SSSpaNG! stellar spectra as sparse, data-driven, non-Gaussian processes

Stephen M. Feeney, Benjamin D. Wandelt, Melissa K. Ness

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **501,** 3258–3271 (2021) doi:10.1093/mnras/staa3586
Advance Access publication 2020 November 18

SSSpaNG! stellar spectra as sparse, data-driven, non-Gaussian processes

Stephen M. Feeney [1,2]★ Benjamin D. Wandelt[1,3,4] and Melissa K. Ness[1,5]

[1]*Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA*
[2]*Department of Physics and Astronomy, University College London, London WC1E 6BT, UK*
[3]*Sorbonne Université, CNRS, UMR 7095, Institut d'Astrophysique de Paris (IAP), 98 bis boulevard Arago, F-75014 Paris, France*
[4]*Sorbonne Université, Institut Lagrange de Paris (ILP), 98 bis boulevard Arago, F-75014 Paris, France*
[5]*Department of Astronomy, Columbia University, Pupin Physics Laboratories, New York, NY 10027, USA*

Accepted 2020 November 12. Received 2020 November 6; in original form 2020 February 10

ABSTRACT

Upcoming million-star spectroscopic surveys have the potential to revolutionize our view of the formation and chemical evolution of the Milky Way. Realizing this potential requires automated approaches to optimize estimates of stellar properties, such as chemical element abundances, from the spectra. The sheer volume and quality of the observations strongly motivate that these approaches should be driven by the data. With this in mind, we introduce SSSpaNG: a data-driven non-Gaussian Process model of stellar spectra. We demonstrate the capabilities of SSSpaNG using a sample of APOGEE red clump stars, whose model parameters we infer using Gibbs sampling. By pooling information between stars to infer their covariance, we permit clear identification of the correlations between spectral pixels. Harnessing this correlation structure, we infer the true spectrum of each red clump star, inpainting missing regions and denoising by a factor of at least two for stars with signal-to-noise ratios of $\sim$20. As we marginalize over the covariance matrix of the spectra, the effective prior on these true spectra is non-Gaussian and sparsifying, favouring typically small but occasionally large excursions from the mean. The high-fidelity inferred spectra produced with our approach will enable improved chemical elemental abundance estimates for individual stars. Our model also allows us to quantify the information gained by observing portions of a star's spectrum, and thereby define the most mutually informative spectral regions. Using 25 windows centred on elemental absorption lines, we demonstrate that the iron-peak and alpha-process elements are particularly mutually informative for these spectra and that the majority of information about a target window is contained in the 10-or-so most informative windows. Such mutual information estimates have the potential to inform models of nucleosynthetic yields and the design of future observations. Our code is made publicly available at https://github.com/sfeeney/ddspectra.

Key words: methods: statistical – stars: abundances – stars: statistics.

1 INTRODUCTION

Surveys such as APOGEE (Majewski et al. 2017), GALAH (De Silva et al. 2015), Gaia-ESO (Gilmore et al. 2012), LAMOST (Newberg et al. 2012), SEGUE (Yanny et al. 2009), and RAVE (Steinmetz et al. 2006) have provided a vast data set of spectroscopic observations that has revolutionized our view of the Milky Way, through corresponding velocity, stellar parameter, individual abundance, and age measurements (e.g. Nidever et al. 2014; Minchev et al. 2014b; Hayden et al. 2015; Kordopatis et al. 2015; Ho et al. 2017b; Frankel et al. 2018; Bland-Hawthorn et al. 2019; Bovy et al. 2019; Mackereth et al. 2019). In the coming years, large spectroscopic surveys such as Sloan V (Kollmeier et al. 2017), WEAVE (Bonifacio et al. 2016), 4MOST (de Jong et al. 2016), PFS (Tamura et al. 2016), *Gaia* RVS (Gaia Collaboration et al. 2016), and MOONS (Cirasuolo et al. 2014) will begin observations, expanding the spectral data we have collected for our Galaxy by orders of magnitude. At present, the large ($>10^5$ star) medium-resolution surveys, such as APOGEE ($R = 22500$), rely on expensive observations, integrating to signal-to-noise ratios (SNRs) of up to 100 per pixel (Zasowski et al. 2013, 2017).

High-SNR spectra have been often regarded as necessary in the pursuit of precision abundances, required for chemical differentiation across the Galaxy. These abundances trace the detailed chemical evolution of the Milky Way, which is driven by an ensemble of stellar explosion and mass-loss activity. In the Galactic disc, where the majority of the stellar-mass resides, abundances provide (Rix & Bovy 2013; Bland-Hawthorn & Gerhard 2016) the record of its inside–out formation over time. The earliest epoch of the Galaxy's formation and its continued interaction with its environment is documented in the chemical composition and characteristics of the stellar halo (e.g. Hawkins et al. 2015; Helmi et al. 2018; Das, Hawkins & Jofre 2019). Current data place strong constraints on the chemical evolution models designed to explain Galactic formation and evolution (e.g. Minchev, Chiappini & Martig 2013, 2014a; Sanderson et al. 2018; Blancato et al. 2019; Clarke et al. 2019; Weinberg et al. 2019). Upcoming data offer the opportunity to refine these models considerably: for example, the disc is also believed to comprise numerous individual birth sites where groups of stars were born. Any prospect of assigning stars to their birth sites via

★ E-mail: stephen.feeney@ucl.ac.uk

their unique chemical signatures (e.g. Bland-Hawthorn, Krumholz & Freeman 2010) requires large stellar numbers and high precision abundance measurements (Mitschang et al. 2013; Ting, Conroy & Goodman 2015; Hogg et al. 2016; Armillotta, Krumholz & Fujimoto 2018).

The large data volumes now in hand have led to the development of new approaches for deriving abundance measurements from spectral data, driven by the need for automatic, efficient means of extracting the full information content of the data. These include data-driven modelling approaches such as The Cannon (Ness et al. 2015), full spectral fitting using physical models as implemented in The Payne (Ting et al. 2018) and deep learning (Leung & Bovy 2019). These approaches improve the precision of abundance measurements significantly, permitting useful abundances to be estimated using 1/4 to 1/9th of the observing time compared to previous approaches. Specifically, abundance precisions on the order of 0.05–0.1 dex can be achieved at SNR ≈ 40 per pixel (Ho et al. 2017a; Ness 2018; Ting et al. 2018; Leung & Bovy 2019). It has also been demonstrated that an ensemble of individual abundances can be derived at medium ($R = 11000$) and low ($R = 1800$) resolution by full spectral modelling (e.g. Casey et al. 2016; Ting et al. 2017; Wheeler et al. 2020). Physically, this is well justified: abundances can be measured from their impact on the entire spectral range as legitimately as from individual elemental lines (e.g. Ting et al. 2018). This methodological advance in particular is relevant for the *Gaia* RVS data ($R = 11000$ spectra for 7 million objects) and, furthermore, the large ensemble of low-resolution data being observed in future surveys. The dramatic and rapid increase in available spectra and availability of increasingly powerful computational resources means we find ourselves in an era of tremendous opportunity for developing new avenues of stellar spectral modelling.

Central to the success of The Cannon and The (Data-Driven) Payne (Xiang et al. 2019) is pooling: sharing information between members of a population to improve our knowledge of individual stars. In The Cannon, pooling is performed in a data-driven fashion by learning the relationship between stellar spectra and individual stellar abundances; in The Payne (during the training step) by calibrating physical models of stellar spectra using labels derived therefrom. In this paper, we seek to generate a data-driven model of the stellar spectra themselves, as opposed to the abundance measurements, formalizing this concept of pooling within a Bayesian hierarchical model (Gelman et al. 2013). By sharing information between stars, we will generate more precise representations of individual spectra, directly infer the correlation structure between spectral pixels and, in the process, gain understanding of the information content of the data. To date, there has been little work on the characterization and interpretation of the correlations between (and the dimensionality of) spectral data (see however Ting et al. 2012; Mitschang et al. 2014; Casey et al. 2019; Price-Jones & Bovy 2019). Our methodology will provide a direct measure of the information content of spectral regions and, correspondingly, elemental abundances.

We use stars observed by the APOGEE survey to build an extremely general and flexible empirical model of a large set of spectral data. Specifically, we implement a Gaussian Process (Rasmussen & Williams 2006) mixture model representation of the APOGEE red clump stars. Unlike typical Gaussian Process analyses, we infer each element of our covariance matrices directly, without assuming a kernel function, and marginalize over the covariances when quoting our inferred true spectra. As a result, and contrary to analyses in which the covariance is fit once and fixed, the prior distributions of our true spectra are highly non-Gaussian, with a sparsifying prior whose negative logarithm is non-convex. Our method is a significant new technical advance in the modelling of stellar spectra and is distinct from, but builds upon, existing progress in data-driven spectral modelling in the regime of large data sets. We use no physical knowledge in constructing our model or selecting priors, and our inference is therefore entirely driven by common trends in the high-dimensional APOGEE data. In successfully pooling information about stars, we achieve the following for the APOGEE spectra:

(i) Prediction of masked (unmeasured or contaminated) regions of the spectra to enable, e.g. abundance estimates that would otherwise be impossible (see Sections 4.1 and 4.2). This is particularly valuable in APOGEE for neutron-capture elements such as Nd and Ce, for which only a handful of weak features exist from which to estimate abundances. Some of these elemental features may fall near one of three chip gaps and therefore be absent in some (but, critically, not all) spectra due to stellar velocities. Our modelling of the data can predict these regions when they are absent.

(ii) Denoising of all spectra, enabling higher precision inference at lower SNR (see Section 4.2). The utility of this feature depends on the size of the effects we wish to discover. Our expectation is that this is particularly useful for weak features on the limits of detection, similar to previous demonstrations using generative modelling (e.g. The Cannon and The Payne).

(iii) Detailed examination of the empirical correlations in the spectra, quantitative measurements of these correlations and identification of which elemental absorption lines are positively and negatively correlated (see Section 4.2).

(iv) Quantification of the information content of the data and determination of the most informative regions of spectra (see Section 4.3). This has consequences for both theory and experimental design. Along with the correlation structure we infer, the information content that we measure should place strong constraints on physical models of nucleosynthesis and chemical evolution. From an experimental design perspective, quantifying the informativeness of regions of spectra can drive the selection of wavelength ranges optimized for specific scientific purposes, answering questions such as whether we can retain sensitivity to abundances by observing a reduced spectral range, or conversely whether we gain significant information on a range of elemental features by observing a particular set of wavelengths.

In the following, we describe the APOGEE data we use in Section 2 and our model and its inference in Section 3. We present our results in Section 4 and discuss their consequences, current limitations and plans for their resolution in Section 5.

