Chlorophyll Level Estimation
Based On Remote Sensing Reflectance

The goal is to develop a prediction model that can estimate the chlorophyll (**CHL**) level based on satellite reflectance data. The remote sensing reflectance (**Rrs**) spectrum at different wavelengths is used as input data. The **Rrs** spectra are recorded at multiple wavelengths as listed in Table 1.

λ (nm)	400	412.5	442.5	490	510	560	620	665	673.75	681.25	708.75	753.75	761.25	764.375	767.5	778.75
Label	WL_400	WL_412	WL_442	WL_490	WL_510	WL_560	WL_620	WL_665	WL_673	WL_681	WL_708	WL_753	WL_761	WL_764	WL_767	WL_778

Table 1: The wavelengths used to record the Rrs spectra.

The training dataset consists of **706339** samples, while the validation dataset has **100** samples. The chlorophyll distributions in the training and validation datasets are presented in Figure 1.

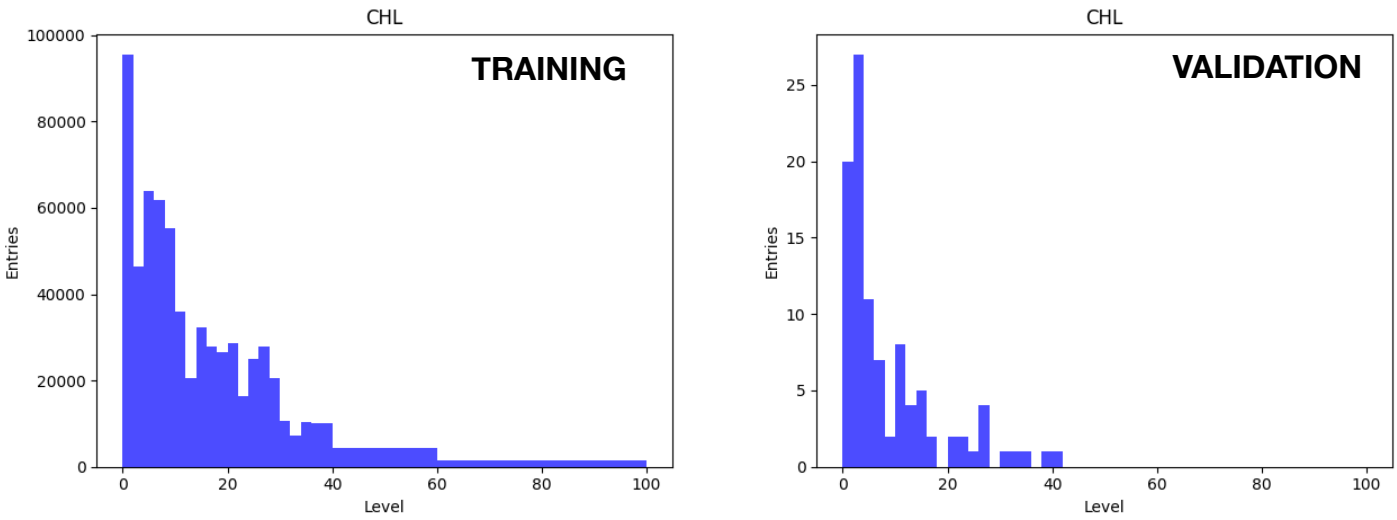

Figure 1: The chlorophyll level in the training and validation dataset.

Both training and validation distributions exhibit a falling spectrum with most of the statistics in the low value range. To reduce the unbalance of the data, further processing is required.

The **Rrs** spectra at three wavelengths (400nm, 560nm and 778.75nm) and their dependence on the chlorophyll levels in the training dataset are presented in Figure 2. All **Rrs** spectra are scaled up to the (0, 1) range to aid the training process. A large number of data samples have both **Rrs** and **CHL** values close to zero. The **Rrs** distribution for 560nm shows the best separation of the data with the peak close to zero clearly visible. Based on the chlorophyll absorption spectrum from Figure 3, my assumption is that most of the direct contribution from chlorophyll to the **Rrs** spectrum is visible for wavelengths in the green band, like the third peak in the **Rrs** distribution at 560nm.

