Mapping circumgalactic medium observations to theory using machine learning

Sarah Appleby [1],[1]★ Romeel Davé,[1,2,3] Daniele Sorini [1],[1,4,5] Christopher C. Lovell [1],[6] and Kevin Lo[1]

[1]*Scottish Universities Physics Alliance (SUPA), Institute for Astronomy, University of Edinburgh, Royal Observatory, Edinburgh EH9 3HJ, UK*
[2]*University of the Western Cape, Department of Physics and Astronomy, Bellville, Cape Town 7535, South Africa*
[3]*South African Astronomical Observatories, Observatory, Cape Town 7925, South Africa*
[4]*Département de Physique Théorique, Université de Genève, 24 quai Ernest Ansermet, CH-1211 Genève 4, Switzerland*
[5]*Institute for Computational Cosmology, Department of Physics, Durham University, South Road, Durham DH1 3LE, UK*
[6]*Institute of Cosmology and Gravitation, University of Portsmouth, Burnaby Road, Portsmouth PO1 3FX, UK*

ABSTRACT

We present a random forest (RF) framework for predicting circumgalactic medium (CGM) physical conditions from quasar absorption line observables, trained on a sample of Voigt profile-fit synthetic absorbers from the SIMBA cosmological simulation. Traditionally, extracting physical conditions from CGM absorber observations involves simplifying assumptions such as uniform single-phase clouds, but by using a cosmological simulation we bypass such assumptions to better capture the complex relationship between CGM observables and underlying gas conditions. We train RF models on synthetic spectra for H I and selected metal lines around galaxies across a range of star formation rates, stellar masses, and impact parameters, to predict absorber overdensities, temperatures, and metallicities. The models reproduce the true values from SIMBA well, with normalized transverse standard deviations of 0.50–0.54 dex in overdensity, 0.32–0.54 dex in temperature, and 0.49–0.53 dex in metallicity predicted from metal lines (not H I), across all ions. Examining the feature importance, the RF indicates that the overdensity is most informed by the absorber column density, the temperature is driven by the line width, and the metallicity is most sensitive to the specific star formation rate. Alternatively examining feature importance by removing one observable at a time, the overdensity and metallicity appear to be more driven by the impact parameter. We introduce a normalizing flow approach in order to ensure the scatter in the true physical conditions is accurately spanned by the network. The trained models are available online.

Key words: galaxies: evolution – galaxies: general – galaxies: haloes – quasars: absorption lines.

1 INTRODUCTION

Over recent years, there has been much effort to characterize the CGM via quasar absorption line studies (see reviews by Putman, Peek & Joung 2012; Tumlinson, Peeples & Werk 2017; Péroux & Howk 2020). Many of the studies probe the strong transitions that exist in the rest ultraviolet (UV) regime and which trace cool or warm gas. Such studies are motivated by the wish to understand the baryon cycle of gas flows in the CGM: accretion on to galaxies from the IGM and satellite galaxies; expulsion of gas via stellar winds and active galactic nuclei feedback; recycling of previously ejected material back on to galaxies.

The physical conditions of the CGM are studied by retrieving kinematics, spatial distributions, metallicities, densities, and temperatures from the absorption features (e.g. Stocke et al. 2013; Lehner et al. 2014, 2018, 2019; Savage et al. 2014; Werk et al. 2014; Wotta et al. 2016, 2019; Keeney et al. 2017; Prochaska et al. 2017; Qu et al. 2022). To extract physical conditions, absorption

systems are commonly fitted with Voigt profiles to model each absorption component and obtain column densities, linewidths, and redshift-space positions. By running ionization models (typically using CLOUDY, Ferland et al. 2017) and varying the input physical parameters, a Bayesian search can be performed across parameter space for the physical conditions of each absorber component using the ensemble of absorption properties as constraints. In such models the clouds are often modelled as plane–parallel slabs of gas with an ionizing flux incident on one face, making the (simplifying) assumption that each cloud is spatially isolated with single-valued properties (e.g. Churchill, Vogt & Charlton 2003; Tripp et al. 2008; Werk et al. 2014; Fumagalli, O'Meara & Prochaska 2016; Keeney et al. 2017; Prochaska et al. 2017).

The analysis and interpretation of CGM observations poses many challenges owing to the complex nature of the halo environment. The shapes of absorption profiles are sensitive to the underlying phase structure and likely contain contributions from different phases, for example due to the motion of gas within the halo and the clumpy gas structure. Even within individual absorber systems the metallicity of the absorbing gas can vary and multiple gas phases may be present (Lehner et al. 2019; Zahedy et al. 2019; Chen et al. 2020; Sankar et al.

★ E-mail: sarahappleby20@gmail.com

2020; Haislmaier et al. 2021; Sameer et al. 2021). Detailed analysis of absorption systems can give relative abundances of different ions that constrain the physical conditions, but this requires high-resolution spectroscopy. Studies that use this technique have moved away from the assumption of a single cloud, by modelling the high and low excitation ions separately (Zahedy et al. 2019; 2021; Haislmaier et al. 2021; Qu et al. 2022), or by modelling the absorption components as arising from multiple clouds (Cooper et al. 2021; Sameer et al. 2021; Nielsen et al. 2022). Interpreting the observational picture is further complicated due to the sensitivity of density and metallicity estimates to the shape of the UVB (Oppenheimer & Schaye 2013; Acharya & Khaire 2022; Gibson et al. 2022). Furthermore, particular ions are not necessarily produced by the same structures and processes at different redshifts due to the evolving UVB (Haardt & Madau 2012; Faucher-Giguère 2020).

Galaxy formation simulations provide a valuable theoretical perspective on these problems as they offer complete particle data and physical properties for the gas that makes up the CGM, making it possible to directly interpret observations. A range of UV metal lines have been used to probe the cool and warm ionized CGM in simulations, testing specific stellar wind implementations (Ford et al. 2013, 2014, 2016; Hummels et al. 2013), the NIHAO simulation suite (Gutcke et al. 2017), EAGLE (Oppenheimer et al. 2016, 2018), IllustrisTNG (Nelson et al. 2020; DeFelippis et al. 2021), FIRE-2 (Li et al. 2021), and SIMBA (Appleby et al. 2021, 2023).

Such simulations can be useful for examining the impact of different line analysis methods on the retrieved CGM gas conditions (e.g. Churchill et al. 2015; Liang, Kravtsov & Agertz 2018). In a recent analysis of a sample of synthetic absorption lines from a cosmological simulation, Marra et al. (2021) tested the accuracy of the single cloud ionization modelling method of retrieving physical gas conditions. The authors find that while there is general agreement between intrinsic conditions and those derived from ionization modelling, such methods capture the average properties of absorbing gas cells, consistent with observational tests by Sameer et al. (2021) comparing single-phase and multiphase modelling. Marra et al. (2022) followed up by testing the assumption of single spatially isolated absorbing clouds in the CGM, showing that several distinct absorbing clouds may be present within a single absorption component. The distinct clouds may arise from gas of different phases that happen to be aligned kinematically. These results demonstrate that the CGM is a complex environment, with non-linear relationships between the underlying CGM conditions and the resulting absorption observables.

Machine learning (ML) algorithms have the capacity to learn complex, non-linear relationships and as such they have been widely applied to astrophysical problems (see review by Fluke & Jacobs 2020). In this paper, we explore a novel approach for cosmological simulations to aid in interpreting CGM absorption observations using ML models. We present a framework for random forest (RF) mapping between synthetic CGM absorption observables from the SIMBA simulation (Davé et al. 2019) and the underlying absorber conditions from particle data. Such a mapping has the potential to be employed as a useful tool in retrieving physical conditions from real, multicomponent absorption observations. This approach eliminates the need for simplifying assumptions about the structure and state of the gas, i.e. whether absorption arise from single or multiple gas phases. Instead the RF mappings implicitly assume the veracity of the SIMBA galaxy formation model and our choice of UVB (Faucher-Giguère 2020) to produce its predictions.

The SIMBA simulations accurately reproduce a variety of observational galaxy properties. At low redshift, these include the star-

forming main sequence, black hole-galaxy co-evolution, radio galaxy populations, dust properties, cold gas properties, and the baryonic Tully–Fisher relation (Davé et al. 2019, 2020; Li, Narayanan & Davé 2019; Thomas et al. 2019; Appleby et al. 2020; Glowacki, Elson & Davé 2020; Lovell et al. 2021; Thomas et al. 2021). On larger mass scales, SIMBA reproduces X-ray scaling relations for massive haloes (Robson & Davé 2020) and low redshift Lyα absorption statistics of the IGM (Christiansen et al. 2020). In previous work, we have shown that SIMBA also broadly reproduces the observed absorption properties of H I (Sorini, Davé & Anglés-Alcázar 2020) and selected metal lines in the CGM (Appleby et al. 2021), and that such absorption arises from physically reasonable gaseous conditions (Appleby et al. 2023); therefore SIMBA is a reasonable choice of simulation with which to explore the capabilities of ML methods to learn relationships in the CGM.

None the less, there is no guarantee SIMBA yields fully accurate and representative circumgalactic media. Indeed, CGM zoom simulations suggest that SIMBA's resolution may be too poor to capture finer details of multiphase gas, particularly for stronger absorbers (e.g. Suresh et al. 2019; van de Voort et al. 2019, though see Nelson et al. 2020). This drawback could be explored via comparing the results of this framework applied to other simulations. We leave this aspect for future work, and here focus on presenting the general framework and its results when applied to the SIMBA model.

In this paper, we train RF ML networks on the low-redshift SIMBA CGM absorber sample presented in Appleby et al. (2023) to produce predictions for the underlying gas conditions in the CGM. This paper is organized as follows. In Section 2, we present the SIMBA simulations. In Section 3, we describe the galaxy selection, spectrum generation, and fitting processes. In Section 4, we describe the RF model and training process. In Section 5, we examine the accuracy of the RF models. In Section 6, we examine the feature importance of the RF models. In Section 7, we present the RF predictions in phase space. Finally in Section 8 we conclude and summarize.

2 SIMULATIONS

SIMBA (Davé et al. 2019) is a suite of state-of-the-art cosmological simulations that is the successor to the MUFASA simulations (Davé, Thompson & Hopkins 2016), with the major additions being the inclusion of two-mode black hole growth and three-mode black hole feedback, along with an on-the-fly dust evolution model. The main simulation, and the one employed in this work, contains 1024^3 gas cells and the same number of dark matter particles within a $(100\ h^{-1}\ \mathrm{Mpc})^3$ volume. This yields a particle mass resolution of $1.8 \times 10^7\ \mathrm{M_\odot}$ per gas cell, and $9.6 \times 10^7\ \mathrm{M_\odot}$ per dark matter particle, with a spatial resolution of $\approx 1\ h^{-1}$ kpc in the densest regions. Since SIMBA has been extensively described in many previous works, and since the primary goal on this work is to present and explore our machine learning framework that is not crucially dependent on which simulation it is applied to, for brevity we do not present all of SIMBA's input physics, but rather refer readers to Davé et al. (2019), Thomas et al. (2019), and Li et al. (2019) for full details.

