Astronomy & Astrophysics manuscript no. tsne_abunds
February 1, 2018

Dissecting stellar chemical abundance space with t-SNE

F. Anders[1,2], C. Chiappini[1,2], B. X. Santiago[3,2], G. Matijevič[1], A. B. Queiroz[3,2], B. Barbuy[4,2]

[1] Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany
e-mail: fanders@aip.de
[2] Laboratório Interinstitucional de e-Astronomia, - LIneA, Rua Gal. José Cristino 77, Rio de Janeiro, RJ - 20921-400, Brazil
[3] Instituto de Física, Universidade Federal do Rio Grande do Sul, Caixa Postal 15051, Porto Alegre, RS - 91501-970, Brazil
[4] Universidade de São Paulo, IAG, Rua do Matão 1226, Cidade Universitária, 05508-900, São Paulo, Brazil

Received February 1, 2018; accepted ...

ABSTRACT

2D chemical-abundance diagrams are important diagnostics of chemo-dynamical evolution in galaxies. However, in the era of industrial Galactic astronomy opened by multi-object spectroscopic stellar surveys, the sample sizes and the number of available abundances have reached dimensions in which it has become difficult to make use of all the available information in an effective manner. Here we demonstrate the use of t-distributed stochastic neighbour embedding (t-SNE) in spectroscopic stellar abundance space of the solar vicinity. By reanalysing high-resolution high-signal-to-noise solar-neighbourhood samples with t-SNE, we find clearer chemical separations of the high- and low-[α/Fe] disc sequences, hints for multiple populations in the high-[α/Fe] population, and a number of chemically peculiar stars, some of which were likely born in dwarf galaxies, others possibly in the Galactic bulge.

Key words. Galaxy: general – Galaxy: abundances – Galaxy: disk – Galaxy: stellar content – Stars: abundances

1. Introduction

One of the major goals of modern Galactic astrophysics is to infer the formation history of our Milky Way. To achieve this goal it is necessary to obtain precise 6D stellar kinematics as well as detailed chemical abundances for large stellar samples. This chemo-kinematical map of the Galactic stellar populations can then be compared to predictions of various Milky-Way models, eventually unveiling the star-formation and dynamical history of our Galaxy.

Massive spectroscopic observing campaigns such as RAVE (Steinmetz et al. 2006), SEGUE (Yanny et al. 2009), APOGEE (Majewski et al. 2017), LAMOST (Deng et al. 2012), GALAH (Martell et al. 2017) and the Gaia-ESO survey (Gilmore 2012) have in the past decade increased both the volume coverage and the statistical sample sizes by more than two orders of magnitude, to $5 \cdot 10^6$ stars distributed from the solar vicinity to the far side of the Galactic bulge and the outer halo. In spite of this recent conquista of the Milky Way in terms of number of spectroscopically analysed stars, detailed multi-abundance chemo-kinematical studies of the immediate solar vicinity (Edvardsson et al. 1993; Fuhrmann 1998, 2011; Fuhrmann et al. 2017; Adibekyan et al. 2012; Bensby et al. 2014; Nissen 2015, 2016; Delgado Mena et al. 2017b, e.g.) remain at least equally important for Galactic Archaeology (see Lindegren & Feltzing 2013 for a quantitative analysis).

The wealth of new data, especially the high dimensionality of chemo-kinematics space, requires new statistical analysis methods to efficiently constrain detailed Milky-Way formation models (including e.g. stellar evolution, stellar chemical feedback, chemical evolution, and dynamical evolution). Traditionally, the metallicity distribution function and 2D chemical-abundance diagrams ([X/Fe] vs. [Fe/H]), and abundance gradients have been used to constrain the chemical evolution of stellar populations (e.g. Pagel 2009). On the other hand, it is also possible to *define* a stellar population by chemistry (e.g. carbon-enhanced metal-poor stars - Beers & Christlieb 2005; the chemical thick disc - Gratton et al. 1996; Fuhrmann 1998; high-[α/Fe] metal-rich stars - Adibekyan et al. 2011), and to then study their structural and chemo-kinematic properties in detail. This is usually done in a simple fashion, by looking at only one 2D abundance diagram. In this paper we explore the possibility of combining the information contained in various measured abundance ratios using t-distributed stochastic neighbour embedding (t-SNE) to define more robust subpopulations and better identify outliers.

In astronomical applications, t-SNE has mainly been used to identify objects with peculiar spectra (e.g. Matijevič et al. 2017; Valentini et al. 2017; Traven et al. 2017; Reis et al. 2017). During the writing of this paper, Kos et al. (2017) demonstrated in a complementary analysis that abundance-space t-SNE is indeed a reliable chemical-tagging tool: the authors were able to recover 7 out of 9 known open and globular clusters with high efficiency and low contamination using 13 chemical abundances from the GALAH survey (Martell et al. 2017), and they also found two new field member stars to known clusters with this technique.

The paper is structured as follows: Sec. 2 introduces t-SNE. Sections 3 and **??** describe and discuss the results for the high-resolution spectroscopic solar-vicinity surveys of Delgado Mena et al. (2017b) and Bensby et al. (2014). We reconsider possible caveats of our results in Sec. **??** and finish with a summary and conclusions in Sec. 4.

