universität
wien

# DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Finding needles in the Galactic haystack:
Towards interpretable machine learning methods
to identify stellar structures in the Milky Way“

verfasst von / submitted by

Dipl.-Ing. Sebastian Ratzenböck, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktor der Technischen Wissenschaften (Dr. techn.)

Wien, 2022 / Vienna 2022

| | |
|---|---|
| Studienkennzahl lt. Studienblatt /<br>degree programme code as it appears on the student record sheet: | UA 786 880 |
| Dissertationsgebiet lt. Studienblatt /<br>field of study as it appears on the student record sheet: | Informatik |
| Betreut von / Supervisor: | Univ.-Prof. Torsten Möller, PhD<br>Univ.-Prof. João Alves, PhD |

# Acknowledgements

I am grateful to Torsten Möller for supporting me during my doctoral studies. Torsten has had a massive impact on my research and has shaped my critical thinking, problem-solving skills, and approach to science. I am also grateful to João Alves, who spent countless hours with me looking at and discussing results, sessions that I always left with a great fascination and passion for physics and our universe, sparked by João's enthusiasm. In addition, I am grateful to Immanuel M. Bomze, who provided valuable feedback and offered great advice throughout the thesis. Their combined contributions extend far beyond this work.

I am incredibly grateful to my family, especially my mother and father, for their unconditional support since day one. This thesis would have also not been possible without the support from colleagues, friends, and especially my partner Marlena who supported me through all ups and downs during the past years. Thank you for being there!

I want to extend a special thank you to the research network Data Science @ Uni Vienna and the University of Vienna, whose generous funding and support enabled me to pursue a PhD.

This work presents results from the European Space Agency (ESA) space mission Gaia. Gaia data are being processed by the Gaia Data Processing and Analysis Consortium (DPAC). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the Gaia MultiLateral Agreement (MLA). The Gaia mission website is https://www.cosmos.esa.int/gaia. The Gaia archive website is https://archives.esac.esa.int/gaia.

# Abstract

This doctoral thesis aims to take a different look at stellar structure censuses in the Milky Way. Specifically, it aims to provide interpretable analysis methods to uncover both previously unknown stellar structures and new members of known stellar populations, providing astronomers with a more complete picture of the different stellar structures in the local Milky Way.

The thesis contributions to the field are twofold: first, it introduces `Uncover`, an extended membership analysis technique that integrates known members of star clusters to search for yet undetected cluster members. `Uncover` is successfully applied to two different use cases, the recently discovered Meingast 1 stream, a Pleiades-age structure covering about 120° of the sky, and the well-studied star-forming region $\rho$ Ophiuchus. For these two very different stellar structures, `Uncover` increased the number of members by tenfold and by about 200, respectively. Second, the thesis introduces Significance Mode Analysis (`SigMA`), an innovative clustering algorithm that studies the topological properties of the density field in multidimensional phase space. The application of `SigMA` to Gaia EDR3 data of the closest young association to Earth, the Scorpio-Centaurus (Sco-Cen) association, finds, for the first time, 48 co-moving and coeval clusters in Sco-Cen, many of them previously unknown. These 48 clusters are independently validated using astrophysical knowledge unknown to `SigMA`.

Both `Uncover` and `SigMA` are formulated in domain-specific language, use expressive hyper-parameters, and allow for result validation to provide confidence in the results. With these tools, we seek to contribute to changing the current culture of blind acceptance of machine learning results and help astronomers build and modify models based on their expertise.

# Contents

# List of Tables

# List of Figures

# 1. Motivation

The ESA/Gaia [44, 42, 43] mission is an unprecedented all-sky survey, providing astronomers with high-precision positions and kinematic measurements 200 times more accurate than the predecessor mission Hipparcos [92]. For the first time, position and kinematic information are available for over 1.6 billion stars, providing enough statistics to study physical processes to a level of detail never seen before. One of Gaia's key science objectives is to disentangle stellar populations in the Milky Way, by studying co-moving and coeval populations, or star clusters. Star clusters[1] are systems of a few dozen up to thousands stars that constitute the elementary building blocks of galaxies [75]. These stars are created in the same formation event and from the same collapsing molecular cloud and carry crucial information on star formation processes imprinted from their birthplaces, like velocity and chemical composition. Star clusters are valuable probes for studying fundamental processes such as the formation and evolution of the Milky Way structure, stellar physics, and exoplanet evolution [86]. Still, pre-Gaia, this meant we could only rely on the few high-contrast clusters, like the Pleiades, Hyades, or the still forming Orion Nebula Cluster. Gaia has completely changed the field, giving access to cluster identification well below the average stellar density of the background [82, 83].

Both empirical evidence and theoretical modeling support the basic notion that stars that were born together move together [61]. Hence, star clusters can be identified via positional and kinematic analyses, which have gained significant traction with the advent of Gaia and its high precision astrometric measurements. However, disentangling and extracting stellar populations is notoriously difficult. Firstly, as a consequence of interactions with the Milky Way potential and giant molecular clouds, these initially quite compact objects are stretched into elongated, sometimes non-convex structures in position space. This "galactic-stretching" leads to a variety of cluster shapes from very compact overdensities (when young), to low-contrast, spread out, s-shaped clusters dominated by the Milky Way tidal forces [81, 96]. Second, due to the low amount of available radial velocities, about 0.4% in Gaia's second data release (DR2 [42][2]), one is, for the

[1] This thesis uses the word "cluster" to describe a stellar structure, or a cluster of stars, in the statistical sense of the word cluster as an enhancement over a background, meaning, it does not discriminate between physically bounded or unbounded cluster, itself a difficult distinction to make observationally. These stellar structures include structures known in the literature as open clusters, associations, moving groups, streams, clusterings, aggregates, subgroups, etc., i.e., coeval and co-moving stellar populations.

[2] This work used Gaia DR2 data, and as soon as it was available, it switched to data from the early data release three (EDR3 [43]). EDR3 replaced the second data release on December 3, 2020. While DR2 is based on observations from the first 22 months of the mission, the EDR3 catalog summarizes measurements from the first 34 months. Regarding measurements related to this work, EDR3 features about 10% more sources with the so-called five-parameter solutions, i.e. celestial positions, parallaxes,

most part, restricted to two tangential velocity axes. Thus, even if we assumed perfectly Gaussian distributed three-dimensional (3D) velocities within these groups, the on-sky projection to the curved two-dimensional (2D) surface can distort original symmetrical convex shapes into arbitrary, non-convex shapes depending on the size, orientation, and distance to the stellar group. Complicating matters further, members of star clusters make up only a tiny subset of the data, with field stars generating background noise that cannot be easily removed in 5D phase space[3]. The (usually) five-dimensional phase space is populated with stellar clusters of various shapes and densities embedded in a sea of noise, making parametric clustering algorithms practically unfeasible. Such distributions are too complex to be modeled precisely in a functional form, and the total number of clusters is also unknown. These circumstances make extracting clusters with a high signal-to-noise ratio a difficult task, especially in the low-density regime.

Nevertheless, Gaia has had a major impact on the discovery and characterization of previously unknown clusters, and even uncovered an entirely new type of stellar structure, Milky Way disk stellar streams [82, 96, 63, 81, 8, 69, 58, 94, 83, 59]. The origin of the new shapes is not yet fully understood and is a hot topic of research, particularly for young stellar clusters.

This work identifies two main modes of how stellar clusters are analyzed in the literature. First, astronomers aim to uncover yet unidentified member stars of already known stellar clusters. Typically, new star clusters' discoveries consist of small high-confidence samples that minimize misclassification of stars. However, larger samples would not only dramatically improve the quality of the derived cluster's physical parameters but also uncover the so far unseen low-density regions of stellar clusters, containing precious information on the cluster formation and evolution [12, 31, 40, 83]. Additionally, such improved membership lists often add low mass stellar members, which are the less prominent cluster members and are often not included in previous membership lists. To know the complete stellar cluster membership, including low-mass stars, allows for improved statistics, in particular on the shape of the initial mass function (IMF), which gives a statistical overview of the probability of the mass distribution in Milky Way stellar clusters. This thesis finds that the existing literature (see Sect. 1.2.1 for a detailed analysis) on identifying new member stars often ignores previously identified members which we aim to incorporate.

The second main mode is to discover yet unidentified stellar groups in the Gaia data

and proper motions. The full six-dimensional phase space is available for 7.2 million stars (about 0.4%) of the entire catalog, as crucial radial velocities are mostly missing (unchanged to DR2). The upcoming Gaia data release three (DR3), planned for June 13th, 2022, aims to increase available radial velocities to about 33 million, facilitating more detailed membership studies.

[3] The standard definition of phase space is a set of coordinates required to determine the state of a particular physical system. Thus, a point in phase space corresponds to a single system state. In mechanics, the phase space is usually six-dimensional (6D) and consists of three position axes and three axes that encode the momentum of a system along these axes. Here, similar to a star's momentum, we consider its precisely determined velocity, the variable of interest in stellar cluster analyses [61]. In case of missing kinematic information, we relax our definition of phase space to a 5D feature space consisting of three positional axes and two velocity axes, as provided by the Gaia astrometry.

set. The meta-study of Kharchenko et al. [64], which was published pre-Gaia, lists 2267 probable stellar clusters in the Milky Way disk. Although the census was believed to be rather complete at the time, Gaia studies found that many established groups were random fluctuations, as well as new clusters [12, 11, 17, 16, 77, 105, 15, 22, 69, 63]. This hints to a large population of yet undetected star clusters which have evaded detection, most likely due to low phase-space density. This thesis finds that currently used clustering tools are not well suited for the task of extracting stellar groups and aims to improve upon existing techniques to provide effective and interpretable methods to identify yet unseen stellar groups.

## 1.1. Research aims

Disentangling and extracting stellar populations is notoriously difficult. However, the payoff is huge: a complete catalog of stellar populations in our Milky Way would serve as a baseline for investigations on several fundamental properties of stellar physics. Star clusters provide probes to tackle problems such as the assembly of the Milky Way, the initial mass function, local gas star formation rate, and timescales for planet formation (see also the reviews by McKee & Ostriker [80] and Kennicutt & Evans [62]).

Instead of a one-size-fits-all solution toward a complete catalog, research often happens in a two-pronged approach: first, star cluster discovery, and second, extended membership analysis. Separating this workflow into two steps is advantageous because practitioners can optimize tools for the specific use case. By improving currently used methods, this work aims to further the journey towards a complete catalog.

This thesis considers both extended membership analysis and cluster identification as machine learning problems. In both cases, the methodological perspective is to develop tools that are interpretable by design while at least maintaining or improving the effectiveness and accuracy of existing methods. This work, rather than directly working on astronomical research questions (RQ), aims to create resources that support sense-making, exploration, and discovery of new knowledge in astronomical data. To change the current culture of blindly trusting machine learning models, we focus on two issues where we see opportunities for improvement. First, we identify the need for comprehensible model selection procedures for existing powerful and efficient machine learning models. Second, we aim to develop innovative tools designed with interpretable hyper-parameters in mind to facilitate the model selection process. Further, this thesis's immediate goal is to validate developed methods on astronomical data directly.

While this thesis focuses on the development of innovative methods in the domain of astronomy, we expect our methods and findings to generalize to other applications with large unlabeled data with high noise, arbitrary cluster shapes, and heteroscedastic measurement errors where modeling the data distribution from first principles is tedious to impossible.

## 1.2. Research questions

Given the vast amount of data Gaia provides and its high-dimensional search space, recent analyses (almost entirely) are based on automated and computerized procedures. These challenges have sparked the employment of a multitude of analysis methods from the fields of data mining and statistical learning and the development of new tools tailored to astronomical data. This thesis critically assesses their performance in extended membership analysis (A) and clustering (B) and identify potential for further development.

### 1.2.1. A. Extended membership analysis

To uncover potentially new cluster members, star clusters are often subject to follow-up studies. Especially the advent of Gaia has sparked many extended membership analysis approaches, as it not only provides more precise positional and kinematic measurements, but also data on very faint stars never seen before.

Current membership analysis approaches can be divided into either unsupervised heuristics (e.g. Clusterix [6], [72], Meingast et al. [82], Röser et al. [95]) or probabilistic models (e.g. Sarro et al. [98], BANYAN $\Sigma$ [41], Cantat-Gaudin et al. [13], Gao [46], Jaehnig et al. [57], ML-MOC [1]) which predominantly focus on modeling stellar clusters as multivariate Gaussians in phase space. However, diverse non-convex star cluster shapes caused by multiple initial conditions and complex interactions with the Milky Way (e.g., tidal tails [82, 96]) introduce significant deviations from Gaussianity. In contrast, unsupervised methods ignore information obtained in previous studies, which could increase the recall and accuracy of searches revisiting known populations.

To the best of our knowledge, we not find in the astronomical literature any occurrence of supervised learning techniques for identifying unseen cluster members. These methods can incorporate a preliminary set of high-fidelity members and (depending on the employed algorithm) provide enough flexibility to estimate highly non-convex decision boundaries in high-dimensional space. However, due to the lack of labeled outlier data, only novelty detection methods (also called one-class algorithms) can successfully train a classifier from a list of cluster members.

A powerful algorithmic choice that adapts to highly non-linear decision scenarios is one-class support vector machines (OCSVM [99]). OCSVMs learn a tight and smooth boundary around a target data set. By applying the kernel trick, this boundary is highly flexible and can describe non-linear, arbitrarily shaped boundary regions. However, its extraordinary versatility quickly becomes its biggest drawback, as its performance depends heavily on the choice of input hyper-parameters. We conjecture that OCSVMs, although powerful, have not been applied to identify star cluster members because model selection is tedious and requires a high level of expert knowledge of the algorithm itself. This thesis identifies a need for comprehensible model selection procedures for OCSVM. These considerations culminate in the following research questions:

**A.1** How can members of previously studied star clusters guide the search for yet undetected member stars?

**A.2** How to apply novelty detection searches in the domain of astronomy with variously shaped groups where the target class is a minority among a sea of outliers and the available training data has unknown contamination from the outlier class.

**A.3** How to effectively provide an overview of the vast space of possible star classification models.

**A.4** How to decompose the membership identification process into small, interpretable steps. Specifically, how to support users to apply their domain expertise to assess the goodness of trained models and effectively build confidence in the final classifier among domain experts?

### 1.2.2. B. Cluster analysis

Finding coeval stellar populations amounts to a needle in a haystack-style search. Less than 5% of all stars are clustered; the rest – so-called field stars – form a non-uniform background in phase space. Various shapes, space densities, and cluster sizes make it a hard challenge to do right. This task is further complicated as (compared to supervised learning) principled model selection techniques (such as cross-validation) cannot be applied. The lack of labeled data means the lack of an optimization criterion. The unsupervised nature also implies that partitioning data into "meaningful" clusters is, in general, an ill-posed problem. Each clustering method comes with individual assumptions about what the clustered space should look like. While parametric clustering algorithms such as Gaussian mixture models (GMM [28]) are easy to interpret, non-parametric methods such as popular star cluster extraction methods DBSCAN [35] and HDBSCAN [10] have more complex selection functions. Although powerful, many heuristics use complex or incomprehensible hyper-parameters. Since no optimization objective exists, many practitioners fall back to manual trial and error searches, effectively aimlessly wandering through the hyper-parameter space. This challenging situation is also reflected in the wide variety of methods used on Gaia data to identify groups. These methods can be roughly separated into the following categories:[4]

- Parametric approaches (e.g., Cantat-Gaudin et al. [13])
- Non-parametric machine learning methods (e.g., DBSCAN [35] by Castro-Ginard et al. [16, 15], Zari et al. [119], Fürnkranz et al. [39], Hunt and Reffert [56]; HDBSCAN [10] by Kounkel et al. [69, 70], Hunt and Reffert [56], Kerr et al. [63]; OPTICS [4] by Ward et al. [114]; SNN [34] by Chen et al. [22]; EnLink [104] by Kos et al. [68] and Chen et al. [22])
- Analysis and machine learning techniques designed toward stellar cluster discovery (e.g. UPMASK [72] by Cantat-Gaudin et al. [12, 11], Peña Ramírez et al. [91]; StarGO [118] by Tang et al. [108], Pang et al. [89, 90]; and other unnamed analysis

[4] The following lists are not exhaustive literature surveys but rather intend to provide a rough overview of the current clustering landscape.

techniques by Kushniruk et al. [73], Oh et al. [87], Galli et al. [45], Meingast et al. [82])

Given the plethora of analysis initiatives that pursue the same goal, we highlight the following research directions. These directions address clustering challenges toward a consolidated star clustering approach: (1) Visual solution space exploration alongside clustering result validation options for domain experts (e.g., the Hertzsprung-Russell diagram (HRD[5] [38]). (2) Meaningful and interpretable hyper-parameters that alleviate or facilitate manual solution space exploration. (3) Internal validation criteria [79] (e.g., Silhouette score [97]) based on heuristic measures such as cluster compactness and separation, optimized for star cluster results, enable automatic model selection.

Of these three options discussed above, this work focuses on the second. It sees a-priori intrepretable tools (or interpretable hyper-parameters) as the fundamental level of model selection techniques. This thesis aims to achieve interpretability by modifying crucial internal algorithmic decisions to conceptually simpler paradigms without sacrificing accuracy. If this modification is not possible, the research directions (1) and (3) mentioned above represent alternative research avenues.

The goal of this work is to facilitate the search for star clusters in phase space. Providing meaningful clustering tools for star cluster extraction faces these challenges:

**B.1** How to increase interpretability and robustness of model agnostic, density-based clustering methods that generalize to variously shaped clusters with variable density, in over 95% background noise with non-uniform background distribution, and heteroscedastic measurement errors?

**B.2** What constitutes a meaningful cluster? Or, how to automatically find the number of groups contained in a data set while using an interpretable measure of "clusteredness".

## 1.3. List of contributions

The main contribution of this thesis is a set of two analysis tools that facilitate the thorough study of stellar clusters in positional and kinematic data sets. In particular, it aims to provide interpretable methods to combat the culture of blind acceptance of a machine learning result. The two methods, which follow the two main modes of stellar cluster analysis, are the following:

[5] The distribution of stars in the positional and kinematic feature space, alongside their distribution in the Hertzsprung Russell diagram (HRD) provides evidence for or against a "true" (coeval) star cluster hypothesis. The HRD shows the evolutionary distribution of stars. It is a scatter plot in which the absolute magnitude of stars, a measure of their brightness, is plotted against the color, a measure of surface temperature of the same stars. The position of a star on the HRD depends on several factors, but notably, on its mass, chemical composition, and age. During its life, a star follows an evolutionary path through the HRD. Stars in stellar clusters are "born" together, originating from large collapsing molecular clouds and thus have the same age and chemical composition. Therefore, star cluster members with different masses are found to lie on and around (due to errors in the measurement process and the variability of – especially young – stars) a curve in the 2D plane.

A) Uncover, an interactive, visual novelty detection framework to identify unseen members of given stellar groups. Suitable models are selected in three steps: First, astronomers define *a priori* knowledge about the star cluster. Second, they use their domain expertise to quantify the goodness of the model. Third, the qualitative assessment by the users creates the opportunity to update prior knowledge and select appropriate models accordingly. Uncover is explained in detail in Ch. 3, in which this thesis aims to answer the research questions **RQ A.1**, **A.2**, **A.3**, and **A.4**. The contributions are summarized in the following points:

   i) This thesis presents Uncover, an innovative model selection method for highly flexible novelty detection models aimed at extended star cluster membership analyzes (see Sect. 3.4).

   ii) This thesis validates membership identification capabilities in two case studies. The application to the recently discovered Meingast 1 stream [82] unveils about 2000 new high-fidelity members, increasing the population size tenfold (see Sect. 3.5). The application to the well-studied $\rho$ Oph region finds 191 new high-fidelity members, demonstrating the effectiveness of Uncover (see Sect. 3.18).

   iii) Using the newly identified Meingast 1 members, this thesis corrects the original age estimate, which was from 1 Gyr to $\sim$ 110 Myr, and determines it's mass to around 2000 $M_{\odot}$, making it by far the most massive stream in the solar neighborhood. In addition, this work can assign several white dwarfs to the Meingast 1 stream (see Sect. 3.5).

   iv) Using the newly identified $\rho$ Oph members, this thesis reveals two main populations that show slightly different ages (see Sect. 3.19).

   v) This thesis embeds Uncover into a visually assisted workflow for cases of vague prior knowledge on the number and distribution of yet unseen member stars and in the presence of training set contamination and high outlier fractions (see Sect., 3.31.2).

   vi) This thesis introduces an analysis and abstraction of data, tasks, and requirements for the star formation domain (see Sect. 3.33).

   vii) This thesis provides a breakdown of the star classification process into small, interpretable steps. This workflow supports users to apply their domain expertise to assess the goodness of trained models, effectively building confidence in the final classifier among domain experts (see Sect. 3.34).

   viii) This thesis validates the visual interface Uncover in two scientific use cases that demonstrate the efficiency and effectiveness of the Uncover interface in finding new stars (see Sect. 3.38).

B) Significant Mode Analysis (SigMA), a density-based clustering method that aims to find modes in the data separated by density dips. The method studies the topological properties of the density field in the multidimensional phase space. The set of critical points in the density field gives rise to the cluster tree, a hierarchical structure in which leaves correspond to modes of the density function. Typically, however, non-parametric density estimation methods lead to an over-clustering of the input data. We propose an interpretable cluster tree pruning strategy by determining minimum-energy paths between pairs of neighboring modes directly in the input space. We tested for deviations from unimodality along these paths, which provides a measure of significance for each pair of clusters. `SigMA` is explained in detail in Ch. 4, in which this thesis aims to answer the research questions **RQ B.1** and **B.2**.

i) This thesis presents `SigMA`, a novel clustering method that takes density peaks, separated by dips, as significant clusters. Using a graph-based approach, it detects peaks and dips directly in the multi-dimensional phase space, providing a measure of significance. The method is able to adapt to non-convex shapes, variable densities, properly incorporates astrometric uncertainties, and is fine-tuned to large-scale surveys in astrophysics (see Sect. 4.4.2)

ii) This thesis identifies about $10^4$ members in the Scorpius–Centaurus association (Sco-Cen) arranged in 48 clusters of co-spatial and co-moving young stars. The HRD for each cluster shows a narrow and well-defined sequence providing a validation test to the ability of `SigMA` to extract coeval and co-moving populations (see Sect. 4.5).

iii) This thesis finds a large fraction of clusters are (tentatively) associated with well-known Sco-Cen massive stars, too bright to be in Gaia EDR3. Because the proposed method is not aware of these massive stars, the association with clusters also constitutes a validation test to `SigMA` (see Sect. 4.5).

iv) When comparing the 48 identified stellar populations in Sco-Cen to previous results from the literature, this thesis finds mostly agreement; however, several discrepancies exist. Visual selection methods used recently on Gaia data of Sco-Cen produce a 15% larger number of candidates when compared to unsupervised methods. On the other hand, the proposed methods are able to find more spatial and kinematical substructure for the same data set, and produce samples with lower contamination levels (see Sect. 4.5.2).

### 1.3.1. Thesis structure

The remaining thesis is loosely structured into two parts following the two main modes of stellar cluster analysis. In sections associated with part A, it discusses methods to uncover *unseen star cluster members.* In sections associated with part B, it covers the identification of *unknown star clusters.* These parts coincide with two main contributions of this thesis, discussed in Sect. 1.3.

In Ch. 2, the state of research is briefly surveyed. Specifically, Sect. 2.1 (A) discusses model selection techniques in one-class situations and Sect. 2.2 (A) highlights related work regarding visual and interactive model selection methods. In Sect. 2.3 (B), density based clustering techniques are surveyed and its challenges are discussed.

Following the state-of-the-art discussion, the main research contributions of this thesis are presented in Ch. 3 (A) and Ch. 4 (B). These chapters include a brief introduction followed by the respective research works, which are presented without modifications to formatting and prefaced by an evaluation of contributions by individual co-authors. Specifically, Ch. 3 introduces the `Uncover` analysis method and its application and validation on astronomical data. Furthermore, a visually assisted workflow for `Uncover` is introduced that incorporates prior knowledge of the number and location of unseen star cluster members. This work proposes strategies for updating (even vaguely formulated) prior beliefs, which, in turn, effectively provide means for model selection. This work has produced the following papers:

> **Ratzenböck, S.**, Meingast, S., Alves, J., Möller, T., & Bomze, I. 2020. "Extended stellar systems in the solar neighborhood - IV. Meingast 1: the most massive stellar stream in the solar neighborhood." *Astronomy & Astrophysics. Supplement Series*, 639, A64.
>
> Grasser, N., **Ratzenböck, S.**, Alves, J., Großschedl, J., Meingast, S., Zucker, C., Hacar, A., Lada, C., Goodman, A., Lombardi, M., Forbes, J. C., Bomze, I. M., & Möller, T. 2021. "The $\rho$ Ophiuchi region revisited with Gaia EDR3 - Two young populations, new members, and old impostors." *Astronomy & Astrophysics. Supplement Series*, 652, A2.
>
> **Ratzenböck, S.**, Obermüller, V. Möller, T. Alves, J. & Bomze, I. 2022. "Uncover: Toward Interpretable Models for Detecting New Star Cluster Members" in *IEEE Transactions on Visualization and Computer Graphics*, doi: 10.1109/TVCG.2022.3172560. *Accepted in April 2022.*

Chapter 4 presents `SigMA`, a hierarchical, density-based clustering method with measures of significance. This clustering technique is presented in the following manuscript:

> **Ratzenböck, S.**, Möller, T., Großschedl, J., Alves, J., Bomze, I. & Meingast, S. 2022. "Significance Mode Analysis (`SigMA`) for hierarchical structures: An application to the Sco-Cen OB association"
> *Major revision decision with Astronomy & Astrophysics in May 2022.*

This thesis concludes with Ch. 5 where a summary of results is presented in Sect. 5.1 and future work is discussed in Sect. 5.2.

# 2. Methodology and state of research

Here, the state of research for each research question is briefly discussed, providing a framework and guideline for this work's contributions. With the aim of interpretable methods for star cluster searches, the research questions have interdisciplinary fields of investigation. On the one hand, the methodological side of this thesis aims to contribute to the field of data science, or more precisely, data visualization and machine learning. On the other hand, it seeks first to validate this thesis' findings in an astronomical context to prove its viability and usefulness.

Related work this thesis builds upon that directly affects the techniques developed in this study are briefly discussed. To address **RQ A.1** and **A.2**, one-class model selection approaches are discussed in the context of stellar clusters in Sect. 2.1. Building upon lessons learned from Ratzenböck et al. [94] and Grasser et al. [51], this work considers model selection techniques in cases where qualitative, visual inspection of inferred stars can guarantee maximal confidence in the model, addressing **RQ A.3** and **A.4** in Sect. 2.2. Finally, in Sect. 2.3, it addresses **RQ B.1** and **B.2** by discussing related work in non-parametric, density based clustering.

Following the discussion on state-of-the-art methods in the domain of astronomy in Sect. 1.2, this chapter will contextualize this work with modern data science practices and achievements.

## 2.1. One-class model selection (A)

Due to the lack of labeled outlier data, traditional model selection techniques such as cross-validation cannot be applied to one-class methods. Since no second class can restrict model growth, models that encompass the whole feature space would achieve a perfect test score. The optimal hyper-parameter selection for one class models remains an open problem to this day [110].

To mitigate the non-trivial selection process of OCSVM hyper-parameters, automatic hyper-parameter selection approaches have been proposed, which should provide suitable results. Automatic strategies either provide selection heuristics, or focus on producing a set of pseudo-outliers [110, 109, 29, 7, 32, 113]. These artificial outliers are subsequently used as an opposing class to the training data during cross-validation. Heuristics are often limited to specific kernel parametrizations. As radial basis function (RBF) kernels bring a high degree of model flexibility most heuristics usually focus on them [36, 65, 112, 116, 48].

Both automatic approaches, however, often assume a problem in which the target class is sufficiently represented while the other class has almost no measurements in comparison [110]. This class imbalance assumption towards the training set is in stark contrast to stellar clustering where the target class is a minority embedded in, and outnumbered by, a background of non-member stars. Furthermore, automatic methods usually provide point estimates for hyper-parameters, providing only a single model to infer new member stars with.

Even in the case of optimal model hyper-parameters, one-class algorithms are shown to exhibit poor performance [109] which can be combated by using also non-optimal learners in an ensemble approach, improving the performance and robustness of the prediction. Additionally, point estimates are not able to adapt to specific user expectations. Moreover, point estimates can also be troubling in the case of noisy training data. Since residual contamination in the training sample from non-member stars is expected, one has to consider that OCSVM classifiers can be sensitive to contamination from outlier data [55, 78]. In this case, the OCSVM classifiers tend to skew toward the anomalies. Amer et al. [2] propose to mitigate the influence of outliers by altering the OCSVM objective function introducing training sample weights. Instead of tweaking the objective function, Ghafoori et al. [47] introduce a pre-processing step which removes anomalies from the training set and simultaneously tries to estimate suitable hyper-parameters. Both approaches, however, need some form of outlier estimate, be it either through the distance to the data centroid [2] or via a k-NN density estimate [47] implying that outliers occur towards the border, or in low density regions of the training set. While this assumption is sufficient for many applications, it does necessarily generalize to star clusters where contamination depends greatly on the training set selection method.

This work will focus on automatic model selection heuristics that are adopted to the domain of star cluster extractions.

## 2.2. Visual model selection (A)

Although no ground truth information is available for individual stars, ensembles of stars can be qualitatively validated by domain experts. The distribution of stars in phase space, alongside their distribution in the HRD[1] provides evidence for or against a “true” star cluster hypothesis.

Since solutions need qualitative verification, the process of finding appropriate and effective models is inherently unsupervised. The difficult problem of hyper-parameter-finding of unsupervised algorithms has been addressed by the visualization community and is known as the paradigm of visual parameter space analysis (VPSA [102]). VPSA aims to replace the tedious manual process with a systematic approach that facilitates the comprehensive visual exploration of the solution space.

A large body of previous work exists on interactive tools to support the exploration of possible models; a summary is listed in the following paragraphs. General purpose tools provide means for exploratory data and cluster analysis. The Hierarchical Clustering Explorer (HCE [103]) is an early example of an interactive visualization tool that

[1] See footnote 3 in Sect. 1.2.2 for more details.

improves the users' understanding of different clusters. HCE organizes the hierarchical cluster structure as a dendrogram with heatmaps. DICON [14] introduced techniques for comparing clustering results across different algorithms and even data sets. To facilitate cluster analysis DICON uses an icon-based cluster visualization that embeds statistical information into a multi-attribute display. Clustrophile 1+2 [27, 18] is a cluster analysis and exploration tool which guides a user through different choices of clustering hyper-parameters and provides interpretable cluster explanations. Clustervision [74], similar to Clustrophile 1+2, is a general purpose clustering tool which performs a meta clustering analysis using multiple different clustering techniques and hyper-parameters. Users can explore these solutions on the basis of five summary statistics measuring cluster compactness and separation with the goal of providing a domain independent comparison of clustering results.

Extensive work has been done on incorporating user feedback into the clustering process. VISTA [23] was developed with the understanding that human interaction is an important factor in clustering. It is designed with the concept in mind, that clustering is not finished without human interaction. Thus, users can interactively refine clusters while improving their understanding of the result and acceptance of the pattern applied. ClusterSculptor [85] enables users to intervene in the clustering processes. Users can iteratively re-organize and interact with clusters using expert knowledge. The system aims to derive clustering rules from these examples. Schreck et al. [100] use user feedback to influence the result of self-organizing map (SOM) clusterings of trajectory data. Matchmaker [76] extends ideas from HCE [103] allowing users to modify clusterings by grouping data dimensions. Open-Box Spectral Clustering [101] is an interactive tool that visualizes mathematical quantities involved in 3D spectral clustering. The system provides hyper-parameter value suggestions and immediately reacts to user feedback to increase the quality of image segmentation. Packer et al. [88] present a distance-based spatial clustering approach and provide a heuristics computation of input hyper-parameters that supports the search for meaningful cluster results. ReVision [117] allows users to steer hierarchical clustering results by utilizing both public knowledge and private knowledge from users. By reformulating this knowledge into constraints, the data items are hierarchically clustered using an evolutionary Bayesian rose tree.

Conceptually similar research to ours include Geono-Cluster [26] and PK-clustering [93]. Geono-Cluster enables biologists to insert their domain expertise into clustering results. The tool displays the expected clustering results to users based on a small subset of data. The system estimates users' intentions and generates potential clustering results. PK-clustering enables users to input prior knowledge and explore the space of clustering results in the context of the provided prior knowledge. The study of consensus between prior assumptions and cluster results allows users to acquire and update their prior knowledge.

In contrast to previous works, this thesis aims to shift the focus from data exploration and insight generation towards effective model generation targeted at a single cluster. Additionally, it finds that currently available systems fail to incorporate previously identified members and no work has been done on visual parameter space exploration on novelty detection methods.

## 2.3. Non-parametric, density based clustering (B)

This thesis considers model agnostic approaches to identify stellar groups in positional and kinematic data sets. Specifically, it aims to study the structure of the density distribution, considering its modes as stellar groups. Related research on non-parametric density-based clustering methods is vast. In the following, a brief introduction is provided and critical literature is discussed along the way. In the following, bold, lower-case variables denote $d$-dimensional vectors.

### 2.3.1. Cluster definition in density-based analyses

The clustering analysis builds on the assumption that observed data $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with $\mathbf{x}_i \in R^d$ are drawn from an unknown density function $f$. The goal is to understand the structure of the underlying density function, which is estimated from data. Wishart [115] provided an early interpretation, which defines clusters as data samples associated with modes in $f$. Koontz et al. [67] proposed a first mode-seeking clustering algorithm, and similarly, later works such as the widely used Mean-Shift algorithm and its variants [24, 25, 111] build upon Wishart's idea.

Hartigan [54] proposed an alternative cluster definition in which clusters are the connected components of the level-sets of $f$. Assuming $f$ has compact support $\mathcal{X}$, the resulting level-sets for the threshold $\lambda$ can be formally written as the following:

$$L(\lambda) := \{\mathbf{x} \in \mathcal{X} \ : \ f(\mathbf{x}) \geq \lambda\} \tag{2.1}$$

Thus, $L(\lambda)$ constitutes a set of connected components that are identified as clusters. The connected components of the level-set $L(c)$ are the resulting clusters, while the remaining data is treated as noise. See the second to last panels in Fig. 4.1 for four clustering solutions, depending on various density threshold levels.

A single threshold, as is, e.g., employed in the DBSCAN [35] algorithm, cannot reveal all peaks for many data sets containing clusters with variable densities. Instead, a hierarchy of clustering solutions emerges by considering all possible threshold values in a hierarchical clustering approach.

#### Hierarchical, density-based clustering

The set of critical points, where $\nabla f = \mathbf{0}$, in the density field gives rise to the cluster tree, a hierarchical structure in which leaves correspond to modes of the density function. Conceptually, the cluster tree is obtained by sweeping the density threshold $\lambda$ from $\infty$ to $-\infty$ and tracking connected components at each step; see Fig. 4.1 for a schematic illustration.

Similar to gradient estimation in mode-seeking algorithms, the computational realization of cluster tree extractions faces several implementational challenges. Estimating the connected components, while easy in one dimension, gets nontrivial in higher dimensions. To approximate connected regions in higher dimensions, implementations and theoretical analyses [5, 107, 19, 71, 20] adopt a graph $G(\lambda)$ over the data samples where vertices and/or edges are filtered according to $\lambda$, thus $\{\mathbf{x} \in X \ : \ \hat{f}(\mathbf{x}) \geq \lambda\}$.

However, Stuetzle & Nugent [107] point out that samples from the same connected component in $L(\lambda)$ may end up in different connected components of $G(\lambda)$. Additionally, density estimates are inherently noisy; both effects, thus, lead to an artificial increase in the number of clusters. To counteract this over-clustering, the resulting graph cluster tree is usually pruned in a post processing step during which spurious clusters are identified and merged back into the "parent cluster" [107, 71, 20].

### Cluster tree pruning

Various methods for cluster tree pruning have been suggested. The popular HDBSCAN [10] algorithm, for example, prunes the cluster tree in two ways; first, clusters that have fewer members than a given threshold value are merged back to the parent mode. Second, HDBSCAN estimates the stability of each cluster in the hierarchy via the concept of relative excess of mass (EOM) [84]. The EOM heuristic measures the lifetime and size of a cluster and favors more prominent and stable clusters that live longer in the cluster tree.

A related pruning heuristic comes from considering each mode's topological stability, or persistence, in $\hat{f}$, introduced by Chazal et al. [21]. Persistence is defined as the lifespan of each connected component, i.e., the difference in density from a mode's birth to its death by merging into its parent mode. The concept of persistence is an effective measure to prune the cluster tree, as it is stable under small perturbations to the underlying density $f$ [33, 120, 49, 21].

Similarly to persistence, Ding et al. [30] present the *saliency* index, a mode's birth to death density ratio. The cluster tree is generated by varying the saliency index between 0 and 1. Cluster configurations that are unchanged for the longest time as the saliency is varied are considered relevant results.

Both heuristics, persistence and salience, have the desirable ability to automatically provide sensible hyper-parameter choices, i.e., largest value ranges where the clustering remains unchanged. However, in the case of many clusters and large data sets, these stable regions typically disappear and selecting the input parameters again warrants a proper parameter search.

Compared to the heuristic notions of stability, there is also growing research to apply statistical methods that test the modality structure of the data. These methods offer the advantage of an interpretable and meaningful parameter $\alpha$, defining the significance level of a corresponding hypothesis test. The null hypothesis $\mathrm{H}_0$ commonly assumes that the data, or subsets of it, are sampled from a uni-modal density, whereas the alternative hypothesis $\mathrm{H}_1$ suggests multi-modality. The null hypothesis is rejected at a significance level $\alpha$ if the p-value from the corresponding test procedure exceeds this significance level.

Hypothesis test procedures have been used to estimate the number of clusters in $k$-means and EM frameworks. G-means [52], PG-means [37], and Dip-means [60] employ the Anderson-Darling [3], Kolmogorov-Smirnov [66, 106], and Hartigan's dip [53] test to estimate the number of clusters, respectively. Burman & Polonik [9] proposed a conceptually similar approach, which examines the modality structure on a straight line path between two candidate modes. Two neighboring peaks are true clusters in the data if there exists no path between them that does not undergo a substantial dip in density.

This work proposes a meaningful cluster tree pruning method by employing a modified version of the modality test by Burman & Polonik [9]. The integration into the hierarchical, density based framework provides a highly accurate clustering algorithm with interpretable hyper-parameters that alleviate practitioners from haphazardly trusting machine learning output.

# 3. Extended membership analysis (A)

“How can members of previously studied star clusters guide the search for yet undetected member stars?” And: “How to decompose the membership identification process into interpretable steps such that practitioners can confidently build powerful novelty detection classifiers themselves?” This chapter answers these questions in three steps.

In the first publication (A1), we[1] developed a heuristic analysis framework we call `Uncover`, which incorporates users’ prior knowledge of yet unseen members to select suitable novelty detection classifiers. We express prior knowledge as ranges of interpretable summary statistics that trained classifiers have to adhere to, instead of directly tuning model hyper-parameters. The method is validated by applying it to the Meingast 1 stellar stream [82], increasing its size roughly tenfold. By considering the selection criteria that led to the stream’s discovery, we can place strict constraints on given summary statistics.

In paper two (A2), we discuss methodological updates in the case of vague and uncertain prior knowledge in the application of `Uncover` to the $\rho$ Oph region. In contrast to Meingast 1, $\rho$ Oph has been studied extensively. Its age and proximity to earth make it a vital star formation probe. The region’s prominence and treatment in the past provide an edge case to the `Uncover` analysis pipeline. Compared to Meingast 1, estimating the number and distribution of yet unseen members is not straightforward. Due to the vague nature of prior knowledge, model selection becomes practically unfeasible as the space of possible solutions cannot be constrained effectively. We propose to sample the space of prior assumption tuples and provide an updated automatic model selection approach which reduces the contamination fraction of inferred stars.

Summary statistics, as demonstrated in the first publication A1, provide an efficient and precise model selection approach if sufficient prior knowledge of undetected group members is available. However, in reality, this knowledge is often vague and abstract. In such cases, we have demonstrated that we can constrain possible models by limiting the contamination fraction, which is determined using radial velocities. A downside of this methodology is its dependence on radial velocity measurements. Less than one percent of data instances across the Gaia DR2 (and EDR3) catalog have radial velocities measurements. Thus, a small and typically uncertain subset of sources determines the goodness of trained models, leading to many false rejections. We aim to find alternative or complementary model selection tools in the case of vague prior knowledge, which considers a more holistic view of the solution space. To do so we have developed a visually supported five-step workflow approach in which we (1) provide a comprehensive overview

[1] Throughout this work, “we” refers to the author team of respective articles. The individual contributions of all co-authors are outlined in the corresponding publication sections, see Sect. 3.1, Sect. 3.14, Sect. 3.29, and Sect. 4.1.

of the vast solution space, (2) give users the opportunity to validate groups of similar models, (3) derive rules for “good” models from users’ judgements at the previous step, (4) facilitate What-If analyzes where users can study the effect of individual rules on the final set of inferred stars. We show in two case studies on $\rho$ Oph and Corona-Australis, see Sect. 3.38.1, and a usability study, see Sect. 3.38.2, that users are effectively, and efficiently building models themselves.

## 3.1. Extended stellar systems in the solar neighborhood - IV. Meingast 1: the most massive stellar stream in the solar neighborhood.

### Full publication details

**Ratzenböck, S.**, Meingast, S., Alves, J., Möller, T., & Bomze, I. 2020. “Extended stellar systems in the solar neighborhood - IV. Meingast 1: the most massive stellar stream in the solar neighborhood.” *Astronomy & Astrophysics. Supplement Series*, 639, A64.

Credit: Ratzenböck et al., A&A, 639, A64, 2020, reproduced with permission © ESO.

### Author contributions

The paper is co-authored by me, Stefan Meingast, João Alves, Torsten Möller, and Immanuel M. Bomze. As the leading author, I conceived and developed `Uncover`, performed the experiments, and wrote the main parts of the paper. The analysis was done closely with Stefan Meingast, the leading author of the Meingast 1 stream discovery paper. Stefan Meingast and João Alves also contributed to the writing and astronomical interpretation of results and discussion. João Alves and Torsten Möller supervised the project and offered suggestions along the way. Immanuel M. Bomze helped revise the final version.

### Information on the Status

Submission date: 27 January 2020
Accepted: 11 April 2020
Published: 07 July 2020

A&A 639, A64 (2020)
https://doi.org/10.1051/0004-6361/202037591
© ESO 2020

Astronomy & Astrophysics

# Extended stellar systems in the solar neighborhood

## IV. Meingast 1: the most massive stellar stream in the solar neighborhood[⋆,⋆⋆]

Sebastian Ratzenböck[1], Stefan Meingast[2], João Alves[1,2,3], Torsten Möller[1,4], and Immanuel Bomze[1,5]

[1] Data Science at University of Vienna, Währinger Straße 29, 1090 Vienna, Austria
e-mail: sebastian.ratzenboeck@univie.ac.at
[2] University of Vienna, Department of Astrophysics, Türkenschanzstrasse 17, 1180 Wien, Austria
[3] Radcliffe Institute for Advanced Study, Harvard University, 10 Garden Street, Cambridge, MA 02138, USA
[4] University of Vienna, Faculty of Computer Science, Währinger Straße 29/S6, 1090 Vienna, Austria
[5] University of Vienna, ISOR/VCOR, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

Received 27 January 2020 / Accepted 11 April 2020

**ABSTRACT**

*Context.* Nearby stellar streams carry unique information on the dynamical evolution and disruption of stellar systems in the Galaxy, the mass distribution in the disk, and they provide unique targets for planet formation and evolution studies. Recently, Meingast 1, a 120° stellar stream with a length of at least 400 pc, was dicovered.

*Aims.* We aim to revisit the Meingast 1 stream to search for new members within its currently known 400 pc extent, using *Gaia* DR2 data and an innovative machine learning approach.

*Methods.* We used a bagging classifier of one-class support vector machines with *Gaia* DR2 data to perform a 5D search (positions and proper motions) for new stream members. The ensemble was created by randomly sampling 2.4 million hyper-parameter realizations admitting classifiers that fulfill a set of prior assumptions. We used the variable prediction frequency resulting from the multitude of classifiers to estimate a stream membership criterion, which we used to select high-fidelity sources. We used the HR diagram and the Cartesian velocity distribution as test and validation tools.

*Results.* We find about 2000 stream members with high fidelity, or about an order of magnitude more than previously known, unveiling the stream's population across the entire stellar mass spectrum, from B stars to M stars, including white dwarfs. We find that, apart from being slightly more metal poor, the HRD of the stream is indistinguishable from that of the Pleiades cluster. For the mass range at which we are mostly complete, $\sim 0.2\,M_\odot < M <\sim 4\,M_\odot$, we find a normal IMF, allowing us to estimate the total mass of stream to be about $2000\,M_\odot$, making this relatively young stream by far the most massive one known. In addition, we identify several white dwarfs as potential stream members.

*Conclusions.* The nearby Meingast 1 stream, due to its richness, age, and distance, is a new fundamental laboratory for star and planet formation and evolution studies for the poorly studied and gravitationally unbound star formation mode. We also demonstrate that one-class support vector machines can be effectively used to unveil the full stellar populations of nearby stellar systems with *Gaia* data.

**Key words.** methods: statistical – open clusters and associations: individual: Meingast 1 – stars: luminosity function, mass function – stars: massive – stars: low-mass – white dwarfs

## 1. Introduction

Coherently moving groups of stars in the Milky Way are unique laboratories where we can coherently study a large variety of astrophysical processes. For instance, the similar birth conditions in nearby moving groups have provided much insight into individual stellar properties (e.g., Torres et al. 2008; Gagné et al. 2014; Riedel et al. 2017, and references therein). Moreover, while older stellar systems experience mass loss due to the gravitational interaction with the Galaxy's gravitational potential (e.g., Meingast & Alves 2019; Röser et al. 2019), young co-moving groups can give us important clues on the governing star formation processes in the Milky Way.

Recently, Meingast et al. (2019), the second installment in this series (hereinafter referred to as Paper II), discovered a 120° stellar stream that is currently traversing the immediate solar neighborhood at a distance of only ∼100 pc. For this paper, the authors determined the age of the system to be 1 Gyr. Their assumption was mostly based on the presence of a single star in their selection, namely the subgiant 42 Ceti. Shortly after the stream's discovery, Curtis et al. (2019) determined stellar rotation periods of stream members to be very similar to stars in the Pleiades. Their application of gyrochronolgy thus sets the age of

[⋆] The full source catalog described in Table G.1 is only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/639/A64

[⋆⋆] In our original discovery paper, we did not name the stream. The authors of the first follow-up paper (Curtis et al. 2019) contacted us regarding a name for the structure but did not agree with our proposed name and decided on their own to name the system the Pisces-Eridanus stream. Their chosen name, however, not only does not capture the true size of the stream (the stream stretches across at least 10 constellations and likely extends beyond these), it is ambiguous as it can lead to confusion with the Pisces moving group (Binks et al. 2018). In general, given the number of new streams being found by *Gaia* and the finite number of constellations, it seems appropriate to move away from using constellations to name streams (e.g., Ibata et al. 2019). An unambiguous remedy to this particular situation is to name the stream after the original discoverer, which we do in this paper, naming the structure Meingast 1.

the stream at close to 120 Myr, implying that the star 42 Ceti is likely an unfortunate interloper.

The search criteria in Paper II were based on the 3D space velocities in a cylindrical coordinate frame derived from astrometric measurements provided with the second *Gaia* data release (*Gaia* DR2; Gaia Collaboration 2016, 2018c). While space velocities provide a robust estimate on membership, evaluating 3D motions of stars requires radial velocity measurements. This requirement substantially limits the identification of members to a small subset of *Gaia* DR2, specifically to stars with $G \lesssim 13$ mag, which in the case of Meingast 1 translate to stellar masses between ~0.5 and 1.5 $M_\odot$.

The goal of this paper is to unveil the stellar population of the Meingast 1 stream, from B stars down to mid-M stars, or the completeness limit of the *Gaia* DR2 data. To this end, we applied state-of-the-art machine learning tools, where we used the previously identified members as a training set. The structure of this paper is as follows: in Sect. 2, we present the data used for the analysis. Section 3 summarizes the method used to select potential stream member sources from the *Gaia* DR2 data set. Finally, in Sect. 4, we present a final high-fidelity source catalog on which we determine the age and mass of the Meingast 1 stream[1].

## 2. Data

For the analysis, we used the 5D position ($\alpha$, $\delta$, $\varpi$) and velocity ($\mu_\alpha$, $\mu_\delta$) information, provided by *Gaia* DR2. Following the data selection in Paper II, we preferred distance estimates provided by Bailer-Jones et al. (2018). The distance limit of the stellar sample is kept at ≤300 pc in accordance with Paper II. This is motivated by the choice of our classifier, which predicts member stars within the limits of the previously determined extent of the stream. Furthermore, the subsequently described method works independently from quality criteria. Therefore, quality filters are only applied for visualisation purposes. This selection results in a data set of 18 692 951 total stars.

For Paper II, the sources were extracted in a 6D parameter space spanned by three spatial ($X$, $Y$, $Z$) and three velocity dimensions ($v_r, v_\phi, v_z$). Specifically, the velocities were represented in a galactocentric cylindrical coordinate system to better represent the bulk motion stars. Consequently, the source identification in Paper II depended on radial velocity measurements, which are scarce in *Gaia* DR2. Within the search region of 300 pc, about 95% of all sources in the catalog were, therefore, not taken into account in Paper II due to missing radial velocity data.

## 3. Member selection

As mentioned above, the bulk of *Gaia* DR2 catalog sources were not used in the original member identification of the stream in Paper II. Omitting the radial velocity component yields a much more complete source list, but at the same time limits any analysis to projected tangential velocities given by the proper motion measurements. While members of spatially confined star clusters can be identified reliably in proper motion space, the recently discovered stream encompasses at least 120° on sky. This large extent introduces significant projection effects in tangential velocities, posing a nontrivial problem for member identification in 5D.

### 3.1. Supervised member selection

To avoid the difficult task of clustering in the 5D position and proper motion space, we pursued a supervised approach based on one-class support vector machines (OCSVM; Schölkopf et al. 2001). Instead of finding a decision boundary between distinct groups in the training sample like a typical SVM (Cortes & Vapnik 1995), an OCSVM constructs a decision surface that attains a maximum separation between the training samples and the origin. Consequently, the algorithm infers the properties of the input samples by enclosing the support of its joint distribution with a hyper surface during the training process. Depending on the position of unseen data points[2] to this surface, a trained predictor acts as a binary function which groups new example points as either resembling the previously seen training data or not. We aim to estimate the extent of the stellar stream by using the OCSVM algorithm and the already classified sources from Paper II as a training set. Subsequently, we predict the membership of unseen stars to the stream within a 300 pc sphere around the Sun (see Sect. 2). In order to find a model that is capable of providing a physically meaningful characterization of the stellar stream in the 5D feature space, the corresponding hyper-parameters of the OCSVM classifier have to be set sensibly.

### 3.2. Parameter tuning

We made use of the libsvm (Chang & Lin 2011) OCSVM implementation, which features two main hyper-parameters for the RBF-kernel[3], $\gamma$ and $\nu$. The parameter $\gamma$ defines a region of influence of the support vectors selected by the model. The variable $\nu$ controls the fraction of possible outliers as well as the fraction of support vectors. Thus, $\gamma$ and $\nu$ are crucial hyper-parameters that define the shape of the enveloping hull.

Additionally, these parameters, and subsequently the classifier shape, depend on the input variable range. Since the parameter $\gamma$ describes a support vector region of influence, different feature ranges lead to a varying model flexibility within each input variable. To mitigate an asymmetric feature weighting, a common approach is to standardize each input variable to a common variance by dividing each feature by its standard deviation. However, as we are dealing with a combined feature space of position and proper motion information a certain weighting towards one of the two feature spaces might be beneficial to properly characterize the joint probability of stream members. Consequently, after scaling the features to unit variance, we added an additional hyper-parameter: $c_x/c_v$. This parameter describes the scaling fraction between positional and proper motion features. When $c_x/c_v = 1$ the variance in both feature spaces is the same. In practice, we set $c_v = 1$ and vary $c_x$ within a certain range.

As we chose a classifier via a set of hyper-parameters, we have to be aware of existing contamination in the training set (estimated to amount to a few percent in Paper II). Additional selection biases caused by the original clustering and parameter choice that influence the final obtained stream selection should be considered. Therefore, only crude estimates about the true joint distribution of the sources in 5D are possible. Nevertheless,

[1] We acknowledge the simultaneous publication by Röser & Schilbach (2020), who have also studied member stars of the Meingast 1 stream.

[2] Stars in the data set are represented as points in a 5D space with three position axes and two proper motion axes constituting the so-called feature space. Thus, in a machine learning context, we refer to stars in the data set as points in a feature space.

[3] We conclude from extensive hyper-parameter searches that the RBF kernel always outperformed the alternative options. Hence we omit the description of other kernel types in this section.

we have information about the resulting classifier shape, which limits the space of possible solutions. Firstly, based on the number of missing radial velocity measurements, we estimate that the total number of member stars should roughly increase twenty-fold. Secondly, due to a lack of a better description we estimate that the true extent is comparable to the original selection in Paper II, which found that the stream is roughly prolate spheroidal with a length of about 400 pc and an equatorial diameter of about 50 pc.

A trained classifier has to be able to capture these prior assumptions. Therefore, we used the above mentioned characteristics to eliminate predictions that seem unfit to describe the stellar stream in 5D. Since we cannot infer the true joint distribution from the available stream members, and our prior assumptions entail some allowable margin of variation, the model parameters cannot be tuned to optimal values. Instead, we aggregated the predictions of multiple models that conform to our prior assumptions into an ensemble of OCSVMs. This procedure is referred to as bootstrap aggregating, also known as bagging (Breiman 1996). A benefit of using multiple aggregated classifiers, in comparison to one single model, is an improvement in prediction stability. Due to its variance-reducing ability, bagging has been successfully applied, especially to noise-prone classifiers, whose predictions vary significantly with small variations in the training data. In Grandvalet (2004), the author suggests that bagging systematically reduces the influence of outlier samples in the training data. Furthermore, by bundling together multiple models, a notion of stability for each star is obtained as different regions of the 5D training space have varying prediction frequencies. Ideally, the ensemble of classifiers has a higher prediction frequency towards the center region of the stellar stream (in 5D) where sources are less likely to be randomly selected field stars. Bagging, therefore, automatically creates a hierarchy from more robust to less robust stream members, which reduces prediction variance compared to a single classifier.

A schematic illustration of a small ensemble classifier is shown in Fig. 1. The black scatter points represent the training set, whereas the colored shapes depict the bounding surfaces of individual OCSVM classifiers trained with different sets of hyper-parameters. The unification of multiple classifiers results in an ensemble classifier where overlapping bounding regions result in different levels of prediction frequency.

The final bagging predictor is obtained in a two step process: Firstly, the actual training phase and, secondly, the validation phase, which rejects models that do not represent our expectations well. In the learning phase (see Appendix A for more details) the model is trained using ten-fold cross validation on a random set of hyper-parameters ($\gamma_i$, $\nu_i$, $(c_x/c_v)_i$). Before deploying the classifier on the full data set, we filtered out models below a mean accuracy score of 0.5, or a standard deviation above 0.15 across the hold-out sets. Models passing this filter criterion enter the validation phase, which assess the classifiers capability of capturing our prior assumptions about the distribution and quantity of predicted sources. We require the model to comply with the following criteria. Firstly, the number of predicted stream members $N_s$ must not exceed a physically sensible range, which is limited to $N_s \in [500, 5000]$. Secondly, the extent of the predicted stream members in position and proper motion space must be similar to the original ones. Thirdly, the cylindrical velocity distribution of the stream members must not deviate too much from the training sample distribution. For a full description on the implementation of these three validation criteria, see Appendix B.

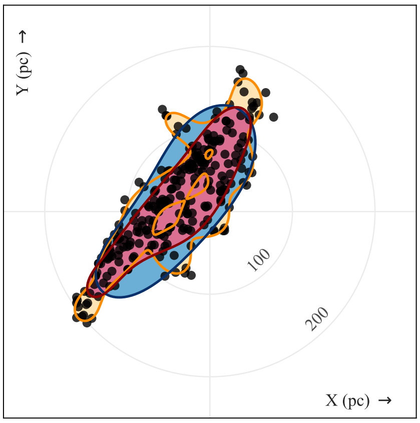


**Fig. 1.** Schematic figure illustrating the effect of different hyper-parameters on the classifier shape in the Galactic X–Y plane. Black points represent the training set, whereas the colored shapes depict the bounding surfaces of individual OCSVM classifiers trained with a different set of hyper-parameters. The unification of multiple classifiers results in an ensemble classifier where overlapping bounding regions result in different levels of stability.

Since we cannot formulate an exact objective function to be minimized, we did not converge to a single, optimal hyper-parameter selection. Instead, the models were assessed as either plausible candidates, which capture out prior assumptions about the distribution of the predicted sources, or not. Therefore, for small ensemble classifiers with only a few models, the prediction depends on the sampling strategy in hyper-parameter space. To reduce the dependency on the search strategy, we iterate through 2.4 million random realizations of ($\gamma_i$, $\nu_i$, $(c_x/c_v)_i$) within their respective range in order to converge to a stable solution. Altogether, the final classifier ensemble consists of a total of 8515 classifiers, which have passed the validation steps. Figure C.1 shows the distribution of accepted models with respect to the hyper-parameters $\nu$, $\gamma$, and $c_x/c_v$. The software used to train the ensemble classifier is publicly available[4].

### 3.3. Limitations and caveats

Any supervised model based on OCSVMs is limited by the provided training data, because the shape of the decision surface is determined by the input training set. As suggested in Paper II, the stream's extent might potentially be much larger due to sensitivity limitations. The method used in this paper is not able to infer the stream membership of stars outside the constructed decision boundary. Finding externally located stream members would require, for example, a transition to unsupervised methods, which are not limited by a fixed training set.

Additionally, the constructed decision boundary depends heavily on the outermost points in the training sample as they are more likely to act as support vectors for the decision surface. As the density of points decreases towards these outer regions (in 5D), the decision boundary depends on random fluctuations of these border points present in the training set. Furthermore, we suspect the fraction of contaminants in stream member stars

[4] https://github.com/ratzenboe/uncover and http://uncover.cs.univie.ac.at/

per unit volume increase towards border regions. Thus, outliers in the border region have an increasing chance of being a support vector defining the shape of the decision surface. These effects, however, are somewhat mitigated by the choice of bagging multiple predictors, which helps to reduce unstable decision surfaces.

While omitting the radial velocity component opens up the possibility to search for more stream members, we lose, at the same time, an additional discriminative dimension. By neglecting the radial velocity distribution of the input data, the implemented classification scheme impacts the contaminant fraction of our final source list. This leads to an increasing recall at the cost of reduced precision.

## 4. Results and discussion

Using no pre-filter selection the classifier ensemble predicts a total of 4243 stream members. This source list does not, however, contain all members from the original training set. Approximately 10% of the training data are not captured by the ensemble classifier. This reduction can be attributed to the model validation phase, where we prioritized more conservative models in an attempt to prevent overfitting. To increase this retrieval rate, we would need to omit the bootstrapping step combined with the subsequent majority voting (see Appendix A) and use the entire sample to train individual classifiers. Also, to be sensitive to more remote points, we would need to include more flexible models in the classifier ensemble. However, these tools and choices have been installed to prevent serious overfitting on the training data and to dampen the influence of outlier samples in the training data. Since an important goal is to find a robust model that minimizes the contamination fraction of the inferred points, we tolerate a slightly reduced retrieval fraction of the original training set points.

To visualize our results, we implemented a series of quality selections described in Appendix D, hereinafter referred to as filter Q1. For a direct comparison to the original training sample, we implemented the filter criteria as in Paper II (excluding the criterion on radial velocities), hereinafter referred to as filter Q2. The quality filters Q1 and Q2 reduce the total number of classified member stars to 2567 and 2913, respectively. This selection contains, however, many sources that are predicted by only a marginal fraction of the 8515 classifiers in the bagging ensemble. Each individual classifier is associated with an individual set of classified stream members. Thus, considering all 8515 classifiers, each source can be assigned a prediction frequency. We define this prediction frequency, hereinafter referred to as stability, as the fraction of classifiers in the bagging ensemble that include a certain star in their prediction set. Figure 2 shows the 5D distribution of the training sample (top row) and the stream members classified by our trained OCSVM (quality filter Q1), where the color indicates the stability of each source for our new classification. We observe that, on average, stability values tend to increase towards the central parts of the stream. Additionally, we find that when inspecting the new source set in the color-absolute magnitude diagram (see Fig. 5), sources with lower stability numbers correlate with a larger scatter, while sources with higher stability values are more compactly distributed around an idealized isochronal curve. Therefore, stability can be used as a measure to filter out potential contaminant sources.

Since the training process includes a validation step, even stars with low stability values can be regarded as potential stream members. Hence, stability constitutes not a probability estimate, but rather a quality feature for which we aim to find a suitable criterion to clean our prediction sample. To determine the reliability of the predicted stellar sample, we estimated the level of contamination at various stability filters.

We measured the contamination via the velocity dispersion in 3D, parametrized via $v_r$, $v_\phi$, and $v_z$. However, due to contributions of random contaminants, the standard error of the prediction set is largely dominated by outliers, regardless of the stability filter criterion. Hence, we describe the variability of the velocity distribution with the median absolute deviation (MAD), which is a robust estimate of statistical dispersion. For reference, the training data distribution measures an MAD in the 3D velocities of 2.1 km s$^{-1}$.

Figure 3 displays the influence of a variable stability filter criterion on the 3D velocity distribution. By moving in the plot from left to right, we gradually added less "stable" sources to the predicted data set. We identified two distinct sections in this curve that are dominated by different slopes. Firstly, the section with stabilities from 100% decreasing to 4% is comprised of a roughly constant growing scatter around the expected 3D Cartesian velocity. Secondly, adding sources with a stability below ~4% results in a rapid growth of the MAD. This sudden increase is most likely caused by adding a significant number of contaminating field stars. Here, we assumed that these contaminating field stars are more likely associated with the outer borders of the stream in the 5D parameter space, which is also where the trained classifier ensemble is less confident about the stream membership of stars. This decrease in stability values of predicted sources towards the outer regions of the stream is also well visible in Fig. 2.

In addition to the sudden increase at 4%, we identify another characteristic property of the MAD distribution in Fig. 3. Starting at about 40%, we observe an extended flat distribution up to 24%. In this range, the amount of scatter remains nearly constant. This filter criterion (`stability` ≥ 24%) yields a very stable subsample to the more lenient `stability` > 4% criterion.

The filter behavior can be observed in more detail in Fig. 4, where the successive cleaning of the prediction set is displayed in each individual velocity component. The solid lines in the figure represent a kernel density estimation of the marginal distributions for various color-coded stability filter criteria. Specifically, we sampled the distributions at constant intervals in stability with a step size of 5%. The hue change from red to shades of blue indicates the transition from a contamination-dominated to a more robust filter regime. In the marginal distributions, the disproportionately large reduction in the amount of scatter around mean velocities by applying the `stability` > 4% filter criterion becomes apparent. For subsequent filter criteria, the contamination outside the training sample distribution (black line) is reduced at a nearly constant rate, particularly in the $v_r$ and $v_\phi$ observables. Moreover, we identify a kinematic substructure in the panel displaying $v_z$ velocities. Sources identified with this substructure have systematically larger vertical velocities by about 5 km s$^{-1}$ compared to the bulk motion of the stream. These sources are only clearly separable in $v_z$ and do not show any obvious correlation in other velocities or can be segregated in spatial coordinates. We note here that this substructure accounts for the high MAD of the predicted sources and is removed only for very conservative stability filter criteria above 90%.

Following the above outlined characteristics in the velocity distributions, we therefore implemented an additional criterion of `stability` > 4% or `stability` > 24% for a more conservative approach. Depending on the quality filter selection, the stability >4% filter criterion reduces the number of predicted stream members to 1869 or 2110 for Q1 and Q2, respectively.

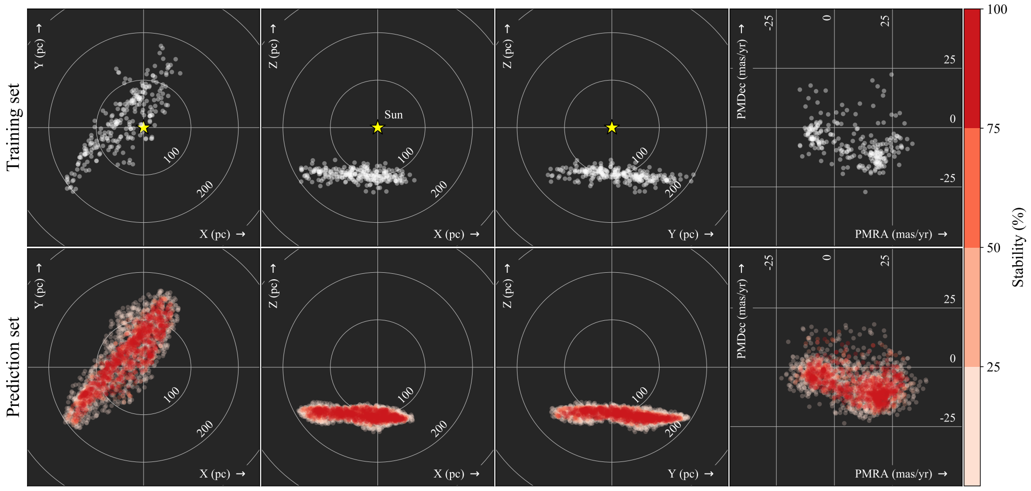


**Fig. 2.** Positional and proper motion projections of the training and prediction set are displayed in the *first and second rows*, respectively. Using a quality pre-selection (see Appendix D), we find a total of 2567 member stars (*bottom row*), compared to 256 in the training set (*top row*). The color information highlights the stability of a given star, which tends to grow towards the central regions of the stream.