2 DATA

For our modelling, we use the APOGEE red clump spectra from DR14 (Bovy et al. 2014; Majewski et al. 2017). These spectra comprise 29502 stars with a mean SNR of 210 and range of SNR of 21–1775. The contamination of red giant branch stars within this sample is of the order of 5–10 per cent (Bovy et al. 2014). While our approach is applicable to any stellar population, we select a largely homogeneous population for this proof of concept, restricting to the narrow temperature and gravity range of the red clump stars. Doing so should reduce the number of components required for our Gaussian Process mixture model, simplifying its inference.

The data have been downloaded from the APOGEE data base having already been shifted in radial velocity back to the rest frame and continuum normalized (see Nidever et al. 2015), with a slight SNR dependence on the continuum normalization that we discover with our Gaussian Process modelling. The spectra

Table 1. The list of the 25 elements that we select for our spectral modelling and their corresponding central wavelength (in a vacuum) corresponding to Fig. 4.

Element	Window centre / Å	Elemental family
Al	16723.500	Light odd-Z (green)
C	15582.101	Light (blue)
Ca	16155.176	Alpha (red)
Ce	15789.063	s-process (brown)
Co	16158.700	Iron peak (orange)
Cr	15684.264	Iron peak (orange)
Cu	16010.023	Iron peak (orange)
Fe	15495.100	Iron peak (orange)
Ge	16764.238	s-process (brown)
K	15167.081	Light odd-Z (green)
Mg	15745.000	Alpha (red)
Mn	15221.867	Iron peak (orange)
N	15321.871	Light (blue)
Na	16378.276	Light odd-Z (green)
Nd	15372.342	r-process (purple)
Ni	15559.517	Iron peak (orange)
O	15760.300	Alpha (red)
P	15715.930	Light odd-Z (green)
Rb	15293.534	s-process (brown)
S	15482.319	Alpha (red)
Si	15964.600	Alpha (red)
Ti	15339.241	Alpha (red)
V	15929.052	Iron peak (orange)
Y	15624.142	s-process (brown)
Yb	16502.973	r-process (purple)

cover the range of 15100.80–16999.81 Å, and comprise $n_\mathrm{b} = 8575$ spectral bins. Repeated inversion of the $n_\mathrm{b} \times n_\mathrm{b}$ covariance matrices required for inference would be prohibitively slow, and we thus restrict our analysis to a set of 25 spectral windows centred on lines confidently assigned to 25 different individual elements. These element windows have been chosen from the set of windows used to drive the APOGEE abundances in consultation with Jon Holtzman and Matthew Shetrone (Holtzman et al. 2015; Shetrone et al. 2015). Specifically, we process all spectral bins within ±1.5 Å of the line centres specified in Table 1, reducing the number of spectral bins to $n_\mathrm{b} = 343$ and hence inversion time by a factor of ∼15000. The elements responsible for these absorption lines can be grouped into the following nucleosynthetic families: iron-peak, alpha-process, r-process, s-process, light and light with odd atomic number. We expect that common production mechanisms should correlate elemental abundances and hence these spectral windows. To examine correlations between and within the nucleosynthetic family members, we colour the elements by their families in relevant figures throughout the paper, setting out these colours in Table 1.

In selecting the windows to examine, we were confronted with a series of choices, each of which ultimately impacted the emphasis of our downstream analysis. Given the potential computational expense of this modelling approach, for our proof-of-concept analysis we adopted only a single-line region for every element, but for as many elements as possible, thereby prioritizing breadth across elements in our correlation and inpainting investigations. The requirements for a line to be selected were that it be identified as one of the windows used by APOGEE in their processing pipeline ASPCAP (García Pérez et al. 2015), as well as being both strong and, where possible, unblended. In some cases, multiple lines fitted these criteria for a single element.

3 METHODS

Gaussian processes are a conceptually simple yet extremely powerful tool for regression and classification (Rasmussen & Williams 2006). Put briefly, a Gaussian process is a set of random variables whose joint distribution is multivariate normal, and is therefore fully specified by a mean function and covariance function. By their (Gaussian) nature, Gaussian processes permit simple, often analytically tractable, inference of their mean and covariance functions given (potentially noisy) observations, yielding flexible non-parametric fits to underlying trends in data and probabilistic predictions for new observations. As a result, Gaussian processes have found use throughout astronomy, from cosmology (Bond & Efstathiou 1987) and cosmography (Shafieloo, Kim & Linder 2012) to models of instrumental systematics (Gibson et al. 2012), exoplanet populations (Foreman-Mackey, Hogg & Morton 2014) and stellar spectra (Czekala et al. 2017).

In this work, we model the underlying 'true' spectrum ($\boldsymbol{s}_i$), of the ith APOGEE red clump star as a draw from a Gaussian process with a mean spectrum ($\boldsymbol{m}$) and covariance (S) to be inferred from the data. In typical Gaussian Process models, the covariance function is taken to be one of a number of standard kernels (Rasmussen & Williams 2006), chosen to reflect known or assumed properties of the observation and/or physical process (e.g. stationarity, isotropy, or periodicity). In the following, we do not assume an analytic form for our covariance function as is traditional in Gaussian Process models. Rather, we infer the correlations between the observed spectral bins, i.e. the individual elements of the covariance matrix. By doing so, we remove any potential for bias induced by a sub-optimal kernel choice incorrectly enforcing stationarity, a single correlation length, or a particular line shape, for example. As a result, we cannot make predictions for the spectra between the observed bins, though this would in principle be possible given stellar spectra observed on shifted or irregularly sampled grids.

We assume the spectral data ($\hat{\boldsymbol{d}}_i$) have been observed with Gaussian noise that is uncorrelated between spectral bins, yielding a diagonal noise covariance matrix (N_i) for each star. Masked pixels are assigned unit flux and (effectively) infinite noise uncertainties. To account for the fact that the red clump might consist of multiple distinct subpopulations (or one population whose distribution of true spectra is non-Gaussian), we allow for multiple classes to exist in our model, each described by its own Gaussian process. We assume non-informative priors on the variables defining these Gaussian processes, adopting an infinite uniform prior on each mean and an inverse-Wishart prior on each class's covariance matrix (Gelman et al. 2013, p. 73). We define the inverse-Wishart prior to have $n_\mathrm{b} + 1$ degrees of freedom, thereby placing a uniform prior on inter-pixel correlations, and a diagonal scale matrix ($\epsilon\,\mathsf{I}$, with $\epsilon = 10^{-6}$), minimizing the impact of the prior relative to the data. We infer the class membership of each star (κ_i), assuming they are sampled from categorical distributions with class probabilities ($\boldsymbol{p}$) drawn, in turn, from a symmetric Dirichlet prior with concentration parameter $\alpha = 1$.[1] These priors state our beliefs that, *a priori*: no location is preferred for the mean spectra; no scale is preferred for the covariance between two spectral bins; and the stars are as likely to be spread evenly between classes as they are to be concentrated in a single class. Our priors make no assumptions about (nor place any constraints on) the

[1]The nth-order Dirichlet distribution is the set of n-dimensional lists with elements in the range 0 to 1 that sum to unity. It describes all ways to partition a data set into n classes, allowing for particular combinations to be preferred over others if desired.

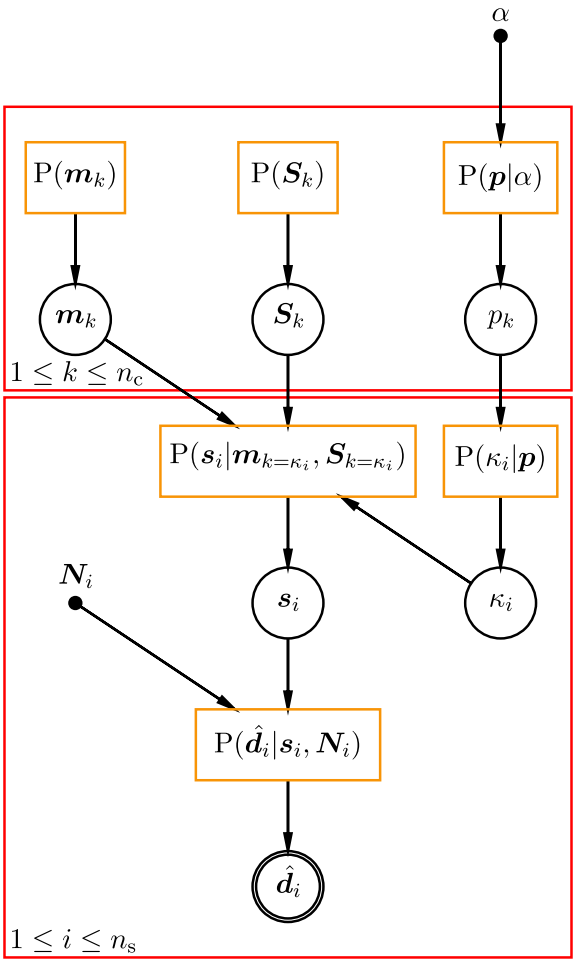

Figure 1. Network diagram for our hierarchical Bayesian model which is a graphical representation of our implemented modelling of the data. See Table 1 for the parameter descriptions.

physics of the data set, reflecting our desire for a purely data-driven inference. Should robust physical priors exist in another setting, it is simple to add them to the analysis.

The data, model parameters, priors and likelihood fully specify our probabilistic model of the APOGEE red clump data set. This model is naturally hierarchical, with some parameters describing populations and others individual stars. This hierarchical nature is made clear in Fig. 1, in which we plot the model as a network diagram. In this diagram, random variables are shown as single black circles, observables as double black circles and fixed parameters as solid black dots. Links between parameters are indicated by arrows, with the probabilistic relationships defining the links contained within orange boxes. The direction of these arrows indicates the order in which parameters must be drawn in order to forward-model the data. Finally, populations of objects are contained within red rectangles or plates, with the indices denoting membership of the population defined in the bottom left of the plate.