2. Dissecting chemistry space with t-SNE

Interpreting the multi-dimensional abundance distributions determined by spectroscopic surveys is not a trivial task, since different abundance diagrams contain different nucleosynthetic information and may be affected by different observational errors. A convenient way to simplify this problem is dimension-

ality reduction, i.e. the projection of the N-dimensional abundance space onto a lower-dimensional space in which the chemical similarity between two stars is reflected by their distance in that space. Possibly the best-known such method is called principal component analysis (PCA), widely used also in astronomical literature. For highly-correlated datasets such as spectral pixel spaces or chemical-abundance spaces, however, more sophisticated non-linear methods like IsoMap or locally linear embedding are known to perform much better (e.g. Matijevič et al. 2012; Ivezić et al. 2013).

In this paper, we reanalyse the high-resolution spectroscopic solar-vicinity surveys of Bensby et al. (2014) and Delgado Mena et al. (2017b) using a machine-learning algorithm called t-distributed stochastic neighbour embedding (t-SNE; Hinton & Roweis 2003; van der Maaten & Hinton 2008). This method is widely used in big-data analytics, and is able to efficiently project complex datasets onto a 2D plane in which the proximity between similar data points is preserved. We use the python implementation of t-SNE included in the `scikit-learn` package (Pedregosa et al. 2012) and refer to the original papers and the online documentation for details about the method and code. In short, the advantage of using t-SNE over other manifold-learning techniques is that it performs much better in revealing structure at many different scales (van der Maaten & Hinton 2008; Matijevič et al. 2017), which is a necessary feature when looking for chemical substructure in the Galactic disc.

How t-SNE works: For a given set of N high-dimensional datapoints $\mathbf{x}_1, \ldots, \mathbf{x}_N$ (images, spectra, or in our case chemical-abundance vectors), t-SNE first computes pairwise similarity probabilities p_{ij} for the points $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$p_{j|i} = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||\mathbf{x}_i - \mathbf{x}_k||^2/2\sigma_i^2)}.$$

To circumvent problems with outliers, the symmetrised similarity of x_j and x_i is defined as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}.$$

In the next step, t-SNE attepts to learn a d-dimensional map $\mathbf{y}_1, \ldots, \mathbf{y}_N$ (in general $d = 2$) that reflects the similarities p_{ij} similarities between two points $\mathbf{y}_i$ and $\mathbf{y}_j$ in the low-dimensional map, defined as

$$q_{ij} = \frac{(1 + ||\mathbf{y}_i - \mathbf{y}_j||^2)^{-1}}{\sum_{k\neq m}(1 + ||\mathbf{y}_k - \mathbf{y}_m||^2)^{-1}}.$$

This metric uses Student's t distribution to avoid crowding problems in the low-dimensional map (van der Maaten & Hinton 2008). Starting from a random Gaussian distribution in the d-dimensional map, the locations of the points $\mathbf{y}_i$ are determined by minimizing the Kullback–Leibler divergence (Kullback & Leibler 1951) between the low- and high-dimensional similarity distributions Q and P:

$$KL(P||Q) = \sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}},$$

using a gradient-descent method. The result of this optimization is a 2D (or 3D) map that reflects the similarities between the high-dimensional inputs (see e.g. Fig. 2).

The method has one main parameter, the so-called perplexity, p, which governs the bandwidth of the Gaussian kernels σ_i appearing in the similarities p_{ij}. As a result, the bandwidth is adapted to the density of the data: smaller values of σ_i are used in denser parts of the data space. The perplexity parameter can be thought of as a guess about the number of close neighbors each point has, and therefore the ideal value for p depends on the sample size. A change in perplexity has in many cases a complex effect on the resulting map, and different values for p should be explored (Wattenberg et al. 2016).

Recently, Linderman & Steinerberger (2017) demonstrated that two other hyper-parameters of t-SNE can be chosen optimally: the learning rate should be set to $\sim$ 1, and the early-exaggeration parameter should be set to $\sim$ 0.1 times the sample size. In the following, we use these recommendations.

In addition, t-SNE, as a genuine machine-learning technique, does have two drawbacks that are relevant for our science case. First, it does not account for individual uncertainties, and may therefore be affected by extremely heteroscedastic errors. Secondly, its current implementations do not allow to treat missing data, so that any star with a missing individual abundance measurement has to be excluded.

3. Re-analysing the HARPS GTO sample

In an extensive series of papers, Adibekyan et al. (2011, 2012); Delgado Mena et al. (2014, 2015); Bertran de Lis et al. (2015); Suárez-Andrés et al. (2017); Delgado Mena et al. (2017b,a) studied the chemical abundances of a sample of 1111 solar-vicinity FGK stars using the very high resolution of the HARPS spectrograph ($R \sim 115,000$). This sample mostly contains metal-rich warm dwarf and subgiant stars, but also includes a wide range of effective temperatures, gravities and metallicities. The HARPS sample initially served to detect and characterise exoplanets and may therefore contain some metallicity-related selection bias; however, e.g. Anders et al. (2014) have shown that the HARPS metallicity distribution (MDF) matches the MDF of high-quality local ($d < 1$ kpc) APOGEE red-giant stars that could be considered less chemically biased.