3 ABSORBER SAMPLE

In this work, we use the sample of $z = 0$ absorbers from our investigation into the physical conditions of absorbing halo gas in Appleby et al. (2023). Here, we summarize the procedure for generating the absorber sample. We select a sample of central galaxies within the fiducial SIMBA volume that evenly sample a range of global galaxy properties. Central galaxies are defined as

the one with the highest stellar mass among those in the halo; in practice, they are mostly within $\lesssim 0.1r_{200}$ from the halo centre. The galaxies fall into three categories based on their star formation rates: star forming, green valley, and quenched. We define star-forming galaxies as with $\log(\text{sSFR/Gyr}^{-1}) > -1.8 + 0.3z$ for consistency with previous work with the SIMBA simulation (e.g. Thomas et al. 2019), define green valley galaxies as within 1 dex below the star-forming galaxy threshold, and define quenched galaxies as those having zero star formation. We further define six $M_\star$ bins of width 0.25 dex, with a minimum of $M_\star > 10^{10}\,\text{M}_\odot$ to ensure well-resolved systems. In Appleby et al. (2023), we selected 12 galaxies from each of the 18 $M_\star$ – SFR bin. Here we select a further 12 galaxies in each to double the underlying galaxy sample (except in the highest mass star forming and green valley bins, which have only 23 and 8 galaxies, respectively) to increase the sample available for training a machine learning mapping.

For each central galaxy, we generate synthetic line of sight (LOS) absorption spectra through the simulation volume at a range of r_{200}-normalized impact parameters ($r_\perp$), probing both the inner and outer CGM ($r_\perp/r_{200} = 0.25, 0.5, 0.75, 1.0, 1.25$). In addition, for each $r_\perp$, we select 8 equally spaced LOS in a circle around the galaxy. Thus, for each galaxy in our sample we generate 40 LOS spectra for each of the following ions, selected to probe a range of excitation energies: H I 1215 Å, Mg II 2796 Å, C II 1334 Å, Si III 1206 Å, C IV 1548 Å, and O VI 1031 Å. This results in a total sample of 16 600 lines of sight.

The spectra are generated along the z-axis of the simulation using the PYGAD analysis package (Röttgers et al. 2020); the procedure is as follows. Gas elements whose smoothing lengths intersect with the LOS are identified and their ionization fractions obtained, using look up tables that are generated with version 17.01 of the CLOUDY cloud simulation code (Ferland et al. 2017) using Cloudy Cooling Tools.[1] We assume a spatially uniform Faucher-Giguère (2020) photoionizing UV background spectrum, since it was shown in Christiansen et al. (2020) to provide the best match to low-redshift Ly α absorption. Self-shielding for H I is applied during the simulation run, but for generating the metal lines we employ the Rahmati et al. (2013) prescription to attenuate the ionizing background strength based on the local density.

Ion densities for each gas element are obtained by multiplying the gas densities by each species' ionization fractions. The mass fractions of each element are individually tracked within SIMBA, based on yields from Type II and Ia supernovae and stellar evolution. Metals are carried out into the CGM primarily by stellar feedback processes, since winds are mass and metal-loaded (Appleby et al. 2021). The ion densities are smoothed along the LOS into pixels of width 2.5 km s^{-1}, using the same spline kernel used in the GIZMO simulation code and the gas elements' individual smoothing lengths and metal masses (for metal lines). Optical depths are then computed from the column densities at a pixel scale, using the oscillator strength for each species. We exclude wind particles since those gas elements are hydrodynamically decoupled from the surrounding gas, which represent a very small fraction of the CGM mass (Appleby et al. 2021). PYGAD also computes column density-weighted physical density, temperature, metallicity, and peculiar velocity in the same manner within the LOS pixels.

We identify regions of absorption within a ± 600 km s^{-1} window centred on the galaxy by computing the detection significance ratio of each pixel, defined as the Gaussian-smoothed flux equivalent width (EW) divided by the Gaussian-smoothed noise EW. Regions are identified as contiguous intervals where the flux drops below the level of the continuum with an overall significance ratio of $>4\sigma$, ensuring that the edges of the regions begin at the continuum and merging nearby regions within 2 pixels of one another.

We fit a superposition of Voigt profiles to each absorption region in order to extract the absorption line observables: the column density N, the Doppler b parameter, the wavelength (or velocity) location along the LOS, and the EW. For the fitting, absorption lines are added to the model fit one at a time, with initial guesses for the line parameters that depend on whether or not the absorption is saturated. For non-saturated absorption, the line is placed at the position of lowest flux, and the initial N and b is based on the depth and velocity width of the local flux minimum. For saturated absorption, the line is placed in the middle of the saturated trough, and the initial N and b are chosen from a coarse grid in order to minimize the reduced chi-square (χ_r^2), computed assuming a signal-to-noise per pixel of 30. This procedure broadly follows that in AutoVP (Davé et al. 1997). The best-fitting Voigt parameters that minimize χ_r^2 are then found using the `scipy.optimize` subpackage.[2] Loose prior bounds on N and b are set based on typical H I and metal line column densities and thermal line widths from $10^4 - 10^7$ K.

If the fit has $\chi_r^2 < 2.5$ then the model is accepted; otherwise we identify the next strongest area of absorption by subtracting the model from the data and place a line at the residual minimum. We repeat the process until an acceptable model is found, up to a maximum of 10 absorption lines per region. Each line must improve the χ_r^2 of the model by at least 5 per cent; the process is halted if 2 consecutive additional lines do not improve the χ_r^2 by at least this margin. If after 10 lines an acceptable model is not found then we adopt the model with the number of lines that performed best. We again check that each line improves the χ_r^2 of the model by iteratively recomputing the χ_r^2 with each line removed; if the χ_r^2 acceptance threshold is reached, or the χ_r^2 increases by less than 5 per cent then the line is removed from the solution. In this way we attempt to obtain a satisfactory fit with the fewest number of absorption lines.

PYGAD not only produces optical depths at each pixel, but also outputs the optical depth-weighted density, temperature, and metallicity (in the relevant element, or the total metallicity in the case of H), as described in Appleby et al. (2023). This enables us to assign physical properties to absorption features. In our case, we assign to each absorber the physical properties associated with the pixel closest to its Voigt profile fitted line centre. We prefer this to a weighting or interpolation scheme, because if there are significant variations in the physical properties between pixels then interpolation can yield values that are less physically meaningful (Ford et al. 2014). None the less, this is an inherently approximate procedure, which is only exact in the unphysical situation that each Voigt profile represents a distinct uniform cloud of absorbing gas. We further note that sub-resolution phenomena such as small-scale turbulence is not accounted for, which could additionally blur the relationship between the absorber temperature and the Voigt profile line width.

The galaxy selection, spectrum generation, and LOS fitting pipeline results in our sample of absorbers. We find that adopting the same strict χ_r^2 limit for all ions results in an incomplete sample. As such we compute the EW directly for each LOS and adopt an upper χ_r^2 threshold for each ion such that we recover 90 per cent of the total EW across all LOS for each species. The sample size and χ_r^2 upper limits

---

[1]https://github.com/brittonsmith/cloudy_cooling_tools

[2]https://docs.scipy.org/doc/scipy/reference/optimize.html

Table 1. Absorber sample properties for the RF models: the number of absorbers below the χ_r^2 limit for each species; the column density completeness limit; the χ_r^2 below which we recover 90 per cent of the total EW; the median χ_r^2 of all absorbers; the excitation energy of the species.

Species	n	$\log(N_{\min}/\text{cm}^{-2})$	$\chi_r^{2.90}$	Median χ_r^2	E(eV)
H I	17 750	12.7	3.5	0.7	13.60
Mg II	5306	11.5	39.8	1.0	15.04
C II	11 062	12.8	15.8	1.3	24.38
Si III	14 119	11.7	35.5	1.9	33.49
C IV	17 463	12.8	6.3	1.2	64.49
O VI	17 463	13.2	4.0	1.2	138.12

($\chi_{r,90}^2$) for each ion are given in Table 1. In practice the typical χ_r^2 for a given region is much lower than these upper limits; the median χ_r^2 of absorption lines in our sample is also shown in Table 1. We also adopt the column density completeness limits from Appleby et al. (2023), which are computed by fitting the power-law portion of the column density distribution function (CDDF) for each ion and identifying where the CDDF falls below 50 per cent of the expectation at low column densities. The completeness limits are given in Table 1. We note that routines to do the spectrum generation, absorption region identification, and Voigt profile fitting are all contained with the publicly available PYGAD package (Röttgers et al. 2020).

4 RANDOM FOREST METHODS

4.1 Random forest regression

RF regression (Breiman 2001) is a supervised, decision tree-based, ensemble method of machine learning. The term 'ensemble method' refers to the process of combining predictions from several machine learning runs (in this case, individual decision trees) in order to more accurately predict the output. Decision trees work in a top–down manner, in which the best split for the data is found by minimizing a cost function. They have the advantage of being easy to interpret and have low bias in their predictions for the training data. However, individual decision trees are prone to overfitting to the training data, hence their predictions for new data have high variance.

The RF algorithm counteracts this effect by constructing many decision trees, each trained on a subset of the data. Random forests may be used for both classification and regression problems; in this work we use RF in its regression mode to deal with our continuous target predictors. The training data subsets are randomly chosen with replacement, and their outputs averaged for an overall prediction in a process known as bootstrap aggregation ('bagging', see Breiman 1996). In this way, RF models retain the low bias of a decision tree, while also minimizing the variance on predictions for new data. Training a single decision tree is considerably faster, however such models are less reliable, particularly when trained on non-linear data (such as the absorber data used here). In this work, we use the Scikit-Learn (Pedregosa et al. 2011) module's RF implementation, RandomForestRegressor.

RF models are widely used in a range of astronomical applications, and have been remarkably successful given the relative simplicity of the approach. The advantage of RF models over other methods (e.g. Neural Network based algorithms) is in the interpretability of the output models, as they indicate the relative importance of the input variables in reaching a prediction. In galaxy formation, RFs (and related tree-based methods) have been widely used for regression problems using both simulation and observational data, for example in predicting the properties of large-scale structure (Lucie-Smith et al.

2018; Li et al. 2022; Lovell et al. 2022) and the properties of galaxies and haloes (Ucci et al. 2017; Nadler et al. 2018; Rafieferantsoa, Andrianomena & Davé 2018; Cohn & Battaglia 2020; Moews et al. 2021; Mucesh et al. 2021; Delgado et al. 2022; McGibbon & Khochfar 2022).

4.2 Input features and target predictors

For each of the ions in our selection, we train a RF model on the data set of simulated CGM absorbers to predict their underlying physical gas conditions. We do this separately for each of the 6 ions we consider, such that the usefulness of this pipeline is not contingent on having line information simultaneously for all 6 ions. We exclude absorbers where the quality of the Voigt profile fit is low (i.e. the fit has a χ_r^2 above the acceptable threshold for that ion) and the column density is below the completeness limit.

For each ion, we use the same set of input features and target predictors. The input features are chosen from among the properties of the CGM absorbers and their central galaxies. Included features which describe the absorbers themselves are: the column density (N), the EW, the linewidth (b), the velocity separation from the host galaxy (dv), and the impact parameter, expressed as a fraction of halo virial radius ($f_{r200} = r_\perp/r_{200}$). In principle, the EW information is fully contained within N and b, but we provide it separately since sometimes combinations of parameters can be easier for the ML to utilize, and also so we can compare between N and EW to see which measure of overall absorption is more important.

Properties of the central galaxy that are included as input features are the stellar mass ($M_\star$), the specific star formation rate (sSFR), and the fraction of kinetic energy contained in rotation ($\kappa_{\rm rot}$, Sales et al. 2012), which Kraljic, Davé & Pichon (2020) found is a reasonable proxy for visual morphology.

From these 8 input features, we predict 3 target gas predictors: the overdensity ($\delta = \rho/\bar{\rho}_m$), temperature ($T$), and metallicity ($Z$). Each of these is a column density-weighted average at the nearest LOS pixel to the absorber, computed at the time of spectral generation and binned along the LOS (see Section 3).

