

Appendix A: Additional Figures

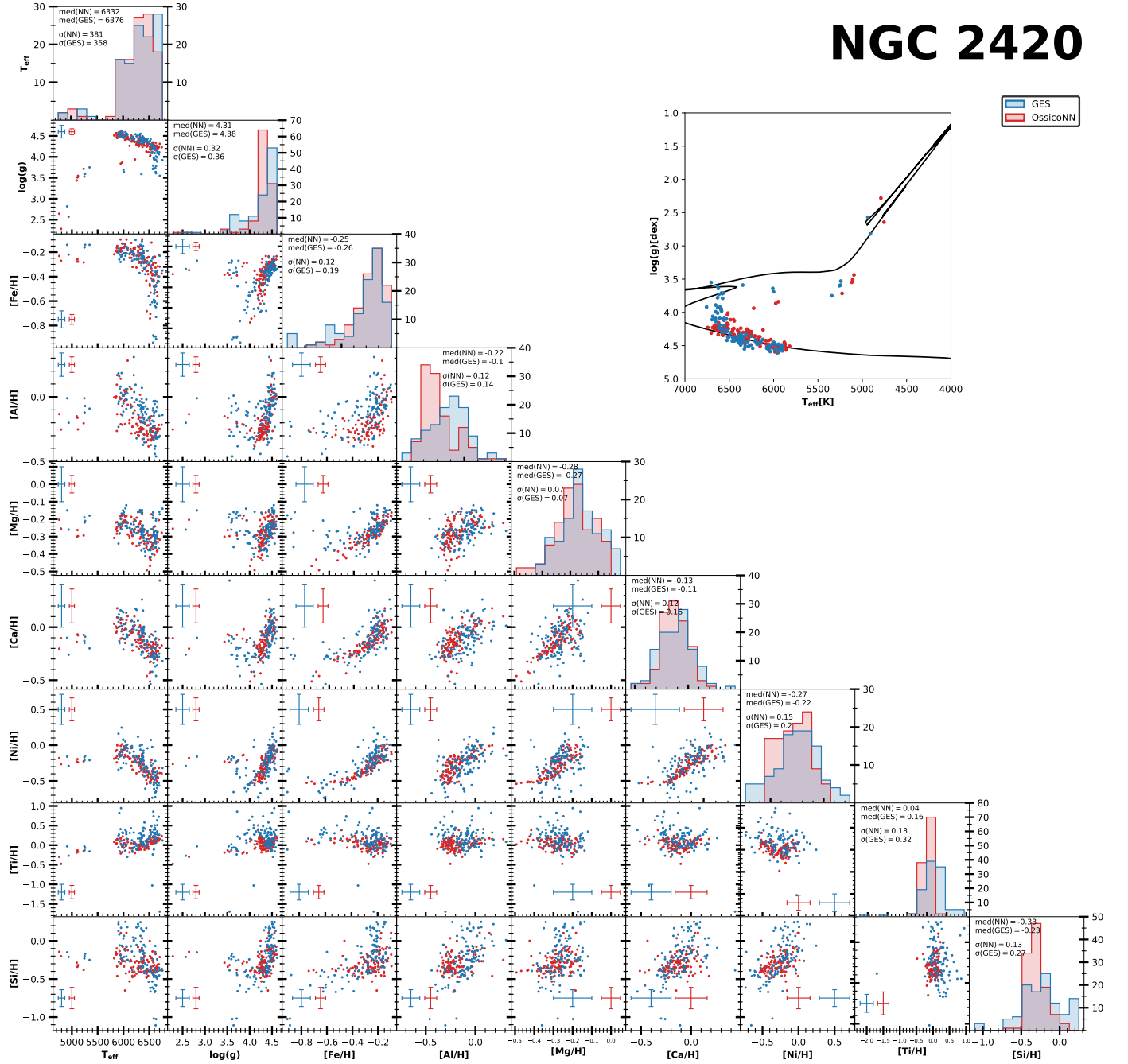


Fig. A.1. Parameters and abundances of stars belonging to the NGC2420 cluster using the GES pipeline and the OssicoNN neural network. The number of GES stars varies across quadrants since it depends on how many stars could be successfully processed through the classical pipelines for the given set of parameters. We also display the Kiel diagram and the isochrone corresponding to the age and metallicity of the cluster measured by Cantat-Gaudin et al. (2020). The isochrone is generated using PARSEC version 1.2S (Bressan et al. (2012)), Chen et al. (2014)).

Br 32

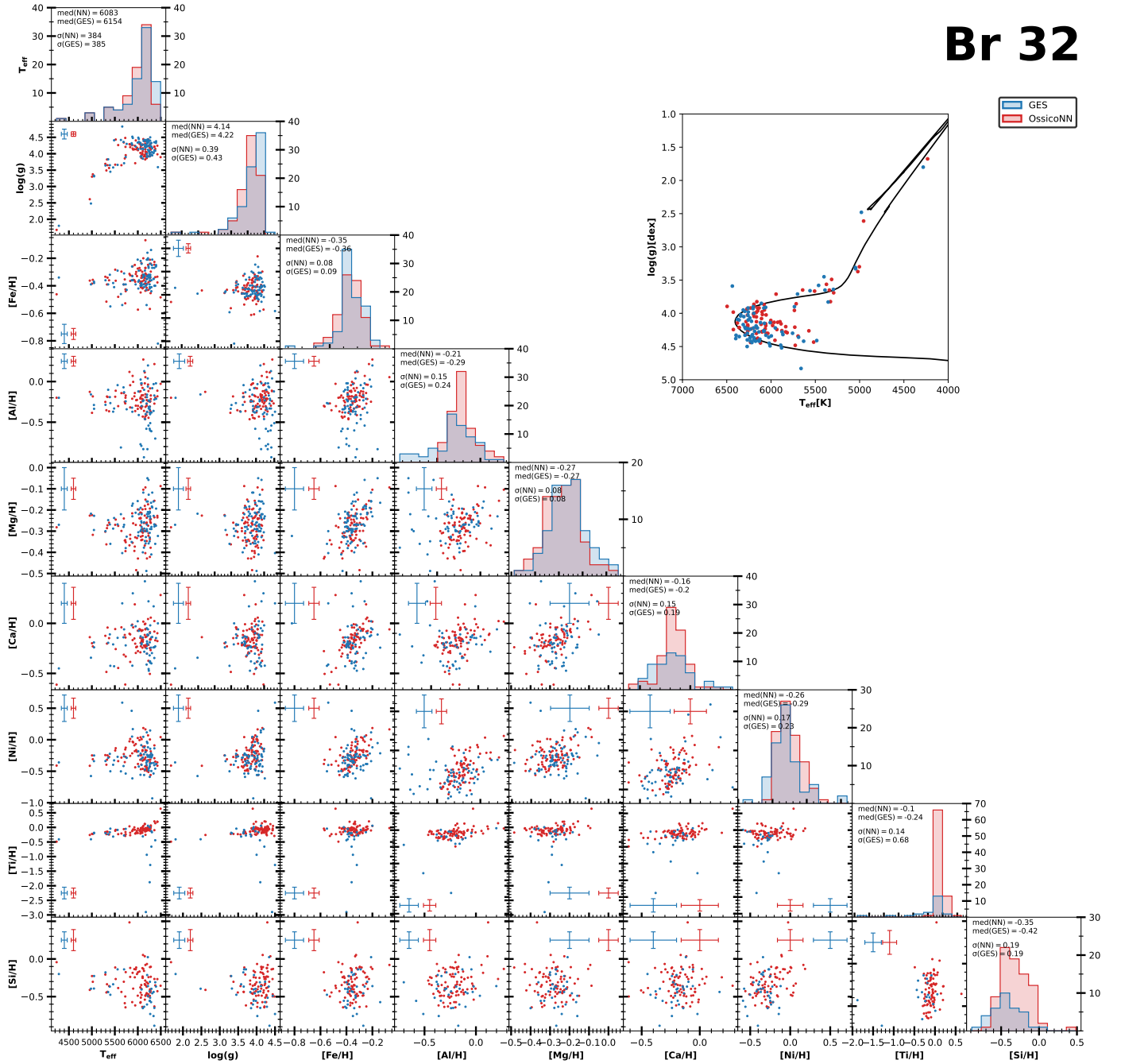


Fig. A.2. Parameters and abundances of stars belonging to the Br32 cluster using the GES pipeline and the OssicoNN neural network. The number of GES stars varies across quadrants since it depends on how many stars could be successfully processed through the classical pipelines for the given set of parameters. We also display the Kiel diagram and the isochrone corresponding to the age and metallicity of the cluster measured by Cantat-Gaudin et al. (2020). The isochrone is generated using PARSEC version 1.2S (Bressan et al. (2012), Chen et al. (2014)).

