SCHOLARLY DEBUT
by Scholarly Review

Winter 2025/2026

Stellar Characteristics of the Solar Neighborhood: Machine Learning Insights for Space Navigation from Gaia Data Release 3

By Ishant Yenamandra, Vishakh Hari, Krish Bahl

AUTHOR BIO

Vishakh Hari is a rising senior at Round Rock High School in Round Rock, TX. He has been interested in astronomy and aerospace/computational engineering for the past three years. He is interested in doing more research in astrophysics and aerospace engineering and plans to pursue engineering after high school.

Krish Bahl is a rising junior at Elkins High School in Missouri City, TX, with a strong interest in astrophysics, aerospace, and mechanical engineering which he hopes to pursue in the future through a career in engineering after high school.

Ishant Yenamandra is a rising junior at Liberal Arts and Science Academy in Austin, Texas and in his personal life, he is heavily interested in computational finance, machine learning, and computer engineering.

ABSTRACT

This study analyzes stars in the solar neighborhood — defined to be a heliocentric sphere with a 500 parsec radius — using stellar data from the Gaia Data Release 3 database. To allow for a feasible dataload with our computing power constraints, we restricted the stars by apparent magnitude, solely including stars within 7.5 <= G <= 13.0, where G is Gaia G-band (apparent) magnitude. We derived our restriction criteria and definition of the solar neighborhood from previous research by Zari et al 2018 and Prisinzano et al 2022. Using the columns for right ascension, declination, parallax, parallax-over-error, G magnitude, BP-RP color, proper motion right ascension, proper motion declination, and radial velocity, we construct interactive 3D spatial and kinematic maps, a luminosity function, and several machine-learning models: Isolation Forest for kinematic anomaly detection, a mass-estimation pipeline (analytic + Random Forest regression), and clustering in combined position–velocity space. Density-based clustering (HDBSCAN) initially returned 49 leaf clusters but with poor separation (silhouette –0.551), motivating a Gaussian Mixture fallback. Scanning k = 2–30 by BIC selected 20 Gaussian components as the preferred segmentation of the 500 pc neighborhood. Kinematic anomaly detection using an Isolation Forest (contamination 0.5%) flags a small, high-velocity subset for follow-up. Insights into local structure and motions may support future stellar navigation systems.

Keywords: *Solar Neighborhood, Computational Astrophysics, Astronomy, Machine Learning, Geometry of Space, Aerospace, Stellar Navigation, Isolation Forest, Anomaly Detection*

SCHOLARLY DEBUT
by Scholarly Review

Winter 2025/2026

INTRODUCTION

The first structured map of the solar neighborhood can be traced back to the 20th century from Kapteyn's model. In 1904, Jacobus Kapteyn constructed the first three-dimensional model of the Milky Way using proper motion studies, star counts, and luminosity magnitudes. Although Kapteyn's original star maps lacked corrections for interstellar dust and showed only approximate distances; they laid the foundation for future improvements in stellar mapping that would become much more accurate in the decades that followed.

The understanding of the solar neighborhood greatly improved in the early 1900s with the creation of the Hertzsprung–Russell diagram (1911–1913), which helped astronomers determine a star's age and stage of life based on its brightness and color. This tool, combined with proper motion and parallax measurements, became key for classifying stars near the Sun and understanding their location in space.

Later, this work was refined significantly by Jan Oort in the 1920s, who introduced a more dynamic framework by incorporating stellar velocities and formulating the concept of Galactic rotation. His analytical derivations, known as Oort constants, were critical in recognizing the Milky Way's disk-like rotation and mass distribution.

In 1989, the European Space Agency launched the Hipparcos satellite: the first space mission designed for measuring star positions from space. Its data, released in 1997, included accurate distances for over 100,000 stars and allowed scientists to map stars and their motions within about 100 parsecs, more precisely than ever before.

