# Feature engineering and machine learning for 21st century astronomy

Etienne Russeil

UNIVERSITÉ Clermont Auvergne

Thèse de Doctorat

École Doctorale des Sciences Fondamentales

*Spécialité: Particules, Interactions, Univers*

*Discipline: Physique/Astronomie*

# Feature Engineering and Machine Learning for 21st Century Astronomy

**Etienne Russeil**

Soutenance publique le 18 octobre 2024, devant le jury composé de :

| | | |
|---|---|---|
| Johan Bregeon | Chargé de recherche (LPSC) | Rapporteur |
| Emmanuel Gangler | Directeur de recherche (LPCA) | Directeur de thèse |
| Julien Donini | Professeur des universités (UCA) | Président de jury |
| Emille Eugênia de Oliveira Ishida | Ingénieur de Recherche (LPCA) | Directrice de thèse |
| Fabrício Olivetti de França | Professeur des universités (UFABC) | Invité |
| David Rousseau | Directeur de recherche (IJCLab) | Rapporteur |
| Paula Sánchez Sáez | Astronome adjointe (ESO) | Invité |
| Karine Zeitouni | Professeur des universités (UVSQ) | Invité |

UNIVERSITÉ Clermont Auvergne

# Résumé

**Title:** Ingénierie des caractéristiques et apprentissage automatique pour l'astronomie du 21e siècle

Les phénomènes transitoires astronomiques comptent parmi les événements les plus énergétiques de l'univers. Afin de percer leurs secrets, des télescopes de plus en plus performants ont été construits pour effectuer des relevés du ciel à grande échelle. Le futur observatoire Vera-C.-Rubin représente l'état de l'art d'une nouvelle génération de tels relevés. Il devrait détecter environ 10 millions de potentiels phénomènes transitoires chaque nuit, et produire une courbe de lumière pour chacun d'entre eux. Compte tenu de ce volume de données sans précédent, l'utilisation de méthodes d'apprentissage automatique pour les analyser est devenu inévitable. Cependant, la performance de la machine est limitée par la qualité des données à partir desquelles elle apprend. C'est pourquoi l'une des étapes les plus cruciales du processus réside dans l'extraction des caractéristiques des courbes de lumière. Idéalement, elle devrait encoder de manière optimale le comportement de l'objet tout en restant interprétable, de sorte qu'elle puisse être utilisée par des experts du domaine. L'objectif de cette thèse est de permettre l'extraction de caractéristiques informatives à partir de courbes de lumière multidimensionnelles. Ce manuscrit présente une série de méthodes conçues pour améliorer les procédures d'extraction de caractéristiques, offrant à la fois un ajustement de courbes contraint par la physique et une modélisation déduite des données. Le procédé Rainbow a été développé pour permettre l'ajustement simultané de courbes de lumière dans toutes les longueurs d'onde sur la base d'hypothèses physiques, générant un espace des paramètres plus informatif. Cette méthode permet à l'expert de sélectionner les modèles paramétriques les mieux adaptés aux spécificités de son domaine. Afin de guider ce choix, je propose la Régression Symbolique Multi-Vues (MvSR). Il s'agit d'une méthode basée sur les données, permettant la construction automatiquement d'un modèle paramétrique à partir d'un ensemble d'exemples. Elle offre à l'expert la possibilité d'élaborer des fonctions analytiques sur mesure. L'applicabilité de MvSR s'étend au-delà de l'astronomie et a été appliquée avec succès à diverses sciences. Enfin, les deux procédés sont combinés au sein d'un pipeline de classification adaptatif, qui prend en compte les propriétés individuelles de chaque courbe de lumière afin de choisir la description paramétrique la plus appropriée. Les résultats montrent que les caractéristiques extraites sont hautement informatives, ce qui permet la séparation de types de phénomènes transitoires pourtant similaires, même tôt dans leur évolution. Toutes les méthodes proposées dans cette thèse contribuent à une vision interdisciplinaire de l'astronomie moderne, dans laquelle les experts du domaine et les scientifiques des données collaborent pour construire des outils efficaces qui profitent à l'ensemble de la communauté.

UNIVERSITÉ Clermont Auvergne

# Abstract

**Title:** Feature engineering and machine learning for 21st century astronomy

Astronomical transients are among the most energetic phenomena in the Universe. In order to unveil their secrets, increasingly better telescopes have been built to perform large scale sky surveys. The upcoming Vera-C.-Rubin Observatory represents the state of the art of a new generation of such surveys. It is expected to detect around 10 million candidate transients each night, producing a light curve for each of them. Given this unprecedented volume of data, the use of machine learning methods to automatically analyze them is unavoidable. However, the machine can only get as good as the quality of the data it is learning from. Hence, one of the most crucial steps of this process lies in the feature extraction of light curves. Ideally, it should optimally encode the object behavior while remaining interpretable, so it can be used by domain experts. The goal of this thesis is to enable meaningful feature extraction from high-dimension light curves. This manuscript presents a series of methods built to enhance state-of-the-art feature extraction procedures, allowing for both, physically motivated modelling and data-driven description. In this context, the Rainbow framework was developed to enable simultaneous multi-wavelength light curve fitting based on physical assumptions, resulting in a more informative parameter space. This method allows the expert to select the best suited parametric models for specific science cases. In order to guide this choice, I propose Multi-view Symbolic Regression (MvSR). It is a data-driven method which automatically constructs a parametric representation from a set of examples, allowing the expert to build tailored analytical functions. The applicability of MvSR extends beyond astronomy and was successfully applied to various sciences. Finally, both frameworks are combined within an adaptive classification pipeline, which takes into account particular characteristics of each light curve to choose the most appropriated parametric description. It demonstrates that the features extracted are highly informative, enabling the separation of similar transient classes, even for poorly sampled light curves. Overall, all methods proposed in this thesis contribute to an interdisciplinary vision of modern astronomy, where domain experts and data scientists collaborate to construct efficient tools that benefit the community.

# Remerciements

En premier lieu, je tiens à remercier mes directeurs de thèse, Emille De Olivera Ishida et Emmanuel Gangler. Manu, merci pour ton soutien, ton enthousiasme et toutes nos discussions éclairantes à débattre de problèmes en tout genre. Je remercie par-dessus tout Emille, qui, en plus d'avoir été une encadrante scientifique de grande qualité, m'a transmis un enseignement philosophique et sociologique sur le monde. Tu as toujours cru en moi et mes idées, tu m'as toujours poussé à faire mieux et à aller plus loin. Tu m'as permis de voyager, de tenter, de rater, d'en rire, de comprendre, de rêver, et surtout de m'améliorer. Pour ça et pour tout le reste, un éternel merci.

Je remercie les membres de mon jury de thèse d'avoir accepté d'évaluer mon travail et d'être présents pour ma soutenance. C'est une chance pour moi d'avoir eu accès à un tel niveau d'expertise. Merci à Julien Donini, Fabrício Olivetti de França, Paula Sánchez Sáez et Karine Zeitouni. En particulier, je remercie les rapporteurs Johan Bregeon et David Rousseau pour leur lecture attentive et leurs retours.

Je tiens également à remercier le laboratoire de physique Clermont Auvergne au sein duquel j'ai pu mener ma recherche et faire des rencontres durant toute ma thèse. Un grand merci à Marine et Cyril, ainsi qu'à toute l'équipe administrative, pour vos sourires, votre légèreté et tous les bons moments passés ensemble. Je remercie tous les membres de l'équipe astro de m'avoir accueilli. Chloé, Corentin et Marie, vous avez été de redoutables adversaires de Mindbugs. Nos soirées, nos discussions et nos délires ont été légendaires, ce fut un plaisir de travailler à vos côtés. Hors du laboratoire, j'ai également partagé de précieux moments avec d'autres collègues. Merci à Miguel et Erwan pour leur accueil chaleureux à Toulouse. Merci Roman, avec qui je suis littéralement allé à l'autre bout du monde. A sincere thank you to Patrick for your hospitality in Urbana, you have been an amazing host. Merci Maxime pour tous nos déjeuners à refaire le monde et la physique. I am very grateful to Fabricio that immediatly trusted me and always brought positivity and momentum to the project. Merci à tous les chimistes de Clermont avec qui j'ai passé de super moments, malgré votre étrange adoration pour la mole. En particulier merci à Miche, sans qui faire tomber Beer-Lambert aurait été impossible. Enfin, je tiens à remercier mon ami Guillaume, qui, fidèle à lui-même, n'a pas hésité une seconde à me rejoindre

dans mes projets farfelus.

J'ai eu l'honneur de faire partie de deux collaborations au sein desquelles j'ai beaucoup appris sur la science, et grâce auxquelles j'ai eu de nombreuses opportunités de présenter mon travail partout dans le monde. Un très grand merci à Fink et à tous les membres de l'équipe, en particulier Julien et Anaïs, pour votre aide et votre accueil chaleureux. Le dynamisme ambiant et la bonne volonté au sein du groupe ont largement contribué à mon épanouissement scientifique. Merci à tous les membres exceptionnels de l'équipe du SNAD. Mon expérience avec vous m'a montré jusqu'où pouvait aller la créativité et l'effervescence intellectuelle. Masha, un profond merci ton accueil en Russie, et pour ta bonne humeur en France. Ce fut un réel plaisir de passer ces années ensemble dans *mon* bureau. I am very grateful to Kostya, that is not only humble and brilliant, but also extraordinarly benevolent. Thank you for your accompaniement and your ideas throughout the thesis.

Je tiens également à remercier mes proches, loin du monde de la physique, mais qui sont pour moi tout aussi essentiels à cette thèse. Merci à mes amis de longue date, Axel, Benjamin, Jordan, Thomas et Vidal, d'avoir été à mes côtés depuis plus de dix ans. Avec vous, j'ai ri aux larmes un nombre incalculable de fois. Vous avez été la solution miracle contre la déprime. Merci également à Adrien, Hugo et Mehdi pour toutes les vacances épuisantes, mais inoubliables qu'on a passé ensemble. Je remercie aussi Kevin, mon frère de cœur depuis toujours, avec qui j'ai pu partager une louable tentative de se mettre au sport tout au long de la thèse. Alexia, merci pour tous les ravitaillements en nourriture italienne. Et merci à Hugo, mon cousin vagabond, qui me soutient sans cesse quels que soient les kilomètres qui nous séparent.

Bien évidemment, je remercie du fond du cœur ma famille proche, et tout particulièrement ma mère, mon père et mon frère. Je suis inimaginablement chanceux d'avoir de vous avoir pour famille. Vous avez toujours cru en moi, vous avez toujours été fiers, et vous m'avez toujours incité à donner le meilleur de moi-même. Merci Jérémy, tu m'as servi d'exemple d'autodiscipline, et c'est grâce à toi que j'ai développé un calme à toute épreuve et une ouverture d'esprit sur le monde. Papa, tu es un soutien indéfectible dans ma vie. Je te suis infiniment reconnaissant de m'avoir accompagné dans toutes mes galères et mes conneries avec autant de bienveillance. Maman, merci pour ton amour inépuisable, tes rires et ta sagesse. Tu m'as transmis le plus important, le grain de folie qui me permet d'appréhender le monde avec légèreté. Vous êtes mes trois piliers, mes fondations.

Pour terminer, j'aimerais remercier mon quatrième pilier, Marion, celle que j'aime et qui partage ma vie depuis le début de l'université. Ensemble, on a vécu des aventures incroyables, on a ri, galéré, appris, grandi, encore ri, et on a même réussi l'exploit d'écrire nos thèses en simultané. Merci d'avoir ensoleillé mon quotidien, de m'avoir supporté en toute circonstance et de m'avoir toujours suivi dans mes projets et mes obsessions. Puissent les aventures durer toute une vie.

Cette thèse est dédiée à Tristan Carl.

# Table of contents

# 1

# Introduction

The arrival of big data, induced by the development of new technologies, has overwhelmed the astronomical community and fostered an era of data-driven scientific development (Zhang and Zhao, 2015). In addition to the challenges and opportunities created by these changes, it is paramount to remind ourselves that the same technology that allowed astronomers to have more data than ever before also made its accessibility non-trivial. Beyond data-rights and specific survey policies, the mere volume of data produced by modern large scale surveys imposes the use of new methods and tools for data filtering and analysis, without which their scientific potential may never come to fruition.

In this context, the adoption of machine learning (ML) methods by the astronomical community should not come as a surprise. There are numerous examples which showcase how the combination of new methods and data have been producing important scientific results (e.g. Domínguez Sánchez et al., 2018, Ishida, 2019, Shallue and Vanderburg, 2018). Nevertheless, whenever possible, we should keep in mind the scientific context of the data at hand and the possibility to use domain knowledge to lead the development of new methods and data analysis strategies. ML should always be used as a tool, built in collaboration with domain experts, to assist science development. Following this principle, this thesis showcases an example of such endeavor in the context of time domain astronomy. I approached the problem of the optimal characterization of transient light curves (i.e. the evolution of a source's brightness over time) by building systems which can automatically process and add value to them, thus lowering the burden on the domain expert and enabling smooth interaction with data from large surveys.

Light curves are generated by time-domain photometric surveys, which repeatedly observe the sky in search for variability. They carry precious information regarding the physical nature of the source, but they are multidimensional, inhomogeneous and noisy time series. Therefore, accessing such information is a non-trivial task, in particular for ML methods. A solution commonly used is to summarize properties of each light curve in a finite number of features. The optimal execution of this step constitutes one of the main challenges of modern transient astronomy. Ideally, this description should translate the source luminosity behavior, while still encompassing the physical reality behind multi-wavelength emissions. In addition, in order to enable allocation of follow-up resources, we should be able to extract information from light curves in real time and as early as possible. Paradoxically, in the era of big data, one of the main challenges is to extract information from a small number of data points per object.

I first tackled the problem of the physical multi-passband fitting of light curves, for which we propose the Rainbow framework. It consists of a multidimensional parametric model, describing the brightness of the object in time and wavelength, and thus offering a description of the bolometric and the temperature properties of the source. It relies on the assumption that the latter is approximately a blackbody emitter. Using phenomenological parametric functions for the bolometric flux and temperature evolution, I obtain a simple and efficient model to characterize and classify transient events. However, the optimal choice of parametric models to be used within Rainbow remains uncertain.

Although multiple functional forms for fitting transient light curves have been proposed and used in literature, none come from first principles. They have been handcrafted by experts to match the expected photometric behavior of the sources. As such, they neither guarantee opti-

mality, i.e. that they fit as best as possible, nor simplicity, i.e. that no simpler equation would yield similar results. I propose Multi-View Symbolic Regression (MvSR), a ML based tool to assist experts in their exploration of possible optimal parametric models. It is based on Symbolic Regression, a data-driven ML method which aims at constructing the optimal mathematical representations of a dataset. Our implementation extends the traditional approach. It enables the use of multiple datasets assumed to be generated by the same phenomenon, to produce a common parametric equation capable of describing them individually. This tool is highly effective in generating models for transient light curves. It proposes solutions of various complexities, including unintuitive forms unlikely to ever be found by experts without the assistance of MvSR.

Incorporating the panel of parametric functions found into the multi-passband framework offered by Rainbow, an optimized representation of transients can be built. Depending on the state of the light curve, more or less complex models can be used to describe more or less complex bolometric and temperature behaviors. Rather than separating the early and late modeling, I present an adaptive approach for classification of transient light curves. Based on the number of observations available across different passbands, optimal parametric functions are chosen, thus enabling coherent choices, preventing overfitting and underfitting. This procedure is suited for very early classification, hence maximizing the potential for future candidate follow-up.

Although the underlying principles driving this work could be applied to a larger range of science cases, they are particularly adapted to the reality of time domain data. Indeed, surveys such as the Zwicky Transient Facility (ZTF, Bellm et al., 2019a) and the Vera C. Rubin Observatory Large Survey of Space and Time (LSST, Ivezić et al., 2019) make (will make) transient time-series publicly available shortly after detection, while the task of filtering and distributing this data to the target scientific communities is left to community brokers [1]. During this thesis, I have been an active member of Fink, one of the seven brokers chosen by LSST to receive the raw alert stream. It follows a decentralized approach to science pipeline construction, where each science module is developed by a different team. This paradigm guarantees that the output of each module fulfills the need of the user, thus placing the domain expert at the center of infrastructure development. This idea is anchored in the thesis. It was even further strengthen by my collaboration with SNAD [2], an international network of researchers specialized in the adaptation of anomaly detection techniques to astronomy. Both collaborations are places where astrophysicists and data scientists are in constant interaction to produce efficient and accessible tools.

The methods presented in this manuscript have been deeply shaped by this vision. They take advantage of astrophysical expertise to build tailored pipelines, able to process light curves in the form of alerts. Such interpretable framework, applied to a public data stream, has already proved its applicability and is currently used by the astronomical community. This manuscript presents in detail Rainbow (Chapter 5), Multi-View Symbolic Regression (Chapter 6), and a complete adaptive classifier (Chapter 7) framework, as well as the scientific foundations on which they are based (Chapters 2, 3, and 4). Analysis and quantitative results associated to them are provided and demonstrate their efficiencies and relevance in the current big photometric data context. Finally, conclusions and perspectives are presented (Chapter 8).

1. https://www.lsst.org/scientists/alert-brokers
2. https://snad.space/

# 2

# Astronomical concepts

## 2.1 Notions of astronomy

This section presents the fundamental concepts of astronomy that are essential for the comprehension of the thesis. Understanding these concepts is crucial, as data must be interpreted with an awareness of how it was generated and its precise significance. Consequently, this section covers both observational and basic astrophysical principles.

### 2.1.1 Magnitude and flux

In the Ancient Greece, the astronomer Hipparchus started to catalog stars depending on their positions and gathered them into constellations. Perhaps more importantly, he took notes of their brightness, grouping them into three categories: brilliant, second degree and faint. This system historically constitutes the first step into photometry. Almost 300 years later, Ptolemy went further into the classification by creating what he called a magnitude scale, ranging from one to six and ordered from the brightest stars (1) to the faintest (6). It is believed that the magnitude was attributed based on how quickly the stars appeared to the naked eye as the night advances (Miles, 2007). This historical definition is at the origin of the current magnitude system.

In modern astronomy, apparent magnitude (hereafter, $m$) is the measure of the perceived brightness of an astronomical source, compared to the brightness of a reference source ($m_{ref}$). The magnitude is unitless and is defined such that, for a given passband (see Section 2.1.2), an increase of 1 in magnitude implies that the source is 2.512 times dimmer than the reference. Reciprocally, a decrease of 1 implies that the source is 2.512 times brighter, such that:

$$m - m_{ref} = -2.5 \log_{10} \left( \frac{F}{F_{ref}} \right). \tag{2.1}$$

Here, $F$ and $F_{ref}$ refers respectively to the flux of light received from the source and from the reference. Its unit is the $J \cdot s^{-1} \cdot m^{-2}$ in the standard system (S.I.), but it is commonly expressed in $erg \cdot s^{-1} \cdot m^{-2}$ (1 erg $= 10^{-7}$ J). The flux is the actual physical value measured by the cameras of modern telescopes (see Section 2.3). The logarithmic nature of magnitude is a consequence of the physical reality of the human eye, for which the perceived brightness of a source does not scale linearly with flux, as illustrated by Figure 2.1. Currently, both flux and magnitude systems are widely used in astronomy.

Different zero points are commonly used as reference ($m_{ref}$=0) to build magnitude scales. The star Vega has been the first historical zero point used in astronomy. However, other magnitude systems exist, such as the AB magnitude, which defines a constant flux emission at any wavelength as the reference value (Bessell, 2005a). Apparent magnitude does not carry information on the intrinsic brightness of an object. A source's brightness can be explained equally well by a faint and nearby, or a luminous and far away object. Therefore, we define the absolute magnitude of a source (hereafter, $M$) as being the apparent magnitude that we would observe if the object was placed at a distance of 10 parsecs (parallax per angular second [1]). Knowing the distance, one can compute the absolute magnitude with:

$$m - M = 5 \log_{10}(d) - 5. \tag{2.2}$$

1. 1 parsec $\simeq$ 3.26 light years

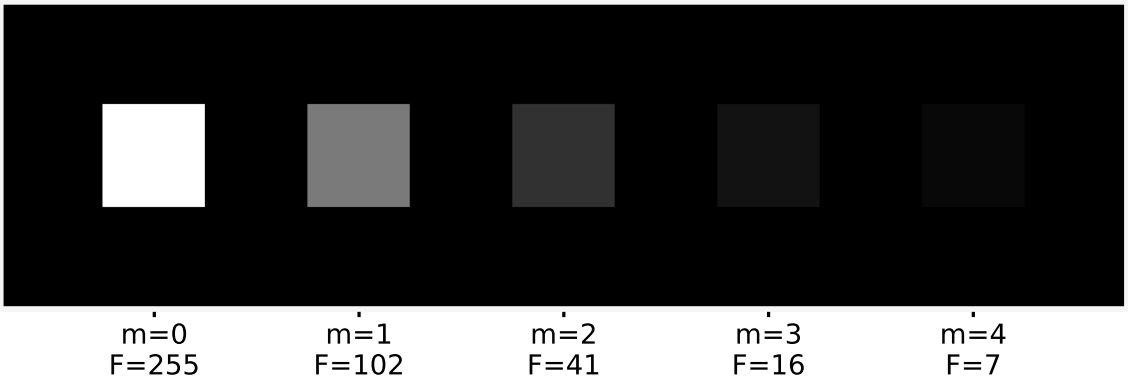



FIGURE 2.1 – Illustration of the brightness at different apparent magnitude. The associated flux corresponds to the intensity of the pixel, varying from 255 (white - left most square) to 0 (black - corresponding to the background).

As an example, the Sun has an apparent magnitude of $m_\odot \simeq -26.7$ [2] mag but an absolute magnitude only of $M_\odot \simeq 4.8$ mag (Measured with a Johnson_B filter by Willmer, 2018).

### 2.1.2 Photometry

Magnitude was historically defined for the visible part of the spectrum. However, astrophysical objects emit electromagnetic radiation in a much broader wavelength range. From highly energetic gamma rays to low frequency radio sources, the study of different emissions carries a lot of information about the physical processes from which they originate. In order to properly characterize the light received, modern photometric telescopes use broad-band filters that block all wavelengths but a specific range, also called passbands. This enables a finer understanding of the properties of the objects. This method of observation is called photometry. It consists in the measurement of the flux of astronomical sources across multiple passbands. In this context, the flux previously defined must always be associated with a passband [3]. Therefore, we define an additional observational property for a source, the bolometric flux (and bolometric magnitude), which corresponds to the total flux emitted across the whole electromagnetic spectrum.

Telescopes are designed to optimally study sub parts of the spectrum, each associated to specific scientific goals (see Section 2.1.7). They commonly observe using a series of filters in order to cover a wider wavelength range. There exist many different filters and filter systems which are currently used. The $UBVRI$ (Bessell, 2005b) historically constitutes the first photometric system. It is frequently used to determine the spectral class of a star through the Hertzsprung-Russell diagram (detailed review by Arp, 1958). Figure 2.2 displays the transmission efficiency of each $UBVRI$ filter. It covers wavelengths ranging from ultraviolet (U) to infrared (I). Sections 2.3.1 and 2.3.2 also present the *ugrizy* photometric system, which is used in the essential of the data presented in this thesis. The specific shape of passband transmission is constrained by the physics of the filter. In most cases [4], ideal filters should look like step functions, with perfect

2. The $\odot$ symbol designates sollar quantities.
3. Historically, the human eye was acting as a filter that allowed only the observation of the visible light.
4. Some passbands can be purposefully designed with gaps in the transmission, e.g. to avoid absorption features of the Earth's atmosphere.

transmission over the desired range and zero transmission everywhere else.

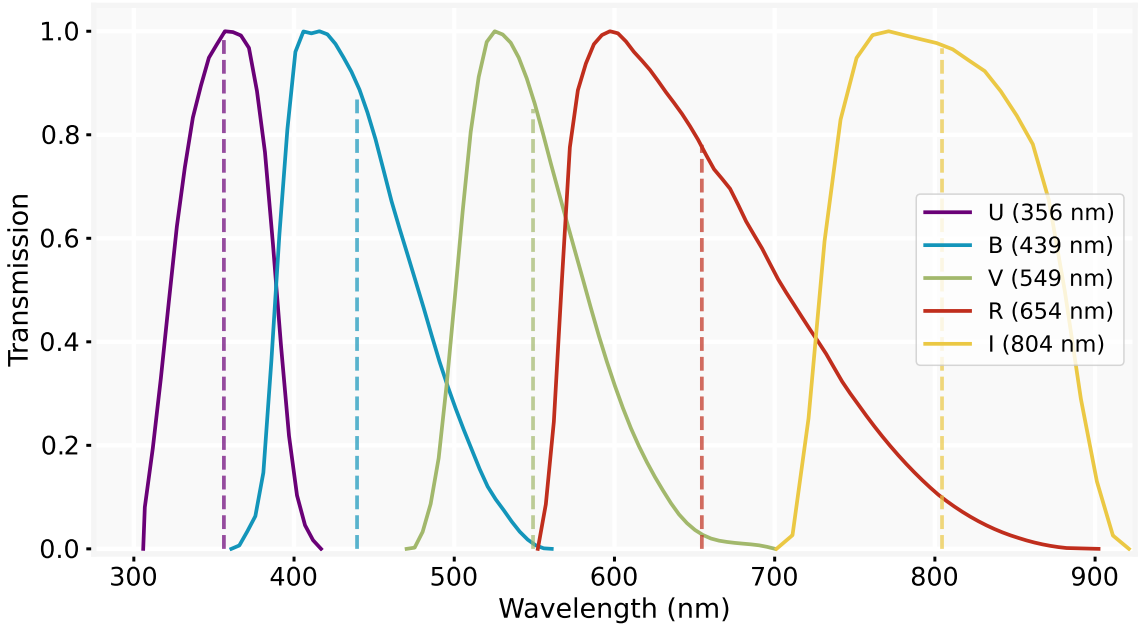



FIGURE 2.2 – Transmission efficiency of the standard $UBVRI$ filters from the `sncosmo` package (Barbary et al., 2016). Each profile is associated with its effective wavelength (inset), which corresponds to the mean wavelength of the detected photons as defined by Koornneef et al. (1986).

Photometry is an invaluable tool to study astrophysical objects. It is used to efficiently study the brightness of sources, and their potential evolution through time (Section 2.2.3). The usage of cameras to acquire pictures of the sky (Section 2.2.1) further increased the importance of photometry. For example, it enabled the study of galactic morphologies (Hubble, 1926), a crucial component to understand the history of our Universe. However, it only provides a broad wavelength overview of the sources, with no detailed information about the physical processes at play. To gain deeper insights, we turn to spectroscopy.

### 2.1.3 Spectroscopy

Spectroscopy is another observational method that can be used to study an astronomical source. It is based on the decomposition of the incoming electromagnetic radiation, which allows measurements with greater details in a large range of wavelength. At such resolution, absorption and emission lines can be observed. They are caused by the chemical elements present along the line of sight. Indeed, specific elements will absorb or emit electromagnetic radiation at precise wavelengths, respectively causing drops or spikes in the spectrum. These lines can often act as signatures for certain astrophysical events, which enables reliable classification. The spectrum of a type Ia supernova (see Section 3.1.2) with a fine binning of half a nanometer is presented in Figure 2.3. In this example, strong Si II absorption line around 600nm, coupled with the absence of hydrogen lines, allows us to identify the object as a supernova of type Ia.

Despite all these advantages, spectroscopy is fundamentally limited by the time required to measure a good quality spectrum. This time is proportional to the resolution of the spectrograph (Burns, 2018). Therefore, in comparison to broad passband photometry, the integration time re-

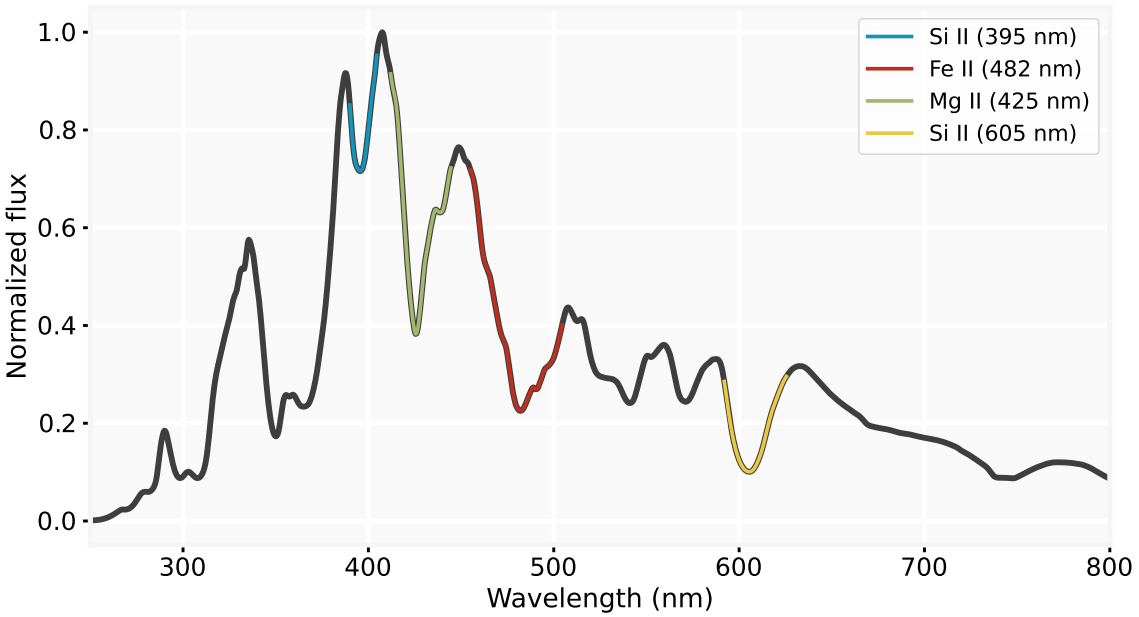



FIGURE 2.3 – Spectra of the supernova Ia SN2002bo $\sim$ 10 days after the explosion (Blondin et al., 2015). The characteristic absorption lines of silicon, iron and magnesium have been highlighted. The flux was normalized to the maximum value.

quired to obtain a good signal-to-noise spectrum can be several orders of magnitude higher. The issue is further enhanced by: the limited field of view of the instruments ; the spatial dispersion of spectra on the camera that may cause sources to overlap ; and the difficulty in obtaining high precision calibration (Weiler, M. et al., 2020).

This represents one of the fundamental limits of observational astronomy. Although it constitutes the most precise information we can acquire, obtaining spectra is so time-consuming that it is a limited resource. Considering the richness of our Universe, the amount of interesting sources to be studied will always be far greater than our spectroscopic capacities. Photometry on the other hand, through the help of automatic surveys, allows fast complete sky coverage, thus enabling the study of brightness evolution for a large variety of astrophysical sources (see Section 2.2). Even if less informative on a single source scale, it can provide high statistics and thus very valuable insights into our Universe. In practice, they are complementary approaches. In case a particular object is of interest based on its photometry, a spectroscopic follow-up can help to understand it in more depth.

### 2.1.4 Blackbody radiation

The electromagnetic emissions from stars and some other astrophysical entities are often approximated as black bodies. That is, objects which perfectly absorb all incoming radiation and for which their emissions purely depend on their temperature. Under this hypothesis, the Stefan-Boltzmann law states that the flux of energy emitted from the surface of the object is proportional to its temperature to the fourth power, such that:

$$F = \sigma T^4, \tag{2.3}$$

with $\sigma$ being the Stefan-Boltzmann constant. The spectral radiance of a blackbody is given by the Planck's law,

$$B_\lambda(T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k_B T}} - 1}, \tag{2.4}$$

where $B_\lambda$ is the spectral radiance expressed in $erg \cdot cm^{-2} \cdot s^{-1} \cdot nm^{-1} \cdot sr^{-1}$, with $\lambda$ the wavelength, $T$ the temperature, $k_B$ the Stephan-Boltzmann constant, $h$ the plank constant and $c$ the speed of light. It represents the energy outputted each second in a wavelength range which is radiated by a surface area into a solid angle of space. This relationship implies that the temperature of the photo sphere (outermost layer) of a blackbody can be deduced purely from its spectral energy distribution (SED), i.e. the distribution of energy emitted at each wavelength at a given time. However, no astrophysical source behaves exactly as a blackbody, therefore it can only be used as a first order model. It is mainly due to the absorption lines in the spectra, but also to the potential non-thermal radiations which are not taken into account in the blackbody model. The blue line in Figure 2.4 presents a spectrum of the Sun from 200 nm to 1750 nm, and illustrates its overall agreement with the blackbody hypothesis.

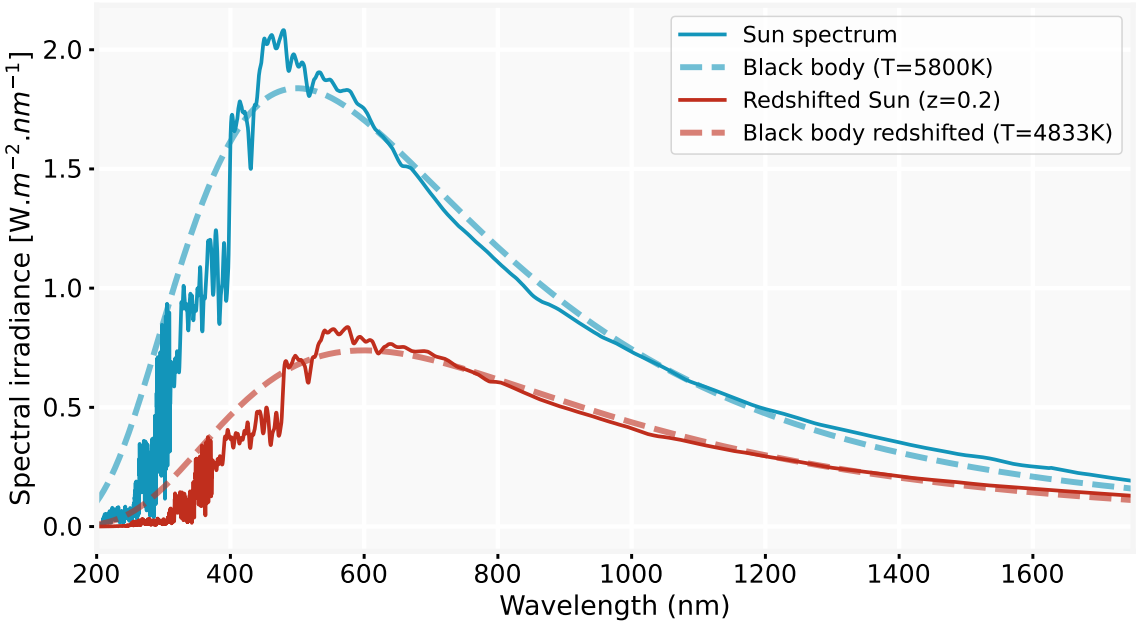



FIGURE 2.4 – Solar irradiance spectrum acquired by using a combination of satellite and sounding rocket observations (Woods et al., 2009). The energy irradiance has been derived from Planck's equation (Equation 2.4) to obtain the blackbody at $T = 5800$ K, which fits the general behavior. A redshifted ($z = 0.2$) spectrum of the sun, along with a fitted cooler ($T = 4833$ K) blackbody model, are shown in red.

An interesting mathematical property of black bodies is that the peak emission depends on the temperature, following the Wien's displacement law:

$$\lambda_{max} \approx \frac{2898\ [\mu m \cdot K]}{T}. \tag{2.5}$$

It implies that sources of high temperature will emit at lower wavelength, and therefore will be bluer. On the contrary, colder objects will produce redder emissions. Based on this law, we can

show that the Sun's peak wavelength emission is around $\approx 500\ nm$.

### 2.1.5 Redshift

The redshift (denoted, $z$) is the phenomenon where the measured wavelength of electromagnetic radiation from an astronomical object is increased ($\lambda_{obs}$) when compared to the same emission in our rest frame ($\lambda_0$), resulting in a shift toward the red end of the spectrum. It is expressed as,

$$1 + z = \frac{\lambda_{obs}}{\lambda_0}. \tag{2.6}$$

It can be attributed to two primary mechanisms. For local sources, peculiar velocities can induce a Doppler redshift effect, which occurs when the source of radiation is moving away from us. To the opposite, sources moving towards us will appear blue shifted. For distant sources, the cosmological redshift prevails. It is caused by the expansion of the Universe, which results in the space between objects being stretched and, thereby, elongates the wavelength of the photons traveling through it. This results in most objects being redshifted to us, with the intensity of the effect increasing with distance. Measuring the redshift (Equation 2.6) automatically gives us the velocity of the source relative to the observer ($v = c/\lambda$). For a given cosmological model, this allows the estimation of the age of the Universe at the moment of emission, and consequently enable the calculation of distances (Hogg, 1999). The red line of figure 2.4 demonstrates the effect of the redshift on the observed spectrum of an object. In this fictional example, we assume that the Sun has been measured at $z = 0.2$. We observe a stretch in wavelength by a factor $1 + z$ (Equation 2.6), shifting its peak emission closer to 600 nm, accompanied by a dimming of the spectrum. A redshifted blackbody behaves exactly like a cooler non-redshifted blackbody, such that:

$$T_{obs} = \frac{T}{(1+z)}. \tag{2.7}$$

In our hypothetical scenario, it would seem as if the Sun was 1000 K cooler than it really is. Therefore, it is important to make a clear distinction of the radiance temperature deduced by a Planck's law fit and the real effective temperature of the astrophysical object. In practice, the redshift of a source is measured spectroscopically by comparing the absorption line wavelengths, $\lambda_{obs}$, of different chemical elements, to their known theoretical counterparts, $\lambda_0$, (see Huggins, 1868).

Comparing the above descriptions of photometry and spectroscopy, we can see that they tackle the same source of information, electromagnetic radiation emitted by a source. Although the tools used for each measurement differ in the wavelength range they cover. Photometry can be described as the integrated signal of a spectrum in a large range of wavelength, convoluted with the filter transmission efficiency. This connection has been extensively explored in the literature. Methods to estimate redshift using broad band photometry have flourished (Butchins, 1981). Recently, given that astronomy enters an era of big photometric data, increasing efforts are put into the improvement and development of new such alternatives (such as Ilbert et al. 2006, Gerdes et al. 2010, de Oliveira et al. 2022 or Tanigawa et al. (2024)).

### 2.1.6 Color index

Spectroscopy can be used to obtain detailed information about a source composition and the distribution of its electromagnetic wavelength emission. Assuming it behaves as a blackbody (Section 2.1.4), the spectrum can also provide the temperature and color of the object. But as stated in section 2.1.3, spectroscopy is a precious resource, and most of the sources we will ever observe will not be associated with a spectrum. Photometric measurements through different filters can be seen, to some extent, as low resolution spectroscopy. That is, a photometric observation of a star using two filters is equivalent to obtaining a spectrum with only two wavelength bins. Assuming a blackbody SED, it is possible to constrain a temperature that fits these observations. Therefore, even if it is less precise than a complete spectrum, any pair of measurements acquired using different filters contains information about the color of the object.

Based on this, photometrical tools have been developed to encode this physical color information, they are called color index. They are computed by subtraction of the observed magnitude from two different filters. Although color index are specially useful to assess the color of blackbodies, they are much more general tools used to characterize any astrophysical object. In the literature, the term color index is often simply expressed as color, but one should keep in mind the difference between the index and the perceived color of an object. Historically, the redder magnitude is subtracted to the bluer magnitude such that a negative color index indicates a blue source, and reciprocally, a positive color index means a red source [5]. Color index computed from the UBVRI photometric system (Section 2.1.2), such as U-B, B-V, V-R or R-I are the most widely used. In particular, B-V is famous for being commonly used in the Hertzsprung-Russell diagram (Russell, 1919), a crucial tool for the description of stellar evolution.

### 2.1.7 Observations across the electromagnetic spectrum

Each photometric telescope is designed to observe a given range of the electromagnetic spectrum. From radio to gamma, the study of each type of emission reveals different facets of our Universe. Various scientific goals revolves around the analysis of specific wavelength ranges. Here we present a few examples of active research topics in the field.

— The measurement of **radio** emissions, with facilities such as the *Atacama Large Millimeter Array* (ALMA, Schloerb et al., 2019), enables the study of star formation otherwise obscured by dust [6]. It is also valuable to understand the nature of sources like active galaxies (Section 3.3), gamma ray burst, tidal disruption events (Section 3.2) and even planetary atmospheres.

— Studying **microwave** primarily reveals the cosmic microwave background, i.e. the thermal signature emitted homogeneously in the Universe $\sim$ 380000 years after the Big Bang. Telescopes, like *Planck* (Planck Collaboration et al., 2020a), have mapped with great details its anisotropies, providing essential insight for cosmological models.

— **Infrared** emissions, observed with telescopes like the *Wide-Field Infrared Survey Explorer* (WISE, Wright et al., 2010) or the *James Webb Space Telescope* (JWST, Gardner et al., 2006), enables the study of thermal sources, in particular cold ones such as brown dwarfs

5. Brighter objects are assigned lower magnitudes, see Section 2.1.1.
6. The interstellar medium is transparent to radio emissions.

or young protostars. In addition, the first stars and galaxies of the Universe appear very redshifted, and can be studied in the infrared. Their observations put constraints on their formation and evolution history, as well as on dark matter and dark energy models.

— The observation of **visible light** constitutes the oldest form of astronomy. Many different instruments have been designed for this purpose, such as the *Sloan Digital Sky Survey* (SDSS, York et al., 2000), the *Hubble Space Telescope* (Freedman et al., 2001), the *Zwicky Transient Facility* (ZTF, Section 2.3.1), or the upcoming *Vera Rubin Observatory* (LSST, Section 2.3.2). They are used to observe a wide variety of objects from stars, galaxies and nebulae to solar system objects, active galaxies and transient phenomena (Chapter 3).

— **Ultra violet** (UV) emissions, studied e.g. by *Far Ultraviolet Spectroscopic Explorer* (FUSE, Moos et al., 2000) or by the *Hubble Space Telescope* (for near UV), are used to understand the properties of the interstellar medium. They also enable the observation of hot astronomical objects, and have been used to characterize the abundance and evolution of the lighter elements of the Universe.

— Measuring **X-ray** radiations enables the study of both extreme temperatures and non-thermal radiation processes. Telescopes, such as *Chandra X-ray Observatory* (CXO, Weisskopf et al., 2000), allow the high-energy observation of objects such as active galaxies, or pulsars. X-ray are also used to constrain our understanding of dark matter by studying large scale structures.

— **Gamma-ray** emissions are the signature of very high energy phenomena, they have been studied with observatories like the *Fermi Gamma-Ray Space Telescope* (Atwood et al., 2009) or the *High Energy Stereoscopic System* (HESS, Aharonian et al., 2006). Their study gives insight about the physical mechanisms at play in active galaxies and supernova remnants. It is also crucial to understand the brief and intense gamma radiation produced by gamma ray burst events, along with their binary merging origin.

This thesis is entirely focused on visible and near infrared observations, with the goal of characterizing and classifying transients in what is known as time domain astronomy. Most of the data used in this work comes from telescopes with passbands ranging from 300 to 1100 nm. The next section introduces the concepts behind time domain astronomy, and presents in detail the two telescopes at the origin of the thesis data (real and simulated).

## 2.2 Time domain astronomy

For most of the history of astronomy, the sky has been seen as immutable. Besides local objects such as the moon, meteors and planets (the word planet comes from the Greek *planetai* which translates to wanderers), most of the Universe appeared never changing. This illusion was due to the limiting magnitude of the human eye, which only allows seeing to a magnitude of up to ∼6 mag, in the best possible observation conditions. Given such limitation, most astrophysical objects are invisible to us. In fact, hints that the Universe is in constant evolution have appeared only a few times in history in the form of nearby supernovae (detailed in Chapter 3) of very high apparent brightness. SN1006 appeared in the sky in 1006 and could be observed for up to two years (Stephenson, 2010). It is believed to be the supernova with the lowest apparent magnitude ever observed, currently estimated around -7.5, and was at first even visible during daytime. This example, along with a few others such as the famous Tycho Brahe supernova in

1572 (Krause et al., 2008), formed the dawn of modern time domain astronomy.

The evolution of astronomical objects through time is at the basis of the study of a wide variety of events. In our galaxy, many stars exhibit periodic variability in their brightness over time, which can be explained intrinsically (e.g. by the pulsation of the celestial body), or extrinsically (e.g. by the orbit of a companion obstructing the line of sight). Other phenomena, such as high proper motion (Eckart and Genzel, 1997) or star flaring (Heyvaerts et al., 1977) can be at the origin of non-periodic variability. In this thesis, the focus will be put on the variability of extragalactic sources and in particular transient events, those with a unique episode of brightness which suddenly appear, and stay observable for a limited time before going extinct forever. Chapter 3 lists the main scientific interests associated to them, along with a description of the physical phenomena at their origin. The change in paradigm, from static to evolving sources, was accompanied by adapted instruments and methods. The following sections present the main tools required for time domain astronomy analysis.

### 2.2.1 CCD imaging

In order to understand the flux temporal evolution of our sources, we must first understand in more details how the flux is measured at a given time. Since their invention in 1970, charge-coupled devices (CCDs, Boyle and Smith, 1970) have become the standard technology used to detect visible to near infrared photons. A CCD is essentially of a semiconductor which converts photons to electrons [7] and stores the charges in potential wells. Subsequently, the electronic charges can be read out to acquire the information numerically and reconstruct a 2D image. Each potential well represents a pixel, and their number on a single CCD can reach up to $10^7$ (Ivezić et al., 2019) . Wide field astronomical cameras are built from the assembly of multiple CCDs. Figure 2.5 shows an example of a picture of the sky taken by the Zwicky Transient Facility camera (ZTF, Bellm et al., 2019a), composed of 16 CCDs, with a total resolution of 600 megapixels (see Section 2.3.1).

In astronomy, point sources appear in reality as blurred extended dots. This effect is a pure physical consequence of the optical system. The mathematical transformation from a point source to a blurry dot is called the Point Spread Function (PSF, Heasley, 1999). Even in ideal conditions it is impossible to perfectly recover a points source, since at the minimum, the aperture will diffract the electromagnetic waves and produce an Airy disc, i.e. a concentric diffraction pattern. However, in practice, this effect is small compared to the contributions from other distortions, like the telescope's optics or the atmosphere. The latter largely constitutes the main contribution to the PSF of ground based telescopes. Since the optical system changes from exposure to exposure, the PSF needs to be re-evaluated each time. In practice, the current and most precise method is to fit a PSF template to the observed images. This delicate procedure implies: automatically finding stars ; properly evaluating the background level ; fitting the PSF template to isolated sources ; and finally applying the fit to all other sources in the image (Heasley, 1999). This evaluation is crucial because the estimation of the flux of a source is done by convolution of the PSF with the electron count of the CCD. Therefore, any subsequent photometric analysis relies on the quality of the PSF.

7. Given a certain efficiency called the quantum efficiency.

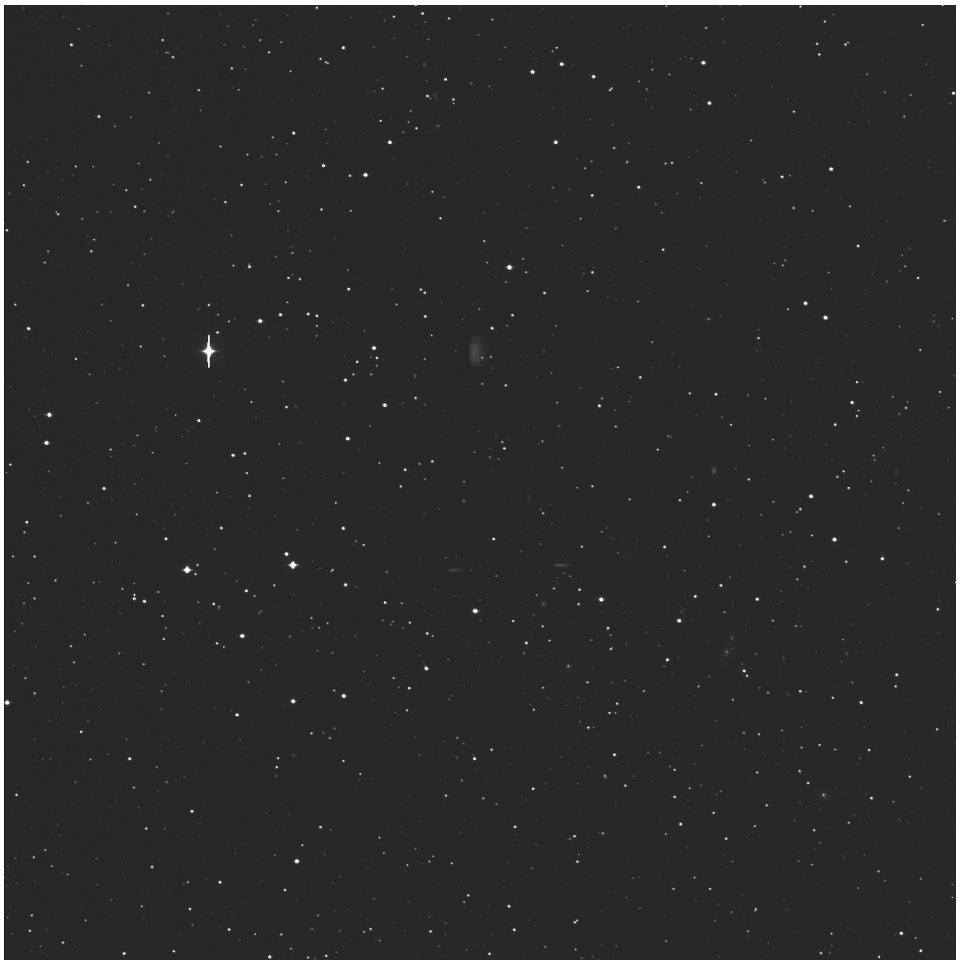


FIGURE 2.5 – Picture of the sky obtained by the Zwicky Transient Facility [8]. It has been acquired by one fourth of a CCD only, using the ZTF-g filter. It corresponds to a field of view of $\sim 0.72$ square degrees.

### 2.2.2 Difference imaging

Rapidly after the first usage of CCD for photometric imaging, astronomers have proposed to perform image subtraction as a mean to study the brightness evolution of sources. The idea, introduced by Tomaney and Crotts (1996) and Alard and Lupton (1998), is to acquire a template (also called reference) image of a patch of the sky, which can later be subtracted to a future image (hereafter, science image) of the same patch, after properly matching the backgrounds and PSF positions. The resulting difference image reveals all the variable sources and removes the stars and galaxies (which vary over longer time periods). It has proven to be extremely effective on densely crowded fields. For the specific problem of finding variable objects, it is more effective to estimate only the portion of a sources brightness that changes from image to image. Using high signal-to-noise reference images, difference imaging can achieve errors close to the theoretical Poisson limit (Bramich, 2008). Figure 2.6 illustrates the image subtraction process used by ZTF (Section 2.3.1). The difference image unambiguously highlights the source that recently appeared in the sky.

9. https://fink-portal.org/ZTF23aandvzg

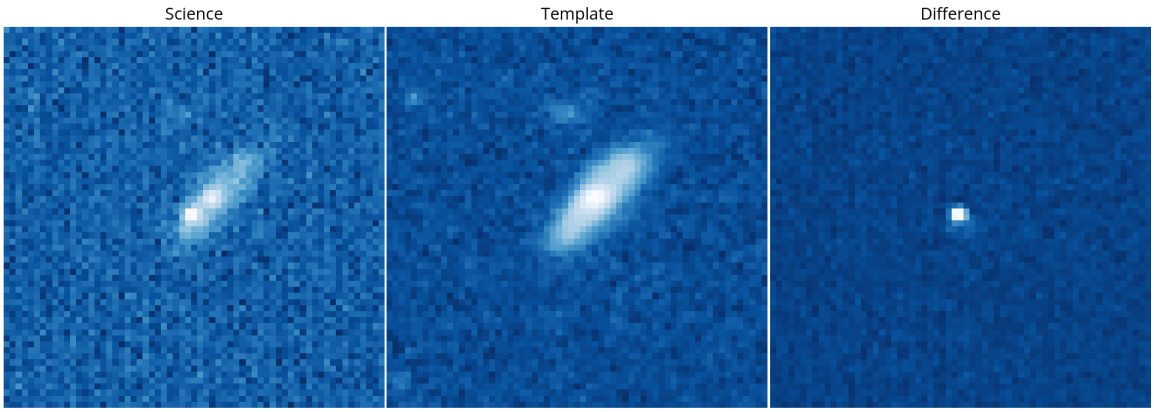



FIGURE 2.6 – Cutout images (63 × 63 pixels around the source) resulting from the difference imaging of the ZTF (ZTF23aandvzg[9]). Template (middle) is the reference image taken in the past, Science (Left) is the image newly acquired, and Difference (right) shows the subtraction of both images.

A decision on whether the new source is real or is purely due to randomness can then be taken by computing the signal-to-noise ratio (e.g. ZTF requires a statistical confidence of $5\sigma$). However, the subsequent analysis of the subtracted image only gives information about the flux variation, and is therefore called difference flux (or difference magnitude). It is a valuable source of information that can be repeatedly taken through time (Section 2.2.3). In order to convert it to an apparent flux, a calibration based on a set of well characterized sources must be performed. It is important to emphasize that difference fluxes acquired with two different filters do not inform about the absolute color index (Section 2.1.6) but only about its evolution.

### 2.2.3 Light curves

Light curves consist in the repeated measurement of the brightness of a source over time. They are commonly represented on a graph displaying flux/magnitude (or difference flux/magnitude) as a function of time, as illustrated by Figure 2.7. The latter represents observations acquired with different filters with different colors.

Light curves constitute the most essential tool of time domain photometry. They provide a record of the source's brightness variation, enabling the study of periodic and transient phenomena such as variable stars, exoplanet transits, or supernovae. Depending on the object studied, important physical information can be extracted from light curves. For example, the intrinsic luminosity of some periodic stars can be deduced from their periods (Leavitt and Pickering, 1912) ; an exoplanet transits carries information about its star's size and density (Seager and Mallén-Ornelas, 2003) ; and the mechanism at the origin of a supernova can be largely constrained by the exact rising and decaying pace of the transient (see Chapter 3). Although they are very informative, light curves are also challenging to work with, because they are:

1. **Multidimensional**. It is the result of the acquisition of data across multiple filters. In addition, the wide variety of existing photometric systems make the direct comparison of light curves acquired with different tools more complicated.

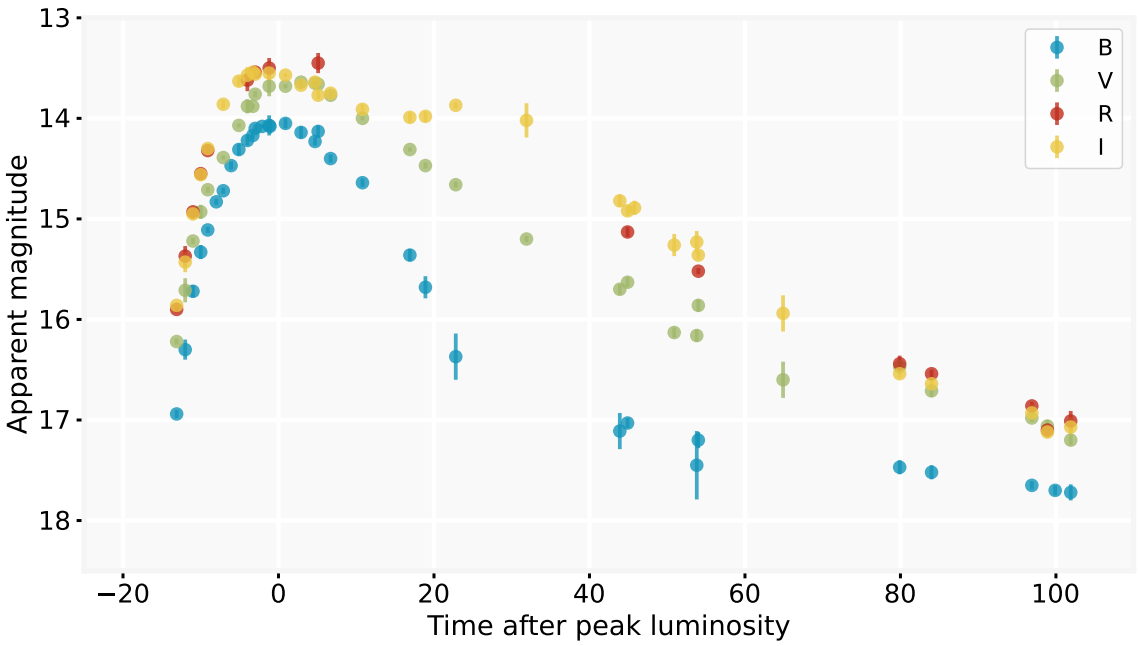



FIGURE 2.7 – Illustration of a light curve acquired over 120 days using $BVRI$ filters. The object is SN2002bo (Benetti et al., 2004), a supernova of type Ia (see section 3.1.2).

2. **Inhomogeneous**. The reality of the observing conditions implies that data points are never regularly acquired. Ground based telescopes in particular are subject to weather and seasons. Therefore, light curves often exhibit gaps (potentially long) with no measurements in addition to their uneven sampling frequency.
3. **Noisy**. Each data point comes with an error associated to the measurement, which must be considered to properly study the light curve. They are not fixed for all observations, but rather decrease with the source brightness, adding an extra layer of complexity in the manipulation of the data. Such uncertainty is unavoidable, as it includes at the minimum an irreducible Poisson noise (Andrae, 2010). In addition, external factors will degrade the observation quality. They include bad weather conditions, read-out noise, or measurement artifacts.

These difficulties are well understood by astrophysical experts who are used to deal with light curves. However, it is a much more challenging task for to use a machine to deal with them. Automatic analysis of light curves, whether it is purely algorithmic or uses machine learning, requires specific procedures to overcome these difficulties (Chapter 4).

## 2.3 Facilities

The invention of CCD (Section 2.2.1) marked the beginning of a new era. Since then, many time domain astronomical surveys such as the *Sloan Digital Sky Survey* (SDSS, York et al., 2000), the *Palomar Transient Facility* (PTF, Law et al., 2009), the *All-Sky Automated Survey for Supernovae* (ASAS-SN, Kochanek et al., 2017), or the *Asteroid Terrestrial-impact Last Alert System* (ATLAS, Tonry et al., 2018) have deeply changed our understanding of the Universe. Most of this thesis work is based on data and simulations related to two facilities, the Zwicky Transient Facility (Section 2.3.1) and the Vera C. Rubin Observatory Legacy Survey of Space

and Time (Section 2.3.2). In order to correctly manipulate astronomical data, it is crucial to understand how it was obtained. Therefore, this section presents these two surveys, from the telescope hardware to the real-time distribution system associated to them.

