Star- and planet formation caught-in-the-act

*G. Marton[1,2], O. Dionatos[3], M. Audard[4], I. Gezer[1,2], D. Hernandez[3], T. Pavlidou[3], J. Roquette[4]

NEMESIS

NEMESIS (**N**ovel **E**volutionary **M**odel for the **E**arly stages of **S**tars with **I**ntelligent **S**ystems) has the ambition to **reshape our understanding** on the formation of stars by employing **artificial intelligence methods** to interpret the **largest, panchromatic data collection of Young Stellar Objects**. Recent evidence suggests that planets form synchronously rather than sequentially to their host stars, indicating a rapid early evolution of star-planet systems. To ascertain these timescales, it is necessary to first determine the characteristic transitions that describe each phase of star formation. The definition of classes for young stellar objects was made possible more than 30 years ago, due to the first space-based infrared sky surveys. Whilst successful in determining global properties, current classification is prone to large uncertainties, and therefore, timescales, which are based on population statistics among different classes in a steady-state evolution, remain dubious.

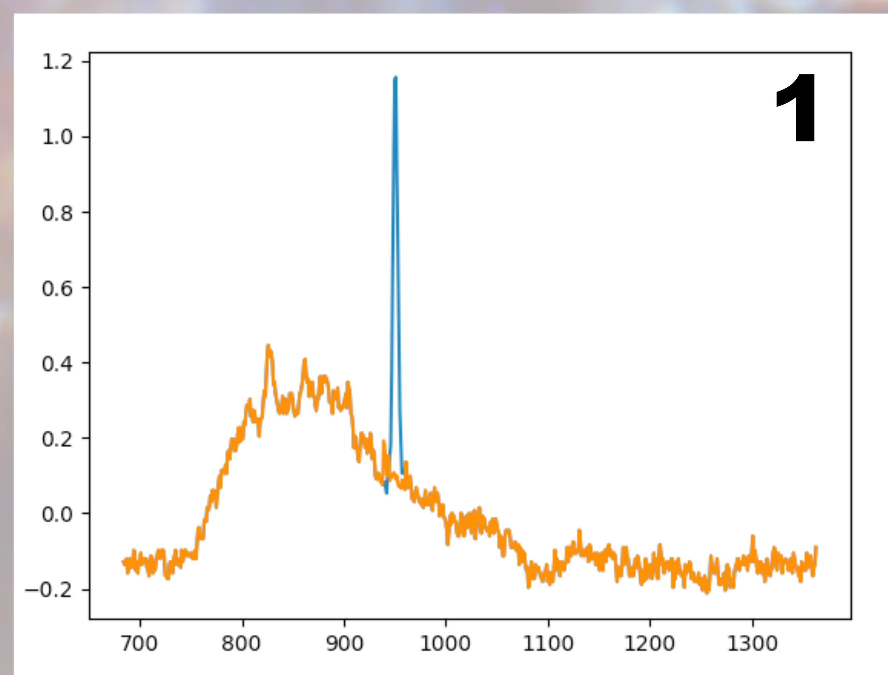
Today, we have in hand significantly higher quality and much larger volumes of data, more advanced computing methods and higher computing power, which allow us to revisit the standard star-formation paradigm and in particular the evolutionary scheme of young stars, using new machine learning techniques applicable to big data.

