Draft version November 4, 2025
Typeset using LaTeX **twocolumn** style in AASTeX7

BALNet: Deep Learning-Based Detection and Measurement of Broad Absorption Lines in Quasar Spectra

Yangyang Li,[1] Zhijian Luo,[1] Shaohua Zhang,[1] Du Wang,[2] Jianzhen Chen,[1] Zhu Chen,[1] Hubing Xiao,[1] and Chenggang Shu[1]

[1] *Shanghai Key Lab for Astrophysics, Shanghai Normal University, Shanghai 200234, China*
[2] *Shanghai Starriver Bilingual School, Shanghai 200233, China*

ABSTRACT

Broad absorption line (BAL) quasars serve as critical probes for understanding active galactic nucleus (AGN) outflows, black hole accretion, and cosmic evolution. To address the limitations of manual classification in large-scale spectroscopic surveys - where the number of quasar spectra is growing exponentially - we propose `BALNet`, a deep learning approach consisting of a one-dimensional convolutional neural network (1D-CNN) and bidirectional long short-term memory (Bi-LSTM) networks to automatically detect BAL troughs in quasar spectra. `BALNet` enables both the identification of BAL quasars and the measurement of their BAL troughs. We construct a simulated dataset for training and testing by combining non-BAL quasar spectra and BAL troughs, both derived from SDSS DR16 observations. Experimental results in the testing set show that: (1) BAL trough detection achieves 83.0% completeness, 90.7% purity, and an F1-score of 86.7%; (2) BAL quasar classification achieves 90.8% completeness and 94.4% purity; (3) the predicted BAL velocities agree closely with simulated ground truth labels, confirming `BALNet`'s robustness and accuracy. When applied to the SDSS DR16 data within the redshift range $1.5 < z < 5.7$, at least one BAL trough is detected in 20.4% of spectra. Notably, more than a quarter of these are newly identified sources with significant absorption, 8.8% correspond to redshifted systems, and some narrow/weak absorption features were missed. `BALNet` greatly improves the efficiency of large-scale BAL trough detection and enables more effective scientific analysis of quasar spectra.

Keywords: Quasars (1319) — Broad-absorption line quasar (183) — Active galaxies (17) — Neural networks (1933) — Astronomy data analysis (1858)

1. INTRODUCTION

Quasars are luminous active galactic nuclei (AGNs) powered by accreting supermassive black holes (SMBHs), and their powerful outflows are generally belived to play a crucial role in galaxy evolution. These outflows carry away huge amounts of material, kinetic energy, and angular momentum from the nuclear region, they may heat or expel interstellar gas, suppress star formation in the host galaxy, and regulate SMBH growth (e.g., P. F. Hopkins & L. Hernquist 2006; P. F. Hopkins & M. Elvis 2010). The most prominent signature of quasar outflows is the presence of blueshifted (up to ~ 0.2c) and broad (at least 2000 km/s) absorption line (BAL) troughs, which appear in both high-ionization species (e.g., C IV, Si IV, and N V; R. J. Weymann et al. 1991; M. S. Brotherton et al. 2001; T. A. Reichard et al. 2003), low-ionization species (e.g., Mg II, Al III, and He I*; A. Tolea et al. 2002; P. C. Hewett & C. B. Foltz 2003; S. Zhang et al. 2010; S. Zhang et al. 2011; W. Liu et al. 2015), and even in the excited of states Fe II and/or Fe III (C. Hazard et al. 1987; R. H. Becker et al. 1997; M. Vivek et al. 2012; S. Zhang et al. 2015). Statistical studies show that approximately 10–20% of ultraviolet- and optically selected quasars exhibit BAL troughs (C. B. Foltz et al. 1990; J. R. Trump et al. 2006; R. Ganguly et al. 2007; R. R. Gibson et al. 2009). It is likely that this fraction depends on the orientation and inner structure of the AGN (A. Tolea et al. 2002; P. C. Hewett & C. B. Foltz 2003), and may also reflect evolutionary phases of quasar activity (D. B. Sanders et al. 1988; F. Hamann & G. Ferland 1993; G. M. Voit et al. 1993). Moreover, the incidence, variability, and strength of

Email: zjluo@shnu.edu.cn, zhangshaohua@shnu.edu.cn

BAL troughs from various ionic species provide crucial diagnostics for the physical and dynamical properties of outflowing gases, as well as valuable insights into their origin and acceleration mechanisms (e.g., N. Filiz Ak et al. 2013; M. Vivek et al. 2014; Z.-C. He et al. 2015; X. Shi et al. 2016a; F. Hamann et al. 2018; Z. Chen et al. 2022). Therefore, the accurate identification and measurement of BAL troughs in quasar spectra constitute a fundamental step in the study of quasar outflows.

Traditional methods for identifying BAL troughs include quasar composite spectrum fitting combined with visual inspection (T. A. Reichard et al. 2003; J. R. Trump et al. 2006), multi-component spectral decomposition (A. Tolea et al. 2002; R. R. Gibson et al. 2009; S. Zhang et al. 2010), principal component analysis (PCA; K. Glazebrook et al. 1998), non-negative matrix factorisation (NMF; J. T. Allen et al. 2008), and quasar spectrum pair-matching technique (S. Zhang et al. 2014; W. Liu et al. 2015). These approaches involve fitting each individual quasar spectrum to reconstruct the intrinsic (unabsorbed) quasar spectrum, and then identifying BAL troughs by comparing the observed and intrinsic spectra. While effective, these methods are time-intensive and not easily scalable to the large and rapidly growing spectroscopic datasets produced by modern astronomical surveys. For example, the Sloan Digital Sky Survey (SDSS; D. G. York et al. 2000) DR16 quasar catalog (DR16Q; B. W. Lyke et al. 2020) has provided 750,414 optical quasar spectra. The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST; X.-Q. Cui et al. 2012) has also identified a total of 56,175 quasars (J. Jin et al. 2023). The ongoing larger survey, the Dark Energy Spectroscopic Instrument (DESI; A. Dey et al. 2019), aims to quadruple the known quasar sample by obtaining spectra of nearly 3 million quasars (E. Chaussidon et al. 2023). The rapid increase in data volume demands the development of more efficient and scalable methods for identifying BAL troughs.

With the advancement of computer technology, machine learning and artificial intelligence methods have been increasingly applied to the classification and feature recognition of astronomical spectral data. Commonly used techniques include K-Nearest Neighbors (KNN; e.g., M.-L. Zhang & Z.-H. Zhou 2007; T. Fushiki 2011; P. Sookmee et al. 2020), Support Vector Machines (SVM; e.g., C. Liu et al. 2015; A. Barrientos et al. 2020), Decision Trees (DT; e.g., J. R. Quinlan 1996; M. Czajkowski et al. 2014), and Artificial Neural Networks (ANN; e.g., K. Wang et al. 2016; N. Busca & C. Balland 2018; F. Rastegarnia et al. 2022). More recently, H. Yang et al. (2023) demonstrated that convolutional neural networks (CNNs) significantly outperform traditional methods in classifying observed astronomical spectra. N. Busca & C. Balland (2018) proposed a deep CNN model, `QuasarNET`, designed for redshift estimation and object classification, including the identification of BAL quasars. In addition, Z. Guo & P. Martini (2019) developed a PCA-enhanced CNN framework for classifying BAL quasars. R. Moradi et al. (2024) introduced an enhanced ResNet-based CNN model, `FNet II`, capable of processing spectra through automated feature extraction, thereby eliminating the need for manual identification of spectral lines. Comparative studies by W. Kao et al. (2024) and S. Pang et al. (2025) have further established the superiority of deep learning approaches, particularly CNN architectures, over traditional machine learning methods like extreme gradient boosting (XGBoost) for BAL quasar identification.

Although existing methods have achieved impressive performance in classifying BAL quasars — with extremely high detection completeness exceeding 97% (e.g., N. Busca & C. Balland 2018; Z. Guo & P. Martini 2019; R. Moradi et al. 2024; W. Kao et al. 2024; S. Pang et al. 2025) — several key challenges remain unresolved. First, most current approaches focus primarily on distinguishing BAL quasars from general quasar populations, but lack the ability to quantitatively characterize BAL troughs, such as their velocity and velocity structure. Second, the training datasets are mainly constructed from previously confirmed BAL quasars, which tend to be biased toward strong C IV BAL troughs with large blueshifted velocities. Although the high completeness and purity of these datasets are notable, they do not truly reflect the performance on real BAL quasars. This leads to a selection bias that limits model sensitivity to shallower troughs or those superimposed on C IV emission lines, thereby reducing the completeness and generalizability of the classification. In addition, the important subclass of redshifted BALs has been largely overlooked. Previous studies have shown that they trace inflowing gases in the nuclear region of AGNs (X. Shi et al. 2017; N.-X. Zhang et al. 2017; H. Zhou et al. 2019), providing key insights into SMBH accretion physics. Therefore, it is necessary to address these limitations from both methodological and training dataset construction perspectives.

In this paper, we introduce `BALNet`, a new automated deep learning framework that integrates a one-dimensional convolutional neural network (1D-CNN) with bidirectional long short-term memory (Bi-LSTM) networks to accurately detect and characterize C IV BAL trough in quasar spectra. Unlike previous methods, `BALNet` not only classifies BAL quasars but also

directly measures the kinematic properties (e.g., velocities) of C IV BAL troughs. For model training and evaluation, we generated a large set of simulated spectra with diverse C IV BAL trough profiles, derived from the SDSS DR16Q dataset.

The remainder of this paper is organized as follows: Section 2 describes the construction of mock spectra used for model training and testing. Section 3 details the architecture of the proposed `BALNet` framework and outlines its training procedure. Section 4 presents the evaluation results of `BALNet` on simulated test data and discusses its performance. Section 5 then applies `BALNet` to the SDSS DR16Q dataset and analyzes its performance. Finally, Section 6 summarizes the main conclusions of this study and outlines future directions.

2. MOCK DATA

The performance of deep learning models depends heavily on both the quantity and quality of the training dataset. A large-scale and accurately labeled training sample is a key factor in improving model performance. It enables deep learning models to effectively learn diverse features, thereby significantly enhancing their generalization ability and overall prediction accuracy. The aim in this study is how to simultaneously identify BAL quasars and measure the BAL troughs. All existing labeled quasar datasets provide only conventional BAL classifications (i.e., whether a quasar is a BAL quasar) and lack detailed annotations of specific BAL velocity structures. As a result, such datasets are inadequate for our research objectives.

