







Significance mode analysis (SigMA) for hierarchical structures

An application to the Sco-Cen OB association★,★★

Sebastian Ratzenböck^{1,2,3} , Josefa E. Großschedl¹ , Torsten Möller^{2,3} , João Alves^{1,2} ,
Immanuel Bomze^{2,4} , and Stefan Meingast¹ 

¹ University of Vienna, Department of Astrophysics, Türkenschanzstraße 17, 1180 Vienna, Austria
e-mail: sebastian.ratzenboeck@univie.ac.at

² University of Vienna, Research Network Data Science at Uni Vienna, Kolingasse 14-16, 1090 Vienna, Austria

³ University of Vienna, Faculty of Computer Science, Währinger Straße 29/S6, 1090 Vienna, Austria

⁴ University of Vienna, ISOR/VCOR, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

Received 31 March 2022 / Accepted 16 May 2023

ABSTRACT

We present a new clustering method, significance mode analysis (SigMA), for extracting 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 validated SigMA on simulated clusters and find that it outperforms competing methods, especially in cases where many clusters are closely spaced. We applied the new method to *Gaia* DR3 data of the closest OB association to Earth, Scorpio-Centaurus (Sco-Cen), and find more than 13 000 co-moving young objects, about 19% of which have a substellar mass. SigMA finds 37 co-moving clusters in Sco-Cen. These clusters are independently validated by their narrow Hertzsprung-Russell diagram sequences and, to a certain extent, by their association with massive stars too bright for *Gaia*, and are hence unknown to SigMA. We compared our results with similar recent work and find that the SigMA algorithm recovers richer populations, is able to distinguish clusters with velocity differences down to about 0.5 km s^{-1} , and reaches cluster volume densities as low as $0.01 \text{ sources pc}^{-3}$. The 3D distribution of these 37 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 more actively star-forming and dynamically complex than previously thought. We confirm that the star-forming molecular clouds in the Sco-Cen region, namely, Ophiuchus, L134/L183, Pipe Nebula, Corona Australis, Lupus, and Chamaeleon, 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 allows an accurate census of the young populations to be constructed, which in turn allows us to quantify their dynamics and recreate the recent star formation history of the local Milky Way.

Key words. methods: data analysis – open clusters and associations: individual: Sco-Cen – solar neighborhood – ISM: clouds

1. Introduction

The ESA *Gaia* mission (Gaia Collaboration 2016, 2018, 2021, 2023a) is transforming our knowledge of the local Milky Way, particularly in regards to the distribution 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, 2019; Castro-Ginard et al. 2018; Cantat-Gaudin et al. 2018a; Galli et al. 2018; Damiani et al. 2019; Meingast et al. 2019, 2021; Kounkel & Covey 2019; Chen et al. 2020; Hunt & Reffert 2021; Olivares et al. 2021). This wide range reflects the rather complex feature space¹ from where the stellar populations are extracted. Firstly, these initially compact objects are stretched into elongated, sometimes non-convex structures

in position space as a consequence of interactions with the Milky Way potential, spiral arms, and giant molecular clouds (e.g., Kamdar et al. 2021). This “galactic stretching” leads to a variety of cluster² shapes, from compact (when young) to low-contrast, spread-out, sometimes S-shaped clusters dominated by Milky Way tidal forces (e.g., Meingast & Alves 2019; Röser et al. 2019; Meingast et al. 2019, 2021; Beccari et al. 2020; Kounkel & Covey 2019; Jerabkova et al. 2019, 2021; Ratzenböck et al. 2020; Kerr et al. 2021; Kamdar et al. 2021). Secondly, due to the low number of available radial velocities, about 2% in the *Gaia* Data Release 3 (DR3) database (Gaia Collaboration 2023a; Katz et al. 2023), one is, for the most part, restricted to two tangential velocity axes plus the spatial three-coordinate axes 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

* Full Table E.1 is only available at the CDS via anonymous ftp to cdsarc.cds.unistra.fr (130.79.128.5) or via <https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/677/A59>

** Interactive Figs. 10, 12, and 13 are available at <https://www.aanda.org>

¹ Stars in the data set are represented as points in a 5D or 6D space with three positional axes and two or three kinematic axes. In a machine learning context, this space is referred to as feature space. The term “feature” is synonymous with dimension or coordinate axis.

² 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 spatially and kinematically coherent structures we find in Sco-Cen. None of the Sco-Cen clusters are expected to be gravitationally bound.

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 background noise that is not easily removable in the 5D space. Thus, the feature space consists of stellar clusters of various shapes and densities embedded in a sea of noise.

To tackle the challenge of identifying subpopulations in a star-forming region, we have developed a method that analyzes the topological structure of the 5D density field spanned by 3D positions and 2D tangential velocities. We applied a fast modality test procedure that introduces a measure of significance to peaks in 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 and apply it to the Scorpius-Centaurus (Sco-Cen) OB association (Kapteyn 1914; Blaauw 1946, 1952, 1964a,b) to identify the different subpopulations and compare results to recent papers with similar goals. Sco-Cen is the closest and best studied OB stellar association (e.g., de Geus et al. 1989; de Geus 1992; de Bruijne 1999; Preibisch & Zinnecker 1999; de Zeeuw et al. 1999; Lépine & Sartori 2003; Preibisch & Mamajek 2008; Makarov 2007a,b; Diehl et al. 2010; Pöppel et al. 2010; Rizzuto et al. 2011; Pecaú et al. 2012; Pecaú & Mamajek 2016; Forbes et al. 2021) and has an age of $\lesssim 20$ Myr (Pecaú et al. 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 interstellar medium and planet formation. Since the advent of large-scale astrometric data from the ESA *Gaia* mission, which started to become available in 2016 (Gaia Collaboration 2016), there has been a renewed interest in this benchmark region, with a focus 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; Kerr et al. 2021; Luhman 2022; Schmitt et al. 2022; Miret-Roig et al. 2022a; Briceño-Morales & Chanamé 2023).

In this paper we present the method SigMA in Sect. 3, using *Gaia* DR3 data (Sect. 2), and validate it in Sect. 4. In Sect. 5 we present an application of SigMA to Sco-Cen, including comparisons to previous work (Sect. 5.2). In Sect. 6 we summarize our findings.

2. Data

In this work we apply the newly developed method presented in this paper, SigMA, to *Gaia* DR3 data at and around the Sco-Cen OB association. To this end, we select a box of about $1.5 \times 10^7 \text{ pc}^3$ from the *Gaia* DR3 Archive (Gaia Collaboration 2023a), 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 (1964a). 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

$$\begin{aligned} & -50 \text{ pc} < X < 250 \text{ pc} \\ & -200 \text{ pc} < Y < 50 \text{ pc} \\ & -95 \text{ pc} < Z < 100 \text{ pc}. \end{aligned} \quad (1)$$

The 3D space positions (XYZ) are derived from the *Gaia* DR3 positions right ascension (α , deg) and declination (δ , deg), and the parallax (ϖ , mas). The distance (d , pc) is estimated from the inverse of the parallax, which is a fairly good approximation of the distance for sources within 200 pc and with low uncertainties (see also Appendix A). The box contains in total 5 587 760 *Gaia* sources when additionally requiring $\varpi > 0$ mas. To reduce the influence of spurious measurements, we applied the following quality criteria to the *Gaia* DR3 data within the selected box:

$$\begin{aligned} & \text{fidelity_v2} > 0.5 \\ & \varpi/\sigma_\varpi \equiv \text{parallax signal-to-noise} (S/N_\varpi) \\ & S/N_\varpi > 4.5. \end{aligned} \quad (2)$$

The parameter fidelity_v2 is a classifier to identify spurious sources in the *Gaia* DR3 and EDR3 (Gaia Collaboration 2021) catalogs, developed by Rybizki et al. (2022), which can be used to select high fidelity astrometry. The `parallax_over_error` cut (similar to an S/N threshold) reduces additional uncertainties in distance. This leaves 980 348 sources inside the box out of 5 587 760 ($\sim 18\%$) to which we applied the SigMA clustering algorithm, which we describe in Sect. 3. In Appendix A we give more details on data retrieval and the choice of the quality criteria. In this paper, the methodology is validated using data with the mentioned quality criteria applied; therefore, if using different criteria, the completeness and contamination estimates (see Sects. 3.5.4, 4.2, and Appendix D) could change as well. Using sources not fulfilling these quality criteria would require updated validation alongside new completeness and contamination estimates.

The clustering is primarily done in the 5D phase space, using the 3D spatial coordinates XYZ in parsecs, and the 2D tangential velocities v_α and v_δ in km s^{-1} , as derived from the observed proper motions ($\mu_\alpha^* = \mu_\alpha \cos(\delta)$, μ_δ) and parallaxes (see Appendix A). The Sun's reflex motion strongly influences the tangential velocities v_α , v_δ . If this is not accounted for, the distribution of Sco-Cen members in tangential velocity space is strongly correlated and depends on a cluster's position and apparent size (see Fig. C.1). Such nonlinear relationships contradict a central underlying assumption of many clustering algorithms. These clustering methods assume a universally valid metric, which implies a global correlation behavior (e.g., the commonly used Euclidean norm assumes no correlation between input features). To avoid formulating a locally adaptive metric, we transform tangential velocities to velocities relative to the local standard of rest (LSR), written as $v_{\alpha,\text{LSR}}$ and $v_{\delta,\text{LSR}}$ in km s^{-1} . This transformation reduces the influence of the reflex motion of the Sun.

We used the barycentric standard solar motion relative to the LSR from Schönrich et al. (2010), while there are different values in the literature, which would give slightly different resulting motions (e.g., Kerr & Lynden-Bell 1986; Francis & Anderson 2009, see also Appendix C in Großschedl et al. 2021). However, the differences are only marginal and irrelevant for our purposes since the main goal is to reduce the strong positional correlation, which is applied consistently to all stars when deciding on one standard Solar motion correction. The effect of the transformation on tangential velocities is highlighted in Appendix C. The LSR conversion of the proper motions is accomplished with Astropy, as outlined in Appendix A. Finally, the different dimensions are scaled to each other, as described in the methods in Sect. 3.3.3.

The final clustering result is obtained from the 5.5D space since we include *Gaia* radial velocities (Katz et al. 2023), when

available, to remove possible field star contamination, as outlined in Sect. 3.5. In this cleaning step, we assign approximate radial velocities (v_r) to all stars in a minimization procedure involving the hypothetical cluster bulk motion. While this procedure also works without access to v_r measurements (see details in Sects. 3.5.2 and 3.5.3), we find in simulations that $\geq 5\%$ of v_r measurements are necessary to achieve no loss in accuracy (see Appendix B.4). Hence, we refer to the used dimensions as the 5.5D space, since v_r are added if available in *Gaia* DR3. *Gaia* DR3 only includes v_r for about 2% of the sources with parallaxes (or about 20% in the selected Sco-Cen box if considering sources with $\sigma_{v_r} < 2 \text{ km s}^{-1}$). Adding auxiliary v_r data from other surveys would improve the statistics but lead to a very inhomogeneous data sample with 6D phase space information. Therefore, we restrict our clustering procedure to the so-called 5.5D 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 5.5D phase space, we can create a method that does not strongly rely on radial velocity information, which can be used more widely on larger data samples.

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 SigMA clustering algorithm in Sect. 3.2, as developed in this work³. An in-depth description of related work underlying SigMA can be found in Appendix B.1.

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. Nowadays, the biggest effort is no longer data collection, but the large sample sizes and high dimensionality significantly impact all parts of the analysis pipeline – storage, processing, modeling, and interpretation. “Big data” usually contain extensive information, diversity, and complexity; thus, we require more complex methods to model its observations. However, many traditional analysis techniques have time and memory complexities that fail to perform 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 deal with large complex data, new interpretive methods need to be tailored to the particular scientific question, in our case, identifying co-moving and coeval clusters of stars inside the 1+ billion stars in the *Gaia* archive. Clustering analysis, or unsupervised machine learning, has recently become essential to identifying coeval stellar structures. Clustering aims to obtain an organization of data points into meaningful clusters. However, due to the lack of labeled data, partitioning into “meaningful” clusters is generally an ill-posed problem (Cornuéjols et al. 2018). The algorithm’s choice and parameters must match the problem at hand. Clustering methods can generally be split into space partitioning algorithms (e.g., K -Means, MacQueen 1967), hierarchical algorithms (e.g., single linkage clustering, Johnson 1967), density-based techniques (e.g., DBSCAN; Ester et al. 1996), and model-based or parametric clustering algorithms

(e.g., expectation maximization; Dempster et al. 1977). An introduction to cluster analysis and classical methods is available, for example, in Jain et al. (1999).

From this list of clustering methods, parametric clustering algorithms and density-based methods are commonly used on astronomical data sets (Kuhn & Feigelson 2019; Hunt & Reffert 2021, 2023). In the following, we give an overview of model-based clustering algorithms in Sect. 3.1.1 with an extension to Bayesian formulation in Sect. 3.1.2. In Sect. 3.1.3 we present density-based clustering methods, which build the foundation for SigMA, discussed in Sect. 3.2.

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, 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. 2018a; Cantat-Gaudin et al. 2019b; Kuhn & Feigelson 2019).

A considerable downside limiting parametric clustering algorithms’ versatility is their dependence on the unknown number k of mixture components. Depending on data characteristics such as dimensionality, the number of samples per cluster, and cluster separation, determining k is a difficult problem. Addressing this problem is paramount, as the resulting model is very sensitive to the choice of k (Celeux et al. 2019).

Although model selection methods such as the Akaike information criterion (AIC; Akaike 1974) and the Bayesian information criterion (BIC; Schwarz 1978) aim to provide a principled approach to selecting k , they make the somewhat restricting assumption that the data are sampled from a model within the collection to be tested. This assumption can result in overestimating the number of k in practical situations (Celeux et al. 2019). Further, BIC and AIC 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. Many of the above-presented problems can be mitigated using a Bayesian analysis approach.

3.1.2. Bayesian clustering approach

The deviance information criterion (DIC) for missing data models (Celeux et al. 2006) is commonly mentioned as a Bayesian model selection criterion. However, in a Bayesian setting, the unknown number k of mixture components can naturally be treated as a random variable estimated jointly with the component-specific parameters.

Notably, two Bayesian formulations of selecting k exist, finite and infinite mixtures models. Finite mixture models often rely on the reversible jump Markov chain Monte Carlo (RJMCMC) technique (Richardson & Green 1997), which can navigate between finite mixture densities with variable k . Similarly, sparse finite Gaussian mixtures (Malsiner-Walli et al. 2016) involve specifying sparse priors on the mixture parameters and can be performed using classical Markov chain Monte Carlo (MCMC) methods. In contrast, nonparametric Bayesian

³ The source code is publicly available via GitHub under: <https://github.com/ratzenboe/SigMA>.

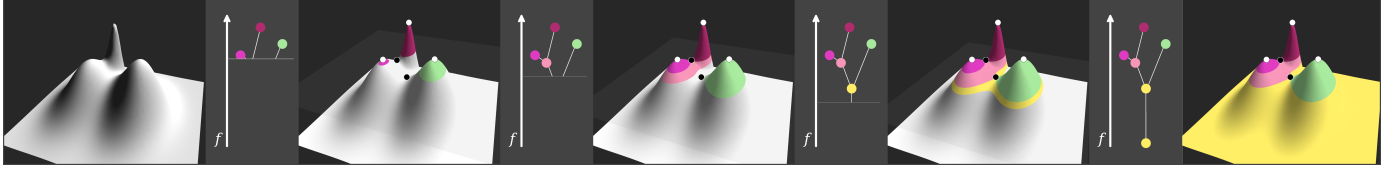


Fig. 1. Level-set method generating a hierarchical merge tree. Via a continuous change of λ from ∞ to $-\infty$, a new component is created at each maximum (white points). At each saddle point (black points), components are merged. The merge tree is fully computed when λ reaches the global minimum.

approaches are based on mixture models with a countably infinite number of components. In this case, the prior over the mixing distribution typically takes the form of a Dirichlet process (Müller & Mitra 2013).

Although Bayesian analysis methods provide well-established methods for identifying the number of clusters k , fitting models to data requires a statistical model that can generate the data set reasonably well (Hogg et al. 2010). The observed morphological structure of co-moving stellar systems in position space is significantly more intricate than simple multivariate Gaussians. In recent years many new cluster shapes such as tidal tails (Meingast & Alves 2019; Röser et al. 2019; Jerabkova et al. 2021), streams (Meingast et al. 2019), strings (Kounkel & Covey 2019), rings (Cantat-Gaudin et al. 2019a), snakes (Wang & Ge 2021), and pearls (Coronado et al. 2022) have been identified. Additionally, the projection of 3D space velocity onto the celestial sphere provides another complexity requiring a flexible clustering scheme.

Nonparametric models are frequently employed when the process that generates data is intricate, and the distribution form is unclear or hard to define. Nonparametric models, unlike parametric models, do not impose strict assumptions about the shape or features of the underlying distribution. Instead, they aim to learn the underlying pattern straight from the data. This allows us to make predictions for exceedingly complex distributions without having to know or presume the shape of the distribution.

3.1.3. Nonparametric, density-based clustering

The premise of nonparametric density-based methods states that the observed data points⁴ $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ are drawn from an unknown density function f . The goal of nonparametric 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 cluster 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} . To increase the robustness of the result, Mean-Shift, for example, 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 nonintuitive 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 sets⁵ of f . Given a data set, X , drawn from an unknown density function, f , that has compact support, X , we can formally write the resulting level sets for the threshold λ as

$$L(\lambda) := \{\mathbf{x} \in X : f(\mathbf{x}) \geq \lambda\}. \quad (3)$$

Thus, $L(\lambda)$ constitutes a set of connected components that we identify as clusters. Varying the parameter λ from ∞ to $-\infty$ gives a hierarchical data summary, called the merge tree. Figure 1 highlights the generation of such a merge tree, which builds the basis for hierarchical density-based clustering. For more details see Appendix B.1.

In the level-set framework, popular clustering algorithms such as DBSCAN can be simply thought of as a single level that is obtained by fixing λ . DBSCAN avoids estimating the data density explicitly, by employing a radius parameter, usually called ϵ , 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 ϵ -sized shells around them.

The connected components of the level set $L(c)$ are the resulting clusters while the remaining data are treated as noise. However, the choice of the parameter λ , which is related to DBSCAN’s ϵ parameter, is ambiguous, a task that 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 ϵ (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 λ cannot reveal all peaks in f . A hierarchy of clustering solutions can be obtained by considering all possible threshold values at once (see Fig. 1).

3.2. SigMA: Significance mode analysis

This section describes our clustering pipeline, SigMA, which builds on several established methods from data mining and statistics. We discuss those methods in further detail in Appendix B.1.

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) that 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

⁴ In the following, bold, lower-case variables denote d -dimensional vectors.

⁵ Often also referred to as super-level sets.

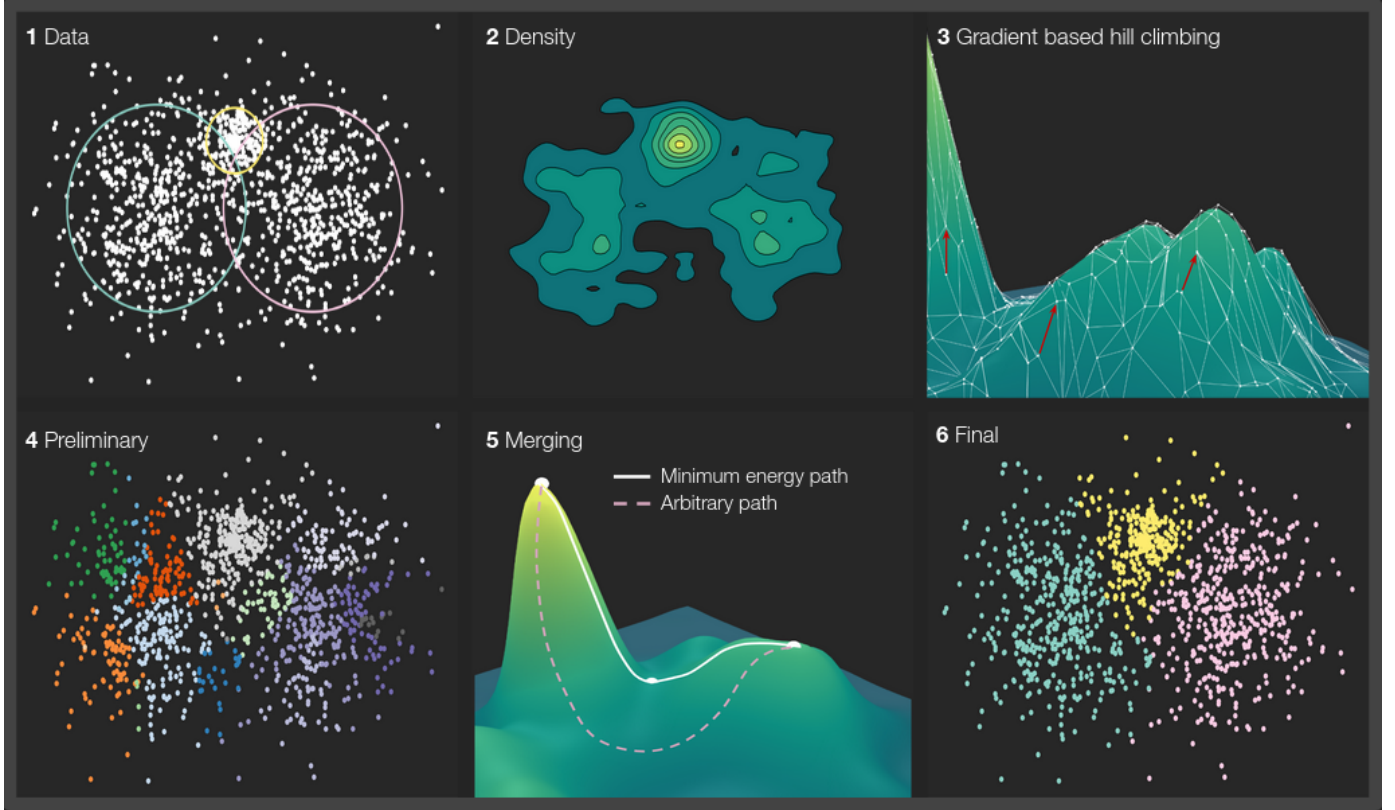


Fig. 2. Proposed clustering process SigMA, highlighted on a 2D toy data set of three Gaussians with variable covariance matrices and means. (1) The generated toy data set consisting of three bivariate Gaussians shown in white alongside 2σ confidence ellipses in color. (2) The clustering procedure starts 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 toward local peaks. (4) This gradient propagation results in a preliminary segmentation of input samples that typically is far too fine-grained. (5) These segmented regions are iteratively merged with a parent mode if a modality test along the MEP detects no significant density dip. (6) The final segmentation retains all three clusters.

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.

3.2.1. A fast modality test procedure

We considered 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 $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ drawn from f and any point \mathbf{r} on a path connecting two modes $\mathbf{c}_i, \mathbf{c}_j$ in f , Burman & Polonik (2009) show that

$$\widehat{SB}(\mathbf{r}) = d \sqrt{k/2} \left[\log d_k(\mathbf{r}) - \max(\log d_k(\mathbf{c}_i), \log d_k(\mathbf{c}_j)) \right] \quad (4)$$

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

$$\widehat{SB}(\mathbf{r}) \geq \Phi^{-1}(1 - \alpha), \quad (5)$$

where Φ is the standard normal cumulative distribution function (CDF). For a more thorough derivation of Eqs. (4) and (5) see Appendix B.2.

Since Eq. (4) processes a single point rather than a complete path, the modality test in Eq. (5) describes a point-wise 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. (5). 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.

Instead of computing the test statistic $\widehat{SB}(\mathbf{r})$ for multiple values of \mathbf{r} , we propose limiting the calculation to only a single realization. Importantly, reducing the number of point-wise evaluations of point-wise tests does not interfere with the distributional assumption of the test statistic itself. Burman & Polonik (2009) show that the test statistic \widehat{SB} along the entire path p with $p = [\mathbf{r}_1, \dots, \mathbf{r}_N]$ follows an N -dimensional multivariate normal distribution with zero mean and identity covariance matrix under the null hypothesis. Thus, the null hypothesis is independent of the number N of point-wise tests performed.

Modifying the modality test procedure to a single evaluation of the test statistics reduces the overall statistical power of the test. Because we under-sample the path between two modes, we decrease the chance of sampling in places where a significant drop in density occurs. Therefore, the probability of type II errors increases, that is, the null hypothesis is not rejected

even though it is false. To maintain statistical power while also extending the test procedure to non-convex cluster shapes, 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 X with the 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.

To effectively reduce the number of point-wise tests without losing all statistical power we need to evaluate Eq. (4) in areas close to the maximum density dip while ignoring other areas irrelevant to the rejection decision. This maximum density dip at point s maximizes the test statistic and, thus, dominates the test procedure. Due to the test statistics' proportionality to the distance $d_k(s)$, 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 point-wise test at the saddle point s connecting the two peaks. As the saddle point governs the modality test, we can assign a p -value that takes the following form:

$$p = 1 - \Phi\left(d\sqrt{k/2}\left[\log d_k(s) - \max(\log d_k(c_i), \log d_k(c_j))\right]\right). \quad (6)$$

At the end of Appendix B.2, we empirically show that these assumptions hold and introduce a small correction factor to the variance of the standard normally distributed test statistic under H_0 that is valid for *Gaia* phase space data.

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 shared border. Using this monotonous property assumption, we aim to provide a fast 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 becomes 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 are separated into m disjoint modal sets $M = \{M_1, \dots, M_m\}$.

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 computed the cluster tree of M . As described in Sect. 3.1.3, the cluster tree is obtained by varying the density threshold λ from $\infty \rightarrow -\infty$ and registering modal regions when λ passes through a peak in \hat{f} and their unification when λ passes through the respective saddle point. To finalize

the cluster tree, we need to identify the saddle points between modal regions of M .

We determine the saddle point between two modes via an edge search in G . Specifically, we consider edges that 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. This edge density approximates the corresponding saddle point density between two adjacent modal regions.

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. (6) is evaluated, and neighboring modal regions are merged if the respective p -value exceeds the significance level α . Therefore, the significance level α provides an immediate and meaningful way to simplify the initial cluster tree.

3.3. Parameter selection

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

3.3.1. Graph

The choice of the graph directly affects the gradient approximation. For instance, 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 dramatically 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 (ERGs) to strike a balance between over and under-connecting points in the data set 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 β -skeleton (Kirkpatrick & Radke 1985) is a one-parameter generalization of an ERG where β 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, respectively. Correa & Lindstrom (2011) find that critical point searches (necessary for topological decomposition, clustering, and gradient estimation) are more accurate with β -skeletons, with $\beta < 1$ compared to k -nearest

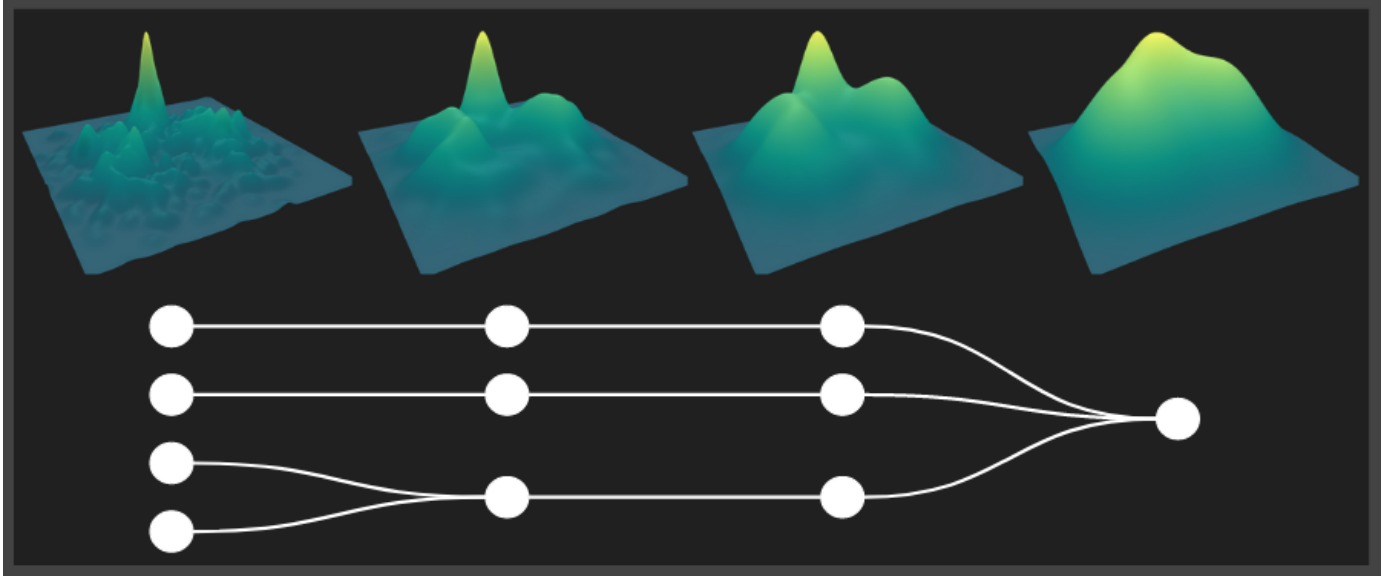


Fig. 3. Schematic figure linking the cluster number to the density estimation process. Applying a smoothing operator generates a family of density fields. This hierarchical family of functions is called a scale space.

neighbor (k -NN) graphs and the Gabriel graph. Since the number of vertices grows very fast as β gets smaller, we chose a value of $\beta = 0.99$.

Adopting a β -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 pruned the initially computed graph in a post-processing step. We removed vertices that show a significant density dip as one moves from one vertex to another. For simplicity, we assumed that the saddle point lies at the arithmetic mean of the two vertex points.

3.3.2. Density estimation

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 employed a model-agnostic, nonparametric estimator for the underlying density.

The most popular nonparametric density descriptors are kernel density estimation (KDE) and k -NN. The KDE technique estimates the density f 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. 3. The k -NN method takes a more naive approach to estimating 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 noncontinuous 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. The nearest

neighbor method is also able to provide a more accurate estimate of high-density regions compared to the kernel method (Burman & Nolan 1992).

In contrast to KDE, the computational cost of nearest neighbor methods is highly efficient due to the use of kd-tree⁶ 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 for determining 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 employed a k -NN estimator to approximate the density function considering these advantages. Specifically, we used a density estimator based on the distance to an empirical measure described by Biau et al. (2011). It is a weighted k -NN estimate, which incorporates distances d_1, \dots, d_k to all nearest neighbors up to k . The distance to an empirical measure 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 as follows,

$$d_m(\mathbf{x}) = \sqrt{\frac{1}{k} \sum_{\mathbf{y}_i \in N_k(\mathbf{x})} \|\mathbf{y}_i - \mathbf{x}\|^2}, \quad (7)$$

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

$$\hat{f}_m(\mathbf{x}) = \frac{1}{nV_d} \left(\frac{\sum_{j=1}^k j^{2/d}}{kd_m^2(\mathbf{x})} \right)^{d/2}, \quad (8)$$

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

⁶ Short for k -dimensional tree.

density values is important, we can ignore constant normalization terms in Eq. (8).

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 tangential velocity dispersion to be 0.5 km s^{-1} 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 s^{-1} and galactic coordinates in pc, both subspaces 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 subspaces.

Scaling factors, c_i , put weight on specific subspaces to increase 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 applied the same scaling c_v to both tangential velocity axes while leaving the positional axes unchanged with $c_x = 1$. SigMA is applied to the following set of dimensions, \mathcal{D} :

$$\mathcal{D} = \{X, Y, Z, c_v \times v_{\alpha, \text{LSR}}, c_v \times v_{\delta, \text{LSR}}\}. \quad (9)$$

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 inversely proportional to parallax, we aim to extract a relationship between a stellar cluster'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 used data on stellar clusters within 500 pc. Cantat-Gaudin & Anders (2020) have compiled a list of open clusters in the Milky Way disk. However, using a single cluster census can introduce a bias in the resulting scaling factor as only a single member selection function was used to obtain the sample. Thus, we substitute and add clusters covered by Gagné et al. (2018b), who have used a multivariate Bayesian model to identify members of young associations within 150 pc⁷.

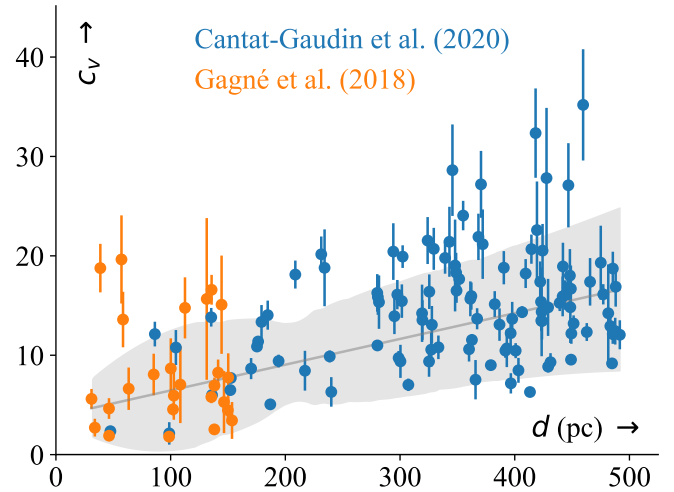


Fig. 4. Empirical distance-scaling relationship using data from Gagné et al. (2018b) and Cantat-Gaudin & Anders (2020). The x -axis represents the distance to stellar clusters; the y -axis shows the dispersion ratio of positional over kinematic subspaces. This dispersion directly corresponds to the velocity scaling factor, c_v , as discussed in Sect. 3.3.3. The gray line represents the best fit linear regression line, while the gray band indicates the 1σ highest density interval obtained from sampling the posterior predictive distribution.

The scaling fraction should account for the distance differences between positional and kinematic subspaces. To quantify this idea, we consider the distance distribution of sources to the cluster's center in each subspace. 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 subspaces as the x - v dispersion ratio. To estimate the uncertainties in the x - v dispersion ratio, we perform bootstrap sampling with 100 resamples for each cluster and compute the mean and standard deviation of computed dispersion ratios across the ensemble.

The dispersion of cluster members in a given feature provides a measure of the scale of stellar populations in that dimension. Since the x - v dispersion ratio is not one (see Fig. 4), we have to prevent an unequal emphasis of one subspace against the other. To compensate for the bias toward positional axes during clustering, the velocity features must be scaled by a factor c_v equal to the observed x - v dispersion ratio.

Figure 4 shows the relation between a cluster's distance and its x - v dispersion ratio (alongside determined uncertainties), which equivalently is our choice of c_v . We identify an approximately linear trend and fit a Bayesian linear model to the data (see Appendix B.3 for a detailed description of the model choice and fitting procedure).

Using this empirical mean model, we find mean suitable scaling factors, c_v , between approximately 5–10, assuming the clusters of Sco-Cen are at a distance of about 100–200 pc. Without the LSR correction, these values translate to a range of 4–7, which are comparable to the correction factors by Kerr et al. (2021, using v_α , v_δ) who used the values 5 and 6 in their clustering approach (see Appendix C for further discussion on these two velocity spaces). At first glance, the model suggests

⁷ We cross-matched the *Gaia* DR1 sources identified by Gagné et al. (2018b) and DR2 sources from Cantat-Gaudin & Anders (2020) with DR3 for more precise astrometry. We opted for the Gagné et al. (2018b) census if a cluster appears in both surveys. Compared to density-based

methods, the mixture of Gaussian densities deals naturally with scaling factors (provided the Gaussian assumption holds). The scaling factors can be compensated (to some extent) in the covariance matrix of the individual Gaussian components.

sampling values in the range of 5–10 or using the mean of 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 clusters we aim to find.

As discussed in Appendix B.3, possible scaling factors can be expressed by the conditional probability integrated over a range of distance values. Given the linear model and associated model uncertainties obtained from the posterior predictive distribution, we find a resulting distribution of scaling factors within distances of 100–200 pc. Keeping the number as small as possible is essential since we must perform a separate clustering run for each sample we draw from the distribution. 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. B.4⁸.

We run the clustering pipeline for each scaling fraction sample, creating an ensemble of 10 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-Shulcoper 2011).

A consensus function aims to produce a result that shares as much information as possible with individual clustering results among the ensemble. In particular, we are interested in robust cluster solutions that exist through multiple velocity scales while ignoring unstable solutions where clusters randomly break apart or merge into others. We outline our consensus clustering approach in Appendix B.5 where Fig. B.6 shows a schematic of the proposed pipeline.

3.4. Measurement uncertainties and multiple hypothesis testing

Rigorous integration of measurement 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 used a Monte Carlo approach that attempts to approximate the sensitivity of the modality structure under statistical uncertainty. We did this by resampling the data N times using a Gaussian distribution centered on each point with an appropriate covariance matrix obtained from *Gaia* data. The resampling procedure creates N “merge” or “do not merge” decisions at each saddle point location. In the following Sect. 3.4.1, we discuss methods for combining the N p -values into a single merge decision. In Sect. 3.4.2 we aim to reduce the chance of falsely rejecting the null hypothesis and thus making incorrect do not merge decisions, the likelihood of which increases as the number of hypothesis tests (i.e., saddle points) grows.

3.4.1. Single merge decision: Combining multiple p -values

Recomputing 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 locally 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 an individual null hypothesis. Instead of focusing on single tests, we need to combine these individual tests and simultaneously test the global null hypothesis that no p -value is significant. As a result, a global hypothesis test can “borrow” information from the other test statistics to gain significance.

A popular method of computing the global p -value is Fisher’s method (Fisher 1934). Fisher’s method assumes that individual p -values are uniformly distributed in the interval $[0, 1]$. Consequently, the negative logarithm follows an exponential distribution: $-\log p_i \sim \text{Exp}(1)$. The test statistic, t , then becomes the sum of the negative log-sum of n p -values, which follows a χ^2 distribution with $2n$ degrees of freedom:

$$t = -2 \sum_{i=1}^n \log p_i \sim \chi^2(2n). \quad (10)$$

Fisher’s method is especially attractive for densely packed cluster agglomerates such as Sco-Cen. If clusters have only marginally different velocities and positions, our point-wise test might produce a p -value slightly larger than the rejection threshold. In such cases finding the precise saddle point position is challenging, leading to a type II error. Although no single test (or very few) can reject the null hypothesis, many small effects in Eq. (10) enable us to reject the global null hypothesis. However, Fisher’s method assumes statistical independence between individual tests. Since the resampled data sets are not independent, this assumption is somewhat violated (for a more detailed discussion, see Appendix B.6).

Instead, we employed the Cauchy combination test (CCT; Liu & Xie 2020), which is similar to Fisher’s method in the sense that it is also able to combine multiple individual p -values that aggregate multiple small effects. Compared to Fisher’s method, the authors show that CCT is still powerful under arbitrary dependence structures among p -values. The test statistic t – which is asymptotically standard Cauchy distributed – is defined in the following:

$$t = \sum_{i=1}^n w_i \tan [(0.5 - p_i)\pi]. \quad (11)$$

The weights w_i must sum to one and can reflect the power of respective hypothesis tests. Since all tests are performed equally, we distribute the weights evenly by choosing $w_i = 1/n$.

In practice, we computed distances to the k th neighbor of each initial modal candidate and corresponding saddle points across resampled data sets. The number of resampled data sets limits the proposed procedure, as data generation is costly. Thus, we restricted the number of samples to $n = 50$, rejecting the global null hypothesis at a 5% significance level, hence $t < 0.05$.

3.4.2. Multiple merge decisions: Reducing false discoveries

The number of random spurious modes and associated saddle points is proportional to the data set size. For large data sets, such as the *Gaia* data, we obtain hundreds of modal candidates, which we prune in a second pass (see Sect. 3.2.2). As the number of modal candidates grows, the chance of falsely rejecting the null hypothesis (a false do not merge decision) increases.

⁸ We point out that the distance notation in the appendix changes from d to r to minimize confusion in the derivation of the final probability density function.

To conceptualize the probability of making one or more type I errors when performing multiple hypothesis tests across a data set, the concepts of family-wise error rate (FWER) and false discovery rate (FDR) have been introduced. FWER aims to control the probability of at least one false positive in the sample, while FDR aims to control the expected proportion of false positives among all positives. Tests from both families involve modifying the significance threshold of statistical tests. FWER correction typically involves more conservative adjustments, such as the Bonferroni correction (Bonferroni 1936), that increase the threshold for statistical significance, while FDR correction typically involves less stringent adjustments. As the number of tests grows, FDR correction provides greater statistical power at the cost of an increased number of type I errors (Shaffer 1995).

Since we aim to identify a small number of significant results among a large number of statistical tests instead of aiming to minimize the total risk of any false positives, tests from the FDR family are preferred over controlling the FWER. The Benjamini-Hochberg (BH; Benjamini & Hochberg 1995) method is a widely used statistical method for controlling the FDR in multiple hypothesis testing. The BH procedure guarantees that the FDR is controlled at the desired significance level α , assuming that the null hypotheses are independent or positively correlated. Since individual tests are performed on spatially disconnected saddle-point and mode pairs, we did not assume any correlation structure among individual tests on a given data set. Thus, once we obtain revised p -values for each saddle-point – mode pair through resampling (see Sect. 3.4.1), we applied the BH procedure to adjust the p -values globally and ensure FDR control.

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 nonoverlapping 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 clusters and field content. This section aims to remove the field star component from the modal region to obtain a final clustering result.