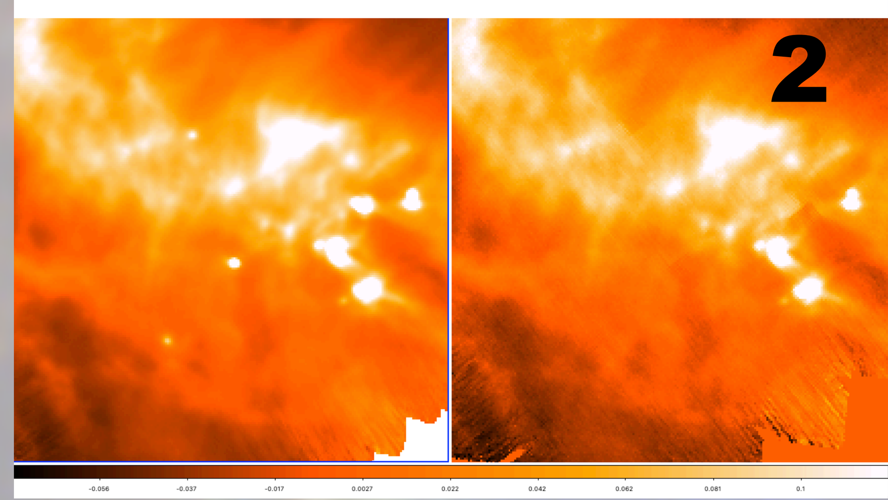
NEMESIS aims to **readjust the current classification scheme** and its characteristic timescales so that it is concurrent with the most recent observational and theoretical constraints. It will reprocess and analyse a unique dataset with **supervised and unsupervised machine learning** algorithms, **deep learning** neural networks for object detection, **clustering and regression analysis** of images in order to advance the analysis and interpretation beyond the current state-of-the-art. Ultimately, NEMESIS brings big data techniques and hybrid machine learning methods to systematically analyse and interpret large data volumes in order to answer some of the most persisting questions, paving the path toward data intensive science applications in modern astrophysics.

NEMESIS is a EU Horizon 2020/SPACE-funded project that was launched in March 2021 with a scheduled duration of **48 months**. It represents a collaboration between the **Department of Astrophysics at University of Vienna** (coordinating institute), the **University of Geneva**, and the **Konkoly Observatory**, Budapest.

Machine learning methods

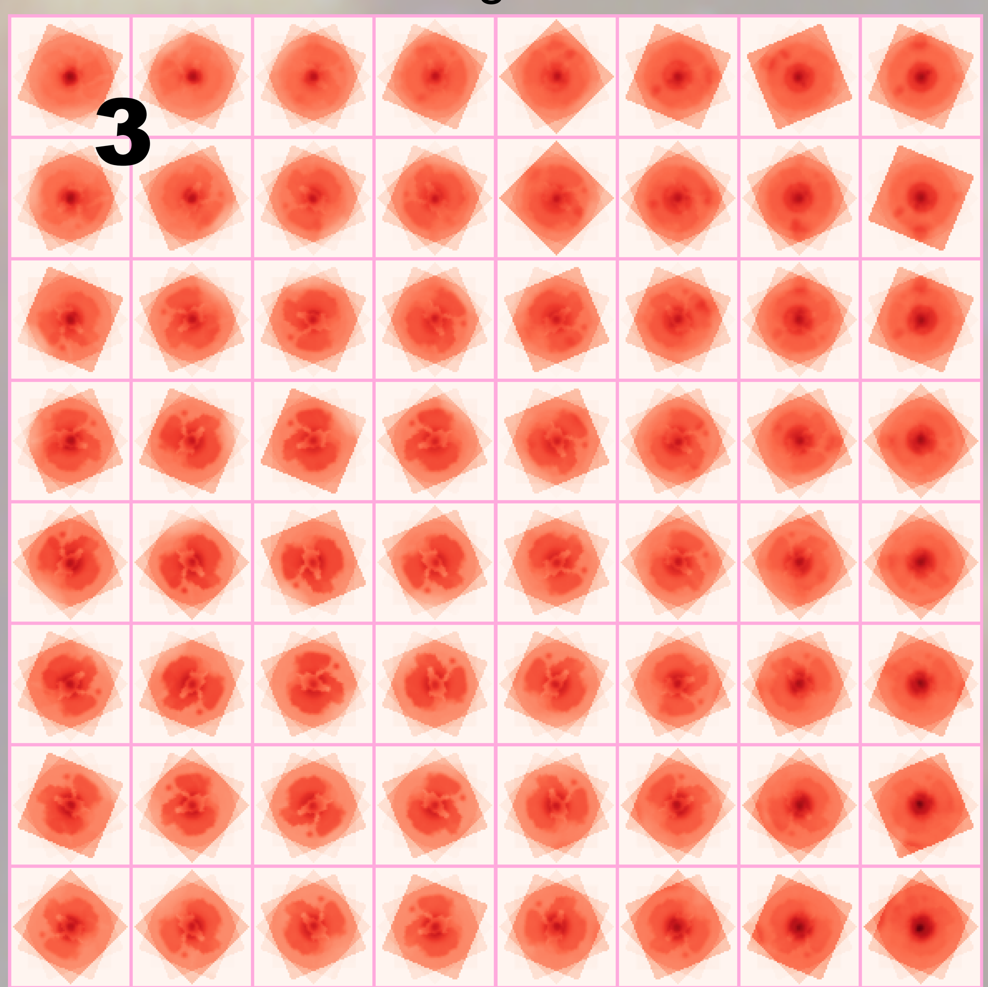
Different astrophysical problems require different machine learning methods. Redefining the star formation timescales is a complicated process. To have a reliable dataset, we not only collect data from astronomy related archives, but also reprocess the so-far best far-infrared dataset, collected by the Herschel Space Telescope. Young Stellar Objects are mostly located in fuzzy, nebulous regions. The background fluctuations not only make the source detection complicated but also heavily affects the photometric accuracy. With NEMESIS we are reprocessing all Herschel photometric observations using different Convolutional Neural Network (CNN) architectures for source identification, and solve the complexity of the photometry in the observational time by separating the source and background signals using Random Forest regression methods. Figure 1 shows the signal of a bolometer pixel crossing a celestial source (blue) and how the timeline is modified after source subtraction (orange). Figure 2 shows the IC348 region before (left) and after (right) point sources are subtracted.





