Unraveling the History of the Milky Way

Unraveling the History of the Milky Way

A dissertation presented
by
Yuan-Sen Ting
to
The Department of Astronomy

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Astronomy & Astrophysics

Harvard University
Cambridge, Massachusetts
December 2016

Thesis advisor: Professor Charlie Conroy Yuan-Sen Ting

Unraveling the History of the Milky Way

Abstract

Understanding physical processes responsible for the formation and evolution of galaxies like the Milky Way is a fundamental problem in astrophysics. However, a key challenge is that the properties and orbits of the stars can only be observed at present. In order to understand what happened in the Milky Way at earlier epochs, one must explore "archaeological" techniques. One idea, "chemical tagging," aims to probe the history of the Milky Way via the unique imprint in elemental abundances of long-disrupted star clusters even the stars are now on widely dispersed orbits and spatially mixed.

In my thesis, I developed the first extensive model of the Milky Way in the context of chemical tagging and explored the opportunities and challenges associated with chemical tagging. My work put the first constraint on the disrupted cluster mass function in the Milky Way and revealed that the Milky Way did not form very massive star clusters in the first five billion years. As part of this effort, I also developed a new technique for measuring stellar properties from large spectroscopic surveys, adopting ideas in machine learning. My technique provides new opportunities for estimating stellar properties more precisely and exploiting the information embedded in low-resolution spectra. This technique has many promises and might hold the key to realizing the full potential of chemical tagging in the future.

Contents

This thesis is dedicated to my family.

Acknowledgments

My four years in graduate school has been a life-changing experience, although it was sometimes rather stressful. Thankfully, I have met many amazing people who helped me to get through this process. First of all, I want to thank my adviser Charlie Conroy for his guidance. It has been a very enjoyable experience working with him. I am also thankful for the input and mentoring from Prof. Alyssa Goodman at Harvard and Prof. Hans-Walter Rix at Max-Planck-Institut für Astronomie (MPIA). I am grateful for many lasting friendships that I have made in these years. My friends never failed to cheer me up during the toughest moments: Fernando Becerra, Stephen Portillo, Pierre Christian, Philip Mocz, Zack Slepian, Jon Clindaniel, Joshua Suresh, Robert Penna, Wen-Fai Fong. Finally, I also want to acknowledge the administrative staff at the Harvard-Smithsonian Center for Astrophysics: Peggy Herlihy, Sarah Block and Rob Scholten for all their encouragement throughout the years. They provided me much needed company, especially appreciated as I am so far away from home.

For my research, I want to thank Benjamin Johnson, Philip Cargile, David W. Hogg, Melissa Ness, Matthias Steinmetz, David Weinberg, Martin Asplund, Daniel Weisz, Joss Bland-Hawthorn, Ken Freeman and Charlie Lada for many illuminating discussions. The research that I present in this thesis was supported by NASA Headquarters under the NASA Earth and Space Science Fellowship Program - Grant NNX15AR83H and was in part supported by the MPIA and the German Research

Foundation through the SFB 881 (A3) grant during my visit at MPIA. Most computational tasks were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University and on the Extreme Science and Engineering Discovery Environment (XSEDE), which is funded by National Science Foundation grant number ACI-1053575.

Finally, I want to thank Instacart, a grocery delivery app in Cambridge, Massachusetts, without which I would not have survived the winters here.

1

Introduction and a guide to this thesis

1.1 Chemical tagging – scientific background & motivation

Understanding the physical processes responsible for the formation and evolution of galaxies is a fundamental yet unsolved problem in astrophysics. Our own galaxy, the Milky Way, offers a unique chance to study the formation and evolution of the structural components of a galactic system (disk, bulge, and halo) because it is the only galaxy for which we can study individual stars in great detail. Despite being the closest galaxy, we still do not have a detailed picture of when, where or how stars in each component of the Milky Way formed, nor of the key processes responsible for the full dynamical and chemical evolution of the Milky Way. One fundamental challenge is that the properties and orbits of stars can only be observed at the present, therefore, in order to understand past properties of the Milky Way, we must rely on "Galactic Archeology" techniques to harness the stellar fossil record.

One of these techniques, "chemical tagging", aims to probe the history of the Milky Way via the unique imprint in elemental abundance space of long-disrupted star clusters which, if successful, will provide completely new ways to study the assembly history of the Milky Way. Stars form in clusters, but these clusters quickly disperse due to dynamical interactions* – only the most massive star clusters can survive more than 100 Myrs while smaller star clusters dissipate soon after they form[124]. Reconstructing star clusters is a key to understanding the Milky Way because each star cluster reveals a distinct star formation event in the Milky Way. Fortunately, some of these disrupted clusters might still remain kinematically intact in phase space for a few dynamical times, and for a short period ap-

*For example, due to n-body interaction and tidal stripping.

pear as "moving groups" (e.g., [37,49,52]). So the kinematics information in some cases can reveal star formation events in the Milky Way. But these kinematic associations will dissolve in less than 2–3 Gyrs. Therefore, the number of stellar siblings that we can reconstruct through kinematic phase space information alone is limited, especially for stars that were born during the early violent phase when the Galactic bulge and globular clusters formed.

To overcome this limitation, chemical tagging, first proposed by Freeman & Bland-Hawthorn [69], provides a complementary way to reconstruct stellar siblings. Chemical tagging presumes that stars born in the same star cluster have near-identical (photospheric) elemental abundances. The elemental abundances of stars reflect the metal enrichment of the molecular clouds from which the stars formed. Depending on the chemical evolution of the Milky Way, the metal enrichment of a molecular cloud can be unique and can be differentiated from other molecular clouds. Moreover, since most photospheric elemental abundances, at least for elements more massive than Na, are not expected to change drastically from stellar evolution (e.g., see [43]), stars should carry these unique stellar "birth markers" throughout their lifetime even they are now on widely dispersed orbits and spatially mixed. The goal of chemical tagging is thus to recognize stellar siblings by looking for stars that have an uncanny similarity in their "chemical fingerprint".

Chemical tagging postulates that stellar siblings should show up as clumps in the elemental abundance space. The right panel of Fig. 1.1 illustrates what we expect the Milky Way would look like in the elemental abundance space if we could fully exploit all of the information in the data. Each clump in the right panel shows an individual star formation event. Identifying these clumps and subsequently reconstructing the birth history of star clusters will enable qualitatively new under-

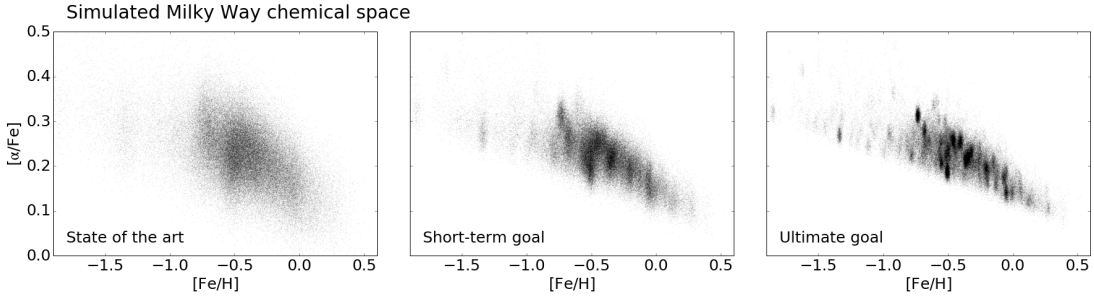

Figure 1.1: Simulated distribution of stars in elemental abundance space. Chemical tagging looks for chemically homogeneous "clumps", with these clumps revealing distinct star formation events in the Milky Way. At present, most substructures are smeared out (left panel) due to the lack of precise elemental abundances and the small sample size. My research aims to realize the full potential of chemical tagging (right panel).

standing of the origin of each star. In particular, it will shed light on the extent to which stars form in small or massive clusters of stars (the cluster mass function, or CMF), and the dynamical history of star clusters, including their formation, disruption, dispersal and orbital migration. Furthermore, the CMF is an indicator of star formation rate and gas mass[1,96] as these qualities set the maximum fragmentation mass in galactic system[115],[†] chemical tagging could also provide an unparalleled window onto past properties of the Milky Way that are not directly observable.

[†]Using a Toomre stability argument, one can derive that a higher gas-to-dynamical mass should entail the formation of more massive star clusters.

1.2 A changing observational landscape and opportunities

An essential requirement of chemical tagging is that star clusters be internally chemically homogeneous. Observationally, open clusters are indeed chemically homogeneous at the level of $\sigma_{[X/H]} < 0.02$ dex[29,131]. Theoretical arguments also suggest that star clusters should be chemically homogeneous, even for loosely bound clusters with 10^{5-7} $M_{\odot}$, as a result of turbulent mixing[63]. In addition to cluster homogeneity, molecular cloud-to-molecular cloud variation is equally important: if all star clusters share the same elemental abundances, it will not be possible to chemically separate the disrupted stars clusters. In Ting et al.[195], I demonstrated that there is a substantial variation in elemental abundances between star clusters using principal component analysis – there are about seven independent nucleosynthetic pathways through which stars in the Milky Way acquire their abundance patterns. I estimated that the number of distinguishable cloud-to-cloud variations is at least 10^3–10^4 for the ongoing large-scale spectroscopic surveys.

In recent years, chemical tagging has garnered more attention because many studies have demonstrated that population-level chemical tagging holds much promise towards identifying distinct stellar populations from various components of the Milky Way. Unlike the original proposal of strong chemical tagging which aims to find stars that were born at the same place and time (both coeval and conatal), population-level chemical tagging only search for stars that formed at roughly the same time (e.g., thick disk stars) or form from the same mechanism (e.g., globular clusters) but not necessarily formed at the same place. For example, stars accreted from different satellite galaxies into the Milky Way show distinct chemistry from the Milky Way's stars (e.g., [127,196]). It has been thought

that these variations could be used to find the remnants of disrupted satellite galaxies. Moreover, α-enhancement[‡] of stars is known to be an excellent probe to separate the younger Milky Way thin disk from the older thick disk population[86]. Apart from the Milky Way disk, Schiavon et al.[177] and Martell et al.[139] studied bulge/halo stars that have similar abundance patterns to globular clusters and estimated the fraction of bulge/halo stars that might come from accreted globular clusters. On top of these, recently Martig et al.[141] found a number of young but surprisingly α-enhanced stars that challenge our current understanding of the chemical evolution of the Milky Way.

While these studies showed that chemical tagging works in a weaker form, they only focus on searching for distinct *populations* of stars through their abundance patterns. The original proposal of chemical tagging[69], which we will call "strong" chemical tagging, aims to look for stars that were born from the same molecular cloud and in the same *star cluster*. In this thesis, I will only focus on strong chemical tagging. Therefore, unless stated otherwise, the term chemical tagging only refers to strong chemical tagging throughout this thesis. Despite its intriguing prospects, strong chemical tagging has never been demonstrated before.

Strong chemical tagging has never been realized partly because studies of high-resolution stellar spectroscopy prior to the year 2015 have been limited to about 1,000 stars[13,36]. With these small samples, one is unlikely to sample enough stars from the same cluster. As a result, chemically homogeneous clumps are hardly visible. Thankfully, the observational landscape is rapidly changing: ongoing surveys, such as as GALAH[53], Gaia-ESO[187] and APOGEE[136]. are collecting high-resolution

[‡]The relative abundances of α-capture elements (e.g., Mg, Si, Ca) compared to iron peak elements (e.g., Fe, Mn, Ni).

spectra, R=$\lambda/\Delta\lambda$=25,000 and S/N $\simeq$ 100, for 10^6 stars in the Milky Way and are measuring 20–35 elemental abundances. The next generation of surveys, including After Sloan 4 (AS4), 4MOST, and Weave, will further increase the sample size to about 10^7 stars starting in the year 2020. Besides the advent of large-scale multiplexed spectroscopic surveys, the spectroscopic data is further complemented by other revolutionary astrometric and photometric surveys such as Gaia, TESS, LSST, and WFIRST, which could provide crucial kinematic information and stellar ages to unravel the history of the Milky Way.[§] With this avalanche of data, we may finally be on the cusp of realizing chemical tagging.

In the pursuit of strong chemical tagging, the research that I will present in this thesis combines cutting-edge ideas in machine learning with classical chemodynamical modeling and spectral fitting. My work identifies challenges of strong chemical tagging, argues why the current status of surveys is not good enough for strong chemical tagging, proposes an intermediate form of chemical tagging that will work with the current status, as well as pushes the boundaries on various fronts to realize strong chemical tagging eventually. In the following, I will summarize the key ideas and results in this thesis.

[§]Either through isochrones/color-magnitude diagrams or asteroseismology.

1.3 Key ideas and results

In this section, I will summarize some key ideas that I will present in this thesis. Chapter 2 focuses on the chemodynamical modeling of the Milky Way tailored to predict chemical tagging signals from the ongoing spectroscopic surveys. The model predicts that with the current sample and elemental abundance precision, it is very challenging to perform strong chemical tagging. Due to this limitation, in Chapter 3, I will describe a new way to constrain the CMF of the Milky Way in the past via an intermediate form of chemical tagging, which we will call statistical-level chemical tagging. In Chapter 4–6, I will present a new spectral fitting technique that optimizes the maximal extraction of spectral information from spectra which is the key to going beyond the intermediate form and realizing strong chemical tagging in the future. And finally, in Chapter 7, I will discuss some of my ongoing work and other future directions.

1.3.1 First extensive Milky Way model for chemical tagging

Recall that star clusters are quickly dispersed after their formation. Assuming a rotational velocity of 200 km/s, it only takes a star about 0.5 Gyr to orbit the Galaxy. When combined with a modest amount of dynamical heating, the implication is that stars from the same star cluster are mostly mixed and distributed throughout the Solar annulus. Since many star clusters are now mixed in the Galaxy, and spectroscopic surveys only collect subsamples of stars within their observed volumes, it is unclear whether we could gather enough stars from the same cluster in the Solar neighborhood and how massive a star cluster has to be in order to be detected through chemical tagging. In ad-

dition, we also have limited resolution in separating clumps in elemental abundance space due to measurement uncertainties of elemental abundances. Due to these many caveats, without chemodynamics models of the Milky Way that directly match the ongoing surveys, it will be difficult to interpret the results of these surveys.

To this end, in Chapter 2 (Ting et al., 2015, ApJ, 807, 104), I will present the first extensive model that directly incorporates many of the key processes relevant for evaluating the chemical tagging surveys. Previous chemodynamical modeling either only focused on the global chemical evolution trends[147,178] or focused on certain stellar groups[23,24] and are not specifically tailored to explore chemical tagging in the Milky Way disk under different evolutionary histories. In the model presented in Chapter 2, I explore various observationally inspired prescriptions for the star formation history, the gas mass distribution, the stellar mass distribution and the radial size growth of the Milky Way through cosmic time and populate the elemental abundance space with simulated star clusters.

My simulations indicate that with the current measurement precision of elemental abundances ($\sigma_{[X/H]}$ = 0.05 dex), we might not be able to resolve individual star clusters through chemical tagging. I found that detectable clumps might not come from single disrupted clusters, even though they could have significantly more stars than the average background. These clumps could form from a stack of many moderately massive clusters at the resolution of the current elemental abundance precision.

Besides the lack of precision measurement of elemental abundances, the interesting "birth similarity" features of chemical tagging are currently smeared out (cf. the left panel of Fig. 1.1) because of other two other major challenges. First, the Milky Way has many more stars that are now spatially

mixed than the sample with high-resolution spectra. Even with current high-resolution spectroscopic surveys of 10^6 stars, my simulations found that we sample only $< 0.01\%$ of stars per star cluster. Since we only sample a few stars from each star cluster, interesting clumps are hardly visible. The second additional challenge is that many star clusters are low mass, meaning that the few stars they contribute are part of a "background" sea of stars. My simulations found that the chemical tagging signals diluted by a sea of "background" stars. I will tackle some of these challenges in Chapter 4–7.

1.3.2 Revealing the stellar cluster mass function in the early Milky Way

While my simulations show that we might not attain strong chemical tagging just yet, I found that the Milky Way's propoerties can still be constrained using an intermediate form of chemical tagging. Even under the current observational constraints where star clusters might not be resolved through chemical tagging, my simulation predicts that different stellar cluster mass functions entail strikingly different stellar distributions in elemental abundance space. In other words, although individual cluster signals might be currently too weak to recover, the global fluctuations in the elemental abundance space from ongoing large-scale spectroscopic surveys can still be used to reconstruct the Milky Way CMF in the past. In between the "population-level" chemical tagging and the "strong" chemical tagging, I propose the "statistical-level" chemical tagging.

In particular, in Chapter 3 (Ting et al., 2016a, ApJ, 816, 10), I will present a work where I applied a new clump search method to the APOGEE DR12 dataset and put into practice the first large-scale chemical tagging analysis of the Milky Way disk. By quantifying the clumpiness of the APOGEE el-

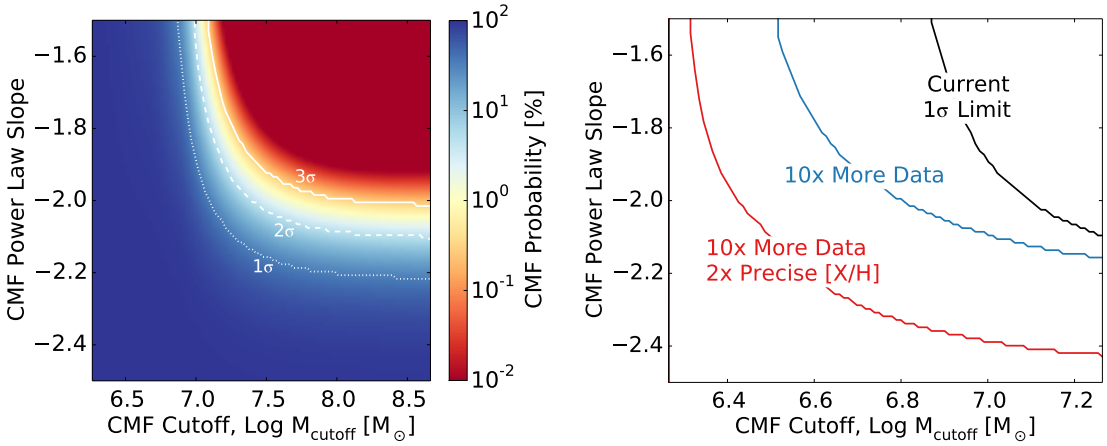

Figure 1.2: Cluster mass function constraints from chemical tagging. Left panel: By measuring the clumpiness of elemental abundance space, my work put the first chemical-tagging based constraint on cluster mass in the Milky Way thick disk. The thick disk did not form any chemically homogeneous star cluster more massive than 10^7 $M_\odot$. Right panel: Forecasts for cluster mass constraints. Upon collection of the factor of ten times more data in the next two years from various large spectroscopic surveys, chemical tagging will be able to determine whether or not 10^6 $M_\odot$ chemically homogeneous star clusters ever existed in the Milky Way.

emental abundance space with ten elements, as shown in the left panel of Fig. 1.2, I will demonstrate that the Milky Way old disk population did not form any chemically homogeneous clusters more massive than 10^7 $M_\odot$. In this work, I forward modeled the APOGEE elemental abundance space and showed that any star cluster more massive than this limit should be detectable by APOGEE via chemical tagging and yet no such signal has been found. This study put the first-ever constraint on how massive a star cluster in the Milky Way could be in the first five billion years after the Big Bang.

Despite this promising result, we note that the CMF constraints presented in Chapter 3 are limited to large cluster masses. This is due to the small available sample size from which the analyses were performed. But the GALAH, Gaia-ESO, and APOGEE 2 surveys will measure more elements (and thus define a larger search space) and a total of ten times more stars than APOGEE DR12 in the next two years. The future prospects of this approach are good. As shown in the right panel of Fig. 1.2, I project through simulations that these surveys will derive a CMF limit that is at least an order of magnitude more constraining. Chemical tagging will reveal whether or not 10^6 $M_\odot$ chemically homogeneous star clusters ever existed in the Milky Way.

1.3.3 Maximal information extraction with new spectral fitting technique

Although I demonstrate that chemical tagging now works in a somewhat limited "statistical" form, the ultimate goal is to perform "strong" chemical tagging and resolve individual disrupted star clusters in elemental abundance space. The current elemental abundance precision from state-of-the-art techniques is not sufficient to achieve strong chemical tagging. Although a cluster is intrinsically homogeneous to a level of better than 0.02 dex [29,131], the cluster will appear smeared out. Better ele-

mental abundance precision, however, will significantly improve the situation. Since chemical tagging deals with a search space with seven-to-nine independent dimensions[195], a small improvement will have an enormous effect. In fact, a factor of two improvement in each elemental abundance will improve the chemical resolving power by a factor of 1,000 (cf. Chapter 2). As a result, chemically homogeneous substructures in the elemental abundance space should be significantly more distinct as a result (cf. the right panel of Fig. 1.1 and Chapter 2).

Rigorous spectral modeling calls for all stellar labels characterizing a star to be fit simultaneously to the full spectrum. For the large datasets at hand, this appears computationally infeasible using the established rectilinear technique[71]. The rectilinear technique requires exponentially more models as the number of parameters increases. Therefore, no analysis of stellar spectra has attempted a self-consistent, simultaneous fitting of all of the relevant parameters to the full spectral range. However, only fitting a portion of the spectral range can limit the precision that is achievable. For example, restricting to unblended strong lines, a simplification that most current techniques adopt, only exploits about 10% of the spectral information, and so the precision obtained is about three times worse than the formal limit (cf. Chapter 4).

To overcome this limitation, in Chapter 4 (Ting et al., 2016b, ApJ, 826, 83) and Chapter 5 (Rix, Ting et al., 2016, ApJ, 826, L25), I will present a new spectral fitting technique that enables the fitting of $>$ 20 parameters simultaneously with minimal loss of information. The defining ideas of this new method are two-fold:

- Only a tiny fraction of the high-dimensional label space is actually occupied by real stars. We consider a convex hull, i.e., the minimal polygon that encompasses all real stars in the label

space, and an adaptive grid approximation for this subspace which has a dramatically smaller volume than a rectilinear grid.

- I directly fit the variation of a spectrum on stellar labels with an explicit functional form. I show that the underlying model spectrum for giants varies mostly quadratically with stellar labels. Only d^2 synthetic models are needed to constrain the quadratic, where d is the number of dimensions. Compare this to the standard interpolation approach, in which the number of models grows as $\exp(d)$, in a 20-dimensional space, my new method reduces the computational cost by a factor of a million.

To demonstrate how well this new spectral fitting technique works, in Chapter 4 and 5, I consider mock synthetic stellar spectra with all 18 labels (effective temperature, surface gravity, microturbulence, and 15 elements) in APOGEE. I show that this method recovers most elemental abundances with a typical systematic error less than 0.03 dex in this 18-dimensional space spanned by the APOGEE giants. This result shows a promising prospect that, with this new technique, we could significantly improve the chemical resolving power by improving the elemental abundance precision by about a factor of two. In Chapter 7, I will discuss some ongoing works, demonstrating this is indeed the case by fitting real APOGEE spectra.

1.3.4 Increasing the sample size through low-resolution spectral fitting

In Chapter 6 (Ting et al., 2016c, submitted to ApJ), I will show that fitting all elements simultaneously with the technique presented in Chapter 4 and 5 also opens a new path for dealing with low-resolution spectra. Due to read noise and low throughput, current high-resolution surveys will be limited to a million bright stars until the next generation of large-scale spectroscopic surveys in the year 2020. However, enabling chemical tagging requires that we go beyond 10^6 stars (cf. Chapter 2).

One way to proceed is by reanalyzing abundantly available low-resolution (but high S/N) spectra. For example, LAMOST and Gaia's Radial Velocity Spectrograph are observing a few million stars with a low-resolution spectrograph ($R \simeq$ 2,000 and 11,000, respectively).

Chemical tagging requires the measurement of multiple elemental abundances. High-resolution spectra with R>20,000 are thought to be indispensable to measure >20 individual elemental abundances, below which spectral lines are blended. But my new technique presented in Chapter 4 and 5 allows for the extraction of information from blended features and hence it hosts the exciting possibility of measuring > 20 elemental abundances with low-resolution spectra. With this idea in mind, in Chapter 6, I will show the result of simulations which degrade model spectra to lower resolutions. Assuming the same exposure time and number of CCD pixels, my study demonstrated that high S/N, low-resolution spectra with R=6,000 can yield elemental abundances as precise as R=100,000 spectra, provided that we fit all stellar labels self-consistently with the new approach. I also verified that the same result holds regardless of stellar type, metallicity, and wavelength coverage. This result might seems surprising at first because the high-resolution spectra should have deeper spectral features. Although the low-resolution spectra have shallower absorption features, I found that the low-resolution spectra could perform equally well because they simply afford a higher S/N and a larger wavelength coverage to compensate for the shallower features.

Comparing the pros and the cons of high-resolution and low-resolution spectroscopy, in Chapter 6, I will argue that for the next generation of large-scale spectroscopic surveys, it may be more advantageous to observe most stars in a low-resolution configuration and only collect a subset of high-resolution spectra for calibration purposes. This work is timely because the next generation

of surveys such as the 4MOST survey will consist of both high-resolution and a low-resolution settings. My study in Chapter 6 lays out a comprehensive theoretical basis to evaluate the advantages of employing each of these two configurations.

In summary, chemical tagging is an incredibly powerful approach that could open new windows to study the Milky Way's properties at earlier epochs. The promise of this technique motivated the collection of a large amount of data. My contribution has been to recognize challenges that currently prohibit chemical tagging (Chapter 2) and overcome these challenges. I have developed innovative ways of extracting statistical indicators from the elemental abundance space (Chapter 3) and more sophisticated methods of obtaining the maximal information from stellar spectra (Chapter 4–6). I am also working on other future directions (Chapter 7) that involve both improving stellar age estimates for millions of stars and exploiting kinematic information from Gaia. Each of these efforts is enormously valuable on its own, but my ultimate goal is too coordinate all of these efforts and eventually bring the full life cycle of the Milky Way into greater focus via chemical tagging.

2

Prospects for chemically tagging stars in the Galaxy

Author list of the original paper: Yuan-Sen Ting, Charlie Conroy, Alyssa Goodman

2.0 Abstract

It is now well-established that the elemental abundance patterns of stars holds key clues not only to their formation but also to the assembly histories of galaxies. One of the most exciting possibilities is the use of stellar abundance patterns as "chemical tags" to identify stars that were born in the same molecular cloud. In this paper we assess the prospects of chemical tagging as a function of several key underlying parameters. We show that in the fiducial case of 10^4 distinct cells in elemental abundance space and $10^5 - 10^6$ stars in the survey, one can expect to detect $\sim 10^2 - 10^3$ groups that are $\geq 5\sigma$ overdensities in the elemental abundance space. However, we find that even very large overdensities in elemental abundance space do not guarantee that the overdensity is due to a single set of stars from a common birth cloud. In fact, for our fiducial model parameters, the typical 5σ overdensity is comprised of stars from a wide range of clusters with the most dominant cluster contributing only 25% of the stars. The most important factors limiting the identification of disrupted clusters via chemical tagging are the number of chemical cells in the elemental abundance space and the survey sampling rate of the underlying stellar population. Both of these factors can be improved through strategic observational plans. While recovering individual clusters through chemical tagging may prove challenging, we show, in agreement with previous work, that different cluster mass functions (CMF) imprint different degrees of clumpiness in elemental abundance space. These differences provide the opportunity to statistically reconstruct the slope and high mass cutoff of CMF and its evolution through cosmic time.

2.1 Background

Despite decades of effort, we still lack a thorough understanding of how galaxies assemble and evolve over cosmic time. This is true not only for distant galaxies but also for our own Milky Way. In the current paradigm, galaxies such as the Milky Way form from smaller pieces (e.g., [182]), driven by the hierarchical growth of dark matter structures (e.g., [158,161]). Much of the most exciting phases of star formation and galaxy assembly appear to have taken place at early times, perhaps before $z \sim 2$. If true, this puts much of the most interesting phases of galaxy formation beyond direct detailed study. For this reason much effort has focused on reconstructing the past based on present-day observations of stars, in particular in the Galaxy. For example, studies of the Galactic stellar halo provides clues to the assembly history of dwarf galaxies (e.g. [57,182]). The properties of stars in the thin and thick disks provide clues to the formation history of these Galactic components. The abundance patterns of the most metal poor stars probe star formation and supernovae conditions during the first generation of stars. And the evolutionary histories of star clusters, both intact, dissolving, and long destroyed, offer clues not only into the star formation process (by reconstructing the CMF), but also the dynamical history of the Galaxy (e.g., [5,110,204]).

However, reconstructing disrupted star clusters is difficult because most star clusters dissolve quickly upon their formation due to dynamical interactions, such as intracluster n-body interaction and tidal stripping. In fact, most clusters are not expected to survive for more than 100 Myrs[124]. For this reason, the study of young embedded clusters (e.g., [16,27,111,160]) is typically restricted to the study of star formation conditions at the present time. Although most star clusters are quickly disrupted,

they retain their identity in kinematic phase space for a longer period of time. Several examples of clusters identified in phase space are known, such as HR1614, the Argus association and the Wolf 360 group (e.g., [37,49,52]), with an age of 2-3 Gyrs. This implies that at least some clusters can maintain their phase space identity for a few disk dynamical times. Within a few dynamical times these groups will phase mix with the background stars, which implies that the timescale over which groups can be identified in phase space is still a small fraction of the age of the Galaxy.

While dynamical information is mostly short-lived, elemental abundances are expected to leave a more permanent fossil record of star clusters. The idea of "chemical tagging", first proposed by Freeman & Bland-Hawthorn [69] (also see [21]), is to use elemental abundances to identify stars that are now widely separated in phase space to a common birth site. If such an association could be made, even for a small fraction of stars, it would provide an extraordinary new view into both the early star formation process and the subsequent dynamical history of the Galaxy.

Observations have shown that satellite galaxies exhibit different chemical evolution histories compared to stars either in the disk, bulge, or halo of the Galaxy (e.g., [127,159,196,201,202]). As a consequence, stars accreted into the Galaxy from different satellite systems should show distinct chemistry from e.g., disk stars. It has been proposed that these variations could be used in chemical tagging to find the remnants of disrupted satellite galaxies [69]. The possibility of reconstructing disrupted satellite galaxies via chemical tagging could for example provide important clues to the missing satellite problem [152].

Previous studies of high-resolution stellar spectroscopy were limited to a few hundred stars (e.g., [10,13,168]). The small samples restricted the possibility of chemical tagging for reasons that will

become clear in later sections. But this situation is rapidly changing. Recent and on-going large-scale surveys, such as GALAH[53], Gaia-ESO[164] and APOGEE[212] aim to observe $10^5 - 10^6$ stars with resolution $R > 20,000$ in order to measure $\sim 15 - 30$ elements for each star. These surveys were motivated, at least in part, by the idea of chemical tagging and the prospects for uncovering the distribution of stars in their n-dimensional elemental abundance space, spanned by the elemental abundances.

There are several conditions that must be met for chemical tagging to work (see[22,23,24,53,69] for details). First, clusters must be internally chemically homogeneous. Open clusters have been found to be chemically homogeneous at the level of $\sigma_{[X/Fe]} < 0.05$ dex (e.g.,[50,51,70,157,194]). Theoretical arguments from Bland-Hawthorn et al.[24] showed that the chemical signature within a protocloud should have sufficient time to homogenize before the first supernova goes off, for clusters with mass $10^5 - 10^7$ $M_\odot$. Simulations by Feng & Krumholz[63] showed that turbulent mixing, even for a loosely bound cluster, could homogenize the elemental abundances of a protocloud. Their simulations showed that turbulent mixing creates an intracluster chemical dispersion at least five times more homogenized than the protocloud. Both observations and theory agree that clusters less massive than $\sim 10^7$ $M_\odot$ should be chemical homogeneous, except perhaps for the confounding internal abundance trends observed in the light elements of all known globular clusters (e.g.,[38,137]), though many globular clusters show a high degree of chemical uniformity (e.g.,[174]) in all heavy elements.

In addition to cluster homogeneity, the existence of substantial cloud-to-cloud variation in elemental abundances is another requirement. For example, if all star clusters shared the same elemental abundances, it would not be possible to separate them in elemental abundance space. We know

that this condition is broadly satisfied given the sizable spread in abundance ratios in existing spectroscopic samples (e.g., [13,56]). Quantitatively, an important parameter is the volume of abundance space that is available for a particular survey. This volume depends both on Galactic chemical evolution and on the particular survey design. The latter is important both in determining the target sample and in the number of elements that can be spectroscopically measured. Combining the available chemical volume with the measurement uncertainty on individual abundances allows us to define the concept of the total number of distinct cells in elemental abundance space. As we will see below, this is a key concept in chemical tagging (see also [69]).

Ting et al. [195] presented an empirical estimate of cloud-to-cloud variation in elemental abundances. They performed principal component analysis and estimated that there are $7 - 9$ independent dimensions among the $\sim$ 25 elements that will be measured by surveys such as GALAH and Gaia-ESO, and $4 - 5$ independent dimensions for an APOGEE-like survey. From this one can estimate the number of distinguishable cloud-to-cloud variations in the elemental abundance space, denoted $N_{\rm cells}$. As discussed in detail in §2.3.5 below, the result is that modern surveys should be able to reach $N_{\rm cells} \sim 10^{3-4}$, at least, implying that there is a decent cloud-to-cloud variation.

The goal of this paper is to explore the prospects for identifying long disrupted star clusters based on their clustering in elemental abundance space. We follow Freeman & Bland-Hawthorn [69], Bland-Hawthorn et al. [23], and De Silva et al. [53] in identifying the global survey parameters and the shape of the CMF as key parameters. Our emphasis on the information contained in the distribution (i.e., clumpiness) of stars in elemental abundance space echoes the results found in Bland-Hawthorn et al. [23]. In the present work we consider a wide array of parameters in order to identify optimal

regions of parameter space for chemical tagging. In addition, for the first time we analyze the local properties of cells in elemental abundance space that appear as high sigma fluctuations and find that in many cases these high overdensities in elemental abundance space are not the result of a single star cluster but instead are comprised of stars from many distinct birth sites.

The rest of this paper is organized as follows. In §2.2 we review several basic arguments relevant for chemical tagging and in §2.3 we describe the model used in the present work. In §2.4 we present the results and discuss how these assumptions and survey strategies affect the chemical tagging detections. In §2.5 we discuss various caveats, limitations and future directions. We conclude in §2.6. It is difficult to present the full set of results from a multidimensional parameter space and so we urge readers to explore the online interactive applet [*] created in the course of this project (see Appendix 2.8 for details).

[*] www.cfa.harvard.edu/~yuan-sen.ting/chemical_tagging.html

2.2 Basic arguments

As we will show quantitatively below, the prospects for chemical tagging largely depends on the number of stars sampled per cluster. This number in turn primarily depends on the number of stars in the survey divided by the integrated star formation rate (SFR), over cosmic history, in the volume sampled by the survey. We will denote the former number as $N_\star$, the latter number as $\mathcal{M}_{\rm annulus}$. Ongoing and upcoming surveys are targeting primarily FGK stars, which have on average $\langle \mathcal{M} \rangle \approx 1 \mathrm{M}_\odot$. This implies that $N_\star$ stars in a survey corresponds to $N_\star$ in solar masses and therefore numerically $\mathcal{M}_{\rm annulus} \approx N_{\rm annulus}$. The ratio of $N_\star$ and $N_{\rm annulus}$ defines the sampling rate. In this section, we motivate why the sampling rate largely defines the number of stars sampled per cluster (see also[53]).

First, let's consider a simple case where there is no radial migration and stellar excursion, i.e., stars stay in the annulus in which they were born. The integrated SFR in the Solar annulus, with a survey width $\Delta R_{\rm survey} = \pm 3$ kpc, is $\sim 2 \times 10^{10}\,\mathrm{M}_\odot$ (see model detail in §2.3).[†] For a survey of 10^6 stars with $\langle \mathcal{M} \rangle = 1\,\mathrm{M}_\odot$, the sampling rate can thus be calculated to be $(10^6\,\mathrm{M}_\odot)/(2 \times 10^{10}\,\mathrm{M}_\odot) = 1/(2 \times 10^4)$. In other words, assuming all stellar mass (including stellar mass loss) is now fully mixed in the annulus, we would have only sampled, on average, $1/(2 \times 10^4)$ of the original zero age mass from each cluster. Thus, we would expect to observe, on average, only one star from a $2 \times 10^4\,\mathrm{M}_\odot$ cluster. If we define the "detection" of a cluster to include the identification of at least 10 stars, then for a survey of 10^6 random stars in the solar annulus we would be able to probe clusters more massive than $2 \times 10^5\,\mathrm{M}_\odot$.

[†]The survey width $\Delta R_{\rm survey}$ defines the Solar annulus by $|R - R_\odot| < |\Delta R_{\rm survey}|$. The survey width should not be confused with the line-of-sight depth from the Sun, which is $|\mathbf{R} - \mathbf{R}_\odot| < 3$ kpc.

In practice, the sample is affected by the process of radial migration (e.g., [24]). Some stars are migrated away from their birth annulus while others that were born outside the Solar annulus will now reside within the Solar annulus. In other words, the number of stars that could end up in the Solar annulus increases with radial migration (another way of thinking of this effect is that the effective volume of the Solar annulus increases as the strength of radial migration increases). Given that the number of stars in the survey stays the same, the sampling rate decreases with radial migration. For a fixed survey strategy, the minimum cluster mass that one can probe increases in the presence of radial migration.

We must also consider the fact that we have limited resolution in separating groups in terms of their elemental abundance variations due to measurement uncertainties on the abundances. Multiple clusters might share the same cell in elemental abundance space (e.g., [23]). If we assume a CMF over the range 50 $M_\odot$ to 10^6 $M_\odot$ and a CMF slope of -2 (see details in §2.3), for a sample of 10^6 stars, in simulations of the latter section, we found that there are about 10^5 clusters contributing to the elemental abundance space. Fully resolving clusters in elemental abundance space would require roughly as many distinct chemical cells [69], but it was argued in the §2.1 that the actual number of chemical cells spanned by the data may be $1-2$ orders of magnitude lower. This suggests that most cells in elemental abundance space will be occupied by many clusters, each with a small number of stars sampled per cluster. One of the key goals of this paper is to understand the distribution of clusters in elemental abundance space under different scenarios.

The simple calculations in this section already demonstrate that key parameters include the number of stars in a survey, $N_\star$, the geometry of the survey (via $\mathcal{M}_{\rm annulus}$), the strength of radial migra-

tion, the shape of the CMF (which sets the typical cluster size), and the number of cells in elemental abundance space (N_{cells}).

2.3 Model description

In this section we describe the ingredients of our model for the Milky Way in some detail. The model is spatially two dimensional (though we assume that stars are uniformly distributed in the azimuthal angle), time-dependent, and statistical in nature. For the present study we are only interested in the disk; the bulge and halo are not included in the model below. We do not follow dynamics nor do we include a treatment of chemical evolution (these will be subjects of future work). The present aim is to build a model that is computationally very fast to allow the exploration of a large multi-dimensional parameter space.

The model specifies the star formation history (SFH) and evolution in time of the size of the Milky Way disk and the gas mass distribution. We define the SFH to be the total SFR in the Milky Way as a function of cosmic time. These quantities are used to model the effects of radial migration and an evolution in the cutoff of the CMF. The model is illustrated in a flow chart in Fig. 2.1. Table 2.1 lists observational constraints that we employ to constrain the model. Free parameters in the model and their adopted fiducial values are listed in Table 2.2. The meaning of other important symbols in this paper are listed in Table 2.3. We now proceed to explain the details of the model.

2.3.1 Star formation history and radial size growth of the disk

The SFH in the Solar neighborhood, $\Sigma_{\rm SFR}(R_\odot, t)$, has been estimated by analyzing the color-magnitude diagram from the Hipparcos catalog, where $R_\odot$ is the Galactocentric radius of the Sun. Results from, for e.g., Hernandez et al.[90] and Bertelli & Nasi[15] showed a rather flat SFH near $R_\odot$, ranging from

Table 2.1: List of constraints in this study.

Property	Value	References
Galactocentric radius of the Sun, $R_\odot$	8 kpc	Ghez et al.[74], Gillessen et al.[75], Reid et al.[169]
Stellar surface density, $\Sigma_\star(R_\odot, z=0)$	$38\,M_\odot pc^{-2}$	Bovy & Rix[30], Flynn et al.[65], Zhang et al.[213]
Gas surface density, $\Sigma_{gas}(R_\odot, z=0)$	$13\,M_\odot pc^{-2}$	Flynn et al.[65]
Total stellar mass in the disk, $\mathcal{M}_\star(z=0)$	$4.5\times10^{10}\,M_\odot$	Binney & Tremaine[18], Bovy & Rix[30], Flynn et al.[65]
Halo virial mass, $\mathcal{M}_{halo}(z=0)$	$10^{12}\,M_\odot$	Kafle et al.[98], Klypin et al.[107] Wilkinson & Evans[206], Xue et al.[209]
Global SFR $(z=0)$	$0.5-2\,M_\odot yr^{-1}$	Chomiuk & Povich[44], Robitaille & Whitney[173] Veneziani et al.[200]
Solar neighborhood SFR, $\Sigma_{SFR}(R_\odot, t)$	$3-6\,M_\odot Gyr^{-1}pc^{-2}$	Bertelli & Nasi[15], Hernandez et al.[90]
Stellar disk scale length, $R_\star(z=0)$	2.2 kpc	Bovy & Rix[30]
SFR scale length, $R_{SFR}(z=0)$	2.6 kpc	Schruba et al.[180] on NGC 6946
Gas scale length, $R_{gas}(z=0)$	4.2 kpc	Schruba et al.[180] on NGC 6946
Radial size growth	$R_\star \propto \mathcal{M}_\star^{0.27}$	van Dokkum et al.[198]

Table 2.2: List of parameters in the model.

Parameter	Fiducial	Range
In-situ fraction, $f_{\mathrm{in-situ}}(\Delta R_{\mathrm{survey}} = \pm 1\,\mathrm{kpc})$	50%	$15\% - 100\%$
Survey width, $\Delta R_{\mathrm{survey}}$	$\pm 3\,\mathrm{kpc}$	$\pm 0.6 - 5\,\mathrm{kpc}$
CMF slope, α	-2.0	-1.5 to -2.5
CMF low mass cutoff, $M_{\mathrm{cluster}}^{\mathrm{min}}$	$50\,\mathrm{M}_{\odot}$	$10 - 100\,\mathrm{M}_{\odot}$
CMF high mass cutoff, $M_{\mathrm{cluster}}^{\mathrm{max}}$	see Fig. 2.2	see Fig. 2.2
Number of chemical cells, N_{cells}	10^4	$10^3 - 10^5$
Number of stars in the survey, $N_{\star}$	10^6	$10^4 - 10^6$

Table 2.3: Meaning of other important symbols in this paper that are not listed in Table 2.1 and 2.2.

Symbols	Meanings
$k_{\rm ch}$	Churning strength in the radial migration prescription
η	Gas fraction; the ratio of gas mass over total dynamical mass
$\sigma_{\rm [X/Fe]}$	Elemental measurement uncertainty in [X/Fe]
σ	Elemental measurement uncertainty along the elemental abundance space principal components
$N_{\rm dim}$	Number of independent/informative dimensions in elemental abundance space
$M_{\rm gas}$	Total gas mass in the Milky Way
$M_{\rm cluster}$	Zero age stellar mass of a star cluster
$M_{\rm annulus}$	Integrated SFR, over cosmic history, in the volume sampled by the survey
$N_{\rm annulus}$	Total number of stars (including stellar mass loss) in the volume sampled by the survey
N_i	Total number of stars sampled in a chemical cell
$N_{\rm mean}$	Average number of stars sampled per chemical cell
$N_{\rm cluster}$	Number of stars sampled from a cluster
$N_{\rm dominant}$	Number of stars sampled from the most dominant cluster in a chemical cell
local S/N	Number of stars sampled from the most dominant cluster over the rest in a chemical cell
$f_{\rm sub}$	Sampling rate of a certain stellar subpopulation

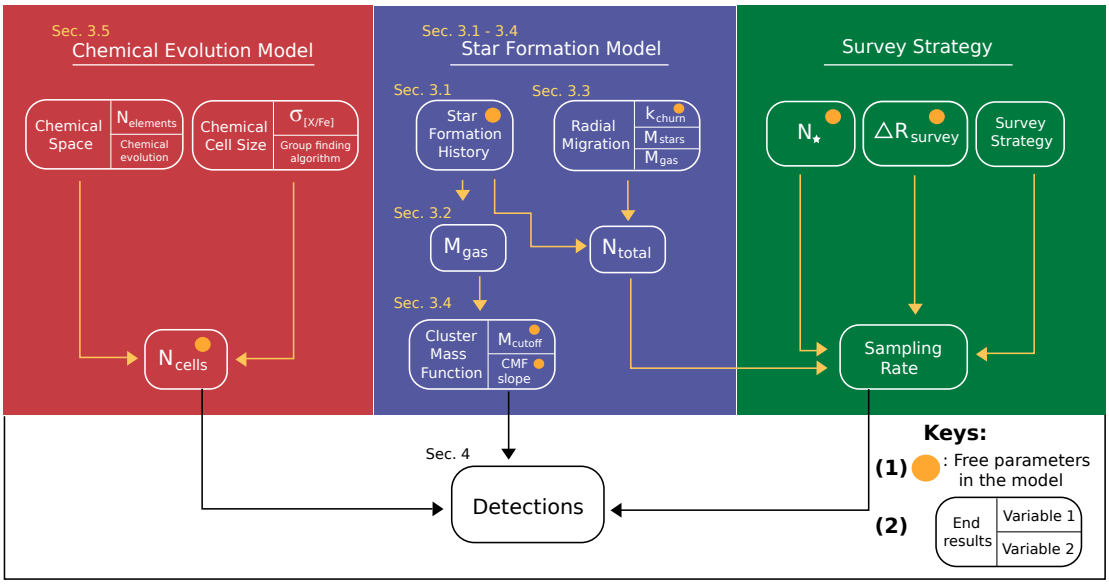

Figure 2.1: Flow chart demonstrating the main components of the model. Sections defining or describing certain components of the model are indicated in the chart.

$3 - 6\,M_\odot Gyr^{-1} pc^{-2}$ through $0 - 8$ Gyr in lookback time. The current total SFR in the Milky Way has been estimated to be $0.5 - 2\,M_\odot yr^{-1}$ from the study of young stellar objects (e.g., [44,173,200]).

In comparison to the Solar neighborhood, the Galactic global SFH is less well understood. We therefore adopt cosmological semi-empirical modeling from Behroozi et al. [12], assuming a Milky Way halo virial mass of $\mathcal{M}_{\rm halo} \equiv \mathcal{M}_{200} = 10^{12}\,M_\odot$ (e.g., [98,107,206,209]). Behroozi et al. [12] investigated the best-fitting global SFH as a function halo mass that is consistent with the observed galaxy stellar mass function, specific SFR, and cosmic SFR. We fit their result for Milky Way-like halos with a Schechter function,

$$\mathrm{SFR}[M_\odot yr^{-1}] = \mathcal{A}\,(t[\mathrm{Gyr}]/C)^B\,\exp(-t[\mathrm{Gyr}]/C). \quad (2.1)$$

Given a global SFH, the stellar mass evolution is calculated assuming the stellar population synthesis code from Conroy et al. [45], with a Kroupa IMF [114] from $0.08 - 125\,M_\odot$. The synthesis code is used to take into account secular stellar mass loss, etc. The normalization of the global SFH is further adjusted such that the present-day stellar mass (long-lived stars + remnant stars) agrees with observations, $\mathcal{M}_\star(z = 0) = 4.5 \times 10^{10}\,M_\odot$ (e.g., [18,30]). In this study, we only trace long-lived stars with $0.5 - 1.5\,M_\odot$ because almost all FGK stars in chemical tagging surveys are within this mass range.

We consider two SFH models in this study, with parameters from equation (1) as follows: (1) $\mathcal{A} = 1.4$, $B = 4.4$, $C = 1.3$, which is the best fitting SFH model from Behroozi et al.; (2) $\mathcal{A} = 15.5$, $B = 2$, $C = 2.7$, which produces better agreement with the observed $\Sigma_{\rm SFR}(R_0, t)$. Both models are

Table 2.4: Summary of the three model variants in this study.

Property	Optimistic	Fiducial	Quiescent
CMF cutoff	$\sim 10^7\,M_\odot$	$\sim 10^6\,M_\odot$	$10^5\,M_\odot$
Global SFR	Peaks in the past	More flat	More flat
$\Sigma_{SFR}(R_o, t)$	Too high in the past	Agrees with observations	Agrees with observations
Integrated SFR	The same	The same	The same

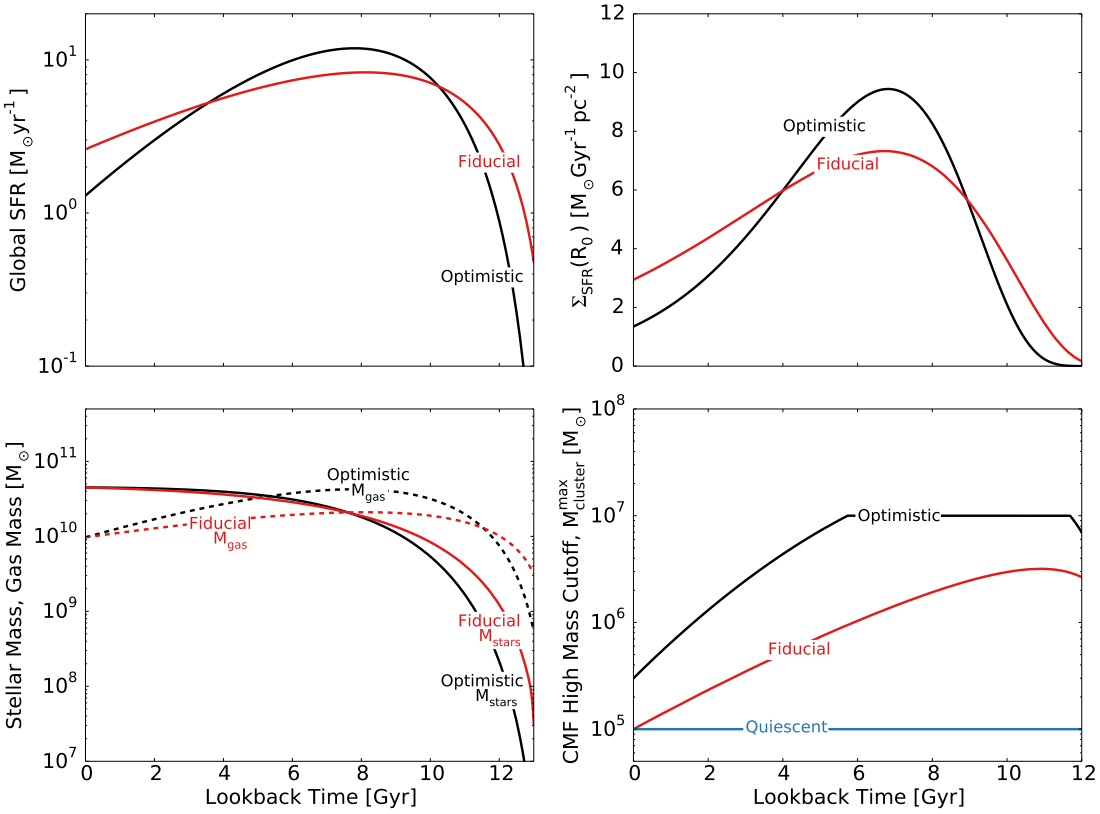

Figure 2.2: Bottom right panel: Evolution of CMF high mass cutoff. The CMF evolves according to Escala & Larson [59]. The CMF cutoff is the main property that defines the quiescent, fiducial and optimistic models that we will discuss throughout this study. For example, the optimistic CMF allows the formation of larger clusters ($\mathcal{M}_{\rm cluster} \sim 10^7\,M_\odot$). We adopt an upper limit of $\mathcal{M}^{\rm max}_{\rm cluster} = 10^7\,M_\odot$, above which clusters are not expected to be homogeneous. Bottom left panel: Stellar and gas mass evolutions. The gas mass at $z = 0$ is calculated from $\Sigma_{\rm gas}(R_0, z = 0) = 13\,M_\odot pc^{-2}$. The gas mass evolution is calculated from the global SFH, following a Kennicutt-Schmidt law with $\alpha_{\rm KS} = 1.5$. Top left panel: Global SFH models in this study, assuming $\mathcal{M}_{\rm halo} = 10^{12}\,M_\odot$ adjusted to produce $\mathcal{M}_\star(z = 0) = 4.5 \times 10^{10}\,M_\odot$. The two SFHs have the same integrated SFR. The SFHs mainly come into play in determining the gas mass evolution and subsequently the CMF cutoff evolution. Since the quiescent CMF cutoff is constant through cosmic time without evolution, employing the optimistic SFH or fiducial SFH for the quiescent model does not change its results as they have the same integrated SFR. We choose to follow the fiducial SFH for the quiescent model as it fits the $\Sigma_{\rm SFR}(R_0, t)$ better. Top right panel: $\Sigma_{\rm SFR}(R_0, t)$ calculated from the global SFHs.

within the uncertainty quoted by Behroozi et al. We adopt the latter as the fiducial model and the former to be the optimistic model (see Fig. 2.2 and Table 2.4). The former coins the term "optimistic model" as its more highly peaked SFR entails a higher total gas mass (see §2.3.2). The higher total gas mass in turn predicts a larger cluster high mass cutoff (see §2.3.4) than the "fiducial model." We emphasize that while the optimistic and fiducial models assume different SFHs, the integrated SFRs of these models over cosmic time are the same. Since the total integrated SFRs are the same, they both produce the same $M_\star(z=0)$ and $\Sigma_\star(R_0, z=0)$. Therefore, the sampling rate is the same for both cases. The global SFR and $\Sigma_{\rm SFR}(R_0, t)$ in these two models are compared in the upper panels in Fig. 2.2. The main differences of these models are summarized in Table 2.4 (the "quiescent model" will be defined in §2.3.4).

With the stellar mass evolution in hand, we then derive the radial size growth of the Milky Way using the empirical relation from van Dokkum et al.[198]. By studying the evolution of galaxies at a fixed comoving number density at different redshifts, van Dokkum et al.[198] found that the effective radius $R_\star$ of Milky Way-like galaxies grow with the total stellar mass according to the relation $R_\star \propto M_\star^{0.27}$.

Finally, to fully specify the star formation at different radii, we also require the star formation scale length, $R_{\rm SFR}$, and its evolution. Unfortunately, determining $R_{\rm SFR}$ for the Milky Way is observationally challenging. Therefore, we resort to $R_{\rm SFR}$ from extragalactic studies where the external vantage point provides an easier measurement of scale lengths. NGC 6946 has long been thought to be a Milky Way counterpart (e.g.,[103]). We find the SFR and the (atomic and molecular) gas mass of NGC 6946 from Schruba et al.[180] can be fitted with an exponential model. We find scale lengths

$R_{\rm SFR}(z=0) = 2.6\,{\rm kpc}$ and $R_{\rm gas}(z=0) = 4.2\,{\rm kpc}$, which we adopt in our model of the Milky Way. To compute the evolution $R_{\rm SFR}(z)$ and $R_{\rm gas}(z)$ through cosmic time, we assume all scale lengths trace the stellar effective radius. We find that this adopted $R_{\rm SFR}(z)$ leads to a stellar disk scale length of $R_\star(z=0) = 2.2\,{\rm kpc}$ and $\Sigma_\star(R_0, z=0) = 38\,{\rm M_\odot pc^{-2}}$. These values agree with existing observations[30,65,213]. Furthermore, the model implies $R_{\rm gas}(z=0) \simeq 2R_\star(z=0)$, agreeing with Bovy & Rix[30].

2.3.2 Gas mass distribution and evolution

The mass of gas in the disk comes into play in two aspects of the model, namely the radial migration prescription and the CMF evolution. We assume $\Sigma_{\rm gas}(R_0, z=0) = 13\,{\rm M_\odot pc^{-2}}$[65], which, when combined with $R_{\rm gas}(z=0) = 4.2\,{\rm kpc}$, yields a total gas mass of $\mathcal{M}_{\rm gas}(z=0) = 9.7\times10^9\,{\rm M_\odot}$. We then estimate the redshift evolution of the gas mass $\mathcal{M}_{\rm gas}(z)$ by inverting the Kennicutt-Schmidt relation with $\alpha_{\rm KS} = 1.5$ and the SFR evolution described in the previous section. The distribution of gas is fully specified by $\mathcal{M}_{\rm gas}(z)$ and $R_{\rm gas}(z)$. The total stellar mass and the total gas mass evolution are shown in the bottom left panel in Fig. 2.2. For this work we do not need to specify the disk scale height because all quantities of interest are related to surface mass densities.

2.3.3 Radial migration

Radial migration describes the phenomenon of stars in the disk moving, either inward or outward, in radius from their birth radius. Studies of processes giving rise to radial migration have a long history. In the past decade, radial migration has gained increasing attention as playing a key role in

driving the chemodynamical evolution of the Milky Way (e.g., [24,55,87,149,178,183]). Due to its role in changing stellar orbiting radii, radial migration provides tentative explanations to some observational puzzles. For example, the upturn in the stellar population age at the outer part of some galaxies (e.g., [9,214]), the wide range of stellar metallicity in the Solar neighborhood (e.g., [87,178]); and perhaps even the formation of the thick disk (e.g., [133]) can be explained by appealing to the process of radial migration.

An important physical process giving rise to radial migration is known as "churning"[183]. In the process of churning, stars that co-rotate with transient non-axisymmetric features can increase their angular momentum while maintaining the ellipticity of the orbit, effectively bumping stars from an orbiting radius to the other. Schönrich & Binney[178] proposed a simple analytic formula for churning that we will adopt in this study. In this prescription, the probability of moving from the i-th to the j-th annulus, P_{ij}, where $j = i \pm 1$, is given by

$$P_{ij} = k_{\rm ch} \frac{\mathcal{M}_j}{\mathcal{M}_{\rm max}}, \tag{2.2}$$

where $\mathcal{M}_j$ denotes the total (stellar + gas) mass of the j-th annulus and $k_{\rm ch}$ is a free parameter governing the strength of the churning.

In the present work we discretize the model galaxy into annuli with width of 0.2 kpc and apply the churning exchange every 0.5 Gyr. We define in-situ fraction, $f_{\rm in-situ}$, as the fraction of stars that were born in-situ in a Solar annulus with $\Delta R_{\rm survey} = \pm 1$ kpc. Clearly, $f_{\rm in-situ}$ depends on the choice of $\Delta R_{\rm survey}$. We choose $\Delta R_{\rm survey} = \pm 1$ kpc to calculate the in-situ fraction, instead of our fiducial

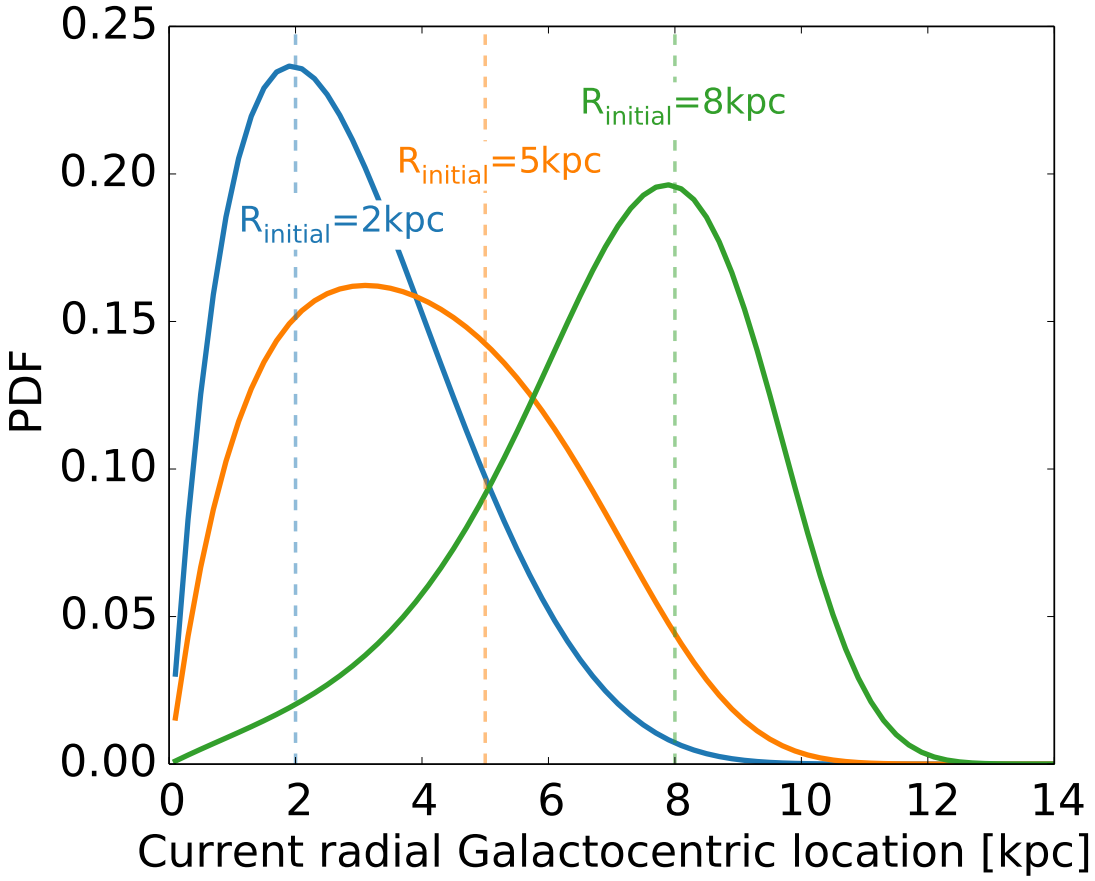

Figure 2.3: Probability of position of a star after evolving over 13 Gyr, assuming $f_{\mathrm{in-situ}} = 50\%$. The solid lines show the final positions, whereas the dashed lines show the corresponding initial positions.

value ± 3 kpc in the model for ease of comparing to hydrodynamics simulations (e.g., [175]). We note that the free parameter $k_{\rm ch}$ maps directly into the variable $f_{\rm in-situ}$, and we choose to express the effect of radial migration in terms of the latter value. We consider a range of $k_{\rm ch}$ corresponding to $f_{\rm in-situ} = 15\% - 100\%$ and we choose $f_{\rm in-situ} = 50\%$ to be the fiducial value, as suggested by simulations (e.g., [81,175]). To illustrate the radial migration prescription adopted in this study, solid lines in Fig. 2.3 show the PDF of the final position of a star after 13 Gyr of evolution starting from various initial positions.

In addition to churning, scattering, e.g., from interactions with molecular clouds, can also diffuse stars from their birth radii. This scattering is known as "blurring" [183]. For simplicity, we do not include blurring in the model. However, we note for our purposes only the fraction $f_{\rm in-situ}$ is important; the details of migration, either through churning or blurring are largely irrelevant in this study.

2.3.4 Cluster mass function evolution

We have discussed in §2.2 that the number of stars sampled per cluster is governed primarily by the sampling rate and the in-situ fraction. However, knowing the detections per cluster is insufficient. To determine the number of detectable groups, we also need to understand the relative number of massive clusters compared to their smaller counterparts. Therefore, the CMF is another key factor (see also [23]). In this study, we assume a CMF that is characterized by a power law slope α, high mass

cutoff $M_{\rm cluster}^{\rm max}$ and low mass cutoff $M_{\rm cluster}^{\rm min}$, where

$$\frac{{\rm d}N}{{\rm d}M} \propto M^{-\alpha}. \tag{2.3}$$

Note that cluster masses refer to zero age masses; clusters will lose at least a factor of two mass after a Hubble time due to stellar evolution effects and the evaporation of stars.