Delgado Mena et al. (2017b) recently reanalysed this sample, employing a revised linelist (Tsantaki et al. 2013), improving the effective temperature calibration, and correcting spectroscopic gravities using the *Hipparcos* parallaxes of van Leeuwen (2007). They report chemical abundances for Mg, Al, Si, Ca, Ti, Fe, Cu, Zn, Sr, Y, Zr and Ba for 1059 stars (Ce, Nd and Eu are available for a substantial subset of these), derived using standard Local Thermodynamic Equilibrium (LTE) analysis using ARES to measure equivalent widths and MOOG to measure abundances by comparing to Kurucz ATLAS9 atmospheres. In this section we test the performance of abundance-space t-SNE on this most recent HARPS GTO sample compilation. The high number of measured abundances, in conjunction with the high precision of the measurements and the reasonable sample size, makes the HARPS sample an ideal test case for our machine-learning algorithm.

Our first tests showed that, in order to obtain reliable t-SNE abundance maps, the sample needed to be analysed in a more restricted temperature range, because certain abundance trends seem to be dominated by underlying temperature trends. Therefore, similar to Delgado Mena et al. (2017b), we chose an effective temperature range of 5300 K $< T_{\rm eff} <$ 6000 K for our analysis. We furthermore restricted surface gravities to $3 < \log g_{\rm HIP} < 5$, and required successful abundance determination for Mg, Al, Si, Ca, TiI, Fe, Cu, Zn, Sr, Y, ZrII, Ce and Ba that we use as input for t-SNE, leaving us with 533 stars.[1] In

[1] Carbon and oxygen estimates are available from previous studies (Suárez-Andrés et al. 2017; Bertran de Lis et al. 2015), but since they

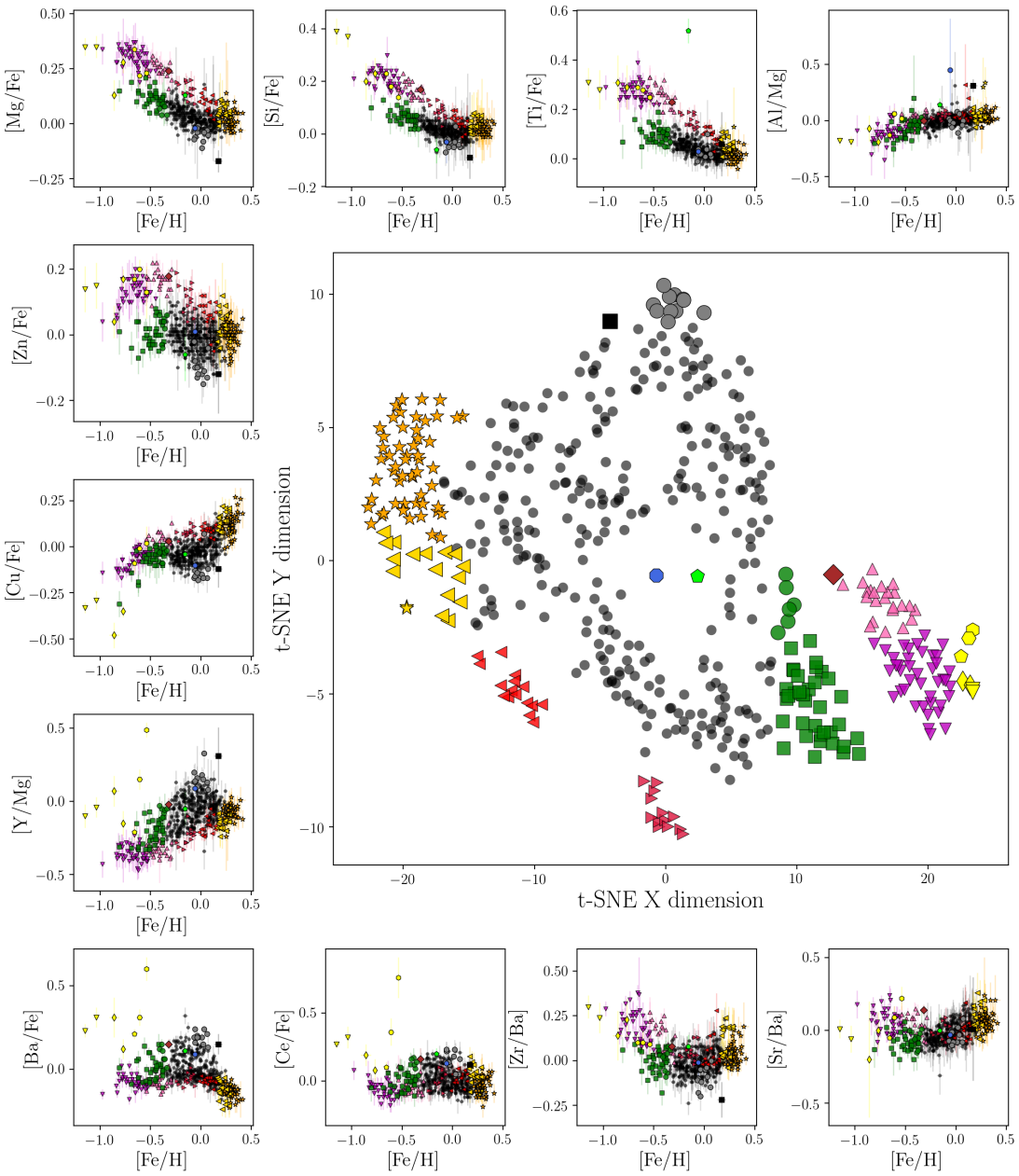

Fig. 1. Illustration of how t-SNE works in abundance space, using the Delgado Mena et al. (2017b) sample. The small panels show eleven of the possible $\sim 20,000$ abundance diagrams that can be created from 13 elements. The resulting reference t-SNE projection of the full abundance space is shown in the big panel, and several identified subgroups are indicated.