Archival data can also introduce unwanted bias or uncertainty, therefore we are inspecting them with machine learning techniques in problems, like spurious source identification in the WISE data products.

Astronomical objects can have different shapes at different wavelengths. This morphology also provides important information. Self Organising Maps (SOMs) provide the solution of morphological classification. Astro-PINK is a very efficient tool to create such SOMs. An example of Spitzer/IRAC images of YSOs organised by PINK is shown on Figure 3.



We are also in the process of exploring the application of an Autoencoder Techinique to uncover hidden structures along with the capacity of our YSO dataset to be used to guess values for stars with missing data. The Autoencoder receives data with parameters described by pairs of measurement+uncertainty which are transformed into a probability distribution for that parameter, with missing data being described by flat probability distributions limited by the maximum and minimum value of that parameter in the whole sample.

Our dataset will be the largest homogenised panchromatic collection of YSOs that will not only contain information on the Spectral Energy Distribution of the sources, but also the spectral features, shape information, environmental parameters and temporal information will be included. This allows us to explore a dataset with many more features than ever before, and allows us to test the current star formation timelines on a statistical basis, compare it to results of supervised machine learning techniques and to apply unsupervised algorithms. The latter techniques have the capability to reveal so far hidden correlations, identify additional groups that may correspond to rapid events during the star- and planet formation, and also to detect outliers, like outbursting YSOs, which are key elements of understanding the chemical evolution of the protoplanetary disks.

Gaia Alerts

Light curves of millions of sources, including YSOs are provided by surveys like YSOVAR, ASAS-SN, ZTF and the Gaia space telescope. Gaia has the advantage of observing the whole sky. It collects photometric observations of some 1.8billion stars down to a faint limit of 20.7th mag in the G band for an average of 80 epochs during the nominal five year mission. Variability in YSOs occurs on timescales spanning a wide range depending on the physical processes. Circumbinary disc occultations last for ~100 days and can cause up to 4 magnitude dimmings. The EX Lup-type outbursts occur on a timescale of couple of hundred days and can also cause 2-4 magnitude brightening, while FU~Ori-type outbursts last even longer and cause even larger brightening, therefore they are also a potential subject to Gaia observations.

