Diffusion Coefficient estimation using ViT

Emilien Silly - emilien.silly@epfl.ch
Data Science Master research project at EPFL, 12 ECTS
Laboratory: Center for imaging - Biomedical Imaging Group
Under the supervision of Daniel Sage

06/06/2025

Abstract

Estimating the diffusion coefficient (D) of single molecules from fluorescence microscopy is essential for probing dynamic cellular processes. However, traditional methods based on mean square displacement (MSD) often fail in noisy or motion-blurred conditions. We propose a Vision Transformer (ViT)-based approach to predict D directly from time-series images, capturing spatial and temporal patterns more effectively than CNNs. We further improve performance by integrating handcrafted features from localized trajectories. To address data scarcity, we train on synthetically generated images and trajectories, enabling fast, adaptable, and robust diffusion analysis across varying imaging conditions.

1 Introduction

Estimating the diffusion coefficient (D) of particles evolving freely following a Brownian motion within cells provides critical insight into cellular states and responses to external stimuli. This is commonly achieved using fluorescence microscopy, where particles of interest are tracked over time. Standard approaches involve nearest-neighbor tracking, sub-pixel localization of particle positions, and diffusion estimation via mean square displacement (MSD). Despite their widespread use, these methods suffer from significant limitations. Localization accuracy degrades with image noise, leading to unreliable D estimates. Furthermore, traditional methods ignore valuable information encoded in the raw images: during the typical $\sim$100ms exposure time, particles continue to move, with faster-moving particles generating more pronounced motion blur. This movement spreads photon intensity across pixels, creating blurry or faint "blob-like" appearances, especially at high diffusion rates, while static particles have a more unblurred and peaked gaussian appearance. While such image artifacts are difficult to interpret manually due to noise and microscope's Point Spread Function (PSF), previous work has shown that convolutional neural networks (CNNs) can partially recover this information from individual frames. In this work, we demonstrate that Vision Transformers (ViTs) outperform CNNs in both accuracy and robustness, and effectively integrate additional inputs such as trajectory-derived features to improve diffusion estimation under challenging imaging conditions.

2 Related Work

Park et al. (2023) introduced Pix2D, a convolutional neural network (CNN) framework for estimating diffusion coefficients directly from fluorescence image stacks, without relying on mean square displacement (MSD) analysis [9]. The model processes small patches centered on single particles and predicts diffusion coefficients for each frame, which are then averaged to produce a final estimate. While this approach demonstrated the feasibility of image-only predictions, it suffers from high uncertainty, particularly when limited temporal data is available.

As part of the EPFL Machine Learning for Science (ML4Science) initiative during the Autumn 2024 semester, our team collaborated with the Biomedical Imaging Group (BIG) on a related project using 3D CNNs (2D spatial + 1D temporal). In this initial work, particles were allowed to move freely across larger image patches, and the model attempted to capture temporal dynamics. While promising for simulated data, the approach showed limited applicability to real experimental conditions due to noise and tracking challenges.

The AnDi Challenge [8] proposed a benchmark for identifying anomalous diffusion (i.e., sub- or super-diffusion) using simulated trajectories. Many top-performing methods leveraged machine learning models trained on handcrafted features extracted from particle tracks, significantly outperforming classical MSD approaches. Notably, the challenge provided a simulator to generate labeled synthetic data under various diffusion regimes—a resource we also employ in this project.

Kæstel-Hansen et al. (2023) presented DeepSPT, a supervised framework that predicts particle diffusion behavior using a compact set of statistical features derived from reconstructed trajectories [3]. Their work

demonstrated high predictive accuracy and successful application to real microscopy data, emphasizing the potential of trajectory-based feature engineering.

Building upon these approaches, our project aims to bridge the gap between image-based and trajectory-based methods. We show that Vision Transformers (ViTs), which better capture spatiotemporal dependencies than CNNs, yield improved performance across noise conditions. Moreover, we propose a hybrid model that integrates both raw image sequences and trajectory-derived features, outperforming methods relying solely on one modality.

3 Method

3.1 Mean Squared Displacement (MSD)

Although this work does not focus on MSD fitting, it is important to understand how particle motion is generated and to recognize the limitations of MSD in the presence of localization uncertainty.