In order to quantify the contamination fraction in our source catalog, we considered the fraction of outliers in the marginal 3D velocity distributions. To do this, we defined, for each velocity component, a region of inliers as the $3\sigma$ around the training sample mean. This definition constitutes a very conservative estimate, as the velocity distribution of the training data is by design very narrow. Furthermore, the kinematic substructure in the $v_z$ component naturally leads to very large contamination fractions. For this reason, we only considered the radial and azimuthal velocity components when estimating the contamination for various stability filter criteria. Figure 6 shows the outlier fraction within each velocity component. Based on our assumptions, we obtain a contamination estimate of roughly 25% and 20% for the stability criteria >4% and >24%, respectively. However, we note again that this is a very conservative estimate that assumes an intrinsic velocity dispersion of only around 1 km s$^{-1}$. By increasing the estimated velocity dispersion to 2 km s$^{-1}$ the contamination drops to roughly $10-15$%, which we suspect to be a more realistic estimate.

Since the ensemble classifier is trained on positional and proper motion data, we can apply it to any survey that provides these measurements. In an effort to increase the source list, especially toward brighter stars, we applied our ensemble classifier to the Hipparcos (van Leeuwen 2007) source catalog, see Appendix F for more details. In total, we find 21 new potential stream members in the Hipparcos catalog, 10 of which we consider to be robust. We added the 10 predicted Hipparcos sources to the HRD plot in Fig. 5. Among the prediction set, we find $\alpha$ Aquarii, the brightest star in the Aquarius constellation. Using the radial velocity information from Soubiran et al. (2008), we find a galactocentric velocity of $\boldsymbol{v} = (-3.15, 229.19, -8.73)$ km s$^{-1}$, which is well within the $3\sigma$ region of the training set. However, a comparison of parallax measurements between *Gaia* and Hipparcos reveals a large systematic discrepancy of a factor of approximately two, which makes $\alpha$ Aquarii a low-fidelity stream member.

Using gyrochronology, Curtis et al. (2019) concluded that the stream has an age comparable to the Pleiades. This contrasted with the isochronal age derived in Meingast et al. (2019), which was hinging on a single star, 42 Ceti, a subgiant. With the new and larger member list, we can now attempt to make a more precise estimate regarding the stream's age.

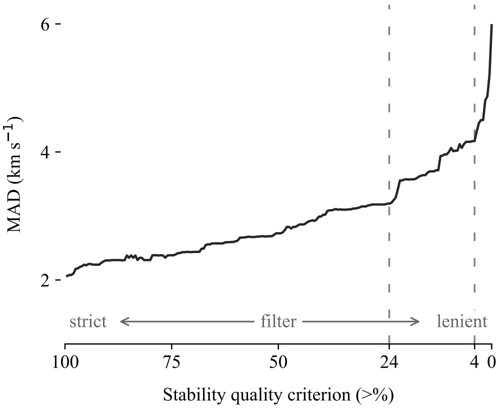


**Fig. 3.** Median absolute deviation of sources from expected 3D velocity as a function of the stability quality filter. The $x$-axis is reversed displaying very strict filter criteria on the leftmost side and lenient filter criteria toward the right side. A trend is visible where the amount of scatter over the stability filter is split into two parts, where each is characterized by a different slope. Suitable quality filters are realized by `stability` > 4% and, more conservatively, `stability` > 24%.

We compared the stream to a selection of the Pleiades members (Gaia Collaboration 2018a). By introducing a slight color offset of ($G_{BP}$–$G_{RP}$ + 0.03) to the stream, we find that the source distributions in the HRD of the Meingast 1 stream and the Pleiades match almost perfectly, as seen in Fig. 7, implying a similar age between the two stellar systems. The small color shift could imply either the presence of dust extinction towards

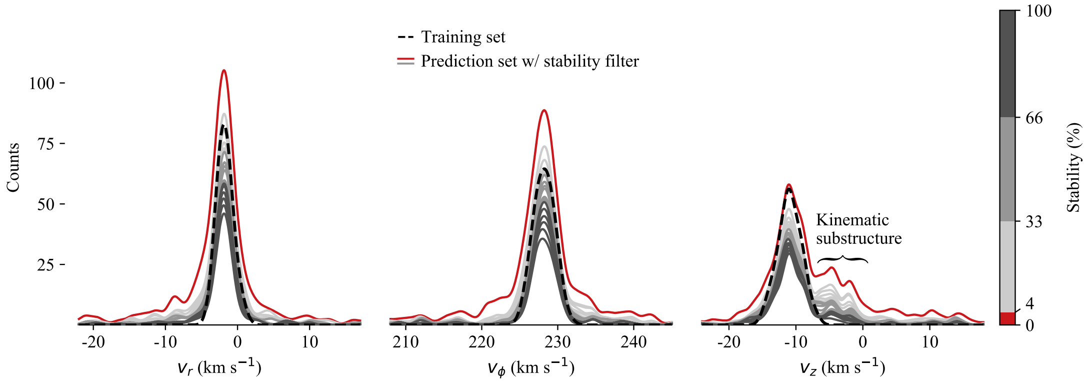


**Fig. 4.** Kernel density estimation of marginal 3D velocity distributions for various stability filter criteria. The individual lines are color-coded by the filter criteria and range from red (`stability` < 4%) to dark blue, which represents the strictest filter criterion. The distributions are sampled at constant intervals in stability with a step size of 5%. The hue change from red to shades of blue indicates the transition from the contamination dominated to the more robust filter regime. In addition, we note a kinematic substructure in the $z$-velocity distribution which is indistinguishable from other sources in all features except $v_z$.

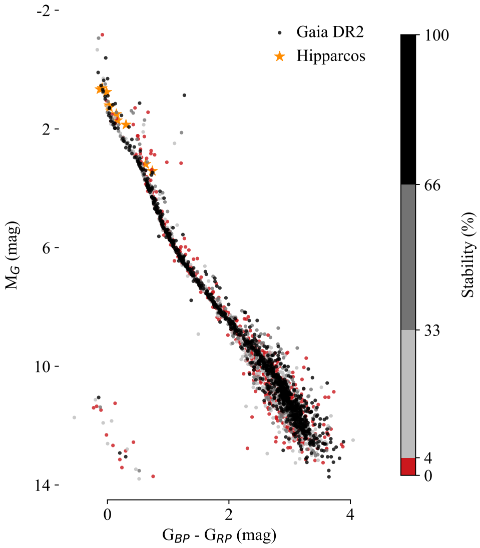


**Fig. 5.** Distribution of predicted sources in color-absolute magnitude diagram. The shades of gray encode the stability information of each source. The hue change in the color map at 4% denotes the transition from robust stream members in gray tones to less reliable sources in red. Additionally, we show 10 new potential stream members, identified by applying the same classifier to the HIPPARCOS catalog.

the Pleiades, or a lower metallicity of the stream, or both. The Pleiades are known to be affected by small amounts of extinction. Additionally, we find a slight metallicity difference between the stream and the Pleiades measured by LAMOST Liu et al. (2015), which is illustrated Fig. E.1. The plot shows a discrepancy between the mean metallicity fraction of the two stellar populations, where sources in Meingast 1 appear to be slightly more metal poor than the ones in the Pleiades, which could help to explain the reddening in color space.

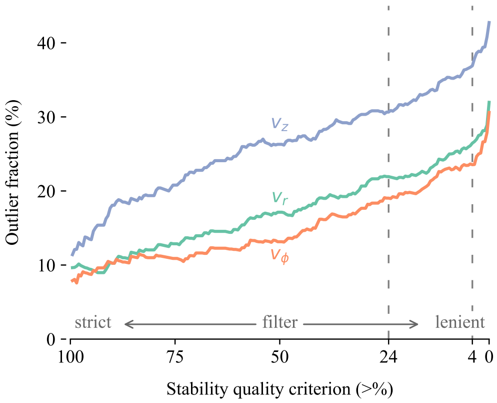


**Fig. 6.** Outlier fraction in individual velocity components for a variable stability filter criterion. Due to a newly identified kinematic substructure in $v_z$, we estimate the contamination only in the radial and azimuthal velocity components (see Sect. 4). Based on this premise, the contamination is estimated to be roughly 25% and 20% for the stability criteria >4% and >24%, respectively.

The three panels in Fig. 7 show the source distributions in the HRD of both, the Meingast 1 stream and the Pleiades, plotted on top of each other and highlighted by different colors. In the left plot, sources in the Meingast 1 stream are highlighted in red, while the Pleiades members selection are kept in gray. The center plot displays both stellar populations, which are shown in gray. The right plot displays the Pleiades in blue on top of Meingast 1 in gray. In order to make a fair comparison, we define the stability filter in such a way that the number of sources of the stream is equal to that of the Pleiades. This results in the following filter criterion: `stability` > 45.9. The particular similarity of the two distributions suggests an approximately identical age. The *Gaia* collaboration (Gaia Collaboration 2018b) estimates the age and metallicity fraction of the Pleiades to be 110 My and $Z = 0.017$, respectively. Therefore, our age estimate

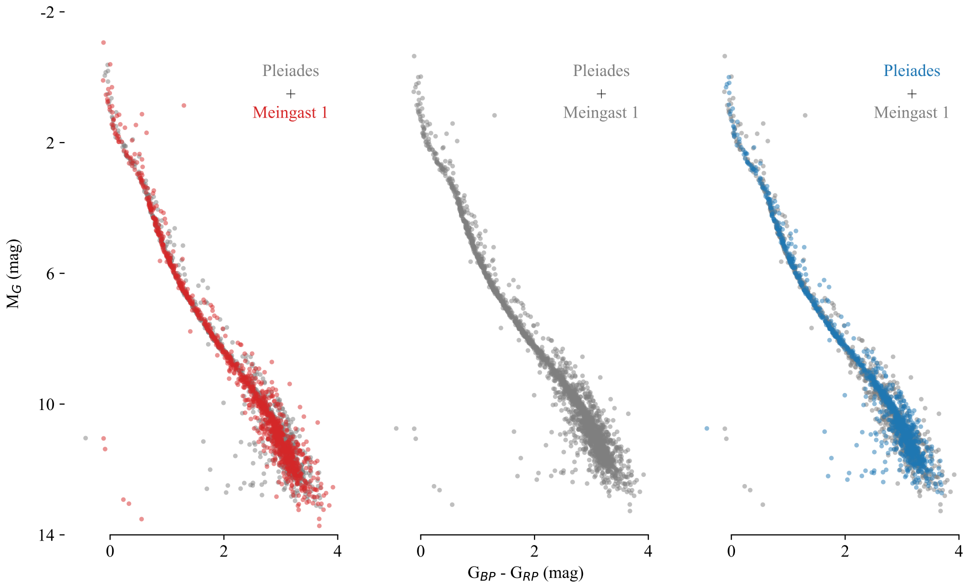


**Fig. 7.** Comparison between predicted stream members and the Pleiades member selection. The three panels show the same two data sets plotted on top of each other and highlighted by different colors. In the *left plot*, the predicted stream members are highlighted in red, while the Pleiades are kept in gray. The *center plot* displays both stellar associations in gray. The *right plot* displays the Pleiades member selection in blue on top of the predicted stellar stream in gray. We chose the stability cut to match the number of sources in the Pleiades sample in order to generate a fair comparison. The CMD distributions of the Pleiades and the predicted stream matches almost perfectly.

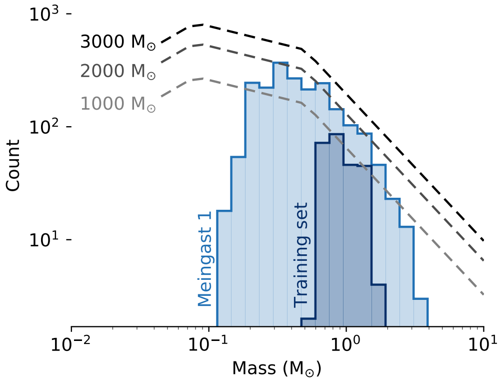


**Fig. 8.** Mass function for Meingast 1 stream sources (light blue) and the training examples (dark blue). The dotted lines indicate model IMFs within a cluster mass range of 1000–3000 $M_\odot$.

is within the expected error range, consistent with Curtis et al. (2019).

We estimated the total mass of the selected sources in accordance with Paper II by using PARSEC isochrones. Using an age estimate of 110 My and a metallicity fraction of $Z = 0.016$ results in the mass distribution shown in Fig. 8. The plot depicts the mass distribution of the training samples (dark blue) versus the predicted samples (light blue). The dotted gray lines indicate IMFs (Kroupa 2001) for clusters masses of 1000 $M_\odot$, 2000 $M_\odot$, and 3000 $M_\odot$. A comparison to the model IMFs suggests an approximate mass of 2000 $M_\odot$, as suggested in Paper II. To our knowledge, this makes the Meingast 1 stream the most massive stellar stream in the solar neighborhood.

Finally, we can speculate on the origin of the Meingast 1 stream. In Paper II, we put forward the possible cluster versus association scenarios for the origin of this extended structure, but opted not to favor one over the other, even though we found evidence for the existence of at least four overdensities in the structure. This ambiguity resulted mainly from the older age derived in Paper II, which made it not obvious to favor one of the two scenarios without a proper simulation. The much younger age determined in Curtis et al. (2019), that we confirm in this work, allowed these authors to favor the association scenario (because ~100 Myr is too short for cluster dissolution). The best and most obvious example is the Pleiades cluster, which is a relatively compact cluster with essentially the same age as Meingast 1. The velocity substructure we found in this paper (see Fig. 4) now allows us to make a stronger case favoring the association scenario as the likely initial configuration of Meingast 1. Unlike compact clusters, stellar associations such as Sco-Cen are known to have velocity substructures of a few to several km s$^{-1}$ (e.g., Wright & Mamajek (2018), Goldman et al. (2018)). A more meaningful look into the origin of Meingast 1, which would require n-body simulations and the effects of the Galactic potential, will enable us to clarify the origin of this mesmerizing structure.

## 5. Summary and conclusion

We revisited the stream discovered in Meingast et al. (2019) to search for new members using *Gaia* DR2 data and a machine-learning approach. Using the original source selection as training data, we deployed a bagging classifier of one-class support vector machines to the full *Gaia* DR2 data, searching for new stream members in position and tangential velocity space. The ensemble classifier is created in a hyper-parameter search combined with a model selection that rejects models that do not meet a set of preconditions. The resulting set of classifiers creates a variable prediction frequency for possible stream member stars, which we used as a criterion to select high-fidelity sources. Subsequently, we validated the newly found sources in the HR diagram and the Cartesian velocity distribution.

In total, we find about 2000 stream high-fidelity member stars, increasing the source population approximately tenfold. As the newly predicted stream members are no longer limited by radial velocity measurements, the new selection substantially extends the main sequence to unveil the stream's population across the entire stellar mass spectrum, from B stars to M stars, including white dwarfs. In a comparison in the color-absolute magnitude diagram, we find that, apart from being slightly more metal poor, the stream is indistinguishable from that of the Pleiades cluster, suggesting a similar age. In the mass range at which we are mostly complete, $\sim 0.2 < M_\odot <\sim 4\,M_\odot$, we identify a normal IMF. This comparison allows us to estimate the total mass of the stream to approximately 2000 $M_\odot$, making it by far the most massive stream we know. Additionally, we find several white dwarfs as members of the stream. We speculate with more confidence, given the velocity substructure found in this work, that Meingast 1 is the likely outcome of a stellar association, but call for a full, state-of-the-art simulation to be done to characterize the origin of this mesmerizing structure.

*Acknowledgements.* We thank the anonymous referee for his insightful comments that helped to improve this study. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This research made use of Astropy, a community-developed core Python package for Astronomy (Astropy Collaboration 2018). We also acknowledge the various Python packages that were used in the data analysis of this work, including NumPy (van der Walt et al. 2011), SciPy (Virtanen et al. 2019), scikit-learn (Pedregosa et al. 2011), Pandas (McKinney et al. 2010), and Matplotlib (Hunter 2007). This research has made use of the SIMBAD database operated at CDS, Strasbourg, France (Wenger et al. 2000)

## References

Astropy Collaboration (Price-Whelan, A. M., et al.) 2018, AJ, 156, 123
Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Mantelet, G., & Andrae, R. 2018, AJ, 156, 58
Binks, A. S., Jeffries, R. D., & Ward, J. L. 2018, MNRAS, 473, 2465
Breiman, L. 1996, Mach. Learn., 24, 123
Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127
Chang, C.-C., & Lin, C.-J. 2011, ACM Trans. Intell. Syst. Technol., 2, 1
Cortes, C., & Vapnik, V. 1995, Mach. Learn., 20, 273
Curtis, J. L., Agüeros, M. A., Mamajek, E. E., Wright, J. T., & Cummings, J. D. 2019, ApJ, 158, 77
Gagné, J., Lafrenière, D., Doyon, R., Malo, L., & Artigau, É. 2014, ApJ, 783, 121
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration 2018a, VizieR Online Data Catalog: J/A+A/616/A10
Gaia Collaboration (Babusiaux, C., et al.) 2018b, A&A, 616, A10
Gaia Collaboration (Brown, A. G. A., et al.) 2018c, A&A, 616, A1
Goldman, B., Röser, S., Schilbach, E., Moór, A. C., & Henning, T. 2018, ApJ, 868, 32
Grandvalet, Y. 2004, Mach. Learn., 55, 251
Hunter, J. D. 2007, Comput. Sci. Eng., 9, 90
Ibata, R. A., Malhan, K., & Martin, N. F. 2019, ApJ, 872, 152
Kroupa, P. 2001, MNRAS, 322, 231
Lindegren, L. 2018, Re-normalising the astrometric chi-square in Gaia DR2, Technical Report GAIA-C3-TN-LU-LL-124-01
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2
Liu, X.-W., Zhao, G., & Hou, J.-L. 2015, Res. Astron. Astrophys., 15, 1089
McKinney, W. 2010, in Data structures for statistical computing in python, eds. S. van der Walt, & J. Millman, Proc. 9th Python Sci. Conf, 51
Meingast, S., & Alves, J. 2019, A&A, 621, L3
Meingast, S., Alves, J., & Fürnkranz, V. 2019, A&A, 622, L13 (Paper II)
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Riedel, A. R., Blunt, S. C., Lambrides, E. L., et al. 2017, AJ, 153, 95
Röser, S., & Schilbach, E. 2020, A&A, 638, A9
Röser, S., Schilbach, E., & Goldman, B. 2019, A&A, 621, L2
Schölkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., & Williamson, R. C. 2001, Neural Comput., 13, 1443
Soubiran, C., Bienaymé, O., Mishenina, T. V., & Kovtyukh, V. V. 2008, A&A, 480, 91
Torres, C. A. O., Quast, G. R., Melo, C. H. F., & Sterzik, M. F. 2008, Young Nearby Loose Associations, 5, 757
van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Comput. Sci. Eng., 13, 22
van Leeuwen, F. 2007, A&A, 474, 653
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2019, SciPy 1.0-Fundamental Algorithms for Scientific Computing in Python
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9
Wright, N. J., & Mamajek, E. E. 2018, MNRAS, 476, 381

## Appendix A: Training process

The training of each individual predictor in the full model ensemble is summarized in the following two steps.

Firstly, we select a random pair of hyper-parameters ($\gamma_i$, $\nu_i$, $(c_x/c_v)_i$) and train a model with tenfold cross validation (CV). Due to a contamination of field stars of a few percent in Paper II, we encourage stricter and more compact descriptions of the stream (in 5D), ignoring potential outliers in the training sample. In a first selection step, we filter models with a low average accuracy across the holdout sets of <0.5 or a standard deviation of above 0.15. The standard deviation filter helps to obtain fairly conclusive predictors for different subsamples on a fixed set of hyper-parameters.

Secondly, models that pass the CV step are deployed on the full data set (see Sect. 2). In an effort to minimize contamination of nearby[5] field stars and thus boost robustness of the prediction, we train the model on 10 bootstrap samples, with a sample size of 80% of the training data size. The union of all 10 predictions is then considered the final model. Before we add the newly trained model (with the hyper-parameter set ($\gamma_i$, $\nu_i$, $(c_x/c_v)_i$) into the final bagging classifier, we validate its performance against our prior beliefs about the approximate model structure described in Sect. 3.2

## Appendix B: Validation process

After training a classifier, we validate its ability to capture important physical aspects about the estimated size and shape of the stellar stream. We require the classifier to capture at least the following criteria:

1. The number of predicted stream members $N_s$ must not exceed a physically sensible range, which is limited to $N_s \in [500, 5000]$.

2. The extent of the predicted stream members in position and proper motion space must be similar to the original ones.

3. The cylindrical velocity distribution of the stream members must not deviate too much from the training sample distribution.

The similarity condition (2.) is achieved by requiring the dispersion of the predicted to the original stream members in position and proper motion space to be approximately equal. We approximate the extent, or dispersion of the stream in both spaces by a single number, namely the mean distance $\overline{d}$ of its member stars to the centroid of the full stream. For a point in position space $\boldsymbol{r} = (x, y, z)$ and its corresponding centroid $\boldsymbol{r}_c$, $\overline{d}_{\boldsymbol{r}}$ is

$$\overline{d}_{\boldsymbol{r}} = \frac{1}{N}\sum_{i}^{N} \|\boldsymbol{r}_i - \boldsymbol{r}_c\|, \tag{B.1}$$

where $N$ is the number of stars belonging to the cluster. Respectively, in proper motion space with a point $\boldsymbol{v} = (\mu_\alpha, \mu_\delta)$ and centroid $\boldsymbol{v}_c$, $\overline{d}_{\boldsymbol{v}}$ is:

$$\overline{d}_{\boldsymbol{v}} = \frac{1}{N}\sum_{i}^{N} \|\boldsymbol{v}_i - \boldsymbol{v}_c\|. \tag{B.2}$$

We use these two structure parameters $\overline{d}_{\boldsymbol{r}}$ and $\overline{d}_{\boldsymbol{v}}$ to determine the extent of the stream in position and proper motion space, respectively. Our aim is to find models whose predicted points retain a similar dispersion to the original ones. To avoid overfitting, we compare the dispersion of the prediction set to the training set which acts as an upper limit:

$$\overline{d}^{\,\mathrm{orig}}_{\boldsymbol{r}/\boldsymbol{v}} > \overline{d}^{\,\mathrm{pred}}_{\boldsymbol{r}/\boldsymbol{v}} . \tag{B.3}$$

Lastly, we control the centroid position of the predicted stream members to avoid systematic shifts. The predicted and original stream centroid must be reasonably close to each other with respect to the average dispersion of training points.

$$\|\boldsymbol{r}_c^{\mathrm{org}} - \boldsymbol{r}_c^{\mathrm{pred}}\| < \overline{d}^{\,\mathrm{org}}_{\boldsymbol{r}} \times 0.1 \tag{B.4}$$

$$\|\boldsymbol{v}_c^{\mathrm{org}} - \boldsymbol{v}_c^{\mathrm{pred}}\| < \overline{d}^{\,\mathrm{org}}_{\boldsymbol{v}} \times 0.1 \tag{B.5}$$

The third condition is implemented by examining the contamination of predicted samples compared to the training sample. To get a rough estimate of the contamination, we compare the galactocentric velocity distribution, meaning $\boldsymbol{v} = (v_r, v_\phi, v_z)$, of the predicted sources to the training sample. Instead of comparing the velocity dispersion of both samples, we characterize the level of contamination by considering the fraction of outlier sources. This way, we try to mitigate the influence of large outliers, which increase the dispersion drastically for such a low number of sources. In order to characterize outlier sources, we consider the training examples. Assuming that almost all sources lie within the $\pm 3\sigma$ range around the mean, we consider the ratio of sources lying outside of the $3\sigma$ range compared to the total amount of sources. A classifier is rejected if on average, across the individual velocity components, more than 25% of sources are considered outliers. The aim of this criterion is to remove models that extend into a region of feature space where the radial velocity distribution does not match our assumption of a co-moving structure.

## Appendix C: Parameter tuning results

The hyper-parameter search in combination with a classifier selection and validation step (see Sect. 3.2) yields a set of approvedparameter triples ($\nu_i$, $\gamma_i$, $(c_x/c_v)_i$) that make up the final OCSVM bagging predictor. The distribution of accepted triples is displayed in Fig. C.1. The color information illustrates the accepted model faction within a certain hyper-parameter bin range. A model is accepted if it passes the quality criteria presented in Sect. 3.2. The model ensemble consists of 8515 individual predictors.

[5] Nearby refers to sources in the vicinity of the stellar stream in the 5D feature space.

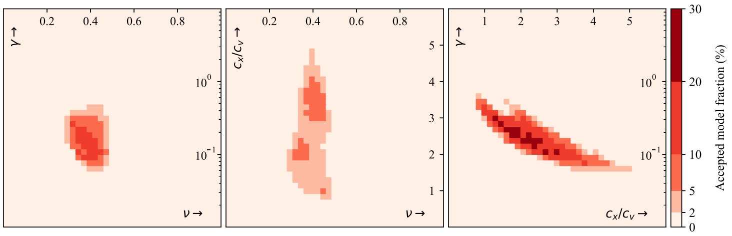


**Fig. C.1.** Hyper-parameter search in parameters $\nu$, $\gamma$, and $c_x/c_v$ yielding the one-class support vector machine bagging predictors. The color information illustrates the accepted model faction within a certain hyper-parameter bin range. A classifier is accepted if it passes the quality criteria presented in Sect. 3.2. The model ensemble consists of 8515 individual predictors.

## Appendix D: Quality criteria

In general, the source identification method we present in this paper is independent of any quality criteria. However, in order to show the distribution of stars in the color magnitude diagram, we apply the following error criteria on data quality. Following the description in Lindegren et al. (2018) the five-parameter solution depends on the number of visibility periods used for a certain source. A visibility period is defined as a group of observations separated from other groups by a gap of at least four days. Since a five-parameter solution is accepted only for `visibility_periods_used` $> 6$, we implement said criterion.

A recommended astrometric quality parameter is the renormalised unit weight error (RUWE) described by Lindegren (2018). It is based on a re-calibration of the unit weight error described in Lindegren et al. (2018). We follow the advice in the technical note (Lindegren 2018) and use the criterion `RUWE` $< 1.4$ to select astrometrically reliable sources. Furthermore, we implement additional astrometric quality measures, `astrometric_sigma_5D_max` $< 0.5$ and $\varpi/\sigma_\varpi > 10$, which reduce the number erroneous measurements.

Finally, we adopt the following photometric quality criteria, `phot_bp_mean_flux_over_error` $> 10$ and `phot_rp_mean_flux_over_error` $> 10$.

## Appendix E: Metal content

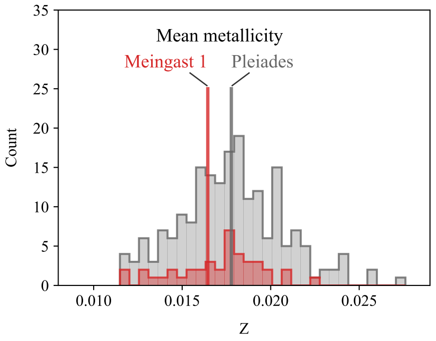


**Fig. E.1.** Comparison of metallicity fraction of Pleiades and Meingast 1 memeber stars. The vertical lines indicate the mean metal content of both populations. We find that the members of the Meingast 1 association are slightly more metal poor than the Pleiades.

Figure E.1 shows a comparison of the metallicity fraction $Z$ between a Pleiades member selection (Gaia Collaboration 2018a) and the stream members. A cross-match of the Pleiades and stream source selections to the LAMOST DR5 Liu et al. (2015) catalog results in 383, and 83 matches, respectively. The conversion from chemical abundance ratios [Fe/H] to the metallicity fraction $Z$ has been made in accordance with the PARSEC (Bressan et al. 2012) solar value of $Z = 0.015$. Subsequently, we filter out the most untrustworthy sources by requiring that the error of the measured chemical abundance ratios [Fe/H] is below 0.05 and [Fe/H] $> -1$. Additionally, we only select sources above an effective temperature of 5000 K. These criteria yield 197 and 44 matched sources for the Pleiades and the Meingast 1 stream, respectively. The metal content distributions of the Pleiades and stream members show a large scatter, but the positions of their respective mean indicate that the Meingast 1 stream members appear to be slightly more metal poor compared to the Pleiades member stars.

## Appendix F: Hipparcos source selection

Compared to the training samples from the *Gaia* DR2 catalog, the Hipparcos sources have larger associated standard errors of measured quantities. Considering the higher uncertainty in the Hipparcos catalog variables, we adopt a more conservative stability filter criterion of `stability` $> 50\%$. Despite a rather high stability cut, a large standard error increases the chance of contaminant stars falling into the selection. Therefore, we adopt a second quality filter where we sample each data point from marginal normal distributions centered on the provided mean value with a standard deviation of the provided standard error of each observable. We then draw 100 samples per source from these marginal distributions and count how often these re-sampled sources are again predicted to be a stream members with `stability` $> 50\%$. Eventually, this quality criterion yields 11 additional sources with a re-sampling fraction of over 50%.

## Appendix G: Table content

**Table G.1.** Contents of the source catalog, which are available online via CDS.

| Column name | Description |
|---|---|
| source_id | *Gaia* DR2 source identification number |
| ra | RA (deg) |
| dec | Declination (deg) |
| $X$ | $x$-Position (pc) |
| $Y$ | $y$-Position (pc) |
| $Z$ | $z$-Position (pc) |
| pmra | $\mu_\alpha$ (mas yr$^{-1}$) |
| pmdec | $\mu_\delta$ (mas yr$^{-1}$) |
| Stability | Stability percentage (%) |
| q1 | Filter criterion Q1 (bool); see Appendix D |
| q2 | Filter criterion Q2 (bool); see Paper II |

**Notes.** The positional data XYZ are measured in Galactic Cartesian coordinates centered on the Sun.

The content of the published source catalog is summarized in Table G.1.

## 3.14. The $\rho$ Ophiuchi region revisited with Gaia EDR3 - Two young populations, new members, and old impostors.

### Full publication details

Grasser, N., **Ratzenböck, S.**, Alves, J., Großschedl, J., Meingast, S., Zucker, C., Hacar, A., Lada, C., Goodman, A., Lombardi, M., Forbes, J. C., Bomze, I. M., & Möller, T. 2021. "The $\rho$ Ophiuchi region revisited with Gaia EDR3 - Two young populations, new members, and old impostors." *Astronomy & Astrophysics. Supplement Series*, 652, A2.

Credit: Grasser et al., A&A, 652, A2, 2021, reproduced with permission © ESO.

### Author contributions

The paper is co-authored by Natalie Grasser, me, João Alves, Josefa Großschedl, Stefan Meingast, Catherine Zucker, Alvaro Hacar, Charles Lada, Alyssa Goodman, Marco Lombardi, John C. Forbes, Immanuel M. Bomze, and Torsten Möller. As the leading author, the Bachelor student Natalie Grasser performed the literature search of $\rho$ Oph sources, compiled a list of high-fideltiy members, and wrote the main parts of the paper. As the second author, I further developed the `Uncover` method to the case of fuzzy prior knowledge, performed the analysis, and wrote the methodology section.

João Alves, Josefa Großschedl, and Stefan Meingast also contributed to the writing and interpretation of results and discussion. João Alves supervised the project and offered suggestions along the way. Catherine Zucker, Alvaro Hacar, Charles Lada, Alyssa Goodman, Marco Lombardi, John C. Forbes, Immanuel M. Bomze, and Torsten Möller helped revise the final version.

### Information on the Status

Submission date: 28 January 2021
Accepted: 6 June 2021
Published: 30 July 2021

A&A 652, A2 (2021)
https://doi.org/10.1051/0004-6361/202140438
© ESO 2021

Astronomy & Astrophysics

# The $\rho$ Ophiuchi region revisited with *Gaia* EDR3

## Two young populations, new members, and old impostors[⋆,⋆⋆]

Natalie Grasser[1], Sebastian Ratzenböck[2], João Alves[1,2,3], Josefa Großschedl[1], Stefan Meingast[1], Catherine Zucker[4], Alvaro Hacar[1], Charles Lada[4], Alyssa Goodman[3,4], Marco Lombardi[5], John C. Forbes[6], Immanuel M. Bomze[2], and Torsten Möller[2]

[1] University of Vienna, Department of Astrophysics, Türkenschanzstrasse 17, 1180 Wien, Austria
e-mail: natalie.grasser@univie.ac.at
[2] Data Science at University of Vienna, Währinger Straße 29, 1090 Vienna, Austria
[3] Radcliffe Institute for Advanced Study, Harvard University, 10 Garden Street, Cambridge, MA 02138, USA
[4] Center for Astrophysics | Harvard & Smithsonian, 60 Garden St., Cambridge, MA 02138, USA
[5] University of Milano, Via Celoria, 16, 20133 Milano, Italy
[6] Flatiron Institute, Simons Foundation, 162 Fifth Avenue, New York, NY 10010, USA

Received 28 January 2021 / Accepted 6 June 2021

### ABSTRACT

*Context.* Young and embedded stellar populations are important probes of the star formation process. Their properties and the environments they create have the potential to affect the formation of new planets. Paradoxically, we have a better census of nearby embedded young populations than of the slightly more evolved optically visible young populations. The high accuracy measurements and all-sky coverage of *Gaia* data are about to change this situation.
*Aims.* This work aims to construct the most complete sample to date of young stellar objects (YSOs) in the $\rho$ Oph region.
*Methods.* We compile a catalog of 1114 Ophiuchus YSOs from the literature and cross-match it with the *Gaia* EDR3, *Gaia*-ESO, and APOGEE-2 surveys. We apply a multivariate classification algorithm to this catalog to identify new, co-moving population candidates.
*Results.* We find 191 new high-fidelity YSO candidates in the *Gaia* EDR3 catalog belonging to the $\rho$ Oph region. The new sources appear to be mainly Class III M stars and substellar objects and are less extincted than the known members, while we find that 28 of the previously unknown sources are YSOs with circumstellar disks (Class I or Class II). The analysis of the proper motion distribution of the entire sample reveals a well-defined bimodality, implying two distinct populations sharing a similar 3D volume. The first population comprises young stars' clusters around the $\rho$ Ophiuchi star and the main Ophiuchus clouds (L1688, L1689, L1709). In contrast, the second population is slightly older (~10 Myr), more dispersed, has a distinct proper motion, and is possibly from the Upper Sco group. The two populations are moving away from each other at about 4.1 km s$^{-1}$ and will no longer overlap in about 4 Myr. Finally, we flag 17 sources in the literature sample as likely impostors, which are sources that exhibit large deviations from the average properties of the $\rho$ Oph population. Our results show the importance of accurate 3D space and motion information for improved stellar population analysis.

**Key words.** astrometry – methods: data analysis – stars: formation – stars: pre-main sequence

## 1. Introduction

Since the development of millimeter-wave receivers and infrared (IR) detectors in the 1970s, local star formation studies have mostly concentrated on the densest star-forming structures in molecular clouds. Successive generations of instruments have opened a fundamental window into molecular cloud structure, cloud fragmentation, and collapse and have unveiled the dust-enshrouded young stellar object (YSO) populations in nearby clouds. This approach has generated an almost paradoxical situation where we currently know more about the very young dust-obscured populations than we know about the more evolved and optically revealed population in nearby star-forming regions.

More evolved YSOs show less IR excess emission and escape detection in IR surveys but are critical to reconstructing a region's star formation history. Therefore, identifying the young optically visible population is essential for reconstructing a star formation event. Moreover, sources in the unobscured environments of nearby star-forming gas include some of the lowest-mass objects (brown dwarfs and planetary-mass objects) and some of the closest proto-planetary disks we can study, the latter becoming important targets for resolved ALMA studies (e.g., ALMA Partnership 2015) in the submillimeter wavelength range.

Optical data from the *Gaia* mission (Gaia Collaboration 2016), with its exquisite sensitivity and all-sky coverage, have changed this situation. With its latest data release, the mission has made a breakthrough in terms of studies of gas shape and motion (Großschedl et al. 2018, 2021) and previously unknown young stellar structures (Meingast et al. 2019, 2021), significantly improving upon its second data release, *Gaia* DR2 (Gaia Collaboration 2018). In this work, we revisit one of the nearest star-forming regions, the $\rho$ Ophiuchi region, by using the newly available *Gaia* Early Data Release 3 (EDR3) data (Gaia Collaboration 2021).

[⋆] Interactive 3D version of Figures 6 and 7 are available online at https://www.aanda.org/10.1051/0004-6361/202140438/olm
[⋆⋆] The final catalog is only available at the CDS via anonymous ftp to cdsarc.u-strasbg.fr (130.79.128.5) or via http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/652/A2

The $\rho$ Ophiuchi ($\rho$ Oph) star-forming region (Wilking et al. 2008) is one of the nearest active star-forming regions, at a distance of approximately 139 pc (Lombardi et al. 2008; Zucker et al. 2020). It comprises the cluster of young stars around the $\rho$ Ophiuchi star (Pillitteri et al. 2016) and the young stars associated with the dense gas in the Ophiuchus cloud complex, mainly the L1688, L1689, and L1709 clouds (Loren 1989a,b). Due to its youth and proximity to Earth, it has played an essential role in many star formation studies, in particular in the definition of the YSO classes (Wilking & Lada 1983; Lada & Wilking 1984; Andre et al. 1993; Greene et al. 1994). The $\rho$ Oph region is located in the foreground of the southeastern edge of Upper Scorpius, which is a subgroup of the Scorpius-Centaurus OB association, and has a distance of around 145 pc (Wilkinson et al. 2018). It has long been suspected that star formation in the $\rho$ Oph region was triggered by feedback from massive stars from Upper Sco (Vrba 1977; Loren & Wootten 1986; Loren 1989a,b; de Geus 1992).

The youngest stars in the region are associated with the densest gas in the Ophiuchus cloud complex, mostly L1688, with an average age of about 0.3 Myr (Greene & Meyer 1995; Luhman & Rieke 1999), while the stellar population on the lower column density surface has an average estimated age of 2–5 Myr (Wilking et al. 2008; Erickson et al. 2011). There are three main dark clouds in the $\rho$ Oph complex, the mentioned Lynds dark clouds L1688, L1689 and L1709 (Lynds 1962; Loren 1989a,b). The large column density toward particular regions in these clouds, where the optical extinction can reach values of up to $A_V$ above 40–50 mag (Wilking & Lada 1983; Wilking et al. 1989; Lombardi et al. 2008), make IR observations essential for studying the embedded young stellar population in the cloud. There is a rich embedded cluster of YSOs in L1688, which is mostly invisible at optical wavelengths, whose stars have not yet dispersed (Ducourant et al. 2017).

In this paper, we apply the recently developed method from Ratzenböck et al. (2020) to *Gaia* EDR3 data, to unveil the most complete sample to date of YSOs toward the $\rho$ Oph region. The method uses the astrometric properties of known YSOs in combination with a bagging classifier of one-class support vector machines (OCSVMs) on *Gaia* EDR3 data to perform a 5D search (3D positions and 2D proper motions) for possible new population members. The algorithm creates a hyper-surface around the positional and proper motion distribution of the input samples in a 5D space to find new sources with similar properties. Radial velocities of the input population are also necessary for constraining the models. We remove models that identify stars with significantly different 3D velocities than the those of training set.

In Sect. 2 we present the data used in this work; this includes known sources from the literature, which we cross-matched with further astronomical surveys. In Sect. 3 we summarize how the classification algorithm operates to identify new sources. We present the results of the algorithm in Sect. 4, including a detailed analysis. In Sect. 5 we discuss some implications of our findings.

## 2. Data

### 2.1. Literature catalog

In this section, we summarize how we compiled our literature catalog of $\rho$ Oph sources. This work is based on studies of $\rho$ Oph and L1688 from 11 papers, which are summarized in Table 1, including the number of sources utilized from each work, which results in a total of 1114 sources. We note that the same source can be presented in more than one work. We assign each paper a digit for citation purposes in our final catalog. Some papers also include sources from IR observations, which are essential for a complete sample due to the high optical extinction in the region and for identifying Class II and earlier Class YSOs. The highest number of sources are provided by Wilking et al. (2008), Cánovas et al. (2019), and Esplin & Luhman (2020). Duplicates were removed with an internal match within a 1.0 arcsec match radius and an internal match on the *Gaia* source IDs. Our result is a final literature table of 1114 unique sources.

Sullivan et al. (2019) provide radial velocities on their sources, while Ducourant et al. (2017) provide proper motions on their sources. Astrometric data (proper motions, parallaxes, and radial velocities) for the remaining sources were obtained by selecting three surveys for cross-matching with our literature sample, which is essential for identifying new sources with the algorithm. The *Gaia* survey provides us with unprecedented astrometry with improved quality and statistics compared to any previous comparable survey, such as HIPPARCOS (Perryman et al. 1997). Therefore, proper motions and parallaxes were obtained from *Gaia* EDR3 (Gaia Collaboration 2021). To complement *Gaia* astrometry and constrain the models of the algorithm, we combined it with radial velocities from APOGEE-2 (Majewski et al. 2017), a large-scale spectroscopic survey conducted in the near-infrared, and *Gaia*-ESO (Gilmore et al. 2012), a spectroscopic survey by the European Southern Observatory (ESO) combined with the *Gaia* astrometry catalog. Radial velocities from these surveys deliver superior resolution and statistics compared to radial velocities from *Gaia*.

A cross-match of the literature sources with data from *Gaia* EDR3 yielded a total of 675 matches, which is 60.5% of the entire literature sample, leaving many sources without *Gaia* equivalents. One explanation for this is that *Gaia* is only sensitive to optical wavelengths, while many of the obtained literature sources are too embedded in the cloud and can only be observed at IR wavelengths. Additionally, several sources, such as from Esplin & Luhman (2020), are brown dwarfs, which are often too faint to be seen by *Gaia*. A cross-match of the total literature sources with APOGEE-2 resulted in 188 matches, while a cross-match with *Gaia*-ESO data yielded 61 matches in our literature catalog. For sources with multiple measurements, higher priority was given to surveys with higher accuracy. Therefore we use *Gaia* proper motions and parallaxes over those obtained from the literature. For sources with multiple radial velocity values, data from *Gaia*-ESO has the highest priority, followed by APOGEE-2 and then *Gaia*.

The distances to the sources were calculated through the inverse of the parallax, which is a good approximation for the relatively close distance to the region of about 130–140 pc (e.g., Luri et al. 2018). Furthermore, the tangential velocities $v_\alpha$ and $v_\delta$, as well as their errors, were calculated through the proper motions and parallaxes, as shown in Eqs. (A.1)–(A.4). For a better overview, we list the symbols and abbreviations of frequent parameters used throughout this paper:

- $\alpha, \delta$ (deg): right ascension and declination
- $l, b$ (deg): galactic longitude and latitude
- $\varpi$ (mas): parallax of the sources
- $d$ (pc): distance to the sources, inverse of parallax
- $\mu_\alpha^*$ (mas yr$^{-1}$): $\mu_\alpha\cos(\delta)$, proper motion along $\alpha$
- $\mu_\delta$ (mas yr$^{-1}$): proper motion along $\delta$
- $v_\mathrm{r}$ (km s$^{-1}$): heliocentric radial velocity
- $v_\alpha, v_\delta$ (km s$^{-1}$): tangential velocities along $\alpha$ and $\delta$
- $v_l, v_b$ (km s$^{-1}$): tangential velocities along $l$ and $b$

**Table 1.** Overview of the literature that was used to collect young stellar members of the $\rho$ Oph region.

| Paper | Method | Sources used | Ref |
|---|---|---|---|
| Greene et al. (1994) | Mid-IR photometric study | 56 | 1 |
| Haisch et al. (2002) | Near- and mid-IR observations | 13 | 2 |
| Padgett et al. (2008) | Multiband Imaging Photometer for *Spitzer* (MIPS) point-sources | 46 | 3 |
| Wilking et al. (2008) | X-ray and IR photometric and spectroscopic surveys | 316 | 4 |
| Evans et al. (2009) | *Spitzer* c2d Legacy survey | 292 | 5 |
| Dunham et al. (2015) | *Spitzer* c2d and GB Legacy surveys | 292 | 6 |
| Rigliaco et al. (2016) | Dynamical analysis with *Gaia*-ESO survey | 45 | 7 |
| Ducourant et al. (2017) | Near-IR observations to determine proper motions | 82 | 8 |
| Cánovas et al. (2019) | Density-based clustering algorithms with *Gaia* DR2 | 831 | 9 |
| Sullivan et al. (2019) | Radial velocity survey with data from IR spectrographs | 34 | 10 |
| Esplin & Luhman (2020) | Astrometry from *Gaia* DR2, proper motions from *Spitzer* | 373 | 11 |

**Notes.** The table lists the used methods and the number of sources we obtained from each paper, resulting in a total of 1114 literature sources. We note that the same source can be presented in more than one work.

- $X, Y, Z$ (pc): positions in Galactic Cartesian coordinates, where $X$, $Y$, and $Z$ point toward the Galactic center, the direction of the Galactic rotation, and the north Galactic pole, respectively
- $U, V, W$ (km s$^{-1}$): velocities in Galactic Cartesian coordinates

### 2.2. Impostors

We have discovered several sources within the literature catalog that have properties that do not fit very well to the region's average astrometric values. In Appendix A we list the interval ranges in which most of the distance, radial velocity, and tangential velocity values in $\rho$ Oph are found, which were used to create a training set (Sect. 3.1). There are 28 sources that have at least one of these values outside our defined intervals and smaller errors than the upper limits listed in Appendix A. However, some of them have values that are still close to the interval limits and could therefore still be a part of $\rho$ Oph, since deviating motions can be caused by interactions in the cluster or by multiple stellar systems. There are, nonetheless, several sources with very large radial velocity deviations from the average. Therefore, we identified all sources with radial velocities $v_\mathrm{r} < -30$ and $v_\mathrm{r} > 20$ and errors <3 as uncertain members and labeled them as impostor candidates in our catalog. We found 17 of such impostor candidates among the literature sources. However, it is important to note that these deviating radial velocities could be caused by multiplicity, such as binary star systems, and could therefore still be members. Due to this uncertainty, and since our intervals are more or less arbitrarily defined, we chose not to remove these impostors from our catalog. Instead, we created a separate column named "Impostors," where they are labeled with a "1" and all others are labeled with a "0".

## 3. Methods

In our work, we applied the classification strategy described in Ratzenböck et al. (2020) for identifying new members of the $\rho$ Oph region in the *Gaia* EDR3 catalog. The goal of Ratzenböck et al. (2020) was to model the extent of the Meingast 1 stellar stream (Meingast et al. 2019) in the combined space of proper motions and positions and subsequently use it to identify new members in *Gaia* DR2, while we use the latest data release EDR3. The model consists of multiple OCSVM classifiers in a bagging ensemble. In the following we refer to sources classified by the OCSVM as members of a stellar population as "predicted" members. Based on the model quality, the prediction set contains known and potentially new candidate sources. In the following we discuss means of selecting high quality models via prior assumption filters.

### 3.1. Training set selection

To provide reliable sources for the classification algorithm, we created a training set by removing outliers and applying quality cuts. The quality cuts are described in Appendix A, where we also present the training set. To guarantee a high-fidelity training set, we limited our selection to sources with radial velocity measurements. Since the hypersurface created by the OCSVM algorithm depends heavily on the distribution of peripheral sources, it is susceptible to outliers. The use of a soft-margin SVM somewhat mitigates this, but to further reduce the effect of potential contaminants on the final model shape, we removed the most extreme outliers from the training set as well. To do so, we estimated the local outlier factor (Breunig et al. 2000) of each source in 5D and removed 5% of the training set with the highest outlier factor. This removal lead to a final training set of 150 sources, which corresponds to 13.5% of the literature sample.

### 3.2. Model selection and prior assumptions

Due to the high model flexibility of OCSVMs, choosing adequate model parameters is critical to guarantee a suitable description of the stellar system. Instead of directly selecting models in the OCSVM hyperparameter space, Ratzenböck et al. (2020) have suggested to constrain the models via prior assumptions they have to adhere to, implicitly tuning the model parameters. In addition, as summary statistics, prior assumptions are usually much easier to interpret compared to the original OCSVM parameters. Each set of prior beliefs corresponds to a distribution of allowed models in the input parameter space, such that there is a mapping from a prior assumption tuple to regions in the OCSVM parameter space that contain models that adhere to the given rule set. Instead of explicitly characterizing this map, we sampled uniformly from the OCSVM hyperparameter space and removed unfit models. To determine a set of prior assumptions for identifying new high-fidelity $\rho$ Oph members, we considered their application in Ratzenböck et al. (2020). The prior assumptions were motivated by the training set selection process. Since only sources with radial velocities were previously identified to be part of the Meingast 1 stream, the authors formulated prior assumptions based on completeness arguments regarding radial velocities. Specifically, the goal was to find still unknown members without radial velocity measurements, which were confined to the training set extent. However, the $\rho$ Oph training set selection function is much more complex as we combined radial velocity information across multiple data surveys. This also means we have much less information about potentially concealed $\rho$ Oph members. Therefore, we adjusted the previous assumptions to the $\rho$ Oph population. In the following, we briefly discuss the selection of the six prior assumptions constraining models via the number and distribution of predicted sources.

#### 3.2.1. Population size

Firstly, we aim to restrict the number of sources a model identifies. Because the $\rho$ Oph population has been studied extensively – with some studies using *Gaia* data as well – we do not expect to find a dramatic increase in overall population size. Based on the number of *Gaia* EDR3 sources in the literature catalog, we estimated a very conservative upper limit of a maximum population size of about twice the number of sources from the literature catalog that have *Gaia* source IDs to be predicted by a single model, setting it to 1400 maximal members. We note here that the prior assumption restrictions only apply to single models, meaning the model ensemble, as a final classifier can exceed individual or multiple prior assumption limits.

#### 3.2.2. Contamination fraction

Secondly, we constrained the contamination fraction of predicted sources across models. The contamination fraction is determined via the 3D velocity distribution of $\rho$ Oph candidate sources. Precisely, we first modeled the 3D velocity distribution of the training samples as a single Gaussian distribution. The mean and covariance matrix were determined by maximizing the likelihood of the training data. Subsequently, we defined the contamination as the fraction of sources outside the $3\sigma$ (99.7%) range of the training set. In practice, we observe very few radial velocities in the predicted set for a single model, and, therefore, the contamination fraction assumption has a minor effect for removing single models. This effect is highlighted in Fig. C.1, where we see an almost constant and maximal number of models adhering to the contamination rule for various maximal values. Since the influence is small across such a large range, we set it to a value of 15%.

#### 3.2.3. Estimated extent and systematic shift

Lastly, we want to constrain the extent of predicted $\rho$ Oph members in position and proper motion space. This was done by measuring the dispersion and systematic shift between training and predicted member distributions. We characterize the dispersion in position and proper motion space by the mean deviation of its member stars to their centroid. The prior assumption corresponds to a constraint on the ratio between the average predicted deviation to the average training deviation. For further details, we refer to Appendix B in Ratzenböck et al. (2020). In the case of $\rho$ Oph, we cannot give a concrete estimate on the expected extent of unknown members in position and proper motion space. Instead, we motivate a range of maximal values. We postulated a constraint on the parameter to be within 1, which constrains the predicted extent to the training set extent, and 2, where models can have twice the dispersion of the training set. We explicitly separated the positional from the proper motion axes since both dispersion measures have physically different meanings, and we might want to restrict one more than the other.

To avoid systematic shifts of the predicted to the training set distribution, we constrained the distance between the centroids of the training and predicted sources. We measured the centroid distance in terms of the mean deviation of the training set sources. A value of one would correspond to a centroid shift with a distance of one mean deviation from the training centroid. Again, finding a precise value is not straightforward, as the value cannot be properly inferred for the unknown $\rho$ Oph population. Therefore, we limited the maximum shift parameter to a range between 0.1 and 0.7, which we consider already a quite large systematic deviation from the training set.

### 3.3. Building the $\rho$ Oph classifier

We subsequently searched for model ensembles within these three parameters, the mean deviation in position, proper motions, and the maximal systematic shift, while keeping the other two prior assumptions, the maximum number of predicted sources, and the maximum contamination fraction, fixed. As stated in Sect. 3.2.1, a prior assumption tuple corresponds to a model ensemble that adheres to the respective beliefs. For each of these ensembles, we determined a stability threshold by minimizing the Kullback–Leibler (KL) divergence (Kullback & Leibler 1951) between the 3D velocity distributions of training and predicted $\rho$ Oph members (see Appendix E for more details). We randomly selected 100 prior assumption tuples within their respective range, resulting in 100 model ensembles with a corresponding stability threshold.

To select single or multiple suitable classifiers from this space of model ensembles, we considered the following. We aimed to maximize the number of predicted $\rho$ Oph sources while minimizing the number of contaminants in our final prediction set. Thus, we studied the distribution of the number of predicted sources over the contamination fraction across the 100 model ensembles. The contamination fraction is determined via the ratio of predicted sources outside the $3\sigma$ range of the training velocity distribution. The distribution of the 100 randomly sampled model ensembles can be seen in Fig. D.1. We observed a clear trend for high-contamination models, which tend to have larger velocity dispersion and interestingly a rather low systematic shift to "good" models. This sample of low-contamination models were identifying possibly new $\rho$ Oph members in a nonsymmetric region around the training set. To construct the final classifier, we combined the predictions of the 90 models with the lowest contamination fraction of <28%[1], corresponding to the two left-most columns of models in the top row of Fig. D.1. Finally, we determined a stability threshold for the final ensemble following the procedure outlined in Appendix E. Doing so, we obtained a stability threshold of 4%. To properly validate the final classifier, we had to consider the previously untouched information, the distribution of sources in the Hertzsprung-Russell Diagram (HRD). In order for the predicted sources to be actual members of the $\rho$ Oph population, they must follow the same isochrone as the training set. Therefore, we determined the residuals of predicted sources to the best fitting isochrone on the training set, where we obtained an age of

[1] The contamination fraction is determined without any quality filters applied.

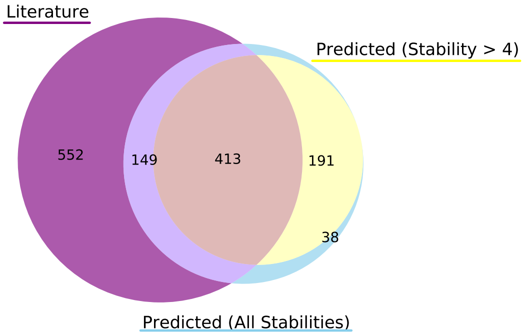


**Fig. 1.** Venn diagram depicting the amount of sources in the literature sample, the predicted sample, and the amount of sources both of them have in common. In total, 791 sources were predicted by the algorithm. Of these, 229 are new sources, with 191 having a stability >4. 562 of the total predicted sources are already among the 1114 literature sources, 413 of those having a stability >4. 552 literature sources were not predicted by the algorithm.

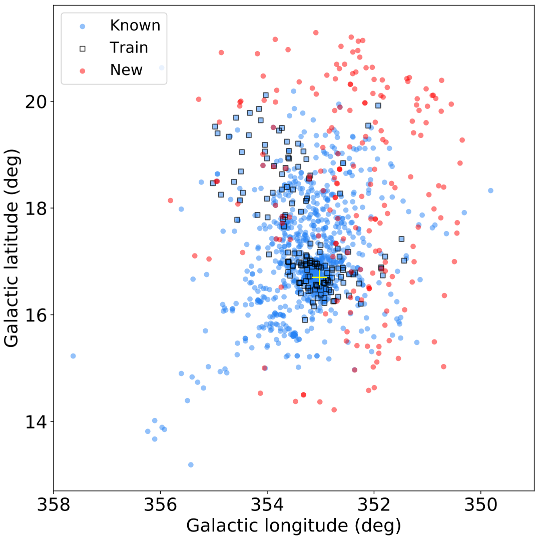


**Fig. 2.** Distribution of $\rho$ Oph sources in galactic coordinates. The known sources from the literature are in blue, while new sources are in red. Sources from the training set are represented by black squares. The approximate location of the extinction peak is marked by a yellow cross.

about 5 Myr, and compared them to the training set residuals. In Fig. F.1, the standard deviation of the training set residuals and predicted residuals can be seen, highlighting an almost perfect agreement with the training data across the full stability range.

## 4. Results

In this section we present the results of the algorithm. Sources from the literature are labeled as "Known" while the new sources are labeled as "New." The following plots in this section show the known sources in blue, the new sources in red, and a control sample in gray, labeled as "Control," which serves as a comparison. The control sample was selected in a relatively dust-free region to the Galactic west of $\rho$ Oph at the same galactic latitude, within 346° $\leq l \leq$ 349° and 15° $\leq b \leq$ 18°.

### 4.1. Predicted sources

A total of 791 sources in the *Gaia* EDR3 catalog were predicted by the algorithm as belonging to the $\rho$ Oph region, based on the properties of the training set. The predicted sources include a total of 229 new sources that are not in the $\rho$ Oph literature catalog. A total of 562 of the predicted sources are already part of the literature sample of 1114 known sources, meaning that 50.4% of the literature sources were recovered by the algorithm.

Only the sources with stability >4%, namely 191 of the new sources, are considered in the following results. These new sources together with the known ones result in 1305 total sources in the $\rho$ Oph region, while when excluding impostor sources we end up with 1288 high probability members. In our final catalog, we also include the new sources predicted by the algorithm with a stability <4%, resulting in a table of 1343 total sources. Figure 1 visualizes the amount of shared sources in the literature and the prediction set in a Venn diagram, showing sources with a prediction stability >4% and all stabilities. More information on the stability can be found in Appendix E. An overview of the final numbers of sources per (sub)sample is given in Table 3. A column overview of the final master catalog of the $\rho$ Oph young stellar members is presented in Appendix H.

A probable reason for the relatively small overlap in Fig. 1 (the algorithm only predicts 50.4% of the known sources from the literature) is the fact that many of the literature sources were obtained through IR surveys since embedded stars in $\rho$ Oph cannot be detected at optical wavelengths. Furthermore, some sources in the literature are impostors, as described in Sect. 2.2. It is also important to note that only 675 literature sources (60.6%) have *Gaia* EDR3 IDs. Therefore, the algorithm has effectively recovered 83.3% of literature sources that are in *Gaia* EDR3. *Gaia* is an optical telescope, hence it is insensitive to high extinction sources in the L1688 dense clump, where the peak of the surface density of YSOs in the cloud complex is located (e.g., Ortiz-León et al. 2017; Ducourant et al. 2017). Sources not visible at optical wavelengths cannot be predicted by the algorithm.

### 4.2. Astrometric properties

Figure 2 shows the distribution of the $\rho$ Oph sources in galactic coordinates with the known sources in blue and the new ones in red. The 150 sources in the training set, labeled as "Train," are included as unfilled black squares for comparison. As can be seen in the figure, the new sources are more dispersed, with many of them being shifted toward the Galactic north, west, and south of the known sources. Hardly any new sources were found near the core of the cloud and toward the Galactic east. The extinction peak of the L1688 cloud, marked by a yellow cross in the figure, lies at around $l \approx 353.0°$ and $b \approx 16.7°$ (Alves et al., in prep.). It is most likely responsible for the lack of new sources in the core since sources with a high optical extinction cannot be detected by *Gaia*. Furthermore, the core region is the most thoroughly studied part of $\rho$ Oph by previous surveys, thus it is unsurprising that few new sources were found near the core region.

Figure 3 shows a histogram of the distances to the $\rho$ Oph sources, which were determined through the inverse of their

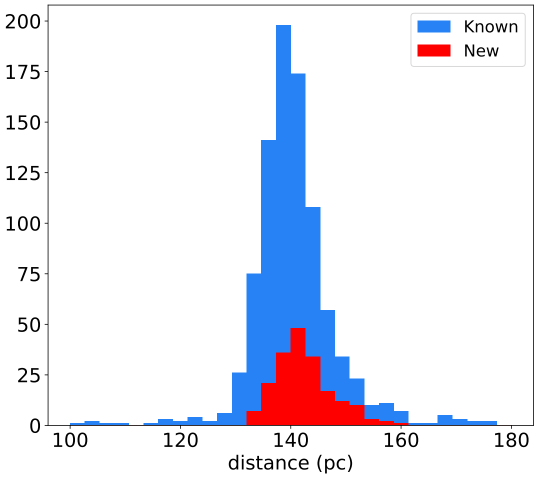


**Fig. 3.** Histogram of distances to the $\rho$ Oph sources. The distribution of the known sources from the literature is in blue, and the new sources are in red.

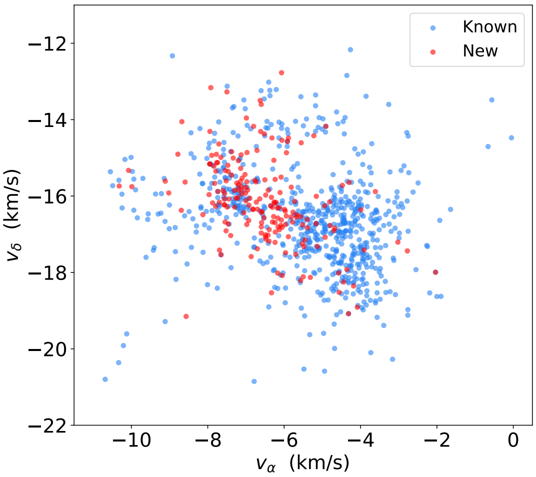


**Fig. 4.** Tangential velocities of the $\rho$ Oph sources. The known sources are shown in blue, and the new sources are shown in red.

parallaxes. Most of the sources are clustered around a mean distance of approximately 140 pc (see Table G.1), which agrees well with the literature value of around 139 pc (Zucker et al. 2020). In general, the average astrometric properties of the known and new sources are very similar and overlap within $\pm 1\sigma$ (see Table G.1), further confirming that they belong to the same region.

Figure 4 shows the tangential velocity distribution of the $\rho$ Oph sources. The impostor sources (see Sect. 2.2) from the literature are not included in the diagram, to avoid the influence of outliers. Although the distribution of the new sources shows an overlap with the bulk of the known sources around $-6 < v_\alpha < -3$ and $-19 < v_\delta < -16$, a large part of the population is shifted toward more negative values of $v_\alpha$ and less negative $v_\delta$, hinting at more than a single population. These two separate dynamical populations can already be recognized in the known sources alone, while the new sources further add to the second dynamical subgroup around $-10 < v_\alpha < -6$ and $-14 < v_\delta < -18$.

For further analysis of this distinct kinematic subgroup, we determined the proper motions in Galactic coordinates and the angles between the Galactic proper motion vectors ($\boldsymbol{\mu}_{l,b}$) and the $l$-axis ($\theta_{l,\mathrm{HEL}}$) in the heliocentric reference frame, and added these values to our table in a new column for all the sources with proper motion measurements. Analyzing these angles in a histogram reveals the two dynamically different populations as two distinct peaks, as can be seen in the histogram in Fig. 5 in the bottom left image. To disentangle these two populations, we use the angle distribution as a visual aid and apply a cut of $\theta_{l,\mathrm{HEL}} < 200°$, resulting in a subgroup of 304 sources for the second population when excluding 2 impostor sources. Using the proper motion angles relative to the local standard of rest ($\theta_{l,\mathrm{LSR}}$) produces a similar result, as shown in the bottom right image of Fig. 5. However, using this method, the separation between the two populations is not as evident, indicating that there might be more than two dynamical populations. For simplicity, we considered only two populations in our work and refer to future studies on Sco-Cen (Ratzenböck et al., in prep.) for a more detailed analysis.

Figure 5 highlights the influence of the Sun's reflex motion on the heliocentric proper motions. The top panels show the direction of motion using the heliocentric velocities (left) and the direction of motion when correcting for the Sun's motion (right), showing velocities relative to the local standard of rest (LSR). The latter show a less clear separation between the two populations. To separate the two populations, we used the heliocentric proper motion to avoid injecting in the final selection uncertainties related to the Sun's motion (Schönrich et al. 2010). In any case, making a selection of the populations in $\theta_{l,\mathrm{LSR}}$ would not change the result significantly.

For further discussion, this second dynamically distinct population shall be referred to as "Pop 2," while the remaining shall be referred as "Pop 1" sources, after excluding impostors (see Sect. 2.2). We define the sources in Pop 1 to be all sources from our $\rho$ Oph catalog excluding impostors and Pop 2 sources. This population comprises the clusters of young stars around the $\rho$ Ophiuchi star and the main Ophiuchus clouds (L1688, L1689, L1709). Concluding, we identify 304 sources in Pop 2 and 1022 in Pop 1 when including sources of all stabilities. When applying a cut at stability >4% for the new sources, we are left with 296 sources in Pop 2 and 993 sources in Pop 1 (see Table 3).

The 304 sources in Pop 2 coincide with the sources whose tangential velocities create the second dynamical structure in Fig. 4. In other words, the two subpopulations seen in this figure and the bimodal angle distribution consist of the same stars. 115 of these 304 sources (37.8%) are new sources identified by the algorithm. Further examination of this subgroup reveals that unlike Pop 1, Pop 2 sources are mostly dispersed and are distributed relatively evenly all around the core of the cloud (see Fig. 11). Their distances exhibit a similar distribution to the other $\rho$ Oph sources, which shows that the two populations occupy approximately the same 3D volume.