Lada & Lada [124] analyzed young embedded clusters within 2.5 kpc from the Sun and found a CMF slope $\alpha \approx -2.0$. We take this as the fiducial value in the model. The fact that $\alpha \approx -2$ is important in chemical tagging. In this case, the total mass in a survey sample coming from clusters within a mass bin δM, can be calculated to be

$$M\,{\rm d}N/{\rm d}M\,\delta M = M^2\,{\rm d}N/{\rm d}M\,\delta\log M \propto \delta\log M. \tag{2.4}$$

Quantitatively, this means that the chance of sampling a star from the logarithmic bin $[10\,{\rm M}_\odot, 100\,{\rm M}_\odot]$ is the same as the probability of sampling from the logarithmic bin $[100\,{\rm M}_\odot, 1000\,{\rm M}_\odot]$, and so forth. Since we adopt a maximum cluster mass $M_{\rm cluster}^{\rm max} = 10^5 - 10^7\,{\rm M}_\odot$ in this model, we have $4-6$ orders of dynamical range in the cluster mass. This large range of cluster mass implies that clusters with $[10\,{\rm M}_\odot, 100\,{\rm M}_\odot]$ contribute only $\sim 10\% - 25\%$ of the total stellar mass. Lada & Lada [124] determined that the CMF low mass cutoff occurs around $M_{\rm cluster} = 50\,{\rm M}_\odot$, which we will adopt as the fiducial value.

Although not shown in this paper, we find that changing the low mass cutoff to $10\,{\rm M}_\odot$ or $100\,{\rm M}_\odot$

has a negligible effect on the results. First, as we have discussed, the small clusters only contribute $\sim 10\% - 25\%$ of the population. Furthermore, changing the low mass cutoff will alter the number of small clusters and hence the background in each cell, however since the signal is concentrated in $\sim 0.1\% - 1\%$ of the chemical cells, as we will show in §2.4.2, only $< 1\%$ of this background change is affecting the signal.

The high mass cutoff $M_{\rm cluster}^{\rm max}$ has a dramatic effect on the results because massive clusters dominate the signal, as shown in later sections. We therefore consider several different scenarios for the high mass cutoff and its evolution with redshift (see the lower right panel of Fig. 2.2). The largest open clusters observed in the Milky Way appear to be Westerlund 1 (e.g., [35]), Berkeley 39 (e.g., [34]) and Arches (e.g., [60]), with a mass few times of 10^4 $M_\odot$. Noting the fact that the cluster could have gone through a period of rapid mass loss in its formation phase (e.g., [124]), we adopt $M_{\rm cluster}^{\rm max} \simeq 10^5$ $M_\odot$ at $z = 0$ as the nominal mass cutoff at $z = 0$ in the Milky Way disk.

A number of arguments suggest that the CMF high mass cutoff could have been higher in the past. For instance, the existence of massive globular clusters with surviving mass of $10^{4.5} - 10^{6.5}$ $M_\odot$ (e.g., [83]) suggests that early conditions in the Galaxy favored the formation of more massive clusters. Observations of high-redshift disk galaxies also suggests a high frequency, relative to $z = 0$, of very massive gas clumps of $10^7 - 10^9 M_\odot$ (e.g., [66,72,97,132]).

Escala & Larson [59] provided a simple model for the maximum cluster mass by studying gravitational instability in disks, similar to Toomre's classic analysis [197]. They calculate the maximum unstable mass to be $M_{\rm cluster}^{\rm max} = \Sigma_{\rm gas}(\lambda_{\rm rot}/2)^2$, where $\lambda_{\rm rot} = \pi^2 G\Sigma_{\rm gas}/\Omega^2$. From this formula, they further found that the maximum cluster mass can be determined by the gas fraction η (i.e., gas mass to the

total gravitational mass) and the total gas mass $M_{\rm gas}$ alone, where

$$M_{\rm cluster}^{\rm max} \propto M_{\rm gas}\eta^2. \tag{2.5}$$

The normalization of this formula depends on a variety of unknown parameters and so we choose instead to fix the normalization by hand at $z = 0$. The dynamics of the Milky Way disk can be explained without appealing to dark matter, at least within the Solar radius. We therefore ignore the influences of dark matter when computing the upper mass cutoff, i.e., we define $\eta = M_{\rm gas}/(M_{\rm gas} + M_\star)$. The evolution of $M_{\rm gas}$ and $M_\star$ follow the discussion in §2.3.1 and 2.3.2.

We consider three scenarios for the evolution of the upper mass cutoff, which we will denote as the quiescent, fiducial and optimistic models (see Fig. 2.2). In the quiescent model, we consider the fiducial SFH and fix $M_{\rm cluster}^{\rm max}(z) = 10^5\ {\rm M}_\odot$ through cosmic time. In the fiducial and optimistic cases, we consider the SFHs labeled as fiducial and optimistic in Fig. 2.2 and allow $M_{\rm cluster}^{\rm max}(z)$ to evolve. We set $M_{\rm cluster}^{\rm max}(z = 0) = 10^5\ {\rm M}_\odot$ for the fiducial case, and $M_{\rm cluster}^{\rm max}(z = 0) = 3 \times 10^5\ {\rm M}_\odot$ for the optimistic case. We use the term "optimistic" because this model allows the formation of very massive clusters, which is favorable for chemical tagging. Finally, we impose a maximum upper limit of $10^7\ {\rm M}_\odot$. Clusters with mass larger than this cutoff are unlikely to be homogeneous[24] in their elemental abundances due to self-enrichment. The evolution of $M_{\rm cluster}^{\rm max}(z)$ in these three cases are plotted in the bottom right panel in Fig. 2.2. The main differences of these three CMF models are summarized in Table 2.4. The range of CMFs we consider is similar to the range explored by Bland-Hawthorn et al. [23], although the authors do not consider a time-dependent CMF as we do here (for

the optimistic and fiducial models).

2.3.5 Elemental abundance space

The last model ingredient is multi-dimensional space of elemental abundances, often referred to as the "elemental abundance space." The elemental abundance space is spanned by the elemental abundances [Fe/H], [X_1/Fe], . . ., [X_n/Fe], where X_1 to X_n are n different elements measured. Since stars that were born together are expected to share the same abundances, they should reside at the same location in elemental abundance space.

As we will show below, the number of chemical cells in elemental abundance space $N_{\rm cells}$ is a key variable in chemical tagging. To understand its importance, let's consider the case where we have an infinite number of chemical cells, in other words we have infinite resolution in the elemental abundance space. In this case, all clusters from various birth sites can be easily identified. However, as the number of cells decreases, the probability that two clusters occupy the same chemical cell increases. In this case, the smaller clusters (in terms of the number of stars sampled per cluster) become contaminants in the detection. They dilute the number of genuine members of the massive clusters.

$N_{\rm cells}$ depends on two ingredients: (a) The elemental abundance space spanned by the sample. This volume is governed by Galactic chemical evolution and survey design, including the number of elements of each star the survey can extract. Note that the volume does not scale in a simple way with the number of elements measured because of the strong correlation between various subgroups of elements. (b) The abundance measurement uncertainty $\sigma_{\rm [X/Fe]}$, which sets the volume of each cell. Regarding (b), in this study, we assume that the width of chemical cell is 1.5 σ, i.e., two different

distinct groups in elemental abundance space can be recovered if their separation is larger than 1.5 σ, where σ represents the uncertainties along the principal components/independent dimensions.[‡] Note that, given an elemental abundance space of $N_{\rm dim}$ (independent) dimensions, the volume of each cell is proportional to $\sigma^{N_{\rm dim}}$. As a consequence, the number of cells is extremely sensitive to the abundance measurement uncertainties. We therefore stress that not only are small uncertainties favorable, but also accurate measurement of the uncertainties and their covariances are equally important.

The elemental abundance space spanned by the sample, in principal, can be modeled through chemodynamical simulations. However, we note that chemical evolution models are still rather uncertain for many elements and are often limited to a relatively small number of elements (e.g., [109,147]). Kobayashi et al. [108] include more elements, but they do not include neutron capture elements. Therefore, we are not aware of an existing chemical evolution model that encompasses all $\sim$ 25 elements measured by the GALAH and Gaia-ESO surveys. For these reasons, and for simplicity, we choose here to adopt empirical results in estimating the volume and defer a chemical modeling approach to future work.

We make use of the estimated elemental abundance space volume of Milky Way disk stars from Ting et al. [195] (also see [6] for a similar study on bulge stars). Using principal components analysis, Ting et al. [195] searched for directions in the elemental abundance space that are orthogonal to each other and contain most variances of the data. These principal components define a n-dimensional

[‡] As these component vectors are comprised of various elements, the uncertainties along these directions require the full covariance matrix of $\sigma_{[X/Fe]}$.

cube spanned by the data. By definition, the number of cells is the volume of the cube divided by the volume spanned by each cell. As for the latter, given the assumption that the width of chemical cell is $1.5\,\sigma$, the volume of the chemical cell is $(1.5\,\sigma)^{N_{\rm dim}}$. The volume of the n-dimensional cube can be estimated from the width of edges in each dimension, which can be calculated from the principal components axial ratios. Here we use the axial ratios of the principal components to estimate the volume that will be spanned by the GALAH data, as an example. The axial ratios of the first 6 dimensions are 1, 0.4, 0.25, 0.25, 0.1, 0.1. Apart from the obvious additional dimension from [Fe/H], Ting et al.[195] speculated that there should be another dimension associated with neutron capture elements. This last dimension was not available in the data analyzed by Ting et al. but will be probed by both GALAH and Gaia-ESO.

We can safely assume that the first principal component spans at least 1.5 dex as it is the diagonal direction of the 17 dimension in study. Let's further assume that [Fe/H] and both of the additional dimensions span 1 dex, and the uncertainties along the independent dimensions are $\sigma = 0.1$ dex. A simple calculation using the axial ratios yields: $N_{\rm cells} = (1.5\,{\rm dex})^6 \times (1 \cdot 0.4 \cdot 0.25 \cdot 0.25 \cdot 0.1 \cdot 0.1) \times (1\,{\rm dex})^2/(1.5\,\sigma)^8 = 10^4$ for GALAH. The Gaia-ESO survey spans a comparable list of elements and should therefore contain a similar number of $N_{\rm cells}$. An APOGEE-like survey should have $2-3$ fewer independent dimensions than GALAH[195]. All other parameters being the same, APOGEE should have $N_{\rm cells} \sim 10^3$.

The above calculations are simple estimates for the number of chemical cells that could easily be off by an order of magnitude. Hence, in the analysis below we consider a wide range in this important parameter, ranging from $10^3 - 10^5$.

2.4 Results

With the model for the Milky Way disk stars now in hand, we turn to using that model to explore what ongoing and future massive spectroscopic surveys of stars may expect to reveal in the context of chemical tagging. In §2.4.1 we investigate how many stars we expect to sample from the same cluster for different number of stars surveyed and both with and without the effect of radial migration. The main results are presented in §2.4.2, where we simulate the number of detectable groups in different scenarios. We study how observations of the distribution of stars in elemental abundance space may encode information on the shape of the CMF. We also investigate whether each detectable group in elemental abundance space is dominated by a single cluster or is comprised of a wide range of clusters.

2.4.1 Number of stars sampled per cluster

In this section we study the number of stars sampled per cluster for several idealized surveys. In particular, we are interested in how many stars will be sampled per cluster after the cluster is dispersed and mixed with the background sea of other clusters, and how the process of radial migration influences the sampling. Note that since we consider quantities as a function of cluster mass in this section, for a fixed $\Sigma_\star(R_\odot, z = 0)$ the results will be independent of the CMF. However, the results do depend on $\Delta R_{\rm survey}$ and $f_{\rm in-situ}$ as these parameters change the sampling rate and the radial migration prescription. Here we assume $\Delta R_{\rm survey} = \pm 3$ kpc and $f_{\rm in-situ} = 50\%$.

In Fig. 2.4, we plot the number of stars sampled per cluster as a function of cluster mass. The

solid lines show the median of the results in each cluster mass bin and the shaded color regions show the 1σ range. In the top panels, we consider the case without radial migration, i.e., stars stay in the orbiting radii that they formed, while the bottom panels show the case with radial migration. The left and right panels show results for $N_\star = 10^5$ and $N_\star = 10^6$. A horizontal line at $N = 10$ stars is meant to serve as a reference point.

While the results in Fig. 2.4 clearly show that the typical sampling rate (within $\pm 1\sigma$ range) per cluster is quite low, except in the case of large $N_\star$ and high cluster mass, we emphasize that the distribution of the number of stars sampled per cluster has a long tail toward high values. We return to this point below.

In the limit where there is no radial migration, the average number of stars (with $\langle \mathcal{M} \rangle = 1\,\mathrm{M}_\odot$) sampled per cluster can be analytically derived (see also[53]). The number of stars sampled per cluster is simply

$$N_{\rm cluster} = \mathcal{M}_{\rm cluster} \frac{N_\star}{\mathcal{M}_{\rm annulus}}. \tag{2.6}$$

Recall that $\mathcal{M}_{\rm annulus}$ is the total integrated SFR in the Solar annulus and $N_\star/\mathcal{M}_{\rm annulus}$ is proportional to the sampling rate. This analytic model is shown in the top panels of Fig. 2.4 and clearly predicts very well the results of the simulations. The grey shaded region demarks the 1σ from this analytic model.

Although illustrative, this analytic formula is unfortunately not applicable when radial migration is included. First, radial migration increases the number of stars that could end up in the Solar annulus, which has the effect of increasing the effective volume of the survey. We can define an effective

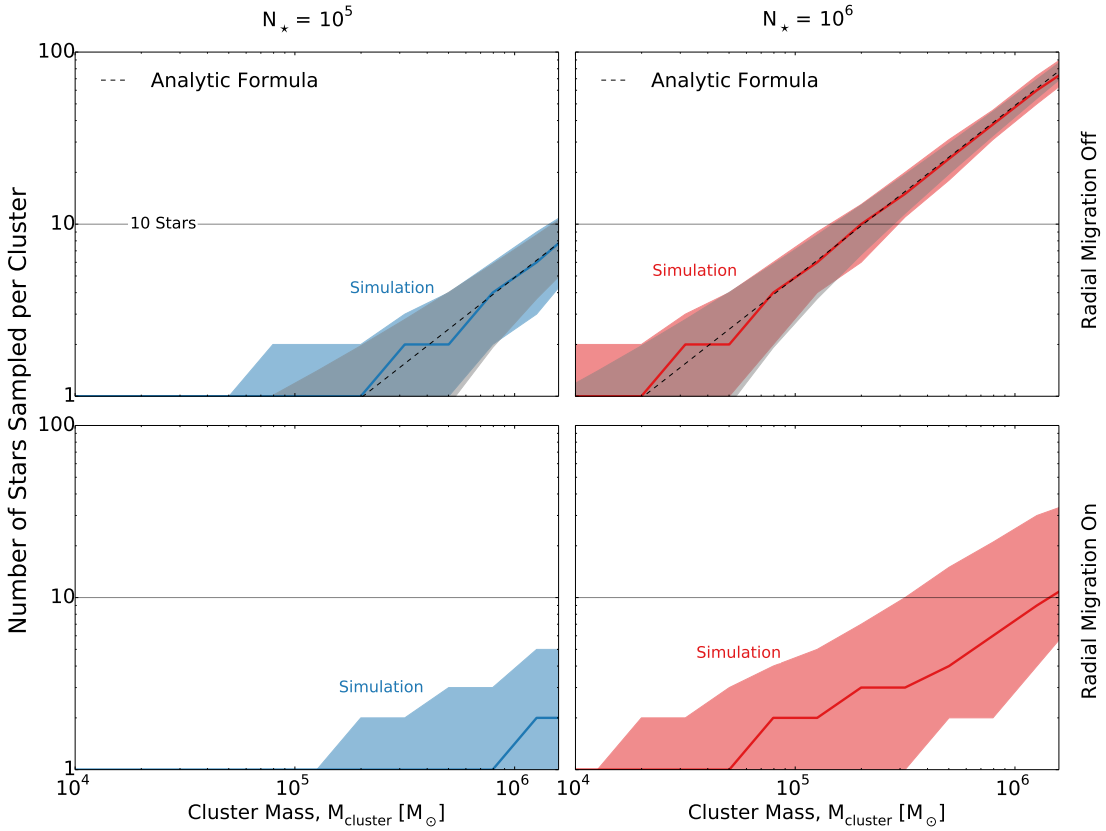

Figure 2.4: Number of stars sampled per cluster as a function of cluster mass, assuming $\Delta R_{\rm survey} = \pm 3\,{\rm kpc}$. The left panels assume $N_\star = 10^5$, whereas the right panels assume $N_\star = 10^6$. The top panels show the cases where there is no radial migration ($f_{\rm in-situ} = 100\%$), while the bottom panels illustrate the cases with radial migration and an in-situ fraction $f_{\rm in-situ} = 50\%$. The solid lines show the median and the shaded regions in color show the 1σ range of the results from simulations. In the limit of no radial migration, the number of stars sampled per cluster can be predicted analytically from equation (6). The predictions from the analytic formula are shown in dashed lines and gray shaded regions. The 1σ range from simulations follows very well the Poisson expectations. However, the analytic formula does not work in the case with radial migration because ex-situ clusters tend to have fewer stars sampled and bring down the number (see text and Fig. 2.5 for details).

radius of the observed annulus to be the mean distance,

$$R_{\text{effective}} = \frac{1}{n} \sum_{i=1}^{n} |R_{i,\text{birth}} - R_{\odot}|, \tag{2.7}$$

where we sum over all the stars in the Solar annulus at the present-day. This equation takes into account the fact that, with radial migration, the actual sampled volume is larger than the observed volume because $|R_{i,\text{birth}} - R_{\odot}| \geq |\Delta R_{\text{survey}}|$. The effective integrated SFR $\mathcal{M}'_{\text{annulus}}$ within this effective volume is strictly larger than the one without radial migration due to the migration of ex-situ population, and therefore the number of stars per cluster will generally be lower than in the case without radial migration.

Moreover, clusters that were born ex-situ are unlikely to have a significant number of stars migrated into the Solar annulus. As shown in Fig. 2.3, while stars born 5 kpc from the Galactic center can move into the Solar annulus at $R_{\odot} = 8$ kpc, only a small fraction of this population is in the Solar annulus. Fig. 2.3 suggests that most of the ex-situ stars, even from massive clusters, will tend to enter as "contaminants" in the sense that they will have only $\mathcal{O}(1)$ stars sampled per cluster. In addition, some stars that were born in-situ will migrate outside the Solar annulus, further diluting the number of members of in-situ clusters. All of these effects work in the same direction of reducing the number of stars per cluster compared to a model without radial migration.

In Fig. 2.5 we show the distribution of the number of stars sampled per cluster for two choices of $N_{\star}$. This figure shows the distribution for a vertical slice in Fig. 2.4 at a cluster mass of $\sim 10^6$ $M_{\odot}$. By separating the in-situ and ex-situ populations, Fig. 2.5 shows that the ex-situ population has on

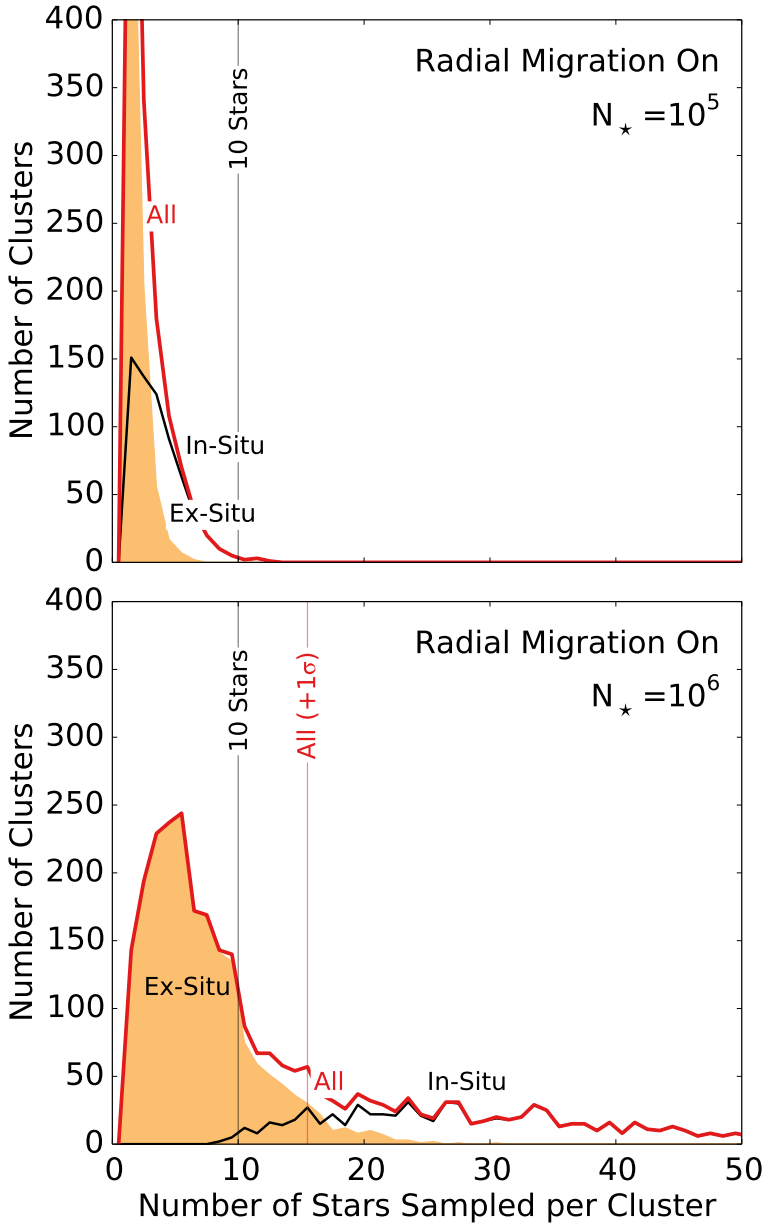

Figure 2.5: Distribution of the number of stars sampled per cluster for $\mathcal{M}_{\rm cluster} = (0.7 - 1.3) \times 10^6\,{\rm M}_\odot$. The top panel shows the result for $N_\star = 10^5$ and the bottom panel shows $N_\star = 10^6$. We assume $\Delta R_{\rm survey} = \pm 3\,{\rm kpc}$ and $f_{\rm in-situ} = 50\%$. We separate the cluster population into two - the in-situ and ex-situ populations. The ex-situ clusters have much smaller number of stars sampled per cluster compared to the in-situ population, indicating that ex-situ stars are mostly contaminants in chemical tagging. The red vertical line shows the 75 percentile of the combined results from in-situ and ex-situ clusters.

average a much smaller number of stars sampled per cluster, in agreement with the arguments described above. Although not shown, we checked that the in-situ population is only marginally influenced by radial migration — only a small fraction of in-situ stars leave the Solar annulus. The mild effect on in-situ clusters is likely due to the fact that we consider a fairly large Solar annulus width of $\Delta R_{\rm survey} = \pm 3$ kpc. In the radial migration prescription in this study, a typical radial migration length is $\sim$ 2 kpc, which is smaller than $|\Delta R_{\rm survey}|$. Although the typical radial migration length is still largely unconstrained from observations, some studies have suggested that since $R_{\odot}$ is beyond the outer Limblad resonance of the Galactic bar[54], a typical radial migration length is < 2 kpc[81].

Another feature evident in Fig. 2.5 is the tail of clusters with a large number of stars sampled per cluster. This highlights that median statistics are not sufficient to capture the full variety of expected behavior. These rare clusters may end up being the most valuable from the standpoint of chemical tagging as they should stand out as strong concentrations of stars in elemental abundance space. The following section explores this effect in detail.

2.4.2 Finding and counting groups in elemental abundance space

Observational uncertainties on elemental abundances impose a finite resolution in elemental abundance space that can have important consequences for chemical tagging[23]. In this section, we simulate observational results by studying detections on a chemical cell-by-cell basis. In the following, for each generated sample, we distribute sampled clusters uniformly (on average) into $N_{\rm cells}$ cells. We perform Monte Carlo simulations and plot the result of the mean from 100 realizations. By jack-knife estimation, we find that the uncertainties from these 100 realizations are $\ll 10\%$ for

$N_\star = 10^5 - 10^6$.

We define several terms that will be important in this section. A cell that contains a high density of stars compared to the average number of stars per chemical cell defines a "group." We distinguish between "group" and "cluster" because the former can be comprised of multiple clusters. The cluster with the most stars sampled in each cell is referred to as the dominant cluster. Stars from the dominant cluster define the "local signal." The rest of the stars in the cell are referred to as "local noise."

Identifying groups in elemental abundance space

If we were to randomly distribute $N_\star$ stars into $N_{\rm cells}$ chemical cells, the number of stars per cell should follow a Poisson distribution with a mean $N_{\rm mean} = N_\star / N_{\rm cells}$ and a 1σ range of $\sqrt{N_{\rm mean}}$. Since stars are born in clusters, there will be clumping in elemental abundance space that is larger than Poisson expectations. The degree of clumpiness depends on several factors, chief among them is the form of the CMF[23].

Operationally we define a cell as containing a "detected" group of stars if that cell deviates from Poisson expectations by at least 5σ and the total number of stars in that cell > 1. Fig. 2.6 shows the deviations from Poisson statistics for different CMFs and numbers of stars in the survey. In the right panel, we assume $N_\star = 10^6$. In this case, both the fiducial and optimistic CMFs show substantial numbers of cells exceeding 5σ from the average. By contrast, when $N_\star = 10^5$ (left panel), only the optimistic CMF shows substantial deviation from Poisson expectations.

Fig. 2.6 demonstrates that the deviation from Poisson is minimal for a quiescent CMF. This lack

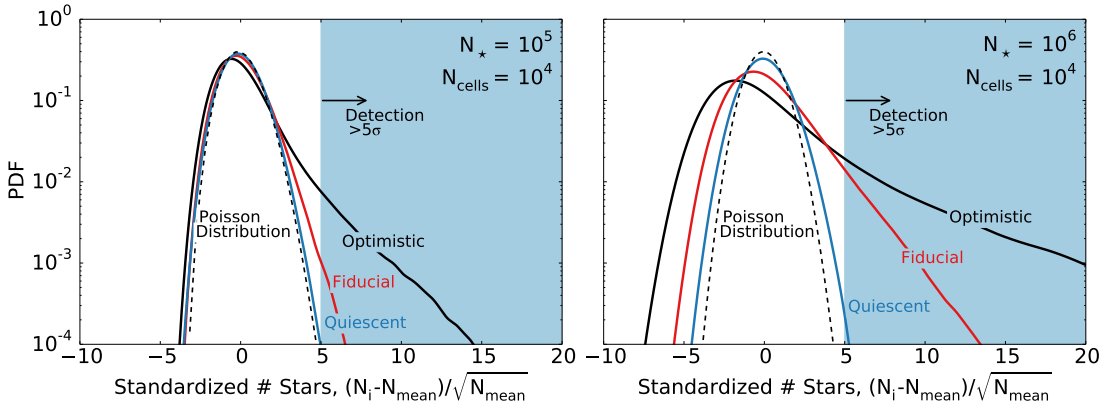

Figure 2.6: Standardized number of stars in each cell compared to a Poisson distribution, where the mean of Poisson distribution is $N_{\rm mean} = N_\star/N_{\rm cells}$ and the standard deviation follows $\sigma = \sqrt{N_{\rm mean}}$. Cells in which the number of stars sampled exceeds 5σ are considered as detectable groups. The y-axis shows the probability of a detectable group having a certain deviation from the Poisson distribution, quantified by the standardized number of stars. The integral under each curve is one. Unless stated of otherwise, we assume fiducial values for all the model parameters, as listed in Table 2.2. Different CMFs show different degrees of deviation from Poisson statistics. The clumpiness of the elemental abundance space may therefore be a useful tool to probe the underlying CMF.

of deviation is not unexpected because clusters with $M_{\rm cluster} < 10^5\,M_\odot$ have $\mathcal{O}(1)$ stars detected per cluster even for $N_\star = 10^6$ (see Fig. 2.4). Hence, randomly distributing clusters in $N_{\rm cells}$ cells for a quiescent CMF is close to randomly distributing $N_\star$ in $N_{\rm cells}$ cells.

Fig. 2.6 also shows that the distribution of deviations can be a sensitive probe of the CMF. CMFs with a higher mass cutoff produce more clumpiness in elemental abundance space. Although not shown, a flatter CMF also entails a larger number of massive clusters and hence a clumpier elemental abundance space, echoing the results of Bland-Hawthorn & Freeman [21] and Bland-Hawthorn et al. [23]. The effect of the CMF on the distribution of deviations could potentially be exploited to reconstruct the CMF (and the physical processes that the CMF depends on, such as the SFH) from observational samples. This will be the subject of future work.

What are groups in elemental abundance space comprised of?

In this section we investigate the properties of the "detected" groups in elemental abundance space (consisting of $> 5\sigma$ fluctuations). Fig. 2.7 shows the distribution of the local "S/N" for those cells exceeding 5σ from Poisson statistics. Recall that the local S/N is defined as the ratio of stars coming from the most massive cluster in the cell to the remaining stars in that cell. A cell dominated by a single massive cluster will have high local S/N. In the left panel, we assume $N_{\rm cells} = 10^4$ and consider three different CMFs. Clearly most of the detectable groups have local S/N < 1, especially for the quiescent and fiducial CMFs.

This result is not surprising in light of the mean number of stars per cell (100 for $N_\star = 10^6$ and $N_{\rm cells} = 10^4$). In this regime, in order for the S/N to be $\gg 1$, we would require that a single domi-

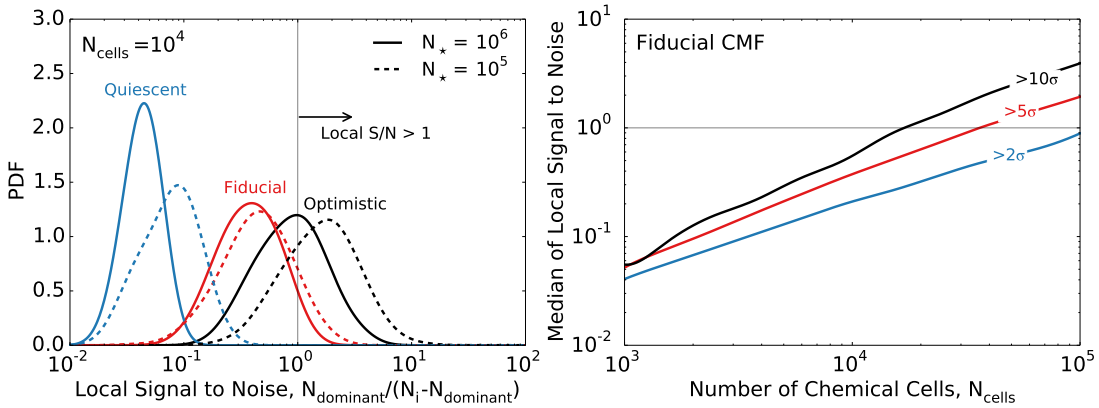

Figure 2.7: Left panel: Local $\mathrm{S/N}$ ratio in chemical cells with 5σ more stars than the average. The number of stars sampled from the dominant cluster is considered signal in each cell, whereas the rest are considered noise. The y-axis shows the probability of a detectable group having a certain local $\mathrm{S/N}$. The integral under each curve is one. We assume $N_{\mathrm{cells}} = 10^4$. In this case, most detectable groups have local $\mathrm{S/N} < 1$, showing that at least half of the stars in the detectable groups are not from dominant clusters. The difference between $N_\star = 10^5$ and 10^6 is small, illustrating that sampling more stars increases the number of stars per cell, but it does not change the $\mathrm{S/N}$. Right panel: Median of local $\mathrm{S/N}$ for different N_{cells}. We assume a fiducial CMF in this panel. Unlike $N_\star$, increasing N_{cells} boosts the local $\mathrm{S/N}$, and hence increases the chance of recovering individual clusters through chemical tagging.

nant cluster contribute $\gg$ 100 stars in a particular cell. However, as shown in Fig. 2.4, the average number of stars sampled per cluster for the most massive clusters is $\sim$ 100 for $N_\star = 10^6$. The relatively low sampling rate, combined with the high average number of stars per cell, essentially guarantees that the local S/N will never be much larger than one. As we discuss in §2.5.1, the prospects for finding higher local S/N cells can be improved by searching in regions of elemental abundance space in which the mean number of stars per cell is low.

The result in the left panel of Fig. 2.7 is fairly insensitive to $N_\star$. Increasing $N_\star$ increases both the number of stars sampled per cluster and the "background" comprised of stars from small clusters and hence the local S/N is left largely unchanged. In fact, the local S/N slightly decreases as we increase $N_\star$. This is not unexpected. As $N_\star$ decreases, it becomes more difficult to exceed the Poisson threshold. Therefore for smaller $N_\star$, the clumping of detectable groups are mostly comprised of more massive clusters (e.g., $\sim 10^7$ $M_\odot$), which implies a better local S/N. By contrast, for a larger $N_\star$, the clumping could either be due to a massive cluster or a few moderately massive clusters (e.g., $\sim 10^4 - 10^6 M_\odot$). While the S/N is somewhat negatively impacted by increasing $N_\star$, the total number of detectable groups greatly increases with increasing $N_\star$, as shown in §2.4.2.

The right panel of Fig. 2.7 shows the median local S/N as a function of the number of chemical cells. Increasing $N_{\rm cells}$ results in a dramatic (almost linear) improvement in the local S/N. An increase in $N_{\rm cells}$ results in a decrease in the local background while keeping the signal unchanged. This panel also shows the effect of changing the definition of a "detected" group from 2σ to 10σ. Increasing the threshold has a modest effect on the local S/N but of course has a dramatic effect on the total number of resulting detected clusters. Although not shown, we have explored the effect of

varying the slope of the CMF from $\alpha = -2.0$ to -1.5. This has only a modest effect on the trends shown in Fig. 2.7.

Note that the (5σ) deviation with respect to Poisson statistics is measurable in reality as it only requires the expected average number of stars in each cell. On the other hand, the local S/N is not measurable.[§] In this paper we only define "detectable groups" according to a measurable parameter, and we emphasize again that we use the term "group" rather than "cluster" when describing clumps in elemental abundance space because of the effect discussed in this section. The ambiguity that can arise, even when a cell deviates by more than 5σ argues strongly that interpretation of the data from ongoing and upcoming surveys will require models such as the one presented in this work.

Number of detectable groups as a function of model parameters

In this section we present the total number of detectable groups in elemental abundance space as a function of a variety of model parameters, including the in-situ fraction $f_{\rm in-situ}$, CMF slope α, survey width $\Delta R_{\rm survey}$, number of chemical cells $N_{\rm cells}$, and number of stars in the survey $N_\star$. We vary one of these model parameters at a time while adopting the fiducial values for the other model parameters (see Table 2.2); modifying more than one parameters at once is allowed in the online applet. The results are presented in Fig. 2.8 and Fig. 2.9.

[§]For readers who want to understand the number of groups that consist mainly a dominant cluster (e.g., having local S/N greater than 1), we urge readers to explore the interactive online applet (see Appendix 2.8 for details). In the applet, we allow users to impose a local S/N criteria.

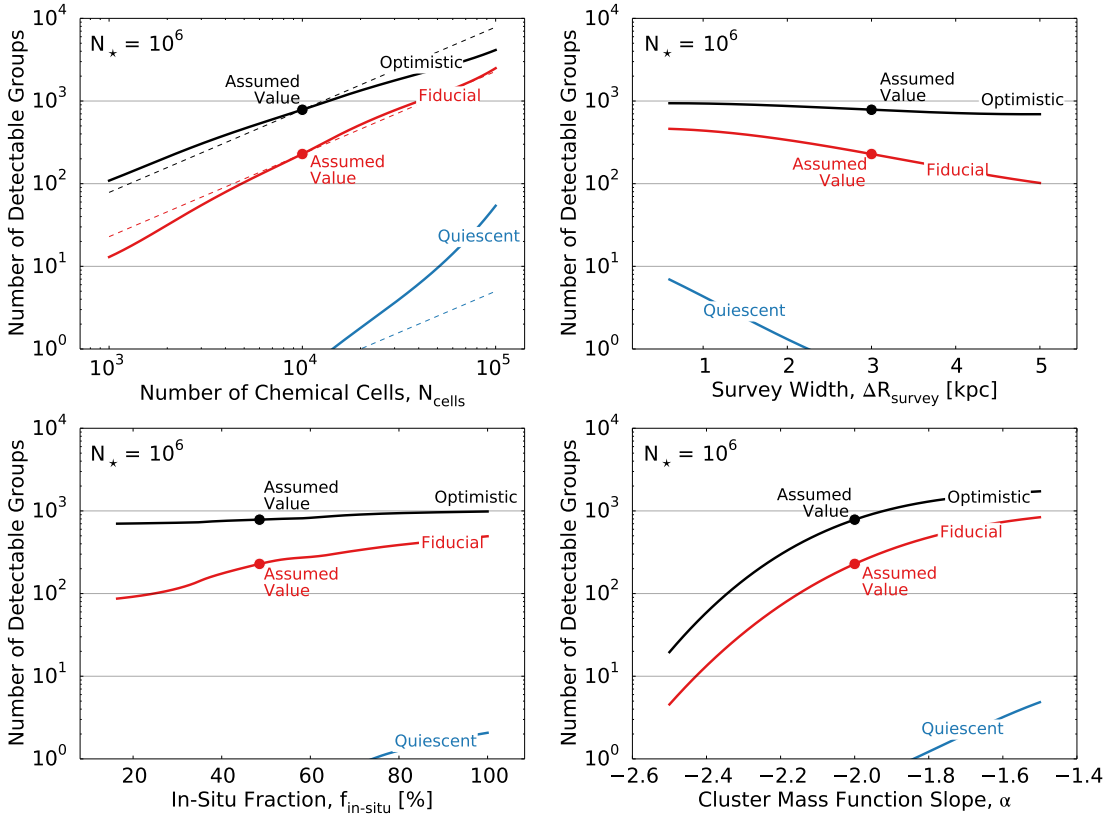

Figure 2.8: Total number of cells that exceed 5σ from Poisson statistics as a function of a variety of model parameters. We vary each of these model parameters while fixing the rest to the fiducial values as listed in Table 2.2. The three different solid lines show results from three CMF evolutions as illustrated in Fig. 2.2. The dashed lines show linear relations for reference. The solid symbols show the results assuming fiducial values for all model parameters. See text for discussion.

Number of chemical cells As the number of chemical cells increases, more moderately massive (e.g., $\sim 10^4 - 10^6 M_\odot$) clusters start to occupy different cells instead of sharing the same cell. The total number of detectable groups thus increases, approximately linearly for the fiducial and optimistic CMFs. However, the gain is more drastic for CMFs with a smaller high mass cutoff. This trend is due to the fact that, given the same $N_\star$, moderately massive clusters are more abundant for CMFs with a smaller high mass cutoff. These clusters might not be detected with a smaller $N_{\rm cells}$. Including more cells benefits these moderate clusters the most.

Since both the number of detectable groups and the local S/N (see §2.4.2) are sensitive to $N_{\rm cells}$, it is clear that $N_{\rm cells}$ is one of the most important parameters in the context of chemical tagging. Recall that the number of cells scales as $\sigma^{-N_{\rm dim}}$, where $N_{\rm dim} \sim 8$ is the number of independent dimensions in the elemental abundance space we can expect for upcoming optical surveys (GALAH and Gaia-ESO). Therefore, if we improve the abundance measurement uncertainties by a factor two, the number of chemical cells is improved by a factor of $2^8 \sim 250$. On the other hand, this also means that the number of chemical cells decreases by a factor ~ 2 for every 10% increase in the measurement uncertainties. Substantial effort should therefore go into decreasing (and characterizing!) the uncertainties in abundance measurements in upcoming spectroscopic surveys.

Survey width As $\Delta R_{\rm survey}$ increases the number of detectable groups decreases. To understand this trend, it suffices to note that as we increase $\Delta R_{\rm survey}$ there are more stars in the annulus. As a result, the chance that we sample from the same cluster decreases (i.e., the sampling rate decreases). Since each cluster is sampled with fewer stars, the chance to observe signal spikes in elemental abun-

dance space also decreases. Therefore, the total number of detectable groups decreases as the survey width widens. In fact, since the volume of the Solar annulus is proportional to ΔR_{survey}, the number of stars in the annulus is also roughly proportional to ΔR_{survey}. Therefore, the sampling rate is, to first order, inversely proportional to ΔR_{survey}.

Interestingly, the survey width has less effect on CMFs with a larger higher mass cutoff. This trend is due to the fact that as we increase the survey width, we also increase the number of clusters, roughly in proportion to ΔR_{survey}. The most massive clusters are the least susceptible to change in sampling rate because a large number of stars from such clusters are already sampled in the fiducial case. For CMFs with a larger high mass cutoff, the decrease in sampling rate caused by an increase in ΔR_{survey} is partly compensated by the increase in the number of massive clusters, resulting in a weak dependence of the number of detectable groups on ΔR_{survey}.

In-situ fraction As the in-situ fraction decreases, the number of cells exceeding 5σ decreases because there are more contaminants from ex-situ clusters (see Fig. 2.5). However, the effect of in-situ fraction is rather marginal for CMFs with a larger high mass cutoff. This effect is best understood from Fig. 2.6. Most of the detectable groups for a quiescent CMF or a fiducial CMF are at the edge of the detection level of 5σ. Hence adding in additional background noise in the form of ex-situ stars can have a much larger effect for model with a quiescent CMF compared to an optimistic CMF, in which many of the cells far exceed the 5σ detection threshold.

CMF SLOPE As we vary the CMF slope, we are essentially redistributing mass between smaller clusters and massive clusters. This has two effects that act in tandem: a shallower CMF results in more massive clusters, which will have more stars sampled per cluster. In addition, a shallower CMF results in fewer low mass clusters that contribute primarily to the "noise" in a cell. The elemental abundance space becomes much clumpier as α increases (also see [23]), and as a result there are many more detectable groups.

NUMBER OF STARS IN THE SURVEY Since the number of stars sampled for massive clusters is roughly proportional to $N_\star$ while the Poisson threshold only grows as $\sqrt{N_{\rm mean}} \propto \sqrt{N_\star}$, increasing $N_\star$ improves the number of detectable groups, as shown in Fig. 2.9. In the left panel, the gain is approximately linear in $N_\star$ for the optimistic and fiducial CMFs. The right panel shows the gain in the number of detectable groups as a function of $N_\star$ and $N_{\rm cells}$. The stochasticity at $N_\star \sim 10^4$ is likely due to the uncertainties in our Monte Carlo procedures.

Selecting subpopulations

As we argued in §2.2, the sampling rate, which is proportional to the number of stars in the survey divided by the number of stars in the survey volume, is a key parameter determining the number of stars sampled per cluster. In the limit where the sampling rate is 100%, the main limiting factor for chemical tagging is the resolution in elemental abundance space. One way to increase the sampling rate is to increase $N_\star$; this was discussed in the previous section. A second way is to decrease the number of stars in the survey volume. The latter will be effective only if one is able to identify a

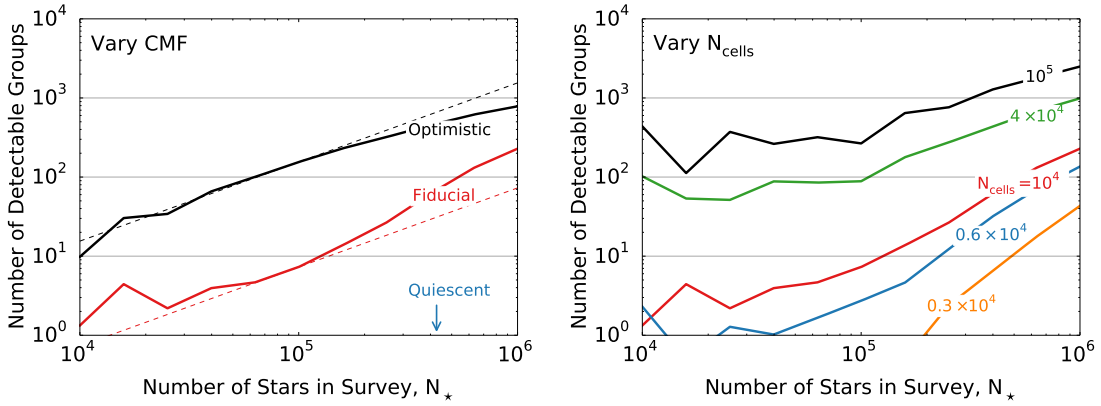

Figure 2.9: Total number of cells exceeding 5σ from Poisson statistics as a function of the number of stars in the survey. We assume a survey width of $\Delta R_{\rm survey} = \pm 3\,{\rm kpc}$ and $f_{\rm in-situ} = 50\%$. The red solid lines in both panels represent the reference results assuming a fiducial CMF and $N_{\rm cells} = 10^4$. The dashed lines show linear relations for reference. Different solid lines in the left panel show the results assuming different CMFs, whereas the right panel shows the results for different $N_{\rm cells}$. See text for discussion.

subpopulation of stars that corresponds to a subpopulation of clusters. For example, selecting on stellar age satisfies this criterion, while selecting a random subsample does not.

Fig. 2.10 considers the case where only stars above certain stellar ages are targeted in a survey. Since the number of older stars is smaller, there are not as many survey candidates compared to the case where we sample all disk stars uniformly. As a consequence, given the same $N_\star$, the chance that we sample from the same cluster improves. In addition to improving the total number of detectable groups, as we consider a more selective stellar subpopulation the number of clusters is reduced. The dominant cluster therefore contributes a greater fraction of the total stars in each detectable group because there are not as many clusters sharing the same cell. As shown in the right panel of Fig. 2.10, if the survey sample is collected randomly from all populations (the red solid line), most of the detectable groups have a local S/N of 0.3. This local S/N value implies that only $0.3/(0.3+1) \simeq 25\%$ of the members of detectable groups are from the dominant cluster. However, if we only target old stars with stellar age > 12 Gyr, the local S/N is ~ 2, indicating that $2/(2+1) \simeq 70\%$ members of each of the detectable groups are from the dominant cluster.

As a caveat, we caution that the interpretation of Fig. 2.10 is complicated by the fact that the selection of older clusters also preferentially selects a population of stars forming from a CMF with a higher mass cutoff (at least for the fiducial model used in the figure). So not only is the sampling rate increasing but so also is the characteristic cluster mass. Future work is required to disentangle these effects.

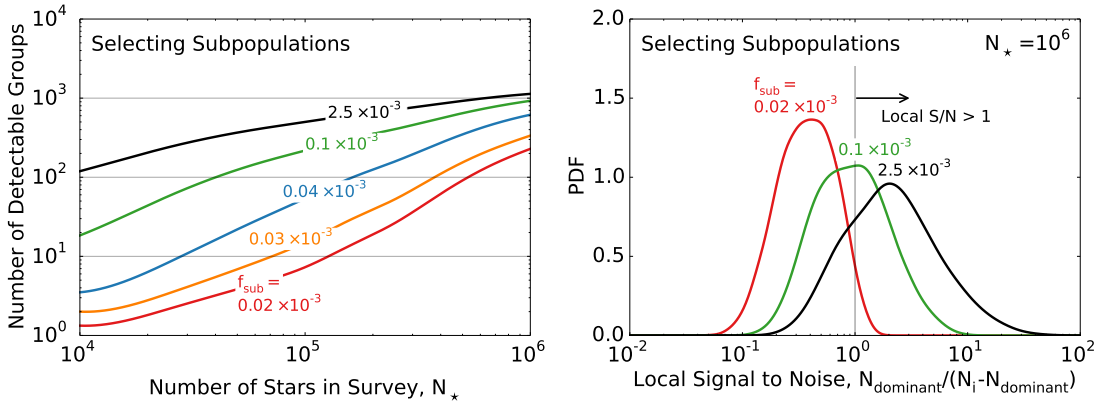

Figure 2.10: Left panel: Total number of cells exceeding 5σ from Poisson statistics as a function of the number of stars in the survey. We assume a fiducial CMF, with $N_{\rm cells} = 10^4$, $\Delta R_{\rm survey} = \pm 3\,{\rm kpc}$ and $f_{\rm in-situ} = 50\%$. Different lines in this panel show the results assuming a variety of subpopulation selections. The subpopulations are selected through the stellar age criteria of $> 0\,{\rm Gyr}$ (the lowest line), $> 3\,{\rm Gyr}$, $> 6\,{\rm Gyr}$, $> 9\,{\rm Gyr}$ and $> 12\,{\rm Gyr}$ (the highest line), respectively. The corresponding sampling rates, $f_{\rm sub}$, for $N_\star = 10^6$ are stated in each line. Right panel: Local ${\rm S/N}$ in each of the detected cells for different subpopulations, assuming $N_\star = 10^6$ and a fiducial CMF. The number of stars sampled from the dominant cluster is considered signal in each cell, whereas the rest are considered noise. See text for discussion.

2.5 Discussion

2.5.1 Summary of the key parameters affecting chemical tagging

The key parameters governing both the ability to detect groups in elemental abundance space and the "purity" of those recovered groups (i.e., the local S/N) are the number of stars in the survey $N_\star$, the number of chemical cells $N_{\rm cells}$, the CMF, and the sampling rate. Table 2.5 presents a summary of the key variables and their effect on various quantities of interest.

Several of these parameters are either outside of the control of the observer, including the form and evolution of the CMF, or are trivially in control of the observer, such as $N_\star$. Others require further consideration. For example, the number of chemical cells depends on both the volume of elemental abundance space and the size of each cell. The former depends on chemical evolution of the stellar population(s) under consideration, and can be influenced by the survey strategy. The latter is proportional to $\sigma^{-N_{\rm dim}}$ where σ is the observational uncertainty on abundance measurements and $N_{\rm dim}$ is the number of effective dimensions in the chemical volume.

Perhaps the most conceptually complex parameter is the sampling rate. For a fixed $N_\star$ the sampling rate is inversely proportional to the total number of stars available within the survey design. The phrase "survey design" was chosen to highlight not only the survey volume but also the subpopulation under consideration. Moreover, with regards to the survey volume, this must be considered in an orbit-averaged sense. For example, a survey targeting stars within 1 kpc of the Sun has a survey volume in this definition that encompasses the entire annulus of the Galactic disk with a width of

Table 2.5: The effects of various survey strategies on chemical tagging detections.

	Improve the number of detectable groups	Improve chance of recovering single cluster	Improve reconstruction of CMF
Increase $N_\star$	✓		✓
Decrease $\sigma_{[\mathrm{X/Fe}]}$	✓	✓	✓
Reduce $\Delta R_{\mathrm{survey}}$	✓		✓
Subpopulations	✓	✓	✓

± 1 kpc. Likewise, a pencil beam survey of bulge stars has a survey volume of the entire bulge. As we showed in §2.4.2, selecting subpopulations of stars can be very effective provided that the selection picks out a subset of clusters. Selecting on stellar age can achieve this, and so will effectively boost the average number of stars sampled per cluster. On top of that, selecting subsample reduces the number of clusters in each cell, and thus improves the local S/N in each detectable group. In contrast, a random subsample of stars will simply result in a smaller number of stars per cluster.

These parameters affect different aspects of chemical tagging. As shown in Table 2.5, increasing the number of stars or reducing the survey volume increases the number of detectable groups and improves the reconstruction of the CMF because it increases the sampling rate, but it has little effect on the local S/N ratio. Even though the sampling rate increases in these cases, both the local signal and noise increase in similar proportions. In contrast, decreasing $\sigma_{[X/Fe]}$ and/or selecting subpopulation reduces the average number of stars per cell, while maintaining the same signal. Therefore the local S/N improves as well.

In this work we focused on idealized surveys of stars in the Milky Way disk. In such situations the ratio of the number of stars in the annulus, $N_{\rm annulus}$ to $N_{\rm cells}$ is $\gg 1$. However, there are regimes in which this ratio can be closer to or less than unity. Bland-Hawthorn et al. [23] considered the regime of metal poor stars in dwarf galaxies. Such subpopulations could easily have a total number less than $N_{\rm cells}$. In this case the mean number of stars per cell will be $\ll 1$ and so significant overdensities in elemental abundance space will much more likely reflect a single cluster, rather than a superposition of multiple clusters (see example in [101]). As argued by Bland-Hawthorn et al. [23], in this regime one can in principle find clusters in elemental abundance space with a relatively modest number of stars

surveyed, provided that the CMF is not too steep. Similarly, for a survey targeting disk stars, one might imagine the first chemical-tagging detections coming from the less populated regime in elemental abundance space with a smaller contaminated background $N_{\rm mean}$ (i.e., outliers), as discussed in Bland-Hawthorn et al. [25].

2.5.2 Strategies for optimizing the potential for chemical tagging

The influence of key parameters on various observables allows us to consider ways in which one could optimize a spectroscopic survey of stars for the purposes of chemical tagging.

A survey that could reach $N_\star \sim 10^6$ and $N_{\rm cells} \gtrsim 4 \times 10^4$ could potentially achieve three major goals: (a) producing a sizable number ($\sim 10^3$) of detectable groups; (b) the detectable groups would consist primarily of a single dominant cluster; and (c) reconstructing the CMF. These goals could be realized if the CMF is somewhere in the range between our "fiducial" and "optimistic" scenarios. The GALAH survey[53] aims to observe $N_\star = 10^6$; a key question will be whether or not the number of chemical cells is closer to 10^4 or 10^5 (see Section 2.5.1 for the key dependencies).

Even if not all three goals are realized in the context of a massive spectroscopic survey, one could imagine a tiered approach. A survey of 10^6 could be used to identify overdensities in the elemental abundance space. One could then follow up those overdensities with higher quality spectroscopy to obtain more precise abundance constraints, or one could appeal to differential techniques to increase the relative abundance precision. One could also use other information to separate multiple clusters within a single cell, e.g., kinematics or color-magnitude diagrams.

Given that both $N_\star$ and $N_{\rm cells}$ affect the number of detectable groups in elemental abundance

space in similar ways, is there an advantage to spending more time collecting greater numbers of stars, or more time obtaining higher quality spectra could lead to smaller $\sigma_{[X/Fe]}$, more elements, and hence larger N_{cells}? In the simplest scenario (assuming for example that one has not already exhausted the input catalog at a particular apparent magnitude), $N_\star$ is roughly proportional to the integration time. On the other hand, since $N_{cells} \propto \sigma^{-N_{dim}}$, there is an enormous gain in N_{cells} for even a modest improvement in the abundance uncertainties. For $N_{dim} \sim 8$ independent dimensions (likely appropriate for e.g., GALAH), one could improve N_{cells} by a factor of two for a 10% reduction in the abundance uncertainties (§2.4.2). Therefore, if the goal is to find as many local peaks in elemental abundance space (i.e., detectable groups) as possible and/or to increase the odds of those peaks being dominated by a single massive cluster, it might be more advantageous to seek strategies that reduce the abundance uncertainties rather than simply acquiring more stars.

An effective way to improve chemical tagging detections is by targeting a stellar subpopulation exclusively. As we have shown in §2.4.2 and discussed in §2.5.1, targeting a subpopulation not only improves the sampling rate but also reduces the number of clusters per chemical cell. It improves chances of the reconstructing the CMF because there are more stars sampled per cluster and more significant deviations from Poisson statistics. It also improves the local S/N and hence the chance of recovering individual clusters within detectable groups in elemental abundance space.

A variety of properties could be used to select special subpopulations from a larger parent sample, including age, metallicity, and kinematics. One could envision pilot surveys at modest spectral resolution designed to select stars in a narrow range in [Fe/H]. Kinematics from Gaia could be used to separate hot and cold components, for example thin and thick disk stars (e.g.,[168]). Stars could also

be selected according to their age once age measurements are available for large samples of stars, e.g., from isochrone fitting and/or asteroseismic constraints. Finally, in an optically selected survey such as GALAH, which is biased to higher Galactic latitudes, it preferentially observes thick disk stars[53]. Since the total number of thick disk stars is smaller than thin disk stars, this preference argues that the sampling rate in these surveys could be larger than the one we assume in this study as we adopt an uniform sampling strategy (see also[25]).

2.5.3 Caveats, limitations and future directions

A variety of assumptions and simplifications were made in this study. Here we highlight the most important limitations and comment on future directions.

When populating the elemental abundance space we assumed that clusters are (statistically) homogeneously distributed in all $N_{\rm cells}$ chemical cells available. From both observations and chemical evolution models we know that this assumption is not true in detail. Of course, there are many more high metallicity stars than low metallicity stars, but also we expect the size of the elemental abundance space to vary systematically with metallicity (for example, due to certain nucleosynthetic pathways, e.g., in AGB stars, that only become important some time after the initial burst of star formation). Because of these complexities, the space cannot be completely described by the parameter $N_{\rm cells}$. A more accurate approach would be to include a model for chemical evolution and then to define overdensities in elemental abundance space with respect to a local background, either using neighboring cells or a more sophisticated group finding algorithm (e.g.,[150,185]).

This study focused on idealized surveys targeting Milky Way disk stars. We did not consider the

bulge, stellar halo, disrupted satellite galaxies, nor nearby dwarf galaxies. Each of these populations offers a unique set of challenges and opportunities. These components will be included in future versions of the model.

We did not follow the actual orbits of stars in a live Galactic potential, and the treatment of radial migration is quite simplistic. One could imagine an extension to the current model that follows the dynamical disruption of star clusters and the sequent orbital histories of the individual stars. This would be very valuable for exploring the potential gains of folding in kinematic information, such as will soon be available from Gaia and/or from the spectroscopic surveys themselves. Mitschang et al. [151] found that kinematics information does not improve the detectability, but it is likely due to the limitation of their small sample with $< 10^3$ stars. As we have demonstrated in this study, detectable groups in small sample are not likely to be co-natal, agreeing with their assessment.

The adopted model for the gas mass is fairly simplistic. However, we emphasize that the gas mass distribution only influences the radial migration prescription and the evolution of the CMF. The former is parameterized via the in-situ fraction, $f_{\rm in-situ}$. In both cases we consider a range of possible scenarios, which in some sense is equivalent to exploring the effects of varying the underlying gas mass model directly.

We assume that the spatial frequency of star formation follows an exponential disk characterized by the scale length $R_{\rm SFR}$. We are aware that this assumption might not be true in detail. At a given time, stars might form in some large scale molecular rings (e.g., [26,76]) or spiral arms (e.g., [17,171]). However, we are only interested in the integrated star formation rate over the cosmic history. Since these transient complexes, at least for the molecular rings, are expected to be short lived and rapidly dissi-

pate ($<$ 100 Myr; e.g.,[11,76]), the smooth star forming assumption is likely to do fine.

2.6 Summary and conclusion

In this study we explored the prospects for chemically tagging stars in idealized spectroscopic surveys of the Solar vicinity. We constructed a simple two dimensional time-dependent model of the Milky Way disk including the effects of radial migration and evolution in the CMF. We explored a number of important parameters affecting the detectability of groups of stars in elemental abundance space and we studied the composition of the detectable groups. We now summarize our principle conclusions.

- The key parameters affecting the number of detectable groups in elemental abundance space, and whether or not those groups are dominated by a single massive cluster, are: the shape and evolution of the CMF; the number of chemical cells; and the survey sampling rate. The sampling rate is proportional to the number of stars in the survey divided by the total number of stars belonging to a particular (sub)population. The latter two parameters are strongly influenced by observational survey design choices.
- The clumpiness in elemental abundance space is strongly influenced by the CMF and by the survey sampling rate. This implies that one can probe the CMF of long disrupted clusters by statistically analyzing the clumpiness in elemental abundance space.
- Confidently identifying individual clusters through chemical tagging will be challenging even for $N_\star = 10^6$, if disk stars are uniformly sampled. Fundamentally this is because the sampling rate is inherently small in such cases ($\sim 10^{-4}$) implying that one expects to collect on average 10 stars per cluster even for clusters with $\mathcal{M}_{\rm cluster} \simeq 10^6\,{\rm M}_\odot$. This is born out by our modeling, where we find that even very large overdensities in elemental abundance space are typically not comprised of stars from a single dominant cluster. In the fiducial case with $N_{\rm cells} = 10^4$, the dominant cluster contributes only 25% of the stars in the detectable group. Additional follow-up of the stars within large overdensities in elemental abundance space may provide additional discriminating power, either by decreasing the measurement uncertainties on the abundances, or by folding in color magnitude diagram or kinematic information.

2.7 Appendix: Sampling algorithm and computational cost

The sampling algorithm used to create a mock sample is illustrated in Fig. 2.11. To summarize, given an SFH, we obtain the stellar mass evolution through the stellar population synthesis code and the gas mass evolution through the inverted Kennicutt-Schmidt relation. The radial size growth is calculated using an observationally estimated mass-radius relation, which we use to predict the evolution of the SFR scale length. After we obtain the SFR scale length, we calculate the SFR at different radii and different cosmic times from the SFH. We spawn stars through cosmic time according to the radial SFR in discrete time bins of 0.1 Gyr. We only trace stars with $0.5 - 1.5\,M_{\odot}$, and we assume a Kroupa IMF.

The gas and stellar masses yield the total mass distribution at different radii and cosmic time. The mass distribution controls the radial migration prescription. The mass distribution is also employed to evaluate the high mass end of the CMF. The CMF is then used to assign a cluster tag to each spawned star, and the radial migration prescription is adopted to mix stars from their birth radii. Note that, we only assign cluster tags after spawning stars in each time bin. We do not generate stars recursively from the CMF although they are both equivalent. In the former case, we avoid a recursive loop in the algorithm and therefore create the mock sample more efficiently. Finally, a mock sample that is within the Solar annulus, given a fixed survey width, is saved for analysis.

