Bar-Ilan University

Statistical Properties of Individual Cell Trajectories Within a Monolayer of Human Bronchial Epithelial Cells

Alon Sela

Submitted in partial fulfillment of the requirements for the Master's Degree in the Department of Mathematics,
Bar-Ilan University

Ramat Gan, Israel 2025

This work was carried out under the supervision of Prof. Gil Ariel,

Department of Mathematics,
Bar-Ilan University.

To my wife, **Megi**.

Thank you for your constant support, love, and patience throughout this journey.

To my children, **Yehonatan** and **Emily.**

Your smiles and hugs gave me strength and motivation along the way.

Contents

Abstract

This thesis investigates the statistical properties of individual trajectories of human bronchial epithelial (HBE) cells within a monolayer, using experimental time-lapse data. The primary goal is to determine whether cell motility follows a Lévy walk—a model of super-diffusion characterized by heavy-tailed step distributions and power-law velocity autocorrelation decay. To this end, cell trajectories were extracted from microscopy images using Trackpy and analyzed through a variety of statistical tools, including mean squared displacement (MSD), displacement distributions, stationarity tests, and velocity auto-correlation functions (ACF).

Initial MSD analysis revealed super-diffusive behavior with scaling exponents close to ballistic motion in the early stages of the experiment, transitioning to near-normal diffusion as cell density increased. However, further analysis of displacement distributions and ACF showed clear deviations from Lévy walk predictions. In particular, displacement data did not fit Lévy alpha-stable or Gaussian distributions, but were well captured by a mixture of two normal distributions. Velocity ACFs displayed exponential rather than power-law decay, further contradicting the Lévy walk model.

Segmented and moving window analyses highlighted a temporal shift in the diffusion regime, with increasing crowding correlating with reduced motility. These findings suggest that while HBE cells initially exhibit super-diffusive motion, their behavior over time is more complex and better modeled by a heterogeneous, density-dependent process rather than a pure Lévy walk.

Part I

Introduction

Random walks are mathematical models that describe paths consisting of a succession of random steps [4, 10]. They are used to model various stochastic processes in fields such as physics, biology, finance, and computer science. Lévy walks are a specific type of random walk that incorporate both short and long steps, following a power-law distribution for the step lengths [1, 10]. This results in super-diffusive behavior, where the mean squared displacement grows faster than linearly with time[1, 9, 10]. In biology, Lévy walks have been observed in the movement patterns of various organisms, including animals searching for food, the spread of pathogens, and the movement of cells and microorganisms [1].

For example, in [1] it has been suggested that bacterial trajectories are subject to super-diffusion characterized by Lévy walks.The goal of my thesis is to characterize the statistical properties of human bronchial epithelial (HBE) cell trajectories. In particular, I aim to test whether they resemble Lévy walks, which implies a non Gaussian super-diffusion. We collaborate with Dr. Victor Yashunsky from Ben-Gurion University of the Negev, who provided the experimental data.

Research Questions In order to characterize the statistical properties of cellular trajectories, my research will focus on answering the following questions:

- Mean squared displacement (MSD) of cellular trajectories.
- How does the diffusion behavior evolve over time?
- What is the distribution of displacements?
- Do the scaled displacements distribution fit Lévy alpha-stable distribution?
- Can the velocity of the cells be considered a stationary process?
- Velocity auto-correlation function.
- Is there a connection between cell density and the diffusion regime?
- Can an alternative model better explain the observed data?

The answers to these questions, as well as any additional insights that may arise, will contribute to a more comprehensive understanding of the characteristics of cells trajectories. By investigating these aspects, we can determine whether the cells exhibit Lévy walk behavior.

Main Results The research reveals that the cellular trajectories exhibit super-diffusive behavior early in the experiment, transitioning toward normal diffusion as time progresses along with cell density increases. Although initial MSD scaling exponent is consistent with Lévy walks, other statistical properties, such as displacement distribution and velocity auto-correlation, contradict the Lévy walks model. Instead, a mixture of two normal distributions provides a superior fit to the data.

Part II

Theoretical background

1 Mean Squared Displacement

Let $r(t)$ be a random variable in $\mathbb{R}^d$, representing the position of a particle at time t. Its initial position at time $t = 0$ is denoted by r_0. The displacement vector, which describes how the particle has moved from its starting point, is given by $\vec{r}(t) - \vec{r_0}$. This vector captures both the direction and magnitude of the motion. To quantify how far the particle has moved from its initial position, we compute the norm of the displacement vector

$$\|\vec{r}(t) - \vec{r_0}\|. \tag{1.1}$$

To simplify the analysis and avoid dealing with square roots, we instead use the squared norm. Taking the expected value of the squared norm across many trajectories yields the Mean Squared Displacement. The Mean Squared Displacement (MSD), given by

$$\left\langle \|\vec{r}(t) - \vec{r_0}\|^2 \right\rangle, \tag{1.2}$$

is a measure used in physics, particularly in the study of particle motion and diffusion. It quantifies the average squared distance between the particle's position at time t and its initial position. MSD is crucial for understanding Brownian motion, diffusion processes, and the dynamics of particles in various systems. The analysis of MSD provides a quantitative measure of how particles diffuse or spread out over time, making it essential for understanding particle dynamics in various fields. The MSD is important because it reveals a lot about the statistical properties of the process [6, 7, 9].

1.1 Diffusion

In the context of a particle trajectory, diffusion refers to the process by which particles spread out over time due to random motion [6]. It can be observed in various fields, such as physical, chemical, and biological systems. A common way to quantify diffusion is by measuring the MSD of the particles over time [4].

In one-dimension, we can describe a diffusion process by Fokker-Planck Equation (FPE), also known as diffusion equation. In the case of pure random fluctuations, without any drift, the FPE is given by

$$\frac{\partial}{\partial t}p\left(x,t\right)=D\frac{\partial^2}{\partial x^2}p\left(x,t\right), \tag{1.3}$$

where $p\left(x,t\right)$ is a probability density function, x denotes the position, and t is time. The solution is a given by

$$p\left(x,t\right)=\frac{1}{\sqrt{4\pi Dt}}e^{-\left(\frac{x}{\sqrt{4Dt}}\right)^2}, \tag{1.4}$$

which is a probability density function of a normally distributed random variable, whose variance grows linearly with time

$$x\sim\mathcal{N}\left(0,2Dt\right). \tag{1.5}$$

This means that as time progresses, the variance increases, implying that the process is non-stationary, as its statistical properties change over time. The parameter D is called the diffusion coefficient. It quantifies the rate at which the particles diffuse and is important for understanding how quickly particles will spread [4, 6]. The larger the value of D, the faster the dispersion. Next, consider the stochastic differential equation (SDE)

$$dx=\sqrt{2D}dB\left(t\right), \tag{1.6}$$

where $B\left(t\right)$ represents Brownian motion. This SDE describes the change in the particle's position. It can be proved that FPE and this SDE are equivalent, namely

$$\frac{\partial p}{\partial t}=D\frac{\partial^2 p}{\partial x^2}\iff dx=\sqrt{2D}dB\left(t\right), \tag{1.7}$$

meaning both equations describe the same process from different perspectives. Integrating the SDE

with respect to time t and computing the variance gives

$$\left\langle (x(t) - x_0)^2 \right\rangle = 2D \left\langle B^2(t) \right\rangle = 2Dt, \tag{1.8}$$

which is the same result obtained earlier from solving FPE. The left hand side is exactly what we defined as the MSD in one-dimension. Therefore, for simple diffusion, such as in the case of particles with trajectories subject to Brownian motion, we say that the MSD grows linearly with time. In higher dimensions we get

$$\left\langle \| \vec{r}(t) - \vec{r_0} \|^2 \right\rangle = 2dDt, \tag{1.9}$$

where d is the number of dimensions. Another way to present the relationship between MSD and time is

$$\left\langle \| \vec{r}(t) - \vec{r_0} \|^2 \right\rangle \propto t, \tag{1.10}$$

which shows that, for simple diffusion process which undergo Brownian motion, the MSD grows linearly with time for large t.

1.2 Anomalous Diffusion

Anomalous diffusion is common in biological environments like cells, where obstacles and active transport processes lead to non-standard particle motion [9, 10]. When referring to the concept of anomalous diffusion, we mean that the MSD is not linear in time. Instead, the asymptotic behavior of the MSD scales with time as

$$\left\langle \| \vec{r}(t) - \vec{r_0} \|^2 \right\rangle \propto t^{\gamma}, \tag{1.11}$$

where γ is a constant such that $\gamma \neq 1$. In the case where $\gamma = 1$ we come back to a normal diffusion process. There are several types of anomalous diffusion.

Sub-Diffusion This is a type of anomalous diffusion where particles spread out slower than expected based on normal diffusion models. In sub-diffusion, the MSD increases with time according to a power law where $0 < \gamma < 1$. In the context of biological systems such as cells, this type of diffusion is characterized by a non-linear relationship between the MSD and time. Specifically, for sub-diffusion, the MSD grows slower than linearly with time.

Super-Diffusion This is a type of anomalous diffusion where particles spread faster than they would under normal diffusion. The MSD scales as a power law where $1 < \gamma < 2$. Super-diffusion can occur in various physical systems and is often observed in complex systems where traditional Brownian motion does not adequately describe particle trajectories.

Ballistic Diffusion This is a type of anomalous diffusion which describes a physical phenomenon where particles move with constant velocity, leading to a MSD that scales quadratically with time, corresponding to $\gamma = 2$. In ballistic diffusion, particles exhibit straight-line motion without experiencing significant random scattering events typically observed in Brownian motion.

Hyper Ballistic Diffusion This type of anomalous diffusion represents an extreme form of particle motion, where the MSD increases faster than quadratically with time, corresponding to $\gamma > 2$. It is characterized by particles moving with continuously increasing velocity and persistent acceleration. In real physical systems, energy resources are finite, making such persistent acceleration impossible.

2 Models

In order to compare the experimental results with models and simulations, I focused on 3 different models which represent the range of diffusion behaviors observed in different systems: Langevin for normal diffusion, fractional Brownian motion for sub-diffusion, and Lévy walk for super-diffusion. Figure 2.1 shows simulation results for the MSD behavior of all three models on a log-log scale.