4.3 Training

Each of the features is transformed into log space; Jo & Kim (2019) showed that transforming quantities into log space improves the accuracy of machine learning predictions for astronomy problems, owing to the wide range of physical scales present in astronomical data. Exceptions to this are dv, f_{r200}, and $\kappa_{\rm rot}$; dv and $\kappa_{\rm rot}$ have nearly uniform intrinsic distributions, while f_{r200} consists of 5 specific values due to our choices of LOS. In addition, we standardize the input and output data by subtracting the mean of the distribution and scaling the variance to unity in each case. We deal with zeros in our data set by setting them to a small non-zero value; in practice this can only be the case for sSFR, whereby we assign sSFR $= 10^{-14}$ $M_\odot$ yr^{-1}. For each ion, we divide the absorber data set into 80 per cent training data used to build the RF model, and 20 per cent test data used to evaluate the performance of the model. Where multiple absorbers arise from the same LOS these can be separated into the training and test data sets; this mitigates overfitting in the model due to galaxy or LOS properties.

We train the RF model separately for each target feature, as we find that this improves the accuracy of the prediction. We separately tune the hyperparameters of each RF model to optimize the model accuracy, using Scikit-Learn's GridSearchCV method to perform an exhaustive grid search over hyperparameter space. The

hyperparameters are the number of trees, the minimum number of data points required in order to split the data, and the minimum number of data points in each resulting split. For each set of hyperparameters, a k-fold cross-validation is performed with $k = 5$, in which the training data is split into k 'folds', and k RF models are iteratively constructed using $k - 1$ folds of the data; the overall score for each set of hyperparameters is the average of each of the k RF models. The coefficient of determination R^2 is used internally at this hyperparameter tuning stage to evaluate the performance of each model, given n data points and true and predicted quantities X_{true} and $X_{\text{predicted}}$:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}, \tag{1}$$

where RSS is the residual sum of squares:

$$\text{RSS} = \sum_{i=1}^{n} (X_{\text{true}}^i - X_{\text{predicted}}^i)^2 \tag{2}$$

and TSS is the total sum of squares:

$$\text{TSS} = \sum_{i=1}^{n} (X_{\text{true}}^i - \langle X_{\text{true}} \rangle)^2. \tag{3}$$

The mean squared error, $\text{MSE} = \text{RSS}/n$, is the cost function used to determine the best decision tree splits. An MSE of zero represents a perfectly accurate prediction.

5 PREDICTIVE ACCURACY

Here, we assess the performance of each of the RF models. Fig. 1 shows the test data RF predictions for the H I absorber physical conditions (δ, T, and Z) against the true values. The colour scale of the hexagonal bins indicates the number of data points in each bin. The black diagonal dashed line represents the 1:1 case of a perfectly predicting model. In each panel, the 1D histograms for the true and predicted values lie along the top and right, respectively. The accuracy of the model is summarized in each panel with three quantities: (1) $\sigma_{\perp,\text{norm}}$, the scatter perpendicular to the perfect 1:1 relation, normalized to the scatter in the true values; (2) the Pearson correlation coefficient ρ_r, given by

$$\rho_r = \frac{\text{cov}(X_{\text{true}}, X_{\text{predicted}})}{\sigma_{X\text{true}} \sigma_{X\text{predicted}}}, \tag{4}$$

where σ_{Xi} is the standard deviation of X_i; and (3) MSE_{norm}, the mean squared error normalized to the scatter in the true values. High correlation is preferred, but does not necessarily indicate an accurate prediction as the outputs could have a systematic offset. We normalize $\sigma_{\perp}$ and MSE in order to compare results between H I and the metals lines, since H I lines trace a wider range of physical conditions.

Beginning with the predictions for H I absorbers, density, and temperature are well-predicted by the ML model. True values are highly correlated with the predictions, and the model predictions have low scatter and error. Density and temperature are physically correlated with one another and have Gaussian distributions. Of the two, temperature ($\sigma_{\perp,\text{norm}} = 0.33$ dex, $\text{MSE}_{\text{norm}} = 0.13$) is predicted more accurately than overdensity ($\sigma_{\perp,\text{norm}} = 0.38$ dex, $\text{MSE}_{\text{norm}} = 0.22$). The RF models for H I density and temperature perform particularly well considering the models' relative simplicity (compared with e.g. a NN-based model). Aside from transforming the features into log space and using the k-fold hyperparameter cross-validation, the model has not been extensively tuned by hand. As such, these results represent a basic model which demonstrate the

capability of RF models to predict gas conditions, which could be improved upon with further tuning. We have also explored alternative ML approaches such as NNs and CNNs, and found that such models do not offer a substantial improvement in terms of predictive accuracy and take considerably longer to run. This has lead us to favour the RF model for its simplicity, speed, and the degree of interpretability in the form of feature 'importances' (see Section 6).

The predictions for H I metallicities are less accurate ($\sigma_{\perp,\text{norm}} = 0.42$ dex, $\text{MSE}_{\text{norm}} = 0.42$). In general, points with $\log Z/Z_{\odot} < -1$ are overpredicted, while points with $\log Z/Z_{\odot} > -1$ are underpredicted. This points to the general tendency of our ML models to output a narrower predicted distribution than in the input data set (this behaviour is also seen to a lesser extent in the density and temperature predictions). This means that the tails of the original distributions are not well captured in the ML model, perhaps as a result of sparse training data at the extremes. Perhaps the poor prediction is unsurprising since H I absorption is not Z-dependent, unlike metal lines which by necessity arise from metal-enriched gas. Therefore it was not obvious that any relationship between H I absorption and metallicity could have been learned from the data. The learned mapping in the metallicity RF model likely arises from the provided galaxy properties and H I absorption strength; we will explore the input feature importance later (Section 6).

The metal line absorber physical conditions are also reasonably well predicted. Figs 2 and 3 show the performance of the RF models for predicting C II and C IV absorber conditions, using the same plot structure as above. The performance for Mg II, Si III, and O VI absorbers are shown in Appendix A. The RF models perform similarly well among all the metal lines, with the same tendency to predict a more concentrated distribution of values than in the original data. In general, for each metal line the predictions are less well correlated with the truth values than for H I; metal line absorber δ, T, and Z have median correlation coefficients of $\rho_r = 0.68, 0.71, 0.68$, respectively, compared with $\rho_r = 0.85, 0.88, 0.81$ for H I. In addition, the scatter is higher in general for the metal line RF models, with median $\sigma_{\perp,\text{norm}} = 0.52, 0.44, 0.52$ for δ, T, and Z, compared with $\sigma_{\perp,\text{norm}} = 0.38, 0.33, 0.42$ for H I. However, the errors in the metal line predictions are comparable with those for H I: median $\text{MSE}_{\text{norm}} = 0.24, 0.13, 0.18$ for metal lines, compared with $\text{MSE}_{\text{norm}} = 0.22, 0.13, 0.42$ for H I. Overall the RF models give reasonable predictions for the physical conditions, and again were not extensively tuned to achieve this.

An interesting feature of the original absorber data set is bimodal metallicity distributions at $\log Z/Z_{\odot} \sim -0.25$ and 0.25, which have not been reported in earlier SIMBA CGM work. The bimodality is apparent in every metal line apart from Mg II, and is broadly reproduced by the RF models. Populations of absorbers in the cool CGM of low redshift Lyman Limit Systems (LLSs) have also been observed to have bimodal metallicity distributions, with both metal-poor and metal-rich absorbers (albeit shifted to lower metallicities, Lehner et al. 2013, 2018, 2019; Wotta et al. 2016, 2019; Berg et al. 2023), suggesting multiple origins for the cool CGM gas, although the metallicities of the observed metal-poor absorbers are much lower than that seen in SIMBA. In future work we will investigate the origin of the bimodal absorber metallicity distribution in SIMBA.

6 FEATURE IMPORTANCE

In this section, we seek insights into the physical origin of the ML-probed correlations by assessing which input features are most useful in predicting the physical conditions.
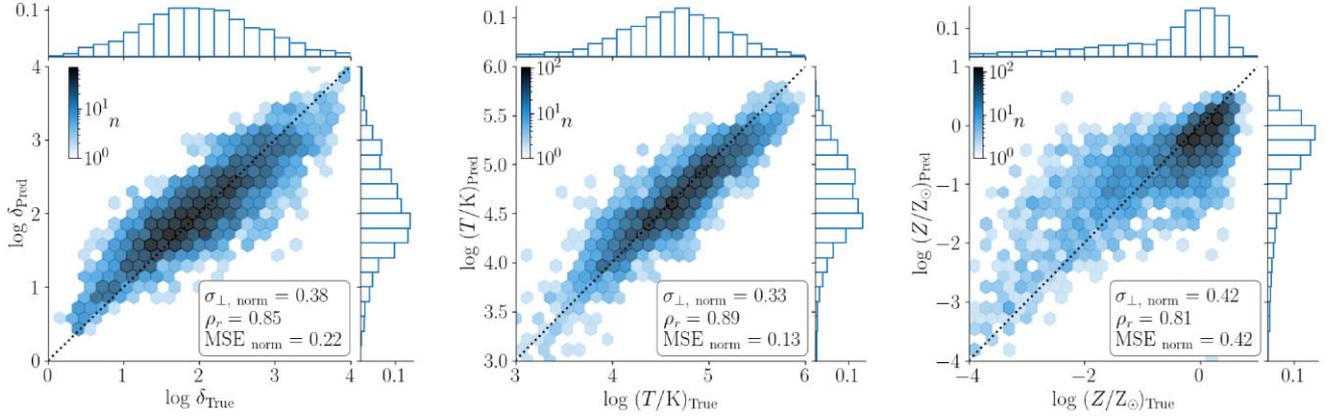
Figure 1. Hexagonal joint histogram of the predicted H I physical conditions from the RF mapping and the true H I physical conditions, including only data in the test set. The number of data points in each bin is shown using colourbars. From left to right, the panels show overdensity, temperature, and metallicity. The diagonal line represents the case where the RF model makes a perfect prediction. The accuracy of the predictions in each panel is summarized by the inset displaying the normalized transverse scatter $\sigma_{\perp,\mathrm{norm}}$, the correlation coefficient ρ_r, and the normalized mean square error, $\mathrm{MSE}_{\mathrm{norm}}$. The 1D histograms of the true and predicted values are shown on the top and side of each panel, respectively.
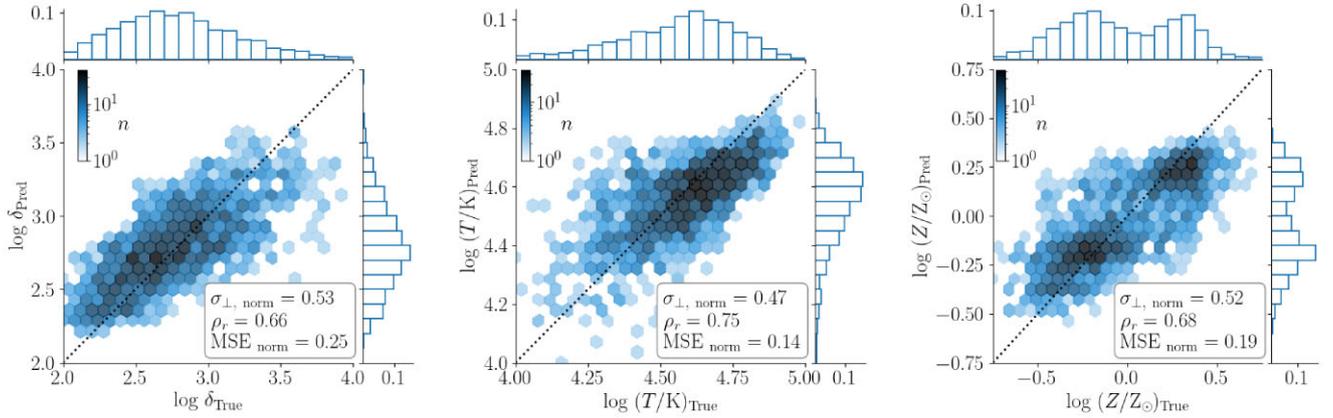
Figure 2. As in Fig. 1, showing the predictions and true values for C II absorbers.
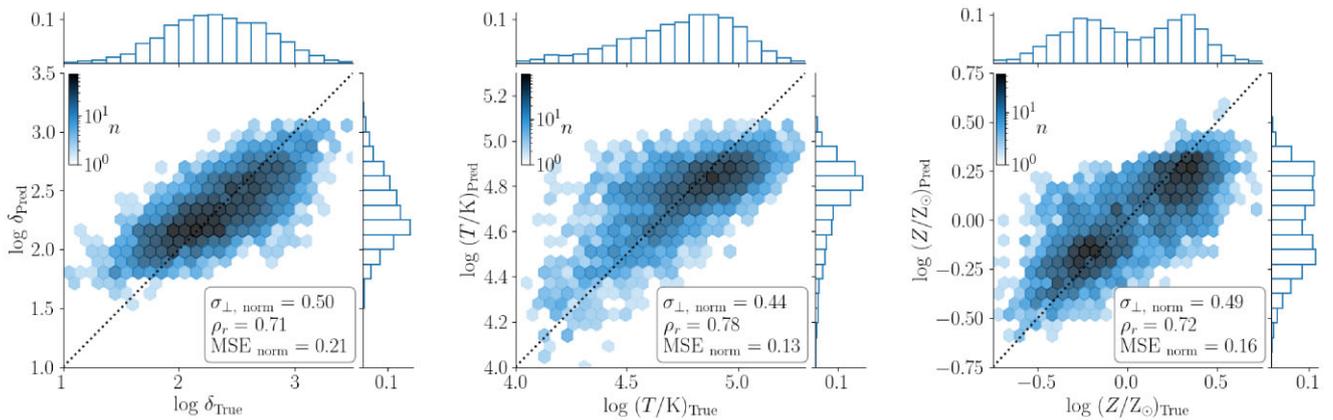
Figure 3. As in Fig. 1, showing the predictions and true values for C IV absorbers.

