moderate redshift—a typical redshift range is $z \in [2.1,\ 2.5]$ [36]—which cover the Lyα and Mg II broad emission lines. These lines strongly constrain the standard pipeline estimates

for quasar redshifts [82]. Since such redshift estimates are a major source of bias in any model which aims to impute spectra, selecting a redshift range which includes such emission features limits our exposure to strong systematics. Approaches such as these often involve a pair of models: the *primary model* which infers the blue-side continua given the red-side spectrum, and the *secondary model* which probes the similarity of high-redshift quasars (our targets for inference) to our moderate-redshift training set.

The primary model predicts the intrinsic (i.e. unabsorbed) continua near Lyα given the redward spectrum. Primary models are often fundamentally non-probabilistic (with a notable exception being the fully Bayesian framework of [36]) though many employ methods such as ensembling or prediction on nearest neighbors to approximate confidence intervals during inference. For example, previous approaches have employed principal component analysis (PCA) to the blue and red sides of the spectrum and subsequently related the coefficients using multiple linear regression [97, 16] or an ensemble of neural networks [24].

The secondary model probes the similarity in spectral characteristics between the moderate-z training set and the high-z quasars which are targets for constraining the epoch of reionization. High similarity assures us that the high-z targets are *in-distribution* and therefore valid inputs to our primary model. Here, similarity is tantamount to likelihood though typically simple models or proxies are used. A number of approaches have been chosen in related studies, for example the reconstruction error of an autoencoder [24] or the likelihood of PCA coefficients under a Gaussian mixture model [16].

In this article we introduce SPECTRE, a fully probabilistic approach to intrinsic continua prediction which utilizes normalizing flows as both the primary and secondary models. We frame the problem as one of conditional density estimation: what is the probability distribution over blue-side continua given the redward spectrum? In doing so, SPECTRE achieves state-of-the-art precision, allows for sampling one thousand plausible continua in less than a tenth of a second on modern GPUs, and can natively provide confidence intervals on the blue-side continua via Monte Carlo sampling. In addition, our secondary model provides the likelihood ratio as a background contrastive score used to measure how in-distribution a given quasar continuum lies, offering a new perspective on the probability of high-redshift quasar spectra under a generative model trained on moderate-redshift quasars. Both primary and secondary models are applied to continua from high redshift ($z > 7$) quasars, ULAS J1120+0641 ($z = 7.09$) [73] and ULAS J1342+0928 ($z = 7.54$) [8], in order to infer the neutral hydrogen fraction of the universe during the epoch of reionization.

This manuscript is organized as follows: §4.3 provides a brief overview of previous approaches to intrinsic continua prediction. In §4.4 we introduce normalizing flows, describe their usefulness in scientific applications, and provide a brief introduction to the particular variety we employ in this work: neural spline flows. Then, in §4.5 we detail our approach by first outlining our data preprocessing scheme, describing our flow-based model and explaining our measurement of the damping wing in quasar spectra. Results are presented in §4.6 for our constraints on the reionization history of the early universe from measuring the damping wing in two $z > 7$ quasars. In

§4.7, we conclude with a summary of our work and further discuss the use of flows for probabilistic modeling in the sciences.

4.3 Related Work

Previous approaches to intrinsic quasar continua prediction have ranged from fully Bayesian models operating purely upon emission features [36] to full-spectrum principal component analyses on blue/red-side continua to learn correlations between the dominant modes of variation in each, as in [16]; [24]. In these approaches, the authors fit models on moderate redshift quasars (typically $z \approx 2$) and apply them to quasars near or at reionization ($z \approx 7$).

In the Bayesian approach of [36], the authors construct a covariance matrix describing the correlations of high-ionization emission line features in moderate redshift quasar spectra. The emission line profiles are compressed into three features: line width, peak height and velocity offset. Each of their chosen emission features (Lyα, Si IV/O IV, C IV, and C III) are modeled with either a single or double component Gaussian profile. The Gaussian components are fit to each spectrum in their training set using a Markov Chain Monte Carlo approach with a χ^2 likelihood function. After fitting each element of their training set, the authors compute the correlation matrix between the emission line features for all included lines. For reconstructing the Lyα emission of high-redshift quasars, the red-side emission features are fit in a manner identical to the training set. The likelihood of its parameter vector is modeled via a high-dimensional Gaussian with

mean equal to the mean parameter vector of the training set and covariance equal to the covariance matrix computed from the training set. By doing so, the authors assume that the marginal distributions of each parameter can be described by a Gaussian. Reconstruction of the blue-side continua then proceeds by collapsing the likelihood function along all dimensions corresponding to the parameters of the red-side emission features, leaving only the 6-dimensional conditional likelihood of the Lyα emission: three parameters for each of its two Gaussian components. Here, a prior on the blue-side emission features is introduced by fitting on unabsorbed quasar spectra at $z \approx 6$. The resulting posterior is a joint distribution over these six parameters which provides a probabilistic model for the blue-side intrinsic continua conditioned on the red-side emission features.

Subsequently, the authors published analyses on the damping wing of hydrogen in two available $z > 7$ QSOs [35, 34] using large-scale epoch of reionization simulations. To do this, they sampled $\sim 10^5$ plausible blue-side continua from their model and multiplied each by $\sim 10^5$ synthetic damping wing opacities from their simulations of the epoch of reionization. These $\sim 10^{10}$ mock spectra were compared to the observed spectrum of the high-redshift QSOs using a χ^2 likelihood function. The final estimate of the neutral fraction was then calculated by weighting the neutral fraction of each synthetic damping wing by its marginal likelihood (over all mock spectra) with reference to the observed spectrum and computing the weighted average.

Other work makes use of principal component analysis (PCA) to reduce the dimensionality of the problem. PCA produces a set of linearly uncorrelated PCA eigenvec-

tors. A spectrum can then be reconstructed with a sum of PCA eigenvectors multiplied by the appropriate coefficients. Among techniques which employ PCA, the number of coefficients may be chosen by hand or to satisfy some criterion on the explained variance (e.g. 99%). Correlations between blue and red-side coefficients are then modeled with either a linear model [97, 83, 26, 25, 16] or an ensemble of neural networks [24]. Inference on high-redshift quasars is then performed by encoding the the red-side spectrum into its PCA coefficients and using the trained model to predict the blue-side coefficients. Using the blue-side coefficients and PCA eigenvectors, the blue-side continua can be reconstructed and used as a prediction of the intrinsic continua. The neutral fraction of hydrogen is then estimated using either a simplified model of the red damping wing (as in [24]) or via full hydrodynamical modeling of the neutral IGM (as in [15]).

4.4 Background

This section describes normalizing flows in moderate detail. For an exhaustive introduction and tutorial refer to [79]—we will adopt a similar notation from here onwards: $\mathbf{x}$ is a D-dimensional[1] random vector (e.g. the measured spectrum of a quasar) generated from an underlying distribution $p_x^*(\mathbf{x})$, and $\mathbf{u} \sim p_u(\mathbf{u})$ is the latent (or hidden) representation of $\mathbf{x}$ in a D-dimensional isotropic Gaussian space. The representations are related via a transformation $T\colon \mathbb{R}^D \to \mathbb{R}^D$ such that $\mathbf{x} = T(\mathbf{u})$. We model the true distribution over $\mathbf{x}$ via a distribution $p_x(\mathbf{x};\ \boldsymbol{\theta})$ produced by a neural network with

[1]The dimensionality D is defined by the data—for spectroscopy, D is determined by the resolution and wavelength range of the spectroscope, although D may change due to subsequent preprocessing.

parameters $\boldsymbol{\theta}$.

4.4.1 Normalizing Flows

Normalizing flows model complex probability densities by mapping samples between a base density and the distribution of interest, i.e. the data distribution. The transformation, T, is composed of a series of invertible mappings T_i, or *bijections*, each of which must be differentiable. The base density is commonly chosen to be Gaussian, with the requirement that its dimensionality be equal to the dimensionality of the data to maintain invertibility. Since the composition of differentiable, invertible mappings $T = T_0 \circ T_1 \circ ... \circ T_N$ is itself differentiable and invertible, the normalizing flow acts as a diffeomorphism between the data space and the Gaussian latent space. In our application, the flow then learns a bijective mapping between Gaussian-distributed latent samples and blue-side quasar continua (conditional on the red-side spectrum).

The density of a random vector in the data space is then well-defined and easily computed via a change of variables—we simply cast our data into the Gaussian latent space where density evaluation is trivial. For a datum $\mathbf{x}$ drawn from a D-dimensional target distribution p_x^*, we can model its density in the following manner:

$$p_x(\mathbf{x}) = p_u(\mathbf{u}) \left|\det J_T(\mathbf{u})\right|^{-1} \tag{4.1}$$

where $\mathbf{x} = T(\mathbf{u})$, p_u is the base density and J_T is the Jacobian of the transformation T which maps samples from the Gaussian space to the data space.

Sampling is similarly uncomplicated: latent vectors are sampled in the Gaus-

sian space and transformed into data space samples via T. Quantities such as confidence intervals can then easily be computed via Monte Carlo sampling and are generally calculable from a single (large batch) forward pass provided the data dimensionality and model size are modest.

In practice, we employ neural networks to parameterize our invertible transformations T_i (which together compose T) and cleverly choose the form of the transformations such that the Jacobian is lower triangular. The neural networks themselves are parameterized by a parameter vector $\boldsymbol{\theta}$. Since the determinant of a lower triangular matrix is simply the product of its diagonal elements, this reduces the complexity of the determinant calculation from $\mathcal{O}(D^3)$ to $\mathcal{O}(D)$.

To train such a model, we minimize the divergence between the target distribution $p_x^*(\mathbf{x})$ and the distribution parameterized by the flow, $p_x(\mathbf{x};\boldsymbol{\theta})$. A common choice of divergence is the Kullback-Leibler (KL) divergence which measures the loss of information when using the model distribution to estimate the target distribution.

$$\begin{aligned}
\mathcal{D}_{KL}\left[p_x^*(\mathbf{x}) \parallel p_x(\mathbf{x};\boldsymbol{\theta})\right] &= \mathbb{E}_{p_x^*(\mathbf{x})}\Big[\log \frac{p_x^*(\mathbf{x})}{p_x(\mathbf{x};\boldsymbol{\theta})}\Big] && (4.2)\\
&= \mathbb{E}_{p_x^*(\mathbf{x})}\Big[\log p_x^*(\mathbf{x}) - \log p_x(\mathbf{x};\boldsymbol{\theta})\Big] && (4.3)\\
&= -\mathbb{E}_{p_x^*(\mathbf{x})}\Big[\log p_x(\mathbf{x};\boldsymbol{\theta})\Big] + \text{const.} && (4.4)\\
& && (4.5)
\end{aligned}$$

where $\mathcal{D}_{KL}$ is the KL-divergence and $\mathbb{E}_{p_x^*(\mathbf{x})}$ denotes the expectation with respect to the target distribution $p_x^*(\mathbf{x})$.

We can then identify an appropriate loss function for training by writing the model density in terms of the flow transformation and its Jacobian, using the fact that the determinant of the inverse of an invertible transformation is the inverse of the determinant of the transformation and recalling that $\mathbf{u} = T^{-1}(\mathbf{x})$.

$$\mathcal{L}(\boldsymbol{\theta}) = -\mathbb{E}_{p_x^*(\mathbf{x})}\Big[\log p_u(T^{-1}(\mathbf{x};\boldsymbol{\theta})) + \log\left|\det J_{T^{-1}}(\mathbf{x};\boldsymbol{\theta})\right|\Big] \tag{4.6}$$

$$\tag{4.7}$$

Given a dataset of samples from the target distribution $\{\mathbf{x}_n\}_{n=1}^N$ (e.g. a collection of quasar spectra) we can approximate the expectation over $p_x^*(\mathbf{x})$ via Monte Carlo.

$$\mathcal{L}(\boldsymbol{\theta}) \approx -\frac{1}{N}\sum_{n=1}^{N}\Big[\log p_u(T^{-1}(\mathbf{x}_n;\boldsymbol{\theta})) + \log\left|\det J_{T^{-1}}(\mathbf{x}_n;\boldsymbol{\theta})\right|\Big] \tag{4.8}$$

$$\tag{4.9}$$

Thus, training a normalizing flow amounts to explicitly maximizing the likelihood of our dataset where the likelihood of each datum is exactly calculable by casting it into a Gaussian latent space where density calculations are trivial. The only caveat is that we must compute the determinant of the Jacobian of the transformation relating the data space to the Gaussian latent space.

The ability to do exact density evaluation make normalizing flows an attractive model for probabilistic modeling. Indeed, deep generative models which admit exact

likelihoods are rare: variational autoencoders (VAEs, see [52]) admit only approximate likelihoods and generative adversarial networks (GANs, see [33]) admit no likelihoods at all. Autoregressive generative models [102, 103, 104] offer exact density evaluation but generate samples via ancestral sampling which requires repeated forward passes through the network. In contrast, normalizing flows can be designed to offer both density estimation and sampling in a single forward pass.

4.4.2 Transforms

Though the mathematical underpinning of normalizing flows is elegant, their practical application is limited by the calculation of a determinant for each bijection T_i during each forward pass. To circumvent this, most flow models employ transformations designed to yield lower triangular Jacobians. Since the determinant of a lower triangular matrix is easily computed by multiplying its diagonal elements, the complexity of the determinant computation then scales linearly in the data dimensionality, D. Two such choices of transformation are the *coupling transform* of [20]; [21] and the *autoregressive transforms* of [54]; [80]. We discuss the merits of each in turn.

4.4.2.1 Coupling Transforms

Coupling transforms operate by dividing an input datum into halves, then using the former half (hereafter the *identity features*) to predict the parameters of an invertible transformation on the latter half (hereafter the *transmuted features*). Invertibility is enforced by restricting our transformations to be strictly monotonic. The

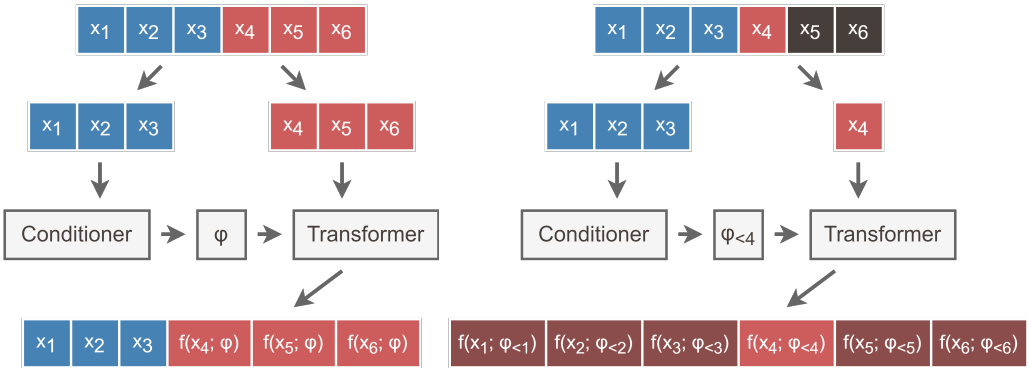


Figure 4.1: **Left:** Example coupling layer transform for a $D = 6$ dimensional input vector, $\mathbf{x}_{1:6}$. Upon splitting the input features into halves, the first half is consumed by the conditioner which outputs a parameter vector, $\boldsymbol{\phi}(\mathbf{x}_{1:3})$. This vector parameterizes a monotonic transformation of the second half features, $\mathbf{x}_{4:6}$. The coupling layer output is then given by the concatenation of the identity features with the transmuted features. **Right:** Example autoregressive layer transform for the same input vector. The input is split into $D = 6$ pieces. Features at each dimension index i undergo a mapping parameterized by all features with dimension index lower than i. Here, we display the process only for element x_4 and use shorthand notation such that $\boldsymbol{\phi}_{<i} = \boldsymbol{\phi}(\mathbf{x}_{<i})$.

identity features remain untransformed as indicated by their name. After each coupling layer, the dimensions of the data are randomly permuted (imposing an arbitrary ordering at the next layer) to allow features of each data dimension an opportunity to be transformed at some layer of the flow. Note that permutations themselves are invertible transformations with a determinant of 1 or -1. The general form of a coupling transform T is shown below (and a diagram is provided in Fig. 4.1, left).

$$\mathbf{y}_{1:d} = \mathbf{x}_{1:d} \tag{4.10}$$

$$\mathbf{y}_{d+1:D} = f(\mathbf{x}_{d+1:D};\ \boldsymbol{\phi}(\mathbf{x}_{1:d})) \tag{4.11}$$

In the normalizing flow literature, $\boldsymbol{\phi}$ is referred to as the *conditioner* and f as the *transformer*. The conditioner is typically an arbitrary neural network and the transformer is any strictly monotonic function.

