UNIVERSITY OF CALIFORNIA
RIVERSIDE

Studying Quasar Spectra with Machine Learning in Sloan Digital Sky Survey

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

Doctor of Philosophy

in

Physics

by

Reza Monadi

September 2023

Dissertation Committee:

Dr. Simeon Bird, Chairperson
Dr. Gabriela Canalizo
Dr. Bahram Mobasher

The Dissertation of Reza Monadi is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my incredible wife, Asieh. Her selflessness and determination in putting her own dreams on hold to support my academic journey have been truly remarkable. Without her unwavering encouragement and motivation, I would not have been able to overcome the challenges along the way.

I would like to extend my heartfelt appreciation to my advisor, Professor Simeon Bird, who undoubtedly deserves recognition as "The Most Optimistic and Friendly Graduate Adviser." As a highly intelligent astrophysicist, Simeon never fails to make his students feel valued and empowered. Moreover, I am grateful for his patience in allowing me to dedicate significant time to my job search during my final year.

I am immensely grateful to my committee members, Professor Bahram Mobasher and Professor Gabriela Canalizo. From Bahram, I have learned to think big and pursue answering significant scientific questions. Gaby has taught me the importance of humbleness and being a compassionate individual while conducting cutting-edge research.

I consider myself extremely fortunate to have been a part of an exceptional research group. Martin Fernandez has been more than just a colleague; he has been like a brother and a trusted advisor. Ming-Feng Ho's expertise in coding, statistics, and even making a perfect cup of coffee has greatly facilitated my academic journey. Bryan Scott has made sure that I feel UCR is my second home, going above and beyond to provide support. I must acknowledge the support and patience of our postdoc, Phoebe Sanderbeck. Also, I want to thank Mahdi's for his KOefforts in organizing our group meetings.

I would also like to express my gratitude to Jessica Doppel for her role in organizing astro coffee and bringing the astro community together in Pierce Hall. Jessica's dedication in reminding everyone about lunch and coffee break times has truly been the glue that holds us all together.

Lastly, I owe a profound debt of gratitude to my parents, Marzieh and Ebrahim, for nurturing my curiosity and enrolling me in a summer astronomy class at the Thaqib Astronomical Association when I was a child. The experiences and opportunities provided by Thaqib have forever shaped my understanding of the importance of science outreach and have positively impacted my life in countless ways.

ABSTRACT OF THE DISSERTATION

Studying Quasar Spectra with Machine Learning in Sloan Digital Sky Survey

by

Reza Monadi

Doctor of Philosophy, Graduate Program in Physics
University of California, Riverside, September 2023
Dr. Simeon Bird, Chairperson

In this thesis, we designed an algorithm to provide robust selection criteria in the parameter space of measured properties of quasars. Our method combines the prior knowledge of an expert observer with what unsupervised machine learning understands about the underlying structures in the data to get a data-driven boundary in the multi-dimensional parameter space of quasar physical properties. We did that by quantifying the dissimilarity of our target group to the majority of the quasars in our data set. Our versatile method can select a cluster of similar data points that are located in statistically significant lower-density regions of the parameter space. We could find more quasars in the class of *extremely red quasars* and show our new sample has even more exotic outflow behavior. Our final selection produces *three* times more quasars with visually verified CIV broad absorption line feature, which is the signature of outflow, than the previous *extremely red quasar* sample. Our method is very useful in selecting the most important follow-up targets for observing red quasars.

In the second project, we could assemble the largest CIV absorption line catalogue to date. By providing a probability for the existence of absorption systems in a quasar spectrum that

is a by-product of our Bayesian model selection and Gaussian Processes methods, we removed the need for visual inspection which is essential in dealing with the upcoming surveys with millions of spectra. After carefully validating our method by comparing a subset of the spectra inspected in the largest visually inspected CIV catalog to what our method predicts, we could find 113,775 CIV absorption systems with at least 95% confidence among 185,425 selected quasar spectra from SDSS DR12. We obtain a posterior distribution for column density, velocity dispersion, and absorption redshift for each investigated spectrum which can be used to get the maximum a posteriori value and the credible interval. Our method is specifically useful when we want to obtain information from low signal-to-noise ratio data.

Contents

List of Figures

Chapter 1

Introduction

We have explored different scales of the Universe during the history of science according to the limits of technology. These limits are set by how strong is our imagination and how big are our ambitions. That is why once our Universe was the Earth, then became our solar system, then we found ourselves on the island of our Milky Way. But now we talk about a supercluster of galaxies and who knows about the future.

Much of the advances in our understanding of the Universe come from our efforts in *Extragalactic Astronomy*. By observing galaxies outside our local group galaxy, we have access to different snapshots of the evolving Universe from when it was younger than now. Through this, we will learn about the formation, history, and evolution of galaxies, which are the building blocks of the Universe. We can learn about galaxy evolution by looking at the space between the galaxies which is not simply void. When these intergalactic spaces have a chance to get illuminated by cosmic flashlights, we realize how rich they are.

1.1 Active Galaxies

Active galaxies are a group of galaxies with enormous emission from their center. Active galaxies can be distinguished by their spectral energy distribution which can not be matched to the cumulative lights of the stars within the galaxy or the thermal black body radiation of a single star (see Figure 1.1). Each feature in these spectral energy distributions has a distinct source. The big blue bump can be related to the accretion disk component of the active galaxy while the infrared bump is more likely coming from the warm dust grain in temperatures above 2000 K. The IR bump is coming from the thermal emission from dust at a wide range of temperatures, *sim* 50 -1000 K. The radio part of radiation in Figure 1.1 is resulted from the interaction of jets with the gas around galaxy.

More detailed observation of active galaxies demonstrated that there is a variety of them and similar to other branches of science, astronomers began to classify them. Seyfert galaxies, the first discovered active galaxies, turned out to show two different classes of emission line properties:

- **Seyfert 1**: They have both narrow and broad allowed lines plus narrow forbidden lines.
- **Seyfert 2**: They have narrow lines which can be forbidden or allowed one.

Moreover, some of these active galaxies were observed on the radio while others were quiet on radio wavelengths. Therefore, we can naturally have radio-loud and radio-quiet quasars. Blazars, being even more exotic, were classified firstly as variable stars since their spectra showed no emission lines but only continuous synchrotron emission with large-amplitude flux variability. By pursuing more precise observations of Seyfert galaxies, astronomers discovered that some of them can have properties of both Seyfert 1 and Seyfert 2 galaxies or even some of them can change their emission

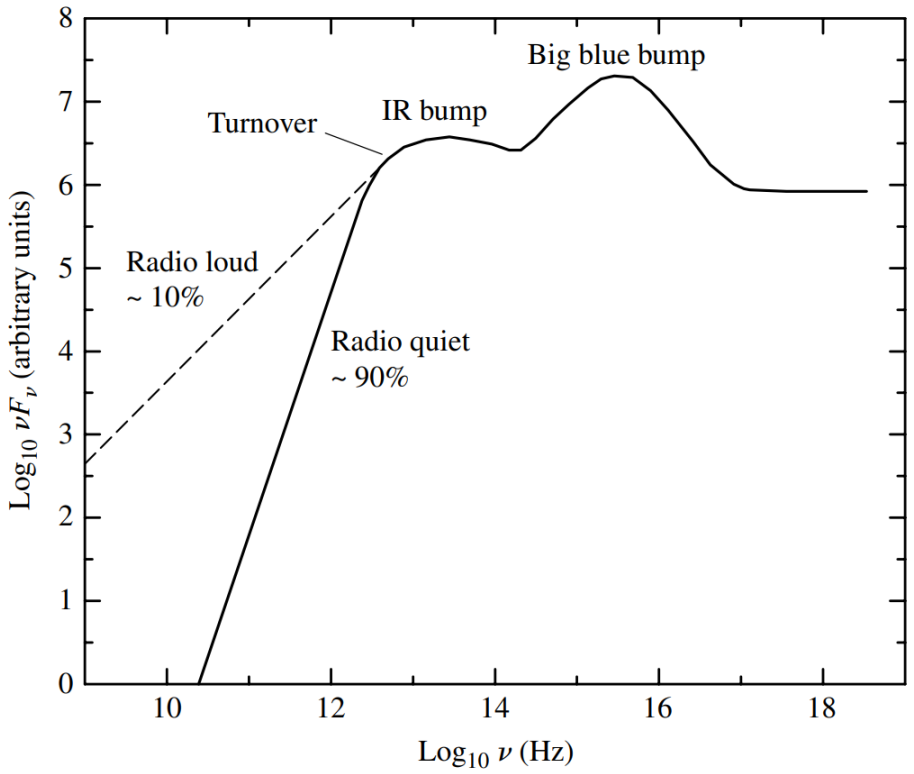

Figure 1.1: Spectral energy distribution of a radio galaxy which is one of the members of the broader group of active galactic nuclei. (Carroll & Ostlie 2017)

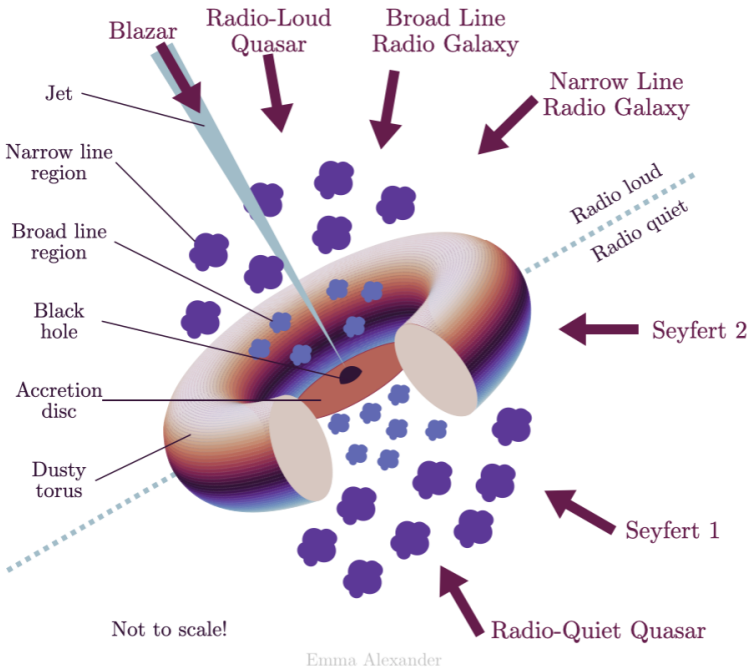

Figure 1.2: Unified model of Active Galactic Nuclei: Different classes of quasars arise from the viewing angle of the observer. (credit: Emma Alexander)

properties over time. The Unified Model of Active Galactic Nuclei (Antonucci 1993) proposed a solution for this puzzle: All of these variations arise from the viewing angle of the observer while the nature of the active galactic nuclei is the same (see Figure 1.2).

As is shown in Figure 1.2, an active galactic nucleus is composed of the following parts:

- **Black Hole:** In the center, there is a super-massive black hole that provides gravitational pull and works like a powerhouse. It is believed that a chain collapse of a cluster of compact stars (Kroupa et al. 2020) or the direct collapse of self-interacting dark matter (Feng et al. 2021)

- **Accretion Disk:** As the black hole pulls the dust and gas inward, they fall in a circular trajectory and create an accretion disk around the black hole. The disk shape is the result of

the combination of the tidal force and the centripetal force on the accreting material. Most of the radiation comes from the accretion disk of the active galaxies

- **Dusty Torus:** It is evident, though it is not possible to directly observe in all active galaxies, that there should be round and the doughnut-shape structure composed of dust around the accretion disk, otherwise we would have been able to directly observe all of the X-ray and UV light emitting from the accretion disk. When the dust particles in the torus absorb the higher energy lights and emit infrared before they get vaporized.

- **Broad Line Region:** There is a clumpy region of gas closer to the black hole, inside the torus, that broad permitted lines (with full width half maxima ~5000 km s^{-1}) can form there. Broad emission lines can respond to the variation of the continuum light very fast (within a month) which is a piece of evidence that they are close to the central engine. Given the high density of this region ($\sim 10^{15}$ m^{-3}) the collisional deexcitation does not give a chance for forbidden lines to form easily.

- **Narrow Line Region:** This region, which contains more mass than the broad line region, but with lower densities ($\sim 10^{10} m^{-3}$), is the origin of narrower lines with full-width half-maxima of ~500 km s^{-1}. As the narrow lines do not usually respond to the variation of the continuum, they should be well far from the central engine. That is why the narrow line region is outside of the dusty torus in Figure 1.2.

1.1.1 Quasars

Quasars are extremely powerful and energetic. But they are roughly the same size as our solar system. They are $\sim 10^5$ more luminous than our own Milky Way. That is why they appear as

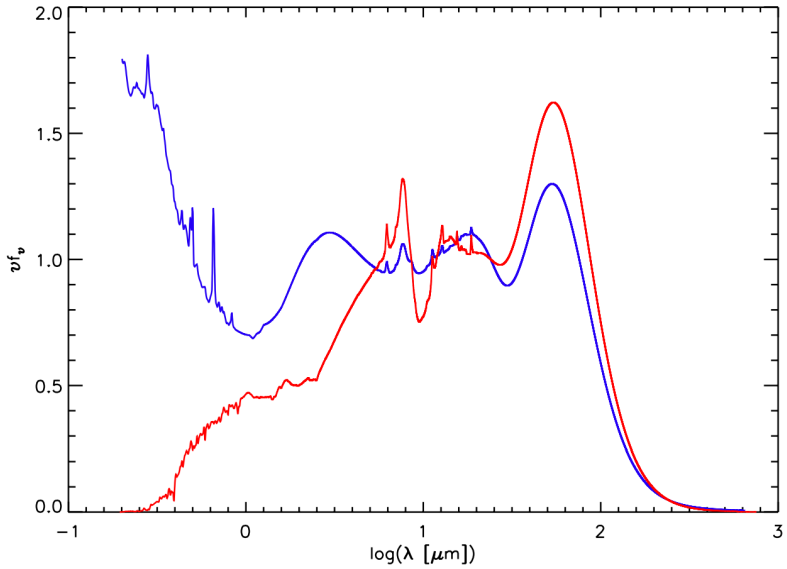

Figure 1.3: The average spectral energy distributions of a sample of six type 1 (blue curve) and six type 2 (red curve) quasars which were selected in the mid-infrared and is modelled at optical to far-infrared wavelengths (Hiner et al. 2009).

bright stars in optical images considering their huge distances from us. Indeed, quasars are the most energetic members of the family of active galaxies. When radio observations became more common, astronomers became interested in the optical counterparts of the observed radio sources. They found optical sources that had emission lines at redshifted wavelengths. They discovered *Quasi-stellar radio sources* or *quasars*.[1]

There are different types of quasars which are usually classified based on their emission line properties (type 1 vs. type 2 similar to Seyfert 1 and Seyfert 2) and their spectral energy distributions (SED) which will be reflected in the various defined colors for a quasar (red quasars vs. blue quasars). Figure 1.3 compares the SED a sample of type 1 and type 2 quasars. The overall

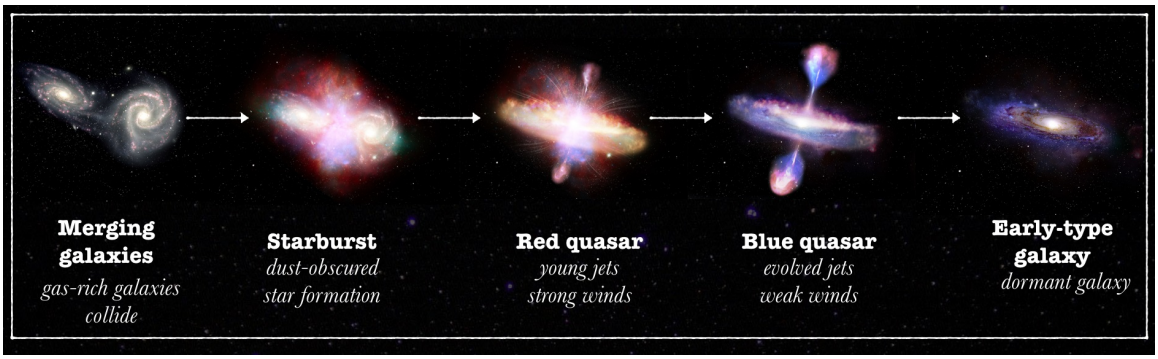

Figure 1.4: An artist's impression of the transitional nature of red quasars in a sequence of events triggered by merging two large and gas-rich galaxies. Credit: Gemini Observatory, GMOS-South, NSF.

shape of the SED is different and for example, there is a bump at 3μm for the average SED of type 1 QSOs which is missing in the SED of type 2 ones. (Hiner et al. 2009)

As Figure 1.4 shows, if two big and gas-rich galaxy merge, their supermassive black holes get combined and we may have a burst of star formation resulting from the tidal force and shocks throughout the process of merging. [2] The provided dust from star formation fuels the accretion disk around the bigger black hole which results in a very luminous and powerfully wind-generating red quasar. Eventually, the winds clear out the dust, quench the star formation, and red quasar into a calmer blue one. Afterwards, when the transitionary quasar phase ends, we will observe an early-type galaxy.

In an attempt to classify quasars considering more observational properties. It has been found that some of the anti-correlation between FeII and [OIII] optical emission is more effective in recognition of a main sequence of quasars, similar to the main sequence of stars (Panda et al. 2017; Marziani et al. 2018). This attempt is somehow in tandem with the evolutionary and transitionary

[1] This naming can be misleading since radio-quite quasars have been observed!

[2] Although a statistical analysis on a large sample of galaxies does not support this idea. (Pearson et al. 2019)

nature of quasars, but at the same time is not as physically interpretable as the traditional distinct classification of quasars. As a result, we need to put some effort into the task of classification of quasars in a way that is both interpretable and is conclusive to as many measured properties of quasars as possible.

1.2 Circumgalactic and Intergalactic

The circumgalactic medium is composed of dust and gas inside the virial radius of a galaxy but outside of the main body of it. For example, the circumgalactic region of our own Milky Way is anywhere 3 kpc up to 200 kpc from the center. The circumgalactic medium of a large galaxy is a mixture of gas composed of different flows of materials. First, there is an outward flow of gas: Supernovae winds and powerful outflows originating from the central active nucleus of the galaxy can eject gas and dust to the outer regions of a galaxy. Second, there is an inflow of gas: The central part of the large galaxy ont only gravitates the outflow mentioned above, it also pulls back the gas stripped from its satellites, and the gas from the intergalactic medium. Figure 1.5 shows these different flows. Moreover, the ejected gas can be contributed in the further star formation and get recycled to the hosting galaxy.

The mixture of different flows in the circumgalactic medium can affect its metalicity. Some galaxies, like starburst ones, show more metalicity (abundance of elements heavier than Helium) since the stellar wind, containing heavier elements, are stronger. However, the inflow of gas from filaments of the intergalactic medium can dilute the metallicity of the circumgalactic gas.

Only 15% of the content of baryons can be found in gravitationally bound systems. The rest of it lies in the intergalactic medium. The intergalactic medium is composed of a collisionally

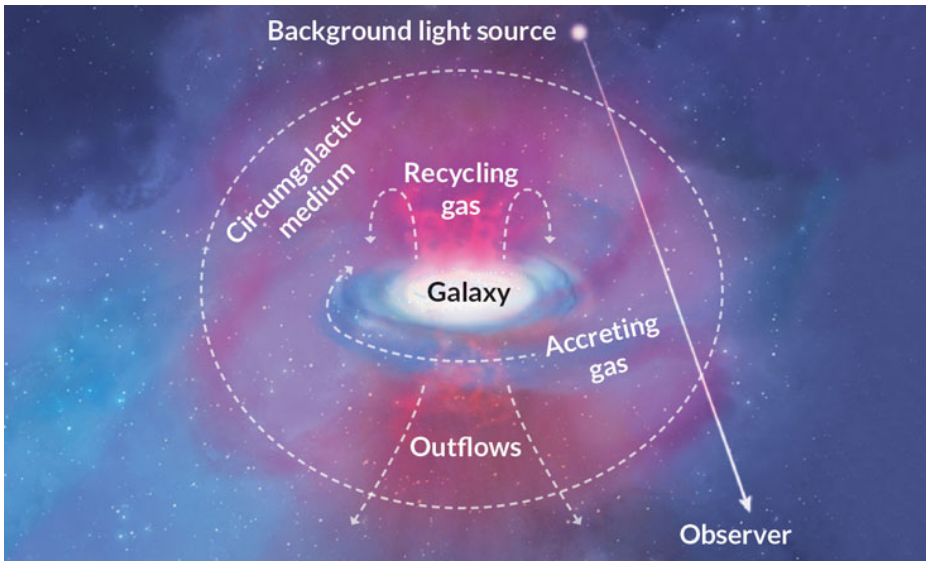

Figure 1.5: An artist's impression of the Circumgalactic Medium which is influenced by outflow and inflow of gas. We can observe the properties of the circumgalactic medium by its imprints in the light coming from the background light sources such as bright quasars. Credit: C. Chang

excited phase which is warm-hot and also includes a cooler diffuse part which can be mostly photo-ionized. It is possible that the intergalactic medium gets enriched by metals and we observe metal absorption lines like CIV in the intergalactic HI gas when it is optically thick (Aracil et al. 2004).

1.3 Quasar absorption lines

The observed flux of a quasar is not just telling us about the emitting source but also about the ionization, chemical, thermal, and structural conditions of the intervening absorbing gas in the line of sight. All of this information come from modelling and measuring the optical depth by comparing the observed flux to the continuum flux which is the flux when we do not expect any absorption:

$$F(\lambda) = F_0(\lambda)e^{-\tau_\lambda}, \tag{1.1}$$

where F is the observed flux from an original flux, F_0, after absorption in a medium with an optical depth of τ_λ. Practically we can think of a *redistribution function* that maps the optical depth τ_λ in Equation 1.1 to different physical properties of the absorbing environments such as density and temperature.

Because the quasar light can interact with anything in its path toward us, when working with the density of the absorption region, we should consider column density which is defined as the density in a cylinder as long as the path length of light with a cross sectional area of 1 cm^{-2}. We can think of the column density with the help of the average number of interactions that may occur between the light and matter along the light-path of L, which is the definition of optical depth τ_λ, in a cylindrical volume with cross-section of $\sigma(\lambda)$ and a number density of $n(s)$:

$$\tau_\lambda = \int_0^L n(s)\sigma(\lambda)ds = N\sigma(\lambda) \tag{1.2}$$

Therefore, a larger density absorbing region will be reflected in a deeper absorption line profile because denser regions have larger optical depths (larger average number of collisions along the path length).

We can also deduce information about the velocity of absorbing gas by measuring the optical depth in Equation 1.1. Different mechanisms can move the absorbing atoms and molecules in the path of light and therefore the observed absorbed wavelength will be shifted to larger and shorter which causes the observed absorption profile to be broadened instead of a sharp delta-function-like feature centered at the transition wavelength. Different ways contribute to broadening the line profile. The absorbing gas particles can collide, especially in larger-density environments, where their velocity changes during the collision. 1548 Å CIV absorption line will be broadened by around 0.015 Å when the absorbing gas has a temperature of 10^6 K. We can also have turbulent

motion in the absorbing gas, caused by gravitational instability or the winds from star formation (Krumholz & Burkhart 2016), which can contribute to the broadening line profile, but usually it is less prominent than thermal broadening. Given the Heisenberg uncertainty principle and considering that the absorption happens in a finite amount of time, there is a natural broadening. For 1548 Å transition of CIV we will have a natural broadening of $\sim 2 \times 10^{-4}$ Å.

1.4 Sloan Digital Sky Survey

Sloan Sky Digital Sky Survey (SDSS) is one the most revolutionary, ambitious, and successful projects in the history of science that maps 1/4 of the full sky: the largest map in human history so that its total quantity of information will rival the materials in all the books of the Library of Congress! SDSS surveys hundreds of millions of celestial objects including more than a million galaxies and *quasars* and provides unique insights into the large-scale structure of the Universe.

SDSS provides photometric data from the observed targets in five different bands: Ultra-violet (3543 Å), Green (4770 Å), Red (6231 Å), Near Infrared (7625 Å), and Infrared (9134 Å). Comparing relative fluxed received from an object is a hint for classifying the type of the observed object and informs further spectroscopic observations.

SDSS revolutionized our understanding of Universe by validating the existing cosmological theories and recalling for new ones to explain the outcomes of accumulative outcomes of this large amount of astronomical data which is more than 600 TB (see Figure 1.6). Moreover, the big data provided SDSS pushed the boundaries of data analysis in the astronomical community by encouraging astronomers to use more modern statistical methods such as Machine Learning versus labor-intensive methods relying on visual inspection of a trained astronomer.

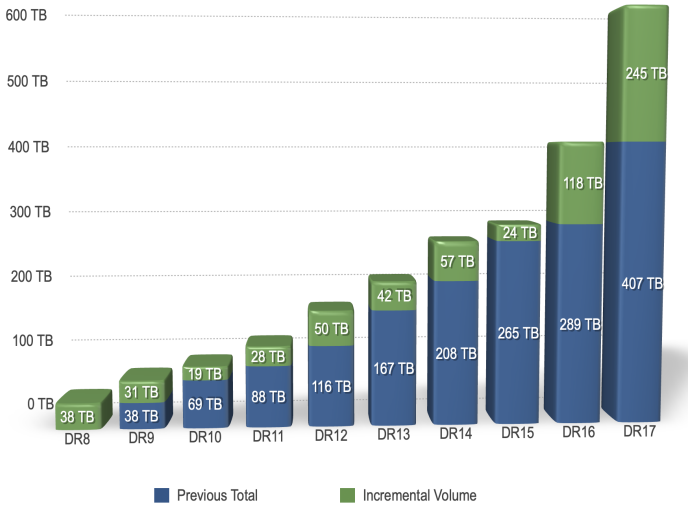

Figure 1.6: Incremental volume of data from SDSS data release 8 (DR8) to SDSS data release 17 (DR17). Credit: SDSS web page.

SDSS has different spectroscopic surveys, including Baryon Oscillation Spectroscopic Survey (BOSS) which mapped 1.5 million luminous red galaxies and 160,000 high redshift quasars. The spectra are produced via 1000 fiber spectrograph that each one correspond to an astronomical object. Then the spectrum is split into red and blue parts and finally are recorded in separate Charged-Coupled Devices (CCD).

1.5 Machine Learning

We have observed millions of galaxies through different surveys like SDSS and will observe billions of them with higher quality and resolutions through upcoming astronomical surveys like Dark Energy Survey Instrument (DESI) and Legacy Survey of Space and Time (LSST) of Vera Rubin Observatory. But why do we need more data? By observing billions of galaxies rather than a million, we might be able to probe the tail of different distributions with more sample points as

the tails are more sensitive to rare data. However, having more data demands automated methods with minimal intervention from humans. In the Galaxy Zoo project, 150, 000 people spent two years doing a very simple task to classify 900,000 galaxies, but we cannot do that with volunteers with this exponentially growing data volumes. Another example is Square Kilometer Array with one exabyte[3] of data per day that critically needs an automated method to reduce the raw data before storing it.

Machine learning helped astronomers in different tasks such as the detection and classification of supernovae via anomaly detection, gravitational lensing based on imaging data with convolutional neural networks, and predicting dark matter galaxy bias by decision tree algorithm (Moriwaki et al. 2023). Moreover, machine learning can assist cosmological simulations by lowering their computation time via *emulators* (Heitmann et al. 2006, 2009; Ho et al. 2023). In what follows, I explain two important categories of machine learning which I used throughout my projects.

1.5.1 Clustering

Clustering is an unsupervised machine learning method which the user lets the algorithm discover knowledge, here finding distinct clusters in the data, by its own with no training. Clustering algorithms group the most similar data points in the input data into *clusters*.

One of the most well-known clustering algorithms is K-Means. For a given number of clusters, the K-Means algorithm finds a set of centroids that minimizes the within-cluster sum of distances of data points:

$$\sum_{i=0}^{n} \min(||x_i - \mu_j||^2), \tag{1.3}$$

[3] If one gigabyte is the size of Earth, then an exabyte is the size of the sun!

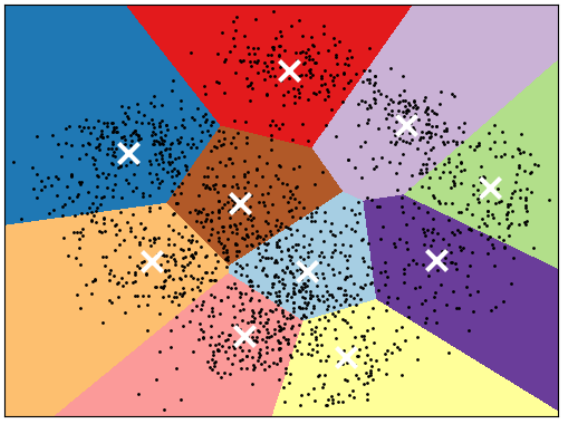

Figure 1.7: Kmeans clustering finds the centroids of the clusters within the data. Credit: scikit-learn

where μ_j is a candidate centroid in the $j-$th interaction and x_i is the $i-$th data point out of total n data points. The algorithm starts from some random seed points and finds the optimum centroids by varying the centroids to minimize the sum in Equation 1.3. Figure 1.7 shows an example of the application of this algorithm in a 2-dimensional parameter space. The shortcoming of K-Means is that it works when the distribution data is composed of distinct bubble-shaped clusters with comparable sizes. Moreover, via brute-force K-Means always outputs some clusters and the user should carefully select the number of clusters.

Another clustering algorithm, which does not need to be told how many clusters to look for, is Density-Based Spatial Clustering of Applications with Noise or DBSCAN for short. Given a minimum number of data points and a minimum radius for the cluster to form, DBSCAN scans all of the possibilities and finds data points in denser regions of the parameter space and assigns them to a cluster. If the data point is located where the density is significantly smaller compared to the density of the clusters, the data point cannot be attributed to any cluster and it will labelled as noise.

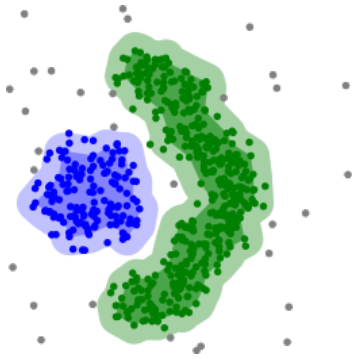

Figure 1.8: DBSCAN clustering finds some clusters within the data. Credit:ELKI

Figure 1.8 shows two clusters found by DBSCAN in green and blue. The grey points are labelled as noise because they cannot be assigned to any clusters.

Agglomerative clustering is a robust clustering method that automatically finds the underlying structure of the data in the parameter space based on the small-scale distance of points from their close by neighbors and the larger scale distance of groups of points from another group of data points. Agglomerative clustering starts with assigning a cluster to any single data point and then merges neighboring data points in each step. The user can stop this process by choosing a desired threshold. Figure 1.9 shows how this algorithm works. In step 1 all data points are one cluster. In step 2, because Point B and C are close together they will be clustered (same as Point D and E), but point A and F are farther away so are not agglomerated in this step. This clustering algorithm keeps agglomerating points in this fashion until all of the data points are attached or a certain threshold is met.

The Local Outlier Factor (LOF) algorithm is another efficient unsupervised anomaly detection method which computes the local density deviation of a given data point for its neighbors

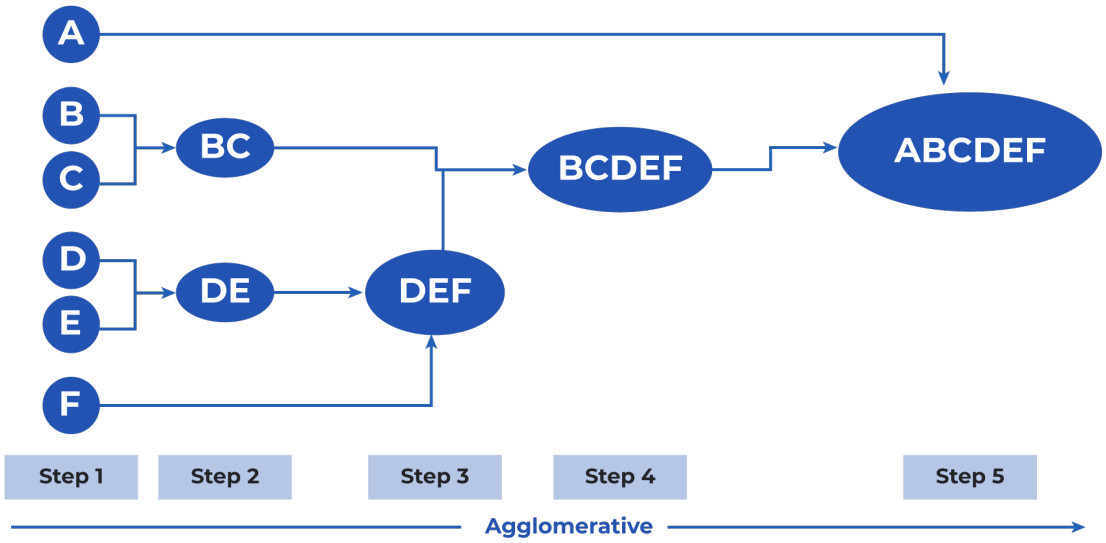

Figure 1.9: Agglomerative clustering finds clusters in each step based on the distance of clusters to each other. Credit:SearchUnify

to a broader region of parameter space where more data points are residing. The algorithm assigns a larger outlier score to the samples that have a lower density than their neighbors (See Section 2.4.3 for more detailed discussion). The user should decide how many data points should be considered as the number of nearest neighbors. Then we can find, for example, top 5% outlier members of the data set by sorting the outlier scores given by local outlier factor analysis. Figure 1.10 illustrates how the local outlier factor works. More outlier points get larger scores whereas data points deep inside a cluster of data will get a less outlier score.

1.5.2 Gaussian Processes

Gaussian processes (GP) can be used as a regression method and lies within the supervised machine learning category. In the usual regression tasks, we always should decide about the complexity of the fitting function. For example, a second-order polynomial vs. a cubic one, and

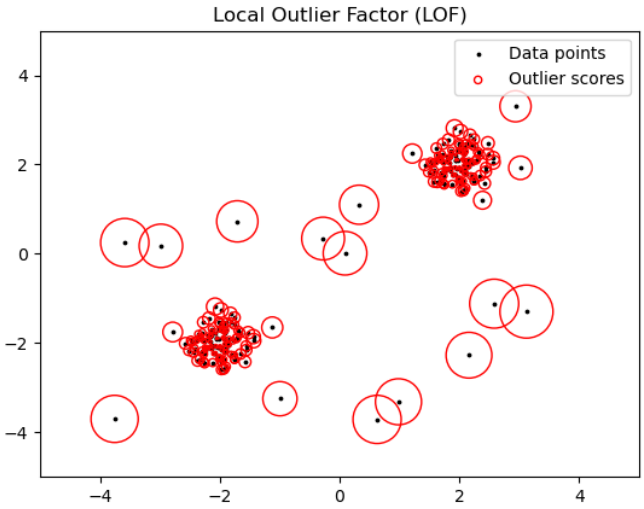

Figure 1.10: Local Outlier FActor analysis. The data points inside denser clusters are less likely to be outliers so have been assigned lower outlier scores (smaller radii). The data points in more sparser regions of the parameter space are certainly more outlier so they got larger outlier scores (larger radii). The x-axis and y-axis are showing the normalized (zero mean and a variance of 1) parameter space. Credit:scikit-learn

then we should choose which model is the best fit using the principle of model selection. But in the regression with Gaussian processes, we do not need to choose a fitting function because we are fitting a family of functions. The mean behavior of the family gives us the best fit and the variations of them provide the uncertainty (see Figure 1.11).

Gaussian Processes allow the data to speak for themselves. Noise in the data will be translated into uncertainty in the prediction which is fantastic for astronomical data which are always noisy. Gaussian Processes, realized more broadly as stochastic processes, is a generalization of probability distribution to *functions*. As working with Gaussian distributions makes computations tractable, Gaussian Processes is practical and robust in many problems. Gaussian Processes give a probabilistic yet much simpler to interpret solution. The most important strength of Gaussian Processes method is its ability of observational uncertainty to the prediction. Gaussian Processes

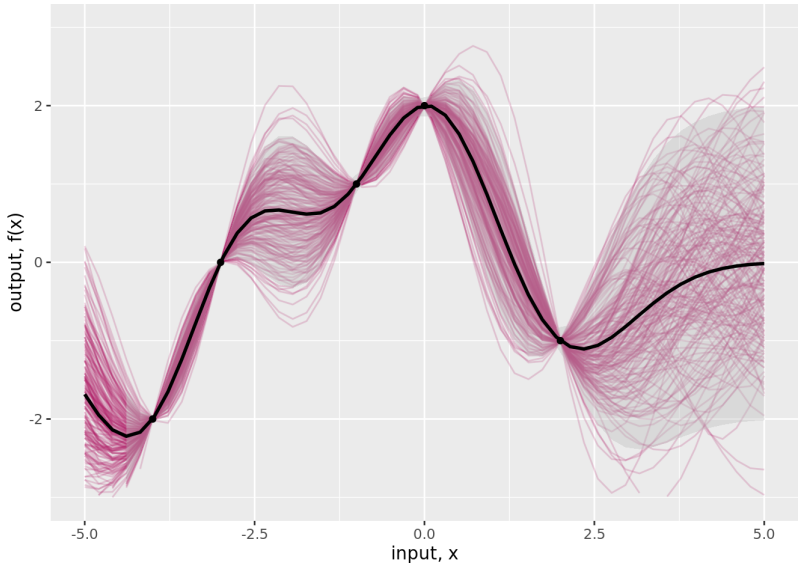

Figure 1.11: Gaussian Processes learns a family of functions (pink curves) given the observed data points (dots). The mean behavior of these learned functions can be used for predicting any data point not included in the observed data with an uncertainty reflected by the variations in the behavior of the learned functions. Credit: RPubs by R Studio

algorithm also provides a robust, interpretable tool for astronomical time series data analysis for the upcoming years (Golovich et al. 2022; Aigrain & Foreman-Mackey 2023).

The key assumption of the Gaussian Processes algorithm is that the observed data:

$$\vec{y}(\vec{x}) = \{y_1(x_1), y_2(x_2), ..., y_n(x_n)\} \tag{1.4}$$

can be thought of as a *single* sample point from a $n-$variate Gaussian distribution and when predicting about a desired new point $y_*(x_*)$ we have (Ebden 2015):

$$\begin{bmatrix} \vec{y}(\vec{x}) \\ y_*(x_*) \end{bmatrix} \sim \mathcal{N}(\vec{\mu}(\vec{x}, x_*), \boldsymbol{K}(\vec{x}, x_*)). \tag{1.5}$$

$\boldsymbol{K}$ here is the covariance matrix function that each element of that is a function of the predicting variable x_*. The covariance function in Equation 1.5 is determined by the observed data, $\vec{y}(\vec{x})$ via maximizing the probability of observing our prediction value given our data: $p(y_*(x_*)|\vec{y}(\vec{x}))$.

By maximizing this probability Gaussian Processes algorithm is actually *learning* the best mean function $\vec{\mu}$ and covariance matrix function $\boldsymbol{K}$. The covariance matrix, by having dimensions as large as the data, can capture the underlying, complex relationship between data in a way that is much easier interpretable than the hidden layers of neural networks.

1.6 Thesis outline

Chapter 2 presents the analysis of using unsupervised machine learning, specifically local outlier factor and kernel density estimation, to determine a statistically significant boundary in the the parameter space of measured properties of quasars in SDSS. This work resulted in a published paper and it is presented here.

In Chapter 3 we talk about the methodology of constructing a catalog of CIV absorption lines in quasar spectra. We show how the Gaussian Processes method which is a supervised machine learning can efficiently and accurately detect absorption lines. This work resulted in a submitted paper which is under review and we expect a publication date in 2023.

Chapter 2

Paper I: Improved selection of extremely red quasars with boxy CIVlines[1]

Abstract: Extremely red quasars (ERQs) are an interesting sample of quasars in the Baryon Oscillation Spectroscopic Sample (BOSS) in the redshift range of $2.0 - 3.4$ and have extreme red colours of $i - W3 \geq 4.6$. Core ERQs have strong CIV emission lines with rest equivalent width of ≥ 100Å. Many core ERQs also have CIV line profiles with peculiar boxy shapes which distinguish them from normal blue quasars. We show, using a combination of kernel density estimation and local outlier factor analyses on a space of the $i - W3$ colour, CIV rest equivalent width (REW) and line kurtosis, that core ERQs likely represent a separate population rather than a smooth transition between normal blue quasars and the quasars in the tail of the colour-REW distribution. We apply our analyses to find new criteria for selecting ERQs in this 3D parameter space. Our final selection produces 133 quasars, which are *three* times more likely to have a visually verified CIV broad

[1]This chapter contains the draft of an article that has been published in the journal of Monthly Notices of the Royal Astronomical Society. (Monadi & Bird 2022a)

absorption line feature than the previous core ERQ sample. We further show that our newly selected sample are extreme objects in the intersection of the WISE AGN catalogue with the MILLIQUAS quasar catalogue in the colour-colour space of ($W1 - W2$, $W2 - W3$). This paper validates an improved selection method for red quasars which can be applied to future datasets such as the quasar catalogue from the Dark Energy Spectroscopic Instrument (DESI).

2.1 Introduction

Quasars are high luminosity active galactic nuclei (AGN), fuelled by gas and dust accreting onto a supermassive black hole (SMBH). Observations show that the growth of a SMBH is correlated with the physical properties of the host galaxy, such as velocity dispersion, stellar mass, and star formation rate, although the mechanisms which correlate these properties are not completely clear (Gebhardt et al. 2000; Tremaine et al. 2002; Kormendy & Ho 2013; Azadi et al. 2015; Graham 2016). The co-evolution of a SMBH and its host galaxy mostly occurs during dusty starbursts resulting in the observation of sub-mm or ultra-luminous infrared galaxies (Sanders et al. 1988; Veilleux et al. 2009). Kroupa et al. (2020) proposed a model whereby SMBHs form when the dynamical collapse of star clusters is accelerated by the accretion of gas from a host galaxy, which can naturally explain the correlation between a host galaxy and SMBH mass, as well as explain quasars found at high redshift.

