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

Vishakh Hari, Krish Bahl, Ishant Yenamandra

The University of Texas at Austin

Computational Astrophysics Research Program

Dr. Shyamal Mitra

July 15, 2025

Stellar Characteristics of the Solar Neighborhood: Machine Learning Insights for Space Navigation from Gaia DR3

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. Using the columns for right ascension, declination, parallax, parallax-over-error, G magnitude, BP-RP color index, proper motion, and radial velocity, we developed several machine learning models and visualizations to perform a comprehensive study of this solar neighborhood. To allow for a feasible data load with our computing power constraints, we restricted the stars by apparent magnitude, solely including stars within $7.5 \leq G \leq 13.0$, where G is Gaia G-band (apparent) magnitude. We derived our restriction criteria and definition of the solar neighborhood from previous research by Zari 2018 and Prisinzano 2022. We first created an interactive three-dimensional map of the stars in the neighborhood, color-coded by estimated spectral type, to understand the spatial stellar distribution of this region. Next, we constructed a luminosity function histogram to analyze the distribution of stellar brightness and a three-dimensional vector velocity field color-coded by estimated spectral type to perform a kinematic analysis of the neighborhood. We then constructed machine learning models for detecting anomaly velocities, mass estimation, and clustering of stellar structures. Our comprehensive study offers insights into the structure and motion of stars in the solar neighborhood and may contribute to future stellar navigation systems for spacecraft.

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

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.

Sources of Data

We queried stellar data from the Gaia Data Release 3 (Gaia DR3) database, limiting our selection of stars within 500 parsecs of the Sun, to fit our definition of the solar neighborhood. This query ensured each star had reliable measurements including positional data (right ascension, declination, parallax); proper motion (*pmra*, *pmdec*); radial velocity; and photometric data (*phot_g_mean_mag, bp_rp*). To ensure data quality, we cleaned all data using Pandas and used *parallax_over_error* to limit stars.

Our main pre-processing involved converting stellar parallaxes to distances in parsecs, computing absolute magnitude, and transforming the right ascension and declination into Cartesian and six-dimensional coordinates. We derived the six-dimensional coordinates from proper motion and radial velocity.

For our analysis and modeling, we used Numpy, Astropy, Pandas, Scikit-learn, HDBSCAN, Plotly, and Matplotlib.

