The needle in a haystack: Hunting for Exo-Comets

Starry Knights:
Ally Baldelli, Joe Bonin, Christopher Montalban, Fernanda Muñoz
Supervisor: Amelia Bayo

Pontificia Universidad Católica de Chile

Columbia Engineering

NOIR Lab

AURA

American Museum of Natural History

Universidad de la Serena, Chile

Background

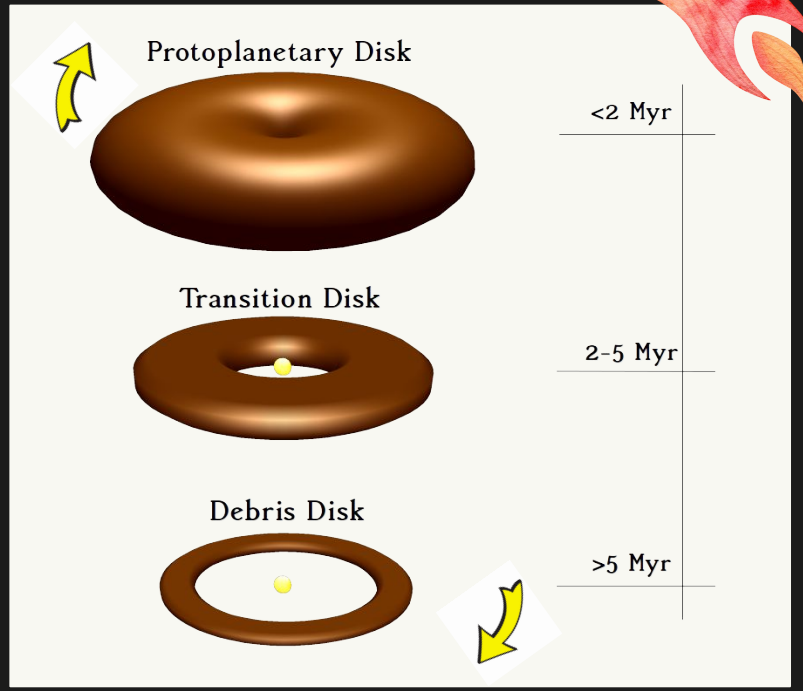
What is an exocomet?

- Small bodies made of rock, ice, dust and gas found outside our solar system
- First observed in 1987 as “falling evaporating bodies” **(FEBs)** in the Beta Pictoris system
- Detected far less frequently than exoplanets

Why study them?

- To provide information on the conditions of formation and evolution of stellar systems

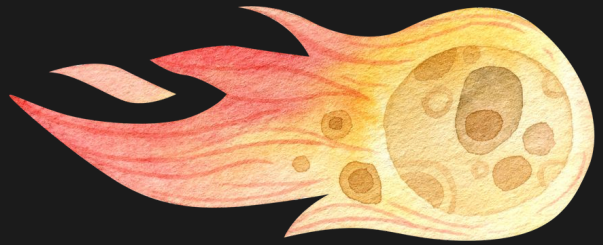
- Relevant in debris disc studies: second-generation dusty discs thought to be devoid of gas

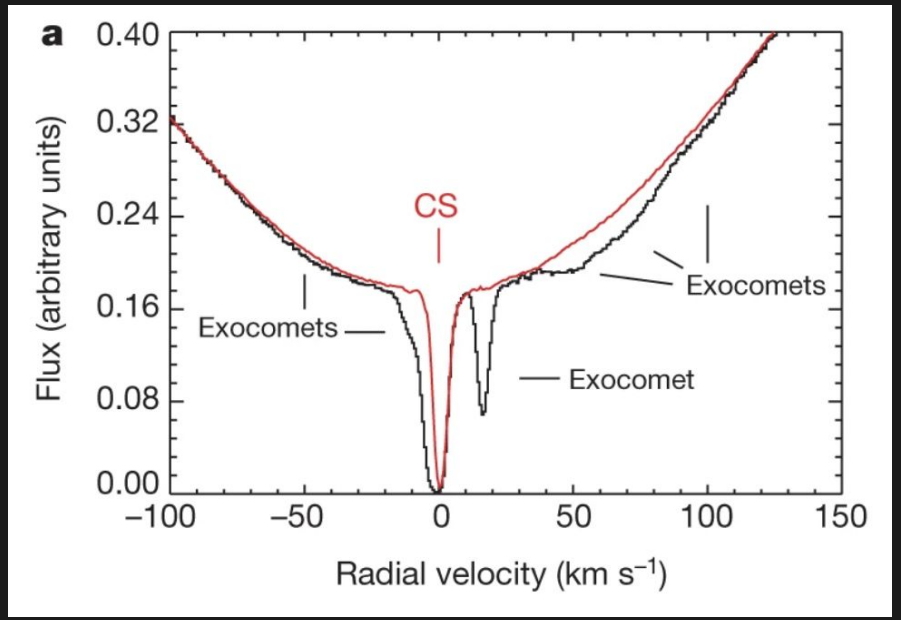


How to study them?

Discoveries rely on observing transient calcium lines in stellar spectra: Ca II K-line at 3,933.66 Å and the Ca II H-line at 3,968.47 Å

These features **are interpreted as exocomets transiting in front of the stellar disk**





A typical Ca II spectrum of β Pic (Ca II K-line) Solid black lines indicates the changes in flux. **Kiefer et al. 2014**

Objectives

- Applying advanced techniques like deep learning and anomaly detection in latent spaces

- Transforming the detection process into an image-based anomaly detection task, potentially enhancing the identification and understanding of exocomets.

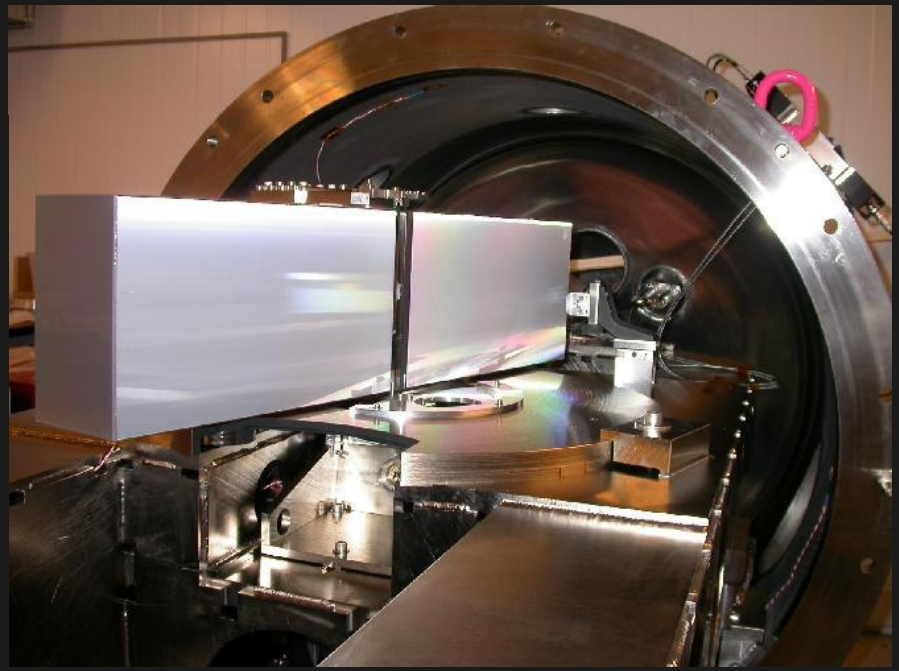
The number of known exocomets **remains limited due to biases** in data collection and analysis.

Data set used

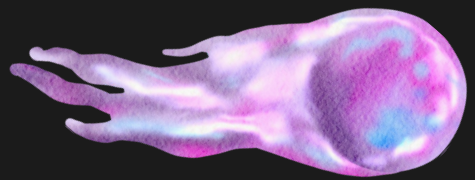
HARPS (High Accuracy Radial velocity Planet Searcher) spectrograph

HARPS radial velocities catalog by Mauro Barbieri released in 2023

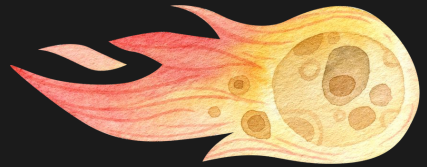
Observations from 2003 to 2023



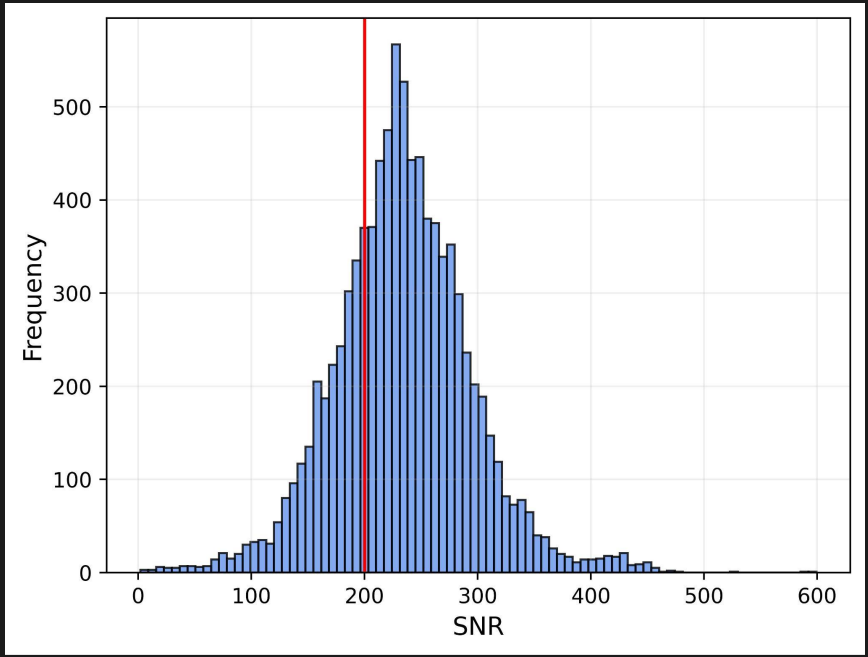
Credits: ESO



Data set used: Beta Pictoris

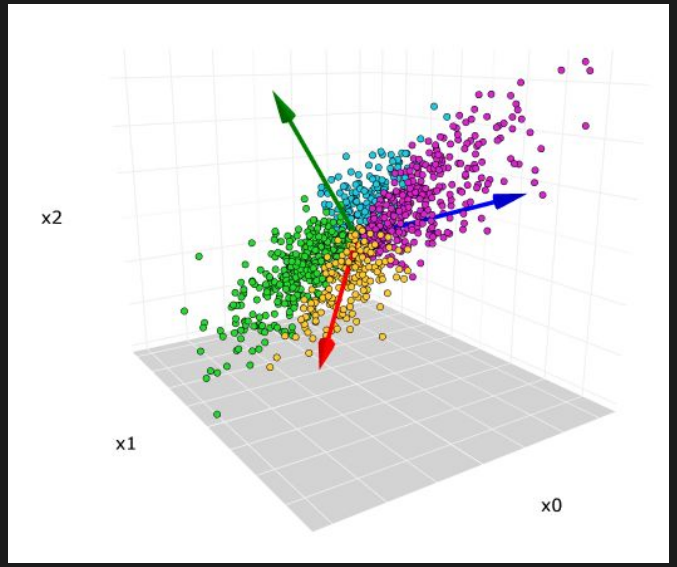


- ADQL query for HARPS Radial Velocities Catalog: 9,093 matching datasets
- Criteria:
  - SNR > 200
  - Cutout region with wavelength range: 3,500 - 4,500 Å
- Matching datasets: 6,676
- DataFrames with spectra: each column represents a wavelength with an associated flux