However, it was the Gaia mission that dramatically reshaped this field. With the release of Gaia Data Release 2 (DR2) in 2018, astronomers obtained full six-dimensional phase-space information (position, motion, and parallax) for millions of stars. Using this data in 2018, Eleonora Zari et al. produced a three-dimensional map of the solar neighborhood focused on young stellar populations — both Upper Main Sequence and Pre-Main Sequence stars. Their methodology applied color–magnitude filters and extinction corrections, revealing well-known objects in space such as Scorpius-Centaurus and Orion. Their study stated: We study the three dimensional arrangement of young stars in the solar neighbourhood using the second release of the Gaia mission (Gaia DR2) and we provide a new, original view of the spatial configuration of the star-forming regions within 500 pc of the Sun. By smoothing the star distribution through a Gaussian filter, we construct three dimensional (3D) density maps for early-type stars (upper main sequence, UMS) (Zari 1).

Soon after, other researchers expanded on this work. In 2018, Cantat-Gaudin et al. — as well as Castro-Ginard et al. et al. in 2020 — applied similar clustering techniques to identify open clusters, using both positional and photometric information. These studies showed that Gaia DR2 could be used not only to find known star groups but also to discover new ones.

With the release of Gaia Early Data Release 3 (Gaia EDR3) in 2020, data quality improved significantly, especially in terms of photometry and calibration. In 2021, Kerr et al. used the HDBSCAN clustering algorithm to detect 27 young stellar groups within 333 parsecs. In 2022, Prisinzano et al. built on earlier work by applying unsupervised machine learning (specifically the DBSCAN algorithm) to identify over 354 star-forming regions and more than 124,000 young stellar objects (YSOs) within 1.5 kpc. As they described, "Gaia EDR3 provides, for the first time, the opportunity to systematically detect and map, in the optical bands, the low-mass populations of star-forming regions in the Milky Way. (Prisinzano 2022)".

Our research builds directly on the recent breakthroughs by Zari et al. In 2018 and Prisinzano et al. In 2022, extending their methods with Gaia DR3 data to produce comprehensive 3D computational machine learning models of the Solar Neighborhood within 500 parsecs—focusing on luminosity, mass, spatial structure, spectral types, and velocity distributions.

DATA AND METHODS

Data and quality cuts

We queried stellar data from the Gaia Data Release 3 (Gaia DR3) database, limiting our selection of stars within 500 parsecs (pc) of the Sun, our definition of the solar neighborhood. The query ensured each star had reliable measurements, including positional data **(ra, dec, parallax)**, proper motion (**pmra, pmdec**), radial velocity, and photometric data (phot_g_mean_mag, bp_rp).

Features and transformation

Feature	*Gaia DR3 Input*	*Units*
Right ascension	ra	deg
Declination	dec	deg
Parallax	parallax	mas
Distance	parallax	pc
Proper Motion	pmra	mas/yr
Proper Motion in declination	pmdec	mas/yr
Total Proper Motion	Pmra, pmdec	mas/yr
Radial Velocity	radial_velocity	km/s
Tangential Velocity	Parallax, total proper motion	km/s
3D Velocity Vector	radial_velocity, pmra, pmdec, ra,	km/s

	dec, parallax	
Cartesian position	ra, dec, distance	pc
Apparent Magnitude	phot_g_mean_mag	mag
Absolute Magnitude	phot_g_mean_mag, distance	mag
Spectral type	bp_rp	categorical

Photometry: M_G from apparent magnitude and parallax

Kinematics: First, we calculated the three-dimensional velocity vectors for each star using radial velocity and proper motion data. Using the relationship

$$v_t = 4.74 \ * \ \mu / \varpi$$

where v_t is the tangential stellar velocity, μ is the proper motion, and ϖ is the parallax, we computed the tangential velocities for all stars in the neighborhood. Radial velocity was directly obtained from Gaia measurements. All the velocity components were then transformed into a Cartesian coordinate frame (*vx*, *vy*, and *vz)* using astrometric conversion factors and dimensional analysis. Then, each velocity vector was paired with a spatial position, as shown in Figure 1, to form a full three-dimensional phase-space representation of both the stellar positions and motion.

The tail of each vector is located at the star's Cartesian position, and the direction and length of each line indicate the star's velocity through space. The vectors are also color-coded by estimated spectral type derived from the BP-RP color index, to explore potential connections between stellar type and motion.

Methods

We generate interactive 3D visualizations of (i) spatial distribution (points scaled by stellar radius proxy; color-coded by estimated spectral type from BP–RP), and (ii) velocity vector fields anchored at stellar positions.