In the following, we discuss the noise removal pipeline schematically highlighted in Fig. 5. The noise removal scheme is based on a density-based member selection technique, which we motivate in Sect. 3.5.1. The pipeline is roughly split into two main parts. In the first part, we aim to assign a radial velocity to each source to transform the data into 6D Galactic Cartesian coordinates (XYZUVW). We determine the cluster's 3D space motion to estimate missing radial velocity information, discussed in Sect. 3.5.2. In the second part, we describe the automated cluster member selection using so-called cluster-noise classifiers (see Sect. 3.5.3). Finally, we discuss contamination and completeness estimates of our member selection procedure in Sect. 3.5.4.

3.5.1. Density-based member selection

Identifying signal and background sources can be formulated using a mixture model approach in which cluster and field star populations are modeled directly in phase space. However, due to complex cluster shapes found in the literature (see Sect. 3.1.2), we cannot create a generative model of the data set at hand. Instead, we return to the density-based formalism of clustering, where we

treat clusters as an enhancement of density over the background. To select cluster members as over-densities in phase space, we reduce the 5D phase space information to the univariate density information. A single density threshold in this univariate space corresponds to an isosurface in the original phase space.

We aim to describe the univariate density distribution as a mixture model to automatically obtain a suitable isosurface threshold to separate the signal from the background. This model should be able to capture the point-wise density distribution of field stars and cluster members. Before fitting a mixture model to data, we must define the number of mixture components and distributions we used.

A mixture of two components in a first approximation seems plausible as the algorithm divides the input space into regions containing a signal and background component. By design, each region consists of a single-density peak in phase space. This distributional condition forces the field star component to lie locally around the cluster while exhibiting no extra peaks. To concur with uni-modality, the background distribution is restricted to uniformity or exactly one density peak that coincides with the signal mode. As the former is more likely, we assume the field component to be approximately uniform in phase space around a cluster. Uniform distributions in N -dimensional feature spaces translate to a single Gaussian in the univariate density space.

The distribution of cluster star densities is harder to model. Cluster members are commonly modeled as multivariate Gaussians (e.g., Gagné et al. 2014; Sarro et al. 2014; Crundall et al. 2019; Riedel et al. 2017). As discussed in Appendix B.7, given a k -NN density estimator, a multivariate Gaussian in phase space approaches a Gaussian distribution in univariate density space as the dimensionality grows. However, observational findings point to more complex morphologies (e.g., Meingast & Alves 2019; Röser et al. 2019; Meingast et al. 2019; Kounkel & Covey 2019; Cantat-Gaudin et al. 2019a; Jerabkova et al. 2021; Wang & Ge 2021; Coronado et al. 2022) and significant mass is contained in the low-density region outside the cluster core (Meingast et al. 2021). As discussed in Sect. 3.1.2, we lack critical information to formulate a generative model for the signal distribution in phase space and, consequently, in univariate density space. Instead of explicitly building a univariate signal model, we employed multiple Gaussian mixture components to describe the point-wise k -NN density distribution. Thus, we did not restrict the number of Gaussian components during the fitting procedure to provide flexibility to capture more complex distributions.

To decide on a proper density threshold ρ_0 , we determine the number of mixture components by minimizing the BIC. The background is automatically identified as the Gaussian component with a low relative mean (i.e., lower point-wise densities), small variance (uniform background component has less relative variance around its mean density than the signal), and large weight (the number of field stars exceeds cluster members by about 100:1). This procedure can be seen in Fig. 6 where we show an example of two Gaussians fitted to the univariate density data (denoted by ρ) of one modal region.

3.5.2. Bulk velocity estimation

The Gaussianity assumption of density components is appropriate only in the original Cartesian coordinate system. Densities computed from tangential velocities suffer from perspective effects, leading to deviations from normality due to the nonlinearity of projections onto the celestial sphere. We find such distortions empirically when analyzing distributions of

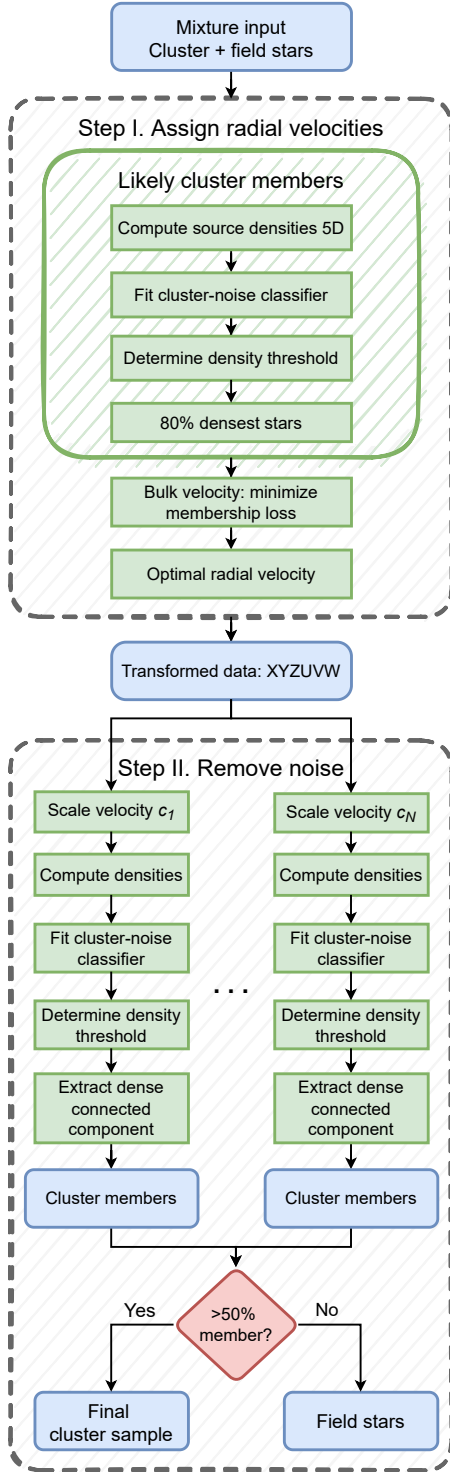


Fig. 5. Noise removal pipeline separating cluster members from field stars. The pipeline is split into two main parts. In the first part, we aim to assign a radial velocity to each source to transform the data into Galactic Cartesian coordinates. In the second step, we used the transformed data to fit several cluster-noise classifiers to separate the signal from the background. Blue represents data products at various steps, green denotes processing steps, and red shows decisions. For more details, see Sect. 3.5.

various modal regions in projected 2D (see the tangential velocity space in Fig. C.1) compared to Cartesian 3D velocities. This effect is reduced by correcting the observed tangential velocities for the Sun’s motion, yielding motions relative to the LSR

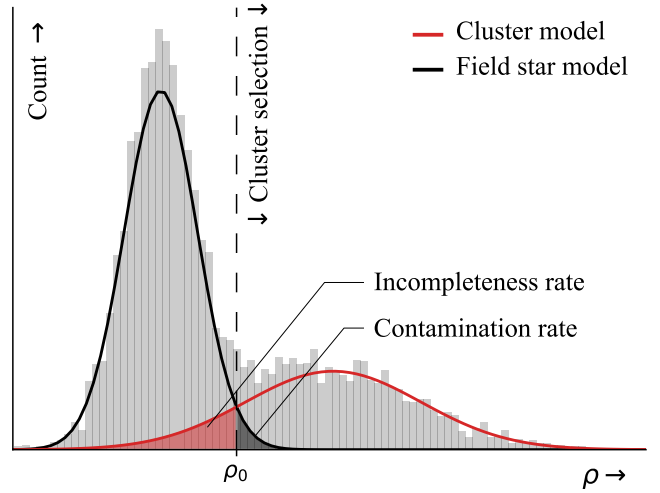


Fig. 6. Noise reduction schematic. We fit the observed uni-variate density distribution, ρ , with a mixture of two Gaussians that model the cluster (red line) and field star (black line) population, respectively. We obtained an approximation to the field star contamination and incompleteness rate in the cluster sample by considering the cluster-noise classifier’s confusion matrix entries, particularly false positives, false negatives, and true positives.

(see Appendix C). However, very nearby clusters, which cover large areas in the sky, are still affected by the observer’s point of view from Earth. We move to the Galactic Cartesian coordinate system to eliminate these observational effects. Thus, we transform the data into the 6D space (XYZUVW) 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, the majority of radial velocity measurements of sources are missing ($\sim 62\%$ in our box, 80% if sources with $\sigma_{v_r} > 2 \text{ km s}^{-1}$ are removed). 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, some clusters lack enough statistics to compute their bulk motion. Before determining individual radial velocities of member stars, we have to estimate the space motion of individual populations. In the following, we describe bulk motion estimation, which provides a mean to estimate an optimal radial velocity. We summarize the process in the first part of Fig. 5.

We determined the space motion, $\tilde{\mathbf{v}}$, of individual populations of size n by minimizing the following loss function, henceforth called membership loss:

$$L(\tilde{\mathbf{v}}) = \sum_{i=1}^n \left(\frac{\Delta v_{\alpha,i}^2}{\sigma_{v_{\alpha,i}}^2} + \frac{\Delta v_{\delta,i}^2}{\sigma_{v_{\delta,i}}^2} + \frac{\Delta v_{r,i}^2}{\sigma_{v_{r,i}}^2} \right) \quad (12)$$

$$\Delta v_{x,i} = v_{x,i}^{\text{obs.}} - \tilde{v}_{x,i}. \quad (13)$$

The minimization is done over the tangential (v_{α} , v_{δ}) and radial (v_r) velocities⁹.

⁹ Compared to the clustering analysis, which assumes a universally valid metric implying a global correlation behavior (see Sect. 2), the optimization procedure is not affected by nonlinear relationships between input features. Thus, to avoid propagating errors through the LSR conversion, we stick to observed tangential and radial velocities.

The delta terms in the membership loss describe the offset between observed and computed values at the specified velocity $\tilde{\mathbf{v}}$. Although we introduce an additional observational error via the parallax uncertainty, we chose the tangential velocities to match the unit of radial velocities, the essential component in the sum in Eq. (12). Each term in the sum is weighted by its respective uncertainty, which decreases the influence of observations with large measurement errors on the membership loss. If all observations lack radial velocities, the last term is set to zero; if only a subset of v_r 's is missing, their values are imputed with the average of its complement.

For a perfectly co-moving population, the membership loss has a global minimum with a value of 0 at the cluster motion. Observational uncertainties, contamination from field stars, and a nonzero velocity dispersion will increase the minimum value accordingly. To search the 3D bulk motion that minimizes the membership loss, we used the quasi-Newton Broyden, Fletcher, Goldfarb, and Shanno method (Nocedal & Wright 1999) with an initial guess of the mean 3D velocity¹⁰. We denote the velocity, which minimizes the membership loss (Eq. (12)), as the optimal bulk motion (\mathbf{v}_{OBM}).

To determine the cluster motion of the co-moving population via our minimization approach, finding \mathbf{v}_{OBM} needs a large and pure selection of cluster sources, meaning truly co-moving stars. We attempt to obtain a relatively clean sample of cluster stars via the aforementioned mixture model approach (see Fig. 6, and “Likely cluster members” in Fig. 5). By fitting a mixture of univariate Gaussians to the density distribution of a modal region, we get a classifier that separates the cluster from field stars¹¹. Since the input density is 1D, the classifier, also called a cluster-noise classifier, becomes a simple threshold classifier.

Sources with a density greater than the threshold ρ_0 are likely 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 \mathbf{v}_{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. Figure 6 shows an example of the contamination estimation.

The optimal bulk motion \mathbf{v}_{OBM} is used to infer an “ideal” radial velocity. The ideal radial velocity minimizes the Euclidean distance between \mathbf{v}_{OBM} and the velocity vector constrained by the measured proper motions. We refer to the computed 3D space motion, which is a combination of measured proper motions and the inferred radial velocity, as the minimally different velocity (\mathbf{v}_{MDV}). On the one hand, genuine cluster members should receive an estimated space velocity close to their true motion (assuming low intra-cluster velocity dispersion of a few km s^{-1}). Field stars (if not an interloper in phase space) show incompatible observations with the co-moving population and are, on average, assigned a different space velocity. Together with sparseness in positional space, field stars consequently show lower densities in phase space.

We infer \mathbf{v}_{MDV} for sources without v_r measurements as well as for sources with large uncertainties of $\sigma_{v_r} > 2 \text{ km s}^{-1}$. After this step, all sources have an associated radial velocity, either measured or inferred. We provide these inferred radial velocities

(\hat{v}_r) in our final catalog¹². Thus, compared to the clustering step of the SigMA pipeline removing the field background explicitly requires radial velocity estimates for all sources (regardless of uncertainty). In Appendix B.4 we estimate the necessary fraction of sources with radial velocity measurements without losing accuracy in \hat{v}_r . Using two simulated data sets, we find that SigMA requires at least 5% of radial velocity measurements in the input catalog to keep errors to a minimum. Due to the need for some v_r measurements, we refer to the SigMA pipeline as operating in 5.5D.

The bulk velocity estimate correlates directly with observed proper motions and radial velocities. Therefore, systematic errors, as in the case of binary or multiple stellar systems, introduce corrupted measurements that potentially bias the final result. However, directly flagging binary stars and removing them from the inference process does not significantly alter the results as only a tiny fraction (0.05%) of *Gaia* sources are identified as multiples (Gaia Collaboration 2023b). Since the work by Gaia Collaboration (2023b) does not represent the entire binary population, we investigated another indicator for multiples in the *Gaia* catalog. For example, the renormalized unit weight error parameter (Lindgren et al. 2018, 2021) can be used as a discriminator, which measures how well the astrometric solution is fitted to a single star model, as also discussed in Penoyre et al. (2022a,b). These authors also show that binaries, which have been observed with longer time baselines (e.g., comparing HIPPARCOS, DR2, and DR3), could still deliver parallaxes and proper motions that are close to the true values (see also Kervella et al. 2022). As a consequence, binaries can still be selected as true-positive members of a cluster if selected with 5D *Gaia* astrometry, as can be seen, for example, by the clear binary sequences in Hertzsprung-Russell diagrams (HRDs; e.g., Meingast et al. 2021). However, even in these cases, the multiple systems do not comprise a significant fraction of the cluster selection. Consequently, we assume that binaries contribute only marginally to the bulk velocity computation.

We applied the presented method to clusters in the Sco-Cen OB association (see Sect. 5) to validate the bulk velocity and 3D velocity estimation procedure. During inference, we randomly remove 95% of radial velocities to facilitate a comparison with observed values. The absolute Δv_r and relative errors δv_r to *Gaia* measurements with $\sigma_{v_r} < 2 \text{ km s}^{-1}$ are shown in Fig. 7. The absolute error is defined as the difference between the estimated radial velocity, \hat{v}_r , and the observed value, v_r :

$$\Delta v_r = \hat{v}_r - v_r. \quad (14)$$

The relative error expresses the magnitude of the absolute error compared with its measured magnitude:

$$\delta v_r = \left| \frac{\hat{v}_r - v_r}{v_r} \right|. \quad (15)$$

We find that 68% of sources (1σ) have absolute errors of less than $\pm 2.35 \text{ km s}^{-1}$ and 95% (2σ) of absolute errors are within $\pm 5.66 \text{ km s}^{-1}$. Thus, the average error is close to the large statistical uncertainties (2 km s^{-1}) in the sample, constituting an approximation for the lower bound for the mean estimation error. The majority (1σ) of relative errors are below 0.55. Thus, inferred radial velocities are in good agreement with observations, validating our method and highlighting its robustness to (not yet fully understood) binary effects and contamination.

The following steps use the space velocity information to determine cluster membership. We pre-filter unlikely members

¹⁰ If no radial velocities are available, our initial guess is the null vector. We empirically find that the optimization converges to the same results for different initial velocities.

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

¹² Given as \mathbf{v}_{ERV} (estimated radial velocity) in Table E.1.

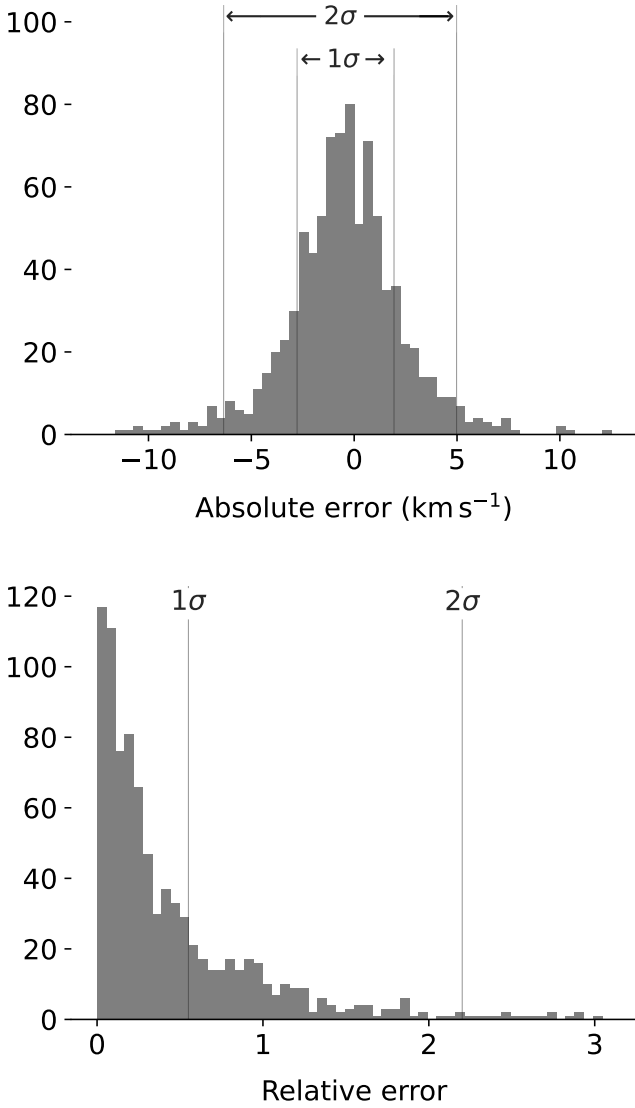


Fig. 7. Absolute and relative error of inferred radial velocities compared to observed values in *Gaia* DR3 with radial velocity errors below 2 km s^{-1} . We randomly removed 95% of available radial velocities during inference to facilitate this comparison. Only inferred values where the *Gaia* observable has been removed are shown. We highlight the 1σ and 2σ percentiles and find that the majority (68%) of absolute errors are within $\pm 2.35 \text{ km s}^{-1}$ and have relative errors below 0.55.

via a kinematic selection before applying the cluster-noise classifier (see Sect. 3.5.3) to a complete 6D phase space, including the computed \hat{v}_r estimates. The pre-filter removes possible contaminant stars that have vastly different 3D motion, namely sources that differ by more than 10 km s^{-1} from the determined bulk motion; hence, $\|\mathbf{v}_{\text{MDV}} - \mathbf{v}_{\text{OBM}}\| < 10 \text{ km s}^{-1}$.

3.5.3. Removing field star contamination

Figure 5 shows the noise removal pipeline, which consists of two main parts. In the previous section, we discussed the first part in which we impute missing radial velocities by assuming the presence of a single dense co-moving population in the input data. In the following, we discuss the second part, in which we aim to fit cluster-noise classifiers to remove the field star content in the combined space of heliocentric Galactic Cartesian position and determined \mathbf{v}_{MDV} 's.

Combining positional and kinematic spaces (here XYZUVW) 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 the density estimation. Instead of selecting a single scaling factor, we chose multiple plausible scaling factors and compute a univariate density distribution ρ for each one. We obtain scaling factors c_1, \dots, c_N by repeating the procedure discussed in Sect. 3.3.3 in Galactic Cartesian phase space (XYZUVW), setting N to 10 (see forking into N subprocesses in Step II in Fig. 5).

We separate the stellar population from the field star component using a cluster-noise classifier (the classifier is motivated and discussed in detail in Sect. 3.5.2). This classifier is applied to the 1D density estimation ρ determined in 6D phase space, using measured and estimated radial velocities (see the x -axis in Fig. 6). This thresholding method results in a global isosurface selection that is independent of positional information of sources in the original feature space (see Fig. 1). To reduce the contamination of random field star components, we employed the β -skeleton as a locality-aware neighborhood graph from which we delete vertices that fall below the computed density threshold ρ_0 , as shown in Fig. 6. More details on this graph-based approach are provided in the related work discussion in Appendix B.1.

Field stars account for the majority of sources in the given samples. Thus, the number of vertices that fall below the density threshold makes up most of the graph. Removing them disconnects the graph and splits it into multiple connected components. We define sources within the densest (and typically the largest) connected component as cluster members. To extract cluster members more robustly, we computed one extraction for a range of scaling parameters (see Sect. 3.3.3). We obtain a final cluster catalog by removing unlikely members that appear in less than half of the N extractions when using different scaling factors, c_i (see Fig. 5).

3.5.4. Contamination and completeness estimate

The cluster-noise classifier is a discriminative model, whose conditional densities (or mixture components) describe the cluster and field star distributions. In combination with the decision threshold ρ_0 , we can internally compute estimates for the field star contamination fraction f_{cont} and the incompleteness fraction f_{inc} for each cluster sample.

In the fitting procedure, the number of mixture components k is a free parameter determined by minimizing the BIC. Thus, the discriminative model can have a different number k of Gaussian components for each cluster. To formalize a consistent definition across a different number of components, we introduce the following notations. We denote the set variables specifying the identity of the mixture component by Z ; its k components are then $Z = \{z_1, \dots, z_k\}$. Individual Gaussian components can then be formulated as densities conditioned on the mixture component:

$$p(x | z_i) = \mathcal{N}(\mu_i, \sigma_i). \quad (16)$$

The mixture density can then be written as

$$p(x) = \sum_{z \in Z} p(z) p(x | z), \quad (17)$$

where $p(z)$ is the prior probability of the mixture component z , also called mixture weight. The k mixture weights must add up to one.

The subset of components that describe the distribution of cluster members is denoted with S . It encompasses all mixture components whose mean (or expected value) exceeds the threshold ρ_0 , and thus $S = \{z \in Z : \mathbb{E}[p(x | z)] > \rho_0\}$. The relative complement of S with respect to Z then contains all components describing the field star distribution; we denote this set as B , defined by $B = Z \setminus S$.

Finally, with this formulation in mind, we can express both the incompleteness fraction f_{inc} and contamination fraction f_{cont} . The incompleteness fraction, as shown in Fig. 6, is the probability of observing a sample from the cluster distribution with a value less than ρ_0 :

$$f_{\text{inc}} = \frac{\sum_{s \in S} \int_{-\infty}^{\rho_0} p(s)p(x | s) dx}{\sum_{s \in S} p(s)}. \quad (18)$$

The contamination fraction is defined as the fraction of false positive samples in the overall cluster sample. Thus, f_{cont} can be expressed as the probability of observing a sample from the field star distribution among all samples with a value larger than the threshold ρ_0 :

$$f_{\text{cont}} = \frac{\sum_{b \in B} \int_{\rho_0}^{\infty} p(b)p(x | b) dx}{\sum_{z \in Z} \int_{\rho_0}^{\infty} p(z)p(x | z) dx}. \quad (19)$$

Both f_{inc} and f_{cont} are schematically shown in Fig. 6 for the example of a two-component mixture.

The contamination and incompleteness are computed for each cluster. We obtain a mean contamination estimate across all clusters in Sco-Cen of 5.3% with a standard deviation of 3.1% across clusters. This value agrees well with photometric contamination estimates via the *Gaia* HRD, as described in Sect. 5 and Appendix D.1. Although we find good agreement, we do not completely trust the stated values due to a lack of knowledge of systematic uncertainties.

The major source of systematic uncertainty is a deviation from Gaussianity of any of the mixture components. Especially, in the case of more than two mixture components, we expect that these internal estimations have increased error rates. By departing from the paradigm of “one mixture component per signal and background” we increase the accuracy of the model (and ideally the obtained cluster members) at the cost of direct model interpretability (and all of its consequences). Further uncertainty is added via the density estimation to which the mixture model is fit. Since we do not have access to the true underlying density f , we inevitably make mistakes by substituting it with our estimate \hat{f} .

We find a mean completeness across clusters of approximately 89.2% with a standard deviation of 8.3%. 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 of local positional and kinematic relationships that might help increase the completeness of our catalog. Further, we estimate the actual value even lower, as we find multiple connected components in the neighborhood graph from 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. Additionally, we evaluate

the estimated completeness fraction by comparing our sample to past extractions in the literature in Sect. 5.2. These comparisons suggest a sample completeness of about 90% (e.g., when compared to Damiani et al. 2019, Luhman 2022, or Schmitt et al. 2022), which agrees with our estimate. However, these surveys can also not be considered complete. On the contrary, a direct comparison shows (see Sect. 5.2) that other applications on Sco-Cen are missing sources that SigMA is able to uncover.

Instead of comparing estimated values to past extractions, we aim to evaluate the accuracy of internal contamination and completeness estimates of SigMA in Sect. 4.2. Using simulations, we find that on average the contamination from field stars can be approximated quite well using our univariate mixture model approach. However, especially in dense cluster environments, our approach seems to overestimate completeness. If a large portion of the cluster exists in the low-density environment outside the cluster core the density approach fails to adequately capture the true number of missing cluster members (see Sect. 4.2.2 for a detailed discussion).

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 Fig. 3 illustrates the dependence of the cluster number on the density estimation process. Applying a smoothing operator generates a family of density fields, called a scale space (Witkin 1987). We used this scale-space concept to study the dependence of extracted clusters on density estimation. Clusters with a long lifetime 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 used 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 cluster¹³ contains the cluster mode of its child. If both conditions for direct and merge links 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 inexact 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 applied 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. 3.

¹³ The parent cluster resides in the $i+1$ st level, whereas the child cluster is from level i .

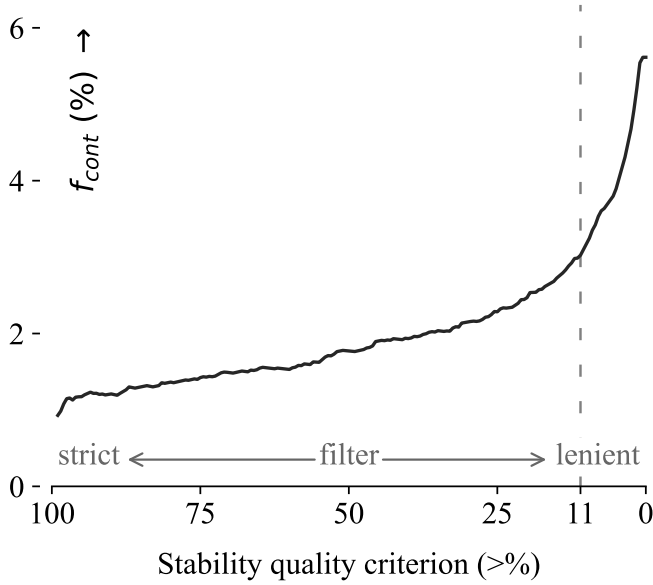


Fig. 8. Stability versus estimated contamination rate. The contamination estimate was determined via source positions in the HRD relative to the 25 Myr isochrone as discussed in Sect. 5, selecting potential contaminants from older populations. We identify a sharp drop in contamination for low stability values that levels off at around 11%.

We extract stable components from the resulting merge tree via a consensus clustering approach, discussed in Sect. 3.3.3. In total, we find 60 stable clusters in the search box, while 37 are discussed in more detail in Sect. 5 as being part of the Sco-Cen association. We find the 23 remaining clusters to be unrelated to the young Sco-Cen association (see Sect. 5). Often these appear as incomplete (or truncated) cluster extractions; in particular, the shape and position of the majority of these clusters suggest that they extend beyond the defined search box.

The consensus approach also lets us characterize how often individual sources appear throughout the cluster ensemble¹⁴. We report this value as “stability” in our cluster catalog. The stability criterion can be used as an effective measure to remove spurious sources from the catalog. Figure 8 highlights the effect of the stability criterion on sources when empirically estimating the contamination from older sources (hence likely unrelated sources) via an HRD. The x -axis shows a given stability criterion where we filter sources with stability $> x$. The y -axis shows the empirical contamination estimate, determined via positions of filtered (older) sources in the HRD. The fraction of sources to the left of the 25 Myr isochrone is used to estimate the false positive rate while sources to the right account for true positives (see further details in Sect. 5 and Appendix D). Based on this result, we recommend a quality criterion of stability $> 11\%$.

However, due to the density-based nature of SigMA, the stability criterion is strongly correlated with density. Especially clusters with minor density enhancement over the field background are short-lived in scale space. Thus, although SigMA detects them clearly, some clusters contain members with overall disproportionately small stability values causing them to fall out of the sample for relatively low stability values (e.g., the cluster Oph-North-Far; see Sect. 5.1.7). Therefore, we do not recommend generally using the stability quality criterion, but

to investigate its behavior per cluster, to get potentially cleaner cluster samples.

3.7. Removing spurious cluster solutions

Unstable cluster solutions are automatically filtered out in our scale space approach in Sect. 3.6. However, the distribution of sources in the Milky Way in phase space is far from uniform. SigMA’s job is to separate the input data space into unimodal regions. This segmentation does not distinguish between compact, clustered over-densities and long-range, low-density modes inherent to the Milky Way distribution. We aim to remove the latter from our cluster sample.

We assume a natural clustering of “real” and “spurious” clusters in 1D density space. That means if the density of all members across the N extracted clusters are plotted we expect a bi-modal distribution. The modes of these distributions then correspond to the real and spurious clusters.

To classify each cluster into any one category, we employed an iterative approach that starts off by assigning all clusters to the category real. Subsequently, we loop through all clusters sorted by their median member density (computed in 5D phase space) in ascending order. In each iteration i , with $i \in \{1, \dots, N\}$, the first i clusters (i clusters with lowest median density) are classified as spurious. At each step, we track the separation and compactness in the 1D density space of individual members across both groups of clusters. The more compact and well separated the members of both groupings are – measured by the Caliński-Harabasz score (Caliński & Harabasz 1974) – the more the classification at step i agrees with our bi-modal assumption.

We used the classification at step i , which maximizes the Caliński-Harabasz score to characterize each cluster as real or spurious. This classification is directly applied to each SigMA clustering solution, hence before obtaining a consensus result across scaling factors and scale space (see last process in SigMA core in Fig. 9).

3.8. The SigMA pipeline

The proposed clustering method SigMA has many components that require sensible choices in order to work together properly. In Sects. 3.2–3.7 we justify and discuss the parameter choices that yield the final analysis pipeline. Figure 9 shows an overview of the full pipeline. It consists of two main parts, the SigMA core, and two consensus clustering steps. In the following, we briefly summarize how these individual components come together.

The SigMA core outputs a clustering result for a given density level i and velocity scale c_j . It iterates through the following steps: First, a k -NN density estimation is computed (see Sect. 3.3.2). Second, a gradient-based hill-climbing step is performed that produces individual modes and saddle point locations. The modal candidates (or clusters) at this point over-segment the data set (see Sect. 3.2.2). Third, at each saddle point, a modality test is applied that determines whether two neighboring modal candidates should be merged or not (see Sect. 3.2.1). Fourth, the field star background is removed from each modal region (see Sect. 3.5). Fifth, spurious clusters are removed from the extraction (see Sect. 3.7).

To guarantee stable results against velocity scaling factors (see Sect. 3.3.3) and density estimation (see Sect. 3.6), we employed a consensus clustering approach, discussed in Sect. 3.3.3. From the cluster ensemble, we can extract a stability value for each source in our final catalog.

¹⁴ The cluster ensemble is the collection of N clustering solutions corresponding to the N density estimates.

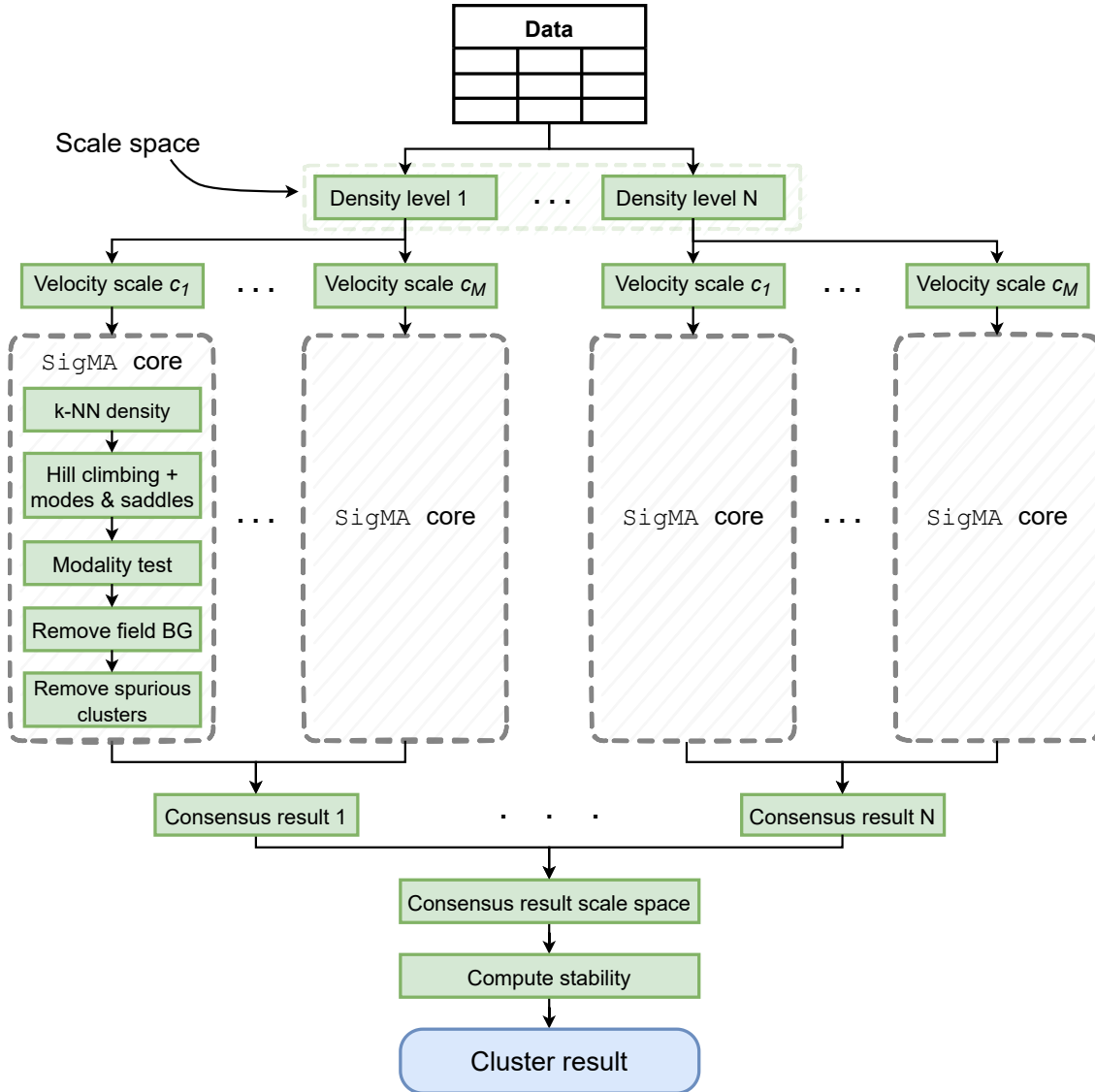


Fig. 9. Schematic illustration of the SigMA pipeline. The pipeline consists of two main parts, the SigMA core and two consensus clustering steps. For a detailed explanation, see the main text of Sect. 3.8 and the references therein.

4. Validation

We followed a two-pronged approach to verify the proposed clustering method SigMA. First, in Sect. 4.1 we validate our clustering technique qualitatively in a case study on the Sco-Cen OB association. We describe the results and comparisons to other studies in Sect. 5. Second, in Sect. 4.2, the algorithm is validated quantitatively on simulated data, where we compare SigMA to other established clustering methods used to identify co-moving clusters.

4.1. Validation using astrophysical knowledge

Two direct observables that can be identified in our application on Sco-Cen (Sect. 5) serve as a validation test of the method. First, and apart from the youngest clusters that are affected by dust extinction, the *Gaia* color-absolute-magnitude diagrams (CMDs; equivalent to observational HRDs) for the stars in each cluster show a narrow (coeval) distribution (see the follow-up work in Ratzenböck et al. 2023). 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 a test is the prominence of massive stars associated in 2D projection with the SigMA identified clusters, while they are often located at a central position within the concerned clusters (e.g., α Sco, β Sco, δ Sco, and ν Sco; see Sect. 5 and Tables 3–5). Nearby massive stars are often too bright to have reliable measurements in the *Gaia* archive, and they are also often part of multiple stellar systems, further reducing the quality of the astrometry. The brightest are not even in *Gaia* (like Antares, Ohnaka et al. 2013), while most have been observed by HIPPARCOS (Perryman et al. 1997). Still, the method finds clusters around many of the massive stars, particularly in the Upper-Sco region. Based on HIPPARCOS astrometry (see Sect. 5), 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.

Table 1. Overview of the clustering algorithms that we test with synthetic data (simulated cluster samples) and compare to SigMA.

Algorithm	Main reference	Short description	For example, used by ^(b)
DBSCAN	Ester et al. (1996)	A density-based Clustering algorithm	Castro-Ginard et al. (2019, 2020, 2022) , Zari et al. (2019) , Fürnkranz et al. (2019) , Hunt & Reffert (2021)
HDBSCAN ^(a)	Campello et al. (2013)	A hierarchical density-based Clustering algorithm	Kounkel & Covey (2019) , Kounkel et al. (2020) , Hunt & Reffert (2021, 2023) , Kerr et al. (2021)

Notes. The results are listed in Table 2. ^(a)See Appendix B.1 for further discussion on the HDBSCAN algorithm. ^(b)The reference lists are not exhaustive literature reviews but are intended to highlight the relevance of the proposed comparison methods.

4.2. Validation using simulations

To objectively investigate the effectiveness of new clustering algorithms, synthetic data with known ground truth information facilitates the comparison to other clustering techniques. Since SigMA is tuned to astrophysical phase space data, we want to test its efficacy on simulations that approximate observational data as closely as possible. Therefore, simulated data should replicate the data model, content, volume, uncertainties, and selection effects of *Gaia* data as closely as possible.

To our knowledge, the *Gaia* Early Data Release 3 (EDR3) mock catalog by [Rybizki et al. \(2020\)](#) best meets these criteria. In particular, the catalog contains a large, realistic open cluster sample with internal rotation and corresponding uncertainties. Although the cluster structure of the open cluster sample differs from Sco-Cen, access to ground truth data and realistic *Gaia* selection effects and content provides a firm validation basis. To test SigMA’s ability to separate clusters in dense cluster environments, commonly found in young OB associations and star-forming regions such as Sco-Cen (Sect. 5) or Orion (e.g., [Chen et al. 2020](#)), we aim to create a derived data set from the original mock catalog that mimics these conditions.

We applied the same error cuts to mock and real data to increase comparability between our qualitative and quantitative tests, as described in Eq. (2). We emphasize that the following results are, thus, conditioned on the given quality criteria. Hence, the reported classifier performances and empirically determined contamination and completeness scores should always be understood in the context of these quality criteria. Therefore, future use of SigMA should aim to meet the same criteria or, in case of modifications, follow the validation protocol discussed in Sects. 4.2.1 and 4.2.2 for the new filter setup.

In the following sections we describe the comparison of results in detail, particularly the data used, the algorithms against which we compare SigMA, and the validation results.

4.2.1. Open cluster sample

The *Gaia* EDR3 mock catalog ([Rybizki et al. 2020](#)) is extensive. As it aims to reproduce *Gaia* data realistically, it contains simulated measurements on over 1.5 billion sources and over 1000 open clusters compiled from the catalogs of [Kharchenko et al. \(2013\)](#) and [Cantat-Gaudin et al. \(2018b\)](#). We need to reduce the data size to a manageable subset to validate SigMA and compare different clustering algorithms. Therefore, we limit the data to a range of 200 pc around the Sun, which yields uncertainty characteristics similar to our Sco-Cen box sample. Applying the quality criteria Eq. (2) results in the final test catalog size of 2 682 883 samples, of which 18 682 are part of 12 open clusters¹⁵. The

mock catalog does not contain the full 5D astrometric uncertainty covariance matrix that SigMA uses. We substituted missing values with real measurements from *Gaia* DR3. Each mock sample was randomly paired with a source within our Sco-Cen box (see Sect. 2), whose values it adopts.

This simulation allows our proposed analysis method to be tested for accuracy. Importantly, we must highlight its capacity to identify clusters in positional and kinematic data in contrast to established analysis methods. We limit our comparison to the two relevant clustering methods (DBSCAN, HDBSCAN) as listed in Table 1, which are considered among the most promising candidates for stellar cluster analysis in a meta-study by [Hunt & Reffert \(2021\)](#).

Since SigMA’s parameters are tuned to deal specifically with *Gaia* data, we employed a grid search to find suitable parametrizations for each of the three clustering algorithms. This strategy measures the peak performance these clustering methods can achieve. A comparison against the best performance results allows for a discussion of methodological advantages and disadvantages rather than reflecting poor parameter selection. A detailed discussion on the parameter search is provided in Appendix B.8.