Appendix B: Latent space study

This section contains figures relating to the study of latent space and internal errors. First, we focus on how the posterior distribution of the parameters (obtained by sampling the latent space) are related in pairs for three stars. We measured the Pearson coefficients for the stars in the test set and obtained the following distributions, which are very little correlated.

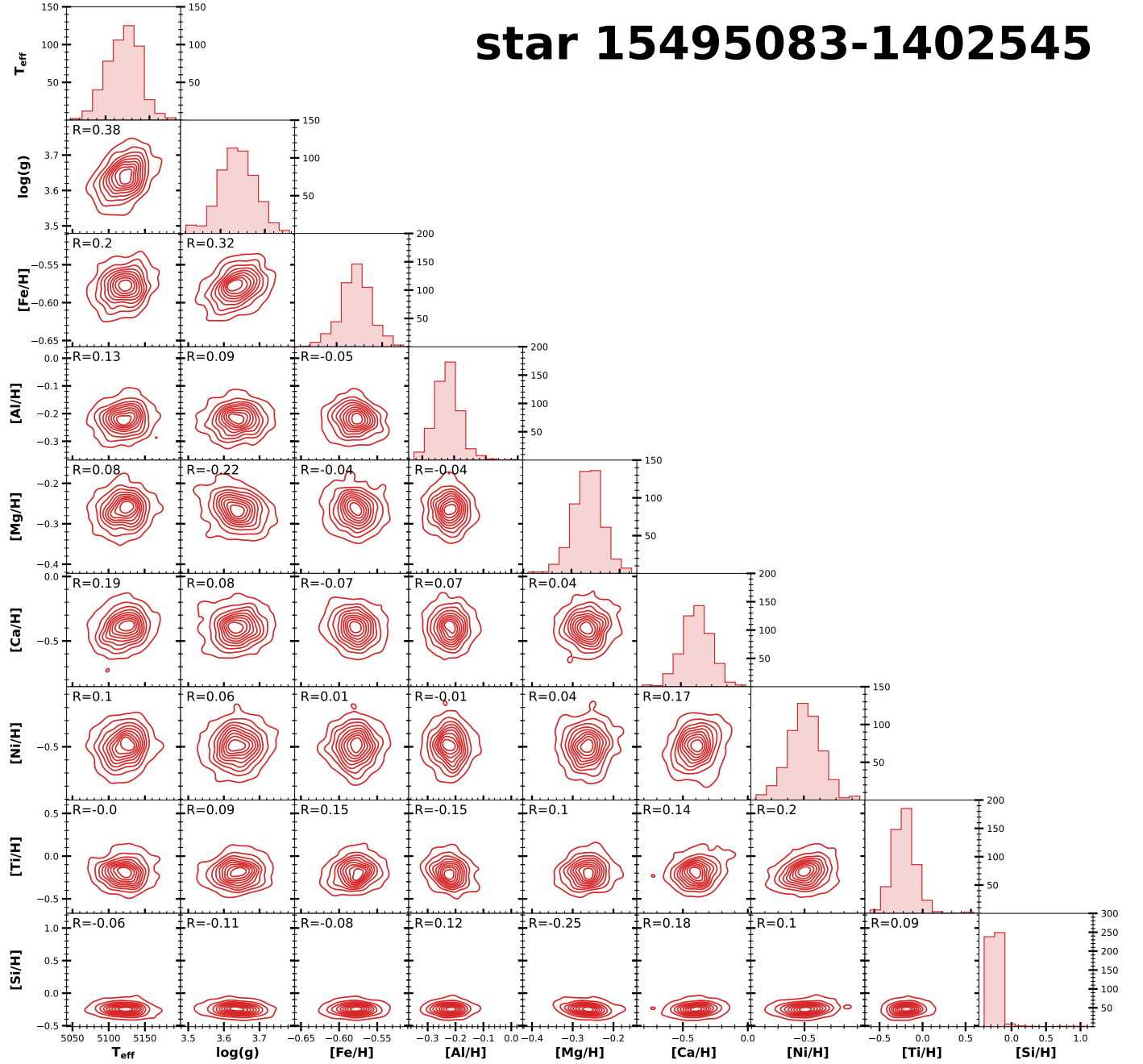


Fig. B.1. Pairwise distribution of all parameters from Latent Space sampling using OssicoNN for star 15495083-1402545 using $N_{\text{internal}} = 500$ samples. Pearson coefficient is computed for each distribution.