6.1 RF model importance

An advantage of the RF method over other ML algorithms (such as neural networks) is it allows some degree of interpretability in the form of the 'importance' of each feature, which arise 'for free' from the structure of the RF model. For an individual decision tree, a feature's importance is computed from the number of times it is used to split the data and how close to the top of the tree the splits are. For an RF model, the importances are the normalized average over all decision trees. However, importance metrics are biased if the input features are highly correlated with one another (Strobl et al. 2007,
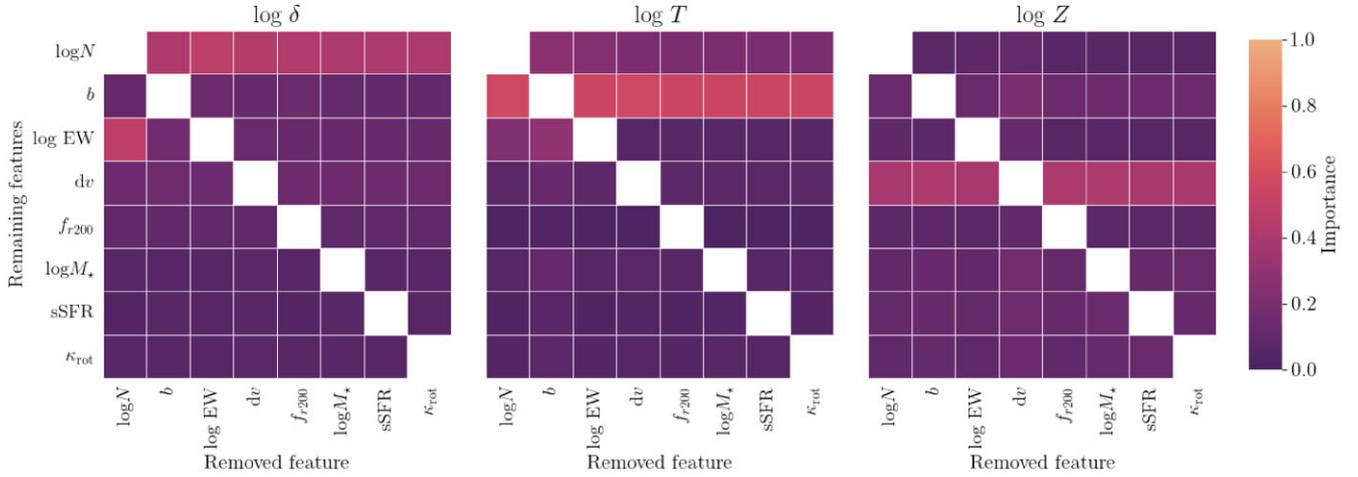
Figure 4. Importance values in predicting H I physical conditions for each remaining input feature, against the input feature removed from the training data. From left to right, the target predictors are δ, T, and Z.

2008) and so they should be treated with caution. Thus we prefer not to use the importance directly reported by RF, but instead compute it more empirically.

To do so, we determine each input feature's importance by iteratively building the RF model and removing each of the features in turn, using the same optimized hyperparameters as in the full feature model. We then retrieve the importance of the remaining features for each model. This process determines whether a feature is genuinely important, or merely defined as such through a fluke of feature combinations. When the most important features are removed, identifying which features take its place as the most important gives an indication of what the RF model is learning.

Fig. 4 shows the feature importance values for predicting H I absorber conditions, against the feature removed from the training data. For predicting overdensity, N is most important feature. When column density is removed, the most important feature is EW; N and EW are correlated with one another and both are correlated with physical density. When predicting temperature, b is the most important feature since the linewidths of individual absorbers in the original spectrum are set in partly by thermal Doppler broadening (with the additional effect of bulk gas motions). When predicting metallicity, the velocity separation is the most important feature. It is not intuitively obvious why this is the case; perhaps due to a dependence on halo velocity dispersion, which is correlated with $M_\star$ and thus the metallicity of the host galaxy that is predominantly responsible for enriching its CGM.

Figs 5 and 6 likewise show the feature importance values for predicting C II and C IV absorber properties. We have examined feature importance for all metals and found that these are representative cases. For the low ion C II, the feature importance rankings for δ and T are similar to that of H I. There is a slightly reduced relative importance of N in predicting δ in favour of b (N and b are correlated features due to their underlying dependence on δ and T). For both ions, the importance of b in predicting T is enhanced compared with H I. In contrast to H I, the most important feature for predicting Z for metal lines is sSFR; when sSFR is removed, the RF model learns from $M_\star$ and $\kappa_{\rm rot}$ instead, indicating that the RF model predicts Z from the galaxy properties. The picture is similar for the high ion C IV, except that in predicting δ the most important features are

instead f_{r200} and sSFR. f_{r200} is perhaps less useful for C II since most of the low ion absorption arises from the inner CGM (Appleby et al. 2023). Galaxies with high star formation have denser gas in their CGM, although this is also the case for H I and C II, so it is unclear why sSFR specifically is an important feature for C IV absorbers.

6.2 Change in predictive accuracy

Arguably, the most meaningful measure of 'importance' to the model is in which features add the most useful information in terms of predictive accuracy. We assess this by iteratively removing each of the features in turn from the training data, and running the RF model as before (with the hyperparameters optimized for the full feature case). In contrast with Section 6.1, we now compute the scatter $\sigma_{\perp,\,{\rm norm}}$ for each new model; if the quality of the predictions are significantly degraded in the absence of a particular feature, this necessarily indicates that this feature encodes crucial information about the physical conditions.

Fig. 7 shows the change in $\sigma_{\perp,\,{\rm norm}}$ resulting from the removal of each input feature in predicting the physical conditions, where a positive change indicates an increase in scatter. Each of the six panels shows the models for a different ion; the rows within each panel show the models for each of δ, T, and Z. The upper plots show H I (left), Mg II (middle), and C II (right) absorber models, while the lower plots show Si III (left), C IV (middle), and O VI (right).

In contrast to the feature importance values, the largest increase in scatter when predicting H I absorber overdensity comes from removing dv and f_{r200}. The loss of accuracy from removing f_{r200} suggests that the RF model is learning the radial density profile; this is also the case for the metal line δ predictions. It is less clear why dv is necessary for an accurate prediction, since halo absorbers can appear at any velocity separation depending on their kinematics. For all ions, removing b causes an increase in scatter in T predictions, confirming that the high feature importance of b reflects the genuine physical relationship with temperature.

