A machine learning approach to distinguish different subdiffusive dynamics in particle tracking

G. Nardi[1,3], M. Santos Sano[1], A. Brelot[2], J.-C. Olivo-Marin[1], T. Lagache[1]

Abstract

This paper presents a novel supervised features-based learning method to classify particle dynamics in biological imaging. To this goal, we consider geometric features computed on trajectories and encoding their intrinsic geometrical characteristics. The method is validated on a dataset simulating different processes: Brownian motion, directed Brownian motion, Orsntein-Ulhenbeck process, fractional Brownian motion, and Continuous-Time Random Walk. The presented approach allows for distinguishing these five dynamical behaviors in a unified framework with high accuracy, and its strength lies in distinguishing several subdiffusive dynamics from free or superdiffusive ones. The robustness to image noise and trajectory length variation is proven, showing the flexibility and reliability of the method in terms of variability due to acquisition techniques. Distinguishing different subdiffusive behaviors strongly impacts particle analysis in biology, and an application is shown to the motion classification for receptors (CCR5) at the cell membrane.

1 Introduction

In the field of biological imaging, the analysis of sub-cellular processes is generally achieved through real-time fluorescence imaging of particles of interest (for instance, molecules, viruses, or organelles) and their automatic tracking over time. To deal with imaging noise and the stochastic motion of imaged particles, elaborate tracking algorithms have been developed over the past two decades [1]. These algorithms automatically detect fluorescent spots in the various images of the sequence and then link the detections into coherent trajectories. The type of stochastic motion associated with the tracked particle is a highly instructive feature to take into account, as it reflects certain biophysical principles underlying the observed biological process. For example, a particle diffusing freely without obstacle or chemical interaction will follow a standard Brownian motion, while chemical binding or trapping in cellular microdomains will lead to *subdiffusive* motion. In contrast, the active transport of biological particles is characterized by *superdiffusive* dynamics.

The classical approach to motion classification allows for the distinction of three diffusion classes: local and confined (*subdiffusive*), free (coinciding with Brownian motion (BM)), and directed and propagated (*superdiffusive*). The standard method to perform this three-class classification is based on the mean square displacement (MSD) function, describing the mean displacement of the particle as a function of the time interval. The MSD method is based on the seminal work [2] proving the linear dependence of the MSD on time for BM. Then, as initially proposed in [3, 4], a polynomial fitting of the empirical MSD function allows for distinguishing three classes (subdiffusive, free, superdiffusive) based on a sublinear, linear, or superlinear (respectively) dependence of the MSD function. However, the MSD criterion has many drawbacks. Accurate diffusion estimation requires very long trajectories, which are difficult to obtain in biological applications, because of particle internalization, or crowded environments preventing reliable tracking over time.

In addition, to distinguish free movements based on the linearity of the MSD function, a confidence interval around the fitting power coefficient must be defined introducing an arbitrary parameter in the statistical analysis.

This has encouraged the introduction of other analytical features to characterize the free motion (BM) of the tracked particles, such as the radius of gyration [5] or the evolution in a bounded domain [6]. [7] analyses features related to the Gaussianity of displacement distribution (moments, trajectory self-similarity, and directional persistence) for diffusion classification. Moreover, to go beyond the MSD criterion and robustly characterize BM, statistical hypothesis tests have been introduced based on specific trajectories features, such as the standardized longest distance traveled by a particle (from its starting point, in a given time interval) [8, 9, 10]. In [11] those hypothesis tests are compared to MSD fitting coefficient and p-variation.

Although these approaches facilitate a more precise analysis of motion diffusion, they do not allow for characterizing the different types of subdiffusive dynamics, a key point for a finer analysis of biological processes. Indeed, several subdiffusive dynamics can coexist at the subcellular scale, with intrinsically different behaviors corresponding to different environmental constraints. The class of subdiffusive motion

[1]Institut Pasteur, Universite de Paris Cité, CNRS UMR 3691, BioImage Analysis Unit, Paris, France
[2]Institut Pasteur, Université de Paris Cité, CNRS UMR 3691, Dynamics of Host-Pathogen Interactions Unit, Paris, France
[3]Corresponding author: giacomo.nardi@pasteur.fr

gathers different families of motion corresponding to different interactions with the related environment and ways to unfurl the space [12]. A more detailed description of subdiffusive movements is, therefore, a major challenge for classification algorithms.

Three main types of subdiffusive processes modeling different biophysical mechanisms are generally used: the Orsntein-Ulhenbeck process (OU) describes confined motion due to the attraction towards an equilibrium point; the fractional Brownian motion (FBM) models motion in a constrained or crowded environment based on its non-independent successive displacements; the Continuous-Time Random Walk (CTRW) describes the motion of particles trapped by an obstacle over some time interval.
Previous work has been devoted to devising suitable features for characterizing subdiffusive processes. For example, [11] used p-variation to distinguish between FBM and CTRW. Various features proposed in the literature, including quantifiers for ergodicity and stationarity, are reviewed in [13], where a theoretical decision tree is also proposed to distinguish subdiffusive motion and fractal constraints. Numerous machine learning approaches that combine different trajectory characteristics to classify subdiffusive movements automatically have recently been developed. In [14], statistical properties of trajectories (Asymmetry and Gaussianity of displacement distribution, fractal dimension, trappedness) are added to MSD, within a supervised ML method for nanoparticle dynamics classification. In [15], an ML approach is developed based on the features proposed in [11] for dynamically classifying G protein-coupled receptors and G proteins. In [16], a supervised ML method is trained using the trajectory itself (as a set of points) to distinguish normal or anomalous (sub- or super-diffusion). However, these approaches follow the standard three-class diffusion classification framework [15] or allow distinguishing a set of specific dynamics [11, 14, 16].

Deep-learning approaches have been developed in parallel to alleviate the need for the user-defined selection of trajectory features. [17] proposes a method based on neural networks to predict the anomalous exponent (MSD fitting coefficient) from single trajectories. [18] develops a deep-learning method to distinguish BM, FBM, and CTRW behavior. Although achieving a higher accuracy than features-based methods, deep-learning approaches, especially on raw trajectories, suffer from standard interpretability issues limiting the understanding of studied phenomena.

The proposed method develops a five-class supervised learning method to classify the previous subdiffusive motion types (OU, FBM and CTRW), standard BM and superdiffusive directed motion (DR). To our knowledge, this is the first method proposed in the related literature, enabling the detection of five motion types in a unified framework. To ensure the explicability of classification, the method is based on geometric features. Four features are collected for every trajectory and a supervised learning method is trained on them. The considered features account for directionality, histogram of angles between successive position triplets, and Ripley's indices on growing balls around the starting points of trajectories. The method has an overall accuracy of 93.6% on trajectories of length 100 (time interval equal to 1/30), meaning that the features reflect the intrinsic geometric properties of previous processes. In particular, beyond the problem of process classification, this enables the method to describe the different ways particles deploy in space. Finally, robustness to noise and length variability is proven. This ensures the method's reliability in biological applications where estimating the noise level and collecting paths of the same length is often difficult. The robust and explainable characterization of subdiffusive behaviors improves the analysis of related environmental constraints on tracked particles under different conditions. This is particularly useful for trajectory analysis in biological imaging, which has as its main goal the description of protein dynamics under different conditions [19, 9, 8]. After validation with synthetic data, we assessed our method on a biological dataset of cell receptor (CCR5) trajectories obtained from fluorescence microscopy videos via a tracking algorithm. We compare the basal state to the PSC-RANTES treatment (inhibiting HIV-1 infection), proving that the latter strongly impacts the receptors' dynamics.

The paper is organized as follows. Section 2 presents the main properties of the five processes considered in this work (BM, DIR, OU, FBM, CTRW). Section 3 presents the geometrical features used to set the supervised learning method. Section 4 describes the proposed method on a simulated dataset composed of trajectories following the five processes considered in this work; the impact of noise and trajectory length variability is studied, and a comparison is performed with a statistical test approach. Application of the method to the classification of cellular receptors (CCR5) on real data is reported in Section 4.6. Finally, Section 5 is dedicated to discussing results and possible improvements.

2 Stochastic processes and their classification

This section recalls stochastic processes' main definitions and properties, presenting their main classification into subdiffusive, free (Brownian), and superdiffusive. We discuss the standard classification method, based on the mean square displacement criterion, and its drawbacks motivating the introduction of alternative approaches discussed in the Introduction. Finally, the rest of the section presents the main stochastic processes widely used to model particle trajectories following different dynamics. A more extensive presentation of these processes is given in the Appendix, intended to serve as a reference for their mathematical properties.

2.1 Standard classification of stochastic processes

General definitions. A stochastic process is a one-parameter family X_t of $\mathbb{R}^2$-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$:

$$X_t \,:\, (t,\omega) \in \mathbb{R}^+ \times \Omega \mapsto X_t(\omega) \in \mathbb{R}^2 \,.$$

We consider the two coordinates of X_t independent random variables in particle tracking. The parameter t corresponds to the time and, for every $\omega \in \Omega$, the related trajectory is defined by the application :

$$t \in \mathbb{R}^+ \mapsto X_t(\omega) \in \mathbb{R}^2 \,.$$

The statistical properties of increments determine the main properties of stochastic processes. A process is said to have *stationary increments* if increments distribution only depends on the time interval and not on the current times (e.g., $X_t - X_s$ and $X_{t-s} - X_0$ follow the same probability distribution for every $s < t \in \mathbb{R}^+$). Similarly, a process is said to be *stationary* if the distribution of X_t does not depend on t. On the other hand, we say that a process has *independent increments* if $X_t - X_s$ and $X_s - X_w$ are independent random variables for every $w < s < t \in \mathbb{R}^+$, implying that each displacement is not correlated to previous realizations. We consider Gaussian processes in the following, meaning the increments follow a Gaussian distribution. Finally, we recall that a process is said to be *continuous* if it has continuous trajectories $t \mapsto X_t$ with probability one on Ω :

$$\mathbb{P}(\{\omega \in \Omega \,|\, \lim_{s \to t} |X_s(\omega) - X_t(\omega)| = 0\}) = 1 \,.$$

The discrete version of a stochastic process can be defined by sampling the continuous paths on a finite set of times $t_0, ..., t_N \in \mathbb{R}^+$, that defines a *discrete-time stochastic process*. A discrete trajectory is defined as a set of successive positions over time

$$\mathbf{X} = (X_{t_1}, ..., X_{t_N}) \,,$$

where $X_i \in \mathbb{R}^2$ with independent components and the time interval between successive positions is constant.

Mean square displacement criterion. The mean squared displacement (MSD) function estimates the relationship between the increments average and the related time interval. Its definition involves the ensemble average computed using the mean in the probability space [12]:

$$\langle X_t^2 \rangle_{ens} \,=\, \mathbb{E}(\|X_{t+\Delta t} - X_t\|^2) \,=\, \int \|X_t - X_0\|^2 \, P(x,t) dx$$

where $P(x,t)$ is the probability to find the particle at position x at time t.

