Anomaly Detection in Astrophysics – VAE Separates Interacting Binary Stars from Normal Red Giants

Abhigyan Acherjee
Georgetown University
Washington, DC
aa3320@georgetown.edu

Savannah Thais
City University of New York
New York, NY
savannah.thais@hunter.cuny.edu

J. L. Sokoloski
Columbia Astrophysics Lab
New York, NY
jeno@astro.columbia.edu

Abstract

Symbiotic binary stars, in which a white dwarf accretes from a red-giant companion can be difficult to detect yet are key to understanding binary stellar evolution and supernova progenitors. We present a novel application of a variational autoencoder (VAE) applied to optical and infrared colors of red giants measured in the SkyMapper, 2MASS, and AllWISE surveys. By learning the latent structure of red giants, the VAE flags anomalies, some of which align with known symbiotic binaries, successfully recovering most above the 90th percentile in anomaly scores. This analysis demonstrates the efficacy of unsupervised anomaly detection for uncovering hidden interacting binaries—and possibly other populations of astronomical objects with small training sets—in forthcoming large surveys such as Rubin Observatory's LSST.

1 Introduction

Astronomical objects whose populations are heterogeneous, or whose physical properties are not fully understood, can be challenging to find and characterize using traditional astronomical methods. For example, until recently, an entire class of binary stars may have been almost completely missed – wide interacting systems called symbiotic binaries, in which a white dwarf accretes matter from a red-giant (RG) companion. Because RG stars are optically bright, their light can hide some of the most obvious optical signatures of an accreting white-dwarf companion. Although one type of symbiotic, with nuclear burning on the surface of the white dwarf, is fairly easy to find with optical spectroscopic survey (examples, references), the other, potentially more important type without nuclear burning on the white-dwarf surface is much harder to discover. But because wide pairs of the low-mass stars that become white dwarfs and RGs are very common, the prevalence of interacting white-dwarf/RG binaries could actually be quite high, with published estimates differing by many orders of magnitudes. Moreover, quantifying the population of these interacting binaries in our galaxy would have far-reaching implications – for binary stellar evolution, for the physics and accretion and nuclear burning, and for the origin of the type of supernovae used to study the expansion of the universe and dark energy.

The research described here is motivated by the impending start of the largest wide-field astronomical survey, to be conducted with the Rubin Observatory, in Chile. Between roughly 2026 and 2036, Rubin Observatory will generate 20 TB of data per night and transform the way astrophysicists do science. It will use its massive camera and unique telescope design to map the southern sky every

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

few nights for ten years, creating the deepest and most detailed image of the sky and also a 'movie' of astronomical objects that move and vary in brightness. Given the volume and velocity of the data, ML methods will be needed to sift through Rubin data for interacting binary stars, and for many if not most other investigations with Rubin data.

In this contribution, we describe a pilot study with data from SkyMapper data-release 2 (DR2) to show that a variational autoencoder (VAE) trained on features constructed from optical and infrared brightness measurements of several hundred thousand red-giant stars can successfully identify known symbiotic binaries as anomalies. VAEs are unsupervised models that compress data, using an encoder model, into a probabilistic latent space and reconstruct it through a decoder model. By learning the distribution of data, they capture its underlying structure. Inputs that deviate from this distribution reconstruct poorly, yielding high reconstruction loss. This property makes VAEs well-suited for anomaly detection in physical science domains such as astrophysics, physics and medical imaging.

We describe our methods in Section 2 and preliminary results in Section 3. We discuss these findings and next steps in Section 4.

