Draft version May 1, 2023
Typeset using LaTeX **twocolumn** style in AASTeX631

Identifying Birth Environments of Isolated Stars With Clustering of Stellar Chemical Abundances

Ian Chow

Advisor: Joshua Speagle

ABSTRACT

Dynamical processes in stellar clusters can disperse stars into the galactic background as isolated stars. Although these isolated stars provide important information about the evolution of clusters and galaxies, tracing them back to their birth clusters is challenging. One technique for doing so, known as chemical tagging, is to compare their chemical abundances to that of a known cluster. While chemical tagging often makes use of dimensionality reduction methods such as Uniform Manifold Approximation and Projection (UMAP) to identify clusters, previous studies have not accounted for observational uncertainty. We investigate how uncertainty impacts chemical tagging by analyzing its effect on UMAP's structure and ability to recover known star clusters. Using a control group of 72 stars from the globular cluster M3 in a sample of 3212 stars from the APOGEE DR17 catalogue, we simulate observational uncertainty using Monte Carlo sampling and compare the UMAP projections and control group recovery fraction to a baseline result. We then investigate whether the Monte Carlo distributions of individual points in UMAP space are Gaussian by using a Kolmogorov-Smirnov (K-S) test and comparing the results to a test set of Gaussian data. We find that the global structure of UMAP is preserved under Monte Carlo sampling, though control group recovery drops significantly. In addition, the distribution of p-values for the K-S tests strongly suggests that UMAP does not propagate Gaussian errow to the low-dimensional embedding. We therefore develop a new dimensionality reduction method based on the variational autoencoder (VAE) neural network architecture, which works directly with Gaussian distributions and thus explicitly propagates Gaussian error by construction. Using the APOGEE DR17 data, we train our proposed model and evaluate its performance. We find that our model is able to robustly reconstruct the original Gaussian distributions while preserving most information in the original data, suggesting a prospective use for probabilistic cluster association in chemical tagging studies of other data sets. In the future, we also hope to generalize our method to directly take as input the spectra from which the abundances are derived.

1. INTRODUCTION

Most stars in the Milky Way generally form in stellar clusters (i.e. globular clusters, open clusters, or stellar associations). Star formation begins when overdensities in giant molecular clouds (GMCs) collapse into dense, gravitationally bound molecular cores which eventually form stellar clusters. During gravitational collapse, these nascent clusters can produce hundreds of stars at a a time (Krumholtz et al. 2014), which dynamical processes such as tidal shocks and two-body relaxation subsequently disperse into the disc of the Milky Way (Krumholtz et al. 2019). Within the dense core of these stellar clusters, three-body interactions between a binary and a single star can cause one star to be ejected from the system. In some cases, these interactions can cause a star to be expelled beyond a cluster's tidal radius entirely, at which point it is no longer gravitationally bound to the cluster and becomes an isolated star. While identifying these isolated stars is crucial in understanding the dynamics and evolution of clusters, connecting them to their birth environments using their position and kinematics alone is difficult due to the manner in which three-body interactions disperse stars. However, stars generally retain the relative elemental abundances from their birth environments when they formed. Assuming a uniform composition of a stellar cluster's parent GMC, we can therefore identify the birth clusters of isolated stars by comparing their chemical "fingerprint" to that of a known cluster. This technique is known as chemical tagging, and has been used to find cluster associations of isolated stars (e.g. Navin et al. (2016), Martell et al. (2016), Schiavon et al. (2017)).

Statistical clustering algorithms such as density-based spatial clustering of applications with noise (DBSCAN) can be used to recover star clusters in high-dimensional abundance data (Price-Jones & Bovy 2019). However, clustering algorithms generally perform poorly on high-dimensional data due to the curse of dimensionality, necessitating the use of dimensionality reduction methods as a preprocessing step. These methods project high-dimensional data into a lower dimension while preserving as much intrinsic information about the original data as possible. Recent developments in machine learning have introduced new dimensionality reduction algorithms for chemical tagging such as t-Stochastic Neighbor Embedding (t-SNE) (van der Maaten & Hinton 2008) and Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018). While t-SNE and UMAP are powerful algorithms that have been used in previous chemical tagging studies (e.g. Grondin et al. (2023), Chun et al. (2020), Anders et al. (2018)), they do not take into account any uncertainty in the data. As such, chemical tagging studies using these algorithms have yet to incorporate observational uncertainty into their analysis. We therefore aim to develop a method for dimensionality reduction and clustering that can propagate uncertainty in a probabilistic manner during chemical tagging, with the end goal of improving the robustness of cluster associations when identifying isolated stars.

We structure our paper as follows: Section 2 introduces the data used in our analysis. Section 3 reviews applications of t-SNE and UMAP to chemical tagging and provides a baseline UMAP projection for later comparison. In section 4.1, we investigate how uncertainty in the data qualitatively impacts the overall structure of UMAP projections as well as its ability to recover known star clusters, using data from the globular cluster M3 as a test case. We then demonstrate in section 4.2 that UMAP does not probabilistically propagate uncertainty from high-dimensional to low-dimensional space in general; specifically, we show that the Gaussian measurement error in the original high-dimensional data is not preserved in the lower-dimensional embedding. Motivated by these results, section 5 proposes a new dimensionality reduction method, based on the variational autoencoder (VAE) neural network architecture, that is constructed to explicitly propagate Gaussian error (unlike UMAP). Using data from the APOGEE DR17 release, we demonstrate that our method is able to preserve most of the information during dimensionality reduction, while reproducing Gaussian error in the low-dimensional representation. We also compare the performance of our model to that of a "naive" VAE using the default architecture. We discuss our results in section 6. Finally, we summarize our conclusions in section 7, including prospects for applying this method in future chemical tagging studies.

2. DATA

In our work, we use data from the DR17 data release (Abdurro'uf et al. 2022) of the Apache Point Observatory Galactic Evolution Experiment (APOGEE) spectroscopic survey, which contains high-resolution ($R \sim 22,500$), high signal-to-noise ratio (> 100), infrared (1.51–1.70μm) spectra for 370,060 stars in the Milky Way (Majewski et al. 2017). Our analysis considers stellar chemical abundances and radial velocities as parameters for dimensionality reduction, as we expect stars in a gravitationally bound cluster to have similar kinematic profiles in addition to chemical abundances. We use the chemical abundance data derived by Leung & Bovy (2019) using their `astroNN` neural network in Python. The data contains 19 `astroNN` chemical abundances in total, as well as radial velocities from the *Gaia* EDR3 data release (Gaia Collaboration 2021). The elements present in the chemical abundance data can be seen in Fig. 1, which is taken from Fig. 3 of Grondin et al. (2023). After filtering stars with duplicate spectra, low signal-to-noise ratio (< 50), lacking chemical abundance data, or otherwise flagged as having potential issues, we obtain a filtered sample of 144,767 stars for our analysis.

In sections 3 and 4, we analyse the impact of uncertainty on UMAP using the globular cluster M3. Following the method in Section 2 of Grondin et al. (2023), we first restrict our selection to those stars within a $10^\circ \times 10^\circ$ field of view around the center of the globular cluster M3, which consists of 3212 unique stars. We then identify an initial control group of 72 M3 members by selecting all stars located within 4 times the half-mass radius of $r_{hm} = 6.34$pc $\approx 0.036^\circ$ (Baumgardt & Hilker 2018) from M3's center in the plane of the sky, following a similar procedure to that outlined by Grondin et al. (2023). A preliminary comparison of the abundance distributions for M3 and the field stars (the $10^\circ \times 10^\circ$ sample) with outlier stars (defined as the most extreme 5% of the data) removed is shown in Fig. 1. Note that we use a more restrictive control group ($4 \times r_{hm}$ compared to $8 \times r_{hm}$) than Grondin et al. (2023) to minimize the chance of including background field stars in our control group, since we do not further select control group stars by chemical abundance. Nevertheless, the chemical fingerprint is clear when qualitatively observing the location of the M3 abundance distributions in parameter space compared to the background in Fig. 1,

supporting our method of selecting a chemically unique control group for subsequent analysis.

For sections 5 and 6, we use all 19 chemical abundances (without the radial velocities this time) of the full filtered set of 144,767 stars to train and evaluate the performance of our neural network. The data is divided into a training set of 115,000 stars and a test set of 29,767 stars for an approximately 75% training and 25% testing split.

3. T-SNE AND UMAP

Methods such as t-SNE and UMAP project high-dimensional data into a lower dimension while preserving as much intrinsic information about the original data as possible. Given a high-dimensional data set, both algorithms first calculate the distance between any two given points using a "similarity score" metric. For t-SNE, the similarity score between a point and any other point in the data is computed using the height of a Gaussian distribution centered on that point, normalized so all similarity scores for a given point sum to 1. In contrast, UMAP only computes similarity scores for the n nearest neighbours of a point, and scales all similarity scores for a point to sum to $\log_2(n)$. Both t-SNE and UMAP then initialize a low-dimensional embedding of the data and then compute similarity scores for any given point in the low-dimensional embedding using a t-distribution centered on that point. The algorithms then iteratively move points around their respective embeddings so that the low-dimensional similarity scores are as close to the high-dimensional scores as possible. Note that while the values of the projected points in t-SNE and UMAP do not have a corresponding physical interpretation, the relative high-dimensional distances between points are preserved through projection to low-dimensional space, allowing us to perform clustering on the low-dimensional embedding.