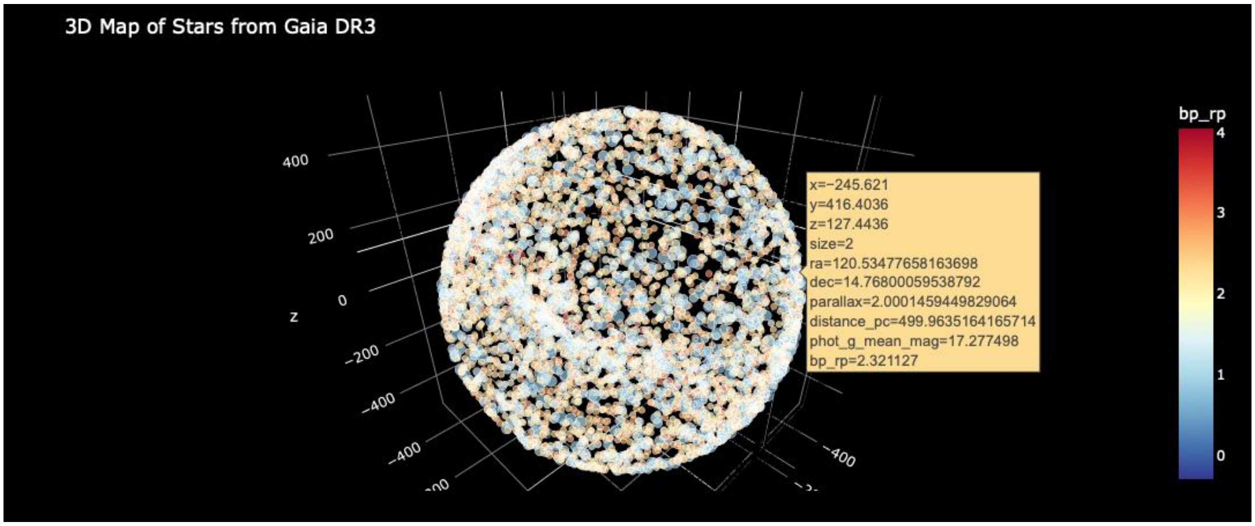
Methods

Figure 1 presents our first visualization in this study: an interactive three-dimensional spatial map of the solar neighborhood. Each point on the map represents a star, with the relative radii of the points directly proportional to the radii of the stars in space. The stars are color-coded by estimated spectral type derived from each star's BP-RP color index. Therefore, a bluer point on the map indicates a star with a higher surface temperature and a spectral type closer to Type

O, and a redder point indicates a star with a lower surface temperature and a spectral type closer to Type M. This plot reveals the overall structure and density of this region, highlighting clustering patterns and variations in spectral type across the neighborhood. This model provided us with a foundational view of the region's stellar spatial distribution and served as a basis for further kinematic and structural analysis.

Figure 1

Interactive Three-Dimensional Spatial Distribution of Stars in the Solar Neighborhood, Color-coded by Estimated Spectral Type

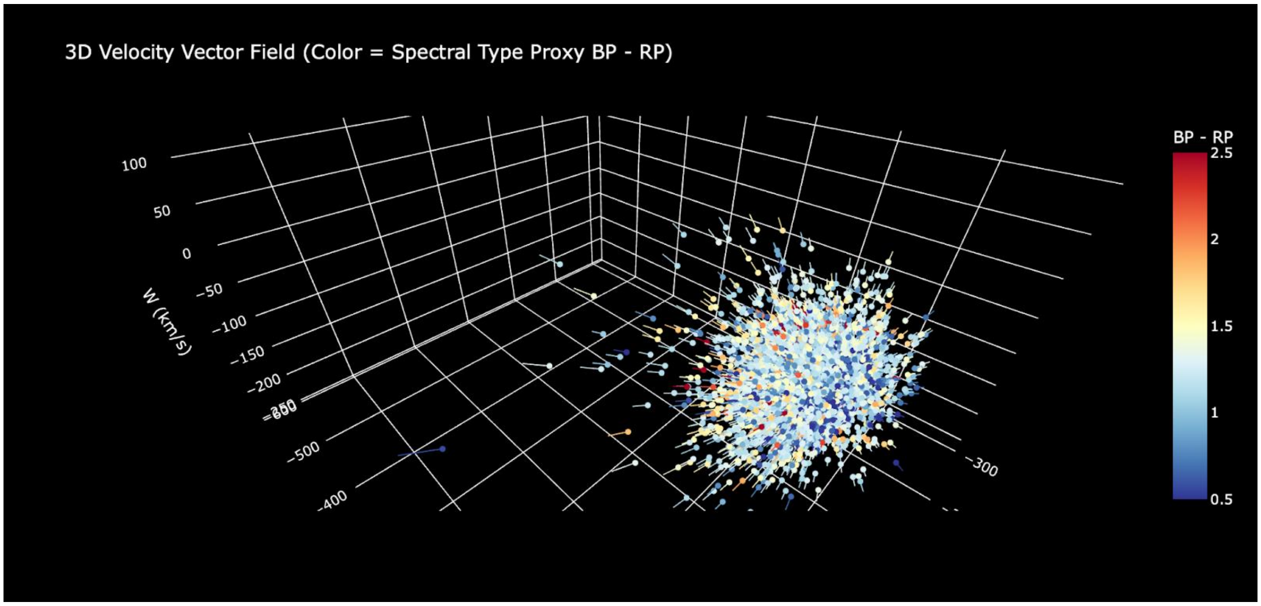


After obtaining an understanding of the positional distribution of stars within the solar neighborhood, we wanted to create a similar distribution to study their kinematic properties. To analyze stellar motions, we constructed a three-dimensional velocity vector field, shown in Figure 2. First, we calculated the three-dimensional velocity vectors for each star using radial velocity and proper motion data. Using the $v_t = 4.74 \times \frac{\mu}{\varpi}$ equation — where v_t is tangential stellar velocity, μ is proper motion, and ϖ is 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 (v_x, v_y, and v_z) using astrometric conversion factors and dimensional analysis. Then, each velocity vector was paired with a spatial position from Figure 1 to form a full three-dimensional phase-space representation of both the stellar positions and motions.

