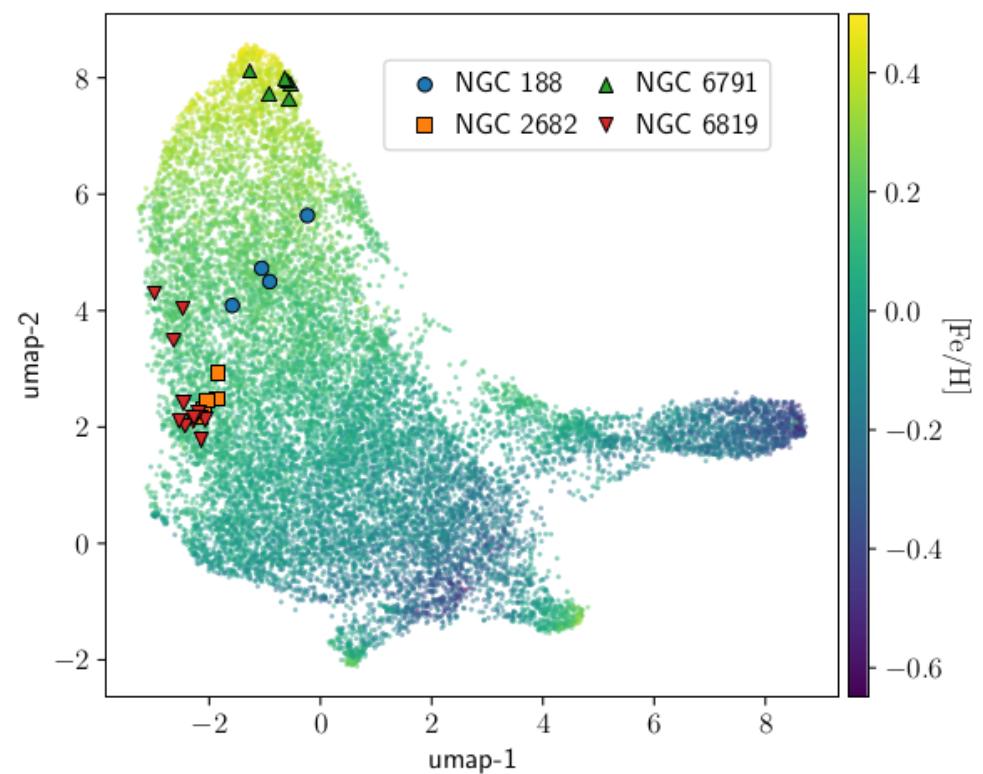


Machine learning in astro: a personal view

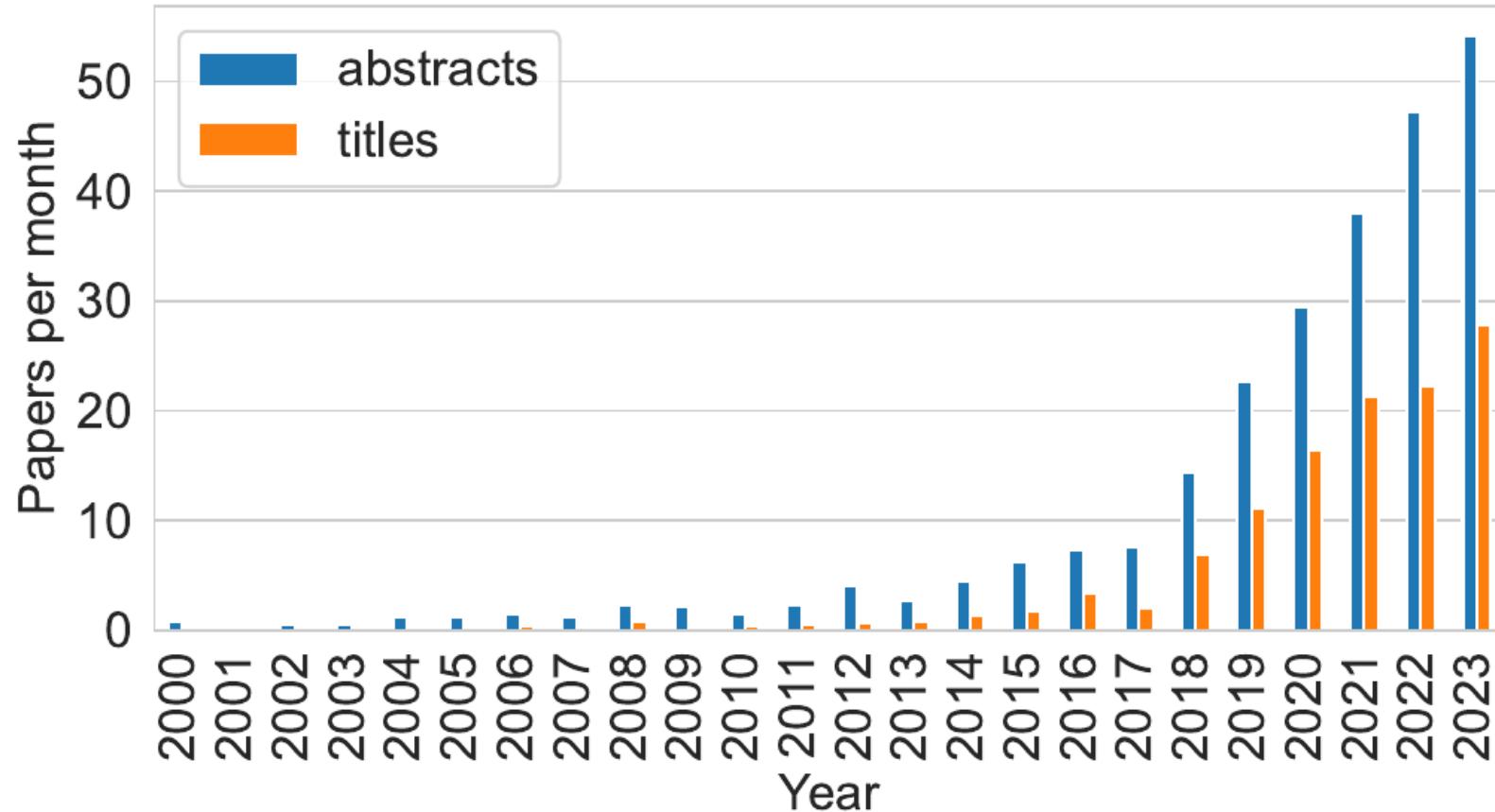
Friedrich Anders, Coimbra, September 2024



gaia

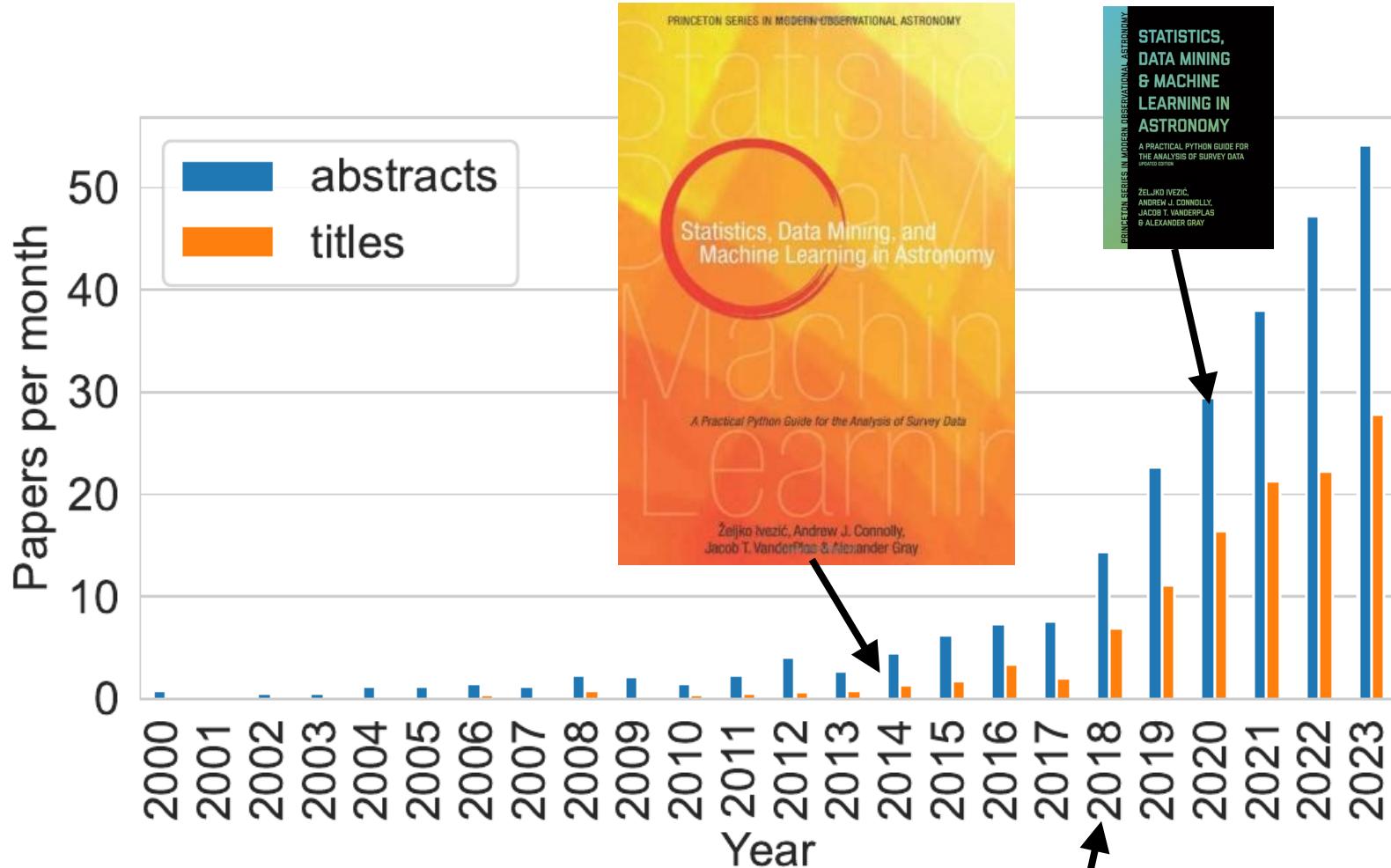


Machine learning in astronomy



Number of refereed publications per month that include the terms "machine learning" or "artificial intelligence"

Machine learning in astronomy



Number of refereed publications per month the last 23 years

Huppenkothen+2023



"intelligence"

Big Data
analysis in
astronomy

„Machine learning in astronomy“ in reality



Machine learning in astronomy: Useful references

Books:

- [Ivezic+2020](#): Statistics, Data Mining, and Machine Learning in Astronomy

Recent reviews:

- [Baron 2019](#): ML in astronomy – a practical overview
- [Fluke & Jacobs 2020](#): Surveying the reach and maturity of machine learning and artificial intelligence in astronomy
- [Reis+2021](#): Effectively using unsupervised machine learning in next generation astronomical surveys
- [Huppenkothen+2023](#): Impactful Machine Learning Research for Astronomy: Best Practices

„7 types of astronomical data“

Fluke & Jacobs 2020

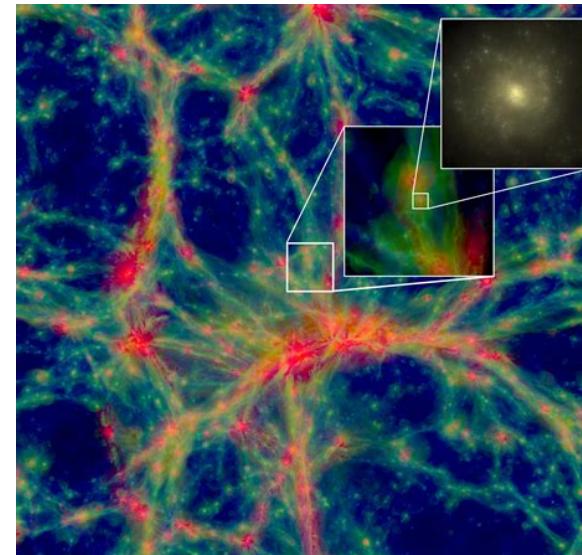
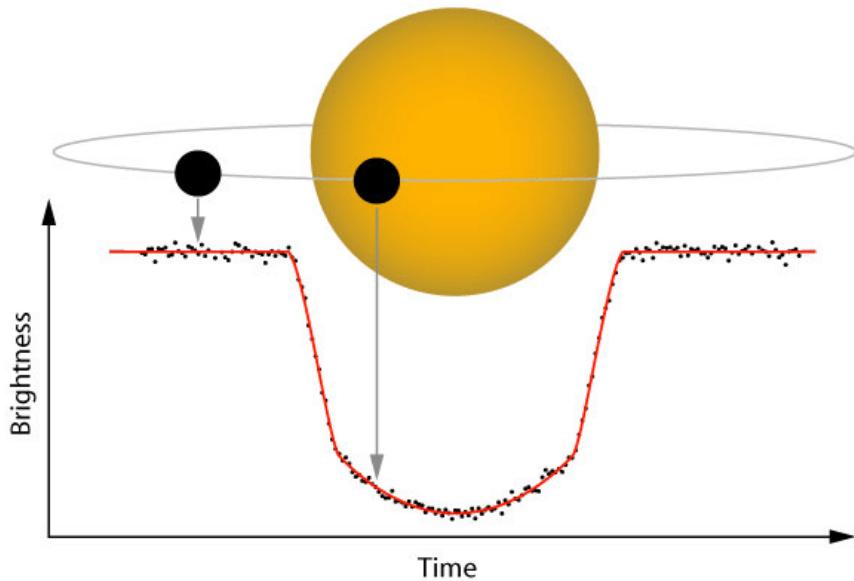
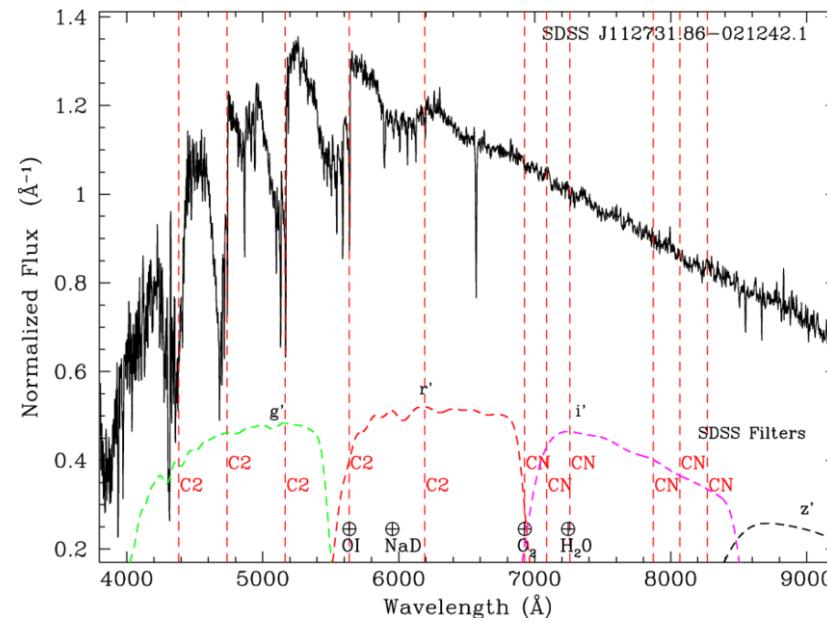


Image
Spectroscopy
Photometry
Light curve
Time Series
Catalogue
Simulation

„7 types of astronomical problems“

Fluke & Jacobs 2020

Classification

Regression

Clustering

Forecasting

Generation

Discovery

Insight

Learn from previous events, and predict or forecast that a similar event is going to occur



Missing information is created, expected to be consistent with the underlying truth



New celestial objects, features, or relationships are identified

Insight is gained into the suitability of applying machine learning, choice of data set, hyperparameters, etc

„7 types of astronomical problems“

Fluke & Jacobs 2020

Nature/Type	Classification	Regression	Clustering	Forecasting	Generation	Discovery	Insight
Image	•			•	•		•
Spectroscopy	•	•	•	•		•	•
Photometry	•	•	•	•		•	•
Light curve	•	•	•	•		•	•
Time Series	•	•	•	•	•	•	•
Catalogue	•	•	•	•	•	•	•
Simulation	•	•		•	•	•	•

„7 types of computational problems“

Ivezic+2020

- 1) **Basic problems:** simple statistics: $O(N)$ or $O(N \log N)$ at worst
- 2) **Generalized N-body problems:** any problem involving distances between tuples of points (nearest-neighbor searches, KDE): typically $O(N^2)$ or $O(N^3)$
- 3) **Linear algebraic problems:** linear systems, eigenvalue problems, and inverses: can be $O(N)$ but in some cases the matrix of interest is $N \times N$, making the computation $O(N^3)$
- 4) **Optimization problems:** from unconstrained ($O(N)$) to constrained (e.g. nonlinear support vector machines: $O(N^3)$ convex and nonconvex).
- 5) **Integration problems:** e.g. estimation of Bayesian models: integration with high accuracy via quadrature has a computational complexity which is exponential in the dimensionality $D \rightarrow$ MCMC
- 6) **Graph-theoretic problems:** probabilistic graphical models or nearest-neighbor graphs for manifold learning
- 7) **Alignment problems:** “cross-matching” in astronomy: The worst-case cost is exponential in $N...$

3 astrophysics science cases

Nature/Type	Classification	Regression	Clustering	Forecasting	Generation	Discovery	Insight
Image	•			•	•		•
Spectroscopy	•	●	•	•		●	●
Photometry	•	●	•	•		●	●
Light curve	•	●	●	•		●	●
Time Series	•	●	●	•		●	●
Catalogue	●	●	●	●		●	●
Simulation	●	●		●	●	●	

ML Methods used in our group: (XD)GMM, kNN, ANN (MLP), CNN, umap, t-SNE, XGBoost

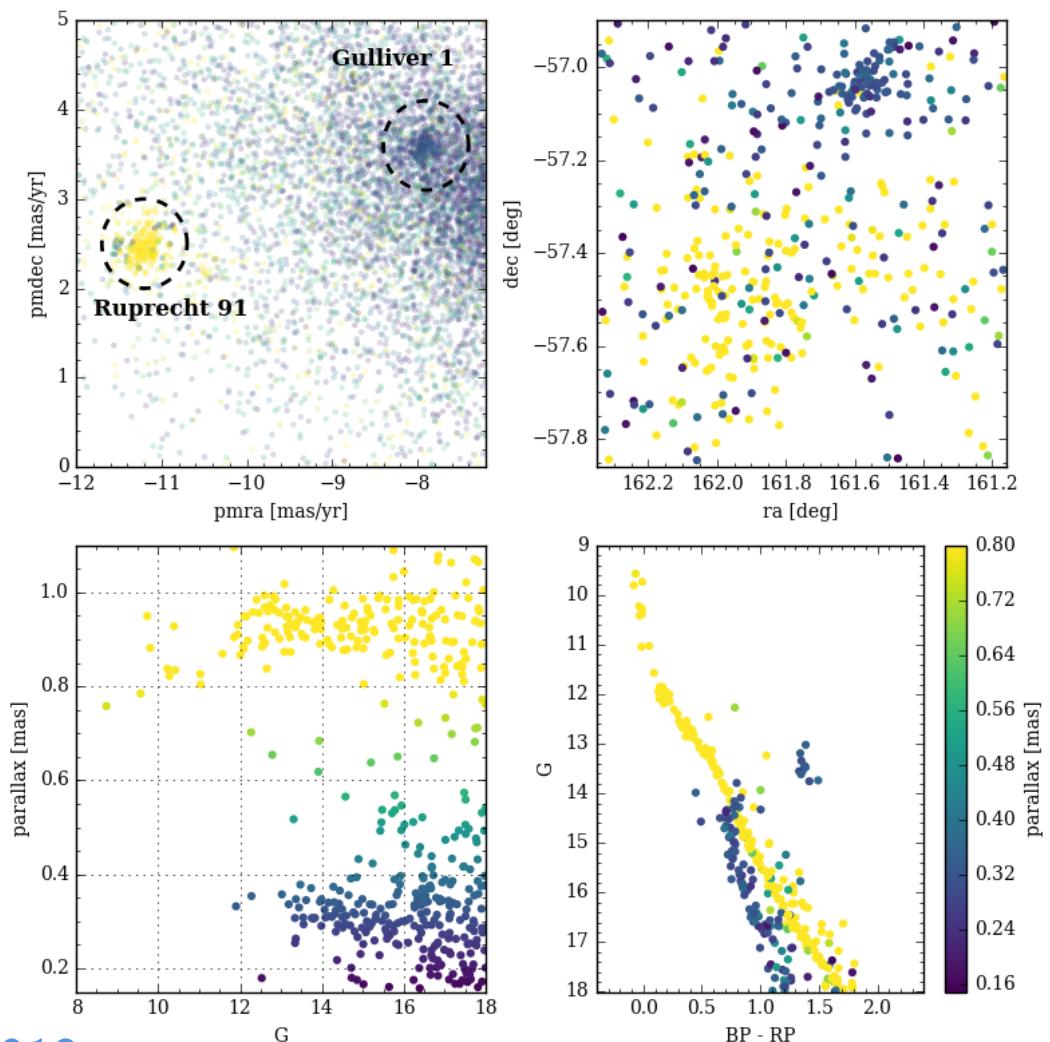
Science case I: The Gaia open cluster census