This definition is adapted to study systems exhibiting independent realizations of the same stochastic process. The ensemble average estimates the mean displacement related to Δt of a set of particles. However, in many biological contexts, several dynamical behaviors can co-exist within the observed trajectories, and classification must be performed for single paths. Then, a time-average version of the MSD function is used, quantifying increments average as a function of the time interval for a given trajectory X_t:

$$\langle X_t^2 \rangle_{T,\Delta t} \,=\, \frac{1}{T - \Delta t} \int_0^{T-\Delta t} \|X_{\tau+\Delta t} - X_\tau\|^2 \, d\tau \,. \tag{2.1}$$

For computational applications (2.1) can be discretized as follows :

$$\mathrm{MSD}_{time}(X_t, \Delta t) = \frac{1}{N - \Delta t + 1} \sum_{k=1}^{N-\Delta t} \|X_{t_k+\Delta t} - X_{t_k}\|^2 \,, \tag{2.2}$$

where N is the number of positions belonging to the trajectory $\mathbf{X}$ and Δt is an integer.

In single-particle analysis, trajectories are classified using time-average mean square displacements via a polynomial fitting of the MSD function (2.2) to $t \mapsto t^\alpha$. This enables the definition of the following MSD criterion [3, 4] for diffusion classification :

$$\text{Diffusion class} = \begin{cases} \text{subdiffusive}, & \text{if } \alpha < 1 \\ \text{free motion}, & \text{if } \alpha = 1 \\ \text{superdiffusive}, & \text{if } \alpha > 1 \end{cases} \tag{2.3}$$

This criterion is used to detect confined, random, and directed motions (respectively), but this is often an inadequate and limited framework for many biological applications, in particular, because of its main drawbacks.

First, a precise estimate of the MSD coefficient needs very long trajectories, which are often unavailable in single-particle tracking. Moreover, a confidence interval is needed to estimate the fitting coefficient α in the case of Brownian motion.

Secondly, in the case $\alpha < 1$, the MSD criterion classifies every local and restricted trajectory as subdiffusive but does not give any information about its intrinsic behavior. As detailed below, subdiffusive processes can reveal different dynamics, such as trapping or constraints, and a more precise classification of subdiffusive behaviors is required. Moreover, beyond identifying the process governing the trajectory, estimating trajectory geometrical parameters can also provide useful information on how the particle unfurls in space (for instance, privileged directions, stopping times, or spreading radius).

The last critical point about the MSD criterion concerns its consistency with ergodicity. In ergodic dynamical systems, particles visit every point of the space uniformly and randomly. This means that the statistical properties of the system can be estimated from typical trajectories rather than ensemble averages. The famous ergodic theorem [20, 21] states that for an ergodic process, $\langle X_t^2 \rangle_{ens}$ and $\langle X_t^2 \rangle_{T,\Delta t}$ exhibit the same functional dependence on t and Δ, respectively. This implies that the MSD criterion based on time-average gives a result consistent with the ensemble-average estimate. However, for non-ergodic processes, time and ensemble-average MSD can exhibit different dependencies on time, making the MSD criterion unreliable. As detailed below, this is the case of continuous-time random walk (CTRW), revealing that the MSD criterion can be inadequate to detect non-ergodic behavior, even for long trajectories.

2.2 Main stochastic processes

This section briefly presents the main stochastic processes studied in this work. Brownian motion (BM) and Directed Brownian motion (DIR) are the standard free and directed diffusion models, respectively. As purely subdiffusive process, we consider the Ornstein-Uhlenbeck process (OU), and the Continuous-Time Random Walk (CTRW), describing confinement and trapping, respectively. Finally, we also consider the fractional Brownian motion (FBM), which can exhibit all types of diffusion depending on its Hurst parameter. This is a very flexible model used in this work to model constrained subdiffusive and superdiffusive motions.

In the following, we expose the main properties of each process, and several examples corresponding to different behaviors are shown in Figure 1.

Brownian motion (BM). Brownian motion, sometimes free or normal, is a continuous Gaussian stochastic process historically used to model random motion [2]. It is denoted by B_t, with stationary and independent increments verifying :

$$(B_t - B_s) \sim \mathcal{N}(0, (t-s)\mathbf{I}_2)$$

for every $t > s$, where $\mathbf{I}_2$ denotes the 2-dimensional identity matrix. However, this process is not stationary as B_t follows a Gaussian distribution with a variance depending on t. Setting $X_t = \sigma B_t$ we can vary the displacement amplitude by a parameter σ, and we get

$$\langle X_t^2 \rangle_{ens} = 2\sigma^2 t\,, \quad \langle X_t^2 \rangle_{T,\Delta t} = 2\sigma^2 \Delta t$$

confirming the well-known Einstein's result [2] that justifies the introduction of MSD criterion. In the related literature, the MSD for BM in dimension two is written as $\langle X_t^2 \rangle_{ens} = \langle X_t^2 \rangle_{T,\Delta t} = 4Dt$ where D is called the diffusion coefficient. For $X_t = \sigma B_t$, we have the relationship $\sigma^2 = 2D$ and, with an abuse of language, σ is also often called diffusion coefficient.

Superdiffusion (DIR). The main model for super-diffusion is the so-called Directed Brownian or Directed motion, which verifies

$$dX_t = \mu dt + \sigma dB_t$$

where the drift component $\mu \in \mathbb{R}^2$ gives a constantly oriented input to the motion. This model describes particles driven by active motors [22], and depending on the ratio $\|\mu\|/\sigma$, the trajectory will have a more linear ($\|\mu\| >> \sigma$) or Brownian ($\|\mu\| << \sigma$) behavior. Like Brownian motion, X_t is a non-stationary continuous Gaussian process with stationary and independent increments.

As the mean of B_t equals zero, a direct computation gives

$$\langle X_t^2 \rangle_{ens} = \|\mu\|^2 t^2 + 2\sigma^2 t\,, \quad \langle X_t^2 \rangle_{T,\Delta t} = \|\mu\|^2 (\Delta t)^2 + 2\sigma^2 \Delta t$$

showing the superdiffusive behavior of X_t according to the MSD criterion.

Ornstein-Uhlenbeck process (OU). The continuous-time OU process [23] verifies

$$dX_t = -\lambda(X_t - \theta)dt + \sigma dB_t \tag{2.4}$$

where θ is the equilibrium point, σ the diffusion coefficient of B_t, and λ represent the drift term. In the following, up to the change of variables $X_t = X_t - \theta$, we consider $\theta = (0,0)$. Assuming that stochastic processes have independent components, the general results for dimension one apply. This process suits

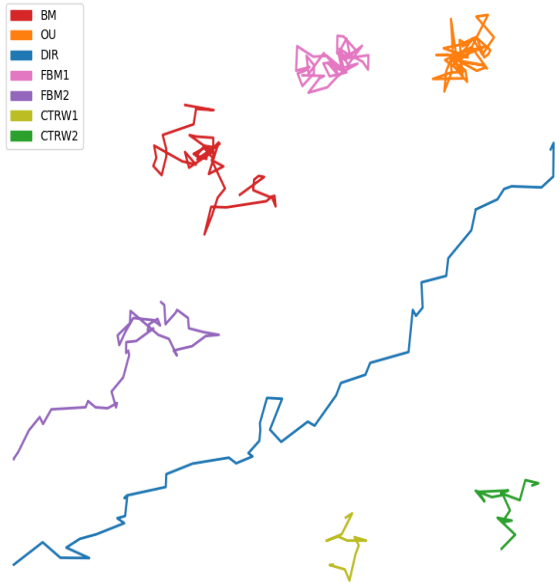


Figure 1: Examples of different types of motions ($N = 50$, $\Delta t = 1$) : BM ($\sigma = 1$), OU ($\lambda = 0.5$, $\sigma = 1$), DIR ($\|u\| = 0.7$, $\sigma = 1$), FBM (FBM1 with $H = 0.2$, and FBM2 with $H = 0.8$), CTRW ($\sigma = 1$, $\gamma = 0.01$ for CTRW1 and $\gamma = 0.9$ for CTRW2.

particles with limited mobility due to an external force attracting them toward an equilibrium point [24]. The previous equation can be solved by variation of constants, obtaining

$$X_t = X_0 e^{-\lambda t} + \sigma \int_0^t e^{-\lambda(t-s)}\, dB_s \tag{2.5}$$

where X_0 is the initial position and the integral term on right side follows the normal distribution $\mathcal{N}(0, \frac{\sigma^2}{2\lambda}(1 - e^{-2\lambda t})\mathbf{I}_2)$. The covariance function [25] is

$$\mathrm{Cov}(X_t, X_s) = \left[\left(\mathrm{Var}(X_0) - \frac{\sigma^2}{2\lambda}\right) e^{-\lambda(t+s)} + \frac{\sigma^2}{2\lambda} e^{-\lambda|t-s|}\right] \mathbf{I}_2 \tag{2.6}$$

showing that the process properties strongly depend on the statistical properties of X_0 (Var denotes the variance of a process).

For example, when assuming that $X_0 \sim \mathcal{N}(0, \frac{\sigma^2}{2\lambda}\mathbf{I}_2)$ is independent of B_t, we get that $X_t \sim \mathcal{N}(0, \frac{\sigma^2}{2\lambda}\mathbf{I}_2)$, proving that X_t is a continuous Gaussian stationary process. The increments are also normally distributed and

$$\mathrm{Cov}(X_t - X_s, X_s) = -\frac{\sigma^2}{2\lambda}(1 - e^{-\lambda(t-s)})\mathbf{I}_2, \quad \forall t, s \in \mathbb{R}^+,\ t > s\,. \tag{2.7}$$

(2.7) proves the increments are stationary but not independent because they are negatively correlated.

Otherwise, X_t is still a continuous Gaussian process with dependent increments, but (2.6) implies that it is not stationary ($\mathrm{Var}(X_t) \propto e^{-2\lambda t}$) and its increments are stationary for t and s significantly larger than $1/\lambda$. This situation corresponds to paths with the same starting point X_0, as in the simulations performed in this work.

In both cases, we note that the variance of X_t is bounded, confirming the OU process's confined behavior: if $\lambda << \sigma$, the motion will present a Brownian-like behavior, while, if $\lambda >> \sigma$, we obtain more confined trajectories.

Finally, by a straightforward computation and using the independence of X_0 and B_t, we obtain

$$\langle X_t^2 \rangle_{ens} = 2\mathrm{Var}(X_0)(1 - e^{-\lambda t})^2 + \frac{\sigma^2}{\lambda}(1 - e^{-2\lambda t})\,,$$

$$\langle X_t^2 \rangle_{T,\Delta t} = \frac{2\sigma^2}{\lambda}(1 - e^{-\lambda \Delta t}) + 2\left(\mathrm{Var}(X_0) - \frac{\sigma^2}{2\lambda}\right)(1 - e^{-\lambda \Delta t})^2 \frac{1 - e^{-2\lambda(T - \Delta t)}}{2\lambda(T - \Delta t)}\,,$$

confirming the OU process's subdiffusive and ergodic behavior [25].

Fractional Brownian Motion (FBM). It is a generalization of Brownian motion characterized by Gaussian increments that are stationary but not independent. This enables the modeling of particles moving in constrained or crowded environments [26].

Formally [27, 28, 29], a fractional Brownian motion is a continuous Gaussian process with increments verifying

$$(X_t - X_s) \sim \mathcal{N}(0, \sigma^2 |t-s|^{2H} \mathbf{I}_2)$$

where $H \in (0,1)$ is the Hurst parameter which determines the intrinsic nature of this process. In particular, for $H = 1/2$ the fractional Brownian motion reduces to Brownian motion.