The tail of each vector is located at the star's Cartesian position, and the direction and length of each line indicates the star's velocity through space. We also color-coded the vectors by estimated spectral type derived from the BP-RP color index, to explore potential connections between stellar type and motion. This visualization allowed us to study kinematic structure, coherence, and motion trends across different regions of the neighborhood and provided a foundation for our machine learning model.

Figure 2

Interactive Three-Dimensional Velocity Vector Field of the Solar Neighborhood, Color-Coded by Estimated Spectral Type from BP-RP Color Index

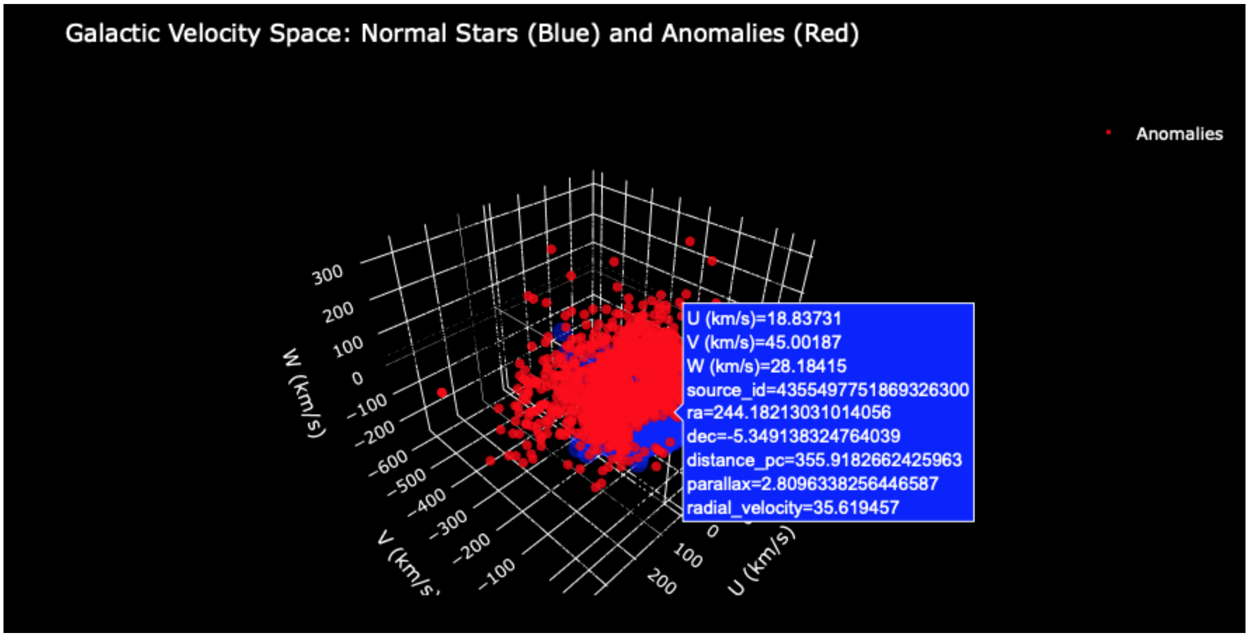


In order to identify stars traveling with unusual kinematic behavior, we implemented an unsupervised machine learning anomaly detection algorithm known as the Isolation Forest. This model recursively partitions data to isolate stellar velocities that are anomalous. For this model, we focused on stellar velocity data derived from the six-dimensional phase space: x, y, z, v_x, v_y, and v_z.

We used the three-dimensional velocity vector components v_x, v_y, and $v_z$as the main input features for this anomaly detection model. We computed these components from the proper motion right ascension, proper motion declination, parallax, and radial velocity data using standard astrometric conversion factors and dimensional analysis to obtain a heliocentric Cartesian velocity space — shown in Figure 3.