Nature/Type	Classification	Regression	Clustering	Forecasting	Generation	Discovery	Insight
Image	•			•	•		
Spectroscopy	•	•	•	•	•	•	•
Photometry	•	•	•	•	•	•	•
Light curve	•	•	•	•	•	•	•
Time Series	•	•	•	•	•	•	•
Catalogue	●	●	●	●	●	●	●
Simulation	●	●		●	●	●	●

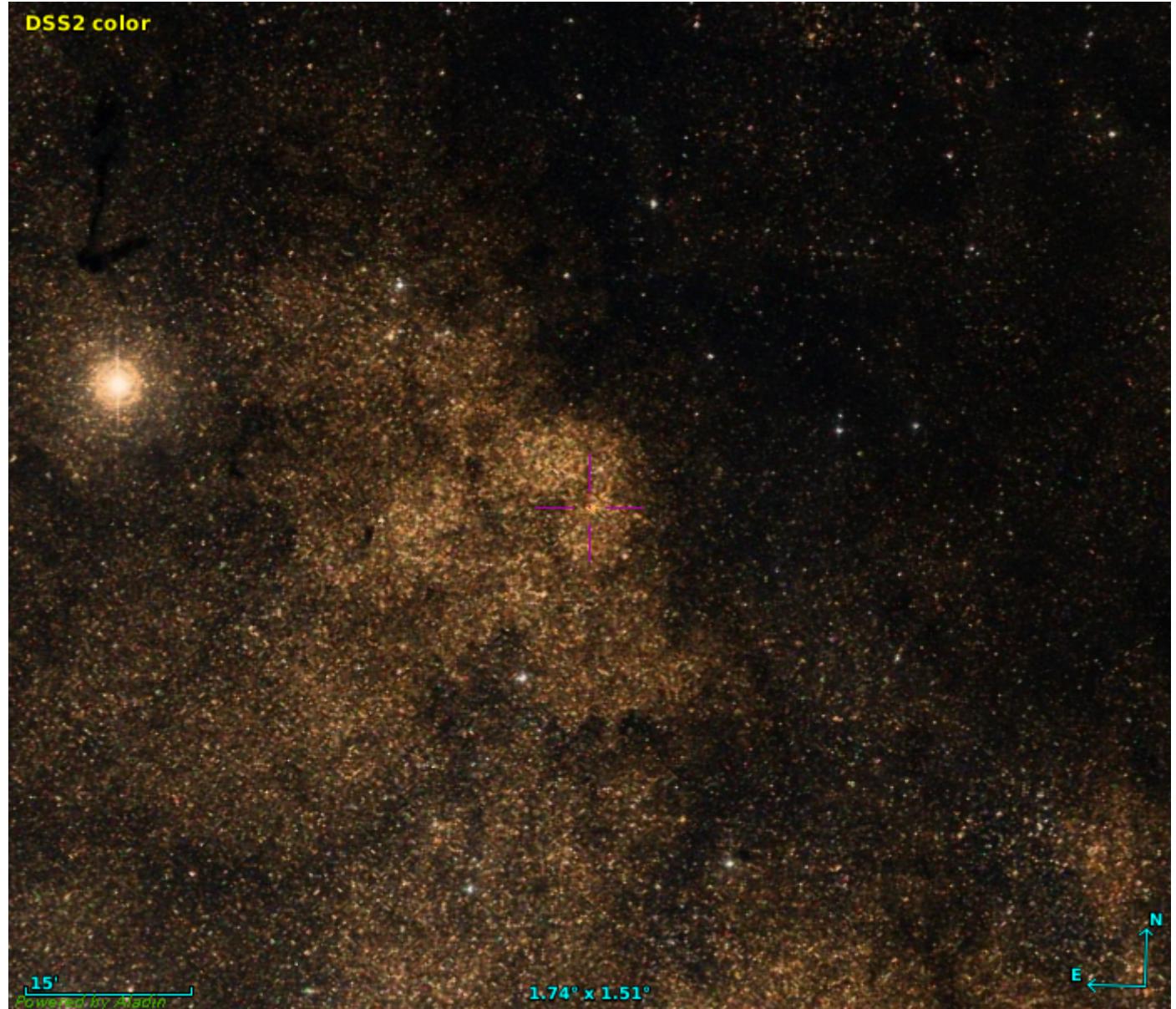
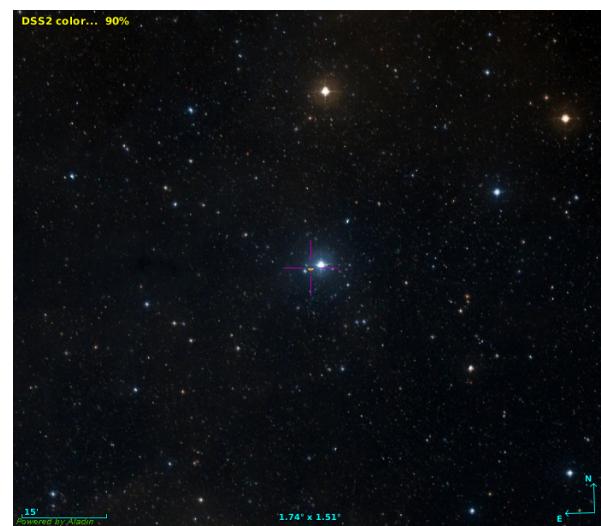
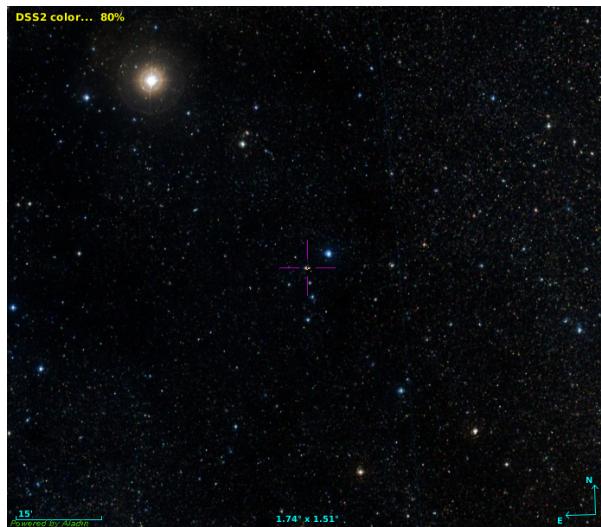
Science case I: Finding star clusters in Gaia data



gaia



What is an open star cluster?



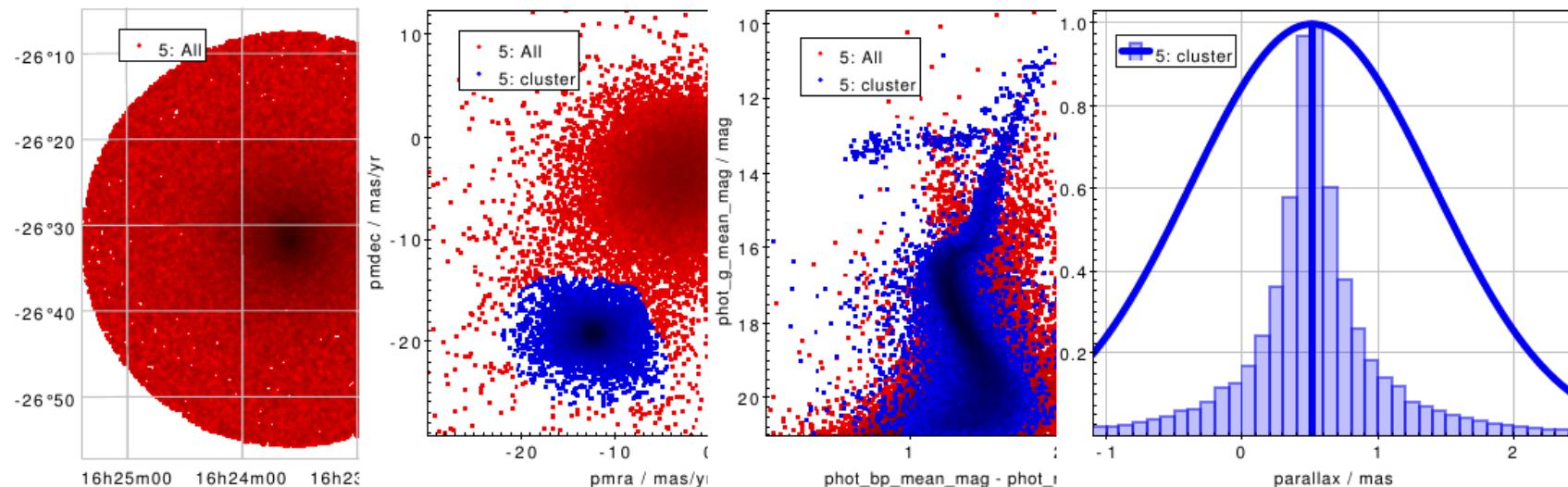
Gaia: cluster finding in phase space

Example: Cluster Identification in 2d

Identify M4 members by proper motion clustering

- Cone Search M4, $r=0.3^\circ$; returns $\sim 62,000$ rows
- Sky Plot — cluster visible
- Proper Motion space plot `pmra` vs. `pmdec`
- Make graphical selection of comoving objects
- Colour-Magnitude diagram `bp_rp` vs. `phot_g_mean_mag`, view selection (**Subsets** tab)
- Plot cluster parallax histogram & Gaussian fit

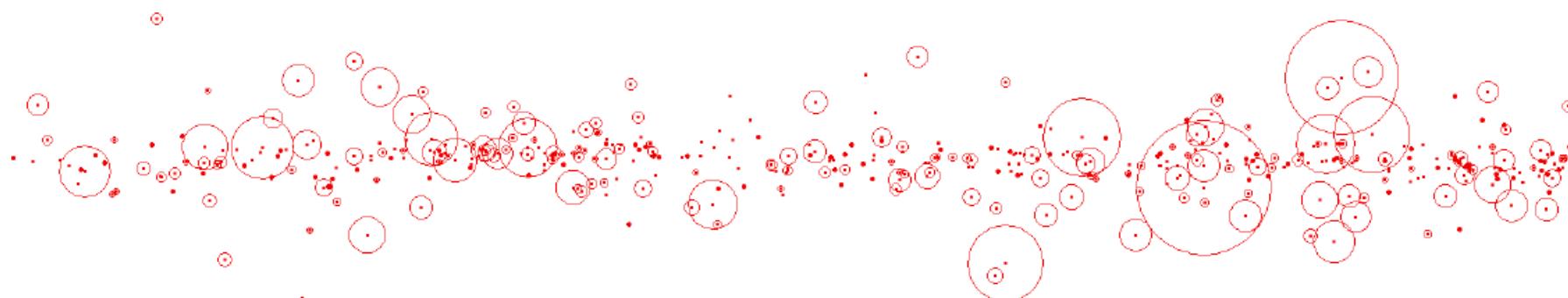
Before DR2, this only worked for the Pleiades!



Gaia DR2: cluster finding in phase space

April 25th 2018, 12:00:01

Started with an automatic query of the Gaia DR2 archive for every single open cluster listed in the catalogues of [Dias et al. \(2002\)](#) and [Kharchenko et al. \(2013\)](#) a total of **3383 cone searches**.



Only used Gaia DR2 sources brighter than G=18.

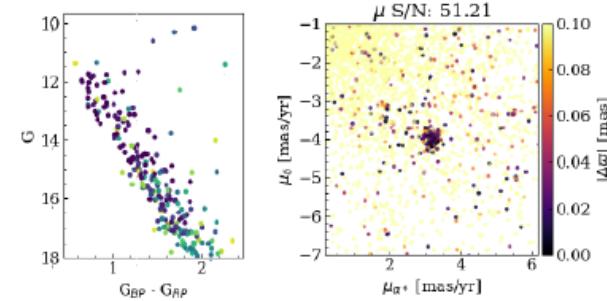
Analysed each investigated field with UPMASK ([Krone-Martins & Moitinho 2014](#)). The code makes groups in astrometric space (**pmra,pmdec,parallax**) and checks if they are more concentrated on the sky than a random distribution. It also takes into account the uncertainties on the astrometric parameters.

Clearly identified 1169 known and 60 previously unknown clusters!

A lot are missing because they are too distant or reddened to be identified on Gaia DR2 alone.

Some clusters recently discarded as “non-existent” are real!

e.g. NGC 886, sparse, high extinction,
but very compact in astrometric space



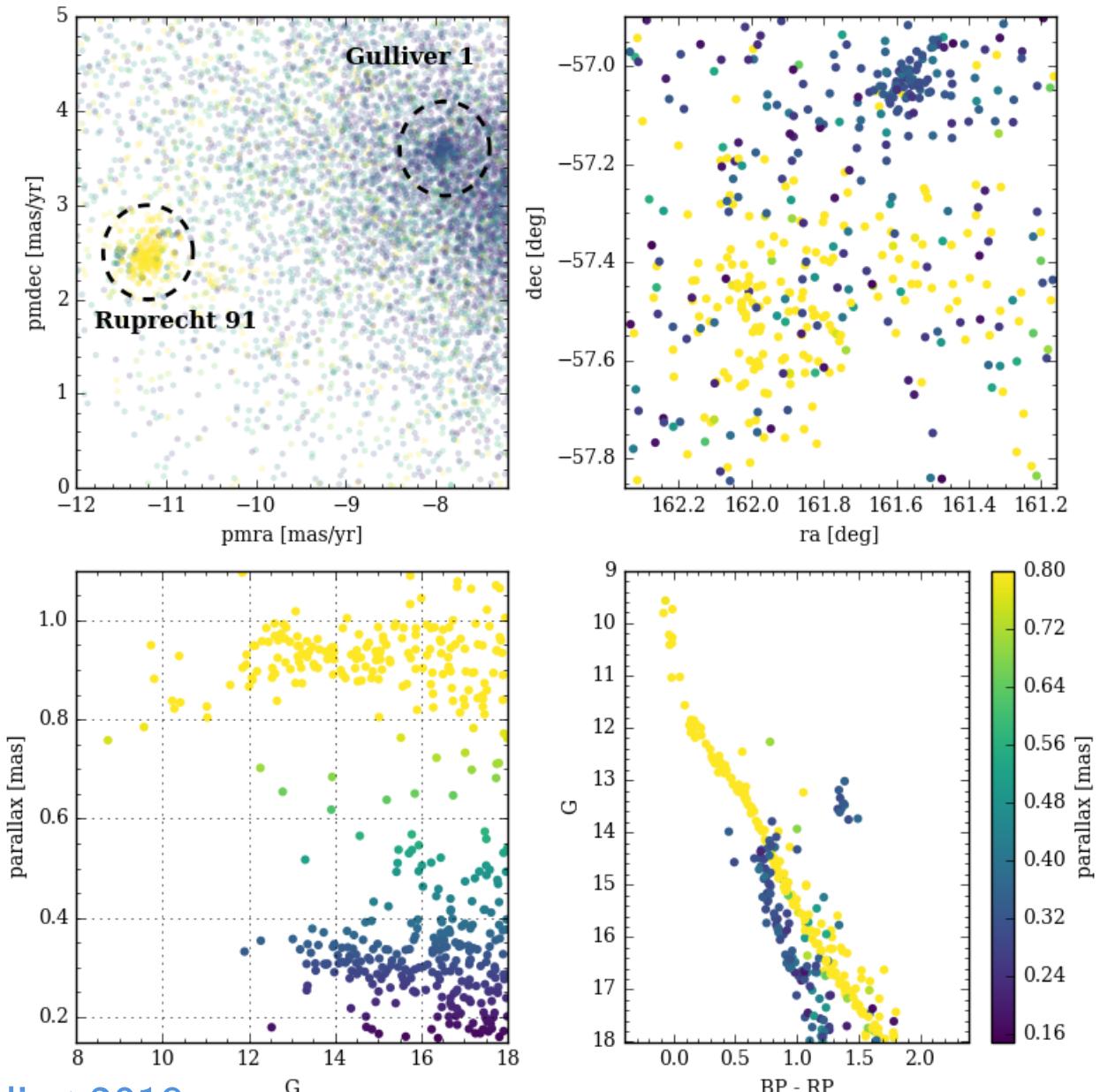
Gaia DR2: cluster finding in phase space Serendipitous discoveries

Basically:

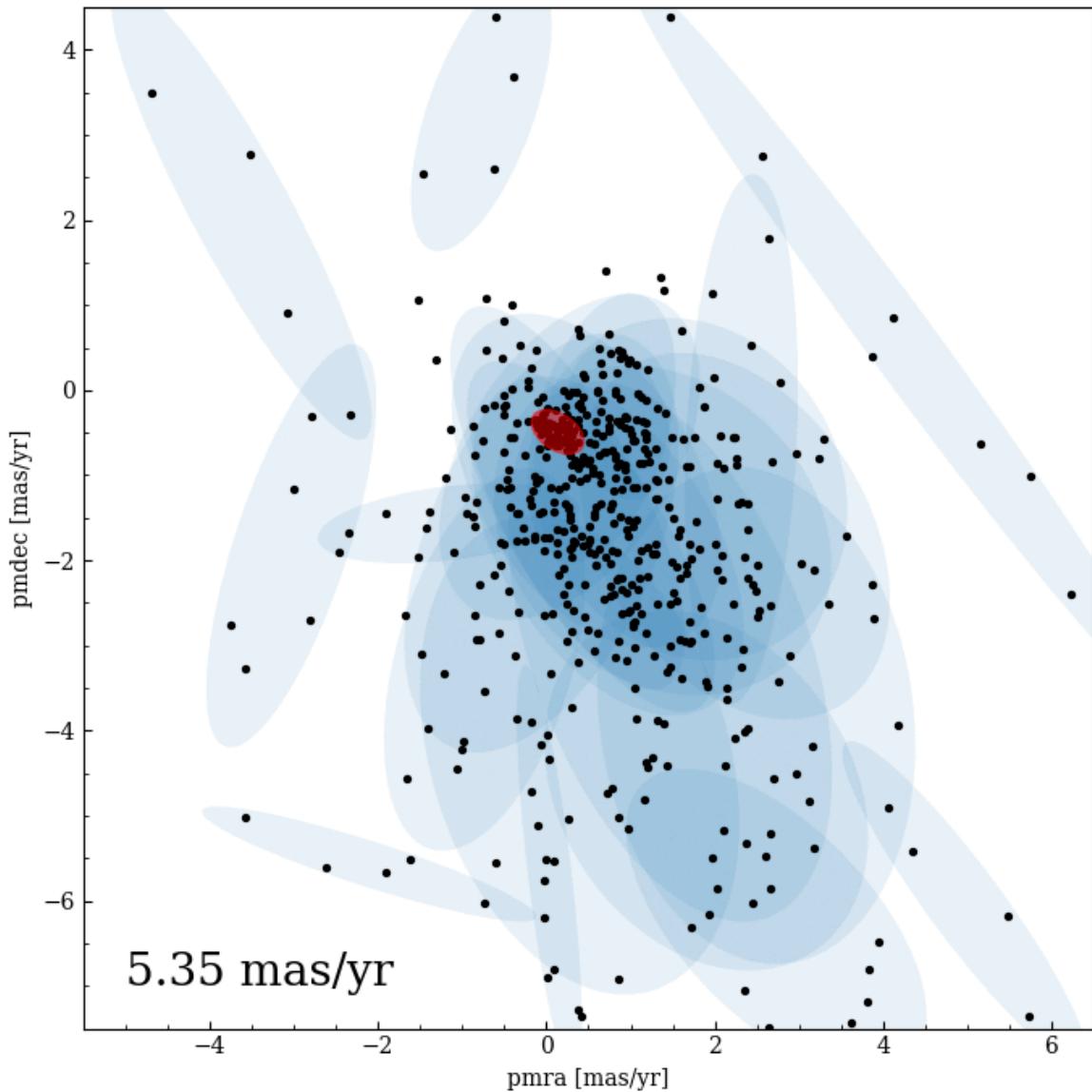
New discoveries could be made **by eye**.