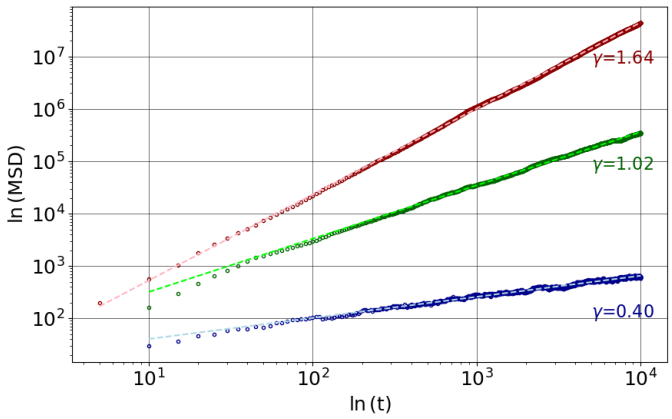


Figure 2.1: Simulations for three different models: Langevin dynamics (green), fractional Brownian motion (blue), and Lévy walk (red). Each simulation has 500 trajectories, total time of 10,000, and Δt equal 5. Additional parameters are: diffusion coefficient $D = 16$, Hurst parameter $H = 0.2$, power law exponent $\alpha = 1.3$, speed $s = 4$. For each model, I calculated the MSD. The figure presents a log-log plot of the results. On the x-axis is the logarithm of time, and on the y-axis is the logarithm of the MSD. The dots represent the actual simulated MSD values (after taking their log). The dashed lines are linear regressions of these values. From these regressions, I have extracted the slopes, which highlight the distinct diffusive regimes for each stochastic process:
• Langevin slope: 1.02 (theoretical value is 1).
• fBm slope: 0.4 (theoretical value is 0.4).
• Lévy walk slope: 1.64 (theoretical value is 1.7).

2.1 Langevin

According to Newton's second law, the net force on a body is equal to its mass multiplied by the body's acceleration

$$F(t) = m\dot{v}(t), \tag{2.1}$$

where m is the mass, and $\dot{v}(t)$ is the acceleration.

The Langevin model is a mathematical framework used to describe the dynamics of particles subjected to both deterministic and random forces [4, 5, 6, 7]. In the absence of an external deterministic force, the Langevin stochastic differential equation for a particle's velocity in 1D is given by

$$m\dot{v}(t) = -m\gamma v(t) + \eta(t). \tag{2.2}$$

The term $-m\gamma v(t)$ originates from friction, where γ is the friction coefficient, and $\eta(t)$ is a Gaussian

white noise with the following properties,

$$\langle \eta(t) \rangle = 0, \qquad \langle \eta(t)\, \eta(t') \rangle = \Gamma \delta(t - t'). \tag{2.3}$$

The function $\delta(t - t')$ is the Dirac delta function, and Γ is a non-negative constant, given by

$$\Gamma = 2m\gamma k_B T. \tag{2.4}$$

Here, k_B is the Boltzmann constant, and T is the temperature of the system.

Since the velocity $v(t)$ is the time derivative of the position $dx(t)/dt = v(t)$, we can find the displacement by integrating the velocity,

$$x(t) - x_0 = \int_{t_0}^{t} v(t')\, dt'. \tag{2.5}$$

Calculating the MSD and by the fluctuation-dissipation theorem [5, 7], we obtain the relation

$$\left\langle (x(t) - x_0)^2 \right\rangle = 2Dt. \tag{2.6}$$

This linear dependence of the MSD on time t characterizes normal diffusive behavior. The term D is called the diffusion coefficient, and is given by $D = k_B T/\gamma m$.

To generalize the above in 2D, let $r(t)$ be the position of a particle at time t, given by

$$r(t) = (x(t), y(t)), \tag{2.7}$$

where $x(t)$ and $y(t)$ are the Cartesian coordinates at time t, each subject to its own Langevin equation

$$m\dot{v}_x(t) = -m\gamma v_x(t) + \eta_x(t), \qquad m\dot{v}_y(t) = -m\gamma v_y(t) + \eta_y(t). \tag{2.8}$$

Here, $v_x(t)$ and $v_y(t)$ are the velocities along the axes. The terms $\eta_x(t)$ and $\eta_y(t)$ are independent Gaussian white noise processes with $\langle \eta_i(t) \rangle = 0$ and $\langle \eta_i(t)\, \eta_i(t') \rangle = 2m\gamma k_B T \delta(t - t')$, $i \in \{x, y\}$.

The corresponding MSD is then given by

$$\left\langle \left(r\left(t\right)-r_0\right)^2 \right\rangle = 4\frac{k_B T}{\gamma m}t = 4Dt. \tag{2.9}$$

2.2 Fractional Brownian Motion

As the name suggests, fractional Brownian Motion (fBm) is a generalization of Brownian motion [8]. It is a continuous Gaussian process with stationary increments. It is common to denote the fBm by $B_H\left(t\right)$, where H is called the Hurst parameter (also known as the Hurst index) where $0 < H < 1$. Unlike Brownian motion, the increments of fBm are dependent, except when $H = 0.5$. In that case, fBm reduces to Brownian motion with independent increments. The value of the Hurst parameter H determines the nature of the dependence.

- When $0 < H < 0.5$, the increments are negatively correlated, indicating anti-persistent behavior where increases (or decreases) are likely to be followed by decreases (or increases).
- When $H = 0.5$, the process behaves as Brownian motion, with no long-range dependence.
- When $0.5 < H < 1$, the increments are positively correlated, indicating persistent behavior where increases (or decreases) in the process are likely to be followed by further increases (or decreases).

Figure 2.2 presents simulation results of fBm for three different values of the Hurst parameter H. The auto-correlation function of fBm is given by

$$\left\langle B_H\left(t\right) B_H\left(s\right)\right\rangle = \frac{1}{2}\left(\left|t\right|^{2H} + \left|s\right|^{2H} - \left|t-s\right|^{2H}\right). \tag{2.10}$$

When $H = 1/2$, the process reduces to a standard Brownian motion. In that case the auto-correlation function is

$$\left\langle B_{1/2}\left(t\right) B_{1/2}\left(s\right)\right\rangle = \frac{1}{2}\left(\left|t\right| + \left|s\right| - \left|t-s\right|\right) = \min\left(s,t\right). \tag{2.11}$$

The MSD for fBm is given by

$$\left\langle B_H^2\left(t\right)\right\rangle = 2Dt^{2H}. \tag{2.12}$$

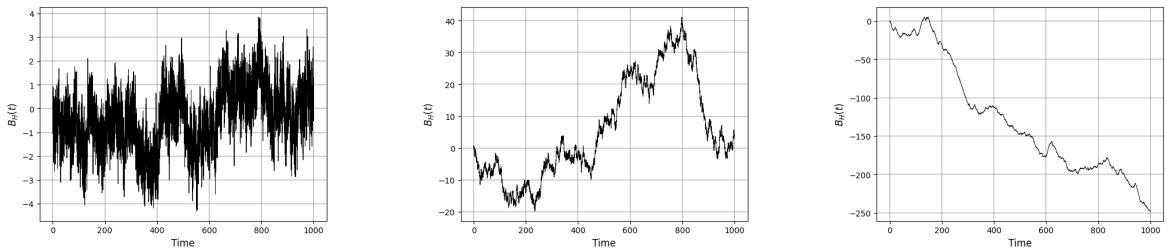
As a result, fBm can be used as a diffusion or anomalous diffusion model, with $0 < H < 0.5$ corresponding to sub-diffusion, $H = 0.5$ for normal diffusion, and $0.5 < H < 1$ for super-diffusion. This versatility of diffusion regimes can be used to model wide range of phenomena.

The fBM can be extended to 2D by considering $r(t) = (x(t), y(t))$ as the position of a particle at time t, where $x(t)$ and $y(t)$ are independent 1D fBm processes with the same Hurst parameter

$$x(t) = B_H^x(t), \qquad y(t) = B_H^y(t). \tag{2.13}$$

The MSD is then given by

$$\left\langle (r(t) - r_0)^2 \right\rangle = 4Dt^{2H}.$$



(a) $H = 0.1$ (b) $H = 0.5$ (c) $H = 0.9$

Figure 2.2: Three different simulations of fBm paths with varying Hurst parameters. (a) $H = 1$, negatively correlated increments, leading to slow spread and sub-diffusion. (b) $H = 0.5$, independent increments leading to normal diffusion. (c) $H = 0.9$, positively correlated increments, leading to fast spread and super-diffusion.

2.3 Lévy Walk

Before delving into the Lévy walk model, it is essential to first understand the concept of Lévy flight.

Lévy Flight This is a continuous-time random walk (CTRW), in which the duration of each step is constant, and the step length l is a random variable drawn from a power-law distribution [9, 10]. The probability density function $f(x)$ of a power-law is given by

$$f(x) = \begin{cases} Ax^{-(\alpha+1)} & x \geq x_{\min} \\ 0 & x < x_{\min} \end{cases}, \tag{2.14}$$

where A is a normalization constant and $x_{\min}$ is the lowest value that the random variable can take $x \in [x_{\min}, \infty)$ [11].

In one-dimensional space, the direction of movement (either left or right) is a random variable u which takes values in the set $\{-1, 1\}$ with equal probabilities. In two-dimensional space, the direction of movement can be represented by an angle θ, which is uniformly distributed $\theta \in [-\pi, \pi]$.

In order for the integral $\int_{x_{\min}}^{\infty} Ax^{-(\alpha+1)}dx$ to converge, we must have $\alpha > 0$. Depending on the value of α, the model exhibits different diffusion behaviors:

- For $0 < \alpha < 1$, the model behaves as ballistic diffusion where $\left\langle (x(t) - x_0)^2 \right\rangle \propto t^2$.
- For $1 < \alpha < 2$, we get super-diffusion with $\left\langle (x(t) - x_0)^2 \right\rangle \propto t^{\gamma}$ where $1 < \gamma < 2$.
- For $\alpha > 2$, it reduces to normal diffusion with $\left\langle (x(t) - x_0)^2 \right\rangle \propto t$.

In order to use Lévy flight as a super-diffusion model, we need $1 < \alpha < 2$, which leads to a divergent second moment

$$\left\langle x^2(t) \right\rangle = \int_{x_{\min}}^{\infty} x^2 \cdot Ax^{-(\alpha+1)}dx = \infty. \tag{2.15}$$

This means that the spread of step lengths from their average value is unbounded, allowing for very large (unbounded) steps. While Gaussian distributions also allow for very large values, their probability is extremely small. In contrast, power-law distributions admit very large values with a high probability.

Let $v = s\hat{u}$ be a velocity vector, where s represents the speed, and $\hat{u}$ is a unit vector indicating the direction of motion. Then, as the connection between length, velocity, and time is given by $l = v \cdot t$,

the velocity is proportional to the step length,

$$s\hat{u} \propto l. \tag{2.16}$$

This implies that the speed is unbounded. Very high speed can occur with significant probability. This unbounded speed characteristic presents a substantial issue when attempting to apply this model to real-world physical scenarios, as it does not account for the finite speed of actual objects. To address this limitation, a more realistic model known as the Lévy walk was introduced.