Unobscured quasars which exhibit blue thermal continuum are the majority of optically selected quasars. *Red quasars*, on the other hand, are a small population of quasars that show a variety of redder near infra-red and optical colours. Several studies have investigated the origin of

the red colour in red quasars (eg. Kim & Im (2018) or Klindt et al. (2019)), however the question still remains unsettled (Calistro Rivera, G. et al. 2021).

One possibility is that red quasars have been obscured and reddened by dust during a brief transition phase between dusty starburst galaxies and blue quasars (Richards 2003; Hopkins et al. 2005; Urrutia et al. 2008; Hopkins et al. 2008; Glikman et al. 2015; Banerji et al. 2015; Assef et al. 2015; Ishibashi & Fabian 2016; Hickox & Alexander 2018). In this model, a quasar is buried in the starburst dust when the host galaxy is young, making the colour of that quasar red. Many ERQs also exhibit extreme line properties which may indicate unusually powerful outflows occurring in a young evolution phase. Quasar-driven outflows may clear out the observer's line of sight, and so at the end of this evolutionary phase, we observe an optical and/or UV luminous quasar.

There are other models; for example, the unified AGN model (Antonucci 1993) suggests that red quasars are viewed with intermediate orientations between Type 1 and Type 2 quasars (for a recent review see Hickox & Alexander (2018)). According to this model, unobscured (type 1) AGN are viewed face-on, while obscured (type 2) AGN are observed edge-on. A red colour is produced by a dusty torus blocking part of the nuclear emission. However, this model has difficulty explaining the extreme line properties seen in some red quasars (Urrutia et al. 2009; Klindt et al. 2019).

Ross et al. (2015) studied a population of red quasars at $0.28 \leq z \leq 4.36$ in the Baryon Oscillation Spectroscopic Survey (BOSS Dawson et al. 2013) of the Sloan Digital Sky Survey-III (SDSS-III Eisenstein et al. 2011a). These red quasars were identified using a simple colour selection originally intended for red galaxies: a magnitude difference of $r_{AB} - W4_{Vega} \geq 14$ between the infra-red band ($W4$ in WISE with effective wavelength of 12μm) and the optical band (r in SDSS with effective wavelength of 6231Å).

Hamann et al. (2016) (hereafter H17) used the sample of Ross et al. (2015) but narrowed down the redshift range to $2.0 \leq z \leq 3.4$ and changed the colour selection to $i - W3 \geq 4.6$ ($\sim$ 3 magnitudes redder than the typical colour of BOSS quasars), calling the sample thus identified Extremely Red Quasars, or ERQs[2]. Interestingly, ERQs showed exotic spectral properties, which motivated H17 to define a smaller core ERQ (CERQ) subsample defined by REW(CIV)$\geq$ 100Å. This criterion was chosen to better correlate red quasars with other extreme line properties: peculiar *boxy* profiles, NV>Lyα, a high incidence of blue-shifted broad absorption lines (BALs), and [OIII] 5007Å outflow speeds reaching > 6000 km/s. ERQs also have an unusually flat UV SED considering their extreme red colour (steep Mid-IR to UV SED), although this may be an artifact of the BOSS selection algorithm, which would not target quasars which are red in all SDSS bands for spectroscopic follow-up.

Perrotta et al. (2019) (hereafter P19) studied a sample of 28 ERQs and found an outflow speed for the [OIII] line of 1992 - 6702 km/s. This is on average three times faster than those of luminosity matched blue quasars. This outflow speed is highly correlated with $i - W3$ colour but not with radio loudness nor Eddington ratios. P19 suggests that this correlation may indicate a connection between reddening and the efficiency of energy and momentum injection from ERQs to the interstellar medium. ERQs may produce more effective feedback in their host galaxies, regulating the star formation rate and SMBH growth more effectively. This is again indicative that some ERQs with extreme line values are connected with an early dusty stage of quasar-galaxy evolution where strong quasar-driven outflows provide important feedback to the host galaxies (P19).

We therefore have a working hypothesis identifying ERQs with an intermediate stage of quasar evolution between dusty galaxies and red quasars. This study is an effort to produce

[2]ERQs are not the reddest $i - W3$ quasars overall, but the reddest ones in BOSS.

quantitative evidence for or against this hypothesis, which is based on the unusual line properties exhibited by the spectra. If such quantitative evidence is forthcoming, we also desire to refine the selection criteria for ERQs in order to better study the outflows connected with this stage of quasar evolution. We will provide selection criteria for objects that exhibit the extreme properties of ERQs, among a sample of quasars with spectroscopic data. We use the existing manually selected sample of ERQs to define a training set, and then provide a modified sample of extremely red quasars in BOSS with more uniform (and more uniformly exotic) properties. In summary, we address the following questions:

- To what extent are ERQs separated from the main locus of BOSS quasars?
- If they are, which selection criteria best produce quasars connected with this intermediate stage of quasar evolution?

We acknowledge the possibility that our sample may be affected by the selection criteria of BOSS, which uses a colour selection to find quasar candidates and thus may discard some red quasars. However, in the absence of another equally large spectroscopic quasar survey this is unavoidable. We will thus analyse BOSS quasars and check for evidence that we are affected by colour selection in Section 2.5.5. We use a standard cosmology throughout this paper. ($H_0 = 67.3$ km s^{-1} Mpc $^{-1}$, $\Omega_m^0 = 0.315$, $\Omega_\Lambda = 0.685$) (Planck Collaboration et al. 2014).

2.2 Quasar samples

In this section, we introduce our quasar samples their selection criteria, summarised in Table 2.1. The primary parent sample is similar to the emission-line catalogue of H17 which results

Table 2.1: The parent samples, sizes and sample selection criteria for the various quasar samples we use. All subsets are taken from a parent sample made in H17 by custom emission line fits. The first sample, T1, is a superset of all the others. These are: type 1 luminosity matched quasars (T1LM), type 1 extremely red quasars (T1ERQ) and type 1 core extremely red quasars (T1CERQ).

Sample	Selection criteria	Size

T1	FWHM(CIV) $\geq$ 2000 km.s^{-1}
	$2 \leq$ z_dr12 ≤ 3.4
	$i - W3 \geq 0.8$
	SNR(REW(CIV)$\geq$3
	SNR(FWHM(CIV)$\geq$4
	SNR(AB$_{W3}$) ≥ 3
	`q_flag =0`
	`cc_flag='0000'`
	`nv_flag=0`
		35,976

T1LM	$10^{46.54}$ erg.s$^{-1} \leq$ L$_{bol} \leq 10^{48.00}$ erg.s^{-1}	29,072
T1ERQ	i-W3 $\geq$ 4.6	154
T1CERQ	i-W3 $\geq$ 4.6 REW(CIV)$\geq$ 100Å	72

from custom fits of CIV and NV emission lines performed on spectra in the SDSS-III BOSS quasar catalogue, The subsamples follow the selections in H17, to which we refer the reader for a detailed explanation of the criteria adopted.

Type 1 sample

Following the H17 sample selection procedure, we first limit the quasar redshift to $2 \leq z \leq 3.4$. This redshift range encompasses most of the BOSS survey, while ensuring that Lyα and NVλ1240 are within the BOSS spectral range.

We require that successful fits to the NV (`nv_flag=0`) and CIV (`q_flag =0`) emission lines are made at reasonable signal to noise (SNR(REW(CIV)) $\geq$ 3 and SNR(FWHM(CIV)$\geq$ 4)). We limit ourselves to Type 1 quasars, defined as FWHM(CIV) $\geq$ 2000 km s^{-1} (Ross et al. 2015).

We also require that the quasars have a good detection in the W3 band (SNR(AB_{W3}) $\geq$ 3), do not exhibit artifacts in the WISE data (`cc_flag='0000'`), and are not excessively blue ($i - W3 > 0.8$).

T1ERQ and T1CERQ samples

Following H17 we use a colour cut of $i - W3 \geq 4.6$ to extract ERQs from the full sample of type 1 quasars. H17 considered several colour cuts, choosing their boundary to produce the most dramatic differences in the median spectral properties of ERQs as compared to blue quasars. H17 also defined a subsample, core type 1 ERQs (T1CERQs), with the additional condition of REW(CIV)$\geq$ 100Å, chosen to be more correlated with the unusual line properties found in some ERQs. These conditions define a natural two dimensional parameter space in $i - W3$ and REW(CIV), which we will use extensively in what follows.

T1LM sample

ERQs are very luminous, with an average bolometric luminosity for T1CERQs of $10^{47.21\pm0.31}$ erg.s^{-1}. For comparison, the full quasar sample has an average luminosity of $10^{46.82\pm0.21}$ erg.s^{-1}. This high luminosity is a selection effect. SDSS cannot detect faint ERQs, because it does not detect objects with an i band magnitude $\lesssim$ 22.

2.3 Exotic Properties of ERQs

In this section we summarise the extensive discussion in H17 of the exotic properties of ERQs. This group of quasars has extreme red colours and strong CIV emission lines, accompanied by unusual boxy CIV emission-lines, unusually large NV/CIV line ratios, narrow CIV lines, and flat

UV spectra. The CIV line properties described here come from the emission line measurements described in H17. They fit two Gaussian components to CIV lines and derived the REW(CIV), kurtosis (see Section 2.3.3), and FWHM(CIV). In the following subsections, we describe these properties.

2.3.1 Extreme red colour

The SED of an ERQ is more luminous in the mid-infrared (Mid-IR) than in the UV part of the SED (see Figure 16 in H17) which gives an extreme red colour to ERQs. There are two possible reasons for this: UV light suppression and Mid-IR light enhancement. As indicated by H17, the mechanisms which lead to MIR light enhancement are not consistent with the extreme red colours of ERQs. However, UV suppression with patchy obscuration seems more consistent with the unusually flat SEDs of ERQs across UV

We use the flux magnitude differences between band-passes in SDSS and/or WISE to measure the colour of a quasar. H17 used $i - W3$, which is a measure of the flux ratio in units of AB magnitude for i and $W3$ bands. The reason why the i band in SDSS was chosen for colour measurement is that shorter wavelength pass-bands in the SDSS-like r-band could be severely contaminated by the CIV line for $z > 2.7$. $W3$ was selected because it has more sensitivity than the $W4$ filter. Only $\sim 40\%$ of the quasars in our sample with a $W3$ detection also have SNR>3 in $W4$. H17 tried other colour spaces, e.g. W3-W4, to distinguish ERQs. However, they found $i - W3$ is more tied to ERQ phenomena.

H17 shows that the exotic line properties of ERQs start to appear at $i - W3 > 4.6 \pm 0.2$. The distribution of ERQs with very high REW(CIV) values (> 150Å) is bimodal and has a dip in $i - W3 \sim 4.6$ (see Figure 4 in H17). Moreover, H17 confirmed that $i - W3 \sim 4.6$ is effective in

separating Type 1, high kurtosis, and high REW(CIV) quasars from other quasars in their sample (see top panel of Figure 5 in H17).

The red colour of T1CERQs is an intrinsic property tied to their extreme physical conditions. The median $i - W3$ for quasars with REW(CIV)>100Å and kt_{80}(CIV)> 0.3 are $\sim 2^m$ redder than the median of $i - W3$ for the full T1LM sample. Thus the strong and boxy CIV line is highly correlated with a red colour.

2.3.2 Strong C IV emission lines

ERQs have very strong CIV emission lines compared to the median T1 quasar in our sample. Non-core ERQs, those with REW(CIV)<100Å, tend to have a spectral energy distribution (SED) consistent with a template SED of Type 1 quasars with a reddening of $E(B - V) = 0.3$ (Figure 16 in H17). Therefore, ERQs with normal REW(CIV)'s tend to be normal quasars reddened by a screen of dust.

The high REW(CIV) in T1CERQs is especially unusual when combined with the high luminosity of ERQs. That is because there is an anti-correlation between luminosity and REW(CIV). This anti-correlation is known as the Baldwin effect (Baldwin 1977), while ERQs stand out from this trend. To see how much ERQs are behaving off-trend, we compared Baldwin REW(CIV)-Luminosity anti-correlation in two different samples: T1 and T1\ERQ[3]. More specifically, we compared the *R-squared*[4] of the regression between luminosity and REW(CIV) in T1 sample with the *R-squared* of the same regression but in T1\ERQ sample.

[3] \ is the set-minus symbol so that $\mathbb{A} \setminus \mathbb{B} = \mathbb{A} - \mathbb{A} \cap \mathbb{B}$.

[4] Note that, R-squared is the ratio between the variance of a model's prediction and the total variance of the dependent variable.

Even though ERQs are only ~ 5% of the T1 sample, we see a 15% increase for *R-squared* in the regression of REW(CIV) and luminosity in T1\ERQ compared to T1 sample. This implies that T1CERQs are very odd in the context of Baldwin effect as they have both high REW(CIV) and very high luminosity values.

2.3.3 Boxy C IV emission-line

The spectrum of an ERQ stands out by a peculiar boxy (wingless) CIV emission line. One of the methods for quantifying the shape of a line profile is the Kurtosis index (Hamann et al. 2016). As we mentioned above, H17 used two Gaussian components for fitting CIV line profiles. After obtaining the fitted line, they measured the ratio of the velocity width after which the line profile reached 80% of the peak height to the velocity width at 20% of the peak height:

$$kt_{80} \equiv \Delta v_{80\%}/\Delta v_{20\%}. \tag{2.1}$$

Kurtosis measures the relative strength of a CIV line's core component to its wing component. A high kurtosis CIV line profile indicates a boxy line, such as occurs in most ERQs the median of kt_{80} is ≥ 0.33 compared to the the T1LM sample with a median of 0.25.

2.3.4 Flat UV SED

The median SED of T1CERQs can distinguish them from other quasars and even other ERQs. T1CERQs' median SED is flat across the UV, yet very red in the Mid-IR. For example, Type 1 quasars with an $E(B-V) = 0.5$ reddening have the same $i - W3$ colour as some ERQs, but a very steep SEDs in UV rest-frame (see Figure 16 in H17). The median SED of T1CERQs is similar to HotDogs (Assef et al. 2015) but with less reddening. Obscuration by a clumpy dusty

torus can suppress UV light without dust reddening converting it to longer wavelengths; this can explain the flat Mid-IR to UV SEDs of ERQs. Obscuration is critical for ERQs, since it is closely related to their peculiar line properties, which can imply a unique physical condition tied to a special evolutionary phase in their host galaxies (H17). However, UV suppression by patchy obscuration seems more consistent with unusually flat SEDs of ERQs across the UV than dust reddening. The $i-z$ colour which is the flux difference between the i-band and the z-band in SDSS, can show the slope of the SED in the region where T1CERQs' median SED is unusually flat. T1CERQs thus have a much bluer $i-z$ than would be predicted for heavily reddened Type 1 quasars (HR1s: Banerji et al. (2013)), which have similar $i-W3$ colour to T1CERQs but much redder $i-z$ colour

2.3.5 Unusual NV to CIV line ratio

T1CERQs have a ⟨NV/CIV⟩ around 2.5 times larger than Type 1 quasars with REW(CIV)>100Å (similar to ERQs) but with no constrain on their colour (unlike ERQs). A large NV/CIV indicates high metallicity, which is common in massive and luminous galaxies (H17). Furthermore, ⟨NV/CIV⟩ in T1CERQs is around 85% of the median NV/CIV line ratio of the luminous normal blue quasars which obey Baldwing effect by having REW(CIV)<30Å. As a result, T1CERQs are a group of unusual NV/CIV line ratio quasars which their NV/CIV are neither similar to similarly high luminous quasars nor like normal blue quasars with similar REW(CIV).

2.3.6 Narrow C IV emission line

We adopted FWHM(CIV)>2000 km s^{-1} as a criterion for selecting Type 1 ERQs. This makes the classification of CERQs based on their FWHM(CIV) ambiguous; they have larger FWHM(CIV) than typical Type 2 quasars, but significantly smaller FWHM(CIV) than most Type 1

quasars. CERQs also have high REW(CIV) lines with boxy profiles and weak continua which resemble Type 2s. By removing Type 2s from our sample, we are removing the effect of geometry from our analyses, because we observe the torus/accretion disk of Type 2 AGN which heavily obscure their central engine. That is why we focus on Type 1 ERQs in this paper as a more homogeneous sample.

One possible explanation for the narrow FWHM(CIV) of T1CERQs is that the CIV line emitting region might be closer to narrow line region (NLR) than is usual for Type 1 quasars. In the standard classification of quasars (Antonucci 1993), Type 1 quasars show broad permitted lines which is indicative of a direct view of the high density, sub-parsec scale, region close to the accretion disk, with fast virial velocities. Type 2 quasars, on the other hand, show narrow forbidden lines coming from a low density environment much farther away from the accretion disk. However, T1CERQs, with significantly smaller FWHM(CIV) compared to the median of T1LMs, may hint at a spatially extended broad line region connected to their lower-density forbidden line regions (H17).

2.4 Analysis Methods

2.4.1 Kurtosis of the CIV line: a third parameter

H17 investigated a large number of unusual emission line properties in the T1ERQ and T1CERQ samples. In particular, they found boxy CIV line shapes, a large NV to CIV line ratio (NV/CIV) and moderately reduced FWHM(CIV) compared to a population of normal blue quasars with similar W3 magnitudes to T1CERQs. However, the NV fits done in H17 did not attempt to deblend the nearby Lyman-α line and so the NV strength may be overestimated. The CIV line is uniquely powerful for our analysis as it is the strongest metal line in the quasar spectrum.

We focus here on the boxy shape of the CIV line, quantified by the kurtosis. Kurtosis (kt_{80}) is defined in H17 as the ratio of the velocity width of the CIV line at 80% of the peak height to the velocity width at 20% of the peak height. A high kurtosis CIV line profile occurs in most ERQs and indicates a boxy line. The median kt_{80} for T1ERQs is 0.35 and for T1CERQs 0.36, while the larger T1LM sample has a median of 0.25.

Figure 2.1 show histograms of kt_{80}(CIV). Each panel is labelled by joint thresholds on REW(CIV) and $i - W3$ colour and the number of quasars satisfying these conditions. A redder colour skews the kt_{80}(CIV) distribution towards more boxy CIV line quasars (higher kt_{80}(CIV)). Increasing the REW(CIV) threshold does not change the overall shape of the kt_{80}(CIV) distribution dramatically, nor its most probable value when compared to the unconditioned sample in the top left panel. However, we see a slightly enhanced population of high kurtosis objects when conditioning on REW(CIV). High kt_{80}(CIV) is thus highly correlated with red colour, but not with REW(CIV), suggesting that it is a good choice for a third parameter, along with $i - W3$ and REW(CIV).

Note that there is a possible confounder in the fitting procedure of H17: weak lines will be fit with a single Gaussian rather than two if the second Gaussian does not improve the fit. A single Gaussian has $kt_{80} = 0.37$. We have checked that this does not significantly affect our results by making a version of Figure 2.1 where spectra with kt_{80}(CIV) > 0.37 have been removed. This reduces the total size of the sample by 34780 spectra and the number of core ERQs by 57. In practice, since most of the removed spectra are not ERQs this cut moderately strengthens the trends we report. Fig. 2.2 shows these high kt_{80}(CIV) objects in the low REW(CIV) and blue part of the parameter space. The median REW(CIV) for kt_{80}(CIV) > 0.37 and kt_{80}(CIV) > 0.36 are 17Å and

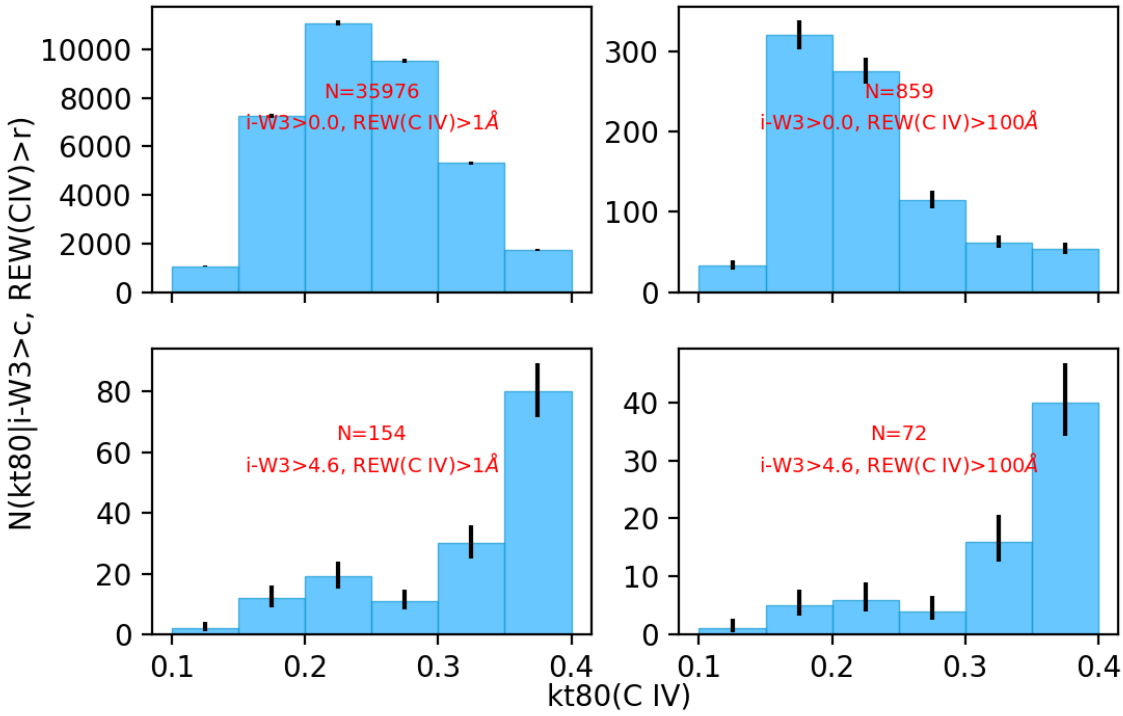

Figure 2.1: Histograms for distribution of kt_{80}(CIV) given conditions on $i-W3$ and REW(CIV). Top left panel shows the unconditioned distribution. Bottom left panel is conditioned on $i-W3>4.6$, and thus shows T1ERQs. Top right panel is conditioned on REW(CIV)> 100Å and bottom right panel is conditioned on both, thus showing T1CERQs. Each panel is labelled by the conditions and the number of quasars satisfying them. c and r in $N(kt_{80}(CIV)|i-W3>c,\ REW(CIV)>r)$ are the colour and REW(CIV) thresholds shown in each panel.

19Å respectively. This indicates that most of the quasars with high kt_{80} are weak CIV line objects, very far from the ERQs in colour space.

2.4.2 Defining T1CERQs with a wedge or a cone

One of our main goals in this study is to examine variations in quasar spectral properties as one moves in parameter space between the main quasar locus and T1CERQs. We define $\vec{v}_{T1CERQ}$, the vector in the 2D parameter space of colour-REW(CIV) between the median of the T1LM sample

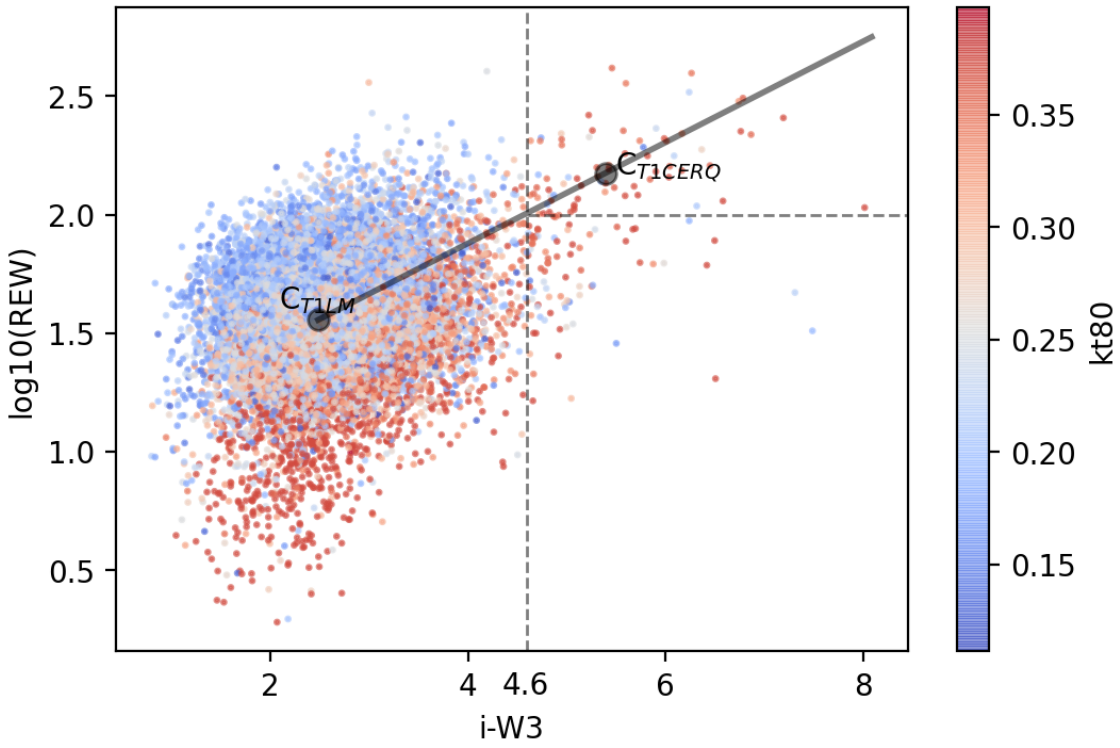

Figure 2.2: Luminosity matched sample distribution in ($i-W3$, REW(CIV), kt_{80}(CIV)) space. Redder points show higher kt_{80}(CIV) and thus higher kurtosis. The vertical line separates the T1ERQ sample from the rest of T1LM sample and the horizontal line separates T1CERQs from other T1ERQs. The black line is along $\vec{v}_{T1CERQ}$ (see Section 2.4.2), which connects the median of the T1LM (C_{T1LM}) sample to the median of the T1CERQ (C_{T1CERQ}) sample.

and the median of the T1CERQ sample, where the black line in Figure 2.2 shows its direction. In order to average over quasar properties, we then define wedges (in 2D) and cones (in 3D), the simplest directional geometric shapes for calculating the median spectra over a region in parameter space. We will use these shapes extensively in the following analysis[5].

A complexity to our definitions of wedges and cones is that we need a dimensionless parameter space within which to define opening angles. We thus normalise all parameters to a dimensionless unit square (cube), based on the range of each parameter. The normalization procedure we choose is the min-max method,[6] which performs a linear transformation to map each coordinate onto a unit square (cube). The maximum value found in the dataset maps to 1 and the minimum maps to 0. For our dataset, $i-W3$ ranges between 0.8 and 8.0, log 10(REW(CIV)) between 0.2 and 2.6, and kt_{80}(CIV) between 0.11 and 0.39.

We define a wedge in a 2D space of $i-W3$ and REW(CIV) along the vector between the median of the T1LM quasar sample and the T1CERQ sample. This vector is $\vec{v}^{2D}_{T1CERQ} = (0.40, 0.26)$ in the normalised space, which corresponds to (2.90, 136Å) once the normalisation is removed. The opening angle for this wedge is calculated by:

$$\theta = \max_i \{\measuredangle \vec{P}_i, \vec{v}_{T1CERQ}\}. \tag{2.2}$$

$\vec{P}_i$ is a vector from the median of the T1LM sample and the ith quasar in the T1CERQ sample. $\measuredangle$ means the angle between two vectors and is always less than 180°. Eq. 2.2 implies that θ is the maximum angle among all deviation angles of T1CERQs from $\vec{v}_{T1CERQ}$. θ is thus the smallest angle for which the wedge covers the entire T1CERQ sample, quasars with $i-W3 \geq 4.6$

[5]It is possible to consider more complex geometries. However this makes the analysis over-complicated and is not necessarily better than a wedge (or cone) which covers an area (or volume), as long as a variety of directions are included.

[6]We used the `MinMaxScaler` function from `sklearn`

and REW(CIV)≥ 100Å. The opening angle in the *normalised* 2D parameter space of $i - W3$ and REW(CIV) for $\vec{v}^{2D}_{T1CERQ}$ is $\theta = 18.5°$.

Equivalently, in the 3D normalised space of ($i - W3$, REW(CIV), kt_{80}(CIV)) we defined a cone towards the median of T1CERQs. But we observed that spectra with low kt80 did not generally exhibit the exotic line properties of T1CERQs. Therefore, we impose a minimum kt_{80}(CIV) ≥ 0.33. This cut excludes quasars with low kt_{80}(CIV) which have large θ angles from $\vec{v}^{3D}_{T1CERQ}$ and keeps the cone focused on the $\vec{v}^{3D}_{T1CERQ}$ direction. This threshold is chosen because it corresponds to the dip in the population distribution of kt_{80}(CIV) conditioned on red colour and strong REW(CIV) (bottom right panel of Figure 2.1). If we do not impose kt_{80}(CIV)≥ 0.33, the cone opening angle will be large, 35.7°, and will include many interloper quasars without the extreme line properties of T1CERQs. The corresponding vector between the median of the T1LM quasar sample and the T1CERQ sample is $\vec{v}^{3D}_{T1CERQ} = (0.40, 0.26, 0.38)$, which is (2.90, 136Å, 0.11) when the normalisation is removed.

The 3D cone with this definition thus includes all quasars with $i - W3 \geq 4.6$, REW(CIV)≥ 100Å, and kt_{80}(CIV)≥ 0.33. It has an opening angle of $\theta = 19.6°$ in the normalised 3D space.

Figure 2.2 visualises the boundaries between the T1LM, T1ERQ, and T1CERQ quasar samples in the parameter space of $i - W3$ and REW(CIV). Colours denote kt_{80}, showing again the large kt_{80} associated with ERQs. The line in Figure 2.2 is along $\vec{v}_{T1CERQ}$, directed from the median of the T1LM sample to the median of the T1CERQ sample.

2.4.3 Local Outlier Factor Analysis

Wishing to investigate whether T1CERQs are a separate sub population of quasars, we applied several clustering methods on our dataset. We tried density-based clustering techniques (e.g. DBSCAN Ester et al. 1996) and hierarchical clustering algorithms (e.g. agglomerative

clustering Day & Edelsbrunner 1984). However, clustering algorithms could not handle the very wide disparity in size between the T1LM sample (29,237 quasars) and the T1CERQ sample (72 quasars). Since T1CERQs are a very small portion (0.25%) of the total quasar sample, clustering methods were either not able to find T1CERQs as a separate cluster or, if they could, the uncertainties in the obtained labels were high.

Instead, we used a Local Outlier Factor (LOF)[7] analysis (Breunig et al. 2000) which quantifies the level of distinctness in T1CERQs. The LOF has had other uses in astronomy: for example, detecting unusual spectra in SDSS (Wei et al. 2013) and distinguishing supernovae candidates from massive galaxies (Tu et al. 2010). LOF measures the extent to which a data point is isolated with respect to its neighbours by comparing the local reachability density of an object to the local reachability density of its k-nearest neighbours using the following score:

$$LOF_k(A) = \frac{1}{\rho_k(A)} \frac{\sum_{B \in \mathbb{N}_k(A)} \rho_k(B)}{\|\mathbb{N}_k(A)\|}. \tag{2.3}$$

This is otherwise called the LOF score for the k-nearest neighbours of point A. $\rho_k(A)$ (or $\rho_k(B)$) is the local reachability density of the k-nearest neighbours of A (or B), defined by:

$$\rho_k(A) = \frac{\|\mathbb{N}_k(A)\|}{\sum_{B \in \mathbb{N}_k(A)} RD_k(A, B)}, \tag{2.4}$$

where $\mathbb{N}_k(A)$ (or $\mathbb{N}_k(B)$) is the set of all k-nearest neighbours of the point A (or B). $\|\mathbb{N}_k(A)\|$ is the number of objects in $\mathbb{N}_k(A)$. $RD_k(A, B)$ is the reachability distance between point A and B defined by:

$$RD_k(A, B) = \max\{\mathbb{D}_k(B),\ d(A, B)\}. \tag{2.5}$$

$d(A, B)$ in Eq. 2.5 is the Euclidean distance between point A and B in the normalised 2 or 3D space. $\mathbb{D}_k(B)$ is the set of all distances between point B and $\mathbb{N}_k(B)$.

[7] We used the LOF implementation in `sklearn` (Pedregosa et al. 2011).

For example, the density around a data point, deep in a dense cluster of points, is very similar to the density of its neighbourhood; this results in LOF$\sim$ 1. If the data point is located somewhere denser than its nearest neighbours, then it has LOF$<$ 1. A point where the average density of the neighbours is higher than that of the point has LOF$>$ 1, corresponding to the expected behaviour for a small cluster separate from the main group.

The LOF is defined as a function of the number of nearest neighbours, k, which sets the scale or resolution of the cluster searched for. Thus translated into our analysis, k provides information about the size of the putative T1CERQ cluster.

Mock Data Analysis in 2D

To better illustrate the behaviour of the LOF(k) score on known distributions of data points and for different k-nearest neighbours, we created 100 mock 2D data sets by making 100 draws from two overlapping Gaussian distributions, G_1 and G_2. To make each mock data set we draw 30000 data points from G_1: $\mathcal{N}(\mu = [0,0], \sigma = [1,0;0,1])$ and 200 data points from (G_2: $\mathcal{N}(\mu = [3,3], \sigma = [1,0;0,1])$), which in total gives us 100 mock data sets consisting of 30200 data points each. We chose the same covariance matrix for G_1 and G_2 for simplicity. However, the distance between the centers of G_1 and G_2 imitates the distance between the median of $i - W3$ in T1LM and the median of $i - W3$ in T1CERQs. Similar to the $i - W3 \geq 4.6$ and REW(CIV)$\geq$ 100Å cuts which define T1CERQs, we define a core G_2 sample (cG_2) with $x, y > 2.5\sigma$. The average population of cG_2 among our 100 mock data sets is 96 (see Figure 2.3). Moreover, on average only 1 data point from G_1 belongs to cG_2, while on average 104 data points from G_2 lay outside of the $x, y > 2.5\sigma$ cuts, showing the level of blending between the G_2 and G_1 populations.

We create a wedge, following the same procedure as Section 2.4.2, for our mock data. We are interested in the behaviour of data in a wedge directed from the median of the bigger population (G_1 in the mock 2D data, T1LM in the real dataset) towards the smaller population (cG_2 in the mock 2D data, T1CERQs in the real dataset). Note that the centre of G_2 is 3σ away from the centre of G_1. The opening angle for the mock wedge (see Section 2.4.2) is 20° to be close to the opening angle in 2D of the real data (18.5°) The corresponding unit vector that is directed from the center of G_1 to the center of G_2 is $\hat{v}_{cG_2} = (1/\sqrt{2}, 1/\sqrt{2})$. To see how LOF_k changes along $\hat{v}_{cG_2}$, we binned the wedge as a function of distance from the center of G_1, with bins shown in Figure 2.3.

We then calculated median LOF scores in each bin of the mock 2D data set using nearest neighbours (ie, cluster size) $k = 40, 50, 100$, and 150 for each of our 100 mock data sets. We plotted the corresponding 68% confidence intervals for median LOF scores within each bin in Figure 2.3. The LOF scores for the mocks have a local minimum in bin 4, well beyond the 68% confidence intervals of bin 3 and bin 5 and also consistent with the confidence interval of the difference between LOF scores of bin 3 and bin 4 (CI(LOF(3)) - CI(LOF(4)) when $k = 40$ or $k = 50$, but not when $k = 100$ or 150.

This local minimum in LOF score is caused by the local over-density in bin 4 from cG_2. A point in bin 6 close to the centre of G_2 is located in a denser region compared to the average point located in bin 3, where the transition between G_1 and G_2 happens; thus the average LOF score in bin 4 is smaller than in bin 3. In the terminology of the literature, bin 4 includes *locally less outlier* data points. Data points in bin 5 are far from the centre of G_2; as a result, data points in bin 4 are on average also *locally less outlier* than data points in bin 5. These two observations explains the dip in the LOF score of bin 4 for $k = 40$ and 50.

However, the local over-density in bin 4 (i.e. local minimum in median LOF score) is less significant when we consider more neighbours (i.e $k = 100$). This is because the population size of cG_2 in the sample is 96 points. A larger cluster includes much of G_1 in addition to G_2 (see Eq. 2.3). Thus LOF_{100} and LOF_{150}, by incorporating more nearest neighbours for a data point in bin 4, do not show a local minimum in the median LOF score. A significant local minimum in the LOF scores in a specific region of parameter space can therefore be used to find the boundary between two populations, even though they have dramatically different sizes.

We confirmed that this local minimum did not occur in other mock datasets without two clearly separated populations. We tested the LOF score variation in a single Gaussian population (G_1) with a normal distribution of $\mathcal{N}(\mu = [0, 0], \sigma = [3, 0; 0, 3])$. We generated LOF_k scores for $k = 40, 50, 100$ and 150 in 100 draws from G_1 each with $30,000$ data points, and used the same binning as for the mock dataset containing two Gaussian distributions. As before we looked at the median LOF_k score and 68% confidence intervals around it. The corresponding plot to Figure 2.4 never showed a local minimum in the LOF_k score.

Mock Data Analysis in 3D

To confirm that this dip also occurs in a mock 3D dataset, we performed a similar analysis for 3D Gaussian distributions $G_1 : \mathcal{N}(\mu = [0, 0, 0], \sigma = [1, 0, 0; 0, 1, 0; 0, 0, 1])$ with 30000 points and $G_2\mathcal{N}(\mu = [3, 3, 3], \sigma = [1, 0, 0; 0, 1, 0; 0, 0, 1])$ with 150 data points. We generated 100 mock data sets and used the same cuts for building cG_2 $(x, y, z \geq 2.5\sigma)$. On average the population of cG_2 in our 100 mock data sets is 50. For comparison, the population size of T1CERQs with the

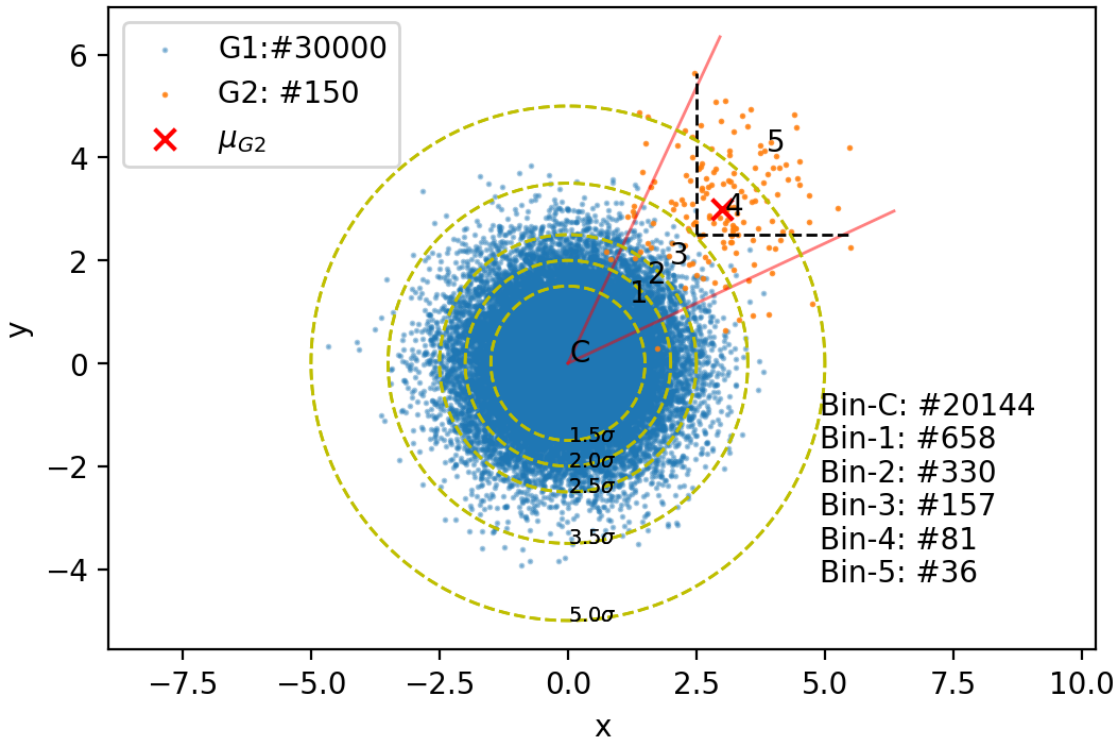

Figure 2.3: 2D density binning on a mock data set composed of two Gaussian populations: G_1: $\mathcal{N}(\mu = [0, 0], \sigma = [1, 0; 0, 1])$ with 30000 points (blue dots) and G_2: $\mathcal{N}(\mu = [3, 3], \sigma = [1, 0; 0, 1])$ with 100 points (orange dots). The red cross shows the center of G_2 sample. The yellow circles are at constant distances from the center of G_1 of 1.5σ, 2σ, 2.5σ, 3.5σ, and 5σ. The data points inside the dashed lines ($x, y > 2.5$) are those which would be selected as mock T1CERQs following the procedure outlined in Section 2.4.2.