To address this, we construct a more comprehensive and accurately labeled mock dataset of quasar spectra for training and evaluating the `BALNet` model. This mock dataset is based on the SDSS DR16Q dataset, limited to quasars with redshifts ranging from 1.5 to 5.7 to ensure that potential C IV BAL troughs fall within the SDSS spectrograph's wavelength coverage. The SDSS DR16Q catalog provides the BI_CIV and AI_CIV parameters, which are used both to distinguish BAL quasars from non-BAL quasars and to quantify the strength of C IV BALs, we then obtain 23,994 BAL quasars with BI_CIV > 0 and 313,739 non-BAL quasars with BI_CIV $= 0$ and AI_CIV $= 0$. In the construction of the mock dataset, spectra of non-BAL quasars are randomly sampled, while C IV BAL troughs are extracted directly from real absorption troughs in observed BAL quasars. The procedures for C IV BAL trough extraction and mock spectrum generation are described in detail in the following two subsections. Before analysis, all spectra are corrected for Galactic extinction using the extinction map of D. J. Schlegel et al. (1998) and the reddening curve of E. L. Fitzpatrick (1999), and then shifted to the rest frame using the primary redshift from the SDSS DR16Q.

2.1. *BAL Trough Extraction*

We employ the quasar spectrum pair-matching method (S. Zhang et al. 2014; W. Liu et al. 2015) to extract the C IV BAL troughs from the spectra of the SDSS DR16Q BAL quasars. The core idea, similar to other traditional approaches, is to construct an unabsorbed model spectrum that closely matches the absorption-free regions of a given BAL quasar candidate. This model is then compared with the observed spectrum to identify potential BAL troughs.

The specific implementation consists of the following three steps: (1) Randomly select 300 non-BAL quasar spectra with redshifts close to that of the BAL quasar candidate under analysis to form a template library. Each template spectrum is smoothed through two iterations of B-spline fitting to remove narrow absorption lines. (2) Each template is reddened using the SMC extinction law and fitted to the absorption-free regions of the given BAL quasar candidate. The scaling factor and the color excess $E(B-V)$ are optimized by the minimizing χ^2 between the model and the observed spectrum. To improve the fitting accuracy and robustness, we restrict the analysis to a rest-frame wavelength range of 1300 - 1700 Å, which ensures coverage of the C IV BAL trough while minimizing contamination from unrelated spectral regions. (3) All model spectra are ranked in ascending order of total χ^2, and the top forty best-fitting models are selected. For each of them, an additional emission-line χ^2 metric, $\chi^2_{\rm EL}$, is calculated over the 1450 - 1650 Å range, excluding pixels affected by BAL troughs. This metric quantifies the similarity between the model and the observed spectrum in the unabsorbed C IV emission-line region. Finally, the composite of the ten models with the lowest $\chi^2_{\rm EL}$ values is adopted as the final unabsorbed model spectrum. This method exhibits strong detection performance for shallow and weak BAL features, and achieves high accuracy in measuring absorption velocities.

We systematically search for BAL troughs in the normalized spectra of BAL quasar candidates. In velocity space, the search is conducted over the range from $v_l = 10{,}000$ km/s to $v_u = -29{,}000$ km/s, where positive velocities indicate redshifted absorption and negative velocities indicate blueshifted absorption. Only features with a continuous absorption over a velocity interval greater than 1,000 km/s for a depth of at least 10% (normalized fluxes smaller than 0.9) are considered valid BAL troughs. After screening 23,994 DR16Q BAL

quasar candidates, we successfully detected 47,267 BAL troughs in 23,107 sources. In Figure 1, we present the number distribution of BAL troughs in these observed BAL quasars (black solid line). The results show that most BAL quasars contain one to three BAL troughs, while a small fraction exhibit four or more, and in rare cases, up to nine troughs. All identified BAL troughs constitute the BAL pattern library, which is used to construct simulated spectra for subsequent model training.

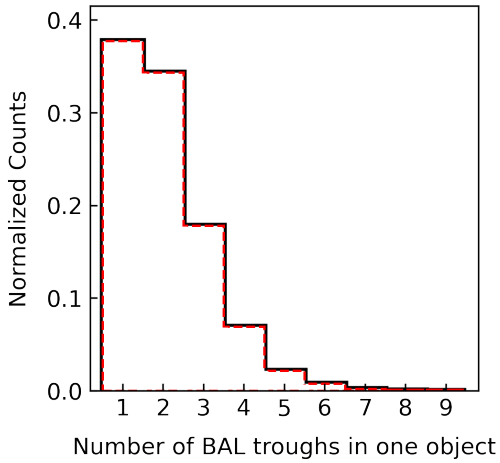

Figure 1. Distributions of the number of C IV BAL troughs per spectrum for 23,107 observed BAL quasars (black solid line) and for 100,000 simulated BAL quasars (red dashed line).

2.2. *Data Construction*

In this subsection, we construct a large-scale mock dataset of simulated BAL quasar spectra by combining the high-quality BAL pattern library with non-BAL quasar samples derived from real observations. These simulated spectra not only encompass a wide variety of BAL trough types, but also include detailed annotations of the corresponding velocity parameters, thereby satisfying the requirements for simultaneous BAL classification and velocity feature extraction. Figure 2 illustrates the procedure for constructing a representative simulated BAL quasar spectrum. The construction process strictly follows the steps outlined below.

Firstly, we randomly select n BAL troughs from the BAL pattern library constructed in the previous subsection, where n is a positive integer (e.g., 1, 2, 3, etc.). These patterns are then placed at random, non-overlapping positions within the velocity range from 10,000 to −29,000 km/s to generate the normalized BAL spectrum ($NS_{BALs}(\lambda)$). In these spectra, wavelength regions without BAL patterns have a flux value of 1. Figure 2 (a) shows a typical normalized BAL spectrum, with gray-shaded areas representing 3 BAL troughs.

Secondly, background spectrum smoothing is performed. We randomly selected an original spectrum from the non-BAL quasar sample and applied a three-point smoothing method. This process aims to reduce the impact of noise and unresolvable absorption lines, thereby minimizing their interference with subsequent model analysis. Figure 2 (b) illustrates an example: the raw spectrum of the SDSS object is shown in black, while its smoothed counterpart, denoted as $f_{non-BAL}(\lambda)$, is shown in blue.

Finally, the background spectrum ($f_{non-BAL}(\lambda)$) is combined with the normalized BAL spectrum ($NS_{BALs}(\lambda)$) to generate the simulated BAL quasar spectrum ($f_{simulated}(\lambda)$), as shown by the black line in Figure 2 (c). The formula is expressed as:

$$f_{simulated}(\lambda) = NS_{BALs}(\lambda) \times f_{non-BAL}(\lambda). \quad (1)$$

To generate the final simulated BAL quasar spectrum data set, we performed the above procedure 100,000 times. The number distribution of BAL troughs in simulated spectra is consistent with the observational data (see Figure 1), thereby ensuring the high credibility of the simulation results. Additionally, to ensure balance and stability in the training dataset, we randomly selected 100,000 spectra from non-BAL quasar sample and combined them with all simulated BAL quasar spectra. The background spectra of the simulated BAL quasars and the non-BAL quasar spectra were not subjected to any signal-to-noise cuts, ensuring that their signal-to-noise distributions are consistent with those of SDSS DR16Q. Through this integration, we created a comprehensive dataset containing a total of 200,000 spectra. This dataset was used to train and test our `BALNet` model.

It should be noted that the training set spectra ultimately constructed encompass various manifestations of BAL troughs, such as blueshifted, redshifted, and multiple BAL troughs. Meanwhile, all spectra in the training set have been processed through interpolation and are evenly distributed across 1165 equally spaced wavelength points within the range of 1300 to 1700 Å. In our dataset, in addition to the spectral data, label data for each spectrum are also required. The most straightforward approach would be to label each wavelength point of the spectrum as to whether it belongs to a BAL trough, for example, using 1 to indicate the presence of a BAL trough and 0 for its absence. However, this method would result in overly complex label data and excessively high computational costs. To avoid this

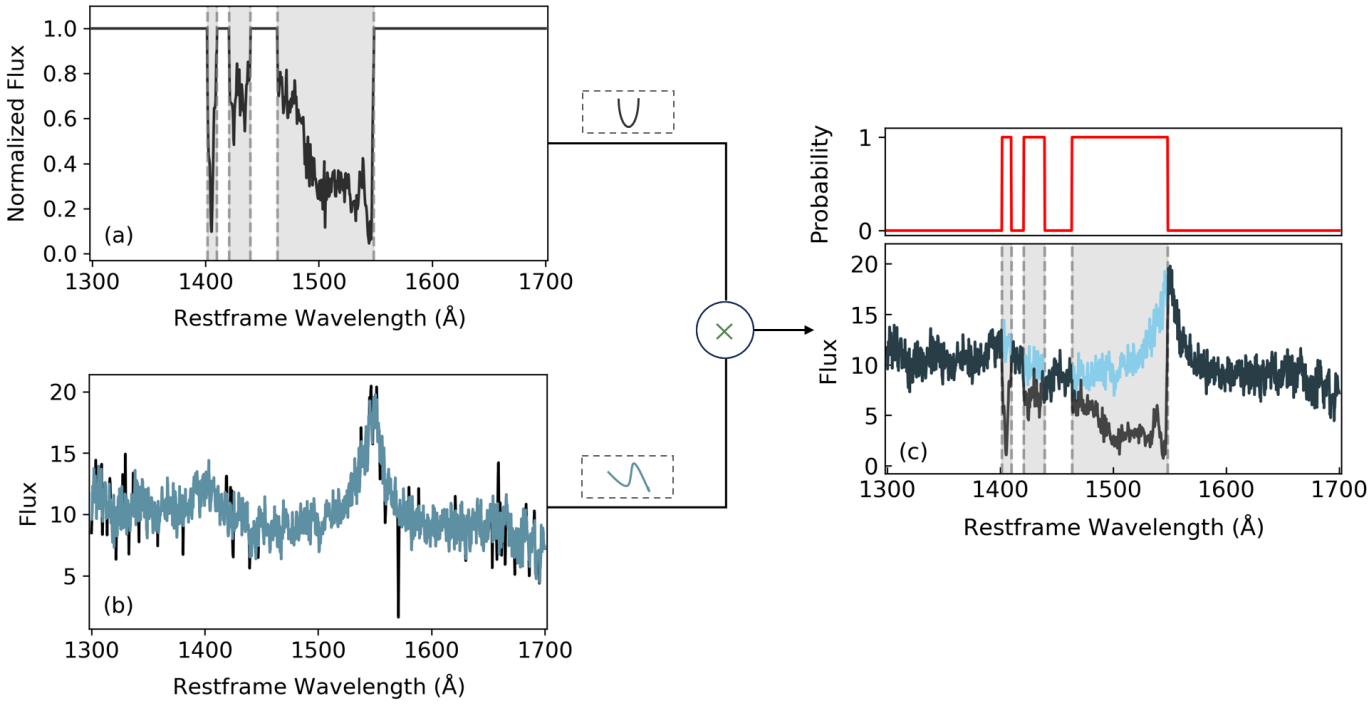

Figure 2. Flowchart illustrating the procedure for constructing simulated spectra. Panel (a) presents a normalized BAL pattern spectrum (black), with the gray-shaded regions marking three randomly selected BAL troughs. Panel (b) displays a real, unabsorbed spectrum (black) and its smoothed version (blue). Panel (c) shows the final simulated spectrum (black) in the bottom part, and the corresponding label vector (red) in the upper part, where BAL regions are labeled as 1 and non-BAL regions as 0.

issue, we defined coarser interval points within the same wavelength range as the simulated spectra, consisting of a total of 387 intervals, with each interval corresponding to three consecutive spectral wavelengths. Subsequently, we constructed a label vector based on these intervals to characterize the distribution of BAL troughs within the spectrum. Specifically, if a BAL trough is present within an interval, the corresponding label value for that interval is marked as 1; otherwise, it is marked as 0 (as indicated by the red line at the top of Figure 2 (c)). In this way, we simplified the structure of the label data while retaining the key information about the BAL troughs in the spectrum, thereby providing more efficient data support for model training.