For H I, the metallicity predictions are degraded by removing dv, f_{r200} or any galaxy property; interestingly the model accuracy improves when absorption-related features are removed. In other
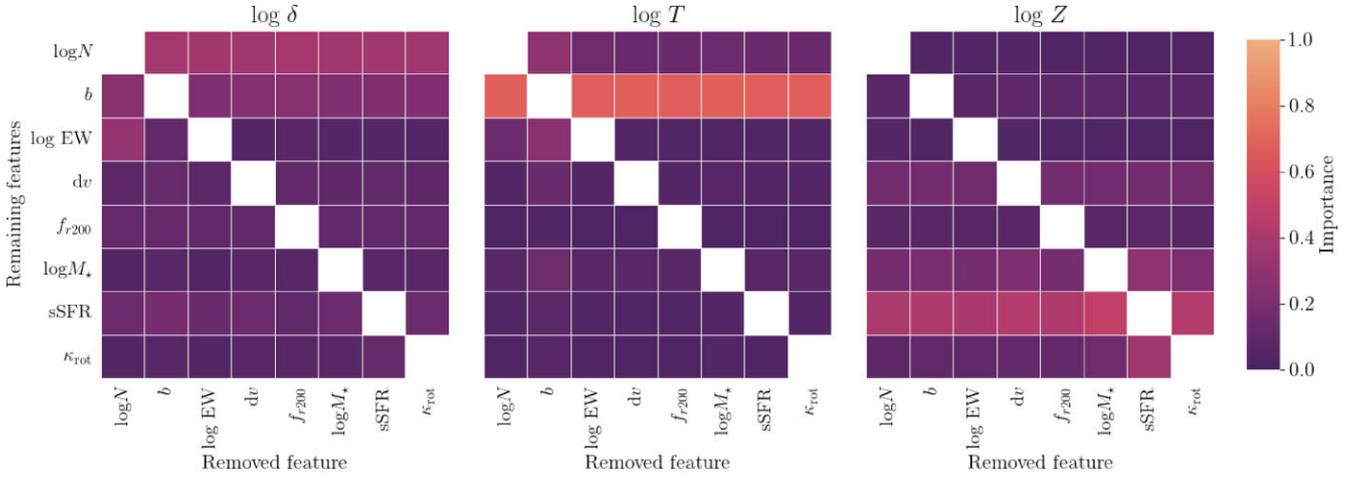
Figure 5. Feature importance values as in Fig. 4, for C II absorbers.
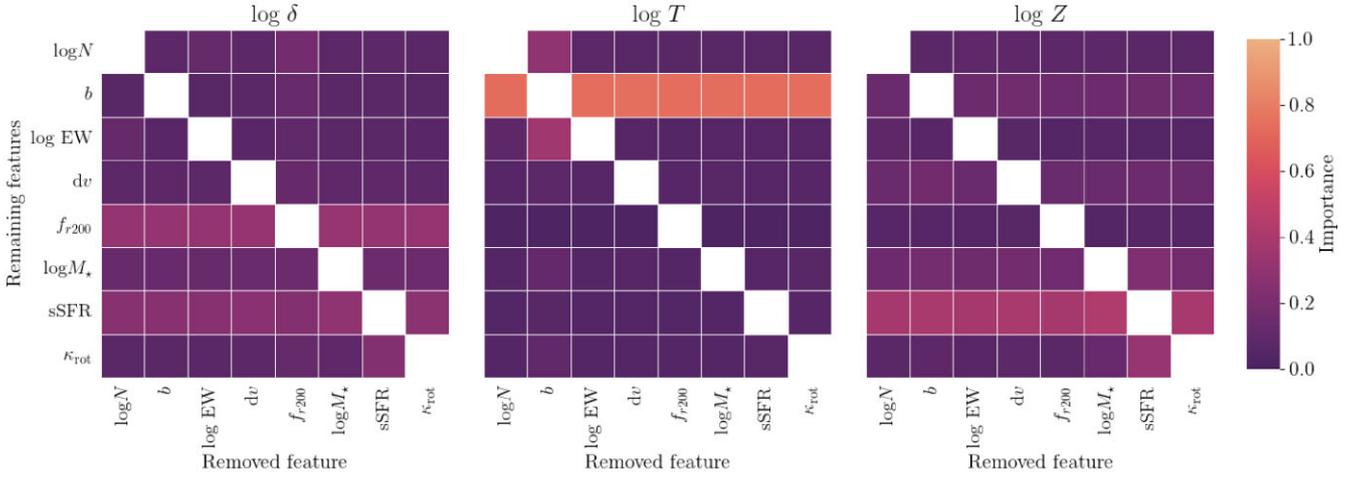
Figure 6. Feature importance values as in Fig. 4, for C IV absorbers.
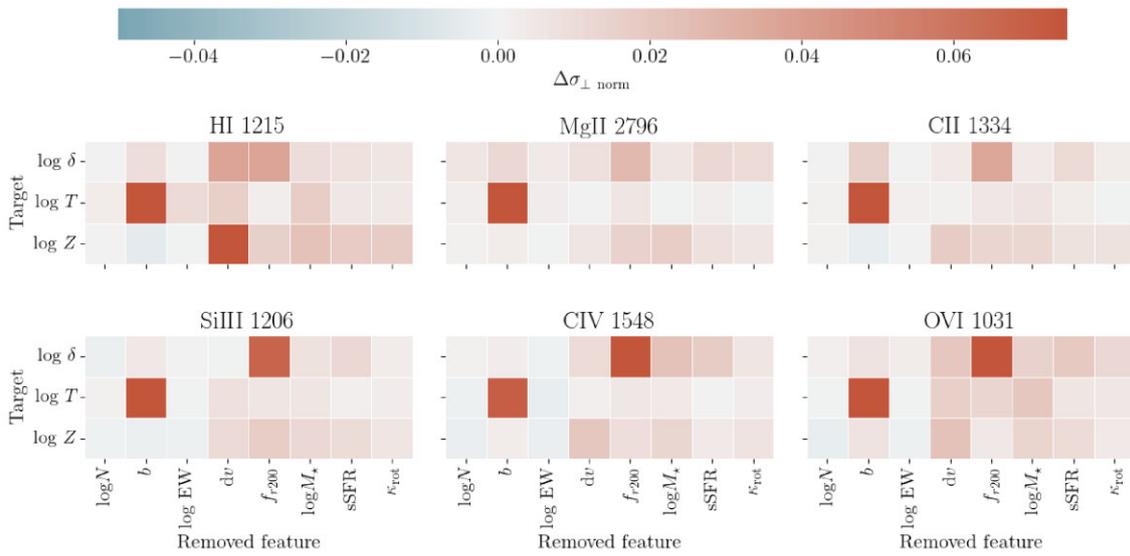
Figure 7. The change in $\sigma_{\perp,\,\mathrm{norm}}$ of the RF models when removing each feature iteratively. Each of the panels shows results for a different ion; the three rows of each panel represent results for each of δ, T, and Z.
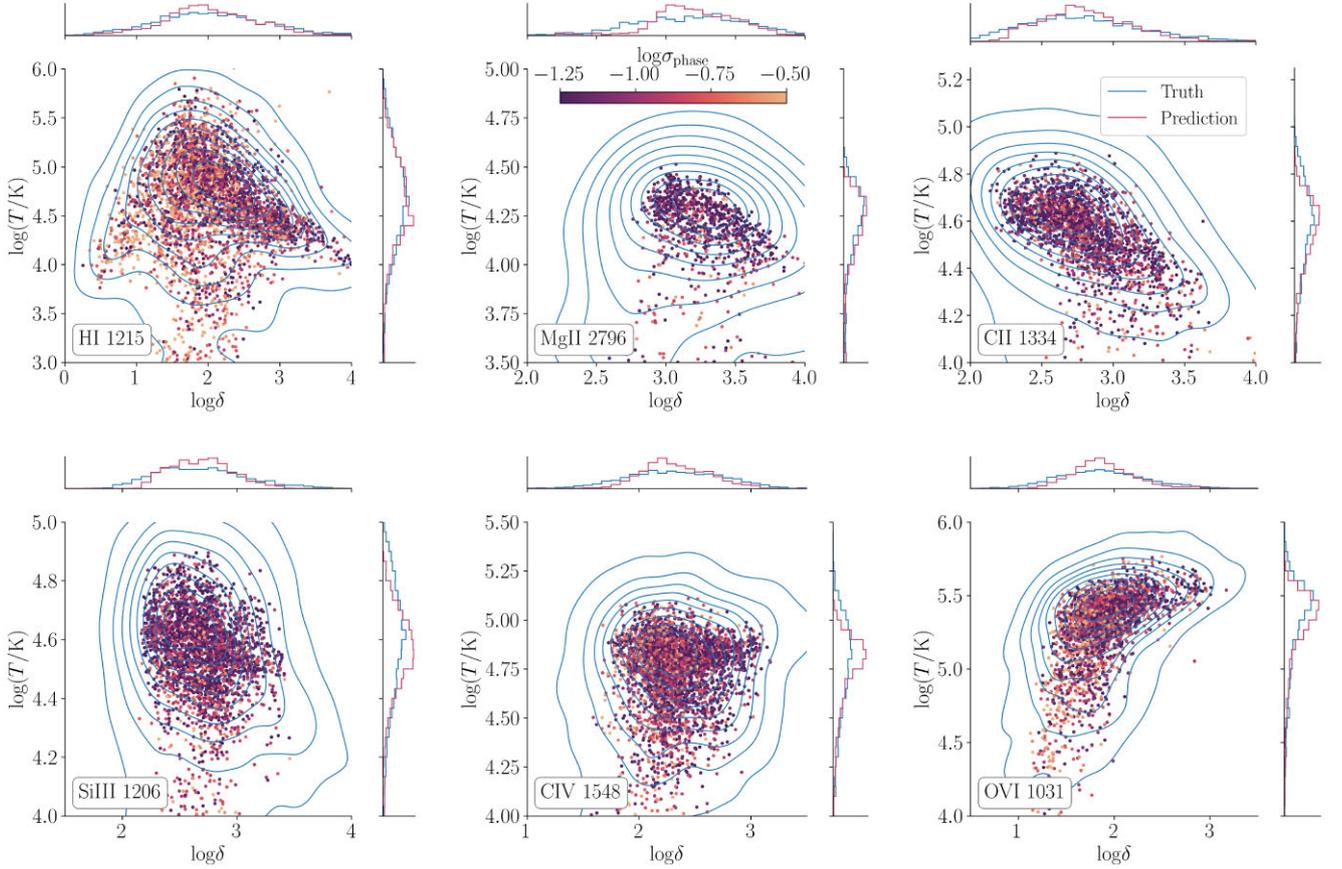
Figure 8. Predicted temperature against predicted overdensity for each of the 6 ions we consider, coloured by overall phase space fractional error. The 1D truth (blue curve) and predicted (pink curve) distributions are shown along the top and right of each panel. The contours show the true distribution in phase space for the test data set. The limits of the plots differ for each ion.

words, gas metallicity in the CGM of a given galaxy can be predicted with reasonable accuracy from only LOS and radial absorber position. The metal lines broadly show the same changes in scatter with removed features, although the changes to the model accuracy are more marginal.

In some cases, $\sigma_{\perp,\,\text{norm}}$ actually reduces with a given input feature removed. Ideally, this should never happen, as more information should always result in a better fit. Thus this indicates that perhaps there is some slight overfitting by the ML algorithm, or else there is some stochasticity in fitting process. The fact that these reductions are generally quite small even when present, much smaller than the typical increases, suggests that this is not a significant issue in the pipeline.

7 PHASE SPACE

Having examined the predictive accuracy and inner workings of the RF models on individual properties, we now ask whether the RF models can reproduce the 2D (δ, T) phase space structure of the absorbers. Although the RF models can separately predict δ and T, this does not guarantee that they reproduce the relationship between these quantities – particularly since the models are trained separately for δ and T, so each RF model has no knowledge of the other target quantities.

Fig. 8 shows the predicted temperature against predicted overdensity for the 6 species we consider. The distributions for the truth and

predicted data are shown along the top and right-hand side of each panel. Note that the plot limits are different for each ion. The points are colour coded by the fractional distance from the true value in δ $- T$ phase space:

$$\sigma_{\text{phase}} = \sqrt{\left(\frac{\delta_{\text{true}} - \delta_{\text{predict}}}{\delta_{\text{true}}}\right)^2 + \left(\frac{T_{\text{true}} - T_{\text{predict}}}{T_{\text{true}}}\right)^2}. \quad (5)$$

By this metric, a higher σ_{phase} (orange in the colour map) represents a larger distance from te truth value. The colour scale indicates that in general the predicted points lie near their truth values in phase space; the metal line absorbers in particular have a low displacement. The contours show the true distribution in phase space for the test dataset. The upper plots show H I (left), Mg II (middle), and C II (right) absorbers, while the lower plots show Si III (left), C IV (middle), and O VI (right). The original structure in phase space between overdensity and temperature is reproduced well by the RF models for each of the ions we consider, a success of the ML approach which (since there was no specific tuning of the model to achieve this) arises because temperature information is encoded in the overdensity data and vice versa. This is an important test of the RF models which verifies that accurate predictions can be made for multiple physical properties per observation.