For $H \neq 1/2$, as the variance of increments depends only on the time interval and not on the current time, the process has stationary increments. However, as the variance of $X_t - X_0$ depends on t, it is not stationary. Moreover, for $t, s \in \mathbb{R}^+$ with $t > s$, it holds

$$\mathrm{Cov}(X_t, X_s) = \frac{\sigma^2}{2}[t^{2H} + s^{2H} - |t-s|^{2H}], \quad \mathbb{E}[(X_t - X_s)X_s] = \frac{\sigma^2}{2}[t^{2H} - s^{2H} - |t-s|^{2H}]$$

showing that the increments are not independent and that the correlation of successive increments depends on H (positive for $H > 1/2$, negative for $H < 1/2$).

Finally, similarly to B_t, we obtain

$$\langle X_t^2 \rangle_{ens} = 2\sigma^2 t^{2H}, \quad \langle X_t^2 \rangle_{T,\Delta t} = 2\sigma^2 (\Delta t)^{2H}$$

which points out the ergodicity of the process and the variety of dynamics represented by fractional Brownian motion: superdiffusive for $H > 1/2$, and subdiffusive for $H < 1/2$.

Continuous-Time Random Walk (CTRW). This process has been introduced [30] to describe trapping for random walks, exhibiting trajectories alternating jumps and waiting times; after every jump, the trajectory maintains the same position for a duration equal to the related waiting time [31, 32].

Denoting by $N(t)$ the number of jumps up to time t and by the $\{t_i\}_{i=0}^{N(t)}$ jumps times, the position of a particle at time t is given by:

$$X_t = \sum_{i=1}^{N(t)} \xi_i$$

where ξ_i denotes the random jump and $\tau_i = t_i - t_{i-1}$ defines the waiting time between the consecutive jumps. We here assume that $t_0 = 0$ and $X_{t_0} = 0$. Jumps $\{\xi_i\}_i$ are independent and identically distributed (iid) random variables following a Gaussian distribution $\mathcal{N}(0, \sigma^2 \mathbf{I}_2)$, while the waiting times $\{\tau_i\}_i$ are iid random variables following a power law distribution on $\psi(t) \sim \frac{\gamma}{\Gamma(1-\gamma)} t^{-\gamma-1}$ with $\gamma \in [0,1]$ and $t \geq 1$ (see [31]). Finally, we assume that the families $\{\xi_i\}_i$ and $\{\tau_i\}_i$ are independent (uncoupled CTRW).

We note that, for large γ, large waiting times are less likely to occur, and CTRW is more similar to Brownian motion.

CTRW is a Gaussian process with independent increments, as X_t is the sum of Gaussian independent random variables. However, the variance of increments explicitly depends on the current time (see Chapter 4 in [31]), and it holds

$$\mathbb{E}[(X_{t+\Delta t} - X_t)^2] = \sigma^2 \mathbb{E}[N(t+\Delta t) - N(t)] \propto [(t+\Delta t)^\gamma - t^\gamma] \tag{2.8}$$

proving that the increments are non-stationary; for the same reason, the process is not stationary either. Furthermore, because of jumps, a CTRW path is not continuous.

Finally, CTRW exhibits the following non-ergodic property [31, 33, 34, 35]:

$$\langle X_t^2 \rangle_{ens} \propto t^\alpha, \quad \langle X_t^2 \rangle_{T,\Delta t} \propto \Delta t\,.$$

Ensemble and time averages point out different behaviors: the former indicates a subdiffusive trajectory, while the latter suggests a normal diffusion.

MSD criterion is then unreliable for CTRW trajectories, classifying them as Brownian motion. Therefore, CTRW is considered a subdiffusive process based on its qualitative dynamic properties rather than the MSD criterion.

Finally, we point out that the properties of CTRW are strongly correlated to the distribution of waiting times. For example, considering an exponential distribution $\psi(t) = e^{-t}$, the process remains no stationary, but its increments become stationary, and the ergodic property holds. This is because, in this case, $\mathbb{E}[N(t)] \cong t$ implies a linear dependence on Δt in (2.8) (see Section 3.3 in [31]).

However, the power-law distribution for waiting times is more realistic for biophysical modeling. This is due to its diverging first moment ($\mathbb{E}[N(t)] = \infty$), implying its subdiffusive and non-ergodic behavior. For these reasons, power-law distribution is preferred to exponential one for biophysical studies.

Process	Continuous	Gaussian	Stationary	Stationary Increments	Independent Increments	Ergodic
BM, DIR	×	×		×	×	×
OU	×	×	(×)	(×)		×
FBM	×	×		×		×
CTRW		×		(×)	×	(×)

Table 1: List of main properties for each stochastic process. Stationarity properties for OU process depend on the initial state, and brackets hold if $X_0 \sim \mathcal{N}(0, \frac{\sigma^2}{2\lambda}\mathbf{I}_2)$; FBM refers to the case $H \neq 1/2$; CTRW properties depend on the distribution of waiting times, brackets hold for exponential distribution and do not for a power-law distribution.

3 Geometrical features

This section defines the main features used in the proposed approach, which are computed on the entire trajectory to summarize its intrinsic geometric properties. The considered features belong to two main families.

The first set of features accounts for directionality, based on analyzing the distribution of angles between successive positions. Several works have already used directionality analysis [7] to characterize the trajectory's persistent, free, or antipersistent dynamic. This helps, in particular, to highlight the recall or propagation behavior characterizing OU or FBM ($H < 1/2$) and DIR or FBM ($H > 1/2$), respectively. On the other hand, a uniform-like angle distribution is associated with BM. Finally, CTRW shows a Dirac-like angle histogram due to its characteristic stopping times.

The second set of features is based on Ripley's ratio on concentric balls to describe how the particle spreads in space. To our knowledge, this work is the first to use these features for motion classification. Considering the ratio of trajectory points leaving in each ball, we can point out local or global (in time) confined behavior. This is particularly important to highlight high propagation and strong confinement, which characterize DIR and OU, respectively.

The results discussed in Section 4 show that these features are sufficient to characterize the intrinsic properties of the processes previously discussed and enable the detection of different types of motion beyond their diffusion classification.

We recall that a discrete trajectory is a set of N positions $X_i \in \mathbb{R}^2$ corresponding to different times $t_1 < ... < t_N \in \mathbb{R}^+$:

$$\mathbf{X} = (X_{t_1}, ..., X_{t_N})\,.$$

In particular, we consider a constant time interval between successive positions, and we suppose that the components of each position are independent.

The directionality of a movement can be studied by analyzing the turning angle of a displacement compared to the previous one [7]. For every three consecutive points of the trajectory $(X_{t_i}, X_{t_{i+1}}, X_{t_{i+2}})$, we consider the angle

$$\theta_{\mathbf{X}}(t_i) = \pi - \angle X_{t_i} X_{t_{i+1}} X_{t_{i+2}}\,,$$

that is considered as positive if $X_{t_{i+1}} - X_{t_i}$ rotates on $X_{t_{i+2}} - X_{t_i}$ in a counterclockwise way. In the case of CTRW, most angles are null since the particle often remains stationary at the same point. Therefore, non-null angles are computed between triplets of not necessarily consecutive displacements.

The histogram of angles $\{\theta_{\mathbf{X}}\}$ defines the related probability density $p_{\theta_{\mathbf{X}}}$, which exhibits a different analytical dependence for different processes (Figure 2). Brownian motion has a uniform distribution of angles, while directed motion exhibits a Gaussian distribution centered at zero, corresponding to a motion with drift. The largest angles are the most probable for OU trajectories, revealing a confined and backward recall dynamic. These examples suggest, in particular, the convexity of the angle distribution as a parameter of interest for distinguishing different types of motions.

To estimate the analytical shape of the angle distribution the following fitting is performed:

$$p_{\theta_{\mathbf{X}}}(x) \propto ax^2\,. \tag{3.1}$$

The value of a defines a geometrical feature of the trajectory linked to its directional variability.

To quantify the existence of a preferred motion direction along the trajectory, similarly to [7], we also consider the following index of directionality

$$P_d(\mathbf{X}) = \mathbb{P}(|\theta_{\mathbf{X}}| < \pi/2) - \mathbb{P}(|\theta_{\mathbf{X}}| \geq \pi/2)\,. \tag{3.2}$$

A positive P_d indicates a persistent movement corresponding to diffusive dynamics. In contrast, a negative P_d indicates an anti-persistent movement corresponding to a confined dynamic due to a tendency to return to an equilibrium point.

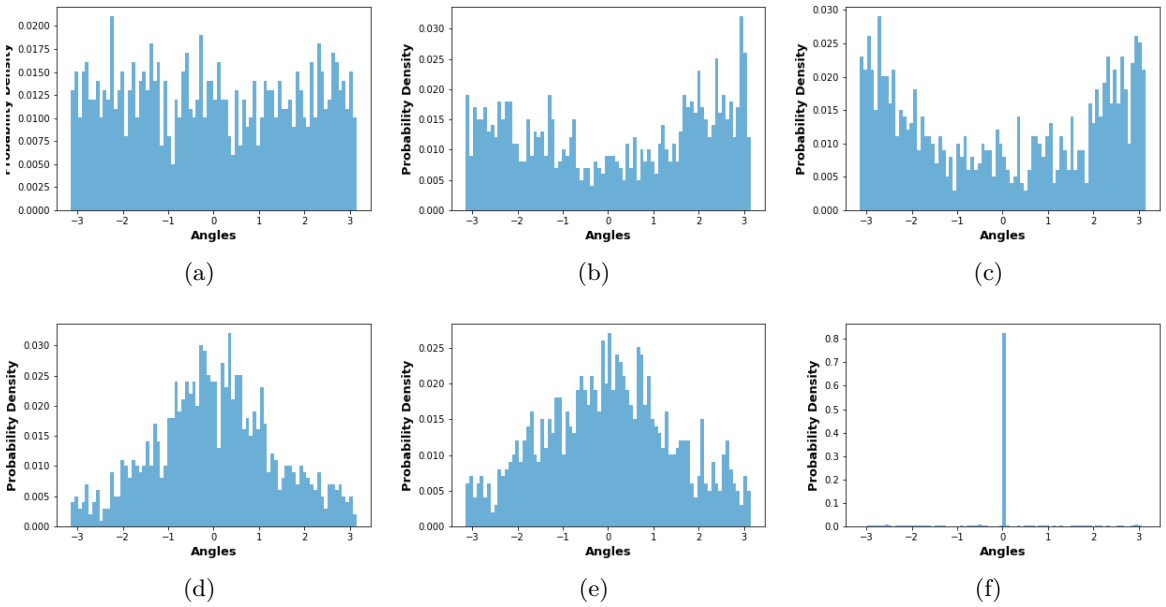


Figure 2: Examples of angle distribution for different processes ($N = 1000$, $\Delta t = 1$): (a) BM ($\sigma = 1$); (b) OU ($\lambda = 0.5$, $\sigma = 1$); (c) FBM with $H = 0.2$; (d) FBM with $H = 0.8$; (e) DIR ($\|u\| = 0.7$, $\sigma = 1$); (f) CTRW with $\sigma = 1$ and $\gamma = 0.9$.