In the case of a single particle undergoing Brownian motion, the Mean Squared Displacement (MSD) quantifies the average squared distance from its initial position over time. It is defined as:

$$\mathrm{MSD}(t) \equiv \left\langle \left|\mathbf{x}(t) - \mathbf{x}_0\right|^2 \right\rangle = \left\langle \sum_{i=1}^{n} \left(x_i(t) - x_i(0)\right)^2 \right\rangle$$

where $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_n(t)]$ represents the position of the particle in n-dimensional space at time t, and $\mathbf{x}_0 = \mathbf{x}(0)$ is the initial position.

For Brownian motion in each dimension, the displacement follows a Gaussian distribution with zero mean and variance $2Dt$, where D is the diffusion coefficient. Therefore, the MSD in each coordinate is $\langle (x_i(t) - x_i(0))^2 \rangle = 2Dt$, and the total MSD becomes:

$$\mathrm{MSD}(t) = 2nDt$$

This linear relationship between MSD and time forms the basis for estimating the diffusion coefficient D by fitting a line to the MSD curve.

However, this approach becomes unreliable in the presence of localization noise, which introduces a bias in the MSD curve, especially at small time lags.

As discussed in [6], the accuracy of D estimation can be significantly improved by assigning more weight to early MSD points in the fitting process. However, Michael et al. [7] derive bounds on the accuracy of MSD-based estimates as a function of the localization error variance σ^2, and highlights that variable or static noise levels can strongly affect the quality of the diffusion coefficient estimate.

3.2 Simulations and Training

During training, particle trajectories are generated using the simulator provided by the AnDi Challenge 2 [8]. In this study, we restrict its use to the case of normal diffusion, corresponding to Brownian motion, by setting the anomalous diffusion exponent $\alpha = 1$. For each trajectory, a diffusion coefficient D (in pixels^2/s) is sampled uniformly from a predefined range. This value serves as the ground truth (D_{GT}) for that particular trajectory and is used to compute the training loss.

It is important to note that the simulator does not allow explicit specification of the temporal resolution (Δt) and the diffusion coefficient is given in pixels^2/s. As such, unit conversion from pixels to physical units (e.g., nanometers) must take this into account when interpreting the predicted diffusion coefficient.

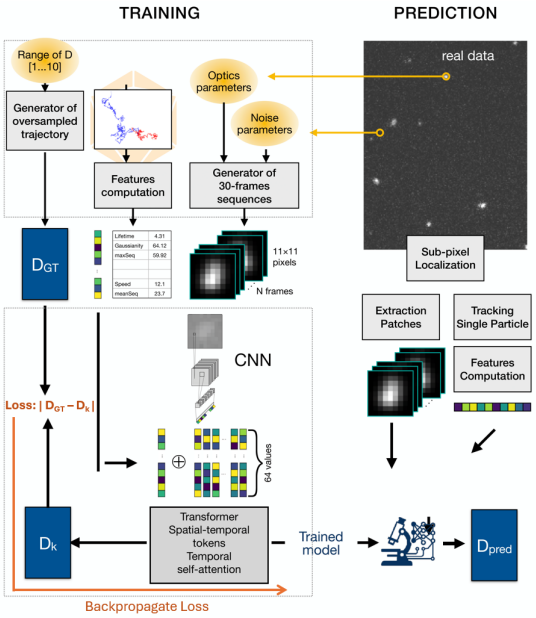

Figure 1: Simulation and Training pipeline

Top left part: Trajectory and image Generation
Bottom left part: Model Training
Right part: Prediction on real data

Once a trajectory is generated, it is transformed into an image sequence that mimics fluorescence microscopy acquisition. To simulate motion blur, we divide each frame into a fixed number of sub-steps (here set to 10). The particle's sub-positions are centered within each frame, convolved individually with the microscope's point spread function (PSF), and then summed. Gaussian noise is added to represent background fluctuations, and Poisson noise is applied to simulate the photon statistics of the acquisition process. Figure 8 shows examples of generated images, with visible motion blur and noise components.

The simulation pipeline is highly configurable: all physical and imaging parameters—such as noise levels, optical resolution, and pixel size—can be tuned to

match specific experimental conditions.