### 2.3.1 Zwicky Transient Facility

The Zwicky Transient Facility (ZTF) is a wide-field astronomical survey that aims at systematically mapping the optical night sky. All the technical details about the survey can be found in Bellm et al. (2019a) and at the official ZTF webpage [10]. It is based at the Palomar observatory, California, and uses a 48-inch Schmidt telescope. It is equipped with 3 optical filters ZTF-g (green), ZTF-r (red) and ZTF-i (infrared) for which the respective transmissions are represented on Figure 2.8. It covers all the visible light, with ZTF-i partly allowing to reach the near infrared emissions. However, only measurements acquired using ZTF-g and ZTF-r are part of the public survey and can be publicly accessed after each observation night [11].

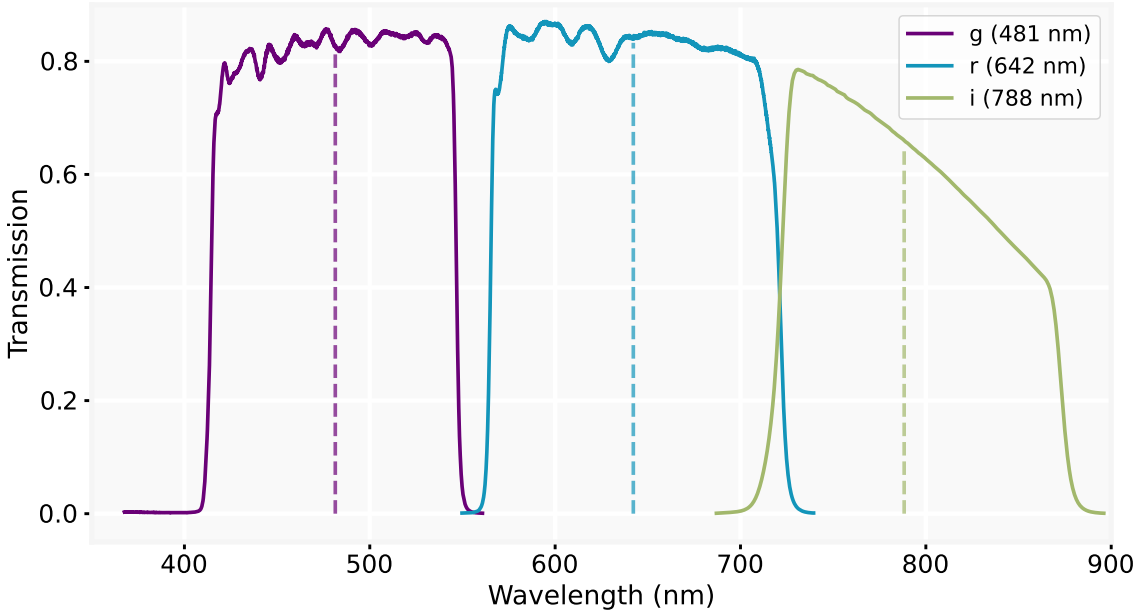



FIGURE 2.8 – Transmission of the ZTF-g, ZTF-r and ZTF-i filters from the `sncosmo` package (Barbary et al., 2016). Each profile is associated with its effective wavelength (vertical dashed lines), which corresponds to the mean wavelength of the detected photons as defined by Koornneef et al. (1986).

The camera is composed of 16 CCDs, covering a field of view of 47 square degrees (Dekany et al., 2020). In total, it has a resolution of 600 megapixels. This design gives to ZTF the widest angle of observation among all optical surveys currently in operation, which easily enables a complete picture of the northern sky within one night. Figure 2.9 illustrates the size of ZTF's field of view in comparison with other modern surveys. The Moon and the Andromeda galaxy are also shown for scale. The survey divides the sky into approximately 1800 fields [12]. Each night, it follows a predefined path across the fields and takes images with 30 seconds exposure time using

10. https://www.ztf.caltech.edu/index.html
11. ZTF-i observations are later available in the form of data releases.
12. https://github.com/ZwickyTransientFacility/ztf_information/blob/master/field_grid/ZTF_Fields.txt

the different filters. It delivers, on average [13], a cadence of one observation of the whole sky per passband every two days (Dekany et al., 2020). A single exposure provides a median limiting magnitude of 20.8 mag in g-band, 20.6 mag in r-band and 19.9 mag in i-band. It means that any source fainter than the limiting magnitude will not be resolved from the background with a signal-to-noise of at least $5\sigma$.

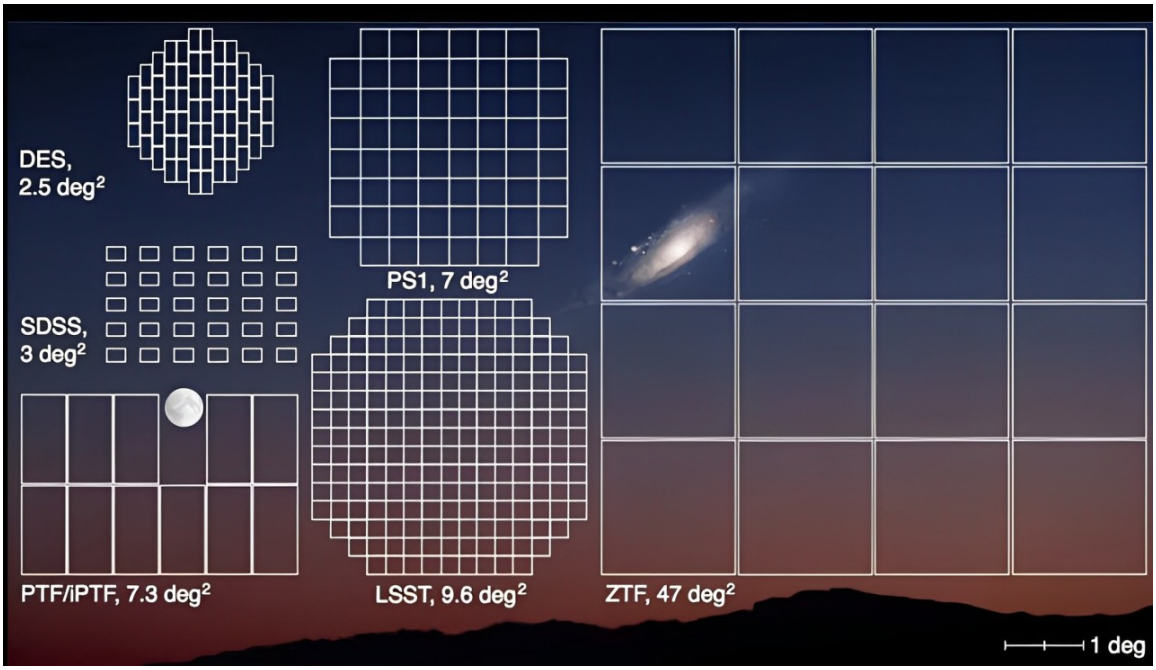



FIGURE 2.9 – Representation of ZTF field of view divided into the 16 CCDs. Other surveys are added for comparison. Original image from Laher et al. (2018).

ZTF has been scanning the sky and releasing data since 2018 and has produced more than 20 data releases (DRs). DRs provide calibrated magnitude for the sources observed, contrarily to the nightly available alerts, that only provide on-the-fly difference magnitudes (see Section 2.3.3). In all its years of observation, ZTF has gathered more than 50 million single-exposure images, which correspond to approximately 800 billion source detections extracted from those images. Once grouped based on their position, it results in the construction of almost 5 billion light curves. This volume of photometric data constitutes one of the largest optical time domain survey sample available, enclosing a huge opportunity for the systematical study of transient objects (detailed in Chapter 3).

### 2.3.2 Vera C. Rubin Observatory

The Vera C. Rubin Observatory [14], currently in late phase of construction, is a modern facility located on top of the Cerro Pachón mountain, in Chile. Its main science goal will be to conduct the Legacy Survey of Space and Time (LSST), a complete sky survey initially expected to operate for 10 years. Over many aspects, it represents the successor of ZTF. All characteristics

13. The telescope is subject to the weather and technical difficulties. We can never perfectly guaranty ground base observations.

14. https://rubinobservatory.org/

of the instrument presented in this section come from LSST (2024). The telescope, equipped with an 8.4 meter primary mirror, can move between two positions in less than 5 seconds. It will be equipped with 6 filters forming the *ugrizY* photometric system, and has a wavelength coverage ranging from 300 nm (near UV) to more than 1000 nm (infrared). Figure 2.10 presents their transmissions.

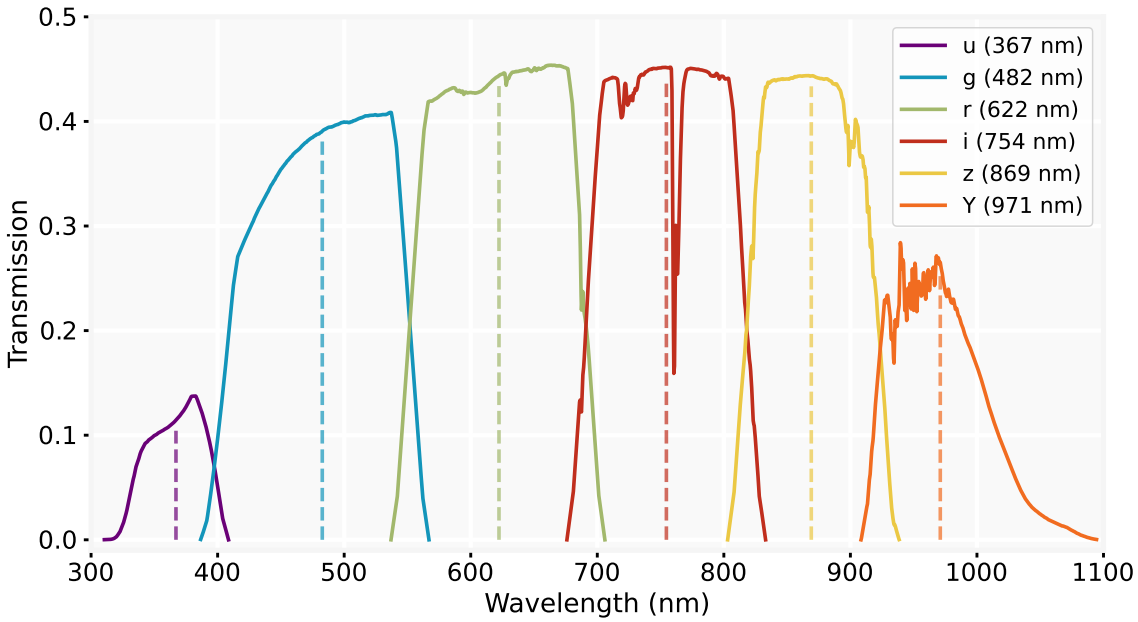



FIGURE 2.10 – Transmission of the LSST-u, LSST-g, LSST-r, LSST-i, LSST-z and LSST-Y filters from the `sncosmo` package (Barbary et al., 2016). Each profile is associated with its effective wavelength (vertical dashed lines), which corresponds to the mean wavelength of the detected photons as defined by Koornneef et al. (1986).

The field of view covered by one image is 9.6 square degrees. As illustrated in Figure 2.9, it represents a large area (about the size covered by 40 full Moons) but it is 5 times smaller than ZTF. Given an exposure time of 30 seconds per image, it will be able to take a picture of the entire visible austral sky every 3–4 nights. Most of the survey time will be divided into two main strategies, the Wide Fast Deep (WFD) and the Deep Drilling Fields (DDF). The WFD is the primary observing strategy, it will cover roughly 18,000 $deg^2$ during $\sim$ 95 % of LSST's operation time (SCOC, 2023). In this strategy, each patch of the sky should be visited on average once every few days. However, for some science cases this cadence is too sparse to allow meaningful data extraction, this is why the DDF has been proposed. It will take from 5 to 7 % of the observing time to scrutinize 5 small patches of the sky at a much higher cadence. Exact cadence numbers still remain to be defined (SCOC, 2023).

The camera used for the imaging is the largest CCD camera ever built. Weighting almost three tons and made of 189 CCDs, it has a total size of 3200 megapixels per image (five times more than ZTF). In addition, the large mirror of LSST enables the observation of the deep sky. Table 2.1 presents the limiting magnitudes per passband for a source detection at $5\sigma$. These values are up to +4 magnitudes when compared to ZTF (for the $g$ passband) meaning that LSST will be able to observe objects 40 times fainter.

| **LSST filter** | u | g | r | i | z | Y |
|---|---|---|---|---|---|---|
| **Limiting magnitude ($5\sigma$)** | 23.7 | 25.0 | 24.5 | 24.1 | 23.6 | 22.6 |

TABLE 2.1 – Theoretical limiting magnitude of LSST per filter for 30 seconds exposures.

Observing dimmer sources implies observing farther objects and therefore earlier into our Universe history. Since the volume of a sphere scales with the cube of its radius, this 4 magnitudes improvement will result in significantly more detected objects. In fact, combining this depth with the ultra-high resolution of the camera, we expect to detect a very large number of sources in each science image, and therefore proportionally more transients. Figure 2.11 illustrates this number by showing a composite image of three individual pictures acquired with different filters, generated by an end-to-end simulation of LSST. The already impressive amount of sources displayed only corresponds to a single CCD among 189.

FIGURE 2.11 – LSST picture of the sky from a single CCD as generated by an end-to-end simulation [15] . The image is a composition of three frames with different filters. It corresponds to a field of view of 13x13 arcminutes.

Such quantity of astronomical information per image will result in large volumes of data being produced daily. It is estimated that each night LSST will produce 20 terabytes of raw data, and therefore more than 70 petabytes will be collected over the 10 years of survey (LSST, 2024).

15. https://www.lsst.org/gallery/image-simulation-1

The difference image analysis will reveal around 10 million varying sources ($5\sigma$) per night, which is approximately ten times more than its predecessor ZTF (Bellm et al. 2020, Patterson et al. 2018). Simply storing, processing and distributing such a data volume to the community poses a great challenge by itself. This explains the need for broker systems, in charge of the real time distribution of the flow of data.

### 2.3.3 Astronomy brokers

A broker is a computing infrastructure in charge of ingesting, processing, and distributing the data collected each night by telescopes. LSST, in particular, will work in collaboration with brokers to provide public access to the stream of alerts. Because of the high bandwidth requirement, only seven official brokers [16] have been selected to receive in real time data from the survey (all brokers are geographically shown in Figure 2.12).


FIGURE 2.12 – Summary map of the official brokers selected to process LSST data stream.

LSST real time data distribution system has been designed similarly to the one proposed by ZTF. In that sense, only information about varying sources are provided in real time. Every time a $5\sigma$ difference imaging detection is registered, a small package of information called an *alert* is created. Every minute, alerts are sent to the brokers in batches. Note that this system implies that multiple alerts can be generated by a single source over time. Each alert contains the following data (Bellm et al., 2020):

— A unique number identifying the alert

— A number identifying the source [17].

— The measurement itself, which includes the difference flux and its associated error, the time of measurement, the filter used, as well as the strategy (WFD or DDF).

16. https://www.lsst.org/scientists/alert-brokers
17. A source is purely identified by its position.

— Past variability of the source, including $5\sigma$ detections and forced photometry, taken from the previous 12 months.

— Stamp images ($30 \times 30$ pixels[18] cutout around the source) of the difference image and template at the source location.

From this alert influx, brokers are expected to provide a variety of functions such as: redistributing the alert packets to the community, cross-correlating alerts with others catalogs, providing a user interface to explore the data, extracting and sharing meaningful information from the alerts, finding anomalous events or providing a classification for sources. Since the LSST alert system is expected to be analogous to ZTF, broker infrastructures are currently developing and testing tools around the ZTF alert pipeline. Despite the data volume being significantly smaller, this is a crucial step in the preparation for LSST.

This thesis work has been conducted as an active member of the Fink broker, therefore in what follows, it will be presented as an example of implementation for broker systems. Although they share a common goal, it is important to note that each broker offers distinct and complementary visions, scientific objectives, and implementations. Fink is a new infrastructure born in 2019 in response to a call released by the LSST collaboration. It is a community broker with a centralized processing, but a distributed scientific development strategy (Möller et al., 2021a). In that sense, the project born in France, has rapidly grown into an international collaboration. The core of Fink resides in optimally harvesting the interdisciplinary nature of the data stream, by allowing any expert team to join and implement a science module within the broker. These modules often take the form of algorithms or machine learning models that output valuable scientific information for the users. They, along with all the relevant information about each source, are accessible via the user-friendly open Science Portal developed by Fink[19] (Figure 2.13).

Currently, the Fink broker proposes a variety of science modules, including classification modules for supernovae (Möller and de Boissière, 2020b), early supernovae of type Ia (Leoni et al., 2022), microlensing events (Godines et al., 2019), kilonovae (Biswas et al., 2023) and hostless transients (Pessi et al., 2024), as well as a satellite tracking system (Karpov and Peloton, 2022). These models are systematically applied to new alerts and generate a probability of the event belonging to a certain class. This extra information is directly accessible for users, e.g. as a way to query only the fraction of the stream satisfying a given classifier. In addition to classifiers, Fink offers an anomaly detection module which assigns an anomaly score to alerts each night and can be used to explore rare/unknown astrophysical events (Pruzhinskaya et al., *in prep.*). Finally, Fink also includes a module to track and reconstruct the trajectories of solar system objects (Le Montagner, R. et al., 2023), which can either be used directly or as a way to remove contamination for extra solar system studies. The broker is in constant evolution and additional modules are under development, such a general classifier for fast transients or a classifier for early tidal disruption events. The latter constitutes an annex work of this thesis and is briefly presented in Appendix D.

18. Still to be defined, but $30 \times 30$ pixels is the minimum, as announced in Bellm et al. (2020)
19. https://fink-portal.org/
20. https://fink-portal.org/ZTF24aalubmp

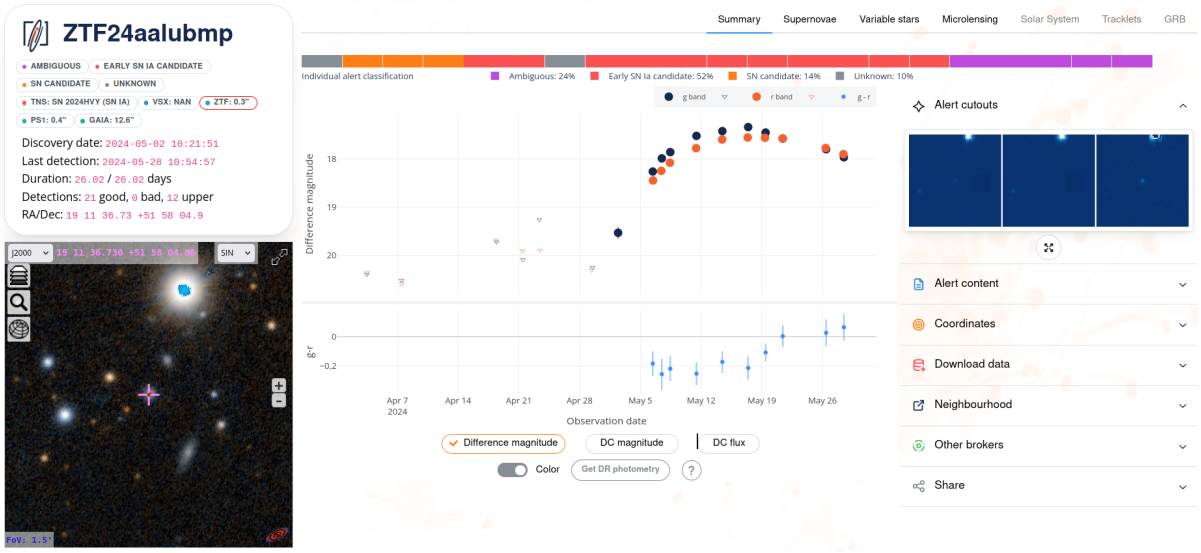



FIGURE 2.13 – Screenshot of the science portal web page for object ZTF24aalubmp [20]proposed by the Fink broker. The top left is a summary of the information of the alert, including ZTF name, classification from various actors, and key numbers about the alert's light curve. The bottom left is an interactive image of the sky at the transient position using the Aladin service (Bonnarel et al., 2000). The middle shows the difference light curve of the source in two passbands. It includes observations (circles) and upper limits from forced photometry (triangles), as well as the color index evolution displayed below. A calibration using the nearest reference source can be performed to obtain DC magnitude/flux, and if available, the complete history of variability can be queried from ZTF DR. The top color bar shows the evolution of the classification of the source by the internal Fink classifiers. The top right images are the template, science, and difference image cutouts. The bottom right menu offers extra information and hyperlinks to other services.

Although this preparation is core for the future LSST stream, it is not enough to prepare solely on ZTF data. It represents the closest survey to our disposal, but it still exhibits major differences beyond the volume of data itself. The cadence of LSST will be significantly lower, the number of filters will increase, covering a wider wavelength range, and the deeper limiting magnitude implies that the distribution of object types will be intrinsically different. For these reasons, LSST simulation data and challenges associated to them have been created. They represent an opportunity to understand the new specificities which will arise with the survey, as well as a way to adapt current tools to the arrival of its data.

### 2.3.4 LSST simulations

The Photometric LSST Astronomical Time-Series Classification Challenge (hereafter PLAsTiCC), was a Kaggle [21] classification challenge that took place in 2018 (Hložek et al., 2020). The goal was to understand how to build optimal machine learning classification models (see Chapter 4) to process LSST light curves. The complete simulated dataset consisted of ∼3.5 million light curves, representing 17 classes of sources that encompassed both transient and variable objects. The task was specifically difficult for machine learning models, since they were required to train with only ∼8000 light curves before being tested on millions. This high imbalance reflects the

21. https://www.kaggle.com/

real situation of astronomical data, where the volume of labeled data [22] is orders of magnitude inferior to that of available light curves.

Realistic light curves were generated using the SuperNova ANAlysis (SNANA) package (Kessler et al., 2009). Originally developed to study supernova of type Ia (Section 3.1.2), it has now become a standard tool in the generation of synthetic light curves, encompassing a large variety of variable sources. For each type of object, given template light curves and spectrum, along with some variable parameters [23], SNANA can generate additional simulated light curves. It has the advantage of incorporating the specificities of the survey such as the filters, cadence, atmospheric transparency, readout noise, and even irregular time interval due to weather conditions.

PLAsTiCC was a great success for the astrophysics community, and even if the challenge is over, the dataset has been (and still is) extensively used for research purposes (e.g. Ishida et al. 2021, Villar et al. 2021, or more recently Demianenko et al. 2023, Khakpash et al. 2024). Although it has provided the first step in the simulation of LSST data, PLAsTiCC is not enough to fully prepare for the future real data. Indeed, it only provides a fix dataset of complete light curves, rather than a stream of alerts. This detail is crucial because the classification of a transient must be performed as soon as possible to enable follow-up observations, rather than waiting for the event to be over. In practice, each new alert will correspond to one additional point on the light curve of a source, and every time, a new classification must be provided. Much more complicated scenarios emerge from this reality. For example, the need to classify a source based on a few detections, or changing the classification of an object based on a new unexpected alert.

In order to address these new challenges, but also for brokers to prepare their pipelines to the LSST alert format, the Extended LSST Astronomical Time-Series Classification Challenge (ELAsTiCC) was recently created (Narayan and ELAsTiCC Team, 2023). It is an extension of PLAsTiCC, meant to provide a rigorous assessment of the performance of different stages of the communication pipeline between the telescope and the brokers. This includes simulating the alert stream, the ingestion and analysis of the alerts by the broker teams, as well as reporting back the classification scores. The alert dataset has again been generated using SNANA. It mimics the equivalent of 3 years of LSST operations, containing 19 classes of transient sources and following an updated observation strategy [24]. In total, approximately 50 million alerts have been generated, corresponding to 3.5 million individual sources. Similarly to PLAsTiCC it first took the form of a competition, but this time restricted to brokers. The Fink team has accepted the challenge and presented the results in Fraga et al. (2024). Part of this thesis has been dedicated to the development of a Fink classification science module for super luminous supernovae (Section 3.1.3) within the ELAsTiCC dataset. It is briefly presented in Annex C.

22. Real light curves for which the classification was determined through spectroscopy.
23. Parameters that generates diversity among light curves in the population.
24. https://community.lsst.org/t/baseline-v3-2-released/7877

# 3

# Extragalactic transients & models

Transients are astronomical sources which appear and fade away over a relatively short period of time [1]. The history of their scientific study is directly linked to the progress of time domain astronomy (Section 2.2). As such, they constitute a recent science [2], and the tools at our disposal to understand them are in constant progress. The term transient encapsulates a wide variety of astrophysical phenomena. Some of them have a galactic origin, i.e. they are generated by objects inside the Milky Way, and would be too faint to be seen inside other galaxies. It is the case of novae (Wiescher et al., 1986) or flaring stars (Heyvaerts et al., 1977) for example. However, they are beyond the scope of this thesis, which focuses on the analysis of extragalactic transients.

In this chapter, we present different types of extragalactic transients, along with the scientific questions and interests related to them. Although they all were considered rare at some point in their history, the progress in the size of telescopes and the development of cameras (Section 2.2.1) considerably improved our understanding of these phenomena. By observing deeper and larger fractions of the Universe, we have been detecting an increasingly larger variety of transients, while also bridging the gaps between classes. Modern photometric facilities, such as ZTF (Section 2.3.1), enabled the discovery of tens of thousands of them, and the future Vera Rubin observatory (Section 2.3.2) will increase this number by more than an order of magnitude. It represents an unprecedented opportunity, but it also comes with challenges. Given that spectroscopy is a precious resource (Section 2.1.3), the nature of a transient and its properties must be identified as much as possible purely based on its light curve. To accomplish this, it is crucial to understand and categorize the photometric behavior of each type of transient.

This chapter details the specific brightness evolution of different transient phenomena. The list is not exhaustive, but covers the main types of objects that are, and will be, studied by ZTF (Section 2.3.1) and LSST (Section 2.3.2). Finally, a review of the different parametric models used in the literature to characterize them is presented in Section 3.4.

## 3.1 Supernovae

Historically, the word nova has been used as a general term to describe the appearance of a new star in the night sky. Despite being now known that it results from the explosion of a star - rather than it's creation - the nomenclature remained. Its usage has evolved during the last century to designated two distinct transient categories: novae and supernovae.

Novae are sudden and temporary large change in the magnitude of white dwarfs from accreting binary systems (Wiescher et al., 1986). This relatively faint process can only be observed within our own galaxy or in the nearest neighboring galaxies. On the other hand, supernovae are much brighter events which can be observed at cosmological distances. The word has been proposed for the first time by Baade and Zwicky (1934), and has since been used as a general term to describe the bright explosion of a star. A classification scheme proposed by Minkowski (1941) divides supernovae into two types, depending on the presence (type II) or absence (type I) of hydrogen lines in their spectra (Section 2.1.3). However, it still includes very diverse pheno-

1. Compared to the typical variability timescale of other celestial objects
2. Although few punctual observations of very bright transient happened in the last millennia (Section 2.2)

mena that can be further divided into finer classes. The main supernovae subtypes are presented below.

### 3.1.1 Core collapse supernovae

When a massive star ($> 8M_{\odot}$) reaches its final state of fusion, it produces heavy elements up to iron (Janka, 2012), from which no additional nuclear energy can be extracted from fusion. It leads to the contraction of the star due to the loss of thermal pressure necessary to counteract the gravitational force. When the mass in the non reacting core approximates the Chandrasekhar limit ($\approx$1.4 $M_{\odot}$, Chandrasekhar, 1931), i.e. the mass at which the electron degeneracy pressure becomes insufficient to compensate the gravitational pressure, it swiftly collapses within milliseconds in what is known as a core collapse event (CC). During this process, the core reaches extreme temperatures, which enables the merging of proton and electron and transforms the core into a proto-neutron star (or a black hole if $> 25M_{\odot}$, Heger et al., 2003), releasing an immense flux of neutrino $\sim 10^{53} erg/s$ (Barwick et al., 2004). The nuclear repulsive forces suddenly halt the collapse, creating an outward pressure shock wave and ejecting matter from the outer layer of the star. Approximately 1% of the neutrino energy is reabsorbed [3] by the outer layer in the form of kinetic and thermal energy, ionizing the photo-sphere and triggering the supernova itself. The heated and opaque outer layer shines across a broad range of wavelengths, dissipating its thermal energy over time, as the recombination occurs.

The core collapse mechanism can occur within a diverse range of progenitors, generating various types of supernovae. Bellow we give further details regarding type II (Section 3.1.1.1), type Ib and Ic (Section 3.1.1.2) and superluminous SN (Section 3.1.3).

#### 3.1.1.1 Supernova II

Supernovae type II (hereafter SNII) are the result of the core-collapse of massive stars ($8M_{\odot} \lesssim M \lesssim 18M_{\odot}$, Smartt et al., 2009) that present an outermost hydrogen layer. Therefore, they are spectroscopically recognized by the presence of Balmer lines, the characteristic absorption lines of hydrogen. A more detailed study of their spectra indicates the existence of two spectroscopic subtypes. If it includes narrow hydrogen emission lines, the supernova belongs to the SNIIn (for narrow) type. This behavior is the result of the interaction with a circumstellar material of high density, suggesting a very massive progenitor star which ejected a fraction of its mass, resulting in a particularly luminous supernova (Kiewe et al., 2012). They represent the brightest subtype of SNII with an average absolute magnitude in the B-band of $\sim -18.5$ mag (Richardson et al., 2014). The most extreme SNIIn stands at the border between super luminous supernovae and classical supernovae (see Section 3.1.3).

Some spectra present very weak hydrogen lines, which completely disappear $\sim$ 10 days after the explosion. We associate this behavior with the SNIIb type (Filippenko et al., 1993). This might indicate that the hydrogen layer of the star was already mostly expelled from its outer layers prior to the supernova (Hoflich et al., 1993, Podsiadlowski et al., 1993). It results in a second brightness peak in the light curve when the hot helium layer is revealed. Figure 3.1 shows an example of double peaked SNIIb (SN2011fu). Spectra acquired past the re-brightening phase

3. The exact mechanism of this process is still not completely understood.

shows strong Helium lines, similarly to SNIb types (see Section 3.1.1.2).

Photometrically, SNII display a fast brightness increase due to the shock breakout, with a typical rise time of $\sim$ 15 days before reaching the peak luminosity. We then distinguish two subtypes based on their decaying behavior, as illustrated in Figure 3.1. Some present a long plateau phase of $\sim$ 100 days (Kou et al., 2020) powered by the recombination of the opaque hydrogen photosphere. During this period, the luminosity of the object stays roughly constant. These are classified as SNIIP (for plateau). Once the recombination process is over, a second declining rate can be observed from the remaining cooling process. Other SNII without plateau display directly the linearly decaying phase, they are classified as SNIIL (for linear). SNIIL typically consist of a constant dimming at a typical decay rate of $\sim$ 0.02 mag/day (Kou et al., 2020). Furthermore, SNIIP and SNIIL are characterized by different absolute peak luminosities, with a typical absolute magnitude in the B-band of $\sim -18.0$ mag and $\sim -16.75$ mag, respectively (Richardson et al., 2014). In reality, the separation between a linear and a plateau behavior is not always so clear, and we observe a continuum of light curves at the border of both types that exhibit short plateau phases (Anderson et al., 2014, Sanders et al., 2015).

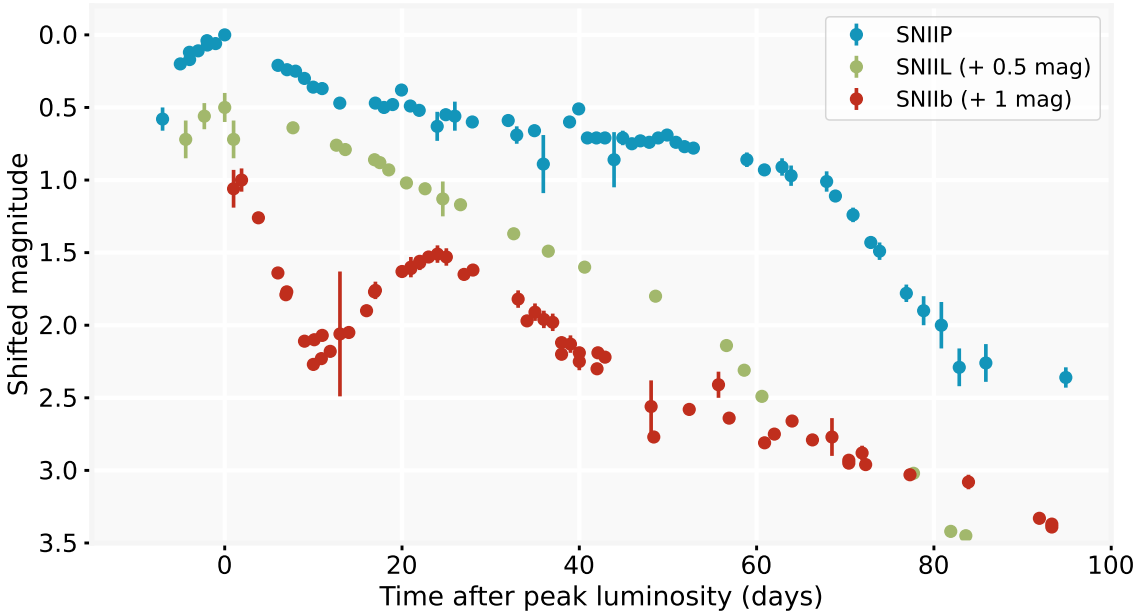



FIGURE 3.1 – Example of SNII light curves observed in the R-filter. In blue, a SNIIP (SN2013fs, Bullivant et al., 2018). In green, a SNIIL (SN1998S, Fassia et al., 2000). In red, a SNIIb (SN2011fu, Morales-Garoffolo et al., 2015). For clarity, magnitudes have been shifted so the observed peaks are at 0 mag (SNIIP), 0.5 mag (SNIIL), and 1 mag (SNIIb).

The study of the metal line in the spectra of SNII has revealed a clear dependence with the metallicity of the environment. This property makes them particularly interesting to measure, as they can be used for probing the abundances of elements in the Universe (Anderson, J. P. et al., 2016). Furthermore, similarly to SNIa (see Section 3.1.2), methods have been developed to standardize their light curve and use them for distance evaluation (Hamuy and Pinto, 2002). However, this approach is challenging due to the large range of variability in their light curve profiles.

### 3.1.1.2 Supernovae Ib & Ic

Supernovae of type Ib and Ic (hereafter SNIb and SNIc) result from the core-collapse of massive stars that have been shedded from their outer layer (Filippenko, 1997) ; they are also referred to as stripped core-collapse supernovae. SNIb are produced by stars that have completely or almost completely lost their hydrogen, thus revealing the helium layer. Therefore, their spectra exhibit strong helium lines and no signature of hydrogen. SNIc come from stars which striped even further their envelope, leaving little to no helium in the photosphere. As of today, the precise mechanism responsible for the stripping of the envelope is still debated among experts. It could be caused by the interaction with a companion in a binary system, but also by strong stellar winds for single massive stars (Maeder, 1996, Woosley et al., 1993).

SNIb and SNIc are often grouped together into the SNIb/c type, as they share many common properties and the frontier which separates them can be blurry. They typically rise in brightness in 2 to 4 weeks, reaching on average an absolute magnitude in the B-band maximum of $\sim -17.0\,\mathrm{mag}$ and $\sim -17.5\,\mathrm{mag}$ for SNIb and SNIc respectively (Richardson et al., 2014). The decaying part is purely driven by the cooling process and last from 2 to 4 months. Due to the absence of the outermost layers, no recombination of hydrogen can happen, excluding any plateau phase. Figure 3.2 displays an example of a SNIc light curve observed by ZTF using two filters. The cooling through thermal radiation is well illustrated by the reddening (the color $g-r$ increases) of the supernova during its dimming.

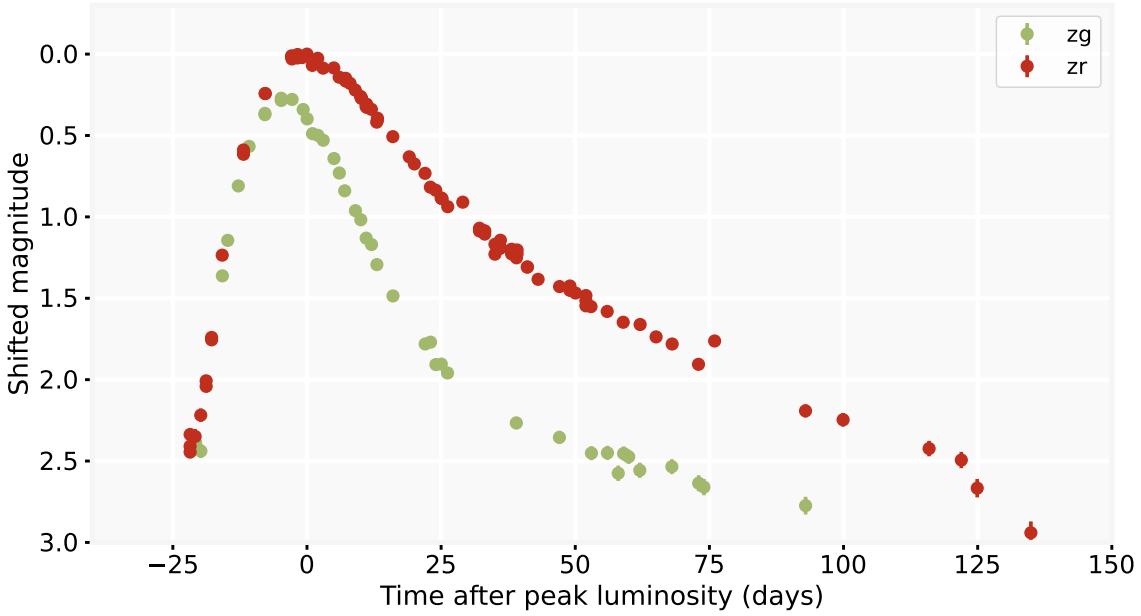



FIGURE 3.2 – Typical light curves for a SNIb/c (SN2019odp, Schweyer et al., 2023) observed by the ZTF telescope in the ZTF-$g$ and ZTF-$r$ filters. Magnitudes have been shifted for so observed peak to is at 0 mag for the ZTF-$r$ filter.

The identification and study of a larger sample of SNIb/c would help constraining the precise mechanism at the origin of the outer layer stripping. It is likely that part of the ejected layer condenses into various chemical species, and it could be a major source of dust production (Sarangi et al., 2018). In addition, light curves from SNIb/c and SNIa (see Section 3.1.2) may

look very similar. The latter, are famously used to measure the expansion rate of the Universe. In this context, SNIb/c constitutes an important contaminant for cosmological studies.

### 3.1.2 Supernovae Ia

Thermonuclear supernovae, also known as supernovae type Ia (hereafter SNIa) are fundamentally different from core-collapse supernovae. Indeed, their main source of luminosity comes from the decay of heavy radioactive elements. They occur in binary systems with a white dwarf orbiting a companion star. Although the question is still debated, it is believed that, from this state, two main paths can lead to a SNIa:

— In the so-called single degenerate scenario, the gradient of the gravitational field of the compact white dwarf is high enough to accrete material from its companion star. The mass gained via this process slowly drives the star towards the Chandrasekhar limit of $\sim 1.4 M_{\odot}$(Whelan and Iben, 1973).

— In the double degenerate scenario, both stars orbiting each other are white dwarfs. The energy is lost in the form of gravitational waves, bring them closer until they merge. The resulting mass of the system will ultimately reach the Chandrasekhar mass (Iben and Tutukov, 1984, Webbink, 1984).

Independently of how the mass increases, the process finishes in the same way. The carbon fusion of degenerate matter ignites a deflagration which triggers the supernovae. The exact ignition mechanism is not completely understood and remains an active research topic (Poludnenko et al., 2019). The quantity of energy released in this short amount of time is enough to completely unbind the star in a supernova event, ejecting the material with a speed of $\sim 10^4$ km/s, most likely leaving no remnant behind (Mazzali et al., 2007). During the thermonuclear process, from 0.2 to $\sim 1 M_{\odot}$ of $^{56}$Ni is produced (Nomoto et al., 1984), which is responsible for the very high luminosity of the SNIa events [4]. Indeed, the $^{56}$Ni will decay and emit thermal radiation following the mechanism:

$$^{56}\mathbf{Ni} \rightarrow {}^{56}\mathbf{Co} \rightarrow {}^{56}\mathbf{Fe}. \tag{3.1}$$

The $^{56}$Ni half life is ∼6 days, powering strongly the supernovae for a few weeks after the peak brightness. Its short lifetime results in a quick exponential decrease in brightness during this phase. Once most of the $^{56}$Ni has decayed, the radioactive $^{56}$Co becomes the main source of energy. Due to its half life of ∼77 days, we then observe a sudden change in the brightness exponentially decaying slope, which becomes slower.

Since SNIa progenitors are white dwarfs, they are easily identifiable spectroscopically. Indeed, they are composed primarily of carbon and oxygen, therefore SNIa exhibit no hydrogen lines, as signified by the type I. Additionally, given the white dwarf's composition, the fusion process creates silicon which constitutes the signature absorption line of SNIa spectra. Their absolute magnitude in the B-band typically reaches $\sim -19.3$ mag, which is significantly more luminous than the standard core collapse supernovae, and can even outshine their entire host galaxy. This property enables their observation at far distances, scouting deep into the Universe history. Most of the energy is emitted in the visible spectrum, producing a light curve composed of a

4. This process also happens within CCSN, but to a much smaller extent because significantly less $^{56}$Ni is produced. Hence, the decay of $^{56}$Ni is a secondary source of luminosity for CCSN.

rising phase of $\sim$ 20 days followed by a dimming period of $\sim$ 100 days exhibiting a double slope characteristic of the nickel decay. However, in the red to near infrared spectrum, SNIa light curves present a second peak of luminosity approximately 20 to 30 days after the first maximum brightness. The exact physical process behind the phenomenon is not completely understood. Kasen (2006) argues that the timing and strength of the rebrightening can be explained by the ionization evolution of the iron-peak elements in the ejecta. Figure 3.3 displays a typical SNIa light curve observed in blue (B) and near infrared (I) filters.

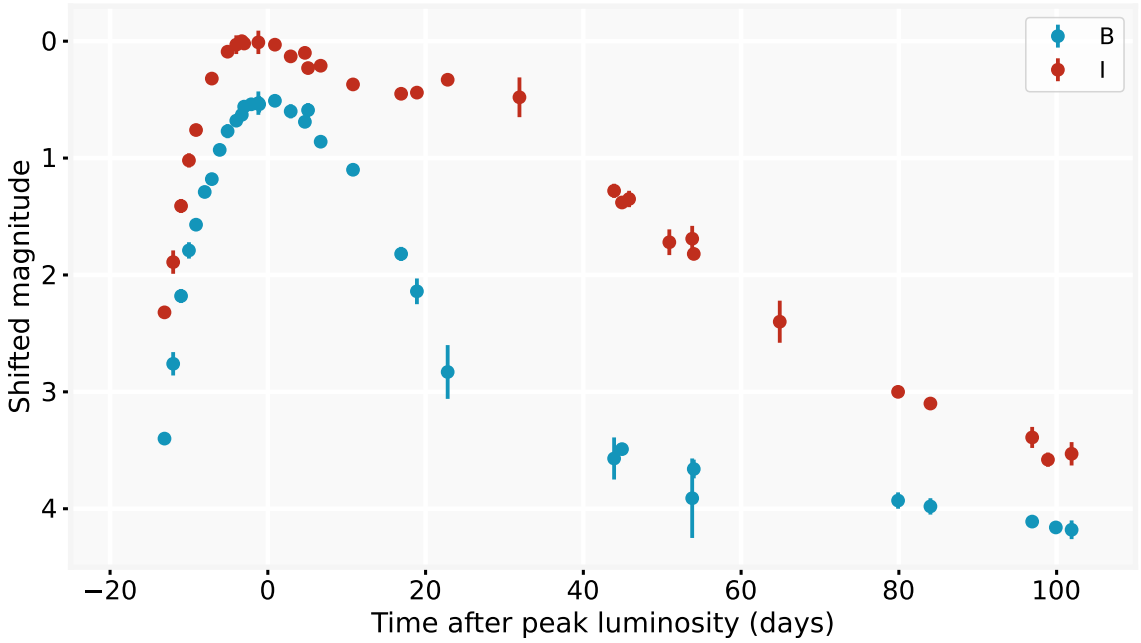



FIGURE 3.3 – Typical light curves for a SNIa (SN2002bo, Benetti et al. 2004) observed in $B$ and $I$ filters. Magnitudes have been shifted so the peak in I is shown at 0 mag.

SNIa events are particularly known for being very similar to each other, both from a photometric and from a spectroscopic aspect. Indeed, since the mechanism that triggers them systematically occurs at a similar mass, we expect their luminosity to always be approximately the same. Thus, they are known to be standardizable candles (Pskovskii, 1977, Rust, 1974). This property enables the evaluation of cosmological distances and has been used as a tool to understand the expansion of the Universe (Riess et al. 1998, Perlmutter et al. 1999).

However, their detailed study has revealed rare SNIa events that behave differently from the homogeneous main sample, they are referred to as peculiar SNIa. To this day, according to the Transient Name Server [5] (TNS), spectroscopic classification, approximately 7% of the $\sim 10^4$ SNIa reported fall under this category. They are further divided into subtypes. For example, the analysis of the supernova SN1991bg (Filippenko et al., 1992a) has lead to the discovery of multiple SNIa that present similar characteristics such as being under-luminous, redder, faster and missing the second peak in the infrared. These objects are commonly classified as SNIa-91bg-like. Similarly, the study of SN1991T (Filippenko et al., 1992b) revealed the existence of longer and brighter than usual SNIa, called SNIa-91T-like. Finally, the discovery of SN 2002cx and similar events lead to a new subclassification among SNIa, the SNIax type (Foley et al.,

5. https://www.wis-tns.org/

2013). Compared to classical SNIa, they are on average bluer, much less bright, show a faster rise time, and they do not present a second maximum in the infrared (Jha et al., 2017).

Many questions are still unsolved regarding SNIa. Understanding the cause of their variability and properly modelling them is essential. The question of their progenitor is still unsolved, and we have no definitive proof for which scenario leads to the explosion, or in which proportion they both do (Rebassa-Mansergas et al., 2019). Moreover, results from using SNIa to measure the rate of expansion of the Universe ($H_0$) are in strong contradiction with estimation made with the cosmic background measurement (Planck Collaboration et al., 2020b) in what is known as the Hubble tension. This tension, also known as the crisis in cosmology, constitutes one of the main question to be solved. It could be explained by hidden physics to be understood, or, as suggested by the newest JWST analysis (Freedman et al., 2024), it could be the result of flaws in calibration procedures. These questions make the observation and classification of large sample of SNIa one of the driving science cases of current and future surveys.

### 3.1.3 Superluminous Supernovae

Superluminous supernovae (here after SLSN) are supernovae whose peak optical absolute magnitude exceeds $\sim -21$ mag in any band (Moriya et al., 2018). They are generally long and slow evolving transients, lasting from hundreds of days to years, with a wide variety of rising and decaying behaviors. Figure 3.4 shows two SLSN exhibiting a fast and a slow evolution.

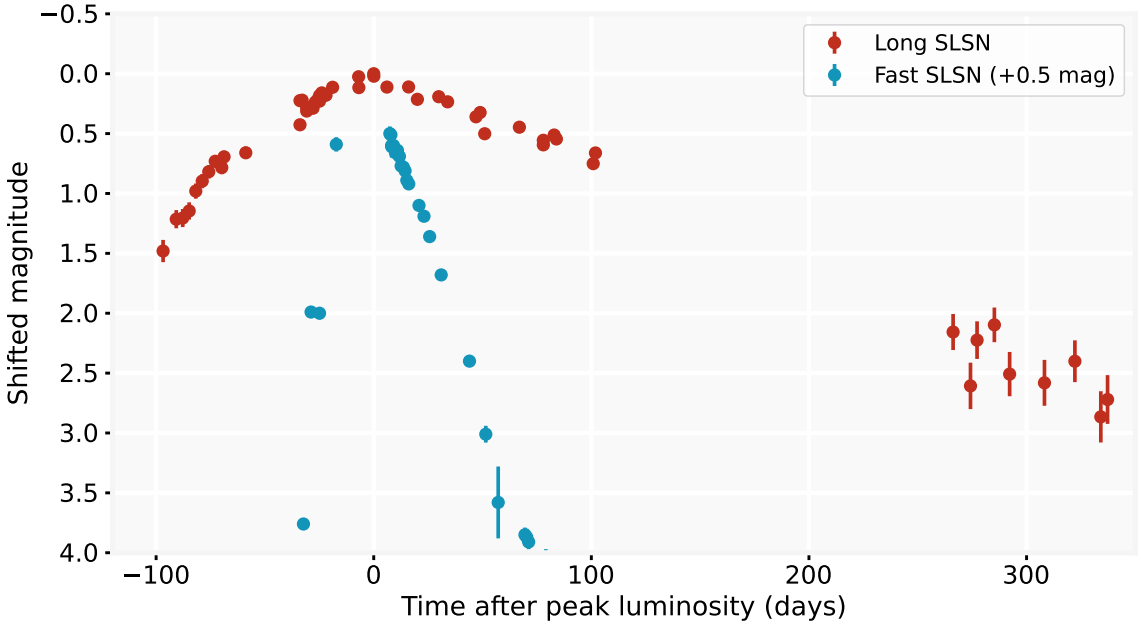



FIGURE 3.4 – Examples of short (SN2011ke, Nicholl et al., 2015, $M = -21.4$ mag) and long (SN2018ibb, Schulze, Steve et al., 2024) light curves of SLSN observed with an $r$ filter. Magnitudes have been shifted so the observed peaks are at 0 mag (Long SLSN), and 1 mag (Fast SLSN). The fast evolving object can only be explained by the magnetar scenario, while the slow one is compatible with the PISN scenario.

Explaining the exact physical process hidden behind the light curve is very challenging. Indeed, they constitute a rare class of events. For example, their rate of occurrence ($z \leq 1$) is $\sim 100$ times smaller than SNIa (Curtin et al., 2019, Desai et al., 2024). Similarly to classical supernovae, they

can be subdivided into type I and type II, depending on the absence or presence of hydrogen. However, this classification is largely insufficient to explain the variety of observed light curves. In particular, the crucial step is to understand the source of additional energy enabling the extreme brightness of these events. Three main mechanisms are invoked to elucidate this question:

1. **Circumstellar material interaction:** Some SLSN show unambiguous signs of strong circumstellar interactions. Such medium, found around core collapse SNe, decelerates the ejecta from the supernova and converts part of its kinetic energy into thermal radiation. This process is recognizable by the presence of narrow hydrogen lines caused by the interaction with the hydrogen rich gas (Fraser, 2020). Therefore, they belong to the SLSN-IIn subtype and could be extreme cases of SNIIn (Moriya et al., 2018) with the excess of luminosity being explained by the higher density of the surrounding medium. It is still not completely clear whether SLSN-IIn and SNIIn form a continuum of the same nature or belong to two separate classes originating from distinct mechanisms. While they all share common properties in terms of brightness, the exact composition of the circumstellar medium introduces a lot of variety in light curve behavior. They typically have a peak absolute luminosity of $\sim -21$ mag and can go up to $\sim -23$ mag (the brightest SLSN-IIn ever observed in the $V$ passband, Drake et al., 2011). The decline duration of the transient can range from 100 days (e.g., SN2008fz, Drake et al., 2010) to multiple years (e.g., SN2003ma, Rest et al., 2011).
2. **Magnetar powered:** The core collapse of a star leads to the formation of a neutron star (or a black hole). In rare cases, a magnetar – a subtype of neutron star presenting an unusually intense magnetic field – can be formed. Depending on initial parameters, such magnetar can act a central engine by transferring its spin into electromagnetic radiation (Dessart et al., 2012). The magnetar model can explain the energy excess of some SLSN given specific choices for spin period, magnetic field strength, and ejecta mass, in particular considering events for which neither the $^{56}$Ni decay nor the circumstellar interaction provides a physical answer. This mechanism can generate both SLSN-I and SLSN-II (Dessart, Luc and Audit, Edouard, 2018).
3. **High $^{56}$Ni mass:** The intense luminosity observed in some SLSN could simply be explained by an unusually large production of $^{56}$Ni (at least $10M_\odot$, Moriya et al., 2018). However, SNIa only generate $\sim 1M_\odot$ and classical core collapse SNe are generally inefficient at producing $^{56}$Ni. A mechanism proposed by Rakavy and Shaviv (1967) can lead to the production of a gigantic mass of radioactive material. For extremely massive stars (140 to $260M_\odot$), once helium has been exhausted in the core, the temperature and density are such that photons are converted to electron-positron pairs. This pair production mechanism results in a loss of radiative pressure support in the core, creating an instability which eventually leads to the collapse of the core and subsequent thermonuclear runaway. This type of supernova is called a pair instability supernova (hereafter PISN). The progenitor's mass required is such that we expect PISN to occur only among the first population of stars in the Universe history [6], which can reach 100 to $300M_\odot$ (Heger and Woosley, 2002). The synthesized $^{56}$Ni mass depends on the mass of the helium core of the progenitor, and can range from $10^{-3}M_\odot$ to more than $40M_\odot$. Hence, not all PISN are SLSN. The high ejecta mass results in long photon diffusion timescales and hence long rise times, up to $\sim$200 days, although some models can accommodate relatively short rise times (Kozyreva

6. Also known as population III stars.

et al., 2017). However, PISN models cannot explain rapidly-declining SLSN light curves, for which the magnetar model is favored. In principle, PISN could generate both SLSN-I and SLSN-II light curves. To this day, there is no confirmed observation of PISN, but promising candidates have been found (Pruzhinskaya et al., 2022, Schulze, Steve et al., 2024).

SLSN are currently considered rare transients, with TNS[5] listing 229 spectroscopically confirmed events. The discovery and classification of any additional event can significantly impact our comprehension of the physical mechanisms behind this population. In particular, confirmed discovery of a PISN would largely impact our understanding of the connection between the structure formation and the chemical evolution of the Universe. Furthermore, it has been shown that, beyond physical information on their own progenitors, SLSN-I constitute excellent candidates to be used as standard candles (Inserra and Smartt, 2014). However, significantly more statistics is required to make precise evaluation of cosmological parameters. Therefore, classifying and characterizing more SLSN will be a crucial task in the future.

## 3.2 Tidal Disruption Events

The density and mass of the supermassive black holes at the center of galaxies is so extreme that objects orbiting near them experience intense tidal forces. This effect implies that there is a large difference in the gravitational field pulling on two points of the same object, effectively deforming it. In some cases, a star orbiting a supermassive black hole could endure so strong tidal effects that it would entirely rip the star apart in what is called a tidal disruption event (hereafter TDE, Lacy et al., 1982). Approximately half of the star's material will spiral down the black hole, while the other half will be ejected away from it. In some special cases, the star could be only partially shredded and continue its orbit to be disrupted again next time it approaches the black hole, this is known as a repeated, or partial, TDE (Lin et al., 2024). The matter spiraling down will form an accretion disk that will temporarily thermally radiate very intensely. However, the TDE spectrum of emission does not exactly match the distribution expected from a pure accretion process. We observe discrepancies in the visible radiations as well as additional X-ray emissions (Gezari, 2021). Understanding the exact processes at play constitutes an active field of research.

The light curve of a TDE displays an increase in brightness over two to four weeks to reach an absolute magnitude from $-19$ to $-21$ mag (in the $r$ passband), sometime comparable with the luminosity of SLSN (Gezari, 2021). The dimming follows a power law, of $t^{-\frac{5}{3}}$ which can be deduced from the dynamics of the fallback of the debris stream. However, partial TDEs should follow a steeper decline following the power law $t^{-\frac{9}{4}}$ (Coughlin and Nixon, 2019). Figure 3.5 shows the light curve of a typical TDE. In general, TDE are long and slow evolving events when compared to classical SNe. Their emission spectrum is generally largely compatible with a high temperature blackbody ($10^4 K$) and therefore they emit mainly blue light.

As of today, TDE constitutes a very rare class of objects, with less than 50 events observed in the visible light and spectroscopically confirmed (Hammerstein et al., 2022). Characterizing more events would give valuable information to understand the origin of the X-ray emission processes and would contribute to the exploration of the diversity in the TDE behavior. Once

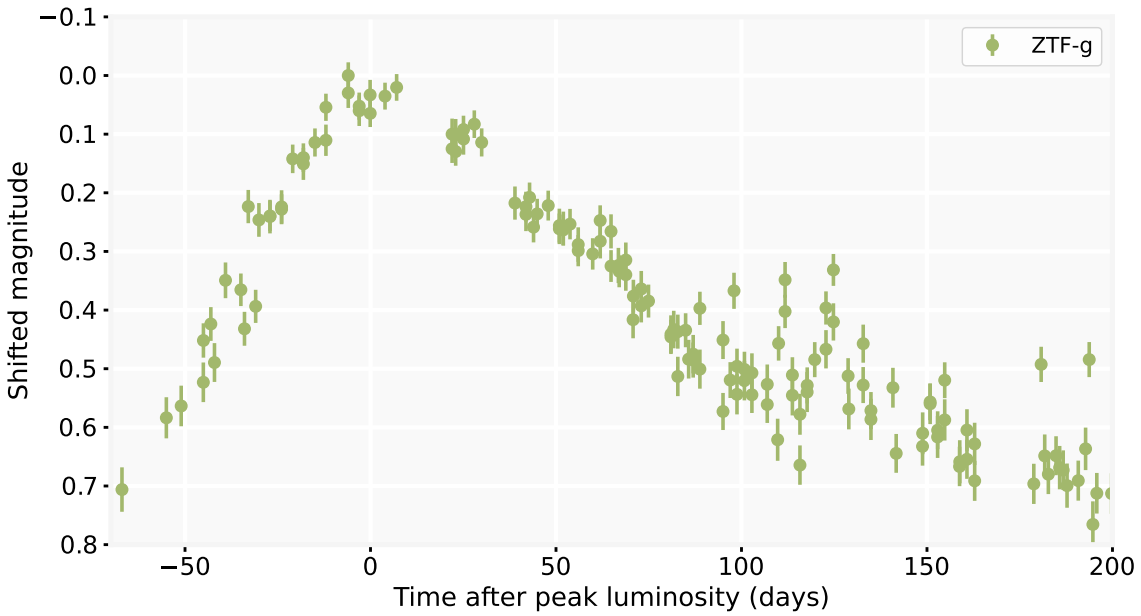



FIGURE 3.5 – Examples of a TDE (AT2020mot, Newsome et al., 2024) observed by ZTF with the $g$ filter. A redshift of 0.07 has been spectroscopically measured.

a larger population is identified, they could also be used to probe the demographics of the supermassive black holes while improving our comprehension of accreting matter around them.