Lévy Walk Unlike the Lévy flight, where the speed of the process can theoretically become infinite, the Lévy walk constrains the speed to a constant value [9]. This adjustment makes the Lévy walk a more accurate representation of natural phenomena, as it better reflects the finite velocity at which physical objects can move. In a Lévy walk, the step length l is still drawn from a power-law distribution, ensuring that occasional long movements occur, but this step is now taken at a constant speed, thereby maintaining a more realistic model of motion in both one-dimensional (1D) and two-dimensional (2D) spaces. As the connection between length, velocity and time is given by $l = v \cdot t$, the time taken for each step is proportional to the step length,

$$t \propto l. \tag{2.17}$$

In the context of a power law, it is worth mentioning that some distributions may only exhibit a power law in their tail, for large enough values. In other words, some distributions behave as power laws asymptotically. Such a distribution can be expressed in the form,

$$f(x) = L(x)\, x^{-(\alpha+1)}, \tag{2.18}$$

where $L(x)$ is a function with two important properties:

1. The limit as $x \to \infty$ of $L(x)$ is a non-negative constant $c > 0$,

$$\lim_{x\to\infty} L(x) = c. \tag{2.19}$$

2. The limit as $x \to \infty$ of $L(ax)/L(x)$equals one,

$$\lim_{x\to\infty} \frac{L\left(ax\right)}{L\left(x\right)} = 1. \tag{2.20}$$

Simulation To simulate a one-dimensional Levy walk, we generate a sequence of walks, each defined by an isotropic random direction $d \in \{-1, 1\}$ (left or right) and a random walk duration drawn from a power-law distribution. A simple way to sample from a power-law distribution is through inverse transform sampling, where we draw a uniform random variable $u \in (0, 1)$ and apply the transformation

$$\tau = \tau_{\min} \left(1 - u\right)^{-1/\alpha}, \tag{2.21}$$

where $\tau_{\min}$ is the minimum possible duration of the walk, $\alpha > 1$ is the power-law exponent, and τ is the resulting walk duration. During each walk, we move at a constant speed s, and the position is updated by $\Delta x = d \cdot \tau \cdot s$.

In the case of two-dimensional simulation, the walk duration τ is generated in the same way. However, the direction is now defined by an angle $\theta \in [-\pi, \pi]$, drawn uniformly at random to ensure isotropic motion in the plane. We then moves a distance $\tau \cdot s$ in the direction specified by θ, resulting in a movement along the axes of $\Delta x = \tau \cdot s \cdot \cos\theta$ and $\Delta y = \tau \cdot s \cdot \sin\theta$.

Part III

Results

3 Experimental Data

The data used in this research was provided by Dr. Victor Yashunsky, whose work focuses in part on the dynamics of Human Bronchial Epithelial (HBE) cells, exploring their collective behavior and motion [2].

In Dr. Yashunsky's experiment, HBE cells were cultured on a glass plate, and their movements were tracked over time using the TrackMate tracking system. TrackMate is an image processing software designed for tracking particles, cells, or other objects in microscopy images and time-lapse image sequences. The user can extract various quantitative measurements such as displacement, velocity, and trajectory features.

The data set comprises detailed tracking information of HBE cells, where the time difference between each frame in the tracking data is 5 minutes. By analyzing this data set, the research aims to test whether HBE cell trajectories are statistically consistent with Lévy walks, contributing to the understanding of cell movement dynamics in a biological context. The analysis will involve calculating the mean squared displacement (MSD), comparison to diffusion and anomalous diffusion simulations, eliminating Gaussian process features, and examining the step length distributions to identify any power-law behavior indicative of Lévy walks.

3.1 Frame Images

The data provided by Dr. Victor Yashunsky includes 5 different sets of experiments, each with 800 frame images from a time-lapse sequence, showing the movement of HBE cells over time. Figure 3.1 shows magnified sections from two different frames of the experiment. The images display a dense field of cells, each with distinct outlines and internal structures. The cells vary in shape and size, appearing to be actively engaged in various stages of movement or interaction.

The high contrast in the image highlights the cell boundaries, making it easier to distinguish in-

dividual cells within the population. The time-lapse images enable a detailed observation of cell dynamics, providing insights into the collective behavior and interactions of the cells.

(a) Frame 1 (b) Frame 413

Figure 3.1: **Frame Image (set number 4).** In order to get a better look, I cut two different frames into 9 equal pieces each, and saved the top-left piece as a new frame. (a) and (b) are actually 1/9 of the original frames. Comparing the first frame with frame number 413, it can be seen that the later frame is much more dense and populated. This increase in cell density is due to cell splitting and multiplying. These observations highlight the dynamic nature of the biological system under study, showcasing the continuous changes and growth within the cellular environment.

3.2 Data Processing

Trackpy, a specialized tool for particle tracking, automates the extraction of cell movement data from the time-lapse images. This section outlines the key steps and functionalities provided by Trackpy in processing the data.

Cell Detection and initialization Using the *trackpy.locate()* function, cells are detected in each frame. This process identifies local maxima in a filtered version of the image, corresponding to the centroid of each detected cell. Adjustable parameters, such as particle diameter and detection threshold, ensure accurate identification of cells while minimizing false positives from noise or background pixels.

Linking Cells Across Frames After detecting cells in individual frames, Trackpy links their positions across frames to form continuous trajectories. The key functions used are:

- *link()*: Connects particles between consecutive frames based on proximity.
- *link_df_iter()*: Iteratively handles multi-frame tracking, with parameters like *search_range* (maximum distance for linking) and *memory* (number of frames to backtrack for reconnecting lost particles).

These linking processes result in the creation of continuous trajectories for each cell, representing their movement over time.

Trajectory Construction and Data Storage The trajectories are stored in a structured PANDAS DATAFRAME, where each row corresponds to a cell's position at a specific frame. Key columns include:

- Frame number: Timestamp of the position.
- (x, y) coordinates: Cell centroid position in the frame.
- Trajectory ID: Unique identifier for each cell's trajectory.

Trackpy supports efficient storage of large datasets in HDF5 format, enabling streamlined access and manipulation for further analysis.

Handling Tracking Challenges Trackpy addresses several challenges in tracking:

- Overlapping particles: Algorithms minimize errors by considering proximity and motion patterns.
- Cell loss: The memory parameter reconnects lost cells across frames.
- Division or fusion events: Although explicit handling of these events is limited, the *memory* parameter helps maintain trajectory continuity.

Exporting Data The resulting trajectory data can be exported in various formats, including PANDAS DATAFRAME, HDF5, and pickle files, facilitating comprehensive analysis. By automating these processes, Trackpy enables efficient large-scale studies of cell dynamics, including assessments of anomalous diffusion behaviors, such as Lévy walks. Table 1 shows a sample of the extracted trajectories data, while Figure 3.2 illustrates cellular trajectories built from the data.

Cell	x	y	Frame
473	343.416	375.762	0
802	816.711	648.121	0
1742	206.706	1470.011	0
1779	171.522	1498.575	0
1813	171.385	1524.101	0

(a)

Cell	x	y	Frame
473	344.207	376.523	1
802	817.865	645.244	1
1742	205.5334	1475.842	1
1779	174.738	1495.427	1
1813	172.344	1520.924	1

(b)

Table 1: **Cell Coordinates Across Frames (set number 4).** The tables display the Cartesian coordinates of five tracked cells in two consecutive frames - frame 0 (a) and frame 1 (b). Each row corresponds to a specific cell, identified by its cell number, and shows the cell's centroid position in a 2D coordinate system. The change in coordinates between the two frames illustrates the displacement of each cell over a 5 minute interval. Each pixel in the frame corresponds to 0.74 microns.

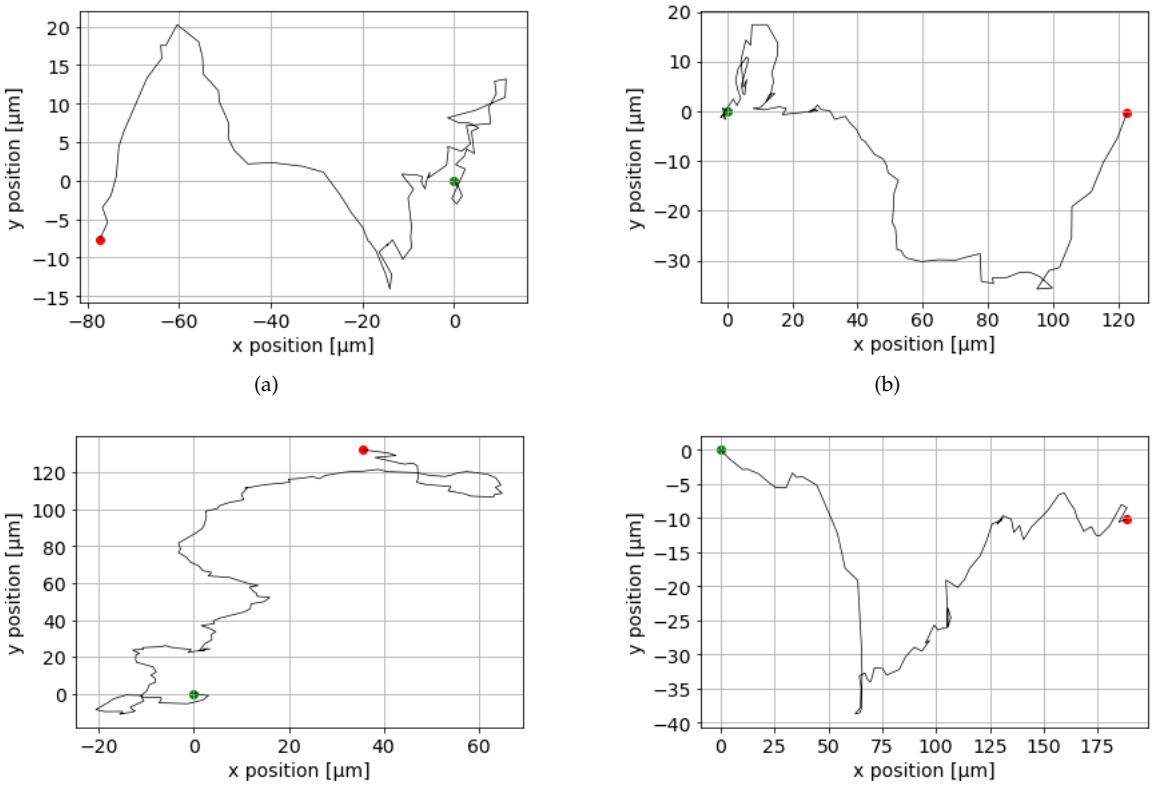


Figure 3.2: **Cell Trajectories (set number 4).** The plot presents four randomly selected cell trajectories. Each black line represents the trajectory of an individual cell, tracked over time. The green dot indicates the starting point of the trajectory, while the red dot marks the endpoint. These trajectories illustrate the diverse movement patterns and displacements of the cells. The long, relatively straight extrusions made us suspect that the trajectories are super-diffusive, consistent with Lévy walks.

4 Data Analysis

4.1 Diffusion Regime

To characterize the diffusion regime of cells, I analyzed their MSD. To ensure reliable analysis, only cell trajectories constructed from at least 75 frames were included, ensuring that sufficient tracking data was available. This threshold was chosen to balance between sample size and trajectory quality. Other thresholds were also tested and yielded similar results, confirming that the conclusions are not

sensitive to this specific cutoff.

For each individual cell, I gathered the (x, y) coordinates of its position across the frames where it appeared, and its trajectory was plotted. Figure 4.1 displays numerous trajectories plotted together. Initially, the time a which the trajectories were observed in the experiment was not taken into consideration, only the absolute trajectory and the total displacement over time were analyzed. For example, one cell's trajectory might span from frame 1 to frame 75 (representing the initial stage of the experiment), while another cell's trajectory might span from frame 328 to frame 427 (occurring during the middle of the experiment). Despite these differences, the specific timing of the trajectory within the experimental timeline was disregarded, and only the number of frames over which the trajectory was observed was considered. This approach allowed me to calculate the MSD across a sufficiently large number of trajectories while ensuring that each trajectory covered a meaningful number of frames. by doing so, I could examine the MSD across a sufficient range of time points, irrespective of when during the experiment the individual trajectories occurred.