our final sample of 530 stars we also discarded 3 stars for which our age determination code, `StarHorse` (Santiago et al. 2016; Queiroz et al. 2017), did not converge. We verified that these choices do not significantly affect the resulting t-SNE maps.

The chemical abundances were complemented by astrometric data (parallaxes, proper motions) from the Gaia/TGAS catalogue Gaia Collaboration et al. (2016), or when these were unavailable (135/1059 stars), from the re-reduced *Hipparcos* data (van Leeuwen 2007). **DESCRIBE StarHorse RUN HERE**.

Fig. 2 again shows our reference t-SNE map for the HARPS sample, but now colour-coded by chemical-abundance ratios, stellar parameters, ages and kinematics. The panels in the first three rows show how t-SNE is grouping the stars with similar abundances in the two-dimensional plane. The panels coloured as a function of stellar parameters demonstrate that the sample is not subject to major systematic abundance shifts, but does show some residual trends with effective temperature, since it preferentially groups cooler stars in slightly different regions of the t-SNE map than hotter ones. Because part of this effect may be due to chemical evolution rather than systematic abundance errors, we decided not to apply any ad-hoc corrections to the abundances.

are based on previous stellar parameter estimates, we decided to only use them in the interpretation. We also did not use Nd and Eu in the t-SNE run, because they were only available for about half of the sample (stars with the highest signal-to-noise ratios). We do, however, use the Nd and Eu results in the interpretation, whenever they are available.