We report the performance of the best model across our search to facilitate a fair comparison. The performance itself is measured using the following clustering validation metrics: the normalized mutual information (NMI) score ([Strehl & Ghosh 2002](#)), adjusted mutual information (AMI; [Vinh et al. 2010](#)), and adjusted rand index (ARI; [Hubert & Arabie 1985](#)). We also report on classification metrics that are easier to interpret, such as true positive rate or recall, precision, accuracy (henceforth denoted as ACC), balanced accuracy (BACC; [Brodersen et al. 2010](#)), and the Matthews correlation coefficient (MCC; [Matthews 1975](#)). In addition, we report the total number of identified clusters, N_{tot} , as well as the number of non-noise clusters, N_{cluster} , that are found to coincide with a toy cluster instead of field members. Similar to precision and recall, we also report contamination and completeness. In contrast to precision and recall, we computed the average cluster contamination and completeness only for clusters that coincide with a true cluster (i.e., for the N_{cluster} identified non-noise clusters). Thus, these measures are not influenced by large false positives. When $N_{\text{cluster}} = N_{\text{tot}}$ the completeness is exactly equal to recall and contamination becomes $1 - \text{precision}$. The resulting numbers are summarized in Table 2.

Only a fraction of sources, less than 1%, are located in clusters. Hence, many of the above-proposed validation metrics will report high values as long as most field stars are clustered in the same group. We remove correctly identified field stars before computing the validation metrics to prevent reporting on artificially inflated clustering scores. By removing this “true negative” component without removing field stars labeled as cluster members (false positives) and cluster members identified as field

¹⁵ As no fidelity information is provided in the EDR3 mock catalog, this quality flag was not reproduced on mock data.

Table 2. Test results on simulated cluster samples.

	Open cluster sample ($N = 12$) ^(a)			Compact cluster sample ($N = 37$) ^(b)		
	SigMA	DBSCAN	HDBSCAN	SigMA	DBSCAN	HDBSCAN
NMI	0.95	0.94	0.86	0.60 ± 0.02	0.39 ± 0.06	0.33 ± 0.16
AMI	0.95	0.94	0.86	0.56 ± 0.01	0.38 ± 0.06	0.32 ± 0.16
ARI	0.96	0.95	0.74	0.38 ± 0.03	0.11 ± 0.03	0.08 ± 0.06
Precision	0.97	0.98	0.74	0.64 ± 0.05	0.23 ± 0.05	0.22 ± 0.15
Recall	0.98	0.96	0.76	0.47 ± 0.04	0.25 ± 0.03	0.20 ± 0.06
Contamination	0.03	0.02	0.08	0.24 ± 0.13	0.13 ± 0.12	0.25 ± 0.22
Completeness	0.98	0.96	0.92	0.76 ± 0.15	0.79 ± 0.24	0.76 ± 0.21
ACC	0.97	0.96	0.76	0.47 ± 0.04	0.25 ± 0.03	0.20 ± 0.06
BACC	0.90	0.88	0.88	0.51 ± 0.04	0.16 ± 0.03	0.18 ± 0.07
MC	0.96	0.95	0.76	0.55 ± 0.04	0.23 ± 0.04	0.16 ± 0.11
N_{tot}	12	12	14	26.8 ± 2.0	10.0 ± 1.7	24.2 ± 4.8
N_{cluster}	12	12	12	24.1 ± 1.7	10.0 ± 1.7	12.0 ± 5.5

Notes. Bold-faced numbers indicate the best performance given a specific evaluation metric. The three clustering methods, SigMA, DBSCAN, and HDBSCAN, are applied to two data sets. ^(a)The open cluster sample contains 12 clusters. It is a subset of the *Gaia* mock EDR3 catalog (Rybizki et al. 2020). ^(b)The compact cluster samples contain 37 clusters. It mimics the cluster environment of Sco-Cen, where groups are densely packed together. The 10 compact cluster samples are generated in a random effects model, and the resulting distribution varies substantially across realizations. To estimate the performance of clustering algorithms on the compact cluster sample, we average performances across ten individual samples. We report the mean and standard deviation of performance scores.

components (false negatives), the reported scores are a conservative estimate of the algorithms' actual performances.

The results are summarized in Table 2. All algorithms can recover the 12 clusters within the data set. We find that the performance of SigMA and DBSCAN are essentially equal – with relatively high evaluation scores – while outperforming HDBSCAN. We find that HDBSCAN (within the parameters we searched) identifies fewer members while also identifying two false positives (i.e., two large extra clusters that entirely contain field stars).

The access to ground truth data also allows us to test internal measurements of contamination and completeness estimates. We find a true mean contamination rate of $2.6 \pm 0.7\%$ across the twelve identified clusters. SigMA's internal estimation is slightly lower than that at a mean contamination rate of $1.1 \pm 0.4\%$. We find an even better agreement between true and estimated completeness values. The true mean completeness rate is $98.3 \pm 0.7\%$, almost identical to SigMA's internal estimation of $98.4 \pm 0.2\%$.

Although the true contamination value is outside the 1σ confidence interval, the estimated value is still very close to the true one in absolute terms. In the open cluster sample, the internal measurements provide a surprisingly good approximation given that we have not explicitly modeled signal and background in univariate density distribution but assumed a simple mixture of Gaussians.

The high reported accuracy across all clustering methods highlights the nature of open clusters. They appear as salient over-densities in phase space, making their detection fairly easy. This situation contrasts with the complex structure that constitutes Sco-Cen. Distinguishing more densely packed clusters from each other is a nontrivial task. SigMA was created with the intention of an interpretable cluster definition, which is put to the test, especially in such environments. Therefore, we aim to create a test data set reproducing densely packed piles to put our analysis tool through its paces.

4.2.2. Tightly packed cluster environment

To our knowledge, there is no realistic (replicating data model, content, volume, uncertainties, and selection effects of *Gaia*

data) simulation of Sco-Cen-like, densely packed associations that can be used to validate SigMA in densely packed cluster environments. Moreover, there are no similar (or any) star-forming regions where a consensus has been reached on the number of true clusters along with their members. Hence, we created a derivative toy data set from the EDR3 mock catalog, which simulates groupings in tightly packed arrangements. We refer to this newly generated mock catalog as the “compact cluster sample”.

The biggest unknown in creating this sample concerns the cluster details. In particular, their number, location, extent, and respective size. However, the application to Sco-Cen has already produced a cluster sample that can be considered when generating toy data, as it provides a candidate set of these quantities.

Two conflicting objectives pose challenges for sampling with known cluster sizes. On the one hand, the cluster sample should avoid strong correlations with results on Sco-Cen (for a discussion of results, see Sect. 5). We want to avoid reproducing previous results as it would favor the SigMA clustering objective over other alternative formulations and inhibit an objective comparison across methods. On the other hand, the compact cluster sample should aim to represent reality faithfully. Since no ground truth exists either on stellar clusters or in the form of dedicated simulations that reproduce such dense cluster structures, we anchor our simulations on reality by considering our Sco-Cen extraction and other literature results. To balance overconfidence in the given results with an accurate description of reality, we perturb the obtained cluster sizes to avoid an unduly high correlation with SigMA and literature results.

In the following we describe how we use these general cluster details as a starting point to generate the compact cluster sample: (1) number of sources, (2a) mean position (in heliocentric Galactic Cartesian coordinate frame, XYZ), (2b) mean space velocity (in the heliocentric Galactic Cartesian coordinate frame, UVW), mean statistical dispersion of objects in the cluster in (3a) positional and (3b) velocity space.

To introduce small and medium-sized deviation from the selected sample, we treat the number of sources and statistical

dispersion as a normal distribution centered on the measured value with a relative variance of 25%. We employed a different strategy to sample new cluster means in position and velocity. Typically, neighboring cluster centroids in the combined positional and velocity space are way within the relative variance of 25%. Thus, updated centroid positions would commonly lie outside the original cluster boundary, drastically interfering with the initial cluster distribution. Instead, we sample centroid positions from a 50% subset of each cluster, introducing variations that guarantee to retain the overall structure.

After sampling a set of cluster quantities, we pair each of the 37 extracted Sco-Cen clusters (see Sect. 5) with an open cluster from the mock EDR3 catalog. We used 15 open clusters within 250 pc from the Sun. This selection provides access to a slightly more diverse cluster sample showing measurement uncertainties similar to the initial Sco-Cen clusters. As different cluster sizes show distinct morphological features – a small cluster can typically not be reproduced by down-sampling a large one to its size – we aim to pair clusters based on member size. A given Sco-Cen cluster, c_s , with n_s sources has the following probability of being paired with one of the N synthetic mock clusters, c_k , with n_k sources, where $k \in [1, \dots, N]$:

$$p(c_k|c_s) = \frac{(n_s - n_k)^2}{\sum_i (n_s - n_i)^2}. \quad (20)$$

Thus, on average Sco-Cen clusters are paired with similar-sized clusters from the mock catalog while maintaining a nonzero probability of being paired with more, unlike clusters. To better understand variations in the number of clusters across random instances, we maintain a fixed cluster count of 37.

For each cluster, the mock counterpart is scaled to the randomly sampled dispersion in position and velocity space (separately), randomly down-sampled to its corresponding (randomly determined) size, and positioned at the corresponding mean in 6D phase space. Subsequently, the synthetic clusters are embedded into the remaining field distribution. Finally, we project the space velocities to the tangential velocity plane, compute right ascension (α , deg), declination (δ , deg), and parallax (ϖ , mas), randomly remove about 62% of radial velocity measurements and apply the coordinate and quality criteria¹⁶ from Eq. (1)–(2) to reflect the clustering conditions of Sco-Cen, as described in Sect. 2.

To evaluate and compare SigMA's clustering performance to alternative algorithms, we generate 10 compact cluster samples and report mean performance scores across these realizations. The results are summarized in Table 2. Compared to results on the open cluster sample, SigMA shows a significantly higher score than competing algorithms, achieving only half of SigMA's performance on average. The performance of DBSCAN and HDBSCAN on the compact cluster sample is approximately similar. Compared to DBSCAN, we find that top-performing HDBSCAN runs again falsely identify clusters of field stars as clusters in the data set. On average, HDBSCAN finds as many false positives as true positives. SigMA, on the other hand, can, on average, identify about twice as many subgroups as other algorithms while keeping the relative number of false positives low.

Although we can highlight SigMA's performance in these compact cluster agglomerates compared to DBSCAN and HDBSCAN, the performance values are drastically worse than in the open cluster sample. We can partially attribute the poor

performance to the tough clustering challenge created by the randomized process. By randomly perturbing mean cluster positions, neighboring clusters easily merge, decreasing the maximally achievable performance. We also find that some clusters (on average, two to three populations) can no longer be identified as their density is indistinguishable from the field. Further, the clusters' extent in 5D is scaled to approximate Sco-Cen deviations, which reliably reproduces the cluster core. However, in some cases, a considerable part of the cluster extends far (up to over 100 pc) beyond traditional Sco-Cen boundaries. The density of these stars compared to the field and their distance to the cluster core makes them impossible to detect with the three tested algorithms. Although a very tough clustering challenge, it still provides a good test bed for algorithms applied to Sco-Cen-like cluster environments.

Besides the clustering performance, we again compare true to estimated contamination and completeness estimates. We find a true mean contamination rate of $23.7 \pm 13.1\%$ across the, on average, 24 identified clusters. SigMA's internal estimation on the compact cluster sample is significantly lower with a mean contamination rate of $6.8 \pm 3.4\%$ (although ~ 7 times higher than in the open cluster sample). As discussed above, merging nearby clusters into a single indistinguishable cluster drastically increases the contamination, a factor that SigMA cannot account for. In contrast, SigMA can only control the contamination of low-density field stars and not cross-contamination (the major contributor) from other clusters. When we ignore cross-contamination from other clusters and focus purely on contamination from field stars, the true contamination fraction becomes $8.2 \pm 4.1\%$, close to the internal value of $6.8 \pm 3.4\%$. SigMA's internal estimate on the real *Gaia* DR3 data is about $5.3 \pm 3.1\%$ (see Sect. 3.5.4), which is well within 1σ uncertainties of both estimated and true values determined from mock data.

The true mean completeness rate is $76.2 \pm 15.2\%$ while SigMA's internal estimation is $89.1 \pm 2.0\%$. Ignoring cross-contamination between clusters, the total fraction of cluster members SigMA is able to pick up is 81.4 ± 1.2 . Although the mean total completeness is only slightly above the average per cluster completeness rate and does not match internal predictions, the result is relatively stable across different resampled data sets. On the compact cluster sample SigMA (and to an even greater extent, DBSCAN and HDBSCAN) cannot find the large source fraction far outside the central cluster region. This fraction is possibly exaggerated, considering the young nature of Sco-Cen sources (because we used open-cluster analogs to build the compact cluster catalog). SigMA's internal completeness estimate on real *Gaia* DR3 data toward Sco-Cen is approximated with $89.2 \pm 8.3\%$ (Sect. 3.5.4), which is slightly over-estimating completeness when compared to the true value of the simulated data, while still comparable within the uncertainties.

The substantial agreement of SigMA's internal metrics with true values in the open cluster sample is in stark contrast to the large discrepancy between estimated and true values in the compact cluster sample¹⁷. While the contamination estimate yields satisfying results if only field star contamination is considered, the completeness estimate likely systematically underestimates the low-density cluster component of stellar clusters. Thus, internal contamination and completeness estimates should

¹⁶ See also Sect. 4.2 for a brief discussion on quality filters and future use.

¹⁷ We want to emphasize here once again that these results are conditioned on a given set of quality criteria. We strongly suggest reaffirming internal contamination and completeness estimates through simulations (as discussed in Sect. 4.2) when modifying these quality criteria in future uses of the SigMA software.

only be used as rough first approximations of the stellar content of detected clusters. To obtain a better understanding, especially of the completeness fraction, we call to consider additional membership analysis tools such as UPMASK (Krone-Martins & Moitinho 2014), BANYAN (Gagné et al. 2018a), or Uncover (Ratzenböck et al. 2020).

5. Application to Sco-Cen

We applied SigMA to *Gaia* DR3 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 subgroups 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 three stellar clusters to the Galactic northeast of Sco-Cen, which we put in the separate Northeast group (NE). 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 clusters 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 follow-up study (Ratzenböck et al. 2023) we discuss in more detail the ages of the individual SigMA clusters and the star formation history of the Sco-Cen complex.

In total SigMA extracts 60 clusters inside the defined search box. Of these, 23 clusters are older populations with ages > 20 Myr, or which are kinematically unrelated. These older clusters include for example, the well-studied IC 2602 (~ 30 Myr; e.g., Randich et al. 1995; Stauffer et al. 1997; Dobbie et al. 2010; Damiani et al. 2019; Meingast et al. 2021), or the Hyades, β Pictoris, Platais 8, Platais 9, Platais 10, IC 2391, Alessi 9, Alessi 13, Tucana-Horologium, Coma-Berenices, Volans-Carina, or NGC 2451A (e.g., Riedel et al. 2017; Gagné et al. 2018b,c; Gagné & Faherty 2018; Sim et al. 2019; Fürnkranz et al. 2019; Cantat-Gaudin & Anders 2020; Meingast et al. 2021; Kerr et al. 2021; Galli et al. 2021a; He et al. 2022). These clusters generally occupy distinct velocity spaces, different from the bulk motion of Sco-Cen. Moreover, the majority of these clusters are truncated by the borders of our defined box; hence, they are incomplete, which is of no consequence to this study. In this work, we focus solely on the young Sco-Cen complex (1) to get a more complete picture of the substructure of this important nearby association, (2) to evaluate the differences to previous studies on Sco-Cen (Sect. 5.2), and (3) to highlight the capability of SigMA to untangle distinct clusters in a dense environment containing overlapping populations in space, which is especially true for young stellar associations like Sco-Cen. The 23 older or unrelated clusters are not discussed further here, although they might be related, or not, to Sco-Cen at larger scales (e.g., “blue streams”; Bouy & Alves 2015). We will discuss these older clusters in future work.

We find that 37 stellar clusters are associated spatially and kinematically with the Sco-Cen OB association, containing in total 13 103 stellar cluster members, which will be discussed in more detail in this section. Figures 10 and 11 show the distribution of the 37 Sco-Cen SigMA clusters projected in Galactic coordinates. Figure 12 shows the distribution of the clusters in 3D space in a heliocentric Galactic Cartesian coordinate frame

(see also the interactive 3D version¹⁸ online and Figs. E.1–E.5 for a better appreciation of individual clusters). The 37 clusters seem to form the continuous body of the Sco-Cen association, beyond Blaauw’s original three subgroups’ boundaries.

Figure 13 shows the location of the SigMA clusters in the tangential velocity plane as observed from the Sun (v_α/v_δ) and also relative to the LSR ($v_{\alpha,\text{LSR}}/v_{\delta,\text{LSR}}$). Since the clusters partially occupy similar velocity spaces in the velocity planes, we also provide online an interactive 2D version of this figure, allowing a better appreciation of 2D kinematical properties of the clusters in Sco-Cen (see also Figs. E.1–E.5). The 37 young clusters all fall on a connected loop-like pattern in tangential velocity space (Fig. 13, left panel), a pattern largely created by the reflex motion of the Sun. This is highlighted in Fig. C.1, showing that these projected motions are expected for clusters at Sco-Cen positions and distances. To avoid this pattern caused by the Sun’s motion, we additionally transform the tangential velocities to velocities relative to the LSR, using the standard solar motion from Schönrich et al. (2010; see Sect. 2). This is shown in Fig. 13 (right panel), where we can see that the clusters now occupy a more compact velocity space. In particular larger clusters, which are stretched over larger areas in the sky, show a smaller velocity dispersion after the LSR conversion (see Appendix C). The 2D kinematical properties of individual clusters can be better appreciated when investigating the online interactive 2D version of the figure, where both velocity spaces can be compared directly.

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 not traditionally associated with Sco-Cen, like Pipe, CrA, Cham, and clusters toward the Galactic northeast (NE), including a cluster connected to the L134/L183 clouds. Moreover, other well-known stellar clusters, traditionally not assigned to Sco-Cen but later suggested to be associated with it, were picked up by SigMA, like ϵ Cham and η Cham (e.g., Mamajek et al. 1999, 2000; Fernández et al. 2008), which are added here to the Sco-Cen complex.

The relatively young β Pictoris stellar cluster (β Pic; e.g., Fernández et al. 2008; Crundall et al. 2019; Miret-Roig et al. 2020, age ~ 18 – 20 Myr) was also picked up by SigMA in the selection box. As mentioned above, we decided not to include this young local association as part of our final sample of 37 stellar clusters. The SigMA extraction of β Pic covers only one side of the known population as defined in Miret-Roig et al. (2020). This is likely due to the larger extent of β Pic in the sky (partially outside of our box boundaries) and 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 (or 5.5D) phase space as used by SigMA in this work as the stars are distributed across the sky as seen from Earth (the Solar system is located inside some of these nearby young associations).

The majority of the 37 clusters can be related to previously identified clusters from the literature, which are often larger scale structures containing several of the SigMA clusters (see comparisons to the literature in Sect. 5.2). The rich substructure identified by SigMA also includes clusters with no clear counterpart in previous works. We decided to name such clusters after their location in the sky (based on constellation) or after the brightest star that is seen in projection to a cluster and where we feel confident that it is part of a cluster (see Sect. 5.1 and Tables 3 and 5). We often find

¹⁸ Clusters can be viewed separately by double-clicking on the cluster name in the legend. By clicking once on another cluster it can be added to the visible clusters, and so on.

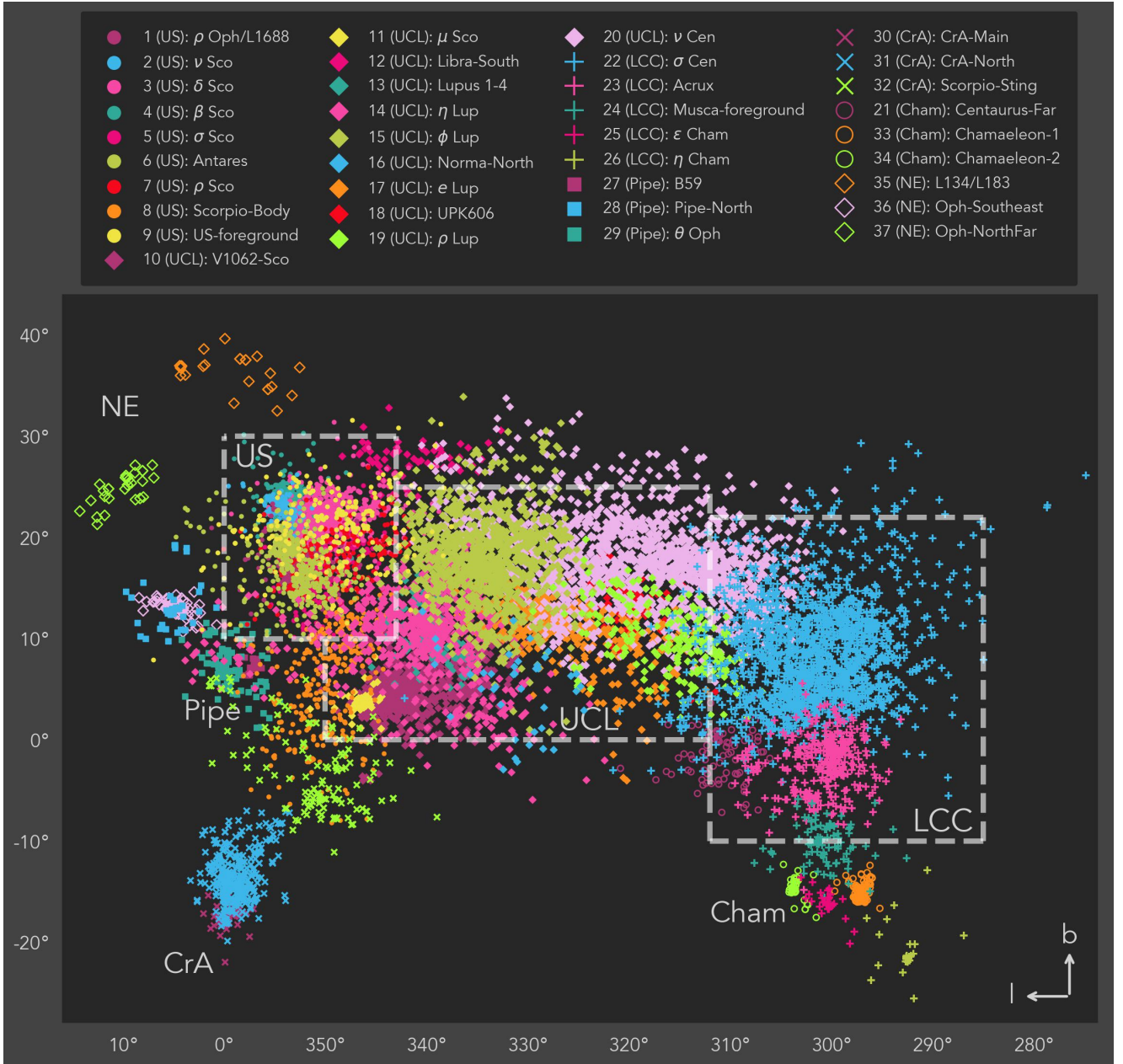


Fig. 10. Distribution of the 37 SigMA clusters in Sco-Cen, projected in Galactic coordinates. Traditionally, the Sco-Cen OB association is separated into US, UCL, and LCC, marked with gray dashed lines. The clusters extracted with SigMA reveal a more complex substructure of Sco-Cen than initially proposed by Blaauw (1946), and they show a more extended spatial distribution that includes the CrA, Pipe, and Cham regions and additional stellar clusters toward the northeast (NE). The clusters are ordered in the legend by region, as given in Table 3. See the interactive 2D version [online](#) or Fig. 11 for a separate view of each cluster. For a better visualization of the clusters' distribution, see the interactive 3D version [online](#) (Fig. 12).

bright B-stars toward cluster centers at approximately the same distance and proper motion, in itself a validation of the SigMA algorithm, as many of these bright stars are not in *Gaia* (but only in HIPPARCOS). 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 5, showing the HIP ID and the HIPPARCOS astrometry. This table allows a direct comparison with the average properties of the SigMA clusters in Tables 3–4. For the cases where there is a reasonable match, we name the cluster with the name of the bright B-star¹⁹. In some

cases, we name the clusters after their location in constellations. Additionally, we index the stellar clusters within this work from 1 to 37 as given in Col. “SigMA” in Table 3.

Figure 14 shows the SigMA cluster members in a *Gaia* CMD (similar to the HRD), confirming the youth of most sources. In Appendix D.1, we give more details on the chosen photometric quality criteria and the selection conditions to estimate the contamination from older populations or field stars. We find an excess of older low-mass sources that visibly separate from the Sco-Cen population, potentially false positive Sco-Cen members. We used a 25 Myr isochrone (to allow for random scatter) to separate “younger” Sco-Cen members from “older” populations or field stars as shown in Fig. 14 (middle panel). This gives a rough estimate for a contamination fraction of

¹⁹ This approach seems valid in particular for US, since also other authors, like Miret-Roig et al. (2022a) or Briceño-Morales & Chanamé (2023), independently decided for similar cluster names.

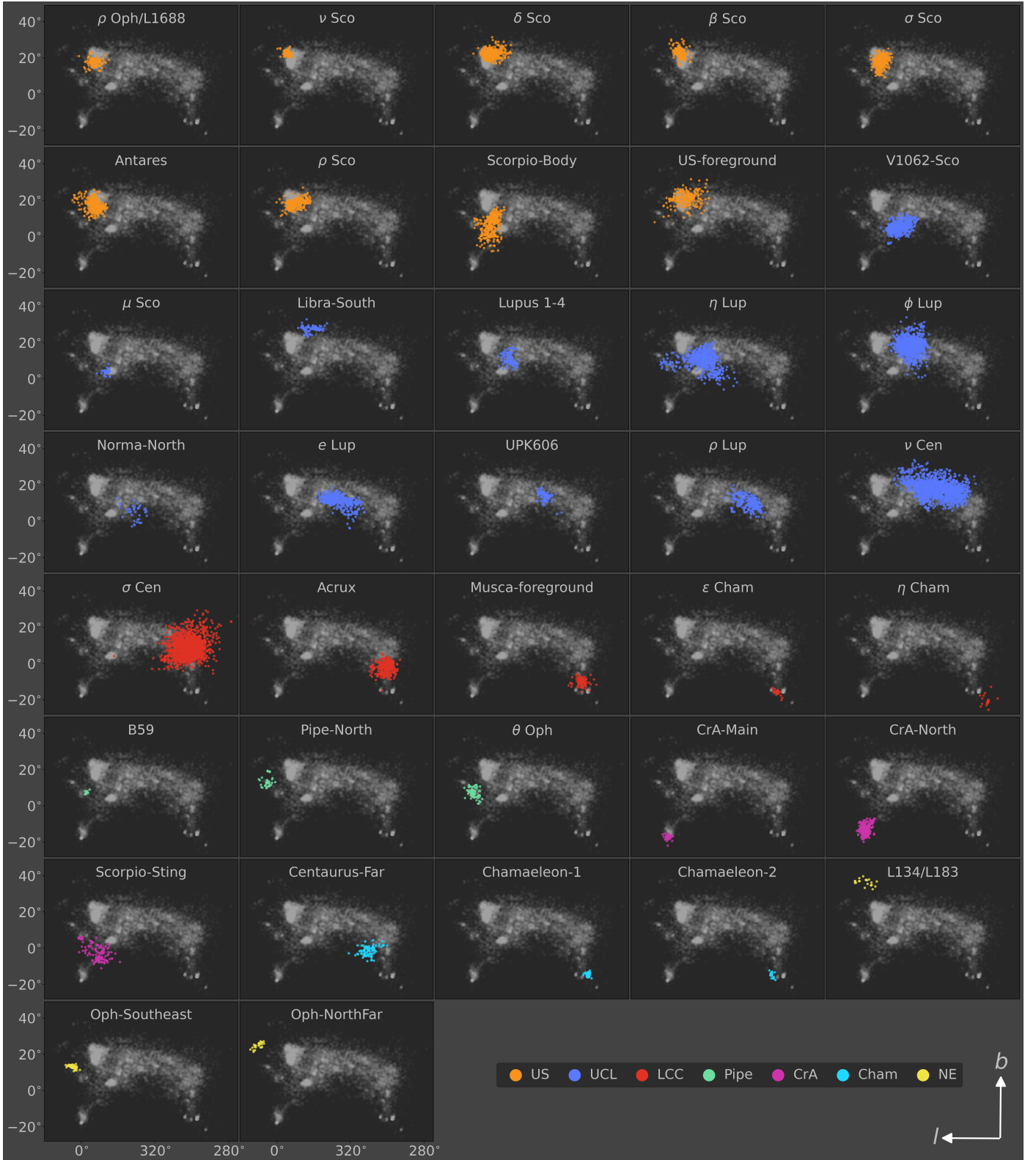


Fig. 11. Distribution of SigMA clusters in Sco-Cen, projected in Galactic coordinates, stratified by cluster membership. Compared to Fig. 10, the small multiples highlight the distribution of individual clusters in the Sco-Cen complex. The color coding represents the seven regions: US (orange), UCL (blue), LCC (red), Pipe (green), CrA (magenta), Cham (cyan), and NE (yellow).

about 4–10%, depending on the photometric quality criteria (see details in Appendix D.1 and Table D.1). This contamination fraction is similar to the estimate in Sect. 3.5.4. The influence of the stability that SigMA assigns each cluster member can be seen

in Fig. 8, where we show how different stability cuts influence the fraction of older sources, as estimated in Fig. 14. The trend in Fig. 8 suggests that a cut at about 11% would give a cleaner cluster membership selection and a lower contamination fraction

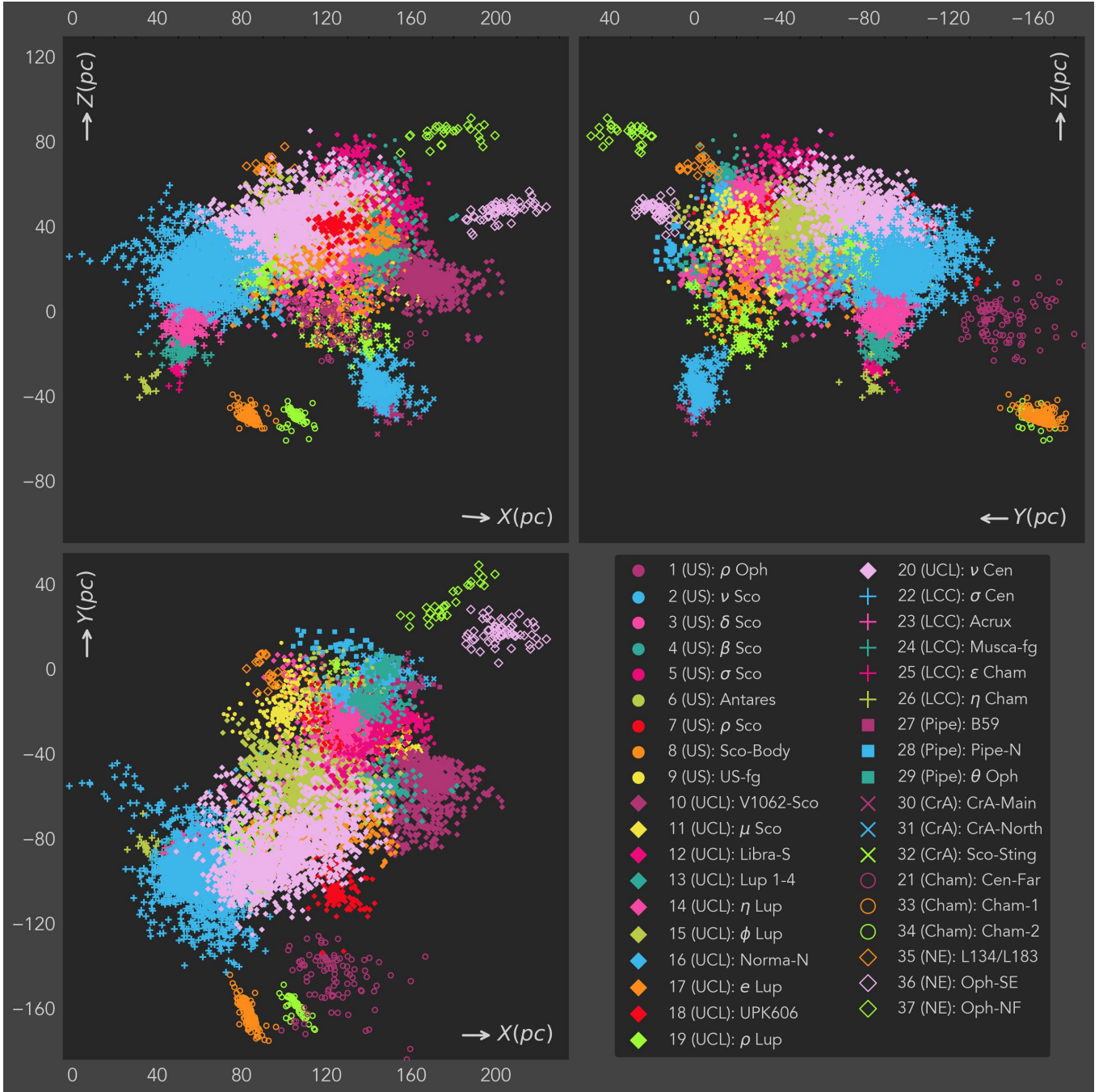


Fig. 12. 3D distribution of the 37 SigMA Sco-Cen clusters in heliocentric Galactic Cartesian coordinates. The Sun is at (0,0,0). Colors and labels are as in Fig. 10. See also the interactive 3D version [online](#) and Figs. E.1–E.5, which allow a better appreciation of individual cluster properties. By double-clicking on a cluster in the legend of the interactive version, the selected cluster can be isolated; by hovering over data points, the cluster membership and observed l , b , d position of a source becomes visible.

(Appendix D.1) since such a cut over-proportionally influences the older sources. However, a stability cut would also deliver an overall less complete sample.

Figure 14 (right panel) shows that there could be up to 19% of substellar candidates in the SigMA Sco-Cen sample, selected with an isomass line at $0.08 M_{\odot}$ from Baraffe et al. (2015, hereafter BHAC15; see Appendix D.2). In the future, more complete samples of the individual clusters can be obtained by using the known members as training sets (e.g., using Uncover, Ratzenböck et al. 2020). Knowing the brown dwarf population will allow the construction of more complete initial mass functions beyond the hydrogen burning limit and a

better characterization of the mass of the individual clusters (e.g., Miret-Roig et al. 2022b).

5.1. Overview of the seven subregions in Sco-Cen

In the following, to help compare SigMA results with the literature, we give a brief overview for each subregion within Sco-Cen (US, UCL, LCC, Pipe, CrA, Cham, and NE). We then give a more detailed comparison to recent works in Sect. 5.2. The listed seven subregions include four regions that are not a traditional part of the Sco-Cen OB association, namely CrA, Pipe, Cham, and NE clusters, while we find them to be part of Sco-Cen

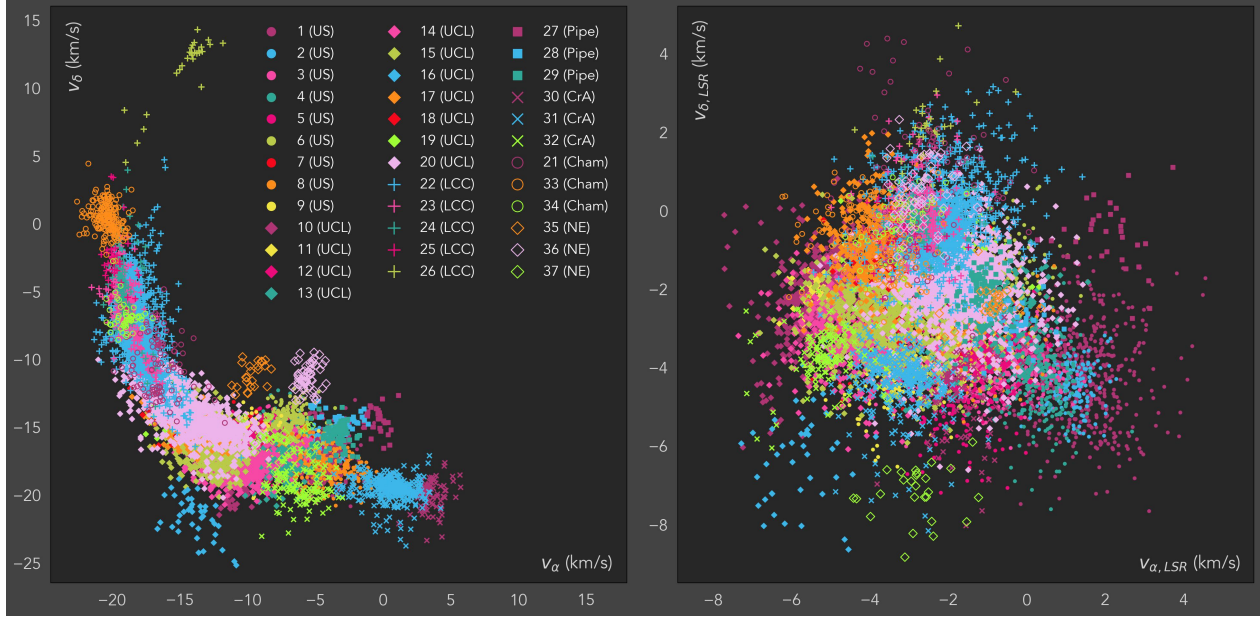


Fig. 13. Tangential velocity distribution of the 37 SigMA clusters. Colors and labels are as in Fig. 10. Left: Observed tangential velocities along α and δ are strongly influenced by the Sun’s reflex motion, while stellar clusters at similar distances and with similar space motions are arranged in a loop-like pattern. Sources at $l \sim 0^\circ$ are located in the lower-right part of the figure, and sources at $l \sim 290^\circ$ in the upper-left part of the figure (see Fig. C.1). Right: Tangential velocities corrected for the Sun’s motion, and hence relative to the LSR. The correction reduces the projection effects of the observed tangential stellar motions. See the interactive 2D version [online](#) and Figs. E.1–E.5 for a better appreciation of the 2D kinematical properties of the clusters in Sco-Cen.

because they are co-moving with the complex itself, and are likely within the 20 Myr age cut we used for the association. Even if we assign each stellar cluster to one of the seven sub-regions, 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.

5.1.1. Upper Scorpius (US)

Toward US we identify nine clusters containing in total 3596 stellar sources, which are partially extending beyond the traditional borders (Fig. 10). Of these nine clusters, seven appear higher surface density and tend to be associated with prominent B-stars, as already pointed out above, namely ρ Oph/L1688, β Sco, δ Sco, ν Sco, σ Sco, Antares, and ρ Sco (see Tables 3–5). The remaining two clusters appear more extended, which we name US-foreground and Scorpio-Body.

The clusters ρ Oph/L1688, Antares, and ρ Sco show significant overlap in the same volume in space, while separating in velocity space. In a recent paper (Grasser et al. 2021) we studied the ρ 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 coincide with the ρ Oph/L1688 and Antares clusters, respectively. In detail, the cross-matched Pop 1 sample contains $\sim 93\%$ of the ρ Oph/L1688 group and few matches with other clusters (Antares, σ Sco, β Sco, δ Sco). The cross-matched Pop 2 sample contains $\sim 75\%$ of the Antares group and few matches with other clusters (ρ Sco, σ Sco, US-foreground). Luhman (2022) point out that “new” ρ 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) refer to sources not previously assigned as members of the young ρ Oph/L1688 star-forming event. The two intertwining distinct populations (both containing new sources) within the same volume have been first

studied in detail in Grasser et al. (2021). In this work, we identify another stellar population, ρ Sco, which also seems to occupy a similar volume in space, partially overlapping with the two populations while having distinct velocities from these.

The group US-foreground is located in front of the more compact clusters, visible in 3D space (Fig. 12), hence the chosen name. Finally, the Scorpio-Body group 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. 5.1.5). It spans the Scorpius constellation’s central body, hence the name. The nine clusters toward US reveal a complex star formation history, which will be further discussed in future work.

5.1.2. Upper Centaurus Lupus (UCL)

We identify rich substructure within UCL separated into 11 SigMA clusters (5935 stellar sources), as listed in Table 3. The most prominent cluster in the region is V1062 Sco (Röser et al. 2018), lying toward 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 2022; see Sects. 5.2.1, 5.2.4). We identify a second cluster close to V1062 Sco, which we call μ Sco, since its members are scattered around the bright B-star μ Sco. We find that the positions and velocities of the two SigMA clusters are very similar, and members of both clusters are part of V1062-Sco-selections in previous work (Sect. 5.2), also named UPK 640 in Cantat-Gaudin & Anders (2020). The star μ Sco, which is the name-giver of μ Sco lies in the center of the cluster, while the star μ Sco 2 is 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.

Lupus 1–4 appears correlated with regions of high dust column-density, matching with previous selections of Lupus–3

Table 3. Overview of the 37 SigMA clusters in Sco-Cen, assigned to seven subregions (Col. 2).