2 Methods

We followed the prescription in Lucy et al. [1] to build a sample of red-giant stars from among the southern-sky astronomical objects in the Skymapper Southern Sky Survey [6] of optical brightness measurements taken between 2014 and 2018 through u, v, g, r, i, and z filters. We selected objects with a set of features identified in Lucy et al. [1], which gave us a sample of likely red giant stars with a range of evolutionary states and/or surface chemistries including S stars, carbon stars, Miras, and post-AGB stars. To build this sample, we extracted optical brightnesses, infrared (IR) brightnesses, and distance estimates from 4 different catalogues–Skymapper [6], 2MASS [8], GAIA [7] and AllWISE [9]. The SQL query that was used by us (and Lucy et al. [1]) is provided in the Appendix A. After making the initial extraction, we made a series of additional cuts to exclude objects whose observing cadence led to unreliable colors. Here, and throughout the document color refers to the difference between the brightness in two observing bands. For instance, $u - v$ refers to the color generated by subtracting v band brightness from u band. Colors are crucial features for distinguishing between types of astronomical objects because they capture underlying physical properties with the distance dependence of intrinsic (rather than observed) brightnesses canceling out. We primarily worked with u, v, g, r, i, and z filters, the WISE bands $w1$,$w2$,$w3$ and $w4$, as well as the j,h,and k bands.

We used AstroPy's [5] TAP-Plus service to execute the SQL query outlined in [1], and then made the cuts on intrinsic J, H and K band brightnesses to remove red dwarfs, which are intrinsically faint, from the sample. This gave us a set of 366,721 likely red giants. To assess the potential of the VAE methodology, we constructed features based on a wide range of colors. Lucy et al. [1] limited the colors used in their cut-based analysis to $u - g$ and $u - v$, to focus on the expected blue excess from any accreting white-dwarf companion. However, Akras et al. [2] suggested that IR colors $w3 - w4$ and $w2 - w1$ from the WISE satellite, along with ground-based $J - H$ colors can also be utilized in the identification of symbiotic stars. We therefore computed the set of all possible SkyMapper colors using the u, v, g, r, i, and z measurements available in SkyMapper, leading to a list of 15 colors just from SkyMapper, and a reduced sample of 242,298 likely red giants for which this full set of colors (including colors generated from IR bands j,h and k as well as the WISE band) was available.

We hypothesized that these colors could also form a useful basis for training a variational autoencoder(VAE) to identify physically meaningful anomalies within our sample. In order to select the specific features to use as inputs to the VAE, we first used principal component analysis (PCA). Using different sets of colors as inputs, we tested the number of principal components that preserved the most variance from the original features. We found that using four principal components consistently preserved a large amount of variance, with diminishing returns seen from including additional principal components, as shown in Figure 1.

We found that different combinations of input features to the PCA yielded approximately similar results, so we chose to move forward with 5 different sets of PCA-derived features. We then implemented a variational autoencoder architecture with a narrow bottleneck consisting of a 3-dimensional latent space. We followed a symmetric encoder-decoder structure that balances model capacity with generalization. We computed the reconstruction loss in principal components space

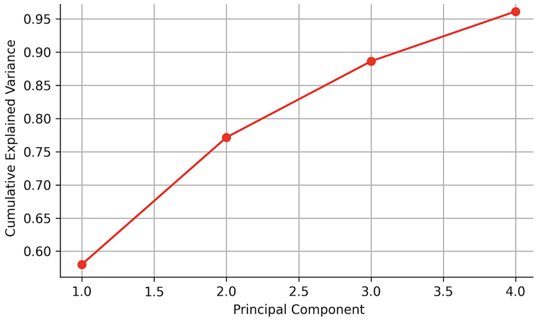

Figure 1: Cumulative variance (y-axis) captured by principal components (x-axis)

rather than original feature (color) space, so that it focused anomaly detection on the most informative dimensions and reduced sensitivity to noise in less important features.

The VAE was trained using standard mean squared error for loss calculation (MSE) along with Adam optimizer, and we used the same seed throughout to preserve reproducibility. We can thus consider objects with large reconstruction losses from the VAE as candidates for unusual red giants or rare astrophysical objects.

For determining the rate at which anomalies corresponded to known symbiotic binaries, we used Lucy et al. [1]'s list of 56 known symbiotic binaries, of which 53 were contained among our sample of 242,298 likely red giants. This list included symbiotics from Merc et al. [4] as well as symbiotics identified by Lucy et al. [1] in their work. We carried out this analysis with different loss thresholds such as 1%, 0.1% and 2%. Because the loss threshold was arbitrary, we also looked at the location of symbiotic binaries within the broader feature space of our set of likely red giants.