Regarding the division of the training and testing sets, we randomly selected 80% of the samples from the final dataset, amounting to 160,000 spectra and their corresponding labels, to form the training set for the model. The remaining 20% of the samples, which is 40,000 spectra and their labels, were then used for testing and evaluating the model.

3. METHODOLOGY

In this section, we will provide a detailed introduction to the `BALNet` model used for detecting BAL troughs in quasar spectra. This model innovatively integrates a one-dimensional convolutional neural network (1D-CNN) and bidirectional long short-term memory networks (Bi-LSTM, M. Schuster & K. K. Paliwal 1997), which enables effective extraction of both local and global features from spectral data. The 1D-CNN layer focuses on capturing local features in the spectrum, such as absorption trough profiles and continuum fluctuations, while the Bi-LSTM layers learn the global contextual dependencies of the spectrum through bidirectional sequence modeling (forward and backward propagation). This combined approach significantly improves the detection performance of BAL trough. In particular, when dealing with quasar spectra that have multiple BAL troughs, the model can more accurately identify and locate these features, thereby improving the accuracy and reliability of detection.

It is important to note that although 1D-CNN has been widely used in previous BAL automatic recognition studies (N. Busca & C. Balland 2018; Z. Guo & P. Martini 2019; S. Pang et al. 2025), the Bi-LSTM integration scheme proposed in this paper is the first sequence modeling-based work in this field. A key advantage of this approach lies in the Bi-LSTM's ability to dynamically learn dependencies across arbitrary distances, a capability not inherent to CNNs. This characteristic is particularly crucial for detecting disjoint BAL

troughs, which are common in our data. While a CNN with varying large kernels could be designed to capture long-range correlations, it requires careful manual tuning of kernel sizes and depths. In contrast, the Bi-LSTM adaptively learns the relevant context lengths from the data, leading to a more efficient and flexible architecture for modeling the complex, long-range dependencies present in BAL spectra. Furthermore, the Bi-LSTM's capacity to simultaneously model both global long-term trends and local short-term fluctuations aligns closely with the comprehensive reasoning process astronomers employ in manual BAL identification. These combined strengths make it especially suitable for handling the multiple absorption features commonly found in quasar spectra. The following sections will first introduce the basic principles of LSTM/Bi-LSTM networks, and then provide a detailed description of the architecture and training strategy of `BALNet`.

3.1. *LSTM Network*

LSTM network is a specialized type of recurrent neural network (RNN) initially proposed by S. Hochreiter & J. Schmidhuber (1997). It has been widely used for processing sequential data, including data analysis of spectral and time series (L. Hu et al. 2022; S. S. Tabasi et al. 2023; Z. Luo et al. 2024a). Unlike traditional RNNs (M. Schuster & K. K. Paliwal 1997), LSTM networks effectively addressed the vanishing and exploding gradient problems encountered when processing long sequences by incorporating memory cell states. Due to its strong memory capacity and ability to capture long-term dependencies, LSTM network is frequently employed in time-series forecasting and reliability prediction.

Similar to RNNs, LSTM networks employ a chain-like structure but feature a more sophisticated design in their repeating modules. By incorporating memory cells and gating mechanisms, LSTMs can effectively capture long-range dependencies. As illustrated in Figure 3, the control flow of a single LSTM timestep involves four core components: the input gate (i), forget gate (f), output gate (o), and candidate memory cell state ($\tilde{C}$). During timestep propagation, these elements operate in concert: the forget gate regulates which historical information to preserve, the input gate selects relevant features from the current input, the cell state synthesizes new and existing information to update memory, while the output gate governs the activation value for the current timestep. This integrated mechanism enables LSTMs to efficiently process sequential data and establish accurate input-output mappings.

By applying the sigmoid (σ) and hyperbolic tangent (tanh) activation functions to the current input sequence

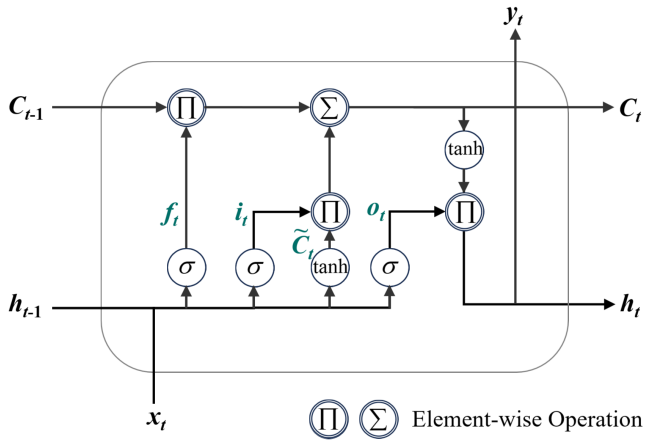

Figure 3. Schematic of the LSTM unit structure, illustrating the four interacting key components (input gate i, forget gate f, output gate o, and candidate memory cell state $\tilde{C}$) and their information flow. The mathematical expressions for the gates are provided in Equation (2), while the temporal update mechanism of the cell state is strictly defined in Equation (3). The arrows in the figure clearly indicate the direction of information flow, reflecting the dynamic gating logic of LSTM when processing sequential data.

x_t and the previous hidden state h_{t-1}, the LSTM operations are updated at each time-step (t) according to the following equations:

$$\begin{aligned} f_t &= \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \\ i_t &= \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right), \\ o_t &= \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right), \\ \widetilde{C}_t &= \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right), \end{aligned} \tag{2}$$

where (W_f, W_i, W_o, W_C) and (b_f, b_i, b_o, b_C) are the weight matrices and bias weights, respectively. $\widetilde{C}_t$ indicates the candidate cell state. Then, utilizing the above equations to update the cell state C_t and hidden state h_t at the current time-step.

$$\begin{aligned} C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \\ h_t &= o_t \odot \tanh\left(C_t\right), \end{aligned} \tag{3}$$

where the $\odot$ symbol denotes an element-wise product. The initial values of C_0 and h_0 are both set to 0. This process illustrates how the LSTM regulates the input of new information and the flow of output information via its gating mechanism, thereby facilitating the effective maintenance of long-term memory and dependencies in sequential data. Through its structural design and gating mechanism, the LSTM enhances its performance and accuracy.