Commonly, neural networks in a coupling layer use the identity features to parameterize an affine transformation on the transmuted features. In such cases, the transformer $\boldsymbol{\phi}$ outputs a set of scale and bias parameters which then act on the latter half of the input.

$$\boldsymbol{\phi}(\mathbf{x}_{1:d}) = \{\boldsymbol{\alpha}(\mathbf{x}_{1:d}),\ \boldsymbol{\beta}(\mathbf{x}_{1:d})\} \tag{4.12}$$

$$f(\mathbf{x}_{d+1:D};\ \boldsymbol{\phi}(\mathbf{x}_{1:d})) = \boldsymbol{\alpha}(\mathbf{x}_{1:d}) \cdot \mathbf{x}_{d+1:D} + \boldsymbol{\beta}(\mathbf{x}_{1:d}) \tag{4.13}$$

Flows employing such transformations have produced promising results in practice but often require an immense number of coupling layers (often hundreds) to

model complicated and high-dimensional probability distributions such as those over natural images [53].

Promising recent approaches [74, 23] in which the identity features are used to predict the parameters of a monotonically increasing piecewise spline have been shown capable of modeling highly multimodal distributions with state-of-the-art results (for flows) in log-likelihood scores. We will make use of such a flow in this work.

Coupling layers can also be made conditional in many ways. Since the conditioner is typically an arbitrary neural network, the output of this network can be conditioned on any additional information by, for example, concatenating the conditioning information onto the identity features before predicting the transformation parameters. Given F-dimensional conditioning information $c_{1:F}$, the transformation parameters are then computed as $\boldsymbol{\phi}(\mathbf{x}_{1:d} \cdot \mathbf{c}_{1:F})$ where $\cdot$ is the concatenation operator. Generally, each layer of the flow would be conditioned in this manner.

4.4.2.2 Autoregressive Transforms

Autoregressive transforms (see Fig. 4.1, right) enforce a lower triangular Jacobian by specifying the following form for their transforms:

$$\mathbf{y}_i = f(\mathbf{x}_i;\ \boldsymbol{\phi}(\mathbf{x}_{<i})) \tag{4.14}$$

With $\boldsymbol{\phi}$ as the *conditioner* and f as the *transformer*. To make this an invertible transformation, the transformer is again chosen to be a monotonic function of $\mathbf{x}_i$. If the transformer and conditioner are flexible enough to represent any function arbitrarily

well (as neural networks are), then autoregressive flows are able to approximate any distribution arbitrarily well (see [79]).

Autoregressive flows can have either one-pass sampling and D-pass density estimation or D-pass sampling and one-pass density estimation [80, 54]. Because flows are usually trained by maximizing the likelihood of the data with respect to model parameters, autoregressive flows are commonly chosen for one-pass density estimation. Autoregressive transforms can also be made conditional by manner similar to coupling transforms: conditional features, $c_{1:F}$, are concatenated onto $\mathbf{x}_{<i}$ before being inputted to the transformer.

4.4.2.3 Choosing a Transform

The choice of transform depends on the task at hand. If one is only interested in density estimation, autoregressive flows are typically chosen because of their capacity to approximate any distribution arbitrarily well with fewer layers than their coupling transform counterparts. If one would like to sample from the model, however, it is often preferable to choose a coupling transform because it offers one-pass density estimation for maximum likelihood training and one-pass data generation. Though, if efficient sampling is the only criterion, inverse autoregressive flows [54] are also an option. Recent work has demonstrated how to model distributions of data which are invariant to transformations of a given symmetry group using equivariant coupling layers [49]. In such a case, using coupling layers for density estimation may provide a better inductive bias since the coupling layer can encode the symmetry.

In this paper, we make use of both kinds of transforms: coupling in the primary model for efficient sampling of blue-side continua, and autoregressive in the secondary model for density estimation.

4.4.3 Neural Spline Flows

In contrast to affine flows where the conditioner produces the parameters of a strictly linear (and thereby inflexible) mapping, neural spline flow conditioners parameterize a piecewise spline which can approximate any differentiable monotonic function in the spline region. The added expressivity of spline layers allow neural spline flows to model complex, multi-modal probability densities with significantly fewer neural network parameters than their affine equivalent.

Neural spline flows make use of the identity features to predict the parameters of a piecewise spline in a region $x,\ y \in [-B,\ B]$ (hereafter the *spline region*) where B is a hyperparameter. The spline is required to be strictly monotonic such that the mapping is one-to-one and thereby invertible. The transformation is piecewise-defined in K different bins spanning the spline region and is linear ($y = x$) beyond. The $K + 1$ bin edges are referred to as *knots*.

Polynomial families of functions are often chosen for the spline—originally up to and including degree two polynomials [74] and subsequently up to and including degree three [22]. Recently, [23] introduced flows which employ rational quadratic splines: a family of functions defined by the division of two quadratic functions. These functions are highly expressive and yet simple to invert. Flows which make use of such transforms

are referred to by [23] as *rational quadratic neural spline flows* (RQ-NSF).

In RQ-NSFs, each of the K bins are assigned monotonic rational quadratic functions parameterized by a neural network. In total, $3K-1$ parameters define a piecewise rational quadratic spline with K bins for a single data dimension: K bin widths (size in x), K bin heights (size in y) and $K-1$ derivatives at the $K-1$ internal knots (since the derivative at the two outer knots must be unity).

For bin k, defining the bin width as $\Delta x_k = x_{k+1} - x_k$, the bin height as $\Delta y_k = y_{k+1} - y_k$, the derivative at knot k as δ_k, the constant $s_k = \Delta y_k / \Delta x_k$ and function $\xi(x) = (x - x_k)/\Delta x_k$, the rational quadratic spline is then defined as follows.

$$r_k(\xi) = \frac{\alpha_k(\xi)}{\beta_k(\xi)} = y_k + \frac{\Delta y_k \big[s_k \xi^2 + \delta_k \xi(1-\xi) \big]}{s_k + \big[\delta_{k+1} + \delta_k - 2s_k \big] \xi(1-\xi)} \tag{4.15}$$

It should be noted that the transformation acts elementwise—a unique spline is parameterized for each dimension of the transmuted features. Thus, for $i = d+1,\ d+2,\ \ldots,\ D$ (indexing the transmuted features) we parameterize a spline r_k^i such that $y_i = r_k^i(x_i)$. Then, the determinant of the Jacobian of the coupling transformation can be written as follows.

$$\det J_T = \det \frac{\partial T_i}{\partial x_j} = \prod_{i=1}^{d} 1 \times \prod_{i=d+1}^{D} \frac{\partial f_i}{\partial x_i} = \prod_{i=d+1}^{D} \frac{\partial r_k^i}{\partial x_i} \tag{4.16}$$

Where the derivative of the spline is shown below.

$$\frac{dr_k}{dx} = \frac{s_k^2\left[\delta_{k+1}\xi^2 + 2s_k\xi(1-\xi) + \delta_k(1-\xi)^2\right]}{\left[s_k + \left[\delta_{k+1} + \delta_k - 2s_k\right]\xi(1-\xi)\right]^2} \tag{4.17}$$

Inverting the spline is possible by inverting equation (4.15) and solving for the roots of the resulting quadratic equation.

4.5 Methods

This section describes our data preprocessing scheme (4.5.1), the implementation of SPECTRE (4.5.2), and our training (4.5.3) and model selection (4.5.4) procedures. Additionally, we describe the likeness of high-z targets to our training dataset (4.5.5) and our measurement of the red damping wing (4.5.6).

4.5.1 Data

4.5.1.1 Training Data

We adopt the data preprocessing scheme of [24], and briefly recount it here. For a detailed description, refer to [24]. A full Python implementation is available in a GitHub repository here: `github.com/DominikaDu/QSmooth`.

We select all quasar spectra from the 14th data release (DR14) of the Sloan Digital Sky Survey (SDSS) quasar catalog [82] within the redshift range `Z_PIPE` $\in$ [2.09, 2.51]. These spectra were captured by the extended Baryon Oscillation Spectroscopic Survey (eBOSS). This redshift range was chosen to minimize our exposure to systematic uncertainties in the SDSS pipeline redshift estimates—quasars within our

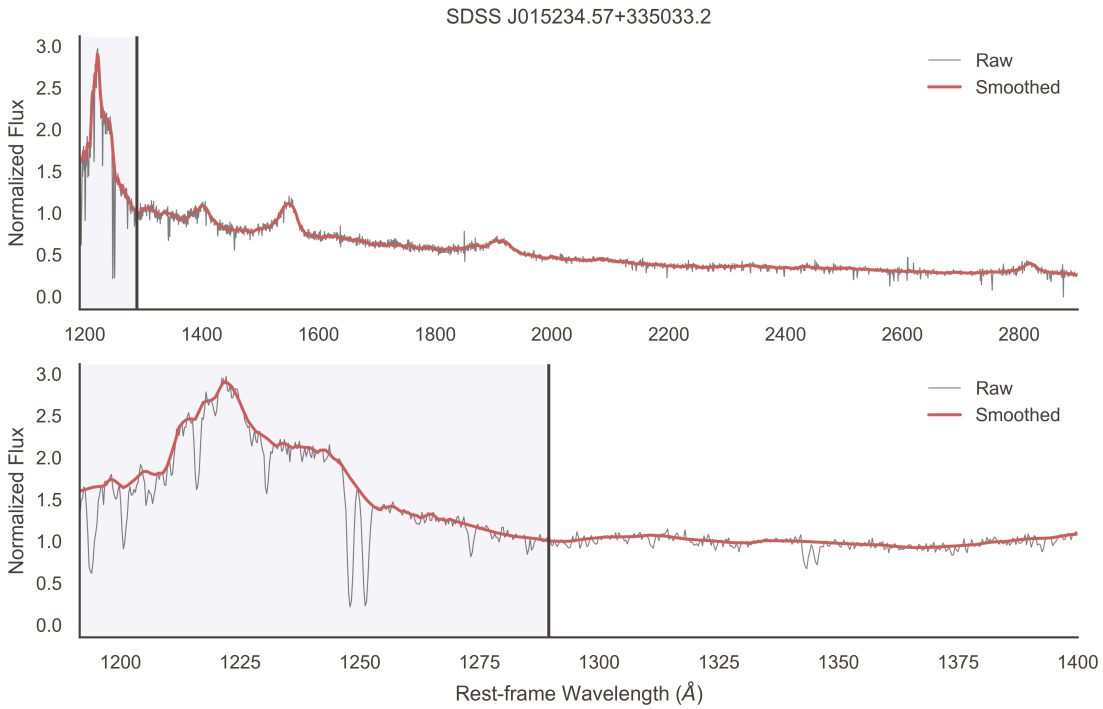


Figure 4.2: **Top**: an example of our preprocessing scheme. Here, we show a smoothed spectrum in red overlaid upon its raw flux counterpart in grey. The region with grey background on the left-hand side is the blue-side of the spectrum whose distribution we attempt to model conditional on the red-side of the spectrum (white background, right). **Bottom**: we zoom in on rest-frame wavelengths near Lyman-α to exhibit the narrow absorption features in the raw flux which have been eliminated in our preprocessing routine by identifying outliers in the difference of the raw flux and its upper envelope. All smoothed spectra are normalized to unity at 1290 Å, the threshold between the red and blue regions.

chosen redshift range include prominent emission features from Lyman-α to Mg II, which strongly constrain the SDSS redshift estimates [36].

We discard all spectra which are flagged as having broad absorption lines (`BI_CIV` $\neq 0$) or tenuous redshift estimates (`ZWARNING` $\neq 0$). We then discard spectra with low signal to noise ratios (`SN_MEDIAN_ALL` ¡ 7.0).

The spectra are subsequently smoothed. We begin by smoothing each spectrum with a median filter of kernel size $k = 50$. Then, a peak-finding algorithm identifies any peaks of the original spectrum lying above the median-smoothed boundary. We interpolate between the peaks to create an upper envelope of the spectrum. The upper envelope is then subtracted from the original spectrum. Absorption features are readily identifiable in these residuals using a RANSAC regressor [28] fit on the residual flux as a function of wavelength. We then interpolate between RANSAC inliers and smooth the resulting spectrum once more with a median filter of kernel size $k = 20$.

After shifting each spectrum to its rest-frame using the SDSS pipeline redshift estimates, each spectrum is normalized such that its flux is unity at $\lambda_{\text{rest}} = 1290\,\text{Å}$. To further clean our dataset, we eliminate any spectra whose normalized flux falls below 0.5 blueward of 1280 Å or below 0.1 redward of 1280 Å. These cuts eliminate spectra with blue-side absorption which may contaminate measurements of the damping wing and spectra with a low signal-to-noise ratio on their red-side, respectively. Each spectrum was then interpolated to a fixed grid of 3,861 wavelengths between 1191.5 Å and 2900 Å spaced uniformly in log space.

The dataset was then filtered using a random forest to cull any remaining

spectra with strong absorption features on the blue-side. After standardizing each spectrum such that the flux in each wavelength bin is z-score normalized, we perform an independent principal component analysis (PCA) on the blue and red-sides, then select the PCA coefficients which together explain 99% of the dataset variance on each side. We train a random forest regressor to predict the blue-side coefficients given the red-side coefficients using 10-fold cross-validation. Outlying spectra (with strong absorption features) preferentially occupy a tail of the reconstruction error distribution which we then select upon to eliminate data points beyond three standard deviations of the mean reconstruction error.

Our final dataset contains 13,703 quasar continua with high signal-to-noise ratios and low contamination from absorption features near Lyman-α. Our blue- and red-side continua are composed of flux values across 345 and 3516 wavelength bins, respectively. The dataset is identical to the dataset used in [24]. We divide the dataset into training, validation and testing partitions using 90/5/5 percent of the data, respectively. The partitions are chosen randomly. An example spectrum and its associated smoothed continuum approximation is shown in Fig. 4.2.

During training and inference, all input spectra were pixel-wise z-score normalized such that the model operated directly on flux z-scores calculated individually in each wavelength bin. Blue-side continua predictions were then inverse transformed before all subsequent analyses.

4.5.1.2 High-z Data

Our inference targets are two high-z quasars: ULAS J1120+0641 ($z = 7.09$) [73] observed by VLT/FORS and Gemini/GNIRS and ULAS J1342+0928 ($z = 7.54$) [8] observed by Magellan/FIRE and Gemini/GNIRS.

The spectra of ULAS J1120+0641 ($z = 7.09$) and ULAS J1342+0928 ($z = 7.54$) contain regions of poor signal-to-noise or missing data which must be imputed in order to predict their blue-side continua. To accomplish this, we again adopt the methods from [24]. We trained two fully connected feed-forward neural networks to fill in missing spectral features, one to be applied on each high-z spectrum individually. Each neural network had three hidden layers of width (55, 20, 11) neurons with exponential linear unit (ELU) activation functions. The networks were trained for 400 epochs with a batch size of 800.

For ULAS J1120+0641, the neural network inputed fluxes from $1660 - 1800\,\text{Å}$ and $2200 - 2450\,\text{Å}$. For ULAS J1342+0928, missing data was imputed in the regions between $1570 - 1700\,\text{Å}$ and $2100 - 2230\,\text{Å}$. After reconstructing the red-side spectra, we apply the same pre-processing pipeline as used on the moderate redshift quasars described above in Sec. 4.5.1.1.

4.5.2 Model

The Spectre architecture is based off of [23]'s original implementation of rational quadratic neural spline flows.

Our network employs 10 layers of spline coupling transforms, each parame-

terized by a residual network conditioner with 256 hidden units. The conditioner uses batch normalization and is regularized via dropout with $p = 0.3$. Each spline is composed of 5 bins in the region $x,\ y \in [-10,\ 10]$ i.e. $B = 10$ though we note little difference for various choices of B so long as it is greater than $\sim$3 (but note this depends on the normalization of your data).

An encoder network is used to extract relevant information from the redward spectrum during training and inference. The encoder is a fully-connected network with 4 layers of 128 hidden units. The dimensionality reduction offered by the encoder allowed us to build very deep conditional flows without reaching our GPU's memory limit.

In summary, SPECTRE produces plausible blue-side continua by transforming random Gaussian samples through a series of ten coupling transforms, each conditioned on the red-side emission. The coupling layers sequentially contort Gaussian-distributed samples to samples from the distribution over blue-side continua. The output of our model is a z-score normalized spectrum which is ultimately re-scaled to produce a candidate sample.