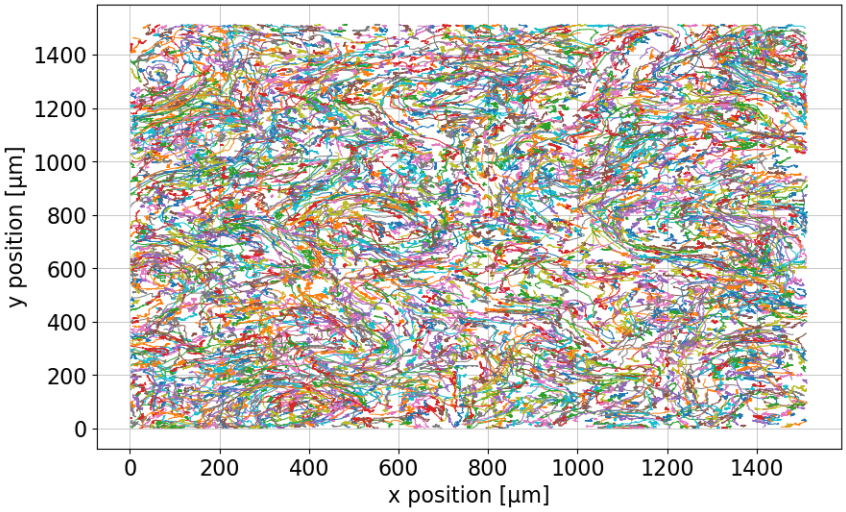


Figure 4.1: **Trajectories (set number 4).** The set consists of 367,673 individual cells, of which 18,092 appear in at least 75 frames. To maintain clarity in visualization, only 4,523 randomly selected trajectories are plotted.

Calculating the MSD The investigation of the diffusion regime began with analyzing the trajectories of individual cells to determine how their movement evolved over time. Each trajectory was examined

step by step, with k representing the number of steps taken from the initial position. Since the initial frame varied between different cells, $k = 0$ corresponded to the first recorded position of a given cell, with subsequent steps indexed accordingly.

The position data, extracted as (x, y) coordinates, were used to compute the displacement at each step. At each step k, the squared displacement was calculated for every cell. These values were then averaged over the entire ensemble to obtain the MSD for step k. The process was repeated for 74 steps to capture the overall MSD behavior. Figure 4.2 presents the results of this analysis.

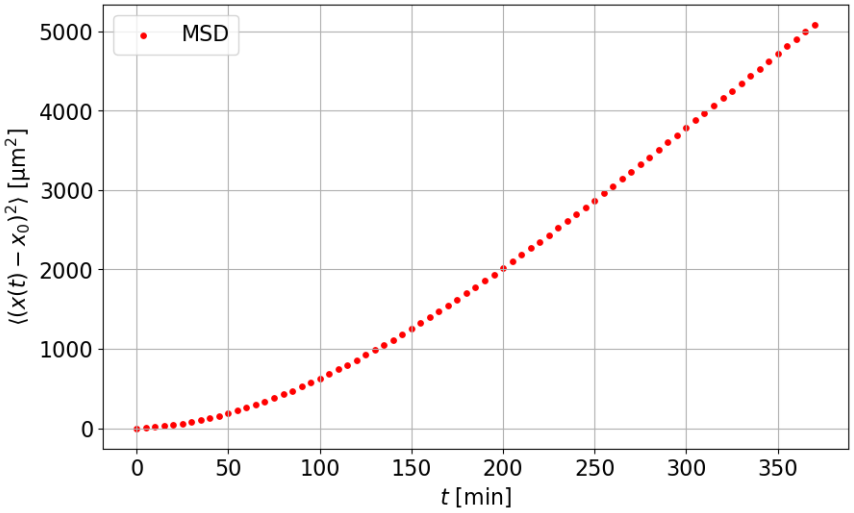


Figure 4.2: **MSD (set number 4).** The plot presents the MSD for cells that appears in at least 75 frames. The x-axis represents time, indicating the progression of the experiment. The y-axis denotes the mean squared displacement value.

Log-Log MSD To characterize the diffusion regime, I examined the relationship between MSD and time using a log-log transformation. This approach transformed the MSD equation into a linear form, where the slope of the resulting line determined the diffusion exponent γ. A linear regression analysis was then performed to estimate this exponent. The results indicates that the cell trajectories followed a super-diffusive behavior with γ in the range of 1.78 to 1.80.

Since cell movement was influenced by dynamic interactions such as collisions, splitting, and environmental factors, I further analyzed whether the diffusion behavior changed over time. To investigate

this, I divided the trajectory data into two segments: the first half and the second half of the experiment. By computing the MSD separately for each segment and applying the same log-log transformation and regression analysis, I obtained diffusion exponents for both phases. The results showed that the diffusion exponent was higher in the first half compared to the second half. Figure 4.3 provides a pictorial overview of the results, while Table 2 presents the numerical analysis for five different experimental sets.

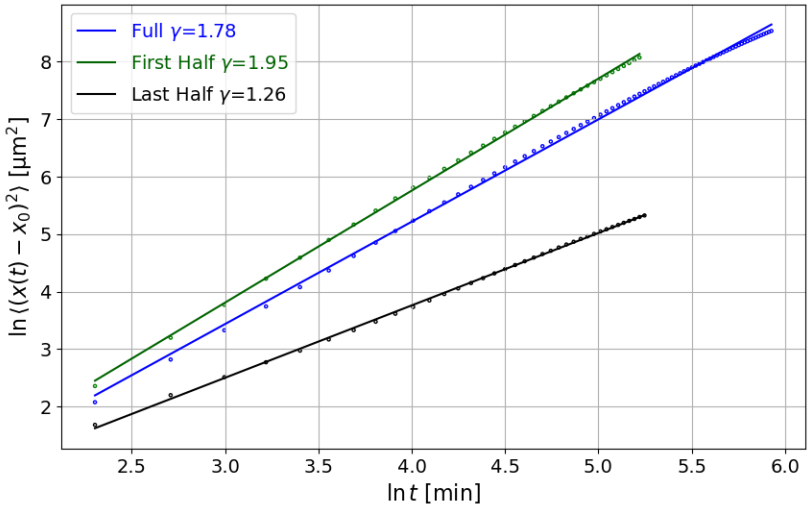


Figure 4.3: **Log-Log of MSD (set number 4).** The plot presents MSD vs. time on a log-log scale, analyzed for the full data set, first half, and last half of the trajectories data. The plot includes three different fits: Full trajectory (blue): The entire data set, with a scaling exponent $\gamma = 1.78$, suggesting super-diffusion. First half (green): The initial part of the trajectories, showing a higher exponent $\gamma = 1.95$, suggesting ballistic behavior. Last half (black): The later part of the trajectories, with a lower exponent $\gamma = 1.26$, indicating a transition towards normal diffusion.

Set	No. Cells	Full	First half	Second half
1	17,197	1.80	1.96	1.16
2	15,732	1.78	1.95	1.14
3	16,874	1.79	1.95	1.25
4	18,092	1.78	1.95	1.26
5	17,709	1.79	1.96	1.27

Table 2: **Scaling Exponent.** The table presents the scaling exponents γ obtained from the log-log MSD analysis for cells appearing in at least 75 frames across five different experiment sets. No. Cells: Refers to the total number of cells included in the full calculation for each set. Full: Represents the diffusion exponent γ when considering the entire trajectory data. First Half: Shows the exponent calculated from the first half of the trajectories, generally exhibiting a higher value, indicating a more ballistic behavior. Second Half: Corresponds to the latter part of the trajectories, where the exponent drops significantly, suggesting a shift toward normal diffusion. Since the first and second halves are analyzed separately, the number of contributing cells in each half is half of the total number listed in the table.

This finding suggested that cell movement was initially more active but slowed down over time. In the early stages of the experiment, cells exhibited a higher degree of displacement, possibly due to lower cell density and fewer interactions, allowing them movement that seems ballistic. As time progressed, increased collisions, crowding effects, or changes in cellular behavior may have contributed to the reduction in the diffusion exponent. Despite this decrease, the overall super-diffusive nature of the movement remained evident throughout the experiment.

4.2 Displacement Distribution

According to the generalized central limit theorem, the sum of i.i.d random variables with a divergent second moment converges to a Lévy alpha-stable distribution [4, 9, 10]. This suggests, that for a cell undergoing a Lévy walk with an exponent in the range $1 < \alpha < 2$, the displacement should follow a symmetric Lévy alpha-stable distribution. Further more, it implies that the components across the axes should follow a symmetric Lévy alpha-stable distribution as well.

To analyze the displacement distribution, for every step k I computed the displacement components of

the j-th cell by

$$\Delta x_j(k) = x_j(k) - x_j(0), \qquad \Delta y_j(k) = y_j(k) - y_j(0), \tag{4.1}$$

where $\Delta x_j(k)$ and $\Delta y_j(k)$ are the displacement components. This process provided displacement values for each step across different trajectories. I then calculated the average displacement at each step k, and used a kernel density estimation to plot the distributions. Figure 4.4 shows the resulting displacement densities at three different time intervals of the experiment.

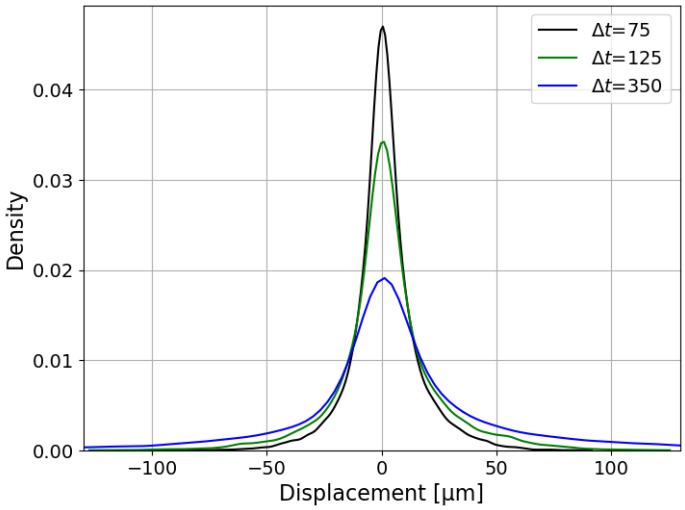
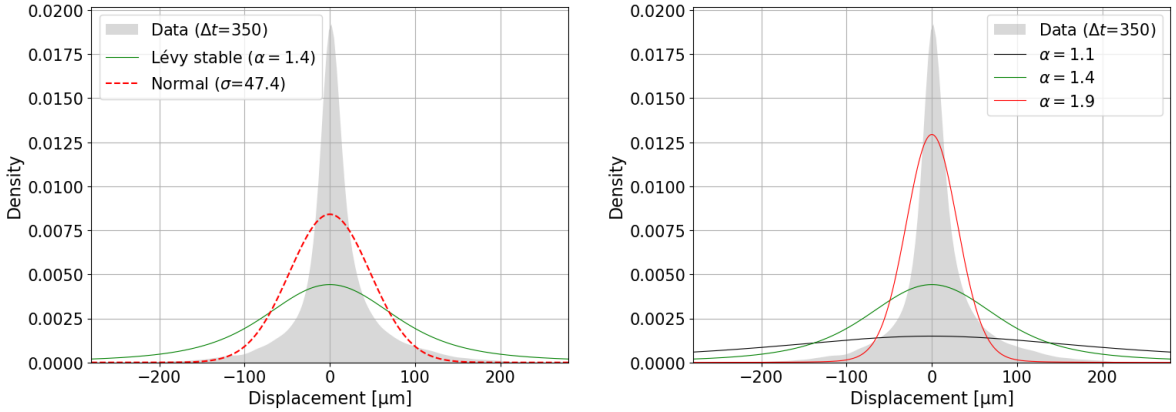


Figure 4.4: **Displacement Densities (set number 4).** The figure presents displacement densities along the x-axis at different time intervals of the experiment**: black** - 75 minutes**, green** - 125 minutes, **blue** - 350 minutes. The results for the y-axis, which aren't plotted here, showed similar results.

