Western Washington University
Western CEDAR

WWU Honors Program Senior Projects | WWU Graduate and Undergraduate Scholarship

Spring 2021

A Neural Network Approach to Identifying YSOs and Exploring Solar Neighborhood Star-Forming History

Aidan McBride
Western Washington University

Ryan Lingg
Western Washington University

Marina Kounkel
Western Washington University

Kevin Covey
Western Washington University

Brian Hutchinson
Western Washington University

Follow this and additional works at: https://cedar.wwu.edu/wwu_honors

 Part of the Stars, Interstellar Medium and the Galaxy Commons

Recommended Citation

McBride, Aidan; Lingg, Ryan; Kounkel, Marina; Covey, Kevin; and Hutchinson, Brian, "A Neural Network Approach to Identifying YSOs and Exploring Solar Neighborhood Star-Forming History" (2021). *WWU Honors Program Senior Projects*. 471.
https://cedar.wwu.edu/wwu_honors/471

This Project is brought to you for free and open access by the WWU Graduate and Undergraduate Scholarship at Western CEDAR. It has been accepted for inclusion in WWU Honors Program Senior Projects by an authorized administrator of Western CEDAR. For more information, please contact westerncedar@wwu.edu.

A Neural Network Approach to Identifying YSOs and Exploring Solar Neighborhood Star-Forming History

Aidan McBride, Ryan Lingg, Marina Kounkel,
Kevin Covey, Brian Hutchinson

Abstract

Stellar ages can act as a marker of birth cluster membership for young stellar objects (YSOs), which allows for an improved understanding of the history of star formation in the solar neighborhood. However, the ages of YSOs have historically been difficult to predict on a large scale. Here, we develop a system of convolution neural network models to differentiate between YSOs and their more-evolved counterparts and predict YSO ages using *Gaia* and 2MASS photometry. The full model and resulting catalog recovers the properties of well-studied young stellar populations to a distance of five kiloparsecs, with significantly higher sensitivity within one kiloparsec, while also identifying new YSO candidate stars. We then explore the resulting catalog's implications for solar neighborhood star formation, and identify several large-scale structures, including two interesting ring or bubble-shaped groupings of young stars which may suggest radially triggered star forming events. Our results support the existence of an inclined Gould's Belt of local star formation, which may coincide with the Local Bubble. In addition, we also identify 26 high velocity "runaway" stars from the Orion Nebula Cluster and characterize their likely origins.

1 Introduction

1.1 Star Formation and Evolution

During their lifetimes, stars undergo a dramatic series of transformations. Starting their lives as gravitationally collapsing overdensities in vast clouds of interstellar gas, protostars gradually accrete material from their environment. Initially bright, but relatively cool, due to their large radii early on, stars dim slightly and increase in temperature as they contract to the main sequence, the state in which inward gravitational pressure balances outward expansion due to the fusion of hydrogen in their cores. The duration of this contraction stage is heavily dependent on the young stellar object's (YSO's) mass, with low-mass objects taking many tens of millions of years to reach the main sequence while the most massive objects may take only a few million years [1]. Once the stars' supplies of

hydrogen are depleted, which occurs much faster in more massive stars, they once again brighten and decrease in surface temperature as internal processes causes changes in the equilibrium pressure and forces an increase in radius, propelling them onto the "red giant" branch.

The temperatures of stars are reflected in their apparent colors, with hotter stars appearing bluer and cooler stars appearing redder. By observing the brightness of stars through various photometric passbands corresponding to specific ranges of color, the relative evolutionary state of stars can be determined by plotting overall stellar brightness against the differences of red and blue passbands. This crucial tool for studying stellar evolution is referred to as the Hertzsprung-Russell (HR) Diagram.

Figure 1 shows an example HR diagram constructed with *Gaia* G (green, wide-band), G_{BP} (blue) and G_{RP} (red) photometric passbands. The horizontal axis shows color and temperature, while the vertical axis shows increasing brightness.

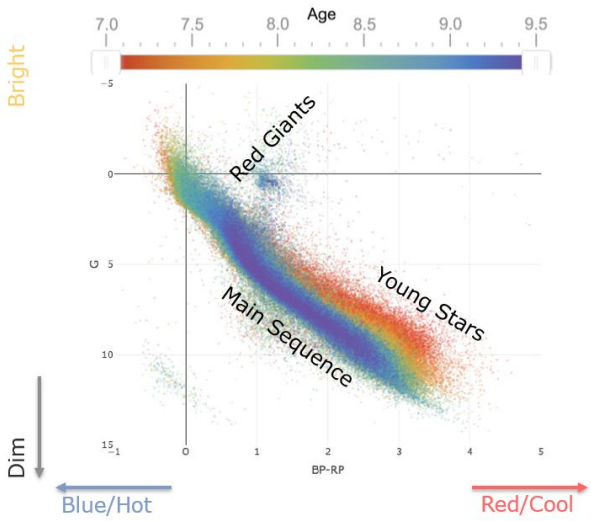

Figure 1: Example of a Hertzsprung-Russell Diagram, a fundamental tool for the study of stellar evolution. The regions of photometric space inhabited by the main evolutionary states of stars are labeled accordingly. Higher-mass stars evolve significantly faster than their lower-mass counterparts, so coeval populations may exhibit multiple regions of the HR diagram at once.