SigMA	Region	Group name	Brightest star ^(a)	Nr.	l (deg)	b (deg)	ϖ (mas)	d (pc)	X (pc)	Y (pc)	Z (pc)
1	US	ρ Oph/L1688	* rho Oph	535	353.20 ^{+0.57} _{-0.81}	17.09 ^{+1.24} _{-0.81}	7.20 ^{+0.23} _{-0.28}	139 ⁺⁶ ₋₄	132 ⁺⁶ ₋₄	-16 ⁺¹ ₋₂	41 ⁺³ ₋₂
2	US	ν Sco	* nu Sco	150	354.46 ^{+0.81} _{-1.03}	22.86 ^{+0.85} _{-0.92}	7.18 ^{+0.16} _{-0.20}	139 ⁺⁴ ₋₃	128 ⁺³ ₋₃	-12 ⁺² ₋₂	54 ⁺² ₋₂
3	US	δ Sco	* b Sco	691	350.25 ^{+2.11} _{-3.66}	22.18 ^{+1.50} _{-1.91}	7.03 ^{+0.20} _{-0.25}	142 ⁺⁵ ₋₄	130 ⁺⁵ ₋₄	-22 ⁺⁵ ₋₉	54 ⁺⁴ ₋₄
4	US	β Sco	HD 142883	285	353.21 ^{+1.48} _{-1.26}	23.53 ^{+1.54} _{-3.08}	6.49 ^{+0.19} _{-0.25}	154 ⁺⁶ ₋₄	141 ⁺⁷ ₋₅	-17 ⁺⁴ ₋₄	61 ⁺⁵ ₋₆
5	US	σ Sco	* c02 Sco	544	351.13 ^{+1.23} _{-2.34}	17.87 ^{+1.92} _{-2.31}	6.29 ^{+0.23} _{-0.26}	159 ⁺⁷ ₋₆	149 ⁺⁸ ₋₇	-24 ⁺⁴ ₋₆	49 ⁺⁴ ₋₆
6	US	Antares	HD 146001	502	352.79 ^{+2.18} _{-2.28}	17.22 ^{+2.83} _{-2.71}	7.21 ^{+0.45} _{-0.44}	139 ⁺⁹ ₋₈	132 ⁺⁹ ₋₉	-17 ⁺⁵ ₋₆	41 ⁺⁵ ₋₆
7	US	ρ Sco	* rho Sco	240	349.31 ^{+3.85} _{-4.77}	18.29 ^{+1.80} _{-2.76}	7.21 ^{+0.47} _{-0.36}	139 ⁺⁷ ₋₈	129 ⁺⁸ ₋₁₀	-24 ⁺⁹ ₋₁₁	43 ⁺⁵ ₋₆
8	US	Scorpio-Body	HD 150638	373	349.36 ^{+3.32} _{-2.49}	7.32 ^{+4.09} _{-7.01}	7.08 ^{+0.91} _{-0.58}	141 ⁺¹³ ₋₁₆	137 ⁺¹¹ ₋₁₅	-26 ⁺⁹ ₋₇	17 ⁺¹³ ₋₁₇
9	US	US-foreground	HD 145964	276	348.81 ^{+5.68} _{-3.80}	21.04 ^{+3.35} _{-3.46}	9.05 ^{+0.67} _{-0.74}	110 ⁺¹⁰ ₋₈	102 ⁺⁹ ₋₇	-20 ⁺⁹ ₋₆	39 ⁺⁸ ₋₇
10	UCL	V1062-Sco	* mu02 Sco	1029	343.11 ^{+1.35} _{-4.41}	4.69 ^{+1.53} _{-1.50}	5.66 ^{+0.23} _{-0.24}	177 ⁺⁸ ₋₇	168 ⁺⁷ ₋₈	-51 ⁺⁵ ₋₁₄	14 ⁺⁵ ₋₄
11	UCL	μ Sco	HD 151726	54	346.19 ^{+0.95} _{-0.65}	3.90 ^{+0.48} _{-0.75}	6.07 ^{+0.20} _{-0.09}	165 ⁺³ ₋₅	160 ⁺³ ₋₅	-40 ⁺³ ₋₁	11 ⁺¹ ₋₂
12	UCL	Libra-South	HD 138343	71	341.77 ^{+3.12} _{-4.68}	27.80 ^{+1.07} _{-1.42}	6.34 ^{+0.30} _{-0.19}	158 ⁺⁵ ₋₆	132 ⁺⁶ ₋₉	-45 ⁺⁹ ₋₈	73 ⁺⁴ ₋₆
13	UCL	Lupus 1-4	* LL Lup	226	339.51 ^{+1.15} _{-2.72}	9.44 ^{+3.59} _{-0.99}	6.27 ^{+0.16} _{-0.22}	160 ⁺⁶ ₋₄	147 ⁺⁶ ₋₄	-55 ⁺³ ₋₃	26 ⁺¹¹ ₋₃
14	UCL	η Lup	* eta Lup	769	339.83 ^{+6.08} _{-4.45}	10.71 ^{+3.13} _{-4.50}	7.37 ^{+1.04} _{-0.46}	136 ⁺⁹ ₋₁₇	124 ⁺⁸ ₋₁₃	-47 ⁺¹⁵ ₋₁₀	25 ⁺⁸ ₋₁₁
15	UCL	ϕ Lup	* phi02 Lup	1114	334.17 ^{+4.39} _{-4.22}	17.69 ^{+3.70} _{-3.78}	7.65 ^{+1.06} _{-0.78}	131 ⁺⁵ ₋₁₆	112 ⁺¹² ₋₁₆	-54 ⁺¹⁰ ₋₉	40 ⁺⁹ ₋₁₀
16	UCL	Norma-North	HD 143215	42	331.12 ^{+4.01} _{-2.98}	6.39 ^{+2.19} _{-6.64}	9.44 ^{+0.80} _{-0.56}	106 ⁺⁷ ₋₈	92 ⁺⁶ ₋₅	-50 ⁺⁹ ₋₉	11 ⁺⁴ ₋₁₂
17	UCL	e Lup	* e Lup	516	327.40 ^{+3.92} _{-6.80}	11.48 ^{+1.93} _{-2.35}	6.89 ^{+0.65} _{-0.49}	145 ⁺¹¹ ₋₁₃	120 ⁺¹³ ₋₂₀	-76 ⁺⁶ ₋₉	29 ⁺⁶ ₋₇
18	UCL	UPK606	HD 125777	131	320.00 ^{+2.17} _{-1.37}	13.72 ^{+1.36} _{-1.68}	5.92 ^{+0.27} _{-0.25}	169 ⁺⁷ ₋₇	125 ⁺⁷ ₋₄	-106 ⁺⁹ ₋₄	40 ⁺³ ₋₄
19	UCL	ρ Lup	* rho Lup	246	315.12 ^{+6.44} _{-3.47}	9.83 ^{+3.78} _{-3.09}	8.16 ^{+0.55} _{-0.57}	123 ⁺⁹ ₋₈	87 ⁺⁷ ₋₈	-82 ⁺⁹ ₋₁₃	21 ⁺⁷ ₋₆
20	UCL	ν Cen	* nu Cen	1737	318.50 ^{+9.60} _{-8.02}	17.45 ^{+4.13} _{-4.08}	7.21 ^{+0.67} _{-0.69}	139 ⁺¹⁵ ₋₁₂	99 ⁺²² ₋₁₉	-87 ⁺¹⁸ ₋₁₂	41 ⁺¹² ₋₉
21	LCC	σ Cen	* sig Cen	1805	301.56 ^{+5.56} _{-5.32}	8.35 ^{+4.99} _{-4.74}	8.71 ^{+0.87} _{-1.05}	115 ⁺¹⁶ ₋₁₀	60 ⁺¹² ₋₁₁	-96 ⁺¹⁰ ₋₁₃	17 ⁺¹¹ ₋₁₀
22	LCC	Acrux	* zet Crux	394	300.27 ^{+3.20} _{-1.71}	-1.98 ^{+2.05} _{-3.53}	9.41 ^{+0.47} _{-0.37}	106 ⁺⁴ ₋₅	54 ⁺⁵ ₋₃	-91 ⁺⁷ ₋₄	-4 ⁺⁴ ₋₆
23	LCC	Musca-foreground	HD 107947	95	300.64 ^{+1.69} _{-2.18}	-10.27 ^{+1.81} _{-1.92}	9.79 ^{+0.30} _{-0.40}	102 ⁺⁴ ₋₃	52 ⁺³ ₋₄	-86 ⁺³ ₋₄	-18 ⁺³ ₋₄
24	LCC	ϵ Cham	* DX Cha	39	300.34 ^{+0.71} _{-0.27}	-15.97 ^{+0.68} _{-0.76}	9.81 ^{+0.18} _{-0.26}	102 ⁺³ ₋₂	50 ⁺¹ ₋₁	-85 ⁺³ ₋₃	-28 ⁺¹ ₋₂
25	LCC	η Cham	* eta Cha	30	292.49 ^{+1.56} _{-0.48}	-21.60 ^{+2.30} _{-0.32}	10.14 ^{+0.40} _{-0.13}	99 ⁺¹ ₋₄	35 ⁺² ₋₁	-85 ⁺² ₋₁	-36 ⁺⁵ ₋₁
26	Pipe	B59	Em* AS 218	32	357.10 ^{+0.38} _{-0.30}	7.11 ^{+0.59} _{-0.63}	6.23 ^{+0.12} _{-0.16}	160 ⁺⁴ ₋₃	159 ⁺⁴ ₋₃	-8 ⁺¹ ₋₁	20 ⁺² ₋₂
27	Pipe	Pipe-North	HD 155427	42	4.92 ^{+1.29} _{-1.15}	12.85 ^{+2.53} _{-1.98}	7.69 ^{+0.69} _{-0.30}	130 ⁺⁵ ₋₁₁	126 ⁺⁶ ₋₁₂	11 ⁺³ ₋₃	29 ⁺⁶ ₋₅
28	Pipe	θ Oph	HD 158704	98	359.71 ^{+1.09} _{-2.75}	7.05 ^{+2.66} _{-1.95}	6.79 ^{+0.29} _{-0.20}	147 ⁺⁵ ₋₆	146 ⁺⁴ ₋₆	-1 ⁺³ ₋₇	19 ⁺⁵ ₋₆
29	CrA	CrA-Main	HD 177076	96	359.87 ^{+0.38} _{-0.66}	-17.65 ^{+0.70} _{-0.32}	6.46 ^{+0.13} _{-0.13}	155 ⁺³ ₋₃	147 ⁺³ ₋₃	0 ⁺¹ ₋₂	-47 ⁺² ₋₁
30	CrA	CrA-North	HD 172910	351	359.02 ^{+1.17} _{-1.77}	-13.97 ^{+2.74} _{-1.91}	6.70 ^{+0.25} _{-0.29}	149 ⁺⁷ ₋₅	145 ⁺⁷ ₋₅	-2 ⁺³ ₋₅	-36 ⁺⁸ ₋₅
31	CrA	Scorpio-Sting	HD 157864	132	350.59 ^{+5.27} _{-3.11}	-3.04 ^{+4.72} _{-3.77}	7.49 ^{+0.66} _{-0.52}	134 ⁺¹⁰ ₋₁₁	131 ⁺⁹ ₋₁₁	-22 ⁺¹² ₋₆	-7 ⁺¹¹ ₋₉
32	Cham	Centaurus-Far	HD 121808	99	310.69 ^{+2.02} _{-3.30}	-1.18 ^{+2.14} _{-2.74}	5.25 ^{+0.35} _{-0.31}	190 ⁺¹² ₋₁₂	122 ⁺¹⁵ ₋₁₂	-142 ⁺⁶ ₋₁₅	-4 ⁺⁷ ₋₉
33	Cham	Chamaeleon-1	V* CR Cha	192	297.22 ^{+0.21} _{-0.52}	-15.52 ^{+0.90} _{-0.37}	5.25 ^{+0.18} _{-0.10}	191 ⁺⁴ ₋₆	84 ⁺² ₋₃	-164 ⁺⁶ ₋₄	-50 ⁺² ₋₂
34	Cham	Chamaeleon-2	V* BF Cha	54	303.69 ^{+0.23} _{-0.37}	-14.72 ^{+0.74} _{-0.38}	5.08 ^{+0.13} _{-0.16}	197 ⁺⁶ ₋₅	105 ⁺⁴ ₋₃	-159 ⁺⁴ ₋₅	-50 ⁺³ ₋₄
35	NE	L134/L183	HD 141569	24	358.40 ^{+5.78} _{-3.05}	36.84 ^{+0.79} _{-2.22}	8.76 ^{+0.24} _{-0.32}	114 ⁺⁴ ₋₃	93 ⁺⁴ ₋₄	-3 ⁺⁹ ₋₅	67 ⁺⁴ ₋₁
36	NE	Oph-Southeast	HD 154922	61	4.54 ^{+1.49} _{-1.01}	13.06 ^{+0.68} _{-0.63}	4.80 ^{+0.19} _{-0.24}	208 ⁺¹¹ ₋₈	202 ⁺¹¹ ₋₈	16 ⁺⁶ ₋₄	48 ⁺³ ₋₄
37	NE	Oph-NorthFar	BD-06 4472	28	9.60 ^{+2.56} _{-1.57}	24.98 ^{+1.32} _{-1.33}	5.06 ^{+0.24} _{-0.37}	198 ⁺¹⁵ ₋₉	176 ⁺¹⁶ ₋₈	29 ⁺¹¹ ₋₅	85 ⁺² ₋₆

Notes. The median cluster positions are listed (see Table 4 for the median velocities). In Cols. 6–12, we list the medians of the positional parameters for each cluster including all cluster members (without considering any stability cut). The given lower and upper uncertainties represent the 1σ scatter around the median. In this scatter, the original measurement uncertainties of single stellar sources are not considered. ^(a)Column 4 lists the brightest star that was selected as a member by SigMA. The star annotation “*” is used as in the SIMBAD astronomical database (Wenger et al. 2000) and helps distinguish stellar names from cluster names throughout the manuscript since some clusters are named after bright stars. Further bright stellar member candidates, which were observed by HIPPARCOS (partially not in *Gaia* DR3), are listed in Table 5.

and 4 stellar members (e.g., Damiani et al. 2019; Kerr et al. 2021), which are merged in our SigMA selection. The average distance to Lupus 1–4 matches well with cloud distance estimates from Zucker et al. (2021, derived from Leike et al. 2020), who report a distance between 155–198 pc, or an average of

about 165 pc for the Lupus 1–4 clouds. Similar distances have been reported in Teixeira et al. (2020) or Galli et al. (2020a).

At the heart of UCL lie the clusters η Lup, ϕ Lup, and e Lup, which likely belong to the oldest parts of Sco-Cen, probably the clusters where the first supernovae in Sco-Cen originated from

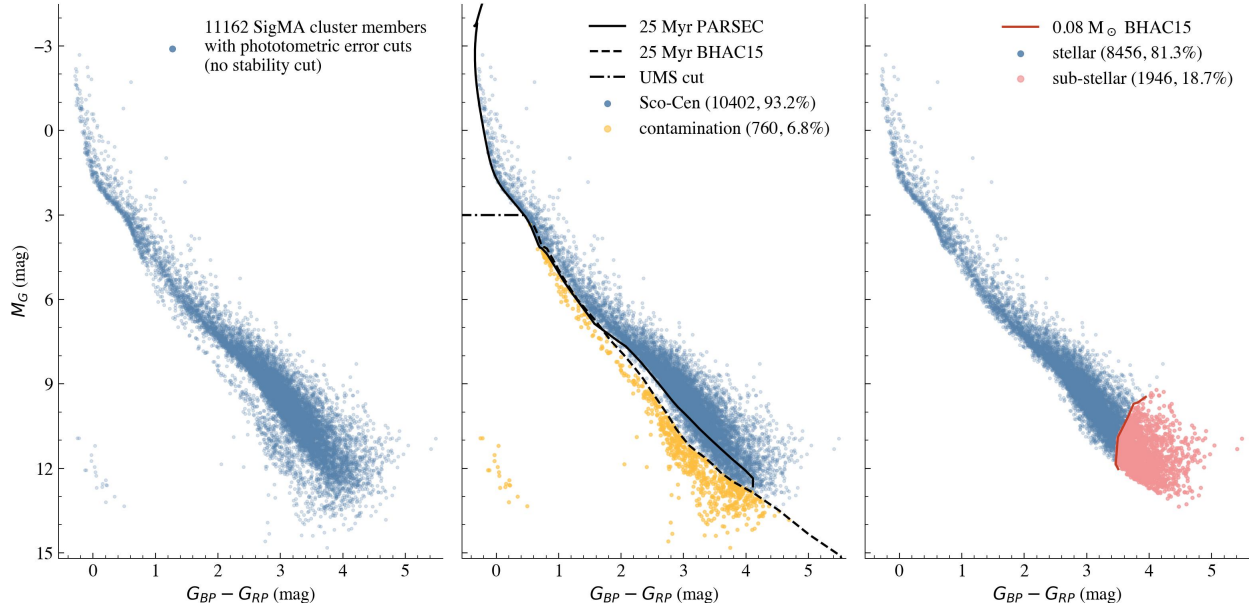


Fig. 14. *Gaia* CMD using M_G versus $G_{BP} - G_{RP}$ showing the SigMA-selected Sco-Cen members. Left: SigMA cluster members that pass the photometric quality criteria as given in Eq. (D.2). Middle: Potential contamination from older sources (orange), selected with a 25 Myr isochrone from PARSEC (black line) and Baraffe et al. (2015, dashed black line) and an additional cut at $M_G > 3$ mag (dashed-dotted black line), which excludes the UMS. The combined conditions indicate contamination from older sources of about 6–7% when using the given photometric quality criteria and no stability cut. Right: Substellar candidates (red dots) selected with a $0.08 M_{\odot}$ isomass line from Baraffe et al. (2015, dark red line) using only the younger source from the middle panel. This cut indicates that roughly 19% of substellar sources are within the SigMA Sco-Cen members when the mentioned cuts and photometric quality criteria are applied and extinction effects are ignored. More details on the quality criteria, the selection borders, and the isochrone models are given in Appendix D.

(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.

There is one cluster slightly in front and to the south of the main UCL body, called Norma-North, named after its location in that constellation. This is a new clustering, which does not have a clear counterpart in the literature. Another SigMA cluster lies to the far side of UCL and the Galactic west of the Lupus constellation. This cluster correlates with UPK 606 when compared to Cantat-Gaudin & Anders (2020; see also Kerr et al. 2021 and Table E.3). Finally, to the Galactic west, UCL connects with LCC via the clusters ν Cen and ρ Lup.

5.1.3. Lower Centaurus Crux (LCC)

We find five SigMA clusters (2363 stellar sources) toward the LCC region (see Table 3), which is now reaching farther below the Galactic plane compared to earlier definitions in the literature. For the SigMA extraction, the young local associations ϵ Cham and η Cham are part of LCC, located at the Southern most tip, confirming the results of Mamajek et al. (1999, 2000) or Fernández et al. (2008). Consequently, the main body of LCC is composed of five subgroups, which seem to constitute an age gradient (e.g., Kerr et al. 2021) from north to south (σ Cen, Acrux, Musca-foreground, ϵ Cham, and η Cham), which we analyze in a follow-up study (Ratzenböck et al. 2023).

5.1.4. Pipe Nebula

Although not traditionally considered part of the Sco-Cen association, we find three SigMA clusters toward the Pipe nebula (172 stellar sources), including B59, Pipe-North, and θ 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 toward cluster and cloud but also by the cloud distance between 147–163 pc (Zucker et al. 2021), compatible with the cluster distance of about 160 pc. The θ Oph cluster, surrounding the B2 star θ Oph, 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 (Gritschneider & Lin 2012). Pipe-North lies slightly in front of the other two clusters (at about 130 pc) and to the Galactic north of the Pipe Nebula, as the name suggests.

5.1.5. 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 count three SigMA clusters to the CrA region, containing a total of 579 sources. We identify a distinct cluster projected on top of the CrA molecular cloud and the embedded Coronet clusters, which we call the CrA-Main group. The cluster distance fits the cloud distance of about 136–179 pc (Zucker et al. 2021). To the Galactic North, we identify a second and more extended group, called CrA-North, which was already discussed in Galli et al. (2020b) or Esplin & Luhman (2022). Additionally, we identify a third group to the Galactic northwest of the two other clusters, 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. 5.2.2 and Table E.3), while they identify a smaller subsample of this group

Table 4. Median velocity parameters for the 37 SigMA clusters in Sco-Cen.

SigMA	Group name (short version)	Nr.	μ_α^* (mas yr ⁻¹)	μ_δ (mas yr ⁻¹)	v_α (km s ⁻¹)	v_δ (km s ⁻¹)	$v_{\alpha,\text{LSR}}$ (km s ⁻¹)	$v_{\delta,\text{LSR}}$ (km s ⁻¹)
1 (US)	ρ Oph	535	-6.81 ^{+1.36} _{-1.56}	-25.91 ^{+1.78} _{-1.58}	-4.50 ^{+0.89} _{-1.01}	-16.95 ^{+0.91} _{-1.28}	1.36 ^{+0.84} _{-0.94}	-4.20 ^{+1.01} _{-1.15}
2	ν Sco	150	-8.47 ^{+0.74} _{-0.81}	-24.50 ^{+0.76} _{-0.98}	-5.62 ^{+0.55} _{-0.56}	-16.22 ^{+0.61} _{-0.73}	1.21 ^{+0.37} _{-0.43}	-4.52 ^{+0.58} _{-0.60}
3	δ Sco	691	-12.24 ^{+2.04} _{-2.54}	-23.92 ^{+1.27} _{-1.35}	-8.22 ^{+1.40} _{-1.99}	-16.27 ^{+0.88} _{-0.75}	-0.77 ^{+0.88} _{-1.09}	-3.88 ^{+0.57} _{-0.66}
4	β Sco	285	-9.39 ^{+1.37} _{-1.07}	-21.69 ^{+1.00} _{-1.73}	-6.87 ^{+0.86} _{-0.77}	-15.86 ^{+0.77} _{-1.50}	0.26 ^{+0.56} _{-0.59}	-4.13 ^{+0.58} _{-0.96}
5	σ Sco	544	-10.68 ^{+1.47} _{-1.37}	-21.79 ^{+1.34} _{-1.87}	-8.12 ^{+1.09} _{-0.87}	-16.49 ^{+0.92} _{-1.23}	-1.68 ^{+0.78} _{-0.61}	-3.48 ^{+0.91} _{-1.31}
6	Antares	502	-11.00 ^{+1.59} _{-1.25}	-23.31 ^{+2.15} _{-2.00}	-7.24 ^{+0.87} _{-0.69}	-15.52 ^{+1.29} _{-0.72}	-1.29 ^{+0.84} _{-0.74}	-2.43 ^{+0.86} _{-1.17}
7	ρ Sco	240	-15.98 ^{+1.96} _{-2.08}	-24.14 ^{+1.65} _{-1.72}	-10.44 ^{+1.22} _{-1.18}	-15.82 ^{+0.74} _{-0.85}	-3.64 ^{+0.52} _{-0.59}	-2.78 ^{+0.62} _{-0.78}
8	Sco-Body	373	-8.04 ^{+2.92} _{-1.91}	-26.82 ^{+3.67} _{-3.55}	-5.30 ^{+2.16} _{-1.70}	-17.64 ^{+1.06} _{-1.08}	-1.38 ^{+0.69} _{-0.39}	-2.83 ^{+0.52} _{-0.67}
9	US-fg	276	-19.97 ^{+2.74} _{-2.86}	-31.25 ^{+4.09} _{-3.02}	-10.63 ^{+1.33} _{-0.96}	-16.28 ^{+1.20} _{-1.07}	-3.33 ^{+1.16} _{-0.73}	-3.86 ^{+0.88} _{-0.77}
10 (UCL)	V1062-Sco	1029	-12.14 ^{+0.96} _{-2.02}	-21.15 ^{+0.93} _{-0.82}	-10.18 ^{+0.75} _{-1.43}	-17.73 ^{+0.75} _{-0.61}	-5.12 ^{+0.41} _{-0.52}	-2.15 ^{+0.66} _{-0.55}
11	μ Sco	54	-11.75 ^{+0.42} _{-0.83}	-22.61 ^{+0.55} _{-0.37}	-9.22 ^{+0.54} _{-0.53}	-17.61 ^{+0.49} _{-0.44}	-5.28 ^{+0.30} _{-0.25}	-2.11 ^{+0.41} _{-0.31}
12	Libra-S	71	-14.87 ^{+1.55} _{-1.76}	-20.92 ^{+1.22} _{-1.03}	-11.15 ^{+1.11} _{-0.86}	-15.44 ^{+0.45} _{-0.71}	-1.00 ^{+0.53} _{-0.62}	-3.71 ^{+0.43} _{-0.42}
13	Lup 1-4	226	-10.54 ^{+1.51} _{-1.67}	-23.39 ^{+1.01} _{-1.02}	-8.07 ^{+1.18} _{-1.41}	-17.77 ^{+0.95} _{-0.74}	-0.76 ^{+0.79} _{-1.07}	-2.85 ^{+0.75} _{-0.92}
14	η Lup	769	-17.63 ^{+2.28} _{-2.01}	-27.84 ^{+2.12} _{-3.93}	-11.21 ^{+1.70} _{-1.34}	-17.83 ^{+0.95} _{-0.83}	-4.15 ^{+0.79} _{-0.85}	-3.18 ^{+0.83} _{-0.65}
15	ϕ Lup	1114	-20.99 ^{+2.67} _{-4.71}	-25.60 ^{+3.00} _{-4.16}	-13.31 ^{+1.48} _{-1.24}	-15.82 ^{+0.90} _{-1.38}	-3.37 ^{+1.05} _{-1.00}	-2.45 ^{+0.85} _{-1.03}
16	Norma-N	42	-27.93 ^{+2.68} _{-2.41}	-42.80 ^{+2.50} _{-4.59}	-14.07 ^{+1.61} _{-1.03}	-21.39 ^{+1.22} _{-1.69}	-5.73 ^{+0.82} _{-0.83}	-6.05 ^{+0.73} _{-1.40}
17	ϵ Lup	516	-20.80 ^{+2.73} _{-4.62}	-21.67 ^{+1.78} _{-1.60}	-14.44 ^{+1.17} _{-1.46}	-15.07 ^{+1.42} _{-0.67}	-3.81 ^{+0.54} _{-0.50}	-1.05 ^{+0.79} _{-0.60}
18	UPK606	131	-20.07 ^{+1.27} _{-0.73}	-17.01 ^{+0.71} _{-1.17}	-15.96 ^{+0.87} _{-0.49}	-13.69 ^{+0.44} _{-0.60}	-3.20 ^{+0.44} _{-0.33}	-1.14 ^{+0.47} _{-0.31}
19	ρ Lup	246	-26.22 ^{+2.43} _{-2.95}	-23.13 ^{+3.50} _{-2.37}	-15.51 ^{+1.52} _{-1.06}	-13.14 ^{+1.35} _{-1.36}	-1.95 ^{+0.45} _{-0.73}	-1.18 ^{+0.41} _{-0.40}
20	ν Cen	1737	-23.33 ^{+5.86} _{-4.69}	-20.27 ^{+2.26} _{-2.61}	-15.28 ^{+2.99} _{-1.84}	-13.53 ^{+1.99} _{-1.64}	-1.78 ^{+0.79} _{-0.73}	-1.71 ^{+0.61} _{-0.82}
21 (LCC)	σ Cen	1805	-33.23 ^{+4.14} _{-3.71}	-13.67 ^{+3.91} _{-5.10}	-18.27 ^{+1.02} _{-0.69}	-7.70 ^{+2.33} _{-2.30}	-2.20 ^{+0.50} _{-0.67}	-0.41 ^{+0.77} _{-0.63}
22	Acrux	394	-37.73 ^{+1.83} _{-1.67}	-11.36 ^{+2.84} _{-4.43}	-19.10 ^{+0.68} _{-0.40}	-5.69 ^{+1.32} _{-2.00}	-2.75 ^{+0.32} _{-0.37}	-0.17 ^{+0.45} _{-0.35}
23	Musca-fg	95	-39.37 ^{+1.82} _{-1.57}	-9.33 ^{+4.36} _{-3.95}	-19.20 ^{+0.61} _{-0.33}	-4.49 ^{+2.05} _{-2.02}	-2.85 ^{+0.31} _{-0.22}	-0.23 ^{+0.44} _{-0.40}
24	ϵ Cham	39	-41.23 ^{+2.27} _{-0.87}	-6.05 ^{+2.04} _{-2.99}	-19.85 ^{+0.73} _{-0.42}	-2.92 ^{+1.01} _{-1.44}	-3.31 ^{+0.50} _{-0.51}	-0.54 ^{+0.49} _{-0.46}
25	η Cham	30	-30.16 ^{+1.93} _{-4.40}	26.86 ^{+1.24} _{-5.90}	-14.06 ^{+0.69} _{-2.04}	12.55 ^{+0.67} _{-3.18}	-2.58 ^{+0.47} _{-0.25}	2.09 ^{+0.59} _{-0.51}
26 (Pipe)	B59	32	-0.49 ^{+0.70} _{-1.16}	-18.84 ^{+0.41} _{-0.60}	-0.37 ^{+0.53} _{-0.89}	-14.48 ^{+0.71} _{-0.76}	2.13 ^{+0.70} _{-0.63}	-0.61 ^{+0.68} _{-0.83}
27	Pipe-N	42	-4.78 ^{+1.51} _{-2.56}	-23.36 ^{+0.62} _{-1.65}	-3.08 ^{+1.04} _{-1.30}	-14.55 ^{+0.79} _{-0.36}	-0.27 ^{+0.44} _{-0.65}	-2.79 ^{+0.38} _{-0.28}
28	θ Oph	98	-4.71 ^{+0.45} _{-0.94}	-21.85 ^{+0.67} _{-2.08}	-3.29 ^{+0.39} _{-0.73}	-15.41 ^{+0.35} _{-0.93}	-1.16 ^{+0.38} _{-0.40}	-2.03 ^{+0.45} _{-0.70}
29 (CrA)	CrA-Main	96	4.57 ^{+1.19} _{-0.70}	-27.11 ^{+0.88} _{-1.33}	3.33 ^{+0.98} _{-0.47}	-19.83 ^{+0.63} _{-1.21}	-1.76 ^{+0.83} _{-0.49}	-4.77 ^{+0.67} _{-1.04}
30	CrA-North	351	0.92 ^{+1.84} _{-2.34}	-27.60 ^{+1.17} _{-1.00}	0.65 ^{+1.32} _{-1.59}	-19.51 ^{+0.68} _{-0.63}	-3.16 ^{+0.59} _{-0.80}	-4.25 ^{+0.54} _{-0.62}
31	Sco-Sting	132	-10.03 ^{+2.98} _{-2.27}	-29.75 ^{+1.78} _{-4.47}	-6.12 ^{+1.33} _{-1.49}	-19.51 ^{+1.49} _{-0.82}	-5.15 ^{+0.52} _{-0.43}	-3.70 ^{+0.65} _{-0.44}
32 (Cham)	Cen-Far	99	-18.40 ^{+1.87} _{-2.36}	-11.75 ^{+2.75} _{-2.67}	-16.64 ^{+1.24} _{-1.26}	-10.51 ^{+2.22} _{-2.15}	-2.79 ^{+1.28} _{-0.83}	0.31 ^{+1.72} _{-1.50}
33	Cham-1	192	-22.55 ^{+0.72} _{-0.91}	0.38 ^{+1.21} _{-1.11}	-20.31 ^{+0.64} _{-0.71}	0.35 ^{+1.06} _{-1.01}	-3.96 ^{+0.74} _{-0.70}	-0.79 ^{+0.86} _{-0.79}
34	Cham-2	54	-20.16 ^{+0.76} _{-0.94}	-7.55 ^{+0.77} _{-0.64}	-18.95 ^{+0.73} _{-0.75}	-7.07 ^{+0.83} _{-0.53}	-3.24 ^{+0.55} _{-0.57}	-0.22 ^{+0.51} _{-0.64}
35 (NE)	L134/L183	24	-17.63 ^{+1.02} _{-0.97}	-20.28 ^{+1.00} _{-1.52}	-9.75 ^{+0.92} _{-0.62}	-11.13 ^{+0.94} _{-0.92}	-0.83 ^{+0.17} _{-0.36}	-2.44 ^{+0.34} _{-0.17}
36	Oph-SE	61	-5.65 ^{+0.61} _{-0.54}	-11.44 ^{+0.66} _{-0.75}	-5.63 ^{+0.54} _{-0.60}	-11.58 ^{+1.26} _{-0.63}	-2.77 ^{+0.51} _{-0.46}	0.30 ^{+1.04} _{-0.69}
37	Oph-NF	28	-8.57 ^{+1.88} _{-1.46}	-16.75 ^{+0.99} _{-1.59}	-7.78 ^{+0.96} _{-1.11}	-15.89 ^{+0.41} _{-0.96}	-2.83 ^{+0.44} _{-0.85}	-7.11 ^{+0.52} _{-0.79}

Notes. In Cols. 4–9, we list the medians of the velocity parameters for each cluster including all cluster members (without considering any stability cut). The given lower and upper uncertainties represent the 1σ scatter (velocity dispersion) around the median. In this scatter, the original measurement uncertainties are not considered. See Table 3 for the positional parameters.

(12 members in Kerr et al. 2021 versus 132 members in this work).

5.1.6. Chamaeleon (Cham)

The well-known star-forming molecular clouds of Chamaeleon are seen through the same line-of-sight as the southern tip of LCC but lie clearly toward the back of LCC when seen in

3D (Fig. 12). We identify two clusters likely associated with the clouds with a total of 246 stellar sources in Chamaeleon 1 & 2, which have already been characterized with *Gaia* (e.g., Roccatagliata et al. 2018; Galli et al. 2021b; Kerr et al. 2021, see also Sect. 5.2.2).

In addition, toward the middle-eastern part of the traditional LCC borders, SigMA extracts another cluster that seems unrelated to the main body of LCC, which we name Centaurus-Far

Table 5. HIPPARCOS astrometry from [van Leeuwen \(2007\)](#) of bright stellar members in Sco-Cen.

HIP	Name	SigMA ^(a)	SpT	HIPPARCOS				Gaia DR3			
				<i>l</i>	<i>b</i>	ϖ	μ_{α}^*	μ_{δ}	ϖ	μ_{α}^*	μ_{δ}
				(deg)		(mas)	(mas yr ⁻¹)	(mas yr ⁻¹)	(mas)	(mas yr ⁻¹)	(mas yr ⁻¹)
80473	* rho Oph	1	B2V	353.69	17.69	9.03 ± 0.90	-5.53 ± 0.89	-21.74 ± 0.93	7.26 ± 0.13	-4.38 ± 0.19	-23.30 ± 0.16
79374	* nu Sco (Jabbah)	2	B2IV	354.61	22.70	6.88 ± 0.76	-7.65 ± 0.71	-23.71 ± 0.47	7.06 ± 0.22	-7.44 ± 0.25	-28.20 ± 0.16
78401	* del Sco (Dschubba)	3	B0.2IV	350.10	22.49	6.64 ± 0.89	-10.21 ± 1.01	-35.41 ± 0.71			
78820	* bet Sco (Acrab)	4	B0.5V	353.19	23.60	8.07 ± 0.78	-5.20 ± 0.92	-24.04 ± 0.64			
80112	* sig Sco (Alniyat)	5	B1III	351.31	17.00	4.68 ± 0.60	-10.60 ± 0.78	-16.28 ± 0.43			
79404	* c02 Sco	5	B2V	348.12	16.84	6.81 ± 0.16	-10.38 ± 0.18	-23.94 ± 0.14			
81266	* tau Sco	6	B0V	351.53	12.81	6.88 ± 0.53	-9.89 ± 0.61	-22.83 ± 0.55			
80763	* alf Sco (Antares)	6	M1Ib + B2.5V	351.95	15.06	5.89 ± 1.00	-12.11 ± 1.22	-23.30 ± 0.76			
78104	* rho Sco (Ikilil)	7	B2IV/V	344.63	18.27	6.91 ± 0.19	-15.68 ± 0.21	-24.88 ± 0.19			
78265	* pi Sco	9	B1V + B2V	347.21	20.23	5.57 ± 0.64	-11.42 ± 0.78	-26.83 ± 0.74			
81477	V* V1062 Sco	10	Ap Si	343.57	5.18	7.54 ± 0.61	-10.25 ± 0.69	-21.59 ± 0.45			
82545	* mu02 Sco (Piprima)	10	B2IV	346.20	3.86	6.88 ± 0.12	-11.09 ± 0.13	-23.32 ± 0.11			
82514	* mu01 Sco (Xamidimura)	11	B1.5IV + B	346.12	3.91	6.51 ± 0.91	-10.58 ± 0.87	-22.06 ± 0.74			
78533	V* LL Lup	13	Ap Si	339.64	11.41	8.61 ± 0.69	-12.65 ± 0.75	-24.14 ± 0.68			
78384	* eta Lup	14	B2.5IV	338.77	11.01	7.38 ± 0.18	-16.96 ± 0.19	-27.83 ± 0.19			
71865	* b Cen	15	B2.5V	325.90	20.10	9.62 ± 0.18	-29.92 ± 0.14	-30.68 ± 0.13			
76945	* psi02 Lup	15	B5V	338.48	16.08	8.97 ± 0.27	-21.37 ± 0.30	-29.98 ± 0.25			
75304	* phi02 Lup	15	B4V	333.84	16.75	6.28 ± 0.20	-18.24 ± 0.22	-20.72 ± 0.16			
71860	* alf Lup	17	B1.5III	321.61	11.44	7.02 ± 0.17	-20.94 ± 0.14	-23.67 ± 0.14			
74449	* e Lup	17	B3IV	327.83	11.43	6.47 ± 0.21	-22.01 ± 0.18	-21.75 ± 0.19			
76371	* d Lup	17	B3IVp	331.02	8.76	7.62 ± 0.43	-20.53 ± 0.33	-21.23 ± 0.31			
71536	* rho Lup	19	B5V	320.13	9.86	10.32 ± 0.16	-28.26 ± 0.12	-28.82 ± 0.13			
72800	V* V1019 Cen	20	B7II/III	327.93	19.11	6.63 ± 0.22	-20.48 ± 0.23	-19.20 ± 0.22			
67464	* nu Cen	20	B2IV	314.41	19.89	7.47 ± 0.17	-26.77 ± 0.12	-20.18 ± 0.08			
63945	* f Cen	21	B5V	305.47	14.34	8.36 ± 0.25	-29.85 ± 0.18	-15.17 ± 0.15			
60823	* sig Cen	21	B3V	299.10	12.47	7.92 ± 0.18	-32.36 ± 0.15	-12.51 ± 0.13			
60009	* zet Cru	22	B2.5V	299.32	-1.36	9.12 ± 0.45	-33.80 ± 0.48	-10.15 ± 0.43			
60718	* alf Cru (Acrux)	22	B0.5IV	300.13	-0.36	10.13 ± 0.50	-35.83 ± 0.47	-14.86 ± 0.43			
58484	* eps Cha	24	B9Vn	300.21	-15.62	9.02 ± 0.36	-40.34 ± 0.38	-8.30 ± 0.40			
42637	* eta Cha	24	B9IV	292.40	-21.65	10.53 ± 0.16	-28.89 ± 0.16	27.21 ± 0.16			
84970	* tet Oph	28	B2IV	0.46	6.55	7.48 ± 0.17	-7.37 ± 0.18	-23.94 ± 0.10			
86670	* kap Sco	31	B1.5III	351.04	-4.72	6.75 ± 0.17	-6.05 ± 0.21	-25.54 ± 0.13			
									10.17 ± 0.07	-29.43 ± 0.18	26.83 ± 0.12

Notes. The *Gaia* DR3 astrometry is given if available. Shown are mostly B-type stars that are either part of the SigMA-selected clusters or which are the name-givers of some clusters. This is not a complete list of B-stars in Sco-Cen. The HIPPARCOS astrometry can be compared to average parameters of the SigMA clusters as derived from *Gaia* astrometry in Tables 3–4, to evaluate possible correlations. The HIPPARCOS parallaxes should be treated with caution, since significant deviations to *Gaia* parallaxes are possible (on the order of about 60 pc when converted to distances), while proper motions show deviations on the order of about ± 2 mas yr⁻¹, when comparing sources that are both in HIPPARCOS and *Gaia* DR3 within about 500 pc from the Sun. ^(a)Column 3 gives the index of the SigMA cluster that is hosting the given star.

(containing 99 sources), since it lies about 60 pc to the back of 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. 5.2.2 and Table E.3). Consistent with [Kerr et al. \(2021\)](#), we count this cluster to the Cham subregion.

Due to their youth, position, and tangential velocities, we assume that the three Cham clusters and the two clouds are part of the Sco-Cen star formation event, but this must be confirmed by tracebacks of the young populations (see, e.g., [Großschedl et al. 2021](#)). Similar suggestions appear in [Lépine & Sartori \(2003\)](#) or [Sartori et al. \(2003\)](#).

5.1.7. Northeast clusters (NE)

We identify three extra clusters to the Galactic north and east of Sco-Cen, which we discuss separately in this section: L134/L183, Oph-southeast, and Oph-North-Far. We assigned these clusters to a separate region, which we call the Northeast clusters (NE), based on their location relative to Sco-Cen in Galactic coordinates, since they do not fit any other of the Sco-Cen subregions.

The cluster L134/L183 is a small, newly identified stellar group to the Galactic north of US (with 24 stellar members). This stellar group is likely associated with the small molecular clouds L134 and L183 (or MBM 36 and 37, [Magnani et al. 1985](#)), which 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 matches the cluster distance of about 114 pc. The presence of the young stellar group close by the clouds 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 “cloud-shine” phenomenon toward this cloud ([Steinacker et al. 2010, 2015](#)).

The cluster Ophiuchus Southeast (Oph-SE, 61 members) lies at a similar projected position as Pipe-North, while being at a farther distance, about 50 pc in the back (hence, we did not count it to the Pipe region). This stellar group was already selected by [Kerr et al. \(2021\)](#) as TLC 4 with 31 members (Sect. 5.2.2). Finally, the group Ophiuchus-North-Far (Oph-NF, 28 members) appears to be a newly identified stellar group, located at a similar distance as Oph-SE. This new group needs more investigations in the future, since the stability of the selected members, as determined by the SigMA algorithm, is generally very low (stability < 11%).