Even though the sampling algorithm is straightforward, the effect of radial migration requires us to spawn stars at all radii in the Milky Way disk. In addition, we need to follow each individual star. Therefore, for each set of parameters, we spawn $\sim 10^{11}$ stars, which is computationally

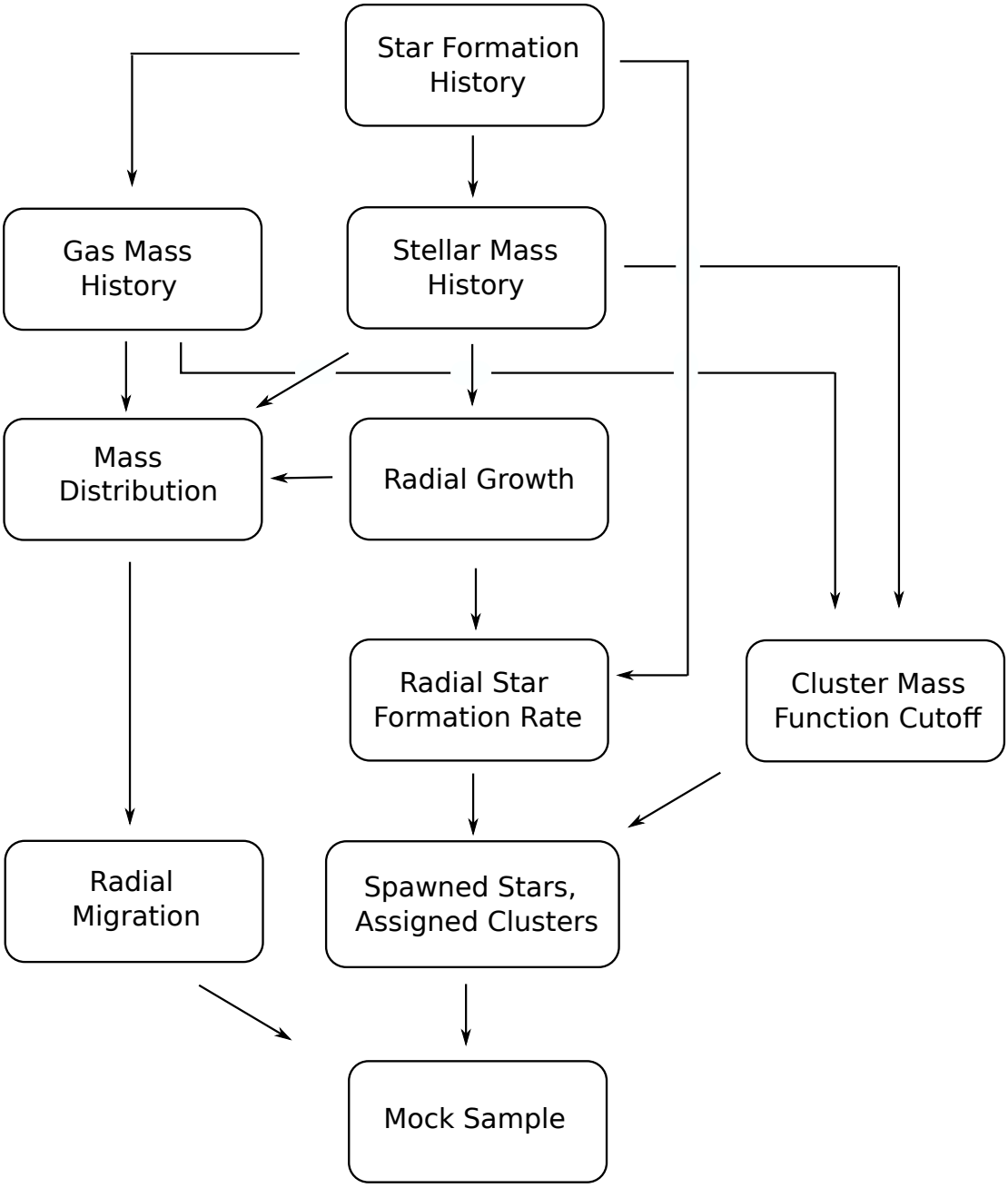

Figure 2.11: Sampling algorithm to create a mock Milky Way dataset in this study.

expensive even for a semi-analytic model. Each parameter set takes a full CPU day and 50 GB of memory per CPU to evaluate. We evaluate a grid of $\sim$ 600 different model parameters. It therefore took $\sim$ 2 CPU years to generate the mock samples. After the mock samples were created, we performed Monte Carlo simulations, distributing them into chemical cells. The Monte Carlo simulations required about the same amount of CPU time. Hence, it took $\sim$ 4 CPU years in total to generate the results in this study. Including a significant amount of exploratory work, this project consumed $\sim$ 40 CPU years of compute time. Obviously, parallelization reduced the total time from $\mathcal{O}$(graduate student lifetime) to $\mathcal{O}$(graduate student year).

2.8 Appendix: Interactive applet

Since we study a large multidimensional grid of simulations, it is challenging to include all results in this paper. We created an online applet to demonstrate results in the multidimensional grid. In the online applet (www.cfa.harvard.edu/∼yuan-sen.ting/chemical_tagging.html) as shown in Fig. 2.12, we plot the cumulative number of detectable groups (exceeding 5σ) as a function of the zero age mass of the dominant cluster. In each detectable group, star cluster with the most stars sampled is considered as the dominant cluster.

The applet allows users to change: the in-situ fraction $f_{\mathrm{in-situ}}$, (i.e., the radial migration prescription); the number of chemical cells N_{cells}; the CMF cutoff $M^{\mathrm{max}}_{\mathrm{cluster}}$, and slope α; the survey depth $\Delta R_{\mathrm{survey}}$; and the number of stars in the survey $N_{\star}$. As demonstrated in Section 2.4.2, these detectable groups do not necessarily comprise of co-natal stars. The online applet also allows users to impose a local S/N selection criteria as defined in Section 2.4.2. For instance, by imposing the criteria local S/N > 1, we select detectable groups that have more stars contributed by the dominant cluster over the combined background from smaller clusters. In the case where no local S/N criteria is imposed, the end point of the cumulative distribution in the applet corresponds to the results in Fig. 2.8 and Fig. 2.9. Finally, there is a "save as reference" button in the applet which allows users to save the current cumulative distribution as a reference and compare with the other choices of parameters

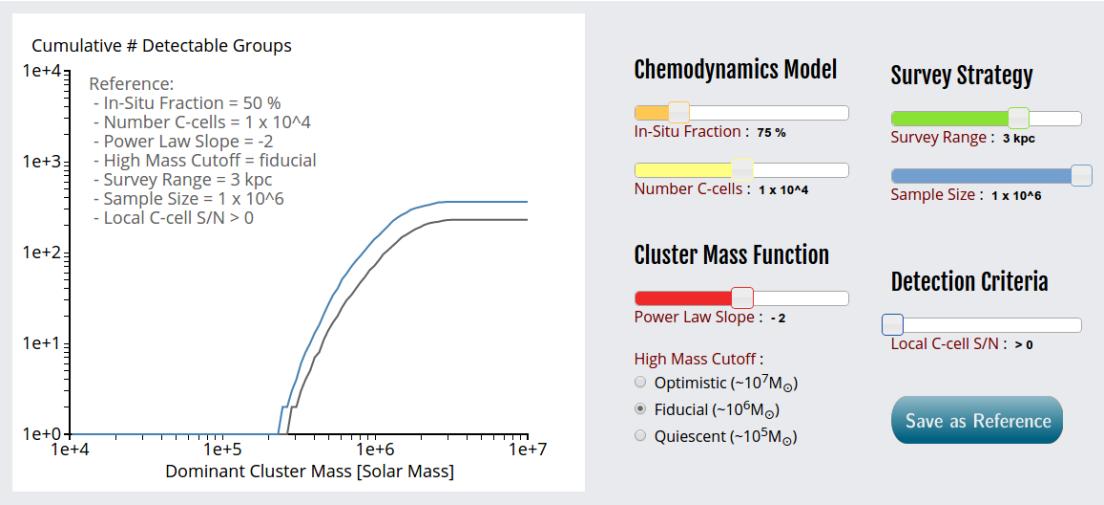

Figure 2.12: A demonstration of the online-applet created in the course of this project.

3

APOGEE chemical tagging constraint on the maximum star cluster mass in the α-enhanced Galactic disk

Author list of the original paper: Yuan-Sen Ting, Charlie Conroy, Hans-Walter Rix

3.0 Abstract

Stars born from the same molecular cloud should be nearly homogeneous in their elemental abundances. The concept of chemical tagging is to identify members of disrupted clusters by their clustering in elemental abundance space. Chemical tagging requires large samples of stars with precise abundances for many individual elements. With uncertainties of $\sigma_{[X/Fe]}$ and $\sigma_{[Fe/H]} \simeq 0.05$ for 10 elements measured for $> 10^4$ stars, the APOGEE DR12 spectra may be the first well-suited dataset to put this idea into practice. We find that even APOGEE data offer only ~ 500 independent volume elements in the 10-dimensional abundance space, when we focus on the α-enhanced Galactic disk. We develop and apply a new algorithm to search for chemically homogeneous sets of stars against a dominant background. By injecting star clusters into the APOGEE dataset we show that chemically homogeneous clusters with masses $\gtrsim 3 \times 10^7\ M_\odot$ would be easily detectable and yet no such signal is seen in the data. By generalizing this approach, we put a first abundance-based constraint on the cluster mass function for the old disk stars in the Milky Way.

3.1 Background

The Milky Way offers a unique opportunity to understand how disk galaxies form, in particular when and where, and in which types of aggregates, or clusters, they formed their stars. As star cluster masses depend on gravitational instabilities (e.g., [59]), the cluster mass function (CMF) indirectly probes the dynamical state of the Milky Way disk over cosmic time.

At least during the intensely star-forming, early phases of the Milky Way, the majority of stars are believed to form in clusters [1,115]. Most of these are rapidly disrupted and dispersed throughout the Galaxy (e.g., [48,112,156]), for a brief while appearing as moving groups (e.g., [37,49,52]). Once the phase space information as a common birth marker is lost, chemical tagging, first proposed by Freeman & Bland-Hawthorn [69], may still betray the common birth origin of stars through their exceptional similarity in elemental abundances. Stars originating from the same star cluster are believed to be homogeneous in their chemistry (e.g., [50,51,63,70,111,194]). Since the photospheric elemental abundances, at least for elements heavier than sodium, are invariant throughout their lifetime, they are permanent tags of the stellar birth origins.

A broad goal of chemical tagging is to reconstruct the stellar CMF, i.e., the relative distribution of (chemically homogeneous) stellar cluster masses when they formed. Although we can investigate the present-day CMF through young star clusters and massive surviving clusters (e.g., [16,27,34,35,124,160]), the Milky Way's CMF in the past is unknown. The key idea in chemical tagging is that massive chemically homogeneous clusters show up as discernible clumps in the multi-dimensional abundance space (e.g., [23,191]).

Besides understanding the CMF, chemical tagging is deemed an essential tool to understand the effect of radial migration in Galactic evolution (e.g., [77,178]). Sellwood & Binney [183] first proposed that stars could migrate significantly from their radial position when resonate with spiral/bar structures (see observational evidence from [87,113,133]). Although simulations concur to the analytic calculations [24,55,81,149,175,176], direct observational evidence of radial migration remains controversial. The ability to recover dispersed star clusters would be fundamental in quantitatively constraining radial migration models.

In recent years, the idea of chemical tagging has garnered more attention. Large spectroscopic surveys, including RAVE [188], APOGEE [212], GALAH [53] and Gaia-ESO [164], are being carried out. Results from these surveys have demonstrated the power of using elemental abundance patterns to identify distinct stellar populations in the Milky Way. Martig et al. [141] found young α-enhanced stars that are difficult to explain within current models of the evolution of the Milky Way. Masseron & Gilmore [143] and Hayden et al. [86] showed that the C/N ratio and the α-elements are good indicators to separate the thin and thick components of the Galactic disk. Schiavon et al. [177] found bulge stars that show abundance patterns similar to globular clusters. These studies focus on finding populations of stars via their abundance patterns, which is a "weak" form of chemical tagging. The goal of the "strong" form of chemical tagging is to identify stars that were born from the same molecular cloud. When we refer to chemical tagging in this paper, we only refer to this "strong" form. In this context, due to these exciting new opportunities, many preparatory works started to explore the capabilities of these surveys and the idea of chemical tagging. For example, Blanco-Cuaresma et al. [20], MacFarlane et al. [135], Mitschang et al. [150,151], Tabernero et al. [189,190] proposed various schemes

and performed numerical experiments on separating open clusters/moving groups in abundance space.

Ting et al.[191] explored the feasibility of chemical tagging by exploring a grid of Galactic evolution parameters. They found that identifying individual clusters through chemical tagging is generally challenging with the on-going surveys. Clumps that show overdensities in abundance space are usually made up of many smaller clusters, i.e., the background contaminants are non-negligible in clump search. One might be able to associate a detected clump as a single disrupted cluster only when the background density in abundance space is low. They showed that a low background can be achieved by either studying a subpopulation that occupies a large volume in abundance space, small abundance uncertainties or a large number of independent elements. Despite all these complications to reconstruct individual clusters, they argued that we can still statistically reconstruct the CMF through the clumpiness in abundance space. The main goal of this paper is to apply that idea to the APOGEE data[92].

In §3.2, we characterize the APOGEE sample that we explore in this study. In §3.3, we introduce our clump search method. The key is to define a robust search sphere that has the highest signal-to-background ratio possible. In §3.4, we apply this method to the APOGEE DR12 data. We discuss how this result can be applied to obtain a tentative constraint on CMF. We conclude in §3.5. By comparing the signal-to-background contrast observed in the data to the simulated contrast from injected clusters, we will argue that no chemically homogeneous clusters more massive than 3×10^7 $M_\odot$ have formed in the α-enhanced disk.

3.2 APOGEE sample properties

We adopt the APOGEE DR12 publicly available sample[92]. Similar to Hayden et al.[86], we consider stars with all elements measured and with reliable abundances, i.e., 4,000 K$<$ $T_{\rm eff}$ $<$ 5,500 K, 1 $<$ $\log g$ $<$ 3.8 and signal to noise ratio $>$ 80. Stars from APOGEE that satisfy these criteria are plotted in Fig. 3.1. In this study, we only focus on the α-enhanced disk as defined via the cut shown in Fig. 3.1. We focus on this subsample as their chemical/spatial modeling is likely to be more straightforward (see §3.4.2 and also see[25,191]). Furthermore, the Milky Way was likely kinematically hotter and more turbulent in the first few billion years (e.g.,[19,28,116]). High redshift (z $\sim$ 2) extragalactic studies have revealed the existence of massive star-forming clumps in star-forming galaxies (e.g.,[73,132]). As a result, star clusters within the α-enhanced disk could be more massive and if so would provide a strong signal in abundance space. The α-enhanced disk also occupies a larger volume in abundance space, i.e., lower background density of stars, which guarantees clumps a better contrast to the background in abundance space.

We checked that our main result presented in this study, namely there is no cluster more massive than 3 $\times$ 10^7 $M_\odot$ formed in the Milky Way, still holds at least to the 1σ level (see §3.4.3), if we choose a selection cut within the shaded region as shown in Fig. 3.1. The result only changes more dramatically if we choose a much steeper cut such as the blue dashed line. In this case, we discard too many low density regions where most chemical tagging signals reside (see §3.3.1 and §3.4.3), and we can only rule out clusters $\gtrsim$ 10^8 $M_\odot$. Using the fiducial selection cut, in total, the α-enhanced sample has 14,002 stars, as shown in the red and black symbols in Fig. 3.1. For reasons and selection

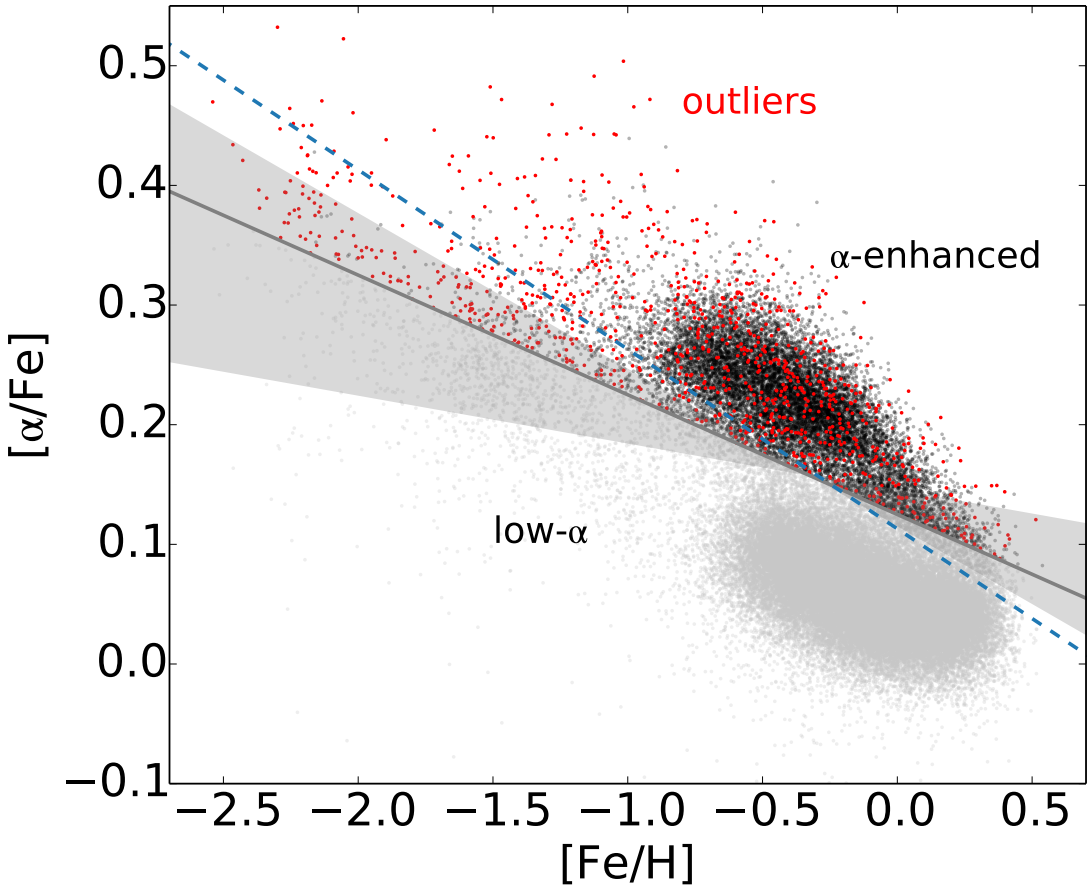

Figure 3.1: Black symbols show 13,000 APOGEE stars selected for this study. We only consider α-enhanced stars, as the volume they occupy in 10-dimensional abundance space is larger than the low-α sequence, so the background will be lower, and therefore detection is more likely. The red symbols show the $\sim$ 7% of outliers in the 10-dimensional abundance space that we do not include in the sample. After we discard these outliers, the 10-dimensional empirical distribution in abundance space is better modeled by an ellipsoid and is easier to deconvolve (see §3.3.2). The solid black line shows the fiducial selection cut. We also examine that the results in this paper are not sensitive to our data selection. If we choose a selection cut within the shaded region, the results in this paper still hold. The results only change more dramatically if we choose a much steeper cut such as the blue dashed line (see text for details).

criteria that will become clear in §3.3.2, we further discard the 7% of most outlying stars, shown as red symbols in Fig. 3.1, and end up with a final sample of 13,000 stars.

Without further kinematic modeling, it is hard to disentangle the halo stars from the disk stars. But the elimination of outliers as shown in Fig. 3.1 culls most of the metal-poor stars with [Fe/H] $<$ -1 and therefore, the bulk population in this study should not be contaminated much by the halo stars. We also performed the same analysis only considering stars with [Fe/H] $>$ -1 and found that the results remain qualitatively the same. For the potential bulge contamination, we find that among the 13, 000 stars, only 3% of them satisfy the bulge stars criteria with $l < 22$, $|b| < 15$ and $R_{\rm gc} < 3$ kpc, where $R_{\rm gc}$ is the isochrone Galactic radius derived in Hayden et al. [86]. Therefore, we will assume throughout this study that the sample only consists of disk stars.

To perform the chemical tagging experiment, we want to consider as many elements as possible in order to maximize the volume in abundance space. In this case the background becomes more diluted, and the signals will therefore have a better chance of standing out from the background. In total, APOGEE measures 15 elements. However, as discussed in Holtzman et al. [92], Na, Ti, V abundances might not be reliable yet in the current release. Furthermore, C and N are expected and seen to evolve through stellar evolution due to the post main sequence dredge up (e.g. [100,143,203]). Therefore, it is complicated to relate C and N to their primitive abundances when the stars formed. Discarding these 5 elements, we consider 10 elements in this study, namely Al, K, O, S, Mg, Si, Ca, Mn, Ni, Fe. Next, we want to define an abundance space out of these 10 elements. As discussed in Ting et al. [195], chemical tagging is better performed in [X/Fe] instead of [X/H]. [X/H] strongly correlate with each other. It is harder to observe the subtle variants among clusters in an abundance space

spanned by [X/H]. Therefore, in this study, we consider an abundance space spanned by [Fe/H] and 9 [X/Fe] from elements beside iron. We denote a vector in this 10-dimensional space to be X in this study.

As we will discuss in more detail in §3.3, to find a chemically homogeneous cluster in abundance space we first need to understand the typical volume that such a cluster occupies, after accounting for the measurement uncertainties that will dominate over the intrinsic scatter. Therefore, to estimate the typical volume, we will evaluate differential uncertainties, or measurement precision (not accuracy), $\sigma_{\mathbf{X}}$ and their correlations, i.e., the empirical point spread function of a chemically homogeneous cluster in abundance space. We estimate this co-variance matrix from known clusters in the APOGEE data and refer to the resulting matrix as the "cluster kernel" below.

We consider the DR10 cluster classification[146] since the DR12 classification is yet to be released. We cross-match the DR10 cluster member identities with DR12 and adopt the elemental abundances from the DR12 release. We only consider clusters with more than 10 cluster members. Three open clusters (NGC6819, NGC2158, M67) satisfy this criterion. Noting that all these clusters are metal-rich with [Fe/H] $\gtrsim -0.1$ and the possibility that abundance determination could be worse at lower metallicity, we also adopt one of the more metal-rich globular clusters, M107 ([Fe/H] $\simeq -1$). We fit M107 members with two Gaussians distributions in the 10-dimensional abundance space to eliminate any possible secondary population from this globular cluster. Including M107, we have a total of 77 cluster members. The primary population of M107 shows measurement uncertainties consistent with other open clusters. Restricting ourselves to the three metal-rich open clusters results in a slightly smaller $\sigma_{\mathbf{X}}$. Therefore, including M107 only makes our results more conservative (see §3.4.3).

We subtract the elemental abundances of each cluster by their means to center clusters at the zero origin. The co-variance matrix of these 77 stars is estimated. This matrix defines an ellipsoid that a typical chemically homogeneous cluster occupies. For each element, we find that $\sigma_{\mathbf{X}} \sim 0.05 - 0.06$ dex; this multivariate Gaussian sets the effective volume that homogeneous clusters occupy in the observed abundance space. This estimate is consistent with Holtzman et al.[92] (see table 6 in the paper, but note that they show measurement uncertainties in $[X/\mathrm{H}]$, but here we evaluate uncertainties in $[X/\mathrm{Fe}] + [\mathrm{Fe}/\mathrm{H}]$). Due to the small sample size of cluster members, we bootstrap the cluster sample and find that the uncertainty on this $\sigma_{\mathbf{X}}$ estimate is about 20%. A larger sample of cluster calibrators would be very helpful as a precise estimate of the cluster kernel is a key ingredient in any chemical tagging analysis.

3.3 Method

In this section, we will describe the challenges in abundance clump searches and our clump search method. Although various schemes have been proposed to separate open clusters/moving groups (e.g., [20,135,185]) in abundance space, a question often not discussed is the estimation and inclusion of background contaminants. Simulations from Ting et al. [191] showed that the background contaminants can be a critical limiting factor in chemical tagging experiments. After extensive experimentation we found that most proposed techniques, such as K-means, Gaussian mixture models, and minimal spanning tree, are only effective in separating clumps in the limit of a small background or when the background can be easily estimated.

Due to this limitation, we have developed a simple new method* geared toward regimes where the background is dominant and has a complex topology in abundance space (read also [62]). The key to our method consists of two parts that we will explain in §3.3.1 and §3.3.3. First, we need to estimate the local density. As we will discuss in more detail in §3.3.3, we define the local density to be the number of stars within a search sphere. The search sphere that we use to distinguish signals from the background should be sufficiently large. It should include a large fraction of a chemically homogeneous cluster but avoid being too wide and should not include too many background contaminants. Secondly, the abundance space distribution is not uniformly distributed. To estimate the detection significance, we have to estimate the expected background at each location.

*Our method can be regarded as a variation of density-based nonparametric clustering techniques.

3.3.1 Abundance space search sphere

In §3.2, we derived the empirical multivariate Gaussian distribution that a typical cluster occupies in abundance space. This distribution defines an ellipsoidal distribution in the 10-dimensional abundance space. Since an optimal search sphere should include a high ratio of cluster objects compared to the background contaminants, the optimal search volume, or search sphere, should follow the same uncertainty ellipsoid. However, it is not convenient to operate with a tilted ellipsoid because a simple Euclidean distance from the center alone is not sufficient to determine whether a star is included in the ellipsoid. Therefore, we linearly transform the abundance space such that the cluster kernel becomes a unit Gaussian distribution. We emphasize that the transformation is only to make calculations more straightforward, leaving all astrophysical implications invariant.

We can now determine an appropriate radius for the search sphere in these coordinates. A unit radius is not a good choice for the search sphere even though clusters follow a unit Gaussian distribution in the transformed coordinates: a box with 2 dex in width in each dimension only captures $(68\%)^{10} = 2\%$ of the clump. Since a unit n-sphere is strictly included in this box, one can show that a unit n-sphere encapsulates an even smaller fraction, 0.02%, of the data. This curse of dimensionality[†] implies that to capture, for example, 68% or 1σ of the cluster members, we require a search sphere with a radius larger than 1 dex in the transformed abundance space. In fact, mathematically, the inclusion fraction of a unit Gaussian within an n-sphere follows the χ^2-distribution. The analytic formula of a χ^2-distribution shows that an n-sphere of 3.4 dex in radius is needed.

[†]Techniques to compactify dimensions, such as PCA (e.g., [195]), do not mitigate this problem because the density of background contaminants also increases accordingly in the compactified space.

One important parameter that will determine the difficulty of chemical tagging detections is the number of separate "chemical cells" in abundance space[69,191]. The number of chemical cells is the ratio between the abundance space volume spanned by the data and the typical volume of a search sphere. We will evaluate this number later. But for now, one way to visualize the number of chemical cells is to compare the effective diameter of the search sphere to the spread in each elemental abundance. We illustrate this comparison in Fig. 3.2. We find that an n-sphere with a radius of 3.4 dex in the transformed abundance space corresponds to an ellipsoid with an effective radius of 0.05 dex in the original abundance space. The effective radius is defined such that an n-sphere with this radius contains the same volume as the ellipsoid. Due to the large dimensionality, we find that if we do not take into account the covariances of the cluster kernel, i.e., if we were to use an n-sphere in the original abundance space, instead of a tilted ellipsoid defined from the cluster members, we estimate that the search volume would be 10,000 times larger and the search sphere would include too many background contaminants.

Even with this optimized search ellipsoid in the original abundance space, as shown in Fig. 3.2, the distribution of each element is typically only 1-3 times the effective search sphere diameter. As a result, it is not possible to search for clusters in the core region of the chemical distribution. In this region, the search sphere would include too many background contaminants. The chemical tagging signals are most likely to come from the peripheral regions of the chemical distribution where the background contaminants are not dominant (also see[23,25,101]). Nonetheless, in a 10-dimensional space, the "surface-to-core" ratio is very large. Therefore, there is a reasonable chance of finding clumps in the peripheral regions. As we will show in §4, all chemical tagging signals indeed come

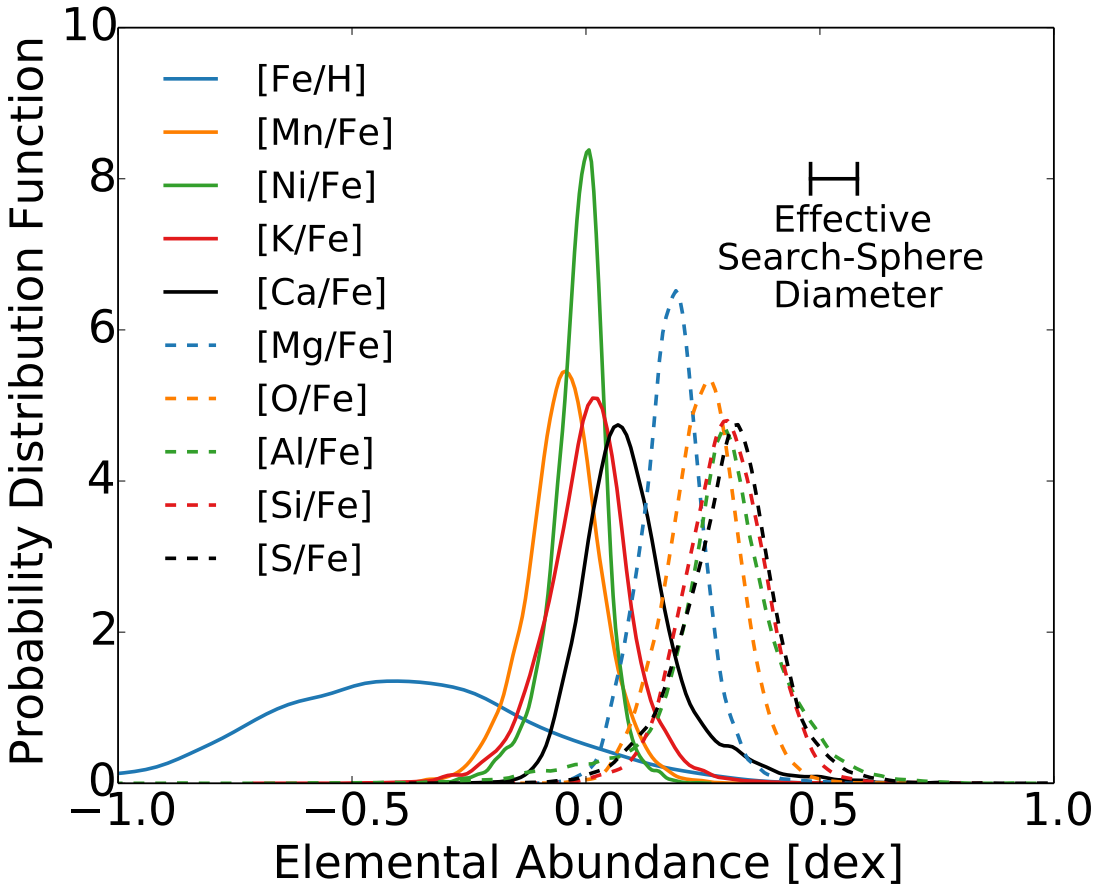

Figure 3.2: Distributions of all 10 elements in this study. By comparing these distributions to the effective search sphere diameter (see §3.3.1 for details), we find the abundance distributions in each dimension to have standard deviation widths that are typically $\sim 1-3$ times of the effective search sphere radius. The relatively small volume in abundance space compared to the search sphere volume highlights the main challenge in chemical tagging.

from the peripheral regions.

3.3.2 The intrinsic abundance distribution of the α-enhanced sample

We now proceed to deconvolve the observed abundance distribution. The intrinsic abundance distribution is required in order to inject mock clusters into the observed APOGEE dataset.

In the previous section, we transformed abundance space such that clusters follow a 10-dimensional unit Gaussian distribution. However, in such coordinates, the overall chemical distribution of the α-enhanced disk will still presumably show co-variances among different coordinate directions because the transformation is only to normalize the cluster point spread function and makes no assumption on the global distribution. Deconvolving such a co-variant distribution directly in 10-dimensions is computationally prohibitive. Therefore, we further rotate the transformed abundance coordinate system to eliminate the co-variances such that the joint 10-dimensional abundance distribution of the α-enhanced disk can be approximated by a product of 10 marginal distributions. Since the cluster kernel, reflecting the measurement uncertainties, is already isotropic in the transformed abundance space, it remains unaffected by further rotation. Upon this rotation, the deconvolution task simplifies to ten independent, one-dimensional deconvolutions on the marginal distributions.

However, this approach only works if the actual ensemble abundance distribution can be well approximated by its marginal distributions after rotation, i.e., if it does not show significant curvature in abundance space. If we consider all 14,002 α-enhanced stars in the sample, we find that this is not a good approximation: the 10-dimensional abundance space has a less regular topology, dominated by a small fraction of outliers in abundance space, as shown in red symbols in Fig. 3.1. To

look for these outliers, we perform a 10-dimensional kernel density estimation, using a unit Gaussian distribution as the smoothing kernel. We rank the data points according to their local density in the kernel density estimation and discard the most outlying 1,002 stars. We check that upon discarding these outliers (7%), injecting clusters according to the joint distribution gives similar statistics to injecting clusters according to product of marginal distributions. We emphasize that discarding outliers shrinks the peripheral regions and makes detection more unlikely. The main purpose of this study is to put an upper limit on the maximum cluster mass (see §3.4.3), discarding outliers only makes our estimate more conservative. We also check that these outliers are not particularly clumped in abundance space and hence are unlikely to be chemical tagging detections.

After breaking down the joint distribution to its marginal distributions, $P_{\mathrm{convolved},i}$, we model each marginal distribution with the sum of two Gaussians (see also [62,144]),

$$P_{\mathrm{convolved},i}(x_i|\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,f) \sim (1-f)\mathcal{N}(x_i|\mu_1,\sigma_1^2)+f\mathcal{N}(x_i|\mu_2,\sigma_2^2). \tag{3.1}$$

where μ and σ are the means and standard deviations of the Gaussian distributions, f shows the relative contribution from each distribution. We require a two-component Gaussian because the marginal distributions often show a core region and a broad wing region. Fitting a single Gaussian will underestimate the total area of low-density wings. As discussed in §3.3.1, the wing regions are the most valuable parts of a chemical distribution as they have the highest chance to detect clumps. In the model, we allow centers of the two Gaussian distributions, μ_1 and μ_2, to be different to ac-

count for an asymmetric distribution. We found that this model provides an excellent fit to each 1D marginalized distribution (although the joint distribution fits are slightly less satisfactory as we will explain below) and the deconvolution can be done analytically. We model the intrinsic chemical distribution $P_{\text{intrinsic}}$ to be

$$P_{\text{intrinsic}}(\mathbf{X}) = \prod_{\text{dim}=1}^{10} P_{\text{intrinsic},i}(x_i), \tag{3.2}$$

where

$$P_{\text{intrinsic},i}(x_i) = P_{\text{convolved},i}(x_i|\mu_1, \mu_2, \sigma_1^2 - 1^2, \sigma_2^2 - 1^2, f). \tag{3.3}$$

Fig. 3.3 shows the intrinsic chemical distribution derived according to the method above. We will use this model to draw mock data of hypothetical clusters, whose abundance probability is drawn from the ensemble distribution. We caution that the 10-dimensional ellipsoidal model does not give a perfect fit to the data, despite the fact that we have eliminated 7% of the outliers. For instance the [Ca/Fe] vs. [Fe/H] distribution, as illustrated in Fig. 3.3, shows a more complex morphology than an ellipsoidal model. In the ideal case, we would draw hypothetical clusters from a deconvolved distribution that displays similar intricate morphology. However, deconvolving such intricate distribution is computationally intractable in 10-dimensional. Nonetheless, we checked that injecting clusters according to the convolved ellipsoidal model shows similar statistics to the empirical convolved distribution. Therefore, in this study, we make the assumption that injecting clusters according to the deconvolved ellipsoidal model gives similar statistics if we were to inject according to the real deconvolved distribution.

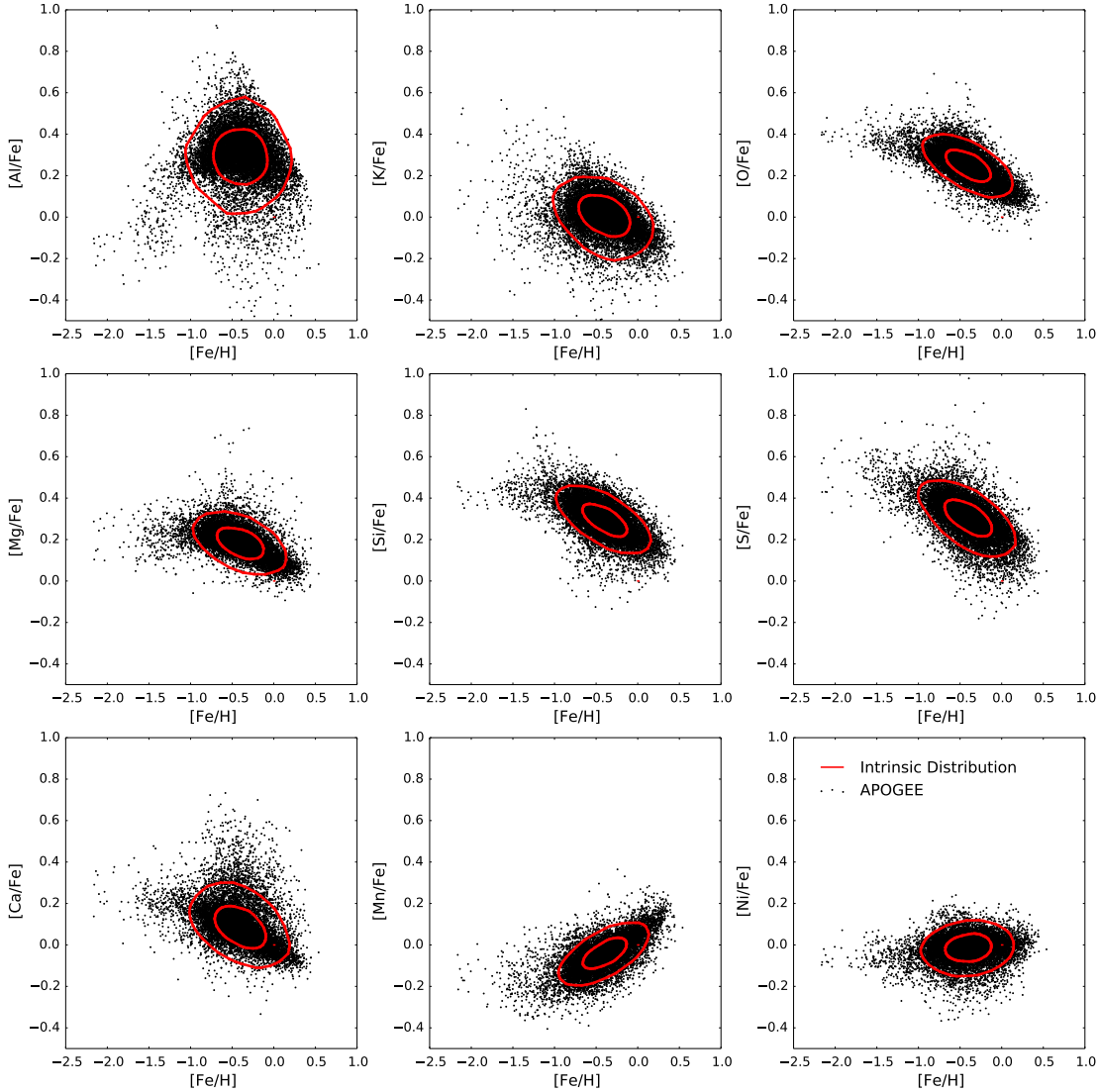

Figure 3.3: Deconvolved (intrinsic) model of the 10 elements in this study, projected onto $[X/\mathrm{Fe}]$-$[\mathrm{Fe}/\mathrm{H}]$ planes. The black symbols show the APOGEE sample adopted, illustrating the empirical distribution. The red contours show the 50 and 90 percentiles of the projected intrinsic distributions. When injecting mock clusters, we draw their center locations from the intrinsic distribution model instead of the empirical APOGEE distribution. Note that although an ellipsoid model provides an acceptable fit to the data, in some cases the abundance distribution is influenced by non-ellipsoidal structures.

3.3.3 Detection significance

So far, we have defined an operative search sphere with radius $r = 3.4$ dex to look for overdensities in the transformed abundance space. We know that the overall abundance distribution, i.e., the background, is not uniform (e.g., [10,13,56,168]). The absolute number of stars within the search sphere is therefore not particularly informative. We want to find regions where the local density within a search volume is significantly higher than its vicinity regions. Therefore, to define the detection significance, we need two ingredients: a local density estimation at each location and the corresponding local background estimation at this location.

Fig. 3.4 shows a schematic illustration of our clump search method. For each star s, we define the local density of this star, n_s, to be the total number of stars located within r distance from this star. Throughout this study, we only consider stars with $n_s > 10$ to avoid the large Poisson fluctuation at small n_s. We estimate the vicinity background, $\langle n_s \rangle$ to be the average of $n_{s'}$ where s' are all stars located within a distance of $r - 2r$ from star s. We define the detection significance to be $\sigma_{\rm detect}(s) = (n_s - \langle n_s \rangle)/\sqrt{\langle n_s \rangle}$, which measures the deviation of the local density from the vicinity background, in units of the Poisson uncertainty of the background.

Although this is a sensible definition of detection significance, there is a complication. If we have a uniform background, provided there is no signal, $\sigma_{\rm detect}$ should center around zero. Unfortunately, this is not the case for an uneven background, especially for high dimensions. At a fixed point in an uneven background, there are always more vicinity regions that have lower densities (toward the valley) than regions that have higher densities (toward the core). As a result, we have $n_s > \langle n_s \rangle$ in gen-

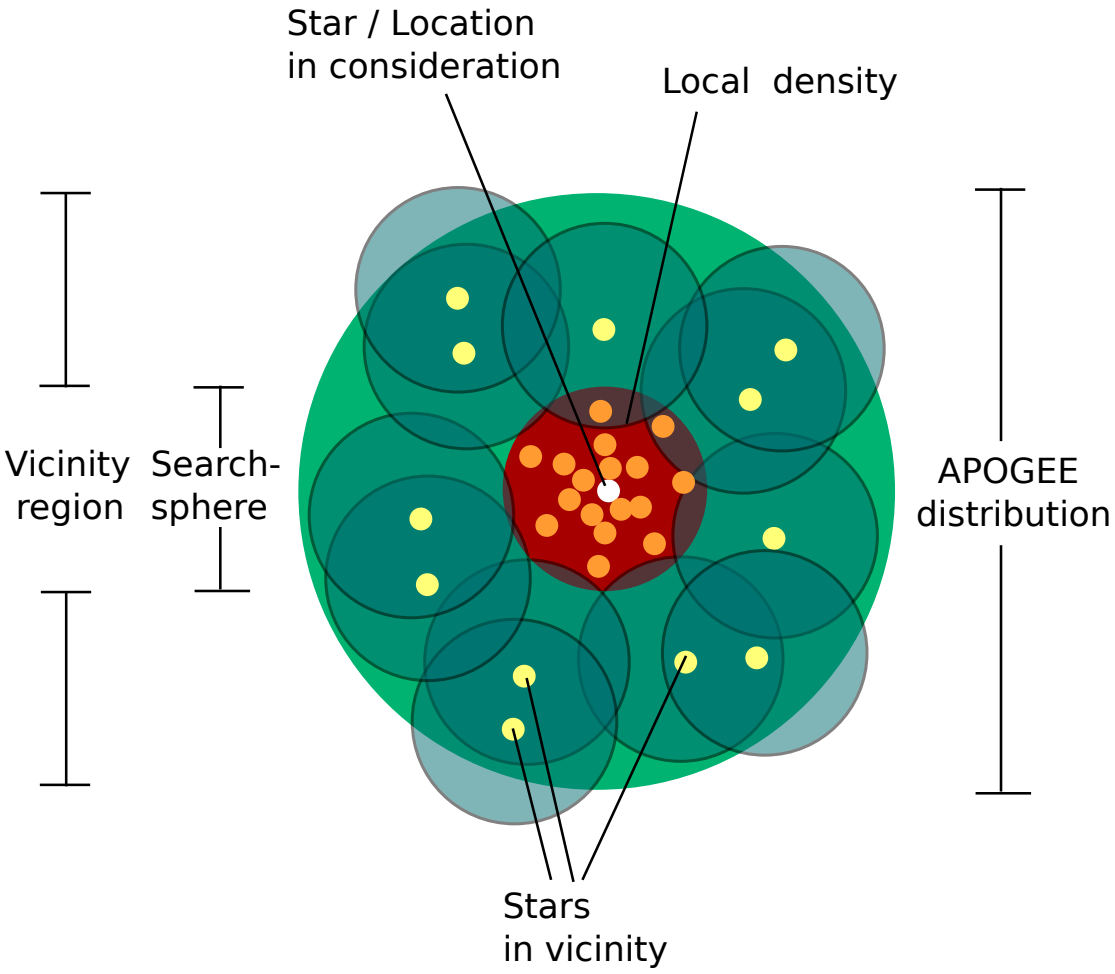

Figure 3.4: Schematic illustration of our clump search method. At each star s (white symbol), we evaluate the number of stars (orange symbols) within the search sphere (red shaded region) that we denote as n_s. The vicinity background, $\langle n_{s'} \rangle$ is calculated by averaging other $n_{s'}$, as shown in the blue shaded regions, where s' are all stars (yellow symbols) that are in the vicinity region (green shaded region). The vicinity region is defined to be the region outside the search sphere but inside two times the search sphere. The detection significance is then defined as $\sigma_{\rm detect}(s) = (n_s - \langle n_{s'} \rangle)/\sqrt{\langle n_{s'} \rangle}$.

eral. This disproportion gets more severe toward the core as the background gradient gets steeper. This disproportion causes $\sigma_{\rm dense}(n_s)$ to be an increasing function of n_s. To overcome this shortcoming and to have $\sigma_{\rm detect}$ centered around zero, we calibrate $\sigma_{\rm detect}$ by the median of $\sigma_{\rm detect}(n_s)$ at each n_s. We denote the calibrated detection significance to be $\overline{\sigma}_{\rm detect}$ and use it to be our operative measure of detection significance in the following and apply this method to the APOGEE data.

3.4 Results

We now explore what we can learn about the number of chemical cells in the APOGEE survey, and about the presence of any clumps in abundance space, that may reflect chemically tagged remnants of disrupted clusters.

3.4.1 The number of chemical cells in the APOGEE observations of the α-enhanced Galactic disk

The number of chemical cells is best estimated in the transformed (and rotated) coordinates where the global chemical distribution has no co-variances between different coordinates and a chemically homogeneous cluster can be represented by a unit Gaussian distribution. To calculate the number of chemical cells, let us estimate the global distribution to be a multivariate Gaussian with no co-variances and with standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_{10}$ in the 10 transformed coordinates. The number of chemical cells, by definition, is the volume ratio between the global distribution over the cluster kernel. Since the cluster kernel has a unit width in all directions, the number of chemical cells in APOGEE can be estimated to be $(\sigma_1 \cdot \sigma_2 \cdots \sigma_{10})/(1\,\mathrm{dex})^{10} \simeq 500$.

This estimate agrees with the prediction in Ting et al.[195] using principle components analysis. The lack of chemical cells despite having 10-dimensions is due the strong correlations among abundances, especially for the α-capture elements and Fe peak elements. The small number of chemical cells emphasizes the challenges in performing chemical tagging with strongly correlated elemental abundances. Nonetheless, as we have discussed in §3.3.2, the APOGEE abundance space has broad

wings that are not captured in a single multivariate Gaussian. Therefore, in the analysis presented below, the signals drawn from a *composite* multivariate Gaussian, as described in equation (3.1), are stronger than a simple multivariate Gaussian with 500 chemical cells.

3.4.2 Relation between N_{inject} and M_{cluster}

To understand whether a cluster will be detected or not, we first need to investigate the number of stars that we would sample in APOGEE from a cluster, given its zero age cluster mass M_{cluster}. We denote the number of stars sampled to be N_{inject} and will use the one-to-one relation between N_{inject} and M_{cluster} in the following discussion. But this conversion is based on some critical assumptions that we will now explain. The relation between N_{inject} and M_{cluster} is one-to-one up to a Poisson uncertainty – more massive clusters have more stars to begin with, and therefore will have more stars sampled in the survey.

In the limit where there is no radial migration, the relation between N_{inject} and M_{cluster} is simple and can be derived analytically. Assuming stars are azimuthally mixed in the annulus, the number of stars sampled from a cluster, N_{inject}, can be approximated (see [25,53,191]) to be

$$N_{\text{inject}} = \frac{M_{\text{cluster}}}{M_{\text{annulus}}} N_{\text{APOGEE}}, \tag{3.4}$$

where M_{annulus} is the total stellar mass (including stellar mass loss) in the annulus and $N_{\text{APOGEE}} = 13{,}000$ is the APOGEE sample size in this study. This formula can be easily understood as the following. Assuming stars in the sample have an average stellar mass $\langle M \rangle \simeq 1\,\mathrm{M}_\odot$, the ratio $N_{\text{APOGEE}}/M_{\text{annulus}}$

gives the stellar mass fraction within the annulus that we would sample in the survey. We denote this ratio to be the sampling rate. When multiplying the sampling rate by a cluster mass, the product gives the stellar mass, and thus the number of stars with $\langle \mathcal{M} \rangle = 1\,\mathrm{M}_\odot$, that we would sample from this cluster.

However, radial migration modifies $\mathcal{M}_{\mathrm{annulus}}$ in a complex way (see details in [191]). Stars born outside the annulus could migrate into the annulus, and stars from the annulus could now appear to be outside the annulus. Due to this complication, to estimate $\mathcal{M}_{\mathrm{annulus}}$, the APOGEE's selection function, as well as a robust Galactic chemical evolution model (e.g., [108,109,147]), is needed. This is clearly beyond the scope of this paper.

To simplify the problem and to only derive a conservative limit on the CMF in §3.4.3, we assume that the α-enhanced disk is completely radially mixed. In other words, the current spatial location of a star is completely random and is independent of their birth radii. In this limit, a star in the sample can be any star from the α-enhanced disk. Therefore, we have $\mathcal{M}_{\mathrm{annulus}} = \mathcal{M}_{\mathrm{total}}$, where $\mathcal{M}_{\mathrm{total}}$ is the total stellar mass of the α-enhanced disk. Although complete mixing is a crude assumption, it is likely to be reasonable for the α-enhanced disk. For example, Hayden et al. [86] showed that there is a universal α-trend irrespective of the Galactocentric radii. A natural explanation of this result is the stars in the α-enhanced disk are well-mixed. Moreover, the APOGEE sample covers a wide range of Galactocentric radius, with $R_{\mathrm{gc}} \simeq 3\,\mathrm{kpc} - 15\,\mathrm{kpc}$ [33,86,154]. It should have sampled the α-enhanced disk from a large fraction of the Milky Way.

We emphasize that the complete mixing assumption gives a conservative limit on the CMF. In the case where the mixing is not complete, we would have sampled more stars from the same clus-

ter, and hence, it would be easier to exceed the APOGEE baseline. On top of the complete mixing assumption, we also assume that the CMF is independent of elemental abundances, and hence the equation (3.4) applies universally to the whole abundance space.

With these assumptions, we only need to properly estimate $\mathcal{M}_{\rm total}$ and apply equation (3.4) to obtain a one-to-one relation between $N_{\rm inject}$ and $\mathcal{M}_{\rm cluster}$. We assume that the α-enhanced disk has an exponential scale length of 3 kpc (e.g., [31]) and consists of 10% stellar mass observed in the solar neighborhood (e.g., [42]). We adopt the stellar density in the solar neighborhood to be 38 $M_\odot pc^{-2}$ (e.g., [30,65,213]), and the solar Galactocentric radius $R_\odot = 8$ kpc (e.g., [74,75,169]). These assumptions yield a present-day α-enhanced disk stellar mass of $\sim 3 \times 10^9$ $M_\odot$. Since the α-enhanced disk is old, massive stars have long since evolved and died. To account for this, we consider a total stellar mass loss of 40% ([45], assuming a Kroupa IMF). Putting all these together, we have $\mathcal{M}_{\rm total} \simeq 6 \times 10^9$ $M_\odot$. We also derive the sampling rate of the current APOGEE sample to be $N_{\rm APOGEE}/\mathcal{M}_{\rm total} = \frac{1}{5\times10^5}$. On average, we would collect one star from a 5×10^5 $M_\odot$ cluster. We will defer the discussion on what this low sampling rate implies in §3.4.3.

3.4.3 Chemical tagging in APOGEE

We apply the clump search method described in §3.3 to the APOGEE sample. The left panel of Fig. 3.5 shows $\overline{\sigma}_{\rm detect}$ as a function of n_s of all 13,000 stars. At face value, it is tantalizing to observe deviations $> 5\sigma$. But we emphasize that the detection significance depends on the various assumptions made, such as the choice of search sphere radius, the minimum number of neighbors requirement, the detection significance calibration and the definition of vicinity region at each data point.

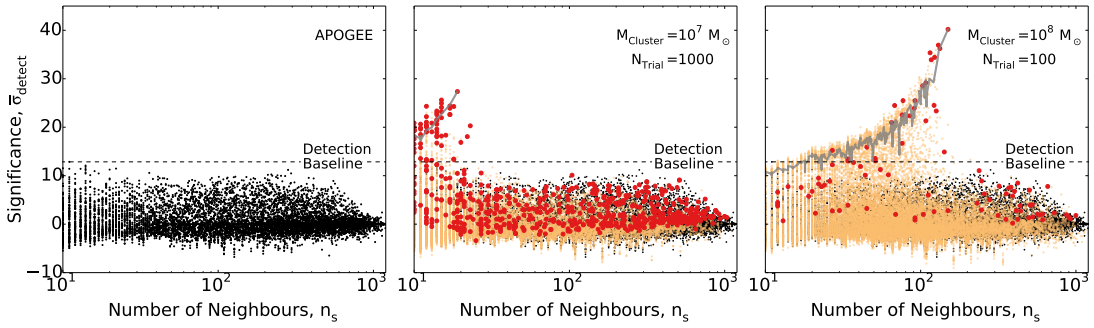

Figure 3.5: Calibrated detection significance as a function of local density in abundance space. The left panel shows the sample observed values. The dashed lines show the observation detection baseline. The overlaid orange symbols in the middle and right panels show results from the injected $10^7\ M_\odot$ and $10^8\ M_\odot$ clusters, respectively. Since we calculate the local density centered at *each* star, there are numerous points for each cluster. We highlight the maximum deviation of each cluster with a bold red symbol. We compile results from 1,000 and 100 trials for the middle and right panels. To demonstrate a typical detection significance distribution within a cluster, we link with solid gray lines all stars from the cluster containing the star with the highest detection significance. The middle and right panels show that if a $10^7 - 10^8\ M_\odot$ cluster formed in the past, some cluster members might show more deviations than the detection baseline.

Without further information such as stellar ages, it is difficult to confirm the origin of these clumps. Furthermore, as shown in Ting et al. [191], most clumps are comprised of many clusters sharing similar elemental abundances.

Due to these uncertainties, instead of interpreting these clumps as detections, we proceed by assuming the APOGEE dataset (left panel of Fig. 3.5) to be the detection baseline. We inject mock clusters of different sizes into the data and estimate the detection significance of these injected objects. By forward modeling, we *rule out cases that are not consistent with the observation baseline*. The middle and right panels show $\overline{\sigma}_{\text{detect}}$ of the injected objects. In the right panel, we combine results of 100 trials, where in each case we inject a $10^8 M_\odot$ ($N_{\text{inject}} \simeq 250$ stars) clump into the data. In the middle panel, we show the results of 1000 trials with $10^7 M_\odot$ clusters ($N_{\text{inject}} \simeq 25$ stars) injected. These two panels show that if $10^7 - 10^8 M_\odot$ clusters have formed in the past, there is a reasonable chance that we would have detected larger deviations than the value observed. A cluster with $10^7 M_\odot$ lies above the detection boundary about $\sim 7\%$ of the time, and a cluster with $10^8 M_\odot$ is detected about $\sim 30\%$ of the time.

Not all high mass clusters will exceed the detection baseline. As shown in the middle and right panels, most clusters, especially at high n_s, blend into the background. To make robust statements, we now proceed to quantify the probability of a cluster exceeding the observation baseline.

Detection probability of individual clusters

In this section, we will quantify the probability of an injected clump exceeding the observation baseline. There are two key parameters that determine this probability: (a) the number of stars injected

as a clump, $N_{\rm inject}$, and (b) the cluster location in abundance space. As for the latter, the clump centers are drawn from the intrinsic distribution model described in §3.3.2. As for the former, we consider a grid of $N_{\rm inject}$, ranging from $3 - 1000$ stars with a step-size of 0.2 in log scale. When injecting a mock cluster with $N_{\rm inject}$ stars, we allow a Poisson fluctuation of $\sqrt{N_{\rm inject}}$. The $N_{\rm inject}$ range in this study roughly corresponds to cluster masses of $10^6 M_\odot - 5 \times 10^8 M_\odot$. We run 10^4 trials for each $N_{\rm inject}$ and find that the Monte Carlo uncertainty is negligible with this many trials.

We model the cluster location by ranking all 10^4 trials by their $P_{\rm intrinsic}(\mathbf{X}_{\rm center})$ value. $\mathbf{X}_{\rm center}$ is the clump center location in abundance space. We put the ranking into a linear scale, which we will denote as $\rho(\mathbf{X}_{\rm center}) \in [0, 100]$, where

$$\rho(\mathbf{X}_{\rm center}) \equiv \frac{\#(\text{trials} < P_{\rm intrinsic}(\mathbf{X}_{\rm center}))}{\#(\text{trials})}. \tag{3.5}$$

If the cluster is located near the background dominated core, it has a higher value in $P_{\rm intrinsic}$ because the background density is very large, and we assign a high $\rho(\mathbf{X}_{\rm center})$. Whereas, if the cluster is located at the peripheral regions, it has a lower $P_{\rm intrinsic}$ value since the background density is low, and we assign a low $\rho(\mathbf{X}_{\rm center})$.

For each trial, we inject a clump and estimate the local density and vicinity background for all objects from the injected clump the same way in §3.3.3. We define a clump to have exceeded the observational baseline if the maximum detection significance of this clump exceeds the baseline as demarcated by the dashed lines in Fig. 3.5. We take the maximum significance of the whole clump because not all injected objects will have high deviations, as shown in the solid gray lines in Fig. 3.5.

Objects located near the surface of a clump will blend into the background. Only the objects near the clump center will have high deviations because the search sphere includes a large fraction of the clump.

The top panel of Fig. 3.6 shows the probability of exceeding the APOGEE baseline. Among all trials that have a similar cluster location and a similar number of stars, we evaluate the fraction of them exceeding the baseline. The x-axis shows $N_{\rm inject}$ and the y-axis shows the cluster location quantified by $\rho(\mathbf{X}_{\rm center})$. This panel illustrates that as the number of injected objects increases or the cluster location is increasingly toward the peripheral regions, the chance of exceeding the baseline improves, consistent with our intuition from Fig. 3.5.

The solid line in the bottom panel of Fig. 3.6 shows the probability marginalized over the cluster location, i.e., the probability of a cluster exceeding the APOGEE baseline as a function of its cluster mass if the cluster location is randomly drawn from the intrinsic abundance distribution. The marginalized probability shows that clusters less massive than $10^7\,{\rm M}_\odot$ have negligible chances of exceeding the baseline, but clusters more massive than $\sim 10^7\,{\rm M}_\odot$ begin to show tension with the deviations observed in APOGEE. The bottom panel also illustrates that even for a cluster as massive as $\sim 5\times10^8\,{\rm M}_\odot$ ($N_{\rm inject}\simeq 1000$), only about half of the time will a cluster exceed the baseline. The lack of significant detection from the other half is not unexpected. As illustrated in the right panel of Fig. 3.5, if a cluster is located in the core region (i.e., high n_s), the background becomes dominant. In this regime, most objects within the search sphere come from background contaminants. Therefore, in the core region, the signal tends to be overwhelmed by the background, regardless of the cluster size.

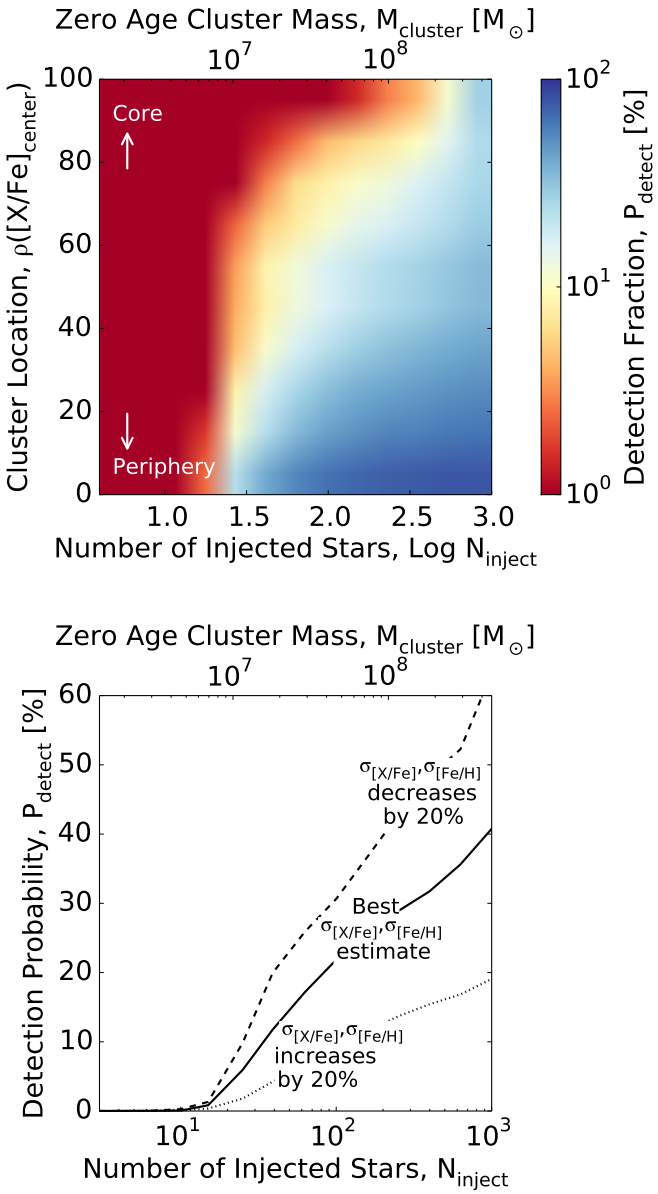

Figure 3.6: Probability of an injected cluster showing more significant detection than the APOGEE data. This probability depends on two parameters: the number of injected stars per cluster and the cluster location in abundance space. The top panel shows the probability as a function of these two parameters. As the number of injected stars increases or the cluster is injected in a more peripheral region of abundance space, the chance to exceed the observation baseline increases. The bottom panel shows the probability marginalized over the cluster location, i.e., the probability of detecting a cluster of a certain cluster mass if the cluster location is randomly drawn from the intrinsic abundance distribution. The solid line shows result assuming the best cluster kernel estimation, $\sigma_{\mathbf{X}}$, as also applied to the top panel. The dashed and dotted lines show the marginal probability assuming $\pm 20\%$ statistical uncertainties of the $\sigma_{\mathbf{X}}$ estimate due to the small sample of cluster members (see §3.2 for details).

Recall that our estimate of the cluster kernel $\sigma_{\mathbf{X}}$ has an uncertainty of 20% due to the small number of cluster stars. Therefore, the concentration of our injected clusters could be off by the same amount. We also explore how this uncertainty might change our results. The dashed and dotted lines show results in cases where our cluster concentration estimate is off by 20%. The dashed line shows the result assuming chemically homogeneous clusters are intrinsically tighter in abundance space by 20%. With a more concentrated signal, the signal will have a better contrast over the background. Therefore, clumps are easier to detect and the probability in Fig. 3.6 increases. However, if clusters are more widely spread, they are more likely to blend into the background. Therefore, the chance of detection decreases, as shown in the dotted line. We defer more detail discussions on how this uncertainty changes our conclusion to §3.4.3.

Limits on the CMF

So far we have only studied the detection probability of an individual cluster injected into the APOGEE data. For example, in the bottom panel of Fig. 3.6, we derived the probability of detecting a cluster as a function of its cluster mass, which we will denote as $P_{\rm detect}(\mathcal{M}_{\rm cluster})$. In this section, we will propagate this individual cluster statistic to constrain the CMF. We derive the total number of clusters of different masses from $\mathcal{M}_{\rm total}$ and the parameters of the CMF. Using this information, we can then evaluate the probability of all these predicted clusters being consistent with the APOGEE observation, which will then place a limit on the CMF. We assume a power-law CMF with a low-end cutoff of 30 $M_\odot$ and then constrain the power-law slope, α, and the high-end cutoff, $\mathcal{M}_{\rm cutoff}$ from the comparison with APOGEE data.

Let's formulate this idea more rigorously. Given a CMF, we know that, on average, there are a total of $\overline{n} = \mathcal{M}_{\rm total}/\overline{\mathcal{M}}$ clusters spawned where $\overline{\mathcal{M}}$ is the mean cluster mass from the CMF. By definition, the cluster masses of $\overline{n}$ clusters follow the CMF, which we will denote as $\mathcal{M}_{\rm cluster,i=1}$, $\mathcal{M}_{\rm cluster,i=2}$, ..., $\mathcal{M}_{\rm cluster,i=\overline{n}}$. The probability $\mathcal{L}(\rm CMF)$ that all these clusters are consistent with the data is

$$\mathcal{L}(\mathrm{CMF}(\alpha, \mathcal{M}_{\rm cutoff})) = \prod_{i=1}^{\overline{n}} (1 - P_{\rm detect}(\mathcal{M}_{\rm cluster,i})), \tag{3.6}$$

i.e., none of these clusters exceeds the observation baseline. In practice, to save computational time and to ensure a well-converged solution, we evaluate the $\mathcal{L}(\rm CMF)$ analytically.