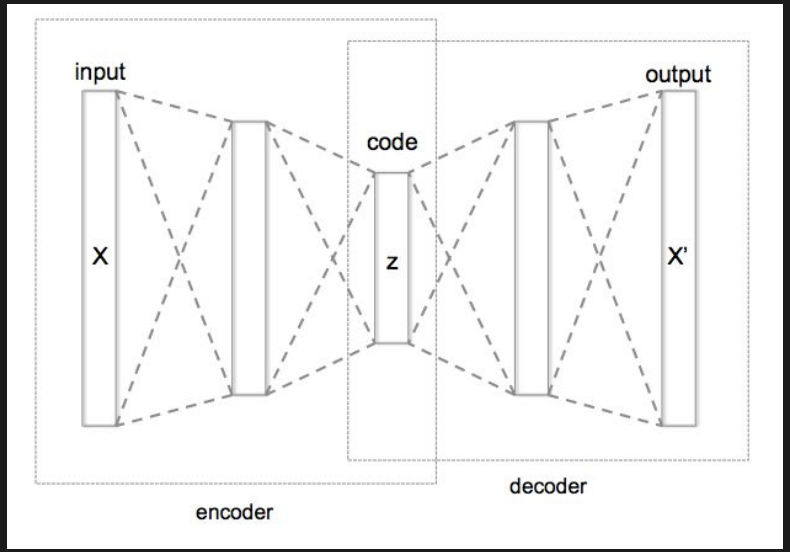


Methods applied to solve the problem

- PCA and Clustering



- Autoencoder



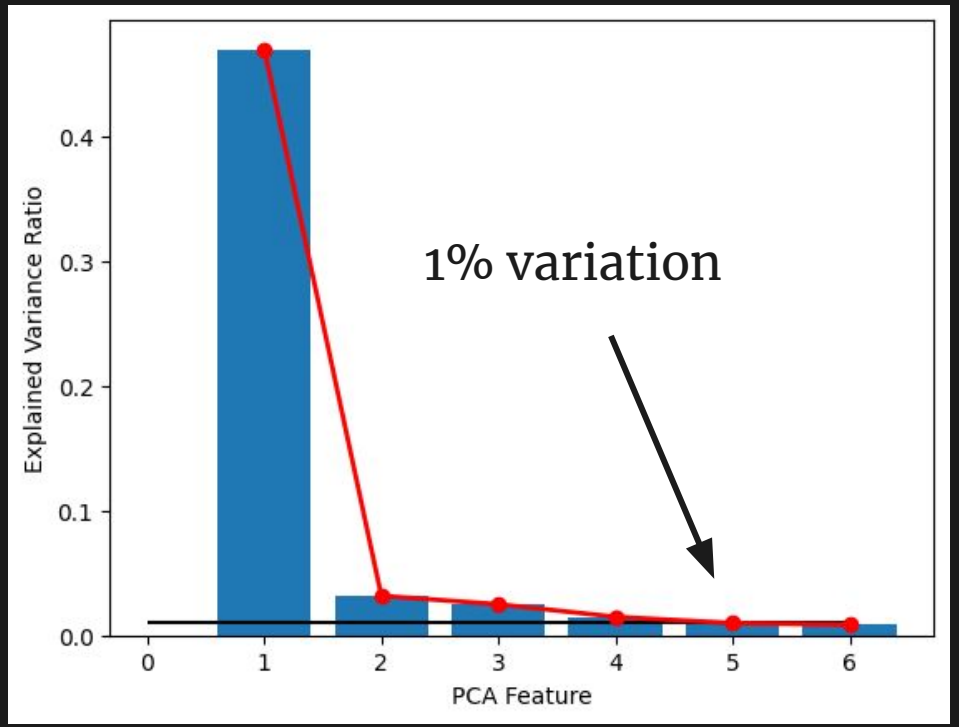
PCA METHOD

PCA Method

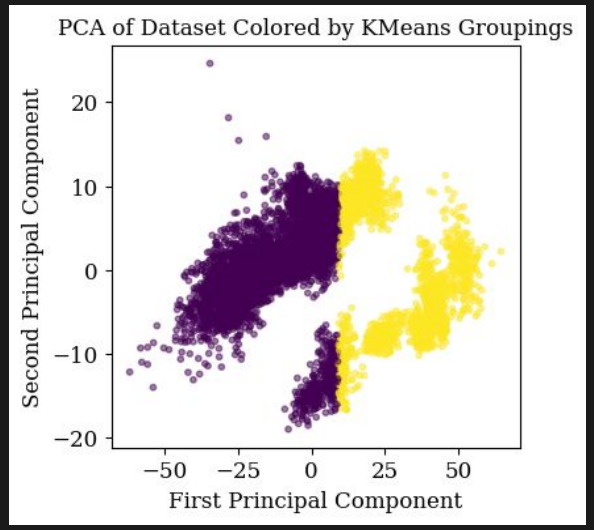
Dimensionality reduction algorithm

We choose to do 5 principal components to reduce variation to one percent

Anymore and there is no variation to create clusters

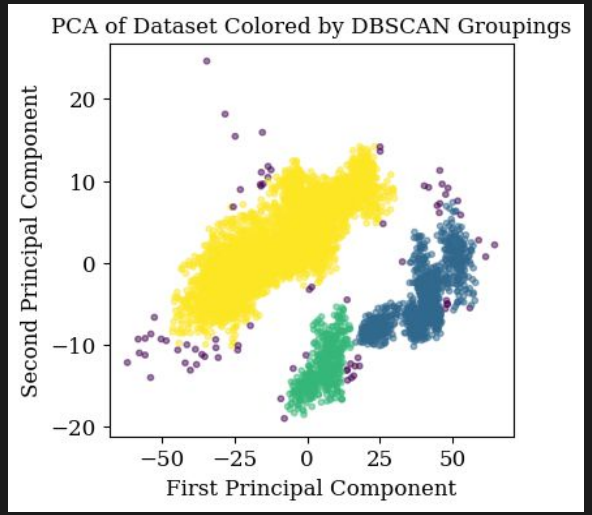


PCA Clustering: Choosing a Method



Pros: Simple method

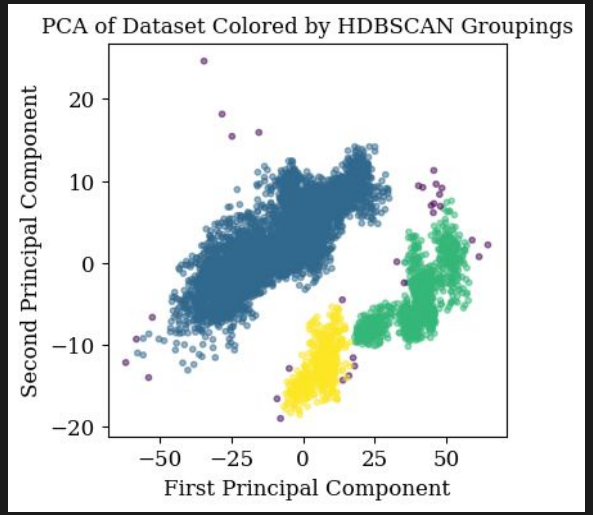
Cons: ... not the best clusters



Pros: Excellent clustering

Cons: Sensitive to densities

Computationally intense



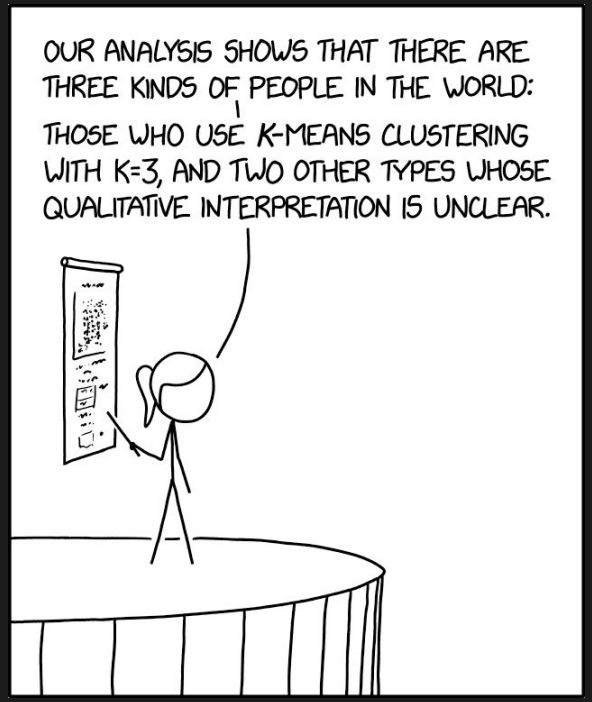
Pros: Excellent clustering

NOT Sensitive to densities

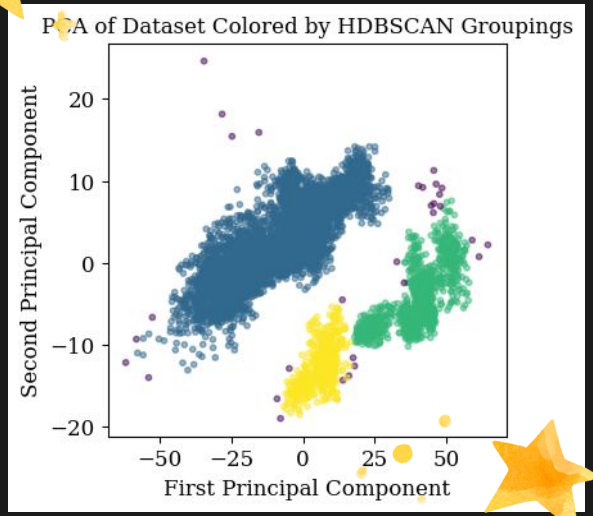
Cons: Computationally intense

Note: Only using two dimensions here for demonstration

Winning Method!



An “unclear” qualitative interpretation

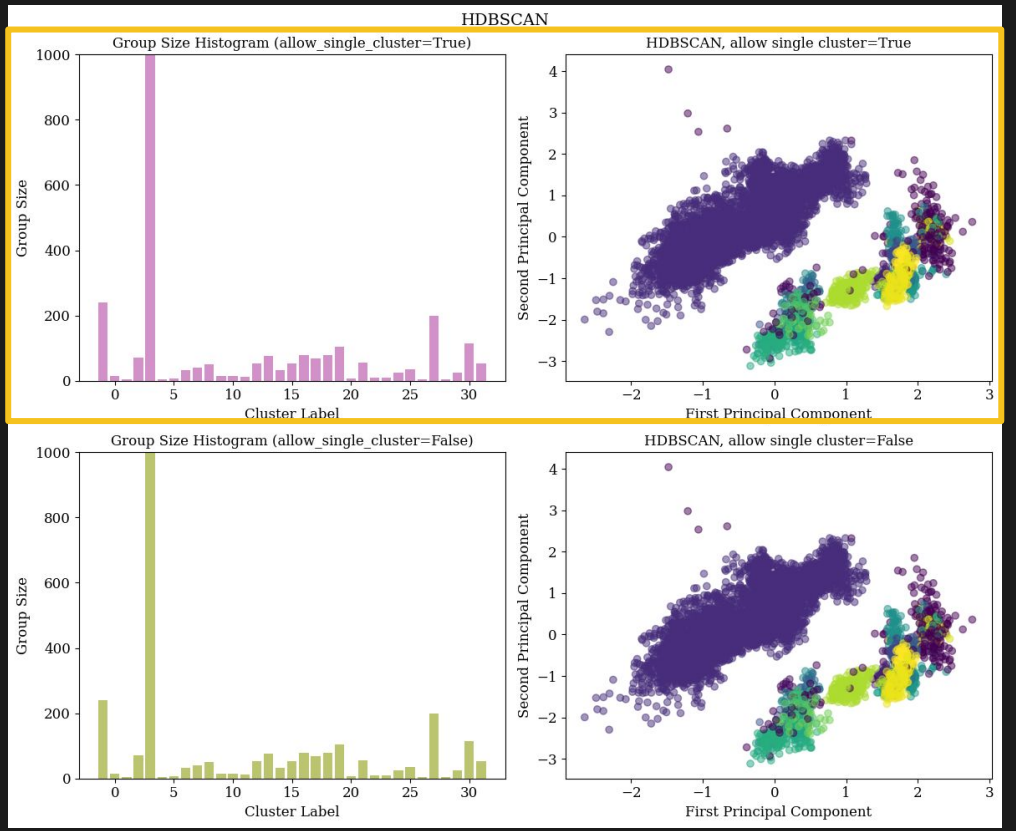


Note: Only using two dimensions here for demonstration

Using HDBSCAN allow single cluster?

We don’t want to force clusters so we choose allow single cluster to True

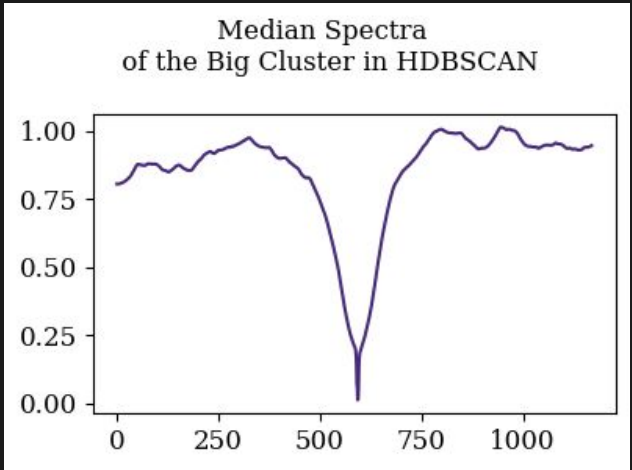
We maintain similar distributions here demonstrating suggesting we have good clusters in our dataset

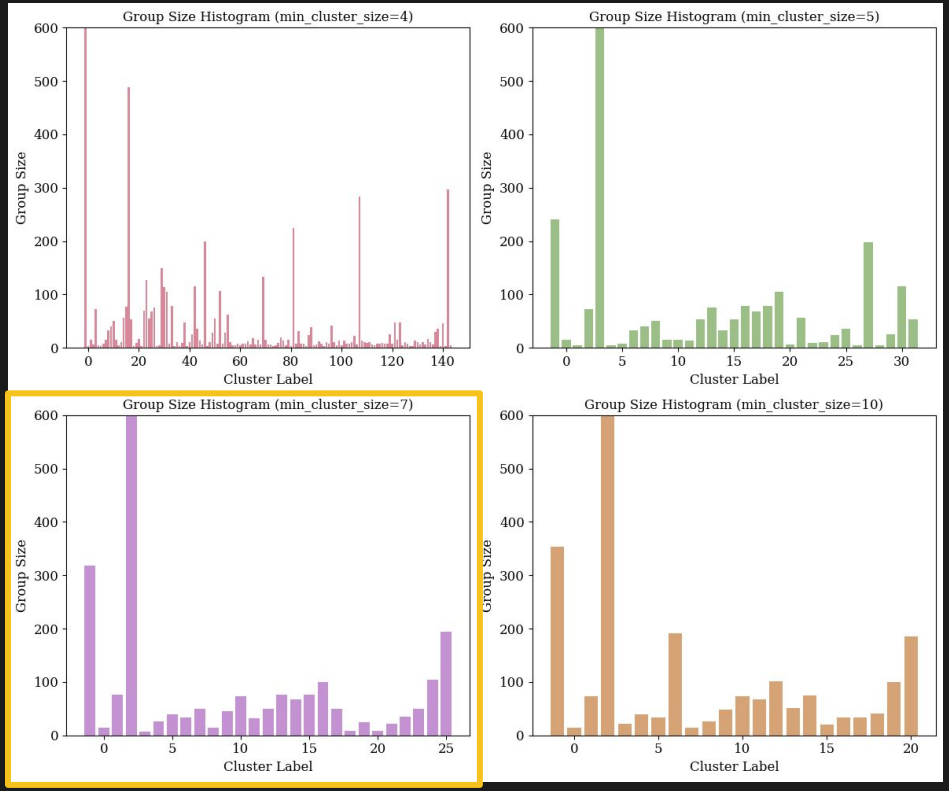


Using HDBSCAN varying minimum cluster size

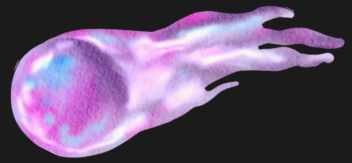
We want smooth groupings

However no matter how it is clustered the largest group remains which is the baseline with theoretically no comets



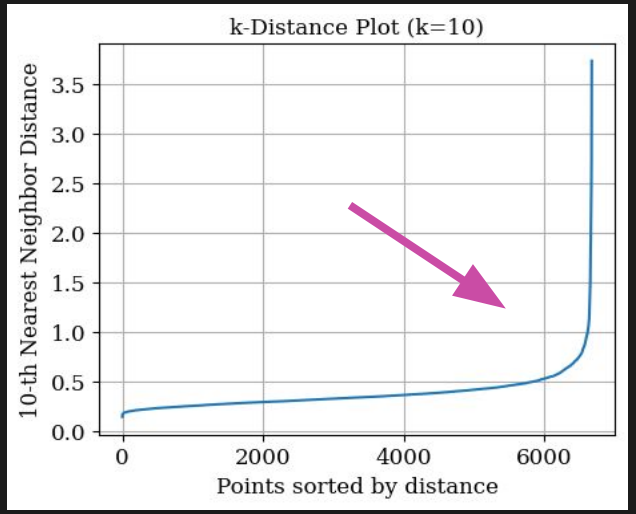


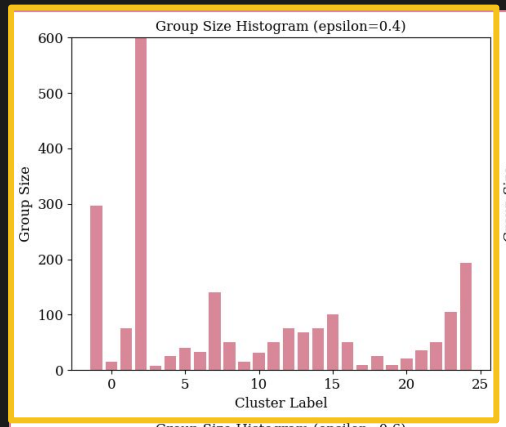
Using HDBSCAN varying epsilon

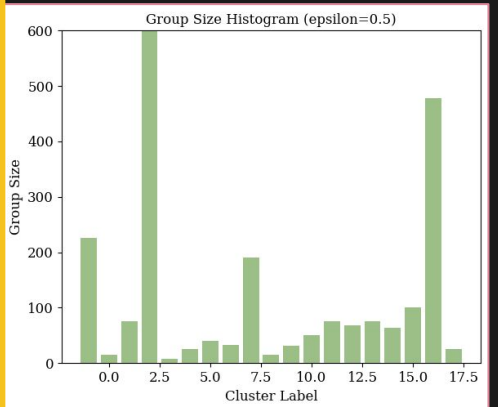


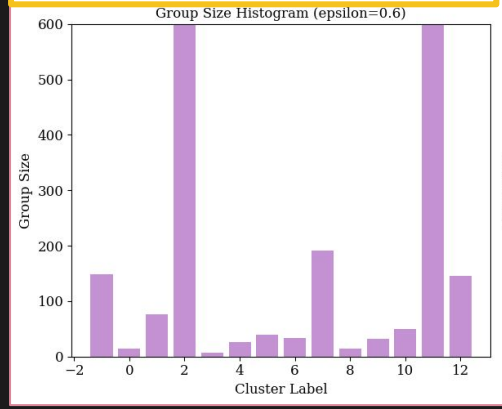
Epsilon - distance around each point that is considered

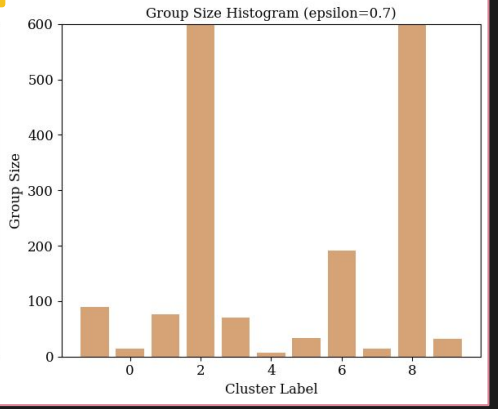
We choose an epsilon based on the k-distance plot.