The Gaia DR2 data are often so good that they look like textbook examples for clustering algorithms.

It was clear that much more could be found...

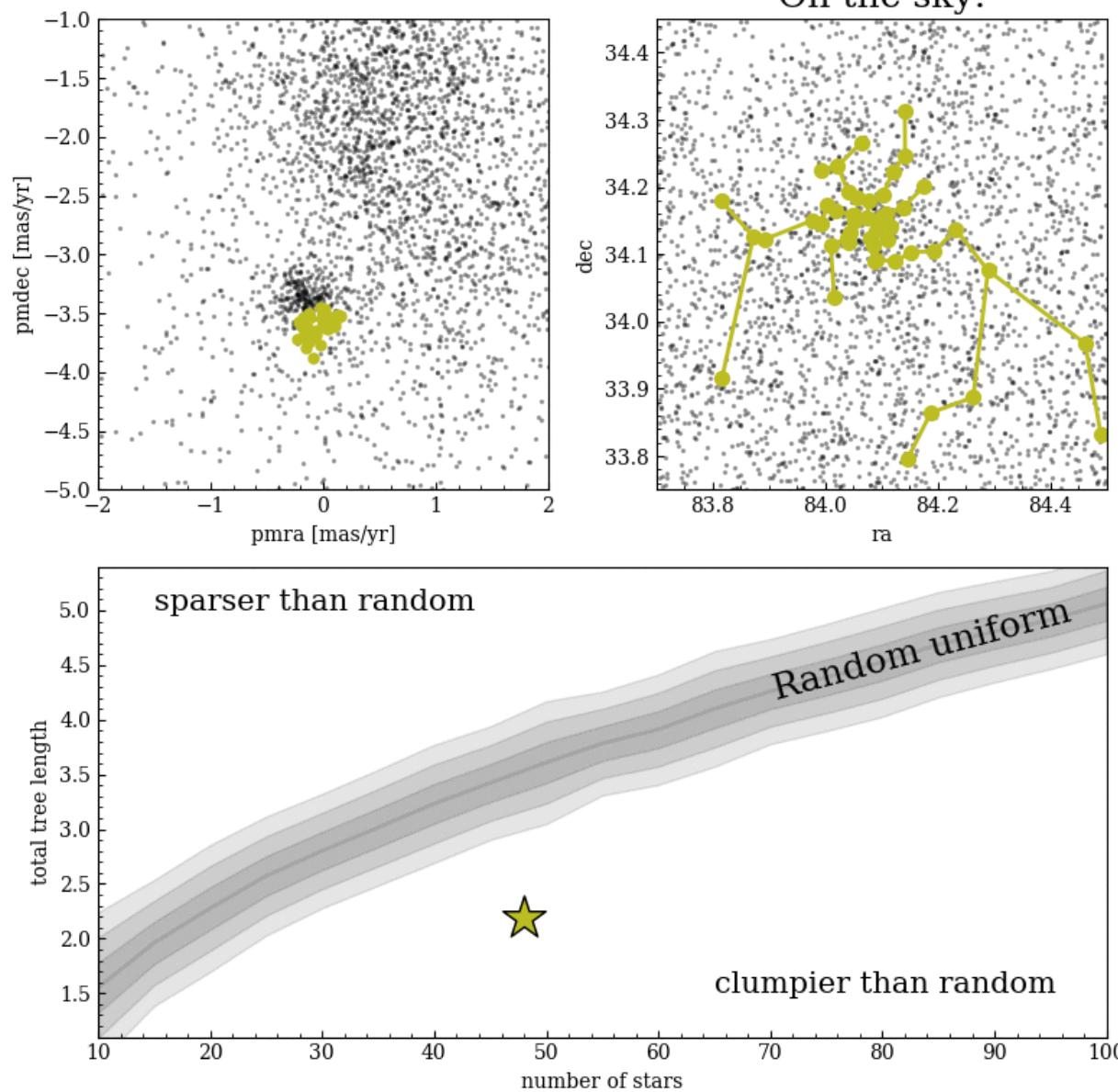


Cantat-Gaudin+2019: GMM + K-Means



Step 1:
Use Gaussian mixture model to look
for sharp peaks in the velocity
distribution
→ cluster candidates

Cantat-Gaudin+2019: GMM + K-Means



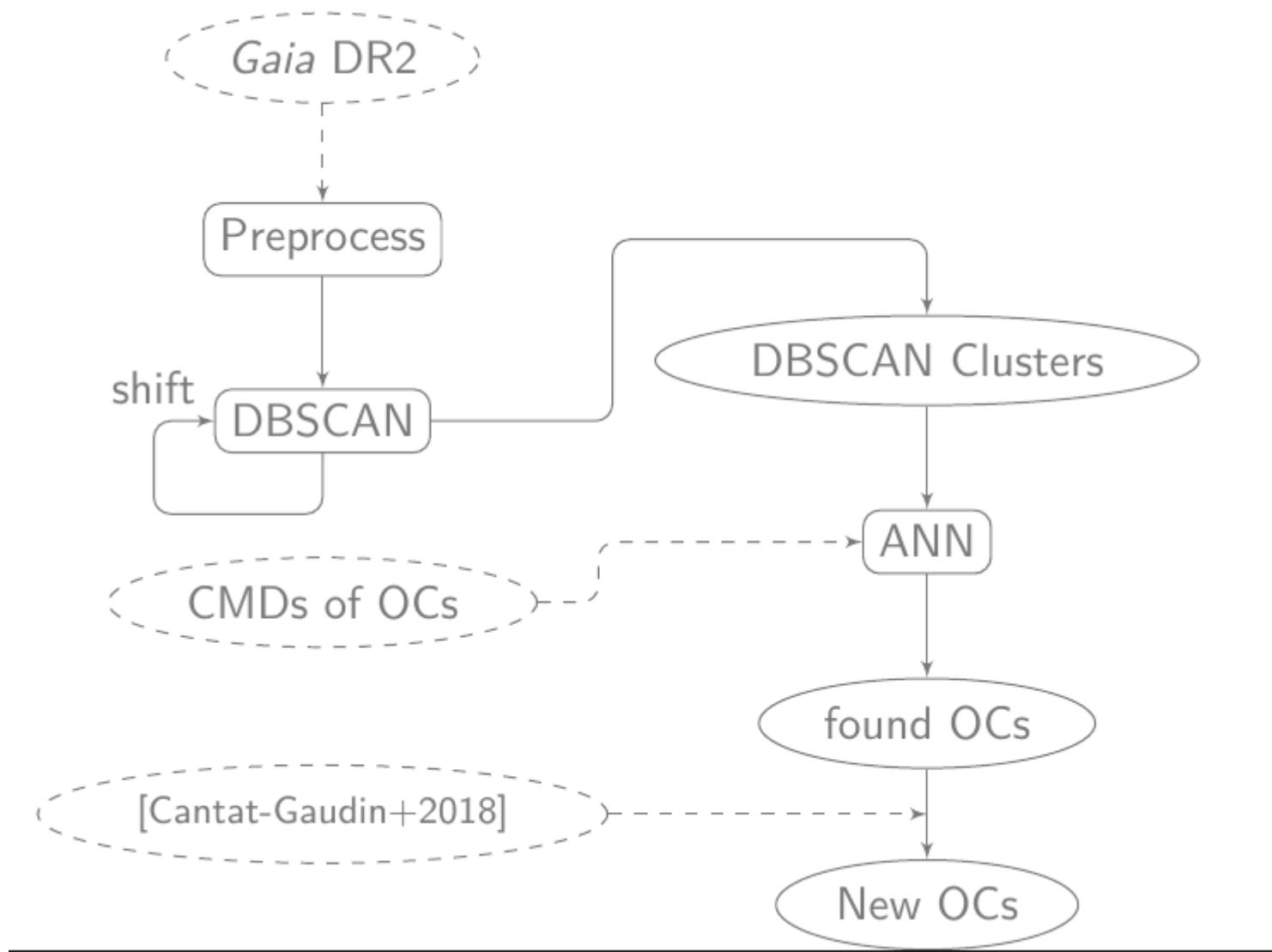
Step 2 (UPMASK):
K-means clustering in 5D astrometric space

Step 3 (UPMASK):
Compare minimum spanning tree distance with random distribution

→ confirm cluster candidates

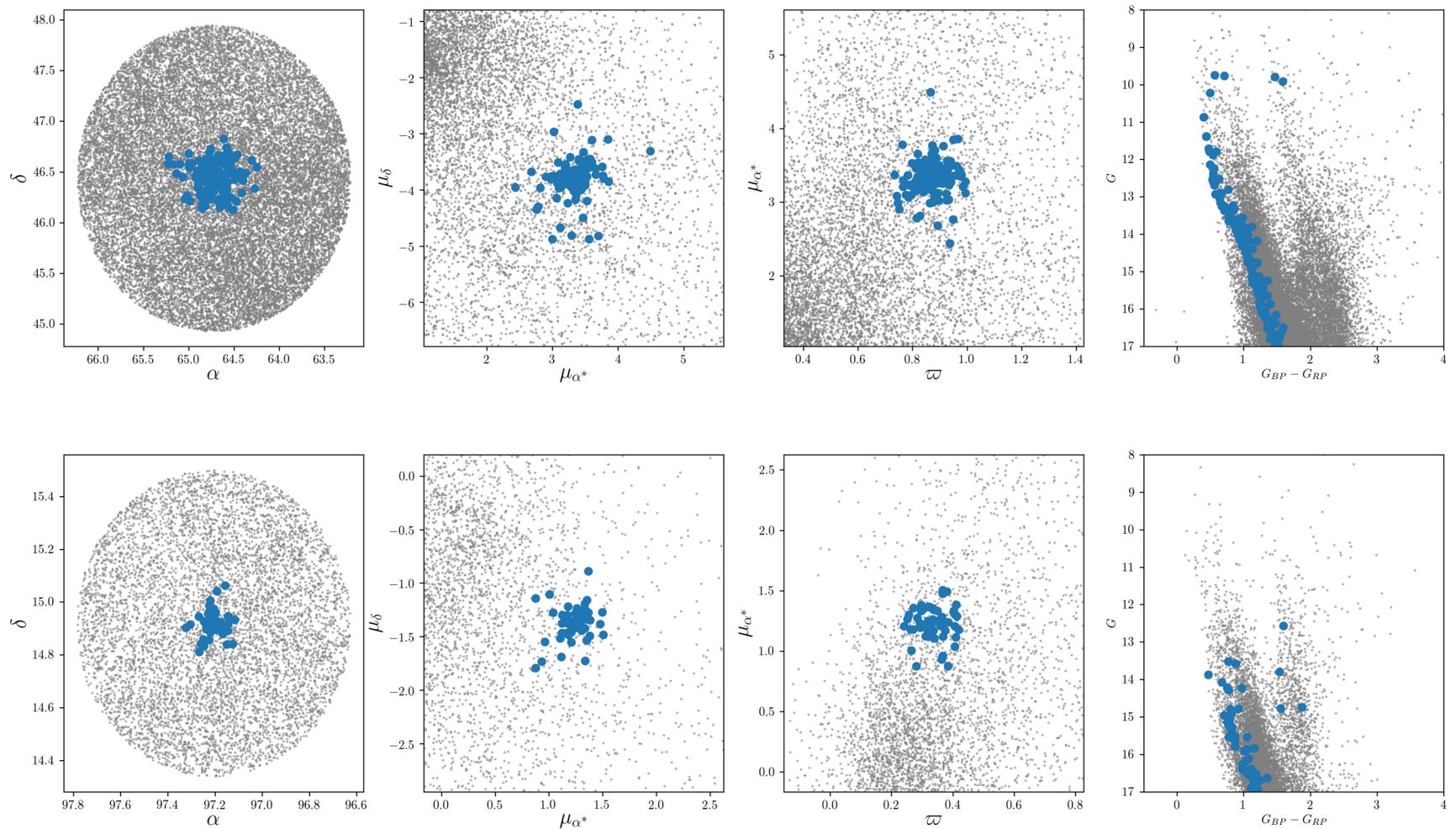
UPMASK:
Krone-Martins & Mointinho 2014

Castro-Ginard+2019, 2020, 2022: Applying DBSCAN to Gaia data



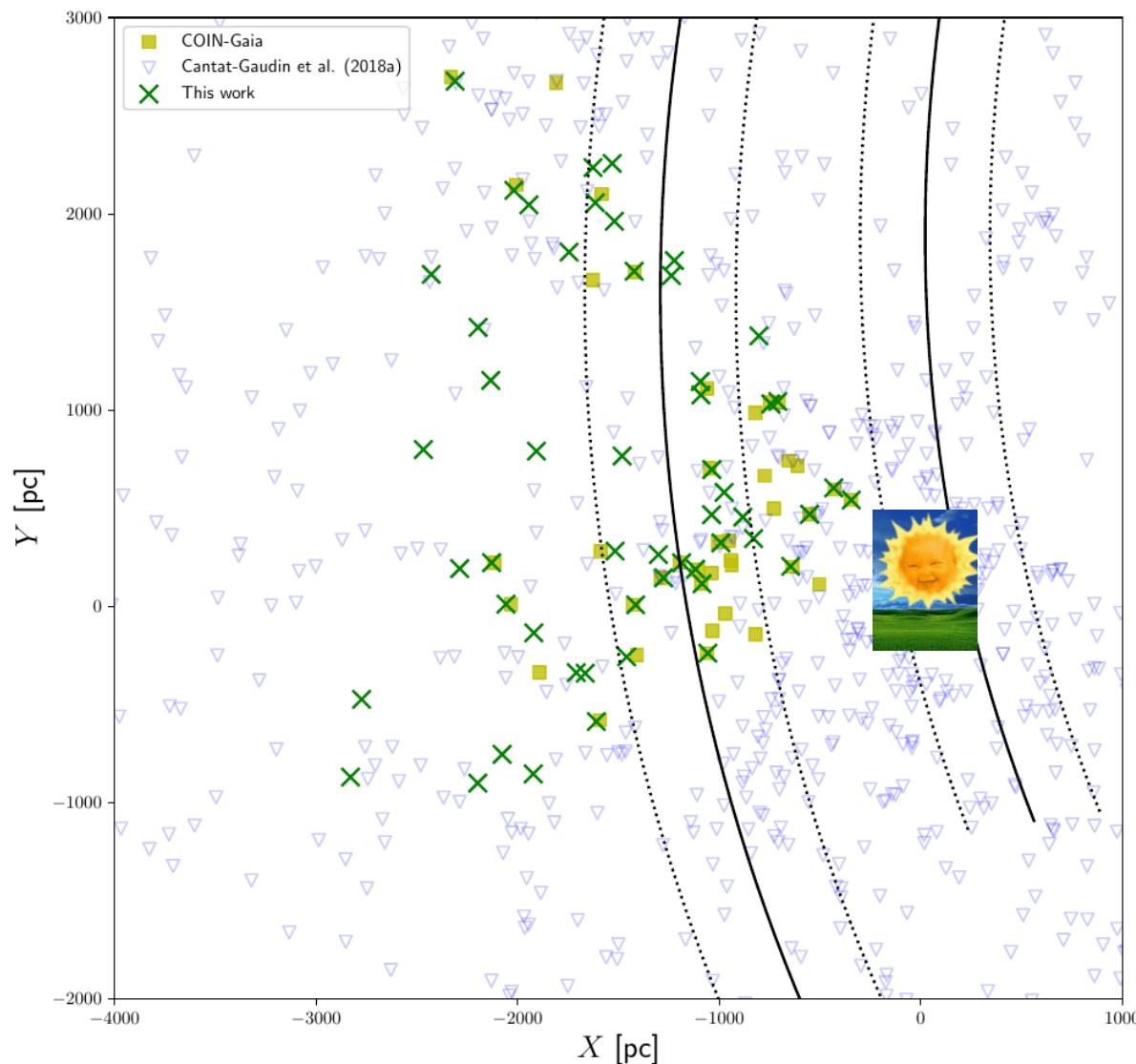
Castro-Ginard+2019, 2020, 2022: Applying DBSCAN to Gaia data

5D clustering + simple ANN classifier (trained with expert humans)
to distinguish cluster and asterism colour-magnitude diagrams (90% accuracy)
(see also [Hunt & Reffert 2023](#))

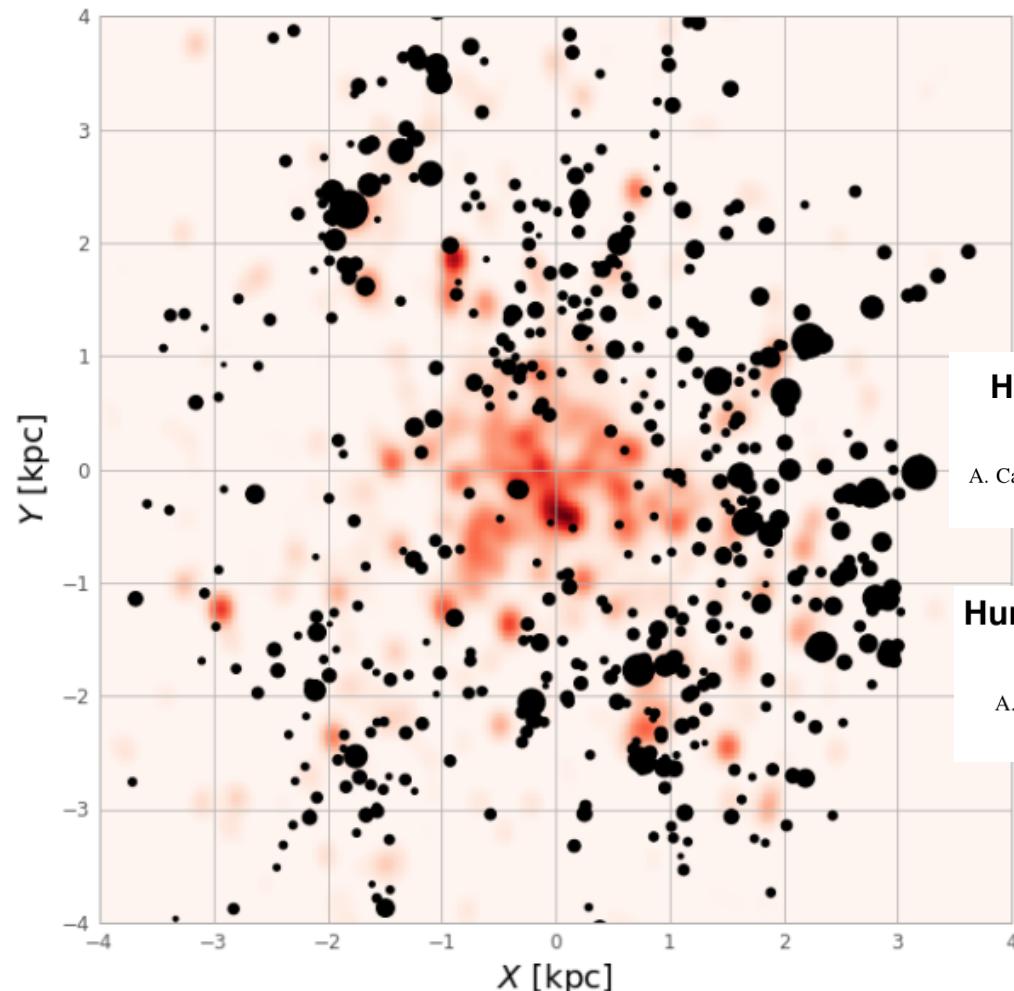


Castro-Ginard+2019: Applying DBSCAN in the Perseus arm

Some in common with [Cantat-Gaudin+2019](#), but also ~40 new ones



Castro-Ginard+2020: Applying DBSCAN to the full Galactic disc



Hunting for open clusters in *Gaia* DR2: 582 new open clusters in the Galactic disc^{*}