The left panel of Fig. 3.7 shows the resulting $\mathcal{L}(\rm CMF)$. The figure demonstrates that if the CMF slope is shallower than $\alpha = -2$, the APOGEE sample is mostly consistent with a high-end cutoff $\lesssim 3 \times 10^7\,\mathrm{M}_\odot$ ($\log \mathcal{M}_{\rm cutoff} = 7.5$). Qualitatively, this result should be expected. As shown in the bottom panel of Fig. 3.6, there is a $\sim 10\%$ chance that a cluster with $\sim 10^7\,\mathrm{M}_\odot$ will exceed the observation baseline. Recall that if the CMF slope $\alpha = -2$, we have equal contributions from all logarithmic mass bins. This implies that the number of clusters with mass $\sim 10^7\,\mathrm{M}_\odot$ is of the order $\sim \mathcal{M}_{\rm total}/10^7\,\mathrm{M}_\odot \sim 100$. Let say there are 50 such clusters, and each cluster only exceeds the baseline $\sim 10\%$ of the time. The probability that all of them would be consistent with the APOGEE observation is still extremely unlikely because $(90\%)^{50} < 1\%$.

This simple illustration also demonstrates two important features. First, the detection probability is very low for individual clusters with masses $< 10^7\,\mathrm{M}_\odot$. If the CMF slope is steeper than -2, most clusters are not massive. In this case, the APOGEE observation provides a very weak constraint on

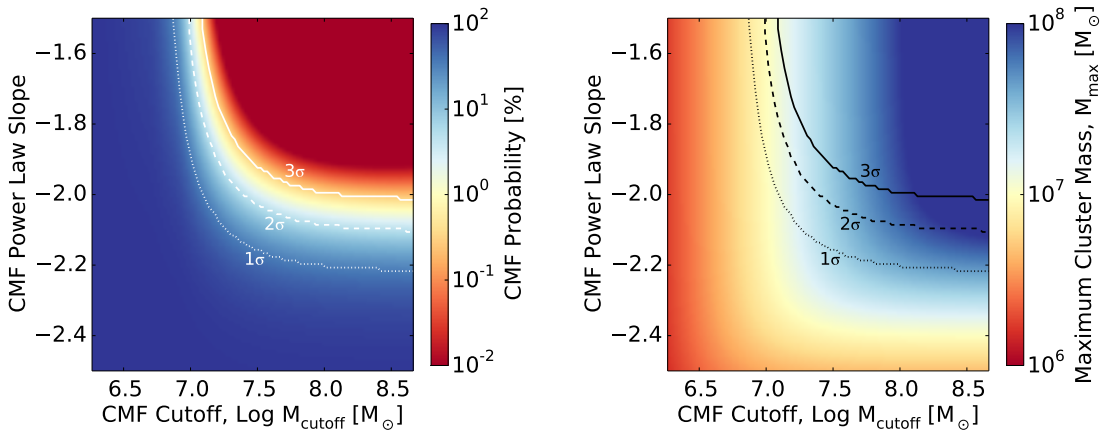

Figure 3.7: Constraint on the α-enhanced disk CMF. The left panel shows the probability of a CMF being consistent with the APOGEE DR12 data as a function of the two CMF parameters, the power-law slope and the upper mass cutoff. The dotted, dashed, and solid lines show the $1-3$ sigma limits, respectively. Unless the CMF power-law slope is steeper than -2, a cluster mass cutoff $\gtrsim 3\times10^7\,M_\odot$ is largely ruled out. The right panel shows the maximum cluster mass, $\mathcal{M}_{\rm max}$, for different CMFs, such that the expected number of clusters $N_{\rm cluster}(>\mathcal{M}_{\rm max})>1$ (see §3.4.3 for details). We overplot the $1-3$ sigma limits calculated from the left panel. The right panel shows that, in most cases, there is on average less than one cluster with $\gtrsim 3\times10^7$ formed in the Milky Way.

the high mass cutoff. As shown in the left panel of Fig. 3.7, if $\alpha \lesssim -2$, we cannot constrain the CMF cutoff. Even though a cluster with mass $\sim 10^8$ $M_\odot$ would easily exceed the baseline, these clusters are extremely rare if the CMF slope is steeper than -2.

Secondly, since the CMF constraint is derived from the product of each detection probability, it is sensitive to $P_{\rm detect}$. As shown in the bottom panel of Fig. 3.6, if our estimate of $\sigma_{\mathbf{X}}$ is off by 20%, it will affect $P_{\rm detect}$, which in turn could dramatically modify our CMF constraint. If clusters are more concentrated in abundance space than we have assumed here, then that will provide a stronger constraint on each detection (dashed line in Fig. 3.6). Therefore, our constraint on the CMF would be conservative. On the other hand, if we have underestimated $\sigma_{\mathbf{X}}$ by 20%, then the clusters would be more widely spread out in abundance space than we have assumed. In this case, most clusters would be harder to detect (dotted line in Fig. 3.6). Although not shown, we have checked that, in this case, we can only rule out CMF with $\mathcal{M}_{\rm cutoff} \gtrsim 10^8$ $M_\odot$ and $\alpha \gtrsim -1.9$.

Nonetheless, independent evidence seems to support our $\sigma_{\mathbf{X}}$ estimate. Mathematically, the rotation that we performed in §3.3.2 is exactly the same as principal components analysis (see appendix in [195]). After the rotation, each coordinate becomes a principal component of the APOGEE abundance space. The APOGEE abundance space has fewer independent dimensions than the observed dimensions [6,195]. Some of these 10 principal components should have very little intrinsic scatter. Therefore, some minor axes of the 10-dimensional ellipsoid are only due to the measurement scatter. Thus, their spreads should be a robust estimate of the measurement uncertainty $\sigma_{\mathbf{X}}$. We find that the widths of these minor axes are consistent with our $\sigma_{\mathbf{X}}$ estimate, showing our estimation of $\sigma_{\mathbf{X}}$ is robust. Therefore, our conservative CMF constraint is likely to hold.

Although not all $\mathcal{M}_{\rm cutoff} \gtrsim 3 \times 10^7\,{\rm M}_\odot$ is ruled out, Fig. 3.7 shows that a very high $\mathcal{M}_{\rm cutoff}$ is only consistent with the data when the CMF slope is steeper than $\alpha = -2$. When the CMF is steep, the number of massive clusters also decreases precipitously. Therefore, the cutoff could be very massive, yet on average there might be less than one such massive cluster in the Milky Way. A high cutoff does not naturally imply the existence of these clusters. Instead of $\mathcal{M}_{\rm cutoff}$, perhaps a more useful constraint is the maximum cluster mass such that we expect to have at least one cluster larger than this mass. We denote this maximum mass to be $\mathcal{M}_{\rm max}$. Assuming $\mathcal{M}_{\rm total} = 6 \times 10^9\,{\rm M}_\odot$, we show $\mathcal{M}_{\rm max}$ as a function of the CMF parameters in the right panel of Fig. 3.7. As expected, this panels shows that when the slope is steep, we have $\mathcal{M}_{\rm max} \ll \mathcal{M}_{\rm cutoff}$, i.e., the cluster mass cutoff is never achieved. When overplotted with the constraints obtained in the left panel, the right panel shows that in most cases, only clusters with masses $\lesssim 3 \times 10^7$ could have formed. As the α-enhanced disk is believed to form in the first 5 billion years (e.g., [88]), our constraints refer to the portion of the disk that formed at $z > 1$.

We have made numerous assumptions in this study, but we emphasize that we have always made the conservative choices. Therefore, our CMF limit should be robust as long as we did not underestimate $\sigma_{\mathbf{X}}$ by 20% and the ellipsoidal intrinsic distribution is a fair representation of the deconvolved distribution. A question remains to be answered: could we obtain a significantly stronger constraint on the CMF using the current APOGEE data? We would argue that the answer is likely no. The bottleneck is intrinsically due to the relatively small number of volume elements in abundance space and the low sampling rate. The former is set by the precision of the abundance measurements and the number of independent dimensions in abundance space sampled by the APOGEE spectra. The

APOGEE sampling rate is of the order $N_{\rm APOGEE}/M_{\rm total} = \mathcal{O}(10^{-5})$, which implies that we would only sample one star from a $\sim 10^5$ $M_\odot$ cluster. The threshold $n_s > 10$ implies that the minimum cluster mass needed is $\sim 10^6$ $M_\odot$. Therefore, in the most optimistic case, we might be able to put a stronger limit by at most an order of magnitude.

How does our stellar CMF limit compare to high redshift observations of star forming galaxies? Recent observations have reported the existence of giant star forming clumps within the disks of star forming galaxies at $z \sim 2$[66,72,73,97,132]. Gas clumps as high as $\sim 10^9$ $M_\odot$ have been observed. There has been much speculation regarding the properties and fate of these giant clumps. Some have argued that they are progenitors of globular clusters[184]. They may migrate by dynamical friction to the centers of galaxies[208]. It is also unclear if the stars forming within these giant clumps contain stars that are coeval or share a common metallicity. What is clear from the results presented in this work is that these giant star forming clumps cannot both be mono-abundance and remain in the α-enhanced disk, at least in the portion of the Milky Way observed by APOGEE (i.e., with a Galactocentric radius of 3 kpc $-$ 15 kpc). Even assuming a total star formation efficiency of 1% (simulations and observations usually show higher values, e.g.[58,61,104,115,117]), these gas clumps would have formed clusters that are at least 10^7 $M_\odot$ and would have stood out in the chemical tagging experiment presented here if they are chemically homogeneous and that they remain in the Milky Way disk, which APOGEE is probing.

3.4.4 Comparison with previous studies

The first chemical tagging experiment on dispersed disk stars was performed by Mitschang et al.[151], and the tagged groups were subsequently studied in Quillen et al.[163]. They studied 714 stars in the solar neighborhood from Bensby et al.[13]. Our results agree with their assessments that the identified groups in these studies are probably not co-natal stars. Each group is unlikely coming from a single disrupted cluster, even though the clump members might be coeval stars as they share similar abundances. In these studies, the sample includes both α-enhanced stars and low-α stars, but the sample size is about ten times smaller than the APOGEE α-enhanced sample. The sampling rate in Bensby et al.[13] is therefore much smaller than the APOGEE α-enhanced sample. Recall that the sampling rate in this study is $\sim \mathcal{O}(10^{-5})$, and thus, we deduce that the sampling rate in these early studies is $\ll \mathcal{O}(10^{-5})/10 = \mathcal{O}(10^{-6})$. If groups detected in these studies were to come from individually disrupted clusters, the parent cluster would have a mass $\gg 10^6 M_\odot$, consistent with the estimates in Quillen et al.[163].

Simulations from Ting et al.[191] also disfavor a co-natal interpretation of the groups identified in these earlier studies. Ting et al.[191] found that even if such large clusters exist, a detected clump in abundance space will still have a sizable background component. More importantly, in the case with a dominant background, the applicability of previous clump search techniques that separate the abundance space into a few distinct regions, such as the one proposed in Mitschang et al.[151], or other tree-based methods (e.g.[135]) is questionable. For those techniques to perform well, the background in abundance space has to be negligible or first be subtracted. As we have explored in this study, the

background estimation can be challenging given its complex topology in high-dimensional space and the fact that the signal is usually overwhelmed by the background contaminants.

3.5 Summary and conclusion

In this study we have exploited the superb APOGEE DR12 data, with typical uncertainties of $\sigma_{[X/Fe]}$ and $\sigma_{[Fe/H]} \simeq 0.05$ for 10 elements measured for $> 10^4$ stars, to put in practice a first large-scale chemical tagging analysis of the α-enhanced disk. Because the number of stars per 10-dimensional abundance volume is lower in the α-enhanced disk, we focused on that portion of abundance space.

This analysis required the development of a new, simple algorithm for identifying clumps in abundance space, and it brought some of the "real life" difficulties of chemical tagging to the fore. Nonetheless, we succeeded in providing the first abundance-based constraints on the masses and mass functions of chemically homogeneous star clusters in the old Galactic disk.

The methodological steps and results can be summarized as follows:

- We determined and applied a coordinate transformation that makes the cluster kernel in abundance space spherical (in 10-dimensions) and have unit variance in each dimension. This kernel enables fast error deconvolutions in this transformed abundance space.
- We generated a model for the intrinsic abundance distribution of the α-enhanced disk, presuming it to be a highly anisotropic and co-variant ellipsoidal distribution in the above 10-dimensional transformed abundance space. After rotating this coordinate system to eliminate the co-variances in this distribution, we modeled each dimension independently as the sum of two Gaussians. Fitting this to the APOGEE data provides a first estimate of the shape and volume of the error-deconvolved abundance space of the α-enhanced Galactic disk.
- We found that despite the unprecedented quality of the APOGEE data, the volume occupied by the stars of the α-enhanced Galactic disk is only ~ 500 times the volume of the cluster kernel. Even with abundance uncertainties of $\sigma_{[X/Fe]}$ and $\sigma_{[Fe/H]} \simeq 0.05$ dex, the cluster kernel spans $> 30\%$ of the abundance width in each elemental abundance dimension. In addition, many of the 10 elemental abundances measured by APOGEE and used herein are highly co-variant.

- We developed an algorithm to detect groups of chemically homogeneous stars, geared toward the background-dominated regime. We found that searching for chemically homogeneous clumps is challenging with high backgrounds. The chemical tagging signals will most likely come from the peripheral regions in abundance space where the background density is relatively low.
- Using APOGEE data as a detection baseline, we were able to constrain the CMF in the Galactic α-enhanced disk. We show that this population is unlikely to have formed clusters more massive than 3×10^7 $M_\odot$ at any point in its history.

Although the current constraints presented in this work are limited to very large cluster masses, the results in this paper vividly demonstrate the potential of chemical tagging in understanding the Milky Way properties in the past. With more data currently being collected by on-going surveys, we should be able to provide much stronger constraints on the CMF in the near future.

4

Accelerated fitting of stellar spectra

Author list of the original paper: Yuan-Sen Ting, Charlie Conroy, Hans-Walter Rix

4.0 Abstract

Stellar spectra are often modeled and fit by interpolating within a rectilinear grid of synthetic spectra to derive the stars' labels: stellar parameters and elemental abundances. However, the number of synthetic spectra needed for a rectilinear grid grows exponentially with the label space dimensions, precluding the simultaneous and self-consistent fitting of more than a few elemental abundances. Shortcuts such as fitting subsets of labels separately can introduce unknown systematics and do not produce correct error covariances in the derived labels. In this paper we present a new approach – CHAT (Convex Hull Adaptive Tessellation) – which includes several new ideas for inexpensively generating a sufficient stellar synthetic library, using linear algebra and the concept of an adaptive, data-driven grid. A convex hull approximates the region where the data lie in the label space. A variety of tests with mock datasets demonstrate that CHAT can reduce the number of required synthetic model calculations by three orders of magnitude in an 8D label space. The reduction will be even larger for higher-dimensional label spaces. In CHAT the computational effort increases only linearly with the number of labels that are fit simultaneously. Around each of these grid points in label space an approximate synthetic spectrum can be generated through linear expansion using a set of "gradient spectra" that represent flux derivatives at every wavelength point with respect to all labels. These techniques provide new opportunities to fit the full stellar spectra from large surveys with $15 - 30$ labels simultaneously.

4.1 Background

Despite many decades of research, many aspects of Milky Way evolution (see review from [170]) and the Local Group galaxies (e.g., [105,205]) remain unsettled. To unravel the formation history of the Milky Way, spectroscopic surveys are currently being carried out to gather elemental abundances and kinematic information of stars across the Galaxy. High-resolution spectra of $10^5 - 10^6$ stars are being collected through surveys such as APOGEE [136], GALAH [53] and Gaia-ESO [187] and with the resolution power of $R \simeq 20{,}000$ and S/N $\simeq 100$. With the exquisite spectra of these stars, the goal is to measure $15 - 30$ elemental abundances of each star as precise as possible. Since most stars are long lived and the elemental abundances of galaxies built up gradually over time, these abundances are tell-tale signs of the Milky Way's evolution. Furthermore, stars that formed together are believed to share exceptionally similar elemental abundances (e.g., [29,50,70,194]). By looking for stars that share similar abundances, one goal in these surveys is to reconstruct star clusters that are now disrupted and dispersed in the Milky Way (e.g., [48,112,124]), an idea commonly known as chemical tagging [23,24,69,191].

Identifying members of disrupted star cluster is an important missing piece to understanding the Milky Way. Stars are believed to have migrated from their birth orbit since they formed either through "radial migration" (see observational evidence from [113,133]) or "blurring" of orbits (e.g., through n-body scattering). For example, it has been proposed that stars could radially migrate when corotating with transient structures such as the Galactic bar and spiral arms [55,81,149,176]. But quantitative, direct observational evidence for radial migration remains scarce, and chemical tagging can provide it.

Chemically tagging disrupted star clusters also informs us about the past star cluster mass function (e.g., [192]). This information is crucial as studies (e.g., [59,115]) have shown that the maximum aggregate size that star formed might depend on the star formation rate and the gas mass in the past. Since most of the clusters are soon disrupted [124] constraining the past cluster mass function through chemical tagging may be the best option [25,192] .

But all this requires efficient and precise abundance determinations from vast sets of observed spectra. This is challenging for two reasons. First, synthetic spectra for large surveys have systematic uncertainties, as 1D models in local thermodynamic equilibrium (LTE) are typically used for generating a synthetic spectral library (e.g., [71,187]). Studies have shown that, at least for metal-poor stars, 3D non-LTE calculations are essential for accurate recovery of labels (e.g., [14]). Second, generating synthetic spectra is computationally expensive. Even for 1D-LTE models, each synthetic spectrum can take hours to generate, which renders the generation of a synthetic library with $15 - 30$ elemental abundances impossible with the rectilinear grid approach.*

In this paper, we will tackle the second challenge by presenting CHAT (Convex Hull Adaptive Tessellation), a set of techniques for fitting stellar spectra by generating a synthetic library using the idea of an adaptive grid and a convex hull. Our method reduces the complicated interpolation-minimization process into a simple series of linear regressions. In §4.2 we will discuss the limitations of the rectilinear grid approach. In §4.3 we describe the idea and the implementation of our method. We present a comparison of our method with the rectilinear grid approach in §4.4 and show that our proposed method here can reduce the number of models by three orders of magnitude in an

*A rectilinear grid is a model grid that has uniform spacing with a fixed interval for each label.

8D label space, and the reduction will be more significant at a higher dimensional label space. This method opens up new possibilities to perform an ab-initio fitting of observed spectra with more labels. We explore some of these possibilities in §4.4, and we conclude in §4.5. We emphasize that these techniques can be used for fitting any set of synthetic spectra to observations. In this paper we focus on 1D LTE models but note that as 3D non-LTE models become computationally affordable, CHAT can be applied to those models as well.

In this paper we focus on techniques directly applicable to automated pipelines for large surveys in which the full spectrum (or portions thereof) are fit to models. However, many of the techniques discussed here are also applicable to the classical technique of fitting equivalent widths of selected spectral features.

4.2 The rectilinear grid and its limitations

For a rectilinear grid, the number of models grows exponentially with the number of dimensions. For example, in the case of the APOGEE survey[92], the rectilinear grid is comprised of 6 main labels including $T_{\rm eff}$, $\log g$, the overall metallicity $[Z/{\rm H}]$, the α-enhancement $[\alpha/Z]$, $[{\rm C}/Z]$ and $[{\rm N}/Z]$. A mere five grid points per dimension would require $5^6 \simeq$ 15,000 models. Therefore, any additional dimension such as microturbulence $v_{\rm turb}$, stellar rotation $v \sin i$ or additional abundances $[X/{\rm H}]$, are very computationally expensive to include. Consequently, the full observed spectra are in practice only fit with these few main labels. Other elements are individually determined in a second step by fitting narrow spectroscopic windows and assuming a fixed underlying atmosphere. But this two-step approach entails several potential problems:

1. As we show in Appendix 4.6, a wide array of elements impact the atmospheric structure. By only considering a subset of important elements when computing grids of atmospheres, one introduces biases in the final spectra. This issue is more apparent for the low-$T_{\rm eff}$ stars (e.g., below $\sim$ 4000 K). Therefore, the two-step approach could introduce non-negligible systematic biases when determining the photosphere structure by only fitting the main labels.

2. Fitting the full spectra with only a few basic labels (e.g., $T_{\rm eff}$, $\log g$, $[Z/{\rm H}]$ and $[\alpha/Z]$) requires assumptions on how the other (not fit) elements trace those labels. A common assumption is that all α-elements trace each other, and all other elements trace [Fe/H]. Although this is a good working assumption, without which the fitting would be much worse, this assumption is not true in detail. Therefore, this simplification incurs systematic offsets in the determinations of the main labels such as $T_{\rm eff}$, $\log g$ and $v_{\rm turb}$.

3. By fitting elemental abundances one at a time, with fixed atmosphere structure, one cannot evaluate their covariances with other labels. For example, elemental abundance determinations depend on $T_{\rm eff}$, and therefore, abundances must be correlated to some level. The covariance matrix is crucial for any chemical tagging studies. Although stars that formed together

are believed to be homogeneous to the level of 0.05 dex (e.g.,[50,70,194]) or better[29,131], the spread of a cluster in elemental abundances space is dominated by the measurement uncertainties. Therefore, to look for overdensities, we would need to evaluate this spread of a chemical homogeneous cluster in the elemental abundances space[130]. Ting et al. [192] showed that ignoring the covariance matrix and using only the marginal uncertainty of each element will increase the background contamination in a search for overdensities by a factor of 10^4 in a 10D space. But any estimate of the covariance matrix requires that all labels are fit simultaneously, again incurring prohibitive computational expense with rectilinear grid fitting.

4. Restricting the fits of individual elemental abundances to "clean" narrow spectral windows excludes other spectral information such as blended lines. From information theory, one can calculate the theoretically achievable precision with the Cramer-Rao bound[46,165]. If we have accurate synthetic spectra and an effective way to fit all elemental abundances simultaneously using the full spectrum, as illustrated in Appendix 4.7, we could in principle achieve a precision of $\sim$ 0.01 dex for APOGEE data if the systematic errors in the models are smaller than this limit. For comparison, the windows currently defined by the APOGEE pipeline only exploit $\sim$ 10% of the spectral information, implying that abundance precision will be $\sqrt{10} \sim 3$ times worse than the formal limit (see Appendix 4.7 for details). We note that these ideal theoretical precisions might not be achievable due to systematic uncertainties in the models, but we should still expect a decent improvement in the precision as we are using more information in the spectra.

5. To improve the interpolation within a rectilinear grid, sophisticated algorithms can be employed. For example, in the case of the APOGEE survey, the interpolation is done with a cubic Bézier function. For this algorithm, many spectra from the rectilinear grid are required for each iteration. This implementation is memory intensive, to the extent that not all wavelength points from each spectrum are saved[71]. Certain compressions are needed, and some information within the spectra is unavoidably discarded in the compression process. As we will discuss in more details in §4.3, the method proposed in this paper reduces the complicated interpolation-minimization process into a simple series of linear regressions. This method is extremely memory effective, and no compression of spectra is needed.

4.3 A new approach

To overcome the challenges discussed in §4.2, we propose a new approach to fitting stellar spectra with ab-initio models, which we call CHAT (Convex Hull Adaptive Tessellation) that has two central elements: (a) around any point in label space for which we have calculated a synthetic spectrum, there exists a (high-dimensional) hypersphere in label space within which the spectrum (i.e., the flux at every wavelength point of a normalized spectrum) varies linearly with changes in any of the labels; as this region is defined through its series expansion, we refer to this region as a "Taylor-sphere."† The number of synthetic spectral models needed to describe any spectrum that lies within the Taylor-sphere of a model grid point grows linearly with the dimensionality of label space, instead of exponentially. (b) Rectilinear grids in high-dimensional spaces are highly inefficient in covering high-dimensional distributions, especially ones that are as correlated and irregular as the distribution of stars in abundance space. Here we develop a data-driven approach to finding a near-minimal set of grid points whose surrounding Taylor-spheres cover all of the relevant label space. Obviously, this requires at least some a priori knowledge of the distribution of stars in the label space. In addition, CHAT simplifies the interpolation-minimization spectral fitting process to a series of linear regression problems around a manageable set of grid points. In §4.3.1, we will expand the basic ideas of the method. In §4.3.2, we discuss some of the attractive properties of this method compared to the rectilinear grid approach, and we move on to the implementation details in §4.3.3.

†Strictly speaking, since we are looking for the best linear interpolations given fixed end points, a more appropriate nomenclature should be "Legendre-sphere", because (a) we are searching for the best multivariate Legendre polynomial approximations to the first order and (b) we are not calculating the gradient spectra with infinitesimal changes in label space.

4.3.1 Basic concepts

Gradient spectra and Taylor-spheres

The predicted continuum-normalized flux of a spectrum, $f_{\rm model}(\lambda|\boldsymbol{\ell})$ of a model that is specified by a set of labels, $\boldsymbol{\ell}$, changes from point to point in that space, but does so "smoothly", or differentiably. The spectrum corresponding to any $\boldsymbol{\ell}$ sufficiently close to a model grid point $\boldsymbol{\ell}_*$ can therefore be described with sufficient accuracy by

$$f^{\rm lin}_{\rm model}(\lambda|\boldsymbol{\ell}_* + \boldsymbol{\Delta\ell}) \simeq f_{\rm model}(\lambda|\boldsymbol{\ell}_*) + \overrightarrow{\nabla}_{\boldsymbol{\ell}} f_{\rm model}(\lambda|\boldsymbol{\ell}_*) \cdot \boldsymbol{\Delta\ell}. \tag{4.1}$$

In $N_{\boldsymbol{\ell}}$-dimensional label space, the calculation of $\overrightarrow{\nabla}_{\boldsymbol{\ell}}$ requires the calculation of (only) $N_{\boldsymbol{\ell}}$ additional model spectra, used to define the vector of "gradient spectra", $\overrightarrow{\nabla}_{\boldsymbol{\ell}} f_{\rm model}(\lambda|\boldsymbol{\ell}_*)$. In this study, we evaluate gradient spectra in finite differences. For each dimension, we derive the 1D gradient spectrum via $\mathrm{d}f_{\rm model}/\mathrm{d}\ell = (f(\ell_1) - f(\ell_2))/(\ell_1 - \ell_2)$. The assumption of linearity implies that gradient spectra are decoupled from one another, and the variation of a spectrum from a label point to another can be approximated by the sum of variation in each dimension. This (only) linear scaling of model numbers with $N_{\boldsymbol{\ell}}$ is one of the key advantages to CHAT. The region in label space for which this 1^{st}-order Taylor expansion is a sufficiently good approximation, we call a "Taylor-sphere." For labels $\boldsymbol{\ell}$ that lie within the Taylor-sphere of $\boldsymbol{\ell}_*$ spectral fitting then becomes a simple regression.

Ness et al. [153] show that even for the important labels such as [Fe/H] and $T_{\rm eff}$ that plausibly have small Taylor-radii, the spectral variation at all wavelength points across the entire label space for all

giants as a function of the labels can be approximated by a quadratic polynomial function. Therefore, we expect the Taylor-radii to cover finite, and not terribly small regions of the relevant label range. In principle, the high-dimensional linear expansion of the space of model spectra could be extended to a general polynomial expansion. Considering 2^{nd}-order expansion, this would require $\sim N_\ell$ times more ab-initio model calculations, but may dramatically increase the size of the corresponding "2^{nd}-order Taylor-sphere", as suggested by the empirical success of *the Cannon*[153], but we will explore in the coming paper.

A data-driven model grid

The next step in CHAT is to find a (near−)minimal set of grid points, $\boldsymbol{\ell}_*$, so that the ensemble of their surrounding Taylor-spheres covers all the relevant label space. If we had such a set of $\boldsymbol{\ell}_*$, and if that set was manageable small, then the entire fitting procedure would be reduced to a set of linear regressions.

To start this, we illustrate concretely how important it is to abandon rectilinear grids in high dimensions. The key point is that the volumes of hyperspheres of unit radius differ drastically from the volumes of hypercubes with unit length in high-dimensional space. Fig. 4.1 illustrates this volume ratio: for a 10-dimensional space, the volume ratio of the hypercube over the hypersphere is $\sim 5 \times 10^2$; and in a 30-dimensional space, the ratio increases to $\sim 5 \times 10^{13}$! Therefore, if we place our synthetic models (at $\boldsymbol{\ell}_*$), only in the (hyper)spherical region where they are needed (rather than in an encompassing hypercube), the density of models grows exponentially with the number of dimensions, for a given total number of models. Furthermore, Ting et al.[195] showed that, in a $\sim$ 25-

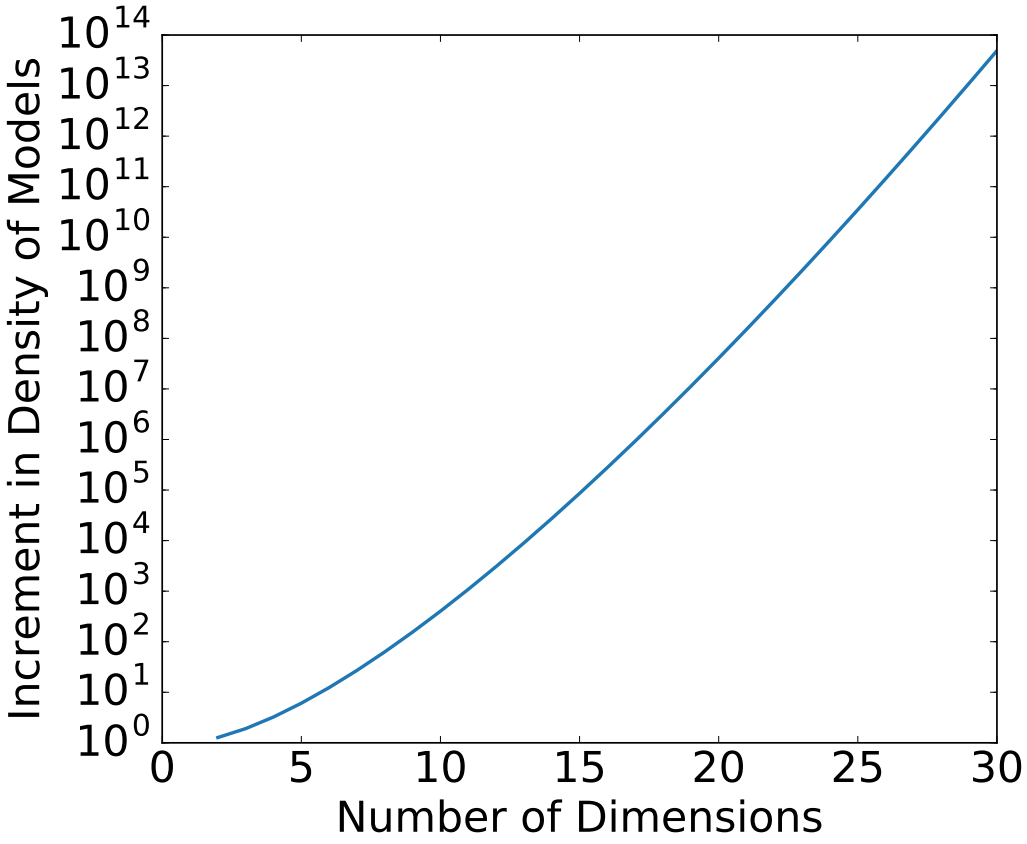

Figure 4.1: The volume ratio of an N-dimensional unit hypercube to a unit hypersphere. This figure shows that using a convex hull significantly improves the density of models in the label space. For example, if the stellar properties only lie in a unit hypersphere in a 10D label space, a rectilinear grid will need 500 times more models than a convex hull approach that only generates models within the hypersphere. If we consider a label space of 30D, the ratio increases to 5×10^{13}. In practice, since stellar elemental abundances only live in a $7 - 9$ dimensional subspace in a 30D space, which is more compact than a hypersphere, the improvement is more significant than the ratio shown in this figure.

dimensional elemental abundances space, stellar elemental abundances are contained in a $7-9$ dimensional subspace, the volume of which is necessarily smaller than the 25-dimensional hypersphere. Therefore, the gain will be much more significant in practice.

Clearly, we only need to calculate spectral models and their gradient spectra in regions of label space containing data. To define such a region, we use the concept of a convex hull. A convex hull is the minimal convex polygon that encompasses all the data points. Of course, for any given survey we do not know which part of label space the data cover. But there is sufficient information from existing surveys that one can define an approximate convex hull (e.g.,[13,92]). In detail, one would want to carefully consider the construction of the convex hull for the specific problem of interest.

After this region is defined, we need to find the set of $N_{\rm mod}$ grid points at $\boldsymbol{\ell}_*$ whose surrounding Taylor-spheres cover this minimal polygon. Our approach is illustrated in Fig. 4.2. The Figure illustrates how we define the convex hull, find Taylor-spheres that cover the convex hull and reduce the spectral fitting problem to a series of linear regressions. The details are described in §4.3.3, but the remarkable result is that the number of necessary grid points $N_{\rm mod}$ remains manageable even for a high-dimensional label space.

4.3.2 CHAT's advantages

CHAT has a number of very attractive properties compared to the rectilinear grid model that we summarize in Table 4.1. First, within the Taylor-sphere of a model grid point, $\boldsymbol{\ell}_*$, the computational expense of fitting only grows linearly with the number of labels to be fit. If $N_{\rm mod}$ model grid points are needed to cover the convex hull with their Taylor-spheres, then we only need to calculate $N_{\rm mod}\times$

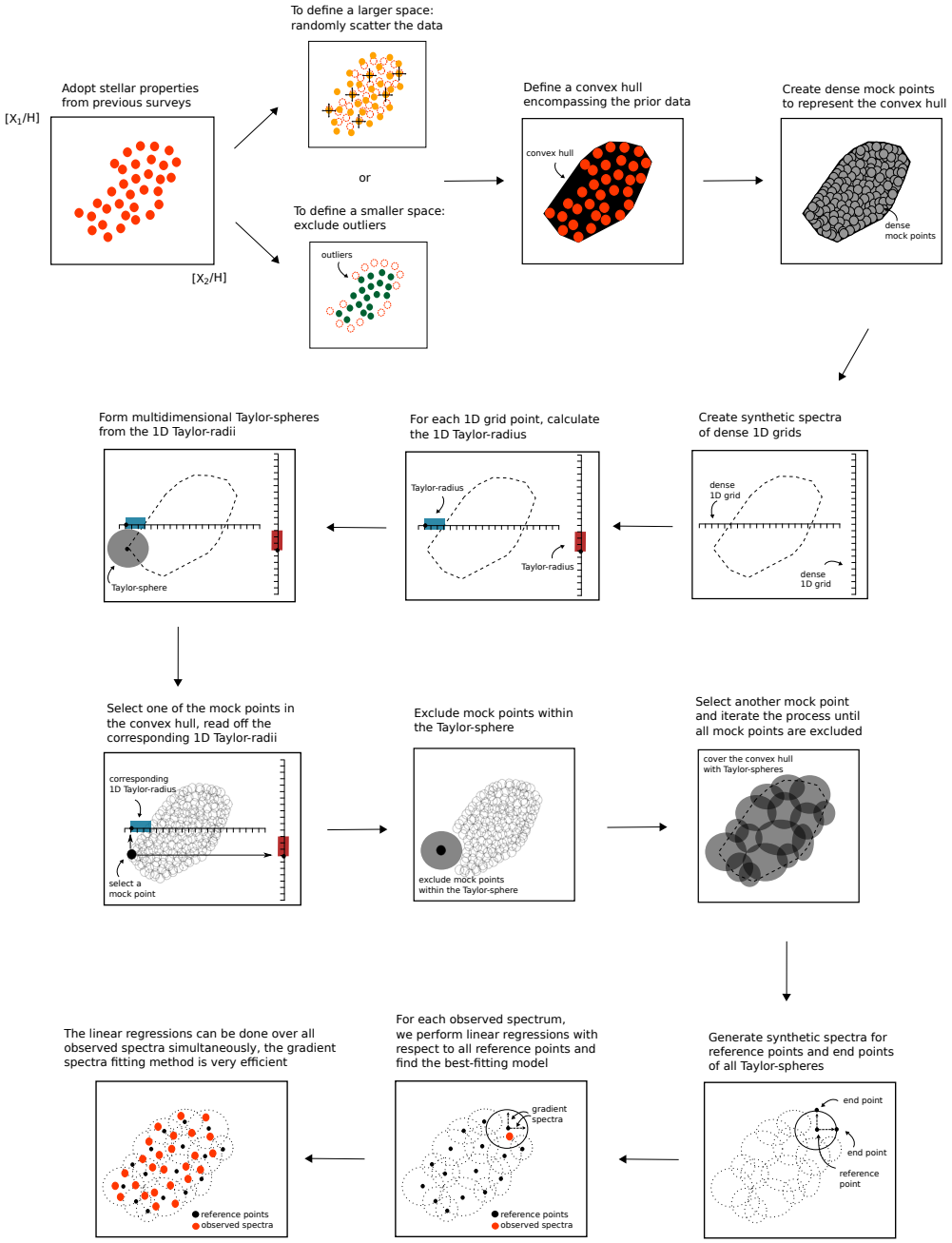

Figure 4.2: A schematic illustration of CHAT and its implementation. In this illustration, we demonstrate a 2D scenario. In practice, we generalize this approach to 10 — 30-dimensional label space using the same idea.

$(N_\ell + 1)$ synthetic models, i.e., for each model grid point, we need to calculate the spectrum at this grid point and additional N_ℓ models to calculate the array of gradient spectra. As N_ℓ increases by one, the size of the array of gradient spectra only increases by one, even though we might need slightly more models $N_{\rm mod}$. This slow growth is different from the rectilinear grid approach, where the number of models always grows as d^{N_ℓ}, where d is the number of grid points in each dimension regardless whether or not the spectrum varies linearly with a label.

Second, decomposing the label space into a series of Taylor-spheres reduces the complicated interpolation-minimization spectral fitting process to a series of linear regressions. This calculation can be easily done on any personal computer after the synthetic library is generated and can be easily parallelized. At the same time, we also circumvent the computational memory problem as discussed in §4.2. As we only need to perform linear regression from each grid point separately, only the gradient spectra of a particular grid point are loaded into the memory each time. The memory requirements for CHAT are very modest.

Aside from the computational speedup, the use of Taylor-spheres to determine where to create models has the important conceptual advantage that the step size in each dimension is determined in a statistically-rigorous manner (such that the error induced by assuming linear interpolation is below a predetermined tolerance).

Fitting models with linear regressions through gradient spectra also provides a natural connection to fully data-driven technniques. In this paper, we calculate the gradient spectra according to synthetic models. But one could replace the theoretical gradient spectra with empirical gradient spectra if we have enough training set to span the label space. As discussed in Ness et al.[153], performing

Table 4.1: Comparisons of the rectilinear grid approach and CHAT.

Properties	Rectilinear grid	CHAT
Defining label space	Results from previous (smaller) surveys are not required	Uses previous results to determine where models are needed
Number of models	Grows exponentially with the number of dimensions	Grows linearly with the number of dimensions in the limit where spectra vary linearly with respect to the additional dimension
Number of dimensions	Limited to $\lesssim$ 7D label space	Can extend to $>$ 10D label space and allows us to perform ab-initio full spectral fitting for multiple elements
Fitting procedure	Requires computationally intensive interpolation-minimization algorithms	Reduces the problem to a series of simple linear regressions. For an 8D space, we reduce the computational time by 100 folds when compared to a quadratic interpolation
Data-driven models	Not compatible with data-driven methods	Since the fitting relies on gradient spectra, the synthetic gradient spectra can potentially be substituted with data-driven empirical gradient spectra
Abundance precision	Limited recovery of the labels due to the sparse density of models. For an 8D space, we would need at least $\mathcal{O}(10^5)$ models	Better recovery because the density of models in the label space increases. We achieve the same precision with $\mathcal{O}(10^2)$ models

spectral fitting in a fully data-driven way is advantageous in some cases because such an approach produces elemental abundances on the same overall scale set by the training set.

4.3.3 Implementation

With the qualitative picture of Fig. 4.2 in mind, we now describe some of the specifics of finding suitably sized Taylor-spheres and filling the data convex hull with these Taylor-spheres.

Defining the Taylor-spheres

So far we have defined the Taylor-spheres qualitatively as the regions around a spectral model at $\boldsymbol{\ell}_*$ within which the vector of gradient spectra describes all model spectral "sufficiently" well. We now describe the quantitative procedure, which starts by determining the 1D Taylor-radii in all the label directions. We generate a fine 1D grid for each of these dimensions with a step size of $\Delta[X/\mathrm{H}] = 0.03$ dex, $\Delta T_{\mathrm{eff}} = 25$ K, $\Delta \log g = 0.05$, $\Delta v_{\mathrm{turb}} = 0.05$ km/s. When evaluating the Taylor-radius in any one label-space coordinate, we adopt fiducial labels for the other coordinates. Focusing here on APOGEE red clump stars, we choose $T_{\mathrm{eff}} = 4{,}750$ K, $\log g = 2.5$, $v_{\mathrm{turb}} = 2$ km/s and Solar metallicity. We checked that the Taylor-radii are relatively insensitive to the choice of the fiducial values for the other coordinates.

To evaluate the Taylor-radius for any one dimension of label space, *for each grid point* ℓ_*, we use the finely-spaced 1D grid to find the maximum distance from the grid point in label space such that all points, $\Delta\ell$ in between the grid point and the end point $\Delta\ell_{\mathrm{max}}$ can be well approximated by linearly interpolating between these two points; the quality of the approximation is judged by the

χ^2 difference between the interpolated spectrum and the directly calculated spectrum at that label point. We define this χ^2 to be

$$\begin{aligned}&\chi^2(\ell_*, \Delta\ell, \Delta\ell_{\rm max})\\ &\equiv \sum_\lambda \frac{\Big(f_{\rm interp}(\lambda \mid \ell_*, \Delta\ell, \Delta\ell_{\rm max}) - f_{\rm model}(\lambda \mid \ell_*, \Delta\ell)\Big)}{\sigma_\lambda^2},\end{aligned} \tag{4.2}$$

where we sum over all the wavelength points, each of which has a typical error of σ_λ in the actual data set. We deem a spectrum to be well enough interpolated if $\chi^2 < \epsilon$, where ϵ is the tolerance that we set. If we were to set ϵ to be the number of fitted labels, the uncertainties of the estimated labels due to interpolation errors would be comparable to uncertainties due to observation noise. Since we typically fit $\mathcal{O}(10)$ labels, we choose $\epsilon \simeq 50$ (assuming S/N=100) [‡] in our study. We also verified that the systematic uncertainties in elemental abundances from interpolation errors (of the order 0.01 dex, see plots in § 4.4.1) are indeed comparable to the one due to observation noise (see Appendix 4.7), justifying this choice. Finally, we define the Taylor-radius of any grid point ℓ_* as the maximum $\Delta\ell_{\rm max}$ for which $\chi^2(\ell_*, \Delta\ell, \Delta\ell_{\rm max}) \leq \epsilon$, for every grid point $\Delta\ell$ where $0 \leq \Delta\ell \leq \Delta\ell_{\rm max}$. We found that $T_{\rm eff}$ has a Taylor-radius of 175 K for $T_{\rm eff} = 4{,}000$ K and a Taylor-radius 350 K for $T_{\rm eff} = 5{,}000$ K; $\log g$ has a $\sim$ constant Taylor-radius of 0.9 dex for $\log g = 1 - 5$; [Fe/H] (an element that has a lot of absorption lines) has a Taylor-radius of 0.5 dex in the low metallicity regime [Fe/H] $= -1$, and the Taylor-radius decreases to 0.3 dex for solar metallicity; [K/H] (a trace

[‡] We could have chosen $\epsilon = 10$, but this requires more models in our test cases. Since the purpose of this paper is to compare CHAT with the rectilinear grid approach, we decided to reduce the computational time by adopting a slightly higher ϵ. As long as we adopt the same number of models for both cases, the comparison is fair.

element) has a Taylor-radius of 2 dex for $[\mathrm{K/H}] = -1$, and 1 dex for solar metallicity.

We have thus far defined the 1D Taylor-radii, which results in a Taylor-sphere encompassing all points $\boldsymbol{\ell}$ with

$$\sum_{n=1}^{N_{\boldsymbol{\ell}}} \left(\frac{\ell_n - \ell_{*,n}}{\Delta\ell_{\mathrm{max},n}} \right)^2 \leq 1, \tag{4.3}$$

There is no guarantee that all points within this Taylor-sphere satisfy the linearity conditions only tested along the coordinate axes. We have tested this issue in several ways. In 2D we verified this assumption by studying grids of random element pairs. For each element pair, we first found the 1D Taylor-radii set by a fixed χ^2 criterion. We found that 2D grid points within an ellipse defined by the 1D Taylor-radii fulfill the same criterion, and grid points outside the ellipse violate the criterion. The ellipse-approximation holds very well for all element pairs that are not strongly correlated in the spectra space. For strongly coupled elements, we performed a few tests with CNO and found that the χ^2 contours appear to be more irregular and do not form a perfect ellipse, even though the simple ellipse approximation still performs reasonably well. We will leave this issue to be explored in more detail in future work.

In Figure 4.3, we demonstrate how well the gradient spectra method works by comparing the gradient spectra reconstructions within a Taylor-sphere and the ab-initio model calculations. We calculate the Taylor-sphere centered at a fiducial reference point of $T_{\mathrm{eff}} = 4{,}750\,\mathrm{K}$, $\log g = 2.5$, $v_{\mathrm{turb}} = 2.0\,\mathrm{km/s}$ and Solar metallicity. We draw random labels within the Taylor-sphere and calculate their ab-initio model spectra. The gradient spectra constructions are done by taking the reference spectrum at the reference point and linearly adding the gradient spectra multiplied by the step

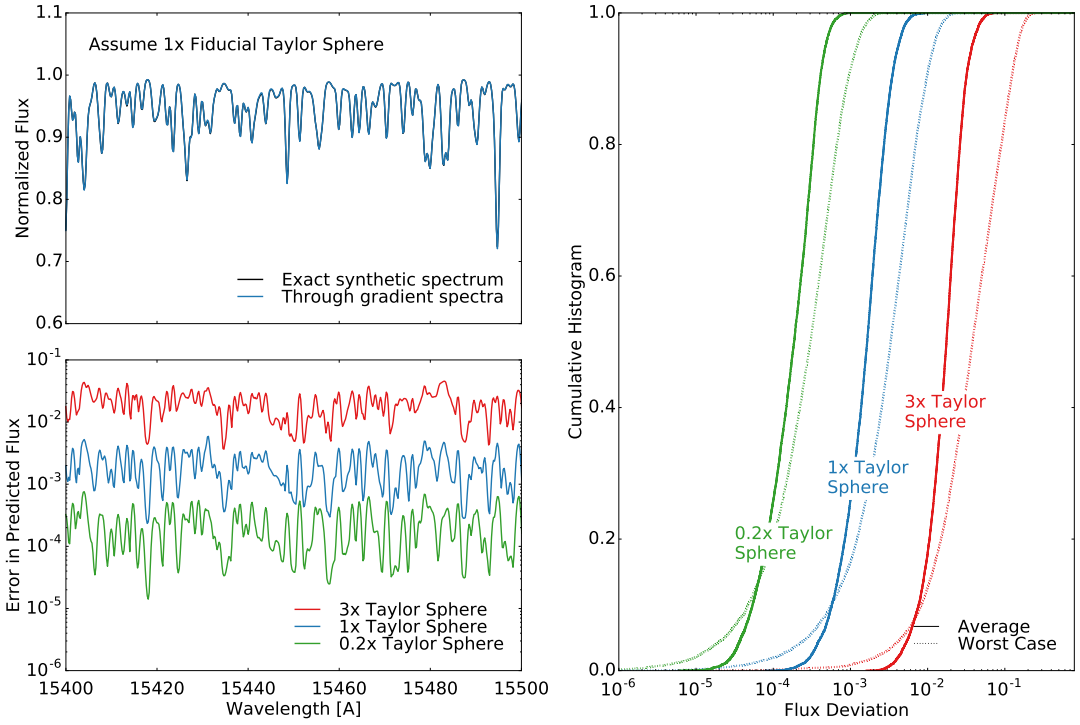

Figure 4.3: Demonstration of the fidelity of synthetic spectrum reconstruction through gradient spectra. We vary all 15 elements in APOGEE and the three stellar parameters, $T_{\rm eff}$, $\log g$ and $v_{\rm turb}$. We consider a χ^2 tolerance $\epsilon = 50$, the tolerance we assume in this study, and evaluate the Taylor-radii and their corresponding Taylor-sphere. The Taylor-radii are evaluated at the fiducial reference point of $T_{\rm eff} = 4{,}750$ K, $\log g = 2.5$, $v_{\rm turb} = 2.0$ km/s and Solar metallicity. To illustrate the behavior as a function of Taylor-radii, we shrink the radii 5 times and expand 3 times which we denote as $0.2\times$ Taylor-sphere and $3\times$ Taylor-sphere, respectively. We generate 100 mock spectra for labels that are within each of these Taylor-spheres and compare the gradient spectra reconstructions to the ab-initio calculations. The top left panel shows the comparison of a segment averaging all 100 trials. The comparison is excellent, justifying our choice of χ^2 tolerance. The bottom left panel shows the mean absolute differences between the exact and gradient-interpolated models assuming three different Taylor-spheres. In the right panel we plot the cumulative histogram of the wavelength-by-wavelength deviations. The solid lines demonstrates the absolute deviations averaged over all 100 trials, and the dashed lines show the worst-case scenarios of the 100 trials. For the fiducial Taylor-sphere, the wavelength-by-wavelength deviation is well below the typical APOGEE S/N of 100.

size in each label dimension. We vary all 15 elements in APOGEE and the three main stellar parameters, $T_{\rm eff}$, $\log g$ and $v_{\rm turb}$. We show that for a tolerance of $\epsilon = 50$, the wavelength-by-wavelength deviation is comfortably below the typical APOGEE S/N of 100. As we will demonstrate in §4.4, we only need a few of these Taylor-spheres to span the relevant label space, which implies that through CHAT, we can reconstruct synthetic spectra with a near-minimal number of synthetic spectra calculations.

Filling the convex hull of label space

We now discuss how to find the set of model grid points, $\boldsymbol{\ell}_*$, so that their surrounding Taylor-spheres fill the convex hull of the pertinent portion of label space. For a high dimensional space, data points that use to determine the convex hull (the minimum polygon) might only cover a small fraction of the volume. In order to make sure that we are covering the convex hull sufficiently well with Taylor-spheres, we start by representing the convex hull by 10^6 uniform mock data points. As we will end up with $<$ 1,000 model grid points (see §4.4), 10^6 mock points will sample the convex hull sufficiently well for our purposes. We construct the convex hull using CONVEXHULL routine in the SCIPY.SPATIAL package. To determine whether a point is within the convex hull, we use DELAUNAY routine and check whether the point is within a simplex of the Delaunay tessellation.

Following Fig. 4.2 we then successively identify model grid points, $\boldsymbol{\ell}_*$. We start by picking a random point from the mock data within the convex hull. We read off their 1D Taylor-radii corresponding to the closest grid points in the finely-spaced 1D grids. Then we ask which of our mock points fall within the Taylor-sphere determined by these 1D Taylor-radii, and eliminate those. Then

we draw another random point from the remaining mock points. We take it to be the next model grid point, consider its surrounding Taylor-sphere, and remove the mock points that lie within that sphere. We emphasize that the Taylor-radii read off from the 1D grids for each random point are difference – we assume ℓ_* that is closest to the random point when evaluating the Taylor-radii. This "adaptive" approach takes into account that the radii could vary, for example, in a low $T_{\rm eff}$ regime compared to a high $T_{\rm eff}$ regime. We repeat this procedure until no mock points remain. The resulting set of $\boldsymbol{\ell}_*$ is the near-minimal set of model grid points, whose Taylor-spheres fill the convex hull.

We refer to this set of grid points as a *near*-minimal set since there are overlaps between Taylor-spheres, so it might not be the absolute minimal set. However, the overlaps of hyperspheres are quite small in a higher dimensional space. For an 8D label space that we will explore in § 4.4.1, through Monte Carlo integration we found that the total volume of all the ellipsoids within the convex hull is only $\sim$ 5 times the volume of the convex hull. So at most, we can only reduce the number of grid points by another factor of five.

This procedure requires no calculation of any model spectra apart from the precalculated spectra of the fine 1D grids in determining the Taylor-radii. After we have the set of $\boldsymbol{\ell}_*$, we calculate the model spectra at each of them, along with the $N_{\boldsymbol{\ell}}$ gradient spectra at each $\boldsymbol{\ell}_*$. These are all the models needed for CHAT.

There are also a number of practical choices to be made to define the convex hull in the first place. If the prior information used to determine the convex hull is very noisy or has outliers, the volume of the convex hull will be larger than the intrinsic volume within which the labels of the sample at hand reside (e.g., [192]). With many dimensions, the volume of the noisy convex hull could be much

larger than the intrinsic volume, "wasting" many model grid points at the periphery of the volume. To address this problem, one could run a kernel density estimation to map out the density of the label space, similar to Ting et al.[192] and cull outliers according to the density map. If, however, a goal of the spectral fitting is to look for rare outliers in label space, one might consider randomly scattering the prior data so that they define a larger volume in label space (see Fig. 4.2). A point to keep in mind is that the particular implementation of CHAT will depend on the problem of interest.

4.3.4 Creating stellar models

The model atmospheres and spectra for this work are computed with the ATLAS12 and SYNTHE programs written and maintained by R. Kurucz[118,119,123]. We adopt the latest line lists provided by R. Kurucz,[§] including line lists for TiO and H_2O, amongst many other molecules. Model atmospheres are computed at 80 zones down to a Rosseland optical depth of 10^3, and each model is automatically inspected for numerical convergence. We adopt the Asplund et al.[7] solar abundance scale. Convection is modeled according to the standard mixing length theory with a mixing length of 1.25 and no overshooting. Spectra are computed with the SYNTHE program and are sampled at a resolution of $R = 300{,}000$ and then convolved to lower resolutions.

4.3.5 Spectral fitting

After calculating all N_{mod} spectra at the model grid points ℓ_*, along with models needed to evaluate their gradient spectra, we have reduced the spectral fitting of one object to a set of N_{mod} linear regres-

[§] http://kurucz.harvard.edu

sions. In practice it may be more advantageous to go to one model grid point, $\boldsymbol{\ell}_*$ and execute a trial linear regression fit on all spectra of the sample. In the end the best fitting labels for each object will be the fit (among the $N_{\rm mod}$) that has the lowest χ^2.

Let us now denote the entire (observed) spectrum of the n^{th} ($n \in [1, N_{\rm sample}]$) object as the vector $\boldsymbol{f}_{{\rm obs},n}$ of length $N_{\rm pix}$, with the vector of its uncertainties denoted as $\boldsymbol{e}_n$, where $N_{\rm pix}$ is the number of wavelength points in the spectrum. The entire model spectrum at a grid point $\boldsymbol{\ell}_*$ can be denoted as the equally long vector $\boldsymbol{f}_{\rm mod}(\boldsymbol{\ell}_*)$. The set of $N_{\boldsymbol{\ell}_*}$ gradient spectra associated with $\boldsymbol{f}_{\rm mod}(\boldsymbol{\ell}_*)$ can be denoted as a $N_{\rm pix} \times N_{\boldsymbol{\ell}_*}$ matrix $\boldsymbol{GL}(\boldsymbol{\ell}_*)$. For linear regression, this defines a covariance matrix $\boldsymbol{\Omega}_n(\boldsymbol{\ell}_*)$:

$$\boldsymbol{\Omega}_n(\boldsymbol{\ell}_*) = \boldsymbol{GL}^{\rm T}(\boldsymbol{\ell}_*) \cdot \left(\boldsymbol{GL}^{\rm T}(\boldsymbol{\ell}_*) \times \boldsymbol{e}_n^{-2}\right)^{\rm T}. \tag{4.4}$$

Then the best fitting labels $\boldsymbol{\ell_n}$, based on the model grid point $\boldsymbol{\ell}_*$ are given by

$$\begin{aligned} &\boldsymbol{\ell_n}(\boldsymbol{\ell}_* \mid \{\text{data}\}_n) \\ &= \boldsymbol{\Omega}_n^{-1}(\boldsymbol{\ell}_*) \cdot \left(\boldsymbol{GL}^{\rm T}(\boldsymbol{\ell}_*) \cdot \left((\boldsymbol{f}_{{\rm obs},n} - \boldsymbol{f}_{\rm mod}(\boldsymbol{\ell}_*)) \times \boldsymbol{e}_n^{-2}\right)\right), \end{aligned} \tag{4.5}$$

where $\text{data}_n \equiv [\boldsymbol{f}_{{\rm obs},n}, \boldsymbol{e}_n]$.

This linear regression fitting process is extremely efficient even though it is performed $N_{\rm mod}$ times because the regression can be done analytically and consumes little memory even for multiple variables. Also the main computational cost in spectral fitting lies in generating a model grid instead of

the fitting process, and for the former, the number of models in CHAT grows much more benignly with more dimensions than an exponential growth. Furthermore, linear regressions around different $\boldsymbol{\ell}_*$ can also be parallelized to speed up the process. Finally, in this regime we have the covariance matrix $\boldsymbol{\Omega}_n$ analytically. This covariance matrix, reflecting the label space error ellipsoid, is a critical component in chemical tagging studies (see[192]).

4.4 Results and implications

We are now in a position to explore what CHAT can deliver in practice. We will cover three related aspects: in §4.4.1 we explore how well CHAT (with $N_{\rm mod}$ grid points, and hence a total of $N_{\rm mod} \times (N_\ell + 1)$ models) does in determining large numbers of labels from data, compared to a sensibly chosen, but rectilinear high-dimensional grid that has a total of $\sim N_{\rm mod} \times (N_\ell + 1)$ grid points. We also explore what systematic errors can occur when fitting a higher dimensional spectra with a lower dimensional effective label space (§4.4.2). In §4.4.3 we explore how to reduce the effective dimensionality of label space, exploiting astrophysical correlations among elemental abundances through a principal component analysis (PCA). Finally, we discuss some limitations of CHAT in §4.4.4.

Throughout these tests we use mock spectral data drawn from synthetic spectral models that are based on up to 18 labels: three stellar parameters $T_{\rm eff}$, $\log g$, $v_{\rm turb}$ and up to 15 APOGEE elemental abundances; the remaining elements, which we will call "trace elements", are assumed to scale with [Fe/H] at solar abundance ratios. In other words, we assume labels (from the APOGEE DR12 red clump stars), create ATLAS12 models with these labels and use these models as our mock data to be fit. That means that there always exists a set of labels that leaves no systematic differences between the data and the model. We defer the application of CHAT to real data to another paper.

4.4.1 Rectilinear grid fitting *vs.* CHAT

In this section we compare how well we recover the input labels from mock spectra using two fitting approaches, a rectilinear grid approach and CHAT using the same number of models. For this test case, we consider an 8D label space: three main stellar parameters – $T_{\rm eff}$, $\log g$, $v_{\rm turb}$ and five main elements, which we will call "primary elements." We chose these elements as they either have a significant influence on the overall atmospheric structure ([Fe/H], [Mg/H], [Si/H],[¶] see Appendix 4.6) for red clump stars (i.e., $T_{\rm eff} = 4{,}500 - 5{,}000$ K), or because they have important molecular features in the H-band APOGEE spectra ([C/H], [N/H]). We define 10 additional elements that are derived in APOGEE DR12 to be "secondary elements" and assume solar metallicity for all other "trace elements." Our adopted element classification nomenclature is summarized in Table 4.2. We sample the 8D label space for the mock spectra by simply adopting as input the labels from the APOGEE red clump sample[33] and $v_{\rm turb} = 2.478 - 0.325 \log g$[92]. We make the same selection cut as Hayden et al.[86], Ting et al.[192] to cull APOGEE values that are not reliably determined. When generating testing data, we add photon noise corresponding to the median wavelength-dependent S/N of the APOGEE red clump spectra.

We assume that the rectilinear grid approach spans the full range of label values in the APOGEE red clump sample, and we take three grid points in each of the eight dimensions, leading to $\sim$ 6,500 model grid points. To interpolate this rectilinear grid we consider both a linear interpolation and

[¶] Another important element is O, but creating a 9D grid is too computationally expensive for the rectilinear grid approach. However, we checked that replacing Si with O does not alter the conclusions in this paper.

Table 4.2: Nomenclature of element classification in this paper.

Nomenclature	Element	Classification
Primary elements	Fe, C, N, Mg, Si	Used in atmosphere calculations[a]
Secondary elements	O, Na, Al, S, K, Ca, Ti, V, Mn, Ni	Held fixed at solar metallicity or trace primary elements when creating model atmospheres
Trace elements	Elements other than the 15 elements in APOGEE DR12 and should have negligible effect on the atmosphere	

[a]These elements have important influences to the atmosphere. See Appendix 4.6.

a quadratic Bézier interpolation[145]. The interpolation codes – FERRE[‖3,4] are adopted from the APOGEE pipeline and provide a direct comparison to the state-of-the-art rectilinear grid approach. FERRE performs a wavelength-by-wavelength interpolation of the flux and finds the best fitting spectrum through χ^2-minimization. For CHAT, we only generate models that are within the convex hull of the APOGEE red clump sample. As the stellar parameters $T_{\rm eff}$, $\log g$ and $v_{\rm turb}$ reflects the evolutionary state of the star, while elemental abundances reflect the chemical evolution of the Galaxy when the star formed, they should be uncorrelated to first approximation (but see[139,143]). We consider the convex hull of these two groups separately and cross-product their model grid points and gradient spectra. We checked that if we were to consider the 8D convex hull directly, we would further reduce the number of models needed as there are non-trivial correlations between these two groups. We choose to consider these two groups separately to have a more direct comparison with the PCA method in §4.4.3. We adjust the size of the Taylor-spheres, by tweaking the $\epsilon \simeq 50$ tolerance in the χ^2 criterion in Eq. 4.2, such that the convex hull is filled by Taylor-spheres formed from $\sim$ 6,500 models (including the model grid points and their associated gradient spectra).

We then determine the best fitting 8D label set for all mock spectra with both methods. Fig. 4.4 shows the comparison of the results. Even in this relatively low-dimensional, 8D label space, CHAT provides better — higher accuracy and precision — label recovery than the rectilinear grid with equally many spectral models. The results also imply that, to achieve the same nominal precision, CHAT will require far fewer models than the rectilinear grid approach.