## 3.3 Active Galactic Nuclei

The centers of active galaxies are extremely bright and constantly evolving regions, they are called active galactic nuclei (hereafter AGN). Their luminosity come from the accretion of matter around the supermassive black hole at the center of the galaxy. Indeed, matter falling into the black hole have considerable angular momentum, thus spiraling around it and forming an accretion disk [7]. The turbulent motion, friction, and the presence of magnetic fields in the disk causes energy to be radiated outward (Shakura and Sunyaev, 1973), eventually leading the material to fall inside the black hole. A significant fraction of this gravitational energy is released in the form of thermal radiation. Because of the high temperature of the disk ($> 10^5$ and up to $10^6$ K), these emissions are mostly X-rays. However, unlike stars and supernovae, AGN cannot be approximated as blackbody emitters. Non-thermal processes, primarily inverse Compton and synchrotron radiation (Torricelli-Ciamponi, G. et al., 2005), are invoked to explain the observed spectrum from AGNs, which also produce visible, infrared and sometime radio emissions. Overall, AGNs are extremely luminous objects that can frequently outshine their whole galaxy. Their absolute magnitude typically ranges from -18 to -24 (in the $i$ passband, Chanchaiworawit and Sarajedini, 2024), and can reach up to -30 mag (in the HK passband, see e.g. Wolf et al., 2018, , the brightest AGN ever observed). Understanding the exact origin of such inflow of matter and modeling the relationship between the AGNs and their galaxies are major topic of research in astrophysics. They are also paramount for the study of accretion and photoionization physics

7. The actual structure of an AGN is much more complex than this simple representation. For example, it includes a torus and jets. However, these considerations are beyond the scope of this manuscript, for which only their visible photometric behavior is relevant.

(Trakhtenbrot et al., 2019), they can trace star formation regions (Masoura et al., 2018) and have the potential to enrich cosmological studies (Martínez-Aldama et al., 2019).

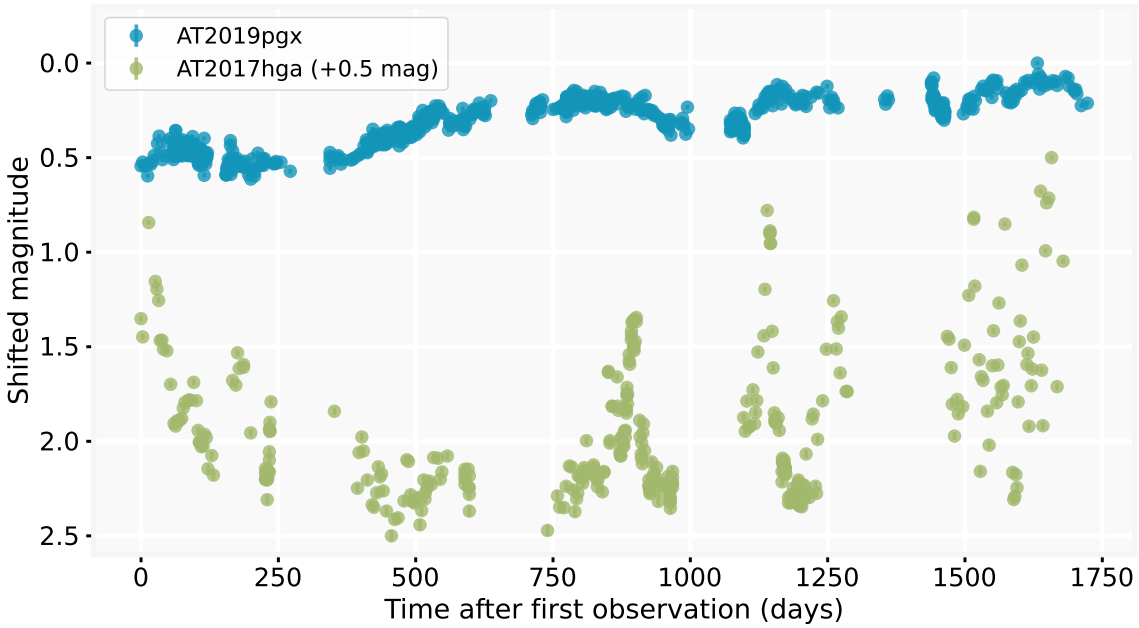



FIGURE 3.6 – Light curves of two distinct AGNs: AT2017hga (Blagorodnova, 2017) and AT2019pgx (Chu et al., 2021), observed by ZTF with the r filter. Magnitudes have been shifted so the observed peaks are shown at 0 mag (AT2019pgx), and 1 mag (AT2017hga).

The luminosity of AGNs varies continuously over many years. This variation is unpredictable because of the complex nature of the accretion processes. Therefore, they are not considered transient but rather stochastic sources. Figure 3.6 displays the light curves of two distinct AGNs [8, 9] and illustrates the slow evolving and chaotic nature of their variability. Despite not being transients, AGNs have been added to this section because they constitute an important source of contamination in the photometric study of transients (e.g. in Hung et al., 2018). The randomness in their evolution, coupled with characteristics of a given observing strategy, can produce light curves that, at a given time, are virtually indistinguishable from those of transients. A sudden burst of luminosity in a distant and otherwise undetectable AGN can also create transient like behaviors to the observer.

## 3.4 Parametric models

Different astrophysical phenomena exhibit different light curves, from which we must extract as much scientifically valuable information as possible. Although each type of transients has its own specific behavior, they can all be described roughly in the same way: they appear, get bright and fade away forever. What is left to determine for each individual source is the rate and the intensity at which each step happens, as well as the presence of potential secondary behaviors (e.g. a second bump or a plateau ...). Parametric model fitting is commonly used as a solution

8. https://ztf.snad.space/dr17/view/1487203100010564
9. https://ztf.snad.space/dr17/view/737202100043140

to individually describe transient light curves. Ideally, the specific choice of the parametric function should encode our expectations regarding the general properties of the specific class under consideration. In this context, identifying the best-fit parameters, which make the connection between the parametric function and the observed light curve, is one of the possible methods for systematically extracting information from photometric observations.

Proper light curve fitting offers very valuable physical information that can be used to better characterize a transient (e.g., Angus et al., 2019, McCully et al., 2022), to interpolate missing observation (Alves et al., 2022), to predict its future behavior (Woosley et al., 1988) or to extract numerical features (e.g., Bloom et al., 2012, Lochner et al., 2016). The latter is of particular interest for machine learning analysis. Indeed, the values of the best fitting parameters, along with a metric which evaluates the quality of the fit, can be used as summarizing features. Feature extraction based on light curve fitting is presented in Section 4.1.

Multiple functional forms have been proposed and are currently used to describe transients. They are all designed to fit light curves expressed in flux, and with a baseline equal to zero. A simple additive parameter can be added to account for non-null baseline. This section presents a list of the most common parametric functions found in the literature, in the context of astronomical transient description.

### 3.4.1 Bazin function

Bazin et al. (2009) proposed a function that is commonly used as an all-purpose transient function. It has originally been proposed as a phenomenological form to describe SNIa (Section 3.1.2) and core-collapse SN (Section 3.1.1) but it is general enough to be applied to other transient events. According to this approach, a light curve in a given filter is described as

$$f(t;\ t_0, A, \tau_{\text{rise}}, \tau_{\text{fall}}) = A \times \frac{e^{\frac{-(t-t_0)}{\tau_{\text{fall}}}}}{1+e^{\frac{t-t_0}{\tau_{\text{rise}}}}}. \tag{3.2}$$

This formulation contains four free parameters: $\tau_{\text{rise}}$ and $\tau_{\text{fall}}$ describe the rising and declining timescales; $t_0$ is a reference time, which is related to the time of peak brightness such that $t_{\text{peak}} = t_0 + \tau_{\text{rise}} \times \ln(\tau_{\text{fall}}/\tau_{\text{rise}} - 1)$, and $A$ is a multiplicative amplitude parameter. The effect of the variation of all parameters is represented in Figure 3.7.

The Bazin function describes an exponential rising phase, followed by an exponential decay. This simple form is enough to describe the general trend of most transients, which explains its wide usage in the literature. For example, Ishida et al. (2019) used it to fit supernovae in the context of active learning, and showed its potential in the optimization of spectroscopic follow-up. McCully et al. (2022) employed it to compare properties of light curves from SN2012Z spanning a decade of observations; Muthukrishna (2022) used its resulting goodness of fit as a proxy for anomaly detection; Kelly et al. (2023) used it to model and calculate time delays from five different light curves from the same strong-lens SN; Corsi et al. (2023) applied it to model broad-line SNIc, and Fulton et al. (2023) used it to obtain a continuous extrapolation for the SN light curve associated with GRB21009A.

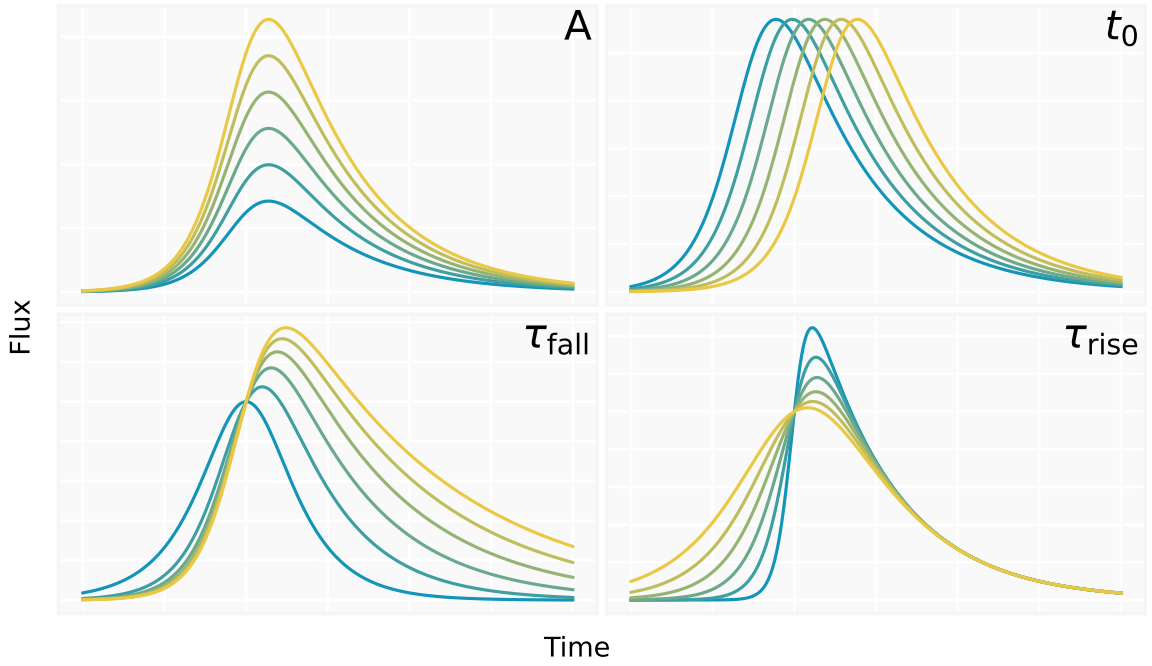



FIGURE 3.7 – Effects of the variation of parameters for the Bazin function. The parameter values increase from blue to yellow. The model's parameters {A, $t_0$, $\tau_{\text{fall}}$, $\tau_{\text{rise}}$} are respectively equal to {1, 0, 10, -2.5}, and panels respectively show their individual variations within bounds {[0.5, 1.5], [-5, 5], [5, 20], [-1, -4]}.

### 3.4.2 Villar function

Villar et al. (2019) proposed an extension to the Bazin function, with the goal of describing more complex behaviors, precisely in the context of feature extraction for machine learning classification. The functional form proposed can be written as

$$f(t;\ t_0, t_1, A, \beta, \tau_{\text{rise}}, \tau_{\text{fall}}) = \begin{cases} \frac{A+\beta(t-t_0)}{1+e^{-\frac{t-t_0}{\tau_{\text{rise}}}}} & \text{if} \quad t < t_1, \\ \frac{(A+\beta(t_1-t_0))\cdot e^{-\frac{t-t_1}{\tau_{\text{fall}}}}}{1+e^{-\frac{t-t_0}{\tau_{\text{rise}}}}} & \text{if} \quad t \geq t_1. \end{cases} \tag{3.3}$$

It contains six free parameters: $\tau_{\text{rise}}$ describes the rising timescale ; $t_0$ is a reference time, $A$ is a multiplicative amplitude parameter, $t_1$ and $\beta$ encodes a plateau phase after the peak and are respectively acting on its duration and its slope, and $\tau_{\text{fall}}$ describes the decay timescale after the plateau phase. The effect of the variation of all parameters is represented on Figure 3.8.

The Villar function enables the description of more subtle transient behavior by allowing the existence of a plateau phase after the peak. Although computationally heavier than Bazin, it offers more accurate models for sufficiently sampled light curves. It has often been used in the astrophysics community. For example, Hosseinzadeh et al. (2020) used it to build a classification pipeline for Pan-STARRS1 supernovae. This pipeline was recently improved and used for ZTF supernovae light curves (de Soto et al., 2024). In Sanchez-Saez et al. (2021), the ALeRCE broker proposed a slightly modified version of Equation 3.3 to replace the discontinuity in the functional form by a smooth sigmoid transition. It was also used for feature engineering and classification of transients.

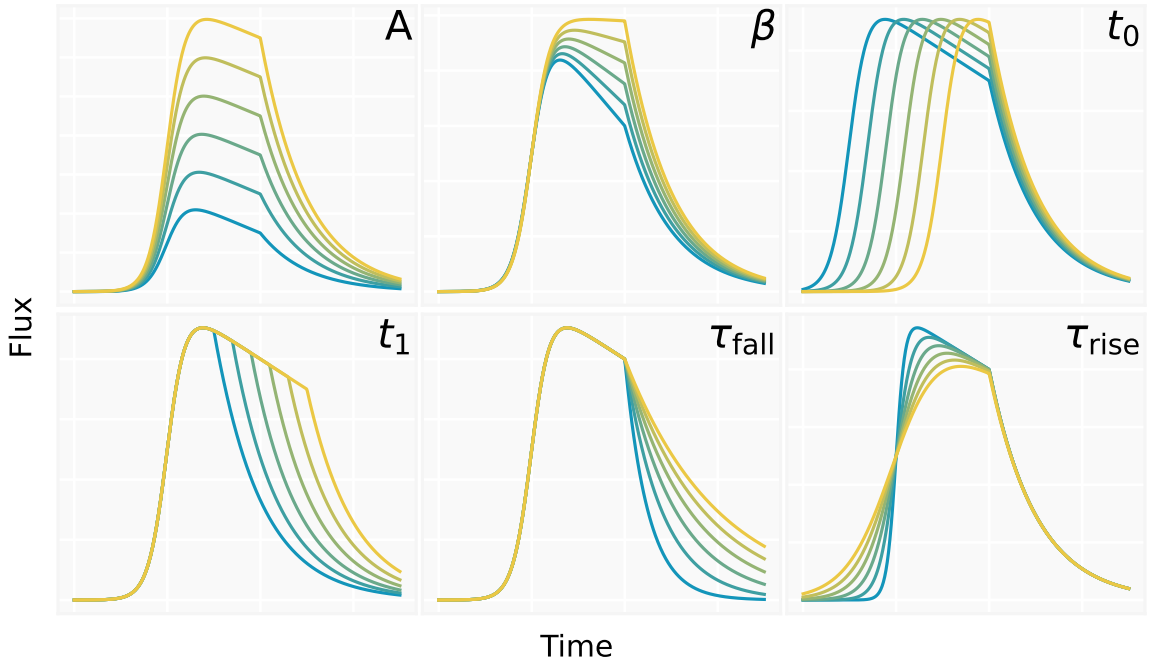



FIGURE 3.8 – Effects of the variation of parameters for the Villar function. The parameter values increase from blue to yellow. The model's parameters {A, $\beta$, $t_0$, $t_1$, $\tau_{fall}$, $\tau_{rise}$} are respectively equal to {1, -0.1, 0, 20, 10, 2}, and panels respectively show their individual variations within bounds {[0.5, 1.5], [-.02, -.001], [-10, 10], [10, 30], [5, 20], [1, 5]}.

### 3.4.3 SALT

Guy et al. (2007) proposed the Spectral Adaptive Light curve Template (SALT) model to phenomenologically describe light curves from SNIa. In opposition to the functions presented above, it performs a two-dimensional fit across time and wavelength by using template spectra to model the flux in different passbands. Its specialization for SNIa, known for their standard behavior, enables a complex description making use of a few parameters,

$$f(t, \lambda;\ t_0, A, x_1, c) = A \times [M_0(t, \lambda;\ t_0) + x_1 \cdot M_1(t, \lambda;\ t_0) + ...] \times e^{(c \cdot CL(\lambda))}. \quad (3.4)$$

It contains 4 free parameters[10]: $t_0$ is the rest-frame time of maximum flux in the B-band, $A$ (usually known as $x_0$) is a multiplicative amplitude parameter, $x_1$ is the stretch parameter, and $c$ is the color parameter. The effect of the variation of all parameters is represented on Figure 3.9.

SALT2 constitutes one of the standard tools in current supernova cosmology analysis. SNIa (Section 3.1.2) are not considered standard candles but rather standardizable candles (e.g. Rust, 1974, Pskovskii, 1977), due to the small intrinsic variability they present. Fitting the stretch, color and amplitude of SNIa light curve enables the correction of these small discrepancies and therefore the precise measurement of $H_0$. Such studies have been performed numerous times, with improved data quality, augmented sample size and finer standardization corrections (e.g. Guy et al., 2007, Betoule et al., 2014, Rigault et al., 2024). The SALT model has also been frequently used for data simulation. Both, the PLAsTiCC and ELAsTiCC, employed it for the

10. Some parameters have been renamed for consistency with previous notations.

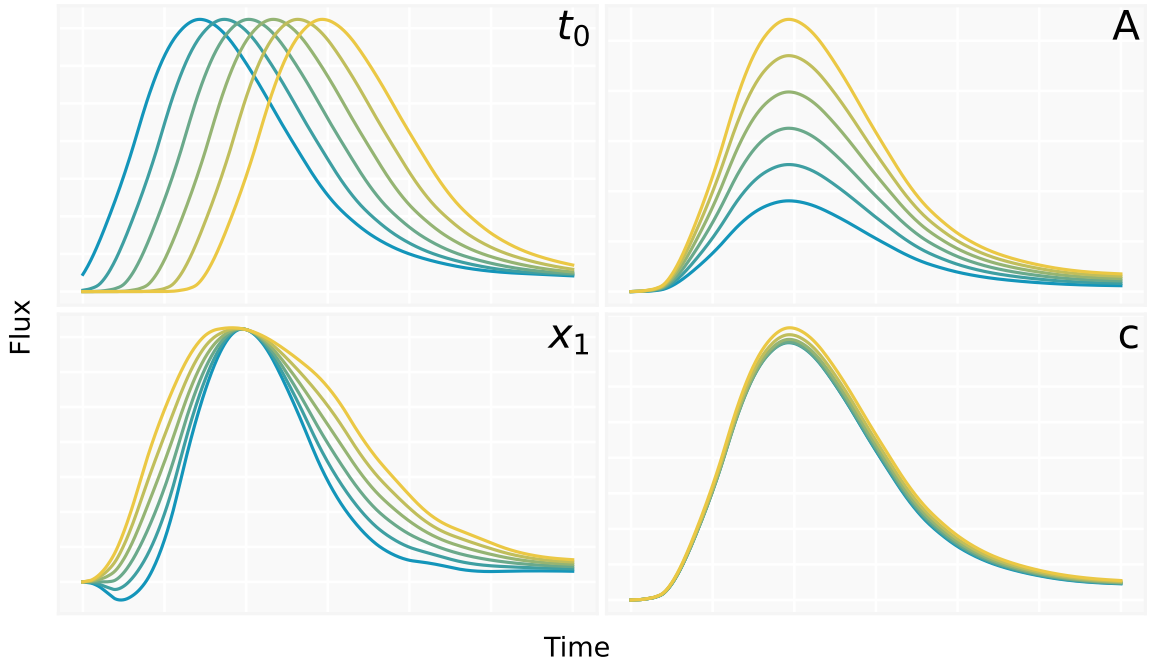



FIGURE 3.9 – Effects of the variation of parameters in the SALT model. The parameter values increase from blue to yellow. The model's parameters $\{t_0, \text{A}, x_1, \text{c}\}$ are respectively equal to {0, 1, 0, 0}, and panels respectively show their individual variations within bounds {[-5, 10], [-0.5, 1.5], [-3, 3], [-0.4, 1]}. The parameter c produces no apparent effect on a single passband (here the B-band is displayed) because it affects the color, i.e. the relationship between passbands.

generation of SNIa light curves (see Section 2.3.4). Although SALT2 remains the most widely used, similar improved SNIa models have been proposed such as the SUpernova Generator And Reconstructor (SUGAR, Léget, P.-F. et al., 2020) that adds two parameters to characterize SNe Ia variability, or SALT3 (Kenworthy et al., 2021), which proposes improved estimation of uncertainties, better separation of color and light curve stretch, and a publicly available training code.

### 3.4.4 Others

Other parametric representation of transients light curves have been proposed in the literature. Newling et al. (2011) proposed a five parameters model – extending the simple Bazin form – which includes a flux tail connected by a cubic spline to the exponential decay. It was developed with the goal of performing photometric supernova classification. Shortly after, Karpenka et al. (2012) also introduced an extension to the Bazin model, employing six parameters, two of which have time unit ($t_0$ and $t_1$). It describes a classical exponentially evolving transient with a potential secondary peak at time $t_1$. Once again, the function was proposed for the feature extraction step preceding a supernovae classification exercise (SNIa vs non-SNIa). Both models have independently been re-used in similar contexts. Vargas dos Santos et al. (2020) used them, along with the Bazin model on the PLAsTiCC dataset, and benchmarked the efficiency of the classifiers resulting from their features. Results showed that the three functions provided equally informative features. Gabruseva et al. (2020) attempted a similar exercise using SALT2, Newling's and Karpenka's equations. Although none yielded better results than SALT2, Newling outperformed Karpenka in the classification of SNIa light curves. Overall, both parametrizations

present intricate forms with high computational instability, which is a challenge for minimizer routines. The unclear benefit of using them over the simple Bazin equation explains their relative low usage in the literature.

Finally, Leoni et al. (2022) reports efforts in classifying supernovae light curves before their peak luminosity. A simple three parameters sigmoid function was used to describe the rising part of the transients. Parameters extracted from the fit were used for the classification of early SNIa. This classifier is currently used by the Fink broker (Section 2.3.3) on the ZTF alert stream.

As shown throughout this section, feature extraction based on parametric fit has been widely used in the literature. It often constitutes the first step of machine learning analysis, which more and more play an important role in time domain astrophysical studies. In the next chapter, we discuss in more depth the role of machine learning in the field, and present the tools that are used in the thesis.

# 4

# Machine learning methods

The term machine learning (ML) designates a family of computer science methods based on a data-driven modeling of the statistical properties of datasets. They can be used on various data structures such as arrays, images, videos, texts, audios or graphs. ML algorithms are often divided into two categories, supervised and unsupervised learning. Supervised learning algorithms determine the relationship between input objects and their output values (labels). They undergo a training phase using a representative set of instances and their associated labels, resulting in the creation of a function that maps inputs to outputs. This function can then be used to predict values for new unlabeled data. Supervised learning includes tasks like regression or classification (Sarker, 2021). On the other hand, unsupervised learning algorithms learn the statistical distribution and its patterns automatically without the need of a training set. It includes methods such as anomaly detection or clustering (Sarker, 2021).

ML emerged as a powerful tool across various research fields due to its ability to process vast amounts of data, identify patterns, and make predictions with minimal human intervention. Astronomy has not been an exception to this trend. Half a century ago, images produced by telescopes were targeted and rare enough to be manually analyzed by astronomers. Nowadays, modern surveys can produce large volumes of complex datasets per night of observation, rendering traditional methods impractical. LSST (and, to a smaller degree, ZTF) represents the ultimate data challenge of visible/near infrared photometry. It will detect $\sim$10 million transients associated to $\sim$ 20 terabytes of images each night (LSST, 2024), hence, a vast majority of the data produced by the telescope will never be seen by a human observer. We have entered the era of big data in astronomy and solutions are being developed to automatically, detect, process, analysis and classify astronomical sources.

Among them, ML has imposed itself as a reliable and efficient solution for many different tasks, in particular for time domain astronomy. Chapter 3 shows in detail how certain light curve behaviors are associated to specific transient types. Therefore, it was natural to use them as an input for machine learning analysis. Historical attempts, such as Filippenko (1997), who characterized light curves visually, have set the building blocks for future automatic learning. McGowan et al. (2003) were among the first to attempt to apply machine learning techniques to light curves. However, it is only $\sim$5 years later, with papers like Mahabal et al. (2008) or Drake et al. (2009), that the time domain community really initiated the transition into the machine learning realm. Today, these methods are omnipresent in the field.

This thesis is anchored in this context, with two supervised ML methods being at the core of this project: Symbolic Regression (Section 4.4) to improve the feature extraction (Section 4.1) of transient light curves (Chapter 6), and tree based classification (Section 4.2.1) to evaluate the quality of the features. Although we chose to analysis the features through the prism of classification, they could be used in different contexts, including unsupervised ML exercises such as anomaly detection (Appendix B).

## 4.1 Feature extraction

In most cases, machine learning models require data of fixed dimension as an input, from which supervised methods can extract statistical properties. In other words, training a classifier to distinguish images from cats and dogs based on their pixel content requires that all images have the same size. An image of different size simply cannot be processed by the model and imposes a transformation step before being classified. In time domain photometry, the objects manipulated are not pictures but rather light curves. They encapsulate physical information that can be extracted and learned by machine learning models. However, as explained in Section 2.2.3, light curves are multidimensional, uneven and noisy. Most ML algorithms require a homogenization step before any subsequent analysis. This procedure is at the basis of the results that will eventually be produced by the ML pipeline. Any valuable information lost during this step will inevitably diminish performances, independently of the complexity of the ML machinery deployed after.

Deep learning methods, such as Recurrent Neural Networks (Rumelhart and McClelland, 1987), have recently been developed to deal with time series without the need to homogenize the data. Such non-parametric treatment has the advantage of being agnostic and purely data driven. However, it can usually be handled properly only by large neural network architectures (e.g. Muthukrishna et al. 2019 or Möller and de Boissière 2020a) and might still benefit from some level of data homogenization (e.g. Fraga et al., 2024, that uses padding for the input size to match the most well sampled light curve). Because of the complexity of these networks, they require large training databases and are computationally expensive to train.

In this thesis, we explore the path of light curve homogenization based on feature extraction (also called feature engineering). It consists in the computation of a set of summarizing values, called features. This set of features must be computable independently of the individual sampling of the light curves, such that the feature extraction step will always yield a vector of the same dimension. The values computed can be simple summary statistics (potentially computed independently for each passband) such as the mean/median flux, the duration, the flux amplitude or the number of observations. They can also be more targeted to probe a specific aspect of the light curves, like the kurtosis, the skewness or an estimate of the period (for periodic sources). Dimensionality reduction methods have also been proposed as a mean to extract light curve features, such as principal component analysis (Ishida and de Souza, 2013) or Fourier decomposition techniques (Deb and Singh, 2009).

Although gathering multiple general statistical features is essential, it is generally not enough to provide a complete description of the photometric behavior of a light curve. A complementary approach consists in fitting an empirical model to the data. The best-fit parameter values, along with a fit quality metric, can be used as features which encode a significant quantity of information. The choice of the model is essential, as it must encode our prior knowledge of the transient behavior. Section 3.4 presents the most common transient parametric models used in the literature. This method has been extensively used in the field of astronomy (e.g. Bloom et al. 2012, Lochner et al. 2016, Ishida et al. 2019), and specialized software packages have even been developed to automatically extract features from light curves (Sanchez-Saez et al. 2021, Malanchev et al. 2021). With the addition of statistical features, this process typically involves the

computation of a dozen to few hundred features (Sanchez-Saez et al., 2021). The construction of an informative parameter space is a non-trivial and delicate step, as it constitutes the basis of any subsequent analysis. A poorly designed extraction can lead to information loss, and/or can generate repeated or uninformative features, all of which may result in a performance decrease of machine learning models. Building the most optimal parameter space is therefore a crucial step for such applications.

In this summarization quest, the choice of parametric model is crucial. Of course, fitting light curves with as much precision as possible is an essential requirement, but it is not enough. Minimization is a mathematical complex task, in particular for non-linear problems (James, 1972). In any attempt to fit a large number of them, it is unavoidable that some fraction will run into local minima, producing meaningless features. This effect can be reduced by choosing functions with as little free parameters as possible, and with minimal degeneracies between parameters. Choosing simpler models will also result in faster minimization time, the extraction of a smaller parameter space, and a reduced risk of overfitting the data. Therefore, an efficient parametric model stands in the compromise between the model accuracy and the simplicity of its description.

In this thesis, we optimize the information collected from the feature extraction of transients based on parametric fits. The improved features will be put to the test by inputting them into transient classification pipelines, a task often tackled in the literature. Indeed, many light curve feature extraction methods have already been built with the goal of classifying large datasets (e.g. Karpenka et al. 2012, Ishida et al. 2019, Sanchez-Saez et al. 2021) and the preparation for the future LSST data maintains this momentum (e.g. Cabrera-Vives et al., 2024b, Fraga et al., 2024).

## 4.2 Tree-based classification

A well-designed feature extraction step enables the summarization of meaningful information contained in light curves into a homogenized dataset representation. This new format is compatible with traditional ML methods, which require a rectangular matrix as an input. In particular, classification models can be built from features. This task represents one of the biggest challenges of a large photometric dataset. Given the number of sources that are, and will be, produced every night by modern surveys, upstream classification pipelines are necessary for the community to fully benefit from the data. Although the goal of this thesis is solely to propose improved methods for the feature extraction of light curves, classification based on these features is performed several times in this manuscript. In this context, the classification results should not be seen as an end in themselves, but rather as a tool to measure the quality of the features extracted. This exercise offers a deep understanding of the parameter space, and how the representation of different type of objects behave in relation to each other.

Many classification methods are available and used in the literature. For example, Vaughan et al. (2024) used a logistic regression to classify galaxy speed of rotations, Kopsacheili et al. (2020) trained a support vector machine model to identify supernova remnants, and Wilson et al. (2023) chose a naive Bayes algorithm to classify young stellar objects. Recently, the astro-

nomy community largely entered the realm of deep learning, and many classifiers now rely on neural networks. They range from classical autoencoders (e.g. Perez-Carrasco et al. 2023, Martinez Galarza et al. 2024) that use a symmetrical architecture to build a reduced latent space, to more complex architectures such as recurrent neural network (e.g. Möller and de Boissière 2020b, Fraga et al. 2024) which internally construct hidden states that carry information from the previous part of a sequence, making them adapted for time series classification (Schmidt, 2019).

In this section, the tree based algorithm used throughout this work is presented. We do not claim that this approach yields the best results of all the possible classification algorithms, and emphasize again that our goal is feature optimization. Therefore the choice of classification algorithm has not been fine-tuned. We choose to use a decision tree based classifier which is generally good, robust, and interpretable, offering a solid understanding of the features and their relative importance. In addition, they are easy to implement, computationally light to train and provide quick classification probabilities.

### 4.2.1 Decision trees

Decision trees (Shalev-Shwartz and Ben-David, 2014) are one of the most classical predictors, used in machine learning. They are general tools used to represent the splitting of a parameter space. Therefore, they are not only used in supervised ML, but also in unsupervised tasks (Liu et al., 2008). Here we focus on decision trees built for supervised classification purposes. In order to illustrate decision trees, let us assume a simple dataset $D$ described in a two-dimensional feature space, from which we attempt to identify two different populations. Figure 4.1 presents the dataset. This is a trivial example, since the two classes are easily separable, even by eye. Note that, in practice, astronomical datasets are much more intricate, with multiple labels, uncertainties, higher dimensionality and significant overlap of populations. However, all concepts explained below still apply to more complex scenarios.

From a given dataset, one can construct a decision tree with the following procedure:

1. Choose one of the features.
2. Choose a threshold for that feature which divides the data into two groups: greater than the threshold and below the threshold.
3. Repeat step 1 with the subsamples, until a stopping condition is reached.

This procedure effectively splits the parameter space into smaller regions. It is convenient to represent successive splits in the form of a tree, hence the name. Figure 4.2 shows the tree representation of the splits shown in Figure 4.1. The first node of a tree is called the root, and the end nodes are the leafs. Once a decision tree is built, it can be used as a predictor, $P$, associating an instance, $X$, to a label, $Y$, by following the tree from its root node to one of its leaf, $P : X \rightarrow Y$. The label Y attributed to an instance X corresponds to the majority class within the leaf.

In order to build an efficient predictor, the features and thresholds chosen during the iterative creation of nodes are driven by the minimization of a metric. Multiple metrics can be used, such as the Gini impurity or the entropy (Shalev-Shwartz and Ben-David, 2014). They are designed to find the optimal split point to separate the classes along a given feature. The Gini impurity is

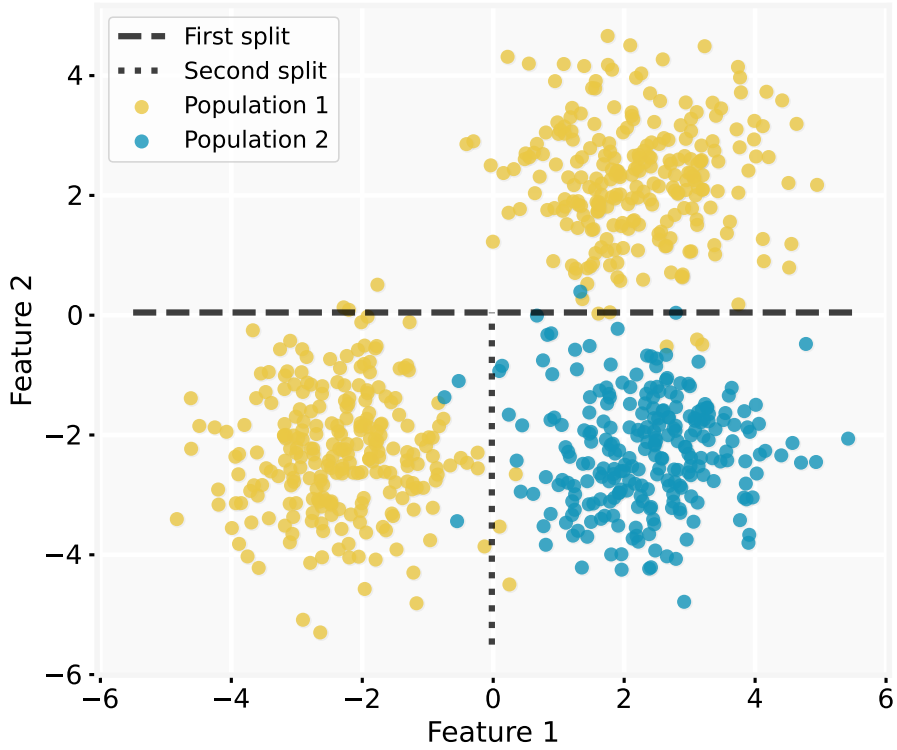



FIGURE 4.1 – Two-dimensional toy dataset containing two populations described by arbitrary features. The long-dash and small-dashed lines represent the split points of the root note and first left-hand split of the decision tree shown in Figure 4.2, respectively. The split points were obtained by minimizing the Gini impurity (Equation 4.1).

the most commonly used for this task. For a dataset containing $k$ labels, and given their relative frequencies $p_i$, with $i \in \{1, 2, ..., k\}$, it is defined as

$$G = 1 - \sum_{i=1}^{k} p_i^2. \tag{4.1}$$

This index provides a measure of how often an instance would be incorrectly classified if a label was attributed randomly based on the distribution of the dataset. The Gini impurity (or Gini index), is inversely proportional to the purity of a give node. Once a split point is chosen, the closer to zero the value of the Gini impurity, the more pure the resulting nodes will be. At each step of the creation of a tree, all features and all split points are examined, and the one minimizing the Gini index is chosen. This step is repeated with the subsets until a stopping condition is reached. It could simply be grown until each leaf is completely pure, however it often leads to overfitting of the training dataset. Other conditions are used to limit its size, such as a maximum tree depth, a Gini value upon which no further nodes are created, or a minimum size of the leaves.

This fully deterministic process has several advantages: it is fast, both to compute and to use as a classifier for new unlabeled data, and it constitutes a directly interpretable predictor. Such interpretability is rare among modern ML algorithms, which are often very complex, and are therefore much more difficult to interpret. However, despite their qualities, decision trees are known to be prone to overfit, and offer poor generalization to unseen data. This issue is

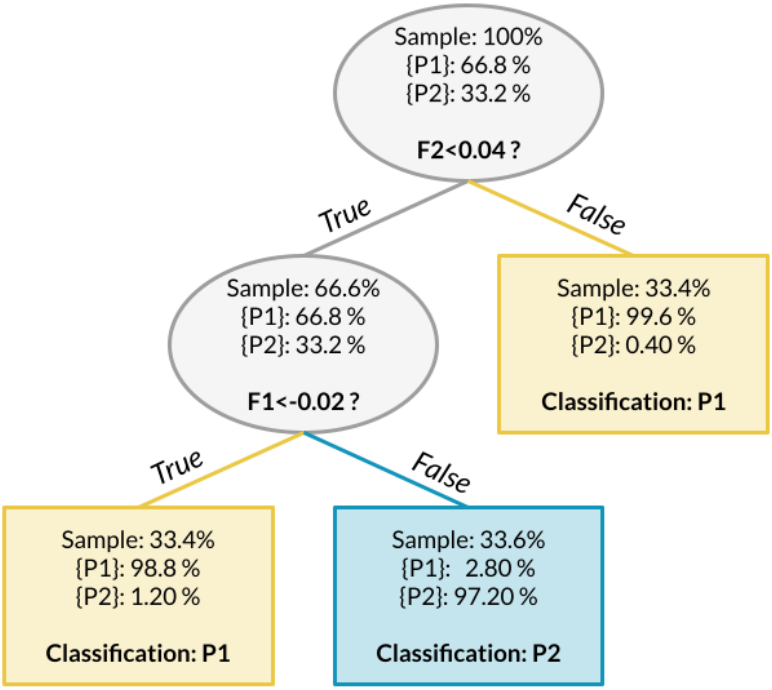



FIGURE 4.2 – Visualization of a decision tree built from the dataset shown in Figure 4.1. *Sample* indicates the percentage of the original dataset contained in the node. The last line in each node (gray ellipse) shows the chosen feature and split point. The percentage of both populations in each node is indicated. The color of a leaf represents the classification associated to it.

commonly addressed by using an ensemble of trees which yields greater statistical power, a forest.

### 4.2.2 Random forest

A Random Forest (RF) is an ensemble method that provides classification based on multiple decision trees (Breiman, 2001). As stated above, decision trees are constructed deterministically, therefore some randomicity must be introduced to generate a population of non-identical predictors. Each tree can, for example, be built from a random subset of the original database, or from a random subset of features at each node (Dietterich, 1998). These procedures ensure that each tree describes a slightly modified version of the original data distribution, and thus, considerably reduces overfitting. The high number [1] of trees enables the construction of a more diverse ensemble with the potential of grasping more details about the statistical distribution of the data than single individuals. A new instance label is determined by predicting its label independently on each tree and using the majority vote as the final answer [2].

Despite its apparent simplicity, Random Forest has proven to be a serious competitor to state-of-the-art classification methods. Its performances have been established through numerous empirical studies (Biau, 2010), and the field of astronomy in particular has widely put

1. The exact number of trees is a parameter defined by the user and should be chosen taking into account the dimensionality and complexity of the dataset.
2. Different weights can be applied for the different trees (El Habib Daho et al., 2014), but this method won't be used in this thesis.

the method to the test throughout the years (e.g. Breiman et al., 2003, Merghadi et al., 2020, Sánchez et al., 2014, Smirnov, 2024). It is interpretable and largely avoids overfitting. It is also statistically proven to handle sparsity, such that the rate of convergence scales only with the informative features and not with the number of noisy variables (Biau, 2010). Random forest models are light and fast to train (scaling linearly with the number of trees), which make them adapted to the large scale data processing of modern photometric surveys. In addition, they provide high performance as general purpose predictors. For these reasons, all classifier models in this thesis are built using a Random Forest algorithm (Chapters 5 and 7).

Because decision trees are based on a succession of feature choice, a feature importance can naturally be extracted from a trained RF model. Multiple methods are used to compute feature importance. One of the most common is the mean decrease in impurity, also called the Gini importance. For a given feature, it is defined as the total reduction in impurity brought by the split on that feature, weighted by the number of sample in the node, and averaged across all trees in the forest (Breiman et al.). The higher the Gini importance, the more discriminative is a given feature. The evaluation of the importance of features is a key element of feature extraction. Not only it may point towards uninformative features that could be dropped out of the analysis, it also allows checking the coherence of the results given our familiarity with the data and the parameter space at hand.

Classification exercises are performed several times within this thesis (Chapter 5 and 7). The results offer a good comprehension of the data itself and the relationship between the transient classes. But most importantly, it enables the evaluation of the information carried by the features. We opted for Random Forest classifiers due to their robustness, ease of implementation and fast computation time. In addition, their interpretability regarding features and their relative importance is key for feature engineering.

## 4.3 Classification metrics

The goal of classification is to separate as best as possible the classes (also called targets) within a dataset. This section presents three standard metrics commonly used to evaluate the performance of classifiers. Although other metrics exist, only these will be used throughout the thesis. The predictions of a classifier applied to a given dataset can be summarized by a confusion matrix, as illustrated on the left panel of Figure 4.3. In this virtual example, a classifier has attributed classes {A, B, C} to the instances of a dataset. The columns represent the targets predicted by the classifier, and the rows represent the true classes. The numbers in the cells are counts of instances.

Therefore, in a confusion matrix, the diagonal (here top-left to bottom-right) represents the true (T) predictions. In opposition, the off-diagonal elements are false (F) predictions. From this, a first metric, the accuracy, can be defined as

$$accuracy = \frac{T}{T+F}. \tag{4.2}$$

It represents the probability that the model will output a true prediction for a randomly drawn instance in the dataset. Although it has the benefit of being simple to interpret, it is

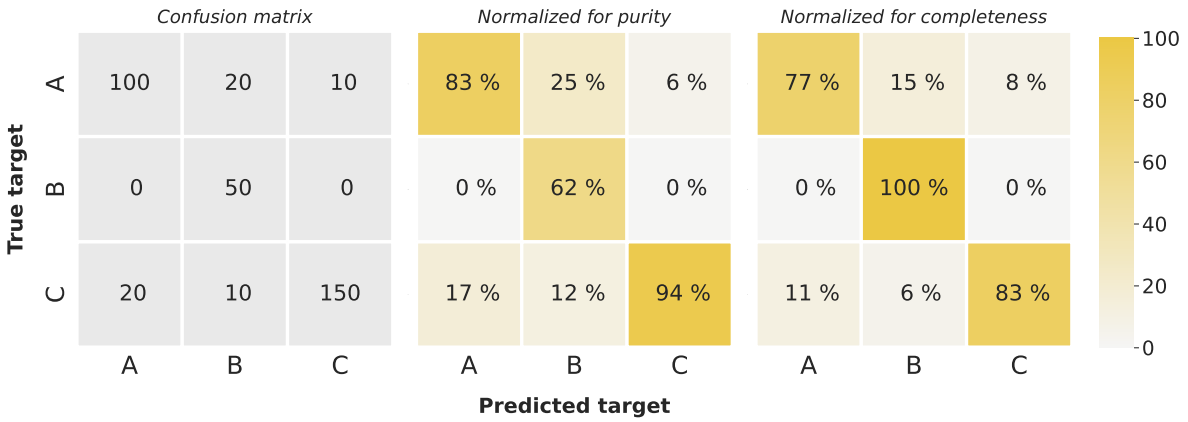



FIGURE 4.3 – The left panel display an example of confusion matrix. The middle and right panels show its normalization respectively on purity ($\sum$ columns = 100%) and completeness ($\sum$ rows = 100%), respectively.

not sufficient to understand the behavior of the classifier for each target. In particular, if the classes are imbalanced, i.e. some are significantly more frequent than others, accuracy will favor the majority class. In order to better characterize a classifier, true and false predictions are subdivided. For a given class, $C$:

— True Positives (TP) are instances belonging to $C$ and predicted as $C$

— True Negatives (TN) are instances not belonging to $C$ and not predicted as $C$

— False Positives (FP) are instances not belonging to $C$ and predicted as $C$

— False Negatives (FN) are instances belonging to $C$ and not predicted as $C$

From this, two class metrics can be computed. The first one is purity (also called precision). It is defined as:

$$purity = \frac{TP}{TP+FP}. \tag{4.3}$$

It corresponds to the probability that the prediction of the classifier is true, given that the prediction is $C$. Hence, purity represents how trustworthy is the model when it predicts a given class. As illustrated in the middle panel of Figure 4.3, a confusion matrix can be normalized such that all diagonal elements display the purity of its class. This is done by dividing each cell by the sum of the elements of its column. The second metric which can be computed is completeness (also called recall). It is defined as:

$$completeness = \frac{TP}{TP+FN} \tag{4.4}$$

It corresponds to the fraction of instances of class $C$ that have been correctly predicted by the classifier. Hence, completeness can be used to assess the fraction of instances that will be missed by the model, given a class C. As illustrated in the right panel of Figure 4.3, a confusion matrix can be normalized such that all diagonal elements display the completeness of its class. This is done by dividing each cell by the sum of the elements of its row.

The combination of both metrics provides a good description of a classifier. Indeed, looking at only one of them can lead to misleading appreciation of its performance. For example, by classifying every instance of a dataset as belonging to $C$, the model is guaranteed to correctly

classify all true $C$. In that sense, it would obtain a completeness of 100% for the class $C$, however purity would be low. Similarly, by being extremely conservative on its attribution of the $C$ class, a model can easily reach 100% purity. But the completeness will be low, since most of the $C$ events would be misclassified.

Despite being very informative, computing these metrics only provides part of the characterization of the results, since no uncertainties are associated to them. In order to estimate the variance of classification scores, resampling methods are generally used (Raschka, 2018). The goal is to create multiple training and testing samples using the original dataset, from which we can compute multiple classification metrics. The statistical properties of the results' distribution enables the computation of uncertainties, by using, for instance, the standard deviation or the percentiles. Various resampling methods are used for this purpose. For example, cross-validation is commonly used (in astronomy e.g. Liang et al., 2023, Perez-Carrasco et al., 2023). It consists in the random division of the dataset in k parts, from which k models are trained, each tested on a single subsample and trained on the others. Another widely used method is bootstrapping (in astronomy e.g. Bazell and Aha, 2001, de Diego et al., 2020). From a dataset made of $n$ instances, the resampling is performed by drawing $n$ individuals with replacement to create a training sample. Individual never chosen constitutes the testing dataset. This operation is repeated k times, from which k models can be trained.

Supervised classification constitutes a classical exercise of machine learning, and one of the most important challenges of time domain surveys. Since they rely on the input features, classifiers, such as Random Forest, can be used to better evaluate the information extracted. The different classification metrics at our disposal enables the understanding of the feature parameter space, in particular of the overlapping regions, i.e. the misclassified events. Such efforts contribute to the optimization of classification pipelines for current and future challenging astronomical data surveys. It represents one approach to machine learning, mostly numerical and often very abstract. It is very far from the original scientific exercise of constructing laws to model phenomenons based on observations. In the next section, Symbolic Regression, a different philosophy over which ML has been developing is presented. It is much closer to the experimental nature of fundamental sciences, while taking full advantage of the power of machine learning. In this thesis, we explore its utilization to optimize the feature extraction step, and we propose to combine it with classical ML approaches to yield the best possible results.

## 4.4 Symbolic Regression

Section 3.4 shows how numerous mathematical models have been proposed to fit transient light curves. Every function presented share a common origin: they have been handcrafted. None of them are generated from first principles. They are phenomenological models, carefully designed by experts based on their knowledge, the available literature of the time, their science interests, and, probably also, an irreducible part of creativity. Such descriptions, have proven their scientific importance by providing a consistent framework where more information can be gathered about a given class even before a full theoretical description is available to explain it. Nevertheless, given the immense space of possible functions to explore, how confident can one be that an optimal solution has been found ? The history of functional forms proposed for

transient light curves has revealed that models can take a wide range of shapes and number of parameters ; that they can be complexified or simplified ; and that they yield various results depending on the type of problem considered. In this context, where the evaluation of all possible representations seems required, the field of Symbolic Regression offers powerful tools to approach the issue. It offers an opportunity to overcome human bias and inspire the exploration of different phenomenological explanations.

Symbolic Regression (SR) is a form of regression method that explores the parameter space of possible equations to optimally fit a dataset. Therefore, it does not require any prior parametric model. It discovers the model structure and optimizes the free parameter values at the same time. It was first introduced by Koza who used Genetic Programming to search for optimal solutions. It has since become the most widely used approach. However, its heavy computational cost has, for a long time, been a drawback to the full implementation of these methods in fundamental science scenarios. Nevertheless, the constant improvement of CPU performances and the optimization of algorithms have given a new momentum to SR.

In this thesis, we explore the use of SR and the development of new methods for an optimal representation of transient light curves. This section presents the algorithm behind SR, an overview of the main implementations available, and the usage of the method in the field of astronomy.

### 4.4.1 Algorithm

Most symbolic regression algorithms are based on genetic programming (first conceptualized by Turing, 1950). It is a branch of computer science based on the generation of an ensemble of random solutions (called a population), to which the general Darwinian principles (Darwin, 1859) are applied, i.e. the survival and mutation of the fittest individuals. Multiple implementation of this same idea have been proposed, from which a selected list is discussed in Section 4.4.2. Although each contains minor specificities, the core concept always remains the same. The goal is to find a mathematical expression with one or more variables that optimally fits a given dataset. It searches for a model $f(x)$ that minimizes a loss function measuring the goodness-of-fit to the input data. Functions are internally encoded as trees [3], as illustrated by the left panel of Figure 4.4. This representation enables efficient numerical computation for the following evolutionary steps:

1. **Hyperparameter selection:** Although we can sometimes read that, SR allows the unbiased and purely data driven discovery of models, this is an oversimplification of the underlying process. Any SR search is constrained by prior hyperparameters set by the expert. The most important choice concerns the operators, i.e. the mathematical building blocks which will be considered when constructing the functional form. They can include the basic set of operations $\{+, -, \times, \div\}$, as well as more complex ones $\{abs, cos, sin, power, log, exp, max, min, ...\}$. The size of the parameter space to explore quickly scales with the number of operators included in the analysis. For practical science use cases, it represents a way to inject prior knowledge about the problem. For example, if the input dataset is a supernova light curve (Chapter 3), prior knowledge indicates that an exponen-

3. Other encoding strategies exist, such as lists (Bartlett et al., 2022).

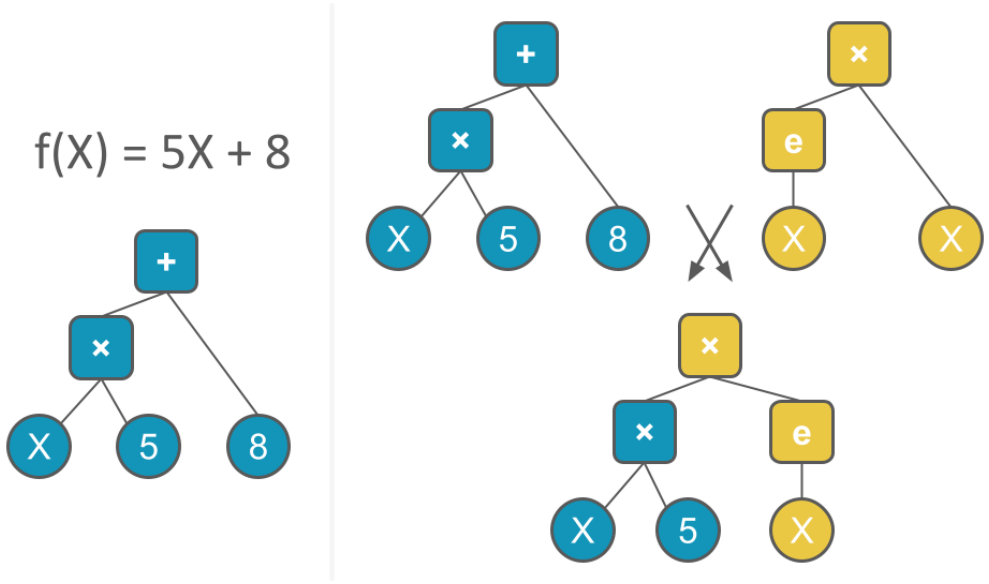



FIGURE 4.4 – The left panel illustrates how mathematical equations can be represented as trees. The right panel illustrates the crossover mutation, i.e. how two trees (blue and yellow) are mixed to create a new one.

tial might be important, while a cosines would make no physical sense. Beside operators, other hyperparameters such as the loss function, the size of a population, the maximum size of trees, or the frequencies of the different mutations (see item 4), can be tuned.

2. **Initialization:** The first step consists in the creation of a set of randomly generated equations (in the form of trees). The ensemble of trees forms the first generation. In all likeliness, the random equations will poorly fit the dataset.
3. **Selection of the fittest:** In order to quantify the quality of the solutions, a goodness-of-fit is computed for each individual using a loss function. A good equation must not only fit the dataset as best as possible, but should also be as simple as possible. This second condition is crucial because a polynomial form of arbitrarily large size can perfectly fit any dataset. The size of the generated equation can be controlled by selecting a goodness-of-fit metric that penalizes the size of the trees. This step will result in a ranking of individuals, from which only the fittest will be kept for the next generation. The exact fraction of the population that survives constitutes a hyperparameter.
4. **Perturbation and recombination:** The best individuals are used to create the next generation. Random modifications, called mutations [4], are applied to their trees to generate new equations. Mutations can take several forms: *point mutation*, i.e. the random modification of a single tree node ; *subtree mutation*, i.e. the random modification of a sub part of the tree ; *hoist mutation*, i.e. the trimming of a random sub part of the tree ; *crossover*, i.e. the random merging of two trees (illustrated in the right panel of Figure 4.4) ; or an absence of mutation, passing the tree as it is. The exact probabilities of each mutation to occur constitute hyperparameters of the analysis.
5. **Repeat until convergence:** Tree mutations result in the random exploration of the space of possible solutions. In order to converge towards an optimal solution, step 3 and

4. This name is inherited from the Darwinian evolution analogy. DNA changes result in the mutation of the individuals

4 should be repeated many times. At each generation, most modifications won't improve the quality of the fit and thus won't be kept for the next generation. However, beneficial changes will remain and quickly spread to the population. Therefore, with time, equations will be increasingly good at describing the input dataset. The evolution stops either once a given number of generation or a minimum goodness-of-fit threshold has been reached. The result takes the form of a Pareto front (Smits and Kotanchek, 2005), that is, a list of the best solutions for each different tree size. This allows the expert to decide up to which point an improvement of the fit is worth the increase in the equation complexity.

### 4.4.2 Implementations

The field of symbolic regression is vast, and many independent implementations have been proposed in the last decade. This section will present an incomplete (largely biased towards physics, open access and python) list of the current most used among them. This overview illustrates the effervescence of the research in this very quickly evolving branch of computer science. All codes presented below have been benchmarked by La Cava et al. (2021), who have provided a detailed study on their relative performances in terms of goodness-of-fit, computation time, resilience to noise and model size.

*GPlearn*[5] is a fully Python based implementation, which stands among the most accessible ones. It largely follows the core concepts presented in Section 4.4.1. It presents the advantage of being very flexible, allowing the usage of custom operators and loss functions. This implementation provides satisfying and small sized solutions for simple problems. However, it displays very inefficient computation speed, and the required time for the evolution quickly grows large for complex problems.

In *AI Feynman*, Udrescu and Tegmark (2020) propose a physics oriented approach to SR. This implementation, which uses a Fortran backend but includes a Python interface, takes advantage of prior physical knowledge such as units and symmetries to constrain the size of the space to be searched. It does not use genetic programming to explore this space, but rather a combination of well-defined algorithmic methods and deep neural network. It has been developed to recover all physical equations found in the famous Richard Feynman *Lectures on Physics* (Feynman et al., 2010), for which it exhibits great performances. But it has been clearly uncapable of performing outside that scope, and presented the worst goodness-of-fit scores of all SR methods when confronted with the challenging (but non-physical) benchmark proposed by La Cava et al. (2021).

To the contrary, *Operon* (Burlacu et al., 2020) has appeared as the best performing SR algorithm of the benchmark. It is developed in pure C++, but recently *pyoperon*, a Python wrapper, has been released. It globally uses genetic programming as described in Section 4.4.1 but with careful resource management and parallelization capacities. This results in a state-of-the-art implementation capable of generating small and accurate models significantly faster than the previous ones. However, it lacks flexibility, making the usage of custom operators/cost functions difficult. It can be a drawback for practical research that often requires freedom of exploration.

5. https://gplearn.readthedocs.io/en/stable/index.html

To solve this issue, Cranmer (2023) recently [6] proposed *PySR*, a modern and efficient SR implementation oriented towards science applications. It has an optimized Julia [7] backend but proposes an accessible Python interface. In particular, it evaluates and mutates individuals inside random subdivisions of the population, allowing for efficient parallelization and independent evolution. Since it was conceived to be used in real science cases, it includes methods to deal with noisy data and uncertainties. Future SR research should help in accessing its relevance for practical research.

### 4.4.3 Symbolic Regression in astrophysics

Despite the great potential of SR in fundamental sciences, its usage has been mostly restricted to the computer science field until recently. NASA ADS [8] only references 76 astrophysics related papers that contain the term Symbolic Regression, of which a significant fraction are simply methodological SR articles which include general physics equations for their benchmarking (such as Udrescu and Tegmark, 2020). Graham et al. (2013) first showed the general applicability of SR to astrophysical datasets by generating relationships for the Hertzsprung-Russell diagram and the fundamental plane of elliptical galaxies. Krone-Martins et al. (2014) later applied SR for concrete astrophysical applications. Using a dataset of galaxies with spectroscopically determined redshift, the authors were able to propose an equation to estimate the photometric redshift with great accuracy. These works were performed using *Eureqa*, a SR implementation that recently fell into the private domain.

However, despite these first efforts, SR has remained almost unused in the field until 2021, which marks a significant acceleration in the interest of the astrophysical community. For example, the CAMELS project (Villaescusa-Navarro et al., 2021), an ambitious state-of-the-art ensemble of cosmological simulations, uses SR to find an analytic expression of the star formation rate density. Matchev et al. (2022) successfully used *PySR* on synthetic dataset to find a formula describing the transit radii of generic hot exoplanets. Bartlett et al. (2022) proposed another approach to SR based on the exhaustion of all possible solutions, guaranteeing that the optimal one will be found. This solution was applied in the field of cosmology to a sample of SNIa, and was able to discover many relationships fitting the data more economically than the Friedmann equation. More recently, Bartlett et al. (2024) used *Operon* to generate an analytical expression for the nonlinear matter power spectrum as a function of redshift and cosmological constants. It offers an interpretable solution to a problem of high importance in cosmology, for which only opaque deep neural networks used to give an accurate approximation.