To verify whether the resulting densities align with the theory, I attempted fitting a symmetric Lévy alpha-stable distribution to the displacement densities across the axes. As the scaling exponent should be $1/\alpha$, I tested several values for α, varying between $1 < \alpha < 2$. The resulting plots showed no fit, suggesting that, although the cells exhibit super-diffusion, their steps do not subject to Lévy walk.

Next, I've attempted to fit a normal distribution, but again, the results showed no fit, indicating that the displacement distribution doesn't follows a normal diffusion either. Figure 4.5 presents the fitting results for the 350 minute time interval.



(a) (b)

Figure 4.5: **Densities Fitting (set number 4).** Both figures present the displacement density of the x-axis after 350 minutes (gray), along with corresponding fitting curves. (a) The figure shows the fitting attempts of a Lévy alpha-stable distribution with $\alpha = 1.4$ (green), and a normal distribution with $\sigma = 47.4$ (red). (b) The figure presents the resulting Lévy alpha-stable fitting with different α values: $\alpha = 1.1$ (black), $\alpha = 1.4$ (green), $\alpha = 1.9$ (red). Similar results were observed for the displacement along the y-axis.

4.3 Velocity Auto-Correlation Function

According to the Lévy walk theory, the velocity assumed to be stationary with auto-correlation function (ACF) that behave as $C(\Delta t) \propto \Delta t^{-\alpha}$. To test whether the experimental trajectories velocity behave as the theory, I first applied the Augmented Dickey-Fuller (ADF) test to check for stationarity in both the velocity's components along the x and y axes, for the first and second moments. Next, I computed the velocity auto-correlation function and tested whether it follows a power-law decay.

Velocity Calculation At each time step k, I computed the velocity components for the j-th cell by

$$v_x^j(k) = \frac{x^j(k+1) - x^j(k)}{\delta t}, \qquad v_y^j(k) \frac{y^j(k+1) - y^j(k)}{\delta t} \tag{4.2}$$

where $v_x^j(k)$ and $v_y^j(k)$ denotes the velocity components of the j-th trajectory at step k, and $\delta t = 5$ minutes, representing the time interval between consecutive frames. By applying these calculations to each trajectory, I obtained a time series of velocity components for every individual cell.

Stationarity Analysis To verify whether the velocity can be considered as a weak sense stationary process, I applied the ADF test in two different ways:

1. Test on Individual Trajectories: I performed the ADF test on each trajectory's velocity time series separately.
2. Test on mean velocity time series: I computed the mean velocity across all trajectories at each step k. I then applied the ADF test to this mean velocity series to assess overall stationarity.

Table 3 summarizes the results of the stationarity test applied to the velocity time series of individual trajectories, while Table 4 and Figure 4.6 present the results of the stationarity test applied to the mean velocity time series.

Moment	First		Second	
Axis	x	y	x	y
Stationary	14,726	14,781	16,306	15,331
Non-stationary	3,366	3,311	1,786	2,761
Stationary %	81.40%	81.70%	90.13%	84.74%
Average ADF	-7.0068	-6.7018	-7.2127	-6.6790
Average p-value	5.43%	5.30%	3.51%	5.65%

Table 3: **Test on Individual Trajectories (set number 4).** The tables presents results of the ADF stationarity test for the first and second moments, applied to the velocity data. Stationary & Non-Stationary: The number of trajectories classified as stationary and non-stationary based on the ADF test. % Stationary: The percentage of trajectories identified as stationary. Average ADF: The mean value of the ADF test statistic across all trajectories. Average p-value: The mean p-value from the ADF test across all trajectories. A p-value less than 5% indicates that the null hypothesis is rejected, suggesting that the time series is stationary.

Moment	First		Second	
Axis	x	y	x	y
ADF	-4.664	-3.900	-5.705	-5.012
p-value	0.0001	0.0020	0.0000	0.00002

Table 4: **Test on Mean Velocity Time Series (set number 4).** The table presents results of the ADF stationarity test for the first and second moments, applied to the velocity data.

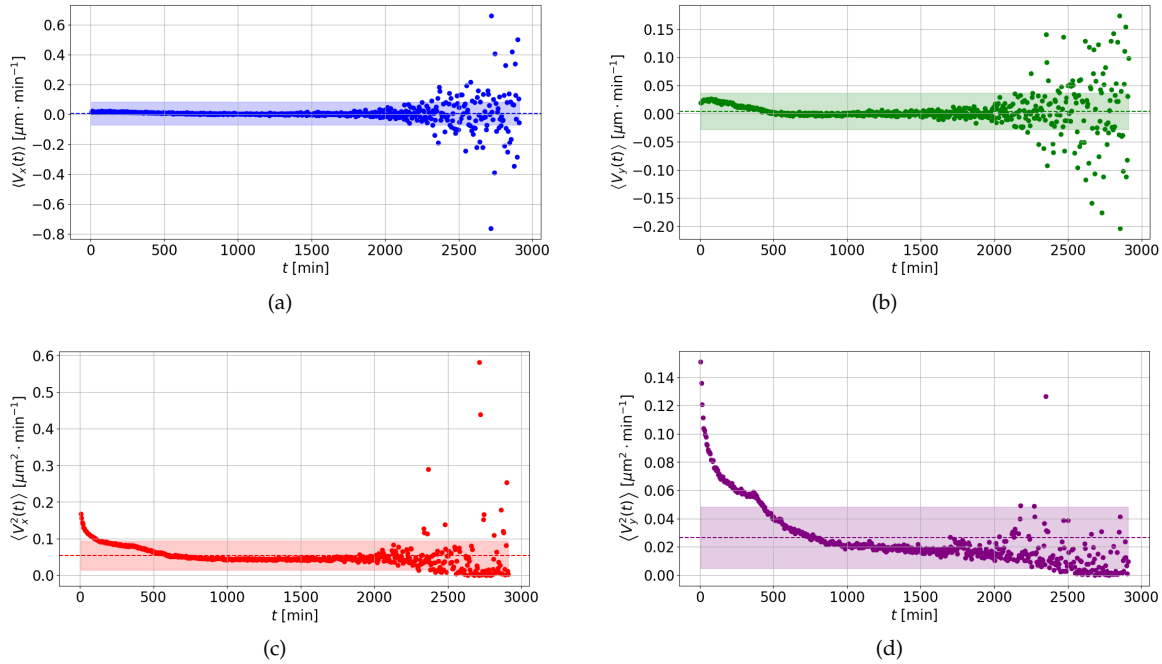


Figure 4.6: **Test on Mean Velocity Time Series (set number 4).** Analysis of the velocity's first and second moments. Average of the first and second moments components for the velocity as a function of time. The dots represent the average value of the moment and the dashed line is the overall mean. the shaded area represents $\pm$ standard deviation.When computing averages for larger time t, fewer trajectories contribute to the calculation. This reduction in sample size leads to increased variability in the estimates, which is visible in the plots. At smaller time t where more trajectories are included, the mean and second moment appear stable. However, as t increases and fewer data points remain for averaging, the estimates become more dispersed, introducing greater statistical fluctuations.

The results obtained from both tests implies that the velocity's first and second moment components can be considered as stationary.

ACF Decay After confirming that the velocity can be treated as a weak sense stationary process, I tested whether the ACF decays as a power law. To do so, I first computed the ACF at different lags by

$$C(\Delta t) = \frac{\left\langle \vec{v}(0)\,\vec{v}(\Delta t)\right\rangle - \left\langle \vec{v}(0)\right\rangle^2}{\left\langle \vec{v}^2(0)\right\rangle - \left\langle \vec{v}(0)\right\rangle^2}, \tag{4.3}$$

where Δt denotes the time lag. The maximum Δt was determined by the shortest trajectory in the data set. Since I examined trajectories with at least 75 frames and the time between frames is $\delta t = 5$, the maximum lag is 370. Figure 4.7 shows the resulting graph.

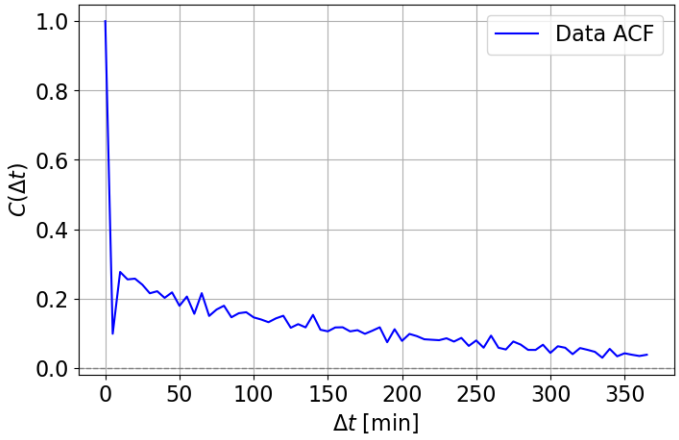


Figure 4.7: **Velocity ACF vs. Time Lag (set number 4).** The x-axis represents the time lag Δt. The y-axis shows the value of the velocity auto-correlation function.

Next, according to the theory, the velocity ACF follows a power law

$$C\left(\Delta t\right) \propto \Delta t^{-\alpha}. \tag{4.4}$$

By applying a log-log transformation, this relationship becomes linear in $\log \Delta t$

$$\log C\left(\Delta t\right) = -\alpha \log \Delta t + \log b. \tag{4.5}$$

Therefore, after computing the ACF for different lags, I performed a log-log transformation, followed by a linear regression and an R^2 analysis. Figure 4.8 shows the resulted linear regression and R^2 value.