star 13193751-3357139

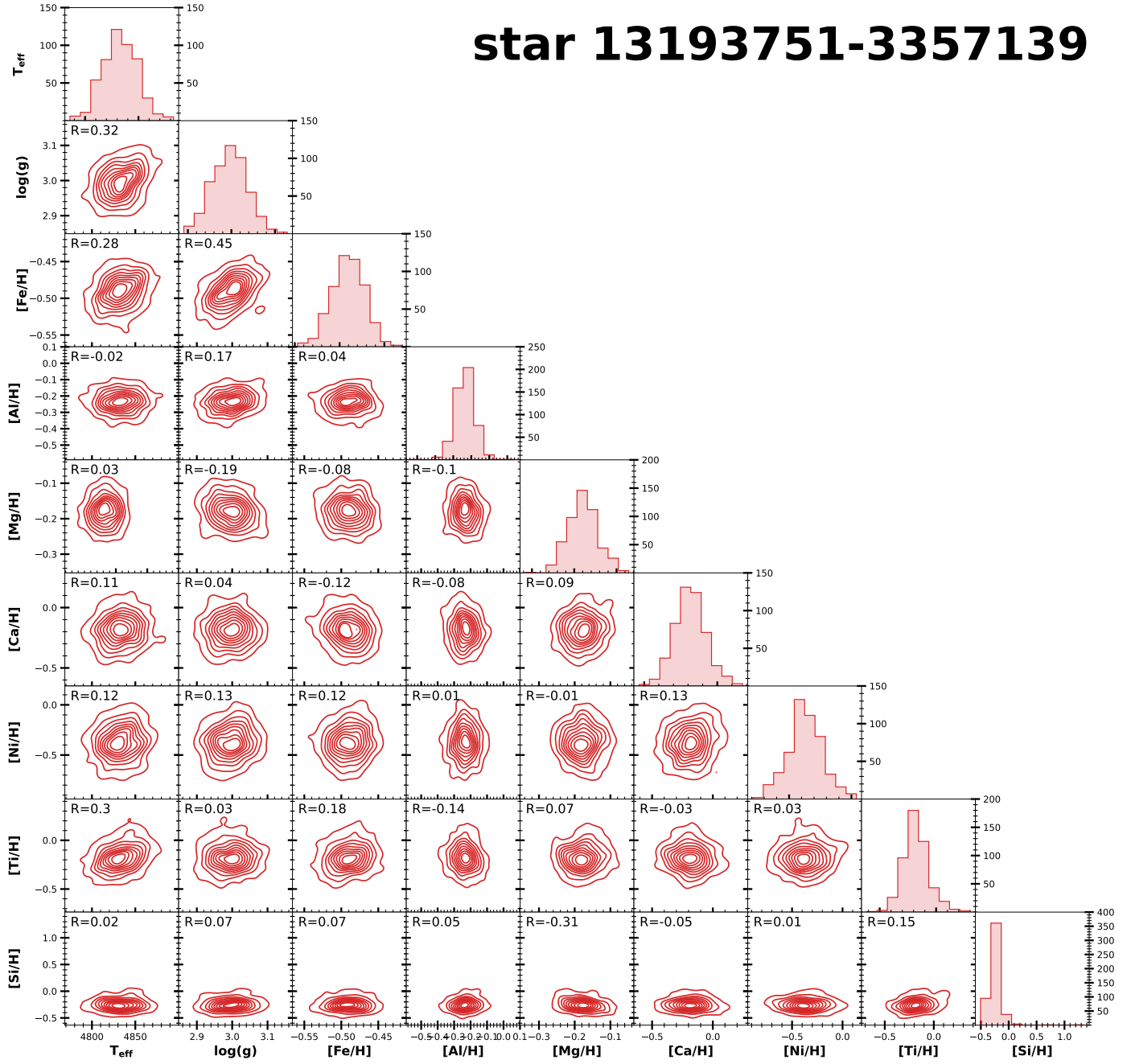


Fig. B.2. Pairwise Distribution of all parameters from Latent Space sampling using OssicoNN for star 13193751-3357139 using $N_{\text{internal}} = 500$ samples. Pearson coefficient is computed for each distribution.

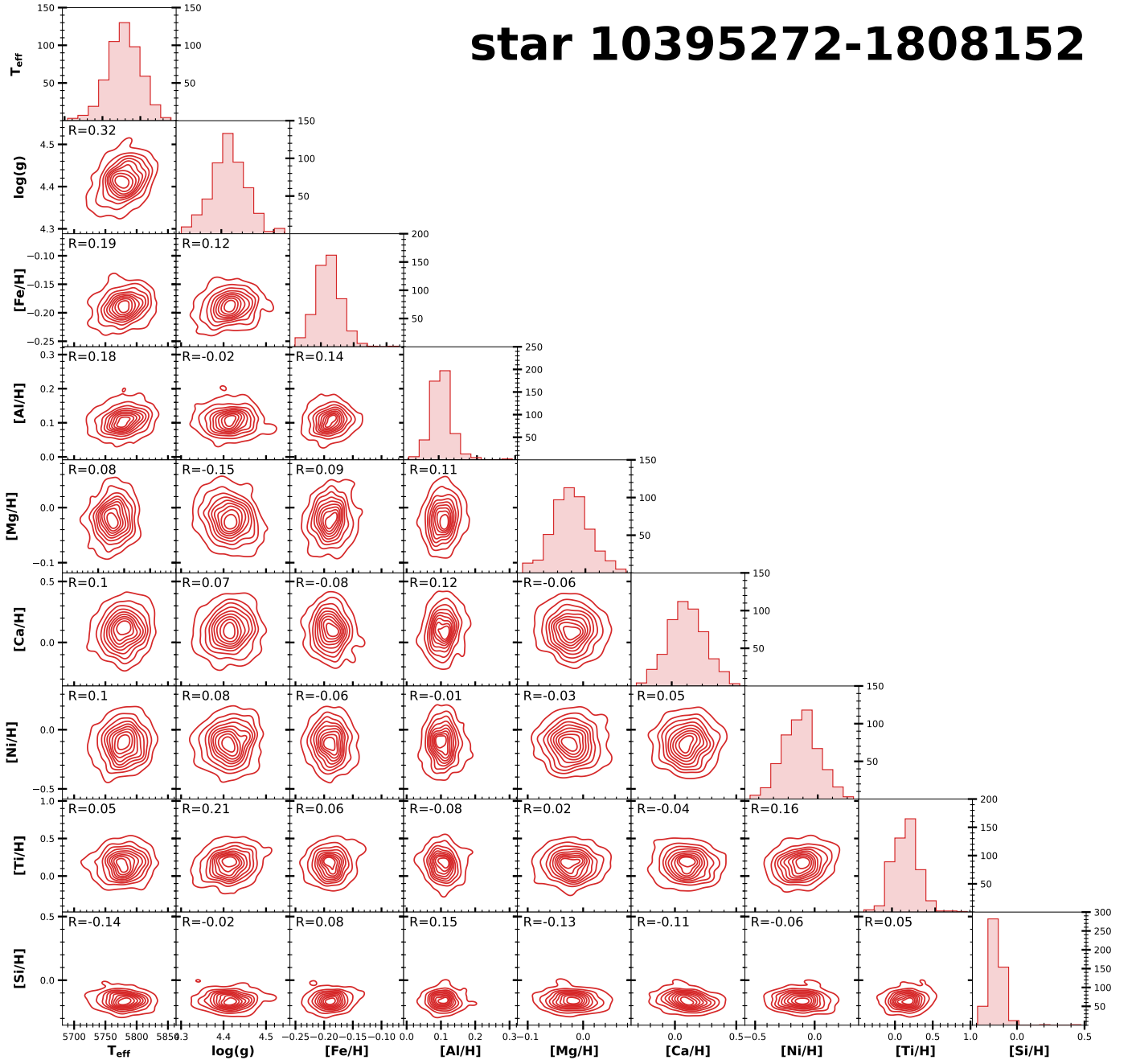
star 10395272-1808152

Fig. B.3. Pairwise Distribution of all parameters from Latent Space sampling using OssicoNN for star 10395272-1808152 using $N_{\text{internal}} = 500$ samples. Pearson coefficient is computed for each distribution.