Although t-SNE and UMAP work in a very similar manner, our analysis in this paper will primarily use UMAP for several reasons. While t-SNE moves every point during every iteration (and thus recomputes all the similarity scores), UMAP moves only a small subset of points (the n nearest neighbours of a given point) at a time. In addition, t-SNE uses a random initialization of its low-dimensional graph, while UMAP performs spectral embedding to initialize the graph into a good starting state. This makes UMAP much faster and less computationally expensive to run compared to t-SNE, and usually results in a more stable projection when run on the same data multiple times or when adding points to the data. Moreover, UMAP is more flexible since it allows projection of data into an arbitrary dimension, while t-SNE is limited to a two-dimensional embedding.

3.1. *Data Preprocessing*

As a preprocessing step before performing any dimensionality reduction, we normalize the data to account for parameters with a large variance so that UMAP and t-SNE will equally weight all chemical abundances and radial velocities when computing a similarity score in the dimensionality reduction process. For unnormalized data, the distances between points in a low-dimensional embedding would be dominated by parameters with large variance, similar to principal components analysis. As we assume all parameters are equally important in the chemical tagging process, normalizing the data ensures that the embedding is not dominated by a few parameters. We therefore normalize the distribution of each parameter in the data by subtracting the parameter mean from every point and dividing by the variance so that $\mu = 0$ and $\sigma = 1$ across all chemical abundances and radial velocities.

3.2. *UMAP Analysis*

We begin by fitting UMAP on the normalized data for all 3212 stars in our sample, including both the control group and field stars, as shown in Fig. 2. The UMAP algorithm is implemented using the `umap-learn` (McInnes et al. 2018) package in Python. The resulting projection shows that the UMAP embedding recovers almost all the M3 control group stars using the group's unique chemical fingerprint. Fig. 2 also provides a baseline for us to compare against in subsequent sections when analyzing the effects of uncertainty on UMAP. We set a minimum distance between embedded points of 0.1 and a local neighbourhood size of $n = 15$, which was empirically determined to balance preserving the overall structure of the data with retaining detail in its local features.

3.3. *t-SNE Analysis*

For completeness, we also fit t-SNE on the same normalized data used for UMAP, to confirm that our M3 control group can be consistently recovered by multiple dimensionality reduction approaches. Fig. 3 shows the t-SNE projection for all 3212 stars in our sample, implemented using the `scikit-learn` (Pedregosa et al. 2011) package in Python. Similar to the UMAP projection, we see that the t-SNE embedding recovers almost all the M3 control group stars, providing confidence that the identified control group has a distinct chemical fingerprint recoverable by dimensionality reduction. For t-SNE, we provide a perplexity value of 30, with other hyperparameters set to the default values used by the

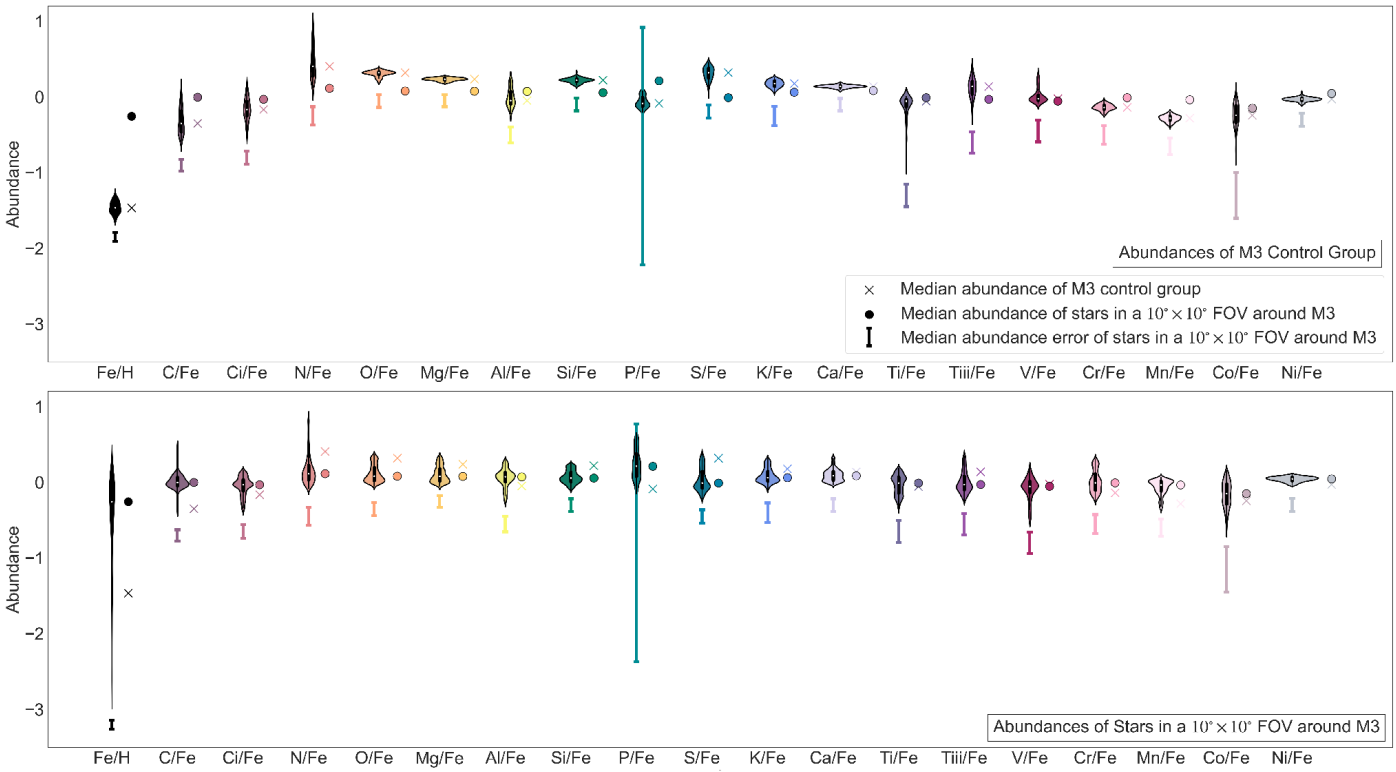


Figure 1. Distributions of relative elemental abundances for 19 elements, in units of $\log_{10}$ of the given ratio, shown as violin plots for the M3 control group (top) and background stars (bottom), taken from Fig. 3 of Grondin et al. (2023). Outlier stars for both the M3 group and the full sample are removed by selecting the central 95% of the data. The median abundances for the M3 control group and background stars are labelled with crosses and circles, respectively. The median abundance errors for background stars are indicated with error bars. Note that Grondin et al. (2023) use a larger control group of 133 stars within 8 times the half-mass radius of M3's center for this plot, while we use a smaller control group of 72 stars within 4 times the half-mass radius in our analysis. Even with this larger group, the chemical fingerprint of the M3 cluster is apparent when comparing the median abundances of the M3 group to the background stars. Moreover, the M3 distributions have a generally lower variance than their corresponding background star distribution.

`scikit-learn` implementation. Similar to neighbourhood size in UMAP, perplexity balances attention between local and global aspects of the data, with our perplexity value again chosen empirically to preserve both large-scale structure and small-scale detail.

4. EFFECTS OF UNCERTAINTY ON UMAP

4.1. *Global Structure*

To qualitatively determine the effects of measurement uncertainty on the global structure of the UMAP embedding, we first simulate uncertainty by performing Monte Carlo sampling on the original data, and then compare the resulting projections of the Monte Carlo samples to the original result in Fig. 2. We produce a single Monte Carlo sample by drawing each parameter of every point from a corresponding Gaussian distribution with mean equal to the observed value and variance equal to the square of the measurement error. The parameter values are then normalized using the means and variances from the original data, as described in section 3.2. UMAP is then performed on the Monte Carlo sample using the same fit as the original data.

Fig. 4 shows the UMAP projections for this procedure on 6 Monte Carlo realizations, while Fig. 5 shows an aggregate plot of the projections on a larger set of 100 realizations. Stars from the M3 control group are labelled on both figures. We see that while Monte Carlo realization generally preserves the global structure of UMAP, it reduces the fraction of M3 points recovered from $\sim 97\%$ in the original data to $\sim 64\%$ under Monte Carlo realization. In addition, the unrecovered M3 points are scattered across different areas of the resulting projection, suggesting that individual points may be highly delocalized in UMAP space.