For clarity, we set out our model's parameters, data and constants in Table 2 and the probability distributions defining each link in the top section of Table 3. The particular set of probability distributions chosen allow for the conditional distributions of each model parameter to be written analytically: these conditional distributions are specified in the bottom section of Table 3. We are therefore able to use Gibbs sampling (Geman & Geman 1984) to estimate the joint

Table 2. Model parameters, data and constants.

quantity	Description
n_s	Number of stars (29502)
n_c	Number of classes (default: 1)
n_b	Number of spectral bins (default: 343)
$\boldsymbol{m}_k$	Mean spectrum of kth class
S_k	Intrinsic spectral covariance of kth class
p_k	kth class probability: fraction of stars in kth class
α	Concentration parameter of Dirichlet prior on class fractions
$\boldsymbol{s}_i$	True spectrum of ith star
κ_i	Class assignment of ith star
$\hat{\boldsymbol{d}}_i$	Observed spectrum of ith star
N_i	Noise covariance matrix of ith star

posterior. Gibbs sampling is a special case of Metropolis–Hastings Monte Carlo (Hastings 1970) in which a single iteration consists of redrawing each parameter in turn from its conditional distribution based on the current state of the sampler. For example, in our case, we first update the class probabilities, then the class memberships, the true spectra, and finally each class's mean spectrum and covariance matrix. Drawing proposed updates from the conditional distributions means the acceptance probability is, by definition, unity, yielding a highly efficient sampler even in high-dimensional settings. By default, we initialize the sampler using the data, generating random class memberships before setting each class's mean spectrum and covariance matrix to the sample mean and covariance of the class members' data, and the true spectrum of each object to its observed data.[2] The resulting sampler is written in PYTHON and made publicly available on Github.[3]

Our Gaussian Process model goes beyond standard approaches. As we sample the individual elements of the signal covariance matrix, the prior for the true spectra is very non-Gaussian. Were we to fit the covariance once and hold it fixed, as is common in the field, the true spectra would be Gaussian-distributed. By marginalizing over the covariance, however, we render these distributions very heavy-tailed, promoting sparse (i.e. typically small but occasionally large) excursions from the mean. As a result, we name the code SSSpaNG: *Stellar Spectra as Sparse, data-driven, Non-Gaussian processes.*

To demonstrate the effectively non-Gaussian nature of the prior on each star's true spectrum we can explicitly marginalize over the true signal covariance S. Limiting ourselves to a single mixture component for clarity, we see that

$$\begin{aligned}\mathrm{P}(\boldsymbol{s}_i|\boldsymbol{m}) &= \int \mathrm{P}(\boldsymbol{s}_i|\boldsymbol{m}, \mathsf{S})\,\mathrm{P}(\mathsf{S})\,d\mathsf{S} \\ &\propto \int |\mathsf{S}|^{-\frac{(2n_\mathrm{b}+3)}{2}} e^{-\frac{1}{2}\mathrm{Tr}\left([(\boldsymbol{s}_i-\boldsymbol{m})\otimes(\boldsymbol{s}_i-\boldsymbol{m})+\epsilon\mathsf{I}]\mathsf{S}^{-1}\right)} d\mathsf{S} \\ &\propto |(\boldsymbol{s}_i-\boldsymbol{m})\otimes(\boldsymbol{s}_i-\boldsymbol{m})+\epsilon\mathsf{I}|^{-\frac{(n_\mathrm{b}+2)}{2}}, \end{aligned} \tag{1}$$

where the integral can be performed by identifying the integrand as an un-normalized inverse-Wishart distribution over S. The result can be rewritten in the following suggestive form

$$\mathrm{P}(\boldsymbol{s}_i|\boldsymbol{m}) \propto e^{-\frac{n_\mathrm{b}+2}{2}\ln|(\boldsymbol{s}_i-\boldsymbol{m})\otimes(\boldsymbol{s}_i-\boldsymbol{m})+\epsilon\mathsf{I}|}, \tag{2}$$

[2]We obtain completely consistent results if we initialize the mean and true spectra to unity and the covariance matrix to the identity matrix.
[3]https://github.com/sfeeney/ddspectra

Table 3. Priors, likelihoods, and conditional distributions for Gibbs sampling. In our simplified notation, U, D, N, and W^{-1} denote uniform, Dirichlet, normal and inverse-Wishart distributions, respectively.

distribution	Form	Process
$\mathrm{P}(\boldsymbol{m}_k)$	$\mathrm{U}(-\infty, \infty)$	Prior on kth class's mean spectrum
$\mathrm{P}(\mathsf{S}_k)$	$\mathrm{W}^{-1}(n_\mathrm{b}+1, \epsilon\,\mathsf{I})$	Prior on kth class's spectrum covariance
$\mathrm{P}(\boldsymbol{p}\|\alpha)$	$\mathrm{D}(\alpha)$	Prior on class probabilities
$\mathrm{P}(\boldsymbol{s}_i\|\boldsymbol{m}, \mathsf{S}, \kappa_i)$	$\mathrm{N}\left(\boldsymbol{m}_{k=\kappa_i}, \mathsf{S}_{k=\kappa_i}\right)$	ith object's spectrum as Gaussian Process
$\mathrm{P}(\kappa_i = k\|\boldsymbol{p})$	p_k	i^{th} object's class membership
$\mathrm{P}\left(\hat{\boldsymbol{d}}_i\|\boldsymbol{s}_i, \mathsf{N}_i\right)$	$\mathrm{N}(\boldsymbol{s}_i, \mathsf{N}_i)$	Noisy, masked spectral measurements
$\mathrm{P}(\boldsymbol{m}_k\|\mathsf{S}_k, \boldsymbol{s}, \boldsymbol{\kappa})$	$\mathrm{N}\left(\frac{1}{n_k}\sum_{\kappa_i=k}\boldsymbol{s}_i, \frac{1}{n_k}\mathsf{S}_k\right)$	Conditional of kth class's mean spectrum
$\mathrm{P}(\mathsf{S}_k\|\boldsymbol{m}_k, \boldsymbol{s}, \boldsymbol{\kappa})$	$\mathrm{W}^{-1}(n_k+n_\mathrm{b}+1, \Gamma_k+\epsilon\,\mathsf{I})$, where $\Gamma_k = \sum_{\kappa_i=k}(\boldsymbol{s}_i-\boldsymbol{m}_k)\otimes(\boldsymbol{s}_i-\boldsymbol{m}_k)$	Conditional of kth class's spectrum covariance
$\mathrm{P}(p_k\|\boldsymbol{\kappa}, \alpha)$	$\mathrm{D}(\boldsymbol{a})$, where $a_k = \alpha + n_k$	Conditional of class probabilities
$\mathrm{P}\left(\boldsymbol{s}_i\|\boldsymbol{m}_{k=\kappa_i}, \mathsf{S}_{k=\kappa_i}, \hat{\boldsymbol{d}}_i, \mathsf{N}_i\right)$	$\mathrm{N}(\boldsymbol{w}_i, \mathsf{W}_i)$, where $\mathsf{W}_i = \left(\mathsf{S}^{-1}_{k=\kappa_i} + \mathsf{N}_i^{-1}\right)^{-1}$ and $\boldsymbol{w}_i = \mathsf{W}_i\left(\mathsf{S}^{-1}_{k=\kappa_i}\boldsymbol{m}_{k=\kappa_i} + \mathsf{N}_i^{-1}\hat{\boldsymbol{d}}_i\right)$	Conditional of ith object's spectrum
$\mathrm{P}(\kappa_i = k\|\boldsymbol{m}, \mathsf{S}, \boldsymbol{p})$	$\frac{\exp\left(-\frac{1}{2}\left[\chi^2_{i,k}+\ln\|\mathsf{S}_k\|\right]+\ln p_k\right)}{\sum_{k'}\exp\left(-\frac{1}{2}\left[\chi^2_{i,k'}+\ln\|\mathsf{S}_{k'}\|\right]+\ln p_{k'}\right)}$, where $\chi^2_{i,k} = (\boldsymbol{s}_i-\boldsymbol{m}_k)^T\,\mathsf{S}_k^{-1}\,(\boldsymbol{s}_i-\boldsymbol{m}_k)$	Conditional of ith object's class membership

from which it becomes clear that this is a highly sparsifying prior whose negative logarithm is non-convex. Conceptually, it strongly prefers spectra close to the class mean, but if a spectrum differs greatly from the mean it is only penalized logarithmically. Note that the covariance prior's scale matrix, $\epsilon\,\mathsf{I}$, acts to soften the prior, providing a small but non-zero floor to the determinant that reduces the preference for spectra exactly matching the mean. This reasoning explains why Gaussian-process modelling can outperform sparse image-reconstruction techniques (Sutter et al. 2014).

4 RESULTS

4.1 Validation of methodology: predicting unmeasured spectral regions

To avoid the complications of comparing data gathered by different spectrographs, we validate our model and code by artificially masking a portion of one of our APOGEE spectra, namely the 15789 Å cerium (Ce II) window of our lowest-SNR star (2M18335753–1302240), with an SNR measurement of SNR = 21, as listed in the APOGEE allStar file. We chose this feature, in particular, as it is a high-value detection in the APOGEE spectral region, being an s-process element. This feature was initially reported in Cunha et al. (2017), who have provided measurements for a handful of the APOGEE stars. Measurements of this element for the full APOGEE survey would build on its chemodynamical reach. This would enable the mapping of the neutron capture family, in addition to the alpha, light and iron-peak elements, across the disc and into the halo and Local Group (e.g. Nidever et al. 2014; Hayden et al. 2015; Majewski et al. 2017; Weinberg et al. 2019). Nine windows were identified in Cunha et al. (2017): we select one (unblended) Ce II window here (the line centred on 15784.75 Å in air, converted to the vacuum scale of the APOGEE spectra) for validation of our methodology.

The measured data for this star in the artificially masked region are plotted in Fig. 2 as a solid black line. The 68 per cent credible interval for the posterior probability on the star's true spectrum is plotted as dark grey, with the corresponding prediction for the observed

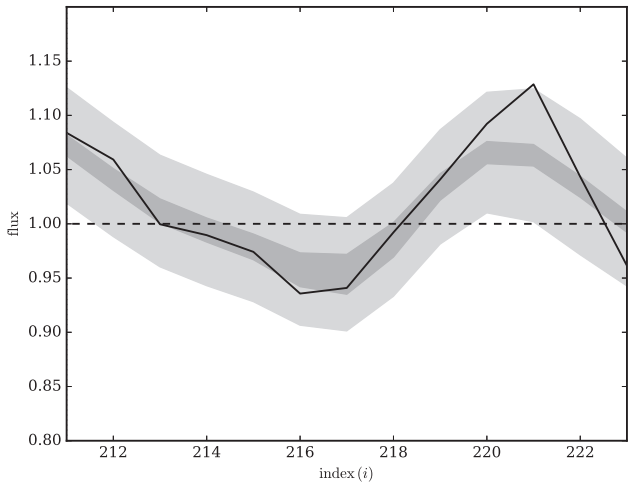

Figure 2. This figure demonstrates the validation of our model and method via the recovery of an artificially masked portion of one star's spectrum: a 3 Å region of spectrum centred on the cerium line at 15789 Å (see Table 1). We select a star with an SNR of 21 for this demonstration to highlight the performance of the model for what would traditionally be considered very low SNR data. The measured spectrum in this region is shown as a solid black line; once masked (dashed line) the flux is set to one, with infinite uncertainty. After fitting our model with the APOGEE data set (including the remainder of this star's measured spectrum), we find that the true spectrum for this star should most likely fall in the dark grey region, and the measured spectrum (i.e. including instrumental noise) should fall in the light grey region. This is in excellent agreement with the data.