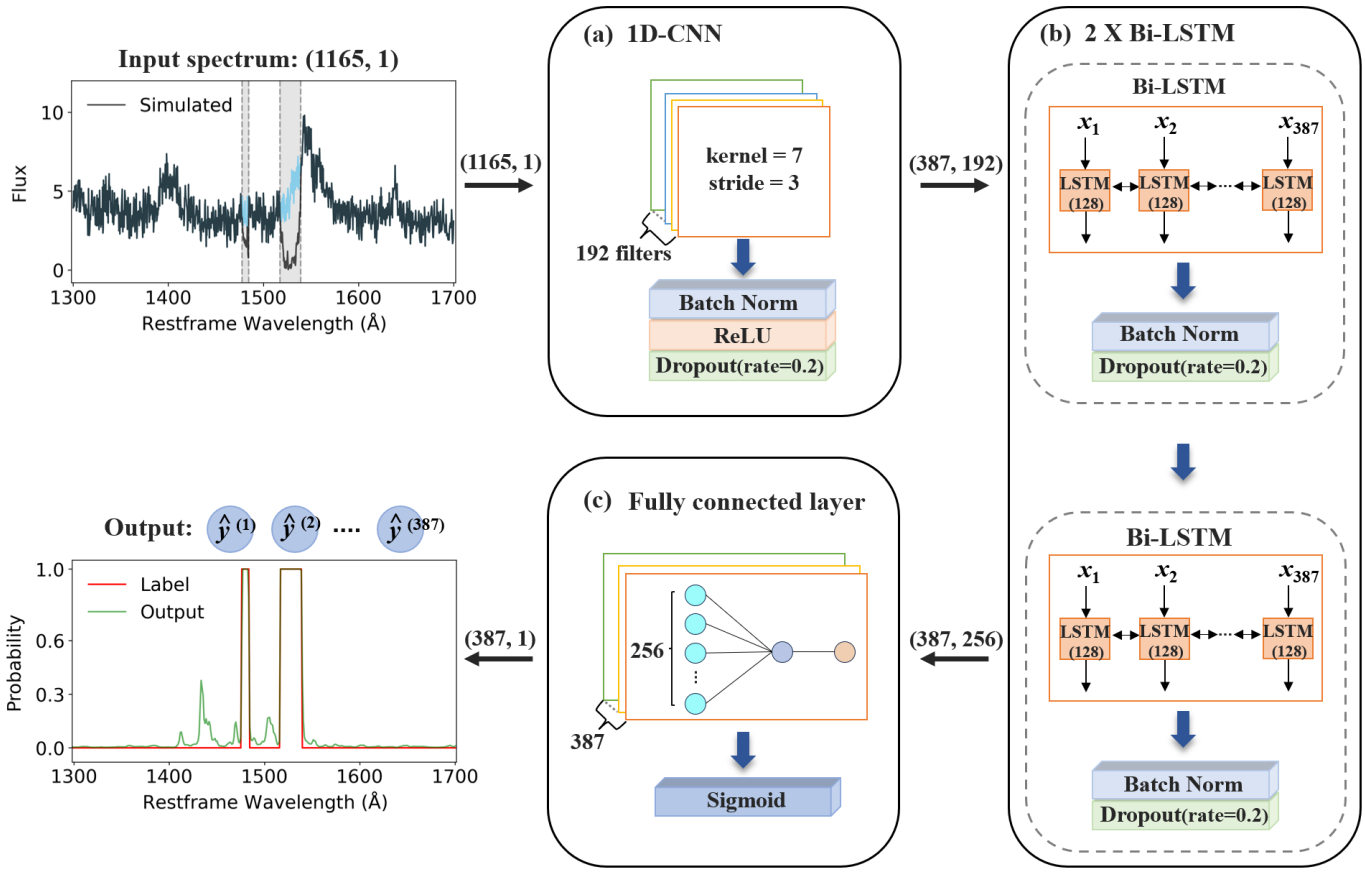

Figure 4. Architecture of the proposed `BALNet` framework, comprising three core modules: (a) The 1D-CNN feature extractor processes the 1165-dimensional input spectrum through convolutional operations (kernel_size=7, stride=3, 192 filters), transforming it into a 387×192 feature matrix, followed by batch normalization, ReLU activation, and dropout (rate=0.2) for feature refinement; (b) The Bi-LSTM module employs two bidirectional LSTM layers (128 hidden units each) to analyze temporal patterns, producing a 387×256 feature matrix. Each LSTM layer is followed by batch normalization and dropout for regularization; (c) The output module generates a 387-dimensional probability vector through a fully connected layer with sigmoid activation, where each element represents the presence probability of BAL troughs at the corresponding spectral position.

In addition, compared with most traditional machine learning methods, LSTM can automatically and effectively learn the dependencies between elements in the input sequence, thereby capturing the long-term dependencies and dynamic characteristics in sequential data. This capability gives LSTM a significant advantage in handling long sequence data.

The Bi-LSTM network is an enhanced variant of LSTM that incorporates two parallel LSTM layers: a forward layer processing the sequence chronologically and a backward layer processing it in reverse order. This bidirectional architecture enables simultaneous capture of both historical and future contextual features, thereby significantly improving the modeling of long-range dependencies in sequential data.

3.2. *`BALNet` Architectures*

The primary objective of this study is to accurately identify BAL troughs in quasar spectra and determine their specific positions within the spectra, in order to obtain the associated velocity information. This task presents significant challenges because BAL troughs exhibit remarkable morphological and positional diversity across different spectra, and are often obscured by interference from other spectral features (such as emission lines and continuum spectra). To address these challenges, we developed a specialized neural network model named `BALNet` which innovatively combines the strengths of 1D-CNN and Bi-LSTM with 1D-CNN extracting local spectral features while Bi-LSTM effectively captures long-term dependencies in spectral sequences. The neural network architecture was implemented using the Keras [3] (F. Chollet & others 2015) and TensorFlow (M. Abadi et al. 2016) frameworks, with Keras serving as the high-level API for TensorFlow. Fur-

[3] https://keras.io/about/

thermore, to ensure reproducibility and enable community access, both the source code and the generated catalogs have been made publicly available on our GitHub repository [4].

The network architecture of `BALNet`, illustrated in Figure 4, follows an end-to-end deep learning approach with three core modules connected in series. First, the 1D-CNN feature extraction module processes the input 1165-dimensional spectral vector. Using a single 1D convolutional layer (kernel_size = 7, stride = 3, 192 filters), it transforms the input into a 387×192 feature matrix, where 387 corresponds to the sequence length and 192 denotes the number of feature channels. This module includes a complete feature optimization process: a Batch Normalization layer (Batch Norm) to stabilize network training, a ReLU activation function to introduce nonlinearity (B. Xu et al. 2015), and a Dropout layer (rate = 0.2) to prevent overfitting (G. E. Hinton et al. 2012; W. Zaremba et al. 2014).

Next, the Bi-LSTM module receives the 387×192 feature matrix and processes it through two layers of bidirectional LSTM. Each LSTM layer contains 128 hidden units and the bidirectional outputs are concatenated to form a 256-dimensional vector, resulting in the final output of a 387×256 feature matrix. This module also applies Batch Norm and Dropout (rate = 0.2) for regularization.

Finally, the output module transforms the 387×256 feature matrix into a 387-dimensional probability output vector through a fully connected layer and a sigmoid activation function. Each element in the vector represents the probability of the presence of a BAL trough at the corresponding position. Typically, a threshold of 0.5 is used for binary classification to determine the existence of a BAL trough point. In addition, in the model, we regard the continuous occurrence of five or more BAL trough points as a confirmation signal for a BAL trough.

The entire network performs end-to-end processing from raw spectral inputs to both classification and positional (velocity) parameter prediction, enabled by strict dimensional control and consistent hyperparameter settings throughout the architecture.

The loss function of the `BALNet` model employs binary cross-entropy loss (BCE) [5] . This loss function optimizes the model's performance by calculating the difference between the model's predicted output and the true labels.

[4] https://github.com/zjluo-code/BALNet

[5] https://www.tensorflow.org/api_docs/python/tf/keras/losses/BinaryCrossentropy

After model construction, we trained the network using 160,000 normalized spectral entries from the training set. Each spectrum was normalized by its maximum value prior to input. All experiments were performed on an NVIDIA RTX 3090 GPU. We employed the Adam optimizer (D. P. Kingma & J. Ba 2017) with $\beta_1 = 0.5$, $\beta_2 = 0.999$, a learning rate of 0.0001, and a batch size of 256. The training completed in 200 epochs, with each epoch requiring ~ 42 seconds, resulting in a total training time of approximately 2.4 hours.

4. MODEL PERFORMANCE EVALUATION

In this section, we evaluate how effectively our trained `BALNet` model performs when detecting BAL troughs and measuring their locations in simulated quasar spectra using the test dataset.

4.1. *Evaluation Metrics*

We quantitatively evaluate `BALNet`'s BAL trough detection performance using three standard metrics: completeness (recall), purity (precision) and F1-score. These metrics constitute a robust evaluation framework where: (1) completeness measures the model's ability to identify all relevant troughs, (2) purity quantifies the correctness of detected troughs, and (3) the F1-score balances these complementary aspects. The metrics are formally defined as:

$$\begin{aligned} \text{Completeness} &= \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ \text{Purity} &= \frac{\text{TP}}{\text{TP}+\text{FP}}, \\ \text{F1-score} &= 2 \times \frac{\text{Purity} \times \text{Completeness}}{\text{Purity} + \text{Completeness}}, \end{aligned} \tag{4}$$

where, TP (true positive) represents the number of BAL troughs that are correctly detected by the model, FP (false positive) represents the number of non-BAL troughs that are incorrectly detected as BAL troughs by the model, and FN (false negative) represents the number of actual BAL troughs that are not detected by the model.

Furthermore, to fully evaluate the performance of `BALNet` in different decision thresholds, we employ the area under the precision-recall curve (AU-PRC) as a complementary evaluation metric. AU-PRC is calculated by integrating the area bounded by the precision-recall (PR) curve and the axes, providing a comprehensive measure of the model's performance stability under varying threshold configurations. This metric ranges from $[0, 1]$, where 1 represents ideal performance and 0.5 corresponds to random guessing. A higher AU-PRC

value indicates that the model maintains high precision while simultaneously achieving high recall, confirming the superior overall detection capability.

To evaluate the performance of our model in estimating BAL velocity parameters based on location measurements, we used the label values of the simulated data as a benchmark and systematically compared the velocity parameters predicted by the model `BALNet` with these true values. To comprehensively quantify the measurement quality, we employed three metrics: the fraction of catastrophic outliers (f_{out}), the normalized median absolute deviation ($\sigma_{\rm NMAD}$), and the bias in the BAL velocity parameter measurements (bias) to assess the quality of the model's velocity parameter measurements (G. B. Brammer et al. 2008; Z. Luo et al. 2024a; Z. Luo et al. 2024b).

Among these metrics, f_{out} measures the proportion of predictions that fall outside an acceptable error range, indicating potential significant deviations from true values. A velocity parameter estimate is considered a catastrophic outlier if it meets the following condition:

$$\frac{|\Delta v|}{1+|v_{\rm true}|} > 0.15, \tag{5}$$

where

$$\Delta v = v_{\rm pred} - v_{\rm true}, \tag{6}$$

$v_{\rm true}$ is the reference velocity used as the "ground truth" benchmark, and $v_{\rm pred}$ is the velocity parameter predicted by the model, both in units of 1000 km/s. $\sigma_{\rm NMAD}$ quantifies the dispersion of predicted velocity values relative to ground truth, serving as a robust measure of estimation precision. It is defined as:

$$\sigma_{\rm NMAD} = 1.48 \times {\rm median}\left(\left|\frac{\Delta v - {\rm median}(\Delta v)}{1+|v_{\rm true}|}\right|\right). \tag{7}$$

The bias assesses whether there is a systematic tendency for the model to overestimate or underestimate the velocity parameters relative to the truth, and it is typically calculated using the following formula:

$${\rm bias} = {\rm median}\left(\frac{\Delta v}{1+|v_{\rm true}|}\right). \tag{8}$$

4.2. *BAL Trough Detection*

As can be seen from the architecture of the `BALNet` (Figure 4), for each quasar spectrum, the model outputs a 387-dimensional probability vector. Therefore, we can identify BAL troughs in quasar spectra by setting an appropriate threshold (PIXEL_PROB). Although a higher threshold can improve the purity of identification, this is often at the expense of completeness. To obtain the optimal value of PIXEL_PROB, we analyzed the PR curves of the model on the training and test sets and calculated the AU-PRC values. As shown in Figure 5, the AU-PRC values reached 0.92. This near-ideal performance (where 1.00 indicates perfect model performance), along with the close alignment between the training and testing PR curves, confirms that our model has good generalization ability and has not overfitted. Considering the balance between precision and recall, we ultimately determined the optimal pixel probability threshold to be 0.3 based on the F1-score.

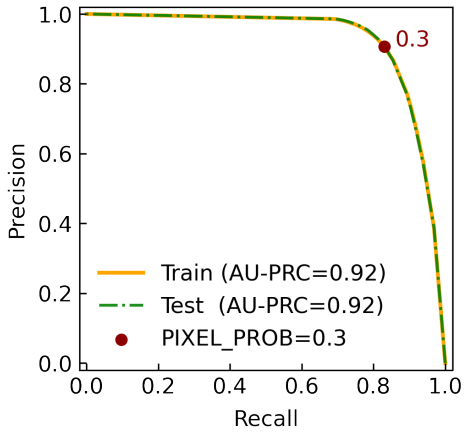

Figure 5. Precision-recall (PR) curves for the training set (orange) and the testing set (green). The brown point marks the probability threshold at which the model achieves its optimal performance (Best F1-score, PIXEL_PROB=0.3). The AU-PRC serves as a quantitative measure of the model's performance.