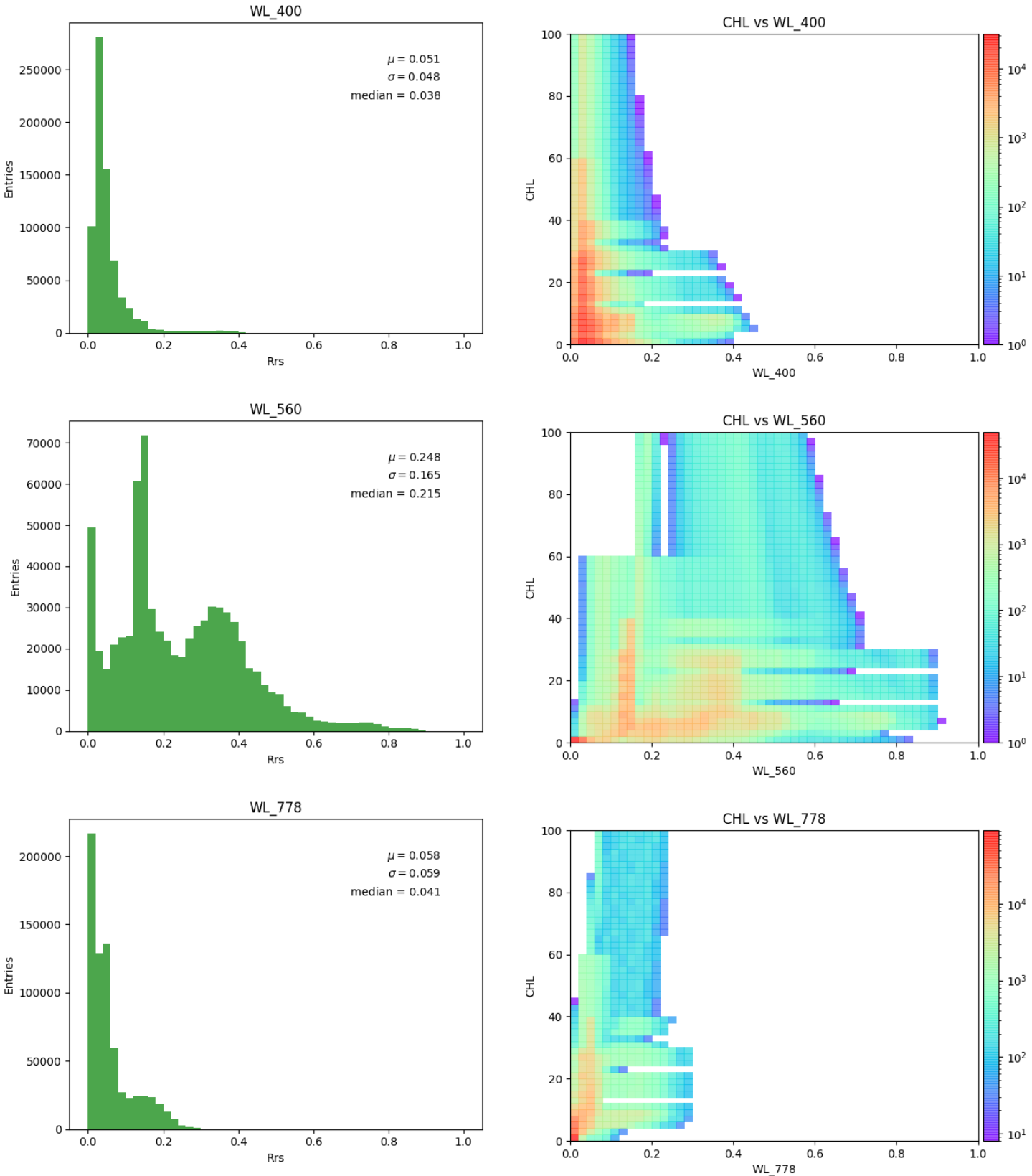

Figure 2: The Rrs spectra and their dependence on the CHL level at different λ.

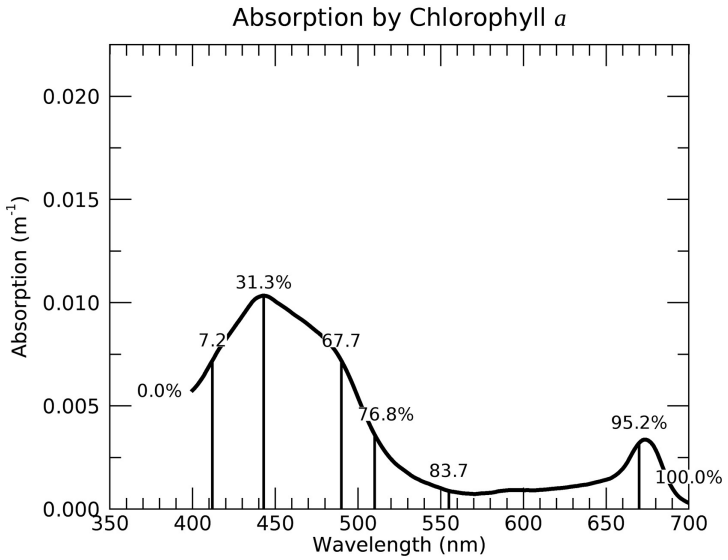

Figure 3: The chlorophyll absorption spectrum.

Based on this information, both training and validation data samples are filtered using the **Rrs** distribution at 560nm. Only data samples that have the **Rrs(560) > 0.018** are used in the training and validation process. Since the data is filtered using this variable, the **Rrs(560)** is completely excluded from the list of input features.

Figure 4 shows the **CHL** distributions in the training and validation data sets after the data is filtered to reduce the unbalance. The distributions exhibit a similar trend and coverage.

