The Astrophysical Journal, 877:26 (14pp), 2019 May 20 https://doi.org/10.3847/1538-4357/ab1906

Probing an X-Ray Flare Pattern in Mrk 421 Induced by Multiple Stationary Shocks: A Solution to the Bulk Lorentz Factor Crisis

Olivier Hervet[1], David A. Williams[1], Abraham D. Falcone[2], and Amanpreet Kaur[2]
[1] Santa Cruz Institute for Particle Physics and Department of Physics, University of California, Santa Cruz, CA 95064, USA; ohervet@ucsc.edu
[2] Department of Astronomy and Astrophysics, Pennsylvania State University, University Park, PA 16802, USA
Received 2019 January 5; revised 2019 April 11; accepted 2019 April 12; published 2019 May 21

Abstract

The common observations of multiple radio VLBI stationary knots in high-frequency-peaked BL Lacs (HBLs) can be interpreted as multiple recollimation shocks accelerating particles along jets. This approach can resolve the so-called "bulk Lorentz factor crisis" of sources with a high Lorentz factor deduced from maximum γ–γ opacity and fast variability and apparently inconsistent slow/stationary radio knots. It also suggests that a unique pattern of the nonthermal emission variability should appear after each strong flare. Taking advantage of the 13 yr of observation of the HBL Mrk 421 by the X-ray Telescope on the *Neil Gehrels Swift Observatory* (*Swift*-XRT), we probe for such an intrinsic variability pattern. Its significance is then statistically estimated via comparisons with numerous similar simulated light curves. A suggested variability pattern is identified, consistent with a main flare emission zone located in the most upstream 15.3 GHz radio knot at 0.38 mas from the core. Subsequent flux excesses in the light curve are consistent with a perturbation crossing all of the downstream radio knots with a constant apparent speed of 45*c*. The significance of the observed variability pattern not arising from stochastic processes is found above three standard deviations, opening a promising path for further investigations in other blazars and with other energy bands. In addition to highlighting the role of stationary radio knots as high-energy particle accelerators in jets, the developed method allows estimates of the apparent speed and size of a jet perturbation without the need to directly observe any motion in jets.

Key words: acceleration of particles – BL Lacertae objects: individual (Markarian 421) – galaxies: jets – radiation mechanisms: non-thermal

1. Introduction

Multiwavelength studies of the variability and modeling of radio-loud active galactic nuclei (AGN) broadband spectral energy distributions (SEDs) attest to a compact emission zone moving with a high Lorentz factor close to the central engine. The particle individual Lorentz factors are often estimated to be above 10^6 for the most energetic blazars, implying long-standing and powerful particle acceleration mechanisms. While the scenario of magnetic reconnection has received considerable attention during recent years, due to recent progress with MHD simulations (Sironi & Spitkovsky 2014), the scenario of acceleration by shocks remains the most studied and accepted for the typical activity state of radio-loud AGN and their common variability (Marscher & Gear 1985; Spada et al. 2001; Fromm et al. 2011).

The shock scenario is supported by multiple observations of gamma-ray flares in coincidence with the emergence of a jet perturbation (or overdensity) in or close to the radio core, mainly seen in flat-spectrum radio quasars (FSRQs) and some low- or intermediate-frequency-peaked BL Lacs (LBLs and IBLs; Jorstad et al. 2001; Marscher et al. 2008; Abeysekara et al. 2018). The formation of recollimation shocks (also referenced as conical standing shocks or reconfinement shocks) in jets is also a phenomenon naturally observed in hydrodynamic and MHD jet simulations as soon as a supersonic, or super-Alfvénic, nonpressured matched flow propagates through an external medium. This pressure mismatch at the interface between the jet inlet and the external medium generates two conical waves, namely a shock wave and a rarefaction wave. The shock wave propagates toward the external medium and is reflected toward the jet axis as it reaches equilibrium with the external medium pressure. The rarefaction wave propagates toward the jet axis, locally dropping the jet pressure and accelerating the flow. The flow is then significantly slowed down after it reaches the reflection point of the conical waves at the jet axis. This process repeats and can produce a string of recollimation shocks until the full dissipation of energy carried out by the waves (e.g., Falle 1991; van Putten 1996; Gómez et al. 1997; Mizuno et al. 2015; Hervet et al. 2017).

Contrary to other blazar types, high-frequency-peaked BL Lacs (HBLs) show mainly stationary or low-speed very long baseline interferometry (VLBI) radio features (radio knots) in their jets, in stark contrast to the high Lorentz factor values deduced from their variability or SED modeling (Hervet et al. 2016; Piner & Edwards 2018). Most of the interpretations of this issue imply two distinct regions between radio knots and high-energy emission zones. Slow/stationary radio knots are assumed to come from a slower and wider jet part than the high-energy emission zone. It can be understood as a strong jet deceleration very close to the core (Georganopoulos & Kazanas 2003) or a stratified jet with differential speeds as nonsteady outflows (Lyutikov & Lister 2010) or spine-layer structure (Ghisellini et al. 2005; Piner & Edwards 2018). We adopt the interpretation of slow/stationary radio knots as a multiple recollimation shock structure, very stable for these sources due to their lower outer jet kinetic power (Hervet et al. 2017).

Following the shock-in-jet model developed by Marscher & Gear (1985), a flare should happen when a perturbation (or moving shock) passes trough a recollimation shock. This scenario was adapted and improved by many further works and is quite successful as a picture of the general broadband blazar flaring behavior (e.g., Komissarov & Falle 1997;

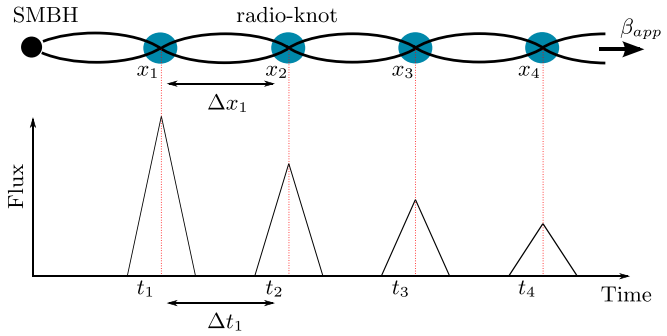

Figure 1. Simplified scheme of the expected light-curve signature of a perturbation crossing the knots x_i with an apparent speed $\beta_{\rm app}$ linking the interknot distance Δx_1 with the delay between two consecutive flares Δt_1.

Türler et al. 2000; Nalewajko & Sikora 2009; Nalewajko et al. 2012; Fromm et al. 2011, 2016; Türler 2011; Marscher 2014). Successive flares are then believed to be triggered by a stochastic injection from the central engine. However, while this approach assumes that one shock at the base of the jet is responsible for the main dissipation process, it does not consider the other potential flares produced by downstream shocks. Here we investigate the possibility of successive flares associated with successive recollimation shocks in relativistic jets. If we relate stationary radio knots to recollimation shocks, we can predict a distinct pattern of variability based on interknot gaps. Thus, after each strong flare occurring at the base of the jet, one should detect several other flares in accordance with the VLBI radio knot distribution in the jet for a given velocity of the flow. The confirmation of such a pattern in HBL light curves would validate the role of stationary radio knots as high-energy particle accelerators and characterize the apparent speed and size of underlying perturbations, which is extremely valuable for constraining the modeling parameters.

In Section 2 we introduce the basic concept of the proposed scenario and the ideal source for its application, Mrk 421. In Sections 3 and 4 we describe how X-ray long-term light curves are handled in view of having the most efficient probe to detect a possible intrinsic post-flare variability pattern. The theoretical models used to check our scenario are developed in Section 5. In Section 6 we describe the method used to create simulated light curves as similar as possible to the real data set and discuss biases induced by these simulations. Results and a general discussion are in Section 7.

Throughout this paper, a flat ΛCDM cosmology is adopted with $H_0 = 69.7$ km s^{-1} Mpc^{-1}, $\Omega_M = 0.286$, and $\Omega_v = 0.714$ (Bennett et al. 2014). It leads to a projected scale of 1 mas = 0.603 pc at the redshift $z = 0.030$ of Mrk 421.

2. Method and Application to Mrk 421

2.1. Concept of the Method

The core of the method is to probe flares associated with the flow passing through the knots, assuming they are stationary shocks. For a given apparent speed $\beta_{\rm app}$, the time delay of the secondary flares can be set by knowing the radio knot positions, as shown in Figure 1.

Considering a constant speed of the flow through a straight jet, the time gap Δt between each successive flare in the light curve should be directly proportional to the observed interknot gap Δx. We have the relation

$$\Delta t_i = (1+z)\frac{\Delta x_i}{c\beta_{\rm app}}. \quad (1)$$

Considering the association of radio knots with recollimation shocks, the underlying flow is expected to accelerate upstream of each shock due to the presence of rarefaction waves locally decreasing the pressure. The speed should then decrease after the shock. The realistic speed profile would be an oscillation, likely with a slower acceleration due to the global conical opening of the jet (Gómez et al. 1997; Komissarov & Falle 1997; Mizuno et al. 2015; Hervet et al. 2017). Throughout this paper, we consider the approximation of an average constant speed of the underlying flow to be valid, with the main motivation being keeping the light-curve model developed in Section 5 as simple as possible. This approximation can be supported with the observed motions in radio jets, which in the majority are well fitted by a constant-speed motion (Lister et al. 2016). As further discussed in Section 5, the theoretical model developed also considers the width of the peaks from the size of the radio knots and a damping factor between successive flares.

2.2. Mrk 421: The Ideal Candidate

Mrk 421 is the brightest X-ray and gamma-ray HBL in the sky in its flaring and average state (Stroh & Falcone 2013). It is one of the most monitored blazars in all wavelengths and shows frequent giant flares (e.g., Aleksić et al. 2015; Abeysekara et al. 2017; Fraija et al. 2017). It is perfectly adapted for this study by also presenting four well-defined VLBI quasi-stationary knots within 5 mas of the radio core at 15.3 GHz, as shown in Figure 2 (left) from the MOJAVE collaboration.[3] All of the observed knots show either nonradial or downward motions. Such motions would be very challenging to describe with a ballistic model but can naturally match low-amplitude shifts/oscillations of quasi-stationary recollimation shocks. The fastest measured knot measured in VLBI (number 6) displays an apparent speed of $0.217 \pm 0.026c$, roughly perpendicular to the jet direction (Lister et al. 2016), and the usual Doppler factor deduced from broadband SED modeling is about 20–25 (Błażejowski et al. 2005; Baloković et al. 2016; Carnerero et al. 2017; Kapanadze et al. 2018a, 2018b), which can be seen as a lower limit, since the Doppler factor is usually constrained from the shortest-variability timescale observed and the maximum possible photon–photon opacity within the emitting region. For a canonical blazar angle with a line of sight of 2°, the SED models lead to a Lorentz factor $\Gamma_{\rm model} \gtrsim 14$, which should be related to the apparent downstream speed of $\beta_{\rm app} \gtrsim 11c$. Mrk 421 is then strongly affected by the bulk Lorentz factor crisis, which is ideal for our study.