4.2. *Local Structure*

We can determine whether the distribution of a point in UMAP space is in general Gaussian by comparing it to a known distribution using a statistical test like the one-sample Kolmogorov-Smirnov (K-S) test, which

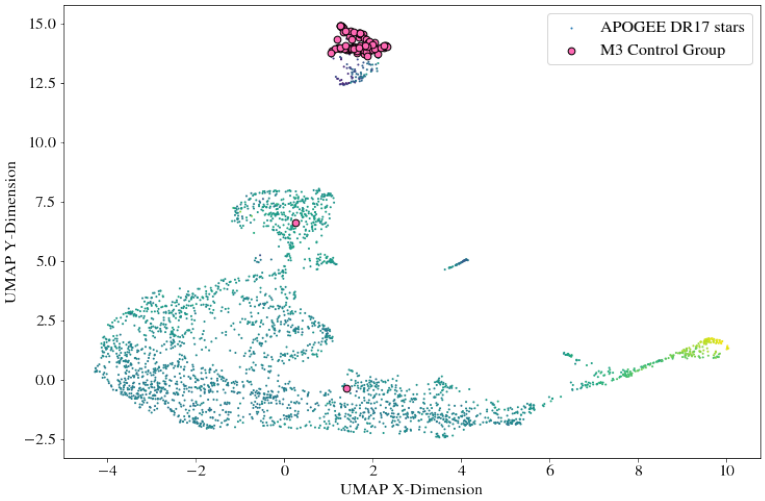


Figure 2. UMAP projection of 3212 APOGEE DR17 stars in a 10° × 10° field of view around M3 using 19 chemical abundances and radial velocities. A minimum distance between embedded points of 0.1 and a local neighbourhood size of $n = 15$ is used. Stars from the M3 control group, as identified in Section 2, are labelled in pink. Field stars are coloured by their C/FE ratio, showing the variation along one axis of the original high-dimensional abundance data in the projection. UMAP successfully recovers almost all the M3 control group stars as a cluster of points at the top of the plot.

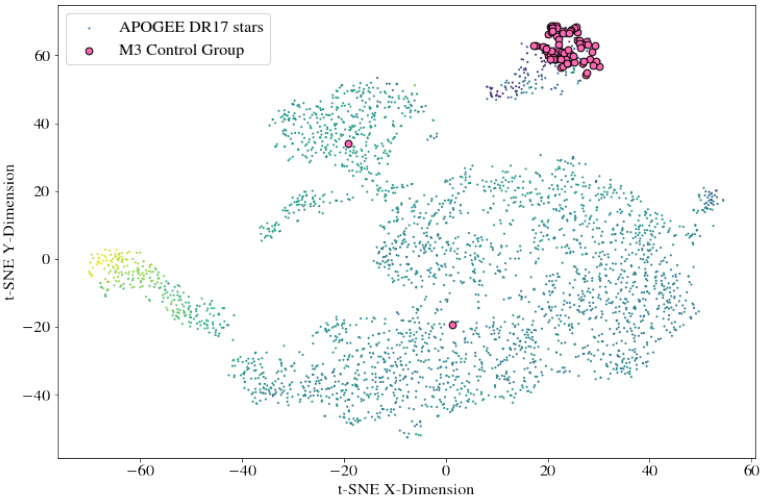


Figure 3. t-SNE projection of the same 3212 APOGEE DR17 stars in a 10° × 10° field of view around M3 using 19 chemical abundances and radial velocities. A perplexity value of 30 is used, with other hyperparameters set to default values. Stars from the M3 control group are labelled in pink. Field stars are coloured by their C/FE ratio, showing the variation along one axis of the original high-dimensional abundance data in the projection. Similar to UMAP, t-SNE is able to recover almost all the M3 control group stars in a cluster of points near the top right of the plot.

computes a test statistic as the maximum distance between a sample and a reference distribution at any x-value. As UMAP maps high-dimensional vectors into 2 dimensions, each point in the original data has a corresponding 2-D distribution in UMAP space under Monte Carlo realization, as shown in the right panel of Fig. 5. However, use of the K-S test (and other widely-used statistical tests for normality such as the Anderson-Darling test) is limited to univariate data, as it is not possible to uniquely order data across two or more axes to compute a maximum distance between two multivariate distribution functions. Therefore, we instead exploit the fact that a p-dimensional multivariate Gaussian-distributed vector $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ can be transformed into a univariate χ^2-distributed variable $Z \sim \chi^2(p)$ according to:

$$Z = (\mathbf{X} - \boldsymbol{\mu})^T \, \mathbf{C}^{-1} \, (\mathbf{X} - \boldsymbol{\mu}) \qquad (1)$$

where $\boldsymbol{\mu}$ is the $p \times 1$ sample mean and $\mathbf{C}$ is the $p \times p$ sample covariance. We transform the 2-dimensional UMAP distribution for a single point (as shown in the right panel of Fig. 5) into a 1-dimensional distribution according to equation 1. The transformed variable Z is then compared to a reference χ^2 distribution with 2 degrees of freedom using a one-sample K-S test, and the corresponding two-tailed p-value of the test is recorded. The p-values of this procedure performed on every point in the UMAP data from Fig. 5 are compared to the p-values for a test set of 2-dimensional Gaussian data in Fig. 6. Note that a uniform or slightly left-skewed distribution of p-values suggests that the data are mostly Gaussian-distributed in UMAP space; conversely, a right-skewed distribution (where most p-values are small) suggests that the data are generally non-Gaussian. As such, the results from Fig. 6 strongly suggest that the data are overwhelmingly non-Gaussian distributed in UMAP space, and thus the original high-dimensional Gaussian distribution in chemical abundance and radial velocity does not correspond to a Gaussian distribution in the UMAP projection.

4.3. *Results*

Our plots of UMAP and t-SNE for the original data in Figs. 2 and 3 show broad agreement in their global structure as well as along a test axis (C/FE ratio) of the data. Moreover, both methods are able to recover almost all the control group points, providing confidence that the chemical tagging process works on our sample

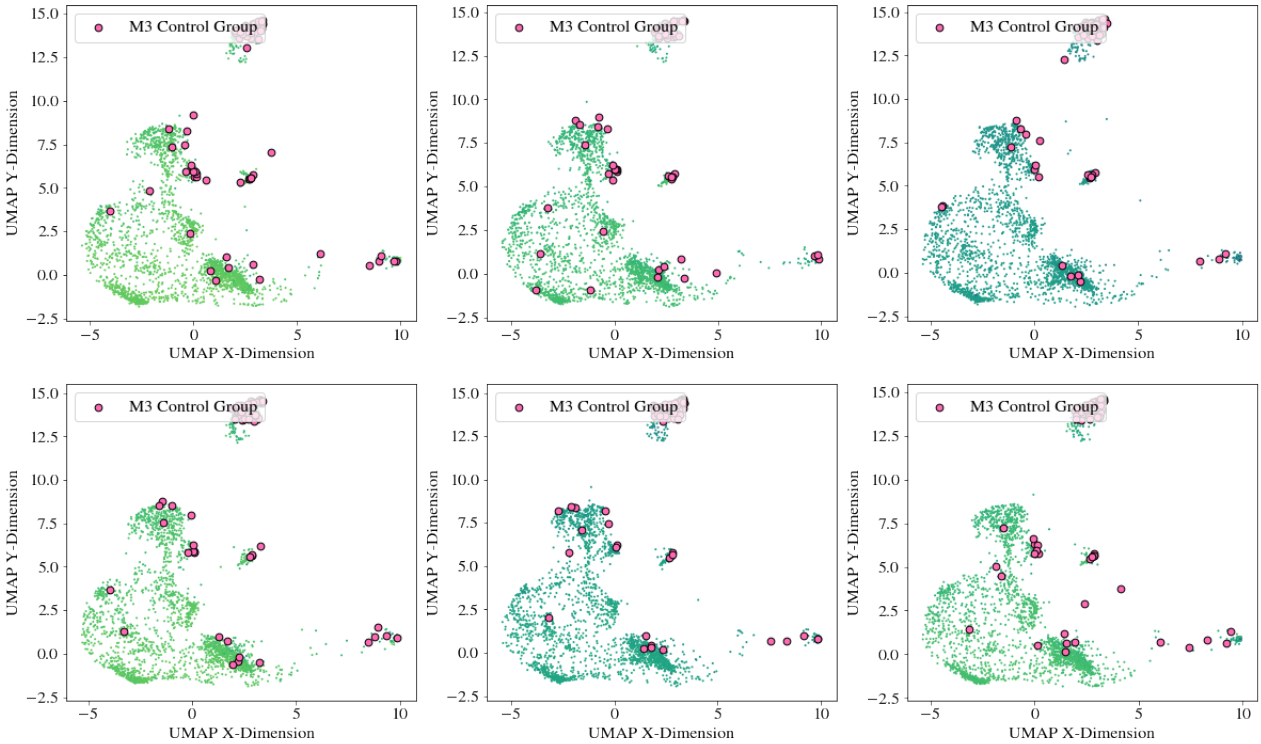


Figure 4. UMAP projections of 6 Monte Carlo samples of the data, with points from the M3 control group labelled in pink. The general structure of the projection, as well as the presence and shape of large-scale features, is largely consistent across the samples; for example, a long horizontal branch extending from the right, a large lobe extending upwards from the main group of stars, and the locations of several overdensities within the main group. Although UMAP still recovers most points from the M3 cluster under Monte Carlo realization, fewer points are recovered compared to the original data in every case.