Table 2 shows the average values of the distances, proper motions, radial velocities, Galactic Cartesian positions $X, Y, Z$ and Galactic Cartesian velocities $U, V, W$, and the standard deviations of these parameters for the two populations (Pop 1 and Pop 2) in the $\rho$ Oph region. The average 3D positions of the two

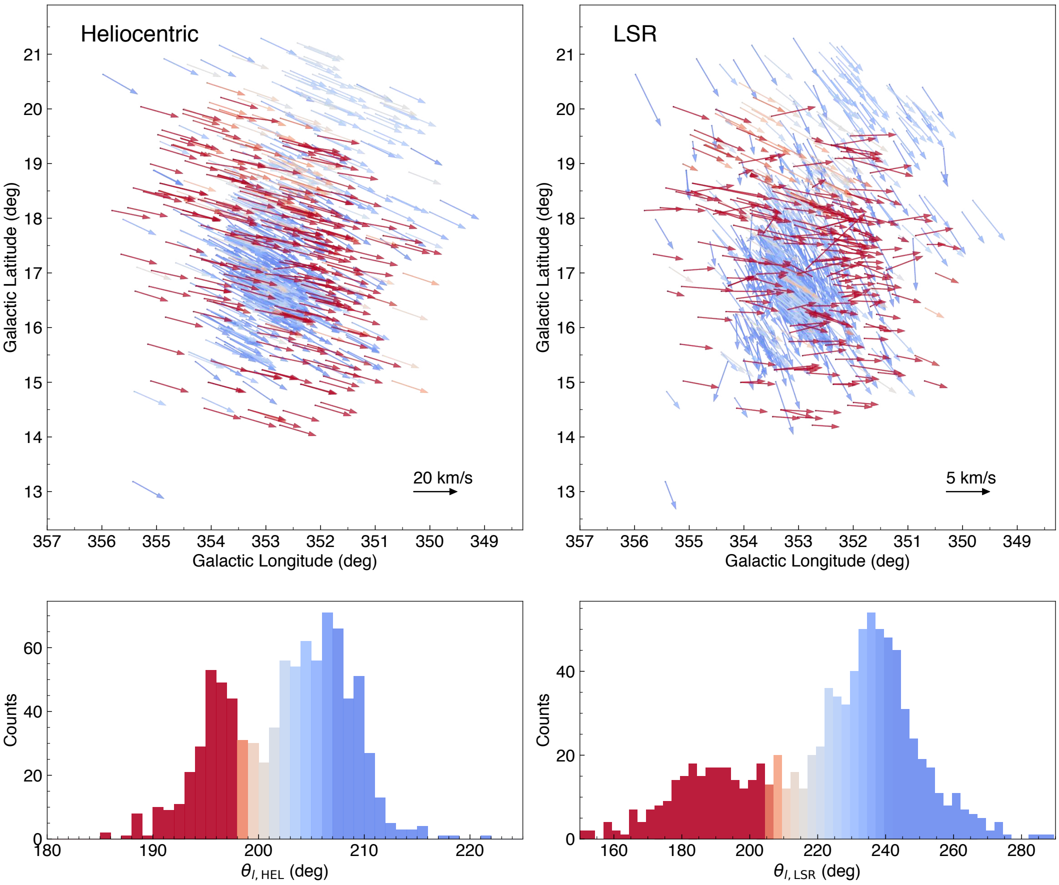


**Fig. 5.** Analysis of the two populations in $\rho$ Oph based on their proper motion. *Top row*: galactic distribution of the known and new $\rho$ Oph members, including all new sources (without stability cut), while impostors are excluded. Arrows represent the tangential velocity vectors, color-coded for the angle between the vectors and the $l$-axis ($\theta_{l,\mathrm{HEL}}$ and $\theta_{l,\mathrm{LSR}}$ in *left and right panel*, respectively). *Left panel*: heliocentric tangential velocity vectors ($\boldsymbol{v}_{\mathrm{HEL}}$), as derived from *Gaia* EDR3 parameters ($v_l$, $v_b$), *right panel*: tangential velocity vectors relative to the local standard of rest ($\boldsymbol{v}_{\mathrm{LSR}}$ based on $v_{l,\mathrm{LSR}}$, $v_{b,\mathrm{LSR}}$). The black arrows in the bottom right corners indicate the vector length for velocities of 20 km s$^{-1}$ and 5 km s$^{-1}$ for $\boldsymbol{v}_{\mathrm{HEL}}$ and $\boldsymbol{v}_{\mathrm{LSR}}$, respectively. These reference vectors have an angle of 180° relative to the $l$-axis. *Bottom row*: histograms showing the distributions of angles $\theta_{l,\mathrm{HEL}}$ and $\theta_{l,\mathrm{LSR}}$ for the sources as in the top panels. The bins in the left histogram have a width of 1° and in the right histogram of 2.5° since $\theta_{l,\mathrm{LSR}}$ covers a larger range of angles. The histograms are color-coded for the angles as in the top panels.

populations only exhibit small deviations, showing that they are not merely a 2D overlap, but mixed in all three spatial dimensions. As can be seen from the proper motions and tangential velocities in Table 2, the Pop 2 sources exhibit slightly different dynamical properties, which set them apart. Although the $U$ and $V$ velocities of the two populations hardly differ from each other, they occupy different regions in the $UVW$ velocity space because of the larger differences in $W$. The bimodality seen in Fig. 4 can also be seen in the $UVW$ space; however, only 55 sources (18.1%) from the second population have $UVW$ velocities. By computing the difference between the $UVW$ vectors of the two populations, we find that they are moving away from each other at about 4.1 km s$^{-1}$ and will no longer overlap in about 4 Myr.

Figure 6 shows the Galactic Cartesian coordinates of the known and new $\rho$ Oph sources for a visualization of their 3D distribution. The literature sources exhibit a more elongated distribution. Previously labeled impostors in Sect. 2.2 are marked with black crosses in Fig. 6. The elongation is most prominent along the line-of-sight, which is mostly caused by the larger errors in the parallax measurements compared to celestial coordinates, while some of the elongation could be caused by outliers. It can be seen in Fig. 6 that the new sources are rather distributed at the outskirts of the main cluster, indicating that they have been missed previously because they are more dispersed in space.

In Fig. 7 we show the same Galactic Cartesian representation as in Fig. 6, this time highlighting the 3D distribution of the Pop 1 and Pop 2 sources. It can be seen that the two populations largely occupy the same space, while there is a lack of Pop 2 sources at very high $Z$ when compared to Pop 1, best visible in the $X$ versus $Z$ panel. This distribution is consistent with the projected Galactic distribution in Fig. 11, where a similar lack of Pop 2 sources to the Galactic north can be seen.

**Table 2.** Average astrometric properties, including their standard deviations ($1\sigma$), for the two populations (Pop 1 and Pop 2) in the $\rho$ Oph region.

| Dimension | Pop 1 | Pop 2 |
|---|---|---|
| $\alpha$ (deg) | 246.4 ± 1.3 | 246.0 ± 1.2 |
| $\delta$ (deg) | −24.2 ± 0.8 | −23.9 ± 1.4 |
| $\varpi$ (mas) | 7.1 ± 0.4 | 7.1 ± 0.4 |
| $d$ (pc) | 140.4 ± 8.0 | 141.3 ± 7.9 |
| $\mu_\alpha^*$ (mas yr$^{-1}$) | −7.2 ± 2.1 | −11.4 ± 1.9 |
| $\mu_\delta$ (mas yr$^{-1}$) | −25.3 ± 2.3 | −23.4 ± 2.1 |
| $v_\alpha$ (km s$^{-1}$) | −4.7 ± 1.1 | −7.6 ± 1.2 |
| $v_\delta$ (km s$^{-1}$) | −17.0 ± 1.4 | −15.7 ± 1.2 |
| $v_\mathrm{r}$ (km s$^{-1}$) | −6.2 ± 4.5 | −3.9 ± 3.3 |
| $X$ (pc) | 132.8 ± 7.7 | 133.5 ± 7.5 |
| $Y$ (pc) | −16.2 ± 2.0 | −16.3 ± 2.9 |
| $Z$ (pc) | 42.3 ± 3.8 | 42.8 ± 3.9 |
| $U$ (km s$^{-1}$) | −5.5 ± 3.4 | −4.1 ± 3.0 |
| $V$ (km s$^{-1}$) | −15.1 ± 1.3 | −16.2 ± 1.5 |
| $W$ (km s$^{-1}$) | −9.4 ± 1.4 | −5.7 ± 1.5 |

### 4.3. Observational HRD

Figure 8 (left) shows an observational Hertzsprung–Russel Diagram (HRD) of the $\rho$ Oph sources with the known sources in blue and the new ones in red. To create the diagram we use the *Gaia* EDR3 passbands $G$ and $G_\mathrm{RP}$, for both the $\rho$ Oph and the control sample. Since *Gaia* EDR3 photometry is affected by systematic errors, corrections were applied to the $G$ band as described in Riello et al. (2021). Using the observed magnitudes $m_G$ in the $G$ band and the individual distances $d$ of the sources, we computed the absolute magnitudes $M_G$ in the $G$ band with $M_G = m_G + 5 - 5\log_{10} d$. Quality cuts as described in Appendix B were applied to the *Gaia* data of the $\rho$ Oph and control sample to include only high quality photometry and astrometry. Isochrones from the PARSEC models (Marigo et al. 2017) for *Gaia* EDR3 photometry are over-plotted in Fig. 8 for 1, 5, and 10 Myr. An extinction vector in the $V$ passband, labeled as $A_V$, is shown to visualize the direction and magnitude of extinction in this color–magnitude space using the reddening law from Cardelli et al. (1989) and O'Donnell (1994) provided by PARSEC. Two equal-mass-curves for sources with 0.09 $M_\odot$ and 1 $M_\odot$ are over-plotted.

The distribution of the known and new sources in the left panel of Fig. 8 overlap, indicating similar ages and luminosities, as also described in Sect. 3.3 and shown in Fig. F.1. This further confirms that they belong to the same region. Their distribution is consistent with earlier work of Luhman & Rieke (1999) and Esplin & Luhman (2020), who find ages of 0.3–6 Myr for $\rho$ Oph sources. Most of the new sources are low-mass stars, similar to the known sources, probably consisting mainly of $M$-type spectral classes or substellar objects.

In the right panel of Fig. 8 we show a similar observational HRD as in the left panel, showing the two dynamical populations in the $\rho$ Oph region. The first population (Pop 1), which comprises the clusters of young stars around the $\rho$ Ophiuchi star and the main Ophiuchus clouds (L1688, L1689, L1709), is shown in red, and the second dynamically distinct population (Pop 2) is shown in yellow. One can see that the second population appears to be slightly older than the first and aligns better with older isochrones. To determine the approximate age of the second population, we compute a least mean square fit to the data, as similarly done in Sect. 3.3, using the $G$, $BP$ and $RP$ passbands, to isochrones with solar metallicity from the PARSEC models (Bressan et al. 2012). We use only high-fidelity sources with stability >4, and quality cuts of ruwe <1.4 and astrometric_sigma5d_max <0.5 (for definitions of used *Gaia* parameters, see Table H.1). With this we obtain an approximate age of 10 Myr for the second population, which is older than the average age of about 5 Myr of the whole sample.

### 4.4. Analysis of infrared colors: Infrared-excess sources

The evolutionary stages of young stars can be estimated by using IR measurements, which reveal the presence of protoplanetary disks and envelopes around the pre-main-sequence stars. Disks and envelopes emit light in IR wavelengths due to their warm dust emission. Cross-matching our complete $\rho$ Oph catalog with data from WISE (Wright et al. 2010), in our case the AllWISE catalog, provides stars with the required IR photometry to analyze IR excesses. We note that not all sources are represented by WISE. The cross-match yielded 1110 sources with WISE data, which is 82.7% of our $\rho$ Oph sources. The $W1$, $W2$, and $W3$ passbands correspond to wavelengths of 3.4 µm, 4.6 µm, and 13 µm, respectively. To use only high quality measurements in our diagram, we only included sources above a specific signal-to-noise ratio (S/N). Sources had to fulfill $w1\mathrm{snr} > 10$, $w2\mathrm{snr} > 10$, and $w3\mathrm{snr} > 7$ for the $W1$, $W2$, and $W3$ passbands. This cut was applied to the $\rho$ Oph and the control sample, leaving 750 sources for the diagram, which is 55.8% of the total $\rho$ Oph sample.

Figure 9 shows a color-color diagram for $W1 - W2$ versus $W2 - W3$, with the known sources in blue and the new ones in red. The control sample is included in gray, and the sources of the second population (Pop 2) are marked by black symbols. The extinction vector in the $K_S$ passband, labeled as $A_K$, was determined using the reddening law for the $W1$, $W2$, and $W3$ passbands as in Meingast et al. (2018). A dashed line, serving as a rough estimate, separates two regions in the diagram, namely those with and without IR excess, as similarly done in Koenig & Leisawitz (2014). The functional form of the dashed line is given by $W1 - W2 = 1.05 - 0.8 \cdot (W2 - W3)$. Sources further to the top and right in the diagram exhibit an IR excess and are therefore most likely YSOs with envelopes or circumstellar disks, Class I or Class II, while Class II are similar to Classical T Tauri stars (Greene et al. 1994).

Most of the new sources have little or no IR excess, which could be the reason why they have not been identified in any previous IR surveys. Sources below and to the left of the dashed line in Fig. 9 are either Class III YSOs or main sequence stars. As Fig. 8 confirms that $\rho$ Oph consists mainly of young stars, this implies that the $\rho$ Oph sources below the line can only be Class III YSOs, which are associated with tenuous disks or bare photospheres, therefore creating no detectable infrared excess (Cánovas et al. 2019).

As can be seen from the red sources above and to the right of the dashed line in Fig. 9, we have found 28 new sources with IR excess, which are likely Class II candidates. This corresponds to a disk fraction of about 19.9% in the new sources, considering the displayed 141 new sources in the diagram. The known sources contain both Class I and Class II candidates. The fraction of sources with IR excess in the known population is roughly 48.6%, considering the 609 known sources within our WISE quality criteria, with 313 Class III YSOs and 296 YSOs with IR excess. Further analysis of the 28 new YSOs with IR excess reveals that they are located further away from the core of the cloud, which might explain why they have not been found

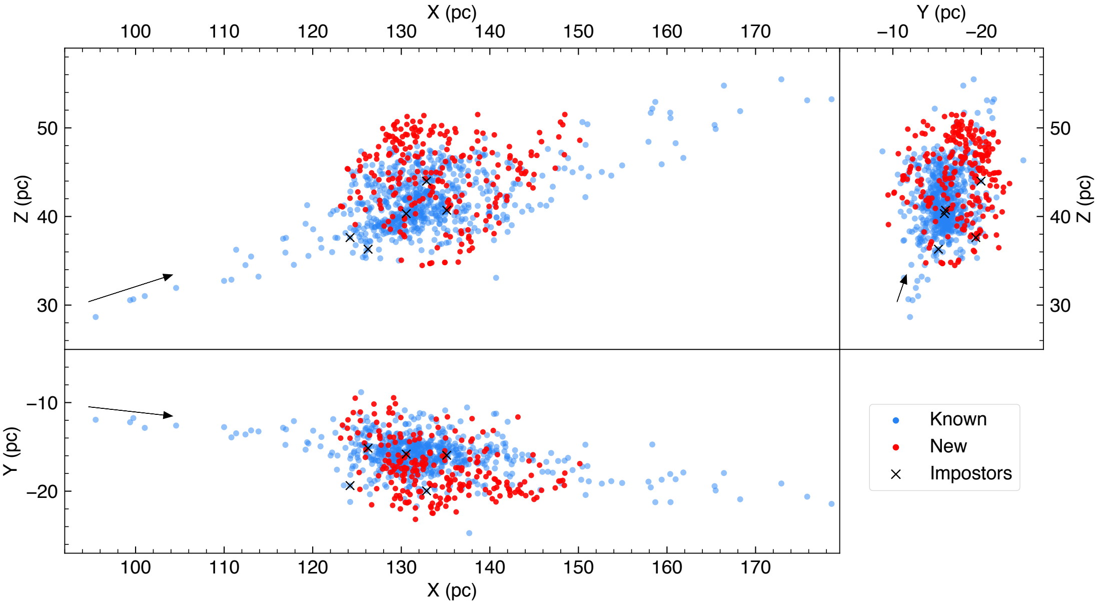


**Fig. 6.** Heliocentric Galactic Cartesian coordinates of the $\rho$ Oph sources. The known sources from the literature are marked with blue dots, the new sources with red dots, and impostor sources (Sect. 2.2) with black crosses (see legend). No quality or stability criteria were applied to the displayed sources. The black arrows in each panel indicate the line-of-sight from the Sun, pointing toward the star $\rho$ Oph and plotted from $d$ = 100–110 pc, which results in different arrow lengths due to projection effects. An interactive 3D version is available online.

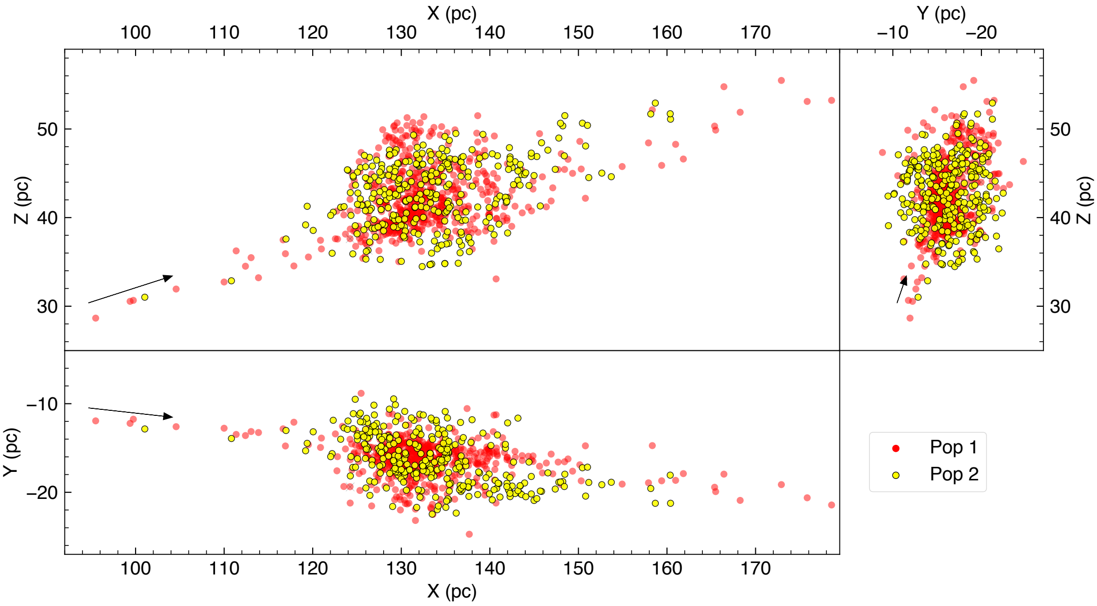


**Fig. 7.** Heliocentric Galactic Cartesian coordinates of the $\rho$ Oph sources, separated into Pop 1 (red) and Pop 2 (yellow). No quality or stability criteria were applied to the displayed sources. As in Fig. 6, the black arrows in each panel indicate the line-of-sight from the Sun. An interactive 3D version is available online.

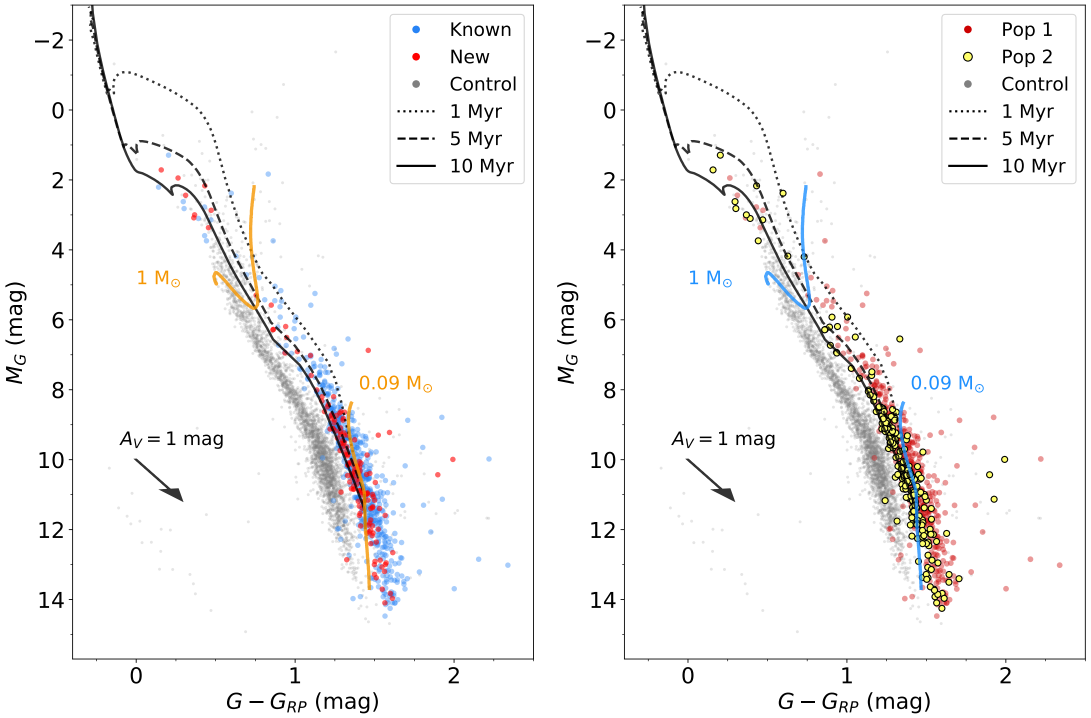


**Fig. 8.** Observational HRDs using the *Gaia G* and $G_{\rm RP}$ passbands, with corrections applied to the *G* passband. An extinction vector with $A_V = 1$ mag is shown as black arrow. The isochrones correspond to ages of 1, 5 and 10 Myr (see legend). The iso-mass lines (*left*: orange, *right*: blue) for 0.09 $M_\odot$ and 1 $M_\odot$ include stars with ages from 0.1 to 100 Myr. A control sample is shown in gray in the back. *Left*: comparing known (blue) and new (red) sources in $\rho$ Oph. *Right*: comparing the two populations in $\rho$ Oph, with Pop 1 in red and Pop 2 in yellow.

in any previous IR study of $\rho$ Oph, which focused mainly on the core region. 19 of the 28 new disk sources are from Pop 1, while 9 of them belong to Pop 2. The positions of the new IR excess sources in the HRD align well with most of the other new sources, showing very little scatter.

The distribution of Pop 2 members in Fig. 9 shows deviations from the average, with only 30 YSOs with IR excess and 154 Class III sources in the diagram, corresponding to a disk fraction of 16.3%, while Pop 1 has 293 YSOs with IR excess and 260 Class III sources in the diagram, resulting in a larger fraction of sources with IR excess of 53.0%. We conclude that Pop 2 contains overall more evolved stellar members and is likely at a later evolutionary stage compared to Pop 1 since the majority do not show any IR excess. This is consistent with the older age of Pop 2 seen in the optical HRD (Fig. 8, right panel). We note that the fraction of sources with IR excess could be overestimated for Pop 1 since even sources without proper motion values were counted to Pop 1, as defined in Sect. 4.2. Therefore, sources without measured astrometry are highly uncertain Pop 1 members since some could belong to Pop 2 or could even be galaxies, which could contaminate an IR-selected YSO sample.

We note that there are two known sources from Pop 2 that show untypically red colors compared to most other Pop 2 sources. The source with the largest $W1 - W2$ value has a *Gaia* source ID of `6049129800518036992`, and it is located near the core of the molecular cloud. Based on its color, it could be a flat-spectrum source or Class I (protostar). The proper motion direction indeed seems to fit to the Pop 2 sample; however, after checking the source in more detail, we find that the source has overall larger errors, indicating that its proper motion and distance, hence the tangential velocity, could be dominated by errors. Therefore, the Pop 2 membership of this source is uncertain, and it could be part of the younger Pop 1. This would reduce the disk percentage of Pop 2 down to 14.8%. The other Pop 2 source with a very significant IR excess, namely the one with the largest $W2 - W3$ value at the right of the diagram, has a *Gaia* source ID of `6050279163829546112`. The IR excess in W3 could indicate that the source is a transition disk.

We conclude that we have found 28 new YSOs with IR excess and 113 new Class III YSOs in Ophiuchus. The fraction of IR excess sources to Class III YSOs is around 0.25 in the new sources, 0.95 in the known sources, and around 0.76 in the entire population. Again, the fraction of known sources with IR excess could be slightly overestimated due to above mentioned reasons. An overview of the final numbers is given in Table 3. All sources with IR excess (Class I or Class II) according to Fig. 9, in total 324, are marked in our final catalog in the column "IR_excess" with a "1", while the remaining sources (Class III) are marked with a "0". Sources not included in Fig. 9 are not classified in this work.

Cross-matching our complete $\rho$ Oph sample with data from 2MASS (Skrutskie et al. 2006) provides us with further IR measurements in the $J$, $H$, and $K_S$ passbands, which correspond to wavelengths of 1.25 µm, 1.65 µm, and 2.17 µm, respectively.

**Table 3.** Final numbers of sources resulting from our $\rho$ Oph stellar member analysis.

| (Sub)sample | *N* |
|---|---|
| Known literature selected sources | 1114 |
| Literature selected sources with measured *Gaia* EDR3 parallax | 682 |
| Literature selected sources without impostors | 1097 |
| Impostor sources in the literature | 17 |
| All new sources without stability cut | 229 |
| New sources with stability cut | 191 |
| New sources with circumstellar disks (Class II) | 28 |
| New Class III sources | 113 |
| Total number of $\rho$ Oph sources without stability cut | 1343 |
| Total number of $\rho$ Oph sources with stability cut for new sources | 1305 |
| Total number of $\rho$ Oph sources without impostors | 1326 |
| Total number of $\rho$ Oph sources with stability cut for new sources and without impostors | 1288 |
| Pop 1 sources without stability cut | 1022 |
| Pop 1 sources with stability cut for new sources | 993 |
| Pop 2 sources without stability cut | 304 |
| Pop 2 sources with stability cut for new sources | 296 |

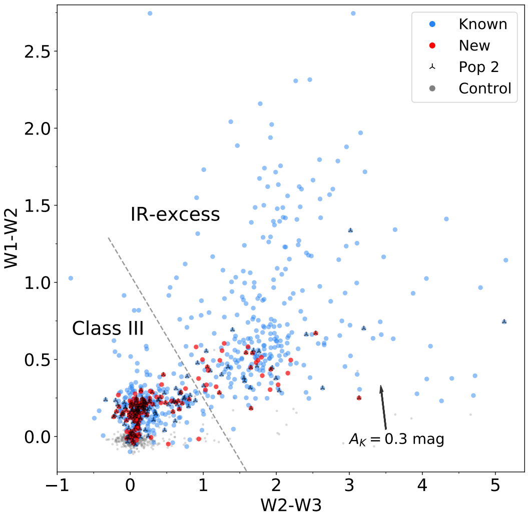


**Fig. 9.** Mid-infrared color-color diagram of the known $\rho$ Oph sources in blue and the new ones in red, including the control sample in gray, using the $W1$, $W2$, and $W3$ passbands from the WISE catalog. The sources comprising the second population (Pop 2) are marked by black symbols. An extinction vector in the $K_S$ passband, labeled as $A_K$, is also included. The sources above the dashed line with $W1-W2 > 1.05-0.8\cdot(W2-W3)$, are YSOs with IR excess due to a circumstellar disk (Class I or Class II), while those below the line are Class III YSOs.

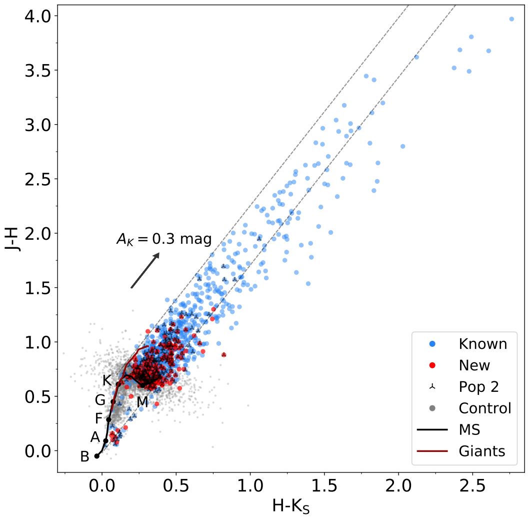


**Fig. 10.** Near-infrared color–color diagram of the know (blue) and new (red) $\rho$ Oph sources, including the control sample in gray and the sources from the second population (Pop 2) as black symbols, using the $J$, $H$ and $K_S$ passbands from 2MASS. The main sequence and the giant branches from Bessell & Brett (1988) are included in the diagram, as well as an extinction vector in the $K_S$ passband, labeled as $A_K$. The two parallel lines with the slope of the extinction vector enclose sources that are reddened mainly due to extinction in this color space.

Figure 10 shows a color–color diagram of $H - K_S$ versus $J - H$. In order to show only high quality measurements, we use the quality cuts *j_cmsig*, *h_cmsig*, *k_cmsig* $< 0.1$. The known, new, and control sources are in blue, red, and gray, respectively, while the sources from the second population (Pop 2) are marked by black symbols. The main sequence (MS) and giant branches are included in the diagram, as determined by Bessell & Brett (1988). The extinction vector in the $K_S$ passband, labeled as $A_K$, was determined using the reddening law for the $J$, $H$ and $K_S$ passbands by Meingast et al. (2018). Two parallel lines with the slope of the extinction vector were added to enclose reddened sources above the main sequence.

As can be seen from their positions in Fig. 10, most of the known and new sources are M stars. However, there are also several higher-mass stars among the new sources, as seen in the bottom left of the diagram. We find that these stars are located relatively far from the core of the cloud, which could explain why they have not been added as members in previous studies. Furthermore, the new sources are, on average, less extincted than the known ones, as would be expected, since they were selected based on the *Gaia* catalog.

## 5. Discussion

In this work we applied the classification strategy developed by Ratzenböck et al. (2020) to identify new members of the $\rho$ Oph region in *Gaia* EDR3. This method yielded 191 new high stability members with similar properties in position and motion to the 1114 known sources from the literature. From these results, we were able to create a master catalog of all known sources in $\rho$ Oph, including our new sources from *Gaia* EDR3. This so far most complete sample of $\rho$ Oph contains 1305 sources (or 1343 when also including the new sources with a stability <4).

### 5.1. The $\rho$ Oph region is a mixture of two young populations

The tangential velocity distribution of the final sample, presented in Fig. 4, reveals structure hinting at the presence of more than one population. The bimodal distribution of the proper motion angles presented in Fig. 5 further asserts the existence of two main populations in the surveyed area, which we call Pop 1 and Pop 2. What is discussed in the literature as the "$\rho$ Oph star-forming region" or "$\rho$ Oph core" is in fact a mixture of at least two populations, with similar but distinct dynamical properties and ages, occupying approximately the same 3D volume. The first (Pop 1), with ages 0.3–6 Myr (Luhman & Rieke 1999; Erickson et al. 2011; Esplin & Luhman 2020), as confirmed in Fig. 8, comprises clusters of young stars around the $\rho$ Ophiuchi star and the main Ophiuchus clouds, namely L1688, L1689, L1709 (see Fig. 11). The second population (Pop 2) appears more dispersed in comparison and has an older age up to ~10 Myr, a disk fraction of ~16.3%, and 3D motions of $U, V, W = -5.5, -16.2, -5.7\,\mathrm{km\,s^{-1}}$. Given that the age, disk fraction, and 3D motion are similar to those of Upper Sco ($U, V, W = -5.1, -16.0, -7.2\,\mathrm{km\,s^{-1}}$, disk fraction ~20%, age ~10 Myr, Pecaut & Mamajek 2016; Luhman & Esplin 2020), it is possible that the 304 Pop 2 sources in the dispersed population originate from the much larger Upper Sco population toward the Galactic north. However, we note that the sources from Pop 2 appear to be cut off toward the Galactic north, as can be seen in Figs. 7 and 11. Considering that Upper Sco lies in the north of $\rho$ Oph, it seems unclear if Pop 2 really originates from there. Still, the proper motion of Pop 2 is essentially the same as the proper motion of Upper Sco (Luhman & Esplin 2020), making it highly unlikely that Pop 2 is not associated with Upper Sco (same age, distance, and motions). More likely, because the training set consists of 77.3% Pop 1 sources, it is possible that this bias caused the algorithm to find fewer Pop 2 sources, causing the apparent cutoff. This will be further examined in future work (Ratzenböck et al., in prep.).

The clear kinematic difference between these two populations, only detectable because of the unprecedented accuracy of *Gaia* EDR3, is the main finding of our study as it sheds light on the genesis of the $\rho$ Oph star-forming region. The proper motion distribution found in Fig. 5, in combination with RVs, translates into a 3D space motion difference between the two populations of about 4.1 km s$^{-1}$. This relative space motion indicates that the regions are moving away from each other and could imply that the origin of the $\rho$ Oph star-forming region is connected to that of the Upper Sco population. A study of the space motion of the two populations is called for as it will give insights on the origin of the different motions.

The closest active star formation region to Earth, the $\rho$ Oph region, remains a natural laboratory for star formation studies, from core formation and collapse to disk formation and evolution into planets. Our work demonstrates how the unprecedented astrometric precision of *Gaia* is revealing the fine dynamical structure of this nearest star-forming regions.

### 5.2. Multiple young populations in star-forming regions

Our finding in this paper of a mixed population in $\rho$ Oph is similar to the discovery of the foreground population in front of the Orion Nebula (Alves & Bouy 2012; Bouy et al. 2014; Chen et al. 2020). Unfortunately, two of the closest benchmark star formation regions to Earth, the $\rho$ Oph region and the Orion Nebula Cluster, are now known to contain multiple young populations, either in projection or intermingled, which complicates the extraction of star formation observables. These two cases are unlikely the exception. Mixed populations are to be expected, for example, in triggered star formation as a previous generation compresses interstellar gas into a new generation of stars. Characterizing the existence of multiple populations in nearby star formation regions is critical because it directly affects the fundamental star formation observables, such as star formation history, rate, efficiency, and the initial mass function (IMF). Looking forward, multiple populations should be looked for in other nearby star-forming regions, and for at least $\rho$ Oph and the Orion Nebula Cluster, they need to be disentangled for a precise description of the basic star formation observables.

### 5.3. Caveats

Some of the literature sources are located off from the center of the cloud, in particular the ones that seem to trace the B44 filament (L1689, L1712, L1759), away from the center of the distribution and toward the lower Galactic east in Fig. 2. These sources might be too far from the cluster center to be considered by the algorithm, since the training set is only located near the center of the distribution (Fig. 2). Still, the sources seen in projection onto B44 are also located at the edge of the proper motion distribution, making them even less likely to be predicted. However, since there are only a handful of sources located so far off, this suggests that the algorithm is not missing a significant number of sources toward the filaments B44 and B45.

### 5.4. Comparison with previous work using Gaia data

Cánovas et al. (2019) applied several clustering algorithms (`DBSCAN,OPTICS,HDBSCAN`) to identify new sources in the $\rho$ Oph region using the *Gaia* DR2 catalog. We have found sources that were not identified as potential members by Cánovas et al. (2019), despite also running our search algorithm on the *Gaia* DR2 catalog before the availability of *Gaia* EDR3. Our search in only *Gaia* DR2 yielded around 150 new members, depending on how strictly we set our prior assumptions. Finding so many new YSOs in the same data set suggest that our approach is an effective tool for searching for new members of co-moving stellar structures.

Esplin & Luhman (2020) used *Gaia* DR2 data and derived proper motions with multi-epoch data from the *Spitzer* Space Telescope to find 155 new young stars, 102 of these associated with the Ophiuchus clouds and 47 with Upper Sco. Unlike our study, Esplin & Luhman (2020) did not use multivariate classification techniques to identify new sources, so we attribute the

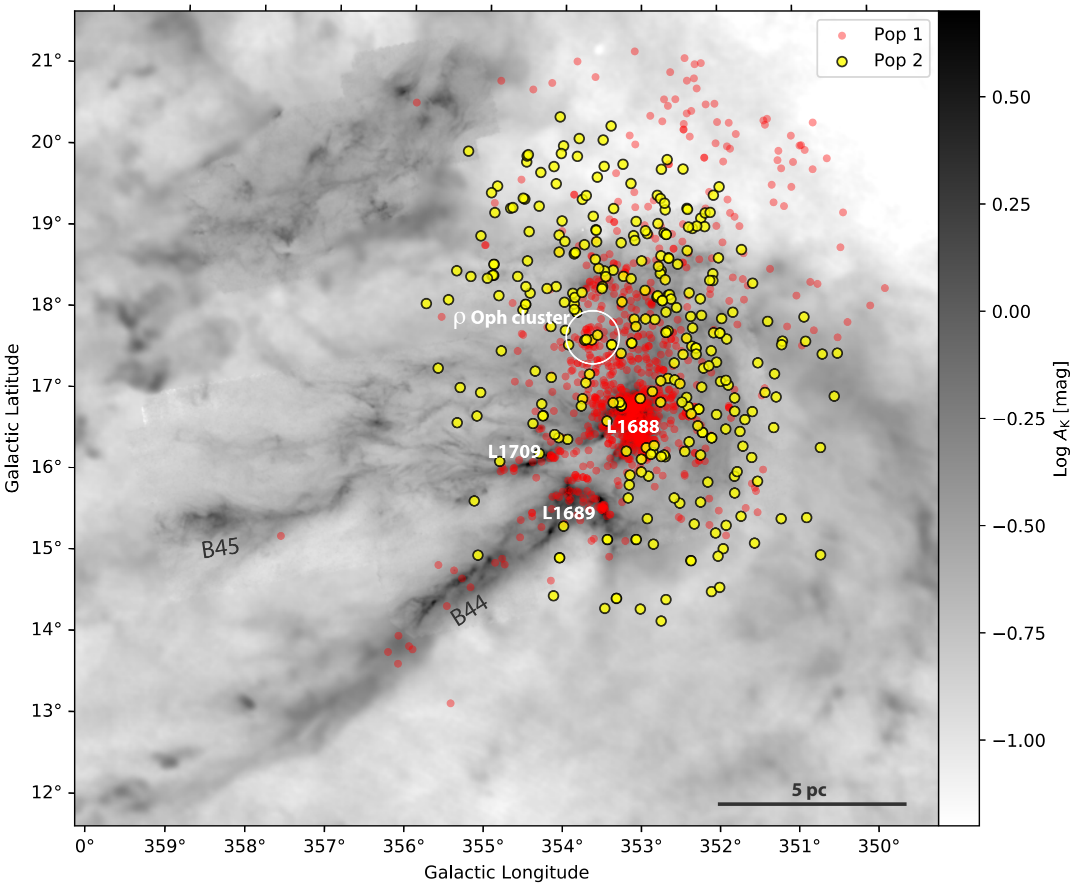


**Fig. 11.** Spatial distribution of the two dynamical populations in $\rho$ Oph in red and yellow circles. The $\rho$ Oph cluster, centered on the $\rho$ Ophiuchi star, is marked by a white open circle. The actively star-forming clouds, L1688, L1689, and L1709, are also marked. Impostors (see Sect. 2.2) are not included in this figure, whereas low stability sources are. The background grayscale is a column density map of Ophiuchus made with *Herschel*, *Planck*, and 2MASS data (Alves et al., in prep.).

discovery of the 191 new YSOs over their search to tailored classification techniques as the one described in this paper, which are powerful tools to disentangle stellar populations in the high-precision *Gaia*-era data.

Concluding, the algorithm from Ratzenböck et al. (2020) has shown to be an effective method for identifying stars belonging to a particular population, based on the properties of a subsample of known sources. The method was able to identify 191 new optically visible sources in $\rho$ Oph, providing more information on the optically revealed population of the region. Therefore, we conclude that our method is a useful tool suitable for similar research in the future.

## 6. Conclusions

The main results from this work can be summarized as follows:

1. We searched the literature to construct a catalog of 1114 known YSOs toward the $\rho$ Ophiuchi region. We cross-match this catalog with the *Gaia* EDR3, *Gaia*-ESO, and APOGEE-2 surveys and use it to feed a classification algorithm designed to find new, co-moving population candidates in *Gaia* EDR3 using a training set of 150 sources.
2. We found 191 new YSO candidates in *Gaia* EDR3 belonging to the $\rho$ Ophiuchi region (229 new YSOs including low-fidelity members). The distribution of the new sources in an HR-diagram is very similar to previously known young stars in the region, validating our selection.
3. The new sources appear to be mainly Class III M stars and substellar objects, and they are generally less extincted than the known members.
4. We found 28 new sources with excess IR emission suggesting the presence of disks.
5. A proper motion analysis of the $\rho$ Ophiuchi region reveals the presence of two main populations: the first population (Pop 1) of 1022 sources comprises clusters of young stars around the $\rho$ Ophiuchi star and the main Ophiuchus clouds (L1688, L1689, L1709), while the second population (Pop 2) of 304 sources is slightly older and more dispersed, with a similar but distinct proper motion from the first. Both populations occupy approximately the same 3D volume. The

second population's age and proper motion suggest that it may have originated from the Upper Sco population.

6. The two populations are moving away from each other at about 4.1 km s$^{-1}$, and will no longer be overlapping in about 4 Myr.
7. Future studies of this benchmark region should treat these two populations separately or risk biasing the star formation observables, such as star formation history, rate, efficiency, or the IMF.
8. The algorithm used in this paper (OCSVM, Ratzenböck et al. 2020) has proven to be an effective method for identifying stars belonging to a particular population, based on the properties of a subsample of known sources.

*Acknowledgements.* We thank the anonymous referee for their thorough and constructive report that has helped us clarify some points and has improved the paper. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 851435). This project has received funding from the Austrian Research Promotion Agency (FFG) under project number 873708. This research has used data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/web/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium), as well as data from the European Southern Observatory (ESO) survey *Gaia*-ESO (Gilmore et al. 2012) and the APO Galactic Evolution Experiment APOGEE (Majewski et al. 2017). This research has also made use of data products from the Two Micron All Sky Survey 2MASS (Skrutskie et al. 2006), which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation, the Wide-field Infrared Survey Explorer WISE (Wright et al. 2010), which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration. The astronomical database SIMBAD (Wenger et al. 2000) has also greatly contributed to this work. Furthermore, this research was made possible thanks to the use of TOPCAT (Taylor 2005), an interactive graphical viewer and editor for tabular data, and Python (https://www.python.org/), in particular Astropy (Astropy Collaboration 2018), a community-developed core Python package for Astronomy, NumPy (van der Walt et al. 2011) and Matplotlib (Hunter 2007), as well as the Aladin Sky Atlas, developed at CDS, Strasbourg Observatory (Bonnarel et al. 2000).

## References

ALMA Partnership (Fomalont, E. B., et al.) 2015, ApJ, 808, L1
Alves, J., & Bouy, H. 2012, A&AS, 547, A97
Andre, P., Ward-Thompson, D., & Barsony, M. 1993, ApJ, 406, 122
Astropy Collaboration (Price-Whelan, A. M., et al.) 2018, AJ, 156, 123
Bessell, M. S., & Brett, J. M. 1988, PASP, 100, 1134
Bonnarel, F., Fernique, P., Bienaymé, O., et al. 2000, A&AS, 143, 33
Bouy, H., Alves, J., Bertin, E., Sarro, L. M., & Barrado, D. 2014, A&AS, 564, A29
Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127
Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. 2000, SIGMOD Rec., 29, 93
Burnham, K. P., & Anderson, D. R. 2002, Model Selection and Multimodel Inference (Springer: New York)
Cánovas, H., Cantero, C., Cieza, L., et al. 2019, A&A, 626, A80
Cardelli, J. A., Clayton, G. C., & Mathis, J. S. 1989, ApJ, 345, 245
Chen, B., D'Onghia, E., Alves, J., & Adamo, A. 2020, Astron. Astrophys. Suppl. Ser., 643, A114
de Geus, E. J. 1992, A&A, 262, 258
Ducourant, C., Teixeira, R., Krone-Martins, A., et al. 2017, A&A, 597, A90
Dunham, M. M., Allen, L. E., Evans, N. J., II, et al. 2015, ApJS, 220, 11
Erickson, K. L., Wilking, B. A., Meyer, M. R., Robinson, J. G., & Stephenson, L. N. 2011, AJ, 142, 140
Esplin, T. L., & Luhman, K. L. 2020, AJ, 159, 282
Evans, N. J., Dunham, M. M., Jorgensen, J. K., et al. 2009, VizieR Online DataCatalog: J/ApJS/181/321
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Gaia Collaboration (Brown, A. G. A., et al.) 2018, A&A, 616, A1
Gaia Collaboration (Brown, A. G. A., et al.) 2021, A&A, 649, A1
Gilmore, G., Randich, S., Asplund, M., et al. 2012, The Messenger, 147, 25
Greene, T. P., & Meyer, M. R. 1995, ApJ, 450, 233
Greene, T. P., Wilking, B. A., Andre, P., Young, E. T., & Lada, C. J. 1994, ApJ, 434, 614
Großschedl, J. E., Alves, J., Meingast, S., et al. 2018, A&A, 619, A106
Großschedl, J. E., Alves, J., Meingast, S., & Herbst-Kiss, G. 2021, A&A, 647, A91
Haisch, K. E., J, Barsony, M., Greene, T. P., & Ressler, M. E. 2002, AJ, 124, 2841
Hunter, J. D. 2007, Comput. Sci. Eng., 9, 90
Kamdar, H., Conroy, C., Ting, Y.-S., et al. 2019, ApJ, 884, L42
Koenig, X. P., & Leisawitz, D. T. 2014, ApJ, 791, 131
Kullback, S., & Leibler, R. A. 1951, Ann. Math. Statist., 22, 79
Lada, C. J., & Wilking, B. A. 1984, ApJ, 287, 610
Lombardi, M., Lada, C. J., & Alves, J. 2008, A&A, 480, 785
Loren, R. B. 1989a, ApJ, 338, 902
Loren, R. B. 1989b, ApJ, 338, 925
Loren, R. B., & Wootten, A. 1986, ApJ, 306, 142
Luhman, K. L., & Esplin, T. L. 2020, AJ, 160, 44
Luhman, K. L., & Rieke, G. H. 1999, ApJ, 525, 440
Luri, X., Brown, A. G. A., Sarro, L., et al. 2018, A&AS, 616, 19
Lynds, B. T. 1962, ApJS, 7, 1
Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2017, AJ, 154, 94
Marigo, P., Girardi, L., Bressan, A., et al. 2017, ApJ, 835, 77
Meingast, S., Alves, J., & Lombardi, M. 2018, A&A, 614, A65
Meingast, S., Alves, J., & Fürnkranz, V. 2019, A&A, 622, L13
Meingast, S., Alves, J., & Rottensteiner, A. 2021, A&A, 645, A84
O'Donnell, J. E. 1994, ApJ, 422, 158
Ortiz-León, G. N., Loinard, L., Kounkel, M. A., et al. 2017, ApJ, 834, 141
Padgett, D. L., Rebull, L. M., Stapelfeldt, K. R., et al. 2008, ApJ, 672, 1013
Pecaut, M. J., & Mamajek, E. E. 2016, MNRAS, 461, 794
Perryman, M. A. C., Lindegren, L., Kovalevsky, J., et al. 1997, A&A, 500, 501
Petersen, K. B., & Pedersen, M. S. 2012, The Matrix Cookbook, version 20121115,
Pillitteri, I., Wolk, S. J., Chen, H. H., & Goodman, A. 2016, A&A, 592, A88
Ratzenböck, S., Meingast, S., Alves, J., Möller, T., & Bomze, I. 2020, A&A, 639, A64
Riello, M., De Angeli, F., Evans, D. W., et al. 2021, A&A, 649, A3
Rigliaco, E., Wilking, B., Meyer, M. R., et al. 2016, A&A, 588, A123
Schönrich, R., Binney, J., & Dehnen, W. 2010, MNRAS, 403, 1829
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Sullivan, T., Wilking, B. A., Greene, T. P., et al. 2019, AJ, 158, 41
Taylor, M. B. 2005, ASP Conf. Ser., 347, 29
van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Comput. Sci. Eng., 13, 22
Vrba, F. J. 1977, AJ, 82, 198
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9
Wilking, B. A., & Lada, C. J. 1983, ApJ, 274, 698
Wilking, B. A., Lada, C. J., & Young, E. T. 1989, ApJ, 340, 823
Wilking, B. A., Gagné, M., & Allen, L. E. 2008, in Star Formation in the $\rho$ Ophiuchi Molecular Cloud, ed. B. Reipurth, 5, 351
Wilkinson, S., Merín, B., & Riviere-Marichalar, P. 2018, A&AS, 618, A12
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Zucker, C., Speagle, J. S., Schlafly, E. F., et al. 2020, A&A, 633, A51

## Appendix A: Training set criteria

In this appendix, we describe the quality cuts determined for the training set, which were used in the classification algorithm to identify new members. We used the tangential velocities $v_\alpha$ and $v_\delta$ and their errors for determining the cuts of the training set. The tangential velocities were calculated through the parallaxes $\varpi$ and proper motions $\mu_\alpha^*$ and $\mu_\delta$ using the following formulas:

$$v_\alpha = 4.74047 \cdot \mu_\alpha^* / \varpi, \tag{A.1}$$

$$v_{\alpha\_\mathrm{err}} = 4.74047 \cdot \sqrt{\mu^{*2}_{\alpha\_\mathrm{err}} / \varpi^2 + \mu^{*2}_\alpha \cdot \varpi^2_\mathrm{err} / \varpi^4}, \tag{A.2}$$

$$v_\delta = 4.74047 \cdot \mu_\delta / \varpi, \tag{A.3}$$

$$v_{\delta\_\mathrm{err}} = 4.74047 \cdot \sqrt{\mu^{*2}_{\delta\_\mathrm{err}} / \varpi^2 + \mu^{*2}_\delta \cdot \varpi^2_\mathrm{err} / \varpi^4}. \tag{A.4}$$

The cuts for the training set were determined by using plots as a visual aid. Figure A.1 shows plots of various properties of the complete literature sample in blue and sources that satisfy our chosen quality cuts in orange. We applied the following quality cuts for constructing the training set:

$$100\,\mathrm{pc} < d < 180\,\mathrm{pc}, \tag{A.5}$$

$$\varpi_\mathrm{err} / \varpi < 0.2, \tag{A.6}$$

$$-15\,\mathrm{km\,s^{-1}} < v_\mathrm{r} < 5\,\mathrm{km\,s^{-1}}, \tag{A.7}$$

$$v_{\mathrm{r\_err}} < 3\,\mathrm{km\,s^{-1}}, \tag{A.8}$$

$$-12\,\mathrm{km\,s^{-1}} < v_\alpha < 2\,\mathrm{km\,s^{-1}}, \tag{A.9}$$

$$v_{\alpha\_\mathrm{err}} < 3\,\mathrm{km\,s^{-1}}, \tag{A.10}$$

$$-22\,\mathrm{km\,s^{-1}} < v_\delta < -11\,\mathrm{km\,s^{-1}}, \tag{A.11}$$

$$v_{\delta\_\mathrm{err}} < 3\,\mathrm{km\,s^{-1}}. \tag{A.12}$$

As the sources are located around a distance $d$ of 140 pc, we applied a symmetrical distance range of 100 to 180 pc for the training set. A relative error-to-value cut was also applied for the parallax $\varpi$. Radial velocities $v_\mathrm{r}$ are mostly around a value of $-5\,\mathrm{km\,s^{-1}}$, so we applied a symmetrical range of $-15$ to $5\,\mathrm{km\,s^{-1}}$. A relative error cut is not sensible for the radial velocities since many of them are close to zero, which could lead to losing sources that actually belong to $\rho$ Oph. Therefore, we applied an absolute radial velocity error cut. Since the errors of the tangential velocities $v_\alpha$ and $v_\delta$ are comparable to the radial velocity errors, similar cuts can be made in all three velocity directions. We applied the same absolute error cut to the tangential velocities, since several $v_\alpha$ values are also close to zero. These conditions select sources that do not deviate much from the average values of the chosen properties, creating a suitable selection for finding new sources with similar properties.

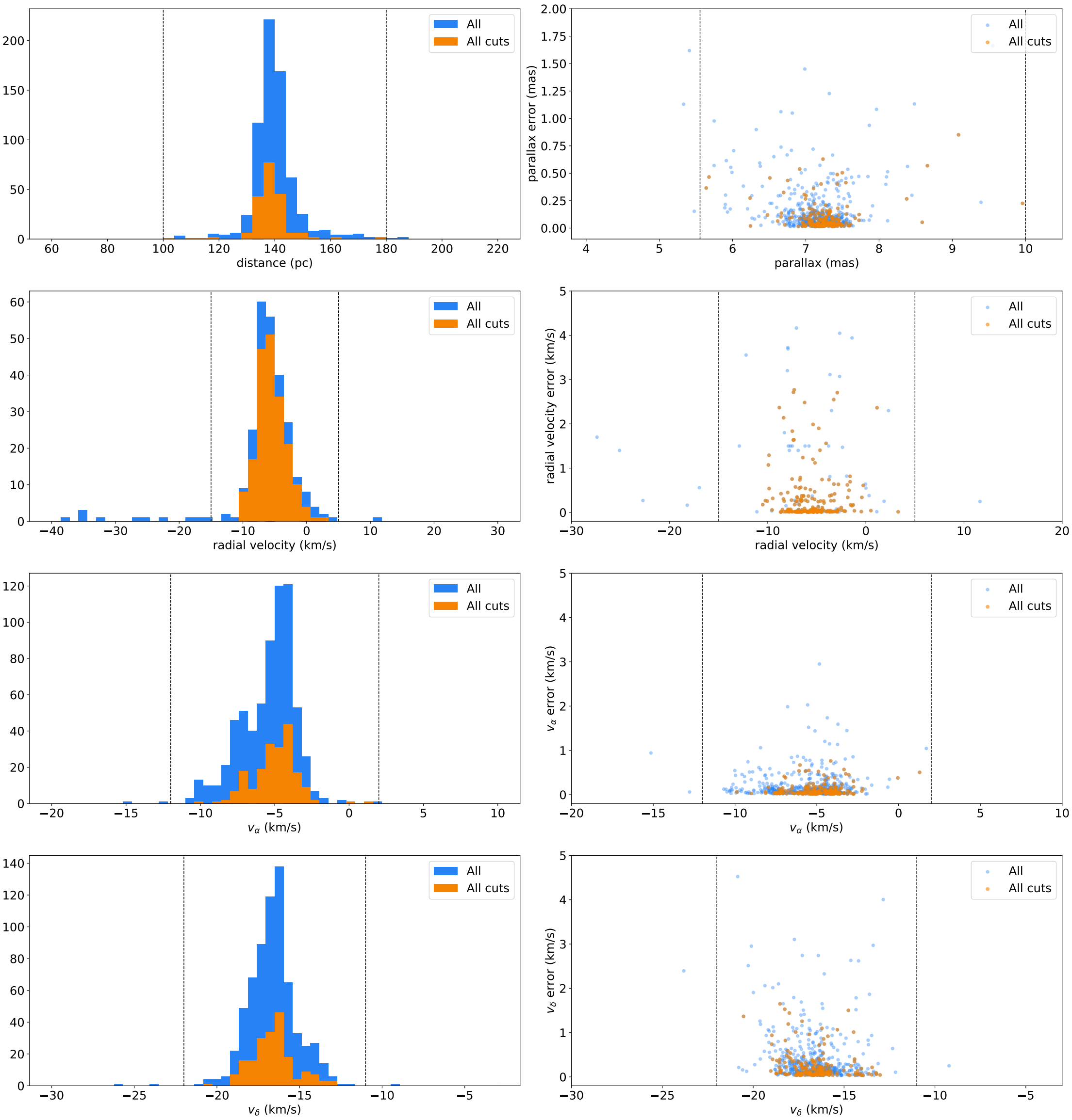


**Fig. A.1.** Various properties of the $\rho$ Oph literature sample shown in blue, while the sources that fulfill all of the quality cuts are shown in orange. These plots were used as a visual aid to determine the cuts for the training set.

## Appendix B: *Gaia* quality criteria

For the observational HRD in Fig. 8, we applied quality cuts to *Gaia* sources in order to reduce contamination by inferior data, similar to the cuts used in Großschedl et al. (2021). Further details on the *Gaia* parameters can be found on the official website of the mission:[2]. We applied the following quality criteria to *Gaia* sources:

$$\varpi_{\rm err}/\varpi < 0.2, \tag{B.1}$$

$$\text{ruwe} < 1.4, \tag{B.2}$$

$$G_{\rm err} < 0.05\,\text{mag}, \tag{B.3}$$

$$\text{visibility\_periods\_used} > 6, \tag{B.4}$$

$$\text{astrometric\_sigma5d\_max} < 1.4. \tag{B.5}$$

The $G_{\rm err}$ value is defined as:

$$G_{\rm err} = 1.0857 \cdot \text{phot\_g\_mean\_flux\_error}/\text{phot\_g\_mean\_flux}. \tag{B.6}$$

[2] https://gea.esac.esa.int/archive/documentation/index.html

## Appendix C: Contamination fraction constraint

Following Ratzenböck et al. (2020), we seek to constrain the contamination fraction of predicted sources across models. As discussed in Sect. 3, the contamination fraction is determined via the 3D velocity distribution of $\rho$ Oph candidate sources. However, for single models, we observed few sources that feature radial velocity measurements in the prediction set, which leads to a marginal effect of the contamination fraction prior assumption on the number of rejected models. This effect is highlighted in Fig. C.1, where we see that over 99% of models adhere to the contamination rule across various maximal threshold values. For each contamination threshold value we sampled 20 models where we have set the maximal number of samples to 800 and sampled the remaining prior assumptions within their respective ranges (see Sect. 3 for more details). The reported accepted model fraction constitutes a mean value across the 20 sampled prior assumption tuples. The standard deviation is negligibly small.

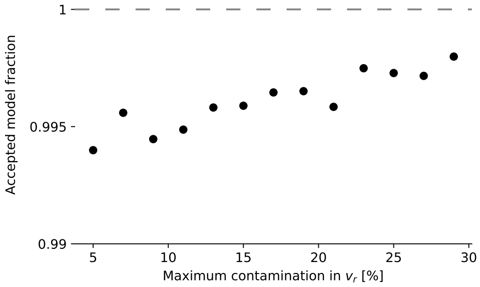


**Fig. C.1.** Accepted model fraction according to various maximal contamination requirements. The prior assumption value was varied between 5% and 30%. We found no significant impact of the contamination fraction restriction for individual models on the number of accepted models.

## Appendix D: Sampling in prior assumption space

Following the discussion in Sect. 3, we randomly sampled 100 prior assumption tuples within their respective range, which resulted in 100 model ensembles. In Fig. D.1 the distribution of the number of predicted sources and contamination fraction space of these ensemble classifiers is shown. The prior assumption space of the maximal positional extent (left column), the maximal velocity extent (middle column) and the maximal systematic shift (right column) was uniformly sampled within their respective ranges. We use color to encode the maximal prior assumption value in this space. On the bottom, the sampled prior assumption distributions for models showing minimal contamination (in purple) and the remaining models (in gray) can be seen. In models with high contamination, we observe a tendency to higher velocity dispersion but low systematic shifts. We observe that "good" models with lower contamination experience sometimes even a drastic systematic shit. This shift is due to the second population we uncovered.

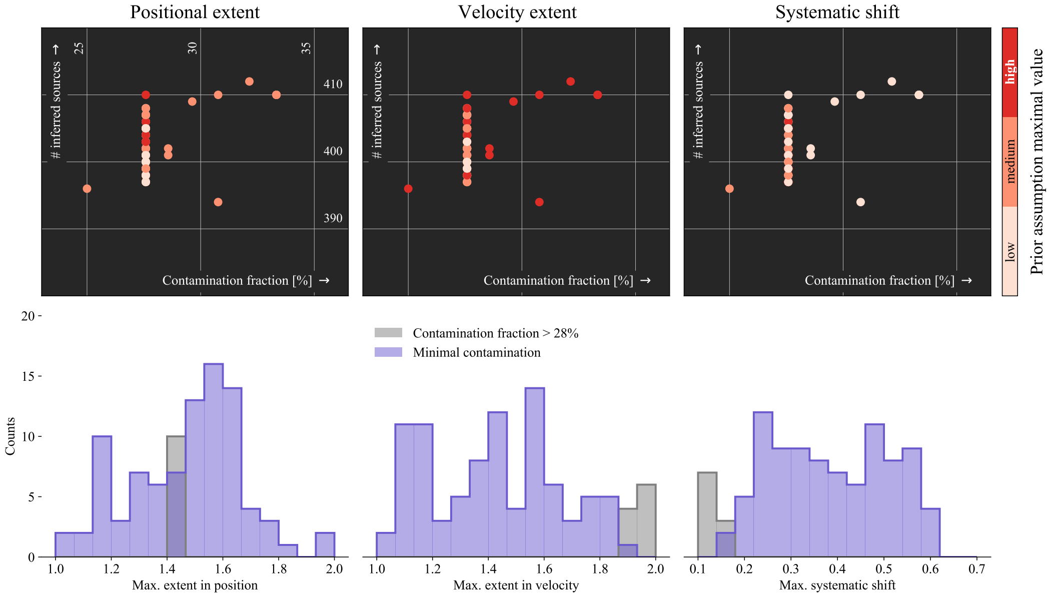


**Fig. D.1.** Distribution of the number of predicted sources and contamination fraction space of these ensemble classifiers. *Top*: distribution of 100 ensemble classifiers trained using various prior assumption constraints in the number of predicted sources and contamination fraction space. We have randomly sampled the prior assumption of the maximal positional extent (*left column*), the maximal velocity extent (*middle column*) and the maximal systematic shift (*right column*) within their respective ranges. The color highlights the maximal prior assumption value. *Bottom*: sampled prior assumption distributions for models showing a contamination of less than 0.28 (in purple) and remaining models (in gray). For models with a higher contamination fraction we observe a tendency to higher velocity dispersion and a small systematic shift.

## Appendix E: Stability

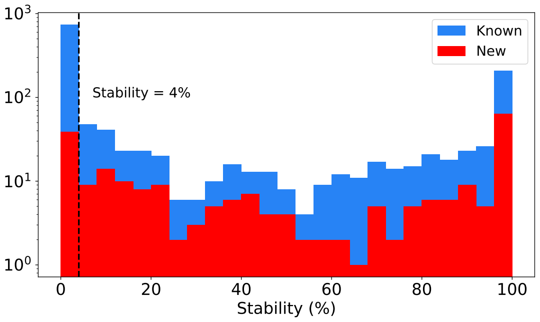


**Fig. E.1.** Stability (in percent) of the known and new sources, as determined by the OCSVM method.

We discuss the stability of the predicted sources as well as the stability cut we chose. Although the model selection process via a set of prior assumptions (see Sect. 3) removed a majority of unsuitable models, the lack of a clear objective function still leaves some contamination in our final prediction sample. To find a set of high-fidelity members, we studied the prediction frequency, or stability, of the predicted sources across the model ensemble. Figure E.1 shows a histogram of the stability of the known and new sources. Both of them show a relatively similar stability distribution. Many of the known sources from the literature are predicted with a stability of 0 because they are not in the *Gaia* EDR3 catalog.

As discussed in Ratzenböck et al. (2020), an appropriate stability threshold should reduce spurious sources while maximizing the number of legitimate cluster members. For this purpose, the authors studied the impact of the stability criterion on the Cartesian velocity dispersion and selected an optimal value by eye. Now we aimed to train multiple model ensembles under different prior assumptions and jointly attempt to characterize each model ensemble corresponding to a single prior belief tuple in terms of a contamination estimate and the number of identified points at their respective optimal stability thresholds (see Sect. 3). Therefore, we intend to automatically determine a threshold value for each model ensemble. To do so we considered the following. The distribution of predicted members and training members in 5D is by design very close and adheres to our prior assumptions, so we cannot infer an independent quality criterion from the prediction in 5D. However, since stars that are born together move together (Kamdar et al. 2019), we can, similarly to Ratzenböck et al. (2020), use the, albeit sparsely available, full 3D velocity information for determining the stability criterion.