A full PyTorch [87] implementation of SPECTRE is available on GitHub (`github.com/davidrei`

4.5.3 Training

SPECTRE was trained on an NVIDIA V100 with a batch size of 32 and an initial learning rate of 5e$-$4. The learning rate was cosine annealed with warm restarts[2]

[2]Cosine annealing is a technique to reduce the learning rate over time. Lower learning rates near the end of training allow a model to settle into minima of the error manifold. Warm restarts reinitialize the learning rate and begin a new annealing schedule. These restarts have empirically been shown to reduce the wall clock time to convergence and in some cases improve model performance.

to a minimum learning rate of 1e−7. The initial annealing period was 5000 batches and grew by a factor of 2 after each restart. After the second restart, the learning rate was annealed to 1e−7 once more, after which it remained constant. SPECTRE's gradient norms were also clipped such that $|\nabla_{\boldsymbol{\theta}}\mathcal{L}| = \min\left(|\nabla_{\boldsymbol{\theta}}\mathcal{L}|,\ 5\right)$. This was enforced prior to each optimizer step and was used to stabilize training by constraining the update step size in parameter space for sizeable gradients. For all of our experiments, we used the Adam optimizer [51] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

4.5.4 Model Selection

Our model hyperparameters were selected via an extensive grid search using a validation set. We found that small batch sizes generally yielded better generalization performance, though when the batch size was too small ($N < 16$) training was often unstable and would occasionally diverge. We note that smaller models tended to perform best (likely due to the limited size of our dataset) though we explored deep conditional flows with up to one hundred coupling layers. We also found that reducing the resolution of our spectra by a factor of 3 improved the performance of our model. This was done by selecting flux values in every third wavelength bin. To verify that emission line profiles were not altered by this reduction in resolution, we compared a sample of low-resolution spectra to their unaltered counterparts and found no such issues. This downsampling cut the dimensionality of our blue- and red-side continua to 115 and 1172, respectively. We hypothesize that performance gains from this modification are due to the reduction in our data dimensionality which makes the task of modeling the density simpler for

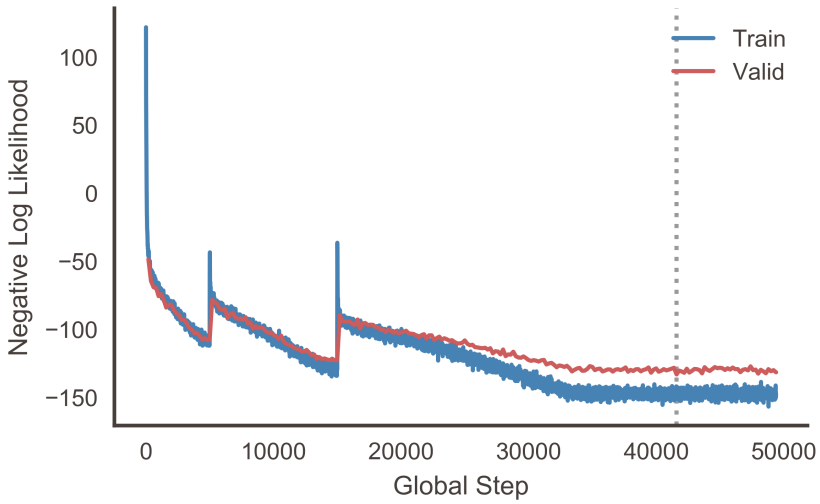


Figure 4.3: SPECTRE's training curve, showing training and validation losses (negative log likelihoods) as a function of global step, where the global step counter is incremented after each parameter update. We employ a cosine annealing learning rate schedule with warm restarts. The annealing period is $\tau = 5000$ global steps and is multiplied by two after each warm restart. The dotted line marks the global step at which our model reached minimum validation error—we use the model from this step in all of our experiments.

the flow. In addition, the unaltered spectra are strongly autocorrelated in such a way that downsampling does not remove a sizeable amount of information. Finally, we experimented with convolutional layers to encode red-side continua, but found they were not as effective as fully connected layers. We postulate that this is due to some underlying global features inherent in the spectra. These features could theoretically be accessed by increasing the receptive field of deep layers in the convolutional encoder, for example by adding additional layers or using dilated convolutions, but in practice we found a simple fully connected encoder worked best. A table denoting our complete model configuration is provided in Appendix 4.8.3.

4.5.5 Likeness of High-z and Moderate-z Spectra

We explore the applicability of our primary model by using our secondary model to quantify the similarity between moderate and high redshift quasar spectra. Since normalizing flows model the likelihood of data explicitly, it naively makes sense to use these likelihoods as a measure of how *in-distribution* a given continuum lies. As pointed out in [76] and [14], however, this can often fail. In [91], the authors show this is an expected failure mode when semantic/informative features are sparse compared to the dimensionality of the data. Our dataset falls in this regime since it is possible to reconstruct spectra with percent-level error by using only tens of PCA components to recreate fluxes across thousands of wavelength bins [16].

To avoid the spurious outlier detection performance of pure likelihoods, we employ the methods of [91] to quantify the notion of a spectrum being in-distribution

with the likelihood ratio:

$$\ell(\mathbf{x}) = \frac{p_{\text{in}}(\mathbf{x})}{p_{\text{out}}(\mathbf{x})} \tag{4.18}$$

where p_{in} is the likelihood of datum $\mathbf{x}$ given by a model trained on in-distribution data and p_{out} is the likelihood of $\mathbf{x}$ given by a model trained on out-of-distribution (OOD) data.

It is instructive to consider the case where semantic and background features are independently generated. In this scenario, one can split the likelihood of a sample, $\mathbf{x}$, into $p(\mathbf{x}) = p(\mathbf{x}_S)p(\mathbf{x}_B)$, where $\mathbf{x}_S$ are semantic features and $\mathbf{x}_B$ are background features. When semantic features are sparse, the likelihood $p(\mathbf{x})$ is dominated by the uninformative background. If both p_{in} and p_{out} give approximately the same density estimate for the background, the full likelihood ratio reduces to a likelihood ratio of semantic information. In the dependent case where background and semantic features are not independently generated, background dependence cannot be eliminated, but the likelihood ratio can still be approximated as:

$$\ell(\mathbf{x}) = \frac{p_{\text{in}}(\mathbf{x}_S|\mathbf{x}_B)p_{\text{in}}(\mathbf{x}_B)}{p_{\text{out}}(\mathbf{x}_S|\mathbf{x}_B)p_{\text{out}}(\mathbf{x}_B)} \approx \frac{p_{\text{in}}(\mathbf{x}_S|\mathbf{x}_B)}{p_{\text{out}}(\mathbf{x}_S|\mathbf{x}_B)} \tag{4.19}$$

In practice, one does not always have access to an OOD dataset. With the correct noise model, however, perturbations can be added to the in-distribution dataset which preserve population-level background statistics, but corrupt in-distribution features.

For our application, the in-distribution data are the cleaned spectra described

in Sec. 4.5.1.1. We tested different noise models to create the out-of-distribution data and we found the dataset containing the unprocessed flux measurements to give the most stringent OOD limits for both ULAS J1120+0641 and ULAS J1342+0928. Results and noise models are summarized in Appendix 4.8.1.

Because this is a density estimation exercise, we choose to use an autoregressive transform for this rational quadratic neural spline flow. We train two such flows — one on the in-distribution dataset and one on the out-of-distribution dataset — to map the red-side spectra onto a multivariate Gaussian which can be evaluated to obtain p_{in} and p_{out}. These models contain the same hyperparameters listed in Table 4.2 aside from the number of encoder layers since there is no conditional information for this task.

In Fig. 4.4 we show the likelihood ratios of training and validation sets along with high-z spectra as calculated by the two flows. Training and validation sets overlap showing the models have not overfit on the training set. As there are no guarantees for OOD detection using this method, we treat a sample's likelihood-ratio percentile as an upper bound on how in-distribution a sample lies. The likelihood ratio of ULAS J1120+0641 ($z = 7.09$) falls in the 61.1 percentile of our training set which means it is well represented by our training data. ULAS J1342+0928 ($z = 7.54$) is in the 10.4 percentile of our training set which, although not an outlier, is less typical of a moderate redshift continuum.

This hierarchy holds across various methods in the literature. In [16] the likelihood of 10 red-side PCA coefficients from a Gaussian mixture model is used to give 15 and 1.5 percentiles to ULAS J1120+0641 and ULAS J1342+0928, respectively. In

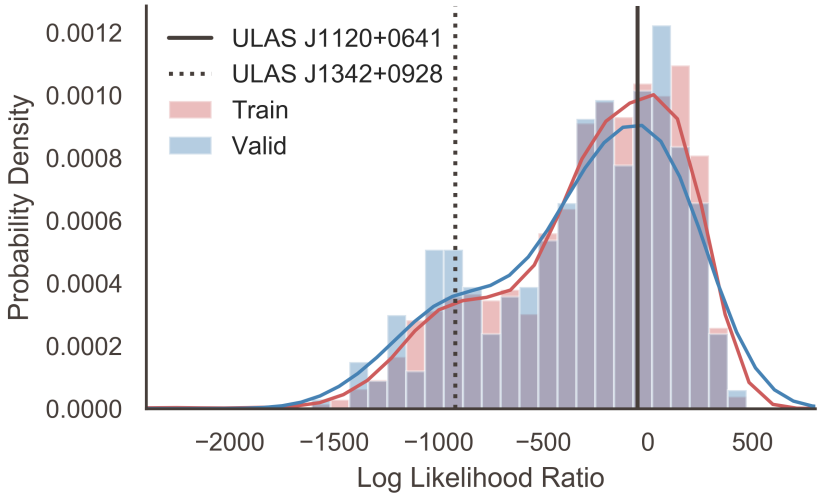


Figure 4.4: We evaluate the likelihood ratios of our training/validation sets and high-z spectra as calculated by two autoregressive rational quadratic neural spline flows. The most stringent out-of-distribution bounds result when one flow is trained on smoothed continua and the other flow is trained on raw flux measurements. We find ULAS J1120+0641 and ULAS J1342+0928 lie in the 61.1 and 10.4 percentiles of our training set, respectively. This implies that both high-z quasars share significant likeness to our training distribution and are valid inputs to our primary model.

[24], an autoencoder is trained to reconstruct 63 red-side coefficients. The reconstruction error of the red-side coefficients then gives a quantifiable measure of how well represented the high redshift continua are by the training set. With this method, they assign 52 and 1 percentiles to ULAS J1120+0641 and ULAS J1342+0928, respectively.

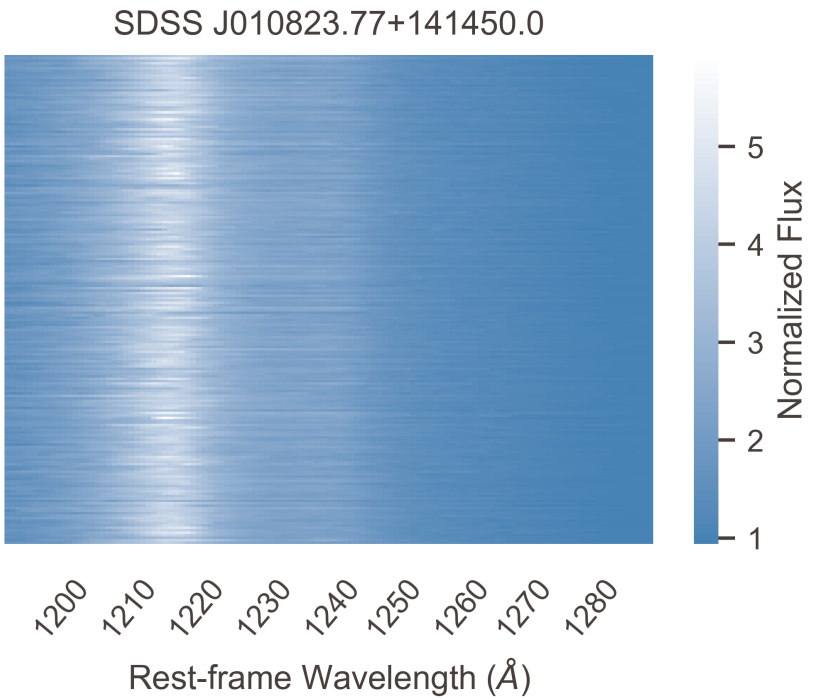


Figure 4.5: A selection of 200 blue-side continuum predictions from SPECTRE for a randomly chosen quasar. Each row corresponds to a single blue-side sample from our model, color-coded to designate its normalized flux at each wavelength. Qualitatively, the model's predictions are consistent with a Lyα peak near 1215.67 Å and a N v emission near 1240.81 Å .

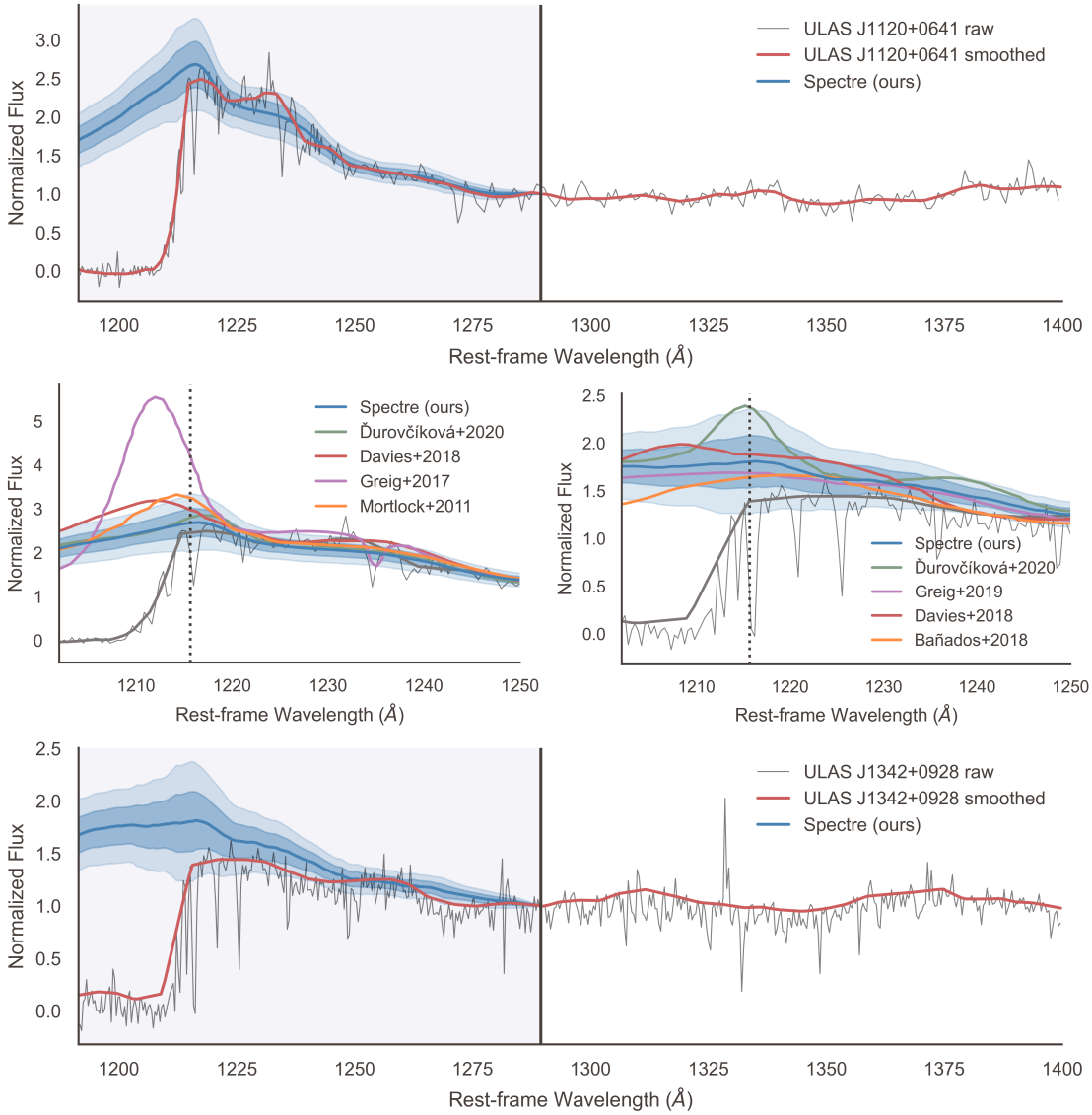


Figure 4.6: Blue-side continua predictions on two high redshift quasars, ULAS J1120+0641 and ULAS J1342+0928. Each solid blue line is the mean of one thousand samples from SPECTRE. The blue and light blue bands reflect one and two standard deviation bands at each wavelength, respectively. **Top**: Our mean prediction with two sigma uncertainty bands overlaid on top of continuum and raw flux measurements of ULAS J1120+0641.

Figure 4.6: **Bottom**: Our mean prediction with two sigma uncertainty bands overlaid on top of continuum and raw flux measurements of ULAS J1342+0928. **Middle Left**: A comparison of predictions from various authors on ULAS J1120+0641. **Middle Right**: A comparison of predictions from various authors on ULAS J1342+0928.

4.5.6 Measurement of the Damping Wing

To estimate the neutral fraction of hydrogen near the epoch of reionization, we measure the damping wing of the Gunn-Peterson trough in two high-redshift quasars: ULAS J1120+0641 at $z = 7.09$ [73] and ULAS J1342+0928 at $z = 7.54$ [8]. Measurement of the damping wing requires knowledge of the intrinsic emission of each quasar, which SPECTRE provides using conditional information from the redward spectrum. We proceed by assuming that the IGM is uniformly neutral from z_n to the blueward edge of the quasar near-zone, $z_{\rm nz} = (1 + z_s)(\lambda_{\rm nz}/\lambda_\alpha) - 1$ where z_s is the redshift of the source. For both ULAS J1120+0641 and ULAS J1342+0928 we use $\lambda_{\rm nz} = 1210\,\text{Å}$ and set $z_n = 6$.