Figure 3

Interactive Three-Dimensional Anomaly Velocity Plot from Isolation Forest Anomaly Detection Machine Learning Model



U (km/s)=18.83731
V (km/s)=45.00187
W (km/s)=28.18415
source_id=4355497751869326300
ra=244.18213031014056
dec=-5.349138324764039
distance_pc=355.9182662425963
parallax=2.8096338256446587
radial_velocity=35.619457

Before inputting the data into the model, we standardized the velocity components to zero mean and unit variance to ensure equal scaling across all dimensions. This preprocessing step improved our model's sensitivity and prevented any one axis from dominating the isolation process. We initialized the Isolation Forest with 100 estimators (trees) and set the contamination parameter to 0.005. 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. Stars with lower anomaly scores were most easily isolated by the Isolation Forest model; 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 3, where we represented normal stars in blue and anomalous stars in red. Each red point, therefore, represents stars with velocities significantly different from the local stellar population. These anomalous stars may be hypervelocity stars ejected through dynamic interactions, halo interlopers, or stars with improperly measured Gaia astrometric parameters. By identifying these anomaly stars in an unsupervised machine learning model, we have enabled future targeted analysis of the dynamics of the solar neighborhood — which can enhance our understanding of local stellar motion and contribute to applications for interstellar navigation and trajectory optimization for spacecraft.

With this foundational kinematic analysis, we began our study of mass clustering patterns in the neighborhood. We began by computing analytic estimates of stellar masses using observed photometry and color. We calculate G-Band Magnitude using $m_G + 5\log_{10}\left(\frac{\varpi}{1000}\right) + 5$ where m_G = *phot_g_mean_mag* and ϖ is the parallax in milliarcseconds. The mass proxy is then taken to be $M_{analytic} = 0.8e^{-0.4(photometry)} \cdot 10^{-0.1M_G}$, where photometry is *bp_rp*. 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 the following scaling before clipping into the range: $\left(\frac{T_{eff}}{5777K}\right)^{0.5}$ To refine the relation, we trained a Random Forest regressor on the three features (*bp_rp*, m_G, and ϖ) with the analytic masses as the target values. We allocated 80% of the data for the training set and 20% for the test set. We 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:

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

As seen in Figure 7, the trees estimated most stars to be below 0.3 solar masses. We believe this was due to the restrictions we placed on magnitude and distance. To confirm accurate estimation, we also had a R^2 score of 0.91.

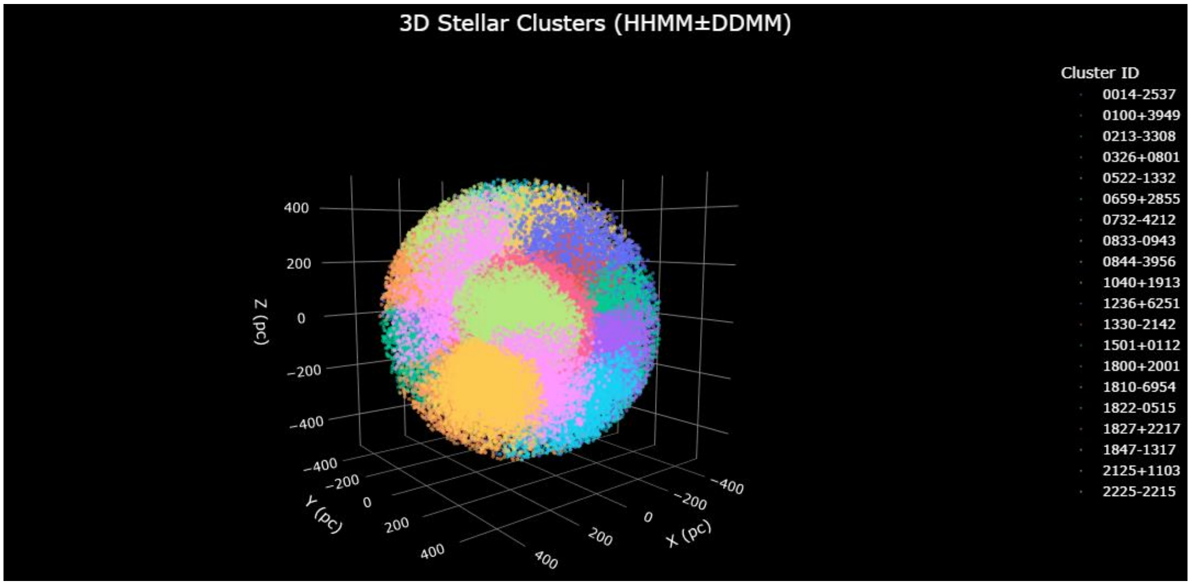
To identify clusters, we combined spatial and kinematic space (x, y, z, v_x, v_y, v_z). We began with a Mini Batch K- Means algorithm with 500 centroids to compute the minimum centroid distance for each star, and sample with probability inversely proportional to that distance. This subsampling preserves sparse features while reducing computing cost. The centroid plot can be seen in Figure 5, in which the centroids were plotted along the Cartesian plane. We labeled the centroids by size, with the largest centroid in the middle of the three-dimensional plot.

