UNIVERSITY OF CALIFORNIA
SANTA CRUZ

APPLICATIONS OF NEURAL PROBABILISTIC MODELING TO HIGH ENERGY AND ASTROPHYSICS

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

DOCTOR OF PHILOSOPHY

in

PHYSICS

by

John Tamanas

March 2022

The Dissertation of John Tamanas
is approved:

Professor Stefano Profumo, Chair

Professor Wolfgang Altmannshofer

Professor Stefania Gori

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures **v**

List of Tables **xv**

Abstract **xvi**

Dedication **xvii**

1 Introduction **1**

2 Fully probabilistic quasar continua predictions near Lyman- with conditional neural spline flows **4**

2.1 Introduction 4
2.2 Related Work 9
2.3 Background 11
2.3.1 Normalizing Flows 11
2.3.2 Transforms 15
2.3.3 Neural Spline Flows 18
2.4 Methods 20
2.4.1 Data 20
2.4.2 Model 24
2.4.3 Training 25
2.4.4 Model Selection 25
2.4.5 Likeness of High-z and Moderate-z Spectra 27
2.4.6 Measurement of the Damping Wing 29
2.5 Results 38
2.5.1 Reionization History Constraints 38
2.5.2 Bias and Uncertainty 39
2.5.3 Uncertainty Assessment 41
2.5.4 Sample Coverage 45
2.6 Conclusion 45

3 Via Machinae: Searching for Stellar Streams using Unsupervised Machine Learning **51**
3.1 Introduction . . . 51
3.2 Data and Input Variables . . . 57
3.3 Via Machinae: The algorithm . . . 62
3.3.1 ANODE: Defining the search regions . . . 62
3.3.2 ANODE: Density estimation . . . 64
3.3.3 Regions of interest . . . 69
3.3.4 Line-finding and stream detection . . . 73
3.3.5 Final Merging and Clustering . . . 75
3.4 Demonstrating the full Via Machinae Algorithm with GD-1 . . . 81
3.5 Conclusions . . . 89
Acknowledgements . . . 91

4 Simulation-Based Inference for efficient sampling of the pMSSM subject to experimental constraints **93**
4.1 Introduction . . . 93
4.2 Simulation-Based Inference (SBI) . . . 95
4.2.1 Likelihood-to-Evidence Ratio Estimation . . . 96
4.2.2 Sequential Neural Likelihood-to-Evidence Ratio Estimation . . . 97
4.3 Phenomenological Minimal Supersymmetric Model (pMSSM) . . . 98
4.4 Benefits from Sequential Training . . . 100
4.4.1 Setup . . . 101
4.4.2 Results . . . 102
4.5 Application to pMSSM . . . 103
4.5.1 SBI Setup . . . 103
4.5.2 Results . . . 104
4.6 Conclusions . . . 109

A Appendix **111**
A.1 Likelihood ratio tests . . . 111
A.2 Additional Samples . . . 112
A.3 Model Hyperparameters . . . 112
A.4 Details of the MAF . . . 115
A.4.1 Training and model selection . . . 115
A.4.2 Hyperparameter Optimization . . . 116
A.5 Globular cluster detection . . . 118
A.6 Comments on stream 2 and disk stars . . . 119
A.7 Training details . . . 121

List of Figures

2.1 To infer the blue-side continuum, SPECTRE samples one thousand predictions conditional on the red-side. The assumed truth, shown in red, is the smoothed continua estimation from our preprocessing scheme, and the raw flux is shown in grey. 5

2.2 **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 $\boldsymbol{\phi}_{<i} = \boldsymbol{\phi}(\mathbf{x}_{<i})$. . . . 14

2.3 **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. 21

2.4 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. 26

2.5 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. 30

2.6 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 . . 31

2.7 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. **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. 32

2.8 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). **Left**: 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. **Right**: 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 [102]. Error bars are omitted since only mean continuum predictions are used. 34

2.9 A comparison of our estimates of the volume-averaged neutral fraction of hydrogen to the Planck constraints [132]. 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. 35

2.10 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 $\bar{x}_{\rm HI} = 0.384 \pm 0.133$. 36

2.11 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. 41

2.12 The relative prediction bias $\langle \epsilon_c \rangle$ and uncertainty $\sigma(\epsilon_c)$ (see Eqn. 2.25) 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. 42

2.13 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 [37] (whose original formulation was introduced in [29]). 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). 43

2.14 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. 46

2.15 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. 47

3.1 A schematic showing an overview of the VIA MACHINAE algorithm. Bolded and boxed terms are defined in Sec. 3.3 (with the exception of *patches*, which are described in Sec. 3.2). First we divide up the sky into evenly-tiled 15° patches. Within each patch, we further divide up the stars into search regions defined by a window in μ_λ, one of the proper motion coordinates (the remaining data features for each star are denoted $\vec{x}$). Then we train the ANODE algorithm on the search regions and their complements, to learn a data-driven measure of local overdensities $R(\vec{x})$. To turn this measure into a stream finder, we further divide up the SRs into regions of interest based on the orthogonal proper motion coordinate μ^*_ϕ. We apply an automated line-finding algorithm based on the Hough transform to the 100 highest-R stars in each ROI. Finally, we combine ROIs adjacent in proper motion that have concordant best-fit line parameters into proto-clusters, and cluster these across adjacent patches of the sky into stream candidates. 55

3.2 The positions in Galactic ℓ and b coordinates used for the centers for the datasets from the *Gaia* DR2 used in our full-sky analysis. The missing grid centers in the Galactic Southern hemisphere are the patches that overlapped with the Magellanic Clouds. The 21 centers which contain the GD-1 stream are shown in red, and the patch used as the worked example in Sec. 3.3 is denoted with a star. 60

3.3 Upper row: Angular position in (ϕ, λ) coordinates (left), proper motion in $(\mu^*_\phi, \mu_\lambda)$ coordinates (center), and photometry (right) of all stars in the patch centered on $(\alpha, \delta) = (148.6°, 24.2°)$. (Note the streaking in angular position due to non-uniform coverage in *Gaia* DR2.) Bottom row: As above, with stars identified by [PWB18] 61

3.4 Left: R distribution for the SR $\mu_\lambda = [-17, -11]$ mas/yr in the patch centered at $(\alpha, \delta) = (148.6°, 24.2°)$. Stars identified as likely members of GD-1 by [PWB18] 66

3.5 Upper row: Angular position in (ϕ, λ) coordinates (left), proper motion in $(\mu_\phi^*, \mu_\lambda)$ coordinates (center), and photometry (right) of all stars (blue) in the $\mu_\lambda \in [-17, -11]$ mas/yr SR of our example patch centered on $(\alpha, \delta) = (148.6°, 24.2°)$. Bottom row: As the upper row, applying the $R > R_{\rm cut}$ cut on the stars in the SR (purple). The GD-1 stream becomes immediately apparent. See text for details. 67

3.6 Normalized histogram of μ_ϕ^* values for stars in the 10° patch centered on $(\alpha, \delta) = (148.6°, 24.2°)$, requiring $2 < |\mu_\lambda| < 4$ mas/yr (blue) and $|\mu_\lambda| < 2$ mas/yr (red). Note that the high density of stars near $\mu_\phi^* \sim 0$ with $|\mu_\lambda| < 2$ mas/yr are not represented in the sample which does not overlap $\mu_\lambda \sim 0$. These very distant stars with near-zero total proper motion are absent as a population from search regions which do not include the zero point of proper motion. 71

3.7 Left: Angular position in (ϕ, λ) coordinates for the 100 highest-R stars (purple) in the $\mu_\phi^* \in [-8, -2]$ mas/year, $\mu_\lambda \in [-17, -11]$ mas/year ROI from our example patch. Right: Associated curves in Hough space for these stars (black lines). The significance $\sigma_L(\theta, \rho)$ of a line oriented at each (θ, ρ) value is shown in color. The region around the point of maximum contrast (as identified by the Via Machinae algorithm) is indicated by the inner white box, with the region defining the background shown as the outer box. 72

3.8 Line significance σ_L versus the central μ_ϕ^* value for each ROI with $\mu_\lambda \in [-17, -11]$ mas/yr in our example patch. Vertical red lines indicate the minimum and maximum μ_ϕ^* values for the candidate GD-1 stars of [PWB18] 76

3.9 A schematic showing how regions of interest (ROIs) are combined into different protoclusters. The different colors denote different seeds, i.e. clusters of ROIs with adjacent μ_λ and the same μ_ϕ^* values. The boxes show how adjacent seeds are combined into protoclusters with different $N_{\rm SR}$. 78

3.10 Histograms of the fraction of stars in the best-fit line of each ROI that were identified as likely GD-1 stars by [PWB18] 79
3.11 Histogram of the σ_L^{tot} values of protoclusters with $N_{\text{SR}} \geq 3$, with each protocluster weighted by the number of ROIs it contains. 82
3.12 The two stream candidates built out of proto-clusters with $N_{\text{SR}} \geq 3$ and $\sigma_L^{\text{tot}} \geq 7.5$. 83
3.13 Scatter plots of the angular positions, proper motions, and color/magnitudes of the 1,688 stars in the more prominent of the two stream candidates identified by Via Machinae, overlaid on the likely GD-1 stars tagged by [PWB18] 84
3.14 Comparison of the likely GD-1 stars from [PWB18] 85
3.15 For each SR, we plot the fraction of total stars N_{total} in an SR which are identified as likely members of GD-1 by [PWB18] 86

4.1 The fraction of points sampled from the posterior that lie within the ranges specified in Eq. 4.5 102
4.2 The expected amount of time (minutes) needed to obtain one sample in the ranges specified in Eq. 4.5. 103
4.3 Samples from the posterior which lie within the experimental constraints specified in Eq. 4.5 obtained by running SNRE with hyperparameters listed in Table A.3 . We note the overdensity of small smuon masses which affect a_μ and Ω_χ by coannhiliations. See Figures 4.5 and 4.4 for corner plots with subsets of these parameters. 104
4.4 Corner plots of μ and gaugino mass parameters created from the same samples as Fig 4.3. Dotted lines correspond to $\mu = M_1$ and $M - 2 = M_1$, respectively. We see a strong bias towards bino-like and wino-like LSPs. Direct detection constraints force μ towards larger values of its range. Additionally, we see a dependence on the relative signs of μ and M_2. . . 105

4.5 Corner plots of smuon and dark matter masses calculated from the same samples as Fig 4.3. Shaded regions do not have a neutralino LSP which, although not explicitly excluded, are not experimentally viable. We note the large concentration of light right-handed smuons. 106
4.6 Samples from the approximate posterior which lie within the experimental constraints specified in Eq. 4.5. The color of each point corresponds to the p-value from compressed spectra constraints reported by ATLAS. The background is shaded according to a gaussian kernel density estimate in order to visualize the concentration of points on these axes. We note the ability of ATLAS to probe models which lie in the neutrino floor. 107
4.7 Same as Fig. 4.6 but plotted on ATLAS constraints. We expect future analyses to constrain much of the viable parameter space with small mass splittings. 108

A.1 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. 113
A.2 Left: SIC curve of signal efficiency ϵ_S to $\epsilon_S/\sqrt{\epsilon_B}$ (for a background efficiency ϵ_B) as a cut is placed on $\log R$, for all hyperparameters tested on the GD-1 example dataset. Right: Density plot of $\log p_{\rm bg}$ versus $\log R$ for stars in the signal region of the GD-1 dataset used for hyperparameter optimization, trained using the neural network parameters that maximize the true-positive over root false-positive rate. 117

A.3 Scatter plots of the angular positions, proper motions, and color/magnitudes of the stars in the second, less prominent stream candidate identified by VIA MACHINAE, overlaid on 2d histograms of all the stars in the circular patch that contains this stream candidate (darker pixels indicate higher density of stars). As in Fig. 3.13, the VIA MACHINAE stars are color-coded by position in α, to facilitate cross referencing between the three individual scatter plots. 120

List of Tables

4.1 Parameter domains for pMSSM. All masses and couplings are in GeV. All squark mass parameters are set to 4 TeV. 99

A.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. 112

A.2 Model and training hyperparameters, where the leftmost column lists the variable name we use in our codebase, 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. 114

A.3 Hyperparameters common to all (S)NRE algorithms tested in all applications. 122

Abstract

Applications of Neural Probabilistic Modeling to High Energy and Astrophysics

by

John Tamanas

Across all fields of science, statistical modeling often involves simplifying assumptions of functional forms in order to make problems tractable. The advent of modern deep learning techniques, however, has begun to alter this approach by replacing overly simplistic functions with universal function approximators that are fast to train and evaluate. In this work, we exhibit three seemingly disparate applications united under the same framework of neural probabilistic modeling.We will begin by using a neural density estimator - known as a normalizing flow - to model intrinsic quasar continua near Lyman-α given the redward spectrum. We use these predictions to estimate the neutral fraction of hydrogen in the spectrum of two z>7 quasars and apply constraints to the timeline of the Epoch of Reionization. Secondly, we show how to use normalizing flows for identifying stellar streams in data from the Gaia telescope. We use anomaly detection techniques developed for High Energy Physics with limited astrophysical assumptions to re-discover GD-1. Finally, we will demonstrate how to use the approximate Bayesian techniques of simulation-based inference to efficiently sample pMSSM models from an experimentally constrained parameter space. Interestingly, the majority of such models are just outside of current experimental bounds.

This is dedicated to my parents, Androula and Kypros, who continue to teach me how to work hard and how to be a good person. To my sisters, Elena and Christa, who constantly show me endless support. To my brothers in law, Joe and Brian who keep me sane. To my nieces and nephews – Alex, Marco, Daphne, Lucas, and Ophelia – who fill me with joy and love. Thank you all for never letting me forget the spaghetti.

Chapter 1

Introduction

The approach of the scientist is to make simplifying assumptions about the nature of a problem of interest, and then to use these assumptions to make predictions. This approach, commonly used in conjuction with Occam's razor, has been extremely successful in the past because it makes extremely complex and difficult problems easier to comprehend and solve. Oftentimes, these assumptions result in the reduced complexity of the mathematical functions used in order to to reduce their computational cost or due to lack of a priori functional forms. This can result in an analysis that is too simplistic to fully describe the dynamics of the problem, thus limiting the discovery potential of the application.

In recent years, however, the emergence of deep learning has been seen as a new paradigm for data-rich analysis. Neural networks, the main ingredient in this approach, have been shown to be able to learn from data in a way that is not only computationally efficient, but is also able to make predictions that are more accurate than the simple assumptions made by the scientist [84].

The deep learning paradigm has given practitioners a language to form their problems in terms of optimization. That is to say, a user is now able to formulate their workflow in terms of an *arbitrary* functions' outputs in downstream tasks. Due to the ability of

neural networks to model the dynamics of any arbitrary function [104], workflows can be shown to be theoretically optimal under assymptotic convergence assumptions.

In this work, we'll pay special attention to problems of high energy and astrophysics that can be interpreted in the language of probability. Specifically, we'll be looking at problems which make use of optimal classifiers and approximate bayesian inference. In both scenarios, it is common in the literature to see assumptions of gaussianity applied to low-dimensional summary statistics of the data (e.g. [52]). When an abundance of data is present, we take the view that simple probability distributions functions (PDFs), e.g. gaussians, can be replaced with a flexible PDF learned from the data (assuming there is no a priori best family of probability distribution functions available).

When a full PDF is required, we will make use of the generative neural network known as a normalizing flow [146]. These models make no assumptions about the underlying distribution of the data, and can be shown to be universal distribution approximators under assymptotic convergence assumptions [35]. These models are especially useful when one would like to perform both density estimation and data emulation. When only density estimation is required, however, we can use arbitrary neural networks to return learned, accurate estimates of the PDF. We will refer to the use of neural networks used in the context of PDFs as *neural probabilistic models*.

In this thesis we show how neural probabilistic models can be used to a variety of problems. In Chapter 2 we will describe the application of normalizing flows to the problem of inferring quasar continua from noisy spectra as there is no known model of the dynamics of quasar continuum emission that can be used. We use the learned model to apply constraints to the timeline of the Epoch of Reionization. Chapter 3 shows an application of neural probabilistic modeling to discover stellar streams in Gaia data. We exhibit how normalizing flows can be used to approximate the optimal decision function between background and assumed signal data en route to re-

discovering GD-1. In Chapter 4 we explore how simulation-based inference can improve analyses of experimentally constrained parameter spaces. Specifically, we show how to efficiently sample the large-dimensional parameter space of the phenomenological minimal supersymmetric model that has yet to be experimentally constrained.

Chapter 2

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

2.1 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 $z \sim 8$ [132], 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

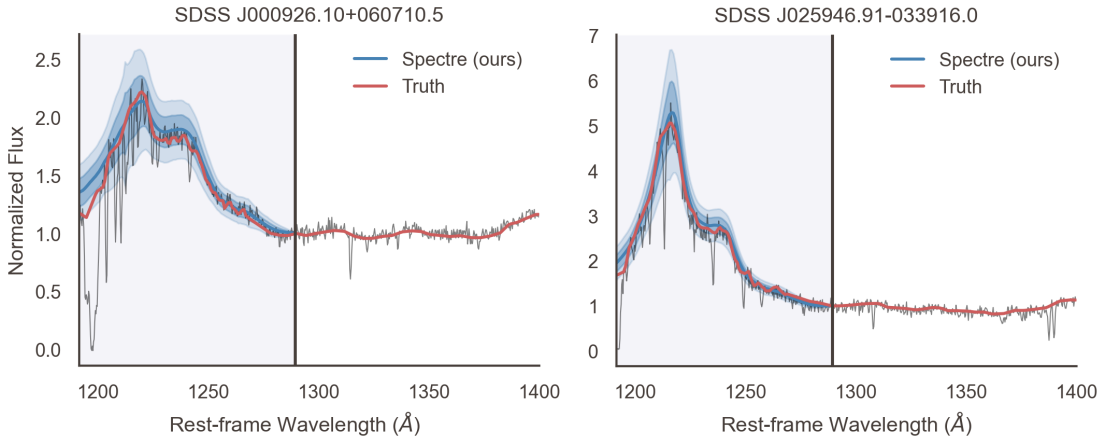

Figure 2.1: To infer the blue-side continuum, SPECTRE samples one thousand predictions conditional on the red-side. The assumed truth, shown in red, is the smoothed continua estimation from our preprocessing scheme, and the raw flux is shown in grey.

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 [55]. 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 [103, 64]. 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 [102]—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 [102], 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 [148].

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 [102], 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 [6], 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 Å $\leq \lambda_{\rm rest} <$ 1290 Å) using information encoded in the red-side spectrum ($\lambda_{\rm rest} \geq$ 1290 Å). This information arises from correlations in the emission features of the blue and red-side spectra [19]. Such models are typically trained on the spectra of quasars at moderate redshift—a typical redshift range is $z \in [2.1,\ 2.5]$ [52]—which cover the Lyα and Mg II broad emission lines. These lines strongly constrain the standard pipeline estimates for quasar redshifts [125]. 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 [52]) 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 [145, 29] or an ensemble of neural networks [37].

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 [37] or the likelihood of PCA coefficients under a Gaussian mixture model [29].

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 ($z = 7.09$) [105] and ULAS J1342+0928 ($z = 7.54$) [7], in order to infer the neutral hydrogen fraction of the universe during the epoch of reionization.

This manuscript is organized as follows: §2.2 provides a brief overview of previous approaches to intrinsic continua prediction. In §2.3 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 §2.4 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 §2.5 for our constraints on the reionization history of the early universe from measuring the damping wing in two $z > 7$ quasars. In §4.6, we conclude with a summary of our work and further discuss the use of flows for probabilistic modeling in the sciences.

2.2 Related Work

Previous approaches to intrinsic quasar continua prediction have ranged from fully Bayesian models operating purely upon emission features [52] to full-spectrum principal component analyses on blue/red-side continua to learn correlations between the dominant modes of variation in each, as in [29]; [37]. 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 [52], 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 [51, 50] 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 eigenvectors. 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 [145, 126, 39, 38, 29] or an ensemble of neural networks [37]. 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 [37]) or via full hydrodynamical modeling of the neutral IGM (as in [28]).

2.3 Background

This section describes normalizing flows in moderate detail. For an exhaustive introduction and tutorial refer to [119]—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 $T : \mathbb{R}^D \to \mathbb{R}^D$ 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 parameters $\boldsymbol{\theta}$.

2.3.1 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

[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.

target distribution p_x^*, we can model its density in the following manner:

$$p_x(\mathbf{x}) = p_u(\mathbf{u}) \left|\det J_T(\mathbf{u})\right|^{-1} \tag{2.1}$$

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 Gaussian 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.

$$\mathcal{D}_{KL}\left[p_x^*(\mathbf{x}) \parallel p_x(\mathbf{x};\boldsymbol{\theta})\right] = \mathbb{E}_{p_x^*(\mathbf{x})}\left[\log \frac{p_x^*(\mathbf{x})}{p_x(\mathbf{x};\boldsymbol{\theta})}\right] \tag{2.2}$$

$$= \mathbb{E}_{p_x^*(\mathbf{x})}\left[\log p_x^*(\mathbf{x}) - \log p_x(\mathbf{x};\boldsymbol{\theta})\right] \tag{2.3}$$

$$= -\mathbb{E}_{p_x^*(\mathbf{x})}\left[\log p_x(\mathbf{x};\boldsymbol{\theta})\right] + \text{const.} \tag{2.4}$$

$$\tag{2.5}$$

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 $\mathbf{u} = T^{-1}(\mathbf{x})$.

$$\mathcal{L}(\boldsymbol{\theta}) = -\mathbb{E}_{p_x^*(\mathbf{x})}\left[\log p_u(T^{-1}(\mathbf{x};\boldsymbol{\theta})) + \log\left|\det J_{T^{-1}}(\mathbf{x};\boldsymbol{\theta})\right|\right] \tag{2.6}$$

$$\tag{2.7}$$

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.

$$\mathcal{L}(\boldsymbol{\theta}) \approx -\frac{1}{N}\sum_{n=1}^{N}\left[\log p_u(T^{-1}(\mathbf{x}_n;\boldsymbol{\theta})) + \log\left|\det J_{T^{-1}}(\mathbf{x}_n;\boldsymbol{\theta})\right|\right] \tag{2.8}$$

$$\tag{2.9}$$

Thus, training a normalizing flow amounts to explicitly maximizing the likelihood

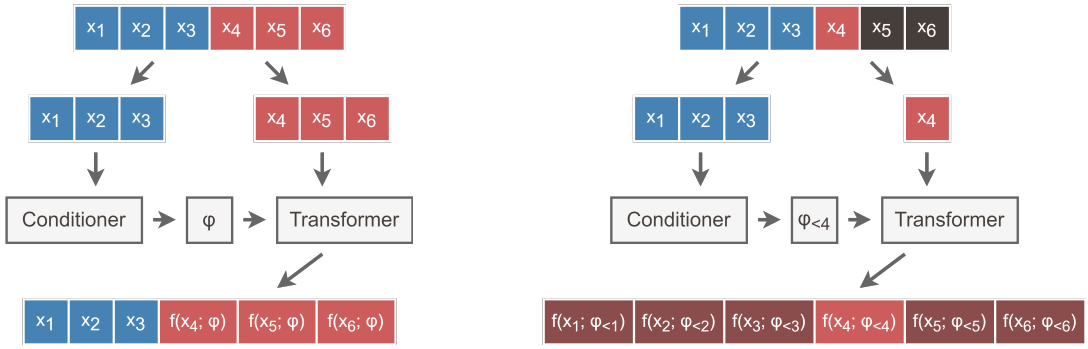

Figure 2.2: Left: Example coupling layer transform for a $D = 6$ dimensional input vector, $\mathbf{x}_{1:6}$. Upon splitting the input features into halves, the first half is consumed by the conditioner which outputs a parameter vector, $\boldsymbol{\phi}(\mathbf{x}_{1:3})$. This vector parameterizes 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 $\boldsymbol{\phi}_{<i} = \boldsymbol{\phi}(\mathbf{x}_{<i})$.

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 [80]) admit only approximate likelihoods and generative adversarial networks (GANs, see [48]) admit no likelihoods at all. Autoregressive generative models [150, 151, 153] 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.

2.3.2 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 [30]; [31] and the *autoregressive transforms* of [82]; [121]. We discuss the merits of each in turn.

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 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. 2.2, left).

$$\mathbf{y}_{1:d} = \mathbf{x}_{1:d} \tag{2.10}$$

$$\mathbf{y}_{d+1:D} = f(\mathbf{x}_{d+1:D};\ \boldsymbol{\phi}(\mathbf{x}_{1:d})) \tag{2.11}$$

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.

$$\boldsymbol{\phi}(\mathbf{x}_{1:d}) = \{\boldsymbol{\alpha}(\mathbf{x}_{1:d}),\ \boldsymbol{\beta}(\mathbf{x}_{1:d})\} \tag{2.12}$$

$$f(\mathbf{x}_{d+1:D};\ \boldsymbol{\phi}(\mathbf{x}_{1:d})) = \boldsymbol{\alpha}(\mathbf{x}_{1:d}) \cdot \mathbf{x}_{d+1:D} + \boldsymbol{\beta}(\mathbf{x}_{1:d}) \tag{2.13}$$

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 [81].

Promising recent approaches [106, 35] 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.

Autoregressive Transforms

Autoregressive transforms (see Fig. 2.2, right) enforce a lower triangular Jacobian by specifying the following form for their transforms:

$$\mathbf{y}_i = f(\mathbf{x}_i;\ \boldsymbol{\phi}(\mathbf{x}_{<i})) \tag{2.14}$$

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 [119]).

Autoregressive flows can have either one-pass sampling and D-pass density estimation or D-pass sampling and one-pass density estimation [121, 82]. 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.

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 [82] 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 [76]. 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.

2.3.3 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 [106] and subsequently up to and including degree three [34]. Recently, [35] 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 [35] 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.

$$r_k(\xi) = \frac{\alpha_k(\xi)}{\beta_k(\xi)} = y_k + \frac{\Delta y_k \left[s_k \xi^2 + \delta_k \xi(1-\xi) \right]}{s_k + \left[\delta_{k+1} + \delta_k - 2s_k \right] \xi(1-\xi)} \tag{2.15}$$

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.

$$\det J_T = \det \frac{\partial T_i}{\partial x_j} = \prod_{i=1}^{d} 1 \times \prod_{i=d+1}^{D} \frac{\partial f_i}{\partial x_i} = \prod_{i=d+1}^{D} \frac{\partial r_k^i}{\partial x_i} \tag{2.16}$$

Where the derivative of the spline is shown below.

$$\frac{dr_k}{dx} = \frac{s_k^2 \left[\delta_{k+1} \xi^2 + 2 s_k \xi(1-\xi) + \delta_k (1-\xi)^2 \right]}{\left[s_k + \left[\delta_{k+1} + \delta_k - 2s_k \right] \xi(1-\xi) \right]^2} \tag{2.17}$$

Inverting the spline is possible by inverting equation (2.15) and solving for the roots of the resulting quadratic equation.

2.4 Methods

This section describes our data preprocessing scheme (2.4.1), the implementation of SPECTRE (2.4.2), and our training (2.4.3) and model selection (2.4.4) procedures. Additionally, we describe the likeness of high-z targets to our training dataset (2.4.5) and our measurement of the red damping wing (2.4.6).

2.4.1 Data

Training Data

We adopt the data preprocessing scheme of [37], and briefly recount it here. For a detailed description, refer to [37]. 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 [125] 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 chosen redshift range include prominent emission features from Lyman-α to Mg II, which strongly constrain the SDSS redshift estimates [52].

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).

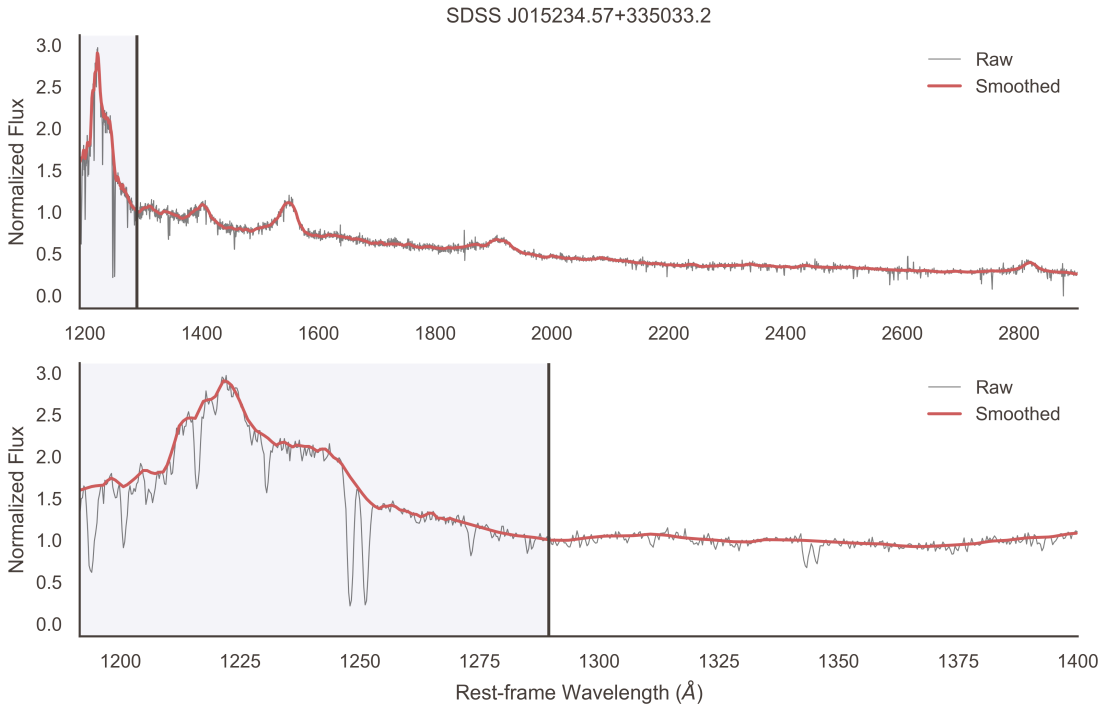

Figure 2.3: Top: an example of our preprocessing scheme. Here, we show a smoothed spectrum in red overlaid upon its raw flux counterpart in grey. The region with grey background on the left-hand side is the blue-side of the spectrum whose distribution we attempt to model conditional on the red-side of 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.

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 [42] 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 $\lambda_{\rm rest} = 1290\,\text{Å}$. 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 [37]. 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. 2.3.

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.

High-z Data

Our inference targets are two high-z quasars: ULAS J1120+0641 ($z = 7.09$) [105] observed by VLT/FORS and Gemini/GNIRS and ULAS J1342+0928 ($z = 7.54$) [7] 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 [37]. 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 Å and 2100 – 2230 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. 2.4.1.

2.4.2 Model

The SPECTRE architecture is based off of [35]'s original implementation of rational quadratic neural spline flows.

Our network employs 10 layers of spline coupling transforms, each parameterized 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 ~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 [128] implementation of SPECTRE is available on GitHub: `github.com/davidreiman/spectre`.

2.4.3 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] 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 $|\nabla_\theta \mathcal{L}| = \min(|\nabla_\theta \mathcal{L}|, 5)$. 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 [77] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

2.4.4 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.

[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.

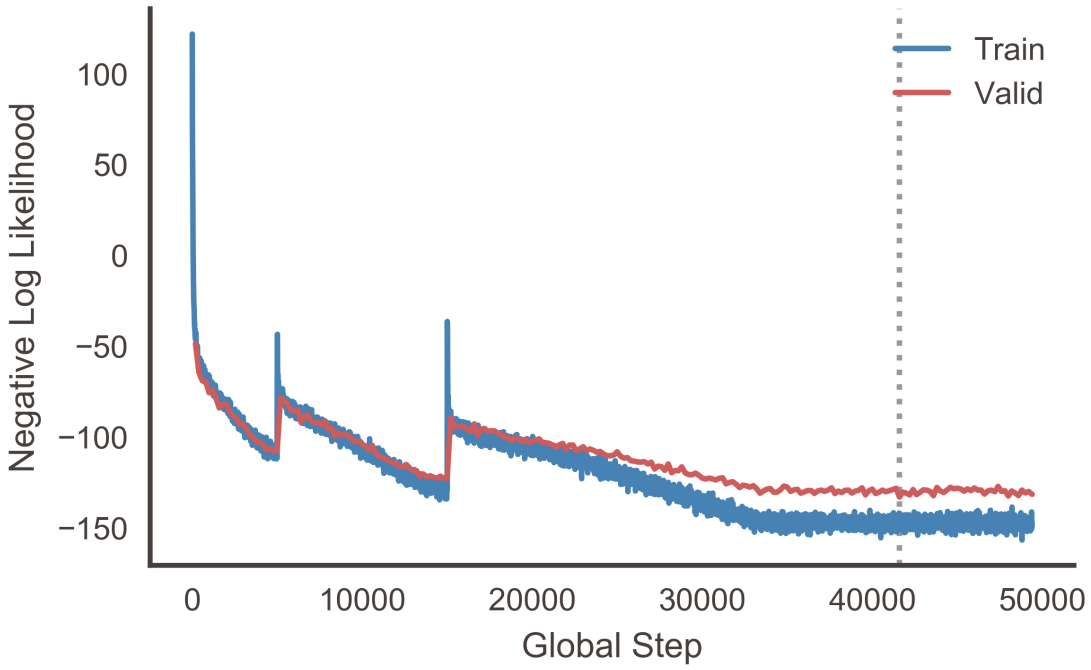

Figure 2.4: SPECTRE's training curve, showing training and validation losses (negative log likelihoods) as a function of global step, where the global step counter is incremented after each parameter update. We employ a cosine annealing learning rate schedule with warm restarts. The annealing period 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.