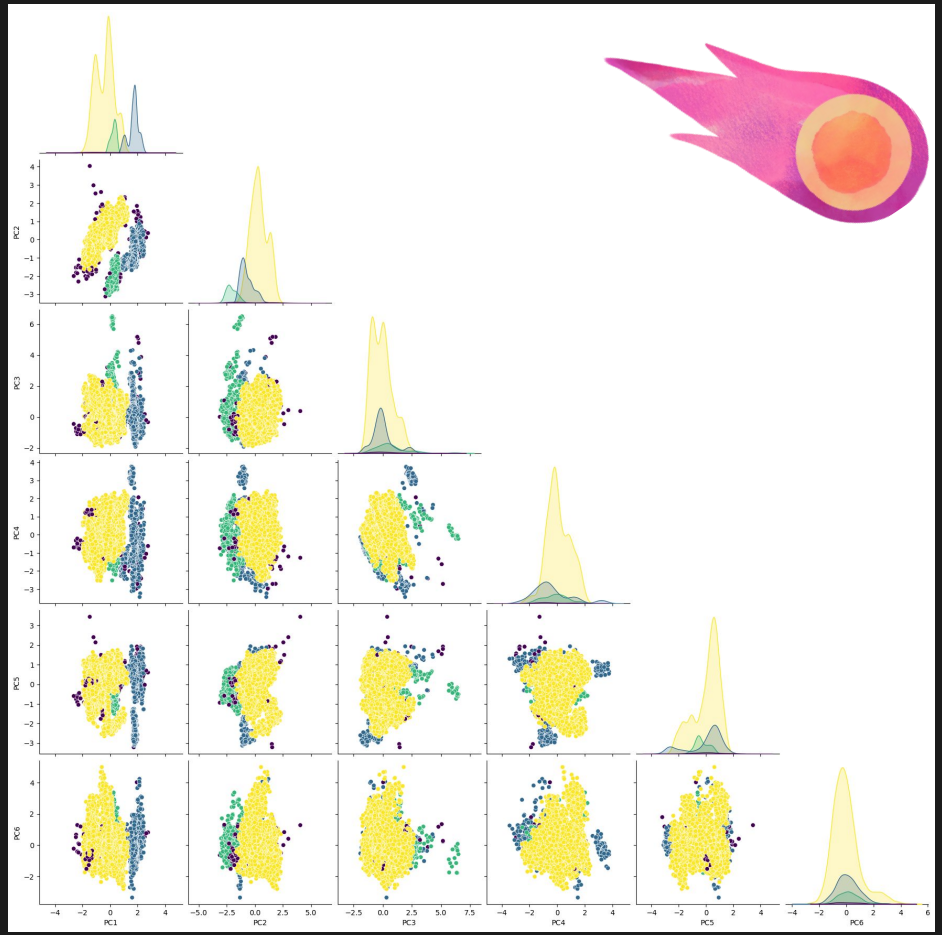


Best HDBSCAN Clustering

We can see all the different groups here

Theoretically each group should demonstrate a different type of observation

- Comet
- Noise
- A base line spectra
- Other variations



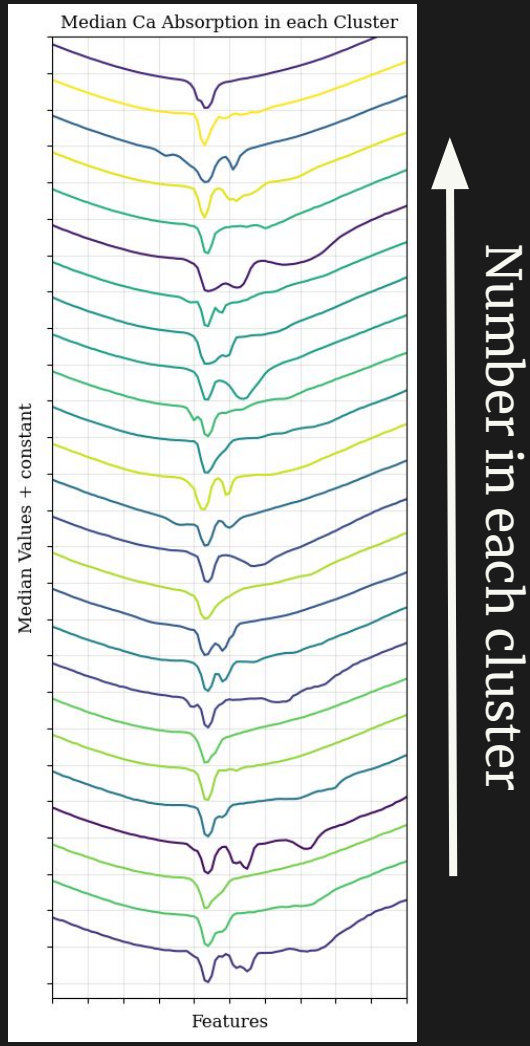
Comparing the Clusters to each other

Examining the individuals clusters we can see:

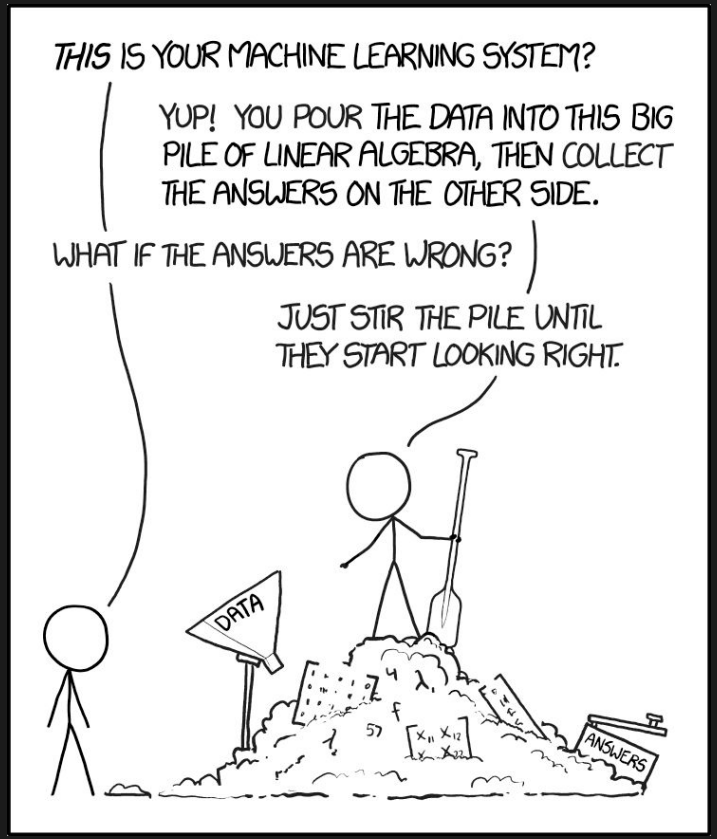
some can be thrown out

others can be combined with each other

This is examined address further in next steps

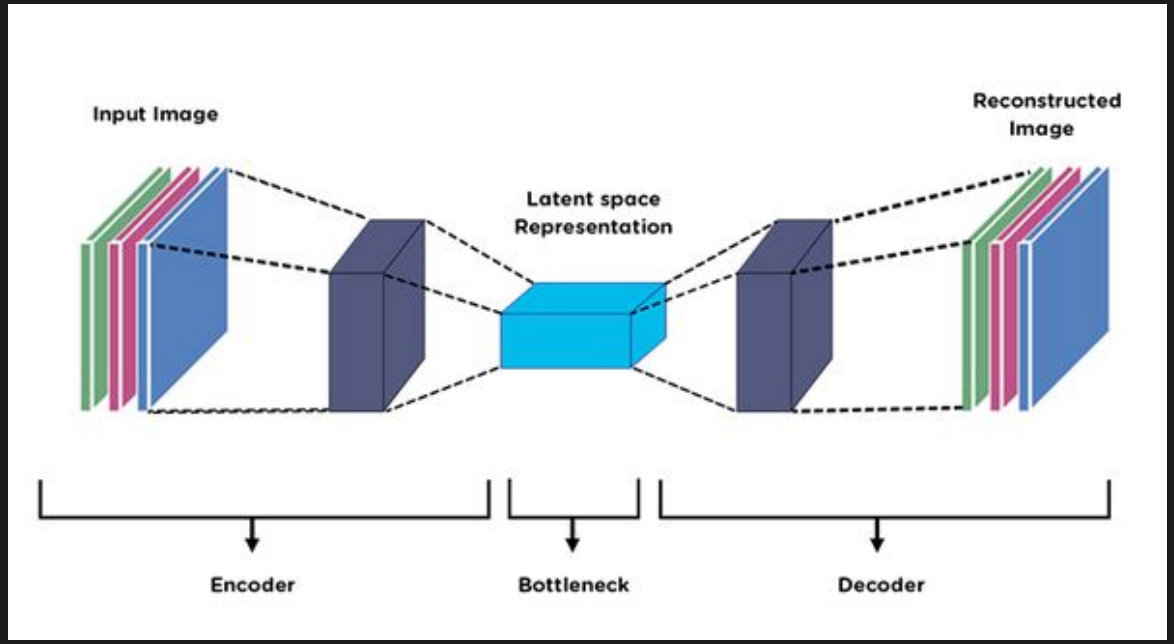


AUTOENCODER METHOD

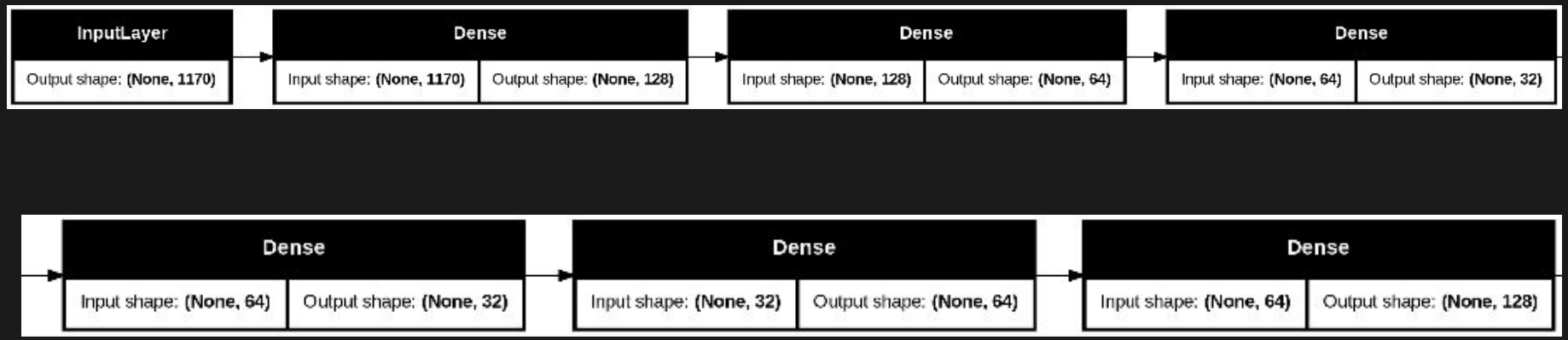


Autoencoder method

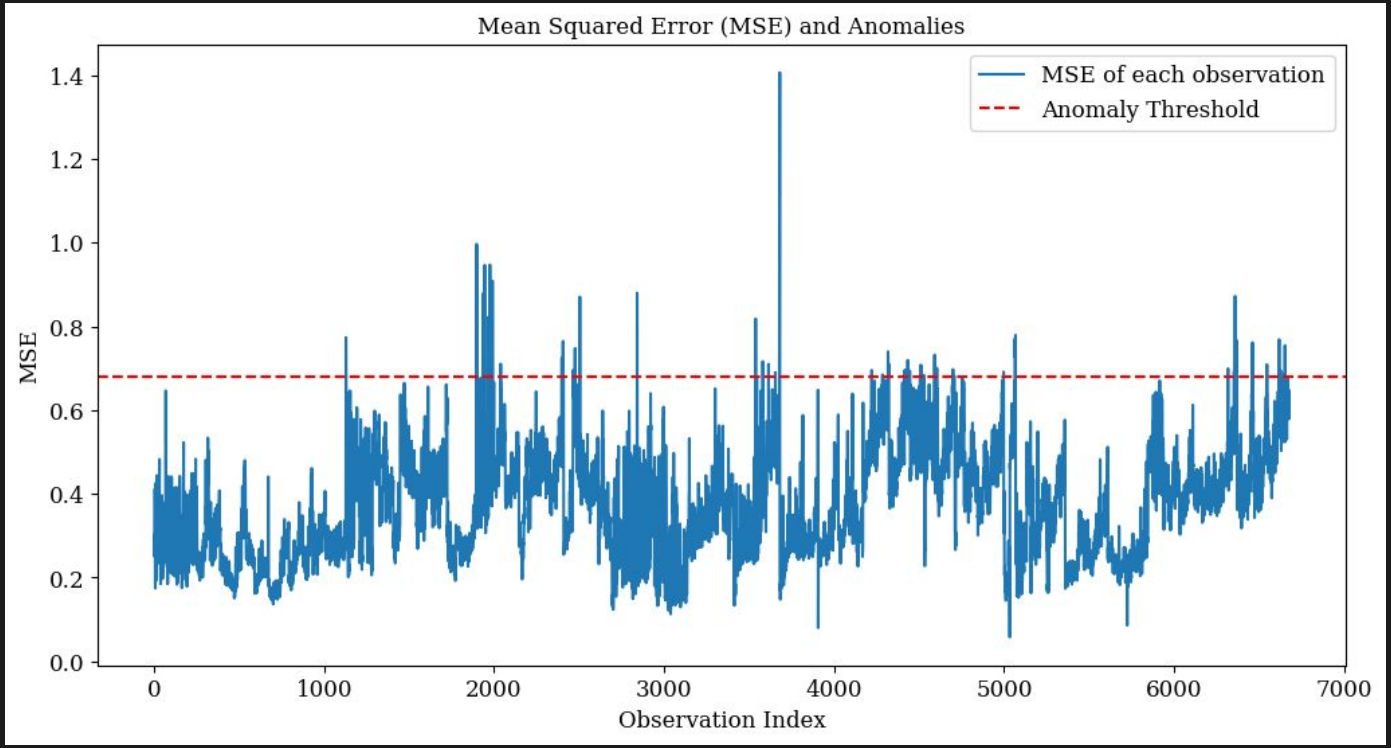
Unsupervised artificial neural networks, such as autoencoders, can be employed to effectively identify anomalous data within a homogeneous dataset



Autoencoder architecture



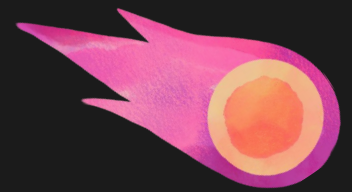
Possible exocomets candidates in Beta Pictoris