Symbolic Regression and other recent machine learning developments have deeply changed our relationship to data. A diverse array of methods has been developed to address various challenges across all scientific disciplines. In the field of astronomy, these methods not only yield excellent results but have often become simply essential. For instance, the time domain surveys ZTF (Section 2.3.1) produces $\sim$ 1 million alerts per night, making visual inspection of the data impractical. The forthcoming LSST (Section 2.3.2) will generate even larger amounts of information by identifying $\sim$ 10 times more variable sources each night. Machine learning

6. It is more recent than the benchmark paper (La Cava et al., 2021) and is therefore not included.
7. https://julialang.org/
8. https://ui.adsabs.harvard.edu/

represents one of the most powerful tool available for analyzing such complex datasets. This manuscript proposes a series of method to optimize the feature extraction of transient light curves. In this context, ML methods are omnipresent. Random forest classification is proposed as a way of evaluating the features and showcasing potential practical applications (Chapter 5 and 7), while a new SR method is developed and used for the elaboration of novel features (Chapter 6). The subsequent chapters presents in details the methods and their performances. They constitute the original contributions of this thesis.

# 5

# Rainbow

*The work presented below has been published in Astronomy & Astrophysics, as part of this thesis work. The content of this chapter was adapted from the original paper, Russeil, E. et al. (2024).*

Extragalactic astronomical transients represent an essential source of information to understand our Universe. They include a wide variety of phenomena (Chapter 3) that we can study, in part, based on the careful examination of their light curves (Section 2.2.3). Properly modelling them not only gives valuable insight about the physics of the transients, it also enables the interpolation of missing observations or the prediction of its future behavior. This step is commonly performed by fitting the light curves using a phenomenological model, several of which have been proposed and used for this task (as detailed in Section 4.1).

Transients are also studied at various wavelengths through the usage of photometric filters (Section 2.1.2). This is essential since the light curves of astrophysical objects are intrinsically 2D surfaces displaying an evolution in time and wavelength. Nevertheless, the parametric functions currently used only describe a one dimensional evolution of the flux with time [1]. This problem is traditionally approached by breaking the 2D nature of the objects and fitting the light curve in each passband separately. For example Zheng and Filippenko (2017) used a broken power-law function to estimate the light curve behavior in each filter for SN2011fe, and Demianenko et al. (2023) compared different machine-learning-based approaches to light curve estimation, with a special emphasis on neural networks.

The independent parametric fitting of light curve passbands has also been used for feature extraction purposes (e.g. Karpenka et al. 2012, Ishida et al. 2019). Hereafter, this procedure will be designated as the MONOCHROMATIC method. This approach ensures a full description of the object without any physical assumption beside the choice of model, but it has three important drawbacks.

1. The number of extracted features scales linearly with the number of passbands. In some cases, the number of filters can be quite large. LSST (Section 2.3.2), for instance, will use 6 passbands, and the current Southern Photometric Local Universe Survey [2] (S-PLUS) uses 12. Combining observations from different telescopes can further increase the number of parameters to fit.
2. The feature extraction of an object can only be performed if it contains enough data points in each passband to produce a meaningful fit. Indeed, a fit can only be mathematically constrained if the number of parameters to minimize is at least equal to the number of observations. If one passband is undersampled, the final description is compromised, and the entire object might be dropped from the analysis. This can significantly affect the size of the sample and amplify the biases toward brighter, well-sampled objects.
3. Considering each filter individually will result in the extraction of highly correlated features because the fluxes in different wavelengths are not completely independent. For instance, the transient peak times in different passbands are linked. In a machine learning context, multiple strongly correlated features should be avoided because it might lead to a decrease

1. Among the options described in Section 3.4, SALT is the only exception.
2. https://www.splus.iag.usp.br/

in the overall performances, and to potential sparsity in the case of small datasets (Yu and Liu, 2003).

The issue of the independent passband paradigm has been investigated in the literature through various methods. For example, Villar et al. (2019) used an iterative Markov chain Monte Carlo (MCMC) fitting procedure that combines the first iteration posterior of each passband and uses it as a prior for the next iteration. Although this method ensures a more coherent parameter set, in particular in the case of filters with significantly fewer data points, it comes at a high computational cost and constitutes only a numerical, rather than an analytical solution. Other works such as Boone (2019) used Gaussian process to describe SNe in the wavelength versus time parameter space and used the GP representation for database augmentation. Kornilov et al. (2023) also used Gaussian process to construct a fully data-driven multidimensional representation of superluminous SNe light curves. Despite their proven applicability, such approaches are non-parametric and computationally expensive, constituting a bottleneck for high volume data processing. Yang and Sollerman (2023) presented an entire software environment for light curve estimation, including options for alignment of observational data from different passbands, and subsequent color and SED estimation through time. Finally, SALT2 (Section 3.4.3) must be mentioned since it offers a 2-dimensional parametric equation for SNIa light curves by using template spectra to model the flux in different passbands. It is a good solution to model SNIa, however it cannot be generalized to other transient events.

The arrival of large-scale sky surveys, such as LSST, will increase the number of available light curves by at least a few orders of magnitude, boosting the potential impact of these studies, but also imposing a new challenge. Ideally, analysis methods should be both physically motivated and computationally efficient. They should thus allow not only easy application to data releases of increasing volume, but also enable immediate integration with community broker systems (Section 2.3.3).

The Rainbow approach, developed during this thesis, aims to give a physically motivated solution to the problem of multi-band light curve fitting at scale. It is composed of three main ingredients: a blackbody-profile hypothesis for the emitted electromagnetic radiation (Section 2.1.4), and two parametric analytical functions describing its temperature evolution and bolometric light curve. The coherent combination of these elements results in an efficient description of the light curve behavior in different wavelengths, even for sparsely sampled light curves from different observational facilities. We demonstrate the performance of the Rainbow method when applied to simulated data from the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC, Section 2.3.4) and real data from the Young Supernova Experiment (hereafter YSE, Aleo et al., 2023).

In Section 5.1 we describe our motivations, experiment design choices (Section 5.1.1 and 5.1.2), and fitting procedure (Section 5.1.3). Section 5.2 introduces the dataset. In Section 5.3 we present the results through several metrics, including goodness of fit (Section 5.3.1), maximum flux time prediction (Section 5.3.2), and full or rising light curve classification (Section 5.3.3). We describe the results of the method applied to the YSE data in Section 5.4. Finally, we present our conclusions in Section 5.5. Complementary figures are available in Appendix A.

## 5.1 Method

We used the blackbody spectral model as a proxy for the thermal-electromagnetic behavior of astrophysical transients. The observed spectral flux density in this model is characterized by any two of three physical quantities (or their combination): the solid angle, temperature, and bolometric flux of the object. We chose to parameterize this via two independent functions of time: one function for the bolometric flux $F_{\rm bol}(t)$, and the other for the temperature $T(t)$. From the Stephan-Boltzmann's law, we express the total bolometric flux emitted by a source as:

$$F_{bol}^{tot} = \sigma_{\rm SB} T^4. \tag{5.1}$$

We also express the total flux per frequency unit, which is computed by the integration of the blackbody over a solid angle, such that:

$$F_{\nu}^{tot} = \int B_{\nu}(\nu) d\Omega = \pi B_{\nu}(\nu). \tag{5.2}$$

Finally, we know that the ratio of the bolometric flux over the flux per frequency unit should be the same for any observer,

$$\frac{F_{\nu}^{obs}}{F_{bol}^{obs}} = \frac{F_{\nu}^{tot}}{F_{bol}^{tot}}. \tag{5.3}$$

Since temperature and bolometric flux are a function of time, $t$, this leads to the following expression for the observed [3] spectral flux density per unit frequency, $\nu$:

$$F_{\nu}(t,\nu) = \frac{\pi\, B(T(t),\nu)}{\sigma_{\rm SB}\, T(t)^4} \times F_{\rm bol}(t), \tag{5.4}$$

where $B$ is the Planck function (Equation 2.4). We do not take cosmological redshift effects into account because the blackbody spectrum retains its properties when it is redshifted (Section 2.1.5). Thus, all the quantities here, including temperature, are assumed to be in the observer frame. For the purpose of comparison with observations, we computed the average of the spectral flux density $F_{\nu}$ for each passband $p$. This was done by incorporating the corresponding filter transmission function, denoted $R(\nu)$,

$$F_p(t) = \frac{\int F_{\nu}(t,\nu)/\nu\, R(\nu)\, {\rm d}\nu}{\int 1/\nu\, R(\nu)\, {\rm d}\nu}\,. \tag{5.5}$$

Notably, instead of integrating over the passband transmission, this method can instead be used with the blackbody intensity at the effective wavelength of each given passband. This approach would be more efficient computationally, but it would be less accurate. In this work, we always integrate over the passbands, as shown by Eq. (5.5) using transmissions provided by the Spanish Virtual Observatory (SVO [4]) filter profile service (Rodrigo et al., 2012).

We called the method Rainbow after its continuous wavelength description. Figure 5.1 illustrates the 2-dimensional continuous surface generated by Rainbow fits. Equation 5.4 is a general form and can be adapted to different problems by choosing appropriate parametric functions to describe the bolometric flux $F_{\rm bol}(t)$ and the temperature evolution $T(t)$. These choices depend on the type of object to describe, and will set the number of free parameters and

3. The super indexes *obs* are dropped for clarity.
4. http://svo2.cab.inta-csic.es/theory/fps/

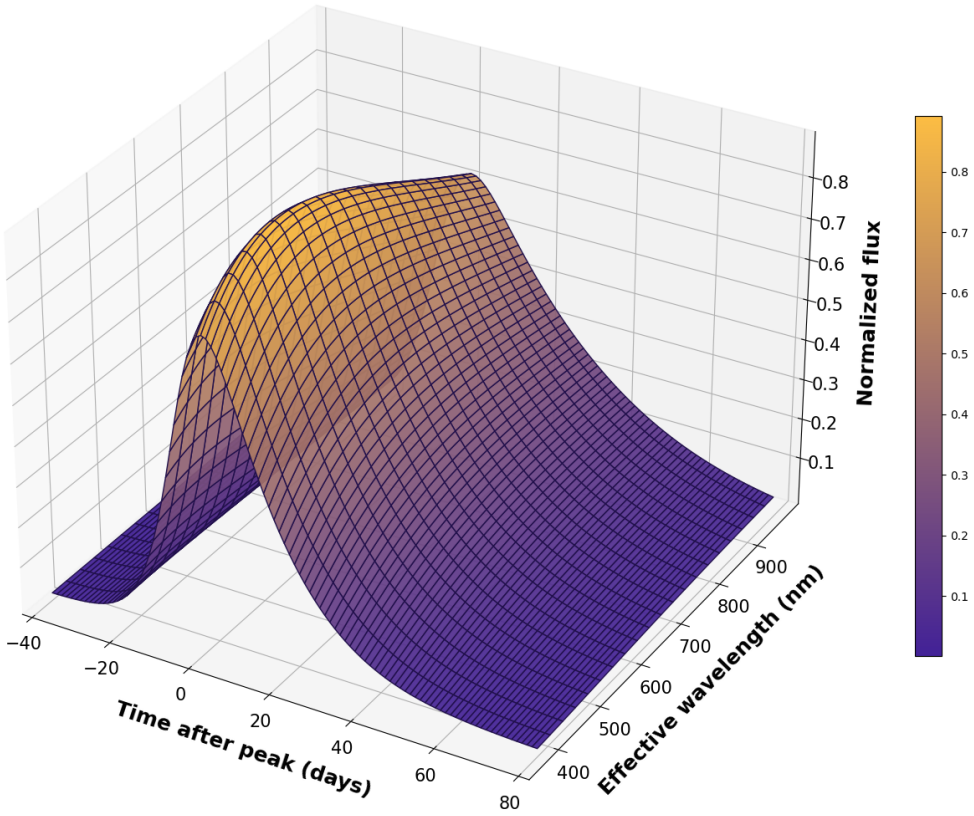



FIGURE 5.1 – Surface plot representation of a RAINBOW fit of a SNII (SN 2020thx) light curve. Representations per passband of the same object with the data points are displayed in Figure 5.10.

the complexity of the equation. The method is particularly well suited to describe poorly sampled data, as the physical assumptions will complement the missing information. It simultaneously addresses the three problems highlighted above:

1. The number of parameters remains constant independently of the number of passbands.
2. An object that is correctly sampled overall, but with sparse data in one or more passbands can still be fitted. Any information from the undersampled filters still helps to constrain the minimization.
3. Because a single fit is performed, repetitive information from the same parameters across different passbands is avoided. The previous small variability encompassed within the repeated parameters is now contained in the temperature evolution parameters.

It is important to emphasize that in a high-cadence scenario, the physical assumptions can impose certain behaviors that will not exactly correspond to the reality of the data. Thus, RAINBOW does not produce optimal results for extremely well-sampled light curves. Moreover, the number of parameters not scaling with the number of passbands comes at the cost of adding the temperature evolution parameters. Therefore, this method is only useful in the case of multi-passband data. Nevertheless, RAINBOW is perfectly suited to process the diverse range of cadences and filter sets adopted by modern datasets.

### 5.1.1 Bolometric flux

In this work, we choose to use the functional form proposed by Bazin to model the bolometric light curve behavior (detailed in Section 3.4.1). This form is the most commonly used and, and although it is not perfect, especially for plateau phases, it is a reasonably good approximation for all different transient types. The relative simplicity of the equation also ensures that the fit can be performed even with a restricted number of observations. We remind here that the Bazin function is expressed as:

$$f(t;\ t_0, A, \tau_{\text{rise}}, \tau_{\text{fall}}) = A \times \frac{e^{\frac{-(t-t_0)}{\tau_{\text{fall}}}}}{1+e^{\frac{t-t_0}{\tau_{\text{rise}}}}}. \tag{5.6}$$

Although this analysis will be limited to the Bazin function, other choices of bolometric flux could be used. The Villar function (Section 3.4.2) would yield more accurate fit at the price of heavier computation and requiring more data points. On the other end, a simpler function like a sigmoid could be considered to model only the rising part of the light curves. These options are later explored in this manuscript. In chapter 7, a panel of parametric functions are used to describe a wide range of bolometric behaviors. In addition, Appendix D presents a side project of this thesis, for which a sigmoid function was used to model early tidal disruption events.

### 5.1.2 Temperature

Despite being used to describe different types of transients, the Bazin function is most commonly applied to SNe. Consequently, we employed a temperature evolution function coherent with SN classes. We used the SUpernova Generator And Reconstructor model (SUGAR ; Léget, P.-F. et al. 2020), a successor of SALT (Section 3.4.3), which provides improved models of the spectral energy distribution of SNIa. We generate spectra from -12 to 48 days since maximum brightness in B band, with a step of 3 days. Each spectrum was fitted with a blackbody to obtain the temperature at a given time, effectively reconstructing the temperature evolution (Figure 5.2).

We visually inspected the results from a sigmoid-like fit (Equation 5.7) and confirmed that it is a good first approximation for the temperature behavior, while still remaining relatively simple and general. Thus, we parameterized the temperature evolution as

$$T(t) = T_{\text{min}} + \frac{\Delta T}{1+\exp\frac{t-t_0}{\tau_{temp}}}, \tag{5.7}$$

which is a four-parameter logistic function behaving as two flat curves linked by an exponential slope. $T_{\text{min}}$ is the minimum temperature the object will reach, $\Delta T$ describes the full temperature amplitude, $\tau_{temp}$ corresponds to the characteristic cooling timescale, and $t_0$ is a reference time parameter that corresponds to the time at half of the slope. Preliminary results using independent $t_0$ for the bolometric and temperature models often resulted in almost equal fit values. Therefore, we assumed that $t_0$ from both equations are equal and merged them into a single parameter. We used Equation 5.7 to describe the temperature evolution in Equation 5.4.

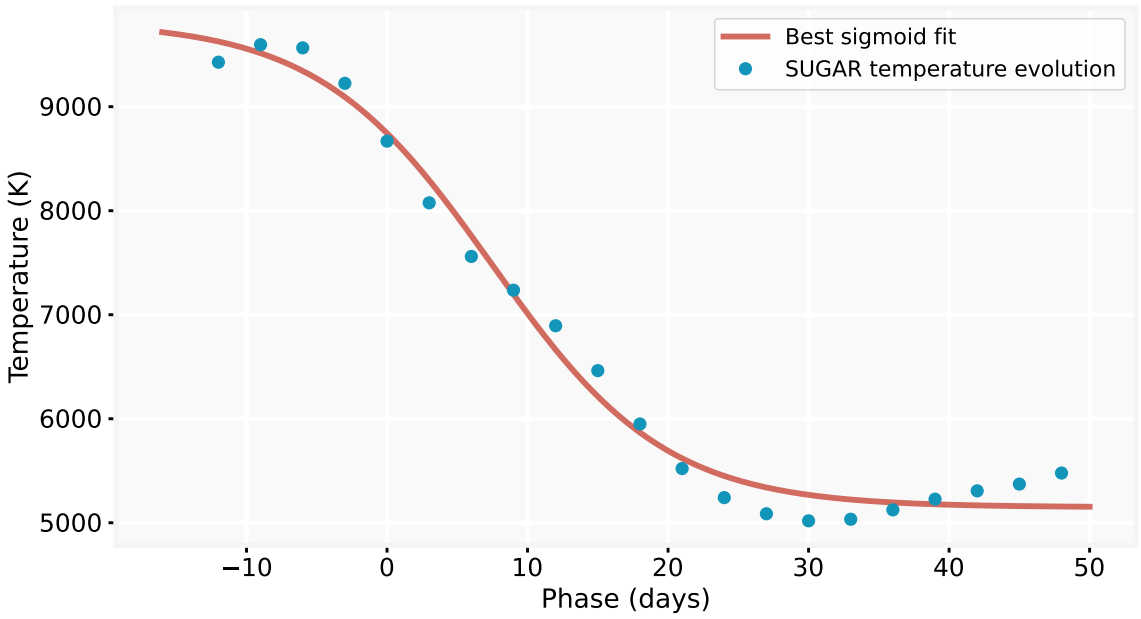



FIGURE 5.2 – Best logistic function fit of the temperature evolution extrapolated from the SUGAR model. The data points have been computed by fitting blackbodies to the SNIa spectra from 3250 to 8650 Å at regular time intervals of 3 days. The phase corresponds to the time since maximum brightness.

### 5.1.3 Feature extraction

As a preprocessing step, each light curve was normalized by the maximum flux measured in LSST-$r$ passband (for the simulated data described in Section 5.2) or ZTF-$r$ (for the real data as described in Section 5.4), which ensures a more uniform dataset. Then, we fit the light curves and extract the resulting best-fit parameters. The minimization step was performed using the least-squares method from the IMINUIT `Python` library [5] (James and Roos, 1975). Table 5.1 summarizes the initial guesses and the parameter bounds used at the minimization step (for the MONOCHROMATIC method, the same choices were applied to the relevant parameters). Regarding units, $t_0$, $\tau_{rise}$, $\tau_{fall}$, and $\tau_{temp}$ are expressed in days, while $T_{min}$ and $\Delta T$ are expressed in kelvin. We report an average run time of 120 ms per fit [6] for the SNe from the PLAsTiCC dataset using the cuts described in Section 5.2.

| | $a$ | $t_0$ | $\tau_{rise}$ | $\tau_{fall}$ | $T_{min}$ | $\Delta T$ | $\tau_{temp}$ |
|---|---|---|---|---|---|---|---|
| Initial guess | *peak* | *peak_time* | $-5$ | 30 | $4 \cdot 10^3$ | $7 \cdot 10^3$ | 4 |
| Upper bound | $10^3$ | $10^3$ | $-0.5$ | 500 | $10^5$ | $10^5$ | 300 |
| Lower bound | 0.1 | $-10^3$ | $-100$ | 0.5 | 100 | 0 | 1.5 |

TABLE 5.1 – Initial guess, lower and upper bounds of the parameters for the minimization step. *peak* is the maximum measured flux over all filters and *peak_time* corresponds to the time of *peak*

5. https://iminuit.readthedocs.io/en/stable/
6. using the RAINBOW function from the public *light-curve* package: https://github.com/light-curve/light-curve-python

When using the Monochromatic approach, we performed a Bazin fit (Equation 3.2) for each filter separately. Given the functional form, we extracted four parameters per passband. Additionally, we collected the least-squares loss as a feature for each passband. The $r$-band flux normalization factor computed previously was also included. Therefore, for a given dataset, we extracted $n_{features} = 5 \times n_{passbands} + 1$ features for each object. Note that in this configuration, if any passband does not contain enough observation to be fitted, the object is discarded from the analysis (see Section 5.2 below).

In the context of the Rainbow method, we performed a single fit of all the filters at the same time using Equation 5.4, resulting in seven best-fit parameters. Additionally, we kept the least-square loss and the normalizing factor. Therefore, we extracted nine features for each object, independently of the dataset considered.

## 5.2 Data

We performed benchmark tests on simulated data from the PLAsTiCC (Hložek et al., 2020) dataset (Section 2.3.4). Each astronomical source was represented by a noisy and inhomogeneously sampled light curve in six different filters[7]: $u$, $g$, $r$, $i$, $z$, and $Y$. The complete dataset consisted of around 3.5 million light curves, representing 17 classes that encompassed both transient and variable objects. In principle, Rainbow could handle any number of passbands, but increasing the number of filters used in the analysis results in greater chances of objects being insufficiently sampled in at least one passband, which implies that this is not suited for the Monochromatic method and thus prevents a coherent comparison of the two paradigms. Therefore, we chose to work with three passbands and discarded $u$, $z$ and $Y$, for which the blackbody approximation of SN is least valid (Pierel et al., 2018).

We selected every well-populated transient types within PLAsTiCC, namely SNIa, SNII, SNIbc, SLSN, and TDEs. The first three SN types were used for all analyses, while TDE and SLSN were only added in the classification tests (Section 5.3.3) with the goal of increasing the complexity of the task. We required each passband to hold at least four photometric true detection points (Kessler et al., 2019). This cut is imposed by the Monochromatic method based on the Bazin equation (Equation 3.2). It is the minimum requirement for the minimization to be reasonable and was therefore applied to both methods. Furthermore, we required that the time of peak luminosity (defined as PEAKMJD inside the PLAsTiCC metadata) was included within the time span of the light curve. Note that the minimum requirement for Rainbow (at least seven true detection points over all passbands) is always true if the Monochromatic requirement is verified.

We used only PLASTiCC data from the Wide Fast Deep (WFD) survey strategy. It represents most of the objects from the simulation and is less well sampled than the Deep Drilling Field (DDF) strategy, making it a perfect test case for sparsely sampled data. The requirements described above are very selective considering the cadence of the dataset, and only 0.4% of the SN objects passed the cuts. This highlights the importance of reducing the required number of points per band, which would increase the number of objects considered in the analysis. We

7. https://www.lsst.org/about/camera/features

performed feature extraction for objects within the WFD sample and stopped when we reached 1000 objects of each type or until we ran out of objects. Following this procedure, we built a database made of 1000 SNIa, 1000 SNII, 468 SNIbc, 1000 SLSN, and 372 TDE.

Additionally, we produced a second, more challenging database using only the rising part of the light curves. We proceeded by removing all points after PEAKMJD, which corresponds either to the time of maximum flux in the $B$ passband in the emitter frame for SNIa, or to the bolometric flux for the other transients [8]. We maintained the same feature extraction method as previously, including the requirement of the minimum number of points. This resulted in a smaller database containing 626 SNIa, 532 SNII, 317 Ibc, 616 SLSN, and 269 TDE.

## 5.3 Results

We evaluated the performance of Rainbow through a comparison against the standard Monochromatic method. This analysis comprises both direct and indirect evaluations. The direct method consists of a measurement of the quality of the light curve reconstruction (Section 5.3.1). However, since the reconstruction quality metric does not necessarily reflect the information loss, we also performed additional indirect tests: the prediction of the peak time (Section 5.3.2), and two classification exercises (Section 5.3.3).

### 5.3.1 Quality of fit

This test was performed independently on all selected SNIa, SNII, and SNIbc following the method proposed in Demianenko et al. (2023). The light curve time span was divided into 5-day-long bins. We randomly selected two bins containing at least one true detection point each, from which all photometric points were removed. This effectively resulted in two random 5-day-long gaps in the light curve. After this step, if enough points remained to fulfill the minimum requirement, they were submitted to the two fitting procedures (Rainbow and Monochromatic). The removed points were used to compute the agreement between the estimated light curve and the measured values. We used the normalized root mean squared error based on observed error (nRMSEo) as a quality metric,

$$\text{nRMSEo} = \sqrt{\frac{1}{m}\sum_i \left[\frac{(y_i - \mu(t_i))^2}{2\epsilon_i^2}\right]}, \tag{5.8}$$

where $t_i$ is the time of measurement, $y_i$ is the observed flux, $\epsilon_i$ is the observed flux error, $\mu(t_i)$ is the estimated flux at the time $t_i$, and $m$ is the number of observations inside the test sample.

Figure 5.3 illustrates the procedure and the resulting fits on a given SNIa example for both methods. It highlights that Rainbow is resilient to information loss. One point removed in the $i$-filter carries much information regarding the maximum flux of the object. In the $i$-filter, the Monochromatic method misses the peak observation by 20 days with half of the intensity, while Rainbow exploits the information brought by the $r$ passband and produces a more realistic light curve.

8. Richard Kessler, private communication

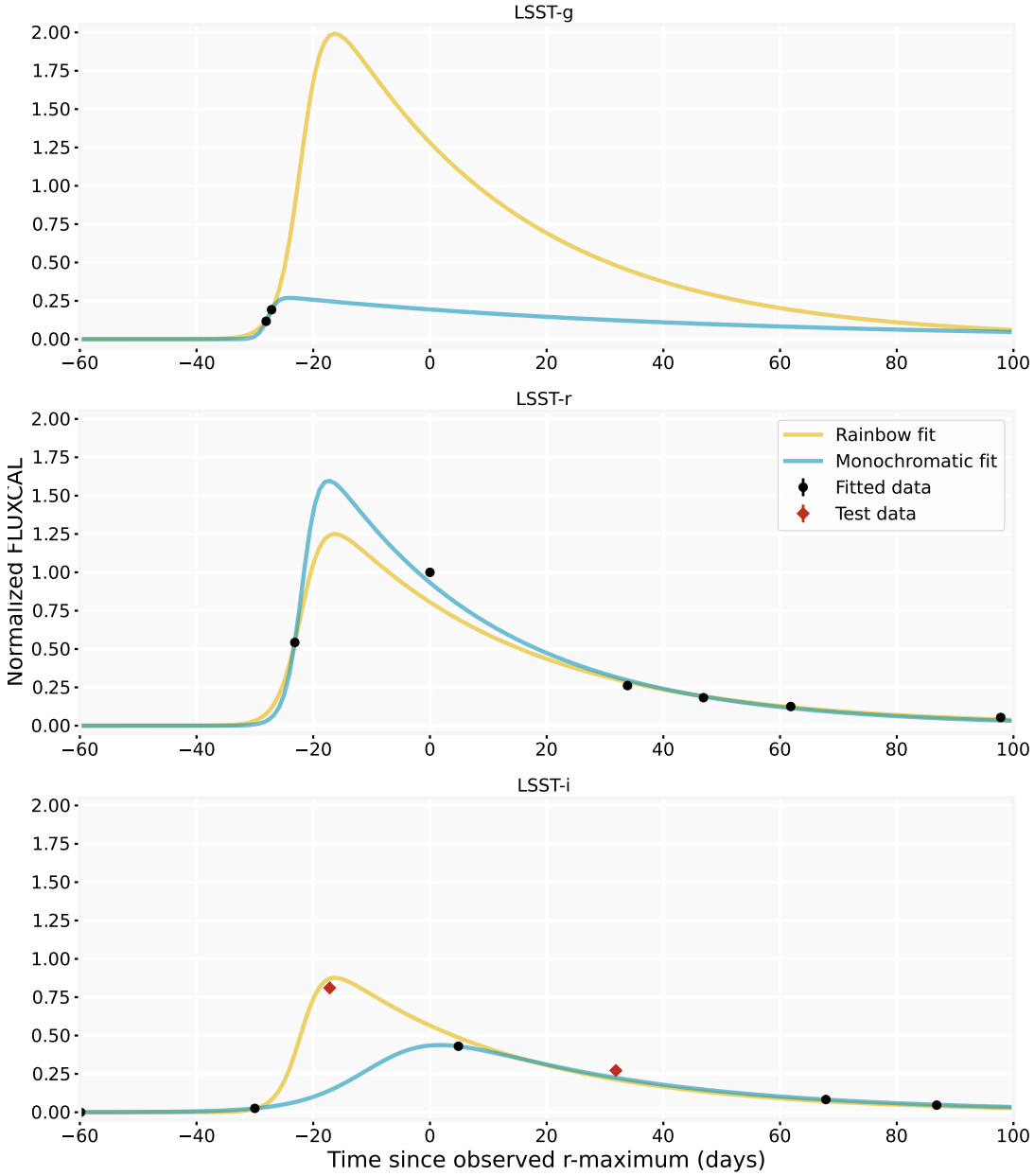



FIGURE 5.3 – Quality of the light curve fit to a PLAsTiCC SNIa light curve ($z = 0.07$). The red diamonds represent points that were randomly removed and were only used to compute the nRMSEo error. The fits were performed considering only points shown as dark blue circles. The error bars are displayed, but are contained within the points.

Table 5.2 displays the median nRMSEo error for the different classes of SNe. It also provides the 25th and 75th percentiles. The results are presented for each separate class of SN, along with the global result (All SNe). In order to facilitate comparison, we estimated, after visual inspection, that an nRMSEo error below 7 can reliably be considered a correct fit for SN-like light curves (the theoretical perfect fit of this metric is $nRMSEo = \frac{1}{\sqrt{2}}$). Rainbow provides a better median error of 4.7 overall, against 5.6 for the Monochromatic method. The improvement in the fit quality is clear for SNIa and SNIbc, while both methods perform similarly for SNII.

| | SNIa | SNII | SNIbc | All SNe |
|---|---|---|---|---|
| Monochromatic | $7.9^{[20.3]}_{[3.3]}$ | $3.7^{[12.2]}_{[1.4]}$ | $4.9^{[12.5]}_{[1.7]}$ | $5.6^{[15.2]}_{[2.1]}$ |
| Rainbow | $6.8^{[14.3]}_{[3.4]}$ | $3.9^{[8.0]}_{[2.0]}$ | $2.8^{[6.9]}_{[1.5]}$ | $4.7^{[10.3]}_{[2.3]}$ |

Table 5.2 – Median nRMSEo (Equation 5.8) for the Monochromatic method and Rainbow. The lower and upper values represent the 25th and 75th percentiles, respectively.

We note that the 25th percentile of errors tends to be higher for the Rainbow method. This means that the very best fits are more often produced by the Monochromatic method. This result is expected since good fits are associated with well-sampled data, and we know that the blackbody assumption is only a first-order approximation. Therefore, when many data points are available, it can act as a constraint on the fit. To the opposite, the 75th percentiles show much lower error values of the Rainbow method overall, which indicates that it produces less extremely incorrect light curves than the Monochromatic method. This was confirmed by computing the mean nRMSEo error (less resilient to outliers), for which we obtained 8.8 and 13.1 for the Rainbow and the Monochromatic method, respectively.

### 5.3.2 Peak time prediction

In this section, we consider a regression task aimed to estimate the time of maximum flux as an indirect approach to evaluate the quality of the light curve reconstructions. PLAsTiCC metadata contain the value PEAKMJD corresponding to the time of the peak flux. It is defined by the time of maximum flux in the $B$ Bessel passband (Thomson, 1949) in the source rest frame for SNIa (in that case, we used PLAsTiCC redshift metadata), or in received bolometric flux for every other transient. We predict this value using a direct and an indirect method.

The direct-prediction method highlights the versatility of Rainbow. The 2D fit function (Equation 5.4) gives access to the flux in any passband at any time. Additionally, Rainbow provides a direct estimate of the bolometric flux of the object. Therefore, in the case of SNIa, we can directly compute the time of maximum flux in the $B$ passband even when no measurements were taken at this wavelength. For the other transients, we can use the bolometric flux explicitly encoded in the equation. In this analysis, we predicted the PEAKMJD for SNIa, SNII, and SNIbc. Figure 5.4 presents the distribution in the difference between the prediction and the reported peak time per class. We fit the distribution with a Gaussian to obtain a mean and a standard deviation. The method provides unbiased results for SNII. The SNIa and SNIbc predictions are biased around 3 days before and after the peak, respectively. Nevertheless, an error

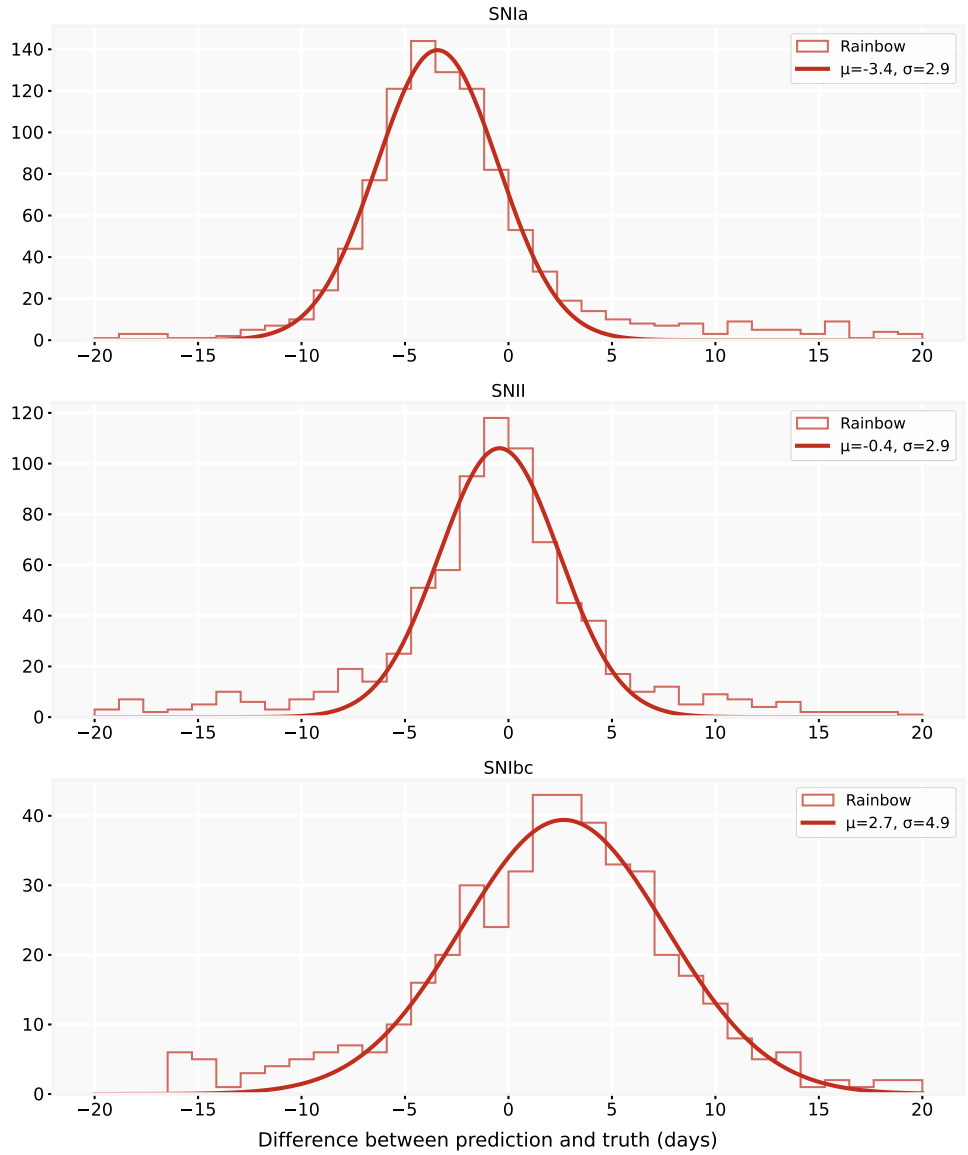



FIGURE 5.4 – Distribution per class of SNe of the difference between the RAINBOW prediction and the reported time of maximum, as given within the PLAsTiCC metadata. The RAINBOW prediction is directly computed from the light curve estimate using the definition of PLAsTiCC PEAKMJD. Additionally, a Gaussian is fitted to the distributions to evaluate its mean and standard deviation.

of 3 days constitutes a small error for a peak prediction of the considered SNe classes. The results show a small spread overall, especially for SNIa and SNII, with a standard deviation of 2.9 days. This method has the advantage of offering a direct computation of the time of maximum, but it can only be computed for the RAINBOW method. Indeed, the MONOCHROMATIC feature extraction procedure does not provide information about the bolometric nor the $B$ passband flux.

For a fair comparison of the quality of prediction, we computed a second indirect metric using the scikit-learn linear regression [9] algorithm trained on the features extracted from both methods. This simple machine learning method enables the training of a model capable of predic-

9. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

ting a numerical value (here the peak time) based on input data (here the features extracted). We evaluated the performances on 100 iterations of bootstrapping (see Section 4.3 and Efron, 1979), which is sufficient to produce robust results (Raschka, 2018). Figure 5.5 displays histograms of all bootstrapping predictions combined for each SN class. Additionally, we fit a Gaussian to the distribution result of each bootstrapping iteration, from which we extracted the standard deviation and the mean. The Gaussians presented in Figure 5.5 represent the mean Gaussian of the bootstrapping iterations. The colored area around the Gaussians represent a $1\sigma$ deviation of the bootstrapping standard deviation distribution.

Results show the excellent predictive power of the Rainbow features for SNIa, with a mean of 0.8 and a standard deviation of 2.3 days (Figure 5.5, top panel). Features from the Mono-

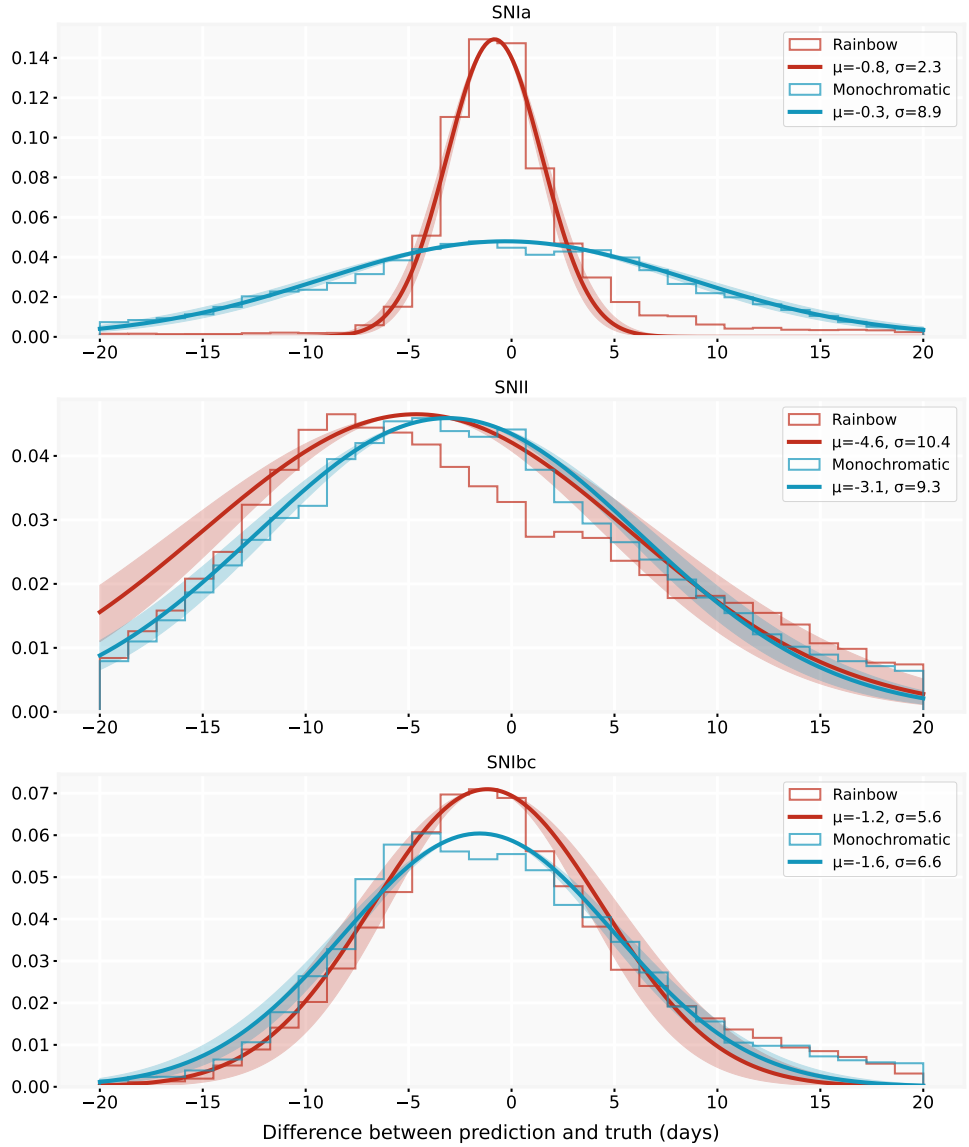



Figure 5.5 – Distribution per class of SNe of the difference between the prediction and true time of maximum as given within the PLAsTiCC metadata. The predictions have been computed based on linear regressor models trained with Rainbow (red) and Monochromatic (blue) features. Additionally, a Gaussian was fitted to the distributions to evaluate their mean and standard deviation.

CHROMATIC approach resulted in a very wide distribution, with a standard deviation of 8.9 days. In the case of SNII, both sets of features result in widely spread predictions. For this particular type, using the direct RAINBOW prediction provides a much better indicator for the time of the flux maximum. For SNIbc, results for the two methods are very similar, with a slight improvement in the standard deviation for the linear regression trained with the RAINBOW features.

### 5.3.3 Classification

A common way to use features extracted from light curves is in ML classification tasks (Section 4.2.2). An informative set of features provides a good summary description of an object and can be used to distinguish several types of classes. We performed a multiclass classification exercise as a way to measure the relative information quality of the features extracted using the RAINBOW method compared to the those resulting from the MONOCHROMATIC procedure. A grid search over the `max_features` and `max_depth` hyperparameters led to no significant improvement of either model over the default hyperparameters. Therefore, we left all hyperparameters at default values. For the classification algorithm, we used the random forest implementation from the `Sklearn` library [10] (Pedregosa et al., 2011). We decided to perform two classification tasks, one with the full light curves, and the other using only the rising part.

For the first exercise, we subsampled the database from Section 5.2 and built a balanced dataset of 300 light curves of each class (SNIa, SNII, SNIb, SLSN, and TDE), thus ensuring a uniform representation of the different light curve classes. We evaluated the results with 100 iterations of bootstrapping. Figure 5.6 displays the differences in median confusion matrix between the RAINBOW features and the *Monochromatic* features. The overall median accuracy values are 88.4% and 81.9%, respectively (see Appendix A, Figures A.1, and A.2 for the individual confusion matrices). Results clearly show that features extracted with the RAINBOW method enclose more discriminative information. Not only the overall accuracy is better, but every single type of transient is more accurately classified. We note that the long-lasting transients, i.e. SNII, SLSN, and TDE, display the largest improvement over the MONOCHROMATIC method.

Figure 5.7 presents the importance of each RAINBOW feature in the classification process. The second and third most important features for separating the different classes are $t_{\rm fall}$ and $t_{\rm rise}$. They come from the bolometric function and are predictably key in the description of a transient type. The most relevant feature is $T_{\rm min}$, which describes the minimum temperature that the object will reach after the event. This underlines the importance of the blackbody approximation within the RAINBOW model.

In order to quantify the difference of the results between the two classifiers, we performed a corrected McNemar (McNemar, 1947) test as proposed in Yu and Liu (2003). From this test, we computed a $\chi^2$ that gives the degree of certainty with which we can reject the null hypotheses, i.e. that the resulting differences are due to random chance and that the two classifiers perform equally well.

10. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

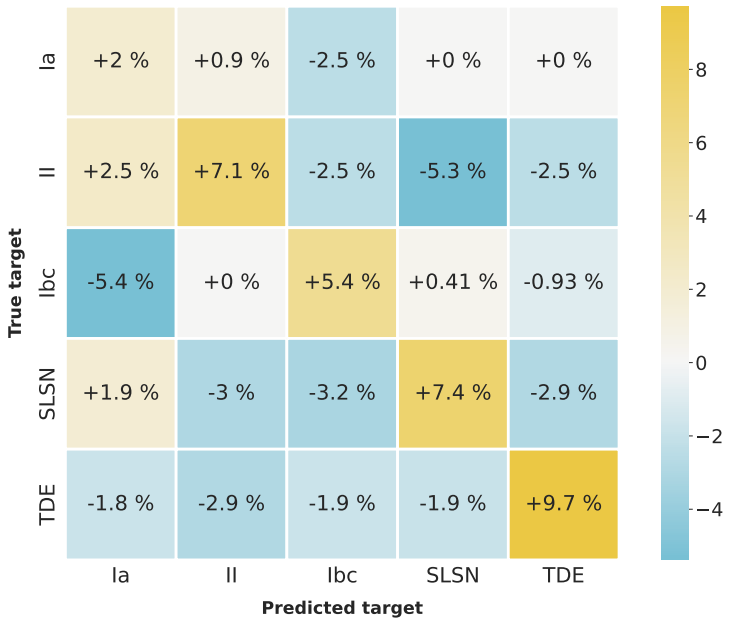



FIGURE 5.6 – Confusion matrix difference between the random forest classifiers trained on RAINBOW and MONOCHROMATIC features (normalized on purity). The dataset is composed of 300 light curves of each class (SNIa, SNII, SNIbc, SLSN, and TDE). The numbers represent the difference (RAINBOW - MONOCHROMATIC) in the median score of 100 iterations of bootstrapping. Individual confusion matrices for each method are given in Appendix A (Figures A.1 and A.2).

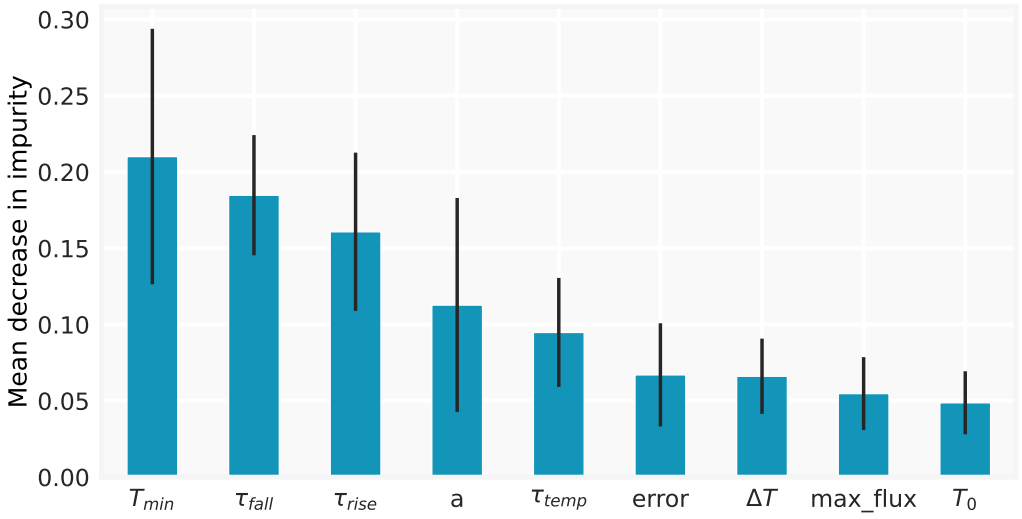



FIGURE 5.7 – Mean feature importance over 100 bootstrapping iterations of the random forest classifier trained with RAINBOW features. The black lines represent the standard deviation of the scores over the 100 bootstrapping iterations.

We computed a 2x2 matrix (Table 5.3) that compares the two model predictions to each other (which should not be confused with a classical confusion matrix result from a given model). This matrix was computed at each bootstrapping step, and the table displays the average over the 100 iterations. The meaningful squares are the top right (TR) and bottom left (BL) since they displayed all the cases where the classifiers output different answers. From this we computed the $\chi^2$ metric as

| | Correct Rainbow | Incorrect Rainbow |
|---|---|---|
| Correct Monochromatic | 416 | 35 |
| Incorrect Monochromatic | 71 | 29 |

Table 5.3 – Mean contingency table of objects correctly and incorrectly classified by the random forest models trained with Rainbow and Monochromatic features over 100 bootstrapping iterations.

$$\chi^2 = \frac{(|TR-BL|-1)^2}{TR+BL}. \tag{5.9}$$

After the computation, we obtained a $\chi^2 = 13$, which for one degree of freedom results in a p-value of $3 \cdot 10^{-4}$. This result shows with a confidence higher than $3\sigma$ that the resulting differences are not due to random chance.

These classification results displayed good performances of the Rainbow method on complete light curves, but in some cases, an early classification of a still rising transient might be required in order to decide about a telescope follow-up (e.g., Leoni, M. et al., 2022). Thus, we evaluated the Rainbow method by repeating the classification exercise in this scenario. From the rising light curve database described in Section 5.2, we randomly built a balanced dataset of 250 of each class (SNIa, SNII, SNIb, SLSN, and TDE). The number of objects per class must be reduced to 250 to maintain a balanced dataset. As previously, we applied 100 iterations of bootstrapping and used a random forest algorithm to build a classifier.

Similarly to Section 5.3.3, we computed the differences in median confusion matrix between the Rainbow and the Monochromatic features, as shown in Figure 5.8. The overall median accuracies are 59.5% and 50.9%, respectively (see Appendix A, Figures A.3 and A.4 for the individual confusion matrices). The results again show that Rainbow features lead to a better classification of each type of rising transients. TDE display the largest difference with an accuracy improved by more than 19%. With the exception of SNII, all other rising SNe are clearly better disentangled using the rainbow features. The type SNII include SNII-P events that are characterized by a slowly decreasing plateau phase. In the case of rising light curves, we lose this determinant information, which could explain the decrease in the accuracy differences when compared to full SNII light curves.

Figure 5.9 presents the importance of each Rainbow feature in the classification process of rising light curves. In this scenario, $\tau_{rise}$ is the most important feature, which is coherent given the early classification challenge. Other features, including those related to temperature, provide second-order information and are all equally useful overall. We also performed a McNemar test for this scenario. The average correct and incorrect classification of the two methods are shown in Table 5.4. Applying Equation 5.9, we obtained a $\chi^2 = 10.7$, which is equivalent to a p-value of $1 \times 10^{-3}$. The result is again statistically robust and leads to a confidence higher than $3\sigma$ that the Rainbow method generates more informative features.

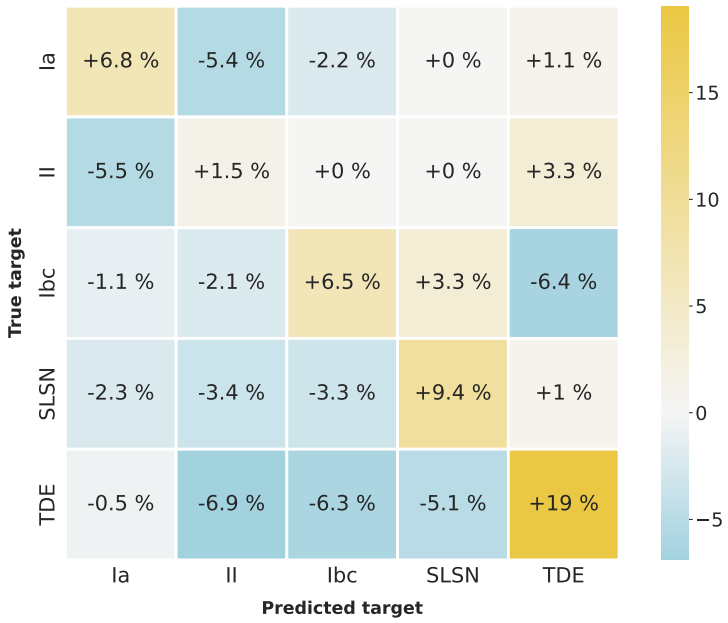



FIGURE 5.8 – Confusion matrix difference between the random forest classifiers trained on RAINBOW features and MONOCHROMATIC features (normalized on purity). The dataset is composed of 250 **rising** light curves of each class (SNIa, SNII, SNIbc, SLSN, and TDE). The numbers represent the difference (RAINBOW - MONOCHROMATIC) in the median score of 100 iterations of bootstrapping. Individual confusion matrices for each method are given in Appendix A (Figures A.3 and A.4).

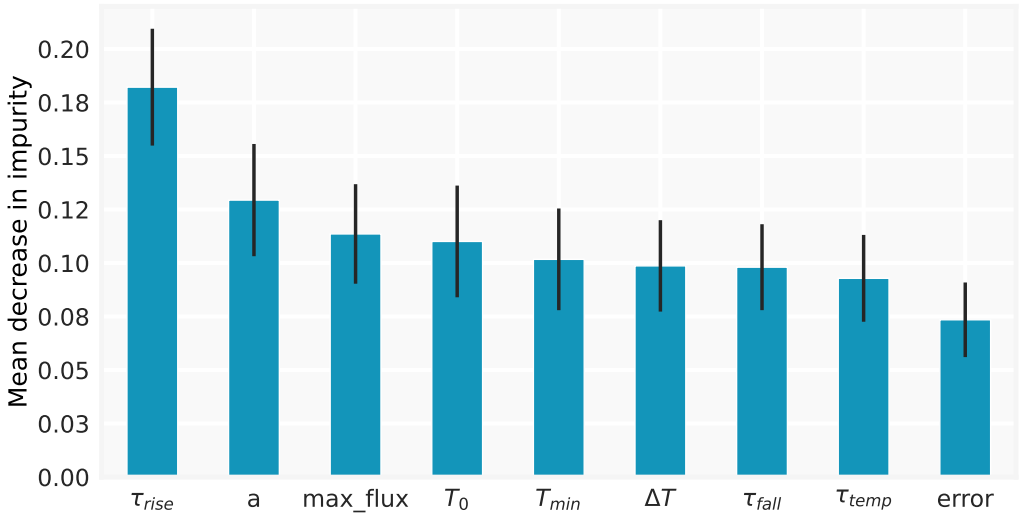



FIGURE 5.9 – Mean feature importance over 100 bootstrapping iterations of the random forest classifier trained with RAINBOW features from rising light curves.

## 5.4 Real data application

In this section, we demonstrate the capabilities of the method when applied to a real data multi-survey scenario, and compare it to the MONOCHROMATIC procedure. We used the Young Supernova Experiment Data Release 1 (YSE DR1, Aleo et al., 2023), which contains final photometry of 1975 transients observed by the Zwicky Transient Facility (ZTF, Bellm et al., 2019b) *gr* and Pan-STARRS1 (PS1, Chambers et al., 2016) *gri* passbands. YSE DR1 is the

| | Correct Rainbow | Incorrect Rainbow |
|---|---|---|
| Correct Monochromatic | 213 | 65 |
| Incorrect Monochromatic | 113 | 148 |

Table 5.4 – Mean contingency table of rising objects correctly and incorrectly classified by the random forest models trained with Rainbow and standard features over 100 bootstrapping iteration.

largest available low-redshift homogeneous multiband dataset of SNe, and thus provides the perfect testing ground for a variety of real objects. It encompasses SNe observations across vast timescales, magnitudes, and redshifts: those that last from a few days to over a year, ranging in apparent magnitudes from 12 mag to 22 mag, and spanning a redshift distribution up to $z \approx 0.5$. PS1 compliments ZTF observations by also observing in $gr$. Additionally, it provides critical $i$ observations to probe red and faint transients typically missed by bluer-sensitive surveys.

Despite adding crucial deep photometry to ZTF data, the PS1 global cadence is well below that of ZTF, making an independent filter fit impossible in many cases due to insufficient number of photometric points. This implies that a Monochromatic feature extraction step will result in a significant loss of objects. This issue can be addressed by considering PS1-$gr$ and ZTF-$gr$ as equivalent, effectively stacking their light curves. This approximation is generally accurate, although small discrepancies can be observed due to differences in passband transmission profiles [11, 12] and differences in photometric pipelines. In this context and in the context of multisurvey analysis in general, Rainbow offers the possibility to perform fits on all data available at once, without any passband approximation.

We compared Rainbow and the Monochromatic method using the evaluation of the quality of the fit described in Section 5.3.1. We used all spectroscopically confirmed SNIa, SNII, and SNIbc, including nondetections, for a total of 254 SNe available in Zenodo [13] (Aleo et al., 2022). Additionally, we manually added two nondetection points in the ZTF-$gr$ passbands at -300 and -400 days before the measured maximum flux time. This provided a flux baseline that helps to constrain the fit, especially when only the falling part of the light curve is available. We show results for two cases: first, where the dataset was entirely used to perform a Rainbow fit (seven parameters per object) ; and second, where the combined PS1-$gr$ and ZTF-$gr$ light curves were fit independently with the Monochromatic method (eight parameters per object). Since PS1-$i$ is poorly sampled (30% of the objects contain fewer than four detection points) and no ZTF-$i$ band is available, we decided to restrict our test to $gr$ passbands. Figure 5.10 illustrates the fits on a SNII (SN 2020thx).