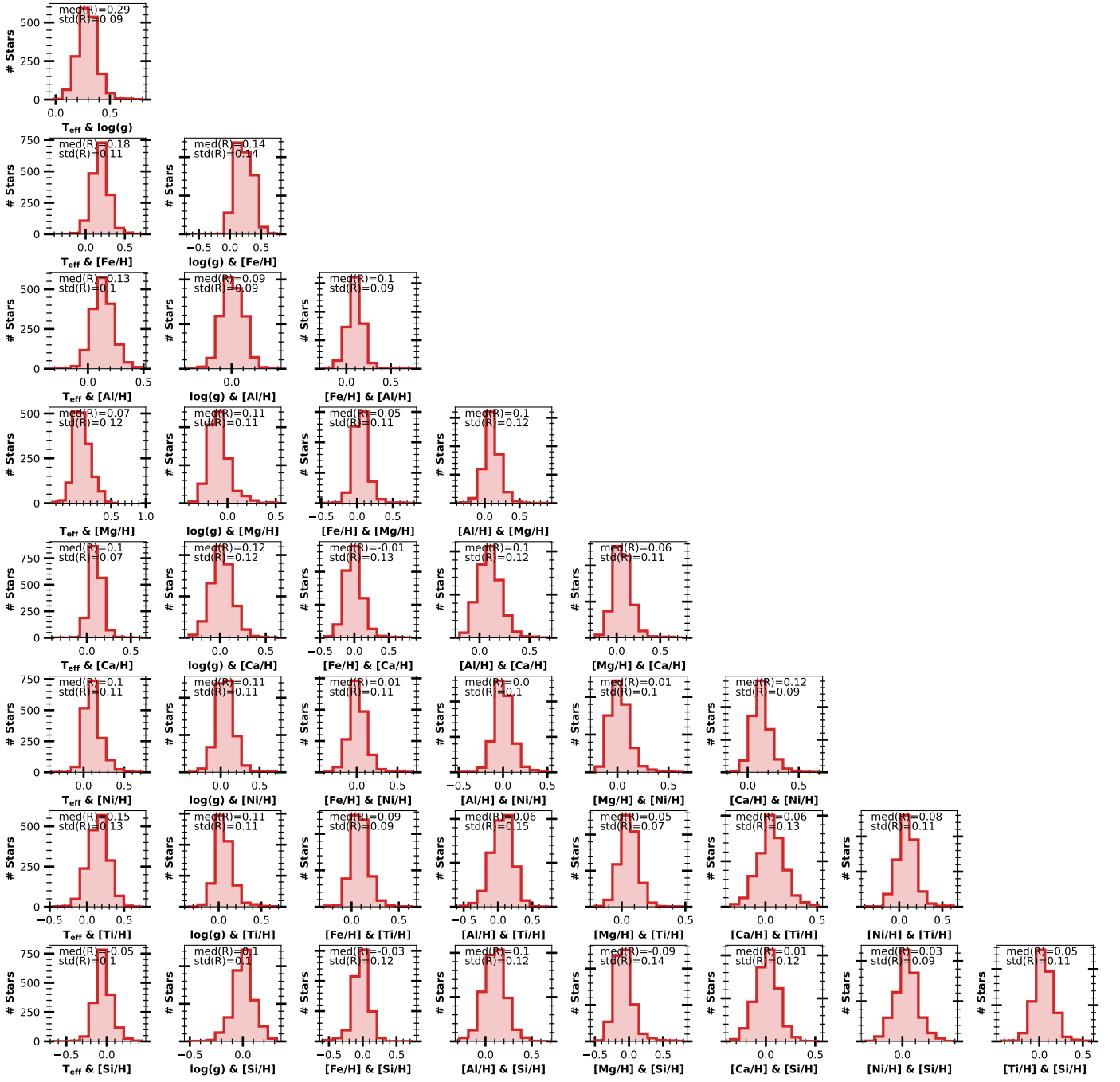


Fig. B.4. Distribution of Pearson coefficients between pairs of parameters for test set stars, derived from sampling latent space to obtain the posterior distribution for one spectrum.

Appendix C: Uncertainty: additional figures

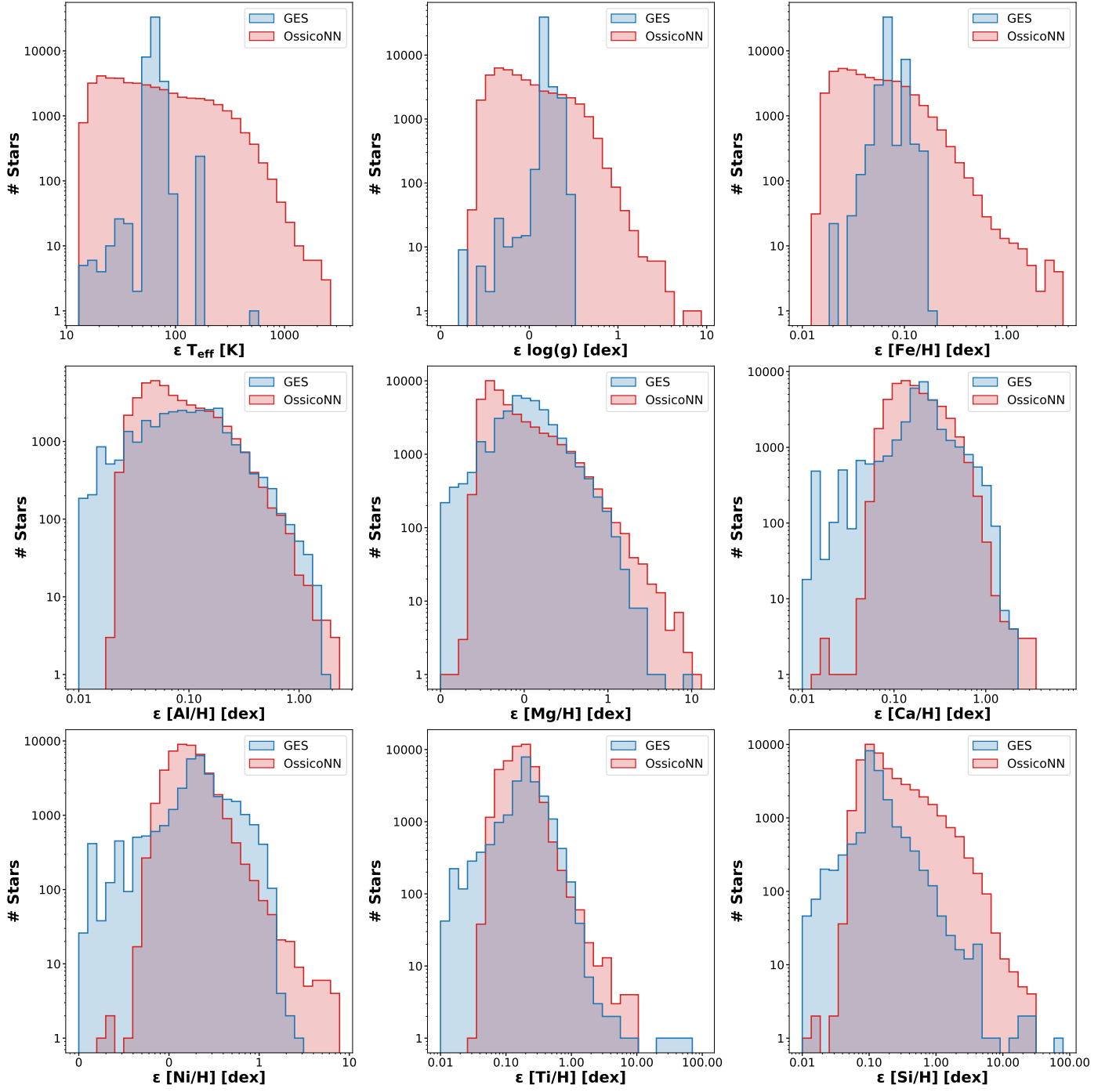


Fig. C.1. Distribution of uncertainties for GES(in blue) and OssicoNN (in red) for the Reduced Catalogue dataset (discuss in section 4.2.)