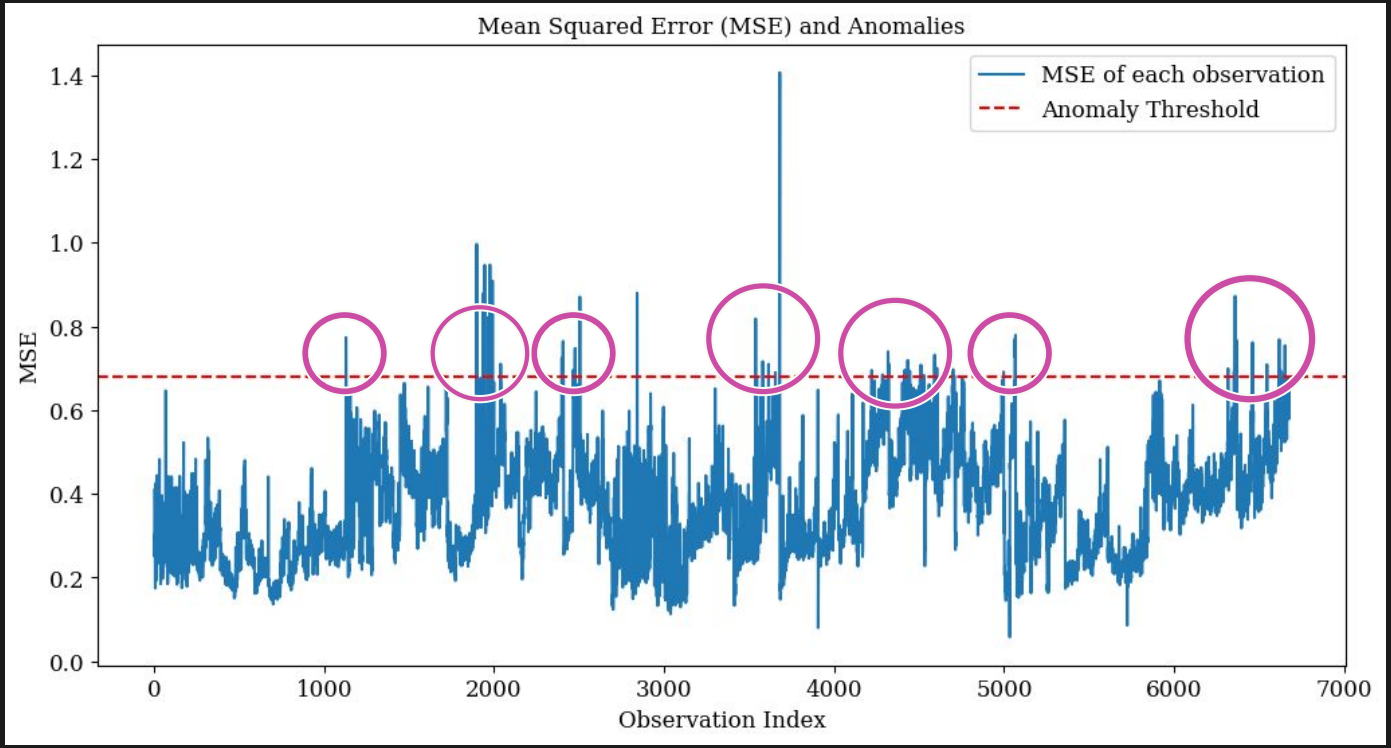
$\bar{x}_{MSE} = 0.371262$

$\bar{x}_{Anomalies} = 0.75544$

$MSE_{3680} = 1.405923$



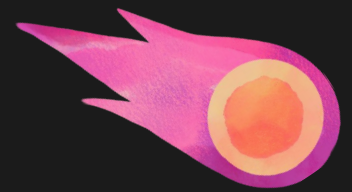
Possible exocomets candidates in Beta Pictoris



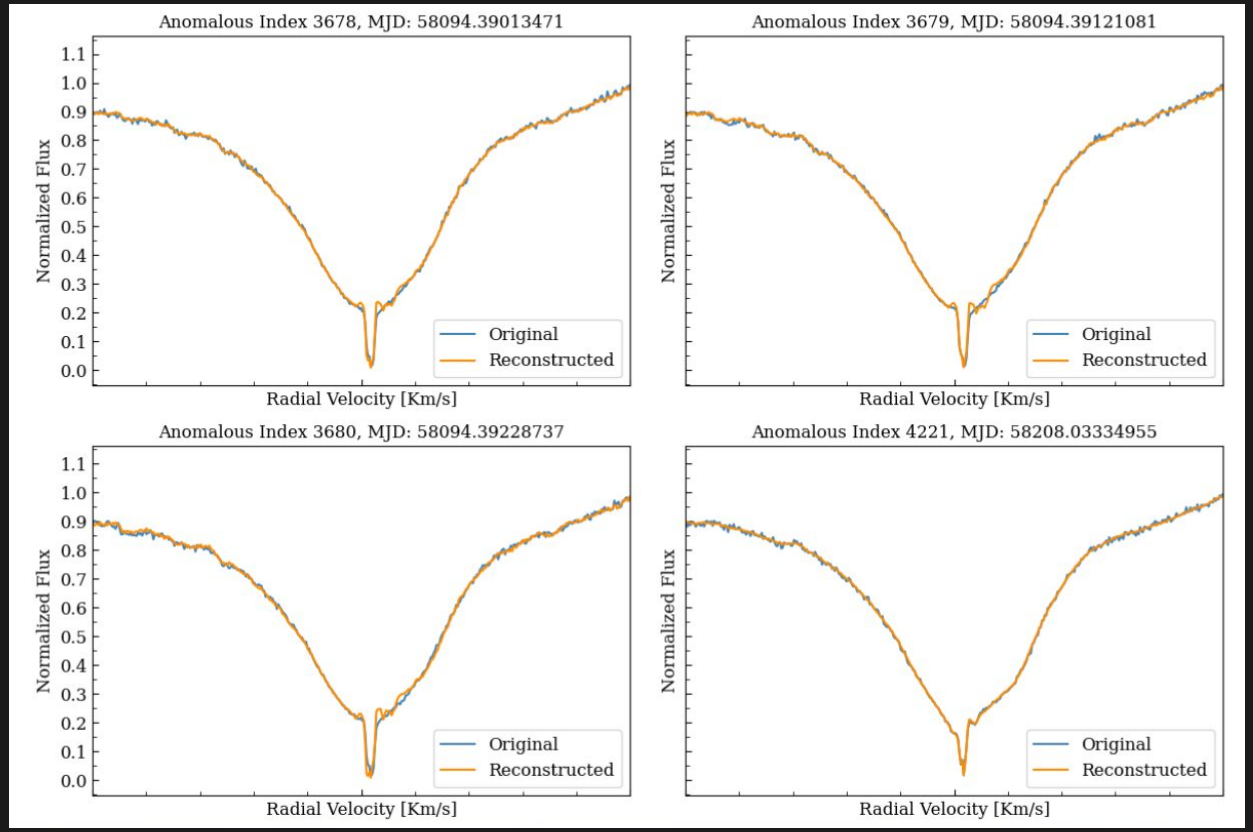
$\bar{x}_{MSE} = 0.371262$

$\bar{x}_{Anomalies} = 0.75544$

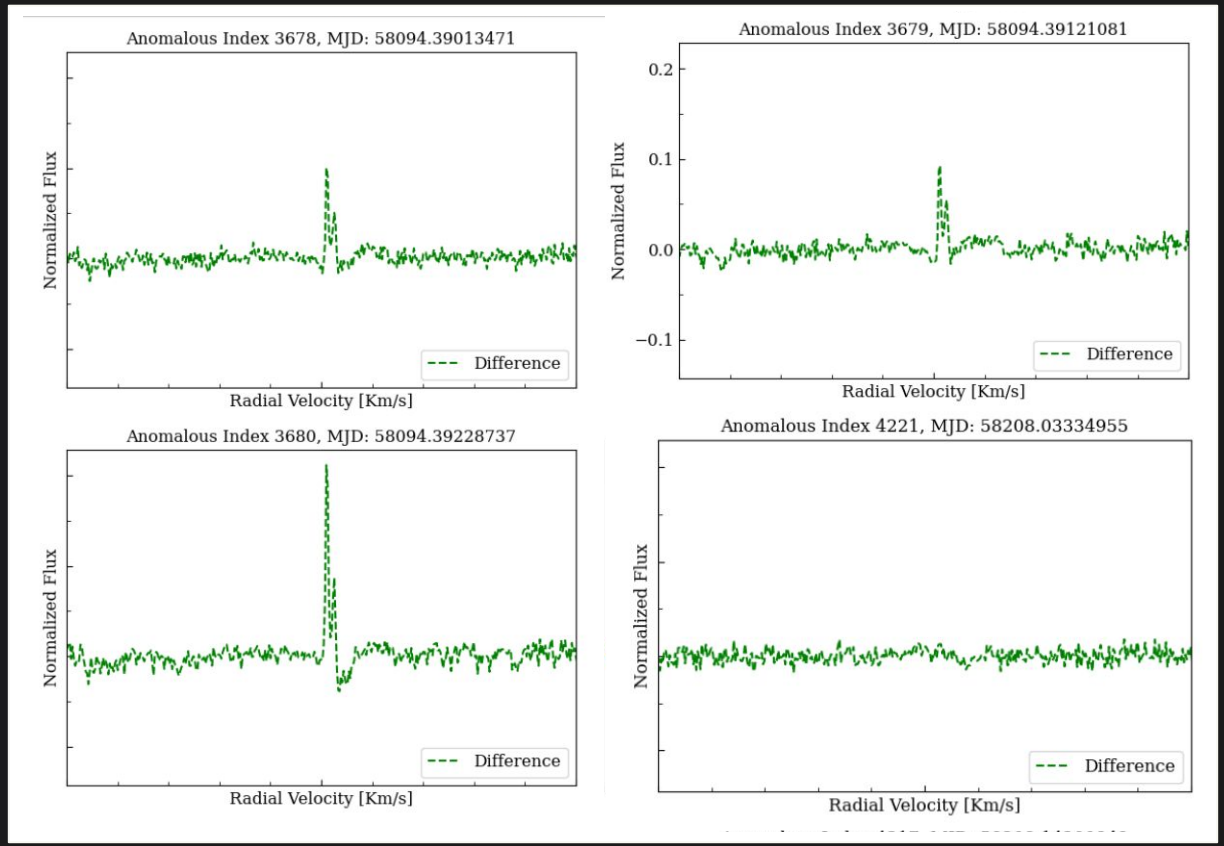
$MSE_{3680} = 1.405923$



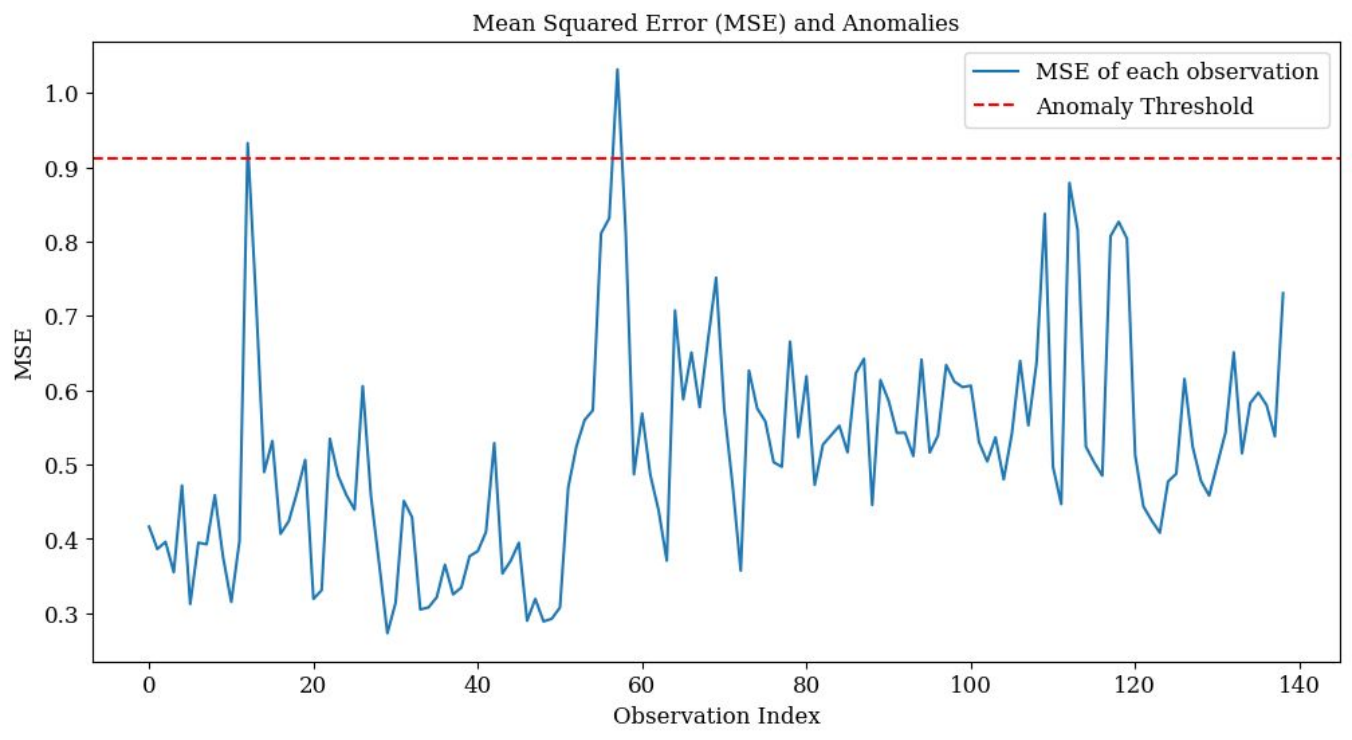
Possible exocomets candidates in Beta Pictoris