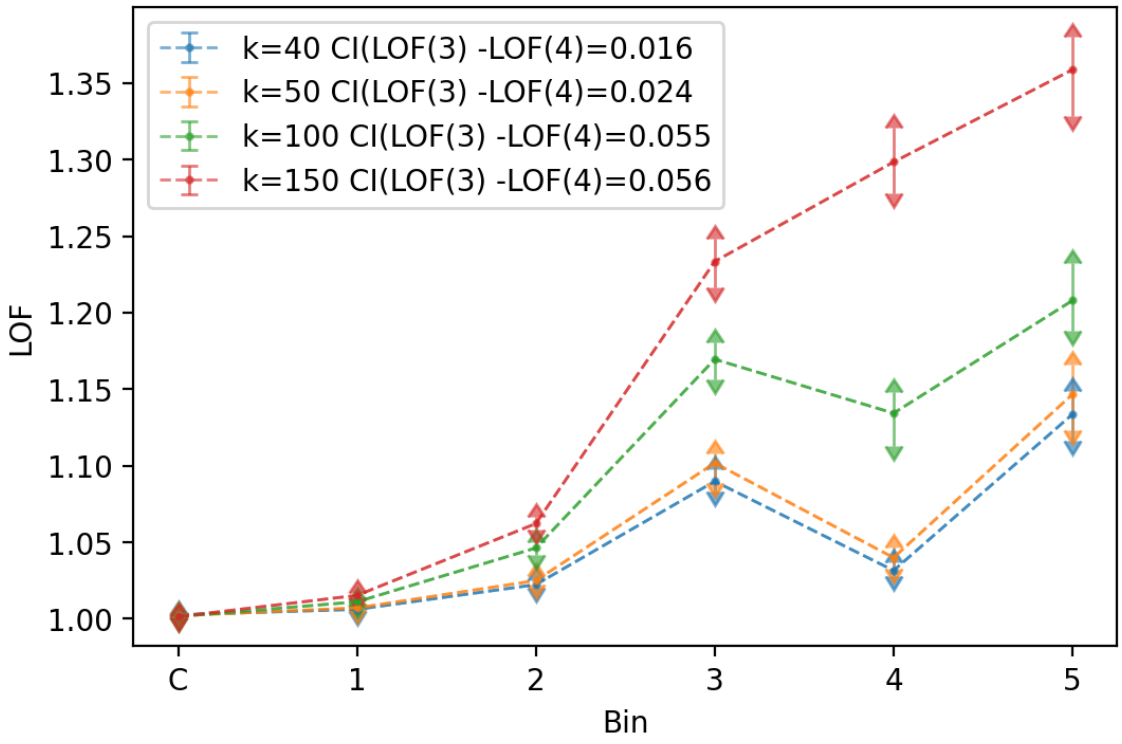

Figure 2.4: Bottom (b): Median LOF scores and their uncertainties in the bins shown in the top panel for nearest neighbours of $k = 40, 50, 100, 150$.

additional kt_{80}(CIV)$\geq$ 0.33 cut is 52. We never have a point from G_1 in the $x, y, z \geq 2.5\sigma$ region, but on average 100 data points from G_2.

We used a simple binning procedure similar to the one used for our mock 2D data. We defined a cone along $\hat{v} = [1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3}]$, directed from the centre of G_1 to the center of G_2 with an opening angle of 20°. We defined a central bin C where $r \leq 1\sigma$. Bin 1-5 are within the cone and between two spheres as follows: bin 1, $1\sigma \leq r \leq 1.5\sigma$; bin 2, $1.5\sigma \leq r \leq 2.5\sigma$; bin 3, $2.5\sigma \leq r \leq 4.8\sigma$; bin 4, $4.8\sigma \leq r \leq 7\sigma$; bin 5, $r \geq 7\sigma$.

Figure 2.5 shows the median LOF score in each bin for different nearest neighbours of $k = 70, 100, 150$, and 200. The 68% confidence intervals for each median LOF score is shown as the error bar. The decrease in the LOF score from bin 3 to bin 4 is more than the 68% confidence intervals of each bin and also more than the 68% confidence interval of the difference between the average LOF score in bin 3 and bin 4. As a result the local dip in the LOF score of bin 6 is significant to at least the 68% level.

Having demonstrated that a signature of two mixed populations (in 2D and 3D) is a dip in the LOF_k computed along the vector towards the smaller population, at a nearest neighbour value smaller than the size of the smaller population, we continue to analyse our real quasar sample.

2.5 Results

2.5.1 Density in the 2D parameter space

Our first objective is to gather evidence to determine whether T1CERQs are part of a separate population or extreme examples of T1 quasars, probing the tail of the main distribution. As

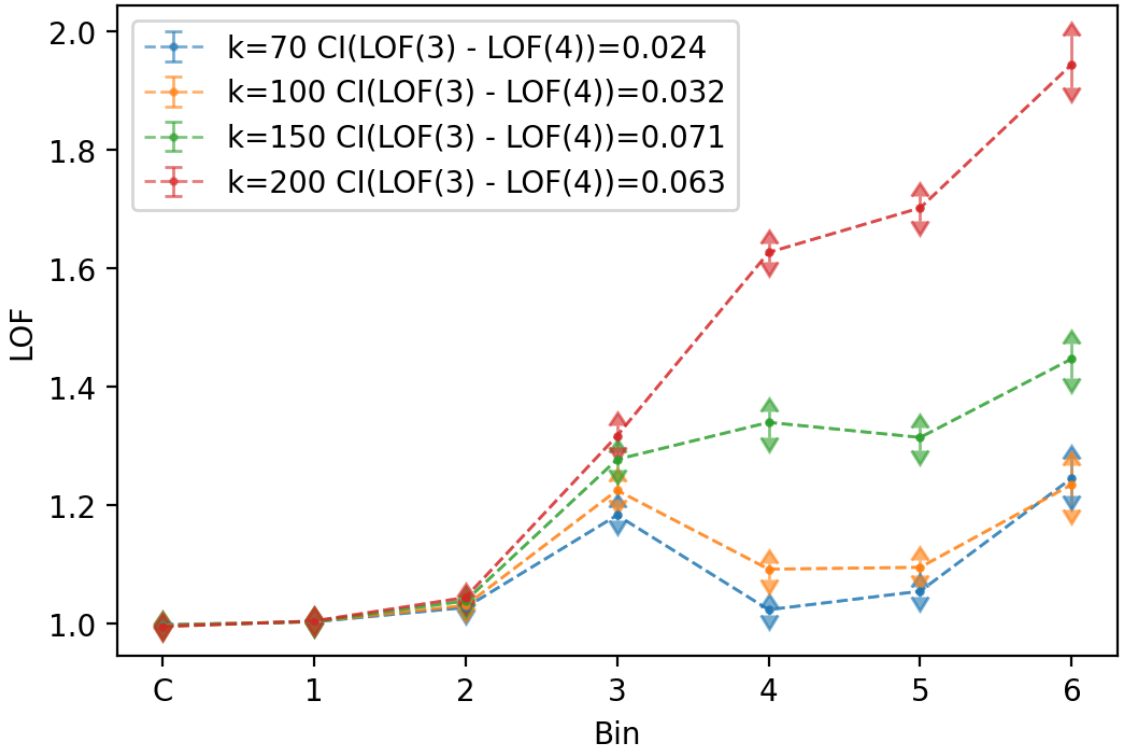

Figure 2.5: Mock 3D with 2 Gaussian population of G_1 : $\mathcal{N}(\mu = [0,0,0], \sigma = [1,0,0;0,1,0;0,0,1])$ with 30000 points and $G_2\mathcal{N}(\mu = [3,3,3], \sigma = [1,0,0;0,1,0;0,0,1])$ with 150 data points for the number of nearest neighbours: $k = 70, 100, 150$, and 200. CI(LOF(3)-LOF(4)) in the legend refers to the the 68% confidence interval for the difference of LOF score between bin 3 and bin 4 for each k.

a first, simple attempt to answer this question, Figure 2.6 visualises the quasars in the 2D parameter space of $i - W3$ colour, REW(CIV), normalised as explained in Section 2.4.2. It is visually apparent that the T1CERQs are over-dense compared to other regions of parameter space at a similar distance from the main quasar locus. To quantify how much, we computed the density of quasars in parameter space using kernel density estimation (KDE) with a Gaussian kernel. We want to compare the density of the parameter space to the high density region near the median of T1LM sample, and so we plotted density contours relative to the maximum density. For the KDE smoothing bandwidth we applied Silverman's rule of thumb to obtain the bandwidth for each dimension separately:

$$h_i = \sigma_i \left(\frac{4}{N(d+2)} \right)^{\frac{1}{d+4}}. \tag{2.6}$$

Here, σ_i is the standard deviation for the i-th dimension of our normalised parameter space, N is the number of objects (29237 for the T1LM sample), and d is the number of dimensions; 2 in 2D parameter space and 3 in 3D parameter space. Given $\sigma_{i-W3} = 0.077$ and $\sigma_{REW(\mathrm{CIV})} = 0.084$, we obtained $h^{2D}_{i-W3} = 0.014$ and $h^{2D}_{REW(\mathrm{CIV})} = 0.015$.

The density contours with $\rho > 0.05\rho_{max}$ in Figure 2.6 are similar in shape, and show the shape of the main quasar locus. However, the contours at lower densities are elongated in the direction of the T1CERQs (blue circles in Figure 2.6), including some mild local density maxima caused by T1CERQs. The low number of samples in this region means that the density contours are somewhat noisy, but it is apparent that the T1CERQs are an over-density in this parameter space. Figure 2.7 shows a similar density trend in the $i - W3$, kt_{80} plane. Here the over-density near the T1CERQs is even more apparent: the lowest density contour is extended at $\mathrm{kt}_{80} \sim 0.35$ towards high $i - W3$.

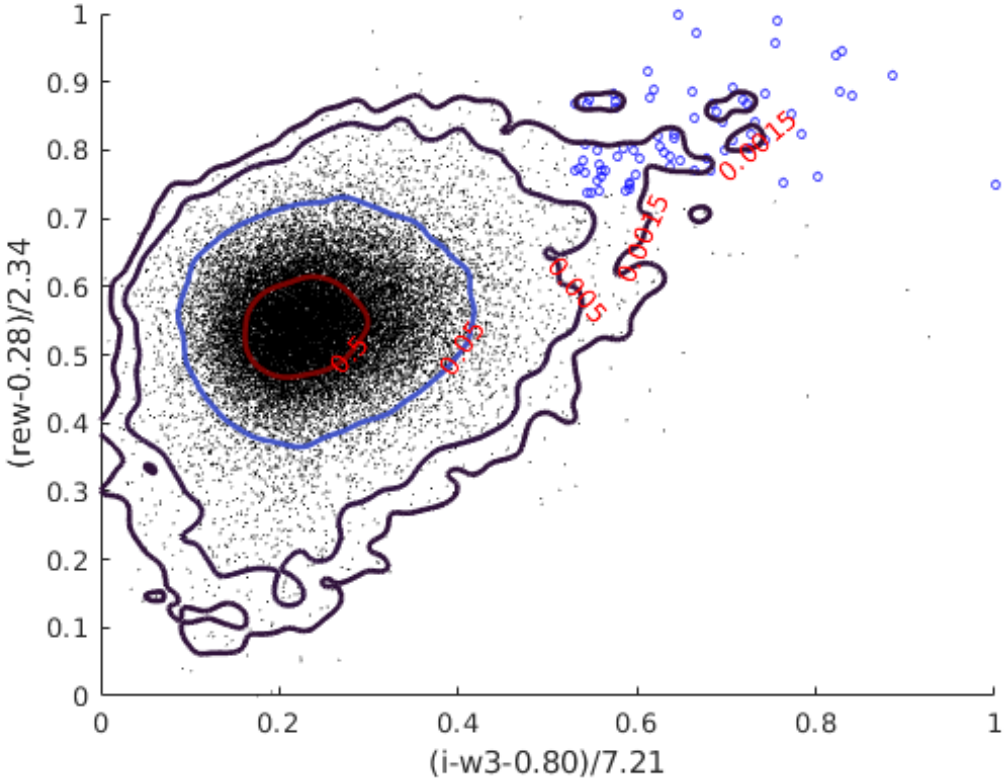

Figure 2.6: Density of quasars in a normalised $i-W3$, REW(CIV) space. Density contours are shown relative to the maximum density at $\rho/\rho_{max} = 0.5, 0.05, 0.005, 0.0015$. Blue circles are T1CERQs. Black dots are the other T1LMs.

A possible explanation for these outer contours is the non-linear effect of dust reddening on the colour distribution of quasars (Richards et al. 2001). However, Fig. 11 of H17 shows that the typical SED of ERQs is very different from the SED of dust-reddened quasars without the strong CIV line characteristic of core ERQs. Core ERQs have SEDs which are much flatter in the rest-frame UV than suggested by their red $i - W3$ colours, while non-core ERQs exhibit a sharp decline in the near UV with only moderately red colours across the near IR, similar to type 1 QSOs reddened by dust extinction (H17).

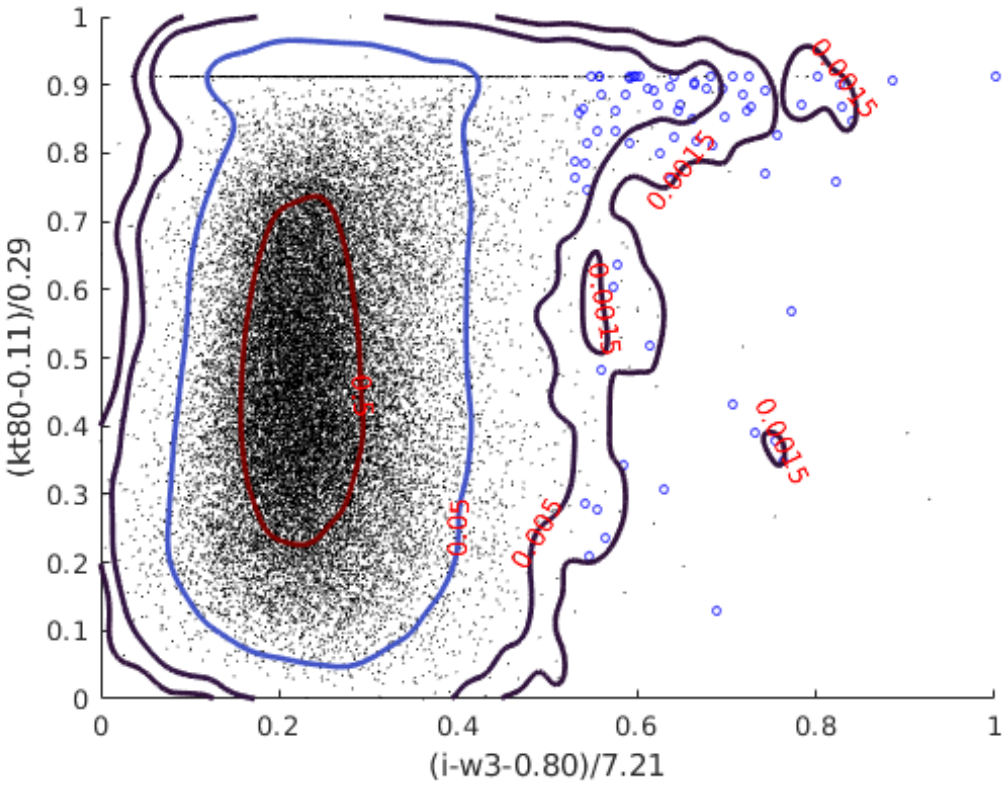

Figure 2.7: Density of quasars in a normalised $i - W3$, kt_{80} space. Density contours are shown relative to the maximum density at $\rho/\rho_{max} = 0.5, 0.05, 0.005, 0.0015$. Blue circles are T1CERQs. Black dots are the other T1LMs.

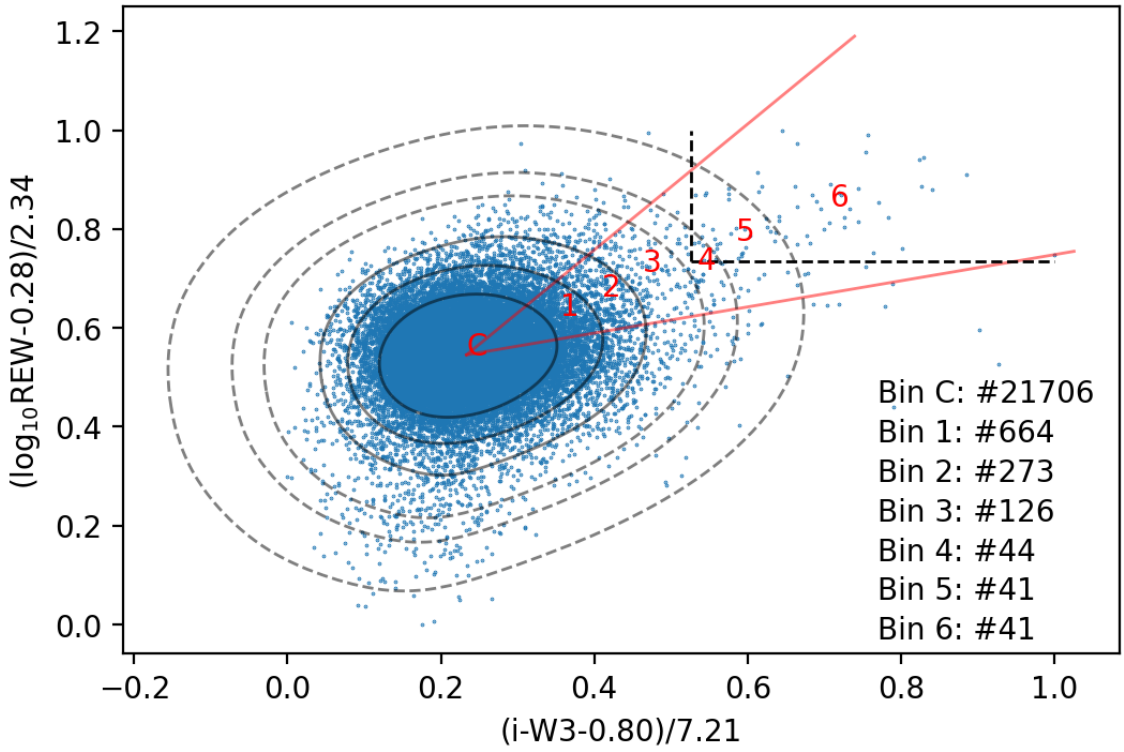

Figure 2.8: A binned wedge along $\vec{v}_{T1CERQ}$ towards the T1CERQ sample, with bins defined by density contours. The population of each bin is provided. Bin-C is enclosed by the innermost solid line contour at $0.3\rho_{max}$. 2nd and 3rd contours are at the levels of 0.03 and 0.01 of ρ_{max}. The three outer dashed line contours are ×1.35, ×1.55, and ×1.95 enlarged version of the biggest solid line contour.

2.5.2 Median Spectra

A second intuitive way to examine exotic line properties is to make median spectra for the T1CERQ sample. Since we have, in Section 2.4.2, constructed vectors towards the T1CERQ sample, together with geometric cones which contain the T1CERQ quasars. We are now able to bin the cones and make median spectra within these bins and examine how the spectra of T1CERQs

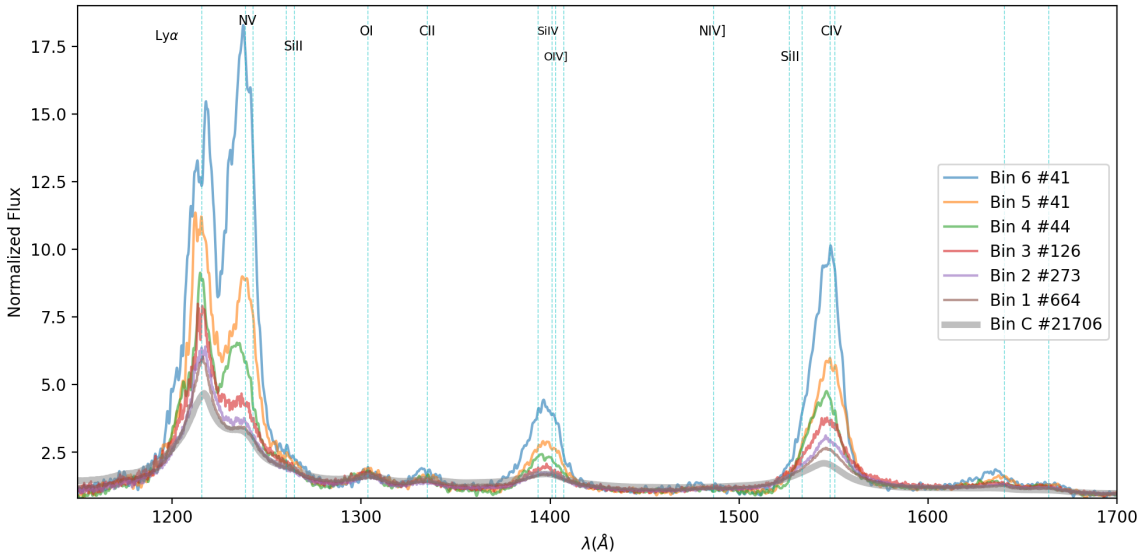

Figure 2.9: Median spectra for bins in the direction of $\vec{v}_{T1CERQ}$. The bin number and the number of quasars in each bin are shown. As a reminder, T1CERQs are found in bins 5 and 6 and part of bin 4.

Table 2.2: List of median physical properties in each bin from Figure 2.8. "C" in the 2nd column denotes the central bin. Column No. shows the number of quasars in each bin. f_{BAL} is the fraction of quasars which contain a visually verified BAL feature near the CIV line. Other columns show the median values and their standard deviations in each bin.

Bin	No.	$i - W3$	REW(CIV)	FWHM(CIV)	kt_{80}(CIV)	NV/CIV	f_{BAL}	Luminosity
C	21706	2.45 ± 0.35	35 ± 10	5400 ± 1500	0.25 ± 0.05	1.54 ± 0.55	0.14	46.83 ± 0.21
1	664	3.37 ± 0.15	55 ± 8	4700 ± 1700	0.24 ± 0.06	1.04 ± 0.46	0.22	46.76 ± 0.19
2	273	3.76 ± 0.18	68 ± 13	4300 ± 1700	0.23 ± 0.07	0.95 ± 0.55	0.29	46.79 ± 0.22
3	126	4.14 ± 0.24	89 ± 22	3900 ± 1500	0.28 ± 0.07	0.99 ± 0.60	0.32	46.86 ± 0.24
4	44	4.65 ± 0.17	92 ± 30	3500 ± 1400	0.35 ± 0.07	1.66 ± 0.78	0.52	47.02 ± 0.28
5	41	5.02 ± 0.26	125 ± 42	3300 ± 1200	0.35 ± 0.06	1.28 ± 0.76	0.29	47.09 ± 0.32
6	41	5.90 ± 0.57	181 ± 80	3100 ± 900	0.36 ± 0.06	1.74 ± 0.58	0.12	47.30 ± 0.29

change in 2- and 3D parameter space. Median spectra are made by stacking, after the following pre-processing steps:

1. Shift the observed flux into the quasar's rest frame.

2. Normalise the spectrum by the median flux between 1680Å and 1730Å in the rest frame. This region was chosen as the quasar spectrum is mostly free from significant line features.

3. Interpolate all fluxes onto a logarithmic grid defined between 800Å and 3000Å.

Since visualisation is easier in 2D, we perform our first median spectrum analysis by binning in the normalised parameter space of $i - W3$ and $\log_{10}$REW(CIV) along the $\vec{v}_{T1CERQ}$ direction described in Section 2.4.2. Figure 2.8 shows the 2D wedge towards $\vec{v}_{T1CERQ}$, together with the density contours around which we define the bins for our median spectra. The bin boundaries are chosen to bring out specific features of the median spectra. Three inner density contours at $0.3\rho_{max}$, $0.1\rho_{max}$, $0.03\rho_{max}$ show the shape of the core of the T1LM sample. The three outermost contours scale the contour at $0.03\rho_{max}$ by 1.35, 1.55, and 1.95, as the number of quasars this far from the main locus is too low to accurately estimate density. The choice of the 1st (1.35) and 2nd (1.55) scale factors gives a bin that covers the boundary of T1CERQs suggested by H17 (the lower left corner of REW(CIV)>100Å and $i - W3$ ≥4.6 box in Figure 2.8). The 3rd scale factor (1.95) is chosen so that bins 5 and 6 are equally populated.

Figure 2.9 and Table 2.2 show that there is an evolution in the line properties of quasars along $\vec{v}_{T1CERQ}$. The median CIV emission line in bins 1 through 3 is symmetric, close to the shape of the median spectrum in bin C, the main T1LM quasar locus (shown by the thick grey spectrum in Figure 2.9).

Table 2.2 shows that there is a jump in the CIV line kurtosis in bin 4: median kt_{80}(CIV) is 0.28 in bin 3 and 0.35 in bins 4-6. Bin 4 also has an unusually large BAL fraction (0.52). To ensure that these properties are due to the T1CERQ vector and not a function of distance from the quasar locus, we also checked the line properties along $-\vec{v}_{T1CERQ}$ and confirmed that kt_{80}(CIV) remained similar to bin C, while the BAL fraction dropped to 0.07. We confirmed these trends by making median spectra along vectors both clockwise and anti-clockwise of $\vec{v}_{T1CERQ}$, again finding that the line kurtosis remained low and confirming that the $\vec{v}_{T1CERQ}$ direction is unique.

There is also a relatively large jump in NV/CIV from ~ 1 in bins $1-3$ to 1.66 in bin 4. We found that NV/CIV also increased along $-\vec{v}_{T1CERQ}$, perhaps indicating that this is not intrinsic to T1CERQs, but in this direction the NV line is weak and the fit is likely to suffer severely from blending with the Lyman-α line.

3D parameter space

Motivated by the success of our 2D analysis, we made median spectra in the 3D normalised parameter space of $i - W3$, $\log_{10}$REW(CIV), and kt_{80}(CIV). The median spectra bins were made within the 3D cone defined in Section 2.4.2. The central bins were again defined by density iso-surfaces relative to the maximum density and computed using a KDE as in Section 2.5.1. The central quasar locus, bin C, was $\rho > 0.5\rho max$. Bin 1 has $\rho = 0.5\rho max - 0.05\rho max$ and bin 2 is $\rho = 0.05\rho max - 0.01\rho max$. As for our 2D binning, we did not use the density iso-surfaces at lower density levels, because the low numbers of spectra in these bins make local density estimates too noisy. Instead, we uniformly enlarged the iso-surface of $0.01\rho_{max}$ by factors of 1.5, 2.1, and 2.5 to make three extra surfaces and used these enlarged surfaces for bins 4 though 6. The expansion factors of 1.5 and 2.1 were chosen so that bin 4 covers H17's boundary for ERQs (the $i - W3 \geq 4.6$

Table 2.3: List of median physical properties in the bins of Figure 2.10. Columns are named as in Table 2.2.

Bin	No.	i-W3	REW(CIV)	FWHM(CIV)	kt_{80}(CIV)	NV/CIV	f_{BAL}	Luminosity
C	8284	2.40 ± 0.21	35 ± 6	5600 ± 1300	0.25 ± 0.03	1.60 ± 0.50	0.11	46.85 ± 0.21
1	386	3.10 ± 0.20	47 ± 7	5400 ± 1400	0.28 ± 0.01	1.22 ± 0.44	0.16	46.79 ± 0.18
2	93	3.69 ± 0.21	58 ± 10	4500 ± 1800	0.30 ± 0.02	1.09 ± 0.55	0.30	46.80 ± 0.22
3	94	4.08 ± 0.33	78 ± 18	3900 ± 1600	0.32 ± 0.03	1.29 ± 0.58	0.52	46.91 ± 0.22
4	64	4.74 ± 0.33	93 ± 24	3500 ± 1300	0.36 ± 0.02	1.68 ± 0.72	0.55	47.19 ± 0.27
5	23	5.43 ± 0.35	142 ± 44	3000 ± 1400	0.37 ± 0.01	1.78 ± 0.71	0.30	47.20 ± 0.33
6	23	6.17 ± 0.61	209 ± 92	3500 ± 1100	0.36 ± 0.01	1.74 ± 0.65	0.04	47.49 ± 0.27

plane in Figure 2.10). The scale factor of 2.5 was chosen so that the last two bins had an equal sized population (23 quasars in each).

All 3D bins are colour coded in Figure 2.10. The transparent box in Figure 2.10 shows $i - W3 \geq 4.6$, $\log_{10}(REW(CIV)) \geq 100$Å, and kt_{80}(CIV)≥ 0.33. Median spectra of these 3D bins are plotted in Figure 2.11 and Table 2.3 summarises the median physical properties in each bin of Figure 2.10. As for the 2D analysis, the kurtosis increments in each bin from 1 to 3 and saturates at 0.33 in bin 4, suggesting bin 4 is a good candidate for a boundary separating a population of red quasars from the main T1LM sample. NV/CIV is larger in bin 4 – 6 compared to bin 1 though bin 3. It is again large in bin C, but this may again be due to blending with Lyman-α. As in 2D, bins 3 and 4 have a high fraction of BALs, although this is not true of bin 6. In general the trends in 3D are similar to those in 2D: this extra parameter, however, will be useful in the next sections.

2.5.3 Local Outlier Factor Analysis

We showed in Section 2.5.1 that T1CERQs are an over-density when compared to other quasars at a comparable distance from the centre of the main population, and in Section 2.5.2 we found that there was an increase in kurtosis around the fourth bin from the central T1LM sample.

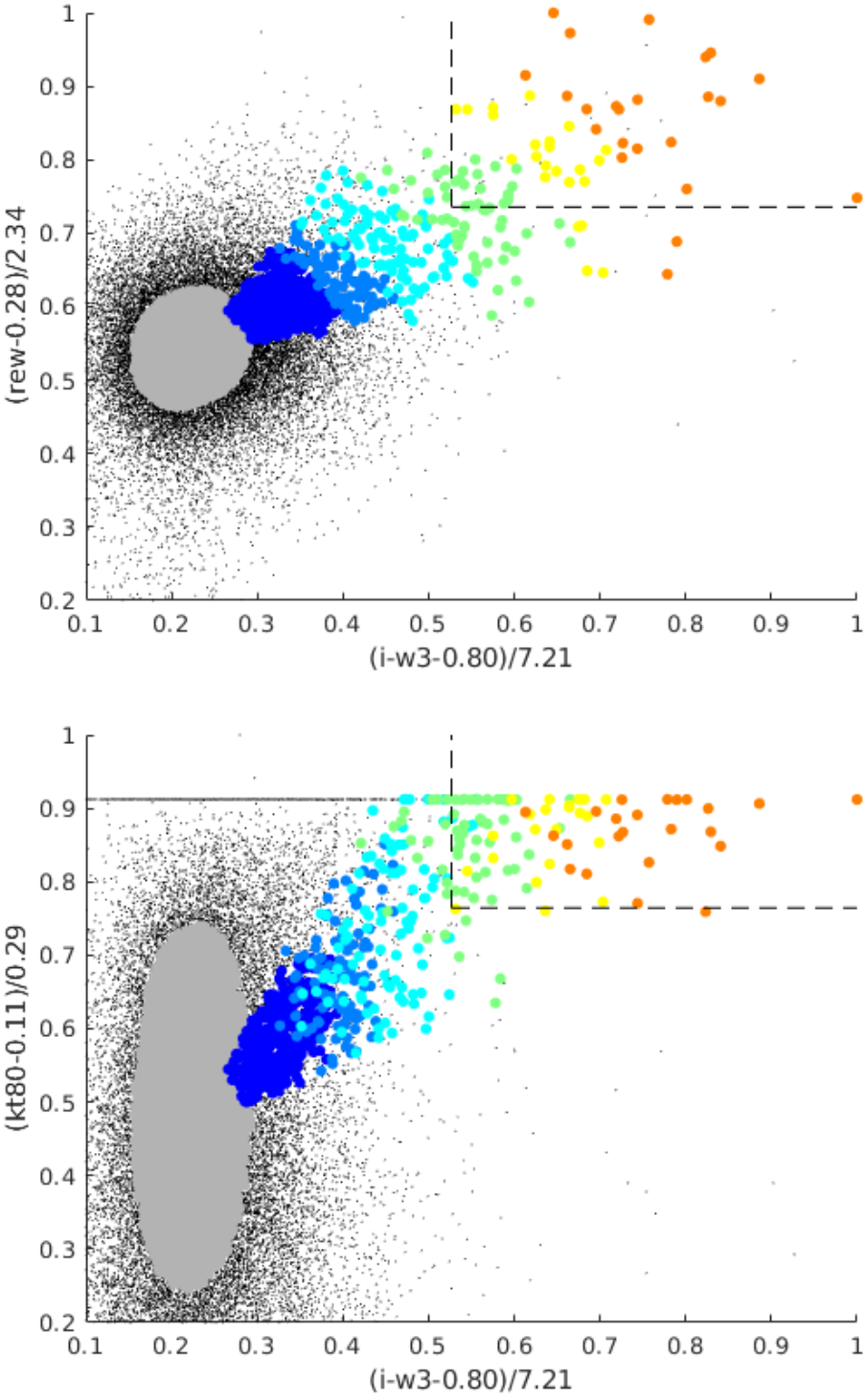

Figure 2.10: 3D bins along a cone around $\vec{v}^{3D}_{T1CERQ}$. The central bin is shown by grey points at the centre. Each bin, separated by density iso-surfaces as described in Section 2.5.2, is painted a different colours. Dashed lines shows the region of $i - W3 \geq 4.6$, $\log_{10}$REW(CIV)$\geq$ 2, and kt_{80}(CIV)$\geq$ 0.33 in the min-max normalised space.

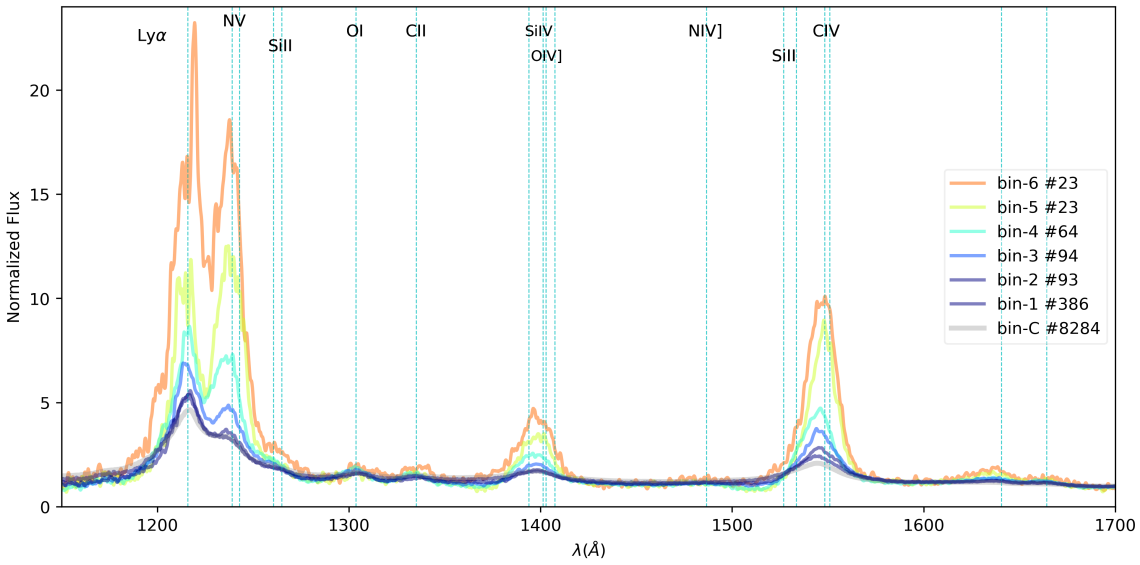

Figure 2.11: Median spectra for the corresponding coloured objects in each bin of the top panel. Spectra colours for each bin match those in Figure 2.10.

In this section, we quantify the distinctness of T1CERQs from the main T1LM sample using LOF, and examine the 2D and 3D candidate boundaries we found in Section 2.5.2. As a reminder, in Section 2.4.3 we showed that a signature of two distinct populations is a dip in the LOF score.

LOF Analysis in 2D

We now proceed to analyse the full T1LM sample with LOF along a vector directed towards the T1CERQ, $\vec{v}_{T1CERQ}$. We use the bins depicted in Figure 2.8, and compute a median LOF score in the *normalised* 2D space of $i - W3$ and $\log_{10}$REW(CIV)). There is a (small) dip in the median LOF score for bin 4. This is interestingly consistent with the results from median spectra in Section 2.5.2, where we saw that bin 4 was also associated with unusual spectral properties. The magnitude of the dip for $k = 40$ is 0.019, which is somewhat less than the 68% uncertainty of the LOF score in our 2D mock data analysis (σ(LOF(bin 3)-LOF(bin 4)) = 0.022). The 2D LOF

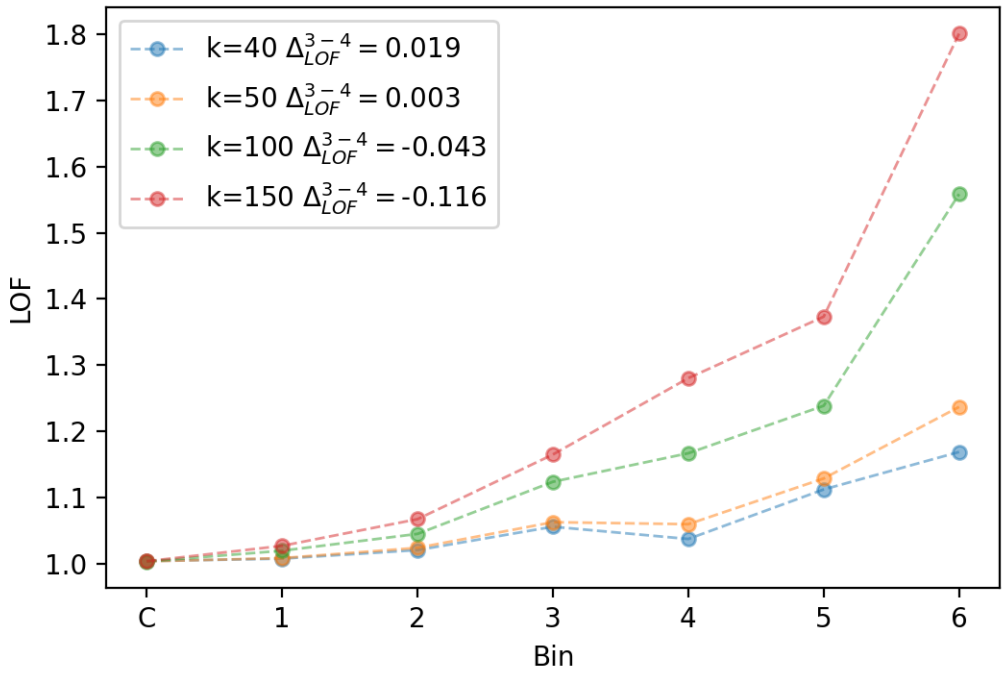

Figure 2.12: Median LOF score in each bin within the wedge of Figure 2.8, along the vector $\vec{v}_{T1CERQ}$ between the centroid of T1LM and T1CERQ, for different numbers of nearest neighbours (k).

analysis thus provides indications that the T1CERQs may be a separate population from the main T1LM, but is by no means definitive.

Figure 2.12 also shows the dependence of LOF score on neighbour number, k. for $k = 40, 50, 100$, and 150. The LOF score falls from bin 3 to bin 4 when the number of nearest neighbours is 40 or 50, and monotonically increases for $k = 100$ or 150. The LOF score thus suggests that a putative separate population of T1CERQs would have a population between 50 and 100, in good agreement with H17, who identified a population of 72 T1CERQs.

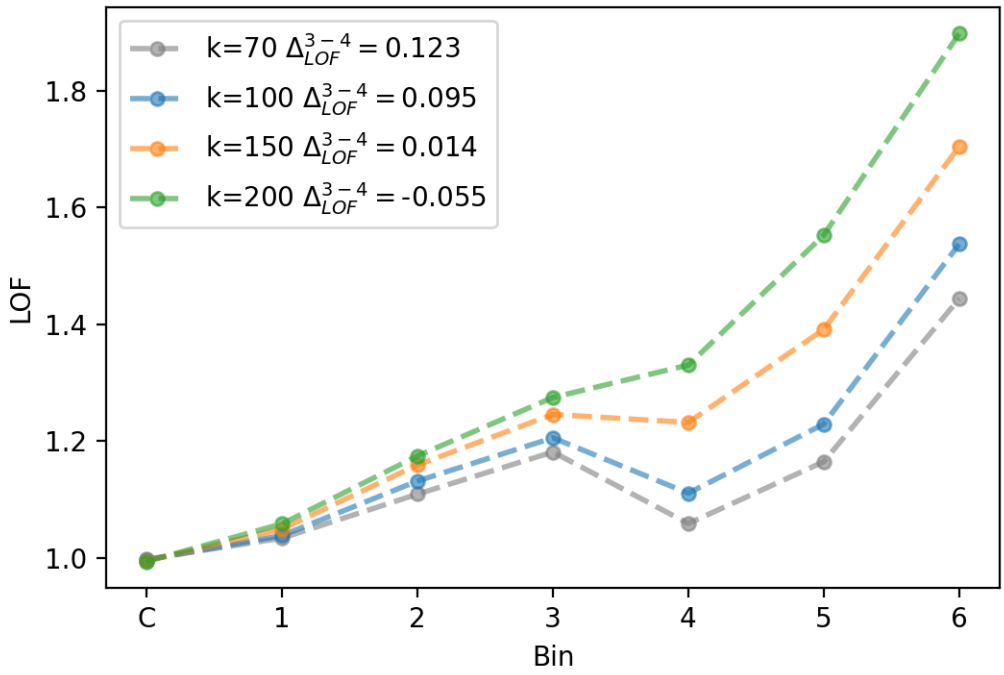

Figure 2.13: Median LOF score in each 3D bin of Figure 2.10, along the 3D vector $\vec{v}^{3D}_{T1CERQ}$ between the centroid of T1LM and T1CERQ, for different numbers of nearest neighbours (k).

LOF Analysis in 3D

We performed a similar LOF analysis in 3D, adding kt_{80}(CIV) to $i-W3$, $\log_{10}$(REW(CIV)) and using a normalised space. Figure 2.10 once again shows a dip in the median LOF score for bin 4. This fall in the LOF score from bin 3 to bin 4 is 0.123 for $k = 70$, significantly larger than the 68% uncertainty of the similar bin in the mock 3D data analysis (σ(LOF(bin 3)-LOF(bin 4)) = 0.025), as Figure 2.5 shows. The LOF analysis in 3D is thus evidence that there is a separate population of quasars along the vector to T1CERQs. Interestingly, Figure 2.5 shows that the dip in the LOF score persists for $k \leq 150$, suggesting that the separate population may be somewhat larger than that found in H17.

2.5.4 Selecting T1BERQs in 3D

H17 found a sub-population of quasars in 2D space, the T1CERQs. Our median spectra and LOF analysis has provided quantitative evidence that this sub-population is distinct from the general trend of the T1LM sample, especially when viewed in the 3D parameter space of $i - W3$, $\log_{10}$(REW(CIV)) and kt_{80}(CIV). There are also indications in the LOF score that the sub-population is moderately larger than the T1CERQ set found by H17. In this section we will design 3D criteria which optimises the selection of these objects. We call our new subset of quasars Type 1 boxy CIV emission line extremely red quasars (T1BERQs).