For this study we consider these four knots as stationary recollimation shocks with their distance to the radio core given by the mean value of the measured distances from the MOJAVE Collaboration. The uncertainties on their distance to the core and radius are given by the standard deviation of the data set. The Mrk 421 knot string follows a conical expansion well, as shown in Figure 2 (right). The knots' radius is fitted by a linear function $f(x) = (0.195 \pm 0.015)\ x + (3.94 \pm 0.76) \times 10^{-2}$ mas, with a reduced χ^2 of 0.28. The radio knot

[3] http://www.physics.purdue.edu/MOJAVE

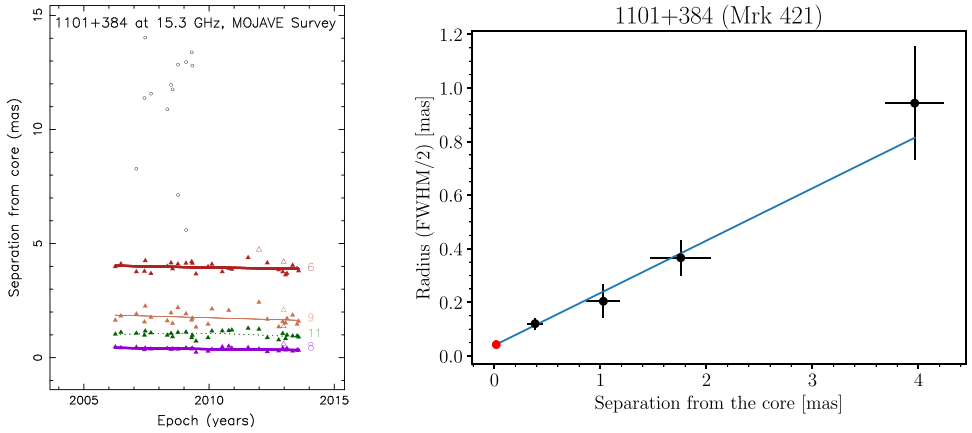

Figure 2. Left: temporal evolution of the knot–core distances of Mrk 421 observed by MOJAVE. Four quasi-stationary knots are firmly detected within 5 mas of the radio core between 2006 and 2014. Gray dots are considered to be nonrobust features to measure the jet kinematics. Right: mean core distance and radius of radio knots with standard deviation, fitted by a linear function. The red point is the radio core. Adapted from Lister et al. (2016).

positions of Mrk 421 were measured in several other studies for different frequencies and epochs. Although the MOJAVE data set is the one that is the most simultaneous with the light curve in our study, it remains relevant to check the consistency of these measurements with the previous observations described in Piner et al. (2010; with an extended data set from Piner et al. 1999; Piner & Edwards 2005) and Lico et al. (2012).

Piner et al. (2010) reported Very Long Baseline Array (VLBA) observations at 22 and 43 GHz of Mrk 421 between 1994 and 2009. They observed knots consistent with the ones detected by MOJAVE; however, they detected a supplementary component between 2008 and 2009 at 43.2 GHz, C8, at ∼0.2 mas from the core. Lico et al. (2012), who performed VLBA observations in 2011, had similar observations. While their 15.36 GHz analysis is consistent with the one presented by MOJAVE, at 23.804 GHz, the first radio knot can be divided into two distinct components named C4a and C4b. Piner et al. (2010) noticed that these 43.2 GHz knots, C7 and C8 (or C4a and C4b from Lico et al. 2012), can be associated with the eastern and western limb-brightened jet structure of the jet (see Figure 3).

The limb-brightened emission is likely an indication of a spine-sheath jet where the outer jet is either more Doppler boosted (due to a smaller angle with the line of sight) or presents a larger intrinsic synchrotron emissivity. Throughout this study, we consider this local limb-brightened emission at high frequencies as a single shock in the inner jet, associated with the position of knot 8. For more clarity, we reference the studied knot positions given by MOJAVE in Table 1 with their associated names from previous studies.

The high-energy emission zone location(s) of radio-loud AGN is still an unresolved question. Multiple studies have highlighted the likely presence of multiple high-energy zones within the jets from broadband emission models and variability studies (e.g., Raiteri et al. 2010; Tavecchio et al. 2011; Nalewajko et al. 2012; Hervet et al. 2015). When comparing the high and very high energy flares with radio VLBI measurements, it appears that flares can be associated with either the radio core or a radio knot outside the core (e.g., Abramowski et al. 2012; Marscher 2014). We note that the radio core is by definition ambiguous and can itself be

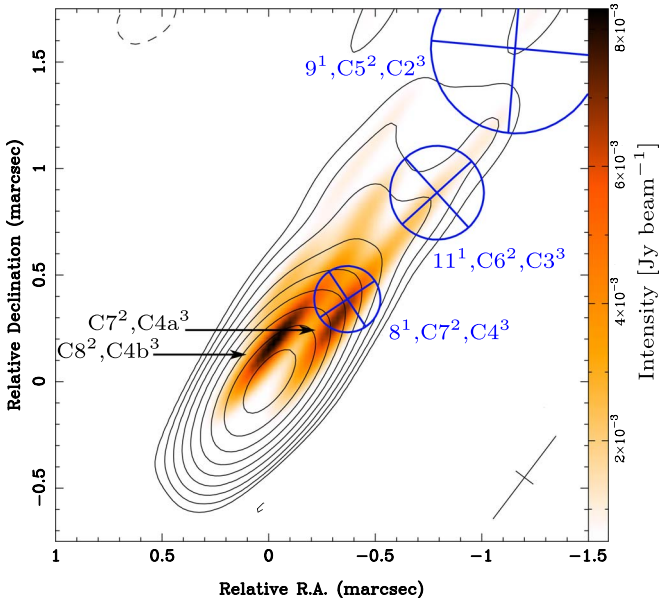

Figure 3. The 43 GHz radio map contour with color image representing the intensity after removing the core emission, observed on 2008 December 8 (Piner et al. 2010). Blue circles are 15.3 GHz components fitted by a 2D Gaussian from 2011 January 14 (Lico et al. 2012). The various knot IDs follow the references: MOJAVE, 1; Piner et al. (2010), 2; Lico et al. (2012), 3. We can note that the 15.3 GHz knots presented here slightly differ in size and position from the 7 yr average values we are using in our study.

Table 1
Projected Distance from the Radio Core of the Four VLBI Quasi-stationary Radio Knots Referenced by MOJAVE with Their Different Associated Names

Knot No. (1)	Knot No. (2)	Knot No. (3)	Core Distance (mas) (1)	Radius (mas) (1)
Core	⋯	⋯	⋯	$4.24 \pm 1.62 \times 10^{-2}$
8	C7	C4	0.38 ± 0.07	$1.20 \pm 0.22 \times 10^{-1}$
11	C6	C3	1.03 ± 0.16	$2.04 \pm 0.63 \times 10^{-1}$
9	C5	C2	1.76 ± 0.29	$3.66 \pm 0.66 \times 10^{-1}$
6		C1	3.96 ± 0.28	$9.44 \pm 0.21 \times 10^{-1}$

Note. 1: MOJAVE; 2: Piner et al. (2010); 3: Lico et al. (2012).

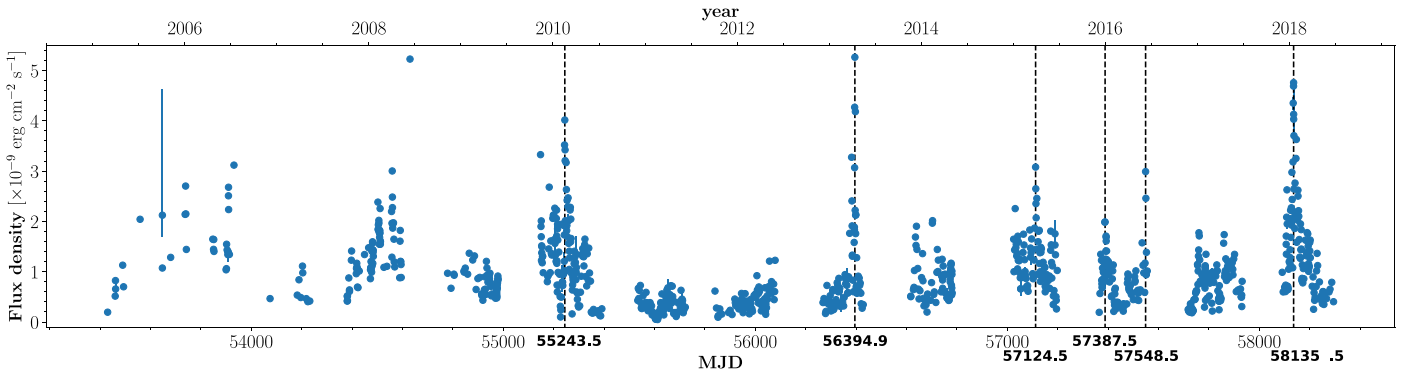

Figure 4. *Swift*-XRT light curve from 2005 March to 2018 May. Vertical dashed lines represent the selected flares as discussed in Section 4.1.

composed of several radio knots when observed with better angular resolution (Gómez et al. 2016). Not knowing if the radio core of Mrk 421 could be associated with a strong first recollimation shock, we probe the following two hypotheses.

1. The biggest observed flares are produced in the radio core, then four following flares are expected in the light curve.
2. The biggest observed flares are produced in the first radio knot, then three following flares are expected in the light curve.

3. *Swift*-XRT Analysis

The X-ray Telescope on the *Neil Gehrels Swift Observatory* (*Swift*-XRT; Burrows et al. 2005) is sensitive in the soft X-ray energy range (0.3–10 keV), which is excellent for measuring flux at or near the synchrotron peak energy for HBLs such as Mrk 421. Large-amplitude flares typically produce copious synchrotron emission in this energy band. *Swift*-XRT has proven to be highly capable of monitoring both long-term flux variability (with a baseline of >14 yr) and large-amplitude flares with precise flux and spectral measurements.[4]

Since Mrk 421 is typically at a high enough count rate to induce pileup of photons in photon-counting (PC) mode, most of the observations were taken in window-timing (WT) mode. The cleaned level 3 event files were used for extracting data products. Initially, a cleaned event file was separated into individual snapshots (i.e., individual pointed observations). Each snapshot was then utilized to extract an image within the 0.3–10 keV energy range. The first 150 s of data were discarded from each snapshot for the WT mode observations in order to exclude data with any spacecraft settling issue that might have occurred during this time interval. A pileup correction was performed using the method described in Romano et al. (2006). The extracted spectrum for each snapshot was obtained by selecting a box with dimensions 40 × 20 pixels (2.″36 pixel^{-1}). The source box region was rotated as per roll angle for the given snapshot. An annular boxed background region rotated at the same angle, with a size of 100 pixels (same height as source region; 20 pixels), was used to obtain the background spectrum.