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 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 A.3.

2.4.5 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 [109] and [25], however, this can often fail. In [138], 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 [29].

To avoid the spurious outlier detection performance of pure likelihoods, we employ the methods of [138] to quantify the notion of a spectrum being in-distribution with the likelihood ratio:

$$\ell(\mathbf{x}) = \frac{p_{\text{in}}(\mathbf{x})}{p_{\text{out}}(\mathbf{x})} \tag{2.18}$$

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:

$$\ell(\mathbf{x}) = \frac{p_{\text{in}}(\mathbf{x}_S|\mathbf{x}_B)p_{\text{in}}(\mathbf{x}_B)}{p_{\text{out}}(\mathbf{x}_S|\mathbf{x}_B)p_{\text{out}}(\mathbf{x}_B)} \approx \frac{p_{\text{in}}(\mathbf{x}_S|\mathbf{x}_B)}{p_{\text{out}}(\mathbf{x}_S|\mathbf{x}_B)} \tag{2.19}$$

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. 2.4.1. 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 A.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 A.3 aside from the number of encoder layers since there is no conditional information for this task.

In Fig. 2.5 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 [29] 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 [37], 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.

2.4.6 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 [105] and ULAS J1342+0928 at z = 7.54 [7]. Measurement of the damping wing requires knowledge of the intrinsic emission of each quasar,

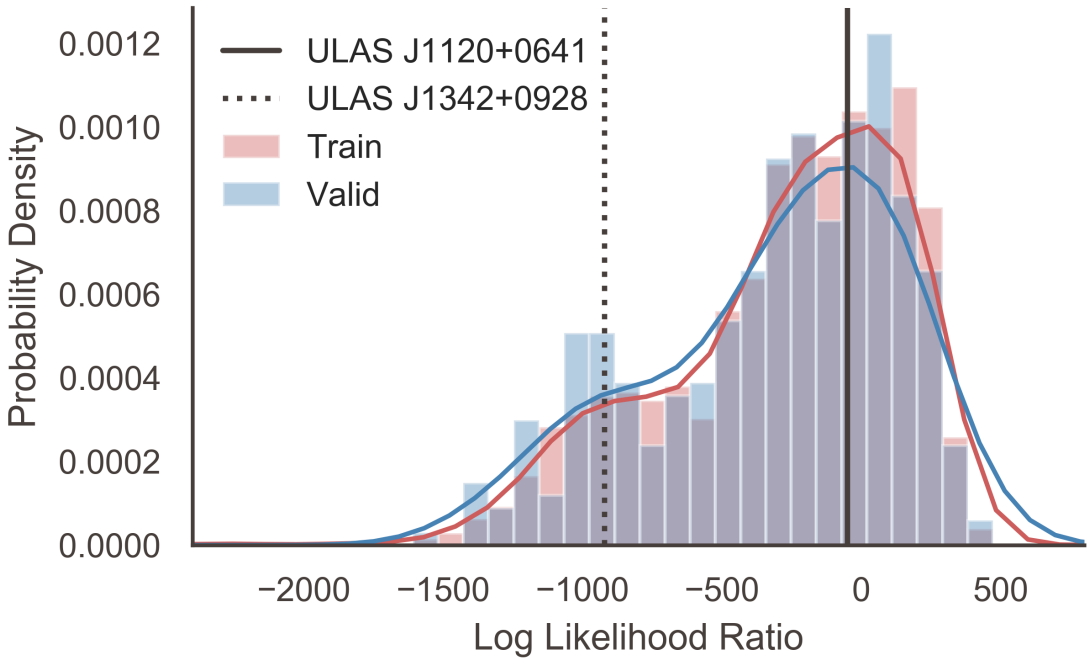

Figure 2.5: We evaluate the likelihood ratios of our training/validation sets and high-z spectra as calculated by two autoregressive rational quadratic neural spline flows. The most stringent out-of-distribution bounds result when one flow is trained on smoothed continua and the other flow is trained 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.

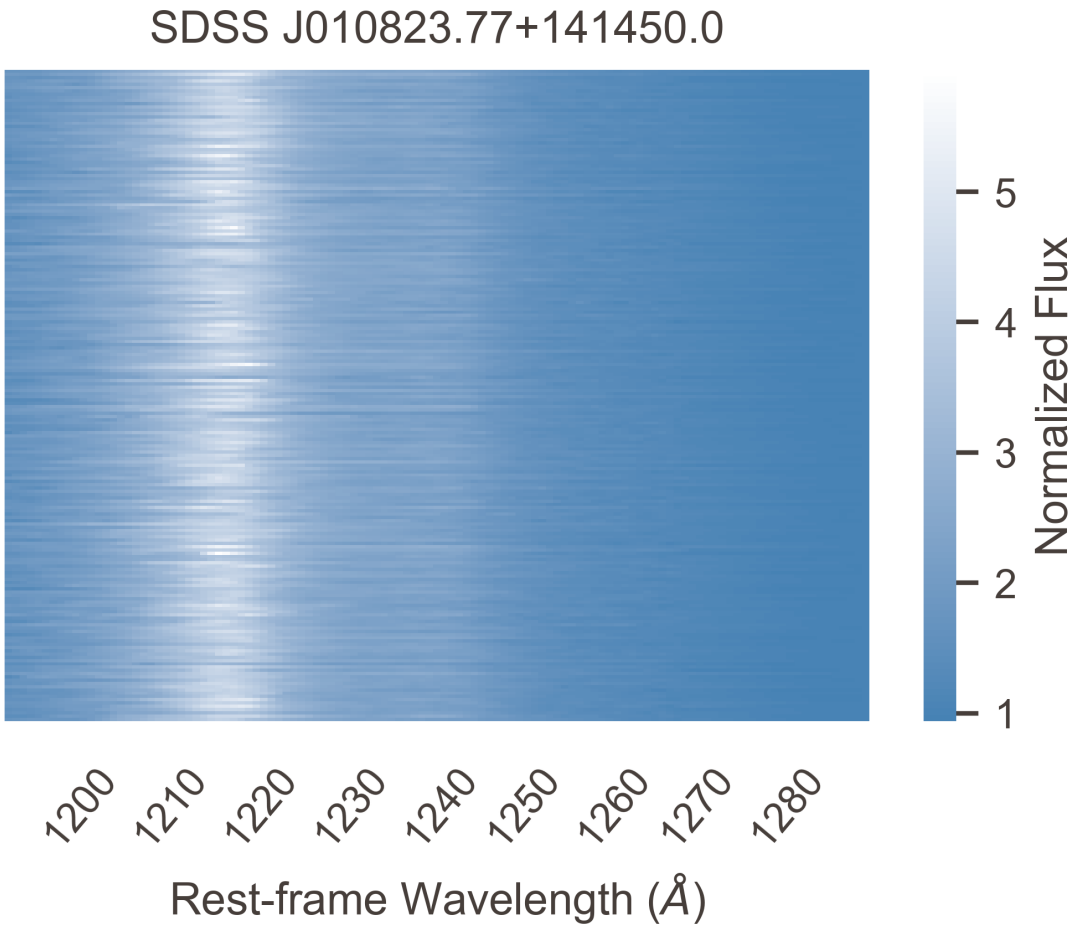

Figure 2.6: A selection of 200 blue-side continuum predictions from SPECTRE for a randomly chosen quasar. Each row corresponds to a single blue-side sample from our model, color-coded to designate its normalized flux at each wavelength. Qualitatively, the model's predictions are consistent with a Lyα peak near 1215.67 Å and a N v emission near 1240.81 Å .

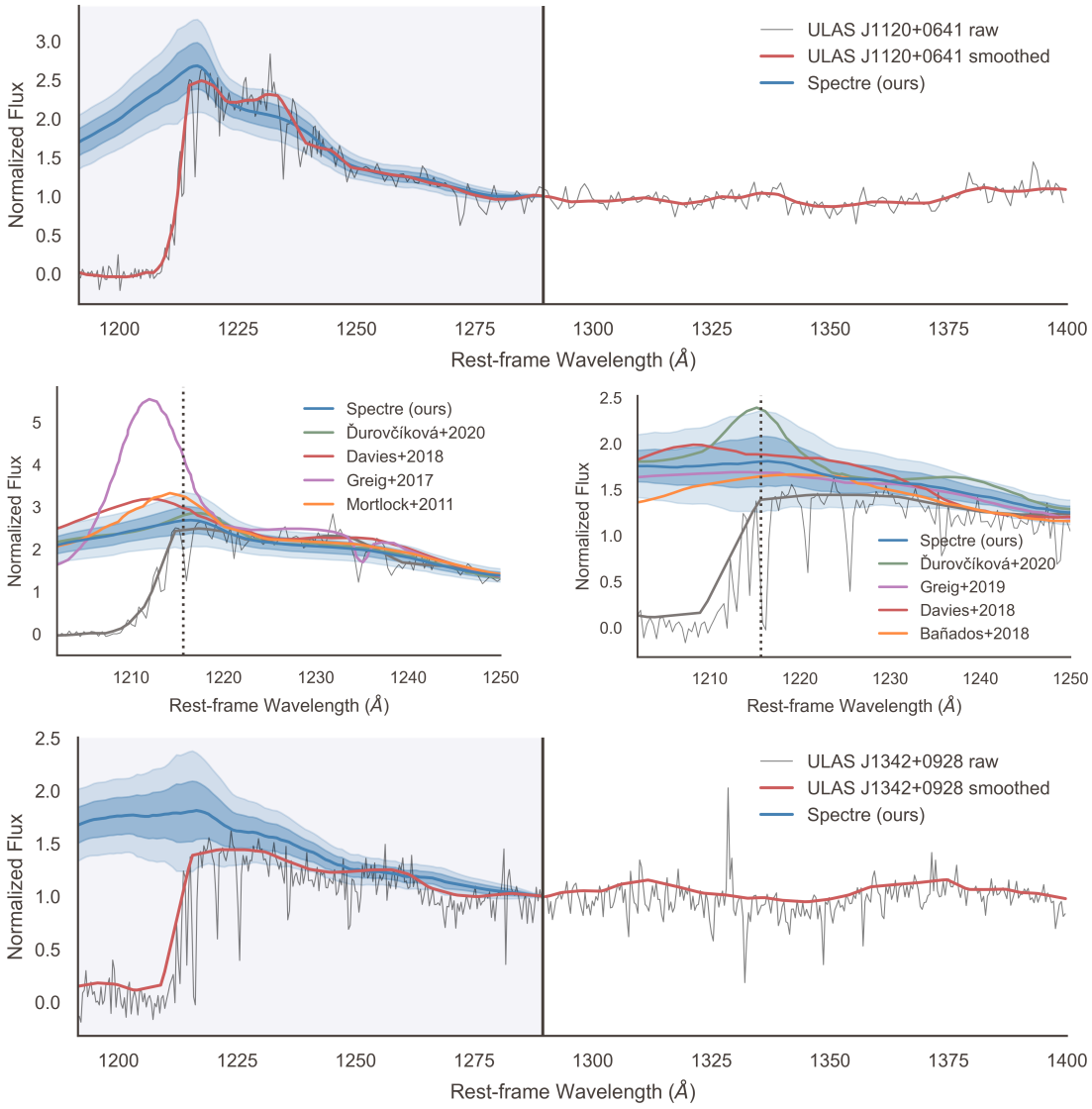

Figure 2.7: Blue-side continua predictions on two high redshift quasars, ULAS J1120+0641 and ULAS J1342+0928. Each solid blue line is the mean of one thousand samples from SPECTRE. The blue and light blue bands reflect one and two standard deviation bands at each wavelength, respectively. **Top**: Our mean prediction with two sigma uncertainty bands overlaid on top of continuum and raw flux measurements of ULAS J1120+0641. **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.

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 [102]:

$$\tau(\Delta\lambda) = \frac{\tau_{\rm GP} R_\alpha}{\pi}(1+\delta)^{3/2}\int_{x_1}^{x_2}\frac{x^{9/2}}{(1-x)^2}dx \tag{2.20}$$

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] $R_\alpha = 2.02\times10^{-8}$, and the Gunn-Peterson optical depth of neutral hydrogen, $\tau_{\rm GP}$, is given by [41] as follows:

$$\tau_{\rm GP}(z) = 1.8\times10^5 h^{-1}\Omega_m^{-1/2}\left(\frac{\Omega_b h^2}{0.02}\right)\left(\frac{1+z}{7}\right)^{3/2}\bar{x}_{\rm HI} \tag{2.21}$$

The integral in Eqn. 2.20 is solvable analytically and its solution is provided in [102] as:

$$I(x) = \frac{x^{9/2}}{1-x} + \frac{9}{7}x^{7/2} + \frac{9}{5}x^{5/2} + 3x^{3/2} + 9x^{1/2} - \frac{9}{2}\log\frac{1+x^{1/2}}{1-x^{1/2}} \tag{2.22}$$

We adopt the Planck 2018 cosmological parameters [132] 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 ~1210 Å.

[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 [102].

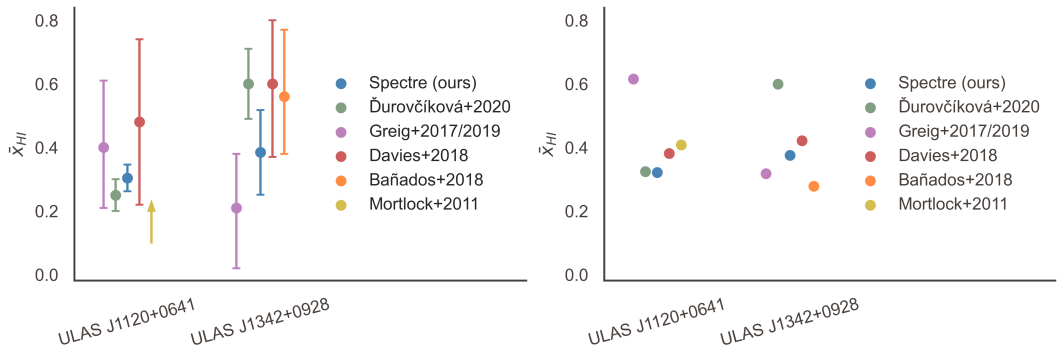

Figure 2.8: A comparison of our estimates of the volume-averaged neutral fraction of hydrogen for ULAS J1120+0641 ($z = 7.09$) and ULAS J1342+0928 ($z = 7.54$). **Left**: Reported results from the literature. These make use of different damping wing models (and some employ full hydrodynamical models of the IGM) which complicates direct comparison. **Right**: 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 [102]. Error bars are omitted since only mean continuum predictions are used.

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 $\lambda_{\rm rest} \in [1210\,\text{Å}, 1250\,\text{Å}]$.

$$\tau_{\rm obs} = -\ln\left(\frac{F_{\rm obs}}{F_{\rm int}}\right) \tag{2.23}$$

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:

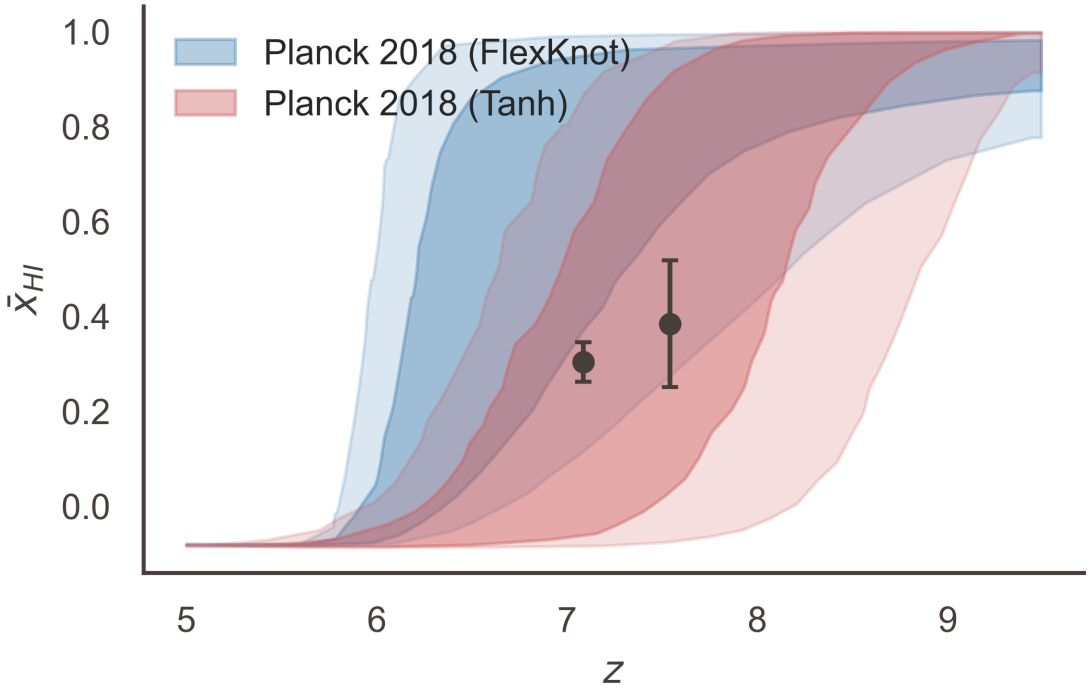

Figure 2.9: A comparison of our estimates of the volume-averaged neutral fraction of hydrogen to the Planck constraints [132]. Planck employs two models for their reionization constraints: the *Tanh* model which assumes a smooth transition from a neutral to ionized universe based on a hyperbolic tangent 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.

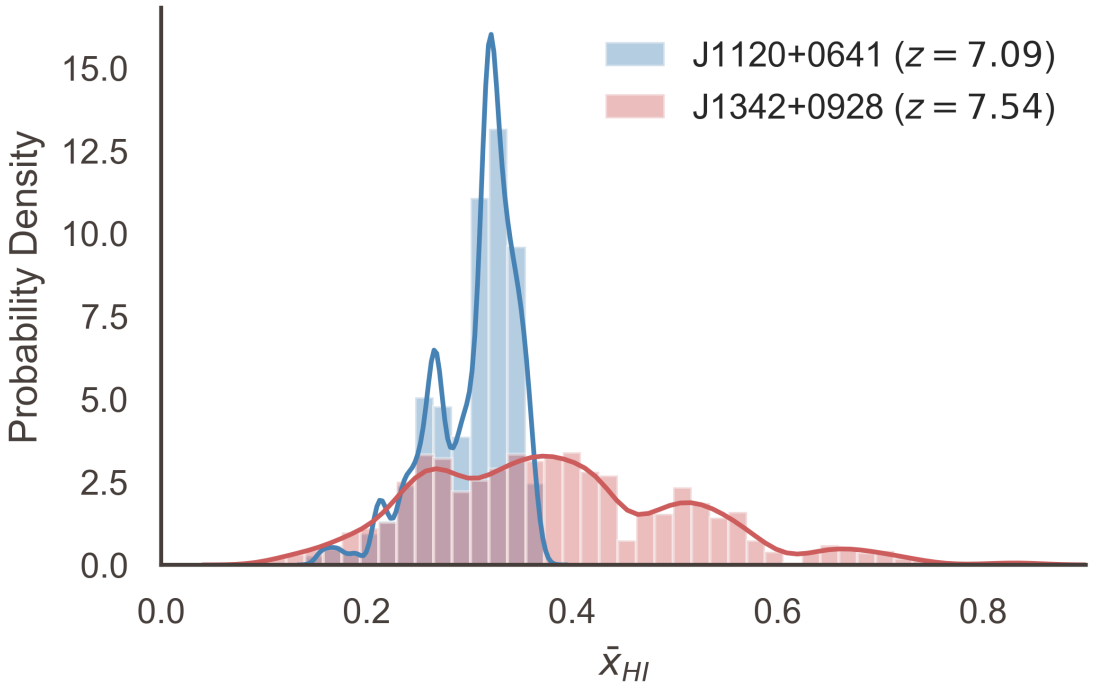

Figure 2.10: Histograms and kernel density estimates depicting the spread in neutral hydrogen fraction predictions over 10000 samples from SPECTRE. From observations of ULAS J1120+0641 we infer $\bar{x}_{\rm HI} = 0.304 \pm 0.042$ while for ULAS J1342+0928 we infer $\bar{x}_{\rm HI} = 0.384 \pm 0.133$.

$$\tau_{\rm obs}(\Delta\lambda) \sim \mathcal{N}(\tau_{\rm GP} R_\alpha/\pi(\Delta\lambda/\lambda)^{-1},\ \sigma_\tau^2) \tag{2.24}$$

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

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. 2.10).

2.5 Results

2.5.1 Reionization History Constraints

We display our predictions of the intrinsic continua of J1120+0641 and J1342+0928 in Fig. 2.7. 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 [37], 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. 2.4.6, 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. 2.8. We also display our results for the volume-averaged neutral fraction of hydrogen in the context of the Planck constraints [132] in Fig. 2.9.

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 [102]. The results are presented on the right of Fig. 2.8 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.

2.5.2 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:

$$\epsilon_c = (F_{\text{model}} - F_{\text{truth}})/F_{\text{truth}} \tag{2.25}$$

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 [37] as it omits the absolute value on the residual term.

In Fig. 2.12, 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 [29] presented in [37] 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 [29], 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 [37], 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. 2.13 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 ~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

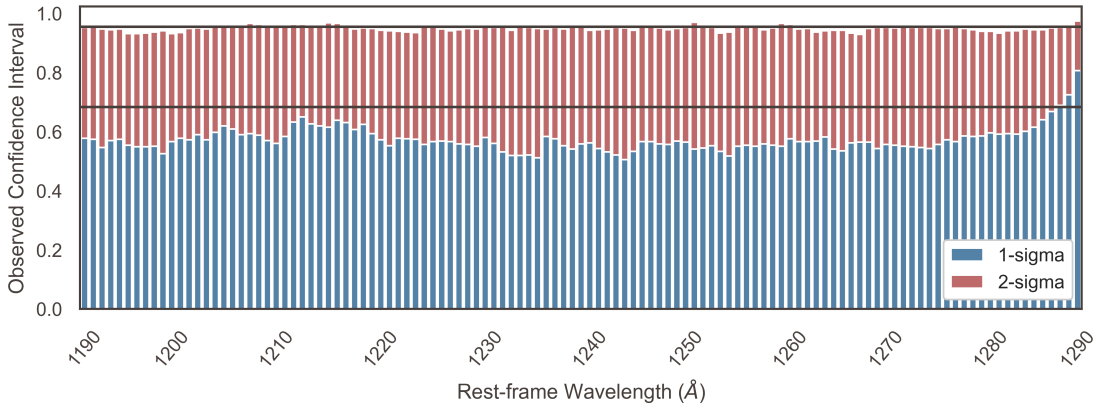

Figure 2.11: Observed confidence intervals as a function of blue-side wavelength. A calibrated model would make predictions of the marginal density over flux in a given wavelength bin such that P% of the observed absolute errors fall within the P% credible interval. We approximate the marginals as Gaussian (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.

and Time (LSST) at the Vera C. Rubin observatory will provide millions of additional quasar spectra [72] which will likely significantly improve SPECTRE's performance.

Additionally, Appendix A.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`.

2.5.3 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:

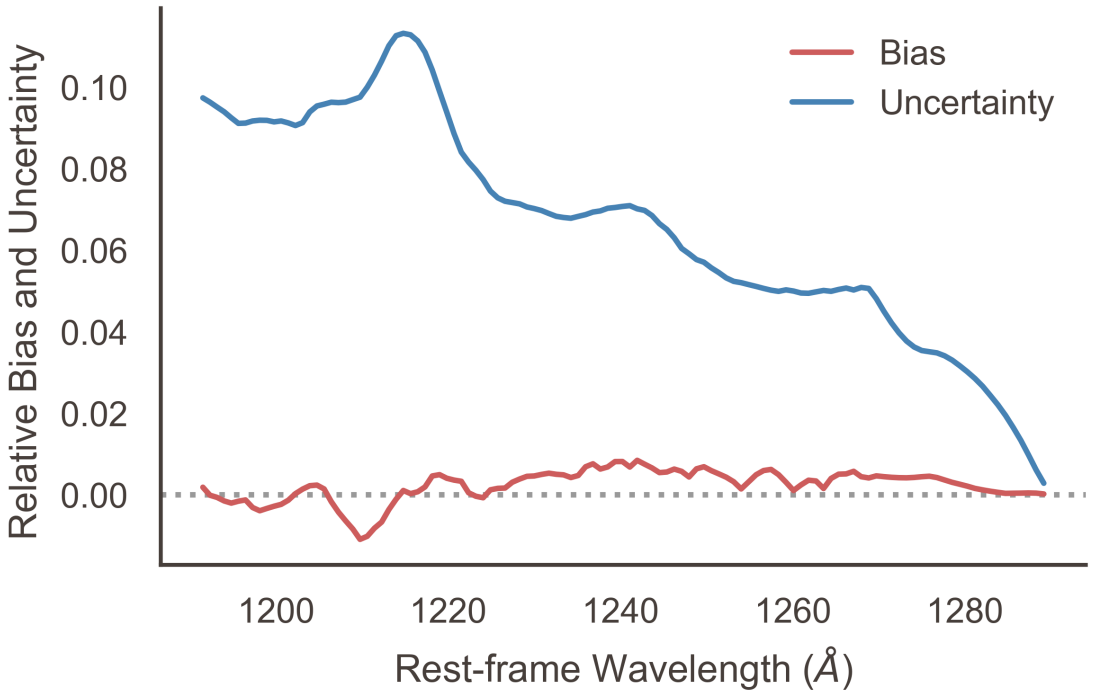

Figure 2.12: The relative prediction bias $\langle \epsilon_c \rangle$ and uncertainty $\sigma(\epsilon_c)$ (see Eqn. 2.25) as a function of blue-side wavelength averaged over our test set. Our average relative prediction uncertainty across all blue-side wavelengths is 6.63%, though we note that this metric is highly sensitive to the preprocessing scheme and is therefore difficult to compare to other methods.

Figure 2.13: 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 [37] (whose original formulation was introduced in [29]). 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).

$$\sigma_j = \sqrt{\sum_{i=1}^{N} \frac{(x_{ij} - \bar{x}_j)^2}{N}} \tag{2.26}$$

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 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. 2.11. 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. 2.15. Though we note a sparsely populated tail of underestimated error (an effect also noted in Fig. 2.11), 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 hyperparameter, 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 [56, 127]. 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 [117].

2.5.4 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. 2.14. 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.

2.6 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

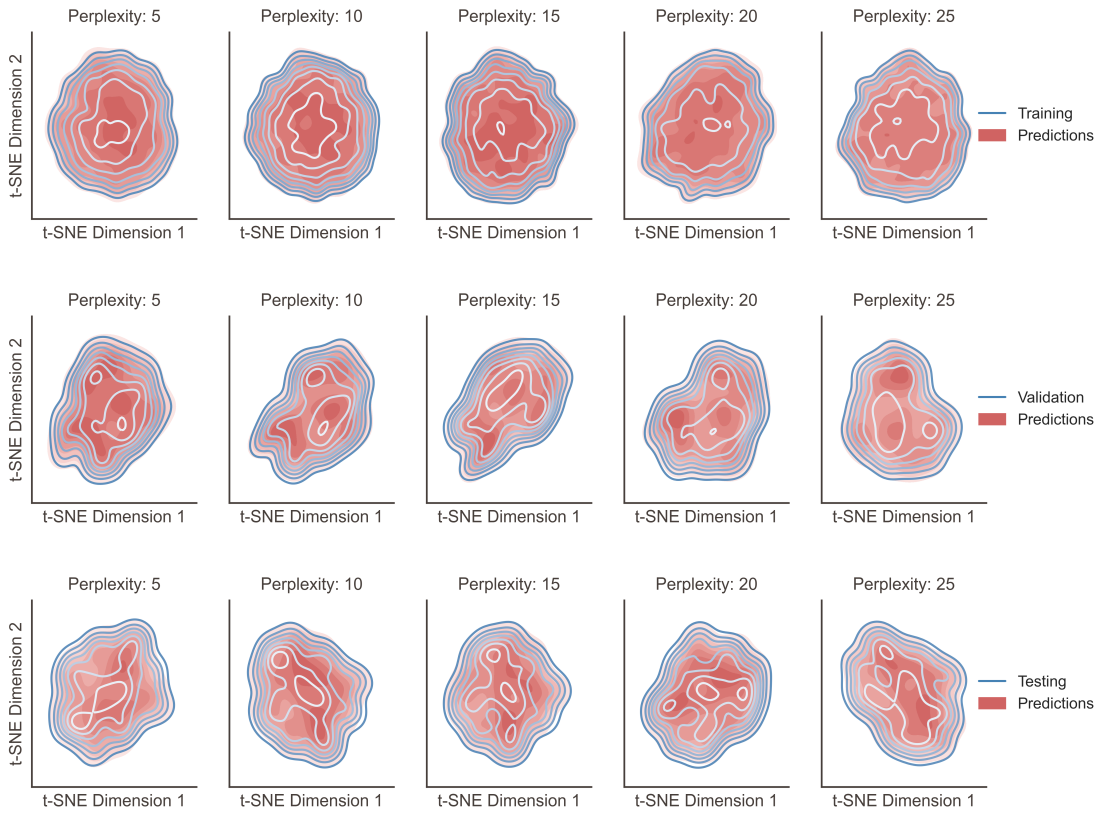

Figure 2.14: Two-dimensional embedding of blue-side spectra produced via t-Distributed Stochastic Neighbor Embedding (t-SNE). **Top**: Comparisons between our training set continua and associated model predictions. **Middle**: Comparisons between validation set continua and associated model predictions. **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.