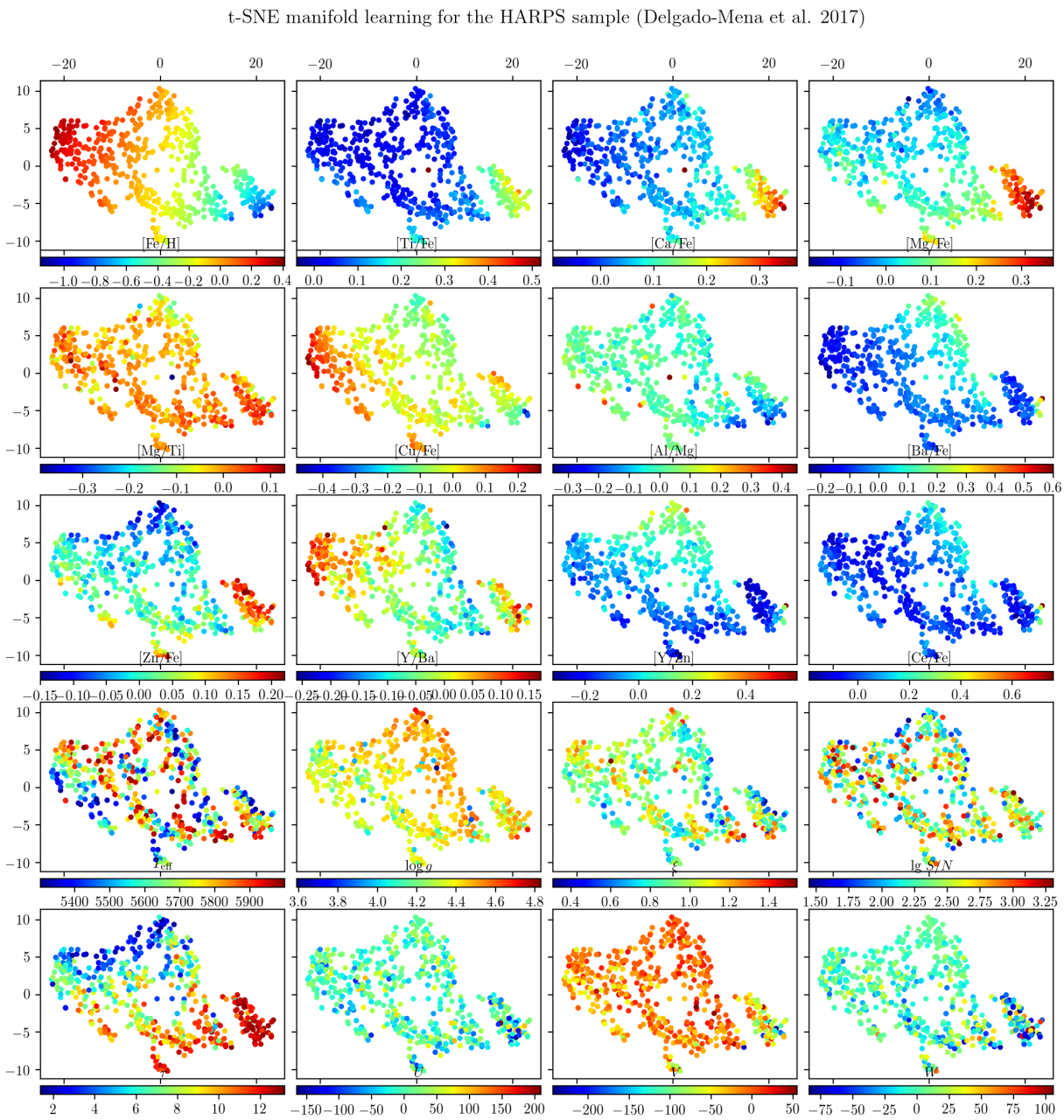

Fig. 2. Fiducial t-SNE projection (p = 40) of the Delgado Mena et al. (2017b) sample, colour-coded by chemical abundances (top three rows), stellar atmospheric parameters and signal-to-noise ratio (fourth row), age (fifth row, first panel) and UVW velocities (fifth row). We note that only [Fe/H] and the [X/Fe] ratios were used as input for the t-SNE run. The distinct populations appearing in these diagrams are studied in detail in Fig. 7.

We identified some of the substructures that appear in Fig. 2 already in Fig. 1. Fig. 3 shows the corresponding [X/Fe] abundance trends versus proton number for each of those substructures.

We now proceed to the discussion of these results.

3.1. The overall appearance of the t-SNE map

3.2. The robustness of the t-SNE results

We tested the robustness of our reference map with a simple Monte-Carlo experiment: For each star, we created 50 mock stars with abundances drawn from a multi-dimensional Gaussian distribution centered on the measured abundance, and variance corresponding to the measured abundance uncertainties. Although t-SNE does not take into account uncertainties in the data, this

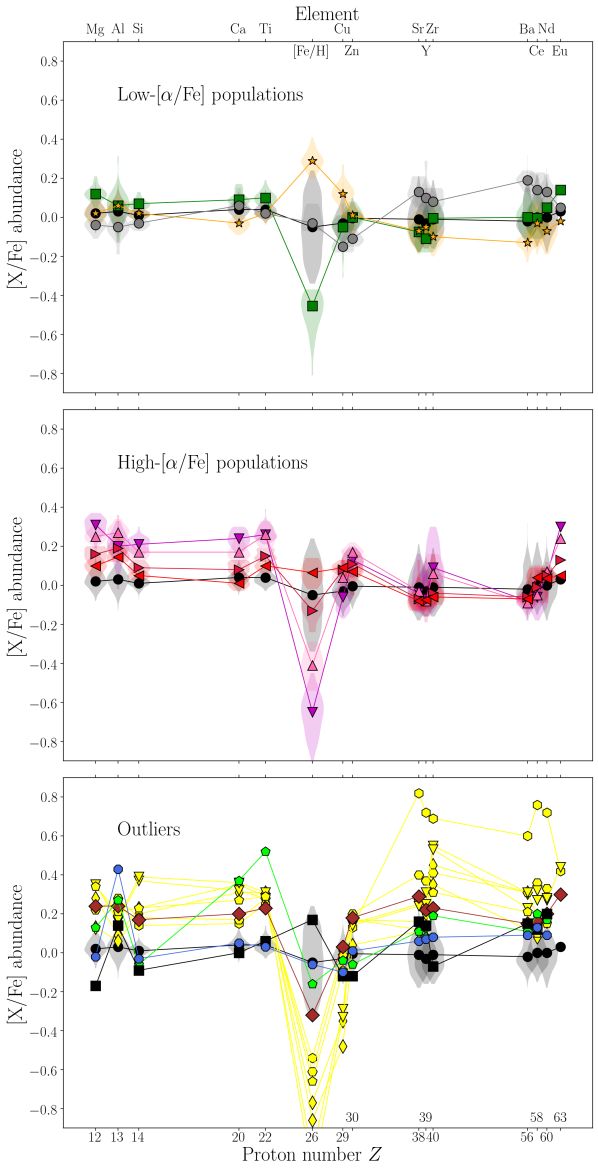

Fig. 3. Chemical-abundance patterns relative to iron for the t-SNE-selected subsamples of the HARPS survey. For visibility, we show only the median abundance ratios of each group.