However, the observational distinction between YSOs and the more evolved main sequence, and between YSOs and the red giant branch, is not always obvious. There is no clearly-defined photometric boundary between pre-main sequence (PMS) young stars and their older main sequence counterparts. Uncertainties in distances to observed stars, unresolved binary star systems, extinction (a reddening effect caused by interstellar dust), and a variety of other sources of ambiguity can shift the position of stars on the HR diagram. Moreover, while low-mass YSOs evolve onto the main sequence over tens or hundreds of millions of years, higher mass stars progress much faster, some reaching the

main sequence in just a few million years. As a result, empirically classifying and deriving the ages of individual young stars has historically been difficult,

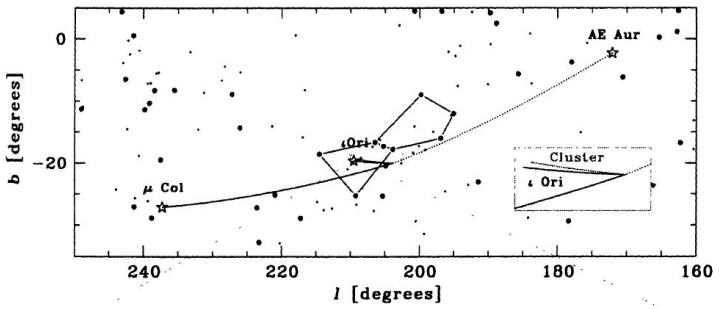

Figure 2: Runaway stars μ Col and AE Aur from Hoogerwerf et al. (2001) and first identified by Blaauw & Morgan (1964), found to project back to the known binary system ι Ori within the Orion Complex of star-forming regions. These stars were identifiable as runaways due to their high masses and clear ejection trajectory; only recently have lower mass runaways with smaller ejection velocites been identifiable.

1.2 Star-Forming Regions and Runaway Stars

The birth of new stars is concentrated within star-forming regions (SFRs), vast clouds of interstellar gas which provide the large volumes of material necessary for stellar accretion. Compared to the density of typical interstellar space, SFRs have significantly more stars per volume, and a much higher likelihood of gravitational interactions between objects.

Due to the exhaustion of the surrounding gas and the dissipation of the clouds due to stellar feedback, localized star formation tends to be coeval, with the dispersion of ages for most stars within a particular region typically being relatively small compared to the average stellar lifespan [2]. As such, stellar ages can provide a reliable indicator of shared birth cluster membership.

Regional clusters of young stars tend to be dynamically cold, meaning that the typical velocity dispersion of sources within a star-forming region is relatively low. For typical SFRs members, another method of determining cluster membership However, due to the complex gravitational interactions within SFRs, some stars may be excited to higher relative velocities; if this velocity is sufficient to eject the stars from their birth clusters, these sources are referred to as "runaways" [3].

Generally, the rate of stellar ejection from their birth clusters is not well understood due to the complexity of these regions. Moreover, while the rate of ejection could be characterized observationally, detecting runaways has historically been difficult. In the past, runaways have been identified from several local SFRs, most notably from the Orion Nebula. This was typically restricted to very high mass, high luminosity stars ejected at very high velocity compared to the average motion of their birth clusters. For example, Hoogerwerf et al. (2001) [4] identified several runaways from the Orion Nebula using data

from the *Hipparcos* space telescope, and also elaborated on work by Blaauw & Morgan (1964) [5] to trace two very massive stars μ Col and AE Aurigae (Figure 2) back to a likely ejection from the ι Ori binary system within the Trapezium grouping in Orion. The latter pair of runaways (shown in Figure 2) is particularly emblematic of the types of ejected stars that historically could be identified due to their large masses, distances from their cluster of origin, and interaction with a dense grouping of massive stars.

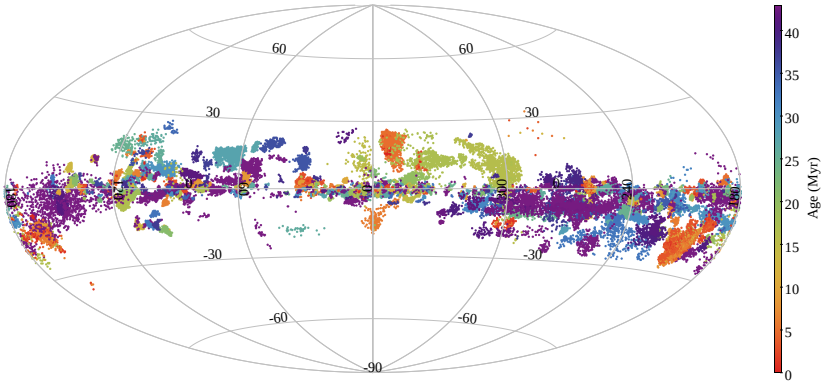

Figure 3: All-sky plot of training catalog stars, color coded by cluster-averaged ages.

Recent advances in ground-based observational techniques have allowed for slightly fainter runaways being detected in small numbers. One notable example is the work of Kounkel (2017) [6], who used radio interferometry to identify three young sources which were ejected from the Orion Nebula due to interactions with some of the region's massive stars. However, these techniques still require relatively precise targeting of candidate YSOs and cannot simply target all stars within a region. For a more sweeping view of runaways from star-forming regions, a wide-scale astrometric sky-survey is required.

1.3 Astrometric Surveys and *Gaia*

Astrometry is the set of astronomical techniques used to measure the positions, solar-centric distances, and motions of stars, typically by fitting a solution to observed stellar positions over time. Aside from the positional coordinates of galactic latitude and longitude, the primary astrometric parameters are stellar parallax and proper motions.

Parallax is a measure of a nearby star's angular deflection against a more distant background when observed from different locations in Earth's orbit; as this deflection is typically less than one arcsecond of angle, a small-angle approximation stellar parallax to serve as an inverse measure of distance. Distances measured using parallax are expressed in units of parsecs, where one parsec ($\sim$ 3.26 lyr) is the distance to an with a parallax of 1 arcsecond.

Proper motions are the measure of a star's angular motion perpendicular to the line of sight, typically measured in miliarcseconds per year. Because stars exhibit both parallax and proper motions as some change in angular position over time, observing proper

motions typically requires several years' worth of measurements of the same object.

Observing stellar parallax and proper motion is a difficult task, as even the nearest stars exhibit parallaxes of only about an arcsecond. In the past, astrometric measurements have been particularly reliable with radio astronomy (e.g. Kounkel 2017) [6] due to the use of interferometry to increase angular resolution and the ability to take observations against very distant "fixed"" reference points. However, as this required targeting of specific stars multiple times over the course of several years, the number of sources which could be targeted for astrometry was typically fairly low.