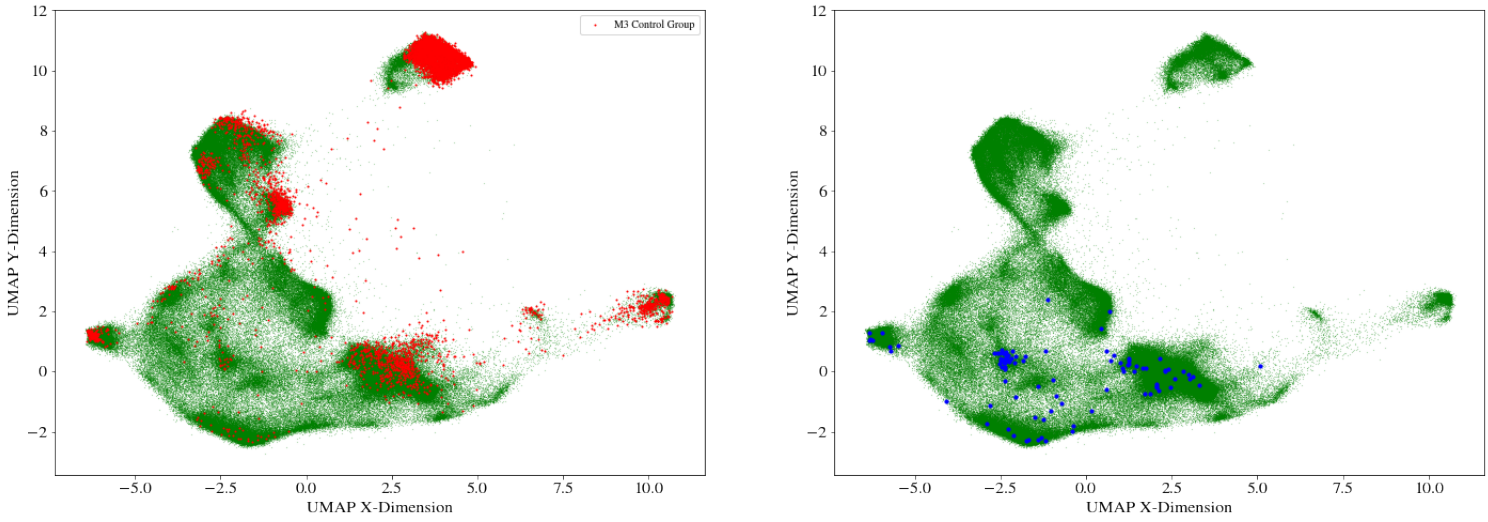
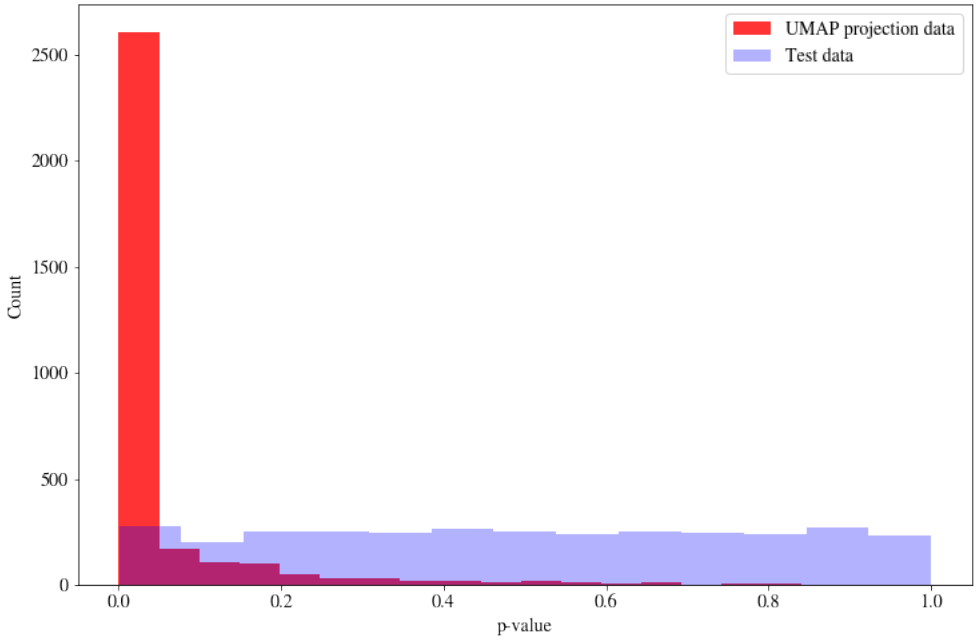


Figure 5. Aggregate of 100 Monte Carlo UMAP projections, showing the distributions of all M3 group points (left panel) and that of a random non-M3 data point (right panel) in UMAP space. UMAP recovers $\sim 64\%$ of the M3 points across all the Monte Carlo realizations, compared to $\sim 97\%$ for the original data. The M3 cluster is clearly visible in the aggregate plot in the left panel, while the remaining points are dispersed across the projection in a generally non-uniform manner, with areas near the edge of the projection showing a higher density of points. The distribution of a single non-M3 point in the right panel appears to be similarly abnormal.



Significance level (α)	**0.001**	**0.01**	**0.05**
Fraction of non-Gaussian points (Test data)	0.0012	0.0087	0.052
Fraction of non-Gaussian points (UMAP data)	0.52	0.68	0.82

Figure 6. Comparison of the p-value distribution obtained using the Kolmogorov-Smirnov (K-S) test for 100 Monte Carlo realizations of the UMAP data with a synthetic test set of known 2-dimensional Gaussian-distributed data. The p-values of the test data are uniformly distributed between 0 and 1, reflecting the fact that the data are actually sampled from a 2-dimensional Gaussian. The p-values of the UMAP data are highly right-skewed, providing strong evidence against our hypothesis that the distributions of the Monte Carlo sampled points in UMAP space are in general Gaussian. The fraction of non-Gaussian distributed points ($p \leq \alpha$) at various significance levels α for both the test and UMAP data are also recorded, showing that most of the UMAP points are not Gaussian distributed.

data. As noted in Section 2, our control group is selected from the filtered APOGEE data solely using its location in celestial coordinates (within $4 \times r_{hm}$ of M3's coordinates) without further processing. It is therefore possible that our control group inadvertently includes background or foreground stars aligned with M3, which could explain the two stars not recovered by UMAP or t-SNE in Figs. 2 and 3.

When simulating uncertainty in the data using Monte Carlo sampling, Figs. 4 and 5 show that UMAP maintains the structure of the data, though it consistently recovers fewer points compared to the original Fig. 2 projection. Our results from Fig. 6 show that the distribution of p-values for the UMAP data are strongly right-skewed, demonstrating that a large majority of points ($\sim$ 82% of points at a significance level $p \leq 0.05$) are non-Gaussian distributed in UMAP space. As UMAP and t-SNE are nonlinear methods, there is no *a priori* reason to believe that they should propagate uncertainty in a Gaussian manner; nevertheless, our results provide strong empirical confirmation that the low-dimensional UMAP embedding does not recover the original Gaussian uncertainty. In addition, the fraction of M3 points recovered is decreased from $\sim$ 97% in the original data to 64% under the Monte Carlo sampling in Fig. 5, demonstrating the challenges posed by uncertainty to chemical tagging analyses. Indeed, some previous authors have expressed doubt in the viability of using chemistry alone to associate stars with their birth clusters. For example, Ting et al. (2015) find that for a synthetic Milky Way disk data set, even dense groups in chemical abundance space are usually comprised of stars from several clusters, concluding that identification of individual clusters through chemical tagging is difficult. Casamiquela et al. (2021) test DBSCAN on known clusters in the APOGEE DR16 RC data (Ahumada et al. 2020) using the same `astroNN` abundances obtained by Leung & Bovy (2019) that we use in our analysis. Casamiquela et al. (2021)

are even more pessimistic about the ability of statistical clustering methods to perform chemical tagging, suggesting that the majority of groups ($> 70\%$) recovered by DBSCAN contain stars from multiple clusters.

However, Figs. 4 and 5 show that the unrecovered M3 points are not uniformly distributed throughout the embedding, but preferentially clustered near the edges of the projection. This suggests that while UMAP is unable to associate these points with the M3 cluster, it is at least able to recognize that their overall chemistry is distinct from the the background stars. Indeed, Casamiquela et al. (2021) observed a similar pattern in the clusters they recovered from the APOGEE DR16 data, noting that UMAP was able to consistently identify clusters located near the edge of the distribution. This motivates us to develop a new dimensionality reduction method capable of propagating uncertainty (i.e. that can estimate a low-dimensional distribution for each input point based on high-dimensional measurement error), which we outline in the following sections.

5. VARIATIONAL AUTOENCODERS

In section 4.3, we determined that the Gaussian error in the chemical abundance data is generally not propagated by UMAP. Here we develop a new dimensionality reduction method that is based on a variational autoencoder (VAE), a type of artificial neural network architecture. We demonstrate that our method is generally able to propagate Gaussian error, unlike UMAP, and that its performance on the APOGEE data compares favorably to the "naive" model that uses the default VAE architecture.