spectrum [which also takes into account the (known) uncertainty on the observations] plotted in light grey. This prediction (strictly speaking, the posterior predictive distribution of the measured data) is in excellent agreement with the measured data, indicating that our model is capable of inpainting masked regions without bias. Note, in addition, that the uncertainty on the true spectrum is much smaller than the measurement noise, demonstrating our method's ability to denoise observed spectra by sharing information between stars [a

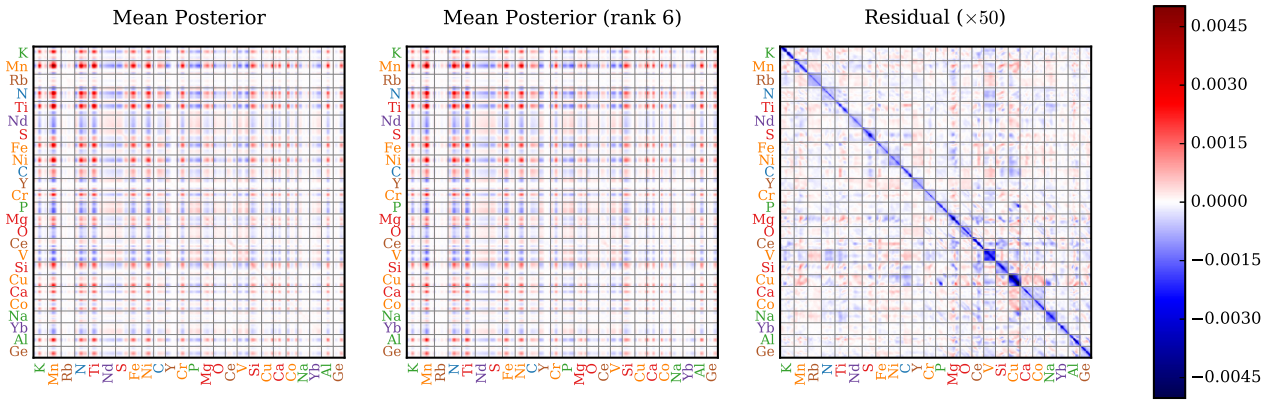

Figure 3. Left: The mean-posterior covariance matrix (S) of the 343 spectral pixels that we model, with the corresponding colour bar giving the magnitude of this covariance (in units of flux^2). The divergent colour map shows the most positive and negative covariances in red and blue, respectively, and zero covariance as white. This matrix demonstrates that the spectral pixels are highly correlated. Centre: A reduced-rank approximation of the mean-posterior covariance matrix, constructed using only those eigenvectors with eigenvalues within 10^{-2} of the largest. This represents a factor of 57 reduction in the number of eigenvectors used to construct the mean-posterior covariance matrix. Right: The residual between the mean-posterior covariance matrix and its reduced-rank approximation, boosted by a factor of 50. This nearly diagonal residual shows that most of the variation between the denoised spectra is strongly correlated between spectra bins.

phenomenon known as *shrinkage* (see e.g. Busemeyer et al. 2015, Chapter 13)].

This denoising property is relevant in the regime of extracting information from both weak lines and lower signal-to-noise data than typically required. In addition to the neutron capture element, Ce, the APOGEE spectral region has been shown to contain a number of r-process neodymium (Nd II) lines, which Hasselquist et al. (2016) estimates are detectable in ≈18 per cent of APOGEE spectra using equivalent width fitting techniques. Our expectation is that this fraction will greatly improve, given our Gaussian process modelling of the spectral lines, which can fit the true spectra of stars with lower uncertainty than the measurement noise.

We note that for this illustration we have inpainted one narrow window of a single star's spectrum, but this is generalizable to inpaint any spectral window, for any star. The predictive power to generate the spectra from the ensemble of all other stars and given prior measured spectral regions is detailed further in Sections 4.2 and 4.3.

4.2 APOGEE inference: feature correlations across the abundance windows

Our inference produces samples of the probability, mean spectrum, and covariance matrix for each class considered, and the true spectrum and class membership of each object. Focusing initially on the single-class case, we plot our covariance and mean inference in Figs 3 and 4, respectively. We plot the mean-posterior covariance matrix in Fig. 3 (left-hand panel). The covariance has strong off-diagonal structure, indicating that certain spectral features are highly correlated and anticorrelated. Its eigenspectrum also decays rapidly: only 239 of 343 eigenmodes have eigenvalues larger than 10^{-4} of the maximum, and only six larger than 10^{-2} of the maximum. A low-rank approximation to the mean covariance retaining only the six largest eigenmodes is plotted in the centre panel of Fig. 3, and the resulting residuals (multiplied by a factor of 50 to render visible) in the right-hand panel. Exploiting this decaying eigenspectrum by assuming the covariance is rank deficient would greatly reduce computation time (by a factor of roughly 187000 if six modes were retained!) but is left for future investigation.

The posterior mean of the mean spectrum is plotted in black in Fig. 4. The mean spectrum is extremely well constrained: its 68 per cent credible interval is narrower than the width of the line. To illustrate the covariance structure captured by our model, we overlay 50 realizations drawn from our Gaussian process model conditioned on the APOGEE data, colour-coded by the value they take in the first spectral bin. These samples can be interpreted as examples of potential noiseless true spectra that could have led to the data. They illustrate the variability permitted by the model and highlight certain clear trends, most notably highly correlated differences in line depths.

We demonstrate our inference of the true spectra of individual stars in Fig. 5, selecting six illustrative examples. From top to bottom, we pick out two spectra whose 15789 Å cerium windows are completely masked; two spectra whose 15372 Å neodymium windows are fully masked; and the two lowest signal-to-noise spectra. The APOGEE IDs for these stars are 2M00014650+7009328, 2M00031631+0042234, 2M04480027+3337594, 2M06053121–0631412, 2M18335753–1302240, and 2M18295507–0340512, with signal-to-noise ratios of SNR = 49, 63, 75, 41, 21, and 23, respectively. Each panel of Fig. 5 contains two shaded regions. The pink shaded area indicates the 1σ deviations from the measured spectra due to noise (these are infinitely wide when the spectrum is masked); the grey, the 68 per cent posterior credible intervals on the true spectra.[4]

Fig. 5 clearly demonstrates our ability to inpaint masked regions of spectra and denoise low signal-to-noise spectra. The inpainting results for the cerium window are particularly encouraging. We are able to make precise (and very different) estimates of these two stars' spectra in the region of the cerium line, permitting, in principle, inference of their cerium abundances where none was previously possible. The same is true for, for example, the aluminium lines of the third, fourth, and fifth stars, along with the oxygen and germanium lines of the second, fifth, and sixth stars. While we are also able to successfully inpaint the neodymium windows for the third and

[4]Recall that we are inferring the true spectra at the measured spectral bins only. In this sense, the smooth grey curves are perhaps misleading, as the posterior uncertainty is strictly infinite between data points.

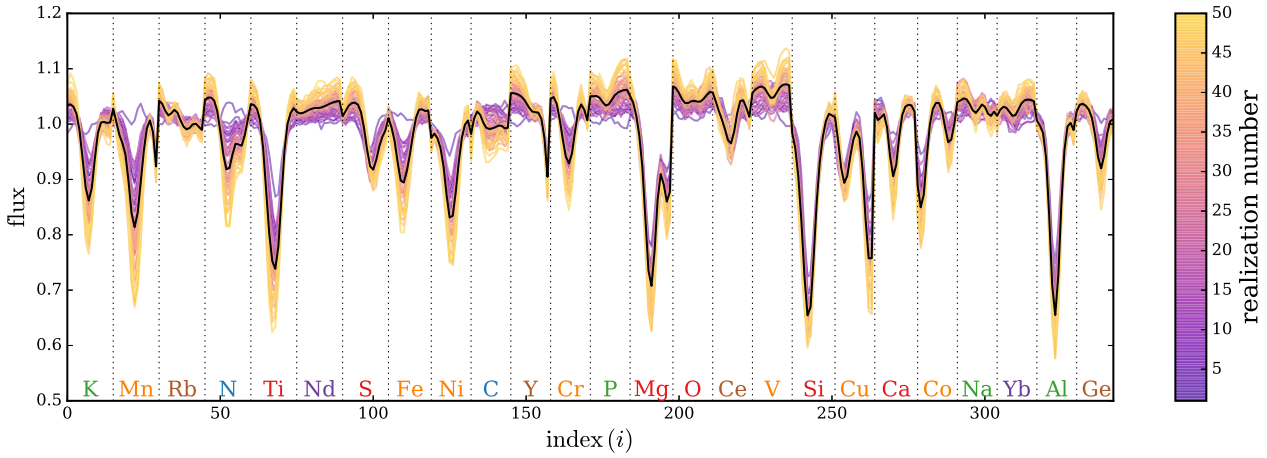

Figure 4. The mean-posterior mean spectrum ($\boldsymbol{m}$) of our Gaussian process model fit using the APOGEE data (black), along with 50 random realizations of potential true spectra ($\boldsymbol{s}$). These draws are coloured from purple to yellow according to their flux in the first spectral bin, and serve to demonstrate the correlations between pixels. Entirely uncorrelated data would show no structure in the colour gradient beyond the first bin; however, we see a clear stratification of yellow to purple as a function of the flux magnitude for most of the pixels.

fourth stars, our model infers very weak line profiles in both cases, making an abundance inference challenging. Our ability to denoise the spectra is obvious for all stars considered: the uncertainties on the true spectra are in all cases smaller than the measurement uncertainty, permitting higher precision abundance determination than previously possible. The sodium line of the last two stars is a particularly good example of the potential for our method to denoise spectra.

The results presented up to this point assume that the APOGEE red clump stars belong to a single class (and their true spectra are therefore realizations of a single Gaussian process). We have experimented with allowing multiple classes, initializing the sampler with random class memberships; however, we find little impact on our final results. The sampler finds slight differences between the classes' mean spectra ($\boldsymbol{m}_k$) and covariances (S_k), but these are driven by the initial randomized class memberships: very few stars leave one class for another during the sampling process, and those that do typically do so only once, in the sampler's first iteration. This is because the probability distribution used for drawing a star's class membership (Table 3, last row) drops exponentially with the squared distance between the star's true spectrum and each class's mean spectrum.[5] In very high dimensions, for almost all stars the distance to a new class is typically much larger than the distance to the current class, and the probability of transitioning to a new class is essentially zero. As such, we believe the class assignments are strongly dependent on the choice of initial state of the Markov chain and hence not meaningful. Exploring these high-dimensional clustering issues is left to future work.