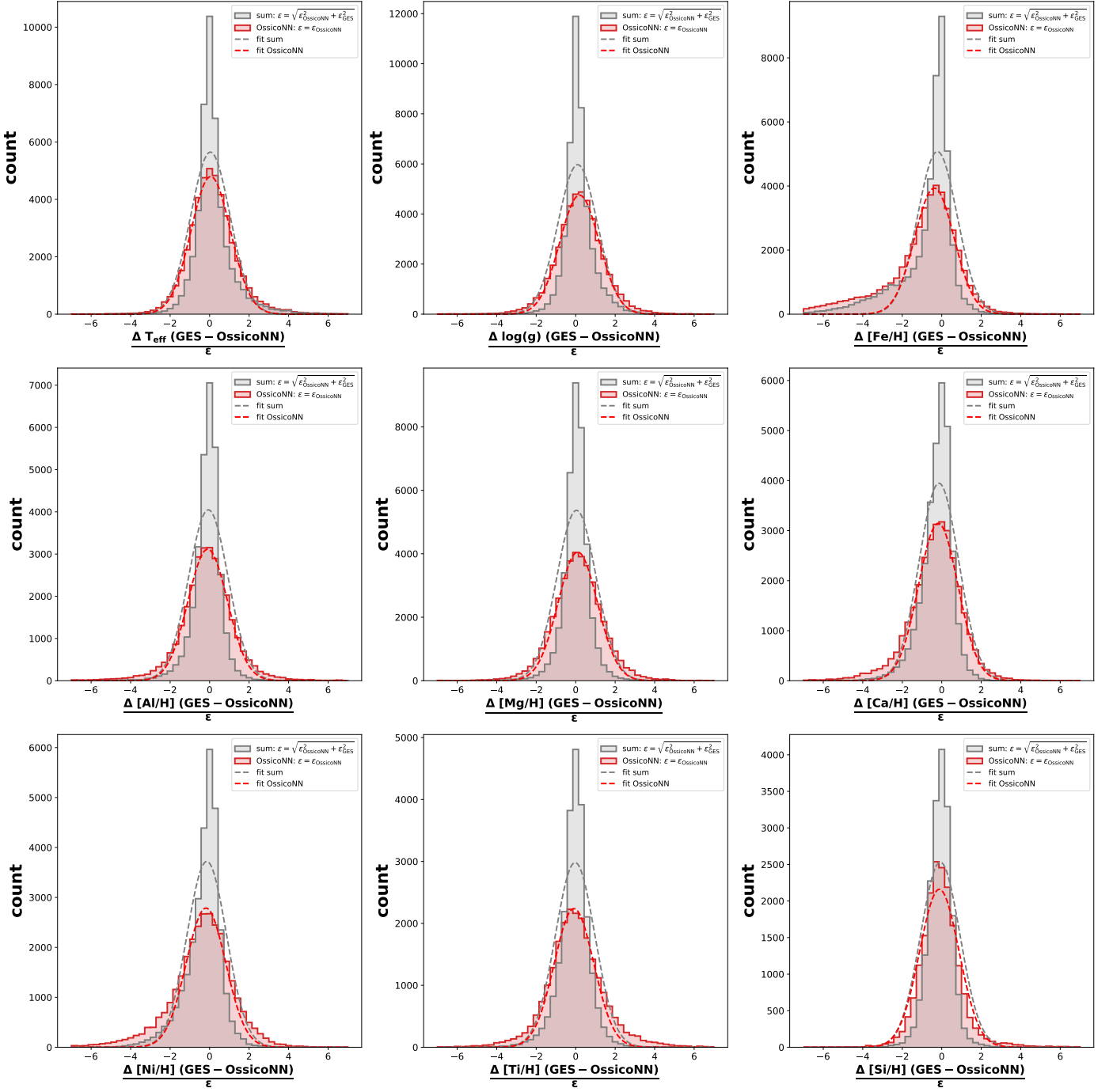


Fig. C.2. Distribution of the residuals between OssicoNN and GES, normalised with respect to the uncertainties of OssicoNN (red-shaded area) and the sum in quadrature of GES and OssicoNN errors (grey-shaded area) for the Reduced Catalogue dataset. The dashed lines represent the fit of the distribution using a Gaussian function with a standard deviation of $\sigma = 1$.

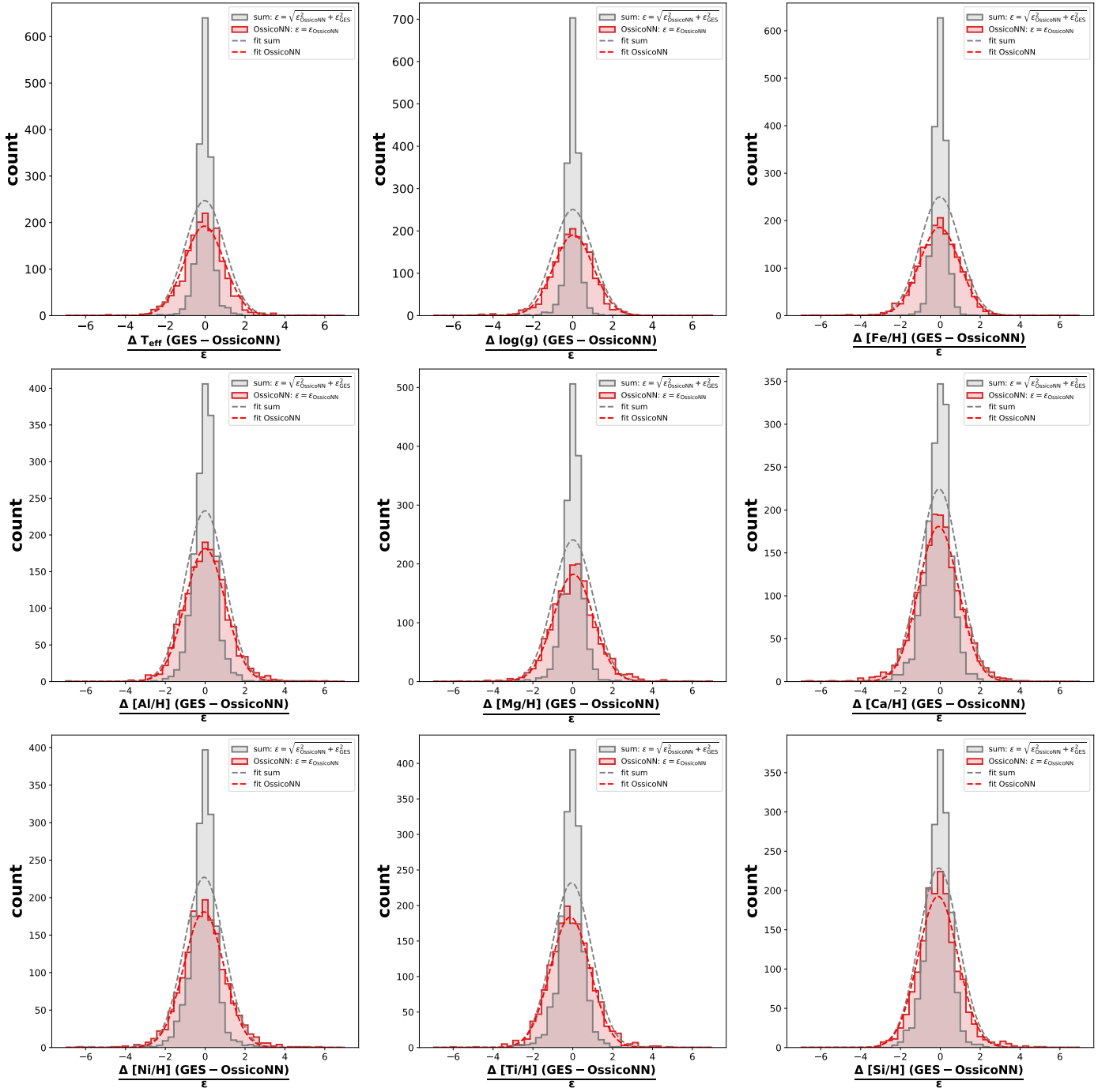


Fig. C.3. Distribution of the residuals between OssicoNN and GES, normalised with respect to the uncertainties of OssicoNN (red-shaded area) and the sum in quadrature of GES and OssicoNN errors (grey-shaded area) for the Test set. The dashed lines represent the fit of the distribution using a Gaussian function with a standard deviation of $\sigma = 1$.