Kinematic Anomaly Detection

In order to identify stars traveling with unusual kinematic behavior, we implemented an unsupervised machine learning anomaly detection algorithm known as the Isolation Forest (Liu et al.). This model recursively partitions data to isolate stellar velocities that are anomalous. For this model, we focused on stellar velocity data derived from the six-dimensional phase space: *x*, *y*, *z*, *vx*, *vy*, and *vz*. We initialized the Isolation Forest with 100 estimators (trees) and set the contamination parameter to 0.005, representing an anomaly rate of 0.5%. We selected a low contamination rate (0.005) to reflect the expectation that genuinely unusual kinematics comprise only a small fraction of nearby stars; this setting focuses the model on the extreme tail of the velocity distribution rather than labeling broad disk substructure as anomalous. This “contamination parameter” determines the threshold for classifying a stellar velocity as an anomaly. The model was trained on the set of velocity vectors shown in Figure 2, and it assigned a stellar anomaly score, ranging from 1(most anomalous) to +1(least anomalous), to each star in the neighborhood. The Isolation Forest model most easily isolated stars with lower anomaly scores; the model classified the top 0.5% of stars with the most anomalous scores as kinematic outliers. We then visualized this model as an interactive three-dimensional anomaly plot, shown in Figure 2, where normal stars are blue and anomalous stars are red. Each red point, therefore, represents stars with velocities significantly different from the local stellar population.

Mass Estimation

We begin by computing an analytic estimate of the stellar mass using the observed photometry and color. We calculate G-band magnitude using:

$$m_G + 5log_{10}(\frac{\varpi}{1000}) + 5$$

Where m_G = phot_g_mean_mag and ϖ is the parallax in milliarcseconds. The mass proxy is then taken to be

$$M_{analytic} = 0.8e^{-0.4(bp\_rp)} \cdot 10^{-0.1\ M_G}$$

We clipped the above equation into a plausible mass range of 0.1,1.0 solar masses. When the effective temperature was available, we applied scaling as

$$(\frac{T_{eff}}{5777K})^{0.5}$$

before clipping into the range.

To refine the relation, we trained a Random Forest regressor on the three features (bp_rp, m_G, ϖ) with the analytic masses as the target values. We split the data 80/20 into training and test sets, fit 100 trees to minimize mean squared error, and evaluated the performance through the out-of-sample R^2 value. We then averaged the analytic mass and the estimated prediction to define the “refined mass” for each star:

$$M_{refined} = 0.5M_{analytic} + 0.5M_{estimated}$$

Clustering in 6D

To identify clusters, we combined spatial and kinematic space (x, y, z, v_x, v_y, v_z).

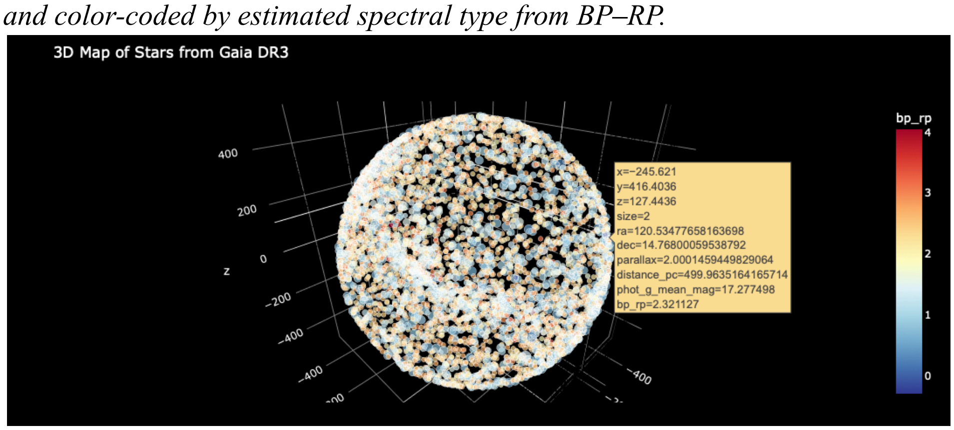
We begin with a MiniBatchKMean with 500 centroids, to compute the minimum centroid distance for each star, and sample with probability inversely proportional to that distance. This subsampling preserves sparse features while reducing computing cost. We use MiniBatchKMeans as a fast way to summarize the full dataset with a fixed number of representative points (“centroids”). With 500 centroids, the algorithm approximates the overall structure of the stellar distribution in phase space without requiring expensive pairwise distance calculations. After fitting, each star is assigned a distance to its nearest centroid; this nearest-centroid distance acts as a simple proxy for how well that region of phase space is represented by the centroids. Stars that are close to a centroid lie in dense, redundant regions of the distribution, while stars that are farther from all centroids are more isolated and tend to lie in low-density or boundary regions. We then convert this distance into a sampling probability and subsample the dataset accordingly. This approach reduces the number of stars that must be processed in downstream analysis while preferentially retaining points that would otherwise be underrepresented by uniform random sampling, helping preserve rare kinematic structures and outliers at substantially lower computational cost.