Part of the explanation for this gain is simple: using Monte Carlo integration, we find that the

[‖] http://www.as.utexas.edu/~hebe/ferre/

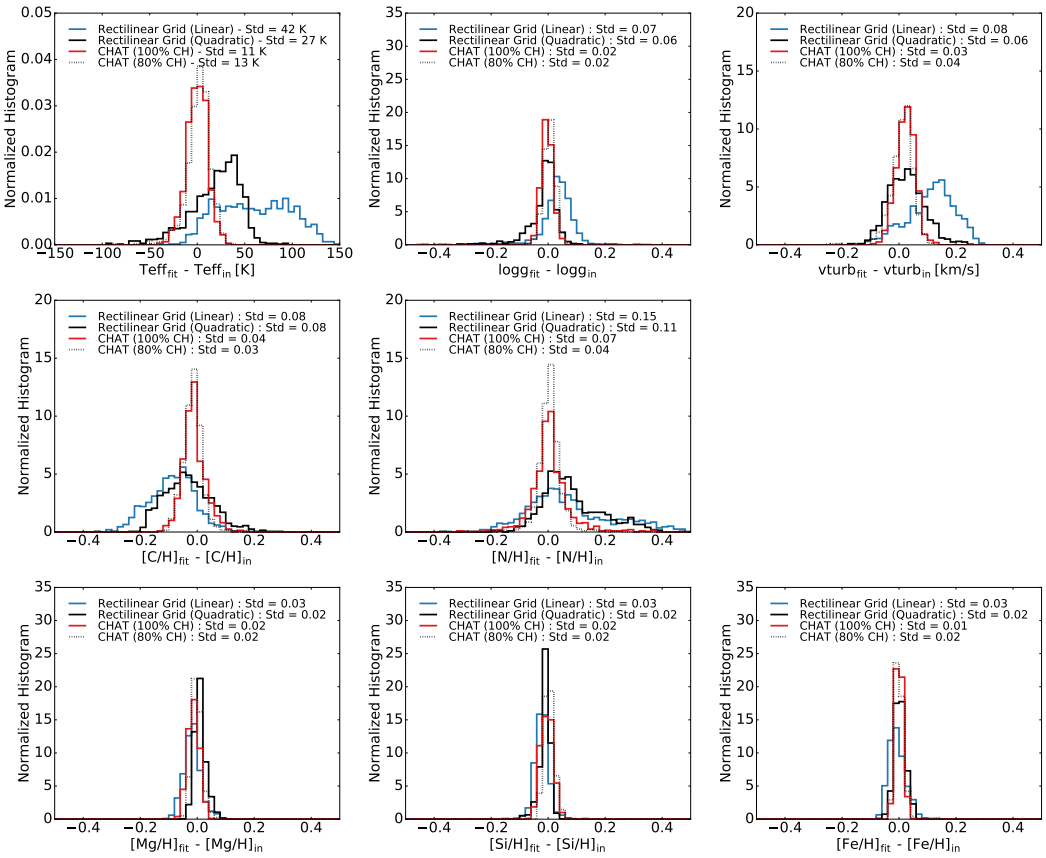

Figure 4.4: Recovery of labels for 1,000 synthetic test models with APOGEE red clump properties. We consider 8D test models, assuming solar abundances for the other elements, and fit the test models with 8D synthetic libraries. The blue lines show the results of a linearly interpolated rectilinear grid with $\sim$ 6,500 grid points. The solid black lines show the results from the same rectilinear grid but the grid is now quadratic Bézier interpolated. To enable a fair comparison with CHAT, the rectilinear grid only spans the range of label values used in this test. The red lines show the results of the standard approach of CHAT with the same number of models as in the rectilinear grid. Labels are recovered very well, in all cases more precisely than when using the rectilinear grid. But CHAT can be further improved by excluding outlying points before defining the convex hull. If we exclude the 20% most outlying points, as shown in the dotted black lines, we can further reduce the number of models needed by a factor of $\sim$ 35 and achieve the same precision. In this case, only 180 models are needed in an APOGEE red clump synthetic library.

volume of the APOGEE red clump 8D convex hull is $\sim$ 100 times smaller than the hypercube spanned by the rectilinear grid. Therefore, we are sampling the label space $\sim$ 100 times denser in CHAT. The rectilinear grid approach would require $\sim$ 650,000 models in order to produce the same recoveries with linear interpolation. For example, if we were to adopt the standard APOGEE spacing in a rectilinear grid, we would need at least $5^8 = 4 \times 10^5$ models. This is consistent with the conventional wisdom that for rectilinear grids it is difficult to go beyond 6D-7D label space. Quadratic/cubic interpolation slightly reduces the number of models needed to achieve the same precision, but the computational efficiency is compromised. In this study, we did not consider a cubic interpolation because a cubic interpolation in the rectilinear grid approach requires at least four grid points per dimension; it requires a minimum of $4^8 = 65{,}000$ models, which is again significantly more models than CHAT. In our test case, using a single CPU and the same number of models, we found that it takes about $\mathcal{O}(10)$ minutes for CHAT to find the best fitting labels for 1,000 spectra, but it takes $\sim \mathcal{O}(1{,}000)$ minutes for a quadratic Bézier interpolation, showing the enormous gain of reducing a complicated interpolation-minimization process to a series of simple linear regressions.

Beyond the label space volume difference, the systematic filling of this (smaller) convex hull with adaptive tessellation results in additional gains, i.e., we put more models in regions where linear interpolations fail. Although not shown, to evaluate the contributions from each of these two aspects, we performed tests by only considering convex hull without adaptive tessellation and an adaptive tessellation in a regular label space without the convex hull. We find that the convex hull plays a more important role because it improves the models density in the label space globally. The adaptive

tessellation plays a smaller role, but it is important for $T_{\rm eff}$ because $T_{\rm eff}$ shows the most nonlinearity in model variation. Low $T_{\rm eff}$ regimes need more models than the high $T_{\rm eff}$ regimes because the synthetic spectra vary more drastically when molecules start to form at low $T_{\rm eff}$.

Remarkably, CHAT can be even more efficient. As discussed in §4.3.3, the periphery of any 8D space comprises most of the volume. And, as illustrated in the right panel at the second last row of Fig. 4.2, covering the convex hull with Taylor-spheres centered at the periphery points covers a region larger than the convex hull itself. Through Monte Carlo integration, we found that only $\sim$ 10% of the total volume of the Taylor-spheres is within the convex hull, the rest of $\sim$ 90% surrounds the convex hull. The outermost points could have already been covered by the Taylor-spheres around points in the interior. Therefore, we eliminate the 20% most outlying points using kernel density estimation before constructing the convex hull around the other 80%. We choose an 80% convex hull because we found that the Taylor-spheres centered at points in this smaller convex hull already cover the full convex hull.

The black dotted lines in Fig. 4.4 demonstrate CHAT's label recovery in this case: we need only 20 model grid points $\boldsymbol{\ell}_*$ and their associated gradient spectra, i.e., a total of 180 spectra (20 model grid points, and 20 $\times$ 8 models to determine the gradient spectra), to fulfill the same $\epsilon - \chi^2$ criterion as before. The resulting label recovery illustrated in the black dotted lines shows that we can achieve the same precision as the standard implementation of CHAT but with 35 times fewer models. The upshot is that by culling 20% of the outlying points when defining the convex hull, we are able to reduce the number of the required synthetic spectrum calculations to fit red clump stars in 8D label space from 6×10^5 in a rectilinear grid approach to only 180 model spectra.

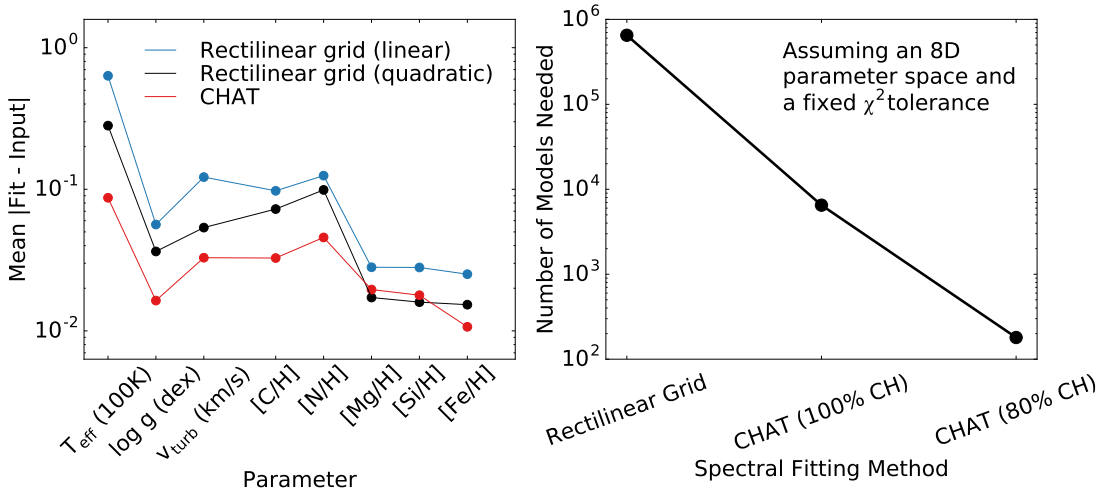

Figure 4.5: A summary of Fig. 4.4 and a comparison of CHAT with the rectilinear grid approach. The left panel shows the mean deviation of the synthetic spectra input recoveries in an 8D label space. In all three cases, we consider $\sim$ 6,500 models. A quadratic interpolation improves the recoveries from a rectilinear grid, but it is still less precise than CHAT and is about 10^2 times slower than CHAT. The right panel shows the estimated number of models needed to have the same recovery precision as CHAT. To have the same density of models in the label space, the rectilinear grid approach requires two orders of magnitude more models, i.e., 6.5×10^5 models. CHAT can be further improved by discarding the 20% outlying data when determining the convex hull. In this case, we can further reduce the number of models by at least another order of magnitude. Only 180 models are needed for the improved CHAT.

We summarize the comparison between the rectilinear method and CHAT in Fig. 4.5 for the case of 8D fitting. Note that, for a rectilinear grid, even if we consider two grid points per dimension, we will require $2^8 = 256$ models. The gradient fitting approach here surpasses the fundamental limit of a rectilinear grid because in the limit where the spectra vary linearly with all labels and are decoupled from one another, the number of models needed in CHAT grows linearly instead of exponentially.

For on-going large spectroscopic surveys, full spectral fitting is limited to a subset of "main" labels. For example, in APOGEE DR12[71,92], a 6D label space of $T_{\rm eff}$, $\log g$, $[Z/{\rm H}]$, $[\alpha/Z]$, $[{\rm C}/Z]$, $[{\rm N}/Z]$ was considered. Even a 6D space with five grid points requires $5^6 \sim$ 15,000 models. To make the computational consumption more affordable, some important fitting labels were not included, such as $v_{\rm turb}$[**] and $v \sin i$. But stellar rotation could play an important role for low-$T_{\rm eff}$ dwarf stars. Omitting them is believed to be the main reason that labels for cool dwarfs were not robust in DR12[92]. We show that even for an 8D space, using CHAT reduces the number of models by a factor of $\sim$ 1,000. The reduction of models opens up the opportunity to expand many more dimensions and allows $v \sin i$ and $v_{\rm turb}$ to be included more easily.

4.4.2 Consequences of fitting a subset of the label space

To fully specify an observed spectrum within its (high-S/N) error bars, one may require the specification of several dozens of labels, encompassing the stellar parameters and all elements that could contribute to the spectrum. However, the rectilinear grid fitting approach is limited to subspaces of

[**] In DR12, APOGEE found a tight $\log g$–$v_{\rm turb}$ relation. Thus, this relation was assumed to reduce the number of models needed in the synthetic library.

much lower dimension. In that case, assumptions have to be made for not-fitted labels. To explore this effect, we create a set of synthetic test spectra that are specified by the 18 labels measured in the APOGEE red clump sample: 15 elements, along with $T_{\rm eff}$, $\log g$ and $v_{\rm turb}$. We assume that all remaining trace elements follow [Fe/H] at solar ratios.

To start, we try to match these mock spectra with CHAT, but fitting only the 8 of the 18 labels, those shown in §4.4.1. Fig. 4.6 (blue lines) illustrates that the label recovery is unsatisfactory. This may not be unexpected, as we only vary five elements in the 8D grid, we necessarily mismatch absorption lines from the other elements, held at Solar ratios. As a consequence, the fit of $T_{\rm eff}$ is compromised, which in turn affects the fit of the other labels. One approach to reducing these systematics, while keeping the number of fitted labels low is to exploit the established astrophysical covariances among elemental abundances: we can, for e.g., assume that all α-elements trace each other, while all other elements scale with the global metallicity, $[Z/{\rm H}]$, an approach followed in the APOGEE pipeline. If we then generate a 7D grid of model spectra, $T_{\rm eff}$, $\log g$, $v_{\rm turb}$, $[Z/{\rm H}]$, $[\alpha/Z]$, $[{\rm C}/Z]$ and $[{\rm N}/Z]$ using CHAT, and with the same number of models, and use them to fit the 18D mock spectra. As shown in red lines in Figure 4.6 the systematic errors in the label recovery are strongly reduced. This demonstrates that label recoveries can be good with low-dimensional (e.g., 7D) fitting, if astrophysical label-correlations are properly exploited.

But even in this latter case systematic offsets in the label recovery remain: about 10 K for $T_{\rm eff}$, 0.1 dex for $\log g$ and 0.1 km/s for $v_{\rm turb}$. Even though the α elements broadly trace each other, this is not true in detail. This suggests the need for a more systematic way to reduce the dimensionality of the label space, an issue which we will address in the next section.

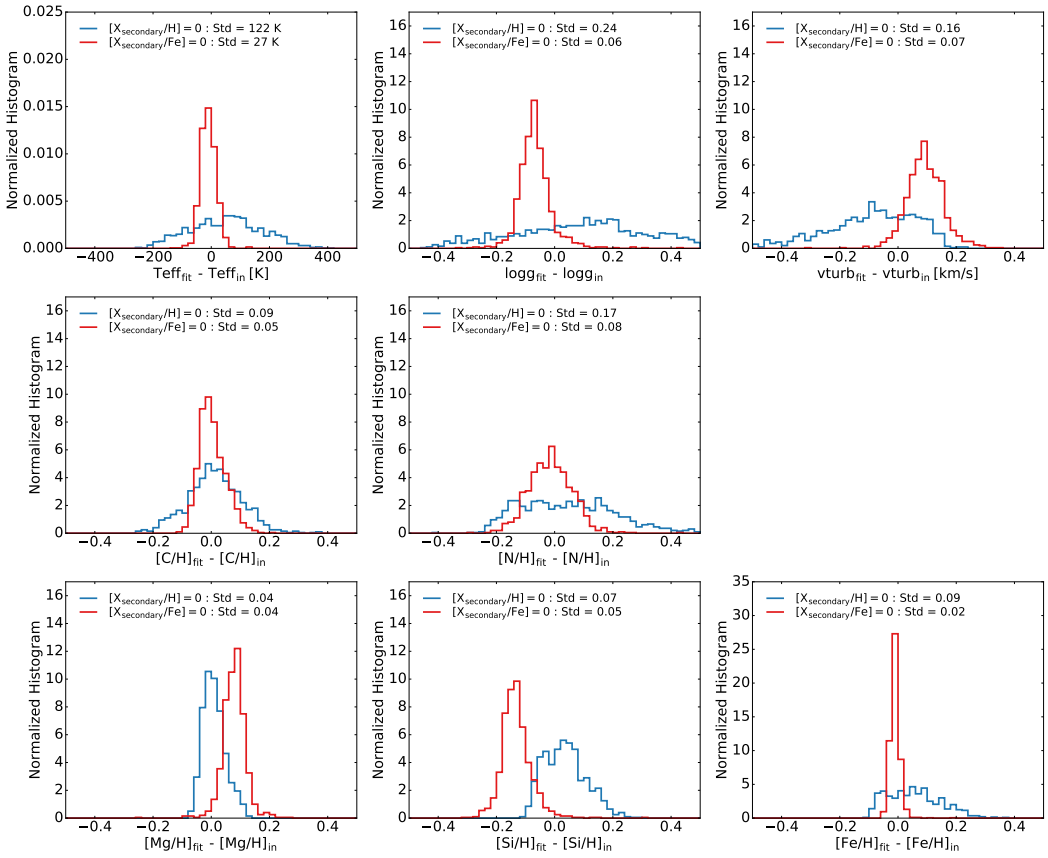

Figure 4.6: Similar to Fig. 4.4, but here we consider 1,000 synthetic test models that vary all 15 elements, $T_{\rm eff}$, $\log g$ and $v_{\rm turb}$, whose labels are drawn from 1,000 APOGEE red clump stars. We fit these test models with lower-dimensional synthetic libraries. In the first case, shown in the blue lines, we consider the $\sim$ 6,500 models 8D CHAT library in Fig. 4.4, fixing the other elements to be solar metallicity. In this case, the fit is unsatisfactory, showing that some assumptions have to be made to approximate high-dimensional spectra with lower-dimensional synthetic libraries. In the second case, shown in the red lines, we assume that the α-elements trace each other and the other elements trace [Fe/H]. We generate the same number of models as the previous case. The fits improve significantly as we take into account the other elements beside the fitted labels. However, these assumptions on the other elements are not true in detail, and hence systematic offsets remain in some of the fits.

4.4.3 Fitting PCA components of elemental abundances

Principle Component Analysis (PCA) provides a simple and systematic way to characterize the astrophysical correlations among elemental abundances. As such, it may be an effective way to set element ratios that are not directly fitted labels. A detailed exposition of PCA in elemental abundances space is beyond the scope of this paper, but details can be found in Andrews et al. [6] and Ting et al. [195].[††] The main points of Ting et al. [195] can be summarized as follows. PCA measures the correlation among elements. Each principal component is a unit vector in the elemental abundances space. The principal components are ordered according to their contributions to the total variances of the data sample. In practice, only the first few principal components are significant and relevant, because additional PCA components are likely dominated by observational noise. Ting et al. [195] showed that for $15 - 30$ measured elements, not all elements provide independent pieces of information. Elements fall into groups that span a much smaller $7 - 9$ dimensional subspace. In turn, measuring the $7 - 9$ dimensional principal components should be sufficient to predict the abundances of all $15 - 30$ elements.

With this idea in mind, a more effective way to fit high-dimensional spectra with a lower-dimensional effective subspace is to consider the coefficient of each main principal component as a fitting label instead of the usual $[X/\mathrm{H}]$. In this case, we should be able to fully characterize an observed spectra with < 13 labels (including the stellar parameters, T_{eff}, $\log g$, v_{turb} and $v \sin i$), but still account for 30 elements. A synthetic library covering 13-dimensional label space is feasible, but only with the

[††]Note that the APOGEE pipeline uses PCA to compactify the spectral space. In our case, we use PCA to compactify the label space of elemental abundances.

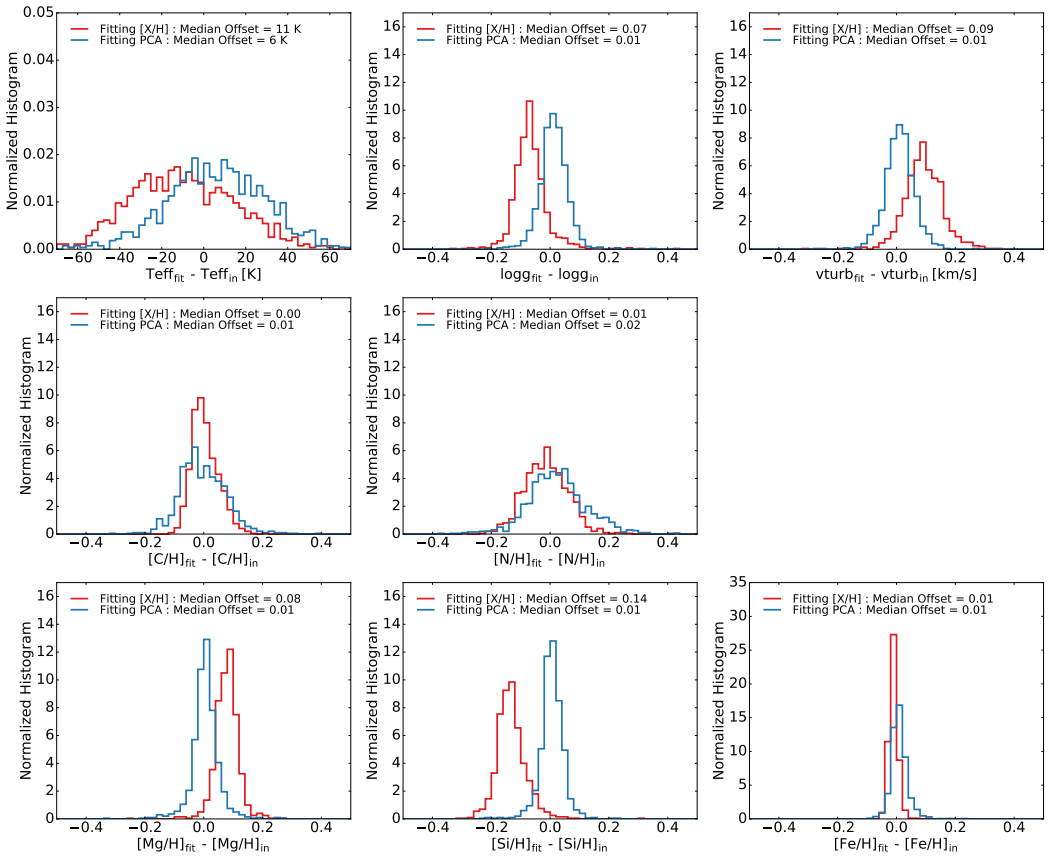

Figure 4.7: An effective way to reduce systematic offsets is to fit the coefficients of PCA components in elemental abundances space instead of fitting $[X/\mathrm{H}]$. The red lines show the same results in Fig. 4.6. We assume that the α-elements trace each other and the other elements trace $[\mathrm{Fe}/\mathrm{H}]$. Since this assumption is not true in detail, fitting the global metallicity $[Z/\mathrm{H}]$ and the α-enhancement leaves some systematic offsets. If we fit the coefficients of principal components instead, as shown in the blue lines, we reduce the systematic offsets because we take into account the correlations of elements more properly.

advantages provided by CHAT (see §4.4.1).

CHAT easily generalizes to the case of fitting PCA components, which are simply taken as the labels in lieu of the direct abundances. In other words, the model grid points and Taylor-spheres are simply determined in the space of stellar parameters and PCA-components. Specifically, We consider an 8D label space comprising $T_{\rm eff}$, $\log g$, $v_{\rm turb}$ and the five most important principal components. As shown in Figure 4.7, standard deviations of the PCA label recovery (transformed into the space of abundances) are about the same as that of the element label recovery because the underlying model density in the label space does not improve by transforming into the PCA space. However, systematic offsets in the label recovery are dramatically lower when fitting principal components, because we take into account the element correlations more properly.

4.4.4 Current limitations and future directions

Despite its attractive properties, CHAT also has limitations. CHAT fundamentally relies on a sensible definition of a convex hull in label space. This should be straightforward for the bulk of Milky Way stars, because we have a basic understanding regarding elemental abundance distributions from previous surveys (e.g., [13,92]). But the construction of a convex hull may be more problematic in other circumstances, such as in the search for extremely metal-poor stars of the Milky Way, or for surveys of other galaxies. In these cases the rectilinear approach or applying CHAT to a large rectangular label space with a lenient tolerance might be applied as a first pass in order to aid in the definition of the convex hull in detail. Because of the use of a convex hull in deciding where to create models, CHAT may not work well as a tool in searching for interesting outliers. In its present form, CHAT

is strictly linear within the Taylor-radius of each label, i.e., it treats the gradient spectrum in each label as independent. We have tested in a few cases that this linear approximation is indeed good within the entire Taylor-sphere, but we have not explored this exhaustively. One way to overcome this limitation is to expand CHAT beyond the 1^{st}-order Taylor-sphere. A 2^{nd}-order Taylor-sphere requires N_ℓ times more models to define both the gradient matrix and the curvature matrix. The fitting process will be slower as the fitting is no longer a simple linear regression, but it is reasonable to assume that a 2^{nd}-order model would cover a much larger label space per Taylor-sphere (see[153]). We are currently exploring this idea and will defer the details to a later paper.

4.5 Summary and conclusion

Major ongoing and planned research initiatives to unravel the chemodynamical formation history of our Milky Way are based on determining the properties of vast numbers of individual stars. This is to be done by taking high-quality spectra of $10^5 - 10^6$ stars in the Milky Way, and then deriving from them extensive sets of labels, i.e., stellar parameters and $10 - 30$ elemental abundances. Rigorous spectral modeling would call for all pertinent labels ($N_{\boldsymbol{\ell}} = 10 - 15$) that characterize one star to be fit simultaneously to its spectrum. For the large data sets at hand this appears, however, computationally infeasible: using established techniques – based on rectilinear model grids in label space – the number of required grid points is prohibitive.

The established response to this quandary is to only fit a few of the labels (typically $4 - 6$) simultaneously, and determine the other labels separately on the basis of this initial fit. In this paper we have shown that this short cut leads to important systematic errors, given the high data quality, and we offer a solution: an approach with several new techniques which we call CHAT.

CHAT's defining ideas and capabilities in determining labels, $\boldsymbol{\ell}$, from spectra can be summarized as follows:

1. Within a sufficiently small patch around a model grid point $\boldsymbol{\ell}_*$ in label space (a "Taylor-sphere"), any spectrum defined by its label vector $\boldsymbol{\ell} \equiv \boldsymbol{\ell}_* + \Delta\boldsymbol{\ell}$, can be described by a linear expansion around that grid point, as:

$$f_{\rm model}(\lambda|\boldsymbol{\ell}_*) + \overrightarrow{\nabla}_{\boldsymbol{\ell}} f_{\rm model}(\lambda|\boldsymbol{\ell}_*) \cdot \boldsymbol{\Delta\ell}, \tag{4.6}$$

   using the "gradient spectra", $\overrightarrow{\nabla}_{\boldsymbol{\ell}} f_{\rm model}$. Within this region of label space, spectral model fitting to data is then reduced to linear regression, which is computationally fast.

2. Only a tiny fraction of the high-dimensional (N_ℓ) label space of stellar parameters and elemental abundances is occupied by real stars. Given prior information, e.g., from existing surveys, an approximate *convex hull* can be constructed for this subspace, of dramatically smaller volume than a rectilinear grid in label space encompassing the relevant label subspace. This much smaller label space volume can then be covered far more densely with synthetic model spectra (for a given computational expense).

3. We have devised an adaptive tessellation of this convex hull with model grid points, $\boldsymbol{\ell}_*$, so that the set of their surrounding Taylor-spheres completely covers label space within the convex hull. Taken together, these elements of CHAT have reduced the daunting task of full spectral fitting to a) the pre-calculation of the grid points, $\boldsymbol{\ell}_*$ and the extent of their "Taylor-spheres", b) the calculation of the model spectra and gradient spectra at this modest number of $\boldsymbol{\ell}_*$, and c) linear regression when actually fitting the spectra. The linear regression also implies that the computational expense for producing the needed synthetic spectra only grows linearly with the dimensionality of label space, N_ℓ, not exponentially.

4. We have tested how well (and how fast) CHAT works in practice by considering the case of an $N_\ell = 8$ label fit to mock APOGEE red clump spectra: we show that the number of required model grid points is reduced from $\mathcal{O}(10^5)$ for rectilinear model grids, to $\mathcal{O}(10^2)$ models for the adaptively tessellated grid within the convex hull, improving the computational cost of generating a synthetic library by three orders of magnitude. We also found that CHAT recovers the best fitting labels 100 times faster than a quadratic Bézier interpolation within a rectilinear grid.

5. The dramatically fewer model points, and only linear growth of computation with N_ℓ makes full spectral fitting with far larger N_ℓ feasible. Specifically, it now seems possible to simultaneously fit additional labels such as micro-turbulence and stellar rotation, along with 15 elemental abundances.

6. We showed explicitly that fitting spectra that were drawn from a high-dimensional label space with a much smaller N_ℓ, can lead to important systematic errors, unless optimal assumptions about the non-fitted labels are made. The usual way of fitting $[Z/\mathrm{H}]$ and $[\alpha/Z]$, assuming other elements are traced by these two characteristics, works well, but some systematic residual offsets remain. If we assume the principal component coefficients to be the fitted labels, the residual offsets are reduced. Fitting principal component coefficients also provides a natural way to perform chemical tagging in a compactified elemental abundances space.

7. With these new techniques, the hope is that we can improve the precision of elemental abundances in these large surveys. As the full elemental abundances space seems to be well-described by $7 - 9$ element groups, improving the abundance precision by a factor of two will improve the resolving power of star clusters by a factor of $2^{7-9} = 100 - 1{,}000$ in elemental abundances space. With such an enormous gain in "resolution", we might be able to chemically tag Milky Way stars to their birth origins and provide a completely new view of the evolution of the Milky Way.

4.6 Appendix: Can we ignore non-primary elements in atmosphere calculations?

Generating stellar synthetic spectra consists of two parts. The first part is the calculation of the model atmosphere. Given stellar parameters and elemental abundances of a star, the temperature, pressure and electron density as a function of the Rosseland opacity of the photosphere can be calculated by solving a system of differential equations (ATLAS12). After this is done, we then proceed with a radiative transfer code (SYNTHE) and generate synthetic spectra by integrating over the photospheric atmosphere. With a restricted range in wavelength, the former step is much more time consuming (for APOGEE, it is $\mathcal{O}(10)$ times slower) than the latter step. Since it is too computationally expensive to generate a rectilinear grid of $15 - 30$ photospheric atmospheres, the standard approach is to make the assumption that the model atmosphere only depends on a few main labels. When determining the spectral variation for secondary elements, only the radiative transfer step is needed.

We put this assumption to the test. We consider fully self-consistent calculations for $[X/\mathrm{H}] = 0.2$ dex, varying one element at a time, and compare to the case where we assume a solar atmosphere, ignoring the enhanced contribution from this element. We calculate the χ^2 of these two spectra assuming $R = 20{,}000$, S/N $= 100$ and wavelength points for an APOGEE-like spectrum. In this calculation, we do not include predicted lines (from the atomic line list), in order to be more conservative in our estimate. The real difference could be larger. In Fig. 4.8, we show that, especially for low-T_{eff} stars, the contributions from secondary elements, as well as v_{turb}, can be significant. The

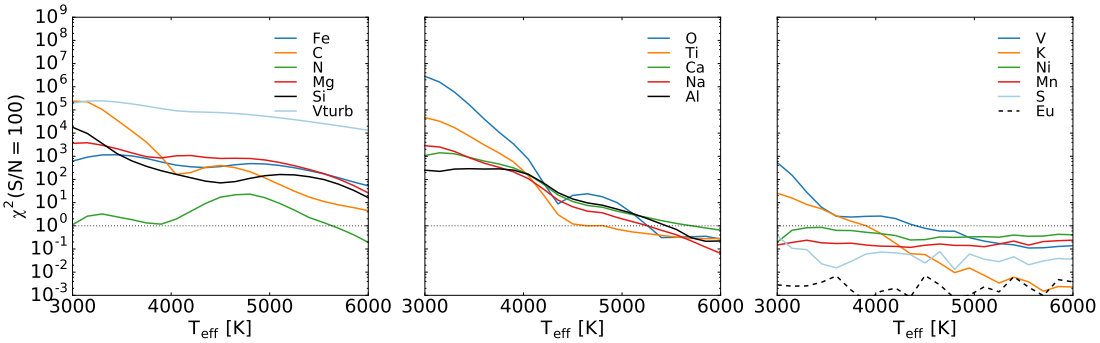

Figure 4.8: Error induced by not creating self-consistent model atmospheres when computing synthetic spectra. We plot the χ^2 between a synthetic spectrum generated with a self-consistent model atmosphere and a spectrum that was computed from a fixed, solar metallicity atmosphere. The underlying reference labels are solar metallicity, $\log g = 2.5$, $v_{\rm turb} = 2\,{\rm km/s}$, $R = 20{,}000$, and ${\rm S/N} = 100$ and the wavelength range and bins of the APOGEE survey. We consider a variation of $[X/{\rm H}] = 0.2$ and $\Delta v_{\rm turb} = 0.5\,{\rm km/s}$. Results are shown as a function of $T_{\rm eff}$. We do not include predicted lines in this calculation. If the χ^2 values are larger than the number of fitted labels (typically $\mathcal{O}(10)$), the variation is important and distinguishable in the APOGEE survey. The figure shows that many elements, as well as $v_{\rm turb}$, affect the atmosphere substantially, with the largest effects at low temperatures. We also calculate the deviation for Eu as a reference. Eu is a trace element in stars and should have no effect on the atmosphere. Nonetheless, the χ^2 for Eu is not strictly zero ($\sim 10^{-3}$), and we checked that this is due to numerical noise in the atmosphere calculation (the atmosphere is precise to the level $0.1\,{\rm K}$ for each Rosseland depth layer). Note that χ^2 depends on the ${\rm S/N}$ quadratically. Hence, assuming $\chi = 10$ to be the threshold, we conclude from the Eu result that for spectra with ${\rm S/N} \lesssim 100 \times \sqrt{10^4} = 10^4$, the numerical noise is negligible for full spectral fitting.

differences in these two cases have χ^2 values larger than the number of fitted labels (typically $\mathcal{O}(10)$). Hence ignoring secondary elements in the atmospheric calculations can bias the abundance determinations for these elements (also read[199]). On the other hand, as shown in the right panel, truly trace elements, such as Eu, indeed have negligible contributions to the atmospheric structure.

4.7 Appendix: Theoretical abundance precisions that could be achieved

Elemental abundances are usually derived from carefully chosen wavelength windows that contain absorption lines that are clean, unblended, and have reliable line parameters (usually calibrated against standards such as Arcturus and the Sun). But more information can be extracted, in principle, if we also consider the blended lines. Furthermore, as illustrated in Appendix 4.6, most elements affect the stellar opacity and atmosphere. Therefore, they will indirectly affect the line formation of the other elements. But how well we can extract this indirect information is more questionable than the blended lines. Nonetheless, it might be interesting to understand how much information there is, in principle, in the high-resolution spectra that are currently being collected. We emphasize that these theoretical precisions are not currently achievable (and may never be!) due to systematic uncertainties in the models. Systematic uncertainties aside, the information content depends on three aspects: (a) the number of uncorrelated and independent wavelength points in each spectrum, (b) the extent to which the features, in our case the depths of the absorption lines, vary as a function of the fitted labels, (c) the measurement uncertainty of the normalized flux at each wavelength bin. The measurement uncertainty could be either due to photon noise or imperfect continuum normalization. Here we only consider the ideal case where the uncertainty due to continuum normalization is negligible.

To measure how much spectral features vary as a function of the fitted labels, we consider gradient spectra with $\Delta[X/\mathrm{H}] = 0.2$, $\Delta T_{\mathrm{eff}} = 200\,\mathrm{K}$, $\Delta \log g = 0.5$, $\Delta v_{\mathrm{turb}} = 0.5\,\mathrm{km/s}$, assuming a reference point at solar metallicity, $T_{\mathrm{eff}} = 4{,}800\,\mathrm{K}$, $\log g = 2.5$ and $v_{\mathrm{turb}} = 2\,\mathrm{km/s}$. We generate spectra

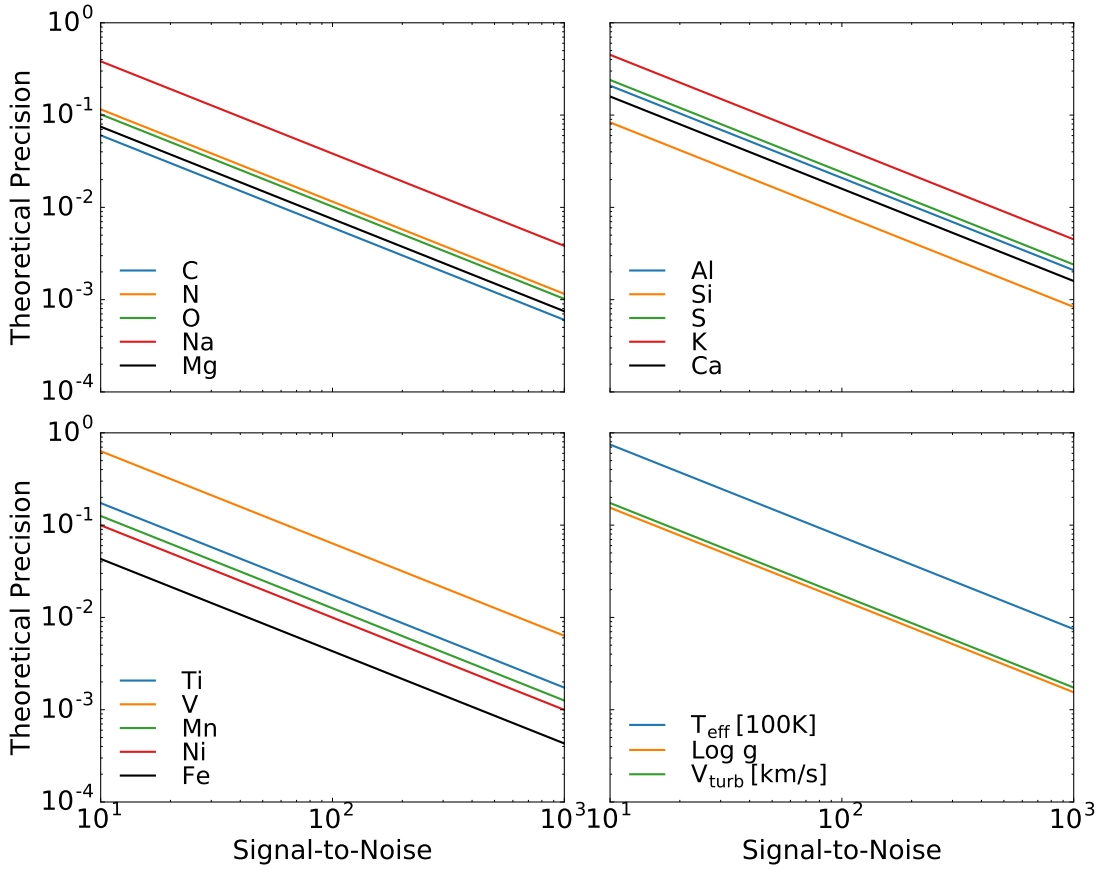

Figure 4.9: Theoretical precision for all 15 elements, $T_{\rm eff}$, $\log g$ and $v_{\rm turb}$ that we could achieve for APOGEE spectra with $R = 20{,}000$ as a function of the spectra S/N. We assume gradient spectra with $\Delta[X/{\rm H}] = 0.2$, $\Delta T_{\rm eff} = 200\,{\rm K}$, $\Delta \log g = 0.5$, $\Delta v_{\rm turb} = 0.5\,{\rm km/s}$ and with respect to the reference point at solar metallicity, $T_{\rm eff} = 4{,}800\,{\rm K}$, $\log g = 2.5$, $v_{\rm turb} = 2\,{\rm km/s}$. We do not include predicted lines in this calculation, and the limit of theoretical precision could be better than what is demonstrated here. If we have robust synthetic models and a way to fit all stellar properties simultaneously, we could, in principle, measure abundances to the precision of $\Delta[X/{\rm H}] \sim 0.01$ dex for all 15 elements, $\Delta \log g \sim 0.01$, $\Delta v_{\rm turb} \sim 0.01\,{\rm km/s}$ and $\Delta T_{\rm eff} \sim 10\,{\rm K}$ with ${\rm S/N} = 100$ APOGEE spectra.

with $R = 20{,}000$ and in the APOGEE wavelength range. We discard wavelength points that have median uncertainties $> 2\%$ in the real APOGEE spectra because these wavelength points are likely affected by telluric or sky lines. We do not include predicted lines when generating the total gradient information in this calculation to be more conservative in our estimates. We denote the variation in spectrum (the gradient spectrum) to be $\overrightarrow{\nabla}_{\ell} f_{\rm model}(\lambda)_i$. For each (i, λ), $\overrightarrow{\nabla}_{\ell} f_{\rm model}(\lambda)_i$ measures the partial derivative of the absorption line at wavelength λ with respect to label i. The Cramer-Rao bound[46,165] predicts that the covariances matrix of the fitted labels, K_{ij} can be calculated from

$$K_{ij}^{-1} = \overrightarrow{\nabla}_{\ell} f_{\rm model}(\lambda_1)_i \; C^{-1}_{\lambda_1,\lambda_2} \overrightarrow{\nabla}_{\ell} f_{\rm model}(\lambda_2)_j, \tag{4.7}$$

where C is the covariance matrix of the normalized flux. The dot product on the right-hand side serves to sum over the contribution from all wavelength points. For example, if we assume S/N $=$ 100 and only consider uncorrelated wavelength points, we have $C \sim \mathrm{diag}(10^{-4}, \ldots, 10^{-4})$. The diagonal entries of K_{ij} are the marginalized uncertainties of each label. We plot these marginalized uncertainties for each label as a function of S/N in Fig. 4.9. The figure shows that for an APOGEE spectrum with S/N $=$ 100, we could achieve a precision of $\sim$ 0.01 dex for most elements.

Finally, Fig. 4.10 shows how much information is missed by focusing on narrow spectroscopic windows. The plot shows the ratio of the information content contained in the narrow spectroscopic windows defined and used for abundance measurement by the APOGEE DR12 pipeline to the full spectral range (masking regions dominated by telluric absorption and sky lines). Fig. 4.10 shows that, for most elements, the spectroscopic windows misses $\sim 90\%$ of the information. Much

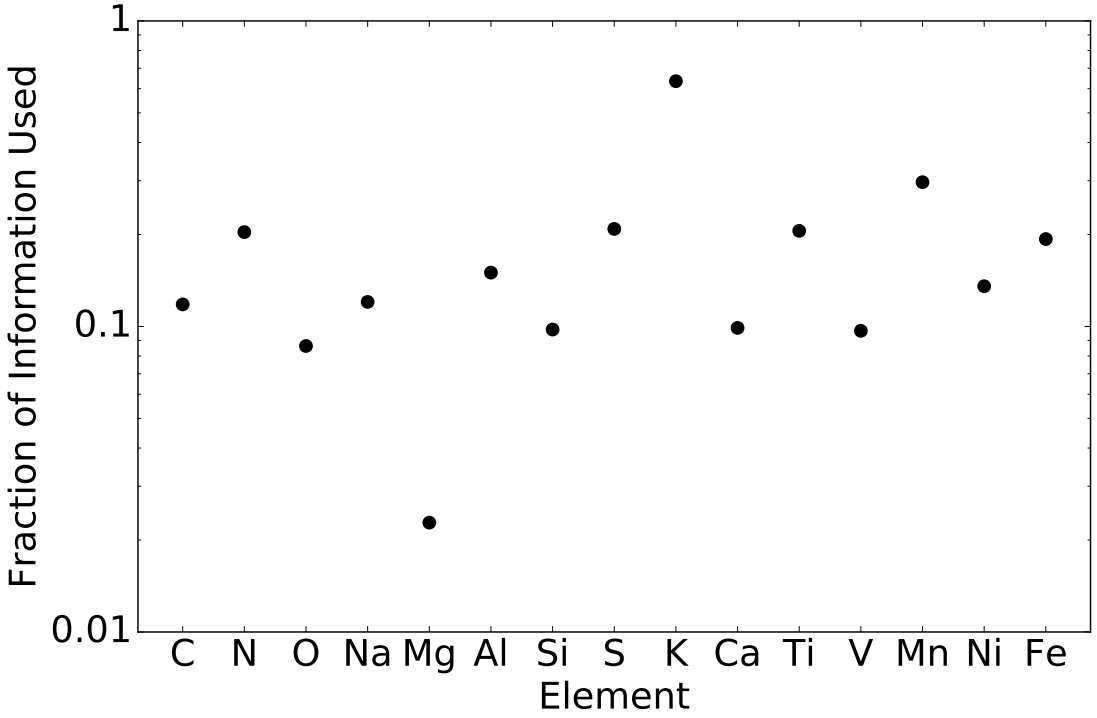

Figure 4.10: Fraction of the total available information used when limiting the fitting to narrow spectroscopic windows compared to fitting the full spectrum. The y-axis shows the ratio of the gradient information contained in the APOGEE DR12 spectral fitting windows to the gradient information in the full H-band spectrum (masking regions dominated by telluric absorption or sky lines). For most elements, restricting to narrow windows only exploits $\sim 10\%$ of the total information in spectra collected by APOGEE. The majority of this extra information is in the form of blended lines, which can be more difficult to interpret.

of the extra information is contained in blended lines and features that do require accurate models to reliably interpret. As illustrated in Eq. 4.7, the measurement precision improves in quadrature with the gradient information. Therefore, we could in principle improve the precision by a factor of three if we can minimize the systematic uncertainties in the models and perform full spectral fitting. This is the task that lies ahead.

5

Constructing polynomial spectral models for stars

Author list of the original paper: Hans-Walter Rix, Yuan-Sen Ting, Charlie Conroy, David W. Hogg

5.0 Abstract

Stellar spectra depend on the stellar parameters and on dozens of photospheric elemental abundances. Simultaneous fitting of these $\mathcal{N} \sim 10 - 40$ model labels to observed spectra has been deemed unfeasible, because the number of ab initio spectral model grid calculations scales exponentially with $\mathcal{N}$. We suggest instead the construction of a polynomial spectral model (PSM) of order $\mathcal{O}$ for the model flux at each wavelength. Building this approximation requires a minimum of only $\binom{\mathcal{N}+\mathcal{O}}{\mathcal{O}}$ calculations: e.g., a quadratic spectral model ($\mathcal{O} = 2$) to fit $\mathcal{N} = 20$ labels simultaneously, can be constructed from as few as 231 *ab initio* spectral model calculations; in practice, a somewhat larger number ($\sim 300 - 1000$) of randomly chosen models lead to a better performing PSM. Such a PSM can be a good approximation only over a portion of label space, which will vary case by case. Yet, taking the APOGEE survey as an example, a single quadratic PSM provides a remarkably good approximation to the exact *ab initio* spectral models across much of this survey: for random labels within that survey the PSM approximates the flux to within 10^{-3}, and recovers the abundances to within ~ 0.02 dex *rms* of the exact models. This enormous speed-up enables the simultaneous many-label fitting of spectra with computationally expensive *ab initio* models for stellar spectra, such as non-LTE models. A PSM also enables the simultaneous fitting of observational parameters, such as the spectrum's continuum or line-spread function.

5.1 Background

The spectra of stars encode an enormous amount of information, mainly about the stars' current physical state and the composition of the chemical elements in their photosphere. But the number of stellar labels* that fully specify a spectrum is large: a handful of stellar parameters and much of the periodic table. We know that stellar spectra with S/N $\sim$ 100 and $R \sim$ 20,000 $-$ 40,000, currently emerging for 10^{4-6} objects from various surveys, contain the information to, constrain $10-40$ labels, at least for stars with favorable effective temperatures, $\sim$ 4,000 K–7,000 K (e.g., [71,186,187]). The accuracy and precision of label estimates for vast stellar samples matters greatly for understanding the formation of the Galaxy, stellar physics, and the origin of the chemical elements (e.g., [68,69,170]).

A principled determination of these stellar labels requires to fit the data with physical model spectra, in which the stellar labels constitute 10–40 model parameters. The calculation of such *ab initio* spectral models through radiative transfer calculations has a storied tradition (for an overview, see [71,187]). Current *ab initio* models vary by the degree of physical simplification they apply: LTE *vs.* non-LTE; plane-parallel *vs.* spherical geometry; 1D, averaged or full 3D; static *vs.* time dependent; and by the extent and robustness of the atomic data that underlie them.

The computation of *ab initio* models is expensive, all the more so if the simplifying assumptions are dropped. This is why "brute force" fitting of spectra with *ab initio* models (of, say, 10–40 labels) is unfeasible for the foreseeable future: most approaches to fitting *ab initio* models to observed spectra have relied on pre-computing grids of *ab initio* spectra in the $\mathcal{N}$-dimensional label space,

*We use the term "labels" to mean the union of stellar parameters and photospheric elemental abundances, because in the current context these two classes of stellar attributes are being treated equivalently.

and then interpolating between them pixel-by-pixel, e.g., quadratically (i.e., $2^{\rm nd}$-order) or cubic (i.e., $3^{\rm rd}$-order), as in Allende Prieto et al.[3,4]. But for any number of grid points, $\mathcal{M} \approx 3-5$, in each label-dimension, the total number of *ab initio* model calculations required grows exponentially with the dimension $\mathcal{N}$ of label space: $N_{\rm tot} \propto \mathcal{M}^{\mathcal{N}} \propto \exp(\mathcal{N} \cdot \ln \mathcal{M})$. Established approaches have coped with this in practice by fitting models spectra first in a $3-6$ dimensional sub-space of $\mathcal{N}$, and subsequently fitting one (or two) further label at a time, holding the initial labels fixed. This approach has important limitations with with state of-the-art data: first, Ting et al.[193] (hereafter T16) has shown that more than just 2 or 3 elemental abundances affect the atmosphere structure, and hence are physically covariant with the basic stellar parameters; second, physical correlations and data-driven covariances are known to exist among (abundance) labels, but cannot be estimated when fitting one label at a time; third, to mitigate against unaccounted covariances, established fitting approaches have often focused on unblended lines, thereby under-exploiting the information content of the data by a large factor (T16).

T16 proposed a way to overcome this impasse by employing more linear algebra in the fitting, to save on *ab initio* model calculations; here we take this idea a step further. T16 proposed to tessellate the space of stellar labels into a finite set of regions (dubbed linear Taylor-spheres, or 1OTS). Within each 1OTS the *ab initio* model flux at each wavelength can be described sufficiently well by a linearized spectral model (LSM), linearized (in all labels) around the *ab initio* model spectrum at a fiducial label value (see also[166]). T16 showed that such LSM can sufficiently approximate the exact model spectra within a 1OTS. Together with the finite number of Taylor-spheres, required to cover any given spectral survey (e.g., $\sim$ 150 for the APOGEE red clumps), this leads to a dramatic

reduction in the total number of *ab initio* model calculations: simultaneous fitting of $10 - 40$ labels should then be feasible.

Here we point out a rather obvious extension of this idea, which yields even greater computational savings: the construction of approximate model spectra, where the predicted flux at each pixel by a polynomial in all labels away from a fiducial model spectrum. This idea had been put forth by Prugniel et al.[162] for empirical spectra, who did, however, not pursue its potential of fitting many labels simultaneously. We denote such approximate *polynomials spectral models* as PSM, to distinguish them from the *ab initio* models themselves. It is important not to think of these PSM as a $\mathcal{O}^{th}$-order interpolation between a pre-calculated grid of *ab initio* models (as e.g., Prugniel et al.[162] did for a quadratic PSM in three labels), as this would still require $\mathcal{M}_{grid}^{\mathcal{N}} \propto \exp\left(\mathcal{N} \cdot \ln \mathcal{M}_{grid}\right)$ *ab initio* model calculations. Instead, one should think of determining the (near)-smallest number of *ab initio* model spectra (specified by $\mathcal{N}$ labels) one needs to calculate in order to construct a $\mathcal{O}^{th}$-order approximation to the *ab initio* model spectra. The simplification and speed-up of such spectral fitting compared to T16 arise from the fact that a single PSM can approximate the *ab initio* model spectra over a much larger volume in label space. While this shares the idea of a polynomial flux approximation with *The Cannon*[153], it is not data-driven model building.

In the subsequent Sections we first derive that the minimal number of *ab initio* models needed to construct a PSM of order $\mathcal{O}$ and then illustrate heuristically how well, and over what volumes in label space, these PSMs approximate the *ab initio* models.

5.2 A polynomial model approximation for spectra of stars

Following T16, we suppose that an *ab initio* modeling "machinery" can predict the normalized flux of a synthetic spectrum, $\mathbf{f}_{a.i.}(\lambda|\boldsymbol{\ell})$, given a set of stellar labels, $\boldsymbol{\ell}$. We assume that the *ab initio* model spectra change from point to point in label-space, but do so smoothly or differentiably at every wavelength. Then the *ab initio* model spectrum at any $\boldsymbol{\ell}$ sufficiently close to an model grid point $\boldsymbol{\ell}_*$ (within a 1$^{\rm st}$-order Taylor-sphere or 1OTS, in the nomenclature of T16) can therefore be described with high accuracy by a linear spectral model (LSM, see T16):

$$\mathbf{f}_{\rm lin}(\lambda|\boldsymbol{\ell}_* + \boldsymbol{\Delta\ell}) \simeq \mathbf{f}_{a.i.}(\lambda|\boldsymbol{\ell}_*) + \boldsymbol{\Delta\ell}^T \cdot \overrightarrow{\mathbf{g}}(\lambda|\boldsymbol{\ell}_*), \tag{5.1}$$

where $\overrightarrow{\mathbf{g}}(\lambda|\boldsymbol{\ell}_*) \equiv \overrightarrow{\nabla}_{\boldsymbol{\ell}} \mathbf{f}_{a.i.}(\lambda|\boldsymbol{\ell}_*)$.

In principle, specifying a LSM merely requires $1 + \mathcal{N}$ model calculations, but T16 showed a a factor of a few more is needed to explore the actual extent of the 1OTS. This LSM approximation, $\mathbf{f}_{\rm lin}$, can obviously be generalized to a polynomial spectra model (PSM):

$$\begin{aligned}\mathbf{f}_{\rm PSM}(\lambda|\boldsymbol{\ell}_* + \boldsymbol{\Delta\ell}) \simeq \mathbf{f}_{a.i.}(\lambda|\boldsymbol{\ell}_*) &+ \boldsymbol{\Delta\ell}^T \cdot \overrightarrow{\mathbf{g}}\,(\lambda|\boldsymbol{\ell}_*) \\ &+ \boldsymbol{\Delta\ell}^T \cdot \underline{\underline{\mathbf{H}}}\,(\lambda|\boldsymbol{\ell}_*) \cdot \boldsymbol{\Delta\ell} + \ldots,\end{aligned} \tag{5.2}$$

where we will focus on 2$^{\rm nd}$-order, both for astrophysical reasons (it may work well enough) and to avoid cumbersome notation. Such a PSM holds for every one of the K wavelengths λ. One may think of it as a model with K 0th-order terms, $\mathbf{f}_{\rm PSM}(\lambda_k|\boldsymbol{\ell}_*)$, then $K \times \mathcal{N}$ 1$^{\rm st}$-order terms, and finally

$K \times \mathcal{N}(\mathcal{N}+1)/2$ 2^{nd}-order terms. The number $\mathcal{N}(\mathcal{N}+1)/2$ arises because of the symmetry of $\underline{\underline{\mathbf{H}}}$. In total that makes for

$$K \times N_{\text{tot}} \equiv K \times (1 + \mathcal{N} \; + \; \mathcal{N} \cdot (\mathcal{N}+1)/2) \tag{5.3}$$

unknown terms. For more general PSM of order $\mathcal{O}$, one has $K \times N_{\text{tot}} = K \times \binom{\mathcal{N}+\mathcal{O}}{\mathcal{O}}$.

If we compute *ab initio* models $\mathbf{f}_{a.i.}(\lambda|\boldsymbol{\ell}_* + \boldsymbol{\Delta\ell})$ at $\binom{\mathcal{N}+\mathcal{O}}{\mathcal{O}}$ different points in label space, $\boldsymbol{\Delta\ell}$, we have created exactly $K \times N_{\text{tot}}$ left-hand-side terms to solve exactly for the terms that specify the PSM. Note that strictly speaking $\overrightarrow{\mathbf{g}}\,(\lambda|\boldsymbol{\ell}_*)$ and $\underline{\underline{\mathbf{H}}}\,(\lambda|\boldsymbol{\ell}_*)$ are not exactly the "gradient" and the "Hessian", but merely the 1^{st} and 2^{nd}-order coefficients that solve the equation.

Compared to the 1OTS, we have to calculate $1 + \mathcal{N}/2$ times more *ab initio* models for any one quadratic PSM. But if the region in label space around $\boldsymbol{\ell}_*$ over with this quadratic PSM works is sufficiently larger, an important speed-up over the (set of) LSM should result. Calculating somewhat more *ab initio* models than this minimum, and solving Eq.5.2 in a least squares sense, makes for a much better conditioned solution for a PSM, as we show below.

5.3 Verification of polynomial spectral model accuracy

Strictly verifying the validity of the PSM approximation, like any approximation to a high-dimensional function, would be of enormous computational expense. Here, too, escaping the curse of dimensionality comes at a price: relying on the physically plausible assumption that spectral flux changes can be approximated by polynomials for modest label changes; and settling for heuristic and approximate ways to explore the extent in label space over which a single PSM is useful.

As in T16, we can set out for a pragmatic test of the PSM approximation, using Kurucz model spectra that resemble in resolution and wavelength coverage of the APOGEE spectra; the arguments should hold qualitatively for other surveys, but need to be tested case-by-case. In total, the DR12 data release of the APOGEE[2,92] provides 17 labels for each star ($T_{\rm eff}$, $\log g$ and 15 elemental abundances) while fixing $v_{\rm macro} = 6$ km/s and adopting a $\log g - v_{\rm turb}$ relation for $v_{\rm turb}$. A quadratic PSM for 19 labels requires a minimum $N_{\rm tot} = 210$ *ab initio* model calculations. We chose the reference label, $\boldsymbol{\ell}_*$, to be the APOGEE DR12 sample median in each of the 19 labels, providing $\mathbf{f}_{a.i.}(\lambda|\boldsymbol{\ell}_*)$ in Eq.5.2. The vast majority of targets in APOGEE are disk stars with all [X/H] > -1, and we restrict our PSM verification to this regime. We then drew 209 $\boldsymbol{\Delta\ell}$ at random from the APOGEE DR12 catalog. For the labels $v_{\rm turb}$ and $v_{\rm macro}$ we adopted the same $\log g - v_{\rm turb}$ relation from APOGEE with a spread of 0.2 km/s, and a distribution in $v_{\rm macro}$ uniform across 3 km/s – 8 km/s. We convolved spectra to the APOGEE resolution assuming the combined LSF from APOGEE and using codes from the APOGEE Python package[29], and continuum normalized spectra the same way as *The Cannon*[153]. This provided the remaining 209 left-hand sides of Eq.5.2 to solve exactly for the $\overrightarrow{\mathbf{g}}(\lambda|\boldsymbol{\ell}_*)$

and $\underline{\underline{\mathbf{H}}}\,(\lambda|\boldsymbol{\ell}_*)$, fully specifies $\mathbf{f}_{\rm PSM}(\lambda|\boldsymbol{\ell}_* + \boldsymbol{\Delta\ell})$ from Eq 5.2.

As expected by construction of the PSM, $\mathbf{f}_{\rm PSM}(\lambda|\boldsymbol{\ell}_* + \boldsymbol{\Delta\ell})$ matches the *ab initio* model at all the 210 $\boldsymbol{\Delta\ell}$ exactly. This minimally constructed PSM also provides good approximations to $\mathbf{f}_{a.i.}(\lambda|\boldsymbol{\ell}_* + \boldsymbol{\Delta\ell})$ for other $\boldsymbol{\Delta\ell}$. Empirical experimentation showed that slightly over constraining Eq 5.2 worked better: we calculated $\mathbf{f}_{a.i.}(\lambda|\boldsymbol{\ell}_* + \boldsymbol{\Delta\ell})$ for 250 and 1,000 $\boldsymbol{\Delta\ell}$ drawn from APOGEE, and solved for the right hand side of Eq.5.2 in a least squares sense to determine the PSM coefficients.

There are two ways in which one can quantify how well the PSM, $\mathbf{f}_{\rm PSM}(\lambda|\boldsymbol{\ell}_* + \boldsymbol{\Delta\ell})$, approximates $\mathbf{f}_{a.i.}(\lambda|\boldsymbol{\ell}_* + \boldsymbol{\Delta\ell})$ for any $\boldsymbol{\Delta\ell}$ drawn from APOGEE: how well do the fluxes match, e.g., in a mean absolute deviation? And, at what accuracy level does the PSM approximation affect the label recovery?

Fig.5.1 illustrates how well an *ab initio* model spectrum of random star in APOGEE can be approximated by the PSM in a mean absolute deviation sense. On average the PSM-predicted flux at any wavelength for a random star within APOGEE is within 10^{-3} or $10^{-3.5}$ of that for its *ab initio* model spectrum, depending on whether we used 250 or 1,000 *ab initio* model calculations to construct the PSM. Fig 5.2 shows how much (or, how little) the PSM approximation, calculated here on the basis of 250 or 1,000 *ab initio* models, affects the label recovery across an APOGEE-like survey. The labels were recovered by a least squares fit of the PSM to noiseless *ab initio* models, fitting all 19 labels *simultaneously*. These were then compared to the actual labels of the respective *ab initio* models. With a single PSM, most labels are recovered as accurate as claimed precisions of current spectral surveys. More quantitatively the quality of the PSM label recovery is well-tracked by the information content that the spectra contain about any one label: following T16, this is quantified by the

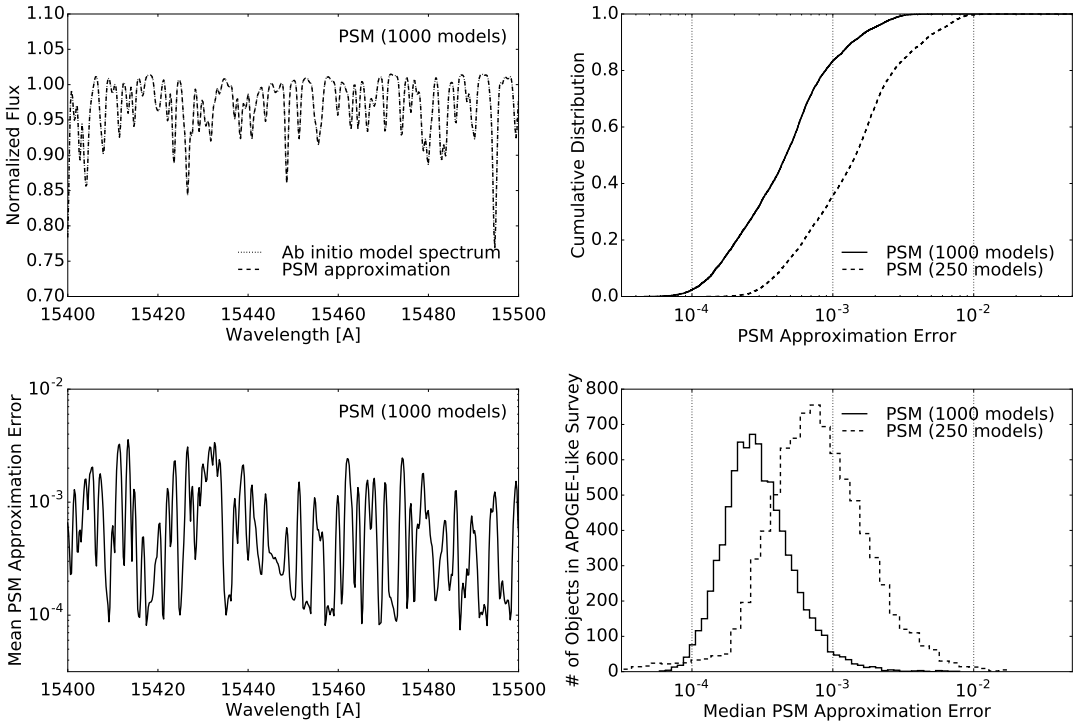

Figure 5.1: Quality of the (quadratic) PSM approximation: a single PSM was constructed using 250 or 1,000 *ab initio* model spectra (*cf.* the absolute minimum number is 231), calculated at label points ("objects") drawn randomly from those in the APOGEE survey[2,92]. The panels illustrate different PSM – *ab initio* model comparisons, for 10,000 other objects drawn from the labels of the APOGEE survey. The top left panel shows for a limited wavelength section the average of the exact *ab initio* model spectra and of the PSM, which appear indistinguishable. The bottom left panel shows the ensemble average (absolute) difference between the *ab initio* model and the PSM flux (the approximation error), as a function of wavelength. For each one of the 10,000 objects there is a pixel-by-pixel distribution of these approximation errors, which is shown in the top right panel for the pixel-by-pixel average approximation error. The bottom right panel finally shows the distribution across all objects of their (pixel-by-pixel) median approximation error. Note that there are rare cases (objects of very high [Fe/H], where the approximation is only good to a median of 10^{-3}. Taken together, however, this shows that a single PSM approximates the exact *ab initio* model spectra typically to within 10^{-3} for objects with a label distribution resembling that of the entire APOGEE survey (which merely serves as an illustration here), over the 10,000 labels of the median. Constructing the PSM from 1,000 instead of 250 random label points leads to a better PSM approximation.