To be co-moving, we postulate that the predicted sources with radial velocity information should be distributed as similarly as possible to the training set 3D velocities. To test this similarity, we modeled the 3D velocity data using a multivariate normal distribution. We determined the mean and covariance by maximizing the likelihood of the training data under the model. To estimate the difference between the trained and predicted sources, we used the Kullback–Leibler (KL) divergence (Kullback & Leibler 1951) $D_{\mathrm{KL}}(p \parallel q)$ where $q$ and $p$ both constitute probability distribution functions. The KL divergence of $p(\boldsymbol{x})$ from $q(\boldsymbol{x})$ of the continuous variable $\boldsymbol{x}$ is defined via

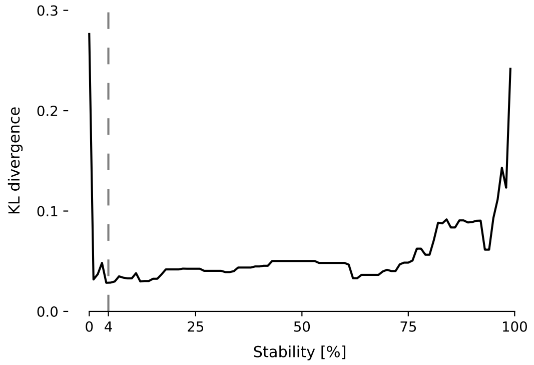


**Fig. E.2.** KL divergence between Cartesian velocity distributions of training and predicted source populations determined across various stability threshold values. We found an optimal threshold criterion of stability >4% for the final ensemble across various prior assumptions that produce minimal contamination (see Sect. 3 for a more detailed discussion).

$$D_{\mathrm{KL}}(p \,\|\, q) = \int_{-\infty}^{\infty} p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right). \tag{E.1}$$

It can be interpreted as the information content that is lost when the true distribution $p$ is substituted by an approximate distribution $q$ (Burnham & Anderson 2002). Here, $p$ represents our training set distribution, while the approximate distribution $q$ describes the distribution of predicted sources. To evaluate $D_{\mathrm{KL}}(p \,\|\, q)$, we modeled $q$, the velocity distribution of the derived members, assuming a single Gaussian. For two multivariate normal distributions, the KL divergence can be written analytically in the following form (Petersen & Pedersen 2012):

$$D_{\mathrm{KL}} = \frac{1}{2}\left[\log\frac{|\Sigma_q|}{|\Sigma_p|} - d + \mathrm{tr}(\Sigma_q^{-1}\Sigma_p) + (\mu_q - \mu_p)^T \Sigma_q^{-1}(\mu_q - \mu_p)\right]. \tag{E.2}$$

Here, $\mu$ and $\Sigma$ refer to the mean and covariance matrices of the multivariate normal distributions, respectively. The variable $d$ describes the number of dimensions, which is in this case $d = 3$. To find the optimal stability threshold we seek to minimize the KL divergence between the Cartesian velocity distribution of training and predicted sample populations, which is illustrated in Fig. E.2. We found an optimal threshold criterion of stability >4%. The stability is included in our final catalog shown in Table H.1.

## Appendix F: Validation of predicted sources in the HRD

As a final validation step, we compare the predicted source distribution to the training set distribution in the HRD. Since both populations should be coeval, we can characterize the HRD distribution by their deviation from the best fitting isochrone on the training set. In Fig. F.1, the standard deviation of residuals between the data and the 5 Myr isochronal curve is shown. We found no significant difference between the training set members and the predicted sources based on their HRD distributions.

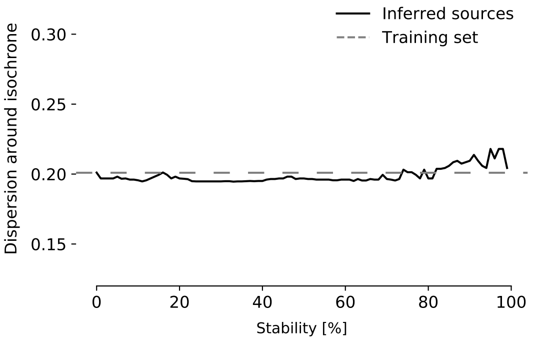


**Fig. F.1.** Comparison between the training and predicted (inferred) sources across the full stability range. The y-axis shows the standard deviation of residuals between the data and an isochrone of 5 Myr, describing the best fit to the training data. We find no significant difference between the training set members and the predicted sources based on their HRD distributions.

## Appendix G: Astrometric properties of known and new sources

**Table G.1.** Average astrometric properties of the known and new sources in $\rho$ Oph.

| Dimension | Known | New | All | $\Delta$ |
|---|---|---|---|---|
| $\alpha$ (deg) | 246.6 ± 1.0 | 245.2 ± 1.6 | 246.3 ± 1.3 | 1.4 |
| $\delta$ (deg) | −24.2 ± 0.8 | −23.8 ± 1.5 | −24.1 ± 0.9 | −0.4 |
| $\varpi$ (mas) | 7.2 ± 0.4 | 7.0 ± 0.3 | 7.1 ± 0.4 | 0.1 |
| $d$ (pc) | 140.0 ± 8.5 | 142.5 ± 5.8 | 140.6 ± 7.9 | −2.4 |
| $\mu_\alpha^*$ (mas yr$^{-1}$) | −8.2 ± 3.0 | −9.6 ± 1.9 | −8.6 ± 2.8 | 1.4 |
| $\mu_\delta$ (mas yr$^{-1}$) | −24.9 ± 2.5 | −23.9 ± 2.0 | −24.7 ± 2.4 | −1.0 |
| $v_\alpha$ (km s$^{-1}$) | −5.4 ± 1.9 | −6.5 ± 1.3 | −5.7 ± 1.8 | 1.1 |
| $v_\delta$ (km s$^{-1}$) | −16.7 ± 1.6 | −16.1 ± 1.1 | −16.5 ± 1.5 | −0.6 |
| $v_{\mathrm{r}}$ (km s$^{-1}$) | −5.8 ± 4.3 | −5.0 ± 5.1 | −0.9 ± 68.5 | −0.8 |
| $X$ (pc) | 132.7 ± 8.1 | 134.1 ± 5.8 | 133.0 ± 7.6 | −1.4 |
| $Y$ (pc) | −15.8 ± 1.9 | −17.4 ± 3.0 | −16.2 ± 2.4 | 1.6 |
| $Z$ (pc) | 41.8 ± 3.4 | 44.6 ± 4.2 | 42.5 ± 3.9 | −2.9 |
| $U$ (km s$^{-1}$) | −5.2 ± 3.3 | −4.9 ± 4.5 | −5.9 ± 5.8 | −0.2 |
| $V$ (km s$^{-1}$) | −15.3 ± 1.4 | −15.9 ± 1.0 | −15.3 ± 1.5 | 0.6 |
| $W$ (km s$^{-1}$) | −8.6 ± 2.1 | −7.3 ± 2.5 | −8.7 ± 2.5 | −1.3 |

**Notes.** The average positional and dynamical values, including their standard deviations ($1\sigma$), were determined for the known and new sources separately, as well as for all of them together. The column $\Delta$ contains the difference of the known and new mean values for comparison of the two.

Table G.1 shows the average astrometric properties of the sources in $\rho$ Oph, such as the distances, proper motions, radial and tangential velocities, Galactic Cartesian positions $X, Y, Z$ and Galactic Cartesian velocities $U, V, W$, as well as the standard deviations ($1\sigma$) of these parameters. These average values were determined for the known and new sources, as well as for all of them together. To avoid the influence of outliers, impostors defined in Sect. 2.2 were not included in the calculations. The column $\Delta$ contains the difference of the known and new mean values for comparison.

There appear to be only small deviations between the properties of the known and new sources, which are not significant within $1\sigma$. This further confirms that, on average, they belong to the same region. The average values of $\rho$ Oph for $\varpi, \mu_{\alpha}^{*}, \mu_{\delta}, X, Y, Z$ agree relatively well with those determined by Cánovas et al. (2019) within $1\sigma$.

## Appendix H: $\rho$ Oph catalog overview

In this appendix we present our final catalog of $\rho$ Oph sources, which is available at the CDS. It includes all known sources from the literature and all sources identified by the OCSVM, even those with a stability <4, resulting in a total of 1343 sources. Table H.1 shows an overview of the column names, their units and their descriptions. In total, our catalog contains 67 columns. The column "Ref" serves as a reference for the literature sources, where each paper is cited by their reference number given in Table 1. Several sources were obtained from more than one paper; therefore, some sources have more than one reference number.

Since the known sources have proper motions and radial velocities obtained from the literature, *Gaia* EDR3, APOGEE-2, or *Gaia*-ESO, we provide the column "Ref_pm_rv" for the reference of the proper motions and radial velocity values, respectively. Each row contains two numbers for citation of these values, where "1," "2," "3," and "4" signify measurements obtained from the literature, *Gaia* EDR3, APOGEE, and *Gaia*-ESO, respectively. "0" implies that a source does not have a corresponding proper motion, parallax or radial velocity measurement.

**Table H.1.** Column overview of the final catalog containing known and new $\rho$ Oph sources.

| Column name | Unit | Description |
|---|---|---|
| source_id_edr3 | – | *Gaia* EDR3 ID |
| RA | deg | Right ascension (J2000) |
| Dec | deg | Declination (J2000) |
| *l* | deg | Galactic longitude |
| *b* | deg | Galactic latitude |
| parallax | mas | Parallax |
| parallax_error | mas | Parallax error |
| distance | pc | Distance, determined from the inverse of the parallax |
| pmra | mas yr$^{-1}$ | Proper motion in ra direction |
| pmra_error | mas yr$^{-1}$ | Error in pmra |
| pmdec | mas yr$^{-1}$ | Proper motion in Dec direction |
| pmdec_error | mas yr$^{-1}$ | Error in pmdec |
| radial_velocity | km s$^{-1}$ | Heliocentric radial velocity |
| radial_velocity_error | km s$^{-1}$ | Error in radial velocity |
| v_alpha | km s$^{-1}$ | Tangential velocity in ra direction |
| v_alpha_error | km s$^{-1}$ | Error in v_alpha |
| v_delta | km s$^{-1}$ | Tangential velocity in dec direction |
| v_delta_error | km s$^{-1}$ | Error in v_delta |
| $X$ | pc | Galactic Cartesian $X$ position component |
| $Y$ | pc | Galactic Cartesian $Y$ position component |
| $Z$ | pc | Galactic Cartesian $Z$ position component |
| $U$ | km s$^{-1}$ | Galactic Cartesian $U$ velocity component |
| $V$ | km s$^{-1}$ | Galactic Cartesian $V$ velocity component |
| $W$ | km s$^{-1}$ | Galactic Cartesian $W$ velocity component |
| ruwe | – | Renormalized unit weight error |
| astrometric_sigma5d_max | mas | Longest principal axis in the 5D error ellipsoid |
| astrometric_params_solved | – | Which parameters have been solved for |
| visibility_periods_used | – | Number of visibility periods in the astrometric solution |
| phot_g_mean_flux | e-/s | $G$-band mean flux |
| phot_g_mean_flux_error | e-/s | Error on $G$-band mean flux |
| phot_g_mean_mag | mag | $G$-band mean magnitude |
| phot_bp_mean_mag | mag | Integrated BP mean magnitude |
| phot_rp_mean_mag | mag | Integrated RP mean magnitude |
| bp_rp | mag | BP–RP color |
| Train | – | =1 for sources in the training set |
| Predict | – | =1 for predicted sources in *Gaia* EDR3 |
| New | – | =1 for new sources in *Gaia* EDR3 |
| Stability | – | Stability of the sources, range: 0–100 |
| Impostors | – | =1 for impostor sources |

**Notes.** Column overview of the final catalog of $\rho$ Oph sources, which includes the known sources from the literature as well as the new sources identified by the algorithm. The complete table is available at the CDS.

**Table H.1.** continued.

| Column name | Unit | Description |
|---|---|---|
| pml | mas $yr^{-1}$ | Proper motion in $l$ direction |
| pmb | mas $yr^{-1}$ | Proper motion in $b$ direction |
| v_l | km $s^{-1}$ | Velocity in $l$ direction |
| v_b | km $s^{-1}$ | Velocity in $b$ direction |
| angle_l_hel | deg | Heliocentric proper motion angle to $l$-axis |
| pml_lsr | mas $yr^{-1}$ | Proper motion in $l$ direction (LSR) |
| pmb_lsr | mas $yr^{-1}$ | Proper motion in $b$ direction (LSR) |
| v_l_lsr | km $s^{-1}$ | Velocity in $l$ direction (LSR) |
| v_b_lsr | km $s^{-1}$ | Velocity in $b$ direction (LSR) |
| angle_l_lsr | deg | LSR proper motion angle to $l$-axis |
| Pop | – | =1 for Pop 1 sources, =2 for Pop 2 sources, =0 if neither |
| IR_excess | – | = 1 for YSOs with IR excess, = 0 for Class III sources |
| designation_2MASS | – | 2MASS ID |
| j_m | mag | $J$-band magnitude |
| j_cmsig | mag | Uncertainty in $J$-band magnitude |
| h_m | mag | $H$-band magnitude |
| h_cmsig | mag | Uncertainty in $H$-band magnitude |
| k_m | mag | K-band magnitude |
| k_cmsig | mag | Uncertainty in K-band magnitude |
| designation_WISE | – | WISE ID |
| w1mpro | mag | WISE $W1$ magnitude |
| w1snr | – | $W1$ S/N |
| w2mpro | mag | WISE $W2$ magnitude |
| w2snr | – | $W2$ S/N |
| w3mpro | mag | WISE $W3$ magnitude |
| w3snr | – | $W3$ S/N |
| Ref | – | Reference for literature sources, see Table 1, range: 1–11 |
| Ref_pm_rv | – | Reference for proper motions and radial velocity: literature=1, *Gaia* EDR3=2, APOGEE=3, *Gaia*-ESO=4 |

## 3.29. Uncover: Toward interpretable models for detecting new star cluster members

### Full publication details

**Ratzenböck, S.**, Obermüller, V. Möller, T. Alves, J. & Bomze, I. 2022. "Uncover: Toward Interpretable Models for Detecting New Star Cluster Members" in *IEEE Transactions on Visualization and Computer Graphics*, doi: 10.1109/TVCG.2022.3172560.
*Accepted to IEEE TVCG in April 2022.*

### Author contributions

The paper is co-authored by me, Verena Obermüller, Torsten Möller, João Alves, and Immanuel M. Bomze. As the leading author, I conceived and developed the workflow and back end of the visualization tool, performed the interviews and case studies, and wrote the paper. The visual interface was mainly implemented by the Bachelor student Verena Obermüller. Torsten Möller supervised the project and offered suggestions along the way. João Alves and Immanuel M. Bomze helped revise the final version.

### Information on the Status

Submission date: 02 June 2021
Accepted: 24 April 2022
Published: TBD

# Uncover: Toward Interpretable Models for Detecting New Star Cluster Members

Sebastian Ratzenböck, Verena Obermüller, Torsten Möller, *Senior Member, IEEE,*
João Alves, and Immanuel M. Bomze

**Abstract**—In this design study, we present Uncover, an interactive tool aimed at astronomers to find previously unidentified member stars in stellar clusters. We contribute data and task abstraction in the domain of astronomy and provide an approach for the non-trivial challenge of finding a suitable hyper-parameter set for highly flexible novelty detection models. We achieve this by substituting the tedious manual trial and error process, which usually results in finding a small subset of passable models with a five-step workflow approach. We utilize ranges of a priori defined, interpretable summary statistics models have to adhere to. Our goal is to enable astronomers to use their domain expertise to quantify model goodness effectively. We attempt to change the current culture of blindly accepting a machine learning model to one where astronomers build and modify a model based on their expertise. We evaluate the tools' usability and usefulness in a series of interviews with domain experts.

**Index Terms**—Interpretable models, model selection, novelty detection, star clusters

## 1 Motivation

STAR clusters constitute the elementary building blocks of galaxies [45]. They provide probes for studying fundamental processes such as galaxy structure formation and evolution, stellar physics, and exoplanet evolution [55]. However, what astronomers know about stellar clusters is limited by the discovery process itself. Due to complex interactions with their dusty birthplaces, the tidal forces from the Milky Way, and unavoidable imperfect measurements and missing data, finding and extracting star clusters is challenging. Typically, new star clusters' discoveries consist of small high-confidence samples that minimize misclassification of stars. These high-fidelity samples are usually restricted to the dense cluster centers. However, larger samples would not only dramatically improve the quality of the derived cluster's physical parameters, but they also uncover the so far unseen low-density regions of stellar clusters. These low-density regions contain essential information on cluster formation and evolution [9], [23], [28], [52]. Although there is no conclusive methodology to identify new cluster members, the advent of deep, space-based all-sky surveys makes it a timely topic.

The search for new stars faces the challenges inherent to unsupervised clustering approaches. The absence of labeled data makes finding an optimal clustering result a highly nontrivial task. The two main challenges are hyper-parameter space exploration and result validation. To search for meaningful solutions, users often fall back to a laborious, manual trial-and-error process.

To mitigate the time spent blindly wandering through the hyper-parameter space, interactive tools such as Tuner [72], and Clustrophile 1+2 [13], [21] provide a systematic approach to hyper-parameter space navigation. Conversely, validation depends on the context of the analysis, the users' goals, and expertise. General purpose systems thus often make efforts to increase the interpretability of results beyond so called *internal validation* measures [48] based on cluster compactness and separation. These scores provide proxies for the goodness of a clustering result. However, since clustering results usually cannot be fully validated, internal validation measures should not be used to optimize clustering results.

This situation changes in the case of star clusters. Although no ground truth information is available for individual stars, systems of multiple stars can be validated by domain experts. General purpose visual cluster analysis tools often focus on data exploration and insight generation rather than generating an effective and accurate clustering result. Moreover, to generalize to a broad range of application scenarios tools such as Clustrophile 2 [13] hardly provide any clustering algorithms that can deal with complex feature spaces. Notably, in the search for new member stars of stellar clusters, we already have a set of previously identified members which currently available systems fail to incorporate. The given set of cluster members provides the chance of employing powerful novelty detection methods

- *Sebastian Ratzenböck is with the Data Science Research Network, Kolingasse 14-16, 1090 Vienna, Austria, and also with the Faculty of Computer Science, Währinger Straße 29/S6, 1090 Vienna, Austria.*
*E-mail: sebastian.ratzenboeck@univie.ac.at*
- *Verena Obermüller was with the Faculty of Computer Science, Währinger Straße 29/S6, 1090 Vienna, Austria.*
- *Torsten Möller is with the Faculty of Computer Science, Währinger Straße 29/S6, 1090 Vienna, Austria, and also with the Data Science Research Network, Kolingasse 14-16, 1090 Vienna, Austria.*
*E-mail: torsten.moeller@univie.ac.at*
- *João Alves is with the Department of Astrophysics, Türkenschanzstraße 17, 1180 Vienna, Austria, and also with the Data Science Research Network, Kolingasse 14-16, 1090 Vienna, Austria.*
*E-mail: joao.alves@univie.ac.at*
- *Immanuel M. Bomze is with the ISOR/VCOR, Oskar-Morgenstern-Platz 1, A-1090 Vienna, Austria, and also with the Data Science Research Network, Kolingasse 14-16, 1090 Vienna, Austria.*
*E-mail: immanuel.bomze@univie.ac.at*

*Manuscript received XXX; revised XXX; accepted XXX.*
*Date of publication XXX; date of current version 20 April 2022.*
*(Corresponding author: Sebastian Ratzenböck.)*
*Digital Object Identifier no. XXX*

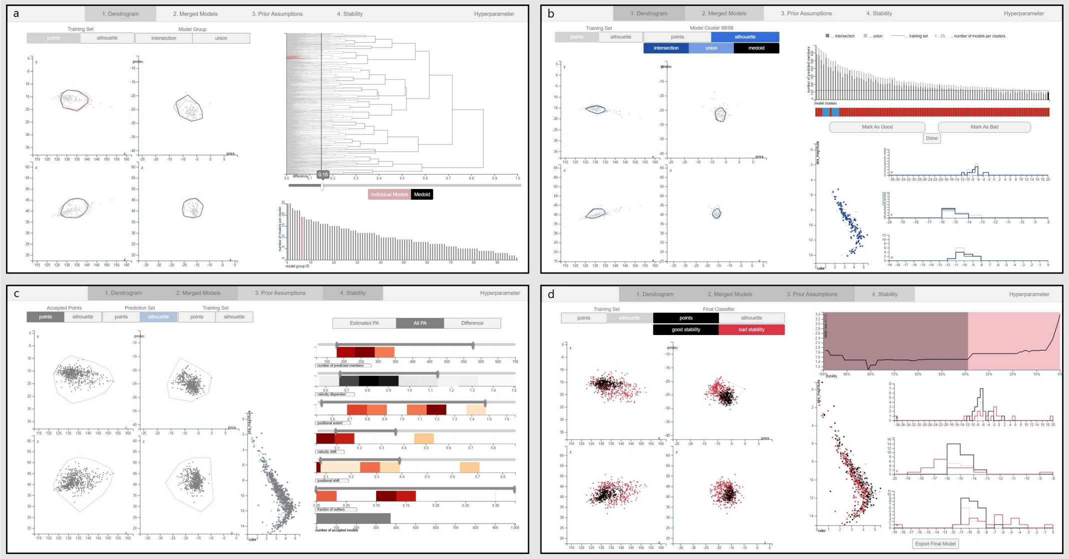


Fig. 1. Uncover Interface. a) Dendrogram Tab showing the silhouettes of model groups from the selected difference threshold. b) Model Group Tab showing the distributions of individual model groups. c) Prior assumption Tab to set accepted ranges for selected summary statistics. d) Stability Tab showing the final model ensemble and the prediction frequency of inferred members.

in which known stars are used as training samples.

Our goal is to enable astronomers to use their domain expertise to assess the quality of novelty detection models and in the process create interpretable (to astronomers) and accurate star classification models.

Given these design considerations, we present a five-stage workflow approach in which users (1) specify a priori knowledge in terms of constrained summary statistic ranges which influence the training of an ensemble of novelty detection models. Models are (2) clustered into user-defined groups which are subsequently (3) judged on their quality by domain experts. The users' quality assessment then updates the range of valid summary statistics. Subsequently, we support users to study and discover the effect of summary statistics on the shape of the predicted distribution in the context of their qualitative assessment. This gives users the opportunity to update their prior knowledge and influence the filter range (4). The updated statistics influence hyper-parameter restrictions on which a final large ensemble classifier is trained. Finally, the user is able (5) to filter out individual stars based on the prediction frequency across models, to finalize the novelty classifier. The contributions can be summarized in the following:

- We present a novel visually assisted workflow for finding appropriate hyper-parameters for highly flexible one-class support vector machines in the presence of training set contamination and extremely high outlier fractions (see Sect. 2).
- We introduce an analysis and abstraction of data, tasks, and requirements for the star formation domain (see Sect. 4).
- We breakdown the star classification process into small interpretable steps. We support users to apply their domain expertise to assess the goodness of trained models, effectively building confidence in the final classifier among domain experts (see Sect. 5).
- We validate our approach in two scientific use cases that demonstrate the efficiency and effectiveness of the Uncover interface in finding new stars (see Sect. 9).

## 2 Algorithmic and Domain Background

Our goal is to enable users to select meaningful models from the vast space of possible star classification solutions. Instead of guiding users through the hyper-parameter space we provide an overview of possible model configurations.

To facilitate model selection, we aim to increase the transparency and interpretability of individual models. To provide trust in selected models, we provide means of validating their outputs. We substitute unintuitive model hyper-parameters with a set of interpretable summary statistics and provide means to study their effect on the model outputs.

In the following, we discuss the necessary expertise to validate star clusters. We highlight and motivate clustering challenges in the context of star clusters more deeply. Subsequently, we discuss one class models and strategies to validate them.

### 2.1 Domain Background

Star clusters are dense groups of at least a few dozen stars. Although it is widely agreed that most stars form in stellar clusters [45] their exact formation history and subsequent evolution is currently subject to ongoing discussion [44], [76]. This discussion on fundamental star formation principles is fueled by the Gaia mission [29], [30], [31] which provides unprecedented positional and kinematic measurements of over 1.6 billion stars in our Milky Way. Since its public release the richness of the Gaia data has sparked a wave of discoveries of star clusters [10], [12], [16], [51]. By studying their size, age, and chemical compositions, stellar clusters provide valuable insights into galaxy formation, structure evolution, and stellar physics.

The precise study of physical processes and inference of physical model parameters is, however, limited by the discovery process. Star clusters appear as stellar overdensities in the space of position and velocity [42]. Due to physical processes such as complex interactions with the galaxy, imperfect measurements, and missing data, finding and extracting star clusters is challenging. Consequently, discoveries of new star clusters are often accompanied by a small high-confidence sample to avoid a high number of misclassified stars. Thus, when a new star cluster is discovered, domain scientists frequently sacrifice recall for high precision.

To infer physical quantities or test hypotheses on stellar physics and/or Galaxy structure and evolution, a sufficiently large sample of stars is needed. In these situations, a high recall is equally important. To uncover potentially new cluster members, star clusters are often subject to follow-up studies [9], [23], [28], [52]. Even though a set of high-fidelity stars already exists, these follow-up studies usually employ fully unsupervised learning, i.e., in data sets without labels indicating a class. Nevertheless, a common aim is to assign new members to previously identified stellar groups. We actually face a gray area between supervised and unsupervised learning, in statistical jargon between classification and clustering (not in the astronomy sense).

Recently however, novelty (or anomaly) detection approaches have been used to search for new member stars [37], [59]. Specifically, one-class support vector machines (OCSVM) [64] are trained on a set of high-fidelity member stars which are then able to identify unseen members. However, OCSVM classifiers are quite tedious to train. Their high flexibility and the lack of labeled outlier data limits their ability to generalize well on account of the provided training data only. Due to the lack of a clear objective function, domain experts usually fall back to manual trial-and-error processes.

Although no ground truth information is available for individual stars, ensembles of stars can be validated by domain experts. The distribution of stars in the positional and kinematic feature space, alongside their distribution in the Hertzsprung Russell diagram (HRD) provides evidence for or against a "true" star cluster hypothesis.

The HRD shows the evolutionary distribution of stars. It is a scatter plot in which the absolute magnitude of stars, a measure of their brightness, is plotted against the color, a measure of surface temperature, of the stars (see left side of Fig. 10). The position of a star on the HRD depends on a number of factors but notably on its mass, chemical composition, and age. During its life a star follows an evolutionary path through the HRD. Stars in stellar clusters are "born" together, originating from large collapsing molecular clouds, and thus have the same age and chemical composition. Therefore, star cluster members with different masses are found to lie on and around (due to errors in the measurement process) a curve in the 2D plane.

We aim to provide a visual interface that enables astronomers to use their domain expertise to search for meaningful star classification results.

### 2.2 Algorithmic Background

In this work, we focus on one-class support vector machines following their recent success in identifying unseen members in star clusters [37], [59]. The OCSVM method is an outlier and novelty detection algorithm which learns a tight and smooth boundary around a target data set. By applying the kernel trick, this boundary is highly flexible and can describe non-linear, arbitrarily shaped boundary regions. However, its extraordinary versatility quickly becomes its greatest drawback, as its performance depends heavily on the choice of input hyper-parameters.

Due to the lack of labeled outlier data, traditional model selection techniques such as cross-validation cannot be applied. Since no second class can restrict model growth, models that encompass the whole feature space would achieve a perfect test score. The optimal hyper-parameter selection for one class models remains an open problem to this day [70].

#### *2.2.1 Summary Statistics Heuristics*

To formally quantify the goodness of a classifier, a set of labeled data instances is needed. In the case of one-class models and unsupervised learning algorithms, principled quantitative validation is impossible. Although the OCSVM approach uses a set of training data in an extended sense, the absence of data instances labeled as abnormal may lead to a trivial model including all observations.

Instead, summary statistics such as the Silhouette score [62] offer an automated model selection heuristics. A set of summary statistics and respective predefined ranges provide straight-forward model filters.

In contrast to the hyper-parameters of the classifier (e.g., the bandwidth parameter $\gamma$ in the kernel function or the relaxation level $\nu$, see below), statistics can be chosen by the domain experts themselves and carry an immediate meaning that can be interpreted by astronomers. Statistics such as velocity dispersion, or the center of mass are metrics already used to quantify star clusters [28], [51]. Such a domain specific model selection heuristics was applied by Ratzenböck et al. [59] who initially motivated and described the use of OCSVMs to search for new member stars. Instead of tuning the model hyper-parameters directly, they compiled six "interpretable" summary statistics and selected models based on a priori defined ranges of these statistics. The final star classification model results then from aggregating the prediction of accepted models.

In the limit of sufficient statistics [27] a set of maximum likelihood estimates for the parameters of the data generating model can be determined. This requires, however, a-priori knowledge on the nature of the joint probability distribution function. In reality, we are left with a set of observed data and insufficient but still informative statistics on the unknown population. Due to the unknown complex interaction and physical model uncertainties, the nature of the underlying star clusters distribution is indeterminate. Domain expertise and a high-fidelity training set can be used to create informed summary statistics for model selection.

A major drawback of using summary statistics for model validation is owed to the vague and abstract nature of prior knowledge. For example, instead of specifying explicitly how many stars domain experts predict to find, a common answer would be: "The population is expected to increase only slightly but not by much." Qualitative feedback provides an effective validation alternative over summary statistics that is much less sensitive to vague knowledge.

#### 2.2.2 Qualitative Validation

In qualitative validation, users directly assess the model predictions. The goodness of star cluster models is tied to the distribution of inferred stars in the positional and kinematic features, respective to the training data. Especially the HRD provides means to support this decision.

Although summary statistics provide a fast model filter approach, visual analysis of inferred stars guarantees maximal confidence in the model. However, the manual inspection of up to millions [59] of models is practically infeasible. We aim to combine the best of both worlds by providing an update scheme on a-priori defined summary statistic ranges informed by manually validated models.

#### 2.2.3 A Combined Approach

To enable astronomers to become model builders themselves, we provide domain experts with a variety of potential model candidates for validation. We derive limits to summary statistics from validated models, which provides an automatic model filter for a subsequent exhaustive model search.

The high flexibility of OCSVM models results in a vast space of possible star classification results. Thus, for consistent results we have to properly sample the space of possible solutions. To deal with a large number of model realizations, we adopt a clustering strategy in which similar models are first grouped and then jointly evaluated. A similar strategy can also be found in FluidExplorer [7] where similar frames in a fluid simulation are grouped together.

To account for different star cluster shapes and sizes we cannot impose a strict clustering rule. Instead, our goal is to enable domain experts to summarize models into user-defined groups. A reasonable and interpretable framework to introduce user control is through hierarchical clustering using a complete-linkage criterion [20]. Compared to other popular linkage criteria such as single or average linkage, the complete-linkage criterion provides an easy to grasp conceptual framework for users. Complete-linkage translates the merge threshold domain experts are able to modify into a maximal difference between individual models in a cluster. In addition, models in a group are expected to show characteristic properties, implying a small intra-group variation. Single-linkage, however, can lead to a very high intra cluster variation as it applies a local merge decision, compared to the complete-linkage criterion.

To represent the distance between two models we choose the symmetric difference cardinality (SDC) between inferred sets of stars. The SDC of two sets A and B is the number of elements which appear in either A or B but not in both. To deal with various cluster sizes we normalize the SDC by the union of both sets, a modification which still preserves the metric quality of the difference measure [81]. This metric measures the relative difference between models, that is the fraction of stars by which models differ. It provides an interpretable difference compared to more complex distances such as the Hausdorff distance [61] that is less sensitive to border point fluctuations.

By using a global-to-local [67] approach we essentially cluster the solutions that allows a domain expert to inspect groups of similar models instead of having to validate each model individually. Each model group summarizes a common classifier trait giving users a much more concise overview of the solution space. Instead of qualitatively inspecting models individually domain experts assess resulting model clusters, thus, scaling to thousands of models.

The number of trained models affects the wait time for the initial training phase and the interpretability of the hierarchical model grouping algorithm in subsequent workflow steps. This is contrasted by the need to properly sample the space of possible star classification results. To cover the hyper-parameter space quickly and evenly, we draw samples from the Sobol sequence [2], [69] until convergence. We stop the sampling process if the majority ($> 90\%$) of the previous $50$ hyper-parameter tuples lack significantly novel models. Model novelty is defined as a normalized SDC of at least $0.05$ from previously trained models.

To improve the chance of finding many suitable models we pre-filter models based on initially defined summary statistic ranges based on a priori assumptions. This step limits the models presented to domain experts to plausible solutions.

Models are then trained according to Ratzenböck et al. [59] who have initially motivated and described the use of OCSVMs to search for new member stars. We briefly summarize the training steps here. To reduce overfitting, models are trained using five-fold cross validation[1], admitting only classifiers above a test accuracy of $50\%$ and a maximum standard deviation of $20\%$ across folds. Although cross validation cannot be used to select an optimal model which generalizes well, we can get rid of models that are unable to identify already known members. To reduce the influence of potential contamination by outliers in the training set, bagging is performed. To do so, individual models are trained on a random subset using $80\%$ of the initial training set.

Subsequently, domain experts are tasked to assess the goodness of self-defined model clusters. We derive updated ranges for the initially defined summary statistics from the user choices during the model validation step. Domain experts then have the option to further examine and modify the proposed ranges. Afterwards, a final training step that can be "run overnight" is performed where a much larger number of models are trained. These models have to comply to the user-informed, updated set of summary statistics.

[1] The training data is randomly shuffled before cross validation.

In the second, more detailed, model training step we are now able to narrow the hyper-parameter space which we derive from user validated models and the final summary statistic range. Therefore, we reduce the number of samples drawn from hyper-parameter space regions with unfit models while densely sampling from hyper-parameter space regions with a high acceptance rate. This strategy drastically reduces training time compared to a manual selection of initially vaguely informed summary statistic ranges [59].

## 3 Related Work

Although OCSVM models require a training step, the lack of labeled outlier data prevents us to quantify the goodness of trained models. Since solutions need a qualitative verification, the process of finding appropriate and effective models is inherently unsupervised.

A large variety of visual tools have been proposed to explore the space of possible classifiers. These tools are often based on visual hyper-parameter space exploration and aim to improve machine learning performance.

Uncover specifically focuses on one-class support vector machines and draws from prior work on optimal hyper-parameter selection.

### 3.1 Visual Clustering Analysis

A large body of previous work exists on interactive tools to support visual clustering analysis. General purpose tools provide means for exploratory data and cluster analysis. The Hierarchical Clustering Explorer (HCE [68]) is an early example of an interactive visualization tool that improves the users understanding of different clusters. HCE organizes the hierarchical cluster structure as a dendrogram with heatmaps. DICON [11] introduced techniques for comparing clustering results across different algorithms and even data sets. To facilitate cluster analysis DICON uses an icon-based cluster visualization that embeds statistical information into a multi-attribute display. Clustrophile 1+2 [13], [21] is a cluster analysis and exploration tool which guides a user through different choices of clustering hyper-parameters and provides interpretable cluster explanations.

Extensive work has been done on incorporating user feedback into the clustering process. ClusterSculptor [54] enables users to intervene in the clustering processes. Users can iteratively re-organize and interact with clusters using expert knowledge. The system aims to derive clustering rules from these examples. Schreck et al. [65] integrate user feedback to influence the result of SOM clusterings of trajectory data. Matchmaker [46] extends ideas from HCE [68] allowing users to modify clusterings by grouping data dimensions. Open-Box Spectral Clustering [66] is an interactive tool that visualizes mathematical quantities involved in 3D spectral clustering. The system provides hyper-parameter value suggestions and immediately reacts to user feedback to increase the quality of image segmentation. Packer et al. [56] present a distance-based spatial clustering approach and provide a heuristics computation of input hyper-parameters that supports the search for meaningful cluster results. ReVision [80] allows users to steer hierarchical clustering results by utilizing both public knowledge and private knowledge from users. By reformulating this knowledge into constraints, the data items are hierarchically clustered using an evolutionary Bayesian rose tree.

Conceptually similar research to ours include Geono-Cluster [18] and PK-clustering [58]. Geono-Cluster enables biologists to insert their domain expertise into clustering results. The tool displays the expected clustering results to users based on a small subset of data. The system estimates users' intentions and generates potential clustering results. PK-clustering [58] enables users to input prior knowledge and explore the space of clustering results in the context of the provided prior knowledge. The study of consensus between prior assumptions and cluster results allows users to acquire and update their prior knowledge.

In contrast to previous works we shift the focus from data exploration and insight generation towards effective model generation targeted at a single cluster. We also incorporate previously identified members which currently available systems fail to consider by using a supervised novelty detection approach.

### 3.2 OCSVM Hyper-parameter Selection

Optimal hyper-parameter selection for one class models remains an open problem [70]. In the following, we discuss automated as well as visually supported model selection approaches.

#### 3.2.1 Automatic Hyper-parameter Selection

To mitigate the non-trivial selection process of OCSVM hyper-parameters, automatic hyper-parameter selection approaches have been proposed, which should provide suitable results. Automatic strategies either provide selection heuristics, or focus on producing a set of pseudo-outliers [4], [22], [24], [70], [71], [74]. These artificial outliers are subsequently used as an opposing class to the training data during cross-validation. Heuristics are often limited to specific kernel parametrizations. As RBF kernels bring a high degree of model flexibility most heuristics usually focus on them [26], [34], [43], [75], [78].

Both automatic approaches, however, often assume a problem in which the target class is sufficiently represented while the other class has almost no measurements in comparison [70]. This class imbalance assumption towards the training set is in stark contrast to stellar clustering where the target class is a minority embedded in, and outnumbered by, a background of non-member stars. Furthermore, automatic methods usually provide point estimates for hyper-parameters, providing only a single model to infer new member stars with.

Even in the case of optimal model hyper-parameters, one-class algorithms exhibit poor performance [71], which we can combat by using non-optimal learners in an ensemble approach. Bagging estimators improve the performance and robustness of the prediction [36]. Additionally, point estimates cannot adapt to specific user expectations and introduce errors in the case of noisy training data. Since residual contamination in the training sample from non-member stars is expected, we have to consider that OCSVM classifiers can be sensitive to contamination from outlier data [39], [47]. In this case, the OCSVM classifiers tend to skew toward the anomalies. Amer et al. [1] propose to mitigate the influence of outliers by altering the OCSVM objective function introducing training sample weights. Instead of tweaking the objective function, Ghafoori et al. [33] introduce a pre-processing step which removes anomalies from the training set and simultaneously tries to estimate suitable hyper-parameters. Both approaches, however, need some form of outlier estimate, be it either through the distance to the data centroid [1] or via a k-NN density estimate [33] implying that outliers occur towards the border, or in low density regions of the training set. While this assumption is sufficient for many applications, we cannot generalize this to star clusters where contamination depends greatly on the training set selection method.

#### 3.2.2 Visual OCSVM Hyper-parameter Estimation

A different and more user-centered approach to find a suitable model was presented by Xie et al. [79] in which the OCSVM classifier is trained in an active learning scenario. User feedback on uncertain samples near the decision boundary updates the decision boundary.

Although active learning is able to adapt to specific user expectations, it fails in the context of star clusters. Data instances can rarely, if ever, be assessed on an individual basis. Conversely, however, it is very much possible for domain experts to discern a genuine star cluster from an incoherent system of stars.

## 4 Data and Task Analysis

We now discuss the data and tasks, and a derived workflow to support the search of new star cluster members. The data flow and workflow are schematically depicted in Fig. 2 and Fig. 3, respectively.

### 4.1 Data

The main data source is the aforementioned Gaia data set [29], [30], [31], a tabular data set containing measurements of over 1.6 billion stars in our Milky Way. Features relevant for this analysis constitute continuous, real-valued measurements of position and velocity, and color and absolute magnitude information which are used for model fitting, and validation, respectively. Users input two separate data sources, a training set and a prediction set. The latter is used to infer cluster membership with trained models. We note here that the full 3D kinematic information is available only for a small subset of stars in the Gaia data set. As discussed in Ratzenböck et al. [59], during training a reduced 2D velocity space is used, called *proper motion* space. Stars that have the full 3D kinematic information are used to validate models.

To speed up the inference process it is advised to provide a small subset of stars in the positional vicinity of the training set where new stars are assumed to lie in. Usually, both the training and prediction set are subsets of the Gaia catalogue. In principle, these two data sources can originate from different star catalogues as long as the feature set is identical[2].

#### 4.1.1 Model Abstraction

The OCSVM model can be abstracted as a basic deterministic input-output model converting input tuples to outputs.

Given the input hyper-parameters $\gamma$, $\nu$, and $\frac{c_x}{c_v}$ and the training set, OCSVM constructs a decision surface that aims to maximize the separation between the training data and the origin. The resulting model is a decision hyper surface enclosing the training data in the input space which constitutes a binary function that classifies new data as in- or outliers. The hyper-parameter $\gamma$ is related to the RBF kernel and controls the region of influence of support vectors. The variable $\nu$ provides an upper bound on the fraction of outliers and at the same time a lower bound on the fraction of support vectors used to construct the decision surface. The hyper-parameter $\frac{c_x}{c_v}$ provides a scaling relationship between positional and proper motion features [59]. Both subspaces are weighted equally when $\frac{c_x}{c_v} = 1$ in which case the variance in both feature spaces is the same.

The kernelized nature of OCSVMs provides an extremely flexible model that adapts well to arbitrary cluster shapes observed in star clusters. In extreme cases, a strongly concave shape is observed resulting from projection effects due to the lack of radial velocities.

Among the outputs are a Boolean member classification for each star in the prediction set and a set of six informative summary statistics derived from the predicted members.

### 4.2 Summary Statistics

Here we make use of the following summary statistics defined by Ratzenböck et al. [59]:

The "number of predicted stream members" is the amount of cluster members a trained model infers from the given prediction set.

The statistics "positional extent" and "velocity dispersion" measure the mean deviation from inferred cluster members from the training set centroid in position and proper motion space, respectively.

The relative position or systematic shift of inferred stars compared to the training set in these two subspaces is characterized by "positional shift" and "velocity shift". These statistics characterize the distance between the centroids of training and inferred stars.

Lastly, "fraction of outliers" utilizes information of stars in the training set and inferred stars that have radial velocity measurements. Models that show significantly different 3D velocities than the training set are considered outliers. This statistic measures the fraction of inferred stars with radial velocities that are outside the $3\sigma$ region of training set stars in marginal 3D velocity distributions.

The authors referred to these summary statistics as *prior assumptions* (PA) which we use synonymously in the following sections.

### 4.3 Task Analysis

We aim to enable astronomers to update vague prior knowledge on the number, location, and movement of unidentified stars, altogether six summary statistics. The assessment of the goodness of multiple models should thereby provide the necessary information to reduce the uncertainty in these summary statistics.

[2] In case two different source catalogues are used, special care must be taken to correctly consider differences in statistical and systematic errors between them.

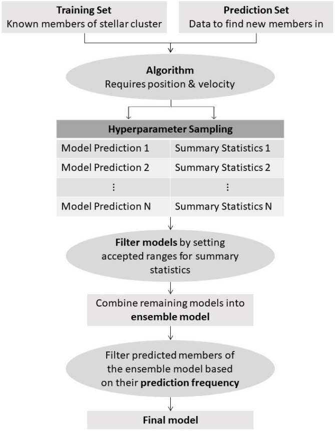

Training Set
Known members of stellar cluster
Prediction Set
Data to find new members in
Algorithm
Requires position & velocity
Hyperparameter Sampling
Model Prediction 1
Summary Statistics 1
Model Prediction 2
Summary Statistics 2
Model Prediction N
Summary Statistics N
Filter models by setting accepted ranges for summary statistics
Combine remaining models into ensemble model
Filter predicted members of the ensemble model based on their prediction frequency
Final model

Fig. 2. Schematic data flow of Uncover.

To facilitate this transition the user has to be able to validate and influence the model selection process down to the individual classifier. With this characterization in mind, we carry out a task analysis. To facilitate comparison to other works, we try to provide abstract reasoning *why* a task is performed [5].

**T1 Verify/Validate** a trained model via its predicted members. To validate models, summary statistics usually provide too little information to inform a confident decision. Instead, domain experts use qualitative judgement to assess the goodness of models, requiring the following. First, users have to be able to assess the distribution of predicted stars in the space of *position and velocity* (**T1.1**) and compare them to the training set. Second, the distribution of stars in the *HRD* provides additional evidence for or against a valid star cluster (**T1.2**).

**T2 Identify** suitable summary statistics ranges. Ranges on summary statistics provide a filter criterion during the full training process (see Sect. 2.2.3) to automatically remove unfit models. We derive updated ranges for each of the six statistics from the users' qualitative model assessment. However, to provide insight into these filters, users have to be able to study and discover their effect on the shape of the predicted distribution (**T2.1**). Users should also be able to explore and analyze the distribution of assessed models in the context of summary statistics (**T2.2**). This gives them the opportunity to update and substantiate their prior knowledge. Finally, users must be able to apply their updated knowledge and interactively refine filter ranges on summary statistics (**T2.3**).

**T3 Explore** the effect of stability filters on the inferred stars. Stability is the prediction frequency of stars across the model ensemble. Stars with high stability are thus inferred by most of the models and vice-versa. Ratzenböck et al. [59] have shown that removing stars with low stability values removes disproportionately more contaminant stars than genuine cluster members, effectively cleaning the sample. We aim to facilitate the exploration of different stability thresholds to study the effects on the ensemble model prediction. Using their domain expertise, users should thereby be able to select a meaningful stability threshold.

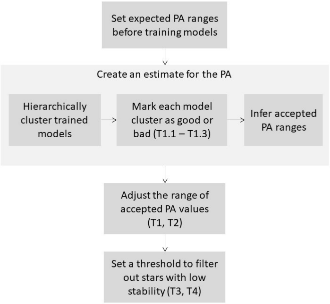

Set expected PA ranges before training models
Create an estimate for the PA
Hierarchically cluster trained models
Mark each model cluster as good or bad (T1.1 – T1.3)
Infer accepted PA ranges
Adjust the range of accepted PA values (T1, T2)
Set a threshold to filter out stars with low stability (T3, T4)

Fig. 3. Schematic workflow of the tool.

**T4 Present** the inferred cluster members of the final ensemble model. To validate the final ensemble model we present the distribution of training and inferred stars in the space of position and velocity, in combination with the HRD. In case domain experts see the final model as unfit, users can go back to previous workflow steps and intervene accordingly.

**T5 Summarize** the model ensemble in terms of their hyper-parameters at different workflow steps. To provide a transparent view on the OCSVM algorithm, users have to be able to inspect the distribution at any time. To understand the model selection effect on the hyper-parameters, we present the distribution of hyper-parameters of models that domain experts deemed fit in comparison to the initially trained, unfiltered models.

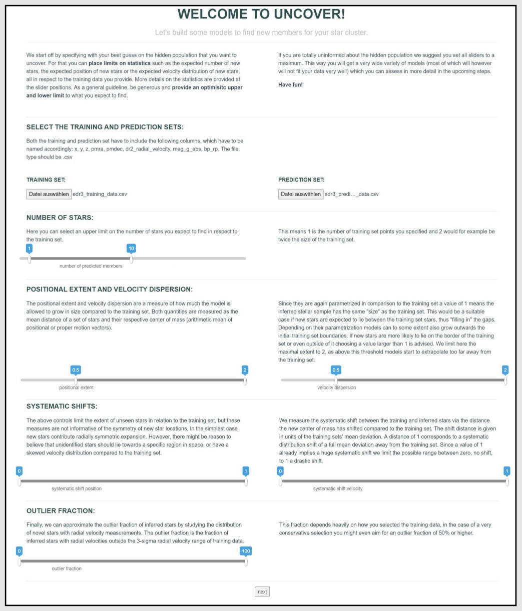
## 5 Uncover Interface

We now discuss the design of the tool starting with the general layout, followed by descriptions of the individual tabs and visualization components.

### 5.1 Layout

The prototype comprises six different views in total, one for each workflow step as well as an additional view for showing information on the hyper-parameters. At the top of each view is a tab-bar, which enables the user to navigate between the different workflow steps and the hyper-parameter view. The tabs are arranged in order of the workflow steps, see Fig. 1 for an overview of the interface from the second to the last workflow steps. The first workflow phase is shown in Fig. 4.

For each of the five tabs, the same general layout (see supplemental material Fig. 2) is used to create a consistent interface throughout the tool. If users can already anticipate where certain information will be presented, users can more



**WELCOME TO UNCOVER!**

Let's build some models to find new members for your star cluster.

We start off by specifying with your best guess on the hidden population that you want to uncover. For that you can **place limits on statistics** such as the expected number of new stars, the expected position of new stars or the expected velocity distribution of new stars, all in respect to the training data you provide. More details on the statistics are provided at the slider positions. As a general guideline, be generous and **provide an optimisitc upper and lower limit** to what you expect to find.

If you are totally uninformed about the hidden population we suggest you set all sliders to a maximum. This way you will get a very wide variety of models (most of which will however will not fit your data very well) which you can assess in more detail in the upcoming steps.

**Have fun!**

SELECT THE TRAINING AND PREDICTION SETS:

Both the training and prediction set have to include the following columns, which have to be named accordingly: x, y, z, pmra, pmdec, dr2_radial_velocity, mag_g_abs, bp_rp. The file type should be .csv

TRAINING SET:

Datei auswählen edr3_training_data.csv

PREDICTION SET:

Datei auswählen edr3_predi…_data.csv

NUMBER OF STARS:

Here you can select an upper limit on the number of stars you expect to find in respect to the training set.

This means 1 is the number of training set points you specified and 2 would for example be twice the size of the training set.

1 10
number of predicted members

POSITIONAL EXTENT AND VELOCITY DISPERSION:

The positional extent and velocity dispersion are a measure of how much the model is allowed to grow in size compared to the training set. Both quantities are measured as the mean distance of a set of stars and their respective center of mass (arithmetic mean of positional or proper motion vectors).

Since they are again parametrized in comparison to the training set a value of 1 means the inferred stellar sample has the same "size" as the training set. This would be a suitable case if new stars are expected to lie between the training set stars, thus "filling in" the gaps. Depending on their parametrization models can to some extent also grow outwards the initial training set boundaries. If new stars are more likely to lie on the border of the training set or even outside of it choosing a value larger than 1 is advised. We limit here the maximal extent to 2, as above this threshold models start to extrapolate too far away from the training set.

0.5 2
positional extent

0.5 2
velocity dispersion

SYSTEMATIC SHIFTS:

The above controls limit the extent of unseen stars in relation to the training set, but these measures are not informative of the symmetry of new star locations. In the simplest case new stars contribute radially symmetric expansion. However, there might be reason to believe that unidentified stars should lie towards a specific region in space, or have a skewed velocity distribution compared to the training set.

We measure the systematic shift between the training and inferred stars via the distance the new center of mass has shifted compared to the training set. The shift distance is given in units of the training sets' mean deviation. A distance of 1 corresponds to a systematic distribution shift of a full mean deviation away from the training set. Since a value of 1 already implies a huge systematic shift we limit the possible range between zero, no shift, to 1 a drastic shift.

0 1
systematic shift position

0 1
systematic shift velocity

OUTLIER FRACTION:

Finally, we can approximate the outlier fraction of inferred stars by studying the distribution of novel stars with radial velocity measurements. The outlier fraction is the fraction of inferred stars with radial velocities outside the 3-sigma radial velocity range of training data.

This fraction depends heavily on how you selected the training data, in the case of a very conservative selection you might even aim for an outlier fraction of 50% or higher.

0 100
outlier fraction

Fig. 4. First workflow step of Uncover.

quickly adapt to a new view and therefore reduce mental overhead [63]. We divide the interface into two equal sized sections, the *scatterplot matrix* and *update* section. The update section in the right half adapts to each workflow step. It contains interaction components which facilitate cluster selection, model navigation and assessment, updating and refining prior knowledge (**T2**), and stability threshold exploration and selection (**T3**). The left section provides a reduced scatterplot matrix, which shows the position and proper motion dimensions separately. This is used to display the multi-dimensional data set (**T1**, **T4**). Depending on the respective workflow step, different data aspects and models are highlighted. This can be the training set, different model groups, medoid models, the models at the minimum and maximum of each summary statistic range, or the final ensemble model. We describe the scatterplot matrix component in more detail when it first appears on the "Dendrogram Tab" in Sect. 5.2.2.

## 5.2 Visualization Components

In this section, the chosen visualizations as well as their intended function for carrying out the corresponding workflow step are discussed in more detail.

### 5.2.1 Dendrogram Tab

Based on training set characteristics, the number of hyperparameter tuples needed to properly cover the space of possible star classification results can be in the hundreds or even thousands. However, users cannot be tasked to assess the quality of each individual model. Instead, we support users to choose groups of similar models that can be assessed together instead of individually.

To summarize possible model clustering configurations the update section of this view, shown in Fig. 9, features a dendrogram. The dendrogram provides an overview of the clustering hierarchy of models resulting from a complete-linkage agglomerated clustering approach. At each step, the two model clusters with the smallest relative difference in predicted points are combined into the same cluster. This difference value is shown on the x-axis of the dendrogram plot. To be able to perceive structure in the dendrogram towards smaller distances, its lines become progressively thinner from 1 to 0 to avoid visual overlaps. The slider can be used to set a threshold for the difference, where merging will stop, so that models with a difference greater than the selected value will remain in separate groups. The bar chart below the dendrogram shows the number of models in each group resulting from the current threshold.

### 5.2.2 Scatterplot Matrix

The scatterplot matrix, seen in the left half of the view (see Fig 1a), shows the model groups resulting from the current cut along with the training set representing the baseline. We aim to provide an overview of the clustering results and thus facilitate a comparison between the resulting groups of models. We choose two summary operands for model groups; the union and the intersection of points inferred by individual models in a group.

The intersection provides a summary of common model features across a group. By comparing the intersection and union of stars inferred by group members we provide an estimate of within-group variation that is easy to understand. The further the two group summaries diverge, the less the models in a group form a coherent cluster. In such cases, a better clustering result can be achieved by reducing the difference threshold.

We choose to summarize models as silhouettes in the scatterplot matrix which shows the maximal extent region of the predicted distribution in each projection. It acts as a visual simplification of a model in the form of a convex hull around the predicted points. Compared to scatter points, silhouettes allow users to easily compare multiple model groups. In this scenario, indicating group identity is nontrivial in scatter points. Not only is the use of color limited to roughly six to seven groups [53] but a large amount of points are also part of multiple groups which drastically increases the amount of unique visual encodings required. Therefore, since examining the stars inferred by individual model groups and assessing their goodness is not the purpose of this workflow step, but of the following one, we omit the display of scatter points here.

To assess a group of models in detail, in order to determine whether they form a meaningful unit, domain experts can explicitly display the convex hull of each model in a given group. Additionally, users can highlight the group medoid, the representative model of the group. It provides an opportunity to identify group characteristics like a certain set of stars that this model group has in common. A comparison with the remaining models should provide further insight into the model variation within the group. By studying the group medoid and the overlap and

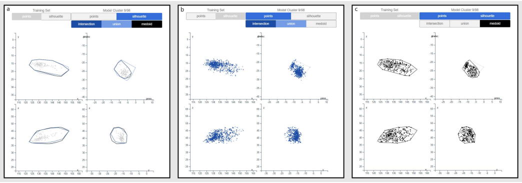


Fig. 5. Different selections in the scatterplot matrix during the model group validation step. On the left (a) three group summaries are highlighted; the union and intersection of predicted members, as well as the group medoid are shown in the form of silhouettes. The middle view (b) shows predicted members as scatter points. In the right view (c) a combination of both model group summaries – points and silhouettes – are used. The training data are displayed as gray scatter points in all three views.

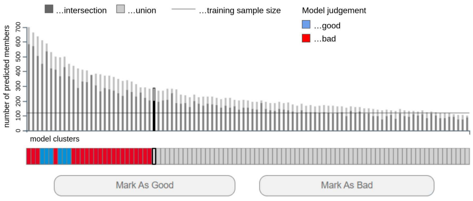


Fig. 6. Visualization components for assessing the model groups. The stacked bar chart shows the number of predicted members resulting from the union and intersection of models in each group. The blue and red cells indicate a good or bad marking of the corresponding model group, respectively.

variation between silhouette shapes, users can determine an appropriate threshold.

### 5.2.3 Model Group Tab

In this workflow step, users are tasked to assess the goodness of model groups defined in the previous step.

The scatterplot matrix view displays the model groups one after the other. The user can choose to plot the training set in the same scatterplot matrix to compare it to the currently shown model group. Depending on the use case, the distribution of inferred member stars in positional and proper motion space in relation to the training can give strong indications towards a good and bad model, see Fig. 5.

To validate the models, positional and kinematic information is provided in the scatterplot matrix (**T1.1**) and an HRD is provided in the update section (**T1.2**). To assess a model, domain experts can verify whether predicted members are distributed in a narrow line in the HRD according to the training set or not. To leverage the kinematic information from the inferred stars for model validation, we provide two kinematic views, see Fig. 1b for more details. First, the proper motion information used for training is displayed in the scatterplot matrix. Second, the Cartesian velocity distribution is displayed in three histograms next to the HRD, see Fig. 10. To verify that the predicted member stars constitute a stellar cluster the Cartesian velocity distribution should roughly follow a normal distribution and not deviate significantly from the training set [42].

For both the training set and the model group, the user can switch between viewing individual data points, which are classified as members, or the silhouette thereof by clicking the buttons labeled accordingly. For each model group, the user can choose to view the union of all inferred members or only the stars that are predicted members across all models in the group. This selection can be done via the buttons labeled "union" or "intersection" above the scatterplot matrix, respectively, see Fig. 5. Additionally, to facilitate the judgement of a group of models, the medoid can be selected as a model representative. Compared to the union and intersection of stars via a model group, the medoid represents an individual model in which characteristic model details become more apparent.

The number of predicted points for both the union and intersection of the models in each group is visualized using a stacked bar chart in the update section. Since users aim to find additional star cluster members, this is the most important summary statistic which provides an overview across model groups.

The number of inferred stars is considered to strongly correlate with model goodness. Depending on the level of prior knowledge, domain experts might be interested in specific ranges of inferred member sizes. Therefore, we sort the bar chart in descending order by union size to support different levels of attention during the users' workflow. This allows users to string together groups that require more attention during the validation process, followed by groups that require less consideration. This attention bias applies, for example, to models that find about the same number or even fewer members compared to the training set. These models typically require less validation effort, as their member size alone indicates a lack of new discoveries. To facilitate a comparison with the training set a horizontal dashed line is drawn indicating its size.

Once the user has come to a decision regarding the suitability of the currently shown model group, the corresponding button in the update section shown in Fig. 6 can be clicked to either mark it as "good" or as "bad". Afterward, the next model group is shown. The bar between the buttons and the bar chart highlights the progress and gives an overview of the model group assessment. Blue and red indicate a good or bad model group, respectively. Model groups which have not been assessed yet are colored in gray. When all the model groups have been evaluated, the button labeled "Done" can be clicked to generate the estimate for the accepted Prior Assumption (PA) ranges.

### 5.2.4 Prior Assumption Tab

The third workflow step, seen in Fig. 1c, supports the analysis and possible adjustment of the PA ranges which result from the previous step.

Each of the six PAs and the corresponding derived ranges are visualized with the help of *scented widgets* [77]. The *widget* is made up of two sliders, one for the minimum and one for the maximum of each PA range. These are positioned on top of the *visual scent* in the form of a bar, which shows the distribution of PA values from all models as a heatmap. The darker the luminance of a cell, the more models have a PA value in the matching range. See Fig. 7 for a detailed view of the PA range interface.

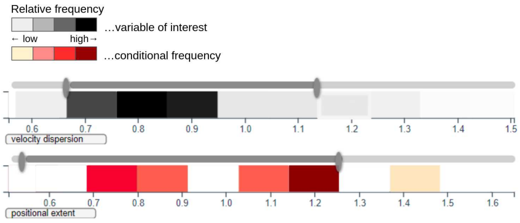


Fig. 7. Scented widget for setting the accepted PA range. The heatmap visualizes the distribution of PA values among either all trained models, all models marked as good, or the difference between the two. Clicking on a heatmap-cell will update the remaining heatmaps to show the models in the selected range in a red colormap.

The heatmap provides a means to analyze the distribution of assessed models in the context of summary statistics (**T2.2**). Users can explore correlations between a selected PA and the remaining PAs. By clicking on a cell of a given PA, all the models whose PA value lies in the selected range will be highlighted in the remaining five heatmaps. To visually separate the distribution of models from a selected heatmap cell in the other summary statistics we choose a red colormap, as can be seen in the update section of Fig. 1c. This interaction supports users to find model trends and correlations.

When first opening the tab, the initial slider position shows the estimate for the accepted PA ranges created in the previous step. For each PA, the sliders are placed according to the minimum and maximum PA value of the models that were marked as good. Additionally, by clicking on the buttons above the heatmaps users can either study the distribution of all initially trained models, under "All PA", the distribution of models assessed as good by the user, under "Estimated PA", and the models judged as bad, under "Difference".

By studying the correlation between models' summary statistics and the distribution of "good" and "bad" models, users can substantiate their prior knowledge and interactively refine PA filter ranges giving them the opportunity to (**T2.3**) precisely control the properties of the final model.

The slider positions in the heatmaps correspond to models shown in the scatterplot matrix. We provide a what-if-analysis where users can isolate the effects of a single PA and study its influence on the inferred stars. At each slider position, stars inferred by models which adhere to the selected filter criterion are shown in the scatterplot matrix. The minimum slider position corresponds to a minimum set of stars that these models can identify. A sensible choice is to require models to at least identify large parts of the training set. The maximum slider represents stars that can be detected by models up to the selected PA value. To illustrate the effect of the whole slider range, we exclusively show stars that can be detected beyond the minimum slider value. Stars associated with the minimum slider position are colored in light blue whereas stars associated with the maximum slider position are highlighted in a darker shade of blue. Light gray points in the background indicate stars outside the maximum slider position which are not inferred by selected models. When no PA is selected, stars inferred by models which adhere to the slider range filters are highlighted in gray in the scatterplot matrix.

By interactively changing the slider position for one or multiple PAs users can study the influence of various summary statistics on the shape and distribution of inferred stars in position, velocity, and the HRD, as well as the correlations between the model behavior and a given summary statistics in more detail (**T.2.1**). This interaction provides additional information for users to update their prior belief and refine given filter ranges (**T2.3**).

The bar visualization at the very bottom of the right half encodes the number of models out of the initially trained ensemble that pass the PA range filter. Thus, it informs the user how restrictive their current ranges are setup. The bar length is updated whenever slider positions are changed.

In this step outlier models can motivate an alternative workflow. As discussed, Uncover is not aimed at providing means for exploratory data analysis, but rather for effective model building. Thus, identifying and characterizing outlier models is not an important task for the user. Especially outlier models which are classified as "bad" require no further investigation on the users' end. Hence, outlier models are not explicitly marked as such in the tool to avoid drawing unnecessary attention to them. However, if an outlier model is considered "good", a user may find few appropriate models in the initially trained model ensemble. In an effort to increase the diversity of "good" models, domain experts might want to restart the training process. This can be done by returning to the first workflow step and modifying initial summary statistic ranges. A sensible choice is to center updated ranges around those of given outlier models. Their respective summary statistics can be analyzed in the heatmap view, see Fig. 7.

### 5.2.5 Stability Tab

The last step of the workflow is dedicated to the final ensemble model and the stability of its predicted members. The final ensemble model is the result of combining the predictions of the models that fulfil the PA restrictions set up in the previous workflow steps. The final predicted distribution of the stellar cluster in question is shown in the scatterplot matrix, the HRD, as well as in the histograms displaying the Cartesian velocity, as shown in Fig. 1d. These views also show the training set to facilitate comparison (**T1**) and allow the user to verify that the final ensemble model creates a suitable prediction.

To switch between viewing the points and the silhouettes, the buttons on top of the scatterplot matrix can be used. However, in the case of the HRD, showing the silhouette of a distribution is not always useful. Stars in different stages of stellar evolution typically occupy distinct subregions of the diagram [30], so a predicted distribution that comprises stars in varying evolutionary phases could form separate clusters with large gaps between them in the HRD. Drawing a silhouette encompassing all the points would then result in a shape that is too coarse and does not reflect the underlying distribution in a useful manner.

The threshold for the stability can be set with the help of a *scented widget* [77] which features a line chart showing the stability in percent and the median absolute deviation (MAD) of predicted members from the expected 3D velocity.

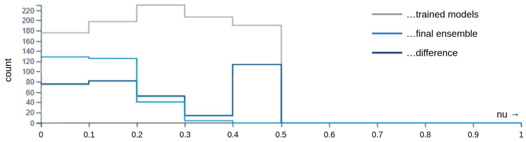


Fig. 8. Histogram of one of the three hyper-parameters. The dark gray line corresponds to all trained models, the bright blue line to all accepted models. The difference between these two is shown by the the dark blue line. The light gray lines show the accepted models from previous settings.

The right side of the brush on the line chart can be moved to set the minimum stability for the final classifier. This also updates the presentation of the predicted distribution in the remaining plots: All points with accepted stability are colored black, while points that will be filtered out because their stability is too low are shown in red, an example of this can be seen in Fig. 1d. If the silhouette-button is selected, the silhouette resulting from the points with acceptable stability is colored black while the silhouette encompassing all predicted members is shown in red.

### *5.2.6 Hyper-parameter Tab*

The previous tab aims to provide supplementary information on the hyper-parameters of the accepted models. Even though the aim of the tool is to relieve the user of having to work directly with the hyper-parameters, information on them should still be available to give the user the possibility to get a better understanding of them. For each of the three hyper-parameters [59] $\gamma$, $\nu$ and $\frac{c_x}{c_v}$, there is one histogram showing the number of accepted models as well as the number of trained models for each possible hyper-parameter value as can be seen in Fig. 8.

## 6 Design Rationale

In this section, the motivations behind the chosen visual encodings are discussed regarding the data and task abstraction.

6.0.0.1 Why a tab-based interface?: An alternative to tabs would be to present the necessary visualizations for carrying out the different workflow steps on a single page. This would remove the need to switch between different views and therefore impose a lower cognitive load on the user [14]. However, the number of required visualizations would not fit onto a single screen in a reasonable size without requiring to scroll the page. The different workflow steps were therefore separated into individual tabs to ensure that the visualizations for each step fit appropriately onto a single screen. To reduce the cognitive load caused by the user, each tab functions as a self-contained unit. This means that every tab contains all necessary visualizations to fully carry out the associated tasks and does not require the user to remember information shown in previous tabs.

To further reduce the mental load when opening a new tab the interface layout (see supplemental material Fig. 2), remains the same throughout the tool. Especially the scatter plot matrix and the actual data displayed on the left-hand side stay the same across the entire tool.

Although the tool supports a linear workflow once the initial model ensemble is trained, the user can decide to go back to any workflow step and modify their decisions. The first step is not part of this tab interface since it amounts to starting the tool up again from the beginning, which requires another time-consuming training step.