4.3 The measured information content in the spectra

We now turn to quantifying the information contained in each elemental window. Our aim is to determine the regions of spectra that are most informative about particular elements of interest. We must note, however, that our elemental windows can contain spectral features in addition to the central absorption line, and thus strong correlations between two windows are not necessarily due solely to the central elements themselves. We start with the mean-posterior covariance within each window, $\bar{\mathsf{S}}_{XX}$, as this describes the fundamental uncertainty with which we can predict the true spectrum of a new red clump star having observed our APOGEE sample. The subscript X here denotes the spectral bins defining the elemental window of interest. We summarize this covariance matrix for six elemental windows ($X = \{\mathrm{C}, \mathrm{Na}, \mathrm{Mg}, \mathrm{Fe}, \mathrm{Yb}, \mathrm{Ce}\}$: one from each elemental family) by plotting the root-mean-square (rms) uncertainty,

$$\boldsymbol{\sigma}_X = \sqrt{\mathrm{diag}\left[\bar{\mathsf{S}}_{XX}\right]}, \tag{3}$$

in black in Fig. 6. For context, we overlay the typical measurement uncertainty[6] as a grey dashed line. This immediately demonstrates that our model of the APOGEE spectra allows us to make sub-noise predictions for some portions of a new star's spectrum without taking further data. The results for the ytterbium window are especially interesting, as the average instrumental noise seems particularly large in this region.

To determine which windows are the most informative, imagine observing one window of this new star's spectrum (corresponding to, say, element Y) *without measurement error*. The long-range correlations present in the inferred covariance matrix (i.e. the fact that elemental abundances are determined by a finite number of physical processes) imply that by doing so we should better constrain the elemental window of interest. To quantify the information gained about element X by (perfectly) observing element Y, we calculate the conditional covariance matrix

$$\mathsf{C}_{XX|Y} = \bar{\mathsf{S}}_{XX} - \bar{\mathsf{S}}_{XY}\bar{\mathsf{S}}_{YY}^{-1}\bar{\mathsf{S}}_{YX}. \tag{4}$$

The conditional covariance contains our full prediction for the uncertainty on window X having observed window Y, but we must compress it in order to construct a useful metric for quantifying information gain. We therefore define our information gain metric to

[5]Specifically, the Mahalanobis distance, or number of 'sigma' the star's spectrum is from the class's mean.

[6]The square root of the average noise variance in each spectral bin, where the average is taken over stars in whose spectra the bin is not masked.

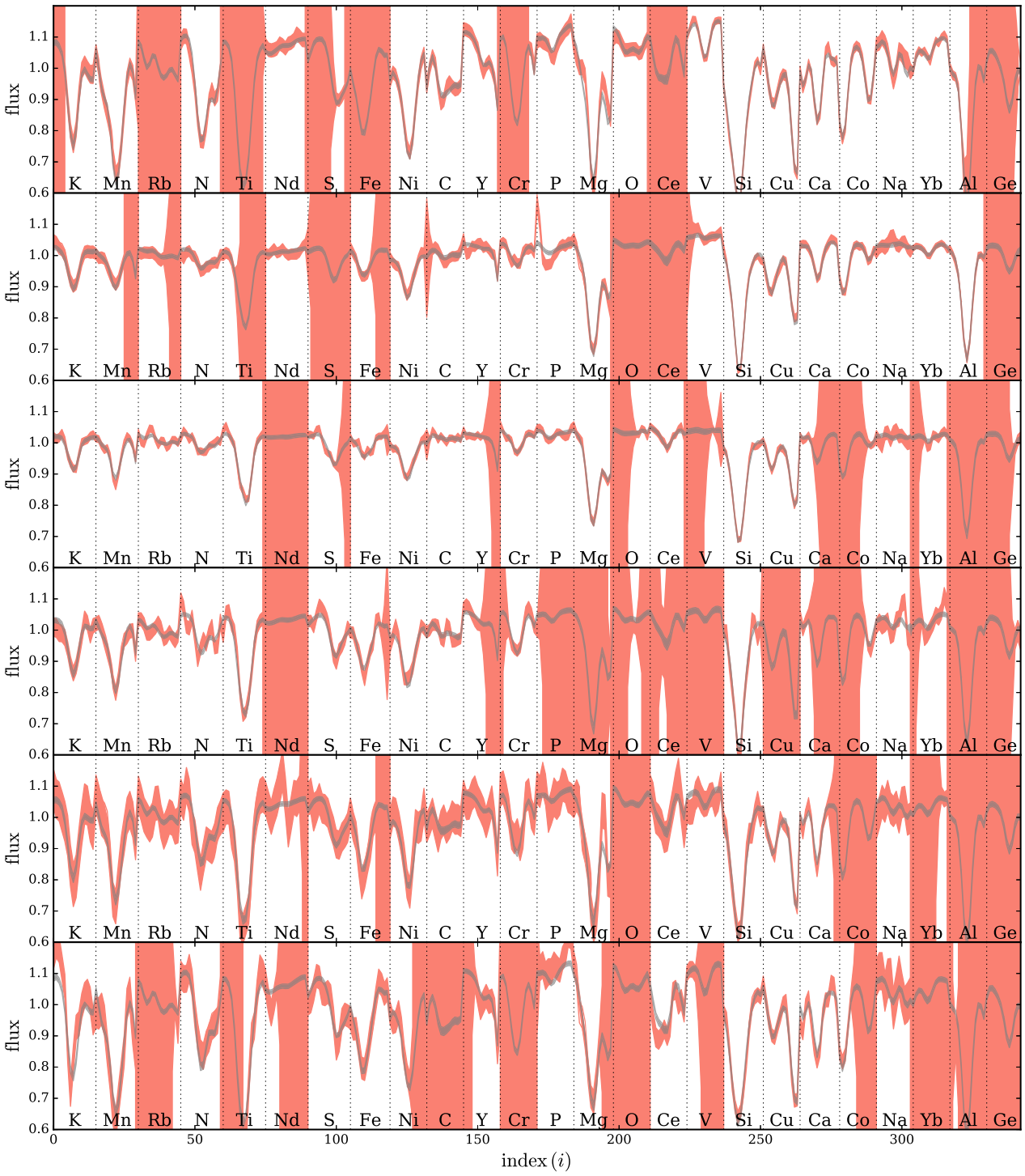

Figure 5. The measured and inferred spectra of six stars, all with low SNR (top to bottom: 49, 63, 75, 41, 21, and 23), selected to demonstrate our ability to both inpaint and denoise the data. The spectral regions shown are 3 Å windows centred on the 25 elemental lines from Table 1. The 68 per cent uncertainties on the observed spectra and inferred 'true' spectra are shown as the pink and grey shaded regions, respectively (note the masked regions in the APOGEE spectra where the measurement uncertainties flare out to infinity). There is excellent agreement between the model and data. The first two spectra have completely masked cerium lines (15789 Å), but our data-driven model makes a high-precision prediction of the cerium abundances for these stars. The middle two stars' neodymium (15372 Å) lines are completely masked. Though the model again inpaints these regions successfully, the weakness of this line means recovery of significant neodymium abundances for these stars remains challenging. All other lines inferred by the model are denoised compared to the raw data, permitting higher precision estimation of the abundances.

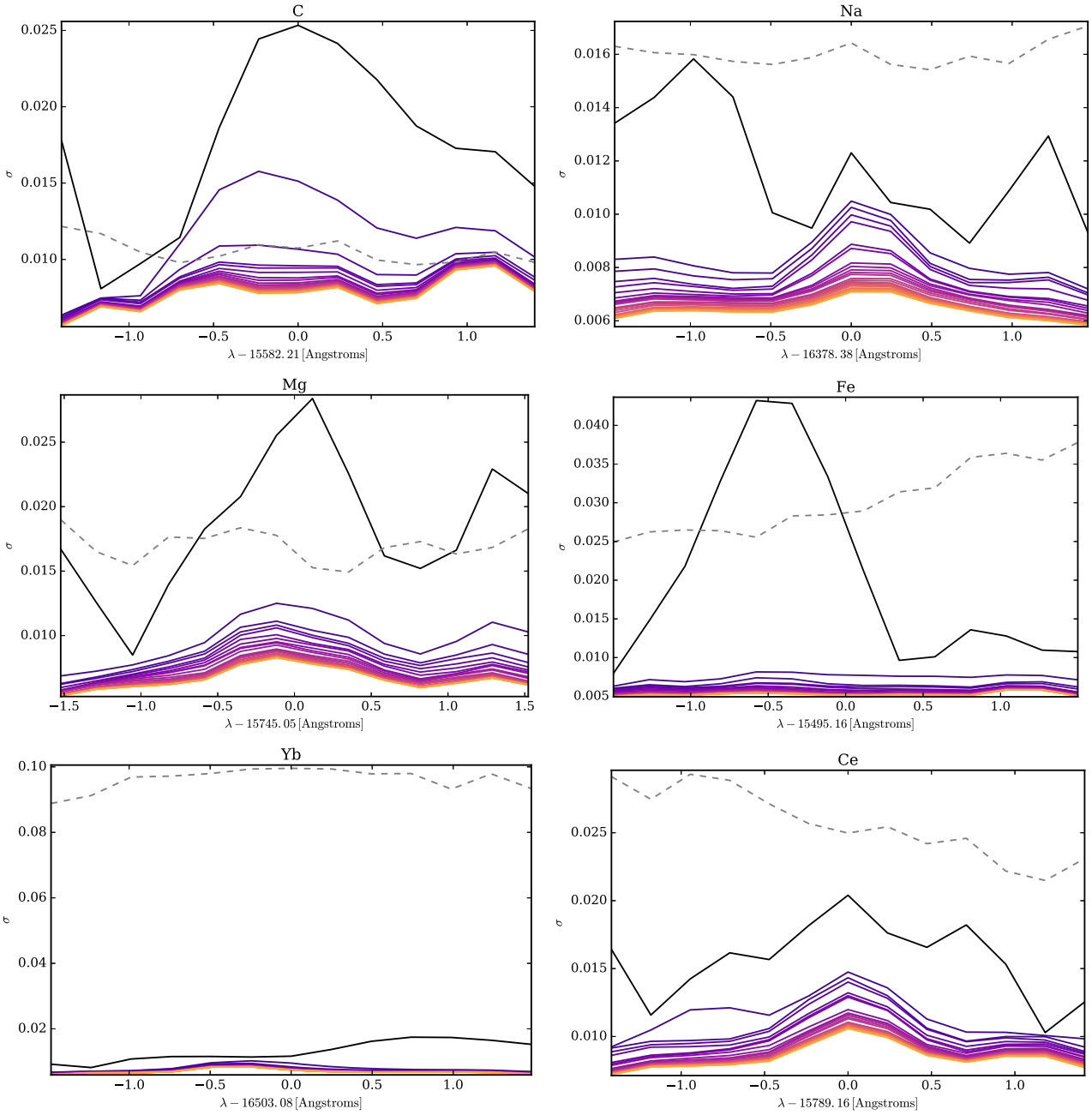

Figure 6. Root-mean-square uncertainties on the spectra within our illustrative set of elemental windows, centred on features due to C, Na, Mg, Fe, Yb, and Ce, respectively. The black line shows the uncertainty on the predicted spectrum of a new APOGEE star having not observed any portion of its spectrum; the grey dashed line indicates the typical uncertainties due to APOGEE noise. The remaining lines show how the uncertainty decreases after having perfectly observed the $1 \leq n \leq 24$ most informative elemental windows of the new star's spectrum, coloured from purple (most informative) to yellow (least informative). The order in which elements are added is plotted in Fig. 7. Note that the impact of adding observations decreases as the information gain curves of Fig. 7 become less steep.