The European Space Agency *Hipparcos* satellite, operational through most of the 1990s, was an exception to the low volume of astrometric data produced by ground-based surveys, measuring astrometric parameters with precision of about 1 mas for about 2 million stars over its lifetime [7]. However, the majority of *Hipparcos* sources with reliable astrometry lie within a distance of $\sim$ 100 pc, which is slightly less than the distance to the nearest major star-forming region.

The biggest leap forward in astrometry has come within just the past few years, with *Gaia*, the ESA's follow-up to *Hipparcos*. *Gaia* was launched in 2013, and its 2016 Data Release 1 (DR1) contained preliminary astrometric solutions for about two million stars, comparable to the *Hipparcos* catalog [8]. Then, its 2018 Data Release 2 (DR2) produced stellar positions, colors, on-sky velocities, and parallaxes for 1.7 billion stars within the Milky Way and neighboring galaxies, with astrometric precision about 100 times greater that *Hipparcos* [9]. This volume and precision of data was unprecedented, particularly among astrometric surveys; while previously, proper motions might be derived for a few precisely targetedd stars, the *Gaia* DR2 catalog now provides these motions for a significant fraction of nearby stars, giving astronomers the most complete three-dimensional view of the Milky Way to date.

1.4 Data Driven Techniques for Astronomy

The advent of highly data-intensive astronomical surveys in the past few decades, such as *Gaia* and other contemporaneous observational missions, has opened the door to new avenues of scientific exploration, but has also required the development new methods to make useful inferences from the sheer volume of data. Increasingly, astronomers have been turning to data-driven machine learning methods as a means of taking advantage of the recent wealth of observational data. Generally speaking, data driven methods require some representative traning sample of "labeled" information, from which algorithms can be tailored to identify and generalize trends to the broader dataset.

Neural networks are a type of machine learning algorithm which have been employed increasingly often within astronomy for a variety of purposes, such as deriving orbital periods (e.g. Olmschenk et al. 2021 [10]), predicting stellar parameters (e.g. Olney et al. 2020 [11]), and identifying young stars (e.g. Vioque et al. 2020 [12]). Conceived to learn in ways analogous to organic neurons, neural networks consist of a system of interconnected layers of initially randomized numerical "neurons", which take some data as input and, by propagating through the layers, produce some numerical output. As the

network is fed labeled training data, its output is compared to the label, and the layers and connections of the network are gradually optimized to best reproduce the expected output. Once the neural network model has been sufficiently trained, it can then be used to produce further outputs for unlabeled data.

Neural networks have the benefit of being able to recognize patterns within data that are higher-dimensional or more entangled than would be possible by a human, and are able to operate on much larger volumes of data than other computational methods. While they require careful selection of training data and can be somewhat computationally intensive to develop, neural networks are fairly flexible, allowing adaptation in architecture or training methods, and can be significantly more efficient than computing parameters through other means.

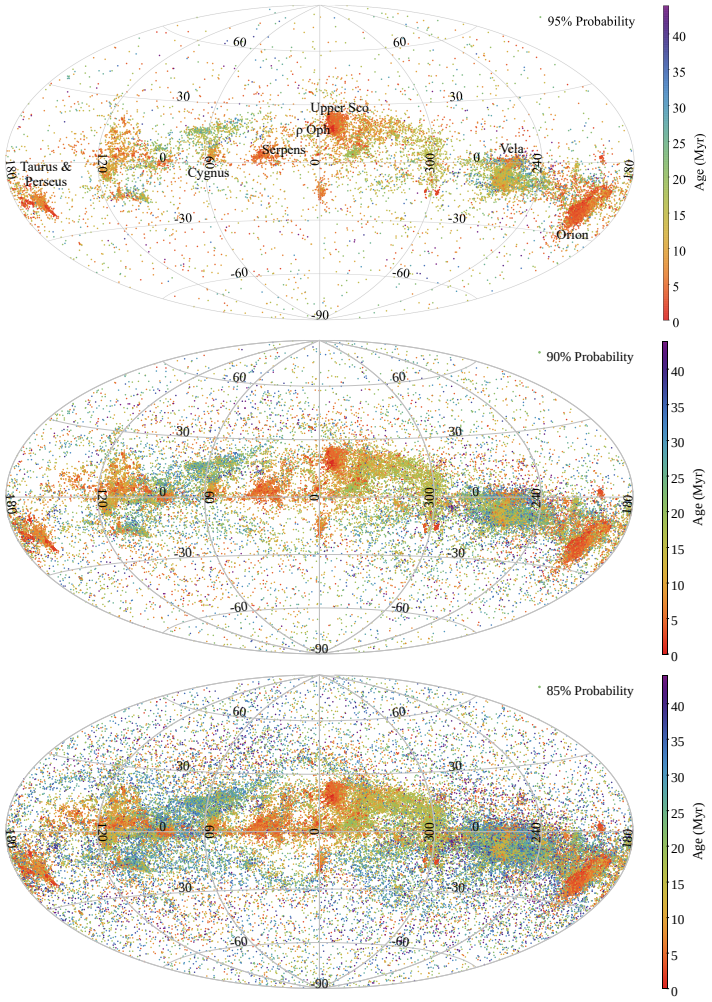

Figure 4: On-sky model outputs in galactic coordinates, showing young stars identified using the model's YSO classifier and stellar ages from the model's age regressor. These maps show three classifier "probability" thresholds, demonstrating how age and unclustered background contamination vary based on the classifier's certainty.

2 Project Overview

Using a set of neural networks trained on data from *Gaia* DR2 as well as the 2 Micron All-Sky Survey, we developed a three-part model which classifies sources into YSOs and more evolved stars, predicting ages for the younger category. We then produced a catalog using this model from all of *Gaia* DR2 within a solar-centric distance of 5 kiloparsecs, which we used to explore the history of star formation in the solar neighborhood.

Additionally, we used the high-precision astrometry of *Gaia* DR2 to search for stars that had been ejected from the Orion Nebula Cluster through gravitational interactions, and identified 26 candidate runaway stars, which we project backwards in time to characterize their origins.