That said, although the RF models succeed in predicting the phase space structure, the predictions are in general too concentrated near the mean of the data. By comparing the predicted distribution with the contours from the original data, it is clear that the predictions
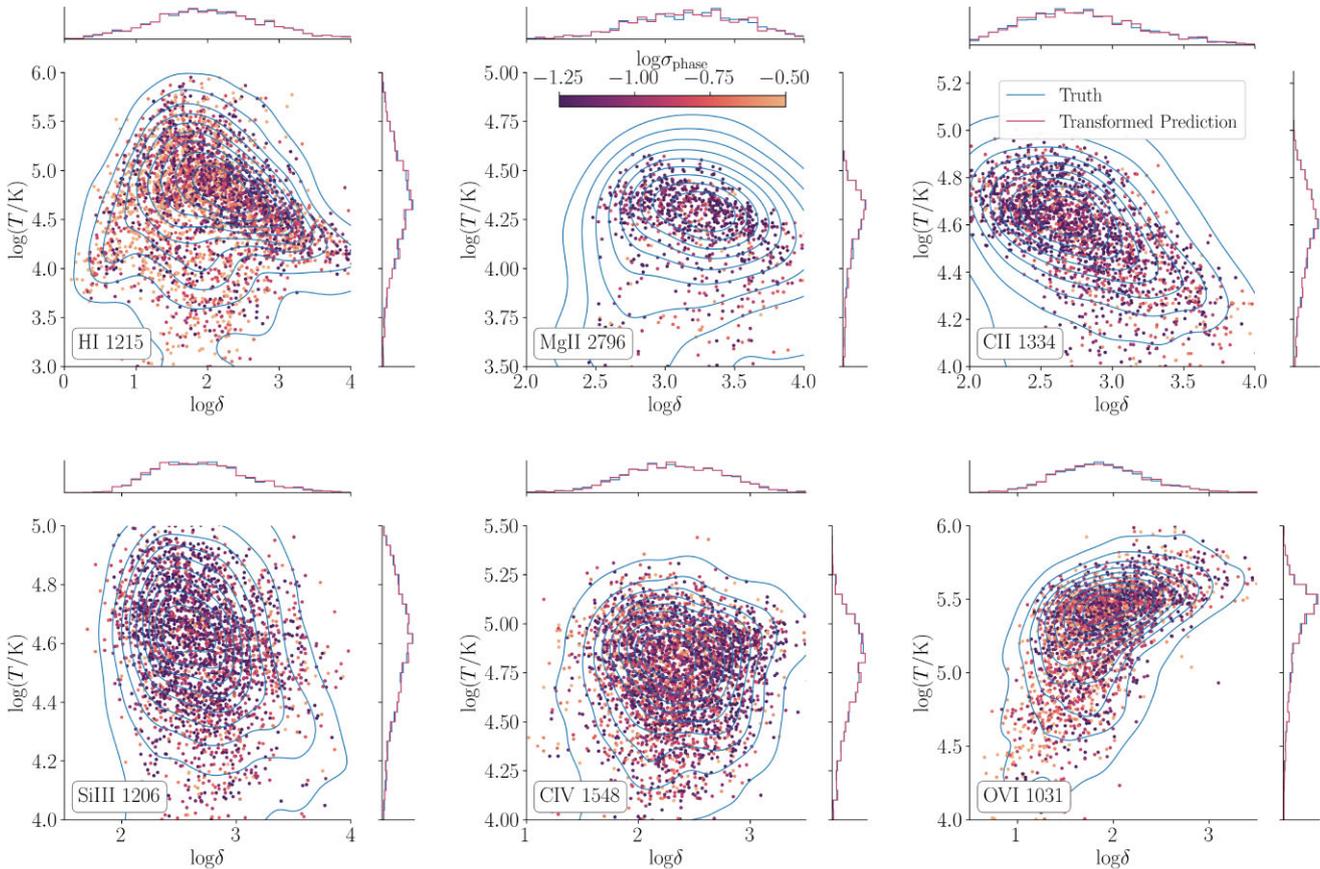
Figure 9. Predicted temperature against predicted overdensity for each of the 6 ions we consider, mapped to the shape of the truth 1D distributions using a quantile transformer. Points are coloured by overall phase space fractional error. The 1D distributions for the truth data (blue curve) and transformed predictions (pink curve) are shown along the top and right of each panel. The contours show the true distribution in phase space for the test data set.

ought to be more spread in phase space. This appears to be a generic feature of the RF models, which is also apparent in the 1D distributions of each feature – in general, the RF models produce predicted distributions that are too concentrated towards the mean. As mentioned in Section 5, this likely arises from sparse training data at the extremes. Thus, the predicted distributions do not capture the important information described by the intrinsic scatter in the original data, which biases the usefulness of these models for observational analysis. Therefore some additional step beyond the basic ML model is required in order to capture the full structure in phase space. Employing an oversampling technique (such as Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise) can boost undersampled regions of phase space and as such mitigate issues that arise as a result of imbalanced training data sets (de Santi et al. 2022).

In order to extend the RF network to also properly capture the full 2D phase space distribution, we develop a new approach based on a normalizing transform. By this, we mean that the predicted and truth data for each feature are mapped on to standard normal distributions, and then the predicted distribution is transformed back on to the shape of the truth data distribution.

To accomplish this we use the quantile transform non-parametric method implemented in `Scikit-Learn` (Pedregosa et al. 2011), `QuantileTransformer`. The method first maps the cumulative distribution of the data on to a standard Gaussian, and then computes the transformed values using a quantile function. The function also

provides the inverse mapping that transforms a distribution back into the original coordinates. The inverse mapping for the truth data distribution (computed from the training data set) is used to reconstruct the predicted distribution. In this way, we can reproduce the larger variance in the truth data without assuming a shape for the predicted data.

Fig. 9 shows the predicted temperature against predicted over-density, using the above normalizing transform approach to map the shape of the truth data. This can be seen in the 1D distributions along the top and right-hand side, which in most cases closely follow the truth data distributions. Crucially, the transformed data also retains the phase space structure of the original predictions. In addition, transforming the predictions on to the shape of the truth data results in no loss of accuracy for the predictions; the MSE_{norm}, ρ and $\sigma_{\perp,norm}$ for the transformed test data sets are very similar to those for the original test data set predictions. As such, this is our preferred method for reproducing the scatter in the truth CGM conditions data.

We initially explored a simpler approach where we added scatter directly to the predicted data. The results in phase space for the additional scatter approach are shown in Appendix B (Fig. B1). We found that this approach substantially washed out non-Gaussian structure in the predicted distribution, such as the anticorrelation between δ and T at $\delta > 10^3$. By instead using the normalizing transform method, we are preserving these structures in phase space as much as possible.
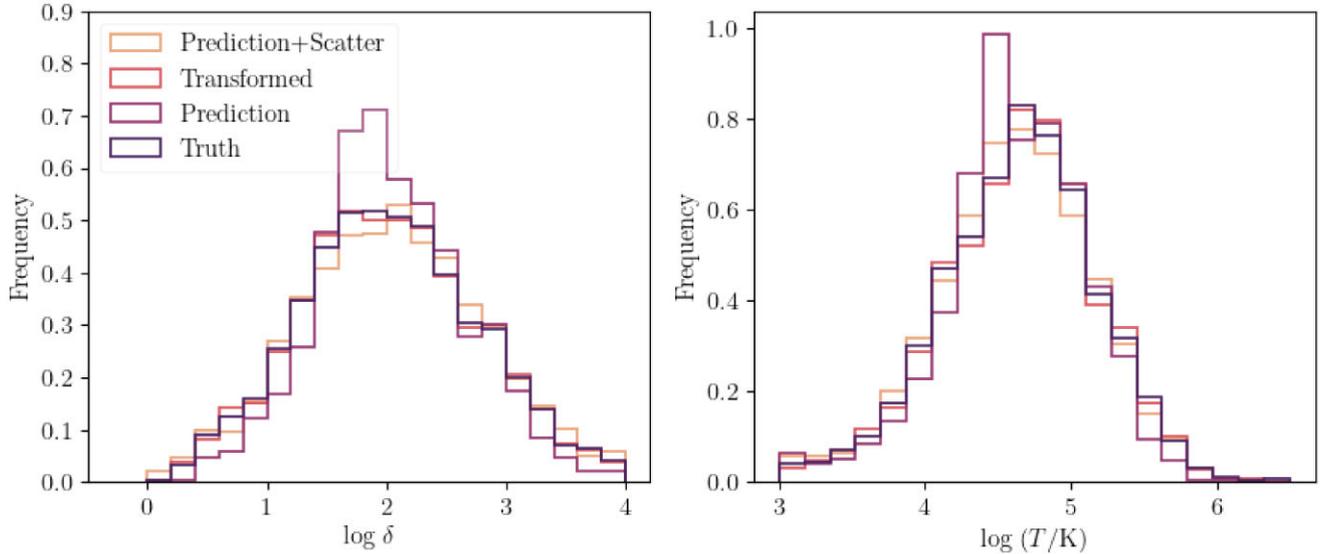
Figure 10. Histograms showing the direct comparison between truth data (dark purple), the predictions (light purple), the transformed predictions (orange), and the predictions with added scatter (yellow) for H I overdensities and temperatures.

Fig. 10 directly compares the distributions for the truth data (dark purple), the predictions from the RF models (light purple), the transformed predictions (orange), and the predictions with additional scatter (yellow) for the H I absorber overdensities and temperatures. The predicted physical conditions from the RF models are clearly too closely concentrated towards the mean. In contrast, applying the normalizing transform approach results in a predicted data set which closely matches the truth data. When additional scatter is instead added to the predictions, the resulting distribution also more closely matches the truth data, although not so precisely. Quantitatively, a two-sample Kolmogorov–Smirnov test with respect to the truth data gives p-values of >0.95 for the transformed overdensity and temperature distributions, but ~ 0.25 for the distributions with added scatter.

There are pros and cons to including the normalizing transform approach to ensure that the phase space scatter is well reproduced. If one wanted to compute distribution functions for physical quantities inferred from absorption line data, not including this post-processing step would result in the distribution functions improperly capturing the tails, which may be important for some applications. However, the additional step necessarily degrades the $\sigma_{\perp,\text{norm}}$ and MSE_{norm} of the predictions, albeit only marginally. The correlation coefficient does not change since we are only scaling the predictions. For example, in the case of H I overdensity, the $\sigma_{\perp,\text{norm}}$ and MSE_{norm} increase from $0.38 \rightarrow 0.40$ and $0.22 \rightarrow 0.25$, respectively, after applying the normalizing transform. For H I temperature, the $\sigma_{\perp,\text{norm}}$ and MSE_{norm} increase from $0.33 \rightarrow 0.34$ and $0.13 \rightarrow 0.14$, respectively. Whether or not to employ the above method thus depends on the application.

To recap, the normalizing transform approach requires setting up an inverse normalizing transform from the training data. To apply this in practice, one then runs the ML pipeline on the input features, produces a predicted distribution, normalizes this distribution using the quantile transformer, and then applies the inverse transform from the training data. This then gives predicted values that includes the additional scatter required to reproduce the spread in the input data. Indeed, the normalizing transform steps can be thought of as part of the ML pipeline itself in order to reproduce closest to the true distribution as possible.

8 CONCLUSIONS

We have produced machine-learnt mappings between CGM absorption observables and the underlying gas conditions for H I and selected metal lines using a Random Forest approach. RF models are preferred over other ML techniques for their relative simplicity and interpretability. These mappings represent a proof of concept for using ML models as part of an analysis pipeline for observational CGM data, which crucially does not make simplifying assumptions about the phase or composition of the absorbing gas. We identify a general tendency of the RF models to output a narrower predicted distribution than in the input data. We demonstrate two methods of reproducing the scatter of the input data: first by adding random Gaussian noise to the predictions, and second by transforming the predictions to the shape of the truth data. Our main results are as follows:

(i) The RF models predict reasonable H I overdensities ($\sigma_{\perp,\text{norm}} = 0.38$ dex, $\text{MSE}_{\text{norm}} = 0.22$) and temperatures ($\sigma_{\perp,\text{norm}} = 0.33$ dex, $\text{MSE}_{\text{norm}} = 0.13$). The predictions of overdensity and temperature are highly correlated with their truth values. Metallicity is less well predicted ($\sigma_{\perp,\text{norm}} = 0.42$ dex, $\text{MSE}_{\text{norm}} = 0.42$); metallicity is not directly traced by H I, therefore the learned relationship likely arises from the correlation with density.

(ii) The RF models also predict reasonable metal absorber conditions and perform to a similar accuracy among all metal lines, with median $\rho_r = 0.68, 0.71, 0.68$, median $\sigma_{\perp,\text{norm}} = 0.52, 0.44, 0.52$ and median $\text{MSE}_{\text{norm}} = 0.24, 0.13, 0.18$ for the overdensity, temperature, and metallicity predictions, respectively.

(iii) We report a bimodality in the absorber metallicity distributions for four of the five metal lines (C II, Si III, C IV, and O VI), suggesting multiple origins for the CGM gas in the Simba model.

(iv) In terms of feature importances, the RF models learn H I absorber overdensity from column density and equivalent width, temperature from the Doppler parameter, and metallicity from the LOS velocity separation. Low ion feature importances are similar to H I, except that metallicities are learned from sSFR and κ_{rot}. High

ion feature importances are similar to the low ions, except that the overdensities are learned from radial distance and galaxy properties.