A. Castro-Ginard¹, C. Jordi¹, X. Luri¹, J. Álvarez Cid-Fuentes², L. Casamiquela³, F. Anders¹, T. Cantat-Gaudin¹, M. Monguió¹, L. Balaguer-Núñez¹, S. Solà², and R. M. Badia²

Hunting for open clusters in *Gaia* EDR3: 628 new open clusters found with OCfinder^{*}

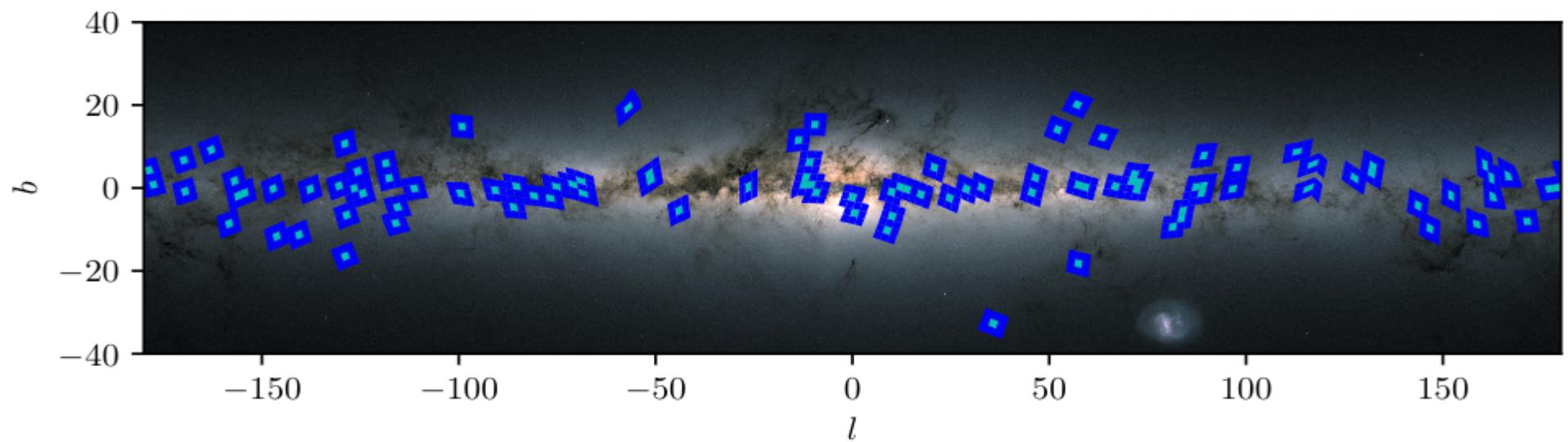
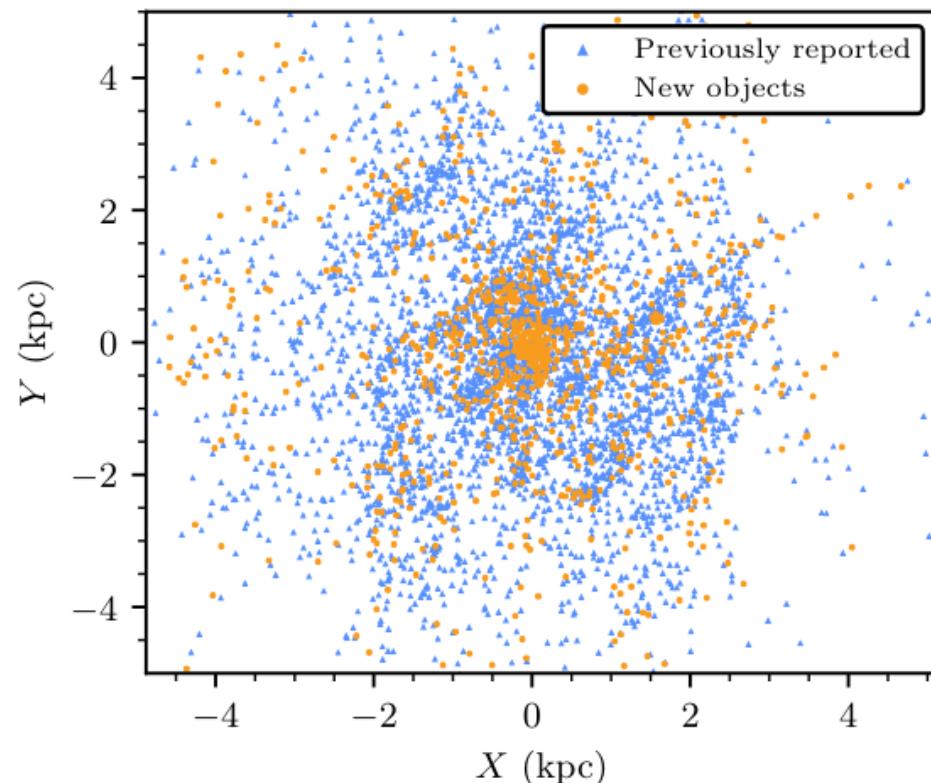
A. Castro-Ginard^{1,2}, C. Jordi², X. Luri², T. Cantat-Gaudin^{2,3}, J. M. Carrasco², L. Casamiquela⁴, F. Anders², L. Balaguer-Núñez², and R. M. Badia⁵

Fig. 5. Distribution of the OCs projected in the $X - Y$ plane. Previously known OCs (CG18, CG19, Cantat-Gaudin et al. 2018, 2019a) are shown as a density map in red. Newly found OCs reported here are shown as black dots, where the size is proportional to the number of members of each cluster.

Hunt & Reffert 2021, 2023: Applying HDBSCAN to the full Galactic disc

Table 1. Algorithms considered for inclusion by this study.

Algorithm	Runtime scaling ^a	Deals with noise	Open-source
KMeans	n	No	sklearn ^b
Affinity propagation	n^2	No	sklearn ^b
Mean-shift	n^2	No	sklearn ^b
Spectral	n^3	No	sklearn ^b
Ward	n^3	No	sklearn ^b
Agglomerative	n^3	No	sklearn ^b
DBSCAN	$n \log n$	Yes	sklearn ^b
OPTICS	n^2	Yes	sklearn ^b
Gaussian mixtures	n	No	sklearn ^b
Birch	n	No	sklearn ^b
Friend of Friends	$n \log n$	No	pyfof ^c
HDBSCAN	$n \log n$	Yes	HDBSCAN ^d



Cantat-Gaudin+2020: Open cluster parameters with an ANN

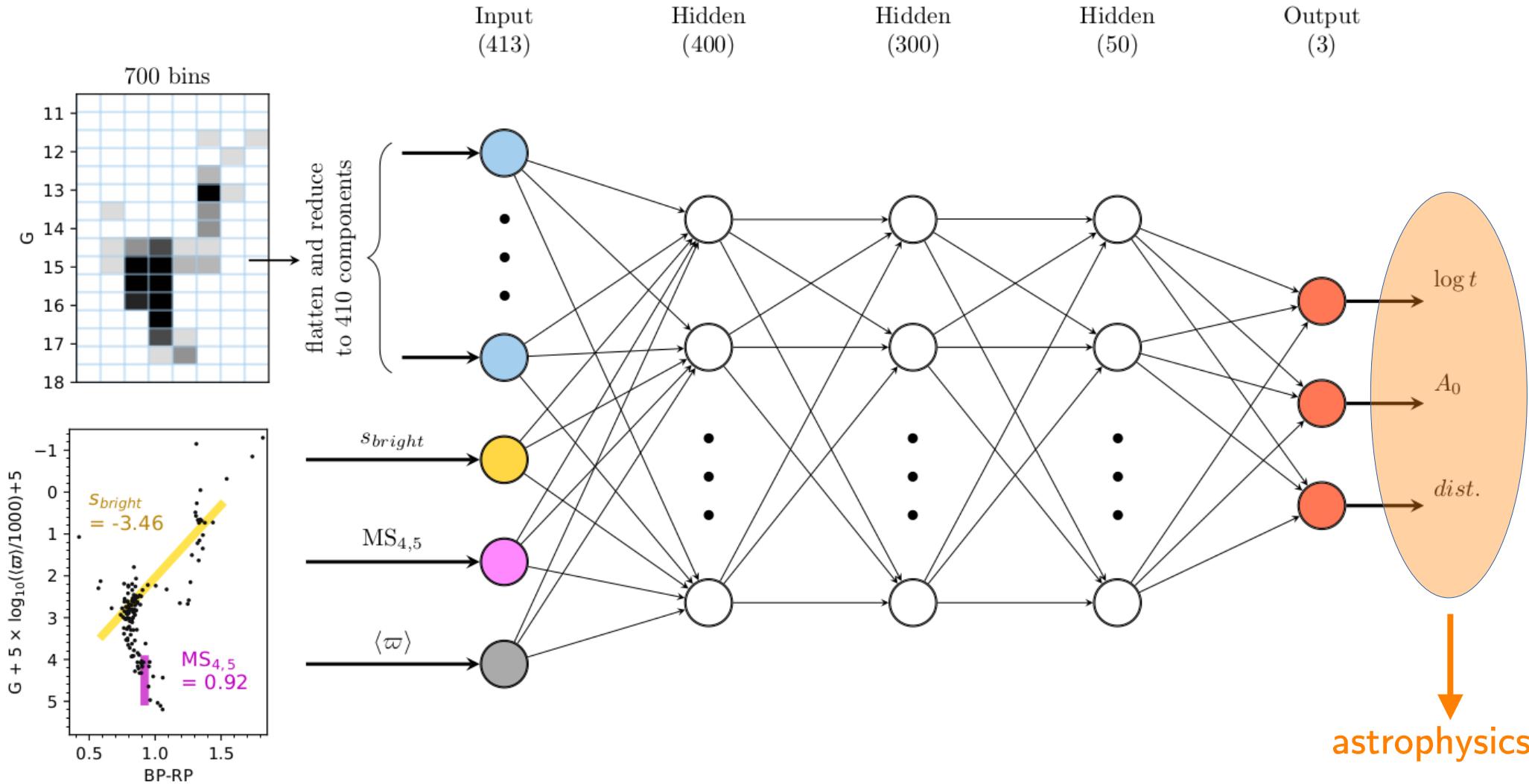
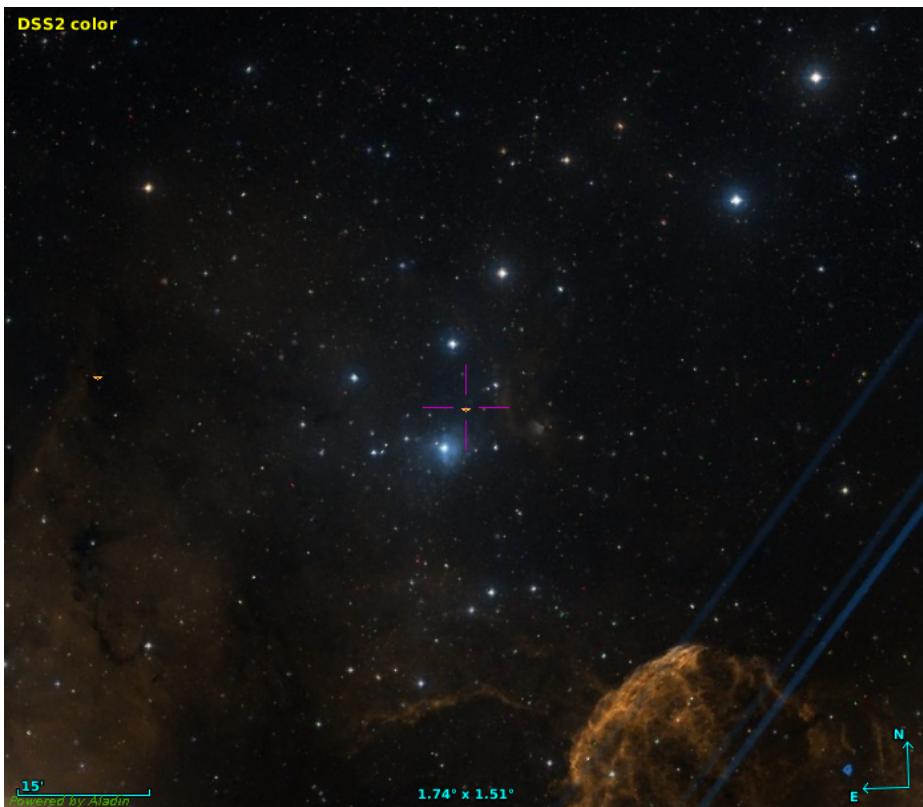


Fig. 2. Architecture of our artificial neural network, indicating the width (number of nodes) of each layer. The example cluster is Haffner 22. The input quantities are described in Sect. 3.1.

Conclusions after Gaia DR2&DR3

- 1) Gaia is an optimal playground for clustering! We are not even doing the most fancy things. Almost every time we try a new algorithm, we find something new.
- 2) **The hard work is not in *finding* candidates - but in vetting and interpreting them.** What is a good metric in 5D astrometric space?
- 3) Don't use published catalogues blindly.



Tristan Cantat-Gaudin @CantatGaudin · 25. Juni

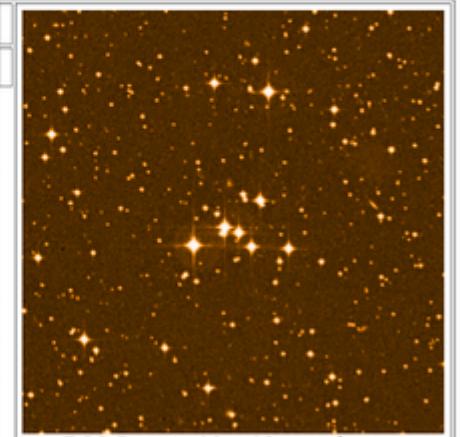
Antwort an @frediferente

My favourite asterism at the moment is Ruprecht 3. So deceitfull!

[Tweet übersetzen](#)

[WEBDA page for open cluster Ruprecht 3](#)

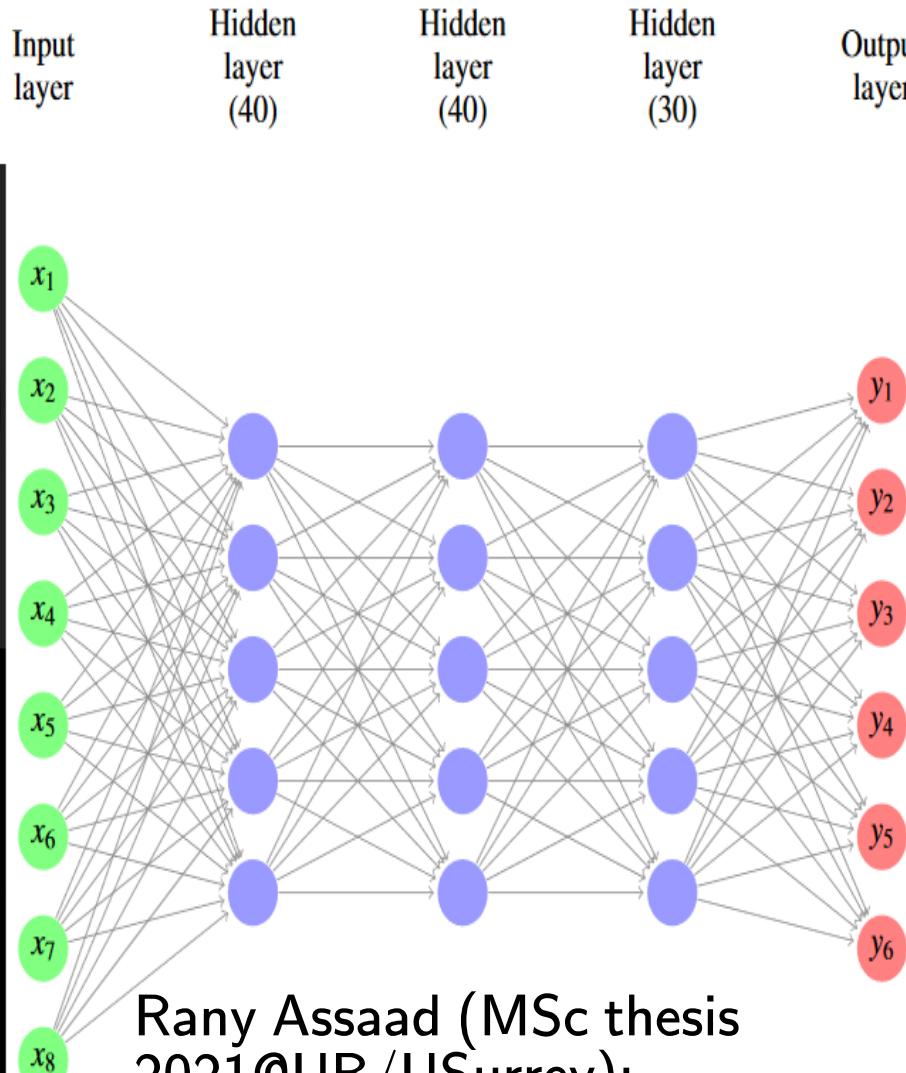
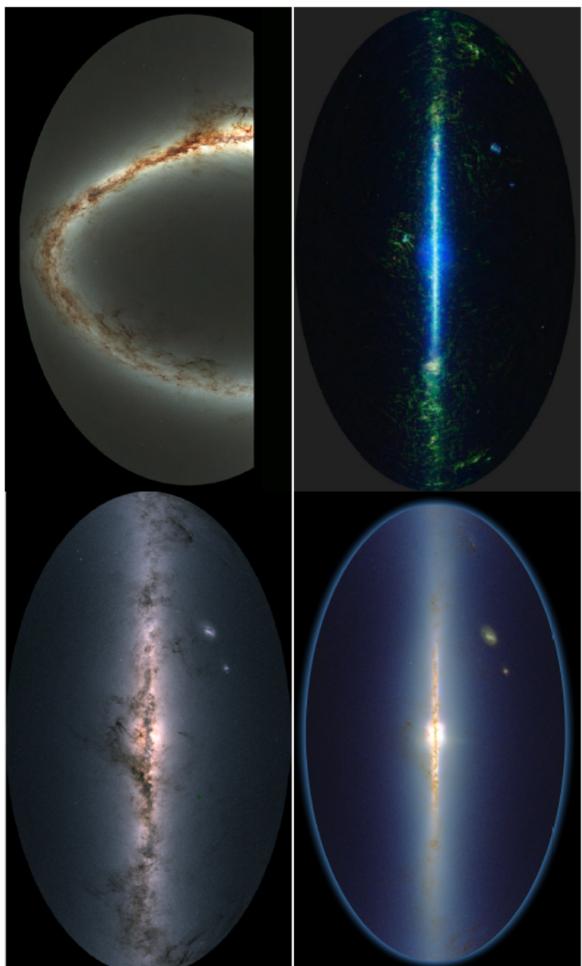
Basic Parameters	
Right Ascension (2000)	06 42 07
Declination (2000)	-29 27 00
Galactic longitude	238.776
Galactic latitude	-14.815
Distance [pc]	
Reddening [mag]	
Distance modulus [mag]	
Log Age	
Metallicity	
Notes	