6.0.0.2 Why scatterplot matrices?: The 3D position and proper motion of the stars are always shown using scatterplot matrices, since this is the standard way of visualizing stellar clusters in the field of astronomy. Other visualization methods for multi-dimensional data were initially considered but found to be unsuitable in this context. Parallel coordinates [40] would be an alternative to scatterplot matrices; However, they are not commonly used in astronomy and would therefore not be very intuitive for the target audience. Additionally, scatterplots can act as 2D projections of the underlying real-world objects described by the data, which reside in a 3D space, and are therefore much more straightforward to interpret. 3D scatterplots were also considered, but showing the data in 3D can result in a variety of problems [53]. The large number of points that need to be presented would make the use of 3D especially challenging, since this would lead to a significant amount of occlusion and thus make it hard to get a full view of the distribution.

Showing the apparent motion of stars as an oriented line anchored at their sky position is a common visual encoding used in astronomy, e.g. de Zeeuw et al. [19] famously showcase three co-moving groups in the nearby Scorpius-Centaurus OB association. The instantaneous velocity of a star is encoded as a small arrow whose origin is at its position. The length of the arrow encodes speed while the angle channel represents the direction of movement.

However, this hybrid visualization presents the following problems. First, available velocity information is limited to proper motion data which may suffer from drastic projection effects. Large stellar populations such as the Meingast 1 stream [51] show significant distortions in proper motion which can lead users to misguided decisions. Second, trained OCSVM models are bound by given training data. Thus, inferred stars will largely have similar velocities which eliminates random background noise that can cause a visual pop-out effect. Third, not only is the angle channel less accurately perceived as the positional channel [49] but it also lacks an absolute scale. Due to variable star cluster positions and projection effects, changes in angle do not carry an unambiguous meaning.

Users have to judge star clusters by considering their positional and kinematic distribution of its members where especially the search for outliers constitutes an essential task. These tasks benefit from the more effective spatial position channel compared to the less accurately perceived angle channel [38]. Combined with discussed projection effect issues we thus refrain from adding velocity information into the positional scatter plot via the angle channel.

6.0.0.3 Why a reduced scatterplot matrix?: The reduced scatterplot matrix, which shows the position and proper motion dimensions separately, was designed to use the available screen space more efficiently. Since it consists of fewer panels than the full scatterplot matrix, it would have the advantage of displaying the individual scatterplots in a bigger size. Both versions were presented to astronomy experts in the course of iterative prototyping.

6.0.0.4 Why silhouettes in addition to points?: An integral part of each step in the workflow is to compare different distributions of stars. This can mean comparing model groups or the final ensemble model to the training set to see if they are a good match or examining the models with the smallest and largest permitted value of each PA to see how much they differ. To facilitate this comparison, the silhouettes of the distributions can provide a summary of their overall shape that is easier to interpret [17].

6.0.0.5 Why scented widgets with heatmaps?: An integral part of the workflow is to set accepted PA ranges that result in a suitable final classifier. To facilitate this task, supplementary information is necessary to help the user make an informed decision about how to best constrain the PA. The corresponding sliders were therefore implemented as *scented widgets* [77], which feature additional visualizations in the form of heatmaps to show the number of models for each PA value. Histograms were considered as an alternative to heatmaps. These would enable the user to read the exact number of models in each bin more accurately. But this comes at the cost of taking up more screen space, since the histograms would need to be shown in an appropriate size to discern the exact length of a bar. However, in this context, communicating the exact number of models in each bin of the PA range is not the goal. Instead, the user should get an idea of the overall distribution of PA values to see if many models are concentrated around a certain range and then set the sliders accordingly. This can be accomplished adequately with the help of heatmaps; therefore histograms would only provide a level of detail that is not necessary in this context at the cost of taking up more screen space.

6.0.0.6 Why histograms and scatter plots to show kinematics?: Kinematic information is used during both training and validation. As discussed, due to largely missing radial velocity measurements, models are trained with two instead of three velocity features. Although star clusters are approximately normally distributed in Cartesian velocity space, the observed 2D velocities, i.e. proper motions, are subject to sometimes drastic projection effects. The observed, potentially highly concave shapes contribute to the difficulties of traditional clustering approaches.

Since very few stars have radial velocity measurements and, thus, 3D velocity information, stellar kinematics is commonly displayed in proper motions space. Typically, proper motion information is displayed in scatter plots as discussed above.

Stars that have 3D velocities are used as model validation. Models that show significantly different 3D velocities than the training set are removed. This information is quantified in the PA "fraction of outliers" which measures the fraction of inferred stars with radial velocities that are outside the $3\sigma$ region in marginal 3D velocity distributions of the training set. To validate models qualitatively, domain experts are tasked to compare the training set distribution against the distribution of inferred star cluster members. To compare the velocity distributions, two design alternatives were considered, scatter plots and histograms.

As discussed above, other designs such as parallel coordinates are unfamiliar to the domain experts and were judged as confusing. Domain experts noted that both design alternatives facilitate the comparison between distributions. Due to the low number of stars, however, users noted that histograms make it easier to reason on the distribution shape. Especially determining if the data are approximately normally distributed, and thus providing means of validating a model, was perceived to be easier with histograms.

Thus, three histograms showing the Cartesian, marginal velocity distributions are provided alongside the HRD to support model validation. We add them to the Model Group Tab and Stability Tab, see sections 5.2.3 and 5.2.5, respectively. In the PA Tab, see Sect. 5.2.4, the velocity histograms are not included as model validation plays a secondary role in this workflow step. Additionally, the summary statistic "fraction of outliers", whose influence the user can interactively explore already supports a quantitative evaluation of 3D velocities.

6.0.0.7 Why histograms for showing hyper-parameters?: The distribution of hyper-parameters for the accepted models could also be presented using the same heatmaps as before. But an additional task is to provide an overview on the OCSVM hyper-parameter distribution at any time. This helps domain experts to gain insights on the effects that model selection via summary statistics has on the model hyper-parameters themselves (**T.5**). Therefore, the chosen visualization type should support displaying multiple distributions at once. When using heatmaps, this can be achieved by juxtaposing several heatmaps to show different distributions [35]. However, length can be judged more accurately than color [53], which would be an advantage of histograms. Instead of juxtaposing several histograms, another option is to add a line corresponding to each distribution that needs to be presented on the same plot as shown in Fig. 8. Superimposing the distributions in this manner also allows for easier comparison between the heights of different bins [41].

## 7 Implementation

The front-end visualization components and interactions are implemented in JavaScript and use d3.js and vue.js. Additional data processing for building the dendrogram and calculating the PA for the trained models has been separated from the front-end and is implemented using python and the web framework Flask.

We made use of the libsvm [15] OCSVM implementation available in scikit-learn [57] library and the Sobol sampling sequence implemented in SciPy [73]. The software is publicly available to foster open science and reproducibility[3].

## 8 Evaluation

In the following, we discuss both formative and summative evaluation steps we performed in the course of this design study.

### 8.1 Formative Evaluations

During progressing from initial paper prototypes to the final implementation, the tool was repeatedly presented to

[3] https://github.com/ratzenboe/uncover-tool

experts in data visualization, statistics, as well as astronomy and subsequently underwent changes based on their feedback.

In the first stage of the design process, the paper prototypes were reviewed in the group of co-authors featuring a visualization expert, a domain expert, and an applied mathematician and statistician. The feedback sessions were held bi-weekly and lasted for 3 months.

After arriving at a final design, the paper prototype was implemented as an interactive wireframe tool which was used during the second review stage. This interactive prototype was then tested and discussed in two interview session with astronomy experts who had no previous involvement in the design process. The two domain experts had different levels of prior knowledge about the underlying algorithm. One test user already had substantial experience using the algorithm and could therefore confidently navigate through the views of the prototype. The domain expert noted that the proposed design would alleviate many challenges she was facing when searching for new member stars. The second user was less familiar with the inner workings of the algorithm, but with the help of additional explanations it was possible to correctly interpret the visualizations and carry out the associated tasks. These interviews suggest that additional documentation for the final tool would be helpful. The user tests also resulted in a number of feature requests, which were taken into consideration when creating the implementation of the final prototype. For a more detailed description of the prototyping process see Sect. 2 of the supplemental material.

### 8.2 Summative Evaluations

To evaluate the usability of the tool, the final implementation was tested by nine domain experts in astronomy. Three of the participants were experts in the field of stellar clusters while the remaining six test users classified their knowledge as intermediary level knowledge of the subject. All test subjects had previous experience in validating stellar clusters via the HRD and 3D velocities. Six of the participants had no previous experience using the algorithm, the other three test users had worked with the algorithm at least once and were familiar with the basic properties of it. One of the users had already tested the interactive prototype in a formative test, the remaining users were new to the tool.

Each test user was given 60 minutes to test the tool. Every session started with a brief introduction to provide some information on the aim of the test as well as the algorithm itself. The participants were then asked to use the tool and instructed to "think-aloud" while doing so. Additional explanations for the individual steps were provided upon request.

All users tested the tool with the same training set as well as a subset of the Gaia DR2 catalogue as the prediction set and were tasked with finding new member stars for the given training data. Since the main purpose of these tests was to assess the usability of the tool and creating a suitable prediction for a stellar cluster might take more refinement than was possible during the given time, the resulting outputs were not checked for their correctness.

The last 20 minutes of each test were reserved for filling out the SUS-questionnaire [6] as well as conducting a short interview. The resulting SUS-score was 78.06 with a standard deviation of 7.89, which would indicate acceptable usability [3].

Participants, who were inexperienced with the underlying algorithm, mentioned that providing more information and explanations as part of the tool would be helpful. Specifically, the statistical foundations of the PA and stability were deemed as hard to interpret without additional explanations. The dendrogram was considered the least intuitive visualization component by test users regardless of their experience level with the algorithm. All users requested extra explanations but after its purpose and use was explained, the information it provides was deemed very helpful by all participants, for more details see supplemental material Sect. 1. The intended purpose of the remaining views was more straightforward to understand without requiring supplementary clarifications. The overall workflow and sequence of steps was judged as well thought-out. They fully cover the necessary functionality for the required data analysis according to all test users. All participants considered the tool a helpful addition to the algorithm and stated that they would prefer it rather than working directly with the algorithm. This suggests that our main goal for the tool was fulfilled. One test user, who had made use of the algorithm before, also expressed interest in using the tool for their future work.

## 9 Scientific Use Cases

In this section, we showcase the efficiency and effectiveness of the Uncover interface in finding new stars to a given stellar cluster in a case study and use case, respectively. More details on the interactive session discussed in the case study can be found in screenshots throughout this paper and in the accompanying video.

### 9.1 Case Study: Searching for New $\rho$ Oph Members

Recently, Grasser et al. [37] have detected over 100 new member stars for the $\rho$ Oph cluster using an ensemble of OCSVM models. Following model selection ideas from Ratzenöck et al. [59] the authors had to limit the result space via prior assumption ranges that models have to adhere to. Since the $\rho$ Oph cluster has been thoroughly investigated in multiple earlier studies [8], [25], [60] their search for new members was highly uninformed. Due to the lack of substantial prior knowledge Grasser et al. had to resort to randomly sampling different prior assumption ranges and analyze the results manually. In the following we repeat this study, using the same training and prediction set, and showcase a more efficient workflow using Uncover.

The target user is an astronomer who aims to find additional sources in the $\rho$ Oph cluster. Upon starting the tool, the user specifies her prior knowledge on the yet unidentified stellar population via range sliders, see Fig. 4. Since the $\rho$ Oph cluster has been studied extensively in the past, she suspects to find new members predominantly outside the currently known cluster region. She limits "positional extend" and "velocity dispersion" to $0.5 - 2$ and "number of predicted members" to $1 - 10$ times the training set size. Having no specific prior knowledge on limiting other

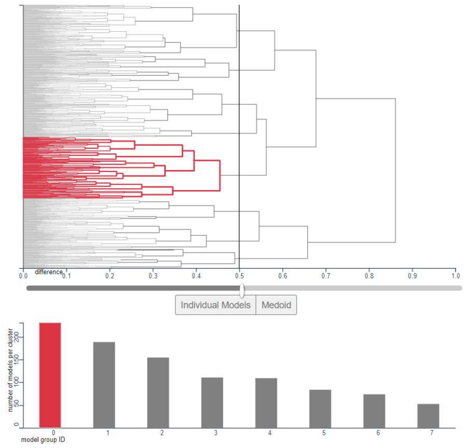


Fig. 9. Dendrogram with corresponding slider and bar chart which shows the number of model groups and their respective size.

summary statistics she leaves the remaining sliders at their initial position (see Fig. 4). On clicking close she arrives at the next workflow step.

In the *Dendrogram Tab* the astronomer aims to group multiple models into a meaningful unit. She explores various difference thresholds in the dendrogram via the slider interface. Looking for a sensible clustering of models she clicks on the leftmost bar in the bar chart to inspect its individual group members in the scatterplot matrix, see Fig.9. She notes that models within each group – represented as silhouettes – highlight very different characteristics of the $\rho$ Oph cluster. To find more meaningful model groups which capture a single model characteristic she gradually decreases the differences threshold while inspecting silhouettes of corresponding group members. At the normalized SDC threshold of $0.15$ she stops her search (see Fig. 1a); not only does she find low variation between individual silhouettes and the group medoid, but also different model groups seem to capture different aspects of the $\rho$ Oph cluster.

In the *Model Group Tab* the astronomer is tasked with assessing the goodness of previously defined model groups. She analyzes the distribution of inferred members in the space of position and proper motions as well as the HRD and three velocity histograms. She finds that the first three models show a large scatter in the HRD and velocities, see Fig.10. Co-evolving stellar groups show a distinguished narrow and well-defined sequence in the HRD as well as a roughly Gaussian distributed 3D velocity. Thus, such increased scatter indicates a large contamination fraction in the sample. Consequently, she marks these groups as "bad".

The next few models are the most interesting ones. They show a second population near the training set, highlighted in gray, in position and kinematic space, but the HRD and 3D velocities indicate a good model. The astronomer conjectures that she just uncovered a second stellar population right next to $\rho$ Oph cluster (see Fig. 5), which Grasser et al. [37] recently discovered. She assesses models featuring this second population as "good". Model groups thereafter do not capture the adjacent population and are thus rejected by the astronomer.

In the *PA tab* the user observes that her initially defined ranges on summary statistics have been updated based on the distribution of accepted models. All updated ranges are rather concentrated towards larger values. By interactively changing the minimum and maximum position of the range slides she learns that the PA "velocity dispersion" and "velocity shift" have a stark influence on the second population as well as the distribution of inferred stars in the HRD. Models with both a lower velocity dispersion and shift are not able to infer stars from the second population. By clicking through the heatmap bins (see Fig. 7) the astronomer finds that a very large fraction of outliers does not correlate with a large "fraction of outliers" scores indicating that these ranges are a good selection. To study the influence of the positional extent statistic on the inferred stars, the astronomer clicks on the corresponding heatmap row. The scatterplot matrix now highlights in dark blue possible stars that can be inferred at the maximum slider position. She increases the slider position and sees a large increase in scatter in position and the HRD. She conjectures that this selection criterion correlates with an increasingly contaminated sample. Thus, she reduces the maximal slider again to exclude likely non-cluster members. By clicking on the next tab she arrives at the *Stability Tab*.

The astronomer explores the influence of various stability thresholds by brushing the line graph representing the 3D velocity dispersion on the right-hand side. She observes a sudden drop in scatter around the training sources in the HRD at about $85\%$ at which coincides with a rapid drop in the 3D velocity dispersion (see Fig 1d). At this threshold both populations seem to be perfectly separated. Not only does the second population, colored in red, show an older age indicated by a shift in the HRD, its 3D velocity distribution is also slightly shifted compared to the training set.

A comparison with the results reported by Grasser et al.[4] yields a $93.3\%$ recall and a relative percentage difference in detected stars of only $3.8\%$[5]. These findings highly coincide with their validated study results [37], which the user was able to replicate with ease in a single session using Uncover.

Finally, she exports the final model with a click on the "Export Final Classifier" button.

### 9.2 Use Case: Finding New Corona-Australis Members

Uncover was used to discover previously unknown members of the Corona-Australis cluster. Due to its proximity and young age Corona-Australis is an important laboratory for studying the star formation process. We chose Corona-Australis specifically, as our collaborators at the Astronomical Institute are interested in finding the most complete sample of the star cluster for follow-up studies. The stellar content of Corona-Australis has recently been studied by

[4] The catalog is publicly available at: https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/652/A2

[5] The results are compared by applying a quality filter in accordance to Grasser et al. [37].

# 4. Clustering analysis (B)

The main research question of this Chapter concerns: "How to consistently search for stellar populations?" And: "How to design interpretable methodologies that robustly find arbitrary shaped clusters in a sea of noise?" To answer these questions this thesis has developed an innovative clustering technique, `SigMA`, which aims to extract clusters with a measure of significance that can scale to millions of data points.

In the following publication (B1) the clustering pipeline `SigMA` is presented. Its main methodological approach is to build clusters from the bottom up. First `SigMA` locates initial cluster candidates in the data set and subsequently iterates over them, merging them in the process if a statistical hypothesis test on uni-modality cannot be rejected. This work adapted the modality test procedure proposed by Burman and Polonik (2009) [9] to handle randomly shaped clusters, astrometric uncertainties, and background noise from a field star population, taking projection effects into account. To increase the robustness of the result, the identified clusters are tracked through a space of smoothed density fields where clusters are identified as structures that persist for a long time.

To showcase the performance of our clustering technique, this study applied `SigMA` to the Scorpius-Centaurus OB association[1]. Although Sco-Cen is the closest and best-studied OB association to Earth this work has uncovered yet unseen detail in our kinematic and positional study. Although Sco-Cen is traditionally subdivided into three main groups, it found 48 groups of co-spatial and co-moving young stars that can each be validated via a narrow and well-defined sequence in the HRD. Compared to earlier studies on the region, `SigMA` is able to un-mix populations previously thought to be single populations. This provides the basis for assigning precise ages to individual groups of stars for the first time ("high-resolution age dating").

[1] The association gets its name from the spectral class of it's most prominent members, O and B stars.

## 4.1. Significance Mode Analysis (SigMA) for hierarchical structures: An application to the Sco-Cen OB association

### Full publication details

**Ratzenböck, S.**, Möller, T., Großschedl, J., Alves, J., Bomze, I. & Meingast, S. 2022. "Significance Mode Analysis (`SigMA`) for hierarchical structures: An application to the Sco-Cen OB association"
*Major revision decision with Astronomy & Astrophysics in May 2022.*

### Author contributions

The paper is co-authored by me, Torsten Möller, Josefa Großschedl, João Alves, Immanuel M. Bomze, and Stefan Meingast. As the leading author, I conceived and developed `SigMA`, performed the analysis, and wrote the first part of the paper including the methodology section. The second part of the paper - the result comparison - was mainly performed by Josefa Großschedl, who also wrote most of it. Torsten Möller supervised the methodological development and provided suggestions and feedback along the way. João Alves supervised the domain application also contributed to the writing and interpretation of results and discussion. Immanuel M. Bomze and Stefan Meingast helped revise the final version.

### Information on the Status

Submission date: 31 March 2022
Major revision decision: 10 May 2022
Accepted: TBD

*Astronomy & Astrophysics* manuscript no. output
March 31, 2022

©ESO 2022

# Significance Mode Analysis (SigMA) for hierarchical structures

## An application to the Sco-Cen OB association

Sebastian Ratzenböck[1,2,3], Torsten Möller[2,3], Josefa E. Großschedl[1], João Alves[1,2], Immanuel Bomze[2,5], and Stefan Meingast[1]

[1] University of Vienna, Department of Astrophysics, Türkenschanzstraße 17, 1180 Vienna, Austria
e-mail: sebastian.ratzenboeck@univie.ac.at
[2] University of Vienna, Data Science at Uni Vienna Research Platform, Austria
[3] University of Vienna, Faculty of Computer Science, Währinger Straße 29/S6, A-1090 Vienna
[4] University of Vienna, ISOR/VCOR, Oskar-Morgenstern-Platz 1, A-1090 Vienna

Received March 31, 2022; accepted...

### ABSTRACT

We present a new clustering method, Significance Mode Analysis (SigMA), to extract co-spatial and co-moving stellar populations from large-scale surveys such as ESA Gaia. The method studies the topological properties of the density field in the multidimensional phase space. We apply the new method to Gaia EDR3 data of the closest OB association to Earth, Scorpio-Centaurus (Sco-Cen), and find about $10^4$ co-moving young objects, about 7% of these sub-stellar. SigMA finds 48 co-moving clusters in Sco-Cen. These clusters are independently validated by their narrow HRD sequences and, to a certain extent, by their association with massive stars too bright for Gaia, hence unknown to SigMA. We compare our results with similar recent work and find that the SigMA algorithm recovers richer populations being able to distinguish clusters with velocity differences down to about 0.3 km/s and reaching cluster volume densities as low as 0.01 stars/pc$^3$. The 3D distribution of these 48 coeval clusters implies a larger extent and volume for the Sco-Cen OB association than typically assumed in the literature. Additionally, we find the association to be more actively star-forming, and dynamically richer than previously thought. We confirm that the mostly star-forming molecular clouds in the Sco-Cen region, namely, Ophiuchus, L134/L183, Pipe Nebula, Corona Australis, Lupus, and Chameleon are part of the Sco-Cen association. The application of SigMA to Sco-Cen demonstrates that advanced machine learning tools applied to the superb Gaia data will allow an accurate census of the young populations, quantify their dynamics, and reconstruct the recent star formation history of the local Milky Way.

**Key words.** Methods: data analysis – (Galaxy:) open clusters and associations: – individual: Sco-Cen – (Galaxy:) solar neighborhood – ISM: clouds

## 1. Introduction

The ESA Gaia mission (Gaia Collaboration et al. 2016, 2018, 2021a) is transforming our knowledge of the local Milky Way, in particular considering the distributing of young stellar populations. However, disentangling and extracting coeval populations remains notoriously difficult. This is reflected in the wide variety of methods applied to the Gaia data (e.g., Oh et al. 2017; Kushniruk et al. 2017; Zari et al. 2017; Castro-Ginard et al. 2018; Cantat-Gaudin et al. 2018; Galli et al. 2018; Zari et al. 2019; Damiani et al. 2019; Meingast et al. 2019b; Kounkel & Covey 2019; Chen et al. 2020; Hunt & Reffert 2021; Olivares et al. 2021; Meingast et al. 2021). The wide range of approaches in the literature reflects the rather complex feature space from where the stellar populations are extracted. Firstly, as a consequence of interactions with the Milky Way potential, spiral arms, and giant molecular clouds, these initially compact objects are stretched into elongated, sometimes concave structures in position space (e.g., Kamdar et al. 2021). This "galactic-stretching" leads to a variety of cluster[1] shapes from compact (when young), to low-contrast, spread-out, sometimes S-shaped clusters dominated by Milky Way tidal forces (e.g., Meingast & Alves 2019a; Röser et al. 2019; Meingast et al. 2019b; Beccari et al. 2020; Kounkel & Covey 2019; Jerabkova et al. 2019; Ratzenböck et al. 2020; Meingast et al. 2021; Jerabkova et al. 2021; Kerr et al. 2021; Kamdar et al. 2021). Secondly, due to the low number of available radial velocities, about 0.4% in Gaia DR2 and EDR3 database (Gaia Collaboration et al. 2018, 2021a), one is, for the most part, restricted to two tangential velocity axes plus the spatial three coordinate axes as derived from Gaia positions, parallaxes, and proper motions (5D phase space). Thus, even under the assumption of perfectly Gaussian distributed 3D velocities within clusters, the projection on the sky distorts the multivariate Gaussian (5D space) into arbitrary shapes depending on the orientation, distance and size of the stellar cluster. To make matters worse, stellar cluster members constitute a minute subset of the Gaia data, with unrelated field stars creating a background noise that is not easily removable in the 5D space. The feature space consists of stellar clusters of various shapes and densities embedded in a sea of noise.

To tackle the challenge of identifying sub-populations in a star-forming region, we developed a method that analyses the topological structure of the 5D density field spanned by 3D positions and tangential velocities. We apply a fast modality test procedure, which introduces a measure of significance to peaks in

[1] In this paper, we use the word "cluster" in the statistical sense, namely, an enhancement over a background. This avoids creating a new word for the spatial/kinematical coherent structures we find in Sco-Cen. None of the Sco-Cen clusters is expected to be gravitationally bound.

the density distribution, thus, providing an interpretable cluster definition. This clustering method is called Significance Mode Analysis, or `SigMA`, and it is designed to extract co-spatial and co-moving stellar populations from large-scale surveys such as ESA Gaia.

The goal of this paper is to present the `SigMA` method, apply it to the Scorpius-Centaurus OB association (Sco-Cen, Blaauw 1946, 1952, 1964a,b) to identify the different sub-populations, and compare results with recent papers attempting similar goals. Sco-Cen is the closest and best studied OB stellar association (e.g., de Geus et al. 1989; de Geus 1992; de Zeeuw et al. 1999; de Bruijne 1999; Preibisch & Zinnecker 1999; de Zeeuw et al. 2001; Lépine & Sartori 2003; Preibisch & Mamajek 2008; Makarov 2007, 2008; Diehl et al. 2010; Pöppel et al. 2010; Rizzuto et al. 2011; Pecaut et al. 2012; Pecaut & Mamajek 2016; Forbes et al. 2021), with an age $\lesssim 20$ Myr (Pecaut, Mamajek, & Bubar 2012). These and many other papers in the literature have established Sco-Cen as an important laboratory for star formation, for the characterization of stellar associations, and for understanding the impact of massive stars on the ISM and planet formation. Since the advent of large-scale astrometric data from the ESA Gaia mission that started in 2016 (Gaia Collaboration et al. 2016), there has been a renewed interest on this benchmark region focusing on the kinematics and 3D structure of the association (Villa Vélez et al. 2018; Wright & Mamajek 2018; Goldman et al. 2018; Damiani et al. 2019; Luhman & Esplin 2020; Grasser et al. 2021; Squicciarini et al. 2021; Schmitt et al. 2021; Kerr et al. 2021; Luhman 2022a).

In this paper we present the method `SigMA` in Sect. 3, using Gaia EDR3 data (Sect. 2), with an application on Sco-Cen discussed in Sect. 4, including comparisons to previous work (Sect. 4.2). In Sect. 5 we give a summary of our findings.

## 2. Data

In this work we apply the newly developed method presented in this paper, `SigMA`, to Gaia data of the Sco-Cen OB association. We select a box of about $10^7$ pc$^3$ from the Gaia EDR3 Archive (Gaia Collaboration et al. 2021a), which extends well beyond the traditional and well studied Sco-Cen regions. Several hints in the literature suggest that the Sco-Cen OB association is a larger complex than traditionally defined by Blaauw (1946) and outlined by de Zeeuw et al. (1999), and it includes several star-forming regions that have originally not been assigned to Sco-Cen (e.g., Lépine & Sartori 2003; Sartori et al. 2003; Bouy & Alves 2015; Kerr et al. 2021; Zucker et al. 2022). The box is defined in a Heliocentric Galactic Cartesian coordinate frame (XYZ) within:

$$\begin{aligned} -50\,\text{pc} &< \text{X} < 200\,\text{pc} \\ -200\,\text{pc} &< \text{Y} < 50\,\text{pc} \\ -95\,\text{pc} &< \text{Z} < 80\,\text{pc} \end{aligned} \tag{1}$$

The 3D space positions (XYZ)[2] are derived from the Gaia EDR3 positions right ascension ($\alpha$, deg) and declination ($\delta$, deg), and the parallax ($\varpi$, mas). The distance ($d$, pc) is derived from the inverse of the parallax, which is a fairly good approximation of the distance for sources within 200 pc and with low errors[3].

The clustering is done in a 5D phase space, using the 3D spatial coordinates XYZ in pc, and the 2D tangential velocities $v_\alpha$ and $v_\delta$ in km s$^{-1}$. The different dimensions are scaled to each other, as described in Sect. 3.3.3. The proper motions ($\mu^*_\alpha = \mu_\alpha \cos(\delta)$, $\mu_\delta$) are transformed from mas yr$^{-1}$ to tangential velocities in km s$^{-1}$ as follows:

$$\begin{aligned} v_\alpha &= 4.74047 \cdot \mu^*_\alpha / \varpi \\ v_\delta &= 4.74047 \cdot \mu_\delta / \varpi \end{aligned} \tag{2}$$

We do not use the third velocity dimension, radial velocity ($v_r$), since Gaia only includes radial velocity measurements for about 0.5% of the sources with parallaxes. Adding auxiliary radial velocity data would improve the statistics, but it would constitute a very inhomogeneous data sample with 6D phase space information. Therefore, we restrict our clustering procedure to the 5D phase space, as provided by Gaia, allowing us to create a homogeneous and more complete overview of the existing clusters in regions like Sco-Cen. Moreover, by focusing on the 5D phase space, we are able to create a method that does not rely on radial velocities, which then can be used more widely on larger data samples. For validation purposes, in Sect. 3.5 we use Gaia DR2 radial velocities[4] (Cropper et al. 2018; Gaia Collaboration et al. 2018, 2021a) to remove noise.

To reduce the influence from spurious measurements, we apply the following quality criteria to the Gaia EDR3 data within the selected box:

$$\begin{aligned} &\text{fidelity\_v2} > 0.9 \\ &\varpi > 0\,\text{mas} \\ &\text{e\_}\varpi/\varpi < 0.2 \\ &\text{e\_}\mu^*_\alpha,\ \text{e\_}\mu_\delta < 2\,\text{mas/yr} \end{aligned} \tag{3}$$

The parameter fidelity_v2 is a classifier to identify spurious sources in the Gaia EDR3 catalog, developed by Rybizki et al. (2022), which can be used to select high fidelity astrometry. The parallax-error and proper-motion-error cuts reduce additional uncertainties in distance and velocities. This leaves 451,127 sources inside the box to which we apply the `SigMA` clustering algorithm, as described in the following Sect. 3. See also Appendix A for details on the data retrieval.

## 3. Methods

In this section, we first give a brief overview of the basic definitions of several widely used clustering algorithms, which leads to detailed explanations on the buildup of the Significance Mode Analysis clustering algorithm (`SigMA`) in Sect. 3.2, as developed in this work.

### *3.1. Clustering algorithms: a brief review*

Understanding the Milky Way, or any object in the Universe is directly linked to the quantity and quality of the available data. Paradoxically, the advent of large, high-dimensional data has led to an apparent problem: the more information we have, the less we seem to be able to grasp the big picture hidden in the data. "Big data" usually contain extensive information, diversity, and complexity and, thus, we require more complex methods to model its observations. However, many traditional analysis techniques have time and memory complexities that fail to perform

[2] The observed positions are transformed to XYZ using `astropy.coordinates.SkyCoord` from Astropy `v4.0`.

[3] For more distant sources, or intrinsically faint sources with high parallax errors, the distance estimate becomes a non-trivial inference problem (e.g., Luri et al. 2018; Bailer-Jones et al. 2021)

[4] The Gaia EDR3 catalog includes the DR2 radial velocities, while updated RVs will be provided in DR3, increasing the RV sample by about a factor 4.6.

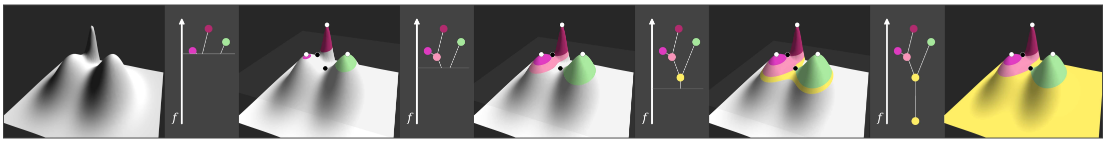


**Fig. 1.** Merge tree generation. Via a continuous change of $\lambda$ from $\infty$ to $-\infty$ a new component is created at each maximum (white points). At each saddle point (black points) components are merged. The merge tree is fully computed when $\lambda$ reaches the global minimum.

under millions or even billions of data samples (Ashok Kumar 2020). Consequently, many studies start with an exhaustive pre-filter step to improve downstream analyses (e.g. Zari et al. 2019; Kerr et al. 2021).

To escape this paradox, new interpretive methods need to be tailored to the particular scientific question, in our case, the identification of co-moving and coeval groups of stars inside the 1+ billion stars in the Gaia archive. Clustering analysis, or unsupervised machine learning, has recently become essential to the identification of coeval stellar structures. The goal of clustering is to obtain an organization of data points into meaningful groups. However, due to the lack of labeled data, partitioning into "meaningful" clusters is in general an ill-posed problem. The choice of the algorithm and its parameters have to match the problem at hand. This constraint applies especially to parametric clustering algorithms.

### 3.1.1. Parametric clustering

Parametric clustering algorithms are appealing because of the probabilistic interpretation of the clusters these algorithms generate. The model-based approach introduces a finite mixture of density functions of a given parametric class. The clustering problem reduces to the parameter estimation of the mixture components, which is typically done using the expectation-maximization (EM) algorithm (Dempster et al. 1977). The EM algorithm tries to find maximum likelihood estimates of given parameters iteratively. A popular approach is to model the mixture components as multivariate Gaussian density (e.g. Gagné et al. 2018; Cantat-Gaudin et al. 2019).

A considerable downside of parametric clustering algorithms is that they will try their best to fit the model to the data even if none exists. In the case of stellar populations, the range of possible shapes of signal and noise, especially considering only the projected 2D velocity information, cannot be modeled accurately by simple distributions, such as multivariate Gaussians.

Moreover, the number of mixture components is unknown, and model selection methods such as Akaike Information Criterion (AIC, Akaike 1974) and Bayesian Information Criterion (BIC, Schwarz 1978) only work well in cases with plenty of data samples, well-separated clusters, and a well-behaved background distribution (Hu & Xu 2003). These circumstances make extracting clusters with a low signal-to-noise ratio difficult, especially in the low-density regime.

### 3.1.2. Non-parametric, density-based clustering

The premise of non-parametric density-based methods states that the observed data points[5] $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ are drawn from an unknown density function $f$. The goal of non-parametric cluster analysis is then to understand the structure of the underlying density function, which is estimated from data. In one of the earliest formulations, Wishart (1969) argues that clusters are data samples associated with modes in $f$. The work proposed by Koontz et al. (1976) and the widely used Mean-Shift algorithm and its variants (Cheng 1995; Comaniciu & Meer 2002; Vedaldi & Soatto 2008) are examples of this *mode-seeking* category.

Mode-seeking methods proceed to group the data by locating local peaks in $f$ and their corresponding attraction basins. Attraction basins are regions in which all gradient trajectories converge into one single peak. However, the gradients and modes are highly dependent on the density function approximation $\hat{f}$. In order to increase the robustness of the result, Mean-Shift seeks to reduce random fluctuations by employing a smoothing kernel to $\hat{f}$. The introduction of an extra parameter shifts the issue to the user, who is tasked to carefully select the non-intuitive smoothing factor in order to obtain a satisfying clustering result. Moreover, the time complexity of at least $O(N^2)$ makes them not great candidates for application to astronomical data sets.

Hartigan (1975) proposed a similar definition of clustering in which a cluster is defined as the connected components of the level-sets[6] of $f$. Given a data set $X$ drawn from an unknown density function $f$ which has compact support $\mathcal{X}$ we can formally write the resulting level-sets for the threshold $\lambda$ as:

$$L(\lambda) := \{\boldsymbol{x} \in \mathcal{X} \;:\; f(\boldsymbol{x}) \geq \lambda\} \tag{4}$$

Thus, $L(\lambda)$ constitutes a set of connected components which we identify as clusters.

In the level-set framework, popular clustering algorithms such as DBSCAN (Ester et al. 1996) can be simply thought of as a single level which is obtained by fixing $\lambda$. DBSCAN avoids estimating the data density explicitly, by employing a radius parameter, usually called $\epsilon$, along with a minimum number of points parameter, `min_points`. Clusters are defined as connected regions of points that contain at least `min_points` within $\epsilon$-sized shells around them.

The connected components of the level-set $L(c)$ are the resulting clusters while the remaining data is treated as noise. However, the choice of the parameter $\lambda$ which is related to DBSCAN's $\epsilon$ parameter, is ambiguous, a task which gets especially challenging when the number of clusters varies greatly between levels. We find a reflection of this difficulty in choosing the right parameters in the astronomical literature, which employs a variety of different heuristics to select the parameter $\epsilon$ (e.g. Castro-Ginard et al. 2018; Zari et al. 2019; Fürnkranz et al. 2019; Hunt & Reffert 2021).

For many data sets containing clusters with variable densities, employing a single threshold $\lambda$ cannot reveal all peaks in $f$. A hierarchy of clustering solutions can be obtained by considering all possible threshold values at once.

[5] In the following, bold, lower-case variables denote $d$-dimensional vectors.

[6] Often also referred to as superlevel-sets.

### 3.1.3. Hierarchical, density-based clustering

The strength of level-set formulation now lies in the natural emergence of a cluster tree, a clustering hierarchy which arises from sweeping the density threshold $\lambda$ from $\infty \to -\infty$. Under a continuous change of $\lambda$, the number of connected components changes when the threshold passes through a critical point in $f$, thus $\nabla f = \mathbf{0}$. A new cluster is born when $\lambda$ reaches the height of a mode in $f$. On the other hand, a cluster dies when $\lambda$ traverses a saddle point or a local minimum, in which case the two connected components merge into a single one. The cluster creation and merging process is schematically shown in Fig. 1.

However, estimating the connected components of level-sets, while easy in one dimension, gets nontrivial in higher dimensions. Consequently, algorithmic realizations of the Hartigan (1975) level-set idea rely on graph heuristics and graph theory in which connected components arise naturally. Early implementations by Azzalini & Torelli (2007) and Stuetzle & Nugent (2010) and subsequent theoretical analyses (Chaudhuri & Dasgupta 2010; Kpotufe & von Luxburg 2011; Chaudhuri et al. 2014) adopt a graph $G(\lambda)$ over the data samples where vertices and/or edges are filtered according to $\lambda$, thus $\{\boldsymbol{x} \in X \ : \ \hat{f}(\boldsymbol{x}) \geq \lambda\}$[7].

However, the use of graphs to represent the connectivity comes with its own limitations. This scheme guarantees that two samples from one connected component of $G(\lambda)$ are to be found in a connected component in $L(\lambda)$. However, as Stuetzle & Nugent (2010) point out, the reverse implication is not necessarily given. This means, samples from the same connected component in $L(\lambda)$ may end up in different connected components of $G(\lambda)$. Since density estimates are inherently noisy, usually too many clusters arise from this iterative filtration procedure. To counteract this over-clustering, the resulting graph cluster tree is usually pruned in a post processing step during which spurious clusters are identified and merged back into the "mother cluster" (Stuetzle & Nugent 2010; Kpotufe & von Luxburg 2011; Chaudhuri et al. 2014).

### 3.1.4. The HDBSCAN algorithm

A well-known algorithm belonging to the family of hierarchical level-set methods is the HDBSCAN algorithm (Campello et al. 2013) (Hierarchical Density-Based Spatial Clustering of Applications with Noise), which recently has been gaining attention in the astronomical community (e.g., Kounkel & Covey 2019; Kounkel et al. 2020; Hunt & Reffert 2021; Kerr et al. 2021). In order to prevent over-clustering, the authors introduce the *minimum cluster size* parameter which provides an interpretable pruning strategy.

At each cluster split decision, the smaller cluster created is merged back into the "mother cluster" if it has less than *minimum cluster size* points, otherwise a new cluster is created. To obtain a flat clustering result from the cluster tree, HDBSCAN estimates the stability of a cluster in the hierarchy via the concept of *relative excess of mass* (EOM). Similar to the concept of excess mass (Muller & Sawitzki 1991), it measures the lifetime and size of a cluster. The heuristic favors more prominent and stable clusters that live longer in the cluster tree. For example, a group that persists for a long time as a single connected component should be preferred over the two small clusters it breaks into that quickly vanish.

However, the EOM criterion tends to produce too large clusters in practice. If a large group persists in the hierarchy for a long enough time, its children are unlikely to exceed the parent's EOM. Alternatively, the HDBSCAN implementation by McInnes et al. (2017) offers the opportunity to extract the leaf nodes from the cluster tree. Since the leaf nodes are extracted only considering the *minimum cluster size* criterion, the resulting clusters lack any stability guarantee; thus, the clustering result is highly susceptible to random density fluctuations. In general, these methods suffer from complex and hard-to-interpret pruning procedures and parameters, which affects the confidence and interpretability of the clustering result.

### 3.1.5. Topological methods

Extracting a flat clustering from the cluster tree requires a notion of cluster stability. As discussed, the concept of relative excess of mass, which inherently depends on the pruning process, can lead to too coarse clusters. A related pruning heuristic comes from considering the topological persistence of each mode in $\hat{f}$, introduced by Chazal et al. (2013). Persistence is defined as the lifespan of each connected component. The notion of persistence is shown to be stable under small perturbations to the inital density $f$ (Edelsbrunner et al. 2000; Zomorodian & Carlsson 2005; Ghrist 2008).

A variation on the persistence formulation is proposed by Ding et al. (2016), who instead of thresholding the cluster lifetime, use cluster *saliency* $\nu$, defined by the ratio of birth and death density, as a cluster stability criterion. By varying $\nu$ between 0 and 1 the cluster tree is revealed and the most stable and long-lived configuration is chosen as an appropriate clustering result.

While easy to interpret, these stability parameters can get quite tedious to select in practice. In the large data and cluster regime, the separation between stable and unstable clusters becomes less apparent. In these limiting cases, selecting the input parameters again warrants a proper parameter search.

### 3.1.6. Extracting stable and significant clusters

Compared to the notion of persistence, there is also growing research to apply statistical methods that test the modality structure of the data. These methods offer the advantage of an interpretable and meaningful parameter $\alpha$, defining the significance level of a corresponding hypothesis test. The null hypothesis $H_0$ commonly assumes that the data, or subsets of it, are sampled from an uni-modal density, whereas the alternative hypothesis $H_1$ suggests multi-modality. The null hypothesis is rejected at a significance level $\alpha$ if the p-value from the corresponding test procedure exceeds this significance level.

We identify first applications of hypothesis test procedures in the clustering literature in the context wrapper methods around the $k$-means and EM frameworks. G-means (Hamerly & Elkan 2004) employs the Anderson-Darling statistic to test the hypothesis that each cluster is generated from a Gaussian distribution. Instead of testing on a per-cluster basis, Pg-means (Feng & Hamerly 2007) tests the whole Gaussian mixture model (GMM) at once. Dip-means (Kalogeratos & Likas 2012) proposes an incremental clustering scheme for selecting $k$ in $k$-means which employs Hartigan's dip statistic (Hartigan & Hartigan 1985). In case the distance distribution of one or more points to their cocluster members exhibit a significant multimodal structure, the cluster is split.

[7] Edges are commonly assigned the minimum density sampled along the path connecting two vertices.

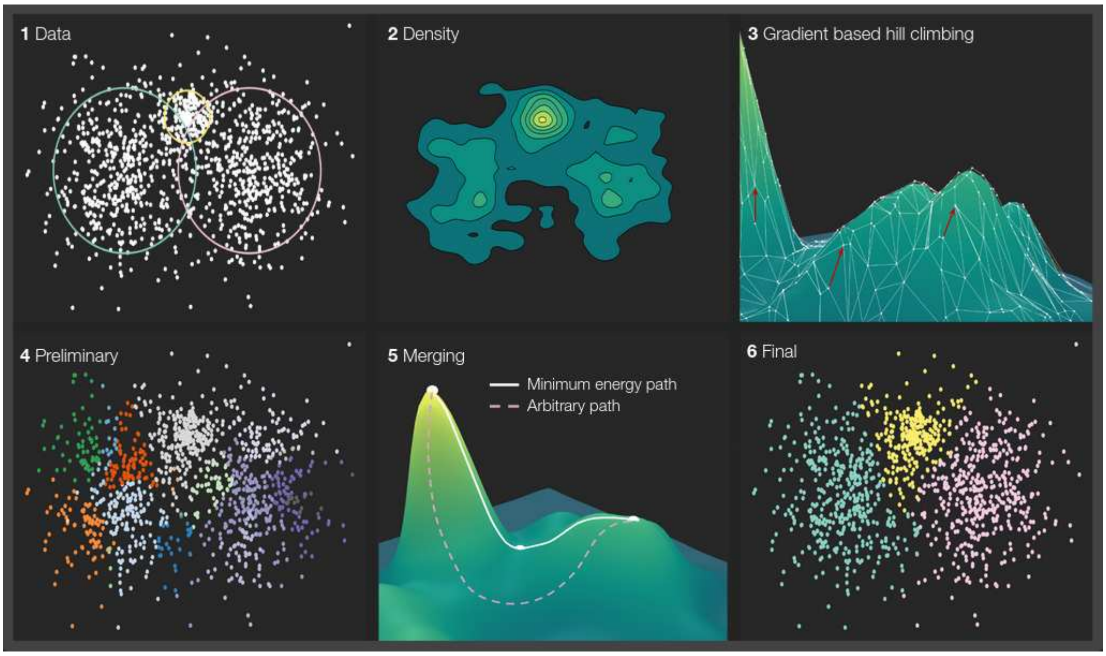


**Fig. 2.** The proposed clustering process `SigMA`, highlighted on a 2D toy data set of three Gaussians with variable covariance matrices and means. **(1)** The generated toy data set consisting of three bivariate Gaussians is shown in white alongside $2\sigma$ confidence ellipses in color. **(2)** The clustering procedure starts off by estimating the density of the input data. **(3)** Next, a graph-based hill climbing step is performed in which points are propagated along gradient lines towards local peaks. **(4)** This gradient propagation results in a preliminary segmentation of input samples which typically is far too fine-grained. **(5)** These segmented regions are iteratively merged with a parent mode if a modality test along the "minimum energy path" detects no significant density dip. **(6)** The final segmentation retains all three clusters.

Skinny-dip (Maurus & Plant 2016) also implements Hartigan's dip test and applies it to one-dimensional linear projections of the data set. Distinct density peaks are to be identified based on the gradient of the projected cumulative distribution function (cdf). By projecting the data iteratively into multiple axes, the samples are partitioned into clusters. Skinny-dip is specifically able to handle background noise very well, however, it considers noise samples to be uniformly distributed and clusters to be axis-parallel.

These algorithms, however, are intrinsically tied to convex or Gaussian cluster assumptions. The recently proposed M-dip (Chronis et al. 2019) is able to deal with arbitrary oriented and shaped clusters, which applies a simulation strategy to approximate values for smallest density dips of uni-modal data sets of the same size and density. However, we do not want to depend on simulations but instead directly obtain a measure of significance from given data.

### 3.2. `SigMA`: Significance Mode Analysis

This section describes our clustering pipeline, `SigMA`, which builds on several established methods, as described above. `SigMA` is tuned to astrometric data provided by Gaia, and aims at producing astrophysically meaningful clustering results. Our technique seeks to identify modal regions in the data set (5D phase space) which are separated by dips. By applying a modality test for each pair of neighboring modes, we obtain a clustering result with measures of significance. The workflow is schematically highlighted in Fig. 2. A modal region is defined as the set of points that all end in a particular mode when following the path tangent to the gradient field at each point. It is important to note that modal regions fully segment the data set, as seen in Fig. 2 (panel 6). Thus, modal regions are a mixture of cluster members and field stars, while the field stars will be removed as noise as outlined in Sect. 3.5 and shown in Fig. 4.

#### 3.2.1. A fast modality test procedure

We consider the hypothesis test introduced by Burman & Polonik (2009) which examines the modality structure of a path between two peaks in the density. Conceptually two neighboring peaks are "true" clusters in the data if there exists no path between them that does not undergo a significant dip in density.

Given the $d$-dimensional data $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn from $f$ and any point $\boldsymbol{r}$ on a path connecting two modes $\boldsymbol{c}_i$, $\boldsymbol{c}_j$ in $f$, Burman & Polonik (2009) show that

$$\widehat{\mathrm{SB}}(\boldsymbol{r}) = d\,\sqrt{k/2}\left[\log d_k(\boldsymbol{r}) - \max(\log d_k(\boldsymbol{c}_i),\ \log d_k(\boldsymbol{c}_j)\right] \tag{5}$$

is asymptotically standard normal distributed. Here $d_k(z)$ denotes the distance to the $k$'th nearest neighbor of the point $z$. The null hypothesis of uni-modality is rejected at significance level $\alpha$ if

$$\widehat{\mathrm{SB}}(\boldsymbol{r}) \geq \Phi^{-1}(1-\alpha) \tag{6}$$

where $\Phi$ is the standard normal cdf. For a more thorough derivation of Eq. (5) and Eq. (6) see Appendix B.

Since Eq. (5) processes a single point rather than a complete path, the modality test in Eq. (6) describes a pointwise procedure. Burman & Polonik (2009) employ the test with samples generated along the straight line connecting two modal candidates to determine the modality for an entire path. The null hypothesis is rejected if any single test fulfills Eq. (6). However, this procedure only applies to convex clusters and does not scale well as tens to hundreds of distance computations along each path increase the run-time drastically.

We aim to minimize the number of distance computations while also extending the test procedure to concave cluster shapes. To do so, we analyze the nature of possible connections between modal candidates in the data. Of all possible paths between two peaks, only the "minimum energy path" (MEP) needs to be considered. The MEP is the optimal solution for the problem of finding the continuous path from one peak to another through input space $\mathcal{X}$ with highest minimal density. Thus, the density dip along the MEP is the minimal possible dip that can exist between two neighboring peaks.

Given a set of initial modal candidate regions in $\hat{f}$ the MEP leads over the connecting saddle point when moving from one mode to another. At the saddle point position, the path reaches its global density minimum. Figure 2 (panel 5) schematically illustrates two possible paths, the MEP and a second arbitrary path.

Instead of evaluating the test statistics in Eq. (5) multiple times, we aim to reduce the calculations to a single one. Since the test procedure is dominated by the point $\boldsymbol{s}$ which maximizes the test statistic it needs to be evaluated only at $\boldsymbol{s}$. Due to the test statistics proportionality to the distance $d_k(\boldsymbol{s})$, its value is maximal when the density is minimal.

For two neighboring modal regions the modality test procedure can, therefore, be reduced to a single pointwise test at the saddle point $\boldsymbol{s}$ connecting the two peaks. As the saddle point governs the modality test, we can assign a $p$-value which takes the following form:

$$p = 1 - \Phi\left(d\sqrt{k/2}\left[\log d_k(\boldsymbol{s}) - \max(\log d_k(\boldsymbol{c}_i),\ \log d_k(\boldsymbol{c}_j)\right]\right) \quad (7)$$

Determining the saddle point is discussed in the following section. If all density minima lie on the boundary of modal regions, the saddle point of two neighboring modes lies at their common border. Using this monotonous property assumption, we aim to provide a fast and yet accurate test procedure to examine the modality structure of the data.

### 3.2.2. Identifying and pruning modal candidates

To identify modal regions from the data set $X$ we implement a graph-based, hill-climbing algorithm analogous to Koontz et al. (1976) where the vertex set of the graph $G$ represents the data $X$. The initial modal search is performed in one pass over the vertices of $G$ sorted in descending $\hat{f}$-order.

A data point is defined as a local mode of $\hat{f}$ if all its neighbor connections have lower densities. Alternatively, points are propagated according to their slope in $\hat{f}$. Each point is iteratively assigned to neighbors with maximum $\hat{f}$-value, see Fig. 2 (panel 3) for a schematic illustration. After this pass the data is separated into $m$ disjoint modal sets $\boldsymbol{M} = \{M_1, \ldots, M_m\}$.

Since graph-based hill-climbing procedures are susceptible to perturbations in $\hat{f}$, a second pass is needed to merge insignificant modal regions into their stable parent mode. To determine the merge order we compute the cluster tree of $\boldsymbol{M}$. As described in Sect. 3.1.2, the cluster tree is obtained by varying the density threshold $\lambda$ from $\infty \rightarrow -\infty$ and registering modal regions when $\lambda$ passes through a peak in $\hat{f}$ and their unification when $\lambda$ passes through the respective saddle point. To finalize the cluster tree we need to identify the saddle points between modal regions of $\boldsymbol{M}$.

We determine the saddle point between two modes via an edge search in $G$. Specifically, we consider edges which connect vertices that lie in different modal sets. We assume extracted modal regions are proper ascending manifolds. Thus, the modal regions are devoid of local minima on the inside, which only lie on the border; consequently, saddle points are found at the common boundary of both regions. The "saddle edge" represents the bridge between two modal regions where the density is maximal. We define edge density as the minimum density along the connecting line segment. To account for density dips along the edge path while limiting the number of distance computations, the edge density is set to be the minimum density between its two vertices and the density at the geometric mean of the vertex positions. The corresponding saddle point density between two adjacent modal regions is approximated by this edge density.

The merging of spurious modes then proceeds by iterating over the set of predetermined saddle points sorted in descending $\hat{f}$-value order. At each step, the uni-modality test in Eq. (7) is evaluated and neighboring modal regions are merged if the respective $p$-value exceeds the significance level $\alpha$. Therefore, the significance level $\alpha$ provides an immediate and meaningful way to simplify the initial cluster tree.

## 3.3. Parameter selection

In the following, we discuss various parameter choices which affect the final clustering result. The presented mode seeking methodology is agnostic to the choice of (1) the graph used in the hill-climbing step, (2) density estimator, and (3) scaling factors between positional and velocity features. In the following, we will explain our decisions on these three algorithmic aspects.

### 3.3.1. Graph

The choice of the graph directly affects the gradient approximation. For example, in a complete graph where every pair of vertices are connected via an edge, the graph-based gradient approximation loses its locality meaning entirely. In this case, the hill-climbing algorithm merges each vertex with the densest point in the data set on the first pass. Thus, over-connected graphs lead to clusters that falsely merge numerous distinct modes in the data set.

Conversely, under-connected graphs such as minimum spanning trees restrict the gradient estimation too much, producing vast amounts of spurious clusters. Furthermore, the low number of neighboring vertices greatly restricts the possible paths between two initially formed modes. Thus, under-connected graphs introduce significant errors in determining saddle points, which drastically compromises the validity of extracted modal regions.

We consider *empty region graphs* (ERG) to strike a balance between over and under connecting points in $X$. In an ERG, a vertex between two points is created if a given region around them does not contain any other point, see Jaromczyk & Toussaint (1992) for a review.

The $\beta$-skeleton (Kirkpatrick & Radke 1985) is a one-parameter generalization of an ERG where $\beta$ determines the size of the empty region. For $\beta = 1$ the graph becomes the Gabriel graph (Gabriel & Sokal 1969), while for $\beta < 1$ and $\beta > 1$ edges are added or removed from it, respectively. Correa & Lindstrom (2011) find that critical point searches (important for topological decomposition, clustering, and gradient estimation) are more accurate with $\beta$-skeletons, with $\beta < 1$ compared to $k$-nearest neighbor graphs and the Gabriel graph. Since the number of vertices grows very fast in size as $\beta$ gets smaller we choose a value of $\beta = 0.95$.

Adopting a $\beta$-skeleton on our 5D data we find that points have on average approximately 50 neighbors. To reduce the chance of separate modal regions being connected via vertices and, thus, erroneously merging in the first hill-climbing step, we prune the initially computed graph in a post-processing step. We remove vertices that show a significant density dip as one moves from one vertex to another. For simplicity, we assume that the saddle point lies at the arithmetic mean of the two vertex points.

### 3.3.2. Density estimation

Such as the graph choice, density estimation is a core part of the algorithmic pipeline that affects gradient propagation and, consequently, the initial mode finding step (see panels 2 and 3 in Fig. 2). Since we cannot describe the complex stellar distribution via parametric models, we employ a model-agnostic, nonparametric estimator for the underlying density.

The most popular non-parametric density descriptors are k-nearest-neighbor ($k$-NN) and kernel density estimation (KDE). KDE models the density by convolving the data with a symmetric kernel function. The bandwidth parameter can be thought of as the standard deviation of the kernel, which determines the smoothing effect of convolution. A gradual increase in bandwidth and its impact on the density is shown in Fig. 5. The $k$-NN method takes a more naive approach to estimate the underlying density. The density value at any given point in the phase space is inversely proportional to the distance to its $k$-th nearest neighbor.

The KDE inherits the smoothness properties of the kernel. Thus, the density becomes infinitely differentiable for a Gaussian kernel. Conversely, the $k$-NN density estimate is not smooth and, in fact, not even continuous. Despite its non-continuous nature, the $k$-NN density estimation method has several advantages for modal clustering. Notably, Dasgupta & Kpotufe (2014) show that point modes of a $k$-NN density estimate approximate the true modes of the underlying density function. Further, the approach has efficient implementations due to fast kd-tree queries that provide desirable memory complexity (Bentley 1975). Further, choosing the number of neighbors $k$ is more straightforward than the bandwidth parameter for KDE. Finally, the locality of the k-NN approach provides a versatile method to determine densities when structures exist at different densities scales. Since KDE employs a constant bandwidth, it can only adapt to a single characteristic density scale. A fixed, "intermediate" bandwidth may adequately resolve medium-density clusters when structures are present at various scales. However, fine-grained and large-scale patterns will be over-smoothed or under-smoothed, respectively.

We employ a $k$-NN estimator to approximate the density function considering these advantages. Specifically, we use a density estimator based on the distance to an empirical measure (DTM) described by (Biau et al. 2011). It is a weighted $k$-nearest neighbor estimate which incorporates distances $d_1, \ldots, d_k$ to all nearest neighbors up to $k$. The DTM is a distance-like function robust to the addition of noise and is used to recover geometric and topological features such as level sets. It is defined in the following:

$$d_m(\boldsymbol{x}) = \sqrt{\frac{1}{k} \sum_{\boldsymbol{y}_i \in N_k(\boldsymbol{x})} \|\boldsymbol{y}_i - \boldsymbol{x}\|^2} \tag{8}$$

where $N_k(\boldsymbol{x})$ is the neighborhood point set of $\boldsymbol{x}$ of size $k$. In other words, the distance to empirical measure takes the form of a mean distance from the point $\boldsymbol{x}$ to its $k$ nearest neighbors. The density estimator is defined via the inverse of this quantity:

$$\hat{f}_m(\boldsymbol{x}) = \frac{1}{nV_d} \left( \frac{\sum_{j=1}^{k} j^{2/d}}{k d_m^2(\boldsymbol{x})} \right)^{d/2} \tag{9}$$

where $V_d$ denotes the volume of the $d$-dimensional unit ball and $n$ is the number of data points.

Since in our use case the order of density values is important, we can ignore constant normalization terms in Eq. (9).

The k-NN algorithm is not only used to estimate the density but also during the modality test procedure, see. Sect. 3.2.1. Since classical k-NN, as employed in the modality test, automatically ignores points within its k-distance, SigMA has a built-in limit to the size of structures it can resolve. This allows us to determine a lower bound on the velocity dispersion of a population that SigMA can identify. We find the minimally resolvable velocity dispersion to be 0.3 km $\mathrm{s}^{-s}$ by analyzing the distribution of k-distances with a lower bound on $k = 15$, which we also assume to be the minimum cluster size. Clusters with lower velocity dispersion get smoothed to at least this minimum dispersion. This value increases as $k$ gets larger.

### 3.3.3. Scaling factors

The clustering analysis of co-moving populations in position and velocity occurs in a combined positional and kinematic phase space. Distance relationships among stars are needed to express densities and build a graph from the input data. Since tangential velocities are measured in km $\mathrm{s}^{-1}$ and galactic coordinates in pc, both sub-spaces have different ranges. Significant range discrepancies between dimensions influence the clustering process as it directly impacts the distance function. Individual 1D distance contributions along feature axes with narrow ranges can be ignored when features with large standard deviations are present. Hence, we consider scaling factors between positional and kinematic feature sub-spaces.

Scaling factors $c_i$ put weight on specific sub-spaces to in- or decrease their importance in the clustering process. The multiplicative factor affects the range of feature axes impacting the distance function. Thus, scaling factors $c_i > 1$ increase the distance to objects in a given dimension $i$, increasing their importance in the process. We apply the same scaling $c_v$ to both tangential velocity axes while leaving the positional axes unchanged; thus, $c_x = 1$. SigMA is applied to the following set of dimensions $\mathcal{D}$:

$$\mathcal{D} = \{X, Y, Z, c_v \times v_\alpha, c_v \times v_\delta\} \tag{10}$$

Theoretical considerations of the scaling relationship $c_x/c_v$ depend on various initial cloud and cluster configurations and interactions. However, the estimation of these influences is plagued by substantial uncertainties. Instead, we aim to determine a suitable scaling factor empirically by considering successful past extractions. Since the tangential velocity is inverse

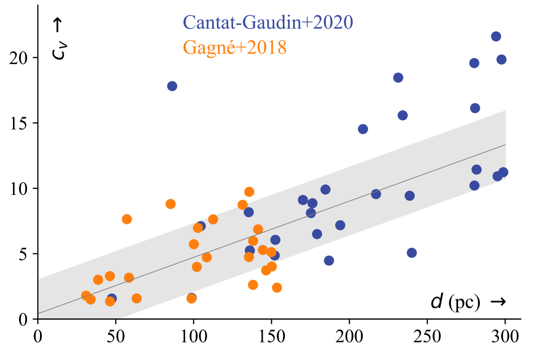


**Fig. 3.** Empirical distance-scaling relationship using data from Gagné et al. (2018) and Cantat-Gaudin & Anders (2020). The x-axis represents the distance to stellar groups; the y-axis shows the dispersion ratio of positional over kinematic sub-spaces. To compensate for unequal feature ranges in favor of the position axes, the velocity scaling factor must be set to the observed dispersion ratio. Seeing a trend, we fit a linear model to this data.

proportional to parallax, our goal is to extract a relationship between a stellar group's distance and its scaling factor.

The Sco-Cen association is at a distance of about 100–200 pc from us. To model the empirical distance-scaling relationship and subsequently apply it to Sco-Cen, we need data on stellar groups within at least 300 pc. Cantat-Gaudin & Anders (2020) provide a survey on open clusters in the Milky Way disk. Since they use the clustering tool DBSCAN to identify groups, their census is also prone to scaling issues. We substitute and add groups covered by Gagné et al. (2018), who have used a multivariate Bayesian model to identify members of young associations within 150 pc.[8]

The scaling fraction should account for the distance differences between positional and kinematic sub-spaces. To quantify this idea, we consider the distance distribution of sources to the cluster's center in each sub-space. Specifically, we compare the median absolute deviation of sources from their centers in position and velocity space, providing a robust statistic for statistical dispersion. We refer to this ratio of observed dispersion in the respective sub-spaces as the x-v dispersion ratio. To compensate for unequal feature ranges in favor of the position axes, the velocity scaling factor $c_v$ must be set to the observed x-v dispersion ratio.

Figure 3 shows the relation between a cluster's distance and its x-v dispersion ratio, which equivalently is our choice of $c_v$. We identify a linear trend and fit a linear model to the data, the gray band indicates a deviation of one standard deviation away from the mean assuming constant Gaussian model uncertainty. Since we observe several outliers, we use the Huber loss (Huber 1964), which is less sensitive to anomalies.

Using this empirical model, we find mean suitable scaling factors $c_v$ between approximately 4–9, assuming the groups of Sco-Cen are at a distance of about 100–200 pc. These values are similar to those of Kerr et al. (2021), who apply correction factors of 5 and 6 in their clustering approach.

At first glance, the model suggests sampling values in the range of 4–9 or using the mean 7.5. However, we also observe a significant scatter around the model that we need to consider. Instead of a single mean scaling factor, we aim to obtain a distribution of values from a given range of distances to the groups we aim to find.

As discussed in Appendix C, possible scaling factors can be expressed by the conditional probability integrated over a range of distance values. Given the linear model and associated Gaussian model uncertainties, we find a resulting distribution of scaling factors within distances of 100–200 pc. Since we need to perform a separate clustering run for each sample that we draw from the distribution, keeping the number as small as possible is essential. We generate ten samples which try to cover the sample space while keeping the underlying probability distribution in mind. The resulting samples can be seen in Fig. C.2.[9]

We run the clustering pipeline for each scaling fraction sample, creating an ensemble of ten clustering solutions. By summarizing the (potentially conflicting) results, we obtain a single consensus clustering solution. The consensus result is more robust against noisy data by aggregating multiple clustering solutions. This aggregation technique creates a meta-solution that usually provides better accuracy than any single clustering result can (Strehl & Ghosh 2002; Vega-Pons & Ruiz-Shulcloper 2011).

A consensus function aims to produce a result which shares as much information as possible with individual clustering results among the ensemble. Thus, the objective is commonly formalized in terms of optimizing the shared mutual information between ensemble labels and the consensus result. Due to the large sample sizes, we make use of the hybrid bipartite graph formulation algorithm introduced by Fern & Brodley (2004), who leverage graph partitioning techniques for an efficient implementation.

### 3.4. The role of uncertainty

Rigorous integration of positional uncertainties into the modality testing procedure of Burman & Polonik (2009) is a highly complex task, primarily due to the heteroscedastic nature of the uncertainties. Instead, we use a Monte Carlo approach that attempts to approximate the sensitivity of the modality structure to positional uncertainties. We do this by resampling the data using a Gaussian distribution centered on each point with an appropriate covariance matrix obtained from Gaia data.

Re-computing the modal structure on each resampled data set individually is computationally expensive. Therefore, we aim to study the effect of deviations on the initially computed modal layout instead. Since every merge decision impacts the final modal structure, we must evaluate the impact of uncertainty at each saddle point. While looping through all saddle points, we re-evaluate the hypothesis test for each resampled modal and saddle point density. However, testing each hypothesis multiple times increases the likelihood of rejecting the null hypothesis. Assuming statistical independence between individual tests, we can introduce a correction term on the significance level $\alpha$.