5.2. Comparison with previous work

In the following we compare the SigMA-selected stellar clusters with recent results from the literature (Table 6), including eight publications. The studies by [Damiani et al. \(2019\)](#), [Schmitt et al. \(2022\)](#), [Luhman \(2022\)](#), and [Žerjal et al. \(2023\)](#) 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). The first three of these studies select members within broad selection borders decided by hand, which we call in this paper visual selection methods. [Squicciarini et al. \(2021\)](#), [Miret-Roig et al. \(2022a\)](#), and [Briceño-Morales & Chanamé \(2023\)](#) focus only on the US region and extract clusters using a combination of *Gaia* astrometry and radial velocities. [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 unsuper-

vised machine learning approach, which is more similar to our work than the aforementioned studies.

The literature samples are cross-matched with the SigMA clusters using the *Gaia* DR3 `source_id`, as specified in Appendix A. We provide: an overview of the discussed literature samples in Table 6, giving the total number of sources of each literature sample; the total number of sources of SigMA Sco-Cen cluster members within the respective studied areas, the field of view (FOV); and the number of total matches. Finally, we list the fraction of sources that we recover or reject (or miss) when compared to the individual literature samples, and the fraction of new sources when comparing the matches to the SigMA sample.

5.2.1. Comparison with [Damiani et al. \(2019\)](#)

[Damiani et al. \(2019, hereafter DPP19\)](#) analyze Sco-Cen using *Gaia* DR2 data and a traditional approach, selecting by hand over-densities in velocity and position space, followed by selecting pre-main-sequence (PMS) stars from an HRD. Their FOV goes slightly beyond the traditional borders of the association (see Table 6). They discuss eight compact clusters, which are prominently peaked in projection and in velocity space (hence, easier to identify with visual selection methods); these are UCL-1, UCL-2, UCL-3, Lupus 3, LCC-1, US-far, US-near, and the well-studied IC 2602. Although SigMA easily detects IC 2602, we do not discuss this cluster since its age (~30 Myr) excludes it as a part of the recent Sco-Cen star formation event, as mentioned above, and it also has distinctly different tangential velocities compared to the bulk motion of Sco-Cen (see Fig. 6 in [DPP19](#)). [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 E.2).

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 with “N”). When cross-matching the [DPP19](#) *Gaia* DR2 sample with DR3 astrometry, we find that 201 stars (1.4%) are rejected when applying the distance criteria from [DPP19](#) ($d < 200$ pc), due to updated parallaxes in DR3. The majority of these sources have not been assigned to any group or belong to one of the diffuse populations. When now considering only the sources in the clustered and diffuse populations within 200 pc (and without IC2602), then there are 10,425 potential Sco-Cen members in [DPP19](#), or 10,421 when additionally applying the box and quality criteria from Sect. 2.

There are in total 9635 cross-matches between the SigMA clusters and [DPP19](#), while 9328 of these (89.5% out of 10,421) belong to either the clustered or diffuse populations (307 are not assigned, “N”). Of the 9328 cross-matches, 7609 belong to one of the four diffuse populations. Comparing this number to their total diffuse population (8689 within 200 pc), we find that about 88% are a match with SigMA clusters. The fact that the majority of sources from so-called diffuse populations are now in clustered populations (at least in the statistical sense) is interesting and calls for future investigations to better understand if physically meaningful “young diffuse stellar populations” actually exist within young stellar associations like Sco-Cen. In most cases, more than one [DPP19](#) group (both clustered or diffuse) fits one of our clusters, and vice versa (see Table E.2). In particular, their diffuse groups each contain subparts of about 10–20 of the SigMA clusters.

Table 6. Overview of the recent literature to which we compare the Sigma Sco-Cen clustering results in more detail.

Reference	Sect.	Data	Studied area	Number statistics					
				Nr. in Ref. ^(a)	SigMA ^(b)	Matches ^(c)	Recovered ^(d)	Rejected ^(d)	New ^(d)
Damiani et al. (2019) ^(e)	5.2.1	DR2	($l = 360^\circ$ to 280° , $b = 0^\circ$ to 30°) OR ($l = 315^\circ$ to 280° , $b = -10^\circ$ to 0°) FOV = 2750 deg^2 , $d < 200 \text{ pc}$	10 421 1732 clustered (17%) 8689 diffuse (83%)	11 796	9328 1719 7609	89.5% 99.2% 87.6%	10.5% 0.8% 12.4%	20.9%
Kerr et al. (2021) ^(f)	5.2.2	DR2	The whole TLC22 stellar group TLC22 without EOM 1–5 22 EOMs of TLC22 (without EOM 1–5)	7394 7138 3453	12 669 12 669 12 669	6270 6270 3447	84.8% 87.8% 99.8%	15.2% 12.2% 0.2%	50.5%
Schmitt et al. (2022) ^(g)	5.2.3	EDR3 and eROSITA	de Zeeuw et al. (1999) borders: US ($l = 343^\circ$ to 360° , $b = 10^\circ$ to 30°) OR UCL ($l = 312^\circ$ to 350° , $b = 0^\circ$ to 25°) OR LCC ($l = 285^\circ$ to 312° , $b = -10^\circ$ to 22°), FOV = 2050 deg^2 , $d \sim 60\text{--}200 \text{ pc}$	6150 ~69% vel-clustered ~26% vel-diffuse ~5% IC 2602	11 348	3385 3385 0 0	55.0% ~80% ~20%	45.0% ~20%	70.2%
Luhman (2022)	5.2.4	EDR3	$l = 2^\circ$ to 285° , $b = -12^\circ$ to 35° , FOV = 3252 deg^2 , $d \sim 90\text{--}250 \text{ pc}$	10 509	12 215	9838	93.6%	6.4%	19.5%
Žerjal et al. (2023)	5.2.5	DR2	$l = 40^\circ$ to 240° , $b = -60^\circ$ to 70° , FOV = $36\,400 \text{ deg}^2$, $d \sim 83\text{--}200 \text{ pc}$	8185	12 943	7671	93.7%	6.3%	40.7%
Squicciarini et al. (2021) ^(h) (only US, subsample with RVs)	5.2.6	EDR3	$\alpha = 236^\circ$ to 251° , $\delta = -29^\circ$ to -16° FOV = 195 deg^2 , $d \sim 125\text{--}175 \text{ pc}$	2745 1442 clustered (53%) 1303 diffuse (47%)	2717	2575 1435 1140	93.8% 99.5% 87.5%	6.2% 0.5% 12.5%	5.2%
Miret-Roig et al. (2022a) ⁽ⁱ⁾ (only US, subsample with RVs)	5.2.7	DR3	$\alpha = 235^\circ$ to 252° , $\delta = -30^\circ$ to -17° FOV = 221 deg^2 , $d \sim 80\text{--}200 \text{ pc}$	2810 2190 5D (78%) 670 6D (24%) 620 Rest (22%)	3089	2683 2145 667 538	95.5% 97.9% 99.6% 86.8%	4.5% 2.1% 0.4% 13.2%	13.1%
Briceño-Morales & Chanamé (2023) ^(j) (only US, subsample with RVs)	5.2.8	EDR3	$l = 343^\circ$ to 360° , $b = 10^\circ$ to 30° FOV = 340 deg^2 , $d < 200 \text{ pc}$	3004 995 clustered (33.1%) 2009 diffuse (66.9%)	3439	2720 989 1731	90.5% 99.4% 86.2%	9.5% 0.6% 13.8%	20.9%

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 13 103 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 given in the literature, then the numbers and matches with these are given below. ^(d)The fraction of recovered, rejected (or missed), and new sources are calculated by comparing the number of matched sources (Col. c) to the sample sizes from the literature (Col. a) or the Sigma sources in the same FOV (Col. b). ^(e)For [DPPI9](#) we only give the number of sources within their clustered or diffuse populations within $1000/\varpi_{\text{EDR3}} < 200 \text{ pc}$ after a cross-match with *Gaia* EDR3, and without IC 2602. ^(f)For [KRK21](#), we do not give the surveyed area since they extracted the clusters from all-sky data within 333 pc from the Sun. Still, we show a comparison to their whole TLC22 group (their main Sco-Cen group). We compare to 12 669 sources without Cham, Oph-SE, and Oph-NE since these are not contained within the TLC22 volume. Below we compare to TLC22, excluding their older EOM groups, and finally to the TLC22 members that are contained in one of the younger EOM groups (again without older EOM groups). ^(g)The X-ray-selected sources from [SCF22](#) include velocity-clustered and velocity-diffuse sources. The separation of these was applied by us by hand, guided by Fig. 7 in [SCF22](#). 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. ^(h)[SGB21](#) only studied the US region, finding sources in a more clustered mode and sources in a more diffuse mode, while the latter is simply the residuals of their clustering procedure. They study a subsample of sources with v_r information in the 6D phase space ($\sim 28\%$), which is not further discussed in this work. ⁽ⁱ⁾[MR22](#) only studied the US region. They study a subsample of sources with v_r information in the 6D phase space ($\sim 30\%$). The numbers of sources in the [MR22](#) bona fide 5D and bona fide 6D samples are listed separately, with “Rest” being sources in neither of the two. Note that the 6D sample is contained within the 5D sample. ^(j)[BMC23](#) only studied the US region. They separate their sample into one diffuse and eight clustered populations.

Focusing on the 1732 [DPP19](#) sources in compact clusters, there are 1719 matches with [SigMA](#) clusters (99%) within 200 pc. The better consensus considering their compact samples highlights the higher robustness of these samples (see also Table 6). For individual samples toward Upper-Sco, we find that their US-near and US-far cannot be assigned clearly to only one of the [SigMA](#) clusters (see Table E.2). US-near correlates best with ρ Oph/L1688 (containing fractions of δ Sco, ν Sco, Antares, and β Sco), and US-far with σ Sco (containing fractions of Antares, β Sco, δ Sco, and ρ Sco). In particular, Antares is distributed almost equally among these two clusters. The Antares group is partially occupying the same volume as ρ Sco and in particular ρ Oph/L1688 (see Sect. 5.1.1 and [Grasser et al. 2021](#)). This highlights the capability of [SigMA](#) to untangle young populations that share the same volume but have different space motions. The rest of the [DPP19](#) compact clusters correlate best with [SigMA](#) clusters as follows: UCL-1 with V1062 Sco and μ Sco, UCL-2 with UPK 606, UCL-3 with ϕ Lup, LCC-1 with Acrux, and Lup III with Lupus 1–4. Finally, about 9% of the [SigMA](#) cluster members correlate with unassigned sources in [DPP19](#) (N), within their FOV and our box criteria.

Concerning the different approaches, comparing the [DPP19](#) visual selection 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. Such an approach will deliver the most prominent clusters. However, somewhat less dense clusters cannot be identified easily when compared to unsupervised machine learning tools, such as [SigMA](#), and their method is less sensitive to possible spatial and kinematical structure in the Sco-Cen population. 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 select a larger number of candidates than a fine-tuned machine learning classifier with also a likely higher number of contaminants.

When focusing on a comparison of the total number of Sco-Cen members in [DPP19](#) stellar clusters (compact and diffuse within 200 pc, 10,421 sources) to the number of matched [SigMA](#) cluster members (9328 within 200 pc; see Table 6), we find that there are 1093 sources only in [DPP19](#), implying that we could be missing about 10% of possible members if all 10,421 sources were good members. We did not perform a detailed comparison but find that the 1093 sources also contain sources that seem to be older than the [SigMA](#) clusters when investigated in a CMD (as in Fig. 14); hence, the difference based on this comparison is likely lower than 10% (see also the comparison with [Luhman 2022](#)). As mentioned above, we expect [SigMA](#) to be missing possible candidates when compared with a method that selects broad regions in various 2D planes of the phase space, but also expect the [SigMA](#) sample to be less contaminated. Nevertheless, the [SigMA](#) sample contains in total 11,796 sources inside the [DPP19](#) FOV, implying that, in the end, we find about 20% more clustered Sco-Cen members. This could partially be caused by the different data sets, DR2 versus DR3, while the different methodologies likely cause more severe disagreements. A deeper analysis is needed, although not warranted in this paper.

5.2.2. 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 Appendix B.1) on *Gaia* DR2 parallaxes and proper motions on a preselected sample of PMS stars with ages $\lesssim 50$ Myr. They

identify 27 top-level clusters (TLCs), including Chameleon as TLC 21 and the Sco-Cen association as TLC 22. The latter was further broken down into another 27 subgroups based on the excess of mass (EOM) method, selecting the most persistent clusters in the clustering tree. Three of these EOM subgroups (EOM 12 Lupus; EOM 17 US; and EOM 27 LCC) were further broken down into leaves, which are nodes of the clustering tree.

TLC 22 covers the main Sco-Cen association, and TLC 21 the Chamaeleon region. Additionally, there were cross-matches with members of the group TLC 4, which is called Ophiuchus Southeast in [KRK21](#). These three TLC groups combined show a similar extent to our Sco-Cen extraction. [SigMA](#) finds in total a slightly lower number of groups toward Sco-Cen (37 in this work versus 45 in [KRK21](#)), while the TLC 22 subgroups in [KRK21](#) also include older or unrelated populations (e.g., β Pic, IC 2602, Platais 8, and EOM-2 & 5), which are not included in our final Sco-Cen sample, as outlined above. Consequently, only 39 of the [KRK21](#) groups toward Sco-Cen fall within the 37 selected [SigMA](#) clusters from this work.

In Table E.3, 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 the TLC 22 main Sco-Cen group (and with TLC 21, Cham; or TLC 4, Oph-SE), while some of our groups also distinctly correspond to EOM subgroups (or leaves). For about 40% of the [SigMA](#) groups, clear accordance with a single EOM group (or leaves group) is not possible due to overlaps with more than one [SigMA](#) group or due to no or only insignificant overlap (see also Table E.3).

Some differences between the [SigMA](#) and [KRK21](#) clustering results might arise from the different data input since we used *Gaia* DR3 and [KRK21](#) used DR2, while this would only create minor deviations. 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 Appendix B.1. 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 Lupus, US, and LCC regions (in TLC 22, EOM-12,17,27) since they found that there are substructures that have not been identified by the EOM method. Some leaf clusters match quite well with [SigMA](#) clusters; in contrast, the [SigMA](#) clusters are significantly richer and mostly more extended, and in some cases, they are differently separated (see details in Table E.3). 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 Appendix B.1 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 leaves in US (ρ Oph/L1688, ν Sco, δ Sco, β Sco) show good agreement with the [SigMA](#) US cluster separations, indicating the robustness of these clusters (see also Sect. 5.2.7).

When comparing all members in the TLC22 group (7394), which contains the main Sco-Cen association, with our [SigMA](#) extraction (12 669 members without Cham, Oph-NE, and Oph-SE) we find 6270 cross-matches in total (Table 6). Hence, 1124 ($\sim 15\%$ of TLC22) sources are only in TLC22, and 6399 are only in [SigMA](#) ($\sim 50\%$ of [SigMA](#)). We find that the [KRK21](#) TLC22

sample contains at least 256 sources from older stellar groups, which gets apparent from their Table 6 (EOM 1–5, including β Pic, IC 2602, and Platais 8), and 456 sources of TLC22 match with sources that are in older SigMA clusters. Combined, this leaves 6895 potentially younger TLC22 Sco-Cen members, and hence 625 possible extra sources ($\sim 8\%$ of TLC22). For these extra sources, a clear separation of the younger Sco-Cen stellar groups as discussed in this work, and the somewhat older groups is not straightforward, since about 50% of the sources in the TLC22 group have not been assigned to a separate subcluster (EOM or leaf). The somewhat older sources can also be estimated when investigating the CMD or the velocity space. In the CMD no clear separation of older or younger sources can be identified. In tangential velocity space, there are sources that have slightly deviating motions from expected Sco-Cen motions or which coincide with velocity spaces of the KRK21 older EOM groups or older SigMA groups. Taking all this into account, the fraction of young TLC22-only sources is likely below 8%.

The reason for these extra potential Sco-Cen members in the KRK21 TLC22 group is similar to the mentioned reasons above (e.g., in Sect. 5.2.1). The TLC22 group represents a cluster root, enveloping the whole Sco-Cen region and somewhat beyond, and no additional substructure was extracted (yet). In the 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. Focusing only on the TLC22 members that are in one of the 22 younger EOM subgroups, we find that we recover 99.8% of these sources as Sco-Cen members. Finally, the TLC22 group seems to be overall more incomplete compared to the SigMA Sco-Cen extraction, since we find in total more members ($\sim 42\%$, 12 669 versus 7394), and also somewhat different substructure. At the same time, the subclusters themselves are significantly richer compared to KRK21.

In conclusion, the comparison with KRK21 highlights the differences that can arise with different unsupervised machine learning tools. Compared to applications of HDBSCAN, we find that SigMA is able to extract similar substructure, however, with only one clustering step (no sub-steps like EOM or leaves are needed), while at the same time extracting significantly higher numbers of members per cluster. The modeling results in Sect. 4.2.2, where we compare the performance of DBSCAN, HDBSCAN, and SigMA, also suggest that SigMA outperforms HDBSCAN. However, a comparison of the KRK21 HDBSCAN results with the model HDBSCAN results cannot be done at face value. In Table 2 we list the performance results for the best-performing HDBSCAN model configuration, where the model parameters were optimized in a grid search, as described in Appendix B.8. The HDBSCAN configuration used by KRK21 is not identical to these model runs; hence, we cannot directly use the performance numbers in Table 2 for a comparison of the two Sco-Cen cluster catalogs (SigMA versus KRK21). Moreover, we did not combine multiple results (KRK21 combines results from EOM and leaf runs) to maximize the performance. Regardless of the difficulties to compare the performances, there is sufficient evidence to conclude that both the direct comparison of the two Sco-Cen cluster catalogs and also the modeling runs indicate that SigMA outperforms HDBSCAN in Sco-Cen-like environments.

5.2.3. Comparison with Schmitt et al. (2022)

Recently, Schmitt et al. (2022, hereafter SCF22) used eROSITA²⁰ (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 X-ray observed sources within the traditional borders (Blaauw 1964a; de Zeeuw et al. 1999), which are Sco-Cen candidate members. They include sources within a distance range of 60 to 200 pc (Table 6), restricted to low-mass stars ($G_{BP} - G_{RP} > 1$, following Pecaut & Mamajek 2013). The 6190 sources include 40 double Gaia sources that match with two different eROSITA sources. We only discuss the 6150 single Gaia sources.

Since X-ray emitting sources are expected to be young (e.g., Schmitt 1997; Neuhäuser 1997; Feigelson & Montmerle 1999; Bouvier et al. 2014), the sources detected by eROSITA in the direction of Sco-Cen, as discussed in SCF22, 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 appear to be more diffuse in velocity space²¹, calling into question search approaches relying on kinematic selections.

We cross-matched the 6150 SCF22 X-ray-selected sources with the SigMA selection in the same FOV (containing 11 348 SigMA sources; see Table 6). We find a total of 3385 cross-matches, while none of these belong to their velocity-diffuse population. This is expected since SigMA only selects clusters confined in position-velocity space, which naturally excludes any such velocity-diffuse sources. SCF22 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 the CMD. However, we see a relatively clear age separation between the velocity-clustered and velocity-diffuse populations. X-ray sources that occupy similar velocity spaces as the 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, when compared to PARSEC model isochrones. 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 coeval part, could be related to a relatively old 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 SCF22 velocity-diffuse young sources are unlikely to be part of Sco-Cen, but represent a mystery that needs to be solved. As SCF22 points out, the sensitivity of eROSITA will allow virtually all young Sco-Cen low-mass members to be detected in the near future. A combination of eROSITA future releases and Gaia data in Sco-Cen will be crucial to increase statistics and better understand the relation between observed X-ray luminosity with distance, age, stellar masses, and the origin of the velocity-diffuse population.

When concentrating on the velocity-coherent sample in SCF22 (without IC 2602), we find that there are about 20% in the whole SCF22 sample that could be additional Sco-Cen candidate members. These are only in SCF22 and have similar velocities as SigMA Sco-Cen members. When investigating these 20% in the CMD to check possible contamination from older stars (similar to Fig. 14), this fraction would reduce to about 10%. These extra potential members might result from the broad selection conditions in SCF22, based on all X-ray detected sources within the Blaauw borders in a distance range of 60 to 200 pc (Table 6). These broad conditions, which do not attempt to identify any

²⁰ Extended ROentgen Survey with an Imaging Telescope Array. A wide-field X-ray telescope on board the Russian-German Spectrum-Roentgen-Gamma (SRG) observatory.

²¹ We applied the separation of kinematically clustered and diffuse populations by hand in v_{α}/v_{δ} space, as indicated in Fig. 7 of SCF22.

underlying clustered structure, will naturally pick up more candidate members, although more false positives too, as discussed in Sects. 5.2.1 and 5.2.9, also because the X-ray detection is not a guarantee to only pick up the youngest sources below 20 Myr.

The mean completeness rate of SigMA in an environment with densely packed clusters as estimated in Sect. 4.2.2 (obtained with mock data), is about 81%, which could explain the missed 10% to 20% of SCF22 sources. On the other hand, there is no completeness estimate given in SCF22, while this sample is likely also incomplete, in particular regarding X-ray faint sources, since we find about 70% more sources in the same FOV. This is an indication that the X-ray luminosity of all young sources is not in general bright enough to be picked up by eROSITA, while future data releases might increase the numbers of observed X-ray Sco-Cen members.

5.2.4. Comparison with Luhman (2022)

Luhman (2022, hereafter L22A) recently investigated the Sco-Cen region using *Gaia* EDR3 data to identify 10 509 kinematic candidate members (see Table 6). L22A includes selections for US, UCL/LCC, V1062 Sco, Ophiuchus, and Lupus (the Southern parts of Sco-Cen are not discussed in L22A), and concentrates on established stellar groups in Sco-Cen to guide the 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 over-densities 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 9838 matches (93.6%), 671 L22A only sources (6.4%), and 2377 SigMA only sources within the L22A studied area (Table 6), where the SigMA sample contains 12 215 sources in total. A more detailed comparison of the SigMA clusters with the L22A subgroups²², which are generally larger scale groups, shows no clear correlation between single groups. Virtually each SigMA cluster has several matches with various L22A subgroups (and vice versa). When investigating the 671 L22A only sources, we find that about 50% of these sources do not show significant signs of being older than 20 Myr, and the majority of the sources do not show significant deviating motions from SigMA Sco-Cen cluster velocities. These extra L22A sources, or part of them, could be Sco-Cen members, meaning we might be missing up to about 6% 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 also discussed in Sect. 5.2.1. Nevertheless, the SigMA samples contain in total more Sco-Cen member candidates within the same FOV.

5.2.5. Comparison with Žerjal et al. (2023)

Žerjal et al. (2023, hereafter ZIC21) present another clustering result for the Sco-Cen association using CHRONOSTAR, a clustering tool developed by Crundall et al. (2019). This is a Bayesian tool to kinematically decompose stellar groups using the full 6D kinematic data, also performing a kinematic age determination. They identify eight distinct kinematic components containing in total 9556 sources²³. The 9556 stellar members are both within

dense and also diffuse stellar groups. They also include two known clusters that we excluded from the final Sco-Cen sample, which are IC 2602 and Platais 8 (H and I in ZIC23). Without these clusters, their sample contains 8185 stars, of which 7671 (94%) match with the SigMA groups.

The groups are C-US, E-US-multi²⁴, D-UCL-V1062-Sco, F-UCL-V1062-Sco, G-UCL-East, T-UCL-West, A-LCC-North, and U-LCC-South. Hence, with their method, they are splitting US, UCL, LCC, and also V1062 Sco, each into two parts. We list all matches of SigMA clusters with ZIC23 in Table E.2. Generally, it can be seen that there is significant mixing of various groups in both the ZIC23 and the SigMA groups. In particular, the ZIC23 groups encompass larger areas, often containing several or up to 20 of the SigMA groups. Concerning the groups D and F, we find that both match with V1062-Sco and μ Sco, while D has a slightly higher correlation with V1062-Sco and F with μ Sco.

Within their FOV also other groups exist (see Fig. 8 in ZIC23), like CrA or the Cham cloud regions, while they were not selected as kinematic members of Sco-Cen. We speculate that their initial sample by Gagné et al. (2018a), which includes sources from US, UCL, and LCC, limits their ability to select these additional groups. Their low signal-to-noise ratio, positional distance, and slightly deviating motions from bulk Sco-Cen sources²⁵ likely prevented their classification as Sco-Cen groups.

Their method fits a mixture of Gaussians to data. Instead of allowing arbitrary covariance matrices, CHRONOSTAR constrains 6D Gaussian distributions in XYZUVW to the following form. Present-day observations are assumed to follow a ballistically evolved Gaussian in Galactic potential. The free-fitting parameters are a cluster's mean birth position in phase space, birth positional and kinematic variance (the covariance matrix is assumed uncorrelated), and age. The fitting is done via a modified EM algorithm where the number of components is determined via the BIC. ZIC23 state that they see evidence for substructure in several groups. Thus, it has to be investigated if the groups found in ZIC23 will eventually break into subgroups. When compared to SigMA we can already see that the individual ZIC23 groups generally contain more than one SigMA cluster (see Table E.2).

5.2.6. Comparison with Squicciarini et al. (2021)

Squicciarini et al. (2021, hereafter SGB21) studied 2745 potential US members (see Table 6) by selecting subgroups solely based on kinematics. They cluster the US region into eight groups that they call the clustered population (1442 stars), and into one older diffuse population (1303), which is, however, differently defined than the diffuse or velocity-diffuse populations in DPP19 or SCF22.

When comparing the SGB21 selection to the SigMA clusters, we find that there are 2575 cross-matches (~94%) in total out of the 2745 sources in SGB21. Hence, we miss 170 SGB21 US candidate members, while 13 sources are lost due to our box and quality criteria (Sect. 2). Focusing on the SGB21 candidate members in the clustered populations (1442), there are only seven sources that are only in SGB21 and not in SigMA, while we miss 163 of the SGB21 diffuse members, which are overall more uncertain members. In total, we find almost a factor of two more clustered sources with SigMA in the same FOV as SGB21, 2717 versus 1442 (see Table 6). The 2575 sources match with nine

²² Subgroups are given in Table 1 of L22A in columns “kin” and “pos.”

²³ 18 sources are outside of our box and quality criteria, leaving 9538 ZIC23 sources. Moreover, there are 25 sources that are outside of the ZIC23 parallax range due to updates from DR2 to DR3.

²⁴ ZIC23 define group E as a complex, multi-population component.

²⁵ Possibly caused by internal feedback mechanisms in the history of Sco-Cen (e.g., Zucker et al. 2022).

of the SigMA clusters, while only seven SigMA clusters have a significant number of matches.

We list the cross-matches of SigMA with SGB21 in Table E.4. We highlight more significant matches here. Groups 1, 2, 3, and 4 match best with ρ Oph/L1688, ν Sco, δ Sco, and β Sco, respectively, while Group 6 also has significant matches with β Sco. Group 5 matches best with σ Sco, while the majority of σ Sco is in the SGB21 diffuse population. Group 7 and Group 8 match best with Antares, while the majority of Antares is also in their diffuse population. Generally, the Antares group seems to split up into more than one cluster, also in other previous work. The two groups US-foreground and ϕ Lup have only a few matches with the diffuse population. The SGB21 diffuse population is largely contained within the SigMA groups σ Sco, Antares, ρ Sco, and δ Sco, with some diffuse members distributed among each mentioned group (see Table E.4). This indicates that the diffuse population is maybe not a separate older group but it contains stars that were not clustered by the methodology in SGB21, while they are clustered in SigMA.

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 semiautomated 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, 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 the SGB21 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 cells²⁶ that 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 multidimensional relationships between data axes. These rather different approaches to extracting clusters in US make it clear that the results cannot be compared at face value, while fractions of the most robust clusters (ρ Oph/L1688, ν Sco, δ Sco, β Sco, σ Sco, and Antares) have been identified by either method.

5.2.7. Comparison with Miret-Roig et al. (2022b)

Miret-Roig et al. (2022a, hereafter MR22) recently applied a Gaussian mixture model (GMM) to *Gaia* DR3 data of the US region. They include radial velocities, when available, to select the stellar groups. They identify seven stellar groups within a FOV of about 220 deg² that is centered on US, containing 2810 sources (see also Table 6). A cross-match with the SigMA members gives 2683 matches; hence, we are missing 127 sources, of which 23 are outside of our box and quality criteria (Sect. 2). This leaves 104 potential members that the SigMA

algorithm did not select. Within the same FOV SigMA contains 3089 Sco-Cen members.

When comparing the SigMA groups to the individual seven MR22 groups in more detail, we find largely good agreement, especially for ρ Oph, ν Sco, δ Sco, and β Sco (see also Table A.1 in MR22 and Table E.4). The SigMA Antares and σ Sco groups seem to be mixed in MR22, with significant fractions in both the MR22 α Sco and σ Sco. Finally, the MR22 π Sco group, which lies largely in the foreground of the other US groups, coincides largely with two of the SigMA groups: US-foreground (which we identified as a foreground population to the traditional US) and ρ Sco. The latter is overlapping in space with Antares and ρ Oph. The larger volume investigated with SigMA allows a more complete sample of the US-foreground to be selected since the MR22 FOV cuts off the outer edges of what they call π Sco.

Since MR22 also use radial velocity information for a subsample (~30%), the classification of these 6D-selected sources might be more robust compared to the 5.5D selection as used by SigMA. Investigating the bona fide 6D-selected members in MR22, we still find some mixing of SigMA clusters within MR22 clusters (and vice versa), especially concerning Antares and σ Sco. These differences need more investigations in the future, and dedicated cluster studies are called for (e.g., with Uncover, Ratzenböck et al. 2020).

5.2.8. Comparison with Briceño-Morales & Chanamé (2023)

Briceño-Morales & Chanamé (2023, hereafter BMC23) present another clustering study on US using *Gaia* EDR3 data. They obtain a clustering solution by first combining the convergent point method (Perryman et al. 1998) with a Gaussian Mixture fit (Pedregosa et al. 2011) to identify kinematic groups with a Bayesian approach. Second, they use OPTICS (Ankerst et al. 1999) to identify spatial substructure in XYZ. This is complemented with age estimates based on *Gaia* photometry. Their astrometrically clean sample obtained in the first step contains 3004 sources (USco kinematic group, KG). In the same FOV, we find 3439 clustered Sco-Cen members (see Table 6).

The USco KG is first broken down into three KGs in a “corrected tangential velocity space”²⁷: ρ Oph KG, USco Young KG, and USco Old KG. This suggests that USco comprises three main kinematic components, with ρ Oph KG mainly correlating with the traditional ρ Oph region, and USco Young KG with the traditional USco region. They argue that their USco Old KG is a new yet unstudied population while pointing out that large fractions of this KG are likely interlopers from UCL. Next, the USco KG is independently substructured with OPTICS in the XYZ space, delivering eight spatial clusters and one diffuse population. The latter are sources that OPTICS did not assign to a cluster. In fact, the majority of sources in the USco KG are contained in this diffuse population (~67%).

Using OPTICS, it can be challenging to identify noise points because there is no explicit noise cluster or appropriate noise threshold. Instead, the noise specification is prone to 1D projection effects of the ordering algorithm (e.g., loss of high-dimensional structure, small perturbations in data point positions produce different orderings), which eventually produces the reachability plot (*R*-plot) from which clusters and noise are determined. In particular, the presence of noise can lead to the creation of many small clusters in the reachability plot, making it challenging to identify clusters.

²⁶ As far as we know, scaling between sky coordinates and tangential velocities was not considered.

²⁷ They use the velocity offset of v_t, v_b relative to expected tangential velocities, derived from sources with known v_r .

When cross-matching the [BMC23](#) and [SigMA](#) samples, we find 2720 sources in common, 90.5% out of the 3004 [BMC23](#) sources or about 79% out of the 3439 [SigMA](#) sources in the same FOV. Hence, we find about 21% additional candidate Sco-Cen members compared to [BMC23](#) (see also Table 6). A detailed overview of the matched clusters is given in Table E.4. We summarize the best matches as follows.

There are five [SigMA](#) clusters that have relatively clear matches with [BMC23](#) spatial clusters, which are ρ Oph, β Sco, ν Sco, σ Sco (named α Sco in [BMC23](#)), US-foreground (named π Sco in [BMC23](#)), and η Lup (named UCL in [BMC23](#)). However, all of these clusters have more or less significant fractions contained in their diffuse population (Table E.4). Interestingly, our δ Sco is equally contained within the [BMC23](#) δ Sco and ω Sco, while the largest part of our δ Sco is in their diffuse population. They also compare these two clusters to [KRK21](#), where they find that both, δ Sco and ω Sco, are contained within EOM17-H (which matches with our δ Sco). In this case, we find that [SigMA](#) and [KRK21](#) agree while [BMC23](#) find further substructure in δ Sco. At this stage, we do not know which result is more physical, and future work is needed to understand this conundrum.

The majority of [SigMA](#)'s Antares, ρ Sco, and Libra-South are contained within their diffuse population; hence, these clusters have not been identified as individual clusterings by [BMC23](#). As mentioned above, their α Sco seems to match best with our σ Sco cluster (not with Antares), while they mention that the two stars *alf Sco and *sig Sco are likely members of this cluster. This connection is of interest since also other works seem to mix the [SigMA](#)-Antares and [SigMA- \$\sigma\$ Sco clusters, which needs further investigations in the future.](#)

In their conclusions, [BMC23](#) find an age-velocity correlation, suggesting that older clusters have increasing deviating tangential motions compared to ρ Oph (which they use as a starting point) as a function of age. Moreover, they find a correlation between cluster density and cluster age, suggesting that these substructures expand at a measurable rate. Finally, they argue that four potential supernovae happened in the US region based on literature information (e.g., [Breitschwerdt et al. 2016](#); [Neuhäuser et al. 2020](#); [Forbes et al. 2021](#)), cluster mass estimates, and on certain “voids” that are identified in the 3D distribution of their cluster extraction. Notably, we do not see such voids in the distribution of the [SigMA](#) clusters.

5.2.9. Concluding remarks on the comparison with the literature

In Table 6, we summarize the comparisons with the literature, stating the overlap size and the percentages of recovered, rejected, and new sources when comparing our sample to the literature samples within the same FOV used in respective publications. The rejected sources can also be seen as missing since they are still potential candidate members of Sco-Cen. We find that we miss sources on the order of a few percent up to about 45%, while this depends on the subsamples we compare to. For instance, some listed publications include compact clusterings and also diffuse populations, which are generally differently defined (e.g., velocity diffuse, remaining sources after clustering steps, etc.). Compared to the more compact clusters, we only miss sources on the order of <1% to a few percent. This underlines the robustness of the more compact clusterings, while the so-called diffuse populations need to be further investigated, particularly in light of their physical nature. The velocity-clustered X-ray observed sources, discussed in [SCF22](#), contain the largest

fraction of missed sources. However, as mentioned in Sect. 5.2.3, some of these members seem to have older ages in the CMD, leaving only about 10% of potentially missed young Sco-Cen members in [SCF22](#).

Compared to other methods described in the literature, [SigMA](#) consistently identifies more Sco-Cen candidate members, surpassing other studies by 5–70%. We considered various estimates of contamination, including 4–10% determined by older sources in the CMD (Appendix D.1), $5.3 \pm 3.1\%$ determined by [SigMA](#)'s internal approximation (Sect. 3.5.4), and $8.2 \pm 4.1\%$ field star contamination determined by Sco-Cen-like cluster simulations (Sect. 4.2.2). Based on these estimates, the true field star contamination rate likely lies between 2 and 12%. Given this contamination estimate and the rate of yet undetected members in comparison to prior work, we expect [SigMA](#) has contributed a significant fraction of new, real member stars throughout Sco-Cen (only exceptions are comparable US samples from [SGB21](#) and possibly [MR22](#)).

The visual selection methods (e.g., [Damiani et al. 2019](#); [Luhman 2022](#)) contain about 6–10% candidate members not detected by [SigMA](#) (Table 6). This is mainly due to these methods' “select-by-eye” approach in projected subspaces of the multidimensional phase space to identify Sco-Cen candidates. However, unsupervised machine learning methods find more spatial and kinematical substructures in the Sco-Cen population and arguably produce samples with lower contamination levels compared to visual selection methods. More importantly, the [SigMA](#) method reveals a more complex velocity structure across the entire Sco-Cen, critical for a physical description of the formation process of OB associations such as Sco-Cen. When comparing [SigMA](#) to other unsupervised methods, which studied the whole Sco-Cen area (in particular [Kerr et al. 2021](#)), we find a compatible substructure. In contrast, the [SigMA](#) clusters are generally richer in members. The additional sources cannot be explained with our contamination estimate (about 2–12%) since this lies below the fraction of new sources when compared to [Kerr et al. \(2021\)](#); about 50% new sources).

Focusing on the US region, we generally find good agreement between cluster selections from [SGB21](#), [KRK21](#), [MR22](#), [BMC23](#), and [SigMA](#). Not surprisingly, the denser clusters in US (ρ Oph, ν Sco, δ Sco, and β Sco) have been recovered well by the different approaches. The Antares, σ Sco, ρ Sco, and US-foreground clusters are slightly more dispersed and have less clear matches across the methods (e.g., merged or distributed differently in different samples). As revealed by *Gaia* data, the newly identified velocity substructure in the US region is relevant to understanding the star formation processes at play in Sco-Cen and will continue to be an obvious target for future studies and surveys. Adding radial velocity information will be critical to characterize these clusters further, as already presented, for example, by [Miret-Roig et al. \(2022a\)](#). Considering the five mentioned studies, the somewhat different clustering results for the US region call for a reanalysis of this important young region to better understand its star formation history.

When considering the cross-contamination between clusters in general, we find that the clustered substructure within Sco-Cen as extracted with [SigMA](#) is generally differing to some degree from other clustering results. The deviations have a different extent depending on which literature sample we compare to (see Tables E.2–E.4). There are some clusters, in particular in USco, where several studies agree, highlighting the robustness of these clusters. However, we rarely find a perfect match, and the final cluster membership of an individual source is often highly uncertain, mainly if located at the outskirts of

a cluster's center in a tightly packed environment surrounded by other clusters. Such behavior is expected since we find in Sect. 4.2.2 that cross-contamination leads to a per-cluster completeness estimate of about 76% of recovered sources when testing on mock data (while this fraction also considers sources lost to the field). Nevertheless, the substructure of Sco-Cen as identified with SigMA is a good starting point for future studies, supported by the narrow CMD sequences per cluster (see the follow-up study, Ratzenböck et al. 2023).

In conclusion, SigMA finds significant numbers of sources not present in other samples (often more than 10% new sources, Table 6) with a contamination fraction of about 2–12%, while missing candidates when compared with visual selection methods (on the order of 10%). More work is needed to understand the sources SigMA misses. A possible way forward toward a most complete sample of Sco-Cen members is to use SigMA cluster members and 3D velocities (by including radial velocities to select the most robust members) as training sets to the Uncover method, a validated bagging classifier of one-class support vector machines (see the application in Ratzenböck et al. 2020 to the Meingast-1 stellar stream, Meingast et al. 2019). 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 dynamic state of the Sco-Cen complex.

6. Summary

In this paper we present SigMA, a method that explores the topological properties of a density field to define significant substructure. To test and validate SigMA, we applied it to *Gaia* DR3 data of the nearest OB association to Earth, Sco-Cen. The main results of this work can be summarized as follows:

1. SigMA is a novel clustering method that interprets density peaks separated by density dips as significant clusters. Using a graph-based approach, the technique detects peaks and dips directly in the multidimensional phase space.
2. SigMA is fine-tuned to large-scale surveys in astrophysics. In this context, 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 appropriately incorporate 5D astrometric uncertainties alongside radial velocity uncertainties, does not require any photometric pre-filtering of stellar populations, and scales to millions of points.
3. SigMA is capable of finding clusters in *Gaia* DR3 data, reaching stellar volume densities as low as $0.01 \text{ sources pc}^{-3}$ and tangential velocity differences between clusters of about 0.5 km s^{-1} .
4. We validated SigMA on two simulated data sets and highlight its merits in relation to established clustering techniques. Our comparison shows that SigMA can significantly outperform competing methods, especially in environments where clusters are densely packed, such as the Sco-Cen OB association. In these dense cluster situations, our simulations show that SigMA has a mean contamination rate of $23.7 \pm 13.1\%$ and a mean completeness rate of $76.2 \pm 15.2\%$. Considering only the field star influence on contamination and completeness (i.e., ignoring cross-contamination from neighboring clusters), these scores improve to $8.2 \pm 4.1\%$ and $81.4 \pm 2.0\%$.
5. SigMA identifies more than 13 000 Sco-Cen members located in 37 individual clusters of co-spatial and co-moving young stars. The CMD for each cluster shows a well-defined sequence. Because SigMA is not aware of a star's brightness or color, the well-defined sequences constitute a validation test for the ability of SigMA to extract coeval populations. A large

fraction of clusters are seen toward well-known Sco-Cen massive stars, too bright to be in *Gaia* DR3, and we (tentatively) associated respective clusters with them. Because SigMA is not aware of these massive stars, their association with SigMA clusters also constitutes a validation test for SigMA.

6. When comparing the 37 SigMA stellar populations in Sco-Cen to previous results from the literature, we find mostly agreement; however, several discrepancies exist. When compared to visual selection methods used recently on *Gaia* data of Sco-Cen, we find that we might be missing roughly 10% of candidates, while at the same time finding a higher total number of stellar members. Unsupervised methods such as SigMA find more spatial and kinematical substructure for the same data set and produce samples with lower contamination levels.