We model the red damping wing using the analytical model of [69]:

$$\tau(\Delta\lambda) = \frac{\tau_{\rm GP} R_\alpha}{\pi}(1+\delta)^{3/2}\int_{x_1}^{x_2}\frac{x^{9/2}}{(1-x)^2}dx \tag{4.20}$$

where $\delta = \Delta\lambda/[\lambda_{\rm nz}(1+z_{\rm nz})]$ and $\Delta\lambda = \lambda - \lambda_{\rm nz}(1+z_{\rm nz})$ is the wavelength offset (in the observed frame) from the Lyman-α transition at the edge of the near-zone. The bounds of the integral are given by $x_1 = (1+z_n)/[(1+z_{\rm nz})(1+\delta)]$ and $x_2 = (1+\delta)^{-1}$. The

constant[3] $R_\alpha = 2.02 \times 10^{-8}$, and the Gunn-Peterson optical depth of neutral hydrogen, $\tau_{\rm GP}$, is given by [27] as follows:

$$\tau_{\rm GP}(z) = 1.8 \times 10^5 h^{-1} \Omega_m^{-1/2} \Big(\frac{\Omega_b h^2}{0.02}\Big) \Big(\frac{1+z}{7}\Big)^{3/2} \bar{x}_{\rm HI} \tag{4.21}$$

The integral in Eqn. 4.20 is solvable analytically and its solution is provided in [69] as:

$$I(x) = \frac{x^{9/2}}{1-x} + \frac{9}{7}x^{7/2} + \frac{9}{5}x^{5/2} + 3x^{3/2} + 9x^{1/2} - \frac{9}{2}\log\frac{1+x^{1/2}}{1-x^{1/2}} \tag{4.22}$$

We adopt the Planck 2018 cosmological parameters [90] of $h = 0.6766 \pm 0.0042$, $\Omega_m = 0.3111 \pm 0.0056$ and $\Omega_b h^2 = 0.02242 \pm 0.00014$. To determine the end of the proximity zone, we employ a common heuristic in the literature: the edge of the proximity zone is where the smoothed spectrum equals one tenth of its magnitude at at Lyman-α. For both high-z quasars, this method suggests a blueward edge of $\sim$1210 Å.

We aim to fit the damping wing model to the observed damping wing where the volume-averaged neutral fraction of hydrogen, $\bar{x}_{\rm HI}$, is our only free parameter. Equipped with our predictions of the intrinsic quasar emission near Lyman-α, $F_{\rm int}$, we measure the observed damping wing by computing the optical depth as a function of wavelength in the range $\lambda_{\rm rest} \in [1210\,\text{Å}, 1250\,\text{Å}]$.

$$\tau_{\rm obs} = -\ln\Big(\frac{F_{\rm obs}}{F_{\rm int}}\Big) \tag{4.23}$$

[3] $R_\alpha = \Lambda/(4\pi\nu_\alpha)$ with Λ the decay constant of the Lyman-α resonance and ν_α the frequency of the Lyman-α line—see [69].

where $F_{\rm obs}$ is the observed flux of the quasar for which we use the smoothed spectrum of the quasar. Note that each blue-side continua prediction sampled from SPECTRE provides a separate estimate of $F_{\rm int}$, and thereby a new measurement of $\tau_{\rm obs}$ and the neutral fraction $\bar{x}_{\rm HI}$. By Monte Carlo sampling plausible continua, we can very easily estimate the distribution over the neutral fraction.

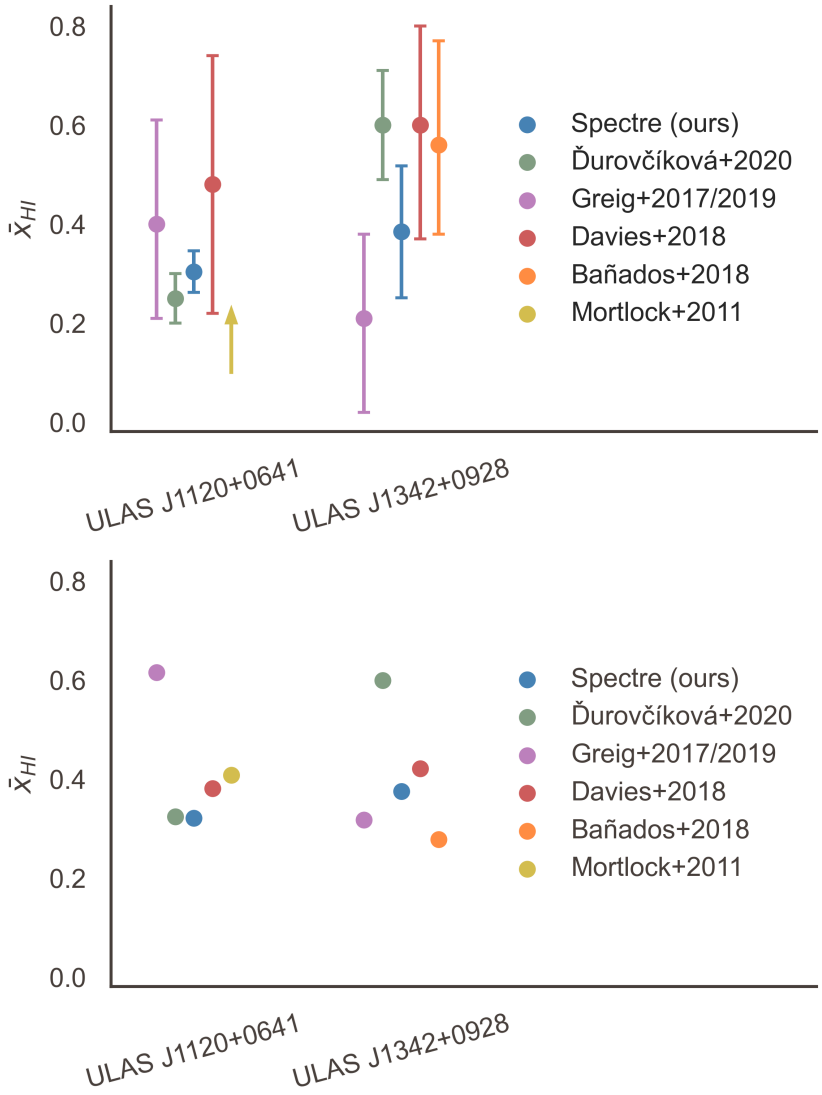
To estimate the neutral fraction itself, we assume that the value of τ in each wavelength bin is distributed according to a Gaussian distribution such that:

$$\tau_{\rm obs}(\Delta\lambda) \sim \mathcal{N}(\tau_{\rm GP} R_\alpha / \pi (\Delta\lambda/\lambda)^{-1},\ \sigma_\tau^2) \tag{4.24}$$

and we perform maximum likelihood inference to estimate the parameter $\bar{x}_{\rm HI}$ which is hidden inside of $\tau_{\rm GP}$. This amounts to a non-linear least squares problem with the additional constraint that $0 \leq \bar{x}_{\rm HI} \leq 1$. Such problems are easily solved with readily available optimization routines in Python, or a simple grid search over the open unit interval.

Uncertainty in cosmological parameters, the redshift of the source, and our estimate of the intrinsic continua all introduce error into our model. In addition, we've made simplifying assumptions: (i) the quasar's proximity zone is entirely ionized, and (ii) the IGM is uniformly dense and neutral beyond the proximity zone. In reality, the proximity zone contains residual neutral hydrogen and the IGM beyond the proximity zone is patchy and uneven, attributable to the growing ionization bubbles surrounding other luminous sources along the line-of-sight.

To estimate the effect of the uncertainty in each of the above factors to our



estimates of the neutral fraction, we use a Monte Carlo approach. We treat the source redshift and the cosmological parameters as Gaussian distributed random variables for which the reported mean is the location of the mode and the uncertainty describes the standard deviation of the mean. We then proceed by:

i. Sampling a source redshift

Figure 4.6: A comparison of our estimates of the volume-averaged neutral fraction of hydrogen for ULAS J1120+0641 ($z = 7.09$) and ULAS J1342+0928 ($z = 7.54$). **Top**: Reported results from the literature. These make use of different damping wing models (and some employ full hydrodynamical models of the IGM) which complicates direct comparison. **Bottom**: Neutral fractions for ULAS J1120+0641 and ULAS J1342+0928 computed from the mean intrinsic continua prediction of all previous approaches and a single damping wing model [69]. Error bars are omitted since only mean continuum predictions are used.

ii. Shifting the red-side spectrum to its rest-frame

iii. Estimating the blue-side continua

iv. Sampling a random vector of cosmological parameters

v. Estimating the neutral fraction

By repeated random sampling, we can empirically estimate the error propagation from each of these sources. We use a thousand samples of plausible blue-side continua from SPECTRE, and for each sample run ten Monte Carlo simulations by drawing random source redshifts and cosmological parameters. We then approximate the distribution over the neutral fraction of hydrogen with the resulting 10,000 estimates. As is typical in the literature, we quote the mean and standard deviation of these 10,000 samples as our prediction and uncertainty, though we note that (especially for ULAS J1120+0641) the distributions are notably non-Gaussian (see Fig. 4.8).

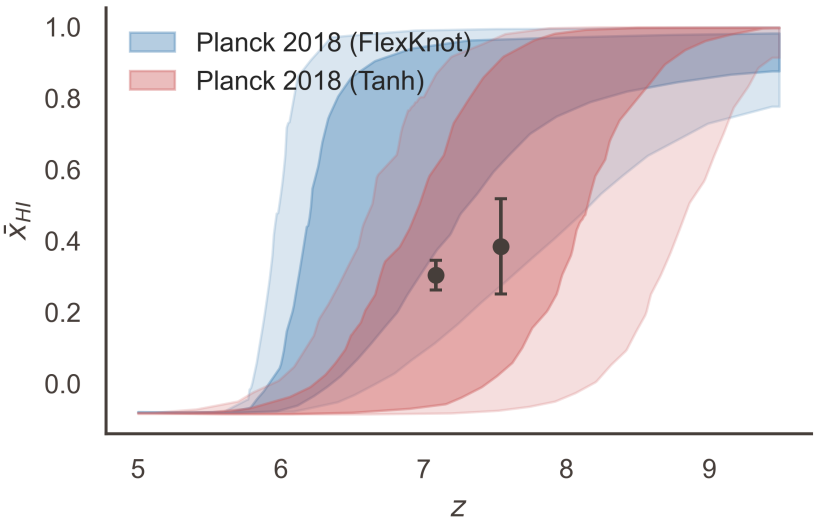


Figure 4.7: A comparison of our estimates of the volume-averaged neutral fraction of hydrogen to the Planck constraints [90]. Planck employs two models for their reionization constraints: the *Tanh* model which assumes a smooth transition from a neutral to ionized universe based on a hyperbolic tangent function and the *FlexKnot* model which can flexibly model any reionization history based on a piecewise spline with a fixed number of knots (though the final result is marginalized over this hyperparameter). SPECTRE's predictions are well within the 1-sigma confidence interval for Planck's Tanh constraints and within the 2-sigma confidence interval for the FlexKnot constraints.

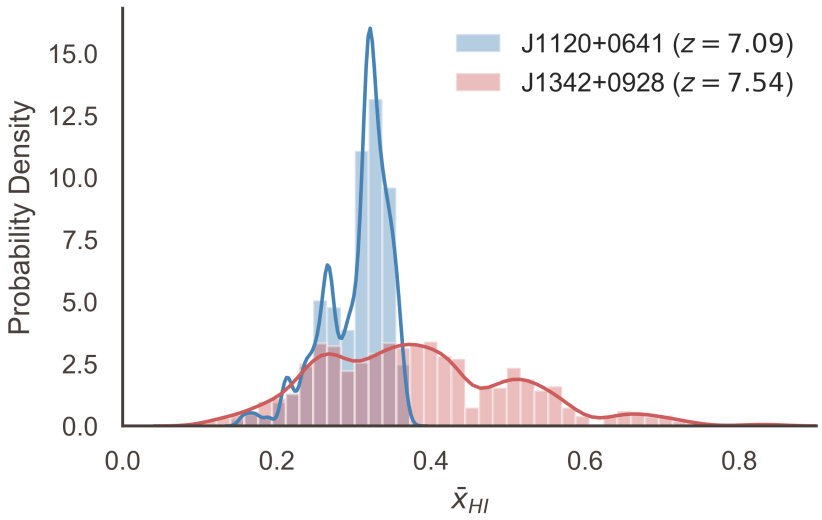


Figure 4.8: Histograms and kernel density estimates depicting the spread in neutral hydrogen fraction predictions over 10000 samples from SPECTRE. From observations of ULAS J1120+0641 we infer $\bar{x}_{\rm HI} = 0.304 \pm 0.042$ while for ULAS J1342+0928 we infer $\bar{x}_{\rm HI} = 0.384 \pm 0.133$.

4.6 Results

4.6.1 Reionization History Constraints

We display our predictions of the intrinsic continua of J1120+0641 and J1342+0928 in Fig. 4.6. For visual comparison with other approaches, we've also included figures which compare our continua predictions to those found in the literature. For ULAS J1120+0641 our intrinsic continua prediction closely matches that of [24], suggesting a modest Lyman-α emission. Of the previous approaches we've considered, we predict the weakest emission. Meanwhile, for ULAS J1342+0928 we predict a moderate Lyman-α emission which places our intrinsic continua prediction approximately midway between those of previous approaches. It should be noted, however, that SPECTRE's uncertainty is greater in its prediction of ULAS J1342+0928 and the mean continua predictions of all previous approaches are captured within our 2-sigma confidence interval though this is markedly untrue for ULAS J1120+0641.

Using the model described in Sec. 4.5.6, we estimate the volume-averaged neutral fraction of hydrogen to be $\bar{x}_{\rm HI} = 0.304 \pm 0.042$ for ULAS J1120+0641 ($z = 7.0851 \pm 0.003$) and $\bar{x}_{\rm HI} = 0.384 \pm 0.133$ for ULAS J1342+0928 ($z = 7.5413 \pm 0.0007$). A comparison between the estimated volume-averaged neutral fraction for our approach and all previous approaches is provided on the left of Fig. 4.6. We also display our results for the volume-averaged neutral fraction of hydrogen in the context of the Planck constraints [90] in Fig. 4.7.

We caution the reader to be wary of direct comparison to previous approaches

in the literature. We note that each previous approach uses very different models of the damping wing. Some employ full hydrodynamical models of the IGM while others (such as ours) make simplifying assumptions. In an attempt to provide a more direct comparison, we've used the mean continuum prediction from each previous approach and computed the neutral fraction with a single damping wing model [69]. The results are presented on the right of Fig. 4.6 and show quite different results in some cases, especially for those that employed full hydrodynamical models of the IGM. This is expected and perhaps sheds some light on the extent to which simplifying assumptions about the state of the foreground IGM biases calculations of the neutral fraction.

4.6.2 Bias and Uncertainty

To evaluate SPECTRE, we measure our continuum bias and uncertainty on a randomly selected validation set from the collection of moderate redshift spectra gathered from eBOSS. These are $N \approx 650$ quasars which were not seen during training. We define the relative continuum error ϵ_c as follows:

$$\epsilon_c = (F_{\text{model}} - F_{\text{truth}})/F_{\text{truth}} \tag{4.25}$$

where we assume the smoothed continuum estimate provided by our preprocessing method is representative of F_{truth} and we take the average over all elements of our validation set. The relative bias is then $\langle \epsilon_c \rangle$ and the relative uncertainty $\sigma(\epsilon_c)$. Here it is important to note that this definition of ϵ_c differs from [24] as it omits the absolute

value on the residual term.

In Fig. 4.10, we show our relative bias and uncertainty as a function of blue-side wavelength averaged over the validation set. We maintain low relative uncertainty at all wavelengths, averaging 6.63% across all blue-side wavelengths. However, we note that this metric is very difficult to compare between approaches since it is strongly dependent upon the preprocessing scheme. The relative uncertainty trends downward as the continuum approaches $\lambda_{\text{rest}} = 1290$ Å, the threshold between the blue and red-side spectrum where the extrapolation becomes trivial. Our model is largely unbiased save for a tendency to very slightly overpredict the continua redward of Lyα. We average a relative bias of 0.34% and are notably not strongly biased near the peak of the Lyα emission itself.

We compare our flow-based model to what is denoted as extended PCA (ePCA): an extension of the original work in [16] presented in [24] where PCA is applied independently (in log space) to the blue and red-side spectra and the resulting coefficients related via a linear model. In the original manuscript which describes the use of PCA to predict intrinsic quasar continua [16], the authors chose six and ten components for the blue and red sides, respectively, finding that inclusion of additional components did not yield better results. In [24], this model is extended to include more PCA components—enough to explain 99% of the variance. The authors find that 36 and 63 principal components are required to meet this criterion on the blue and red side, respectively. Fig. 4.11 shows the mean absolute percentage error of Spectre and ePCA as a function of blue-side wavelength. Spectre reduces the mean absolute percentage error by $\sim$5%

while offering direct calculation of confidence intervals without ensembling or otherwise. Like other deep learning models, we expect SPECTRE's performance to improve with dataset size. Near-future surveys such as the Legacy Survey of Space and Time (LSST) at the Vera C. Rubin observatory will provide millions of additional quasar spectra [45] which will likely significantly improve SPECTRE's performance.