3 Training Data

Neural network models require a representative catalog of training data with reliable ground-truth values in order to generalize trends within the data. For this purpose, we relied heavily on the catalog produced by Kounkel et al., (2019) [13] . This catalog used the HDBSCAN [14] clustering algorithm to identify coherent groups of stars based on their *Gaia* DR2 positions, parallaxes, and proper motions within a distance of five kiloparsecs. Because YSOs are typically much more densely clustered than other stars, this mostly recovered young stars as clustered sources. To calculate the per-cluster age, photometry of sampled stars from each cluster were compared to age-dependent grids of theoretical stellar isochrone models, as well as a neural network trained on thousands of synthetic cluster populations. Notably, however, due to the use of motions as a basis for clustering, this method is inherently not sensitive to any sources moving with peculiar velocities, such as small comoving groups or runaway stars.

Instead of relying on clustering, we want to extract YSOs from the field by their photometric parameters alone. To produce a photometry-reliant model, we compiled a training dataset from this catalog by combining the included Gaia DR2 parallax, G, B_{BP}, and G_{RP} photometric passbands with further infrared photometry in 2MASS J, H, and K passbands. The training targets were the previously-derived ages and a binary flag indicating whether each source was a member of an identified star-forming group. This training catalog contains 982,004 stars, of which 62,484 are flagged as YSOs. Figure 3 shows the on-sky distribution of YSOs used in model training. We further included quality cuts described by Lindgren et al. (2018) [15] in order to be sure that all data used for model training was precise enough to negate systematic observational effects.

3.1 Augmentation

After initial results suggested that the models were not fully generalizing on the dataset, we also employed methods to augment the real data with synthetically produced photometry. Copying the measurements for a sample of stars in the original dataset, we then

added a random offset to the stellar parallax chosen to synthetically shift the apparent distance of the star within a maximum radius of 5 kpc. Using this synthetic distance, we recalculated their photometric magnitudes as would be observed at this new distance.

By augmenting the data using synthetic observations, we were able to afford the model more variety in its inputs, and artificially increase the proportion of young stars to evolved stars that the model was exposed to. We were also able to slightly extend the distance to which the model was effective by bolstering the number of YSOs represented in the training catalog at larger distances. It should be noted, however, that this does not necessarily allow for better detection of especially dim YSOs at large distances, as these stars are typically beyond *Gaia*'s detection limits.

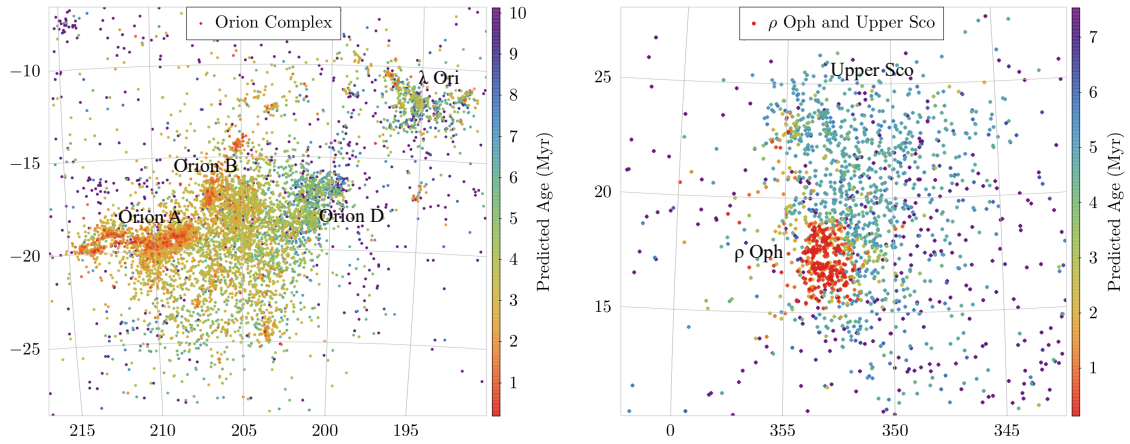

Figure 5: Local detail of catalog results for the Orion Molecular Complex (left) and ρ Oph/Upper Sco (right). Probability, age, and distance cutoffs have been selected to best isolate these regions from their background.

4 Neural Network Model Development

After some experimentation, my collaborators and I found that a set of three neural networks with similar architectures but different functions was best suited to the overall functionality of our model.

4.0.1 A_V Map

We trained the first neural network to act as an extinction map, taking individual stars' on-sky positions and outputting the expected A_V, a photometric parameter corresponding to the amount of interstellar dust along the line of sight. *Gaia* galactic latitude and longitude (l and b) and stellar parallax (π) from the training dataset were taken as inputs, while the target, A_V, was inferred from extinction measurements present in the *Gaia* catalog. The outputs of this extinction map are used as inputs to the next two

neural networks in order to negate contamination from "extinguished" (photometrically reddened) background stars due to the effects of dust.

4.0.2 YSO Classifier

We trained the second neural network to output a 0-1 probability (referred to as "YSO Probability" or "YSO Certainty") indicating the certainty that individual stars are young stellar objects. This model takes *Gaia* stellar parallax, A_V as predicted by the previous model, and the photometric passbands of G, BP, RP (*Gaia*) and J, H, K (2MASS) were taken as inputs, with the training target set to the binary flag described in section 3. During training, we employed data augmentation (see section 3.1) to oversample from the YSO class, and to increase the number of stars apparent at higher distances.