We standardize positions and velocities separately, multiplying the velocity axes by 0.5 to balance their influence. We scaled the velocity dimensions by 0.5 to reduce the contributions of kinematic difference to the distance metric (by a factor of 0.25 in squared distance). Geometrically, this stretches neighborhoods along velocity axes favoring clusters that are spatially coherent while tolerating moderate velocity dispersion, and mitigating over-fragmentation caused by small variations. The primary clustering algorithm is HDBSCAN configured with a minimum cluster size of 30, samples at 2, and selection epsilon at 0.2. We used the leaf selection method, and HDBSCAN assigns each star either to a cluster label or as noise. We end with a computation of the silhouette score on a random subset to assess true cluster separation.

If the silhouette score falls below 0.25, indicating poorly separated clusters, we invoke a Gaussian Mixture Model (GMM) fallback. We fit standard GMMs for component counts k=2 to 30, each with full covariance and three random initializations, and select the model minimizing the Bayesian Information Criterion (BIC). The chosen GMM then reassigns every star to one of its k Gaussian components. The final cluster labels are stored, yielding an adaptive, statistically grounded partition of the solar neighborhood into coherent stellar groups.

RESULTS

Figure 1. Spatial distribution (500 pc).
Three-dimensional map of stellar positions in the solar neighborhood. Points are scaled by a radius proxy and color-coded by estimated spectral type from BP–RP.



The spatial map reveals non-uniform density with recognizable over-densities and voids consistent with known local structures and continued dominance of faint, red dwarfs. Color coding shows expected gradients in stellar type across features.

Figure 2. Velocity vector field.
Three-dimensional velocity vectors (v_x, v_y, v_z) *anchored at stellar positions, color-coded by estimated spectral type. Vector lengths indicate stellar speeds, revealing distinct kinematic patterns where hotter (blue) and cooler (red) stars show differing motion trends across the Solar Neighborhood.*