The choice of our kt_{80}(CIV) parameter space is also motivated by Figure 2.1, which suggests that there are a small number of low kt_{80}(CIV) objects within the T1CERQ class and that a minimum kt_{80} condition will produce a purer sample. Here we outline our recipe for selecting T1BERQs, summarizing steps introduced in earlier sections:

1. Normalise the parameter space of ($i - W3$, $\log_{10}$(REW(CIV)), kt_{80}(CIV)) with a Min-Max scaler (discussed in Section 2.4.2).

2. Define $\vec{v}^{3D}_{T1CERQ}$, a vector from the median of the T1LM sample to the median of those points satisfying $i - W3 \geq 4.6$, REW(CIV)$\geq$ 100Å, and kt_{80}(CIV)$\geq$ 0.33), in the normalised space.

3. Find a cone along $\vec{v}^{3D}_{T1CERQ}$, with a tip located at the median point of T1LM, and an opening angle so that the cone includes all quasars satisfying $i - W3 \geq 4.6$, REW(CIV)$\geq$ 100Å, kt_{80}(CIV)$\geq$ 0.33.

4. Using KDE, find density iso-surfaces and bin the cone in the previous step by successive iso-surfaces. One of the bins passes through an initial guess about the boundary of the desired population (here $i - W3 \geq 4.6$, REW(CIV)$\geq$ 100Å, and kt_{80}(CIV)$\geq$ 0.33).

5. Calculate the LOF score in each bin.

6. Find the bin showing a local minimum in LOF score.

7. Repeat steps (4) to (6) varying the inner and outer boundaries of the candidate bin found in step (6) and find the bin which shows the largest decrease in LOF scores as compared to the neighbour bin located closer to the centre of the T1LM sample.

8. Find a plane perpendicular to $\vec{v}^{3D}_{T1CERQ}$ and tangent to the inner boundary of the optimum bin in step (7).

9. Define the boundaries of the T1BERQs using the common region between the plane of step (8) and the cone of step (3).

Following this procedure, we found all quasars in a cone with a tip at (in our normalised 3D parameter space) $(0.23, 0.54, 0.47)$ and an opening angle of 19.6°, the same cone as in Section 2.4.2. The bin with a minimum LOF score was bin 4, as in Section 2.5.4, and the optimised, adjusted boundary between bins 4 and 3 was expanded by a factor of 1.5 from the bin boundaries of Section 2.5.2. The change in the LOF score across this bin boundary increased moderately to LOF(bin 3) - LOF(bin 4) = 0.130 for $k = 70$. Thus, the plane in step (8) of our procedure passes though a point $(0.50, 0.72, 0.75)$ in the *normalised* space of ($i - W3$, $\log_{10}$(REW(CIV)), kt_{80}(CIV)) with a normal vector of $\hat{n} = (0.64, 0.41, 0.64)$. We thus define T1BERQs by the following

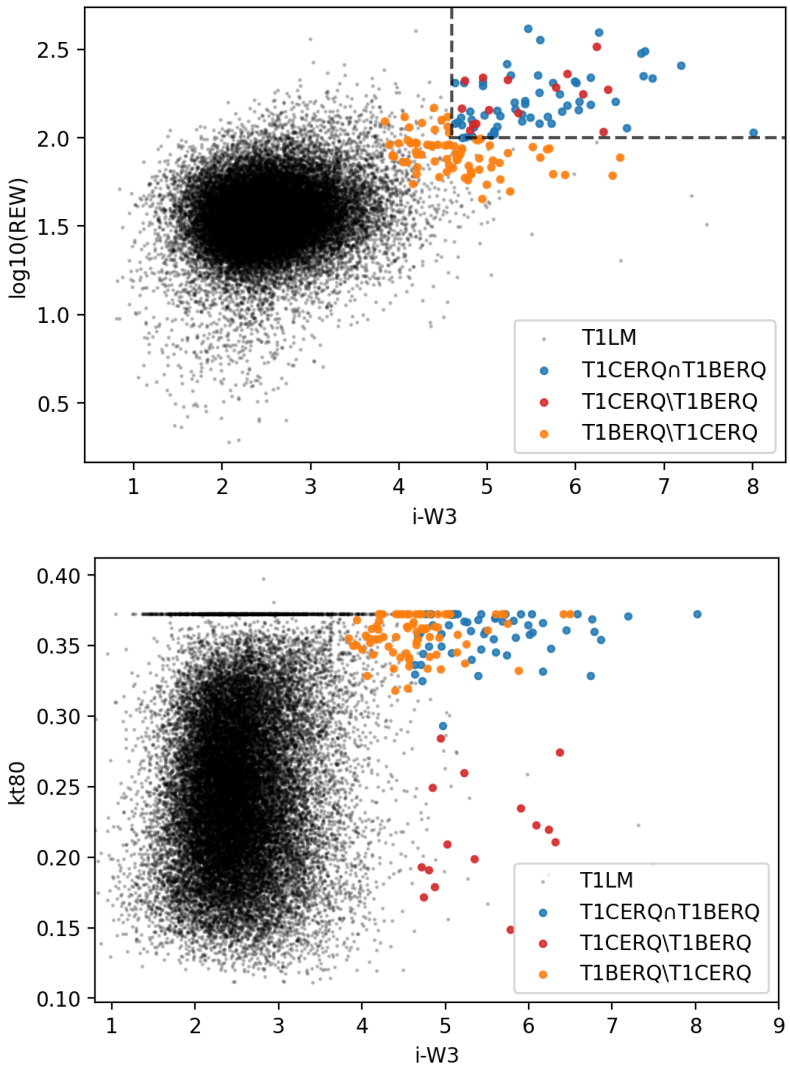

Figure 2.14: (Left) Projection of the 3D selection of T1BERQs into (i-W3, $\log_{10}$(REW(CIV)) space. (Right) Projection of the 3D selection of T1BERQs into (i-W3, kt_{80}(CIV)) space. Blue dots belong to the intersection of T1CERQs and T1BERQs. Red dots are T1CERQs which are not T1BERQs. Orange dots are T1BERQs which are not T1CERQs. T1LMs which are not T1CERQS or T1BERQs are shown by black dots.

inequalities in the 3D *normalised* parameter space:

$$0.64(i - W3)+0.41 \log_{10} REW + 0.64 kt_{80}(C\textsc{iv}) - 1.10 \geq 0. \quad (2.7)$$

$$\theta \leq 19.6^{\circ}, \quad (2.8)$$

where θ is defined in Eq. 2.2.

Figure 2.14 visualises the resulting set of quasars, T1BERQs, in 2D projections of the 3D space. Quasars are colour coded to show those which would be selected by both the T1CERQ and the T1BERQ criteria, by one but not the other, or by neither. Quasars selected by T1CERQ but not T1BERQ (15 quasars) are those with low kt_{80}: they are red and possess strong but not boxy CIV lines. Quasars selected by T1BERQ but not T1CERQ (76 quasars) are those which our local outlier factor selection algorithm judged to be closer to the ERQ subset than the main quasar locus. They are generally somewhat less red than the other ERQs and have weaker CIV lines, but exhibit the same extreme line properties. Overall, the T1BERQ selection produces 133 quasars.

Figure 2.15 compares the median spectra of T1BERQs to T1CERQs. As expected given the selection criterion, the T1BERQ sets have higher average kt_{80}, but lower average $i - W3$ and lower REW(CIV). However, they also exhibit the other unusual line properties associated with T1CERQs, to a stronger extent. In particular, the 76 quasars in T1BERQs but not T1CERQs have a high BAL fraction of $f_{BAL} = 0.62$, roughly three times larger than T1CERQs (see Fig. 2.16 for a clearer comparison). The low kt_{80} quasars which were removed were also those with the lowest BAL fraction. The FWHM of the CIV line is larger in the newly selected T1BERQs, strengthening the general trend shown in Table 2.3. Finally the NV line is strong, as shown by the high NV/CIV, and visually in the median spectra, where NV strength is comparable to the Lyman-α emission line.

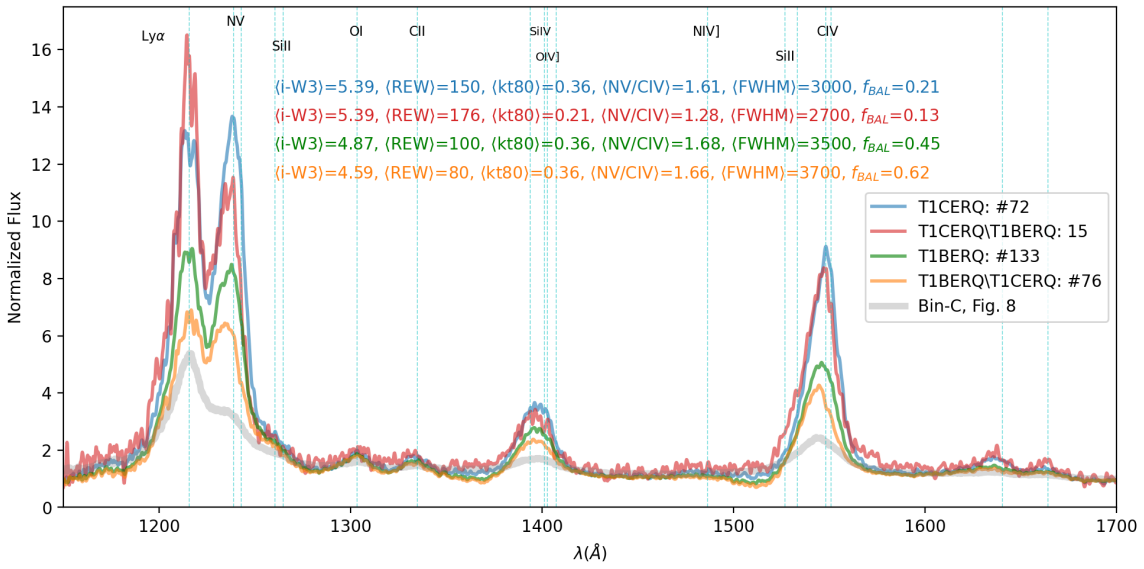

Figure 2.15: The median spectrum of T1CERQs is shown by the blue curve. The median spectrum of T1CERQs which are not among T1BERQs is plotted with the red curve. The median spectrum of those T1BERQs which are not among T1CERQs is shown in orange. The thick grey curve shows the median spectrum of all quasars in the dense region within bin C in Figure 2.10.

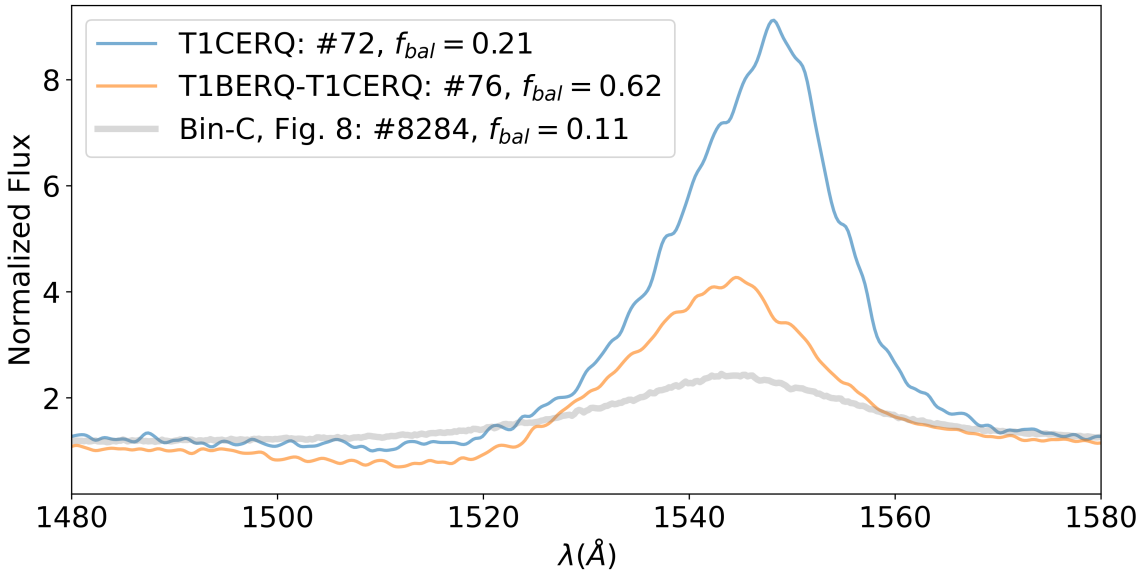

Figure 2.16: This figure shows a zoom in view of the median spectra around CIV BAL region. Our newly classified objects (i.e. T1BERQs which are not among T1CERQs) are compared to T1CERQs and to the quasars in the central bin of Fig. 2.10 and Fig. 2.11. Our 76 newly classified quasars have a higher visually verified BAL fraction.

2.5.5 T1BERQs in WISE AGN catalogue

To determine whether our sample of T1BERQs are extremely red only within the SDSS colour selection criteria or are extreme also as a part of other quasar samples, we performed a parallel analysis using the MILLIQUAS catalogue (Flesch 2021). We cross-matched the quasars in MILLIQUAS with the WISE AGN catalogue (Assef et al. 2018), as the infrared flux measurements of WISE are well-suited to studying red quasars like ERQs (H17). MILLIQUAS is a compendium of extant spectra with a high likelihood of being quasars. As SDSS is the largest spectral survey in existence, most, but not all, spectra in MILLIQUAS come from SDSS. If the colour selection function of SDSS were truncating the ERQ distribution, we would expect that the set of objects in MILLIQUAS but not in SDSS would extend substantially further towards the locus of ERQs in WISE colour space.

Figure 2.17 shows our selected sample of T1BERQs in the $(w1 - w2, w2 - w3)$ colour-colour space[8] of the WISE catalogue. We also show our comparison sample (i.e. crossmatched WISE AGN and MILLIQUAS quasars that have spectroscopic redshift but are not listed in SDSS). MILLIQUAS does extend the quasar locus moderately towards low $w2-w3$, but this is in the opposite direction to the T1BERQs. There is no evidence that colour selection effects are skewing our sample. We also show the histogram of T1BERQs in colour-colour space, as well as the histogram of the MILLIQUAS catalogue. These two histograms clearly have separate centers, indicating that even though T1BERQs are originally selected in the SDSS, they are extreme even in the MILLIQUAS catalogue *after excluding SDSS quasars.*

[8]Unfortunately, the $i-$band magnitudes for the quasars in our comparison sample are not available; otherwise it would be very illustrative to compare T1BERQs with our comparison sample in a 3D colour-colour-colour plot of $(i - w3, w1 - w2, w2 - w3)$.

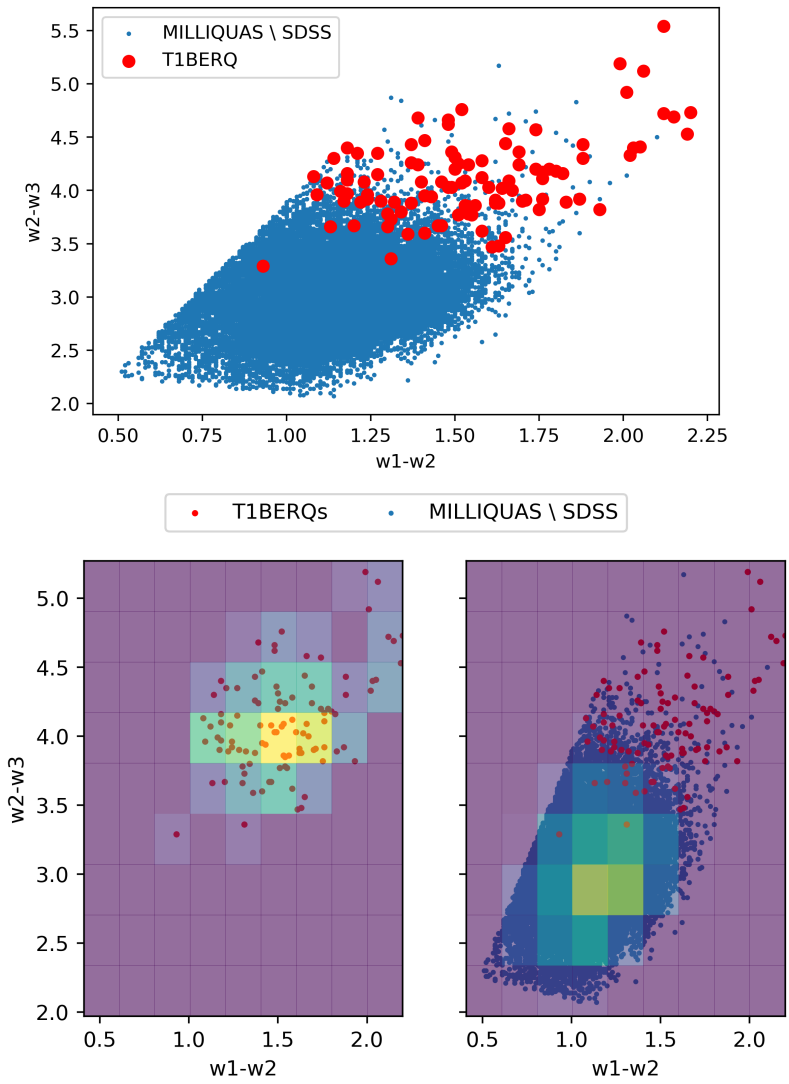

Figure 2.17: (Upper): blue dots are all quasars in the MILLIQUAS catalogue, but not in the SDSS catalog, with a spectroscopic redshift. Red dots are T1BERQs. (lower left) 2D histogram of T1BERQs in this colour space. (Lower right): 2D histogram of all quasars in the upper panel, with T1BERQs shown as red dots.

2.6 Conclusions

We have studied the phenomenon of extremely red quasars (ERQs), found by Hamann et al. (2016) (H17) to have red colour, large REW(CIV) and unusual emission line properties, including boxy CIV line profiles a high incidence of blue-shifted BALs, and high [OIII] 5007Å outflow speeds up to 6702 km s^{-1} Perrotta et al. (2019). These properties are consistent with ERQs being consistent with an early dusty stage of quasar-galaxy evolution, where strong quasar-driven outflows provide important feedback to the host galaxies. In this paper, we have used data driven techniques to understand whether the ERQs, when mapped into spectral parameter space, represent a separate population and, if so, where the boundaries of this population lie.

We applied a kernel density estimation in the space of $i-W3$ colour and CIV rest equivalent width identified by H17 to show that the ERQs produce overdensities. We computed the local outlier factor, previously calibrated on mock data, to assess whether these overdensities could be explained as statistical fluctuations at large distance from the median sampled quasar. The signature of two separate populations, as we showed using mock data, is a dip in the local outlier factor at the boundary between the populations. In two dimensions there was a dip near the boundary of the ERQs, but it was not strong enough to rule out a statistical fluctuation. We therefore considered higher dimensionality space, adding a third parameter, kt_{80}, defined by H17 as a measurement of line kurtosis or boxiness, and correlated with the presence of an ERQ. In the three dimensional space of $i-W3$ colour, CIV rest equivalent width and kt_{80}, we found a strong dip in the local outlier factor around the boundaries of ERQs, with a cluster size of 100-150. The dip in the local outlier factor provides evidence that ERQs are connected to a distinct phase of quasar formation, rather than being

part of a smooth transition from normal blue quasars to the tail of colour-REW distribution towards redder colours and larger REW(CIV)s.

We refined the selection criteria for T1CERQs, resulting in a new sample of 'boxy' ERQs (T1BERQs). The idea behind these selection criteria is to use line emission properties to better align the boundary of the T1CERQ sample with the onset of the special phase of quasar formation that leads to these exotic quasar properties. To do this, we made use of the 'boxy' shape of the CIV line, defining a boundary in 3D which maximised the depth of the dip in the LOF score. Our final sample defined T1BERQs by the inequalities 2.7 and 2.8, which refer to the common region between a plane and a cone obtained by finding the largest minimum of an LOF score in a bin.

There are 15 quasars in the sample of T1CERQs which are not in T1BERQs. Despite having very red colour and extremely strong CIV lines, these quasars have much lower kt_{80}(CIV) compared to the average T1CERQ and are thus excluded on the basis of their non-boxy line profile. On the other hand, there are 76 quasars which are within the T1BERQ sample, but are not T1CERQs. Selected on the basis of their kt_{80}(CIV) as well as their red $i - W3$ and high REW(CIV), these quasars have more extreme spectral properties, exclusive of the selection criteria, than the T1LM sample. Our T1BERQ selection criteria produced NV lines which were strong compared to the CIV and Lyman-α, and CIV lines with a greater FWHM than expected for the quasars' colour. The T1BERQs also had a high BAL fraction of $f_{BAL} = 0.62$, roughly three times larger than T1CERQs. If ERQs are associated with an early dusty stage of quasar formation, we would expect strong metal lines and a high fraction of BAL, associated with a dense accretion disc. The final result of our paper is thus improved selection criteria which produce a purer sample of these interesting objects. This will help to identify ERQs more efficiently in up-coming large quasar surveys such as the Dark Energy

Spectroscopic Instrument (DESI) DESI Collaboration et al. (2016) or HETDEX Hill et al. (2008), and select the best targets for follow-up observations investigating quasar and galaxy evolution.

Data Availability

Our underlying quasar line catalogue is from Hamann et al. (2016) and is available as the Supplemental BOSS Emission Line Catalog[9]. Our analysis scripts and the sample of T1BERQs as a fits table are publicly available in `GitHub`[10].

Acknowledgements

We are extremely grateful to Fred Hamann for his important contribution to this paper. We are also grateful to Serena Perrotta, Marie Wingyee Lau, Ming-Feng Ho, and Jarred Gillette for their insightful comments and suggestions. The authors appreciate the constructive comments of the anonymous reviewer and we would lie to thank Joseph Mazzarella for his guidance about using NED/iPac data bases. SB was supported by NSF grant AST-1817256. RM thanks Fred Hamann for supporting him for part of this work from NSF grant AST-1911066.

[9] https://datadryad.org/stash/dataset/doi:10.6086/D1H59V

[10] https://github.com/rezamonadi/ExtremelyRedQuasars

Chapter 3

Paper II: CIV absorbers in SDSS DR12: detection with Gaussian processes[1]

Abstract We assemble the largest CIV absorption line catalogue to date, leveraging machine learning to remove the need for visual inspection. We also provide a probability to classify the reliability of the absorption system within a quasar spectrum. We used Gaussian processes to train a quasar continuum model to detect CIV absorbers. Our training set was a subsample of DR7 spectra that had no detectable CIV absorption in the previous largest (visually inspected) CIV absorption catalogue. We used Bayesian model selection to decide between our continuum model and our absorption-line models. Our catalogue provides maximum *a posteriori* values and credible intervals for CIV redshift, column density, and Doppler velocity dispersion. Using a random hold-out sample of 1301 spectra from all of the 26,030 investigated spectra in DR7 CIV catalogue, we validated our pipeline and obtained an area under the curve (AUC) of 0.87 for the true positive rate versus false positive rate

[1]This chapter contains a draft of an article that has been submitted for publication by Monthly Notices of the Royal Astronomical Society and is under review. (Monadi et al. 2023)

curve (aka Receiver Operator Characteristic curve). We found good purity and completeness values, both $\sim 80\%$, when a probability of $\sim 95\%$ is used as the threshold. We obtained similar CIV redshifts and rest equivalent widths with our pipeline compared to our training set. Applying our algorithm to 185,425 selected quasar spectra from SDSS DR12, we produce a catalogue of 113,775 CIV doublets with at least 95% probability. We detect CIV absorption systems with a redshift range of 1.37–5.1, including 33 systems with a redshift larger than 5 and 549 absorbers systems with more than 95% reliability and a rest equivalent widths greater than 2 Å. Our catalogue can guide high resolution follow-up observations and may be cross-matched with galaxy catalogues or other quasar absorption line catalogues to investigate the properties of the circumgalactic medium.

3.1 Introduction

Metals, elements heavier than helium, are formed in the hearts of massive stars and recycled into the interstellar medium (ISM) by supernovae and stellar winds. Ultimately some of these metals are transported into the circumgalactic medium (CGM) or even intergalactic medium (IGM). Measurements of the abundance of metals in the Universe over time allow us to study the cycling of baryons through galaxies and, thus, the formation and evolution of galaxies (Tumlinson et al. 2017; Péroux & Howk 2020).

Quasar absorption lines enable us to measure the abundance of elements and their ionization states within the intergalactic gas. Particularly useful is the CIV $\lambda\lambda 1548, 1550$ doublet. This doublet is caused by a strong transition of an abundant metal that redshifts into optical bands at $z \sim 1.5 - 5.2$, with lower redshifts observable in the UV (e.g., Cooksey et al. 2009; Shull et al. 2014; Hasan et al. 2022) and higher in the IR (e.g., Simcoe et al. 2011; Ryan-Weber et al. 2009; Becker

et al. 2009; Davies et al. 2023). The rest wavelengths of the CIV doublet make it detectable outside the HI Lyα forest. Moreover, CIV has an unsaturated doublet ratio of 2 : 1 for $W_{r,1548} : W_{r,1550}$, easing automated line detection methods (Churchill 2020). [2]

CIV is a resonance line doublet that is useful for studying many physical properties of the IGM and CGM over cosmic time. It has been extensively studied; here we will provide an abbreviated overview, and the interested reader is referred to (Péroux & Howk 2020, and references therin) for a more comprehensive review.

Studying statistical properties of CIV absorption systems, such as their rest equivalent width distribution, sheds light on all of the processes that contribute to the formation and propagation of this metal throughout the IGM and CGM (Songaila 2005; D'Odorico et al. 2010; Simcoe 2011; Hasan et al. 2020, 2022). The metalicity and enrichment history of the CGM was studied using the CIV/HI line ratio (Ellison et al. 2000). The ratio of CIV to other metal lines can constrain the ionization state of the IGM (Boksenberg & Sargent 2015). Also the ratios of different carbon ions (CII/CIV) can be used to infer the ionization state of the absorbing gas in the IGM at a redshift where neutral hydrogen absorption is saturated (Cooper et al. 2019).

One can measure or constrain the temperature and kinematics of CIV absorbers to analyze the physics of the IGM (Rauch et al. 1996; Appleby et al. 2023). The study of the characteristics of metal lines, such as CIV, offers valuable information for developing models of contamination in baryon acoustic oscillation measurements of the Lyman-α forest (Yang et al. 2022). The auto-correlation (clustering) of CIV absorbers systems will constrain the IGM metallicity and enrichment topology (Chen et al. 2001; Scannapieco et al. 2006; Tie et al. 2022). Close quasar-galaxy pairs connect CIV absorbers to galactic halos and provide a tool for studying galaxy evolution (Adelberger

[2] W_r stands for the rest equivalent width of the absorption line.

et al. 2005; Bordoloi et al. 2014; Rubin et al. 2015; Burchett et al. 2015, 2016). Also, CIV absorbers have been observed at $z > 5$, probing the tail end of the reionization epoch (Becker et al. 2009; Ryan-Weber et al. 2009; Simcoe et al. 2011; D'Odorico et al. 2013; Codoreanu et al. 2018; Doughty & Finlator 2023).

Most relevant to our current work, Cooksey et al. (2013) detected strong CIV absorbers in the low signal-to-noise spectra of the Sloan Digital Sky Survey (SDSS) (Adelman-McCarthy et al. 2008; Eisenstein et al. 2011b). [3] On the theory side, CIV has been associated with enriched gas surrounding galactic halos in cosmological simulations (Haehnelt et al. 1996; Bird et al. 2016).

The above surveys and catalogues of CIV were assembled by visual inspection of quasar spectra by trained astronomers, sometimes supplemented by template fitting to discover candidate absorbers. However, this visual inspection is prohibitively time-consuming with the large size of modern quasar surveys. The largest CIV catalogues are from SDSS: Cooksey et al. (2013) used Data Release (DR) 7 and Chen et al. (2014) used DR9. The visually inspected quasar catalogue of SDSS DR12 contains 185,541 quasars (Ross et al. 2012), which can potentially have CIV absorbers. The upcoming Dark Energy Spectroscopic Instrument (DESI DESI Collaboration et al. 2016) will obtain spectra for more than 30 million galaxies and quasars. DESI will observe more than ten times the number of galaxies observed by SDSS and $\sim 10^7$ quasars. Leveraging the increase in quasar spectra for CIV studies is best served by an automated detection algorithm. SDSS DR12 contains the largest extant quasar spectral catalogue with *visually verified redshifts*. However, it has a relatively low spectroscopic resolution and a low median signal-to-noise ratio (SNR). This makes the detection of

[3] Chen et al. (2014) assembled a CIV catalogue from SDSS DR9 quasar spectra. However, we did not use their candidate absorbers as a detailed comparison to previous CIV catalogues was missing.

an absorption line, like the CIV doublet, quite challenging. However, our Bayesian approach based on Gaussian processes is capable of extracting reliable information even from noisy data.

Our automated CIV detection pipeline is based on the technique for detecting Damped Lyman-α absorbers (DLAs) from Garnett et al. (2017), which was extended to multiple absorbers by Ho et al. (2020). A Gaussian process model with a bespoke learned kernel is built for the quasar spectrum in the absence of absorption, and Bayesian model selection is used to determine whether an absorber is preferred over the no-absorption (i.e., continuum) model given the quasar instrumental noise. The pipeline is built using a Bayesian framework, allowing us to make probabilistic statements even about the noisiest observed data. Detection probabilities can be used to further refine the catalogue to increase purity or completeness. Furthermore, as a fully Bayesian pipeline, it provides a posterior distribution for the column density, redshift, and Doppler velocity dispersion for each absorber.

The rest of this paper is structured as follows. In Section 3.2, we summarise the data we used for different stages in our pipeline. In Section 3.3, we detail the mathematical framework for obtaining our absorption models, our Gaussian process model for quasar emission, and our Bayesian approach to search for absorbers in the quasar spectra. We validate our approach by testing our algorithm in a hold-out sub-sample of our training set in Section 3.4. The resulting CIV catalogue is presented and discussed in Section 3.5. We summarise and discuss potential future applications of our catalogue in Section 3.6.

3.2 Data

Our primary dataset was SDSS quasar spectra; we followed Cooksey et al. (2013) in designating quasars with their spectroscopic modified Julian date, fibre identification number, and plate number. We trained our absorption-free model on a subset of SDSS DR7 (Adelman-McCarthy et al. 2008) filtered to avoid CIV absorbers as detected by the so called "Precious Metals" (PM) catalogue (Cooksey et al. 2013).[4] The PM catalogue did not search for absorbers in spectra that did not meet certain criteria (see Table 1 in Cooksey et al. for more details). Excluded were spectra of a broad absorption line quasar, spectra with insufficient wavelength coverage for CIV, and spectra with low median SNR ($\langle S/N \rangle < 4\,\mathrm{pix}^{-1}$). Our training set is based on the PM CIV catalogue, so starting from SDSS DR7, we also exclude quasar spectra not searched by Cooksey et al. (2013). The initial DR7 quasar catalogue contains 105,783 quasars, of which 26,030 were searched for CIV absorption. Our training set further excluded the 10,861 spectra which contain one or more CIV absorbers in the PM catalogue. Our null model was thus trained on 15,169 "CIV-free" spectra, meaning spectra that either were not found as a CIV *candidate* (as defined by Cooksey et al. (2013)) or did not pass the *visual verification* check.

Before training a continuum model on all of the 15,169 spectra in C13, we train a number of candidate continuum models on 95% of our training set and then validate these candidate continuum models on a random hold-out sample of 5% of all searched spectra in the DR7 catalogue, which contains 1301 quasars. This is our *validation set* that we used as a tool to find the optimum values of the parameters needed to train a candidate continuum model. These tuning parameters include: flux normalization wavelength range, the minimum number of non-NaN pixels in a training spectrum,

[4] We obtained the list of spectra from `igmabsorbers.info` and downloaded the spectra from http://das.sdss.org/spectro/1d_26

the dimension of the covariance matrix (see Equation 3.10), etc. After applying our pipeline on the validation set, we assessed the performance of the classification (i.e. classifying a given spectrum as having CIV absorber(s) or otherwise) using the PM catalogue as a "ground truth". We found the best candidate continuum model by maximising the classification score (see Section 3.4.2) and purity/completeness (see Section 3.4.3). At this point, we took the parameters of the best candidate continuum model and built our final model from all of the 15,169 "CIV-free" DR7 spectra investigated in the PM catalogue pipeline.

We applied our algorithm on a subset of the SDSS DR12 quasar catalogue (Alam et al. 2015) to build our new CIV catalogue. We chose our working quasar sample starting from the SDSS-DR12 quasar catalogue.[5] We kept only quasars with rest-frame wavelength coverage between 1310 Å–1548 Å, the region of potential CIV absorption (avoiding both the Lyα forest and the potential for false positives of CIV from OIλ1302 or SiIIλ1304). This means quasars with redshifts satisfying 1310 Å(1+$z_{\rm QSO}$)> 3650 Å (or $z_{\rm QSO} > 1.7$) and 1548 Å(1+$z_{\rm QSO}$) < 10400 Å (or $z_{\rm QSO} < 5.7$). We removed detected broad absorption line quasars (BAL) using the SDSS BAL catalogue[6]. After these selections, we downloaded the list of quasar spectra from the SDSS-III Baryon Oscillation Spectroscopic Survey Science Archive Server[7].

We converted all observed spectra to the emission rest-frame using the visually inspected quasar redshift estimate from the SDSS pipeline, which we assume to be exact.[8] Missing or otherwise masked flux values (e.g., from a bad pixel) were denoted by `NaN` and were not used in our pipeline.

[5] http://data.sdss3.org/sas/dr12/boss/qso/DR12Q/DR12Q.fits

[6] http://data.sdss3.org/sas/dr12/boss/qso/DR12Q/DR12Q_BAL.fits

[7] https://data.sdss.org/sas/dr12/boss/spectro/redux/

[8] We used Z_VI, column 8 of SDSS DR12 quasar catalogue.

3.3 Method

We modified the pipeline introduced in Garnett et al. (2017) and Ho et al. (2020) to look for CIV absorbers in SDSS DR12. We learnt an *a priori* distribution for the shape of the quasar emission spectra without CIV using SDSS DR7 spectra classified by the PM catalogue from Cooksey et al. (2013). The null model, M_N, was learned from SDSS DR7 spectra identified as 'non-detection' (i.e., no CIV *candidate* in the PM study). Each iteration, we did a Bayesian model selection between the null model, a model for a CIV doublet model (M_D), and a model for an 'interloper' singlet absorption line (M_S) to compute the posterior probability of CIV absorption. We searched for up to seven CIV absorbers in each spectrum, reporting probabilities for each. There were six main changes since Ho et al. (2020). First, the absorption profile was updated to model a CIV doublet, instead of a DLA. Second, a model for singlet line absorbers was introduced, which serves a similar role to the sub-DLA model in Ho et al. (2020). Without this singlet absorption line model, the pipeline produces excessive false positives, as it has no other way to match absorption except a CIV doublet. Third, in addition to sampling absorber redshift and column density, we sampled the Doppler velocity dispersion, which allows more accurate fits. Forth, we no longer model the Lyman-α forest in the null model, as it does not overlap our CIV absorption region. Fifth, instead of a fixed instrumental broadening profile, we used the reported Gaussian wavelength-dependent dispersion from SDSS.[9] Sixth, we report an individualised probability for each absorber we detect, rather than the joint probability of observing at least a certain number of absorbers, as in Ho et al. (2020). Figure 3.1 shows an overview of our pipeline, as described in this Section.

[9] Column 6 of the fits files of SDSS spectra, see the SDSS data model.

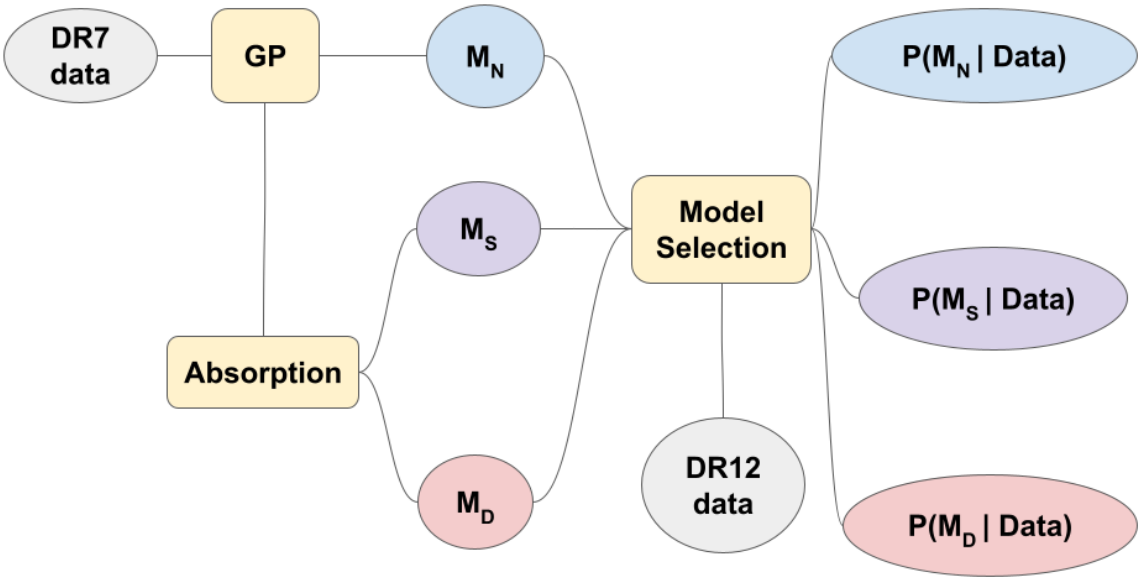

Figure 3.1: This is a flow chart for our pipeline. Training spectra from SDSS DR7 are used to train a Gaussian process kernel with which to model the quasar continuum (i.e., null model, M_N). Analytic Voigt profiles are used to construct models for absorption from a CIV doublet (M_D) or a generic singlet absorber (M_S). Conditioning on DR12 spectra produces a posterior probability estimate for each model that can be used to decide if there is a CIV absorber in the given spectrum or not. Moreover, for the absorber models, M_D and M_S, we have a posterior distribution for each model parameter: absorber redshift, Doppler velocity dispersion for the absorption profile, and the absorber column density.

Section 3.2 described our initial training data, a subset of SDSS DR7. Section 3.3.1 explains the Voigt-profile model for any absorber detection. Section 3.3.2 summarises our null (aka absorption-free or continuum) model, M_N, for the quasar emission function, which uses a bespoke Gaussian Process kernel. Section 3.3.3 describe two analytic absorption models, M_S and M_D, which are generated by convolving M_N with a singlet or doublet Voigt profile, respectively. In addition, we need model priors, $Pr(M)$, for each model (see Section 3.3.4). The model likelihood is discussed in Section 3.3.5. Section 3.3.6 explains our technique for deciding how many CIV absorbers to search for.

3.3.1 Absorption function

Voigt profiles are useful for modelling the absorption effect in the observed spectrum of an emitting source such as quasars (Churchill 2020). A Voigt profile is given by:

$$\phi(\mathrm{v};\sigma_{\mathrm{CIV}},\gamma_{\ell\mathrm{u}}) = \int \frac{d\mathrm{v}}{\sqrt{2\pi}\sigma_{\mathrm{CIV}}} \exp\left(-\mathrm{v}^2/2\sigma_{\mathrm{CIV}}^2\right) \frac{4\gamma_{\ell u}}{16\pi^2[\nu-(1-\mathrm{v}/c)\nu_{\ell u}]^2+\gamma_{\ell u}^2}, \tag{3.1}$$

which is a convolution between Lorentzian and Gaussian profiles. The former computes the natural broadening and the latter thermal broadening (Draine 2011). The velocity, v, in Equation 3.1 is given by:

$$\mathrm{v}(\lambda) = \mathrm{c}\left(\frac{\lambda}{\lambda_{\ell\mathrm{u}}(1+\mathrm{z}_{\mathrm{CIV}})} - 1\right). \tag{3.2}$$

A negative (positive) velocity refers to a position in λ-space that is red-ward (blue-ward) of the observed CIV absorption in $\lambda_{\ell u}(1+z_{\mathrm{CIV}})$. The Lorentzian broadening contribution is:

$$\gamma_{\ell u} = \frac{\Gamma\lambda_{\ell u}}{4\pi}, \tag{3.3}$$

where Γ is the damping constant. The Doppler velocity dispersion for a CIV absorber, σ_{CIV}, is:

$$\sigma_{\mathrm{CIV}} = \sqrt{\frac{kT}{6m_p+6m_n}}, \tag{3.4}$$

where k, T, m_p and m_n are the Boltzmann constant, gas temperature, proton mass, and neutron mass, respectively. The Doppler velocity dispersion controls the width of the absorption profile as a function of temperature. For the CIV doublet at $\lambda = 1548$ Å, $\Gamma = 2.643\times10^8$ s^{-1} and for $\lambda = 1550$ Å, $\Gamma = 2.628\times10^8$ s^{-1}. Lorentzian broadening is thus small ($\gamma_{\ell u}/\sigma_{\mathrm{CIV}} \sim 0.01$ for $T \sim 10^4$K) and the Voigt profile is close to Gaussian.

The optical depth, τ, itself is a function of observed frequency ($\nu = c/\lambda$) given: absorber column density $N_{\rm CIV}$ which controls the depth of the profile, absorber redshift $z_{\rm CIV}$ which sets the wavelength where we observe the absorption, and Doppler velocity dispersion $\sigma_{\rm CIV}$. The optical depth is given by:

$$\tau_{\ell u}(\lambda; z_{\rm CIV}, N_{\rm CIV}, \sigma_{\rm CIV}) = \frac{N_{\rm CIV}\pi e^2 f_{\ell u}\lambda_{\ell u}}{m_e c}\phi(\mathrm{v}(\lambda), \sigma_{\rm CIV}, \gamma), \tag{3.5}$$

where c is the speed of light, e is the elementary charge, m_e is the mass of the electron and $\lambda_{\ell u}$ is the transition wavelength for the lower state (ℓ) and the upper state (u) and $f_{\ell u}$ is the oscillator strength of the transition. Using spectroscopic notation (Tennyson 2019), the 1548 Å absorption line is a transition from $2^2S_{\frac{1}{2}}$ to $2^2P^o_{\frac{1}{2}}$ and the 1550 Å absorption line is a transition from $2^2S_{\frac{1}{2}}$ to $2^2P^o_{\frac{3}{2}}$. The absorption profile is related to the optical depth via:

$$a_{\ell u}(\lambda; z_{\rm CIV}, N_{\rm CIV}, \sigma_{\rm CIV}) = \exp\left(-\tau_{\ell u}(\lambda; z_{\rm CIV}, N_{\rm CIV}, \sigma_{\rm CIV})\right), \tag{3.6}$$

where the ℓu subscript can refer to either 1548 Å or 1550 Å transitions. The doublet model $\mathrm{M_D}$ will be built by convolving the null model with an absorption profile that considers both 1548 Å or 1550 Å. The singlet model $\mathrm{M_S}$, on the other hand, only considers the 1548 Å transition.