For observations taken in PC mode, first, a circular source region with a size of 20 pixels and an annular background region were chosen to extract spectra. If the source counts were found to be >0.6 counts s^{-1}, a pileup correction was performed. In order to correct for pileup, an appropriate annular region was selected as the source region for the final spectrum extraction, ensuring that the count rate drops to at least 0.6 counts s^{-1}.

Fluxes and spectra are extracted with 1 day binning. We utilized XSpec (Arnaud 1996) to fit all spectra with a model comprised of a log parabola combined with absorption as specified in the Tuebingen–Boulder interstellar medium (ISM) absorption model. This X-ray spectral shape of Mrk 421 is confirmed by previous studies (Massaro et al. 2004). The hydrogen column density was fixed to 0.019 cm^{-2} , which was derived from the LAB survey (Kalberla et al. 2005). Within XSpec, we utilized cflux to determine the unabsorbed flux in the 0.3–10 keV band.

The full *Swift*-XRT light curve is shown in Figure 4.

4. Formatting the Data Set

4.1. Major Flare Selection

The brighter a flare is, the more we expect it to be associated with an ejection through the main recollimation shock. Therefore, we want to select the brightest flares in X-rays as input to our method. The way the flares are selected can impact the results of the study. Selecting too few flares will not bring enough constraints on the method, with the risk of being biased against the typical behavior of the source by selecting "exceptional" events. On the other hand, selecting many weak flares increases the risk of injecting intrinsic stochastic fluctuations into the method and burying any possible variability pattern in noise.

As a middle ground, we select a flare only if the peak of the flux is above the 90th percentile of the distribution, giving a threshold value of 1.90 × 10^{-9} erg cm^{-2} s^{-1} for the *Swift*-XRT data set. Later sections discuss the impact of using a different flux threshold (FT) to select flares. No flare is considered if it has a significance less than 3σ above the median flux. Also, in order to have confidence that a high measured flux is the flare peak, a flare is selected only if it has at least one data point in the 10 days before it and three data points in the 10 days after it. This ensures having a temporal estimation (∼day) of a flare, which is relevant for the light-curve analysis method, as developed in the following section.

Finally, we intend to select the first flare that starts a sequence and to avoid having two series too close to each other, which can mislead the method. It can be done by selecting only the strongest flare in a given time range. Thus, a flare is not selected if it happens during the 100 days before or after a stronger flare. This exclusion zone of 100 days applies even if the stronger flare is not selected for our method (due to

[4] https://www.swift.psu.edu/monitoring/

a bad timing estimation). This cut, however, has some limitations. In case of low apparent speeds of the flow ($\lesssim 5c$), the time gap expected between the nearby radio knots of Mrk 421 is more than 100 days, making our method less sensitive for those speeds. And having too big a time gap would lead to a limited number of flares. The impact of the choice of these cuts on the final results is quantified in Section 5.4 to estimate systematic errors.

The date of the selected flares and their associated positions in the light curve are shown in Figure 4.

4.2. Light-curve Stacking

This study aims to see if there is a regular intrinsic pattern in the light curve after a strong flare. However, each X-ray flare of Mrk 421 is different, with a variability apparent afterward that we assume is in part due to strong and unpredictable turbulences within the jet. Also, many observing gaps after flares, which can be as big as several months, make the definition of a variability pattern even more difficult in a flare-by-flare study.

Hence, we stack all of the selected flares on the dates given in Figure 4. By working on this stacked light curve, we expect that the pure stochastic variability will play a reduced role and that we have a typical post-flare data set without large observing gaps. There is a risk that an unevenly stacked data set creates a misleading pattern not associated with any physical process. This issue is addressed in Section 5.4 where we quantify the systematic errors associated with such a method, and in Section 6 where we apply the same stacking process on simulated light curves.

The final stacked light curve is from 40 days before the selected flares to 600 days after them, which theoretically allows us to probe apparent speeds as low as $0.5c$ for the main flare in the radio core and $0.8c$ for the main flare in the upstream radio knot (at those low apparent speeds, a perturbation would take $\sim$600 days to reach the next downstream knot). In order to have a clear picture of a possible variability pattern, each flare is normalized to the strongest one. We apply a normalization factor only on fluxes above the full light-curve median to not alter the flux baseline of the source. The normalized flux for the stacked light curve i applied to a data point j takes the form

$$\text{If } (F_{i,j} - m) > 0:$$
$$F_{\mathrm{norm},i,j} = \frac{F_{\max} - m}{F_{\max,i} - m}(F_{i,j} - m) + m, \tag{2}$$

where m and $F_{\max}$ are the median value and biggest flare, respectively, in the original light curve; $F_{i,j}$ is the original flux point in the light curve i; and $F_{\max,i}$ is the maximum flux of the light curve i. The error bars are adapted accordingly to keep the same error/flux ratio.

The resulting stacked light curve is presented in Figure 5. At first sight, we notice the great dispersion of data points, which is in part due to the duplicate of flares inherent to the stacking method. These flare duplicates also have their fluxes amplified by the normalization process. But mostly, this dispersion points toward strong stochastic X-ray flux variations of Mrk 421, making it hard to discern a possible intrinsic variability pattern.

For display purposes, a clearer view is given by rebinning the data. Since the stacked data set is unevenly sampled, with a concentration of points around the stacked flares, we adopt a binning by keeping a constant number of data points per bin. The binned data, as well as the rms dispersion within each bin, are shown in Figure 5. Two excesses at 11 and 23 days after the main flares, and a possible one around 64 days, suggest a post-flare variability pattern. Also, the amplitude of these excesses is decreasing with time, which is consistent with adiabatic (expansion) and radiative losses. In order to evaluate the significance of these suggested features, simulations are required to assess the impact of the various effects, such as binning, sampling, and stacking; this is discussed in later sections.

The strong excess seen in the last bin is intriguing. We consider it unlikely to be associated with the process we want to probe. First, this very long delay seems unlikely to be associated with the real flow speed of Mrk 421, which is known to show a strongly Doppler-boosted radiation. Also, the amplitude of such an excess is close to the ones of the selected flares, leading to a noncooling jet over long periods. It is, however, possibly highlighting a long-term periodicity of Mrk 421 flares, possibly linked to the accretion disk timescale.

The HBLs are known to be the least powerful blazars and have been associated with a weak accretion mode known as the "advection-dominated accretion flow" (ADAF). Approximating the gas flow angular frequency Ω as the Keplerian angular frequency Ω_k (Manmoto et al. 1996), we have

$$\Omega = \left(\frac{GM_{\bullet}}{r}\right)^{1/2} \frac{1}{r - r_g}, \tag{3}$$

with the black hole mass $M_{\bullet} = 1.7 \times 10^8 M_{\odot}$ estimated from fundamental-plane-derived velocity dispersion (Woo et al. 2005) and the associated Schwarzschild radius $r_g = 5.03 \times 10^{13}$ cm. Then, an accretion disk perturbation with an orbital period of 600 days would be located at a distance $r = 233 r_g$ from the black hole, which could correspond to the interface between the ADAF and the outer standard thin disk structure (Esin et al. 1997).

5. Theoretical Models

5.1. Multi-Gaussian

The purpose of the presented model is not to simulate the particle physics processes of perturbation crossing shocks, such as particle acceleration, cooling, or radiative transfer. Several former studies addressed this approach via MHD-based and semi-analytic models (e.g., Türler et al. 2000; Mimica et al. 2009; Fromm et al. 2011, 2016; Türler 2011). While these models shed light on the shock mechanisms in jets, they would be unfit to statistically probe the existence of a light-curve pattern induced via multiple shocks due to degeneracies between numerous parameters or inadequately long computation times. Instead, we want to probe a signature of successive shocks with the simplest possible function and maximum physical constraints given by VLBI observations in order to reduce the number of free parameters.

We first consider a general flux baseline $B(t)$ as a linear function in order to picture a possible long-term flux variation of the 640 days' stacked light curve not associated with the

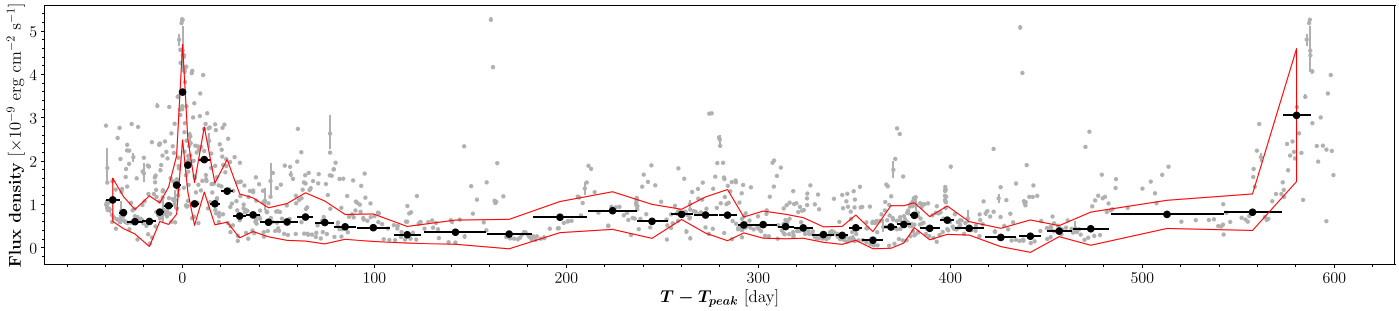

Figure 5. Flare-stacked light curve used to probe a post-flare variability pattern. For clarity, we show a binned data set with 18 data points per bin. The red lines picture the rms range associated with the flux dispersion of stacked light curves.

multiple flares probed:

$$B(t) = f_0 + f_1 t. \tag{4}$$

On top of this baseline, a multi-Gaussian function is implemented: a five-Gaussian function for the radio core-flare hypothesis and a four-Gaussian function for the upstream radio knot hypothesis. The time gap Δt_i between each peak depends on the free parameter of the apparent speed and the interknot gaps measured in the VLBI observation, as expressed in Equation (1). The timing of each expected peak t_i can be expressed from Equation (1) by

$$t_i = 7.26 \times 10^2 \frac{(1+z)x_i}{\beta_{\rm app}} \text{ days}. \tag{5}$$

The spread of the Gaussian $\sigma_{G,i}$ is scaled to the size of the radio knots R_i following the formula

$$\sigma_{G,i} = \frac{7.26 \times 10^2}{\sqrt{2\log(2)}} \frac{(1+z)R_i}{\beta_{\rm app}} S + C \text{ days}, \tag{6}$$

with R_i the knot radius given in Table 1 and Figure 2 (right). Here S and C are scaling factors. The coefficient $\sqrt{2\log(2)}$ is used to convert the measured size of the radio knots expressed in Gaussian FWHM/2 into a standard deviation. Since we consider a constant apparent speed, the Gaussian spread in days is strictly proportional to a unit of size (see Equation (24)).