We standardized positions and velocities separately, multiplying the velocity axes by 0.5 to balance their influence. 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 assigned each star either to a cluster label or classified it as noise. We ended with a computation of the silhouette score on a random subset to assess true cluster separation.

If the silhouette score fell below 0.25, indicating poorly separated clusters, we invoked a Gaussian Mixture Model (GMM) fallback. Although we used DBSCAN and identified 49 clusters, the silhouette score was –0.51, as seen in Figure 8. This was considered poor, and we resorted to the GMM fallback. We fitted standard GMMs for component counts k = 2 to k = 30, each with full covariance and three random initializations, and selected 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. Through this modeling, we plotted a three-dimensional comprehensive heliocentric model with clusters labeled as seen in Figure 4. This shows the well separated clusters due to the GMM model fallback. Furthermore, to confidently assess the clusters, we analyzed simple statistics as seen in Figure 6. Most clusters were larger than 2000 stars, which followed our goal to map moving clusters and smaller clusters. The largest cluster was identified with about 15000 stars, which follows the general size of globular clusters. This comprehensive study of stellar clustering lays the foundation for further study of the solar neighborhood to aid stellar navigation applications.

Figure 4

Three-Dimensional Plot of Clusters Mapped in Cartesian Coordinates



Note. On the right side, you can see the color-coded unique cluster IDs, based on HHMMDDDM format.