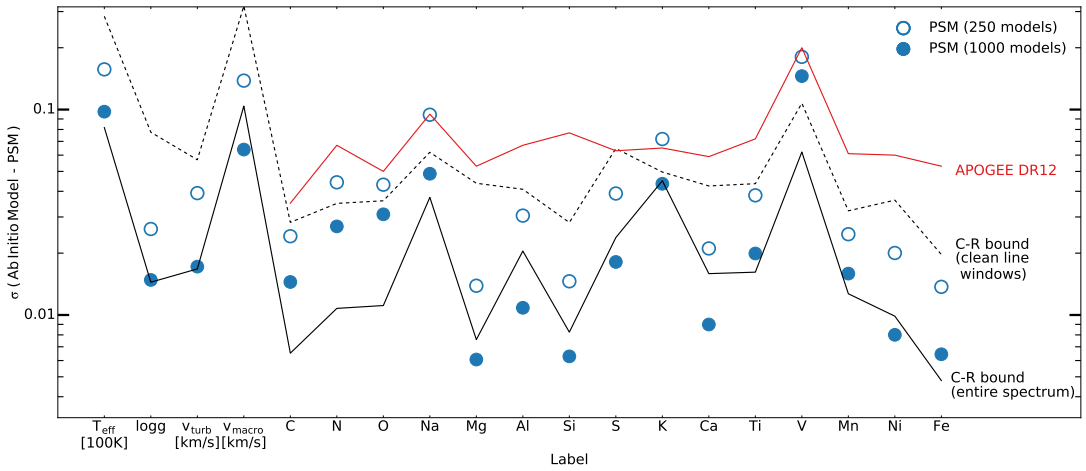

Figure 5.2: Quality of the label recovery using the same PSM approximation as in Fig.5.1, based on 250 (open circles) and 1,000 (full circles) *ab initio* model calculations, respectively. Shown is the *rms* difference between the labels of the PSM approximation that best matches the exact *ab initio* model spectrum in a χ^2-sense, and the actual labels of the exact spectrum: PSM- induced errors in the label recovery by the PSM approximation are typically 0.02 dex (when considering the label range of the APOGEE survey). The dashed and solid lines show the theoretically achievable label precision at S/N $=$ 100 (the Cramer-Rao bound; see T16), when using the APOGEE wavelength windows, or the full spectrum. A single PSM approximation can be used for fitting all labels simultaneously across much of the APOGEE survey, without inducing serious systematic errors. The red line indicates typical APOGEE DR12 precisions. The quality of the label recovery remains (to within $\sim$ 10% of each label's accuracy), even if a number of spectral continuum and line-spread parameters are also fit simultaneously.

Cramer-Rao bound (for S/N$\sim$ 100) using either only the APOGEE wavelength windows for certain elements or the whole spectrum.

The PSM appears heuristically as a better approximation when calculated on the basis of more *ab initio* model calculations, presumably for two reasons: the system of linear equations in Eq.5.2 becomes better conditioned; and a better sampling of label-space better mitigates any break-down of the polynomial approximation. Both factors must play a role: when we restrict the label range over which we first construct and then test the PSM, the PSM label recovery is even closer to the exact solution. Yet, the PSM constructed on the basis of 1,000 (compared to 250) *ab initio* model calculations is still performing better. How many models to calculate for the PSM construction, and over which portion of label space to apply it will therefore depend in practice on the computational expense of the *ab initio* models and the desired label accuracy. Nonetheless, Fig.5.1 and Fig.5.2 demonstrate that with calculating only 250 (or 1,000) *ab initio* models one can construct a single (quadratic) PSM that performs remarkably well in approximating results from exact model spectra at random 19-dimensional label location across much of APOGEE survey.

5.4 Summary and conclusion

We have shown the advantages for spectral model fitting of generalizing the local linear expansion of *ab initio* model spectra laid out in T16 to higher order, constructing polynomial spectral models (PSM) that approximate the variations of the predicted spectral flux at each wavelength as a polynomial function of the labels. This reduces the calculation of the model spectra needed in simultaneous fitting of many stellar labels to observed spectra to linear algebra. Compared to established approaches that first calculate grids and then interpolate, the dramatic gain in constructing a PSM comes from the much more benign scaling of the computational effort with increasing label dimension: $\propto \binom{\mathcal{N}+\mathcal{O}}{\mathcal{O}}$, or $\propto \mathcal{N}^2$ for a quadratic model with $\mathcal{O} = 2$, as opposed to $\propto \exp(\mathcal{N} \cdot \ln \mathcal{M})$. The way these PSM are constructed are mathematically very much analogous to data-driven *The Cannon*[153], where a quadratic spectral model is derived form observed spectra. The arguments here provide a systematic guidance for the size of the required training set in *The Cannon*: we should expect the training set size to scale as (a multiple) $\binom{\mathcal{N}+2}{2}$, or $\propto \mathcal{N}^2$; this makes it plausible that *The Cannon* could constrain 19 labels from a training set of 10,000[40].

The heuristic verification of the PSM approximation, along with the framework laid out in T16, means that there should be no longer serious technical obstacles to determining stellar labels in large surveys to what amounts to fitting all labels with *ab initio* model spectra simultaneously. The accommodation of label correlation facilitates the extraction of abundance information from blended spectral features. We find from the gradient spectra, that 80% of the spectrum's information on a label is spread over typically 30% of all pixels, and is not just in narrow spectral windows (T16). For

any given data set this should allow higher precision and accuracy. PSM also allows to treat parameters of the experimental set-up, such as the continuum fit or the spectral line-spread function (LSF) quasi as stellar labels, and fit them simultaneously.

Of course, constructing PSMs is not a panacea: while a single PSM appears to suffice for the much of APOGEE survey, this is presumably because APOGEE has targeted stars in a rather restricted portion of label space: giant stars in a narrow temperature range. Yet, even there, constructing a separate PSM for the metal-poor regime may be advisable, as small model flux differences cause larger label recovery errors. Second, it is probably worth exploring the PSM approach to higher order in at least some of the labels. Perhaps most importantly, any fitting based on *ab initio* spectral models can only work as well as the physics behind them. Insufficient atomic data or the restrictions of the LTE approximation remain untouched by the ideas laid out here. Nonetheless, we feel that T16 and this paper lay out a path that may help in doing justice to the enormous information content of present and future stellar spectroscopy surveys.

6

Measuring abundances of more than 20 elements with low-resolution stellar spectra

Author list of the original paper: Yuan-Sen Ting, Charlie Conroy, Hans-Walter Rix. Phillip Cargile

6.0 Abstract

Understanding the evolution of the Milky Way requires precise abundances of many elements for many stars. A common perception is that deriving more than a few elemental abundances ([Fe/H], [α/Fe], perhaps [C/H], [N/H]) requires medium-to-high spectral resolution, $R \gtrsim 10{,}000$, mostly to overcome the effects of line blending. In recent work[172,193] we have shown how one can efficiently model the full stellar spectrum, fitting an arbitrarily large number of stellar labels simultaneously. In this paper we quantify to what precision the abundances of many different elements can be recovered as a function of spectroscopic resolution and wavelength range. In the limit of perfect spectral models we show that the precision of elemental abundances is nearly independent of resolution for a fixed exposure time and number of detector pixels; low-resolution spectra simply afford much higher S/N per pixel and generally larger wavelength range in a single setting. We also show that estimates of most stellar labels are not strongly correlated with one another once $R \gtrsim 1{,}000$. Modest errors in the line spread function do not affect these conclusions. These results, to be confirmed with an analysis of observed low-resolution data, open up new possibilities for the design of large spectroscopic stellar surveys and for the re-analysis of archival low-resolution datasets.

6.1 Background

Massively multiplexed stellar spectroscopic surveys are a central part of the current astronomy landscape, aimed at understanding stellar physics, the genesis of elements in the cosmos and the chemical/dynamical evolution of the Milky Way. This field is currently undergoing a revolution in the quality and quantity of spectra (e.g., see review from[170]): current surveys aim to collect high quality spectra for millions of stars. But these extensive datasets bring new analysis and modeling challenges. Novel approaches are emerging (e.g., [40,153,172,193]) for turning these massive datasets into precise stellar labels, encompassing stellar parameters and elemental abundances of stars.

Spectral resolution, R, is a key parameter characterizing spectroscopic surveys, and the goal of this paper is to determine the resolution required to measure stellar labels to a desired precision. Traditionally, stellar spectroscopy has parsed itself into three resolution regimes: low-resolution with $R \lesssim 10{,}000$, medium-resolution with $10{,}000 \lesssim R \lesssim 50{,}000$, and high-resolution with $R \gtrsim 50{,}000$. Low-resolution $R \simeq 2{,}000$–$10{,}000$ spectroscopic surveys, such as SEGUE[211], RAVE[188], Gaia Radial Velocity Spectrometer (RVS)[167], and LAMOST[134], have aimed at deriving fundamental stellar parameters such as $T_{\rm eff}$, $\log g$ and radial velocity. However, because essentially all stellar spectral lines are blended at low-resolution, only measurements of [Fe/H] and a few elements such as $[\alpha/\mathrm{Fe}]$, [C/H], [N/H] have been attempted systematically (e.g., [67,106,125,126,210]). But even with a limited number of stellar labels, these surveys are crucial because they can amass the statistical samples necessary to provide a global view of the Galaxy. For example, metallicity distribution functions can be used to infer star formation histories (e.g., [39,86]); metallicity gradients provide a window into

the global dynamical history of stars (e.g., [78,102,178]) and the inside-out formation of the Milky Way (e.g., [148,179]), and stellar age-metallicity-kinematic dispersion relations identify the extent to which stars become kinematically dispersed over time due to dynamical heating (e.g., [8,79,140]).

Medium-resolution spectroscopic surveys such as the on-going APOGEE [136], GALAH [53] and Gaia-ESO [187] surveys and the forthcoming 4MOST survey [82] aim to collect stellar spectra with $R \simeq 24{,}000$. These surveys are designed to overcome the perceived shortcomings of their low-resolution counterparts, and aim to measure detailed elemental abundances of $10-40$ elements. Precise abundances for many elements are a key to understanding the chemical evolution of the Milky Way, as well as stellar nucleosynthesis. For example, core-collapse supernovae from massive stars produce relative overabundances of α-capture elements (e.g., [128,207]), whereas type Ia supernovae produce overabundances of iron-peak elements (e.g., [93] , and also review from Nomoto et al. [155]). Mass loss from AGB stars adds additional complexity to the chemical evolution of the Milky Way (e.g., [100,203]). Ting et al. [195] used principal component analysis and demonstrated there are at least seven pathways for galaxies to collect their metals. One goal of deriving multi-elemental abundances for many stars is to unravel the contributions from these different channels at various evolutionary stages of the Milky Way.

High-resolution spectra, with $R \gtrsim 50{,}000$, are the gold standard for measuring precise and accurate stellar parameters and detailed abundances. At this resolution many of the strong stellar absorption lines are unblended. However, such spectra make high demands on instrumentation and exposure time. For this reason high-resolution surveys (e.g., [13,36,64,89,95,94]) contain far fewer stars than medium and low-resolution surveys.

An exciting application of precise abundance measurements for many elements is the concept of

chemical tagging[69]: if stars born from the same molecular cloud share the same or very similar elemental abundances (as suggested by recent observational works[29,131]), then elemental abundances can serve as a birth-tag for each star. The goal of "strong" chemical tagging is to look for stellar siblings by searching for similarities in chemical space[69]. "Strong" chemical tagging has proven to be challenging and is yet to be realized in practice, in part because it requires a vast sample size and very precise elemental abundances[130,191]. But a weaker form of chemical tagging has been demonstrated to be viable (e.g., [91,139,163]). "Weak" chemical tagging uses precise measurement of elemental abundances to separate various groups of stars. For example, dwarf galaxies in the Milky Way are separable both from each other and from the Milky Way stellar halo in [α/Fe]-[Fe/H] space (e.g., [202]); globular cluster stars have unique abundance patterns that allow their identification in the Milky Way bulge and stellar halo (e.g., [138,177]); and the thick disk, thin disk and halo stars can be well separated with their α-enhancement measurements (e.g., [85,86]).

With strong scientific motivation for precisely measured elemental abundances of many ($\gtrsim$ 20) elements in many ($N > 10^6$) stars, it is worth revisiting the optimal survey configuration to achieve these goals. In this paper, we will demonstrate that – at a given exposure time or survey speed, and for a fixed number of detector pixels, low-resolution spectra can constrain comparably many elements and at the same precision as medium-resolution spectra. Moreover, the estimates of elemental abundances show little correlation, once $R \gtrsim$ 1,000, even though the spectral lines are severely blended at low-resolution. These conclusions apply in the limit of very high quality spectral models, although the influence of bad pixels is smaller than often assumed. These results suggest new strategies for designing future generations of spectroscopic surveys.

This paper is structured as follows – in Section 6.2, we provide an overview of basic concepts and intuition related to spectra as a function of resolution, and describe how to model spectra in high dimensional space. We explore the information content of low-resolution spectra in Section 6.3, and in Section 6.4 we perform spectral fitting on synthetic spectral data with characteristics similar to the APOGEE survey. In Section 6.5, we discuss some implications of these results and highlight several caveats. We conclude in Section 6.6.

6.2 Basic ideas

In this section we present the basic arguments for why there need not be significant loss of abundance information when choosing a spectroscopic survey with $R \sim 1{,}000$, instead of $R \sim 100{,}000$. The arguments presented in this section turn out to be fairly insensitive to the detailed wavelength range chosen for any spectroscopic survey. A key assumption throughout this work is the assumption of sufficiently accurate spectral models. All commonly used *ab initio* stellar spectral models (e.g., [80,84,120,121,122]) have significant deficiencies, e.g., arising from incomplete and/or inaccurate atomic and molecular line parameters and the assumptions of 1D LTE (see also [187]). Therefore some of the results we present speak at present to the information content, in principle, of low-resolution spectra, but may require data-driven models [40,153] and/or improved *ab initio* spectral models in order to apply these results to real data. Moreover, in future work we will directly test many of our conclusions by fitting models to observations of stars with spectra obtained with a variety of spectral resolutions and wavelength ranges.

6.2.1 The advantage of high-resolution spectra

Photospheric elemental abundances are encoded in the strengths of atomic and molecular absorption lines. The most common classical methods of measuring elemental abundances rely on the measurement of the equivalent width of absorption lines with well-known atomic parameters. Equivalent width measurements can then be placed on a curve-of-growth, given a stellar model, in order to derive the abundance of the species giving rise to the observed feature. One must also know the

effective temperature and surface gravity, as these parameters have large and complicated effects on the strengths of lines. Frequently the effective temperature is determined by independent (and not necessarily self-consistent) means, e.g., by color-temperature relations (e.g., [39]).

Photospheric lines are broadened by various processes, including pressure broadening, rotation, and macro-/micro-turbulence. These sources of line broadening, combined, are typically of the order $v_{\rm broad} \simeq 1 - 10\,{\rm km/s}$, which translates to an intrinsic spectral resolution of star of $R = \lambda/\Delta\lambda = c/v_{\rm broad} \simeq 10^4 - 10^5$. In order to resolve spectral lines one would therefore want to obtain spectra at $R \gtrsim 10^4$. At lower resolution the lines blend together, at least in cool and metal rich stars (the most common stars in most large stellar spectroscopy surveys). It is then no longer possible to simply measure equivalent widths. The most straightforward way to make progress in such cases, while preserving the full information content of the spectra, is to self-consistently fit entire spectral regions, which is the approach taken here.

Another advantage of operating at high-resolution is that one can isolate and focus on the spectral lines whose atomic parameters are well known from laboratory work, and discard spectral regions that are not well-modeled, as a way to mitigate systematic uncertainties in the models. This is a clear advantage of working at high-resolution (but see also [47]), although the extent to which this issue can be mitigated or addressed at low-resolution has not been thoroughly addressed.

Finally, there are very subtle effects in the spectra of stars that would be entirely invisible at low-resolution, such as isotope ratios. For example the measurement of the ^{24}Mg/^{25}Mg isotopic ratio, which induces a shift of ~ 0.01 nm in the MgH spectral lines, or the ^{6}Li/^{7}Li isotopic ratio [129], which requires 3D NLTE models to properly model the line shapes and derive reliable isotopic ratios.

6.2.2 Quantifying information content with gradient spectra

In order to understand why low-resolution spectroscopy could possibly perform comparably well, we need a metric for the theoretically achievable uncertainties for each stellar label. A compact but mathematically rigorous way to do this is the Cramer-Rao bound[46,165], which we introduced in this particular context in Ting et al.[193]. How well we can estimate a stellar label depends on two things, (a) how much a spectrum varies as we vary the stellar label, i.e., the response function of a spectrum, and (b) the flux uncertainties at each wavelength pixel and their covariances. Let C be the covariance matrix of the normalized flux. The Cramer-Rao bound predicts that the covariance matrix of the stellar labels K_{ij} can be calculated via

$$K_{ij}^{-1} = \overrightarrow{\nabla}_{\ell} f_{\rm model}(\lambda_1)_i \; C^{-1}_{\lambda_1,\lambda_2} \overrightarrow{\nabla}_{\ell} f_{\rm model}(\lambda_2)_j. \tag{6.1}$$

For each (i, λ), the "gradient spectrum" $\overrightarrow{\nabla}_{\ell} f_{\rm model}(\lambda)_i$ measures the variation of the spectral flux at wavelength pixel λ with label i. Eq. 6.1 then essentially takes the quadrature sum of the variations across different wavelength pixels, weighted by the uncertainties of the observed flux. For example, if the spectral response to label changes is steep, we have large values for $\overrightarrow{\nabla}_{\ell} f_{\rm model}(\lambda)_i$ and hence small values for K_{ij} – more precise measurements. Similarly, if the observed spectra have a higher S/N, the values for C will be smaller which will also result in smaller values for K_{ij}. The sum extends over the available pixels in the spectrum. Throughout this paper we assume that the wavelength sampling is always $\lambda/3R$ (we adopt a factor of 3, following the sampling of the APOGEE survey spectra).*

*Note that for most high-resolution echelle spectrographs the wavelength sampling and spectral resolu-

We simplify the calculation in Eq. 6.1 by assuming that there are no correlations between adjacent wavelength points, i.e., C^{-1} is a diagonal matrix.

The covariance matrix K_{ij} of the stellar labels and the gradient spectra are the quantities on which we base the majority of our results in this study. Not only it is a mathematically robust way to represent how much spectral information there is in the spectra, it also predicts which elements can be detected above a given significance threshold and the covariance between different stellar label estimates. Clearly, the calculation of K_{ij} depends on the chosen resolution and wavelength range. As we vary the resolution, we will be summing up from different wavelength pixels, and the gradient spectra will also change with resolution. In short, in order to evaluate how low-resolution spectroscopy performs compared to medium-resolution spectroscopy, we will study how K_{ij} varies as a function of resolution, spectral type, and wavelength range.

6.2.3 Many stellar labels from low-resolution spectra

We will now examine, at first qualitatively, how the uncertainties in stellar label estimates vary as a function of resolution in order to gain some basic intuition. It is qualitatively clear that at the same wavelength range and the same S/N per resolution element, a high-resolution spectrum must contain much more information than a low-resolution spectrum. But for spectroscopic surveys there are two important boundary conditions that need to be taken into account for a sensible comparison of spectra taken at different resolutions: the first one is the exposure time per object, which sets the survey speed; the second is the number of available detector pixels onto which each spectrum

tion are not necessarily directly connected in this way.

can be mapped. As detector "real estate" is an important boundary condition in highly multiplexing spectroscopic surveys, higher resolution generally forces the choice of a proportionally smaller wavelength range (assuming a fixed number of pixels per resolution element). As a consequence, the spectra at lower resolution will have higher S/N per resolution element and larger wavelength range (with a greater chance to enclose key diagnostic lines of different elements). Both effects work in favor of the low-resolution spectra. We can now evaluate analytically how K_{ij} changes as we lower the resolution:

1. The *rms* depths of narrow spectral lines decrease inversely proportional to the width of the line spread function (LSF) kernel. As a result, the *rms* values of the gradient spectra scale as R. Seen another way, the equivalent width (the total integral) of spectral lines is constant at different resolutions, but the size of a resolution element is proportional to $1/R$. Therefore, the *rms* depth per wavelength pixel sampled at each resolution element must scale as R so that the equivalent width (the sum of Δ resolution element width $\times$ gradient) is constant.
2. On the other hand, for fixed exposure time and object flux, the S/N per pixel will improve by $1/\sqrt{R}$ due to Poisson statistics.
3. Furthermore, for a given number of detector pixels, the wavelength range scales as $1/R$. Assuming the spectral lines are evenly distributed, we will collect $1/R$ times more spectral lines at low-resolution. As information adds in quadrature, having R times more lines will improve the information content by a factor of $\sqrt{R}$, thus the precision improves proportional to $1/\sqrt{R}$.

These simple arguments show that to first order low-resolution and medium-resolution spectra should achieve the same uncertainty for stellar labels, given the sensible boundary condition of equal survey speed. In fact, provided that we have robust models and the ability to fit all stellar labels simultaneously, the uncertainty should be entirely independent of resolution, at least so long as the assumptions above hold. We will show with simulations in Section 6.3 that this insensitivity to reso-

lution holds over a perhaps surprisingly large range in R. However in practice, the label uncertainty is not entirely independent of resolution, especially at the highest and lowest resolutions:

- For elements that have only few spectral lines, expanding the wavelength range does not necessarily generate more information. The newly included wavelength range might be devoid of spectral lines for some elements. So, for low-resolution spectra, we might lose information by a factor of $\sqrt{R}$. However, including a wider wavelength range also implies that low-resolution spectra can detect more elements that might have no detectable lines in medium-resolution spectra of necessarily narrower wavelength range.
- Once the LSF at very high-resolution becomes narrower than the intrinsic broadening of most lines, further increasing the resolution does not improve the gradient spectra. Therefore, for a fixed exposure time the information content of an observed spectrum will decrease at higher resolution. However, such high-resolution will in some cases be critically important for dealing with systematic issues, e.g., identifying and removing telluric features, which are intrinsically narrower than most stellar lines.
- At very low-resolution (e.g., $R \sim 100$ as for Gaia's BP/RP spectra), estimates for stellar labels become more correlated. Mathematically, the covariance matrix K_{ij}^{-1} becomes less diagonal. In other words, once the estimates of different stellar labels become highly degenerate, their individual estimates become less precise.
- When modeling low-resolution spectra one is forced to fit the full spectrum and one must therefore have knowledge of the line spread function (LSF) across a wide wavelength range. This can introduce additional challenges that are not as severe when modeling equivalent widths of individual features from high-resolution data.

6.2.4 Fitting multiple stellar labels simultaneously

We have argued that low-resolution spectra contain the same amount of information for a fixed number of pixels and at fixed exposure time, but we can extract this information only if we are able

to fit all stellar labels simultaneously. Generating state-of-the-art model spectra over a wide wavelength range takes several CPU hours for a given set of stellar labels. In a parameter space of $20-60$ labels, it is computationally prohibitive to search for the best-fitting stellar labels through minimization – each step in the minimization process will take several CPU hours. The standard approach to this problem is to create a synthetic library on an approximately rectilinear grid in the stellar label space, creating models at each grid point and then interpolating between them (e.g., [71]). However, in this method, the number of models needed grows exponentially with the number of labels, implying insurmountable computational cost for fitting $20-60$ labels. We tackled this problem in Rix et al. [172], Ting et al. [193] and devised a new algorithm – polynomial spectral models (PSM) – that can fit $20-60$ labels simultaneously. In essence, PSM constructs a model for the predicted flux at each wavelength point in the label space that is a polynomial function of all labels. But it does so in a way that requires only $N^2 \sim 1{,}000$ *ab initio* models, for $N = 20-60$ labels. The success of such modeling depends of course on whether the stellar spectra to be fit are well approximated by a PSM. In Rix et al. [172], we found that a single second order expansion captures almost all the label space spanned by the APOGEE sample of giants with $T_{\rm eff} > 4{,}000$ K. Rix et al. [172] found that the median deviation of the normalized flux between *ab initio* calculated APOGEE models with 18 parameters and the PSM models is only 0.001. Such an "interpolation error" is negligible as it is an order of magnitude smaller than the typical S/N of an observed spectrum (S/N $\simeq$ 100). Furthermore, finding the best-fitting models with PSM is also extremely efficient because it regularizes the likelihood space in a χ^2-minimization. For instance, we found that PSM fitting 100,000 APOGEE spectra with 20 parameters requires less than 100 CPU hours. PSM therefore appears to provide a practical solution to

the requirement in low-resolution spectra of fitting all labels simultaneously. We will demonstrate how PSM can be used to fit low-resolution spectra in Section 6.4.

6.3 The information content of low-resolution spectra

In Section 6.2.3 we discussed analytically, and qualitatively, why stellar parameter estimation should not depend strongly on spectral resolution under certain conditions. In this section we explore this issue in more detail by using synthetic model spectra and evaluating how uncertainties on stellar labels, calculated with Eq. 6.1, vary as a function of spectral resolution. We use model spectra to calculate the gradient spectra, $\overrightarrow{\nabla}_{\ell} f_{\rm model}(\lambda)_i$, in Eq. 6.1 and the label covariance matrices that reflect the label uncertainties, under the assumption that the models are a good description of the data.

6.3.1 Setup

We compute 1D LTE model atmospheres from the ATLAS12 code maintained by R. Kurucz[118,119,123]. We adopt the latest line list provided by R. Kurucz,[†] which include TiO, H_2O, CH, CN, CO, OH, MgH amongst other molecules. We evaluate the atmospheric structure with 80 zones of Rosseland optical depth, $\tau_{\rm R}$, with the maximum depth of $\tau_{\rm R} = 1{,}000$. We automate the numerical convergence inspection for each calculated atmosphere and adopt the solar abundances from Asplund et al.[7]. We adopt the standard mixing length theory with a mixing length of 1.25 and no overshooting for convection. Spectra are evaluated with the radiative transfer code SYNTHE with a nominal resolution $R = 300{,}000$ and are subsequently convolved to lower resolutions assuming a normal distribution with a FWHM of λ/R.

To calculate approximate gradient spectra, we consider the differences of two spectra that differ

[†] http://kurucz.harvard.edu

by $\Delta T_{\rm eff} = 250$ K, $\Delta \log g = 0.5$, $\Delta v_{\rm turb} = 0.5$ km/s, $\Delta[X/{\rm H}] = 0.2$ with respect to a chosen reference point; Ting et al.[193] elaborated why that is a sensible approximation. For any stellar label i and reference point, we calculate the gradient spectra as:

$$\nabla_{\boldsymbol{\ell}} f_{\rm model}(\lambda, \boldsymbol{\ell}_i) = \frac{f_{\rm model}(\lambda, \boldsymbol{\ell}_i + \Delta\boldsymbol{\ell}_i) - f_{\rm model}(\lambda, \boldsymbol{\ell}_i)}{\Delta \boldsymbol{\ell}_i}, \tag{6.2}$$

where $f_{\rm model}$ is the normalized flux of a model spectrum. In this study, we always perform full-consistent calculations – we re-evaluate the atmospheric structure whenever we vary a stellar label, even though in many cases, e.g., for Eu, this is unnecessary (see[193] for details). This is an important point because many elements have a significant effect on the atmospheric structure, which in turn can affect the emergent spectrum. So for example an enhancement in Na not only affects the atomic Na I lines but also, at a lower amplitude, large regions of the spectrum owing to the change in the atmospheric structure (Na is a major electron donor in cool stars).

To study how the results vary for different stellar types, we consider a few reference points in this study, namely:

- M-giants: $T_{\rm eff} = 3{,}800$ K, $\log g = 0.5$
- K-giants: $T_{\rm eff} = 4{,}800$ K, $\log g = 2.5$
- G-dwarfs: $T_{\rm eff} = 5{,}800$ K, $\log g = 4.5$
- F-dwarfs: $T_{\rm eff} = 6{,}800$ K, $\log g = 4.5$

We adopt the following relation between $v_{\rm turb}$ and $\log g$[92]:‡

$$v_{\rm turb} = 2.478 - 0.325 \log g \quad {\rm km/s}, \tag{6.3}$$

The top panels of Fig. 6.1 illustrate gradient spectra for three elements – C, Fe, and K – assuming a K-giant, solar metallicity reference point. We consider three different resolutions – the nominal model resolution at $R = 300{,}000$, a medium-resolution mode, $R = 24{,}000$, and a low-resolution mode, $R = 6{,}000$. We also show the normalized spectra with and without enhancements in the abundances of these three elements in the lower panels. At $R = 300{,}000$ and $R = 24{,}000$, some of the spectral lines are resolved and unblended. These lines are typically selected to measure equivalent widths and derive elemental abundances. Carbon has many more lines due to molecular contributions. Elements such as potassium have far fewer lines. However, at $R = 6{,}000$, all lines are blended. If we wish to derive the elemental abundance of potassium, for example, we will need to model other elements contributing to the blends at the same time. Therefore, to extract spectral information at $R = 6{,}000$, we need to model the blended lines by fitting all relevant stellar labels simultaneously.

The top panels of Fig. 6.1 reveal a few interesting features. For e.g., at $R = 300{,}000$, the global depths of spectral lines are not exactly $300{,}000 / 6{,}000 = 50$ times deeper than $R = 6{,}000$. There are three effects in play: (a) at $R = 300{,}000$, the intrinsic broadening is larger than the LSF broad-

‡APOGEE calibrated this relation with giants, so this relation might not apply to the broad range of stellar types in this study. But the goal here is to have a wide variety of $\log g$ and $v_{\rm turb}$ as our reference points, so the exact relation between these two parameters does not impact our results.

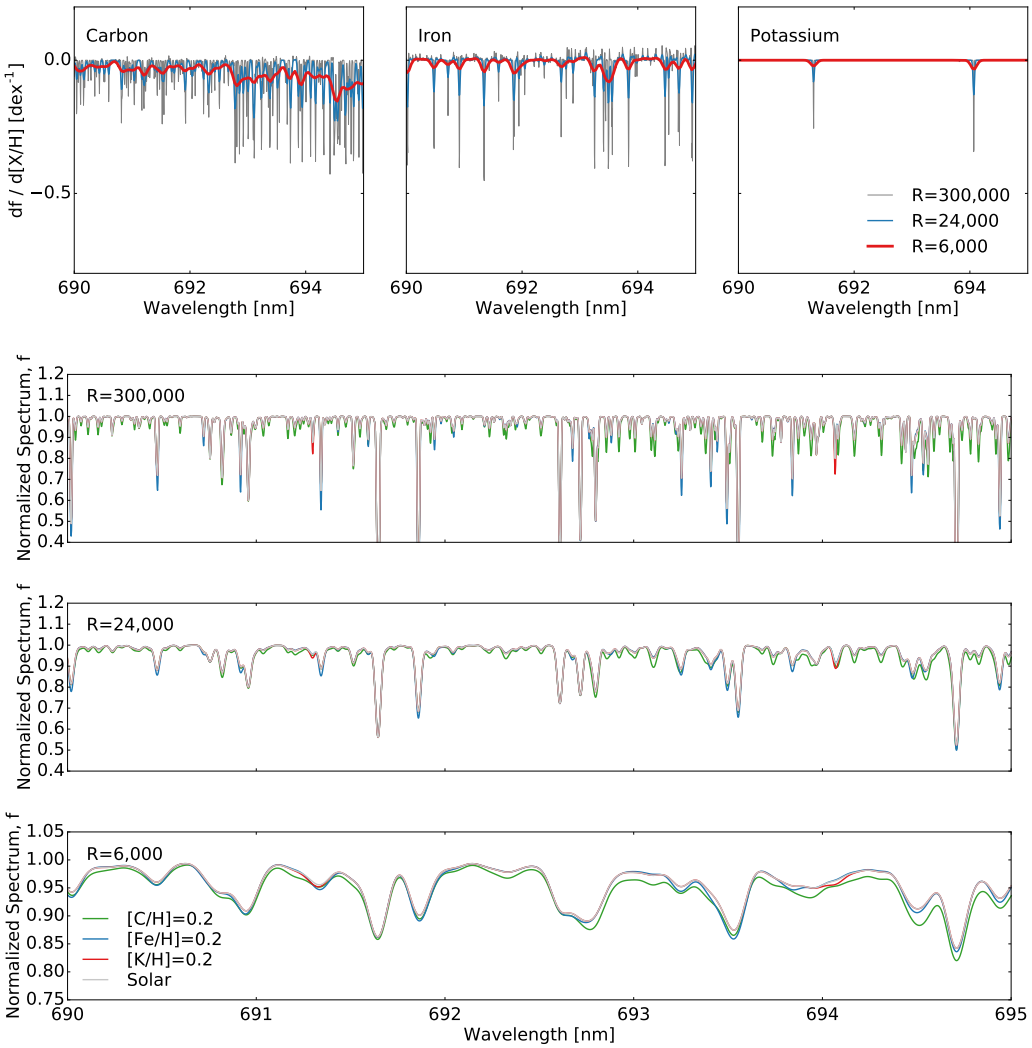

Figure 6.1: Illustration of the effects that changing abundances have on stellar spectra at different resolutions. The top panels show small wavelength segments of the gradient spectra of a solar metallicity, K-giant with respect to the abundances of C, Fe and K. We consider three different resolutions, $R = 300{,}000$, $24{,}000$ and $6{,}000$. In the top panels, we calculate the gradient spectra by evaluating the difference between the solar spectrum and the spectrum with $\Delta[X/\mathrm{H}] = 0.2$ in each of these elements. The three lower panels show these enhanced normalized spectra (with respect to a solar metallicity spectrum). In each of these three lower panels, we show different spectra that are individually enhanced in C, Fe and K. At $R = 300{,}000$ and $R = 24{,}000$, some of the spectral lines remain unblended, however at $R = 6{,}000$, all spectral lines are blended. Nonetheless, even at $R = 6{,}000$, spectra that are enhanced in different elements show distinct features. But in order to extract elemental abundances at $R = 6{,}000$, we must model the blended lines by fitting all stellar labels simultaneously.

ening. As we have discussed in Section 6.2.3, over-resolving lines does not improve the gradients. (b) When there are many overlapping lines, such as the carbon and iron lines, gradients do not degrade as much at low-resolution. One way to think of this is that overlapping/blended features have larger effective widths, so that convolution does not degrade the gradients in the same way as isolated lines. (c) Since we are convolving a spectral profile instead of a delta function, although the rms depth is proportional to R, the minimum point of the convolved profile alone does not necessarily scale exactly with R. The last effect has no influence on our arguments in Section 6.2.3, but the first two effects work in favor of low-resolution spectroscopy. They imply that the gradients only degrade linearly with R at certain restricted conditions. For example, the potassium lines at $R = 24{,}000$ and $R = 6{,}000$ are less affected by these two effects and show a close-to-linear gradient degradation. But going from $R = 300{,}000$ to $R = 24{,}000$, especially for the carbon and iron lines, the gradients do not degrade proportionally with R.

Beside studying how the results vary for different stellar types, we also consider different wavelength ranges in this study. We will assume the wavelength ranges of the APOGEE, GALAH, Gaia-ESO[§], 4MOST, Gaia RVS, RAVE, SEGUE and LAMOST surveys. Their wavelength ranges and spectral resolutions are listed in Table 6.1 (and are visualized in Appendix 6.7). Note that 4MOST plans to work at two configurations. The low-resolution configuration will operate on a larger wavelength range than the medium-resolution configuration. We will show in the following subsections that, regardless of the wavelength range and stellar type, low-resolution spectroscopy can measure

[§]We assume the GIRAFFE HR10 and HR21 settings, with which most sample of Gaia-ESO will be observed

equally many elements with the same precision as medium-resolution spectroscopy.

6.3.2 Stellar label estimates as a function of spectral resolution

In this section we evaluate how uncertainties of stellar labels vary as a function of spectral resolution, R, considering $T_{\rm eff}$, $\log g$, $v_{\rm turb}$ and all elements with atomic numbers from 3 to 99 as stellar labels. We calculate the theoretical uncertainties of these stellar labels using Eq. 6.1. The output covariance matrix K_{ij} has the size of 100 $\times$ 100, showing the covariances of all stellar labels. The diagonal entries of K_{ij} show variances that one can achieve for each stellar label, and the square roots of these values give the theoretical uncertainties that we will explore in this section. Clearly, the gradient spectra depends on stellar type, wavelength range and metallicity. Therefore, we calculate K_{ij} for different wavelength ranges, different stellar types and two metallicities – $[Z/{\rm H}] = 0$ and -2.

We also verified K_{ij} by numerical simulations. We modify a reference spectrum with linear combinations of gradient spectra from all stellar labels and noise up the spectrum. We perform full spectral fitting (using PSM) via χ^2-minimization and find that K_{ij} gives the exact estimate of the covariance matrix of stellar labels. Finally, to study how theoretical uncertainties vary with R, we convolve gradient spectra to various resolutions, and recalculate K_{ij} for each R. We define an element to be detectable if its uncertainty is better than 0.1 dex at $R = 24{,}000$ and ${\rm S/N} = 100$.

Fig. 6.2 show the uncertainties as a function of R of all detectable stellar labels (including $T_{\rm eff}$, $\log g$ and $v_{\rm turb}$). The figure overplots results from all stellar types, wavelength ranges and metallicities. We normalize the value of y-axis to be unity at $R = 6{,}000$. Since uncertainty scales linearly with S/N (see Eq. 6.1), the ratio of uncertainties plotted in Fig. 6.2 is independent of S/N. The left panel shows

Table 6.1: Wavelength ranges, survey resolutions and approximation resolutions adopted in this study of the various surveys

Survey	Wavelength range approximate (nm)	Survey resolution	Resolution adopted here	Survey status
GALAH	470–490, 565–585, 650–675, 760–790	28,000	24,000	On-going
APOGEE	1,500–1,700	22,500	24,000	On-going
Gaia-ESO	534–562, 848–900	20,000	24,000	On-going
4MOST (medium-res.)	390–435, 515–575, 605–675	24,000	24,000	Planned
4MOST (low-res.)	390–885	6,000	6,000	Planned
Gaia RVS	840–880	11,500	8,000	On-going
RAVE	840–880	7,000	8,000	Completed
SEGUE/BOSS	390–900	2,000	2,000	Completed
LAMOST	390–900	1,800	2,000	On-going

that, assuming the same S/N per pixel and the same wavelength range, the uncertainty degrades mostly linearly with $1/R$, regardless of stellar label, stellar type, wavelength range and metallicity, as explained in Section 6.2.3: since the absolute values of gradient spectra decrease proportionally to $1/R$, the uncertainties should also degrade linearly with $1/R$.

However, given the same exposure time, low-resolution spectra will have a S/N per pixel that is $1/\sqrt{R}$ higher than high-resolution spectra. In the middle panel, we take this into account and rescale the uncertainties in the left panel by $\sqrt{R/6,000}$.

In the right panel, we also account for the larger wavelength range afforded by low-resolution spectra (given a fixed total number of pixels and a fixed number of pixels per resolution element) and further scale the uncertainties by another factor of $\sqrt{R/6,000}$, assuming spectral information distributes uniformly over the entire wavelength range. As expected from the arguments in Section 6.2.3, this factor compensates for the lower-resolution. *Remarkably, regardless of stellar type, wavelength range and metallicity, the achievable stellar label uncertainties are indeed nearly independent of spectral resolution, if we have robust models and can fit all stellar labels simultaneously (Fig. 6.2, right panel).*

But Fig. 6.2 also quantifies how these simple trends are violated at both the very low-resolution and high-resolution ends. At the low-resolution end (e.g., $R \lesssim 1{,}000$), stellar label estimates become nearly degenerate, resulting in reduced precision, as we have discussed in Section 6.2.3. As for the high-resolution end, spectral lines are eventually resolved, so further increasing the resolution does not improve the information content. By visual inspection, we found that for our model grid, most spectral lines indeed have intrinsic broadening of the order of $R \simeq 10^{4.5}$. Over-resolving lines be-

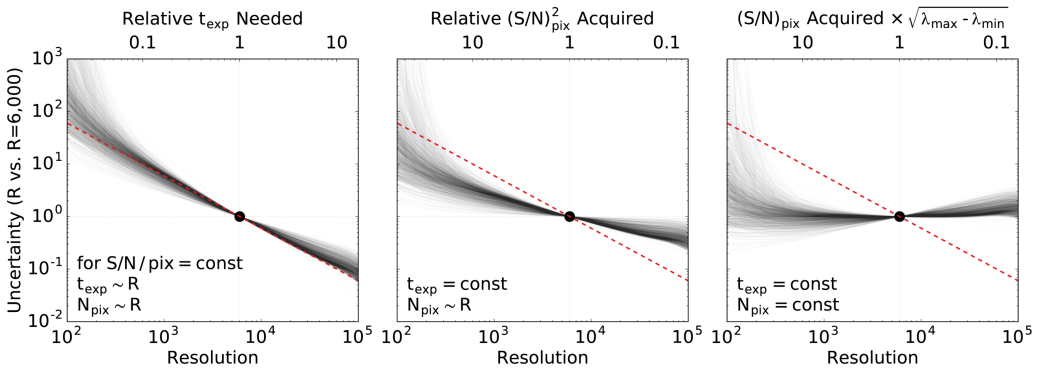

Figure 6.2: Uncertainties of stellar labels as a function of spectral resolution, relative to $R = 6{,}000$. We overplot results for all detectable stellar labels (stellar parameters and elemental abundances) from different stellar types, wavelength ranges and metallicities (see Section 6.3.2 for details) because the result is general and independent of these choices. The relative label uncertainties depend very much on the boundary conditions under which spectra of different resolutions are compared, as illustrated in the three panels. The left most one is a commonly used approach to such a comparison; the right most panel shows the comparison that is most pertinent to large spectroscopic surveys. Specifically, the left panel assumes that all resolution configurations have the same S/N per wavelength pixel (or resolution element). In this case, high-resolution spectra outperform low-resolution spectra, following a $1/R$ linear trend as depicted with the red dashed line, at the cost of significantly longer exposure times for higher resolution data. The middle panel assumes the same exposure time (and thus higher S/N for low-resolution spectra), and identical wavelength range (which would require R times more detector real estate for high-resolution spectra). The right panel assumes the same exposure time and the same number of detector pixels (low-resolution spectra thus have wider wavelength range); as spectral diagnostic information is contained throughout the near-UV to near-IR spectra of the most common stellar types, broad wavelength range is very important. The right panel shows that, given the same exposure time and detector pixels, going to higher resolution (beyond $R \gtrsim 1{,}000$) no longer improves the measurement uncertainties of stellar labels: the higher S/N per pixel, and the more extensive wavelength range compensate for lower spectral resolution, and vice versa. At very low-resolution with $R \lesssim 1{,}000$, stellar label estimates become degenerate and deviate from the linear trend.

yond this resolution does not improve the gradients and causes the high-resolution to deviate from the linear trend.

4MOST survey as a case study

One can adopt simple arguments to rescale the uncertainties in Fig. 6.2 for a particular survey design. For example, one can assume that spectral line information is uniformly distributed throughout all wavelengths and derive an $\sqrt{R}$ improvement in uncertainty when going from the middle to the right panel of Fig. 6.2. This assumption might be a good approximation for elements that have many spectral lines such as Fe, and α-capture elements. But for trace elements, such as Li, K, that have only a few spectral lines, expanding wavelength range does not necessarily improve the information content. To work out a concrete example, we compare the two proposed resolution configurations of the 4MOST survey. Here we consider, for the same exposure time and a larger wavelength range, the tradeoffs in the low- *vs.* medium-resolution setups for this particular survey.

4MOST proposes a medium-resolution configuration with a shorter wavelength range of $390 - 435$ nm, $515 - 575$ nm, $605 - 675$ nm and a low-resolution configuration with a wider wavelength range of $390 - 885$ nm. These two configurations serve as a perfect case study to evaluate how uncertainty of stellar label changes when comparing low S/N, shorter wavelength range and medium-resolution spectra to high S/N, larger wavelength range but low-resolution spectra. In Fig. 6.3, we assume $R = 8{,}000$ for the low-resolution configuration and $R = 20{,}000$ for the medium-resolution configuration. These resolutions are chosen such that both configurations consume an equal number of detector pixels when compensated with the difference in wavelength range. Also, for the

same exposure time, the low-resolution configuration will have a higher S/N – we assume the low-resolution configuration has a better S/N per pixel by a factor of $\sqrt{20,000/8,000}$.

Fig. 6.3 shows the ratio of uncertainties for all detectable elements and stellar parameters of the two configurations. Note that since we are plotting the ratio of uncertainties, the result is independent of the absolute values of S/N per pixel. In the x-axis, we sort elements by their uncertainties in the medium-resolution configuration. If the two scaling relations as assumed in Fig 6.2 are exact, in particular, spectral line information is uniformly distributed throughout all wavelengths – for e.g., information from stellar parameters: $T_{\rm eff}$, $\log g$ and $v_{\rm turb}$ – the ratio should be close to unity. However if the information is concentrated only in a small wavelength range, expanding wavelength range does not collect more spectral information, and in this case, the low-resolution configuration will have a worse uncertainty by a factor of $\sqrt{20,000/8,000}$. The upper dashed line shows this value as the upper limit. Fig. 6.3 shows a clear trend – stellar parameters and elemental abundances that have better uncertainties, such as Fe, Mg, generally have more lines, thus the uniform distribution of spectral information is a more valid approximation, resulting in ratios closer to unity. For elements that are less precisely measured, they are mostly elements that only have a small number of lines that reside in the wavelength region of the medium-resolution configuration. Thus expanding the wavelength range at low-resolution does not help in this case. Some elements (e.g., Na, Ba, K, Zn, S) are better measured (ratio $<$ 1) at low-resolution. Expanding the wavelength range includes more lines from these elements that would otherwise be missed by the medium-resolution configuration. The last four elements in Fig. 6.3 (Pt, Ge, Rb, In) highlight the scenario in which the medium-resolution configuration does not cover any transitions of these elements, and so these

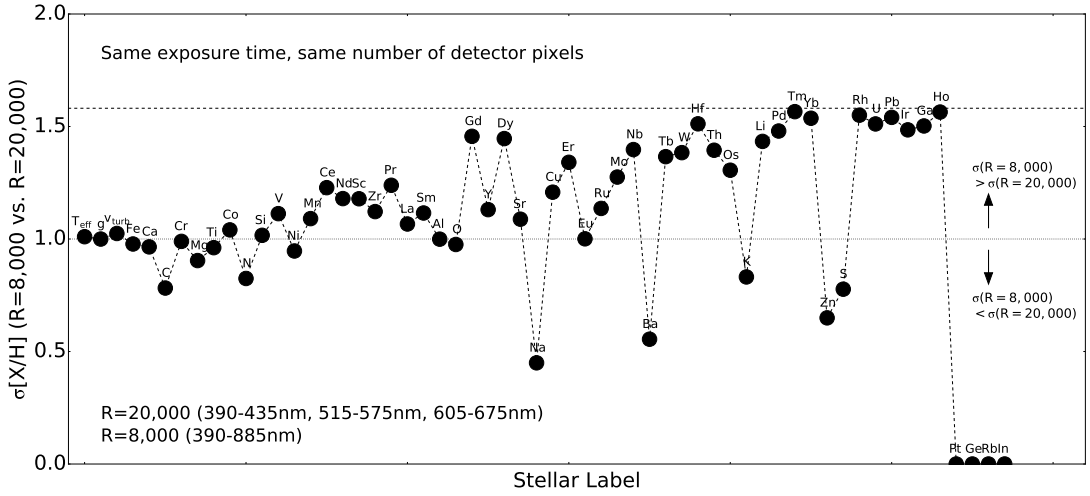

Figure 6.3: Comparison of stellar label uncertainties for the two different resolution configurations of the 4MOST survey, assuming the same exposure time and number of detector pixels. We sort elements along x-axis according to their uncertainties in the medium-resolution configuration; we have also included three stellar parameters ($T_{\rm eff}$, $\log g$, $v_{\rm turb}$). For elements with numerous lines, such as Fe, Mg, the lower resolution appears fully compensated by the expanded wavelength range: high and low-resolution configurations perform equally well. For trace elements with only a few lines there are two regimes: if the signal comes from few, or even just one line, high-resolution spectra perform better, by up to the theoretical factor of $\sqrt{20,000/8,000}$ (the upper dashed line). In sharp contrast, there are elements, where the wavelength range of the medium-resolution configuration misses the (only) diagnostic lines (Pt, Ge, Rb, In); obviously, the low-resolution configuration performs far better in that case.

elements are unmeasureable for this particular medium-resolution configuration.

Although some elements perform worse at low-resolution even with the same exposure time and number of detector pixels, note that the ratio of uncertainties is bounded by an upper limit of $\sqrt{20,000/8,000}$. We can compensate this loss if we spend 20,000 / 8,000 $=$ 2.5 times more exposure time with the low-resolution configuration. Since low-resolution spectrographs are generally more accessible and medium-resolution spectrographs have other downsides, such as lower instrumental throughput and more restrictive read noise limitations for faint targets, (see discussion in Section 6.5), it would still seem that low-resolution stellar spectroscopy with $R \simeq 6{,}000$ and properly chosen wavelength range is the optimal strategy to design large-scale stellar spectroscopic surveys.

6.3.3 Number of detectable elements for various surveys

In the previous section we discussed how the ratio of uncertainties vary as a function of spectral resolution. In this section, we will study the absolute uncertainties – i.e., the square root of diagonal entries of K_{ij} in Eq. 6.1 – given a fixed S/N per pixel, and determine how many elements we can, in principle, detect for various surveys. We assume a S/N per pixel of 100 in this section. We do not show the other values of S/N because the uncertainty scales linearly with S/N (cf. Eq. 6.1). We emphasize again that these uncertainties can only be attained if the model spectra are perfect, or nearly so.

Fig. 6.4 shows the theoretically achievable uncertainties of detectable elements and stellar parameters for various surveys, assuming solar metallicity and K-giants. For each survey setup, we assume

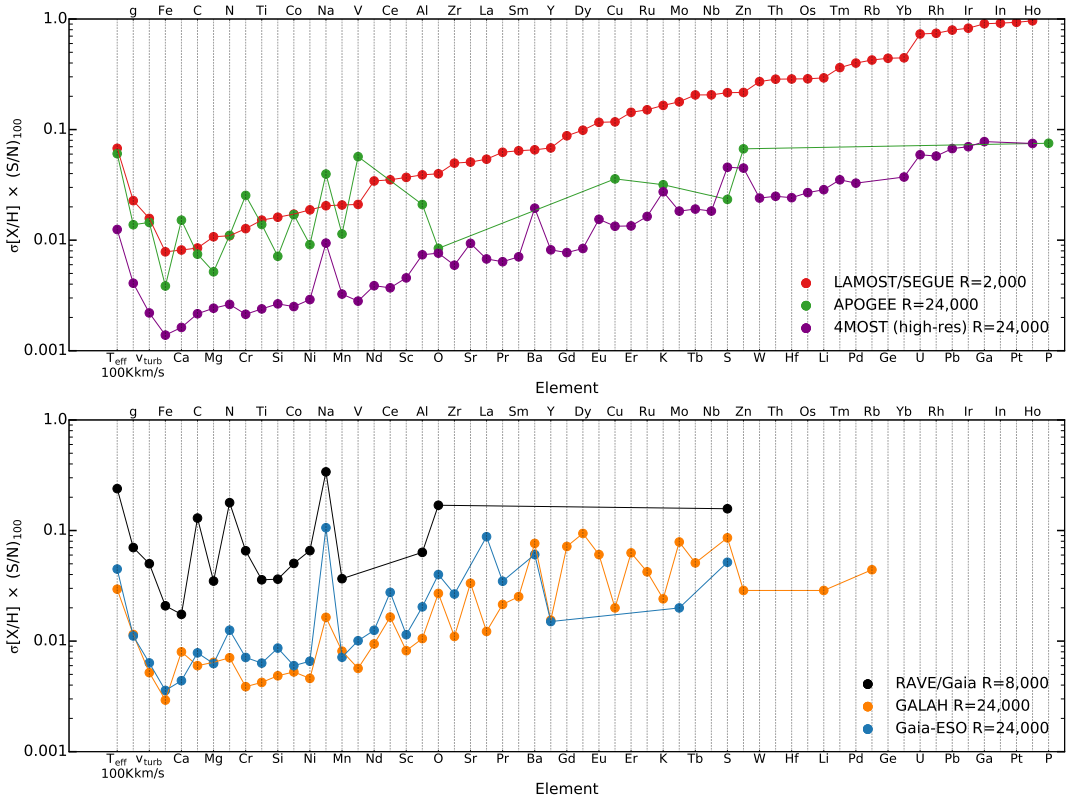

Figure 6.4: Theoretical (best-case scenario) uncertainty of stellar labels for various spectroscopic survey configurations. Stellar abundances are sorted according to their uncertainties in low-resolution spectra ($R \simeq 2{,}000$) that cover most of optical wavelength range (such as LAMOST, or SEGUE). For all surveys, we adopt the resolutions listed in Table 6.1, solar metallicity, K-giants and a S/N per pixel $= 100$. If the synthetic models are robust, the y-values show the minimal uncertainties (Cramer-Rao bound) that we can achieve for stellar parameters and elemental abundances when fitting all stellar labels simultaneously. For elements where there are no useful spectral diagnostics in a survey's wavelength range the (filled) symbols have been omitted. Optical surveys like 4MOST, GALAH, Gaia-ESO, but also SEGUE and LAMOST can detect up to $50 - 55$ elements and infrared surveys like APOGEE can detect up to 20 elements. More strikingly, if we have robust models, even with small wavelength ranges, RAVE and Gaia RVS can detect about 15 elements. The $(\mathrm{S/N})_{100}$ in the y-axis label is to remind that theoretical uncertainty scales linearly with S/N per pixel (cf. Eq. 6.1), therefore for other S/N values, it suffices to scale the uncertainties as shown in the y-axis accordingly.

the adopted resolutions as stated in Table 6.1. Optical surveys like 4MOST, GALAH, Gaia-ESO, SEGUE and LAMOST can measure up to $50 - 55$ elements. Strikingly, even for low-resolution spectra like SEGUE and LAMOST that has only $R \simeq 2{,}000$, in principle, we can still measure as many elements as medium-resolution spectra, provided that we can fit all stellar labels simultaneously and have robust stellar models.

Infrared surveys, such as APOGEE, contain less information (also see Appendix 6.7) and can "only" detect up to 20 elements, consistent with the APOGEE pipeline[92,181]. Not surprisingly, given the same resolution, surveys that have larger wavelength ranges such as 4MOST have smaller uncertainties than surveys that have more limited wavelength ranges like GALAH and Gaia-ESO. But interestingly, even for small wavelength ranges and low-resolution spectra from RAVE or Gaia, we can, in principle, detect about 15 elements. Measuring multi-elemental abundances with RAVE and Gaia RVS is an important application of PSM that we are currently exploring. Fig. 6.4 also suggests that high S/N spectra, such as stacked spectra from LAMOST and SEGUE, could detect many more elements than are currently being measured.

Instead of plotting uncertainties of individual elemental abundances, we can also compress this information and plot the cumulative distribution of uncertainties for all elemental abundances, as shown in Fig. 6.5. The y-axis of Fig. 6.5 shows the cumulative number of elements that we can detect that have smaller theoretical uncertainties than the threshold shown in the x-axis. We consider two resolution configurations – a medium-resolution configuration of $R = 24{,}000$ and a low-resolution configuration of $R = 6{,}000$. Each row in Fig. 6.5 has three separate panels, showcasing three different possible comparisons between the low-resolution and medium-resolution configurations, the

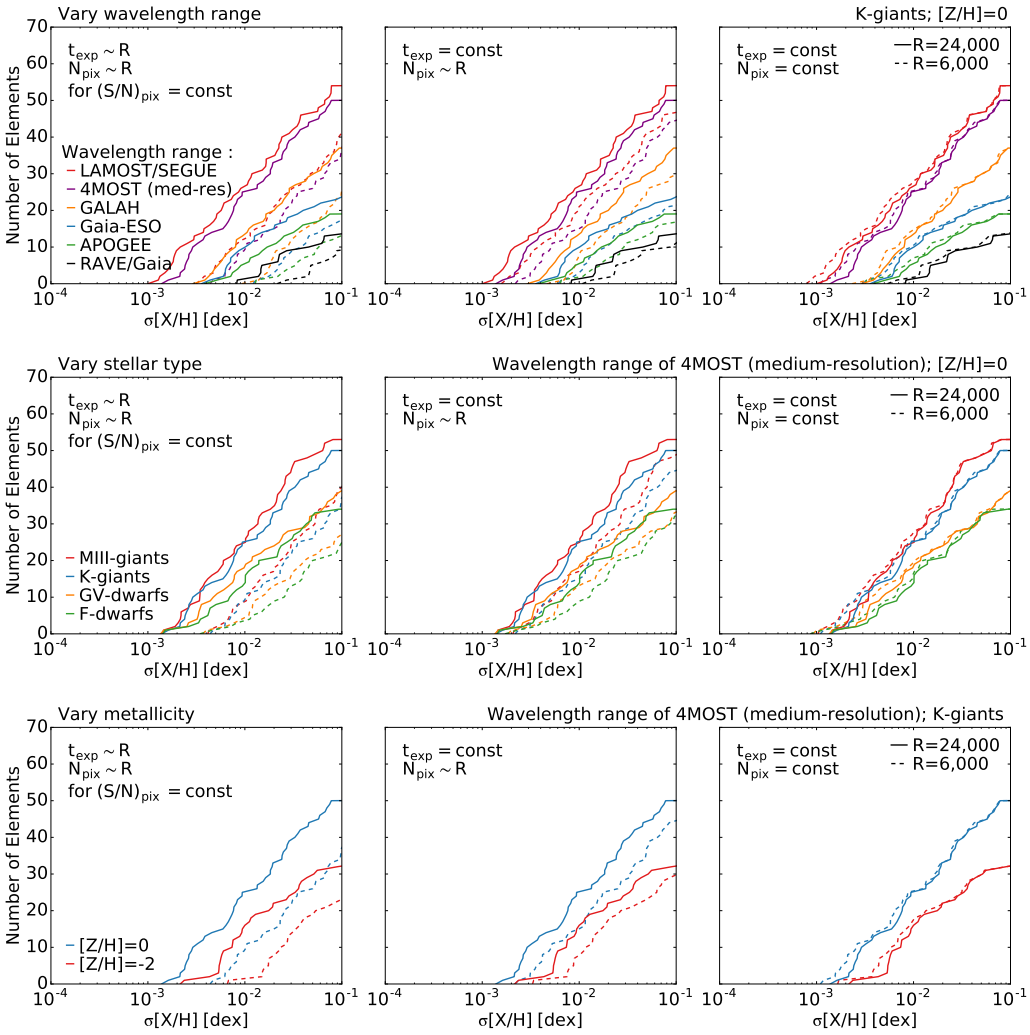

Figure 6.5: The number of different elements (y-axis) for which abundances can be obtained with a certain (theoretical) uncertainty (x-axis), as a function of stellar type, wavelength range and metallicity. The solid lines assume a medium-resolution configuration of $R = 24{,}000$ and the dashed lines assume a low-resolution configuration of $R = 6{,}000$. Panels from left to right illustrate three different comparisons between the medium-resolution configuration and the low-resolution configuration. We assume a fixed $\mathrm{S/N} = 100$ per pixel and a fixed wavelength range for the medium-resolution configuration, and vary the properties of the low-resolution configuration from left to right. The left panels assume the same S/N per wavelength pixel and the same wavelength range; the middle panels assume the same exposure time (higher S/N for the low-resolution configuration). The right panels further assume the same number of detector pixels (more extensive wavelength range for the low-resolution configuration), which we deem the most pertinent comparison. In that latter case, low-resolution spectra can detect as many elements as the medium-resolution spectra with similar precision. Panels from top-to-bottom assume three different comparisons of survey targets and survey configurations: the top panels compare different wavelength ranges from various surveys; the middle panels compare different stellar types; the bottom panels compare stars with different metallicities.

same way as Fig. 6.2. To recap, the panels on the left assume the same S/N per pixel and the same wavelength range. The middle panels assume the same exposure time and the right panels further assume the same number of detector pixels. For the middle panels and the right panels, we rescale the uncertainties of low-resolution spectra in the left panels by a factor of $\sqrt{6,000/24,000}$ and 6,000 / 24,000, respective, following Fig. 6.2.

The top panels of Fig. 6.5 show the number of detectable elements at $R = 24{,}000$ and $R = 6{,}000$ for various wavelength ranges, assuming solar metallicity and K-giants. For surveys that have these resolutions, such as 4MOST, GALAH and Gaia-ESO, these panels are just compact representations of Fig. 6.4. But we caution that for surveys that operate at a much lower resolution, such as LAMOST and SEGUE ($R \simeq 2{,}000$), results in the top panels might not be directly applicable – these panels only show the number of detectable elements if LAMOST and SEGUE were to operate in $R = 24{,}000$ and $R = 6{,}000$. Not surprisingly, at a given resolution and assuming the same S/N, the top panels show that a larger wavelength range, such as LAMOST and 4MOST, can detect more elements. These panels also show that, generally speaking, optical wavelength ranges contain more information and can measure more elements than the infrared. But more important, as shown in the right panel, if we assume the same exposure time and the same number of detector pixels, the dashed lines coincide with the solid lines, showing that $R = 6{,}000$ spectra can detect as many elements as the $R = 24{,}000$ spectra, echoing our earlier conclusions. This conclusion also holds true for the other comparisons that will we discuss next.

Thus far, we have only discussed how the detectability of elements vary as a function of wavelength range. But the detectability also depends on stellar type and metallicity. The middle panels

of Fig. 6.5 show the number of detectable elements for different stellar types, assuming a wavelength range of the 4MOST (medium-resolution) survey and solar metallicity. These panels show that cooler stars (e.g., M-giants) can detect more elements than hotter stars (e.g., F-dwarfs). This result is not surprising because cooler stars have more spectral lines, especially contributions from molecular lines. In fact. M-giants almost double the number of detectable elements compared to F-dwarfs. Since part of these cooler features come from molecular contributions and noting the composite nature of molecular features, this demonstrates the importance of full spectral fitting over many stellar labels simultaneously, without which we will not be able to extract information from molecular lines.

Finally, the bottom panels show the number of detectable elements in two different metallicity regimes, $[Z/\mathrm{H}] = 0$ and $[Z/\mathrm{H}] = -2$, assuming the wavelength range of 4MOST (medium-resolution) and K-giants. We calculate the K_{ij} matrix using gradient spectra with respect to reference points at these two different metallicities. Metal-poor stars have smaller gradient spectra which in turns predict a smaller number of detectable elements. Nonetheless, for optical surveys like 4MOST, although the number of elements is more restricted at the metal-poor regime, the bottom panels show that we can still detect up to 30 elements. Studies in Appendix 6.7 also indicate that there is still sufficient spectral information at low metallicity in the optical wavelength. But spectral information is more limited in the infrared, Although not shown, we found that we can only detect about 5 elements at $[Z/\mathrm{H}] = -2$ with an APOGEE-like setup.

6.3.4 Stellar parameter correlation as a function of spectroscopic resolution

Thus far we have only considered the diagonal entries of K_{ij} in Eq. 6.1 – i.e., the theoretical uncertainties of stellar labels. However, there is more information in K_{ij}. This matrix also infers the correlations of stellar labels. More precisely, for each label pair (i, j), the submatrix $\widetilde{K_{ij}}$ from the (i, j) rows and columns of K_{ij} shows the covariance of the i-th and j-th stellar labels, from which we calculate their correlation via

$$C_{ij} \equiv \widetilde{K_{ij}} / \sqrt{\widetilde{K_{ii}}\widetilde{K_{jj}}}. \tag{6.4}$$

Uncorrelated estimations of stellar labels are crucial for Galactic studies because correlated estimates could make astrophysical interpretations difficult when looking for trends among stellar labels or searching for structures in chemical space[192]. In this section we will study how the correlation C_{ij} varies as a function of spectral resolution.