Each peak is then defined as

$$P_i(t) = \frac{1}{\sigma_{G,i}\sqrt{2\pi}} \exp\left[\frac{-(t-t_i)^2}{2\sigma_{G,i}^2}\right]. \tag{7}$$

Finally, at constant power, the peaks should have a flux decrease roughly proportional to the volume of the emission zones. We then express a Gaussian amplitude decrease as

$$A_i = \alpha/\sigma_{G,i}^3. \tag{8}$$

The full theoretical model, including the baseline, is thus given by

$$G_m(t) = \sum_{i=n}^{5} [A_i P_i(t)] + B(t), \tag{9}$$

with $n = 1$ or 2 following the radio core or radio knot hypothesis.

The function $G_m(t)$ contains only six free parameters; in order to obtain a realistic model, we constrain the parameter space of some of them. For minimal accuracy of the method, we want to be able to probe at least two peaks associated with post-flare events in the light curve, which sets a minimal apparent speed of $\sim 2c$ considering a secondary peak at the maximum delay of 500 days. The minimal apparent speed is deduced from the closest consecutive knots associated with this delay (we considered a maximum delay shorter than the 600 days probed to be sure to have good resolution of such a peak). The maximal measured apparent speed in a blazar is $\sim 50c$, in the jet of PKS 0805–07 (Lister et al. 2016). We consider $\beta_{\rm app} \in [2, 70]$.

All of the selected X-ray flares in the original light curve (considered as the first flare of the sequence) have a duration well below 50 days. We consider this value as a constraint on $\sigma_{G,1}$ from Equation (6). In this equation, we assume that the width of the Gaussian cannot grow faster than the width of knots along the jet. Indeed, it is safe to assume that the high-energy shock zone is only a portion of the observed radio knots. Due to the energy loss along the jet propagation, it is likely that this shock zone will not occupy a relatively larger area in the downstream knots. We set the boundaries of $C \in [0, 50]$. So, following the constraint on $\sigma_{G,1}$ and C, we set the parameter space of $S \in [0, 3.8]$ for a flare in the radio core and $S \in [0, 1.4]$ for a flare in the upstream radio knot.

The parameter space of all parameters is summarized in Table 2.

5.2. EMG

Blazar flare profiles may present skewness, for which the decay is usually longer than the rise time. Such a skewness is most often modeled by a combination of two exponential functions (e.g., Abdo et al. 2010; Chatterjee et al. 2012). We consider a typical flare profile as an exponentially modified Gaussian (EMG) function, which has similar properties as the two exponential ones and the same number of free parameters. The EMG has the specificity to rise as a Gaussian function and decay as an exponential one.

We use the EMG function expression

$$\text{EMG}(t) = \frac{h\sigma}{\tau}\sqrt{\frac{\pi}{2}} \exp\left(\frac{\sigma^2}{2\tau^2} - \frac{t-\mu}{\tau}\right) \times \text{erfc}\left[\frac{1}{\sqrt{2}}\left(\frac{\sigma}{\tau} - \frac{t-\mu}{\sigma}\right)\right] + B(t), \tag{10}$$

where h is the amplitude, σ is the Gaussian standard deviation, the mean $m = \mu + \tau$ is set at zero, and τ is the exponential relaxation time.

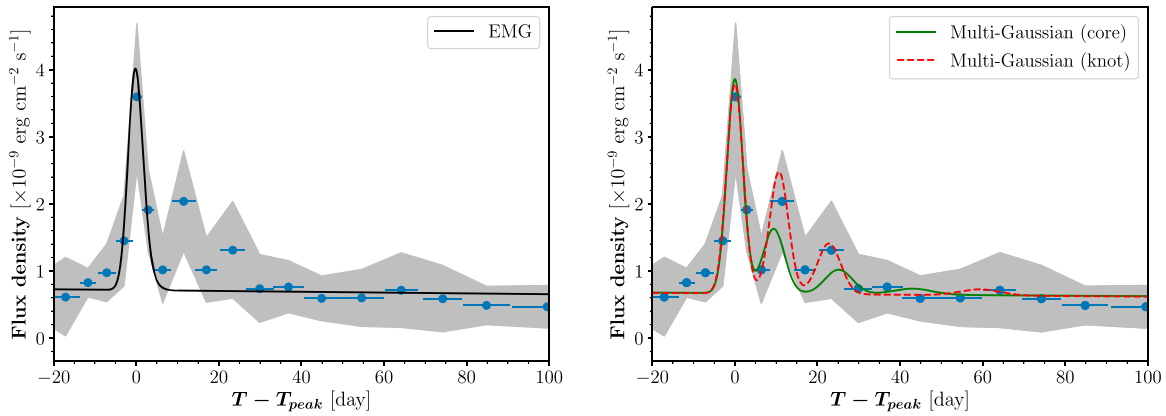

Figure 6. Comparison of the different flares models: EMG, multi-Gaussian from a main flare in the radio core, and multi-Gaussian from a main flare in the upstream radio knot. The models are represented on top of the stacked light curves. For clarity, we show a binned data set with 18 data points per bin. The gray band is the rms range associated with the flux dispersion of stacked light curves.

Table 2
Parameter Boundaries Applied to the Light-curve Models

	Parameter	Boundaries	Unit
Baseline	f_0	$[0, \infty]$	erg cm^{-2} s^{-1}
	f_1	$[-\infty, \infty]$	erg cm^{-2} s^{-1} day^{-1}
Multi-Gaussian	α	$[0, \infty]$	day^4 erg cm^{-2} s^{-1}
	$S_{\rm core}$	[0, 3.8]	⋯
	$S_{\rm knot}$	[0, 1.4]	⋯
	C	[0, 50]	day
	$\beta_{\rm app}$	[2, 70]	c
EMG	h	$[0, \infty]$	erg cm^{-2} s^{-1}
	σ	[0.5, 50]	day
	τ	[0.5, 50]	day

As for the multi-Gaussian model, the EMG model takes into account a linear baseline $B(t)$. Hence, the full EMG model has five free parameters.

5.3. Model Comparison

The best fits of the three models are presented in Figure 6. The data point dispersions in the stacked light curves (associated with intrinsic stochastic variations) have larger amplitudes that the measurement errors associated with each observation. This large data dispersion leads to extremely high values of χ^2, whatever the model used. The models presented do not aim to describe each variation of the fluxes in the light curve but look for an intrinsic regular pattern within the stochastic noise. While the fit quality cannot validate a given model by itself, it can, however, be used to compare the performance of each model.

All of the fitted models show excesses above the baseline within a period of 100 days after the stacked flares. Considering only the range where at least one model is above 1% of the baseline, $[t_{\rm flare}-7, t_{\rm flare}+70]$, the fit qualities improve, as well as the relative difference between models (see Table 3).

The EMG function has the worst χ^2. Although it has a visually good representation of the main flare, it does not describe the excesses above the baseline after the flare, contrary to the multi-Gaussian. Both multi-Gaussian functions, core and knot, are pointing toward second and third peaks located at ∼9–11 and ∼22–26 days, respectively, after the main flare. However, the knot scenario is favored with the lowest χ^2, and each of its expected peaks matches the observed flux excesses well. Thus, in the following, we focus on the theoretical model of a main flare from the upstream radio knot.

Table 3
Fitting Results for the Different Proposed Models

	χ^2/dof $t \in [-20, +100]$	χ^2/dof $t \in [-7, +70]$
EMG	$7.86 \times 10^5/325$	$6.05 \times 10^5/233$
G_m, core	$7.16 \times 10^5/324$	$5.26 \times 10^5/232$
G_m, knot	$6.74 \times 10^5/324$	$4.83 \times 10^5/232$

Note. The first column is for the time range presented in Figure 6, while the second column considers the range where at least one model is above 1% of the baseline.

5.4. Statistical and Systematic Uncertainties

The statistical uncertainties on the fitted model parameters are estimated from the covariance matrix calculation done with the python `scipy.optimize.curve_fit` method.[5] The data dispersion being much larger than the error associated with each point, the original covariance matrix is scaled to the reduced χ^2 of the best fit to avoid an obvious underestimation of the statistic uncertainties. This process scales the original error bars to match the sample variance of the residuals after the fit.

While being a reasonable method, we raise a warning that the statistical uncertainties estimated this way are likely close to, but not exactly, the true ones (e.g., by assuming a normal distribution of the fit residuals).

The way flares are selected in the X-ray light curve plays a role in the fitting results, leading to associated systematic uncertainties. We determine the systematic uncertainties of the model parameters by applying different cuts in the flare selection. As defined in Section 4.1, three cuts are applied to select flares: the FT; the minimum time gap between two selected flares, $\Delta_{\rm Flares}$, and the time range around a given flare where we want a minimum amount of data taken, $\Delta t_{\rm data}$. In order to estimate systematic uncertainties, we consider the effects of applying a much looser and harder set of cuts. The loose cuts select many more flares (13), while the hard ones

[5] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.html

Table 4
Cut Sets Applied to Select Flares in the X-Ray Light Curve

Cuts	Default	Loose	Hard	Unit
FT	90%	80%	95%	flux percentile
$\Delta_{\rm Flares}$	100	75	150	days
$\Delta t_{\rm data}$	10	20	5	days

Note. Systematic uncertainties are estimated from the loose and hard cuts.

select fewer (five) but better-defined flares. The different cuts are summarized in Table 4.

The systematic loose cuts uncertainties for each parameter are calculated as $\Delta_{\rm sys,loose} = \text{loose} - \text{default}$. The same is applied for hard cuts. If the values of loose and hard cuts are not bracketing a default parameter value, only the larger $\Delta_{\rm sys}$ is taken into account. The default parameter values and the systematic uncertainties for the different models are given in Table 5.

These two alternative cut sets do not impact the favored interpretation of the strongest flares originating from the upstream radio knot. Indeed, the knot-flare model always has the lowest χ^2 value, whatever the cut choice.

6. Light-curve Simulation

The significance of the knot-flare scenario against the null hypothesis can be estimated via comparisons with multiple realistic simulated light curves of Mrk 421. By applying the exact same method on simulated light curves, one can estimate the probability that the observed post-flare variability pattern is from pure stochastic noise.

The conditions we want to fulfill for the simulated light curves compared to the original one are

1. similar power spectrum density (PSD),
2. similar time sampling, and
3. similar flux distribution.

6.1. PSD

The *Swift*-XRT PSD is produced using the LombScargle package of Astropy.[6] The frequency range considered to build the PSD is delimited by the total light-curve length, $\nu_{\rm min} = 1/T$, with T the 13.3 yr span of the total light curve, and the Nyquist frequency, defined as $\nu_{\rm max} = N/(2T)$, with N the number of data points (e.g., Uttley et al. 2002).