3 Results and discussion

Cross-matching anomalies that the VAE identified in our distribution with known symbiotic binaries from Lucy et al. [1] confirmed that VAE can successfully identify these interacting binaries as anomalous. In our VAE analysis, all five groups of PCA-derived features performed well in identifying anomalies using the reconstruction loss, based on a cross-match with the list of known symbiotics [1]. For example, using the 2% threshold, 30 of the 53 known symbiotics in our sample were identified as anomalous, for the group of variables $u - g$, $g - i$, $r - z$, $j - h$, $h - k$, and $w1 - w2$, indicating that the list of anomalous objects generated by the reconstruction loss is potent for highlighting and identifying real symbiotics. Table 1 demonstrates that almost 86% of symbiotics appeared above the 90th percentile, 64% appeared above the 95% percentile, and only a small minority appeared at or below the 50th percentile.

Percentile	Number of Symbiotics	% of Symbiotics
0-50th	4	7.5%
50-90th	8	15.1%
90-95th	7	13.2%
95-99th	22	41.5%
99-100th	12	22.6%

Table 1: Number of known symbiotics at each loss percentile for feature set u-g,g-i,r-z,j-h,h-k,w1-w2.

The success of the VAE methodology was robust, and suggested that anomaly-detection algorithms can provide insight into underlying characteristics of astronomical objects. The groups of features we fed into the VAE pipeline with a 1% threshold were not only able to identify a significant number of symbiotics with a wide range of color properties. For instance, the group of features consisting of the colors $u - g$, $g - i$, $r - z$, $j - h$, $h - k$, and $w1 - w2$ (top row of Table 2) yielded symbiotic stars that appear in the middle of the distribution in the $u - g$ vs $u - v$ plot, as shown in Figure 2.

Table 2 shows the success of the VAE for different sets of color features. The higher or lower rate of success with which different sets of color features lead to the identification of known symbiotics as

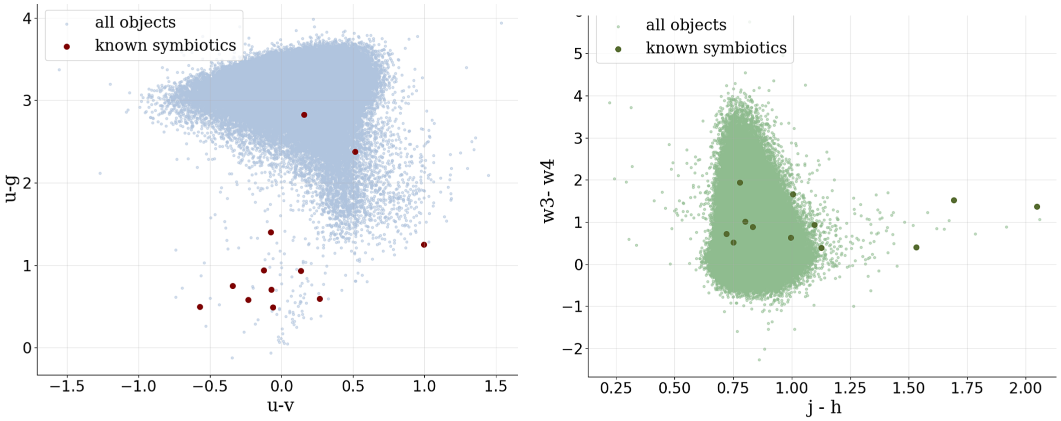

Figure 2: The distribution of normal red giants and symbiotic binaries identified as anomalous by the VAE on two color-color plots. *Left:* Scatterplot for $u - g$ vs $u - v$. The small blue points correspond to normal red giants, and the larger red points are known symbiotics that overlap with the anomalies generated by the VAE loss reconstruction. *Right:* Scatterplot for $w3 - w4$ vs $j - h$. The small green points correspond to normal red giants, and the larger blue point are known symbiotics identified as anomalous by the VAE.

anomalous has the potential to provide physical insight, especially as we move to larger data sets with more features (even beyond color features). The success of our VAE methodology at identifying recently confirmed but difficult to find interacting binaries suggests that other anomalies identified by the VAE are excellent candidates for the discovery of additional interacting binaries.