The number of timepoints in each trajectory determines the number of frames in the corresponding image sequence. The localized position of the trajectory for each frame is sampled from the trajectory by averaging the positions of a frame, and added with gaussian localization error. These localized positions are then used to generate 25 trajectory features following a subset of the features presented by [3]. During training, each sequence is first normalized in intensity alongside the features associated to the sequence, then passed through the model, which outputs a predicted diffusion coefficient D_{k}. The loss is computed as the discrepancy between D_{k} and the ground truth D_{GT}, and the model is updated via backpropagation.

Once the model is trained, it is usable for prediction on real data. To do so, particles are detected, then tracked and patches around the particle's position are extracted. Sub-pixel localization for each particle are estimated and features are computed from each trajectory. The patches and features are fed into the model which predicts one Diffusion value for each particle as D_{Pred}

3.3 Models

To assess the effectiveness of deep learning approaches in predicting diffusion coefficients, we implemented and compared several models, including both previously published architectures and newly proposed ones.

We re-implemented Pix2D, a convolutional neural network (CNN) architecture proposed by Park et al. [9], using a ResNet-based backbone instead of a CNN for faster training and convergence as described in [2]. Additionally, we designed a baseline two-layer multilayer perceptron (MLP) that operates solely on handcrafted trajectory features, mimicking traditional machine learning approaches.

To explore multimodal integration, we introduced a lightweight CNN+MLP hybrid model, which combines both image data and trajectory features as input. This architecture allows us to compare the benefits of multimodal input with those of the more advanced transformer-based approach.

The main contribution of this work is a vision transformer (ViT)-based architecture tailored for diffusion analysis. In our design, each frame in the image sequence is treated as a token. Initially, each image is processed through a multi-layer CNN (similar to a shallow ResNet) to extract feature embeddings. These embeddings form a sequence that is passed to a transformer encoder composed of multiple layers of self-attention and feedforward sub-modules, following the standard architecture introduced by Vaswani et al. [11].

To enable regression of the diffusion coefficient, we prepend a dedicated regression token to the sequence, following the strategy proposed in [1]. This token aggregates information across all frames through the self-attention mechanism. Additionally, trajectory features—when available—are concatenated to this token at the input stage, allowing them to be directly integrated into the attention process.

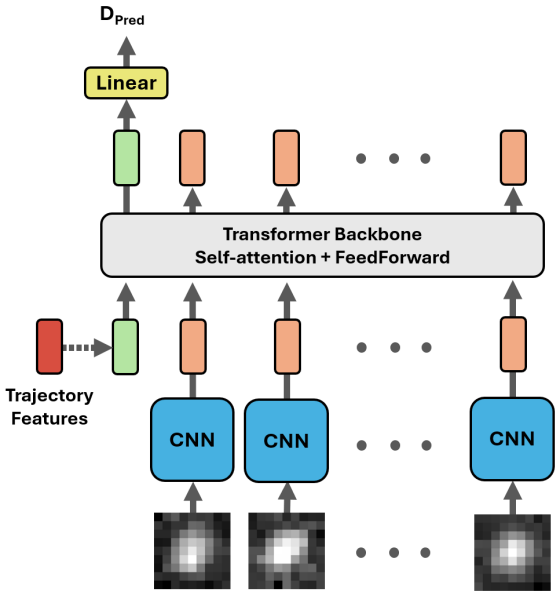

Figure 2: Transformer Model Architecture

Green box: Regression token;
Orange boxes: Embedded image tokens.

The architecture is fully configurable in terms of depth, embedding dimension, and number of attention heads. In this study, we use a compact configuration: 6 transformer layers, each with 4 attention heads and a hidden size of 128. This results in a model with approximately 500k parameters—comparable in scale to the ResNet-based CNN (320k parameters), making it both efficient and expressive.

4 Experiments

4.1 Classical MSD-Based Estimation

We first evaluated traditional methods based on Mean Squared Displacement (MSD), which assume that particle trajectories have been accurately localized. To simulate such data, we generated trajectories with $N \times 10$ steps, then averaged positions in blocks of 10 to mimic frame acquisition. Gaussian noise was added to emulate localization error typically encountered in experimental tracking.