We use the Bonferroni correction (Bonferroni 1936), which compensates for an increased rejection probability by dividing the significance level of $\alpha$ by the number of $n$ comparisons. If at least one test is rejected with a p-value $< \alpha/n$, adjacent modal regions are not merged. In practice, we compute distances to

[8] We cross match the Gaia DR1 sources identified by Gagné et al. (2018) and DR2 sources from Cantat-Gaudin & Anders (2020) with EDR3 for more precise astrometry. If a cluster appears in both surveys, we opted for the Gagné et al. (2018) census to reduce the influence of scaling issues with DBSCAN clustering.

[9] We want to point out that the distance notation in the appendix changes from $d$ to $r$ to minimize confusion in the derivation of the final pdf.

the $k$'th neighbor of each initial modal candidate and the corresponding saddle points for each resampled data set. The number of resampled data sets limits the proposed procedure, as data generation is costly. Thus, we restrict the number of samples to 50, which means that two modal regions are kept separate if at least one p-value falls below $\alpha/50$ which results in 0.001 for a significance level of 5%.

### 3.5. Noise removal

Following the procedures described above, we obtain a data set segmentation into prominent peaks by iteratively merging modal regions separated by insignificant dips in density. This segmentation yields a list of non-overlapping areas in the data set without a noise characterization in mind. In principle, each modal region contains a dense core and background population corresponding to the stellar group and field content. In this section we aim to remove the field star component from the modal region to obtain a final clustering result.

We aim to remove the field star component in each modal region separately. By assuming the density of field stars and cluster stars to be approximately Gaussian distributed, we can model the observed density distribution in each modal region via a mixture of two univariate Gaussians. Thus, one Gaussian component describes the distribution of field star densities and the other one describes the stellar group densities. In Figure 4 we show an example of two Gaussians fitted to the density data of one modal region.

#### 3.5.1. Bulk velocity estimation

The Gaussianity assumption of density components is appropriate only in the original Cartesian coordinate system. Densities computed from proper motions suffer from perspective effects leading to deviations from normality due to the non-linearity of projections. We find such distortions also empirically when analyzing distributions of various modal regions in projected 2d (see, e.g., the tangential velocity space in Fig. 8) compared to Cartesian 3D velocities. Thus, we aim to transform all data into the six-dimensional parameters space (3D positions and 3D velocities) to facilitate efficient signal and background models.

A transformation from proper motion space to a 3D Cartesian velocity space is only possible if radial velocity information is available. However, only less than 5% of all sources in our sample have radial velocity measurements from Gaia. Nevertheless, we can exploit the co-moving property of stellar populations. We aim to adopt a similar strategy to Meingast et al. (2021), inspired by convergence point ideas (e.g., van Leeuwen 2009). The expected radial velocity value can be determined when the 3D bulk motion of stars alongside their positions is known.

However, compared to the method proposed by Meingast et al. (2021), we cannot determine the bulk 3D velocity for all groups. Hence, before we can determine the individual radial velocities in the first place, we have to estimate the space motion of individual populations. We determine the space motion of individual populations of size $n$ by minimizing the following loss function:

$$L(\tilde{\mathbf{v}}) = \sum_{i=1}^{n} \left( \frac{\Delta v_{\alpha,i}^2}{\sigma_{v_{\alpha,i}}^2} + \frac{\Delta v_{\delta,i}^2}{\sigma_{v_{\delta,i}}^2} + \frac{\Delta v_{r,i}^2}{\sigma_{v_{r,i}}^2} \right) \tag{11}$$

$$\Delta v_{x,i} = v_{x,i}^{\text{obs.}} - \tilde{v}_{x,i} \tag{12}$$

The minimization is done over the tangential ($v_\alpha$, $v_\delta$) and radial ($v_r$) velocities. The delta terms describe the offset between observed and computed values at the specified velocity $\tilde{\mathbf{v}}$. Although we introduce an additional observational error due to the parallax uncertainty, we choose the tangential velocities to match the unit of radial velocities, the essential component in the sum in Eq. (11). Each term in the sum is weighted by its respective uncertainty, which decreases the influence of observations with large measurements errors. If all observations lack radial velocities, then the last term is set to zero; if only a subset of $v_r$'s are missing, their values are imputed with the average of its complement.

For a perfectly co-moving population, the loss in Eq. (11) has a global minimum with a value of 0 at the group motion. Observational uncertainties, contamination from field stars, and a non-zero velocity dispersion will increase the minimum value accordingly. To search the 3D bulk motion that minimizes the proposed loss, we use the quasi-Newton method of Broyden, Fletcher, Goldfarb, and Shanno (BFGS) (Nocedal & Wright 1999) with an initial guess of the mean 3D velocity[10]. We denote the velocity which minimizes the loss (11) as the optimal bulk motion (OBM).

To determine the group motion of the co-moving population via our minimization approach, finding the OBM needs a large and pure selection of cluster sources; meaning truly co-moving stars. We attempt to obtain a rather clean sample of cluster stars via the aforementioned mixture model approach (Fig. 4). By fitting a mixture of two univariate Gaussians to the density distribution of a modal region we get a classifier that roughly separates cluster from field stars[11]. Since the input density is one-dimensional, the classifier – also referred to as cluster-noise classifier – becomes a simple threshold classifier. Sources with a density greater than the threshold are classified as cluster members. As the classifier is trained on densities determined in the 5D space which experiences projection distortion, we only use the 80% most dense stars in the cluster sample to determine the OBM. This density filter is designed to remove likely field star contaminants (false positives) which are typically expected to be less dense than cluster members. In Figure 4 we show an example of the contamination estimation.

The OBM is used to infer an "ideal" radial velocity which minimizes the Euclidean distance between the OBM and the velocity vector constrained by the measured proper motions. We call the 3D motion resulting from measured proper motions and the computed radial velocity the *minimally different velocity* (MDV) which we infer for sources without $v_r$ measurements as well as for cases with large relative uncertainties of $v_r/\sigma_{v_r} < 2$. After this step, all sources have an associated radial velocity, either measured or inferred.

The 3D velocity information is used to determine cluster membership in the following steps. We pre-filter unlikely members via a kinematic selection before applying the cluster-noise classifier to the now complete 6D phase space including the computed $v_r$ estimates. The pre-filter removes possible contaminant stars that have vastly different 3D motion, namely more than 10 km s$^{-1}$ from the OBM.

[10] If no radial velocities are available, our initial guess is the null vector.

[11] We use a simple threshold classifier where both mixture components have equal class (posterior) probability. The likelihoods and class fractions are estimated using a univariate GMM.

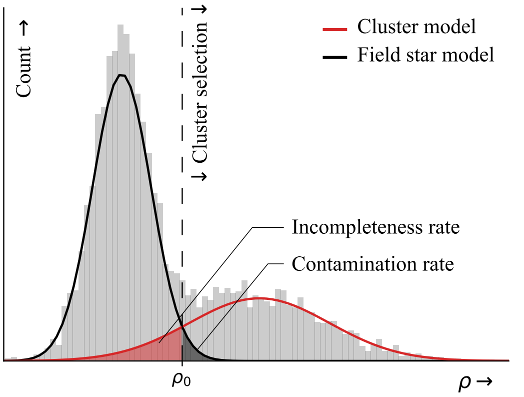


**Fig. 4.** Noise reduction schematic. We fit the observed 1D density distribution $\rho$ with a mixture of two Gaussians modeling the cluster (red line) and field star (black line) population, respectively. We obtain an approximation to the field star contamination, and incompleteness rate in the cluster sample via the cluster-noise classifier's false positive, and false negative rate, respectively.

### 3.5.2. Removing field star contamination

To remove field star contamination, we create a combined 6D Cartesian position and velocity space using determined MDVs to obtain a separation of signal and background content. Combining the space directly puts an emphasis on one of the subspaces (position or velocity) due to different value ranges; see Sect. 3.3.3 for more details. Large axis ranges automatically dominate the extraction as distances along these dimensions are penalized, more drastically impacting density estimation. Thus, we scale one subspace over the other and subsequently compute a density distribution for a given modal region.

We separate the stellar population from the field star component using a cluster-noise classifier. This classifier is applied to the 1D density estimation $\rho$ determined in 6D phase space; see the x-axis in Fig. 4. By boiling down the high-dimensional information into a single variable, the classifier ignores positional information of sources in the original feature space. Thus, random over-densities in the field might be extracted alongside the cluster. To reduce the contamination of random field star components, we introduce a neighborhood graph from which we delete vertices which fall below the computed density threshold $\rho_0$, as shown in Fig. 4. We define sources within the densest (and typically the largest) connected component as cluster members. To extract cluster members more robustly we compute one extraction for a range of scaling parameters, see Sect. 3.3.3. We obtain a final cluster catalog by removing sources which appear in less than half of these extractions.

### 3.5.3. Contamination and completeness estimate

The applied noise reduction scheme presents a direct way of estimating the field star ($F$) contamination fraction $f_{FP}$ (false positive rate) in the cluster ($C$) sample, as well as the incompleteness fraction $f_{FN}$ (false negative rate) of unidentified members. The contamination fraction or false positive rate is simply the probability of observing a sample from the field star distribution with a value larger than the density threshold $\rho_0$:

$$f_{FP} = P_F(\rho > \rho_0) \tag{13}$$

Similarly, we estimate the incompleteness fraction or false negative rate as the probability of observing a sample from the cluster distribution with a value less than $\rho_0$:

$$f_{FN} = P_C(\rho < \rho_0) \tag{14}$$

The determination of $f_{FP}$ and $f_{FN}$ is schematically shown in Fig. 4. We compute both values for each group, which gives us a distribution of contamination and incompleteness rates. Eventually, we obtain a mean contamination estimate across all groups in Sco-Cen of 3% with a standard deviation of 2% across groups (see also Sect. 4 for an independent contamination estimate using astrophysical knowledge). This estimate does not take care of systematic uncertainties. One source of systematic uncertainty is a possible deviation from Gaussianity of any of the mixture components. Further uncertainty is added via the density estimation to which the mixture model is fit. Since we do not have access to $f$, we inevitably make mistakes by substituting it with our estimate $\hat{f}$.

Although these factors of uncertainty are not considered in our simple contamination estimate, these findings seem to coincide quite well with the empirical HRD contamination estimate shown in Fig. 9 in Sect. 4.

We find the mean completeness across groups of approximately 83% with a standard deviation of 10%. Similarly to the contamination fraction, determining the incompleteness depends on the mixture components and density approximation. Still, compared to the contamination fraction, the incompleteness estimate is relatively high. A caveat of our noise reduction procedure is that we reduce high-dimensional phase space information into a univariate variable that is used to filter the data. This univariate formulation lacks descriptions about local positional and kinematic relationships that might help to increase the completeness of our catalog. Further, we estimate the actual value even lower, as we find multiple connected components in the neighborhood graph of which we only extract the main component. We also only admit stars that pass a threshold of 50% across different scaling fractions. All these decisions increase the precision of our sample at the cost of a reduced recall.

A lower completeness fraction compared to our initial estimate is also what we find when comparing our sample to past extractions in the literature in Sect. 4.2. These comparisons suggest sample completeness towards 75%.

We consider increasing the completeness fraction as an important future work. Until then, tools such as BANYAN (Gagné et al. 2018), or `Uncover` (Ratzenböck et al. 2020) can help to improve our presented membership list.

### 3.6. Multi-scale clustering

The density field is the main parameter of the proposed clustering method. Its topology is affected by the estimation process, which impacts the final result. Especially the smoothing parameter can create, on the one hand, a very rough and, on the other hand, an over-simplified density field. The schematic Figure 5 illustrates the dependence of the cluster number to the density estimation process. Applying a smoothing operator generates a family of density fields, called a scale space (Witkin 1987). We use this scale space concept to study the dependence of extracted clusters on the density estimation. Clusters with a long lifetime

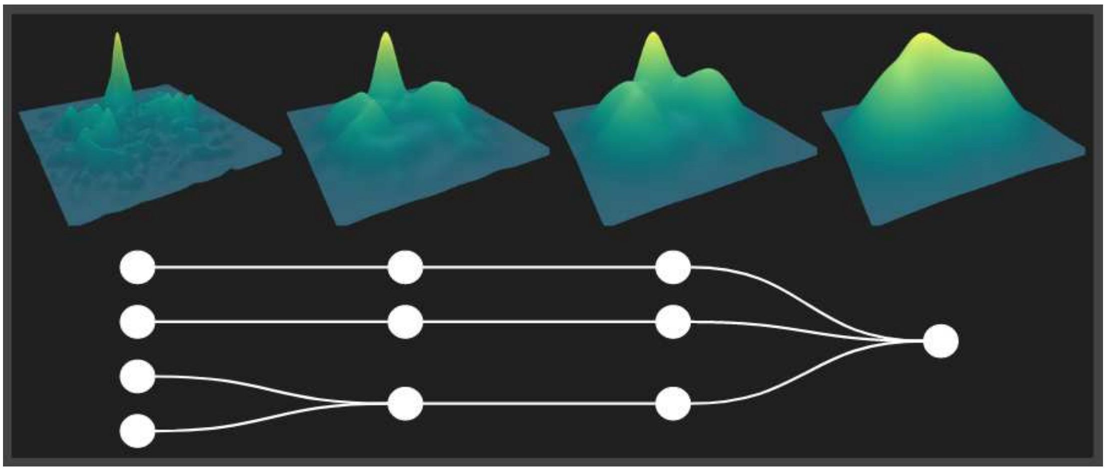


**Fig. 5.** Schematic figure linking the cluster number to the density estimation process. Applying a smoothing operator generates a family of density fields. This hierarchical family of functions is called a scale space.

in the scale space are preferred over, for example, "short lived" children.

We approximate the scale space by running SigMA $N$ times obtained by progressively smoothing the initial density field. Given an ensemble of $N$ density estimates $\{\hat{f}\}_i, i \in [0, N]$, we track clusters through various density filters. To track clusters through different levels of scale space we use three cluster connection rules based on cluster modes which we approximate by the densest point in a modal region. The connections we define are the following: *direct link*, *merge*, and *split*.

A direct link connection denotes a connection between two modal regions whose Jaccard similarity is larger than 50% and both cluster modes lie in the intersection set. A merge connection is a weaker condition and is only placed if no direct link can be established. A merge link is made when a parent cluster[12] contains the cluster mode of its child. If both conditions for direct and merge link are not satisfied, a split connection is placed between a parent and child cluster if the child contains the cluster mode from its parent.

The emergence of critical points, or additional clusters, in smoother versions of the scale space is a result of the non-exact nature of our density estimation (Reininghaus et al. 2011; Lifshitz & Pizer 1990) as well as due to randomness introduced by our Monte Carlo strategy. In the absence of noise, smoother density filters result in a simplified topology. Thus, we apply the pruning strategy introduced by Reininghaus et al. (2011) to the resulting merge-split graph which generates a simplified merge tree. The merge tree for our running toy data set is schematically illustrated in Fig. 5. The resulting merge tree was analyzed visually, from which we selected the 48 most stable clusters in Sco-Cen (Sect. 4).

### 3.7. Validation using astrophysical knowledge

Two direct observables, that can be identified in our application on Sco-Cen (see Sect. 4), serve as a validation test of the method. First, and apart from the youngest groups that are affected by dust extinction, the Gaia color-absolute-magnitude diagrams (equivalent to observational Hertzsprung-Russell Diagrams, HRDs) for the stars in each group show a narrow (coeval) distribution (see Fig. 9 and Paper II). There is no procedural reason why this should be the case, the method does not know about the brightness and colors of the stars. Only a meaningful selection of co-moving stellar siblings can produce the observed narrow sequences in the HRDs. Another observable that serves as test is the prominence of massive stars associated in 2D projection with the groups identified, while they are often located at a central position within the concerned clusters (e.g., $\alpha$ Sco, $\beta$ Sco, $\delta$ Sco, $\nu$ Sco; see Sect. 4 and Table 2). These massive stars are too bright to have reliable measurements in the Gaia archive and the brightest are not even in Gaia (like Antares, Ohnaka et al. 2013), still, the method finds groups around them. Based on Hipparcos astrometry (Table 2) we find strong evidence that many of these bright stars share similar parallaxes and proper motions as the clusters they seem to belong to in projection. This is an astrophysically relevant result (massive stars do not form alone and are often found at central positions) and it serves as another direct validation of the method.

## 4. Application to Sco-Cen

We apply SigMA to Gaia EDR3 data inside a box of about $10^7$ pc$^3$ containing the Sco-Cen OB association, as defined in Sect. 2. The box was chosen to include the classical Blaauw definition of Sco-Cen, including the classical sub-groups Upper-Scorpius (US), Upper-Centaurus-Lupus (UCL), and Lower-Centaurus-Crux (LCC), and to go beyond them and include the molecular cloud complexes of Pipe, Corona Australis (CrA), Chameleon (Cham), and L134/L183. Some of these regions were tentatively associated with Sco-Cen in the past (e.g., Lépine & Sartori 2003; Sartori et al. 2003; Preibisch & Mamajek 2008; Bouy & Alves 2015; Kerr et al. 2021).

In this paper we discuss the SigMA extracted young stellar groups in Sco-Cen, which are part of the $\lesssim$ 20 Myr Sco-Cen star formation event (Pecaut et al. 2012), and their connection to previous work. In a future paper (Ratzenböeck et al. in prep, Pa-

[12] The parent cluster resides in the $i+1$'th level, whereas the child cluster is from level $i$.

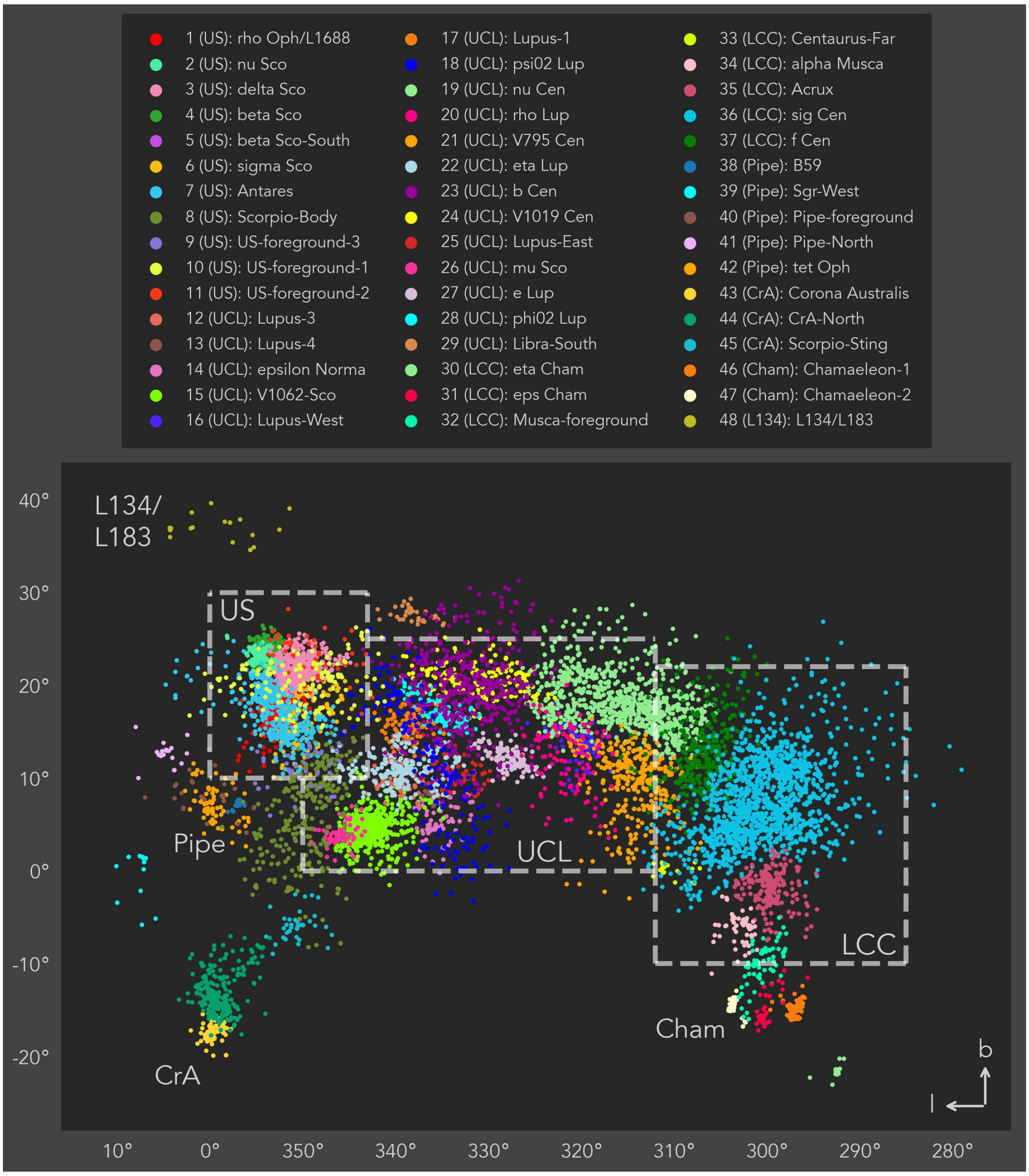


**Fig. 6.** The distribution of the 48 SigMA clusters in Sco-Cen projected in Galactic coordinates. Traditionally, the Sco-Cen OB association was separated into US, UCL, and LCC, marked with gray dashed lines. The SigMA extracted clusters reveal a more complex substructure of Sco-Cen than initially proposed by Blaauw (1946), and they show a more extended spatial distribution that includes the CrA, Pipe, Cham, and L134/L183 regions. The clusters are ordered in the legend by region, as given in Table 1. See here an interactive 2D version. For a better visualization of these clusters see the interactive 3D version (see Fig. 7).

per II) we discuss in more detail the ages of the individual SigMA groups and the star formation history of the Sco-Cen complex.

In total SigMA extracts about 70 clusters inside the defined search box. Of these, approximately 20 clusters are older populations with ages > 20 Myr, for example, IC 2602 (~30 Myr, e.g., Dobbie et al. 2010; Damiani et al. 2019). Groups older than 20 Myr are not discussed further here, although they might be related to Sco-Cen at larger scales (e.g., "blue streams", Bouy & Alves 2015). We will discuss these older groups in future work.

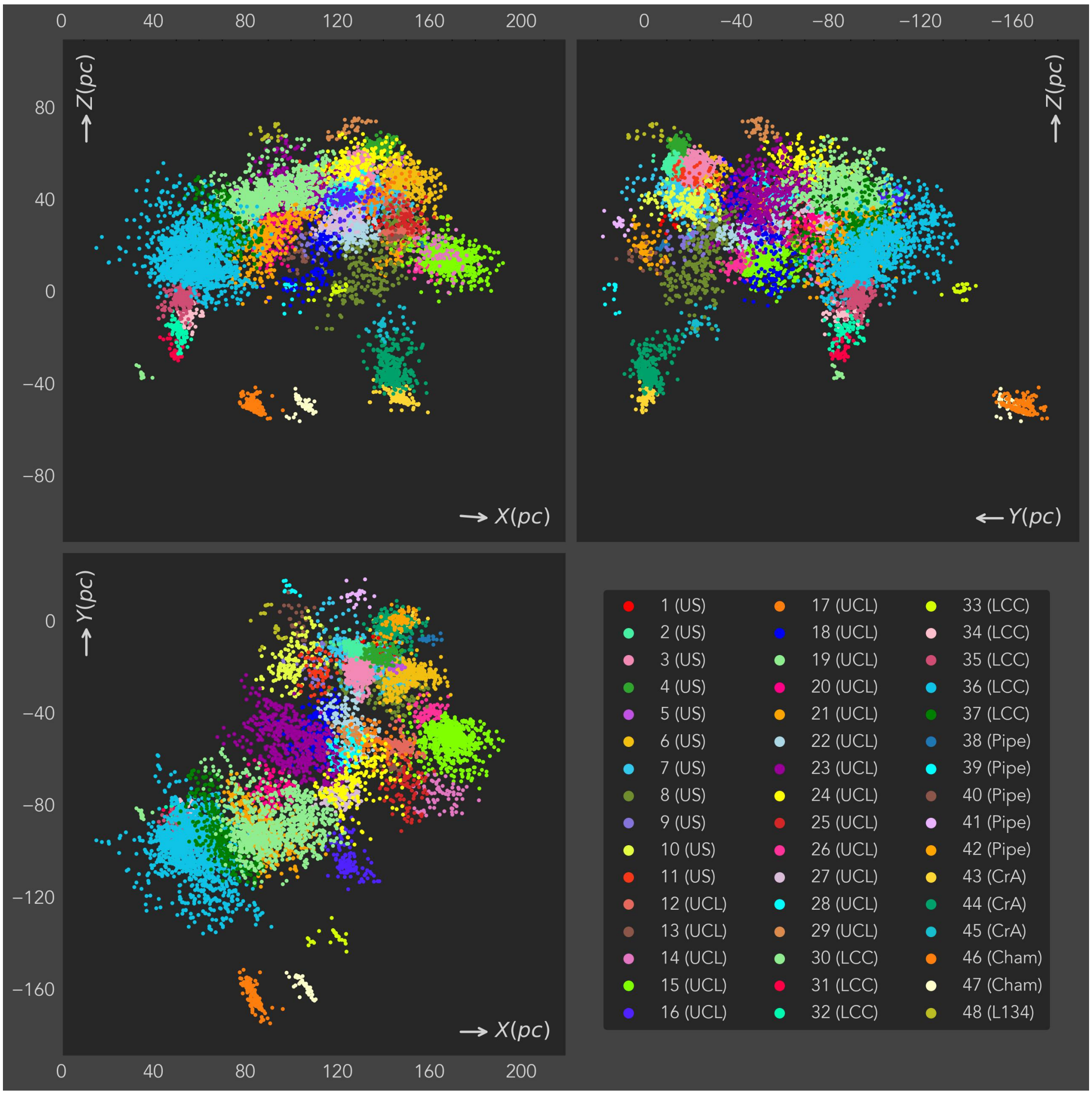


**Fig. 7.** 3D distribution of the 48 `SigMA` Sco-Cen clusters in Heliocentric Galactic Cartesian coordinates. The Sun is at (0,0,0). Colors and labels are as in Fig. 6. See also the interactive 3D version of the figure, which allows a better separation of the clusters (by double-clicking on a cluster in the legend of the interactive version, the selected cluster can be isolated).

We find that 48 stellar groups are associated spatially and kinematically with the Sco-Cen OB association, containing in total 9810 stellar cluster members, which will be discussed in more detail in this paper. In Figure 6 we show the distribution of the 48 Sco-Cen `SigMA` clusters projected in Galactic coordinates, and in Fig. 7 in 3D space using a Heliocentric Galactic Cartesian coordinate frame (see also the interactive version of the 3D figure). The 48 clusters seem to form the continuous body of the Sco-Cen association, beyond Blaauw's original three subgroups boundaries.

In Figure 8 we show the location of the `SigMA` clusters in the tangential velocity plane ($v_\alpha/v_\delta$). Since the clusters partially occupy similar velocity spaces in the $v_\alpha/v_\delta$ plane, we also provide an interactive version of this figure, allowing a better appreciation of 2D kinematical properties of the clusters in Sco-Cen. The 48 young clusters all fall on a connected loop-like pattern in tangential velocity space, a pattern largely created by the reflex motion of the Sun. This is highlighted in Fig. D.1 in Appendix D, showing that these projected motions are expected for stellar groups at Sco-Cen positions and distances, since they follow the theoretical Galactic orbits of sources at these Galactic

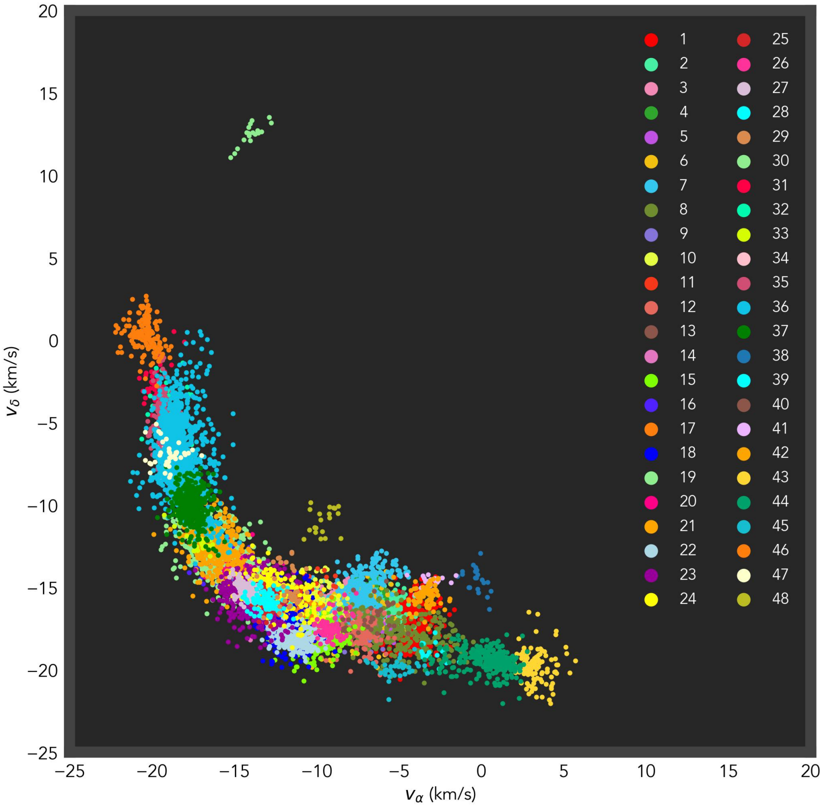


**Fig. 8.** Tangential velocity distribution of the 48 `SigMA` clusters. The observed tangential velocities relative to $\alpha$ and $\delta$ are strongly influenced by the Sun's reflex motion, while stellar groups at similar distances and with similar space motions are arranged along a loop-like pattern. Sources at $l \sim 0°$ are located in the lower right part of the figure, and sources at $l \sim 290°$ in the upper left part of the figure (see also Fig. D.1 in Appendix D). See the interactive version of the figure for a better appreciation of 2D kinematical properties of the clusters in Sco-Cen.

positions when assuming the velocity of the local standard of rest (LSR, Schönrich et al. 2010).

The Sco-Cen association, as extracted with `SigMA`, reaches well below the Galactic plane, as was indicated by previous works (e.g., Kerr et al. 2021) and is now further confirmed here. This includes regions that are not traditionally associated with Sco-Cen, like Pipe, CrA, Cham, and L134/L183. Moreover, other well know stellar groups, traditionally not assigned to Sco-Cen but later suggested to be associated with it, were picked up by `SigMA`, like the $\epsilon$ Cha and $\eta$ Cha (e.g., Mamajek et al. 1999, 2000; Fernández et al. 2008) or the $\beta$ Pictoris moving-group ($\beta$ Pic, e.g., Fernández et al. 2008; Miret-Roig et al. 2020).

We decided to not include the young nearby moving group $\beta$ Pic as part of our final sample of 48 stellar groups. The `SigMA` clustering extraction of $\beta$ Pic covers only one side of the known population as defined in Miret-Roig et al. (2020). This is likely due to the relatively close distance to the Sun (average distance of about 40 pc) which makes it more difficult to extract members from the 5D phase space as used by `SigMA` in this work. Generally, the observed proper motions of sources very close to the Sun are highly influenced by the reflex motion of the Sun (see also Appendix D), and sources of such nearby populations are scattered all over the sky, therefore, one can only confirm their dynamical membership by knowing the true 3D space velocities.

The majority of the 48 groups can be related to previously identified groups from the literature, which are often larger scale structures containing several of the `SigMA` clusters (see Sect. 4.2). The rich sub-structure as identified by `SigMA` also includes clusters with no clear counterpart in previous works. We decided to name such clusters after their location in a constellation, or after the brightest star that is part of a cluster or the brightest star that is seen in projection to a cluster. We often find bright B-stars towards cluster centers, at approximately the same distance and proper motion. We used Hipparcos astrometry (van Leeuwen 2007) to tentatively associate bright B-stars to the new clusters and list them and their astrometric properties in Table 2, showing the HIP ID and the Hipparcos astrometry. This table allows a direct comparison with the average properties of the `SigMA` clusters in Table 1. For the cases where there is a reasonable match, we name the cluster with the name of the bright B-star. Additionally, we index the stellar groups within this work from 1 to 48 as given in Col. "`SigMA`" in Table 1.

In Figure 9 we show the `SigMA` cluster members in a Gaia HRD (see Appendix E for details), confirming the youth of the majority of the sources. We find an excess of older low-

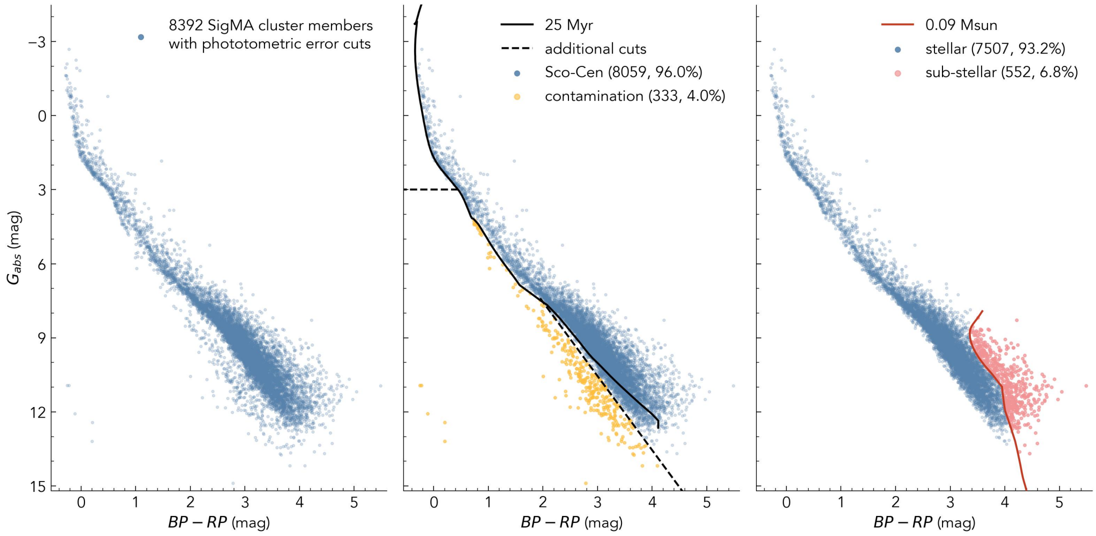


**Fig. 9.** Gaia color-absolute-magnitude diagram (HRD) $G_{\text{abs}}$ versus $BP - RP$ of the SigMA stellar cluster members. *Left:* SigMA cluster members that pass the photometric quality criteria as given in Appendix E. *Middle:* Potential contamination from older sources (orange), selected with a 25 Myr isochrone from PARSEC (black line) and with two additional cuts (black dashed lines). The cut at $G_{\text{abs}} = 3$ mag excludes the upper–main-sequence. The bottom black dashed slope accounts for the larger scatter of faint sources. Looking at the left panel, one can see a clear separation of an older sequence, which we attempt to separate with this cut, since the 25 Myr isochrone from PARSEC would be too conservative at the low-mass end. This indicates a contamination from older sources of about 4%. *Right:* Potential sub-stellar candidates (red dots) are selected with a 0.09 $M_{\odot}$ iso-mass line (dark-red line) from PARSEC. This cut indicates that there are roughly 7% of sub-stellar sources in the young `SigMA` Scon-Cen clusters. More details on the quality criteria, the selection borders, and the used PARSEC models are given in Appendix E.

mass sources that clearly separate from the Sco-Cen population, which are likely false positive `SigMA` sources. We use a 25 Myr isochrone (to allow for random scatter) plus two additional cuts, as shown in Fig. 9 (middle panel), and explained in Appendix E, to have a rough estimate for the fraction of contaminants to be about 4%. This contamination fraction is similar to the estimate in the methods section (Sect. 3.5.3). In Appendix E we give more details on the chosen photometric quality criteria and the selection conditions. In a follow up paper (Paper II) we will investigate the individual ages of each `SigMA` cluster in more detail with the help of isochrone fitting, allowing a more detailed investigation of the star formation history of the Sco-Cen complex.

When further investigating the young `SigMA` Sco-Cen members in the HRD in Fig. 9 (right panel), we find that there are about 6–7% sub-stellar objects (brown-dwarf candidates) within our sample (see also Appendix E). In the future, more complete samples of the individual clusters can be obtained by using the known members as training sets (e.g., as demonstrated in Ratzenböck et al. 2020), allowing to get more complete initial mass functions and a better characterization of the sub-stellar population (e.g., Miret-Roig et al. 2022).

### *4.1. Overview of the seven subregions in Sco-Cen*

In the following, and to help comparing `SigMA` results with the literature, we give a brief overview for each sub-region within Sco-Cen (US, UCL, LCC, Pipe, CrA, Cham and L134/L183). We then give a more detailed comparison to recent works in Sect. 4.2. The listed seven subregions include four regions that are not a traditional part of the Sco-Cen OB association, namely CrA, Pipe, Cham, and L134/L183, while we find them to be comoving with the larger Sco-Cen complex. Even if we assign each stellar group to one of the seven subregions, we stress that this classification should not be seen as physically distinct regions inside Sco-Cen, but simply to help compare our results with the literature.

#### 4.1.1. Upper Scorpius (US)

Toward US we identify 11 clusters (2558 stellar sources), which are partially extending beyond the traditional borders (Fig. 6). Of these 11 clusters, seven appear higher surface density and tend to be associated with prominent B-stars, as already pointed out above, namely $\rho$ Oph/L1688, Antares, $\beta$ Sco, $\beta$ Sco-South, $\delta$ Sco, $\nu$ Sco, and $\sigma$ Sco (see Tables 1 & 2).

Antares is the most extended among these clusters, showing substructure in velocity space. Moreover, Antares and $\rho$ Oph/L1688 are the only clusters showing significant overlap in the same volume in space within the `SigMA` clusters. In a recent paper (Grasser et al. 2021) we studied the $\rho$ Oph/L1688 cluster with Gaia EDR3 data and identified two kinematically distinct populations within the same volume (Pop 1 and Pop 2). These two populations are coincident with the $\rho$ Oph/L1688 and Antares groups, respectively. In detail, the cross-matched Pop 1 sample contains ~85% of the $\rho$ Oph/L1688 group (and ~4% of Antares, ~9% of $\delta$ Sco). The cross-matched Pop 2 sample contains ~85% of the Antares group and also a fraction of the $\sigma$ Sco group (~11%). Luhman (2022a) point out that "new" $\rho$ Oph/L1688 members in Grasser et al. (2021) have already been identified previously by other literature as being part of US. We clarify here that the "new" sources in Grasser et al. (2021) refers to sources that have not been assigned previously as members

**Table 1.** Average parameters of the 48 SigMA clusters in Sco-Cen.

| SigMA | Region | Group Name | Brightest star | Nr. | $l$ | $b$ | $\varpi$ | $d$ | $\mu_\alpha^*$ | $\mu_\delta$ | $v_\alpha$ | $v_\delta$ | $X$ | $Y$ | $Z$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | (deg) | | (mas) | (pc) | (mas/yr) | | (km/s) | | (pc) | (pc) | (pc) |
| 1 | US | $\rho$ Oph/L1688 | i Sco | 463 | 353.20 | 16.99 | 7.195 | 139 | -6.80 | -25.97 | -4.49 | -16.98 | 131.69 | -15.82 | 40.70 |
| 2 | US | $\nu$ Sco | nu Sco | 139 | 354.56 | 22.89 | 7.181 | 139 | -8.48 | -24.35 | -5.62 | -16.12 | 127.55 | -12.12 | 54.20 |
| 3 | US | $\delta$ Sco | ome Sco | 388 | 350.65 | 22.08 | 7.049 | 142 | -11.69 | -23.99 | -7.83 | -16.20 | 129.50 | -21.40 | 53.29 |
| 4 | US | $\beta$ Sco | HD144273 | 147 | 353.40 | 23.84 | 6.540 | 153 | -9.67 | -21.54 | -7.01 | -15.63 | 138.85 | -16.09 | 61.80 |
| 5 | US | $\beta$ Sco-South | HD146367 | 28 | 352.21 | 20.26 | 6.449 | 155 | -8.33 | -23.57 | -6.15 | -17.37 | 144.22 | -19.80 | 53.85 |
| 6 | US | $\sigma$ Sco | c02 Sco | 354 | 351.16 | 17.82 | 6.271 | 159 | -10.60 | -21.56 | -8.03 | -16.45 | 149.80 | -23.50 | 48.90 |
| 7 | US | Antares | HD146001 | 449 | 352.81 | 17.21 | 7.209 | 139 | -11.16 | -23.68 | -7.35 | -15.62 | 131.14 | -16.51 | 41.24 |
| 8 | US | Scorpio-Body | HD154310 | 315 | 349.15 | 7.21 | 6.997 | 143 | -8.08 | -26.64 | -5.44 | -17.69 | 139.78 | -26.86 | 17.21 |
| 9 | US | US-foreground-3 | HD151012 | 46 | 349.47 | 11.57 | 8.696 | 115 | -17.71 | -32.12 | -9.43 | -17.47 | 110.58 | -19.29 | 22.84 |
| 10 | US | US-foreground-1 | HD145964 | 170 | 349.32 | 20.59 | 9.151 | 109 | -20.53 | -31.72 | -10.54 | -16.59 | 100.72 | -19.45 | 38.26 |
| 11 | US | US-foreground-2 | HD140968 | 59 | 348.60 | 23.25 | 8.109 | 123 | -18.83 | -26.67 | -11.15 | -15.40 | 111.25 | -22.55 | 48.17 |
| 12 | UCL | Lupus-3 | LL Lup | 139 | 339.56 | 9.44 | 6.277 | 159 | -10.13 | -23.48 | -7.56 | -17.86 | 147.34 | -54.68 | 26.20 |
| 13 | UCL | Lupus-4 | MY Lup | 23 | 336.09 | 8.35 | 6.245 | 160 | -11.11 | -23.51 | -8.39 | -17.87 | 144.53 | -63.75 | 23.43 |
| 14 | UCL | $\epsilon$ Norma | eps Nor | 69 | 335.65 | 5.02 | 5.512 | 181 | -14.38 | -19.82 | -12.26 | -17.34 | 164.14 | -75.19 | 16.45 |
| 15 | UCL | V1062 Sco | mu02 Sco | 794 | 343.30 | 4.58 | 5.658 | 177 | -12.01 | -21.23 | -10.08 | -17.83 | 168.52 | -50.84 | 14.15 |
| 16 | UCL | Lupus-West | HD125777 | 112 | 319.89 | 13.74 | 5.960 | 168 | -20.16 | -17.01 | -15.98 | -13.68 | 124.16 | -105.68 | 39.97 |
| 17 | UCL | Lupus-1 | HD140817 | 110 | 338.78 | 15.76 | 6.728 | 149 | -16.72 | -23.44 | -11.74 | -16.60 | 133.55 | -52.20 | 40.25 |
| 18 | UCL | $\psi$02 Lup | psi02 Lup | 229 | 336.01 | 11.44 | 7.833 | 128 | -19.62 | -28.99 | -11.84 | -17.41 | 113.27 | -49.25 | 25.19 |
| 19 | UCL | $\nu$ Cen | nu Cen | 897 | 315.41 | 17.77 | 7.365 | 136 | -25.00 | -19.89 | -15.99 | -12.88 | 90.51 | -90.76 | 41.40 |
| 20 | UCL | $\rho$ Lup | rho Lup | 116 | 321.22 | 12.45 | 8.237 | 121 | -25.42 | -24.86 | -14.56 | -14.26 | 91.09 | -73.64 | 26.41 |
| 21 | UCL | V795 Cen | V795 Cen | 351 | 314.28 | 10.11 | 7.652 | 131 | -26.25 | -20.62 | -16.16 | -12.67 | 87.99 | -91.35 | 22.77 |
| 22 | UCL | $\eta$ Lup | HD 143699 | 242 | 339.42 | 10.92 | 7.244 | 138 | -17.53 | -27.57 | -11.45 | -18.07 | 126.57 | -47.55 | 26.16 |
| 23 | UCL | $b$ Cen | b Cen | 546 | 332.51 | 19.59 | 8.133 | 123 | -24.00 | -26.35 | -13.98 | -15.39 | 101.98 | -53.45 | 41.88 |
| 24 | UCL | V1019 Cen | HD132238 | 188 | 330.89 | 20.76 | 6.406 | 156 | -16.61 | -20.63 | -12.27 | -15.21 | 127.34 | -70.68 | 55.23 |
| 25 | UCL | Lupus-East | HD143022 | 87 | 334.34 | 10.44 | 5.875 | 170 | -16.10 | -19.57 | -12.98 | -15.86 | 149.69 | -71.93 | 30.86 |
| 26 | UCL | $\mu$ Sco | HD150641 | 80 | 346.02 | 3.87 | 6.065 | 165 | -11.74 | -22.46 | -9.22 | -17.53 | 159.68 | -39.90 | 11.16 |
| 27 | UCL | $e$ Lup | e Lup | 139 | 327.85 | 12.07 | 6.851 | 146 | -20.57 | -21.89 | -14.34 | -15.17 | 121.38 | -76.19 | 30.70 |
| 28 | UCL | $\phi$02 Lup | HD137432 | 116 | 334.99 | 16.98 | 7.028 | 142 | -19.43 | -23.55 | -13.02 | -15.83 | 122.77 | -57.51 | 41.21 |
| 29 | UCL | Libra-South | TYC-6762-602-1 | 38 | 338.31 | 27.51 | 6.558 | 152 | -16.33 | -20.81 | -11.60 | -15.33 | 123.84 | -50.13 | 70.33 |
| 30 | LCC | $\eta$ Cham | eta Cha | 20 | 292.46 | -21.64 | 10.137 | 99 | -30.05 | 26.89 | -14.06 | 12.56 | 34.92 | -84.67 | -36.33 |
| 31 | LCC | $\epsilon$ Cham | DX Cha | 41 | 300.23 | -15.59 | 9.807 | 102 | -41.03 | -5.76 | -19.79 | -2.81 | 49.64 | -85.09 | -27.52 |
| 32 | LCC | Musca-foreground | HD104600 | 67 | 300.48 | -10.02 | 9.748 | 103 | -39.37 | -9.40 | -19.20 | -4.50 | 51.54 | -86.34 | -17.73 |
| 33 | LCC | Centaurus-Far | HD121808 | 24 | 311.18 | 0.03 | 5.523 | 181 | -19.77 | -14.27 | -17.00 | -12.23 | 118.93 | -137.99 | 0.10 |
| 34 | LCC | $\alpha$ Musca | HD112383 | 64 | 302.47 | -5.87 | 9.700 | 103 | -38.10 | -13.67 | -18.63 | -6.69 | 55.41 | -86.51 | -10.54 |
| 35 | LCC | Acrux | zet Cru | 215 | 299.76 | -1.51 | 9.347 | 107 | -37.92 | -10.56 | -19.20 | -5.33 | 52.62 | -92.70 | -2.87 |
| 36 | LCC | $\sigma$ Cen | sig Cen | 1417 | 300.59 | 7.23 | 8.795 | 114 | -33.82 | -12.66 | -18.40 | -7.15 | 57.79 | -96.82 | 14.11 |
| 37 | LCC | $f$ Cen | f Cen | 326 | 306.63 | 12.69 | 8.258 | 121 | -30.71 | -17.18 | -17.62 | -10.03 | 69.67 | -94.86 | 26.35 |
| 38 | Pipe | B59 | AS 220 | 21 | 357.09 | 7.07 | 6.218 | 161 | -0.35 | -18.94 | -0.26 | -14.42 | 159.67 | -8.20 | 19.76 |
| 39 | Pipe | Sgr-West | HD163296 | 15 | 7.24 | 1.36 | 9.963 | 100 | -5.95 | -39.55 | -2.80 | -18.93 | 99.17 | 12.71 | 2.43 |
| 40 | Pipe | Pipe-foreground | CD-25-12033 | 29 | 0.13 | 8.03 | 9.646 | 104 | -13.37 | -33.67 | -6.49 | -16.66 | 102.62 | 0.23 | 14.30 |
| 41 | Pipe | Pipe-North | HD155427 | 22 | 4.85 | 12.73 | 7.554 | 132 | -4.93 | -23.36 | -3.11 | -14.60 | 128.37 | 10.87 | 28.91 |
| 42 | Pipe | $\theta$ Oph | HD158704 | 82 | 359.88 | 7.04 | 6.768 | 148 | -4.71 | -21.70 | -3.28 | -15.37 | 146.72 | -0.32 | 18.49 |
| 43 | CrA | Corona Australis | HD176270 | 124 | 359.86 | -17.66 | 6.485 | 154 | 4.36 | -27.16 | 3.20 | -19.81 | 147.07 | -0.37 | -46.80 |
| 44 | CrA | CrA-North | HD172910 | 265 | 359.05 | -13.66 | 6.753 | 148 | 0.84 | -27.67 | 0.58 | -19.43 | 144.02 | -2.41 | -35.12 |
| 45 | CrA | Scorpio-Sting | HD159807 | 36 | 350.63 | -5.96 | 6.973 | 143 | -7.65 | -29.54 | -5.20 | -19.82 | 140.06 | -22.72 | -14.71 |
| 46 | Cham | Chamaeleon-1 | CV Cha | 148 | 297.23 | -15.44 | 5.231 | 191 | -22.54 | 0.30 | -20.32 | 0.27 | 83.99 | -164.35 | -50.00 |
| 47 | Cham | Chamaeleon-2 | Hen 3-854 | 40 | 303.68 | -14.71 | 5.085 | 197 | -20.22 | -7.52 | -18.95 | -7.01 | 105.35 | -158.52 | -49.73 |
| 48 | L134 | L134/L183 | HD141569 | 20 | 358.12 | 36.92 | 8.854 | 113 | -17.53 | -20.28 | -9.69 | -10.85 | 90.57 | -3.05 | 67.43 |

of the young $\rho$ Oph/L1688 star-forming event. In fact, the two intertwining distinct populations within the same volume have been mentioned the first time in Grasser et al. (2021).

The three groups US-foreground-1,2,3 are located in front of the more compact clusters, visible in 3D space (Fig. 7), hence the chosen names. Finally, the group called Scorpio-Body extends from US toward the Galactic South, beyond the traditional borders of US, with a significant fraction located in UCL and in the direction of CrA (Sect. 4.1.5). It spans across the central body of the Scorpius constellation, hence the name. The 11 clusters toward US reveal a complex star formation history, which will be further discussed in a follow-up paper, where we analyze the ages of the SigMA clusters (Paper II).

#### 4.1.2. Upper Centaurus Lupus (UCL)

We identify rich substructure within UCL, containing 18 SigMA clusters (4276 stellar sources), as listed in Table 1. The most prominent cluster in the region is V1062 Sco (Röser et al. 2018), lying towards the far side of Sco-Cen. This cluster was picked up easily by visual selection methods (e.g., by Damiani et al. 2019 or Luhman 2022a; see Sects. 4.2.1, 4.2.5). We identify a second cluster close to V1062 Sco, which we call $\mu$ Sco, since its members are scattered around that bright B-star. We find that the positions and velocities of the two SigMA clusters are very similar, and members of both groups are part of V1062-Sco-selections in previous work. The star $\mu$01 Sco, which is the name giver of $\mu$ Sco lies in the center of the cluster, while the star $\mu$02 Sco is

**Table 2.** Hipparcos astrometry from van Leeuwen (2007) of bright stellar members in Sco-Cen.

| HIP | Name | `SigMA`[a] | SpT | $l$ | $b$ | $\varpi$ | $d$[b] | $\mu_\alpha^*$ | $\mu_\delta$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | (deg) | | (mas) | (pc) | (mas/yr) | |
| 80473 | rho Oph | 1 | B2V | 353.69 | 17.69 | 9.03 | 111 | -5.53 | -21.74 |
| 79374 | nu Sco | 2 | B2IV | 354.61 | 22.70 | 6.88 | 145 | -7.65 | -23.71 |
| 78401 | del Sco | 3 | B0.2IV | 350.10 | 22.49 | 6.64 | 151 | -10.21 | -35.41 |
| 78820 | bet Sco | 4 | B0.5V | 353.19 | 23.60 | 8.07 | 124 | -5.20 | -24.04 |
| 80112 | sig Sco | 6 | B1III | 351.31 | 17.00 | 4.68 | 214 | -10.60 | -16.28 |
| 80763 | Antares | 7 | M1Ib+B2.5V | 351.95 | 15.06 | 5.89 | 170 | -12.11 | -23.30 |
| 80582 | eps Nor | 14 | B4V | 336.00 | 0.98 | 6.15 | 163 | -13.68 | -19.89 |
| 81477 | V1062 Sco | 15 | ApSi | 343.57 | 5.18 | 7.54 | 133 | -10.25 | -21.59 |
| 82545 | mu02 Sco | 15 | B2IV | 346.20 | 3.86 | 6.88 | 145 | -11.09 | -23.32 |
| 76945 | psi02 Lup | 18 | B5V | 338.48 | 16.08 | 8.97 | 111 | -21.37 | -29.98 |
| 67464 | nu Cen | 19 | B2IV | 314.41 | 19.89 | 7.47 | 134 | -26.77 | -20.18 |
| 71536 | rho Lup | 20 | B5V | 320.13 | 9.86 | 10.32 | 97 | -28.26 | -28.82 |
| 69618 | V795 Cen | 21 | B4Vne | 314.13 | 3.96 | 6.77 | 148 | -23.80 | -20.92 |
| 78384 | eta Lup | 22 | B2.5IV | 338.77 | 11.01 | 7.38 | 136 | -16.96 | -27.83 |
| 71865 | b Cen | 23 | B2.5V | 325.90 | 20.10 | 9.62 | 104 | -29.92 | -30.68 |
| 72800 | V1019 Cen | 24 | B7II/III | 327.93 | 19.11 | 6.63 | 151 | -20.48 | -19.20 |
| 82514 | mu01 Sco | 26 | B1.5IV+B | 346.12 | 3.91 | 6.51 | 154 | -10.58 | -22.06 |
| 74449 | e Lup | 27 | B3IV | 327.83 | 11.43 | 6.47 | 155 | -22.01 | -21.75 |
| 75304 | phi02 Lup | 28 | B4V | 333.84 | 16.75 | 6.28 | 159 | -18.24 | -20.72 |
| 42637 | eta Cha | 30 | B9IV | 292.40 | -21.65 | 10.53 | 95 | -28.89 | 27.21 |
| 58484 | eps Cha | 31 | B9Vn | 300.21 | -15.62 | 9.02 | 111 | -40.34 | -8.30 |
| 61585 | alf Mus | 34 | B2IV-V | 301.66 | -6.30 | 10.34 | 97 | -40.20 | -12.80 |
| 60718 | Acrux | 35 | B0.5IV | 300.13 | -0.36 | 10.13 | 99 | -35.83 | -14.86 |
| 60823 | sig Cen | 36 | B3V | 299.10 | 12.47 | 7.92 | 126 | -32.36 | -12.51 |
| 63945 | f Cen | 37 | B5V | 305.47 | 14.34 | 8.36 | 120 | -29.85 | -15.17 |
| 84970 | tet Oph | 42 | B2IV | 0.46 | 6.55 | 7.48 | 134 | -7.37 | -23.94 |

**Notes.** Shown are mostly B-type stars that are either part of the `SigMA` selected clusters, or which are the name-givers of some clusters. [a] Col. 3 gives the index of the `SigMA` cluster, that is likely related to the given star. [b] The distance is simply $1000/\varpi$ from Hipparcos, shown here for completeness for an easier comparison with average distances of the `SigMA` clusters as derived from Gaia parallaxes in Table 1. The Hipparcos distance estimates should be treated with caution, since significant deviations to Gaia distance estimates are possible, while proper motions show deviations on the order of about ±2 km s$^{-1}$ when comparing sources which are both in Hipparcos and Gaia EDR3.

part of the `SigMA` selected members for V1062 Sco, located at the periphery of this cluster. This suggests a possible connection between the two clusters, but this statement is tentative at this point. To the West of V1062 Sco, and also located towards the far side of Sco-Cen, are the clusters $\epsilon$ Norma and Lupus-East.

The Lupus molecular clouds are located within UCL (Teixeira et al. 2020). `SigMA` extracts at least three clusters that might be related to the clouds, which are Lupus-1,3,4. Lupus-3 and 4 appear better correlated with regions of high dust column-density, matching with previous selections of Lupus-3 and 4 stellar members (e.g., Damiani et al. 2019; Kerr et al. 2021). The Lupus-4 cloud matches well with estimated cloud distances from Zucker et al. (2019) (both are located at about 160 pc; the dust-distance to Lupus-3 was not as directly measured as to Lupus-4). Lupus-1 is a more dispersed population, while members at its center seem to correlate with the Lupus-1 cloud (at about 155 pc from Leike et al. 2020; Zucker et al. 2021, while the `SigMA` cluster is at about 149 pc).

At the heart of UCL lie the clusters $e$ Lup and $\phi$02 Lup, which likely belong to the oldest parts of Sco-Cen, probably the clusters where the first supernovae in Sco-Cen originated from (Zucker et al. 2022). To the North of the traditional UCL borders we find a clustering, which has not been isolated in previous works, named Libra-South, based on its location within that constellation. The cluster lies at the northern borders of our investigated XYZ box, hence it needs more investigations in the future, since it might be a larger cluster than the `SigMA` extracted cluster.

### 4.1.3. Lower Centaurus Crux (LCC)

We find eight `SigMA` clusters (2174 stellar sources) toward the LCC region (see Table 1), which is now reaching farther below the Galactic plane compared to most of the work in the literature. For the `SigMA` extraction, the young local associations $\epsilon$ Cha and $\eta$ Cha are part of LCC, located at the Southern most tip, confirming the results of Mamajek et al. (1999, 2000) or Fernández et al. (2008). Toward LCC, `SigMA` extracted a cluster that seems unrelated to the main body of LCC, which we name Centaurus-Far since it lies about 60 pc further away from it, at a distance similar to that of the Chamaeleon clouds. This cluster was already identified in Kerr et al. (2021), as part of the TLC21 group (Cham-group) as EOM3, and named Cen-South (see Sect. 4.2.4 and Table F.2).

### 4.1.4. Pipe Nebula

Although not traditionally considered part of the Sco-Cen association, we find five `SigMA` clusters toward the Pipe nebula (169 stellar sources), including B59, Sgr-West, Pipe-foreground, Pipe-North, and $\theta$ Oph. The group B59 seems to be closely related to the star forming B59 cloud (e.g., Lombardi et al. 2006; Brooke et al. 2007; Román-Zúñiga et al. 2007, 2010). This is supported not only by projection in the sky towards cluster and cloud but also by the cloud distance of about 147–154 pc (Zucker et al. 2021), compatible within the uncertainties with the cluster distance about 161 pc. The $\theta$ Oph cluster, surrounding the $\theta$ Oph B2 star, is located at about the same distance to B59 and is close to the stem of the Pipe Nebula cloud, giving ground to studies of a possible interaction between the B2 star and the cloud (Gritschneder & Lin 2012). The other three groups in the Pipe Nebula are more dispersed and mostly located in-front of the Pipe Nebula cloud.

#### 4.1.5. Corona Australis (CrA)

The possible physical connection between CrA and the Sco-Cen association was already pointed out in previous studies (e.g., Mamajek & Feigelson 2001; Preibisch & Mamajek 2008; Kerr et al. 2021) and confirmed by our work. We identify a distinct cluster projected on top of the CrA molecular cloud and the embedded Coronet clusters, which we simply call the Corona Australis or CrA group. To the north we identify a second more extended group, called CrA-North, which was already discussed in Galli et al. (2020) or Esplin & Luhman (2022). Additionally, we identify a third group to the north-west of the two other groups, apparently building a bridge to the main body of Sco-Cen. This group we name Scorpio-Sting since its projected location matches the sting of the Scorpio constellation. Sco-Sting has only one clear counterpart in the literature, namely the TLC22/EOM7 group in Kerr et al. (2021) (see Sect. 4.2.4 and Table F.2), while they identify a smaller sub-sample of this group (12 members in Kerr et al. 2021 versus 36 members in this work). In total, the three stellar groups contain 425 stellar sources.

#### 4.1.6. Chamaeleon (Cham)

The well-known star-forming molecular clouds of Chamaeleon are seen through the same line-of-sight as the southern tip of the LCC, but lie clearly towards the back of LCC when seen in 3D (Fig. 7). We identify two clusters with a total of 188 stellar sources, Chamaeleon-1,2, which are likely directly related to the two molecular clouds of the same name and are already characterized with Gaia (e.g. Roccatagliata et al. 2018; Galli et al. 2020; Kerr et al. 2021, see also Sect. 4.2.4). Due to their youth, position, and tangential velocities we assume that the Cham clusters and clouds are part of the Sco-Cen star formation event, but this must be confirmed by tracebacks of the young population (see, e.g., Großschedl et al. 2021). Similar suggestions appear in Lépine & Sartori (2003) or Sartori et al. (2003).

#### 4.1.7. L134/L183

The cluster L134/L183 is a small, newly identified group to the Galactic North of US (with 20 stellar members). We assign this group to a separate region, since it does not fit to any other of the predefined Sco-Cen subregions. This stellar group is likely associated with the small molecular clouds L134 and L183 (or MBM 36 and 37, Magnani, Blitz, & Mundy 1985), that are currently non star-forming (Pagani et al. 2003, 2004, 2005). The distances to the clouds in Zucker et al. (2019) are about 105–120 pc, which match the cluster distance of about 113 pc. The presence of the close-by young stellar group suggests that (1) the clouds are remnants of a larger cloud that formed the newly identified `SigMA` cluster and (2) that the newly identified sources might be playing a role in the observed “cloudshine” phenomenon towards this cloud (Steinacker et al. 2010, 2015).

### 4.2. Comparison with previous work

In the following we compare the `SigMA` selected stellar groups with recent results from the literature, including Damiani et al. (2019), Schmitt et al. (2021), Squicciarini et al. (2021), Kerr et al. (2021), and Luhman (2022a). The studies by Damiani et al. (2019), Schmitt et al. (2021), and Luhman (2022a) discuss the whole Sco-Cen region, slightly extending beyond the traditional Sco-Cen borders, while excluding the regions to the Galactic South (CrA and Cham). These three studies select members within broad selection borders decided by hand, which we call in this paper visual selection methods. Squicciarini et al. (2021) focus only on the US region and extract clusters using a combination of a machine learning method and visual inspection. Kerr et al. (2021) present an all-sky study of young stars within 333 pc, hence covering the new extended view of the Sco-Cen association, using an unsupervised machine learning approach, which is more similar to our work then the aforementioned studies. The literature samples are cross-matched with the `SigMA` clusters using the Gaia EDR3 `source_id`, as specified in Appendix A. We provide an overview of the discussed literature samples in Table 3, giving the total number of sources of each literature sample, the total number of sources in `SigMA` Sco-Cen clusters within the respective studied areas, and the number of total matches.

#### 4.2.1. Comparison with Damiani et al. (2019)

Damiani et al. (2019) (hereafter, DPP19) analyzed Sco-Cen with the help of Gaia DR2 data. They used a traditional approach, selecting by hand over-densities in velocity space and position space, followed by selecting pre–main-sequence (PMS) stars from an HRD. Such an approach will deliver the most prominent clusters. However, somewhat less dense clusters can not be identified easily, when compared to unsupervised machine learning tools, like `SigMA`, and their method is less sensitive to possible spatial and kinematical structure in the Sco-Cen population. Their field of view (FOV) was slightly extended beyond the traditional borders of the association (see Table 3). They discuss eight compact clusters, which are prominently peaked in projection and in velocity space; these are UCL-1, UCL-2, UCL-3, Lupus 3, LCC-1, US-far, US-near, and the well studied IC 2602 (e.g. Randich et al. 1995; Stauffer et al. 1997; Dobbie et al. 2010; Meingast et al. 2021). Although `SigMA` easily detects IC 2602, we do not discuss this cluster since its age ($\sim$ 30 Myr) excludes it as a part of the recent Sco-Cen star formation event (that we define as $\lesssim$ 20 Myr, as in Pecaut et al. 2012). DPP19 also discuss four diffuse populations (D1, D2a, D2b, US-D2), which are generally distributed across large parts of the traditional Blaauw Sco-Cen OB association. Moreover, their catalog includes sources, which have not been assigned to any group (labeled with “N” in Table F.1). DPP19 were not able to further substructure the diffuse populations with their methods.