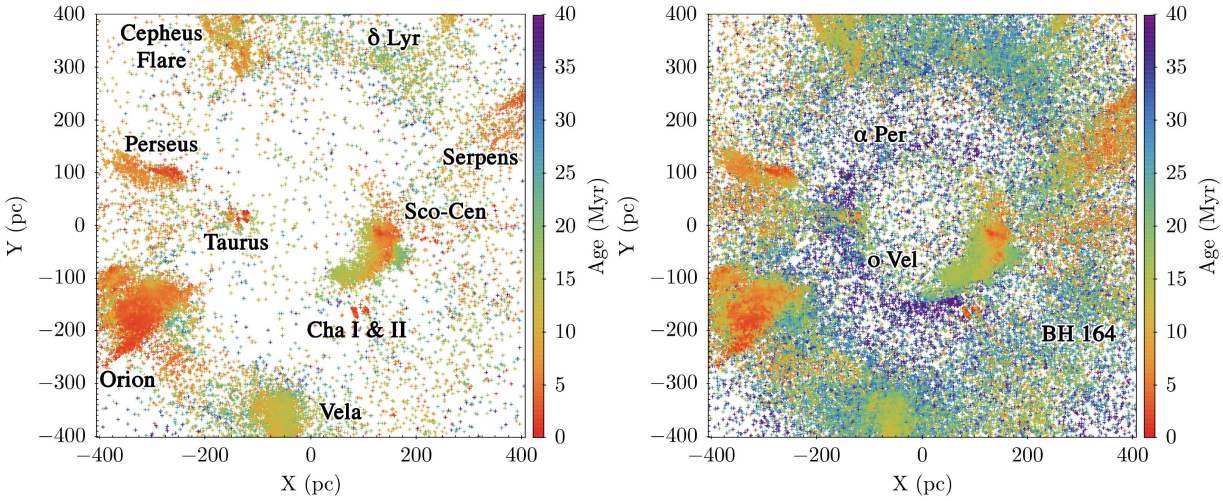

Figure 6: Distribution of PMS stars (up to a confidence thresholds of 95% - left, and 85% - right) in a heliocentric rectangular reference frame, color coded by age, showing the bubble-like structure present amount nearby SFRs.

4.0.3 Age Regressor

We trained the third neural network to regress stellar ages for all sources classified as YSOs. This model takes the same network inputs as in section 4.0.2 were used, with the target set to the known ages of the training set, in the form of $\log\big(\text{age}\big)$. However, because photometric ages cannot be derived for main sequence stars, we restricted the training dataset for this model to YSOs only. As with the YSO classifier, we employed data augmentation to balance the number of stars at relatively low ($\log\big(\text{age}\big) < 6.5$) ages with stars of higher ($6.5 < \log\big(\text{age}\big) \lessapprox < 8$) ages, and to increase the number of stars at higher distances.

4.1 Model Training

Each model was implemented in the software package PyTorch with identical structure, apart from the dimension of the input layer. The basic network architecture consists of several stages; a set of four convolution layers take a one-dimensional input (with variable size) and gradually increase the output size to a maximum of 128 channels, each separated by ReLU and batch normalization operations. Next, the parameter space is again reduced by a fifth convolution layer to a one-dimensional array, fed into three fully-connected (deep) neural network layers, and finally reduced to a scalar-valued output.

To reduce the risk of the models over-relying on any particular input parameter, all inputs and training targets were normalized between their minimum and maximum values to lie within the range $[-1, 1]$. After models were trained, their outputs were automatically mapped to their original parameter space.

During model training, one-dimensional input and target tensors were constructed from the training set and ingested into the network input. Using a mean-square-error loss function and the Adam optimizer (all native functions available within PyTorch), the models were trained over a variable number of epochs (iterations) until convergence had been achieved.

5 Model Results

5.1 Overall Results

To 70% confidence, we identify 197,315 YSOs in *Gaia* DR2 within a distance of five kiloparsecs, which we have made available as a downloadable catalog with separate columns indicating the classifier's output YSO certainty and regressor's predicted age (in log years). Figure 4 shows an all-sky view of the catalog for a variety of YSO certainty thresholds.

5.2 Individual SFRS

By isolating known star-forming regions according to their positions, distances, and expected ages, we use our produced catalog to explore the local-scale history of star formation in several noteworthy solar-neighborhood SFRs.

5.2.1 Orion Molecular Cloud Complex

The Orion Molecular Complex is one of the largest star forming regions in the solar neighborhood and has been particularly well studied [16]. Orion contains a large number of smaller subregions and a has complex history of star formation, with some regions older than 10 Myr and others still within the first few Myr of their star forming lifespan, including the well-known Orion Nebula [17].

The first panel of Figure 5 shows our catalog's results for the OMC at YSO probability of 90%. Especially noteworthy in this region is the age gradient from the older subregions

λ Ori and Orion D to the younger, actively star-forming subregions Orion B and, in particular, Orion A, which contains the Orion Nebula Cluster (ONC). The Orion Nebula's stellar population is well studied, with an average age of $\sim$ 3 Myr. Prior to this work, about 5000 stars were known to be members of its cluster [18]. Using our model, we recover approximately 300 stars in the region of the Orion Nebula which have not previously been identified as cluster members, but are classified as young to YSO certainty of 95%.

5.2.2 Upper Scorpius and ρ Ophiuchi

Upper Scorpius is relatively young component of the much larger Scorpius-Centaurus complex of stellar associations that exist towards the Galactic center; ρ Ophiuchi is a subregion of Upper Sco which exists at high galactic latitude and has an average age of less than 1 Myr, making it perhaps the youngest nearby SFR in which YSOs can be observed. [19].

The second panel of Figure 5 shows our catalog's results for the Upper Scorpius region, making the disparity between the the $<$ 1 Myr ρ Ophiuchi SFR and the surrounding $\sim$ 5 Myr population very apparent. As one of the youngest regions in the training catalog, ρ Ophiuchi contains stars that lie well above the main sequence, making them particularly easy to identify photometrically.

5.3 Large-Scale Structures

In the distribution of solar neighborhood star-forming regions, an apparent inclined disk structure has been previously identified, reaching its highest galactic latitudes towards the galactic center near ρ Ophiuchi and dipping below the plane of the galaxy towards the Galactic anticenter near Orion. This structure has historically been referred to as the Gould's Belt, with some hypotheses for its origin suggesting multiple supernovae (e.g. Pöppel & Maronetti 2000 [20]) or infalling collisions with the plane of the galaxy (e.g. Bekki 2009 [21]); however, some recent work, such as Zari (2018) [22] has found no coherent connections between Gould's Belt regions that would suggest a common cause.