Additionally, Appendix 4.8.2 includes a selection of blue-side continua predictions on the test set labeled with their SDSS designation for reference. These predictions were generated at random and not chosen by hand. More random predictions on the test set can be found at SPECTRE's GitHub repository: `github.com/davidreiman/spectre`.

4.6.3 Uncertainty Assessment

In this section, we explore the quality of SPECTRE's uncertainty estimates. We define the pixelwise uncertainty in N random generations from SPECTRE as follows:

$$\sigma_j = \sqrt{\sum_{i=1}^{N} \frac{(x_{ij} - \bar{x}_j)^2}{N}} \tag{4.26}$$

where i indexes the samples generated from SPECTRE and j denotes the pixel or wavelength bin. This makes the assumption that the marginal distribution over flux in each wavelength bin is Gaussian-distributed though in our experiments we find that this is true to a high degree of accuracy.

To measure the calibration of SPECTRE's uncertainty estimates, we make predictions on all spectra in the test set and compare the observed confidence intervals

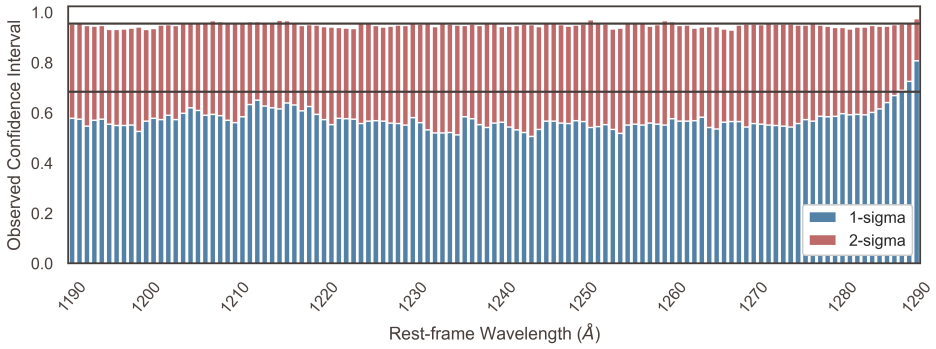


Figure 4.9: Observed confidence intervals as a function of blue-side wavelength. A calibrated model would make predictions of the marginal density over flux in a given wavelength bin such that $P\%$ of the observed absolute errors fall within the $P\%$ credible interval. We approximate the marginals as Gaussian (which we find to be true to a high degree of accuracy) and show here the observed confidence intervals for 1- and 2-sigma (e.g. $P = 68$ and $P = 95$). The solid horizontal lines correspond to the expected confidence intervals. We note that SPECTRE tends to be over-confident at the 1-sigma level but is well-calibrated at the 2-sigma level.

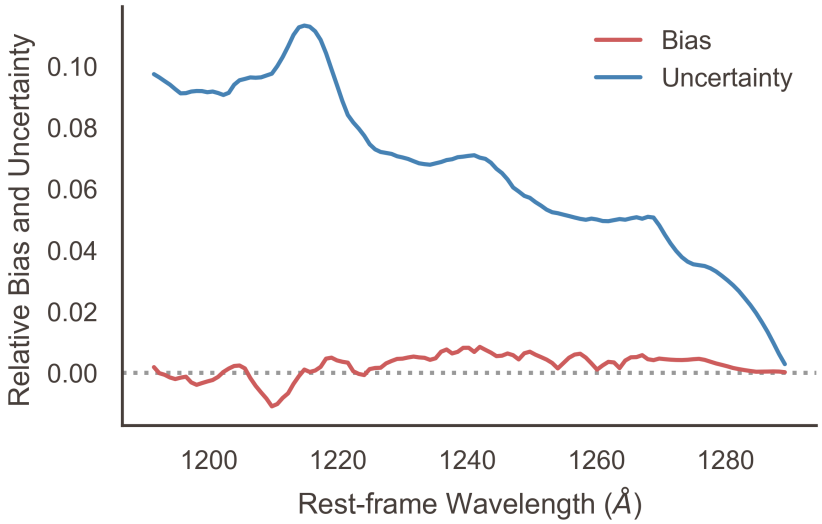


Figure 4.10: The relative prediction bias $\langle \epsilon_c \rangle$ and uncertainty $\sigma(\epsilon_c)$ (see Eqn. 4.25) as a function of blue-side wavelength averaged over our test set. Our average relative prediction uncertainty across all blue-side wavelengths is 6.63%, though we note that this metric is highly sensitive to the preprocessing scheme and is therefore difficult to compare to other methods.

Figure 4.11: Mean absolute percentage error over the test set as a function of blue-side wavelength for both SPECTRE and the extended PCA (ePCA) method of [24] (whose original formulation was introduced in [16]). Error is calculated between the smoothed continua (assumed truth) and SPECTRE's mean blue-side continua prediction. SPECTRE performs similarly to ePCA while providing an estimation of the full distribution over blue-side continua given the red-side spectrum, allowing for density estimation and sampling (and thereby Monte Carlo estimates of confidence intervals).

to the expected confidence intervals. That is, for a calibrated model we would expect to find $P\%$ of the absolute errors within the $P\%$ confidence interval predicted by the model. We can quantify our calibration by checking if this is indeed true. We do so for each pixel (wavelength bin) on the blue-side of all test-set spectra and present the results in Fig. 4.9. At the 1-sigma level, we find that on average slightly less than 68% (approximately 57%) of the absolute errors lie within SPECTRE's confidence interval which suggests that our model is slightly over-confident. However, at the 2-sigma level SPECTRE is highly calibrated, as on average ˜95% of the absolute errors fall within SPECTRE's 2-sigma confidence interval.

We also test the quality of SPECTRE's uncertainty predictions by computing the joint probability distribution of SPECTRE's absolute prediction error and predicted uncertainty over all elements of the validation set. The results are presented in Fig. 4.13. Though we note a sparsely populated tail of underestimated error (an effect also noted in Fig. 4.9), SPECTRE's uncertainty is generally strongly correlated with its own error in its predictions.

There is a growing body of literature on the tuning of an additional hyper-parameter, temperature, which modifies the base distribution after training. This is referred to as temperature-scaling and is used to increase the fidelity of model samples and calibrate uncertainties [38, 84]. We do not explore these here as this is an active field of research and it is not yet clear which prescription is most reliable [77].

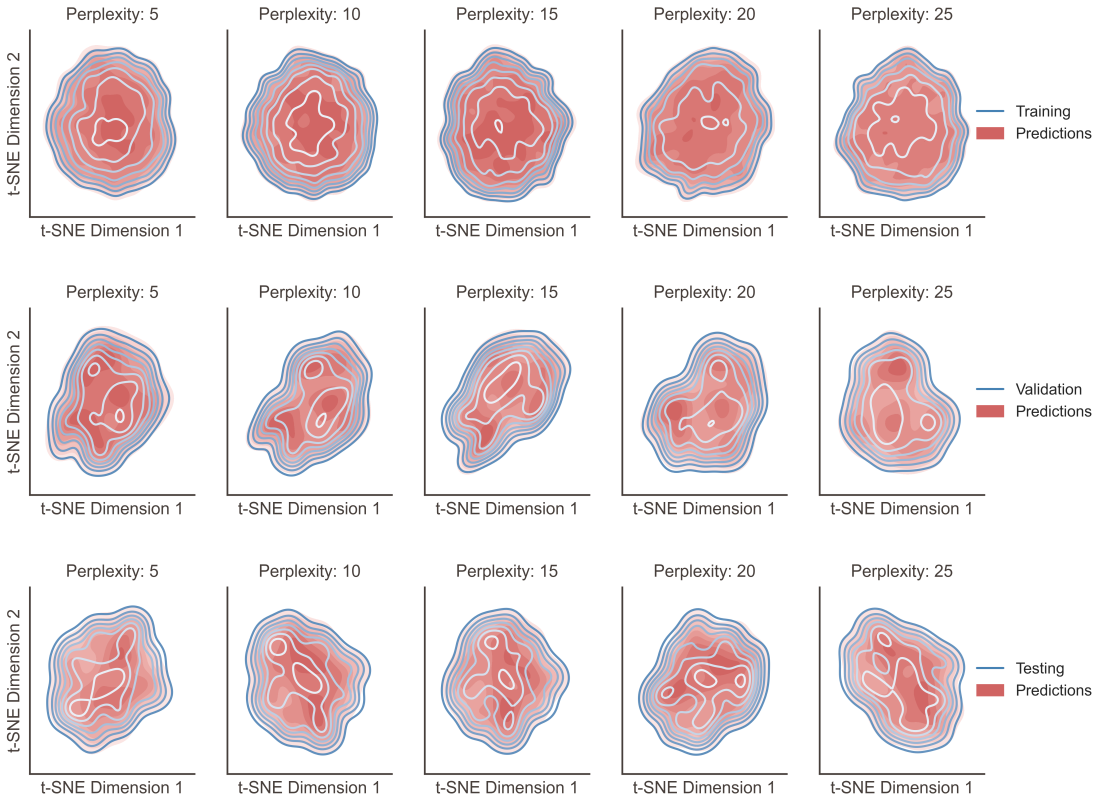


Figure 4.12: Two-dimensional embedding of blue-side spectra produced via t-Distributed Stochastic Neighbor Embedding (t-SNE). **Top**: Comparisons between our training set continua and associated model predictions. **Middle**: Comparisons between validation set continua and associated model predictions. **Bottom**: Comparisons between test set continua and associated model predictions. We display the results for a series of perplexity choices and find that training/validation set distributions consistently overlap with our model samples. Qualitatively, this implies our model's samples accurately cover the full data distribution.

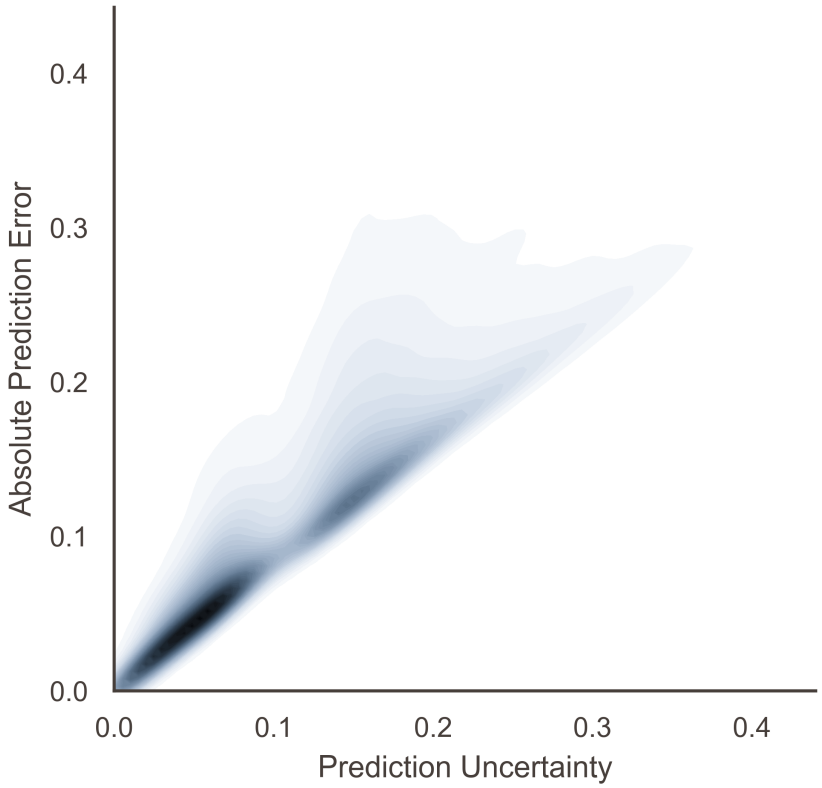


Figure 4.13: Kernel density estimate of the joint distribution over absolute error and predictive uncertainty in SPECTRE samples over all wavelengths and spectra in the test set.

4.6.4 Sample Coverage

To ensure that our model achieves full coverage of the training data distribution, we cast the full training and validation sets down to two-dimensional representations with a dimensionality reduction technique known as t-Distributed Stochastic Neighbor Embedding or t-SNE. We then produce a similar number of random samples from SPECTRE and compute the embeddings of these samples for comparison to the training and validation sets. Visualizations of the results are provided in Fig. 4.12. Both figures show unique t-SNE embeddings for increasing values of the t-SNE *perplexity*, a hyperparameter which can loosely be interpreted as an initial guess on the number of close neighbors each data point will have. Since the t-SNE algorithm is known to produce very different results for different choices of perplexity, we've shown the results for a variety of choices for completeness. We achieve full coverage of both the training and validation sets. However, we note that the location of the modes are slightly offset though generally overlapping.

4.7 Conclusion

In this manuscript, we have introduced normalizing flows as a powerful and expressive tool for probabilistic modeling in the sciences. Flows boast the ability to perform exact density evaluation, compute uncertainty intervals and carry out one-pass density estimation or sampling (provided one chooses an appropriate flow transform). Many problems in astronomy and beyond can benefit from the use of generative models

and such tasks benefit from uncertainty quantification, which other popular models (such as generative adversarial networks) cannot provide.

Among deep generative models, flows are the only models which offer both exact density evaluation and one-pass sampling (provided coupling layer or inverse autoregressive flows are used). Apart from flows, autoregressive models (which factorize high-dimensional joint probability distributions into a product of conditionals via the probability chain rule) are the only other deep generative model capable of exact density evaluation. However, sampling from a D-dimensional distribution with an autoregressive model requires D forward passes since sampling is ancestral.

Flows can also be used to learn priors over data distributions for Bayesian modeling. In maximum a posteriori inference, naive priors are often chosen which don't capture the true complexity of the data at hand. Instead, unconditional flows can provide much more realistic priors on the data given a sizeable dataset. Additionally, flows find use in likelihood-free inference techniques where they are used to approximate the intractable likelihood of a complicated and/or black-box simulator [81]. This likelihood can then be integrated to obtain the posterior.

Commonly, efficient sampling is the primary model criterion for scientists. Coupling layer or inverse autoregressive flows satisfy this criterion and have recently been used for more efficient sampling in all-purpose numerical integrators [30] and in the estimation of the expectation values of physical observables in lattice quantum chromodynamics [49].

In general, flows are capable of density estimation for a wide variety of contin-

uous or discrete-valued data. They have been successfully applied as generative models for images [53] and text [101] and used to perform anomaly detection in particle physics [75].

We have applied a specific flow variant—rational quadratic neural spline flows—to the task of intrinsic quasar continua prediction and provided a fully probabilistic model which is readily applicable to current and future high-redshift quasars. Our model consumes the red-side ($\lambda_{\text{rest}} > 1290\,\text{Å}$) spectrum to estimate a distribution over the blue-side ($1190\,\text{Å} < \lambda_{\text{rest}} < 1290\,\text{Å}$) continua. In contrast to previous approaches in the literature, SPECTRE directly models the full probability distribution over blue-side continua and therefore can be resampled arbitrarily many times to generate new plausible blue-side continua and estimate quantities such as confidence intervals without the use of ensembles.

We have also provided two new measurements of the neutral fraction of hydrogen at redshifts $z > 7$. Our results are compatible with reionization constraints from Planck and in agreement with most previous approaches. Our results support a rapid end to ionization however it is difficult to make bold claims on the topic as the available $z > 7$ data is extremely sparse and more robust modeling of the IGM would be prudent. Our modeling of the damping wing makes multiple simplifying assumptions that are untrue: (i) the quasar's proximity zone is entirely ionized, and (ii) the IGM blueward of the proximity zone is uniformly dense and neutral. These concerns can be addressed with future work using full hydrodynamical IGM modeling e.g. [15] in combination with continua predictions from SPECTRE.

Acknowledgements

We acknowledge and thank Daniel Mortlock and Eduardo Bañados for providing the spectra of ULAS J1120+0641 and ULAS J1342+0928, respectively. We would also like to thank Vanessa Boehm, Kyle Cranmer, Frederick B. Davies, Conor Durkan, and Brian Maddock for useful comments and discussion.

We acknowledge use of the Lux supercomputer at UC Santa Cruz, funded by NSF MRI grant AST 1828315.

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.

4.8 Appendix

4.8.1 Likelihood Ratio Tests

Here we describe the noise models used when we performed out-of-distribution detection using the likelihood ratio method described in Sec. 4.5.5. All results are summarized in Table 4.1. For reference, the high-z spectra both lie in the 99.9 percentile of in-distribution likelihoods.

The dataset created with the SDSS noise model is simply composed of the raw flux measurements of quasars whose continua lie in our in-distribution dataset. $\mathcal{N}(0,\ \sigma)$ refers to adding uncorrelated noise sampled from a Gaussian with mean 0 and standard deviation, σ, to our processed spectra. The noisy PCA model decomposes all continua into PCA components. Uncorrelated noise sampled from Gaussians with variance equal to each components' explained variance is added to each component

Table 4.1: A summary of out-of-distribution detection results on high-z spectra. The lowest likelihood ratio percentiles are shown in bold. We find an out-of-distribution model trained on SDSS flux measurements to provide the best likelihood ratio constraints.