At this threshold, our trained model achieved a BAL trough identification accuracy of 90.7% on the testing set. Table 1 summarizes the detailed performance metrics of the model on both the training and testing sets, including completeness, purity, and F1-score. As shown in the table, the `BALNet` model performs consistently well across both datasets, achieving a completeness of around 83%, a purity of approximately 91%, and an F1-score of about 87%. These results demonstrate the model's robustness and effectiveness in detecting BAL troughs in quasar spectra. Figure 6 presents representative examples in which the trained model successfully identifies spectra containing zero, one, or multiple BAL troughs.

The presence of BAL troughs in quasar spectra serves as a key diagnostic for identifying BAL quasars. Accordingly, our model can be applied to spectral data to classify sources as BAL or non-BAL quasars. Evaluation on the testing set shows that the model achieves a completeness of 90.8% and a purity of 94.4% in BAL quasar classification. These results indicate that the model ef-

Table 1. Evaluation metrics of BAL trough detection by `BALNet`, using the optimal PIXEL_PROB threshold of 0.3.

	Completeness (%)	Purity (%)	F1-score (%)	AU-PRC
Train	83.1	90.5	86.6	0.92
Test	83.0	90.7	86.7	0.92

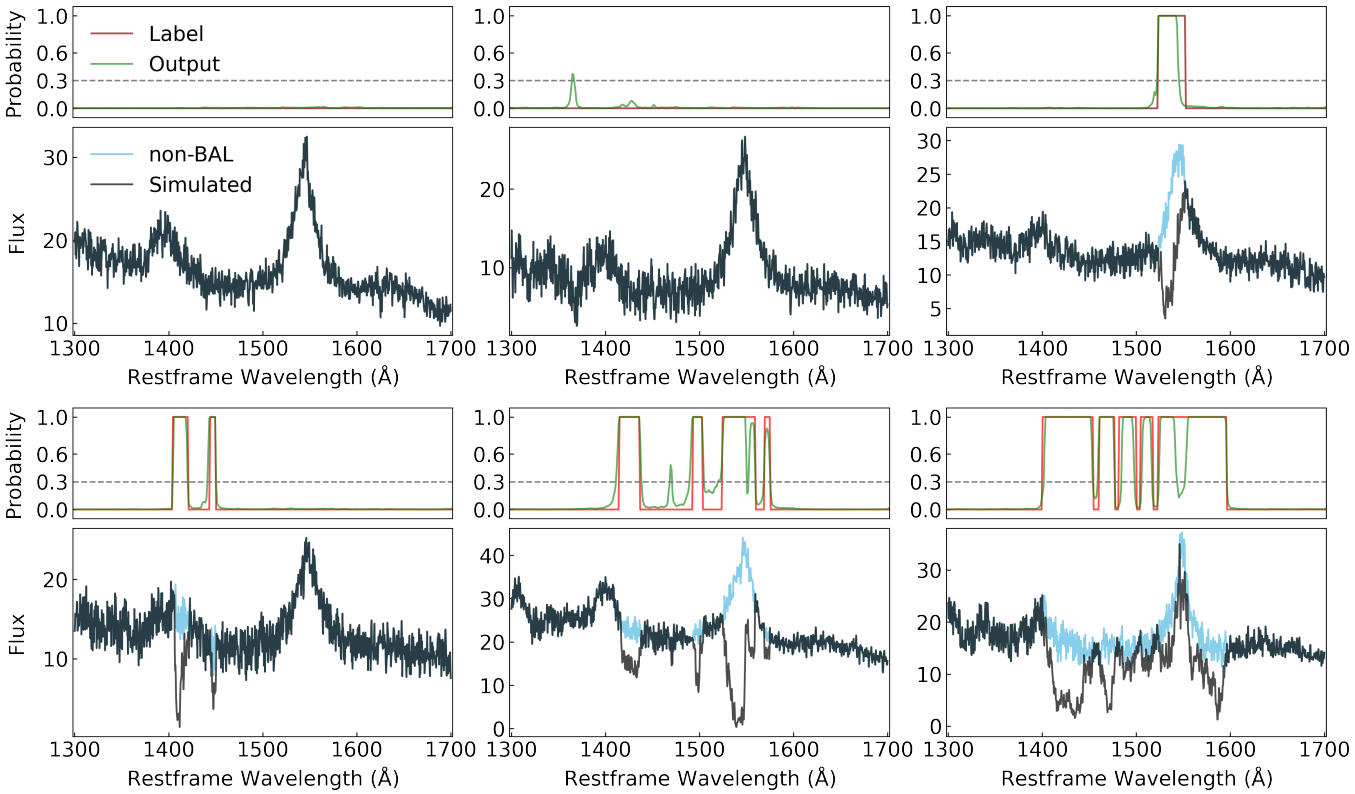

Figure 6. Examples from the test set and corresponding model predictions. In each panel, the top row displays the ground-truth label vectors in red and the model-predicted probabilities in green, scaled between 0 and 1. The bottom row shows the simulated spectra in black and their corresponding non-BAL quasar background spectra in blue; the difference between the two reveals the BAL troughs embedded within the simulated spectra.

fectively detects the majority of true BAL quasars while maintaining a very low false-positive rate, demonstrating its robustness and reliability in distinguishing BAL quasars from general quasar populations.

4.3. *BAL Trough Velocity Measurements*

Based on the output vector of the `BALNet` model, each element corresponds to a specific position in the input spectrum. This positional correspondence enables the determination of both the location and the width of BAL troughs in quasar spectra, from which three key velocity parameters are extracted: $V_{\rm max}$ and $V_{\rm min}$, and $V_{\rm ave}$. Specifically, $V_{\rm max}$ and $V_{\rm min}$ denote the maximum (blueward) and minimum (redward) velocities of the BAL trough, respectively, while $V_{\rm ave}$ represents its average velocity. These parameters quantify the kinematic properties of BAL outflows and serve as important diagnostics of quasar physical conditions, such as the dynamics and geometry of the outflowing material.

Figure 7 presents the comparison between the predicted and "ground-truth" label values of the velocity parameters on the testing set. As shown, the predicted values closely track the overall trend of the label values, indicating that the model effectively captures the kinematic properties of BAL troughs. Quantitatively, the fraction of catastrophic outliers ($f_{\rm out}$) dose not exceed approximately 9.0%, suggesting that while a small number of outliers exist, the model maintains high predictive accuracy overall. Furthermore, the normalized median absolute deviation ($\sigma_{\rm NMAD}$) is below 0.03, and the systematic bias in the measurement of the velocity (bias) is less than 10^{-5}, further validating the robustness and high reliability of the model in the predicting of the velocity parameters.

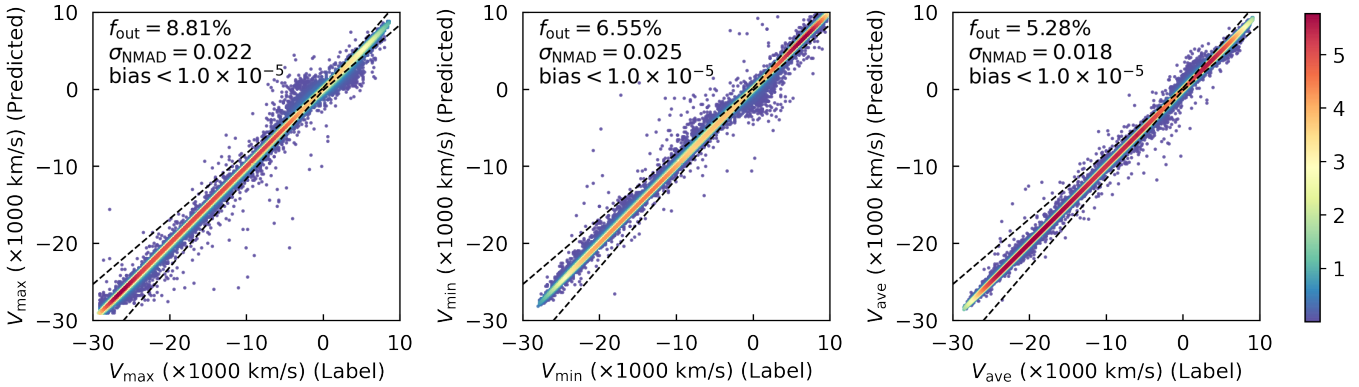

Figure 7. Comparison between the "ground-truth" labels and the predicted BAL velocity parameters from simulated spectra. Model performance is assessed using three metrics: the fraction of catastrophic outliers (f_{out}), the normalized median absolute deviation (σ_{NMAD}), and the systematic velocity bias (bias). The solid line indicates perfect agreement between predicted and label values, while the dashed lines mark the ±0.15 threshold set by the spectral resolution; data points lying outside this range are considered catastrophic outliers. The color scale reflects the density distribution of data points.

Table 2. Comparing the completeness and purity of `BALNet` with previous works in BAL quasar classification. Note that different studies employ distinct datasets and labeling strategies.

Model	Completeness (%)	Purity (%)	Train: Test	Data	Reference
`BALNet` (CNN + Bi-LSTM)	90.8	94.4	8:2	Mock data	this work
`QuasarNet` (CNN)	98.0	77.0	8:2	DR12	N. Busca & C. Balland (2018)
PCA-echenced CNN	97.4	40.0	9:1	DR12	Z. Guo & P. Martini (2019)
`FNet II` (ResNet + CNN)	99.0	-	9:1	DR16/DR17	R. Moradi et al. (2024)
PCA + XGBoost	97.7	96.2	9:1	DR16	W. Kao et al. (2024)

4.4. *Compare With Other Works*

In previous studies (see Table 2), `QuasarNet` (N. Busca & C. Balland 2018) and `FNet II` (R. Moradi et al. 2024) have been proposed as four-class classifiers designed to distinguish stars, galaxies, quasars, and BAL quasars. These models fundamentally rely on the BAL flags provided in the SDSS catalog to label training samples and evaluate performance by comparing predictions against these flags. However, known inaccuracies in the BAL flags in SDSS, combined with significant class imbalance—BAL quasars being far less numerous than other object types—may compromise the validity of the resulting performance metrics. For these reasons, although `QuasarNet` achieves a high completeness of 98.0%, its relatively low purity of 77.0% reflects a significant false positive rate, likely caused by the scarcity of BAL training examples. Meanwhile, W. Kao et al. (2024) found that among various dimensionality-reduction methods and machine-learning classifiers, the combination of PCA and the XGBoost classifier represents the pinnacle of efficacy in the BAL quasar classification task, boasting impressive accuracy rates of 97.60% by 10-fold cross-validation and 96.92% on the external testing set. To improve label accuracy, Z. Guo & P. Martini (2019) refined their labels through multiple iterations involving training, prediction, and visual re-inspection of ambiguous cases. Their PCA-based CNN method achieves completeness comparable to that of N. Busca & C. Balland (2018), while effectively capturing narrower absorption troughs with widths less than 2000 km/s, as well as troughs extending to the center of the C IV emission line.