Figure 6 shows our catalog results for all local star-forming regions from a heliocentric perspective. We find that the majority of major SFRs near the sun lie within one of two rings, one with a radius of roughly 150 pc and the other with a radius of roughly 400 pc. The space between younger SFRs does appear to be bridged by older populations, which are more represented at lower classifier certainties. lending credence to the suggestion of a coherent Gould Belt.

One possible explanation for these observed large-scale structures is that they are related the Local Bubble, a known underdensity in the interstellar medium in the vicinity of the Sun [23]. The local bubble has a radius of roughly 150 pc, which coincides well with the inner of the two rings. If the Local Bubble was caused by some radially expanding shock, such as a supernova explosion, this shock may also have played a part in triggering star formation within these rings.

6 Runaways

This section reiterates work from McBride & Kounkel (2019) [18], published in The Astrophysical Journal

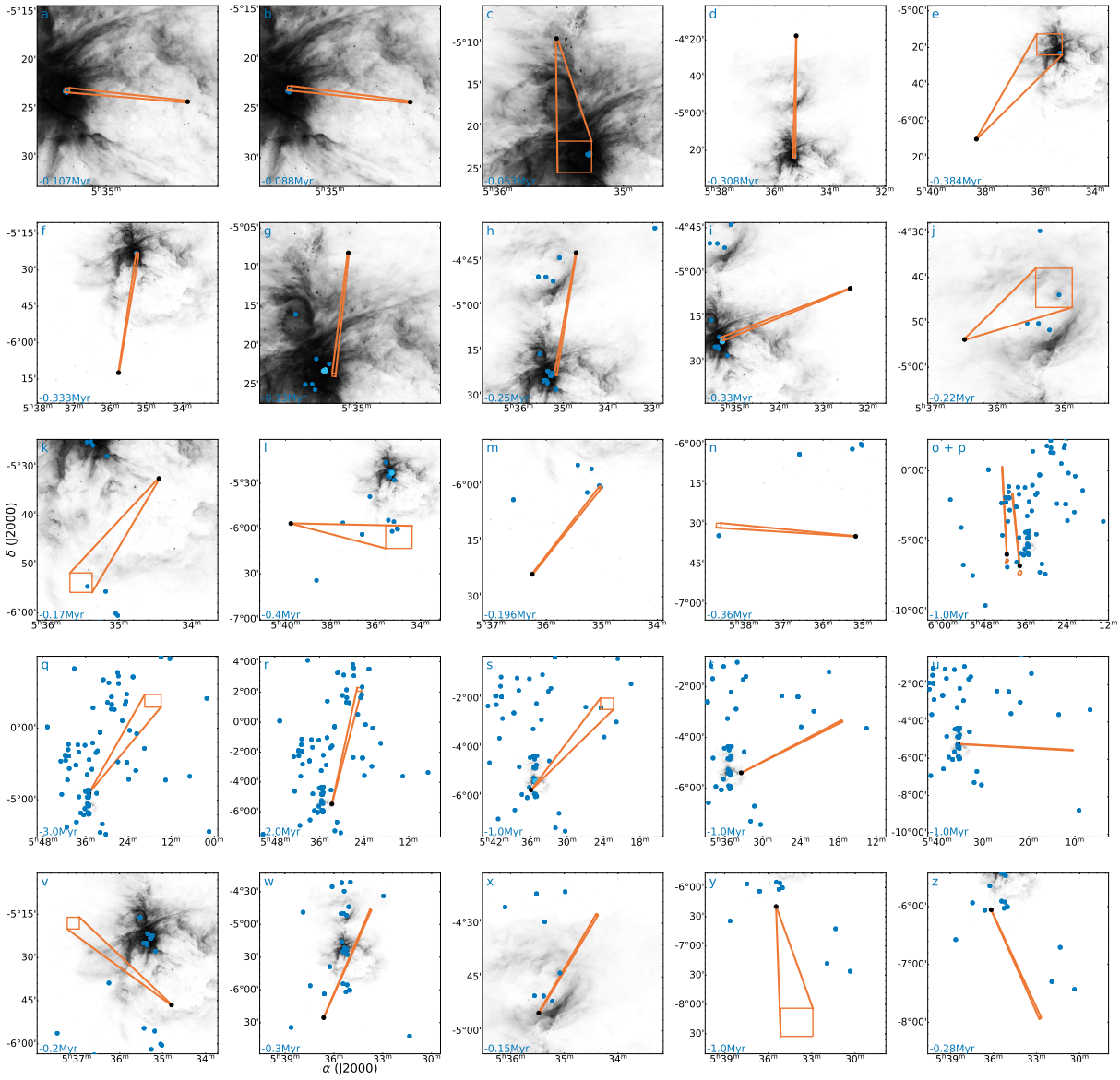

Figure 7: The 26 identified runaway sources from the Orion Nebula Cluster found using *Gaia* DR2

The advent of *Gaia* DR2 has also allowed for previously impossible advances in the search for runaway stars from nearby star-forming regions. Where previously, candidate runaways would have to be targeted for astrometric observation, *Gaia* now allows for the motions of many stars to be characterized at once.

However, the unprecedented volume of data available also presents challenges, as it is difficult to distinguish between the proper motion of a young star ejected from its birth cluster and the innate motion of an older, unrelated star in the background. While *Sagitta* should help confirm the youth of young stars in the future, initial searches for runaways were carried out before its existence.

As the Orion Nebula Cluster (ONC) has already been well-studied as a SFR which has prouced notable runawways, it is the ideal candidate for a follow-up study with *Gaia* DR2 astrometry. To solve the issue of possible background contamination, we compiled a catalog of all confirmed members of the Orion Nebula from the literature [figure out citations!]. We then crossmatched this catalog with sources in *Gaia* DR2. In total, this catalog contained 5988 YSOs.