be

$$I = \frac{1}{2} \log \frac{|\mathsf{S}_{XX}|}{|\mathsf{C}_{XX|Y}|} \geq 0. \tag{5}$$

This can be interpreted in two ways. From an information theory perspective, the differential entropy of an n-dimensional multivariate normal distribution with covariance S_{XX} is $\frac{n}{2}\left[1 + \log 2\pi\right] + \frac{1}{2}\log|\mathsf{S}_{XX}|$ (see e.g. Cover & Thomas 2006, Chapter 9). Changing the covariance matrix to $\mathsf{C}_{XX|Y}$ as we do by observing additional windows therefore changes the differential entropy of the system (i.e. adds information to it) by precisely I nats. From a geometric perspective, note that the determinant of a matrix is the hypervolume of the ellipsoid whose major axes are the eigenvectors of the matrix and have length of the eigenvalues. The square root of the determinant of a *covariance* matrix is therefore the volume of the error ellipsoid on the quantities of interest, up to a constant prefactor. Our metric I can therefore also be interpreted as the logarithmic factor of improvement in predicting window X's true spectrum obtained by

observing window Y. Regardless of the interpretation, observing a new window can only add information, contracting the covariance matrix (or, in the worst-case scenario, leaving it unchanged), and thus I cannot be less than zero.

With this metric in hand, we can take each elemental window in turn and determine the information gained by observing each other window. The window Y^1 with the most negative I is the most informative about our target window X; indeed, as our metric I is symmetric, these two elements are the most informative about each other. We then repeat this process, conditioning on Y^1 *and* each other window in order to find the second most informative window, Y^2, continuing to add windows until we find the optimal order in which to build up information on the element of interest. We denote the list of the n most informative elements $\boldsymbol{Y}^n = \{Y^1, Y^2, \ldots, Y^n\}$; the covariance in window X conditioned on these elements is $\mathsf{C}_{XX|\boldsymbol{Y}^n}$.

We plot the results of this process for the six illustrative elements in Figs 6 and 7. In Fig. 6, we demonstrate how the rms uncertainty within each elemental window shrinks as we condition on more and more information, now taking the rms uncertainty to be

$$\boldsymbol{\sigma}_X = \sqrt{\mathrm{diag}\left[\mathsf{C}_{XX|\boldsymbol{Y}^n}\right]}. \tag{6}$$

We plot the rms uncertainties after conditioning on the $1 \leq n \leq n_\mathrm{b} - 1$ most informative windows as a series of solid curves, coloured from purple to yellow. Observing the most informative window, Y^1, significantly improves the uncertainty on the spectral window of interest, and conditioning on additional windows continues to add information, albeit with diminishing returns. After observing all other windows, the rms uncertainty at the centre of the window of interest (i.e. directly over the elemental absorption line) has been reduced by a factor of roughly 2–5.

In Fig. 7, we plot the most informative windows for our six elements of interest first, along with the information gained by observing each additional window moving to the right on the x-axis. The windows' labels are coloured by their elemental family, with members of the target element's family picked out in bold. Recall that our information gain metric can be interpreted as the logarithm of the fractional reduction in volume of the error ellipses on the true spectrum. These plots cover the rough range $1.4 \leq I \leq 5.3$, corresponding to reducing the error volume by factors of 4 to 200. Reflecting the qualitative results of Fig. 6, each of the curves in Fig. 7 flattens as more elements are observed, indicating that the single greatest information gain is provided by observing the most informative elemental window and the bulk of the information is provided by the first 10 or so elements. None of the curves plateau, however, and thus all elements provide information on the window of interest. It is perhaps interesting to note that the most informative element is not, in general, from the same family as the element of interest (though this is true for magnesium). We caution over-interpretation of this point, however, for two reasons: (1) this conclusion applies only to this specific set of spectral windows and (2) these windows are broader than the elemental features they are designed to capture, and can therefore contain information about a number of elements.

Having discussed our detailed findings for the six illustrative elemental windows, we now summarize the results for all of the elemental windows. In Fig. 8, we plot the information gains for every pair of windows; that is, for each elemental window we plot the information we would gain by observing each other window perfectly. As we have demonstrated in Figs 6 and 7, there is much information to be gained by adding further observations, but given there are 24! ways of ordering them we will have to make do with the first. In doing so, we at least discover the most informative elemental pairs. We present the complete set of information gains in two ways. In the left-hand panel of Fig. 8, we group the elements by their families, sorting within each family by increasing atomic number. The most informative elemental pairs (the brightest yellow pixels) are Ni-Mn (both iron-peak), Mg-Si (both alpha) and Fe-Ti (iron-peak and alpha), and this trend is generically true of the families as a whole: the iron-peak and alpha elements predict both themselves and each other well. Indeed, these elements also predict the other families well.

There is considerable structure in this matrix, with patterns of predictivity common to multiple elements: for example, the majority of alpha-element and iron-peak rows look very similar. We make a first pass at sorting using this structure in the right-hand panel of Fig. 8. We quantify the similarity between the ith and jth rows in the plotted matrix of information gains using the distance

$$d(i \leftrightarrow j) = \sum_k |I_{ik} - I_{jk}|, \tag{7}$$

where I_{ik} is the information gain for the ith element from observing the kth.[7] To sort the elements by similar predictivity, we use a simple greedy algorithm, approximating the global optimum through a series of locally optimal choices. To start, we pick an initial value of i, then find the most similar element by determining the row j that minimizes $d(i \leftrightarrow j)$. We then take element j as the comparator, calculating distances ($d(j \leftrightarrow k)$) to find the most similar of the remaining elements, and repeat until no elements remain. This approach is not guaranteed to find the global optimum, and indeed depends on the first element chosen. We therefore repeat the process with each element as the starting point and select the sorted matrix whose total distance between rows is minimal.

The resulting sorted matrix, plotted in the right-hand panel of Fig. 8, has much clearer structure than when sorted by elemental family. On the whole, the iron-peak elements are most similar as well as most predictive, closely followed by the alpha elements; copper, vanadium, and oxygen are, however, notable exceptions to these patterns. There is also a fairly clean break around rubidium and nitrogen, beyond which the information gains drop noticeably. Note, however, that aluminium and copper are moderately informative about titanium, silicon, magnesium, and cobalt. Our ordering placed them beyond the Rb-N break; this may well be due to the sub-optimality of the greedy algorithm.

As cautioned above, all of the conclusions reached thus far are conditional on the precise definitions of the elemental windows set out in Table 1. To gain an impression of how generic these conclusions are, we repeat the above analysis using broader, 5 Å windows, presenting a version of Fig. 8 for these windows in Fig. 9. There are numerous notes to make on this figure. First, the scale extends to larger information gains: these windows are broader, contain more features and are therefore more predictive. The choice of first element that minimizes the total distance between rows in the plot is now neodymium, not nickel, but the structure is still similar: the most informative elements are from the iron-peak and alpha group, and these elements' similar predictivities mean they cluster in the plot. There is, again, something of a drop in information gains at nitrogen; however, the iron-peak and alpha elements now predict the other families better than before. Somewhat surprisingly, for these windows Y-Ni is the most informative pair. This is, however,

[7] Note that we use an absolute distance metric here: using a Euclidean distance metric instead yields similar results.