Possible exocomets candidates in Beta Pictoris



Possible exocomets candidates in HD 172555

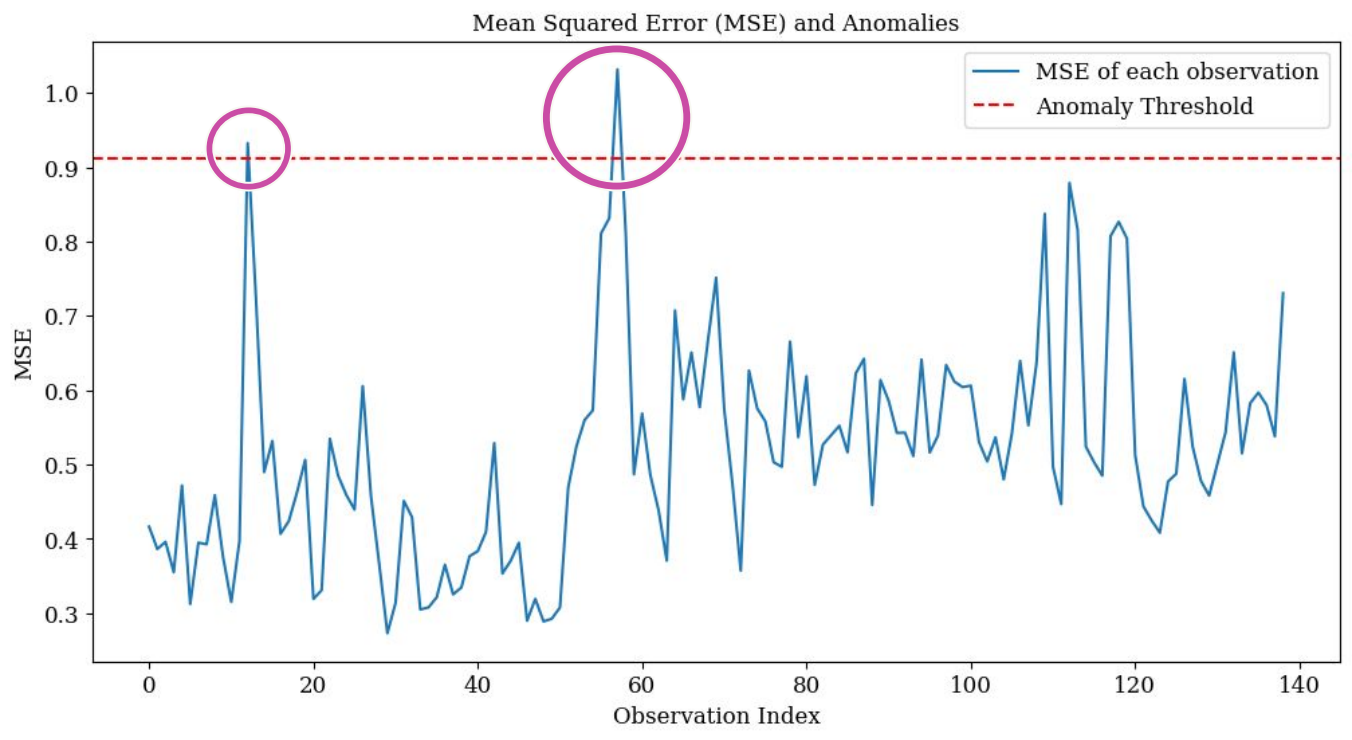


$\bar{x}_{MSE} = 0.517751$

$MSE_{12} = 0.932405$

$MSE_{57} = 1.031454$

Possible exocomets candidates in HD 172555

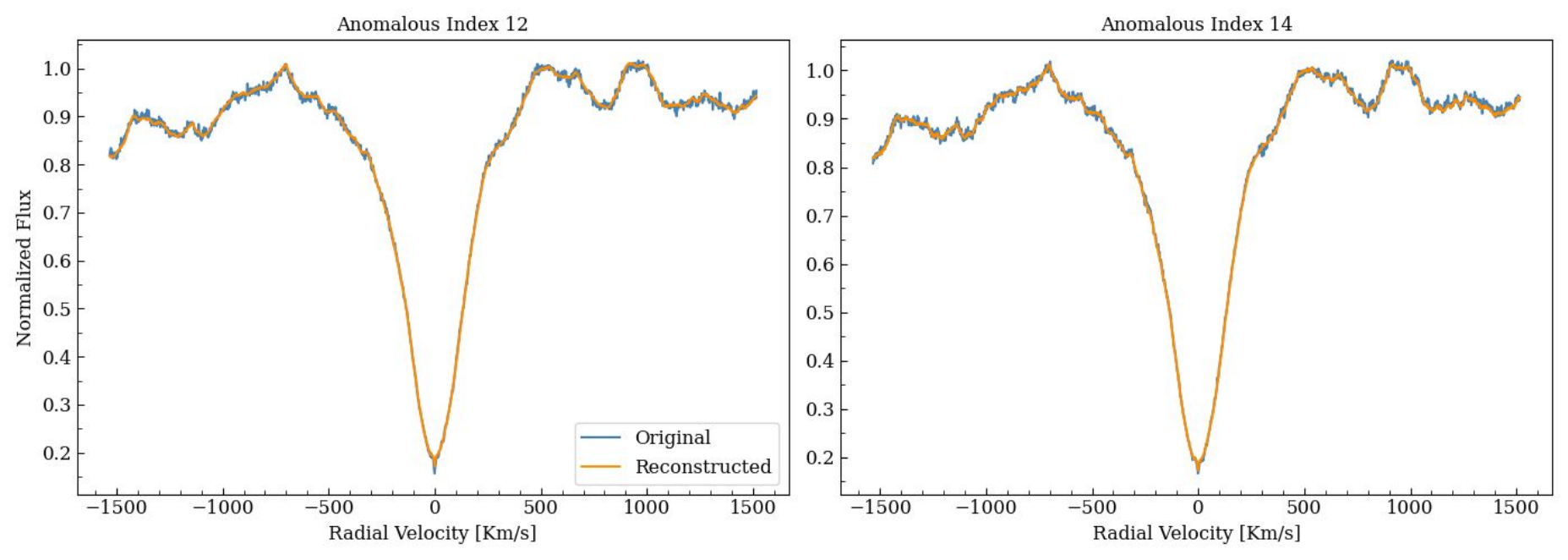


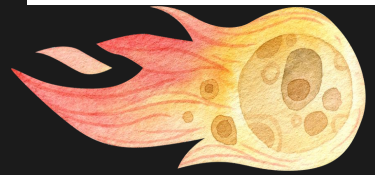
$\bar{x}_{MSE} = 0.517751$

$MSE_{12} = 0.932405$

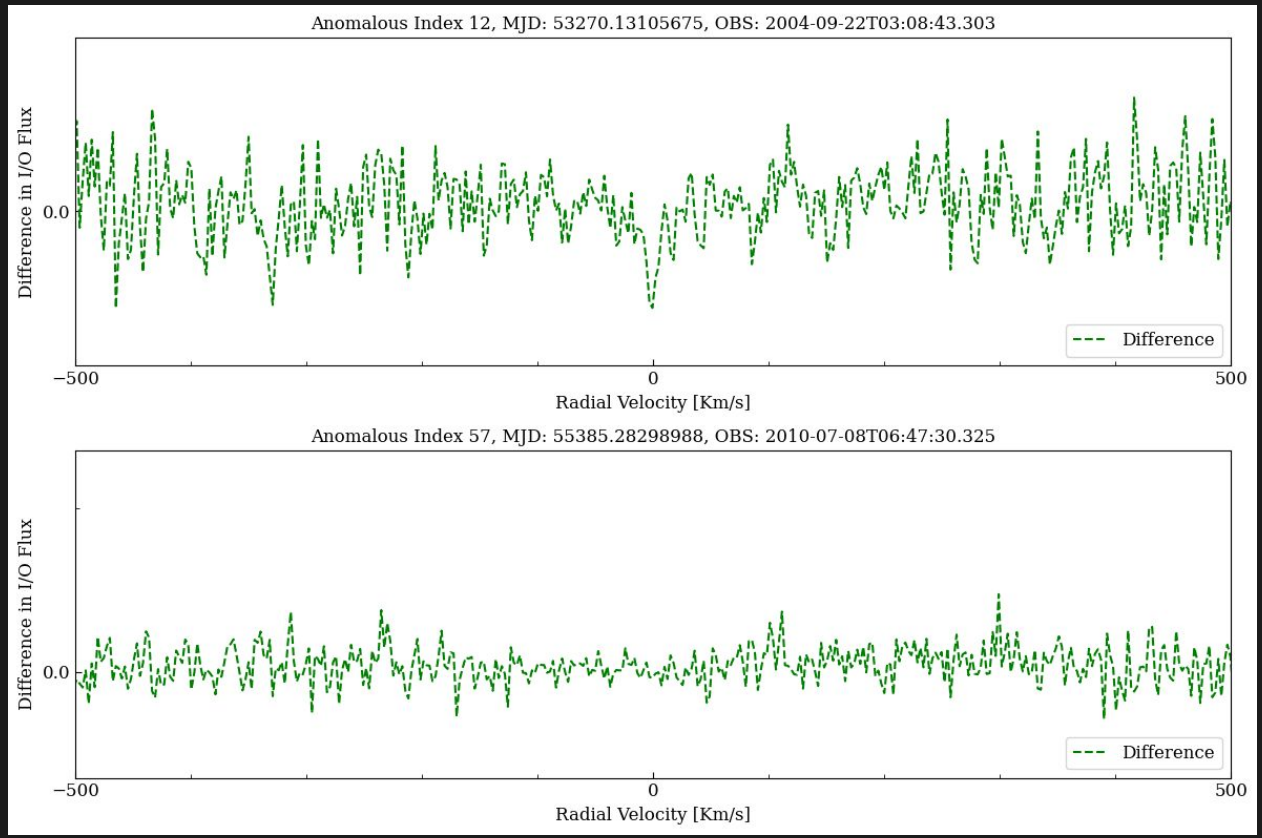
$MSE_{57} = 1.031454$

Possible exocomets candidates in HD 172555



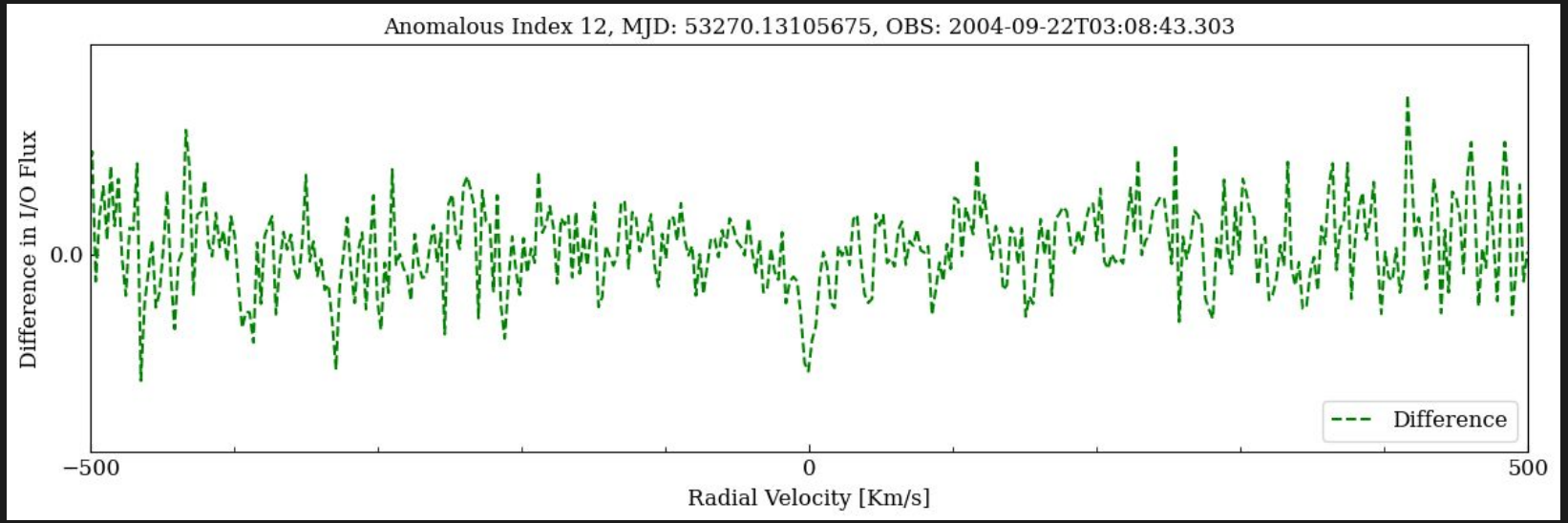


Possible exocomets candidates in HD 172555



Possible exocomets candidates in HD 172555

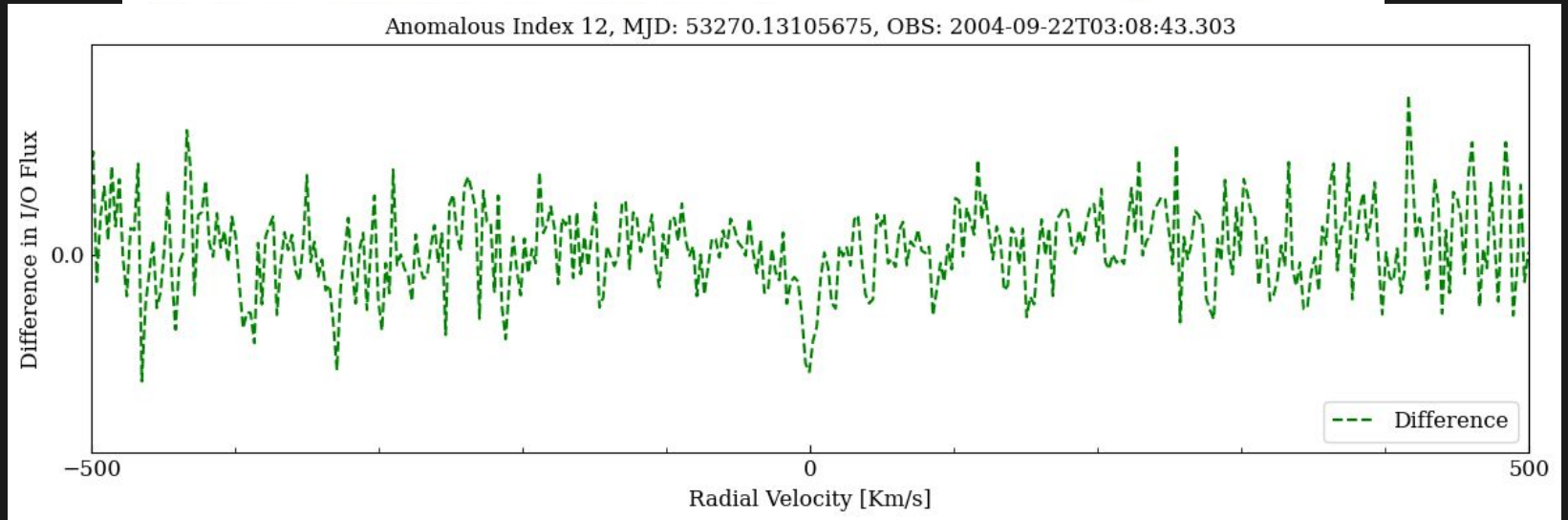
Using the autoencoder with HD 172555 and comparing with **Kiefer et al. 2014:**



Possible exocomets candidates in HD 172555

Table 3. Fit parameters with 1-sigma error bars.