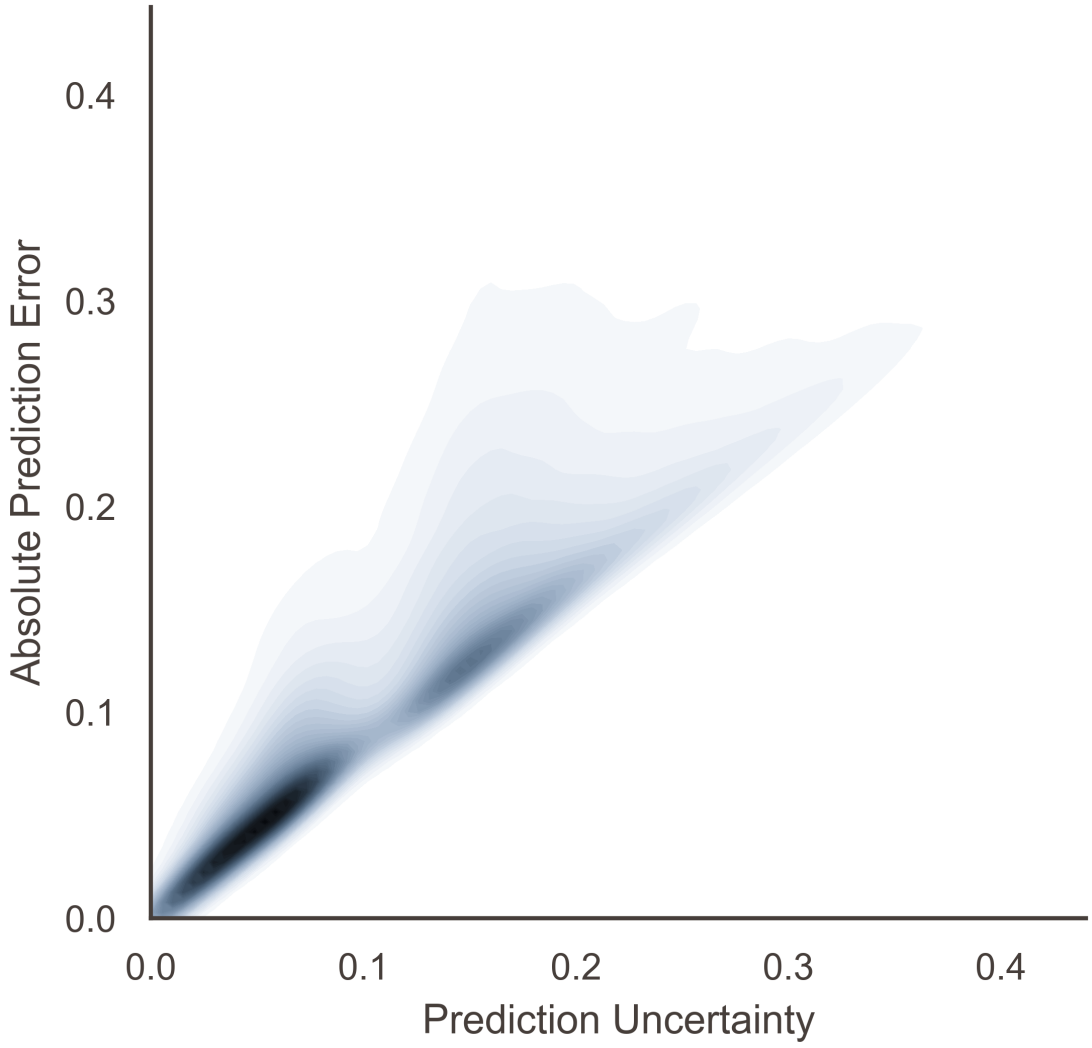

Figure 2.15: Kernel density estimate of the joint distribution over absolute error and predictive uncertainty in SPECTRE samples over all wavelengths and spectra in the test set.

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 [123]. 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 [47] and in the estimation of the expectation values of physical observables in lattice quantum chromodynamics [76].

In general, flows are capable of density estimation for a wide variety of continuous or discrete-valued data. They have been successfully applied as generative models for images [81] and text [149] and used to perform anomaly detection in particle physics [107].

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 [28] 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.

Chapter 3

Via Machinae: Searching for Stellar Streams using Unsupervised Machine Learning

3.1 Introduction

Stellar streams, the tidally-stripped remnants of dwarf galaxies and globular clusters, provide a unique window into the properties of the Milky Way and its formation history. Streams trace the historical record of the mergers that built the Milky Way [73, 61, 13, 60, 12, 98]. Their orbits allow measurements of the underlying gravitational potential of the Milky Way [75, 69, 83, 113, 152, 141, 86, 93, 137]. The presence of gaps and density perturbations within streams can inform the population of dark matter substructure, and subsequently the properties of dark matter [23, 142, 40, 18, 8, 17]. They can also be used to empirically track the underlying distribution of dark matter [136, 112].

Starting with the Sloan Digital Sky Survey (SDSS) [154], numerous surveys have increased the number of cataloged stellar streams [116, 114, 53, 13, 144]. Most recently,

the *Gaia* Space Telescope [45, 87] has opened a new frontier of Galactic kinematics and thus new opportunities for the discovery and study of stellar streams.

Numerous successful stream-finding techniques have been applied to the *Gaia* data [92, 95, 155, 99, 20, 100, 70]. In some cases cross-referencing *Gaia* with other spectroscopic catalogs can provide additional kinematic or spectroscopic information, although statistically limiting the sample size (see e.g. StarGo [155], which identifies streams in the cross match of *Gaia* DR2 with LAMOST DR5 [91]). Of the methods relying exclusively on *Gaia*, the Streamfinder algorithm [92, 95] leverages the fact that stars within a stellar stream would have similar orbits through the Galaxy. By searching for stars occupying the same "hypertubes" through six-dimensional position/velocity space, Streamfinder has discovered a number of new stellar streams [96, 71, 94, 70]. In order to construct these orbits, Streamfinder must assume a form for the Galactic potential, and search for stars on an isochrone as part of a kinematically cold stream.

In this paper, we present Via Machinae, a new algorithm for automated stellar stream searches with *Gaia* data. Based on unsupervised machine learning techniques, we identify streams as local overdensities in the angular position, proper motion, and photometric space of stars in *Gaia* DR2. Importantly, *we do not assume the stars in question lie on a particular orbit or stellar isochrone.* In fact, the initial (and most computationally-intensive) machine learning training steps of Via Machinae are designed to find *all anomalous* structures first, in an agnostic manner. Only then do we implement selections based on prior knowledge of the properties of known stream candidates (particularly that the stars are distributed in an approximately linear structure over small angles on the sky). Such choices can be modified to target structures with other distributions in stellar photometry and proper motion, for example globular clusters or debris flow [88, 85].[1] This flexibility may allow our technique to be sensitive to a

[1] Debris flow refers to structure localized in velocity space, but incoherent in physical space [61, 89, 85]. This is usually the case for older mergers, e.g. the *Gaia* Sausage/Enceladus [111].

wider variety of stellar streams than previous methods, and can be generalized to other anomalous features within the *Gaia* dataset (or other astrophysical surveys).

VIA MACHINAE has two main components: an anomaly finding algorithm, and a line finding algorithm. The first component is the ANODE (ANOmaly detection with Density Estimation) algorithm ([108] hereafter referred to as [NS20]). Originally developed to search for new physics at the Large Hadron Collider, ANODE is a general machine learning algorithm for finding localized overdensities in any dataset. To accomplish this, ANODE leverages recent advances in density estimation using neural networks, specifically the idea of *normalizing flows* (for a recent review and original references, see e.g. [120]). In this paper, following the original ANODE work [NS20], we use Masked Autoregressive Flows (MAF) [122] to estimate the probability densities of stars in the *Gaia* dataset.

The ANODE algorithm begins by slicing up the dataset into search regions and their complements, the control regions. As kinematically cold streams are expected to be fully localized in both proper motions, we choose to split the dataset into search regions consisting of slices in one of the proper motion coordinates. Then we use the MAF to estimate the probability distribution in position/proper motion/color/magnitude space of the stars in each search region in two different ways: (1) directly with the stars in the search region; and (2) indirectly with the stars in the control region, followed by interpolation into the search region. The interpolation step is a "free" byproduct of the density estimation, because we actually learn a *conditional* probability density conditioned on the proper motion used to define the search region. If the search region contains a stream while the control region does not, then (2) can be thought of as a data-driven estimate of the probability density of the "background" (i.e. non-stream) halo stars in the search region. Taking the ratio of these two density estimates forms a discriminant R, which is sensitive to anomalous overdensities (or underdensities) in

the search region. By selecting the stars with the largest likelihood ratios, we can preferentially enhance the presence of stream stars vs. background stars in any given search region.

After performing such a selection in each search region, we are left with a much reduced set of stars spread across the sky. Only some of these stars will correspond to stellar streams. The rest may be other interesting structures (e.g. globular clusters or debris flow) or spurious false positive fluctuations of the ANODE algorithm. This leads to the second major component of the VIA MACHINAE algorithm: an automated method to search for linear features in a collection of stars in an angular patch of the sky. Simply fitting the stars to a line using (for example) least squares regression yields extremely unsatisfactory results, owing the presence of noise and outliers (i.e. in a collection of stars, only a small fraction might belong to the stream). Instead, we have developed a method based on the Hough transform. This is an age-old machine learning technique that was originally developed for finding lines and edges in photographs [67, 33], but which we adapt here to accomplish the same purpose in scatter plots.[2] The idea of the Hough transform is to convert the problem of line finding to counting intersections of curves in an auxiliary parameter space (the Hough space). In this way, one can also give a (rough) figure-of-merit to the best-fit line detection, based on the contrast between regions of high and low curve density in Hough space.

The major steps and key terms of VIA MACHINAE are summarized in Fig. 3.1. Moving from left to right in this figure:

- We divide the sky into overlapping *patches* of stars, each a circular region of radius 15°.

- These patches are then divided into overlapping *search regions* based on one proper motion coordinate. The estimated probability ratio R for each star in each

[2]The Hough transform has also been proposed for stellar stream identification in [130, 129] in the context of M31.

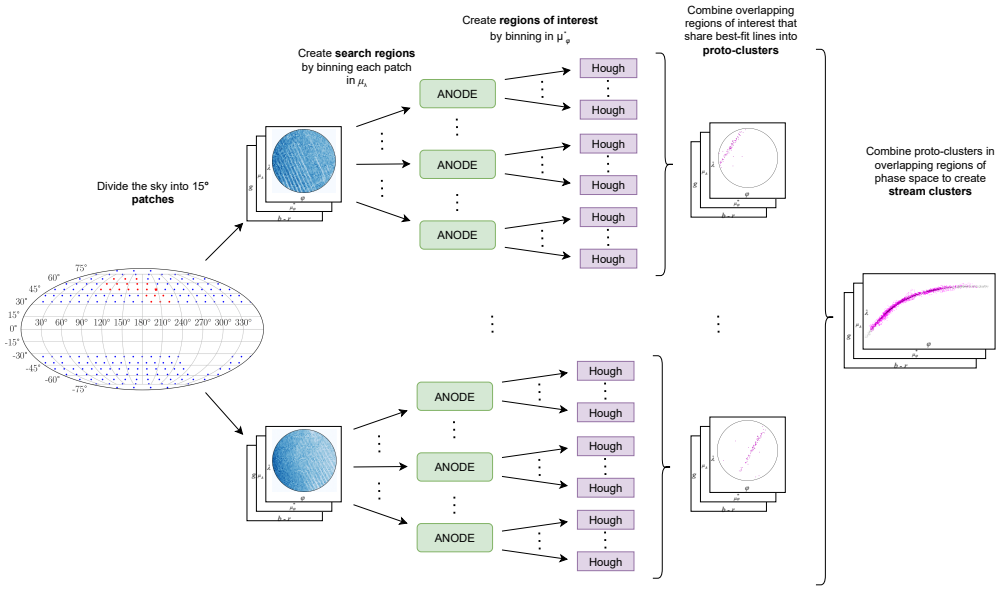

Figure 3.1: A schematic showing an overview of the VIA MACHINAE algorithm. Bolded and boxed terms are defined in Sec. 3.3 (with the exception of *patches*, which are described in Sec. 3.2). First we divide up the sky into evenly-tiled 15° patches. Within each patch, we further divide up the stars into search regions defined by a window in μ_λ, one of the proper motion coordinates (the remaining data features for each star are denoted $\vec{x}$). Then we train the ANODE algorithm on the search regions and their complements, to learn a data-driven measure of local overdensities $R(\vec{x})$. To turn this measure into a stream finder, we further divide up the SRs into regions of interest based on the orthogonal proper motion coordinate μ^*_ϕ. We apply an automated line-finding algorithm based on the Hough transform to the 100 highest-R stars in each ROI. Finally, we combine ROIs adjacent in proper motion that have concordant best-fit line parameters into proto-clusters, and cluster these across adjacent patches of the sky into stream candidates.

search region is obtained by ANODE training. We then limit ourselves to the inner 10° of the patch to avoid edge effects (among other fiducial cuts).

- Each search region is then subdivided into *regions of interest* using the orthogonal proper motion coordinate which was not used to define the search region. In order to further purify signal to noise, a cut on color is imposed to focus on old, metal-poor stars that comprise the majority of known streams.

- The 100 stars with the highest R values in each region of interest are mapped to Hough space and the most line-like feature is assigned a significance σ_L.

- In overlapping regions of interest, we combine coincident lines and σ_L values to obtain a proto-cluster for the patch, with an accompanying total significance $\sigma_L^{\rm tot}$.

- Proto-clusters in neighboring patches are combined into a stream candidate.

In this paper, we will use the GD-1 stream to illustrate the steps of the Via Machinae algorithm. GD-1 [54], is an exceptionally long and dense stellar stream located at ~ 10 kpc, most likely originating as a globular cluster of mass $\sim 2 \times 10^4 M_\odot$ [83]. When first detected using SDSS, GD-1 was thought to span $\sim 60°$ in the sky. Using the second data release of *Gaia* (*Gaia* DR2), it has been extended by as much as 20° [135] (hereafter [PWB18]), and was found to include gaps that could be evidence for dark matter substructure [135, 18, 9, 97, 10]. Though most stellar streams are not nearly as long, dense, narrow, or well-defined as GD-1, it nevertheless provides an excellent testbed for Via Machinae, as stellar membership of the stream has been extensively studied (see e.g. [135, 18, 17]), and its distinctiveness allows for clear demonstrations of the utility of the algorithm.

This paper is organized as follows: In Sec. 3.2, we introduce the *Gaia* data and its processing into inputs that will be used for anomaly detection. We then present the algorithm in Sec. 3.3, with each step illustrated by its action on a segment of the

GD-1 stream. In Sec. 3.4, we apply Via Machinae to the entire length of the GD-1 stream. Finally, in Sec. 3.5 we conclude with a summary and a list of interesting future directions motivated by this work. In a subsequent work [143], hereafter Paper II, we will apply our technique across the full *Gaia* DR2 dataset, and demonstrate its ability to detect other known streams, and present new stream candidates.

3.2 Data and Input Variables

Before introducing the Via Machinae algorithm, we must first describe the data upon which it will be applied, and the pre-processing required.

Starting with the *Gaia* DR2 dataset,[3] we limit ourselves to distant stars with measured parallax less than 1 mas (corresponding to stars beyond 1 kpc). We do not correct for the *Gaia* DR2 zero-point parallax offset; varying the parallax cut by ±0.05 results in only a $\sim 3\%$ change in the number of stars and so is highly unlikely to affect our algorithm. We tile the sky with 15° patches using Healpy [49, 156] (with nside = 5). This patch size was selected to have a tractable number of stars for the machine learning training step of the algorithm, as will be described in Sec. 3.3.2. The patches are also large enough to capture significant portions of most known streams if they should pass through them. As stars in the Galactic disc would overwhelm the training, we limit the analysis to high Galactic latitudes $|b| > 30°$. We also exclude all patches that overlap with the LMC or SMC. The final result is 200 patches in total.

For stars within a patch, our data consists of two position, two kinematic, and two photometric parameters: the angular position on the sky (e.g., right ascension [ra, α] and declination [dec, δ]), the corresponding angular proper motions ($\mu_\alpha \cos\delta$ and μ_δ),

[3] As this work was being completed, *Gaia* EDR3 [46] was released. While our results likely would have been improved by using this new dataset, re-running the ANODE method on *Gaia* EDR3 proved to be too computationally expensive (the full-sky scan of *Gaia* DR2 took $O(10^5)$ NERSC-hours). We plan to apply our method to *Gaia* EDR3 in a future publication.

the magnitude of the star in the *Gaia* G-band (g), and the difference in the G_{BP} and G_{RP} *Gaia* bands ($b - r$). Throughout this work, we will not correct for dust or extinction; especially since we confine ourselves to high Galactic latitudes, these corrections are generally small ($\lesssim 0.1$ for $b - r$) and do not vary much across a patch. Since we are only interested in local overdensities in each patch and will select a wide range of color for our final analysis, dust and extinction corrections should not significantly affect our results.

The (α, δ) coordinates do not have a Euclidean distance metric across the sky, and the resulting distortions across the patch, especially at high latitudes, could negatively affect our neural density estimation.[4] Therefore, for each patch (defined by a circle centered on $(\alpha, \delta) = (\alpha_0, \delta_0)$ in angular position), we rotate the positions and proper motions using Astropy [5, 4] into a new set of centered longitude and latitude coordinates (ϕ, λ) so that $(\alpha_0, \delta_0) \to (0^\circ, 0^\circ)$. The unit vectors for the rotated coordinate system, $(\hat{\phi}, \hat{\lambda})$, are aligned with those of the previous unit vectors $(\hat{\alpha}, \hat{\delta})$. Within each patch, we will calculate angular distances using a simple Euclidean metric in (ϕ, λ). For notational simplicity, we will define the new proper motion coordinate $\mu_\phi \cos\lambda$ as μ_ϕ^* for the remainder of the work (similarly $\mu_\alpha^* \equiv \mu_\alpha \cos\delta$).

The patches defined above will be used as input for the ANODE method, as will be described in Sec. 3.3, using the features $(\phi, \lambda, \mu_\phi^*, \mu_\lambda, b - r, g)$. After training ANODE on each patch, we impose a set of additional fiducial cuts on the data. As we will describe in more detail in Sec. 3.3.2, these cuts are driven by the limitations of the MAF density estimator. Specifically, to avoid edge effects in the neural network output, the post-ANODE fiducial region studied in this paper is the inner 10° of each patch with a magnitude cut of $g < 20.2$. Above this magnitude cut, the completeness drops rapidly [21]; this choice also helps reduce (but does not completely eliminate) streaking in the

[4] Density estimation on spheres and other non-Euclidean manifolds is an active area of research, see e.g. [139]. We do not use these techniques in this work.

data and other artifacts due to incomplete coverage of the dimmest stars in the *Gaia* DR2 data [45].

As described in the Introduction, in this work we focus on demonstrating the Via Machinae algorithm using GD-1 as a worked example. Therefore, we limit ourselves here to patches of the sky that are known to contain portions of the GD-1 stream. We find that 21 patches in our all-sky sample include stars which have been identified by [PWB18] as possible members of the GD-1 stream, for a total of 1,985 candidate GD-1 stars. Before (after) the ANODE fiducial cuts, the patches containing GD-1 have various numbers of stars, ranging from 8.0×10^5 (2.7×10^5) in the patch with the least number of stars, to 2.1×10^6 (7.0×10^5) stars in the patch with the most number of stars. Fig. 3.2 shows the locations of all 200 patch centers we use to tile the sky as well as the 21 patches containing GD-1 stars.

We will use the stream membership labels of [PWB18] (which can be downloaded at [134]) as our point of comparison throughout this work. These were derived through relatively simple means: a visual inspection of the data, combined with polygonal cuts on proper motion, color and magnitude and a parallel strip cut (the "stream track") in angular position. Thus we do not take them as "absolute truth" labels – indeed, some level of background contamination within this sample is certainly visible by eye. Nevertheless, the GD-1 candidate labels of [PWB18] still furnish a very useful and powerful point of comparison.

In Sec. 3.3, we will use one of these 21 patches containing GD-1, centered on $(\alpha_0, \delta_0) = (148.6°, 24.2°)$, to provide a worked example of each stage of Via Machinae. Within this patch's fiducial region, there are 334,376 stars, of which 276 have been identified as candidate members of GD-1 by [PWB18]. The position, proper motion, and photometry of these stars is shown in Fig. 3.3. In this patch, the candidate GD-1 stars lie in the range $\mu_\lambda \in [-14.6, -8.6]$ mas/yr.

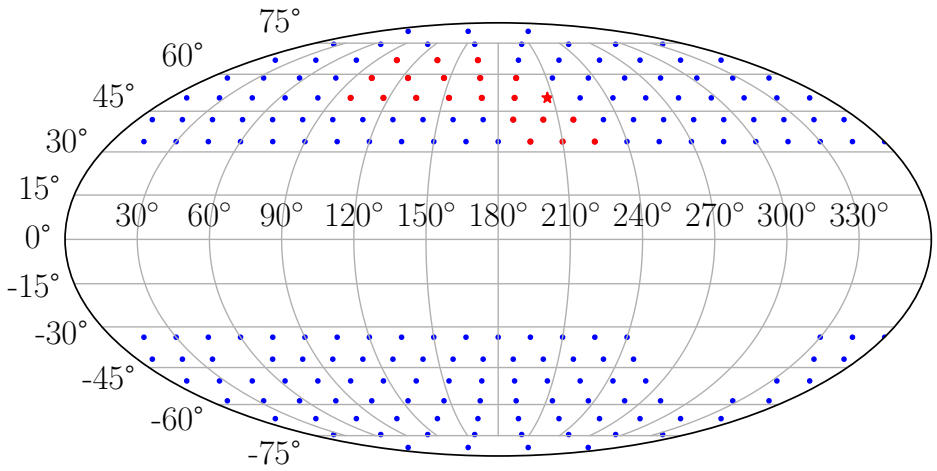

Figure 3.2: The positions in Galactic ℓ and b coordinates used for the centers for the datasets from the *Gaia* DR2 used in our full-sky analysis. The missing grid centers in the Galactic Southern hemisphere are the patches that overlapped with the Magellanic Clouds. The 21 centers which contain the GD-1 stream are shown in red, and the patch used as the worked example in Sec. 3.3 is denoted with a star.

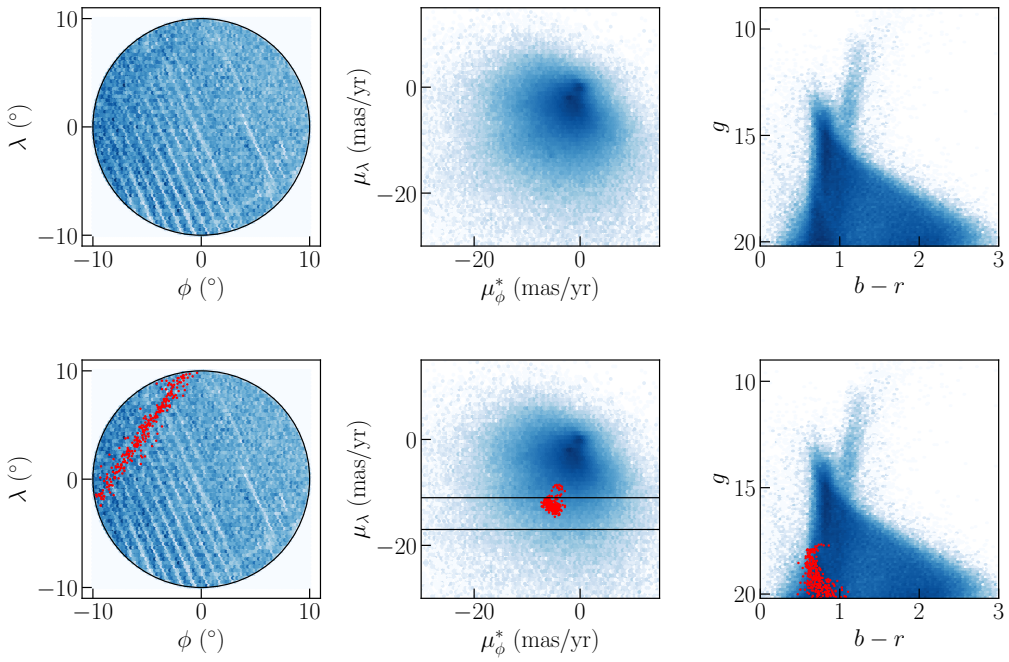

Figure 3.3: Upper row: Angular position in (ϕ, λ) coordinates (left), proper motion in $(\mu_\phi^*, \mu_\lambda)$ coordinates (center), and photometry (right) of all stars in the patch centered on $(\alpha, \delta) = (148.6°, 24.2°)$. (Note the streaking in angular position due to non-uniform coverage in *Gaia* DR2.) Bottom row: As above, with stars identified by [PWB18] as likely GD-1 stars shown in red, along with an example search region $\mu_\lambda \in [-17, -11]$ mas/yr in proper motion.

3.3 Via Machinae: The algorithm

3.3.1 ANODE: Defining the search regions

As described in the Introduction, the first part of Via Machinae is based on the ANODE method [NS20]. The starting point of ANODE is the subdivision of the stars within a single patch into *search regions* (SRs) which are windows in one feature of the dataset. The complement of the search region is called the *control region* (CR). The feature and the width of the window should be chosen so that, if a stream is present, there exists (at least) one SR which fully (or nearly fully) contains the entire stream. As we will explain in the next subsection, this is to enable accurate background estimation from the CR. Defining the SRs by strips of angular position, for example, would not satisfy this requirement, unless the strips coincidentally aligned with the direction of the stream within the patch. However, stellar streams are kinematically cold and so are concentrated in both proper motion coordinates. Thus, we define our SRs using one of the proper motion coordinates. (Selecting SRs based on both proper motion coordinates is possible, but would greatly increase our total training time.) Since streams are localized in both proper motions, in principle it should not matter which one we choose; for this study, we choose μ_λ to be the proper motion coordinate defining the SRs.[5]

Based on the proper motion properties of known streams we find that a choice of a window in μ_λ of width 6 mas/yr is optimal. Streams like GD-1, located $O(10\ \text{kpc})$ from the Earth, have proper motion dispersions of ~ 2 mas/yr. Such streams would be completely contained within our SRs at distances larger than $2-3$ kpc (which is more

[5]The choice of proper motion coordinate can affect the performance of the algorithm through the number of background stars in the SR. For example, if the stream stars have small values of μ_λ but large values of μ_ϕ, then defining the SR in terms of μ_λ would lead to more background stars for the same number of stream stars, and hence a lower S/B, decreasing the stream detection probability. In Paper II we will also incorporate the results of a scan over SRs defined using μ_ϕ and show how this can achieve complementary results.

or less commensurate with the parallax cut we placed on our dataset). We also note that the stream does not have to be completely enclosed within a given SR for the algorithm to function. Proper functioning of ANODE requires only that the relative distribution of stars within the SR differ significantly from that in the CR; since the CR contains many more stars than the SR, a leakage of stream stars into the CR will not typically invalidate our approach.

Since we do not know *a priori* which SR contains a stream, we must scan over all regions. In practice, we define a series of SRs by stepping in units of $\mu_\lambda = 1$ mas/yr, with each SR then defined by the choice of $[\mu_\lambda^{\min}, \mu_\lambda^{\max}]$:

$$[\mu_\lambda^{\min}, \mu_\lambda^{\max}] = \ldots, [-10, -4], [-9, -3], \ldots, [3, 9], [4, 10], \ldots \tag{3.1}$$

in units of mas/yr. The complement of the proper motion window (i.e. all the stars in the same patch that are not in the SR) defines the control region (CR) for each SR.

Each of these choices of $(\alpha_0, \delta_0, \mu_\lambda^{\min})$ furnishes a search region and control region pair for the ANODE training step. Overlapping the SRs in this way allows us to fully capture potential streams in at least one μ_λ window when performing a blind search – if the SRs were not overlapping, then a stream could easily fall at the edge of two SRs, diluting the signal in each. By selecting SRs which are wide enough in proper motion to fully contain a kinematically cold stream and overlapping them by shifts which are smaller than the proper motion width of a typical stream, we minimize the possibility of this dilution.

SRs with fewer than 20k stars or more than 1M stars (before the fiducial cuts) are rejected for ANODE training. The former requirement is because too few stars in the SR results in poor density estimation performance, and the latter requirement is to avoid overly-long training times. In addition, SRs that contained a GC candidate (identified using a simple algorithm described in App. A.5) were cut from the analysis, as the

presence of the GC would completely overwhelm the training (i.e. in an SR containing a GC, the GC would correspond to such a large, delta-function-like overdensity, that ANODE would be unable to identify any other overdensity in the SR, such as one coming from a stream). In the end, we are left with a total of 545 SRs across the 21 patches of the sky containing GD-1.

To provide an example of an SR, we turn to our sample GD-1 patch defined in the previous section, centered on $(\alpha_0, \delta_0) = (148.6°, 24.2°)$. We select the SR defined by $\mu_\lambda \in [-17, -11]$ mas/yr, which encompasses the majority of the GD-1 stars contained within this patch. This SR is shown in Fig. 3.3 and contains 34,823 stars in total, of which 252 are tagged by [PWB18] as possible GD-1 members.

3.3.2 ANODE: Density estimation

Having defined the search regions, we turn to the probability density estimation step of the ANODE algorithm. As discussed in Sec. 3.2, the stars in our dataset are characterized by two position coordinates, two proper motion coordinates, color, and magnitude. Having set aside one of the proper motion coordinates μ_λ to define the search regions with, the remaining features $(\phi, \lambda, \mu^*_\phi, b-r, g)$ we will refer to collectively as $\vec{x}$.

Suppose the stars in a patch consist of “signal stars" coming from a cold stellar stream, and “background stars" coming from the stellar halo. Let the conditional probability density of the background stars be $P_{\text{bg}}(\vec{x}|\mu_\lambda)$, and the conditional density for the data (consisting of background stars plus signal stream stars) be $P_{\text{data}}(\vec{x}|\mu_\lambda) = (1-\alpha)P_{\text{bg}}(\vec{x}|\mu_\lambda) + \alpha P_{\text{sig}}(\vec{x}|\mu_\lambda)$ where α is a measure of the signal strength. Then the

optimal test statistic for distinguishing data from background is [115]:[6]

$$R(\vec{x}|\mu_\lambda) = \frac{P_{\text{data}}(\vec{x}|\mu_\lambda)}{P_{\text{bg}}(\vec{x}|\mu_\lambda)}. \tag{3.2}$$

If the signal is small ($\alpha \ll 1$) but sufficiently localized in feature space (i.e. a local overdensity), then we expect $R \gg 1$ where the signal is localized and $R \approx 1$ everywhere else. Since R can be computed without knowing α or P_{sig}, selecting data points with high R can purify signal to background in a model-agnostic way.

Probability density estimation of arbitrary distributions is a difficult problem, and so ANODE is only made feasible through recent advances in machine learning. In this paper, as in [NS20], we employ the MAF architecture [122] for the density estimation task. The MAF uses a specially-structured neural network to learn a bijective mapping from the original feature space into a latent space where the data is described by a unit multivariate normal distribution.[7]

Although it is relatively straightforward to train the MAF directly on the stars in the SR to learn $P_{\text{data}}(\vec{x}|\mu_\lambda, \mu_\lambda \in \text{SR})$ (the numerator of the likelihood ratio Eq. (3.2)), estimating the background density $P_{\text{bg}}(\vec{x}|\mu_\lambda)$ takes more consideration. Calculating the denominator P_{bg} from first principles often proves impossible. Instead, one of the key ideas of the ANODE method is to use sideband interpolation from the CR (the complement of the SR) to estimate the background density in the SR. More precisely, we train a second MAF on the CR to learn $P_{\text{data}}(\vec{x}|\mu_\lambda, \mu_\lambda \in \text{CR})$. If there is no stream in the CR, then

$$P_{\text{data}}(\vec{x}|\mu_\lambda, \mu_\lambda \in \text{CR}) = P_{\text{bg}}(\vec{x}|\mu_\lambda, \mu_\lambda \in \text{CR}). \tag{3.3}$$

[6]Note that this will in general not be the optimal statistic for distinguishing any particular signal hypothesis from the background, rather it is the optimal test for distinguishing the background-only hypothesis from the data-driven probability distribution. For more discussion of the meaning of optimality in the context of anomaly detection, see the Appendix to [NS20].

[7]Our selection of hyperparameters is described in App. A.4.2.

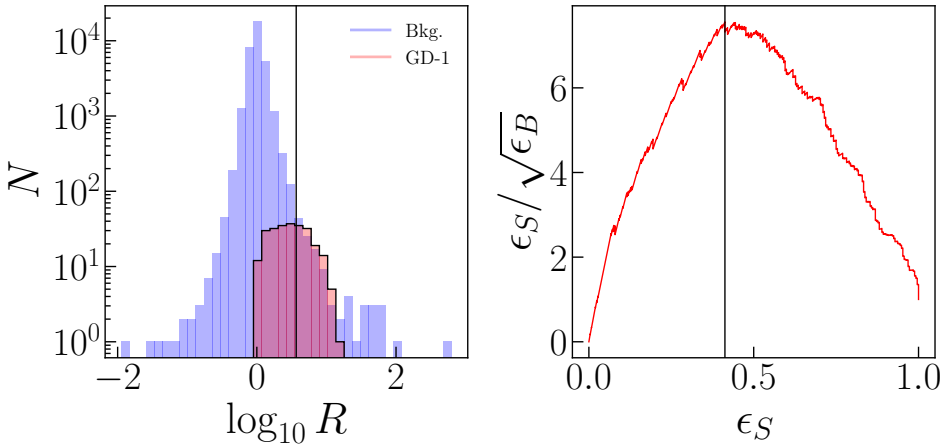

Figure 3.4: Left: R distribution for the SR $\mu_\lambda = [-17, -11]$ mas/yr in the patch centered at $(\alpha, \delta) = (148.6°, 24.2°)$. Stars identified as likely members of GD-1 by [PWB18]
are shown in red, while the "background" stars (those not tagged as likely GD-1 members by [PWB18]) are in blue. Right: Significance Improvement Characteristic (SIC) curve for the same SR, showing the signal efficiency ϵ_S and the significance improvement (signal efficiency over square root of background efficiency, $\epsilon_S/\sqrt{\epsilon_B}$) as the cut on R is varied. The vertical lines in both plots designate the R value that maximizes the SIC curve.