As said above, the rest of the features relate to how the particle unfurls in space during its trajectory, which can be described by analyzing its displacements through concentric balls.

We consider the Ripley's index K_r in a ball $B(X_{t_1}, r)$ of radius r centered at the starting point:

$$K_r = |\{X_{t_i} \in \mathbf{X} \,|\, X_{t_i} \in B(X_{t_1}, r)\}| \,/\, N \,,$$

accounting for the number of trajectory points living in that ball. To make this computation consistent with the trajectory dynamic, we define the reference radius

$$R = \frac{1}{N} \sum_{i=1}^{N-1} \|X_{t_{i+1}} - X_{t_i}\|$$

and we compute the vector

$$K_{\mathbf{X}} = (K_R, ..., K_{kR}, ..., K_{NR}) \,.$$

Figure 3 shows the vector $K_{\mathbf{X}}$ for different processes, suggesting that its analytical shape enables differentiating different dynamic behaviors. The analytical shape of $K_{\mathbf{X}}$ is characterized by an increasing function reaching a final plateau. This characteristic plateau starts at the first radius larger than the maximum distance the particle reaches from its starting point along the trajectory. It is reached since small radii for confined trajectories, while, superdiffusive trajectories (DIR or FBM with $H > 1/2$) exhibit an initial slower increase.

This suggests estimating the analytical law of $K_{\mathbf{X}}$, as a function of index k, using the following fitting:

$$K_{\mathbf{X}} \propto 1 - e^{-br} \tag{3.3}$$

and considering b as the third feature of interest.

Moreover, beyond the global shape graph of $K_{\mathbf{X}}$, some local plateaus can appear, for instance between radii kR and $(k + k_0)N$ ($k_0 < N$), depending on particle displacements after having passed the boundary of $B(X_{t_1}, kR)$. This reveals local (in time) confined evolutions or, on the contrary, sporadic long displacements.

Figure 3 shows several dynamic behaviors that can lead to local plateaus. For example, this can happen for the CTRW process because of longer waiting times, which can lead to a path with no point inside $B(X_{t_1}, (k+k_0)R) \setminus B(X_{t_1}, kR)$. This can also happen for superdiffusive trajectories whenever a displacement is larger than $k_0 R$, making the particle spread through the boundary of several successive balls.

The quantification of local plateaus allows describing these local irregularities. This can be estimated by the ratio of points with non-null derivative:

$$P_p(\mathbf{X}) = \mathbb{P}(K'_{\mathbf{X}} \neq 0) \,, \tag{3.4}$$

which defines the last feature considered in this work.

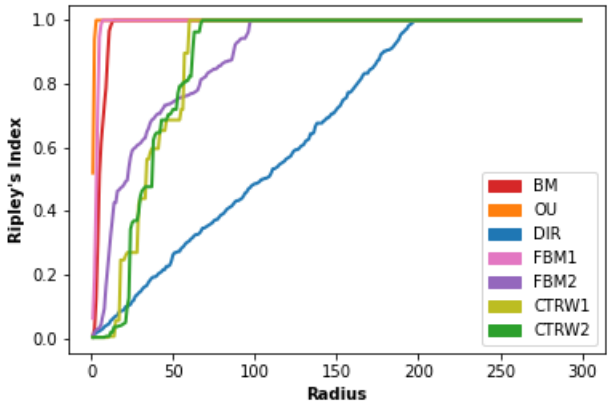


Figure 3: Curves of Ripley's indices for different processes ($N = 300$, $\Delta t = 1$, radius is expressed in terms of indices ks'): BM ($\sigma = 1$), OU ($\lambda = 0.5$, $\sigma = 1$), DIR ($\|u\| = 0.7$, $\sigma = 1$), FBM (FBM1 with $H = 0.2$, and FBM2 with $H = 0.8$), CTRW ($\sigma = 1$, $\gamma = 0.01$ for CTRW1 and $\gamma = 0.9$ for CTRW2).

4 Method and Results

This section presents a supervised motion classification method based on the previous section's features. The first part is dedicated to presenting the method on a simulated dataset. In this context, we study the influence of image noise and trajectories' length on the method's accuracy. The second part applies the method to biological data to study the dynamical behavior of receptors (CCR5) at the cell membrane.

4.1 Trajectory simulation.

The method is set on a dataset of trajectories simulating the stochastic processes presented in Section 2. A dataset is generated for a fixed trajectory length, simulating 1000 trajectories per process.

The time interval Δt between two consecutive points within the trajectory depends on the observed phenomenon and the acquisition method. The choice of this parameter has to be consistent with the acquisition process because it defines the statistical properties of several processes (for instance, the variance and covariance of increments in BM, FBM, and OU). Then, it influences the related previously defined features. This work considers $\Delta t = 1/30$, a common value in confocal imaging in biology.

Brownian motion is generated by iterating random Gaussian data generation with σ uniformly sampled between 0.1 and 10.

OU process is generated using σ between 1 and 10, and for each of them, we set $\lambda = r\sigma/\sqrt{\Delta t}$ where r is a ratio parameter uniformly sampled between 0.2 and 1. Trajectories are simulated using the DiffusionProcess class available in the stochastic package of Python [36].

Similarly, directed trajectories are simulated by summing iteratively a Gaussian vector with the drift one. The diffusion coefficient σ is uniformly distributed between 1 and 10, the drift vector has fixed direction $(1, 1)$, and its norm is defined by $\mu = r\sigma/\Delta t$ with r sampled between 0.2 and 1.

Considering Δt in the parameter computation for OU and DIR motions ensures that r represents the ratio of their deterministic component to the random one.

To simulate FBM trajectories, we use the Hosking algorithm [29], implemented in the fbm Python package [37], with Hurst parameters H uniformly sampled on $[0.15, 0.3] \cup [0.7, 0.85]$. This choice enables the analysis of FBM trajectories with a behavior strongly distinct from a Brownian ($H = 1/2$), confined ($H \sim 0$), or superdiffusive ($H \sim 1$) one.

Finally, CTRW trajectories are simulated using a Gaussian distribution for jumps with σ uniformly sampled between 0.1 and 10. For each value, the waiting times are simulated according to an exponential distribution with parameter γ uniformly sampled on $]0, 1[$. In particular, waiting times are sampled on the interval $[1, N/5]$ to avoid too large times compared to the trajectory length. Once the set of parameters is fixed, the path simulation is straightforward [32]. A sufficient number of waiting times and jumps must be generated using the related distributions, and the desired path is defined by alternating obtained jumps and waiting times.

4.2 Classification method

We develop a supervised learning method for classifying five processes described in Section 2.2 based on the geometric features described in Section 3. For each trajectory, the features dataset collects the parameters defined in (3.1), (3.2), (3.3), and (3.4). The dataset is split into training and test sets following the ratio of

70%-30% in a balanced way to models and parameters. Then, a Random Forest model (ten trees) is trained using the model's name as labels (BM, OU, DIR, FBM, CTRW) and validated on the test set. Table 2 and Table 3 present the method results for different trajectory lengths.

As expected, the method performance improves for larger lengths. This shows that the geometric characteristics of movements need time to assert themselves distinctively, confirming the difficulty of the motion classification problem for short trajectories. Most misclassifications concern short FBM trajectories (classified as OU or BM) due to the variability of their dynamic behaviors depending on H. This error decreases significantly with increasing length, proving that FBM describes an intrinsically different dynamic fully characterized by our features. On the other hand, CTRW behavior is easily learned, also for short trajectories, due to successive waiting times and the limit configuration of $p_{\theta_{\mathbf{X}}}$ and $K_{\mathbf{X}}$. Figure 4 summarizes the method's accuracy against the trajectory length. Through this paper, the error bars shown in graphs represent the related Binomial proportion confidence interval with confidence level set to 95%.

Length	**BM**	**OU**	**DIR**	**FBM**	**CTRW**
N=40	82.7	80	80	60.7	99.6
N=70	89.9	89.7	90.2	80	100
N=100	94.2	92.3	92.5	89.1	99.9
N=200	98.9	98.2	97.4	96.3	100
N=300	98.8	99	99.6	98.6	100

Table 2: Results of the machine learning method: Recall by motion class is shown for different trajectory lengths.

Length	**BM**	**OU**	**DIR**	**FBM**	**CTRW**
N=40	69.9	81.5	81.7	70.6	100
N=70	82.3	89.3	90.9	87.8	100
N=100	88	92.9	94.1	93.5	100
N=200	95.8	97.6	99	98.5	100
N=300	98.4	98.2	99.5	99.9	100

Table 3: Results of the machine learning method: Precision by motion class is shown for different trajectory lengths.

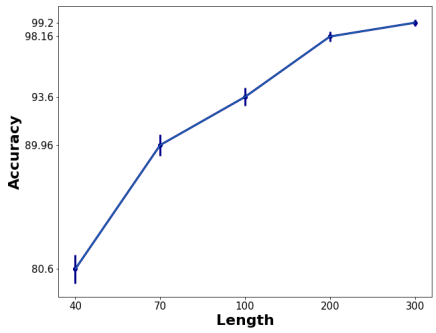


Figure 4: Method accuracy depending on trajectory length.

Finally, Figure 5 shows the features' importance for each class based on the mean decrease accuracy method. A random permutation of values is performed for each feature, and the trained model is applied to the new dataset. For each class, the difference between the original and the new recall estimates how that feature discriminates the class (because of randomness, we average the results of ten independent permutations). The computation is made on the training set for the model with a length of $N = 100$.

Although the importance scores vary with the trajectory length, we can point out some common trends. The fitting coefficient of Ripley's curve is important to identify OU process; this is due to the confined

behavior of OU paths resulting in a rapidly increasing Ripley's curve. Ripley's plateaux distinguish DIR motions because, depending on the drift component, not each ball (used for Ripley's indices calculation) contains new points. CTRW is classified based on the angle histogram close to a Dirac distribution at zero for this process. A more uniform mix of features identifies BM and FBM. In particular, FBM classification is the most impacted by the directionality feature, confirming that it owns intrinsic geometric properties compared to its confined and directed *alter ego* (OU, DIR).