The PSD index is extracted from a power-law fit with a best value of $\eta = 1.35 \pm 0.01$ ($P_\nu \propto \nu^{-\eta}$). The power-law function has a good fit with $\chi^2_{\rm red} = 0.39$ for the logarithmically binned PSD shown in Figure 7.

6.2. Sampling

A simulated light curve is produced considering power-law noise with the index η by the `astroML.time_series.generate_power_law` tool[7] (Vanderplas et al. 2012), based on the method developed by Timmer & Koenig (1995).

In order to avoid the red noise leak (transfer of variability power from the low to high frequencies due to the finite length of observations), we simulate light curves 100 times larger than the observed one, then clip them to the original length. Also, the *Swift*-XRT data set is far from evenly sampled, mostly due to the observations being taken as "targets of opportunity" (ToOs). Since having a different sampling in the simulated light curves would bias a fair statistical test, we resample the simulated light curves by taking the interpolated fluxes corresponding to each observing date of *Swift*-XRT.

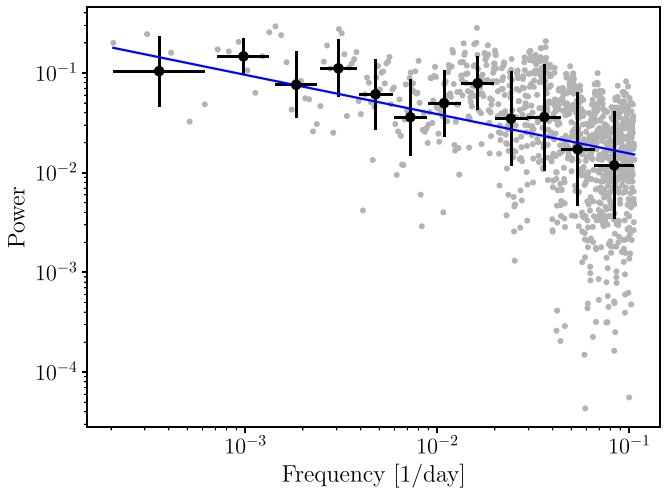

Figure 7. *Swift*-XRT PSD fitted by a power-law function. Black points are logarithmically binned data.

Table 5
Parameter Values for the Different Studied Models with Associated Systematic Uncertainties

Parameter	Value	Uncertainty
	EMG	
f_0	$(7.13 \pm 0.27) \times 10^{-1}$	$^{+0.40}_{-0.32} \times 10^{-1}$
f_1	$(-6.06 \pm 0.90) \times 10^{-4}$	$^{+2.5}_{-1.7} \times 10^{-4}$
τ	1.03 ± 0.42	$^{+0}_{-0.45}$
σ	1.63 ± 0.21	$^{+0}_{-0.25}$
h	3.77 ± 0.55	$^{+0.64}_{-0.71}$
	Core-flare Model	
f_0	$(6.67 \pm 0.28) \times 10^{-1}$	$^{+0.36}_{-0.25} \times 10^{-1}$
f_1	$(-4.80 \pm 0.91) \times 10^{-4}$	$^{+2.1}_{-1.4} \times 10^{-4}$
α	$(-1.09 \pm 0.25) \times 10^{2}$	$^{+0.48}_{-0} \times 10^{2}$
S	$(4.1 \pm 0.9) \times 10^{-1}$	$^{+1.3}_{-1.3} \times 10^{-1}$
C	1.56 ± 0.13	$^{+0}_{-0.25}$
$\beta_{\rm app}$	30.3 ± 1.6	$^{+3.9}_{-1.2}$
	Knot-flare Model	
f_0	$(6.63 \pm 0.27) \times 10^{-1}$	$^{+0.36}_{-0.23} \times 10^{-1}$
f_1	$(-4.69 \pm 0.90) \times 10^{-4}$	$^{+1.9}_{-0.7} \times 10^{-4}$
α	$(-1.22 \pm 0.26) \times 10^{2}$	$^{+0.53}_{-0} \times 10^{2}$
S	$(2.4 \pm 0.6) \times 10^{-1}$	$^{+1.1}_{-0.5} \times 10^{-1}$
C	1.57 ± 0.15	$^{+0.03}_{-0.39}$
$\beta_{\rm app}$	44.6 ± 1.2	$^{+3.8}_{-0.3}$

6.3. Producing a Realistic Lognormal Distribution

Mrk 421 is known to show a lognormal flux distribution from radio to very high energies (Tluczykont et al. 2010; Sinha et al. 2016; Kushwaha et al. 2017).

[6] http://docs.astropy.org/en/stable/stats/lombscargle.html
[7] http://www.astroml.org/modules/generated/astroML.time_series.generate_power_law.html

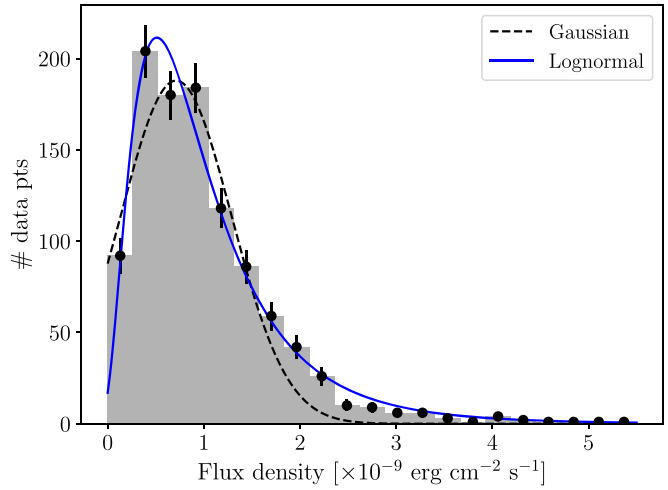

Figure 8. Comparison between Gaussian and lognormal functions fitted to the *Swift*-XRT flux distribution.

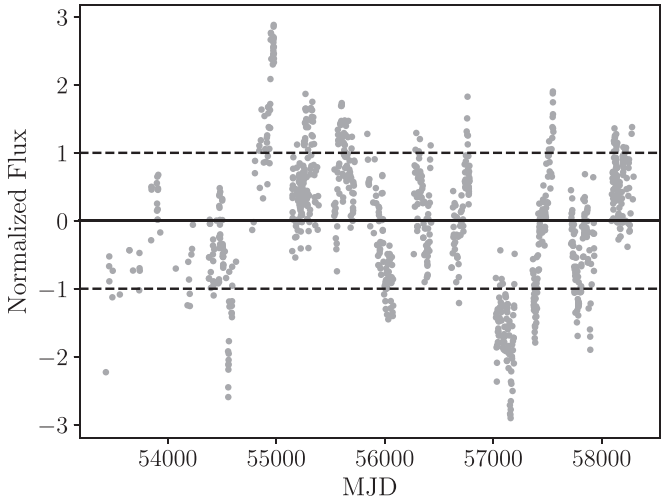

Figure 9. Example of a resampled and normalized simulated light curve. The black solid and dashed lines are the mean and the variance of the data set, respectively.

We confirm this behavior in our *Swift*-XRT data set by testing a lognormal against a normal distribution hypothesis. Both have 19 degrees of freedom. The reduced χ^2 of the lognormal function shows a better fit, with $\chi^2_{\rm red,Lognorm} = 1.58$ and $\chi^2_{\rm red,Gauss} = 4.85$. Assuming a usual p-value acceptance limit of 0.05, the lognormal function is accepted with $P_{\rm Lognorm} = 5.2 \times 10^{-2}$, while the Gaussian assumption is strongly rejected with $P_{\rm Gauss} = 1.3 \times 10^{-11}$. These two fits are shown in Figure 8.

Before adjusting the simulated light curves to the one with a realistic distribution, we need to normalize their variance $V(\Phi(t))$ to 1 and mean value $\langle\Phi(t)\rangle$ to zero. An example of such a resampled and normalized light curve, $\Phi_{\rm sim,norm}(t)$, is given in Figure 9.

Then, the distribution can be transformed to lognormal following the equation

$$\Phi_{\rm sim,LN}(t) = \exp((\Phi_{\rm sim,norm}(t) \times a) + b), \tag{11}$$

with $a = \sigma_{\rm sim}$ and $b = \mu_{\rm sim}$ of the normally distributed logarithm $\log(\Phi_{\rm sim,LN}(t))$. This comes from the fact that the mean and variance of $\Phi_{\rm sim,norm}(t)$ are zero and 1, respectively.

These two parameters, a and b, can be observationally constrained considering that observed and simulated light curves should have similar mean values, as well as similar variability amplitudes $F_{\rm var}$.

The variability amplitude, as defined by Rodríguez-Pascual et al. (1997), is expressed as

$$F_{\rm var} = \frac{\sqrt{V(\Phi(t)) - \Delta^2}}{\langle\Phi(t)\rangle}, \tag{12}$$

with Δ^2 the mean square value of the uncertainties. At this point, the simulated data set does not yet have associated uncertainties, so $F_{\rm var,sim}$ can be expressed only from the variance and the mean. For a lognormal distribution, they have these forms:

$$\langle\Phi(t)\rangle = \exp(\mu + \sigma^2/2) \tag{13}$$

and

$$V(\Phi(t)) = (e^{\sigma^2} - 1)\langle\Phi(t)\rangle^2. \tag{14}$$

So, $F^2_{\rm var,sim} = e^{\sigma^2_{\rm sim}} - 1$.

Knowing that the coefficient $a = \sigma_{\rm sim}$, and we want similar observed and simulated $F_{\rm var}$, we can write a as

$$a = \sqrt{(\log(F^2_{\rm var,obs}) + 1)}. \tag{15}$$

Then, following Equation (13), and given the assumption of similar observed and simulated mean values $\langle\Phi(t)\rangle$, the coefficient b takes the form

$$b = \langle\Phi_{\rm obs}(t)\rangle - a^2/2. \tag{16}$$

From the amplitude variability $F_{\rm var,obs} = 0.68$, we deduce the values $a = 0.62$ and $b = 3.69$.

Instead of having similar $F_{\rm var}$, one can choose to have similar median values between the observed and simulated $\Phi(t)$. The median value of a lognormal distribution is defined as

$$\text{median}(\Phi(t)) = e^{\mu}. \tag{17}$$

Then, we have the corresponding values of $a = 0.52$ and $b = 3.74$.

Finally, by directly doing a Gaussian fit to $\log(\Phi_{\rm obs}(t))$, we obtain the coefficients $a = 0.61 \pm 0.03$ and $b = 3.79 \pm 0.03$.

We can explain the differences between these three estimations by considering that the *Swift*-XRT light curve does not exactly follow a lognormal distribution, and $F_{\rm var,obs}$ has intrinsic uncertainties (Vaughan et al. 2003).

For the simulated light curves, we consider the middle ground between these three estimations by taking the average values of $a = 0.59$ and $b = 3.74$.