(v) In terms of predictive accuracy, the radial distance and LOS separation provide the most useful information for predicting H I overdensity; the radial distance is also most useful for the metal line overdensities. The Doppler parameter is again the most important feature for predicting temperature for all lines. The LOS separation provides the most useful information for predicting H I metallicity; the predictions for all lines are degraded by removing galaxy properties.

(vi) The predictions for overdensity and temperature reproduce the phase space structure seen in the original data for all six ions, despite being trained for separately in the RF models. This verifies that accurate predictions can be made for multiple physical properties per observation.

(vii) By mapping the predicted data distributions on to the shape of the input distributions using a quantile normalizing transformer, we can reproduce the intrinsic scatter in the CGM phase space conditions with no loss of predictive accuracy or phase space structure.

Although we have considered H I and the metal ions separately, future work on this topic could explore RF models using combinations of absorption lines to assess whether predictions may be improved by using information from multiple ion species. A shortcoming of the ML models presented here is the that the predictions are too concentrated towards the mean of the distribution. Further development would be needed on the pipeline in order to reproduce the scatter in the original data without losing information in phase space.

The motivation of this project is to develop a useful analysis tool for the astronomical community to aid in interpreting absorption observations of the CGM, assuming the galaxy formation model of the SIMBA simulations and the Faucher-Giguère (2020) UVB. The next phase of this work is to test the method by applying the ML mappings to real observational data and comparing to results derived from ionization modelling. As such, the trained models produced for this work are available online and we encourage others to test the RF models on their own observational data.

A natural extension of this project will be to develop additional ML mappings using absorber data from other simulations such as EAGLE and IllustrisTNG to assess the impact of galaxy formation models on the predicted conditions for the CGM. Training the RF models on data from one simulation and testing on data from another would provide a robust test of the impact of galaxy formation model on our results. In addition, developing mappings using absorbers from the CAMELS project (Villaescusa-Navarro et al. 2021) would test the dependence of our results on both astrophysical and cosmological models.

DATA AVAILABILITY

The simulation data and galaxy catalogues underlying this article are publicly available.[3] The software used in this work is available on Github[4] and the derived data will be shared on request to the corresponding author.

REFERENCES

Acharya A., Khaire V., 2022, MNRAS, 509, 5559
Appleby S., Davé R., Kraljic K., Anglés-Alcázar D., Narayanan D., 2020, MNRAS, 494, 6053
Appleby S., Davé R., Sorini D., Storey-Fisher K., Smith B., 2021, MNRAS, 507, 2383
Appleby S., Davé R., Sorini D., Cui W., Christiansen J., 2023, MNRAS, 519, 5514
Berg M. A. et al., 2023, ApJ, 944, 101
Breiman L., 1996, Mach. Learn., 24, 123
Breiman L., 2001, Mach. Learn., 45, 5
Chen H.-W. et al., 2020, MNRAS, 497, 498
Christiansen J. F., Davé R., Sorini D., Anglés-Alcázar D., 2020, MNRAS, 499, 2617
Churchill C. W., Vogt S. S., Charlton J. C., 2003, AJ, 125, 98
Churchill C. W., Vander Vliet J. R., Trujillo-Gomez S., Kacprzak G. G., Klypin A., 2015, ApJ, 802, 10
Cohn J. D., Battaglia N., 2020, MNRAS, 491, 1575
Cooper T. J. et al., 2021, MNRAS, 508, 4359
Davé R., Hernquist L., Weinberg D. H., Katz N., 1997, ApJ, 477, 21
Davé R., Thompson R., Hopkins P. F., 2016, MNRAS, 462, 3265
Davé R., Anglés-Alcázar D., Narayanan D., Li Q., Rafieferantsoa M. H., Appleby S., 2019, MNRAS, 486, 2827
Davé R., Crain R. A., Stevens A. R. H., Narayanan D., Saintonge A., Catinella B., Cortese L., 2020, MNRAS, 497, 146
DeFelippis D., Bouché N. F., Genel S., Bryan G. L., Nelson D., Marinacci F., Hernquist L., 2021, ApJ, 923, 56
de Santi N. S. M., Rodrigues N. V. N., Montero-Dorta A. D., Abramo L. R., Tucci B., Artale M. C., 2022, MNRAS, 514, 2463
Delgado A. M., Wadekar D., Hadzhiyska B., Bose S., Hernquist L., Ho S., 2022, MNRAS, 515, 2733
Faucher-Giguère C.-A., 2020, MNRAS, 493, 1614
Ferland G. J. et al., 2017, RMxAA, 53, 385
Fluke C. J., Jacobs C., 2020, WIREs Data Min. Knowl. Discov., 10, e1349
Ford A. B., Oppenheimer B. D., Davé R., Katz N., Kollmeier J. A., Weinberg D. H., 2013, MNRAS, 432, 89
Ford A. B., Davé R., Oppenheimer B. D., Katz N., Kollmeier J. A., Thompson R., Weinberg D. H., 2014, MNRAS, 444, 1260
Ford A. B. et al., 2016, MNRAS, 459, 1745

[3] https://simba.roe.ac.uk
[4] https://github.com/sarahappleby/cgm_ml

Fumagalli M., O'Meara J. M., Prochaska J. X., 2016, MNRAS, 455, 4100

Gibson J. L., Lehner N., Oppenheimer B. D., Howk J. C., Cooksey K. L., Fox A. J., 2022, AJ, 164, 9

Glowacki M., Elson E., Davé R., 2020, MNRAS, 498, 3687

Gutcke T. A., Stinson G. S., Macciò A. V., Wang L., Dutton A. A., 2017, MNRAS, 464, 2796

Haardt F., Madau P., 2012, ApJ, 746, 125

Haislmaier K. J., Tripp T. M., Katz N., Prochaska J. X., Burchett J. N., O'Meara J. M., Werk J. K., 2021, MNRAS, 502, 4993

Hummels C. B., Bryan G. L., Smith B. D., Turk M. J., 2013, MNRAS, 430, 1548

Jo Y., Kim J.-h., 2019, MNRAS, 489, 3565

Keeney B. A. et al., 2017, ApJS, 230, 6

Kraljic K., Davé R., Pichon C., 2020, MNRAS, 493, 362

Lehner N. et al., 2013, ApJ, 770, 138

Lehner N., O'Meara J. M., Fox A. J., Howk J. C., Prochaska J. X., Burns V., Armstrong A. A., 2014, ApJ, 788, 119

Lehner N., Wotta C. B., Howk J. C., O'Meara J. M., Oppenheimer B. D., Cooksey K. L., 2018, ApJ, 866, 33

Lehner N., Wotta C. B., Howk J. C., O'Meara J. M., Oppenheimer B. D., Cooksey K. L., 2019, ApJ, 887, 5

Li Q., Narayanan D., Davé R., 2019, MNRAS, 490, 1425

Li F. et al., 2021, MNRAS, 500, 1038

Li Q., Han J., Wang W., Cui W., De Luca F., Yang X., Zhou Y., Shi R., 2022, MNRAS, 514, 5890

Liang C. J., Kravtsov A. V., Agertz O., 2018, MNRAS, 479, 1822

Lovell C. C., Geach J. E., Davé R., Narayanan D., Li Q., 2021, MNRAS, 502, 772

Lovell C. C., Wilkins S. M., Thomas P. A., Schaller M., Baugh C. M., Fabbian G., Bahé Y., 2022, MNRAS, 509, 5046

Lucie-Smith L., Peiris H. V., Pontzen A., Lochner M., 2018, MNRAS, 479, 3405

Marra R. et al., 2021, MNRAS, 508, 4938

Marra R., Churchill C. W., Kacprzak G. G., Nielsen N. M., Trujillo-Gomez S., Lewis J. G., 2022, preprint (arXiv:2202.12228)

McGibbon R. J., Khochfar S., 2022, MNRAS, 513, 5423

Moews B., Davé R., Mitra S., Hassan S., Cui W., 2021, MNRAS, 504, 4024

Mucesh S. et al., 2021, MNRAS, 502, 2770

Nadler E. O., Mao Y.-Y., Wechsler R. H., Garrison-Kimmel S., Wetzel A., 2018, ApJ, 859, 129

Nelson D. et al., 2020, MNRAS, 498, 2391

Nielsen N. M., Kacprzak G. G., Sameer, Murphy M. T., Nateghi H., Charlton J. C., Churchill C. W., 2022, MNRAS, 514, 6074

Oppenheimer B. D., Schaye J., 2013, MNRAS, 434, 1043

Oppenheimer B. D. et al., 2016, MNRAS, 460, 2157

Oppenheimer B. D., Schaye J., Crain R. A., Werk J. K., Richings A. J., 2018, MNRAS, 481, 835

Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825

Péroux C., Howk J. C., 2020, ARA&A, 58, 363

Prochaska J. X. et al., 2017, ApJ, 837, 169

Putman M. E., Peek J. E. G., Joung M. R., 2012, ARA&A, 50, 491

Qu Z. et al., 2022, MNRAS, 516, 4882

Rafieferantsoa M., Andrianomena S., Davé R., 2018, MNRAS, 479, 4509

Rahmati A., Pawlik A. H., Raičević M., Schaye J., 2013, MNRAS, 430, 2427

Robson D., Davé R., 2020, MNRAS, 498, 3061

Röttgers B., Naab T., Cernetic M., Davé R., Kauffmann G., Borthakur S., Foidl H., 2020, MNRAS, 496, 152

Sales L. V., Navarro J. F., Theuns T., Schaye J., White S. D. M., Frenk C. S., Crain R. A., Dalla Vecchia C., 2012, MNRAS, 423, 1544

Sameer et al., 2021, MNRAS, 501, 2112

Sankar S., Narayanan A., Savage B. D., Khaire V., Rosenwasser B. E., Charlton J., Wakker B. P., 2020, MNRAS, 498, 4864

Savage B. D., Kim T. S., Wakker B. P., Keeney B., Shull J. M., Stocke J. T., Green J. C., 2014, ApJS, 212, 8

Sorini D., Davé R., Anglés-Alcázar D., 2020, MNRAS, 499, 2760

Stocke J. T., Keeney B. A., Danforth C. W., Shull J. M., Froning C. S., Green J. C., Penton S. V., Savage B. D., 2013, ApJ, 763, 148

Strobl C., Boulesteix A.-L., Zeileis A., Hothorn T., 2007, BMC Bioinf., 8, 25

Strobl C., Boulesteix A.-L., Kneib T., Augustin T., Zeileis A., 2008, BMC Bioinf., 9, 307

Suresh J., Nelson D., Genel S., Rubin K. H. R., Hernquist L., 2019, MNRAS, 483, 4040

Thomas N., Davé R., Anglés-Alcázar D., Jarvis M., 2019, MNRAS, 487, 5764

Thomas N., Davé R., Jarvis M. J., Anglés-Alcázar D., 2021, MNRAS, 503, 3492

Tripp T. M., Sembach K. R., Bowen D. V., Savage B. D., Jenkins E. B., Lehner N., Richter P., 2008, ApJS, 177, 39

Tumlinson J., Peeples M. S., Werk J. K., 2017, ARA&A, 55, 389

Ucci G., Ferrara A., Gallerani S., Pallottini A., 2017, MNRAS, 465, 1144

Villaescusa-Navarro F. et al., 2021, ApJ, 915, 71

Werk J. K. et al., 2014, ApJ, 792, 8

Wotta C. B., Lehner N., Howk J. C., O'Meara J. M., Prochaska J. X., 2016, ApJ, 831, 95

Wotta C. B., Lehner N., Howk J. C., O'Meara J. M., Oppenheimer B. D., Cooksey K. L., 2019, ApJ, 872, 81

Zahedy F. S., Chen H.-W., Johnson S. D., Pierce R. M., Rauch M., Huang Y.-H., Weiner B. J., Gauthier J.-R., 2019, MNRAS, 484, 2257