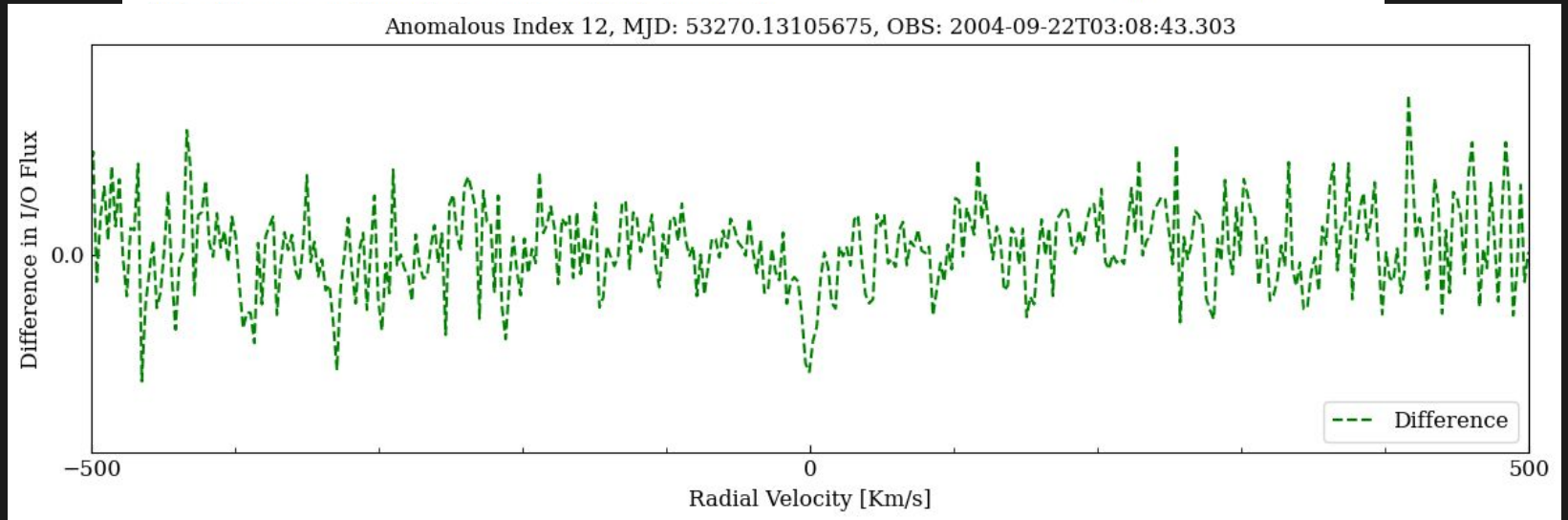
Date (MJD)	Date (D/M/Y)	K line depth	Velocity (km s^{-1})	*FWHM* (km s^{-1})	Surface ratio α	Optical depth
53 269.996	22/09/04	0.059 ± 0.003	0.35 ± 0.37	13.9 ± 0.9	≳0.9	0.061 ± 0.003
53 270.134	22/09/04	0.072 ± 0.006	2.3 ± 0.5	19.5 ± 1.2	≳0.84	0.075 ± 0.006
53 603.145	21/08/05	0.034 ± 0.004	13.5 ± 1.5	38.4 ± 9.7	$0.04^{+0.04}_{-0.01}$	1.69 ± 1.05
55 385.285	08/07/10	0.029 ± 0.006	1.26 ± 1.07	$18.4^{+7.9}_{-10.1}$	≳0.024	≲10.3
55 723.248	11/06/11	0.037 ± 0.002	−1.6 ± 0.5	22.5 ± 3.2	0.04 ± 0.01	1.90 ± 0.63
55 723.280	11/06/11	0.032 ± 0.002	−2.44 ± 0.51	24.9 ± 3.3	0.04 ± 0.01	1.48 ± 0.51
55 723.309	11/06/11	0.030 ± 0.003	−3.24 ± 0.73	24.2 ± 4.7	$0.04^{+0.02}_{-0.01}$	1.59 ± 0.78



Possible exocomets candidates in HD 172555

Table 3. Fit parameters with 1-sigma error bars.

Date (MJD)	Date (D/M/Y)	K line depth	Velocity (km s^{-1})	*FWHM* (km s^{-1})	Surface ratio α	Optical depth
53 269.996	22/09/04	0.059 ± 0.003	0.35 ± 0.37	13.9 ± 0.9	≳0.9	0.061 ± 0.003
53 270.134	22/09/04	0.072 ± 0.006	2.3 ± 0.5	19.5 ± 1.2	≳0.84	0.075 ± 0.006
53 603.145	21/08/05	0.034 ± 0.004	13.5 ± 1.5	38.4 ± 9.7	$0.04^{+0.04}_{-0.01}$	1.69 ± 1.05
55 385.285	08/07/10	0.029 ± 0.006	1.26 ± 1.07	$18.4^{+7.9}_{-10.1}$	≳0.024	≲10.3
55 723.248	11/06/11	0.037 ± 0.002	−1.6 ± 0.5	22.5 ± 3.2	0.04 ± 0.01	1.90 ± 0.63
55 723.280	11/06/11	0.032 ± 0.002	−2.44 ± 0.51	24.9 ± 3.3	0.04 ± 0.01	1.48 ± 0.51
55 723.309	11/06/11	0.030 ± 0.003	−3.24 ± 0.73	24.2 ± 4.7	$0.04^{+0.02}_{-0.01}$	1.59 ± 0.78

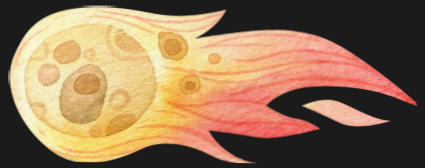


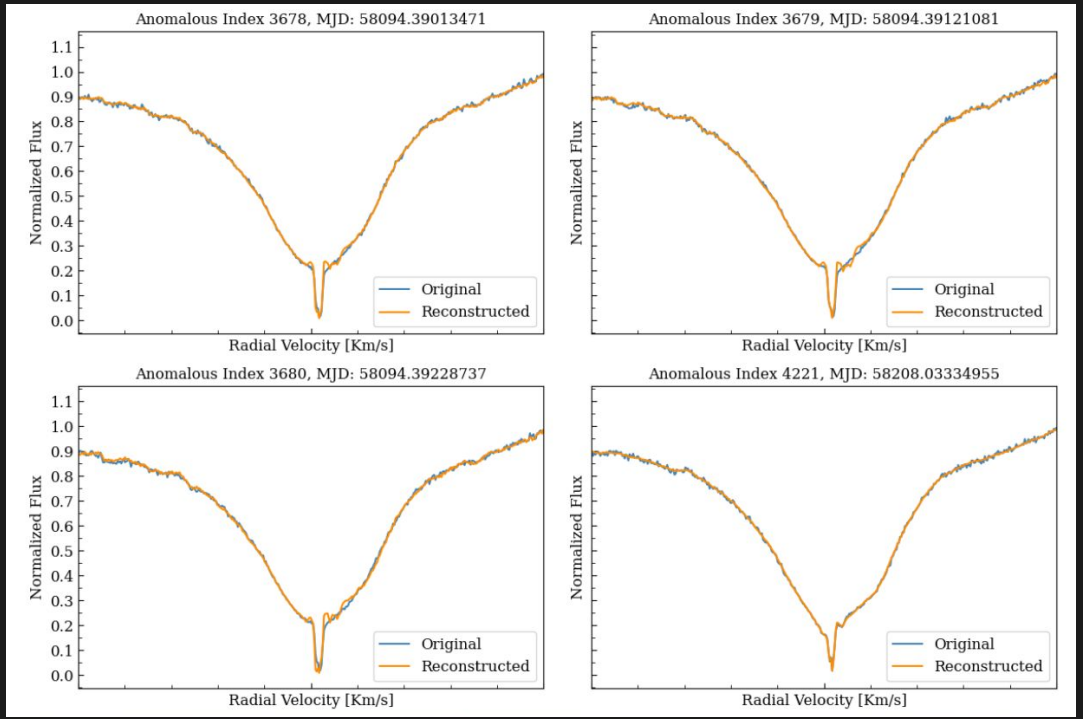
Comparing Methods

Comparing Results in Beta Pictoris

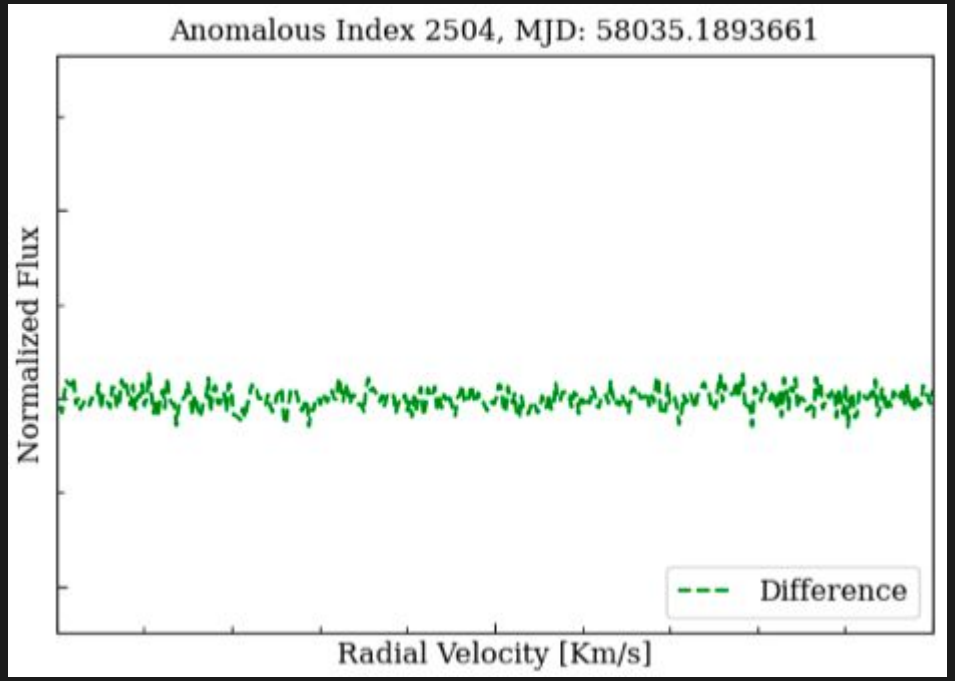
The autoencoder found multiple anomalies, some of which may be exo comets

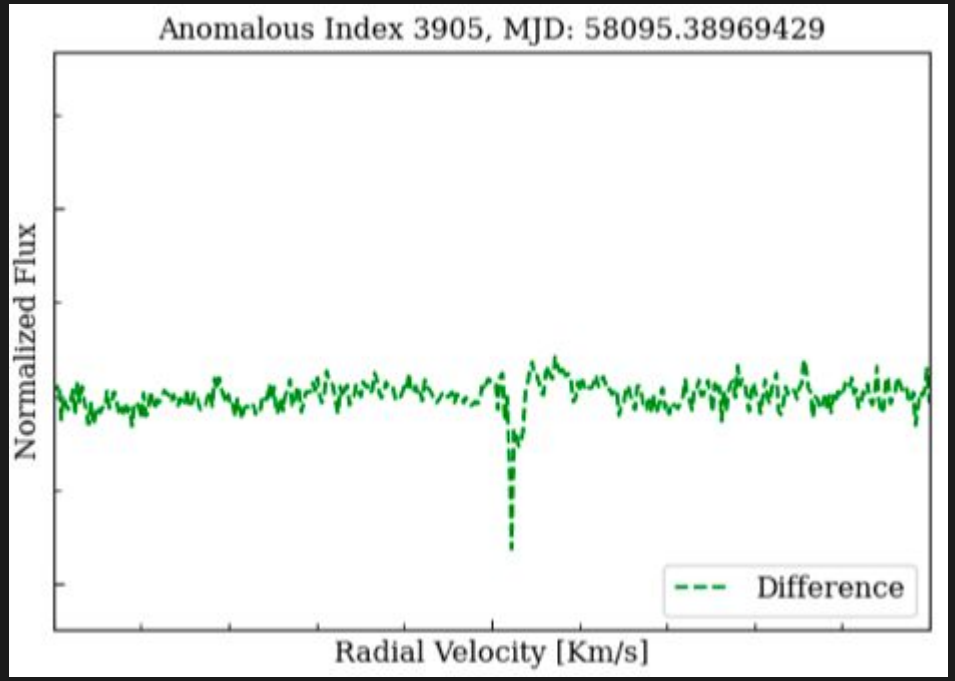
We can (and should) compare with the PCA clustering to check if both methods agree





Comparing Results in Beta Pictoris

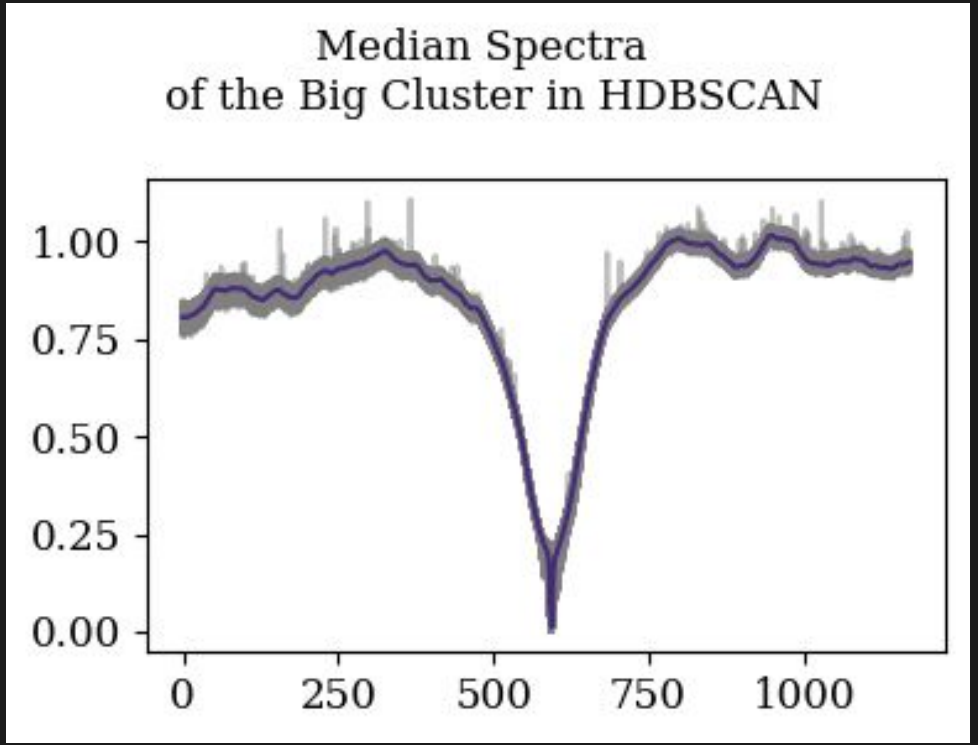




Comparing Anomalies in Beta Pictoris

Almost all anomalies found in the autoencoder were located in the same cluster (Cluster 2)

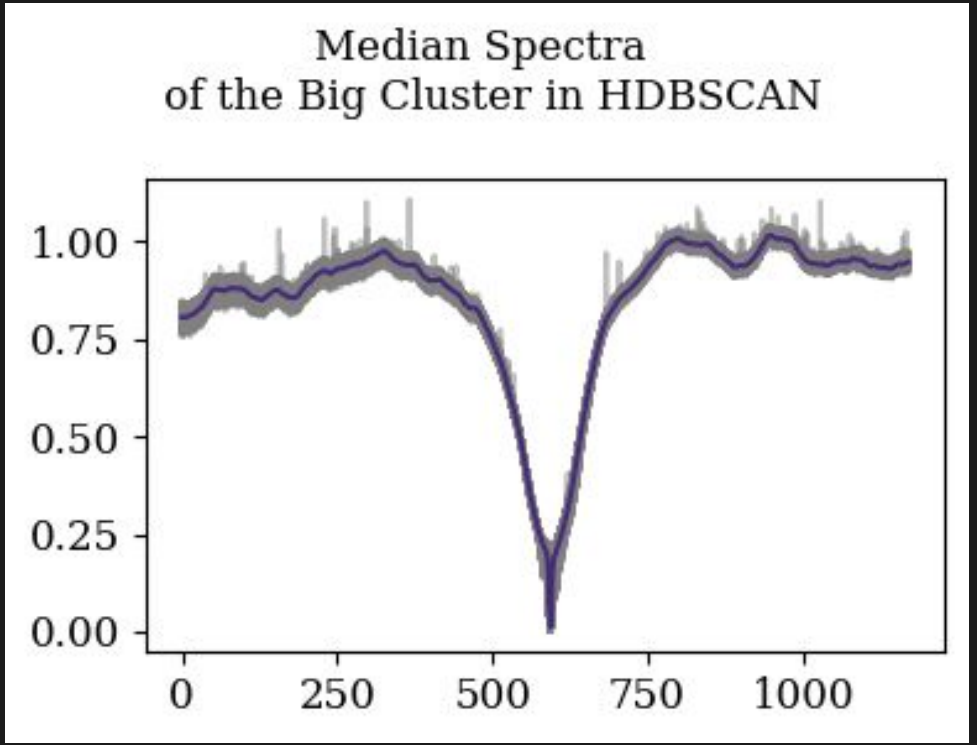
This cluster is by far the largest, and can be interpreted as a baseline for the star

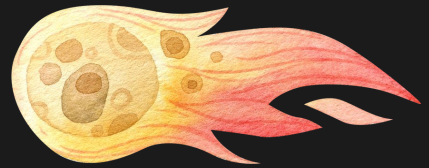


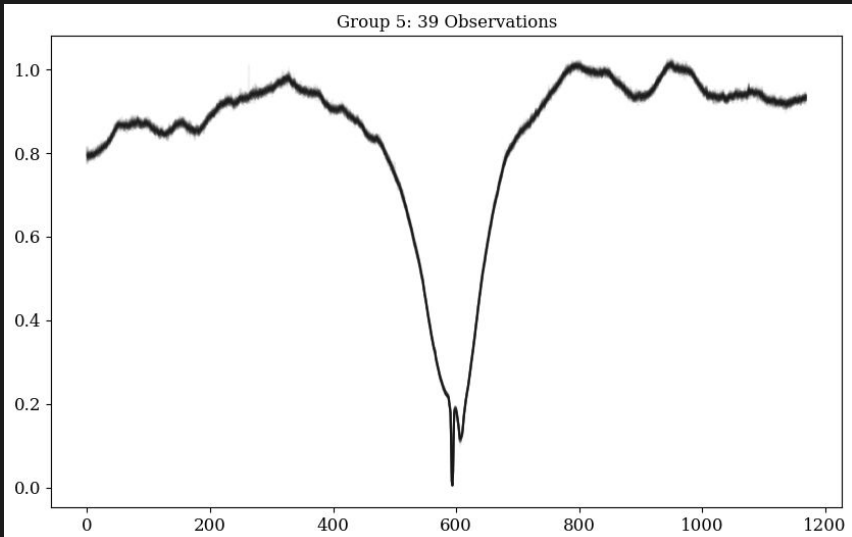
Why is this happening?