6.4. Simulated Errors

The simulated errors on fluxes should also be realistic. We notice the absence of significant correlation between the *Swift*-XRT fluxes and associated uncertainties, with a Pearson correlation coefficient of $r = 0.0059$ and the p-value $P = 0.85$. Since the simulated light curves have the same number of data points as the original one, we simply associate each of the simulated light curves with the observed uncertainties randomly shuffled. This method ensures the exact same distribution of uncertainties for all simulations. Finally, each point is randomly projected following a normal distribution, with its standard deviation given by the error bar.

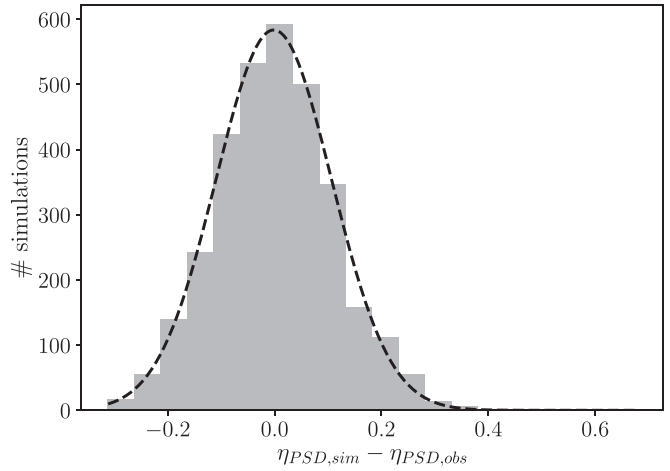

Figure 10. Distribution of the differences between reconstructed and original PSD indexes for 3200 simulated light curves after applying the correction factor of $\Delta\eta_{\rm PSD} = +3.3 \times 10^{-2}$. The distribution has a mean value of $(-1.7 \pm 1.8) \times 10^{-3}$ with a standard deviation of $(1.08 \pm 0.01) \times 10^{-1}$.

6.5. Checking the Simulated Light Curves

After all of the processes described above, the simulated light curves have PSD indexes that differ from the original one. The distribution of the reconstructed PSD index of a large number of simulations ($\eta_{\rm PSD,sim}$–$\eta_{\rm PSD,obs}$) is checked by fitting this distribution with a Gaussian. The resulting mean value of $(-9.6 \pm 1.5) \times 10^{-3}$ highlights a significant bias, about 6σ, that simulated light curves show on average lower PSD indexes.

We correct this bias by iteratively testing various values of $\eta_{\rm PSD}$ used to reconstruct the light curves and stopping the iteration when converging toward a $<1\sigma$ discrepancy, corresponding to a correcting factor of $\Delta\eta_{\rm PSD} = +3.3 \times 10^{-2}$, giving consistent results between observed and simulated indexes, as shown in Figure 10.

The reconstructed light curves being based on Monte Carlo simulations with potential strong alterations due to the resampling process, we perform further checks to ensure that all simulations are realistic enough to be used for our statistical comparison. Simulated light curves are considered good when they have a reconstructed PSD index and a lognormal distribution (μ and σ) within three standard deviations of those of the original. An example of such a simulated light curve passing all of the checks is shown in Figure 11.

6.6. Bias of ToO Observations

As discussed in Section 6.2, the fact that *Swift*-XRT mostly observes Mrk 421 as a ToO introduces a non-even sampling of the data set, which is fully considered in the simulated light curves by the resampling process. However, it induces another bias that cannot be easily simulated. Working in response to a ToO means better sampled observations when a flare is occurring. Following the ToO criteria, denser observations are taken when a flux reaches a given threshold defined by the observers.

This is not the case for simulated light curves, which leads to fewer and weaker flares passing the selection cuts, on average. It has the effect of reducing the data dispersion of the fit residuals in stacked simulated light curves, thus leading to lower reduced χ^2, which biases the statistic test in favor of the simulations.

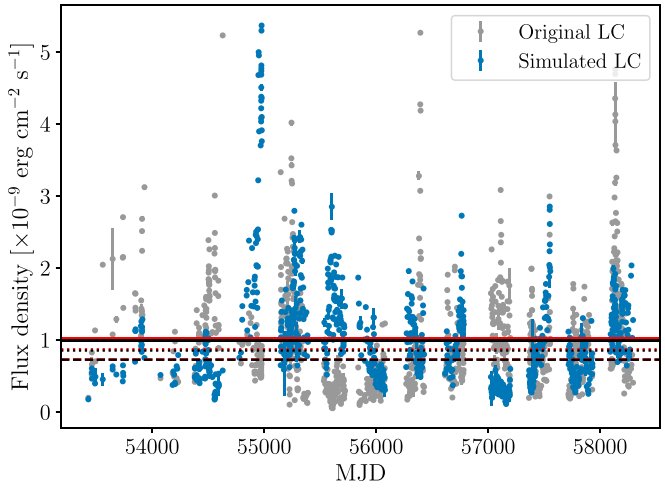

Figure 11. Example of a simulated light curve passing all of the checks, compared to the original one. The solid, dashed, and dotted lines are the mean, median, and standard deviation, respectively. Black lines are for the original light curve, while red lines are for the simulated one.

Table 6
Post-simulation Cuts to Select Light Curves with Enough Selected Flares and Sufficiently High Average Fluxes $M_{F,\rm flares}$ to Ensure a Fair Comparison with the *Swift*-XRT Data Set

	Default	Loose	Hard
Original Data Set			
nb flares	6	13	5
$M_{F,\rm flares}$[a]	3.68	2.74	4.02
Simulations			
Cut (nb flares)	$\geqslant$4	$\geqslant$11	$\geqslant$4
Cut ($M_{F,\rm flares}$)[a]	$\geqslant$3.3	$\geqslant$2.5	$\geqslant$3.5

Note. These cuts are adjusted for the three flare selections described in Section 5.4.
[a] Fluxes in 10^{-9} erg cm^{-2} s^{-1}.

This bias can be taken into account by applying a selection cut on the simulated light curves based on the minimum number and flux average $M_{F,\rm flares}$ of selected flares. By working on a large number of simulations, we adjust these two cuts to produce results as close as possible to those of the original light curve. We do not want the simulations to have a higher number of selected flares and $M_{F,\rm flares}$, on average, which would bias the statistical test in the other way. Keeping these average values slightly below the original ones ensures having a conservative estimate of the probed model significance. These cut values are shown in Table 6.

7. Results and Discussion

7.1. Significance of the Multiple-shock Scenario

From the simulations described in the previous section, we can now provide a fair comparison with the original data set. At the end, only a small portion of the simulated light curves ($\sim$1/10–1/20) are passing all of the cuts to be considered realistic enough for a statistical test. Several million light curves are then produced to have enough statistics. The fraction of simulated light curves $f_{\rm sim}$ having a knot-flare model fit worse than the one of the original data set can be converted to

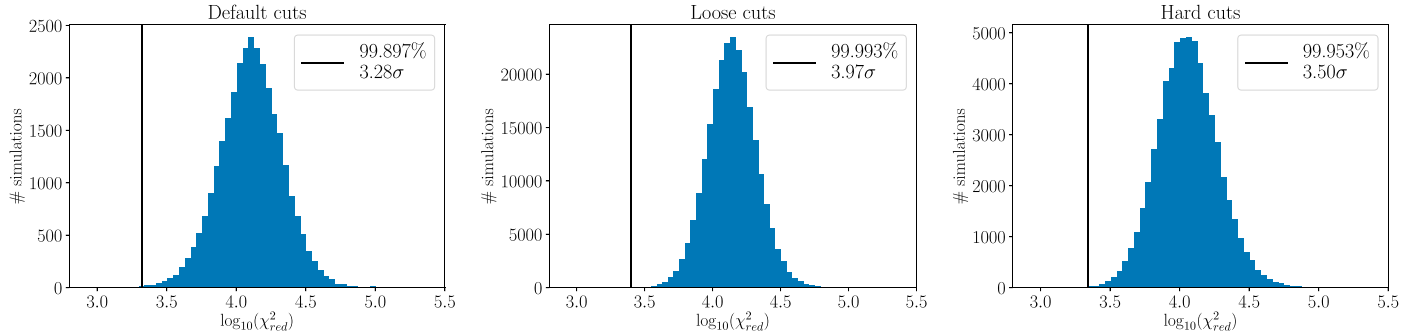

Figure 12. Distribution of the multi-Gaussian fit results ($\chi^2_{\rm red}$) on large samples of simulated light curves, considering the default, loose, and hard flare selection cuts. The black lines are the result for the original *Swift*-XRT light curve. The intrinsic multi-Gaussian post-flare pattern of Mrk 421 is validated above a 3σ level against stochastic fluctuations for all three sets of cuts.

the significance of the intrinsic post-flare pattern result against stochastic fluctuations. We use this expression:

$$\sigma_{\rm result} = {\rm erf}^{-1}(f_{\rm sim})\sqrt{2}\,. \tag{18}$$

Due to the varying number of degrees of freedom in each stack of simulated light curves, the reduced χ^2 is used as an estimator of the fit quality. Also, the post-flare series probed are mostly occurring in a small temporal region of the 640 day stacked light curves. Comparing the $\chi^2_{\rm red}$ on these 640 days would give too much importance to the baseline fit quality instead of the probed post-flare scenario. Hence, we consider the $\chi^2_{\rm red}$ only for the time range between the first and last Gaussian. This time range is defined between the first and last data point where the multi-Gaussian model is 1% above the baseline. The $\chi^2_{\rm red}$ values associated with default, loose, and hard cuts are $\chi^2_{\rm red,Default} = 2.09 \times 10^3$, $\chi^2_{\rm red,Loose} = 2.50 \times 10^3$, and $\chi^2_{\rm red,Hard} = 2.18 \times 10^3$, respectively.

The significances of the knot-flare scenario against a stochastic process from light-curve simulations for the three sets of cuts are between 3.28σ and 3.97σ (see Figure 12). The biggest significance of 3.97σ is found for the loose cuts. The decrease of the fit quality of the *Swift*-XRT data associated with the noise induced by 13 selected flares in the loose cuts is less than the average one of simulations, leading to a better significance than the default cuts with six selected flares. This suggests that the intrinsic post-flare pattern is also present in weaker flares.

7.2. Characterization of the Jet and Perturbation

The deduced apparent flow speed of the VLBI jet of Mrk 421 of $\beta_{\rm app} = 44.6^{+4.0}_{-1.2}$ gives a physical constraint on the maximum angle with the line of sight θ as

$$\theta < 2\arctan(1/\beta_{\rm app}), \tag{19}$$

leading to $\theta \leqslant 2.^\circ 69$ when considering a 90% confidence level limit.