Additionally, we decided to look at an approach that used just the variational autoencoder to find anomalies, bypassing the dimensionality reduction of the PCA. The results of these for each group variables is provided in Table 2 . Each number represents the average number of anomalies that were confirmed as known symbiotic stars, from an average of 5 runs of our VAE pipeline. We see that the VAE only pipeline yields a larger number of symbiotic matches than the VAE plus PCA pipeline. This is due to the dual information loss that occurs during the dimensionality reduction of the PCA, followed by the reduction due to the VAE.

Group of Parameters	Number of Symbiotics (PCA+VAE)	Number of Symbiotics (VAE)
u-g,g-i,r-z,j-h,h-k,w1-w2	12	28
g-i,w3-w4,r-z,j-h,h-k,j-k	12	8
g-r,w1-w2,g-i,j-h,h-k,j-k	10	13
v-g,w3-w4,r-z,j-h,h-k,j-k	7	7
u-g,w1-w2,u-v,j-h,h-k,j-k	23	27

Table 2: Number of known symbiotics identified by VAE loss of each feature set with a 1% threshold (rounded to nearest integer based on the average of 5 runs), when considered with PCA dimensionality reduction prior to VAE and when considered without PCA dimensionality reduction.

Our results from this run led us to believe that using just a variational auto-encoder to generate anomalies leads to more matched symbiotic stars. Since our initial set of experiments consisted of groups of 6 variables each, we decided to increase that to include all possible colors that could be generated by the band u,v,g,r,i,z, the infrared bands of j, h and k and the WISE band colors w1-w2 and w3-w4. We chose these particular WISE band colors due to their usefulness in identifying anomalies in diagnostic color-color diagrams. We also wanted to look at the effect of varying the number of dimensions that we allowed the VAE to reduce the latent space to, and look at the subsequent results. These are shown in Table 3 . For purposes of statistical robustness, we computed the number of matched symbiotics which were the average of 5 runs. This group included the following colors: $u - g$, $u - v$, $u - i$, $u - r$, $u - z$, $v - g$, $v - r$, $v - i$, $v - z$, $g - r$, $g - i$, $g - z$, $r - i$, $r - z$, $i - z$, $j - h$, $h - k$, $j - k$, $w1 - w2$, $w3 - w4$.

Number of Dimensions	Number of Symbiotics (Merc Galatic and Skymapper Confirmed)
3	46.80
4	44.20
5	40
6	36.80

Table 3: Number of known symbiotics identified by VAE loss of the feature set u-g,u-v,u-i,u-r,u-z,v-g,v-r,v-i,v-z,g-r,g-i,g-z,r-i,r-z,i-z,j-h,h-k,j-k,w1-w2,w3-w4 with a 1% threshold, for different dimensions

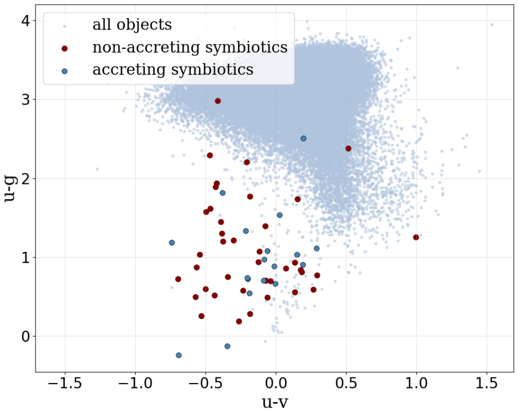

Figure 3: Distribution of accreting vs non-accreting symbiotic stars from the Merc and Skymapper directories.