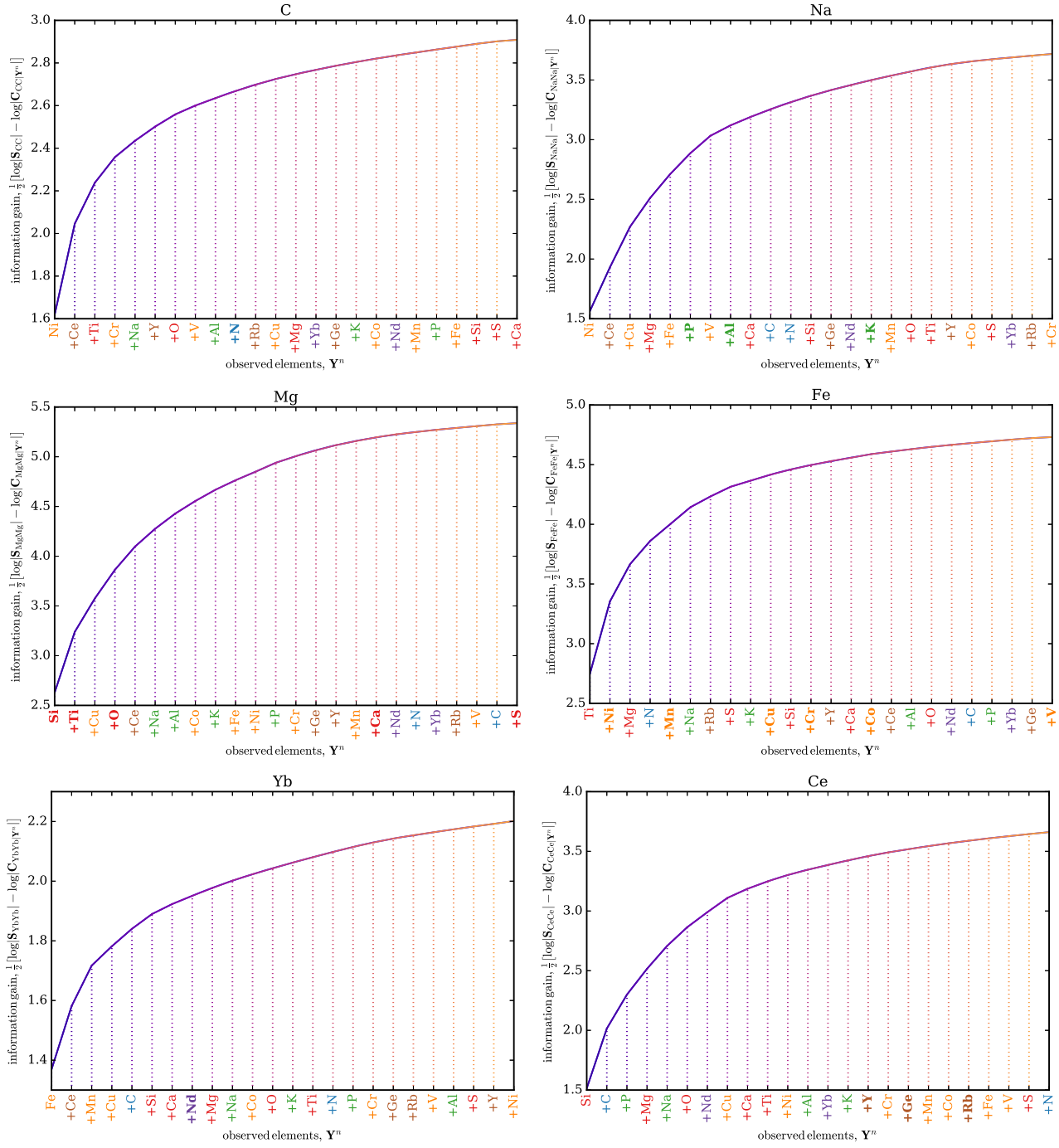

Figure 7. Information gains for our illustrative elemental windows obtained by observing the other 24 windows. The *x*-axis of each panel lists, from left to right, the window that would provide the most information on the element of interest assuming all previous windows have been observed. To take the top left-hand panel as an example: one would learn the most about the C window by observing Ni, then adding Ce, Ti, Cr, etc. The *y*-axis quantifies the resulting information gain, and can be interpreted as a change in entropy of the system or the factor of reduction in the total uncertainty on the target window's predicted spectrum provided by observing the other windows. Note the different *y*-axis ranges for the six different elements (the most extreme being Yb and Mg): the larger the overall information gain, the better the elemental window is predicted by the rest of the spectrum. Note also that while the gain from observing successive elements decreases it does not entirely flatten: each individual element adds information on the target element. Finally, the finite range of these plots indicates that, though elements are highly correlated, no one perfectly predicts another.

due to an iron line (at around 15626 Å) that appears in the yttrium window when it is extended to 5 Å. Finally, note that increasing the bandwidth to 5 Å causes our cobalt and calcium windows to merge. These two last points serve to highlight again the fact that our conclusions derive from and apply to the full spectrum within each window, not necessarily solely to the element whose line defines the window centre. Careful consideration should be made of how to define and label windows in future work.

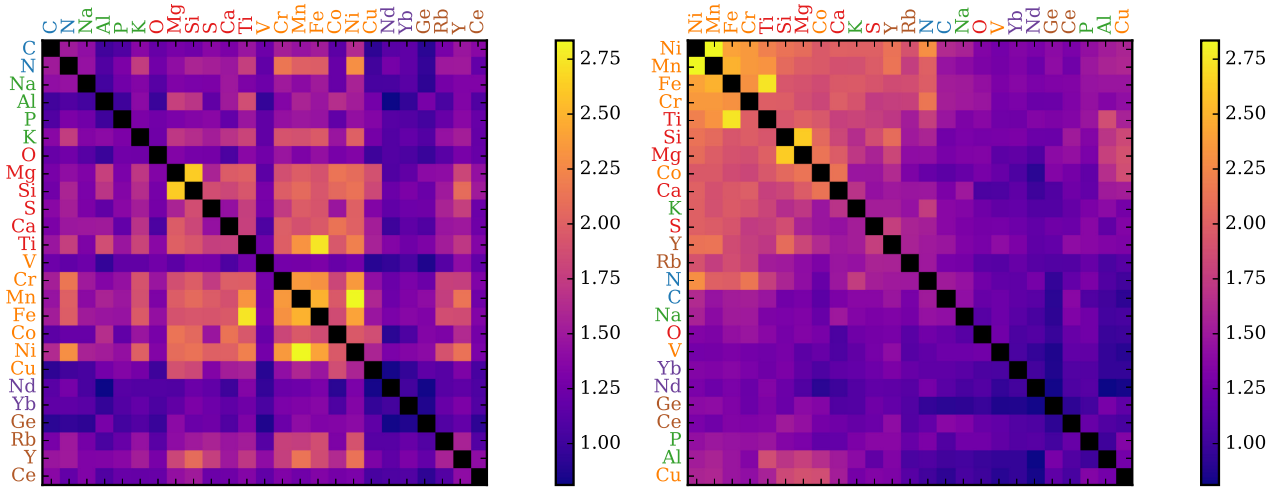

Figure 8. Information gains for pairs of elemental windows, colour coded from purple to yellow in order of increasing information gain. In the left-hand panel, the elemental windows are grouped according to their nucleosynthetic family, as indicated by the colour of their label. The iron-peak family of elements (and Ni, Mn, Fe, and Cr in particular) are the most predictive, followed by the alpha elements (Ti, Si, and Mg in particular). In the right-hand panel, the windows are sorted to minimize the difference between adjacent rows, thereby clustering elemental windows with similar information content. Note that this does not discretely separate elements into their nucleosynthetic families, particularly beyond the iron-peak and alpha elements.

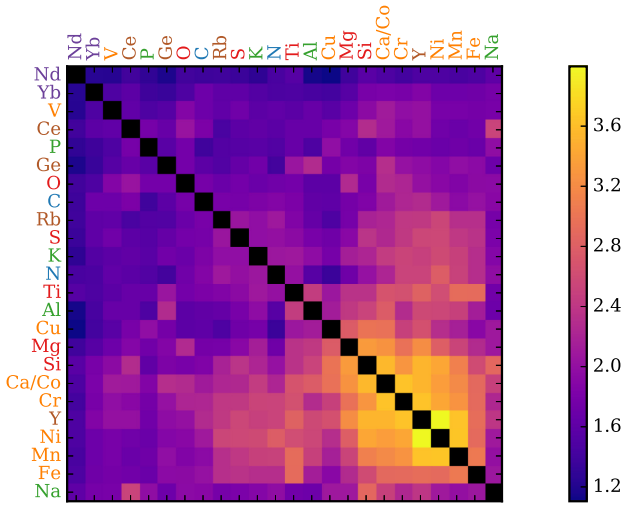

Figure 9. Information gains for pairs of elemental windows, as in the right-hand panel of Fig. 8 (with elements grouped by similar information content), now using 5 Å windows in place of our standard 3 Å windows. This figure demonstrates the impact of the precise window definitions on the information gain: the magnitude of the gains has increased and the ordering of the windows has changed, though the iron-peak and alpha elements remain most predictive and grouped as before.

5 DISCUSSION AND CONCLUSIONS

In this work, we have demonstrated how to pool information from ensembles of stellar spectra in order to denoise and inpaint individual observations, introducing a method we call SSSpaNG. This has been done with the goal of optimizing the quality and quantity of measurements that can be made from stellar spectra, including chemical abundances and ages, from upcoming million-star spectroscopic surveys. We do so by modelling the distribution of 29502 APOGEE red clump stars' spectra as a high-dimensional Gaussian Process whose covariance matrix describes the variations in spectra within the population. Inferring the elements of this covariance matrix directly, we have shown that this completely data-driven model is capable of capturing the correlations between spectral pixels. These correlations can be harnessed to yield improved estimates of individual spectra, along with precise *and* accurate predictions for unobserved spectral pixels. By marginalizing over the covariance, we effectively place a non-Gaussian, highly sparsifying prior on these inferred spectra, strongly preferring spectra close to the population mean, but penalizing large deviations only logarithmically. We produce complete spectra with decreased uncertainties for each member of the population (reducing flux errors by a factor of 2–3 for stars with SNR $\approx$ 20). This denoising will enable improved abundance inference precision, for all elements, for every star. In particular, this provides significant opportunity for far higher fidelity abundance determinations for low SNR spectra. Our method therefore significantly enhances the precision of abundance estimation from data in hand. Equivalently, it suggests that precision abundance estimates can be achieved with less telescope time per spectrum.

We have demonstrated our method's potential using the recently discovered 15789 Å cerium line, a high-value APOGEE target due to its s-process provenance (Cunha et al. 2017). Our model makes accurate and precise predictions for this line in low-SNR stars in which the line is completely masked, permitting confident estimates of cerium abundances where they would previously have not been possible.

Modelling the red clump stars' spectra as a Gaussian Process also allows us to quantify the information gained by observing portions of a star's spectrum, and thereby define the most mutually informative regions of spectra. We have done so for windows centred on 25 elemental absorption lines in the APOGEE wavelength range, demonstrating that the iron-peak and alpha-process elements are particularly mutually informative. Harnessing this information, we are able to predict the spectrum in all but one of our example windows with uncertainty less than the APOGEE noise given high-precision observations of the single most-informative window. While we are unable to perfectly predict the flux in any single elemental window by observing a combination of other windows, we find that the majority of information about a target window is typically contained in the 10-or-so most informative windows. This is a

clear demonstration of the power of using the data themselves to drive our understanding of the diversity of (and relationships between) different nucleosynthetic channels. Indeed, the correlation structure and information content that we can measure directly should place strong constraints on the physical processes that control chemical evolution. These relationships could inspire new, data-driven approaches to chemical evolution modelling (also see Casey et al. 2019), replacing current theoretical approaches that fail to reproduce observed elemental yields in detail (e.g. Rybizki, Just & Rix 2017; Blancato et al. 2019). Our information gain results also have important repercussions for the design of future observations, motivating the targeting of carefully selected, restricted spectral windows that yield strong predictions on a range of unobserved elements.

It is important at this point to address the current limitations of this method. The computational cost of the method is dominated by the matrix inversions required, which scale as the number of spectral pixels cubed. For each iteration of the Gibbs sampler, we must perform one inversion per star and one inversion per class: too many for us to process the complete APOGEE data set given available resources. In this work, we have restricted ourselves to narrow windows around our target elements; however, our results (most notably Figs 7 and 9) clearly show that there is significant value in including more of the spectra if possible. There are two obvious ways to achieve this: by throwing greater computational resources at the problem, or by exploiting the decaying eigenspectrum of the covariance matrices we find to infer a low-rank approximation to the covariance.

In the first approach, we can exploit the manifest parallelism in our algorithm. With access to the same number of CPUs as stars in the sample one could reduce the number of inversions per CPU per Gibbs sample to two at most.[8] Walltimes for our current 343-pixel runs are roughly 7 h on 48 Intel Xeon CPUs; with 29502 CPUs the full data set could therefore be processed in 7.5 d, though RAM-usage considerations might also affect this calculation. While clearly computationally heavy, this is feasible on existing large computing facilities.