An autoencoder consists of two neural networks working together: an encoder $q_\phi(\mathbf{z}\|\mathbf{x})$ that maps a high-dimensional input vector $\mathbf{x} \in X$ to a lower-dimensional latent representation $\mathbf{z} \in Z$, and a decoder $p_\theta(\mathbf{x}\|\mathbf{z})$ that attempts to reconstruct the original input from the corresponding latent representation. The autoencoder is then trained to minimize a loss function $\mathcal{L}(\mathbf{x}, \mathbf{x}')$, which quantifies the reconstruction loss of the autoencoder. For the "naive" VAE model, a popular function for reconstruction loss (which we use as well) is the mean squared error between $\mathbf{x}'$ and $\mathbf{x}$:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{n}(\mathbf{x_i} - \mathbf{x_i'})^2$$

which is summed over all N observations.

A *variational* autoencoder alters this structure by having the encoder map every input vector $\mathbf{x}$ to a distribution $\mathcal{N}(\mathbf{z}, \sigma_\mathbf{z})$ in latent space (rather than a point), with the parameters $\mathbf{z}$ and $\sigma_\mathbf{z}$ learned by the encoder. The decoder then samples the latent vector $\mathbf{z}$ from the corresponding distribution when reconstructing the original input. The loss function is then modified to include an additional structure penalty $D_{KL}(P||Q)$, which is the Kullback-Leibler (KL) divergence between the latent distribution $P = \mathcal{N}(\mathbf{z}, \sigma_\mathbf{z})$ and the unit Gaussian $Q = \mathcal{N}(0, 1)$:

$$D_{KL}(P||Q) = \frac{1}{2}\left(\mathbf{z}^2 + \sigma_\mathbf{z}^\mathbf{2} - \ln\sigma_\mathbf{z}^\mathbf{2} - 1\right) \tag{2}$$

This structure penalty helps the network estimate the shape of an underlying distribution that produces the original input data. The following subsection outlines the modifications we make to the VAE architecture for our model.

5.1. *Method*

The chemical abundance measurements and errors $\mathbf{x}, \sigma_\mathbf{x}$ are input to the VAE, and two encoders for the mean and variance of the corresponding latent distribution are trained separately. Note that in our implementation of the model, we parameterize using log-variance rather than variance so that the model does not learn negative values for variance. The first encoder $E_{\phi 1} : \mathbf{x} \rightarrow \mathbf{z}$ maps the input measurement $\mathbf{x}$ to a point $\mathbf{z}$ in latent space, representing the mean of its corresponding latent distribution. The second encoder $E_{\phi 2} : (\mathbf{x}, \sigma_\mathbf{x}) \rightarrow \sigma_\mathbf{z}$ maps the input measurement as well as the error σ_x to the variance of the latent distribution, σ_z. By separating the mean and variance in this manner, we ensure that the latent space vector $\sigma_\mathbf{z}$ can be directly interpreted as a lower-dimensional encoding of the information contained in the original measurements (i.e. errors in a given measurement are propagated through to the latent space, but do not affect their position in latent space). In standard VAE usage, the sampling step is meant to estimate an unknown distribution that underlies the original observation $\mathbf{x}$. However, in our procedure, we already know the distribution of the original observation and are instead interested in propagating it into (and reconstructing it from) the latent space. Thus, we eliminate the step of sampling from the latent space distribution entirely, and instead directly input the latent distributions (parameterized using $\mathbf{z}, \sigma_\mathbf{z}$ as $\mathcal{N}(\mathbf{z}, \sigma_\mathbf{z})$) into the decoder. The decoder architecture therefore follows the same structure as the encoder; two decoders for the mean and variance of the reconstructed abundances are trained separately (again using log-variance). The first decoder $D_{\phi 1} : \mathbf{z} \rightarrow \mathbf{x}'$ maps the latent vector $\mathbf{z}$ to the reconstructed measurement $\mathbf{x}'$, while the second decoder $D_{\phi 2} : (\mathbf{z}, \sigma_\mathbf{z}) \rightarrow \sigma'_\mathbf{x}$ maps the latent vector and error σ_z to the reconstructed error $\sigma'_\mathbf{x}$. The reconstruction penalty for each abundance in

an observation is then computed as the KL divergence between the original distribution $X = \mathcal{N}(\mathbf{x}, \sigma_{\mathbf{x}})$ and the reconstructed distribution $Y = \mathcal{N}(\mathbf{x}', \sigma'_{\mathbf{x}})$, which for two multivariate Gaussians is computed as:

$$D_{KL}(X||Y) = \frac{1}{2}\left(\ln\left(\frac{\sigma_{\mathbf{x}}^{\mathbf{2}}}{\sigma_{\mathbf{x}}'^{\mathbf{2}}}\right) + \frac{\sigma_{\mathbf{x}}'^{\mathbf{2}} + (\mathbf{x}' - \mathbf{x})^2}{\sigma_{\mathbf{x}}^{\mathbf{2}}} - 1\right)$$

The structure penalty for each abundance in an observation is computed as described in Eq. 2; it is the KL divergence between the standard normal distribution $Q = \mathcal{N}(0, 1)$ and the latent distribution $P = \mathcal{N}(\mathbf{z}, \sigma_{\mathbf{z}})$. The total loss function $\mathcal{L}$ to optimize over is therefore the sum of the reconstruction and structure penalties, summed over all N observations and averaged over all M chemical abundances:

$$\mathcal{L} = \frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\left(D_{KL}(X_{ij}||Y_{ij}) + D_{KL}(P_{ij}||Q_{ij})\right)$$

Unlike UMAP/t-SNE, our method therefore directly reproduces the original Gaussian error in the latent space by construction; both the encoder (dimensionality reduction) and decoder (reconstruction) work directly with distributions rather than points. The latent space and reconstructed distributions are also assumed to be Gaussian by the model, as both the structure and reconstruction penalties are computed as the KL divergence between two Gaussian distributions. The Gaussian error in the original data is therefore propagated through every step of the process.

5.2. *Naive Model*

We begin by first training the "naive" VAE using the original architecture described in section 5, without making the modifications we describe in section 5.1. This is to confirm that the VAE architecture can be used for dimensionality reduction in chemical tagging analysis, and to provide a comparison for our modified variant.

Our naive model uses the full set of filtered stars from the APOGEE DR17 database, divided into training and test sets as described in section 2 The neural network was implemented using the `TensorFlow` (Abadi et al. 2015) package in Python, using a 2-dimensional latent space similar to the UMAP projection. The encoders and decoders are each composed of 4 dense fully-connected layers with 100 neurons, and a final output layer that uses 2 (for the encoder) or 19 (for the decoder) neurons, corresponding to the number of dimensions in the output. The dense layers use a Gaussian Error Linear Unit (GeLU) activation function, and the final output layers use a linear activation function. The neural network is trained for 100 epochs using batch gradient descent with the `Adam` optimizer. We set a learning rate of $\alpha = 10^{-7}$ and a batch size of 256, to improve stability when training.

Fig. 7 shows the original metallicities (Fe/H) compared to the reconstructed values given by our naive VAE, as well as the residuals. We see that our naive VAE is able to reconstruct the original metallicity measurements with good accuracy, although the reconstruction compresses the metallicities into a smaller range than the original values, particularly for very metal-poor or metal-rich stars. This is reflected in the residual plot, which shows some correlation with the original values for very high or low values of metallicity but are otherwise unbiased and normally distributed. Fig. 8 shows the full results for the other 18 chemical abundances, and displays the same loss of information observed with the metallicity, in which the reconstructed abundance values are compressed into a smaller range than the original. Nevertheless, the naive model is quite capable at recovering the original abundance measurements, with the residuals of each abundance being quite robust. The results in Figs. 7 and 8 will serve as a baseline comparison for our modified VAE model in section 5.3.

5.3. *Modified VAE Model*

We train our proposed model in section 5.1 and evaluate its ability to perform dimensionality reduction and to reconstruct the original data. The modified VAE is trained on the same dataset as the naive VAE and uses 2 encoders and decoders as described in section 5.1, each with the same internal structure as the naive VAE. As this model takes into account the measurement errors in the observations, we use a 2-dimensional latent space for both the means and variances (4 latent variables in total), similar to the UMAP projection. We also train the model using the same hyperparameters that we used for the naive model.