Figure 5

Comprehensive Three-Dimensional Map with Centroids of Clusters Plotted

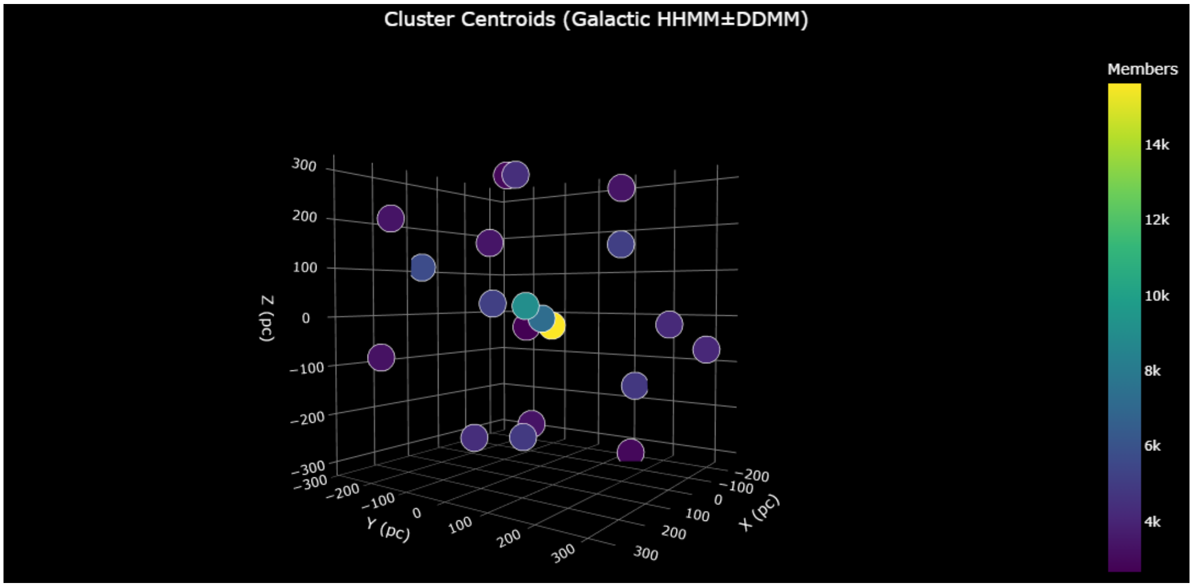
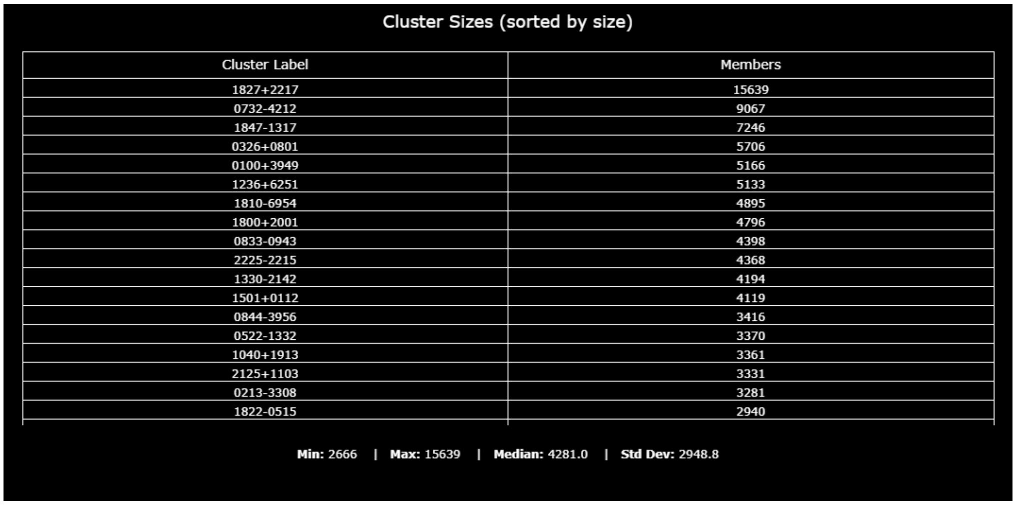


Figure 6

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



Cluster Sizes (sorted by size)

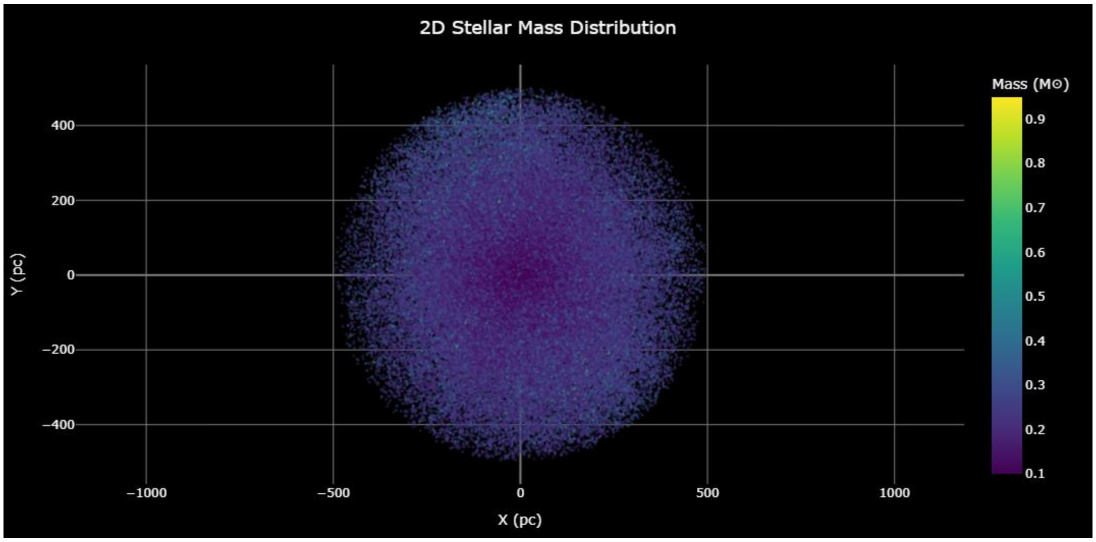
Cluster Label	Members
1827+2217	15639
0732-4212	9067
1847-1317	7246
0326+0801	5706
0100+3949	5166
1236+6251	5133
1810-6954	4895
1800+2001	4796
0833-0943	4398
2225-2215	4368
1330-2142	4194
1501+0112	4119
0844-3956	3416
0522-1332	3370
1040+1913	3361
2125+1103	3331
0213-3308	3281
1822-0515	2940

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

Note. The smallest cluster contained 2666 stars, and the largest cluster contained 15639 stars.