If the background distribution in the CR is a smooth and slowly varying function of μ_λ, then the MAF provides an automatic interpolation into the SR and yields an estimate for $P_{\text{bg}}(\vec{x}|\mu_\lambda, \mu_\lambda \in \text{SR})$, the denominator of Eq. (3.2).[8]

An important point to note is that the MAF (along with most, if not all unsupervised density estimators) has difficulty matching rapid or discontinuous changes in the probability density as a function of the features $\vec{x}$. This is not a problem for the proper motion and $b-r$ features, which smoothly go to zero. However, in position-space, the selection of stars within a circular patch on the sky results in a sharp cutoff in density at the edge of the patch. Similarly, at high magnitude g, the sensitivity of the *Gaia* satellite drops

[8] If there is signal in the CR, then by assumption it will be a very small perturbation to $P_{\text{data}}(\vec{x}|\mu_\lambda, \mu_\lambda \in \text{CR})$ (i.e. we assume there are many more background stars than signal stars in the CR). Then Eq. (3.3) will still be approximately true, and the signal contamination in the CR should not greatly affect the R statistic in the SR.

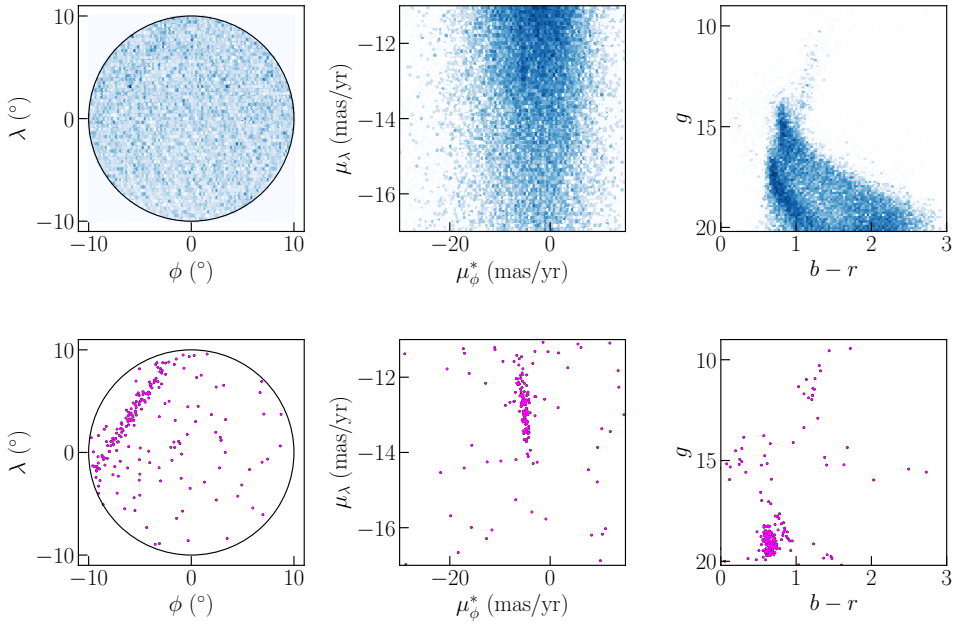

Figure 3.5: Upper row: Angular position in (ϕ, λ) coordinates (left), proper motion in $(\mu_\phi^*, \mu_\lambda)$ coordinates (center), and photometry (right) of all stars (blue) in the $\mu_\lambda \in [-17, -11]$ mas/yr SR of our example patch centered on $(\alpha, \delta) = (148.6°, 24.2°)$. Bottom row: As the upper row, applying the $R > R_{\rm cut}$ cut on the stars in the SR (purple). The GD-1 stream becomes immediately apparent. See text for details.

rapidly. The result is spuriously large R values near the edge of the patch in position space and at large g. To avoid this, we train on a larger dataset than the fiducial region in which we perform the subsequent stream-finding steps. As previously discussed in Sec. 3.2, after running ANODE, we define a fiducial region of 10° around the center of the patch in (ϕ, λ) position space and a magnitude cut of $g < 20.2$.

In Fig. 3.4 (left), we show a histogram of the ANODE probability ratio R for the stars in the $\mu_\lambda \in [-17, -11]$ mas/yr SR within the GD-1 example patch. We see that the likely GD-1 stars identified by [PWB18] are disproportionately represented at the high-R tail of the ANODE distribution. By cutting on R, the resulting sample of stars would be enriched with stream stars compared to the full sample. For a given value of R, the signal efficiency ϵ_S is the fraction of candidate stream stars passing the cut on R, and the background efficiency ϵ_B, is the fraction of non-stream-candidate stars passing the threshold. In Fig. 3.4 (right), we show the significance improvement characteristic (SIC) curve, comparing ϵ_S to $\epsilon_S/\sqrt{\epsilon_B}$ as R is varied. We see that cutting on the ANODE output can greatly improve the purity of the sample and enhance the significance of the stream detection. For the sake of illustration, we have indicated in Fig. 3.4 the optimal $R_{\rm cut}$ value, defined to be the cut on R that maximizes the significance improvement in Fig. 3.4 (right). (In more general settings, without stream-labeled stars, the optimal cut on R would not be known, see the next subsection for further discussion of this.) Starting with 252 stars out of 34,823 identified as candidate GD-1 members by [PWB18], the optimal $R_{\rm cut}$ value (corresponding to $\log_{10} R_{\rm cut} = 0.57$ for this SR) selects 206 stars, of which 103 are candidate GD-1 stars (corresponding to $\epsilon_S = 0.41$). This nominally increases the statistical significance of the stream (i.e. $S/\sqrt{B}$) by more than a factor of 7. We emphasize that the R ratio was learned in a completely data-driven, unsupervised manner, and at no point in the training were the stream candidate labels from [PWB18] ever used. Here the labels are just used to illustrate the efficacy of the ANODE R-ratio

in identifying stream stars.

In Fig. 3.5 (top), we show all the stars in the $\mu_\lambda \in [-17, -11]$ mas/yr SR, and (bottom) those stars passing the optimal R cut. GD-1 is an exceptionally dense and distinct stream: unlike other known streams it is visible, albeit barely, before the cut on R. Performing the cut of $R > R_{\rm cut}$, as shown in the lower panel, drastically increases the significance of the stream, as expected.

Finally, we comment on the issue of streaking that can clearly be observed in the position space plots of the stars in many patches (Fig. 3.3 and Fig. 3.5 are prime examples). These streaks are artifacts due to *Gaia*'s scan pattern and incomplete coverage of the sky in DR2. They might seem concerning for the ANODE method, as they appear as line-like overdensities in the angular coordinates, just like stellar streams would. However, we find no evidence that ANODE is incorrectly selecting for these spurious features. The reason is that ANODE looks for evidence of a *local* overdensity by comparing the stars in one proper motion slice with the stars outside of it. The streaking patterns are largely uncorrelated with proper motion; therefore, the overdensity they correspond to will actually *cancel* in the construction of the R ratio, and these streaking stars will not be selected for by the ANODE algorithm.

3.3.3 Regions of interest

Up to this point, our method has been largely agnostic to the astrophysics of stellar streams (beyond the choice to use proper motion as our SR-defining feature). Stars tagged as anomalous by the ANODE training may be streams, globular clusters, debris flow, or some other structure localized in the Milky Way's velocity-space. The steps in this and subsequent subsections are designed specifically to find cold stellar streams similar to the ones identified previously in data; different cuts and/or choices of parameters could be used to focus on other interesting astrophysical structures. The cuts we

choose are:

- First, we remove all stars within a box around zero proper motion of width 2 mas/yr. That is, we require

$$|\mu_\lambda| > 2 \text{ mas/yr} \text{ OR } |\mu_\phi^*| > 2 \text{ mas/yr}. \tag{3.4}$$

  Recall that the ANODE training identifies stars within the SR that are anomalous compared to the interpolation into the SR of the CR density estimate. Stars with proper motion near zero are predominantly distant stars; this population is not well-represented in a CR that does not contain $(\mu_\phi^*, \mu_\lambda) \sim (0, 0)$ mas/yr. An example can be seen in Fig. 3.6. If the SR contains this zero point, the distant stars are (correctly) identified as anomalous relative to the population in the control regions, but their sheer number completely overwhelms any other signal in the SR, requiring their removal after training is complete.

- Cold stellar streams, produced by tidally stripped globular clusters or dwarf galaxies, are predominantly composed of old, low metallicity stars. Many existing stream-finding algorithms leverage this by fitting stars in the stream candidate to isochrones appropriate to this assumption (see e.g. [92]). Although the ANODE training is agnostic to such assumptions, in this work we are specifically interested in identifying cold streams, and not all anomalous overdensities. To that purpose, we now select stars in a specific color range in order to further purify signal to background. We require our stream candidates to lie in the broad range of colors $(b-r) \in [0.5, 1]$. This range of colors was chosen so that it will contain (nearly) all of GD-1 and every stream found by Streamfinder in [96, 71]. (Streamfinder targeted globular cluster streams composed of stars with ages $\sim$ 10 Gyr and metallicities -2 dex$\lesssim$ [Fe/H]$\lesssim$-1 dex). But being broader and more general than

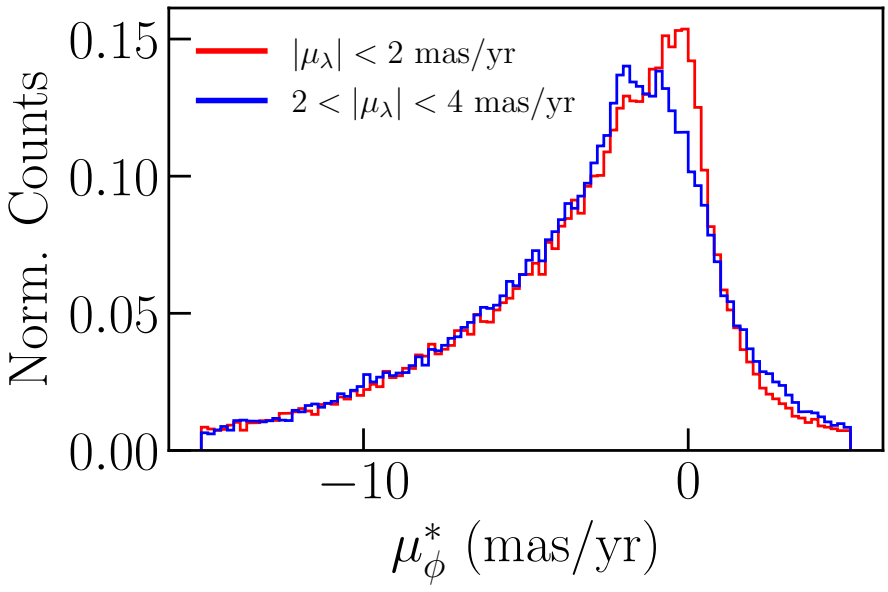

Figure 3.6: Normalized histogram of μ^*_ϕ values for stars in the 10° patch centered on $(\alpha, \delta) = (148.6°, 24.2°)$, requiring $2 < |\mu_\lambda| < 4$ mas/yr (blue) and $|\mu_\lambda| < 2$ mas/yr (red). Note that the high density of stars near $\mu^*_\phi \sim 0$ with $|\mu_\lambda| < 2$ mas/yr are not represented in the sample which does not overlap $\mu_\lambda \sim 0$. These very distant stars with near-zero total proper motion are absent as a population from search regions which do not include the zero point of proper motion.

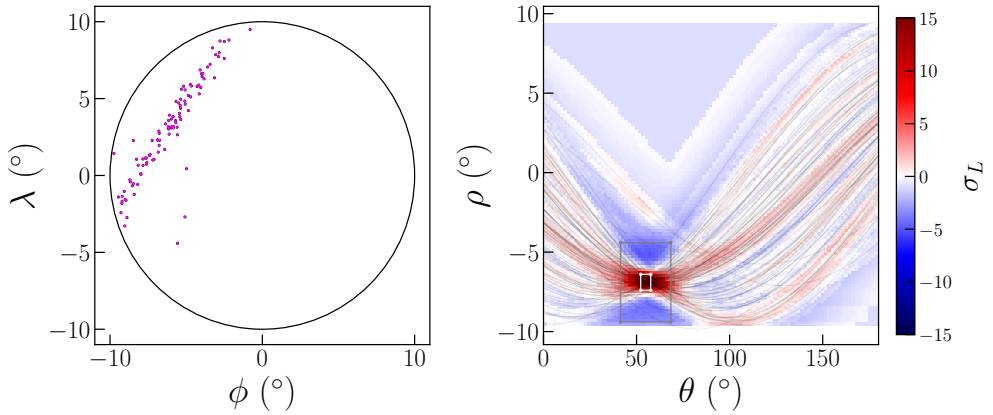

Figure 3.7: Left: Angular position in (ϕ, λ) coordinates for the 100 highest-R stars (purple) in the $\mu_\phi^* \in [-8, -2]$ mas/year, $\mu_\lambda \in [-17, -11]$ mas/year ROI from our example patch. Right: Associated curves in Hough space for these stars (black lines). The significance $\sigma_L(\theta, \rho)$ of a line oriented at each (θ, ρ) value is shown in color. The region around the point of maximum contrast (as identified by the VIA MACHINAE algorithm) is indicated by the inner white box, with the region defining the background shown as the outer box.

fitting to specific isochrones, we hope it will also enable the discovery of new streams. There may be interesting anomalous structures outside of this color range, which will be investigated in a future work.

- To further isolate any potential streams, we subdivide the SRs defined by windows of μ_λ into overlapping windows of μ_ϕ^*, with width 6 mas/yr and a stride of 1 mas/yr. We call these windows *regions of interest* (ROIs) and they are labeled by $(\alpha_0, \delta_0, \mu_\lambda^{\rm min}, \mu_\phi^{*\rm min})$. We exclude any ROI that has fewer than 200 stars as we need larger statistics to determine the presence of a stream.

Applying these cuts and further subdivision of the data to the 21 patches of the sky containing GD-1, we obtain 17,563 ROIs in total.

Within each ROI, we must decide how to apply the cut on the ANODE overdensity function $R(\vec{x}|\mu_\lambda)$. Many different types of cuts are possible, for instance setting a threshold as a percentile cut in each ROI, or a fixed value of R across all ROIs. We

have empirically found that selecting the 100 highest R stars in each ROI is effective at finding known streams (more on this in Paper II). An example of this is shown in the left panel of Fig. 3.7. It is possible that another cut (e.g. the 1000 highest R stars in an ROI) would also be effective or would find other, qualitatively different streams. This would be interesting to explore in future work.

3.3.4 Line-finding and stream detection

Over large angles on the sky, most streams form arcs in (α, δ) rather than lines (and streams with large line-of-sight velocities may not appear to form lines at all). However, the deviation from a line for the stars in the stream is small across a 10° radius circle on the sky.

Given the large number of ROIs – $\mathcal{O}(10^4)$ for the 21 patches of the sky containing GD-1 alone – we need an automated procedure for line finding. To do so, we adapt a long-standing technique from the field of computer vision based the Hough transform [67, 33]. A line passing through a point on the plane (ϕ, λ) can be expressed in terms of the distance ρ of closest approach to the origin, and the angle θ between the ϕ axis and the perpendicular from the line to the origin:[9]

$$\rho = \phi \sin\theta - \lambda \cos\theta. \tag{3.5}$$

Viewing this equation another way leads to the idea of the Hough transform for line finding: for a single point, the collection of lines that pass through it will form a sinusoidal curve in the (θ, ρ) Hough space described by Eq. (3.5). If we consider two points in the plane, then their curves in Hough space will intersect for the values of θ and ρ that define a line passing through both points. For a set of points in the plane, a subset of points on a line will manifest itself as overdensity in the (θ, ρ) space as many

[9] Note ρ can take negative values – there is a periodicity in Hough space of the form $(\rho, \theta) \sim (-\rho, \theta \pm \pi)$.

such curves intersect.

In Fig. 3.7, we show an example of the Hough transform on position data (left panel) of the 100 highest-R stars in the ROI with $\mu_\phi^* \in [-8, -2]$ mas/yr, $\mu_\lambda \in [-17, -11]$ mas/year from our example patch. As can be seen in the right panel, the Hough curves for the stars on the line all cross at the same point, corresponding to the θ and ρ values of the line on which the stream falls. The Hough transform therefore converts the problem of finding a line among a set of 2-dimensional points to the problem of finding the point with the highest density of curves in a 2-dimensional plane. Although this overdensity is obvious by eye in the example shown in Fig. 3.7, this is an extreme case and most overdensities will not be as clear-cut.

We automate the line-finding by identifying the region in Hough space with the highest contrast in density compared to the region surrounding it. We define a filter function which is applied to a box centered on a location (θ, ρ) of width w_θ and height w_ρ. The filter counts the number of stars whose Hough curves pass through the box, allowing us to define a number of curves at each point $n(\theta, \rho)$. We then redo the filtering with a larger box (subtracting the curves which also pass through the initial box) to estimate the "background" curve count, $\bar{n}(\theta, \rho)$ (being careful to renormalize the counts for the different areas of the patch covered by the two regions in Hough space). Examples of these two filtering regions are shown in Fig. 3.7. The large and small box dimensions are "hyperparameters" of the Hough transform line detection method and must be tuned based on known stellar streams to maximize detection efficiency. In this work, we will specialize to $w_\theta = 5.4°$ and $w_\rho = 1°$ for the inner box, and an outer box five times larger. This was found to be optimal for detecting relatively narrow streams such as GD-1. In Paper II we will also explore other hyperparameters for the line finder that are sensitive to wider streams.

From the filter function count of Hough curves and background estimate at each

point (θ, ρ), we define the line detection significance to be

$$\sigma_L(\theta, \rho) = \frac{n(\theta, \rho) - \bar{n}(\theta, \rho)}{\sqrt{\bar{n}(\theta, \rho)}} \tag{3.6}$$

We search in Hough space for the parameters that maximize this significance. Concretely, we bin the $(\theta - \rho)$ plane in two dimensions, using a grid of 100 bins for $0 \leq \theta \leq \pi$ and 100 bins for $-10^\circ \leq \rho \leq 10^\circ$. We then select the bin that maximizes σ_L and return this as our line detection in each ROI.[10]

When a stream is present in the SR and within the proper motion range of an ROI, we expect the resulting σ_L value to be much larger than those of ROIs without linear structures. As an example of this, in Fig. 3.8 we show the σ_L values for every ROI in the SR as a function of the central μ_ϕ^* value defining each ROI, with vertical red lines indicating the maximum and minimum ROIs which contain any GD-1 stars. As can be seen, the high-significance lines fall only in the ROIs containing GD-1 stars. By cutting on σ_L, we are to be able to distinguish ROIs that contain an actual stream in the high-R stars from those without.

3.3.5 Final Merging and Clustering

After selecting the 100 highest R stars in each ROI and applying the Hough transform line finder, we obtain the line parameters (θ, ρ) with the highest significance σ_L in each ROI. We wish to use the significances of these lines to select only the most promising stream candidates. However, cutting on the raw σ_L of an individual ROI is not effective in identifying a tractable number of likely stream candidates, because of the large trials factor (the so-called "look elsewhere effect"). Across only the 21 patches containing GD-1 there are already $O(10^4)$ ROIs, and random fluctuations could result in spurious

[10] We are implicitly assuming here that each ROI will contain at most one stream. We believe this is a safe assumption, since ROIs are fully localized in both proper motions and angular position.

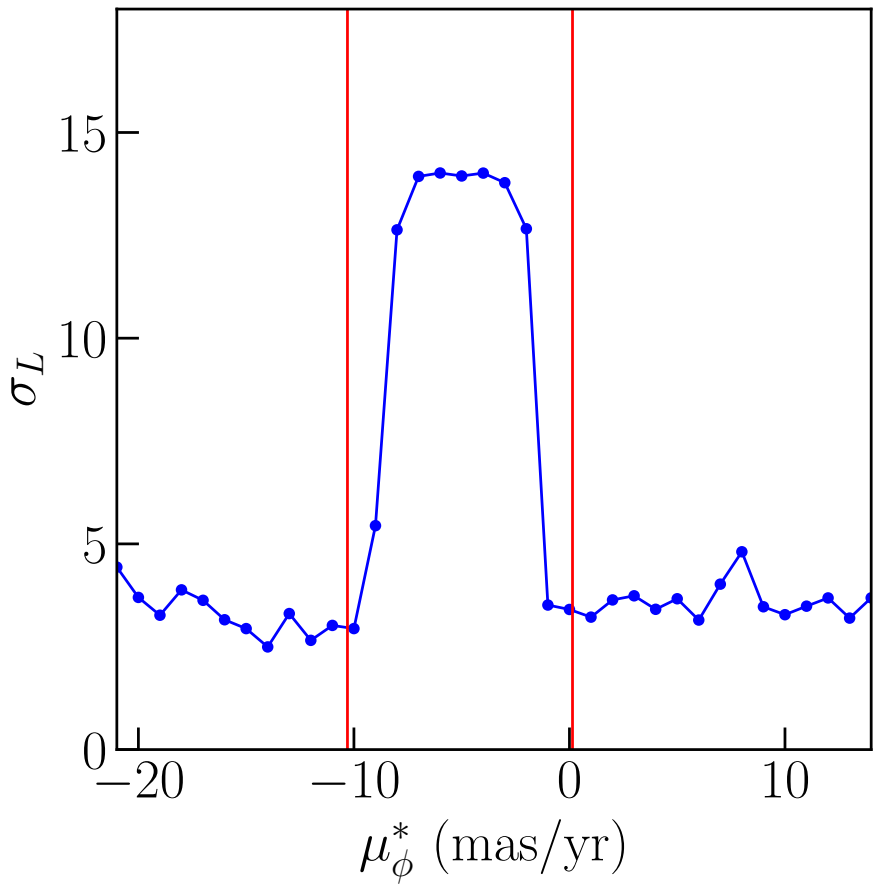

Figure 3.8: Line significance σ_L versus the central μ_ϕ^* value for each ROI with $\mu_\lambda \in [-17, -11]$ mas/yr in our example patch. Vertical red lines indicate the minimum and maximum μ_ϕ^* values for the candidate GD-1 stars of [PWB18]

.

line-like features in the background stars. This essentially dilutes the significance of a individual line detection by a correction factor, which may not be entirely trivial to estimate in the presence of correlations between ROIs.

To obtain a meaningful line detection, we use the fact that a stream is likely to be found in multiple ROIs – since the SRs are highly overlapping, each star generally has more than one R value attached to it. Therefore, we aim to cluster the ROIs that have concordant best-fit line parameters, across proper motions in a given patch, and across patches.

To perform this combination of overlapping ROIs, we have developed a three-step clustering algorithm (see Fig. 3.1 for a graphical illustration of these steps):

1. In a given patch, we consider all ROIs with the same value of μ_ϕ^*. We group together ROIs adjacent in μ_λ which have concordant line parameters.[11] In this way, all ROIs in a patch are clustered into *seeds* which have the same μ_ϕ^* and consecutive values of μ_λ. For each seed, we add the line significances of its ROIs in quadrature to form a combined line significance σ_L^{tot}.[12]

2. Next, we group together seeds at adjacent μ_ϕ^* based on the same criteria for concordance of line parameters. This forms *proto-clusters*, as shown in the second-to-last step of Fig. 3.1.

3. Finally, we merge together proto-clusters across adjacent patches using the same criteria for concordance of line parameters. This produces our final *stream candidates*, as shown in the final step of Fig. 3.1.

[11] To be precise, we require the line parameters to be within $\Delta\theta = \pi/10$ and $\Delta\rho = 2^\circ$ of each other.

[12] We are careful not to interpret σ_L^{tot} as a meaningful statistical significance in this work; rather we think of it more loosely as a figure of merit or an anomaly score for stream detection. At best, σ_L^{tot} would be a *local* significance (i.e. ignoring an enormous and difficult-to-quantify look-elsewhere-effect), and would be based on the assumption (probably not completely true) that separate ANODE runs in neighboring SRs return completely uncorrelated, random values of R on background-only stars.

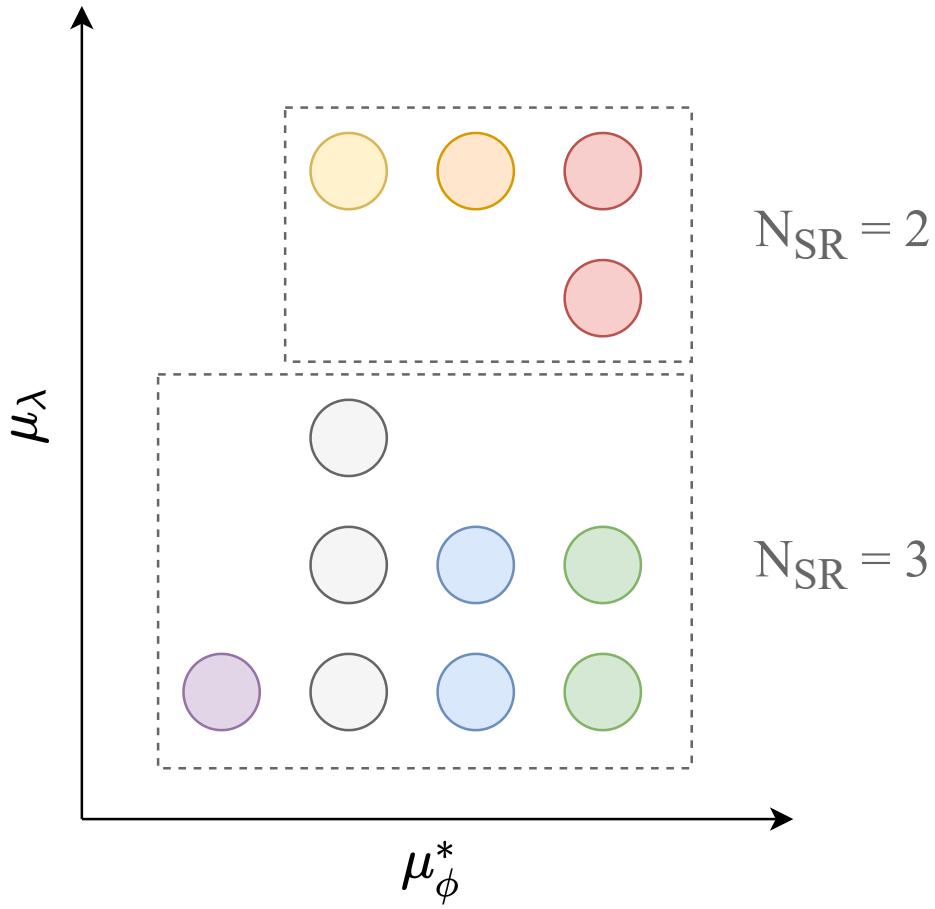

Figure 3.9: A schematic showing how regions of interest (ROIs) are combined into different protoclusters. The different colors denote different seeds, i.e. clusters of ROIs with adjacent μ_λ and the same μ_ϕ^* values. The boxes show how adjacent seeds are combined into protoclusters with different N_{SR}.

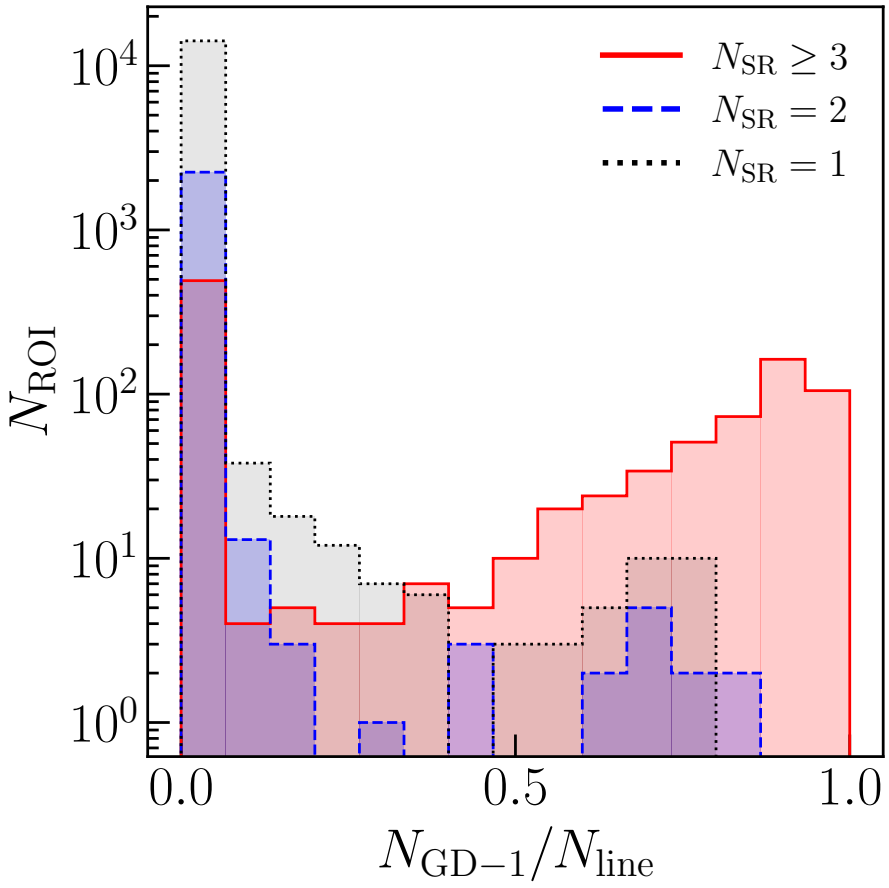

Figure 3.10: Histograms of the fraction of stars in the best-fit line of each ROI that were identified as likely GD-1 stars by [PWB18]
, for ROIs which are part of proto-clusters with $N_{\rm SR}$ = 1 (black, dashed), $N_{\rm SR}$ = 2 (blue, dotted) and $N_{\rm SR} \geq 3$ (red,solid). We see that requiring $N_{\rm SR} \geq 3$ greatly increases the fraction of candidate GD-1 stars in the best-fit line.

A schematic of steps (i - ii) is shown in Fig. 3.9. There we see 12 hypothetical ROIs that are colored by seed. Neighboring seeds are combined into protoclusters which are denoted by dashed boxes. The number of signal regions, $N_{\rm SR}$, in each protocluster is the number of ROIs in its largest seed.

In step (i), the rationale for grouping in μ_λ and not μ_ϕ^* is that in a given patch, ROIs with the same μ_λ but different μ_ϕ^* represent different, highly-overlapping slices of the same SR, with each star that appears in multiple ROIs having the same R values from ANODE. On the other hand, ROIs with the same μ_ϕ^* and different μ_λ represent different SRs, and each SR represents an independent ANODE training. Although the SRs are highly overlapping, the ANODE training is sufficiently stochastic that we take the outcome in different SRs to be quasi-independent. This motivates the adding in quadrature of the line significances of the ROIs in each seed.

In step (ii), for each proto-cluster, we characterize its significance by the seed with the highest $\sigma_L^{\rm tot}$ that it contains. The size of this seed we will call $N_{\rm SR}$ and is another measure of the significance of the proto-cluster. Note that we do not add the $\sigma_L^{\rm tot}$ values of different seeds in a proto-cluster together in quadrature, since these are highly correlated.

Applying the final merging and clustering steps to the 21 patches containing GD-1, we find that the 17,563 ROIs are clustered into 10,955 proto-clusters. Of these, 10,267 have $N_{\rm SR} = 1$; 606 have $N_{\rm SR} = 2$; and 82 have $N_{\rm SR} \geq 3$.

All else being equal, we expect real streams to have higher values of $\sigma_L^{\rm tot}$ and $N_{\rm SR}$. We show in Fig. 3.10 histograms of the fraction of stars within the best-fit line of each ROI that have been identified as candidate GD-1 stars by [PWB18], for ROIs that belong with proto-clusters with different values of $N_{\rm SR}$. As can be seen, the fraction of candidate GD-1 stars (i.e. the “purity" of the best-fit line) is significantly improved when we require $N_{\rm SR} \geq 3$.