In the future, in particular when combined with 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 promises²⁸ the unveiling of detailed cluster distributions such as the one presented here but for all the near 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 thank the anonymous referee for their detailed and very helpful comments. We thank Núria Miret-Roig, Maruša Žerjal, Vito Squicciarini, and J. H. M. M. Schmitt for providing their data ahead of final publication. S. Ratzenböck acknowledges funding by the Austrian Research Promotion Agency (FFG, <https://www.ffg.at/>) under project number F0999892674. Additionally, S. Ratzenböck extends his gratitude to the research network Data Science at Uni Vienna and Petra Schönfelder for excellent administrative support. J. Großschedl acknowledges funding by the Austrian Research Promotion Agency (FFG) under project number 873708. 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 2013, 2018), NumPy (van der Walt et al. 2011), Matplotlib (Hunter 2007), Galpy (Bovy 2015), and Plotly (Plotly 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 Trans. Autom. Control*, **19**, 716
 Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. 1999, *SIGMOD Rec.*, **28**, 49
 Ashok Kumar, G. 2020, *Int. J. Sci. Technol. Res.*, **9**, 6
 Astropy Collaboration (Robitaille, T. P., et al.) 2013, *A&A*, **558**, A33
 Astropy Collaboration (Price-Whelan, A. M., et al.) 2018, *AJ*, **156**, 123
 Azzalini, A., & Torelli, N. 2007, *Stat. Comput.*, **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
 Baraffe, I., Homeier, D., Allard, F., & Chabrier, G. 2015, *A&A*, **577**, A42
 Beccari, G., Boffin, H. M. J., & Jerabkova, T. 2020, *MNRAS*, **491**, 2205
 Benjamini, Y., & Hochberg, Y. 1995, *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289
 Bentley, J. L. 1975, *Commun. ACM*, **18**, 509
 Biau, G., Chazal, F., Cohen-Steiner, D., Devroye, L., & Rodríguez, C. 2011, *Electron. J. Stat.*, **5**, 204
 Blaauw, A. 1946, *Publ. Kapteyn Astron. Lab. Groningen*, **52**, 1
 Blaauw, A. 1952, *Bull. Astron. Inst. Netherlands*, **11**, 414

²⁸ See Sect. 4.2 for a brief discussion on quality filters and future use.

- Blaauw, A. 1964a, *ARA&A*, **2**, 213
- Blaauw, A. 1964b, in *The Galaxy and the Magellanic Clouds*, ed. F. J. Kerr 20, 50
- Boch, T., & Fernique, P. 2014, in *Astronomical Data Analysis Software and Systems XXIII*, eds. N. Manset, & P. Forshay, *ASP Conf. Ser.*, **485**, 277
- Bonferroni, C. 1936, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali 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*, eds. 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
- Breitschwerdt, D., Feige, J., Schulreich, M. M., et al. 2016, *Nature*, **532**, 73
- Bressan, A., Marigo, P., Girardi, L., et al. 2012, *MNRAS*, **427**, 127
- Briceno-Morales, G., & Chanamé, J. 2023, *MNRAS*, **522**, 1288
- Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. 2010, in *2010 20th International Conference on Pattern Recognition*, 3121
- Brooke, T. Y., Huard, T. L., Bourke, T. L., et al. 2007, *ApJ*, **655**, 364
- Burman, P., & Nolan, D. 1992, *J. Multivariate Anal.*, **40**, 132
- Burman, P., & Polonik, W. 2009, *J. Multivariate Anal.*, **100**, 1198
- Burrows, A., Hubbard, W. B., Lunine, J. I., & Liebert, J. 2001, *Rev. Modern Phys.*, **73**, 719
- Calinski, T., & Harabasz, J. 1974, *Commun. Stat.*, **3**, 1
- Campello, R. J., Moulavi, D., & Sander, J. 2013, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (Springer), 160
- Cantat-Gaudin, T., & Anders, F. 2020, *A&A*, **633**, A99
- Cantat-Gaudin, T., Jordi, C., Vallenari, A., et al. 2018a, *A&A*, **618**, A93
- Cantat-Gaudin, T., Vallenari, A., Sordo, R., et al. 2018b, *A&A*, **615**, A49
- Cantat-Gaudin, T., Mapelli, M., Balaguer-Núñez, L., et al. 2019a, *A&A*, **621**, A115
- Cantat-Gaudin, T., Krone-Martins, A., Sedaghat, N., et al. 2019b, *A&A*, **624**, A126
- Castro-Ginard, A., Jordi, C., Luri, X., et al. 2018, *A&A*, **618**, A59
- Castro-Ginard, A., Jordi, C., Luri, X., Cantat-Gaudin, T., & Balaguer-Núñez, L. 2019, *A&A*, **627**, A35
- Castro-Ginard, A., Jordi, C., Luri, X., et al. 2020, *A&A*, **635**, A45
- Castro-Ginard, A., Jordi, C., Luri, X., et al. 2022, *A&A*, **661**, A118
- Celeux, G., Forbes, F., Robert, C. P., & Titterton, D. M. 2006, *Bayesian Anal.*, **1**, 651
- Celeux, G., Frühwirth-Schnatter, S., & Robert, C. P. 2019, *Handbook of Mixture Analysis*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods (CRC Press), 117
- Chaudhuri, K., & Dasgupta, S. 2010, in *Advances in Neural Information Processing Systems*, eds. J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Curran Associates, Inc.), 23
- Chaudhuri, K., Dasgupta, S., Kpotufe, S., & von Luxburg, U. 2014, *IEEE Trans. Inf. Theor.*, **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
- Cheng, Y. 1995, *IEEE Trans. Pattern Anal. Mach. Intell.*, **17**, 790
- Chen, Y., Girardi, L., Bressan, A., et al. 2014, *MNRAS*, **444**, 2525
- Chen, Y., Bressan, A., Girardi, L., et al. 2015, *MNRAS*, **452**, 1068
- Chronis, P., Athanasiou, S., & Skiadopoulos, S. 2019, in *2019 IEEE International Conference on Data Mining (ICDM)*, 91
- Comaniciu, D., & Meer, P. 2002, *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**, 603
- Cornuéjols, A., Wemmert, C., Gańczarski, P., & Bennani, Y. 2018, *Inf. Fusion*, **39**, 81
- Coronado, J., Fürnkranz, V., & Rix, H.-W. 2022, *ApJ*, **928**, 70
- Correa, C., & Lindstrom, P. 2011, *IEEE Trans. Visual. Comput. Graphics*, **17**, 1852
- Crundall, T. D., Ireland, M. J., Krumholz, M. R., et al. 2019, *MNRAS*, **489**, 3625
- 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*, eds. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Curran Associates, Inc.), 27
- 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
- Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977, *J. R. Stat. Soc. Series B Stat. Methodol.*, **39**, 1
- de Zeeuw, P. T., Hoogerwerf, R., de Bruijne, J. H. J., Brown, A. G. A., & Blaauw, A. 1999, *AJ*, **117**, 354
- 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
- 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
- Evans, D. W., Riello, M., De Angeli, F., et al. 2018, *A&A*, **616**, A4
- Feigelson, E. D., & Montmerle, T. 1999, *ARA&A*, **37**, 363
- Feng, Y., & Hamerly, G. 2007, in *Advances in Neural Information Processing Systems*, eds. B. Schölkopf, J. Platt, & T. Hoffman, 19 (MIT Press)
- Fernández, D., Figueras, F., & Torra, J. 2008, *A&A*, **480**, 735
- Fisher, R. A. 1934, *Breakthroughs in Statistics* (Springer), 66
- Forbes, J. C., Alves, J., & Lin, D. N. C. 2021, *Nat. Astron.*, **5**, 1009
- Francis, C., & Anderson, E. 2009, *New A*, **14**, 615
- Freytag, B., Allard, F., Ludwig, H. G., Homeier, D., & Steffen, M. 2010, *A&A*, **513**, A19
- Freytag, B., Steffen, M., Ludwig, H. G., et al. 2012, *J. Comput. Phys.*, **231**, 919
- Frühwirth-Schnatter, S., Celeux, G., & Robert, C. 2019, in *Handbook of Mixture Analysis*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods (CRC Press)
- Fürnkranz, V., Meingast, S., & Alves, J. 2019, *A&A*, **624**, L11
- Gabriel, K. R., & Sokal, R. R. 1969, *System. Biol.*, **18**, 259
- Gagné, J., & Faherty, J. K. 2018, *ApJ*, **862**, 138
- Gagné, J., Lafrenière, D., Doyon, R., Malo, L., & Artigau, É. 2014, *ApJ*, **783**, 121
- Gagné, J., Mamajek, E. E., Malo, L., et al. 2018a, *ApJ*, **856**, 23
- Gagné, J., Roy-Loubier, O., Faherty, J. K., Doyon, R., & Malo, L. 2018b, *ApJ*, **860**, 43
- Gagné, J., Faherty, J. K., & Mamajek, E. E. 2018c, *ApJ*, **865**, 136
- Gaia Collaboration (Brown, A. G. A., et al.) 2016, *A&A*, **595**, A2
- 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
- Gaia Collaboration (Vallenari, A., et al.) 2023a, *A&A*, **674**, A1
- Gaia Collaboration (Arenou, F., et al.) 2023b, *A&A*, **674**, A34
- Galli, P. A. B., Joncour, I., & Moraux, E. 2018, *MNRAS*, **477**, L50
- Galli, P. A. B., Bouy, H., Olivares, J., et al. 2020a, *A&A*, **643**, A148
- Galli, P. A. B., Bouy, H., Olivares, J., et al. 2020b, *A&A*, **634**, A98
- Galli, P. A. B., Bouy, H., Olivares, J., et al. 2021a, *A&A*, **654**, A122
- Galli, P. A. B., Bouy, H., Olivares, J., et al. 2021b, *A&A*, **646**, A46
- Ghrist, R. 2008, *Bull. Am. Math. Soc.*, **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
- Großschädl, 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*, eds. S. Thrun, L. Saul, & B. Schölkopf, 16 (MIT Press)
- Hartigan, J. A. 1975, *Clustering Algorithms*, 99th edn. (USA: John Wiley & Sons, Inc.)
- Hartigan, J. A., & Hartigan, P. M. 1985, *Ann. Stat.*, **13**, 70
- He, Z., Wang, K., Luo, Y., et al. 2022, *ApJS*, **262**, 7
- Hogg, D. W., Bovy, J., & Lang, D. 2010, *ArXiv e-prints* [arXiv:1008.4686]
- Hu, X., & Xu, L. 2003, in *Intelligent Data Engineering and Automated Learning*, eds. J. Liu, Y. M. Cheung, & H. Yin (Berlin, Heidelberg: Springer), 195
- Hubert, L., & Arabie, P. 1985, *J. Classification*, **2**, 193
- Hunt, E. L., & Reffert, S. 2021, *A&A*, **646**, A104
- Hunt, E. L., & Reffert, S. 2023, *A&A*, **673**, A114
- Hunter, J. D. 2007, *Comput. Sci. Eng.*, **9**, 90
- Jain, A. K., Murty, M. N., & Flynn, P. J. 1999, *ACM Comput. Surveys*, **31**, 264
- Jaromczyk, J., & Toussaint, G. 1992, *Proc. 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
- Johnson, S. C. 1967, *Psychometrika*, **32**, 241
- Kalogeratos, A., & Likas, A. 2012, in *Advances in Neural Information Processing Systems*, eds. F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Curran Associates, Inc.), 25
- Kamdar, H., Conroy, C., Ting, Y.-S., & El-Badry, K. 2021, *ApJ*, **922**, 49
- Kapteyn, J. C. 1914, *ApJ*, **40**, 43
- Katz, D., Sartoretti, P., Guerrier, A., et al. 2023, *A&A*, **674**, A5
- Kerr, F. J., & Lynden-Bell, D. 1986, *MNRAS*, **221**, 1023
- Kerr, R. M. P., Rizzuto, A. C., Kraus, A. L., & Offner, S. S. R. 2021, *ApJ*, **917**, 23
- Kervella, P., Arenou, F., & Thévenin, F. 2022, *A&A*, **657**, A7
- Kharchenko, N. V., Piskunov, A. E., Schilbach, E., Röser, S., & Scholz, R. D. 2013, *A&A*, **558**, A53

- Kirkpatrick, D. G., & Radke, J. D. 1985, in *Machine Intelligence and Pattern Recognition*, ed. G. T. Toussaint (North-Holland), [Comput. Geom.](#), **2**, 217
- Koontz W. L. G., Narendra P. M., & Fukunaga K. 1976, *IEEE Trans. Comput.*, **25**, 936
- Kounkel, M., & Covey, K. 2019, [AJ](#), **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 Machine Learning, ICML'11* (Madison, WI, USA: Omnipress), 225
- Krone-Martins, A., & Moitinho, A. 2014, [A&A](#), **561**, A57
- Kuhn, M. A., & Feigelson, E. D. 2019, *Handbook of Mixture Analysis, Chapman & Hall/CRC Handbooks of Modern Statistical Methods* (CRC Press), 463
- Kushniruk, I., Schirmer, T., & Bensby, T. 2017, [A&A](#), **608**, A73
- Lange, K. L., Little, R. J. A., & Taylor, J. M. G. 1989, *J. Am. Stat. Assoc.*, **84**, 881
- Leike, R. H., Glatzle, M., & Enßlin, T. A. 2020, [A&A](#), **639**, A138
- Lépine, J. R. D., & Sartori, M. J. 2003, *Astrophys. Space Sci. Libr.*, **299**, 63
- Lifshitz, L., & Pizer, S. 1990, *IEEE Trans. Pattern Anal. Mach. Intell.*, **12**, 529
- Lindgren, L., Hernández, J., Bombrun, A., et al. 2018, [A&A](#), **616**, A2
- Lindgren, L., Klioner, S. A., Hernández, J., et al. 2021, [A&A](#), **649**, A2
- Liu, Y., & Xie, J. 2020, *J. Am. Stat. Assoc.*, **115**, 393
- Lombardi, M., Alves, J., & Lada, C. J. 2006, [A&A](#), **454**, 781
- Luhman, K. L. 2022, [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
- MackQueen, J. B. 1967, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, eds. L. M. L. Cam, & J. Neyman (University of California Press), 1, 281
- Magnani, L., Blitz, L., & Mundy, L. 1985, [ApJ](#), **295**, 402
- Makarov, V. V. 2007a, [ApJS](#), **169**, 105
- Makarov, V. V. 2007b, [ApJ](#), **670**, 1225
- Malsiner-Walli, G., Frühwirth-Schnatter, S., & Grün, B. 2016, *Stat. comput.*, **26**, 303
- Mamajek, E. E., & Feigelson, E. D. 2001, in *Young Stars Near Earth: Progress and Prospects*, eds. T. R. Jayawardhana, & T. Greene, *ASP Conf. Ser.*, **244**, 104
- 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
- Matthews, B. 1975, *Biochim. Biophys. Acta (BBA) - Protein Structure*, **405**, 442
- 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
- McInnes, L., Healy, J., & Astels, S. 2017, *J. Open Source Softw.*, **2**, 205
- Meingast, S., & Alves, J. 2019, [A&A](#), **621**, L3
- Meingast, S., Alves, J., & Fürnkranz, V. 2019, [A&A](#), **622**, L13
- Meingast, S., Alves, J., & Rottensteiner, A. 2021, [A&A](#), **645**, A84
- Merloni, A., Nandra, K., & Predehl, P. 2020, *Nat. Astron.*, **4**, 634
- Miret-Roig, N., Galli, P. A. B., Brandner, W., et al. 2020, [A&A](#), **642**, A179
- Miret-Roig, N., Galli, P. A. B., Olivares, J., et al. 2022a, [A&A](#), **667**, A163
- Miret-Roig, N., Bouy, H., Raymond, S. N., et al. 2022b, *Nat. Astron.*, **6**, 89
- Muller, D. W., & Sawitzki, G. 1991, *J. Am. Stat. Assoc.*, **86**, 738
- Müller, P., & Mitra, R. 2013, *Bayesian Anal.*, **8**, 269
- Neuhäuser, R. 1997, *Science*, **276**, 1363
- Neuhäuser, R., Gießler, F., & Hambaryan, V. V. 2020, *MNRAS*, **498**, 899
- Nocedal, J., & Wright, S. J. 1999, in *Numerical Optimization* (New York, NY: Springer)
- 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., Lagache, G., Bacmann, A., et al. 2003, [A&A](#), **406**, L59
- Pagani, L., Bacmann, A., Motte, F., et al. 2004, [A&A](#), **417**, 605
- Pagani, L., Pardo, J. R., Apponi, A. J., Bacmann, A., & Cabrit, S. 2005, [A&A](#), **429**, L181
- 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
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *J. Mach. Learn. Res.*, **12**, 2825
- Penoyre, Z., Belokurov, V., & Evans, N. W. 2022a, *MNRAS*, **513**, 2437
- Penoyre, Z., Belokurov, V., & Evans, N. W. 2022b, *MNRAS*, **513**, 5270
- Perryman, M. A. C., Lindgren, L., Kovalevsky, J., et al. 1997, [A&A](#), **323**, L49
- Perryman, M. A. C., Brown, A. G. A., Lebreton, Y., et al. 1998, [A&A](#), **331**, 81
- Plotly, T. I. 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, 5, 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
- Ratzenböck, S., Großschedl, J. E., Alves, J., et al. 2023, [A&A](#), in press, <https://doi.org/10.1051/0004-6361/202346901>
- Reininghaus, J., Kotava, N., Guenther, D., et al. 2011, *IEEE Trans. Visual. Comput. Graphics*, **17**, 2045
- Richardson, S., & Green, P. J. 1997, *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **59**, 731
- Riedel, A. R., Blunt, S. C., Lambrides, E. L., et al. 2017, [AJ](#), **153**, 95
- 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., Franciosi, E., & Randich, S. 2018, [A&A](#), **617**, L4
- Román-Zúñiga, C. G., Lada, C. J., Muench, A., & Alves, J. F. 2007, [ApJ](#), **664**, 357
- Román-Zúñiga, C. G., Alves, J. F., Lada, C. J., & Lombardi, M. 2010, [ApJ](#), **725**, 2232
- Röser, S., Schilbach, E., Goldman, B., et al. 2018, [A&A](#), **614**, A81
- Röser, S., Schilbach, E., & Goldman, B. 2019, [A&A](#), **621**, L2
- Rybicki, J., Demleitner, M., Bailer-Jones, C., et al. 2020, *PASP*, **132**, 074501
- Rybicki, J., Green, G. M., Rix, H.-W., et al. 2022, *MNRAS*, **510**, 2597
- Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. 2016, *PeerJ Comput. Sci.*, **2**, e55
- Sarro, L. M., Bouy, H., Berihuete, A., et al. 2014, [A&A](#), **563**, A45
- 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. 2022, [A&A](#), **661**, A40
- Schönrich, R., Binney, J., & Dehnen, W. 2010, *MNRAS*, **403**, 1829
- Schwarz, G. 1978, *Ann. Stat.*, **6**, 461
- Shaffer, J. P. 1995, *Ann. Rev. Psychol.*, **46**, 561
- Sim, G., Lee, S. H., Ann, H. B., & Kim, S. 2019, *J. Korean Astron. Soc.*, **52**, 145
- 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., Pagani, L., Bacmann, A., & Guieu, S. 2010, [A&A](#), **511**, A9
- Steinacker, J., Andersen, M., Thi, W. F., et al. 2015, [A&A](#), **582**, A70
- Strehl, A., & Ghosh, J. 2002, *J. Mach. Learn. Res.*, **3**, 583
- Stuetzle, W., & Nugent, R. 2010, *J. Comput. Graphical Stat.*, **19**, 397
- Taylor, M. B. 2005, in *Astronomical Data Analysis Software and Systems XIV*, eds. P. Shopbell, M. Britton, & R. Ebert, *ASP Conf. Ser.*, **347**, 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, *Comput. Sci. Eng.*, **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*, eds. D. Forsyth, P. Torr, & A. Zisserman (Berlin, Heidelberg: Springer), 705
- Vega-Pons, S., & Ruiz-Shulcloper, J. 2011, *Int. J. Pattern Recogn. Artif. Intell.*, **25**, 337
- Vehtari, A., Gelman, A., & Gabry, J. 2017, *Stat. Comput.*, **27**, 1413
- Villa Vélez, J. A., Brown, A. G. A., & Kenworthy, M. A. 2018, *Res. Notes Am. Astron. Soc.*, **2**, 58
- Vinh, N. X., Epps, J., & Bailey, J. 2010, *J. Mach. Learn. Res.*, **11**, 2837
- Wang, K., & Ge, Y. 2021, *Astrophysics Source Code Library* [record ascl:2102.002]
- Watanabe, S. 2010, *J. Mach. Learn. Res.*, **11**, 3571
- Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, [A&AS](#), **143**, 9
- Wishart, D. 1969, in *Proceedings of the Colloquium in Numerical Taxonomy*, ed. A. J. Cole (New York: Academic Press), 282
- Witkin, A. P. 1987, in *Readings in Computer Vision*, eds. M. A. Fischler, & O. Firschein (San Francisco, CA: Morgan Kaufmann), 329
- 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
- Zari, E., Rix, H. W., Frankel, N., et al. 2021, [A&A](#), **650**, A112
- Žerjal, M., Ireland, M. J., Crundall, T. D., Krumholz, M. R., & Rains, A. D. 2023, *MNRAS*, **519**, 3992
- Zomorodian, A., & Carlsson, G. 2005, *Discrete Comput. Geom.*, **33**, 249
- Zucker, C., Speagle, J. S., Schlafly, E. F., et al. 2019, [ApJ](#), **879**, 125
- 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

Appendix A: *Gaia* DR3 data retrieval and details on quality criteria

The *Gaia* DR3 data were downloaded from the *Gaia* Archive²⁹ using the following ADQL query:

```
SELECT * FROM gaiadr3.gaia_source
WHERE (1000./parallax*COS(1*PI()/180)*COS(b*PI()/180))>-50.
AND (1000./parallax*COS(1*PI()/180)*COS(b*PI()/180))<250.
AND (1000./parallax*SIN(1*PI()/180)*COS(b*PI()/180))>-200.
AND (1000./parallax*SIN(1*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_over_error>4.5
```

The first six expressions give the XYZ box conditions, while X is positive toward the Galactic center, Y is positive in the direction of Galactic rotation, and Z points toward the Galactic north-pole. XYZ can also be calculated using `astropy.coordinates.SkyCoord` from *Astropy* v4.0. The parameter `fidelity_v2` from Rybizki et al. (2022) was retrieved with the following ADQL query, using the Topcat TAP Query and the GAVO service³⁰:

```
SELECT gaia.*
FROM gedr3spur.main AS gaia
JOIN tap_upload.t1 AS mine
USING (source_id)
```

We support our quality criteria choices from Sect. 2 as follows. Using `fidelity_v2 > 0.5` is suggested as separator into “good” and “bad” sources by Rybizki et al. (2022) (see also Zari et al. 2021). Using a threshold of 0.9 would give slightly cleaner data; however, few sources lie in the range between 0.5 and 0.9 (~2% in the box), and we opted for the less conservative value. The additional cut using `parallax_over_error`, which is similar to the signal-to-noise ratio (S/N_{ϖ}), is used to reduce further parallax uncertainties. The choice of the threshold $S/N_{\varpi} > 4.5$ is further supported by Rybizki et al. (2022), where they apply different classifiers (high- and low- S/N classifiers) for sources above and below this threshold. To avoid inhomogeneous data, we decided to include this S/N threshold. Moreover, we want to avoid too high parallax errors. As mentioned, we used the inverse of the parallax to estimate the distance to a source, which gets unreliable if the uncertainties are too large. For more distant sources, or intrinsically faint sources with high parallax errors, the distance estimate becomes a nontrivial inference problem (e.g., Luri et al. 2018; Bailer-Jones et al. 2021). In Table A.1 we list the typical uncertainties of the various parameters for sources inside the box after the applied quality criteria, and also separately for sources that are selected as members of the 37 Sco-Cen clusters (given in brackets in Table A.1). It can be seen that the majority of the sources in the box (2σ) have parallax uncertainties below 0.6 mas.

For SigMA we used the 5.5D phase space, as mentioned in Sect. 2, for which we used tangential velocities in km s^{-1} . The proper motions ($\mu_{\alpha}^* = \mu_{\alpha} \cos(\delta)$, μ_{δ}) are transformed from mas yr^{-1} to tangential velocities in km s^{-1} as follows, where the conversion constant 4.74047 is in units of km yr s^{-1} :

$$\begin{aligned} v_{\alpha} &= 4.74047 \cdot \mu_{\alpha}^* / \varpi \\ v_{\delta} &= 4.74047 \cdot \mu_{\delta} / \varpi. \end{aligned} \quad (\text{A.1})$$

Moreover, we corrected the tangential motions for the Sun’s reflex motion, resulting in tangential motions relative to the LSR,

Table A.1. Typical parameter uncertainties for sources used in our analysis.

Uncertainties ^a	1, 2, 3 σ percentiles		
	68.3%	95.5%	99.7%
σ_{ϖ} (mas) <	0.13 (0.08)	0.56 (0.34)	1.17 (0.94)
S/N_{ϖ} >	47.6 (83.2)	9.3 (20.8)	4.8 (7.5)
σ_d (pc) <	+3.8(+1.8) −3.6(−1.7)	+26.0(+7.4) −21.0(−6.7)	+65.2(+23.8) −43.7(−18.0)
$\sigma_{\mu_{\alpha}}$ (mas yr^{-1}) <	0.13 (0.09)	0.60 (0.39)	1.47 (1.06)
$\sigma_{\mu_{\delta}}$ (mas yr^{-1}) <	0.12 (0.08)	0.54 (0.32)	1.34 (0.89)
$\sigma_{v_{\alpha}}$ (km s^{-1}) <	+0.4(+0.2) −0.4(−0.2)	+3.2(+0.6) −2.7(−0.6)	+12.1(+2.1) −8.9(−1.7)
$\sigma_{v_{\delta}}$ (km s^{-1}) <	+0.4(+0.2) −0.4(−0.2)	+3.3(+0.8) −2.8(−0.7)	+12.1(+2.4) −8.9(−1.9)
σ_{v_r} (km s^{-1}) <	3.3 (5.8)	8.3 (13.0)	26.2 (37.3)
σ_X (pc) <	+2.3(+1.4) −2.2(−1.4)	+20.1(+6.3) −16.5(−5.7)	+54.6(+20.3) −36.8(−16.0)
σ_Y (pc) <	+1.6(+0.6) −1.5(−0.6)	+13.4(+2.9) −11.2(−2.7)	+41.5(+12.1) −28.1(−9.8)
σ_Z (pc) <	+3.9(+0.4) −0.9(−0.4)	+6.5(+1.8) −5.4(−1.7)	+19.6(+6.4) −13.4(−5.1)

Notes. The listed uncertainties give the threshold below which the given percentage of sources can be found in the whole box (or in the SigMA Sco-Cen selection, in parentheses). ^aGiven are the uncertainties for the following parameters: parallax, distance, proper motions in the direction of α and δ , tangential velocities in the direction of α and δ , radial velocities from *Gaia* DR3, and XYZ. The radial velocities are available for a subsample of 367,127 sources (37%) inside the box (or 4967, 40%, in the SigMA Sco-Cen selection). For the derived parameters we give lower and upper uncertainty thresholds, since the errors, gained via a Monte Carlo approach, get asymmetric for large parallax errors.

using the values by Schönrich et al. (2010). This conversion is accomplished with the help of *Astropy*, by defining the below sky coordinates. For a conversion of heliocentric proper motions to proper motions relative to the LSR, the radial velocity can be set to an arbitrary value in the sky-coordinate definition of *Astropy* (here set to 0), since different RV values do not change the outcome of this conversion.

```
from astropy.coordinates import ICRS, LSR
from astropy import units as u
```

```
skyicrs = ICRS(ra = ra * u.deg,
               dec = dec * u.degree,
               distance = 1000./parallax * u.pc,
               pm_ra_cosdec = pmra * u.mas/u.yr,
               pm_dec = pmdec * u.mas/u.yr,
               radial_velocity = 0. * u.km/u.s)
```

```
pma_lsr = skyicrs.transform_to(LSR()).pm_ra_cosdec.value
pmd_lsr = skyicrs.transform_to(LSR()).pm_dec.value
```

```
v_a_lsr = 4.74047 * pma_lsr / parallax
v_d_lsr = 4.74047 * pmd_lsr / parallax
```

In Sect. 5.2 we compare the SigMA clusters with recent literature samples. To this end, we cross-match the samples using the *Gaia* DR3/EDR3 `source_id`. This cross-match is straight forward for the samples of Schmitt et al. (2022), Squicciarini et al. (2021), Luhman (2022), and Miret-Roig et al. (2022a) who used *Gaia* DR3/EDR3 data, which allows a direct match with the `source_id`. In the case of Damiani et al. (2019), Kerr et al. (2021), and Žerjal et al. (2023), who used *Gaia* DR2 data, we first retrieve the *Gaia* DR3 `source_id` using the `gaiadr3.dr2_neighbourhood` catalog from the *Gaia* Archive, since the DR2 and DR3 `source_ids` are not generally the same. Such a cross-match delivers a few sources that have several possible matches of DR3 with DR2 sources (see Torra et al. 2021;

²⁹ <https://gea.esac.esa.int/archive/>

³⁰ <https://dc.zah.uni-heidelberg.de/>, German Astrophysical Virtual Observatory

Gaia Collaboration 2021). In such cases, we chose the closest match, using the provided `angular_distance` parameter.

Appendix B: Detailed information and background on the methods

In this appendix, we give in-depth and background information on some aspects of our methodology from Sect. 3.

B.1. Related work: Cluster analysis

In the following, we present a cross section of related work upon which SigMA rests. From the vast corpus of data mining and statistics literature, we focus specifically on identifying stable groups using density-based clustering methods.

Hierarchical, density-based clustering

The strength of level-set formulation (see Eq. (3) and further discussions in Sect. 3.1.3) lies in the natural emergence of a cluster tree, a clustering hierarchy that arises from sweeping the density threshold λ from $\infty \rightarrow -\infty$. With a continuous change in λ , the number of connected components changes when the threshold passes through a critical point in f , and thus $\nabla f = \mathbf{0}$. A new cluster is born when λ reaches the height of a mode in f . On the other hand, a cluster dies when λ 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 λ , and thus $\{x \in X : \hat{f}(x) \geq \lambda\}$ ³¹.

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).

The HDBSCAN algorithm

A well-known algorithm belonging to the family of hierarchical level-set methods is the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) algorithm

(Campello et al. 2013), 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 that provides an interpretable pruning strategy.

At each cluster split decision, the smaller cluster created is merged back into the mother cluster if it has fewer 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 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 and which 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 affect the confidence and interpretability of the clustering result.

Topological methods

Extracting a flat clustering from the cluster tree requires a notion of cluster stability. As discussed, the concept of relative EOM, 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 initial 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, ν , defined by the ratio of birth and death density, as a cluster stability criterion. By varying ν 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.

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 α , 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 a unimodal density, whereas the alternative hypothesis H_1

³¹ Edges are commonly assigned the minimum density sampled along the path connecting two vertices.

suggests multi-modality. The null hypothesis is rejected at a significance level α 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 GMM at once. Dip means (Kalogeratos & Likas 2012) proposes an incremental clustering scheme for selecting k in k -means that employs Hartigan's dip statistic (Hartigan & Hartigan 1985). If the distance distribution of one or more points to their co-cluster members exhibits a significant multimodal structure, the cluster is split.

Skinny-dip (Maurus & Plant 2016) also implements Hartigan's dip test and applies it to 1D linear projections of the data set. Distinct density peaks are to be identified based on the gradient of the projected 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 arbitrarily oriented and shaped clusters, which applies a simulation strategy to approximate values for the smallest density dips of unimodal data sets of the same size and density. However, we do not want to depend on simulations but directly obtain a measure of significance from given data.

B.2. Testing for uni-modality

Here we highlight the work of Burman & Polonik (2009) more closely, whose 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 formally define the concept of such a path.

A formal description of the test procedure

We consider directed, continuous paths from \mathbf{x}_1 to \mathbf{x}_2 through input space \mathcal{X} . By assuming there exists a parametrization $\mathbf{r}(t)$, with $t \in [0, 1]$, the path becomes the image of $\mathbf{r}(t)$. 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 $\mathbf{x}_1 = \mathbf{r}(0)$ and $\mathbf{x}_2 = \mathbf{r}(1)$, respectively.

Let f be the underlying density function and \mathbf{x}_1 and \mathbf{x}_2 two candidate modes of f . We assume, without loss of generality, that $f(\mathbf{x}_1) < f(\mathbf{x}_2)$. If all possible paths undergo a density dip when moving from \mathbf{x}_1 to \mathbf{x}_2 , both points are found in two distinct modal regions:

$$\exists t \in (0, 1) : f(\mathbf{r}(t)) < f(\mathbf{x}_1) \quad (\text{B.1})$$

Conversely, if we can find a path between \mathbf{x}_1 and \mathbf{x}_2 where all points have a higher density than \mathbf{x}_1 , both points are part of the same modal region:

$$f(\mathbf{r}(t)) \geq f(\mathbf{x}_1) \quad \forall t \in (0, 1]. \quad (\text{B.2})$$

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

$$f(\mathbf{r}(t)) \geq \min(f(\mathbf{x}_1), f(\mathbf{x}_2)) \quad \forall t \in (0, 1]. \quad (\text{B.3})$$

An equivalent and useful formulation is obtained by taking the logarithm on both sides; after that, the left side is subtracted from the inequality:

$$\text{SB}(t) := -\log f(\mathbf{r}(t)) + \min(\log f(\mathbf{x}_1), \log f(\mathbf{x}_2)) \quad (\text{B.4})$$

Using the variable $\text{SB}(t)$, we can formulate the null hypothesis as follows:

$$H_0 : \text{SB}(t) \leq 0 \quad \forall t \in (0, 1). \quad (\text{B.5})$$

Rather than testing H_0 across the entire path, a point-wise test $H_{0,t} : \text{SB}(t) \leq 0$ for some values of t is employed.

Since we do not have access to the underlying density f , we cannot directly test the hypothesis in Eq. (B.5). 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 $\text{SB}(t)$ as the number data samples approaches infinity:

$$\widehat{\text{SB}}(t) = d \sqrt{k/2} [\log d_k(\mathbf{r}(t)) - \max(\log d_k(\mathbf{x}_1), \log d_k(\mathbf{x}_2))]. \quad (\text{B.6})$$

Here $d_k(\mathbf{x})$ denotes the distance to the k th nearest neighbor of the point \mathbf{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 $\widehat{\text{SB}}(t)$ is approximately standard normally distributed, the null hypothesis is rejected at significance level α if

$$\widehat{\text{SB}}(t) \geq \Phi^{-1}(1 - \alpha), \quad (\text{B.7})$$

where Φ is the standard normal CDF. Therefore, if any $t \in (0, 1)$ fulfills condition (B.7), H_0 is rejected.

Due to the employment of the k -NN technique, this test procedure applies naturally to multivariate data without the need to project the data onto a 1D 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.

Empirical results

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, enough samples must be tested along the path to detect significant dips reliably.

We provide a natural extension to the presented procedure, which applies to arbitrary cluster shapes while reducing the number of test evaluations to one. In Sect. 3.2, we describe the modification in more detail and argue that reducing the original method to a single point-wise evaluation at the saddle point leaves the test unchanged.

To substantiate this statement, we empirically validate our method on simulated data. In particular, we aim to show that these changes do not affect the test statistic distribution under

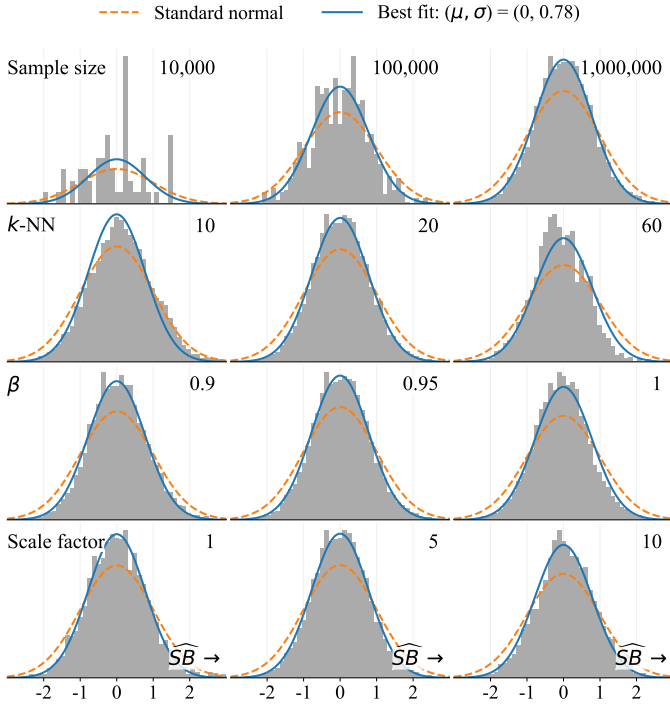


Fig. B.1. Test statistic distribution, \widehat{SB} , of our reduced test procedure on a unimodal data set. The distribution is stable to variations in different parameters and respective values (as shown in the corresponding rows and columns).

the null hypothesis. According to [Burman & Polonik \(2009\)](#), the test statistic \widehat{SB} is standard normally distributed under H_0 , which assumes uni-modality between two points in the input space.

Figure B.1 shows the test statistic distribution \widehat{SB} of our reduced test procedure on a unimodal data set. To faithfully test the distribution under H_0 , we require a unimodal data set with *Gaia* and *Sco-Cen*-like positions and kinematic properties, realistic errors, and suitable size. Since the *Gaia* EDR3 mock catalog ([Rybizki et al. 2020](#)) reproduces kinematic features found in the Milky Way, its uni-modality is guaranteed.

Instead, we directly use data within the box defined around *Sco-Cen* in Eq. (1). We randomly shuffle the observations in each feature to remove any local over-densities. This procedure leaves marginal distributions unchanged while fully de-correlating the data. Since all marginal distributions are unimodal, the joint distribution (implicitly constructed as a factorization of marginals) is equally unimodal.

To gauge parameter effects on the test statistic distribution, we vary the sample size and three *SigMA* parameters across different values³². The parameters are the overall sample size, the k parameter of k -NN density estimation method (see Sect. 3.3.2), the β parameter of the underlying β -skeleton graph (see Sect. 3.3.1), and the velocity scaling factor (see Sect. 3.3.3).

As shown in Fig. B.1, varying the given parameters within a sensible range does not modify the test statistic distribution. Due to its stability, a single test statistic distribution under H_0 can be universally assumed over different parametrizations of *SigMA*.

The \widehat{SB} distribution on unimodal data closely follows a zero-mean Gaussian. However, in our tests, the standard deviation differs slightly from unity as stated by [Burman & Polonik \(2009\)](#).

We update its value to the average standard deviation across our tests of $\sigma = 0.78$.

B.3. Determining velocity scaling factors

Here we discuss the derivation of the scaling factor distribution, which we used to weigh the velocity subspace in the clustering process. A more detailed justification is provided in 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).

B.3.1. Statistical model

Figure 4 shows the relation between a cluster's distance and the scaling factor y alongside determined uncertainties. We observe an approximately linear relationship between the cluster distance and the scale factor that we aim to model. We observe that determined uncertainties for the i th data point σ_i cannot account for deviations from any hypothetical regression line. Thus, we include the simple assumption in the model that reported variances are underestimated by some factor, f . The scaled variance for the i th sample becomes the following:

$$s_i^2 = \sigma_i^2 + f^2(mx_i + b). \quad (\text{B.8})$$

Here the parameters m and b represent the slope and intercept of the regression line, respectively.

We assume Cauchy distributed deviations from the regression line to further reduce the influence of outliers. The Cauchy distribution provides a robust statistical model that naturally handles observations with considerable deviations from the mean with its longer-than-normal-tails ([Lange et al. 1989](#)). The scale parameter becomes the adjusted variance in Eq. (B.8). In this model, given a distance to a stellar cluster, r_i , a modified scaling factor uncertainty, s_i , a fractional amount, f , a slope, m , and an intercept, b , the density of observed velocity scaling factors $p(y_i | r_i, s_i, f, m, b)$ becomes

$$p(y_i | r_i, s_i, m, b) = \left[\pi s_i \left(1 + \frac{(y_i - mr_i - b)^2}{s_i^2} \right) \right]^{-1}. \quad (\text{B.9})$$

By factoring these conditional probabilities for the N data points (assuming statistical independence between them), we obtain the likelihood:

$$\mathcal{L} = \prod_{i=1}^N p(y_i | r_i, s_i, m, b). \quad (\text{B.10})$$

Using Bayes' theorem, we obtain the posterior probability density function (PDF):

$$p(m, b, f | \{y_i\}_{i=1}^N, I) \propto p(\{y_i\}_{i=1}^N | m, b, f, I) p(m, b, f | I). \quad (\text{B.11})$$

The PDF $p(m, b, f | I)$ describes our prior knowledge of the line parameters (m, b) and the scaled variance f , while I summarizes all the prior knowledge of the r_i and σ_i . We employed weakly informative priors (as the problem is relatively low dimensional, this does not affect the inference outcome) with normal, zero-mean prior densities for m and b with a standard deviation of 20. The prior PDF for f follows a half-Cauchy distribution with $\beta = 10$. The Half-Cauchy distribution provides a suitable prior PDF as it is truncated to have nonzero probability density on \mathbb{R}^+ .

³² While varying one parameter, the remaining ones are set to their default values as described in Sect. 3.3

To obtain samples from the posterior PDF, we used PyMC (Salvatier et al. 2016), a publicly available code that implements the MCMC method. For each line parameter (m, b), we computed its maximum a posteriori position, representing the best-fitting line. We generate samples from the posterior predictive distribution to estimate the uncertainty in y around the mean at each location r . Figure 4 and B.3 show the 1σ credible interval determined via computing the 68% high-density interval (HDI) of generated posterior predictive samples.