Science case II: Stellar parameter estimation for Gaia

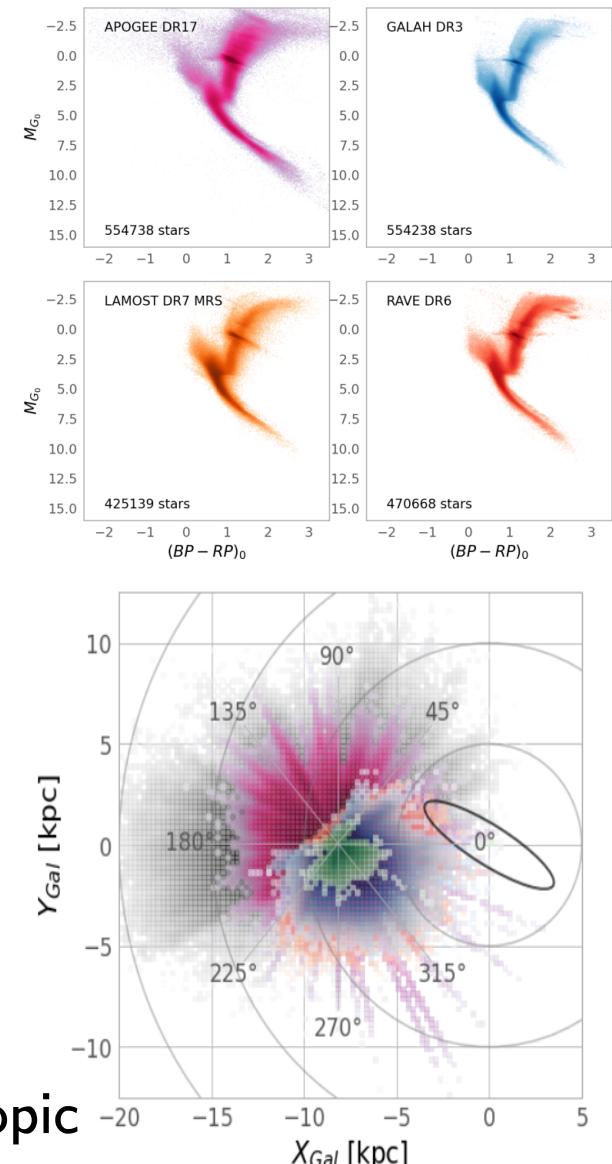
Nature/Type	Classification	Regression	Clustering	Forecasting	Generation	Discovery	Insight
Image	•			•	•		•
Spectroscopy	•	●	•	•		●	●
Photometry	•	●	•	•		•	•
Light curve	•	●	•	•		•	•
Time Series	•	●	•	•	•	•	•
Catalogue	•	●	•	•		●	●
Simulation	•	●		•	•	•	•

„Label transfer“: Spectroscopy (<10M stars) → Gaia + photometry (>300M stars)



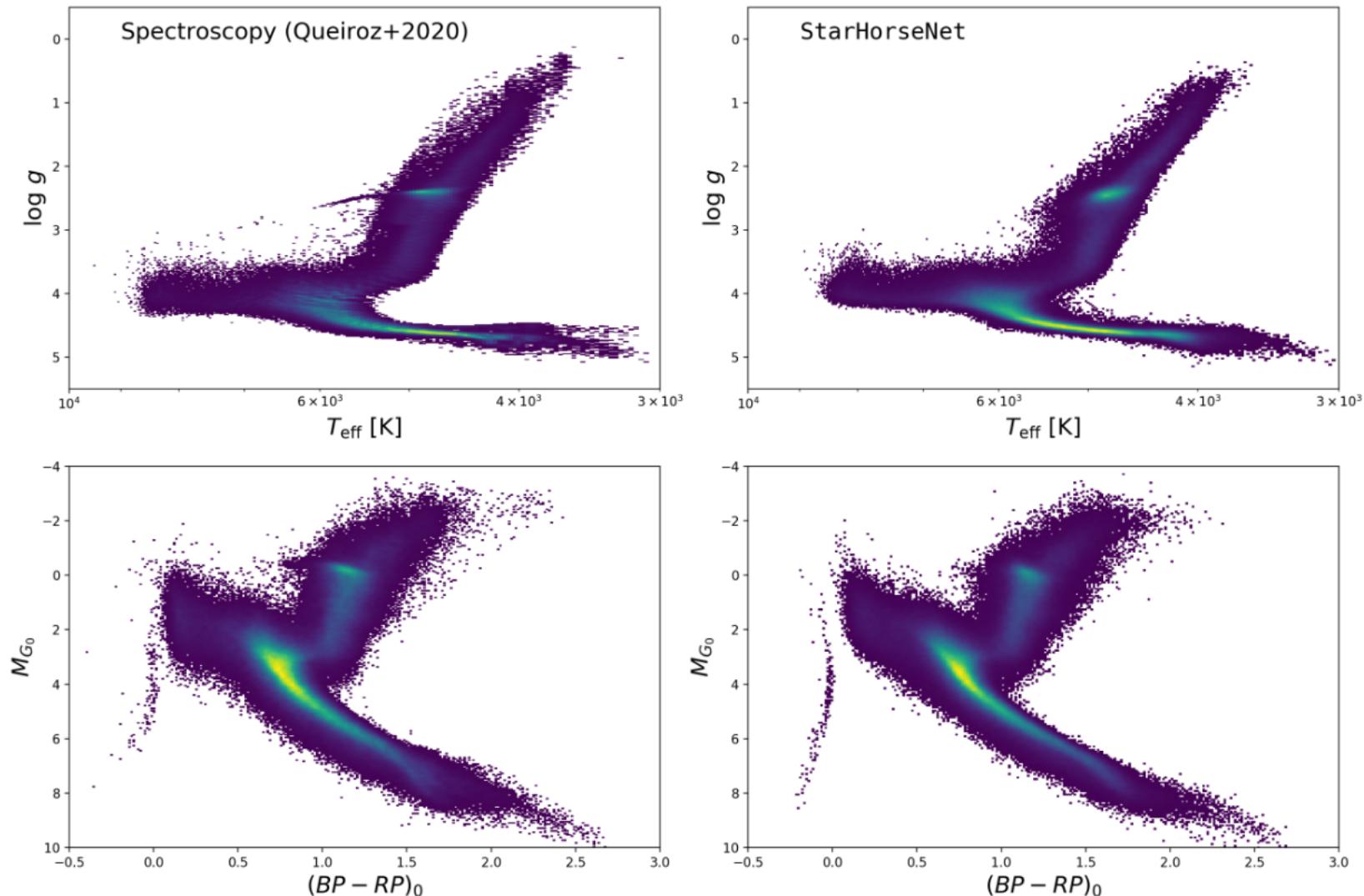
Rany Assaad (MSc thesis
2021@UB/USurrey):

- Tested **ANN regression** of Gaia DR2 + photometry to spectroscopic labels ([Queiroz+2020](#))



„Label transfer“: Spectroscopy (<10M stars) → Gaia + photometry (>300M stars)

Test dataset:



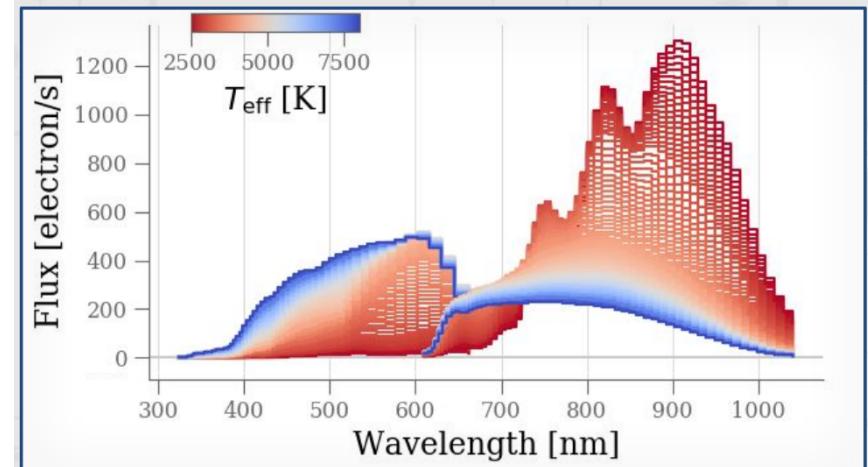
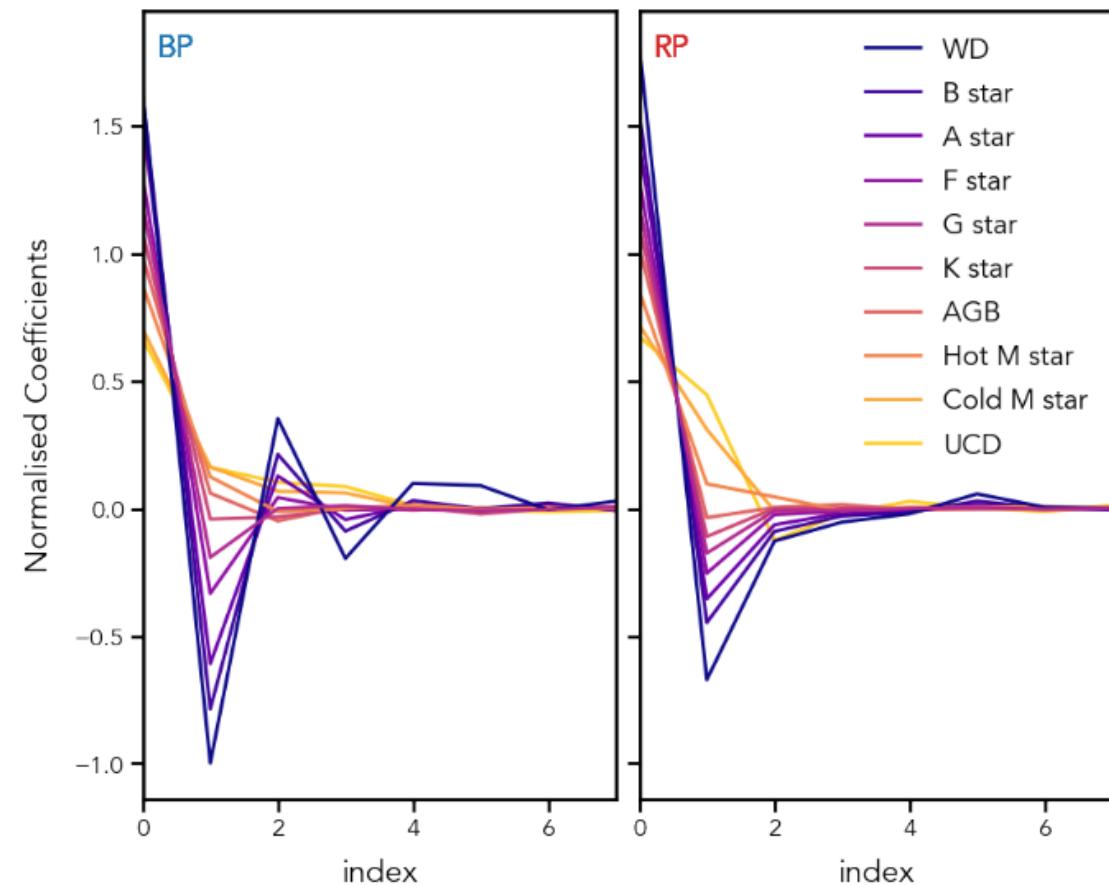


gaia DR3: 220M BP/RP spectra!

CU8 presentation
O. Creevey EAS 2021



Use of mean Bp and Rp Spectra



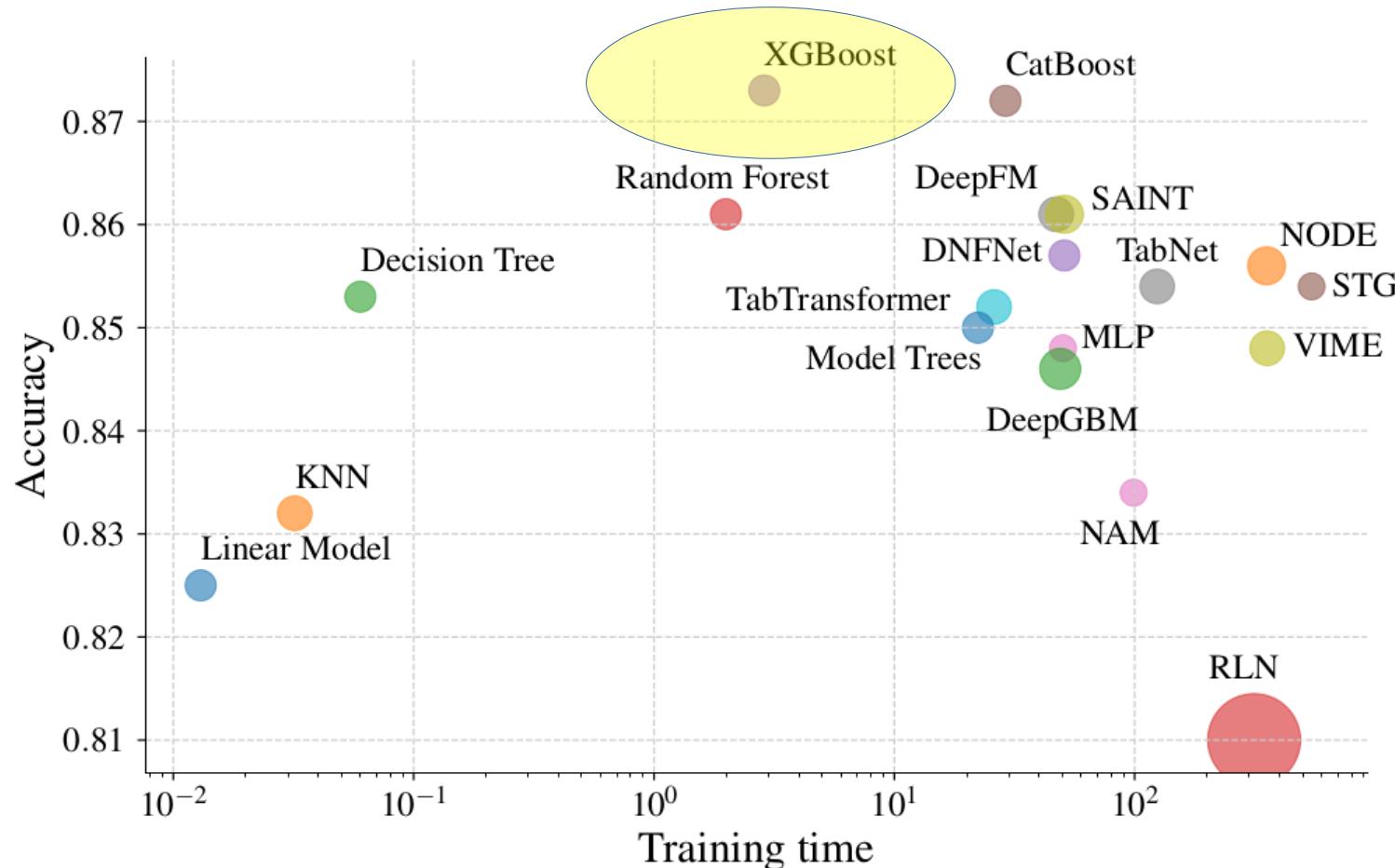
Bp and Rp spectra produced by Gaia-DPAC-CU5/DPCI
Astrophysical Parameters (APs) produced by Gaia-DPAC-CU8/DPCC

- De Angeli+2023:**
- First eight XP coefficients (see Carrasco+2021) of the continuous representation in BP and RP



gaia DR3 StarHorseNet → SHBoost

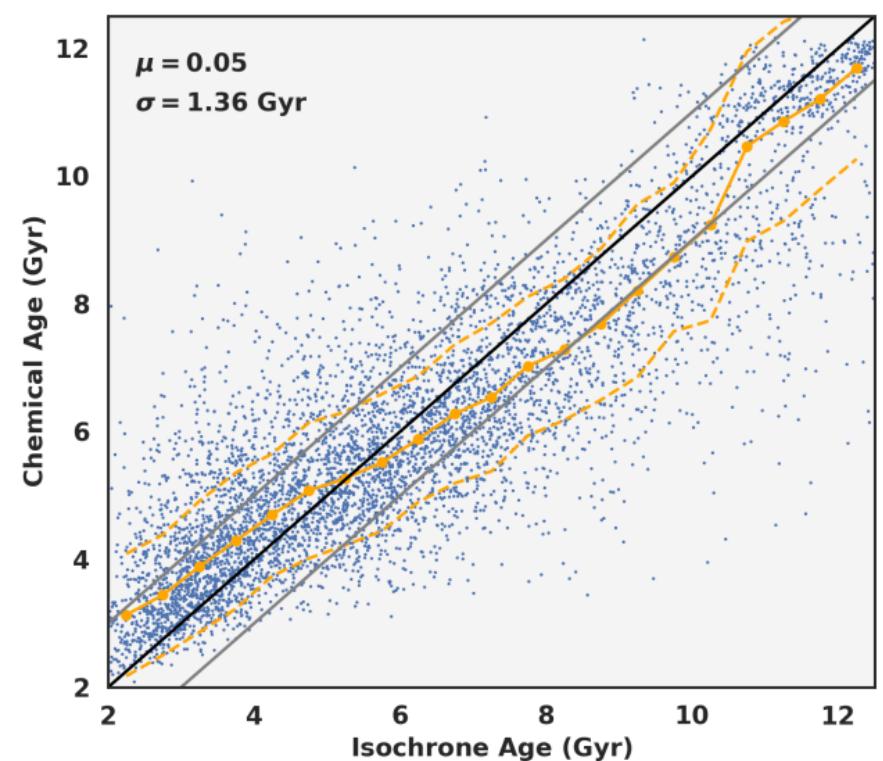
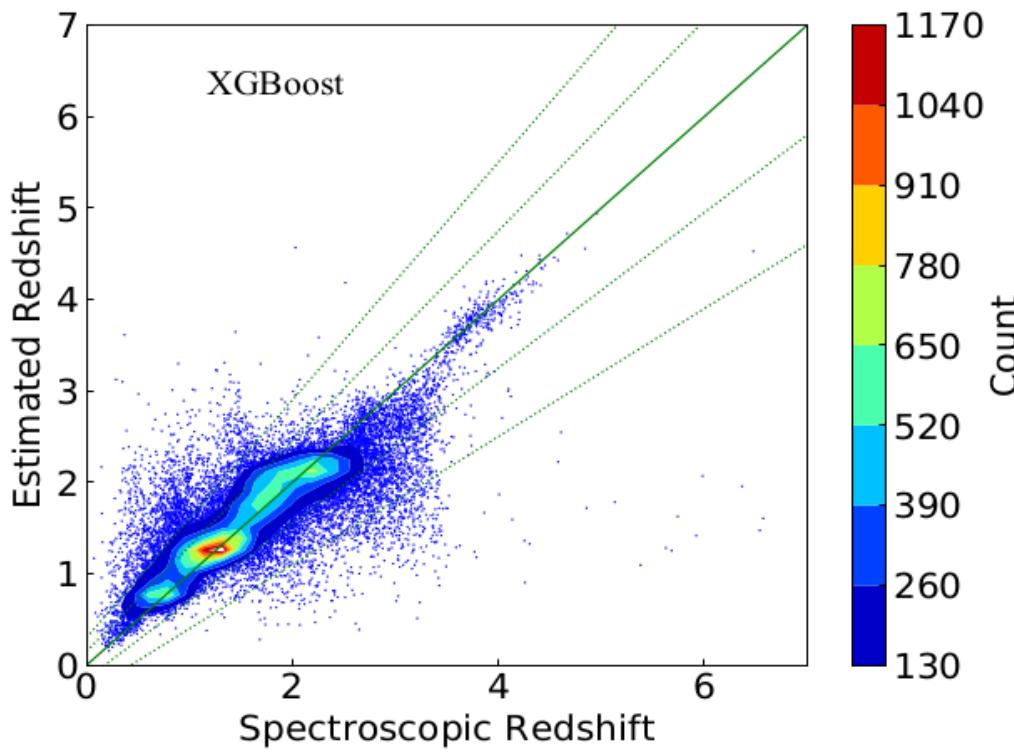
- DR3 includes 220M XP spectra in the form of a coefficient table...
→ project to derive higher-precision stellar labels (with/without XP spectra)