With a lower-dimensional latent space , the VAE is constrained in its ability to accurately reconstruct the diverse population of luminous red giants. As the latent dimension increases, the VAE gains representational capacity, allowing it to learn more sophisticated encodings that better capture the complex patterns present in the normal population. This leads to a tighter, more concentrated distribution of reconstruction losses for the majority class, while the reconstruction errors for symbiotic stars also decrease but to a lesser extent. When using a percentile-based threshold to identify anomalies, this compression of the loss distribution means that fewer objects exceed the threshold, resulting in fewer detected anomalies overall. Consequently, some symbiotic stars that were previously flagged as anomalies in the 3-dimensional model may no longer stand out sufficiently in the 6-dimensional model, as their reconstruction errors become less distinguishable from the improved normal object reconstructions.

Finally, to visualize the anomalies better, we decided to plot the population of accreting and non-accreting symbiotic stars from merc and skymapper on the same color-color plot. The results are shown in the Figure 3. We note that certain Skymapper anomalies appear to exist outside the population of red giants in the u-g vs u-v diagnostic color-color diagram.

4 Conclusions

Selection criteria based directly on color features are often insufficient for identifying symbiotic binary stars. This pilot study demonstrated that ML anomaly detection methods such as VAE show promise for finding such astronomical objects using large astronomical datasets. Our VAE identified some symbiotic binaries as anomalous compared to normal red giants even when those symbiotics did not inhabit the fringes of standard color-color plots. It successfully identified many symbiotic binaries as anomalous despite only using color features, whereas other searches have required optical spectroscopy and/or measures of rapid brightness variability. This initial work has strong promise for the discovery of previously unidentified symbiotics. We also believe that incorporating a time domain feature into our VAE pipeline can yield better matches with symbiotics that don't appear anomalous in our current feature set.

Acknowledgments and Disclosure of Funding

We would like to thank Dr. Adrian Lucy for their support and continual clarifications provided during the process of acquiring and cleaning the SkyMapper data. This work was funded by a seed grant from the Columbia University Data Science Institute.

References

[1] A B Lucy, J L Sokoloski, G J M Luna, K Mukai, N E Nuñez, D A H Buckley, H Breytenbach, B Paul, S B Potter, R Manick, D A Howell, C Wolf, C A Onken (2025). A new way to find symbiotic stars: accretion disc detection with continuum survey photometry. *Monthly Notices of the Royal Astronomical Society*. doi:10.1093/mnras/staf1351

[2] Akras, S., Leal-Ferreira, M. L., Guzman-Ramirez, L., & Ramos-Larios, G. (2019). A machine learning approach for identification and classification of symbiotic stars using 2MASS and WISE. *Monthly Notices of the Royal Astronomical Society*, **483**(4), 5077–5104. doi:10.1093/mnras/sty3359

[3] Lucy, A., Sokoloski, J., Luna, G., Mukai, K., Nuñez, N., Buckley, D., Breytenbach, H., Paul, B., Potter, S., Manick, R., Howell, D., Wolf, C., & Onken, C. (2024). A new way to find symbiotic stars: accretion disc detection with optical survey photometry. arXiv preprint arXiv:2412.00855. doi: 10.48550/arXiv.2412.00855.

[4] Merc, J., Gális, R., & Wolf, M. (2019). First Release of the New Online Database of Symbiotic Variables. *Research Notes of the AAS*, **3**(2), 28. doi: 10.3847/2515-5172/ab0429.

[5] Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. (2013). Astropy: A community Python package for astronomy. *Astronomy & Astrophysics*, **558**, A33. doi: 10.1051/0004-6361/201322068

[6] Onken, C. A.; Wolf, C.; Bessell, M. S.; Chang, S.-W.; Da Costa, G. S.; Luvaul, L. C.; Mackey, D.; Schmidt, B. P.; Shao, L., 2019, "SkyMapper Southern Survey: Second data release (DR2)," *Publications of the Astronomical Society of Australia*, **36**, e033, August 2019. doi: 10.1017/pasa.2019.27.