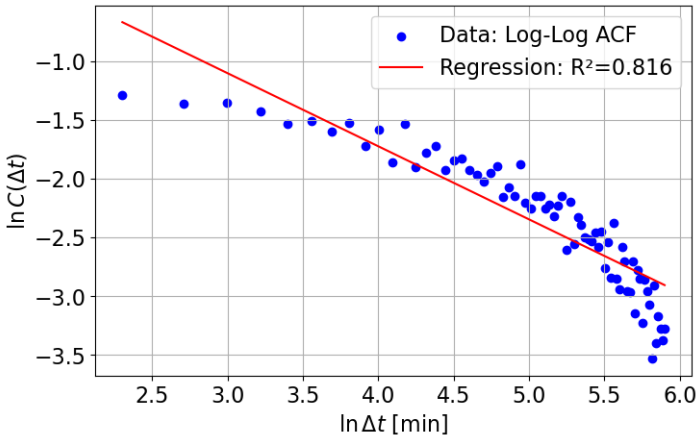


Figure 4.8: **Velocity Log-Log ACF (set number 4).** The x-axis represents $\log \Delta t$, and the y-axis represents $\log C(\Delta t)$. The blue dots correspond to the computed ACF values at different time lags after applying a log transformation. The red line represents the linear regression fit to the data. If the velocity ACF followed a power-law decay, the blue dots would align well with the red line. However, the spread in the data and the R^2 value suggest that the power-law model does not provide a perfect fit.

However, the results didn't aligned well with a power law decay, as predicted by the Lévy walk model. To further investigate, I tested for an exponential decay

$$C(\Delta t) \propto e^{-\alpha \Delta t}, \tag{4.6}$$

which appears linear in Δt after a log-log transformation

$$\log C(\Delta t) = -\alpha \Delta t + \log b, \tag{4.7}$$

and repeated the linear regression with an R^2 analysis. Figure 4.9 shows the improved fit.

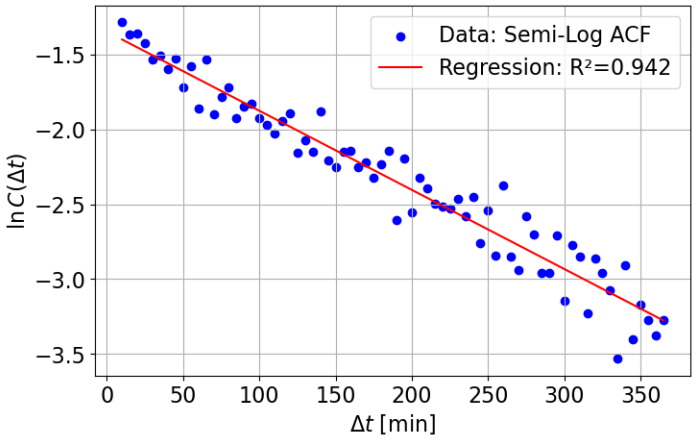


Figure 4.9: **Velocity Semi-Log ACF (set number 4).** The x-axis represents Δt, and the y-axis represents $\log \Delta t$. The blue dots correspond to the log transformed ACF values at different time lags. The red line represents the linear regression fit, which appears to align well with the data. The higher R^2 value compared to the log-log plot suggests that an exponential decay model provides a better fit than a power-law decay.

The improved R^2 value confirmed that the velocity ACF does not follow a power law decay but instead exhibits an exponential decay. Table 5 presents the R^2 results comparison for the different experimental sets.

Set	log-log R^2	semi-log R^2
1	0.8336	0.9865
2	0.8488	0.9868
3	0.6979	0.8178
4	0.8165	0.9425
5	0.8090	0.9510
Average	**0.8017**	**0.9369**

Table 5: **Linear Regression's R^2 Values.** The table shows the R^2 values obtained from linear regression analyses, performed on both log-log and semi-log transformed data, for different experiment sets. The results shows that the exponential decay model provides a much better fit for the velocity ACF decay behavior compared to the power law decay.

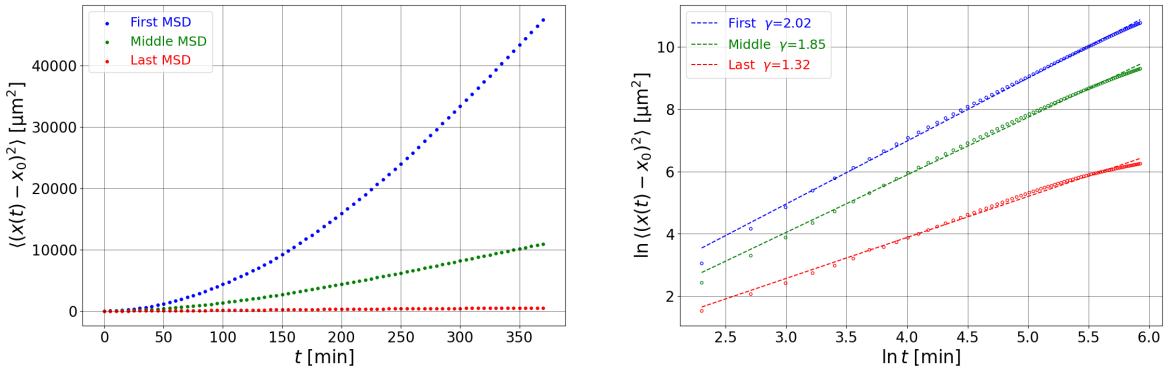
The results implies that the velocity ACF isn't decaying as power-law, but rather exponentially. By so, it contradicts the initial assumption that the HBE cells follow a Lévy walk.

4.4 Segment Analysis

Upon reviewing the initial analysis, it became apparent that the data did not align with the expected behavior of Lévy walks across the entire data set. To better understand the temporal evolution of the diffusion properties, I decided to break the data into three distinct segments: first, middle, and last. This approach allowed me to investigate potential changes in the diffusion regime over time, treating each segment as an independent experiment.

4.4.1 MSD

For each segment, I selected only cell trajectories that lasted at least 75 frames to ensure robust statistical analysis. I then repeated the same analytical steps to each segment separately, including calculating the MSD, applying a log-log transformation, and performing a linear regression to determine the scaling exponent γ. Figure 4.10 provides a visual comparison across the three segments. In addition, I calculated the number of cells at each segment.



(a) (b)

Figure 4.10: **Segment Analysis of MSD (set number 4).** The figure shows the difference between the three segments. (a) MSD plot. The x-axis represents time, and the y-axis is the mean squared displacement. The curves correspond to different segments of the experiment. The first segment (blue) shows the highest values, followed by the middle segment (green) and the last segment (red). (b) Log-Log MSD. The x-axis represents the log of time, and the y-axis represents the log of MSD. The dots are the experimental data, and the dashed lines are the linear regression fitting, with slopes of $\gamma = 2.02$ for the first segment, $\gamma = 1.85$ for the middle, and $\gamma = 1.32$ for the last, showing decreasing in diffusion as the experiment progresses.

The results showed notable differences between the three segments, suggesting that the statistical

Segment	First		Middle		Last	
set	no. cells	γ	no. cells	γ	no. cells	γ
1	261	1.99	4,027	1.86	12,291	1.24
2	148	2.01	2,966	1.9	12,011	1.26
3	250	2.01	4,332	1.87	12,121	1.33
4	266	2.02	5,113	1.85	12,742	1.32
5	299	2.00	4,731	1.86	12,514	1.32
Average	**244.8**	**2.006**	**4,233.8**	**1.868**	**12,335.8**	**1.294**

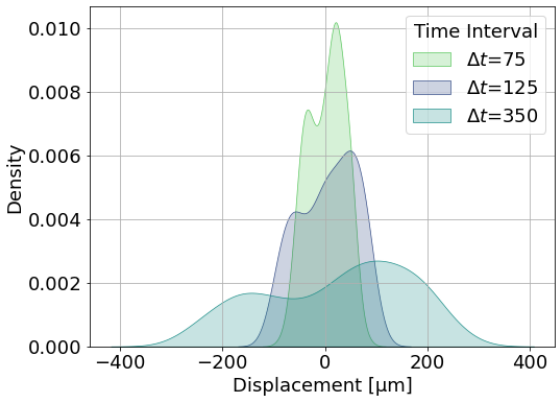
Table 6: **Segment Analysis of Scaling Exponent.** The table shows the number of cells and the diffusion scaling exponent γ across three experimental segments: first, middle, and last. Initially, the number of cells is relatively low (244.8 on average), but it increases significantly over time, reaching 12,335.8 on average in the last segment, due to cell splitting. At the same time, the diffusion exponent γ decreases from 2.0006 (average) in the first segment (indicating super-diffusive motion) to 1.29 (average) in the last segment, suggesting a transition toward normal diffusive behavior. This trend implies that as cell density increases, crowding and collisions limit movement, leading to reduced displacement.

properties of the trajectories are changing over time. The results make intuitive sense, as the scaling exponent was approximately $\gamma = 2$ (ballistic) in the early stage with fewer cells, and decreasing towards $\gamma = 1$ (normal diffusion) in the later stage as cell density increased. This shift from ballistic to normal can be attributed to the effects of growing density and crowding, which lead to more frequent collisions and, consequently, a reduction in movement. Table 6 summarizes the numerical results for each segment across five different experimental sets.

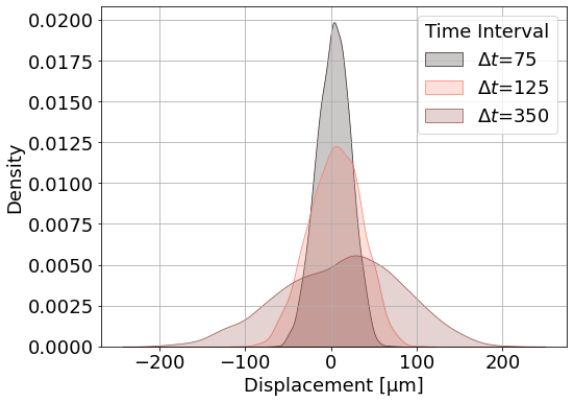
4.4.2 Displacement Analysis

In addition to the MSD analysis, I performed a density analysis to the three segments, in order to further investigate the temporal evolution of cell movement. For each segment, I selected the same three time intervals relative to its starting point: $\Delta t = 75, 125, 350$ minutes. These time intervals represent different stages within each segment, allowing for a detailed comparison of movement patterns over time. For each selected time interval, I calculated the displacement distribution and plotted it using a kernel density estimation. This resulted in several density distributions, enabling two types of comparisons: within a segment at different time intervals, and across segments at the same time interval. Figure 4.11 shows the displacement distributions within each segment across different time intervals, while Figure 4.12 compares the displacement distributions across segments at a fixed

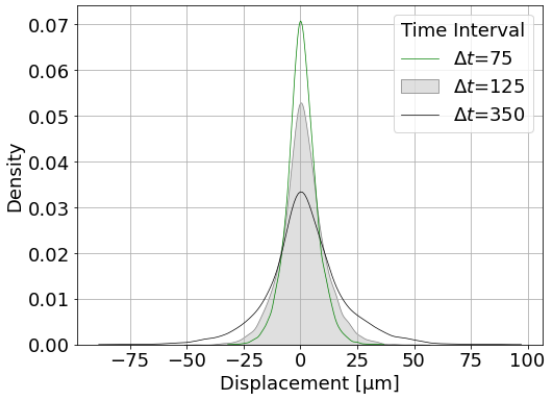
time interval of 350 minutes.