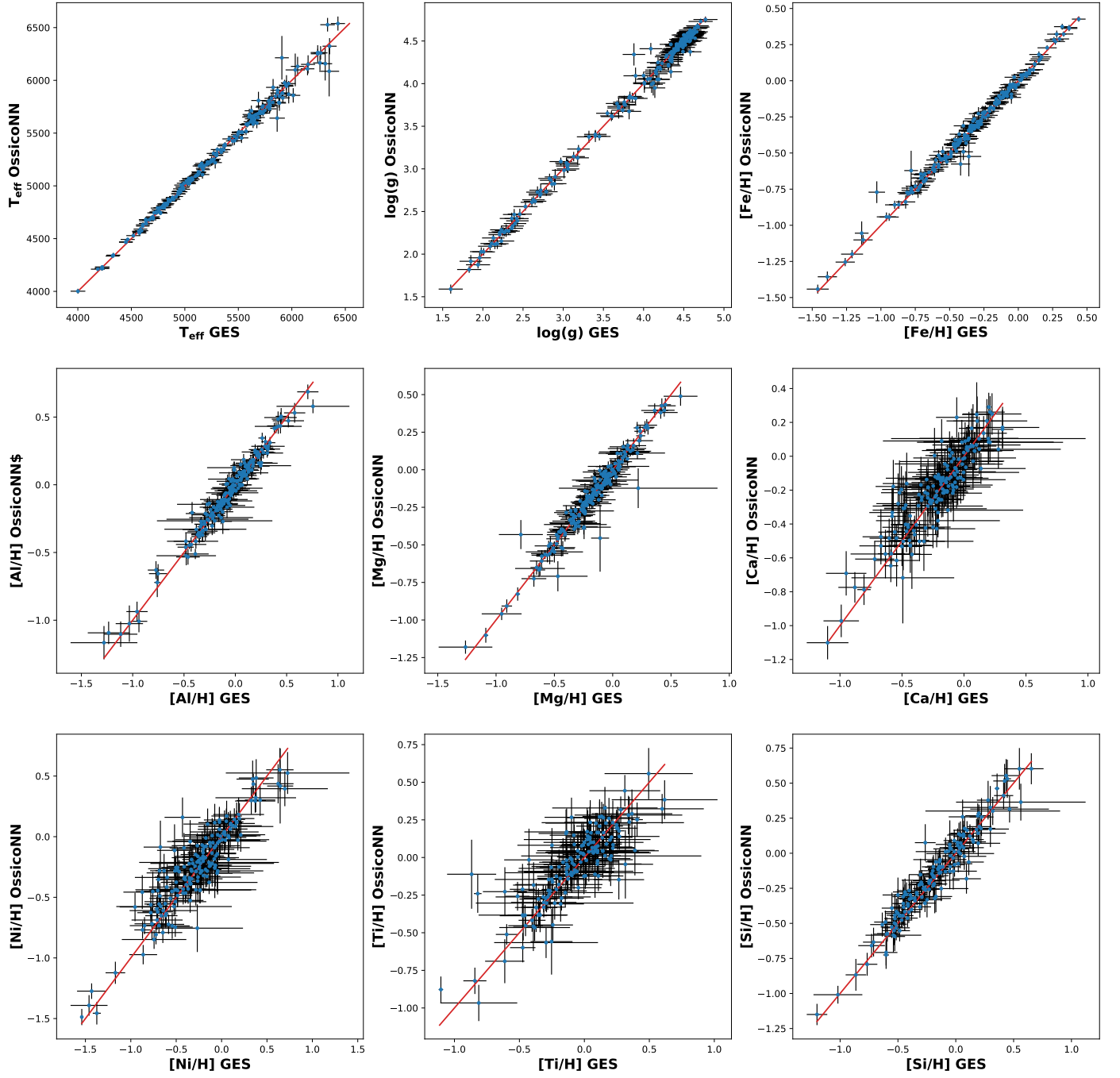


Fig. C.4. Parameter estimates for 200 randomly selected stars from the Reduced Catalogue dataset (with restrictions discussed in Section 4.2) by GES and OssicoNN, along with their respective uncertainties.

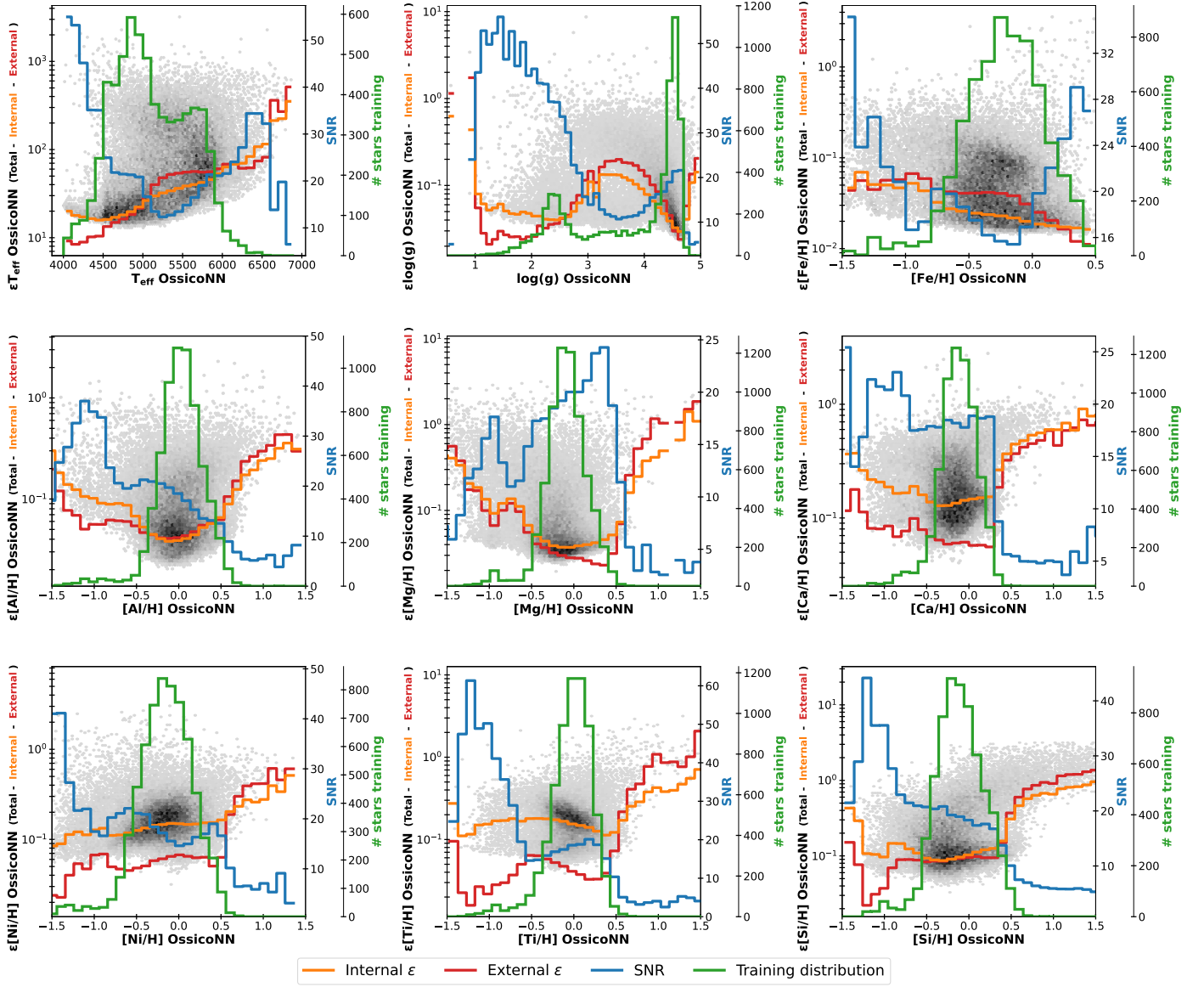


Fig. C.5. Density distribution of total OssicoNN uncertainties relative to the inferred parameters for the Reduced Catalogue dataset. The red and orange lines indicate the median external and internal uncertainties across parameter bins. The blue and green distributions, with scales on the right, represent the median S/N and the training set distribution per parameter bin, respectively.