The jet Doppler and Lorentz factors can both be expressed in functions of the apparent speed and the angle with the line of sight following these formulas:

$$\delta = \sqrt{1 - \left(\frac{\sin\theta}{\beta_{\rm app}} + \cos\theta\right)^{-2}}\left(1 + \frac{\beta_{\rm app}}{\tan\theta}\right) \tag{20}$$

and

$$\Gamma = \frac{1}{\sqrt{1 - \left(\frac{\sin\theta}{\beta_{\rm app}} + \cos\theta\right)^{-2}}}. \tag{21}$$

This parameter space can have an additional constraint from the jet opening angle of Mrk 421. Indeed, a canonical relation links the apparent jet full opening angle $\alpha_{\rm app}$ with the Lorentz factor, which can be expressed as

$$\Gamma = \frac{2\rho}{\alpha_{\rm app}\sin\theta}. \tag{22}$$

This equation can be seen as an approximation of relativistic jet gas dynamics, where the Lorentz factor depends on the opening angle and the ratio of pressure between the jet core and the external medium $P_{\rm ext}/P_0$ (Daly & Marscher 1988; Jorstad et al. 2005). The deduced value of $\rho = 0.17 \pm 0.08$ from multiple jet radio VLBI measurements (opening angle, apparent speed, and variability) by Jorstad et al. (2005) leads to $P_{\rm ext}/P_0 \simeq 1/3$, which corresponds to a case where jets naturally form standing recollimation shocks (Daly & Marscher 1988), fully consistent with the probed multiple-shock scenario.

The apparent opening angle can be deduced from the slope ϕ of the linear fit shown in Figure 2 (right) as $\alpha_{\rm app} = 2\arctan(\phi)$. Thus, as shown in Figure 13, the system can be resolved within the parameter ranges $\theta \in$ [0.38–1.8] deg, $\Gamma \in$ [43–66], and $\delta \geqslant 31$.

This Doppler factor lower limit is relatively high compared to previous estimations of Mrk 421 from SED modeling with $\delta \sim$ 20–25 (Katarzyński et al. 2003; Aleksić et al. 2015; Baloković et al. 2016) but consistent with the range of $\delta \in$ [15–35] deduced by Tavecchio et al. (1998) from broadband SED parameterization. We can note that the maximum Doppler value is quite difficult to estimate from SED models due to the known degeneracy between the parameters.

The width of the multiple Gaussian given by Equation (6) provides valuable information to constrain the general features of the perturbation crossing the shocks. In the following, we assume that particle acceleration and cooling times are shorter than the shock crossing time of a perturbation. This assumption implies that the duration of a flare is roughly equal to the duration of the perturbation crossing a shock.

We consider that each Gaussian peak $_{Pi}$ is defined as the convolution product of a Gaussian perturbation P_p crossing a Gaussian shock P_s. The standard deviations can then be

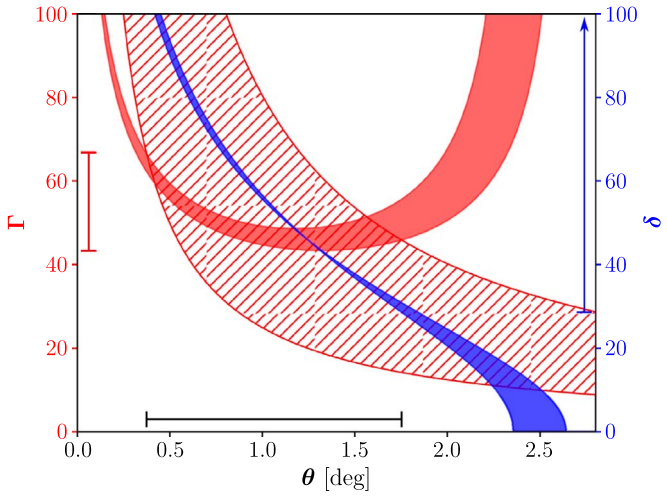

Figure 13. Lorentz factor (red) and Doppler factor (blue) as a function of the angle with the line of sight θ. The plain bands are calculated given the uncertainty on $\beta_{\rm app}$, while the red-hashed band is calculated given the uncertainties on ρ and $\alpha_{\rm app}$. The segments are showing the likely range for each parameter.

written as

$$\sigma_{G,i} = \sqrt{\sigma_p^2 + \sigma_{s,i}^2}. \tag{23}$$

The width of the perturbation, expressed as the Gaussian FWHM, takes the form

$$W_p = 2\sqrt{2\log(2)}\,\sqrt{\sigma_{G,i}^2 - \sigma_{s,i}^2}\,\frac{c\beta_{\rm app}}{1+z}. \tag{24}$$

The shock standard deviation $\sigma_{s,i}$ can be constrained from zero for a perpendicular shock with no width to an upper limit at the size of the radio knots:

$$0 \leqslant \sigma_{s,i} \leqslant \frac{R_i}{\sqrt{2\log(2)}}\,\frac{1+z}{c\beta_{\rm app}}. \tag{25}$$

Then, we can determine the perturbation width from the boundaries on $\sigma_{s,i}$. The first shock gives the strongest constraints, leading to a value of $W_p = 3.9^{+1.5}_{-3.0} \times 10^{17}$ cm, taking into account the uncertainties on the knot measured radius and fit parameters (statistical and systematic). The comoving intrinsic width can be written as

$$W_{p,\rm int} = \frac{W_p}{\Gamma \sin\theta}. \tag{26}$$

Given the values of Γ and θ deduced above, the perturbation intrinsic width lies within the range $W_{p,\rm int} \in [0.43\text{–}19] \times 10^{17}$ cm.

7.3. A New Look at Mrk 421 Emission Scenarios

Mrk 421 is known to present a flux–flux correlation between X-rays and gamma-rays, specifically strong in the very high energy (VHE) regime ($E > 100$ Gev). This correlation has been observed in flares and short-timescale variability (Fossati et al. 2008; Horan et al. 2009; Acciari et al. 2011), as well as in periods of months to years (Acciari et al. 2014). It was also noticed that this correlation even extends to the lowest observed fluxes of Mrk 421 (Baloković et al. 2016). It indicates that gamma-rays and X-rays are coming from the same emission zones, whatever the activity state of the source. In the context of the present study, it means that the flaring gamma-ray emission zones are located inside the radio knots.

Mrk 421 is also known to present strong and fast outbursts in X-rays and gamma-rays, on timescales of ∼15 minutes (Gaidos et al. 1996; Paliya et al. 2015). At first sight, this is not compatible with our scenario, where the size and speed of the perturbation are fitted for about a day-to-day timescale variability. However, we did not consider that these perturbations should naturally be very turbulent environments. Small-scale turbulence crossing a shock is well suited to produce fast flares, as simulated by Marscher (2014).

Prior to this study, the likely possibility of multiple high-energy emission zones in the Mrk 421 jet were discussed in many works (e.g., Błażejowski et al. 2005; Baloković et al. 2016; Carnerero et al. 2017; Kapanadze et al. 2018a, 2018b). While having a general good broadband SED representation, these studies highlighted that the single-zone synchrotron self-Compton (SSC) scenario is strongly challenged by some observed variability patterns and has difficulty modeling the hard TeV spectrum.

Due to the high frequency of the synchrotron peak, the SSC interaction falls into the Klein–Nishina regime at TeV energies, preventing any strong radiation (Fossati et al. 2008). This is a common issue of the so-called "extreme blazars" (EHBLs or UHBLs), including Mrk 421 (Ghisellini 1999). This issue can be resolved if we consider another radiation field in VHE. It encouraged the development of (lepto-)hadronic scenarios, where this additional radiation can be produced by protons (synchrotron or inverse Compton) or secondary particle emission. Several of these models were applied to the study of Mrk 421 (Abdo et al. 2011; Mastichiadis et al. 2013; Zech et al. 2017).

A natural leptonic explanation can, however, be proposed in the framework of the multiple-shock scenario. If we consider that a small fraction of the particles accelerated in the first shock are not fully cooled before reaching other shocks, they will be reaccelerated. Consecutive shocks then have the potential to push the spectra up to the highest energies in AGN, as shown by Meli & Biermann (2013), and can explain an excess in TeV spectra with respect to the one-zone leptonic approach. It is also interesting to note that this spectral issue mostly occurs in HBLs, which were observed to be the most likely sources to have multiple quasi-stationary knots in their jets (Hervet et al. 2016; Piner & Edwards 2018). It then makes HBLs the best candidates for such a particle reacceleration.

As a last point, we can highlight that variability induced by a change of the thermal and nonthermal particle density crossing a shock (or similar to a shock crossing different density regions) was proposed in various studies of Mrk 421. From the evolution of Mrk 421 flares, Fossati et al. (2008) noted that it is "very suggestive of acceleration or injection of the higher energy end of the electron population," as expected by such a multiple-shock reacceleration process. It was also highlighted by Garson et al. (2010) that the variation likely comes from a change of the local density encountered in the shock environs. In this view, radiative shock scenarios (whether single, multizone, semi-analytic, or MHD-based) are promising, such as those of Chen et al. (2011), Moraitis & Mastichiadis (2011), Marscher (2014), Fromm et al. (2016), and Bodo & Tavecchio (2018).

8. Conclusion

In this paper, we show evidence for a possible regular pattern of post-flare variability in Mrk 421. The time delay of the suggested post-flare excesses in the Mrk 421 stacked light curve is consistent with a scenario of the propagation of jet perturbations with roughly similar sizes and constant speeds crossing the multiple stationary VLBI radio knots.

The favored interpretation is a main emission zone in the most upstream VLBI radio knot at 0.38 mas from the core and secondary emission zones from the three other downstream radio knots. This interpretation is preferred at a 3σ level to stochastic fluctuations, as reproduced by numerous realistic simulated light curves.

From our multiple-flare model fitted to the data set, we deduce an apparent speed of the flow of $\beta_{\rm app} = 45^{+4}_{-2}c$. It leads to a jet angle with the line of sight $\theta \in$ [0.38–1.8] deg, associated with a Lorentz factor $\Gamma \in$ [43–66], Doppler factor $\delta \geqslant 31$, and typical intrinsic size of the perturbations crossing the jet $W_{p,\rm int} \in [0.43\text{–}19] \times 10^{17}$ cm. These physical quantities shed new light on the jet physics of Mrk 421 by providing strong constraints that are not based on the usual broadband SED models or from direct observed motions in jets.

The multiple-shock scenario probed brings a natural and simple solution to the blazar bulk Lorentz factor crisis. Stationary radio knots are interpreted as stationary shocks (likely recollimation shocks) and thus are not direct markers of the jet flow speed. The deduced Lorentz and Doppler factors from the multiple-shock scenario are relatively high but not in disagreement with the SSC broadband models and observed fast variability presented in previous studies. We also note that a very recent study performed by Banerjee et al. (2019) on a time-dependent modeling of Mrk 421 in an internal shock scenario leads to beaming parameters of Mrk 421 that are fully consistent with our estimations ($\theta = 1.3$ deg, $\delta \in$ [40–44], $\Gamma \in$ [28–40], and $W_{p,\rm int} = 1.1 \times 10^{17}$ cm). These similar results from a totally independent study and method strengthen the relevance of our approach.