Figure 7

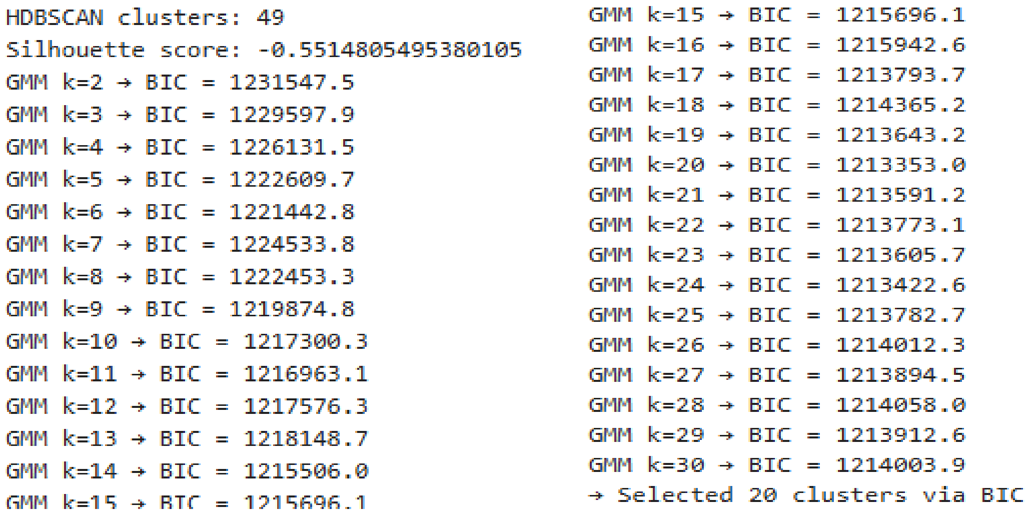
A Two-Dimensional Projection of the Three-Dimensional Heliocentric Model, Showing the Distribution of Estimated Masses



Note. Most stars lied between 0.1-0.4 solar masses.

Figure 8

Our DBSCAN and GMM Outputs, with DBSCAN at 49 Clusters and GMM at 20 Clusters.



```
HDBSCAN clusters: 49
Silhouette score: -0.5514805495380105
GMM k=2 → BIC = 1231547.5
GMM k=3 → BIC = 1229597.9
GMM k=4 → BIC = 1226131.5
GMM k=5 → BIC = 1222609.7
GMM k=6 → BIC = 1221442.8
GMM k=7 → BIC = 1224533.8
GMM k=8 → BIC = 1222453.3
GMM k=9 → BIC = 1219874.8
GMM k=10 → BIC = 1217300.3
GMM k=11 → BIC = 1216963.1
GMM k=12 → BIC = 1217576.3
GMM k=13 → BIC = 1218148.7
GMM k=14 → BIC = 1215506.0
GMM k=15 → BIC = 1215696.1
```

```
GMM k=15 → BIC = 1215696.1
GMM k=16 → BIC = 1215942.6
GMM k=17 → BIC = 1213793.7
GMM k=18 → BIC = 1214365.2
GMM k=19 → BIC = 1213643.2
GMM k=20 → BIC = 1213353.0
GMM k=21 → BIC = 1213591.2
GMM k=22 → BIC = 1213773.1
GMM k=23 → BIC = 1213605.7
GMM k=24 → BIC = 1213422.6
GMM k=25 → BIC = 1213782.7
GMM k=26 → BIC = 1214012.3
GMM k=27 → BIC = 1213894.5
GMM k=28 → BIC = 1214058.0
GMM k=29 → BIC = 1213912.6
GMM k=30 → BIC = 1214003.9
→ Selected 20 clusters via BIC
```

Note. The silhouette score shows how the poor clustering from DBSCAN was changed for a GMM model, which had better separation in clusters.