In the second approach, the simplest way to reduce the rank of the inferred signal covariance matrix is to project the data on to the largest $m < n_b$ principal components of the *sample* covariance matrix prior to inference. Unfortunately, as the sample covariance contains both noise and signal its principal components are suboptimal for this task, severely degrading the inference. A natural solution would be to amend our model to explore only covariance matrices with a restricted structure (e.g. diagonal plus low-rank, along the lines of Zhang, Sarkar & Mallick 2013). We leave such extensions to future work.[9]

In the meantime, we are restricted to carrying out the analysis in windows as in this work. As the results depend entirely on the windows selected, the set of windows should be carefully optimized for the task at hand. In this proof-of-concept paper, we simply selected the strongest well-defined lines for a range of interesting elements, using a fixed bandwidth for all windows. For targeted applications, our information gain metric provides a well-motivated tool with which to optimize both the positions and widths of the elemental windows used. We have demonstrated in this work that restricting to a subset of windows still permits significant denoising and inpainting. This performance can be adapted to particular goals through careful definition of the windows; however, cutting the spectra clearly penalizes our ability to make serendipitous discoveries of new lines. We have shown here the method's ability to discover weak lines in noisy and masked spectra, but this is only possible because some stars have observed the relevant wavelengths. The loss of discovery space is a cost that must be weighed against improved performance in future applications of this work.

The final current limitation of this method is the poor sampling performance we observe when inferring the properties of multiple populations in our very high-dimensional APOGEE data. For the moment, we have chosen to model the red clump with a single class, asserting that the stars' binned spectra are distributed as a multivariate normal. As such, our handling of contaminants (or outliers) is suboptimal. Contaminants will manifest as non-Gaussianity or multimodality in the bulk population, and will therefore increase the variance of the inferred true spectra and covariance matrix if incorrectly modelled as a single Gaussian population. We do not expect contaminants to impact our results strongly, as they are estimated to make up only 5–10 per cent of our red clump sample (Bovy et al. 2014), but the same cannot be said for more diverse data sets. We know that different stellar populations have different spectral correlation structures: globular clusters, for example, have known abundance anticorrelations that are not seen in the disc and field halo stars (e.g. Kraft et al. 1997; Gratton et al. 2015; Pancino et al. 2017; Carretta 2019). Demonstrating that our sampler can efficiently and accurately fit multiple classes will allow us to not only model data sets containing different, potentially non-Gaussian populations completely, but also discover new populations. This is particularly interesting as it ties into, for example, a method of understanding chemodynamical classes in the Galactic halo, which is expected to consist of discrete chemical sub-systems with different elemental correlations. As with the other limitations, investigating modifications to the sampler (simulated annealing, for example) to address this issue, is left to future work.

ACKNOWLEDGEMENTS

The Flatiron Institute is supported by the Simons Foundation. SMF is supported by the Royal Society. MKN is supported in part by the Alfred P. Sloan Foundation. We would like to thank Brice Menard (JHU) who was instrumental in bringing our team together to perform this work.

DATA AVAILABILITY

The data employed in this article are available for download from the Sloan Digital Sky Survey's Value Added Catalogs, at https://www.sdss.org/dr14/data_access/value-added-catalogs/?vac_id=apogee-red-clump-rc-catalog.

REFERENCES

Armillotta L., Krumholz M. R., Fujimoto Y., 2018, MNRAS, 481, 5000
Blancato K., Ness M., Johnston K. V., Rybizki J., Bedell M., 2019, ApJ, 883, 34

[8]We must invert each class's covariance matrix in order to sample the class memberships and true stellar spectra. We must also invert the sum of each star's inverse class covariance matrix and inverse noise covariance matrix in order to update its true spectrum. While we can parallelize the loops over classes and stars, the loops must be carried out sequentially, and thus some CPUs will always perform two inversions. If multiple CPUs were available for each star, these inversions could also be parallelized, further reducing walltime.

[9]The structure of the covariance matrix also implies that certain kernels could potentially serve as useful covariance functions. Exploration of the utility of, for example, rational quadratic, Gibbs or mixtures of covariance functions (Rasmussen & Williams 2006) is also left to future work.

Bland-Hawthorn J., et al., 2019, MNRAS, 486, 1167
Bland-Hawthorn J., Gerhard O., 2016, ARA&A, 54, 529
Bland-Hawthorn J., Krumholz M. R., Freeman K., 2010, ApJ, 713, 166
Bond J. R., Efstathiou G., 1987, MNRAS, 226, 655
Bonifacio P. et al., 2016, in Reylé C., Richard J., Cambrésy L., Deleuil M., Pécontal E., Tresse L., Vauglin I., eds, SF2A-2016: Proceedings of the Annual Meeting of the French Society of Astronomy and Astrophysics, p. 267
Bovy J. et al., 2014, ApJ, 790, 127
Bovy J., Leung H. W., Hunt J. A. S., Mackereth J. T., Garcia-Hernandez D. A., Roman-Lopes A., 2019, preprint (arXiv:1905.11404)
Busemeyer J. R., Wang Z., Townsend J. T., Eidels A., 2015, The Oxford Handbook of Computational and Mathematical Psychology. Oxford Univ. Press, Oxford, UK
Carretta E., 2019, A&A, 624, A24
Casey A. R. et al., 2019, ApJ, 887, 73
Casey A. R., Hogg D. W., Ness M., Rix H.-W., Ho A. Q., Gilmore G., 2016, preprint (arXiv:1603.03040)
Cirasuolo M. et al., 2014, Ground-based and Airborne Instrumentation for Astronomy V. Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham WA, USA, p. 91470N
Clarke A. J. et al., 2019, MNRAS, 484, 3476
Cover T. M., Thomas J. A., 2006, Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, New York
Cunha K. et al., 2017, ApJ, 844, 145
Czekala I., Mandel K. S., Andrews S. M., Dittmann J. A., Ghosh S. K., Montet B. T., Newton E. R., 2017, ApJ, 840, 49
Das P., Hawkins K., Jofre P., 2020, MNRAS, 493, 5195
de Jong R. S. et al., 2016, Ground-based and Airborne Instrumentation for Astronomy VI. Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham WA, USA, p. 99081O
De Silva G. M. et al., 2015, MNRAS, 449, 2604
Foreman-Mackey D., Hogg D. W., Morton T. D., 2014, ApJ, 795, 64
Frankel N., Rix H.-W., Ting Y.-S., Ness M. K., Hogg D. W., 2018, ApJ, 865, 96
Gaia Collaboration et al., 2016, A&A, 595, A2
García Pérez A. E. et al., 2015, AJ, 151, 144
Gelman A., Carlin J. B., Stern H. S., Rubin D. B., 2013, Bayesian Data Analysis, 3rd edn., Chapman and Hall/CRC, Boca Raton, FL, USA
Geman S., Geman D., 1984, IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6, 721
Gibson N. P., Aigrain S., Roberts S., Evans T. M., Osborne M., Pont F., 2012, MNRAS, 419, 2683
Gilmore G. et al., 2012, Messenger, 147, 25
Gratton R. G. et al., 2015, A&A, 573, A92
Hasselquist S. et al., 2016, ApJ, 833, 81
Hastings W. K., 1970, Biometrika, 57, 97
Hawkins K., Jofré P., Masseron T., Gilmore G., 2015, MNRAS, 453, 758
Hayden M. R. et al., 2015, ApJ, 808, 132
Helmi A., Babusiaux C., Koppelman H. H., Massari D., Veljanoski J., Brown A. G. A., 2018, Nature, 563, 85
Hogg D. W. et al., 2016, ApJ, 833, 262
Holtzman J. A. et al., 2015, AJ, 150, 148
Ho A. Y. Q., Rix H.-W., Ness M. K., Hogg D. W., Liu C., Ting Y.-S., 2017b, ApJ, 841, 40
Ho A. Y. Q., Rix H.-W., Ness M. K., Hogg D. W., Liu C., Ting Y.-S., 2017a, ApJ, 841, 40
Kollmeier J. A. et al., 2017, preprint (arXiv:1711.03234)
Kordopatis G. et al., 2015, A&A, 582, A122
Kraft R. P., Sneden C., Smith G. H., Shetrone M. D., Langer G. E., Pilachowski C. A., 1997, AJ, 113, 279
Leung H. W., Bovy J., 2019, MNRAS, 483, 3255
Mackereth J. T. et al., 2019, MNRAS, 489, 176
Majewski S. R. et al., 2017, AJ, 154, 94
Minchev I. et al., 2014b, ApJ, 781, L20
Minchev I., Chiappini C., Martig M., 2013, A&A, 558, A9
Minchev I., Chiappini C., Martig M., 2014a, A&A, 572, A92
Mitschang A. W., De Silva G., Sharma S., Zucker D. B., 2013, MNRAS, 428, 2321
Mitschang A. W., De Silva G., Zucker D. B., Anguiano B., Bensby T., Feltzing S., 2014, MNRAS, 438, 2753
Ness M., 2018, PASA, 35, e003
Ness M., Hogg D. W., Rix H.-W., Ho A. Y. Q., Zasowski G., 2015, ApJ, 808, 16
Newberg H. J. et al., 2012, in Aoki W., Ishigaki M., Suda T., Tsujimoto T., Arimoto N., eds, Astronomical Society of the Pacific Conference Series Vol. 458, Galactic Archaeology: Near-Field Cosmology and the Formation of the Milky Way. ASP Conference Proceedings, San Francisco, CA, p. 405
Nidever D. L. et al., 2014, ApJ, 796, 38
Nidever D. L. et al., 2015, AJ, 150, 173
Pancino E. et al., 2017, A&A, 601, A112
Price-Jones N., Bovy J., 2019, MNRAS, 487, 871
Rasmussen C., Williams C., 2006, Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge
Rix H.-W., Bovy J., 2013, A&AR, 21, 61
Rybizki J., Just A., Rix H.-W., 2017, A&A, 605, A59
Sanderson R. E. et al., 2018, ApJS, 246, 6
Shafieloo A., Kim A. G., Linder E. V., 2012, Phys. Rev. D, 85, 123530
Shetrone M. et al., 2015, ApJS, 221, 24
Steinmetz M. et al., 2006, AJ, 132, 1645
Sutter P. M. et al., 2014, MNRAS, 438, 768
Tamura N. et al., 2016, Ground-based and Airborne Instrumentation for Astronomy VI. Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham WA, USA, p. 99081M
Ting Y.-S., Freeman K. C., Kobayashi C., De Silva G. M., Bland-Hawthorn J., 2012, MNRAS, 421, 1231
Ting Y.-S., Conroy C., Goodman A., 2015, ApJ, 807, 104
Ting Y.-S., Rix H.-W., Conroy C., Ho A. Y. Q., Lin J., 2017, ApJ, 849, L9
Ting Y.-S., Conroy C., Rix H.-W., Cargile P., 2018, ApJ, 879, 69
Weinberg D. H. et al., 2019, ApJ, 874, 102
Wheeler A. et al., 2020, ApJ, 898, 58
Xiang M. et al., 2019, ApJS, 245, 34
Yanny B. et al., 2009, AJ, 137, 4377
Zasowski G. et al., 2013, AJ, 146, 81
Zasowski G. et al., 2017, AJ, 154, 198
Zhang L., Sarkar A., Mallick B. K., 2013, preprint (arXiv:1310.4195)

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