B.3.2. Model caveat

The relationship between the dispersion in position and velocity space is, in principle, not affected by the observer. We conjecture that the observed correlation likely stems from different magnitudes of observational uncertainties between tangential velocities and the parallax. Figure B.2 shows the distribution of the ratio between measured distance and tangential velocity uncertainties as a function of distance³³. We identify an almost perfect linear trend when plotting the rolling median (window size 5 pc) for sources within 500 pc. This trend suggests that, on average, the distance error (i.e., parallax error) grows faster than the tangential velocity error when moving away from the observer (which is to be expected). This relationship suggests a faster growth rate in positional cluster dispersion than in on-sky motions caused by the convolution with (on average) larger uncertainties. It directly affects the density by smearing out the unobserved source distribution in respective positional and kinematic subspaces at different scales as a function of distance, which we aim to counteract using scaling factors conditioned on distance.

This relationship suggests deriving the scaling factors directly from the observed error ratio distribution. However, this entails propagating the uncertainties through the complex selection functions of clustering algorithms. Instead, we aim to derive the relationship directly from data from extracted clusters in the Gaia data set (see Sect. 3.3.3 and Fig. 4).

As a final validation to choosing a linear model (assuming a correlation between distance and the dispersion relation) over a constant one (assuming variable independence), we fit a constant and linear model to the data. As the linear model, the constant model assumes Cauchy distributed deviations from the mean and assumes that reported variances are underestimated by some factor f . We employed the same weakly informative priors for the intercept and factor parameters (b, f). Using samples drawn from the posterior distribution and the computed log point-wise posterior predictive density, we applied leave-one-out cross-validation (Vehtari et al. 2017) and the widely applicable information criterion (Watanabe 2010) to perform model selection. We find that both criteria significantly favor the linear over the constant model.

³³ The uncertainties in distance and tangential velocity have been obtained by propagating the uncertainties in parallax and proper motions. We restrict the sample to sources with small relative uncertainties to guarantee minor uncertainty approximation errors, using quality criteria discussed in Eq. (2). Modifying the quality criteria affects the range of y values but does not eliminate the linear trend between distance and the presented uncertainty ratio when analyzing the rolling median.

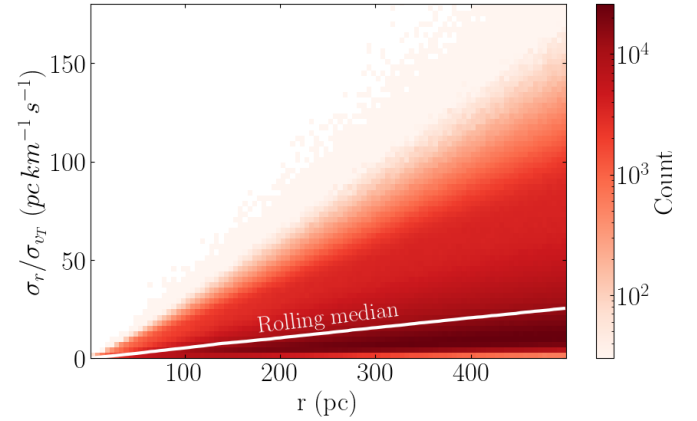


Fig. B.2. Frequency distribution of the ratio between measured distance and tangential velocity uncertainties as a function of distance. We observe a linear trend (suggested by the rolling median, window size 5 pc) between this ratio and distance. We hypothesize that the empirical distance-scaling relationship (see Fig. 4) is caused by this trend.

B.3.3. A distribution over scaling factors

We aim to obtain the distribution $f(y | r_0 \leq r \leq r_1)$, which describes the scaling factor (y) behavior for a given range of distances to clusters 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 $f(r, y)$ ³⁴.

As we are dealing with an improper PDF, we consider the following proportionality condition and handle the normalization of the left-hand side later:

$$\begin{aligned} f(y | r_0 \leq r \leq r_1) &\propto \int_{r_0}^{r_1} f(r, y) dr \\ &\propto \int_{r_0}^{r_1} f(y | r) f(r) dr. \end{aligned} \quad (\text{B.12})$$

Since $f(r) \propto 1$ is approximately independent of 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

$$f(y | r_0 \leq r \leq r_1) \propto \int_{r_0}^{r_1} f(y | r) dr. \quad (\text{B.13})$$

Thus, to obtain a solution to Eq. (B.13), we need an expression for the conditional PDF $f(y | r)$. We assume that inlier data are Gaussian distributed around the linear model with a variable standard deviation $\sigma(r)$, which we estimate from the 1σ HDI of generated posterior predictive samples. The distribution of scaling parameters y conditioned on the distance r can then be approximated via the following expression:

$$f(y | r) \propto \exp\left(-\frac{(y - g(r))^2}{2\sigma(r)^2}\right). \quad (\text{B.14})$$

Figure B.3 schematically shows the integrating process where the conditional PDFs $f(y | r)$ are shown for $r = 100$ and $r = 200$.

Since we do not have any analytic expression for $\sigma(r)$, we numerically approximate the integral in Eq. (B.14). The top part

³⁴ The marginal distribution $f(r)$ is approximately uniform, i.e., $f(r) \propto 1$ over \mathbb{R}^+ . Thus, the joint distribution $f(r, y)$ is improper as it does not integrate to unity.

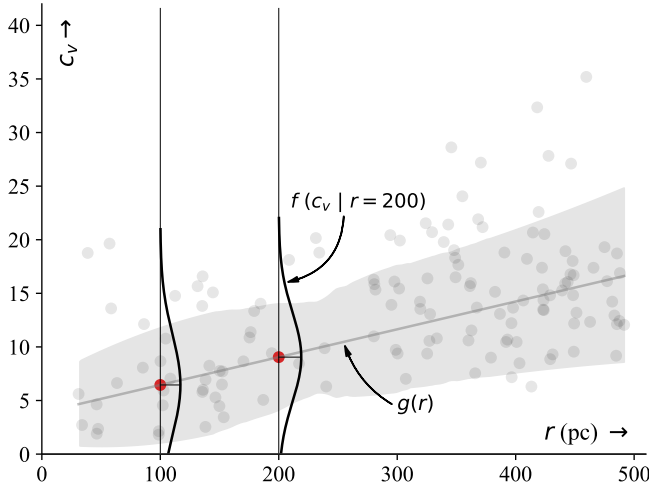


Fig. B.3. Scaling factor determination via the empirical distance-scaling relationship. The scaling factor distribution for clusters at a distance between 100 and 200 pc depends on the conditional distribution of scaling factors at a given distance, $f(c_v | r)$.

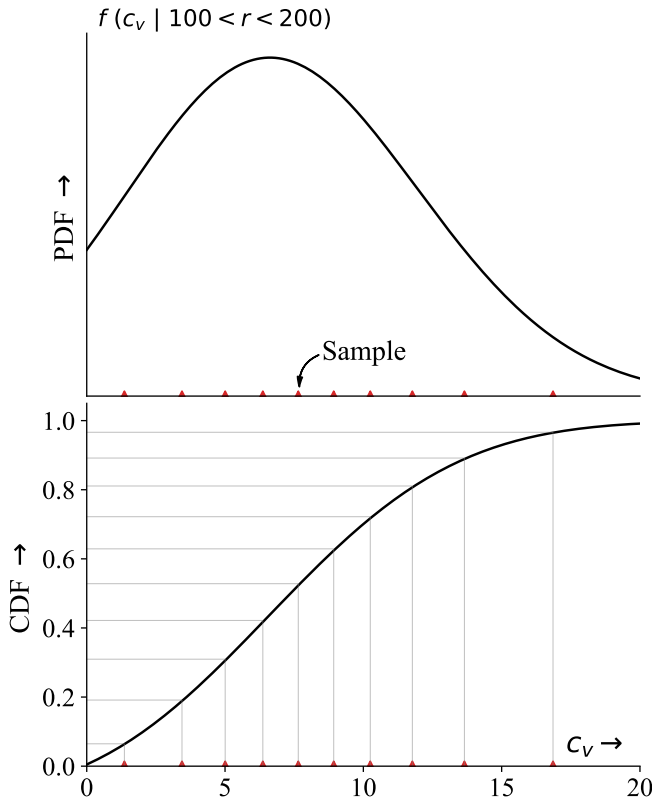


Fig. B.4. PDF and CDF of the scaling factor conditioned on a given range of distances. The ten red scatter points indicate samples drawn from the ten-quantile splitting procedure. We separate the PDF into ten continuous intervals with equal probabilities from which we derive samples as the mean position of these intervals.

of Fig. B.4 shows the resulting PDF when we solve $f(y | r_0 \leq r \leq r_1)$ for sources in Sco-Cen, where we assume a minimum distance of $r_0 = 100$ and a maximum distance of $r_1 = 200$. Here an immediate caveat of our simple symmetric model uncertainty assumption becomes apparent; the resulting distribution has infinite support and, thus, a nonzero probability density for

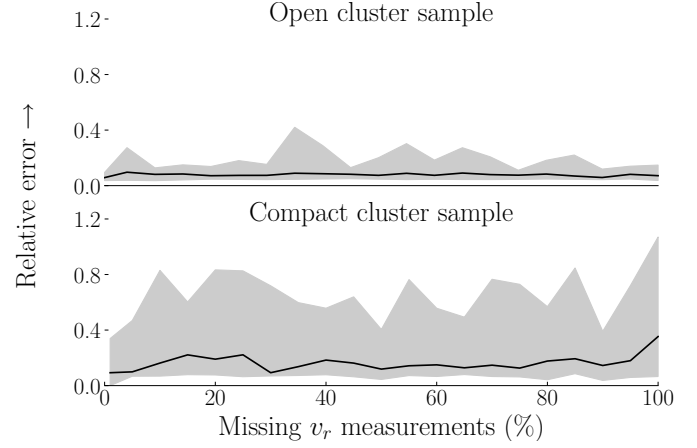


Fig. B.5. Simulating the impact of the fraction of missing v_r measurements on the relative error of inferred radial velocities. The top panel shows relative errors of inferred v_r on the open cluster sample as a function of missing-ness. The bottom panel shows the result of this study using the compact cluster sample. We find twice larger errors on the compact cluster sample. Additionally, the estimation procedure seems almost independent of the number of missing radial velocity measurements in both samples, except in a small region of the compact cluster sample, where the error increases if fewer than 5% of sources have v_r measurements.

$f(y < 0 | r)$. Since negative scaling is physically meaningless, we limit the distribution to \mathbb{R}^+ , thus setting $f(y < 0 | r) = 0$ and normalizing the PDF to integrate to unity in \mathbb{R}^+ .

We consider sampling strategies to obtain scaling factors for 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 ten samples that represent ten quantiles of the PDF. We separate the PDF into ten continuous intervals with equal probabilities from which we derive samples as the mean position of these intervals.

To compute the quantiles, we numerically determine the CDF of $f(y | r_0 \leq r \leq r_1)$. The CDF for $r \in [100, 200]$ is shown in the bottom part of Fig. B.4. The ten red scatter points³⁵ 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 $F^{-1}(y | r_0 \leq r \leq r_1)$ we used a numerical approximation³⁶.

B.4. Influence of missing radial velocities on the bulk velocity determination

In this section we examine the influence of missing radial velocities v_r on the accuracy of determined bulk velocities. As discussed in Sect. 3.5.2, the bulk velocity of a sample is determined by minimizing observed proper motions and radial velocities against theoretical observables for a given cluster bulk motion. Since radial velocities are not always available (in our sample $\sim 20\%$ of sources have v_r measurements), especially as more distant objects are studied, we estimate how the fraction of missing v_r affects the accuracy of determined bulk velocities and

³⁵ The velocity scaling values are:

$c_v = \{1.36, 3.44, 5.01, 6.37, 7.65, 8.93, 10.25, 11.77, 13.65, 16.86\}$.

³⁶ We made use of the open source library `pynverse v0.1.4.4` to calculate the numerical inverse of the CDF.

consequently inferred radial velocities of cluster members. For this purpose, we simulate a given fraction of missing v_r in our open cluster and compact cluster samples (see Sect. 4.2 for detailed descriptions of these samples) and compare the inferred radial velocities of cluster members to known true values. We repeat this process for several fractions of missing v_r , ranging from 1% missing to 100% missing, and calculate the mean relative error (and the 1σ quantile) between the resulting calculated radial velocities (via the inferred bulk motion; see Sect. 3.5.2) and the ground truth. Figure B.5 shows the influence of various fractions of missing v_r s on the relative error.

First, we find that relative errors of inferred v_r on the open cluster sample are, on average, around twice (1.97 ± 0.51) as low compared to inferred v_r from the compact cluster sample. This difference is consistent with results reported in Sect. 4.2, where we report very low contamination in clustering solutions on the open cluster sample compared to the compact cluster sample. Further, the cluster sizes in the open cluster sample are, on average, four times larger, providing significantly better statistics to determine the bulk motion from the minimization procedure in Eq. (12).

Second, we identify negligible correlation between the relative error in inferred radial velocities and the fraction of missing v_r measurements. In particular, the open cluster sample shows no performance loss as the number of missing radial velocities goes to 100%. From this result, we conjecture that given large enough clusters, the bulk velocity determination discussed in Sect. 3.5.2 is sufficiently constrained by the observed proper motions and, hence, independent of radial velocity measurements. Although the relative errors also seem mostly independent of the number of missing radial velocity measurements, we observe an increased error of inferred v_r in the compact cluster sample as the missing-ness goes toward 100%. When no radial velocities are available, the relative uncertainty of inferred v_r grows around twofold (mean relative error of 0.38 versus 0.21). At the same time, the statistical dispersion (measured via the 1σ range) appears to be almost stable when compared to cases with access to v_r measurements. Since the v_r estimation procedure seems independent of the number of missing radial velocity measurements for less than 95% missing-ness, we recommend using SigMA on data with at least 5% of available v_r measurements to keep radial velocity estimation errors to a minimum. Since SigMA identifies clusters in 5D phase space and determines members most effectively if at least 5% of input stars have radial velocity measurements, we refer to the entire SigMA pipeline working in 5.5 dimensions (5.5D).

B.5. Consensus clustering

This section discusses algorithmic means to identify stable cluster solutions from an ensemble of clustering results. To identify robust solutions, we represent all clusters across the clustering ensemble in a graph. Each cluster of a single solution from the ensemble is represented by one node. We connect two nodes via an edge if two clusters share at least one common point. Figure B.6 highlights this step in the first two frames. The ensemble comprises three clustering solutions. The individual runs A, B, and C contain three, two, and four clusters, respectively. Each source is classified into a single cluster for a given run, resulting in disjoint sets. Thus, edges connect clusters from different runs.

Edges in the graph are weighted by the corresponding Jaccard similarity, which measures their common overlap. The Jaccard similarity is defined as the ratio of the intersection of two

sets over their union. Typically, a 0.5 or greater value indicates a high similarity between two sets. Since this linkage is fundamental for determining the consensus result, it should avoid connecting dissimilar clusters. Thus, we remove edges with a weight below 0.5 (see Fig. B.6, step 3). This threshold is quite conservative, as it can separate similar cluster solutions, for example if one cluster is a subset of the other. We relax the cut criterion in the following way to avoid over-pruning the graph. Two clusters a and b with respective sizes n_a and n_b , where $n_a > n_b$, are linked if the n_b densest points of a amount to a Jaccard similarity of 0.5 or greater with points from cluster b . This criterion guarantees connectivity between cluster extractions at different isosurface thresholds while separating ties to clusters that randomly fragment into multiple subclusters.

Robust clusters throughout the ensemble will have strong connections to their counterparts from different runs while having none or very weak connections to other nodes in the graph. Thus, a robust cluster solution builds a strong clique in this graph³⁷. In contrast, unstable clusters will have many weak connections to many other clusters from different runs but none or very few strong ones; see Fig. B.6, panel 4.

To extract all stable cluster solutions, we aim to identify all strongly connected cliques in the graph (i.e., the consensus result). Since individual sources and clusters can be part of several cliques, we employed a voting strategy to determine the final data partitioning (Vega-Pons & Ruiz-Shulcloper 2011). Each clique is represented as a vector of length N , where N is the number of sources in the data set. The N values correspond to the sum of the individual clusters represented as 0–1 vector, where all entries are 1 for sources inside a cluster and 0 elsewhere. Each source is then associated with a single clique by maximizing the respective entry at the source’s position (in the vector). Thus, the larger a clique, the more likely it wins a vote. To favor robust cluster solutions, we multiply each vector by the median of its connection strengths. This number is maximized if the cluster is unchanged throughout different runs. This step is summarized in panel 5 in Fig. B.6.

B.6. Testing the independence assumption of resampled Gaia data

When testing for multi-modality, SigMA simulates multiple realizations of the *Gaia* data set to see how robust density dips between pairs of neighboring peaks are. In Sect. 3.4, we discuss how these realizations are used in a global hypothesis test that determines whether samples of pairwise adjacent density enhancements are part of a single underlying mode. This global hypothesis test combines individual tests from each realization and evaluates the global null hypothesis that no p -value is significant.

However, the choice of combining different p -values is affected by the correlation structure between distinct tests. Typically, statistical tests require independence across tests to guarantee proper levels of specificity and sensitivity. In the case of the commonly used Fisher’s combination test (Fisher 1934), positively correlated p -values increase the chance of type I errors (rejects a true null hypothesis). To investigate the independence assumption of p -values across resampled data sets, we consider the effects that resampling has on the modality test procedure.

The modality test is fully described by the dimensionality of the data and, most importantly, by the k -distances of

³⁷ A clique in a graph is a subset of nodes that are all connected with each other. Thus, every pair of nodes in a clique is joined by an edge.

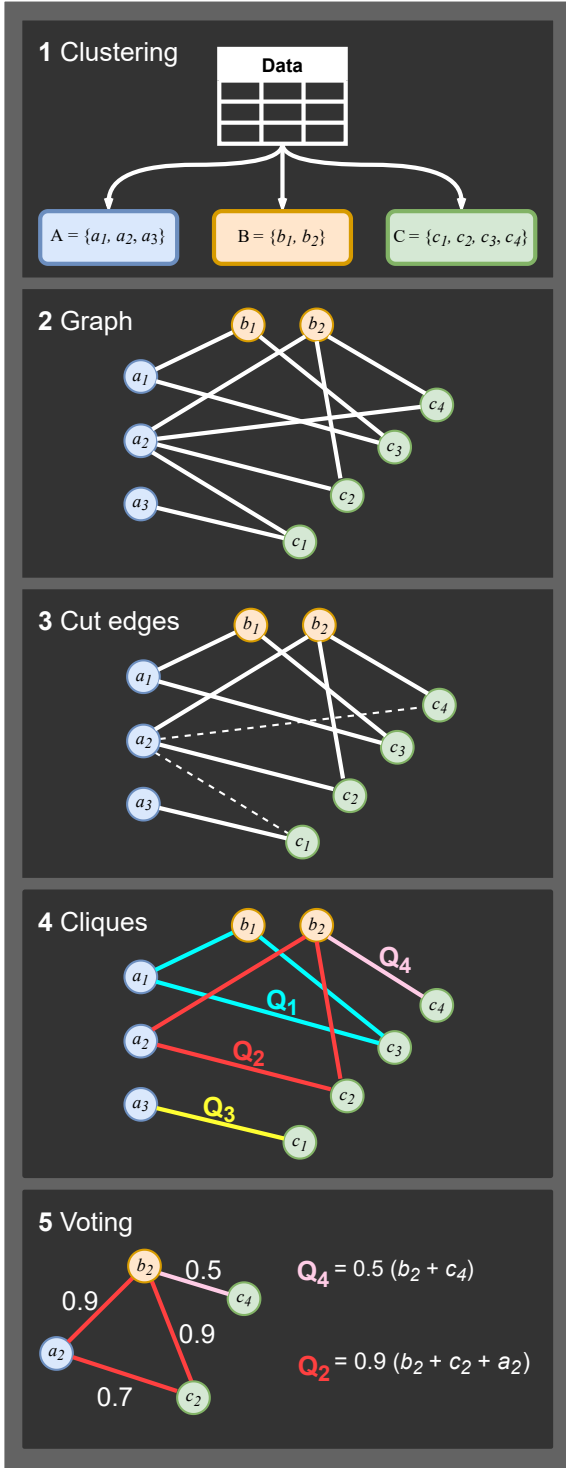


Fig. B.6. Consensus clustering pipeline for a simple example using an ensemble of three clustering solutions: A , B , and C . (2) Clusters from the ensemble are linked in a graph based on overlapping points. (3) Edges between clusters are removed if the overlap between their members is insufficient to assume a common cluster solution. (4) Cliques that represent stable cluster solutions, i.e., consensus clusters, are extracted. (5) A voting strategy determines the assignment of individual sources to cliques.

modal and saddle points. Resampling the data set influences these k -distances. The sampling of new data points is done with Gaussians centered on mean astrometric observables with

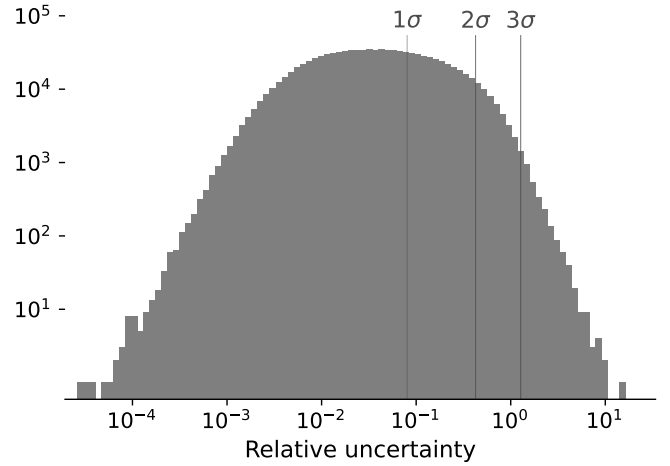


Fig. B.7. Log-log histogram showing the relative uncertainty of data points, i.e., the ratio of positional uncertainty given by the error covariance matrix over the nearest neighbor distance. The majority of data points are far below unity. The concentration of relative uncertainty values at zero indicates that k -distances across resampled data sets are strongly correlated. The 1σ , 2σ , and 3σ lines indicate the percentiles containing 68, 95, and 99.7% of the respective distribution.

heteroscedastic error covariance matrices available in the *Gaia* database. We aim to use the covariance matrix in relation to the nearest neighbor distance to estimate the interdependence of p -values.

As the entries of the covariance matrix shrink to zero, the resampled data points converge to the original data points until they eventually coincide. At the same time, the p -values across different data sets approach the same value, leading to a perfect correlation between them. On the other hand, if the standard deviation along its principle axis (or any other direction for that matter) extends far beyond a point nearest neighbor or even its k -distance, the resampled data differs substantially from the original one, de-correlating the p -values.

Figure B.7 illustrates this relationship for data in the Sco-Cen box. The histogram shows the relative uncertainty of data points (i.e., the ratio of positional uncertainty given by the error covariance matrix over the nearest neighbor distance). The distances are computed in the space of observed astrometric quantities where the uncertainties are assumed Gaussian. Most data points are far below unity (i.e., errors are small relative to their absolute values). The majority (68%, i.e., 1σ) has a relative uncertainty below 0.08 (see the 68, 95, and 99.7% percentiles in Fig. B.7 marked as 1σ , 2σ , and 3σ , respectively). The concentration of relative uncertainty values at zero indicates that k -distances across resampled data sets are strongly correlated. Thus, we cannot assume independence of p -values across different samples.

B.7. Distribution of point-wise densities

Instead of directly recovering clusters from phase space data, SigMA extracts modal regions, a mixture of field stars and cluster members. To separate signal from background, we employed a density-based classifier that selects cluster members as an over-density over the background. We consider the distribution of field stars and cluster members in univariate density to determine a density threshold automatically.

As discussed in Sect. 3.5, we approximately can treat the field star content as a uniform distribution locally around a

cluster in phase space. A uniform distribution in phase space translates to a Gaussian distribution in 1D density space (see Sect. 3.5 for a more detailed discussion).

Cluster members are commonly modeled as multivariate Gaussians (e.g., Gagné et al. 2014; Sarro et al. 2014; Crundall et al. 2019; Riedel et al. 2017). Although observational findings point to more complex morphologies (e.g., Meingast & Alves 2019; Röser et al. 2019; Jerabkova et al. 2021; Meingast et al. 2019; Kounkel & Covey 2019; Cantat-Gaudin et al. 2019a; Wang & Ge 2021; Coronado et al. 2022), the Gaussianity assumption provides a good starting point to consider the point-wise univariate density distribution. Figure B.8 shows multiple distributions of point-wise density estimates of 100,000 samples drawn from an N -dimensional Gaussian³⁸. The left column shows the likelihood of individual samples. It provides an assessment of the local density under the true model. The number of samples with a relatively high likelihood decreases exponentially as the dimension N increases from one (top row) to six (bottom row).

The curse of dimensionality plagues neighborhood queries in high dimensions. As the dimensionality grows, points are increasingly isolated, making neighborhoods no longer local. This effect can already be seen for moderate dimensions in the right column of Fig. B.8. It shows point-wise density estimates obtained via the k -NN technique. Around the fourth dimension, the distribution of k -distances starts to converge to a normal distribution, which incorrectly suggests an underlying uniform distribution in N -dimensional input space.

We computed point-wise density estimations in five and six dimensions. Although we cannot specifically write down a generative model for stellar clusters in phase space, we can assume that neighborhood queries are strongly affected by the given dimensionality. Thus, we model background and signal contributions as Gaussians in univariate density space.

B.8. Parameter optimization

The parameter choices of our proposed SigMA analysis pipeline are tuned to *Gaia* data (see Sect. 3.3). In contrast, DBSCAN and HDBSCAN, which we used to compare and test our clustering technique, are general clustering techniques whose parameters we must set. Instead of using subjective, error-prone prior knowledge to determine suitable parametrizations, we search the space of possible clustering results for the best result. This strategy measures the peak performance (in case of comprehensive/absolute prior knowledge) a clustering algorithm can achieve. Thus, a comparison against the best results allows for a discussion on methodological advantages and disadvantages rather than reflecting poor parameter selection.

To search the space of possible model configurations, we employed a grid search to evaluate the clustering algorithm on a regular grid in parameters space. A grid search has significant benefits in this scenario compared to other parameter tuning methods, such as random search or Bayesian optimization. First, a large portion of the parameter space is discrete. Thus, a finite step size can cover the entire parameter space in these subspaces. Second, the parameter spaces are low dimensional (maximally 4D), allowing densely spaced samples in each parameter axis. Third, compared to random search and Bayesian optimization, a grid search is deterministic and, thus, provides reproducibility. Further, as the grid is predetermined (compared to Bayesian

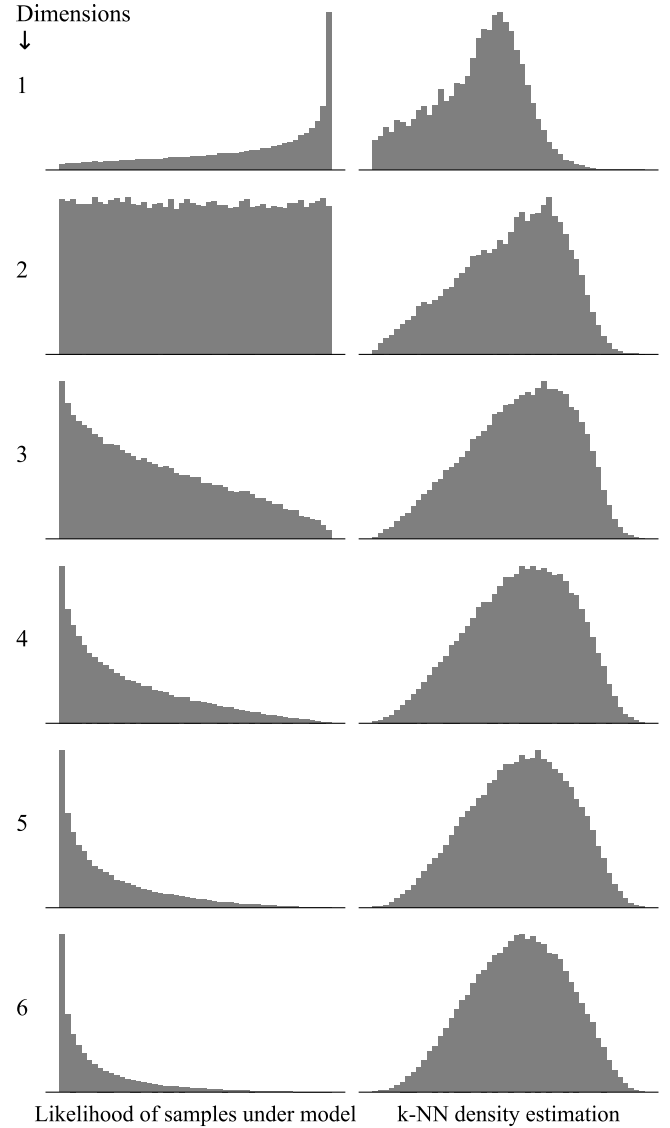


Fig. B.8. Distribution of point-wise density estimates of samples drawn from an N -dimensional Gaussian. The left column shows the likelihood of individual samples. It provides an assessment of the local density under the true model. The right column shows point-wise densities estimated via the k -NN technique. Increasing the dimensionality, N , from one to six (top to bottom row), the distribution of point-wise densities approaches a Gaussian.

optimization), it allows the computation to be fully parallelized, thus guaranteeing dense sampling in reasonable times.

To evaluate the performance of a clustering model (i.e., parameter choice), a range of metrics are used, including both clustering validation metrics (such as NMI, AMI, and ARI) and classification metrics (such as recall, precision, accuracy, balanced accuracy, and MCC). The best-performing model is determined by selecting the model with the highest score across these metrics. Since all of these metrics report on slightly different model summaries, some scores might show large outlying (either abnormally high or low scores) values. To remove their influence, the best parameter choice is determined as maximizing the median across these eight metrics.

While we determine optimal parameters on the “open cluster sample” (see Sect. 4.2.1) by simply running the grid search, the “compact cluster sample” (see Sect. 4.2.2) is a realization

³⁸ The Gaussian is chosen to have zero mean and identity covariance matrix

of a random effects model, making the model parameters themselves stochastic variables. In the latter case, we aim to find expected values for respective clustering parameters across ten realizations of the compact cluster sample. Optimal parametrization is expected to fluctuate between data realizations. However, the compact cluster sample is a single cluster region to which a unique parametrization should be applied. Thus, after searching optimal parameter sets across the ten individual data samples, we took the median (to reduce the effect) of the resulting parameters and reran the algorithm on every sample to obtain the clustering score reported in Table 2.

In the following subsections, we describe algorithm-specific parameter choices of the employed grid search. We also search optimal values for each algorithm for the velocity scaling factor. We adopt values discussed in Appendix B.3.

DBSCAN

DBSCAN has two main parameters, `epsilon` and `min_samples`. The parameter `epsilon` describes a neighborhood radius; in particular, it is the maximum distance within which two samples are considered neighbors. The parameter `min_samples` denotes the minimum number of samples required in an `epsilon`-neighborhood to be considered a cluster (or specifically a core point that seeds a cluster).

Since we normalize the velocities to the spatial subspace XYZ we can treat `epsilon` as a distance in parsecs. Together with the given minimum number of points, it defines a minimum density needed to be considered a cluster. We search for optimal results within a range of `epsilon` $\in [2, 25]$ pc with a step size of 0.5 pc. At the same time, we vary the minimum number of samples in the following range: `min_samples` $\in [4, 40]$ with a step size of 2.

We find optimal parameters for the open cluster sample to be (`epsilon`, `min_samples`, c_v) = (9.5, 6, 5.95). For the compact cluster sample, we find an optimal solution with (`epsilon`, `min_samples`, c_v) = (9.0, 20, 5.95)

HDBSCAN

HDBSCAN has three main parameters, `min_cluster_size`, `min_samples`, and `cluster_selection_method`. Intuitively, `min_cluster_size` determines the smallest cluster sizes that HDBSCAN considers. Which points are still associated with a cluster is determined by `min_samples`. By increasing `min_samples`, clusters are progressively forced into denser areas leaving more points to be declared as noise. The parameter `cluster_selection_method` determines how clusters are selected from the cluster tree hierarchy (see Appendix B.1 for a detailed discussion).

We search for optimal results within a range of `min_cluster_size` $\in [20, 100]$ with a step size of 2. At the same time, we vary the minimum number of samples `min_samples` in the same range and step size while requiring `min_cluster_size` \geq `min_samples`³⁹. The selection method parameter can take the following values: `cluster_selection_method` \in ["leaf", "EOM"].

We find optimal parameters for the open cluster sample to be (`min_cluster_size`, `min_samples`, `cluster_selection_method`, c_v) = (60, 60, "EOM", 8.51). For the compact cluster sample, we find an opti-

mal solution with (`min_cluster_size`, `min_samples`, `cluster_selection_method`, c_v) = (24, 18, "EOM", 5.95).

Appendix C: Projected velocities

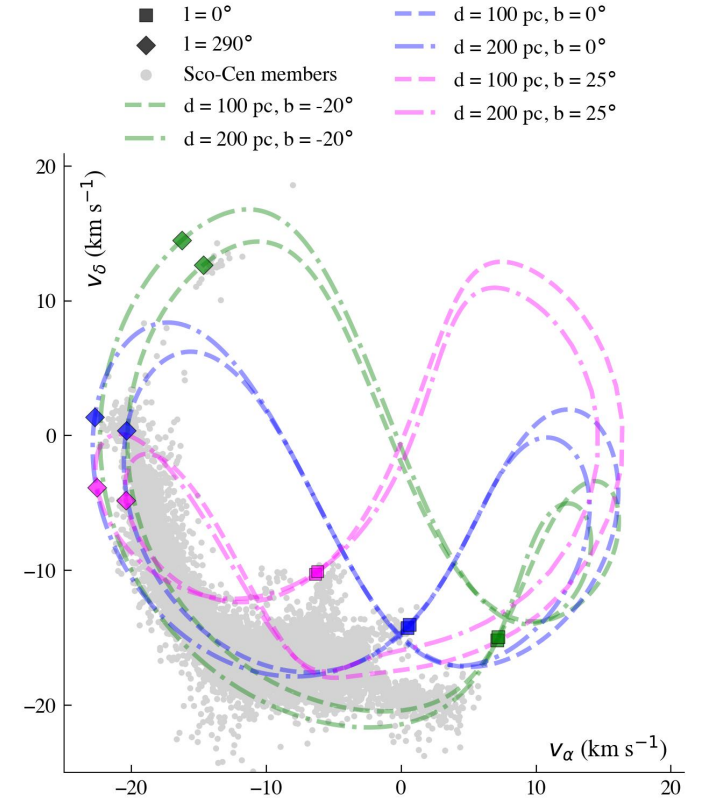


Fig. C.1. Tangential velocities in the v_α/v_δ plane, showing the theoretical locations of sources with circular Galactic orbits and LSR velocities. Six different cases are shown, while each line represents sources at all l positions. The six cases are for two different distances (100 pc, dashed lines; 200 pc, dash-dotted lines), and for three different b positions ($b = -20^\circ$, green; $b = 0^\circ$, blue; $b = 25^\circ$, magenta). The indicated longitude positions at $l = 0^\circ$ (box symbols) and $l = 290^\circ$ (diamond symbols) roughly mark the eastern and western borders of Sco-Cen. The SigMA-selected Sco-Cen members are shown as gray dots (without stability cut). See also Fig. 13 for a separation of the clusters and for the $v_{\alpha, \text{LSR}}/v_{\delta, \text{LSR}}$ plane.

The reflex motion of the Sun influences how the observed tangential velocities are distributed in v_α/v_δ space. Figure C.1 shows the theoretical positions of objects if they follow a circular orbit around the Galactic center at 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 LSR velocity from Schönrich et al. (2010). The projected motions are given for all Galactic longitude (l) positions at two distances (d) of 100 pc and 200 pc and at three Galactic latitudes (b) of -20° , 0° , and 25° . These d and b ranges encompass the Sco-Cen region, which reaches from about $l = 0^\circ$ to 290° . The members of Sco-Cen within the selected SigMA clusters are plotted as gray dots in Fig. C.1.

Overall, the young stellar clusters in Sco-Cen seem to roughly follow expected motions in our Galaxy, assuming they

³⁹ This is an intrinsic requirement of the algorithm.

follow the LSR velocity. The figure additionally highlights the issues with the projected tangential velocity plane v_α/v_δ , which is a function of position in the sky and the Sun's motion. Very nearby stellar clusters, like in nearby young local associations, could cover large areas in the sky and consequently also in the tangential velocity plane while being simultaneously confined in 3D velocity space (UVW). This effect can be reduced if computing the tangential velocities relative to the LSR $v_{\alpha,\text{LSR}}/v_{\delta,\text{LSR}}$, which eliminates the influence of the reflex motion of the Sun (see also Sect. 2 and Appendix A). A comparison of the two velocity spaces (v_α/v_δ and $v_{\alpha,\text{LSR}}/v_{\delta,\text{LSR}}$) is shown in Fig. 13.

Appendix D: The *Gaia* DR3 CMD

In this section we first describe our procedure to estimate the fraction of possible contaminants from older stellar populations (or field stars) in the SigMA-selected Sco-Cen sample using the *Gaia* CMD, and second, we estimate the fraction of substellar objects (brown dwarf candidates).

D.1. Estimating the contamination from older sources

Figure 14 in Sect. 5 shows a *Gaia* CMD using the magnitudes from the *Gaia* DR3 passbands G , G_{BP} , and G_{RP} . Not all sources have detections in all three passbands; within the SigMA-selected Sco-Cen members, 12,724 (97%) sources have an entry in all three passbands. The absolute magnitude M_G is calculated with the distance modulus using the inverse of the parallax as distance. To estimate how many SigMA-selected Sco-Cen members are consistent with the expected ages ($\lesssim 20$ Myr) and which sources could be contaminants from older populations (or field stars), we first need to apply photometric quality criteria. Inferior photometric quality mostly affects the fainter low-mass sources and shifts them to the left in the *Gaia* CMD. We used the magnitude errors and the photometric flux excess factor C , which are defined as follows:

$$\begin{aligned} G_{\text{err}} &= 1.0857/\text{phot_g_mean_flux_over_error} \\ G_{\text{RP,err}} &= 1.0857/\text{phot_rp_mean_flux_over_error} \\ G_{\text{BP,err}} &= 1.0857/\text{phot_bp_mean_flux_over_error} \\ C &= \text{phot_bp_rp_excess_factor} = (I_{\text{BP}} + I_{\text{RP}})/I_G \\ C^* &= \text{corrected } C. \end{aligned} \quad (\text{D.1})$$

The flux excess factor, C (Evans et al. 2018; Riello et al. 2021), gives the flux excess in the $G_{\text{BP}} - G_{\text{RP}}$ color relative to the G band flux. It is recommended to use the corrected C (denoted as C^*) as given in Riello et al. (2021), reducing color dependence. Using these parameters, we applied the following quality criteria to the photometry:

$$\begin{aligned} G_{\text{err}} &< 0.007 \text{ mag} \\ G_{\text{RP,err}} &< 0.03 \text{ mag} \\ G_{\text{BP,err}} &< 0.15 \text{ mag} \\ C^* &< 0.3. \end{aligned} \quad (\text{D.2})$$

These cuts reduce the number of Sco-Cen members from 12,724 with complete photometric information to 11,162 (leaving 88%), mainly reducing the number of the fainter low-mass stars. If not applying any quality criteria, many sources would be shifted toward older ages only because of unreliable *Gaia* colors.

Figure 14 shows two isochrones. The dashed line is a 25 Myr isochrone from Baraffe et al. (2015) (BHAC15⁴⁰) for *Gaia*

Table D.1. Overview of the contamination fraction from older stars as estimated with different photometric quality criteria.

Used Cuts	All	Young	Old
No Stability cut			
In G , G_{BP} , G_{RP} (no cuts)	12,724	11,122 (87.4%)	1,602 (12.6%)
Eq. D.3 (looser cut)	11,906	10,700 (89.9%)	1,206 (10.1%)
Eq. D.2 (used in Fig. 14)	11,162	10,402 (93.2%)	760 (6.8%)
Eq. D.4 (stricter cut)	9,636	9,227 (95.8%)	409 (4.2%)
Stability > 11%			
In G , G_{BP} , G_{RP} (no cuts)	11,213	10,211 (91.1%)	1,002 (8.9%)
Eq. D.3 (looser cut)	10,621	9,873 (93.0%)	748 (7.0%)
Eq. D.2 (used in Fig. 14)	10,014	9,607 (95.9%)	407 (4.1%)
Eq. D.4 (stricter cut)	8,692	8,528 (98.1%)	164 (1.9%)

Notes. In parentheses we give the fraction of young and old stellar candidates relative to the number of sources after the individual photometric error cuts, given in column “All.” The top four rows show the estimates using sources with no stability cut, and the bottom four rows show the estimates using a cut at stability > 11%. We stress that this is not a generally suggested cut to get cleaner samples, since the maximum stability varies per cluster.

DR3/EDR3 passbands. The solid line is a 25 Myr isochrone from PARSEC⁴¹ for *Gaia* DR3/EDR3 passbands (e.g., Bressan et al. 2012; Chen et al. 2014, 2015; Marigo et al. 2017; Riello et al. 2021), including the upper-main-sequence (UMS), which is missing in the BHAC15 models. We used both models since BHAC15 models deliver a better representation of low-mass stars.