We compared four approaches to estimate the diffusion coefficient D from MSD curves:

1. Estimating D from the pairwise displacement between consecutive positions;

2. Fitting only the optimal number of MSD values as stated in [7]; Here 10% of points

3. Applying the weighted fitting method described in [6];

4. Fitting the full MSD curve.

Our results show that fitting the full MSD curve yields the least reliable and least precise estimates of D. The weighted fitting method offers only marginal improvements but remains prone to error. In contrast, limiting the fit to the early portion of the MSD curve (first 10%) significantly improves precision. However, when localization noise is present, relying on a single displacement step (method 1) becomes more error-prone due to noise sensitivity, despite its lower inherent uncertainty. This will be important for later analysis, since we will compare our models against this baseline to capture the benefit of our method.

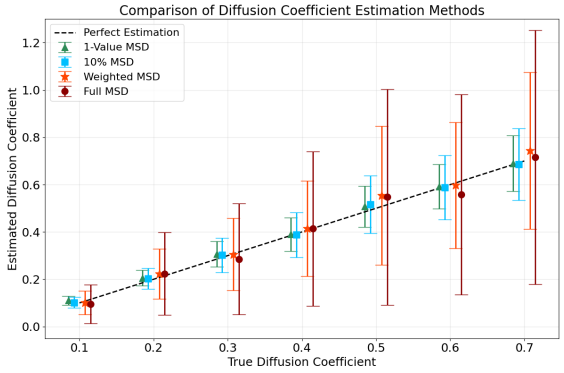

Figure 3: Comparison of Classical MSD Methods

4.2 Transformer Image Encoding and Model Size

Classical Vision Transformers (ViTs) typically divide input images into smaller patches, each treated as an individual token. Positional encodings are then added to these tokens, allowing the transformer to capture spatial relationships and outperform traditional CNNs, especially for long-range dependencies.

In our case, the input images are already small (typically 7×7 to 15×15 pixels), making patching unnecessary. Instead, we treat each image in the temporal sequence as a separate token. However, transforming each image into a meaningful vector representation remains critical for downstream performance. To this end, we evaluated several image embedding strategies:

- **Linear Projection:** A standard ViT approach where the flattened image is linearly projected into the transformer input dimension.
- **1-Layer CNN:** A shallow convolutional layer where the number of filters matches the input dimension of the transformer.
- **Deep ResNet CNN:** A multi-layer residual CNN where the final output dimension matches the transformer input. Residual connections help stabilize training and accelerate convergence.

Model Size Considerations

Since the model is intended to be retrained for different microscope configurations, we aim for an architecture that balances performance with training efficiency. We tested three model capacities:

Model	Layers	Heads	E/H Dim	# Params
Small	3	2	32 / 64	326k
Normal	6	4	64 / 128	514k
Large	12	8	128 / 256	1.93M

Table 1: Transformer model configurations tested for trade-offs between model size and training efficiency.

Each model size was combined with each embedding type and trained on the same dataset to ensure fair comparisons. As shown in Figure 4, the Normal-sized model with a deep ResNet embedding consistently outperformed the others. Both the Small and Large models showed inferior results: the former likely due to limited capacity, and the latter due to slower convergence and significantly longer training and inference times. In practice, the Large model required up to *3*-times longer to complete predictions, making it impractical for high-throughput applications.

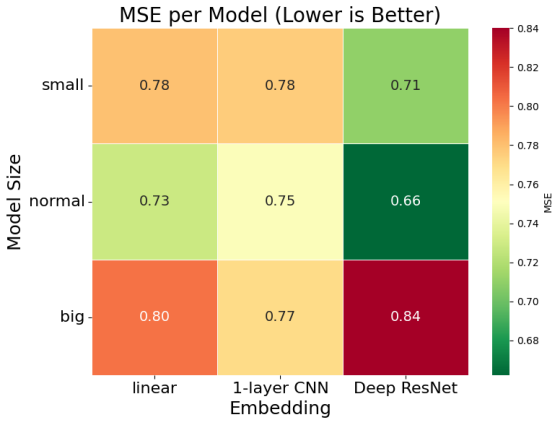

Figure 4: Performance comparison of different model sizes and embedding strategies.