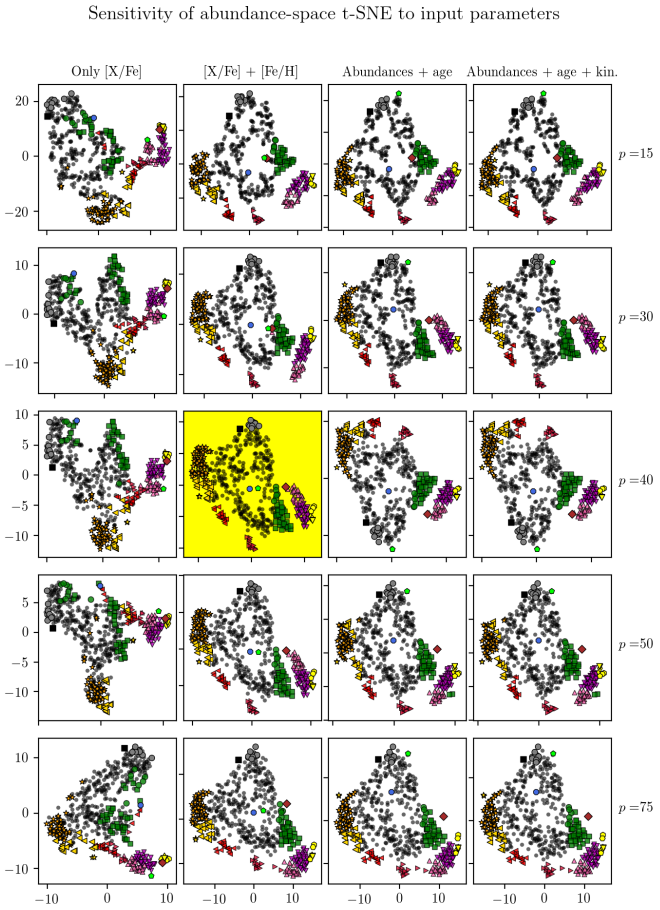

Fig. 4. t-SNE representations of the chrono-chemo-kinematics space spanned by the Delgado Mena et al. (2017b) sample. Each column row represents a combination of input information, while each row corresponds to a particular perplexity value, as indicated on the right side of the figure. The panel highlighted in yellow represents the results that we analyse in detail in this paper by defining chemical subpopulations based on this map.

procedure allows us to assure that the groups that we identified in the t-SNE map are not due to chance groupings.[2]

3.3. The thin-thick disc dichotomy

As discussed in the works of Adibekyan et al. (2011, 2012) or Delgado Mena et al. (2017b), there is a clear discontinuity between the high- and the low-[α/Fe] sequences in the [Mg/Fe] vs. [Fe/H] diagram. This discontinuity is reflected in a very clear manner in the t-SNE projection: We find a clear and obvious gap between the chemical thin- and thick-disc populations in the t-SNE diagram that remains very robust for different choices of the t-SNE hyper-parameters. Primarily, this means that the chemical patterns of thin and thick disc are indeed distinct, and can be disentangled by high-resolution spectroscopy. Secondly, our analysis of the full chemical information results in a much more accurate division of the chemically-thin and thick populations. Indeed, if one only relies on one diagnostic, such as the [Mg/Fe] vs. [Fe/H] diagram (Adibekyan et al. 2011; Delgado Mena et al. 2017b), several thin-disc stars would (most probably incorrectly) be identified as belonging chemical thick disc (see Fig. 1).

[2] In general, adding uncertainties to measured (i.e. already noisy) data will blur the true values even more. This means that if a signal disappears in the Monte-Carlo test, the test does not rule out the existence of the signal. On the other hand, if the signal persists, the signal is very unlikely to be due to a chance grouping.

3.4. Sub-populations and chemically peculiar stars

Thick-disc sub-populations: Adibekyan et al. (2011) first discovered a clear discontinuity between the metal-poor and metal-rich [α/Fe]-enhanced (or hαmr) disc populations. In our t-SNE analysis of the Bensby et al. (2014) sample, similar to the original paper, we only see a hint of a difference between the two populations (dubbed Thick Disc I and II in Fig. **??**). The bimodality is better seen when slightly higher perplexity values are chosen ($p \sim 30$). Even if ages and/or kinematics are included as additional dimensions in the analysis, this picture does not change much.

Thin-thick transition stars:

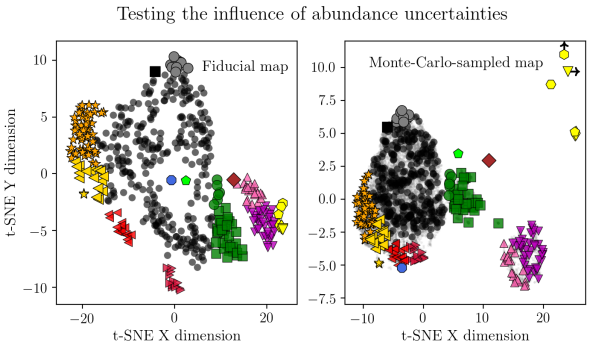

Fig. 5. Robustness test of our t-SNE-selected subsamples. The right panel shows the fiducial map, while the left panel shows the result of our Monte-Carlo test. For each star, 50 random stars were drawn from the a Gaussian centered on the measured abundance, and with widths corresponding to the measured uncertainties. The resulting map demonstrates that our selected subgroups are robust to measurement errors.

Super-metal-rich stars:
The metal-poor thin disc:
Satellite debris

3.5. Abundance trends with age

3.6. Kinematic trends

4. Conclusions

YES WE CAN use t-SNE to better define subpopulations in abundance space. However, the non-parametric non-linear behaviour of the technique makes it difficult to estimate the significance of found subgroups or clusters. The method could, however, be coupled to a genuine cluster finding algorithm.

Potential for weak chemical tagging demonstrated in this paper; the viability of t-SNE for strong chemical tagging (finding dispersed members of open clusters) is still not completely clear, but see Kos et al. (2017).