The left panel of Fig. 6.6 shows the cumulative correlations from all detectable stellar labels. The y-axis indicates the fraction of label pairs that have correlations smaller than the threshold value in the x-axis. We only consider the global distribution of correlations from all (detectable) label pairs in this section and refer interested readers to Appendix 6.9 for correlations of each label pair. The left panel shows that many label pairs have large correlations at $R = 100$ – in fact, more than half of the label pairs have correlations larger than 0.4. Strong degeneracies of stellar labels are expected at $R = 100$. At this resolution, there are only 30 wavelength pixels in the APOGEE wavelength range, so most stellar labels are contributing to most of the pixels.

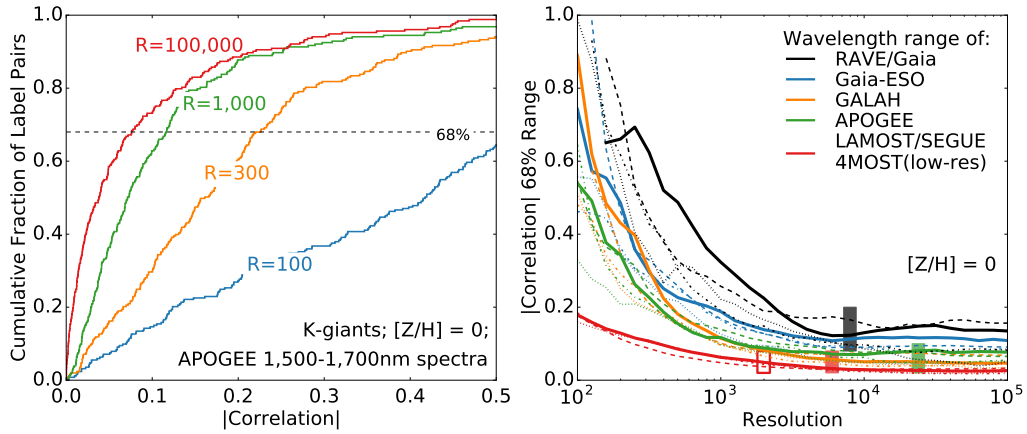

Figure 6.6: Statistics of the correlation between the label estimates, as a function of spectral resolution and wavelength range. The left panel shows the cumulative distribution of correlations among all (detectable) label pairs. We assume a wavelength range of the APOGEE survey, solar metallicity and K-giants. The lines in different colors show the correlations assuming various spectral resolutions. Going from $R = 100$ to $R = 1{,}000$ produces much more uncorrelated stellar label estimates yet, going from $R = 1{,}000$ to $R = 100{,}000$ barely reduces the correlations. The right panel shows how the typical level of label correlation (68% of pairs; see dashed line in the left panel) depends on the spectral resolution for the adopted wavelength range of different surveys (line colors). The right panel also illustrates the (weak) dependence of these correlations on spectral type: the solid, dashed, dashed-dotted and dotted lines assume stellar types of K-giants, M-giants, G-dwarfs, and F-dwarfs, respectively. Regardless of stellar type and wavelength range, the right panel shows that the label estimate correlations are generally modest, or even small, for $R \gtrsim 1{,}000$; however, $\log g$ and $T_{\rm eff}$ are crucial labels that remain substantially correlated, even at high-resolution. The black box shows the survey resolutions of RAVE and Gaia; the green box indicates the resolutions of APOGEE, GALAH, Gaia-ESO and 4MOST (medium-resolution); the shaded red box shows the resolution of 4MOST at low-resolution, and the hollow red box shows the survey resolutions of LAMOST and SEGUE. If one can fit all stellar labels simultaneously for these surveys, then most abundance estimates will not be seriously correlated, even though the spectral lines are blended.

However, increasing resolution to $R = 1{,}000$ already removes or strongly diminishes the correlations between labels. About 80% of the label pairs have correlations smaller than 0.15. In detail, as shown in Appendix 6.9, only prominent stellar labels that contribute to most pixels, namely $T_{\rm eff}$, $\log g$, $v_{\rm turb}$, Fe, C, N, O are strongly correlated beyond $R = 1{,}000$. The green and red lines in the left panel shows that going to a even higher resolution, such as $R = 100{,}000$, no longer decreases the correlations by much. In practice, the correlations at low-resolution should be even smaller compared to high-resolution. Fig. 6.6 assumes a fixed wavelength range, but as we have discussed earlier, given the same number of detector pixels, low-resolution spectra will have a much more extensive wavelength range, which allows further disentanglement of different contributions from various stellar labels.

The right panel in Fig. 6.6 shows that this result is general and is independent of stellar type and wavelength range. Instead of plotting the cumulative distributions as in the left panel, the right panel plots the correlation values corresponding to the 68% percentile of the cumulative distributions as a function of spectral resolution. The solid, dashed, dashed-dotted and dotted lines assume different stellar types – K-giants, M-giants, G-dwarfs, and F-dwarfs, respectively, and the lines in different colors show results from various wavelength ranges. All these lines concur with the previous conclusion that stellar labels are not strongly correlated beyond $R \simeq 1{,}000$, with the exceptions of the RAVE and Gaia RVS wavelength ranges. RAVE's and Gaia RVS's labels are only not strongly correlated beyond $R \sim 6{,}000$ because RAVE and Gaia RVS have a limited wavelength range ($\lambda = 840 - 880$ nm). With this limited wavelength range, below $R \sim 6{,}000$, there are too few wavelength pixels to distinguish contributions from various stellar labels.

The lines in the right panel show correlations at various resolutions, but for a spectroscopic survey, there is a well-defined survey resolution. An important question then is, are stellar labels correlated at the nominal survey resolutions? To answer this question, we label the survey resolutions with boxes in the right panel of Fig. 6.6. The black box shows the resolutions of RAVE and Gaia RVS. The green box shows the resolutions of APOGEE, GALAH, Gaia-ESO and 4MOST (medium-resolution). The shaded red box shows the resolution of 4MOST in the low-resolution configuration; and the hollow red box shows the resolutions of LAMOST and SEGUE. All boxes are in the region where the correlation curves have already plateaued, indicating that stellar labels will not be strongly correlated from these surveys.

6.4 Fitting and deriving 18 stellar labels with $R = 6{,}000$ spectra

Thus far we have shown that, given the same exposure time and the same number of detector pixels, spectral information remains largely independent of resolution. But there is still one crucial question yet to be answered, since most spectral lines at low-resolution are blended, can we model these blended features by fitting all stellar labels simultaneously? In other words, even though we know the spectral information is there, can we extract it? In this section we will generate and fit mock spectra at $R = 6{,}000$ and show that we can recover multi-elemental abundances at this resolution, even in the presence of bad pixels, imperfections of LSF modeling, and flux uncertainties.

We choose to study the wavelength range of APOGEE ($\lambda = 1{,}500$–$1{,}700$ nm) as our test case. We generate flux-normalized synthetic models at $R = 300{,}000$ and subsequently convolve them to $R = 6{,}000$ with a Gaussian kernel. We follow APOGEE DR12/DR13 and assume a wavelength sampling of $\lambda/3R$. With this sampling there are $\sim$ 1800 wavelength pixels at $R = 6{,}000$. Here we only study flux-normalized models and will discuss the potential problem with continuum normalization at low-resolution in Section 6.5.2.

We adopt the same PSM approach as in Rix et al.[172] and perform full spectral fitting. Here we briefly summarize the idea of PSM. Instead of interpolating spectra, PSM constrains explicit quadratic functions that define how flux varies as a function of stellar labels at each wavelength. One can regard PSM as a second order Taylor expansion of a spectrum. How well PSM performs depends on the "convergence radii" of the Taylor sphere. The key to the success of PSM (and related data-driven models such as the Cannon, cf.[40,153]) is that the Taylor sphere encompasses most of the stellar label

space that matters – i.e., the region of stellar label space where stars typically occupy. Previously in Rix et al.[172], we tested that we can fit all 18 stellar labels ($T_{\rm eff}$, $\log g$, $v_{\rm turb}$ and 15 elements) simultaneously and recover these stellar labels at $R = 24{,}000$. Our aim here is to extend that analysis to $R = 6{,}000$.

In order to test how well we can recover realistic stellar labels, we consider labels from the APOGEE DR12 catalog[92], restricting to objects with $4{,}000\,\mathrm{K} < T_{\rm eff} < 5{,}500\,\mathrm{K}$ and $1 < \log g < 4$. We tested that a single PSM region performs sufficiently well within this $T_{\rm eff} - \log g$ range. Going beyond this range might require multiple PSM spheres to cover the full relevant label space (see[193]), but will not alter our conclusions. We also remove objects that have not measured abundances for all elements in the APOGEE DR12 catalog. We randomly choose 1,000 stars and use their stellar labels to generate our training set and constrain the PSM functions. We randomly choose another 1,000 stars to generate our testing set. The testing set is used to determined how well we can recover their input parameters. Note that to fully define a PSM for 18 stellar labels, we only need a minimal training set of $(18 \times 19)/2 = 171$ spectra. But Rix et al.[172] found that overconstraining PSM with more training set, whenever it is still computational feasible, produces a better result. Therefore, we choose to constrain the PSM with 1,000 training models.

Fig. 6.7 shows the recovery of input parameters for the testing spectra. Gaussian random errors are included assuming a S/N per pixel of 200 and 500. We also assign random values to 10% of the testing spectra pixels and assign large uncertainties to these "bad pixels." These mimic pixels affected by skylines, cosmic rays or other possible contaminants. They also mimic pixels that are not well-modeled in real life applications and have to be subsequently clipped from spectral fitting. The gray

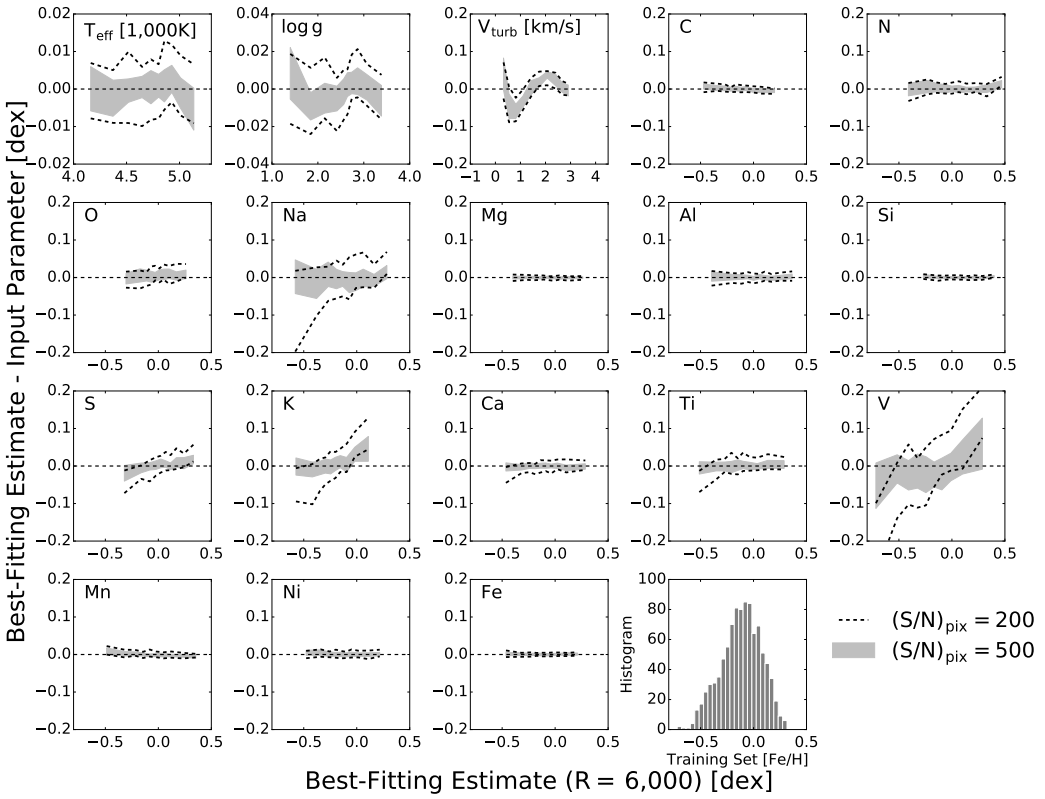

Figure 6.7: Stellar label precision resulting from the simultaneous fitting of all labels with a PSM model[172] to spectra of APOGEE's wavelength range, but at a resolution of only $R = 6{,}000$. We assume synthetic model spectra and adopt stellar labels from the APOGEE DR12 catalog. We consider stellar labels that have $4{,}000\mathrm{K} < T_{\rm eff} < 5{,}500\,\mathrm{K}$ and $1 < \log g < 4$, from which we generate 1,000 models to construct the PSM. The PSM is then used to fit another 1,000 testing models that have similar APOGEE stellar labels. Each panel shows the differences between the most likely PSM label estimates and the input labels. To mimic actual complications in spectral analyses, we also assume 10% of the testing pixels to have large uncertainties, whose values are not used in spectral fitting. The gray band shows the 1σ range assuming $\mathrm{S/N} = 500$ per wavelength pixel, and the dashed lines assume $\mathrm{S/N} = 200$ per wavelength pixel. Even with noised-up spectra and 10% of bad pixels, almost all elemental abundances are recovered better than $0.01 - 0.05$ dex from APOGEE-like spectra at $R = 6{,}000$. However, we found that systematics in the PSM label estimates can become important at lower S/N; this is for elements that have limited number of lines or only very weak signatures in the wavelength range, such as K, V or Na. The last panel shows the [Fe/H] distribution of the training spectra used in the construction of the PSM. Note that the $T_{\rm eff}$ and $v_{\rm turb}$ subplots assume different units than shown in the y-axis.

shaded region in each panel illustrates the 1σ range and demonstrates that, with robust models, we can recover 18 stellar labels with a precision of 0.01−0.05 dex at $R = 6{,}000$. The recovery of $v_{\rm turb}$ has a non-monotonic systematic. This suggests that the PSM model is not a perfect representation of the variation of flux as a function of $v_{\rm turb}$. A more complicated function might be needed to describe the variation, but even with a simple quadratic function, the systematic is small (< 0.1 km/s).

Another potential source of systematic uncertainty is our imperfect knowledge of the line spread function (LSF). To study how sensitive our results are with imperfect adopted broadening kernel, we model mock spectra that are further convolved with an additional broadening of $0.5 - 50$ km/s. Fig. 6.8 shows the scatter between the best-fit and input stellar label as a function of additional broadening. As before, we also assume 10% bad pixels and adopt S/N per pixel of 200 and 500. The figure shows that, at $R = 6{,}000$, the estimates are not severely affected by an additional broadening < 10 km/s. But if the LSF errors are larger than 10 km/s, the PSM estimates are biased. As spectral features become shallower and broader with additional broadening, we will overestimate $T_{\rm eff}$ and $\log g$, which in turn generally causes overestimations of $[X/{\rm H}]$ to compensate for the higher temperature. In contrast, although not shown, we checked that bad pixels (with large uncertainties assigned) and flux uncertainties, as studied in Fig. 6.7 and Fig. 6.9, do not bias the stellar label estimates at high S/N. On the flip side, the weak dependence with additional broadening shows that we cannot measure $v_{\rm macro} < 10$ km/s at $R = 6{,}000$ because $v_{\rm macro}$ broadening is completely dominated by the instrumental LSF broadening at low-resolution. On the other hand, $\log g$ and $v_{\rm turb}$ can be recovered at low-resolution because their broadening effects are not simple convolutions with a kernel and so can be distinguished from the broadening due to the LSF.

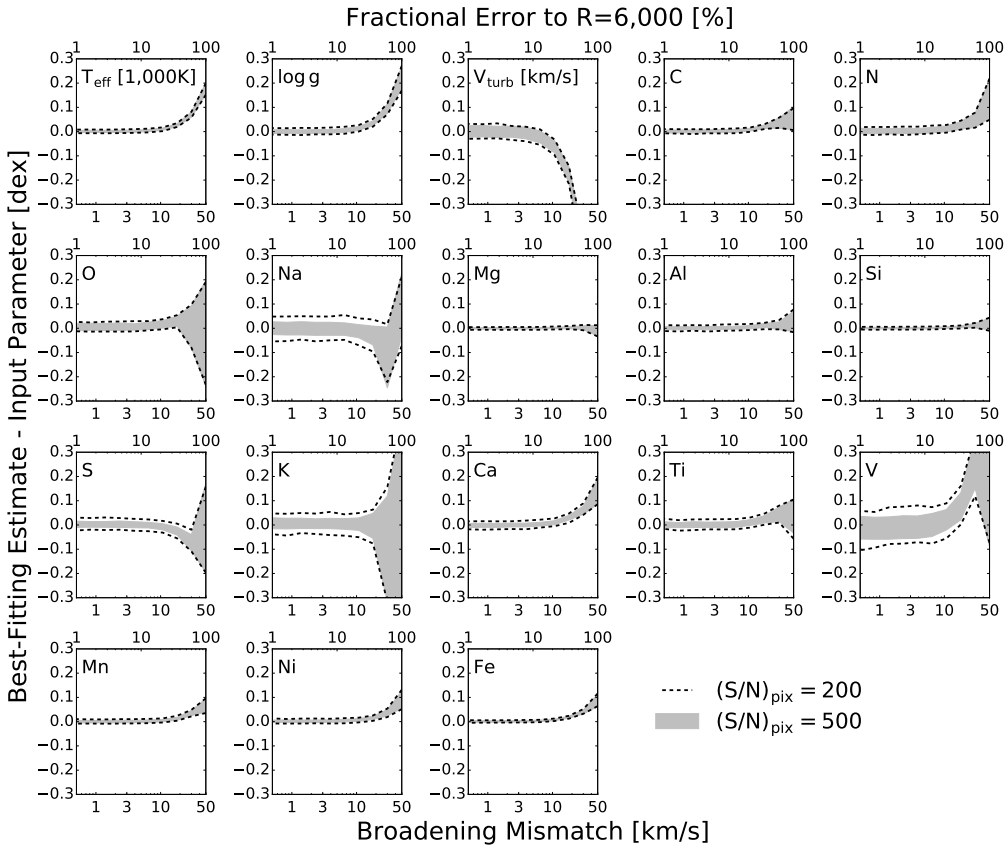

Figure 6.8: Sensitivity of low-resolution label estimates to mismatches in the assumed spectral line-spread function (LSF). The plot layout is the same as Fig. 6.7, but we study the deviation (averaging over the full range of the label) as a function of additional (and erroneous) LSF broadening in the PSM fitting. We also assume 10% of bad pixels and S/N of 200 and 500 per wavelength pixel. The figure shows that the label estimates are insensitive to even substantive differences between adopted and true LSF. An additional of broadening of $< 10\,\mathrm{km/s}$ has negligible effects on the estimates. In turn, this implies that low-resolution spectroscopy cannot recover v_{macro}, modeled as a Gaussian convolution of the spectrum, for stars to the level of $\sim 10\,\mathrm{km/s}$. But with a more severe additional broadening than $10\,\mathrm{km/s}$, as the lines become shallower and broader we will overestimate T_{eff} and $\log g$, which in turn generally causes overestimations of $[X/\mathrm{H}]$ to compensate for the higher T_{eff}.

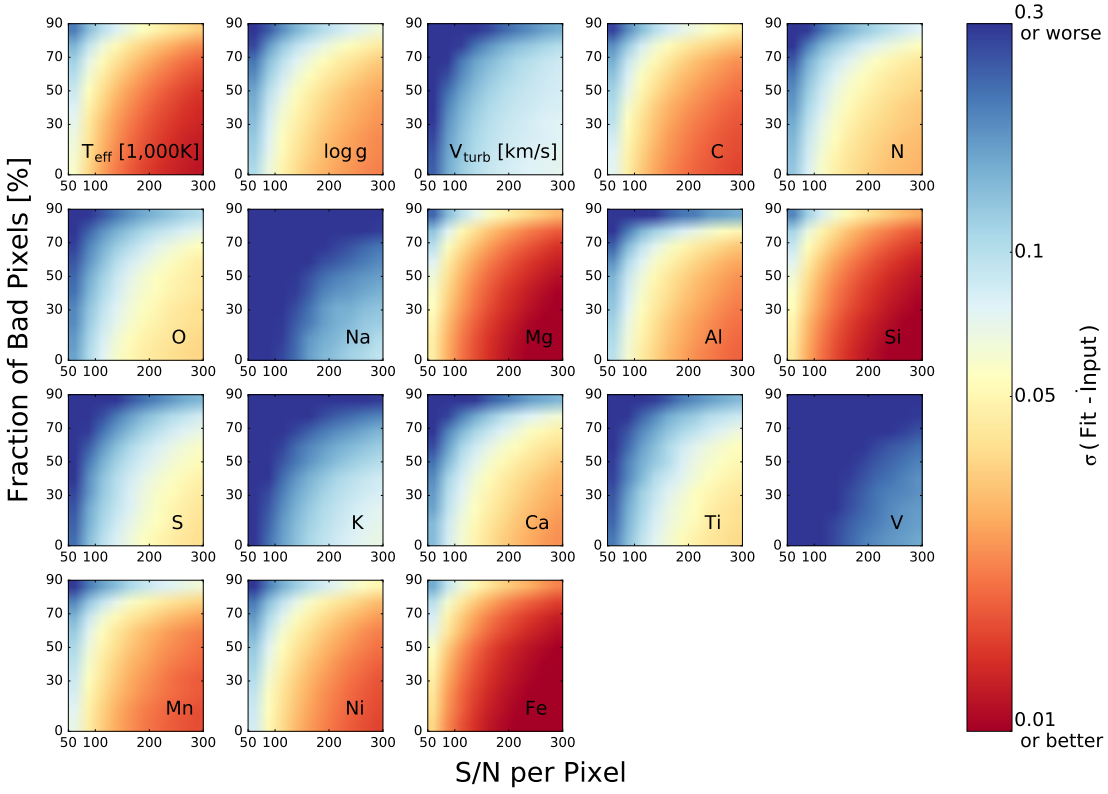

Figure 6.9: Variance of different stellar label estimates obtained through PSM of APOGEE-like spectra, but presumed to be at $R = 6{,}000$. The variances are calculated as a function of S/N per pixel and fraction of "bad pixels." At S/N per pixel $>$ 100, we can recover most stellar labels better than 0.1 dex. For stellar labels that have many spectral features across all wavelengths, such as Fe, Mg and Si, their recoveries depend less on the fraction of bad pixels. For stellar labels that have only weak gradients, such as $v_{\rm turb}$, V, Na, K, their recoveries are more compromised at $R = 6{,}000$.

Fig. 6.9 shows the scatter in the stellar label recovery as a function of S/N and the fraction of assumed bad pixels. With S/N per pixel > 100, PSM can recover most stellar labels better than 0.1 dex, even with $\sim 50\%$ of bad pixels. The weak dependence with the fraction of bad pixels is not surprising – since information only adds in quadrature, having a single reliable line can carry a lot of weight. Thus, for elements that have many spectral lines, such as Fe, Mg and Si, a high fraction of bad pixels does not substantially change the results. On the other hand, the measurement of stellar labels that only have weak gradients, such as $v_{\rm turb}$, V, Na, K, at low-resolution, can be seriously compromised by a large fraction of bad pixels and/or large flux uncertainties.

6.5 Discussion

In this paper we have demonstrated that it is possible to derive precise many elemental abundances with low-resolution spectra if one is not limited by systematic shortcomings of spectral models. Perhaps more remarkably we show that – at given exposure time and number of detector pixels – low-resolution spectra can yield elemental abundances as precise as high-resolution spectra, and without strong correlations among stellar labels. In this section we discuss several important caveats to these conclusions and additional complications, both for high and low-resolution analyses.

6.5.1 Some drawbacks to high-resolution spectroscopy

Practicalities aside, if there is little or no gain in label precision at a given survey speed between resolutions $R \sim 1{,}000$ and $R \sim 100{,}000$, one might then wonder what, if any, downsides exist to pursuing a survey at the upper end of this range? First, as discussed in previous sections, the general independence of elemental abundance precision on resolution assumes that information is spread uniformly throughout the spectrum. This is generally not the case, especially for important classes of elements such as r- and s-process neutron-capture elements. Because of this fact, there is some minimum wavelength range that is necessary to cover in order to probe a given set of elements. This fact would tend to work against collecting high-resolution data as multiple instrument configurations are required and hence the number of detector pixels required is not fixed but increases with resolution.

Spectrographs with very high spectral resolution are not suited for even moderately faint objects:

they tend to have lower throughput than low-resolution spectrographs, and the exposure time to overcome the read-noise dominated regime for faint objects is often prohibitive.

6.5.2 Limitations of low-resolution stellar spectroscopy

The fundamental limitation in analyzing low-resolution spectra is the reliance on the models being of high quality, e.g., without significant shortcomings in the line list. At high-resolution one can focus on lines with very accurate atomic data and that are known to form in relatively well-understood layers of the atmosphere. At low-resolution every "feature" is in reality a blend of many lines and so it is difficult to isolate the good from the bad regions of the model spectra. However, it is worth emphasizing that a single (or several) bad pixels on their own will not necessarily compromise the fits at low-resolution. In most cases there is a vast amount of redundant information in the spectrum, and so even if for example one or more iron lines are in error, the many other good iron lines will dominate the determination of the final iron abundance. Also, data driven approaches such as *The Cannon*[153] can construct spectral models for low-resolution spectra (based on accurate and precise training labels, e.g., from high-resolution spectra) that are – almost by construction – without substantive systematic errors.

There are a variety of ways that one can mitigate the effects of model imperfections when fitting low-resolution spectra. At a minimum, one can (and should) fit ultra high-resolution spectral atlases of standard stars such as the Sun and Arcturus. The residuals in the fits to the standards can be convolved to low-resolution and used to down-weigh spectral regions that are poorly described by the models. A more ambitious approach would be to collect a sample of ultra high-resolution

spectra and tune the models to fit those data (e.g., by astrophysically calibrating the atomic line parameters, which are often not known to high precision). Ideally the sample for which ultra high-resolution spectra are available should span the full range of parameter space that one is interested in studying at low-resolution. These tuned models should then by design provide excellent fits to low-resolution data.

There are other important aspects of fitting low-resolution spectra that are related to the data quality and characteristics. In principle, with perfectly flux calibrated data, one could choose to fit the fluxed spectrum directly. In practice spectra are often not flux calibrated to the precision required and so some methods for continuum normalization are adopted. At high-resolution one can either measure equivalent widths or fit polynomials to regions of the spectrum that are free of (strong) absorption lines. At low-resolution there are no wavelength ranges that probe only the continuum, and so the method of normalization is more model-dependent. Another advantage of high-resolution data is that it is much easier to subtract bright sky lines, which are intrinsically very narrow. This is more of a concern in the NIR where there is a forest of bright sky lines. Yet another advantage of working at high-resolution is that equivalent widths are independent of the LSF and so the precise wavelength-dependent instrumental resolution need not be modeled. At low-resolution the LSF must be accurately modeled in order to derive reliable parameters. In practice this means that a parameterized LSF should become part of the model.

Subtle effects even at high-resolution, such as asymmetric line profiles, due for example to 3D effects or spot modulation, and small shifts in line centers, due for example to isotopic ratio effects, Zeeman splitting, gravitational redshifting, or convective motions, are just several examples of effects

that are unlikely to be detectable, even with perfect models, at low-resolution (detection of these effects with perfect models would also require perfect knowledge of the wavelength solution and LSF). Finally, although not directly related to deriving stellar parameters, another obvious advantage of high-resolution is precision radial velocity measurements, which have been instrumental to studying exoplanet populations.

6.6 Summary and conclusions

Large spectroscopic surveys such as APOGEE, GALAH and Gaia-ESO are now collecting several orders of magnitude more stellar spectra in the Milky Way than all previous surveys combined. But the key to unravel the evolution of the Milky Way depends on how well we can turn stellar spectra into stellar labels – stellar parameters such as $T_{\rm eff}$, $\log g$, $v_{\rm turb}$ and many elemental abundances. At resolutions below $R \lesssim 20{,}000$ most spectral lines are blended. Deriving reliable stellar parameters therefore requires simultaneously fitting dozens of stellar labels in order to model the blended features. Fitting dozens of stellar labels simultaneously has only recently been demonstrated to be possible with the aid of polynomial spectral models (PSM). In light of this new technique, in this paper we explored how the information content of stellar spectra varies as a function of resolution and explored the possibility of deriving multi-elemental abundances from low-resolution spectra. Our findings are summarized below:

- We explore the information content in spectra covering 300–2,400 nm, considering different wavelength ranges of past, on-going, and future spectroscopic surveys – APOGEE, GALAH, Gaia-ESO, 4MOST, Gaia RVS, RAVE, SEGUE and LAMOST – and different stellar types, from M-giants to F-dwarfs. Assuming that the underlying models (whether *ab initio* or data-driven) are without systematic errors, we find that optical surveys can measure $50 - 55$ elements, and infrared survey can measure about 20 elements, even with low-resolution, $R \simeq 6{,}000$, high S/N spectra. Even smaller wavelength ranges associated with the RAVE and Gaia RVS surveys can potentially measure up to 15 elements at high S/N.
- Assuming the same exposure time per star and same number of detector pixels (e.g., $R \cdot (\lambda_{\rm max} - \lambda_{\rm min}) = $ constant, and a constant number of pixels per resolution element), the derived uncertainties on stellar labels are essentially independent of resolution for $1{,}000 \lesssim R \lesssim 100{,}000$.

- Even though spectral lines are blended at low-resolution, most stellar labels are not correlated at $R \gtrsim 1{,}000$. This holds generically for elements that produce detectable features at more than one location in the observed spectrum.

- We demonstrate that it is possible to recover 18 labels from low-resolution $R = 6{,}000$ APOGEE-like model spectra, even in the presence of a significant fraction of bad pixels, imperfections in modeling the LSF, and realistic observational uncertainties.

- Deriving precise many elemental abundances from low-resolution spectra could open up new windows for Galactic archeology, and in particular, chemical tagging because the latter requires a vast sample size, which is generally more challenging to obtain at medium- or high-resolution. We suggest that, in order to optimize scientific returns, a strategy for future spectroscopic surveys would be to collect a small number of high-resolution ($R \simeq 100{,}000$) spectra for model calibration purposes but to carry out the main survey at much lower resolution.

6.7 Appendix: Information content of stellar spectra

In this section, we explore the total spectral information content as a function of wavelength by adopting the idea of gradient spectra, i.e., how much a spectrum changes as we vary elemental abundances. We calculate gradient spectra for elements with atomic numbers from 3 to 99 (Li to Es), from $\lambda = 300 - 2{,}400$ nm, at $R = 300{,}000$, and $\Delta[X/\mathrm{H}] = 0.2$. For the purpose of illustration, the gradient spectra are subsequently boxcar-smoothed with a bin size of 10 nm. Despite exploring an exhaustive list of elements, we find many elements to have zero gradient spectra because there are no significant atomic lines for these elements. We compare the information content at two different metallicities – $[Z/\mathrm{H}] = 0$ and $[Z/\mathrm{H}] = -2$, and four different stellar types – M-giants, K-giants, G-dwarfs, F-dwarfs.

Fig. 6.10 shows the sum of gradient spectra from all elements, illustrating the total spectral information. Since the resolution element is proportional to λ/R, a bluer wavelength has a smaller resolution element – in other words, we sample more wavelength pixels at bluer wavelengths. Taking that into account, we further divide the sum of gradients by the wavelength in Fig. 6.10. Therefore, the y-axis has an unit of $\mathrm{dex}^{-1}\mathrm{nm}^{-1}$. But it is the relative amount of information that matters, the absolute scale of the y-axis is not important. We note that the total information does not directly infer the number of detectable elements. For example, molecules such as TiO and CN can have an enormous amount of spectral lines, but yet there are not many elements involved. To overcome this shortcoming, Fig. 6.11 provides another view of the information content (also see [22] for a similar analysis). We separate the wavelength range into portions of 10 nm, and evaluate how many

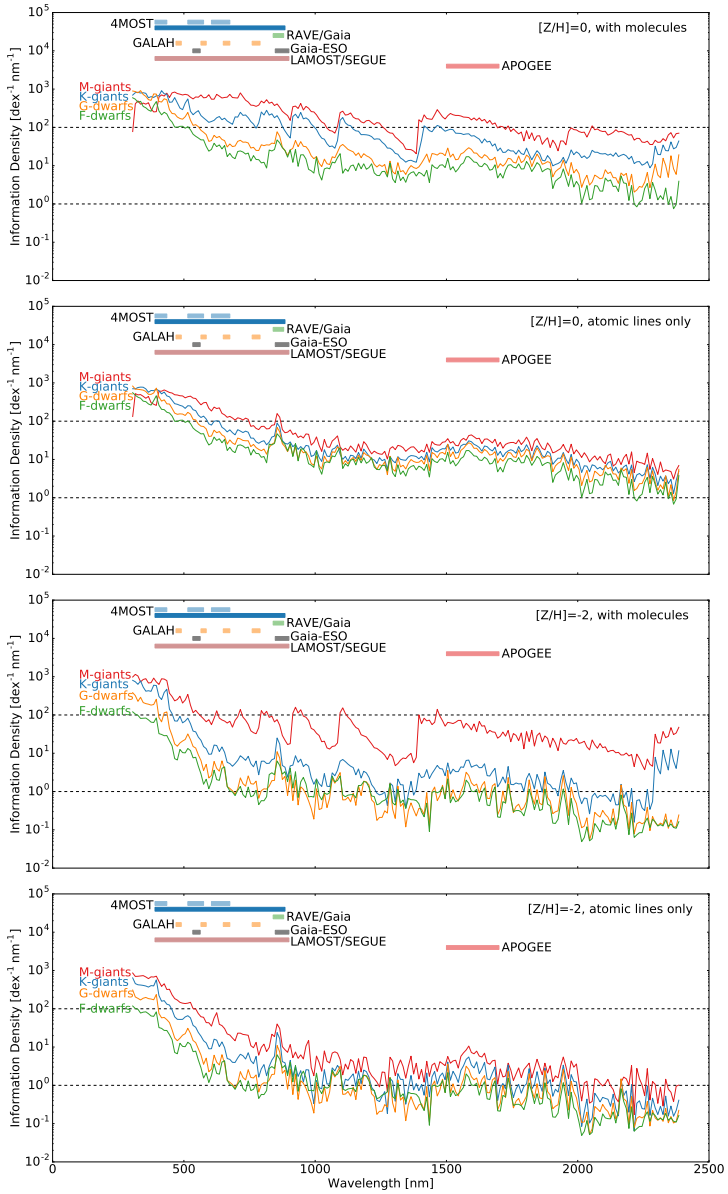

Figure 6.10: Spectral information for all elemental abundances, as a function of wavelength, spectral type (line color) and metallicity (top *vs.* bottom panels): each line shows the sum of all gradient spectra from elements with atomic numbers from 3 to 99 and take into account that bluer wavelengths have smaller resolution elements. The colored horizontal bars show the wavelength ranges of various large spectroscopic surveys. Lines in different colors illustrate different stellar types – from M-giants to F-dwarfs. The top two panels show the information content of $[Z/\mathrm{H}] = 0$, and the lower two panels assume $[Z/\mathrm{H}] = -2$. Within each of two panels at $[Z/\mathrm{H}] = 0$ and $[Z/\mathrm{H}] = -2$, the lower one excludes molecular lines in the model. This quantifies that the information content about many elements increases towards shorter wavelengths and cooler spectral types. Also, molecular features, often omitted from analyses because of their complexity, have large information content.

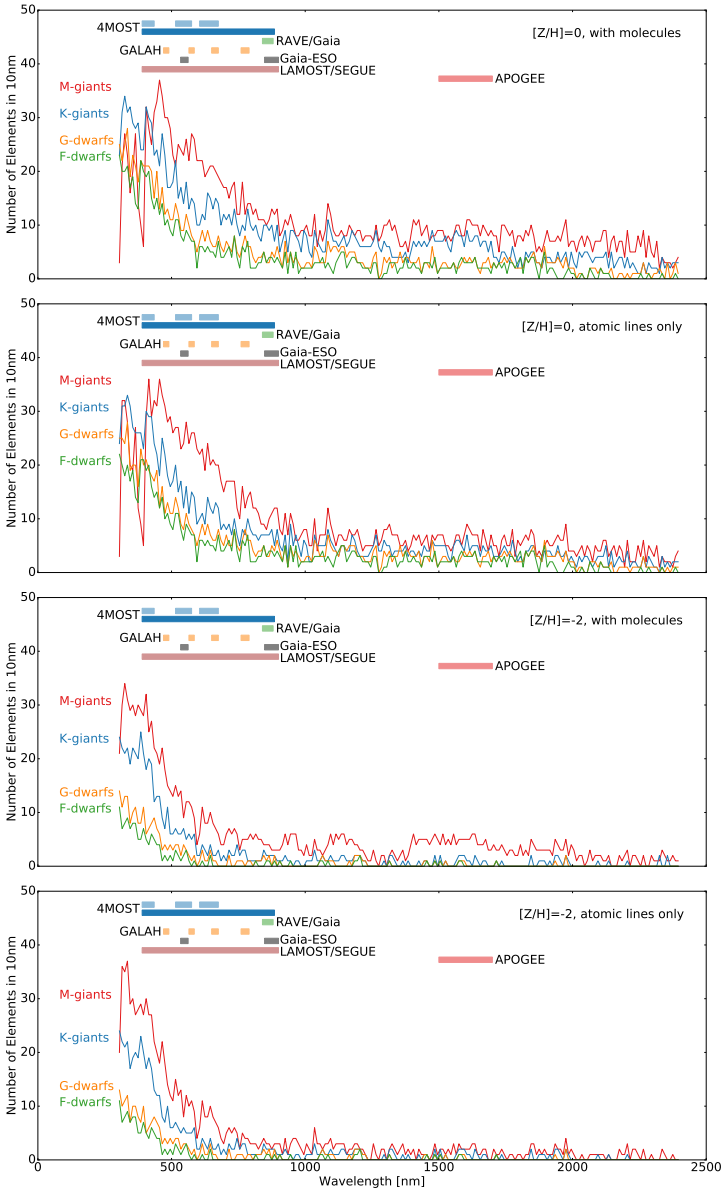

Figure 6.11: Number of elements with detectable spectral signatures within any 10 nm portion of the spectrum. The panel layout is the same as Fig. 6.10. We define an element to have detectable spectral signatures if there is at least a spectral line with gradient greater than 0.025 dex^{-1} at the resolution of $R = 6{,}000$. Note that the total number of detectable elements (cf. Fig. 6.5) is necessarily larger than the values shown in the y-axis because different elements contribute at different wavelengths. Only a few important molecules contribute to most of the information in the infrared. Therefore, despite its high information content as shown in Fig. 6.10, the number of elements with detectable spectral signatures in the infrared is much smaller than in the optical.

elements have detectable spectral signatures in each of these portions. We define an element has detectable spectral signatures if there is at least a spectral line with a gradient greater than 0.025 dex^{-1} at $R = 6{,}000$. The elements that have detectable spectral signatures at each portion can be different, therefore the total number of detectable elements is larger than the value in the y-axis. We refer readers to Fig. 6.4 for the total number of detectable elements of each survey. The horizontal bars in these figures illustrate the wavelength ranges of various spectroscopic surveys. For 4MOST, the long bar shows the wavelength range of the low-resolution configuration, and the split short bars show wavelength range of the medium-resolution configuration.

The top two panels show the information content of $[Z/\mathrm{H}] = 0$, and the bottom two panels show the information content of $[Z/\mathrm{H}] = -2$. As expected, metal-rich stars contain more information and can detect more elements than metal-poor stars because there are more spectral lines. In each of these two panels, we include the molecular contributions in the top panel and leave them out in the bottom panel. Each panel in Fig. 6.10 and Fig. 6.11 shows a similar monotonous increment of information for cooler stars. The difference in total information content for an M-giant and an F-dwarf can differ up to two orders of magnitude and 20 elements depending on the wavelength. It is not surprising that cooler stars have much more information because many spectral lines form at a lower temperature, especially from molecular contributions in the infrared. However, since most information in the infrared comes from a few important molecules, the number of elements with detectable spectral signatures is much lower for the infrared, despite its high information content. The number of detectable elements per 10 nm range is $10 - 30$ in the optical but is fewer than 10 elements in the infrared.

Since much of the information in the infrared comes from molecules, and due to the composite nature of molecules, this also vividly demonstrates the importance of methods like the PSM to fit all stellar labels simultaneously. Although optical wavelength contains more information, extinction is much more significant in the optical, therefore, optical surveys are typically limited to the solar neighborhood. Infrared surveys, like APOGEE, are better able to cover a larger region of the Milky Way. Clearly, depending on the science goal, the wavelength range of a spectroscopic survey should be carefully chosen. Interestingly, at optical wavelengths, there are about 10–30 elements per 10 nm range, showing that for surveys that are restricted to small wavelength ranges, such as GALAH, we can still easily measure more than 30 elements. Even for surveys like RAVE or Gaia RVS that have very limited wavelength ranges, the information content suggests that, with robust models, we should be able to detect $\sim$ 15 elements. Finally, below 400 nm, spectral lines in M-giants become so dense that they form an absorption trough with zero stellar flux. We have virtually zero gradient spectra for most elements in this trough, and as a result, both the information content and the number of elements with detectable spectral signatures drop precipitously for M-giants at wavelengths bluer than 400 nm.

6.8 Appendix: Stellar label uncertainty as a function of spectral resolution

We show in Section 6.3 and in Fig. 6.2 that, given the same exposure time and the same number of detector pixels, beyond $R \gtrsim 1{,}000$, stellar label uncertainties are largely independent of spectral resolution. The gain from a higher S/N and a larger wavelength range for low-resolution spectroscopy compensates the linear trend of uncertainty with resolution when assuming the same S/N and wavelength range. In this appendix, we will study the absolute uncertainties of a few stellar labels to demonstrate this result in more detail. Similar to Section 6.3, we assume an anchor point at $R = 6{,}000$, i.e., at $R = 6{,}000$, we adopt the wavelength range as assumed and a S/N per wavelength pixel of 100. For other resolutions, we assume a wavelength range inversely proportional to the resolution and a $\sqrt{R/6,000}$ time better/worse in photon noise so that they consume the same number of detector pixels and exposure time. We also assume that spectral information is uniformly distributed over the entire wavelength range, so the larger/smaller wavelength range changes the uncertainty by a factor $\sqrt{R/6,000}$.

Fig. 6.12 considers the wavelength ranges of the 4MOST survey (optical, $\lambda = 390 - 885$ nm) and the APOGEE survey (infrared, $\lambda = 1{,}500$–$1{,}700$ nm), adopting spectra for solar metallicity K-giants. We only choose a few stellar labels for the purpose of illustration. Although not shown, the other stellar labels follow roughly the same trend. The green and red filled symbols show the survey resolutions of 4MOST and APOGEE. Since LAMOST and SEGUE share a similar wavelength range as 4MOST, we overplotted their survey resolutions as red hollow symbols.

Beside the weak dependence of uncertainty with spectral resolution, Fig. 6.12 also shows that

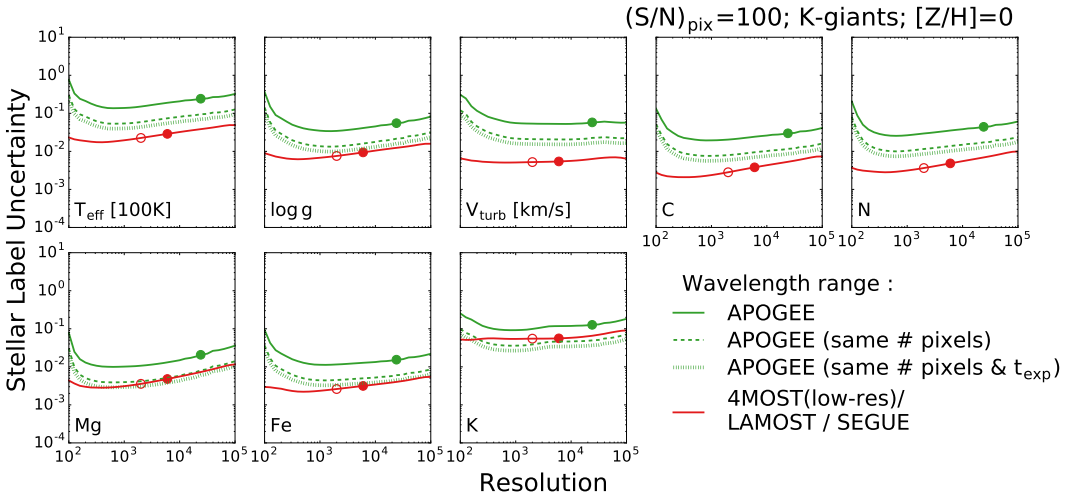

Figure 6.12: Stellar label uncertainties as a function of spectral resolution, assuming the same exposure time and the same number of detector pixels. We assume an anchor point at $R = 6{,}000$, i.e., we show at $R = 6{,}000$ the stellar label uncertainties with $\mathrm{S/N} = 100$ per wavelength pixel and the wavelength range of the spectroscopic surveys. For other spectral resolutions, we vary the S/N according to Poisson statistics assuming the same exposure time (low-resolution spectra have higher S/N) and further scale the label uncertainties by $\sqrt{R/6,000}$, taking into account that low-resolution spectra have a more extensive wavelength range for the same detector real estate. Shown are only a few stellar labels for the purpose of illustration. We find the stellar uncertainties to have a weak dependence on spectral resolution. The solid green lines adopt the wavelength range of APOGEE and the solid red lines adopt the wavelength range of 4MOST (in the low-resolution configuration), SEGUE, and LAMOST. The green symbols, red filled symbols, and red hollow symbols demonstrate the actual survey resolutions of APOGEE, 4MOST (low-resolution), and LAMOST/SEGUE, respectively. Since 4MOST has a larger wavelength range and a smaller resolution element than APOGEE, and red giants are brighter in the infrared than the optical, the green dashed and dotted lines account for these differences in order to have a fairer comparison for the optical and the infrared. The green dashed lines assume that APOGEE has the same number of wavelength pixels as the optical surveys, and the green dotted lines further assume the same exposure time (APOGEE achieves higher S/N because giants are brighter in the infrared).

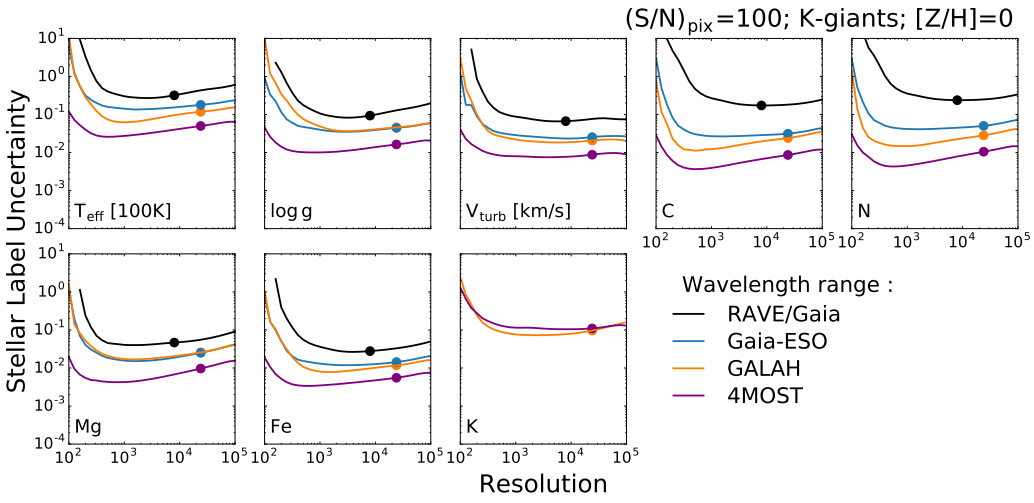

Figure 6.13: Same as Fig. 6.12, but for the other spectroscopic surveys under consideration here. Note that RAVE and Gaia RVS do not detect K, so we omit RAVE and Gaia RVS from the potassium subplot.

4MOST has better uncertainties than APOGEE. However, 4MOST also has a larger wavelength range than APOGEE and more wavelength pixels per wavelength range because the resolution element in the bluer wavelength is shorter. Furthermore, red giants are brighter in the infrared than in the optical. For example, the mean flux of a K-giant in the APOGEE wavelength range is about twice of the mean flux in the 4MOST wavelength range. Therefore given the same exposure time, the S/N for the APOGEE survey is about $\sqrt{2}$ better. In order to have a fairer comparison, the green dashed and dotted lines take into account these differences by scaling the APOGEE uncertainties accordingly. Since spectral information adds in quadrature, in the dashed lines, we scale the APOGEE uncertainties in the green solid lines by the square root of the ratio of number of pixels between 4MOST and APOGEE. The dotted lines further scale the uncertainties by a factor of $\sqrt{2}$ due to the brighter flux in the infrared. Fig. 6.12 shows that even compared to the scaled uncertainties, 4MOST still achieves better precision, demonstrating that the optical wavelength indeed has more spectral information than the infrared, consistent with our assessments in Appendix 6.7.

Fig. 6.13 shows similar results, but assuming wavelength ranges from other spectroscopic surveys. Both figures show a weak dependence of uncertainty with resolution beyond $R \gtrsim 1{,}000$ demonstrating that this trend is generic and is independent of wavelength range. Although not shown in this appendix (cf. Fig. 6.2), we also tested that this trend remains the same for other stellar types and metallicities. But on top of these, Fig. 6.13 also illustrates that surveys having shorter wavelength ranges, e.g., RAVE, Gaia RVS, and GALAH, tend to deviate more from this trend. These surveys have fewer wavelength pixels, as a result, stellar labels become more degenerate and produce deviations from the flat trend seen at higher resolutions.

Throughout this study, we often assume S/N = 100 per pixel. But we emphasize that S/N plays no role in most of our discussions because theoretical uncertainty exactly scales linearly with S/N (cf. Eq. 6.1). Since we focuses on the relative uncertainties in this study, the contributions from S/N cancel out. Therefore, our general conclusions in this study are completely independent of S/N.

6.9 Appendix: Correlation of stellar labels as a function of spectral resolution

In Section 6.3.4 and Fig. 6.6, we studied the global distribution of correlations from all detectable stellar labels. Here, we show more details in this Appendix. Fig. 6.14—Fig. 6.16 show each of the pairwise correlations that comprise the global distribution. We assume the wavelength range of APOGEE, solar metallicity, and K-giants. We define an element to be detectable if its uncertainty at $R = 24{,}000$ is better than 0.01 dex. This definition makes a total 23 stellar labels (20 detectable elements). Fig. 6.14–Fig. 6.16 show the pairwise correlations assuming $R = 100$, 1,000 and 24,000, respectively – each panel shows the correlation of a label pair. We shade each panel with the correlation value to guide the eye, and adopt a color scheme in log scale to increase the contrast since most label pairs have moderate correlations between 0.2–0.4. We also tested the other wavelength ranges from different surveys and found that the results remain qualitative the same.

Fig. 6.14 shows that almost all stellar labels are degenerate at $R = 100$. At this resolution, there are only 30 wavelength pixels in the APOGEE wavelength range. Most stellar labels contribute to the same set of pixels. However, when increasing the resolution to $R = 1{,}000$, as shown in Fig. 6.15, most labels are already not strongly correlated. Only stellar labels that contribute to most pixels – $T_{\rm eff}$, $\log g$, $v_{\rm turb}$, Fe, C, N, O – have strong correlations. Going to an even higher resolution, e.g., $R = 24{,}000$ as shown in Fig. 6.16, no longer decreases the correlations of stellar labels significantly. Stellar labels that have consistent contributions to all pixels continue to correlate in most cases even at the highest resolution. Although not shown, we find that the correlations at $R = 100{,}000$ remain

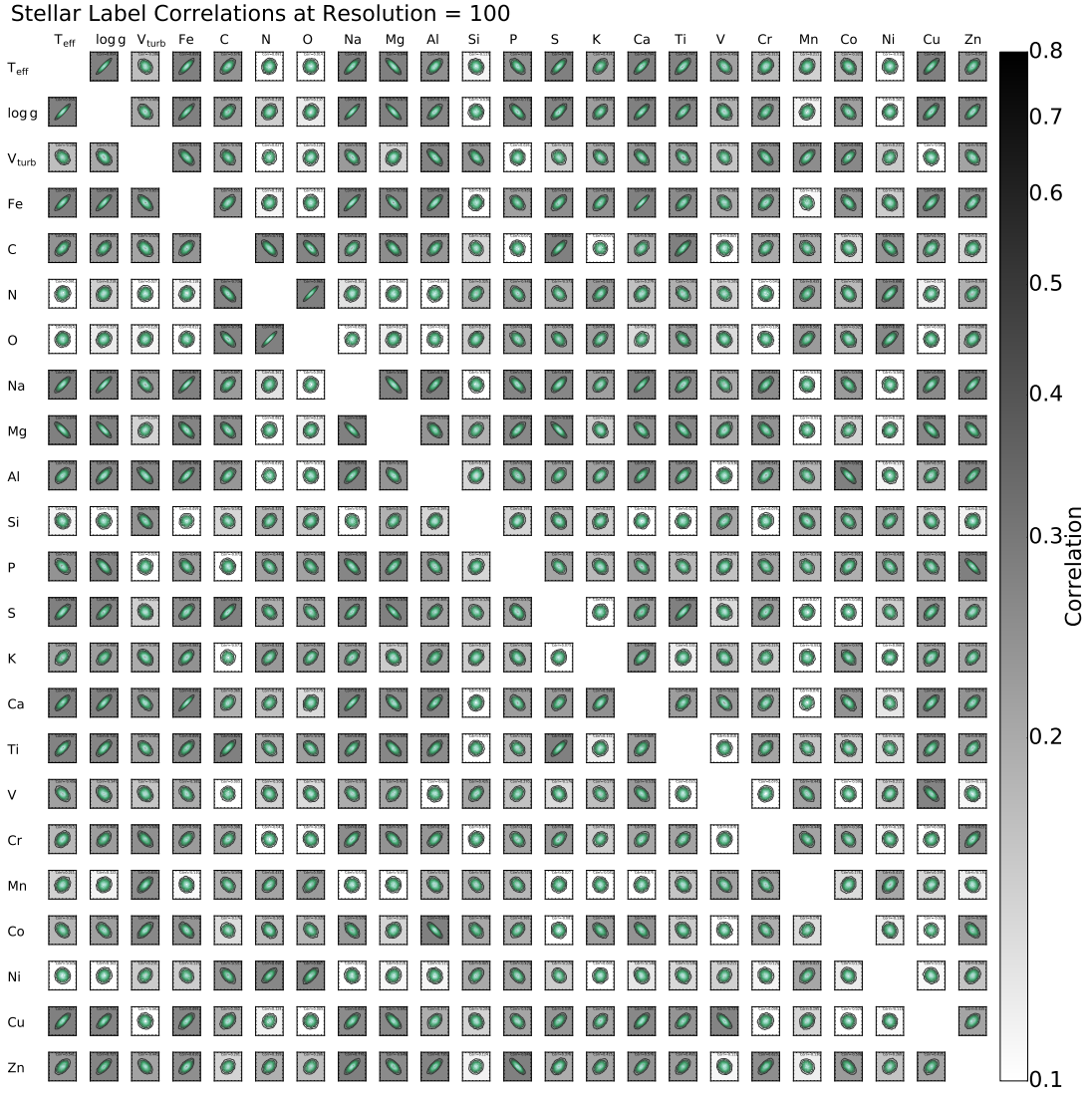

Figure 6.14: Correlations in the estimates of all detectable stellar labels, assuming the wavelength range of APOGEE but at an assumed $R = 100$. Each panel shows the correlation of a different label pair, with darker shade indicating a stronger correlation. At $R = 100$, most stellar labels are strongly correlated.

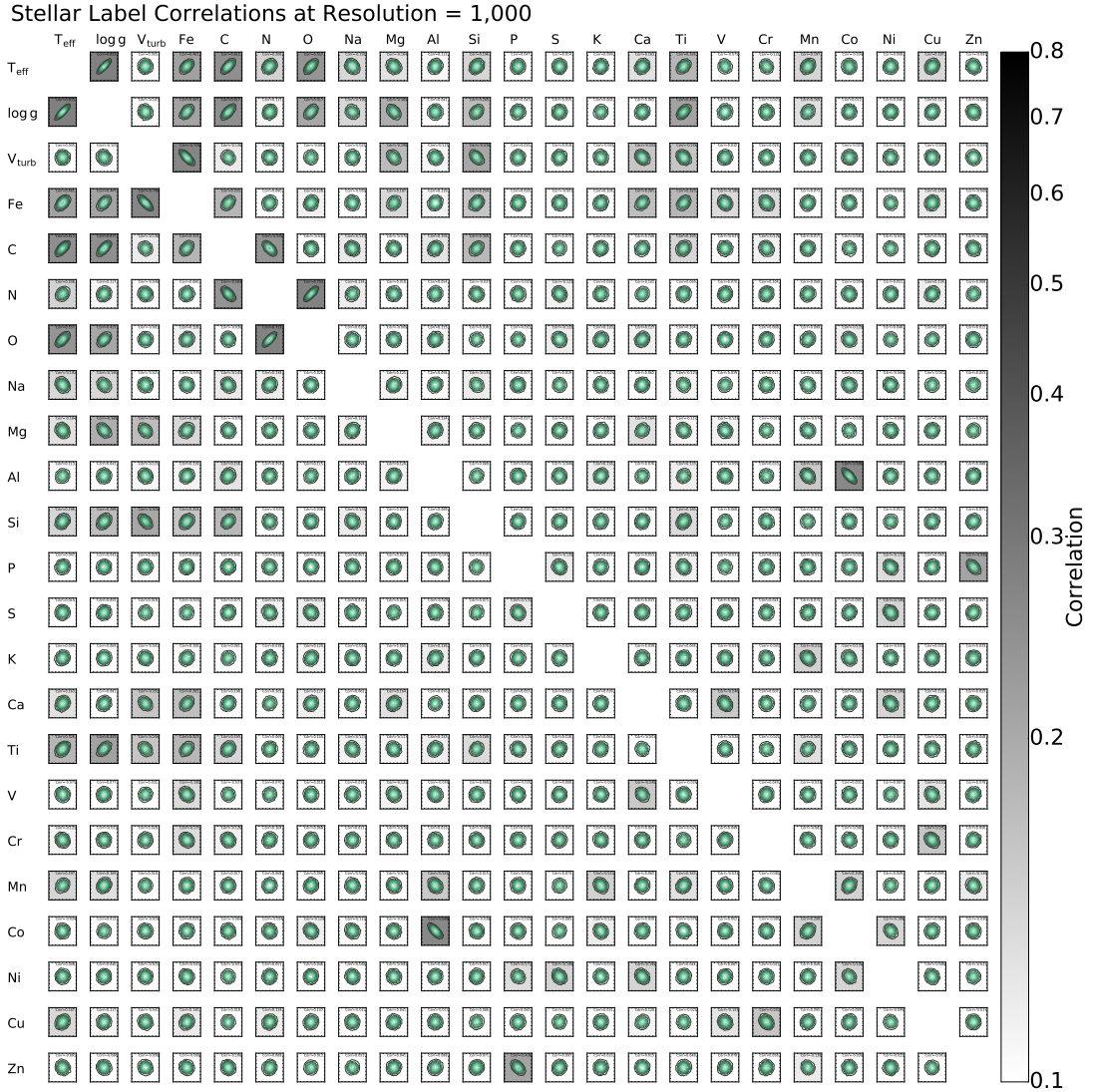

Figure 6.15: Analogous to Fig. 6.14, but for an assumed resolution of $R = 1{,}000$. At $R = 1{,}000$, most stellar label estimates are largely uncorrelated. Only those stellar labels that contribute to most wavelength pixels, such as $T_{\rm eff}$, $\log g$, $v_{\rm turb}$, Fe, C, N, O, have strong – and well-known – correlations.

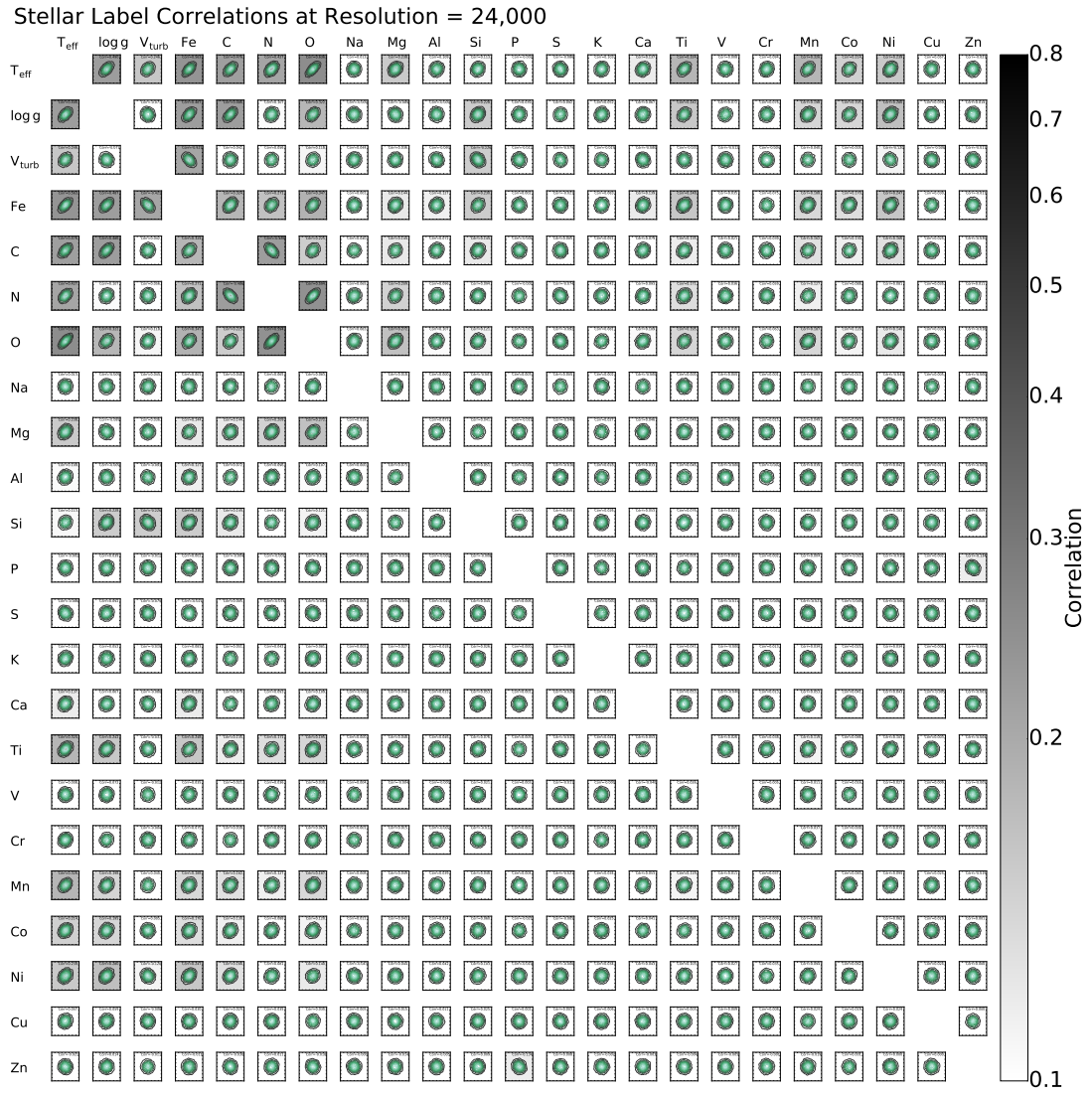

Figure 6.16: Analogous to Fig. 6.14, but here we assume $R = 24{,}000$. Increasing spectral resolution from $R = 1{,}000$ to $R = 24{,}000$ has only minimal effect on the correlations in stellar labels.

practically the same as $R = 24{,}000$. Finally, we note that the correlations evident at high-resolution are often missed by stellar characterization methods that do not solve for all parameters at once, e.g., classical equivalent width based techniques.