From this catalog, the average proper motion of all stars in the ONC, in spatial coordinates of right ascension (RA or α) and declination (DEC or δ), were calculated. We applied a photometric cut initially described in Kounkel et al. 2018 [24], isolating only YSOs with clear signs of age less than roughly 15 Myr. This restricted the final list of clearly defined cluster members to 1867 sources.

The typical dispersion of stellar velocities within the nebula is small, typically less than 5 mas/yr from the cluster average. Any star travelling at speeds significantly higher than the typical dispersion of the cluster was likely accelerated from some gravitational interaction and ejected as a runaway. To isolate these stars from the list of cluster members, we simply disregarded all sources that exhibited velocities within the typical dispersion and kept the remaining members. This yielded a final total of 26 runaway stars, for which we attempted to identify an originating star or system of stars.

We compiled a list of the positions of all OB-type (a stellar classification indicating bright and blue photometry) stars within the vicinity of the Orion Nebula. These stars are the most massive sources within this region and would be responsible for the vast majority of gravitational instability required to excite runaway stars to ejection velocities; while lower-mass stars could factor in to some number of gravitational ejections, there are far too many to account for all possible interactions. Of particular interest among the compiled O and B stars is the Trapezium, a dense group of massive stars that makes up the gravitational core of the ONC.

To identify the likely cause of each ejected runaway, we projected them back over a maximum timespan of 1 Myr, stopping at the first OB star which they came within astrometric error of contact with. The trajectories of these stars can be seen in Figure 7. Of the 26 runaway sources, nine appear to originate from the Trapezium group, four appear to originate from other OB stars in the ONC, four appear to project back to other nearby clusters, and the remaining five have no single obvious point of origin.

7 Conclusion

Understanding the star-forming history of the solar neighborhood is crucial for the study of stellar evolution in general. Characterizing how stars evolve together and how the

dynamic lifetime of young stellar populations unfolds over time allows for a much richer view of the nearby Galaxy and can help us develop a history of our own Sun's lifetime.

The preponderance of data now available from *Gaia* and other major astronomical surveys allows for significant improvements in our understanding of the star forming history and dynamics of the Galaxy, provided that the techniques we use are robust enough to be applied on a wide scale. Machine learning techniques, when applied in tandem with other validation criteria, are a particularly effective way of making sense of the large volumes of data from modern surveys.

We train a data-driven machine learning model for identifying young stars and estimating their ages based on their photometric measurements within a distance of 5 kpc and apply it to the large and highly precise *Gaia* DR2 dataset, producing a catalog with nearly 200,000 potential young stars. Our catalog allows for the certainty of selecting YSOs to be selected depending on the application's tolerance for contamination and required age ranges.

Using our catalog, we characterize the stellar populations of several nearby star-forming regions, which in some cases has previously only been possible for small numbers of stars within the larger region. We also examine the broader distribution of star-forming regions in the solar neighborhood and find that they trace the previously identified Gould Belt, an inclined planar disk structure roughly centered near the sun. We note the coincidence of the Local Bubble in the interstellar medium with the inner radius of the Gould Belt and propose that they may be correlated.

Additionally, we identify 26 high-velocity stars within the Orion Nebula Cluster, which we project back in time to attempt to characterize their origins. We find that of these stars, roughly half follow trajectories originating from massive stars within the cluster, whereas the remaining half either appear to originate from other nearby star-forming regions or have no obvious source.

As our produced catalog of young stars allows the effects of contamination from more evolved sources to be mitigated and cluster origin to be characterized via the proxy of stellar ages, in the future we hope to carry out similar analyses of runaway stars from other local star-forming regions which were not previously as well-studied as the Orion Nebula. Additionally, members of our catalog have been selected for spectroscopic measurement by the Sloan Digital Sky Survey, which will provide further details about the evolutionary state of these stars and the regions they formed within.

[1] B. Reipurth. *Handbook of Star Forming Regions, Volume I: The Northern Sky*. Number v. 1 in ASP Monographs. Astronomical Society of the Pacific, 2008.

[2] R. D. Jeffries, S. P. Littlefair, Tim Naylor, and N. J. Mayne. No wide spread of stellar ages in the Orion Nebula Cluster. *MNRAS*, 418(3):1948–1958, December 2011.

[3] Christina Schoettler, Richard J. Parker, Becky Arnold, Liam P. Grimmett, Jos de Bruijne, and Nicholas J. Wright. Dynamical evolution of star-forming regions: III. Unbound stars and predictions for Gaia. *MNRAS*, 487(4):4615–4630, August 2019.

[4] R. Hoogerwerf, J. H. J. de Bruijne, and P. T. de Zeeuw. On the origin of the O and B-type stars with high velocities. II. Runaway stars and pulsars ejected from the nearby young stellar groups. *A&A*, 365:49–77, January 2001.

[5] A. Blaauw. The O Associations in the Solar Neighborhood. *ARA&A*, 2:213, 1964.

[6] M. Kounkel, L. Hartmann, L. Loinard, G. N. Ortiz-León, A. J. Mioduszewski, L. F. Rodríguez, S. A. Dzib, R. M. Torres, G. Pech, P. A. B. Galli, J. L. Rivera, A. F. Boden, N. J. Evans, II, C. Briceño, and J. J. Tobin. The Gould's Belt Distances Survey (GOBELINS) II. Distances and Structure toward the Orion Molecular Clouds. *ApJ*, 834:142, January 2017.

[7] E. Høg, C. Fabricius, V. V. Makarov, S. Urban, T. Corbin, G. Wycoff, U. Bastian, P. Schwekendiek, and A. Wicenec. The Tycho-2 catalogue of the 2.5 million brightest stars. *A&A*, 355:L27–L30, March 2000.

[8] Gaia Collaboration, G. Clementini, L. Eyer, V. Ripepi, M. Marconi, T. Muraveva, A. Garofalo, L. M. Sarro, M. Palmer, X. Luri, and et al. Gaia Data Release 1. Testing parallaxes with local Cepheids and RR Lyrae stars. *A&A*, 605:A79, September 2017.