Borisov+2021 benchmark-tested regression algorithms for tabular data

XGBoost regression in astronomy

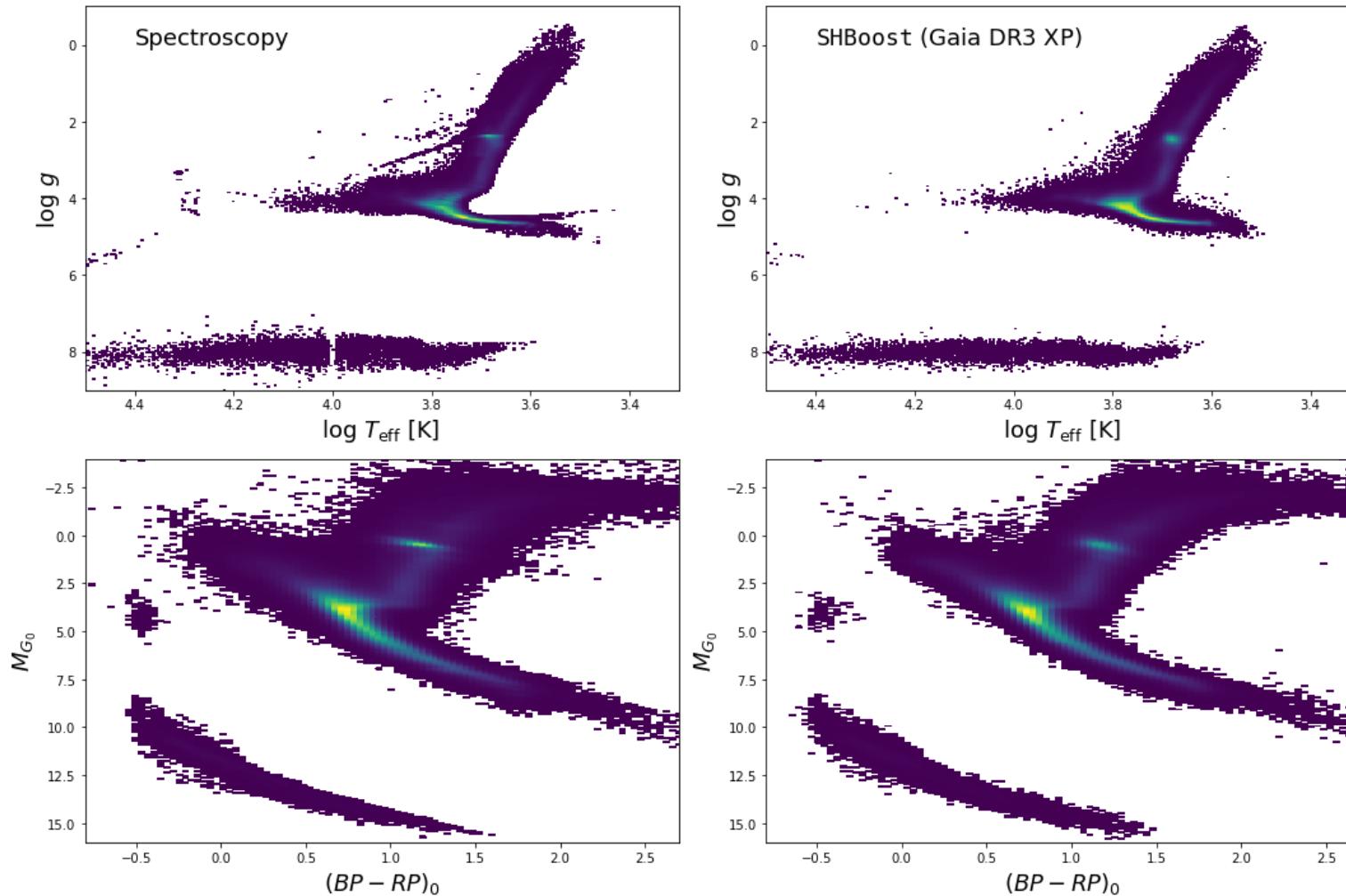
- Extreme Gradient-Boosted Trees *python* package *xgboost*: „scalable, portable and accurate“ implementation of boosted trees (Friedman 2001, [Chen & Guestrin 2016](#))
- Widely used for classification in astro: e.g. [Bethapudi & Desai 2018](#); [Yi et al. 2019](#); [Li et al. 2019](#); [Cunha & Humphrey 2022](#)
- Examples for regression:
 - **Photometric redshifts** ([Chong & Yang 2019](#); [Li+2022](#); [Humphrey+2023](#))
 - **Number of sunspots** ([Dang+2022](#))
 - **Spectroscopic stellar ages** ([Hayden+2022](#); [Anders+2023](#))





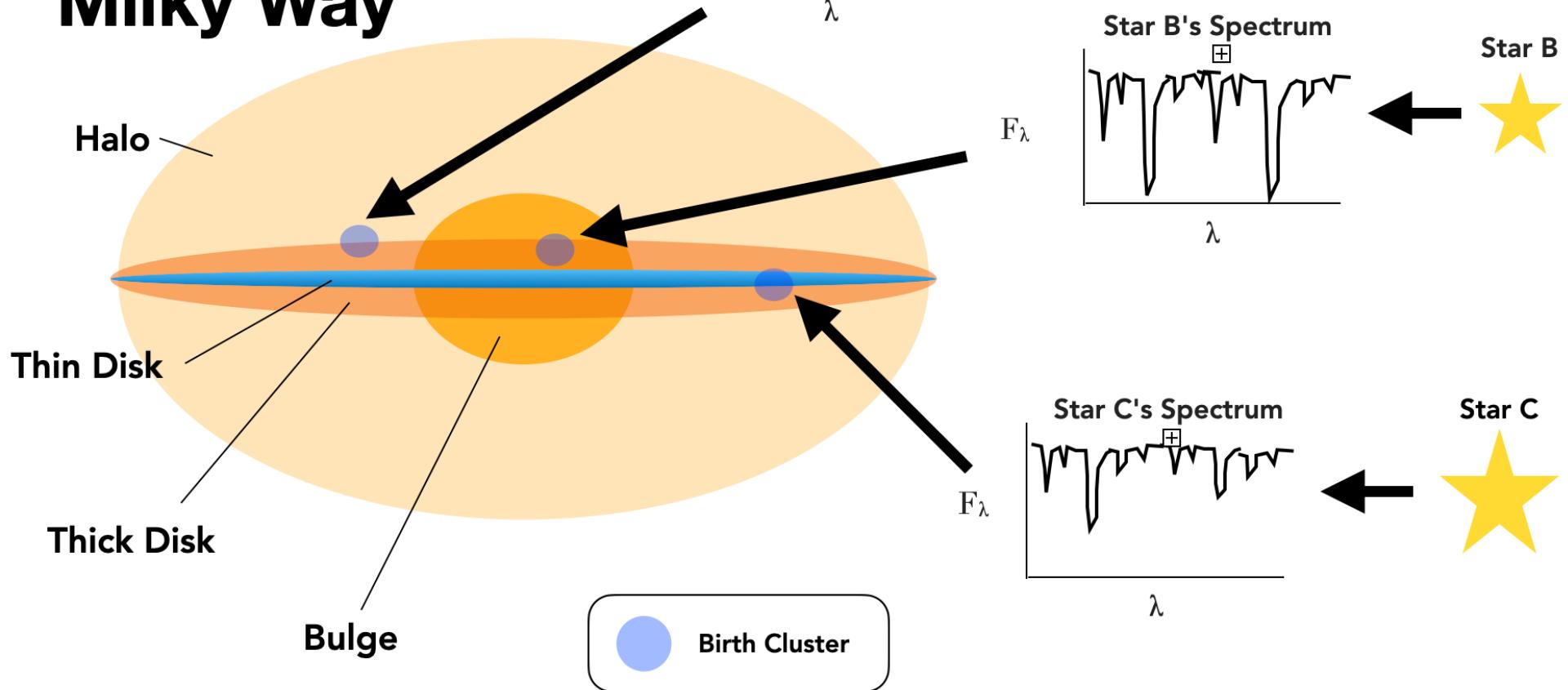
gaia DR3 StarHorse → SHBoost

- XGBoost works very nicely (even without parameter tuning) to predict StarHorse-like output parameters (d , Av , T_{eff} , $\log g$, $[\text{M}/\text{H}]$, mass)
- Much better metallicities (~ 0.15 dex) thanks to Gaia XP spectra (also: [Andrae+2023](#))
- Also works well for white dwarfs and hot stars (larger training set)



Science case III: Chemical tagging

Chemical Tagging in the Milky Way

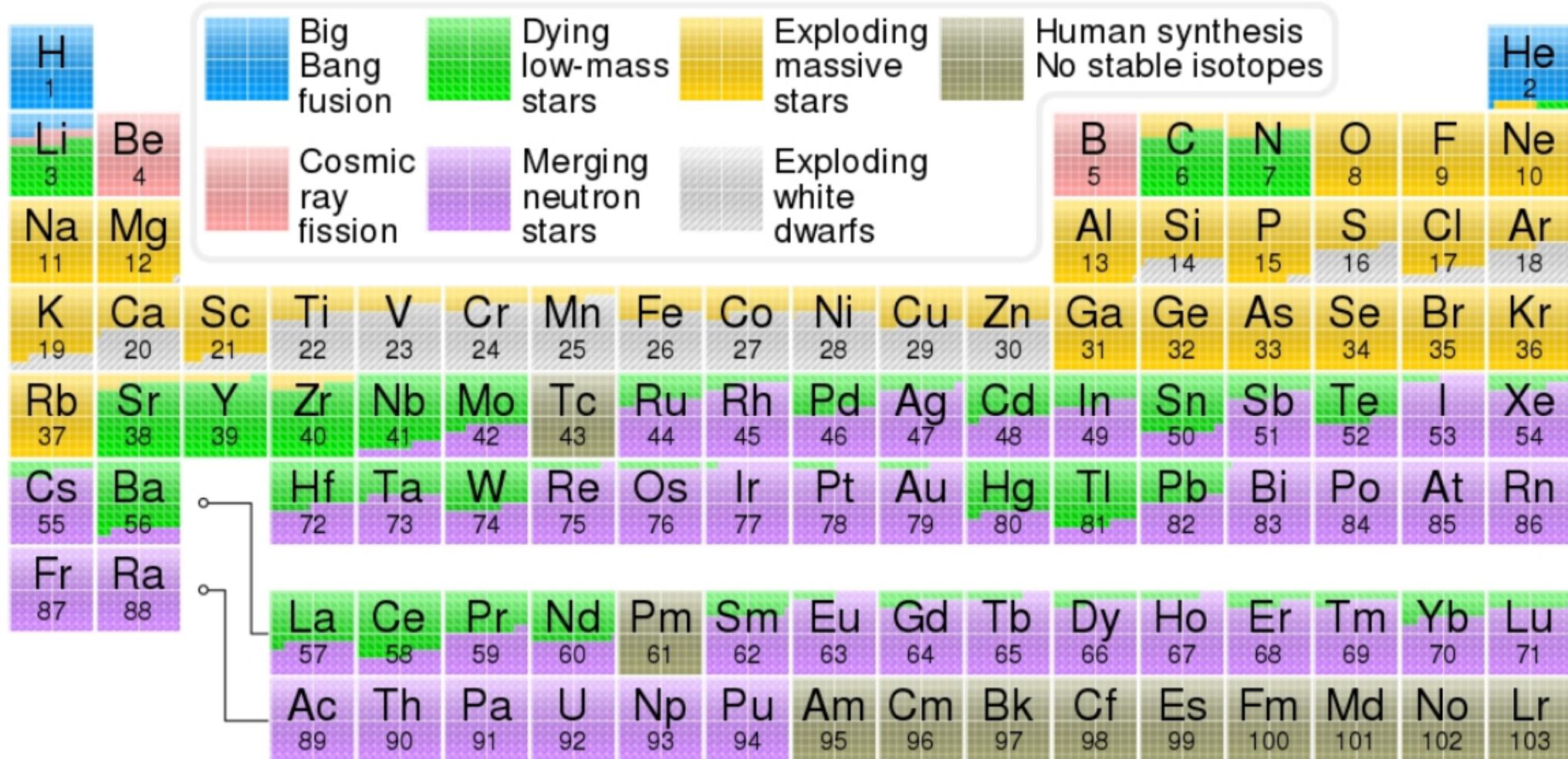


Science case III: Chemical tagging

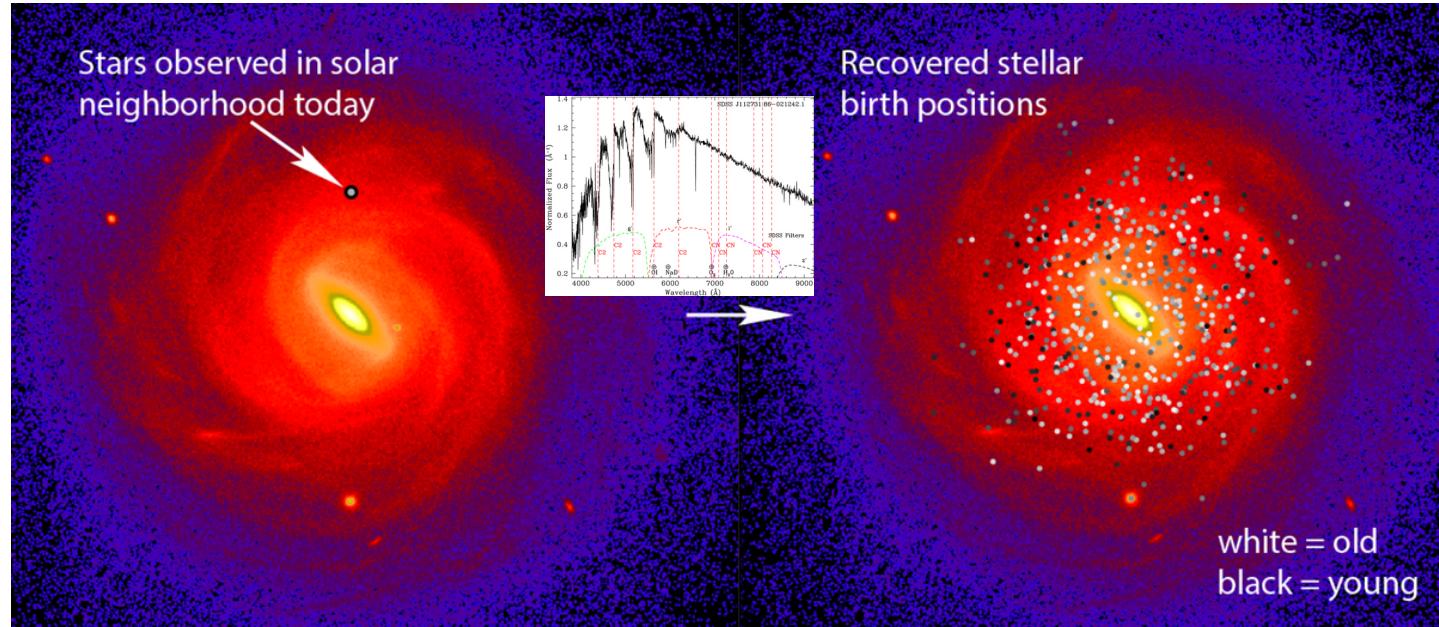
Nature/Type	Classification	Regression	Clustering	Forecasting	Generation	Discovery	Insight
Image	•			•	•		•
Spectroscopy	•	•	•	•	•	•	•
Photometry	•	•	•	•		•	•
Light curve	•	•	•	•		•	•
Time Series	•	•	•	•	•	•	•
Catalogue	•	•	•	•	•	•	•
Simulation	•	•		•	•	•	•

Science case III: Chemical tagging

The origin of the elements



Science case III: Chemical tagging



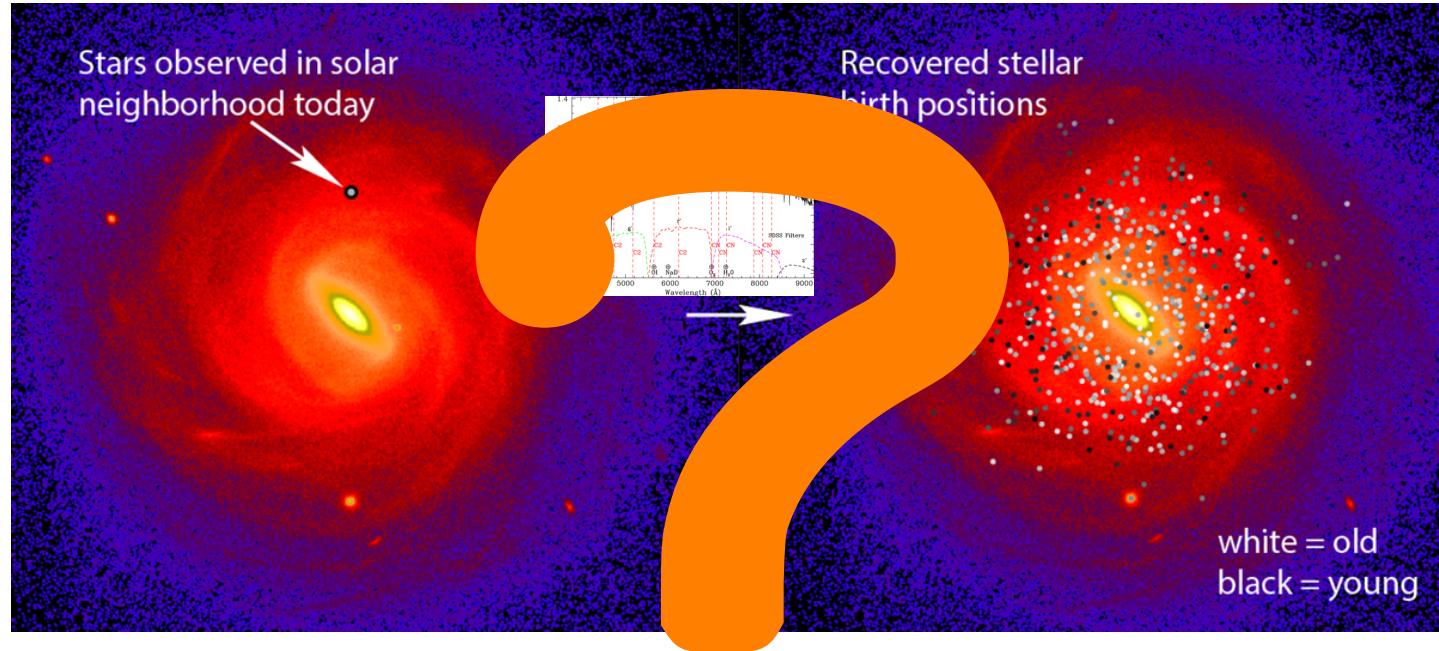
Chemical Signatures

Minchev+2018 PR

A major goal of near-field cosmology is to tag or to associate individual stars with elements of the protocloud. For many halo stars, and some outer bulge stars, this may be possible with phase space information provided by Gaia. But for much of the bulge and the disk, secular processes cause the populations to become relaxed (i.e., the integrals of motion are partially randomized). In order to have any chance of unravelling disk formation, we must explore chemical signatures in the stellar spectrum. Ideally, we would like to tag a large sample of representative stars with a precise time and a precise site of formation.