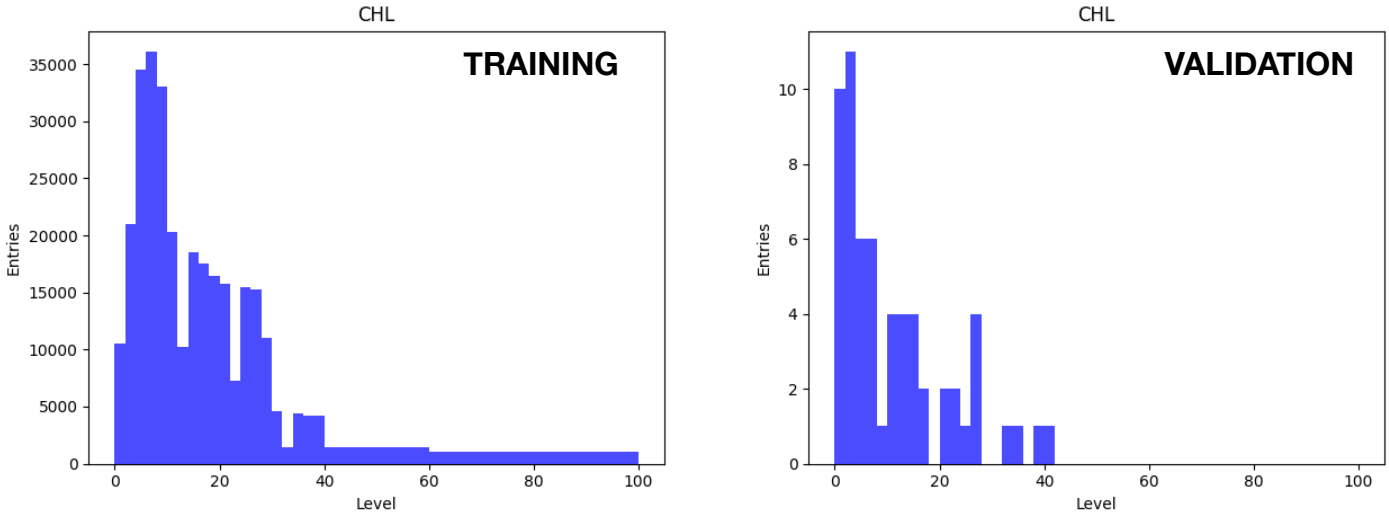

Figure 4: The chlorophyll level in the training and validation dataset after filtering.

For comparison, the **Rrs** spectra at three wavelengths (400nm, 560nm and 778.75nm) and their dependence on the chlorophyll levels in the training dataset, after filtering the data, are presented in Figure 5.

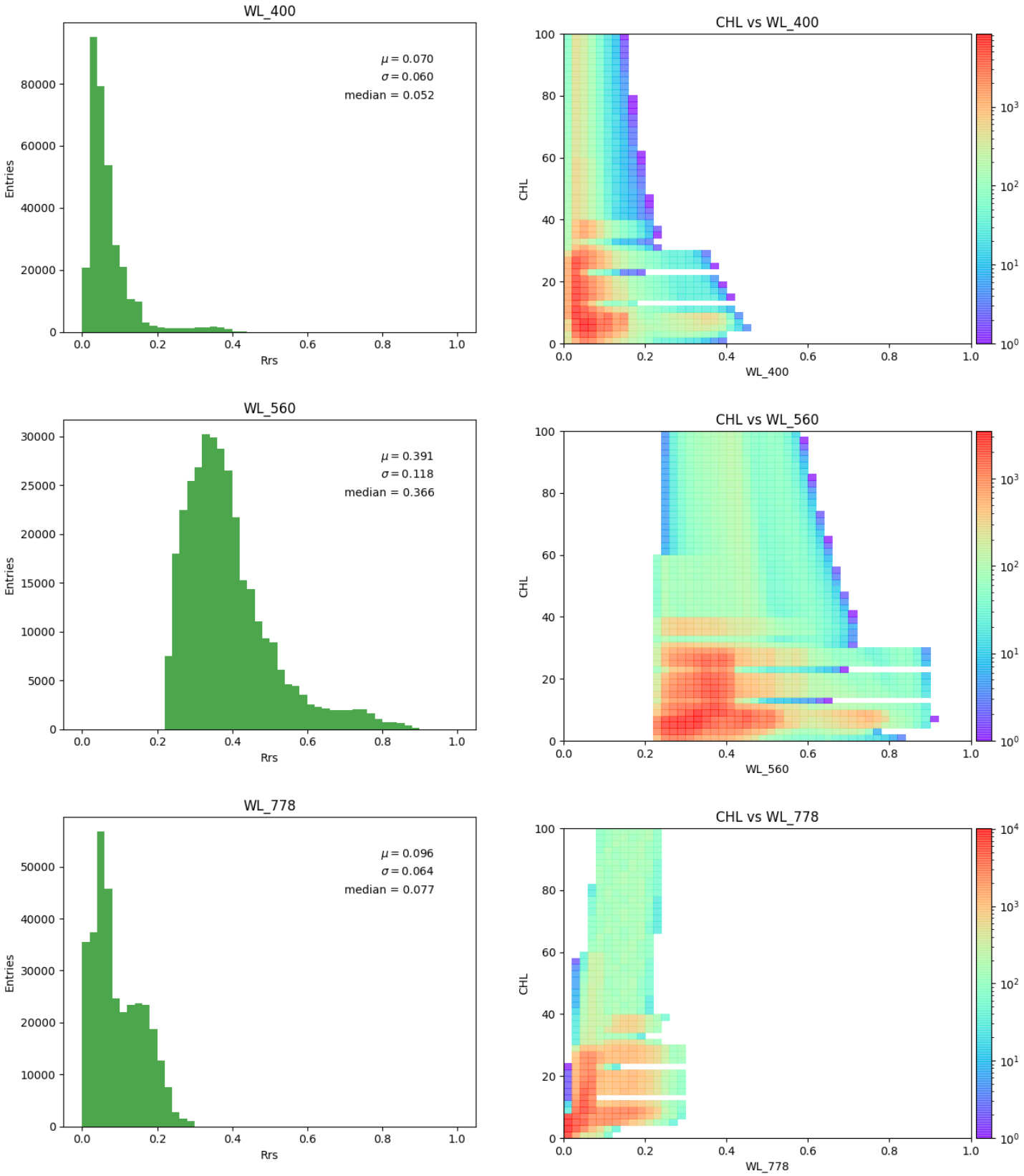

Figure 5: The Rrs spectra and their dependence on the CHL level at different λ after filtering.