[9] Gaia Collaboration, A. G. A. Brown, A. Vallenari, T. Prusti, J. H. J. de Bruijne, C. Babusiaux, C. A. L. Bailer-Jones, M. Biermann, D. W. Evans, L. Eyer, and et al. Gaia Data Release 2. Summary of the contents and survey properties. *A&A*, 616:A1, August 2018.

[10] Greg Olmschenk, Stela Ishitani Silva, Gioia Rau, Richard K. Barry, Ethan Kruse, Luca Cacciapuoti, Veselin Kostov, Brian P. Powell, Edward Wyrwas, Jeremy D. Schnittman, and Thomas Barclay. Identifying Planetary Transit Candidates in TESS Full-frame Image Light Curves via Convolutional Neural Networks. *AJ*, 161(6):273, June 2021.

[11] Richard Olney, Marina Kounkel, Chad Schillinger, Matthew T. Scoggins, Yichuan Yin, Erin Howard, K. R. Covey, Brian Hutchinson, and Keivan G. Stassun. APOGEE Net: Improving the Derived Spectral Parameters for Young Stars through Deep Learning. *AJ*, 159(4):182, April 2020.

[12] M. Vioque, R. D. Oudmaijer, M. Schreiner, I. Mendigutía, D. Baines, N. Mowlavi, and R. Pérez-Martínez. Catalogue of new Herbig Ae/Be and classical Be stars. A machine learning approach to Gaia DR2. *arXiv e-prints*, page arXiv:2005.01727, May 2020.

[13] Marina Kounkel and Kevin Covey. Untangling the Galaxy. I. Local Structure and Star Formation History of the Milky Way. *AJ*, 158(3):122, Sep 2019.

[14] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), mar 2017.

[15] L. Lindegren, J. Hernández, A. Bombrun, S. Klioner, U. Bastian, M. Ramos-Lerate, A. de Torres, H. Steidelmüller, C. Stephenson, D. Hobbs, U. Lammers, M. Biermann, R. Geyer, T. Hilger, D. Michalik, U. Stampa, P. J. McMillan, J. Castañeda,

M. Clotet, G. Comoretto, M. Davidson, C. Fabricius, G. Gracia, N. C. Hambly, A. Hutton, A. Mora, J. Portell, F. van Leeuwen, U. Abbas, A. Abreu, M. Altmann, A. Andrei, E. Anglada, L. Balaguer-Núñez, C. Barache, U. Becciani, S. Bertone, L. Bianchi, S. Bouquillon, G. Bourda, T. Brüsemeister, B. Bucciarelli, D. Busonero, R. Buzzi, R. Cancelliere, T. Carlucci, P. Charlot, N. Cheek, M. Crosta, C. Crowley, J. de Bruijne, F. de Felice, R. Drimmel, P. Esquej, A. Fienga, E. Fraile, M. Gai, N. Garralda, J. J. González-Vidal, R. Guerra, M. Hauser, W. Hofmann, B. Holl, S. Jordan, M. G. Lattanzi, H. Lenhardt, S. Liao, E. Licata, T. Lister, W. Löffler, J. Marchant, J.-M. Martin-Fleitas, R. Messineo, F. Mignard, R. Morbidelli, E. Poggio, A. Riva, N. Rowell, E. Salguero, M. Sarasso, E. Sciacca, H. Siddiqui, R. L. Smart, A. Spagna, I. Steele, F. Taris, J. Torra, A. van Elteren, W. van Reeven, and A. Vecchiato. Gaia Data Release 2. The astrometric solution. *A&A*, 616:A2, August 2018.

[16] L. A. Hillenbrand. On the Stellar Population and Star-Forming History of the Orion Nebula Cluster. *AJ*, 113:1733–1768, May 1997.

[17] M. Kounkel, L. Hartmann, L. Loinard, A. J. Mioduszewski, L. F. Rodríguez, G. N. Ortiz-León, M. D. Johnson, R. M. Torres, and C. Briceño. VLBA Observations of Strong Anisotripic Radio Scattering Toward the Orion Nebula. *AJ*, 155:218, May 2018.

[18] Aidan McBride and Marina Kounkel. Runaway Young Stars near the Orion Nebula. *ApJ*, 884(1):6, October 2019.

[19] T. Preibisch and E. Mamajek. *The Nearest OB Association: Scorpius-Centaurus (Sco OB2)*, volume 5, page 235. 2008.

[20] W. G. L. Pöppel and P. Marronetti. The kinematical characteristics of the CNM at mid b mid $gt=$ 10degr and the hypothesis of a local explosive event. *A&A*, 358:299–309, June 2000.

[21] K. Bekki. Dark impact and galactic star formation: origin of the Gould belt. *MNRAS*, 398:L36–L40, September 2009.

[22] E. Zari, H. Hashemi, A. G. A. Brown, K. Jardine, and P. T. de Zeeuw. 3D mapping of young stars in the solar neighbourhood with Gaia DR2. *A&A*, 620:A172, December 2018.

[23] Amin Farhang, Jacco Th. van Loon, Habib G. Khosroshahi, Atefeh Javadi, and Mandy Bailey. A three-dimensional map of the hot Local Bubble using diffuse interstellar bands. *Nature Astronomy*, 3:922–927, July 2019.

[24] Marina Kounkel, Kevin Covey, Genaro Suárez, Carlos Román-Zúñiga, Jesus Hernandez, Keivan Stassun, Karl O. Jaehnig, Eric D. Feigelson, Karla Peña Ramírez, Alexandre Roman-Lopes, Nicola Da Rio, Guy S. Stringfellow, J. Serena Kim, Jura Borissova, José G. Fernández-Trincado, Adam Burgasser, D. A. García-Hernández, Olga Zamora, Kaike Pan, and Christian Nitschelm. The APOGEE-2 Survey of the Orion Star-forming Complex. II. Six-dimensional Structure. *AJ*, 156(3):84, Sep 2018.