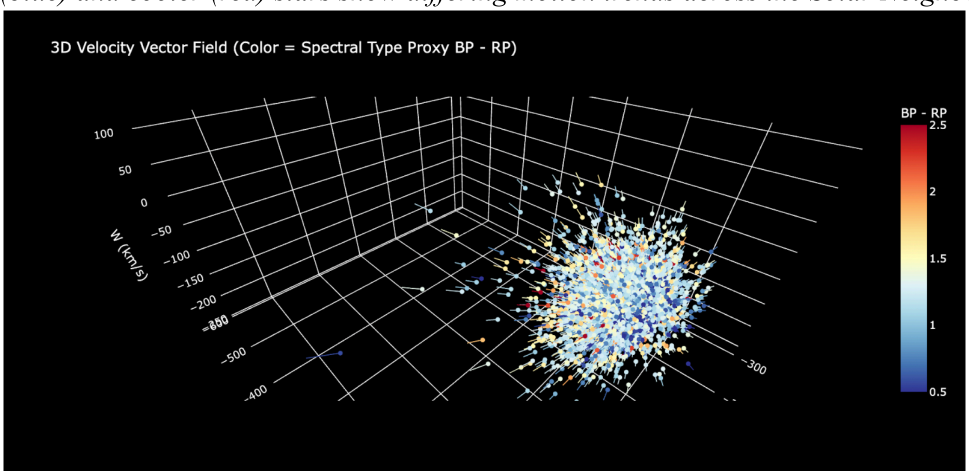
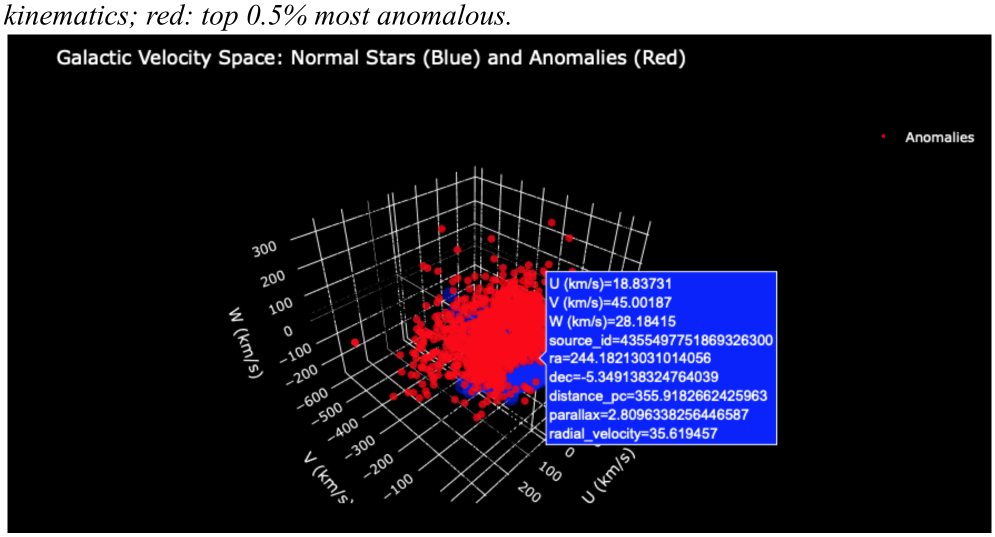


Figure 3.
Kinematic anomaly candidates. Isolation Forest classification of stellar velocities. Blue: nominal kinematics; red: top 0.5% most anomalous.



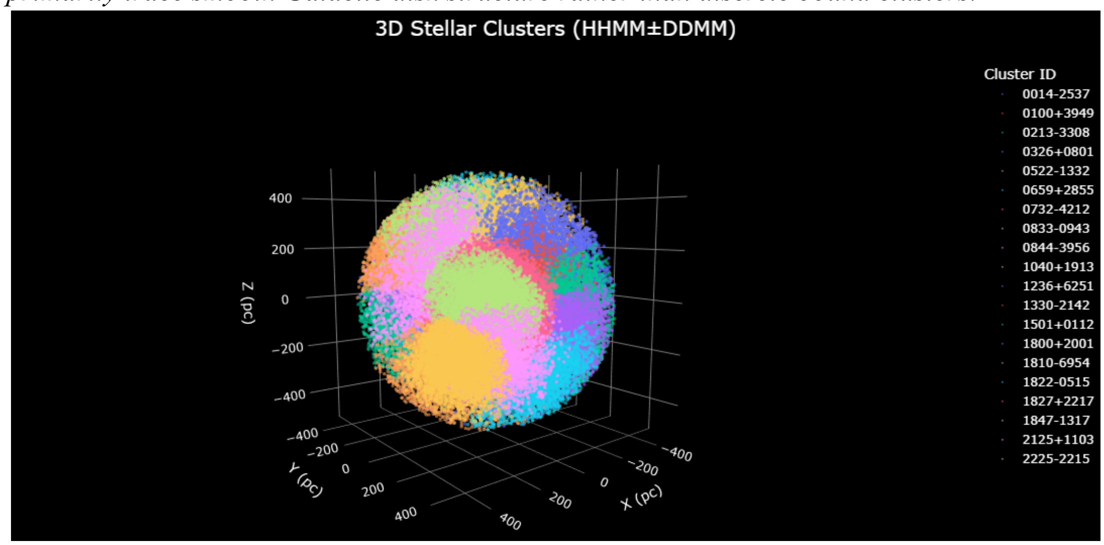
SCHOLARLY DEBUT
by Scholarly Review

Winter 2025/2026

Anomaly candidates cluster in localized regions of velocity space and occasionally in position–velocity neighborhoods, indicating potential runaways, halo interlopers, or measurement/systematic edge cases.