The model under consideration is a fully connected neural network developed in TensorFlow. The basic design of the model consists of:

- **one input layer**
- **two hidden layers**
- **one output layer**

For this model, the following hyper-parameters are considered:

- **number of nodes in the hidden layers**
- **activation function**
- **batch size**
- **learning rate**
- **regularization rate**
- **dropout rate**

The model is trained using the:

- **ADAM optimizer**
- **mean squared error (MSE) as the loss function**

The optimum combination of hyper-parameters is obtained using a grid-based search approach by scanning the hyper-parameter space listed in Table 2 over 20 epochs.

	V1	V2	V3	Best
Number of Nodes (N)	16	32	64	**16**
Activation Function (AF)	relu	linear	sigmoid	**linear**
Batch Size (BS)	20	40	80	**20**
Learning Rate (LR)	0.01	0.001	0.0001	**0.0001**
Regularization Rate (RR)	0.01	0.001	0.0001	**0.001**
Drpout Rate (DR)	0.1	0.2	0.4	**0.1**

Table 2: The hyper-parameter search space.

The results of the hyper-parameter search are presented in Figures 7-12. The initial search starts with **32** nodes with **relu** activation functions and a batch size of **20**. The hyper-parameter in red is the one under scrutiny while the ones in blue are already fixed.

Given the selected architecture, the final model is obtained by using the "best" hyper-parameters from Table 2, and train the model for about 40 epochs, as presented in Figure 6. This is the model used on the test dataset.

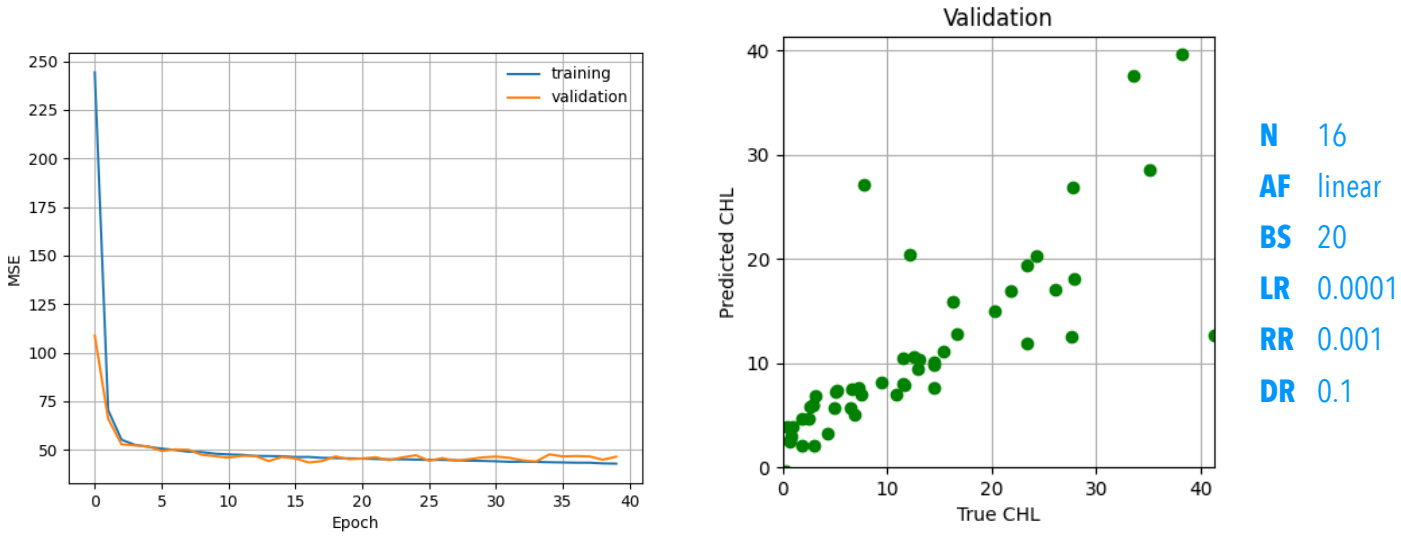

Figure 6: The efficiency of the final model.

The next step would be to try either different neural network architectures or other algorithms like random forest or SVM.