The accuracy of the method can be naturally improved by having long monitoring after strong flares; the larger the data set, the better an intrinsic post-flare pattern can be distinguished. It would also be improved by better radio VLBI monitoring. More radio data will reduce the uncertainties on the size and position of radio knots.

This first study probing a post-flare variability pattern in Mrk 421 has considerable potential to be extended in multiple ways. Given the strong X-ray–VHE correlation of Mrk 421, a natural continuity would be to check this pattern in the VHE light curves of suitable observatories, such as VERITAS, MAGIC, or FACT.

As soon as a blazar is identified with multiple stationary knots and has a multiyear dense monitoring in an energy band associated with a great variability (usually in the energy range of its synchrotron or inverse Compton peaks), it is theoretically possible to perform the same study. Confirming such a pattern in multiple other sources would lead to a great leap forward in our knowledge of AGN jet physics and the origin/location of the high-energy emission zones.

We thank Jonathan Biteau for his helpful suggestions and advice on light-curve simulation and Pranati Modumudi for assistance with analysis of the *Swift*-XRT data. We acknowledge support from the US NASA *Swift*-GI grants NNH17ZDA001N and NNX16AN78G. We also thank the US National Science Foundation for support under grants PHY-1307311 and PHY-1707432. This research has made use of data from the MOJAVE database, which is maintained by the MOJAVE team (Lister et al. 2018).

Software: XSpec (Arnaud 1996), astroML (Vanderplas et al. 2012), Astropy (Astropy Collaboration et al. 2013), SciPy (Jones et al. 2001), NumPy (Walt et al. 2011), Matplotlib (Hunter 2007).

ORCID iDs

Olivier Hervet https://orcid.org/0000-0003-3878-1677
David A. Williams https://orcid.org/0000-0003-2740-9714
Amanpreet Kaur https://orcid.org/0000-0002-0878-1193

References

Abdo, A. A., Ackermann, M., Ajello, M., et al. 2010, ApJ, 722, 520
Abdo, A. A., Ackermann, M., Ajello, M., et al. 2011, ApJ, 736, 131
Abeysekara, A. U., Archambault, S., Archer, A., et al. 2017, ApJ, 834, 2
Abeysekara, A. U., Benbow, W., Bird, R., et al. 2018, ApJ, 856, 95
Abramowski, A., Acero, F., Aharonian, F., et al. 2012, ApJ, 746, 151
Acciari, V. A., Aliu, E., Arlen, T., et al. 2011, ApJ, 738, 25
Acciari, V. A., Arlen, T., Aune, T., et al. 2014, APh, 54, 1
Aleksić, J., Ansoldi, S., Antonelli, L. A., et al. 2015, A&A, 578, A22
Arnaud, K. A. 1996, in ASP Conf. Ser. 101, Astronomical Data Analysis Software and Systems V, ed. G. H. Jacoby & J. Barnes (San Francisco, CA: ASP), 17
Astropy Collaboration, Robitaille, T. P., & Tollerud, E. J. 2013, A&A, 558, A33
Baloković, M., Paneque, D., Madejski, G., et al. 2016, ApJ, 819, 156
Banerjee, B., Joshi, M., Majumdar, P., et al. 2019, MNRAS, in press (arXiv:1905.01043)
Bennett, C. L., Larson, D., Weiland, J. L., & Hinshaw, G. 2014, ApJ, 794, 135
Błażejowski, M., Blaylock, G., Bond, I. H., et al. 2005, ApJ, 630, 130
Bodo, G., & Tavecchio, F. 2018, A&A, 609, A122
Burrows, D. N., Hill, J. E., Nousek, J. A., et al. 2005, SSRv, 120, 165
Carnerero, M. I., Raiteri, C. M., Villata, M., et al. 2017, MNRAS, 472, 3789
Chatterjee, R., Bailyn, C. D., Bonning, E. W., et al. 2012, ApJ, 749, 191
Chen, X., Fossati, G., Liang, E. P., & Böttcher, M. 2011, MNRAS, 416, 2368
Daly, R. A., & Marscher, A. P. 1988, ApJ, 334, 539
Esin, A. A., McClintock, J. E., & Narayan, R. 1997, ApJ, 489, 865
Falle, S. A. E. G. 1991, MNRAS, 250, 581
Fossati, G., Buckley, J. H., Bond, I. H., et al. 2008, ApJ, 677, 906
Fraija, N., Benítez, E., Hiriart, D., et al. 2017, ApJS, 232, 7
Fromm, C. M., Perucho, M., Mimica, P., & Ros, E. 2016, A&A, 588, A101
Fromm, C. M., Perucho, M., Ros, E., et al. 2011, A&A, 531, A95
Gaidos, J. A., Akerlof, C. W., Biller, S., et al. 1996, Natur, 383, 319
Garson, A. B., III, Baring, M. G., & Krawczynski, H. 2010, ApJ, 722, 358
Georganopoulos, M., & Kazanas, D. 2003, ApJL, 594, L27
Ghisellini, G. 1999, APh, 11, 11
Ghisellini, G., Tavecchio, F., & Chiaberge, M. 2005, A&A, 432, 401
Gómez, J. L., Lobanov, A. P., Bruni, G., et al. 2016, ApJ, 817, 96
Gómez, J. L., Martí, J. M., Marscher, A. P., Ibáñez, J. M., & Alberdi, A. 1997, ApJL, 482, L33
Hervet, O., Boisson, C., & Sol, H. 2015, A&A, 578, A69
Hervet, O., Boisson, C., & Sol, H. 2016, A&A, 592, A22
Hervet, O., Meliani, Z., Zech, A., et al. 2017, A&A, 606, A103
Horan, D., Acciari, V. A., Bradbury, S. M., et al. 2009, ApJ, 695, 596
Hunter, J. D. 2007, CSE, 9, 90
Jones, E., Oliphant, T., Peterson, P., et al. 2001, SciPy: Open Source Scientific Tools for Python, http://www.scipy.org/
Jorstad, S. G., Marscher, A. P., Lister, M. L., et al. 2005, AJ, 130, 1418
Jorstad, S. G., Marscher, A. P., Mattox, J. R., et al. 2001, ApJ, 556, 738
Kalberla, P. M. W., Burton, W. B., Hartmann, D., et al. 2005, A&A, 440, 775
Kapanadze, B., Vercellone, S., Romano, P., et al. 2018a, ApJ, 858, 68
Kapanadze, B., Vercellone, S., Romano, P., et al. 2018b, ApJ, 854, 66
Katarzyński, K., Sol, H., & Kus, A. 2003, A&A, 410, 101
Komissarov, S. S., & Falle, S. A. E. G. 1997, MNRAS, 288, 833
Kushwaha, P., Sinha, A., Misra, R., Singh, K. P., & de Gouveia Dal Pino, E. M. 2017, ApJ, 849, 138
Lico, R., Giroletti, M., Orienti, M., et al. 2012, A&A, 545, A117

Lister, M. L., Aller, M. F., Aller, H. D., et al. 2016, AJ, 152, 12
Lister, M. L., Aller, M. F., Aller, H. D., et al. 2018, ApJS, 234, 12
Lyutikov, M., & Lister, M. 2010, ApJ, 722, 197
Manmoto, T., Takeuchi, M., Mineshige, S., Matsumoto, R., & Negoro, H. 1996, ApJL, 464, L135
Marscher, A. P. 2014, ApJ, 780, 87
Marscher, A. P., & Gear, W. K. 1985, ApJ, 298, 114
Marscher, A. P., Jorstad, S. G., D'Arcangelo, F. D., et al. 2008, Natur, 452, 966
Massaro, E., Perri, M., Giommi, P., & Nesci, R. 2004, A&A, 413, 489
Mastichiadis, A., Petropoulou, M., & Dimitrakoudis, S. 2013, MNRAS, 434, 2684
Meli, A., & Biermann, P. L. 2013, A&A, 556, A88
Mimica, P., Aloy, M.-A., Agudo, I., et al. 2009, ApJ, 696, 1142
Mizuno, Y., Gómez, J. L., Nishikawa, K.-I., et al. 2015, ApJ, 809, 38
Moraitis, K., & Mastichiadis, A. 2011, A&A, 525, A40
Nalewajko, K., & Sikora, M. 2009, MNRAS, 392, 1205
Nalewajko, K., Sikora, M., Madejski, G. M., et al. 2012, ApJ, 760, 69
Paliya, V. S., Böttcher, M., Diltz, C., et al. 2015, ApJ, 811, 143
Piner, B. G., & Edwards, P. G. 2005, ApJ, 622, 168
Piner, B. G., & Edwards, P. G. 2018, ApJ, 853, 68
Piner, B. G., Pant, N., & Edwards, P. G. 2010, ApJ, 723, 1150
Piner, B. G., Unwin, S. C., Wehrle, A. E., et al. 1999, ApJ, 525, 176
Raiteri, C. M., Villata, M., Bruschini, L., et al. 2010, A&A, 524, A43
Rodríguez-Pascual, P. M., Alloin, D., Clavel, J., et al. 1997, ApJS, 110, 9
Romano, P., Campana, S., Chincarini, G., et al. 2006, A&A, 456, 917
Sinha, A., Shukla, A., Saha, L., et al. 2016, A&A, 591, A83
Sironi, L., & Spitkovsky, A. 2014, ApJL, 783, L21
Spada, M., Ghisellini, G., Lazzati, D., & Celotti, A. 2001, MNRAS, 325, 1559
Stroh, M. C., & Falcone, A. D. 2013, ApJS, 207, 28
Tavecchio, F., Becerra-Gonzalez, J., Ghisellini, G., et al. 2011, A&A, 534, A86
Tavecchio, F., Maraschi, L., & Ghisellini, G. 1998, ApJ, 509, 608
Timmer, J., & Koenig, M. 1995, A&A, 300, 707
Tluczykont, M., Bernardini, E., Satalecka, K., et al. 2010, A&A, 524, A48
Türler, M. 2011, Mem. Soc. Astron. Italiana, 82, 104
Türler, M., Courvoisier, T. J.-L., & Paltani, S. 2000, A&A, 361, 850
Uttley, P., McHardy, I. M., & Papadakis, I. E. 2002, MNRAS, 332, 231
van Putten, M. H. P. M. 1996, ApJL, 467, L57
Vanderplas, J., Connolly, A., Ivezić, Ž., & Gray, A. 2012, in Conf. on Intelligent Data Understanding (CIDU) (New York: IEEE), 47
Vaughan, S., Edelson, R., Warwick, R. S., & Uttley, P. 2003, MNRAS, 345, 1271
Walt, S. V. D., Colbert, S. C., & Varoquaux, G. 2011, CSE, 13, 22
Woo, J.-H., Urry, C. M., van der Marel, R. P., Lira, P., & Maza, J. 2005, ApJ, 631, 762
Zech, A., Cerruti, M., & Mazin, D. 2017, A&A, 602, A25