This dataset provides well-sampled light curves with 15 and 19 detection points on average in the $g$ and $r$ passbands from ZTF and PS1 combined, respectively. Since Rainbow is particularly

11. http://svo2.cab.inta-csic.es/svo/theory/fps3/index.php?mode=browse&gname=Palomar&gname2=ZTF&asttype=
12. http://svo2.cab.inta-csic.es/svo/theory/fps3/index.php?mode=browse&gname=PAN-STARRS&asttype=
13. https://zenodo.org/record/7317476

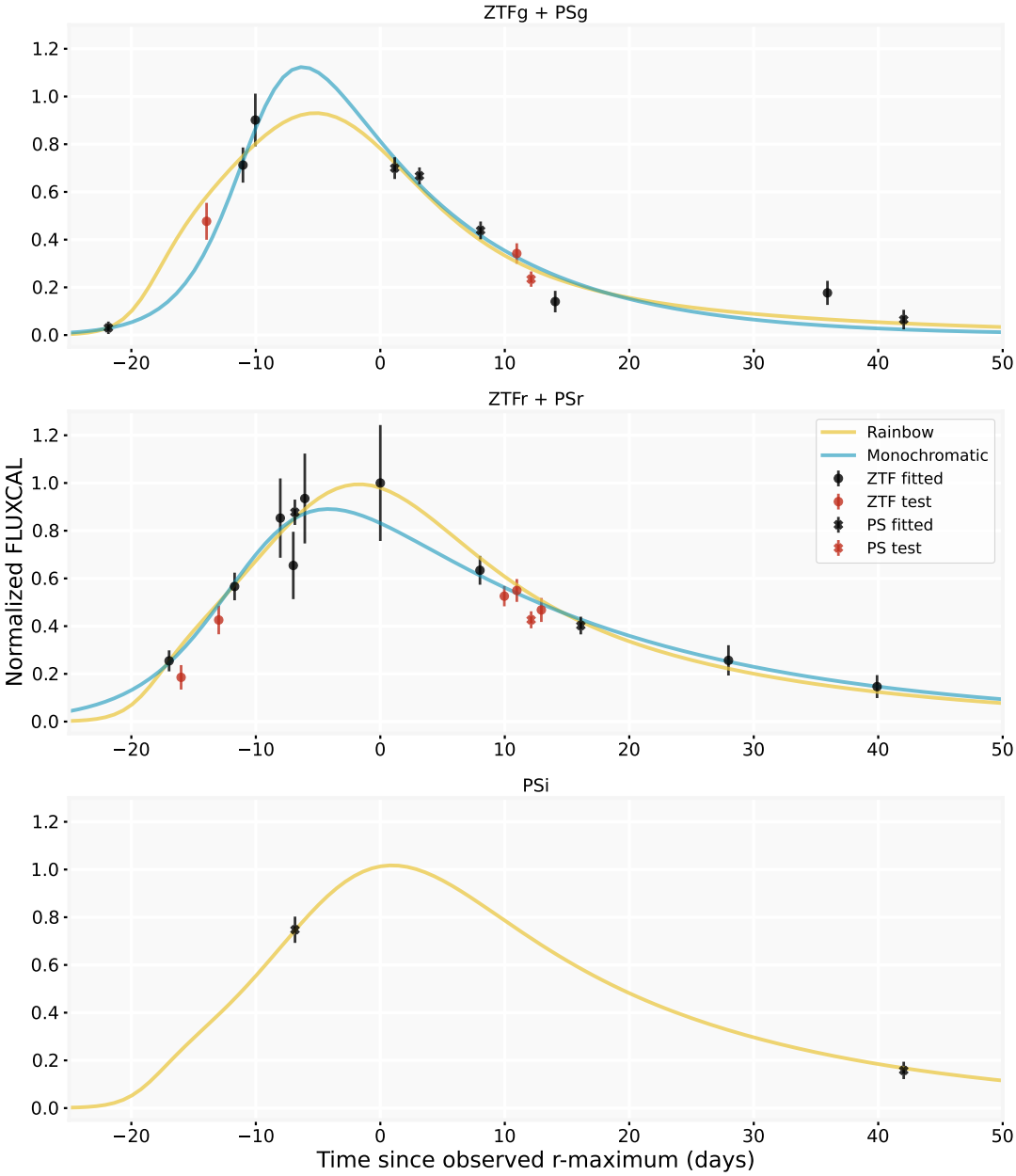



FIGURE 5.10 – Quality-of-fit evaluation process on a YSE DR1 SNII light curve (SN 2020thx) sampled at 50%. The red points were randomly removed and used to compute the nRMSEo error. The fits were performed only considering the dark points.

efficient at compensating the lack of information from poorly sampled passbands, we expect both methods to perform equally well in this scenario. Therefore, we compared the methods for different sampling levels. We created eight datasets by randomly sampling from 30% to 100% of the points in steps of 10%. In order to provide uncertainties, we performed this procedure ten times with different seeds. Since we still require a minimum number of points per passband (see Section 5.2), the datasets contained from 123 SNe for 30% sampling to 254 SNe for the complete sample on average. We display the average median nRMSEo error (Equation 5.8) per sampling level in Figure 5.11. The error bars correspond to one standard deviation of the median error over the ten seeds. The area of each point is proportional to the mean number of SNe in each sample. RAINBOW is clearly superior in all situations where the number of points is limited. However,

both methods are equivalent in terms of predicting the light curve profile when using 90% or more of the original number of points. This result indicates that the small additional information provided by PS-$i$ together with the physical assumptions of the Rainbow framework make a crucial difference on sparse transient datasets.

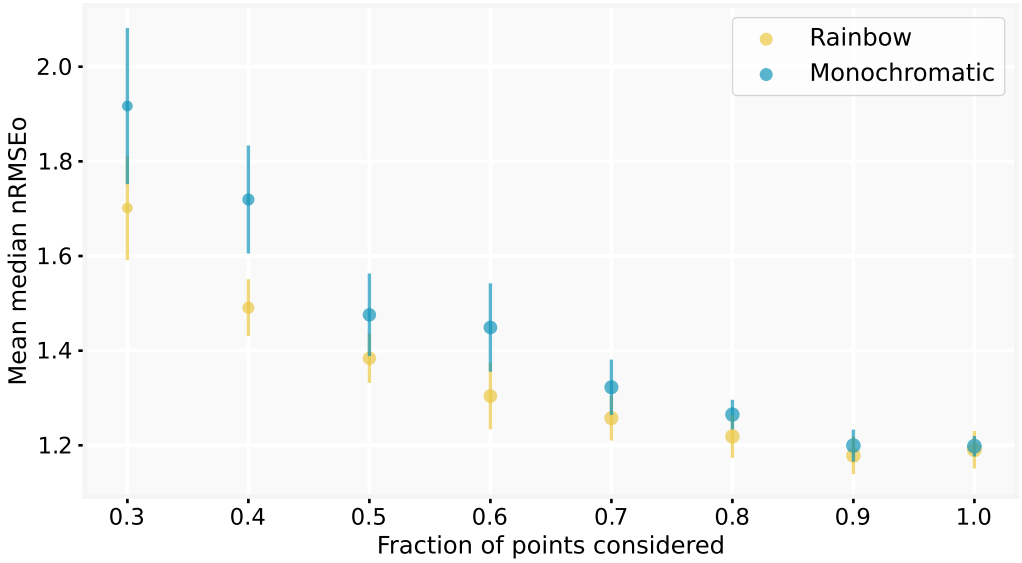



Figure 5.11 – Evolution of the median nRMSEo error for different sizes of the subsample of the YSE dataset. The error bars represent one standard deviation of the ten random subsamples. The data point surface is proportional to the mean number of SNe in the samples, from 123 for 30% to 25 for 100%.

## 5.5 Conclusion

Estimating a continuous light curve behavior from sparse and noisy photometric measurements is of crucial importance for the study of astronomical transients. The Rainbow framework presented here enables the fit of a continuous 2D surface in time and wavelength (Appendix 5.1) even when the light curves are sparsely sampled in a number of different passbands. It starts by assuming that the thermal-electromagnetic behavior of the transient can be approximated by a blackbody, and it incorporates user-defined parametric functions representing the temperature evolution and bolometric light curve behavior. As a result, it uses information in the available bands to inform the reconstruction in other wavelengths, providing a simple and robust framework for light curve estimation and feature extraction, which is especially adapted when the data are sparse or spread across a set of passbands.

We used simulated data from PLAsTiCC to demonstrate the effectiveness of the method in a series of tests: goodness of fit (Section 5.3.1), estimation of the time of peak brightness (Section 5.3.2), and using the best-fit parameter values as input to machine-learning-based classifications (Section 5.3.3). In all these, we compared Rainbow to the more traditional approach of fitting a parametric function independently to each passband (the Monochromatic method). The results show that Rainbow outperforms or equals the results obtained from the Monochromatic method, with the advantage of being applicable to significantly more sparse light curves.

Nevertheless, the method also inherits the drawbacks of the blackbody assumption, which may not be suitable for specific science cases. For example, the energy distribution of a supernova is far from being a blackbody in the ultraviolet and infrared spectrum (e.g., Faran et al. 2018). The Rainbow method should be used in all wavelengths and epochs for which the blackbody assumption holds. However, in the task of classification, using a Rainbow fit on a non-blackbody source can be beneficial, as the poor fit score will indicate that the hypothesis does not hold for this event.

Caution should also be applied for very well sampled light curves. Even when the blackbody hypothesis is valid, it remains an approximation to physically describe the gaps in the data. In the case of high-cadence measurements, it acts as a constraint that can erase important information. Whenever a large number of data points is available, allowing the necessary number of parameters to be fit by independently fitting a parametric function to each passband will result in a better light curve approximation [14]. We demonstrated this issue using real data from YSE DR1 (Section 5.4). The results showed that Rainbow provides significantly more informative reconstructions for a small number of observations. Finally, the blackbody spectrum approach presented in this paper can be generalized to any parametric SED model. For instance, power-law SEDs are commonly observed in various types of objects: those populating synchrotron emission (e.g., optical afterglows of gamma-ray bursts and UV emission of stellar flares), accretion disk outbursts (e.g., dwarf novae and active galaxy nuclei), and others. For a better performance of the ML pipelines, multiple models based on different spectral parameter evolution laws could be used, while keeping the parameter bound ranges broad enough to cover objects of different astrophysical classes.

In the context of modern astronomical surveys, not only the volume of data will pose an important challenge. The quality and complexity of the data gathered by new surveys will also evolve. LSST will push the limits of detection to even fainter magnitudes in six different passbands, but the majority of the survey strategy (wide-fast-deep) will probably consist of significantly sparse light curves for at least a few of the passbands (Lochner et al., 2018). In this context, Rainbow represents an efficient option to enable modeling and analysis of multidimension light curves. Rainbow has been incorporated into a well-established feature extraction package [15] (Malanchev et al., 2021) that is already used by three different community brokers: ANTARES (Matheson et al., 2021), AMPEL (Nordin et al., 2019), and Fink (Möller et al., 2021b). Thus, it can be immediately used by the community to analyze alert data.

Although Rainbow has proven its efficiency, the question of which optimal parametric model should be used remains. The analysis presented in this chapter used the particular combination of a Bazin and a logistic functions, but any parametric forms could be used instead (Section 3.4). Rainbow offers two degrees of freedom to represent the temperature and bolometric light curve behavior. Hence, the challenge is doubled. The utilization of Symbolic Regression (SR, Section 4.4) could enable an efficient and completely data driven exploration of the potential parametric models. The next chapter shows how the quest for optimal function for light curve fitting lead to the creation of a whole new SR framework, which enables automatic discovery of parametric representations from a set of examples.

14. Note that better reconstruction does not guarantee more informative features.
15. https://github.com/light-curve/light-curve-python

# 6

# Multi-view Symbolic Regression

*The work presented below has been published and presented in the Genetic and Evolutionary Computation Conference (GECCO) 2024* [1]*, as part of this thesis work. This chapter constitutes an adaptation of the original publication, Russeil et al. (2024).*

Rainbow proposes an efficient and physically motivated framework to process multi-passband transient light curves. However, the choices of parametric model describing the bolometric and temperature evolution still remains. As shown in Section 3.4, multiple answers have been proposed in the literature. Although they generally offer good light curve description, they all have been empirically handcrafted to fit the data. It implies that the current parametric models are biased by the expert's preconceived ideas on transient parametric forms. It results in models easily understandable in terms of structure, with directly interpretable parameters, but with no guarantee regarding the fact that they provide an optimal description of a family of transient. In this chapter, we propose to use Symbolic Regression (Section 4.4) to build a completely data-driven method to generate parametric equations.

This idea led to the development of a completely new SR method that we call Multi-View Symbolic Regression (MvSR). The goal is to search for a general parametric model that can simultaneously describe multiple datasets generated by the same underlying mechanism. Hence, it expands the traditional SR, which only aims at describing a specific dataset, to produce a specific description, rather than a general model. This project is by essence interdisciplinary, since its development required discussion and collaboration with computer science experts. Moreover, given its vast applicability, MvSR extends beyond the particular problem of light curves description for which it was initially created. It constitutes a general tool to discover parametric models, independently of the data origin. As a consequence, it led to interdisciplinary collaborations to acquire data and expertise from other scientific fields, demonstrating the general purpose of this tool. This chapter maintains this spirit, and applications on chemistry and economy datasets are proposed in addition to the transient light curve use case.

Section 6.1 describes the adaptation of the SR algorithm. In Section 6.2, we describe the experimental methods used to assess the validity of our approach. Section 6.3 shows the obtained results, followed by a discussion. In Section 6.4 we apply MvSR to real scientific datasets from different fields and discuss the functional forms proposed. Finally, in Section 6.5 we conclude with some final thoughts and future perspectives.

## 6.1 Multi-View Symbolic Regression

The core of our modern scientific knowledge is based on the careful production and analysis of experimental data. In a traditional scenario, the researcher's task is to give meaning to measurable results, and analyze them in the light of a hypothesis. Frequently, the goal of this exercise is to find a mathematical description which can, at the same time, describe the recorded outcomes according to the state of the art of the field, and predict results from future similar experiments. Scientists became experts in the highly non-linear thought process required to interpret and translate scientific knowledge into suitable mathematical expressions. Due to the increase in data complexity, this task has also been approached through a machine learning

1. https://gecco-2024.sigevo.org/HomePage

perspective with the goal of automating the thought process.

Among them, Symbolic Regression (Section 4.4) produces a mathematical expression with one or more variables that optimally fits a given dataset. This method has been successfully applied to simulated scientific datasets in physics (Cranmer et al., 2020), chemistry (Hernandez et al., 2019), medicine (La Cava et al., 2023) and social sciences (de França et al., 2023), to cite a few. In many real scientific applications, the researcher is faced with different sources of data describing the same model but acquired from different setups. This is particularly the case for observational sciences such as astronomy, for which no two objects are studied under the same exact conditions. In practice, this means that we have more data available to fit the model but, though they share the same functional structure, they may differ in parameter values.

In this chapter, the idea of Multi-View Symbolic Regression (MvSR) is introduced. It allows the practitioner to use information from multiple sources and control the desired number of parameters to guide the search process towards a parametric model that contains the correct number of parameters. Thus, it is neither too flexible (i.e., a universal approximator) nor too rigid (i.e., fitting only a single data source). This stimulates the discovery of scientific models, which facilitates further analysis and interpretation of the phenomena. Specifically, we propose an adaptation to the Operon (Burlacu et al., 2020) software package. It calculates the quality of a candidate solution by calculating its score individually on each data source, and then returning an aggregated score over all sources.

The goal of MvSR can be summarized through the functional forms used to model transient light curves (Section 4.1). Even though transients of the same type are collectively described by the same dynamics, the physical properties of the object, its exact brightening mechanism, the redshift, the distance or the filter of observation will produce a whole range of behavior. It results that they all behave following a specific $f(x;\theta)$, but with different values of $\theta$. We argue that the combination of multiple single examples (hereafter single-view) can help to constrain the search space of hypothesis.

### 6.1.1 Algorithm

In summary, MvSR finds a model $f(x;\theta)$ minimizing the aggregated error when fitting the parameters $\theta$ independently for each experiment, while applying a constraint on the number of parameters. More formally, given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{p}$ with $p$ data points, $x_i \in \mathcal{R}^d, y_i \in \mathcal{R}$, SR seeks a parametric model $f(x;\theta)$ that minimizes a loss function $\mathcal{L}$ given the optimal parameter values $\theta$:

$$\min_{f,\theta} \mathcal{L}(f(x;\theta), y).$$

However, with MvSR using $k$ datasets, the objective becomes:

$$\min_{f} agg_{i=1..k}\left(\min_{\theta^i} \mathcal{L}(f(x^i;\theta^i), y^i\right),$$

where the superscript $i$ refers to the index of each dataset, $agg$ is an aggregation function such as $max, avg, med$. For example, $max$ implies that the worst fit among all views will be used

as the final score, favoring solutions that are minimally good for every view. However, *avg* will compute the mean score of all fits, which implies that solutions providing excellent fit on some views, but poor description on others may be selected. Additionally, we impose the constraints

$$x \in \mathbb{R}^m$$
$$\theta \in \mathbb{R}^n$$

where $m$ is the number of independent variables ($x$), $n$ is the number of parameters in the model ($\theta$) with $n$ being bounded by a finite subset of the natural numbers.

Overall, given $k$ different datasets, we want to find the function $f$ with a limited number of parameters that minimizes the aggregated value of $\mathcal{L}$ for each dataset when independently adjusting the value of $\theta$ for each set. The purpose for these constraints is to find a model that can correctly adjust to the data without underfitting nor overfitting, while containing the smallest number of parameters that summarizes each dataset. A complete and ideal implementation of MvSR should allow to:

1. Control the maximum number of parameters.
2. Allow reuse of parameters in the symbolic expression (i.e., $\theta_0$ can appear multiple times).
3. Receive multiple datasets as input.
4. Optimize parameters independently for each dataset.
5. Use an aggregation function to compute a global loss.
6. Penalize solutions based on the number of parameters used.

In this work, we present a version of MvSR where only points 3, 4 and 5 of the list above have been fulfilled. The other points remain crucial for an optimal model generation, especially for practical scientific use, however as a first proof of concept we chose to implement only the minimal requirements. Our current implementation (detailed in Section 6.1.2) slightly modifies a pre-existing Symbolic Regression algorithm at the evaluation phase. Given $k$ datasets, the algorithm will:

1. Generate a parametric functional form
2. Find the best-fit parameters which optimize the description of each of the $k$ datasets;
3. Calculate the $k$ losses
4. Combine the losses using the aggregation function.

Subsequently, it will use this result as a metric for reproduction (within the context of genetic programming, GP).

### 6.1.2 Implementation details

In this work we use Operon (Burlacu et al., 2020), a high-performance C++ framework supporting single and multi-objective GP with non-linear optimization of the parameters using the Levenberg-Marquardt algorithm (Levenberg, 1944). This framework was reported to perform

well with respect to the runtime and overall quality of the results (La Cava et al., 2021). The framework also has a Python module counterpart, called PyOperon, which presents all the flexibility of the C++ version.

The implementation was adapted to support the independent evaluation of datasets. The symbolic expression is independently fitted to each of the inputted datasets, such that the expression can have different adjusted parameters for each view from which we evaluate the loss function (mean squared error). The collection of all the losses from all the view is then aggregated using an *aggregation function*. This final value is then used during selection and reproduction stages to calculate the probability of survival.

The available aggregation functions are *average, median, min, max, harmonic mean*. Notice that since Operon always minimizes the fitness, the *min* and *max* aggregation functions represent the best and worst fits, respectively. We use *max* as the aggregate function, which leads the evolution towards the creation of a function $f$ that minimizes the worst fit among the different sources of data. Therefore, the goal is to find the best comprise, so that the function generated is able to fit all views. The best expression found is then converted into a parametric solution using SymPy (Meurer et al., 2017). The expression is simplified before replacing each float by a free parameter and converting it into a python function. This float replacement procedure is imposed by the Operon implementation. It is not ideal and future implementation of MvSR should directly handle parameters.

## 6.2 Experiments

To demonstrate the advantages of MvSR when multiple data sources are available, we devised an experimental design using data artificially created from the same generating function with either different parameters or covering different regions of its domain. We highlight the benefit of using this approach instead of a traditional SR. Beside artificial benchmarks, we applied MvSR to three real-world datasets from different scientific fields, showing that it can be used to generate efficient solutions to various problems.

### 6.2.1 Data generation

For the artificial data, we set up a series of challenging benchmarks based on standards from the SR literature. For this purpose, we chose three generating functions:

$$f_1(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 \tag{6.1}$$

$$f_2(x) = \sin\left(\theta_0 x_0 x_1\right) + \theta_1 (x_2 - \theta_2)^2 + \theta_3 x_3 + x_4 \tag{6.2}$$

$$f_3(x) = \sqrt{\theta_0 x_0^2 + \left(\theta_1 x_1 x_2 - \frac{\theta_2}{(\theta_3 x_1 x_3 + 1)}\right)^2} \tag{6.3}$$

Equation 6.1 ($f_1$) is a third order polynomial function which constitutes a simple one-dimensional case for which results can easily be visualized and interpreted. Equations 6.2 ($f_2$) and 6.3 ($f_3$)

are based on the Friedman functions (Friedman, 1991) [2] for which free parameters have been added by replacing some constant values. They are multidimensional and constitute a more challenging task.

From the parametric functions, we generate examples which individually carries incomplete information about their parent function. For the polynomial function ($f_1$) we test two separate cases: i) each dataset is generated with two parameters set to 0 (see Views 1 to 4 in Table 6.1), ii) each dataset uses the same parameters (see partial view in Table 6.1) but displays only a narrow part of the behavior from which extrapolation of the original function is challenging. For $f_2$ and $f_3$ we will only test the first case where, by setting some coefficients to 0, we mischaracterize the original function with the goal of rebuilding it using the different views. Each of these benchmarks will contain a total of 4 datasets (i.e., views) with the sample sizes fixed for each view. Datasets for $f_1$ consist of 20 points equally spaced $x \in [-2, 2]$. For the partial view scenario, we keep the same number of points, but sample each dataset into evenly spaced domains $x \in [-2, -1], [-1, 0], [0, 1]$ and $[1, 2]$. The dataset for $f_2$ consists of 100 points uniformly distributed on the interval $x_0, x_1, x_2, x_3, x_4 \in [0, 1]$ [3]. The dataset for $f_3$ consists of 100 points uniformly distributed on the intervals $x_0 \in [0, 100]$, $x_1 \in [40\pi, 560\pi]$, $x_2 \in [0, 1]$, $x_3 \in [1, 11]$ [4].

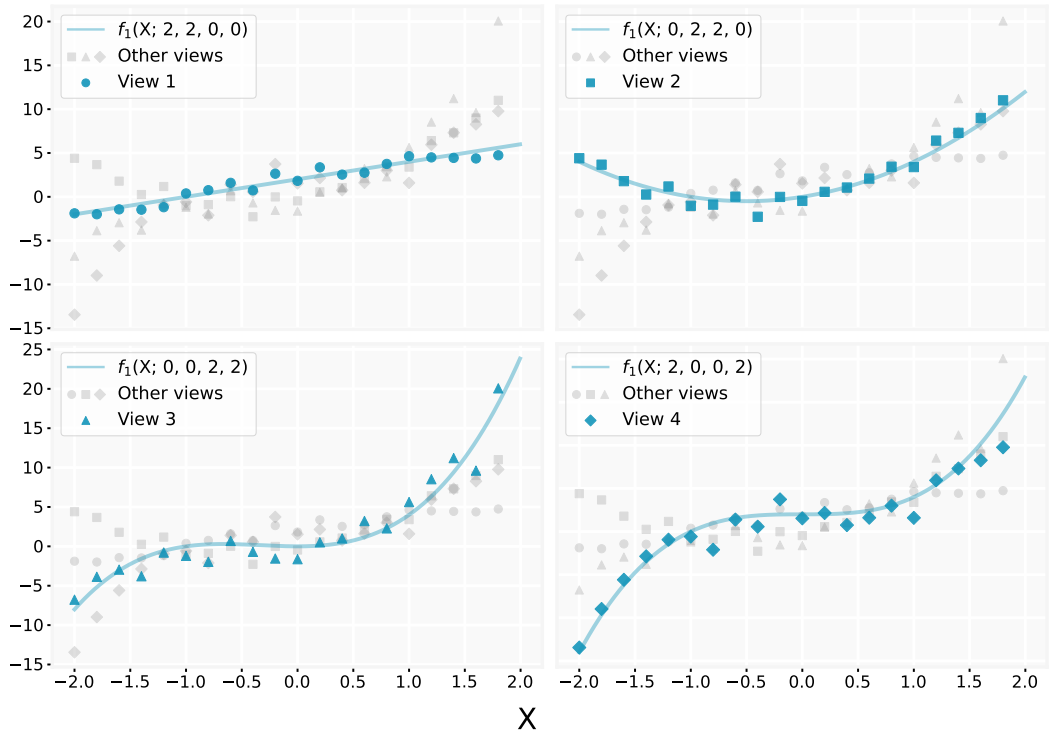
TABLE 6.1 – Parameter values used for each view. In each column, two parameters are set to 0 to depict the extreme situation where they have no effect into the data. The partial view of $f_1$ is also presented. It has no parameters equal to 0 but each example is very restricted in the sampling range.

| | View 1 | View 2 | View 3 | View 4 | Partial view ($f_1$) |
|---|---|---|---|---|---|
| $\theta_0$ | 2 | 0 | 0 | 2 | 2 |
| $\theta_1$ | 2 | 2 | 0 | 0 | $-2$ |
| $\theta_2$ | 0 | 2 | 2 | 0 | 2 |
| $\theta_3$ | 0 | 0 | 2 | 2 | 2 |

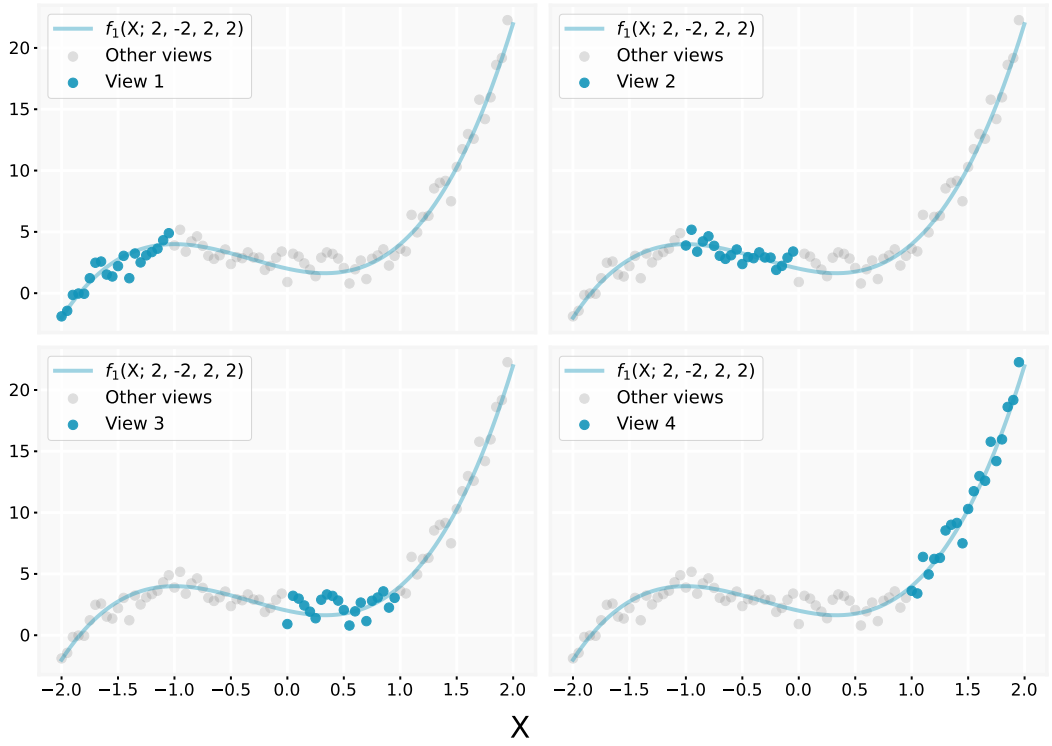
The challenge is visually illustrated with the polynomial function ($f_1$) in Figure 6.1a, which displays the four noisy (see below) views generated using the parameter values from Table 6.1. They all exhibit part of the full behaviors, from which the complete functional form must be extrapolated. Figure 6.1b illustrates the partial domain exercise (last column of Table 6.1). Here, each view exhibits the full behavior, however only a reduced domain is available. It illustrates the difficulty to extrapolate the correct underlying function from noisy datasets. In both cases, MvSR will make use of the four examples at once, while classical SR can only manipulate single datasets.

In order to homogenize results between different benchmarks, we scaled the target variable $y$ of each dataset by applying the transformation $y'_i = 10 \cdot \frac{y_i}{\max(\text{abs}(y_i))}$. The factor 10 was arbitrarily chosen to multiply the absolute error, thus avoiding all scores to be clustered around 0 independently of the fit quality. This procedure introduces an extra scaling parameter for functions $f_2$ and $f_3$, increasing the true number of free parameters to 5. For each generative function,

2. created by the statistician Jerome H. Friedman, not to be mistaken with the physicist Alexander Friedmann
3. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_friedman1.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_friedman2.html




(a) Null parameters scenario with noise level $\tau = 0.066$




(b) Partial domain scenario with noise level $\tau = 0.033$

FIGURE 6.1 – Illustration of the artificial benchmarks generated from $f_1$ (Equation 6.1), using the four first columns (6.1a) and the last column (6.1b) from Table 6.1 as parameter values. The blue lines show the correct generating models. The other views are shown in light gray to illustrate the multi-view nature of the problem.

we create multiple datasets using different noise rates, $\{0.000, 0.033, 0.066, 0.100\}$, in order to verify the robustness of this approach when faced with different noise levels. The noisy target is sampled from the normal distribution $\mathcal{N}(y, \sigma_y\sqrt{\frac{\tau}{1-\tau}})$, where $\tau$ is the noise rate.

### 6.2.2 Operon Hyperparameters and Post-processing

As mentioned in Section 6.1.2, we used an adapted version of pyOperon [5] supporting the use of multiple datasets and aggregation functions. For the following experiments, we used the hyperparameters depicted in Table 6.2. Additionally, we varied the hyperparameter `max_tree_size` from 5 to 25 with increments of 2. This experiment will serve two purposes: i) having a baseline of models simpler than the original generating function since the functions require a minimum size of $7, 11, 14$ to be correctly represented; ii) test whether MvSR is prone to overfitting if given the freedom to expand the expression to larger sizes, thus providing opportunity to fit the noise term as well.

Table 6.2 – List of the fixed hyperparameters used in MvSR experiments.

| **Parameter** | **Value** |
|---|---|
| population size | 1000 |
| number of evaluations | 100000000 |
| pool size | 5 |
| error metric | $MSE$ |
| prob. cx | 1.0 |
| prob. mut. | 0.25 |
| max depth. | 10 |
| optim. iterations | 100 |
| aggregation function | max |
| operators | add, sub, mul, div, square, exp, sqrt, sin ($f_2$ only) |

Each one of the 176 combinations [6] of benchmark functions ($f_1$ partial domains, and $f_1, f_2, f_3$ multiple views), noise level and maximum size represents a single instance of our set of benchmarks. For each experiment, we ran the original pyOperon SR module for each one of the four views independently, as well as the MvSR adaptation on all views at the same time. The procedure was repeated with 100 different random seeds to assess the variance of the results. After each run, the string representation of the best symbolic model is processed with Sympy to replace the numerical values with parameter variables and the corresponding expressions are stored for post-processing. These expressions are individually refitted to the noiseless version of each dataset minimizing the least squares with the python package iminuit (Dembinski and et al., 2020). The final score is calculated using the mean squared error (MSE). Therefore, if

5. https://github.com/heal-research/pyoperon/releases
6. 4 noises × 4 datasets × 11 tree sizes = 176 combinations

the correct expression (or any equivalent) is generated the score will be 0, even for functions generated on noisy datasets.

## 6.3 Artificial Benchmark Results

Figures 6.2 shows heatmap plots of the obtained results for each benchmark function for every combination of noise level and maximum size. In these heatmaps, the color scheme displays the median MSE of each combination, lighter color meaning better results. Any values higher than 5 are depicted as the darkest color in order to keep the contrast in the visualization of the results. In each plot, we show the worst and best single-view results (using the average of the plotted values as a choice criterion), as well as the MvSR results. In order to demonstrate that the usage of multiple datasets improves the capacity to recover the correct expression, we must show that MvSR performs on average better than the best single-view.

The polynomial (first row) summarizes well the advantages of MvSR compared to the traditional approach. For noiseless data, we see that no single view finds the correct solution. This was to be expected because they individually lack information about the full form. Since they are easy to recover and noiseless, the solutions are always simplified such that it can only describe the specific view it was trained on. However, MvSR finds the full correct form as soon as the tree size allows. In this simple exercise, the addition of noise actually improves results for the single-views. Indeed, the additional complexity tends to favor larger overfitted expressions, which by chance will perform better on the other views. MvSR appears resilient to noise and consistently outputs the correct solution. In the partial polynomial exercise (second row), single-views are enough to recover the full expression from noiseless data, despite the reduced domain range of study. However, even a small addition of noise results in the degradation of the results. As the single-view has access to only one part of the polynomial curve, it can often fit the training data with a lower degree polynomial. This does not happen with the MvSR as it recovers a correct solution in almost every combination of noise and maximum size. The Friedman equations (third and fourth rows) are the most difficult to recover. No methods reliably recovers the exact solution. It could be explained by a too small evolution time, or by the very limited number of data points, considering the dimensionality of the problem. Nevertheless, in every combination MvSR returns a near-optimal model, while even the best single-views struggle with the lack of information and higher noise levels. We observe that MvSR is resilient against noise and successfully prevents overfitting even when the maximum size is larger than the original expression.

An analysis of the number of free parameters used in the diverse configurations reveals that single views and MvSR tend to over-parametrize the solutions. On average, over all the scenario studied, MvSR and singles views respectively produce models with $\sim 2.5$ and $\sim 3$ extra parameters compared to the original generative function. This issue comes from the current implementation, lacking a penalty term on the number of parameters. The full implementation of MvSR proposed in Section 6.1.2 should improve this point. Nevertheless, this results shows that the improved performances of MvSR fits cannot be explained by an over-parametrization compared to the single view approach.

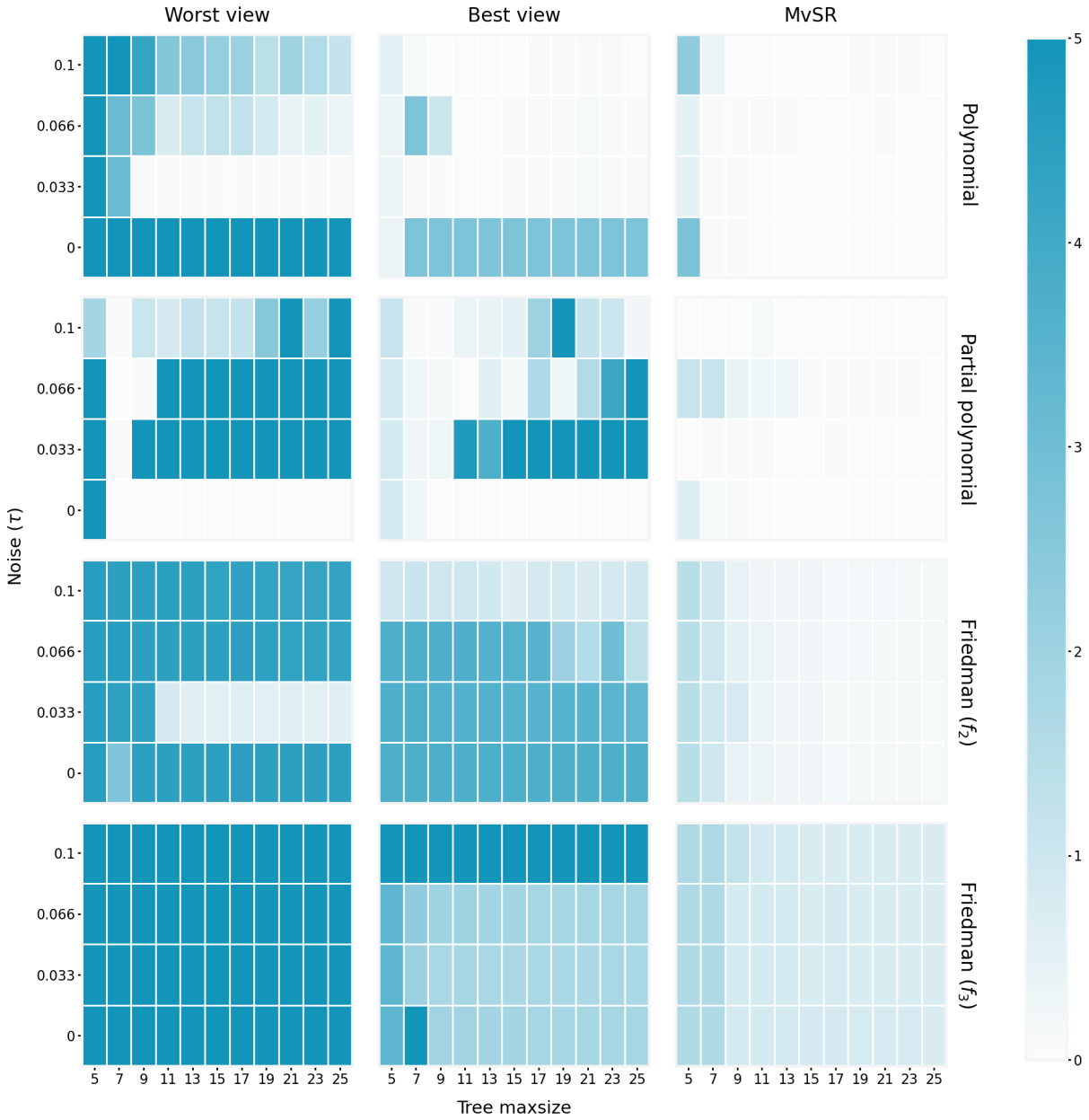



FIGURE 6.2 – Heatmap of the median MSE of the tested combinations of functions, noises, and maximum expression sizes. Row show results for the $f_1$, $f_1$ partial domains, $f_2$ and $f_3$ benchmarks, respectively. Columns represent the worst single-view (left), best single-view (center), and MvSR results (right). The color bar represents the median MSE over 100 iterations for that configuration, ranging from 0 (white) to a clipped value of 5 blue. The clipping improves the comparison of small values.

Overall, we see that MvSR consistently performs better than the best single-view SR. It demonstrates how the addition of multiple example helps to constrain the search towards an optimal solution. In addition, the worst view column illustrates how an unlucky choice of single view can result in extremely poor generalization. Therefore, although in some cases a single example can be enough to find the correct answer, if available the usage of multiple datasets should always be preferred. From these conclusions, we attempt to apply MvSR on various scientific datasets. It constitutes an additional test for the method that will be confronted to real measurements, intrinsically more complex than artificially generated data. It is also more

difficult to evaluate the results, since the true generative functions are unknown. Therefore, this exploration also represents an opportunity to discover interesting functional forms for various phenomenons.

## 6.4 Scientific application

In this section we apply MvSR to real experimental datasets coming from 3 scientific fields, namely chemistry, finance and astrophysics. Such problems represent a significantly harder challenge for SR algorithms as the data generally have no absolute "correct" generative model, are irregularly sampled and display non-Gaussian noise. However, it constitutes perfect testing grounds for MvSR, since one cannot know *a priori* if the chosen examples contain enough information to generalize to all other examples. Therefore, in such cases, the most conservative approach is to always use MvSR in order to build general laws. The code used to produce all the results presented in this work is publicly available [7].

### 6.4.1 Chemistry dataset

Beer's law is an empirical law widely used in chemistry which relates the attenuation of light to the properties of the material through which the light is travelling (Swinehart, 1962). The attenuation of a beam of light going through a liquid is presumed to be only due to absorption, as solutions do not scatter light of wavelengths frequently used in analytical spectroscopy. In UV-visible spectroscopy, we characterize a solution by its transmittance, T. This corresponds to the ratio of light intensity before and after passing through the sample. We express the absorption of a solution as $A = -\log(T)$. We experimentally observe that all molecules display a common pattern: for $A \leq 1$, the absorption rises linearly with the concentration, while for $A > 1$ the linearity breaks and the absorption increases more and more slowly until it reaches a plateau.

The Beer's law is used to describe the properties of chemical species when $A \leq 1$. It states that the absorptive capacity of a dissolved substance is directly proportional to its concentration in the solution, and is expressed as $A = \epsilon l c$ where A is the absorbance, $\epsilon$ is the molar extinction coefficient, $l$ is the optical path length and $c$ the concentration. A handful of alternative models have been proposed. They attempt to build a general Beer's law capable of fitting a larger range of absorption. For example, Bozdoğan (2022) suggested using a quadratic polynomial equation to compensate the positive or negative deviations from the linearity. Yeh et al. (2023) proposed to extend the law by adding two exponent parameters on $l$ and $c$, thus adding flexibility to describe the deviations. Both approaches extend the range of validity of the law but do not provide a solution general enough to properly characterize the absorption at any $A$ value.

We propose to use MvSR on a set of measurements to find such a general parametric solution. In order to proceed, several wavelength scans were carried out using the Hitachi double-beam spectrophotometer UH3500 from 800 nm to 200 nm. Four molecules were analyzed at various concentrations in dichloromethane: a commercial coumarin, two bodipy (Song et al., 2012, Tran et al., 2023) and a porphyrin (Alan et al., 1967) which was previously synthesized in order to measure their absorbance $A$ as a function of the solution's concentration. Ideally, the parametric

7. https://github.com/erusseil/MvSR-analysis

model should recover their extinction coefficient $\epsilon$.

After a small hyperparameter exploration we produce a simple accurate parametric function using MvSR. It was produced with a max tree length of size 15, with *add*, *sub*, *mul*, *div*, *exp* and *log* operators allowed, resulting in:

$$f(x;\mu,\epsilon) = \log\left(\frac{1}{\mu+\mathrm{e}^{-\epsilon x}}\right). \tag{6.4}$$

Figure 6.3 presents the best fits of Equation 6.4 on the datasets. The latter display high non-linearity behaviors, in particular the absorption tend to reach a plateau around $A = 3$, nevertheless, the extended Beer's law proposed by MvSR carries ideal properties to fit the observations. Indeed, in the case where the linearity is respected, the $\mu$ parameter can be set to 0 and the equation simplifies into the original Beer's law. In such case, $\epsilon$ carries exactly the same information as the standard method. In the general case, $1/\mu$ characterizes the plateau that the absorption will reach at high concentrations. MvSR shows that an exponential transition between the linear evolution and the plateau provides an accurate fit to the data. Additionally, Figure 6.3 displays the classical Beer's law fitted to the data points for which $A \leq 1$. We also observe that the lower the extinction coefficient is, the further apart the two models are at low absorption. This suggests a deviation from Beer-Lambert's law at low $\epsilon$, and would require further investigation.

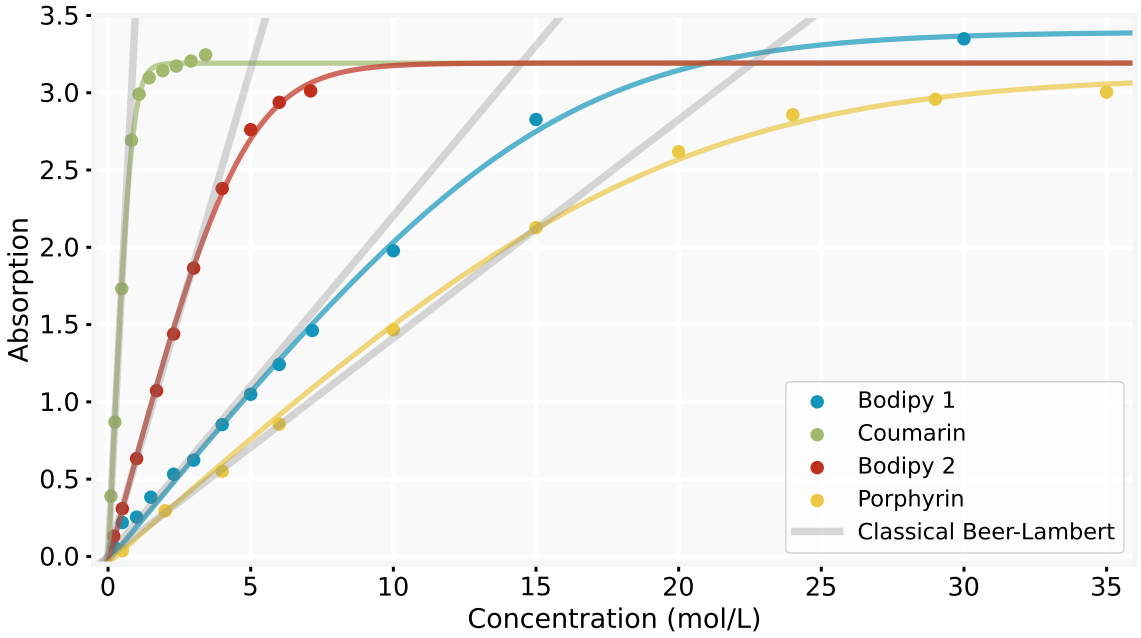



FIGURE 6.3 – Best MvSR fit (Equation 6.4) of the absorption as a function of the molar concentration for 4 different molecules. Gray lines correspond to the Beer's law fitted to the data points for which $A \leq 1$.

In summary, Equation 6.4 offers the possibility of computing molar extinction coefficients without being strictly restricted to the linear regime. The functional form proposed by MvSR contains two parameters and thus requires only a few data points to be constrained. The effect of the variation of the parameters is shown in Figure 6.4. It could be used as an easier and more general alternative for the determination of intrinsic properties of chemical species.

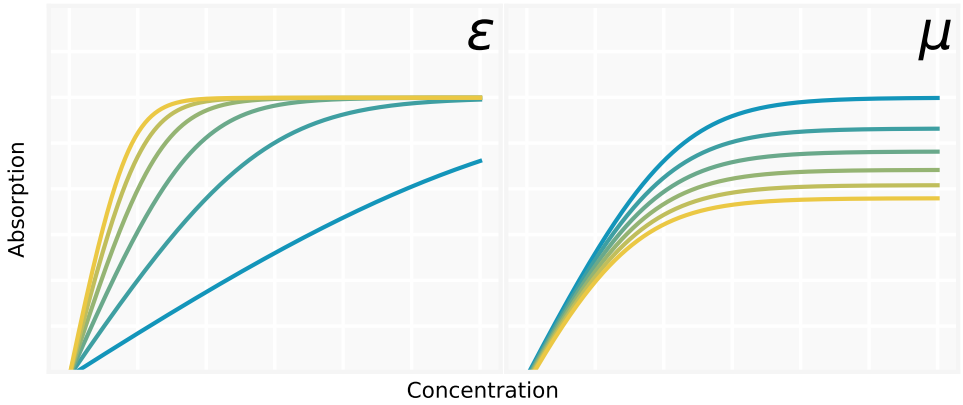



FIGURE 6.4 – Effects of the variation of parameters for the Equation 6.4, the extended Beer's law. The parameter values increase from blue to yellow. The model's parameters $\{\epsilon, \mu\}$ are respectively equal to $\{0.3, 0.05\}$, and panels respectively show their individual variations within bounds $\{[0.1, 0.75], [0.05, 0.15].\}$

### 6.4.2 Finance dataset

In finance, an asset represents a resource owned or controlled by an economic entity. In the context of MvSR different assets can be used as different views to study the complex emergent behaviors of financial markets. We define the return, $r$, as the difference between the price, $p$, of an asset at time $t$ and time $t+1$ such that $r_t = p_{t+1} - p_t$. At first approximation, we can describe the distribution of returns by a Brownian motion (Bachelier, 1900). The field of econophysics, which applies tools from statistical physics to study these stochastic processes, seeks to identify probability distribution functions that go beyond such Gaussian approximation. Those initial models led to the development of the renowned Black-Scholes equation (Black and Scholes, 1973), today known to neglect the significance of rare and extreme events due to their lack of fat tails. Indeed, the distribution of these specific events can be fitted with a *power-laws* (Mantegna and Stanley, 1995).

More recent models, as presented in Mantegna and Stanley (1999), focus on Lévy processes. These include, for example, the Gaussian distribution as well as the Cauchy distribution, the second being an already improved solution with its fat tails. Such models were proposed after studying the statistical properties that most data exhibit, such as the famous $S\&P500$ [8], and then finding distributions that possess such properties. However, our Multi-view approach enables us to identify a common distribution for all assets by considering each of them individually. In this section, we show that MvSR rediscovers some of the presented distributions and propose new models that better fit the real return distributions.

We analyze time series data consisting of multiple prices generated by various assets. We use a publicly available Kaggle dataset [9] containing data from 491 companies. We use 10 random companies as views for MvSR, while the remaining are only used for testing purposes. Values of assets are taken each day at open market time, over a period of 5 years starting from January $1^{st}$, 2018. For each asset, we study the distribution of its returns with a sampling of 100 bins of equal

8. $S\&P500$ regroups the assets of 500 big companies.
9. https://www.kaggle.com/datasets/iveeaten3223times/massive-yahoo-finance-dataset

width. We normalize the data as described in Section 6.2.1. The data shown in Figure 6.5 gives two examples of asset distribution. This type of data exhibits common statistical properties. A positive mean corresponds to the overall economic growth ; a leptokurtic profile which indicates more extreme events than a normal distribution would produce ; and a negative skewness marks an asymmetric shape that informs rare events are more likely to be crises than economic booms. The first and last properties are only observed for datasets covering long time periods (months or years) such as ours.

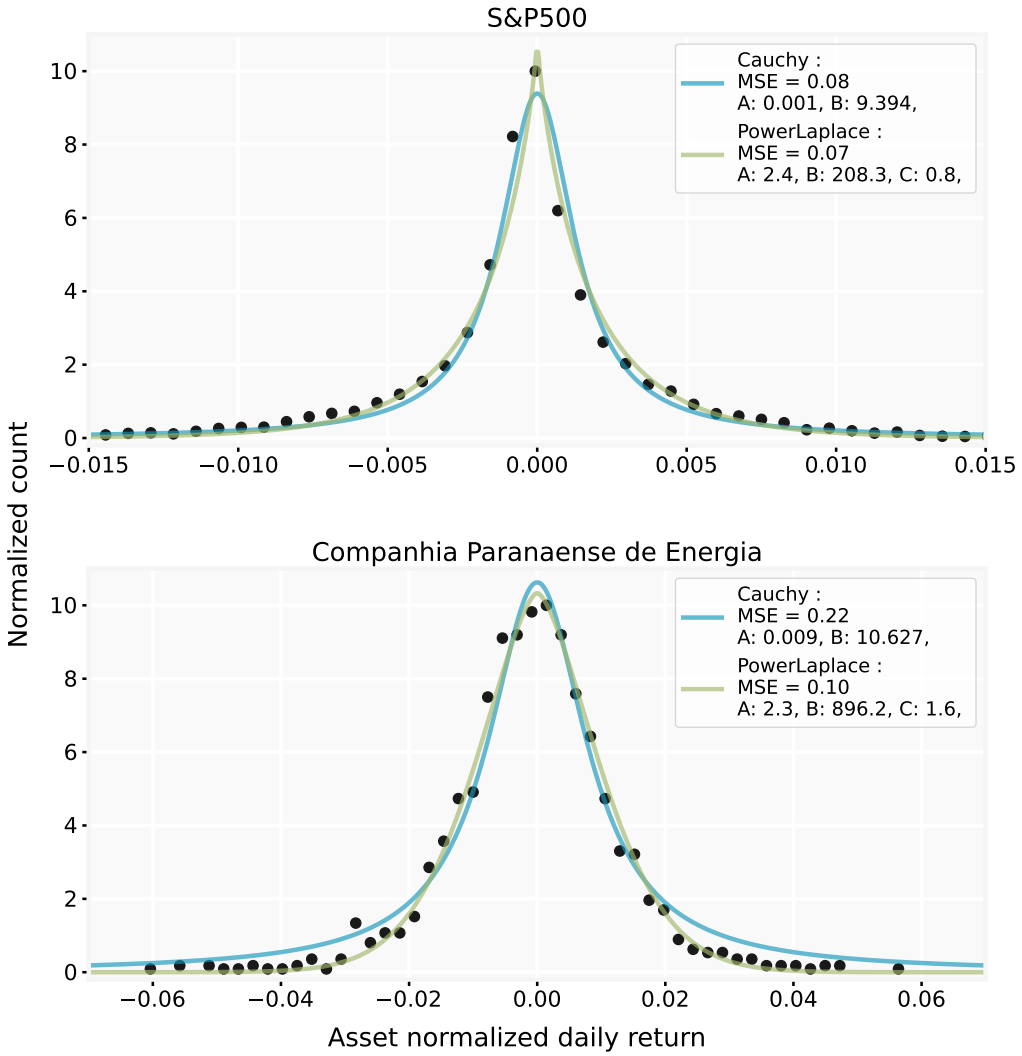



FIGURE 6.5 – Normalized (following the procedure of Section 6.2.1) distribution of returns for 2 assets, fitted by the Cauchy model and the best MvSR solution (Power-Laplace).

Under those conditions, we explore multiple seeds, tree lengths (ranging from size 8 to 20) and operators (using various combinations of *add*, *sub*, *mul*, *div*, *exp*, *abs* and *power*) to obtain multiple possible parametric solutions with MvSR. In Table 6.3 we present 6 parametric models generated by MvSR. We measure the performance of all functions by fitting them to each asset, and display the MSE value as a comparative metric in the last two columns of Table 6.3.

| *Models* | Equation f(x) | $med(MSE)$ | $MSE_{S\&P}$ |
|---|---|---|---|
| Gaussian Bachelier (1900) | $A \cdot \mathrm{e}^{-\frac{x^2}{B}}$ | 0.363 | 0.260 |
| Laplace Kou (2002) | $A \cdot \mathrm{e}^{-B\lvert x\rvert}$ | 0.342 | 0.084 |
| Cauchy Liu et al. (2012) | $A \cdot B^2/(x^2+B^2)$ | 0.305 | 0.079 |
| Linear-Laplace | $(A-Bx) \cdot \mathrm{e}^{-C\lvert x\rvert}$ | 0.327 | 0.065 |
| Exp-Laplace | $A \cdot \mathrm{e}^{Bx-C\lvert x\rvert}$ | 0.328 | ***0.063*** |
| Power-Laplace | $A \cdot \mathrm{e}^{B\lvert x\rvert^C}$ | ***0.246*** | 0.075 |

TABLE 6.3 – Best functions generated by MvSR. The last two columns respectively show the median MSE score of the functions fitted on individual normalized assets and the score when fitted on normalized the $S\&P500$ dataset. Bold numbers correspond to the best score of the column.

In our experiment, MvSR noticeably recovers 3 widely used parametric forms: the Gaussian, Cauchy and *Laplace* distributions. In addition to these results, we obtain solutions similar to Laplace with some variations. These three distributions are presented in Table 6.3. We can see from their scores that those parametric forms present fits of similar quality compared to the solutions already present in the literature. In particular, the median MSE shows a significant improvement when using the *Power-Laplace* function. We name it after it's composition $h(x) = g(f(x))$ with the Laplace distribution $g(x) = e^{-a\times|x|}$ and a power function $f(x) = x^b$.

We illustrate this by showing the distributions of returns of two assets, alongside the fit given by *Power-Laplace* and *Cauchy* (the literature function performing the best on our dataset). The first asset corresponds to the aggregation of $S\&P500$. The second asset is the one for which we observe the biggest improvement in using *Power-Laplace* compared to *Cauchy*. It illustrates the advantages of the new function. It performs well because the power is always a root [10], which accentuates the fat tail effect of the distribution, allowing a better fit in the tails while maintaining a good peaked fit in the center. The effect of the variation of the parameters is shown in Figure 6.6. The last column of Table 6.3 shows the scores of all the functions fitted on the $S\&P500$ index. This index is an aggregated measure of the value of the assets of 500 large companies. As such, the distribution of its returns should behave like an average distribution of these assets. For these data, the best distribution is the one in the second to last line of Table 6.3. Like for the previous application on multiple assets, this shows our new distributions are also better fits than those present in the literature.

We see that distributions from the literature are particularly relevant for the $S\&P500$, as they all perform well (except the basic Gaussian). However, their scores do not differ much. This shows that they are almost all equivalent fits for this aggregated index. On the other hand, their median scores for a variety of assets are much more disparate. In this second case, our algorithm produced much more efficient distributions that outperformed all the others. This indicates that our multi-view approach is an efficient strategy for finding a unique distribution that gives good fits for as many assets as possible.

10. We always find $0 < b < 1$.

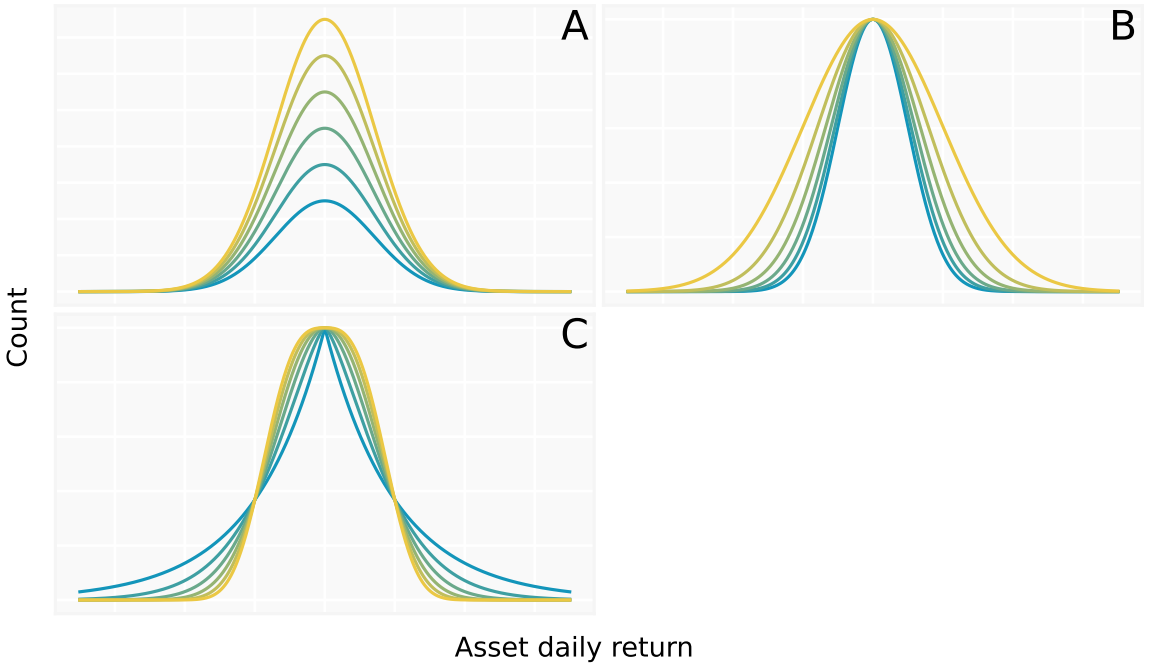



FIGURE 6.6 – Effects of the variation of parameters for the *Power-Laplace* equation (Table 6.3). The parameter values increase from blue to yellow. The model's parameters {A, B, C} are respectively equal to {1, -1, 2}, and panels respectively show their individual variations within bounds {[0.5, 1.5], [-2, -0.5], [1, 3]}.

We showed that applying MvSR to a set of assets all at once, rather than looking at them separately or using an aggregated index, is an efficient approach to gaining new insights into markets behavior and improving characterization of the stochastic process that governs them. Results presented here could be fine-tuned by adding more functions to the MvSR. For example, the *Gamma function* could allow us to generate more distributions with similar properties to those found in the literature, such as the Gamma variance process (Madan et al., 1998) and the Student's t-distribution (Bouchaud and Potters, 2000), which are more general forms of the distributions we already found using MvSR. Finally, further investigation of the distributions found here may prove insightful for future modelling in the field.

### 6.4.3 Astrophysics dataset

Astrophysics constitutes an ideal testing environment for MvSR as it encompasses a diverse array of non-linear equations aimed at modeling observations of large-scale phenomenon within the Universe. Classical SR has been widely used to characterize various relationships (see Section 4.4.3 for more details) such as the Hertzsprung–Russell diagram, the plane of elliptical galaxies (Graham et al., 2013) or photometric redshifts (Krone-Martins et al., 2014). After this project was completed, we also became aware of Tenachi et al. (2023), who used a similar multi-dataset approach when dealing with the treatment of stellar streams.