To get a measure for the contamination from older sources (older than the expected ~ 20 Myr), we select sources to the left of the two 25 Myr isochrones, allowing for random scatter around the 20 Myr isochrone (in particular young stars often show higher variability than main-sequence stars). Additionally, we did not consider sources at the UMS since there the trend reverses (younger sources are to the left of the UMS). Hence, we applied a cut at $M_G > 3$ mag, only including fainter sources to select older stellar candidates.

The combined conditions deliver 760 candidate contaminants (and 10,402 young Sco-Cen candidates) out of 11,162 sources with applied photometric quality criteria. This is about 6–7% possible contaminants from older populations within the SigMA clusters. Considering the chosen borders, we stress that this separation can only be seen as a rough estimate. In particular, we did not consider any possible contaminants in the UMS regime, where it is more difficult to distinguish young members from older stellar populations. Moreover, the chosen isochrone models have intrinsic uncertainties, and any change in metallicity or extinction is ignored in our test. Finally, we only examine sources from the CMD in Fig. 14 with the applied photometric quality criteria from Eq. (D.2). Hence, we cannot make any statement for sources with inferior photometric quality, which often suffer from higher astrometric uncertainties and which could be shifted in the CMD space. Consequently, such sources could also have a higher probability of having a wrong cluster membership solely due to the generally larger scatter in various dimensions.

To better understand the influence of the quality criteria, we repeat the selection of old star contamination with different photometric quality cuts. First, we consider the case of applying no cuts, using the 12,724 sources with entries in all three *Gaia* bands, delivering a contamination fraction from older sources of

⁴⁰ <http://perso.ens-lyon.fr/isabelle.baraffe/BHAC15dir/>

⁴¹ <http://stev.oapd.inaf.it/cgi-bin/cmd>; assuming solar metallicity (metal fraction $z = 0.0152$) and no extinction.

about 13%. Next, we applied somewhat looser and also stricter quality criteria than given in Eq. (D.2) as follows, first showing the looser cuts (D.3) and then the stricter cuts (D.4):

$$\begin{aligned} G_{\text{err}} &< 0.01 \text{ mag} \\ G_{\text{RP,err}} &< 0.045 \text{ mag} \end{aligned} \quad (\text{D.3})$$

$$\begin{aligned} G_{\text{BP,err}} &< 0.25 \text{ mag} \\ C^* &< 0.5 \end{aligned}$$

$$\begin{aligned} G_{\text{err}} &< 0.004 \text{ mag} \\ G_{\text{RP,err}} &< 0.015 \text{ mag} \\ G_{\text{BP,err}} &< 0.05 \text{ mag} \\ C^* &< 0.3. \end{aligned} \quad (\text{D.4})$$

With the looser cuts, we get a contamination fraction of about 10%, and with the stricter cuts about 4% (see overview in Table D.1). It gets clear that the fraction of sources to the left of the chosen isochrones decreases significantly (from about 13% to 4%) when using superior photometry, which indicates that many sources indeed get erroneously shifted to older ages when not considering the influence of photometric uncertainties. In conclusion, we estimate the contamination fraction from older populations or field stars to be between 4–10%.

Additionally, we investigate the influence of the membership stability as delivered by the SigMA algorithm (Sect. 3.6). Figure 8 shows the influence of different stability cuts on the “old star contamination fraction” when using the quality criteria from Eq. (D.2). It can be seen that for low stability ($\leq 11\%$) there is also a significant increase in the contamination fraction. When only using sources with stability $> 11\%$, we would get a old star contamination fraction in the range of about 2–9% for the cases of strict to no photometric cuts (see Table D.1). Hence, 2–4% can be considered as the lower limit for contamination from older stellar populations (or field stars), while about 10% is likely the upper limit. We conclude that the majority of the SigMA-selected Sco-Cen members (likely more than 90%) are sources with young Sco-Cen-typical ages and therefore likely not contaminants from interloping older populations or field stars. The stability can be used to get more reliable members, while the stability cut needs to be decided individually per cluster.

D.2. Estimating the fraction of substellar sources

To get an estimate of substellar sources in our sample (brown dwarf candidates), we used a $0.08 M_{\odot}$ isomass line in Fig. 14 (right panel), which is extracted from BHAC15 models using ages from 0.5–30 Myr. We select sources below $0.08 M_{\odot}$, which is defined as the approximate hydrogen-burning limit (e.g., Baraffe et al. 1998; Burrows et al. 2001; Freytag et al. 2010, 2012; Dieterich et al. 2014). The uncertainties at the low-mass regime make this selection only a rough estimate, in combination with the uncertainties of the stellar models (e.g., unknown

metallicity, neglected extinction, different models give different results). Additionally, all the uncertainties mentioned above in Sect. D.1 (e.g., chosen error or stability cuts) should be considered.

Using a cut $0.08 M_{\odot}$ and the quality criteria from Eq. (D.2) we find that there are 1946 out of 10,402 (18.7%) potential substellar sources when considering only the younger sources from the middle panel in Fig. 14. This selection indicates a fraction of substellar objects of about 19% within the SigMA clusters. If applying less strict error cuts (no cuts or Eq. (D.3)), the fraction stays at about 19%, and if applying more strict error cuts (Eq. (D.4)) the fraction decreases to about 12%. This is expected since more conservative photometric error cuts affect in particular faint sources. Changing the stability criteria does not influence these different fractions significantly, since sources both in the stellar and substellar regime seem to be affected almost equally. Concluding, we estimate that there are about 19% of brown dwarf candidates in the SigMA-selected Sco-Cen sample, which can be considered as an upper limit (if correcting for extinction would likely deliver a lower fraction).

Appendix E: Auxiliary tables and figures

In Table E.1 we give an overview of the contents of the source catalog containing all Sco-Cen members as selected in this work, including labels for cluster membership. The full version of the table is available online as a machine-readable version.

We provide three additional tables, giving an overview of the literature comparisons between the SigMA clusters and other Sco-Cen samples. In Table E.2 we compare to Damiani et al. (2019) and Žerjal et al. (2023), in Table E.3 to Kerr et al. (2021), and in Table E.4 we compare to Squicciarini et al. (2021), Miret-Roig et al. (2022a), and Briceño-Morales & Chanamé (2023). More details on the comparisons can be found in Sect. 5.2.

Finally, we provide additional figures of the position and velocity spaces of the individual 37 SigMA clusters. This allows a better appreciation of the individual cluster source distribution in each parameter space. The Figs. E.1–E.5 are constructed as follows. The cluster names are given in the left panel (l, b panel) of each row. Each column shows one of the six different parameter spaces for one cluster. The parameter spaces are the same as in Figs. 10–13, namely l versus b (deg), X versus Y (pc), X versus Z (pc), Y versus Z (pc), v_{α} versus v_{δ} (km s $^{-1}$), and $v_{\alpha,LSR}$ versus $v_{\delta,LSR}$ (km s $^{-1}$). These axes labels are given at the top of each column. We note that Col. 5 (tangential velocities) shows a larger velocity range than Col. 6 (tangential velocities relative to the LSR), where the clusters actually occupy a smaller velocity space and hence show a smaller velocity dispersion. All SigMA-selected Sco-Cen members are plotted in gray in all panels and the given cluster is over-plotted with red dots. See also Figs. 10–13 for an alternative view of the 37 SigMA clusters, and the interactive 2D and 3D versions [online](#).

Table E.1. Catalog of the 13,103 Sco-Cen members labeled for cluster membership as identified with SigMA.

Column name	Unit	Column description
dr3_source_id		The source ID from <i>Gaia</i> DR3.
SigMA		SigMA membership label for each source, as defined in Table 3.
stability		Membership stability of each source between 0–100%
distance	pc	Distance derived from inverse of the parallax
e_d_upper	pc	Upper 1σ distance uncertainty determined from the 68.3 percentile of the sampled d distribution
e_d_lower	pc	Lower 1σ distance uncertainty determined from the 68.3 percentile of the sampled d distribution
X	pc	Heliocentric Galactic Cartesian coordinate, X-axis grows positive toward the Galactic center
Y	pc	Heliocentric Galactic Cartesian coordinate, Y-axis grows positive in direction of Galactic rotation
Z	pc	Heliocentric Galactic Cartesian coordinate, Z-axis grows positive toward the Galactic North-pole
e_X_upper	pc	Upper 1σ X uncertainty determined from the 68.3 percentile of the sampled X distribution
e_Y_upper	pc	Upper 1σ Y uncertainty determined from the 68.3 percentile of the sampled Y distribution
e_Z_upper	pc	Upper 1σ Z uncertainty determined from the 68.3 percentile of the sampled Z distribution
e_X_lower	pc	Lower 1σ X uncertainty determined from the 68.3 percentile of the sampled X distribution
e_Y_lower	pc	Lower 1σ Y uncertainty determined from the 68.3 percentile of the sampled Y distribution
e_Z_lower	pc	Lower 1σ Z uncertainty determined from the 68.3 percentile of the sampled Z distribution
v_alpha	km s ⁻¹	Tangential velocity in the direction of α
v_delta	km s ⁻¹	Tangential velocity in the direction of δ
e_v_alpha_upper	km s ⁻¹	Upper 1σ v_α uncertainty determined from the 68.3 percentile of the sampled v_α distribution
e_v_delta_upper	km s ⁻¹	Upper 1σ v_δ uncertainty determined from the 68.3 percentile of the sampled v_δ distribution
e_v_alpha_lower	km s ⁻¹	Lower 1σ v_α uncertainty determined from the 68.3 percentile of the sampled v_α distribution
e_v_delta_lower	km s ⁻¹	Lower 1σ v_δ uncertainty determined from the 68.3 percentile of the sampled v_δ distribution
v_alpha_LSR	km s ⁻¹	Tangential velocity in the direction of α and relative to the LSR
v_delta_LSR	km s ⁻¹	Tangential velocity in the direction of δ and relative to the LSR
v_ERV	km s ⁻¹	Estimated radial velocity, given as \hat{v}_r in Sect. 3.5.2

Notes. The full machine-readable version of the catalog is available online, while a column overview is given here. We list all relevant derived parameters. Original *Gaia* DR3 parameters can be queried from the *Gaia* Archive by using the dr3_source_id.

Table E.2. Comparing the SigMA clusters with stellar group selections from Damiani et al. (2019) and from Žerjal et al. (2023).

SigMA	Name (SigMA)	Nr ^a	Matches with DDP19 ^b	Matches with ZIC23 ^c
1	rho Oph	535	US-f(3)US-n(316)US-D2(68)N(20)	C-USco(405)E-USco-multi(13)
2	nu Sco	150	US-n(62)US-D2(65)N(1)	C-USco(127)E-USco-multi(1)
3	delta Sco	691	US-f(33)US-n(74)D1(88)D2a(8)D2b(2)US-D2(387)N(6)	G-UCL-East(3)C-USco(489)E-USco-multi(24)
4	beta Sco	285	US-f(58)US-n(17)US-D2(152)N(8)	C-USco(189)E-USco-multi(19)
5	sigma Sco	544	US-f(180)US-n(3)D1(14)D2a(3)US-D2(227)N(7)	G-UCL-East(2)C-USco(105)E-USco-multi(184)
6	Antares	502	US-f(67)US-n(40)D1(19)D2a(5)US-D2(249)N(29)	C-USco(78)E-USco-multi(252)
7	rho Sco	240	US-f(14)US-n(2)D1(159)D2a(3)US-D2(10)N(15)	G-UCL-East(48)C-USco(1)E-USco-multi(7)
8	Sco-Body	373	D1(2)D2a(221)US-D2(34)N(7)	E-USco-multi(291)
9	US-fg	276	D1(188)D2b(1)US-D2(7)N(12)	G-UCL-East(16)E-USco-multi(29)
10	V1062-Sco	1029	UCL-1(554)D1(11)D2a(222)D2b(4)N(10)	D-UCL-V1062-Sco(499)F-UCL-V1062-Sco(228)G-UCL-East
11	mu Sco	54	UCL-1(36)D1(2)D2a(5)	D-UCL-V1062-Sco(7)F-UCL-V1062-Sco(30)
12	Libra-S	71	D1(1)D2a(8)D2b(13)US-D2(32)	E-USco-multi(4)
13	Lup 1-4	226	LupIII(67)D2a(47)D2b(65)N(4)	G-UCL-East(6)T-UCL-West(1)E-USco-multi(109)
14	eta Lup	769	UCL-3(3)D1(549)D2a(43)D2b(10)US-D2(1)N(14)	G-UCL-East(419)E-USco-multi(15)
15	phi Lup	1114	UCL-3(48)D1(627)D2a(62)D2b(148)N(38)	G-UCL-East(652)T-UCL-West(28)E-USco-multi(17)
16	Norma-N	42	D1(1)N(6)	
17	e Lup	516	D1(319)D2a(18)D2b(80)N(5)	G-UCL-East(349)T-UCL-West(15)
18	UPK 606	131	UCL-2(50)D1(2)D2b(57)N(1)	G-UCL-East(54)T-UCL-West(9)
19	rho Lup	246	D1(17)D2b(189)N(2)	A-LCC-North (45)G-UCL-East(10)T-UCL-West(110)
20	nu Cen	1737	UCL-2(2)D1(54)D2a(12)D2b(1270)US-D2(3)N(50)	A-LCC-North (70)G-UCL-East(116)T-UCL-West(790)E-US
21	sig Cen	1805	LCC-1(1)D1(45)D2b(1384)N(43)	A-LCC-North (1077)U-LCC-South(56)T-UCL-West(66)
22	Acrux	394	LCC-1(89)D1(11)D2b(242)N(4)	A-LCC-North (25)U-LCC-South(316)
23	Musca-fg	95	D2b(35)N(2)	U-LCC-South(76)
24	eps Cham	39		U-LCC-South(25)
25	eta Cham	30		U-LCC-South(3)
26	B59	32	D2a(1)N(20)	E-USco-multi(2)
27	Pipe-North	42		E-USco-multi(38)
28	tet Oph	98	D2a(37)US-D2(6)N(2)	E-USco-multi(87)
29	CrA-Main	96		E-USco-multi(2)
31	Sco-Sting	132	D1(22)D2a(1)	
32	Cen-Far	99	D2b(41)N(1)	
35	L134/L183	24		E-USco-multi(16)

Notes. Only those SigMA clusters that have cross-matches with either of the two literature samples are given here. ^aNumber of sources from this work, for a direct comparison with the number of cross-matches given in brackets in Cols. 4–5. ^bThe DDP19 group shortcuts are given for eight compact clusterings (UCL-1, UCL-2, UCL-3, Lupus 3, LCC-1, US-far, US-near), four diffuse populations (D1, D2a, D2b, US-D2), and sources that have not been assigned to any of these groups (N). The number in brackets gives the number of matches with the respective SigMA cluster. See details in Sect. 5.2.1. ^cZIC23 report eight subgroups in Sco-Cen (C, E, D, F, G, T, A, U) and two additional older groups (H, I; IC 2602 and Platais 8), while there are no matches of H or I with the 37 SigMA clusters. Again, the number of matches is given in brackets. The groups contain the following numbers of sources in ZIC23: C-USco 1432, E-USco-multi 1483, D-UCL-V1062-Sco 506, F-UCL-V1062-Sco 273, G-UCL-East 1713, T-UCL-West 1057, A-LCC-North 1234, and U-LCC-South 487. See further details in Sect. 5.2.5.

Table E.3. 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	535	22(308)	17(272)	I(109)	US-I-rho Oph
2	nu Sco	150	22(91)	17(90)	E(54)	US-E
3	delta Sco	691	22(414)	17(378)	F(1)H(102)I(1)	US-H
4	beta Sco	285	22(167)	17(137)	G(29)	US-G
5	sigma Sco	544	22(296)	17(248)	A(1)C(17)D(22)	US-C/D
6	Antares	502	22(292)	17(239)	A(1)B(11)C(1)F(24)	US-B/F
7	rho Sco	240	22(128)	17(64)	A(9)	US-A
8	Scorpio-Body	373	22(193)	16(12)17(45)		EOM-16/US
9	US-foreground	276	22(111)	13(30)		EOM-13
10	V1062-Sco	1029	22(503)	14(20)15(347)		LowerSco/EOM-14
11	mu Sco	54	22(28)	15(23)		LowerSco
12	Libra-South	71	22(23)			
13	Lupus-1-4	226	22(143)	12(102)	A(46)B(14)	Lupus-IV/III
14	eta Lup	769	22(411)	9(15)22(102)23(6)		EOM-9/22/23
15	phi Lup	1114	22(391)	17(4)19(10)23(6)		EOM-19/23
16	Norma-North	42	22(4)			
17	e Lup	516	22(257)	11(1)20(76)21(8)		EOM-20
18	UPK606	131	21(1)22(53)	11(32)		UPK606
19	rho Lup	246	22(123)	21(2)25(10)		EOM-25
20	nu Cen	1737	22(573)	11(3)21(2)24(108)26(14)27(2)		EOM-24
21	sig Cen	1805	22(987)	25(1)26(26)27(421)	C(4)D(12)E(48)	EOM-26/LCC-D/E
22	Acrux	394	22(258)	27(208)	B(1)C(96)	LCC-C-Crux S
23	Musca-foreground	95	22(65)	27(46)	B(16)	LCC-B
24	eps Cham	39	22(23)	27(20)	A(17)	LCC-A-eps Cha
25	eta Cham	30	22(18)	18(17)		eta Cha
26	B59	32	22(14)	6(13)		Pipe
27	Pipe-North	42	22(19)			
28	tet Oph	98	22(49)	10(28)		Theia67
29	CrA-Main	96	22(53)	8(52)		CrA
30	CrA-North	351	22(207)	7(1)8(195)		CrA
31	Scorpio-Sting	132	22(62)	7(11)		EOM-7
32	Centaurus-Far	99	21(36)	3(30)		Cen-South
33	Chamaeleon-1	192	21(101)	1(101)		Cha-1
34	Chamaeleon-2	54	21(30)	2(30)		Cha-2
35	L134/L183	24	22(6)			
36	Oph Southeast	61	4(20)			Oph Southeast

Notes. Only those SigMA groups that have matches with KRK21 are given here, while Oph-North-Far is the only SigMA cluster without matches. ^aNumber of sources from this work, for a direct comparison with the number of cross-matches as given in brackets in Cols. 4–6. ^bCol. 4 lists the TLC group labels if there are matches with SigMA, with the number of cross-matches in brackets. There have been only matches with the TLC groups 4, 21, and 22. ^cCol. 5 lists the EOM group labels if there are matches with SigMA, with the number of cross-matches in brackets, while each EOM represents a subclustering within a lower level TLC group. ^dThe letters in Col. 6 correspond to LEAF subgroups, with the number of cross-matches in brackets, while a LEAF group represents a subclustering within a lower-level EOM group. Leaves only exist for the EOM groups 12 (Lupus), 17 (US), and 27 (LCC). ^eGroup names from KRK21, if there is a significant overlap with SigMA. Only the (sub)group with the most significant number of cross-matches is given (in a few cases more than one), as apparent from the numbers in brackets in Cols. 5 & 6. See more details in Sect. 5.2.2.

Table E.4. Comparing the SigMA clusters with stellar group selections from Squicciarini et al. (2021), Miret-Roig et al. (2022a), and Briceño-Morales & Chanamé (2023).

SigMA	Name (SigMA)	Nr ^a	Matches with SGB21 ^b	Matches with MR22 ^c	Matches with BMC23 ^d
1	ρ Oph/L1688	535	G1(428)G2(2)G3(1)G4(1)G8(2)D(51)	α Sco(3) δ Sco(9) ν Sco(1) σ Sco(67) ρ Oph(370)	D(184) ρ Oph(225) ν Sco(1) α Sco(1)
2	ν Sco	150	G1(1)G2(110)G6(10)D(23)	β Sco(2) δ Sco(2) ν Sco(110) σ Sco(22)	D(28) ω Sco(6) ν Sco(84) β Sco(3)
3	δ Sco	691	G1(19)G2(2)G3(390)G4(9) G5(50)G6(2)G7(1)G8(2)D(136)	α Sco(27) β Sco(8) δ Sco(410) ν Sco(29) π Sco(38) σ Sco(90) ρ Oph(6)	D(398) ω Sco(59) ν Sco(22) δ Sco(66) α Sco(1) β Sco(1)
4	β Sco	285	G1(1)G4(141)G5(11)G6(46)D(57)	α Sco(2) β Sco(169) σ Sco(70)	D(86) α Sco(26) β Sco(107)
5	σ Sco	544	G1(2)G4(2)G5(104)G8(3)D(377)	α Sco(268) δ Sco(3) π Sco(7) σ Sco(163)	D(254) α Sco(192)
6	Antares	502	G1(13)G3(5)G7(44)G8(33)D(313)	α Sco(290) π Sco(29) σ Sco(52) ρ Oph(37)	D(322) ρ Oph(29) ω Sco(1) ν Sco(1) α Sco(5)UCL(2)
7	ρ Sco	240	D(168)	α Sco(3) π Sco(180)	D(172) π Sco(2) ρ Oph(7) α Sco(2)
8	Scorpio-Body	373		α Sco(2) σ Sco(2)	D(81)UCL(3)
9	US-foreground	276	D(13)	α Sco(1) π Sco(194)	D(102) π Sco(96)
10	V1062-Sco	1029			D(2)
12	Libra-South	71		σ Sco(1)	D(24)
13	Lupus 1-4	226			D(7)
14	η Lup	769		π Sco(12)	D(55)UCL(47)
15	ϕ Lup	1114	D(2)	π Sco(3)	D(9)
20	ν Cen	1737		σ Sco(1)	
28	θ Oph	98			D(7)

Notes. Only those SigMA groups that have cross-matches with one of the three literature samples are given here. ^aNumber of sources from this work, for a direct comparison with the number of matches as given in brackets in Cols. 4–6. ^bThe SGB21 groups (G) are numbered from 1 to 8, and their diffuse population is given with D. 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– ν Sco B; G3– b Sco; G4–HD 144273; G5–HIP 77900; G6–HIP 78968; G7–HIP 79910; G8–HD 146467. See details in Sect. 5.2.6. ^cComparison to the seven groups in US as identified by MR22. Again, the number of cross-matches is given in brackets. See details in Sect. 5.2.7. ^dComparison to the eight clusters and one diffuse population (D) in US as identified by BMC23. Again, the number of cross-matches is given in brackets. See details in Sect. 5.2.8.

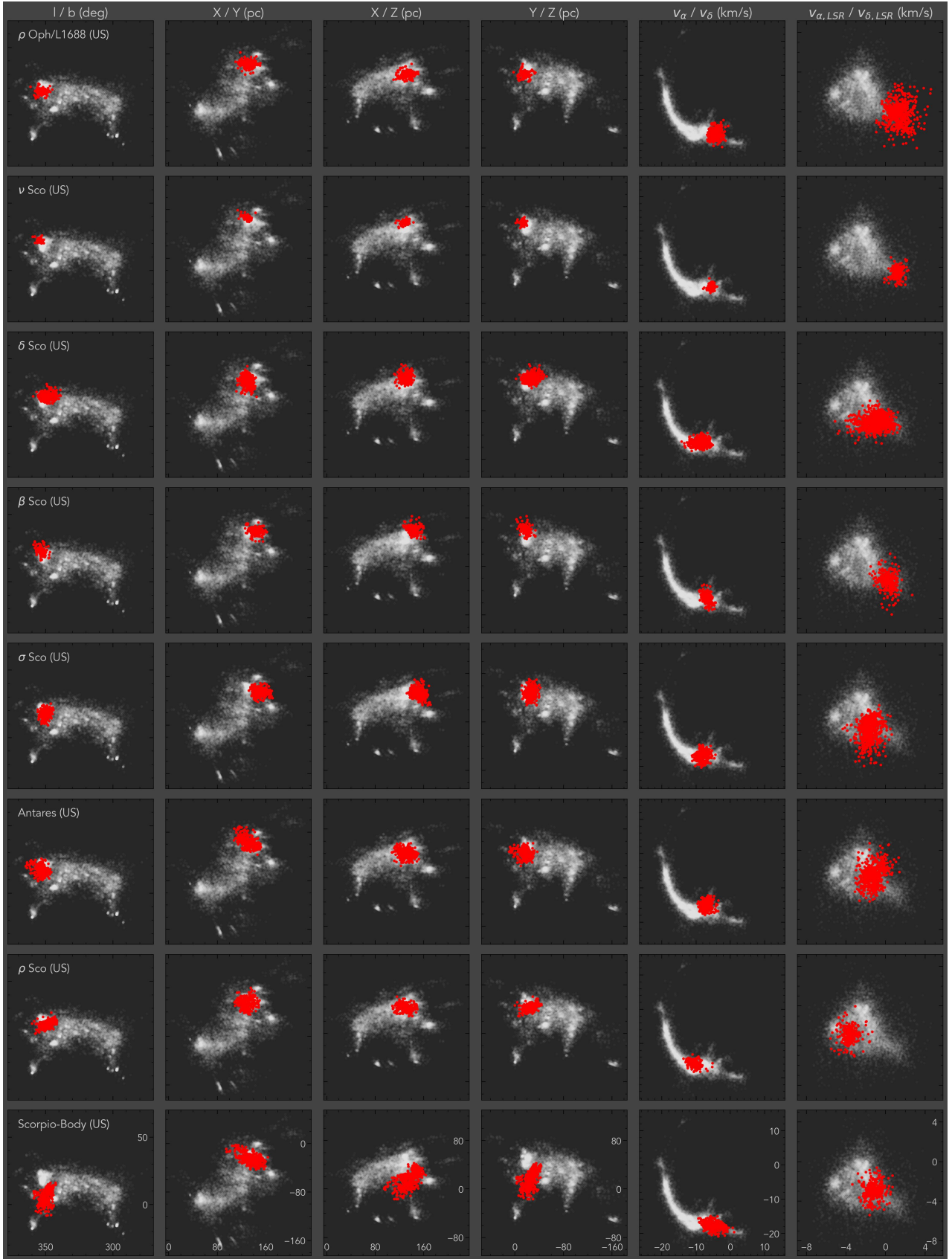


Fig. E.1. Six parameter spaces, with the individual clusters highlighted in red. Shown are clusters SigMA 1–8 (part of US). The gray background sources are all SigMA-selected Sco-Cen members. Cluster names are given in the left panels of each row. The used xy-axes are given as titles at the top of each column. Tick labels are only given in the bottom row. Note that the v_{LSR} space shows a smaller velocity range compared to the tangential velocity space in Col. 5 and hence a lower velocity dispersion. See also Figs. 10–13 and the main text for more details.

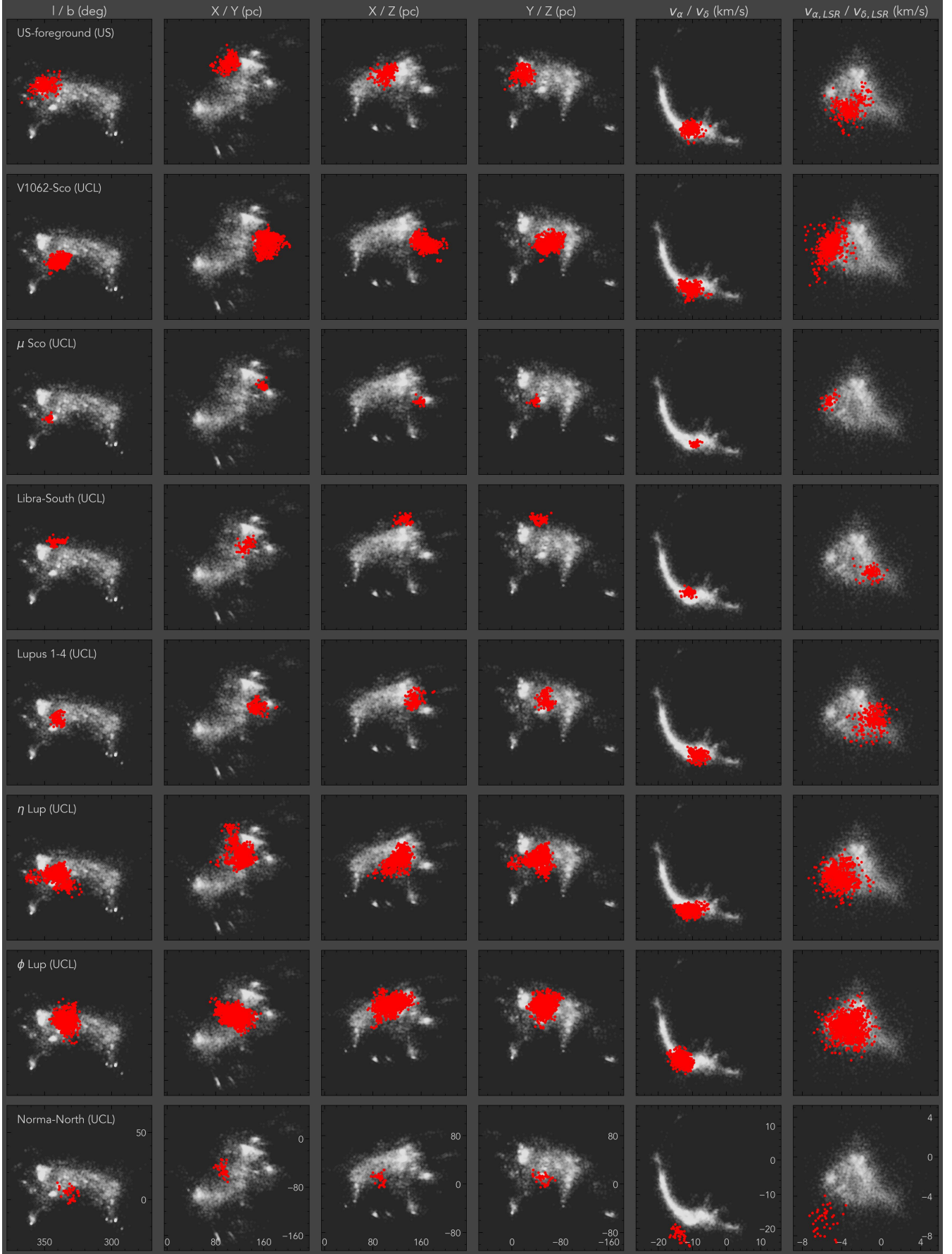


Fig. E.2. Same as Fig. E.1, but for clusters Sigma 9–16 (part of US and UCL).

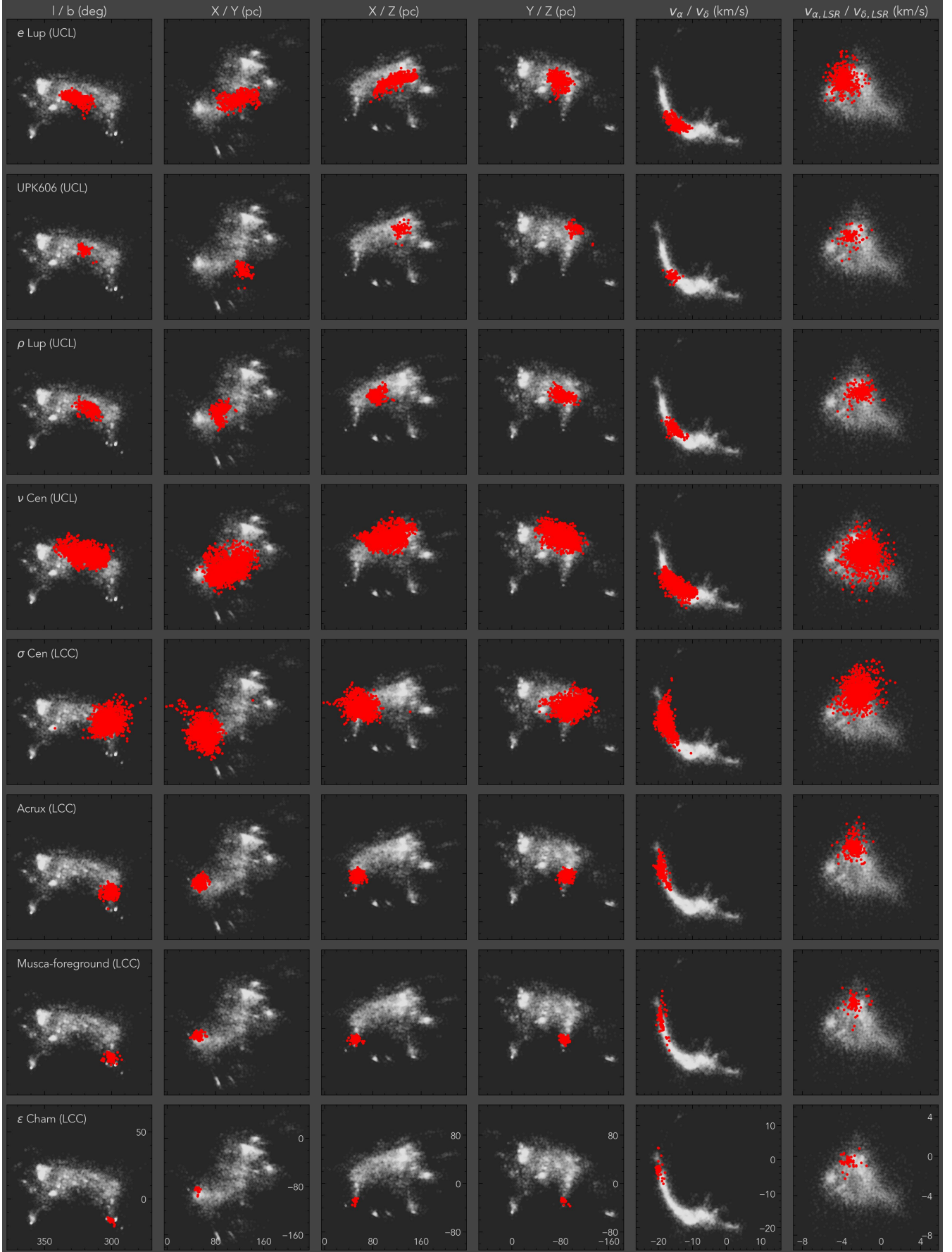


Fig. E.3. Same as Fig. E.1, but for clusters Sigma 17–24 (part of UCL and LCC).

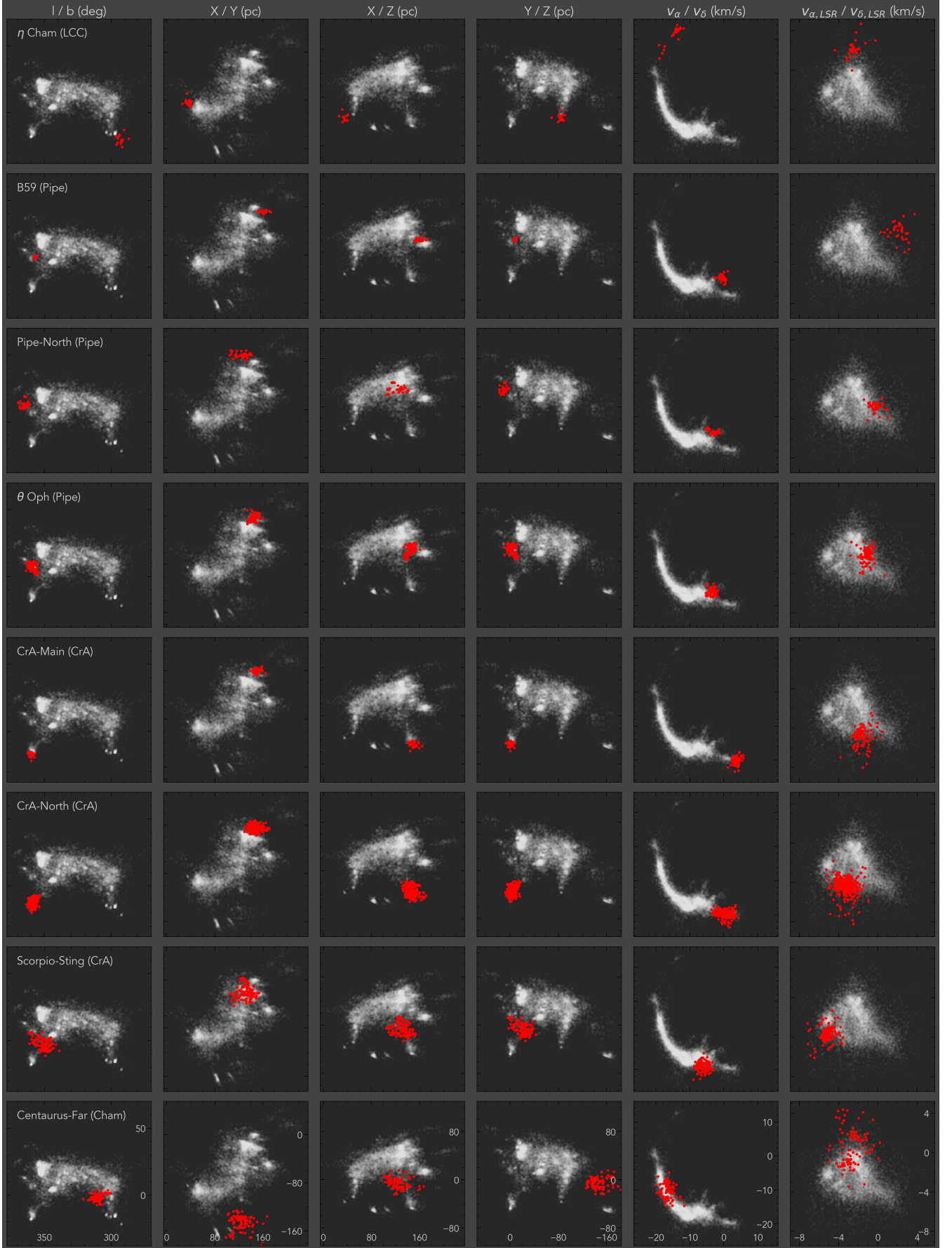


Fig. E.4. Same as Fig. E.1, but for clusters Sigma 25–32 (part of LCC, Pipe, CrA, and Cham).

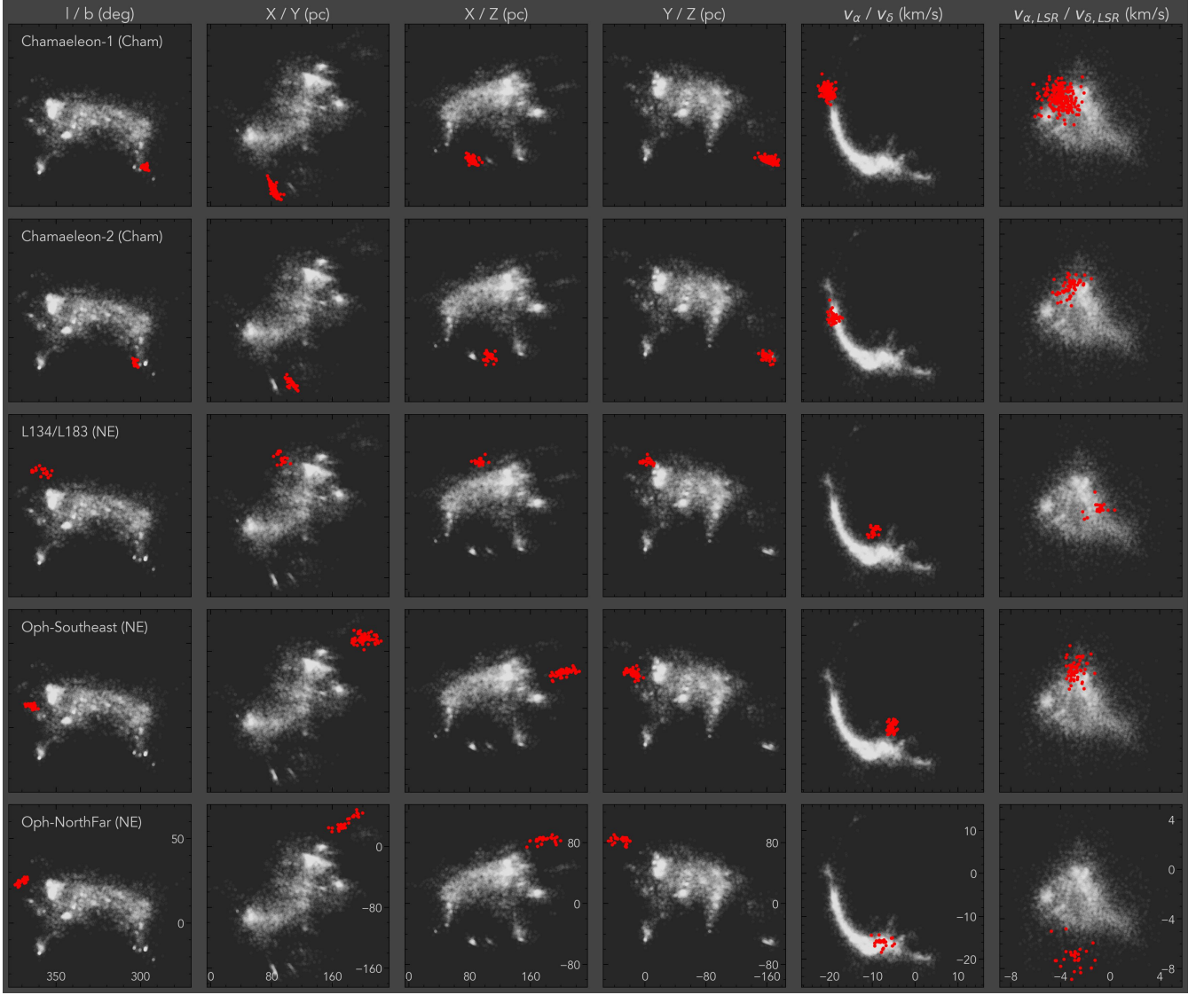


Fig. E.5. Same as Fig. E.1, but for clusters Sigma 33–37 (part of Cham and NE).