Freeman & Bland-Hawthorn 2002

Science case III: Chemical tagging



Chemical Signatures

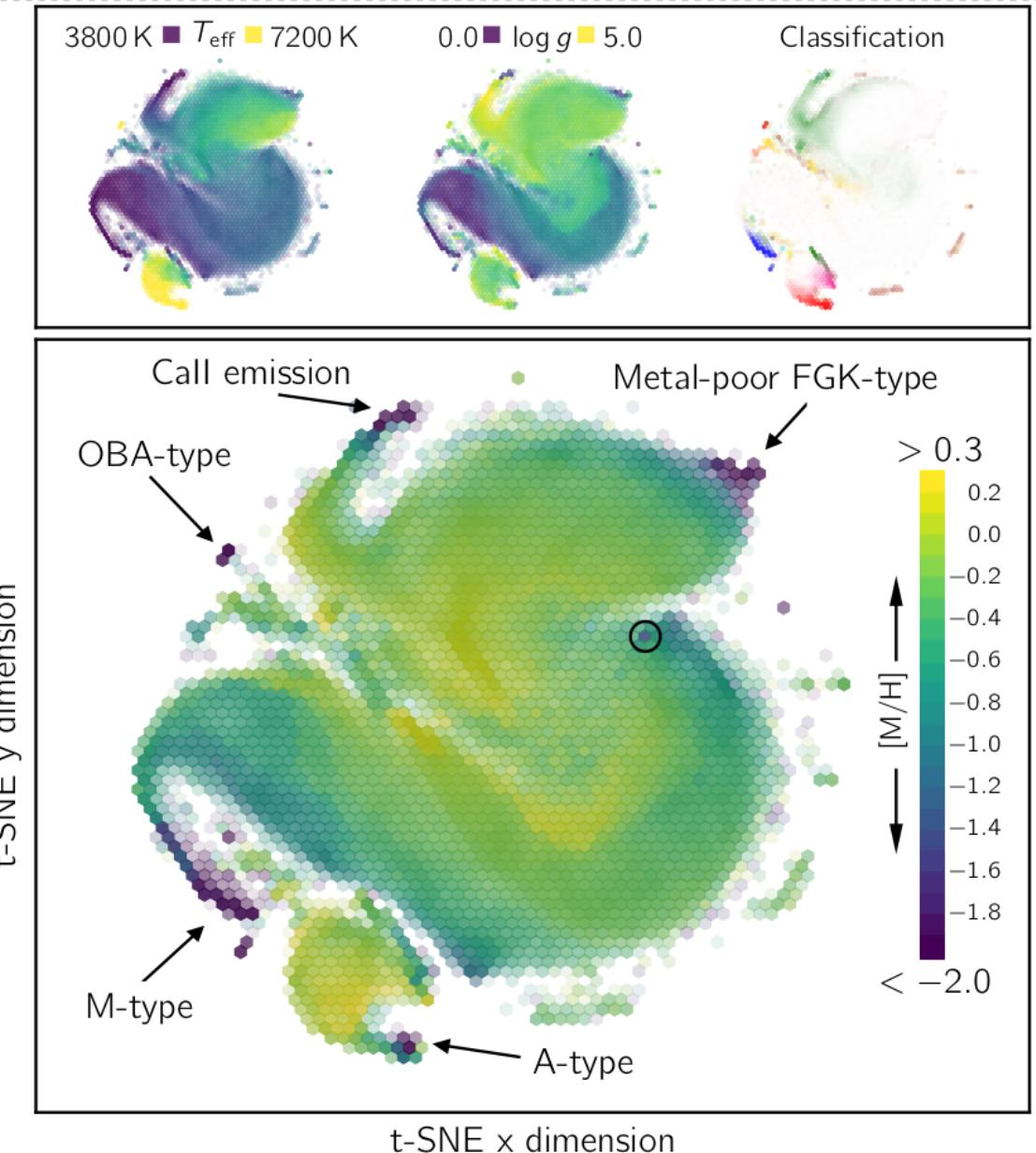
Minchev+2018 PR

A major goal of near-field cosmology is to tag or to associate individual stars with elements of the protocloud. For many halo stars, and some outer bulge stars, this may be possible with phase space information provided by Gaia. But for much of the bulge and the disk, secular processes cause the populations to become relaxed (i.e., the integrals of motion are partially randomized). In order to have any chance of unravelling disk formation, we must explore chemical signatures in the stellar spectrum. Ideally, we would like to tag a large sample of representative stars with a precise time and a precise site of formation.

Freeman & Bland-Hawthorn 2002

An „easy“ example: Finding very metal-poor stars in surveys

Matijevic+2017:
First t-SNE analysis of stellar spectra



An „easy“ example: Finding very metal-poor stars in surveys

Matijevic+2017

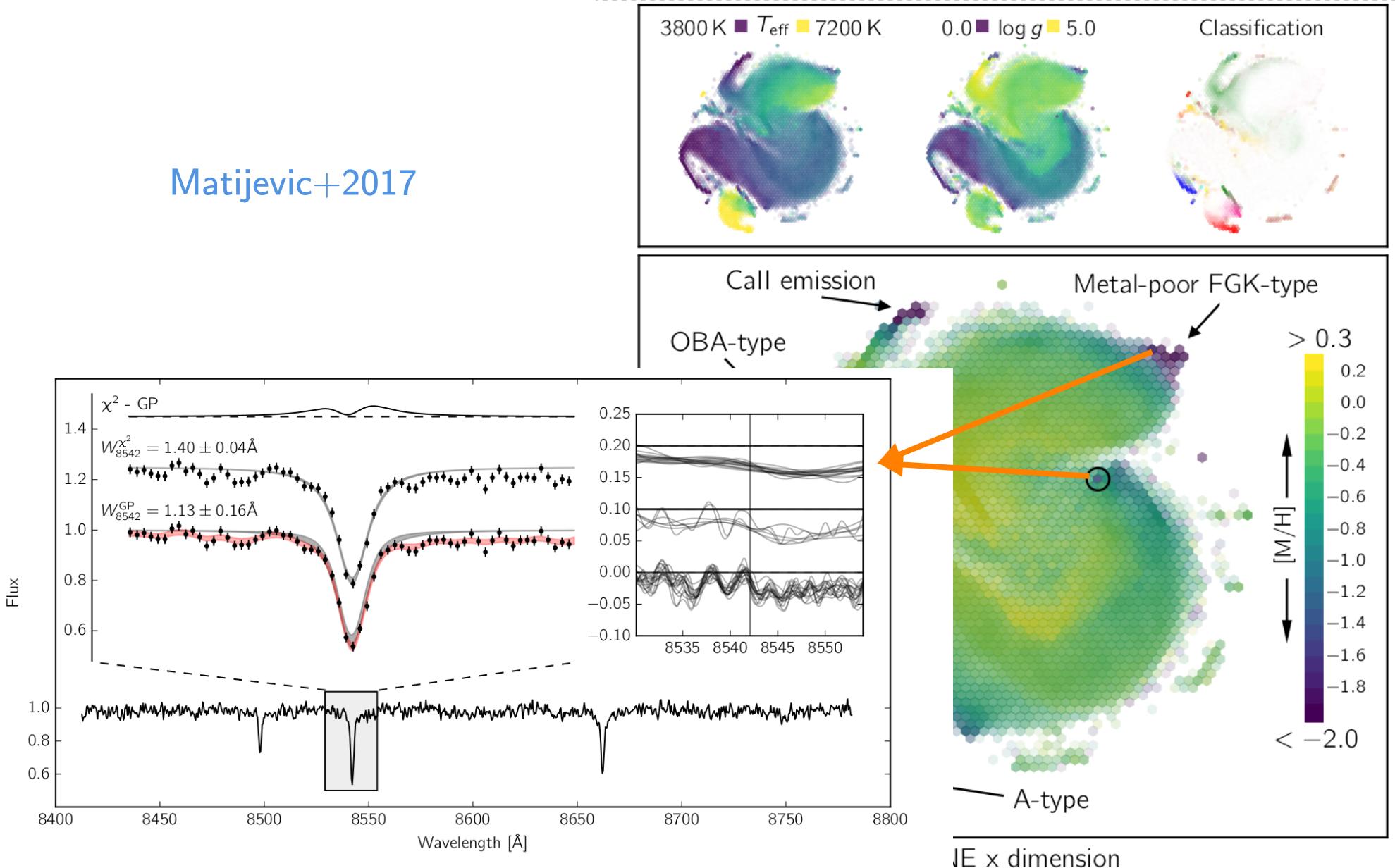


Fig. 4. Bottom part: one extremely metal-poor star spectra of RAVE. The metallicity of this star was estimated by our study to be $[\text{Fe}/\text{H}] = -3.04 \pm 0.22 \text{ dex}$ and the measured S/N per pixel is ~ 52 . The gray box marks the region around the central Ca II line for which the two solutions

Testing the idea of strong chemical tagging with HDBSCAN

Casamiquela+2021 Idealised scenario:

- All sample stars are open cluster members
- High-resolution ($R > 45,000$), high SNR (> 70) spectra
- Only red-clump stars \rightarrow no abundance trends with temperature or gravity

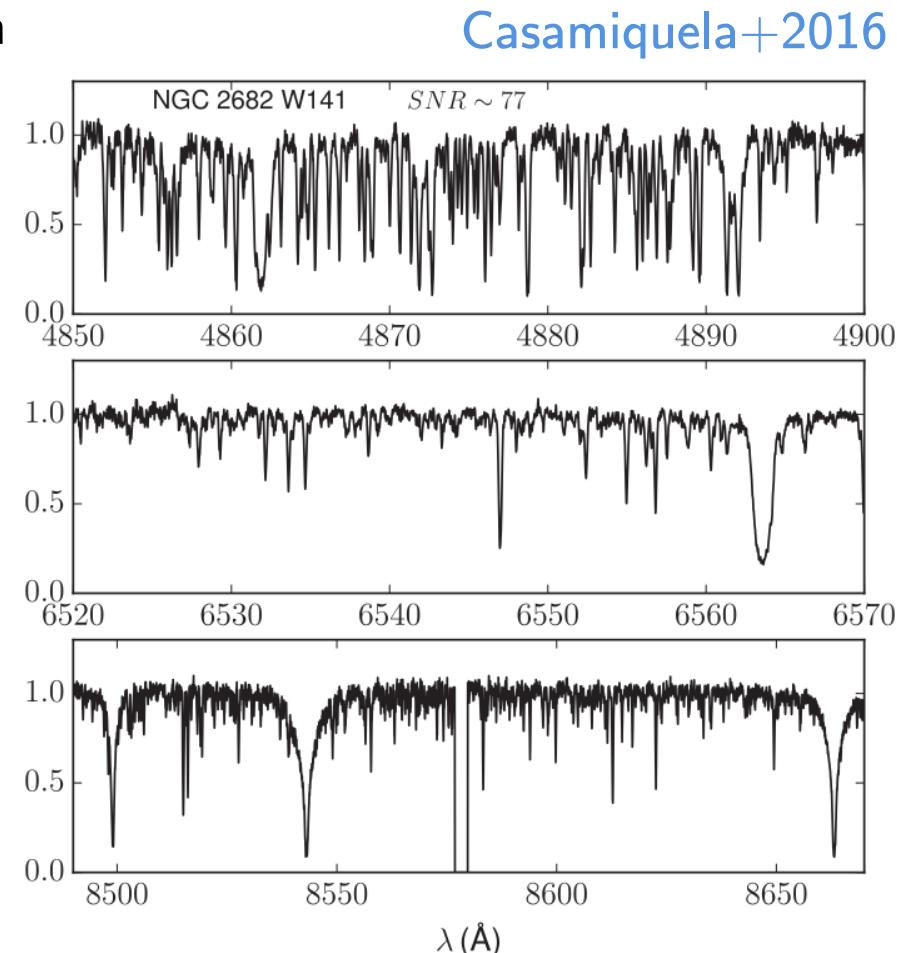
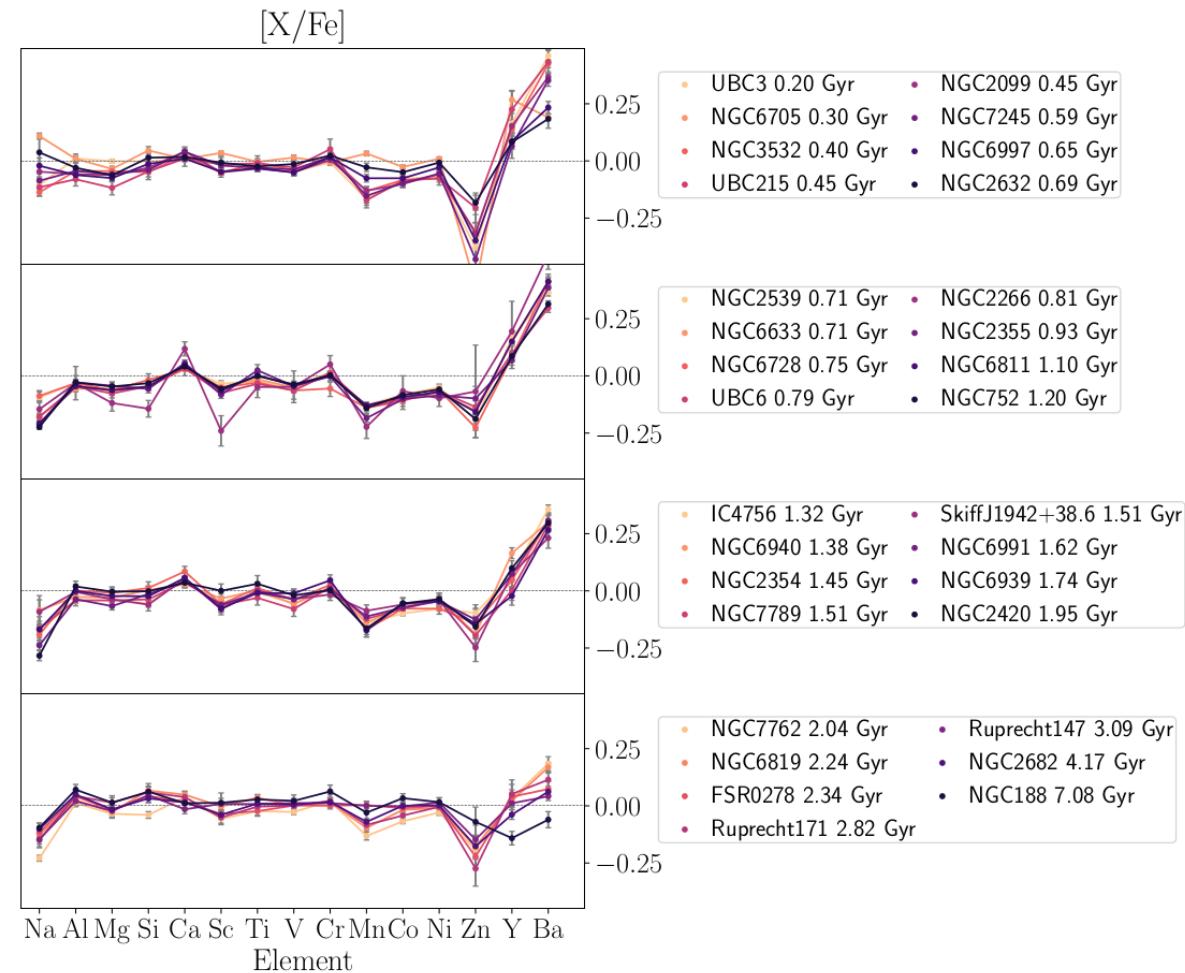
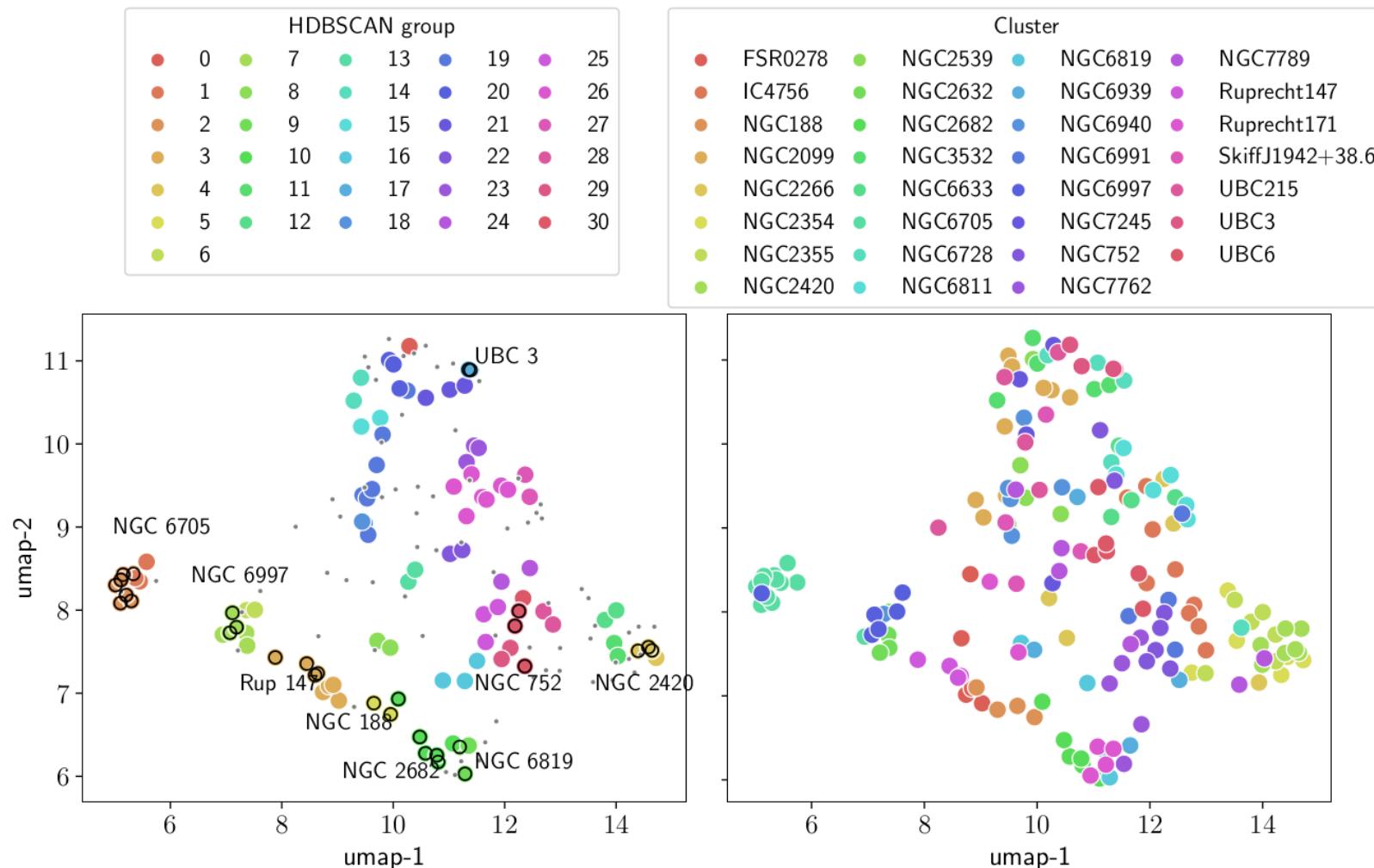


Figure 3. The Ca triplet (bottom), H α (middle) and H β (top) regions of the final combined and normalized spectrum of the star NGC 2682 W141 observed with HERMES ($\text{SNR} \sim 77$). A small gap from the order merging

Testing the idea of strong chemical tagging with HDBSCAN

Casamiquela+2021 Idealised scenario:

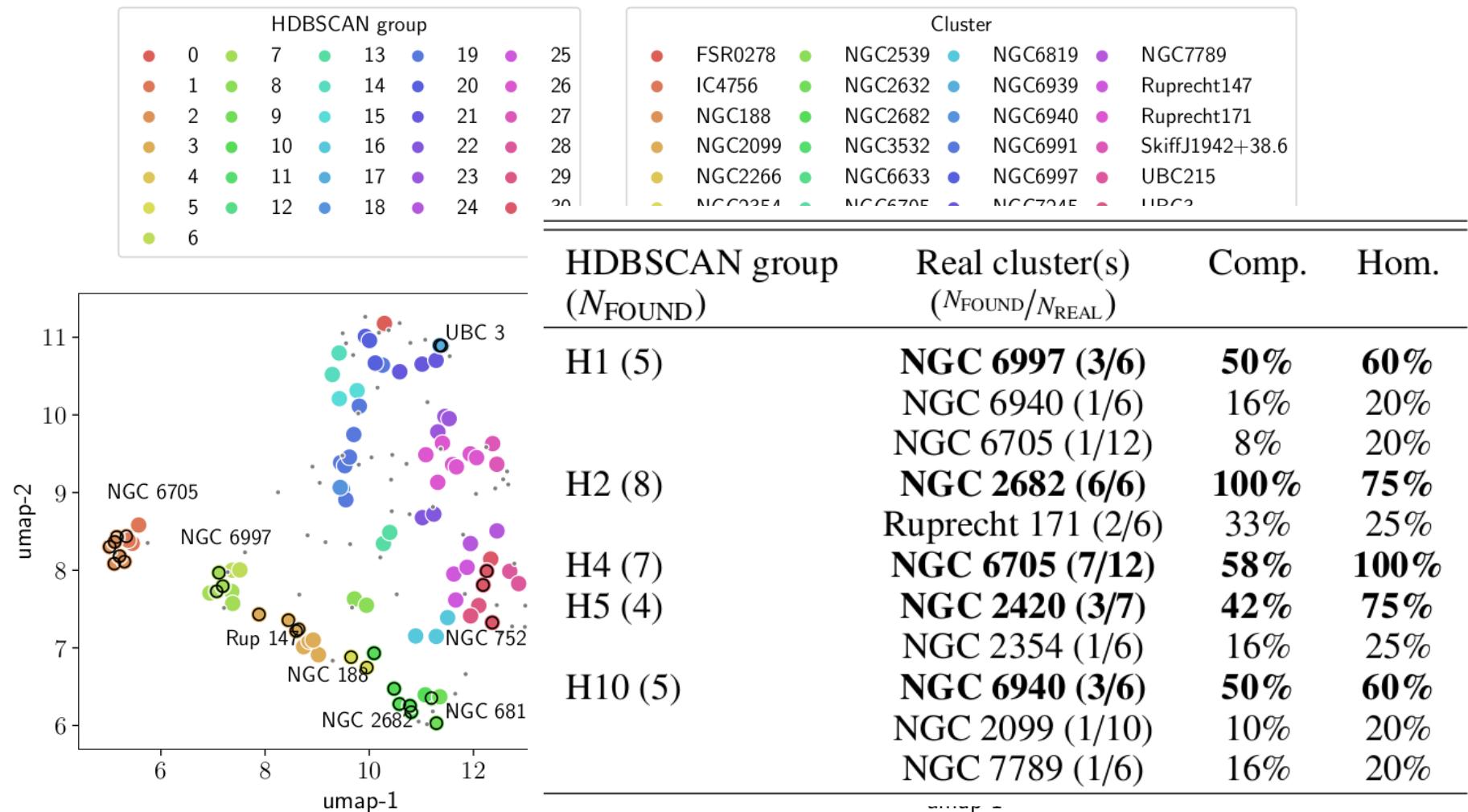
- HDBSCAN finds groups in abundance space → **no correspondence** with physical clusters
 - umap (similar to t-SNE) only used for visualisation



Testing the idea of strong chemical tagging with HDBSCAN

Casamiquela+2021 Idealised scenario:

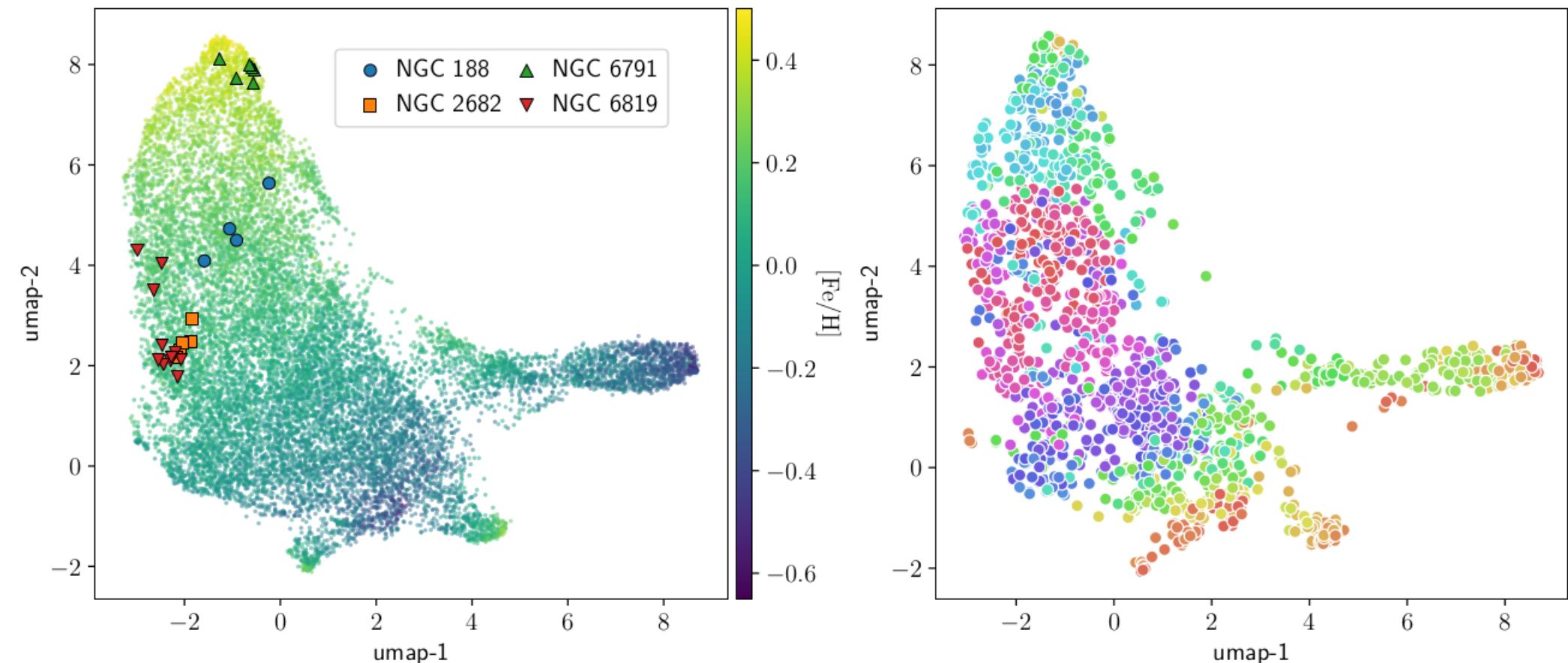
- HDBSCAN finds groups in abundance space → **never 1:1 correspondence** with physical clusters



Testing the idea of strong chemical tagging with HDBSCAN

[Casamiquela+2021](#) More realistic scenario (APOGEE survey):

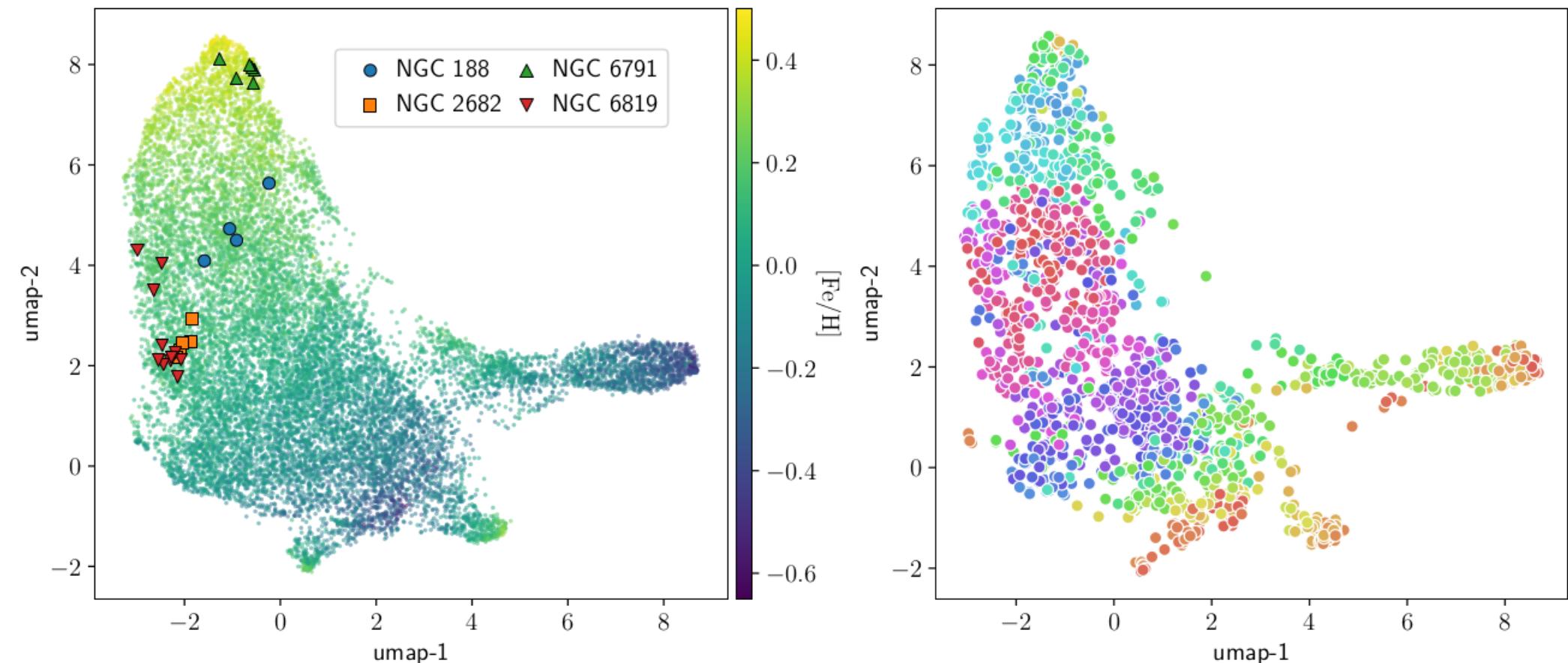
- As expected, HDBSCAN finds groups, but the physical correspondence is even worse



Testing the idea of strong chemical tagging with HDBSCAN

Casamiquela+2021 More realistic scenario (APOGEE survey – including field stars):

- As expected, HDBSCAN finds groups, but the physical correspondence is even worse



Summary

Nature/Type	Classification	Regression	Clustering	Forecasting	Generation	Discovery	Insight
Image	•			•	•		
Spectroscopy	•	•	•	•	•	•	•
Photometry	•	•	•	•	•	•	•
Light curve	•	•	•	•	•	•	•
Time Series	•	•	•	•	•	•	•
Catalogue	•	•	•	•	•	•	•
Simulation	•	•			•	•	•

Summary

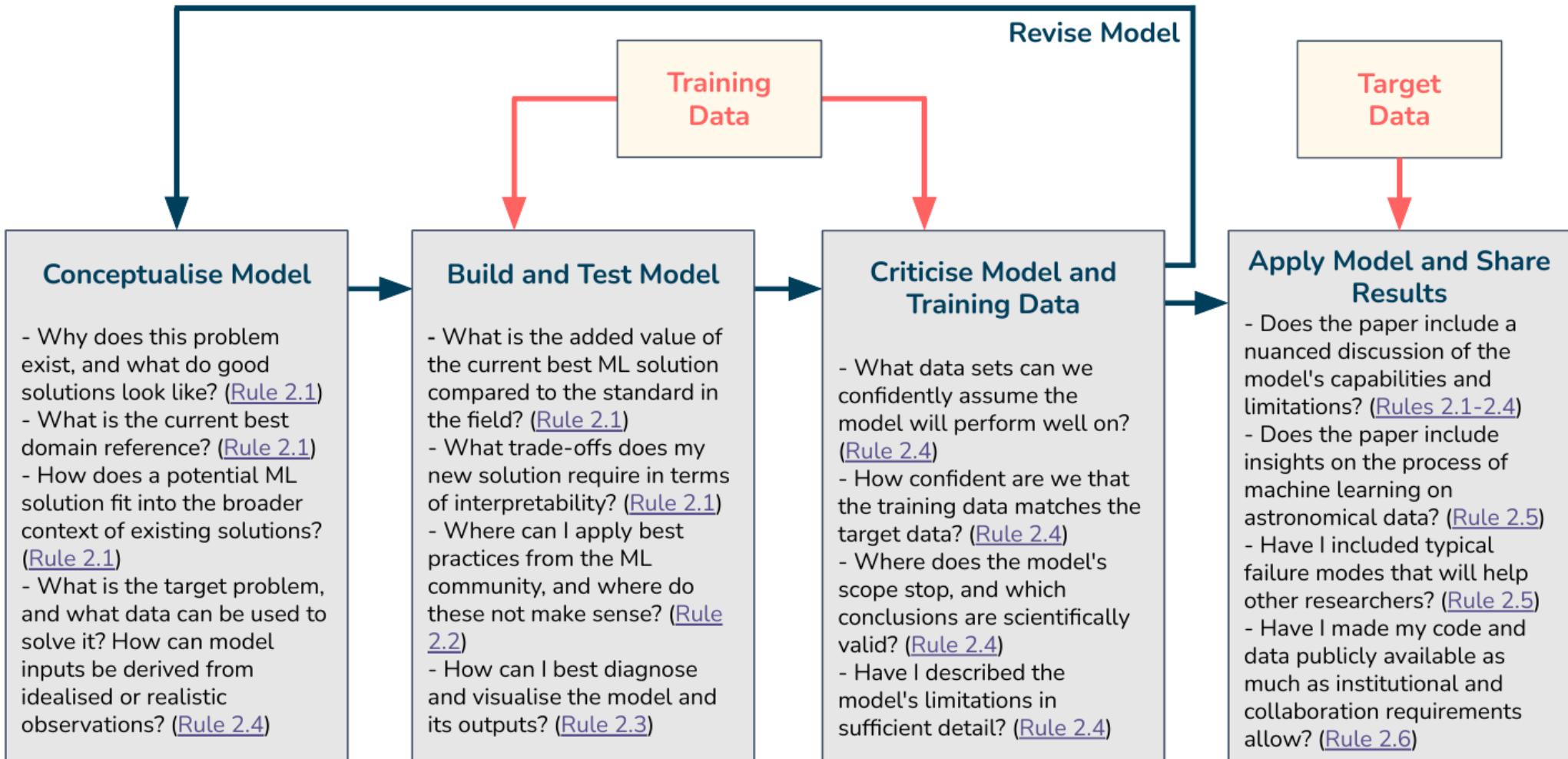


Figure 2. Box's loop for ML in astronomy.

Summary

A. A QUICK-START GUIDE FOR ASSESSING ML ASTRONOMY RESEARCH

Interdisciplinary results can be difficult to assess because they require a deep understanding not only of the scientific domain, but also of the methodology. This quick-start guide is intended as a starting point for readers and referees to assess new research for which they have a domain understanding but may lack methodological context. For referees, we caution that this guide is not absolutely prescriptive, nor exhaustive. Referees should consult journal expectations (e.g. [American Astronomical Society 2022](#); [MNRAS 2022](#); [Nature 2022](#)) and more general refereeing references ([Wager et al. 2002](#); [Nicholas & Gordon 2011](#); [Raff 2013](#); [Ntampaka et al. 2022](#)) for guidelines on best practices for providing evaluations of manuscripts. The following considerations do not replace refereeing best practices; instead, they are additions to best practices that are specific to evaluating ML astronomy research.

1. Compare against a domain reference and put results in the broader context.

- (a) Are the results put in the appropriate context? For example, [if the method replaces an existing “traditional” technique, are results \(accuracy, compute cost, robustness, etc.\) from the traditional technique used as a comparison?](#)
- (b) In some cases, the new ML method enables an analysis that was not possible before; no traditional benchmark exists. In this case, it may be more difficult to put the results in context.
- (c) If the outputs of the model are likely to be used downstream (e.g. for population-level analyses), [do the authors consider how biases in their model might propagate into these analyses?](#)

2. Adopt best practices from the ML community.

- (a) Have the authors included citations in their literature review to summarize particular best practices applied in their work?

Talk to your colleagues & organise something!

Machine Learning at the ICCUB

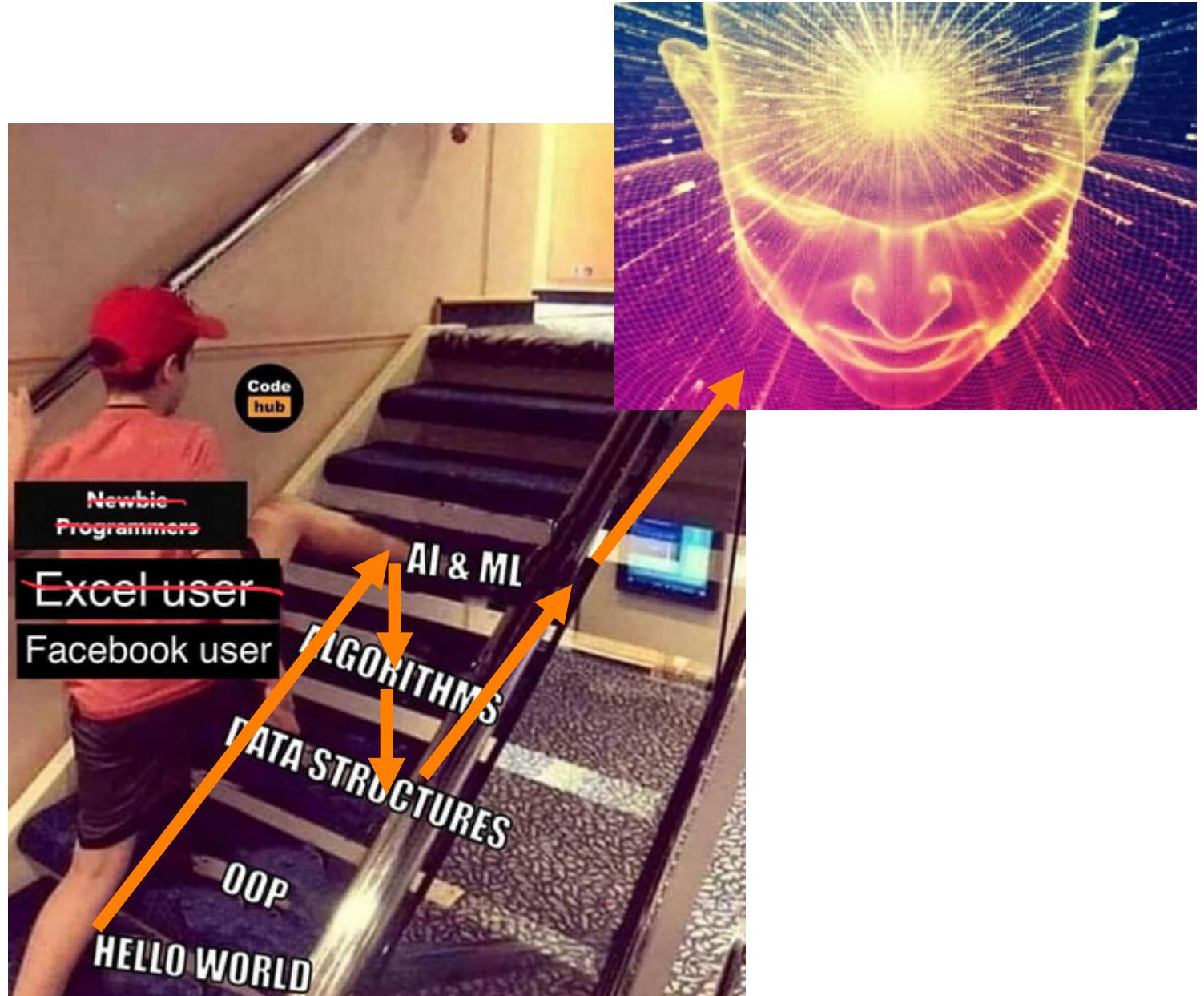
Oct 23–24, 2023

UB Physics Faculty, Sala de Graus Eduard Fontserè

Europe/Madrid timezone

10:00	Transforming Healthcare: Machine Learning and Computer Vision at the Forefront (Invited Talk)	Prof. Piella Fenoy	
	Machine learning in LHCb	Lukas Calefice	
	<i>Sala de Graus Eduard Fontserè, UB Physics Faculty, Sala de Graus Eduard Fontserè</i>	11:30 - 12:00	
12:00	Quasar classification and resdhift estimation	Ignasi Pérez	
14:00	Data Science Approaches to Physics and Complex Systems (Invited talk)	Dimitri Marinelli	
15:00	Real vs. bogus classifier in time-domain surveys	Dr Nadia Blagorodnova	
	<i>Sala de Graus Eduard Fontserè, UB Physics Faculty, Sala de Graus Eduard Fontserè</i>	15:00 - 15:30	
	Search for lensed gravitational-wave events with machine learning	Mr Roberto Bada Nerín	

→ Small but very successful workshop:)



Contact:
fanders@fqa.ub.edu
<https://fjaellet.github.io>