Impact of Embedding Depth

Another key observation is that deeper CNN embeddings (e.g., ResNet-based) significantly outperform both linear and shallow CNN embeddings. This improvement is likely due to the ability of deeper models to learn more robust spatial features, as opposed to linear projections, which capture limited structural information.

To evaluate robustness under data scarcity, we assessed model performance as a function of the number of input images. Figure 5 shows that the deep embedding converges faster and achieves lower mean squared error (MSE) with fewer examples. For instance, it achieves an MSE below 1 with only 5 images, whereas other embeddings require at least twice as many to reach similar performance.

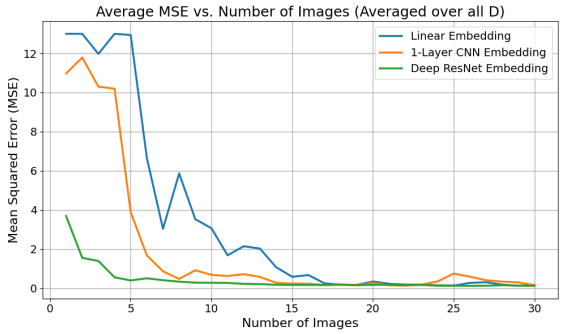

Figure 5: MSE as a function of the number of images for different embedding strategies.

Based on these experiments, we selected the Normal-sized transformer with deep CNN embedding as our default architecture. It offers the best trade-off between accuracy, robustness, and computational cost, and is well-suited for retraining across different microscopy settings.

4.3 Image Deconvolution

A major limitation of image-only models is the high noise inherent in microscopy data, stemming from both Gaussian and Poisson noise. This noise severely reduces visibility of motion blur, which is the key feature we aim to capture, while introducing artifacts that can be mistaken for real signals. Compounding this, the microscope's point spread function (PSF) causes light emitted by the particle to spread spatially. Since the PSF is well-characterized and integrated into our image generation process, we hypothesized that applying deconvolution to reverse its effects might improve model performance.

To test this, we trained models on several image variants:

- Noiseless synthetic images (to establish an upper-bound on performance),
- Realistic noisy images (baseline),
- Noisy images smoothed with a Gaussian filter,
- Images deconvolved using the Richardson–Lucy (RL) algorithm with 2, 5, and 10 iterations.

The results show no performance gain from deconvolution. None of the RL or Gaussian-filtered variants outperformed the model trained directly on noisy images. All remained far from the optimal performance achieved with noiseless data. Our interpretation is straightforward: the noise level is simply too high.

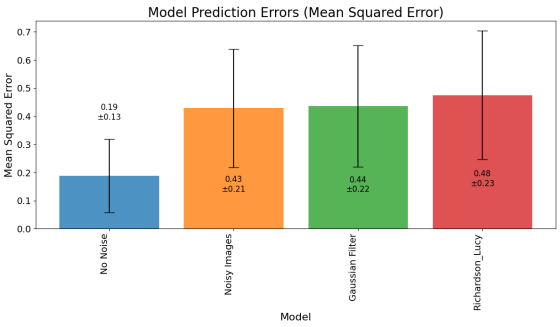

Figure 6: MSE vs deconvolution method used

While RL removes some background noise, it also amplifies or invents features where noise resembles the particle, degrading motion-blur fidelity.

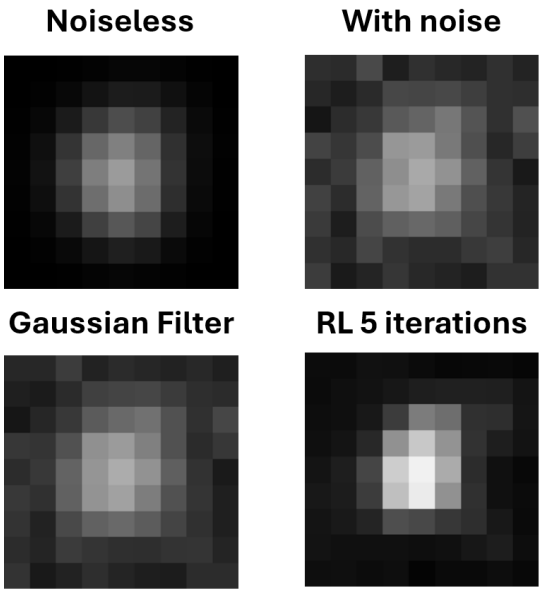

Figure 7: Deconvolution example

Figure 7 illustrates this failure case. RL deconvolution suppresses edge noise but also introduces misleading structures due to overfitting to noisy features, which is detrimental when motion blur is the signal of interest.