SDSS resolution is insufficient for detailed modelling of CIV absorption systems as is done with high-resolution spectra (e.g. Hasan et al. 2020). Indeed, strong CIV absorption at SDSS resolution can be reasonably modelled by a single Voigt profile with appropriate choice of $z_{\rm CIV}$, $N_{\rm CIV}$, and $\sigma_{\rm CIV}$, as we do in this work (see Section 3.3.5). We acknowledge that the same absorption at higher resolution would reveal finer structure and require multiple Voigt profiles, with different combinations of $z_{\rm CIV}$, $N_{\rm CIV}$, and $\sigma_{\rm CIV}$ that would be strong constraints on the physical conditions of the gas giving rise to the absorption. The $N_{\rm CIV}$ and $\sigma_{\rm CIV}$ values returned by our algorithm may

not be as tightly constrained as the z_{CIV} measurements. Remember that N_{CIV} and σ_{CIV} control the Voigt profile shape in our absorption model that is compared to the observed flux deficit in the SDSS spectra (see Table 3.1).

3.3.2 Quasar emission function

The physics of quasar emission is not fully understood, and there is considerable variety in observed quasar spectra. Thus we used an empirical model for the quasar emission function (aka continuum) based on the observed spectra. We modelled the emission function of a quasar, f, in the absence of any absorption (including CIV absorption) using *Gaussian processes* that generate a distribution over functions. Gaussian Processes result from a generalisation of a multivariate Gaussian distribution to infinite domains (Rasmussen C. E. 2006). As the standard library of kernels is insufficiently flexible to model the complicated correlations between different emission lines in a quasar spectrum, we used a customised kernel learned directly from the training set.[10] We described the training set in Section 3.2.

Here we briefly summarise the technique. Our model was similar to Garnett et al. (2017) where the process of obtaining a Gaussian process model for quasar emission spectra is described in more detail. However, unlike Garnett et al. (2017), we did not model the Lyman-α forest as we were looking for CIV absorbers outside of the forest. We trained a CIV-free model between 1310 Å and 1555 Å, which produced the best results during the validation phase. This range is close to the rest frame CIV absorption wavelength searched in the PM catalogue. Figure 3.2 shows an example learned quasar continuum together with the observed flux and noise.

[10] Our training set consisted of all of the spectra investigated in the PM CIV catalogue and classified as not containing CIV absorbers.

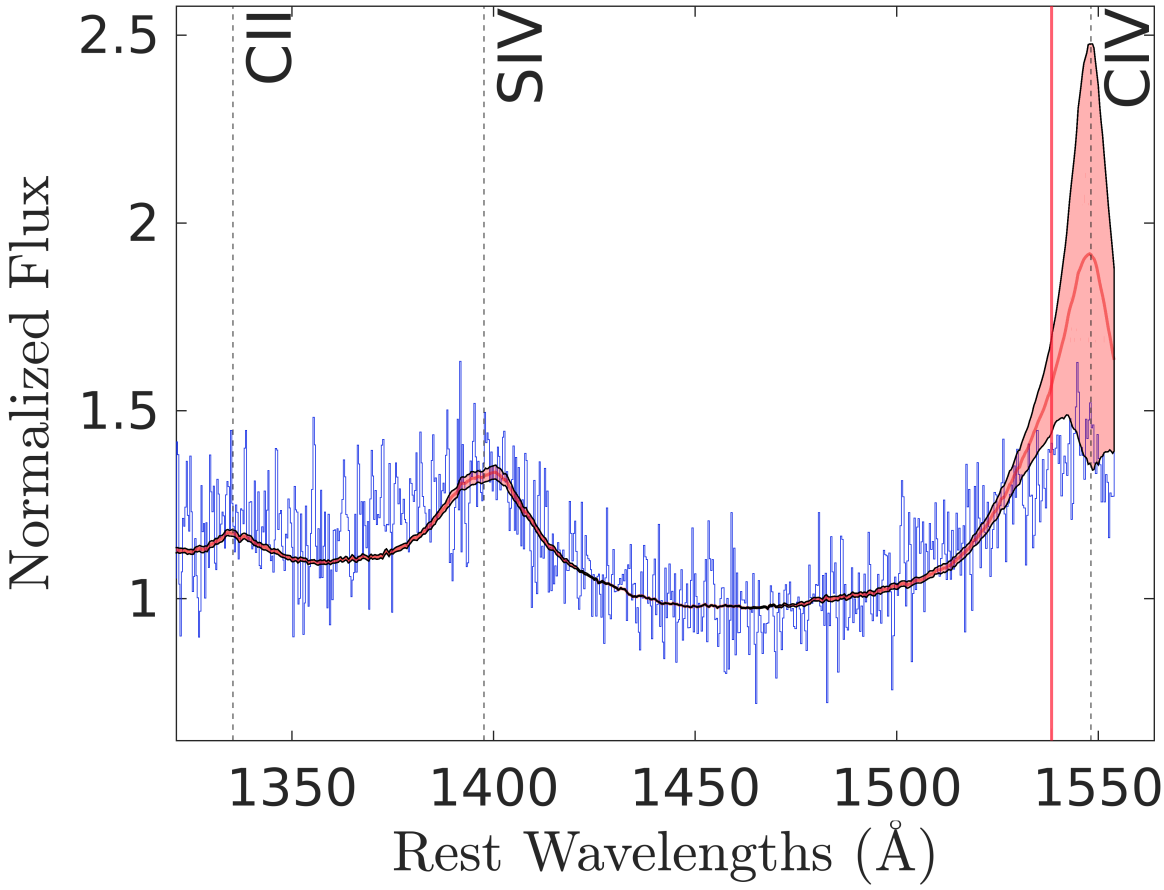

Figure 3.2: An example learned quasar emission function (red curve) with the normalised observed smoothed flux (blue curve). The shaded red region shows 1σ uncertainties. The SDSS DR7 quasar has QSO-ID: 51630-0266-280 and redshift 2.57. Note that we search for absorbers starting 3000 km s^{-1} red-ward of the quasar's redshift (shown by the solid red vertical line), so the moderate failure to match the quasar CIV emission line in this case does not lead to an artificial preference for CIV absorption. Prominent emission lines are marked by dashed vertical lines.

Even from low SNR spectra, our method extracts some statistical information, so we do not enforce a minimum SNR in our search. Our pipeline naturally gives low likelihoods to low-SNR spectra during the training. We can completely specify a Gaussian process by its mean and correlation functions (analogous to the first two moments of a Gaussian distribution). We specify the mean function μ by:

$$\mu(\lambda) = \langle y(\lambda) \rangle, \tag{3.7}$$

where λ is the rest-frame wavelength and $y(\lambda)$ is the observed rest-frame flux for the training-set spectra, after applying a mask for missing pixels; angle brackets ($\langle\rangle$) denote an average over wavelengths. Before computing this average over the training set, we have normalised the quasar flux and the flux variance so that they have a median value of unity in the normalisation range. This normalisation was needed so that the Gaussian process model is insensitive to variations in (observed) quasar brightness. We chose the range from 1420 Å to 1475 Å as it contains no prominent emission lines (Zhu & Ménard 2013; Hamann et al. 2016; Monadi & Bird 2022b). We also confirmed empirically that this normalization range produces the best score when applied to our validation set. We remind the reader that the validation set is a random subset (1301 spectra) of all candidate DR7 spectra in the PM catalogue (see Section 3.2).

The Gaussian process covariance function describes the correlation between flux values at two separate wavelengths, λ and λ'. Most applications of Gaussian processes assume a simple kernel for the covariance, such as the exponential squared kernel (Rasmussen C. E. 2006). However, the complex correlation between features in quasar continua is hard to describe using the simple/standard covariance functions like the radial basis function. Instead, our algorithm directly

learned a covariance function:

$$K(\lambda, \lambda') = \mathrm{cov}[f(\lambda), f(\lambda')], \tag{3.8}$$

by considering all of the cumulative information contained in the spectra of our training set: all of the flux measurements and noise measurements given at the observed wavelengths.[11] We need to maximise the joint likelihood of generating the whole training set given that the underlining model is the null model (i.e. absorption-free). We assume our observations (i.e. flux and noise given at each observed wavelength in the training set) are independent and drawn from a Gaussian distribution with width corresponding to the observed noise of the SDSS pipeline. Next we maximise the likelihood (see section 5.3 of (Garnett et al. 2017) for details) and learn the quasar mean function (Equation 3.7) and quasar covariance function (Equation 3.8). Optimising this joint likelihood function was done using `minFunc`: a Matlab function for unconstrained optimization of differentiable real-valued multivariate functions using line-search methods.[12]

We binned quasar spectra linearly in wavelength, from $1310-1555$ Å, with a bin size of $\Delta\lambda$. This gave us the number of bins as:

$$N_{\mathrm{bin}} = \frac{1555 - 1310}{\Delta\lambda}. \tag{3.9}$$

If we input the binned wavelength grid, $\boldsymbol{\lambda}$, to Equation 3.7 we get the learned mean vector $\boldsymbol{\mu}$, with N_{bin} elements. The covariance matrix, $\mathbf{K}$, an $N_{\mathrm{bin}} \times N_{\mathrm{bin}}$ matrix, is calculated on two discretized wavelength grids, $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda'}$, using Equation 3.8. A very fine $\Delta\lambda$ is not desirable because it increases the size of $\boldsymbol{\mu}$ and $\mathbf{K}$ and thus is more computationally expensive. On the other hand a coarse $\Delta\lambda$ cannot capture enough information from the quasar spectra. The optimum $\Delta\lambda$ in Garnett et al. (2017)

[11] The third column of the SDSS fits tables for observed spectra contains inverse noise variance $(\sigma(\lambda))^{-2}$.

[12] https://www.cs.ubc.ca/ schmidtm/Software/minFunc.htm

and Ho et al. (2020) was 0.25 Å. We empirically found that $\Delta\lambda = 0.5$ Å is the optimum value for the redder spectral region we examine here which gives us $N_{\rm bin} = 490$. Without further structural assumptions on $\mathbf{K}$, our algorithm would have to learn a matrix of $N_{\rm bin}^2 \sim 2.4 \times 10^5$ elements. To circumvent this, we used a low rank decomposition:

$$\mathbf{K} = \mathbf{M}\mathbf{M}^\top, \tag{3.10}$$

where $\mathbf{M}$ is a $N_{\rm bin} \times k$ matrix, for any positive integer k. Larger-k models allow for higher fidelity modelling of $\mathbf{K}$. Following Garnett et al. (2017), we set $k = 20$. We also checked $k = 19$, 21, and 22, finding that our results were insensitive to this choice. Figure 3.3 shows the learned covariance matrix. This covariance matrix describes how likely the quasar emission spectrum is to vary around the mean spectrum. It encodes the information contained in the spectra of our training set, the "CIV-free" spectra from the PM catalogue.

Having learned the mean quasar vector $\boldsymbol{\mu} = \mu(\boldsymbol{\lambda})$ (see Equation 3.7) and the lower rank decomposition matrix $\mathbf{M}$ in Equation 3.10 which gives us the covariance matrix $\mathbf{K}$ (Equation 3.8), we can write the Gaussian processes model for the quasar emission function, trained on the observed spectra, as a multivariate Gaussian distribution:

$$p(f(\boldsymbol{\lambda})) = \mathcal{GP}(\mu(\boldsymbol{\lambda}), K(\boldsymbol{\lambda}, \boldsymbol{\lambda}')) = \mathcal{N}(f(\boldsymbol{\lambda}); \mu(\boldsymbol{\lambda}), K(\boldsymbol{\lambda}, \boldsymbol{\lambda}')), \tag{3.11}$$

where $\mathcal{GP}$ denotes a Gaussian process. We remind the reader that a Gaussian process is a Gaussian distribution over functions. Therefore, we can write the Gaussian process for the quasar emission function f given our learned mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{K}$ as:

$$\mathcal{N}(f; \boldsymbol{\mu}, \mathbf{K}) = \frac{1}{\sqrt{(2\pi)^d \det(\mathbf{K})}} \exp\Big(-\frac{1}{2}(f - \boldsymbol{\mu})^\top \mathbf{K}^{-1} (f - \boldsymbol{\mu})\Big), \tag{3.12}$$

where d is the dimension of the quasar emission function f.

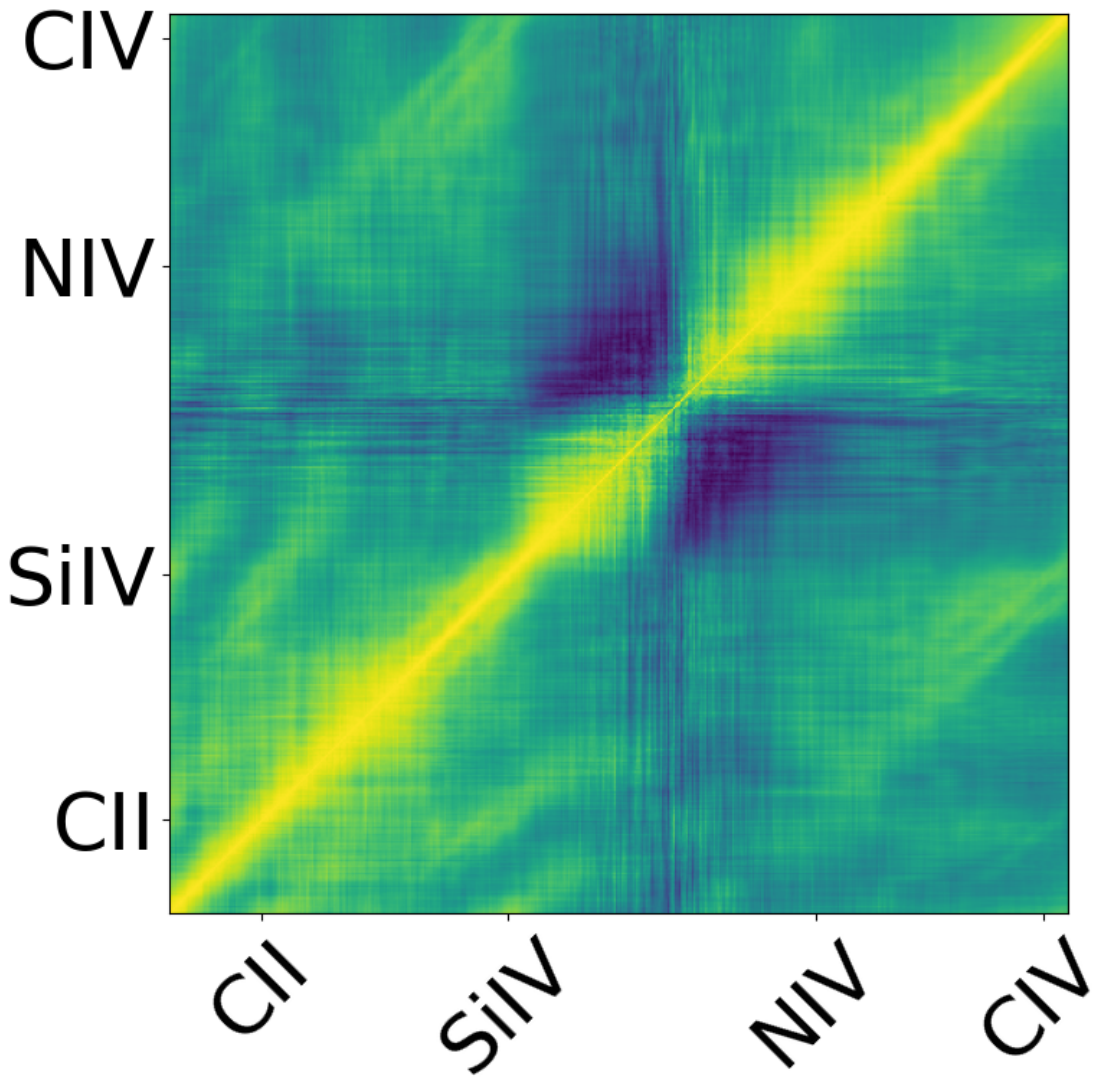

Figure 3.3: Learned covariance matrix **K** (see Equation 3.8 and Equation 3.10) for our null (continuum) model. This matrix is built up by considering the observed flux and noise from our CIV-free training set (see Section 3.2). Brighter pixels show stronger correlations and darker regions weaker ones. The wavelengths of prominent emission lines are labelled. The bright diagonal implies stronger correlations between pixels at smaller wavelength separation.

3.3.3 Absorption line models

We want to find the probability of a CIV doublet in the observed spectrum of a quasar given the observed rest-frame flux $\boldsymbol{y}(\boldsymbol{\lambda})$, under our null (aka absorption-free or continuum) GP model $\mathrm{M_N}$. Our data were composed of the observed wavelengths $\boldsymbol{\lambda}$, their corresponding observed quasar flux $\boldsymbol{y}(\boldsymbol{\lambda})$, and their corresponding observed noise $\boldsymbol{\sigma}(\boldsymbol{\lambda})$. We define the data as:

$$\mathcal{D} = \{\boldsymbol{\lambda}; \boldsymbol{y}(\boldsymbol{\lambda}), \boldsymbol{\sigma}(\boldsymbol{\lambda})\}. \tag{3.13}$$

Bayes' rule gives the *model posterior*, the probability of each model given the data:

$$P(\mathrm{M_i}|\mathcal{D}) = \frac{\mathrm{P}(\mathcal{D}|\mathrm{M_i})\mathrm{Pr}(\mathrm{M_i})}{\sum_\mathrm{j} \mathrm{P}(\mathcal{D}|\mathrm{M_j})\mathrm{Pr}(\mathrm{M_j}),}. \tag{3.14}$$

We defined three models:

- $\mathrm{M_N}$ models the quasar continuum without absorption (Equation 3.11).
- $\mathrm{M_D}$ is a model containing exactly one CIV doublet. $\mathrm{M_D}$ is built by convolving $\mathrm{M_N}$ with the absorption function (Equation 3.6) for all observed wavelengths.

$$\mathrm{M_D} \rightarrow \mathrm{convolve}(\mathrm{a}_{1548,1550}(\boldsymbol{\lambda}), \mathrm{M_N}) \tag{3.15}$$

- $\mathrm{M_S}$ is a singlet model containing exactly one generic singlet absorption line. For simplicity, we implemented $\mathrm{M_S}$ using the same Voigt profile as $\mathrm{M_D}$ but including only the 1548Å absorption line.

$$\mathrm{M_S} \rightarrow \mathrm{convolve}(\mathrm{a}_{1548}(\boldsymbol{\lambda}), \mathrm{M_N}) \tag{3.16}$$

We added this singlet model, in addition to the CIV-free and CIV-doublet models, so that our Bayesian framework is not forced to give a high probability of a CIV doublet if there is a strong singlet line in

the spectrum and nearby noise happens to be similar to a CIV doublet. For example, a broad singlet line like SiII1526, FeIIλ1608, or AlIIλ1670, can be mis-identified as a CIV doublet, if we have only two models (i.e. M_N and M_D). The singlet model, M_S, provides an alternative to both M_N and M_D for such lines.

Figure 3.4 shows an example, the application of our pipeline to QSO-ID: 51608-0267-264 with $z_{QSO} = 1.89$. Here a noise fluctuation and a strong line happen to have a velocity separation similar to a CIV doublet. For this spectrum, we have:

$$\begin{aligned} \log(P(M_N|\mathcal{D})) &= -297.8216, \\ \log(P(M_S|\mathcal{D})) &= -250.1609, \text{ and} \\ \log(P(M_D|\mathcal{D})) &= -257.1906. \end{aligned}$$

Although the doublet model is not a very good fit, the null model is even worse. Thus without the singlet model, M_S, our pipeline would incorrectly prefer the doublet model and detect a CIV absorber.

A sampling problem arises due to the low resolution of the SDSS spectrograph. Real spectrographs measure the total integrated flux across the spectral pixel. A simple estimate for this is to evaluate the Voigt profile at the center of the pixel. However, at the low resolution of the SDSS spectrograph, this can be a poor estimate, leading to unphysical doublet ratios. For this reason we compute the integrated flux by first evaluating the Voigt profile on a grid of pixels which is finer than the grid in the SDSS spectrum by a factor of n_{ave}. We found by experiment that the model accuracy does not improve for $n_{ave} > 20$ sub-samples.

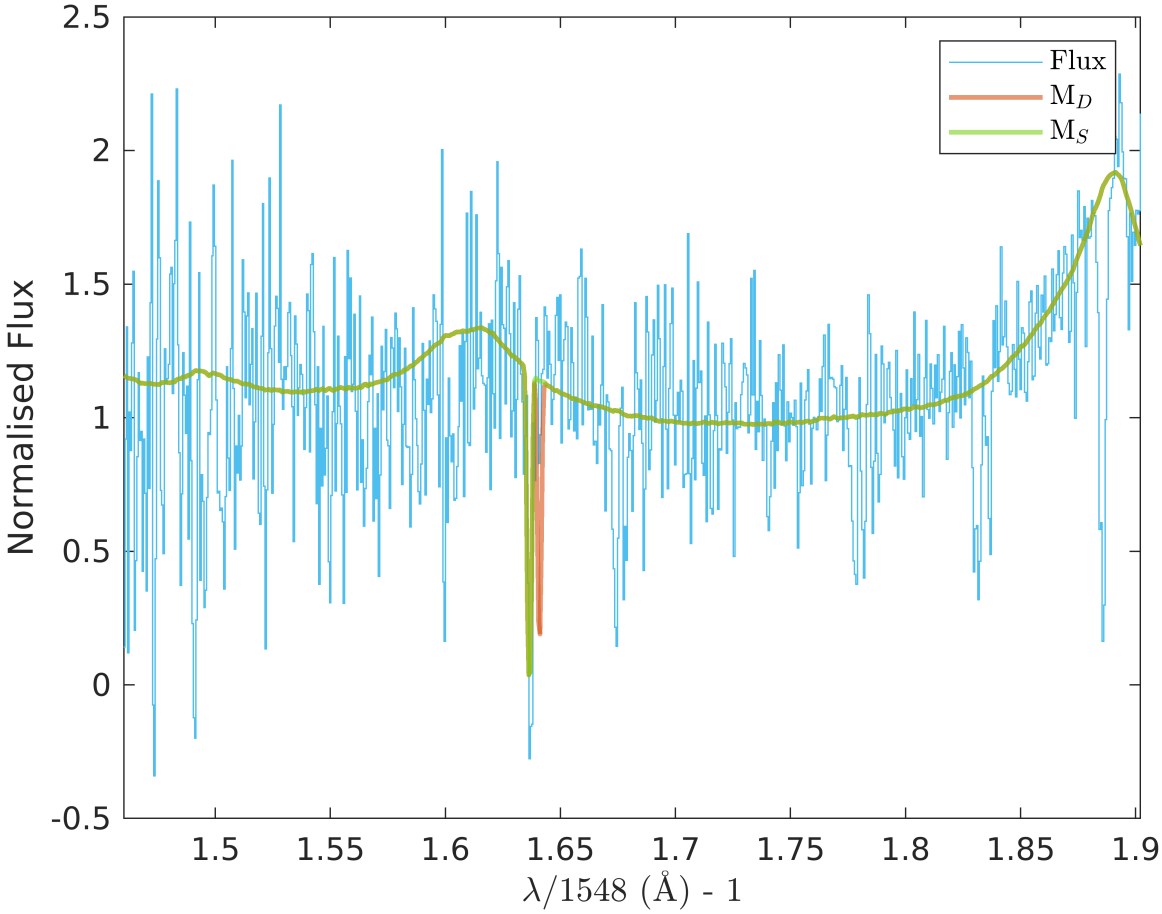

Figure 3.4: The figure shows the spectrum of QSO-ID: 51608-0267-264 with $z_{\rm QSO}$=1.89 (blue) where the singlet model (green) is preferred over the CIV doublet model (red), which is in turn preferred over the null model. If we did not have M_S, our pipeline would have incorrectly detected a CIV absorber at $z_{\rm CIV}$ = 1.635.

3.3.4 Model priors

To calculate the model posterior (Equation 3.14), we need model priors, $Pr(\mathrm{M})$, for each of the three models. We set priors for the CIV doublet, $\mathrm{M_D}$, using population statistics from our training set, the PM catalogue of Cooksey et al. (2013). We counted the fraction of spectra with absorbers at $z_{CIV} < z_{\mathrm{QSO}} - 30000/c$, where c is the speed of light in km s^{-1}. This small decrease in our upper limit for the absorption redshift accounts for any possible error in estimating the redshift from the SDSS pipeline. For simplicity, we used the same prior for the singlet and doublet line models, i.e., $Pr(\mathrm{M_S}) = \mathrm{Pr}(\mathrm{M_D})$. There are no single-line catalogues for these data, and using equal priors ensures that whichever model is the best-fit will be used.

The prior for the CIV-free model can be obtained by:

$$Pr(\mathrm{M_N}(\mathrm{k\ CIV})) = 1 - \mathrm{Pr}(\mathrm{M_D}(\mathrm{k\ CIV})), \tag{3.17}$$

where "k *CIV*" denotes some integer number k of CIV systems. We did not include $Pr(\mathrm{M_S})$ in Equation 3.17 to enable a pointwise model comparison between $\mathrm{M_D}$, $\mathrm{M_S}$, and $\mathrm{M_N}$. Especially when searching for multiple absorbers, our main purpose is deciding the probability of detection or non-detection of CIV absorbers in a spectrum. Furthermore, the small shift in the normalization of model priors is several orders of magnitude smaller than the effect of normalising the model posteriors in Equation 3.14.

When searching for additional absorbers in spectra where there is already a detection, we use the prior probability of spectra with $(k-1)$ CIV absorbers having k absorbers:

$$\begin{aligned} Pr(k\ C\textsc{iv}) &= P(k\ C\textsc{iv}|(k-1)\ C\textsc{iv}) \\ &= \frac{P((k-1)\ C\textsc{iv} \cap k\ C\textsc{iv})}{P((k-1)\ C\textsc{iv})} \\ &= \frac{P(k\ C\textsc{iv})}{P((k-1)\ C\textsc{iv})}\,. \end{aligned} \tag{3.18}$$

The equality follows as the intersection between the set with k CIV and the set with $(k-1)$ CIV will be the set of quasars with k CIV absorbers. $Pr(k\ C\textsc{iv})$ is guaranteed to be less than 1, because there are always fewer spectra with more absorption systems.

Figure 3.5 shows the M_D priors we used for different searches as a function of z_{QSO}. When the redshift increases, all of the priors reach a plateau after $z_{QSO} \sim 3$. There is a decrease from $Pr(1\ C\textsc{iv})$ to the subsequent priors so that $Pr(7\ C\textsc{iv}) < 15\%$. CIV absorbers cluster (eg. Boksenberg et al. 2003), so the prior for detecting k absorbers in a spectrum given a redshift is larger than $Pr(1\ C\textsc{iv})^k$: when there is no clustering and the absorbers are perfectly independent.

3.3.5 Model likelihood

The model likelihood, $P(\mathcal{D}|\mathrm{M})$, in Equation 3.14 is the probability that the observed data, $\mathcal{D}$, have been generated by a considered model, M, after marginalising the model parameters. We do the marginalisation over a prior distribution for each parameter in the model:

$$P(\mathcal{D}|\mathrm{M}) = \int \mathrm{P}(\mathcal{D}|\mathrm{M},\theta)\mathrm{P}(\theta|\mathrm{M})\mathrm{d}\theta. \tag{3.19}$$

Here $P(\mathcal{D}|\mathrm{M},\theta)$ is the likelihood of the spectra being generated by model M, if the model has a certain set of parameters θ. We use the prior probability distribution of $P(\theta|\mathrm{M})$ from Equation

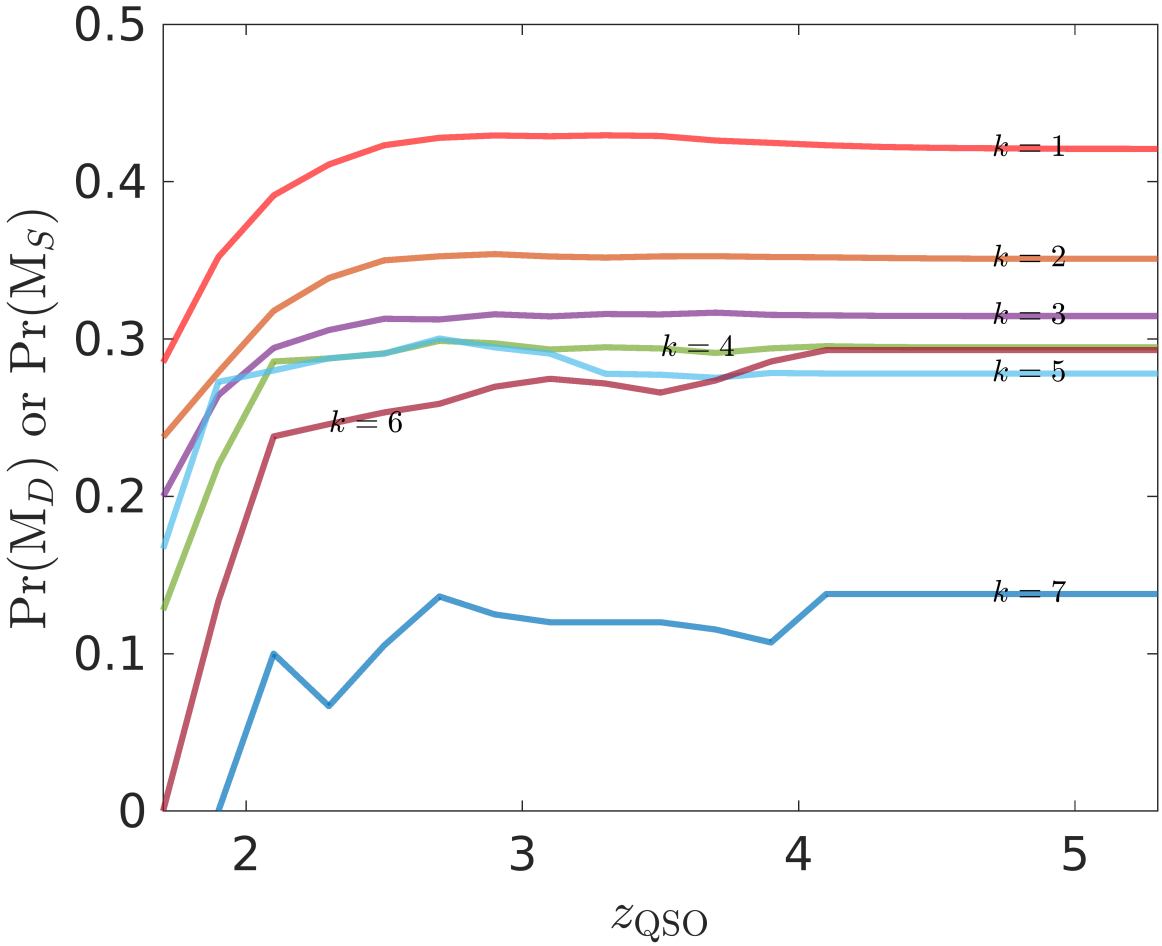

Figure 3.5: Prior probability for a spectrum containing k CIV absorbers as a function of quasar redshift, for $k = 1-7$. We use the average number of absorbers in the PM spectrum in our wavelength search range. CIV is *a priori* more likely as z_{QSO} increases but reaches a plateau at $z_{\mathrm{QSO}} \sim 2.5-3$. This is because the CIV wavelength coverage is shorter for low z_{QSO} as the 1548 Å emission line pushes to the blue-end of the SDSS spectral range. Note that we assume the same prior for the singlet model for $k = 1-7$.

3.19 to integrate out all of the possible θ and obtain a parameter-independent model likelihood. The null model $\mathrm{M_N}$ has no free parameters, but $\mathrm{M_D}$ and $\mathrm{M_S}$ have three free parameters each: 1) absorption redshift (z_{CIV}), 2) column density of *C*IV (N_{CIV}), and 3) Doppler velocity dispersion (σ_{CIV}). As mentioned in Section 3.3.1, it is sufficient for our purposes to model an absorption line at SDSS resolution with a single Voigt profile (defined by z_{CIV}, N_{CIV}, and σ_{CIV}) and use these values to measure a rest equivalent width; however, only redshift and the rest equivalent width are well-constrained by the data.

We need to have priors for each of these parameters to perform the integral in Equation 3.19. A parameter prior is a probability distribution which we know *a priori* might be true for given possible values of a parameter in a model. In implementations of the Bayesian approach for detecting DLAs in quasar spectra (Ho et al. 2020, 2021), the prior distribution for column density was learned from previous DLA catalogues, and they used a uniform absorber red-shift prior distribution.

One of the input parameters in the Voigt profile[13] is the absorber column density, N_{CIV}. Following Garnett et al. (2017) and Ho et al. (2020), we need to sample a column density distribution to perform the integral in Equation 3.19 and obtain the model likelihood. The column density range detected by Cooksey et al. (2013) was $\log N_{CIV} \approx 13$ to > 15.8. After some experimentation, we chose a slightly larger range: $12.5 < \log_{10} N_{CIV} < 16.1$, which maximised the performance on our validation set. We probed larger column densities than PM catalogue because: first, their column densities are often lower limits as they used the apparent optical depth method (Savage & Sembach 1991) and a lot of the absorption systems were saturated. Second, the larger size of SDSS DR12 gives us a longer survey pathlength which increases our chances of finding the exponentially rare strong systems. We searched for lower column densities than the PM catalogue

[13] See `voigt_IP.c` in https://github.com/rezamonadi/GaussianProcessCIV

since our catalogue could potentially be more sensitive to weaker absorbers. We thus used a mixture probability density function consisting of: (1) the N_{CIV} probability density function (obtained by kernel density estimation) from the reported values in the PM catalogue and (2) a uniform probability density function in the same range. We have also confirmed that our column density prior sample reproduces a rest equivalent width ($W_{r,1548}$) distribution in reasonable agreement with the PM catalogue for the 1548 Å line.

We also need a prior for the Doppler velocity dispersion, σ_v. The typical temperature for the intergalactic medium is $\sim 10^4 - 10^5$ K, which gives a $\sigma_v \sim 2.6 - 8.3$ km s^{-1} for CIV. However, at the low resolution of the SDSS spectra (~ 150 km s^{-1}), it is impossible to detect an absorption line with this velocity dispersion. Fortunately, CIV absorbers cluster (Boksenberg et al. 2003) and blend into a broader absorption profile with larger effective σ_{CIV}. By experimenting with different ranges for σ_{CIV} we chose lower and upper bounds for σ_{CIV} to be 35 km s^{-1}and 115 km s^{-1}, respectively. This range enables our process to be sensitive to similar rest equivalent widths as the PM catalogue.

We imposed a uniform prior distribution on the absorber redshift, z_{CIV}. The lower limit is the redshift at which the 1548 Å line is observed at $1310(1 + z_{QSO})$,[14] or the blue end of our input spectrum whichever is larger. Therefore:

$$1 + z_{min} = \max\left[\frac{\min(\lambda_{obs})}{1548}, \frac{1310(1 + z_{QSO})}{1548}\right]. \tag{3.20}$$

We also require a small velocity separation between the absorber and the quasar, to ensure that we are not finding intrinsic CIV absorbers around the host galaxy of the quasar:

$$z_{max} = z_{QSO} - \frac{\delta v}{c}(1 + z_{QSO}). \tag{3.21}$$

[14]To avoid possible confusion with any OI, SiII absorption pairs, see C13.

We considered δv = 1000 to 5000 km s^{-1}, and achieved the best validation performance when δv = 3000 km s^{-1}, which matches the minimum velocity separation between quasar and absorbers in the PM catalogue.

We assumed that N_{CIV} and σ_{CIV} are independent from z_{QSO}, although z_{CIV} depends on z_{QSO} as described in Equation 3.20 and Equation 3.21. We calculated the marginalised model likelihood by integrating the absorption-model priors $M_{D/S}$[15] as:

$$P(\theta|z_{QSO}) \propto P(z_{CIV}|z_{QSO})P(N_{CIV})P(\sigma_{CIV}). \tag{3.22}$$

Then we performed the integral for our absorption models, M_D and M_S, in Equation 3.19:

$$P(\mathcal{D}|z_{QSO}) \propto \int P(\vec{y}|\theta, z_{QSO})P(\theta|z_{QSO})d\theta. \tag{3.23}$$

However, Equation 3.23 is intractable, so we approximated it with a quasi-Monte Carlo method. This method selected 10,000 samples of $\{N_{CIV}, \sigma_{CIV}, z_{CIV}\}$ at which to calculate the model likelihood. The samples were drawn from a Halton sequence to ensure an approximately uniform spatial distribution. We approximate the model evidence by the sample mean:

$$P(\mathcal{D}|M_{D/S}, z_{QSO}) \simeq \frac{1}{N}\sum_{i=1}^{N} P(\mathcal{D}|\theta_i, z_{QSO}, M_{D/S}). \tag{3.24}$$

We integrated out the parameters, $\theta = \{z_{CIV}, N_{CIV}, \sigma_{CIV}\}$, with a given parameter prior $P(\theta|z_{QSO}, M_{D/S})$. We use 10,000 samples: lower sample sizes under-sample the likelihood function, while larger sample sizes cause the code to run slower. We considered 10,000–50,000 samples in the validation phase and found that increasing the number of samples did not significantly improve the validation performance. Note that using more samples increases the run-time cost of processing a quasar. In calculating the model evidence for the singlet model, M_S, we used a single component Voigt profile

[15] Either the doublet model or the singlet model.

centred on 1548 Å (Equation 3.16) while for calculating the model evidence for the doublet model, M_D, we use a double component Voigt profile centred at 1550 Å and 1548 Å (Equation 3.15). We used the same parameter priors for both the singlet and doublet models for simplicity.

3.3.6 Multiple absorber search

In this paper, instead of reporting probabilities for multiple CIV absorbers as Ho et al. (2020) did for DLAs, we simplified and reported the probability that there is an absorber at a given redshift. For example, the posterior probability for the $k = 3$ model in Ho et al. (2020) does not indicate which of these three absorbers in $M_{DLA(3)}$ is most probable, instead reporting the probability that a given spectrum contains some combination of three absorbers.

Here, we wish to find multiple absorbers in a spectrum. We proceed iteratively, noting that at any point the best-fit may be a singlet or a doublet, and mask out the most likely absorber each time. We mask 350 km s^{-1}around 1548 Å(MAP(z_{CIV}) + 1) and 350 km s^{-1}around 1550 Å(MAP(z_{CIV}) + 1), where MAP(z_{CIV}) is the maximum *a posteriori* value for z_{CIV}. For single-line absorbers, we mask 350 km s^{-1}around 1548 Å, again at MAP(z_{CIV}). Our procedure is as follows:

1) Fit our three models $M_{N/S/D}$ on an observed spectrum.

2) If M_N (the null, CIV-free, model) has the highest posterior for any search, there is no CIV absorption in the given spectrum. Stop any further searches. Otherwise go to step 3.

3) If either M_S or M_D has the highest posterior, mask the spectral region around the most probable absorption profile. Return to step 1 to search for subsequent absorbers if no more than seven searches previously have been done. Otherwise stop any further searches.

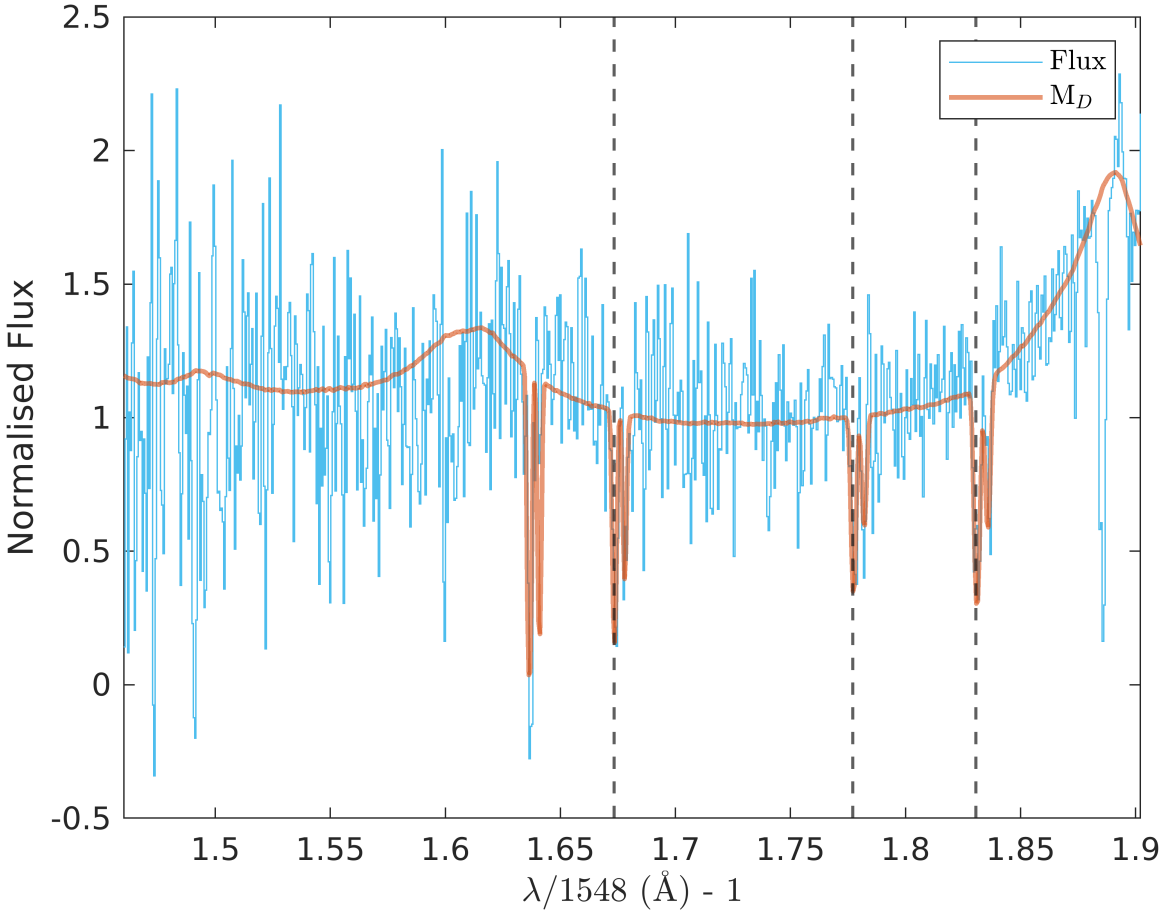

Figure 3.6: Example SDSS DR7 spectrum with QSO-ID: 51608-0267-264 and $z_{\rm QSO} = 1.89$. Both PM and our pipeline find three absorbers between $z_{C\rm IV} = 1.65$–1.85. We also find an absorber at $z_{C\rm IV} = 1.489$ (probability 92%) that was not detected by PM, due to noise in this part of the spectrum (specifically, the 1550 line was not automatically detected with their parameters, thus the doublet was not visually inspected). The probabilities that our pipeline provides for the existence of the first, second, third, and forth CIV absorber are $P(C\rm IV) = [1.00, 1.00, 1.00, 0.92]$, respectively, our maximum a posteriori absorber redshift values are $z_{C\rm IV} = [1.829, 1.672, 1.775, 1.489]$, and our rest equivalent widths from Voigt profile integration (see Equation 3.30) are $W^{\rm GP}_{r,1548} = [1.37, 0.87, 0.90, 0.79]$ Å. In the PM-catalogue the absorber redshifts are $z_{\rm PM} = [1.831, 1.673, 1.777]$ with corresponding $W^{\rm PM}_{r,1548} = [1.21 \pm 0.18, 1.40 \pm 0.20, 0.94 \pm 0.19]$ Å

Figure 3.6 shows an example quasar spectrum (SDSS DR7 QSO-ID: 51608-0267-264 and $z_{\rm QSO}$ = 1.89) within which both the PM and GP pipelines find three absorbers. Moreover, the GP pipeline finds an absorber at z_{CIV} = 1.489 (probability 92%) that was not detected by PM, due to noise in this part of the spectrum. Specifically, the 1550 line was not automatically detected with their parameters, thus the doublet was not visually inspected.