	Likelihood ratio percentile	
Noise model	ULAS J1120+0641	ULAS J1342+0928
SDSS	**61.1**	**10.4**
$\mathcal{N}(0,\ \sigma = 0.07)$	61.5	17.5
$\mathcal{N}(0,\ \sigma = 0.2)$	61.4	20.1
Noisy PCA	61.4	18.8

before reconstructing the continuum.

4.8.2 Additional Samples

We show here (Fig. 4.14) a random selection of predictions on the test set spectra complete with SPECTRE's 1- and 2-sigma uncertainties for inspection by the reader. More random predictions are available at SPECTRE's GitHub repository: `github.com/davidreiman/spectre`.

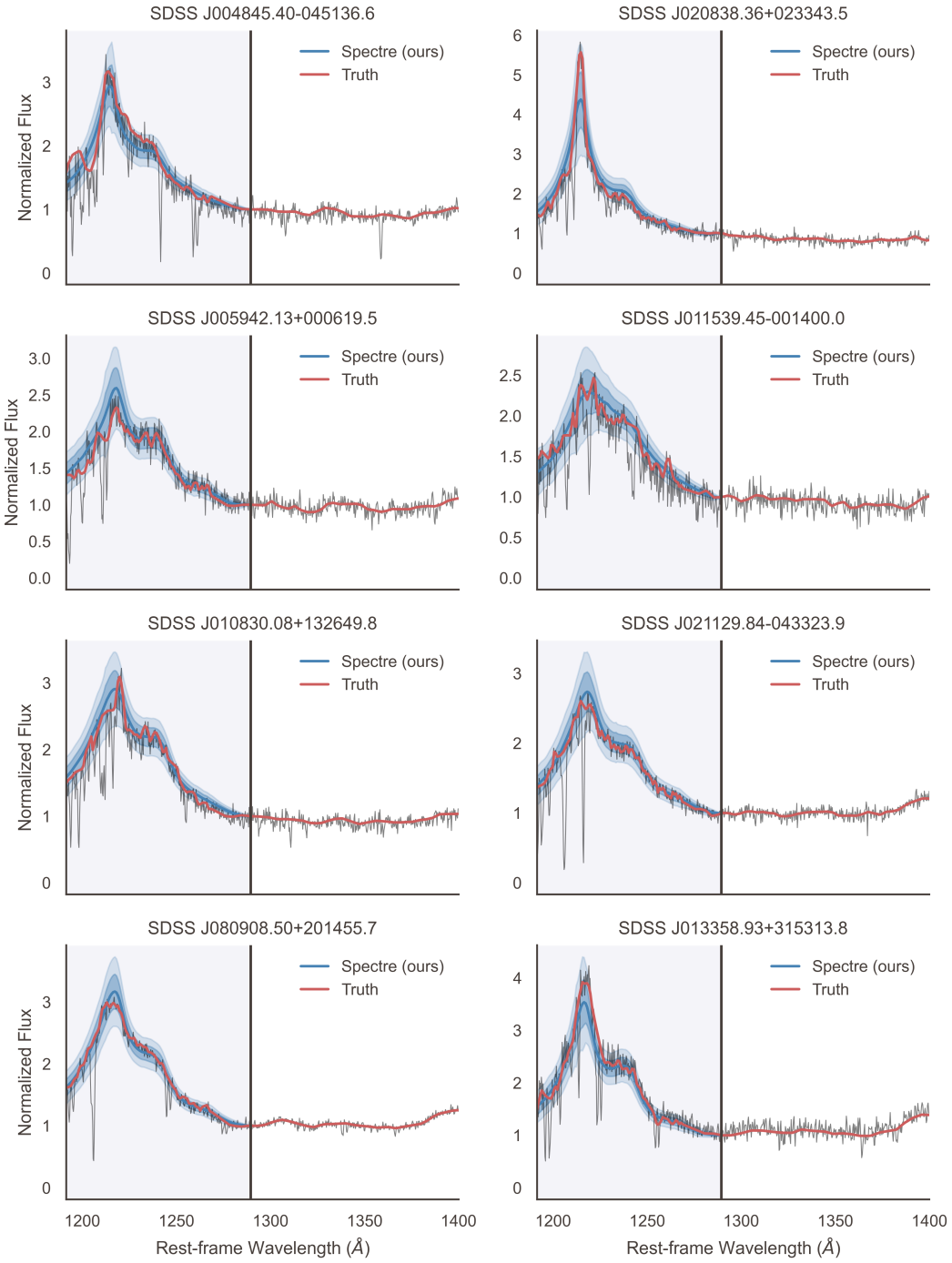


Figure 4.14: A random selection of eight predictions on moderate redshift test set spectra from eBOSS. SPECTRE's mean predictions are displayed in blue alongside its 1- and 2-sigma estimates. The assumed truth, shown in red, is the smoothed continua estimation from our preprocessing scheme, and the raw flux is shown in grey. The object in each panel is denoted by its official SDSS designation.

4.8.3 Model Hyperparameters

In Table 4.2 we list the hyperparameters of the model used for all experiments and analyses in this manuscript. These hyperparameters were chosen via grid search on a randomly selected validation set of moderate-z quasar spectra from eBOSS.

Table 4.2: Model and training hyperparameters, where the leftmost column lists the variable name we use in our codebase (see `github.com/davidreiman/spectre`), the middle column offers a brief description of the hyperparameter's function, and the rightmost column lists the value we used in our final model.

Hyperparameter	Description	Value
`n_layers`	Number of coupling layers in flow	10
`hidden_units`	Number of hidden units in conditioner	256
`n_blocks`	Number of residual blocks in conditioner	1
`tail_bound`	(x, y) bounds of spline region	10.0
`tails`	Spline function type beyond bounds	linear
`n_bins`	Number of bins in piecewise spline	5
`min_bin_height`	Minimum spline bin extent in y	0.001
`min_bin_width`	Minimum spline bin extent in x	0.001
`min_derivative`	Minimum spline derivative at knots	0.001
`dropout`	Dropout probability in flow coupling layers	0.3
`use_batch_norm`	Use batch normalization in coupling layers	`True`
`unconditional_transform`	Unconditionally transform identity features	`False`
`use_cnn_encoder`	Use a CNN encoder for redward spectrum	`False`
`encoder_units`	Number of hidden units in encoder	128
`n_encoder_layers`	Number of encoder layers	4
`encoder_dropout`	Dropout probability in encoder layers	0.0
`subsample`	Degree at which to subsample in wavelength	3
`log_transform`	Log transform data	`False`
`standardize`	Standardize data by wavelength	`True`
`learning_rate`	Initial learning rate	5e-04
`min_learning_rate`	Minimum learning rate	1e-07
`anneal_period`	Learning rate annealing period (in batches)	5000
`anneal_mult`	Annealing period multiplier after each restart	2
`n_restarts`	Total number of warm restarts	2
`batch_size`	Batch size during training	32
`eval_batch_size`	Batch size during evaluation	256
`eval_n_samples`	Number of samples to draw from flow during evaluation	1000
`grad_clip`	Maximum gradient norm during training	5.0
`n_epochs`	Maximum number of training epochs	200

Part IV

Ongoing Work

Chapter 5

Recovering the local density field in the vicinity of distant galaxies with attention-based graph neural networks

The environment of a galaxy is powerfully indicative of its properties and its evolution. By environment, one can mean either the galaxy's location in the cosmic web or the value of the nonlinear local density field in the vicinity of the galaxy. These two ideas are not independent—for example galaxies located at the intersection of cosmic filaments are typically members of an overdense *cluster*. However, the value of the density field is the primary determinant of nearly all galaxy properties [60, 99] save galaxy alignment (which is typically oriented along the cosmic web).

For this reason, recovering the value of the local density field near distant galaxies may allow us to probe their evolution and provide a new window into interesting

physics governing the formation and evolution of galaxies in general. In this ongoing work, we seek to do so using recently developed techniques from the field of geometric deep learning.

Our goal is to recover the value of the continuously varying nonlinear density field (which is dominated by dark matter) given only discrete points of luminous matter (galaxies). The problem is complicated by the fact that we operate in redshift space—the exact locations of each galaxy is in general unknown since the radial distance is estimated based on a noisy redshift estimate. These redshift estimates can be expensive and precise (spectroscopic redshifts) or inexpensive and noisy (photometric redshifts). We typically have an order of magnitude more photometric redshifts than spectroscopic redshifts. The problem is then: given a galaxy survey with a combination of mostly photometric redshifts and some spectroscopic redshifts, how can we make optimal use of this data to recover the local density field around each galaxy?

In the astronomy literature, common local density estimators typically operate by approximating the local density field with number densities e.g. by counting galaxies in a fixed aperture, computing surface densities on a dynamic 2-D surface using distances to Nth nearest neighbors, or using Voronoi tessellation volumes. This is often acceptable since generally we're more interested in density quantiles (to distinguish among high, mid or low density environments) rather than exact density values. Each of these techniques has its benefits and drawbacks—for example, Nth nearest neighbor is dynamic in that the surface area changes based on how close neighboring galaxies are to one another and therefore typically does best in dense clusters but poorly in void

regions. For each of these approaches, there is no clear way to improve the results by integrating other observed features such as galaxy photometry, morphology, effective radius, etc. which are known to correlate with local density.

Meanwhile, neural networks can learn an optimal approximator of local density using 3-D coordinates as well as any additional observables. Here, we frame the problem as a classification task where the neural network is trained to map the features of a query galaxy (and the features of all galaxies in a fixed aperture about the query galaxy) to the density quantile the query galaxy belongs to. Thus far, we have chosen to work in quartiles.

A useful inductive bias for the problem is that of graph neural networks (GNNs), permutation-invariant neural networks for use with graph-structured data such as point clouds. That is, a graph neural network will return the same answer no matter the order the points are presented to it. This is a symmetry of our problem since the density at the query point is invariant to which point in the aperture the neural network considers first. It should be noted that this invariance is rare in neural networks which typically operate on grid-based data such as images, videos, or sentences where changing the ordering is catastrophic to learning.

Our model operates by treating each galaxy (the query point and all its neighbors in aperture) as nodes of a graph, where each node is connected to every other node (including itself) via an edge. Each node is assigned a feature vector which contains the galaxy's properties (position in 3D space and any additional observables). At each layer of a graph neural network, messages are passed between nodes via connected edges and

the representation of each node is updated. In this manner, nodes disperse information through the graph and can store information about themselves and their neighbors.

The message-passing mechanism is based on attention. Attention allows each node to learn an attention distribution (a categorical distribution) over all other nodes (including itself) such that it attends to each node with varying strengths. In essence, each node is learning to weight the edges connecting itself to all other nodes such that it can strongly attend to relevant nodes while ignoring nodes that are irrelevant. Our implementation makes use of the Transformer [104] architecture by noting that removal of the positional encoding layer makes Transformers completely invariant to permutations of the input.

We replicated the results of Malavasi et al. [64] using the mock Euclid lightcone from Merson et al. [67] and achieved significant increases in quartile assignment accuracy with no included spectroscopic reference galaxies (see Fig. 5.1).

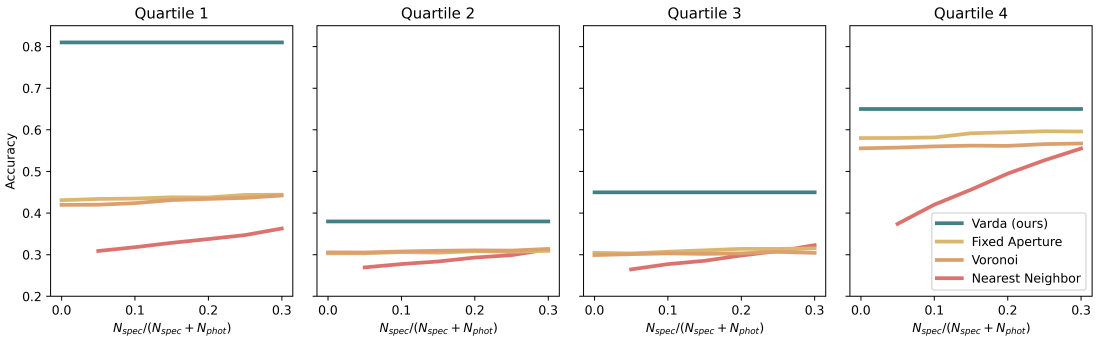


Figure 5.1: Quartile assignment accuracy as a function of spectroscopic redshift fraction for fixed aperture, Nth nearest neighbor, Voronoi, and our graph Transformer approach (Varda). Note that Varda was only evaluated in the purely-photometric redshift regime and we expect it to perform better as we introduce spectroscopic reference galaxies (hence the horizontal line should be interpreted as a lower bound).

Chapter 6

Likelihood-free inference for astrophysical simulator models with intractable likelihoods

Astronomers often have access to a very high-fidelity forward simulator of an astrophysical process of interest whose simulations depend on a vector of simulation parameters (e.g. a cosmological N-body simulation.) In a Bayesian sense, the simulator acts as a stochastic sampler of the likelihood function but often does not permit explicit evaluation of the likelihood itself—typically either because the simulator is a black-box with inaccessible inner workings or the likelihood is implicitly defined via an intractable integral over unobserved (or latent) variables [78].

Most often the intent of this modeling is to develop an inference procedure on the simulation parameters, as these often correspond to physical quantities of interest

such as constants in physical laws. As a concrete example, let us consider the measurement of the galaxy stellar mass function (GSMF), a key tool for understanding galaxy evolution. The GSMF describes the number density of galaxies as a function of mass and provides astronomers a window into understanding how intrinsic galaxy properties such as mass affect the galaxy's evolution.

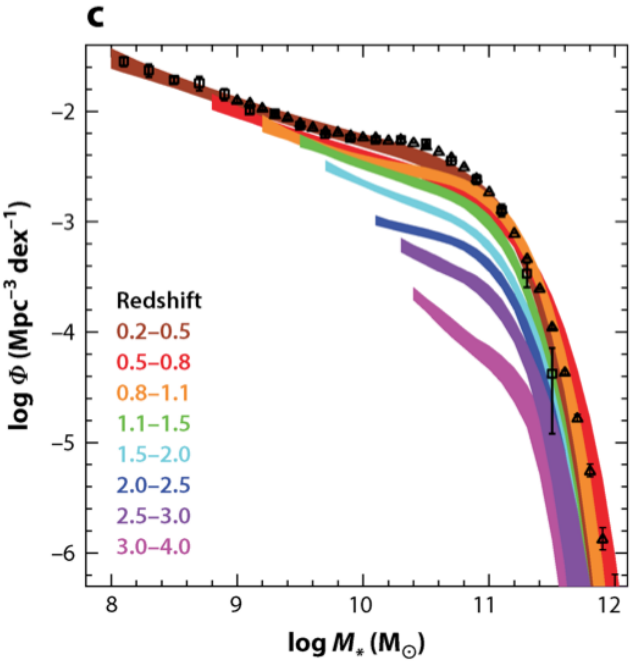


Figure 6.1: Evolution of the galaxy stellar mass function with redshift. Taken from Cosmic Star-Formation History (Madau & Dickinson 2014) [63] and used with permission from the authors.

We imagine the GSMF as a stochastic process measured in a fixed number of bins such that it can be represented as a finite random vector we call $\mathbf{x}$. The probability density which describes and generates $\mathbf{x}$ depends on a vector of simulation parameters,

in this case cosmological parameters $\boldsymbol{\theta}$ input into a large N-body simulation.

We additionally include latent variables $\mathbf{z}$ which describe the random state of the simulator. For cosmological or astrophysical purposes, this random state often controls the noise modeling of the simulator which mocks the effect our instruments (and external factors such as atmospheric seeing) have on our observations. Equipped with this knowledge we can write that $\mathbf{x} \sim p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{z})$.

We can then generate universes with any choice of parameters $\boldsymbol{\theta}$ via a cosmological N-body simulation and subsequently measure $\mathbf{x}$. Let us imagine that our goal is to constrain the cosmology of the Universe via maximum a posteriori (MAP) inference on the cosmological parameters. Then, according to Bayes' theorem we can write the following expression.

$$p(\boldsymbol{\theta}|\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})/p(\mathbf{x}) \tag{6.1}$$

If the simulator is inexpensive, and knowledge of the likelihood is of little use, we can simply generate pairs of $\{\mathbf{x}, \boldsymbol{\theta}\}$ and fit the posterior $p(\boldsymbol{\theta}|\mathbf{x})$ directly by way of a neural network $\mathbf{g}$. The network may parameterize a parametric distribution representing $p(\boldsymbol{\theta}|\mathbf{x})$ (in which case the output of the network is a vector of such parameters) or model the conditional distribution via a non-parametric approach such that $\mathbf{g}(\boldsymbol{\theta}, \mathbf{x}) \approx p(\boldsymbol{\theta}|\mathbf{x})$ (both cases of which would constitute a *conditional model*), or it may output simply its best estimate of the true value of $\boldsymbol{\theta}$ given $\mathbf{x}$, for which in the optimal case: $\mathbf{g}(\mathbf{x}) \approx \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x})$. Such an approach is described in the literature as a *discriminative model*.