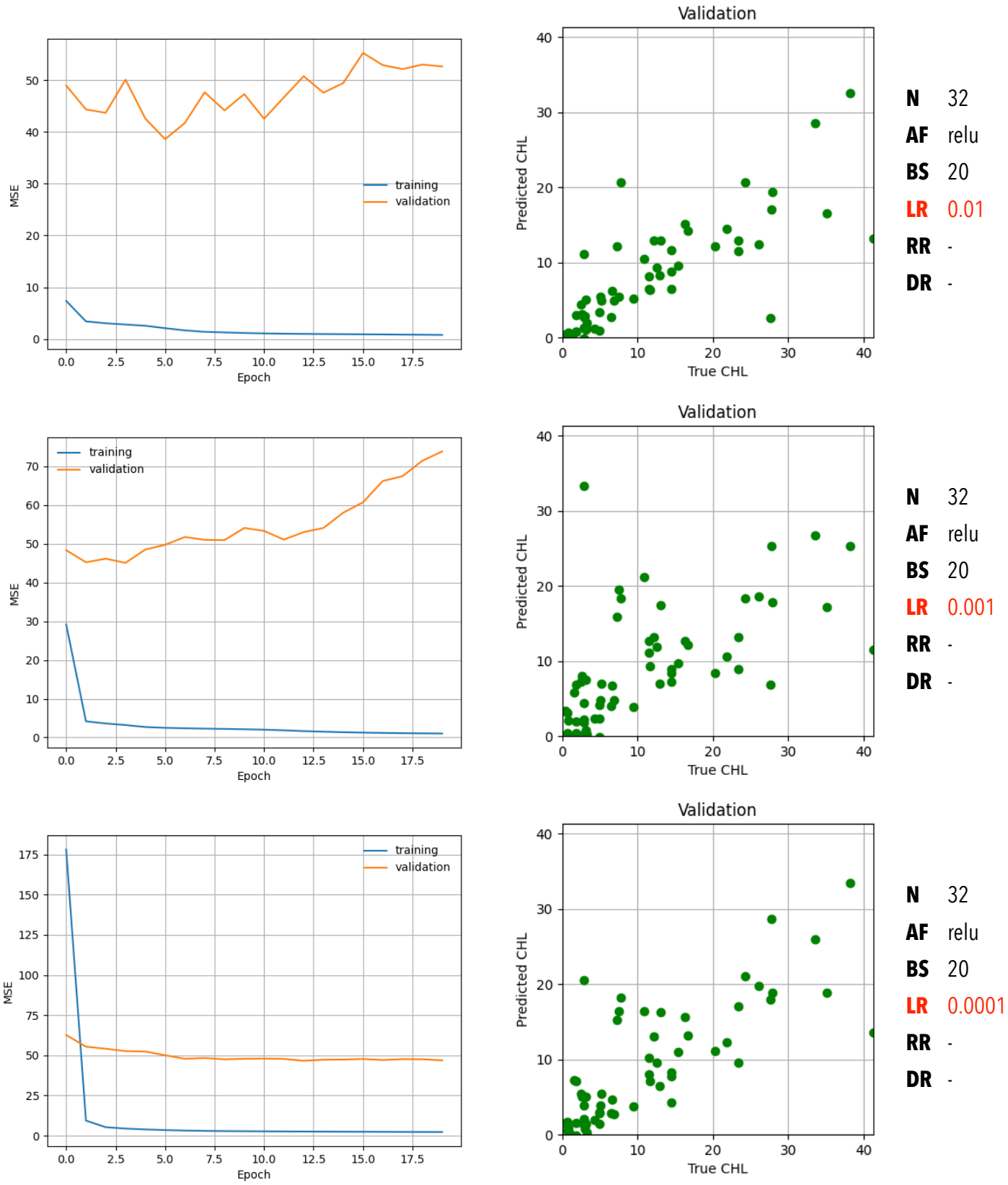

Figure 7: The variation of the learning rate (select LR = 0.0001).