In Fig. 3.11, we show the distributions of $\sigma_L^{\rm tot}$ values across the ROIs (the total line significance is for the proto-cluster that the ROI has been clustered into), for $N_{\rm SR} \geq 3$. We see that there is clearly a bulk distribution at low $\sigma_L^{\rm tot}$ and then a tail of outliers, with the separation occurring around $\sigma_L^{\rm tot} = 8$. It is reasonable to suppose that the majority of these low-$\sigma_L^{\rm tot}$ corresponds to false positives, while the tail could correspond to real stream detections that should be subjected to more in-depth investigation.

3.4 Demonstrating the full Via Machinae Algorithm with GD-1

Having described all the steps of the Via Machinae algorithm, we now demonstrate the full algorithm on the 21 patches of the sky that contain GD-1. For the first step of the algorithm (ANODE), we used the "Haswell" processors at NERSC, for a total of approximately 10,000 CPU-hours to analyze all 21 patches. For the subsequent steps of Via Machinae (line finding, forming protoclusters, and forming stream candidates), we used the local HEP cluster at Rutgers, for a total of approximately 50 CPU-hours.

Motivated by the discussion in the previous subsection, we focus on only those proto-clusters with $N_{\rm SR} \geq 3$ and $\sigma_L^{\rm tot} \geq 8$. This leaves only 16 proto-clusters. Merging these results in only two stream candidates, shown in Fig. 3.12. One might have expected far more stream candidates, given the enormous trials factors involved (e.g. $O(10^4)$ ROIs that we started with). This is a sign that the cuts on $N_{\rm SR}$ and $\sigma_L^{\rm tot}$ that we have chosen are indeed effective at reducing the false positive rate.

The less prominent stream candidate, shown in blue, is built from a single proto-cluster representing 16 ROIs with $\sigma_L^{\rm tot} = 9.5$. It comes from the patch centered at $(\alpha, \delta) = (138.8°, 25.1°)$. The stream candidate does not correspond to any known stream, and *a priori* it may be a real stream or a spurious detection. Closer inspection

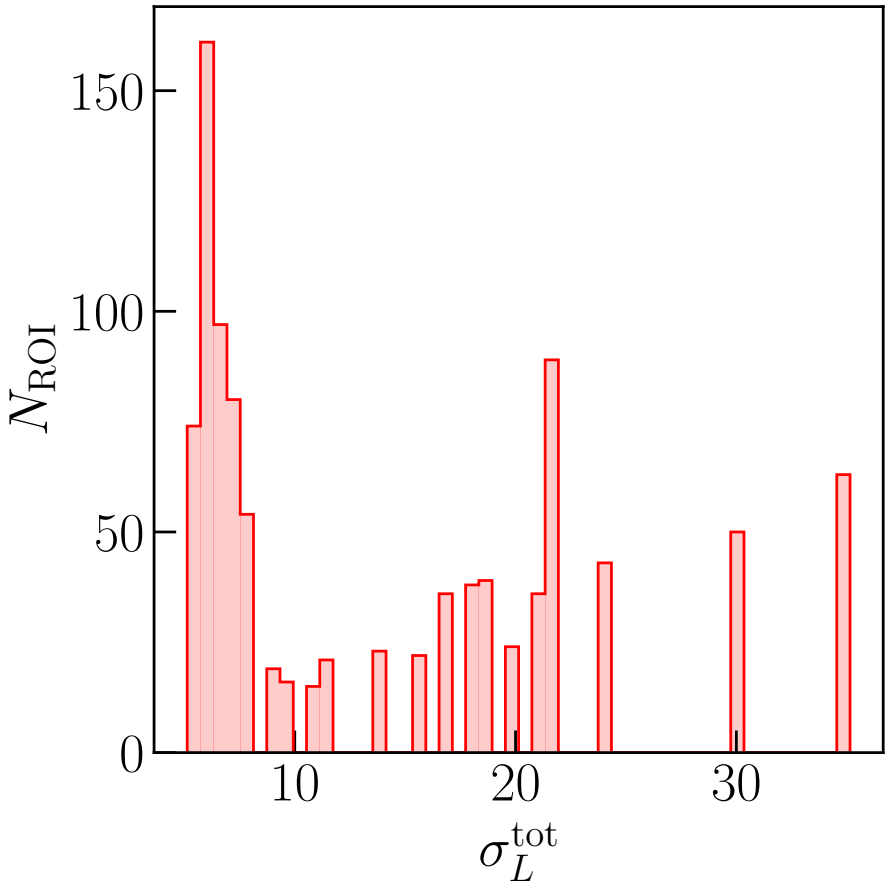

Figure 3.11: Histogram of the σ_L^{tot} values of protoclusters with $N_{\mathrm{SR}} \geq 3$, with each protocluster weighted by the number of ROIs it contains.

reveals that all of the high-R stars identified by ANODE are tightly clustered at the edge of the circular patch, almost perfectly aligned with the direction of the Galactic disk (and on the same side of the patch as the disk). Although this patch is $\gtrsim 30°$ off the Galactic plane, we still observe a strong density gradient towards and aligned with the disk. Therefore, we suspect that ANODE has identified disk stars in this case, and not a stellar stream. We discuss this further in App. A.6.

The second, more prominent stream shown in red in Fig. 3.12, is composed of 15 proto-clusters representing 518 ROIs, and is clearly GD-1. In Fig. 3.13 we show

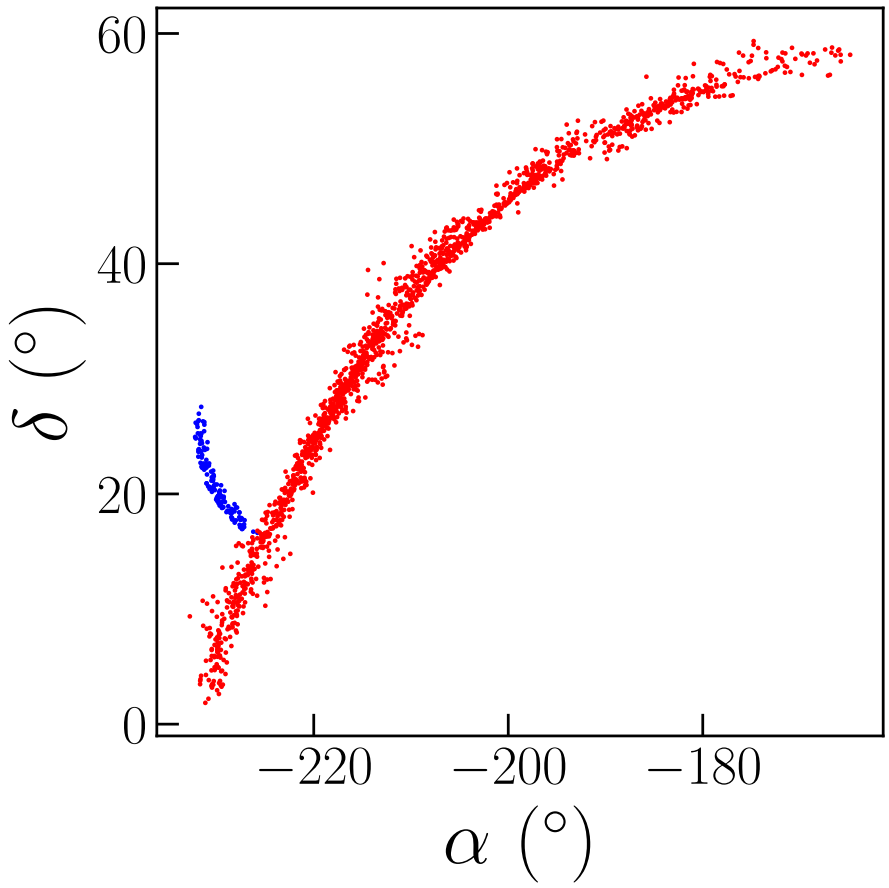

Figure 3.12: The two stream candidates built out of proto-clusters with $N_{\rm SR} \geq 3$ and $\sigma_L^{\rm tot} \geq 7.5$.

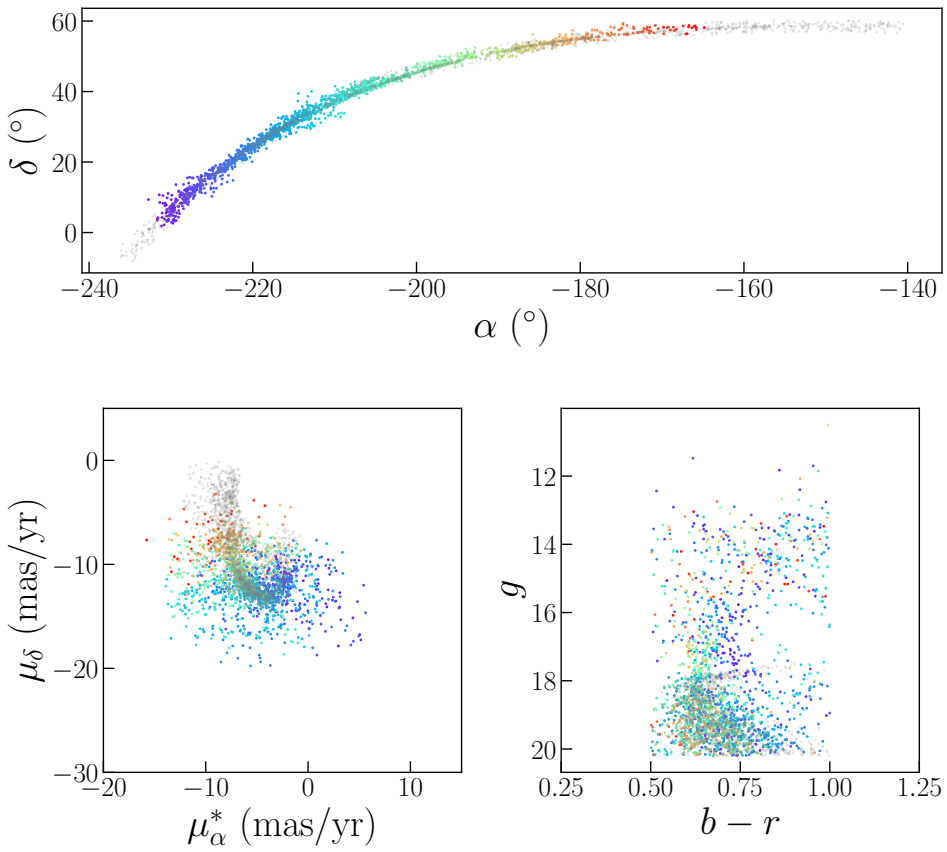

Figure 3.13: Scatter plots of the angular positions, proper motions, and color/magnitudes of the 1,688 stars in the more prominent of the two stream candidates identified by Via Machinae, overlaid on the likely GD-1 stars tagged by [PWB18] (gray) in the same region of g and $b-r$ space. The Via Machinae stars are color-coded by position in α, to facilitate cross referencing between the three individual scatter plots.

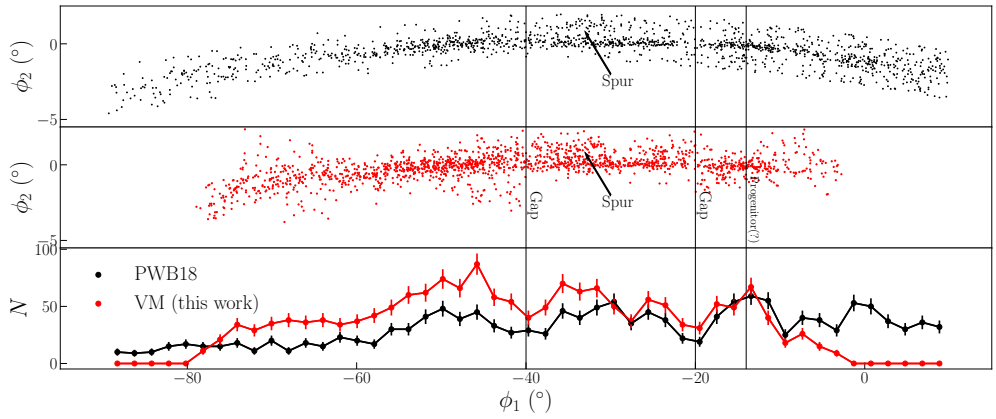

Figure 3.14: Comparison of the likely GD-1 stars from [PWB18] (top, black) and the stream candidate stars identified by Via Machinae (middle, red), in the GD-1 stream-aligned coordinate system (ϕ_1, ϕ_2) [83]. The location of previously identified features of GD-1 (two gaps, the possible progenitor location, and the spur) are indicated. Bottom row shows the number of candidate stream stars identified by [PWB18] (black) and Via Machinae (red) in ϕ_1 bins of width 2°; the error bars are purely statistical (Poissonian). In top and bottom panels, a cut on $g < 20.2$ and $0.5 < b - r < 1$ has been applied to the stars from [PWB18] so that a direct comparison can be made with the stars in this analysis.

the positions, proper motions, and photometry of the 1,688 stars in this stream candidate, overlaid on the locations of the stars tagged as likely GD-1 stream members by 2018ApJ...863L..20P. In Fig. 3.14, we present another look at the comparison between Via Machinae and [PWB18], this time using the coordinate system aligned with the GD-1 stream [83].[13] in this section, we show the stars of the latter without correction for extinction, and apply the same cuts on their (uncorrected) magnitudes and colors of $g < 20.2$ and $0.5 < b - r < 1$ as we do for our fiducial sample.

Broadly speaking, we see that Via Machinae has done an excellent job finding the GD-1 stars across the 21 patches of the sky considered in this work. Some notable features and caveats which deserve consideration are as follows:

- Fig. 3.14 shows that Via Machinae has successfully reproduced some famous features of GD-1, including both gaps, the possible progenitor, and the “spur"

[13] To allow for direct comparison of our results with [PWB18]

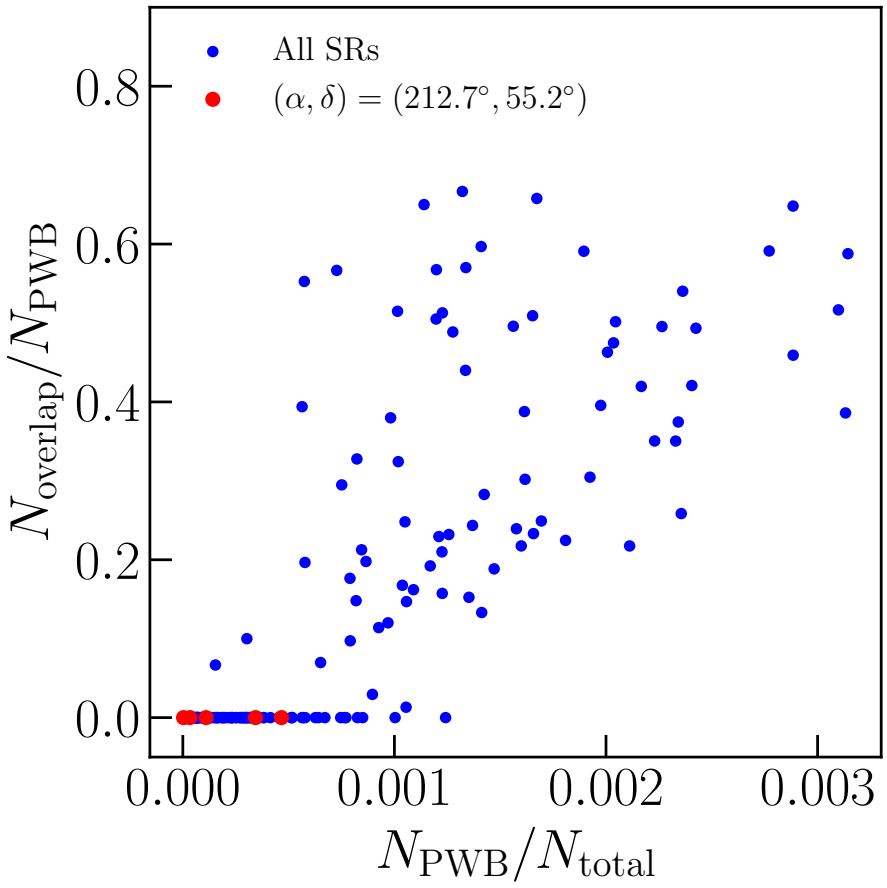

Figure 3.15: For each SR, we plot the fraction of total stars N_{total} in an SR which are identified as likely members of GD-1 by [PWB18]
(N_{PWB}), compared to the fraction of N_{PWB} which are *also* identified by Via Machinae as likely members of GD-1 (N_{overlap}). The SRs which lie in the patch centered on $(\alpha, \delta) = (212.7°, 55.2°)$ are shown in red. This patch contains the majority of the right-hand side of GD-1 which is not identified in our analysis.

[PWB18].

- We see that Via Machinae confirms most of the additional 20° of GD-1 discovered in [PWB18] (corresponding to $\alpha \lesssim -220°$, or $\phi_1 \lesssim -60°$). The left-most end of GD-1 ($\alpha \lesssim -235°$, $\phi_1 \lesssim -80°$) is missing from our stream candidate; this is because those patches were not included in our analysis as they were deemed too close to the disk ($|b| < 30°$).

- On the right side of GD-1, we see that we are also missing stars compared to [PWB18]. Closer inspection of this missing region reveals that this segment of the stream was captured by only a single patch, centered on $(\alpha, \delta) = (212.7°, 55.2°)$, and the proper motion of GD-1 on this end of the stream is closer to $\mu_\lambda = 0$, increasing the number of background stars in the relevant SRs. We will return to this point and elaborate on it further below.

- The feature protruding from the stream at $\alpha \sim -215°$ and $\delta \sim 40°$ (see Fig. 3.13) is most likely an artifact of our line finding procedure.

- Of the 1,985 stars identified as likely members of GD-1 from [PWB18], 1,519 are in our fiducial (color and magnitude) region, and 738 (49%) overlap with the membership of our stream candidate.

- The remaining 950 stars in our stream candidate were not tagged by [PWB18]. Some of these may very well be additional members of GD-1. However, the proper motion and color-magnitude plots of Fig. 3.13 make clear that our method also picks up a significant number of non-stream stars, see e.g. the group of bright stars that are clearly not associated with the main GD-1 isochrone.

Regarding the stars missing from the right-most end ($\alpha \gtrsim -160°$ or $\phi_1 \gtrsim -10°$) of GD-1, it is notable that this side of the stream has values of μ_λ closest to zero (this is

apparent in Fig. 3.13 after taking into account that $\mu_\lambda \approx \mu_\delta$ for these patches). Hence, this segment of GD-1 falls primarily in SRs with an increased number of background stars compared to the rest of the stream. The fact that we do not recover this part of GD-1 strongly suggests that successful stream-finding with Via Machinae may require a minimum admixture of stream stars in the SR. In Fig. 3.15, we use the stream candidates from [PWB18] as a proxy for the GD-1 stars, and plot the fraction of [PWB18]-tagged stars which are also identified as stream candidates by Via Machinae (in each SR) versus the fraction of stars in each SR which are tagged by [PWB18]. As can be seen, the fraction of [PWB18] stars also identified as stream members by Via Machinae is strongly correlated with the fraction of stream stars in the SR. In particular, the overlap fraction drops precipitously when the stream makes up $\lesssim 0.1\%$ of the total stars in the SR. All of the SRs through which the missing right-hand side of the GD-1 stream pass have a low fraction of stream stars. We believe this goes a long way toward explaining why Via Machinae missed these members of GD-1. Further work is needed (including a study of other streams beyond GD-1) to determine if this threshold is a more general requirement of ANODE and Via Machinae for stream detection.

Apart from the apparent required minimum S/B detection thresholds for ANODE and Via Machinae, the fact remains that our stream candidate does not include all of the likely GD-1 stars tagged by [PWB18] (completeness), and appears to include a substantial number of non-GD-1 stars (purity). However, this is not necessarily a drawback of the method. Rather, it reflects the emphasis placed by Via Machinae on *stream discovery* rather than *stream membership*. When designing our algorithm, our choices were motivated to identify stream candidates at a sufficiently high statistical significance to overcome the random background. Decisions such as the number of high-R stars to include in each ROI and the line width in the Hough transform were made with this in mind, rather than maximizing accurate stream membership of the

candidate. Loosening these criteria would likely recover more of the tagged stream stars than the 49% identified here – this would have to be weighed against increasing the number of false-positive stream candidates identified across the full sky. Thus, the resulting stream candidates should be taken as signs for discovery, rather than an accurate membership study of particular stars and whether they belong to a stream. After stream discovery, the candidate must be considered individually, loosening or eliminating some of the algorithmic choices that are part of Via Machinae. The density estimates from ANODE may continue to aid in this a posteriori analysis, but this is beyond the scope of the current paper.

3.5 Conclusions

In this work, we described a new machine learning-based algorithm called Via Machinae for stellar stream detection using *Gaia* DR2 data, and applied this technique to identify the GD-1 stream. As a particularly distinct stream with readily available membership catalogs to use for detailed comparisons, GD-1 is an excellent testbed for our algorithm.

The core of our technique is ANODE, a data-driven, unsupervised machine-learning algorithm that uses conditional probability density estimation to identify anomalous data points in a search region without having to explicitly model the background distribution. This approach is made possible by advances in deep learning that approximate the probability densities in an unsupervised way. We take as input for the ANODE training the angular position, proper motion, and photometry of the stars in *Gaia* DR2. No astrophysical knowledge is embedded into ANODE, other than in our choice to condition the probability estimation on one of the proper motion coordinates μ_λ – which is also used to define the search regions. This allows us to identify potential anomalies while remaining agnostic to the Galactic potential, orbits, or stellar composition of the streams.

The output of ANODE is a likelihood ratio R, with large R values corresponding to stars whose phase space density in the search region is larger than expected based on interpolation from the control regions. To turn these anomalies into stream detection, Via Machinae engages in a number of additional steps. Some of these steps – concentrating on old, metal-poor stars (identified with a cut on $b - r$, without requiring the stars to lie on an isochrone), further slicing the data into regions of interest based on the other proper motion coordinate μ_ϕ^* – are designed to improve signal-to-noise of stream detection, without sacrificing too much of the model-independence. Other steps – automated line finding using the Hough transform, merging concordant best-fit lines in adjacent ROIs and patches of the sky – are intended to build a stronger case for a stream detection (as opposed to some other anomalous structure or spurious false positive). The upshot is that Via Machinae produces a list of stream candidates that have been found in multiple overlapping search regions with high significance.

Using this method, we recover the GD-1 stream across the 21 patches in our full-sky scan that include it. Although GD-1 is an atypically dense, cold and narrow stream, it is still non-trivial that our method is able to recover it in an unsupervised and fully automated way. Moreover, we chose to focus on GD-1 in this paper as it provides a clear, step-by-step introduction to our algorithm. The application of Via Machinae to other known streams and to the full-sky dataset will be discussed in a forthcoming Paper II.

This initial application of unsupervised density estimators for stellar stream discovery, suggests other potentially interesting directions which recent advances in deep learning have made possible. Most obviously, our method can in principle be adapted to look for other interesting cold objects in the Milky Way, such as debris flow, tidal tails, and other stellar substructure [74, 73, 140, 44, 43, 59]. Other methods of density estimation beside the MAF may also prove to be useful: we used the MAF because

it is reasonably fast and easy to train and was demonstrated to perform well in the ANODE anomaly detection task in [NS20]. However, neural autoregressive flows [68], neural spline flows [36], and mixture density networks [16] may possibly have improved performance in some or all contexts.

Having data-driven measures of "signal" and "background" densities may prove to be useful for problems beyond discovery. Sampling from these density estimators is possible, and might be a way to construct mock catalogues. The density estimates themselves might be useful for answer questions of stream membership. This could help in going beyond the Via Machinae discovery steps outlined in this paper, and further establish the validity and accuracy of the proposed stream candidates.

Acknowledgements

We would like to thank A. Bonaca, D. Hogg, S. Pearson, A. Price-Whelan for helpful conversations; and Ting Li, Ben Nachman and Bryan Ostdiek for comments on the manuscript. MB and DS are supported by the DOE under Award Number DOE-SC0010008. LN is supported by the DOE under Award Number DESC0011632, the Sherman Fairchild fellowship, the University of California Presidential fellowship, and the fellowship of theoretical astrophysics at Carnegie Observatories. LN is grateful for the generous support and hospitality of the Rutgers NHETC Visitor Program, where this work was initiated.

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/`

consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

Data Availability

This paper made use of the publicly available *Gaia* DR2 data. For the GD-1 stars identified through our analysis, please email the corresponding author.

Chapter 4

Simulation-Based Inference for efficient sampling of the pMSSM subject to experimental constraints

4.1 Introduction

The pursuit of understanding beyond the Standard Model (BSM) physics theories in the context of experimental results is the cornerstone of much research in high energy physics (HEP). Although weak-scale observables can be calculated with increasingly precise accuracy, it is non-trivial to perform the inverse calculation, i.e. determining the regions of parameter space responsible for these results.

This is especially the case for theories with many free parameters – the most famous of which is the phenomenological minimal supersymmetric model (pMSSM). The typical method in HEP literature is to define a search space in pMSSM parameters, uniformly sampling it, and then rejecting all samples which do not conform with theoretical or experimental constraints. The number of required samples, however, grows

exponentially with the dimension of the search space, so these methods are extremely computationally intensive, if not intractable when calculations are slow.

There has been a recent upsurge of developments in the literature regarding the inverse problem of restricting parameter spaces of a forward model purely through sampling (i.e. calculating observables from model parameters). The introduction of neural networks to simulation-based inference (SBI) frameworks, in particular, has lead to explosive improvements in accuracy and precision and ablation studies comparing them all [124][118][62][90].

In this work, we introduce the sequential neural ratio estimation (SNRE) algorithm to the problem of sampling from an experimentally constrained pMSSM. Although not explicitly stated, recent work [66] has used methods from the simulation-based inference literature – namely neural likelihood estimation. We show how that method fits into the larger SBI framework and demonstrate an alternative, though related approach that makes use of likelihood-to-evidence ratio neural estimation. Additionally, we demonstrate sequential versions of two algorithms that can significantly reduce the number of model evaluations required.

Specifically, we are interested in sampling from the pMSSM parameter space which produces experimentally viable predictions for the relic density of dark matter (Ω_χ), Higgs mass (m_h), anomalous magnetic moment of the muon (μ), and WIMP-nucleon cross sections (σ^{SI}).

This manuscript is organized as follows: § 4.2 details the general SBI framework and describes the SNRE algorithm and § 4.3 details the search space of the pMSSM to which we will be applying SNRE. In § 4.4 we demonstrate the improved sampling efficiency of sequential algorithms. An application of SNRE to an experimentally constrained pMSSM parameter space is shown in § 4.5. Finally, in § 4.6 we summarize our findings and further discuss the uses of SBI.

4.2 Simulation-Based Inference (SBI)

The goal of this work is to make use of the simulation-based inference (SBI) framework to sample from a large-dimensional parameter space while evaluating the model as few times as possible. There are essentially four overarching approaches to SBI: rejection sampling with approximate bayesian computation (ABC), posterior estimation (NPE), likelihood-to-evidence ratio estimation (NRE), and likelihood estimation (NLE). For a full review of simulation-based inference,and other methods therein, we refer the reader to [26].

We begin by stating Bayes' rule:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{4.1}$$

where $P(X|\theta)$ is the likelihood of the data given the model parameters, $P(\theta)$ is the prior probability of the model parameters, and $P(X)$ is the evidence. In this work, θ refers to supersymmetric parameters and X refers to various observables of our choosing, e.g. the Higgs mass.

The NRE approach begins by approximating the likelihood-to-evidence ratio, $r(X, \theta) = \frac{P(X|\theta)}{P(X)}$, and then using it to sample from the posterior distribution of the model parameters with Hamiltonian Monte Carlo (HMC). HMC can be thought of as a Markov Chain Monte Carlo (MCMC) [101] [57] algorithm where a chain's position and momentum are sampled from a joint distribution and then evolve according to Hamilton's equations. HMC makes use of derivatives to calculate its transitions, so it cannot be run on microMEGAS itself, but can be used with neural networks. For a full review of HMC, see [110].

4.2.1 Likelihood-to-Evidence Ratio Estimation

We will be following the approach of [62] to approximate the likelihood-to-evidence ratio. Assume we have two classes of simulations, Y_0 and Y_1, each of which are composed of pairs of $\{X, \theta\}$. We define Y_0 and Y_1 such that $Y_0 = \{X, \theta\} \sim P(X)P(\theta)$ and $Y_1 = \{X, \theta\} \sim P(X|\theta)$. We use curly brackets to denote a set of values. The optimal classifier (one which minimizes the binary cross-entropy loss) between Y_0 and Y_1 is given by

$$d^*(X, \theta) = \frac{P(X|\theta)}{P(X|\theta) + P(X)P(\theta)} \tag{4.2}$$

which can in turn be used to express the likelihood-to-evidence ratio:

$$r^*(X, \theta) = \frac{P(X|\theta)}{P(X)} = \frac{P(X, \theta)}{P(X)P(\theta)} = \frac{d^*(X, \theta)}{1 - d^*(X, \theta)} \tag{4.3}$$

where we used the fact that the joint distribution $P(X, \theta) = P(X|\theta)P(\theta)$. Thus, approximating $r^*(X, \theta)$ is equivalent to finding the optimal classifier $d^*(X, \theta)$, which we can accomplish by training a classifier to distinguish between Y_0 and Y_1. For the rest of this work, we'll use $r(X, \theta)$ to denote the approximate likelihood-to-evidence ratio.

In practice, we produce samples of Y_1 by first drawing θ from the prior, $P(\theta)$, and then passing it through our simulator (microMEGAS) to generate X. We begin sampling from Y_0 in the same way: we draw θ from the prior, plug it into our simulator to generate X. After this, however, we *re-sample* $\theta' \sim P(\theta)$ and use the new value in its place, i.e. $(X, \theta') \sim Y_1$.

We will use a neural network to model $r(X, \theta)$ which allows us to take advantage of the vast literature and resources available for deep learning applications. More specifically, we will use an ensemble of multi-layer perceptrons (MLPs). Ensembling models and taking their average was shown in [63] to produce more accurate, conservative

posteriors than would be obtained by simply training a single model.

Although not applicable for this work, it is worth pointing out that the data can be of virtually any format. The authors of [62] have shown the ability of this algorithm to ingest 128×128 pixel images of strongly-lensed galaxies and produce accurate posteriors of its Einstein radius.

4.2.2 Sequential Neural Likelihood-to-Evidence Ratio Estimation

If the prior is not carefully crafted to be a good approximation of the posterior to start with, we expect many of the samples to produce observables far from the experimental values of interest. These can be seen as wasted evaluations of the model, the very thing we are trying to limit. However, with a trained NRE, we have a better approximation of the posterior available to us. We therefore elect to iterate the training procedure outlined in the previous section, but replace the sampling of θ from the prior with a sampling from the intermediate posterior. The full algorithm is outlined in Algorithm 1.

Empirically, it has been shown that sequential versions of SBI tend to converge to the true posteriors with fewer simulator evaluations than their non-sequential counterparts [90]. Intuitively, this is because the approximate posterior will be most accurate in the most dense regions of the parameter space *as specified by the training data*. When we are sampling from the posterior, however, we are most interested in its regions of highest probability. Therefore, by iteratively closing in on the true posterior, we can reduce the number of model evaluations required. We demonstrate this in Section 4.4 where we run trials to determine how efficient (the fraction of simulations which are not excluded) naive, NRE, and SNRE sampling are. Our results are summarized in Figures 4.1 and 4.2.

Finally, it is important to note that either version of SBI is not guaranteed to converge to the true posterior, and in fact, can often produce overconfident bounds on parameter

space with exceedingly high computational budgets required to calibrate [63]. This problem is somewhat alleviated by using an ensemble of classifiers, which we do for all of our applications.