We will articulate our experiment around supernovae light curve modelization. Given the high intrinsic variability within these events, MvSR appears as a good solution to generate data-driven versatile models. Many functional forms have already been proposed (Section 3.4) for this task. It is important to understand to which extent MvSR is able to re-build the known

solutions. In addition, the potential discovery of new functions better describing the light curves can result in more efficient further analysis.

We choose three SNIa (Section 3.1.2) with a good time sampling from ZTF Data Release 17: SN2019fck, SN2018aye and SN2021mwb (Bellm et al., 2019c, Malanchev et al., 2023). We use observations in $g$ and $r$ photometric filters independently, resulting in a dataset of six examples for MvSR. The light curve shape of SNIa in the $g$ filter could be roughly described as an exponentially rising brightness followed by an exponential fading. Such behavior is typically well modeled by parametric equations found in the literature (Bazin et al. 2009, Villar et al. 2019, Sanchez-Saez et al. 2021). However, being observed in a redder $r$ filter, SNIa light curves display a secondary bump. This particular behavior is much harder to describe with a simple parametric equation. We choose to use both $g$ and $r$ bands to provide a wide variety of examples to MvSR. We apply data quality cuts and select only data points with a signal-to-noise ratio greater than 20. The data was shifted so that the observed peak time is at $t = 0$ and the entire light curve was normalized by its maximum flux. Finally, we consider only data points with a time ranging from -50 to 150 days from the observed peak.

We use a maximum tree length of 12, include the *add*, *sub*, *mul*, *div*, *exp*, *square* and *power* operators and explore multiple seeds to obtain a panel of possible parametric solutions. Table 6.4 displays a selection of the 3 best parametric solutions found by MvSR along with the mean $R^2$ score of the best fit on the SNe.

| Equation f(t) | $< R^2 >$ | No. parameters |
|---|---|---|
| $e^{-At\cdot(B-e^{-Ct})}$ | 0.990 | 3 |
| $\frac{A}{(B\cdot e^{Ct}+e^{-Dt})}$ | 0.987 | 4 |
| $\frac{A^{Bt}}{Ct+(-Dt+e^{Et})^2}$ | 0.992 | 5 |

TABLE 6.4 – Summary of the best parametric functions generated using MvSR on SNIa light curves. The second column corresponds to the mean $R^2$ score over the 6 examples provided.

Noticeably, we recognize the second equation as the Bazin function, a model widely used in the literature (Section 3.4.1). However, the standard form includes a $t_0$ parameter appearing twice which is added to $t$, effectively encoding for a time shift. The form generated by MvSR is mathematically equivalent but uses parameters appearing only once. The top panel of Figure 6.7 shows a Bazin fit of one of the three SNe (in $g$ and $r$). The last equation in Table 6.4 requires 5 parameters and provides the most accurate description of the SNe. MvSR can produce solutions of arbitrary size with increasingly good fit, however we choose to limit it to 5 parameters maximum in order to prevent overfitting. This choice is coherent with the functions used in the literature (Section 3.4).

Finally, we highlight the first solution presented in Table 6.4. It is characterized by an unusual intricate exponential form. Despite lacking similar counterparts in the literature, it provides excellent fits and even outperforms the standard Bazin function. The bottom panel of Figure 6.7) displays its best fit on of one of the three SNe. Although it is not describing the

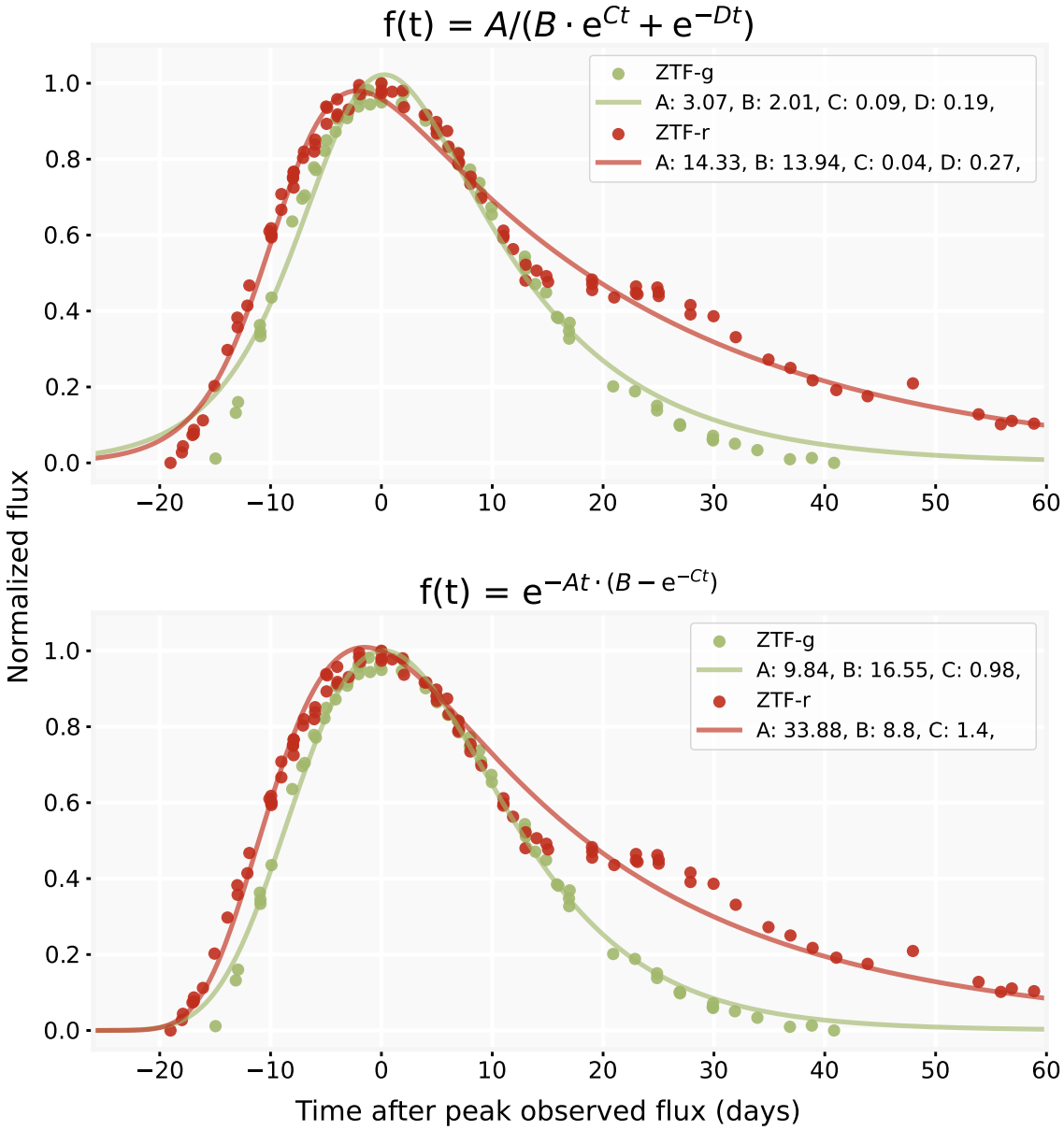



FIGURE 6.7 – Best fit of two parametric functions found by MVSR on SN2021mwb in the $g$ and $r$ filters. The top panel corresponds to the Bazin function commonly used in the literature.

second bump, it shows a clear improvement in the description of the rising and decaying tails of the event, as well as in the peak phase. Such improvement is particularly impressive considering that one less parameter was required to describe the transient. However, in order to be used in non-normalized contexts, the function should be expended by the addition of two standard parameters: a time shift and an amplitude. Hence, after adding the parameters and rearranging its terms, we propose the *Doublexp* function [11], defined as,

$$f(t; A, t_0, \tau_1, \tau_2, p) = A \cdot \mathrm{e}^{\frac{t-t_0}{\tau_1} \cdot (p - \mathrm{e}^{-\frac{t-t_0}{\tau_2}})}, \tag{6.5}$$

11. Named after its intricate exponential form

where $A$ is the amplitude, $t_0$ a reference time, $\tau_1$ and $\tau_2$ timescales parameters, and $p$ a unitless adjustment parameter. Unlike any previously handcrafted transient models (Section 3.4), the parameters were not built to encode intuitive behaviors, and thus are much less trivial to interpret. Figure 6.8 shows the effect of the variation of each parameter. From this, we can interpret $\tau_1$ and $\tau_2$ as being respectively related to a characteristic decay and rise time, while $p$ is linked to a change in the decay rate during the falling phase. We also notice that they are partly degenerated with the amplitude and time shift parameters [12].

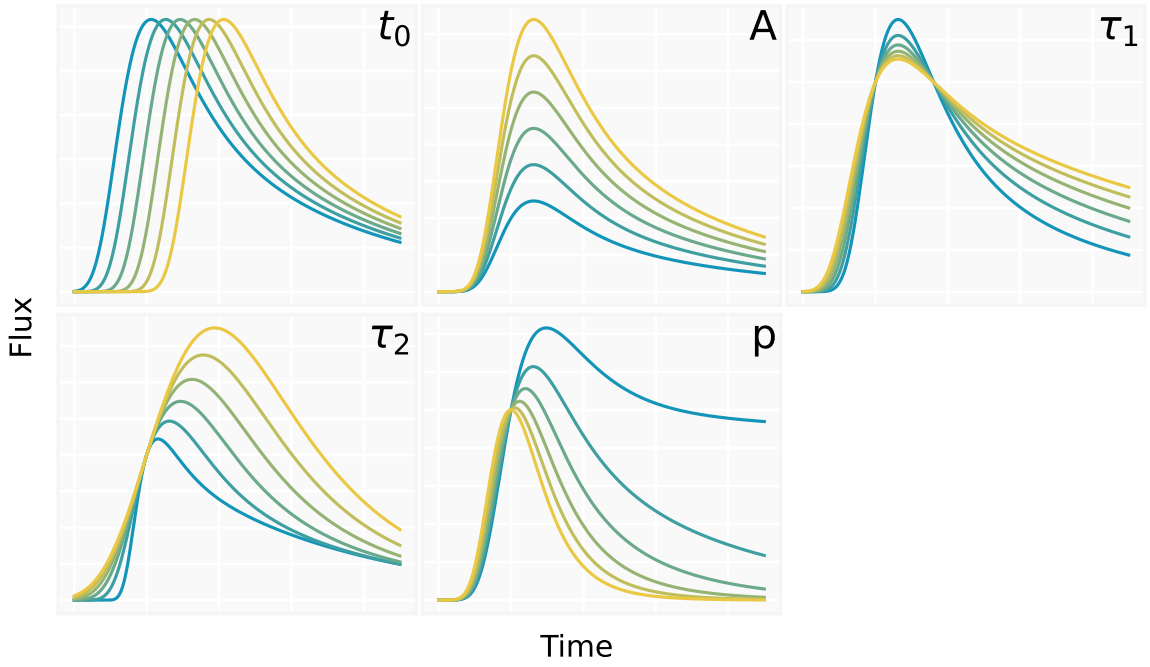



FIGURE 6.8 – Effects of the variation of parameters for the Equation 6.5, the *Doublexp* function. The parameter values increase from blue to yellow. The model's parameters $\{t_0$, A, $\tau_1$, $\tau_2$, p$\}$ are respectively equal to {0, 1, 10, 10, 0.2}, and panels respectively show their individual variations within bounds {[-5, 15], [0.5, 1.5], [8, 10], [5, 30], [0.01, 1]}.

Overall, MvSR was able to generate multiple good models to describe the behavior of SN Ia. It recovered a solution from the literature and even proposed improved solutions in terms of goodness of fit and/or number of parameters used. However, the generated models struggle with the same problem as the equations from the literature: they don't provide a simple description of the second bump of the SNIa in the $r$ band. It may also highlight a limitation of the current MvSR implementation, which doesn't include parameter repetition, the second point of the complete implementation stated in Section 6.1. Given that a repeated time shift parameter is standard in models used in the literature, this result highlights the importance of a more complete MvSR implementation for practical scientific applications.

### 6.4.4 Astrophysical dataset (early development)

Prior to the efficient MvSR (presented in Section 6.1.2), a rough implementation built from `gplearn` [13] (Section 4.4.2) was developed. Due to the low computational speed and accuracy

12. The amplitude and time shift parameters remain necessary nonetheless for non-normalized light curves.
13. https://gplearn.readthedocs.io/en/stable/

of *gplearn* (La Cava et al., 2021) this code was rapidly obsolete and replaced by the pyOperon based version. However, it had the advantage of fulfilling all the 6 core points of a full MvSR implementation (Section 6.1). Among the experimentation done with this first version, we attempted to find a good functional representation of the light curve of a SLSN (Section 3.1.3). SNAD160 [14] (ZTF18aautop) behavior has been of particular interest because of its extremely slow evolution of ∼2 years, evoking it as a PISN candidate (see Pruzhinskaya et al. (2022) and Appendix B). Using the $g$ and $r$ passbands as two separate views, MvSR generated the following model:

$$f(t; A, t_0, \tau) = A \cdot (t - t_0) \cdot \mathrm{e}^{\frac{t-t_0}{\tau}}. \tag{6.6}$$

It contains 3 parameters: the amplitude $A$, $t_0$ a time shift and $\tau$ the timescale for the decaying phase. The full implementation of MvSR enabled the repeated usage of $t_0$ in the equation. The model is interesting as it allows modeling a transient from the first rising point until complete exponential extinction with very few parameters. This is due to a constraint on the rising part, which is roughly approximated as linear with a slope depending on $\tau$, and for which no previous baseline is modeled. We name this function *Linexp* after its combination of linear rise and exponential decay. Figure 6.9 shows the effect of the variation of each parameter.

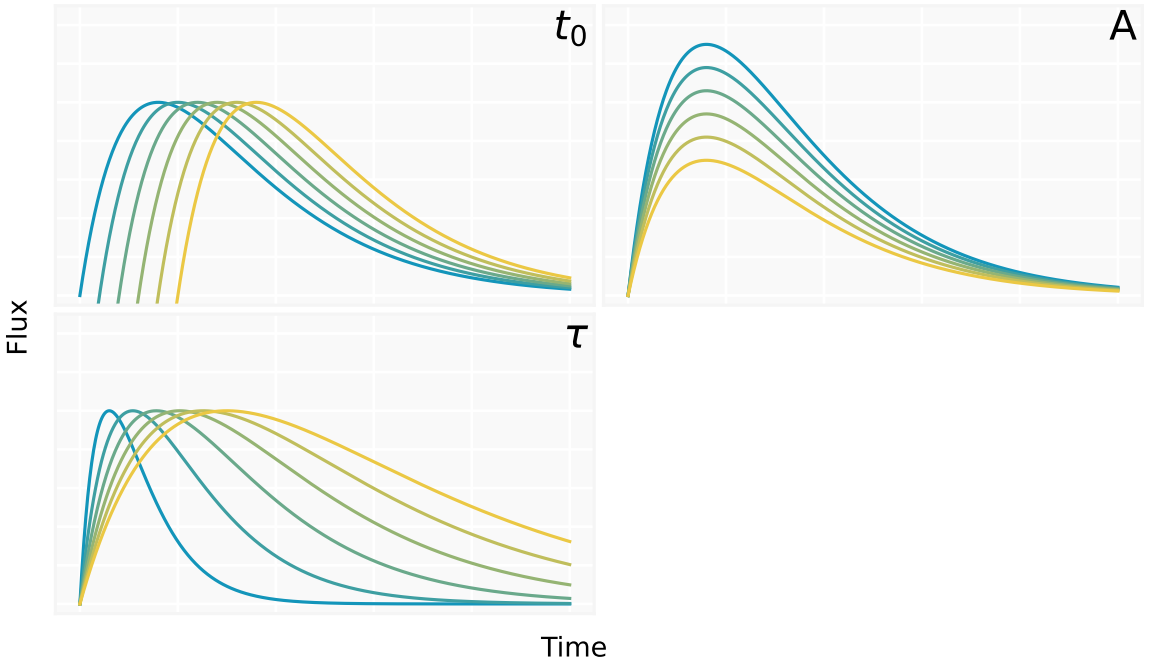



FIGURE 6.9 – Effects of the variation of parameters for the Equation 6.6, the *Linexp* function. The parameter values increase from blue to yellow. The model's parameters $\{t_0, A, \tau\}$ are respectively equal to {0, -1, 8}, and panels respectively show their individual variations within bounds {[0, 10], [-1.3, -0.7], [3, 15]}.

It illustrates how the model irrevocably links the rise and fall timescales. This property enables the first order modelization of SLSN. In particular, the quality of the fit carries information on whether this property is or not relevant for a given transient. Appendix C explains how it has been used for the feature extraction and classification of SLSN light curves in the ELAsTiCC challenge (Fraga et al., 2024).

14. https://ztf.snad.space/dr17/view/821207100004043

## 6.5 Conclusions

Symbolic Regression (SR) has proven to be extremely efficient in searching for mathematical expressions that describe the relationship between a set of explanatory and response variables. In its traditional form, it translates the behavior of one such dataset into an analytical function which can be used for further analysis. Nevertheless, in a realistic scenario, the researcher is frequently faced with multiple outcomes from the same experiment. These may correspond to different experimental setups, initial conditions or domain coverage, but are all generated by the same underlying mechanism, one which the researcher aims to describe.

In this work, we proposed Multi-View Symbolic Regression (MvSR), a framework that exploits this scenario by extending the scope of traditional SR, allowing the user to provide multiple examples as input. The algorithm searches for the best parametric function which simultaneously describes all the input data provided. This is achieved by fitting each input dataset independently with the same regression model and aggregating their individual fitness into a single one. For this purpose, we included an aggregation fitness function in Operon that supports different aggregators such as mean, maximum, median, or harmonic mean.

We have tested this approach with four different challenging artificial benchmarks with added noise and composed of extreme situations where some of the parameters are set to 0 or the domain coverage is limited. When compared to the traditional approach, we report that MvSR is capable of correctly retrieving the original expression in most configurations with a higher accuracy than SR, even in a presence of a strong noise. Additionally, we stress-tested our method by using real-world experiments from three different areas: chemistry, finance and astrophysics. MvSR was not only capable of recovering well known models from the literature, but it also found new alternatives that are promising in these fields.

Our results showcase the potential enclosed in applying MvSR in real scientific scenarios. These could be made even better with further functionalities, like enabling a maximum number of parameters to be used in the model, either as a hard constraint or through a penalization term, and allowing the same parameter to appear more than once in the final expression. Such additional features would not only result in more flexible functions, with smaller number of parameters, but it would also allow the researcher to indirectly tailor the final result, thus increasing the chances of a parametric form which can inspire interpretability. In this context, a full implementation of MvSR would have much broader applications than the ones described here. We intend to further explore its ramifications in dedicated studies focused on the particularities of each science case. Nevertheless, in order to fully exploit its potential and popularize its use within a broad range of scientific areas, a user-friendly implementation of all the above-mentioned functionalities is paramount. Our collaborators from computer science are engaged in this exercise, and we expect to have a full implementation available to the community soon.

What is most relevant for the broad science problem being tackled in this thesis is that the application of MvSR on supernovae light curves has yielded excellent results. Not only it was able to recover the Bazin function, it has also generated two interesting parametric forms: *doublexp* (Equation 6.5) and *linexp* (Equation 6.6). They both offer efficient light curve description with respect to the number of free parameter they require. As such, they both constitute

excellent candidates to be used for the bolometric description of a Rainbow fit analysis (Chapter 5). However, they clearly do not serve the same purpose. While *linexp* is useful to describe transients with a poorly constrained rising light curve, *doublexp* is suited to model the complete behavior of well sampled ones. This specialization of models due to specific characteristics of the data is unavoidable. It does not only apply to the functions generated by MvSR, but also to those already proposed in the literature (Section 3.4). Each of them have their own requirements regarding the number of observations, are specialized in early or late light curves, and their corresponding results achieve different levels of precision. None of them are intrinsically better, they are simply suited for different data contexts.

Therefore, in the context of big photometric surveys, rather than settling on a single function that will suboptimally fit certain light curves, we propose to make the best use of the library of models at our disposal. It can be done individually for each light curve, by choosing the most adapted model, considering its state of evolution. This procedure will enable the optimal parametric description of transients in the dataset, while ensuring the earliest fit possible by including simple models. These qualities are crucial for large time domain surveys, like LSST, which require early and reliable classification to enable spectroscopic follow-up. Therefore, the next chapter presents our proposition of such an adaptive fitting, using the models generated by MvSR within the Rainbow framework. We showcase the performance of the method on the exercise of transient classification from LSST-like simulated data.

# 7

# Adaptive transient classifier

Results obtained using the Rainbow framework (Chapter 5) showcase the importance of a multi-passband approach in characterizing transient light curves. Beyond enabling a physically motivated description, it also reduces the number of observations required to perform a fit by taking into account the correlation between filters. This is crucial in the context of multi-band surveys like LSST (Section 2.3.2). Given its 6 filters, and the complex decisions behind its final observation strategy [1], it is safe to assume that only a small fraction of light curves will have enough points to mathematically model filters independently (Section 3.4). In this context, Rainbow proved to be successful in satisfactorily describing light curves while demanding a significantly lower number of observations.

Nevertheless, one issue that is yet to be addressed is the relevance of using one single functional form to describe the bolometric behavior of all objects in a diverse dataset. Although we may reasonably fit complete light curves (i.e. fully displaying the rise, peak and decay parts) with a single versatile function, the description of partial ones can certainly be optimized using other alternatives. For example, even for a well sampled light curve, there is little sense in describing only its rising phase with a complex model capable of encoding post-maximum stages like second bump or plateau behaviors. Moreover, if we subsequently use the minimized parameter values as features to train a machine learning model (e.g. Ishida et al., 2019, Karpenka et al., 2012, Sanchez-Saez et al., 2021, as well as Section 5.3.3), the absence of constraints from the data will lead to a non-informative coverage of the parameter space.

This issue may be negligible when analyzing complete light curves in well curated catalogs, however, it becomes of primary importance when dealing with real time alert streams generated by time domain surveys such as ZTF and LSST (Section 2.3.3). In this context, each alert carries a light curve composed of the last detection in addition to a finite photometric observation history of the transient [2]. Thus, the same astrophysical source may generate many sequential alerts, each one containing one additional observed point. Using a complex model to fit an early transient will not only result in overfitting, but also in a not-so-early characterization. This may cause significant impact in our ability to quickly identify fast transients for subsequent follow-up.

Ideally, we would like to construct a flexible parameter space suited for training traditional machine learning models, and capable of matching the information content available in each light curve. In this chapter, we propose such an adaptive feature extraction pipeline. Taking advantage of the flexibility allowed by Rainbow, we made available three options for depicting the temperature evolution, as well as 6 different possible functional forms for describing bolometric light curves. Among the latter, 2 were proposed by Multi-view Symbolic Regression (MvSR, Chapter 6). Given this initial set-up, for each light curve, the algorithm goes through a series of decisions in order to choose the most appropriate combination of temperature and bolometric light curve descriptions. Thus, each light curve is modelled with just enough complexity to allow meaningful feature extraction. Once the entire dataset is processed, the best-fit parameter values are aggregated into a single matrix, which is subsequently used to train and test a Random Forest classifier (Section 4.2.2). The complete pipeline is evaluated for the case of extragalactic transient alerts classification. The exact methodology is explained in Section 7.1 and Section 7.2

1. https://www.lsst.org/content/charge-survey-cadence-optimization-committee-scoc
2. The light curve within each alert package is limited to 30 days in ZTF. This threshold is expected to be of 1 year for LSST.

details the dataset used to evaluate the method. Section 7.3 presents the classification results and Section 7.4 shows our conclusions.

## 7.1 Adaptive Feature Extraction

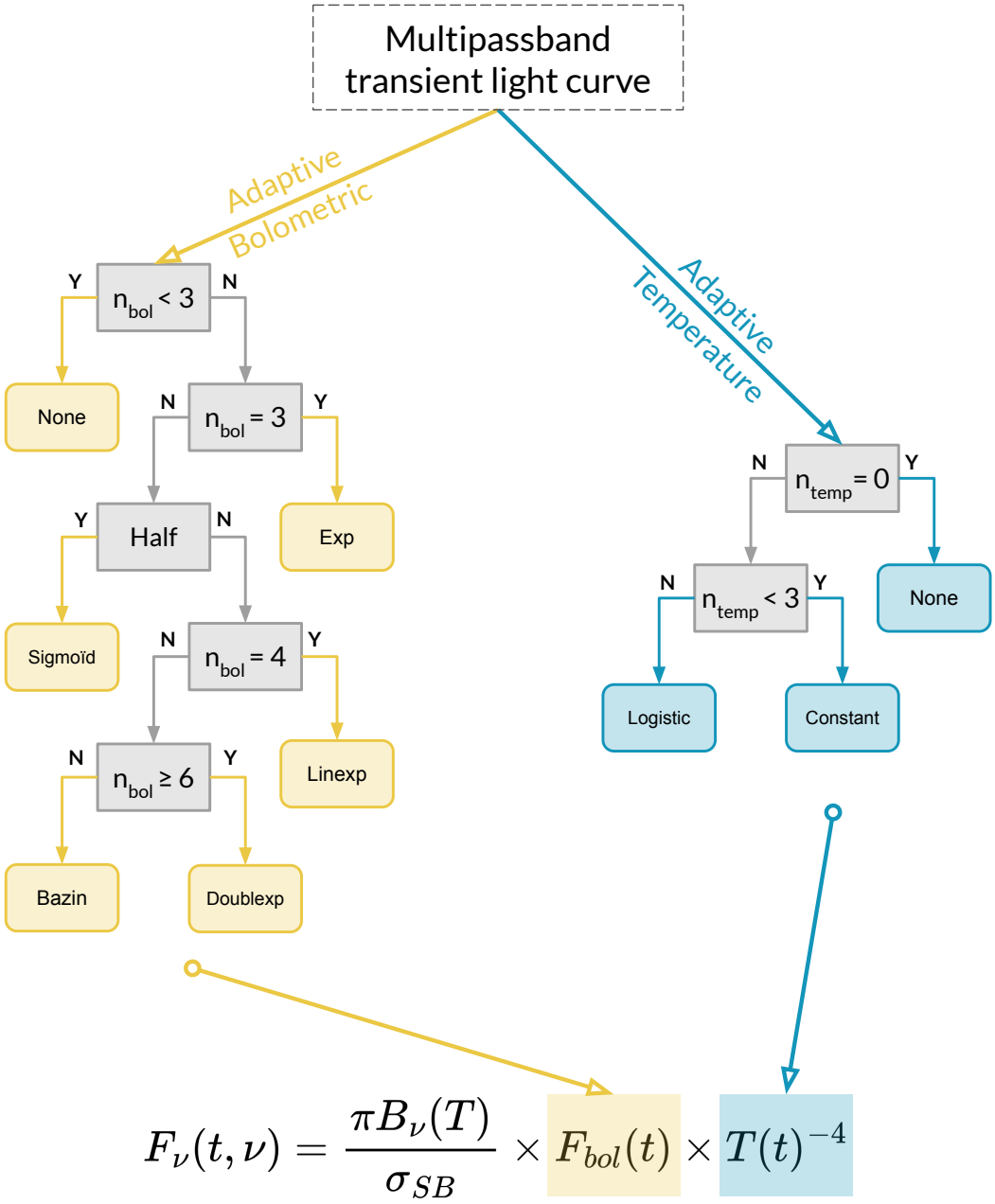



FIGURE 7.1 – Summarizing chart of the adaptive feature extraction pipeline. The bolometric model is chosen based on the number of observations in most sampled passband, $n_{bolometric}$, and the state of evolution of the light curve, *half* (Section7.1.1). The temperature model is chosen based on the number of observations in the second most sampled passband, $n_{temperature}$ (Section7.1.2). Both models are combined into a RAINBOW framework (Equation 5.4), which is used to fit the light curve and extract the best-fit parameters.

The methodology proposed in this section is adaptive, in the sense that the model chosen to fit the data depends on the data itself. This choice is unambiguously determined based on the number of detections and the state of evolution of the transient. In the context of RAINBOW, two adaptive components are available: the bolometric flux (Section 7.1.1) and the temperature (Section 7.1.2) models. Gathering knowledge from the literature, MVSR generated solutions, as well as basic mathematical functions, we propose a list of possible models, ensuring that each light curve is described using a suitable mathematical description given the information available. Such a tailored model prevents underfitting, i.e. using an unnecessarily simple model for a complex light curve, and overfitting, i.e. using a complex over-parametrized model to fit a handful of observations. The pipeline is summarized in Figure 7.1. We present below details about the available choices for bolometric flux (Section 7.1.1) and temperature (Section 7.1.2) evolution. Examples of reconstructions for different stages of a light curve are shown in Section 7.1.3 and the construction of the feature matrix is detailed in Section 7.1.4.

### 7.1.1 Bolometric flux

The function describing the bolometric light curve must be chosen according to the number of points available in the most well sampled passband. In the context of RAINBOW, a first approximation of the light curve behavior can be deduced purely based on one passband. If a single band has enough points to satisfactorily constraint the fit, additional observations from different filters do not necessarily result in better constraints. Indeed, two observations using different filters, but acquired at the same time (or at close time interval), would not carry information about the bolometric light curve evolution. Therefore, we define $n_{bolometric}$ as being the number of points in the most well sampled passband. If available, the last non-detection point should be included, since it helps to constrain the rising time. To ensure that the fit is meaningful, a function with $k$ parameters requires that $n_{bolometric} \geq k + 1$.

However, $n_{bolometric}$ alone is not enough to choose an adapted functional form. Even with many detections, a complex model can be unconstrained if the light curve behavior consists of a pure rise or a pure decay. We define *half* transients as those which are purely rising or decaying. Similarly, a *full* transient is characterized by a complete light curve, showing the rising, peak and decaying stages of evolution. Fitting a *half* transient with a function tailored to describe *full* light curves would result in a completely arbitrary description of the missing half, which can, in turn, lead to non-informative features. Therefore, a simple check is performed to evaluate if the transient exhibits a *half* or *full* behavior. Inspired by the work developed in searching for early Tidal Disruption Events (Appendix D), a light curve is considered to be in its rising phase if, for all passbands, the last measured flux value, $f(t_{last})$, is not significantly lower than any of the previous ones, $f(t_i)$, taking into account corresponding uncertainties, $\sigma(t_i)$. Mathematically, this translates into

$$f(t_{last}) + \sigma(t_{last}) > f(t_i) - \sigma(t_i), \qquad \forall t_i < t_{last}. \tag{7.1}$$

Similarly, a light curve is considered to have purely a decaying phase if the first point among all passbands is significantly brighter (within error bars) than all the following observations. This means that

$$f(t_{first}) + \sigma(t_{first}) > f(t_i) - \sigma(t_i), \qquad \forall t_i > t_{first}. \tag{7.2}$$

This simple method ensures, in most cases, a correct evaluation of the state of evolution of the transient. Based on this estimate (*half* or *full*) and the value of $n_{bolometric}$, we propose 6 parametric functions, adapted to describe a diverse set of light curve scenarios (Table 7.1). Note that exponential, sigmoid and *Linexp* are symmetrical functions, thus they can be used to describe both rising or decaying light curves.

| **Function** | **Equation**, $f(t)$ | $\mathbf{n_{bolometric}}$ | **State** |
|---|---|---|---|
| None | — | $< 3$ | — |
| Exponential | $e^{\frac{t-\mathbf{t_0}}{\tau_{\mathbf{half}}}}$ | 3 | Half |
| Sigmoid | $\frac{\mathbf{A}}{1+e^{\frac{t-\mathbf{t_0}}{\tau_{\mathbf{half}}}}}$ | $\geq 4$ | Half |
| *Linexp* | $\mathbf{A} \cdot (t - \mathbf{t_0}) \cdot e^{\frac{A}{\lvert A \rvert} \cdot \frac{t-\mathbf{t_0}}{\tau_{\mathbf{half}}}}$ | 4 | Full |
| *Bazin* | $\frac{\mathbf{A}}{e^{\frac{-(t-\mathbf{t_0})}{\tau_{\mathbf{rise}}}}+e^{\frac{t-\mathbf{t_0}}{\tau_{\mathbf{fall}}}}}$ | 5 | Full |
| *Doublexp* | $\mathbf{A} \cdot e^{\frac{t-\mathbf{t_0}}{\tau_1} \cdot (\mathbf{p}-e^{-\frac{t-\mathbf{t_0}}{\tau_2}})}$ | $\geq 6$ | Full |

TABLE 7.1 – Set of expressions and conditions used to choose a bolometric light curve model. From left to right: the first column shows the nomenclature used in the text. The ones in *italic* correspond to expressions found by MVSR, as described in Chapter 6. The second displays the functional forms explicitly. The third column shows the conditions imposed on $n_{bolometric}$ and the fourth describes the light curve states. *Half* implies that the light curve exhibits only a rising or falling behavior. *Full* indicates a complete light curve. The left branch of Figure 7.1 shows a decision tree version of this table.

If $n_{bolometric} < 3$, no fit is performed, since there is not much information to be extracted. For light curves with $n_{bolometric} = 3$ the information available is minimal, however, a simple exponential model can provide valuable information regarding the rising/decaying rate of the transient. Once additional points are available, and if the light curve is still exhibiting a *half* behavior, the sigmoid fit offers a simple and efficient model to describe the transition between the observed peak and the rising/decaying exponential phase. However, if the light curve displays a *full* behavior, with $n_{bolometric} = 4$, the *Linexp* function (Section 6.4.4) can provide some additional insights. It models an exponential evolution of the rising/decaying part of the light curve, an estimate of the maximum flux and a rough approximation of the transition to the other poorly constrained half. For *full* light curves with $n_{bolometric} = 5$, the data becomes informative enough to allow for a complete description of its behavior. In this case, a Bazin fit (Section 3.4.1) enables a first order exponential modelling of both the rising and falling parts of the light curve. Finally, when $n_{bolometric} \geq 6$, the *Doublexp* function (Section 6.4.3) provides a more precise description, while still using a relatively simple expression with only 5 degrees of freedom.

### 7.1.2 Temperature

The temperature functions must be chosen based on the amount of color information available. In practice, to deduce how the temperature changes with time, given a blackbody hypothesis, one must know how any two passbands evolve relative to each other. Therefore, we define $n_{temperature}$ as being the number of points in the second most well sampled passband. In the context of RAINBOW, it allows us to estimate the temperature by observing its relative evolution with respect to the most well sampled passband. Based on the value of $n_{temperature}$, we propose a series of temperature evolution functions, capable of modelling a versatile range of color behaviors (Table 7.2).

| **Function** | **Equation** $T(t) =$ | $\mathbf{n_{temperature}}$ |
|---|---|---|
| None | — | 0 |
| Constant | $\mathbf{T}$ | 1 or 2 |
| Logistic | $\mathbf{T_{min}} + \frac{\mathbf{T_{max}} - \mathbf{T_{min}}}{1 + e^{\frac{t - \mathbf{t_0}}{\tau_{\mathbf{color}}}}}$ | $\geq 3$ |

TABLE 7.2 – Set of expressions and conditions used to select a functional form for the temperature evolution. From left to right: the first column shows the nomenclature used in the text. The second column displays the explicit functional forms and the third shows the number of points in the second most populated passband guiding the choice of function. Figure 7.1 shows a decision tree version of this table (right branch).

If observations are available in only one passband, no information about the color of the source can be deduced. In such case, a complete RAINBOW fit cannot be performed, and a classical single passband description is used. This scenario is important to be taken into account because a transient with many observations in a single band can exhibit a very identifiable pattern that should not be overlooked. If $n_{temperature} \in \{1, 2\}$, we can evaluate the temperature at a given time. However, the available information is not yet enough to properly model its time evolution. Therefore, for lack of better options, a constant temperature is used as a first approximation. In case $n_{temperature} \geq 3$, a logistic function can be used to estimate the temperature variation of the transient. This simple approximation, proposed in Chapter 5, has shown great results in fitting the cooling phase of many transient events. The hypothesis used in Section 5.1 is maintained, and $t_0$ is used as a common parameter shared with the chosen bolometric function.

### 7.1.3 Examples

The combination of bolometric and temperature functions presented above allows us to model a broad range of light curves. Two of them are illustrated below on concrete SNIb/c examples. The data was taken from the ELAsTiCC dataset using the LSST-griz passbands (Section 7.2). Prior to the fit, light curves are normalized by dividing fluxes by the observed maximum over all bands, and defining it to be at time zero. Since for many types of transients, observations

in LSST-u and LSST-Y are often incoherent with the blackbody assumption, we decided not to use them for the fit. However, they are still used for computing statistical features (see Section 7.1.4). ELAsTiCC proposes data in the form of an alert stream, closely matching the future LSST data format. It implies that each object generates an alert (and therefore a light curve) every time a change in brightness is detected. Therefore, multiple light curves containing incrementally more observations can be associated to a single source.

Figure 7.2 presents the most intuitive scenario. All light curves come from alerts generated by the same object but at different stages of evolution, from early (top panel) to late (bottom panel). The first panel corresponds to a very early stage of the transient. Given the small amount of information, the best description consists of an exponential rise ($n_{bolometric} = 3$). The source has already been observed across three passbands, allowing for a first constant estimate of the temperature ($n_{temperature} = 2$). In the second panel, three observations have been added, enabling more precise description. Since the light curve is still on the rise in all passbands, the bolometric model chosen is a sigmoid ($n_{bolometric} = 4$). This new function enables the description of the peak brightness. In addition, the temperature is now sufficiently constrained to use a logistic temperature evolution. In the third panel, the last alert added an observation significantly dimmer than the previous ones in the LSST-g passband. The transient is therefore not considered rising anymore, and given that $n_{bolometric} = 5$, the model chosen to describe the light curve is the Bazin function. This enables a crude estimation of the peak time (that will later be confirmed to be accurate within a few days) and a description of the beginning of the decaying slope. We can also see that the logistic temperature plays a vital role in describing the cooling process (given the blackbody hypothesis), characterized by the reddening of the source. Finally, the last panel shows the transient at late stages of evolution. Here, enough photometric information is available to allow the *doublexp* function to be used. Overall, it is correctly characterized with proper description of the rise and fall timescales. In particular, we see that the fit is compatible with the relative intensities of the different passbands, indicating that the blackbody assumption is a valid model for this object.

Figure 7.3 presents another scenario. In this case, the first three observations are already dimming. Therefore, the transient is considered to exhibit a $half$ behavior and functions will be chosen accordingly. Given that $n_{bolometric} = 3$, a simple decaying exponential is used as the bolometric model. Moreover, since only one passband is available, no temperature can be approximated for the transient yet. The second panel shows two new observations in different filters, enabling a first constant temperature estimation. In the third panel, a significant amount of data was added such that $n_{bolometric} = 6$. However, given that only half of the transient must be described, a sigmoid function is the most complex model that will be used for the bolometric description. Passbands are also sufficiently sampled to use the logistic temperature evolution model. These two functions are the most complex representation used for $half$ transients. As illustrated by the last panel, they are enough to provide a correct characterization of the decaying part of simple transients.

These examples illustrate how the adaptive fit prevents over fitting by choosing functional forms according to characteristics of the data. This enables earlier fit while still describing in detail later stages of evolution. In addition, it correctly describes light curves with many observations that do not require a complex representation, such as the one presented on the

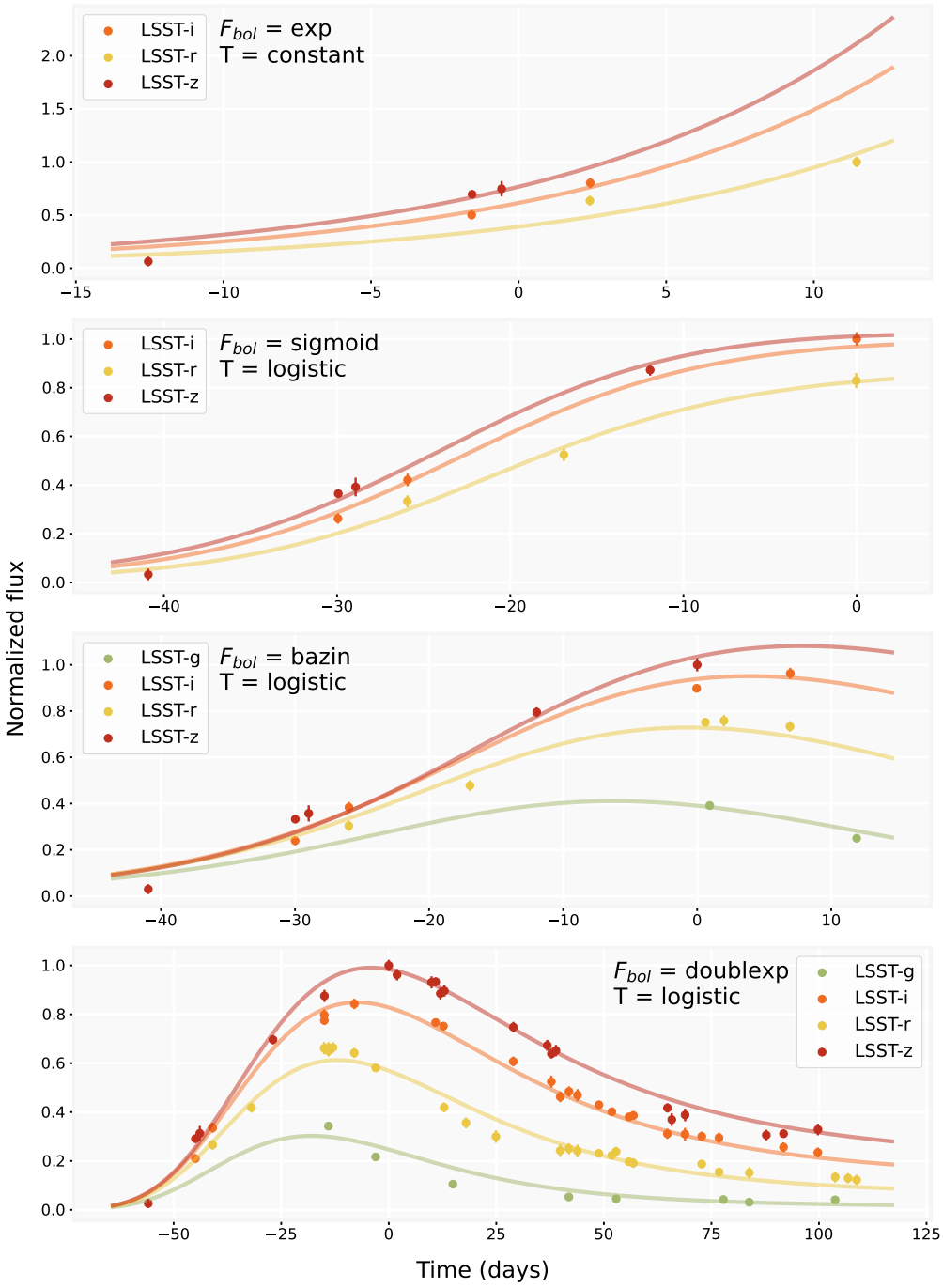



FIGURE 7.2 – Illustration of the adaptive fitting procedure on a *full* SNIb/c object from the ELAsTiCC dataset. Each panel shows the same object in different stages of the light curve, from early (top) to late (bottom). Each color corresponds to a different LSST filter. The parametric models chosen to describe the bolometric flux, $F_{\rm bol}$, and temperature, $T$, are shown within each panel.

third and fourth panels of Figure 7.3. These advantages make the method particularly suited for classification pipelines based on feature extraction.

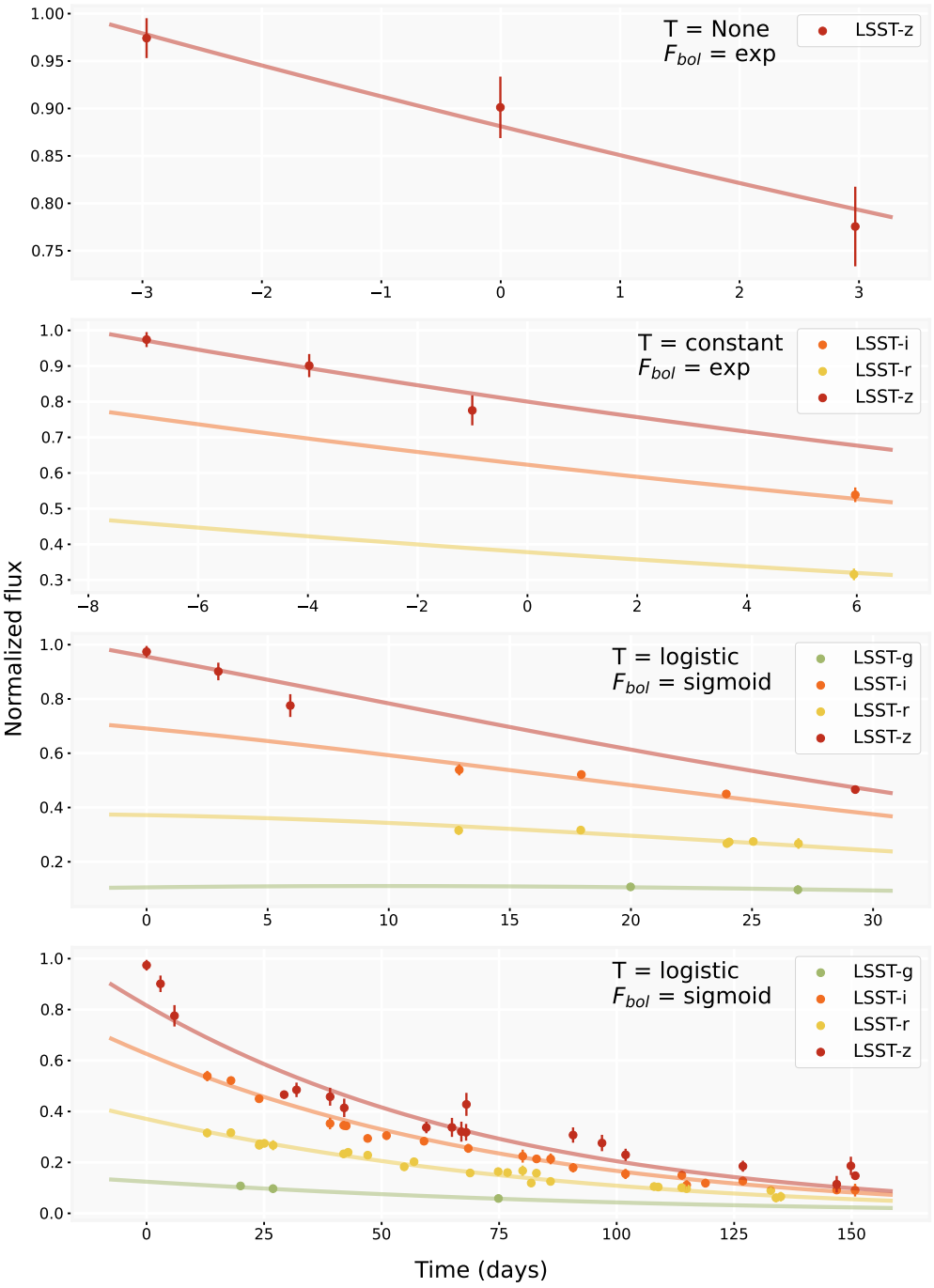



FIGURE 7.3 – Illustration of the adaptive fitting procedure on a *half* SNIb/c object from the ELAsTiCC dataset. Each panel shows the same object with a different number of observations, from low (top panel) to high (bottom panel). Each color corresponds to a different LSST filter. The parametric models chosen for bolometric flux, $F_{\text{bol}}$, and temperature, $T$, are shown within each panel.

### 7.1.4 Feature matrix

The proposed adaptive pipeline enables a tailored and physically motivated parametric description of each light curve, while ensuring an early reasonable fit. The best-fit parameters carry great summarization power, translating the light curve behavior into a low dimensional representation. Hence, similarly to Section 5.3.3, we use them as features to construct a homogeneous

input matrix which will be used to train machine learning classifiers. Each light curve is subjected to the complete feature extraction pipeline (Figure 7.1). Once the combination of bolometric and temperature functions is chosen, the model is optimized using the least square minimization from `iminuit` [3] (Dembinski and et al., 2020). As a result, we store the best-fit parameter values, their associated uncertainties and the final least squared error.

For the construction of the feature matrix, each line corresponds to one light curve and the first 10 columns are reserved for the best-fit parameters associated to the bolometric flux (equivalent to the maximum possible number of parameters and their associated uncertainties). The first column, hereafter *bolometric_0*, is reserved for reference times ($t_0$). The second column, *bolometric_1*, is reserved for the first timescale parameter ($\tau_{\rm half}$ for exponential, sigmoid or *Linexp*, $\tau_{rise}$ for *Bazin*, and $\tau_1$ for *Doublexp*). The third column, *bolometric_2* holds amplitude values ($A$ for sigmoid, *Linexp, Bazin* or *Doublexp*). The fourth column, *bolometric_3*, stores the second timescale parameter, i.e. $t_{\rm fall}$ for *Bazin* and $\tau_2$ for *Doublexp*. Finally, *bolometric_4* only contains $p$, the last parameter of *Doublexp*. Columns number six to ten are reserved for their associated uncertainties, in the same order.

Similarly, six additional columns are reserved for the parameters related to temperature and their uncertainties. *temperature_0* encapsulates $T$ for constant or $T_{\rm min}$ for logistic, followed by *temperature_1* and *temperature_2* which respectively hold $T_{\rm max}$ and $\tau_{\rm color}$ from the logistic model [4]. The least squared error of the fit is stored as an additional column. We also compute the following statistical features for each light curve, assigning each to its own column in the feature matrix:

— *Rising*: a feature describing the state of the transient: 1 if it is purely rising. -1 if it is purely decaying. 0 else ;

— The number of observations in each passband ;

— The peak observed flux in each passband ;

— The mean and standard deviation of the flux across all bands ;

— The mean and standard deviation of the signal-to-noise flux across all bands and

— The duration of the light curve in days.

These statistical properties play an essential role in the extraction of information. In particular, they enable the feature extraction with a very restricted number of observations, even in the extreme case where the light curve contains only 1 observed point. The classifier accuracy for such data will be limited by the information content available, hence, a strong performance evolution should be expected as the number of observations increases. For each light curve (row), cells corresponding to parameters which were not used are populated with -999 value, which account for the lack of information while still enabling the computation of decision trees. Since each function holds a different number of parameters, this configuration ensures that light curves with different levels of complexity will populate completely separate regions of the parameter space [5].

3. https://scikit-hep.org/iminuit/

4. The reference parameter $t_0$ being shared with the function depicting bolometric flux.

5. Both, sigmoid and *Linexp*, contains 3 parameters, but since one is used for *half* and the other for *full* light curves, the *Rising* feature acts to differentiate them.

The final feature matrix is composed of 35 dimensions (columns). This value is small when compared to similar state-of-the-art feature extraction procedures. For example, Sanchez-Saez et al. (2021) proposes 169 features per light curve for the characterization of ZTF light curves. More recently, Cabrera-Vives et al. (2024b) extends the previous features for the ELAsTiCC dataset and produces 429 features per light curve to build a multi-class classifier. This relatively low number of features enables the use of simple, computationally cheap and interpretable ML methods such as Random Forests (Section 4.2.2). Despite all these advantages, it still remains to be proven if the proposed feature matrix holds enough information to enable a successful machine learning classifier. This is done in the next sections.

## 7.2 Dataset

Following Fraga et al. (2024), we use the testing sample released in the first version of ELAsTiCC (Narayan and ELAsTiCC Team, 2023). This dataset was originally built to be used as a test for broker infrastructures (see Section 2.3.4), and it constitutes the best LSST-like alert stream simulation currently available. It uses the 6 LSST passbands (ugrizY) and implements a realistic observation strategy [6]. In addition, the data is formatted into alert packets. This implies that a source can generate many light curves at different stages of its evolution, offering multiple opportunities to classify it. This is considerably more challenging than characterizing an object based on its full light curve, as proposed by its predecessor, PLAsTiCC (Hložek et al., 2020, see also Section 2.3.4).

The ELAsTiCC simulation covers the equivalent of 3 years of LSST operations, and includes ∼52 million alerts, or ∼4 million objects. Figure 7.4 shows the fraction of light curves with more than $n$ observations within LSST-griz passbands. We can see that a quarter of them hold 5 data points or less, and only half contain more than 10 points. In this paradigm, each source should be classified as early as possible, which would enable potential follow-up observation of astrophysically interesting objects. Given the multi-passband nature of the observations and the need to generate early models, this dataset constitutes an ideal testing ground for our adaptive classifier.

The dataset contains 19 classes of variable objects divided into 5 broader categories: SN-like, Periodic, Non-periodic, Long and Fast. In this work, we consider only extra galactic transient sources, with the goal of disentangling them despite their similarities. Potential transient contaminants such as AGN (Section 3.3) are excluded. We work under the hypothesis that highly accurate broad classifiers (e.g. CATS from Fraga et al., 2024) are used upstream to pre-filter alerts before they are subjected to our model. Similar strategies have already been used successfully employed by other science modules within the broker (e.g. Leoni, M. et al., 2022, Pessi et al., 2024), allowing us to take advantage of the available infrastructure and focus on specific science cases.

In total, 8 classes of transients are included in our analysis: SNIa, SNII, SNIb/c, SLSN, TDE, PISN, SNIax, and SN91-bg. We randomly sample up to 1 million alerts of each type

6. https://community.lsst.org/t/baseline-v3-2-released/7877

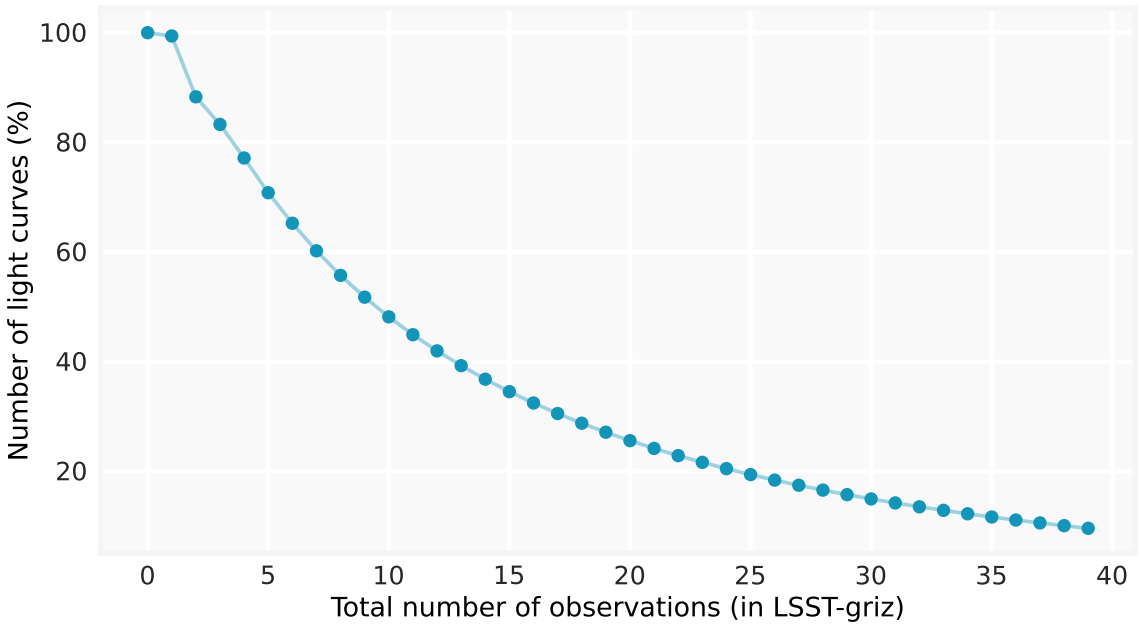



FIGURE 7.4 – Percentage of light curves with at least n observations (in LSST-griz), with n ranging from 0 to 40.

in the analysis[7]. The left panel of Figure 7.5 shows the population distribution in our initial sample. Since only a few alerts are available, kilonovae were excluded from the analysis. In total ∼3.7 million alerts are used, corresponding to ∼1.2 million unique sources. Following the method proposed by Sanchez-Saez et al. (2021), we create a set of 20 training and testing samples. For each of them, objects are randomly split such that 80 % are used for the training of the classifier and 20 % are used for the testing. It ensures that unique sources are only present in one of the two samples. This procedure enables the evaluation of variance in classification results.

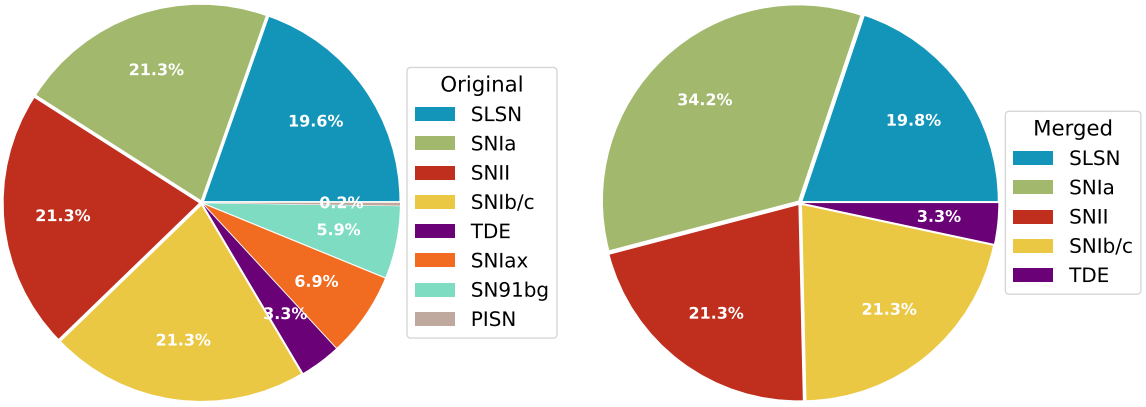



FIGURE 7.5 – Alert distribution of the ELAsTiCC extragalactic transient classes. The left panel shows the original classes. The right panel shows the new distribution after merging similar classes together: {SNIa, SNIax, SN91bg} into SNIa and {SLSN, PISN} into SLSN.

7. For classes with less than 1 million alerts available, i.e. SLSN, TDE, SNIax, SN91bg and PISN, all alerts were included.

A simple analysis has been performed using the original ELAsTiCC classes and is presented in Section 7.3.4. However, an alternative taxonomy is used for the main analysis (Section 7.3.1) because the original one comes with several issues. Not only PISN are much rarer than other classes, they are also technically a subtype of SLSN (see Section 3.1.3). Therefore, similarly to Fraga et al. (2024), the SLSN and PISN are merged into a single class. Additionally, SNIax and SN91bg both represent rare subtypes of SNIa (Section 3.1.2). Thus, they are also combined into a single class. This results in the creation of a dataset, for which the class distribution is shown in the right panel of Figure 7.5. Considering these labels, 4 out of the 5 classes are well represented. TDE constitutes the rarest class, with only 3.3 % of the alerts. It still represents more than 100k events, which is sufficient to inform the model regarding its statistical properties.