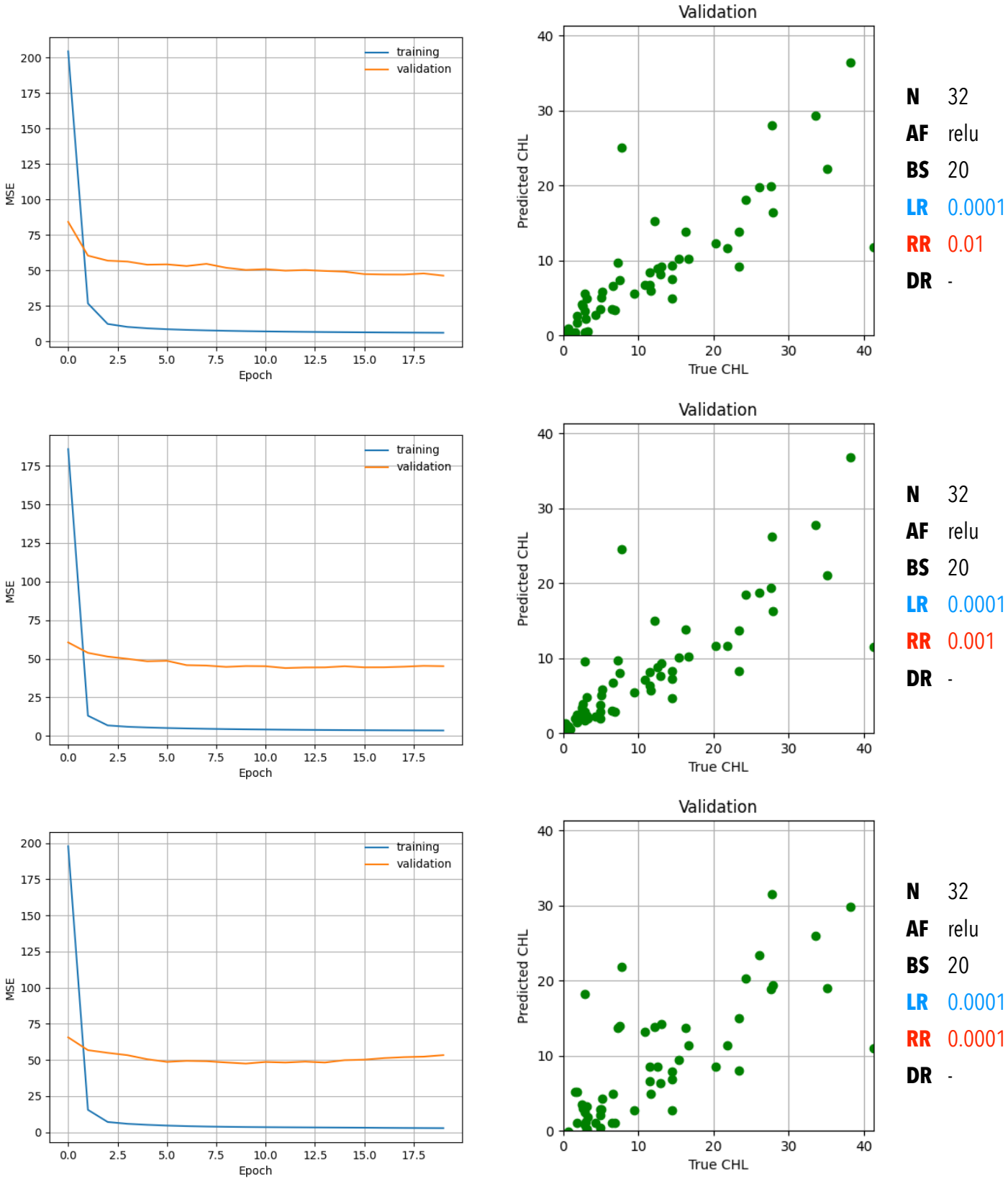

Figure 8: The variation of the regularization rate (select RR = 0.001).

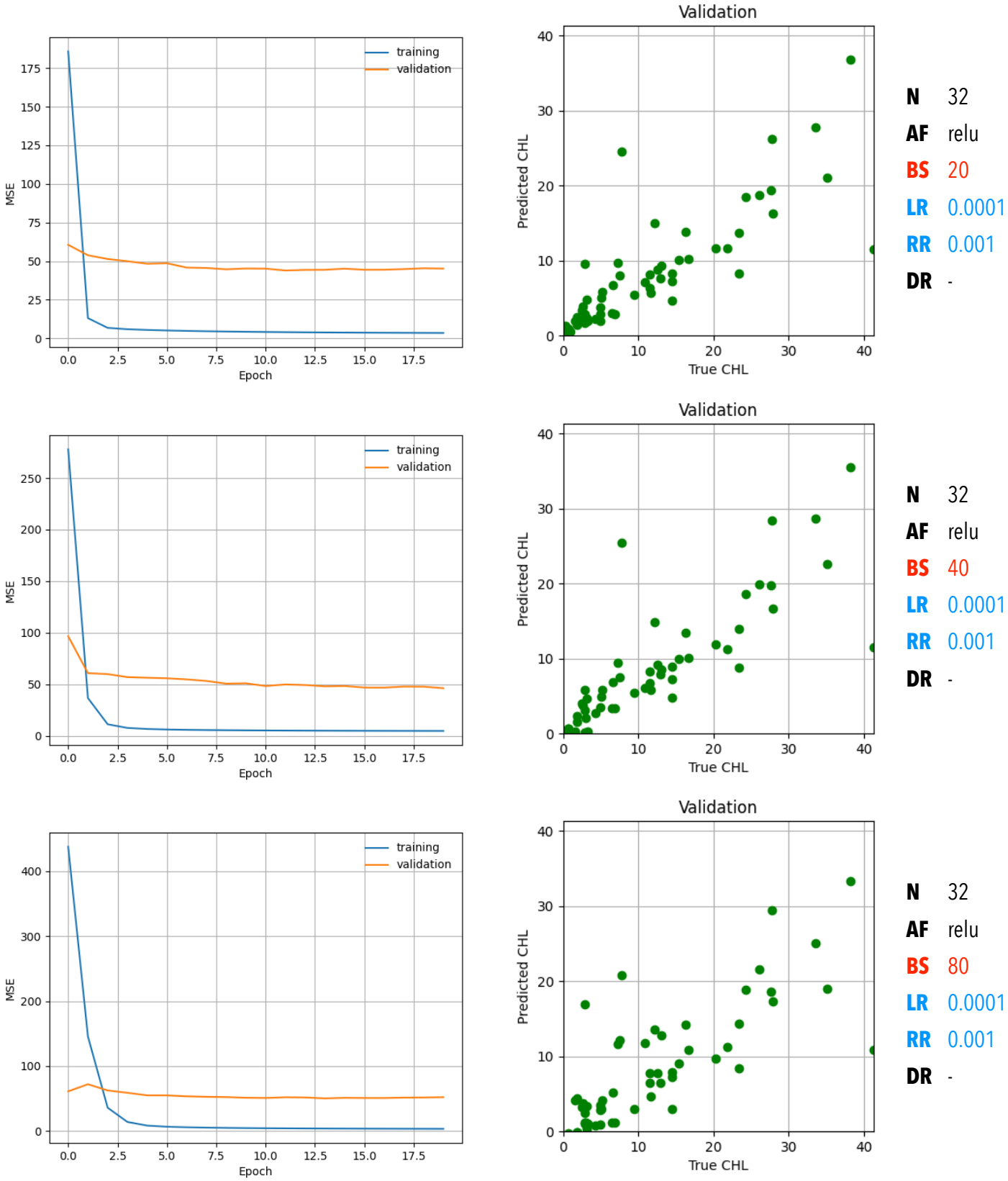

Figure 9: The variation of the batch size (select BS = 20).