Future Work

We would like to partner with aerospace engineers to explore further research we can perform to make a greater impact on stellar navigation for spacecraft. We are waiting for the Gaia Data Release 4 database to be released in 2026. We will apply our work to the new database and expand the scope of our research to an even more comprehensive study of the solar neighborhood.

Acknowledgements

We would like to thank our mentor Dr. Shyamal Mitra for his guidance and support during this research: his coaching and timely assistance were greatly beneficial for our research process. We also thank the Computational Astrophysics research program at the University of Texas at Austin for providing this incredible opportunity to conduct research with esteemed professionals in astrophysics. We would also like to thank Dr. Karl Gebhardt, Chair of the Astronomy Department at UT Austin, for his insights. Finally, this study utilized the Gaia Data Release 3 database, made publicly available by the European Space Agency’s Gaia Mission.

Author Biographies

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 has pursued STEM subjects such as multivariable calculus, differential equations, machine learning, and physics. 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. As a member of the FBISD Engineering Academy, he takes advanced courses like Aerospace Engineering, Engineering Science, and AP Physics. He also enjoys building and launching rockets and is especially fascinated by the physics behind space travel and propulsion, 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. He has been deeply interested in machine learning, specifically in natural language processing and sentiment analysis. He has published research in machine learning previously, and is working to publish his paper in a novel metric relating to computational linguistics. In his personal life, he is heavily interested in computational finance, machine learning, and computer engineering.

References

Cantat-Gaudin, T., Jordi, C., Vallenari, A., Bragaglia, A., Balaguer-Núñez, L., Soubiran, C., Bossini, D., Moitinho, A., Castro-Ginard, A., Krone-Martins, A., Casamiquela, L., Sordo, R., & Carrera, R. (n.d.). A Gaia DR2 view of the open cluster population in the Milky Way. *Astronomy and Astrophysics*, *618*, A93. https://doi.org/10.1051/0004-6361/201833476

Castro-Ginard, A., Jordi, C., Luri, X., Cid-Fuentes, J. Á., Casamiquela, L., Anders, F., Cantat-Gaudin, T., Monguió, M., Balaguer-Núñez, L., Solà, S., & Badia, R. M. (2020). Hunting for open clusters in Gaia DR2: 582 new open clusters in the Galactic disc. *Astronomy and Astrophysics*, *635*, A45. https://doi.org/10.1051/0004-6361/201937386

European Space Agency (ESA). (1997). *The Hipparcos and Tycho Catalogues* [Data set]. https://www.cosmos.esa.int/web/hipparcos/catalogues

Kerr, R. M. P., Rizzuto, A. C., Kraus, A. L., & Offner, S. S. R. (n.d.). Stars with Photometrically Young Gaia Luminosities Around the Solar System (SPYGLASS). I. Mapping Young Stellar Structures and Their Star Formation Histories. *The Astrophysical Journal*, *917*(1), 23. https://doi.org/10.3847/1538-4357/ac0251

Kragh, H. S. (2015). Jacobus Cornelius Kapteyn : born investigator of the Heavens. *Research Portal Denmark*, *416*, 503–504. https://local.forskningsportal.dk/local/dki-cgi/ws/cris-link?src=ku&id=ku-c9ca19aa-cf37-49d0-9abd-cdfa7b0bf2d3&ti=Jacobus%20Cornelius%20Kapteyn%20%3A%20Born%20Investigator%20of%20the%20Heavens

Oort, J. (n.d.). Observational evidence confirming Lindblad's hypothesis of a rotation of the galactic system. *Bulletin of the Astronomical Institutes of the Netherlands*, *3*, 275. https://scholarlypublications.universiteitleiden.nl/access/item%3A2728648/view

Paul, E. R. (1986). J. C. Kapteyn and the early Twentieth-Century universe. *Journal for the History of Astronomy*, *17*(3), 155–182. https://doi.org/10.1177/002182868601700301

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

Vallenari, A., et al. (Gaia Collaboration). (2023). Gaia Data Release 3. Summary of the content and survey properties. *Astronomy & Astrophysics*, *674*, A1. https://doi.org/10.1051/0004-6361/202243940.

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