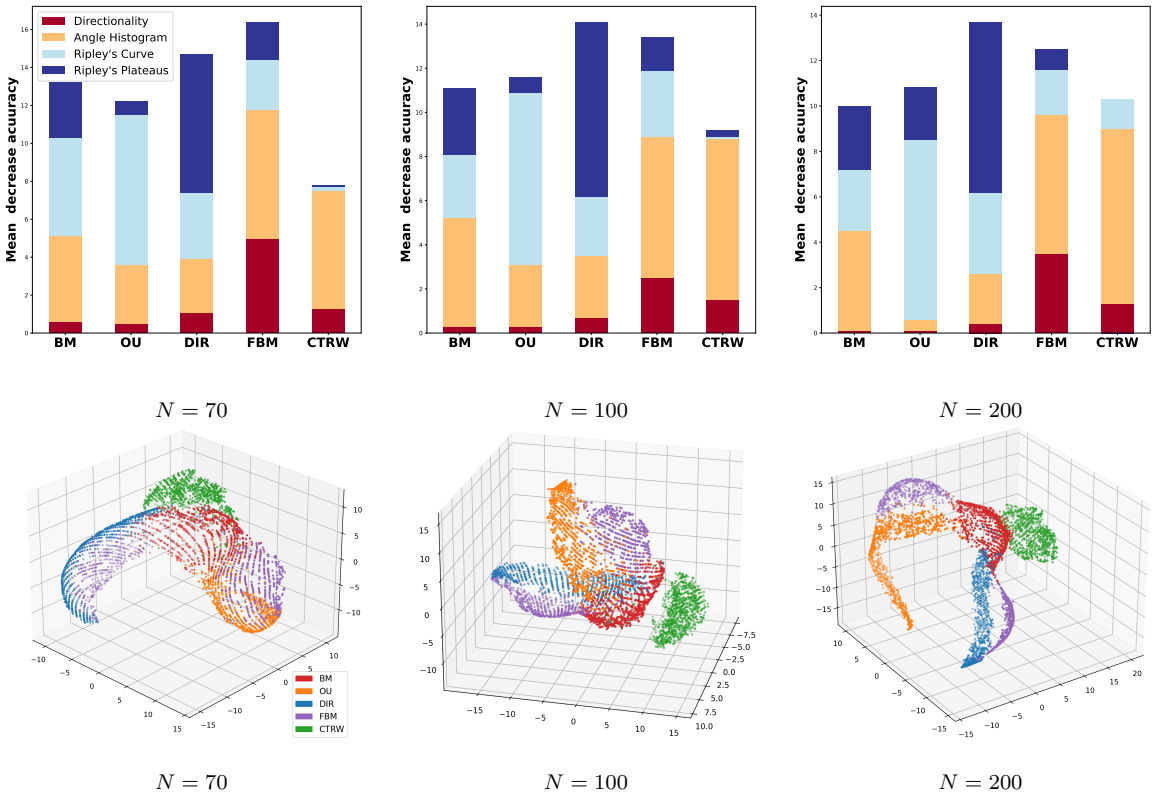


Figure 5: Features analysis on the train set for several trajectory lengths. Top : Feature-specific mean decrease accuracy per motion class. Bottom : tSNE representation in 3D of the feature set.

4.3 Impact of noise and immobility threshold

The main problem in confocal movie analysis is dealing with the associated multi-source image's noise. For instance, electronic noise and intensity variation over time can affect the precision of particle detection. The choice of particle barycenter as a related position also introduces a noise due to particle detection error. In conclusion, tracking algorithms work with noisy estimations of particle positions, and classification methods should be able to overcome this issue. To study our method's noise robustness, we construct a noisy dataset in which each trajectory position is perturbed by adding a Gaussian noise.

Following [15, 14], we consider a Gaussian noise with zero mean and variance σ_n verifying

$$L_n = \begin{cases} \frac{\sqrt{D\Delta t + \mu(\Delta t)^2}}{\sigma_n} & \text{for DIR} \\ \frac{\sqrt{D\Delta t}}{\sigma_n} & \text{otherwise} \end{cases} \tag{4.1}$$

where L_n denotes signal-to-noise ratio (SNR), D is the diffusion coefficient estimated via the fitting formula for the MSD ($\propto 4Dt^\alpha$), and μ is the norm of the drift component used to simulate directed paths (DIR). Ranging L_n from 1 (high noise) to 9 (low noise), the previous equation allows the computation of $\sigma_n = \sigma_n(L_n)$ and the generation of different noisy datasets to the impact on the model.

We test the noise robustness considering the model trained on pure trajectories of length 100 and collecting its performances on the previously generated noisy data.

However, in our case, this standard approach to noise analysis needs a preliminary step. As mentionned above, the estimation of particle barycenter can be affected by electronic noise or intensity variation, impacting the immobility of particles, which is the main characteristic of CTRW paths. An immobility threshold

should be applied to trajectory points to detect immobile particles over different time intervals. To do this, we compare each point with the previous one, and if their distance is smaller than the given threshold, its position is set equal to the previous one. This threshold is empirically estimated in biological applications and depends on the image resolution and particle size. For the simulated data, we perform tests with two thresholds: $\sqrt{2}\sigma_n(L_n)$ corresponding to the exact noise variance, and $\sqrt{2}\sigma_n(7)$ representing an arbitrary threshold corresponding to a low noise. The threshold is applied to each trajectory before performing motion classification.

Figure 6 reports the noise analysis results with the previously defined immobility thresholds. In the bottom panel, the arbitrary threshold $\sqrt{2}\sigma_n(7)$ is applied, and a strong impact is observed on the detection of highly noisy CTRW (SNR=1,2). These cases correspond to an analysis without immobility correction, which strongly affects the CTRW classification. However, these results prove the model's good noise robustness for the other motion classes. In the top panel, the exact threshold $\sqrt{2}\sigma_n(L_n)$ is applied before performing classification. This improves CTRW classification but, for highly noisy trajectories (SNR=1,2), such a correction negatively affects the accuracy of the other motion classes. This is because, for SNR=1,2, the diffusion coefficient used for pure trajectory simulation is similar to the noise variance, leading to the overestimation of immobility in BM or FBM trajectories. The pure trajectory diffusion coefficient for lower SNR becomes larger than the noise variance, avoiding immobility overestimation and improving performance for the two immobility thresholds. Independently on the immobility criterion, starting at SNR=3, the accuracy becomes stable around the value obtained testing on pure trajectories, showing that the features used by the proposed method are strongly robust to noise.

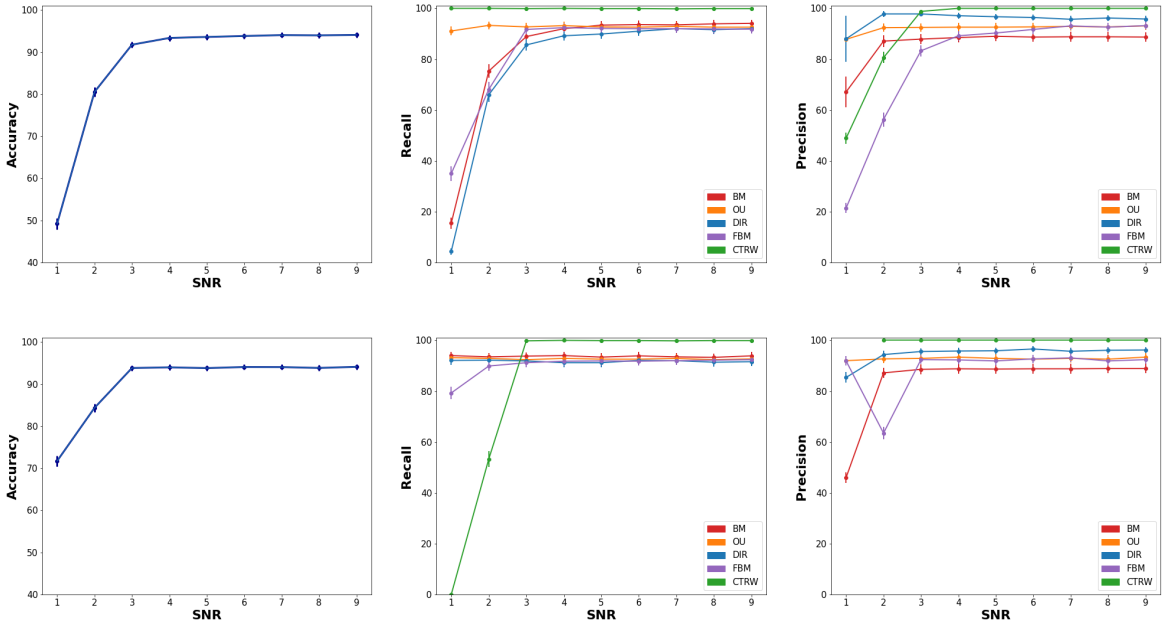


Figure 6: Analysis of accuracy, recall and precision for the method trained on pure trajectories ($N = 100$, $\Delta t = 1/30$) and tested on noisy trajectories. Results show a different behavior depending on the immobility threshold used to correct noisy trajectories. Top: the immobility threshold $\sqrt{2}\sigma_n(L_n)$ depends on the noise applied to the pure trajectory; Bottom: the threshold is arbitrary fixed to $\sqrt{2}\sigma_n(7)$.

4.4 Impact of length inequality

Working with fixed-length trajectories is often difficult and results from post-processing routines. As our method is trained on pure trajectories of a fixed length N, it is important to study how length inequality in the test dataset can affect the method's performance.

To do this, we trained the proposed method on the dataset of pure trajectories of length 100 and tested it on several datasets of pure trajectories with lengths from 70 to 130. The results are reported in Figure 7.

For instance, using the model on shorter trajectories leads to misclassifying directed paths predicted as FBM ones, whereas misclassified FBM trajectories are mostly labeled as OU. Surprisingly, BM motion is the most sensitive to length variation and is mostly confused with the DIR class. OU and CTRW paths are uniformly well classified on shorter trajectories.

On the other hand, results slightly improve if the model is tested on longer trajectories, suggesting that it is convenient to underestimate the length used to train the model.

Finally, by limiting the length variance to $\Delta N = 10$ in the shortening sense ($N > 90$), we register a maximum gap of the accuracy of 1.6% compared to accuracy on the dataset with the same length (N=100, accuracy = 93.6%). This proves the good flexibility of our method, allowing particle tracking analysis based on the collection on more paths, even with a slightly different length.

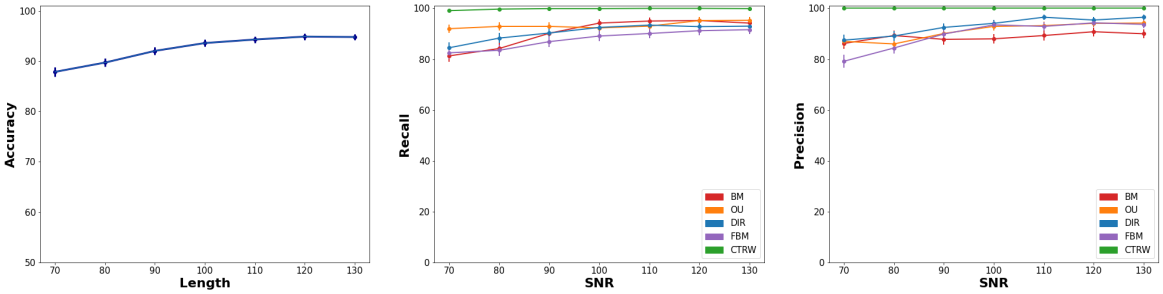


Figure 7: Results of accuracy, recall, and precision for the method trained on pure trajectories of length 100 and tested on pure trajectories with different lengths ($\Delta t = 1/30$).

4.5 Comparison to statistical test method

As detailed in the Introduction, alternative methods have been proposed in recent years to overcome the drawbacks related to the MSD criterion. In particular, [8] develops a statistical test method for diffusion classification (subdiffusive, free (BM), superdiffusive) based on the standardized maximal distance of the particle from its starting point along its trajectory:

$$T_N = \frac{\max\limits_{k=0,\dots,N} \|X_{t_k} - X_{t_0}\|}{\left[\frac{1}{2}\sum_{i=1}^{N} \|X_{t_i} - X_{t_{i-1}}\|^2\right]^{\frac{1}{2}}}$$

Via the Monte-Carlo method, the distribution of T_N is simulated for free motion (BM) of a given length N. Then, the quantiles $q_{2.5}$ and $q_{97.5}$ of its distribution are computed allowing to set the following three-hypothesis-test procedure:

$$\text{Diffusion classification} = \begin{cases} \text{subdiffusive}, & \text{if } T_N < q_{2.5}, \\ \text{superdiffusive}, & \text{if } T_N > q_{97.5}, \\ \text{free motion}, & \text{otherwise.} \end{cases} \tag{4.2}$$

A similar approach is used in [9] to classify the motion of CCR5 receptors at the cell membrane.