Zahedy F. S. et al., 2021, MNRAS, 506, 877

van de Voort F., Springel V., Mandelker N., van den Bosch F. C., Pakmor R., 2019, MNRAS, 482, L85

APPENDIX A: RESULTS FOR OTHER METALS

For completeness, here we present Fig. A1, A2, A3 (similar to Fig. 1, for Mg II, Si III, and O VI). The general trends are already captured by the plots in the main text for C II and C IV. However, there are some interesting notable point. For instance, O VI shows significantly higher temperatures, as expected since it is a higher ionization line. The metallicity bimodality is still slightly present for O VI, although at a much lower significance than for the lower ions. Si III has the largest scatter in the recovered T, and also shows some bias such that high-T absorbers are underpredicted while low-T ones are overpredicted. This may be because Si III seems to have absorbers spanning the widest range in temperatures from among the metal ions considered. In terms of the RF performance, however, these ions tell a similar story, which is encouraging since it means the RF methodology is widely applicable with similar efficacy across a range of commonly observed low-z UV ions.
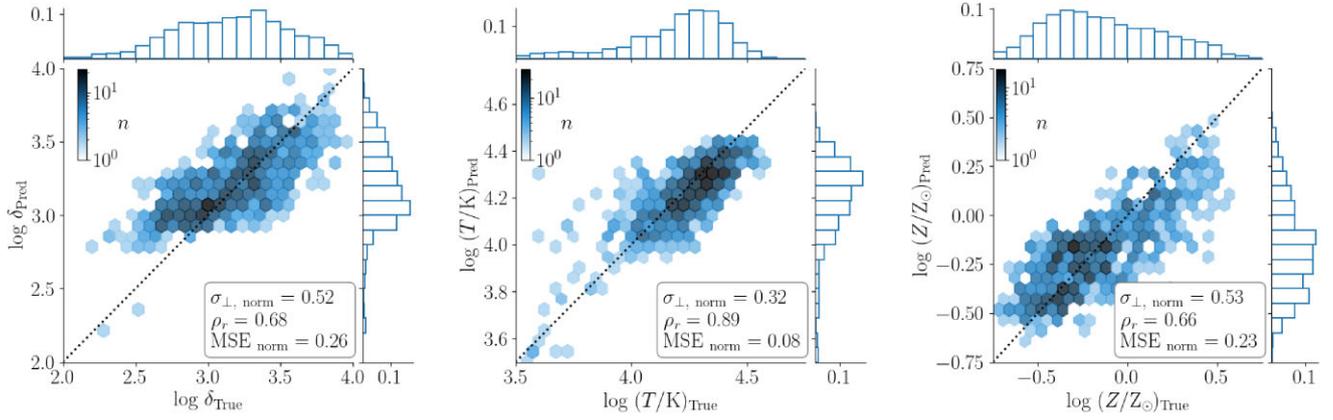
Figure A1. As in Fig. 1, showing the predictions and true values for Mg II absorbers.
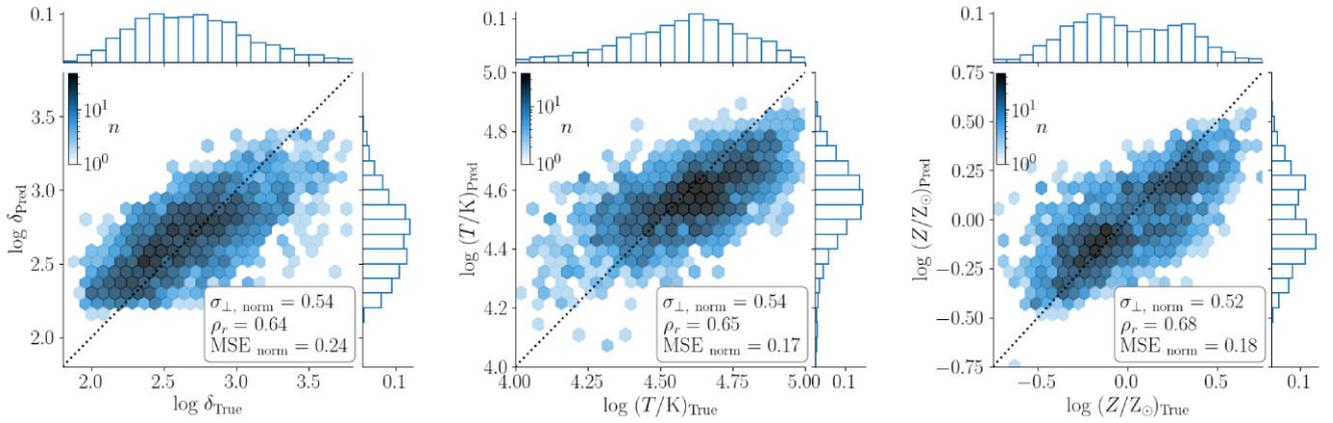
Figure A2. As in Fig. 1, showing the predictions and true values for Si III absorbers.
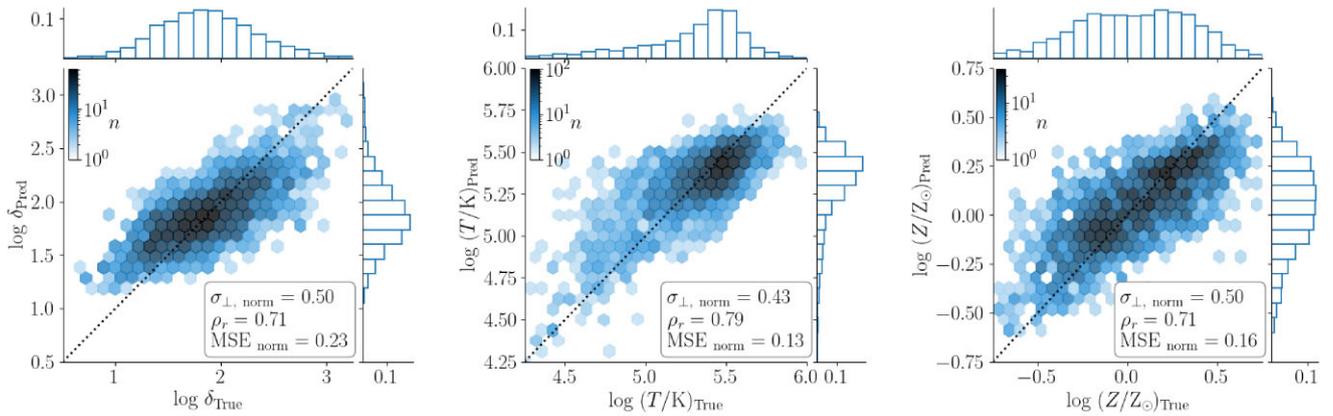
Figure A3. As in Fig. 1, showing the predictions and true values for O VI absorbers.

APPENDIX B: DIRECT GAUSSIAN APPROACH TO ADDING PHASE SPACE SCATTER

In Section 7, we described our normalizing transform-based approach to adding scatter to the RF predictions in order to match the 2D truth distributions in phase space. A more straightforward approach is to directly add Gaussian scatter to the predicted δ and T distributions to

match the truth without first applying a normalizing transformation. However, the results were less satisfactory.

Fig. B1 shows the results, which can be compared to Fig. 9. It is clear that the simpler approach causes features within the true phase space to be more washed out, and substantially degrades the predictive accuracy. Thus we prefer the normalizing transform approach presented in Section 7.
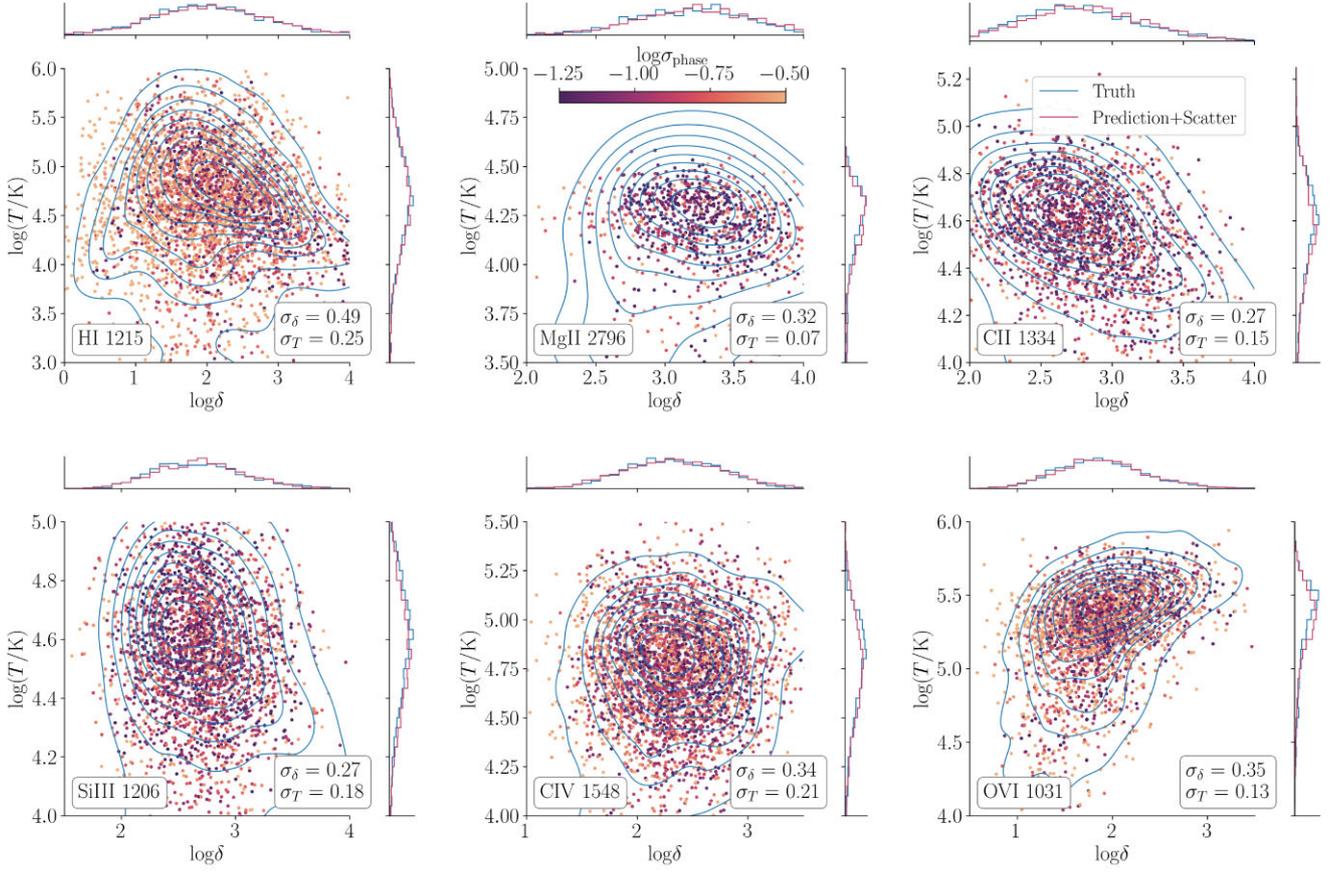
Figure B1. Predicted temperature against predicted overdensity for each of the 6 ions we consider, with added random Gaussian noise to reproduce the original 1D distributions. Points are coloured by overall phase space fractional error. The 1D truth and predicted distributions are shown along the top and right of each panel. The contours show the true distribution in phase space for the test data set. The widths of the random noise Gaussians for δ and T are shown in the bottom right of each panel. Compared to the normalizing transform results shown in Fig. 9, this approach washes out features in phase space substantially more.

This paper has been typeset from a TEX/LATEX file prepared by the author.