It is better to confine the analysis to narrow regions in atmospheric-parameter space to avoid spurious abundance trends induced by differences in atmospheric parameters.

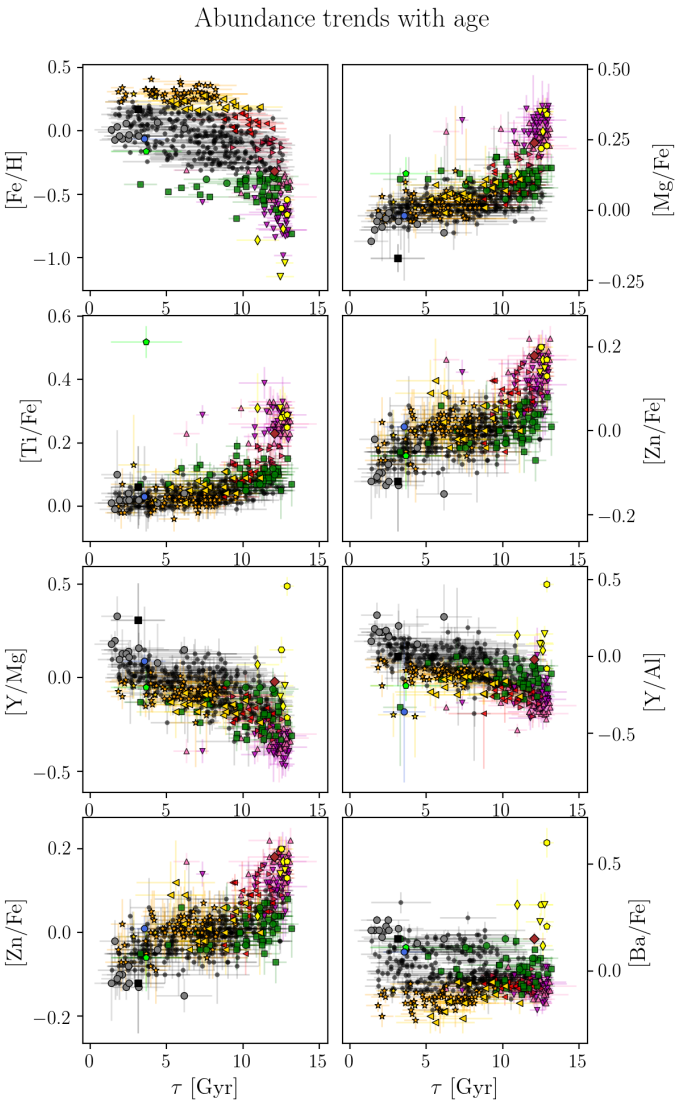

Fig. 6. Abundance trends with stellar age, measured with the `StarHorse` code (Queiroz et al. 2017).

References

Adibekyan, V. Z., Santos, N. C., Sousa, S. G., & Israelian, G. 2011, A&A, 535, L11
Adibekyan, V. Z., Sousa, S. G., Santos, N. C., et al. 2012, A&A, 545, A32
Anders, F., Chiappini, C., Santiago, B. X., et al. 2014, A&A, 564, A115
Beers, T. C. & Christlieb, N. 2005, ARA&A, 43, 531
Bensby, T., Feltzing, S., & Oey, M. S. 2014, A&A, 562, A71
Bertran de Lis, S., Delgado Mena, E., Adibekyan, V. Z., Santos, N. C., & Sousa, S. G. 2015, A&A, 576, A89
Delgado Mena, E., Bertrán de Lis, S., Adibekyan, V. Z., et al. 2015, A&A, 576, A69
Delgado Mena, E., Israelian, G., González Hernández, J. I., et al. 2014, A&A, 562, A92
Delgado Mena, E., Tsantaki, M., Adibekyan, V. Z., et al. 2017a, ArXiv e-prints
Delgado Mena, E., Tsantaki, M., Adibekyan, V. Z., et al. 2017b, ArXiv e-prints
Deng, L.-C., Newberg, H. J., Liu, C., et al. 2012, Research in Astronomy and Astrophysics, 12, 735
Edvardsson, B., Andersen, J., Gustafsson, B., et al. 1993, A&A, 275, 101
Fuhrmann, K. 1998, A&A, 338, 161
Fuhrmann, K. 2011, MNRAS, 414, 2893
Fuhrmann, K., Chini, R., Kaderhandt, L., & Chen, Z. 2017, MNRAS, 464, 2610
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2016, A&A, 595, A2
Gilmore, G. 2012, in Astronomical Society of the Pacific Conference Series, Vol. 458, Galactic Archaeology: Near-Field Cosmology and the Formation of the Milky Way, ed. W. Aoki, M. Ishigaki, T. Suda, T. Tsujimoto, & N. Arimoto, 147
Gratton, R., Carretta, E., Matteucci, F., & Sneden, C. 1996, in Astronomical Society of the Pacific Conference Series, Vol. 92, Formation of the Galactic Halo...Inside and Out, ed. H. L. Morrison & A. Sarajedini, 307
Hinton, G. E. & Roweis, S. T. 2003, in Advances in neural information processing systems, 857–864
Ivezić, Ž., Connolly, A., VanderPlas, J., & Gray, A. 2013, Statistics, Data Mining, and Machine Learning in Astronomy
Kos, J., Bland-Hawthorn, J., Freeman, K., et al. 2017, MNRAS, submitted, arXiv:1709.00794
Kullback, S. & Leibler, R. A. 1951, Ann. Math. Statist., 22, 79
Lindegren, L. & Feltzing, S. 2013, A&A, 553, A94
Linderman, G. C. & Steinerberger, S. 2017
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94
Martell, S. L., Sharma, S., Buder, S., et al. 2017, MNRAS, 465, 3203
Matijevič, G., Chiappini, C., Grebel, E. K., et al. 2017, A&A, 603, A19
Matijevič, G., Zwitter, T., Bienaymé, O., et al. 2012, ApJS, 200, 14
Nissen, P. E. 2015, A&A, 579, A52
Nissen, P. E. 2016, A&A, 593, A65
Pagel, B. E. J. 2009, Nucleosynthesis and Chemical Evolution of Galaxies
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2012, ArXiv e-prints
Queiroz, A. B. A., Anders, F., Santiago, B. X., et al. 2017, MNRAS, submitted, arXiv:1710.09970
Reis, I., Poznanski, D., Baron, D., Zasowski, G., & Shahaf, S. 2017, ArXiv e-prints
Santiago, B. X., Brauer, D. E., Anders, F., et al. 2016, A&A, 585, A42
Steinmetz, M., Zwitter, T., Siebert, A., et al. 2006, AJ, 132, 1645