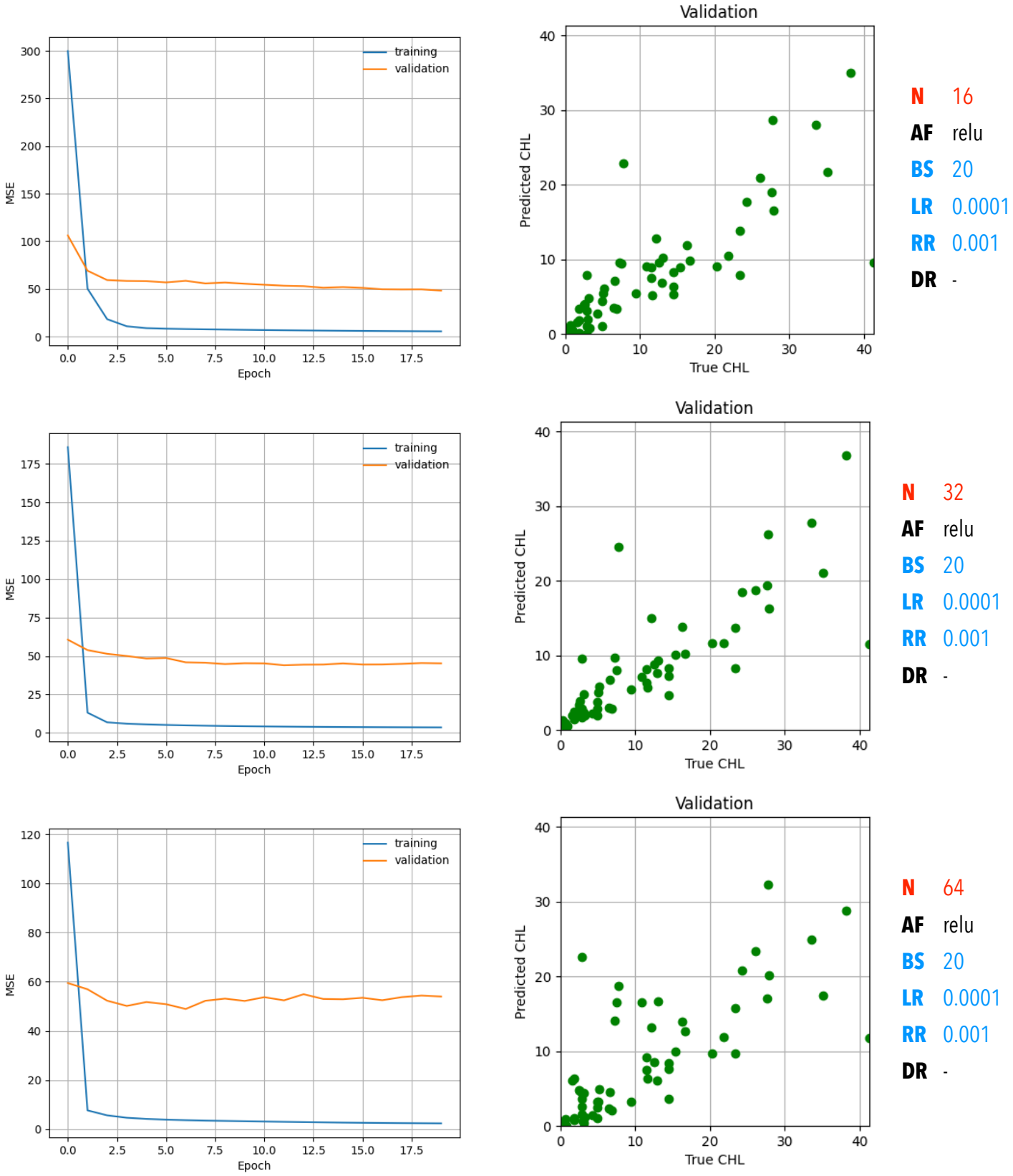

Figure 10: The variation of the number of nodes (select N = 16).