The DPP19 catalog contains in total 14,437 sources, of which 1734 are in their seven clustered Sco-Cen populations (350 in IC 2602), 8727 are in their four diffuse populations, and the rest 3626 have not been assigned to any population (labeled

**Table 3.** Overview of the recent Literature to which we compare our results in more detail.

| Reference | Data | Studied Area | Number statistics | | |
|---|---|---|---|---|---|
| | | | Ref.[a] | SigMA[b] | Matches[c] |
| Damiani et al. (2019)[d] | Gaia DR2 | ($l$ = 360° to 280°, $b$ = 0° to 30°) ∨ | 10,185 | 8919 | 7221 |
| | | ($l$ = 315° to 280°, $b$ = −10° to 0°) | 1703 clustered (∼17%) | | 1539 |
| | | FOV = 2750 deg$^2$, $d$ < 200 pc | 8482 diffuse (∼83%) | | 5682 |
| Kerr et al. (2021)[e] | Gaia DR2 | The whole TLC22 stellar group | 7394 | 9598 | 5135 |
| Schmitt et al. (2021)[f] | Gaia EDR3 & | de Zeeuw et al. (1999) borders: | 6190 | 8593 | 2614 |
| | eROSITA | US ($l$ = 343° to 360°, $b$ = 10° to 30°) ∨ | ∼65% vel-clustered | | 2614 |
| | | UCL ($l$ = 312° to 350°, $b$ = 0° to 25°) ∨ | ∼30% vel-diffuse | | 0 |
| | | LCC ($l$ = 285° to 312°, $b$ = −10° to 22°), | ∼5% IC 2602 | | 0 |
| | | FOV = 2050 deg$^2$, $d$ ∼ 60–200 pc | | | |
| Luhman (2022a) | Gaia EDR3 | $l$ = 2° to 283°, $b$ = −12° to 35°, | 10,509 | 9155 | 7713 |
| | | FOV = 3252 deg$^2$, $d$ ∼ 90–250 pc | | | |
| Squicciarini et al. (2021)[g] | Gaia EDR3 | $\alpha$ = 236° to 251°, $\delta$ = −29° to −16° | 2745 | 1918 | 1857 |
| (only US) | (subsample | FOV = 195 deg$^2$, $d$ ∼ 125–175 pc | 1442 clustered (∼53%) | | 1199 |
| | with RVs) | | 1303 diffuse (∼47%) | | 658 |

**Notes.** [(a)] Number of stellar members from the given reference. If there was a distinction in the literature between members in a more clustered or diffuse mode (which are generally differently defined in each reference), then the numbers are given below. [(b)] Number of stellar cluster members from SigMA in the given studied area (volume), out of the total 9810 SigMA stellar cluster members. [(c)] Number of matches between the given reference and the SigMA clusters. If a distinct comparison with clustered or diffuse sources was possible, then the matches with these are given below. [(d)] For DPP19 we only give the number of sources within their clustered or diffuse populations within $1000/\varpi_{\rm EDR3} < 200$ pc after a cross-match with Gaia EDR3, and without IC 2602. [(e)] For KRK21 we do not give the surveyed area, since they extracted the clusters from all-sky data within 333 pc from the Sun. We show the comparison to their whole TLC22 group, while this group also includes somewhat older stellar groups, like IC 2602, as furhter explained in the text. [(f)] The X-ray selected sources from SCF21 included velocity-clustered and velocity-diffuse sources. The separation of these was applied by us by hand, guided by Fig. 7 in SCF21. Hence, the fractions are only given roughly. The fraction of potential IC 2602 members is also given. The SigMA clusters have only matches with their velocity-clustered population. [(g)] SGB21 only studied the US region, finding sources in a more clustered mode and sources in a more diffuse mode, while the latter are simply the residuals of their clustering procedure. They study a subsample of sources with $v_r$ information in the 6D phase space (∼28%), which is not further discussed in this work.

with "N"). When cross-matching the DPP19 Gaia DR2 sample with EDR3 astrometry, we find that 654 stars (4.5%) are rejected when applying the distance criteria from DPP19 ($d$ < 200 pc), due to updated parallaxes in EDR3. The majority of these are sources that have not been assigned to any group or that belong to their diffuse populations. When now considering only the sources in the clustered and diffuse populations within 200 pc, then there are 10,185 potential Sco-Cen members in DPP19.

In total there are 7419 cross-matches between the SigMA clusters and DPP19, while 7221 of these belong to either the clustered or diffuse populations (198 are not assigned, "N"). Of the 7221 cross-matches, 5682 belong to one of the four diffuse populations. Comparing this number to their total diffuse population (8482 within 200 pc), we find that about 2/3 are a match with the SigMA clusters. In most cases, more than one DPP19 group (both clustered or diffuse) fits to one of our groups (see Table F.1), and vice versa. In particular, their diffuse groups each contain sub-parts of about 10 to 20 of the SigMA groups.

Focusing on the DPP19 compact groups (1539 matches out of 1703 within 200 pc), we find that their US-near and US-far can not be assigned clearly to only one of the SigMA groups (see Table F.1). US-near is most closely related to $\rho$ Oph/L1688, while also containing significant fractions of $\nu$ Sco, $\delta$ Sco, and the Antares group. US-far correlates best with $\sigma$ Sco, while also containing significant parts of $\delta$ Sco, $\beta$ Sco, and Antares. In particular, Antares is distributed almost equally among these two groups. The Antares group is indeed quite extended in space, partially occupying the same volume as $\rho$ Oph/L1688 (see Sect. 4.1.1 and Grasser et al. 2021). The case of the Antares and $\rho$ Oph/L1688 groups highlights the capability of SigMA to untangle young populations that share the same volume but have slightly different space motions.

The rest of the DPP19 compact groups are more clearly correlated with the SigMA groups, with UCL-1 matching with V1062 Sco and $\mu$ Sco, UCL-2 with Lupus-West, UCL-3 with $\phi$02 Lup, LCC-1 with Acrux, and Lup III with Lupus 3. The unassigned sources in DPP19 (N) correlate with a large fraction (about 80%) of our SigMA clusters, within the DPP19 FOV.

Regarding the comparison between the DPP19 visual section method and the SigMA unsupervised clustering method, we first note that the method used by DPP19 starts with a selection of stars by hand in velocity space, followed by a selection by hand of PMS stars on the HRD. This approach will always find more candidates than an unsupervised method. For example, a look at Figs. 2, 3, and 4 in DPP19 will make clear that the total number of member candidates using this approach is a strong function of the size of the selection shapes used in tangential velocity space and the HRD. These selection borders will necessarily select a larger number of true positives than an unsupervised method, while also the total number of false positives is likely higher.

We compare the number of sources in DPP19 stellar groups (compact and diffuse within 200 pc, 10,185 sources) to the number of matched SigMA cluster members (7221 within 200 pc) (see Table 3). There are 2964 sources only in DPP19, implying that we could be missing about 29% of possible members if all 10,185 sources were good members. We did not perform a detailed comparison but find that the 2964 sources also contain sources that seem to be older then the SigMA clusters when investigated in an HRD (similar as in Fig. 9), hence the incompleteness based on this comparison is likely lower than 29% (see also comparison with Luhman 2022a). As mentioned above, we expect SigMA to be missing possible candidates when compared with the method followed by DPP19 (which is a method that selects broad regions in various 2D planes of the phase space) but also expect the SigMA sample to be less contaminated. A deeper analysis is needed, although not warranted in this paper.

#### 4.2.2. Comparison with Schmitt et al. (2021)

Recently, Schmitt et al. (2021) (hereafter, SCF21) used eROSITA[13] (Merloni et al. 2020) to search for low-mass Sco-Cen members by cross-correlating the eRASS1 source catalog with the Gaia EDR3 catalog. They discuss 6190 possible Sco-Cen members within the traditional Blaauw borders (de Zeeuw et al. 1999), which are both observed by eROSITA and Gaia. Since X-ray emitting sources are expected to be young (e.g., Schmitt 1997; Feigelson & Montmerle 1999; Favata & Micela 2003; Bouvier et al. 2014), the sources detected by eROSITA, as discussed in SCF21, are potential members of Sco-Cen. They found X-ray sources down to about 0.1 $M_\odot$, and, unexpectedly, they also found the existence of a population of young X-ray emitting stars that appears to be more diffuse in velocity space[14], calling into question search schemes relying on kinematic selections.

We cross-matched the 6190 SCF21 X-ray selected sources with the SigMA selection. We find in total 2614 cross-matches in their studied area (see Table 3), while none of these belonging to their velocity-diffuse population. The latter is expected, SigMA only selects groups which are confined in position-velocity space, which naturally excludes any such velocity-diffuse sources. SCF21 claim that the diffuse population is largely composed of young stars, only somewhat older compared to the kinematically confined Sco-Cen members. We confirm the general youth of the sources by inspecting the two populations in an HRD, however, we see a relatively clear age separation between the velocity-clustered and velocity-diffuse populations. X-ray sources in the velocity space of the majority of Sco-Cen members have ages between 0.1–20 Myr, while X-ray sources that are velocity-diffuse have ages between 10–1000 Myr, with the majority at about 30–100 Myr. While these are technically young stars, they seem too old to be related to the Sco-Cen association.

The origin of this co-spatial but velocity diffuse population remains mysterious. Since these sources are older than Sco-Cen, they are unlikely to result from stellar interactions in Sco-Cen (an a priori unlikely process given the low stellar density of Sco-Cen). The diffuse population, or the co-eval part of it, could be related to a relatively older star-formation episode, sharing today the volume space of Sco-Cen, a plausible scenario in the Milky Way (Fürnkranz et al. 2019). We posit here that the SCF21 velocity-diffuse young sources are unlikely to be part of Sco-Cen, but represent a mystery that needs to be solved. As SCF21 point out, the sensitivity of eROSITA will allow in the near future to detect virtually all young Sco-Cen low-mass members. A combination of eROSITA future releases and Gaia data in Sco-Cen will be crucial to increase statistics and lead to a better understanding of the relation between observed X-ray luminosity with distance, age, stellar masses, and the origin of the velocity-diffuse population.

Finally, when concentrating on the velocity-coherent sample in SCF21 (without IC 2602), we find that there are about 35% in the whole SCF21 sample that could be additional Sco-Cen candidate members, which are only in SCF21 and have similar velocities as SigMA Sco-Cen members. When investigating additionally older-star possible contaminants in an HRD (similar to Fig. 9), this fraction would reduce to about 30%. This relatively high number is of interest, which might result from the broad selection conditions in SCF21, based on all X-ray detected sources within the Blaauw borders in a distance range of 60 to 200 pc (Table 3), while restricted to low-mass stars (BP−RP > 1, according to Pecaut & Mamajek 2013). These broad conditions, which do not attempt to identify any underlying clustered structure, will naturally pick up more members, while also more false positives, as also discussed in Sect. 4.2.1 and 4.2.6.

#### 4.2.3. Comparison with Squicciarini et al. (2021)

Squicciarini et al. (2021) (hereafter, SGB21) studied 2745 potential US members (see Table 3) by selecting subgroups solely based on kinematics. They divided the region into eight groups which they call the clustered population (1442 stars), and into an older diffuse population (1303), which is, however, differently defined then the velocity-diffuse population in Schmitt et al. (2021).

When comparing their selection to the SigMA clusters, we find that there are 1857 cross-matches in total out of the 2745 sources in SGB21, matching with 11 of the SigMA clusters, while only seven SigMA clusters have significant cross-matches. We list the cross-matches of SigMA with SGB21 in Table F.1 in Appendix F. We highlight more significant cross-matches here: Group 1, 2, and 3, match best with $\rho$ Oph/L1688, $\nu$ Sco, and $\delta$ Sco, respectively. Group 4 matches best with $\beta$ Sco and $\beta$ Sco-South, while also Group 6 has significant matches with $\beta$ Sco. Group 5 matches best with $\sigma$ Sco, while the majority of $\sigma$ Sco is in the SGB21 diffuse population. Group 7 and Group 8 match best with Antares, while the majority of Antares is also in the SGB21 diffuse population. Generally, the Antares group seems to split up into more than one cluster, also in other previous work. The four groups Sco-Body, $\psi$02 Lup, and US-foreground-1 & 2 have only a few matches with the diffuse population. The SGB21 diffuse population is largely contained within the $\sigma$ Sco and Antares groups, with some diffuse members distributed among each mentioned group (see Table F.1). This suggests that the diffuse population is not a separate older group but stars that were not clustered by the methodology in SGB21.

Focusing on the SGB21 candidate members in the clustered populations, there are 243 sources only in SGB21 (~9% out of their total, or 17% out of their clustered). However, in total we find more clusteres sources toward US (1918 in this work versus 1442 in SGB21 in the same volume, see Table 3).

The differences in the final cluster definition in US likely arise from the different clustering methodologies. To better understand the SGB21 approach we outline the basics here. SGB21 use a semi-automated approach based on iterative k-means clustering on a 4D sample, using 2D sky positions and 2D tangential velocities. The authors propagate the sky positions 15 Myr into the past and future, producing a new 4D data set at each step; tangential velocities are constant throughout individual data sets. By studying the sky distribution of each slice, SGB21 visually identify over-densities. These over-densities are extracted via k-means clustering in 4D space at a given time step. Subsequently,

[13] Extended ROentgen Survey with an Imaging Telescope Array. A wide-field X-ray telescope on-board the Russian-German "Spectrum-Roentgen-Gamma" (SRG) observatory.

[14] We applied the separation of kinematically clustered and diffuse populations by hand in $v_\alpha/v_\delta$ space, as indicated in Fig. 7 in SCF21, since the selection conditions are not clearly outlined by the authors.

the clustered data points are removed from the data set, and the process of looking for over-densities starts anew. The clustering process terminates when the authors cannot find any apparent density peaks in the sky distribution.

Besides the feature space difference, SigMA has significant differences compared to SGB21's iterative clustering approach. First, the k-means algorithm cannot deal with the observed non-convex cluster shapes in projected coordinates. The extracted clusters are 4D Voronoi cells[15] which can have very elongated shapes. Second, SGB21 analyze 2D projections of the high-dimensional data to identify clusters visually. Thus, cluster selection is influenced by projection effects and human judgment. Conversely, SigMA employs a modality test directly in the high-dimensional phase space, taking into account multi-dimensional relationships between data axes. These rather different approaches to extract clusters in US make it clear that the results can not be compared at face value, while fractions of the most robust clusters ($\rho$ Oph/L1688, $\nu$ Sco, $\delta$ Sco, $\beta$ Sco, $\sigma$ Sco, and Antares) have been identified by either method.

#### 4.2.4. Comparison with Kerr et al. (2021)

Recently, Kerr et al. (2021) (hereafter, KRK21) presented a study of nearby young stellar populations within 333 pc from the Sun. They use the HDBSCAN clustering algorithm (see Sect. 3.1.4) on Gaia DR2 parallaxes and proper motions, on a pre-selected sample of PMS stars with ages $\lesssim$ 50 Myr. They identify 27 *top-level clusters* (TLC), including Chameleon as TLC 21 and the Sco-Cen association as TLC 22. The latter was further broken down into another 27 sub-groups based on the *excess of mass* (EOM) method, selecting the most persistent clusters in the clustering tree. Three of these EOM sub-groups (EOM 12, Lupus; EOM 17, Upper-Sco; and EOM 27, LCC) where further broken down into *leafs*, which are nodes of the clustering tree.

The TLC 22 covers the Sco-Cen region in KRK21 and TLC 21 covers the Chamaeleon region. These two groups combined show a similar extent to our Sco-Cen extraction. SigMA finds more groups (48 in this work versus 44 in KRK21), while the TLC 22 sub-groups in KRK21 also include older populations (e.g., IC 2602 or Platais 8), which are not in our final Sco-Cen sample, since we do not discuss older groups (yet). Therefore, only 38 of the KRK21 groups toward Sco-Cen fall within the younger selected SigMA clusters from this work.

In Table F.2 in Appendix F we show an overview of the matches of SigMA groups with corresponding KRK21 groups. Overall, the SigMA Sco-Cen groups are more richly populated compared to the KRK21 groups. In most cases, there is at least some overlap between our groups and their main TLC 22 group (and with TLC 21, Cham), while some of our groups also distinctly correspond to EOM subgroups (or leafs). For about 40% of the SigMA groups, a clear accordance with a single EOM group (or leaf group) is not possible, due to overlaps with more than one SigMA group, or due to no or only insignificant overlap.

Some differences of the SigMA and KRK21 clustering results might arise from the different data input, since we use Gaia EDR3 and KRK21 use DR2, while this would only create minor deviations if DR2 data would have been used for SigMA. Although both HDBSCAN and SigMA approximate the hierarchical cluster tree, we expect discrepancies in clustering results. The primary reason for this difference is the cluster tree pruning strategy discussed in Sect. 3.1.4. The EOM heuristic prioritizes large clusters over their children when they maintain a long lifetime in the density hierarchy. The resulting children fail to exceed the parent's EOM. Conversely, our pruning strategy does not depend on cluster lifetimes but only cares about substantial density valleys between neighboring density peaks.

The additional leaf separations in KRK21 were applied to the US, Lupus, and LCC regions, since they found that there are substructures that have not been identified by the EOM method. Their leaf clustering is, however, often not a good match with the SigMA clustering, especially concerning LCC. The SigMA clusters are differently separated within the larger LCC and they are also richer and mostly more extended when compared to KRK21 clusters. Compared to the EOM heuristic or SigMA's multi-modality considerations, leaf clusters do not come with statistical guarantees. The clustering result is highly susceptible to random density fluctuations since leaf nodes are extracted only considering the minimum cluster size criterion (Stuetzle & Nugent 2010); see Sects. 3.1.3 and 3.1.4 for more details. Without any additional pruning strategy which deals with spurious clusters, leaf clustering results need to be taken with a grain of salt. Nevertheless, some of the leafs in US ($\rho$ Oph/L1688, $\nu$ Sco, $\delta$ Sco, $\beta$ Sco) show good agreement with the SigMA US cluster separations, indicating the robustness of these clusters.

When comparing the TLC22 group (7394) with our Sco-Cen SigMA extraction (9598 without Cham) we find 5135 cross-matches in total (Table 3). Hence, 2259 ($\sim$31%) sources are only in TLC22, and 4463 are only in SigMA. We find that the KRK21 only sample contains older stellar groups, which gets also apparent from their Table 6 (including, e.g., $\beta$ Pic or IC 2602). However, a clear separation of the younger Sco-Cen stellar groups as discussed in this work, and the somewhat older groups is not straight forward, since about 50% of the sources in the TLC22 group have not been assigned to a separate sub-cluster (EOM or leaf). The somewhat older sources can also be estimated when investigating the HRD[16] or the velocity space. There are sources that have deviating motions from SigMA Sco-Cen members, which largely coincide with the KRK21 older EOM groups. We try to estimate the "older-star contamination" in KRK21 by taking into account all these points, resulting in a lower fraction of TLC22 only sources, which could be young candidate members missed by SigMA ($\sim$23%). The reason for these extra source in the KRK21 TLC22 group is similar to the mentioned reasons above (e.g., in Sect. 4.2.1). The TLC22 group represents a cluster root, enveloping the whole Sco-Cen region and somewhat beyond, and no additional substructure was extracted (yet). In a following step KRK21 use the EOM and leaf methods to identify individual clusters, while in this step they lose almost 50% of the original TLC22 group, as mentioned above.

Generally, the TLC22 group seems to be overall more incomplete compared to the SigMA Sco-Cen extraction, since we find in total more members, while also finding more substructure. In conclusion, the comparison with KRK21 highlights the differences that can arise with different unsupervised machine learning clustering tools, and a careful choice of the appropriate clustering algorithm should be considered for the scientific question at hand.

#### 4.2.5. Comparison with Luhman (2022)

Luhman (2022a) (hereafter, L22A) recently investigated the Sco-Cen region containing selections for US, UCL/LCC, V1062 Sco, Ophiuchus, and Lupus (the Southern parts of Sco-Cen are not

[15] As far as we know, scaling between sky coordinates and tangential velocities was not considered.

[16] The KRK21 PMS selection includes sources up to about 50 Myr.

discussed in L22A), and using Gaia EDR3 data to identify 10,509 candidate members of Sco-Cen (see Table 3). L22A concentrates on established stellar groups in Sco-Cen to guide their selection. The visual selection approach of L22A is not suitable to separate the underlying kinematical substructure of the Sco-Cen population. For example, it is clear from Fig. 4 in L22A (bottom panel) that the UCL/LCC group contains several overdensities in $l/b$ space, but these are not extracted or identified. The L22A selection is based on global kinematic criteria, extracting candidates exhibiting proper motions similar to expected proper motions of known members.

Cross-matching the 10,509 L22A Sco-Cen candidate members with the SigMA clusters gives a total of 7713 matches, 2796 L22A only sources, and 1442 SigMA only sources within the L22A studied area (Table 3), where the SigMA sample contains 9155 sources in total. When investigating the 2796 L22A only sources, we find that they do not show significant signs of being older than 20 Myr, or of having significant deviating motions from SigMA Sco-Cen cluster velocities. These extra sources, or part of them, could be Sco-Cen members, meaning we might be missing up to about 1/4 of the candidates in L22A. This is not surprising because methods based on visual selection, using broad selection borders, will naturally find more candidates as discussed in Sect. 4.2.1.

#### 4.2.6. Concluding remarks on the comparisons

In general, the visual selection methods used recently on Gaia data of Sco-Cen (Damiani et al. 2019; Luhman 2022a), produce in total a ~15% larger number of candidates (Table 3) when compared to unsupervised machine learning methods within the same area (e.g., Kerr et al. 2021, and this work). This is mainly because these methods select by eye broad regions in projected sub-spaces of the multi-dimensional phase space to identify Sco-Cen candidates. On the other hand, unsupervised machine learning methods find more spatial and kinematical substructure in the Sco-Cen population, and produce samples with lower contamination level when compared with visual selection methods.

When compared to other unsupervised methods that studied the whole Sco-Cen area (in particular Kerr et al. 2021), the SigMA clusters are often richer. More importantly, for describing the formation process of OB associations such as Sco-Cen, the SigMA method reveals not only more clusters but a more complex velocity structure across the entire Sco-Cen.

Focusing on the US region, we find generally good agreement for US clusters from SGB21, KRK21, and SigMA. Not surprisingly, the denser clusters in US ($\rho$ Oph/L1688, $\nu$ Sco, $\delta$ Sco, $\beta$ Sco) have been all recovered to some extend by the different approaches. The Antares cluster and also the $\sigma$ Sco cluster are slightly more dispersed, especially in SigMA, and they have less clear matches across the methods. The newly identified velocity substructure in the US region, as revealed with Gaia data, is relevant to understand the star formation processes at play in OB associations like Sco-Cen and will be an obvious target with future Gaia releases, where the additional radial velocity information (increasing by a factor of about 5 for Gaia DR3) will be critical to further characterize these clusters.

Finally, SigMA appears to miss candidates when compared with visual selection methods (of the order of 25%), while, at the same time, finding significant numbers of sources not present in those samples. More work is needed to understand the sources SigMA misses, but at face value, the way forward toward a most complete sample of Sco-Cen members is to use 3D velocities (by including radial velocities) and SigMA cluster members as training sets to the Uncover method (Ratzenböck et al. 2020), a validated bagging classifier of one-class support vector machines (see application in Ratzenböck et al. 2020 to Meingast-1, Meingast et al. 2019b). In the near-future, improved membership lists will allow a more precise analysis of the star formation history of Sco-Cen, the initial mass function of each cluster, and the dynamical state of the Sco-Cen complex.

## 5. Summary

In this paper, we present SigMA, a method that explores the topological properties of a density field to define significant structure. To test and validate SigMA, we apply it to Gaia EDR3 data of the nearest OB association to Earth, Sco-Cen. The main results of this work can be summarized as follows:

1. We present SigMA, a novel clustering method that takes density peaks, separated by dips, as significant clusters. Using a graph-based approach, we detect peaks and dips directly in the multi-dimensional phase space.
2. SigMA is fine-tuned to large-scale surveys in astrophysics. This new method is able to identify co-spatial and co-moving groups with non-convex shapes and variable densities, with a measure of significance. SigMA is able to properly incorporate 5D astrometric uncertainties, does not need any photometric pre-filtering, and scales to millions of points.
3. SigMA is capable of finding clusters in Gaia EDR3 data, reaching stellar volume densities as low as 0.01 stars/pc$^3$ and tangential velocity differences of about 0.3 km/s between clusters.
4. SigMA identifies about $10^4$ Sco-Cen members arranged in 48 clusters of co-spatial and co-moving young stars. The HRD for each cluster shows a narrow and well-defined sequence. Because SigMA is not aware of a star's brightness nor color, the well-defined stellar sequences in the HRD constitute a validation test to the ability of SigMA to extract coeval and co-moving populations.
5. A large fraction of clusters is seen towards well-known Sco-Cen massive stars, too bright to be in Gaia EDR3, and are (tentatively) associated with them. Because SigMA is not aware of these massive stars, the association with clusters also constitutes a validation test to SigMA.
6. When comparing the 48 SigMA stellar populations in Sco-Cen to previous results from the literature we find mostly agreement, however, several discrepancies exist. Visual selection methods used recently on Gaia data of Sco-Cen produce a ~15% larger number of candidates when compared to unsupervised methods. On the other hand, unsupervised methods like SigMA find more spatial and kinematical substructure for the same data set, and produce samples with lower contamination levels.

In the future, in particular with the radial velocities in the upcoming Gaia DR3 data release (plus auxiliary radial velocity surveys), a detailed comparative study of the different clustering methods is fully warranted. The application of SigMA to upcoming Gaia data releases promises the unveiling of detailed clusters distributions like the one presented here but for all the nearest benchmark star-forming regions. Reconstructing an accurate and high-spatial resolution Star Formation History of the last 50 Myr in the Local Milky Way with Gaia data is within reach.

*Acknowledgements.* We acknowledge fruitful discussions with Núria Miret-Roig, Catherine Zucker, Alyssa Goodman, and Charlie Lada. J. Großschedl acknowledges funding by the Austrian Research Promotion Agency (FFG) under project number 873708 (https://www.ffg.at/). This work has used data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. This research has made use of Python, https://www.python.org, of *Astropy*, a community-developed core Python package for Astronomy (Astropy Collaboration et al. 2013, 2018), NumPy (van der Walt et al. 2011), Matplotlib (Hunter 2007), Galpy (Bovy 2015), and Plotly (Plotly Technologies Inc. 2015). This research has used the SIMBAD database operated at CDS, Strasbourg, France (Wenger et al. 2000), of the VizieR catalog access tool, CDS, Strasbourg, France (Ochsenbein et al. 2000), and of "Aladin sky atlas" developed at CDS, Strasbourg Observatory, France (Bonnarel et al. 2000; Boch & Fernique 2014). This research has made use of TOPCAT, an interactive graphical viewer and editor for tabular data (Taylor 2005).

## References

Akaike, H. 1974, IEEE Transactions on Automatic Control, 19, 716

Ashok Kumar, G. 2020, International Journal of Scientific & Technology Research, 9, 6

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33

Azzalini, A. & Torelli, N. 2007, Statistics and Computing, 17, 71

Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Demleitner, M., & Andrae, R. 2021, AJ, 161, 147

Baraffe, I., Chabrier, G., Allard, F., & Hauschildt, P. H. 1998, A&A, 337, 403

Beccari, G., Boffin, H. M. J., & Jerabkova, T. 2020, MNRAS, 491, 2205

Bentley, J. L. 1975, Commun. ACM, 18, 509–517

Biau, G., Chazal, F., Cohen-Steiner, D., Devroye, L., & Rodríguez, C. 2011, Electronic Journal of Statistics, 5, 204

Blaauw, A. 1946, Publications of the Kapteyn Astronomical Laboratory Groningen, 52, 1

Blaauw, A. 1952, Bull. Astron. Inst. Netherlands, 11, 414

Blaauw, A. 1964a, ARA&A, 2, 213

Blaauw, A. 1964b, in The Galaxy and the Magellanic Clouds, ed. F. J. Kerr, Vol. 20, 50

Boch, T. & Fernique, P. 2014, in Astronomical Society of the Pacific Conference Series, Vol. 485, Astronomical Data Analysis Software and Systems XXIII, ed. N. Manset & P. Forshay, 277

Bonferroni, C. 1936, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8, 3

Bonnarel, F., Fernique, P., Bienaymé, O., et al. 2000, A&AS, 143, 33

Bouvier, J., Matt, S. P., Mohanty, S., et al. 2014, in Protostars and Planets VI, ed. H. Beuther, R. S. Klessen, C. P. Dullemond, & T. Henning, 433

Bouy, H. & Alves, J. 2015, A&A, 584, A26

Bovy, J. 2015, ApJS, 216, 29

Bressan, A., Marigo, P., Girardi, L., et al. 2012, MNRAS, 427, 127

Brooke, T. Y., Huard, T. L., Bourke, T. L., et al. 2007, ApJ, 655, 364

Burman, P. & Polonik, W. 2009, Journal of Multivariate Analysis, 100, 1198

Burrows, A., Hubbard, W. B., Lunine, J. I., & Liebert, J. 2001, Reviews of Modern Physics, 73, 719

Campello, R. J., Moulavi, D., & Sander, J. 2013, in Pacific-Asia conference on knowledge discovery and data mining, Springer, 160–172

Cantat-Gaudin, T. & Anders, F. 2020, A&A, 633, A99

Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al. 2018, A&A, 618, A93

Cantat-Gaudin, T., Krone-Martins, A., Sedaghat, N., et al. 2019, A&A, 624, A126

Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, A&A, 618, A59

Chaudhuri, K. & Dasgupta, S. 2010, in Advances in Neural Information Processing Systems, ed. J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta, Vol. 23 (Curran Associates, Inc.)

Chaudhuri, K., Dasgupta, S., Kpotufe, S., & von Luxburg, U. 2014, IEEE Transactions on Information Theory, 60, 7900

Chazal, F., Guibas, L. J., Oudot, S. Y., & Skraba, P. 2013, J. ACM, 60

Chen, B., D'Onghia, E., Alves, J., & Adamo, A. 2020, A&A, 643, A114

Chen, Y., Bressan, A., Girardi, L., et al. 2015, MNRAS, 452, 1068

Chen, Y., Girardi, L., Bressan, A., et al. 2014, MNRAS, 444, 2525

Cheng, Y. 1995, IEEE Transactions on Pattern Analysis and Machine Intelligence, 17, 790

Chronis, P., Athanasiou, S., & Skiadopoulos, S. 2019, in 2019 IEEE International Conference on Data Mining (ICDM), 91–100

Comaniciu, D. & Meer, P. 2002, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 603

Correa, C. & Lindstrom, P. 2011, IEEE Transactions on Visualization and Computer Graphics, 17, 1852

Cropper, M., Katz, D., Sartoretti, P., et al. 2018, A&A, 616, A5

Damiani, F., Prisinzano, L., Pillitteri, I., Micela, G., & Sciortino, S. 2019, A&A, 623, A112

Dasgupta, S. & Kpotufe, S. 2014, in Advances in Neural Information Processing Systems, ed. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger, Vol. 27 (Curran Associates, Inc.)

de Bruijne, J. H. J. 1999, MNRAS, 310, 585

de Geus, E. J. 1992, A&A, 262, 258

de Geus, E. J., de Zeeuw, P. T., & Lub, J. 1989, A&A, 216, 44

de Zeeuw, P. T., Hoogerwerf, R., de Bruijne, J., Brown, A., & Blaauw, A. 2001, in Encyclopedia of Astronomy and Astrophysics, ed. P. Murdin (IOP Publishing Ltd), 1915

de Zeeuw, P. T., Hoogerwerf, R., de Bruijne, J. H. J., Brown, A. G. A., & Blaauw, A. 1999, AJ, 117, 354

Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977, J. R. Stat. Soc. Series B Stat. Methodol., 39, 1

Diehl, R., Lang, M. G., Martin, P., et al. 2010, A&A, 522, A51

Dieterich, S. B., Henry, T. J., Jao, W.-C., et al. 2014, AJ, 147, 94

Ding, J., Shah, S., & Condon, A. 2016, Bioinformatics, 32, 2567

Dobbie, P. D., Lodieu, N., & Sharp, R. G. 2010, MNRAS, 409, 1002

Edelsbrunner, H., Letscher, D., & Zomorodian, A. 2000, in Proceedings 41st Annual Symposium on Foundations of Computer Science, 454–463

Esplin, T. L. & Luhman, K. L. 2022, AJ, 163, 64

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96 (AAAI Press), 226–231

Favata, F. & Micela, G. 2003, Space Sci. Rev., 108, 577

Feigelson, E. D. & Montmerle, T. 1999, ARA&A, 37, 363

Feng, Y. & Hamerly, G. 2007, in Advances in Neural Information Processing Systems, ed. B. Schölkopf, J. Platt, & T. Hoffman, Vol. 19 (MIT Press)

Fern, X. Z. & Brodley, C. E. 2004, in Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04 (New York, NY, USA: Association for Computing Machinery), 36

Fernández, D., Figueras, F., & Torra, J. 2008, A&A, 480, 735

Forbes, J. C., Alves, J., & Lin, D. N. C. 2021, Nature Astronomy, 5, 1009

Fürnkranz, V., Meingast, S., & Alves, J. 2019, A&A, 624, L11

Gabriel, K. R. & Sokal, R. R. 1969, Systematic Biology, 18, 259

Gagné, J., Mamajek, E. E., Malo, L., et al. 2018, ApJ, 856, 23

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021a, A&A, 649, A1

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021b, A&A, 650, C3

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2016, A&A, 595, A2

Galli, P. A. B., Bouy, H., Olivares, J., et al. 2020, A&A, 634, A98

Galli, P. A. B., Joncour, I., & Moraux, E. 2018, MNRAS, 477, L50

Ghrist, R. 2008, Bulletin of the American Mathematical Society, 45, 61

Goldman, B., Röser, S., Schilbach, E., Moór, A. C., & Henning, T. 2018, ApJ, 868, 32

Grasser, N., Ratzenböck, S., Alves, J., et al. 2021, A&A, 652, A2

Gritschneder, M. & Lin, D. N. C. 2012, ApJ, 754, L13

Großschedl, J. E., Alves, J., Meingast, S., & Herbst-Kiss, G. 2021, A&A, 647, A91

Hamerly, G. & Elkan, C. 2004, in Advances in Neural Information Processing Systems, ed. S. Thrun, L. Saul, & B. Schölkopf, Vol. 16 (MIT Press)

Hartigan, J. A. 1975, Clustering Algorithms, 99th edn. (USA: John Wiley & Sons, Inc.)

Hartigan, J. A. & Hartigan, P. M. 1985, The Annals of Statistics, 13, 70

Hu, X. & Xu, L. 2003, in Intelligent Data Engineering and Automated Learning, ed. J. Liu, Y.-m. Cheung, & H. Yin (Berlin, Heidelberg: Springer Berlin Heidelberg), 195–202

Huber, P. J. 1964, The Annals of Mathematical Statistics, 35, 73

Hunt, E. L. & Reffert, S. 2021, A&A, 646, A104

Hunter, J. D. 2007, Computing in Science and Engineering, 9, 90

Jaromczyk, J. & Toussaint, G. 1992, Proceedings of the IEEE, 80, 1502

Jerabkova, T., Boffin, H. M. J., Beccari, G., & Anderson, R. I. 2019, MNRAS, 489, 4418

Jerabkova, T., Boffin, H. M. J., Beccari, G., et al. 2021, A&A, 647, A137

Kalogeratos, A. & Likas, A. 2012, in Advances in Neural Information Processing Systems, ed. F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger, Vol. 25 (Curran Associates, Inc.)

Kamdar, H., Conroy, C., Ting, Y.-S., & El-Badry, K. 2021, ApJ, 922, 49

Kerr, R. M. P., Rizzuto, A. C., Kraus, A. L., & Offner, S. S. R. 2021, ApJ, 917, 23

Kirkpatrick, D. G. & Radke, J. D. 1985, in Machine Intelligence and Pattern Recognition, Vol. 2, Computational Geometry, ed. G. T. TOUSSAINT (North-Holland), 217–248

Koontz, Narendra, & Fukunaga. 1976, IEEE Transactions on Computers, C-25, 936

Kounkel, M. & Covey, K. 2019, AJS, 158, 122

Kounkel, M., Covey, K., & Stassun, K. G. 2020, AJ, 160, 279
Kpotufe, S. & von Luxburg, U. 2011, in Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11 (Madison, WI, USA: Omnipress), 225–232
Kushniruk, I., Schirmer, T., & Bensby, T. 2017, A&A, 608, A73
Leike, R. H., Glatzle, M., & Enßlin, T. A. 2020, A&A, 639, A138
Lépine, J. R. D. & Sartori, M. J. 2003, in Astrophysics and Space Science Library, Vol. 299, Astrophysics and Space Science Library, ed. J. Lépine & J. Gregorio-Hetem, 63
Lifshitz, L. & Pizer, S. 1990, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12, 529
Lombardi, M., Alves, J., & Lada, C. J. 2006, A&A, 454, 781
Luhman, K. L. 2022a, AJ, 163, 24
Luhman, K. L. & Esplin, T. L. 2020, AJ, 160, 44
Luri, X., Brown, A. G. A., Sarro, L. M., et al. 2018, A&A, 616, A9
Magnani, L., Blitz, L., & Mundy, L. 1985, ApJ, 295, 402
Makarov, V. V. 2007, ApJ, 670, 1225
Makarov, V. V. 2008, ApJS, 169, 105
Mamajek, E. E. & Feigelson, E. D. 2001, in Astronomical Society of the Pacific Conference Series, Vol. 244, Young Stars Near Earth: Progress and Prospects, ed. R. Jayawardhana & T. Greene, 104–115
Mamajek, E. E., Lawson, W. A., & Feigelson, E. D. 1999, ApJ, 516, L77
Mamajek, E. E., Lawson, W. A., & Feigelson, E. D. 2000, ApJ, 544, 356
Marigo, P., Girardi, L., Bressan, A., et al. 2017, ApJ, 835, 77
Maurus, S. & Plant, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York, NY, USA: Association for Computing Machinery), 1055–1064
McInnes, L., Healy, J., & Astels, S. 2017, The Journal of Open Source Software, 2
Meingast, S. & Alves, J. 2019a, A&A, 621, L3
Meingast, S., Alves, J., & Fürnkranz, V. 2019b, A&A, 622, L13
Meingast, S., Alves, J., & Rottensteiner, A. 2021, A&A, 645, A84
Merloni, A., Nandra, K., & Predehl, P. 2020, Nature Astronomy, 4, 634
Miret-Roig, N., Bouy, H., Raymond, S. N., et al. 2022, Nature Astronomy, 6, 89
Miret-Roig, N., Galli, P. A. B., Brandner, W., et al. 2020, A&A, 642, A179
Muller, D. W. & Sawitzki, G. 1991, Journal of the American Statistical Association, 86, 738
Nocedal, J. & Wright, S. J. 1999, Numerical Optimization (Springer, New York, NY)
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23
Oh, S., Price-Whelan, A. M., Hogg, D. W., Morton, T. D., & Spergel, D. N. 2017, AJ, 153, 257
Ohnaka, K., Hofmann, K. H., Schertl, D., et al. 2013, A&A, 555, A24
Olivares, J., Bouy, H., Sarro, L. M., et al. 2021, A&A, 649, A159
Pagani, L., Bacmann, A., Motte, F., et al. 2004, A&A, 417, 605
Pagani, L., Lagache, G., Bacmann, A., et al. 2003, A&A, 406, L59
Pagani, L., Pardo, J. R., Apponi, A. J., Bacmann, A., & Cabrit, S. 2005, A&A, 429, 181
Pecaut, M. J. & Mamajek, E. E. 2013, ApJS, 208, 9
Pecaut, M. J. & Mamajek, E. E. 2016, MNRAS, 461, 794
Pecaut, M. J., Mamajek, E. E., & Bubar, E. J. 2012, ApJ, 746, 154
Plotly Technologies Inc. 2015, Collaborative data science, Montreal, QC, https://plot.ly
Pöppel, W. G. L., Bajaja, E., Arnal, E. M., & Morras, R. 2010, A&A, 512, A83
Preibisch, T. & Mamajek, E. 2008, in Handbook of Star Forming Regions, Volume II, ed. B. Reipurth, Vol. 5 (Reipurth, B.), 235
Preibisch, T. & Zinnecker, H. 1999, AJ, 117, 2381
Randich, S., Schmitt, J. H. M. M., Prosser, C. F., & Stauffer, J. R. 1995, A&A, 300, 134
Ratzenböck, S., Meingast, S., Alves, J., Möller, T., & Bomze, I. 2020, A&A, 639, A64
Reininghaus, J., Kotava, N., Guenther, D., et al. 2011, IEEE Transactions on Visualization and Computer Graphics, 17, 2045
Riello, M., De Angeli, F., Evans, D. W., et al. 2021, A&A, 649, A3
Rizzuto, A. C., Ireland, M. J., & Robertson, J. G. 2011, MNRAS, 416, 3108
Roccatagliata, V., Sacco, G. G., Franciosini, E., & Randich, S. 2018, A&A, 617, L4
Román-Zúñiga, C. G., Alves, J. F., Lada, C. J., & Lombardi, M. 2010, ApJ, 725, 2232
Román-Zúñiga, C. G., Lada, C. J., Muench, A., & Alves, J. F. 2007, ApJ, 664, 357
Röser, S., Schilbach, E., & Goldman, B. 2019, A&A, 621, L2
Röser, S., Schilbach, E., Goldman, B., et al. 2018, A&A, 614, A81
Rybizki, J., Green, G. M., Rix, H.-W., et al. 2022, MNRAS, 510, 2597
Sartori, M. J., Lépine, J. R. D., & Dias, W. S. 2003, A&A, 404, 913
Schmitt, J. H. M. M. 1997, A&A, 318, 215
Schmitt, J. H. M. M., Czesla, S., Freund, S., Robrade, J., & Schneider, P. C. 2021, arXiv e-prints, arXiv:2106.14549
Schönrich, R., Binney, J., & Dehnen, W. 2010, MNRAS, 403, 1829
Schwarz, G. 1978, The Annals of Statistics, 6, 461
Squicciarini, V., Gratton, R., Bonavita, M., & Mesa, D. 2021, MNRAS, 507, 1381
Stauffer, J. R., Hartmann, L. W., Prosser, C. F., et al. 1997, ApJ, 479, 776
Steinacker, J., Andersen, M., Thi, W. F., et al. 2015, A&A, 582, A70
Steinacker, J., Pagani, L., Bacmann, A., & Guieu, S. 2010, A&A, 511, A9
Strehl, A. & Ghosh, J. 2002, Journal of machine learning research, 3, 583
Stuetzle, W. & Nugent, R. 2010, Journal of Computational and Graphical Statistics, 19, 397
Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29
Teixeira, P. S., Scholz, A., & Alves, J. 2020, A&A, 642, A86
Torra, F., Castañeda, J., Fabricius, C., et al. 2021, A&A, 649, A10
van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science and Engineering, 13, 22
van Leeuwen, F. 2007, A&A, 474, 653
van Leeuwen, F. 2009, A&A, 497, 209
Vedaldi, A. & Soatto, S. 2008, in Computer Vision – ECCV 2008, ed. D. Forsyth, P. Torr, & A. Zisserman (Berlin, Heidelberg: Springer Berlin Heidelberg), 705–718
Vega-Pons, S. & Ruiz-Shulcloper, J. 2011, International Journal of Pattern Recognition and Artificial Intelligence, 25, 337
Villa Vélez, J. A., Brown, A. G. A., & Kenworthy, M. A. 2018, Research Notes of the American Astronomical Society, 2, 58
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9
Wishart, D. 1969, Mode analysis, a generalization of nearest neighbour which reduces chaining
Witkin, A. P. 1987, in Readings in Computer Vision, ed. M. A. Fischler & O. Firschein (San Francisco (CA): Morgan Kaufmann), 329–332
Wright, N. J. & Mamajek, E. E. 2018, MNRAS, 476, 381
Zari, E., Brown, A. G. A., de Bruijne, J., Manara, C. F., & de Zeeuw, P. T. 2017, A&A, 608, A148
Zari, E., Brown, A. G. A., & de Zeeuw, P. T. 2019, A&A, 628, A123
Zomorodian, A. & Carlsson, G. 2005, Discrete & Computational Geometry, 33, 249
Zucker, C., Goodman, A., Alves, J., et al. 2021, ApJ, 919, 35
Zucker, C., Goodman, A. A., Alves, J., et al. 2022, Nature, 601, 334
Zucker, C., Speagle, J. S., Schlafly, E. F., et al. 2019, ApJ, 879, 125

## Appendix A: Gaia EDR3 data retrieval

The Gaia EDR3 data was downloaded from the Gaia Archive[17] using the following ADQL query:

```
SELECT * FROM gaiaedr3.gaia_source
WHERE (1000./parallax*COS(l*PI()/180)*COS(b*PI()/180))>-50
AND (1000./parallax*COS(l*PI()/180)*COS(b*PI()/180))<250
AND (1000./parallax*SIN(l*PI()/180)*COS(b*PI()/180))>-200
AND (1000./parallax*SIN(l*PI()/180)*COS(b*PI()/180))<50
AND (1000./parallax*SIN(b*PI()/180))>-95
AND (1000./parallax*SIN(b*PI()/180))<100
AND (parallax_error/parallax)<0.2
AND parallax>0.
AND pmra_error<2.
AND pmdec_error<2.
```

The parameter fidelity_v2 from Rybizki et al. (2022) was retrieved with the following ADQL query, using the Topcat TAP Query and the GAVO service[18]:

```
SELECT mine.source_id, gaia.*
FROM gedr3spur.main AS gaia
JOIN tap_upload.t1 AS mine
USING (source_id)
```

In Sect. 4.2 we compare the SigMA clusters with recent literature samples. To this end we cross-match the samples using the Gaia EDR3 source_id. This cross-match is straight forward for the samples of Schmitt et al. (2021), Squicciarini et al. (2021), and Luhman (2022a), who also used Gaia EDR3 data. In the case of Damiani et al. (2019) and Kerr et al. (2021), who used Gaia DR2 data, we first retrieve the Gaia EDR3 source_id using the gaiaedr3.dr2_neighbourhood catalog from the Gaia Archive, since the DR2 and EDR3 source_ids are not generally the same. Such a cross-match delivers few sources that have several possible matches of EDR3 wit DR2 sources (see Torra et al. 2021; Gaia Collaboration et al. 2021a). In such cases we choose the closer match, using the provided angular_distance parameter.

## Appendix B: Modality test procedure by Burman & Polonik (2009)

Here we highlight the work of Burman & Polonik (2009) more closely, who's modality test procedure we adopt in this work. The modality procedure is tied to the notion of a density dip along a path between two points in the data set. In the following, we aim to define the concept of such a path formally.

We consider directed, continuous paths from $\boldsymbol{x}_1$ to $\boldsymbol{x}_2$ through input space $\mathcal{X}$. By assuming there exists a parametrization $\boldsymbol{r}(t)$, with $t \in [0, 1]$, the path becomes the image of $\boldsymbol{r}(t)$. With this map, we can uniquely express every point on the path via the parameter $t$. For example, its start and endpoints are given by $\boldsymbol{x}_1 = \boldsymbol{r}(0)$ and $\boldsymbol{x}_2 = \boldsymbol{r}(1)$, respectively.

Let $f$ be the underlying density function and $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ two candidate modes of $f$. We assume, without loss of generality, that $f(\boldsymbol{x}_1) < f(\boldsymbol{x}_2)$. If all possible paths undergo a density dip when moving from $\boldsymbol{x}_1$ to $\boldsymbol{x}_2$, both points are found in two distinct modal regions:

$$\exists\, t \in (0, 1) : f(\boldsymbol{r}(t)) < f(\boldsymbol{x}_1) \tag{B.1}$$

Conversely, if we can find a path between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ where all points have a higher density than $\boldsymbol{x}_1$, both points are part of the same modal region:

$$f(\boldsymbol{r}(t)) \geq f(\boldsymbol{x}_1) \quad \forall\, t \in (0, 1] \tag{B.2}$$

Eq. (B.2) describes the case of single-modality, which constitutes the null hypothesis we aim to reject. For general pairs of modal candidates it becomes:

$$f(\boldsymbol{r}(t)) \geq \min(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)) \quad \forall\, t \in (0, 1] \tag{B.3}$$

An equivalent and useful formulation is obtained by taking the logarithm on both sides; after that the left side is subtracted from the inequality.

$$\mathrm{SB}(t) := -\log f(\boldsymbol{r}(t)) + \min(\log f(\boldsymbol{x}_1),\ \log f(\boldsymbol{x}_2)) \tag{B.4}$$

Using the variable SB($t$), we can formulate the null hypothesis as follows:

$$\mathrm{H}_0 : \mathrm{SB}(t) \leq 0 \qquad \forall\, t \in (0, 1) \tag{B.5}$$

Rather than testing $\mathrm{H}_0$ across the full path a point-wise test $\mathrm{H}_{0,t}$ : $\mathrm{SB}(t) \leq 0$ for some values of $t$ is employed.

Since we do not have access to the underlying density $f$, we cannot test the hypothesis in Eq. (B.5) directly. Instead, we have a data set of $d$-dimensional random variables drawn from $f$. Given proper normalization of the coordinate axes (see Sect. 3.3.3), Burman & Polonik (2009) show that the following expression is asymptotically standard normal distributed and converges – up to a constant factor – to SB($t$) as the number data samples approaches infinity:

$$\widehat{\mathrm{SB}}(t) = d\,\sqrt{k/2}\,[\log d_k(\boldsymbol{r}(t)) - \max(\log d_k(\boldsymbol{x}_1),\ \log d_k(\boldsymbol{x}_2))] \tag{B.6}$$

Here $d_k(\boldsymbol{x})$ denotes the distance to the $k$'th nearest neighbor of the point $\boldsymbol{x}$. The distance is an approximation to the density $f$. Due to their inverse proportionality the sign is flipped between Eq. (B.4) and Eq. (B.6); and the minimum is replaced with the maximum function.

Since the corresponding test statistic $\widehat{\mathrm{SB}}(t)$ is approximately standard normally distributed, the null hypothesis is rejected at significance level $\alpha$ if

$$\widehat{\mathrm{SB}}(t) \geq \Phi^{-1}(1 - \alpha) \tag{B.7}$$

where $\Phi$ is the standard normal cdf. Therefore, if any $t \in (0, 1)$ fulfills condition (B.7), $\mathrm{H}_0$ is rejected.

Due to the employment of $k$ nearest neighbor technique, this test procedure applies naturally to multivariate data without the need of projecting the data onto a one-dimensional line, as is the case for most modality tests. Furthermore, nearest neighbor queries have access to very efficient algorithms such as the Kd-tree (Bentley 1975) which reduces neighbor searches to only $O(\log N)$ distance computation. Thus, these considerations allow us to study the modality structure of the data set at Gaia data scales without careful projection loss considerations.

Burman & Polonik (2009) describe the iterative application of the test procedure to modal candidates to cluster the data into significant modal regions. However, the test is employed along the straight line path connecting two modes, which limits the procedure to convex cluster shapes only. Moreover, to detect significant dips reliably, enough samples need to be tested along the path.

We aim to provide a natural extension to the presented procedure, which applies to arbitrary cluster shapes while reducing the number of point-wise tests to a single one; see Sect. 3.2 for a detailed description.

[17] https://gea.esac.esa.int/archive/

[18] German Astrophysical Virtual Observatory, https://dc.zah.uni-heidelberg.de/

## Appendix C: Scaling factor distribution

Here we discuss the derivation of the scaling factor distribution, which we use to weigh the velocity sub-space in the clustering process. For a more detailed motivation see Sect. 3.3.3.

We replace the scaling factor variable $c_v$ with $y$ to simplify and shorten the reading flow. Additionally, compared to the main text, we denote the distance to a cluster with $r$ instead of $d$. This notation makes the integration alongside the differential $dr$ easier to read (otherwise the differential would be $dd$).

Our goal is to obtain the distribution $f(y \mid r_0 \leq r \leq r_1)$, which describes the behavior of the scaling factor $y$ for a given range of distances to groups of interest. A simple way to find this distribution is to interpret the empirical linear model $g(r)$ and associated Gaussian uncertainties as an improper probability function $f(r, y)$[19].

As we are dealing with an improper pdf, we consider the following proportionality condition and handle the normalization of the left hand side later.

$$\begin{aligned} f(y \mid r_0 \leq r \leq r_1) &\propto \int_{r_0}^{r_1} f(r, y)\, dr \\ &\propto \int_{r_0}^{r_1} f(y \mid r) f(r)\, dr \end{aligned} \tag{C.1}$$

Since $f(r) \propto 1$ is independent on the distance $r$ we can add it to the yet unknown constant normalization factor and move it out of the integral. Hence we can write the target distribution as:

$$f(y \mid r_0 \leq r \leq r_1) \propto \int_{r_0}^{r_1} f(y \mid r)\, dr \tag{C.2}$$

Thus, to obtain an analytic solution to Eq. (C.2) we need an expression for the conditional pdf $f(y \mid r)$. Assuming that the data are Gaussian distributed around the linear model with a constant standard deviation $\sigma$[20], we can write the following expression:

$$f(y \mid r) \propto \exp\left(-\frac{(y - g(r))^2}{2\sigma^2}\right) \tag{C.3}$$

Figure C.1 schematically shows the integrating process where the conditional pdfs $f(y \mid r)$ are shown for $r = 100$ and $r = 200$.

By substituting Eq. (C.3) into Eq. (C.2) and solving the integral we obtain:

$$f(y \mid r_0 \leq r \leq r_1) \propto \operatorname{erf}\left(\frac{y - g(r_0)}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{y - g(r_1)}{\sqrt{2}\sigma}\right) \tag{C.4}$$

where erf($x$) is the error function. To normalize the probability density in Eq. (C.4), we compute its integral. Since both summands are of the same type, we only need to solve the following integral:

$$I(c) = \int_{-\infty}^{+\infty} \operatorname{erf}\left(\frac{y - c}{\sqrt{2}\sigma}\right) dy \tag{C.5}$$

The variable $c$ represents the constants $g(r_0)$ and $g(r_1)$. The integral in Eq. (C.5) evaluates to:

$$I(c) = \exp\left(\frac{-(y - c)^2}{2\sigma^2}\right) \sqrt{\frac{2\sigma^2}{\pi}} + (y - c) \operatorname{erf}\left(\frac{y - c}{\sqrt{2}\sigma}\right) \Bigg|_{-\infty}^{+\infty} \tag{C.6}$$

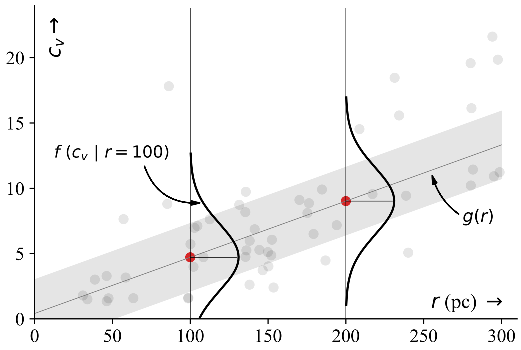


**Fig. C.1.** Scaling factor determination via the empirical distance-scaling relationship. The scaling factor distribution for groups at a distance between 100 – 200 pc depends on the conditional distribution of scaling factors at a given distance $f(c_x/c_v \mid r)$.

Thus, the integral of Eq. (C.4) can be expressed in the following form:

$$\int_{-\infty}^{+\infty} f(y \mid r_0 \leq r \leq r_1) = n \times [I(g(r_0)) - I(g(r_1))] \overset{!}{=} 1 \tag{C.7}$$

The factor $n$ represents the normalization factor. Rearranging the resulting terms by function type, we get the following:

$$I(g(r_0)) - I(g(r_1)) = h(y) + l(y) \Bigg|_{-\infty}^{+\infty} \tag{C.8}$$

The functions $h(y)$ and $l(y)$ describe a sum of exponential and error functions, respectively. The functions are defined in the following:

$$\begin{aligned} h(y) &= \sqrt{\frac{2\sigma^2}{\pi}} \left[\exp\left(\frac{-(y - g(r_0))^2}{2\sigma^2}\right) - \exp\left(\frac{-(y - g(r_1))^2}{2\sigma^2}\right)\right] \\ l(y) &= (y - g(r_0)) \operatorname{erf}\left(\frac{y - g(r_0)}{\sqrt{2}\sigma}\right) - (y - g(r_1)) \operatorname{erf}\left(\frac{y - g(r_1)}{\sqrt{2}\sigma}\right) \end{aligned} \tag{C.9}$$

We evaluate the summands of this primitive integral at the border individually. First, sum Gaussians in $h(y)$ goes to zero as $y$ approaches negative and positive infinity:

$$\lim_{y \to \pm\infty} h(y) = 0 \tag{C.10}$$

The sum of error functions can be rearranged into the following form:

$$\begin{aligned} l(y) = y &\left[\operatorname{erf}\left(\frac{y - g(r_0)}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{y - g(r_1)}{\sqrt{2}\sigma}\right)\right] \\ &+ g(r_1) \operatorname{erf}\left(\frac{y - g(r_1)}{\sqrt{2}\sigma}\right) - g(r_0) \operatorname{erf}\left(\frac{y - g(r_0)}{\sqrt{2}\sigma}\right) \end{aligned} \tag{C.11}$$

Evaluating $l(y)$ at the borders results in the following:

$$\lim_{y \to \pm\infty} y \left[\operatorname{erf}\left(\frac{y - g(r_0)}{\sqrt{2}\sigma}\right) - \operatorname{erf}\left(\frac{y - g(r_1)}{\sqrt{2}\sigma}\right)\right] = 0 \tag{C.12}$$

and

$$\begin{aligned} \lim_{y \to \pm\infty} &\left[g(r_1) \operatorname{erf}\left(\frac{y - g(r_1)}{\sqrt{2}\sigma}\right) - g(r_0) \operatorname{erf}\left(\frac{y - g(r_0)}{\sqrt{2}\sigma}\right)\right] = \\ &= \pm(g(r_1) - g(r_0)) \end{aligned} \tag{C.13}$$

[19] Since the marginal distribution $f(r) \propto 1$ is a uniform distribution over $\mathbb{R}^+$, the joint distribution $f(r, y)$ is improper as it does not integrate to unity.

[20] We observe the standard deviation to be approximately constant for the range of interest; $r \in [100, 200]$

This last term is the only non-zero contribution to the integral. Its evaluation at the lower edge results in the same but negative value to the upper edge. Thus, the area under the curve is twice that value. The normalization factor $n$ then becomes the following:

$$I(g(r_0)) - I(g(r_1)) = 2[g(r_1) - g(r_0)]$$
$$n = \frac{1}{2[g(r_1) - g(r_0)]} =: \frac{1}{2\Delta g} \tag{C.14}$$

The function value difference, $\Delta g$, is always positive since $g(r)$ is a strictly monotonically increasing function and $r_1 > r_0$, see Fig. C.1. Thus, $n$ is a proper normalization factor that is non-zero and positive for all pairs $r_0$ and $r_1$. Using this normalization constant, the conditional pdf can be written as:

$$f(y \mid r_0 \leq r \leq r_1) = \frac{1}{2\Delta g}\left[\mathrm{erf}\left(\frac{y - g(r_0)}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{y - g(r_1)}{\sqrt{2}\sigma}\right)\right] \tag{C.15}$$

The top part of Figure C.2 shows the resulting pdf when applying Eq. (C.4) to sources in Sco-Cen, where we assume a distance range of $r \in [100, 200]$. Here we can see an immediate caveat of our simple constant model uncertainty assumption; the resulting distribution has infinite support, thus, non-zero probability density for $f(y < 0 \mid r)$. Although physically meaningless, the total probability of such events is small and, as seen below, does not drastically influence the final sample set.

We consider sampling strategies to obtain a set of scaling factors to use in the clustering process. Random sampling can generate almost identical realizations, so the possible solution space might not be covered evenly. Since we need to perform a separate clustering, run for each sample drawn, keeping the number as small as possible is essential. To cover the space evenly while considering the underlying probability distribution, we select a set of 10 samples that represent 10 quantiles of the pdf. We separate the pdf into 10 continuous intervals with equal probabilities from which we derive samples as the mean position of these intervals.

To compute the quantiles, we determine the cdf by solving the integral over the conditional pdf in Eq. (C.15); using functions $h(y)$ and $l(y)$ defined in Eq. (C.9) the cdf becomes:

$$\begin{aligned} F(y \mid r_0 \leq r \leq r_1) &= \int f(y \mid r_0 \leq r \leq r_1)\, dy \\ &= \frac{1}{2\Delta g}[h(y) + l(y)] + C \end{aligned} \tag{C.16}$$

To obtain a proper cdf from Eq. (C.16), we set the constant of integration $C$ to $1/2$. Thus, the cdf becomes:

$$F(y \mid r_0 \leq r \leq r_1) = \frac{1}{2}\left(1 + \frac{1}{\Delta g}[h(y) + l(y)]\right) \tag{C.17}$$

The cdf defined in Eq. (C.16) for $r \in [100, 200]$ is shown in the bottom part of Fig. C.2. The ten red scatter points[21] indicate samples drawn from the 10-quantile splitting procedure where horizontal lines indicate equal probability intervals. To invert the cdf and obtain scaling fraction samples from $F^{-1}(y \mid r_0 \leq r \leq r_1)$ we used a numerical approximation[22].

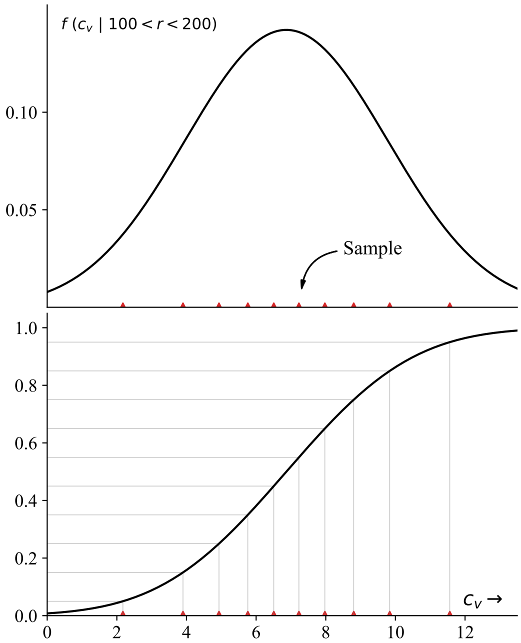


**Fig. C.2.** Pdf and cdf of scaling factor conditioned on a given range of distances. The ten red scatter points indicate samples drawn from the 10-quantile splitting procedure. We separate the pdf into 10 continuous intervals with equal probabilities from which we derive samples as the mean position of these intervals.

## Appendix D: Projected velocities

The reflex motion of the Sun influences how the observed tangential velocities are distributed in $v_\alpha/v_\delta$ space. In Figure D.1 we show theoretical positions of objects if they follow a circular orbit around the Galactic center at the given positions within the Galactic potential. The orbits are estimated within a Milky Way potential including a disk, bulge, and halo component, using the python package `galpy` by Bovy (2015) (`galpy.potential.MWPotential2014`; `galpy.potential.vcirc`) and assuming the local standard of rest (LSR) velocity from Schönrich et al. (2010). The projected motions are given for all Galactic longitude ($l$) positions at distances ($d$) of 100 pc and 200 pc and at Galactic latitudes ($b$) of −20°, 0°, and 25°. These $d$ and $b$ ranges encompass the Sco-Cen region, which reaches from about $l = 0°$ to 290°. The members of Sco-Cen within the selected SigMA clusters are plotted as gray dots in Fig. D.1. Overall, the young stellar groups in Sco-Cen seem to roughly follow expected motions in our Galaxy assuming LSR velocities. The figure additionally highlights the issues the come with the projected tangential velocity plane $v_\alpha/v_\delta$, which is a function of position in the sky and distance of a source. Very nearby sources, like in nearby young local associations, cloud cover large areas of this plane,

[21] The velocity scaling values are: $c_v = \{2.17, 3.9, 4.93, 5.76, 6.51, 7.23, 7.97, 8.8, 9.84, 11.56\}$.

[22] We made use of the open source library `pynverse v0.1.4.4` to calculate the numerical inverse of the cdf.

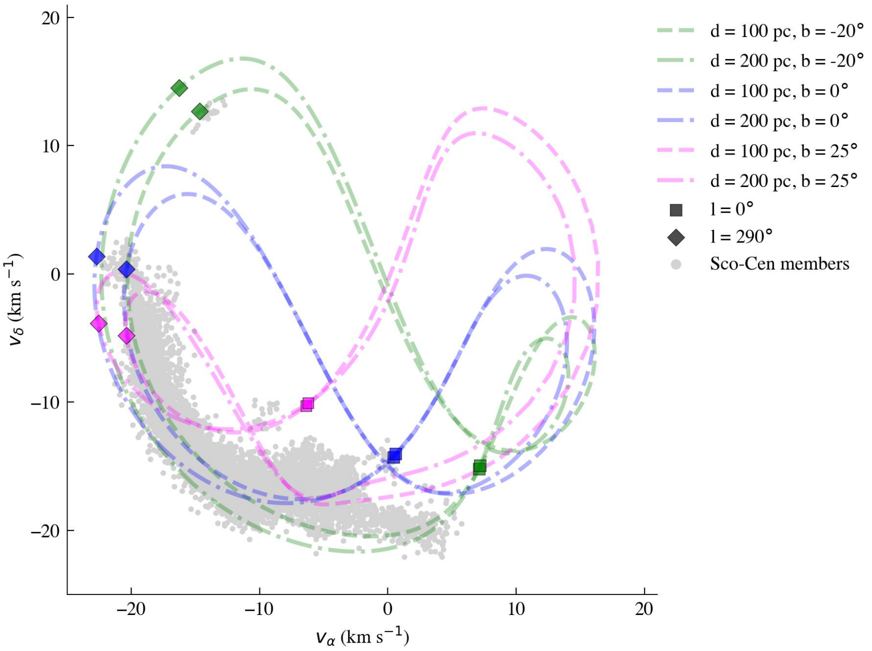


**Fig. D.1.** Tangential velocities in the $v_\alpha/v_\delta$ plane of theoretical sources with circular Galactic orbits and LSR velocities. Shown are six different cases, while each of the lines represents sources at all $l$ positions. The six cases are for two different distances (100 pc, dashed lines; 200 pc, dash-dotted lines), and for three different $b$ positions ($b = -20°$, green; $b = 0°$, blue; $b = 25°$, magenta). The indicated longitude positions at $l = 0°$ (box symbols) and $l = 290°$ (diamond symbols) roughly mark the eastern and western borders of Sco-Cen. The `SigMA` selected Sco-Cen members are shown with gray dots. See also Fig. 8 for a separation of the clusters.

while being at the same time confined in 3D velocity space (UVW).