Appendix D: Astrophysical relation: additional figures

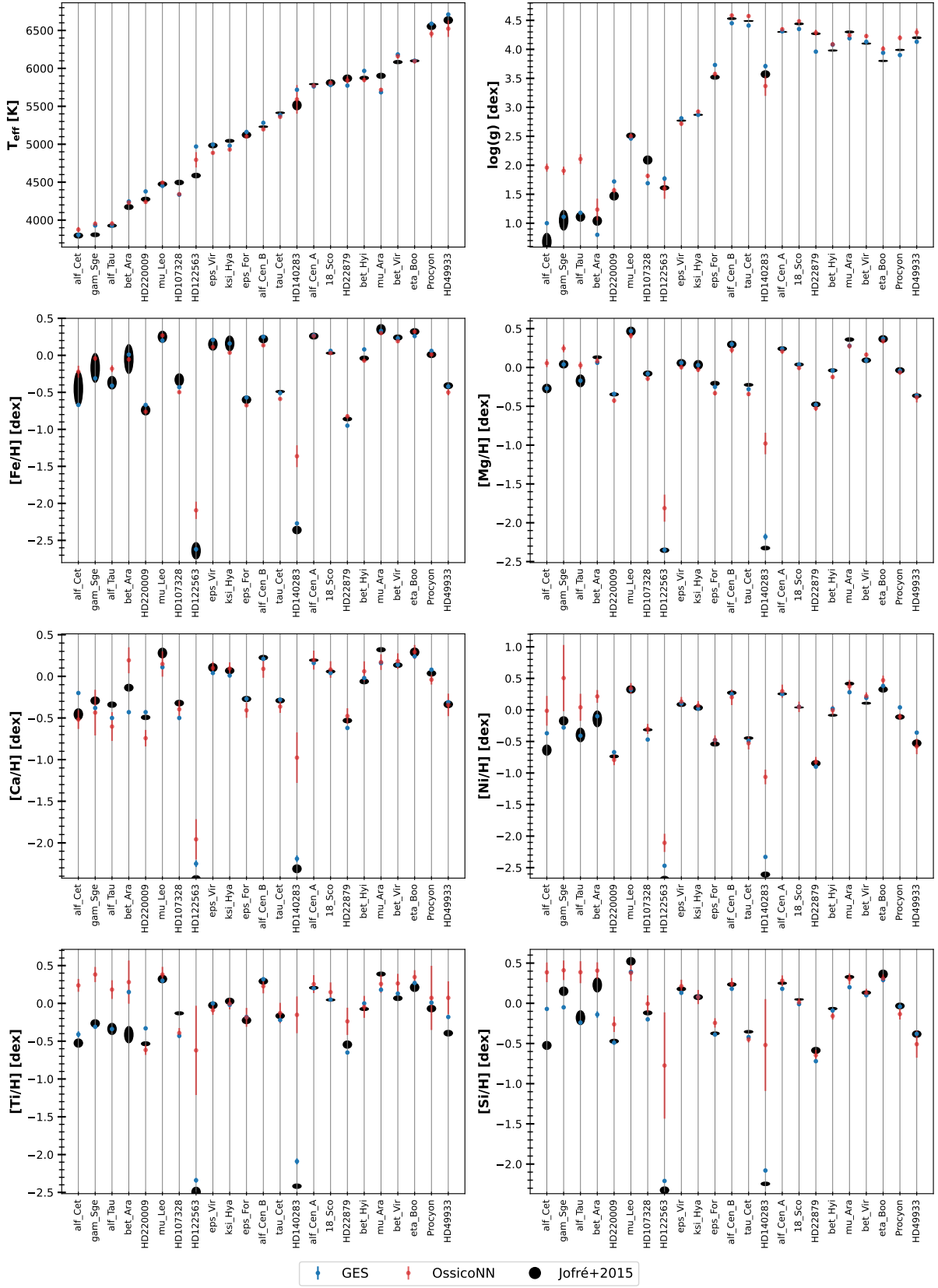


Fig. D.1. Parameter estimation for selected Benchmark stars using the GES pipeline, the OssicoNN neural network and Jofré et al. (2015), which combines spectral and spectral-independent analysis. The vertical half-axes of the Jofré+2015 points indicate the uncertainty of the estimate.

Appendix E: Maximum likelihood loss

Ardizzone et al. (2019b) suggest training cINNs by formulating a loss function that considers both the input and latent space parameters, along with the model's parameters (θ). This loss is constructed based on the likelihood function. Herein, we provide the problem definition and the calculations necessary to derive equation 3.

Let $p_Z(\mathcal{Z})$ and $p_X(\mathcal{X})$ be probability densities defined on the spaces of \mathcal{Z} and \mathcal{X} , respectively. Let θ denote the parameters of a neural network, and let y be a conditioning variable. Then, the neural network function is given by $f(x|\theta, y)$, which maps an input x to an output depending on θ and y .

Given a set of training data $(x_i)_{i=1}^n \in \mathcal{X}^n$, we can formulate the likelihood function as:

$$L(\theta) \doteq \prod_{i=1}^n p_{\theta}(x_i). \quad (\text{E.1})$$

The likelihood function measures how well the model parameters θ fit the observed data. A common approach to estimate θ is to maximise the likelihood function. The change-of-variable formula enables us to establish a connection between the probabilities in the data space and the probabilities in the latent space:

$$p_X(\mathbf{x}; \mathbf{c}, \theta) = p_Z(\mathbf{z}; \mathbf{c}, \theta) \left| \det \left(\frac{\partial f}{\partial x} \right) \right| \quad (\text{E.2})$$

$$= p_Z(f(\mathbf{x}; \mathbf{c}, \theta)) \left| \det \left(\frac{\partial f}{\partial x} \right) \right| \quad (\text{E.3})$$

Bayes' theorem can be used to determine the posterior along the model parameters:

$$p(\theta; \mathbf{x}, \mathbf{y}) \propto p_X(\mathbf{x}, \mathbf{y}; \theta) p_{\theta}(\theta). \quad (\text{E.4})$$

Starting from the likelihood minimisation and applying the formulas above we obtain :

$$\begin{aligned} \mathcal{L} &= -\log(L(\theta)) \\ \mathcal{L} &= -\log \left(\prod_{i=1}^n p_{\theta}(x_i) \right) \\ \mathcal{L} &= \mathbb{E}_i [-\log(p_{\theta}(x_i))] \\ \mathcal{L} &= \mathbb{E}_i [-\log(p_X(x_i; \theta, y_i) p_{\theta}(\theta))] \\ \mathcal{L} &= \mathbb{E}_i [-\log(p_X(x_i; \theta, y_i) - \log(p_{\theta}(\theta))] \\ \mathcal{L} &= \mathbb{E}_i \left[-\log(p_Z(f(x_i; \theta, y_i)) - \log \left(\left| \det \left(\frac{\partial f}{\partial x} \right) \right| \right) - \log(p_{\theta}(\theta)). \right] \end{aligned} \quad (\text{E.5})$$

Assuming that Z and θ follow a normal distribution

$$\begin{aligned} p_Z(z) &= \exp(-z^2/2) \\ p_{\theta} &= \exp(-\theta^2/2\sigma^2), \end{aligned} \quad (\text{E.6})$$

the result is:

$$\mathcal{L} = \mathbb{E}_i \left[\frac{\|f(x_i; \theta, y_i)\|^2}{2} - \log \left(\left| \det \left(\frac{\partial f}{\partial x} \right) \right| \right) + \frac{1}{2\sigma^2} \|\theta\|^2 \right]. \quad (\text{E.7})$$

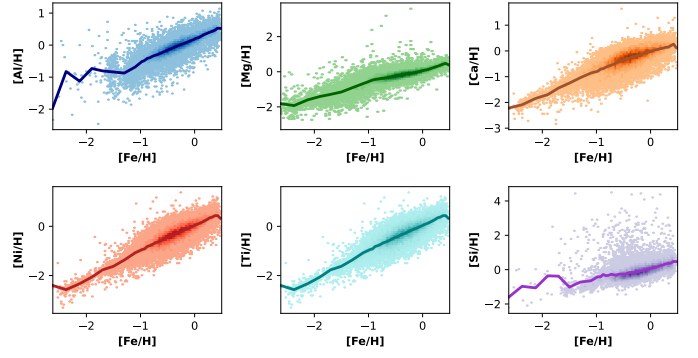


Fig. F.1. Fit of the element distribution with respect to metallicity, using parameters inferred from classical pipelines and the GES 5.1 data. To fit this distribution, we calculate the average per bin across 39 bins of varying lengths. Specifically, there are 3 bins spanning a range of 0.25 dex between -2.75 and -2 dex, followed by 5 bins between -2 and -1 dex, and finally 31 bins, each with a width of 0.05 dex, covering the range from -1 to 0.6 dex. The average (fit) is represented by the solid line.