Let $p_r(\theta|X_0)$ be the posterior in round r. Let $p_{r=0}(\theta|X_0) = P(\theta)$. $r < R$ set prior to $p_r(\theta|X_0)$ $n \leq N$ Sample $\theta_n \sim p_r(\theta|X_0)$ Sample $X_i \sim P(X|\theta_n)$ Add $\{\theta_n, X_n\}$ to $\mathcal{D}$ $d \leq D$ Sample $\{\theta_A, X_A\}_d \sim \mathcal{D}$ Sample $\{\theta_B, X_B\}_d \sim \mathcal{D}$ Assign labels $y = 1$ for $\{\theta_A, X_A\}_d$ and $\{\theta_B, X_B\}_d$ Assign labels $y = 0$ for $\{\theta_B, X_A\}_d$ and $\{\theta_A, X_B\}_d$

(re-)train $r(X, \theta)$ to classify $\{\theta_A, X_A\}_D$, $\{\theta_B, X_B\}_D$, $\{\theta_B, X_A\}_D$, and $\{\theta_A, X_B\}_D$ by minimizing the binary cross-entropy loss. $p_r(\theta|X_0) = r(X, \theta)P(\theta)$

4.3 Phenomenological Minimal Supersymmetric Model (pMSSM)

The unconstrained MSSM consists of over free 100 parameters. By enforcing no flavor changing neutral currents, no new sources of CP violation, and first and second generation universality, these parameters are restricted to a 19-dimensional parameter space [32]. As we are interested in supersymmetric contributions to the anomalous magnetic moment of the muon, we expect light smuons be more probable than their first generation counterparts. Therefore, we relax this final assumption in the slepton sector of the pMSSM, which adds two more parameters to this space. To simplify our search, we set all squark masses to 4 TeV, large enough so that they have negligible effects in our calculations.

To utilize the SBI framework, one must have a forward model of the likelihood, i.e. a function which takes in a set of parameters, θ and produces an observable, X, from some underlying distribution. The underlying distribution need not be known explicitly for our goal is to learn it from the samples. Additionally, we must provide a prior

Parameter	Domain	Description
$\|\mu\|$	[100, 4000]	Higgs mixing parxameter
$\|M_1\|$	[50, 1000]	Bino soft-SUSY mass
$\|M_2\|$	[100, 4000]	Wino soft-SUSY mass
M_3	[400, 4000]	Higgsino soft-SUSY mass
M_{L_1}	[100, 4000]	Left-handed selectron mass
M_{L_2}	[100, 1000]	Left-handed smuon mass
M_{L_3}	[100, 4000]	Left-handed stau mass
M_{r_1}	[100, 4000]	Right-handed selectron mass
M_{r_2}	[100, 1000]	Right-handed smuon mass
M_{r_3}	[100, 4000]	Right-handed stau mass
M_A	[400, 4000]	Pseudoscalar Higgs mass
$\tan\beta$	[1.0, 60]	Ratio of Higgs VEVs
$\|A_t\|$	[0, 4000]	Trilinear Higgs-stop coupling
$\|A_b\|$	[0, 4000]	Trilinear Higgs-sbottom coupling
$\|A_\tau\|$	[0, 4000]	Trilinear Higgs-stau coupling

Table 4.1: Parameter domains for pMSSM. All masses and couplings are in GeV. All squark mass parameters are set to 4 TeV.

distribution on the parameters. For our applications, we will use a uniform prior on the parameters, with bounds chosen to be consistent with the literature.

In the following appications, we will aim to produce posterior distributions conditioned on a combination of observables calculated by microMEGAS. Although not infinitely precise, microMEGAS presents a deterministic output, so we assume gaussian error bars on these calculations. Thus, our likelihood is given by $\mathcal{N}(X(\theta), \sigma_X)$, where $\mathcal{N}$ is the normal distribution and $X(\theta) = (\Omega_{\rm DM}(\theta), a_\mu(\theta), m_h(\theta), p_{\rm Xenon1T})$ is the vector of calculated observables from microMEGAS. The variance of the distribution is $\sigma_X^2 = \sigma_{\rm exp}^2 + \sigma_{\rm th}^2$ with theoretical gaussian uncertainties: $\sigma_X = (\sigma_\Omega, \sigma_{a_\mu}, \sigma_{m_h}, \sigma_{p_{\rm Xenon1T}}) = (0.01, 65 \times 10^{-11}, 1.0\ {\rm GeV}, 10^{-5})$. pMSSM parameters, θ are sampled from a uniform prior, with bounds shown in table 4.1. When sampling from our posterior, we set the observed values of X to their experimental results $X_0 = (0.12, 125\ {\rm GeV}, 251 \times 10^{-11}, 0.0455)$ [133] [1] [24] [2] [14]. As Xenon1T has not yet observed WIMP dark

matter, we set its p-value equal to $\mathrm{CDF}_N(2\sigma)$ and treat all p-values greater than this value identically via the pre-processing shown in App. A.3.

4.4 Benefits from Sequential Training

In practice, one is often looking for the minimal amount of time required to obtain N samples from a distribution. We can approximate the amount of time spent computing, T, via

$$T = Ns + Rt + (R-1)v + w \tag{4.4}$$

where N is the total number of microMEGAS calculations performed, s is the time per microMEGAS evaluation, R is the number of sequential rounds of training, t is the time to train the network in each sequential round, v is the amount of time to sample from the intermediate posteriors during SNRE training, and w is the amount of time to sample from the final SNRE posterior. In this application, we limit R, t, v, w, so that the majority of the time is spent evaluating microMEGAS, i.e. $T \approx Ns$. N is often chosen to be large enough so that the number of samples surviving the constraints is larger than some number, M. Thus, we can write $N = M/\varepsilon + D$ where ε is the efficiency of our microMEGAS samples, and D is the number of samples used to train our models (for traditional methods, $D = 0$). Thus for very small ε, as is the case when sampling θ from the prior, this can be computationally prohibitive. The goal of SBI is thus to minimize N by maximizing ε. We illustrate this effect by calculating the efficiency of over a range of training sample sizes and SNRE training rounds.

4.4.1 Setup

The benefits of the SBI framework become clear when looking at the efficiency of sampling, i.e. the fraction of samples which satisfy the given constraints of the problem. To illustrate this, we perform the pMSSM posterior approximation exercise while enforcing:

$$\begin{aligned} 0.08 < \Omega_\chi < 0.14 \\ 122\ \text{GeV} < m_h < 128\ \text{GeV} \\ 56 \times 10^{-11} < a_\mu < 445 \times 10^{-11} \\ 0.0455 < p_{\text{Xenon1T}} \end{aligned} \tag{4.5}$$

where p_{Xenon1T} is the p-value of the model as determined by Xenon1T.

We choose to run the SBI pipeline on 3 different choices of R. Specifically we set $R = 1$ for NRE, $R = 3$ for SNRE_3, and $R = 5$ for SNRE_5. We compare these with the "prior" baseline which uses samples from the prior to calculate observables and then subjects them to the same filtering as shown in Eq. 4.5. We emphasize that samples from the prior do not correspond to samples from the posterior, and are only used as a point of comparison.

We allot a budget of 150,000 microMEGAS calculations and split them into training sets and inference sets. The training sets are used to train the networks, and the inference sets are the evaluations from which our final samples come. Ideally, the training sets will be as small as possible so that we can maximize the number of posterior samples in our inference set. Increasing the size of the training set, however, could increase the efficiency of posterior samples, thus enabling equal performance with the a weaker algorithm producing a larger inference set. For each method, we run a trial on a training set of size 10,000, 31,000, and 100,000 with the remaining calculations left for inference.

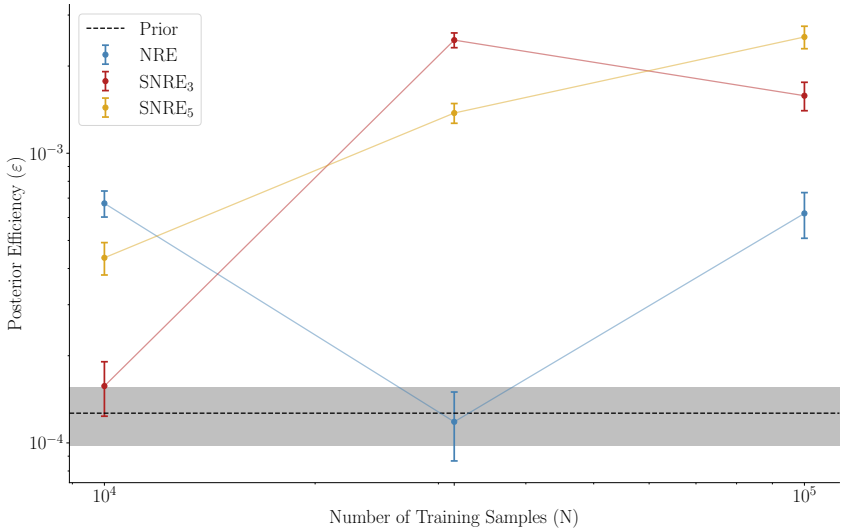

Figure 4.1: The fraction of points sampled from the posterior that lie within the ranges specified in Eq. 4.5

For SNRE$_R$, the calculations are equally distributed amongst the R rounds.

4.4.2 Results

The results demonstrate several interesting effects. First, in Figure 4.1 we observe the expected trend amongst the sequential algorithms, as the number of points used to train increases, the number of points sampled from the approximate posteriors which lie within the ranges specified in Eq. 4.5 also increases. This is not true for NRE, however, which performed worse for the medium-sized training set. This is likely due to overfitting which was not prevented by early-stopping. As the efficiency is highest for SNRE$_5$, it is the preferred model if additional computational resources are available after the training phase, e.g. in follow up work.

Shown in Figure 4.2, we determine the average amount of time from start to finish required to generate a valid sample from the final posterior. We determine that, although slower overall, the efficiency of SNRE$_R$ gives it a factor of up to 3 times faster sampling

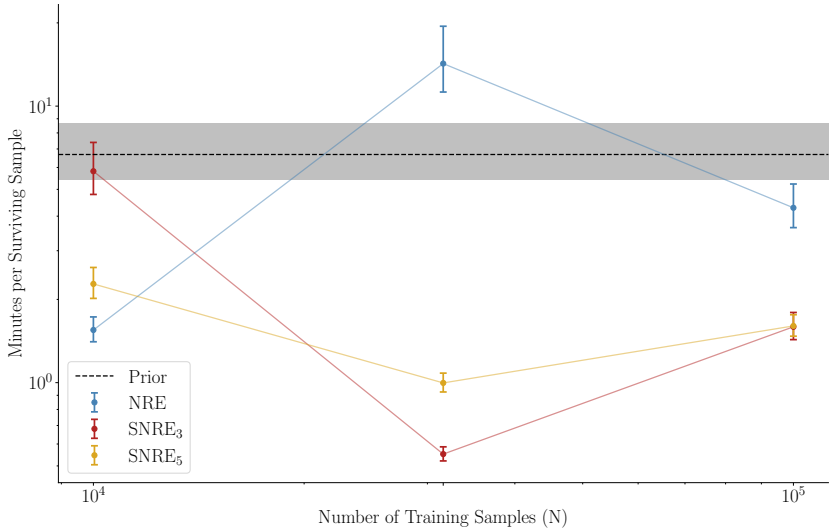

Figure 4.2: The expected amount of time (minutes) needed to obtain one sample in the ranges specified in Eq. 4.5.

compared to NRE even when the inference set is significantly smaller. The results suggest sequential algorithms are consistently the most efficient when the $\mathcal{O}(10\%)$ of the allotted calculations are used for training.

4.5 Application to pMSSM

4.5.1 SBI Setup

We now examine the posteriors and experimental observables produced by an application to the pMSSM. In this search, we are interested in finding the regions of parameter space which are most likely to contain the correct relic density, Higgs mass, anomalous magnetic moment of the muon, and are not excluded by Xenon1T.

Inspired by the results of section 4.4, we run SNRE for a total of 10 rounds, each with 20,000 new samples calculated with microMEGAS. The likelihood-to-evidence

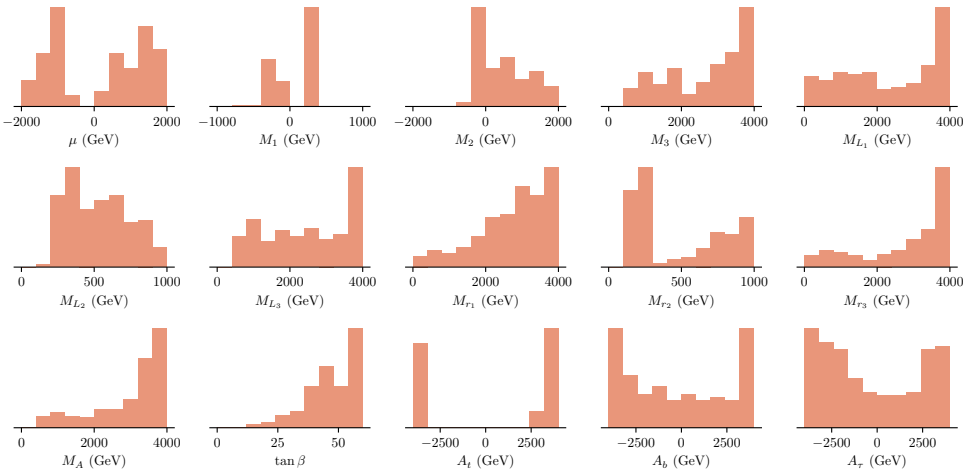

Figure 4.3: Samples from the posterior which lie within the experimental constraints specified in Eq. 4.5 obtained by running SNRE with hyperparameters listed in Table A.3 . We note the overdensity of small smuon masses which affect a_μ and Ω_χ by coannhiliations. See Figures 4.5 and 4.4 for corner plots with subsets of these parameters.

approximators are made up of an ensemble of 5 binary classifiers whose inputs are a concatenation of X and θ and whose outputs are the logit of the classification probability. Each network has 3 hidden layers with 256 units each, and is trained to minimize the binary cross-entropy loss between samples from the likelihood, $p(X|\theta)$, and the joint probability, $p(X)p(\theta)$. During inference time, the ensemble outputs are averaged together to produce approximate likelihood-to-evidence ratio of a given sample. When we sample with HMC during intermediate training steps, we use 20 chains with 2000 warmup steps. During final inference time, we use 32 chains with 2000 warmup steps.

4.5.2 Results

Samples from the approximate posterior are shown as one-dimensional histograms in Fig. 4.3. The multi-modal and irregularly shaped distribution of points shows the utility of flexible neural methods which are able to hone in on these regions. The results exhibit several features existent in the literature, but were not enforced a priori in the

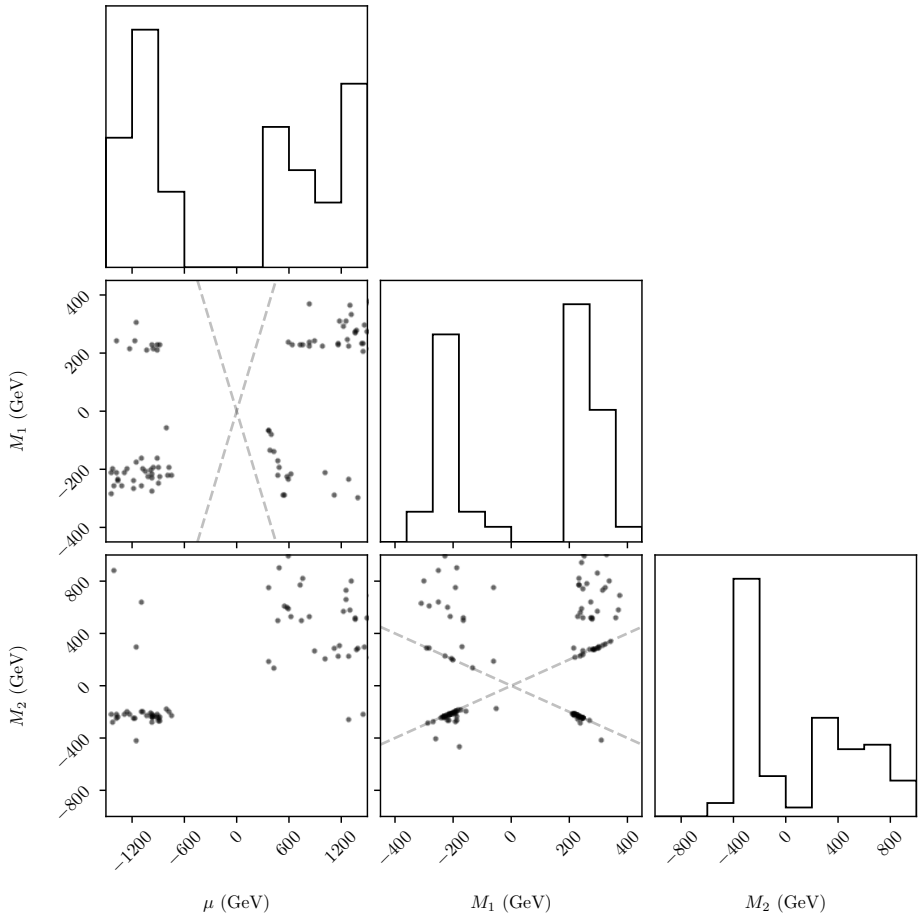

Figure 4.4: Corner plots of μ and gaugino mass parameters created from the same samples as Fig 4.3. Dotted lines correspond to $\mu = M_1$ and $M - 2 = M_1$, respectively. We see a strong bias towards bino-like and wino-like LSPs. Direct detection constraints force μ towards larger values of its range. Additionally, we see a dependence on the relative signs of μ and M_2.

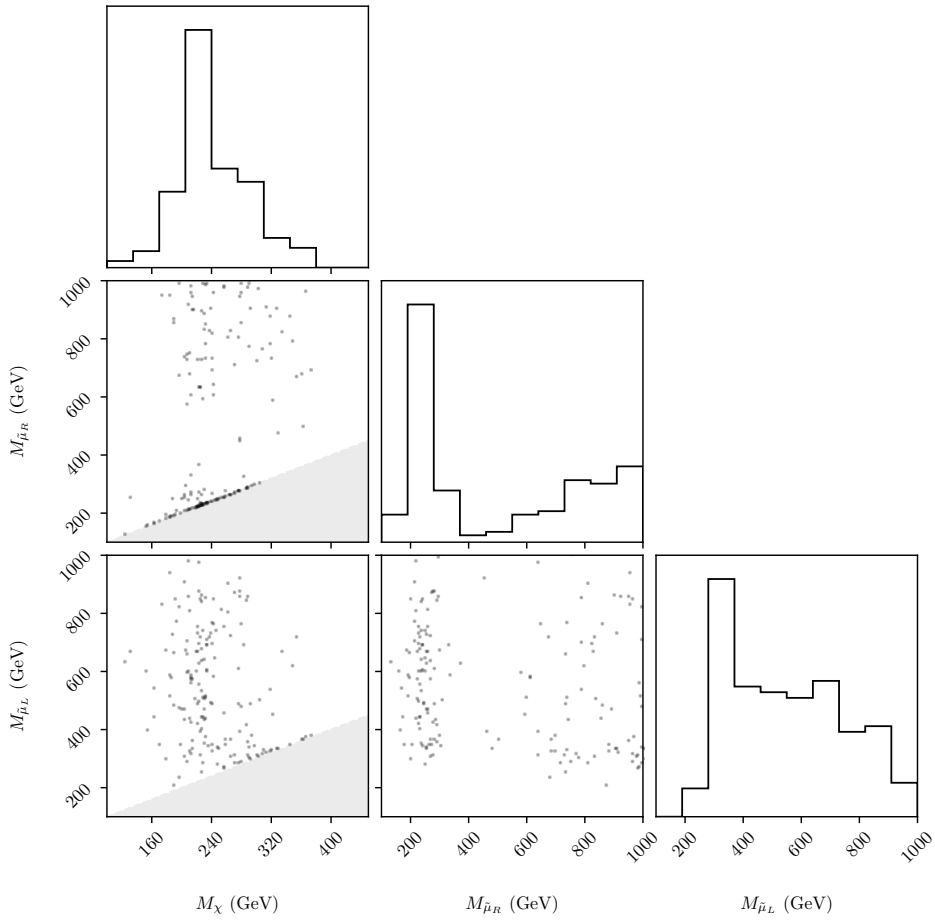

Figure 4.5: Corner plots of smuon and dark matter masses calculated from the same samples as Fig 4.3. Shaded regions do not have a neutralino LSP which, although not explicitly excluded, are not experimentally viable. We note the large concentration of light right-handed smuons.

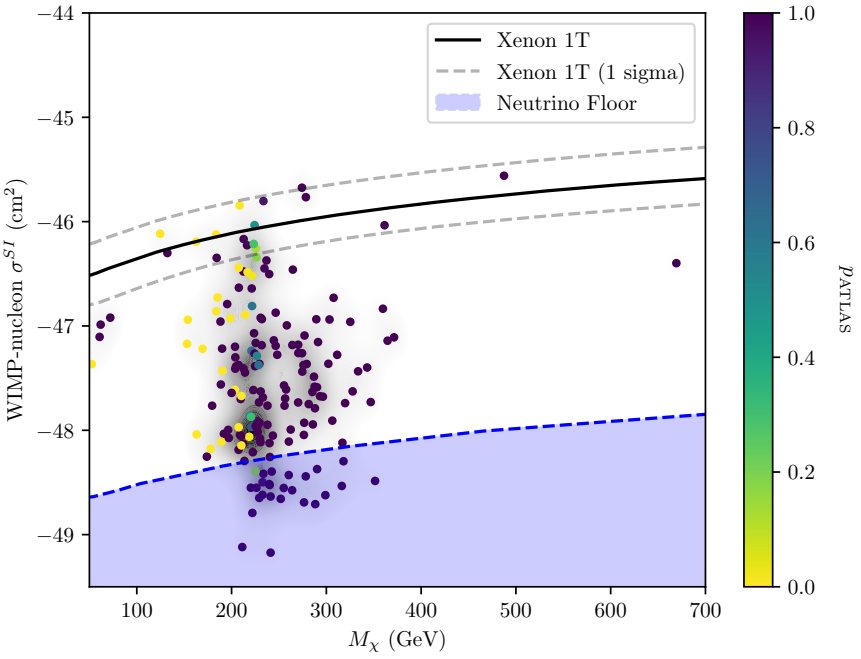

Figure 4.6: Samples from the approximate posterior which lie within the experimental constraints specified in Eq. 4.5. The color of each point corresponds to the p-value from compressed spectra constraints reported by ATLAS. The background is shaded according to a gaussian kernel density estimate in order to visualize the concentration of points on these axes. We note the ability of ATLAS to probe models which lie in the neutrino floor.

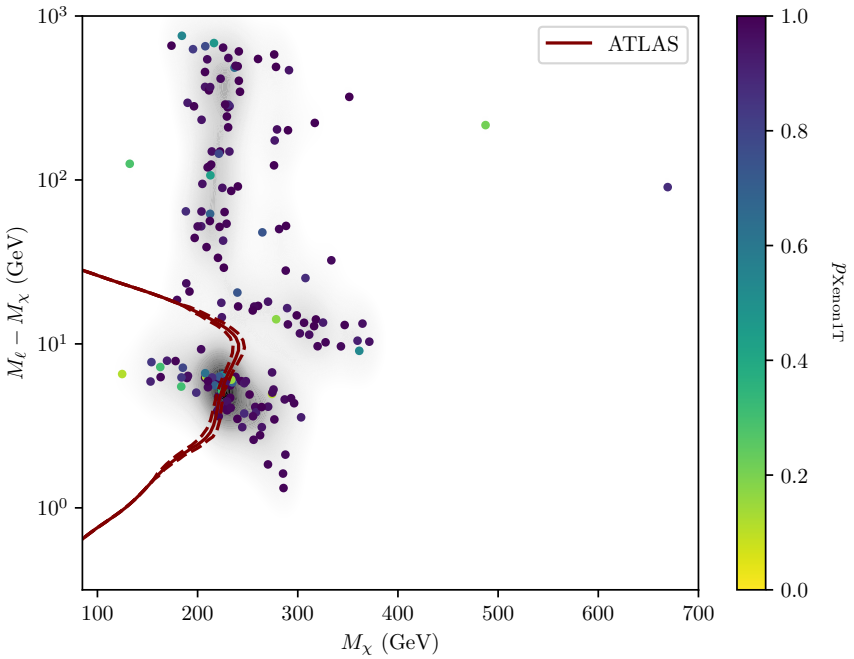

Figure 4.7: Same as Fig. 4.6 but plotted on ATLAS constraints. We expect future analyses to constrain much of the viable parameter space with small mass splittings.

search, outside of choosing bounds on the prior. Additionally, we plot two-dimensional histograms for a subset of the dimensions in Figures 4.4 and 4.5.

We can see in these, for example, the learned dependence on the relative signs of μ and the gaugino parameters. In addition, we find that most models have a bino-like LSP with light smuons whose masses are near the LSP mass. These are expected as bino-like LSPs have smaller WIMP-nucleon cross sections, and lighter sleptons contribute more to a_μ. We note the generally larger values of M_{L_2} compared to M_{r_2}. This is mostly due to the sneutrino requiring a larger value of M_{L_2} because its mass is typically lighter than the left-handed smuon. The right-handed smuon mass, on the other hand, is generally equal to M_{r_2}.

Of the 96,000 samples generated from the approximated posterior, 179 survive the cuts in Eq 4.5 resulting in an efficiency of 0.00186, similar to that shown in Fig 4.1. These remaining points are shown on Xenon1T and ATLAS constraints in Figures 4.6 and 4.7. Interestingly, we see the experiments are complimentary: regions of low direct detection constraining power exist in regions of high ATLAS constraining power, and vice versa. We expect future analyses to probe the majority of the surviving parameter space.

4.6 Conclusions

In this work we introduce the simulation-based inference framework to analyses of parameter spaces as defined by beyond the standard model theories. We show how equivalent amounts of computational time result in orders of magnitude more pMSSM models of experimental interest as compared to naive sampling from the prior. Additionally, by performing a light scan over hyperparameters, we demonstrate how sequential sampling methods are even more efficient than non-sequential counterparts.

Finally, we use these methods to sample from the parameter space of the pMSSM

which is responsible for dark matter relic density, Higgs mass, and the anomalous magnetic moment of the muon which has not yet been excluded by direct detection experiments. We find the most likely regions of this space are tantalizingly close to current ATLAS bounds, and are likely to be covered in future experiments.

Future exercises like this one can make use of calibration methods presented [27] to make sure constraints on the approximated posterior are not overly confident (and therefore excluding viable regions of parameter space). Additionally, higher-dimensional observables may be taken into consideration. Rather than summaries of observed weak-scale quantities, it is potentially of interest to operate on the raw obtained from experiment.

Appendix A

Appendix

A.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. 2.4.5. All results are summarized in Table A.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 before reconstructing the continuum.

Table A.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

A.2 Additional Samples

We show here (Fig. A.1) 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`.

A.3 Model Hyperparameters

In Table A.3 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.

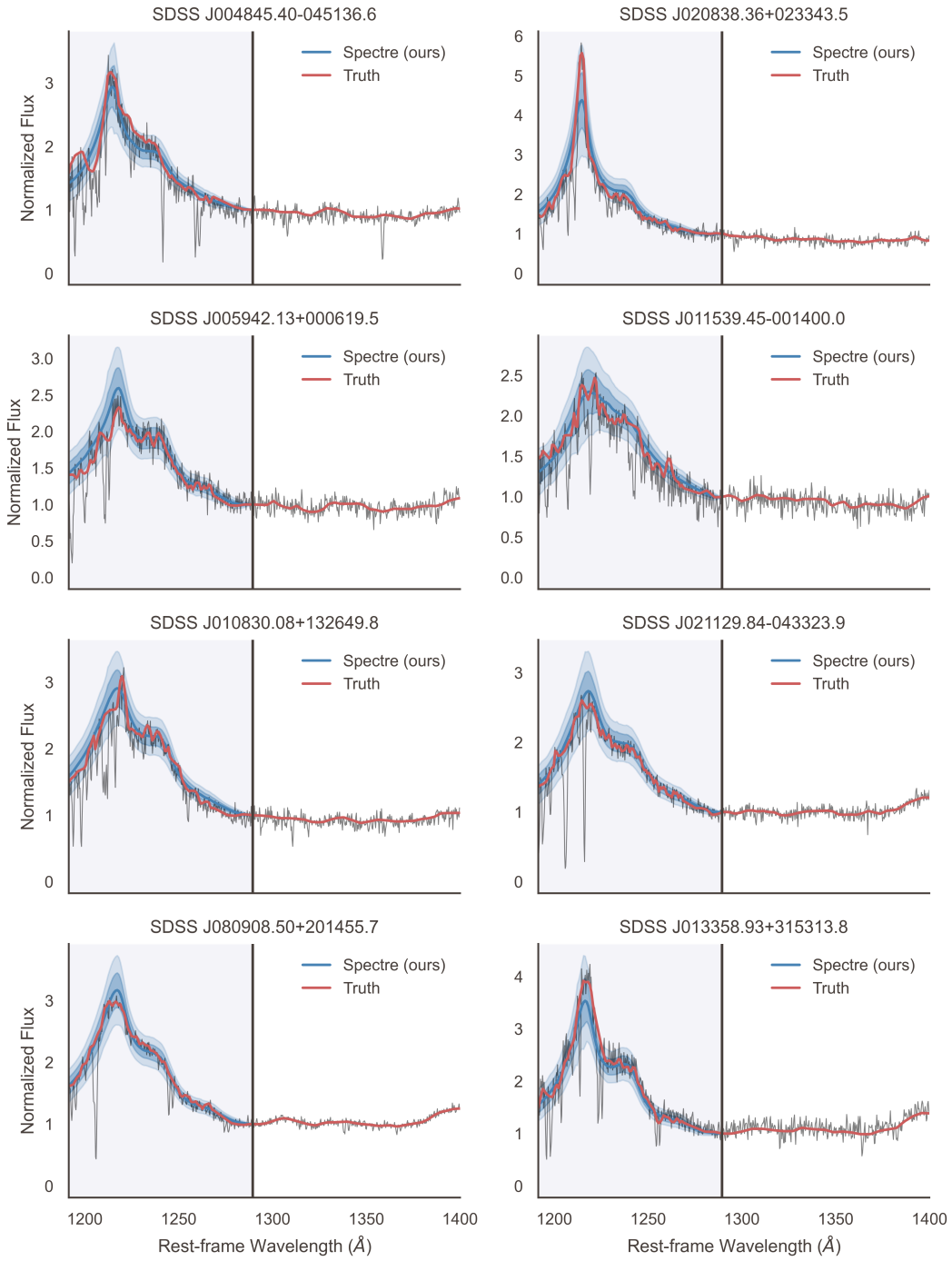

Figure A.1: A random selection of eight predictions on moderate redshift test set spectra from eBOSS. SPECTRE's mean predictions are displayed in blue alongside its 1- and 2-sigma estimates. The assumed truth, shown in red, is the smoothed continua estimation from our preprocessing scheme, and the raw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

Table A.2: Model and training hyperparameters, where the leftmost column lists the variable name we use in our codebase, 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

A.4 Details of the MAF

A.4.1 Training and model selection

For each SR – defined by a patch of the sky and a slice in μ_λ – we train two separate MAFs, one on the stars $\mu_\lambda \in [\mu_\lambda^{\rm min}, \mu_\lambda^{\rm max}]$ in the SR and one on the stars in its complement (the CR), $\mu_\lambda \notin [\mu_\lambda^{\rm min}, \mu_\lambda^{\rm max}]$. Before training, the data is standardized by shifting the mean in each feature to zero and normalizing the standard deviation to unity.

We opt not to divide the data up into training and validation sets, as doing so would dilute the significance of any stream detection. Based on direct inspection, we do not find any evidence for overfitting. For the density estimation, overfitting would typically correspond to $p(x)$ degenerating into a set of delta functions centered on each point in the training data. This is generally not a concern for the MAF, and in fact it generally has the opposite problem (not being able to fit extremely sharp distributions). Also, the lower bound on dataset size (SRs must have at least 20,000 stars, otherwise they are rejected) should be sufficient to mitigate overfitting for the dimensionality of the feature space.

For each MAF, we train for 150 epochs using the Adam optimizer [79]. The learning rate is a hyperparameter that will be included in the scan to be described in Sec. A.4.2. This number of epochs seemed to be sufficient for convergence, and training for significantly longer is computationally prohibitive. To smooth out fluctuations in the MAF from epoch to epoch arising from stochastic gradient descent, we calculate a running average for each star's probability density over the output of 20 consecutive training epochs.

To select the best model for each MAF, we employ the following approach. On general grounds, we expect the $\log R$ distribution to be roughly symmetric around 0 in

the absence of any signal; any deviation from $R = 1$ is due to random fluctuations in the MAFs estimating the numerator or the denominator of the likelihood ratio. The better the performance of the density estimation, the more sharply peaked the R distribution should be around $R = 1$. Furthermore, we expect astrophysical signals (such as streams) to typically correspond to *over*densities, not underdensities. Putting all this together, we select the "best" epoch by considering the $\log R$ distribution for $\log R < 0$, reflecting this across 0, and choosing the epoch with the smallest standard deviation in this symmetrized distribution. Strictly speaking, we only perform this for the MAF trained on the CR; for the MAF trained on the SR, we take the last 20 epochs, as we found this led to the best performance on tests with the labelled GD-1 stars.