Fig. 9 shows the original metallicities compared to the reconstructed values and measurement errors from our model (with the residuals displayed for both as well). The modified VAE is able to reconstruct the original measurements with much greater accuracy than the naive model, with the standard deviation of residuals being more than an order of magnitude lower (~ 0.15 for the naive model compared to ~ 0.011 for the modified VAE). In addition, the information loss of the modified VAE along the metallicity axis is much smaller than for the naive one, as the reconstructed metallicities have a wider range. This is also visible from the residuals, which are more narrowly distributed around 0 than for the naive model and appear largely uncorrelated with

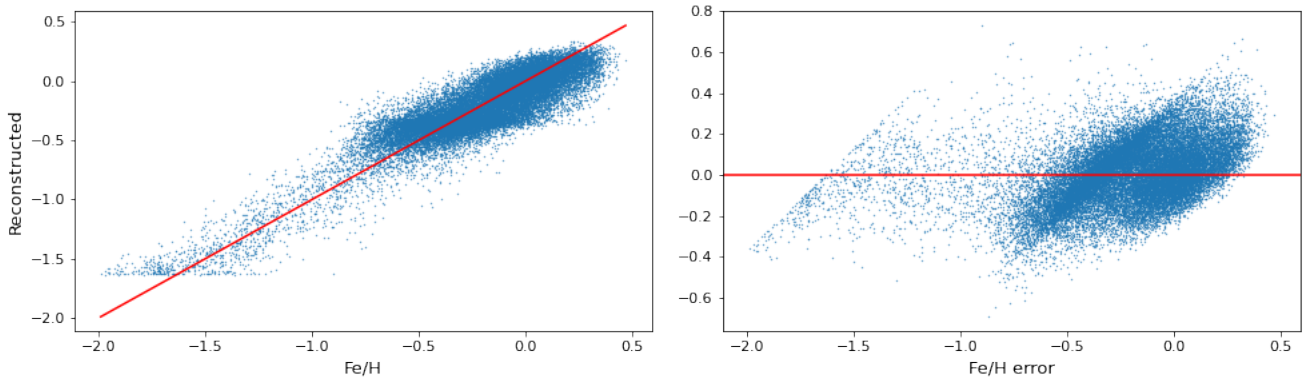


Figure 7. Left: Plot of the observed Fe/H ratios $\mathbf{x}$ (metallicity) on the x-axis against the reconstructed values $\mathbf{x}'$ on the y-axis, obtained using our naive VAE model. Both axes are in units of $\log_{10}$ metallicity. The red line $y = x$ represents perfect reconstruction of the original values by the model, with deviation from the line representing some amount of reconstruction error for that measurement. The reconstructed metallicities are compressed into a smaller range than the original values, with some information being lost. Right: Plot of residuals for the metallicities shown on the left. The residuals for extreme high or low values of metallicity are correlated with the measurements (i.e. negative residuals for very low metallicity and positive for high metallicity), reflecting how the model compresses the range of reconstructed metallicities. However, the residuals are otherwise unbiased and normally distributed about 0, suggesting our model is quite robust.

the measurements, unbiased and normally distributed, suggesting our model is quite robust. Moreover, the naive VAE does not take into account the measurement errors at all, while our modified VAE reconstructs the measurement errors quite accurately overall, lending confidence in our model's ability to propagate the original Gaussian distributions to the latent space and reconstruct them while preserving most of the information.

The full results for the other 18 chemical abundances and their corresponding measurement errors are shown in Figs. 10 and 11, respectively. Fig. 10 demonstrates that our model is able to reconstruct the abundance measurements more accurately than the naive VAE, with residuals generally being more narrowly distributed around 0 for most abundances. Although the same information loss observed in the naive VAE (the reconstructed abundances being compressed into a smaller range) is still present, it is much less severe overall. The residuals also appear more robust for most of the abundances as well, underlying the improvements yielded by our modified VAE compared to the naive model. Similarly, the plots in Fig. 11 show our modified VAE model also reconstructs the rest of the abundance errors accurately, suggesting that the Gaussian uncertainties in the other abundances can be propagated to a lower dimension and then reconstructed. Indeed, the overall performance of the model on the test data is very promising and suggests that it could be used to propagate Gaussian uncertainty for dimensionality reduction.

6. DISCUSSION

The results of Fig. 10 demonstrate that our modified VAE model produces a more accurate reconstruction for all the abundances compared to the naive VAE shown in Fig. 8, providing confidence in our overall approach. Note that the plots in both figures have the same axes.

As noted in the previous sections, most of the reconstructed chemical abundances for both the naive and modified VAE (as well as the measurement errors for the latter) have their range compressed, due to some information being lost during the encoding and decoding process. As each measurement is weighted the same amount by the reconstruction loss, we expect the model to generally emphasize more accurate reconstruction along the axes with the largest range, analogous to other dimensionality reduction methods like principal component analysis. Figs. 9 and 10 demonstrate this phenomenon; the most accurate reconstructed abundance is the metallicity, since it has the largest range of any of the 19 parameters (from roughly -2.0 to 0.5). While some information is always lost when projecting to a lower dimension (regardless of the method), Figs. 10 and 11 demonstrate that our modified VAE model can still encode the data in a way where the VAE can reconstruct the original abundances and measurement errors while preserving most of the information contained in the original data. This is crucial, since (as noted in section 5.1) our modified VAE works directly with Gaussian distributions during the entire process, and therefore explicitly propagates Gaussian error during the encoding and decoding. In contrast, methods like t-SNE and UMAP

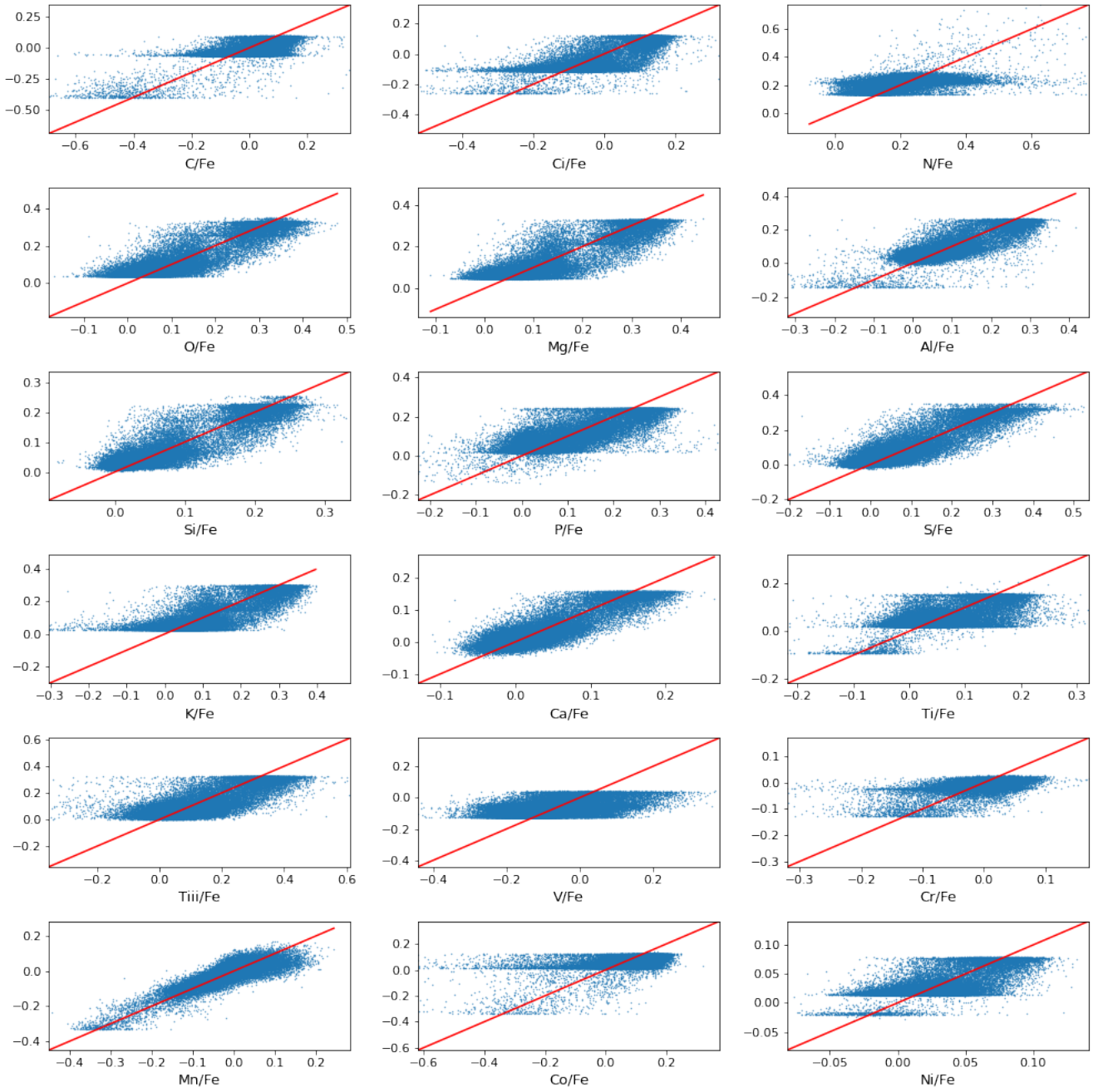


Figure 8. Measurements of the other 18 chemical abundance ratios plotted against the reconstructed values obtained from the naive VAE model. All axes are in units of $\log_{10}$ abundance. As with metallicity, the VAE similarly compresses the reconstructed chemical abundances into a smaller range than the original values, with this loss of information being more severe for some abundances than others. Nevertheless, the naive model is able to recover most of the information in the original abundance measurements, and the residuals for the abundances generally appear unbiased and normally distributed.

