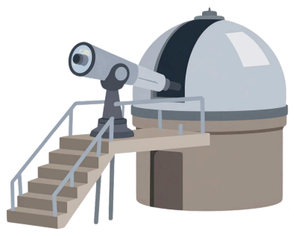

Anomaly Detection in Astrophysics – VAE Separates Interacting Binary Stars from Normal Red Giants

Abhigyan Acherjee[1], Savannah Thais[2], and J.L. Sokoloski[3]

01: AIM

We aim to develop an unsupervised anomaly-detection framework using a variational autoencoder (VAE) trained on optical and infrared colors of a population of luminous red objects from SkyMapper, 2MASS, and AllWISE. By learning the latent structure of typical red giants, this approach will identify anomalous objects—recovering known symbiotic binaries . This framework will help reveal rare binary populations in current datasets and scale to forthcoming large surveys such as Rubin Observatory's LSST survey.

02: DATA

We used the following catalogues as data sources in our experiments, but our focu s was primarily on the Skymapper catalogue.
Population of luminous red objects: *2MASS, ALLWISE, GAIA* and *Skymapper*.
Confirmed Symbioitcs: *Merc Catalogue* of confirmed symbiotic stars.
Distance metrics: *Bailer Jones* catalogue.

03: DATA MINING

The data was primarily present in dataset. To extract, it we followed the following steps:

Data Mining
Constructed a red-giant sample by querying SkyMapper, 2MASS, GAIA, and AllWISE, applying cuts to remove unreliable observations and red dwarfs, yielding ~366k likely red giants (later narrowed to ~242k with full color coverage).

Cleaning Data
Cleaned data and removed non-regular night captures and extracted a set of optical and IR brightnesses to capture diverse stellar properties and enable broader anomaly detection than prior cut-based methods.

Merge catalogues
Merged catalogues and the u,v,g,r,i,z bands to get our full sample of ~242k luminous red objects.

04: METHODOLOGY

Our main findings from the data are outlined below. We notice wide variations and state level differences in the analysis

Color Feature Construction & Selection
Generated a wide set of optical, infrared, and WISE color features, tested multiple combinations, and used PCA to identify the no. of principal components that preserved most variance for use as compact VAE inputs. (Fig 2)

VAE Architecture & Training:
Implemented a symmetric encoder–decoder VAE with a 3-dimensional latent bottleneck, trained using MSE reconstruction loss in PCA space (or original color space for VAE-only), with anomalies defined by high reconstruction errors.

PCA+VAE vs. VAE-Only Evaluation:
Compared anomaly detection performance across 5 feature sets, once combining PCA with VAE and once without. Additonally, varied the number of dimensions and included multiple colors to compare matches with confirmed symbiotic stars.

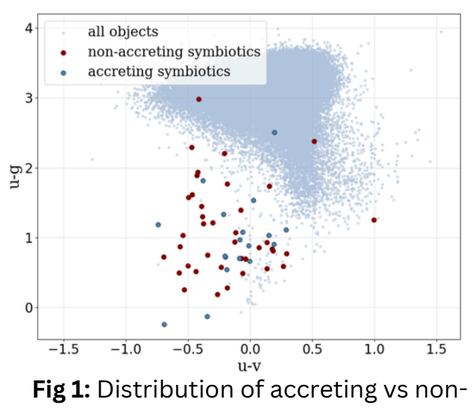

Fig 1: Distribution of accreting vs non-accreting symbiotics.