In contrast, `BALNet` exhibits a more balanced performance, achieving a completeness of 90.8% and a purity of 94.4%, which demonstrates its robustness in distinguishing BAL quasars from non-BAL quasars in mock data. The remarkable performance of `BALNet` can be attributed to its innovative architecture, which cleverly integrates 1D-CNN and Bi-LSTM networks. The 1D-CNN specializes in extracting local spectral features such as absorption troughs and continuum variations, whereas the Bi-LSTM effectively captures global spectral patterns by modeling long-range dependencies. This dual feature extraction strategy not only enhances feature robustness but also significantly improves the accuracy of BAL trough detection. More importantly, unlike general-purpose classifiers like `QuasarNet` and `FNet`

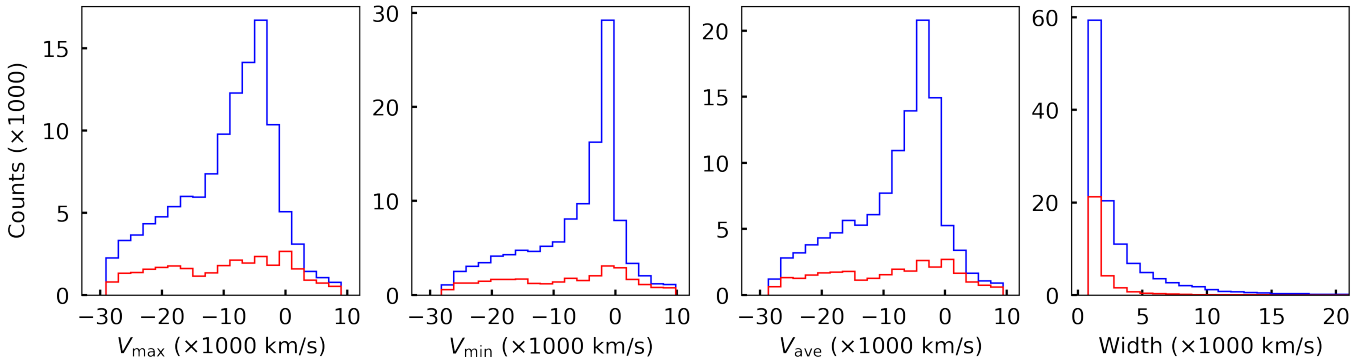

Figure 8. Velocity and width distributions of all BAL troughs (blue) identified by `BALNet` in the SDSS DR16Q spectra. The red curves highlight the BAL troughs corresponding to newly identified BAL quasars that are absent from the SDSS DR16Q catalog.

II, `BALNet` is explicitly designed for BAL trough detection—a specialized design that is crucial to its superior purity. Nevertheless, it is important to emphasize that the completeness and purity metrics reported in Table 2 reflect only the model's performance on the training dataset, and the quality of the labels in the training data directly determines the model's generalizability to real observational data. In this work, the BAL troughs and unabsorbed wavelength regions in the datasets used for `BALNet` training and testing are precisely labeled.

5. APPLICATION TO DR16Q

As demonstrated in the evaluation using the mock dataset, `BALNet` exhibits excellent performance in both BAL trough detection and velocity parameter estimation. To further assess its practical effectiveness, we applied the trained model to the quasar spectral data released by the SDSS DR16Q. The trained `BALNet` model (with the optimal probability threshold PIXEL_PROB = 0.3) was applied to a total of 446,839 quasar spectra from the SDSS DR16Q dataset within the redshift range of $1.5 \leq z \leq 5.7$. All spectra were preprocessed by performing linear interpolation to 1165 uniformly spaced sampling points within the wavelength range [1300, 1700] Å, followed by three-point moving average smoothing, to ensure consistency with the format of the simulated training set.

A total of 117,626 BAL troughs have been identified by `BALNet` in 91,164 quasars, indicating that 20.4% of the sources in the SDSS DR16Q sample are classified as BAL quasars. The velocity and width distributions of these BAL troughs are shown in Figure 8. As shown, our search covers a velocity range from blueshifts of 29,000 km/s to redshifts of 10,000 km/s, which is significantly broader than the velocity ranges typically explored in previous studies, where the focus was generally on blueshifts between 25,000 km/s and 3,000 km/s or up to 0 km/s. As a result, our sample includes a more diverse set of BAL troughs, comprising both high-velocity blueshifted and redshifted BAL troughs. For example, we identify 3,379 high-velocity blueshifted BAL troughs with mean velocities (V_{ave}) exceeding 25,000 km/s, as well as 10,370 redshifted BAL troughs with $V_{ave} \geq 0$ km/s.

According to the SDSS DR16Q catalog, B. W. Lyke et al. (2020) identified a total of 99,856 BAL quasars with BAL_PROB ≥ 0.75. These objects can be categorized into two distinct groups based on their C IV BAL properties: (1) 23,994 quasars with BI_CIV > 0, and (2) 75,702 quasars with BI_CIV $= 0$ and AI_CIV > 0. The former group serves as the parent sample from which we extract C IV BAL troughs, as described in Section 2.1. Each trough in the spectra of these sources is precisely measured using the pair-matching method. Among these BAL quasars, `BALNet` successfully recovered 98.3% of the sources. Of the 412 missed cases, the pair-matching method determined that 153 objects do not exhibit any BAL features, implying that the true missing rate of `BALNet` is only 1.1%. For the latter, comparison with the `BALNet` catalog reveals that 33,243 sources — accounting for 43.9% of the total — were rejected. According to the definitions of BI_CIV and AI_CIV in B. W. Lyke et al. (2020), these sources exhibit absorption troughs confined to the blueshifted velocity range of 0 - 3,000 km/s. This suggests that `BALNet` may still have limitations in detecting BAL troughs at low velocities, particularly those overlapping the C IV emission line. In addition, `BALNet` also newly identified 25,123 BAL quasars. The velocity and width distributions of their BAL troughs are shown as red curves in Figure 8. As illustrated, the newly detected BAL troughs are relatively uniformly distributed in velocity space, while the majority (77.8%) have widths ($V_{min} - V_{max}$) below 2,000 km/s.

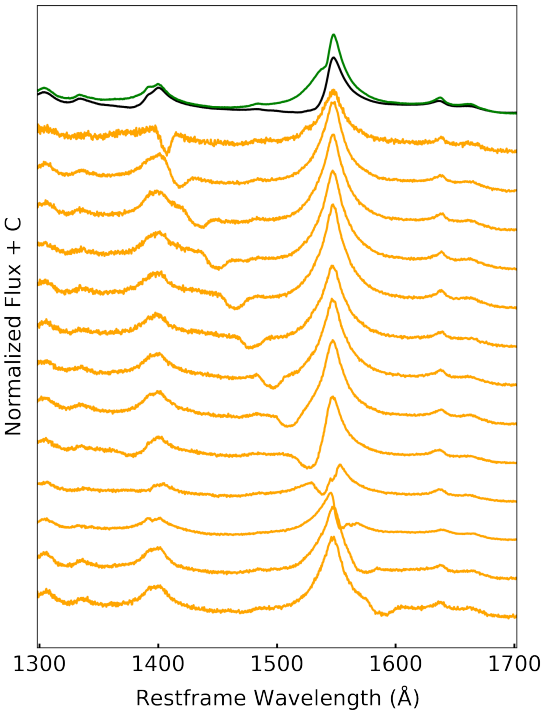

Figure 9. Composite spectra of different BAL quasar subsamples. The black and green curves represent the BAL quasars in SDSS DR16Q that are recovered and rejected by `BALNet`, respectively. The orange curves show a sequence of composite spectra for the BAL quasars newly identified by `BALNet`.

To further assess the reliability of BAL quasar identification, Figure 9 compares the composite spectra of the aforementioned BAL quasar subsamples. The black curve represents the composite spectrum of 66,041 BAL quasars identified by both `BALNet` and SDSS DR16Q, displaying prominent BAL troughs. In contrast, the green curve — corresponding to 33,655 BAL quasars classified by SDSS DR16Q but rejected by `BALNet` — shows only an extremely weak absorption feature on the blue wing of the C IV emission line. Given the relatively uniform velocity distribution of the newly detected BAL troughs by `BALNet`, we construct a series of composite spectra for thses BAL quasars with different average velocities ($V_{\rm ave}$), using velocity bins of 3,000 km/s. These are shown in orange. As shown, BAL troughs appear sequentially in the composite spectra from high to low velocities, transitioning from blueshifted to redshifted regions. This clearly confirms that the newly detected features are real and significantly more prominent than those in the `BALNet`-rejected sample, affirming their nature as genuine BAL troughs in the observed spectra. Figure 10 shows spectra randomly selected from these newly identified BAL quasars, which exhibit clearly significant BAL troughs.

Although the composite spectra have revealed that the rejected BAL quasars display only weak absorption features on the blue wing of the C IV emission line, while the newly identified sources exhibit prominent BAL troughs — demonstrating the robustness and reliability of the `BALNet` method — we still aim to further evaluate potential selection biases inherent in this approach. As described in Section 2.1, we identified a total of 47,267 BAL troughs and constructed the BAL pattern library. In Figure 11, we compare the velocity and width distributions of the troughs recovered by `BALNet` with those that were missed. It is clear that the missed troughs are not uniformly distributed in velocity space. A disproportionately large fraction of them occur within the velocity range dominated by the C IV emission line and near the extreme high-velocity end close to the measurement limits. This suggests that, when generating mock datasets, it may be necessary to increase the number of targets in these specific velocity regions to ensure sufficient sampling. Moreover, the missed troughs generally exhibit weaker absorption strengths (AI) and narrower widths, with approximately 90% having widths below 2,000 km/s, and nearly half of them narrower than 1,200 km/s. In addition, the spectra of these missed absorption troughs generally have lower spectral quality, with 55% of them exhibiting a signal-to-noise ratio below 3.0. During detection processing, we applied a three-point moving average smoothing to the original spectra, and to reduce computational costs, the size of the probability vector labels was also compressed when generating the training data. These compromises reduced the effective velocity resolution, causing some narrow troughs to fall below the detection threshold and be discarded. Additionally, the spectral signal-to-noise ratio also partially affects the detection results.