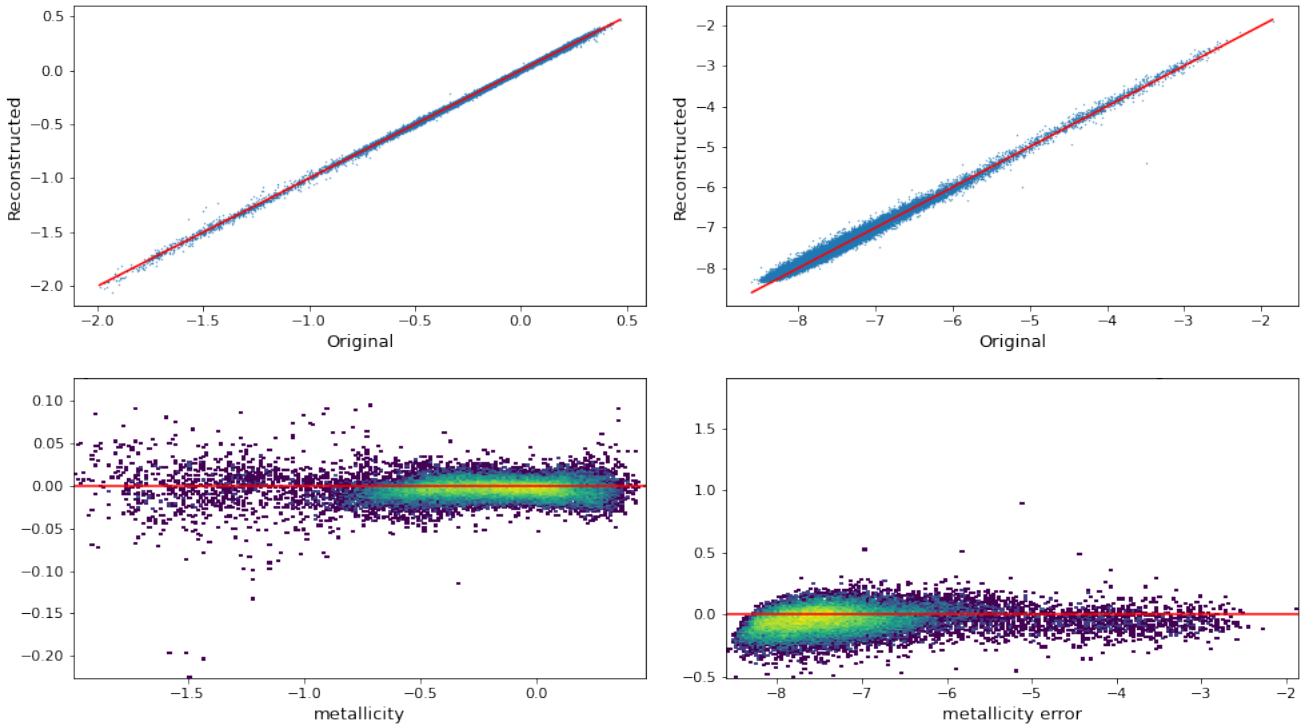


Figure 9. Top left: Observed metallicities are plotted against the reconstructed values, with both axes in units of $\log_{10}$ and with the same scale as in Fig. 7. Compared to Fig. 7, our modified VAE model produces a significantly more accurate reconstruction with much less information loss, as this model is able to reconstruct the entire metallicity range. Top right: The associated measurement errors of the metallicities are similarly plotted against their reconstructions. Our modified VAE reconstructs the errors quite accurately, suggesting that it can propagate the original Gaussian distributions to the latent space and reconstruct them while preserving most information. Bottom: The residuals of the observed metallicities (left) and their associated measurement errors (right) are plotted as 2-D histograms, with denser areas represented using lighter shades. The metallicity reconstruction is much more accurate than for the naive model, with residuals distributed more tightly around 0 (note the relative scale of the axes in both figures). Both sets of residuals also appear uncorrelated, unbiased and normally distributed, suggesting our method is quite robust.

work with data points throughout and do not take into account measurement uncertainty at any step.

Our work therefore suggests a more robust method for propagating Gaussian measurement uncertainty in chemical tagging analysis, unlike widely-used existing dimensionality reduction methods like UMAP and t-SNE.

7. CONCLUSIONS

In this paper, we analyzed chemical abundance and radial velocity data from the APOGEE DR17 survey to determine how measurement uncertainty in high-dimensional data, simulated using Monte Carlo sampling, affects both the global structure of UMAP and its ability to recover a known control group of stars, the M3 globular cluster, as shown in Fig. 2. The UMAP results for 6 Monte Carlo realizations drawn from a multivariate Gaussian, as seen in Fig. 4, qualitatively demonstrate that while global structure and large-scale features in the embedding are stable under Monte Carlo sampling, recovery of the M3 control group is adversely affected. Computing the UMAP projections of 100 Monte Carlo samples in Fig. 5, we estimate that $\sim 64\%$ of M3 points are recovered by UMAP under Monte Carlo sampling, compared to $\sim 97\%$ in the original projection. Moreover, the unrecovered stars are irregularly distributed throughout the projection, suggesting that individual points are delocalized in UMAP space.

We then analyzed whether the Monte Carlo distribution of individual points in UMAP space is in general Gaussian, to determine whether UMAP propagates Gaussian measurement error from the original high-dimensional data to its lower-dimensional embedding. We transformed the distribution of each point in the data (generated by the 100 Monte Carlo samples) according to Equation 1, which converts a p-dimensional multivariate Gaussian distribution into a univariate χ^2 distribution with p degrees of freedom. The transformed

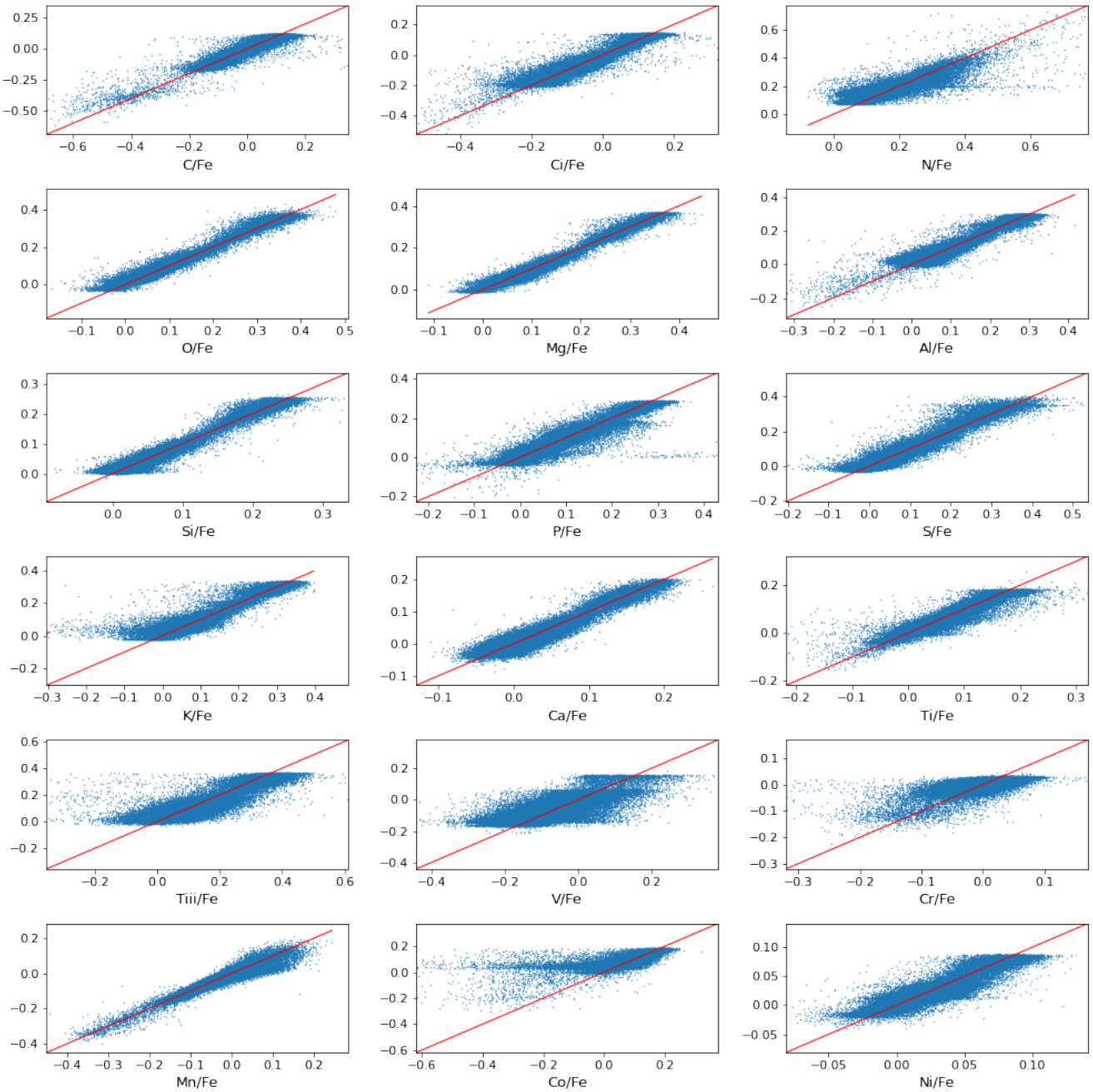


Figure 10. The other 18 chemical abundance ratios are plotted against the reconstructed values obtained with the modified VAE model, with the same axes as used in Fig. 8. As with metallicity, our modified VAE model produces a more accurate reconstruction of the other chemical abundances compared to the naive model, with residuals distributed more tightly around 0. While the same loss of information observed in the naive VAE (where the reconstructed abundances are compressed into a smaller range) is still present, it is much less severe overall.

distribution for every point was then compared to a reference χ^2 distribution using a one-sample Kolmogorov-Smirnov test (corresponding to comparison of the original distribution to a Gaussian) and the test p-values were recorded. We then contrasted the distribution of p-values for the Monte Carlo UMAP data with a synthetic test set of 2-dimensional Gaussian data. Our results in Fig. 6 provide strong evidence that points in UMAP space are generally non-Gaussian distributed, and therefore UMAP does not propagate Gaussian error.

