Abstract:

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 (Park et al.) [1] 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 (Kæstel-Hansen et al.) [2] 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 (Muñoz-Gil et al.) [3] 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.

1700 characters without space

2047 characters with space

[1] Park, H.H., Wang, B., Moon, S. *et al.* Machine-learning-powered extraction of molecular diffusivity from single-molecule images for super-resolution mapping. *Commun Biol* **6**, 336 (2023). <https://doi.org/10.1038/s42003-023-04729-x>

[2] Kæstel-Hansen, J., de Sautu, M., Saminathan, A. *et al.* Deep learning-assisted analysis of single-particle tracking for automated correlation between diffusion and function. *Nat Methods* **22**, 1091–1100 (2025). <https://doi.org/10.1038/s41592-025-02665-8>

[3] Muñoz-Gil, G., Volpe, G., Garcia-March, M.A. *et al.* Objective comparison of methods to decode anomalous diffusion. *Nat Commun* **12**, 6253 (2021). <https://doi.org/10.1038/s41467-021-26320-w>

**Our Proposal**

[graphique pipeline]

Simulator:

Input: Range of D values, optics parameters, noise settings.

The simulator samples a value of D in range, generates a trajectory with wanted D, computes features from averaged positions and generates K frames by combining multiple sub-positions in one image.

Output: The real D\_GT used to generate the trajectory, associated features and K images

Trainer:

Input: D\_GT, associated features and images

Images go through a CNN that generates K embedding vectors, which are concatenated to the feature vector and inputted into the Transformer. The transformer’s output D\_Pred is compared to the D\_GT and the error is backpropagated using a loss function.

Simulation + Training is repeated until loss converges

Output: trained Model

Predictor:

Input: Real data image with multiple particles

Particles are tracked individually, for each particle compute trajectory features and extract image patches. Input into trained model and compare prediction to classical methods.

**Shape and Trajectory Improves Performances:**

A graph of a model

AI-generated content may be incorrect.A graph with red dots and green dots

AI-generated content may be incorrect.

Results from left to right:

Existing methods in shades of red/orange. Prediction from localized position using MSD (MSD Localized), prediction from features computed on localized position (Feat only; used by Kæstel-Hansen), Average prediction of CNN model on images (CNN Only; used by H.H. Park).

New methods in shades of green: Temporal transformer on images encoded by CNN (Transf(CNN)), Multilayer Perceptron on CNN features + Features (CNN + Feat), Temporal transformer on images encoded by CNN and features (Transf(CNN + Feat)).

While image methods can outperform standard and feature-based methods, image transformer have lower error and deviation rates. Combining features and images outperforms using them individually, and a transformer model outperforms classical MLP.

Classical methods and small models (Feat only) are very fast but lack accuracy, which can be obtained by using a CNN and/or Transformer model

**Towards a Real System**

Currently models are performing well on simulated data but need some fine-tuning to work on real data. A real challenge is estimating the noise of real images, to replicate it closely in simulations. Otherwise, the models’ predictions are not accurate due to out of distribution data and predictions cannot be trusted. The lack of ground truth on real data makes verifiability of results harder.

Another subject of interest for biologists is detecting changes in diffusion coefficient. Additional work is required to apply it on real data, but transformer architecture outperforms standard CNN due to the temporal self-attention mechanism when trained accordingly.