The Gaia Science Alerts (GSA) has been designed to produce notifications for transient phenomena, that is to say any event which would benefit from a timely reaction, and thus to avoid a potential science loss. The number of alerting sources varies from 1000 to several thousand per day, which go through a pipeline to minimise contamination but maximise the science return. This pipeline includes a careful eyeballing of the alert candidates. We designed an automated classifier based on a human voting system that is able to mimic human decisions in the process of publishing alerts with an accuracy >95%. This machine learning based method is now being tested in the GSA system and helps us to identify YSO events which are relevant to the science goals of NEMESIS, and allows us to discriminate objects where important processes affecting the star- and planet formation are seen in front of our eyes.

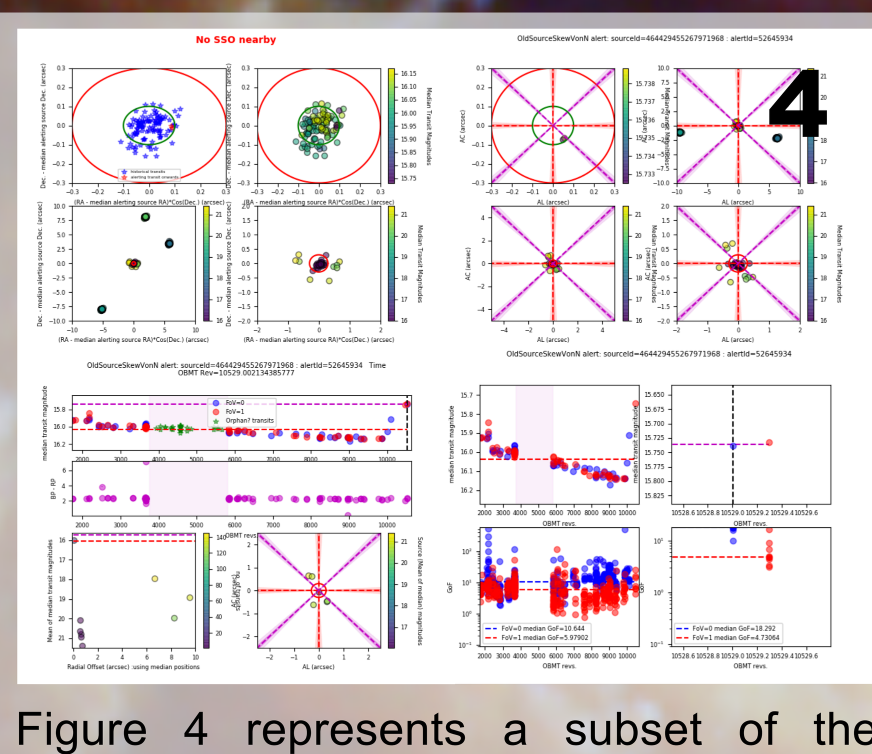


Figure 4 represents a subset of the information used to identify the sources that should be published for the wider community. The method uses PyTorch framework for an image based classification, and also uses Random Forest for a parameter based classification.

*email: marton.gabor@csfk.org
1) Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, Eötvös Loránd Research Network, Konkoly Thege Miklós út 15-17., H-1121, Hungary
2) CSFK, MTA Centre of Excellence, Budapest, Konkoly Thege Miklós út 15-17., H-1121, Hungary
3) Department of Astrophysics, University of Vienna, Türkenschanzstrasse 17, A-1180 Vienna, Austria
4) Department of Astronomy, University of Geneva, Chemin Pegasi 51, 1290 Versoix, Switzerland

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004141