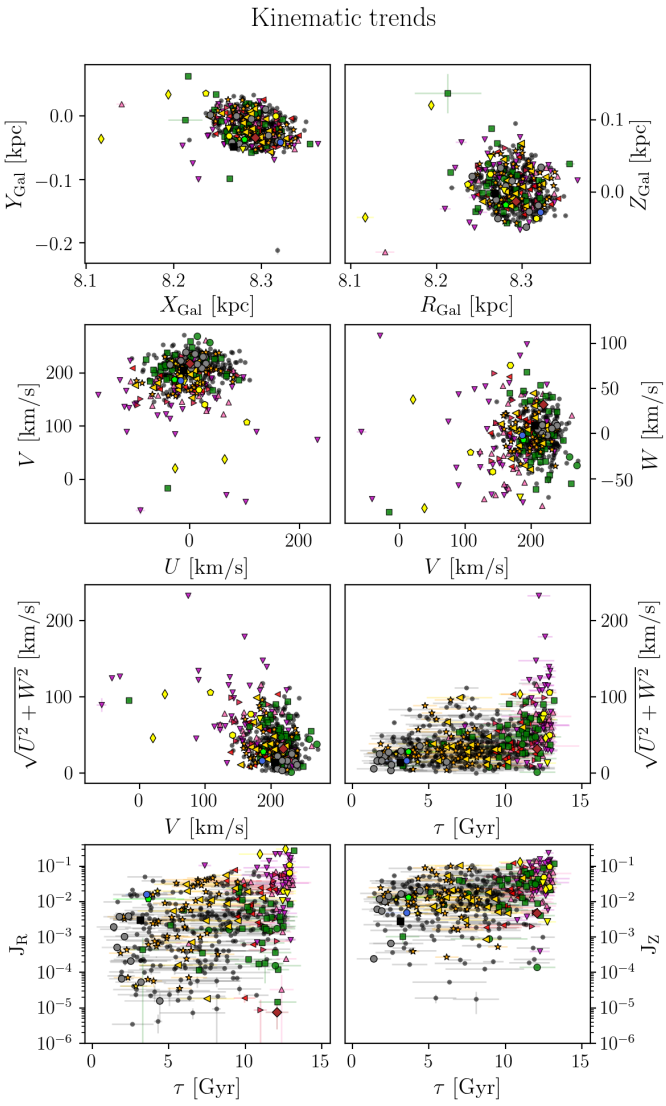

Fig. 7. Abundance trends with stellar age, measured with the `StarHorse` code (Queiroz et al. 2017).

Suárez-Andrés, L., Israelian, G., González Hernández, J. I., et al. 2017, A&A, 599, A96
Traven, G., Matijevič, G., Zwitter, T., et al. 2017, ApJS, 228, 24
Tsantaki, M., Sousa, S. G., Adibekyan, V. Z., et al. 2013, A&A, 555, A150
Valentini, M., Chiappini, C., Davies, G. R., et al. 2017, A&A, 600, A66
van der Maaten, L. & Hinton, G. 2008, The Journal of Machine Learning Research, 9, 85
van Leeuwen, F., ed. 2007, Astrophysics and Space Science Library, Vol. 350, Hipparcos, the New Reduction of the Raw Data
Wattenberg, M., Viégas, F., & Johnson, I. 2016, Distill
Yanny, B., Rockosi, C., Newberg, H. J., et al. 2009, AJ, 137, 4377

Acknowledgements. FA would like to thank Elisa Delgado-Mena for sharing the re-reduced HARPS-GTO data, and Guillaume Guiglion, Katia Cunha, Paula Jofré, Bertrand Lemasle and the other participants of the IAU symposium 334 in Potsdam, as well as David W. Hogg, for their encouragement and critical thoughts.