3.4 Validation

For validation, we trained a CIV-free model, M_N, on a reduced training set of 95% of the inspected spectra in the PM catalogue Cooksey et al. (2013). We then *validated* our algorithm with the remaining 5% of the inspected (1301) spectra in the PM catalogue to check the agreement between the PM catalogue and our method. Note that when we applied our algorithm to DR12 spectra, we re-trained our model using all SDSS DR7 spectra inspected in the PM catalogue without a reliable CIV absorber.

Our model is compared to the CIV absorbers as rated in the PM catalogue. Cooksey et al. 2013 rated their automatically detected CIV candidates from 0 (definitely not CIV), 1, 2, and 3 (definitely CIV), thus providing a rough estimate of confidence in an absorber. Absorbers with a ranking ≥ 2 are considered real CIV absorbers in the PM catalogue. We construct a "ground truth" sample of the PM CIV with rating ≥ 2. Within a spectrum, we enforce that our GP-detected absorber is within 350 km s^{-1} of a PM-detected system to be considered as a "match" between catalogues (see Figure 3.6 for examples of matched absorbers); this cutoff is roughly $3 \times \max(\sigma_{CIV})$ (where σ_{CIV} is measured by the GP),[16] which ensures we are not detecting a complex/blended system in

[16] For reference, in Cooksey et al. (2013), CIV absorbers were grouped into a single system if they were within 250 km s^{-1} of each other.

two successive iterations (see Section 3.3.6). Moreover, we obtained a better purity/completeness (see Section 3.4.3) with a 350 km s^{-1} masking window.

3.4.1 Velocity separation

The velocity separation between absorbers detected in both the GP and PM catalogues is:

$$\delta v_{\rm PM,GP} = \frac{z^{\rm PM}_{CIV} - z^{\rm GP}_{CIV}}{1 + z^{\rm PM}_{CIV}} c. \tag{3.25}$$

Figure 3.7 shows that absorber redshifts obtained by our pipeline in the validation set are almost always consistent with the PM catalogue at the level of the SDSS spectral resolution, i.e., $|\delta v_{\rm GP,PM}| \lesssim 150\,{\rm km\,s^{-1}}$. Very few points in Figure 3.7 lie outside of the ± 150 km s^{-1}horizontal lines. Our pipeline produces z_{CIV} on average slightly greater/ redder than the PM catalogue, with a median offset $\delta v^{\rm med}_{\rm PM,GP}\approx$-50 km s^{-1}. This is not a significant difference; by comparison an SDSS pixel is 69 km s^{-1}.

We visually inspected the 9 spectra in our validation set of 1301 spectra with $\delta v_{\rm PM,GP} \geq 50\,{\rm km\,s^{-1}}$: most of them were in a complex/blend system and some of them were close to the QSO where the GP continuum was not perfect. We also investigated the 14 spectra in the validation set that show $\delta v_{\rm PM,GP} \leq -150\,{\rm km\,s^{-1}}$: most of them belong to a complex system or even a mini-BAL system. In some cases the GP continuum fit is not good. As a reference, we investigated 17 spectra with $\delta v_{\rm PM,GP} \sim -50\,{\rm km\,s^{-1}}$: these spectra are usually high SNR and/or the GP continuum fit is very good, especially around the detected absorption system. Moreover, there is no significant correlation between the strength of the absorber systems and PM-GP velocity separation (Equation 3.25) as shown by Figure 3.8.

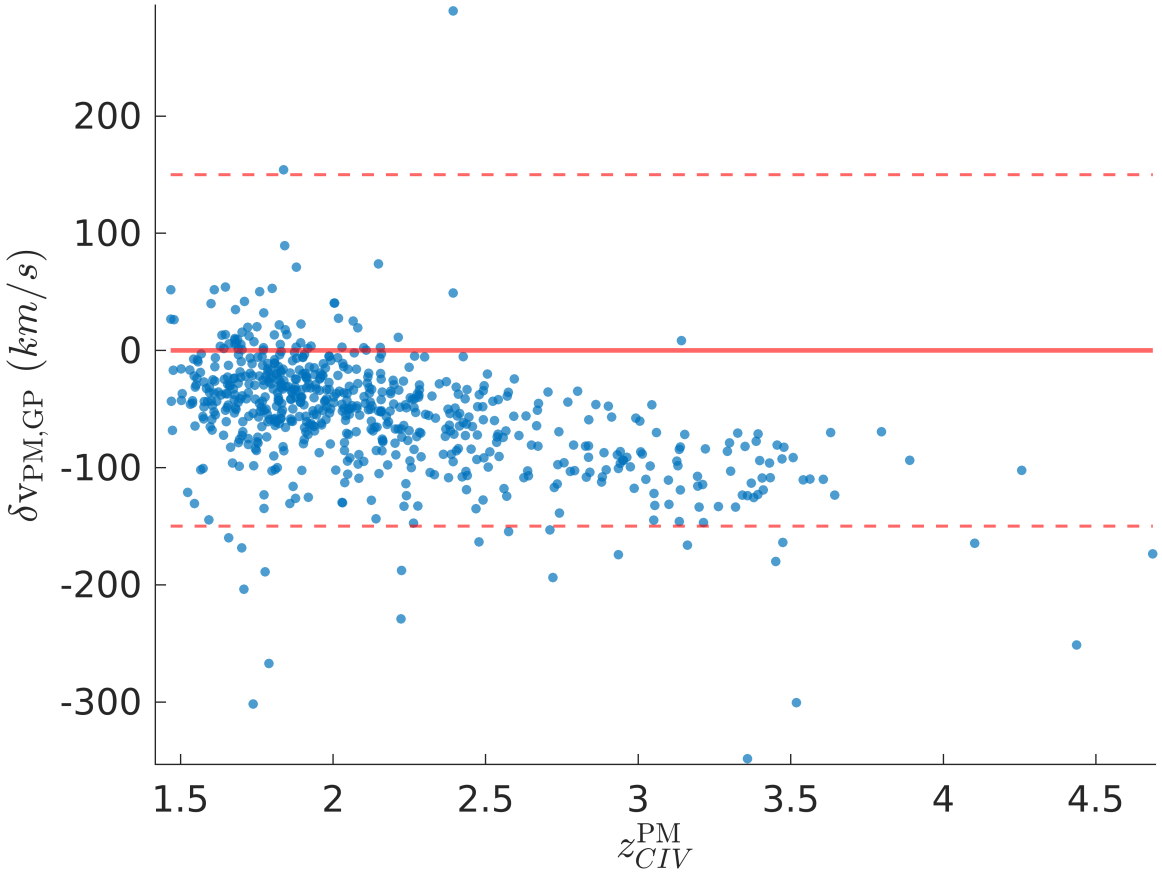

Figure 3.7: Velocity difference between the detected absorbers in the GP pipeline with $P(M_D) \geq 0.95$ in the validation set and the absorbers in the PM catalogue. Only absorber pairs closer than 350 km s^{-1}are shown. The thick red line shows $\delta v_{\mathrm{PM,GP}} = 0$ and the dashed lines are $\delta v_{\mathrm{PM,GP}} = \pm 150$ km s^{-1}(the SDSS spectral resolution). The median offset is $\delta v_{\mathrm{PM,GP}}^{\mathrm{med}} \approx$-50 km s^{-1}, which is less than an SDSS pixel (69 km s^{-1}).

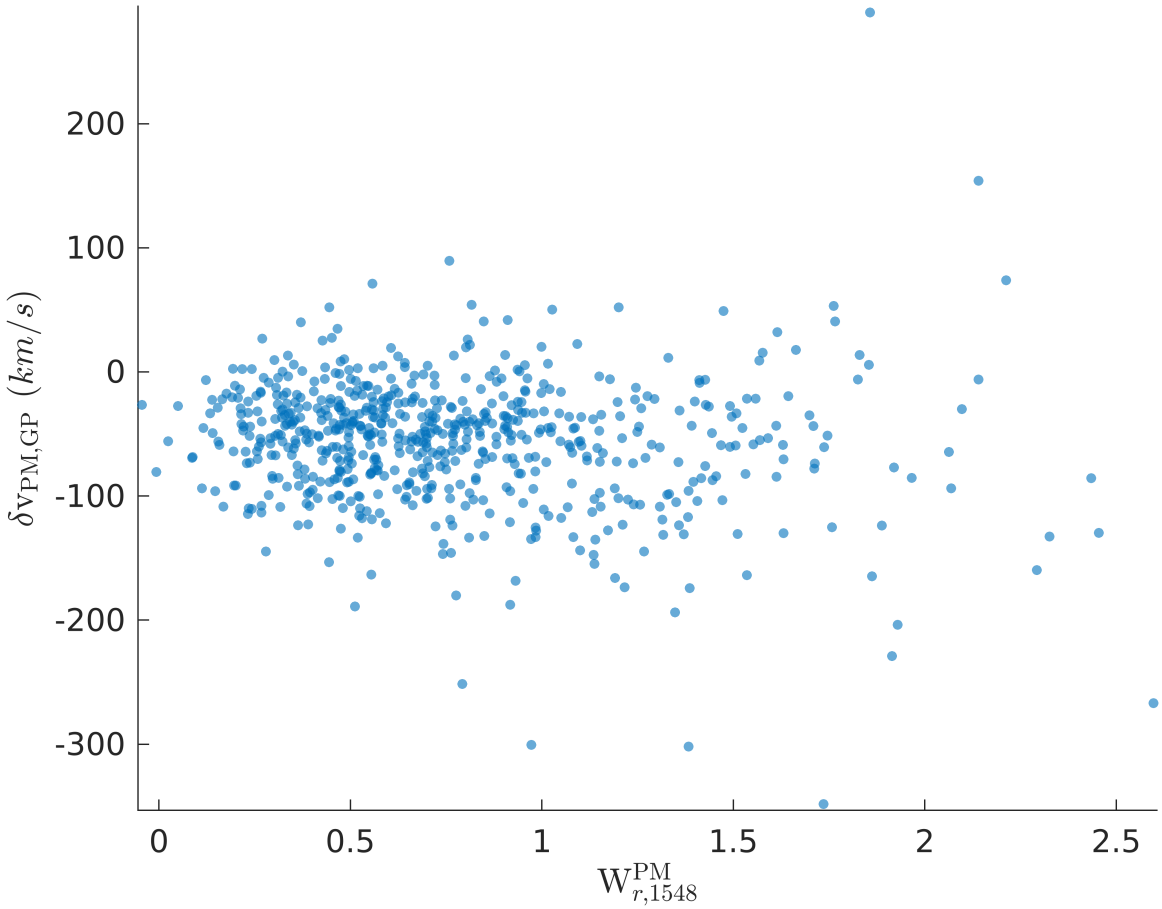

Figure 3.8: Velocity separation (Equation 3.25) between GP and PM detected CIV absorption systems is shown versus the reported rest equivalent width values for 1548 Å in the PM catalogue ($\mathrm{W}^{\mathrm{PM}}_{r,1548}$). There is no correlation between the velocity separation and the strength of detected absorbers.

3.4.2 Receiver Operator Characteristic (ROC) curve

We use the Receiver Operator Characteristic (ROC) curve (Figure 3.9), which is the true positive rate versus false positive rate for any classification threshold: $0 \leq P(\mathrm{M_D}) \leq 1$ to obtain a score out of 1 for the performance of our classification (no CIVabsorber versus CIVabsorbers). The Y-axis of the ROC curve in Figure 3.9, the true positive rate, is the ratio of the number of CIV absorbers in our catalogue to the total number of of absorbers in the PM catalogue with a ranking ≥ 2. CIV absorbers in our catalogue are defined to be those with posterior probability greater than a threshold, $\mathrm{P(M_D)}$, between 0 and 1. They must also be less than 350 $\mathrm{km\,s^{-1}}$apart from an absorber in the PM catalogue with a ranking≥ 2. The X-axis of the ROC curve in Figure 3.9, the false positive rate, is the ratio of CIV absorbers in our catalogue that do not have any matching absorber with ranking ≥ 2 in the PM catalogue (given any $\mathrm{P(M_D)}$ threshold between 0 and 1) to the total number of absorbers in the PM catalogue with a ranking ≥ 2.

A higher classification performance (i.e. in each search run over a spectrum we classify it as CIV-free or having a CIV absorber) is reflected in a larger area under the curve (AUC) for the ROC curve. We obtain a quite reasonable AUC = 0.87. Note that here "true positive" refers to a PM CIV absorber recovered by the GP algorithm in the training set, and "false positive" is a GP CIV absorber not in the PM catalogue. However, as seen in Figure 3.6, the GP procedure *can* find real/true CIV absorption not identified in the PM survey; hence, "false positives" may be better considered "GP unique". This also means that the classification performance (AUC = 0.87) we obtained here might underestimate the true performance.

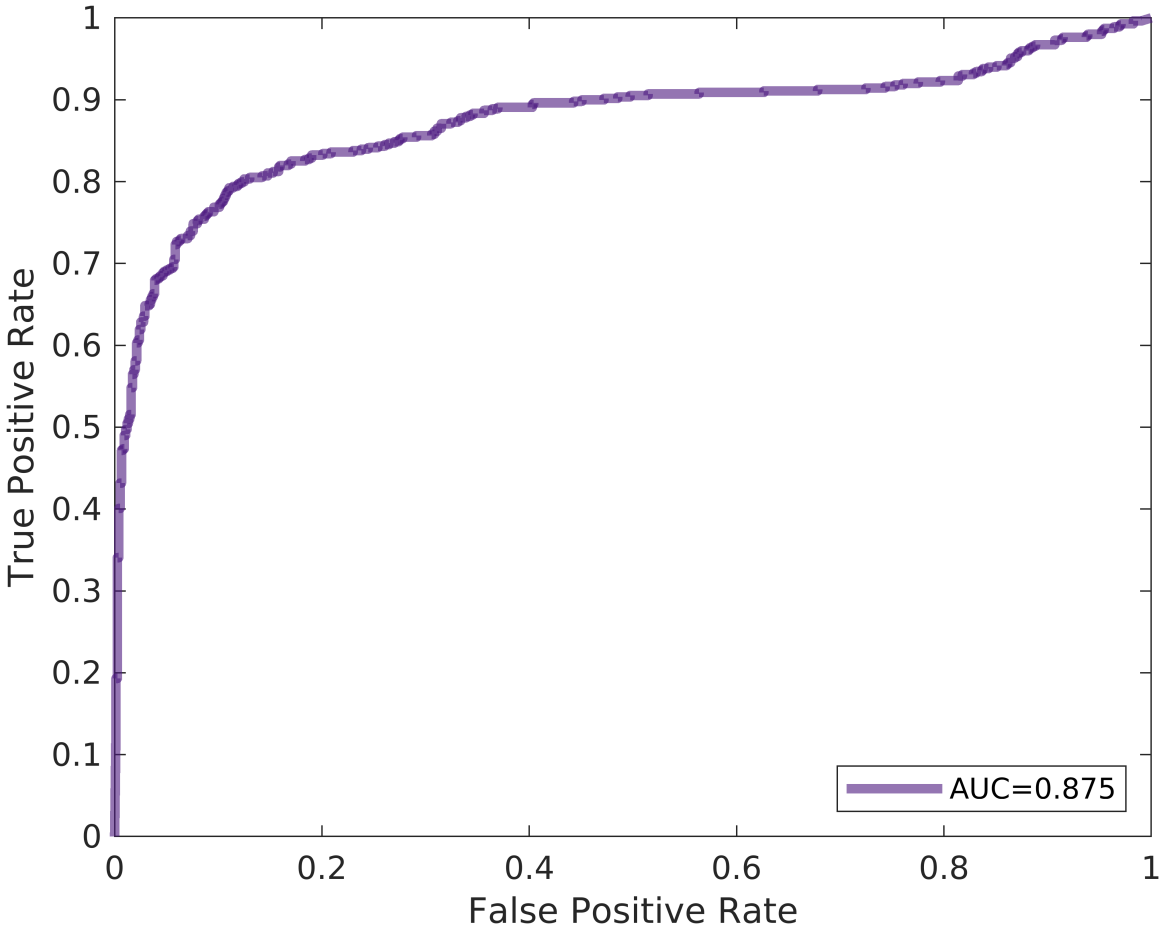

Figure 3.9: Receiver Operator Characteristic (ROC) curve for our DR7 validation. True Positive Rate is plotted versus False Positive Rate. True positives are CIV systems in our catalogue at least 350 km s^{-1}apart from an absorber in the PM catalogue with ranking $\geq$2 given any P(M_D) threshold between 0 and 1. False positives are those absorbers in our catalogue that do not have any matching absorber in the PM catalogue; though they may be real CIV absorbers (see Figure 3.6). Above a relatively small False Positive Rate ($\sim$ 0.2), our algorithm procedure obtains True Positive Rate above 80% and, hence, is a successful way to identify CIV absorbers. The area under the ROC curve (AUC) is a quantitative metric for the equality of the GP algorithm; we get AUC = 0.87, a reasonable value compared to an ideal classification that gives AUC = 1.00.

3.4.3 Purity and Completeness

We assessed our algorithm's performance by comparing individual absorption systems. We can compare our GP catalogue for various CIV posterior probabilities to the 'ground truth' sample of the PM catalogue. We define the purity of our GP catalogue as the fraction of the GP catalogue also in the PM catalogue:

$$\text{Purity} = \frac{\text{GP} \cap \text{PM}}{\text{GP}}. \tag{3.26}$$

The completeness is the fraction of the PM catalogue also in the GP catalogue:

$$\text{Completeness} = \frac{\text{GP} \cap \text{PM}}{\text{PM}}. \tag{3.27}$$

Figure 3.10 shows completeness and purity as a function of threshold value. One should choose a threshold that gives the best possible combination of purity and completeness, around the point where the curves intersect. We thus choose a threshold of 95%, which Figure 3.10 shows produces purity and completeness of $\sim$ 80% in a roughly equal balance. However, our catalogue reports posterior probabilities, so the user may choose a different threshold as desired for their application. One may sacrifice purity for completeness or vice versa.

3.4.4 Rest equivalent width comparison

We can evaluate our algorithm by comparing 1548 Å rest equivalent width $W_r^{\text{GP,flux}}$ between the GP and PM catalogues. $W_r^{\text{GP,flux}}$ is obtained by integrating the normalised flux deficit from our GP continuum (M_N) in a wavelength integration window corresponding to $4 \times \sigma_{CIV}$ around the maximum *a posteriori* z_{CIV} for the 1548 Å line. We impose that the flux integration window does not exceed the midpoint of the 1550 Å and 1548 Å lines.

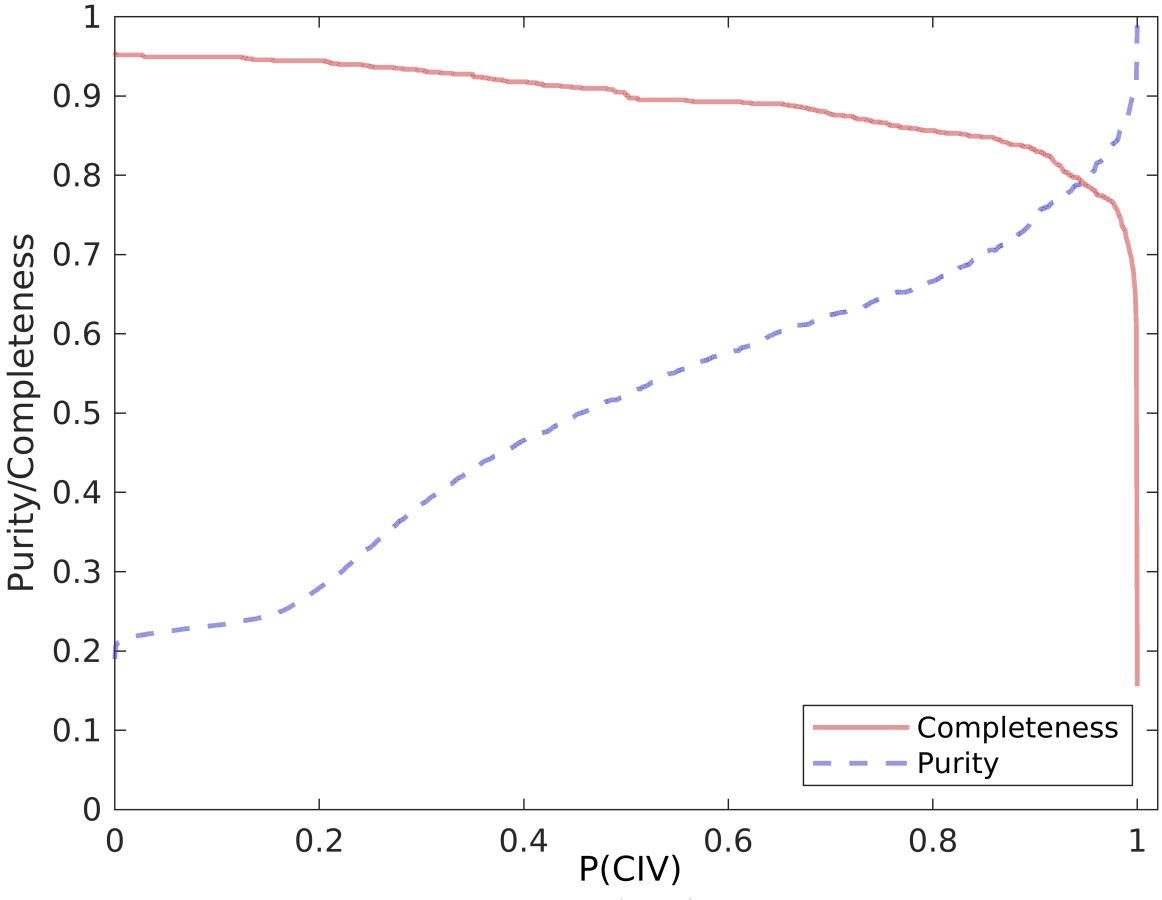

Figure 3.10: Purity (Equation 3.26) and completeness (Equation 3.27) of the GP catalogue compared to the PM catalogue for different CIV posterior probability (Equation 3.14) thresholds. The maximum allowed velocity separation between our catalogue and the PM catalogue absorbers is 350 km s^{-1}. The intersection of the purity (dashed blue curve) and completeness (solid red curve) at a threshold of $\sim$ 95% gives us a balanced purity/completeness of $\sim$ 80%.

Figure 3.11 shows the difference ratio in our validation set between the rest equivalent width of the 1548Å line in the GP catalogue and the PM catalogue, scaled by the maximum error (because the rest equivalent width errors from the PM and GP catalogues are highly correlated):

$$\frac{\mathrm{W}_{r,1548}^{\mathrm{GP,flux}} - \mathrm{W}_{r,1548}^{\mathrm{PM}}}{\mathrm{err}_{\mathrm{max}}}.$$

The maximum error, $\mathrm{err}_{\mathrm{max}}$, is obtained by comparing the rest equivalent width error from the GP pipeline to that from the PM catalogue:

$$\mathrm{err}_{\mathrm{max}} = \max\{\mathrm{err}(\mathrm{W}_{r,1548}^{\mathrm{GP,flux}}), \mathrm{err}(\mathrm{W}_{r,1548}^{\mathrm{PM}})\}. \tag{3.28}$$

We obtained $\mathrm{err}(\mathrm{W}_{r,1548}^{\mathrm{GP,flux}})$ by considering the observed noise in each pixel included in the integration window described above.

Around 94% of the data points in Figure 3.11 have $\left|\left(\mathrm{W}_{r,1548}^{\mathrm{GP,flux}} - \mathrm{W}_{r,1548}^{\mathrm{PM}}\right)/\mathrm{err}_{\mathrm{max}}\right| \leq 2$, which shows a reasonable consistency between the GP and PM rest equivalent widths.

We visually inspected all of the 21 absorbers with $\left(\mathrm{W}_{r,1548}^{\mathrm{GP,flux}} - \mathrm{W}_{r,1548}^{\mathrm{PM}}\right)/\mathrm{err}_{\mathrm{max}} < -2$: they mostly have continuum issues and a low GP continuum. For QSO 53886-1823-377, a triplet[17] CIV system at $z_{CIV} = 1.838$ caused a lower 1548 rest equivalent width in the GP catalogue than in the PM catalogue. In the spectrum of QSO 53083-1757-529 our algorithm finds an absorber at the end of the spectrum, which also yields a lower rest equivalent width when compared to the PM catalogue. Looking at 13 absorbers with $\left(\mathrm{W}_{r,1548}^{\mathrm{GP,flux}} - \mathrm{W}_{r,1548}^{\mathrm{PM}}\right)/\mathrm{err}_{\mathrm{max}} > 2$, we realised that most of these absorbers belong to a triplet CIV system. As a reference we also checked absorbers with $\left|\mathrm{W}_{r,1548}^{\mathrm{GP,flux}} - \mathrm{W}_{r,1548}^{\mathrm{PM}}\right|/\mathrm{err}_{\mathrm{max}} < 0.025$: these spectra were mostly high SNR and the GP continuum fit the observed quasar very well.

[17] When a lower redshift absorber's 1550 Å line blends with the 1548 Å line of the higher redshift absorber.

The colour bar in Figure 3.11 shows the *maximum a posteriori* σ_{CIV} that the GP algorithm produces for each absorber. There is a correlation between larger *maximum a posteriori* σ_{CIV} and absorbers where the GP rest equivalent width, $W^{GP,flux}_{r,1548}$, is larger than the PM rest equivalent width, $W^{PM}_{r,1548}$. We visually inspected these systems and found that many of them are triplet or mini-BAL systems, for which the GP is more likely to give a large σ_{CIV}.

The difference in rest equivalent widths of the 1550 Å line between the GP and PM catalogues behaves similarly. We find that 518 (86%) of GP absorbers with a PM absorber system at a redshift offset less than 350 km s^{-1} away showed $\left|W^{GP,flux}_{r,1500} - W^{PM}_{r,1500}\right| / err_{max} \leq 2$. The 1550 Å line is weaker than the 1548 Å line, leading to a generally lower detection significance. However, for strong absorbers it is useful because it is less saturated.

The GP pipeline finds 822 absorbers in the validation set spectra with $P(M_D) \geq 95\%$. In the PM catalogue the validation set spectra contain 829 absorbers with a ranking ≥ 2. We can divide these absorbers into four different categories with the following statistics:

1. PM & GP: absorbers with a ranking$\geq$ 2 in the PM catalogue, $P(M_D) \geq 0.95$ in the GP catalogue, and $\delta v_{PM,GP} \leq 350$ km s^{-1}. This category contains 647 absorbers, $\sim$ 78% of the PM absorbers and $\sim$ 79% of the GP absorbers among the 1301 spectra in the validation set.

2. GP only: absorbers with $P(M_D) \geq 0.95$ but no absorber in the PM catalogue with ranking ≥ 2 and a velocity offset less than 350 km s^{-1}. This category includes 175 absorbers (21% of the GP absorbers in the validation set). Some of these absorbers are true CIV which fell beneath the sensitivity of the candidate search in Cooksey et al. (2013), and some are other doublet lines which our GP model has incorrectly classified as CIV.

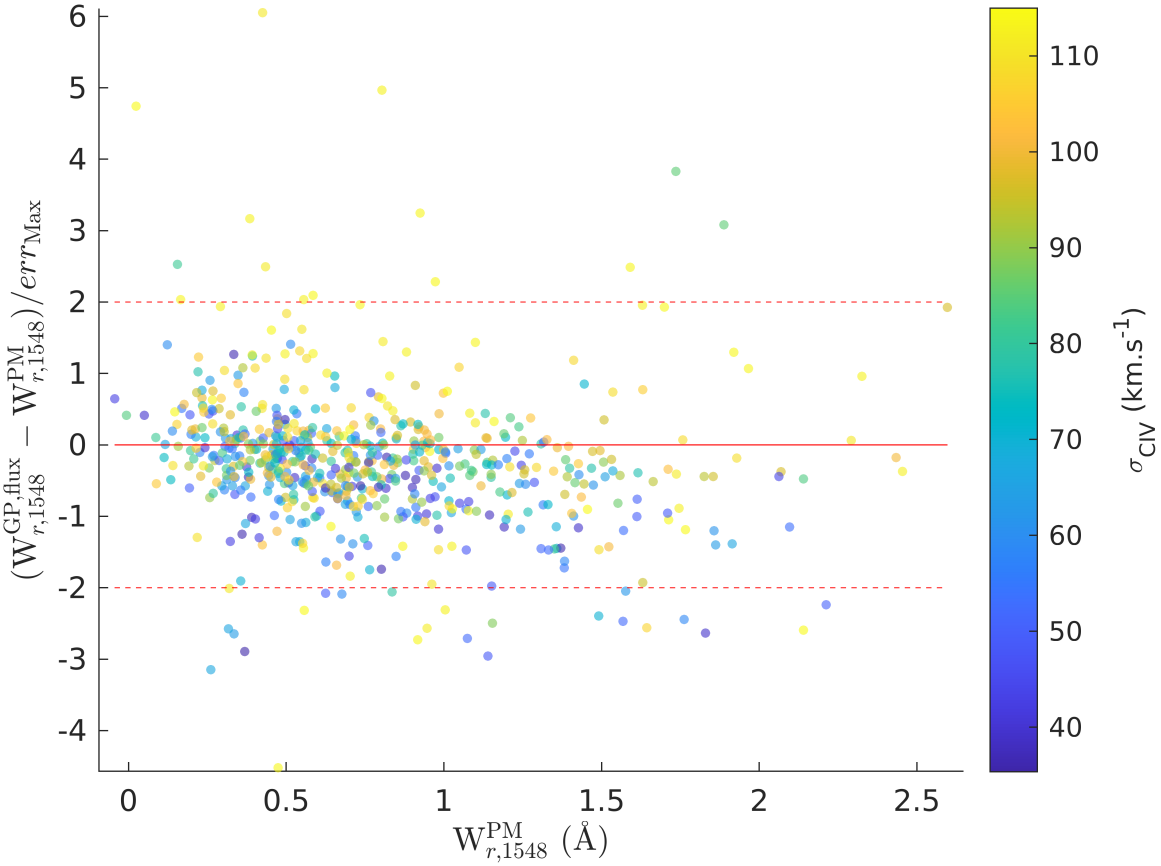

Figure 3.11: The ratio of the difference between rest equivalent width from our pipeline with boxcar flux summation ($W^{GP,flux}_{r,1548}$) and rest equivalent width from the PM catalogue ($W^{PM}_{r,1548}$) to the total error (see Equation 3.28) from the PM catalogue and our pipeline for $W_{r,1548}$. The data points here are those absorption systems in the validation set where our pipeline reports an absorber with $P(M_D) \geq 0.95$ and for which there is an absorber with ranking ≥ 2 in the PM catalogue at a redshift offset less than 350 km s^{-1}(GP & PM in Section 3.4.4). As the colour bar shows, there is a trend towards larger maximum *a posteriori* σ_{CIV} when the GP rest equivalent width is larger than the rest equivalent width from the PM catalogue.

3. GP uncertain: absorbers with a ranking ≥ 2 in the PM catalogue, and an absorber from the GP catalogue with a velocity offset less than 350 km s^{-1} but $P(M_D) < 95\%$. There are 142 of these absorbers, $\sim$ 17% of the PM absorbers in the validation set spectra. Note that 85 ($\sim$ 60%) of these GP uncertain absorbers have $P(M_D) \geq 50\%$ in the GP catalogue and that the authors of the PM catalogue resolved ambiguous absorbers by inspecting other metal lines from the same system.

4. PM only: 40 (4.8%) of 829 absorbers with ranking ≥ 2 in the PM catalogue validation set had no GP absorber candidates within 350 km s^{-1}in the GP catalogue. The GP pipeline thus misses these absorbers in its successive searches of the validation set spectra. Two absorbers were assigned to this category because there were two PM absorbers in the spectrum closer than 350 km s^{-1} to each other. The GP catalogue found one, and the region containing the second was masked. Note that these absorbers are the reason why Figure 3.10 does not show a completeness of 1, even with a threshold of $P(M_D) = 0$.

Figure 3.12 shows the distribution of rest equivalent widths for 40 PM only absorbers, 142 GP uncertain absorbers, 175 GP only absorbers, and 647 PM & GP absorbers in the four categories described above. Figure 3.12 also demonstrates that absorbers in these four categories have similar rest equivalent width distributions.

There are 17 strong absorbers ($W_{r,1548} \geq 1.2$Å) in the PM only or GP uncertain categories. 11 of these absorbers are triplet/mini-BAL systems where the GP pipeline gives $P(M_S) \sim P(M_D)$. The complex shape of these absorption systems are not a good match to either model, so the GP pipeline is not able to distinguish between them. Two absorbers have $P(M_D) \geq 0.95$ but a velocity

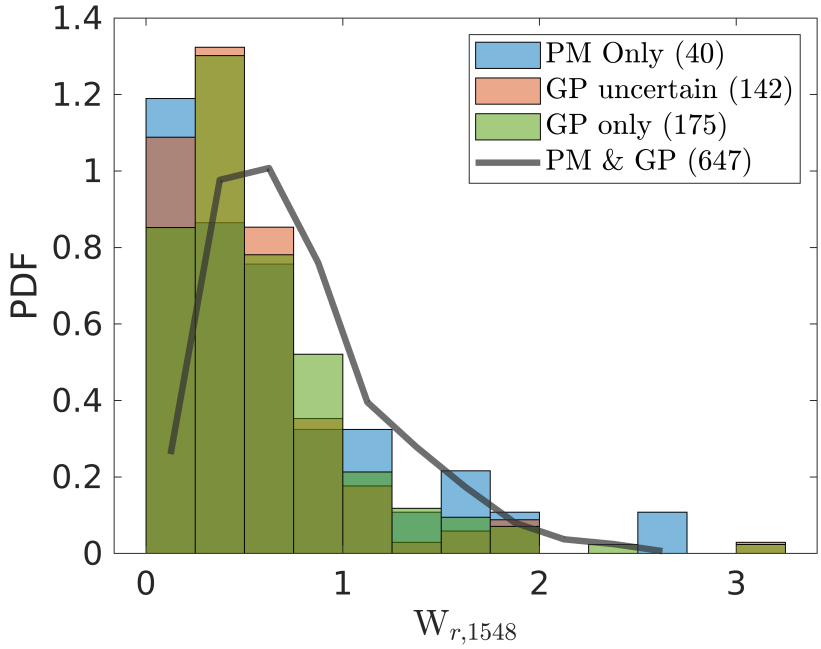

Figure 3.12: Distribution of $W_{r,1548}$ for absorbers in four categories described in Section 3.4.4: detected in both the GP and PM catalogues (thick black line), in the GP uncertain (brown), in GP only (green), in the PM catalogue only (blue). The rest equivalent width distribution is similar for all categories. There are some strong absorbers with ($W_{r,1548} > 1.2$Å) classified as "PM only". Visual inspection of the spectra of these systems indicates that they are part of a triplet/complex absorber or a broad mini-BAL system.

separation more than 350 km s^{-1}. One of these is also close to a complex absorber system. Five absorbers are detected by the GP catalogue as a singlet, and so have $P(\mathrm{M_S}) \gg \mathrm{P(M_D)}$.

Thus most missed strong absorbers are caused either by CIV triplets or by the GP pipeline preferring a singlet fit to a doublet in cases where the doublet structure is not well resolved. Note that when conducting visual inspection of CIV absorbers, an observer may resolve ambiguous lines using information from other metal line transitions associated with the same system, whereas our GP pipeline uses only information from the CIV transition.

3.4.5 Example absorbers

In this section we examine example absorbers from the GP only, GP uncertain and PM only categories discussed above. Figure 3.13 shows an example of a spectrum where the GP catalogue shows two absorbers with probability more than 95%, but the PM catalogue has zero detections. In this case, the GP CIV at $z = 2.288$ is actually AlII $\lambda 1670$ from a strong, multi-component system at $z = 2.05$ with MgII $\lambda\lambda 2796, 2803$; FeII $\lambda\lambda\lambda 2344, 2374, 2384$ and $\lambda\lambda 2586, 2600$; and AlIII$\lambda\lambda 1854, 1862$. The latter was flagged by the GP algorithm as $z_{CIV} = 2.650$. The $z = 2.288$ "CIV" was detected as a candidate in Cooksey et al. (2013) but visual inspection revealed its true identification; the $z = 2.650$ "CIV" was not even a candidate in Cooksey et al. (2013) because the would-be 1550 line was not detected by the automated candidate finder (i.e., it fell below their sensitivity threshold). Thus some of the absorbers in the GP only category are simply missed by the PM pipeline and some are false detections of other doublets.

Figure 3.14 shows an example spectrum where the GP pipeline is uncertain about an absorber detected in the PM catalogue. Both pipelines find the CIV absorber at $z^{\rm GP}_{\rm CIV} = 1.827$. However, the PM catalogue identifies a second absorber at $z^{\rm PM}_{\rm CIV} = 1.822$. This absorber is also detected by the GP pipeline. However, the GP pipeline is unable to distinguish between the doublet and singlet models as the 1550 Å line is blended with the higher redshift absorber. It thus assigns both models equal probability, hence $P(M_D) = 49\%$. P uncertain category that I explained above.

Figure 3.15 illustrates QSO-ID: 51943-0300-475, which contains an example of an absorber in the PM only category. The PM catalogue contains two absorbers at $z^{\rm PM}_{\rm CIV} = [3.5309, 3.5389]$. The GP pipeline finds an absorber in the first CIV-search at $z^{GP}_{\rm CIV} = 3.540574$ with a posterior prob-

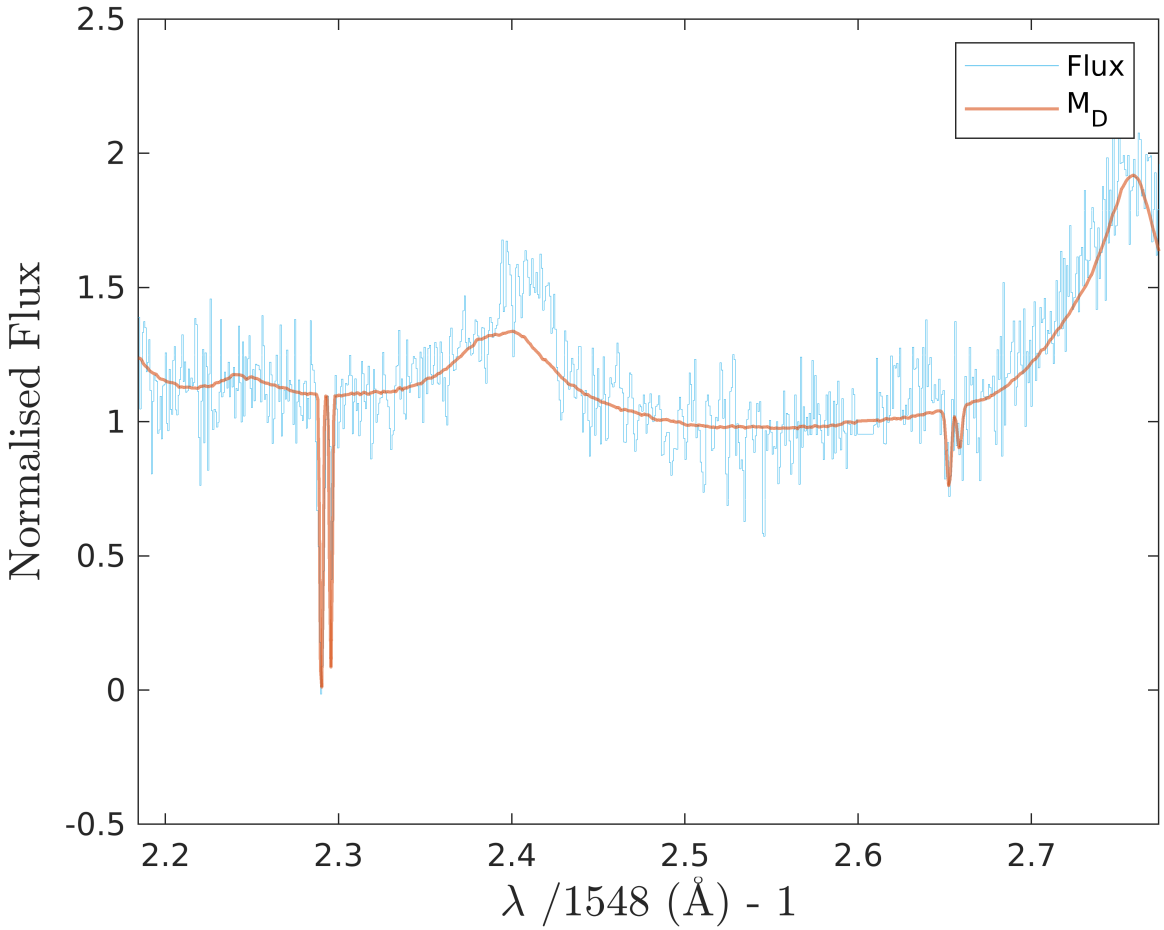

Figure 3.13: Example spectrum with two CIV absorbers found by GP with high confidence but not included in the PM catalogue. The QSO-ID is 51994-0309-592 and $z_{\rm QSO} = 2.76$. Posterior probabilities for the two searches are $\mathrm{P(M_D)} = [1.00, 0.98]$. The maximum *a posteriori* absorption redshifts are $z_{CIV} = [2.288, 2.650]$, and the rest equivalent widths are $\mathrm{W}^{\rm GP,flux}_{r,1548} = [0.90, 0.32]$ Å. These two "CIV" systems are actually non-CIV absorption lines from a strong, complex system at lower redshift. The PM pipeline identified the $z = 2.288$ lines as a CIV *candidate* but ranked it zero; the $z = 2.650$ "CIV 1550 Å" line fell below the PM detection threshold.