As an alternative, we considered—but did not have time to implement—a learned denoising approach. Since we can generate both clean and noisy synthetic images, a model could be trained to map noisy inputs to clean targets followig the method proposed by Spotiflow [5]. This might recover motion blur more effectively and even allow sub-pixel localization. Still, the fundamental limitation remains: at low signal-to-noise ratios, no method—classical or learned—can reliably recover signal when it is indistinguishable from noise.

4.4 Impact of PSF Size and Noise

As previously discussed, image noise heavily degrades model performance. Another critical factor is the point

spread function (PSF) size: a wider PSF spreads the particle signal over more pixels, reducing the visibility of motion blur—the key feature our model relies on.

To systematically evaluate these limitations, we trained and tested models across a range of PSF sizes (from 1.25 to 2.5 pixels) and signal-to-noise ratios (SNRs) from noise-free (ideal) conditions down to an SNR of 5—a level typically considered near the lower bound of usability. Figure 8 shows examples of such images, with 3 examples of each kind of variation.

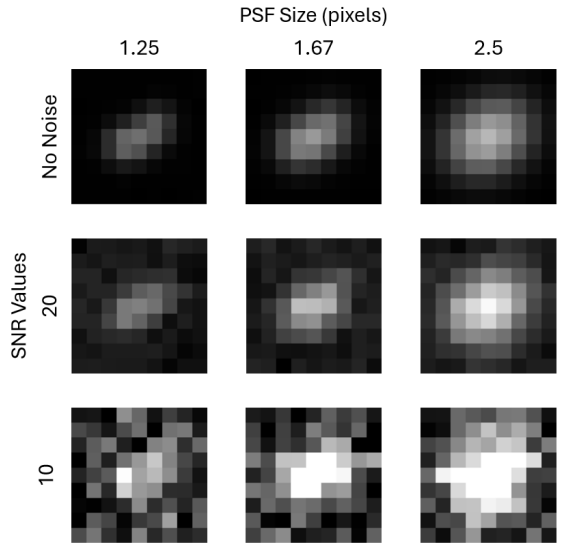

Figure 8: PSF Size and SNR Example images

As shown in Figure 9, the trend is unambiguous: performance drops as PSF size increases and SNR decreases. A doubling of the PSF size can cause up to a 30% increase in error, depending on the noise level. However, improvements from reducing PSF size plateau quickly—going from a mid-sized PSF to the smallest tested one yields only marginal gains.

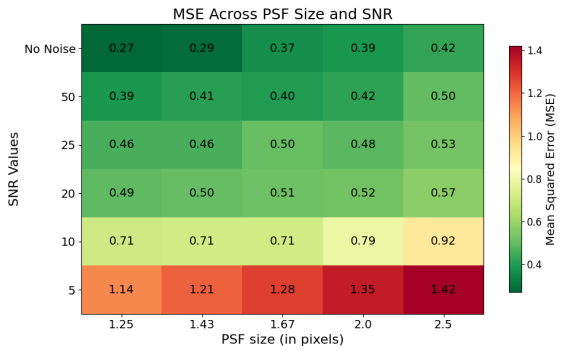

Figure 9: MSE Vs PSF Size and SNR ratio

Noise, on the other hand, has a dramatic effect. The model trained on clean, noiseless images achieves 4× lower error than the one trained at an SNR of 5, and 2× better than at SNR 10. In practical terms, most real microscopy data fall between SNRs of 5–10 and use a PSF around 1.5 pixels. This means that even under best-case acquisition scenarios today, image-only models are operating far from their theoretical performance ceiling.

This experiment confirms a hard limit: unless microscope technology drastically improves—especially in noise reduction—there's a ceiling to how much we can squeeze from visual-only models. However, we need to keep in mind that less noise wouldn't just help our model, it would also greatly improve localization estimates and thus positively impact accuracy of MSD computation.