(a) First Segment



(b) Middle Segment



(c) Last Segment

Figure 4.11: **Displacement Distribution (set number 4)**. The figure presents the displacement distributions along the x-axis across three different segments of the experiment: first (a), middle (b), and last (c). Within each segment, the same time intervals were sampled: $\Delta t = 75, 125, 350$ minutes. It can be seen that, as the experiment progresses, the distributions become narrower, reflecting a decrease in cell movement. Comparing the same time interval across the different segments shows a clear transition in diffusion behavior, from broader heavy tailed distributions in the early segment, to narrower and more concentrated distributions in the later stages - indicating a reduction in motility.

The density plots revealed distinct differences in the shape of the displacement distributions across segments, for the same time interval. In the first segment, the distribution was broad and spread out, with a notable presence of large displacements. At the middle and last segments, the distributions became narrower, reflecting a transition towards what seems as normal diffusion. This shift is likely

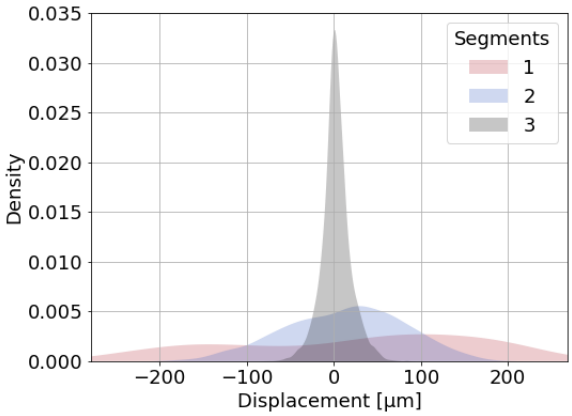
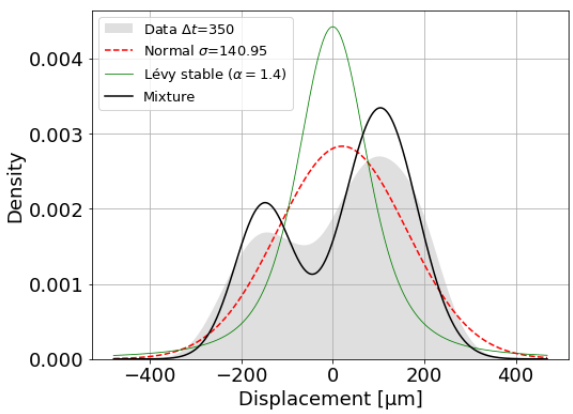


Figure 4.12: **Displacement Distributions Across Segments (set number 4).** The figure presents the kernel density estimation of displacement for the first (pink), middle (purple), and last (gray) segments over time interval of 350 minutes. It can be seen that the width and height of the densities are quit different, indicating different diffusion regime for each segment.

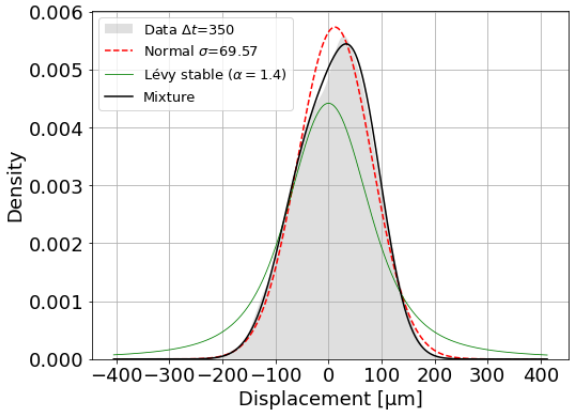
due to increased crowding and cell colliding, which constrain movement and reduce the occurrence of large jumps.

Density Fitting Next, within each segment, I attempted fitting a symmetric Lévy alpha-stable distribution across the axes. As before, I've done so with several values of α, varying in the range $1 < \alpha < 2$. However, none of the tested values provided a good fit to the data. As the density distributions became (seemingly) Gaussian-like over time, I then attempted to fit a normal distribution. The results showed that, while the normal distribution was not a good fit in the first segment, it provided a better approximation in the middle and last segments. Given these findings, I considered the possibility that the displacement distribution is more complex in nature and requires a more flexible model for accurate fitting.

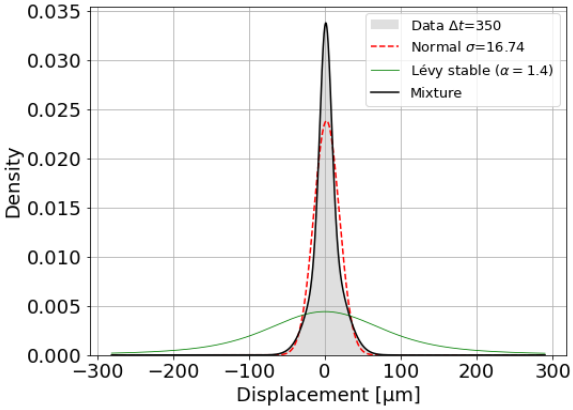
To address this, I applied a mixture of two normal distributions, aiming to capture both the broad spread and the pronounced peaks observed in the data. The mixture model was fitted to the displacement data using a maximum likelihood approach. The results showed that the mixture model provided a significantly improved fit compared to both the single normal and Lévy alpha-stable distributions. Figure 4.13 shows the fitting results across the three segments at $\Delta t = 350$ minutes, comparing the empirical data with the fitted models.



(a) First Segment



(b) Middle Segment



(c) Last Segment

Figure 4.13: **Density Fitting (set number 4).** Comparison of displacement densities across different segments at the same time interval of $\Delta t = 350$ minutes. Each plot presents the empirical data (gray), fitted normal distribution (dashed red), Lévy alpha-stable with $\alpha = 1.4$ (green), and a mixture of two normal distributions (black). It can be seen that the mixture model consistently provides the closest fit to the data across all segments.

Examining the density plots, it is evident that the normal and Lévy alpha-stable distributions fail to capture the full complexity of the empirical data, particularly in the first and last segments. In contrast, the mixture model closely aligns with the observed distributions, effectively capturing the heterogeneity in cell movement and providing a more accurate representation of the underlying motility dynamics.

Akaike Information Criterion Test To quantitatively assess which model best describes the displacement distributions across the axes, I applied the Akaike Information Criterion (AIC)[12], given by

$$\mathrm{AIC} = 2k - 2\log\mathcal{L}, \tag{4.8}$$

where k is the number of parameters of the fitted model, and $\mathcal{L}$ is the log-likelihood of the distribution. It evaluates model quality based on both goodness of fit and model complexity, helping to determine the most appropriate statistical representation of the observed data.

For each segment I fitted a symmetric Lévy alpha-stable distribution, Gaussian distribution, and Mixture of two normal distribution. The Lévy alpha-stable is characterized by four parameters: stability index α, skewness β, location μ, and scale c. Since the focus was on symmetric distribution, β and μ were set to zero, leaving $k = 2$. Similarly, the Gaussian distribution, which is characterized by its mean μ and standard deviation σ, also resulted in $k = 2$. The Mixture of two normal distribution required five parameters: two means μ_1, μ_2, two standard deviations σ_1, σ_2, and a weight w_1 (as w_2 is given by $1 - w_1$), resulting in $k = 5$.

AIC values were computed for each model across different time intervals and segments. The model with the lowest AIC was considered the best fit for the displacement data. The results consistently showed that the Mixture of two normal distributions had the lowest AIC across all segments and time points, making it the most suitable model. Table 7 presents the AIC results obtained across different segments and time intervals from one selected experimental set.

Segment	Time	Lévy Stable	Gaussian	Mixture
1	75	2,955	2,819	2,790
	125	3,243	3,064	3,030
	350	3,822	3,551	3,501
2	75	49,184	47,965	47,946
	125	53,667	52,557	52,520
	350	62,084	60,974	60,898
3	75	111,926	92,544	92,092
	125	121,179	101,530	100,871
	350	139,406	115,650	114,329

Table 7: **Akaike Information Criterion (set number 4).**The table presents AIC values for three different models: Lévy stable, Gaussian, and Mixture of two normal distribution, applied to the displacement distributions along the x-axis. The scaling exponent used was $\alpha = 1.4$. Results are shown across three segments and at different time intervals $\Delta t = 75, 125, 350$ minutes. In every case, the Mixture model yielded the lowest AIC, indicating it provided the best fit for the data. The y-axis showed similar results.

This suggests that the displacement distribution is more complex than can be captured by a single, uni-modal distribution. Instead, it likely reflects the presence of multiple dynamic populations or movement modes, such as cells moving in opposite directions, or a combination of slower and faster movement behaviors. This might explain the requirement of a more flexible, multi-component model to accurately describe the data.

4.5 Moving Window Analysis

It became evident that the diffusion regime is not constant over time, but rather exhibits dynamic transitions throughout the course of the experiment. To investigate this temporal heterogeneity in greater detail, I implemented a floating window analysis. This approach involved dividing the experiment into several overlapping segments, allowing a more detailed assessment of the dynamics. Each floating window spanned a fixed number of 241 frames, with a stride of 25, ensuring overlap between consecutive segments. For instance, the first window covered frames 0 to 240, the second covered frames 25 to 265, and so on. As with the threes segments analysis, each floating window was treated as an experiment on its own.

This method provided a finer temporal resolution, enabling the detection of subtle transitions in behavior that might otherwise be missed with an overall or large-scope view. Within each floating window, I repeated the entire analysis pipeline - calculating the MSD, making a log-log transformation, and applying a linear regression to estimate the scaling exponent γ. In addition, I computed the average number of cells per frame. This gave an indication about the relationship between the density of the cells and their ability to move freely. The goal was to capture time variations in movement behavior and diffusion regime, that are not observable when calculating the overall scaling exponent over the entire experiment.

The results reveal a systematic decrease in the scaling exponent γ over time, combined with increasing cell density. Initially, γ values were close to 2, indicating ballistic motion. As the experiment progressed and cell density increased, γ steadily declined, reaching values near 1.2 in the final windows, suggest approaching normal diffusion. By plotting γ as a function of the average number of cells per frame, a clear relationship emerged, demonstrating how increased density limits cells mobility and affects the diffusion regime. Figure 4.14 illustrates this dynamic relationship between the average number of cells and the scaling exponent γ.