A.4.2 Hyperparameter Optimization

Here we describe the hyperparameter optimization for the MAF neural network used for density estimation in this work. For the MAF architecture, these hyperparameters include the number of blocks in the neural network that make up the affine transformations, the number of hidden layers in the network, and the number of nodes in each hidden layer. Then there are the usual hyperparameters involved in training (mini-batch size, learning rate, etc.). The optimal values for these hyperparameters are not derivable from first principles; instead, we must perform a scan over the hyperparameters and select the configuration that maximizes the performance of the neural network. To measure performance, we will use the GD-1 labelled stars from [PWB18] and quantify the signal/background discrimination power of the ANODE method as in Sec. 3.3.2.

To optimize the hyperparameters, we used a 15° patch of the sky centered on $(\alpha, \delta) = (140°, 30°)$, which contains a segment of the GD-1 stream (this patch was hand-selected, and is not one of the 200 centers described in Sec. 3.2). The patch contains 1.2×10^6 stars, of which 574 were tagged as stream stars by [PWB18]. The

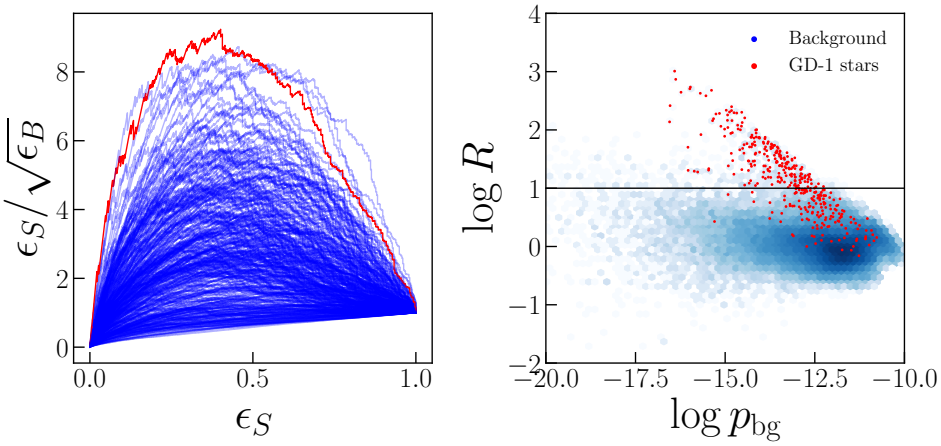

Figure A.2: Left: SIC curve of signal efficiency ϵ_S to $\epsilon_S/\sqrt{\epsilon_B}$ (for a background efficiency ϵ_B) as a cut is placed on $\log R$, for all hyperparameters tested on the GD-1 example dataset. Right: Density plot of $\log p_{\rm bg}$ versus $\log R$ for stars in the signal region of the GD-1 dataset used for hyperparameter optimization, trained using the neural network parameters that maximize the true-positive over root false-positive rate.

inner 10° fiducial region has 4.3×10^5 stars, 374 of which are stream-tagged.

Using these tagged stars, we hand-pick an SR defined by $\mu_\alpha^* \in [-8.75, -15]$ mas/yr, which contains all the stream stars and 1.7×10^5 total stars (6.4×10^4 in the fiducial region).

We varied the hyperparameters over batch size, number of epochs, learning rates, number of blocks, and number of hidden layers with:

$$\begin{aligned} \text{batch size} &= [256, 512, 1024] \\ \text{num. blocks} &= [16, 18, 20, 25] \\ \text{num. hidden} &= [16, 32, 64] \\ \text{num. epochs} &= [125, 150, 175] \\ \text{learning rate} &= [5 \times 10^{-5}, 7 \times 10^{-5}, 2 \times 10^{-5}, 4 \times 10^{-5}]. \end{aligned} \tag{A.1}$$

We train the MAF for each combination of these five parameters. For each hyperparameter set in the scan, we calculate a $\log R$ value for each star in the fiducial (inner 10°) region. After training, we use the last epoch to construct the significance improvement characteristic (SIC) curve by varying a cut on $\log R$. The SIC curves for each hyperparameter configuration in the scan are plotted in the left panel of Fig. A.2, with the optimal choice that maximizes $\epsilon_S/\sqrt{\epsilon_B}$ highlighted in red. On the right panel of Figure A.2, we show the distribution of $\log p_{\text{bg}}$ versus $\log R$ for stars in the SR for the optimal set of hyperparameters. The optimal hyperparameters – 150 epochs, a batch size of 512, 20 blocks, 64 hidden blocks, and a learning rate of 7×10^{-5} – are then used for all MAF trainings in this work.

A.5 Globular cluster detection

Here we describe the simple algorithm we use to remove SRs that contain a suspected globular cluster. The presence of such overdensities in an SR is enough to distort the density estimation; the MAF cannot fit the delta function that is a globular cluster while simultaneously accurately describing the rest of the patch.

Based upon inspecting many patches pre- and post-ANODE, we find that the GCs that spoil the MAF are usually visible as a single bright pixel in a simple two-dimensional (2D) density plot of the stars' positions in an SR. Given that there are thousands of SRs to sift through, we make a 2D histogram of the latitude and longitude of all the stars in an SR (recall, the patch size is 15°). With some tuning, we find that a good resolution is 120×120 bins across the $15° \times 15°$ region. We then compute the mean number of counts $\bar{N}$, the max number of counts N_{max}, and the standard deviation of the number of counts (as measured by the inter-quartile range) σ. We declare the SR to contain a

likely GC if

$$\frac{N_{\rm max} - \bar{N}}{\sigma} > 4 \qquad \text{and} \qquad N_{\rm max} > 25. \tag{A.2}$$

In other words, the bin with the maximum number of stars had to have at least 25 stars, and had to be at least "4σ" significant over the background stellar distribution.

Using these simple criteria, we find 1,381 (out of 6,117) SRs in the full-sky dataset contain a GC candidate. We have visually inspected all of the SRs containing GC candidates and confirmed that the selections appear to be reasonable.

A.6 Comments on stream 2 and disk stars

Here we elaborate further on the second, less prominent stream candidate tagged by the full Via Machinae algorithm in the 21 patches containing GD-1. As described in Sec. 3.4, this stream candidate is contained completely in a single patch (centered on $(\alpha, \delta) = (138.8^\circ, 25.1^\circ)$), and all of the high-$R$ stars follow tightly the edge of the circular patch, aligned and on the same side as the Galactic disk. We illustrate this further in Fig. A.3, which show the stars of the stream candidates in angular position, proper motion and color/magnitude space, overlaid on top of density plots of all the stars in the patch containing the stream candidate. This shows clearly how the stream candidate is aligned with the density gradient in the patch (which in turn is aligned with the Galactic disk, which one can check by transforming to Galactic coordinates (ℓ, b)). We also see that the stream candidate is clustered in proper motion space close to $\mu_\alpha^*,\ \mu_\delta \sim 0$, which as we have noted in Sec. 3.3.3 is a significant source of false positives for the ANODE method. Finally, we note that (unlike for GD-1 and other known streams), there is no noticeable correlation between the position along the stream and the proper motion. Taken together, we view this as strong evidence that this second stream candidate is likely to be a false positive.

More generally, we observe a strong gradient in stellar density towards the Galactic disk in many patches and SRs. There is also likely a strong correlation between disk stars and proper motion within a patch.[1] Therefore, it is potentially concerning that Via Machinae could systematically misidentify disk stars as stream stars.

A conservative approach to avoid this misidentification is to reject all ROIs where the line-finder returned best fit parameters that are at the edge of the patch closest to the Galactic disk and parallel to it. Specifically, we propose to cut out all ROIs whose best-fit line radius has $|\rho| > 9.5°$, slope less than 0.2 radians in Galactic ℓ, b coordinates (that is, aligned with the disk), and are localized on the side of the patch nearest to the disk. This requirement removes only 91 ROIs (out of $\approx 17,000$) from our sample. Such cut would eliminate the second stream that we find in Sec. 3.4, but it would not affect the GD-1 stream candidate at all.

[1] Understanding this correlation requires modeling stellar orbits in the Milky Way, and a detailed understanding of projection and line-of-sight effects. This is beyond the scope of the present work.

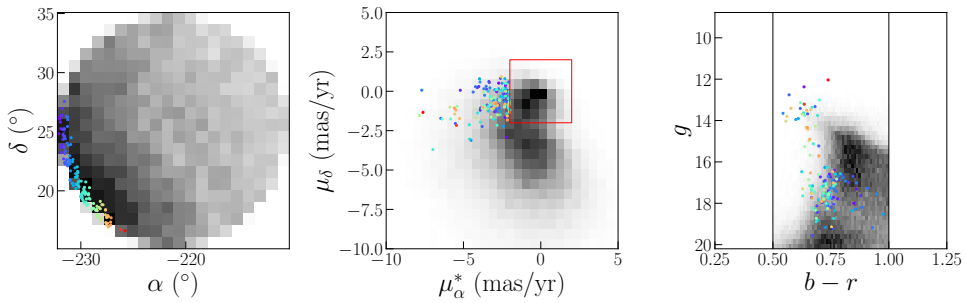

Figure A.3: Scatter plots of the angular positions, proper motions, and color/magnitudes of the stars in the second, less prominent stream candidate identified by Via Machinae, overlaid on 2d histograms of all the stars in the circular patch that contains this stream candidate (darker pixels indicate higher density of stars). As in Fig. 3.13, the Via Machinae stars are color-coded by position in α, to facilitate cross referencing between the three individual scatter plots.

A.7 Training details

Prior to entering neural network, the data are pre-processed to restrict the dynamic range of the inputs.

$$\begin{aligned} \Omega_\chi &\to 10 \log_{10} \left(\text{clip}(\Omega_\chi, 10^{-4}, \infty) \right) \\ m_h &\to \frac{\text{clip}(m_h, 118, \infty) - 123.864}{2.2839} \\ a_\mu &\to \log_{10} \text{clip}(a_\mu, 10^{-11}, 10^{-8}) + 9.5 \\ p_{\text{Xenon1T}} &\to \text{clip}(p_{\text{Xenon1T}}, 0, 0.0455) \times 10 \end{aligned} \tag{A.3}$$

where

$$\text{clip}(x, a, b) = \begin{cases} a & \text{if } x < a \\ b & \text{if } x > b \\ x & \text{otherwise} \end{cases} \tag{A.4}$$

We make use the LBI package [147] with JAX [22] and Flax [58] backends to intialize,fit the models in order to perform approximate bayesian inference. We use the optax [65] repository's implementation of Adam [78] to optimize our models. We use numpyro's implementation of HMC [131] [15] to sample from the posterior distribution of the model parameters. We run our simulations using microMEGAS v5.2.13 [11] with SOFTSUSY v.4.1.7 [3] backend. Our entire codebase is open-source and can be found here: `https://www.github.com/jtamanas/MSSM`. All computations were run on a machine with an AMD Ryzen 7 3700X processor, 16 GB of RAM, and no GPU.

Hyperparameter	Value
Hidden layers	3
Hidden layer size	256
Ensemble size	5
Learning rate	3×10^{-4}
Weight decay	0
Max gradient norm	10^{-3}
Validation interval	50
Patience	200
Batch size	256
Training split	0.95

Table A.3: Hyperparameters common to all (S)NRE algorithms tested in all applications.

Bibliography

[1] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A.A. Abdelalim, O. Abdinov, R. Aben, B. Abi, M. Abolins, et al. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1–29, 2012.

[2] B. Abi, T. Albahri, S. Al-Kilani, D. Allspach, L.P. Alonzi, A. Anastasi, A. Anisenkov, F. Azfar, K. Badgley, S. Baeßler, et al. Measurement of the positive muon anomalous magnetic moment to 0.46 ppm. *Physical Review Letters*, 126(14), Apr 2021.

[3] B.C. Allanach. Softsusy: A program for calculating supersymmetric spectra. *Computer Physics Communications*, 143(3):305–331, Mar 2002.

[4] Astropy Collaboration, A. M. Price-Whelan, B. M. SipHocz, H. M. G"unther, P. L. Lim, S. M. Crawford, S. Conseil, D. L. Shupe, M. W. Craig, N. Dencheva, A. Ginsburg, J. T. Vand erPlas, L. D. Bradley, D. P'erez-Su'arez, M. de Val-Borro, T. L. Aldcroft, K. L. Cruz, T. P. Robitaille, E. J. Tollerud, C. Ardelean, T. Babej, Y. P. Bach, M. Bachetti, A. V. Bakanov, S. P. Bamford, G. Barentsen, P. Barmby, A. Baumbach, K. L. Berry, F. Biscani, M. Boquien, K. A. Bostroem, L. G. Bouma, G. B. Brammer, E. M. Bray, H. Breytenbach, H. Buddelmeijer, D. J. Burke, G. Calderone, J. L. Cano Rodr'iguez, M. Cara, J. V. M. Cardoso, S. Cheedella, Y. Copin, L. Corrales, D. Crichton, D. D'Avella, C. Deil, 'E. Depagne, J. P. Dietrich, A. Donath, M. Droettboom, N. Earl, T. Erben, S. Fabbro, L. A. Ferreira, T. Finethy, R. T. Fox, L. H. Garrison, S. L. J. Gibbons, D. A. Goldstein, R. Gommers, J. P. Greco, P. Greenfield, A. M. Groener, F. Grollier, A. Hagen, P. Hirst, D. Homeier, A. J. Horton, G. Hosseinzadeh, L. Hu, J. S. Hunkeler, Z. Ivezi'c, A. Jain, T. Jenness, G. Kanarek, S. Kendrew, N. S. Kern, W. E. Kerzendorf, A. Khvalko, J. King, D. Kirkby, A. M. Kulkarni, A. Kumar, A. Lee, D. Lenz, S. P. Littlefair, Z. Ma, D. M. Macleod, M. Mastropietro, C. McCully, S. Montagnac, B. M. Morris, M. Mueller, S. J. Mumford, D. Muna, N. A. Murphy, S. Nelson, G. H. Nguyen, J. P. Ninan, M. N"othe, S. Ogaz, S. Oh, J. K. Parejko, N. Parley, S. Pascual, R. Patil, A. A. Patil, A. L. Plunkett, J. X. Prochaska, T. Rastogi, V. Reddy Janga, J. Sabater, P. Sakurikar, M. Seifert, L. E. Sherbert, H. Sherwood-Taylor, A. Y. Shih, J. Sick, M. T. Silbiger, S. Singanamalla, L. P.

Singer, P. H. Sladen, K. A. Sooley, S. Sornarajah, O. Streicher, P. Teuben, S. W. Thomas, G. R. Tremblay, J. E. H. Turner, V. Terr'on, M. H. van Kerkwijk, A. de la Vega, L. L. Watkins, B. A. Weaver, J. B. Whitmore, J. Woillez, V. Zabalza, and Astropy Contributors. The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package. *aj*, 156(3):123, September 2018.

[5] Astropy Collaboration, T. P. Robitaille, E. J. Tollerud, P. Greenfield, M. Droettboom, E. Bray, T. Aldcroft, M. Davis, A. Ginsburg, A. M. Price-Whelan, W. E. Kerzendorf, A. Conley, N. Crighton, K. Barbary, D. Muna, H. Ferguson, F. Grollier, M. M. Parikh, P. H. Nair, H. M. Unther, C. Deil, J. Woillez, S. Conseil, R. Kramer, J. E. H. Turner, L. Singer, R. Fox, B. A. Weaver, V. Zabalza, Z. I. Edwards, K. Azalee Bostroem, D. J. Burke, A. R. Casey, S. M. Crawford, N. Dencheva, J. Ely, T. Jenness, K. Labrie, P. L. Lim, F. Pierfederici, A. Pontzen, A. Ptak, B. Refsdal, M. Servillat, and O. Streicher. Astropy: A community Python package for astronomy. , 558:A33, October 2013.

[6] 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.

[7] 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.

[8] Nilanjan Banik and Jo Bovy. Effects of baryonic and dark matter substructure on the Pal 5 stream. *MNRAS*, 484(2):2009–2020, April 2019.

[9] Nilanjan Banik, Jo Bovy, Gianfranco Bertone, Denis Erkal, and T. J. L. de Boer. Novel constraints on the particle nature of dark matter from stellar streams. *arXiv e-prints*, page arXiv:1911.02663, November 2019.

[10] Nilanjan Banik, Jo Bovy, Gianfranco Bertone, Denis Erkal, and T. J. L. de Boer. Evidence of a population of dark subhalos from Gaia and Pan-STARRS observations of the GD-1 stream. , January 2021.

[11] G. Bélanger, F. Boudjema, A. Pukhov, and A. Semenov. micrOMEGAs: A tool for dark matter studies. *Nuovo Cimento C Geophysics Space Physics C*, 33(2):111–116, March 2010.

[12] V. Belokurov, D. Erkal, N. W. Evans, S. E. Koposov, and A. J. Deason. Co-formation of the disc and the stellar halo. *MNRAS*, 478(1):611–619, July 2018.

[13] V. Belokurov, D. B. Zucker, N. W. Evans, G. Gilmore, S. Vidrih, D. M. Bramich, H. J. Newberg, R. F. G. Wyse, M. J. Irwin, M. Fellhauer, P. C. Hewett, N. A.

Walton, M. I. Wilkinson, N. Cole, B. Yanny, C. M. Rockosi, T. C. Beers, E. F. Bell, J. Brinkmann, Ž. Ivezić, and R. Lupton. The Field of Streams: Sagittarius and Its Siblings. *Astrophys. J. Lett.*, 642(2):L137–L140, May 2006.

[14] G. W. Bennett, B. Bousquet, H. N. Brown, G. Bunce, R. M. Carey, P. Cushman, G. T. Danby, P. T. Debevec, M. Deile, H. Deng, et al. Final report of the e821 muon anomalous magnetic moment measurement at bnl. *Physical Review D*, 73(7), Apr 2006.

[15] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.

[16] Christopher M Bishop. Mixture density networks. technical report ncrg/94/004, aston university, 1994.

[17] Ana Bonaca, Charlie Conroy, David W. Hogg, Phillip A. Cargile, Nelson Caldwell, Rohan P. Naidu, Adrian M. Price-Whelan, Joshua S. Speagle, and Benjamin D. Johnson. High-resolution Spectroscopy of the GD-1 Stellar Stream Localizes the Perturber near the Orbital Plane of Sagittarius. *Astrophys. J. Lett.*, 892(2):L37, April 2020.

[18] Ana Bonaca, David W. Hogg, Adrian M. Price-Whelan, and Charlie Conroy. The Spur and the Gap in GD-1: Dynamical Evidence for a Dark Substructure in the Milky Way Halo. *Astrophys. J.*, 880(1):38, July 2019.

[19] 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.

[20] Nicholas W. Borsato, Sarah L. Martell, and Jeffrey D. Simpson. Identifying stellar streams in Gaia DR2 with data mining techniques. *MNRAS*, 492(1):1370–1384, February 2020.

[21] Douglas Boubert and Andrew Everall. Completeness of the Gaia verse II: what are the odds that a star is missing from Gaia DR2? , 497(4):4246–4261, October 2020.

[22] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[23] R. G. Carlberg, C. J. Grillmair, and Nathan Hetherington. The Pal 5 Star Stream Gaps. *Astrophys. J.*, 760(1):75, November 2012.

[24] S. Chatrchyan, V. Khachatryan, A.M. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, et al. Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc. *Physics Letters B*, 716(1):30–61, 2012.

[25] 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.

[26] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *arXiv e-prints*, page arXiv:1911.01429, November 2019.

[27] Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating Likelihood Ratios with Calibrated Discriminative Classifiers. *arXiv e-prints*, page arXiv:1506.02169, June 2015.

[28] 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.

[29] 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.

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

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

[32] A. Djouadi, S. Rosier-Lees, M. Bezouh, M. A. Bizouard, C. Boehm, F. Borzumati, C. Briot, J. Carr, M. B. Causse, F. Charles, X. Chereau, P. Colas, L. Duflot, A. Dupperin, A. Ealet, H. El-Mamouni, N. Ghodbane, F. Gieres, B. Gonzalez-Pineiro, S. Gourmelen, G. Grenier, Ph. Gris, J. F. Grivaz, C. Hebrard, B. Ille, J. L. Kneur, N. Kostantinidis, J. Layssac, P. Lebrun, R. Ledu, M. C. Lemaire, Ch. LeMouel, L. Lugnier, Y. Mambrini, J. P. Martin, G. Montarou, G. Moultaka, S. Muanza, E. Nuss, E. Perez, F. M. Renard, D. Reynaud, L. Serin, C. Thevenet, A. Trabelsi, F. Zach, and D. Zerwas. The Minimal Supersymmetric Standard Model: Group Summary Report. *arXiv e-prints*, pages hep–ph/9901246, January 1999.

[33] Richard O. Duda and Peter E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, January 1972.

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

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

[36] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows, 2019.

[37] Dominika Ďurovčíková, Harley Katz, Sarah EI Bosman, Frederick B Davies, Julien Devriendt, and Adrianne Slyz. Reionization history constraints from neural network based predictions of high-redshift quasar continua. *Monthly Notices of the Royal Astronomical Society*, 493(3):4256–4275, 2020.

[38] 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.

[39] 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.

[40] Denis Erkal, Sergey E. Koposov, and Vasily Belokurov. A sharper view of Pal 5's tails: discovery of stream perturbations with a novel non-parametric technique. *MNRAS*, 470(1):60–84, September 2017.

[41] 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.

[42] 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.

[43] A. S. Font, I. G. McCarthy, R. A. Crain, T. Theuns, J. Schaye, R. P. C. Wiersma, and C. Dalla Vecchia. Cosmological simulations of the formation of the stellar haloes around disc galaxies. , 416(4):2802–2820, October 2011.

[44] Andreea S. Font, Kathryn V. Johnston, James S. Bullock, and Brant E. Robertson. Phase-Space Distributions of Chemical Abundances in Milky Way-Type Galaxy Halos. , 646(2):886–898, August 2006.

[45] Gaia Collaboration, A. G. A. Brown, A. Vallenari, T. Prusti, J. H. J. de Bruijne, C. Babusiaux, C. A. L. Bailer-Jones, M. Biermann, D. W. Evans, L. Eyer, F. Jansen, C. Jordi, S. A. Klioner, U. Lammers, L. Lindegren, X. Luri, F. Mignard, C. Panem, D. Pourbaix, S. Randich, P. Sartoretti, H. I. Siddiqui, C. Soubiran, F. van Leeuwen, N. A. Walton, F. Arenou, U. Bastian, M. Cropper, R. Drimmel, D. Katz, M. G. Lattanzi, J. Bakker, C. Cacciari, J. Castañeda, L. Chaoul, N. Cheek, F. De Angeli, C. Fabricius, R. Guerra, B. Holl, E. Masana, R. Messineo, N. Mowlavi, K. Nienartowicz, P. Panuzzo, J. Portell, M. Riello, G. M. Seabroke, P. Tanga, F. Thévenin, G. Gracia-Abril, G. Comoretto, M. Garcia-Reinaldos, D. Teyssier, M. Altmann, R. Andrae, M. Audard, I. Bellas-Velidis, K. Benson, J. Berthier, R. Blomme, P. Burgess, G. Busso, B. Carry, A. Cellino, G. Clementini, M. Clotet, O. Creevey, M. Davidson, J. De Ridder, L. Delchambre, A. Dell'Oro, C. Ducourant, J. Fernández-Hernández, M. Fouesneau, Y. Frémat, L. Galluccio, M. García-Torres, J. González-Núñez, J. J. González-Vidal, E. Gosset, L. P. Guy, J. L. Halbwachs, N. C. Hambly, D. L. Harrison, J. Hernández, D. Hestroffer, S. T. Hodgkin, A. Hutton, G. Jasniewicz, A. Jean-Antoine-Piccolo, S. Jordan, A. J. Korn, A. Krone-Martins, A. C. Lanzafame, T. Lebzelter, W. Löffler, M. Manteiga, P. M. Marrese, J. M. Martín-Fleitas, A. Moitinho, A. Mora, K. Muinonen, J. Osinde, E. Pancino, T. Pauwels, J. M. Petit, A. Recio-Blanco, P. J. Richards, L. Rimoldini, A. C. Robin, L. M. Sarro, C. Siopis, M. Smith, A. Sozzetti, M. Süveges, J. Torra, W. van Reeven, U. Abbas, A. Abreu Aramburu, S. Accart, C. Aerts, G. Altavilla, M. A. Álvarez, R. Alvarez, J. Alves, R. I. Anderson, A. H. Andrei, E. Anglada Varela, E. Antiche, T. Antoja, B. Arcay, T. L. Astraatmadja, N. Bach, S. G. Baker, L. Balaguer-Núñez, P. Balm, C. Barache, C. Barata, D. Barbato, F. Barblan, P. S. Barklem, D. Barrado, M. Barros, M. A. Barstow, S. Bartholomé Muñoz, J. L. Bassilana, U. Becciani, M. Bellazzini, A. Berihuete, S. Bertone, L. Bianchi, O. Bienaymé, S. Blanco-Cuaresma, T. Boch, C. Boeche, A. Bombrun, R. Borrachero, D. Bossini, S. Bouquillon, G. Bourda, A. Bragaglia, L. Bramante, M. A. Breddels, A. Bressan, N. Brouillet, T. Brüsemeister, E. Brugaletta, B. Bucciarelli, A. Burlacu, D. Busonero, A. G. Butkevich, R. Buzzi, E. Caffau, R. Cancelliere, G. Cannizzaro, T. Cantat-Gaudin, R. Carballo, T. Carlucci, J. M. Carrasco, L. Casamiquela, M. Castellani, A. Castro-Ginard, P. Charlot, L. Chemin, A. Chiavassa, G. Cocozza, G. Costigan, S. Cowell, F. Crifo, M. Crosta, C. Crowley, J. Cuypers, C. Dafonte, Y. Damerdji, A. Dapergolas, P. David, M. David, P. de Laverny, F. De Luise, R. De March, D. de Martino, R. de Souza, A. de Torres, J. Debosscher, E. del Pozo, M. Delbo, A. Delgado, H. E. Delgado, P. Di Matteo, S. Diakite, C. Diener, E. Distefano, C. Dolding, P. Drazinos, J. Durán, B. Edvardsson, H. Enke, K. Eriksson, P. Esquej, G. Eynard Bontemps, C. Fabre, M. Fabrizio, S. Faigler, A. J. Falcão, M. Farràs Casas, L. Federici, G. Fedorets, P. Fernique, F. Figueras, F. Filippi, K. Findeisen, A. Fonti, E. Fraile, M. Fraser, B. Frézouls, M. Gai, S. Galleti, D. Garabato, F. García-Sedano, A. Garofalo, N. Garralda, A. Gavel, P. Gavras,

J. Gerssen, R. Geyer, P. Giacobbe, G. Gilmore, S. Girona, G. Giuffrida, F. Glass, M. Gomes, M. Granvik, A. Gueguen, A. Guerrier, J. Guiraud, R. Gutiérrez-Sánchez, R. Haigron, D. Hatzidimitriou, M. Hauser, M. Haywood, U. Heiter, A. Helmi, J. Heu, T. Hilger, D. Hobbs, W. Hofmann, G. Holland, H. E. Huckle, A. Hypki, V. Icardi, K. Janßen, G. Jevardat de Fombelle, P. G. Jonker, Á. L. Juhász, F. Julbe, A. Karampelas, A. Kewley, J. Klar, A. Kochoska, R. Kohley, K. Kolenberg, M. Kontizas, E. Kontizas, S. E. Koposov, G. Kordopatis, Z. Kostrzewa-Rutkowska, P. Koubsky, S. Lambert, A. F. Lanza, Y. Lasne, J. B. Lavigne, Y. Le Fustec, C. Le Poncin-Lafitte, Y. Lebreton, S. Leccia, N. Leclerc, I. Lecoeur-Taibi, H. Lenhardt, F. Leroux, S. Liao, E. Licata, H. E. P. Lindstrøm, T. A. Lister, E. Livanou, A. Lobel, M. López, S. Managau, R. G. Mann, G. Mantelet, O. Marchal, J. M. Marchant, M. Marconi, S. Marinoni, G. Marschalkó, D. J. Marshall, M. Martino, G. Marton, N. Mary, D. Massari, G. Matijevič, T. Mazeh, P. J. McMillan, S. Messina, D. Michalik, N. R. Millar, D. Molina, R. Molinaro, L. Molnár, P. Montegriffo, R. Mor, R. Morbidelli, T. Morel, D. Morris, A. F. Mulone, T. Muraveva, I. Musella, G. Nelemans, L. Nicastro, L. Noval, W. O'Mullane, C. Ordénovic, D. Ordóñez-Blanco, P. Osborne, C. Pagani, I. Pagano, F. Pailler, H. Palacin, L. Palaversa, A. Panahi, M. Pawlak, A. M. Piersimoni, F. X. Pineau, E. Plachy, G. Plum, E. Poggio, E. Poujoulet, A. Prša, L. Pulone, E. Racero, S. Ragaini, N. Rambaux, M. Ramos-Lerate, S. Regibo, C. Reylé, F. Riclet, V. Ripepi, A. Riva, A. Rivard, G. Rixon, T. Roegiers, M. Roelens, M. Romero-Gómez, N. Rowell, F. Royer, L. Ruiz-Dern, G. Sadowski, T. Sagristà Sellés, J. Sahlmann, J. Salgado, E. Salguero, N. Sanna, T. Santana-Ros, M. Sarasso, H. Savietto, M. Schultheis, E. Sciacca, M. Segol, J. C. Segovia, D. Ségransan, I. C. Shih, L. Siltala, A. F. Silva, R. L. Smart, K. W. Smith, E. Solano, F. Solitro, R. Sordo, S. Soria Nieto, J. Souchay, A. Spagna, F. Spoto, U. Stampa, I. A. Steele, H. Steidelmüller, C. A. Stephenson, H. Stoev, F. F. Suess, J. Surdej, L. Szabados, E. Szegedi-Elek, D. Tapiador, F. Taris, G. Tauran, M. B. Taylor, R. Teixeira, D. Terrett, P. Teyssand ier, W. Thuillot, A. Titarenko, F. Torra Clotet, C. Turon, A. Ulla, E. Utrilla, S. Uzzi, M. Vaillant, G. Valentini, V. Valette, A. van Elteren, E. Van Hemelryck, M. van Leeuwen, M. Vaschetto, A. Vecchiato, J. Veljanoski, Y. Viala, D. Vicente, S. Vogt, C. von Essen, H. Voss, V. Votruba, S. Voutsinas, G. Walmsley, M. Weiler, O. Wertz, T. Wevers, Ł. Wyrzykowski, A. Yoldas, M. Žerjal, H. Ziaeepour, J. Zorec, S. Zschocke, S. Zucker, C. Zurbach, and T. Zwitter. Gaia Data Release 2. Summary of the contents and survey properties. *Astron. & Astrophys.*, 616:A1, August 2018.