7

Ongoing work and future directions

With the advent of multiple current and planned large-scale multiplexed spectroscopic surveys of the Milky Way, the fundamental question that must be answered is what will be learned by collecting spectra from millions of stars in the Milky Way that is not possible with a smaller sample. Recent results from the APOGEE survey have provided a partial answer to this question. The APOGEE survey showed that collecting 10^5 stars from a large area of the Milky Way is crucial to having a complete view of the Milky Way[86]. For example, the APOGEE survey revealed that the α-enhanced sequence of the Milky Way shows the same trend everywhere in the disk, supporting a short timescale formation of the thick disk. The APOGEE survey has undoubtedly launched the dawn of a revolution in the field of Galactic archaeology with big data. However, it is important to show *quantitatively* that collecting an additional order of magnitude more stars, i.e., 10^{6-7} stars, can indeed provide a new understanding of a galactic system. More precisely, in what context the number of stars in a survey matters the most. For example, the Disco proposal for AS4 aims to observe 5 million stars; it is crucial to figure out what would be the optimal survey strategy for such a survey.

To tackle these questions, my thesis work has focused on the challenges and opportunities of the powerful technique known as chemical tagging. Chemical tagging is an interesting test case because, in principle, chemical tagging *always* benefits from having more stars as the chance of collecting two stellar siblings simply grows quadratically with the number of stars. Demonstrating the feasibility of chemical tagging (or thematically similar variations of this technique) and quantifying the gain with the number of stars sampled in a survey will provide a strong case for future large-scale spectroscopic surveys.

In this thesis work, I showed that the interesting "birth similarity" features of chemical tagging

are currently smeared out because:

1. The Milky Way has many more stars than the current sample of 10^5 stars. We only sample a few stars from each star cluster, and interesting clumps are hardly visible. This limitation argues that 10^6 or more stars are necessary (though not sufficient) to realize strong chemical tagging.
2. The current elemental abundance measurements have limited precision; most clusters are unresolved in the elemental abundance space given the current $[X/\mathrm{H}]$ precision of about 0.05 dex.
3. Many star clusters are low mass, meaning that the few stars they contribute are part of a "background" sea of stars. The dominating background implies that even an overdensity in the elemental abundance space does not necessarily translate into a group of conatal stars.

The successful realization of chemical tagging thus requires that we push the boundaries on multiple fronts. I have laid out groundwork to mitigate some of these difficulties (cf. Chapter 4–6). In this final chapter, I will describe other ongoing work and future directions that will further tackle these three challenges. It is important to emphasize that although these efforts are primarily aimed at realizing strong chemical tagging, other forms of chemical tagging will also benefit.

Finally, we note that even in the pessimistic limit where clumps remain unresolved in the elemental abundance space, the extent to which stars form in small or massive clusters still imprints clear signatures on the statistical properties of the elemental abundance space (cf. Chapter 3). This "statistical-level" chemical tagging will benefit tremendously from a larger sample in the future (see Fig 1.2) because a larger sample will reduce the Poisson errors of the global statistical fluctuations sought by this method. Collecting spectra from $3 - 10$ million stars in the future planned spectroscopic surveys (e.g., Weave, 4MOST, Disco in AS4) will be highly valuable in this pursuit. In a way, the idea of "statistical-level" chemical tagging resembles the study of baryon acoustic oscillation in

cosmology – even though individual fluctuations can be small and are hard to interpret, the global statistics of the datasets is nevertheless incredibly rich. Moreover, in the limit where we are not systematically limited by chemodynamics models of the Milky Way, the results of "statistical-level" chemical tagging techniques will always improve with a larger sample.

7.1 Pushing the boundaries on the spectral fitting technique

Realizing chemical tagging requires the maximal extraction of information from stellar spectra, i.e., we should attempt to attain the theoretically achievable precision (what is known as the Cramer-Rao bound, cf. Chapter 4–6), using all information in the spectra. But this is not possible without a self-consistent, simultaneous fitting of all relevant stellar parameters and elemental abundances to the full spectral range (cf. Chapter 4). To this end, I introduced a new spectral fitting technique (Chapter 4–6), but the method presented in these chapters relies on the assumption that a quadratic function is a good approximation of the variation of flux with stellar labels. While this assumption works sufficiently well when dealing only with the giants (also see[153]), the approximation becomes less precise if we try to fit the full range of stellar parameters with a single model. For example, some lines only form at a low temperatures (e.g., $T_{\rm eff} < 4{,}000$K), and the opacity of the stellar atmosphere changes more dramatically in this regime. As a result, a quadratic approximation must fail and a more accurate approach requires stitching together multiple grids for different spectral types of stars. The stitching approach is what I adopted in Chapter 4. This approach is adequate, but it is not ideal since it requires making additional choices on how to pick the best models within the stitching regions.

Furthermore, the quadratic assumption is more problematic for photometric data as well as spectra with very low resolution ($R \lesssim 1{,}000$). Exploring how to build a model with the technique we presented for photometric data may be crucial in Galactic archaeology with the data from LSST, Gaia BP/RP and WFIRST arriving in the next few years. But in these low-resolution limits, a com-

bination of many blended features can contribute to a single pixel or a photometry band. The quadratic assumption works because it represents well the curve of growth of a single absorption line in certain limits, but this approximation is likely to be less true as we combine more lines. Therefore, the simplistic quadratic assumption can limit the achievable precision as we try to generalize the method to fit both giants and dwarfs self-consistently and extend to other applications (e.g., galactic spectra, SED fitting). Even considering the giants alone, we have already shown that some stellar parameters, such as microturbulence, are not well approximated (Fig. 6.7) by a quadratic function.

In the ongoing work described in the following sections, I have improved the method by considering non-parametric options and replacing the quadratic models. I tested with model spectra and showed that a single non-parametric model can predict the variation of flux for a much larger range of spectral types, covering spectral types from M-giants to F-dwarfs ($T_{\rm eff} = 3{,}000\mathrm{K} - 6{,}000\mathrm{K}$). The non-parametric regression method predicts the variation of flux by finding the best smooth "function" that represents well the training data without restricting to an ad-hoc function as before. The results that I will present next adopt support vector regression for the non-parametric extension but it has some shortcoming. Similar to Gaussian processes, support vector regression is computationally slow because the method is still intrinsically an interpolation scheme that requires loading a large amount of data for the predictions. I am currently exploring neural networks as a substitute for support vector regression since neural networks are capable of overcoming the interpolation limitation and are very computationally efficient once the neural networks are trained.

7.2 Pushing the boundaries of elemental abundance precision

Strong chemical tagging requires measuring 7–9 elements from different elemental abundance groups[195] with a relative (not absolute) precision of 0.03 dex for more than a million stars (cf. Chapter 2–3). Chemically homogeneous substructures should be significantly more apparent in the elemental abundance space if this criterion can be attained. The studies of open clusters suggest that stars born in the same star cluster should be chemically homogeneous at a level better than 0.02 dex[29,131]. However, they appear vastly smeared out because the current precision of elemental abundances provided by most massive spectroscopic surveys is 0.05 dex[181], making the accurate identification of stellar siblings much more challenging.

I demonstrated that restricting to strong unblended lines in the current pipeline only exploits about 10% of the spectral information (cf. Chapter 4). Although some lines are excluded because their atomic parameters are not very well understood, some other lines are excluded merely because they are blended. These blended lines should provide more spectral information and should yield better precision. I have tested how well the new spectral fitting technique, which we plan to rebrand as "The Payne",* works by fitting real APOGEE spectra with all 19 labels – effective temperature, surface gravity, microturbulence, C^{12}/C^{13} and 15 elemental abundances.

Fig. 7.1 shows the dispersion of the open cluster M67 from its members with $T_{\rm eff} = 4{,}500\,\mathrm{K} -$

*The philosophy and idea behind my technique resemble another technique in the field – the Cannon – which has recently garnered much attention. But unlike the Cannon[153], which is a data-driven model, my method "trains" on ab-initio calculated synthetic models. Since this approach is complementary to the Cannon and is a theoretical version of the Cannon, it might be appropriate to name this method "the Payne" in honor of Cecilia Payne-Gaposchkin, another great astronomer.

5,000 K. The orange line shows the elemental abundance estimates calculated using the Payne, and the blue line shows the estimates from APOGEE DR13 (the latest data release). Since open clusters should be very chemically homogeneous, the internal dispersions of their elemental abundances are a good probe of measurement uncertainties. The figure shows that the Payne method, at least for this limited temperature range, yields a precision of 0.03 dex, which is better than 0.05 dex from APOGEE DR13. Without resorting to a sophisticated differential analysis, the Payne demonstrates that open clusters, in this case, M67, are indeed chemically homogeneous to a level better than 0.03 dex. However, due to the imperfectness of synthetic models, I found that to achieve a similar precision beyond this temperature range requires a robust model of the residual as a function of stellar parameters, i.e., we need to adjust the imperfectness of the stellar models empirically from the fit. This is a direction that I am currently pursuing.

Besides empirically correcting stellar models by studying the residuals of the fits, a more ambitious future direction that I intend to pursue is the calibration of a line list beyond the Sun and Arcturus. Both of these stars are hot and hence many spectral lines are missing in the current calibration of the stellar line list. We note that the Payne can be extended naturally to fit for both the stellar labels as well as the line list parameters. This may be the ideal way to explore the massive spectroscopic datasets in hand: generating a near perfect atomic line list for synthetic stellar spectra through empirical data. In practice, one possible way to do this is by adopting the linear method presented in Chapter 4. If the atomic parameters are only slightly biased, the line list can be corrected by studying the linear response function of each atomic parameter. In this limit, the problem of improving the line list becomes a simple linear algebra problem and can be solved via linear regression.

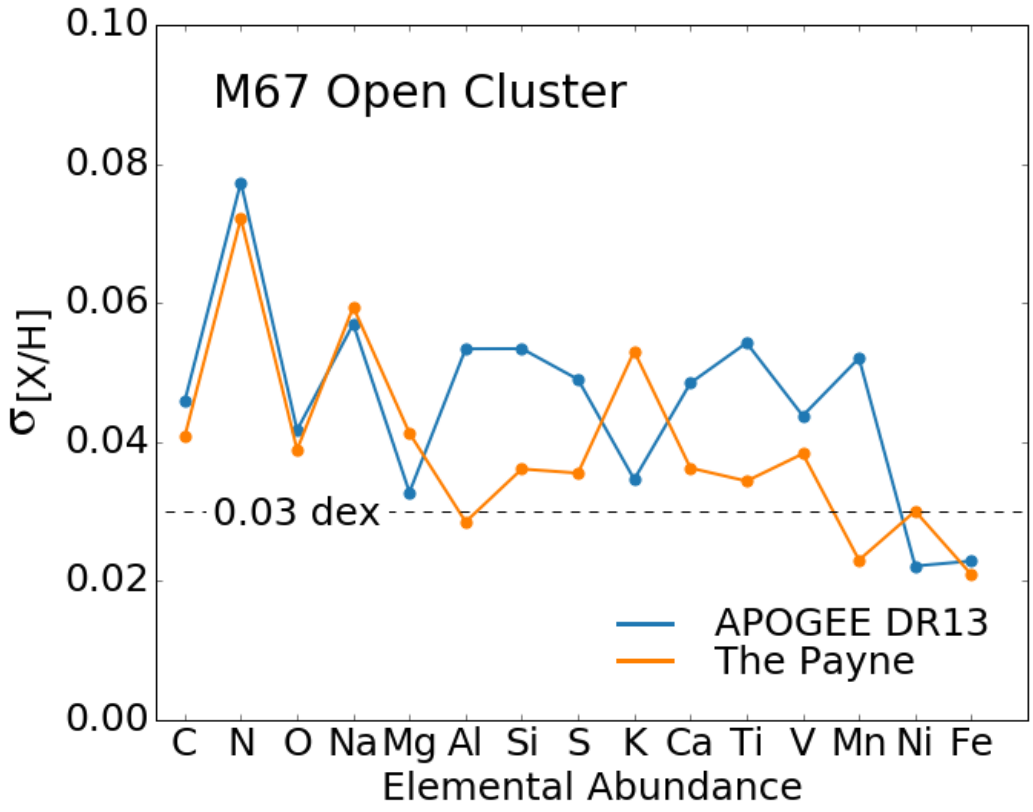

Figure 7.1: Demonstration of the power of my new spectral fitting technique (the Payne). I fit the members of the open cluster M67 within the temperature range of $T_{\rm eff} = 4{,}500\,{\rm K} - 5{,}000\,{\rm K}$. Since open clusters are thought to be very homogeneous, the internal dispersion of the open cluster shows the measurement uncertainties of different methods. With just 2,000 training models in the synthetic library, as shown with the orange line, the Payne recovers most elemental abundances with a precision of $\sim 0.03\,{\rm dex}$ which is better than $0.05\,{\rm dex}$ from the APOGEE DR13 pipeline as shown with the blue line.

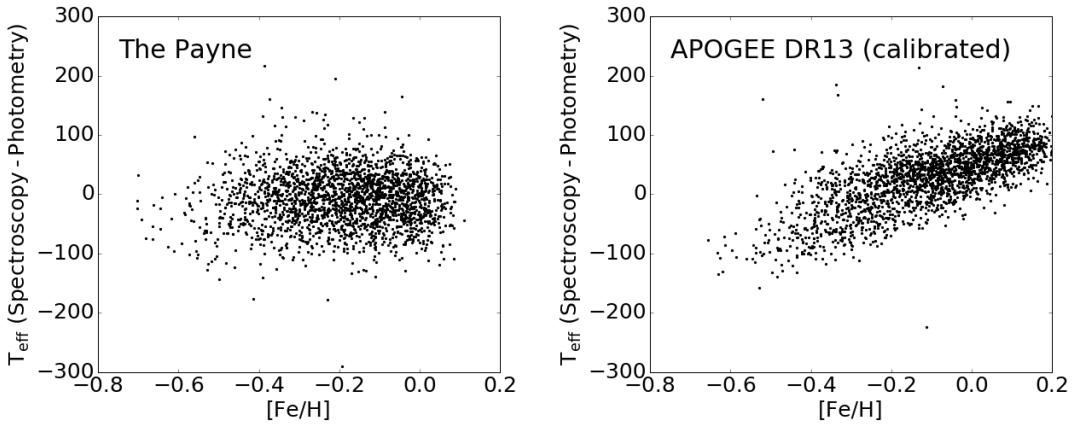

Figure 7.2: The spectroscopic $T_{\rm eff}$ from the Payne is more consistent with the "ground truth" photometry $T_{\rm eff}$. We consider a subsample of APOGEE stars that have asteroseismic data from Kepler and show the differences between the spectroscopic $T_{\rm eff}$ and the photometric $T_{\rm eff}$. The latter is estimated using the asteroseismic masses and the MIST isochrones. We also tested that using the IRFM (infrared flux method) $T_{\rm eff}$ as the photometry $T_{\rm eff}$ instead does not alter the conclusion. The left panel assumes the best-fitting spectroscopic $T_{\rm eff}$ with the Payne and the right panel adopts $T_{\rm eff}$ from APOGEE DR13. The APOGEE $T_{\rm eff}$ has a prominent systematic bias with metallicity. Conversely, $T_{\rm eff}$ estimates from the Payne do not show such a trend and agree better with the photometry $T_{\rm eff}$.

Finally, besides achieving a better precision of 0.03 dex, perhaps another major achievement of the Payne is that it has also potentially resolved the photometry-spectroscopy temperature discrepancy seen in the APOGEE survey. To show this, I fit the APOGEE data that have asteroseismic estimates from the Kepler satellite. The asteroseismic data from Kepler provides a unique probe of the temperature that is completely independent from the spectroscopic estimates. Combining the bolometric luminosities estimated using the asteroseismic masses, the APOGEE DR13 metallicities and the MIST isochrones [43], I estimate another stellar $T_{\rm eff}$ which , for simplicity I will refer to as the "photometry temperature".[†] We also checked our results remain the same if we were to use the temperature calculated using the infrared flux method (IRFM). Fig. 7.2 shows the differences between the photometry temperature and the spectroscopic temperature estimates from the Payne versus the APOGEE DR13. The APOGEE DR13 temperature shows a prominent systematic deviation from the photometric temperature as a function of metallicity, but this puzzling trend disappears if we fit everything self-consistently with the Payne.

Due to its many promises, future APOGEE data releases may implement the Payne method. As I am a member of the GALAH, APOGEE, and 4MOST surveys and have strong connections with the Gaia-ESO and LAMOST surveys, I plan to incorporate this technique into these surveys moving forward. A combination of the Payne and the data-driven Cannon [153,40] could be a very powerful combination to extract the most information out of every spectrum from these surveys, homogenize

[†]This calculation is done by Jieun Choi. We also note that this temperature should, strictly speaking, be called the MIST temperature because it relies on the MIST isochrones instead of an SED fitting. But we found that the MIST temperature agrees fairly well with the IRFM temperature. For simplicity and avoid cluttering of nomenclatures, we simply refer to it as the photometry temperature.

the elemental abundances across different surveys and improve the atomic line list.

7.3 Pushing the boundaries on spectroscopic ages

Another way to improve strong chemical tagging is to expand the search space beyond elemental abundances. For example, stellar age can serve as an additional dimension for the chemical tagging searches because stars that were born together by definition should have the same age. Increasing the dimensionality of the search space could better resolve different disrupted clusters that might otherwise be unresolved in a crowded lower-dimensional projection. However, measuring precise stellar ages for giants is non-trivial. The isochrones in the $T_{\rm eff}$-log g space and the color-magnitude diagrams for giants only span narrow dynamical ranges. Small uncertainties in the observables can translate into substantial changes in estimated stellar ages for giants due to the lack of discriminatory power from these methods.

Thankfully, stars only spend a short time in the giant phase and a much longer period on the main sequence. Therefore, we can obtain precise stellar age estimates if we know the main sequence lifetime, which is determined by the stellar mass. Measuring precise stellar masses are possible for the giants. A recent study from Martig et al.[142] demonstrated the C/N ratio is good mass indicator for giants because in the giant phase, stars undergo dredge ups, bringing material from the core of the stars to the surface. The enhancement of the nitrogen abundance (at the expense of the carbon abundance) happened during this dredging process depends on: (a) the depth of the convective layer and (b) the details of the CNO cycle in the stars. Both of these quantities are sensitive to stellar mass which in turns makes the C/N ratio a good mass (and hence age) indicator for giants. Also note that in my previous studies (e.g. Chapter 3), I did not include the abundances of CNO in the chemical

tagging searches. Even if stars born from the same cluster are chemically homogeneous, unlike other heavier elements, the CNO abundances of these stars may be inhomogeneous as they evolve, a result of stellar evolution and dredge ups. Measuring stellar ages using CNO thus provides a way to use these additional, currently unused elemental abundances in chemical tagging.

Different lines in Fig. 7.3 show the C/N ratio as a function of $T_{\rm eff}$ from the MIST isochrones [43]. An important point here is that there is a substantial separation in [C/N] for different isochrones showing that the C/N ratio is indeed a good age indicator for giants, in agreement with the preceding discussion. Plotted on top of these isochrones are the [C/N] values of the open cluster NGC6819's members. The left panel shows the values estimated with the Payne and the right panel shows the values from APOGEE DR13. Kalirai et al. [99] estimated from photometry that NGC6819 has an age of $\sim$ 2.5 Gyr and our spectroscopic determination agrees very well with their age estimate. Conversely, the APOGEE values do not coincides with the photometry age estimate and have a larger spread, showing that the Payne estimates are not only more precise, but they are also potentially more accurate. The better performance of the Payne is not surprising. Specifically, the abundances of C and N are degenerate with other main stellar parameters such as $T_{\rm eff}$ and $\log g$. As we have shown in Fig. 7.2, the APOGEE $T_{\rm eff}$ is likely to be systematically biased and this problem could affect the fitting of the C and N abundances. This result also demonstrates that it is possible to attain very precise stellar ages (with an uncertainty $\lesssim$ 2 Gyr) from spectroscopic data alone because the spread of the Payne estimates for this particular cluster is tighter than the differences between the 2.5 Gyr and the 4 Gyr isochrones.

Finally, besides measuring more accurate C/N ratios, a particularly exciting prospect of the Payne

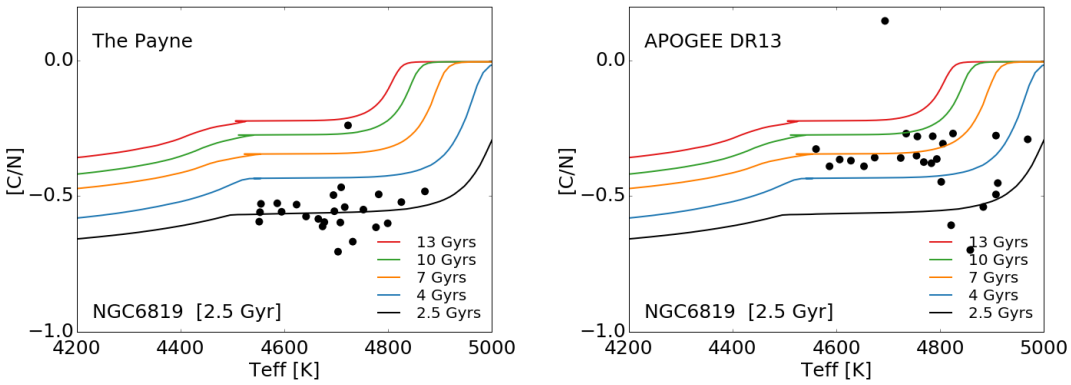

Figure 7.3: The Payne may provide better stellar age estimates directly from spectroscopic data alone. The solid lines show the MIST isochrones of different ages and the data points show [C/N] and $T_{\rm eff}$ of individual stars in the open cluster NGC6819. The left panel plots the estimates from the Payne, fitting the APOGEE spectra, and the right panel adopts values from APOGEE DR13. Photometric data suggests that NGC6819 is an $\sim 2.5\,\mathrm{Gyr}$ cluster[99]. The spectroscopic estimates from the Payne agree better with the $2.5\,\mathrm{Gyr}$ isochrone compared to the APOGEE DR13 values, suggesting that Payne elemental abundance estimates are likely to be more accurate.

– and another of my ongoing projects – is to measure the C^{12}/C^{13} ratio using the blended CN and CO lines for the APOGEE sample. Similar to the C/N ratio, the C^{12}/C^{13} is also a prominent age indicator because C^{13} also gets dredged up at the expense of C^{12} during the giant phase. Combining the C^{12}/C^{13} measurements with the C/N measurements may make the stellar age estimates even more precise.

7.4 Pushing the boundaries on spectral resolution needed for surveys

In Chapter 6, I demonstrated that low-resolution spectra contain the same amount of spectral information as high-resolution spectra, given the same exposure time and the same number of CCD pixels. I showed that it is possible to measure >20 individual abundances even with low-resolution R=2,000–6,000 spectra with the Payne. If this result holds, a reanalysis of the abundantly available low-resolution spectra will open an entirely new avenue for chemical tagging because they will increase the sample size for chemical tagging significantly. That said, low-resolution spectra are certainly more susceptible to the systematics of stellar model spectra, which are often far from perfect. Furthermore, one also has to overcome the problem of determining the continuum accurately for low-resolution spectra. In Chapter 6, I only showed that the Payne works for simulated stellar spectra, but the further step of feasibly extracting multiple (>20) elements from real observed low-resolution spectra is the key to settle these concerns.

Recent work from Casey et al. [41] showed, encouragingly, that it is possible to extract seven elements with RAVE spectra at $R = 6{,}000$ using data-driven models with the Cannon. While the number of elements in their study is limited because the RAVE survey has a short wavelength coverage, their result vividly demonstrates that the idea presented in Chapter 6 can work, at least for the data-driven models. At this point, it might be useful to remind the readers the differences between the Cannon and the Payne (our approach). The Payne aims to fit observed spectra with ab-initio calculated model spectra, whereas the Cannon selects a subset of "standard" stars and uses the observed spectra of these standard stars to be the model spectra. Since the Cannon assumes observed

spectra as model spectra, almost by definition, the major advantage of using the Cannon is that the model spectra are nearly free of systematics – there is no ab-initio calculation of models from first principles involved. Although the idea presented in Chapter 6 has shown to work in the limit of data-driven models, the data-driven approach has serious shortcoming too. Since data-driven models only "transfer" labels, the Cannon will suffer from whatever biases that the training set has. For example, unlike the Payne, it cannot solve the spectroscopic-photometry $T_{\rm eff}$ bias that we have seen in the right panel of Fig. 7.2 – if the training set shows this systematic trend, the inferred stellar labels for the testing set will show the same trend. Therefore, showing that we can obtain > 20 using ab-initio calculated models with the Payne is still an important next step.

In particular, I will continue to collaborate with Charlie Conroy and use the Payne to fit low-resolution spectra of the Gaia benchmark stars currently being collected with the FAST spectrograph. Since these Gaia benchmark stars are well studied, I can then compare the resulting abundance estimates with detailed abundances derived from high-resolution spectra for each of these stars. In addition, I will collaborate with Dan Weisz at Berkeley to demonstrate the efficacy of this idea with the rich and large set of low-resolution archival observations of stars taken with the Keck-DEIMOS spectrograph over the past two decades.

7.5 Refining chemical tagging with kinematic data from Gaia

Finally, even with more sample and more precise abundances, chemical tagging signals can still be difficult to interpret if we do not clean up the background contaminating stars. Recall that an overdensity in the elemental abundance space can have many contributions from the background interlopers. To clean up the background, first it is important to understand what stars contribute to the background. In Chapter 2, I showed that the strongest chemical tagging signals come from "in-situ" star clusters – star clusters that are born within the Solar annulus. As shown in Fig. 7.4, in contrast, ex-situ star clusters (born outside the Solar annulus) generally have fewer than ten stars per cluster in the sample. Stars from ex-situ clusters are noise in clump-finding and prohibit chemical-tagging. In-situ clusters tend to have a higher sampling rate than ex-situ clusters because most of the cluster members are still within the Solar annulus. The implication from my simulations is that if we can exclude ex-situ stars by using the orbital information, for example, from Gaia, the resulting chemical tagging signals will be stronger. This section is dedicated to argue that there should be enough chemical-kinematics information to achieve just that.

It is also important to note that while it is critical to collect a large sample of stars such that we will obtain a sizable sample per cluster, a larger sample size will not automatically resolve the "clump-to-background" ratio problem discussed above. As we collect more stars, both the signals and the background increase in about the same proportion (cf. Chapter 2). Exploring the connection between the chemistry and the kinematics of the stars may be the only way to solving the background problem. Gaia is currently measuring positions and velocities of a billion Milky Way stars, and the

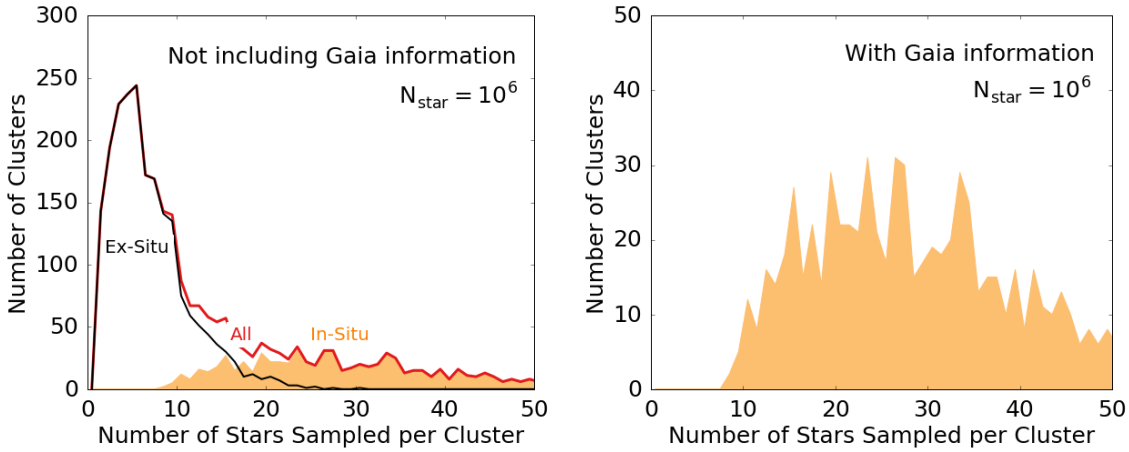

Figure 7.4: Including kinematic information greatly reduces the background of contaminating stars. Left panel: Simulations (with radial migration) of how many stars are sampled per disrupted star cluster in a 10^6 stars survey. We define a cluster to be in-situ if it forms 5–11 kpc from the Galactic center, and ex-situ otherwise. We sample many more stars per cluster from in-situ clusters because most cluster members are still within the Solar annulus. Ex-situ clusters typically contribute fewer than ten stars and so only contribute to a sea of background contaminating stars. Right panel: With kinematic information from the Gaia data, we can significantly reduce ex-situ stars in the sample. In this regime, we will collect more than ten stars for each cluster and chemical tagging signals will be much stronger.

full catalog will become available in a year's time (in late year 2017 or early year 2018). Despite its importance, detailed modeling of how chemical tagging can benefit from kinematic data is mostly absent. Although stars are now phase-mixed, not all information is lost. Some quantities such as angular momentum and integrals of motion of the stars are conserved in a slowly varying galactic potential and can still betray the origins of the stars. For example, stars that are born ex-situ may have more eccentric orbits. The eccentricity of these orbits will manifest itself in the integrals of motion. Furthermore, stars that were born in the inner galaxy and migrated out to the Solar annulus may also have different chemical signatures due to the metallicity gradients of the Milky Way disk.

To test this idea, I have begun a collaboration with Phil Hopkins and Andrew Wetzel at Caltech and the Carnegie Observatories to explore their extremely high-resolution LATTE simulation. Although this simulation is not an exact perfect replica of the real Milky Way, the goal of this exploration is to show that there is an enormous amount of information from the kinematics and the chemistry of the stars that shows tell-tale signatures of the stellar birth radius. A simulation is a powerful tool in this case because the birth radius from the Galactic center of each star particle is known beforehand. It allows us to explore whether or not we can find an optimal mapping that relates the current stellar chemical and kinematics properties to the birth radius properties.

One way to find such a function is through neural networks, a commonly used machine learning technique well-suited to search for a non-parametric mapping from a set of properties to another. I have trained a neural network to map the kinematics properties (6D phase space) and elemental abundances (C, N, O, Ne, Mg, Ca, Si, S, Fe) at $z = 0$ in the LATTE simulation to a in-situ/ex-situ indicator. For the latter, I assigned 1 to a star particle if it is born in-situ and 0 otherwise. I used half

of the star particles within 3 kpc from the Sun as the training set and cross-validated the other half as my testing set. For each testing star particle, we can then assign an "in-situ" probability inferred from the neural network – i.e. how likely this testing star particle is born within the Solar annulus given its current properties. The y-axis in Fig. 7.5 shows the remaining fractions of the in-situ or ex-situ populations if we make an in-situ probability threshold cut shown in the x-axis. The plot demonstrates that, for example, if we make a cut with an in-situ probability threshold of 20% (the dashed line on the left), we will exclude 90% of the ex-situ stars while only scarifying 40% of the in-situ stars. And if we set the in-situ probability threshold to be 40% (second dashed line), we will clean up 97% of the ex-situ stars while discarding only 60% of the in-situ stars. I also checked that the neural network is indeed picking up information about the integrals of motion and the metallicity gradients and use them to discriminate the ex-situ stars from the in-situ stars, confirming my prior intuitions.

The exact numbers mentioned above are not important because they are likely model-dependent. The bottom line is that there is indeed a substantial information about the stellar birth radius encoded in the chemical and kinematic properties. As a next step, I plan to explore the empirical data from APOGEE and build a hierarchical Bayesian model for stellar birth radius. More precisely, using the APOGEE data, Bovy et al.[32] have constructed the stellar surface density of the Milky Way as a function of Galactocentric radius, $\Sigma_*(R)$ for populations with different $[\alpha/\mathrm{Fe}]$–[Fe/H] and are constructing the same for populations with different age–[Fe/H] (J. Bovy, priv. comm.). I plan to extend my models in Chapter 2. Given a star formation rate as a function of Galactocentric radius and cosmic time, $\mathrm{SFR}(R, t)$, a hierarchical Bayesian model can be built to constrain the radial migra-

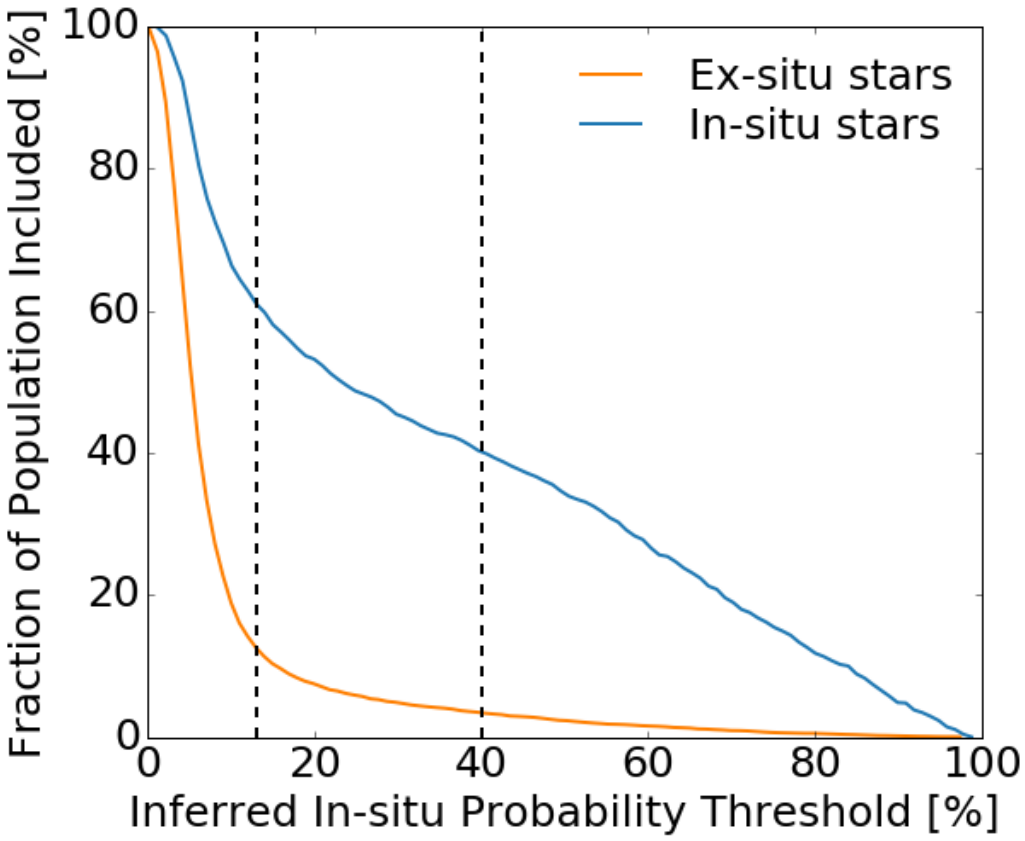

Figure 7.5: Demonstration that current phase space properties and elemental abundances of stars contain information about the birth radius of the stars. A neural network is trained to map the 6D phase space properties and the nine elemental abundances from the $z = 0$ snapshot of the LATTE simulation to the probability which shows whether or not a star is born "in-situ" (i.e., within the Solar annulus). I trained the neural network using half of the star particles within 3 kpc from the Sun and cross-validated with the other half. The y-axis shows the remaining fraction of the in-situ or the ex-situ populations of the testing set if we assume an in-situ probability threshold cut shown in the x-axis. In particular, the two dashed lines illustrate that we can exclude 90% (97%) of the ex-situ stars by discarding only 40% (60%) of the in-situ populations.

tion at different radii and times such that the resulting $\Sigma_*(R, t)$ matches the empirical profiles for various populations. The best-fitting Milky Way model can then be used to provide a birth radius probability distribution function for each star with its current kinematics and chemical properties. Such a model can be very valuable in the Gaia era: not only it will exclude background contaminating stars in the chemical tagging searches, it can also make future chemical tagging surveys more efficient by allowing for preselection of stars that were likely born in-situ.

7.6 Final remarks

In this thesis work, I have discussed the opportunities of strong chemical tagging, the difficulties that this technique faces and ways to mitigate these challenges. In the limit where strong chemical tagging cannot be realized, I presented an alternative approach: using the statistical properties of the elemental abundance space (cf. Chapter 3). I showed that even in the limit where clusters are not entirely resolved in the elemental abundance space, its clumpiness nevertheless contains much information about the Milky Way's past. In this thesis, for the simplicity of arguments, I presented these two regimes as if they are two entirely different forms of chemical tagging. But in practice, there should be a continuous spectrum between these two versions of chemical tagging. In fact, to complete the picture, one should also include the "population-level" chemical tagging.

In the limit where the sampling rate is low, and the measurement precision of elemental abundances is not exquisite, the only form of chemical tagging that is viable is the "population-level" chemical tagging. As we have discussed in the introduction, elemental abundances have shown to be a powerful tool to separate stars from the thin disk, the thick disk, the halo and the dwarf galaxies. Population-level chemical tagging is largely insensitive to the sample size. Therefore, it is a well-established method that has shown to work and has been the dominant form of chemical tagging in the last decade.

But as we are now moving towards sample sizes of $10^5 - 10^6$ stars, we should collect a sizable number of stars from disrupted massive star clusters in the Milky Way. Hence, if these clusters ever existed, their effect should be detectable, at least statistically, in the elemental abundance space (cf.

Chapter 2–3). The statistical-level chemical tagging goes beyond exploiting only the global trend of elemental abundances. Instead, it studies the local properties (e.g., clumpiness) of the elemental abundance space. As more and more data will soon become available, more effort should be invested to further develop statistical-level chemical tagging techniques. This is a largely unexplored area of research in Galactic archaeology, and it may contain many surprises.

The ultimate goal of chemical tagging is to fully resolve disrupted clusters in the elemental abundance space, i.e., strong chemical tagging. As I have discussed at length in this thesis, the realization of strong chemical tagging will require more than just a large sample size. It also requires very precise elemental abundances and refining the signals by excluding the sea of background stars. An important thing to keep in mind is that in practice, we are always somewhere between statistical-level chemical tagging and strong chemical tagging. The Milky Way is a very complex system with many chemodynamics processes in play. It is impossible to claim a reconstruction of disrupted cluster without forward-modeling the Milk Way. Any overdensity in the elemental abundance space has to be interpreted with a proper Milky Way model that includes an exact characterization of the data selection function and the measurement precision. In a way, strong chemical tagging should be regarded as an optimal version of statistical-level chemical tagging, instead of a separate concept.

To put all these together, looking for stellar siblings in the elemental abundance space is like looking for twins in a big room.[‡] The chance of finding two people having the same birthday in a big room is non-negligible even though they might not be twins. These interloping pairs resemble the

[‡]This analogy is modified from a comment made by David W. Hogg in the conference summary of the Galactic Archeology and Stellar Physics conference in Canberra on Nov 21–25th, 2016.

background contaminating stars in chemical tagging. But even in this limit, one can still learn a lot about the number of twins – if we can properly quantify the number of people in the room (selection function) and the number of days in a year (chemical evolution of the Milky Way), we can estimate the number of interloping "twins" that are mistaken by chance. This idea sets the basis for statistical-level chemical tagging. One way to refine the signals is via measuring a more accurate "birth time" instead of a "birthday". As we improve (and can better characterize) the precision of the birth time, we will be able to make stronger statements about the number of real twins, eventually moving toward the regime of strong chemical tagging. The contribution of my thesis work has been to point out that we can study a lot about twins even with the presence of interloping pairs through properly modeling the Milky Way. But more work has to be done. The challenge that lies ahead is to measure the birth time more precisely (more precise elemental abundances), build a better demographic model of the population (in-situ/ex-situ stars) and understand better how many days are there in a year (a better chemodynamical Milky Way model). Achieving these three criteria will hopefully propel us to the strong chemical tagging limit in the near future where each pair that we find has a reasonably good chance of being real twins.

References

[1] Adamo, A., Kruijssen, J. M. D., Bastian, N., Silva-Villa, E., & Ryon, J. 2015, MNRAS, 452, 246

[2] Alam, S., Albareti, F. D., Allende Prieto, C., et al. 2015, ApJS, 219, 12

[3] Allende Prieto, C., Beers, T. C., Wilhelm, R., et al. 2006, ApJ, 636, 804

[4] Allende Prieto, C., Fernández-Alvar, E., Schlesinger, K. J., et al. 2014, A&A, 568, A7

[5] Allison, R. J. 2012, MNRAS, 421, 3338

[6] Andrews, B. H., Weinberg, D. H., Johnson, J. A., Bensby, T., & Feltzing, S. 2012, AcA, 62, 269

[7] Asplund, M., Grevesse, N., Sauval, A. J., & Scott, P. 2009, ARA&A, 47, 481

[8] Aumer, M., Binney, J., & Schönrich, R. 2016, MNRAS, 462, 1697

[9] Bakos, J., Trujillo, I., & Pohlen, M. 2008, ApJ, 683, L103

[10] Barklem, P. S., Christlieb, N., Beers, T. C., et al. 2005, A&A, 439, 129

[11] Bastian, N., Gieles, M., Efremov, Y. N., & Lamers, H. J. G. L. M. 2005, A&A, 443, 79

[12] Behroozi, P. S., Wechsler, R. H., & Conroy, C. 2013, ApJ, 770, 57

[13] Bensby, T., Feltzing, S., & Oey, M. S. 2014, A&A, 562, A71

[14] Bergemann, M., Kudritzki, R.-P., Gazak, Z., Davies, B., & Plez, B. 2015, ApJ, 804, 113

[15] Bertelli, G., & Nasi, E. 2001, AJ, 121, 1013

[16] Bica, E., Dutra, C. M., Soares, J., & Barbuy, B. 2003, A&A, 404, 223

[17] Bik, A., Lamers, H. J. G. L. M., Bastian, N., Panagia, N., & Romaniello, M. 2003, A&A, 397, 473

[18] Binney, J., & Tremaine, S. 2008, Galactic Dynamics: (2nd ed.; Princeton, NJ: Princeton Univ. Press)

[19] Bird, J. C., Kazantzidis, S., Weinberg, D. H., et al. 2013, ApJ, 773, 43

[20] Blanco-Cuaresma, S., Soubiran, C., Heiter, U., et al. 2015, A&A, 577, A47

[21] Bland-Hawthorn, J., & Freeman, K. 2014, in The Origin of the Galaxy and Local Group, Saas-Fee Advanced Course, Vol. 37 (Berlin: Springer)

[22] Bland-Hawthorn, J., & Freeman, K. C. 2004, PASA, 21, 110

[23] Bland-Hawthorn, J., Karlsson, T., Sharma, S., Krumholz, M., & Silk, J. 2010a, ApJ, 721, 582

[24] Bland-Hawthorn, J., Krumholz, M. R., & Freeman, K. 2010b, ApJ, 713, 166

[25] Bland-Hawthorn, J., Sharma, S., & Freeman, K. 2014, in in EAS Publications Ser., Vol. 67, The Milky Way Unravelled by Gaia, ed. C. Soubrian et al. (Cambridge: Cambridge Univ. Press), 219

[26] Block, D. L., Bournaud, F., Combes, F., et al. 2006, Natur, 443, 832

[27] Borissova, J., Bonatto, C., Kurtev, R., et al. 2011, A&A, 532, A131

[28] Bournaud, F., Elmegreen, B. G., & Martig, M. 2009, ApJ, 707, L1

[29] Bovy, J. 2016a, ApJ, 817, 49

[30] Bovy, J., & Rix, H.-W. 2013, ApJ, 779, 115

[31] Bovy, J., Rix, H.-W., Liu, C., et al. 2012, ApJ, 753, 148

[32] Bovy, J., Rix, H.-W., Schlafly, E. F., et al. 2016b, ApJ, 823, 30

[33] Bovy, J., Nidever, D. L., Rix, H.-W., et al. 2014, ApJ, 790, 127

[34] Bragaglia, A., Gratton, R. G., Carretta, E., et al. 2012, A&A, 548, A122

[35] Brandner, W., Clark, J. S., Stolte, A., et al. 2008, A&A, 478, 137

[36] Brewer, J. M., Fischer, D. A., Basu, S., Valenti, J. A., & Piskunov, N. 2015, ApJ, 805, 126

[37] Bubar, E. J., & King, J. R. 2010, AJ, 140, 293

[38] Carretta, E., Bragaglia, A., Gratton, R., & Lucatello, S. 2009, A&A, 505, 139

[39] Casagrande, L., Schönrich, R., Asplund, M., et al. 2011, A&A, 530, A138

[40] Casey, A. R., Hogg, D. W., Ness, M., et al. 2016a, ArXiv e-prints, arXiv:1603.03040

[41] Casey, A. R., Hawkins, K., Hogg, D. W., et al. 2016b, ArXiv e-prints, arXiv:1609.02914

[42] Cheng, J. Y., Rockosi, C. M., Morrison, H. L., et al. 2012, ApJ, 752, 51

[43] Choi, J., Dotter, A., Conroy, C., et al. 2016, ApJ, 823, 102

[44] Chomiuk, L., & Povich, M. S. 2011, AJ, 142, 197

[45] Conroy, C., Gunn, J. E., & White, M. 2009, ApJ, 699, 486

[46] Cramer, H. 1946, Mathematical Methods of Statistics (Princeton, NJ: Princeton Univ. Press)

[47] Czekala, I., Andrews, S. M., Mandel, K. S., Hogg, D. W., & Green, G. M. 2015, ApJ, 812, 128

[48] Dalessandro, E., Miocchi, P., Carraro, G., Jílková, L., & Moitinho, A. 2015, MNRAS, 449, 1811

[49] De Silva, G. M., D'Orazi, V., Melo, C., et al. 2013, MNRAS, 431, 1005

[50] De Silva, G. M., Freeman, K. C., Asplund, M., et al. 2007b, AJ, 133, 1161

[51] De Silva, G. M., Freeman, K. C., & Bland-Hawthorn, J. 2009, PASA, 26, 11

[52] De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., Asplund, M., & Bessell, M. S. 2007a, AJ, 133, 694

[53] De Silva, G. M., Freeman, K. C., Bland-Hawthorn, J., et al. 2015, MNRAS, 449, 2604

[54] Dehnen, W. 2000, AJ, 119, 800

[55] Di Matteo, P., Haywood, M., Combes, F., Semelin, B., & Snaith, O. N. 2013, A&A, 553, A102

[56] Edvardsson, B., Andersen, J., Gustafsson, B., et al. 1993, A&A, 275, 101

[57] Eggen, O. J., Lynden-Bell, D., & Sandage, A. R. 1962, ApJ, 136, 748

[58] Elmegreen, B. G. 2002, ApJ, 577, 206

[59] Escala, A., & Larson, R. B. 2008, ApJ, 685, L31

[60] Espinoza, P., Selman, F. J., & Melnick, J. 2009, A&A, 501, 563

[61] Evans, II, N. J., Dunham, M. M., Jørgensen, J. K., et al. 2009, ApJS, 181, 321

[62] Everitt, B. S., Landau, S., Leese, M., & Stahl, D. 2010, Cluster Analysis (5th ed.; New York: Wiley)

[63] Feng, Y., & Krumholz, M. R. 2014, Natur, 513, 523

[64] Fischer, D. A., & Valenti, J. 2005, ApJ, 622, 1102

[65] Flynn, C., Holmberg, J., Portinari, L., Fuchs, B., & Jahreiß, H. 2006, MNRAS, 372, 1149

[66] Förster Schreiber, N. M., Genzel, R., Bouché, N., et al. 2009, ApJ, 706, 1364

[67] François, P., Monaco, L., Bonifacio, P., et al. 2016, A&A, 588, A7

[68] Frebel, A., & Norris, J. E. 2015, ARA&A, 53, 631

[69] Freeman, K., & Bland-Hawthorn, J. 2002, ARA&A, 40, 487

[70] Friel, E. D., Donati, P., Bragaglia, A., et al. 2014, A&A, 563, A117

[71] García Pérez, A. E., Allende Prieto, C., Holtzman, J. A., et al. 2016, AJ, 151, 144

[72] Genzel, R., Tacconi, L. J., Eisenhauer, F., et al. 2006, Natur, 442, 786

[73] Genzel, R., Tacconi, L. J., Kurk, J., et al. 2013, ApJ, 773, 68

[74] Ghez, A. M., Salim, S., Weinberg, N. N., et al. 2008, ApJ, 689, 1044

[75] Gillessen, S., Eisenhauer, F., Trippe, S., et al. 2009, ApJ, 692, 1075

[76] Gordon, K. D., Bailin, J., Engelbracht, C. W., et al. 2006, ApJ, 638, L87

[77] Gould, A., & Rix, H.-W. 2015, ArXiv e-prints, arXiv:1502.05709

[78] Grand, R. J. J., Kawata, D., & Cropper, M. 2015, MNRAS, 447, 4018

[79] Grand, R. J. J., Springel, V., Gómez, F. A., et al. 2016, MNRAS, 459, 199

[80] Gustafsson, B., Edvardsson, B., Eriksson, K., et al. 2008, A&A, 486, 951

[81] Halle, A., Di Matteo, P., Haywood, M., & Combes, F. 2015, A&A, 578, A58

[82] Hansen, C. J., Ludwig, H.-G., Seifert, W., et al. 2015, AN, 336, 665

[83] Harris, W. E., & Pudritz, R. E. 1994, ApJ, 429, 177

[84] Hauschildt, P. H., Allard, F., Ferguson, J., Baron, E., & Alexander, D. R. 1999, ApJ, 525, 871

[85] Hawkins, K., Jofré, P., Masseron, T., & Gilmore, G. 2015, MNRAS, 453, 758

[86] Hayden, M. R., Bovy, J., Holtzman, J. A., et al. 2015, ApJ, 808, 132

[87] Haywood, M. 2008, MNRAS, 388, 1175

[88] Haywood, M., Di Matteo, P., Lehnert, M. D., Katz, D., & Gómez, A. 2013, A&A, 560, A109

[89] Heiter, U., Jofré, P., Gustafsson, B., et al. 2015a, A&A, 582, A49

[90] Hernandez, X., Valls-Gabaud, D., & Gilmore, G. 2000, MNRAS, 316, 605

[91] Hogg, D. W., Casey, A. R., Ness, M., Rix, H.-W., & Foreman-Mackey, D. 2016, ArXiv e-prints, arXiv:1601.05413

[92] Holtzman, J. A., Shetrone, M., Johnson, J. A., et al. 2015, AJ, 150, 148

[93] Iwamoto, K., Brachwitz, F., Nomoto, K., et al. 1999, ApJS, 125, 439

[94] Jofré, P., Heiter, U., Soubiran, C., et al. 2014, A&A, 564, A133

[95] —. 2015, A&A, 582, A81

[96] Johnson, L. C., Seth, A. C., Dalcanton, J. J., et al. 2016, ApJ, 827, 33

[97] Jones, T. A., Swinbank, A. M., Ellis, R. S., Richard, J., & Stark, D. P. 2010, MNRAS, 404, 1247

[98] Kafle, P. R., Sharma, S., Lewis, G. F., & Bland-Hawthorn, J. 2012, ApJ, 761, 98

[99] Kalirai, J. S., Richer, H. B., Fahlman, G. G., et al. 2001, AJ, 122, 266

[100] Karakas, A. I., & Lattanzio, J. C. 2014, PASA, 31, 30

[101] Karlsson, T., Bland-Hawthorn, J., Freeman, K. C., & Silk, J. 2012, ApJ, 759, 111

[102] Kawata, D., Grand, R. J. J., Gibson, B. K., et al. 2017, MNRAS, 464, 702

[103] Kennicutt, R. C., & Evans, N. J. 2012, ARA&A, 50, 531

[104] Kennicutt, Jr., R. C. 1998, ApJ, 498, 541

[105] Kirby, E. N., Bullock, J. S., Boylan-Kolchin, M., Kaplinghat, M., & Cohen, J. G. 2014, MNRAS, 439, 1015

[106] Kirby, E. N., Guhathakurta, P., Simon, J. D., et al. 2010, ApJS, 191, 352

[107] Klypin, A., Zhao, H., & Somerville, R. S. 2002, ApJ, 573, 597

[108] Kobayashi, C., Karakas, A. I., & Umeda, H. 2011, MNRAS, 414, 3231

[109] Kobayashi, C., Umeda, H., Nomoto, K., Tominaga, N., & Ohkubo, T. 2006, ApJ, 653, 1145

[110] Kollmeier, J. A., & Gould, A. 2007, ApJ, 664, 343

[111] Koposov, S. E., Glushkova, E. V., & Zolotukhin, I. Y. 2008, A&A, 486, 771

[112] Koposov, S. E., Rix, H.-W., & Hogg, D. W. 2010, ApJ, 712, 260

[113] Kordopatis, G., Binney, J., Gilmore, G., et al. 2015, MNRAS, 447, 3526

[114] Kroupa, P. 2002, Sci, 295, 82

[115] Kruijssen, J. M. D. 2012, MNRAS, 426, 3008

[116] Krumholz, M. R., Dekel, A., & McKee, C. F. 2012, ApJ, 745, 69

[117] Krumholz, M. R., & McKee, C. F. 2005, ApJ, 630, 250

[118] Kurucz, R. L. 1970, SAOSR, 309

[119] —. 1993, SYNTHE Spectrum Synthesis Programs and Line Data, ed. R. L. Kurucz (Cambridge, MA: Smithsonian Astrophysical Observatory) http://adsabs.harvard.edu/abs/1993sssp.book.....K

[120] Kurucz, R. L. 1996, in ASP Conf. Ser., Vol. 108, M.A.S.S., Model Atmospheres and Spectrum Synthesis, ed. S. J. Adelman, F. Kupka, & W. W. Weiss (San Francisco: ASP), 2

[121] Kurucz, R. L. 2003, in IAU Symposium, Vol. 210, Modelling of Stellar Atmospheres, ed. N. Piskunov, W. W. Weiss, & D. F. Gray (San Francisco: ASP), 45

[122] —. 2005, MSAIS, 8, 14

[123] Kurucz, R. L., & Avrett, E. H. 1981, SAOSR, 391

[124] Lada, C. J., & Lada, E. A. 2003, ARA&A, 41, 57

[125] Lee, Y. S., Beers, T. C., Allende Prieto, C., et al. 2011, AJ, 141, 90

[126] Lee, Y. S., Beers, T. C., Masseron, T., et al. 2013, AJ, 146, 132

[127] Letarte, B., Hill, V., Tolstoy, E., et al. 2010, A&A, 523, A17

[128] Limongi, M., & Chieffi, A. 2003, ApJ, 592, 404

[129] Lind, K., Melendez, J., Asplund, M., Collet, R., & Magic, Z. 2013, A&A, 554, A96

[130] Lindegren, L., & Feltzing, S. 2013, A&A, 553, A94

[131] Liu, F., Yong, D., Asplund, M., Ramírez, I., & Meléndez, J. 2016, MNRAS, 457, 3934

[132] Livermore, R. C., Jones, T., Richard, J., et al. 2012, MNRAS, 427, 688

[133] Loebman, S. R., Roškar, R., Debattista, V. P., et al. 2011, ApJ, 737, 8

[134] Luo, A.-L., Zhao, Y.-H., Zhao, G., et al. 2015, RAA, 15, 1095

[135] MacFarlane, B. A., Gibson, B. K., & Flynn, C. M. L. 2016, in ASP Conf. Ser., Vol. 507, Multi-Object Spectroscopy in the Next Decade: Big Questions, Large Surveys, and Wide Fields, ed. I. Skillen, M. Barcells, & S. P. Trager (San Francisco: ASP), 79

[136] Majewski, S. R., APOGEE Team, & APOGEE-2 Team. 2016, AN, 337, 863

[137] Marino, A. F., Villanova, S., Milone, A. P., et al. 2011, ApJ, 730, L16

[138] Martell, S. L., & Grebel, E. K. 2010, A&A, 519, A14

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

[140] Martig, M., Minchev, I., & Flynn, C. 2014, MNRAS, 443, 2452

[141] Martig, M., Rix, H.-W., Aguirre, V. S., et al. 2015, MNRAS, 451, 2230

[142] Martig, M., Fouesneau, M., Rix, H.-W., et al. 2016, MNRAS, 456, 3655

[143] Masseron, T., & Gilmore, G. 2015, MNRAS, 453, 1855

[144] McLachlan, G., & Peel, D. 2000, Finite Mixture Models (1st ed.; New York: Wiley)

[145] Mészáros, S., & Allende Prieto, C. 2013a, MNRAS, 430, 3285

[146] Mészáros, S., Holtzman, J., García Pérez, A. E., et al. 2013b, AJ, 146, 133

[147] Minchev, I., Chiappini, C., & Martig, M. 2013, A&A, 558, A9

[148] —. 2014, A&A, 572, A92

[149] Minchev, I., & Famaey, B. 2010, ApJ, 722, 112

[150] Mitschang, A. W., De Silva, G., Sharma, S., & Zucker, D. B. 2013, MNRAS, 428, 2321

[151] Mitschang, A. W., De Silva, G., Zucker, D. B., et al. 2014, MNRAS, 438, 2753

[152] Moore, B., Ghigna, S., Governato, F., et al. 1999, ApJ, 524, L19

[153] Ness, M., Hogg, D. W., Rix, H.-W., Ho, A. Y. Q., & Zasowski, G. 2015, ApJ, 808, 16

[154] Nidever, D. L., Bovy, J., Bird, J. C., et al. 2014, ApJ, 796, 38

[155] Nomoto, K., Kobayashi, C., & Tominaga, N. 2013, ARA&A, 51, 457

[156] Odenkirchen, M., Grebel, E. K., Dehnen, W., et al. 2003, AJ, 126, 2385

[157] Önehag, A., Gustafsson, B., & Korn, A. 2014, A&A, 562, A102

[158] Peebles, P. J. E. 1971, Physical Cosmology (Princeton, NJ: Princeton Univ. Press)

[159] Pompéia, L., Hill, V., Spite, M., et al. 2008, A&A, 480, 379

[160] Porras, A., Christopher, M., Allen, L., et al. 2003, AJ, 126, 1916

[161] Press, W. H., & Schechter, P. 1974, ApJ, 187, 425

[162] Prugniel, P., Vauglin, I., & Koleva, M. 2011, A&A, 531, A165

[163] Quillen, A. C., Anguiano, B., De Silva, G., et al. 2015, MNRAS, 450, 2354

[164] Randich, S., Gilmore, G., & Gaia-ESO Consortium. 2013, Msngr, 154, 47

[165] Rao, C. R. 1945, Bull. Calcutta Math. Soc., 37, 81

[166] Recio-Blanco, A., Bijaoui, A., & de Laverny, P. 2006, MNRAS, 370, 141

[167] Recio-Blanco, A., de Laverny, P., Allende Prieto, C., et al. 2016, A&A, 585, A93

[168] Reddy, B. E., Lambert, D. L., & Allende Prieto, C. 2006, MNRAS, 367, 1329

[169] Reid, M. J., Menten, K. M., Brunthaler, A., et al. 2014, ApJ, 783, 130

[170] Rix, H.-W., & Bovy, J. 2013, A&A Rev., 21, 61

[171] Rix, H.-W., & Rieke, M. J. 1993, ApJ, 418, 123

[172] Rix, H.-W., Ting, Y.-S., Conroy, C., & Hogg, D. W. 2016, ApJ, 826, L25

[173] Robitaille, T. P., & Whitney, B. A. 2010, ApJ, 710, L11

[174] Roederer, I. U., & Thompson, I. B. 2015, MNRAS, 449, 3889

[175] Roškar, R., Debattista, V. P., Quinn, T. R., Stinson, G. S., & Wadsley, J. 2008, ApJ, 684, L79

[176] Roškar, R., Debattista, V. P., Quinn, T. R., & Wadsley, J. 2012, MNRAS, 426, 2089

[177] Schiavon, R. P., Zamora, O., Carrera, R., et al. 2016, MNRAS, arXiv:1606.05651

[178] Schönrich, R., & Binney, J. 2009, MNRAS, 396, 203

[179] Schönrich, R., & McMillan, P. 2016, ArXiv e-prints, arXiv:1605.02338

[180] Schruba, A., Leroy, A. K., Walter, F., et al. 2011, AJ, 142, 37

[181] SDSS Collaboration, Albareti, F. D., Allende Prieto, C., et al. 2016, ArXiv e-prints, arXiv:1608.02013

[182] Searle, L., & Zinn, R. 1978, ApJ, 225, 357

[183] Sellwood, J. A., & Binney, J. J. 2002, MNRAS, 336, 785

[184] Shapiro, K. L., Genzel, R., & Förster Schreiber, N. M. 2010, MNRAS, 403, L36

[185] Sharma, S., & Johnston, K. V. 2009, ApJ, 703, 1061

[186] Sheinis, A., Anguiano, B., Asplund, M., et al. 2015, ArXiv e-prints, arXiv:1509.00129

[187] Smiljanic, R., Korn, A. J., Bergemann, M., et al. 2014, A&A, 570, A122

[188] Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645

[189] Tabernero, H. M., Montes, D., & González Hernández, J. I. 2012, A&A, 547, A13

[190] Tabernero, H. M., Montes, D., Gonzalez Hernandez, J. I., & Ammler-von Eiff, M. 2014, ArXiv e-prints, arXiv:1409.2348

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

[192] Ting, Y.-S., Conroy, C., & Rix, H.-W. 2016a, ApJ, 816, 10

[193] —. 2016b, ApJ, 826, 83

[194] Ting, Y.-S., De Silva, G. M., Freeman, K. C., & Parker, S. J. 2012b, MNRAS, 427, 882

[195] Ting, Y.-S., Freeman, K. C., Kobayashi, C., De Silva, G. M., & Bland-Hawthorn, J. 2012a, MNRAS, 421, 1231

[196] Tolstoy, E., Hill, V., & Tosi, M. 2009, ARA&A, 47, 371

[197] Toomre, A. 1964, ApJ, 139, 1217

[198] van Dokkum, P. G., Leja, J., Nelson, E. J., et al. 2013, ApJ, 771, L35

[199] VandenBerg, D. A., Bergbusch, P. A., Dotter, A., et al. 2012, ApJ, 755, 15

[200] Veneziani, M., Elia, D., Noriega-Crespo, A., et al. 2013, A&A, 549, A130

[201] Venn, K. A., & Hill, V. M. 2008, Msngr, 134, 23

[202] Venn, K. A., Irwin, M., Shetrone, M. D., et al. 2004, AJ, 128, 1177

[203] Ventura, P., Karakas, A. I., Dell'Agli, F., et al. 2015, MNRAS, 450, 3181

[204] Webb, J. J., Harris, W. E., Sills, A., & Hurley, J. R. 2013, ApJ, 764, 124

[205] Weisz, D. R., Dolphin, A. E., Skillman, E. D., et al. 2015, ApJ, 804, 136

[206] Wilkinson, M. I., & Evans, N. W. 1999, MNRAS, 310, 645

[207] Woosley, S. E., & Weaver, T. A. 1995, ApJS, 101, 181

[208] Wuyts, S., Förster Schreiber, N. M., Genzel, R., et al. 2012, ApJ, 753, 114

[209] Xue, X. X., Rix, H. W., Zhao, G., et al. 2008, ApJ, 684, 1143

[210] Yang, L., Kirby, E. N., Guhathakurta, P., Peng, E. W., & Cheng, L. 2013, ApJ, 768, 4

[211] Yanny, B., Rockosi, C., Newberg, H. J., et al. 2009, AJ, 137, 4377

[212] Zasowski, G., Johnson, J. A., Frinchaboy, P. M., et al. 2013, AJ, 146, 81

[213] Zhang, L., Rix, H.-W., van de Ven, G., et al. 2013, ApJ, 772, 108

[214] Zheng, Z., Thilker, D. A., Heckman, T. M., et al. 2015, ApJ, 800, 120