Appendix F: Augmented dataset

One of the challenges of our machine learning approach is the limited size of the training set. To address this issue, we apply a data augmentation technique that recovers some of the spectra with incomplete parameters. This enables us to include stars that have only a few missing parameters in our training set. However, we also need to ensure that the quality and reliability of the training set are not compromised by adding noisy or inaccurate data. Therefore, we impose some selection criteria based on the importance and availability of the parameters. Temperature, gravity and metallicity are essential for stellar parameter estimation using the classical method, so we exclude any spectra that lack one of these parameters. Likewise, we discard any stars that have more than three missing values out of the six remaining parameters, to avoid introducing too many artificial values. Lastly, we apply a median S/N threshold of 25 to match the quality of the original training set. By applying these criteria, we are able to recover 7,038 spectra, which doubles the size of our training set to 13,726 spectra. We note that these augmented spectra are only used for training purposes and not for testing or validation.

To estimate the missing parameters, we adopt a simple interpolation method based on the metallicity distribution of our data set. We divide the 67,046 spectra into 39 bins according to their metallicity $[\text{Fe}/\text{H}]$ values, and compute the average element abundance for each bin (Fig. F.1). Then, we assign this average value to any missing abundance in a spectrum that belongs to that bin. This way, we preserve the correlation between metallicity and element abundance in our data set.

Using the augmented dataset not only increases the density of the training data but also extends the range of the training parameters. The temperature range now spans from 3676 to 7205 K, surface gravity ranges from 0.48 to 4.96 dex, and metallicity varies from -2.62 to 0.47 dex. Similarly, the range for aluminium extends from -2.34 to 1.06 dex, magnesium from -2.40 to 0.85 dex, calcium from -2.64 to 1.27 dex, nickel from -2.56 to 1.19 dex, titanium from -2.37 to 2.14 dex, and silicon from -2.21 to 4.39 dex. Augmenting the training set does not impact the precision of the test set, the noise in the dataset, or the uncertainty metrics. The primary changes are observed in the HR10 & HR21 full dataset, particularly in the Kiel diagram. Previously, the Kiel diagram was highly accurate for high S/N stars (see Fig. F.2), but

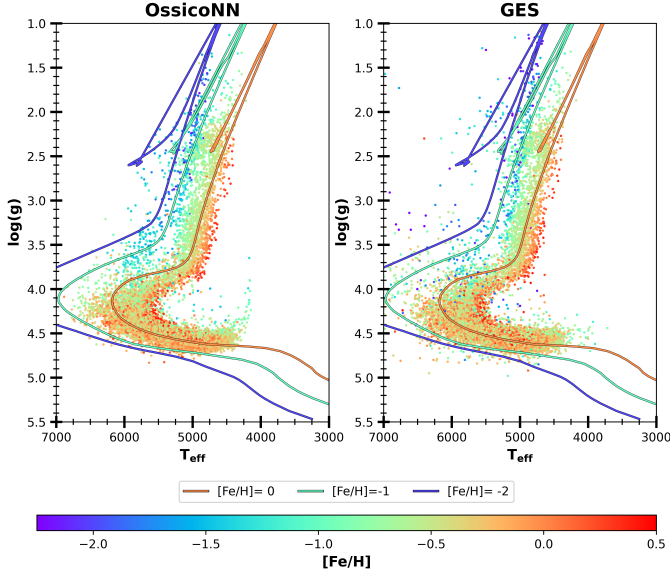


Fig. F.2. Kiel diagram: $\log g$ versus effective temperature, colour-coded by metallicity, for stars in the Milky Way field with $S/N > 25$, as derived using OssicoNN astrophysical parameters (left panel) and the GES recommended values (right panel). Superimposed are three isochrones for an age of 5 Gyr and three metallicities ($[\text{Fe}/\text{H}] = 0$: orange, $[\text{Fe}/\text{H}] = -1$: green, $[\text{Fe}/\text{H}] = -2$: blue) with colours that match the colormap). The isochrones are generated using PARSEC version 1.2S (Bressan et al. (2012), Chen et al. (2014)).

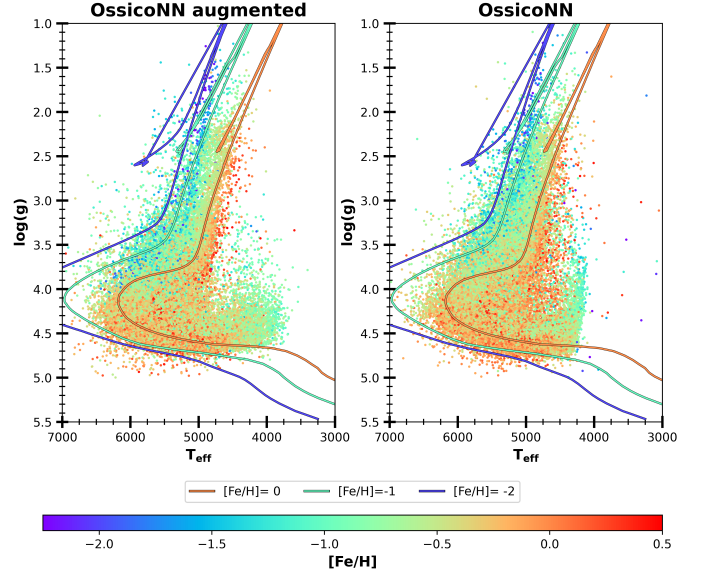


Fig. F.3. Kiel diagram: $\log g$ versus effective temperature, colour-coded by metallicity, for stars in the Milky Way field as derived using OssicoNN astrophysical parameters with the augmented dataset (left panel) and OssicoNN with the normal training set (right panel). Superimposed are three isochrones for an age of 5 Gyr and three metallicities ($[\text{Fe}/\text{H}] = 0$: orange, $[\text{Fe}/\text{H}] = -1$: green, $[\text{Fe}/\text{H}] = -2$: blue) with colours that match the colormap). The isochrones are generated using PARSEC version 1.2S (Bressan et al. (2012), Chen et al. (2014)).

when considering all stars (Fig. 10), certain populations were not well reproduced, such as the end of the red giant branch and the beginning of the main sequence. Additionally, the gradient of the red giant branch was poorly defined for sub-solar metallicities. Fig. F.3 demonstrates that the augmented dataset resolves these issues, with all star populations, including those at the edges, now well defined.

The same phenomenon is observed for the benchmark stars, where all the parameters agree with Jofré et al. (2015)'s estimates. This is particularly evident for HD122563 and HD140283, which have very low metallicities (-2.64 and -2.36 , respectively) and were previously estimated inaccurately as > -1.5 . With the augmented dataset OssicoNN_agm, the new estimates for HD122563 and HD140283 are -2.64 and -2.27 , respectively. However, we did not retain this model because, by measuring the correlation coefficients between elements, we found that the augmentation led to deviations from GES by approximately 0.10 for silicon and titanium, the two elements for which we interpolated the most values. Nevertheless, this comparison between augmented and non-augmented datasets underscores that the main issues identified by OssicoNN are due to the training set being too small. These problems are expected to be resolved in new surveys, where both the quantity and quality of data will be significantly higher.