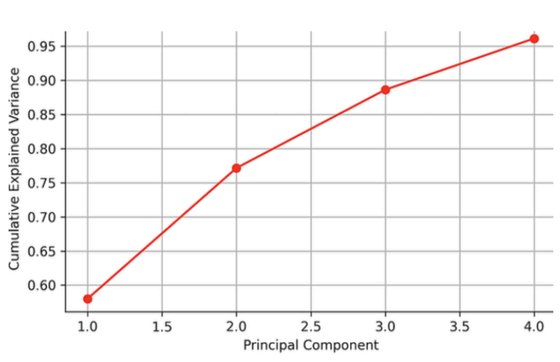

Fig 2 : Cumulative variance (y-axis) captured by principal components (x-axis)

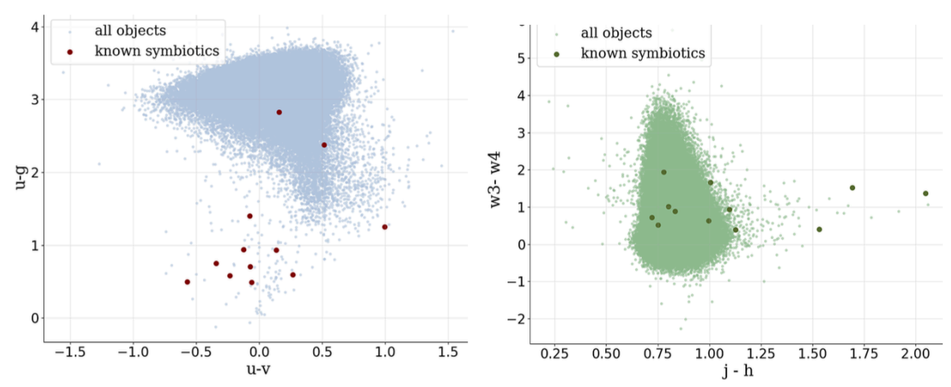

Fig 3: Left: Scatterplot for u− g vs u− v. The small blue points correspond to normal red giants, and the larger red points are known symbiotics that overlap with the anomalies generated by the VAE loss reconstruction. **Right:** Scatterplot for w3− w4 vs j− h. The small green points correspond to normal red giants, and the larger blue point are known symbiotics identified as anomalous by the VAE.

05:RESULTS

Our findings were as follows:

- The VAE reliably identified known symbiotic binaries as anomalies, with ~86% of known systems scoring above the 90th percentile in reconstruction loss, demonstrating strong separation between interacting binaries and normal red giants even when symbiotics occupy the center of traditional color–color diagrams rather than their extremes.
- Across five initial feature sets, anomaly recovery remained robust, but experiments showed that bypassing PCA and using the full color space directly in the VAE nearly always recovered more symbiotics. This is shown in Table 2. This suggests that PCA introduces information loss that suppresses subtle but physically meaningful color signatures.
- Expanding the VAE input to all 20+ optical and IR colors and varying latent dimensionality revealed that smaller latent spaces (≈3 dimensions) produced the strongest anomaly signals, recovering up to ~47 known symbiotics.
- Color–color visualizations confirmed that VAE-flagged anomalies include both accreting and non-accreting symbiotic stars(Fig 1) , with some anomalies lying outside the red-giant locus in diagnostic diagrams—highlighting the VAE's ability to detect physically unusual systems that traditional color cuts fail to isolate.

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 dimensionalityreduction prior to VAE and when considered without PCA dimensionality reduction.

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.

06:CONCLUSION

In conclusion, we demonstrated that variational autoencoders provide a powerful, unsupervised framework for identifying interacting symbiotic stars hidden within large populations of red giants. By learning the underlying latent structure , the VAE effectively flagged known symbiotic binaries as anomalies, even when they did not occupy extreme regions of traditional color–color diagrams. Our experiments further showed that the choice of feature sets and latent dimensionality plays a critical role in anomaly separability. In particular, increasing the latent dimension reduced separability by tightening the reconstruction-loss distribution around the normal red-giant population, causing symbiotic binaries to blend in and fall below anomaly thresholds. As upcoming surveys such as Rubin Observatory's LSST generate unprecedented volumes of photometric data, this approach, incorporating time-domain features—may enable more complete censuses of interacting binaries and other rare stellar populations.

1: Georgetown University 2:CUNY-Hunter College 3: Columbia University