We compare the proposed method, based on motion classification, to the hypothesis test (4.2), performing diffusion classification. We consider the test dataset for trajectories of length $N = 100$, and predictions are made both via our method (trained on pure trajectories of the same length) and via the three-decision-test (4.2) (the quantiles are computed via Monte-Carlo simulation of 100001 Brownian trajectories of length $N = 100$).

In Figure 8, we report the normalized (with respect to motion classes) confusion matrices of motion classification versus motion or diffusion classes. BM, OU, and DIR are correctly classified in terms of diffusion. FBM is correctly split into two classes of subdiffusive ($H < 1/2$) and superdiffusive ($H > 1/2$) paths. However, compared to our method, a larger part of the FBM track is misclassified by the test (4.2). Misclassified paths are labeled as free motion (BM), although they have been simulated with Hurst coefficients far from $1/2$ ($H \in [0.15, 0.3] \cup [0.7, 0.85]$). Finally, the test (4.2) misclassifies the totality of CTRW trajectories as free motion. Although this is consistent with the Gaussianity of the jumps distribution, this shows that the (4.2) is unadapted to the detection of CTRW paths.

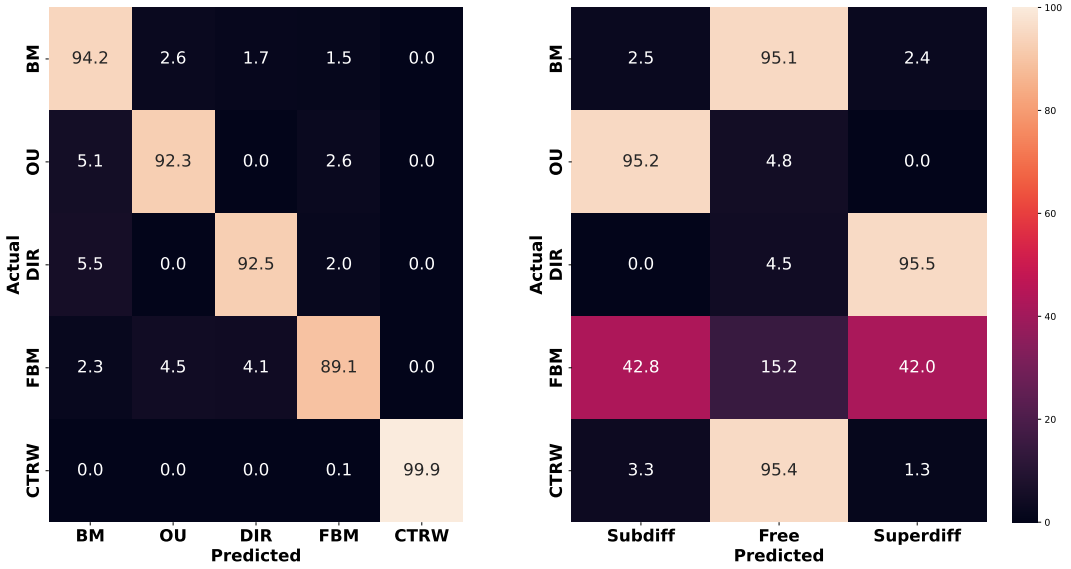


Figure 8: Performance comparison of machine learning and three-hypothesis-test (4.2) [8] tested on pure trajectories of length 100. Left: Normalized confusion matrix for the proposed model trained on pure trajectories of length 100. Right: Normalized confusion matrix for the diffusion classification (4.2) compared to motion labels.

4.6 Mixing motions

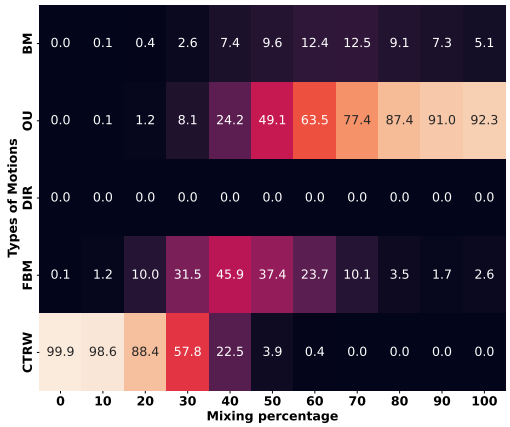
This section shows the method's performance on trajectories mixing several types of motion. To construct these trajectories we consider the pure trajectories of the test dataset with length $N = 100$ and mix them according to several percentages. Let m_1, m_2 be two types of motions, and let T_{m_1}, T_{m_2} the sets of corresponding paths (1000 for each motion) in the test dataset. For a given percentage p and for $i = 1, ..., 1000$, we consider $\mathbf{X}^{1,i} \in T_{m_1}, X^{2,i} \in T_{m_2}$ and define the following mixed trajectory

$$\mathbf{X}^i_{mixed} = \left(\mathbf{X}^{1,i}_0, ..., \mathbf{X}^{1,i}_p, \tilde{\mathbf{X}}^{2,i}_{p+1}, ..., \tilde{\mathbf{X}}^{2,i}_N\right)$$

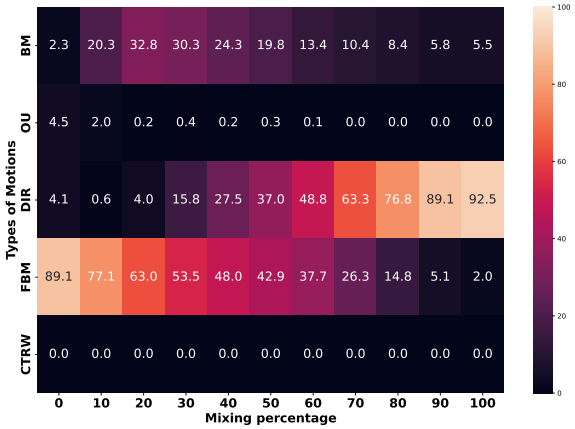
where $\tilde{\mathbf{X}}^{2,i} = \mathbf{X}^{2,i} + (\mathbf{X}^{1,i}_p - \mathbf{X}^{2,i}_p)$.

Considering several percentages (from 10% to 90%), this defines, for each p, a set of 1000 trajectories whose first p steps are governed by the m_1 motion and the rest by the m_2 motion.

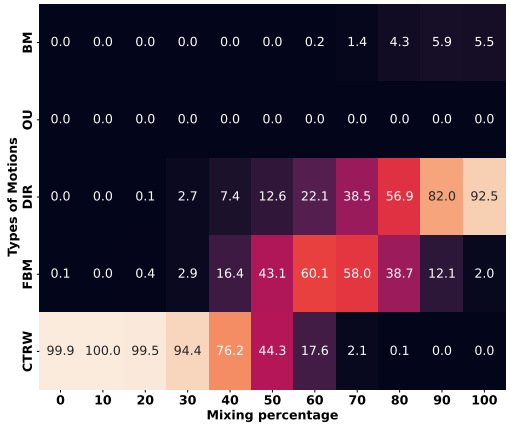
For each pair of motions and each percentage, the corresponding mixing trajectories are analyzed using the model trained on pure trajectories of length $N = 100$. This allows testing of how the detection of the majority motion is influenced by the coupling association and the corresponding percentage. Figure 10 shows the main results on trajectories mixing Brownian motion with other processes, while Figure 9 concerns the other couples of motions.