Often, even if the simulator is expensive, we still have interest in modeling the likelihood function since it provides a complete description of the data $\mathbf{x}$ given $\boldsymbol{\theta}$ and can be useful for other tasks such as generating new data instances or imputing corrupted data. In such cases, we may choose to approximate the likelihood function instead of the posterior, and instead access the posterior by defining a prior over $\boldsymbol{\theta}$ and employing Markov Chain Monte Carlo (MCMC) [68]. This approach is contrasted with the discriminative approach, and is referred to as a *generative model*.

Returning to the problem at hand, in our attempt to infer cosmological parameters from the GSMF we find that the likelihood term $p(\mathbf{x}|\boldsymbol{\theta})$ is inaccessible (as is common for forward simulators in astrophysics and cosmology) due to its implicit definition via an integral over the random internal state of the simulator.

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{z}) p(\mathbf{z}) d\mathbf{z} \tag{6.2}$$

Simplifying assumptions may be made such that the likelihood can be represented in analytic form but these often fail to capture the underlying complexity of the data-generating process. In this regime, approximate or likelihood-free inference methods are required and in some cases can even provide a means of efficiently sampling the likelihood if the simulator is expensive (as is the case for large cosmological N-body simulations which require millions of CPU-hours to run).

Traditional likelihood-free (or *simulation-based*) methods such as Approximate Bayesian Computation (ABC) proceed by drawing a set of simulator parameters from the prior $p(\boldsymbol{\theta})$ with which a dataset is generated via the simulator. The simulated

dataset is compared to a true dataset via a distance measure based on the summary statistics of each. If the discrepancy is within a user-defined threshold the vector of simulator parameters is accepted and stored. After many rounds, the set of accepted parameter vectors serves as a finite sample from the approximate posterior which can then be used to estimate the posterior mode, for example. However, Approximate Bayesian Computation suffers from strong dependence on the user's choice of tolerance and summary statistics to include in the distance calculation.

Likelihood-free inference methods which instead employ neural networks as flexible function approximators to approximate the likelihood function have shown great promise and are a general purpose solution to inference problems where data is plentiful or the simulator is inexpensive to sample. In fact, even in such cases where the simulator is expensive to sample, so-called *sequential* approaches which generate new data on a round-by-round basis [81] perform competitively. Neural networks built for likelihood approximation are functions $\mathbf{h}$ parameterized by a parameter vector $\boldsymbol{\phi}$ such that $p(\mathbf{x}|\boldsymbol{\theta}) \approx \mathbf{h}_{\boldsymbol{\phi}}(\mathbf{x}, \boldsymbol{\theta})$.

In this project, we're attempting to constrain the timeline of the Epoch of Reionization via likelihood-free inference of the neutral fraction of hydrogen given observed quasar flux. We have a simulator of the damping wing which does not admit a tractable likelihood due to its definition in terms of an implicit integration over the unobserved transmission field.

$$p(\mathbf{f}|\mathbf{S}, \boldsymbol{\theta}, \boldsymbol{\sigma}) = \int \mathcal{N}(\mathbf{f}; \mathrm{T} \cdot \mathbf{S}, \boldsymbol{\sigma}) p(\mathbf{T}|\boldsymbol{\theta}) d\mathbf{T} \tag{6.3}$$

Thus far we've focused on a separate but related problem of inferring the exponent, γ, in the temperature-density relation of the IGM ($T = T_0(\rho/\bar{\rho})^{\gamma-1}$) given transmission skewers in the Lyman-α forest where the likelihood is tractable. This provides a useful testbed for the problem where we can compare our approximate likelihood methods to the ground truth.

Formally, relative fluctuations of the density field from its mean value ($\delta = (\rho - \bar{\rho})/\bar{\rho}$) are described by a Gaussian random field. In fixed bins, this means the overdensity field is multivariate Gaussian distributed, and when exponentiated this gives the optical depth of the IGM such that $\boldsymbol{\tau} = e^{\boldsymbol{\delta}}$ up to a normalization constant. The transmission field is then related to the optical depth via $\mathbf{T} = e^{-\boldsymbol{\tau}}$ such that $\mathbf{T} = e^{-e^{\boldsymbol{\delta}}}$. Since these exponential transformations are bijective, computing the likelihood of the transmission field is completely tractable in terms of the underlying density field.

$$p(\gamma|\mathbf{T}) \propto p(\mathbf{T}|\gamma) p(\gamma) \tag{6.4}$$

Therefore, since the likelihood is simply a multivariate Gaussian distribution, we can then produce forward simulations of the likelihood function $p(\mathbf{T}|\gamma)$ and our goal is to then use a normalizing flow (described in 4.4) to learn a data-driven approximation of the likelihood itself which we can compare to the true likelihood. We can also then use our approximate likelihood and Markov Chain Monte Carlo to do inference on γ.

Our initial results are promising—generated samples from our approximate likelihood match the moments of samples from the simulator (as measured by the 1-point and 2-point correlation functions, see Fig. 6.2) and with MCMC we can constrain the value of γ to within 2% error.

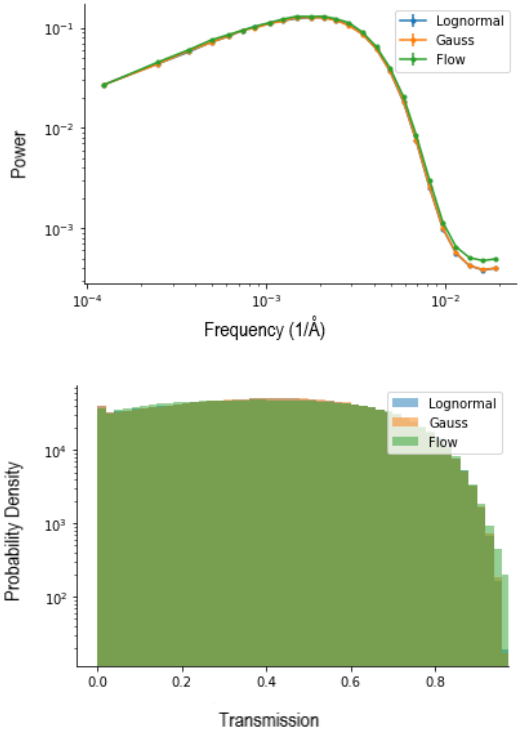


Figure 6.2: Fourier space 2-point correlation function (power spectrum) and 1-point correlation function for samples generated from the true likelihood and samples generated from our flow-based approximate likelihood.

Part V

Conclusion

In this thesis, I've presented my work using deep neural networks to approach challenging problems in astronomy and astrophysics including deblending galaxy images from near-future surveys, learning fully probabilistic models for intrinsic quasar continua, recovering the local density field in the vicinity of distant galaxies, and approximating intractable likelihood functions for astrophysical simulator models.

Deep learning is a powerful tool for data-driven probabilistic modeling in astronomy and astrophysics. Neural networks have the capacity to approximate highly complex functional relationships that are difficult or impossible to write down a priori, and can replace strong (and often incorrect) assumptions about how our data is distributed with data-driven likelihoods.

The tools discussed in this thesis have begun to revolutionize the field of astronomy, and when used carefully are powerful instruments for furthering our knowledge of the Universe and all of its constituent parts.

Bibliography

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[2] T. M. C. Abbott, F. B. Abdalla, S. Allam, A. Amara, J. Annis, J. Asorey, S. Avila, O. Ballester, M. Banerji, W. Barkhouse, L. Baruah, M. Baumer, K. Bechtol, M. R. Becker, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, J. Blazek, S. Bocquet, D. Brooks, D. Brout, E. Buckley-Geer, D. L. Burke, V. Busti, R. Campisano, L. Cardiel-Sas, A. C arnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, R. Cawthon, C. Chang, C. Conselice, G. Costa, M. Crocce, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, R. Das, G. Daues, T. M. Davis, C. Davis, J. De Vicente, D. L. DePoy, J. DeRose, S. Desai, H. T. Diehl, J. P. Dietrich, S. Dodelson,

P. Doel, A. Drlica-Wagner, T. F. Eifler, A. E. Elliott, A. E. Evrard, A. Farahi, A. Fausti Neto, E. Fernandez, D. A. Finley, M. Fitzpatrick, B. Flaugher, R. J. Foley, P. Fosalba, D. N. Friedel, J. Frieman, J. García-Bellido, E. Gaz tanaga, D. W. Gerdes, T. Giannantonio, M. S. S. Gill, K. Glazebrook, D. A. Goldstein, M. Gower, D. Gruen, R. A. Gruendl, J. Gschwend, R. R. Gupta, G. Gutierrez, S. Hamilton, W. G. Hartley, S. R. Hinton, J. M. Hislop, D. Hollowood, K. Honscheid, B. Hoyle, D. Huterer, B. Jain, D. J. James, T. Jeltema, M. W. G. Johnson, M. D. Johnson, S. Juneau, T. Kacpr zak, S. Kent, G. Khullar, M. Klein, A. Kovacs, A. M. G. Koziol, E. Krause, A. Kremin, R. Kron, K. Kuehn, S. Kuhlmann, N. Kuropatkin, O. Lahav, J. Lasker, T. S. Li, R. T. Li, A. R. Liddle, M. Lima, H. Lin, P. López-Reyes, N. MacCrann, M. A. G. Maia, J. D. Maloney, M. Manera, M. March, J. Marriner, J. L. Marshall, P. Martini, T. McClintock, T. McKay, R. G. McMahon, P. Melchior, F. Menanteau, C. J. Miller, R. Miquel, J. J. Mohr, E. Morganson, J. Mould, E. Neilsen, R. C. Nichol, D. Nidever, R. Nikutta, F. Nogueira, B. Nord, P. Nugent, L. Nunes, R. L. C. Ogando, L. Old, K. Olsen, A. B. Pace, A. Palmese, F. Paz-Chinchón, H. V. Peiris, W. J. Percival, D. Petravick, A. A. Plazas, J. Poh, C. Pond, A. Por redon, A. Pujol, A. Refregier, K. Reil, P. M. Ricker, R. P. Rollins, A. K. Romer, A. Roodman, P. Rooney, A. J. Ross, E. S. Rykoff, M. Sako, E. Sanchez, M. L. Sanchez, B. Santiago, A. Saro, V. Scarpine, D. Scolnic, A. Scott, S. Serrano, I. Sevilla-Noarbe, E. Sheldon, N. Shipp, M. L. Silveira, R. C. Smith, J. A. Smith, M. Smith, M. Soares-Santos, F. Sobre ira, J. Song, A. Stebbins, E. Suchyta, M. Sullivan, M. E. C. Swanson, G. Tarle, J. Thaler, D. Thomas,

R. C. Thomas, M. A. Troxel, D. L. Tucker, V. Vikram, A. K. Vivas, A. R. Walker, R. H. Wechsler, J. Weller, W. Wester, R. C. Wolf, H. Wu, B. Yanny, A. Zenteno, Y. Zhang, and J. Zuntz. The Dark Energy Survey Data Release 1. *ArXiv e-prints*, January 2018.

[3] R. G. Abraham and S. van den Bergh. The Morphological Evolution of Galaxies. *Science*, 293:1273–1278, August 2001.

[4] J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. Allende Prieto, K. S. J. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, C. A. L. Bailer-Jones, I. K. Baldry, J. C. Barentine, B. A. Bassett, A. C. Becker, T. C. Beers, E. F. Bell, A. A. Berlind, M. Bernardi, M. R. Blanton, J. J. Bochanski, W. N. Boroski, J. Brinchmann, J. Brinkmann, R. J. Brunner, T. Budavári, S. Carliles, M. A. Carr, F. J. Castander, D. Cinabro, R. J. Cool, K. R. Covey, I. Csabai, C. E. Cunha, J. R. A. Davenport, B. Dilday, M. Doi, D. J. Eisenstein, M. L. Evans, X. Fan, D. P. Finkbeiner, S. D. Friedman, J. A. Frieman, M. Fukugita, B. T. Gänsicke, E. Gates, B. Gillespie, K. Glazebrook, J. Gray, E. K. Grebel, J. E. Gunn, V. K. Gurbani, P. B. Hall, P. Harding, M. Harvanek, S. L. Hawley, J. Hayes, T. M. Heckman, J. S. Hendry, R. B. Hindsley, C. M. Hirata, C. J. Hogan, D. W. Hogg, J. B. Hyde, S.-i. Ichikawa, Ž. Ivezić, S. Jester, J. A. Johnson, A. M. Jorgensen, M. Jurić, S. M. Kent, R. Kessler, S. J. Kleinman, G. R. Knapp, R. G. Kron, J. Krzesinski, N. Kuropatkin, D. Q. Lamb, H. Lampeitl, S. Lebedeva, Y. S. Lee, R. French Leger, S. Lépine, M. Lima, H. Lin, D. C. Long, C. P. Loomis, J. Loveday, R. H.

Lupton, O. Malanushenko, V. Malanushenko, R. Mandelbaum, B. Margon, J. P. Marriner, D. Martínez-Delgado, T. Matsubara, P. M. McGehee, T. A. McKay, A. Meiksin, H. L. Morrison, J. A. Munn, R. Nakajima, E. H. Neilsen, Jr., H. J. Newberg, R. C. Nichol, T. Nicinski, M. Nieto-Santisteban, A. Nitta, S. Okamura, R. Owen, H. Oyaizu, N. Padmanabhan, K. Pan, C. Park, J. Peoples, Jr., J. R. Pier, A. C. Pope, N. Purger, M. J. Raddick, P. Re Fiorentin, G. T. Richards, M. W. Richmond, A. G. Riess, H.-W. Rix, C. M. Rockosi, M. Sako, D. J. Schlegel, D. P. Schneider, M. R. Schreiber, A. D. Schwope, U. Seljak, B. Sesar, E. Sheldon, K. Shimasaku, T. Sivarani, J. Allyn Smith, S. A. Snedden, M. Steinmetz, M. A. Strauss, M. SubbaRao, Y. Suto, A. S. Szalay, I. Szapudi, P. Szkody, M. Tegmark, A. R. Thakar, C. A. Tremonti, D. L. Tucker, A. Uomoto, D. E. Vanden Berk, J. Vandenberg, S. Vidrih, M. S. Vogeley, W. Voges, N. P. Vogt, Y. Wadadekar, D. H. Weinberg, A. A. West, S. D. M. White, B. C. Wilhite, B. Yanny, D. R. Yocum, D. G. York, I. Zehavi, and D. B. Zucker. The Sixth Data Release of the Sloan Digital Sky Survey. , 175:297–313, April 2008.

[5] J. Amiaux, R. Scaramella, Y. Mellier, B. Altieri, C. Burigana, A. Da Silva, P. Gomez, J. Hoar, R. Laureijs, E. Maiorano, D. Magalhães Oliveira, F. Renk, G. Saavedra Criado, I. Tereno, J. L. Auguères, J. Brinchmann, M. Cropper, L. Duvet, A. Ealet, P. Franzetti, B. Garilli, P. Gondoin, L. Guzzo, H. Hoekstra, R. Holmes, K. Jahnke, T. Kitching, M. Meneghetti, W. Percival, and S. Warren. Euclid mission: building of a reference survey. In *Space Telescopes and Instru-*

mentation 2012: Optical, Infrared, and Millimeter Wave, volume 8442 of , page 84420Z, September 2012.

[6] S. Andreon, G. Gargiulo, G. Longo, R. Tagliaferri, and N. Capuano. Wide field imaging - I. Applications of neural networks to object detection and star/galaxy classification. , 319:700–716, December 2000.

[7] Stanislaw Bajtlik, Robert C Duncan, and Jeremiah P Ostriker. Quasar ionization of lyman-alpha clouds-the proximity effect, a probe of the ultraviolet background at high redshift. *The Astrophysical Journal*, 327:570–583, 1988.

[8] Eduardo Bañados, Bram P Venemans, Chiara Mazzucchelli, Emanuele P Farina, Fabian Walter, Feige Wang, Roberto Decarli, Daniel Stern, Xiaohui Fan, Frederick B Davies, et al. An 800-million-solar-mass black hole in a significantly neutral universe at a redshift of 7.5. *Nature*, 553(7689):473–476, 2018.

[9] Julian E Bautista, Mariana Vargas-Magaña, Kyle S Dawson, Will J Percival, Jonathan Brinkmann, Joel Brownstein, Benjamin Camacho, Johan Comparat, Hector Gil-Marín, Eva-Maria Mueller, et al. The sdss-iv extended baryon oscillation spectroscopic survey: baryon acoustic oscillations at redshift of 0.72 with the dr14 luminous red galaxy sample. *The Astrophysical Journal*, 863(1):110, 2018.

[10] E. Bertin and S. Arnouts. SExtractor: Software for source extraction. , 117:393–404, June 1996.