4.5 Simulation in Mitochondria

To test our models on real data, we collaborated with Jose Requejo-Isidro (CSIC), using microscopy images from his lab involving particles diffusing within the cristae of mitochondria. The goal was not only to estimate the diffusion coefficient but also to infer the diffusion direction—crucial for distinguishing between motion within cristae membranes versus along the cristae rims.

This shifted the problem from standard free diffusion to constrained 1D diffusion shaped by mitochondrial geometry, requiring a custom simulator. It also forced a change in the training objective: beyond predicting D, the loss function had to incorporate angular information.

Despite several iterations of simulator development, the extreme noise levels in the real images made training impractical. The effort required to handle the data exceeded what was justifiable for this project's scope. As a result, we dropped this sub-topic. However, the simulator code is publicly available in the project's GitHub repository should future work revisit it.

5 Results

We introduce two transformer-based models for predicting diffusion coefficients from microscopy data:

- Transf(CNN): a vision-only model using a CNN encoder followed by a transformer;

- Transf(CNN + Feat): the same model extended with additional localized features as input, yielding significantly better performance.

These models were trained on simulated images matching real experimental conditions (SNR 10). At evaluation, we also included a classical MSD-based method operating on localized particle trajectories, as well as several baselines:

A feature-only ML model using 25 hand-crafted trajectory descriptors, following [3];

The CNN-only model from Park et al. [9];

A hybrid CNN + Feature model without attention.

As shown in Figure 10, traditional MSD estimation is clearly outperformed by ML methods, even feature-only approaches. Our image-only transformer surpasses

the baseline CNN, demonstrating the benefit of temporal self-attention. More importantly, fusing images and features leads to consistent improvements across all architectures. Our full model, Transf(CNN + Feat), outperforms all tested baselines by a significant margin.

Figure 10: Mean Squared Error (MSE) of all models

We also compared model inference times in figure 11 on a batch of 10,000 samples. The classical MSD method is fastest, as expected, since it operates on preprocessed (localized) trajectories. However, it also performs worst. Our image-only transformer yields a 10% accuracy gain over a comparable CNN, at a 20× increase in runtime. Interestingly, adding features to either model greatly boosts accuracy without significantly increasing cost—image processing remains the main computational bottleneck. This comforts us in the idea that a Hybrid model using images and features is the right path towards significant improvement in prediction accuracy.

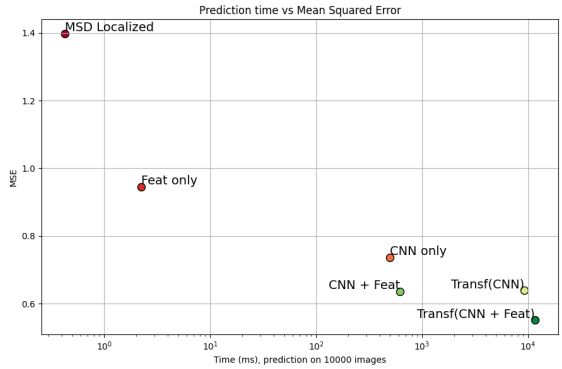

Figure 11: Model MSE vs. Inference Time (log scale)

5.1 Detecting Changes in Diffusion Coefficient

For many biological applications, the primary interest is not the absolute value of the diffusion coefficient D, but rather the detection of changes in D over time. These transitions can signal events such as drug interactions, conformational changes in the molecule, or shifts in environmental conditions. There are two main strategies to address this problem.

The first approach, partially explored in this project, leverages models that remain accurate even with a small number of input images. With such models, it becomes feasible to use a sliding window over the particle's trajectory and detect abrupt variations in predicted diffusion coefficients.

To evaluate this, we tested our model on a synthetic trajectory where a particle first diffuses with $D = 3$, then transitions to $D = 7$ after 30 frames. Using a sliding window of 10 frames, the model was able to track this change in real time, as illustrated in Figure 12. The prediction curve follows the underlying ground truth, showing that the model adapts its estimates based on the particle's instantaneous behavior. This result indicates that a simple thresholding mechanism on the model output could be sufficient to detect such transitions.