We then proposed a new dimensionality reduction method based on a modification of the variational autoencoder (VAE) neural network architecture. Unlike UMAP, our modified VAE works entirely with distributions (rather than points) throughout the dimensionality reduction process, and thus directly propagates Gaussian error to a lower-dimensional latent space by construction. Using the APOGEE DR17 data set, Figs. 10 and 11 show that our model can recover most of the information in the original Gaussian distributions, and that the reconstruction is more accurate than a naive model that uses the default VAE architecture. Indeed, the results strongly suggest that our method can robustly propagate Gaussian error, unlike other dimensionality reduction methods like t-SNE/UMAP or the naive VAE model. This would in turn allow for probabilistic identification of cluster associations in chemical tagging analysis, which has not been done previously.

7.1. *Next Steps*

In the future, we plan to apply this method to chemical tagging studies of other astronomical data sets, to probabilistically determine cluster associations of stars in a way that existing methods like UMAP do not allow. We could also consider further modifying the VAE architecture used in our model. One possible change could be to completely decouple the encoder/decoder tracks for measurements and errors, so they would be independently encoded and decoded from the latent space. This would allow direct interpretation of the latent space errors in the same way as the latent vectors. We could also consider adjusting hyperparameters like the internal structure of the encoders/decoders or the batch size for gradient descent to improve model performance, as both were chosen arbitrarily. Finally, as our model currently takes in chemical abundances as input, we hope to generalize it so it can be applied directly to stellar spectra without requiring the preprocessing step of converting the spectra to chemical abundances first.

Software: `UMAP` (McInnes et al. 2018), `scikit-learn` (Pedregosa et al. 2011), `TensorFlow` (Abadi et al. 2015)

REFERENCES

Abadi, M. et al. 2015, TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abdurro'uf et al. 2022, ApJS, 259, 35

Ahumada, R. et al. 2020, ApJS, 249, 3

Anders, F., Chiappini, C., Santiago, B. X., et al. 2018, A&A, 619, A125

Baumgardt, H., & Hilker, M. 2018, MNRAS, 478(2), 1520-1557

Bovy, J., Hogg, D.W., & Roweis, S.T. 2011, Annals of Applied Statistics, 5(2B), 1657-1677

Casamiquela, L., Castro-Ginard, A., Anders, F., & Soubiran, C. 2021, A&A, 654, A151

Chun, S.-H., Lee, J.-J., & Lim, D. 2020, ApJ, 900, 146

Gaia Collaboration (Brown, A.G. A., et al.) 2021, A&A, 649, A1

Grondin, S. M., Webb, J. J., Leigh, N. W. C., et al. 2023, MNRAS, 518(3), 4249-4264

Krumholtz, M. R., Bate, M. R., Arce, H. G., et al. 2014, Protostars and Planets VI, eds. H. Beuther, R. S. Klessen, C. P. Dullemond & T. Henning, Tucson, AZ: University of Arizona Press, 243-266

Krumholtz, M. R., McKee, C. F., Bland-Hawthorn, J. 2019, ARA&A, 57, 227 - 303

Leung, H. W., & Bovy, J. 2019, MNRAS, 483(3), 3255-3277

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94

Martell, S. L., Shetrone, M. D., Lucatello, S., et al. 2016, ApJ, 825, 146

McInnes, L., Healy, J., & Melville, J. 2018, arXiv e-prints, arXiv:1802.03426

Navin, C. A., Martell, S. L., Zucker, D. B. 2016, ApJ, 829, 123

Pedregosa, F., et al. 2011, Journal of Machine Learning Research, 12(85), 2825-2830

Price-Jones, N., & Bovy, J. 2019, MNRAS, 487(1), 871-886

Schiavon, R. P., Zamora, O., Carrera, R., et al. 2017, MNRAS, 465(1), 501-524

Ting. Y.-S., Conroy, C., & Goodman, A. 2015, ApJ, 807, 104

van der Maaten, L., Hinton, G. 2008, Journal of Machine Learning Research, 9(86), 2579-2605

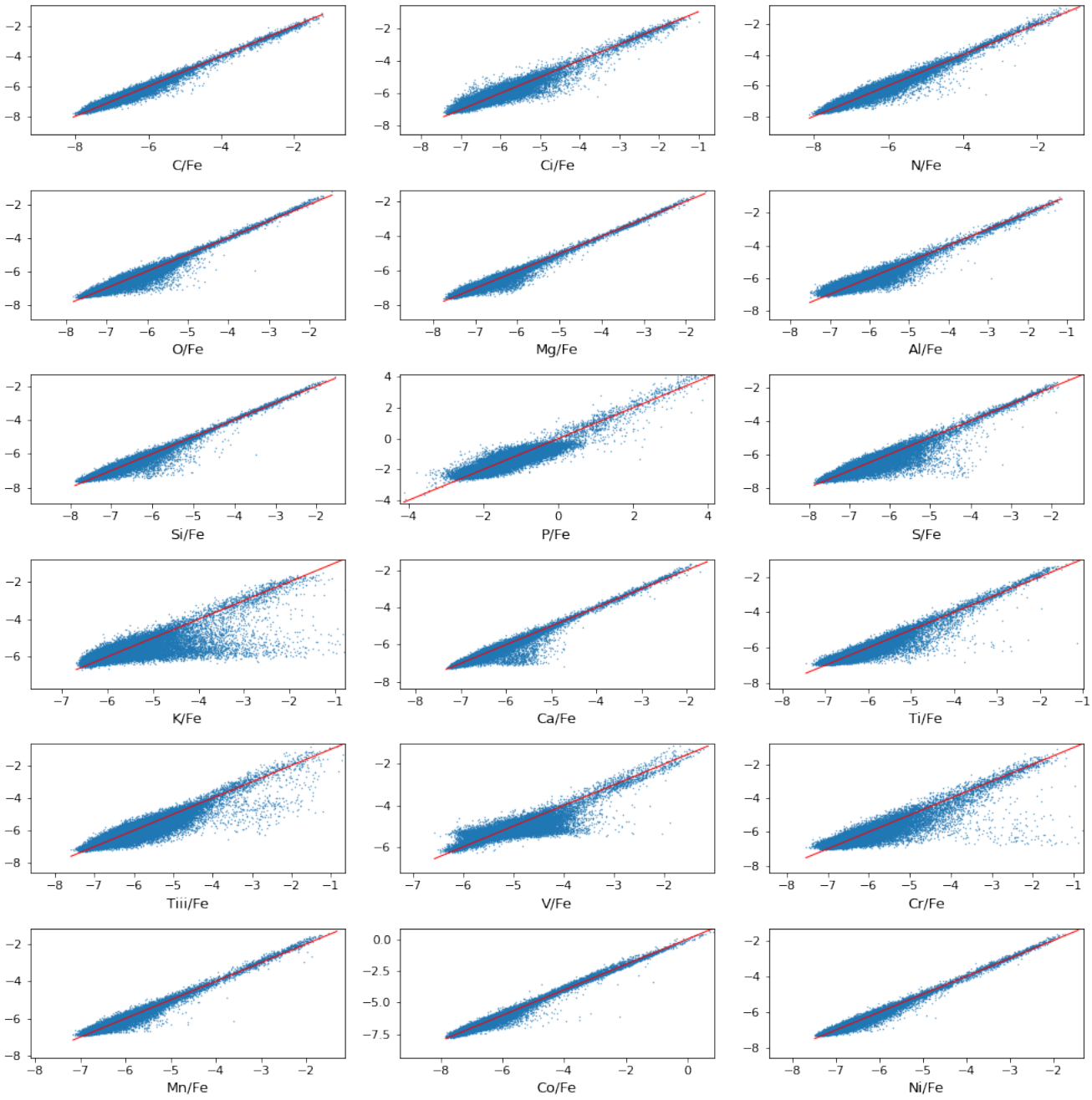


Figure 11. Measurement errors for the other 18 abundances are plotted against the reconstructed errors obtained with our modified VAE. While our model is able to reconstruct the abundance errors well overall, some of the residual plots appear biased, with the reconstructed errors being lower than the original values. This bias is particularly visible in the K/Fe and Tiii/Fe plots. Nevertheless, the model is capable of reconstructing the original Gaussian distributions for most of the abundances (unlike UMAP), and the residuals for most of the abundances are still robust.