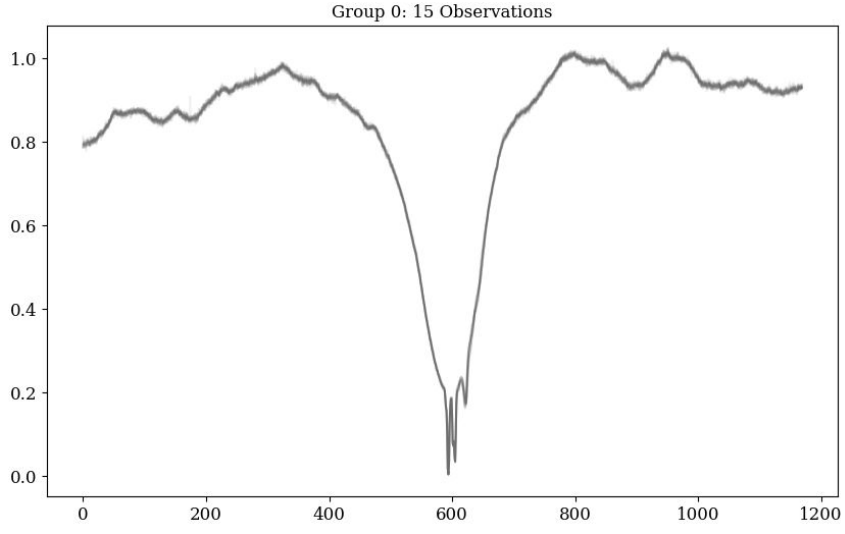
The cluster is misinterpreting an exocomet anomaly as noise!

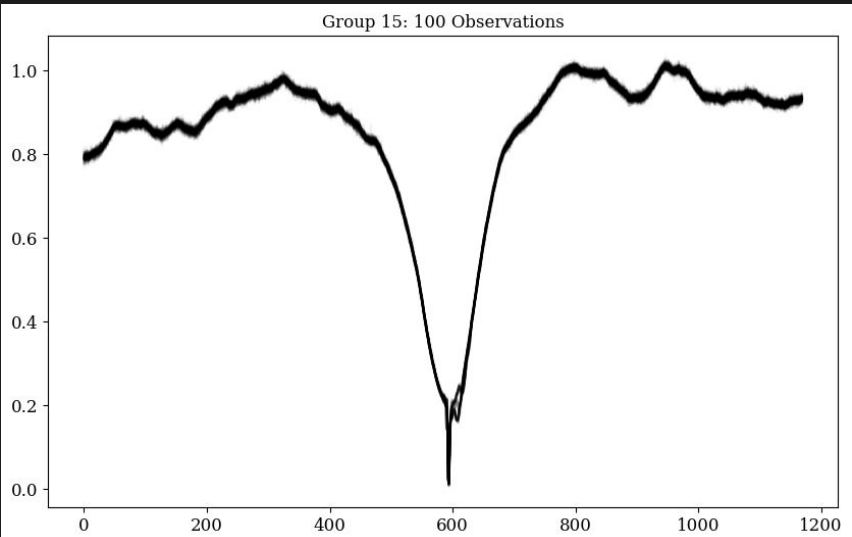
To fix this, we can instead cluster with our 5 PCA components AND the SNR

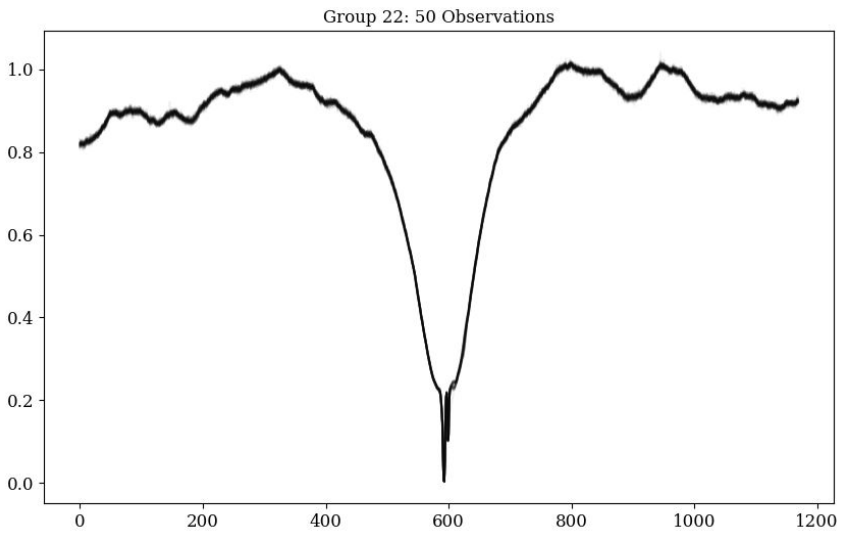





Group 5: 39 Observations
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200

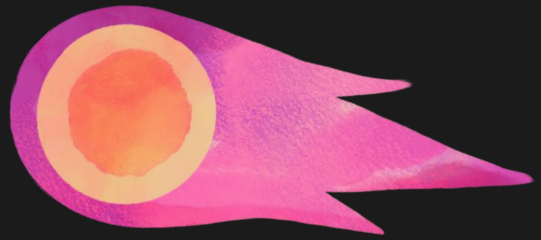

Group 0: 15 Observations
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200


Group 15: 100 Observations
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200


Group 22: 50 Observations
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200

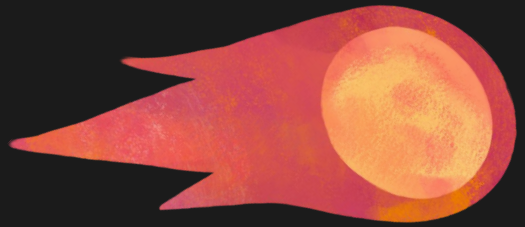
Next steps

- Compare PCA and autoencoder methods by checking reconstruction error
- Apply PCA and autoencoder to other stars
- Examine the clusters and understand the distinctions of each