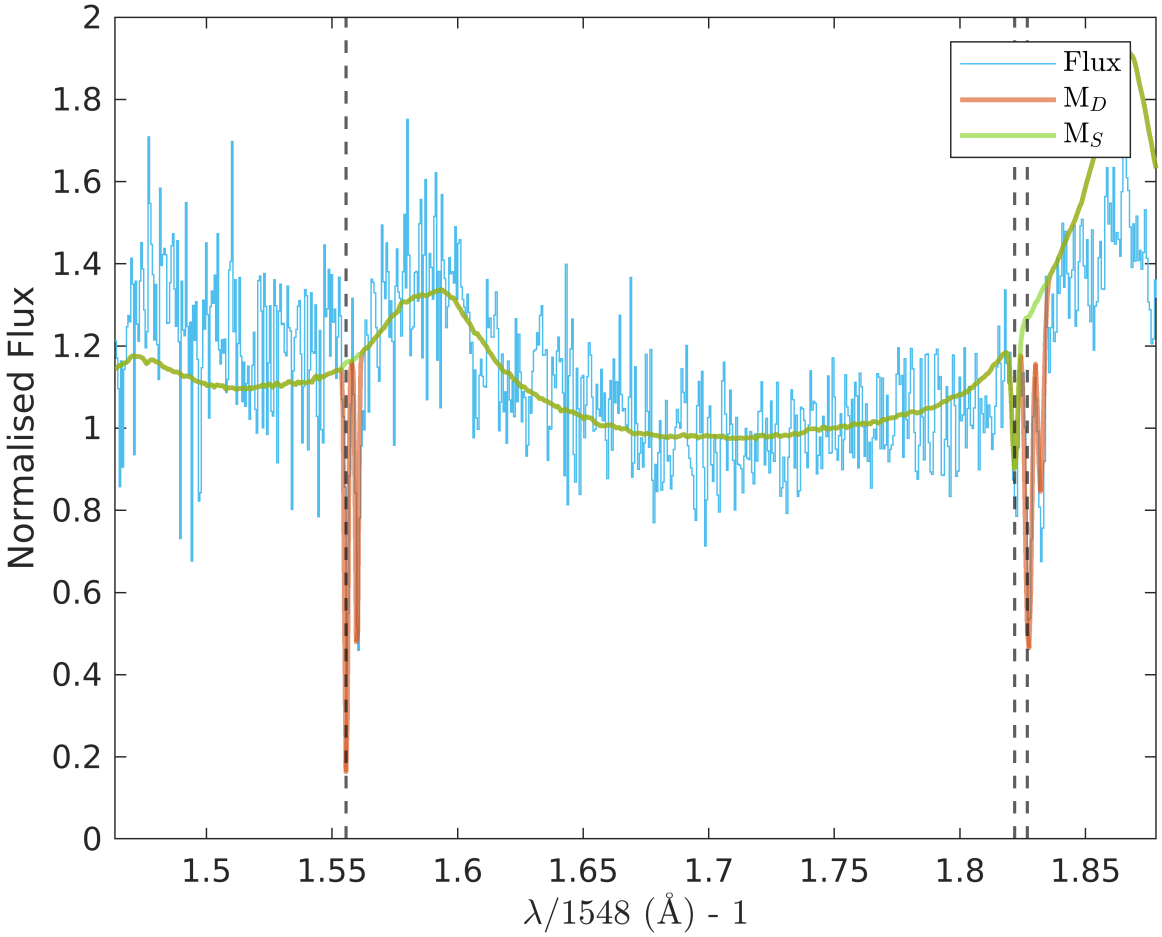

Figure 3.14: Example of an absorber at $z^{\rm PM}_{CIV} = 1.822$ detected by the PM catalogue, but assigned a relatively low probability ($P(M_D) = 49\%$) by the GP catalogue. The QSO-ID for this spectrum is 52367-0332-585, and the quasar redshift is 1.87. The vertical dashed lines show the position of PM absorbers. The posterior absorption probabilities are $P(M_D) = [1.00, 1.00, 0.49, 0.15]$, with maximum a posterior absorber redshifts of $z_{CIV} = [1.556, 1.827, 1.822, 1.693]$, and the rest equivalent widths are $W^{\rm GP,flux}_{r,1548} = [0.528 \pm 0.37, 1.21 \pm 0.25, 0.55 \pm 0.30, 0.05 \pm 0.35]$ Å. The PM catalogue reported absorbers at $z^{\rm PM}_{CIV} = [1.556, 1.827, 1.822]$ with $W^{\rm PM}_{r,1548} = [0.88 \pm 0.12, 0.88 \pm 0.08, 0.40 \pm 0.10]$ Å.

ability of 1 for the doublet model. According to our multi-absorber finding procedure (see Section 3.3.6) we mask the observed flux 350 km s^{-1}around the found absorber and do the next search. However, since the other reported absorber in the PM catalogue ($z^{\rm PM}_{\rm CIV}$ = 3.5309) is offset only 110 km s^{-1}from the absorber found in the first GP search, it is in a masked region and not identified by the GP pipeline in the second search.

3.5 Results for SDSS DR12

We applied our model to find CIV absorbers in a subset of the SDSS DR12 quasar catalogue. We searched quasars with rest-frame wavelength coverage between 1310 Å and 1548 Å ($1.7 < z_{\rm QSO} < 5.7$), and without detected BALs. This leaves 185,425 quasar spectra (see Section 3.2). For each spectrum, the GP pipeline provides (shown as columns in Table 3.1): posterior probability of CIV absorption, maximum *a posteriori* values for our absorption model parameters (z_{CIV}, N_{CIV}, and σ_{CIV}), together with their 95% confidence intervals, and rest equivalent widths (for 1548 Å and 1550 Å) and their 95% confidence intervals. Maximum *a posteriori* values and 95% confidence intervals for our absorption model parameters summarise the likelihood distribution, $P(\mathcal{D}|\theta_i, z_{\rm QSO}, {\rm M_D})$, of our 10,000 parameter samples (see Equation 3.24). Each of these results are contained in a $185,425 \times 7$ array. If the search terminated finding fewer than seven absorbers, we report a `NaN` value for the columns associated with all further absorbers. Table 3.1 shows a snapshot of our search results for the first 10 absorbers with $\rm P(M_D) \geq 0$.

Figures 3.16 through 3.19 illustrate the four CIV searches done by the GP pipeline on QSO-56265-6151-936 with $z_{\rm QSO}$ = 2.4811, and we briefly explain these iterations here. We found

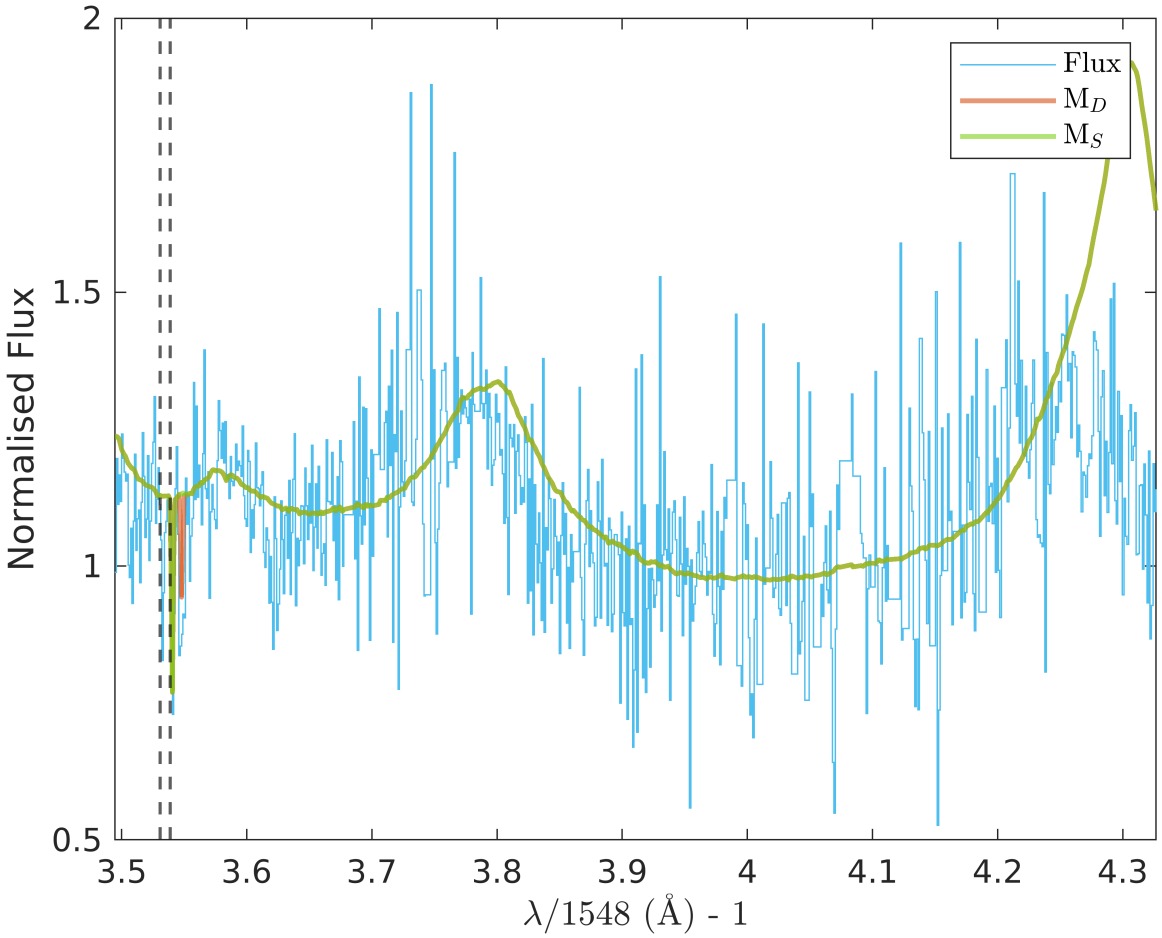

Figure 3.15: Example spectrum containing a PM only absorber for QSO-ID: 51943-0300-475 and $z_{\rm QSO} = 4.31$ where $z^{\rm PM}_{\rm CIV} = [3.5309, 3.5389]$ (vertical dashed lines). GP assigns P(M_D) = 1 to $z^{GP}_{\rm CIV} = 3.540574$ which is offset by only 110 km s^{-1}from $z^{\rm PM}_{\rm CIV} = 3.5389$. Before the second search, we mask 350 km s^{-1}around the first absorber and thus are unable to detect the second PM catalogue absorber.

Table 3.1: For each sight-line, identified by Column 1 and 2, we report the absorber's redshift (Column 3), column density in $\log(\mathrm{cm}^{-2})$ (Column 4), Doppler velocity dispersion in $\mathrm{km\,s}^{-1}$(Column 5), rest equivalent width for 1548 Å $W_{r,1548}$ (Column 6), rest equivalent width for 1550 Å $W_{r,1550}$ (Column 7), the posterior probability of the CIV absorber $P(M_D)$ (Column 8), and the posterior probability of the singlet absorber $P(M_S)$ (Column 9). We show only absorbers with $P(M_D)\neq$NaN. This table demonstrates a portion of the full table for the first ten rows. Note that those measurements with large errors are uncertain (i.e. low absorption model posterior probability). The full table with 445,765 rows is available at https://doi.org/10.5281/zenodo.7872725.

(1) QSO-ID	(2) z_{QSO}	(3) z_{CIV}	(4) $\log(N_{CIV})$ $\log(\mathrm{cm}^{-2})$	(5) σ_{CIV} $\mathrm{km\,s}^{-1}$	(6) $W_{r,1548}$ (Å)	(7) $W_{r,1550}$ (Å)	(8) $P(M_D)$	(9) $P(M_S)$
56238-6173-528	2.3091	1.91039±0.00370	15.66±0.84	52.24±0.63	1.306±1.601	1.177±1.759	0.63	0.18
	2.3091	1.91039±0.00370	15.66±0.84	52.24±0.63	1.306±1.601	1.177±1.759	0.63	0.18
	2.3091	2.21620±0.00380	14.75±1.34	104.74±0.65	1.328±4.342	0.838±3.956	0.40	0.00
56268-6177-595	2.4979	2.11727±0.00113	13.96±0.32	58.93±0.18	0.280±0.599	0.152±0.367	0.23	0.61
	2.4979	1.99557±0.00263	13.99±0.67	39.18±0.43	0.261±0.968	0.148±0.725	0.41	0.00
55810-4354-646	2.3280	1.90383±0.00134	13.89±0.35	48.29±0.23	0.230±0.515	0.125±0.314	0.30	0.36
	2.3280	1.94502±0.00355	13.77±0.96	55.32±0.61	0.187±1.387	0.099±0.980	0.31	0.00
56565-6498-177	2.3770	1.95293±0.00380	14.52±1.10	49.38±0.64	0.636±1.921	0.418±1.781	0.29	0.32
56268-6177-608	3.7120	3.33339±0.00063	14.49±0.10	75.29±0.08	0.763±0.378	0.460±0.284	1.00	0.00
	3.7120	3.51346±0.00163	14.21±0.28	96.87±0.20	0.530±0.939	0.289±0.586	0.76	0.00
	3.7120	3.22375±0.00499	14.06±0.98	62.16±0.60	0.344±1.880	0.189±1.506	0.28	0.00

a CIV absorber at $z_{CIV} = 2.13682$. In the first search, the null model had $P(M_N) = 0.0$, the single line model $P(M_S) = 0.0$, and the CIV doublet model $P(M_D) = 1.0$. We thus masked the CIV doublet model 350 $\mathrm{km\,s}^{-1}$around the CIV absorber at $z_{CIV} = 2.13682$ in the first CIV search and commenced the second search, shown in Figure 3.17. Our second search found an absorber at $z_{CIV} = 2.15132$ with $P(M_D) = 1.0$, $P(M_S) = 0.0$, and $P(M_N) = 0.0$. For the third CIV search we masked 350 $\mathrm{km\,s}^{-1}$around each of the absorbers found in the previous steps and found a third absorber at $z_{CIV} = 2.42670$, again with with $P(M_D) = 1.0$, $P(M_S) = 0.0$, and $P(M_N) = 0.0$. The fourth search, with regions around all three previous absorbers masked, found $P(M_D) = 0.27$, $P(M_S) = 0.0$ and $P(M_N) = 0.73$. Since the null model now had the largest model posterior, this was the final CIV absorber search in this spectrum (see Section 3.3.6).

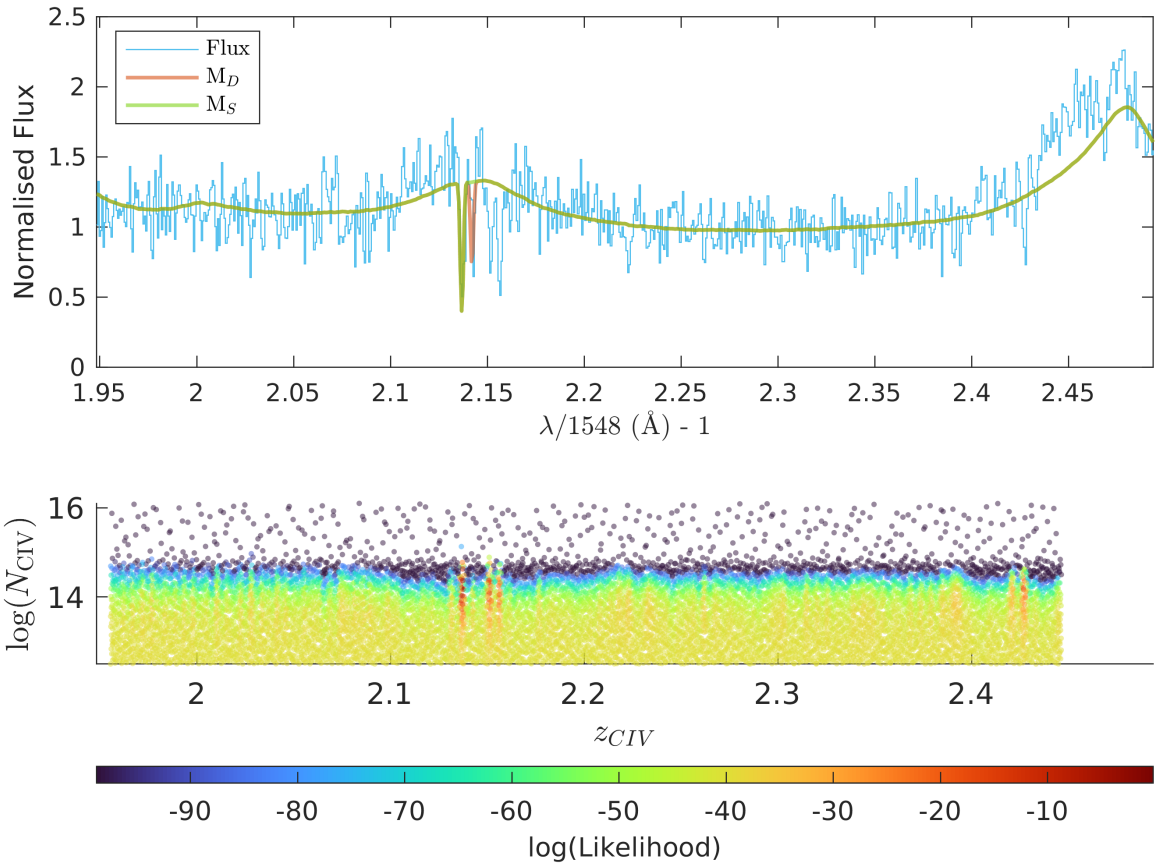

Figure 3.16: The first CIV search on QSO-56265-6151-936 (z_{QSO} = 2.4811). The upper panel shows the normalised flux (light blue), CIV model (M_D, red curve), and the single line model (M_S, green curve) as a function of CIV redshift. The lower panel shows the likelihood function value for M_D as a colour map for each of the 10,000 z_{CIV} samples (x-axis) and N_{CIV} samples. The third parameter (σ_{CIV}) is projected onto this 2D space. Our GP pipeline gives the following results for the first search: P(M_D)=1.00, z_{CIV}=2.13682±0.00049, log(N_{CIV})=14.42±0.20, σ_{CIV}=64.55±0.08 km s^{-1}, $W_{r,1548}$=0.568±0.372 Å, $W_{r,1550}$=0.072±0.386 Å.

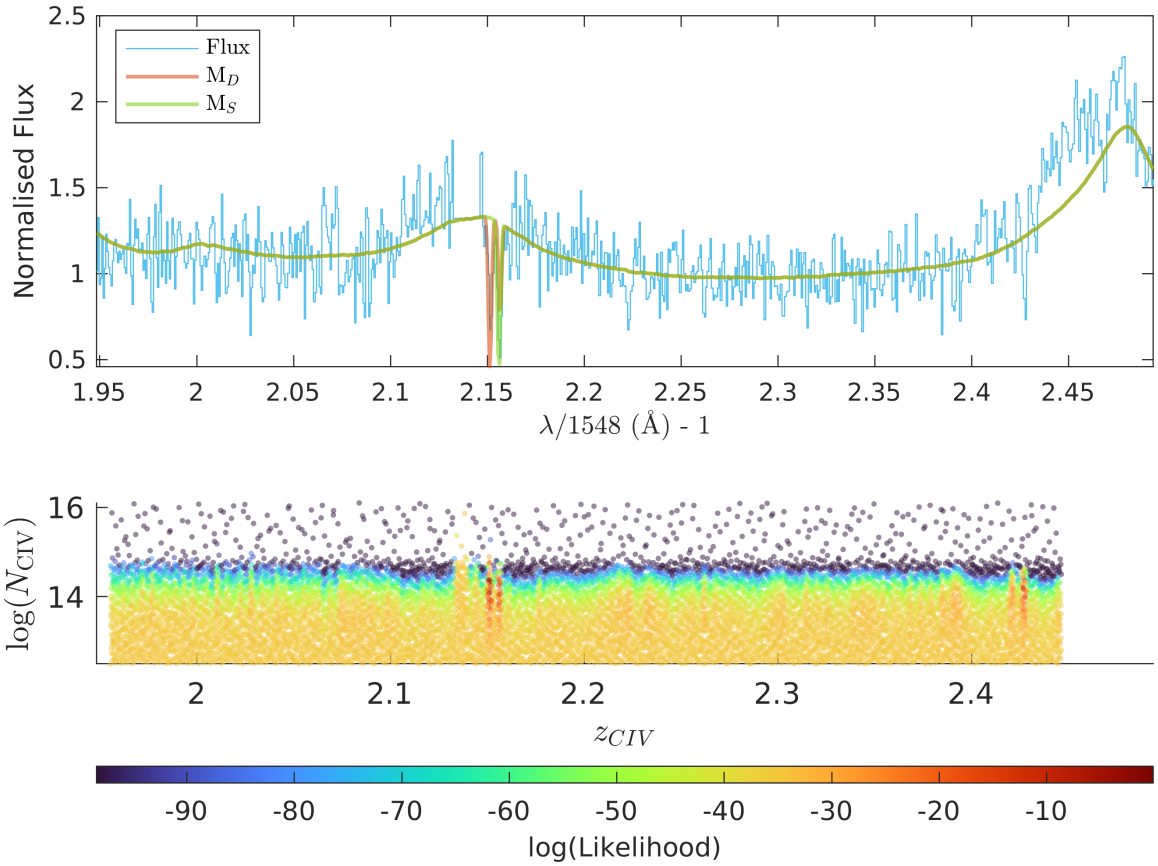

Figure 3.17: The second CIV search on QSO-56265-6151-936 (z_{QSO} = 2.4811). The upper panel is similar to Figure 3.16. However, we masked 350 km s^{-1}around the absorber found in the first CIV search at z_{CIV} = 2.13682. The lower panel shows the likelihood function values for M_D after masking the region around the absorber found in the first step. Our GP pipeline gives the following results for the second CIV search: P(M_D)=1.00, z_{CIV}=2.15132±0.00076, log(N_{CIV})=14.38±0.21, σ_{CIV}=64.55±0.08 km s^{-1}, $W_{r,1548}$=0.615±0.365 Å, $W_{r,1550}$=0.707±0.376 Å.

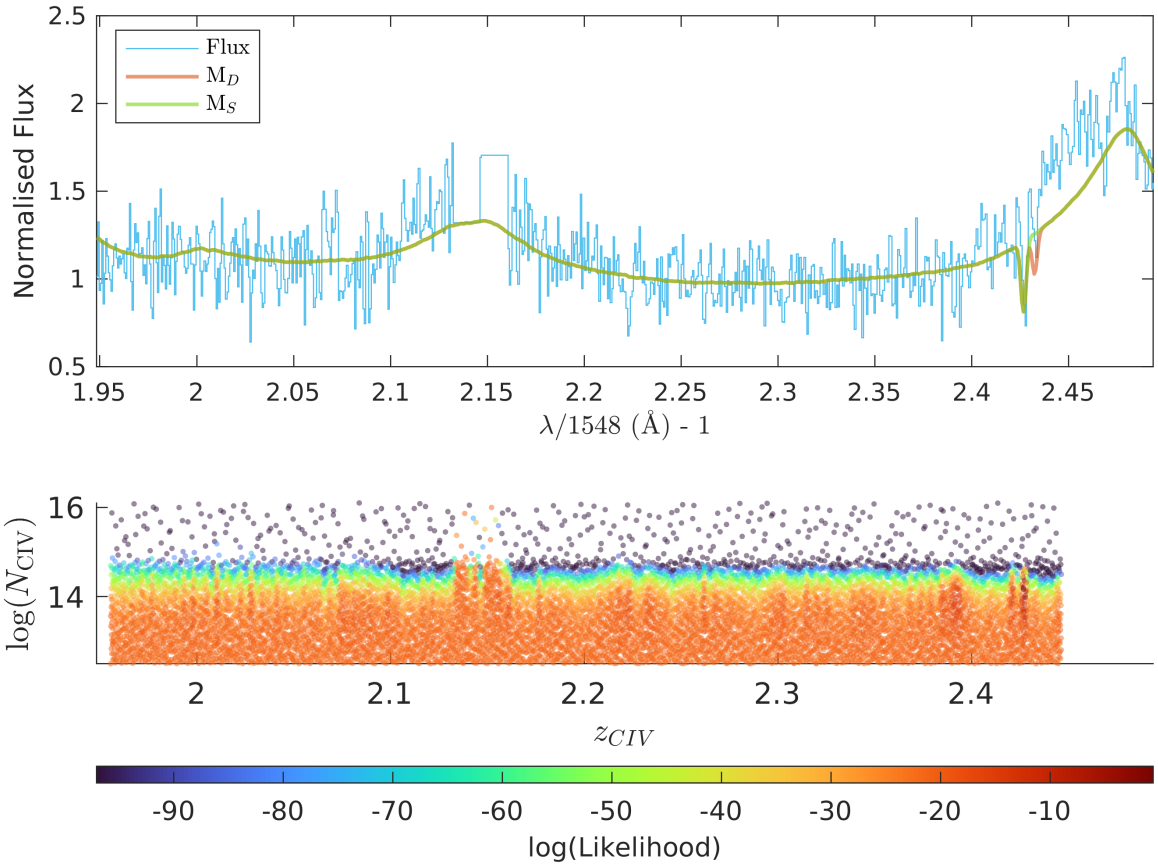

Figure 3.18: The third CIV search on QSO-56265-6151-936 (z_{QSO} = 2.4811). The upper panel is similar to Figure 3.17 but with 350 km s^{-1}around the two absorbers found in the first and second CIV searches at z_{CIV} = 2.13682 and z_{CIV} = 2.15132 masked. The lower panel shows the likelihood function value as a colour map after masking both absorbers. Our GP pipeline gives the following results for the third search: P(M_D)=1, z_{CIV}=2.42670±0.00006, log(N_{CIV})=14.17±0.02, σ_{CIV}=111.81±0.01 km s^{-1}, $W_{r,1548}$=0.164±0.407 Å, $W_{r,1550}$=-0.602±.396 Å.

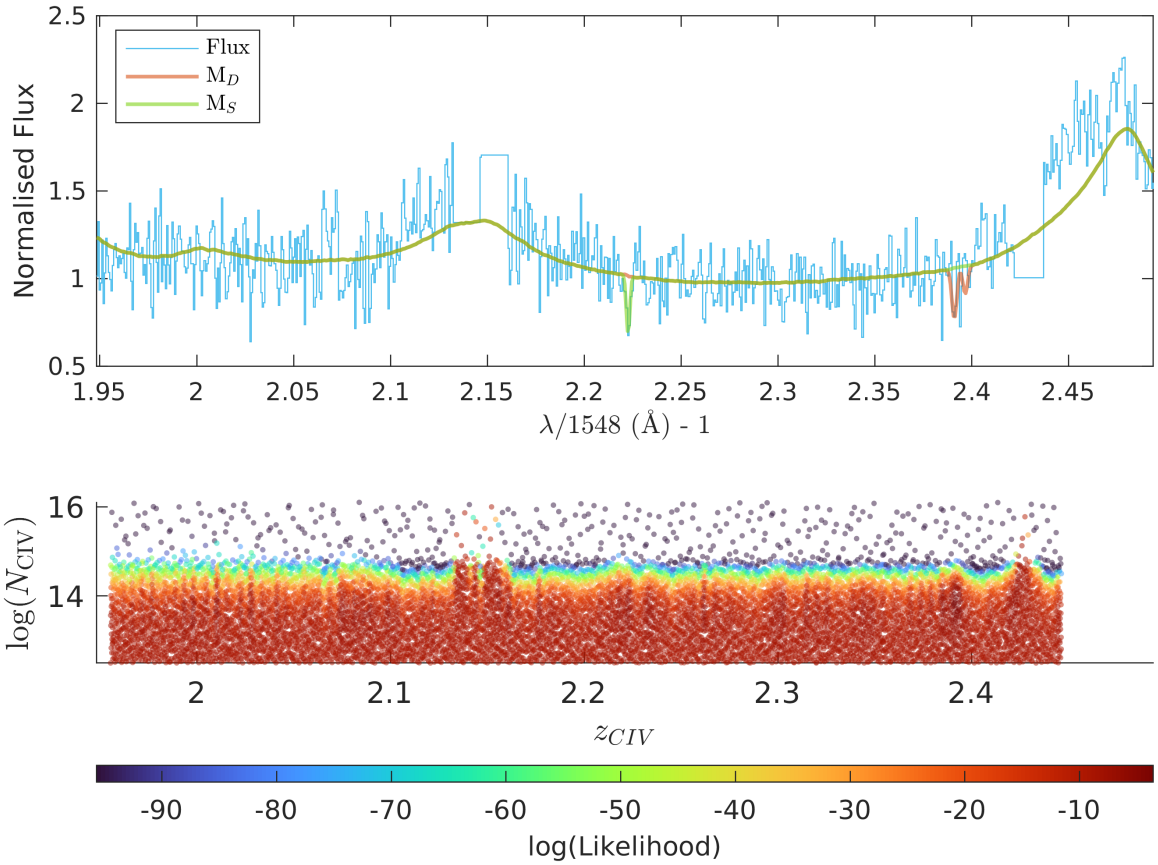

Figure 3.19: The fourth and final CIV search on QSO-56265-6151-936 ($z_{\rm QSO}$ = 2.4811). The upper panel is similar to Figure 3.18 but with 350 km s^{-1}around the absorbers found by the previous three searches masked. The lower panel shows the likelihood function value as a colour map for each of the 10,000 $z_{\rm CIV}$ samples (x-axis) and $N_{\rm CIV}$ samples. Our GP pipeline gives the following results for the final search: P(M$_{\rm D}$)=0.27, $z_{\rm CIV}$=2.39100± 0.00341, log($N_{\rm CIV}$)=14.04±0.82, $\sigma_{\rm CIV}$=105.99±0.56 km s^{-1}, W$_{r,1548}$=0.248±0.434, W$_{r,1550}$=-0.085±0.424. Note that since the highest probability in the fourth search was P(M$_{\rm N}$)=0.73, the algorithm performs no further searches (see Section 3.3.6).

Table 3.2 summarises the reported posterior probabilities for our catalogue. Around 66% of spectra have no CIV absorbers detectable at more than 85% confidence. Around 15% of spectra have one doublet and around 8% two doublets, each with a confidence more than 85%. The probability for detecting two independent absorbers in a spectrum is:

$$\mathrm{P}(2\ C\mathrm{IV}) = \mathrm{P}(1\ C\mathrm{IV}) \times \mathrm{P}(1\ C\mathrm{IV}) = 0.15^2 \sim 2.2\%. \tag{3.29}$$

The actual probability of two CIV absorbers in a spectrum is higher, $\sim$ 8%, demonstrating that absorbers are not independent but strongly correlated. Furthermore, we find five or more doublets at > 85% confidence in 3.1% of spectra.

We detected a single line absorber in $\sim$ 10% of sight-lines. If single line absorbers were independent, we would expect 0.1^2 or 1% of spectra to contain two singlet line absorbers. The actual probability of finding two singlets in a single sight-line was $\sim$ 2%, so the correlation between single line absorbers is much weaker than for CIV.

Figure 3.20 shows the distribution of (maximum *a posteriori*) absorber redshifts. There is a peak around $z \sim 2$, mirroring the distribution of quasar redshifts. Overall, we have detected an order of magnitude more absorbers than the PM catalogue, reflecting the larger size of our sight-line sample. There are 33 absorbers in DR12 with $\mathrm{P(M_D)} > 0.85$ and a redshift higher than 4.68, the maximum reported z_{CIV} in the PM catalogue.

In Figure 3.21 we show the distribution of maximum *a posteriori* Doppler velocity dispersion, σ_{CIV}, for the absorbers detected in the DR12 spectra. While the adopted prior for σ_{CIV} was a flat distribution between 35 $\mathrm{km\,s^{-1}}$and 115 $\mathrm{km\,s^{-1}}$, we see that the posterior distribution is moderately bimodal. The peak at larger σ_{CIV} values is connected with larger column densities. We

Table 3.2: The number of spectra containing different numbers of CIV absorbers for various doublet model probability thresholds, $P(M_D)$. The first column shows the number of CIV absorbers found within each spectrum (see Section 3.3.6) . The second through fourth columns show probability thresholds of> 65%, 85%, and 95% respectively. Cells show the number of quasar spectra falling in each category, together with the corresponding percentage of the 185,425 spectra in our SDSS DR12 sample.

N(CIV)	$P(M_D) > 0.65$	$P(M_D) > 0.85$	$P(M_D) > 0.95$
0	113142 61.0%	123994 66.8%	131767 71.0%
1	31163 16.8%	27733 14.9%	24981 13.5%
2	17526 9.4%	14533 7.8%	12777 6.9%
3	10020 5.4%	8424 4.5%	7218 3.9%
4	6112 3.3%	4960 2.6%	4176 2.3%
5	4155 2.2%	3342 1.8%	2656 1.4%
6	2426 1.3%	1771 0.9%	1348 0.7%
7	881 0.5%	668 0.4%	502 0.3%

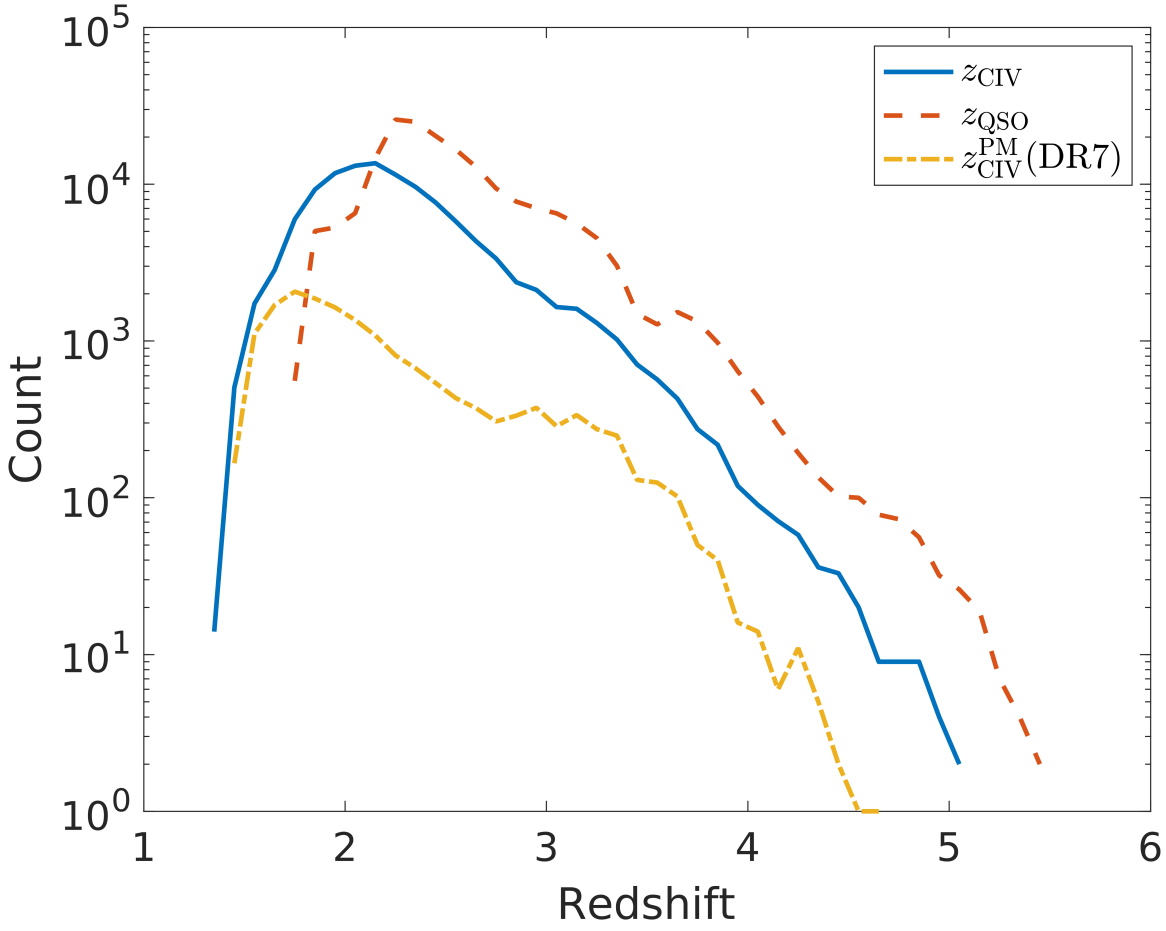

Figure 3.20: The redshift distributions of DR12 quasars (red dashed), high-probability ($P(M_D)\geq 0.95$) DR12 GP CIV absorbers (blue solid), and DR7 PM CIV absorbers with ranking≥2 (yellow dot-dashed). The quasar redshift is offset towards redder values than the absorber redshift, as expected, since absorbers cannot be more redshifted than the quasar. The GP catalogue finds absorbers outside of the absorber redshift range reported in the PM catalogue.

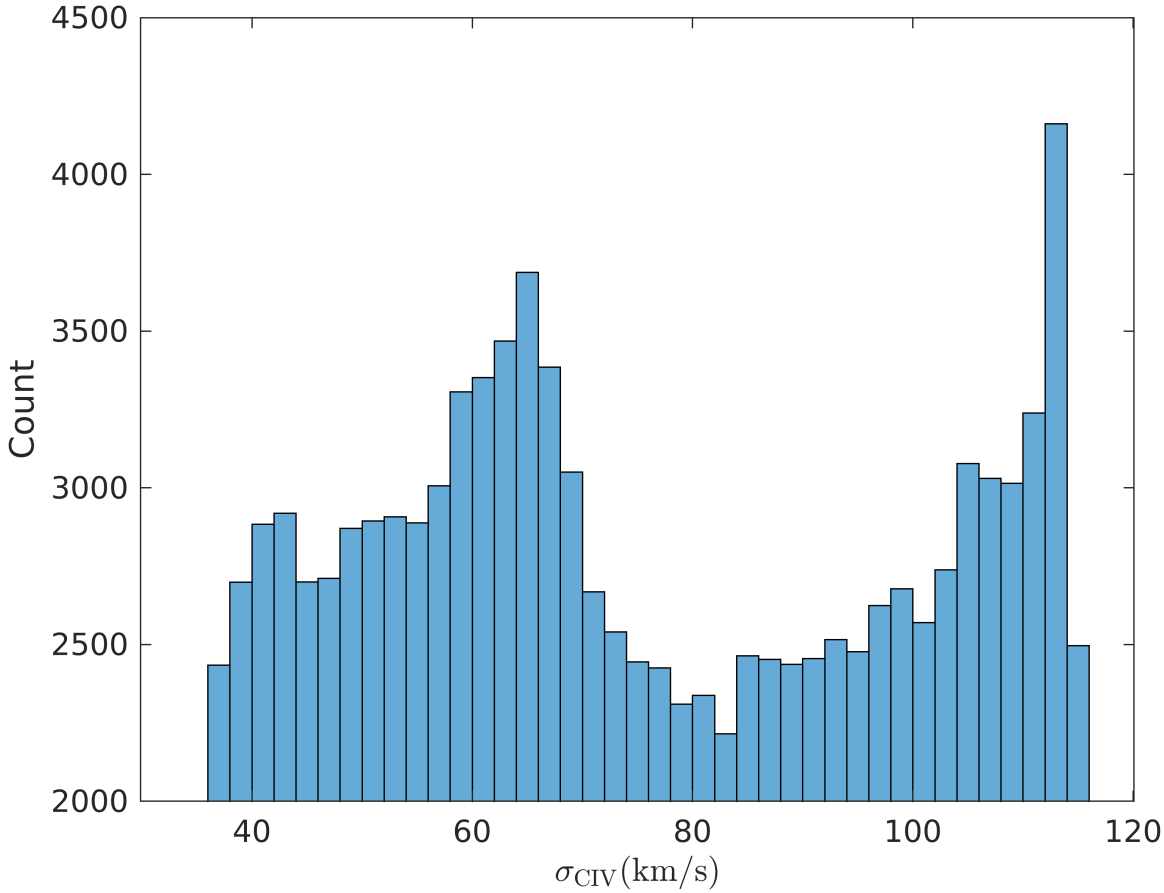

Figure 3.21: Distribution of the maximum *a posteriori* Doppler velocity dispersion values for absorbers detected in SDSS DR12 with P(M_D)≥ 0.95. Our prior distribution for Doppler velocity dispersion was uniform between 35 km s^{-1}and 115 km s^{-1}but the posterior distribution is bimodal. The larger σ_{CIV} posterior values are mostly associated with CIV absorbers found near low SNR pixels.

examined the spectra of detected absorbers with $\log N_{CIV} > 16$ and $\sigma_{CIV} > 110$ km s^{-1}. Most of these spectra are noisy and in some cases the line is heavily blended. The mean S/N around detected CIV absorbers with probability larger than 85% is 6.8 pix^{-1}, compared to a mean S/N of 2.5 pix^{-1} for spectra that contain absorbers with $\sigma_{CIV} > 80$ km s^{-1}and $\log N_{CIV} > 15$.