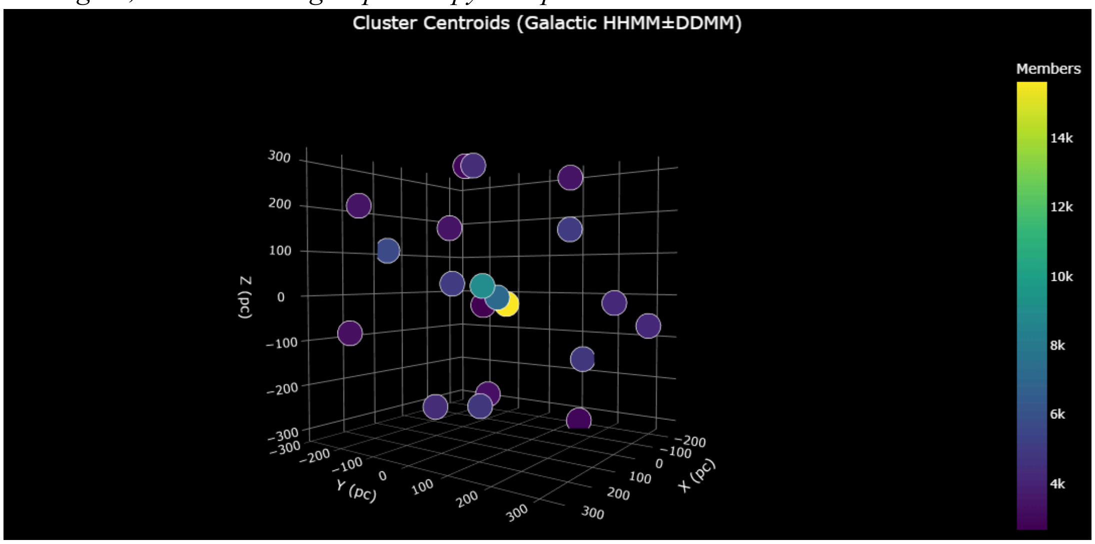
Figure 4

HDBSCAN partitions the 500 pc solar-neighborhood sample into broad spatial components that primarily trace smooth Galactic disk structure rather than discrete bound clusters.



Clustering. The various colours are denoted by the Cluster ID in the right-most side of the figure. These show location of the clusters based on HHMM+-DDD format, where hh and mm are hours and minutes of right ascension, and ddd is degrees of declination.

Figure 5
Cluster centroids and member counts show that the largest components concentrate near the densest local disk region, while smaller groups occupy the sparser outskirts.



Clustering. HDBSCAN identified coherent groups with a silhouette value above the fallback threshold in our baseline run (values depend on the final sample after quality cuts). Where local density is low or boundaries are diffuse, HDBSCAN labels noise; in such cases, the GMM fallback partitions the data into Gaussian components with reasonable BIC behavior, providing a complementary segmentation.

Figure 6.
Basic Statistics of Cluster Sizes, Sorted by Largest to Smallest.

Cluster Sizes (sorted by size)

Cluster Label	Members
1827+2217	15639
0732-4212	9067
1847-1317	7246
0326+0801	5706
0100+3949	5166
1236+6251	5133
1810-6954	4895
1800+2001	4796
0833-0943	4398
2225-2215	4368
1330-2142	4194
1501+0112	4119
0844-3956	3416
0522-1332	3370
1040+1913	3361
2125+1103	3331
0213-3308	3281
1822-0515	2940

Min: 2666 | **Max:** 15639 | **Median:** 4281.0 | **Std Dev:** 2948.8

Mass estimation. The Random Forest improved stability over the analytic proxy, especially at color extremes. The refined masses remain within the intended 0.1-1.0 stellar mass range for this magnitude-limited sample and track expected mass–color–luminosity trends.

DISCUSSION

The velocity field reveals flows and localized shear that are consistent with nearby groups and associations within a 500 pc radius, as defined. The anomaly candidates, though in small part, may include high-velocity stars or members of dynamically heated populations. Focused follow-up using spectroscopic metallicities and complementary surveys would help distinguish between these possibilities and clarify whether the most extreme outliers are halo interlopers, runaways, or artifacts.