[7] Vallenari, A.; Brown, A. G. A. et al., 2023, "Gaia Data Release 3: Summary of the content and survey properties," *Astronomy & Astrophysics*, **674**, A1, June 2023. doi:10.1051/0004-6361/202243940

[8] Skrutskie, M. F.; Cutri, R. M.; Stiening, R.; Weinberg, M. D.; Schneider, S.; Carpenter, J. M.; Beichman, C.; Capps, R.; Chester, T.; Elias, J.; Huchra, J.; Liebert, J.; Lonsdale, C.; Monet, D. G.; Price, S.; Seitzer, P.; Jarrett, T.; Kirkpatrick, J. D.; Gizis, J. E.; Howard, E.; Evans, T.; Fowler, J.; Fullmer, L.; Hurt, R.; Light, R.; Kopan, E. L.; Marsh, K. A.; McCallon, H. L.; Tam, R.; Van Dyk, S.; Wheelock, S., 2006, "The Two Micron All Sky Survey (2MASS)," *The Astronomical Journal*, **131**(2), 1163–1183, February 2006. doi:10.1086/498708

[9] Wright, E. L.; Eisenhardt, P. R. M.; Mainzer, A. K.; Ressler, M. E.; Cutri, R. M.; Jarrett, T.; Kirkpatrick, J. D.; Padgett, D.; McMillan, R. S.; Skrutskie, M.; Stanford, S. A.; Cohen, M.; Walker, R. G.; Mather, J. C.; Leisawitz, D.; Gautier, T. N. III; McLean, I.; Benford, D.; Lonsdale, C. J.; Blain, A.; Mendez, B.; Irace, W. R.; Duval, V.; Liu, F.; Royer, D.; Heinrichsen, I.; Howard, J.; Shannon, M.; Kendall, M.; Walsh, A. L.; Larsen, M.; Cardon, J. G.; Schick, S.; Schwalm, M.; Abid, M.; Fabinsky, B.; Naes, L.; Tsai, C., 2019, "AllWISE Source Catalog," NASA/IPAC Infrared Science Archive (IRSA1) DataSet, January 2019. doi:10.26131/IRSA1

A Technical Appendix

```
SELECT
   m.object_id, g.source_id, m.raj2000,  m.dej2000, m.glon,
   m.glat, m.u_ngood, m.u_nclip, m.v_ngood, m.g_ngood,
   m.r_ngood, m.i_ngood, m.z_ngood, w.w1mpro, w.w1sigmpro,
   w.w2mpro, w.w2sigmpro, t.j_m, t.j_msigcom, t.h_m,
   t.h_msigcom, t.k_m, t.k_msigcom, m.u_psf, m.e_u_psf, m.v_psf,
   m.e_v_psf, m.g_psf, m.e_g_psf, m.r_psf, m.e_r_psf, m.i_psf,
   m.e_i_psf, m.z_psf, m.e_z_psf, m.prox, t.prox AS tprox,
   m.allwise_dist, g.parallax, g.parallax_error,
   g.astrometric_excess_noise, g.astrometric_excess_noise_sig,
   g.pmra, g.pmra_error, g.pmdec, g.pmdec_error, m.ebmv_sfd
FROM
    dr2.master m
JOIN
    ext.twomass_psc t ON m.twomass_key=t.pts_key
JOIN
    ext.gaia_dr2 g ON m.gaia_dr2_id1=g.source_id
JOIN
    ext.allwise w ON m.allwise_cntr=w.cntr
WHERE
    m.twomass_dist < 2.0 /* cross-matching radii (arcsec) */
    AND m.gaia_dr2_dist1 < 2.0
    AND m.allwise_dist < 3.0
    AND m.prox > 6.0
    AND t.ph_qual = ’AAA’ /* quality cuts */
    AND t.gal_contam = 0
    AND t.ext_key IS NULL
    AND t.cc_flg = ’000’
    AND m.class_star > 0.9
    AND flags_psf = 0
    AND m.u_ngood > 0
    AND m.v_ngood > 0
    AND m.g_ngood > 0
    AND m.nch_max = 1
    AND (t.j_m - t.k_m) > 0.85 /* 2MASS initial color cut */
    AND t.j_m < 14.0 /* Select for high 2MASS SNR */
```

Figure 4: SQL query used to create list of red giants