ELAsTiCC also contains some metadata information associated to each alert. They include sky position, Milky Way photometric extinction, photometric redshift estimation and information about the host galaxy. Although they contain discriminative information for classification purposes, they have not been included as features for two reasons. Primarily because the goal of this analysis is to understand how much information can be extracted purely from light curves. Adding metadata would improve classification results, but would, at the same time, make it more difficult to identify whether the information is coming from the method or the metadata. Secondly, in a real scenario, it is unrealistic to systematically expect this amount of metadata for every single alert, in particular regarding the host galaxy information.

In what follows, we present the classification results of models trained on ELAsTiCC. This dataset is relatively recent, and therefore only two published papers propose its classification, from which no comparison can be drawn. I have directly contributed to the first one [8], Fraga et al. (2024). This work summaries the efforts developed by the Fink community in preparing binary classifiers using decision trees and deep learning strategies, as well as multi-class classifiers using neural networks. However, none of the feature based classifiers were used in a multi-class task. Moreover, all multi-class classifications considered only the broad ELAsTiCC classes, thus a very different taxonomy than the one used here. The second one, Cabrera-Vives et al. (2024a), provides a global multi-class classifier. However, the authors used exclusively the training sample from ELAsTiCC [9], which consists of full light curves. Hence, a very different exercise than considering the alerts themselves. In addition, the metadata is systematically included, which prevents a direct comparison with our results. Finally, since every dataset is deeply different, we won't attempt to extrapolate transient classification results from other contexts. Results proposed in this chapter may be used as a baseline for the evaluation of future alert classifiers focused on extragalactic sources.

## 7.3 Results

Random forest classifiers [10] (Section 4.2.2) are trained separately on each of the 20 training subsets. In order to produce light and efficient models less likely to overfit the training

8. This work is described in more detail in Appendix C.
9. https://portal.nersc.gov/cfs/lsst/DESC_TD_PUBLIC/ELASTICC/TRAINING_SAMPLES/
10. Using the scikit-learn implementation https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

sample, the maximum tree depth is limited to 40 [11] and each forest was limited to 250 trees. To deal with imbalanced labels, weights inversely proportional to the class frequencies are applied. Performance metrics are reported in the form of the median, the 5th percentile and the 95th percentile of the scores across the 20 iterations.

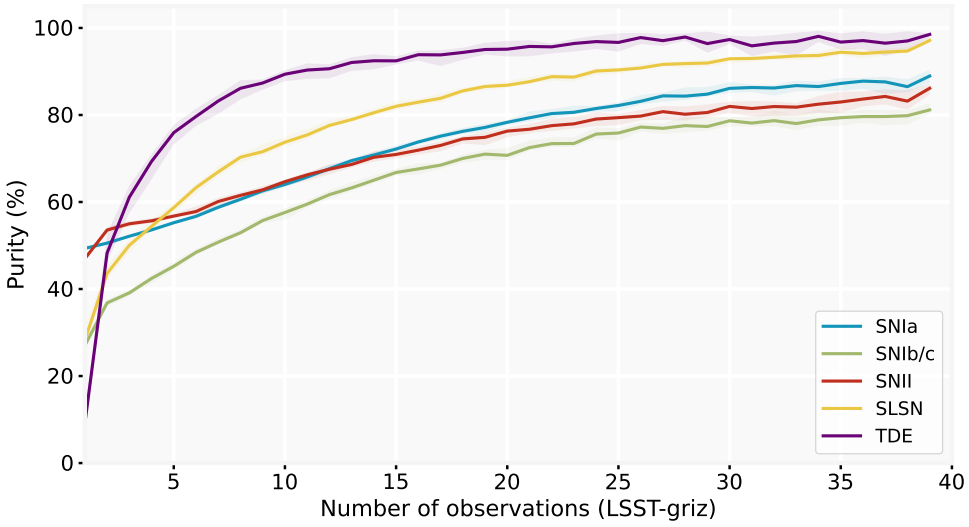
Figure 7.6 shows how model performances evolve depending on the number of available observations, from $n = 1$ to $n \geq 40$. A clear improvement in purity and completeness is observed as the number of observations increases, in particular for the first 15. For well sampled light curves ($> 25$) the classifier reaches very high performance for all classes, with purity ranging from 80% (SNIb/c) to almost 100% (TDE) and completeness ranging from 75% (SNII and SNIb/c) to more than 90% (SLSN). For the least well sampled alerts, the model relies partly on the statistical distribution of the classes. This is clearly translated into the early high completeness of SNIa, the main class in the sample (34.2%).

Based on the performance evolution, we define two scenarios. The first one is the early [12] classification, for light curves containing between 5 and 12 observations [13] (across LSST-griz filters). This category represents by itself a third of all alerts (see Figure 7.4). It is a very challenging task, since it involves making predictions with 1 to 3 observations per passband on average. Nevertheless, given that they represent a significant fraction of the data, it is very important to provide even a crude classification.

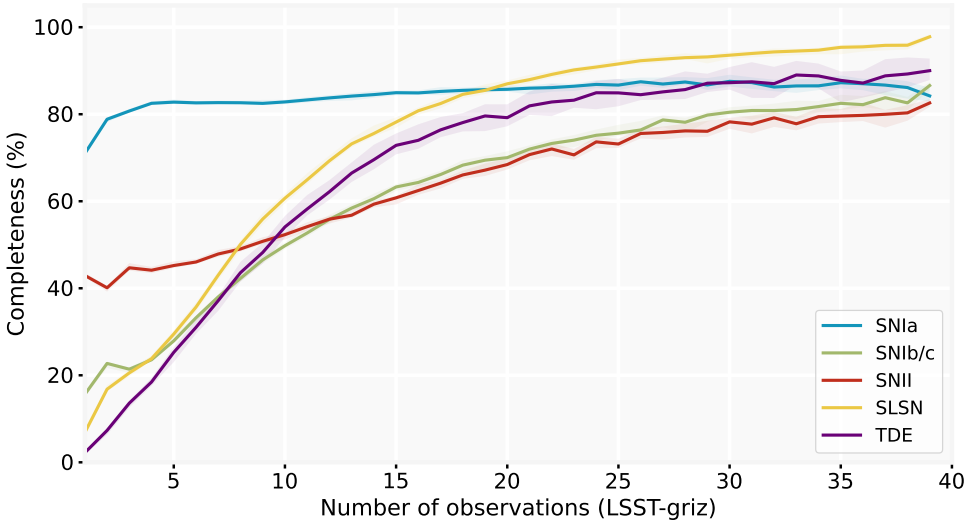
The second scenario concerns alerts with 12 or more observations (LSST-griz), which represents 42% of the alerts. It enables an understanding of the classifier properties within more standard conditions. Indeed, state-of-the-art light curve classifiers based on feature extraction usually require a minimal number of observations. For example, in the exercise of SN classification on LSST-like data, Dai et al. (2018) requires at least one observation before and two after the peak in each passband. Sanchez-Saez et al. (2021) requires at least 6 observations in one of the two public ZTF passband, while de Soto et al. (2024) imposes a minimum of 5 observations in each passband. Therefore, the choice of using light curves with 12 or more total data points is coherent with current expectations. Given that 4 passbands are used for the fit, this corresponds to a minimum average situation where three observations are available in each filter.

In what follows, results on the testing sample are separated, and presented along theses two scenarios. In Section 7.3.1 the classification results with 12 or more observations per light curve are shown. Early classification for alerts holding between 5 and 12 observations are presented in Section 7.3.2. A feature importance analysis of the classifier as a whole is proposed in Section 7.3.3. Finally, Section 7.3.4 presents a brief extra analysis, aiming at classifying alerts using the original ELAsTiCC taxonomy.

11. Deeper trees resulted in no performance improvement.
12. We emphasize that, in this context, "early" relates to the time when the first alert was generated for a given object. It does not necessarily mean the transient light curve itself is in a rising state.
13. We start at 5 because most light curves with less observations won't be fitted.




(a) Evolution of the purity.



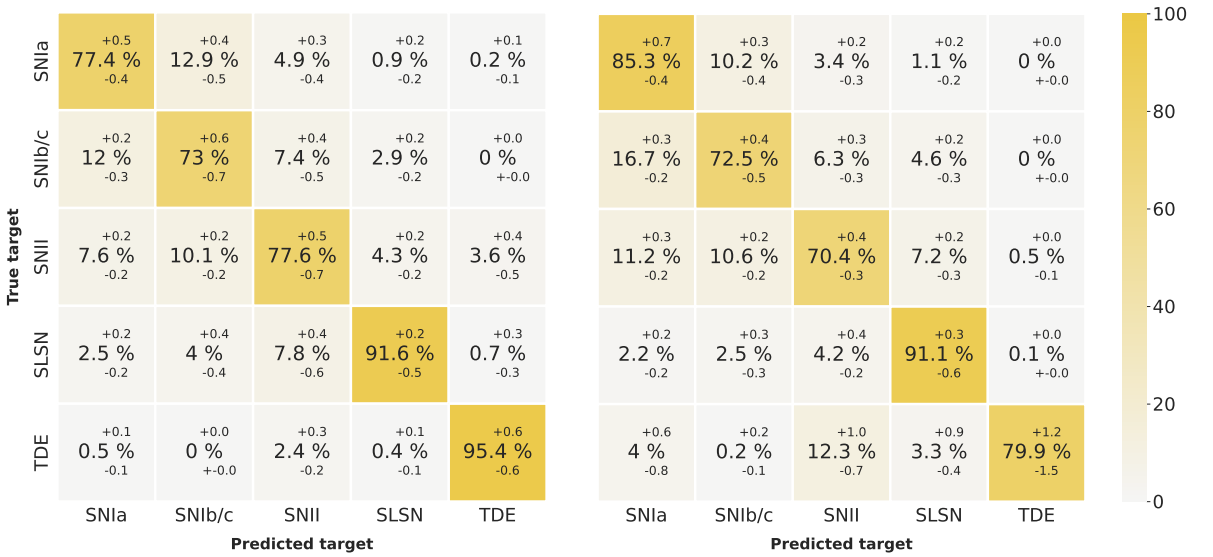

(b) Evolution of the completeness.

FIGURE 7.6 – Evolution of the median score per class as a function of the number of observations in LSST-griz. The uncertainty areas are delimited by the 5th and 95th percentiles.

### 7.3.1 Classification

Figure 7.7 shows the median score of the 20 classifiers when tested on light curves with at least 12 observations (LSST-griz), normalized respectively on purity (Figure 7.7a) and completeness (Figure 7.7b). Each cell also reports the 5th (lower) and 95th (upper) percentile of the results distributions over 20 samples. For common SN types (SNIa, SNII and SNIb/c), purity ranges from 73% to ∼77.5%, with most of the contaminants being other common SN types. An 8% misclassification of SLSN to SNII can be observed. This is most likely due to light curve plateau behaving similarly to slow SLSN events. For TDE and SLSN, results are particularly good, with more than 90% purity. Again, most contamination comes from SNII which exhibits slower evolution than other SN types. Regarding completeness, the model is highly efficient for SNIa with 85% and SLSN with 91%. Other common SN types have completeness above ∼70%, with most missed objects being classified as SNIb/c or SNIa. Finally, the classifier shows 80%

completeness on TDE. This result is particularly good given the great imbalance between classes in the dataset. Although they only represent 3% of the sample, the feature extraction was informative enough to learn the properties from this class. The confusion with SNII is again observed for TDE (as well as SLSN), with 12 % of the missed TDE being classified as SNII (and 3% being assigned to the SLSN class). Overall, the classifiers have relatively low variance across the 20 iterations. The 5th and 95th percentiles displays low variation of the results, typically below $\pm 0.5\%$ indicating that the results are robust over changes of subsamples. The highest variations are associated to the completeness of TDE, for which they reach $\pm$ 1.5%. This amplitude is due to the smaller size of the TDE sample.




(a) Confusion matrix normalized on purity.

(b) Confusion matrix normalized on completeness.

FIGURE 7.7 – Confusion matrices for light curves with at least 12 observations (in LSST-griz). For each cell, the central value indicates the median score of the 20 independent models. Lower and upper values within cells indicate 5 and 95 percentiles, respectively.

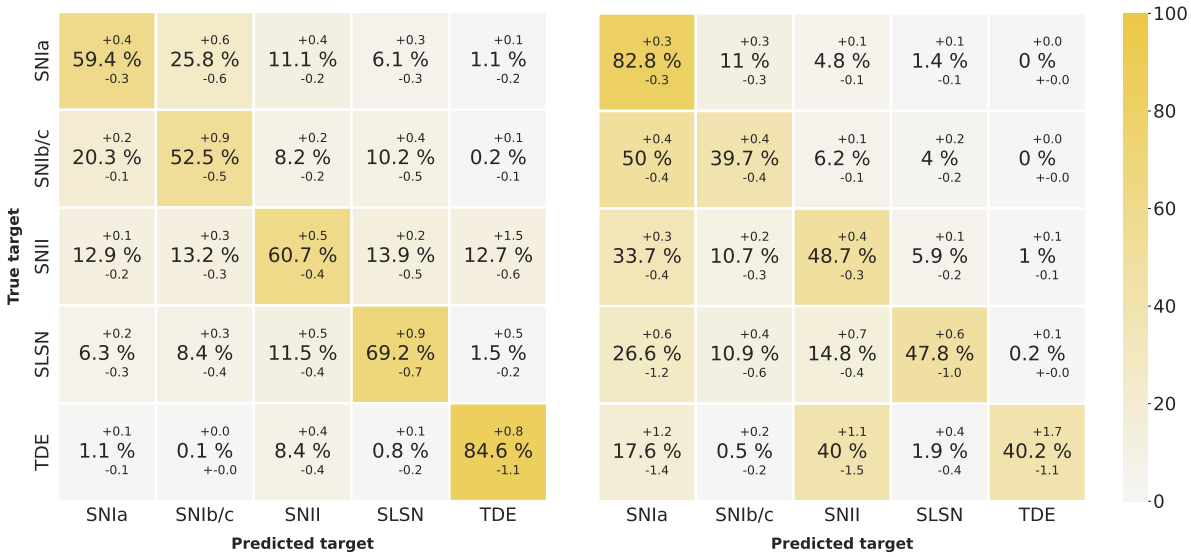
These results illustrate the high descriptive power encoded in the features extracted from light curves with 12 or more observations. Despite the similar behavior of most transient classes, predictions can be trusted with a high confidence. Moreover, the reported miss-classifications are expected within given known light curves characteristics. For example, SNII and SLSN, or SNIa and SNIb/c are classes known to behave similarly. Although they are wrong predictions, they indicate that the properties learned from the features are physically coherent. These results are especially good considering that no metadata information was included, In particular this high confidence does not require redshift information.

### 7.3.2 Early classification

Figure 7.8 shows confusion matrices from the classifiers tested on light curves containing 5 to 12 observations (LSST-griz), normalized respectively on purity (Figure 7.8a) and completeness (Figure 7.8b). As expected, the significantly lower information content in each light curve translates to a less precise characterization. For all classes but one, the purity is $\gtrsim 60\%$, reaching 85% for TDE. SNIb/c shows the lowest purity, 52.5%. An overall confusion between SNIa, SNIb/c

and SNII is observed, with a strong misclassification of SNIa to SNIb/c (26%). Similarly to results presented before (Figure 7.7), a confusion between SNII and slow evolving classes (SLSN and TDE) is observed. In particular, objects wrongly classified as TDE are almost always SNII.

Concerning completeness, the early classification relies partly on the class statistics. Since SNIa are the most common transient in the sample, any light curve with no information suggesting a class will preferentially be classified as SNIa. This bias, that could already be observed in Figure 7.6b, leads to an inflated SNIa completeness of 83 %. All other classes reach between 40 and 48 % completeness, meaning that a little less than half of the alerts of each type are identified correctly. Beyond the global misclassification towards SNIa, the confusion between SNII, SLSN and TDE is observed. A surprising confusion of 11% of SLSN to SNIb/c is also present, suggesting that over shorter time scales these transients are more difficult to distinguish. The score variation over the 20 samples is relatively small, reaching at maximum ± 1.5% for TDE, showing that the classification results are robust despite the smaller sample size.




(a) Confusion matrix normalized on purity.

(b) Confusion matrix normalized on completeness.

FIGURE 7.8 – Confusion matrices for early light curves with 5 to 12 observations (in LSST-griz). For each cell, the central value indicates the median score over 20 independent models. Lower and upper values indicate 5th and 95th percentiles, respectively.

Although early classification is very challenging, the model offers satisfying results. Overall, the classifier has a 60% accuracy, suggesting that the feature extraction was particularly efficient at summarizing relevant early light curve information. The purity matrix indicates that, even in such context, any class prediction is more likely than not to be correct. The high purity for TDE classification in particular is crucial. Indeed, it could lead to follow-up observations, which would be extremely valuable given the little observational data available on TDE [14]. Overall, the classifier enables reliable classification of early alerts. Although predictions should be considered with more care, they are essential since early events represent one third of the total alerts available.

14. In Appendix D, an extra work on early feature extraction based TDE classification inspired by this method is presented.

### 7.3.3 Feature importance

The importance of features in the classification process are shown in Figure 7.9. The most relevant feature in the classification of transient is the duration of the light curve. This is not surprising, since it represents an immediate way to separate longer classes (SNII, SLSN and TDE) from shorter ones (SNIa and SNIb/c) when the light curve is sampled enough. The series of simple statistical features based on flux (peak flux per band, and standard deviation/mean of the flux and on the signal-to-noise-ratio) stand among the most informative features. They are general and they directly summarize the observations without prior, hence they should always be used for light curve classification. These values are computed from all 6 passbands, in opposition to the fit that is performed only on LSST-griz. Therefore, they encode some information about the transients only accessible through them.

The third most relevant feature is the *temperature*_0 parameter. It reveals the importance of the RAINBOW procedure. Taking into account the multidimensional nature of the light curves appears as essential information. The two other temperature parameters also constitute important features (17th and 18th). The quality of the fit (12th) also plays a major role in the classification process. Indeed, it indicates whether the transient matches our blackbody assumption or not. In combination with bolometric parameters, it also describes the evolution of the transient and to which point the fit is reliable. We observe that *bolometric*_4 (34th), and *bolometric*_3 (28th) to a smaller extent, are significantly less informative than the first ones (13th, 14th, and 16th). It was to be expected since it corresponds to the sparser bolometric columns (i.e. they contain the most -999), and characterizes only the most well sampled *full* light curves. Finally, minor information are brought by the number of points per passband and the uncertainties on the optimized parameters.

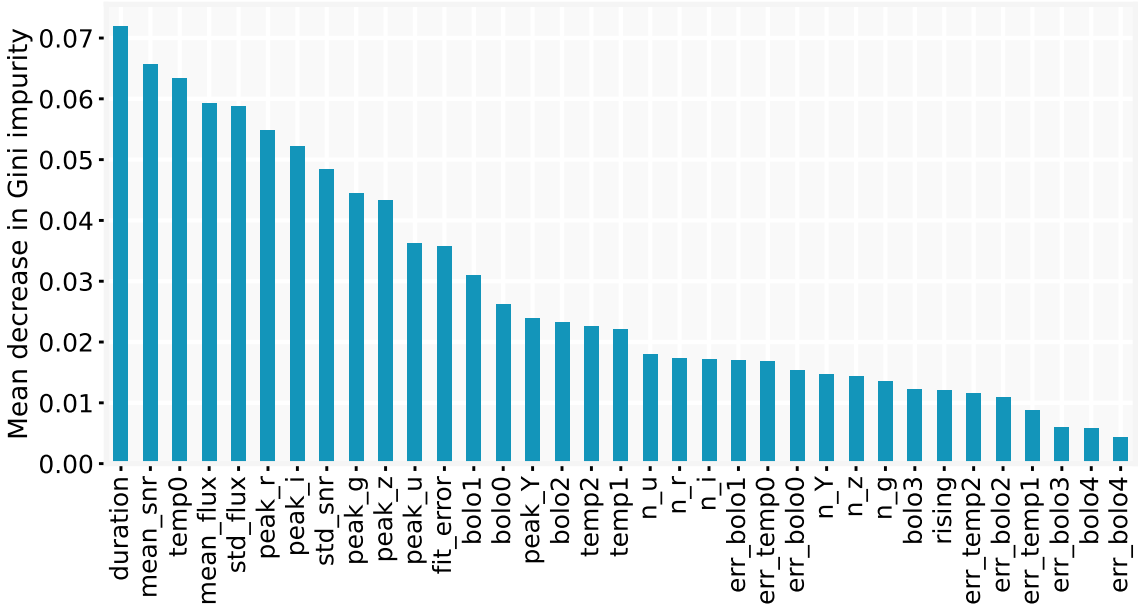



FIGURE 7.9 – Feature importance of the random forest classifier trained on merged classes.

### 7.3.4 Original taxonomy

The original taxonomy proposed in the ELAsTiCC dataset comes with few issues that justified the merging of classes used in the Section 7.3.1 and 7.3.2. However, this section still presents an extra analysis of a separate classifier trained with the original classes. Although it would be less realistic to use in practice, studying its performances can provide valuable information about the classifier itself and the representation of transients in the feature space.

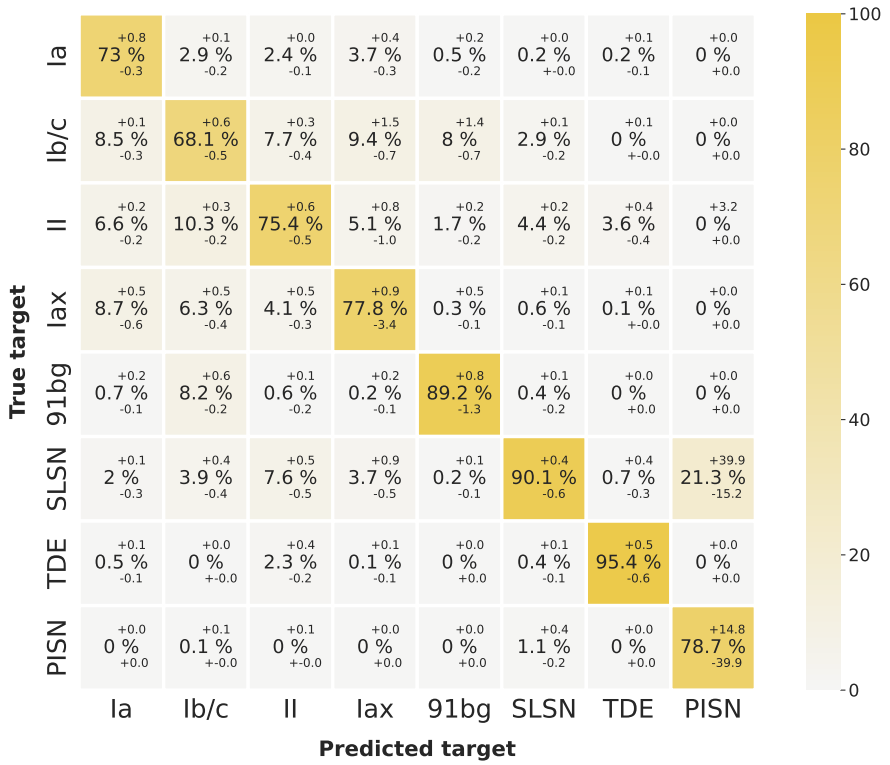
Figures 7.10a and 7.10b show confusion matrices of the classifier on light curves containing at least 12 observations (LSST-griz), normalized respectively on purity and completeness. Overall, using this original taxonomy mostly affects the classification results for classes that are more similar among each other. For example, almost no decrease in performance is observed for SNII, SLSN and TDE. But, it results in a $\sim 5$ % decrease in purity and completeness for SNIb/c. For SNIa however, completeness improves from 85 to 92.5 %. This is due to the separation between standard SNIa and rarer more difficult to classify subtypes (SNIax and SN91bg-like), making the SNIa class more consistent. The rare SNIa subtypes are identified with high confidence, 78% for SNIax and 89% for SN91bg. Most objects wrongly classified as a SNIa subtype are, in reality, SNIb/c. This is coherent with their dimmer physical nature (see Section 3.1.2), which makes their light curve more similar to SNIb/c. This confusion explains the clear drop in classification performance regarding SNIb/c.

Although purity results from SNIa subtypes is high, they are often missed. The classifier has 46% completeness on SN91bg, with almost all missed ones being classified as SNIb/c. An even lower completeness of 18% is reached for SNIax, with 35% being classified as regular SNIa, 29% as SNIb/c and 15% as SNII. Regarding the PISN class, results are clear. Almost all of them ($\sim$90%) are misclassified as SLSN. The classifier purity on PISN is high, but given the associated variance among different subsamples, no conclusion can be drawn. This effect is due to the very small fraction of PISN in the sample (0.3%). For very rare events, a classifier is not an adapted tool, and methods such as anomaly detection should be considered instead (see Appendix B). However, the fact that PISN overlap SLSN in the parameter space is physically coherent (Section 3.1.3).

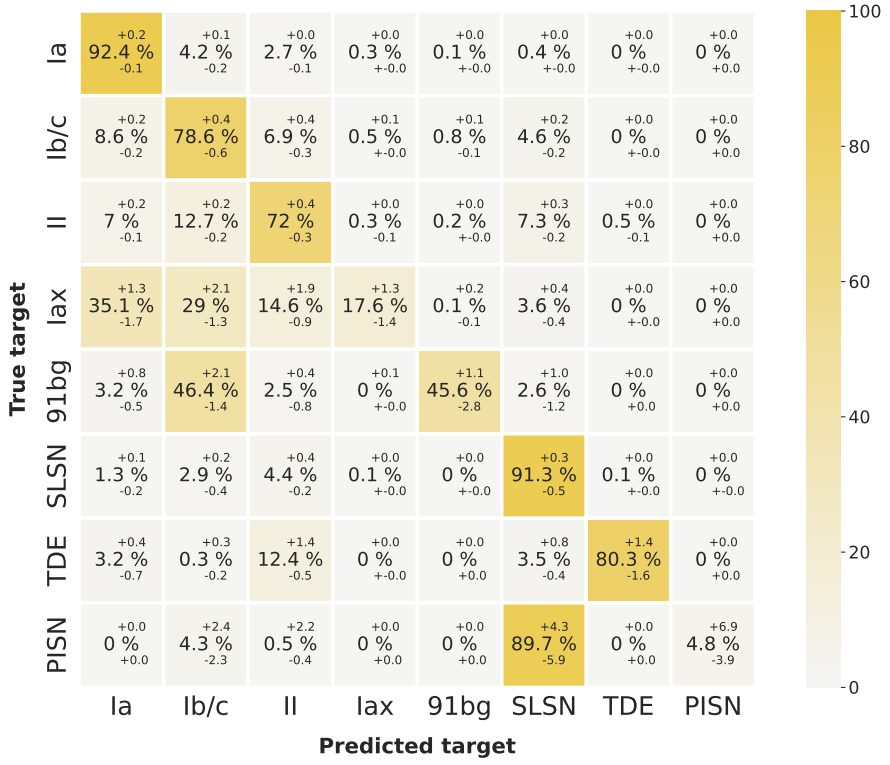
## 7.4 Conclusion

Parametric light curve fitting offers an efficient and interpretable way to summarize the information of light curves into a small set of features. In this thesis, efforts have been put into the improvement of this procedure. MvSR constitutes an automated method to generate parametric models suited for different types of light curve behaviors, while Rainbow provides a solid framework to deal with multi passband analysis, and requires fewer observations to characterize an object. Despite these improvements, other challenges remains to be tackled. The choice of parametric model for individual light curves is one of them. Ideally, the function used should have the right complexity to neither overfit nor underfit the transient, resulting in the extraction of highly informative features.

This chapter proposes an adaptive fit as a possible answer to this question. The method is based on the evaluation of the state of evolution of each light curve, enabling the correct




(a) Confusion matrix normalized on purity.




(b) Confusion matrix normalized on completeness.

FIGURE 7.10 – Confusion matrices for light curves with at least 12 observations (in LSST-griz) using the original ELAsTiCC taxonomy. For each cell, the central value indicates the median score over 20 independent testing samples. Lower and upper values indicate 5th and 95th percentiles, respectively.

parametrization given the amount of information available. Using the ELAsTiCC dataset, we extracted features from extragalactic transient classes using the adaptive fit procedure, with the goal of evaluating the quality of the features on a classification exercise. For this task, we used a simple Random Forest algorithm, but more complex ML methods could be used in the future to optimize the classification score. We show that the method yields high separation between classes. In particular, the adaptive procedure is able to describe objects with very few observations. In this context, despite the difficulty of disentangling such similar events with a reduced information content, the classifier achieves good performances, that could be used as solid hints to further investigate interesting candidates. However, we observe that for early predictions, the model relies on the statistics of the training sample. We chose not to focus on the optimization of the classifier itself, but methods such as active learning (Settles 2012, Leoni, M. et al. 2022) could be used to further improve the results. By crafting a smaller balanced dataset, selecting only the most informative light curves, the model would not only be more robust to statistical bias but also lighter, easier to deploy for real time alert stream and open to the possibility of training with a small number of real light curves and spectroscopically confirmed labels.

Photometric time domain surveys, like ZTF or LSST, provide information in an incremental manner, with each new alert adding additional clues about the nature of an astrophysical event. In order to extract as much value as possible from light curves, our methods should be designed to deal with this reality. In this context, methods based on parametric fit are, generally, not ideal since they are built upon rigid models and are unable to describe all possible scenarios in an accurate and simple manner. The adaptive feature extraction represents an effort to overcome this challenge, intrinsically linked to the study of transient sources. This type of approach enables early characterization of light curves, which is mandatory to allow subsequent follow-up. The arrival of LSST, increasing by two orders of magnitude the volume of data produced each night, will further strengthen the necessity to adapt our methods. The number of candidates that will be available, coupled with our limited spectroscopic time, implies that classifiers must provide the best performances possible to make optimal use of our resources. However, classifiers are ultimately meant to be used by astrophysics experts interested in a particular type of object. Hence, the decision process should also be as accessible and interpretable as possible, fostering the collaboration between machine learning and astrophysics experts. The work presented in this chapter is a practical example on what results can be achieved by such interaction.

# 8

# Conclusion

*“Odi panem quid meliora“*

– Loth., *2006*

Feature extraction is one of the standard methods used for the characterization of light curve from time domain surveys. In the study of transient events, a common procedure consists in fitting a model to the data and using the minimized parameters as features to be input in machine learning applications. This thesis implements a series of methods built to optimize the amount of information extracted from this procedure. In Chapter 5, I presented Rainbow a physically motivated parametric framework able to model the multidimensional nature of transient light curves. I showed that it largely improved their characterization compared to approaches treating passbands independently. In Chapter 6, I tackled the question of the construction of the parametric functions themselves, and proposed MvSR, a Symbolic Regression method capable of automatically generating models based on examples. This versatile tool can be used to construct functional forms tailored to the specificities of a problem. Its data driven nature overcomes human biases regarding the model construction and is able to propose efficient novel functional forms. Finally, in Chapter 7, I proposed a solution to overcome the issue of fitting light curves at different stages of evolution. Using the Rainbow framework and combining models generated with MvSR and from the literature, I proposed an adaptive method, able to choose an adequate description based on the amount of information available for each light curve. Such approach appears essential when dealing with alert data-stream, as it enables early classification of events, and therefore follow-up observations.

Feature extraction optimization largely contributes to the machine learning community efforts of classifying light curves. However, it must be acknowledged that some deep learning methods, such as Recurrent Neural Networks, do not require features and may directly classify the raw data. Despite that, working with features still appears essential. First because they are not antagonist with deep learning methods, and can easily be combined to construct powerful pipelines. Then because the parameter space built from feature extraction are generally less complex, and can thus be learned even with a limited amount of labeled data. Hence, it is adapted to a broad range of real data scenarios. These two strategies build upon intrinsically different assumptions regarding data characteristics and the role of models, thus providing significantly different views of the data. Whenever possible, combining results from both approaches is beneficial (e.g. Leoni, M. et al. 2022, Fraga et al. 2024) and such efforts should be encouraged. However, this exercise is out of the scope of this work.

All methods developed in this manuscript share a common philosophy regarding the processing of astronomical data. Machine learning and astrophysics experts are not seen as two groups impervious to each other, but rather as cooperating from the conceptual stages of all projects. I conducted this thesis as an active member of the Fink broker and the SNAD team, both networks which share this collaborative vision. In enabling the development of science modules by domain experts, we ensure that outcomes from implemented pipelines will be used in practice by the community. The work performed in this thesis concerned both, the construction of such pipelines (Fraga et al., 2024, and Appendix D) and the elaboration of tools easily usable for future science cases. Indeed, by being able to decide which model to use in which context, visually inspect the light curve fits, interpret the parameters values, and analyze the relative feature importance, the domain expert has the opportunity to guide development towards a specific science goal.

Such interdisciplinarity inevitably generates more accessible tools, expanding their usage to a larger audience. This is highlighted by MvSR, which have been created from our collaboration with computer scientists. The versatility of the method quickly lead to the expansion of the analysis to include other science cases. Although they are deeply different from astronomical data, the discussions they brought yielded valuable insights for the understanding and improvement of MvSR. Moreover, as proven by the interest it generated in various fields, MvSR has the potential to be used in even more scenarios. Improved versions are currently being developed, with particular efforts being put towards the construction of an accessible user-friendly interface.

The flexible nature of Rainbow also shows the importance of interdisciplinary tools. It is implemented within a well-established feature extraction package [1] already used by three different community brokers: ANTARES, AMPEL, and Fink. Originally developed for a precise science case, the right expert knowledge has rapidly revealed its potential in other contexts. It is currently being implemented within two Fink early classification science modules, for SNIa (Appendix C), and for TDE (Appendix D). Pushing the accessibility, future development from Fink will also integrate an option to produce a Rainbow fit on any alert light curve, allowing anyone to quickly evaluate the evolution of a transient.

Astronomy is quickly entering a new era, where the data produced is of such complexity that it requires both a high level of expertise to be understood and intricate machine learning models to be processed. In this context, collaboration is the key to achieving meaningful scientific goals. The work proposed in this thesis constitutes a modest step in this direction.

1. https://github.com/light-curve/light-curve-python

# A

## Rainbow confusion matrices

Classification confusion matrix of Random Forest models trained on Rainbow (Figures A.1 and A.3) and Monochromatic features (Figures A.2 and A.4). These results were used to construct those shown in Figure 5.6 and Figure 5.8.

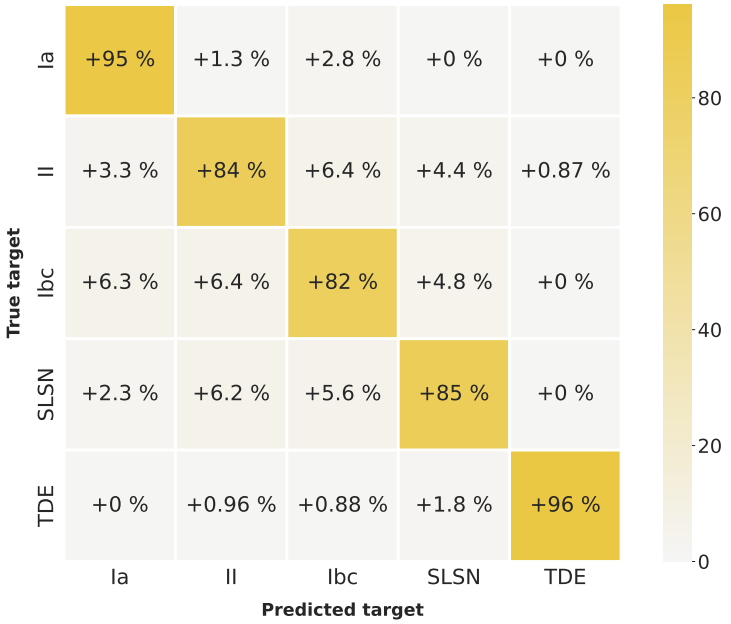



Figure A.1 – Full light curve scenario. Confusion matrix normalized on purity of the Random Forest classifier trained on Rainbow features. The dataset is composed of 300 light curves of each class (SNIa, SNII, SNIbc, SLSN and TDE). Numbers represent the median score of 100 iterations of bootstrapping.

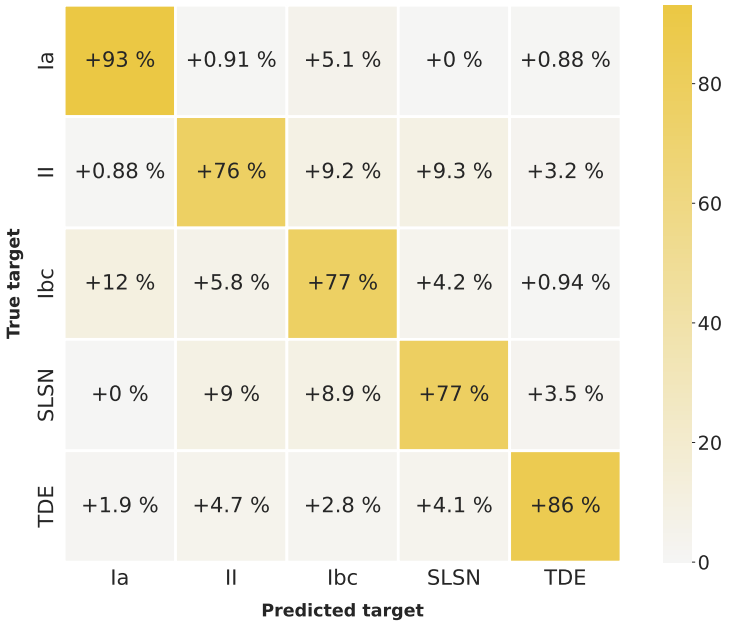



Figure A.2 – Full light curve scenario. Confusion matrix normalized on purity of the Random Forest classifier trained on Monochromatic features. The dataset is composed of 300 light curves of each class (SNIa, SNII, SNIbc, SLSN and TDE). Numbers represent the median score of 100 iterations of bootstrapping.

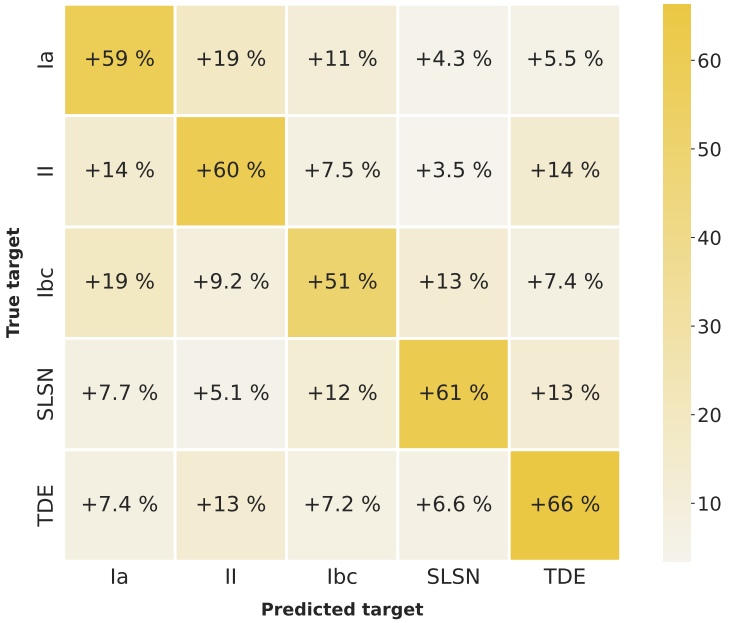



FIGURE A.3 – Rising light curve scenario. Confusion matrix normalized on purity of the Random Forest classifier trained on RAINBOW features. The dataset is composed of 250 light curves of each class (SNIa, SNII, SNIbc, SLSN and TDE). Numbers represent the median score of 100 iterations of bootstrapping.

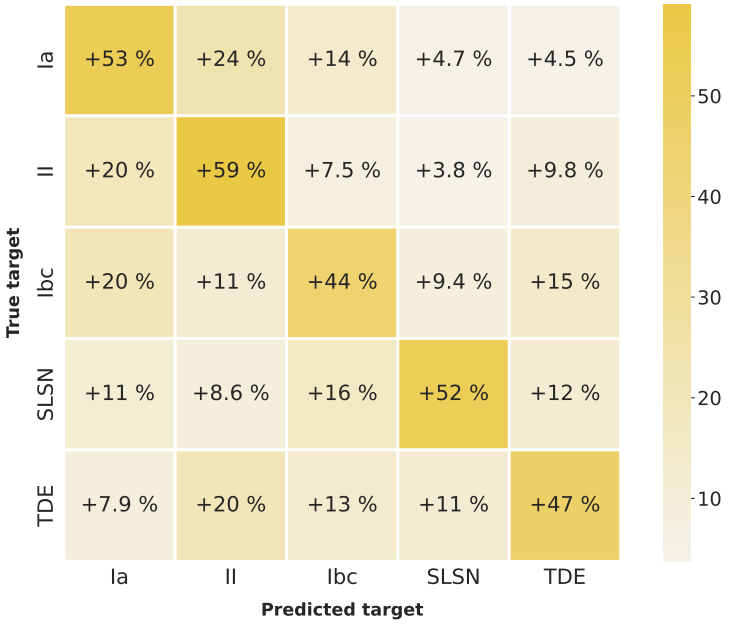



FIGURE A.4 – Rising light curve scenario. Confusion matrix normalized on purity of the Random Forest classifier trained on MONOCHROMATIC features. The dataset is composed of 250 light curves of each class (SNIa, SNII, SNIbc, SLSN and TDE). Numbers represent the median score of 100 iterations of bootstrapping.

# B

## SNAD

During this thesis, I have been an active member of the SNAD collaboration [1], an international group of researchers mainly working on anomaly detection for astronomical datasets. The core algorithm used for this purpose is the isolation forest (Liu et al., 2008), an unsupervised ML technique based on tree ensemble. It isolates observations by recursively partitioning the data through random selection of features and split values. The underlying assumption is that anomalies, being rare and distinct, require fewer partitions to be isolated compared to normal data points, resulting in shorter average path lengths in the constructed trees. As a result, an anomaly score is attributed to each instance based on the average path length in a forest of trees.

Both Rainbow (Chapter 5) and MvSR (Chapter 6) started as SNAD projects. Although they do not involve anomaly detection directly, they aim at improving the feature extraction, which is at the basis of many ML analysis. These projects emerged from the propositions, discussions and debates that occurred at the 2022 SNAD-V workshop [2]. This annual event is one of the three [3, 4] SNAD workshops that I have attended. From these environments, where the discussions and inter-collaborations are constant, multiple thesis side projects have emerged. Although I was not the leading author, I briefly summarize below the published works that I have contributed to:

— In Pruzhinskaya, M. V. et al. (2023), we explore the potential of active learning applied to anomaly detection on astronomical data. Indeed, anomalies discovered by classical isolation forest methods are statistically rare (called outliers) but not always scientifically interesting. Real telescope measurements can be subject to artifacts like cosmic particles, CCD leakage or satellite tracks, such that the resulting light curve will be very different from normal ones. In large datasets, this type of outlier tend to dominate the set of objects with high anomaly scores. Furthermore, the definition of an interesting astrophysical anomaly is not absolute, but rather, relative to the interest of the expert analyzing the data. For these reasons, we used Active Anomaly Detection (AAD) algorithm proposed by Das et al. (2017). It is an adaptive learning strategy, whereby at each iteration, a binary reply from the expert is incorporated into the weight calculation of an isolation forest model, producing updated anomaly scores. After several iterations, the model adapts and outputs more frequently anomalies which are interesting to the expert.
We apply this method on the data release 3 (DR3) of ZTF (see Section 2.3.1), considering as anomalies any light curves that resemble those of supernovae (SNe) and searching for uncatalogued or anomalous transients. From the 2100 objects visually inspected, we found 104 SN-like events, 57 of which were reported for the first time. Among the newly found transients, we reported three objects (SNAD121, SNAD160 and SNAD187) with broad, slowly evolving light curves that stand as promising superluminous supernova candidates. The results presented here confirm the effectiveness of adaptive learning approaches in filtering large astronomical datasets for expert analysis. In this work, I contributed to the effort of visually inspecting and classifying the candidates proposed by the AAD loop. I also participated in the writing and the production of plots for the publication.

— Pruzhinskaya et al. (2022) is a research note investigating in more details the hypothesis that SNAD160, the SLSN found in Pruzhinskaya, M. V. et al. (2023), could be a pair

1. https://snad.space/
2. https://snad.space/2022/
3. https://snad.space/2021/
4. https://snad.space/2023/

instability supernova (see Section 3.1.3). We find that it is largely compatible with PISN models, however without deeper spectroscopic analysis, other SLSN models cannot be ruled out.

— In Malanchev et al. (2023), we present in technical details the SNAD viewer, a web portal for astronomers which presents a centralized view of sources from ZTF and includes data gathered from multiple publicly available astronomical archives. It also proposes a set of features automatically extracted from light curves for the expert to scrutinize. This set of feature is computed using the python *light-curve* package [5]. Initially, it has been developed by the team to provide efficient expert feedback in the context of adaptive machine learning, enabling various SNAD projects such as Malanchev et al. (2021), Pruzhinskaya, M. V. et al. (2023) and Aleo et al. (2023). However, it grew beyond the original goal and is now used by an international community of astronomers. In addition, it is currently integrated into the ANTARES and Fink brokers (Section 2.3.3), as well as into the Young Supernova Experiment marshal (Coulter et al., 2022). I contributed to the work by implementing the Rainbow (Chapter 5) framework within the *light-curve* package, enabling the computation of features in the form of optimized parameters. Although the features are not yet available on the SNAD viewer, the development is ongoing, and they will be accessible in the near future.

— In Korolev et al. (in prep.) [6], we propose the *pineforest* algorithm [7], a new active anomaly detection algorithm. Similarly to AAD (Das et al., 2017), it is based on an adaptive isolation forest. In our version, at each iteration of the loop, rather than reweighing branches, trees that proposed outliers which are not considered interesting by the expert are replaced by new random ones. We show on toy datasets as well as ZTF light curves that our method is more efficient than AAD at learning the type of outlier requested by the expert, i.e. for a given amount of loops, *pineforest* finds significantly more anomalies. Our benchmark also shows that *pineforest* is computationally faster. I contributed to the development by testing the method on various datasets. I further helped by developing tutorials and notebooks to demonstrate the effectiveness of the method for non-astronomy audiences.

5. https://github.com/light-curve/light-curve-python
6. https://github.com/snad-space/coniferest
7. https://github.com/snad-space/coniferest

# C

## ELASTICC

The ELAsTiCC challenge, described in Section 2.3.4, not only constitutes the most up-to-date database of LSST-like alerts, but was also presented as a classification challenge for broker teams. As a member of Fink, I joined the collaborative effort to propose an answer to the challenge. In Fraga et al. (2024), we present a series of classifiers based on various ML methods and targeting different goals. It includes two broad [1] deep learning classification frameworks based on recurrent neural network: an adapted version of SuperNNova (Möller and de Boissière, 2020a), and the CBPF Alert Transient Search (CATS). The latter is a multi-class classifier whose goal is to disentangle the five broad classes: SN-like, Fast, Long, Periodic and Non-Periodic. SuperNNova proposes an ensemble of binary classifiers for each broad class, as well as a multi-target one. In addition, we propose two specific models that constitute my contribution in the collaborative work (all details are available in Fraga et al., 2024).

## C.1 Superluminous Supernova classifier

The superluminous supernova (SLSN) classifier is based on feature extraction of normalized alerts followed by a random forest classifier. Within ELAsTiCC taxonomy, SLSN and PISN (Section 3.1.3) are two distinct classes, however, since their predicted light curves can have similar morphology, we use a common classifier for both and call it the SLSN classifier. For each filter, we compute the following set of features: maximum and standard deviation of the flux; mean signal-to-noise ratio and number of points. We also added the following metadata information: right ascension (`ra`), declination (`dec`), host galaxy photometric redshift (`hostgal_zphot`), host galaxy photometric redshift error (`hostgal_zphoterr`) and distance between the host and the transient (`hostgal_snsep`). In addition, parametric fits of the light curves in passbands $r$ and $i$ are computed using the *Linexp* equation:

$$f(t) = A(t - t_0) \times e^{-\frac{t-t_0}{t_{fall}}}, \tag{C.1}$$

which depends on amplitude ($A$), a time offset ($t_0$) and a characteristic time of decay ($t_{fall}$). This functional form was founds using MvSR on a ZTF light curves from the SLSN candidate SNAD160 (Pruzhinskaya et al., 2022, see Section 6.4.4). We found this simple functional form to describe accurately both SLSN-I and PISN. The optimized parameters fitted on passbands $r$ and $i$ and the root-mean-square error are included as features for the classifier, thus we impose for each alert to contain at least 3 observed points in passbands $r$ and $i$. Finally, the standard deviation and maximum absolute value of the $r - i$ color were calculated by using the interpolation from the fit. In total, 39 features are extracted for each alert [2].

The classifier is based on a random forest algorithm trained using the active learning (AL) procedure, i.e. the training sample is built by iterative selection of the most informative alert (exact procedure described in Leoni, M. et al. (2022). This strategy allows the classifier to focus on the relevant boundaries between SLSN and similar transients, rather than simply learning the global distribution between very unbalanced classes. This procedure tends to favor purity over

1. ELAsTiCC classification is hierarchical, with broad classes (e.g. SN) above specific classes (e.g. SNIa).
2. This footnote is added at the last minute before the manuscript final submission. The independent $r$ and $i$ fit has been replaced by a Rainbow fit. This resulted in vastly superior classification results, in particular completeness-wise. Definitive updated results will soon be made publicly available.

completeness, which is reasonable given the volume of alerts that will be produced by LSST. The model was trained on a sample of 21000 alerts chosen optimally via AL from the first year of available data. The other two years are used as a validation sample, it represents almost 5 million alerts corresponding to ∼3 million distinct objects. The performances of the classifier are shown on the left panel of Figure C.1. It provides an excellent purity of 91 % with a relatively low recall of 27.4 %. In the context of the very large ELAsTiCC data set, we favor this high precision low recall asymmetry as it would still result in more than 100K SLSN alerts being classified with high confidence. By using the CATS model upstream, and passing only the Long objects into the SLSN classifier, we can reach a purity of ∼ 97%.

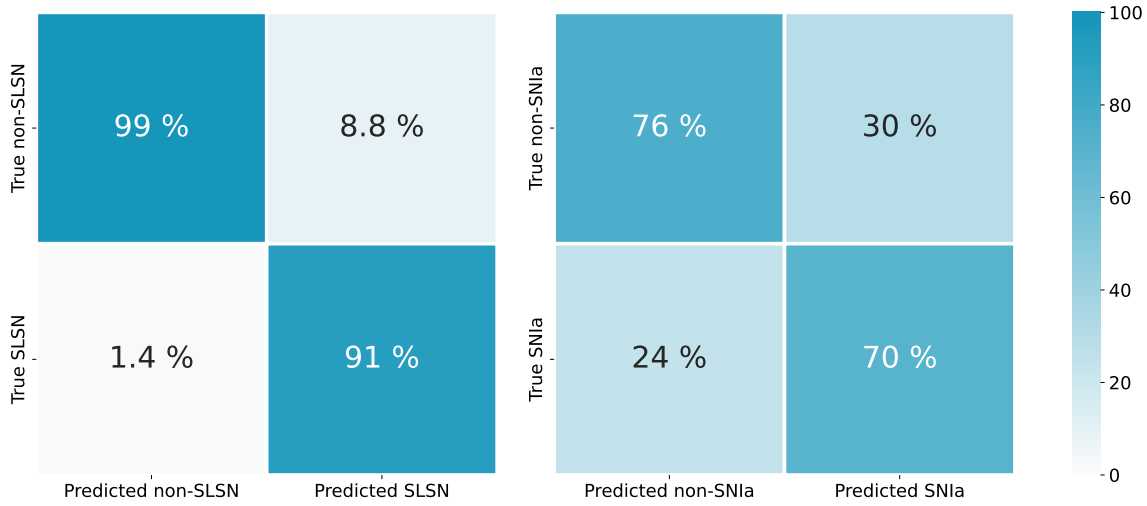



FIGURE C.1 – Confusion matrices for the SLSN (left) and the Early SNIa (right) classifiers. Results are normalized on purity.

## C.2 Early Supernova Ia classifier

The second classifier is the early SNIa classifier, an adaptation from the current science module running on ZTF alerts (Leoni, M. et al., 2022, , hereafter, EarlySNIa). It offers a binary classification for early – i.e. before peak luminosity – SNIa. In the context of real data, labelling is an extremely expensive process, and ideally we would like to discover such transients early enough so they are still sufficiently bright to allow spectroscopic classification. EarlySNIa is based on independent feature extraction for each of the 2 ZTF passbands and a random forest classifier enhanced by AL. However, in the context of LSST, with 3 times more passbands and a considerably sparser cadence, the module required significant modifications.

In order to allow classification with a lower number of points per filter and, at the same time, take into account color information, we implemented the RAINBOW feature extraction (Chapter 5) to comply with the characteristics of the new data set. A parametric model was simultaneously fitted to the light-curves in all available passbands, and the best-fit parameter values were used as features, thus given as input to the random forest classifier. This approach enables early description even when the number of observations in each filter is significantly limited. The preprocessing for each alert included:

1. Averaging all observations within the same night.
2. Removing any intra-night flux measurements lower than -10 (FLUXCAL > -10).
3. Requiring a minimum of 7 points per object, in any filter, including forced photometry.
4. Ensuring that intra-night flux measurements are consistently increasing within at least 2 passbands

Thus, considering that only rising alerts survived such selection cuts, we described the bolometric evolution of our light-curves with a sigmoid function. The temperature evolution was described with the falling logistic function given in Section 5.1.2, equation 5.7. Beyond these, we also use the measurement of the quality of the fit and the maximum measured flux as features. As a result, each alert is represented by 7 values. To this we added the mean signal-to-noise ratio (`FLUXCAL/FLUXCALERR`) ; the number of points in all passbands before intra-night smoothing (`nobs`) ; separation between the host and the transient (`hostgal_snsep`) and the host photometric redshift (`host_photoz`). Thus resulting in 11 parameters per alert.

The first year of ELAsTiCC was used as a training sample. Approximately one third of the alerts passed the cuts ($\sim$ 6.5 million), among which one third are SNIa. From these we selected a sample of 114 701 alerts for designing our experiments. We trained a binary classifier whose positive class (i.e. SNIa) corresponds to approximately a third of the full sample. We used half of the unique object identifiers (`diaObjectId`) for training and the other half for testing to avoid information leak between the two subsets. We trained a random forest model and report the scores on the validation sample (2 unseen years) in the right panel of Figure C.1.

One caveat to keep in mind is that the module is only interested in classifying rising light curves. Thus, several alerts are eliminated by selection cuts, never being classified at all. Results presented here correspond to alerts that survived the feature selection. Among these, the module was able to achieve $\sim$ 70% purity and completeness. Given that the classifier was adapted from Leoni, M. et al. (2022), which achieved > 80% purity and $\sim$ 50% completeness on ZTF data, we consider the results very successful. We manage to largely maintain the classification quality, lowering the purity but largely rising the completeness, even though ELAsTiCC is significantly more complex and sparse. It highlights the importance of the efficient multi-passband modelling offered by Rainbow . Note that this classifier has a majority of contaminants within the SN-like broad class, hence the usage of broad classifier upstream is not expected to improve results.

# D

# Early Tidal Disruption Event classification

The RAINBOW framework presented in Chapter 5 displays excellent results. It has been demonstrated that the additional information brought by the blackbody temperature and the removal of deeply correlated parameters produces more robust results. RAINBOW fits (and features extracted from them) are on average better at light curve reconstruction, peak time prediction and machine learning classification accuracy. The latter has shown general improvement by the usage of RAINBOW on the PLAsTiCC dataset, but it has been particularly significant in the exercise of classifying early light curves of TDE. It displayed 66% precision instead of 47% with the MONOCHROMATIC method, for a total of 19% improvement (Figure 5.8). This improvement can be explained by the singular temperature profile of TDEs. They are expected to be purely thermal objects, completely compatible with the blackbody assumption used in RAINBOW (Chapter 5.1). In addition, they should be significantly hotter than supernovae, and should not change their temperature over time. All these properties can be encoded within the temperature parameters of RAINBOW, which will constitute highly discriminant features. Although this result was encouraging for the future of TDE classification, PLAsTiCC is only a simulation. It encompasses a single well-defined TDE model, and the leap to real data condition is often challenging in comparison. Further tests were needed to assess the potential of RAINBOW for the study and discovery of TDEs.

It is in this context that we started a collaboration with TDE specialists from the Institut de Recherche en Astrophysique et Planétologie (IRAP). The goal was to construct an efficient classifier for TDE candidates within the Fink broker infrastructure (Section 2.3.3). Similarly to an already existing science module for the early classification of SNIa, the challenge is to provide an answer before transient reaches its peak brightness. This would enable spectroscopic follow-up of the sources, which can lead to major advancements in the field given the little observational data gathered on TDEs. By constructing a dedicated science module, we would generate a probability of being a TDE associated to each alert. This information would then be publicly available through the Fink portal. Given the rarity of these events, and the challenge of distinguishing them from SN so early, the objective is not to reach high purity but rather high completeness, i.e. that we should not miss any TDE, even at the price of false positive events. The module itself is an ongoing project. In summary, it is a combination of an early description using a sigmoid function and excluding falling transients, a RAINBOW fit using a constant temperature model, cuts based on the prior physical knowledge provided by the experts, and a random forest classifier trained using known TDE light curves.

We use alert light curves from ZTF and work with the public $ZTF-g$ and $ZTF-r$ bands only. We use a sample of 3 months of ZTF alerts, from June to August 2020. In addition, we manually add all TDEs discovered by the ZTF-I Survey (Hammerstein et al., 2022). This sample was visually inspected to create a golden sample, leaving only 10 well sampled and photometrically identifiable TDEs [1]. The main challenge to differentiate TDE from SN light curves comes from their high resemblance during the rising part. A proper usage of RAINBOW is therefore a crucial step, since temperature and rise time are the only photometric information to disambiguate them. We expect them to have a rather slow rise compared to most SNe. The temperature is modeled with a constant temperature, as expected from TDEs. From this, we expect them to be hotter and generally provide a better fit when compared to SNe. Similarly to

1. Their ZTF identifiers are: ZTF17aaazdba, ZTF19aabbnzo, ZTF19aapreis, ZTF19aarioci, ZTF19abhhjcc, ZTF19abzrhgq, ZTF20abfcszi, ZTF20abjwvae, ZTF20acitpfz and ZTF20acqoiyt.

the method proposed in Chapter 7, the objective is not to over-parametrize a simple problem. As such, we use RAINBOW with a sigmoid model for the bolometric flux, since a simple exponential rise is enough to describe rising TDEs, such that:

$$F_\nu(t,\nu;\ A,t_0,t_{rise},T) = \frac{\pi\, B(T(t),\nu)}{\sigma_{\rm SB}\, T^4} \times \frac{A}{1+\mathrm{e}^{\frac{t-t_0}{t_{rise}}}} \tag{D.1}$$

This simple model requires 4 free parameters to describe the light curves. However, before any fit, we preprocess and filter the light curves using the following criteria:

— We keep only light curves that show a rise in at least one filter.