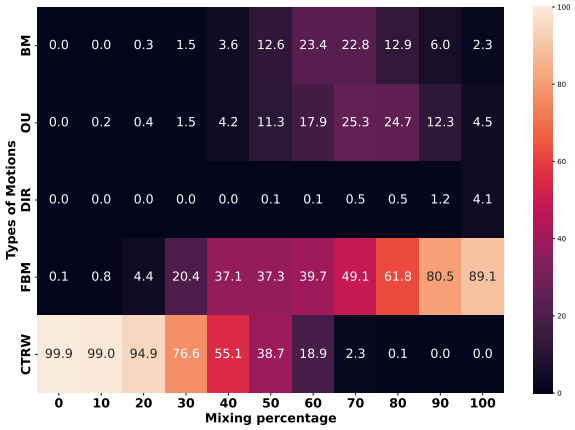
$m_1 = \text{OU}$, $m_2 = \text{CTRW}$



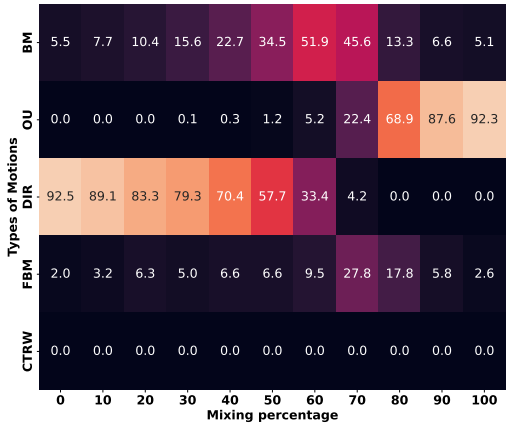
$m_1 = \text{DIR}$, $m_2 = \text{FBM}$



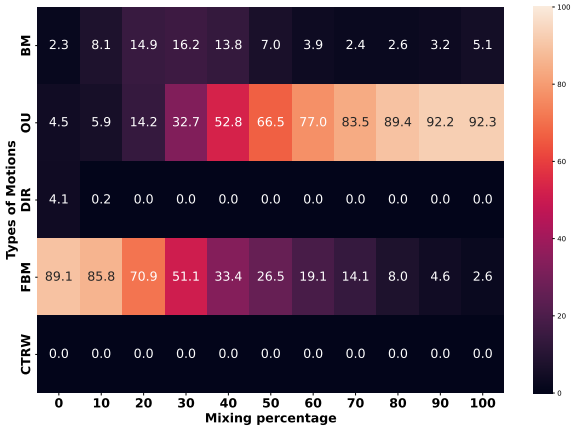
$m_1 = \text{DIR}$, $m_2 = \text{CTRW}$



$m_1 = \text{FBM}$, $m_2 = \text{CTRW}$

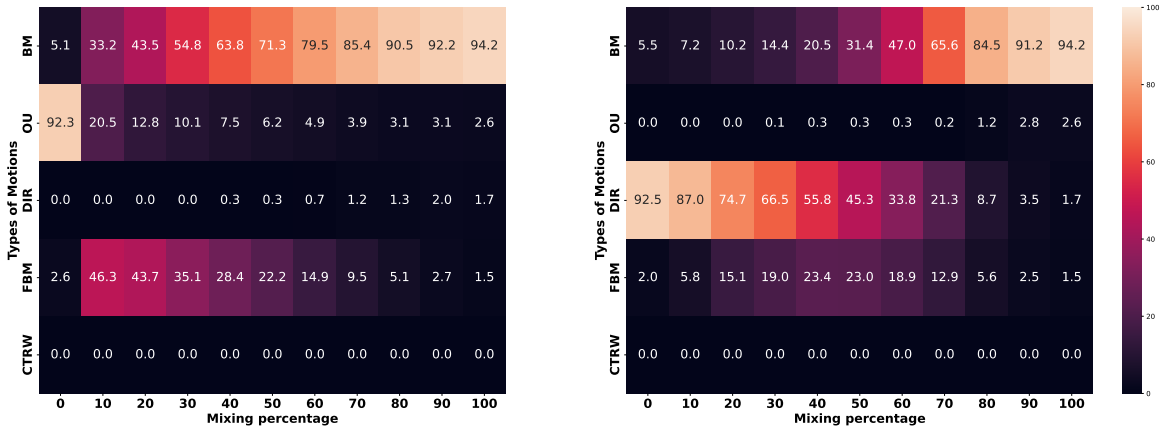


$m_1 = \text{OU}$, $m_2 = \text{DIR}$

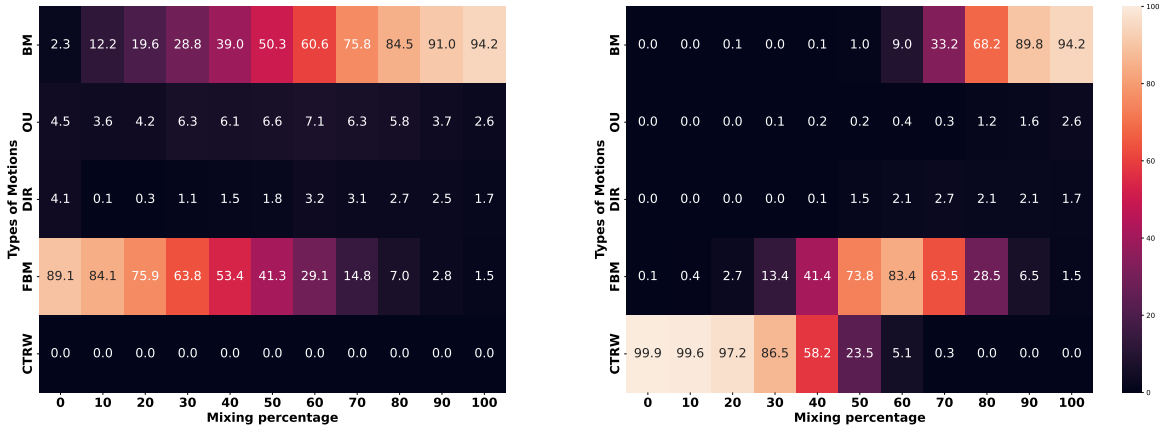


$m_1 = \text{OU}$, $m_2 = \text{FBM}$

Figure 9: Trajectories mixing subdiffusive and directed motions are analyzed with the model trained on pure trajectories ($N = 100$). The prediction performance is shown for each couple of motions depending on the mixing percentage.



$m_1 = \text{BM}$, $m_2 = \text{OU}$ $m_1 = \text{BM}$, $m_2 = \text{DIR}$



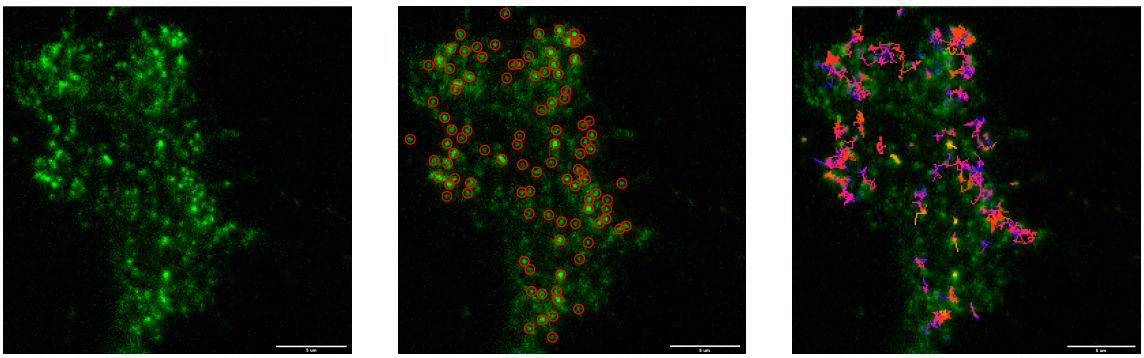
$m_1 = \text{BM}$, $m_2 = \text{FBM}$ $m_1 = \text{BM}$, $m_2 = \text{CTRW}$

Figure 10: Trajectories mixing Brownian motion with the other types of movement. The mixed trajectories are analyzed with the model trained on pure trajectories. The prediction performance is shown for each couple of motions depending on the mixing percentage.

4.7 Application to biological data

Motion classification algorithms are widely used in biology to study the dynamic behavior of sub-cellular particles. In this work, we are interested in analyzing the cell membrane receptors CCR5, which are involved in HIV infection. Images are acquired using confocal microscopy, enabling visualization of a thin outer layer of the cell membrane (200 nm). We collected movies with 30 fps (which is consistent with the developed method with $\Delta t = 1/30$) by imaging several entire cells. The images are processed using the *Spot Detector* and *Spot Tracking* plugins of the Icy software [38], which allows the detection of the receptors and the reconstruction of related trajectories. Figure 11 shows the different steps of the tracking analysis on Icy. As the receptor size corresponds to three pixels (equivalent to three microns), the immobility threshold of one pixel is applied to correct trajectory position and highlight immobility.

To track CCR5 on long trajectories without ambiguities, we worked with cells expressing a low receptor level at the cell surface. We used cells expressing CCR5 under the control of the RUSH system (retention using selective hooks) developed in [39, 40]. It allows the synchronization and the study of proteins, which follow the biosynthesis/secretion pathway. It is based on the intracellular retention of a protein of interest (here CCR5) and its release by induction (using biotin). RUSH-CCR5 expressing cells, even if not biotin-induced, express CCR5 at a very low level, which corresponds to a leak from the system but which is ideal for our monitoring. The RUSH-CCR5 construct allows the cell surface expression of a CCR5 protein fused to the fluorescent protein GFP. Then, the presence of GFP-CCR5 at the cell surface is detected by labeling the cells with an anti-GFP-AF647 booster. This labeling overcomes the background noise linked to the



(a) (b) (c)

Figure 11: Different steps of tracking algorithm: (a) original image, (b) spot detection, (c) trajectory reconstruction (scale bar = 5 μm).

presence of GFP-CCR5 along the secretion pathways.

This experimental technique allows imaging receptors in a low-density environment, facilitating detection and trajectory reconstruction. This enables both the minimization of tracking errors and the collection of longer trajectories. We can finally collect a dataset of a thousand trajectories with lengths between 90 and 120 time steps. According to the previous analysis on the impact of length inequality and noise, the dataset can be analyzed using a model trained on pure trajectories of length 100.

In the following, we perform motion classification to compare the basal state to the treatment by PSC-RANTES, which displays potent anti-HIV-1 activity. The exceptional capacity of PSC-RANTES to inhibit infection is related to its ability to increase CCR5 down-regulation. PSC-RANTES acts as a superagonist by recognizing a larger array of CCR5 conformational states than native chemokines [41].

The results in Figure 12 reveal several motion subpopulations governed by different processes highlighting that several groups of receptors coexist, facing different environmental constraints and fates. Moreover, these results show the strong impact of the PSC-RANTES stimulation on the nature of the receptors' dynamic.

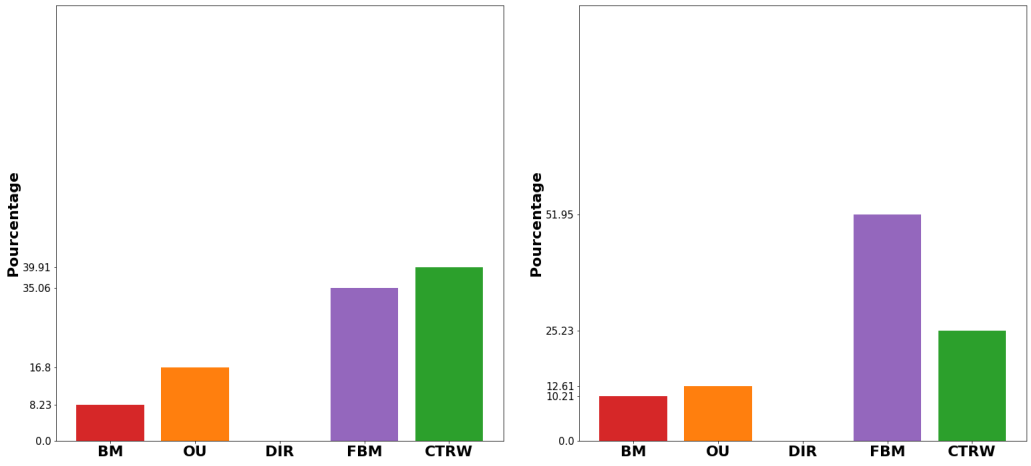
In the case of cellular receptors, subdiffusive processes correspond to different situations: OU allows describing attraction between receptors, implying confined evolutions; FBM describes constraint movement across the cell membrane, which is viscoelastic and inhabited by other protein assemblies; CTRW refers to the case where receptors are immobilized because a ligand or a temporary change of their polymerization state.

Figure 12 shows that CCR5 dynamics are governed by CTRW and FBM dynamics. In the basal state, the majority of tracks exhibit either intermittent motion (CTRW) or dynamics with constrained increments (FBM). In particular, this analysis reveals that free motion is better described by CTRW than by pure BM, as is often assumed. On the other hand, after stimulation by PSC-RANTES, CTRW dynamics are less represented in favor of FBM motion. This highlights the impact of PSC-RANTES stimulation on receptors' behavior, prompting them to move from free movement, characterized by jumps and pauses, to constrained spreading at the cell membrane.

5 Discussion

This paper addresses the problem of motion classification for particle tracking and related applications in biology. Instead of looking at the trajectory in terms of diffusion, as in the standard approach, the proposed method performs motion classification based on stochastic processes. This work presents a unified framework to recognize the five standard dynamics used in particle dynamic modeling (BM, OU, DIR, FBM, CTRW), which is a step forward in the field, especially in distinguishing different behaviors in the subdiffusive regime. This is particularly useful for studying the dynamics of cell receptors, which essentially hold subdiffusive dynamics, to distinguish different subpopulations responding to different environmental and biological constraints.