6. SUMMARY

In this study, we develop `BALNet`, an innovative deep learning framework that integrates a 1D-CNN and Bi-LSTM networks to automatically detect C IV BAL troughs in quasar spectra. Unlike conventional approaches that are typically limited to BAL quasar classification, `BALNet` not only identifies BAL quasars but also directly measures the velocities of their BAL troughs. To ensure robust training and evaluation, we constructed a more comprehensive simulated dataset by combining non-BAL quasar spectra and BAL troughs, both meticulously derived from the SDSS DR16Q data. Quantitative evaluations demonstrate the excellent performance of `BALNet` on the testing set, achieving a completeness of 83.0% and a purity of 90.7% in BAL trough detection (F1-score = 86.7%), as well as 90.8% complete-

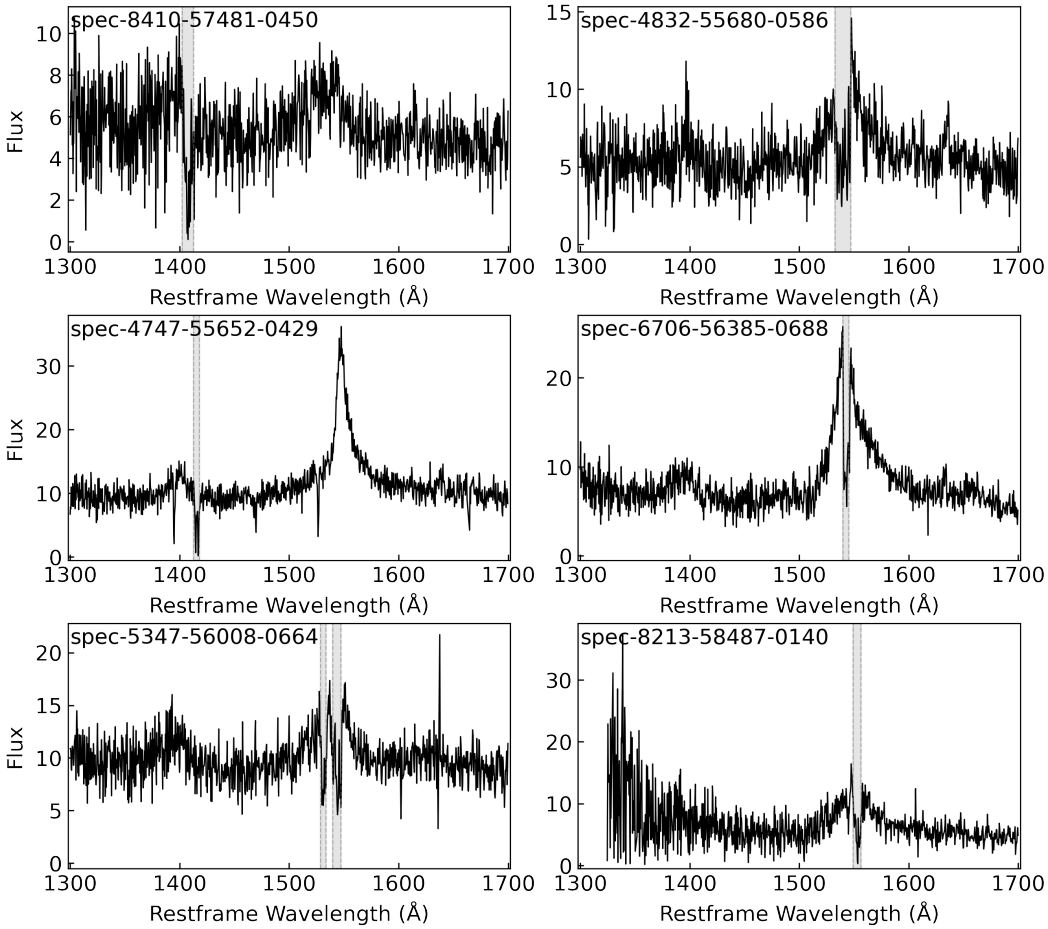

Figure 10. Examples of newly identified BAL quasars. Each panel displays the observed spectrum in black, with the BAL troughs detected by `BALNet` highlighted in gray-shaded regions; the last panel features a redshifted absorption trough.

ness and 94.4% purity in BAL quasar identification. The velocity measurements of the detected BAL troughs exhibit strong consistency with the ground-truth labels, confirming both the accuracy and reliability of `BALNet`.

Applied to 446,839 quasar spectra within the redshift range of $1.5 \leq z \leq 5.7$ from the SDSS DR16Q, `BALNet` identifies 91,164 BAL quasars, accounting for 20.4% of the sample and providing more comprehensive BAL trough coverage (8.8% are redshifted BAL systems). Compared to the DR16Q-classified BAL quasars (BAL_PROB $\geq$ 0.75), `BALNet` demonstrates excellent recovery performance, achieving a rate of 98.3% for sources with BI_CIV > 0, but only 56.1% for those with BI_CIV $= 0$ and AI_CIV > 0. These results suggest that while the model is highly effective and comparable to existing automated methods in identifying BAL quasars overall, significant discrepancies persist for sources with weak and narrow BAL troughs. Such troughs are particularly sensitive to spectral preprocessing (e.g., interpolation and smoothing) and to spectral resolution (e.g., the compression of the probability vector labels). In future work, targeted improvements in these two aspects are expected to significantly enhance the detection of narrow and weak absorption. Furthermore, `BALNet` newly identified 25,123 BAL quasars, whose composite spectra exhibit prominent absorption troughs, further validating the model's capability to detect previously unrecognized BAL features.

This sample can be used to systematic investigate the cosmic evolutions of BAL properties (e.g., M. Bischetti et al. 2022; M. Bischetti et al. 2023), as traced by C IV BAL troughs, as well as the dependence of the BAL fraction with quasar nuclear properties (e.g., continuum slope, luminosity, black hole mass and accretion rate) (e.g., J. R. Trump et al. 2006; R. Ganguly et al. 2007; S. Zhang et al. 2014). In addition, a subset comprising 20.7% of our sample with at least two-epoch observations provides a valuable opportunity to investigate the BAL trough variability (e.g., N. Filiz Ak et al. 2013; S. Zhang et al. 2015; Z. He et al. 2017). Monitoring changes in their strength, equivalent width, and velocity structure enables probing of the physical mechanisms

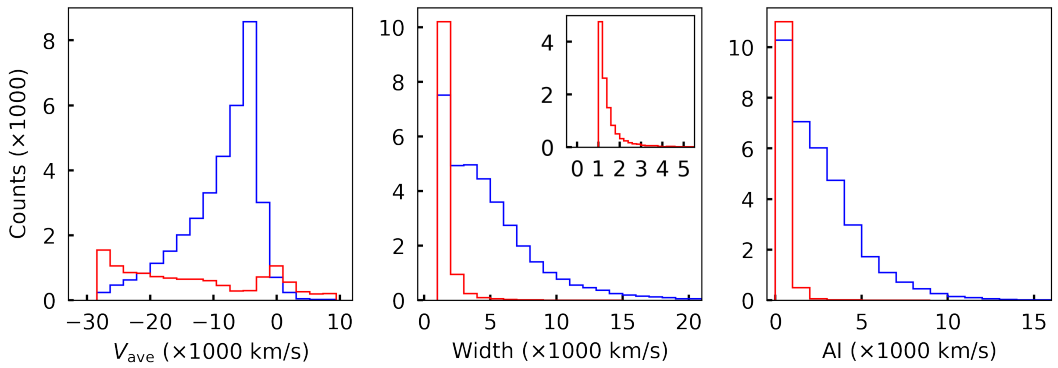

Figure 11. Comparison between BAL troughs recovered (blue) and those missed (red) by `BALNet` in the BAL pattern library. In the middle panel, the upper-right inset presents the results obtained from a finer binning (200 km/s) of the missed BAL troughs.

governing outflows, such as acceleration, deceleration, or transverse motion. Notably, redshifted BAL troughs account for up to 8.8% of all BAL troughs (10.9% of the sources), representing the first large-scale, systematic survey specifically targeting these rare features. Previous studies indicate that these redshifted BALs may originate either from rotating outflows close to the central black hole or from infalling (i.e., accreting) gas (e.g., P. B. Hall et al. 2013; X. Shi et al. 2016b; H. Zhou et al. 2019; G. Li et al. 2021), thus providing important clues to the gas kinematics near AGNs. Therefore, these systems offer valuable diagnostics for investigating black hole accretion processes. A comprehensive multi-wavelength analysis of BAL troughs—including Balmer lines as well as metastable He I and optical Fe II and Mg II in both optical and NIR bands—will be crucial for revealing the microphysical mechanisms that regulate black hole feeding.

In conclusion, the `BALNet` framework represents a significant advancement in BAL quasar research. It simultaneously identifies BAL quasars and measures their BAL troughs, while enhancing the efficiency of fitting unabsorbed spectra by leveraging the predicted properties of BAL troughs. Additionally, `BALNet` can select troughs with varying confidence levels by setting different PIXEL_PROB thresholds, thereby providing robust support for specific investigations. With its demonstrated reliability in BAL trough detection and characterization, `BALNet` provides astronomers with a powerful tool for analyzing large quasar spectral datasets from current and future surveys.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grant No. 12573009, 12173026 and 12141302), the Innovation Program of Shanghai Municipal Education Commission (Grant No. 2025GDZKZD04), and the scientific research grants from the China Manned Space Project (Grant No. CMS-CSST-2025-A06, CMS-CSST-2025-A07). S.H.Z. acknowledges the support from the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning and the Shuguang Program (23SG39) of Shanghai Education Development Foundation and Shanghai Municipal Education Commission. This work makes use of data from SDSS-IV. Funding for SDSS has been provided by the Alfred P. Sloan Foundation and Participating Institutions. Additional funding toward SDSS-IV has been provided by the U.S. Department of Energy Office of Science. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS website is www.sdss.org.