— We exclude light curves that start decaying in at least one of their band.

— We require at least a combined total of 5 data points between both bands.

— We keep only objects that have been varying for at least a week and no more than a year

— The light curves, in both $g$ and $r$ bands, are normalized by the maximum flux value of both bands

After performing these cuts, we select for each object only the last alert that survived the filtering. To these alerts, we perform a feature extraction step based on the RAINBOW fit (see Section 3.4) using the iminuit package (Dembinski and et al., 2020). We use both the minimized parameters and their associated uncertainties [2] as features. The following additional features are also added: the $\chi^2$ quality of the fit, the standard deviation of the flux and of the signal-to-noise ratio in each passband, the normalization factor, and the number of points. Finally, an extra feature consisting of the difference of the last data point time ($t_{last}$) to the center of the sigmoid ($t_{ref}$) relative to the rise time ($\tau_{rise}$) is given by computing $sigmoid_{dist} = (t_{last} - t_{ref})/\tau_{rise}$.

This simple feature is enough to differentiate potential AGNs with pseudo rise behavior, but that will never decay. After the feature extraction, we further filter out some light curves based on the following criteria:

— The SNR (value/error) of the rise time and amplitude values from the fit must be larger than 1.5.

— The $sigmoid_{dist}$ value should be positive to avoid fits with unconstrained rise, but no larger than 8 to avoid fits with long non-decaying plateaus.

— Based on the parameter space analysis, the temperature and rise time are required to be $> 10^4$ and $< 100$ respectively. We also drop those that cluster around the upper limit of the amplitude (10), suggesting a poor fit.

— The reduced $\chi^2$ must be smaller than 100 to drop very bad fits.

The light curves surviving these cuts are then analyzed with a nearest neighbor algorithm [3]. This procedure will construct the distance (in the feature space) between all alerts, and find their nearest neighbors. The examination of the neighbors of the known TDEs will yield similar objects: TDE candidates.

As expected, many rising supernovae were detected as potential TDEs. Although it is wrong, it confirms that the pipeline is behaving as intended. Multiple object found could be compatible both with a SN or a TDE explanation, but are spectroscopically unclassified. Unfortunately,

2. Uncertainties were added in the form of a signal-to-noise ratio feature.
3. https://scikit-learn.org/stable/modules/neighbors.html

the alerts considered are past, and the sources are extinct already, leaving their true nature unknown forever. However, this is very encouraging regarding the usage of the pipeline as a science module for the Fink broker. Finally, ZTF19abuwgfg [4], a very peculiar event shown in Figure D.1, has been discovered during the data exploration.

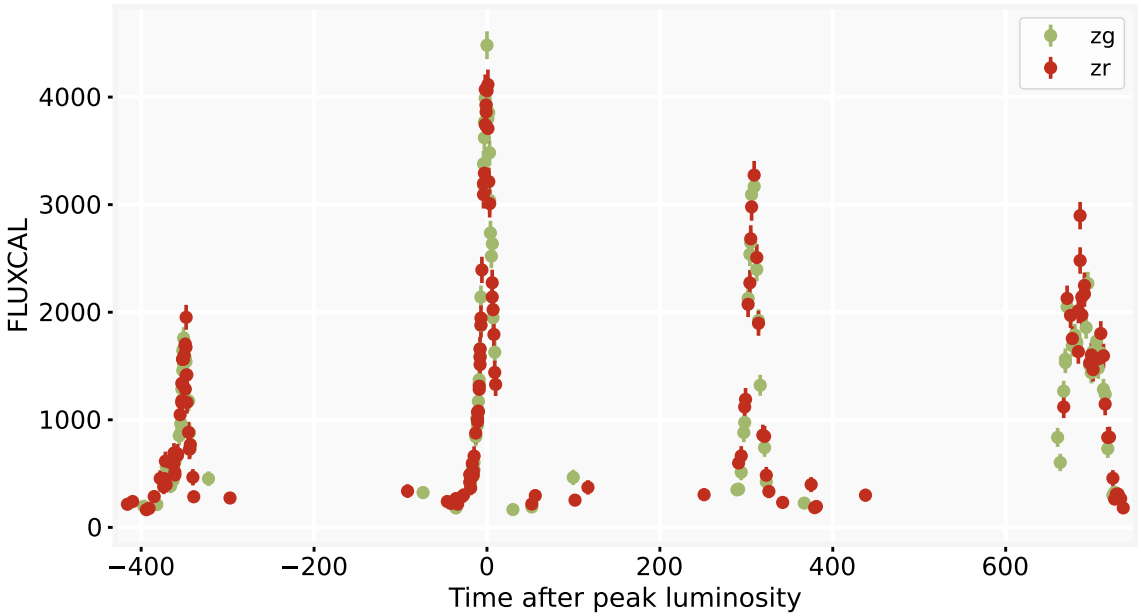



FIGURE D.1 – Light curve of ZTF19abuwgfg, discovered using the TDE-pipeline. The transient is currently unclassified, but is compatible with the very rare scenario of repeated partial tidal disruption events.

The light curve consist of repeating outbursts, each looking like a potential TDE. Assuming the emission is thermal, the temperature is $\sim$ 11 000 K, which is compatible with TDE models. The host is faint (mag $\sim$ 22), and its extra-galactic or galactic nature is unclear. If extra-galactic, the source could be a repeated partial Tidal Disruption Event, an extremely rare subtype of TDE for which only a couple of events have been confirmed Lin et al. (2024). Stars having a strongly elliptical orbit around black holes will be partially disrupted every time they approach too closely, producing this pseudo periodic bursts. However, the light curve alone is not enough to prove that the transient is a partial TDE, and further investigations should be conducted to understand the phenomenon. Independently of its exact nature, it highlights the potential of the pipeline for discoveries of rare events. Further developments of the project are ongoing, including the improvement of the cuts and features used, the follow-up of additional interesting candidates, and the integration of the module in the Fink broker.

4. https://fink-portal.org/ZTF19abuwgfg

# Bibliographie


F. Aharonian et al. The H.E.S.S. Survey of the Inner Galaxy in Very High Energy Gamma Rays. *ApJ*, 636(2):777–797, Jan. 2006. doi: 10.1086/498013.

A. Alan et al. A simplified synthesis for meso-tetraphenylporphine. *Journal of organic chemistry.*, 32(2), 1967. ISSN 0022-3263.

C. Alard and R. H. Lupton. A Method for Optimal Image Subtraction. *ApJ*, 503(1):325–331, Aug. 1998. doi: 10.1086/305984.

P. D. Aleo et al. The Young Supernova Experiment Data Release 1 (YSE DR1) Light Curves, Nov. 2022. If you have any questions about this data set, please reach out to Patrick D. Aleo at paleo2@illinois.edu.

P. D. Aleo et al. The Young Supernova Experiment Data Release 1 (YSE DR1): Light Curves and Photometric Classification of 1975 Supernovae. *ApJS*, 266(1):9, May 2023. doi: 10.3847/1538-4365/acbfba.

C. S. Alves et al. Considerations for Optimizing the Photometric Classification of Supernovae from the Rubin Observatory. *ApJS*, 258(2):23, Feb. 2022. doi: 10.3847/1538-4365/ac3479.

J. P. Anderson et al. Characterizing the V-band Light-curves of Hydrogen-rich Type II Supernovae. *ApJ*, 786(1):67, May 2014. doi: 10.1088/0004-637X/786/1/67.

Anderson, J. P. et al. Type ii supernovae as probes of environment metallicity: observations of host hii regions. *A&A*, 589:A110, 2016. doi: 10.1051/0004-6361/201527691.

R. Andrae. Error estimation in astronomy: A guide. *arXiv e-prints*, art. arXiv:1009.2755, Sept. 2010. doi: 10.48550/arXiv.1009.2755.

C. R. Angus et al. Superluminous supernovae from the Dark Energy Survey. *MNRAS*, 487(2): 2215–2241, Aug. 2019. doi: 10.1093/mnras/stz1321.

H. C. Arp. The Hertzsprung-Russell Diagram. *Handbuch der Physik*, 51:75, Jan. 1958. doi: 10.1007/978-3-642-45908-5_2.

W. B. Atwood et al. The Large Area Telescope on the Fermi Gamma-Ray Space Telescope Mission. *ApJ*, 697(2):1071–1102, June 2009. doi: 10.1088/0004-637X/697/2/1071.

W. Baade and F. Zwicky. On super-novae. *Proceedings of the National Academy of Sciences*, 20(5):254–259, 1934. doi: 10.1073/pnas.20.5.254.

L. Bachelier. Théorie de la spéculation. In *Annales scientifiques de l'École normale supérieure*, volume 17, pages 21–86, 1900.

K. Barbary et al. SNCosmo: Python library for supernova cosmology. Astrophysics Source Code Library, record ascl:1611.017, Nov. 2016.

D. J. Bartlett et al. Exhaustive Symbolic Regression. *arXiv e-prints*, art. arXiv:2211.11461, Nov. 2022. doi: 10.48550/arXiv.2211.11461.

D. J. Bartlett et al. syren-halofit: A fast, interpretable, high-precision formula for the ΛCDM nonlinear matter power spectrum. *arXiv e-prints*, art. arXiv:2402.17492, Feb. 2024. doi: 10.48550/arXiv.2402.17492.

S. W. Barwick et al. APS Neutrino Study: Report of the Neutrino Astrophysics and Cosmology Working Group. *arXiv e-prints*, art. astro-ph/0412544, Dec. 2004. doi: 10.48550/arXiv.astro-ph/0412544.

D. Bazell and D. W. Aha. Ensembles of Classifiers for Morphological Galaxy Classification. *ApJ*, 548(1):219–223, Feb. 2001. doi: 10.1086/318696.

G. Bazin et al. The core-collapse rate from the Supernova Legacy Survey. *A&A*, 499(3):653–660, June 2009. doi: 10.1051/0004-6361/200911847.

E. Bellm et al. Plans and policies for lsst alert distribution. Technical Report LDM-612, September 2020.

E. C. Bellm et al. The Zwicky Transient Facility: System Overview, Performance, and First Results. *PASP*, 131(995):018002, Jan. 2019a. doi: 10.1088/1538-3873/aaecbe.

E. C. Bellm et al. The Zwicky Transient Facility: System Overview, Performance, and First Results. *PASP*, 131(995):018002, Jan. 2019b. doi: 10.1088/1538-3873/aaecbe.

E. C. Bellm et al. The Zwicky Transient Facility: System Overview, Performance, and First Results. *Publications of the Astronomical Society of the Pacific*, 131(995):018002, Jan. 2019c. doi: 10.1088/1538-3873/aaecbe.

S. Benetti et al. Supernova 2002bo: inadequacy of the single parameter description. *MNRAS*, 348(1):261–278, Feb. 2004. doi: 10.1111/j.1365-2966.2004.07357.x.

M. S. Bessell. Standard photometric systems. *Annual Review of Astronomy and Astrophysics*, 43(Volume 43, 2005):293–336, 2005a. ISSN 1545-4282. doi: https://doi.org/10.1146/annurev.astro.41.082801.100251.

M. S. Bessell. Standard photometric systems. *Annual Review of Astronomy and Astrophysics*, 43(Volume 43, 2005):293–336, 2005b. ISSN 1545-4282. doi: https://doi.org/10.1146/annurev.astro.41.082801.100251.

M. Betoule et al. Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples. *A&A*, 568:A22, Aug. 2014. doi: 10.1051/0004-6361/201423413.

G. Biau. Analysis of a Random Forests Model. *arXiv e-prints*, art. arXiv:1005.0208, May 2010. doi: 10.48550/arXiv.1005.0208.

B. Biswas et al. Enabling the discovery of fast transients. A kilonova science module for the Fink broker. *A&A*, 677:A77, Sept. 2023. doi: 10.1051/0004-6361/202245340.

F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, May 1973. ISSN 1537-534X. doi: 10.1086/260062.

N. Blagorodnova. Blagorodnova, N. Transient Classification Report for 2017-10-11. *Transient Name Server Classification Report*, 2017-1102:1, Oct. 2017.

S. Blondin et al. A one-dimensional Chandrasekhar-mass delayed-detonation model for the broad-lined Type Ia supernova 2002bo. *MNRAS*, 448(3):2766–2797, Apr. 2015. doi: 10.1093/mnras/stv188.

J. S. Bloom et al. Automating Discovery and Classification of Transients and Variable Stars in the Synoptic Survey Era. *PASP*, 124(921):1175, Nov. 2012. doi: 10.1086/668468.

F. Bonnarel et al. The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources. *A&AS*, 143:33–40, Apr. 2000. doi: 10.1051/aas:2000331.

K. Boone. Avocado: Photometric Classification of Astronomical Transients with Gaussian Process Augmentation. *AJ*, 158(6):257, Dec. 2019. doi: 10.3847/1538-3881/ab5182.

J.-P. Bouchaud and M. Potters. *Theory of financial risks*, volume 12. Cambridge University Press, Cambridge From Statistical Physics to Risk . . ., 2000.

W. S. Boyle and G. E. Smith. Charge coupled semiconductor devices. *The Bell System Technical Journal*, 49(4):587–593, 1970. doi: 10.1002/j.1538-7305.1970.tb01790.x.

A. E. Bozdoğan. Polynomial equations based on bouguer–lambert and beer laws for deviations from linearity and absorption flattening. *Journal of Analytical Chemistry*, 77(11):1426–1432, Nov 2022. ISSN 1608-3199. doi: 10.1134/S1061934822110028.

D. M. Bramich. A new algorithm for difference image analysis. *MNRAS*, 386(1):L77–L81, May 2008. doi: 10.1111/j.1745-3933.2008.00464.x.

L. Breiman. Random forests. 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

L. Breiman et al. *Classification and Regression Trees*. Chapman and Hall/CRC. ISBN 978-1-315-13947-0. doi: 10.1201/9781315139470.

L. Breiman et al. Random Forests: finding quasars. In E. D. Feigelson and G. J. Babu, editors, *Statistical Challenges in Astronomy*, pages 243–254. 2003.

C. Bullivant et al. SN 2013fs and SN 2013fr: exploring the circumstellar-material diversity in Type II supernovae. *Monthly Notices of the Royal Astronomical Society*, 476(2):1497–1518, 01 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty045.

B. Burlacu et al. Operon c++: An efficient genetic programming framework for symbolic regression. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*, GECCO '20, page 1562–1570, internet, July 8-12 2020. Association for Computing Machinery. doi: doi:10.1145/3377929.3398099.

S. Burns. *A Practical Guide to Observational Astronomy*. CRC Press, 2018.

S. A. Butchins. Predicted redshifts of galaxies by broadband photometry. *A&A*, 97(2):407–409, Apr. 1981.

G. Cabrera-Vives et al. ATAT: Astronomical Transformer for time series And Tabular data. *arXiv e-prints*, art. arXiv:2405.03078, May 2024a. doi: 10.48550/arXiv.2405.03078.

G. Cabrera-Vives et al. ATAT: Astronomical Transformer for time series And Tabular data. *arXiv e-prints*, art. arXiv:2405.03078, May 2024b. doi: 10.48550/arXiv.2405.03078.

K. C. Chambers et al. The Pan-STARRS1 Surveys. *arXiv e-prints*, art. arXiv:1612.05560, Dec. 2016. doi: 10.48550/arXiv.1612.05560.

K. Chanchaiworawit and V. Sarajedini. Ensemble Variability Properties of Active Galactic Nuclei in the SDSS DR17. *ApJ*, 969(2):131, July 2024. doi: 10.3847/1538-4357/ad479a.

S. Chandrasekhar. The Maximum Mass of Ideal White Dwarfs. *ApJ*, 74:81, July 1931. doi: 10.1086/143324.

M. Chu et al. ZTF Transient Classification Report for 2021-07-22. *Transient Name Server Classification Report*, 2021-2534:1–2534, July 2021.

A. Corsi et al. A Search for Relativistic Ejecta in a Sample of ZTF Broad-lined Type Ic Supernovae. *ApJ*, 953(2):179, Aug. 2023. doi: 10.3847/1538-4357/acd3f2.

E. R. Coughlin and C. J. Nixon. Partial stellar disruption by a supermassive black hole: Is the light curve really proportional to t-9/4. *The Astrophysical Journal Letters*, 883(1):L17, sep 2019. doi: 10.3847/2041-8213/ab412d.

D. A. Coulter et al. YSE-PZ: An Open-source Target and Observation Management System, Nov. 2022. D. A. Coulter acknowledges support from the National Science Foundation Graduate Research Fellowship under Grant DGE1339067.

M. Cranmer. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl. *arXiv e-prints*, art. arXiv:2305.01582, May 2023. doi: 10.48550/arXiv.2305.01582.

M. Cranmer et al. Discovering symbolic models from deep learning with inductive biases. *Advances in Neural Information Processing Systems*, 33:17429–17442, 2020.

C. Curtin et al. First Release of High-redshift Superluminous Supernovae from the Subaru HIgh-Z SUpernova CAmpaign (SHIZUCA). II. Spectroscopic Properties. *ApJS*, 241(2):17, Apr. 2019. doi: 10.3847/1538-4365/ab07c8.

M. Dai et al. Photometric classification and redshift estimation of LSST Supernovae. *MNRAS*, 477(3):4142–4151, July 2018. doi: 10.1093/mnras/sty965.

C. Darwin. Origin of the species. In *British Politics and the Environment in the Long Nineteenth Century*. Routledge, 1859. ISBN 978-1-00-319465-1. Num Pages: 9.

S. Das et al. Incorporating Feedback into Tree-based Anomaly Detection. *arXiv e-prints*, art. arXiv:1708.09441, 2017. doi: 10.48550/arXiv.1708.09441.

J. A. de Diego et al. Galaxy classification: deep learning on the OTELO and COSMOS databases. *A&A*, 638:A134, June 2020. doi: 10.1051/0004-6361/202037697.

F. O. de França et al. Understanding conflict origin and dynamics on twitter: A real-time detection system. *Expert Systems with Applications*, 212:118748, 2023.

F. M. F. de Oliveira et al. Data-driven photometric redshift estimation from type Ia supernovae light curves. *arXiv e-prints*, art. arXiv:2212.14668, Dec. 2022. doi: 10.48550/arXiv.2212.14668.

K. M. de Soto et al. Superphot+: Realtime Fitting and Classification of Supernova Light Curves. *arXiv e-prints*, art. arXiv:2403.07975, Mar. 2024. doi: 10.48550/arXiv.2403.07975.

S. Deb and H. P. Singh. Light curve analysis of variable stars using Fourier decomposition and principal component analysis. *A&A*, 507(3):1729–1737, Dec. 2009. doi: 10.1051/0004-6361/200912851.

R. Dekany et al. The Zwicky Transient Facility: Observing System. *PASP*, 132(1009):038001, Mar. 2020. doi: 10.1088/1538-3873/ab4ca2.

H. Dembinski and P. O. et al. scikit-hep/iminuit. Dec 2020. doi: 10.5281/zenodo.3949207.

M. Demianenko et al. Understanding of the properties of neural network approaches for transient light curve approximations. *A&A*, 677:A16, Sept. 2023. doi: 10.1051/0004-6361/202245189.

M. Demianenko et al. Supernova light curves approximation based on neural network models. *Journal of Physics: Conference Series*, 2438(1):012128, feb 2023. doi: 10.1088/1742-6596/2438/1/012128.

D. D. Desai et al. Supernova rates and luminosity functions from ASAS-SN I: 2014-2017 Type Ia SNe and their subtypes. *MNRAS*, 530(4):5016–5029, June 2024. doi: 10.1093/mnras/stae606.

L. Dessart et al. Superluminous supernovae: 56Ni power versus magnetar radiation. *Monthly Notices of the Royal Astronomical Society: Letters*, 426(1):L76–L80, 10 2012. ISSN 1745-3925. doi: 10.1111/j.1745-3933.2012.01329.x.

Dessart, Luc and Audit, Edouard. Super-luminous type ii supernovae powered by magnetars. *A&A*, 613:A5, 2018. doi: 10.1051/0004-6361/201732229.

T. G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7):1895–1923, 10 1998. ISSN 0899-7667. doi: 10.1162/089976698300017197.

H. Domínguez Sánchez et al. Improving galaxy morphologies for SDSS with Deep Learning. *MNRAS*, 476(3):3661–3676, Feb. 2018. doi: 10.1093/mnras/sty338.

A. J. Drake et al. First Results from the Catalina Real-Time Transient Survey. *ApJ*, 696(1):870–884, May 2009. doi: 10.1088/0004-637X/696/1/870.

A. J. Drake et al. Discovery of the Extremely Energetic Supernova 2008fz. *ApJ*, 718(2):L127–L131, Aug. 2010. doi: 10.1088/2041-8205/718/2/L127.

A. J. Drake et al. The Discovery and Nature of the Optical Transient CSS100217:102913+404220. *ApJ*, 735(2):106, July 2011. doi: 10.1088/0004-637X/735/2/106.

A. Eckart and R. Genzel. Stellar proper motions in the central 0.1 PC of the Galaxy. *MNRAS*, 284(3):576–598, Jan. 1997. doi: 10.1093/mnras/284.3.576.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979. doi: 10.1214/aos/1176344552.

M. El Habib Daho et al. Weighted vote for trees aggregation in random forest. volume 0, 04 2014. doi: 10.1109/ICMCS.2014.6911187.

T. Faran et al. The evolution of temperature and bolometric luminosity in Type II supernovae. *MNRAS*, 473(1):513–537, Jan. 2018. doi: 10.1093/mnras/stx2288.

A. Fassia et al. Optical and infrared photometry of the Type IIn SN 1998S: days 11–146. *Monthly Notices of the Royal Astronomical Society*, 318(4):1093–1104, 11 2000. ISSN 0035-8711. doi: 10.1046/j.1365-8711.2000.03797.x.

R. P. Feynman et al. *The Feynman lectures on physics ; New millennium ed.* Basic Books, New York, NY, 2010. Originally published 1963-1965.

A. V. Filippenko. Optical Spectra of Supernovae. *ARA&A*, 35:309–355, Jan. 1997. doi: 10.1146/annurev.astro.35.1.309.

A. V. Filippenko et al. The Subluminous, Spectroscopically Peculiar Type 1a Supernova 1991bg in the Elliptical Galaxy NGC 4374. *AJ*, 104:1543, Oct. 1992a. doi: 10.1086/116339.

A. V. Filippenko et al. The Peculiar Type IA SN 1991T: Detonation of a White Dwarf? *ApJ*, 384:L15, Jan. 1992b. doi: 10.1086/186252.

A. V. Filippenko et al. The "Type IIb" Supernova 1993J in M81: A Close Relative of Type Ib Supernovae. *ApJ*, 415:L103, Oct. 1993. doi: 10.1086/187043.

R. J. Foley et al. Type Iax Supernovae: A New Class of Stellar Explosion. *ApJ*, 767(1):57, Apr. 2013. doi: 10.1088/0004-637X/767/1/57.

B. M. O. Fraga et al. Transient Classifiers for Fink: Benchmarks for LSST. *arXiv e-prints*, art. arXiv:2404.08798, Apr. 2024. doi: 10.48550/arXiv.2404.08798.

M. Fraser. Supernovae and transients with circumstellar interaction. *Royal Society Open Science*, 7:200467, 07 2020. doi: 10.1098/rsos.200467.

W. L. Freedman et al. Final Results from the Hubble Space Telescope Key Project to Measure the Hubble Constant. *ApJ*, 553(1):47–72, May 2001. doi: 10.1086/320638.

W. L. Freedman et al. Status Report on the Chicago-Carnegie Hubble Program (CCHP): Three Independent Astrophysical Determinations of the Hubble Constant Using the James Webb Space Telescope. *arXiv e-prints*, art. arXiv:2408.06153, Aug. 2024. doi: 10.48550/arXiv.2408.06153.

J. H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1 – 67, 1991. doi: 10.1214/aos/1176347963.

M. D. Fulton et al. The Optical Light Curve of GRB 221009A: The Afterglow and the Emerging Supernova. *ApJ*, 946(1):L22, Mar. 2023. doi: 10.3847/2041-8213/acc101.

T. Gabruseva et al. Photometric Light Curves Classification with Machine Learning. *Journal of Astronomical Instrumentation*, 9(1):2050005-3986, Jan. 2020. doi: 10.1142/S2251171720500051.

J. P. Gardner et al. The James Webb Space Telescope. *Space Sci. Rev.*, 123(4):485–606, Apr. 2006. doi: 10.1007/s11214-006-8315-7.

D. W. Gerdes et al. ArborZ: Photometric Redshifts Using Boosted Decision Trees. *ApJ*, 715 (2):823–832, June 2010. doi: 10.1088/0004-637X/715/2/823.

S. Gezari. Tidal Disruption Events. *ARA&A*, 59:21–58, Sept. 2021. doi: 10.1146/annurev-astro-111720-030029.

D. Godines et al. A machine learning classifier for microlensing in wide-field surveys. *Astronomy and Computing*, 28:100298, 2019. ISSN 2213-1337. doi: https://doi.org/10.1016/j.ascom.2019.100298.

M. J. Graham et al. Machine-assisted discovery of relationships in astronomy. *Monthly Notices of the Royal Astronomical Society*, 431(3):2371–2384, 03 2013. ISSN 0035-8711. doi: 10.1093/mnras/stt329.

J. Guy et al. SALT2: using distant supernovae to improve the use of type Ia supernovae as distance indicators. *A&A*, 466(1):11–21, Apr. 2007. doi: 10.1051/0004-6361:20066930.

E. Hammerstein et al. The final season reimagined: 30 tidal disruption events from the ztf-i survey. *The Astrophysical Journal*, 942(1):9, dec 2022. doi: 10.3847/1538-4357/aca283.

M. Hamuy and P. A. Pinto. Type ii supernovae as standardized candles. *The Astrophysical Journal*, 566(2):L63, jan 2002. doi: 10.1086/339676.

J. N. Heasley. Point-Spread Function Fitting Photometry. In E. R. Craine et al., editors, *Precision CCD Photometry*, volume 189 of *Astronomical Society of the Pacific Conference Series*, page 56, Jan. 1999.

A. Heger and S. E. Woosley. The Nucleosynthetic Signature of Population III. *ApJ*, 567(1):532–543, Mar. 2002. doi: 10.1086/338487.

A. Heger et al. How Massive Single Stars End Their Life. *ApJ*, 591(1):288–300, July 2003. doi: 10.1086/375341.

A. Hernandez et al. Fast, accurate, and transferable many-body interatomic potentials by symbolic regression. *npj Computational Materials*, 5(1):112, Nov 2019. ISSN 2057-3960. doi: 10.1038/s41524-019-0249-1.

J. Heyvaerts et al. An emerging flux model for the solar phenomenon. *ApJ*, 216:123–137, Aug. 1977. doi: 10.1086/155453.

R. Hložek et al. Results of the Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC). *arXiv e-prints*, art. arXiv:2012.12392, Dec. 2020. doi: 10.48550/arXiv.2012.12392.

P. Hoflich et al. SN 1993J : explosion of a massive cool supergiant with a small envelope nass ? *A&A*, 275:L29–L32, Aug. 1993.

D. W. Hogg. Distance measures in cosmology. *arXiv e-prints*, art. astro-ph/9905116, May 1999. doi: 10.48550/arXiv.astro-ph/9905116.

G. Hosseinzadeh et al. Photometric classification of 2315 pan-starrs1 supernovae with superphot. *The Astrophysical Journal*, 905(2):93, dec 2020. doi: 10.3847/1538-4357/abc42b.

E. P. Hubble. Extragalactic nebulae. *ApJ*, 64:321–369, Dec. 1926. doi: 10.1086/143018.

W. Huggins. Further Observations on the Spectra of Some of the Stars and Nebulae, with an Attempt to Determine Therefrom Whether These Bodies are Moving towards or from the Earth, Also Observations on the Spectra of the Sun and of Comet II., 1868. *Philosophical Transactions of the Royal Society of London Series I*, 158:529–564, Jan. 1868.

T. Hung et al. Sifting for Sapphires: Systematic Selection of Tidal Disruption Events in iPTF. *ApJS*, 238(2):15, Oct. 2018. doi: 10.3847/1538-4365/aad8b1.

J. Iben, I. and A. V. Tutukov. Supernovae of type I as end products of the evolution of binaries with components of moderate initial mass. *ApJS*, 54:335–372, Feb. 1984. doi: 10.1086/190932.

O. Ilbert et al. Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey. *A&A*, 457(3):841–856, Oct. 2006. doi: 10.1051/0004-6361:20065138.

C. Inserra and S. J. Smartt. Superluminous Supernovae as Standardizable Candles and High-redshift Distance Probes. *ApJ*, 796(2):87, Dec. 2014. doi: 10.1088/0004-637X/796/2/87.

E. E. O. Ishida. Machine learning and the future of supernova cosmology. *Nature Astronomy*, 3:680–682, Aug. 2019. doi: 10.1038/s41550-019-0860-6.

E. E. O. Ishida and R. S. de Souza. Kernel PCA for Type Ia supernovae photometric classification. *MNRAS*, 430(1):509–532, Mar. 2013. doi: 10.1093/mnras/sts650.

E. E. O. Ishida et al. Optimizing spectroscopic follow-up strategies for supernova photometric classification with active learning. *MNRAS*, 483(1):2–18, Feb. 2019. doi: 10.1093/mnras/sty3015.

E. E. O. Ishida et al. Active anomaly detection for time-domain discoveries. *A&A*, 650:A195, June 2021. doi: 10.1051/0004-6361/202037709.

Ž. Ivezić et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *ApJ*, 873(2):111, Mar. 2019. doi: 10.3847/1538-4357/ab042c.

F. James. Function minimization, 1972.

F. James and M. Roos. Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations. *Comput. Phys. Commun.*, 10:343–367, 1975. doi: 10.1016/0010-4655(75)90039-9.

H.-T. Janka. Explosion mechanisms of core-collapse supernovae. *Annual Review of Nuclear and Particle Science*, 62(Volume 62, 2012):407–451, 2012. ISSN 1545-4134. doi: https://doi.org/10.1146/annurev-nucl-102711-094901.

S. W. Jha et al. Type Iax Supernovae. In *American Astronomical Society Meeting Abstracts #229*, volume 229 of *American Astronomical Society Meeting Abstracts*, page 410.01, Jan. 2017.

N. V. Karpenka et al. A simple and robust method for automated photometric classification of supernovae using neural networks. *Monthly Notices of the Royal Astronomical Society*, 429 (2):1278–1285, 12 2012. ISSN 0035-8711. doi: 10.1093/mnras/sts412.

S. Karpov and J. Peloton. Impact of satellite glints on the transient science on ZTF scale. *arXiv e-prints*, art. arXiv:2202.05719, Feb. 2022. doi: 10.48550/arXiv.2202.05719.

D. Kasen. Secondary Maximum in the Near-Infrared Light Curves of Type Ia Supernovae. *ApJ*, 649(2):939–953, Oct. 2006. doi: 10.1086/506588.

P. L. Kelly et al. The Magnificent Five Images of Supernova Refsdal: Time Delay and Magnification Measurements. *ApJ*, 948(2):93, May 2023. doi: 10.3847/1538-4357/ac4ccb.

W. D. Kenworthy et al. SALT3: An Improved Type Ia Supernova Model for Measuring Cosmic Distances. *ApJ*, 923(2):265, Dec. 2021. doi: 10.3847/1538-4357/ac30d8.

R. Kessler et al. SNANA: A Public Software Package for Supernova Analysis. *PASP*, 121(883): 1028, Sept. 2009. doi: 10.1086/605984.

R. Kessler et al. Models and Simulations for the Photometric LSST Astronomical Time Series Classification Challenge (PLAsTiCC). *PASP*, 131(1003):094501, Sept. 2019. doi: 10.1088/1538-3873/ab26f1.

S. Khakpash et al. Multi-filter UV to NIR Data-driven Light Curve Templates for Stripped Envelope Supernovae. *arXiv e-prints*, art. arXiv:2405.01672, May 2024. doi: 10.48550/arXiv.2405.01672.

M. Kiewe et al. Caltech Core-Collapse Project (CCCP) Observations of Type IIn Supernovae: Typical Properties and Implications for Their Progenitor Stars. *ApJ*, 744(1):10, Jan. 2012. doi: 10.1088/0004-637X/744/1/10.

C. S. Kochanek et al. The All-Sky Automated Survey for Supernovae (ASAS-SN) Light Curve Server v1.0. *PASP*, 129(980):104502, Oct. 2017. doi: 10.1088/1538-3873/aa80d9.

J. Koornneef et al. Synthetic photometry and the calibration of the Hubble Space Telescope. *Highlights of Astronomy*, 7:833–843, Jan. 1986.

M. Kopsacheili et al. A diagnostic tool for the identification of supernova remnants. *MNRAS*, 491(1):889–902, Jan. 2020. doi: 10.1093/mnras/stz2594.

M. V. Kornilov et al. Reduction of supernova light curves by vector Gaussian processes. *MNRAS*, Sept. 2023. doi: 10.1093/mnras/stad2645.

S. Kou et al. A new method to classify type iip/iil supernovae based on their spectra. *The Astrophysical Journal*, 890(2):177, feb 2020. doi: 10.3847/1538-4357/ab6601.

S. G. Kou. A jump-diffusion model for option pricing. *Management Science*, 48(8):1086–1101, Aug. 2002. ISSN 1526-5501. doi: 10.1287/mnsc.48.8.1086.166.

J. R. Koza. Genetic programming as a means for programming computers by natural selection. 4(2):87–112. ISSN 1573-1375. doi: 10.1007/BF00175355.

A. Kozyreva et al. Fast evolving pair-instability supernova models: evolution, explosion, light curves. *MNRAS*, 464(3):2854–2865, Jan. 2017. doi: 10.1093/mnras/stw2562.

O. Krause et al. Tycho Brahe's 1572 supernova as a standard typeIa as revealed by its light-echo spectrum. *Nature*, 456(7222):617–619, Dec. 2008. doi: 10.1038/nature07608.

A. Krone-Martins et al. The first analytical expression to estimate photometric redshifts suggested by a machine. *Monthly Notices of the Royal Astronomical Society: Letters*, 443(1): L34–L38, 06 2014. ISSN 1745-3925. doi: 10.1093/mnrasl/slu067.

A. Krone-Martins et al. The first analytical expression to estimate photometric redshifts suggested by a machine. *MNRAS*, 443:L34–L38, Sept. 2014. doi: 10.1093/mnrasl/slu067.

W. La Cava et al. Contemporary symbolic regression methods and their relative performance. 07 2021.

W. G. La Cava et al. A flexible symbolic regression method for constructing interpretable clinical prediction models. *npj Digital Medicine*, 6(1):1–14, June 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00833-8.

J. H. Lacy et al. The nature of the central parsec of the Galaxy. *ApJ*, 262:120–134, Nov. 1982. doi: 10.1086/160402.

R. R. Laher et al. Processing Images from the Zwicky Transient Facility. *Robotic Telescope, Student Research and Education Proceedings*, 1(1):329–336, Oct. 2018. doi: 10.48550/arXiv.1708.01584.

N. M. Law et al. The Palomar Transient Factory: System Overview, Performance, and First Results. *PASP*, 121(886):1395, Dec. 2009. doi: 10.1086/648598.

Le Montagner, R. et al. Enabling discoveries of solar system objects in large alert data streams. *A&A*, 680:A17, 2023. doi: 10.1051/0004-6361/202346905.

H. S. Leavitt and E. C. Pickering. Periods of 25 Variable Stars in the Small Magellanic Cloud. *Harvard College Observatory Circular*, 173:1–3, Mar. 1912.

Léget, P.-F. et al. Sugar: An improved empirical model of type ia supernovae based on spectral features. *A&A*, 636:A46, 2020. doi: 10.1051/0004-6361/201834954.

M. Leoni et al. Fink: Early supernovae Ia classification using active learning. *A&A*, 663:A13, July 2022. doi: 10.1051/0004-6361/202142715.

Leoni, M. et al. Fink: Early supernovae ia classification using active learning. *A&A*, 663:A13, 2022. doi: 10.1051/0004-6361/202142715.

K. Levenberg. A method for the solution of certain non-linear problems in least squares. 2(2): 164–168, 1944. ISSN 0033-569X, 1552-4485. doi: 10.1090/qam/10666.

R. Liang et al. Kilonova-Targeting Lightcurve Classification for Wide Field Survey Telescope. *Universe*, 10(1):10, Dec. 2023. doi: 10.3390/universe10010010.

Z. Lin et al. The unluckiest star: A spectroscopically confirmed repeated partial tidal disruption event AT 2022dbl. *arXiv e-prints*, art. arXiv:2405.10895, May 2024. doi: 10.48550/arXiv.2405.10895.

F. T. Liu et al. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. doi: 10.1109/ICDM.2008.17.

T. Liu et al. An intermediate distribution between gaussian and cauchy distributions. *Physica A: Statistical Mechanics and its Applications*, 391(22):5411–5421, Nov. 2012. ISSN 0378-4371. doi: 10.1016/j.physa.2012.06.035.

M. Lochner et al. Photometric Supernova Classification with Machine Learning. *ApJS*, 225(2): 31, Aug. 2016. doi: 10.3847/0067-0049/225/2/31.

M. Lochner et al. Optimizing the LSST Observing Strategy for Dark Energy Science: DESC Recommendations for the Wide-Fast-Deep Survey. *arXiv e-prints*, art. arXiv:1812.00515, Nov. 2018. doi: 10.48550/arXiv.1812.00515.

LSST. Rubin observatory system & lsst survey key numbers, 2024.

D. B. Madan et al. The variance gamma process and option pricing. *Review of Finance*, 2(1): 79–105, Apr. 1998. ISSN 1573-692X. doi: 10.1023/a:1009703431535.

A. Maeder. The Conti scenario for forming WR stars: past, present and future. In J. M. Vreux et al., editors, *Liege International Astrophysical Colloquia*, volume 33 of *Liege International Astrophysical Colloquia*, page 39, Jan. 1996.

A. Mahabal et al. Towards Real-time Classification of Astronomical Transients. In C. A. L. Bailer-Jones, editor, *Classification and Discovery in Large Astronomical Surveys*, volume 1082 of *American Institute of Physics Conference Series*, pages 287–293. AIP, Dec. 2008. doi: 10.1063/1.3059064.

K. Malanchev et al. The SNAD Viewer: Everything You Want to Know about Your Favorite ZTF Object. *Publications of the Astronomical Society of the Pacific*, 135(1044):024503, Feb. 2023. doi: 10.1088/1538-3873/acb292.

K. L. Malanchev et al. Anomaly detection in the Zwicky Transient Facility DR3. *Monthly Notices of the Royal Astronomical Society*, 502(4):5147–5175, 02 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab316.

K. L. Malanchev et al. Anomaly detection in the Zwicky Transient Facility DR3. *MNRAS*, 502 (4):5147–5175, Apr. 2021. doi: 10.1093/mnras/stab316.

R. N. Mantegna and H. E. Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 376(6535):46–49, July 1995. ISSN 1476-4687. doi: 10.1038/376046a0.

R. N. Mantegna and H. E. Stanley. *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press, 1999.

M. L. Martínez-Aldama et al. Can Reverberation-measured Quasars Be Used for Cosmology ? *The Astrophysical Journal*, 883(2):170, Oct. 2019. doi: 10.3847/1538-4357/ab3728.

J. Martinez Galarza et al. X-ray transient discoveries with machine learning. In *AAS/High Energy Astrophysics Division*, volume 21 of *AAS/High Energy Astrophysics Division*, page 108.11, May 2024.

V. A. Masoura et al. Disentangling the AGN and star formation connection using XMM-Newton. *Astronomy and Astrophysics*, 618:A31, Oct. 2018. doi: 10.1051/0004-6361/201833397.

K. T. Matchev et al. Analytical Modeling of Exoplanet Transit Spectroscopy with Dimensional Analysis and Symbolic Regression. *ApJ*, 930(1):33, May 2022. doi: 10.3847/1538-4357/ac610c.

T. Matheson et al. The ANTARES Astronomical Time-domain Event Broker. *AJ*, 161(3):107, Mar. 2021. doi: 10.3847/1538-3881/abd703.

P. A. Mazzali et al. A Common Explosion Mechanism for Type Ia Supernovae. *Science*, 315 (5813):825, Feb. 2007. doi: 10.1126/science.1136259.

C. McCully et al. Still Brighter than Pre-explosion, SN 2012Z Did Not Disappear: Comparing Hubble Space Telescope Observations a Decade Apart. *ApJ*, 925(2):138, Feb. 2022. doi: 10.3847/1538-4357/ac3bbd.

K. E. McGowan et al. Searching for Short-term Variations in Miras using Machine Learning. In *American Astronomical Society Meeting Abstracts*, volume 203 of *American Astronomical Society Meeting Abstracts*, page 08.02, Dec. 2003.

Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, Jun 1947. ISSN 1860-0980. doi: 10.1007/BF02295996.

A. Merghadi et al. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth Science Reviews*, 207:103225, Aug. 2020. doi: 10.1016/j.earscirev.2020.103225.

A. Meurer et al. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, Jan. 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.103.

R. Miles. A light history of photometry: from Hipparchus to the Hubble Space Telescope. *Journal of the British Astronomical Association*, 117:172–186, Aug. 2007.

R. Minkowski. Spectra of Supernovae. *PASP*, 53(314):224, Aug. 1941. doi: 10.1086/125315.

A. Möller and T. de Boissière. SuperNNova: an open-source framework for Bayesian, neural network-based supernova classification. *MNRAS*, 491(3):4277–4293, Jan. 2020a. doi: 10.1093/mnras/stz3312.

A. Möller and T. de Boissière. SuperNNova: an open-source framework for Bayesian, neural network-based supernova classification. *MNRAS*, 491(3):4277–4293, Jan. 2020b. doi: 10.1093/mnras/stz3312.

A. Möller et al. FINK, a new generation of broker for the LSST community. *MNRAS*, 501(3): 3272–3288, Mar. 2021a. doi: 10.1093/mnras/staa3602.

A. Möller et al. FINK, a new generation of broker for the LSST community. *MNRAS*, 501(3): 3272–3288, Mar. 2021b. doi: 10.1093/mnras/staa3602.

H. W. Moos et al. Overview of the Far Ultraviolet Spectroscopic Explorer Mission. *ApJ*, 538 (1):L1–L6, July 2000. doi: 10.1086/312795.

A. Morales-Garoffolo et al. SN 2011fu: a type IIb supernova with a luminous double-peaked light curve. *MNRAS*, 454(1):95–114, Nov. 2015. doi: 10.1093/mnras/stv1972.

T. J. Moriya et al. Superluminous Supernovae. *Space Sci. Rev.*, 214(2):59, Mar. 2018. doi: 10.1007/s11214-018-0493-6.

D. Muthukrishna. Real-time detection of anomalies in large-scale transient surveys. In *American Astronomical Society Meeting Abstracts*, volume 54 of *American Astronomical Society Meeting Abstracts*, page 215.06, June 2022.

D. Muthukrishna et al. RAPID: Early Classification of Explosive Transients Using Deep Learning. *PASP*, 131(1005):118002, Nov. 2019. doi: 10.1088/1538-3873/ab1609.

G. Narayan and ELAsTiCC Team. The Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC). In *American Astronomical Society Meeting Abstracts*, volume 241 of *American Astronomical Society Meeting Abstracts*, page 117.01, Jan. 2023.

J. Newling et al. Statistical classification techniques for photometric supernova typing. *Monthly Notices of the Royal Astronomical Society*, 414(3):1987–2004, 06 2011. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2011.18514.x.

M. Newsome et al. Probing the Subparsec Dust of a Supermassive Black Hole with the Tidal Disruption Event AT 2020mot. *ApJ*, 961(2):239, Feb. 2024. doi: 10.3847/1538-4357/ad036e.

M. Nicholl et al. On the diversity of superluminous supernovae: ejected mass as the dominant factor. *MNRAS*, 452(4):3869–3893, Oct. 2015. doi: 10.1093/mnras/stv1522.

K. Nomoto et al. Accreting white dwarf models for type I supernovae. III. Carbon deflagration supernovae. *ApJ*, 286:644–658, Nov. 1984. doi: 10.1086/162639.

J. Nordin et al. Transient processing and analysis using AMPEL: alert management, photometry, and evaluation of light curves. *A&A*, 631:A147, Nov. 2019. doi: 10.1051/0004-6361/201935634.

M. T. Patterson et al. The zwicky transient facility alert distribution system. *Publications of the Astronomical Society of the Pacific*, 131(995):018001, nov 2018. doi: 10.1088/1538-3873/aae904.

F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

M. Perez-Carrasco et al. Alert Classification for the ALeRCE Broker System: The Anomaly Detector. *AJ*, 166(4):151, Oct. 2023. doi: 10.3847/1538-3881/ace0c1.

S. Perlmutter et al. Measurements of $\Omega$ and $\Lambda$ from 42 High-Redshift Supernovae. *ApJ*, 517(2):565–586, June 1999. doi: 10.1086/307221.

P. J. Pessi et al. ELEPHANT: ExtragaLactic alErt Pipeline for Hostless AstroNomical Transients. *arXiv e-prints*, art. arXiv:2404.18165, Apr. 2024. doi: 10.48550/arXiv.2404.18165.

J. D. R. Pierel et al. Extending supernova spectral templates for next-generation space telescope observations. *Publications of the Astronomical Society of the Pacific*, 130(993):114504, oct 2018. doi: 10.1088/1538-3873/aadb7a.

Planck Collaboration et al. Planck 2018 results. VI. Cosmological parameters. *A&A*, 641:A6, Sept. 2020a. doi: 10.1051/0004-6361/201833910.

Planck Collaboration et al. Planck 2018 results. VI. Cosmological parameters. *A&A*, 641:A6, Sept. 2020b. doi: 10.1051/0004-6361/201833910.

P. Podsiadlowski et al. The progenitor of supernova 1993J: a stripped supergiant in a binary system ? *Nature*, 364(6437):509–511, Aug. 1993. doi: 10.1038/364509a0.

A. Y. Poludnenko et al. A unified mechanism for unconfined deflagration-to-detonation transition in terrestrial chemical systems and type Ia supernovae. *Science*, 366(6465):aau7365, Nov. 2019. doi: 10.1126/science.aau7365.

M. Pruzhinskaya et al. Could snad160 be a pair-instability supernova ? *Research Notes of the AAS*, 6(6):122, jun 2022. doi: 10.3847/2515-5172/ac76cf.

Pruzhinskaya, M. V. et al. Supernova search with active learning in ztf dr3. *A&A*, 672:A111, 2023. doi: 10.1051/0004-6361/202245172.

I. P. Pskovskii. Light curves, color curves, and expansion velocity of type I supernovae as functions of the rate of brightness decline. *Soviet Ast.*, 21:675, Dec. 1977.

G. Rakavy and G. Shaviv. Instabilities in Highly Evolved Stellar Models. *ApJ*, 148:803, June 1967. doi: 10.1086/149204.

S. Raschka. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv e-prints*, art. arXiv:1811.12808, Nov. 2018. doi: 10.48550/arXiv.1811.12808.

A. Rebassa-Mansergas et al. Where are the double-degenerate progenitors of Type Ia supernovae ? *MNRAS*, 482(3):3656–3668, Jan. 2019. doi: 10.1093/mnras/sty2965.

A. Rest et al. Pushing the Boundaries of Conventional Core-collapse Supernovae: The Extremely Energetic Supernova SN 2003ma. *ApJ*, 729(2):88, Mar. 2011. doi: 10.1088/0004-637X/729/2/88.

D. Richardson et al. Absolute-magnitude distributions of supernovae. *The Astronomical Journal*, 147(5):118, apr 2014. doi: 10.1088/0004-6256/147/5/118.

A. G. Riess et al. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *AJ*, 116(3):1009–1038, Sept. 1998. doi: 10.1086/300499.

M. Rigault et al. ZTF SN Ia DR2: Study of Type Ia Supernova lightcurve fits. *arXiv e-prints*, art. arXiv:2406.02073, June 2024. doi: 10.48550/arXiv.2406.02073.

C. Rodrigo et al. SVO Filter Profile Service Version 1.0. IVOA Working Draft 15 October 2012, Oct. 2012.

D. E. Rumelhart and J. L. McClelland. *Learning Internal Representations by Error Propagation*, pages 318–362. 1987.

E. Russeil et al. Multi-View Symbolic Regression. *arXiv e-prints*, art. arXiv:2402.04298, Feb. 2024. doi: 10.48550/arXiv.2402.04298.

Russeil, E. et al. Rainbow: A colorful approach to multipassband light-curve estimation. *A&A*, 683:A251, 2024. doi: 10.1051/0004-6361/202348158.

H. N. Russell. Some problems of sidereal astronomy*. *Proceedings of the National Academy of Sciences*, 5(10):391–416, 1919. doi: 10.1073/pnas.5.10.391.

B. W. Rust. *Use of supernovae light curves for testing the expansion hypothesis and other cosmological relations*. 1974.

C. Sánchez et al. Photometric redshift analysis in the Dark Energy Survey Science Verification data. *MNRAS*, 445(2):1482–1506, Dec. 2014. doi: 10.1093/mnras/stu1836.

P. Sanchez-Saez et al. Alert Classification for the ALeRCE Broker System: The Light Curve Classifier. *The Astronomical Journal*, 161(3):141, Mar. 2021. doi: 10.3847/1538-3881/abd5c1.

N. E. Sanders et al. Toward Characterization of the Type IIP Supernova Progenitor Population: A Statistical Sample of Light Curves from Pan-STARRS1. *ApJ*, 799(2):208, Feb. 2015. doi: 10.1088/0004-637X/799/2/208.

A. Sarangi et al. Dust in Supernovae and Supernova Remnants I: Formation Scenarios. *Space Sci. Rev.*, 214(3):63, Apr. 2018. doi: 10.1007/s11214-018-0492-7.

I. H. Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):160, Mar 2021. ISSN 2661-8907. doi: 10.1007/s42979-021-00592-x.

P. Schloerb et al. A Decade of US Community Access to the Large Millimeter Telescope Alfonso Serrano. In *Bulletin of the American Astronomical Society*, volume 51, page 148, Sept. 2019.

R. M. Schmidt. Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. *arXiv e-prints*, art. arXiv:1912.05911, Nov. 2019. doi: 10.48550/arXiv.1912.05911.

Schulze, Steve et al. 1100 days in the life of the supernova 2018ibb - the best pair-instability supernova candidate, to date. *A&A*, 683:A223, 2024. doi: 10.1051/0004-6361/202346855.

T. Schweyer et al. SN 2019odp: A Massive Oxygen-Rich Type Ib Supernova. *arXiv e-prints*, art. arXiv:2303.14146, Mar. 2023. doi: 10.48550/arXiv.2303.14146.

T. R. O. S. C. O. C. SCOC. Survey cadence optimization committee's phase 2 recommendations. Technical Report PSTN-055, Febuary 2023.

S. Seager and G. Mallén-Ornelas. A Unique Solution of Planet and Star Parameters from an Extrasolar Planet Transit Light Curve. *ApJ*, 585(2):1038–1055, Mar. 2003. doi: 10.1086/346105.

B. Settles. *Active Learning*. Morgan & Claypool Publishers, 2012. ISBN 1608457257.

N. I. Shakura and R. A. Sunyaev. Black holes in binary systems. Observational appearance. *A&A*, 24:337–355, Jan. 1973.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 978-1-10-705713-5.

C. J. Shallue and A. Vanderburg. Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90. *AJ*, 155(2):94, Feb. 2018. doi: 10.3847/1538-3881/aa9e09.

S. J. Smartt et al. The death of massive stars - I. Observational constraints on the progenitors of Type II-P supernovae. *MNRAS*, 395(3):1409–1437, May 2009. doi: 10.1111/j.1365-2966.2009.14506.x.

E. Smirnov. A comparative analysis of machine learning classifiers in the classification of resonant asteroids. *Icarus*, 415:116058, June 2024. doi: 10.1016/j.icarus.2024.116058.

G. F. Smits and M. Kotanchek. *Pareto-Front Exploitation in Symbolic Regression*, pages 283–299. Springer US, Boston, MA, 2005. ISBN 978-0-387-23254-6. doi: 10.1007/0-387-23254-0_17.

H. Song et al. A tailor designed fluorescent 'turn-on' sensor of formaldehyde based on the bodipy motif. *Tetrahedron Letters*, 53(37):4913–4916, 2012. ISSN 0040-4039. doi: https://doi.org/10.1016/j.tetlet.2012.06.117.

F. R. Stephenson. SN 1006: the brightest supernova. *Astronomy & Geophysics*, 51(5):5.27–5.32, 10 2010. ISSN 1366-8781. doi: 10.1111/j.1468-4004.2010.51527.x.

D. F. Swinehart. The beer-lambert law. *Journal of Chemical Education*, 39(7):333, Jul 1962. ISSN 0021-9584. doi: 10.1021/ed039p333.

S. Tanigawa et al. HAYATE: photometric redshift estimation by hybridizing machine learning with template fitting. *MNRAS*, 530(2):2012–2038, May 2024. doi: 10.1093/mnras/stae411.

W. Tenachi et al. Class Symbolic Regression: Gotta Fit 'Em All. *arXiv e-prints*, art. arXiv:2312.01816, Dec. 2023. doi: 10.48550/arXiv.2312.01816.

W. Thomson. Delay networks having maximally flat frequency characteristics. *Proceedings of the IEE - Part III: Radio and Communication Engineering*, 96:487–490(3), November 1949. ISSN 0369-8947.

A. B. Tomaney and A. P. S. Crotts. Expanding the Realm of Microlensing Surveys with Difference Image Photometry. *AJ*, 112:2872, Dec. 1996. doi: 10.1086/118228.

J. L. Tonry et al. ATLAS: A High-cadence All-sky Survey System. *PASP*, 130(988):064505, June 2018. doi: 10.1088/1538-3873/aabadf.

Torricelli-Ciamponi, G. et al. Non-thermal emission from agn coronae. *A&A*, 438(1):55–69, 2005. doi: 10.1051/0004-6361:20041010.

B. Trakhtenbrot et al. A new class of flares from accreting supermassive black holes. *Nature Astronomy*, 3:242–250, Jan. 2019. doi: 10.1038/s41550-018-0661-3.

A. Tran et al. Carboxylate bodipy integrated in mof-5: easy preparation and solid-state luminescence. *J. Mater. Chem. C*, 11:14896–14905, 2023. doi: 10.1039/D3TC02581K.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. ISSN 00264423.

S.-M. Udrescu and M. Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631.

M. Vargas dos Santos et al. On the cosmological performance of photometrically classified supernovae with machine learning. *MNRAS*, 497(3):2974–2991, Sept. 2020. doi: 10.1093/mnras/staa1968.

S. P. Vaughan et al. The SAMI galaxy survey: predicting kinematic morphology with logistic regression. *MNRAS*, 528(4):5852–5863, Mar. 2024. doi: 10.1093/mnras/stae409.

F. Villaescusa-Navarro et al. The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. *ApJ*, 915(1):71, July 2021. doi: 10.3847/1538-4357/abf7ba.

V. A. Villar et al. Supernova photometric classification pipelines trained on spectroscopically classified supernovae from the pan-starrs1 medium-deep survey. *The Astrophysical Journal*, 884(1):83, oct 2019. doi: 10.3847/1538-4357/ab418c.

V. A. Villar et al. A Deep-learning Approach for Live Anomaly Detection of Extragalactic Transients. *ApJS*, 255(2):24, Aug. 2021. doi: 10.3847/1538-4365/ac0893.

R. F. Webbink. Double white dwarfs as progenitors of R Coronae Borealis stars and type I supernovae. *ApJ*, 277:355–360, Feb. 1984. doi: 10.1086/161701.

Weiler, M. et al. Spectrophotometric calibration of low-resolution spectra. *A&A*, 637:A85, 2020. doi: 10.1051/0004-6361/201936908.

M. C. Weisskopf et al. Chandra X-ray Observatory (CXO): overview. In J. E. Truemper and B. Aschenbach, editors, *X-Ray Optics, Instruments, and Missions III*, volume 4012 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 2–16, July 2000. doi: 10.1117/12.391545.

J. Whelan and J. Iben, Icko. Binaries and Supernovae of Type I. *ApJ*, 186:1007–1014, Dec. 1973. doi: 10.1086/152565.

M. Wiescher et al. Explosive hydrogen burning in novae. *A&A*, 160:56–72, May 1986.

C. N. A. Willmer. The absolute magnitude of the sun in several filters. *The Astrophysical Journal Supplement Series*, 236(2):47, jun 2018. doi: 10.3847/1538-4365/aabfdf.

A. J. Wilson et al. A naive Bayes classifier for identifying Class II YSOs. *MNRAS*, 521(1):354–388, May 2023. doi: 10.1093/mnras/stad301.

C. Wolf et al. Discovery of the Most Ultra-Luminous QSO Using GAIA, SkyMapper, and WISE. *Publ. Astron. Soc. Australia*, 35:e024, June 2018. doi: 10.1017/pasa.2018.22.

T. N. Woods et al. Solar irradiance reference spectra (sirs) for the 2008 whole heliosphere interval (whi). *Geophysical Research Letters*, 36(1), 2009. doi: https://doi.org/10.1029/2008GL036373.

S. E. Woosley et al. Supernova 1987A: Six Weeks Later. *ApJ*, 324:466, Jan. 1988. doi: 10.1086/165908.

S. E. Woosley et al. The Evolution of Massive Stars Including Mass Loss: Presupernova Models and Explosion. *ApJ*, 411:823, July 1993. doi: 10.1086/172886.

E. L. Wright et al. The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. *AJ*, 140(6):1868–1881, Dec. 2010. doi: 10.1088/0004-6256/140/6/1868.

S. Yang and J. Sollerman. Haffet: Hybrid analytic flux fitter for transients. *The Astrophysical Journal Supplement Series*, 269(2):40, nov 2023. doi: 10.3847/1538-4365/acfcb4.

Y.-C. Yeh et al. A novel model extended from the bouguer-lambert-beer law can describe the non-linear absorbance of potassium dichromate solutions and microalgae suspensions. *Frontiers in Bioengineering and Biotechnology*, 11, 2023. ISSN 2296-4185. doi: 10.3389/fbioe.2023.1116735.

D. G. York et al. The Sloan Digital Sky Survey: Technical Summary. *AJ*, 120(3):1579–1587, Sept. 2000. doi: 10.1086/301513.

L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings, Twentieth International Conference on Machine Learning*, volume 2, pages 856–863, 01 2003.

Y. Zhang and Y. Zhao. Astronomy in the Big Data Era. *Data Science Journal*, 14:11, May 2015. doi: 10.5334/dsj-2015-011.

W. Zheng and A. V. Filippenko. An Empirical Fitting Method for Type Ia Supernova Light Curves: A Case Study of SN 2011fe. *ApJ*, 838(1):L4, Mar. 2017. doi: 10.3847/2041-8213/aa6442.