The proposed method follows a features-based supervised approach (Random Forest) to guarantee the geometrical characterization of trajectories and the explicability of predictions. The results on simulated data are proven for trajectories of length 100 with a time interval $\Delta t = 1/30$, a common frame-per-second rate in TIRF microscopy, obtaining an overall accuracy of 93.6%. As discussed in Section 4.2, the method accuracy strongly depends on the length of trajectories. This confirms that a specific motion needs time to deploy its intrinsic properties to be described by statistical estimators. This reaffirms the difficulty of short-trajectory classification and encourages novel observation techniques to avoid them. This is the main reason for the choice of Rush system for imaging receptors in Section 4.7: this enables imaging receptors



(*a*) (*b*)

Figure 12: Results of motion classification for CCR5 trajectories of length approximatively equal to 100 time steps ($\Delta t = 1/30$). Comparison between the basal state (*a*) and after PSC-RANTES stimulation (*b*).

in less dense environments, leading to fewer tracking errors and, finally, longer trajectories. Although this acquisition technique allows the collection of reliable trajectories, our method performs analysis at a fixed time scale without addressing the complex question of dynamic change over time. Particle dynamics can vary along their trajectories in response to environmental constraints, and switching times represent useful information for biological studies. Classification methods using fixed-length trajectories can lose this information, as motion switching can happen within the fixed temporal window. Detecting dynamic changes remains a difficult problem involving the definition of estimators for different behaviors to test over time. We strongly believe that the features-based approach developed in this work represents a preliminary framework to analyze dynamic consistency at different time scales.

The robustness to image noise appearing because of several experimental constraints (for instance, fluorescence variability inducing localization errors) is studied in Section 4.3. The results prove that the selected features represent an intrinsic signature of motion that resists low noise levels. In this context, the impact of an immobilization threshold, used to correct detected position and highlight immobility, is studied, showing its impact on accuracy results.

Moreover, the impact of length inequality for test trajectories is discussed in Section 4.4. Considering that trajectories have, at most, ten steps less, with respect to the length used for training, the method keeps performing in the same range. This guarantees, in particular, flexibility and reliability in biological applications. Finally, as shown in Section 4.5, our approach produces a motion classification consistent with the diffusion-based one, in addition to distinguishing CTRW that would be classified as BM otherwise.

In conclusion, our method defines an accurate and explicable approach to observing trajectories at a given temporal scale and recognizing their evolutions. Beyond stochastic process recognition, the approach distinguishes different ways to unfurl in space using geometric features encoding the intrinsic geometrical properties of particle dynamics. The robustness to noise and length inequality makes it an ergonomic and reliable approach for dynamical classification and related application to biological imaging.

References

[1] N. Chenouard, I. Smal, F. De Chaumont, M. Maška, I. F. Sbalzarini, Y. Gong, J. Cardinale, C. Carthel, S. Coraluppi, M. Winter, *et al.*, "Objective comparison of particle tracking methods," *Nature methods*, vol. 11, no. 3, pp. 281–289, 2014.

[2] A. Einstein *et al.*, "On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat," *Annalen der physik*, vol. 17, no. 549-560, p. 208, 1905.

[3] H. Qian, M. P. Sheetz, and E. L. Elson, "Single particle tracking. analysis of diffusion and flow in two-dimensional systems," *Biophysical journal*, vol. 60, no. 4, pp. 910–921, 1991.

[4] M. J. Saxton and K. Jacobson, "Single-particle tracking: applications to membrane dynamics," *Annual review of biophysics and biomolecular structure*, vol. 26, no. 1, pp. 373–399, 1997.

[5] J. Rudnick and G. Gaspari, "The shapes of random walks," *Science*, vol. 237, no. 4813, pp. 384–389, 1987.

[6] M. J. Saxton, "Lateral diffusion in an archipelago. the effect of mobile obstacles," *Biophysical journal*, vol. 52, no. 6, pp. 989–997, 1987.

[7] N. Gal, D. Lechtman-Goldstein, and D. Weihs, "Particle tracking in living cells: a review of the mean square displacement method and beyond," *Rheologica Acta*, vol. 52, pp. 425–443, 2013.

[8] V. Briane, C. Kervrann, and M. Vimond, "Statistical analysis of particle trajectories in living cells," *Physical Review E*, vol. 97, no. 6, p. 062121, 2018.

[9] F. Momboisse, G. Nardi, P. Colin, M. Hery, N. Cordeiro, S. Blachier, O. Schwartz, F. Arenzana-Seisdedos, N. Sauvonnet, J.-C. Olivo-Marin, *et al.*, "Tracking receptor motions at the plasma membrane reveals distinct effects of ligands on ccr5 dynamics depending on its dimerization status," *Elife*, vol. 11, p. e76281, 2022.

[10] K. Hubicka and J. Janczura, "Time-dependent classification of protein diffusion types: A statistical detection of mean-squared-displacement exponent transitions," *Physical Review E*, vol. 101, no. 2, p. 022107, 2020.

[11] A. Weron, J. Janczura, E. Boryczka, T. Sungkaworn, and D. Calebiro, "Statistical testing approach for fractional anomalous diffusion classification," *Physical Review E*, vol. 99, no. 4, p. 042149, 2019.

[12] V. Briane, M. Vimond, and C. Kervrann, "An overview of diffusion models for intracellular dynamics analysis," *Briefings in bioinformatics*, vol. 21, no. 4, pp. 1136–1150, 2020.

[13] Y. Meroz and I. M. Sokolov, "A toolbox for determining subdiffusive mechanisms," *Physics Reports*, vol. 573, pp. 1–29, 2015.

[14] T. Wagner, A. Kroll, C. R. Haramagatti, H.-G. Lipinski, and M. Wiemann, "Classification and segmentation of nanoparticle diffusion trajectories in cellular micro environments," *PloS one*, vol. 12, no. 1, p. e0170165, 2017.

[15] J. Janczura, P. Kowalek, H. Loch-Olszewska, J. Szwabiński, and A. Weron, "Classification of particle trajectories in living cells: Machine learning versus statistical testing hypothesis for fractional anomalous diffusion," *Physical Review E*, vol. 102, no. 3, p. 032402, 2020.

[16] G. Muñoz-Gil, M. A. Garcia-March, C. Manzo, J. D. Martín-Guerrero, and M. Lewenstein, "Single trajectory characterization via machine learning," *New Journal of Physics*, vol. 22, no. 1, p. 013010, 2020.

[17] S. Bo, F. Schmidt, R. Eichhorn, and G. Volpe, "Measurement of anomalous diffusion using recurrent neural networks," *Physical Review E*, vol. 100, no. 1, p. 010102, 2019.

[18] N. Granik, L. E. Weiss, E. Nehme, M. Levin, M. Chein, E. Perlson, Y. Roichman, and Y. Shechtman, "Single-particle diffusion characterization by deep learning," *Biophysical journal*, vol. 117, no. 2, pp. 185–192, 2019.

[19] T. Sungkaworn, M.-L. Jobin, K. Burnecki, A. Weron, M. J. Lohse, and D. Calebiro, "Single-molecule imaging reveals receptor–g protein interactions at cell surface hot spots," *Nature*, vol. 550, no. 7677, pp. 543–547, 2017.

[20] G. D. Birkhoff, "Proof of the ergodic theorem," *Proceedings of the National Academy of Sciences*, vol. 17, no. 12, pp. 656–660, 1931.

[21] U. Krengel, *Ergodic theorems*, vol. 6. Walter de Gruyter, 2011.

[22] G. Ruan, A. Agrawal, A. I. Marcus, and S. Nie, "Imaging and tracking of tat peptide-conjugated quantum dots in living cells: new insights into nanoparticle uptake, intracellular transport, and vesicle shedding," *Journal of the American Chemical Society*, vol. 129, no. 47, pp. 14759–14766, 2007.

[23] G. E. Uhlenbeck and L. S. Ornstein, "On the theory of the brownian motion," *Physical review*, vol. 36, no. 5, p. 823, 1930.

[24] N. Monnier, S.-M. Guo, M. Mori, J. He, P. Lénárt, and M. Bathe, "Bayesian approach to msd-based analysis of particle motion in live cells," *Biophysical journal*, vol. 103, no. 3, pp. 616–626, 2012.

[25] Y. Mardoukhi, A. Chechkin, and R. Metzler, "Spurious ergodicity breaking in normal and fractional ornstein–uhlenbeck process," *New Journal of Physics*, vol. 22, no. 7, p. 073012, 2020.

[26] G. Guigas, C. Kalla, and M. Weiss, "Probing the nanoscale viscoelasticity of intracellular fluids in living cells," *Biophysical journal*, vol. 93, no. 1, pp. 316–323, 2007.

[27] G. F. Lawler, "Stochastic calculus: An introduction with applications," *American Mathematical Society*, 2010.

[28] L. Coutin, "An introduction to (stochastic) calculus with respect to fractional brownian motion," in *Séminaire de Probabilités XL*, pp. 3–65, Springer, 2007.

[29] T. Dieker, *Simulation of fractional Brownian motion.* Masters Thesis, Department of Mathematical Sciences, University of Twente, 2004.

[30] E. W. Montroll and G. H. Weiss, "Random walks on lattices. ii," *Journal of Mathematical Physics*, vol. 6, no. 2, pp. 167–181, 1965.

[31] J. Klafter and I. M. Sokolov, *First steps in random walks: from tools to applications.* OUP Oxford, 2011.

[32] G. Germano, M. Politi, E. Scalas, and R. L. Schilling, "Stochastic calculus for uncoupled continuous-time random walks," *Physical Review E*, vol. 79, no. 6, p. 066102, 2009.

[33] A. Lubelski, I. M. Sokolov, and J. Klafter, "Nonergodicity mimics inhomogeneity in single particle tracking," *Physical review letters*, vol. 100, no. 25, p. 250602, 2008.

[34] J.-H. Jeon and R. Metzler, "Analysis of short subdiffusive time series: scatter of the time-averaged mean-squared displacement," *Journal of Physics A: Mathematical and Theoretical*, vol. 43, no. 25, p. 252001, 2010.

[35] T. Neusius, I. M. Sokolov, and J. C. Smith, "Subdiffusion in time-averaged, confined random walks," *Physical Review E*, vol. 80, no. 1, p. 011109, 2009.

[36] C. Flynn, *Generate realizations of stochastic processes in python.* https://github.com/crflynn/stochastic.

[37] C. Flynn, *Exact methods for simulating fractional Brownian motion and fractional Gaussian noise in Python.* https://github.com/crflynn/fbm.

[38] F. De Chaumont, S. Dallongeville, N. Chenouard, N. Hervé, S. Pop, T. Provoost, V. Meas-Yedid, P. Pankajakshan, T. Lecomte, Y. Le Montagner, *et al.*, "Icy: an open bioimage informatics platform for extended reproducible research," *Nature methods*, vol. 9, no. 7, pp. 690–696, 2012.

[39] G. Boncompain, S. Divoux, N. Gareil, H. De Forges, A. Lescure, L. Latreche, V. Mercanti, F. Jollivet, G. Raposo, and F. Perez, "Synchronization of secretory protein traffic in populations of cells," *Nature methods*, vol. 9, no. 5, pp. 493–498, 2012.

[40] G. Boncompain, F. Herit, S. Tessier, A. Lescure, E. Del Nery, P. Gestraud, I. Staropoli, Y. Fukata, M. Fukata, A. Brelot, *et al.*, "Targeting ccr5 trafficking to inhibit hiv-1 infection," *Science Advances*, vol. 5, no. 10, p. eaax0821, 2019.

[41] J. Jin, P. Colin, I. Staropoli, E. Lima-Fernandes, C. Ferret, A. Demir, S. Rogée, O. Hartley, C. Randriamampita, M. G. Scott, *et al.*, "Targeting spare cc chemokine receptor 5 (ccr5) as a principle to inhibit hiv-1 entry," *Journal of Biological Chemistry*, vol. 289, no. 27, pp. 19042–19052, 2014.