Thank You

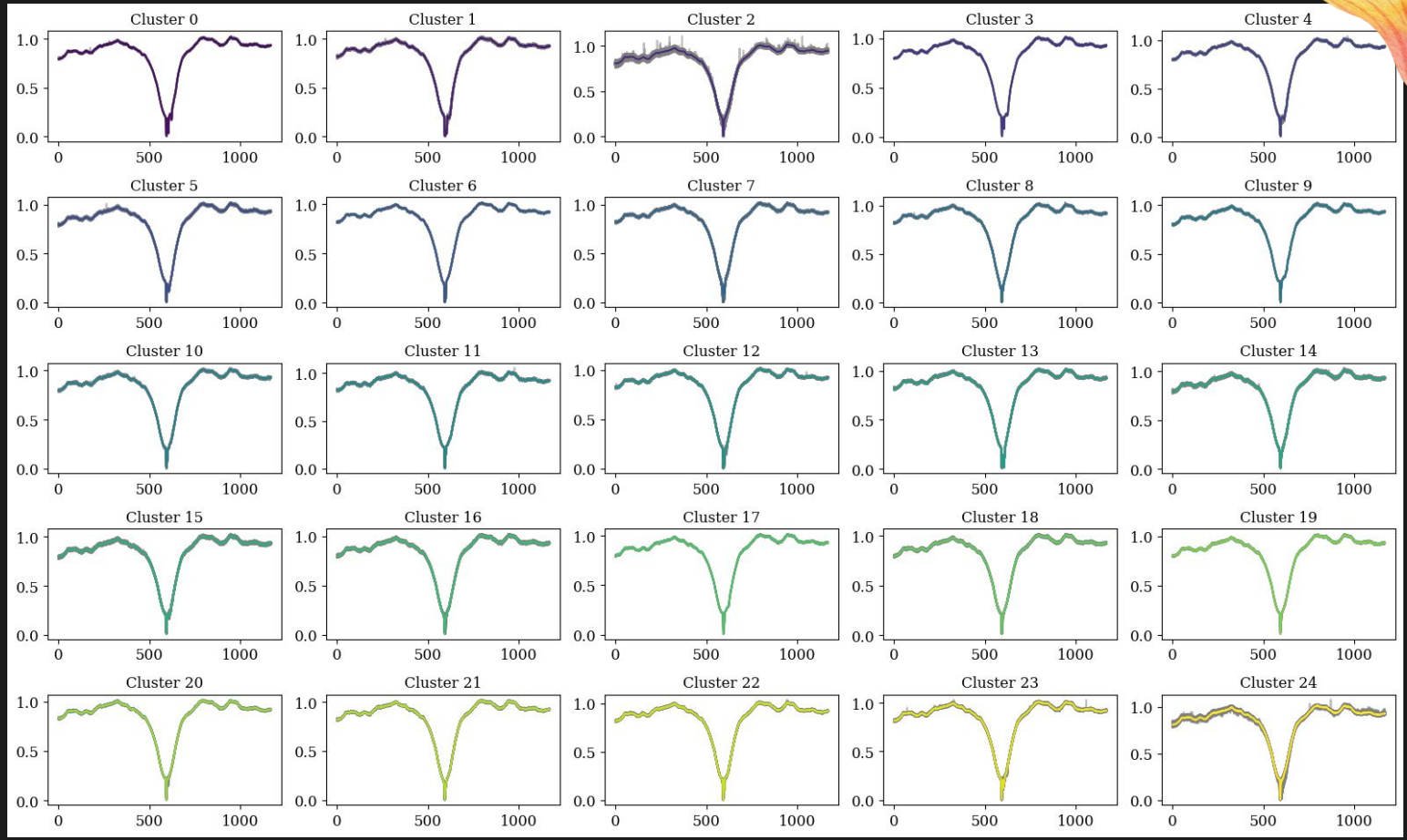
Gracias



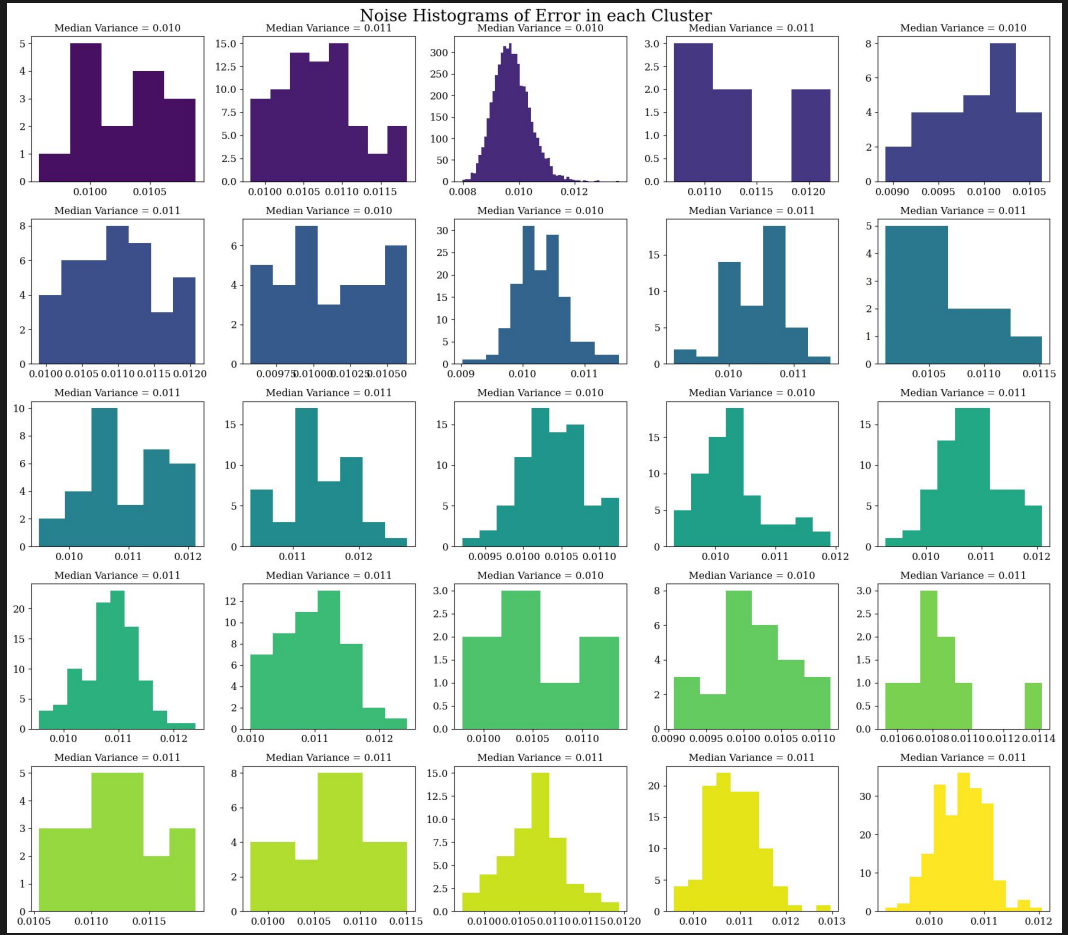
HARPS

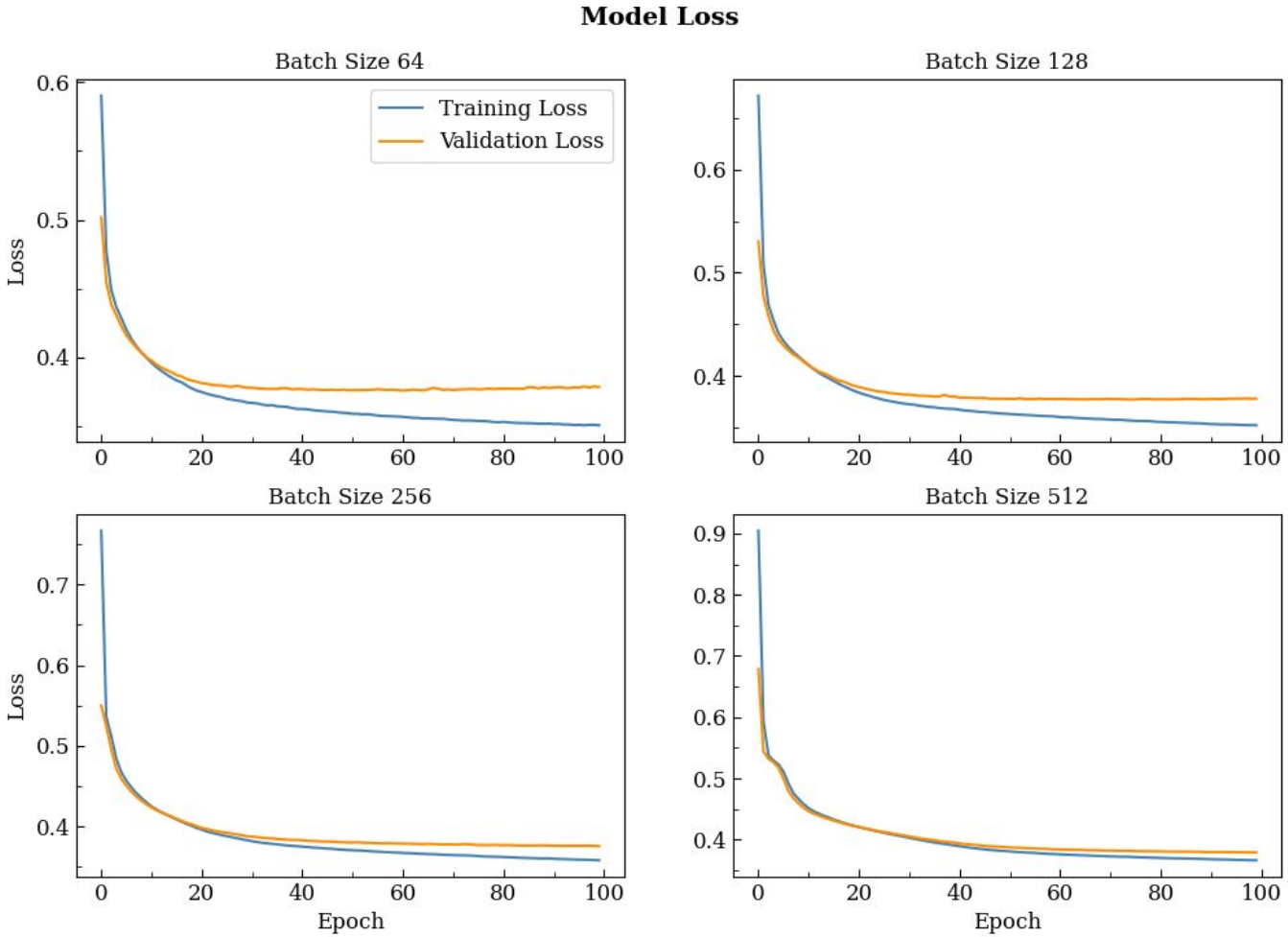
- Instrument installed on the 3.6 meter telescope at La Silla, Chile
- High resolution: ~115,000
- Wavelengths range: from 3,800 to 6,900 A
- Stability
- Comprises 289,843 observations of 6,488 unique astronomical objects
- 9,098 observations for Beta Pictoris

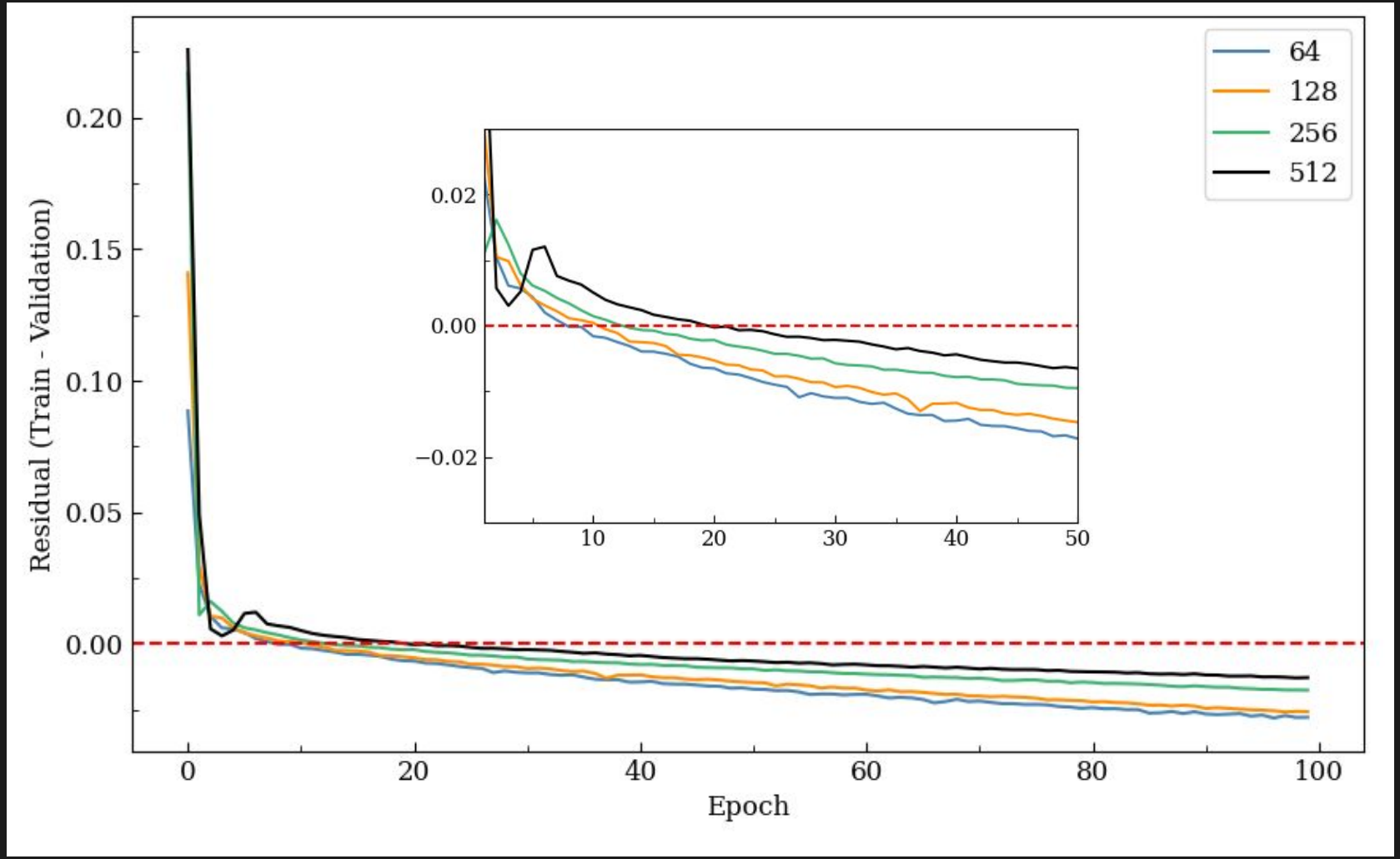
Variation in PCA Clusters

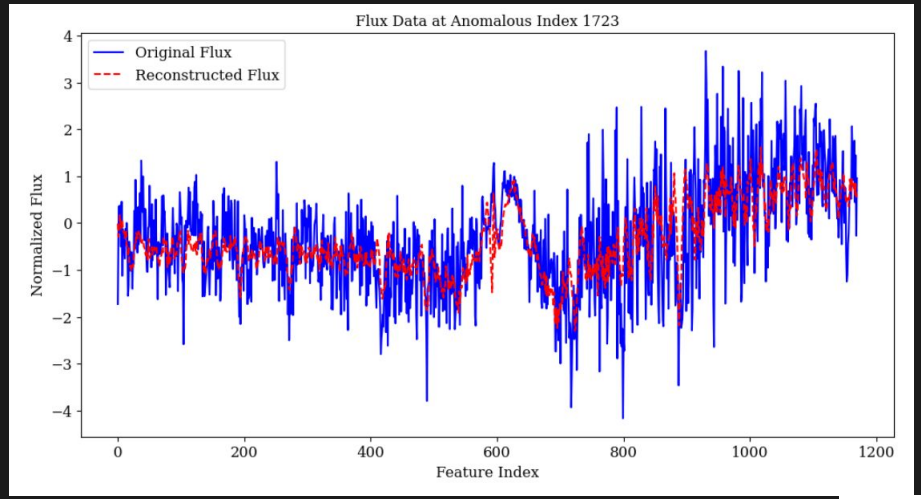


PCA group Variation



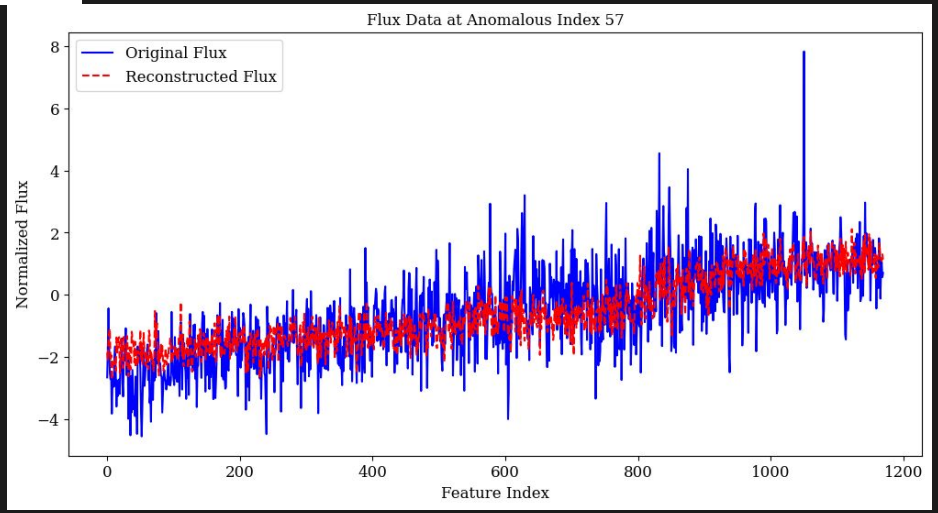

Model Loss
Batch Size 64
Batch Size 128
Batch Size 256
Batch Size 512
Training Loss
Validation Loss
Loss
Epoch
0.4
0.5
0.6
0.7
0.8
0.9
0
20
40
60
80
100

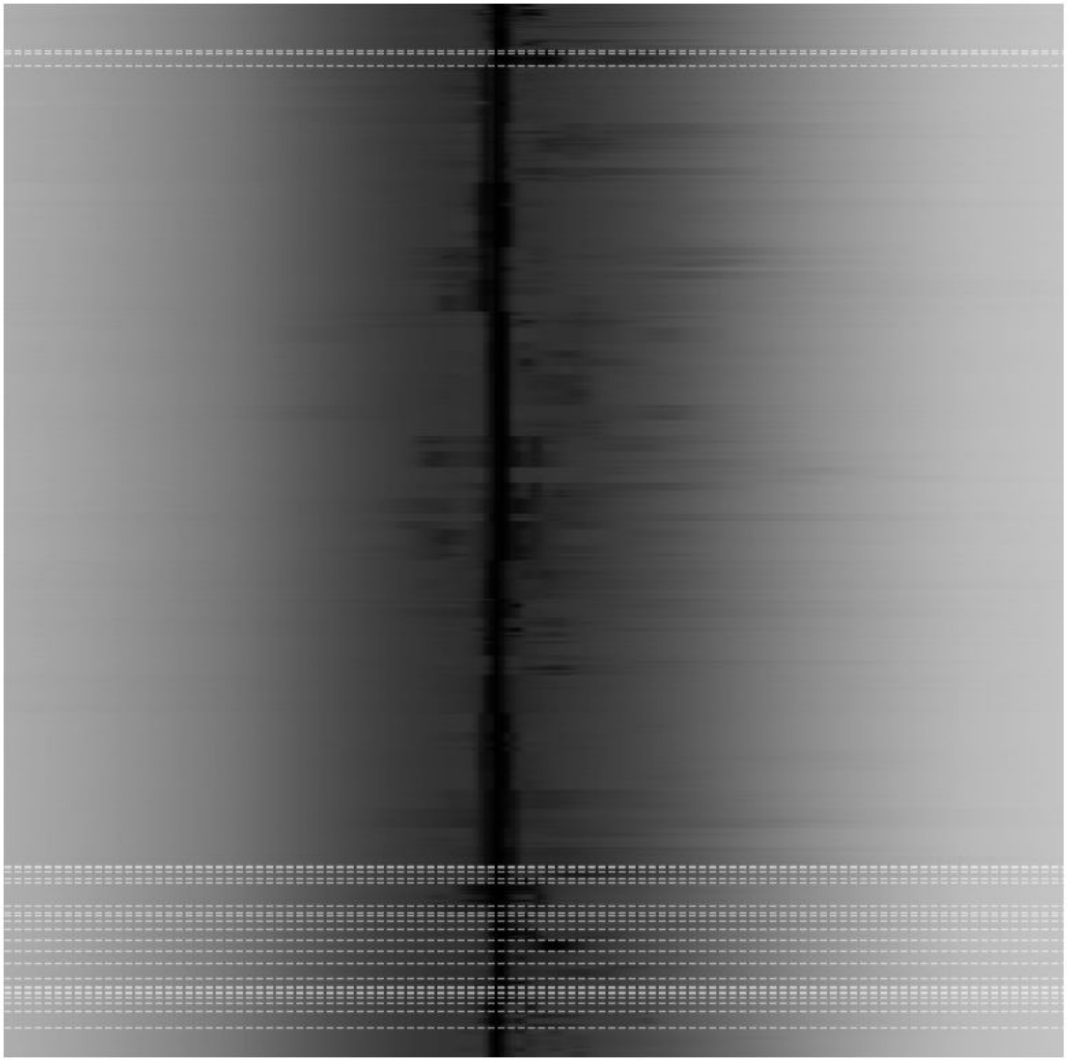

64
128
256
512
Residual (Train - Validation)
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100
Epoch
0.02
0.00
−0.02
10
20
30
40
50



- Beta Pictori

- HD 172555





Autoencoder architecture

•[18]:

```
# Encoder
encoded = Dense(128, activation='relu')(input_layer)
encoded = Dense(64, activation='relu')(encoded)
encoded = Dense(32, activation='relu')(encoded)

#Dense: A fully connected neural network layer. The numbers (128, 64, 32) represent the number of neurons in each layer.
#activation='relu': Uses the ReLU (Rectified Linear Unit) activation function, which introduces non-linearity.

# Decoder
decoded = Dense(64, activation='relu')(encoded)
decoded = Dense(128, activation='relu')(decoded)
decoded = Dense(input_dim, activation='linear')(decoded) # Here the activation function was sigmoid, now is linear

#The decoder mirrors the encoder but in reverse.
#activation='sigmoid': The final layer uses the sigmoid function to ensure the output values are between 0 and 1,
#matching the normalized input data. We change the normalization from MinMax to Standard scale so we use Linear function to compensate
#the shape

# Autoencoder Model
autoencoder = Model(inputs=input_layer, outputs=decoded)
autoencoder.compile(optimizer='adam', loss='mse') #Mean Squared Error

# 5. Train the Autoencoder
history = autoencoder.fit(X_train, X_train,
                epochs=50,
                batch_size=256,
                shuffle=True,
                validation_data = [X_test, X_test],
                validation_split=0.2)

#fit: Trains the model.

#    flux_normalized: Both the input and target are the same since it's an autoencoder.
#    epochs=50: The model will iterate over the entire dataset 50 times.
#    batch_size=256: Number of samples per gradient update.
#    shuffle=True: Shuffles the data before each epoch to reduce overfitting.
#    validation_split=0.2: Uses 20% of the data for validation to monitor the model's performance on unseen data
```