REFERENCES

Abadi, M., Barham, P., Chen, J., et al. 2016, in 12th USENIX symposium on operating systems design and implementation (OSDI 16), 265–283, doi: 10.48550/arXiv.1605.08695

Allen, J. T., Hewett, P. C., Belokurov, V., & Wild, V. 2008, in American Institute of Physics Conference Series, Vol. 1082, Classification and Discovery in Large Astronomical Surveys, ed. C. A. L. Bailer-Jones (AIP), 85–91, doi: 10.1063/1.3059090

Barrientos, A., Solar, M., & Mendoza, M. 2020, in Astronomical Society of the Pacific Conference Series, Vol. 522, Astronomical Data Analysis Software and Systems XXVII, ed. P. Ballester, J. Ibsen, M. Solar, & K. Shortridge, 385

Becker, R. H., Gregg, M. D., Hook, I. M., et al. 1997, ApJL, 479, L93, doi: 10.1086/310594

Bischetti, M., Feruglio, C., D'Odorico, V., et al. 2022, Nature, 605, 244, doi: 10.1038/s41586-022-04608-1

Bischetti, M., Fiore, F., Feruglio, C., et al. 2023, The Astrophysical Journal, 952, 44, doi: 10.3847/1538-4357/accea4

Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503, doi: 10.1086/591786

Brotherton, M. S., Tran, H. D., Becker, R. H., et al. 2001, ApJ, 546, 775, doi: 10.1086/318309

Busca, N., & Balland, C. 2018, arXiv e-prints, arXiv:1808.09955, doi: 10.48550/arXiv.1808.09955

Chaussidon, E., Yèche, C., Palanque-Delabrouille, N., et al. 2023, The Astrophysical Journal, 944, 107, doi: 10.3847/1538-4357/acb3c2

Chen, Z., He, Z., Ho, L. C., et al. 2022, Nature Astronomy, 6, 339, doi: 10.1038/s41550-021-01561-3

Chollet, F., & others. 2015, GitHub repository. https://github.com/fchollet/keras

Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, Research in Astronomy and Astrophysics, 12, 1197, doi: 10.1088/1674-4527/12/9/003

Czajkowski, M., Grześ, M., & Kretowski, M. 2014, Artificial Intelligence in Medicine, 61, 35, doi: https://doi.org/10.1016/j.artmed.2014.01.005

Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168, doi: 10.3847/1538-3881/ab089d

Filiz Ak, N., Brandt, W. N., Hall, P. B., et al. 2013, ApJ, 777, 168, doi: 10.1088/0004-637X/777/2/168

Fitzpatrick, E. L. 1999, Publications of the Astronomical Society of the Pacific, 111, 63, doi: 10.1086/316293

Foltz, C. B., Chaffee, F. H., Hewett, P. C., Weymann, R. J., & Morris, S. L. 1990, in Bulletin of the American Astronomical Society, Vol. 22, 806

Fushiki, T. 2011, Statistics and Computing, 21, 137, doi: 10.1007/s11222-009-9153-8

Ganguly, R., Brotherton, M. S., Cales, S., et al. 2007, The Astrophysical Journal, 665, 990, doi: 10.1086/519759

Gibson, R. R., Jiang, L., Brandt, W. N., et al. 2009, ApJ, 692, 758, doi: 10.1088/0004-637X/692/1/758

Glazebrook, K., Offer, A. R., & Deeley, K. 1998, ApJ, 492, 98, doi: 10.1086/305039

Guo, Z., & Martini, P. 2019, ApJ, 879, 72, doi: 10.3847/1538-4357/ab2590

Hall, P. B., Brandt, W. N., Petitjean, P., et al. 2013, MNRAS, 434, 222, doi: 10.1093/mnras/stt1012

Hamann, F., Chartas, G., Reeves, J., & Nardini, E. 2018, MNRAS, 476, 943, doi: 10.1093/mnras/sty043

Hamann, F., & Ferland, G. 1993, ApJ, 418, 11, doi: 10.1086/173366

Hazard, C., McMahon, R. G., Webb, J. K., & Morton, D. C. 1987, ApJ, 323, 263, doi: 10.1086/165823

He, Z., Wang, T., Zhou, H., et al. 2017, The Astrophysical Journal Supplement Series, 229, 22, doi: 10.3847/1538-4365/aa647a

He, Z.-C., Bian, W.-H., Ge, X., & Jiang, X.-L. 2015, MNRAS, 454, 3962, doi: 10.1093/mnras/stv2114

Hewett, P. C., & Foltz, C. B. 2003, AJ, 125, 1784, doi: 10.1086/368392

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. 2012, arXiv e-prints, arXiv:1207.0580, doi: 10.48550/arXiv.1207.0580

Hochreiter, S., & Schmidhuber, J. 1997, Neural Computation, 9, 1735, doi: 10.1162/neco.1997.9.8.1735

Hopkins, P. F., & Elvis, M. 2010, MNRAS, 401, 7, doi: 10.1111/j.1365-2966.2009.15643.x

Hopkins, P. F., & Hernquist, L. 2006, ApJS, 166, 1, doi: 10.1086/505753

Hu, L., Chen, X., & Wang, L. 2022, ApJ, 930, 70, doi: 10.3847/1538-4357/ac5c48

Jin, J., Wu, X., Fu, Y., et al. 2023, ApJS, 265, 25, doi: 10.3847/1538-4365/acaf89

Kao, W., Zhang, Y., & Wu, X. 2024, PASJ, 76, 653, doi: 10.1093/pasj/psae037

Kingma, D. P., & Ba, J. 2017, https://arxiv.org/abs/1412.6980

Li, G., Shi, X., Tian, Q., et al. 2021, ApJ, 916, 86, doi: 10.3847/1538-4357/ac06c8

Liu, C., Cui, W.-Y., Zhang, B., et al. 2015, Research in Astronomy and Astrophysics, 15, 1137, doi: 10.1088/1674-4527/15/8/004

Liu, W., Zhou, H., Ji, T., et al. 2015, The Astrophysical Journal Supplement Series, 217, 11, doi: 10.1088/0067-0049/217/1/11

Luo, Z., Li, Y., Lu, J., et al. 2024a, MNRAS, 535, 1844, doi: 10.1093/mnras/stae2446

Luo, Z., Tang, Z., Chen, Z., et al. 2024b, Monthly Notices of the Royal Astronomical Society, 531, 3539, doi: 10.1093/mnras/stae1397

Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, ApJS, 250, 8, doi: 10.3847/1538-4365/aba623

Moradi, R., Rastegarnia, F., Wang, Y., & Mirtorabi, M. T. 2024, MNRAS, 533, 1976, doi: 10.1093/mnras/stae1878

Pang, S., Kong, H., Li, Z., Kao, W., & Zhang, Y. 2025, Applied Sciences, 15, doi: 10.3390/app15031024

Quinlan, J. R. 1996, ACM Comput. Surv., 28, 71–72, doi: 10.1145/234313.234346

Rastegarnia, F., Mirtorabi, M. T., Moradi, R., Vafaei Sadr, A., & Wang, Y. 2022, Monthly Notices of the Royal Astronomical Society, 511, 4490, doi: 10.1093/mnras/stac076

Reichard, T. A., Richards, G. T., Schneider, D. P., et al. 2003, The Astronomical Journal, 125, 1711, doi: 10.1086/368244

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

Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, The Astrophysical Journal, 500, 525, doi: 10.1086/305772

Schuster, M., & Paliwal, K. K. 1997, IEEE Transactions on Signal Processing, 45, 2673, doi: 10.1109/78.650093

Shi, X., Jiang, P., Wang, H., et al. 2016b, ApJ, 829, 96, doi: 10.3847/0004-637X/829/2/96

Shi, X., Zhou, H., Shu, X., et al. 2016a, ApJ, 819, 99, doi: 10.3847/0004-637X/819/2/99

Shi, X., Pan, X., Zhang, S., et al. 2017, ApJL, 843, L14, doi: 10.3847/2041-8213/aa725e

Sookmee, P., Suwannajak, C., Techa-Angkoon, P., Panyangam, B., & Tanakul, N. 2020, in 2020 17th International Joint Conference on Computer Science and Software Engineering (JCSSE), 98–103, doi: 10.1109/JCSSE49651.2020.9268348

Tabasi, S. S., Salmani, R. V., Khaliliyan, P., & Firouzjaee, J. T. 2023, ApJ, 954, 164, doi: 10.3847/1538-4357/ace03f

Tolea, A., Krolik, J. H., & Tsvetanov, Z. 2002, The Astrophysical Journal, 578, L31, doi: 10.1086/344563

Trump, J. R., Hall, P. B., Reichard, T. A., et al. 2006, ApJS, 165, 1, doi: 10.1086/503834

Vivek, M., Srianand, R., Petitjean, P., et al. 2014, MNRAS, 440, 799, doi: 10.1093/mnras/stu288

Vivek, M., Srianand, R., Petitjean, P., et al. 2012, MNRAS, 423, 2879, doi: 10.1111/j.1365-2966.2012.21098.x

Voit, G. M., Weymann, R. J., & Korista, K. T. 1993, ApJ, 413, 95, doi: 10.1086/172980

Wang, K., Guo, P., & Luo, A.-L. 2016, Monthly Notices of the Royal Astronomical Society, 465, 4311, doi: 10.1093/mnras/stw2894

Weymann, R. J., Morris, S. L., Foltz, C. B., & Hewett, P. C. 1991, ApJ, 373, 23, doi: 10.1086/170020

Xu, B., Wang, N., Chen, T., & Li, M. 2015, arXiv e-prints, arXiv:1505.00853, doi: 10.48550/arXiv.1505.00853

Yang, H., Zhou, L., Cai, J., et al. 2023, MNRAS, 518, 5904, doi: 10.1093/mnras/stac3292

York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, AJ, 120, 1579, doi: 10.1086/301513

Zaremba, W., Sutskever, I., & Vinyals, O. 2014, arXiv e-prints, arXiv:1409.2329, doi: 10.48550/arXiv.1409.2329

Zhang, M.-L., & Zhou, Z.-H. 2007, Pattern Recognition, 40, 2038, doi: 10.1016/j.patcog.2006.12.019

Zhang, N.-X., Brandt, W. N., Ahmed, N. S., et al. 2017, ApJ, 839, 101, doi: 10.3847/1538-4357/aa6842

Zhang, S., Wang, H., Wang, T., et al. 2014, The Astrophysical Journal, 786, 42, doi: 10.1088/0004-637X/786/1/42

Zhang, S., Wang, H., Zhou, H., Wang, T., & Jiang, P. 2011, Research in Astronomy and Astrophysics, 11, 1163, doi: 10.1088/1674-4527/11/10/005

Zhang, S., Wang, T., Wang, H., et al. 2010, The Astrophysical Journal, 714, 367, doi: 10.1088/0004-637X/714/1/367

Zhang, S., Zhou, H., Wang, T., et al. 2015, ApJ, 803, 58, doi: 10.1088/0004-637X/803/2/58

Zhou, H., Shi, X., Yuan, W., et al. 2019, Nature, 573, 83, doi: 10.1038/s41586-019-1510-y