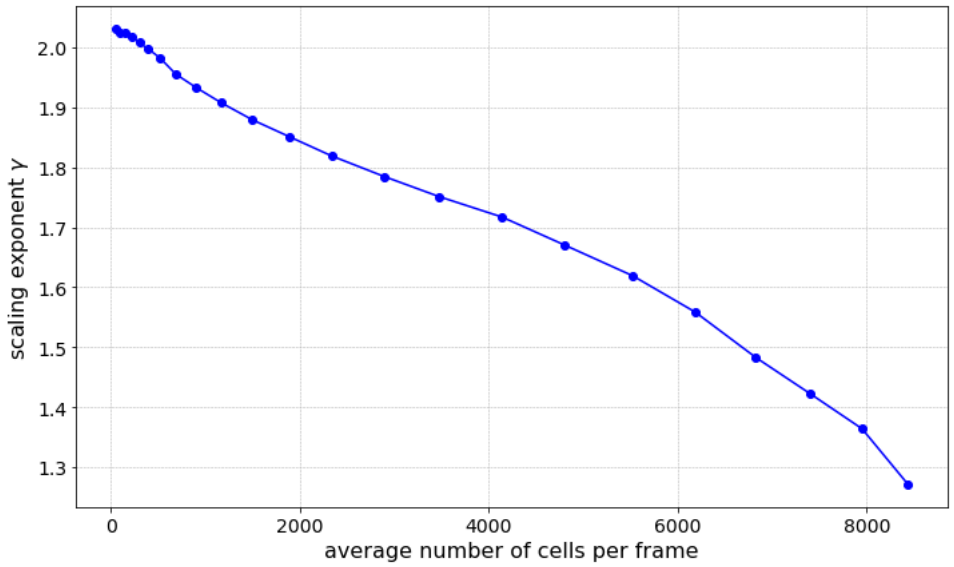


Figure 4.14: **Floating Window Scaling Exponent (set number 4).**The plot shows the scaling exponent γ as a function of cell density, measured by the average number of cells per frame. The x-axis represents the average number of cells per frame, while the y-axis shows the corresponding scaling exponent γ, obtained from log-log regression of the MSD within each time window. A clear monotonic decline in γ is observed as the average number of cells increases, indicating a transition from super-diffusive (ballistic-like) motion at low density, to normal diffusive behavior at higher densities. This trend highlights the effect of crowding on cell movement.

Part IV

Discussion

The goal of this research was to investigate the statistical properties of HBE cell trajectories and evaluate whether their motion could be characterized by a Lévy walk model. Using time-lapse imaging data and analytical techniques, including Trackpy for trajectory extraction and various statistical methods for analysis, we explored multiple aspects of cell movement across both time and space.

Initial analysis of the MSD across all cell trajectories revealed super-diffusive behavior, with scaling exponents γ in the range of $1.78 - 1.80$. This observation initially suggested the possibility of Lévy walk dynamics, which are known to produce super-diffusion with similar exponents. By dividing each data set into two equal halves and calculating the scaling exponent separately, we started investigating the temporal evolution of the diffusion behavior. Interestingly, in the first half of each set, we observed very strong super-diffusion with $\gamma \approx 1.96$ approaching ballistic motion. In contrast, the second half of the data consistently exhibited a significant drop in the exponent, with γ values in the range $1.14 - 1.27$, closer to normal diffusion. This difference suggests a time-dependent change in motility rather than a stationary stochastic process. Further analyses showed inconsistencies with the theoretical properties of Lévy walks.

The displacement distributions, when tested against both symmetric Lévy alpha-stable distributions and Gaussian distributions, failed to fit well to either model. Notably, even at long time intervals where heavy-tailed behavior would typically be most prominent, the empirical data did not exhibit the power-law characteristics required for Lévy walks. Instead, a mixture of two normal distributions provided a significantly better fit, as evidenced by lower AIC values across all experimental segments and time points. This suggests that the cell motion likely consists of multiple movement modes or dynamic sub-populations, which cannot be adequately captured by a single Lévy or Gaussian model.

The velocity ACF analysis provided additional evidence against the Lévy walk hypothesis. Lévy walks are expected to yield a power-law decay in the velocity ACF. However, our results showed a significantly better fit to an exponential decay model. The ADF stationarity tests supported the

assumption of weak stationarity for the first and second moments, but the decay behavior of the ACF strongly contradicted the theoretical Lévy walk dynamics.

To further investigate the temporal evolution of diffusion behavior, we performed a three-segment analysis as well as a moving window approach. Both methods consistently revealed a gradual shift in motility over time. In the early stages of the experiment, the motion was nearly ballistic, with scaling exponents $\gamma \approx 2$. As time progressed, the diffusion behavior transitioned toward values of $\gamma \approx 1.2$, indicative of almost normal diffusion. This temporal trend showed a clear correlation with cell density. As the number of cells increased, the measured γ values decreased. This suggests that as cells became more densely packed, their movement was increasingly constrained.

The mixture of two normal distributions, which consistently provided a superior fit compared to both single Gaussian and Lévy stable models, capture the complexity of cell displacement distributions. While this statistical improvement does not directly imply the existence of two distinct sub-populations, it does suggest heterogeneity in the displacement behavior that a single-mode model fails to capture. Interestingly, a video created from the time-lapse frame images visually reveals what appears to be two main opposing streams of cells, moving in different directions. Although this observation is qualitative, it offers a possible explanation for the improved fit of the mixture model. The displacement distributions may be shaped by directional flows or transient collective motion, rather than distinct motility modes. Further quantitative analysis would be required to confirm whether these visually observed streams correspond to separable motion patterns within the data.

Despite these insights, several important questions remain unanswered and point toward potential avenues for future work:

Censoring Effects: How do limitations in trajectory duration—caused by cells moving out of frame or segmentation issues—affect statistical measurements such as MSD scaling or displacement distributions? A more systematic approach is needed to quantify potential bias introduced by data censoring.

Interpretation of the Mixture Model: Why does a mixture of two Gaussian distributions outperform both single Gaussian and Lévy models? Is this simply a statistical artifact, or does it reflect biological reality such as underlying sub-populations, transient directional motion, or collective effects?

Time vs. Density Dependence: The observed decrease in motility (γ values) over time correlates with increasing cell density. However, it remains unclear whether this trend is driven primarily by mechanical constraints due to crowding or intrinsic cellular changes over time, such as senescence or metabolic slowing.

Super-Diffusion Despite Exponential ACF: Super-diffusive behavior is often associated with long-time correlations, yet the velocity auto-correlation function decays exponentially rather than as a power law. This raises the question: what mechanisms are responsible for the observed super-diffusion, and can they be reconciled with short-range memory effects?

Addressing these questions will be critical for constructing more accurate models of cell motility and understanding the underlying biological mechanisms at play.

References

[1] Gil Ariel, Amit Rabani, Sivan Benisty, Jonathan D. Partridge, Rasika M. Harshey, and Avraham Be'er. "Swarming bacteria migrate by Lévy Walk." Nature Communications 6:8396 (2015).

[2] Carles Blanch-Mercader, Victor Yashunsky, Simon Garcia, Guillaume Duclos, Luca Giomi, and Pascal Silberzan. "Turbulent dynamics of epithelial cell cultures." Physical Review Letters 120 (2018).

[3] Crispin William Gardiner. "Handbook of stochastic methods for physics, chemistry and the natural sciences." Springer Series in Synergetics (1985).

[4] Venkataraman Balakrishnan. "Mathematical physics: applications and problems." Springer Nature (2020).

[5] Venkataraman Balakrishnan. "Fluctuation-dissipation theorems from the generalized Langevin equation." Pramana 12, 301-315 (1979).

[6] Gerrit Gerhartz. "Langevin and Fokker-Planck equation." In Seminar Theoretical Statistical Physics, Heidelberg University (2023).

[7] Patrick H. Diamond. "Brownian Motion and Langevin Equations." Lecture Notes for Physics 210B. https://courses.physics.ucsd.edu/2020/Fall/physics210b/lecture.html (2020).

[8] Yat Chun Chester Wong, and Paul Bilokon. "Simulation of Fractional Brownian Motion and Related Stochastic Processes in Practice: A Straightforward Approach." Available at SSRN (2024).

[9] Vasily Zaburdaev, Sergey Denisov, and Joseph Klafter. "Lévy walks." Reviews of Modern Physics 87, 483-530 (2015).

[10] Ralf Metzler, and Joseph Klafter. "The random walk's guide to anomalous diffusion: a fractional dynamics approach." Physics Reports 339, 1-77 (2000).

[11] Aaron Clauset, Cosma R. Shalizi, and M. E. Newman. "Power-law Distributions in Empirical Data." ArXiv, https://doi.org/10.1137/070710111 (2007).

[12] Hamparsum Bozdogan. "Akaike's information criterion and recent developments in information complexity." Mathematical Psychology 44, 62-91 (2000).

Hebrew Abstract

המחקר הזה בוחן את התכונות הסטטיסטיות של מסלולי תאים בודדים בשכבה דקה של תאי אפיתל סימפונים אנושיים, באמצעות נתוני טיימלפס ניסיוניים. המטרה המרכזית היא לבדוק האם תנועת התאים מתנהלת לפי מודל של הליכת לוי - מודל של סופר־דיפוזיה המאופיין בהתפלגויות צעדים עם זנבות כבדים ודעיכה של פונקציית הקורלציה העצמית של המהירות לפי חוק חזקתי.

לשם כך, מסלולי התאים הוצאו מתמונות מיקרוסקופ באמצעות תוכנת טראקפי ונותחו על ידי כלים סטטיסטיים שונים, ביניהם ממוצע ריבועי ההעתק, התפלגויות ההעתק, מבחני נייחות ופונקציות הקורלציה העצמית של מהירות. ניתוח ראשוני של תוחלת ריבוע ההעתק חשף התנהגות סופר־דיפוסיבית עם מעריכים הקרובים לתנועה בליסטית בשלבים המוקדמים של הניסוי, אשר עוברת לדיפוזיה כמעט רגילה עם העלייה בצפיפות התאים. עם זאת, ניתוח מעמיק יותר של התפלגויות ההעתק ופונקציה הקורלציה העצמית הראה סטיות ברורות מהחיזוי של מודל הליכת לוי. במיוחד, הנתונים לא התאימו להתפלגות לוי אלפא־יציבה או להתפלגות גאוסית, אלא תוארו טוב יותר על ידי מיקס של שתי התפלגויות נורמליות. פונקציית הקורלציה העצמית של המהירות הציגה דעיכה מעריכית במקום דעיכה לפי חוק חזקתי, מה שמנוגד למודל הליכת לוי.

ניתוח מקטעים וחלונות זמן נעים הדגישו שינוי במהלך הזמן במשטר הדיפוזיה, כאשר גידול בצפיפות התאים הראה מתאם לצמצום התנועה. ממצאים אלה מצביעים על כך שבזמן שהתאים מציגים תחילה תנועה סופר־דיפוסיבית, ההתנהגות שלהם לאורך זמן מורכבת יותר ומתוארת טוב יותר על ידי תהליך הטרוגני התלוי בצפיפות, ולא על ידי הליכת לוי.

עבודה זו נעשתה בהדרכתו של פרופ׳ גיל אריאל,
מן המחלקה למתמטיקה,
של אוניברסיטת בר־אילן.

אוניברסיטת בר־אילן

תכונות סטטיסטיות של מסלולי תאים בודדים בשכבה דקה של תאי אפיתל סימפונים אנושיים

אלון סלע

עבודה זו מוגשת כחלק מהדרישות לשם קבלת תואר מוסמך במחלקה למתמטיקה

של אוניברסיטת בר־אילן

רמת גן תשפ"ה