Methodologically, we believe that the Isolation Forest contamination rate is too conservative at 0.5%. Varying the threshold may lead to more precision or recall, depending on future changes. The blended mass estimation strategy, taking an analytic proxy with a Random Forest regressor trained on available features, improves stability and color and luminosity extremes without overfitting. Similarly, the two-step clustering method combines both non-parametric cluster shapes and a low-constraint regime where Gaussian components are a practical approximation. The flexibility of HDBSCAN with a Gaussian Mixture fallback ensures that weakly separated structures are not messed up when density-based criteria underperform.

The magnitude limitations of 7.5 to 13 bias the sample towards brighter, typically nearer or intrinsically located luminous stars, affecting luminosity and mass functions. Although parallax-over-error filters mitigate the worst effects of astrometric noise, uncertainty spreads nonlinearly into distance and velocity components. This may have widened structures in space, as well as inflated anomaly scores near the boundary. Extinction and reddening of stars were ignored, which removes bias from BP-RP color. In this scenario. It was used as a temperature and type proxy, but could be biased towards sight lines with dust, affecting both mass estimations and spectral type colorings.

Despite these caveats, the integrated framework of 3D mapping, velocity-field construction, anomaly detection, and adaptive clustering provides a scalable approach to local Galactic cartography. For stellar navigation, moving groups serve as waypoints that offer positional catalogs, while anomalous stars help lead baseline navigation and travel. If we extended the pipeline to fainter magnitudes, incorporating extinction metrics and chemical abundances (redness), and validating clusters against known clusters, it would help advance the paper past basic stellar mapping.

CONCLUSION

We present a machine-learning analysis of Gaia DR3 stars within 500 pc, producing 3D spatial and kinematic maps, a refined mass estimator, kinematic anomaly candidates, and 6D clustering. Results highlight coherent structures and outliers consistent with known features of the solar neighborhood and demonstrate a reproducible pipeline for discovery. Future work will (i) extend to fainter magnitudes, (ii) incorporate extinction and metallicity/abundance information, and (iii) cross-validate clusters against known associations to refine membership and ages.

ACKNOWLEDGEMENTS

We thank Dr. Shyamal Mitra for mentorship and the UT Austin Computational Astrophysics program for research support. We are grateful to Dr. Karl Gebhardt for his insights. This workuses data from ESA's Gaia mission (DR3).

REFERENCES

Bobylev, V. V., Bajkova, A. T., & Mylläri, A. A. (2010). Analysis of peculiarities of the stellar velocity field in the solar neighborhood. *Astronomy Letters*, *36*(1), 27–43. https://doi.org/10.1134/s1063773710010044

Prisinzano, L., Damiani, F., Sciortino, S., Flaccomio, E., Guarcello, M. G., Micela, G., Tognelli, E., Jeffries, R. D., & Alcalá, J. M. (n.d.). Low-mass young stars in the Milky Way unveiled by DBSCAN and Gaia EDR3: Mapping the star forming regions within 1.5 kpc. *Astronomy and Astrophysics*, *664*, A175. https://doi.org/10.1051/0004-6361/202243580

Tang, X.-C., Tian, H., Li, J., Bing-qiu Chen, Chen, Y.-R., Liu, C., & Qiu, D. (2024). Detailed Mapping of the Galactic Disk Structure in the Solar Neighborhood through LAMOST K Dwarfs. *MNRAS*, *000*, 1–16.

Zari, E., Hashemi, H., Brown, A. G. A., Jardine, K., De Zeeuw, P. T., ESO, Leiden Observatory, Leiden University, & Consultant, Radagast Solutions. (2018). 3D mapping of young stars in the solar neighbourhood with Gaia DR2. In *Astronomy & Astrophysics* (Vol. 620, pp. A172–A172). https://doi.org/10.1051/0004-6361/201834150

Zucker, C., Alves, J., Goodman, A., Meingast, S., & Galli, P. (n.d.). The Solar Neighborhood in the Age of Gaia. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2212.00067

Liu, Fei Tony, et al. "Isolation Forest." *2008 Eighth IEEE International Conference on Data Mining*, Dec. 2008, https://doi.org/10.1109/icdm.2008.17.