[11] Todd A Boroson and Richard F Green. The emission-line properties of low-redshift

quasi-stellar objects. *The Astrophysical Journal Supplement Series*, 80:109–135, 1992.

[12] RJ Bouwens, GD Illingworth, PA Oesch, J Caruana, B Holwerda, R Smit, and S Wilkins. Reionization after planck: the derived growth of the cosmic ionizing emissivity now matches the growth of the galaxy uv luminosity density. *The Astrophysical Journal*, 811(2):140, 2015.

[13] C. Chang, M. Jarvis, B. Jain, S. M. Kahn, D. Kirkby, A. Connolly, S. Krughoff, E.-H. Peng, and J. R. Peterson. The effective number density of galaxies for weak lensing measurements in the LSST project. , 434:2121–2135, September 2013.

[14] Hyunsun Choi, Eric Jang, and Alexander A. Alemi. WAIC, but Why? Generative Ensembles for Robust Anomaly Detection. *arXiv e-prints*, page arXiv:1810.01392, October 2018.

[15] Frederick B Davies, Joseph F Hennawi, Eduardo Bañados, Zarija Lukić, Roberto Decarli, Xiaohui Fan, Emanuele P Farina, Chiara Mazzucchelli, Hans-Walter Rix, Bram P Venemans, et al. Quantitative constraints on the reionization history from the igm damping wing signature in two quasars at z¿ 7. *The Astrophysical Journal*, 864(2):142, 2018.

[16] Frederick B Davies, Joseph F Hennawi, Eduardo Bañados, Robert A Simcoe, Roberto Decarli, Xiaohui Fan, Emanuele P Farina, Chiara Mazzucchelli, Hans-

Walter Rix, Bram P Venemans, et al. Predicting quasar continua near lyα with principal component analysis. *The Astrophysical Journal*, 864(2):143, 2018.

[17] W. A. Dawson, M. D. Schneider, J. A. Tyson, and M. J. Jee. The Ellipticity Distribution of Ambiguously Blended Objects. , 816:11, January 2016.

[18] J. T. A. de Jong, G. A. Verdoes Kleijn, K. H. Kuijken, and E. A. Valentijn. The Kilo-Degree Survey. *Experimental Astronomy*, 35:25–44, January 2013.

[19] Sander Dieleman, Kyle W. Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. , 450:1441–1459, June 2015.

[20] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[21] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[22] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Cubic-spline flows. *arXiv preprint arXiv:1906.02145*, 2019.

[23] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, pages 7509–7520, 2019.

[24] Dominika Ďurovčíková, Harley Katz, Sarah EI Bosman, Frederick B Davies, Julien Devriendt, and Adrianne Slyz. Reionization history constraints from neural net-

work based predictions of high-redshift quasar continua. *Monthly Notices of the Royal Astronomical Society*, 493(3):4256–4275, 2020.

[25] Anna-Christina Eilers, Frederick B. Davies, and Joseph F. Hennawi. The Opacity of the Intergalactic Medium Measured along Quasar Sightlines at z $\sim$ 6. , 864(1):53, September 2018.

[26] Anna-Christina Eilers, Frederick B. Davies, Joseph F. Hennawi, J. Xavier Prochaska, Zarija Lukić, and Chiara Mazzucchelli. Implications of z $\sim$ 6 Quasar Proximity Zones for the Epoch of Reionization and Quasar Lifetimes. , 840(1):24, May 2017.

[27] Xiaohui Fan, Michael A Strauss, Robert H Becker, Richard L White, James E Gunn, Gillian R Knapp, Gordon T Richards, Donald P Schneider, J Brinkmann, and Masataka Fukugita. Constraining the evolution of the ionizing background and the epoch of reionization with z 6 quasars. ii. a sample of 19 quasars. *The Astronomical Journal*, 132(1):117, 2006.

[28] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[29] Levi Fussell and Ben Moews. Forging new worlds: high-resolution synthetic galaxies with chained generative adversarial networks. *arXiv e-prints*, page arXiv:1811.03081, November 2018.

[30] Christina Gao, Joshua Isaacson, and Claudius Krause. i-flow: High-dimensional integration and sampling with normalizing flows. page arXiv:2001.05486, 1 2020.

[31] Neil Gehrels, D Spergel, WFIRST SDT, et al. Wide-field infrared survey telescope (wfirst) mission and synergies with lisa and ligo-virgo. In *Journal of Physics: Conference Series*, volume 610, page 012007. IOP Publishing, 2015.

[32] I. Goodfellow. NIPS 2016 Tutorial: Generative Adversarial Networks. *ArXiv e-prints*, December 2017.

[33] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *ArXiv e-prints*, June 2014.

[34] Bradley Greig, Andrei Mesinger, and Eduardo Bañados. Constraints on reionization from the z= 7.5 qso ulasj1342+ 0928. *Monthly Notices of the Royal Astronomical Society*, 484(4):5094–5101, 2019.

[35] Bradley Greig, Andrei Mesinger, Zoltan Haiman, and Robert A Simcoe. Are we witnessing the epoch of reionization at z= 7.1 from the spectrum of j1120+ 0641? *Monthly Notices of the Royal Astronomical Society*, 466(4):4239–4249, 2017.

[36] Bradley Greig, Andrei Mesinger, Ian D McGreer, Simona Gallerani, and Zoltán Haiman. Lyα emission-line reconstruction for high-z qsos. *Monthly Notices of the Royal Astronomical Society*, 466(2):1814–1838, 2017.

[37] James E Gunn and Bruce A Peterson. On the density of neutral hydrogen in intergalactic space. *The Astrophysical Journal*, 142:1633–1641, 1965.

[38] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[39] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *ArXiv e-prints*, December 2015.

[40] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *ArXiv e-prints*, February 2015.

[41] Lars Hernquist, Neal Katz, David H Weinberg, and Jordi Miralda-Escude. The lyman-alpha forest in the cold dark matter model. *The Astrophysical Journal Letters*, 457(2):L51, 1996.

[42] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[43] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv e-prints*, February 2015.

[44] M. J. Irwin. Automatic analysis of crowded fields. , 214:575–604, June 1985.

[45] Željko Ivezić. Lsst survey: millions and millions of quasars. *Proceedings of the International Astronomical Union*, 12(S324):330–337, 2016.

[46] J. F. Jarvis and J. A. Tyson. FOCAS - Faint Object Classification and Analysis System. , 86:476–495, March 1981.

[47] Danilo Jimenez Rezende, George Papamakarios, Sébastien Racanière, Michael S. Albergo, Gurtej Kanwar, Phiala E. Shanahan, and Kyle Cranmer. Normalizing Flows on Tori and Spheres. *arXiv e-prints*, page arXiv:2002.02428, February 2020.

[48] R Joseph, F Courbin, and J-L Starck. Multi-band morpho-spectral component analysis deblending tool (muscadet): Deblending colourful objects. *Astronomy & Astrophysics*, 589:A2, 2016.

[49] Gurtej Kanwar, Michael S Albergo, Denis Boyda, Kyle Cranmer, Daniel C Hackett, Sébastien Racanière, Danilo Jimenez Rezende, and Phiala E Shanahan. Equivariant flow-based sampling for lattice gauge theory. *arXiv preprint arXiv:2003.06413*, 2020.

[50] W. C. Keel, A. M. Manning, B. W. Holwerda, M. Mezzoprete, C. J. Lintott, K. Schawinski, P. Gay, and K. L. Masters. Galaxy Zoo: A Catalog of Overlapping Galaxy Pairs for Dust Studies. , 125:2, January 2013.

[51] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *ArXiv e-prints*, December 2014.

[52] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arxiv 2013. *arXiv preprint arXiv:1312.6114*, 2013.

[53] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10215–10224. Curran Associates, Inc., 2018.

[54] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.

[55] Stephen Cole Kleene. Representation of events in nerve nets and finite automata. Technical report, RAND PROJECT AIR FORCE SANTA MONICA CA, 1951.

[56] Jan Kremer, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Steenstrup Pedersen, and Christian Igel. Big universe, big data: machine learning and image analysis for astronomy. *IEEE Intelligent Systems*, 32(2):16–22, 2017.

[57] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[58] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.

[59] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *ArXiv e-prints*, September 2016.

[60] Christoph T Lee, Joel R Primack, Peter Behroozi, Aldo Rodríguez-Puebla, Doug Hellinger, and Avishai Dekel. Properties of dark matter haloes as a function of local environment density. *Monthly Notices of the Royal Astronomical Society*, 466(4):3834–3858, 2017.

[61] C. Lintott, K. Schawinski, S. Bamford, A. Slosar, K. Land, D. Thomas, E. Edmondson, K. Masters, R. C. Nichol, M. J. Raddick, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. , 410:166–178, January 2011.

[62] LSST Science Collaboration, P. A. Abell, J. Allison, S. F. Anderson, J. R. Andrew, J. R. P. Angel, L. Armus, D. Arnett, S. J. Asztalos, T. S. Axelrod, and et al. LSST Science Book, Version 2.0. *ArXiv e-prints*, December 2009.

[63] Piero Madau and Mark Dickinson. Cosmic star-formation history. *Annual Review of Astronomy and Astrophysics*, 52:415–486, 2014.

[64] Nicola Malavasi, Lucia Pozzetti, Olga Cucciati, Sandro Bardelli, and Andrea Cimatti. Reconstructing the galaxy density field with photometric redshifts-i.

methodology and validation on stellar mass functions. *Astronomy & Astrophysics*, 585:A116, 2016.

[65] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[66] Peter Melchior, Fred Moolekamp, Maximilian Jerdee, Robert Armstrong, Ai-Lei Sun, James Bosch, and Robert Lupton. scarlet: Source separation in multi-band images by constrained matrix factorization. *arXiv preprint arXiv:1802.10157*, 2018.

[67] Alexander I Merson, Carlton M Baugh, John C Helly, Violeta Gonzalez-Perez, Shaun Cole, Richard Bielby, Peder Norberg, Carlos S Frenk, Andrew J Benson, Richard G Bower, et al. Lightcone mock catalogues from semi-analytic models of galaxy formation–i. construction and application to the bzk colour selection. *Monthly Notices of the Royal Astronomical Society*, 429(1):556–578, 2013.

[68] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[69] Jordi Miralda-Escudé. Reionization of the intergalactic medium and the damping wing of the gunn-peterson trough. *The Astrophysical Journal*, 501(1):15, 1998.

[70] Jordi Miralda-Escudé, Renyue Cen, Jeremiah P Ostriker, and Michael Rauch.

The lyα forest from gravitational collapse in the cold dark matter+ λ model. *The Astrophysical Journal*, 471(2):582, 1996.

[71] Satoshi Miyazaki, Yutaka Komiyama, Satoshi Kawanomoto, Yoshiyuki Doi, Hana Furusawa, Takashi Hamana, Yusuke Hayashi, Hiroyuki Ikeda, Yukiko Kamata, Hiroshi Karoji, Michitaro Koike, Tomio Kurakami, S Miyama, Tomoki Morokuma, Fumiaki Nakata, Kazuhito Namikawa, Hidehiko Nakaya, Kyoji Nariai, Yoshiyuki Obuchi, and Hideo Yokota. Hyper suprime-cam: System design and verification of image quality. *Publications of the Astronomical Society of Japan*, 70, 01 2018.

[72] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, pages 2924–2932, 2014.

[73] Daniel J Mortlock, Stephen J Warren, Bram P Venemans, Mitesh Patel, Paul C Hewett, Richard G McMahon, Chris Simpson, Tom Theuns, Eduardo A Gonzáles-Solares, Andy Adamson, et al. A luminous quasar at a redshift of z= 7.085. *Nature*, 474(7353):616–619, 2011.

[74] Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. Neural importance sampling. *arXiv preprint arXiv:1808.03856*, 2018.

[75] Benjamin Nachman and David Shih. Anomaly detection with density estimation. *arXiv preprint arXiv:2001.04990*, 2020.

[76] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lak-

shminarayanan. Do Deep Generative Models Know What They Don't Know? *arXiv e-prints*, page arXiv:1810.09136, October 2018.

[77] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, 2019.

[78] George Papamakarios. Neural density estimation and likelihood-free inference. *arXiv preprint arXiv:1910.13233*, 2019.

[79] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

[80] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

[81] George Papamakarios, David C. Sterratt, and Iain Murray. Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows. *arXiv e-prints*, page arXiv:1805.07226, May 2018.

[82] Isabelle Pâris, Patrick Petitjean, Éric Aubourg, Adam D Myers, Alina Streblyanska, Brad W Lyke, Scott F Anderson, Éric Armengaud, Julian Bautista, Michael R

Blanton, et al. The sloan digital sky survey quasar catalog: fourteenth data release. *Astronomy & Astrophysics*, 613:A51, 2018.

[83] Pâris, I., Petitjean, P., Rollinde, E., Aubourg, E., Busca, N., Charlassier, R., Delubac, T., Hamilton, J.-Ch., Le Goff, J.-M., Palanque-Delabrouille, N., Peirani, S., Pichon, Ch., Rich, J., Vargas-Magaña, M., and Yèche, Ch. A principal component analysis of quasar uv spectra at z 3. *A&A*, 530:A50, 2011.

[84] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, and Alexander Ku. Image transformer. *CoRR*, abs/1802.05751, 2018.

[85] J. Pasquet, E. Bertin, M. Treyer, S. Arnouts, and D. Fouchez. Photometric redshifts from SDSS images using a Convolutional Neural Network. *ArXiv e-prints*, June 2018.

[86] J. Pasquet, E. Bertin, M. Treyer, S. Arnouts, and D. Fouchez. Photometric redshifts from SDSS images using a Convolutional Neural Network. *ArXiv e-prints*, June 2018.

[87] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[88] W. J. Pearson, L. Wang, F. F. S. van der Tak, P. D. Hurley, D. Burgarella, and

S. J. Oliver. De-blending deep Herschel surveys: A multi-wavelength approach. , 603:A102, July 2017.

[89] Gualtiero Piccinini. The first computational theory of mind and brain: a close look at mcculloch and pitts's "logical calculus of ideas immanent in nervous activity". *Synthese*, 141(2):175–215, 2004.

[90] Planck Collaboration. Planck 2018 results. vi. cosmological parameters. *arXiv preprint arXiv:1807.06209*, 2018.

[91] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood Ratios for Out-of-Distribution Detection. *arXiv e-prints*, page arXiv:1906.02845, June 2019.

[92] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[93] B. T. P. Rowe, M. Jarvis, R. Mandelbaum, G. M. Bernstein, J. Bosch, M. Simet, J. E. Meyers, T. Kacprzak, R. Nakajima, J. Zuntz, H. Miyatake, J. P. Dietrich, R. Armstrong, P. Melchior, and M. S. S. Gill. GALSIM: The modular galaxy image simulation toolkit. *Astronomy and Computing*, 10:121–150, April 2015.

[94] K. Schawinski, C. Zhang, H. Zhang, L. Fowler, and G. K. Santhanam. Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit. , 467:L110–L114, May 2017.

[95] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and

Z. Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *ArXiv e-prints*, September 2016.

[96] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[97] Nao Suzuki, David Tytler, David Kirkman, John M. O'Meara, and Dan Lubin. Predicting QSO Continua in the Lyα Forest. , 618(2):592–600, January 2005.

[98] Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[99] Masayuki Tanaka, Tomotsugu Goto, Sadanori Okamura, Kazuhiro Shimasaku, and Jon Brinkmann. The environmental dependence of galaxy properties in the local universe: dependences on luminosity, local density, and system richness. *The Astronomical Journal*, 128(6):2677, 2004.

[100] Tomonori Totani, Kentaro Aoki, Takashi Hattori, George Kosugi, Yuu Niino, Tetsuya Hashimoto, Nobuyuki Kawai, Kouji Ohta, Takanori Sakamoto, and Toru Yamada. Probing intergalactic neutral hydrogen by the lyman alpha red damping wing of gamma-ray burst 130606a afterglow spectrum at z= 5.913. *Publications of the Astronomical Society of Japan*, 66(3):63, 2014.

[101] Dustin Tran, Keyon Vafa, Kumar Agrawal, Laurent Dinh, and Ben Poole. Dis-

crete flows: Invertible generative models of discrete data. In *Advances in Neural Information Processing Systems*, pages 14692–14701, 2019.

[102] Aaron Van den Oord, N Kalchbrenner, and K Kavukcuoglu. Pixel recurrent neural networks. arxiv 2016. *arXiv preprint arXiv:1601.06759*, 2016.

[103] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.

[104] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[105] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[106] Y. Zhang, T. A. McKay, E. Bertin, T. Jeltema, C. J. Miller, E. Rykoff, and J. Song. Crowded Cluster Cores: An Algorithm for Deblending in Dark Energy Survey Images. , 127:1183, November 2015.

[107] Yuanyuan Zhang, Timothy A. McKay, Emmanuel Bertin, Tesla Jeltema, Christopher J. Miller, Eli Rykoff, and Jeeseon Song. Crowded cluster cores: An algorithm for deblending in dark energy survey images. *Publications of the Astronomical Society of the Pacific*, 127(957):1183–1196, 2015.