Looking at the distribution of each model posterior probability gives us an insight into how the spectra have been classified. Figure 3.22 shows the distribution doublet model posterior

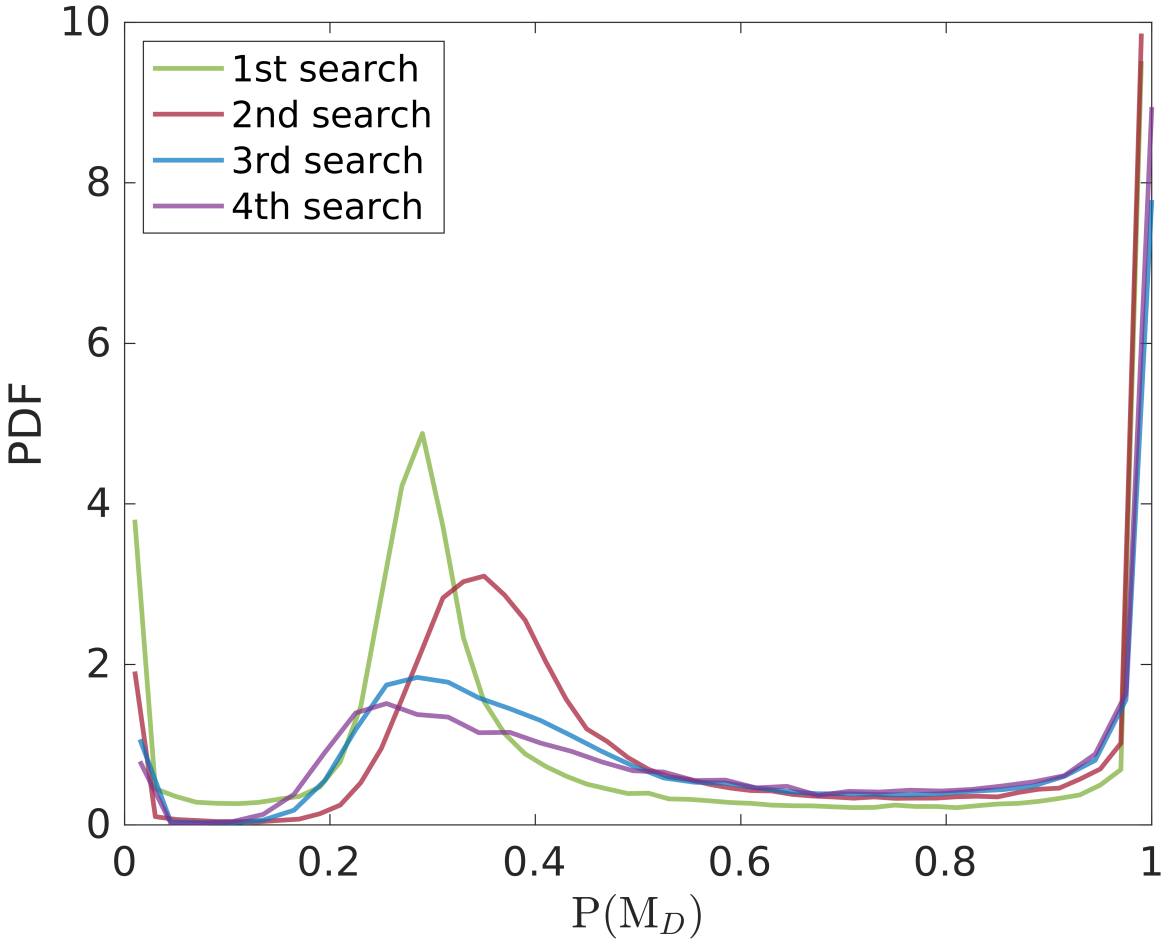

Figure 3.22: Distribution of $P(M_D)$ for the first to fourth searches. We do not show the fifth to seventh searches as they find very few absorbers (see Table 3.2). The peak around $P(M_D) \sim 0.3$ comes from low SNR spectra where the posterior probabilities of our three models are dominated by their priors.

probabilities for the first four CIV absorber searches in the DR12 spectra. For each search, we see a peak in the posterior probability distribution around 30%, stronger for earlier searches. We also examined $P(M_S)$ and $P(M_N)$ for the first search and found that many of these ambiguous CIV absorbers also have $P(M_S) \sim 0.3$. In addition, these absorbers are generally in lower S/N spectra. Thus this peak occurs when the spectra are weakly constraining and the posterior absorber probability is dominated by the model priors.

Figure 3.23 shows the 1548 Å rest equivalent widths from our SDSS DR12 catalogue. We use rest equivalent widths derived from the parameters of the Voigt profile doublet model, $W_{r,1548}^{\mathrm{GP,Voigt}}$, which are computed using:

$$W_{r,1548}^{\mathrm{GP,Voigt}} = \int (1 - a_{1548}) d\lambda. \tag{3.30}$$

Here a_{1548} is the absorption function for our 1548 Å line, and we compute the rest equivalent width from the maximum *a posteriori* values of $z_{C\mathrm{IV}}$, $N_{C\mathrm{IV}}$, and $\sigma_{C\mathrm{IV}}$ under $\mathrm{M_D}$. Figure 3.23 also shows the DR7 PM catalogue rest equivalent width distribution for comparison. The larger sample of spectra in SDSS DR12 allows us to probe higher rest equivalent widths, with 110 absorbers at larger rest equivalent widths than 3.15 Å, the highest rest equivalent width in the PM catalogue. Note however, that our GP pipeline contains models only for singlet and doublet absorbers, and so may mis-classify blended or mini-BAL systems, which do not strongly resemble the models it uses.

Figure 3.24 shows the distribution of the doublet ratio, $\mathrm{W}_{r,1548}^{\mathrm{GP,Voigt}}/\mathrm{W}_{r,1550}^{\mathrm{GP,Voigt}}$, obtained by integrating the Voigt profile at the maximum *a posteriori* values of $z_{C\mathrm{IV}}$, $N_{C\mathrm{IV}}$, and $\sigma_{C\mathrm{IV}}$. The doublet ratio varies from two on the linear portion of the curve of growth to one when both lines are saturated. The DR12 distribution indicates the vast majority of doublets are moderately saturated (i.e., ratio less than two), with a subset of very strong (i.e., saturated) doublets with a ratio of unity.

3.6 Conclusions

We trained a quasar continuum model to detect CIV absorbers using a Gaussian process. The training was done on a sample of DR7 spectra which were labelled as CIV free in the Precious Metals catalogue of Cooksey et al. (2013). We used Bayesian model selection to compare our

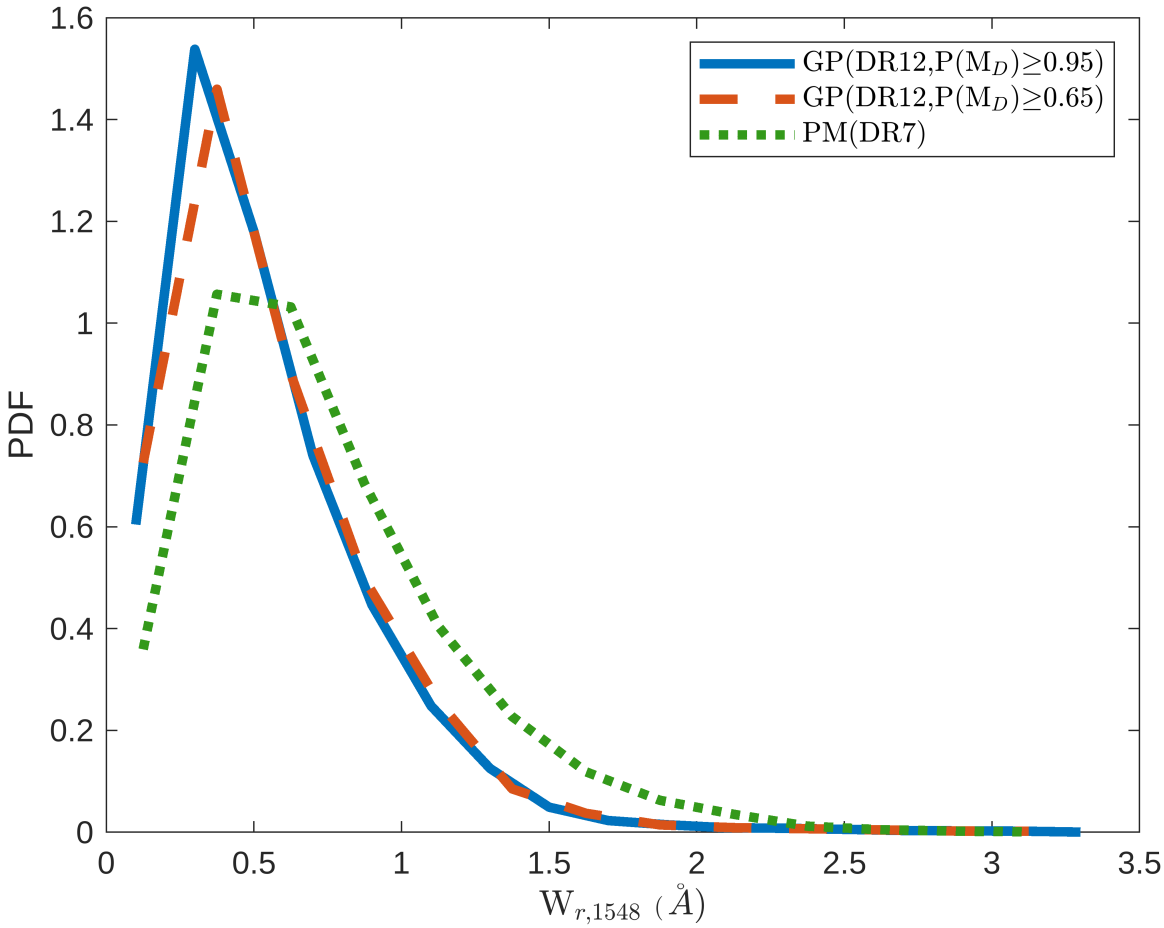

Figure 3.23: The distribution of the rest equivalent width of the 1548 Å ($W_{r,1548}^{GP,Voigt}$) line obtained by Voigt profile integration (Equation 3.30). We show $W_{r,1548}^{GP,Voigt}$ for all detected DR12 absorbers with $P(M_D) \geq 0.95$ (blue curve) and $P(M_D) \geq 0.65$ (dashed red curve). We also show for comparison $W_{r,1548}^{PM}$ values from the PM (DR7) catalogue in the dotted green curve.

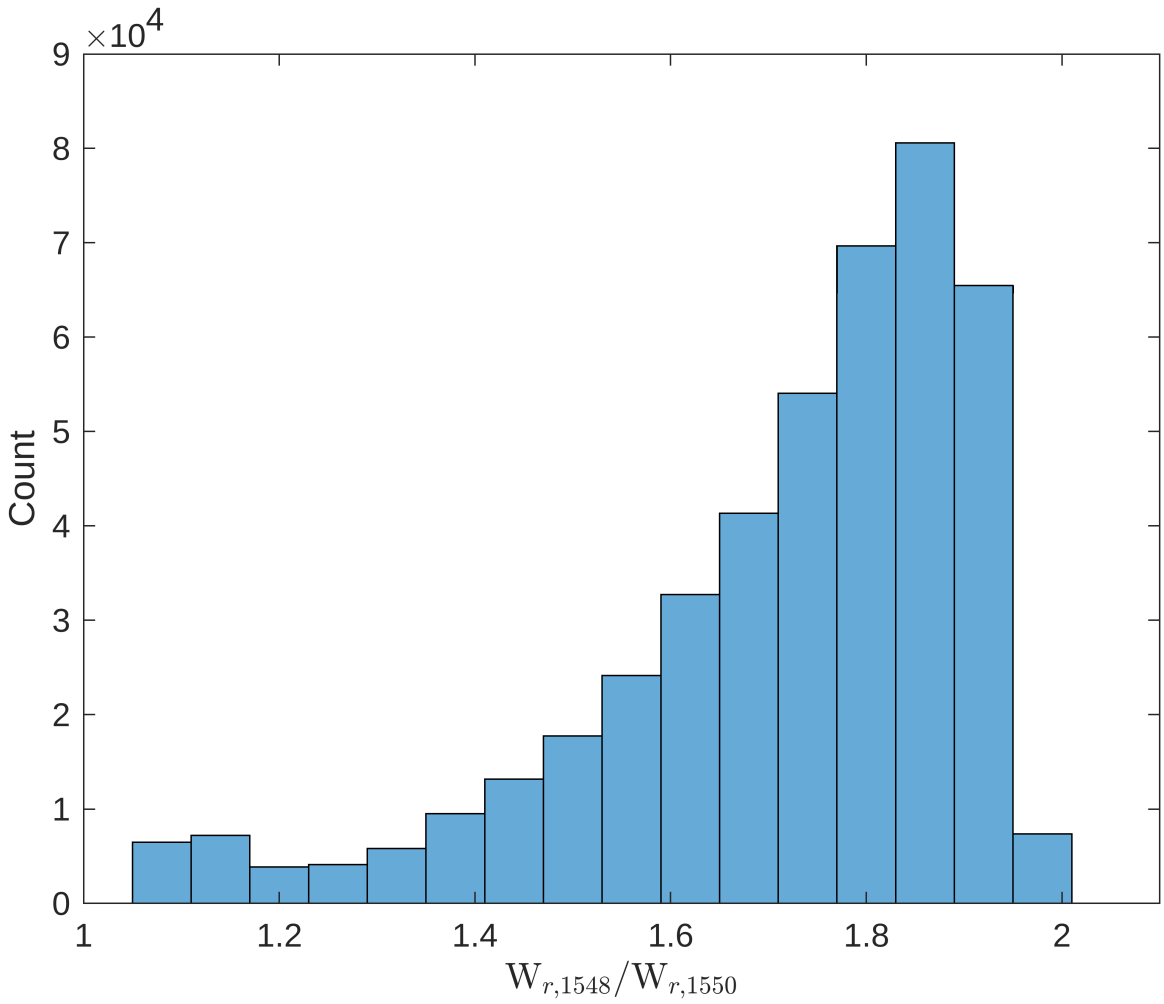

Figure 3.24: The distribution of the doublet ratio, $W^{GP,Voigt}_{r,1548}/W^{GP,Voigt}_{r,1550}$, both measured by GP pipeline in SDSS DR12 according to Equation 3.30 for those detected absorbers with $P(M_D) \geq 95\%$. The rest equivalent widths values are calculated based on our Voigt profile integration. The distribution of doublet ratio is in agreement with the theoretical range of 1–2. The existence of a sub-population of absorbers with saturated lines (doublet ratio ~ 1) are obvious.

continuum model to models containing one to seven CIV doublets. We added an extra model for single line absorbers, to avoid confusion from interloping metal lines. The prior distribution was taken from our training catalogue and flat parameter priors were used for the CIV redshift, z_{CIV} and Doppler velocity dispersion σ_{CIV}. We searched for up to 7 absorbers in each tested spectrum and provide a comprehensive catalogue containing CIV detection probability, as well as maximum a posteriori values and credible intervals for z_{QSO}, N_{CIV} and σ_{CIV}. We validated our pipeline by applying it to a hold-out sample of 1301 spectra from the PM catalogue. Our pipeline produced similar results to the PM catalogue and has good purity and completeness. Generally the two catalogues produced similar CIV redshifts and rest equivalent widths.

Thus validated, we applied our model to SDSS DR12, and produced the largest CIV absorption catalogue yet seen. Among the total 185,425 selected quasar spectra in SDSS DR12, we found 113,775 CIV doublets with $> 95\%$ confidence. Note that the user may pick the desired confidence threshold in our catalogue, thanks to our reported posterior probabilities for each absorber. We detected CIV absorption up to $z \sim 5$, including 33 systems at higher redshift than seen in DR7. We also detect 110 absorbers in DR12 with a rest equivalent width larger than the maximum in the DR7 catalogue.

Our method is good for detecting unblended CIV absorbers. However, when absorption systems are complex and blended, the line may be a poor match to both the singlet and doublet models. In these cases, our pipeline is sometimes unable to distinguish between genuine CIV doublet absorption and other interloper metal lines.

Potential applications of our catalogue include: 1) finding targets for high-resolution follow-up of complex CIV systems (e.g. Galbiati et al. 2023) 2) Cross-matching with galaxy

catalogues to find the properties of the galactic circumgalactic medium within which our CIV absorbers lie. 3) Cross-matching with a Damped Lyman-α catalogue to investigate the relationship between the highly ionised carbon and neutral hydrogen in the circumgalactic medium.

Finally, the statistical properties of our catalogue can be computed and compared to the outputs of cosmological simulations to test and improve models for galactic feedback.

Our technique can also be applied to later, larger quasar catalogues such as those from the SDSS DR16 and the upcoming Dark Energy Spectroscopic Instrument quasar survey.

3.7 Data availability

All of our codes are available publicly in GitHub [18] and our final catalogue can be found in Zenodo. [19]

3.8 Appendix

As a consistency check, we compared the rest equivalent width from the maximum *a posteriori* values of our model fit, using Equation 3.30, to the rest equivalent width from integrating the flux around the absorber, as in the validation phase. Figure 3.25 shows the difference between the two rest equivalent width estimates, normalised by the error estimate. Figure 3.25 shows a symmetric unit Gaussian distribution centered at zero, demonstrating that our model parameters are both approximately unbiased and have well-calibrated error estimates.

[18] https://github.com/rezamonadi/GaussianProcessCIV

[19] https://doi.org/10.5281/zenodo.7872725

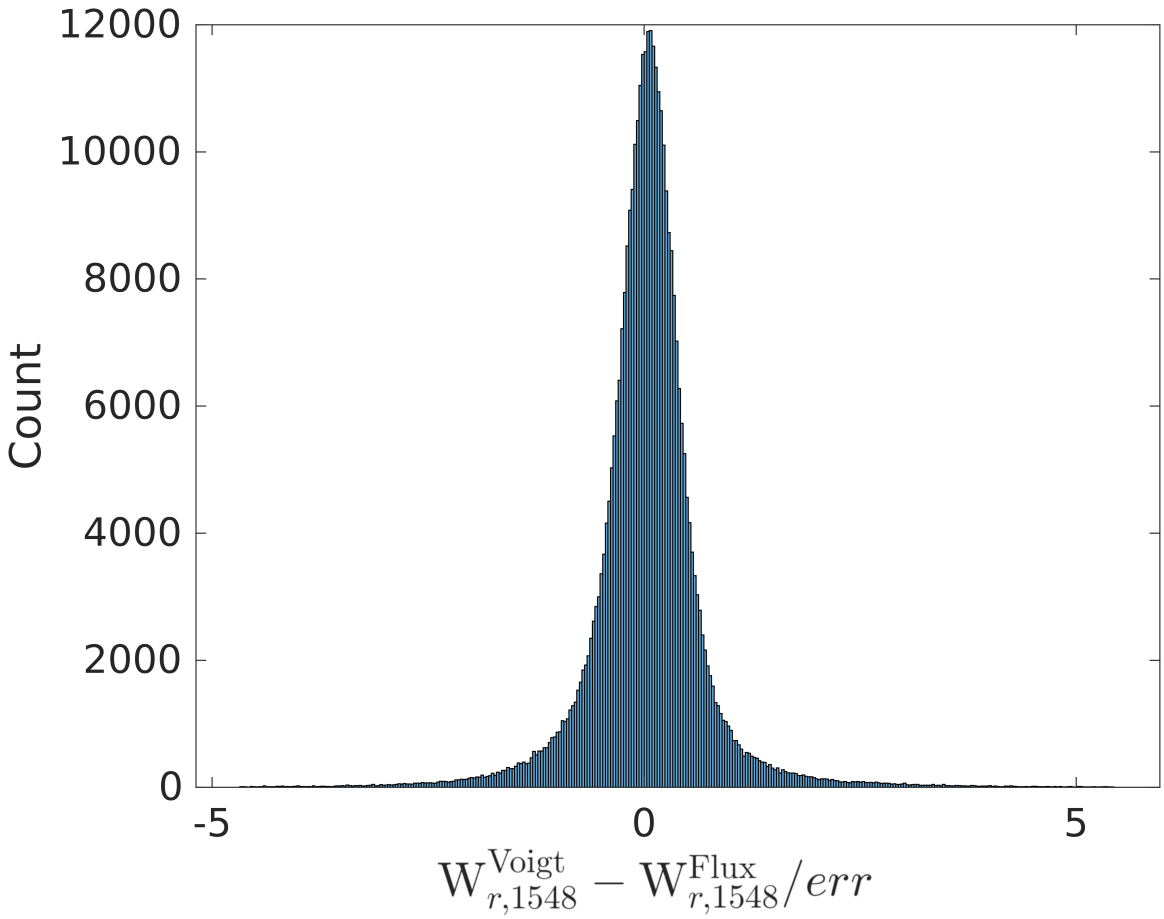

Figure 3.25: The difference between the two rest equivalent width estimates for the 1548 Å line explained in the text. These are using the maximum *a posteriori* model parameters and integrating the flux around the detected CIVabsorber. Differences are normalised by the expected error from the model parameter posteriors, and show the expected Gaussian distribution.

Table 3.3: Table of probabilities P(M_S): first column shows the number of single line absorbers. The 2nd through 4th columns show the number of single absorbers with probabilities > 65%, 85%, and 95% respectively.

CIV	P(M_S) > 0.65	P(M_S) > 0.85	P(M_S) > 0.95
0	155905 (84.0%)	162533 (87.6%)	166675 (89.9%)
1	25441 (13.7%)	19159 (10.3%)	15366 (8.3%)
2	3210 (1.7%)	2914 (1.6%)	2626 (1.4%)
3	675 (0.4%)	637 (0.3%)	583 (0.3%)
4	161 (0.09%)	152 (0.08%)	147 (0.08%)
5	29 (0.02%)	26 (0.01%)	24 (0.01%)
6	4 (0.00%)	4 (0.00%)	4 (0.00%)

In the validation phase (see Section 3.4.4), we calculated the rest equivalent width by integrating the flux around the absorber, in order to compare to the rest equivalent widths from the PM catalogue. However, we prefer to estimate rest equivalent widths for our SDSS DR12 catalogue directly from our maximum *a posteriori* model parameters, as these are less sensitive to noisy pixels in the integration range.

Table 3.3 shows the number of candidate absorbers for the single line absorber model in SDSS DR12. Note that our training set does not label these absorbers and so we have not validated the potential detections.

3.9 Acknowledgement

R.M. was supported by Higher Education Emergency Relief Funds. RM thanks Fred Hamann for supporting him for part of this work from NSF grant AST-1911066.S.B. SB was supported by NASA ATP 80NSSC22K1897. We used the HPCC cluster at UC Riverside and AWS credits provided under an amazon machine learning research award. M.H. was supported by NASA FINESST grant 80NSSC21K1840. KLC acknowledges partial support from NSF AST-1615296.

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

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

Chapter 4

Conclusions and Future Paths

We used machine learning methods to extract information from the observed spectra of quasars in the SDSS database. As the data volume and the number of astronomical objects we are observing are growing exponentially, we need automatic methods to do science easier and more feasible.

In the first project, we could define an efficient and robust method to find some needles in a haystack by harnessing the power of unsupervised machine learning. We went beyond intuitive thresholds in the parameter space of the measured properties of quasars with the help of combining local outlier factor analysis and kernel density estimation. Our method can be adapted to find a small cluster of data in a region of parameter space that is sparsely occupied by data points. This was very challenging. By letting the data speak themselves, we could capture more extremely red quasars with outstanding outflow behaviours. This project help to learn about red quasars which are very important in understanding about quasars life in general and their interactions with galaxies to shed light on galaxy evolution problem.

One of the future directions can be using dimensionality reduction methods, such as t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten & Hinton 2008) to visualize the data with a large number of dimensions to a 2-D or 3-D space. These dimensions can be any measured properties of quasars from spectroscopy (eg. the FWHM of CIV line) and any one from photometric data (eg. colors). Then, we can look at the clusters that naturally show up in the reduced and mapped space. This is interesting because we may serendipitously find a group of special astronomical objects.

The second project we obtained the largest CIV absorption catalog to date from SDSS quasar spectra. It was an example of using supervised machine learning in a task that traditionally trained observers would do by spending some amount of time. We trained a quasar continuum function with the help of Gaussian Processes and used Bayesian model selection to decide if there are absorbers within the quasar spectrum. Our method can process each absorption system in around 10 seconds when we use 32 CPU cores in UCR's High Performance Computer Cluster. This is significantly faster than visual inspection, considering that the inspector needs some breaks and her performance change during the time! Our pipeline also provides a dynamic catalog so that the user has the freedom to choose a level of confidence for the absorption systems that one wants to study. Moreover, we provided a posterior distribution for the physical properties of absorbers in a quasar spectrum, namely column density, velocity dispersion, and absorber redshift. These posterior distributions can be used to infer the most likely measurements and their corresponding credible intervals. Our method is specifically effective to obtain information even from the noisy spectra. Note that DESI decided to use a Gaussian Processes detection method, similar to what our group did in Ho et al. (2021), for detecting Damped Lyman-α systems (Karaçaylı et al. 2023). Having metal

lines catalog like what we made here is very important in understanding the distribution of heavy elements in cosmos. With a good understanding about metals in cosmos, we can put constraints on the galaxy evolution models that predict their production and propagation.

There are different paths for future work in this regard. We are working on a similar catalog for MgII absorption lines. Cross-correlating metal absorbers, especially when the number statistics is large, can tell us a lot about the circumgalactic medium and galaxy evolution in a bigger picture. Moreover, we are planning to use our validated method to make our final catalog of CIV absorbers using the latest quasar catalog of SDSS DR16. Another possibility is looking at the clustering of CIV absorbers to constrain the distribution and evolution of carbon in the cosmos. Having a large catalog of Damped Lyman-α absorbers from our group (Ho et al. 2021), we also can look at the ratio of Hydrogen and Carbon during cosmic time. Our catalogue can guide high-resolution follow-up observations and may be cross-matched with galaxy catalogues or other quasar absorption line catalogues to investigate the properties of the circumgalactic medium.

Bibliography

Adelberger, K. L., Shapley, A. E., Steidel, C. C., et al. 2005, The Astrophysical Journal, 629, 636, doi: 10.1086/431753

Adelman-McCarthy, J. K., Agüeros, M. A., Allam, S. S., et al. 2008, apjs, 175, 297, doi: 10.1086/524984

Aigrain, S., & Foreman-Mackey, D. 2023, Annual Review of Astronomy and Astrophysics, 61, null, doi: 10.1146/annurev-astro-052920-103508

Alam, S., Albareti, F. D., Prieto, C. A., et al. 2015, The Astrophysical Journal Supplement Series, 219, 12, doi: 10.1088/0067-0049/219/1/12

Antonucci, R. 1993, araa, 31, 473, doi: 10.1146/annurev.aa.31.090193.002353

Appleby, S., Davé, R., Sorini, D., Cui, W., & Christiansen, J. 2023, Monthly Notices of the Royal Astronomical Society, 519, 5514, doi: 10.1093/mnras/stad025

Aracil, B., Petitjean, P., Pichon, C., & Bergeron, J. 2004, aap, 419, 811, doi: 10.1051/0004-6361:20034346

Assef, R. J., Stern, D., Noirot, G., et al. 2018, apjs, 234, 23, doi: 10.3847/1538-4365/aaa00a

Assef, R. J., Eisenhardt, P. R. M., Stern, D., et al. 2015, apj, 804, 27, doi: 10.1088/0004-637X/804/1/27

Azadi, M., Aird, J., Coil, A. L., et al. 2015, apj, 806, 187, doi: 10.1088/0004-637X/806/2/187

Baldwin, J. A. 1977, apj, 214, 679, doi: 10.1086/155294

Banerji, M., Alaghband-Zadeh, S., Hewett, P. C., & McMahon, R. G. 2015, Monthly Notices of the Royal Astronomical Society, 447, 3368, doi: 10.1093/mnras/stu2649

Banerji, M., McMahon, R. G., Hewett, P. C., Gonzalez-Solares, E., & Koposov, S. E. 2013, mnras, 429, L55, doi: 10.1093/mnrasl/sls023

Becker, G. D., Rauch, M., & Sargent, W. L. W. 2009, The Astrophysical Journal, 698, 1010, doi: 10.1088/0004-637X/698/2/1010

Bird, S., Rubin, K. H. R., Suresh, J., & Hernquist, L. 2016, mnras, 462, 307, doi: 10.1093/mnras/stw1582

Boksenberg, A., & Sargent, W. L. W. 2015, apjs, 218, 7, doi: 10.1088/0067-0049/218/1/7

Boksenberg, A., Sargent, W. L. W., & Rauch, M. 2003, arXiv e-prints, astro, doi: 10.48550/arXiv.astro-ph/0307557

Bordoloi, R., Tumlinson, J., Werk, J. K., et al. 2014, The Astrophysical Journal, 796, 136, doi: 10.1088/0004-637X/796/2/136

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. 2000, 93

Burchett, J. N., Tripp, T. M., Prochaska, J. X., et al. 2015, apj, 815, 91, doi: 10.1088/0004-637X/815/2/91

Burchett, J. N., Tripp, T. M., Bordoloi, R., et al. 2016, apj, 832, 124, doi: 10.3847/0004-637X/832/2/124

Calistro Rivera, G., Alexander, D. M., Rosario, D. J., et al. 2021, AandA, 649, A102, doi: 10.1051/0004-6361/202040214

Carroll, B. W., & Ostlie, D. A. 2017, An Introduction to Modern Astrophysics, 2nd edn. (Cambridge University Press), doi: 10.1017/9781108380980

Chen, H.-W., Lanzetta, K. M., & Webb, J. K. 2001, apj, 556, 158, doi: 10.1086/321537

Chen, Z.-F., Qin, Y.-P., Pan, C.-J., et al. 2014, apjs, 210, 7, doi: 10.1088/0067-0049/210/1/7

Churchill, C. W. 2020, Cosmological absorption line spectroscopy. http://astronomy.nmsu.edu/cwc/CWC/CUP/book.pdf

Codoreanu, A., Ryan-Weber, E. V., García, L. Á., et al. 2018, mnras, 481, 4940, doi: 10.1093/mnras/sty2576

Cooksey, K. L., Kao, M. M., Simcoe, R. A., O'Meara, J. M., & Prochaska, J. X. 2013, apj, 763, 37, doi: 10.1088/0004-637X/763/1/37

Cooksey, K. L., Thom, C., Prochaska, J. X., & Chen, H.-W. 2009, The Astrophysical Journal, 708, 868, doi: 10.1088/0004-637X/708/1/868

Cooper, T. J., Simcoe, R. A., Cooksey, K. L., et al. 2019, apj, 882, 77, doi: 10.3847/1538-4357/ab3402

Davies, R. L., Ryan-Weber, E., D'Odorico, V., et al. 2023, mnras, 521, 314, doi: 10.1093/mnras/stad294

Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2013, aj, 145, 10, doi: 10.1088/0004-6256/145/1/10

Day, W. H., & Edelsbrunner, H. 1984, Journal of classification, 1, 7

DESI Collaboration, Aghamousa, A., Aguilar, J., et al. 2016, arXiv e-prints, arXiv:1611.00036, doi: 10.48550/arXiv.1611.00036

D'Odorico, V., Calura, F., Cristiani, S., & Viel, M. 2010, Monthly Notices of the Royal Astronomical Society, 401, 2715, doi: 10.1111/j.1365-2966.2009.15856.x

D'Odorico, V., Cupani, G., Cristiani, S., et al. 2013, mnras, 435, 1198, doi: 10.1093/mnras/stt1365

Doughty, C. C., & Finlator, K. M. 2023, mnras, 518, 4159, doi: 10.1093/mnras/stac3342

Draine, B. T. 2011, Physics of the interstellar and intergalactic medium

Ebden, M. 2015, arXiv e-prints, arXiv:1505.02965, doi: 10.48550/arXiv.1505.02965

Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011a, aj, 142, 72, doi: 10.1088/0004-6256/142/3/72

—. 2011b, aj, 142, 72, doi: 10.1088/0004-6256/142/3/72

Ellison, S. L., Songaila, A., Schaye, J., & Pettini, M. 2000, The Astronomical Journal, 120, 1175, doi: 10.1086/301511

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. 1996, 96, 226

Feng, W.-X., Yu, H.-B., & Zhong, Y.-M. 2021, apjl, 914, L26, doi: 10.3847/2041-8213/ac04b0

Flesch, E. W. 2021, arXiv e-prints, arXiv:2105.12985, doi: 10.48550/arXiv.2105.12985

Galbiati, M., Fumagalli, M., Fossati, M., et al. 2023, mnras, doi: 10.1093/mnras/stad2087

Garnett, R., Ho, S., Bird, S., & Schneider, J. 2017, mnras, 472, 1850, doi: 10.1093/mnras/stx1958

Gebhardt, K., Bender, R., Bower, G., et al. 2000, apjl, 539, L13, doi: 10.1086/312840

Glikman, E., Simmons, B., Mailly, M., et al. 2015, apj, 806, 218, doi: 10.1088/0004-637X/806/2/218

Golovich, N., Dawson, W., Bartolić, F., et al. 2022, apjs, 260, 2, doi: 10.3847/1538-4365/ac5969

Graham, A. W. 2016, in Astrophysics and Space Science Library, Vol. 418, Galactic Bulges, ed. E. Laurikainen, R. Peletier, & D. Gadotti, 263, doi: 10.1007/978-3-319-19378-6_11

Haehnelt, M. G., Steinmetz, M., & Rauch, M. 1996, apjl, 465, L95, doi: 10.1086/310156

Hamann, F., Zakamska, N. L., Ross, N., et al. 2016, Monthly Notices of the Royal Astronomical Society, 464, 3431, doi: 10.1093/mnras/stw2387

Hasan, F., Churchill, C. W., Stemock, B., et al. 2022, The Astrophysical Journal, 924, 12, doi: 10.3847/1538-4357/ac308c

—. 2020, The Astrophysical Journal, 904, 44, doi: 10.3847/1538-4357/abbe0b

Heitmann, K., Higdon, D., Nakhleh, C., & Habib, S. 2006, apjl, 646, L1, doi: 10.1086/506448

Heitmann, K., Higdon, D., White, M., et al. 2009, apj, 705, 156, doi: 10.1088/0004-637X/705/1/156

Hickox, R. C., & Alexander, D. M. 2018, araa, 56, 625, doi: 10.1146/annurev-astro-081817-051803

Hill, G. J., Gebhardt, K., Komatsu, E., et al. 2008, 399, 115, doi: 10.48550/arXiv.0806.0183

Hiner, K. D., Canalizo, G., Lacy, M., et al. 2009, apj, 706, 508, doi: 10.1088/0004-637X/706/1/508

Ho, M.-F., Bird, S., Fernandez, M. A., & Shelton, C. R. 2023, arXiv e-prints, arXiv:2306.03144, doi: 10.48550/arXiv.2306.03144

Ho, M.-F., Bird, S., & Garnett, R. 2020, mnras, 496, 5436, doi: 10.1093/mnras/staa1806

—. 2021, mnras, 507, 704, doi: 10.1093/mnras/stab2169

Hopkins, P. F., Hernquist, L., Cox, T. J., & Kereš, D. 2008, The Astrophysical Journal Supplement Series, 175, 356, doi: 10.1086/524362

Hopkins, P. F., Hernquist, L., Cox, T. J., et al. 2005, The Astrophysical Journal, 630, 705, doi: 10.1086/432438

Ishibashi, W., & Fabian, A. C. 2016, Monthly Notices of the Royal Astronomical Society, 463, 1291, doi: 10.1093/mnras/stw2063

Karaçaylı, N. G., Martini, P., Guy, J., et al. 2023, arXiv e-prints, arXiv:2306.06316, doi: 10.48550/arXiv.2306.06316

Kim, D., & Im, M. 2018, aap, 610, A31, doi: 10.1051/0004-6361/201731963

Klindt, L., Alexander, D. M., Rosario, D. J., Lusso, E., & Fotopoulou, S. 2019, mnras, 488, 3109, doi: 10.1093/mnras/stz1771

Kormendy, J., & Ho, L. C. 2013, araa, 51, 511, doi: 10.1146/annurev-astro-082708-101811

Kroupa, P., Subr, L., Jerabkova, T., & Wang, L. 2020, mnras, 498, 5652, doi: 10.1093/mnras/staa2276

Krumholz, M. R., & Burkhart, B. 2016, mnras, 458, 1671, doi: 10.1093/mnras/stw434

Marziani, P., Dultzin, D., Sulentic, J. W., et al. 2018, Frontiers in Astronomy and Space Sciences, 5, doi: 10.3389/fspas.2018.00006

Monadi, R., & Bird, S. 2022a, mnras, 511, 3501, doi: 10.1093/mnras/stac294

—. 2022b, mnras, 511, 3501, doi: 10.1093/mnras/stac294

Monadi, R., Ho, M.-F., Cooksey, K. L., & Bird, S. 2023, arXiv e-prints, arXiv:2305.00023, doi: 10.48550/arXiv.2305.00023

Moriwaki, K., Nishimichi, T., & Yoshida, N. 2023, Reports on Progress in Physics, 86, 076901, doi: 10.1088/1361-6633/acd2ea

Panda, S., Czerny, B., & Wildy, C. 2017, Frontiers in Astronomy and Space Sciences, 4, doi: 10.3389/fspas.2017.00033

Pearson, W. J., Wang, L., Alpaslan, M., et al. 2019, aap, 631, A51, doi: 10.1051/0004-6361/201936337

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, the Journal of machine Learning research, 12, 2825

Péroux, C., & Howk, J. C. 2020, araa, 58, 363, doi: 10.1146/annurev-astro-021820-120014

Perrotta, S., Hamann, F., Zakamska, N. L., et al. 2019, mnras, 488, 4126, doi: 10.1093/mnras/stz1993

Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2014, aap, 571, A16, doi: 10.1051/0004-6361/201321591

Rasmussen C. E., W. C. K. I. 2006, Gaussian Processes for Machine Learning (MIT Press, Cambridge, MA)

Rauch, M., Sargent, W. L. W., Womble, D. S., & Barlow, T. A. 1996, The Astrophysical Journal, 467, L5, doi: 10.1086/310187

Richards, G. T., Fan, X., Schneider, D. P., et al. 2001, aj, 121, 2308, doi: 10.1086/320392

Richards, G. T. e. a. 2003, aj, 126, 1131, doi: 10.1086/377014

Ross, N. P., Myers, A. D., Sheldon, E. S., et al. 2012, apjs, 199, 3, doi: 10.1088/0067-0049/199/1/3

Ross, N. P., Hamann, F., Zakamska, N. L., et al. 2015, Monthly Notices of the Royal Astronomical Society, 453, 3932, doi: 10.1093/mnras/stv1710

Rubin, K. H. R., Hennawi, J. F., Prochaska, J. X., et al. 2015, apj, 808, 38, doi: 10.1088/0004-637X/808/1/38

Ryan-Weber, E. V., Pettini, M., Madau, P., & Zych, B. J. 2009, Monthly Notices of the Royal Astronomical Society, 395, 1476, doi: 10.1111/j.1365-2966.2009.14618.x

Sanders, D. B., Soifer, B. T., Elias, J. H., et al. 1988, apj, 325, 74, doi: 10.1086/165983

Savage, B. D., & Sembach, K. R. 1991, apj, 379, 245, doi: 10.1086/170498

Scannapieco, E., Pichon, C., Aracil, B., et al. 2006, Monthly Notices of the Royal Astronomical Society, 365, 615, doi: 10.1111/j.1365-2966.2005.09753.x

Shull, J. M., Danforth, C. W., & Tilton, E. M. 2014, The Astrophysical Journal, 796, 49, doi: 10.1088/0004-637X/796/1/49

Simcoe, R. A. 2011, The Astrophysical Journal, 738, 159, doi: 10.1088/0004-637X/738/2/159

Simcoe, R. A., Cooksey, K. L., Matejek, M., et al. 2011, The Astrophysical Journal, 743, 21, doi: 10.1088/0004-637X/743/1/21

Songaila, A. 2005, The Astronomical Journal, 130, 1996, doi: 10.1086/491704

Tennyson, J. 2019, Astronomical Spectroscopy: An Introduction to the Atomic and Molecular Physics of Astronomical Spectroscopy (World Scientific)

Tie, S. S., Hennawi, J. F., Kakiichi, K., & Bosman, S. E. I. 2022, Monthly Notices of the Royal Astronomical Society, 515, 3656, doi: 10.1093/mnras/stac2021

Tremaine, S., Gebhardt, K., Bender, R., et al. 2002, apj, 574, 740, doi: 10.1086/341002

Tu, L., Luo, A., Wu, F., & Zhao, Y. 2010, Science China Physics, Mechanics and Astronomy, 53, 1928

Tumlinson, J., Peeples, M. S., & Werk, J. K. 2017, araa, 55, 389, doi: 10.1146/annurev-astro-091916-055240

Urrutia, T., Becker, R. H., White, R. L., et al. 2009, apj, 698, 1095, doi: 10.1088/0004-637X/698/2/1095

Urrutia, T., Lacy, M., & Becker, R. H. 2008, apj, 674, 80, doi: 10.1086/523959

Van der Maaten, L., & Hinton, G. 2008, Journal of machine learning research, 9

Veilleux, S., Rupke, D. S. N., Kim, D. C., et al. 2009, apjs, 182, 628, doi: 10.1088/0067-0049/182/2/628

Wei, P., Luo, A., Li, Y., et al. 2013, Monthly Notices of the Royal Astronomical Society, 431, 1800

Yang, L., Zheng, Z., du Mas des Bourboux, H., et al. 2022, The Astrophysical Journal, 935, 121, doi: 10.3847/1538-4357/ac7b2e

Zhu, G., & Ménard, B. 2013, The Astrophysical Journal, 770, 130, doi: 10.1088/0004-637X/770/2/130