[46] Gaia Collaboration, A. G. A. Brown, A. Vallenari, T. Prusti, J. H. J. de Bruijne, C. Babusiaux, M. Biermann, O. L. Creevey, D. W. Evans, L. Eyer, A. Hutton, F. Jansen, C. Jordi, S. A. Klioner, U. Lammers, L. Lindegren, X. Luri, F. Mignard, C. Panem, D. Pourbaix, S. Randich, P. Sartoretti, C. Soubiran, N. A. Walton, F. Arenou, C. A. L. Bailer-Jones, U. Bastian, M. Cropper, R. Drimmel,

D. Katz, M. G. Lattanzi, F. van Leeuwen, J. Bakker, C. Cacciari, J. Castañeda, F. De Angeli, C. Ducourant, C. Fabricius, M. Fouesneau, Y. Frémat, R. Guerra, A. Guerrier, J. Guiraud, A. Jean-Antoine Piccolo, E. Masana, R. Messineo, N. Mowlavi, C. Nicolas, K. Nienartowicz, F. Pailler, P. Panuzzo, F. Riclet, W. Roux, G. M. Seabroke, R. Sordo, P. Tanga, F. Thévenin, G. Gracia-Abril, J. Portell, D. Teyssier, M. Altmann, R. Andrae, I. Bellas-Velidis, K. Benson, J. Berthier, R. Blomme, E. Brugaletta, P. W. Burgess, G. Busso, B. Carry, A. Cellino, N. Cheek, G. Clementini, Y. Damerdji, M. Davidson, L. Delchambre, A. Dell'Oro, J. Fernández-Hernández, L. Galluccio, P. García-Lario, M. Garcia-Reinaldos, J. González-Núñez, E. Gosset, R. Haigron, J. L. Halbwachs, N. C. Hambly, D. L. Harrison, D. Hatzidimitriou, U. Heiter, J. Hernández, D. Hestroffer, S. T. Hodgkin, B. Holl, K. Janßen, G. Jevardat de Fombelle, S. Jordan, A. Krone-Martins, A. C. Lanzafame, W. Löffler, A. Lorca, M. Manteiga, O. Marchal, P. M. Marrese, A. Moitinho, A. Mora, K. Muinonen, P. Osborne, E. Pancino, T. Pauwels, J. M. Petit, A. Recio-Blanco, P. J. Richards, M. Riello, L. Rimoldini, A. C. Robin, T. Roegiers, J. Rybizki, L. M. Sarro, C. Siopis, M. Smith, A. Sozzetti, A. Ulla, E. Utrilla, M. van Leeuwen, W. van Reeven, U. Abbas, A. Abreu Aramburu, S. Accart, C. Aerts, J. J. Aguado, M. Ajaj, G. Altavilla, M. A. Álvarez, J. Álvarez Cid-Fuentes, J. Alves, R. I. Anderson, E. Anglada Varela, T. Antoja, M. Audard, D. Baines, S. G. Baker, L. Balaguer-Núñez, E. Balbinot, Z. Balog, C. Barache, D. Barbato, M. Barros, M. A. Barstow, S. Bartolomé, J. L. Bassilana, N. Bauchet, A. Baudesson-Stella, U. Becciani, M. Bellazzini, M. Bernet, S. Bertone, L. Bianchi, S. Blanco-Cuaresma, T. Boch, A. Bombrun, D. Bossini, S. Bouquillon, A. Bragaglia, L. Bramante, E. Breedt, A. Bressan, N. Brouillet, B. Bucciarelli, A. Burlacu, D. Busonero, A. G. Butkevich, R. Buzzi, E. Caffau, R. Cancelliere, H. Cánovas, T. Cantat-Gaudin, R. Carballo, T. Carlucci, M. I. Carnerero, J. M. Carrasco, L. Casamiquela, M. Castellani, A. Castro-Ginard, P. Castro Sampol, L. Chaoul, P. Charlot, L. Chemin, A. Chiavassa, M. R. L. Cioni, G. Comoretto, W. J. Cooper, T. Cornez, S. Cowell, F. Crifo, M. Crosta, C. Crowley, C. Dafonte, A. Dapergolas, M. David, P. David, P. de Laverny, F. De Luise, R. De March, J. De Ridder, R. de Souza, P. de Teodoro, A. de Torres, E. F. del Peloso, E. del Pozo, M. Delbo, A. Delgado, H. E. Delgado, J. B. Delisle, P. Di Matteo, S. Diakite, C. Diener, E. Distefano, C. Dolding, D. Eappachen, B. Edvardsson, H. Enke, P. Esquej, C. Fabre, M. Fabrizio, S. Faigler, G. Fedorets, P. Fernique, A. Fienga, F. Figueras, C. Fouron, F. Fragkoudi, E. Fraile, F. Franke, M. Gai, D. Garabato, A. Garcia-Gutierrez, M. García-Torres, A. Garofalo, P. Gavras, E. Gerlach, R. Geyer, P. Giacobbe, G. Gilmore, S. Girona, G. Giuffrida, R. Gomel, A. Gomez, I. Gonzalez-Santamaria, J. J. González-Vidal, M. Granvik, R. Gutiérrez-Sánchez, L. P. Guy, M. Hauser, M. Haywood, A. Helmi, S. L. Hidalgo, T. Hilger, N. Hładczuk, D. Hobbs, G. Holland, H. E. Huckle, G. Jasniewicz, P. G. Jonker, J. Juaristi Campillo, F. Julbe, L. Karbevska, P. Kervella, S. Khanna, A. Kochoska, M. Kontizas, G. Kordopatis, A. J. Korn,

Z. Kostrzewa-Rutkowska, K. Kruszyńska, S. Lambert, A. F. Lanza, Y. Lasne, J. F. Le Campion, Y. Le Fustec, Y. Lebreton, T. Lebzelter, S. Leccia, N. Leclerc, I. Lecoeur-Taibi, S. Liao, E. Licata, E. P. Lindstrøm, T. A. Lister, E. Livanou, A. Lobel, P. Madrero Pardo, S. Managau, R. G. Mann, J. M. Marchant, M. Marconi, M. M. S. Marcos Santos, S. Marinoni, F. Marocco, D. J. Marshall, L. Martin Polo, J. M. Martín-Fleitas, A. Masip, D. Massari, A. Mastrobuono-Battisti, T. Mazeh, P. J. McMillan, S. Messina, D. Michalik, N. R. Millar, A. Mints, D. Molina, R. Molinaro, L. Molnár, P. Montegriffo, R. Mor, R. Morbidelli, T. Morel, D. Morris, A. F. Mulone, D. Munoz, T. Muraveva, C. P. Murphy, I. Musella, L. Noval, C. Ordénovic, G. Orrù, J. Osinde, C. Pagani, I. Pagano, L. Palaversa, P. A. Palicio, A. Panahi, M. Pawlak, X. Peñalosa Esteller, A. Penttilä, A. M. Piersimoni, F. X. Pineau, E. Plachy, G. Plum, E. Poggio, E. Poretti, E. Poujoulet, A. Prša, L. Pulone, E. Racero, S. Ragaini, M. Rainer, C. M. Raiteri, N. Rambaux, P. Ramos, M. Ramos-Lerate, P. Re Fiorentin, S. Regibo, C. Reylé, V. Ripepi, A. Riva, G. Rixon, N. Robichon, C. Robin, M. Roelens, L. Rohrbasser, M. Romero-Gómez, N. Rowell, F. Royer, K. A. Rybicki, G. Sadowski, A. Sagristà Sellés, J. Sahlmann, J. Salgado, E. Salguero, N. Samaras, V. Sanchez Gimenez, N. Sanna, R. Santoveña, M. Sarasso, M. Schultheis, E. Sciacca, M. Segol, J. C. Segovia, D. Ségransan, D. Semeux, S. Shahaf, H. I. Siddiqui, A. Siebert, L. Siltala, E. Slezak, R. L. Smart, E. Solano, F. Solitro, D. Souami, J. Souchay, A. Spagna, F. Spoto, I. A. Steele, H. Steidelmüller, C. A. Stephenson, M. Süveges, L. Szabados, E. Szegedi-Elek, F. Taris, G. Tauran, M. B. Taylor, R. Teixeira, W. Thuillot, N. Tonello, F. Torra, J. Torra, C. Turon, N. Unger, M. Vaillant, E. van Dillen, O. Vanel, A. Vecchiato, Y. Viala, D. Vicente, S. Voutsinas, M. Weiler, T. Wevers, Ł. Wyrzykowski, A. Yoldas, P. Yvard, H. Zhao, J. Zorec, S. Zucker, C. Zurbach, and T. Zwitter. Gaia Early Data Release 3. Summary of the contents and survey properties. , 649:A1, May 2021.

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

[48] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 4(5):6, 2014.

[49] K. M. Górski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *Astrophys. J.*, 622:759–771, April 2005.

[50] 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.

[51] 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.

[52] 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.

[53] C. J. Grillmair. Detection of a 60°-long Dwarf Galaxy Debris Stream. , 645(1):L37–L40, July 2006.

[54] C. J. Grillmair and O. Dionatos. Detection of a 63° Cold Stellar Stream in the Sloan Digital Sky Survey. *Astrophys. J.*, 643(1):L17–L20, May 2006.

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

[56] 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.

[57] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970.

[58] Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2020.

[59] Amina Helmi. Streams, Substructures, and the Early History of the Milky Way. , 58:205–256, August 2020.

[60] Amina Helmi, Carine Babusiaux, Helmer H. Koppelman, Davide Massari, Jovan Veljanoski, and Anthony G. A. Brown. The merger that led to the formation of the Milky Way's inner stellar halo and thick disk. *Nature*, 563(7729):85–88, October 2018.

[61] Amina Helmi and Simon D. M. White. Building up the stellar halo of the Galaxy. *MNRAS*, 307(3):495–517, August 1999.

[62] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free MCMC with Amortized Approximate Ratio Estimators. *arXiv e-prints*, page arXiv:1903.04057, March 2019.

[63] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, and Gilles Louppe. Averting A Crisis In Simulation-Based Inference. *arXiv e-prints*, page arXiv:2110.06581, October 2021.

[64] 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.

[65] Matteo Hessel, David Budden, Fabio Viola, Mihaela Rosca, Eren Sezener, and Tom Hennigan. Optax: composable gradient transformation and optimisation, in jax!, 2020.

[66] Jacob Hollingsworth, Michael Ratz, Philip Tanedo, and Daniel Whiteson. Efficient sampling of constrained high-dimensional theoretical spaces with machine learning. *The European Physical Journal C*, 81(12), Dec 2021.

[67] P. V. C. Hough. Machine Analysis of Bubble Chamber Pictures. *Conf. Proc. C*, 590914:554–558, 1959.

[68] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows, 2018.

[69] Rodrigo Ibata, Geraint F. Lewis, Michael Irwin, Edward Totten, and Thomas Quinn. Great Circle Tidal Streams: Evidence for a Nearly Spherical Massive Dark Halo around the Milky Way. *Astrophys. J.*, 551(1):294–311, April 2001.

[70] Rodrigo Ibata, Khyati Malhan, Nicolas Martin, Dominique Aubert, Benoit Famaey, Paolo Bianchini, Giacomo Monari, Arnaud Siebert, Guillaume F. Thomas, Michele Bellazzini, Piercarlo Bonifacio, Elisabetta Caffau, and Florent Renaud. Charting the Galactic acceleration field I. A search for stellar streams with Gaia DR2 and EDR3 with follow-up from ESPaDOnS and UVES. *arXiv e-prints*, page arXiv:2012.05245, December 2020.

[71] Rodrigo A. Ibata, Khyati Malhan, and Nicolas F. Martin. The Streams of the Gaping Abyss: A Population of Entangled Stellar Streams Surrounding the Inner Galaxy. *Astrophys. J.*, 872(2):152, February 2019.

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

[73] Kathryn V. Johnston. A Prescription for Building the Milky Way's Halo from Disrupted Satellites. *Astrophys. J.*, 495(1):297–308, March 1998.

[74] Kathryn V. Johnston, Lars Hernquist, and Michael Bolte. Fossil Signatures of Ancient Accretion Events in the Halo. , 465:278, July 1996.

[75] Kathryn V. Johnston, HongSheng Zhao, David N. Spergel, and Lars Hernquist. Tidal Streams as Probes of the Galactic Potential. *Astrophys. J. Lett.*, 512(2):L109–L112, February 1999.

[76] 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.

[77] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[78] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, page arXiv:1412.6980, December 2014.

[79] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

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

[81] 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.

[82] 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.

[83] Sergey E. Koposov, Hans-Walter Rix, and David W. Hogg. Constraining the Milky Way Potential with a Six-Dimensional Phase-Space Map of the GD-1 Stellar Stream. , 712(1):260–273, March 2010.

[84] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[85] Michael Kuhlen, Mariangela Lisanti, and David N. Spergel. Direct Detection of Dark Matter Debris Flows. *Phys. Rev. D*, 86:063505, 2012.

[86] Andreas H. W. Küpper, Eduardo Balbinot, Ana Bonaca, Kathryn V. Johnston, David W. Hogg, Pavel Kroupa, and Basilio X. Santiago. Globular Cluster Streams as Galactic High-Precision Scales—the Poster Child Palomar 5. *Astrophys. J.*, 803(2):80, April 2015.

[87] L. Lindegren, J. Hernández, A. Bombrun, S. Klioner, U. Bastian, M. Ramos-Lerate, A. de Torres, H. Steidelmüller, C. Stephenson, D. Hobbs, U. Lammers, M. Biermann, R. Geyer, T. Hilger, D. Michalik, U. Stampa, P. J. McMillan, J. Castañeda, M. Clotet, G. Comoretto, M. Davidson, C. Fabricius, G. Gracia, N. C. Hambly, A. Hutton, A. Mora, J. Portell, F. van Leeuwen, U. Abbas, A. Abreu, M. Altmann, A. Andrei, E. Anglada, L. Balaguer-Núñez, C. Barache, U. Becciani, S. Bertone, L. Bianchi, S. Bouquillon, G. Bourda, T. Brüsemeister, B. Bucciarelli, D. Busonero, R. Buzzi, R. Cancelliere, T. Carlucci, P. Charlot, N. Cheek, M. Crosta, C. Crowley, J. de Bruijne, F. de Felice, R. Drimmel, P. Esquej, A. Fienga, E. Fraile, M. Gai, N. Garralda, J. J. González-Vidal, R. Guerra, M. Hauser, W. Hofmann, B. Holl, S. Jordan, M. G. Lattanzi, H. Lenhardt, S. Liao, E. Licata, T. Lister, W. Löffler, J. Marchant, J. M. Martin-Fleitas, R. Messineo, F. Mignard, R. Morbidelli, E. Poggio, A. Riva, N. Rowell, E. Salguero, M. Sarasso, E. Sciacca, H. Siddiqui, R. L. Smart, A. Spagna, I. Steele, F. Taris, J. Torra, A. van Elteren, W. van Reeven, and A. Vecchiato. Gaia Data Release 2. The astrometric solution. *Astron. & Astrophys.*, 616:A2, August 2018.

[88] Mariangela Lisanti and David N. Spergel. Dark Matter Debris Flows in the Milky Way. *Phys. Dark Univ.*, 1:155–161, 2012.

[89] Mariangela Lisanti and David N. Spergel. Dark matter debris flows in the Milky Way. *Physics of the Dark Universe*, 1(1-2):155–161, November 2012.

[90] Jan-Matthis Lueckmann, Jan Boelts, David S. Greenberg, Pedro J. Gonçalves, and Jakob H. Macke. Benchmarking Simulation-Based Inference. *arXiv e-prints*, page arXiv:2101.04653, January 2021.

[91] A. Li Luo, Yong-Heng Zhao, Gang Zhao, Li-Cai Deng, Xiao-Wei Liu, Yi-Peng Jing, Gang Wang, Hao-Tong Zhang, Jian-Rong Shi, Xiang-Qun Cui, Yao-Quan Chu, Guo-Ping Li, Zhong-Rui Bai, Yue Wu, Yan Cai, Shu-Yun Cao, Zi-Huang Cao, Jeffrey L. Carlin, Hai-Yuan Chen, Jian-Jun Chen, Kun-Xin Chen, Li Chen, Xue-Lei Chen, Xiao-Yan Chen, Ying Chen, Norbert Christlieb, Jia-Ru Chu, Chen-Zhou Cui, Yi-Qiao Dong, Bing Du, Dong-Wei Fan, Lei Feng, Jian-Ning Fu, Peng Gao, Xue-Fei Gong, Bo-Zhong Gu, Yan-Xin Guo, Zhan-Wen Han, Bo-Liang He, Jin-Liang Hou, Yong-Hui Hou, Wen Hou, Hong-Zhuan Hu, Ning-Sheng Hu, Zhong-Wen Hu, Zhi-Ying Huo, Lei Jia, Fang-Hua Jiang, Xiang Jiang, Zhi-Bo Jiang, Ge Jin, Xiao Kong, Xu Kong, Ya-Juan Lei, Ai-Hua Li, Chang-Hua Li, Guang-Wei Li, Hai-Ning Li, Jian Li, Qi Li, Shuang Li, Sha-Sha Li, Xin-Nan Li, Yan Li, Yin-Bi Li, Ye-Ping Li, Yuan Liang, Chien-Cheng Lin, Chao Liu, Gen-Rong Liu, Guan-Qun Liu, Zhi-Gang Liu, Wen-Zhi Lu, Yu Luo, Yin-Dun Mao, Heidi Newberg, Ji-Jun Ni, Zhao-Xiang Qi, Yong-Jun Qi, Shi-Yin Shen, Huo-Ming Shi, Jing Song, Yi-Han Song, Ding-Qiang Su, Hong-Jun Su, Zheng-Hong Tang, Qing-Sheng Tao, Yuan Tian, Dan Wang, Da-Qi Wang, Feng-Fei Wang, Guo-Min Wang, Hai Wang, Hong-Chi Wang, Jian Wang, Jia-Ning Wang,

Jian-Ling Wang, Jian-Ping Wang, Jun-Xian Wang, Lei Wang, Meng-Xin Wang, Shou-Guan Wang, Shu-Qing Wang, Xia Wang, Ya-Nan Wang, You Wang, Yue-Fei Wang, You-Fen Wang, Peng Wei, Ming-Zhi Wei, Hong Wu, Ke-Fei Wu, Xue-Bing Wu, Yu-Zhong Wu, Xiao-Zheng Xing, Ling-Zhe Xu, Xin-Qi Xu, Yan Xu, Tai-Sheng Yan, De-Hua Yang, Hai-Feng Yang, Hui-Qin Yang, Ming Yang, Zheng-Qiu Yao, Yong Yu, Hui Yuan, Hai-Bo Yuan, Hai-Long Yuan, Wei-Min Yuan, Chao Zhai, En-Peng Zhang, Hua-Wei Zhang, Jian-Nan Zhang, Li-Pin Zhang, Wei Zhang, Yong Zhang, Yan-Xia Zhang, Zheng-Chao Zhang, Ming Zhao, Fang Zhou, Xu Zhou, Jie Zhu, Yong-Tian Zhu, Si-Cheng Zou, and Fang Zuo. The first data release (DR1) of the LAMOST regular survey. *Research in Astronomy and Astrophysics*, 15(8):1095, August 2015.

[92] Khyati Malhan and Rodrigo A. Ibata. STREAMFINDER - I. A new algorithm for detecting stellar streams. *MNRAS*, 477(3):4063–4076, July 2018.

[93] Khyati Malhan and Rodrigo A. Ibata. Constraining the Milky Way halo potential with the GD-1 stellar stream. *MNRAS*, 486(3):2995–3005, July 2019.

[94] Khyati Malhan, Rodrigo A. Ibata, Raymond G. Carlberg, Michele Bellazzini, Benoit Famaey, and Nicolas F. Martin. Phase-space Correlation in Stellar Streams of the Milky Way Halo: The Clash of Kshir and GD-1. *Astrophys. J. Lett.*, 886(1):L7, November 2019.

[95] Khyati Malhan, Rodrigo A. Ibata, Bertrand Goldman, Nicolas F. Martin, Eugene Magnier, and Kenneth Chambers. STREAMFINDER II: A possible fanning structure parallel to the GD-1 stream in Pan-STARRS1. *MNRAS*, 478(3):3862–3870, August 2018.

[96] Khyati Malhan, Rodrigo A. Ibata, and Nicolas F. Martin. Ghostly tributaries to the Milky Way: charting the halo's stellar streams with the Gaia DR2 catalogue. *MNRAS*, 481(3):3442–3455, December 2018.

[97] Khyati Malhan, Monica Valluri, and Katherine Freese. Probing the nature of dark matter with accreted globular cluster streams. , 501(1):179–200, January 2021.

[98] Khyati Malhan, Zhen Yuan, Rodrigo Ibata, Anke Arentsen, Michele Bellazzini, and Nicolas F. Martin. Evidence of a dwarf galaxy stream populating the inner Milky Way Halo. *arXiv e-prints*, page arXiv:2104.09523, April 2021.

[99] Stefan Meingast and João Alves. Extended stellar systems in the solar neighborhood. I. The tidal tails of the Hyades. *Astron. & Astrophys.*, 621:L3, January 2019.

[100] Stefan Meingast, João Alves, and Verena Fürnkranz. Extended stellar systems in the solar neighborhood . II. Discovery of a nearby 120° stellar stream in Gaia DR2. *Astron. & Astrophys.*, 622:L13, February 2019.

[101] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of State Calculations by Fast Computing Machines. , 21(6):1087–1092, June 1953.

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

[103] 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.

[104] Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the Number of Linear Regions of Deep Neural Networks. *arXiv e-prints*, page arXiv:1402.1869, February 2014.

[105] 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.

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

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

[108] Benjamin Nachman and David Shih. Anomaly Detection with Density Estimation. *Phys. Rev. D*, 101:075042, 2020.

[109] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do Deep Generative Models Know What They Don't Know? *arXiv e-prints*, page arXiv:1810.09136, October 2018.

[110] Radford Neal. MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. 2011.

[111] Lina Necib, Mariangela Lisanti, and Vasily Belokurov. Inferred Evidence for Dark Matter Kinematic Substructure with SDSS-Gaia. , 874(1):3, March 2019.

[112] Lina Necib, Mariangela Lisanti, Shea Garrison-Kimmel, Andrew Wetzel, Robyn Sanderson, Philip F. Hopkins, Claude-André Faucher-Giguère, and Dušan Kereš. Under the FIRElight: Stellar Tracers of the Local Dark Matter Velocity Distribution in the Milky Way. , 883(1):27, September 2019.

[113] Heidi Jo Newberg. Lessons from Tidal Debris in the Halo of the Milky Way. In *AAS/Division of Dynamical Astronomy Meeting #41*, volume 41 of *AAS/Division of Dynamical Astronomy Meeting*, page 5.01, May 2010.

[114] Heidi Jo Newberg, Brian Yanny, Connie Rockosi, Eva K. Grebel, Hans-Walter Rix, Jon Brinkmann, Istvan Csabai, Greg Hennessy, Robert B. Hindsley, Rodrigo Ibata, Zeljko Ivezić, Don Lamb, E. Thomas Nash, Michael Odenkirchen, Heather A. Rave, D. P. Schneider, J. Allyn Smith, Andrea Stolte, and Donald G. York. The Ghost of Sagittarius and Lumps in the Halo of the Milky Way. , 569(1):245–274, April 2002.

[115] J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London Series A*, 231:289–337, January 1933.

[116] Michael Odenkirchen, Eva K. Grebel, Constance M. Rockosi, Walter Dehnen, Rodrigo Ibata, Hans-Walter Rix, Andrea Stolte, Christian Wolf, Jr. Anderson, John E., Neta A. Bahcall, Jon Brinkmann, István Csabai, G. Hennessy, Robert B. Hindsley, Željko Ivezić, Robert H. Lupton, Jeffrey A. Munn, Jeffrey R. Pier, Chris Stoughton, and Donald G. York. Detection of Massive Tidal Tails around the Globular Cluster Palomar 5 with Sloan Digital Sky Survey Commissioning Data. , 548(2):L165–L169, February 2001.

[117] 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.

[118] George Papamakarios and Iain Murray. Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation. *arXiv e-prints*, page arXiv:1605.06376, May 2016.

[119] 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.

[120] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference, 2019.

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

[122] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation, 2018.

[123] 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.

[124] 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.

[125] 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.

[126] 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.

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

[128] 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.

[129] Sarah Pearson, Susan E. Clark, Alexis J. Demirjian, Kathryn V. Johnston, Melissa K. Ness, Tjitske K. Starkenburg, Benjamin F. Williams, and Rodrigo A. Ibata. The hough stream spotter: A new method for detecting linear structure in resolved stars and application to the stellar halo of m31, 2021.

[130] Sarah Pearson, Tjitske K. Starkenburg, Kathryn V. Johnston, Benjamin F. Williams, Rodrigo A. Ibata, and Rubab Khan. Detecting Thin Stellar Streams in External Galaxies: Resolved Stars and Integrated Light. , 883(1):87, September 2019.

[131] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

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

[133] Planck Collaboration, N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak,

R. Battye, K. Benabed, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, J. Carron, A. Challinor, H. C. Chiang, J. Chluba, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, J. M. Delouis, E. Di Valentino, J. M. Diego, O. Doré, M. Douspis, A. Ducout, X. Dupac, S. Dusini, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, Y. Fantaye, M. Farhang, J. Fergusson, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, F. K. Hansen, D. Herranz, S. R. Hildebrandt, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J. M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, P. Lemos, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. López-Caniego, P. M. Lubin, Y. Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Matarrese, N. Mauri, J. D. McEwen, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. A. Miville-Deschênes, D. Molinari, L. Montier, G. Morgante, A. Moss, P. Natoli, H. U. Nørgaard-Nielsen, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J. L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, C. Sirignano, G. Sirri, L. D. Spencer, R. Sunyaev, A. S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, L. Valenziano, J. Valiviita, B. Van Tent, L. Vibert, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacchei, and A. Zonca. Planck 2018 results. VI. Cosmological parameters. , 641:A6, September 2020.

[134] Adrian M. Price-Whelan and Ana Bonaca. Gaia data, Pan-STARRS photometry, and stream selection masks for the region around the GD-1 stream, June 2018.

[135] Adrian M. Price-Whelan and Ana Bonaca. Off the Beaten Path: Gaia Reveals GD-1 Stars outside of the Main Stream. , 863(2):L20, August 2018.

[136] Chris W. Purcell, Andrew R. Zentner, and Mei-Yu Wang. Dark matter direct search rates in simulations of the Milky Way and Sagittarius stream. , 2012(8):027, August 2012.

[137] Stella Reino, Elena M. Rossi, Robyn E. Sanderson, Elena Sellentin, Amina Helmi, Helmer H. Koppelman, and Sanjib Sharma. Galactic potential constraints from

clustering in action space of combined stellar stream data. *arXiv e-prints*, page arXiv:2007.00356, July 2020.

[138] 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.

[139] Danilo Jimenez Rezende, George Papamakarios, Sebastien Racaniere, Michael S. Albergo, Gurtej Kanwar, Phiala E. Shanahan, and Kyle Cranmer. Normalizing flows on tori and spheres, 2020.

[140] Brant Robertson, James S. Bullock, Andreea S. Font, Kathryn V. Johnston, and Lars Hernquist. Λ Cold Dark Matter, Stellar Feedback, and the Galactic Halo Abundance Pattern. , 632(2):872–881, October 2005.

[141] Jason L. Sanders and James Binney. Stream-orbit misalignment - I. The dangers of orbit-fitting. *MNRAS*, 433(3):1813–1825, August 2013.

[142] Jason L. Sanders, Jo Bovy, and Denis Erkal. Dynamics of stream-subhalo interactions. *MNRAS*, 457(4):3817–3835, April 2016.

[143] David Shih, Matthew R. Buckley, Lina Necib, and John Tamanas. Finding streams in gaia dr2 using via machinae (in preparation), 2021.

[144] N. Shipp, A. Drlica-Wagner, E. Balbinot, P. Ferguson, D. Erkal, T. S. Li, K. Bechtol, V. Belokurov, B. Buncher, D. Carollo, M. Carrasco Kind, K. Kuehn, J. L. Marshall, A. B. Pace, E. S. Rykoff, I. Sevilla-Noarbe, E. Sheldon, L. Strigari, A. K. Vivas, B. Yanny, A. Zenteno, T. M. C. Abbott, F. B. Abdalla, S. Allam, S. Avila, E. Bertin, D. Brooks, D. L. Burke, J. Carretero, F. J. Castander, R. Cawthon, M. Crocce, C. E. Cunha, C. B. D'Andrea, L. N. da Costa, C. Davis, J. De Vicente, S. Desai, H. T. Diehl, P. Doel, A. E. Evrard, B. Flaugher, P. Fosalba, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. Hartley, K. Honscheid, B. Hoyle, D. J. James, M. D. Johnson, E. Krause, N. Kuropatkin, O. Lahav, H. Lin, M. A. G. Maia, M. March, P. Martini, F. Menanteau, C. J. Miller, R. Miquel, R. C. Nichol, A. A. Plazas, A. K. Romer, M. Sako, E. Sanchez, B. Santiago, V. Scarpine, R. Schindler, M. Schubnell, M. Smith, R. C. Smith, F. Sobreira, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, D. L. Tucker, A. R. Walker, R. H. Wechsler, and DES Collaboration. Stellar Streams Discovered in the Dark Energy Survey. , 862(2):114, August 2018.

[145] 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.

[146] 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.

[147] John Tamanas. LBI: Likelihood-based Inference with JAX, 3 2022. Available at `https://github.com/jtamanas/LBI`, version 1.0.0.

[148] 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.

[149] Dustin Tran, Keyon Vafa, Kumar Agrawal, Laurent Dinh, and Ben Poole. Discrete flows: Invertible generative models of discrete data. In *Advances in Neural Information Processing Systems*, pages 14692–14701, 2019.

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

[151] 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.

[152] A. Varghese, R. Ibata, and G. F. Lewis. Stellar streams as probes of dark halo mass and morphology: a Bayesian reconstruction. *MNRAS*, 417(1):198–215, October 2011.

[153] 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.

[154] Donald G. York, J. Adelman, Jr. Anderson, John E., Scott F. Anderson, James Annis, Neta A. Bahcall, J. A. Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, William N. Boroski, Steve Bracker, Charlie Briegel, John W. Briggs, J. Brinkmann, Robert Brunner, Scott Burles, Larry Carey, Michael A. Carr, Francisco J. Castander, Bing Chen, Patrick L. Colestock, A. J. Connolly, J. H. Crocker, István Csabai, Paul C. Czarapata, John Eric Davis, Mamoru Doi, Tom Dombeck, Daniel Eisenstein, Nancy Ellman, Brian R. Elms, Michael L. Evans, Xiaohui Fan, Glenn R. Federwitz, Larry Fiscelli, Scott Friedman, Joshua A. Frieman, Masataka Fukugita, Bruce Gillespie, James E. Gunn, Vijay K. Gurbani, Ernst de Haas, Merle Haldeman, Frederick H. Harris, J. Hayes, Timothy M. Heckman, G. S. Hennessy, Robert B. Hindsley, Scott Holm, Donald J. Holmgren, Chi-hao Huang, Charles Hull, Don Husby, Shin-Ichi Ichikawa, Takashi Ichikawa, Željko Ivezić, Stephen Kent, Rita S. J. Kim, E. Kinney, Mark Klaene, A. N. Kleinman,

S. Kleinman, G. R. Knapp, John Korienek, Richard G. Kron, Peter Z. Kunszt, D. Q. Lamb, B. Lee, R. French Leger, Siriluk Limmongkol, Carl Lindenmeyer, Daniel C. Long, Craig Loomis, Jon Loveday, Rich Lucinio, Robert H. Lupton, Bryan MacKinnon, Edward J. Mannery, P. M. Mantsch, Bruce Margon, Peregrine McGehee, Timothy A. McKay, Avery Meiksin, Aronne Merelli, David G. Monet, Jeffrey A. Munn, Vijay K. Narayanan, Thomas Nash, Eric Neilsen, Rich Neswold, Heidi Jo Newberg, R. C. Nichol, Tom Nicinski, Mario Nonino, Norio Okada, Sadanori Okamura, Jeremiah P. Ostriker, Russell Owen, A. George Pauls, John Peoples, R. L. Peterson, Donald Petravick, Jeffrey R. Pier, Adrian Pope, Ruth Pordes, Angela Prosapio, Ron Rechenmacher, Thomas R. Quinn, Gordon T. Richards, Michael W. Richmond, Claudio H. Rivetta, Constance M. Rockosi, Kurt Ruthmansdorfer, Dale Sandford, David J. Schlegel, Donald P. Schneider, Maki Sekiguchi, Gary Sergey, Kazuhiro Shimasaku, Walter A. Siegmund, Stephen Smee, J. Allyn Smith, S. Snedden, R. Stone, Chris Stoughton, Michael A. Strauss, Christopher Stubbs, Mark SubbaRao, Alexander S. Szalay, Istvan Szapudi, Gyula P. Szokoly, Anirudda R. Thakar, Christy Tremonti, Douglas L. Tucker, Alan Uomoto, Dan Vanden Berk, Michael S. Vogeley, Patrick Waddell, Shu-i. Wang, Masaru Watanabe, David H. Weinberg, Brian Yanny, Naoki Yasuda, and SDSS Collaboration. The Sloan Digital Sky Survey: Technical Summary. , 120(3):1579–1587, September 2000.

[155] Zhen Yuan, Jiang Chang, Projjwal Banerjee, Jiaxin Han, Xi Kang, and M. C. Smith. StarGO: A New Method to Identify the Galactic Origins of Halo Stars. , 863(1):26, August 2018.

[156] Andrea Zonca, Leo Singer, Daniel Lenz, Martin Reinecke, Cyrille Rosset, Eric Hivon, and Krzysztof Gorski. healpy: equal area pixelization and spherical harmonics transforms for data on the sphere in python. *Journal of Open Source Software*, 4(35):1298, March 2019.