## Appendix E: The Gaia EDR3 HRD

The HRD in Fig. 9 in Sect. 4 shows a color-absolute-magnitude diagram using the magnitudes from the Gaia EDR3 passbands $G$ versus $BP - RP$. The $G$ band is corrected as recommended in Gaia Collaboration et al. (2021b)[23], and the absolute magnitude $G_{\rm abs}$ is calculated with the distance modulus using the inverse of the parallax as distance. We applied the following quality criteria to the photometry, which mainly affects faint sources.

$$\begin{aligned} & G_{\rm err} < 0.006 \\ \wedge\, & BP_{\rm err} < 0.1 \\ \wedge\, & RP_{\rm err} < 0.02 \end{aligned} \tag{E.1}$$

This cut reduces lower quality measurements, mostly found scattered at the low-mass regime. The magnitude errors are calculated as follows:

$$\begin{aligned} G_{\rm err} &= 1.0857/\texttt{phot\_g\_mean\_flux\_over\_error} \\ BP_{\rm err} &= 1.0857/\texttt{phot\_bp\_mean\_flux\_over\_error} \\ RP_{\rm err} &= 1.0857/\texttt{phot\_rp\_mean\_flux\_over\_error} \end{aligned} \tag{E.2}$$

The isochrone in Fig. 9 shows a 25 Myr PARSEC isochrone[24] for Gaia EDR3 passbands (e.g., Bressan et al. 2012; Chen et al. 2014, 2015; Marigo et al. 2017; Riello et al. 2021), assuming solar metallicity (metal fraction $z = 0.0152$) and no extinction. To get a measure for the contamination from older sources (older than the expected $\lesssim 20$ Myr), we select sources to the left of a 25 Myr isochrone, allowing for random scatter around the 20 Myr isochrone. Additionally, we do not consider sources at the upper–main-sequence (UMS), since there the trend reverses (younger sources are to the left of the UMS). Hence, we apply a cut at $G_{\rm abs} > 3$ mag, only selecting fainter sources as older-source candidates. An additional selection cut is applied at the low-mass regime, where we find that a sequence of older sources is clearly discernible, while the majority of low-mass stars is scattered around the 25 Myr low-mass–isochrone and at younger ages. Since the models for low-mass stars are not as well developed as for intermediate- and high-mass stars, and the overall uncertainties for fainter sources are generally higher, we apply an additional cut, as shown by the bottom dashed slope in Fig. 9 (middle panel). This slope is defined as follows, selecting older sources to the left of it:

$$G_{\rm abs} > 3 \cdot (BP - RP - 7.4) + 1.95 \tag{E.3}$$

The combined conditions deliver 333 out of 8392 sources with applied photometric quality criteria, hence about 4%, which are possible contaminants from older populations within the SigMA clusters. Considering the chosen borders, we like to stress that this separation can only be seen as rough estimate, in particular because we made the additional slope cut by hand (Eq. (E.3)), hence, the contamination by old sources is likely at least about 4% or somewhat higher. Without the cut in Eq. (E.3)

[23] https://github.com/agabrown/gaiaedr3-6p-gband-correction

[24] http://stev.oapd.inaf.it/cgi-bin/cmd

and using only the 25 Myr isochrone plus the cut at the UMS, the contamination fraction from older sources could be up to 15%.

Finally, to get an estimate of sub-stellar sources in our sample, we use a 0.09 $M_\odot$ iso-mass line in Fig. 9 (right panel), which is again extracted from PARSEC models using ages from $10^4$ to $10^{10}$ yr, to get a wide range. The lowest masses in the PARSEC models are given for 0.09 $M_\odot$, hence we select sources below this line to get sources with masses of about 0.08 $M_\odot$ (hydrogen-burning limit) and lower (e.g., Baraffe et al. 1998; Burrows et al. 2001; Dieterich et al. 2014). The uncertainties at the low-mass regime make this selection only a rough estimate. With the cut we find that there are 552 out of 8392 (6.6%), or out of 8059 (6.8%) sources (considering either all sources from the left panel or only the younger sources from the middle panel in Fig. 9). This selection indicates a fraction of sub-stellar objects of about 6–7% within the SigMA clusters.

## Appendix F: Comparisons with selected literature samples

Here we provide two additional tables, giving an overview of the literature comparisions between the SigMA clusters and the Sco-Cen samples in Damiani et al. (2019) and Squicciarini et al. (2021) in Table F.1, and Kerr et al. (2021) in Table F.2. More details on the comparissons can be found in the main part of this paper in Sect. 4.2.

**Table F.1.** Comparing the SigMA clusters with stellar group selections from Damiani et al. (2019) and from Squicciarini et al. (2021). Only those SigMA groups which have cross-matches with either of the two literature samples are given here.

| SigMA | Name (SigMA) | Nr[a] | Matches with DDP19[b] | Matches with SGP21[c] |
|---|---|---|---|---|
| 1 (US) | $\rho$ Oph/L1688 | 463 | US-f(3)US-n(287)US-D2(57)N(14) | G1(404)G4(1)D(27) |
| 2 | $\nu$ Sco | 139 | US-n(59)US-D2(62)N(2) | G2(105)G3(2)G6(8)D(19) |
| 3 | $\delta$ Sco | 388 | US-f(22)US-n(49)D1(29)US-D2(254)N(1) | G1(6)G2(1)G3(330)G4(3)G5(10)G6(2)G8(1)D(26) |
| 4 | $\beta$ Sco | 147 | US-f(51)US-n(1)US-D2(80)N(3) | G3(4)G4(100)G6(33)D(5) |
| 5 | $\beta$ Sco-South | 28 | US-f(3)US-n(10)US-D2(9) | G4(19)G5(1)D(6) |
| 6 | $\sigma$ Sco | 354 | US-f(141)US-n(2)D1(5)D2a(1)US-D2(162)N(3) | G4(1)G5(79)D(250) |
| 7 | Antares | 449 | US-f(68)US-n(48)D1(50)D2a(4)US-D2(210)N(15) | G1(11)G3(3)G7(41)G8(27)D(308) |
| 8 | Scorpio-Body | 315 | D1(2)D2a(208)US-D2(22)N(5) | D(1) |
| 9 | US-foreground-3 | 46 | D1(37)N(1) | |
| 10 | US-foreground-1 | 170 | D1(124)US-D2(1)N(10) | D(1) |
| 11 | US-foreground-2 | 59 | D1(35)US-D2(2)N(8) | D(12) |
| 12 (UCL) | Lupus-3 | 139 | LupIII(67)D2a(25)D2b(27)N(1) | |
| 13 | Lupus-4 | 23 | D2a(2)D2b(21) | |
| 14 | $\epsilon$ Norma | 69 | D2a(58)D2b(1)N(1) | |
| 15 | V1062 Sco | 794 | UCL-1(528)D1(8)D2a(109)N(8) | |
| 16 | Lupus-West | 112 | UCL-2(47)D1(4)D2b(47)N(2) | |
| 17 | Lupus-1 | 110 | UCL-3(1)D1(55)D2a(19)D2b(12)N(3) | |
| 18 | $\psi$02 Lup | 229 | D1(169)D2a(8)D2b(12)N(1) | D(3) |
| 19 | $\nu$ Cen | 897 | D1(23)D2b(684)N(33) | |
| 20 | $\rho$ Lup | 116 | D1(32)D2b(64)N(4) | |
| 21 | V795 Cen | 351 | D1(48)D2b(247)N(4) | |
| 22 | $\eta$ Lup | 242 | D1(214)D2a(2)D2b(2)N(2) | |
| 23 | $b$ Cen | 546 | UCL-3(2)D1(354)D2a(1)D2b(82)N(18) | |
| 24 | V1019 Cen | 188 | D2a(12)D2b(141)N(2) | |
| 25 | Lupus-East | 87 | D1(5)D2a(48)D2b(19) | |
| 26 | $\mu$ Sco | 80 | UCL-1(51)D1(2)D2a(7)N(1) | |
| 27 | $e$ Lup | 139 | D1(122)D2b(4) | |
| 28 | $\phi$02 Lup | 116 | UCL-3(40)D1(48)D2a(3)D2b(13)N(3) | |
| 29 | Libra-South | 38 | D1(5)D2a(6)D2b(15)US-D2(2) | |
| 32 (LCC) | Musca-foreground | 67 | D2b(28)N(1) | |
| 33 | Centaurus-Far | 24 | D2b(20) | |
| 34 | $\alpha$ Musca | 64 | D1(1)D2b(55) | |
| 35 | Acrux | 215 | LCC-1(84)D1(1)D2b(107)N(1) | |
| 36 | $\sigma$ Cen | 1417 | D1(25)D2b(1116)N(41) | |
| 37 | $f$ Cen | 326 | D1(8)D2b(265)N(3) | |
| 38 (Lupus) | B59 | 21 | N(15) | |
| 40 | Pipe-foreground | 29 | D1(8)D2a(1) | |
| 42 | $\theta$ Oph | 82 | D2a(32)US-D2(2)N(2) | |

**Notes.** [(a)] Number of sources from this work, for a direct comparison with the cross-matches as given in brackets in Cols. 4–5. [(b)] The DPP19 group shortcuts are given for eight compact clusterings (UCL-1, UCL-2, UCL-3, Lupus 3, LCC-1, US-far, US-near), for four diffuse populations (D1, D2a, D2b, US-D2), and for sources that have not been assigned to any of these groups (N), while the number in brackets gives the cross-matches with the respective SigMA cluster. [(c)] The SGB21 groups (G) are numbered from 1 to 8, and their diffuse population is given with D. Again, the number of cross-matches is given in brackets. The eight groups in SGB21 are associated with the brightest star in each group as follows: G1–$i$ Sco; G2–$\nu$ Sco B; G3–$b$ Sco; G4–HD 144273; G5–HIP 77900; G6–HIP 78968; G7–HIP 79910; G8–HD 146467.

**Table F.2.** Comparing the SigMA clusters with Kerr et al. (2021) clusters toward Sco-Cen.

| SigMA | Name (SigMA) | Nr[a] | TLC[b] | EOM[c] | LEAF[d] | Name (KRK21)[e] |
|---|---|---|---|---|---|---|
| 1 | $\rho$ Oph/L1688 | 463 | 22(275) | 17(249) | I(102) | UpperSco-I/$\rho$ Oph |
| 2 | $\nu$ Sco | 139 | 22(85) | 17(83) | E(53) | UpperSco-E |
| 3 | $\delta$ Sco | 388 | 22(260) | 17(258) | H(90)I(1) | UpperSco-H |
| 4 | $\beta$ Sco | 147 | 22(106) | 17(105) | G(28) | UpperSco-G |
| 5 | $\beta$ Sco-South | 28 | 22(12) | 17(10) |  | UpperSco |
| 6 | $\sigma$ Sco | 354 | 22(228) | 17(194) | C(16)D(22) | UpperSco-C,D |
| 7 | Antares | 449 | 22(309) | 17(257) | A(3)B(10)F(25) | UpperSco-A,B,F |
| 8 | Scorpio-Body | 315 | 22(164) | 16(12)17(41) |  | UpperSco |
| 9 | US-foreground-3 | 46 | 22(20) | 9(1) |  |  |
| 10 | US-foreground-1 | 170 | 22(76) | 13(26) |  | EOM13 |
| 11 | US-foreground-2 | 59 | 22(27) | 13(3)17(1) |  |  |
| 12 | Lupus-3 | 139 | 22(95) | 12(81) | A(46) | Lupus-IV |
| 13 | Lupus-4 | 23 | 22(20) | 12(19) | B(14) | Lupus-III |
| 14 | $\epsilon$ Norma | 69 | 22(47) | 14(17) |  | EOM14 |
| 15 | V1062 Sco | 794 | 22(409) | 14(3)15(337) |  | LowerSco |
| 16 | Lupus-West | 112 | 22(50) | 11(33) |  | UPK606 |
| 17 | Lupus-1 | 110 | 22(59) | 23(5) |  |  |
| 18 | $\psi$02 Lup | 229 | 22(110) | 17(7) |  | UpperSco |
| 19 | $\nu$ Cen | 897 | 22(294) | 11(1)24(107)26(3) |  | EOM24 |
| 20 | $\rho$ Lup | 116 | 22(53) |  |  |  |
| 21 | V795 Cen | 351 | 22(186) | 11(2)21(11)25(11) |  |  |
| 22 | $\eta$ Lup | 242 | 22(161) | 22(101) |  | EOM22 |
| 23 | $b$ Cen | 546 | 22(182) | 23(1) |  |  |
| 24 | V1019 Cen | 188 | 22(45) |  |  |  |
| 25 | Lupus-East | 87 | 22(50) |  |  |  |
| 26 | $\mu$ Sco | 80 | 22(37) | 15(31) |  | LowerSco |
| 27 | $e$ Lup | 139 | 22(85) | 20(73) |  | EOM20 |
| 28 | $\phi$02 Lup | 116 | 22(37) | 23(5) |  |  |
| 29 | Libra-South | 38 | 22(10) |  |  |  |
| 30 | $\eta$ Cham | 20 | 22(16) | 18(16) |  | $\eta$ Cham |
| 31 | $\epsilon$ Cham | 41 | 22(27) | 27(23) | A(16) | LCC-A/$\epsilon$ Cham |
| 32 | Musca-foreground | 67 | 22(46) | 27(37) | B(13) | LCC-B |
| 33 | Centaurus-Far | 24 | 21(18) | 3(18) |  | Cen-South |
| 34 | $\alpha$ Musca | 64 | 22(44) | 27(36) | B(1)C(10) | LCC-C/Crux-South |
| 35 | Acrux | 215 | 22(153) | 27(132) | B(1)C(80) | LCC-C/Crux-South |
| 36 | $\sigma$ Cen | 1417 | 22(823) | 27(441) | C(4)D(12)E(47) | LCC-C,D,E |
| 37 | $f$ Cen | 326 | 22(191) | 26(36)27(2) |  | EOM26 |
| 38 | B59 | 21 | 22(10) | 6(9) |  | Pipe |
| 39 | Sgr-West | 15 | 22(3) |  |  |  |
| 40 | Pipe-foreground | 29 | 22(13) | 9(12) |  | EOM9 |
| 41 | Pipe-North | 22 | 22(10) |  |  |  |
| 42 | $\theta$ Oph | 82 | 22(41) | 10(28) |  | Theia67 |
| 43 | Corona Australis | 124 | 22(71) | 8(70) |  | CrA |
| 44 | CrA-North | 265 | 22(173) | 7(1)8(162) |  | CrA |
| 45 | Scorpio-Sting | 36 | 22(17) | 7(10) |  | EOM7 |
| 46 | Chamaeleon-1 | 148 | 21(93) | 1(93) |  | Chamaeleon-1 |
| 47 | Chamaeleon-2 | 40 | 21(26) | 2(26) |  | Chamaeleon-2 |
| 48 | L134/L183 | 20 | 22(5) |  |  |  |

**Notes.** [(a)] Number of sources from this work, for a direct comparison with the number of cross-matches as given in brackets in Cols. 4–6. [(b)] The numbers give the KRK21 TLC group, with the number of cross-matches in brackets. There have been only cross-matches with the TLC groups 21 and 22. [(c)] The numbers give the KRK21 EOM sub-group, with the number of cross-matches in brackets, wile each EOM represents a sub-clustering within the lower level TLC group. [(d)] The letters give the KRK21 LEAF sub-group, with the number of cross-matches in brackets, wile a LEAF group represents a sub-clustering within the lower level EOM group. [(e)] Group names from KRK21, if significant overlap with SigMA clusters was present. Only the (sub)group with the most significant number of cross-matches is given, as apparent from the numbers in brackets in Cols. 5–6.

# 5. Conclusion

In the following, the main results of this thesis are summarized in Sect. 5.1 and an outlook on future work and follow-up projects are discussed in Sect. 5.2.

## 5.1. Summary of results

The main results of this thesis are two-fold: First, the thesis proposes two analysis techniques for the detailed study of stellar structures. Second, the application of these tools to Gaia DR2 and EDR3 update stellar cluster catalogs from which several domain results can be derived.

1. In this thesis, the analysis pipeline `Uncover` was developed, facilitating the use of powerful one-class support vector machines for extensive membership searches. Since principled model selection for one-class models remains an open problem [110], this work proposed selection heuristics involving interpretable summary statistics. In Ratzenböck et al. [94], we defined six complementary summary statistics based on the number and distributional characteristics of yet unidentified cluster members. These statistics, such as expected ranges on velocity dispersion and positional extent, are specified in relation to high-fidelity members used to train OCSVM models. This work showed that by sampling random hyper-parameters and rejecting models if they do not adhere to a priori-defined summary statistics ranges, effectively builds a bagging classifier of one-class support vector machines[1].

2. Building on results from Ratzenböck et al. [94], `Uncover` was extended to work with vague prior knowledge. As summary statistics ranges could not be determined a priori, in Grasser et al. [51] we searched for suitable model ensembles and the corresponding summary statistics ranges. This work determined an objective that aimed to maximize the number of inferred members while minimizing the contamination fraction defined by comparing the 3D velocity distribution of training and inferred members[2]. Finally, suitable ensembles were selected in the "number of

[1] In cases of high model flexibility and unknown field star contamination content in the training set, bagging improves accuracy and reduces variance in the prediction [50]

[2] The contamination fraction is determined by comparing the 3D velocity distribution of training members to inferred candidate sources. Precisely, the training samples are first modeled as a 3D Gaussian distribution in velocity space (mean and covariance matrix are determined by maximizing the likelihood of the training data). The contamination fraction is the number of candidate members outside the $3\sigma$ (99.7%) region of training sources compared to the total number of inferred sources (with a valid radial velocity measurement). Since only stars with radial velocity measurements have access to the full 3D velocity information, the contamination fraction is a very rough estimate of the

candidates" and "contamination fraction" plane. The selected model ensemble, i.e., the final classifier, was validated by considering the distribution of inferred sources in the HRD, which was previously untouched information. Inferred members can also be validated via their isochronal age in comparison to training set sources. The residual distribution of inferred candidates and training set sources to the best fitting isochrone (to the training set) provided strong evidence that uncovered sources are actual members of the $\rho$ Oph system.

3. The application of `Uncover` to the recently discovered Meingast 1 stream [82] found about 2000 high-fidelity stream members, increasing the source population approximately tenfold. As the newly predicted stream members are no longer limited by radial velocity measurements (as was the case in the discovery paper), the new selection substantially extended the main sequence to unveil the stream's population across the entire stellar mass spectrum, from B stars to M stars[3], including white dwarfs. The comparison in the HRD of the newly identified stream members with the Pleiades cluster (apart from being slightly more metal poor) suggested a similar age, correcting the original age estimate which was from 1 Gyr to $\sim 110$ Myr. In the mass range of $\sim 0.2 < M_\odot < \sim 4\,M_\odot$, this work identified a normal IMF which allowed an estimation of the total mass of the stream to approximately 2000 $M_\odot$, making it by far the most massive stream in the solar neighborhood. In addition, this work was able to assign several white dwarfs to the Meingast 1 stream.

4. The application of `Uncover` to the $\rho$ Oph region found 191 new young stellar object (YSO) candidates in Gaia EDR3 belonging to the $\rho$ Oph system. An analysis of stellar types revealed that these new sources appear to be mainly Class III M stars and substellar objects. A total of 28 new members showed excess infrared emission suggesting the presence of circumstellar dusty disks. The proper motion analysis of the $\rho$ Oph region revealed a bi-modal structure, suggesting the presence of two main populations: the first population (1022 sources) comprises clusters of young stars around the $\rho$ Oph star and the main Ophiuchus clouds (L1688, L1689, L1709). The second population (304 sources) is slightly older and more dispersed, with a similar but distinct proper motion from the first. Both populations occupy approximately the same 3D volume. The second population's age and proper motion suggested that its origin may have originated from the Upper Scorpius (US) population. Finally, the velocity difference of about 4.1 km/s between the two populations suggested a de-mixing of both populations in 3D space in about 4 Myr.

5. In this thesis a design study was performed which resulted in the visualization tool `Uncover`, a further refinement of previously developed extensive membership analysis methods. This tool expands on the quantitative model selection process of previous versions of `Uncover` by introducing astronomers' qualitative judgement of inferred

true contamination.

[3] The spectral classification of stars is subdivided into seven groups using the letters O, B, A, F, G, K, and M. This sequence describes a gradual decrease in temperature, mass, and size from hottest (O type) to the coolest (M type) stars.

cluster candidates into the analysis pipeline. Although no ground truth information is available for individual stars, systems of multiple stars can be validated by domain experts (e.g., using tools such as the HRD). Typically, qualitative model assessment is able to create maximal trust in the final classifier. Further, although we have demonstrated that model selection can be achieved by limiting the contamination fraction, this methodology depends on radial velocity measurements, which are scarce in Gaia EDR3.

To support interactively building interpretable and powerful models in unsupervised scenarios where qualitative model validation is possible, this thesis devised a workflow with the following general guidelines: first, provide an overview of possible model solutions. In the design study, the vast space of possible model configurations was concisely summarized using a hierarchical clustering approach. Different model clustering solutions could be explored by users who control the granularity of model groups. Second, to support the model evaluation additional validation tools such as HRD and 3D kinematic information were provided. Third, `Uncover` facilitates prior knowledge declaration and updating on yet unseen cluster members via interpretable summary statistics ranges. To update and substantiate the initial, potentially vague prior knowledge, this work provided the following: the users' qualitative model assessment is translated into updating initial summary statistics rules. Further, users were able to explore correlations between summary statistics via linked heatmaps and perform What-If analyzes to study the effect of individual summary statistics on inferred stars.

In a usability study with nine domain experts and two use cases, users were able to efficiently build effective and high-performance novelty detection models. Further, in a case study we efficiently recovered the second population discovered in Grasser et al. [51] in a single session of `Uncover`.

6. This thesis developed an innovative clustering algorithm `SigMA` that identifies density peaks, separated by substantial dips, as clusters. By using a graph-based approach, `SigMA` detects peaks and dips directly in the multi-dimensional phase space. The method tracks clusters though a family of gradually smoothed density fields, creating a scale space of clustering solutions. The clustering solution is obtained by identifying unchanged and stable clusters in scale-space that are independent from a single density estimate. To integrate observational uncertainties into the clustering procedure, `SigMA` employs a re-sampling strategy from which density deviations in the dip depth across samples are derived. `SigMA` deals with field stars in a two-step approach: first, the cluster's bulk 3D motion is determined. This work defined an objective function that measures the differences in observed proper motions and idealized proper motions given a random bulk motion. By minimizing this objective function `SigMA` is able to approximate the cluster's 3D bulk velocity. Second, the bulk motion is used to compute optimal radial velocities for cluster candidates, which permits a full 6D phase-space analysis to remove field stars based on their phase-space density. Thus, `SigMA` is fine-tuned to large-scale surveys in astrophysics. As shown in Ch. 4, this new method is specialized to identify

co-spatial and co-moving groups with non-convex shapes and variable densities, with a measure of significance. `SigMA` does not need any photometric pre-filtering and scales to millions of points. It is capable of finding clusters in Gaia EDR3 data, reaching stellar volume densities as low as 0.01 stars/pc$^3$ and tangential velocity differences of about 0.3 km/s between clusters.

7. `SigMA` identified about $10^4$ Sco-Cen members arranged in 48 clusters of co-spatial and co-moving young stars. The HRD of each cluster showed a narrow and well-defined sequence. Because `SigMA` is not aware of a star's brightness nor color, the well-defined stellar sequences in the HRD constitute a validation test to the ability of `SigMA` to extract coeval and co-moving populations. This work found that a large fraction of clusters towards Sco-Cen have massive stars, too bright to be in Gaia EDR3, which are (tentatively) associated with them. Because `SigMA` is not aware of these massive stars, the association with clusters also constitutes a validation test to `SigMA`, based on the fact that massive stars are often found at the centers of rich clusters. When comparing the 48 `SigMA` stellar populations in Sco-Cen to previous results from the literature we found mostly agreement, however, several discrepancies exist. Manual selection heuristics (via on-sky and proper motion cuts) used recently on Gaia data of Sco-Cen produce a ∼15% larger number of candidates when compared to unsupervised methods. On the other hand, unsupervised methods like `SigMA` found more spatial and kinematical substructure for the same data set, and produce samples with lower contamination levels.

## 5.2. Future work

The tools and stellar cluster catalogs originated from this thesis have already sparked some potential future work which will be presented in the following section. Several of these projects are currently (as of May 2022) in active development and their preliminary results are briefly outlined.

In Sect. 5.2.1 age determination of the groups identified in Sco-Cen is discussed. In Sect. 5.2.2 the value of `Uncover` and `SigMA` for future clustering applications to the local Milky Way is highlighted. In Sect. 5.2.3 possible extensions to the membership analysis approach of `Uncover` are presented. Finally, in Sect. 5.2.4 further work on `SigMA` in the direction of visual hyper-parameter space exploration is highlighted.

### 5.2.1. The star formation history of Sco-Cen

Sub-populations encode different star formation events and offer a path to understand how the formation process proceeds in time and space inside a cloud, as well as an understanding of the origin of the global velocity dispersion in clusters and associations, critical for the dispersal of young populations into the Galactic field. Knowing the age, motion, size, and mass of these sub-populations will open a new window on how nature forms bounded clusters and associations, by allowing a reconstruction of the sequence of events in a star formation region.

In a next step, the SigMA identified clusters can be used for a precise age study of the 48 stellar groups to reveal the star formation history of the OB association. In a first pilot study we find that the HRD of each cluster shows a narrow and well-defined sequence from which we extract an isochronal age. Thus, we produce a high-resolution age map of the association revealing an older population at the core of Upper Centaurus-Lupus (UCL), and sequential, age ordered branches reaching to outer edges of the 3D distribution of sources in Sco-Cen. Further, we can now precisely date stars inside US, solving its age controversy. What is normally taken in the literature as US consists of 12 clusters with ages between 4 and 17 Myr, naturally explaining the wide age spread. Finally, using this high-resolution age map we compile a catalog with over thousands of Sco-Cen brown dwarf candidates.

### 5.2.2. Application of SigMA and Uncover to the local Milky Way

SigMA can disentangle populations that are moving with velocities as small as 0.3 km/s. In a complementary approach to identifying stellar structures, the application of Uncover ensures a complete source catalog; together with isochronal age dating, these tools can provide an unseen high-resolution age map of the local kpc. Together with Gaia DR3, the legacy value of such a catalog would be huge; its accurate time scales could feed many science cases such as the study of the initial mass function (IMF), star formation history, the origin of associations, timescales for planet formation, and the dispersion of clusters into the Galactic field.

### 5.2.3. Iterative model design

As shown throughout this thesis, OCSVMs are a powerful novelty detection method for extensive analysis of star cluster membership. Their ability to incorporate previously identified star cluster members into the search for new candidates gives them a powerful advantage over fully unsupervised searches. However, the training set also limits the model itself. The decision boundary created during the training process can reach beyond the given training set only to a certain extent. In cases where the training set covers a small sub-region of the entire population, Uncover is likely not able to find large portions of the remaining sources.

To adapt the current analysis workflow to these situations, an iterative and expanding model procedure may be promising. A straight-forward way to facilitate an expanding model is to allow the training set itself to grow in size. Thus, when the classifier infers new candidate members after each training iteration, a second workflow step is added in which a set of new high-fidelity cluster members is determined. To automate the process, some quantitative measures (e.g., based on external factors such as HRD position or 3D velocity in relation to the initial training set) are needed to identify these new training set members.

If and under which conditions this procedure converges needs to be evaluated in future research.

### 5.2.4. Comprehensive visual parameter selection

When discussing the research goals in Sect.1.2.2, this thesis discussed potential research avenues toward a consolidated star clustering approach. To recap, these were: (1) **Visual solution space exploration** alongside clustering result validation options for domain experts (e.g., HRD), (2) **meaningful and interpretable hyper-parameters** that alleviate or facilitate manual solution space exploration, and (3) **internal validation criteria** [79] optimized for star cluster results that enable automatic model selection.

This work focused on the second avenue, which resulted in the development of the innovative `SigMA` analysis pipeline. Although model selection becomes easier with interpretable hyper-parameters, depending on the complexity of input data, the output of `SigMA` still needs some context to handle properly. In particular, exploring the scale-space hierarchy and the influence of different alpha values can become overwhelming without proper visual presentation.

As discussed in Sect. 2.2, this thesis finds that available visualization tools are not transparent about the effects of different input parameters, but rather focus on the clustering results itself. Since a blind trust in machine learning methods and their results can lead to erroneous interpretations of data, interpretable "white box" tools are highly needed in the scientific community. This work identifies the potential to implement the clustering tool `SigMA` in a visual support system. The tool should support the analysis of the sensitivity of different density smoothing parameters, i.e., the scale space, as well as the influence of different significance levels $\alpha$ on the clustering solution.

By providing the environment mentioned above, astronomers may be able to properly reflect on the machine learning approach and the results it provides. In October 2021, a pilot study was started together with master student Johannes Preisinger, in which a series of interviews with eight domain scientists were conducted, which resulted in a list of requirements such a tool has to fulfill. In end of April 2022, a first high-fidelity prototype was presented to a group of 7 (4 completely new to the tool) domain experts who interacted with the tool and provided further feedback towards a successful final version. Future work is needed to deploy a final working version to facilitate large-scale cluster analysis of Gaia data.

# Bibliography

[1] M. Agarwal, K. K. Rao, K. Vaidya, and S. Bhattacharya, “ML-MOC: Machine Learning (kNN and GMM) based Membership determination for Open Clusters”, *Monthly Notices of the Royal Astronomical Society*, vol. 502, no. 2, pp. 2582–2599, 2021.

[2] M. Amer, M. Goldstein, and S. Abdennadher, “Enhancing one-class support vector machines for unsupervised anomaly detection”, in *ODD ’13: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, Chicago, Illinois: ACM, 2013, 8–15.

[3] T. W. Anderson and D. A. Darling, “Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes”, *The Annals of Mathematical Statistics*, vol. 23, no. 2, pp. 193–212, 1952.

[4] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, “Optics: Ordering points to identify the clustering structure”, in *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD ’99, Philadelphia, Pennsylvania, USA: ACM, 1999, 49–60.

[5] A. Azzalini and N. Torelli, “Clustering via nonparametric density estimation”, *Statistics and Computing*, vol. 17, no. 1, pp. 71–80, 2007.

[6] L. Balaguer-Núñez, M. López del Fresno, E. Solano, D. Galadí-Enríquez, C. Jordi, *et al.*, “Clusterix 2.0: A virtual observatory tool to estimate cluster membership probability”, *Monthly Notices of the Royal Astronomical Society*, vol. 492, no. 1, pp. 5811–5843, 2019.

[7] A. Bánhalmi, A. Kocsor, and R. Busa-Fekete, “Counter-Example Generation-Based One-Class Classification”, in *Machine Learning: ECML 2007. Lecture Notes in Computer Science, vol 4701.*, J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič, *et al.*, Eds., Berlin, Heidelberg: Springer, 2007, pp. 543–550.

[8] G. Beccari, H. M. J. Boffin, and T. Jerabkova, “Uncovering a 260 pc wide, 35-Myr-old filamentary relic of star formation”, *Monthly Notices of the Royal Astronomical Society*, vol. 491, no. 2, pp. 2205–2216, 2019.

[9] P. Burman and W. Polonik, “Multivariate mode hunting: Data analytic tools with measures of significance”, *Journal of Multivariate Analysis*, vol. 100, no. 6, pp. 1198–1218, 2009.

[10] R. J. G. B. Campello, D. Moulavi, and J. Sander, “Density-based clustering based on hierarchical density estimates”, in *Advances in Knowledge Discovery and Data Mining*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–172.

[11] T. Cantat-Gaudin and F. Anders, “Clusters and mirages: cataloguing stellar aggregates in the Milky Way”, *Astronomy & Astrophysics*, vol. 633, A99, 2020.

[12] T. Cantat-Gaudin, C. Jordi, A. Vallenari, A. Bragaglia, L. Balaguer-Núñez, *et al.*, “A Gaia DR2 view of the open cluster population in the Milky Way”, *Astronomy & Astrophysics*, vol. 618, A93, 2018.

[13] T. Cantat-Gaudin, A. Krone-Martins, N. Sedaghat, A. Farahi, R. S. de Souza, *et al.*, “Gaia DR2 unravels incompleteness of nearby cluster population: new open clusters in the direction of Perseus”, *Astronomy & Astrophysics*, vol. 624, A126, 2019.

[14] N. Cao, D. Gotz, J. Sun, and H. Qu, “DICON: Interactive Visual Analysis of Multidimensional Clusters”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2581–2590, 2011.

[15] A. Castro-Ginard, C. Jordi, X. Luri, J. Álvarez Cid-Fuentes, L. Casamiquela, *et al.*, “Hunting for open clusters in Gaia DR2: 582 new open clusters in the Galactic disc”, *Astronomy & Astrophysics*, vol. 635, A45, 2020.

[16] A. Castro-Ginard, C. Jordi, X. Luri, T. Cantat-Gaudin, and L. Balaguer-Núñez, “Hunting for open clusters in Gaia DR2: the Galactic anticentre”, *Astronomy & Astrophysics*, vol. 627, A35, 2019.

[17] A. Castro-Ginard, C. Jordi, X. Luri, F. Julbe, M. Morvan, *et al.*, “A new method for unveiling open clusters in gaia. new nearby open clusters confirmed by DR2”, *Astronomy & Astrophysics*, vol. 618, A59, 2018.

[18] M. Cavallo and Ç. Demiralp, “Clustrophile 2: Guided Visual Clustering Analysis”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 267–276, 2019.

[19] K. Chaudhuri and S. Dasgupta, “Rates of convergence for the cluster tree”, in *Advances in Neural Information Processing Systems*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23, Curran Associates, Inc., 2010.

[20] K. Chaudhuri, S. Dasgupta, S. Kpotufe, and U. von Luxburg, “Consistent procedures for cluster tree estimation and pruning”, *IEEE Transactions on Information Theory*, vol. 60, no. 12, pp. 7900–7912, 2014.

[21] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba, “Persistence-Based Clustering in Riemannian Manifolds”, *Journal of the ACM (JACM)*, vol. 60, no. 6, pp. 1–38, 2013.

[22] B. Chen, E. D'Onghia, J. Alves, and A. Adamo, "Discovery of new stellar groups in the Orion complex. Towards a robust unsupervised approach", *Astronomy & Astrophysics*, vol. 643, A114, 2020.

[23] K. Chen and L. Liu, "A visual framework invites human into the clustering process", in *15th International Conference on Scientific and Statistical Database Management, 2003.*, 2003, pp. 97–106.

[24] Y. Cheng, "Mean shift, mode seeking, and clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, 1995.

[25] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[26] S. Das, B. Saket, B. C. Kwon, and A. Endert, "Geono-cluster: Interactive visual cluster analysis for biologists", *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 12, pp. 4401–4412, 2021.

[27] Ç. Demiralp, "Clustrophile: A Tool for Visual Clustering Analysis", in *KDD Workshop on Interactive Data Exploration and Analytics*, ACM, 2016, pp. 37–45.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[29] H. Deng and R. Xu, "Model Selection for Anomaly Detection in Wireless Ad Hoc Networks", in *2007 IEEE Symposium on Computational Intelligence and Data Mining*, IEEE, 2007, pp. 540–546.

[30] J. Ding, S. Shah, and A. Condon, "densityCut: an efficient and versatile topological approach for automatic clustering of biological data", *Bioinformatics*, vol. 32, no. 17, pp. 2567–2576, 2016.

[31] C. Ducourant, R. Teixeira, A. Krone-Martins, S. Bontemps, D. Despois, *et al.*, "Proper motion survey and kinematic analysis of the $\rho$ Ophiuchi embedded cluster", *Astronmy & Astrophysics*, vol. 597, A90, 2017.

[32] C. Désir, S. Bernard, C. Petitjean, and L. Heutte, "One class random forests", *Pattern Recognition*, vol. 46, no. 12, pp. 3490–3506, 2013.

[33] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological persistence and simplification", in *Proceedings 41st Annual Symposium on Foundations of Computer Science*, 2000, pp. 454–463.

[34] L. Ertöz, M. Steinbach, and V. Kumar, "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data", in *Proceedings of the 2003 SIAM International Conference on Data Mining (SDM)*, ser. Proceedings, Society for Industrial and Applied Mathematics, 2003, pp. 47–58.

[35] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96, Portland, Oregon: AAAI Press, 1996, 226–231.

[36] P. F. Evangelista, M. J. Embrechts, and B. K. Szymanski, "Some Properties of the Gaussian Kernel for One Class Learning", in *Artificial Neural Networks – ICANN 2007. Lecture Notes in Computer Science book series (LNCS, volume 4668)*, J. M. de Sá, L. A. Alexandre, W. Duch, and D. Mandic, Eds., Berlin, Heidelberg: Springer, 2007, pp. 269–278.

[37] Y. Feng and G. Hamerly, "PG-means: learning the number of clusters in data", in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., vol. 19, MIT Press, 2007.

[38] P. J. Flower, "Transformations from Theoretical Hertzsprung-Russell Diagrams to Color-Magnitude Diagrams: Effective Temperatures, B-V Colors, and Bolometric Corrections", *The Astrophysical Journal*, vol. 469, p. 355, 1996.

[39] V. Fürnkranz, S. Meingast, and J. Alves, "Extended stellar systems in the solar neighborhood - III. like ships in the night: The coma berenices neighbor moving group", *Astronomy & Astrophysics*, vol. 624, p. L11, 2019.

[40] J. Gagné and J. K. Faherty, "BANYAN. XIII. A First Look at Nearby Young Associations with Gaia Data Release 2", *The Astrophysical Journal*, vol. 862, no. 2, p. 138, 2018.

[41] J. Gagné, E. E. Mamajek, L. Malo, *et al.*, "BANYAN. XI. The BANYAN $\Sigma$ Multivariate Bayesian Algorithm to Identify Members of Young Associations with 150 pc", *The Astrophysical Journal*, vol. 856, no. 1, p. 23, 2018.

[42] Gaia Collaboration, A. G. A. Brown, A. Vallenari, T. Prusti, J. H. J. de Bruijne, *et al.*, "Gaia Data Release 2 - Summary of the contents and survey properties", *Astronomy & Astrophysics*, vol. 616, A1, 2018.

[43] Gaia Collaboration, A. G. A. Brown, A. Vallenari, T. Prusti, J. H. J. de Bruijne, *et al.*, "Gaia Early Data Release 3. Summary of the contents and survey properties", *Astronomy & Astrophysics*, vol. 649, A1, 2021.

[44] Gaia Collaboration, T. Prusti, J. H. J. de Bruijne, A. G. A. Brown, A. Vallenari, *et al.*, "The Gaia mission", *Astronomy & Astrophysics*, vol. 595, A1, 2016.

[45] P. A. B. Galli, I. Joncour, and E. Moraux, "Three-dimensional structure of the Upper Scorpius association with the Gaia first data release", *Monthly Notices of the Royal Astronomical Society: Letters*, vol. 477, no. 1, pp. L50–L54, 2018.

[46] X. Gao, "5D memberships and fundamental properties of the old open cluster NGC 6791 based on *Gaia*-DR2", *Astrophysics and Space Science*, vol. 365, no. 2, p. 24, 2020.

[47] Z. Ghafoori, S. M. Erfani, S. Rajasegarar, J. C. Bezdek, S. Karunasekera, *et al.*, “Efficient Unsupervised Parameter Estimation for One-Class Support Vector Machines”, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 5057–5070, 2018.

[48] Z. Ghafoori, S. Rajasegarar, S. M. Erfani, S. Karunasekera, and C. A. Leckie, “Unsupervised Parameter Estimation for One-Class Support Vector Machines”, in *Advances in Knowledge Discovery and Data Mining*, J. Bailey, L. Khan, T. Washio, G. Dobbie, J. Z. Huang, *et al.*, Eds., Cham: Springer, 2016, pp. 183–195.

[49] R. Ghrist, “Barcodes: The persistent topology of data”, *Bulletin of the American Mathematical Society*, vol. 45, no. 1, pp. 61–75, 2008.

[50] Y. Grandvalet, “Bagging equalizes influence”, *Machine Learning*, vol. 55, no. 3, pp. 251–270, 2004.

[51] N. Grasser, S. Ratzenböck, J. Alves, J. Großschedl, S. Meingast, *et al.*, “The $\rho$ Ophiuchi region revisited with Gaia EDR3 - Two young populations, new members, and old impostors”, *Astronomy & Astrophysics*, vol. 652, A2, 2021.

[52] G. Hamerly and C. Elkan, “Learning the k in k-means”, in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds., vol. 16, MIT Press, 2004.

[53] J. A. Hartigan and P. M. Hartigan, “The dip test of unimodality”, *The Annals of Statistics*, vol. 13, no. 1, pp. 70–84, 1985.

[54] J. A. Hartigan, *Clustering Algorithms*, 99th. USA: John Wiley & Sons, Inc., 1975.

[55] H. Hoffmann, “Kernel PCA for novelty detection”, *Pattern Recognition*, vol. 40, no. 3, pp. 863–874, 2007.

[56] E. L. Hunt and S. Reffert, “Improving the open cluster census. I. Comparison of clustering algorithms applied to Gaia DR2 data”, *Astronomy & Astrophysics*, vol. 646, A104, 2021.

[57] K. Jaehnig, J. Bird, and K. Holley-Bockelmann, “Membership lists for 431 open clusters in gaia DR2 using extreme deconvolution gaussian mixture models”, *The Astrophysical Journal*, vol. 923, no. 1, p. 129, 2021.

[58] T. Jerabkova, G. Beccari, H. M. J. Boffin, M. G. Petr-Gotzens, C. F. Manara, *et al.*, “When the tale comes true: multiple populations and wide binaries in the Orion Nebula Cluster”, *Astronomy & Astrophysics*, vol. 627, A57, 2019.

[59] T. Jerabkova, H. M. J. Boffin, G. Beccari, G. de Marchi, J. H. J. de Bruijne, *et al.*, “The 800 pc long tidal tails of the Hyades star cluster - Possible discovery of candidate epicyclic overdensities from an open star cluster”, *Astronomy & Astrophysics*, vol. 647, A137, 2021.

[60] A. Kalogeratos and A. Likas, “Dip-means: An incremental clustering method for estimating the number of clusters”, in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012.

[61] H. Kamdar, C. Conroy, Y.-S. Ting, A. Bonaca, M. C. Smith, *et al.*, "Stars that Move Together Were Born Together", *The Astrophysical Journal*, vol. 884, no. 2, p. L42, 2019.

[62] R. C. Kennicutt and N. J. Evans, "Star Formation in the Milky Way and Nearby Galaxies", *Annual Review of Astronomy and Astrophysics*, vol. 50, no. 1, pp. 531–608, 2012.

[63] R. M. P. Kerr, A. C. Rizzuto, A. L. Kraus, and S. S. R. Offner, "Stars with Photometrically Young Gaia Luminosities Around the Solar System (SPYGLASS). I. Mapping Young Stellar Structures and Their Star Formation Histories", *The Astrophysical Journal*, vol. 917, p. 23, 2021.

[64] N. V. Kharchenko, A. E. Piskunov, E. Schilbach, S. Röser, and R.-D. Scholz, "Global survey of star clusters in the Milky Way - II. The catalogue of basic parameters", *Astronomy & Astrophysics*, vol. 558, A53, 2013.

[65] S. Khazai, S. Homayouni, A. Safari, and B. Mojaradi, "Anomaly Detection in Hyperspectral Images Based on an Adaptive Support Vector Method", *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 4, pp. 646–650, 2011.

[66] A. Kolmogorov, "Sulla determinazione empirica di una lgge di distribuzione", *Istituto Italiano degli Attuari*, vol. 4, pp. 83–91, 1933.

[67] W. Koontz, P. M. Narendra, and K. Fukunaga, "A Graph-Theoretic Approach to Nonparametric Cluster Analysis", *IEEE Transactions on Computers*, vol. C-25, no. 9, pp. 936–944, 1976.

[68] J. Kos, J. Bland-Hawthorn, M. Asplund, S. Buder, G. F. Lewis, *et al.*, "Discovery of a 21 Myr old stellar population in the Orion complex", *Astronomy & Astrophysics*, vol. 631, A166, 2019.

[69] M. Kounkel and K. Covey, "Untangling the Galaxy. I. Local Structure and Star Formation History of the Milky Way", *The Astronomical Journal*, vol. 158, no. 3, p. 122, 2019.

[70] M. Kounkel, K. Covey, and K. G. Stassun, "Untangling the galaxy. II. structure within 3 kpc", *The Astronomical Journal*, vol. 160, no. 6, p. 279, 2020.

[71] S. Kpotufe and U. von Luxburg, "Pruning nearest neighbor cluster trees", in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11, Bellevue, Washington, USA: Omnipress, 2011, 225–232.

[72] A. Krone-Martins and A. Moitinho, "UPMASK: unsupervised photometric membership assignment in stellar clusters", *Astronomy & Astrophysics*, vol. 561, A57, 2014.

[73] I. Kushniruk, T. Schirmer, and T. Bensby, "Kinematic structures of the solar neighbourhood revealed by gaia DR1/TGAS and RAVE", *Astronomy & Astrophysics*, vol. 608, A73, 2017.

[74] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, *et al.*, "Clustervision: Visual supervision of unsupervised clustering", *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 142–151, 2018.

[75] C. J. Lada and E. A. Lada, "Embedded clusters in molecular clouds", *Annual Review of Astronomy and Astrophysics*, vol. 41, no. 1, pp. 57–115, 2003.

[76] A. Lex, M. Streit, C. Partl, K. Kashofer, and D. Schmalstieg, "Comparative analysis of multidimensional, quantitative data", *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1027–1035, 2010.

[77] L. Liu and X. Pang, "A Catalog of Newly Identified Star Clusters in Gaia DR2", *The Astrophysical Journal Supplement Series*, vol. 245, no. 2, p. 32, 2019.

[78] W. Liu, G. Hua, and J. R. Smith, "Unsupervised One-Class Learning for Automatic Outlier Removal", in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2014, pp. 3826–3833.

[79] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of Internal Clustering Validation Measures", in *2010 IEEE International Conference on Data Mining*, IEEE, 2010, pp. 911–916.

[80] C. F. McKee and E. C. Ostriker, "Theory of star formation", *Annual Review of Astronomy and Astrophysics*, vol. 45, no. 1, pp. 565–687, 2007.

[81] S. Meingast and J. Alves, "Extended stellar systems in the solar neighborhood - I. The tidal tails of the Hyades", *Astronomy & Astrophysics*, vol. 621, p. L3, 2019.

[82] S. Meingast, J. Alves, and V. Fürnkranz, "Extended stellar systems in the solar neighborhood - II. Discovery of a nearby 120° stellar stream in Gaia DR2", *Astronomy & Astrophysics*, vol. 622, p. L13, 2019.

[83] S. Meingast, J. Alves, and A. Rottensteiner, "Extended stellar systems in the solar neighborhood - V. Discovery of coronae of nearby star clusters", *Astronomy & Astrophysics*, vol. 645, A84, 2021.

[84] D. W. Muller and G. Sawitzki, "Excess mass estimates and tests for multimodality", *Journal of the American Statistical Association*, vol. 86, no. 415, pp. 738–746, 1991.

[85] E. J. Nam, Y. Han, K. Mueller, A. Zelenyuk, and D. Imre, "ClusterSculptor: A Visual Analytics Tool for High-Dimensional Data", in *2007 IEEE Symposium on Visual Analytics Science and Technology*, IEEE, 2007, pp. 75–82.

[86] E. R. Newton, A. W. Mann, A. L. Kraus, *et al.*, "TESS hunt for young and maturing exoplanets (THYME). IV. three small planets orbiting a 120 myr old star in the Pisces–Eridanus stream*", *AJS*, vol. 161, no. 2, p. 65, 2021.

[87] S. Oh, A. M. Price-Whelan, D. W. Hogg, T. D. Morton, and D. N. Spergel, "Comoving Stars in Gaia DR1: An Abundance of Very Wide Separation Comoving Pairs", *The Astronomical Journal*, vol. 153, no. 6, p. 257, 2017.

[88] E. Packer, P. Bak, M. Nikkilä, V. Polishchuk, and H. J. Ship, "Visual Analytics for Spatial Clustering: Using a Heuristic Approach for Guided Exploration", *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2179–2188, 2013.

[89] X. Pang, Y. Li, S.-Y. Tang, M. Pasquato, and M. B. N. Kouwenhoven, "Different Fates of Young Star Clusters after Gas Expulsion", *The Astrophysical Journal*, vol. 900, no. 1, p. L4, 2020.

[90] X. Pang, Z. Yu, S.-Y. Tang, J. Hong, Z. Yuan, *et al.*, "Disruption of Hierarchical Clustering in the Vela OB2 Complex and the Cluster Pair Collinder 135 and UBC 7 with Gaia EDR3: Evidence of Supernova Quenching", *The Astrophysical Journal*, vol. 923, no. 1, p. 20, 2021.

[91] K. Peña Ramírez, C. González-Fernández, A.-N. Chené, and S. Ramírez Alegría, "The VVV open cluster project. Near-infrared sequences of NGC 6067, NGC 6259, NGC 4815, Pismis 18, Trumpler 23, and Trumpler 20", *Monthly Notices of the Royal Astronomical Society*, vol. 503, no. 2, pp. 1864–1876, 2021.

[92] M. Perryman, L. Lindegren, J. Kovalevsky, E. Hoeg, U. Bastian, *et al.*, "The HIPPARCOS catalogue", *Astronomy & Astrophysics*, vol. 323, no. 1, pp. 49–52, 1997.

[93] A. Pister, P. Buono, J. D. Fekete, C. Plaisant, and P. Valdivia, "Integrating Prior Knowledge in Mixed-Initiative Social Network Clustering", *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1775–1785, 2021.

[94] S. Ratzenböck, S. Meingast, J. Alves, T. Möller, and I. Bomze, "Extended stellar systems in the solar neighborhood. IV. Meingast 1: the most massive stellar stream in the solar neighborhood", *Astronomy & Astrophysics*, vol. 639, A64, 2020.

[95] S. Röser and E. Schilbach, "A census of the nearby Pisces-Eridanus stellar stream. Commonalities with and disparities from the Pleiades", *Astronomy & Astrophysics*, vol. 638, A9, 2020.

[96] S. Röser, E. Schilbach, and B. Goldman, "Hyades tidal tails revealed by Gaia DR2", *Astronomy & Astrophysics*, vol. 621, p. L2, 2019.

[97] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53–65, 1987.

[98] L. M. Sarro, H. Bouy, A. Berihuete, E. Bertin, E. Moraux, *et al.*, "Cluster membership probabilities from proper motions and multi-wavelength photometric catalogues. I. Method and application to the Pleiades cluster", *Astronomy & Astrophysics*, vol. 563, A45, 2014.

[99] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution", *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[100] T. Schreck, J. Bernard, T. Tekusova, and J. Kohlhammer, “Visual cluster analysis of trajectory data with interactive Kohonen Maps”, in *2008 IEEE Symposium on Visual Analytics Science and Technology*, IEEE, 2008, pp. 3–10.

[101] T. Schultz and G. L. Kindlmann, “Open-Box Spectral Clustering: Applications to Medical Image Analysis”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2100–2108, 2013.

[102] M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Möller, “Visual Parameter Space Analysis: A Conceptual Framework”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2161–2170, 2014.

[103] J. Seo and B. Shneiderman, “Interactively exploring hierarchical clustering results [gene identification]”, *Computer*, vol. 35, no. 7, pp. 80–86, 2002.

[104] S. Sharma and K. V. Johnston, “A GROUP FINDING ALGORITHM FOR MULTIDIMENSIONAL DATA SETS”, *The Astrophysical Journal*, vol. 703, no. 1, pp. 1061–1077, 2009.

[105] G. Sim, S. H. Lee, H. B. Ann, and S. Kim, “207 New Open Star Clusters within 1 kpc from Gaia Data Release 2”, *The Korean Astronomical Society*, vol. 52, 5.

[106] N. Smirnov, “Table for Estimating the Goodness of Fit of Empirical Distributions”, *The Annals of Mathematical Statistics*, vol. 19, no. 2, pp. 279–281, 1948.

[107] W. Stuetzle and R. Nugent, “A generalized single linkage method for estimating the cluster tree of a density”, *Journal of Computational and Graphical Statistics*, vol. 19, no. 2, pp. 397–418, 2010.

[108] S.-Y. Tang, X. Pang, Z. Yuan, W. P. Chen, J. Hong, *et al.*, “Discovery of tidal tails in disrupting open clusters: Coma berenices and a neighbor stellar group”, *The Astrophysical Journal*, vol. 877, no. 1, p. 12, 2019.

[109] D. M. J. Tax and K.-R. Muller, “A consistency-based model selection for one-class classification”, in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3, IEEE, 2004, pp. 363–366.

[110] D. M. J. Tax and R. P. W. Duin, “Uniform Object Generation for Optimizing One-Class Classifiers”, *The Journal of Machine Learning Research*, vol. 2, 155–173, 2002.

[111] A. Vedaldi and S. Soatto, “Quick shift and kernel methods for mode seeking”, in *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 705–718.

[112] S. Wang, J. Yu, E. Lapira, and J. Lee, “A modified support vector data description based novelty detection approach for machinery components”, *Applied Soft Computing*, vol. 13, no. 2, pp. 1193–1205, 2013.

[113] S. Wang, Q. Liu, E. Zhu, F. Porikli, and J. Yin, “Hyperparameter selection of one-class support vector machine by self-adaptive data shifting”, *Pattern Recognition*, vol. 74, pp. 198–211, 2018.

[114] J. L. Ward, J. M. D. Kruijssen, and H.-W. Rix, “Not all stars form in clusters – Gaia-DR2 uncovers the origin of OB associations”, *Monthly Notices of the Royal Astronomical Society*, vol. 495, no. 1, pp. 663–685, 2020.

[115] D. Wishart, “Mode analysis, a generalization of nearest neighbour which reduces chaining”, in *Numerical Taxonomy*, A. J. Cole, Ed., London: Academic Press, 1969, 282–311.

[116] Y. Xiao, H. Wang, and W. Xu, “Parameter Selection of Gaussian Kernel for One-Class SVM”, *IEEE Transactions on Cybernetics*, vol. 45, no. 5, pp. 941–953, 2015.

[117] W. Yang, X. Wang, J. Lu, W. Dou, and S. Liu, “Interactive steering of hierarchical clustering”, *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 10, pp. 3953–3967, 2021.

[118] Z. Yuan, J. Chang, P. Banerjee, J. Han, X. Kang, *et al.*, “StarGO: A new method to identify the galactic origins of halo stars”, *The Astrophysical Journal*, vol. 863, no. 1, p. 26, 2018.

[119] E Zari, A. G. A. Brown, and P. T. de Zeeuw, “Structure, kinematics, and ages of the young stellar populations in the Orion region”, *Astronomy & Astrophysics*, vol. 628, A123, 2019.

[120] A. Zomorodian and G. Carlsson, “Computing persistent homology”, *Discrete & Computational Geometry*, vol. 33, no. 2, pp. 249–274, 2005.

# List of abbreviations

$G$ Graph over the data sample in which data points build the vertices.

$L$ Level-set.

$M_\odot$ Solar mass.

$X$ The observed data.

$\mathbf{x}$ A $d$-dimensional data point.

$\alpha$ Significance level of a statistical hypothesis test.

$\hat{f}$ The estimated density function from data.

$\lambda$ A real valued density threshold.

$\mathcal{X}$ The (compact) support of underlying dentity function $f$.

$f$ The unknown data generating probability density function.

**2D** Two-dimensional.

**3D** Three-dimensional.

**5D** Five-dimensional.

**6D** Six-dimensional.

**B** Spectral classification of the second hottest stars..

**BANYAN** Bayesian Analysis for Nearby Young AssociatioNs.

**DBSCAN** Density-Based Spatial Clustering of Applications with Noise.

**DICON** Dynamic ICON-based interactive visual analysis of multidimensional clusters.

**DR2** Data Release two.

**DR3** Data Release three.

**EDR3** Early Data Release three.

**EM** Expectation–Maximization.

**EOM** Excess Of Mass.

**ESA** European Space Agency.

**G** Gaussian.

**GMM** Gaussian Mixture Models.

**Gyr** Giga year; one billion years.

**$H_0$** Null hypothesis.

**$H_1$** Alternative hypothesis.

**HCE** Hierarchical Cluster Explorer.

**HDBSCAN** Hierarchical DBSCAN.

**HRD** Hertzsprung–Russell Diagram.

**IMF** Initial Mass Function.

**M** Spectral classification of the coolest stars..

**ML-MOC** ML-MOC: Machine Learning based Membership determination for Open Clusters.

**Myr** Mega year; one million years.

**NN** Nearest Neighbors.

**OCSVM** One-Class Support Vector Machines.

**OPTICS** Ordering Points To Identify the Clustering Structure.

**pc** Parsec; 1 pc $\approx 3.086^{16}$ meters.

**PG** Projected Gaussian.

**PK** Prior Knowledge.

**RBF** Radial Basis Function.

**RQ** Research Question.

**Sco-Cen** Scorpius–Centaurus association.

**SigMA** Significant Mode Analysis.

**SNN** Shared Nearest Neighbor.

**SOM** Self-Organizing Map.

**TBD** To Be Determined.

**UCL** Upper Centaurus-Lupus.

**US** Upper Scorpius.

**VPSA** Visual Parameter Space Analysis.

**YSO** Young Stellar Object.

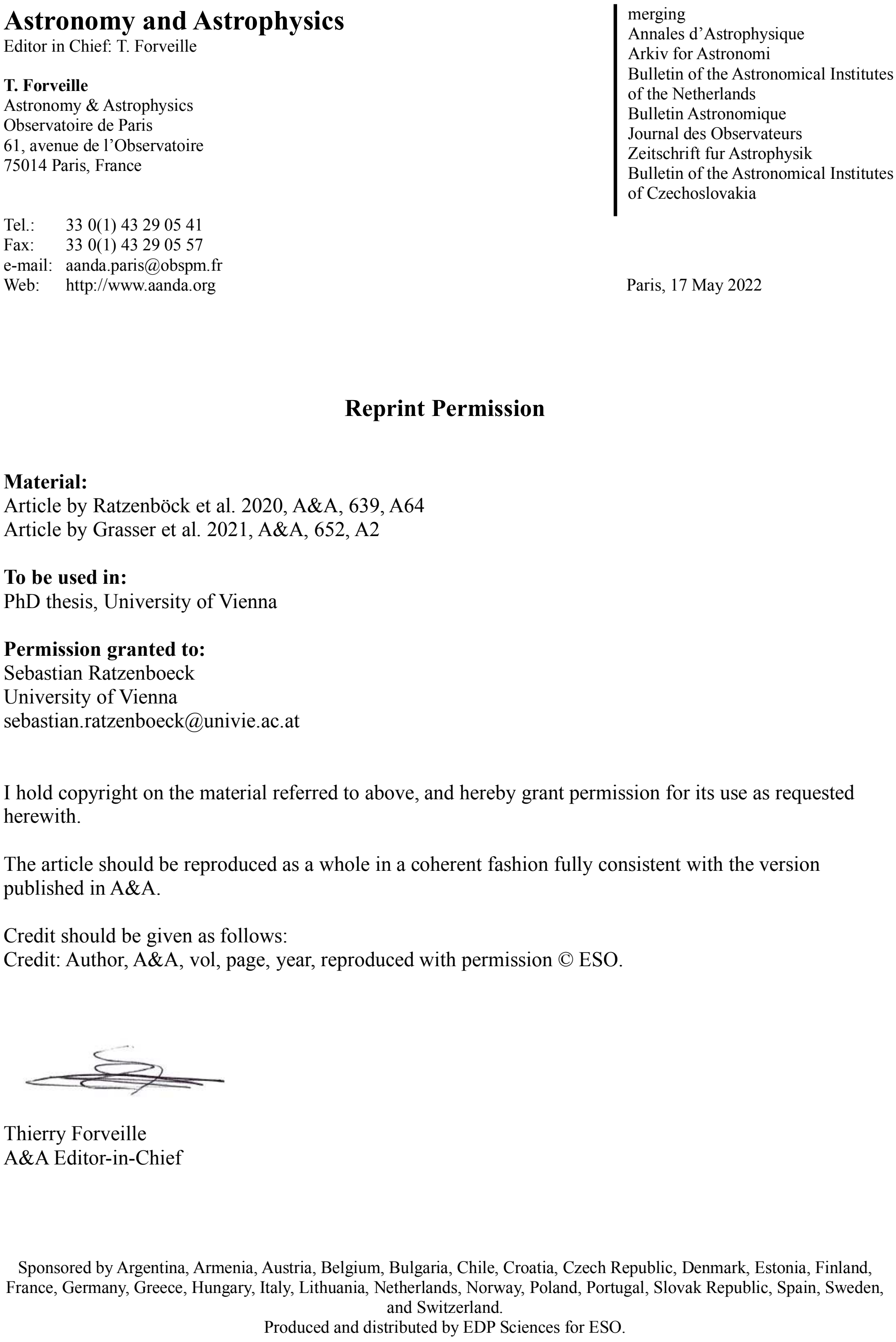
# Appendix

## I. Kurzfassung

Die vorliegende Doktorarbeit beschäftigt sich mit dem Auffinden stellarer Gruppen in der Milchstraße und beabsichtigt es, neue Blickwinkel auf gebräuchliche Methoden in diesem Bereich zu eröffnen. Insbesondere sollen im Zuge der Arbeit transparente Analysemethoden bereitgestellt werden, um bisher unbekannte Sternhaufen sowie neue Mitglieder bekannter stellarer Populationen aufzudecken. Ziel ist es, Techniken für Astronom*innen bereitzustellen, welche ein vollständigeres Bild verschiedener Sternhaufen in der lokalen Milchstraße darstellen können.

Im Rahmen der Dissertation wird zunächst das Verfahren `Uncover` vorgestellt. `Uncover` ist ein Verfahren zur umfangreichen Mitgliederanalyse stellarer Gruppen und kann zuvor entdeckte zugehörige Sterne dieser Gruppen einbeziehen, um nach bisher unentdeckten Sternen zu suchen. Die Methode konnte erfolgreich in zwei Anwendungsfällen durchgeführt werden: bei der kürzlich entdeckten Meingast 1 Gruppe – einem Sternhaufen, das gleichzeitig mit den Plejaden geformt wurde und etwa 120° des Himmels einnimmt – und bei der bereits sehr eingehend erforschten Sternentstehungsregion $\rho$ Ophiuchus. Für diese beiden sehr unterschiedlichen Sternsysteme konnte `Uncover` die Anzahl der gefundenen zugehörigen Sterne um das Zehnfache, somit um etwa 200 Sterne erhöhen. Bei der zweiten Methode zur Auffindung stellarer Gruppen handelt es sich um einen innovativen Clustering-Algorithmus, Significance Mode Analysis (`SigMA`), der die topologischen Eigenschaften der Dichteverteilung im mehrdimensionalen Phasenraum untersucht. Durch die Anwendung von `SigMA` auf Gaia-EDR3-Daten der Scorpius-Centaurus-Assoziation (Sco-Cen) konnten zum ersten Mal 48 sich gemeinsam bewegende und gleichaltrige Cluster in Sco-Cen gefunden werden, von denen viele bisher unbekannt waren. Diese 48 Haufen wurden unabhängig voneinander mit Hilfe von astrophysikalischem Wissen validiert.

Sowohl `Uncover` als auch `SigMA` sind in einer domänenspezifischen Sprache formuliert, verwenden aussagekräftige Hyperparameter und ermöglichen eine Ergebnisvalidierung, um zuverlässige Ergebnisse sicherzustellen. Mit diesen Werkzeugen möchten wir dazu beitragen, die derzeitige Kultur des blinden Vertrauens in Machine-Learning-Tools zu verändern und Astronom*innen dabei helfen, Modelle auf Grundlage ihrer Fachkenntnisse zu erstellen und zu modifizieren.



**Astronomy and Astrophysics**
Editor in Chief: T. Forveille

**T. Forveille**
Astronomy & Astrophysics
Observatoire de Paris
61, avenue de l'Observatoire
75014 Paris, France

Tel.: 33 0(1) 43 29 05 41
Fax: 33 0(1) 43 29 05 57
e-mail: aanda.paris@obspm.fr
Web: http://www.aanda.org

merging
Annales d'Astrophysique
Arkiv for Astronomi
Bulletin of the Astronomical Institutes
of the Netherlands
Bulletin Astronomique
Journal des Observateurs
Zeitschrift fur Astrophysik
Bulletin of the Astronomical Institutes
of Czechoslovakia

Paris, 17 May 2022

**Reprint Permission**

**Material:**
Article by Ratzenböck et al. 2020, A&A, 639, A64
Article by Grasser et al. 2021, A&A, 652, A2

**To be used in:**
PhD thesis, University of Vienna

**Permission granted to:**
Sebastian Ratzenboeck
University of Vienna
sebastian.ratzenboeck@univie.ac.at

I hold copyright on the material referred to above, and hereby grant permission for its use as requested herewith.

The article should be reproduced as a whole in a coherent fashion fully consistent with the version published in A&A.

Credit should be given as follows:
Credit: Author, A&A, vol, page, year, reproduced with permission © ESO.

Thierry Forveille
A&A Editor-in-Chief

Sponsored by Argentina, Armenia, Austria, Belgium, Bulgaria, Chile, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Lithuania, Netherlands, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, and Switzerland.
Produced and distributed by EDP Sciences for ESO.