Single-Particle Tracking (SPT) is a powerful technique to detect the dynamic behaviour of molecules in biological systems. By estimating the diffusion coefficient (D) of individual particles, cell states and ongoing biochemical processes can be inferred. However, standard SPT methods based on Mean Square Displacement (MSD) fail in high-noise fluorescence imaging, especially when the particle moves during image acquisition. This motion causes blur and spreads photons across pixels, making localization unreliable—but also encoding useful information.

Recent work (citer Park ?) showed that convolutional neural networks (CNNs) can extract this information directly from raw images to predict D without relying on trajectories. These promising results lack the usage of temporal and trajectory-based information, leading to high error rates in predictions. In this project, we explore the use of Vision Transformers (ViTs) for predicting molecular diffusion coefficients from fluorescence time-series images. Our approach demonstrates that ViTs outperform CNNs in multiple signal-to-noise conditions, capturing long-range dependencies and temporal dynamics more effectively. Moreover, other ML methods predict D from incorporated handcrafted features extracted from reconstructed trajectories. We show that incorporating these features alongside the image data leads to further performance gains, especially when trajectories are partially available or corrupted.

To overcome data scarcity and labelling challenges, we adopt a self-supervised training paradigm. Simulated trajectories and corresponding noisy images are generated synthetically, approximating a wide range of experimental conditions. The training process is computationally efficient, taking approximately one hour on a single GPU, and the resulting models can be easily adapted to different microscope settings or acquisition protocols,

#making our framework both practical and scalable for real-world applications.