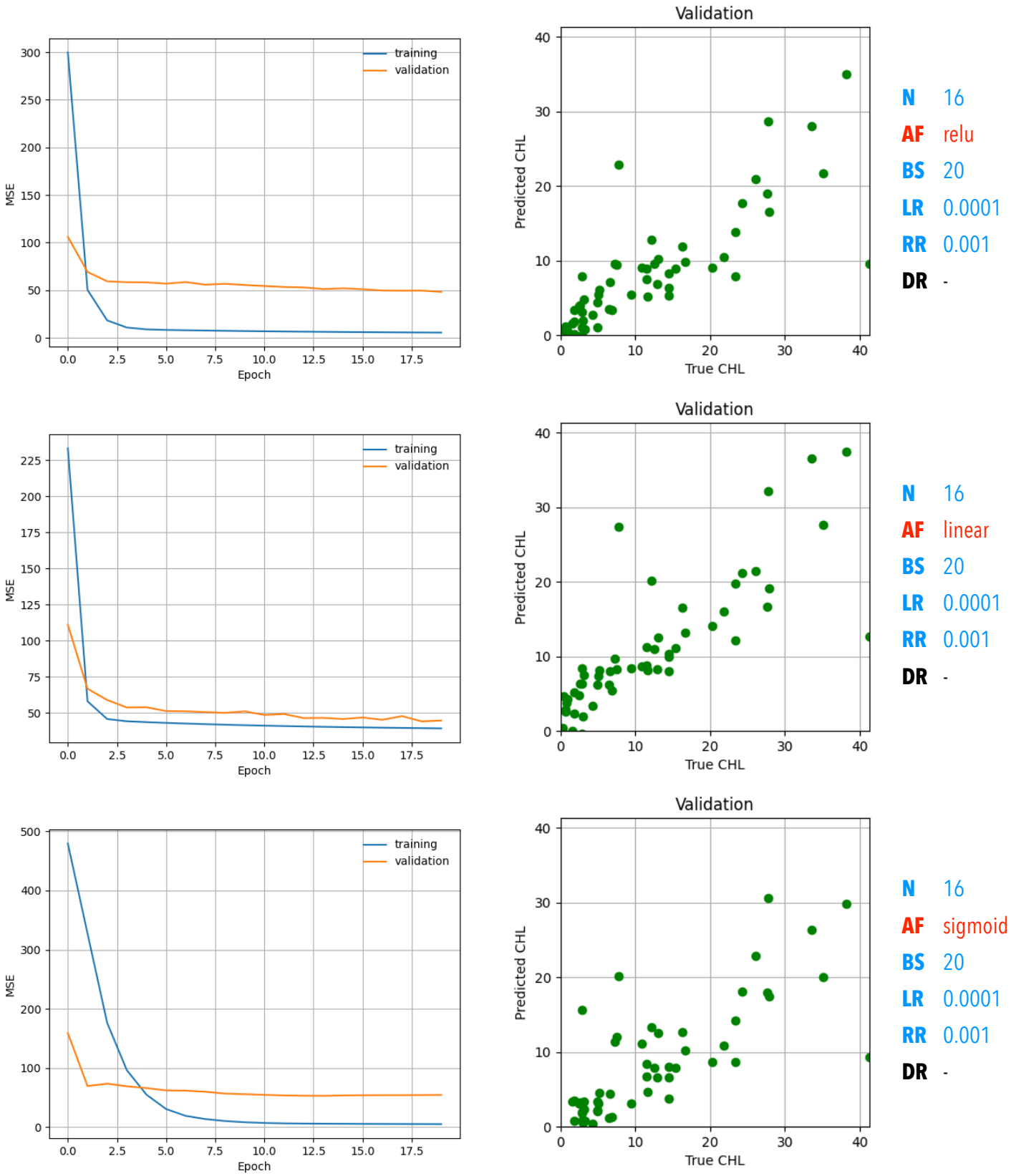

Figure 11: The variation of the activation function (select AF = linear).

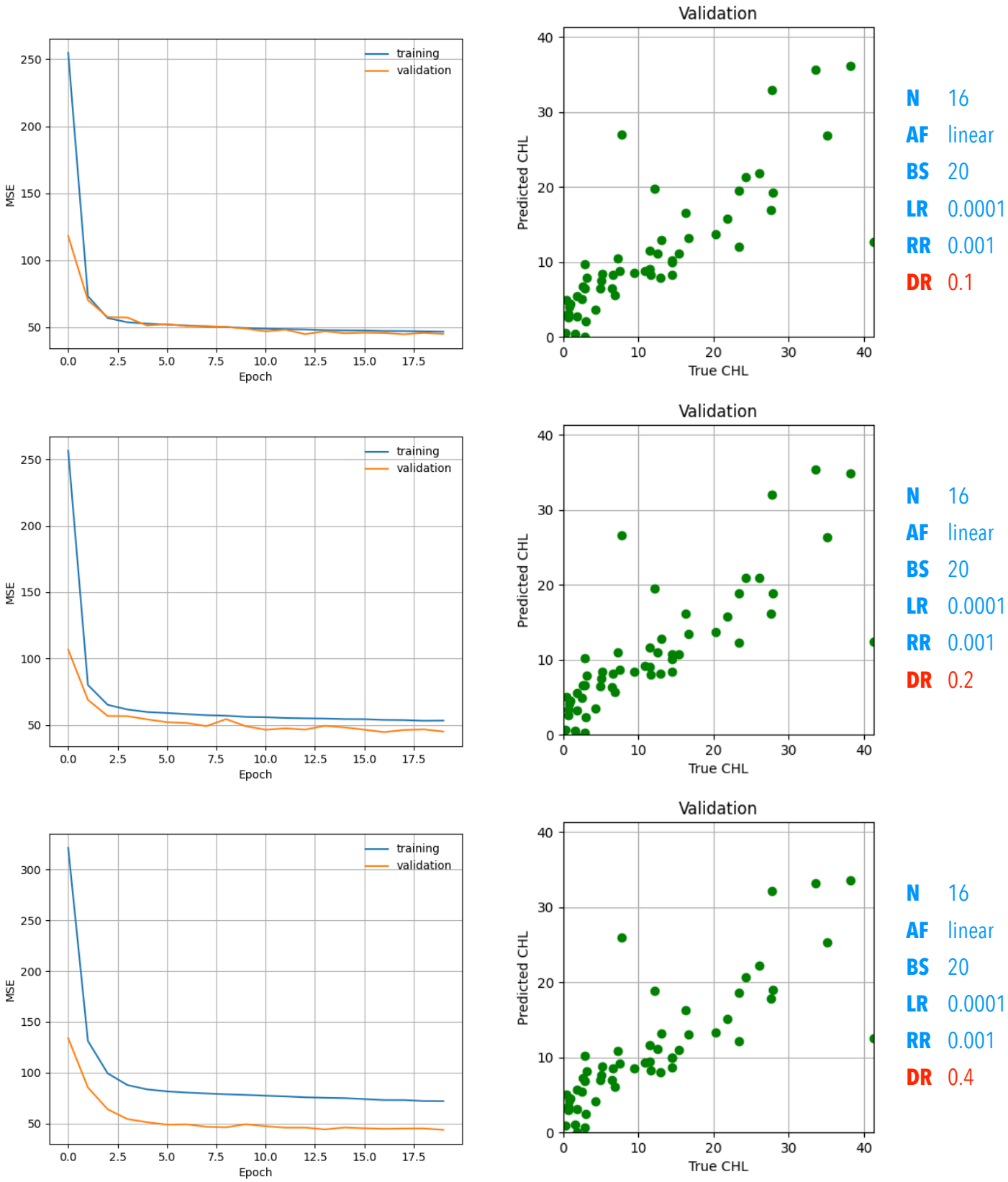

Figure 12: The variation of the dropout rate (select DR = 0.1).