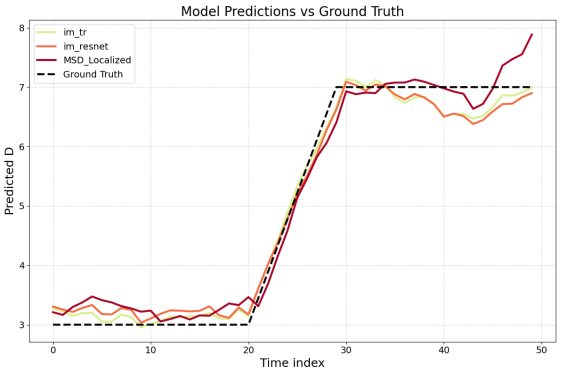

Figure 12: Model predictions on a particle trajectory with a change in diffusion coefficient at frame 30.

A second strategy involves explicitly training a model to detect changes in D. To this end, we constructed trajectories with known transitions (e.g., from low to high diffusion and vice versa) and modified the model architecture to output a prediction for each frame instead of a global regression token. While this change-enabled model could successfully detect transitions, it exhibited increased prediction noise and reduced accuracy on stationary trajectories.

Overall, we conclude that a robust model capable of accurate estimation using short windows is a more practical and effective approach. It simplifies training and avoids trade-offs between global accuracy and change sensitivity.

5.2 Rotation and data augmentation

For image-only models, we explored techniques to boost inference performance without requiring additional training data. Specifically, we applied test-time augmentations that preserve the physical properties of diffusion while altering the image's orientation. Two classes of transformations were used: rotations by multiples of 90° and horizontal/vertical flips. These operations do not affect the underlying diffusion characteristics—since the images are centered, a blurred particle

moving rightward is diffusion-wise equivalent to one rotated in another direction.

By averaging predictions over multiple augmented versions of the same input, we expected a performance gain via ensembling multiple unbiased estimators. As shown in Figure 13, this strategy leads to a modest improvement in prediction accuracy, particularly for the ResNet model, which is less expressive than the Vision Transformer. However, the benefit is limited.

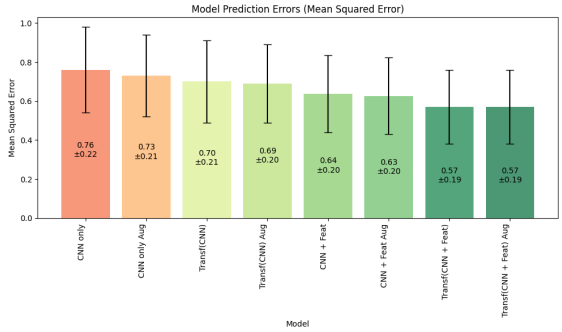

Figure 13: Comparison of models receiving normal input vs models receiving augmented input sequences (suffix Aug)

When handcrafted trajectory features are incorporated, the additional value of image augmentations becomes negligible. This suggests that the hybrid model is already extracting most of the usable information, leaving little room for further improvement via simple transformations.

Although these augmentations could also be applied during training or image generation to expand the dataset, we opted against it. Overusing the same base images through augmented copies could lead to overfitting. Moreover, applying multiple augmentations at inference introduces significant computational overhead—up to 8× longer per prediction—without a proportionate performance gain, making it impractical in many scenarios.

5.3 Prediction on Real Data

A key objective of this project was to apply our model to real microscopy data, ideally in a plug-and-play fashion directly on video sequences. However, this turned out to be significantly more challenging than anticipated. We successfully implemented basic tracking and linking using algorithmic approaches: Difference of Gaussian (DoG) filtering for particle detection and Nearest Neighbor linking for frame-to-frame matching. On the other hand, sub-pixel localization, which is essential for calculating MSD and trajectory-based features, proved more problematic. A simple Gaussian maximum likelihood approach frequently failed on noisy frames, where the signal was barely distinguishable from the background.

Despite these limitations, we applied our trained models to real tracked particles. For the cases where the model confidently produced a prediction, we were unable to validate its correctness. Real data lacks ground-truth labels, making any form of direct accuracy assessment impossible. Traditional MSD-based approaches offer limited utility for cross-validation, as they too suffer from poor reliability under noisy conditions. Moreover, in many instances, the model's predictions were either clearly wrong or entirely uninformative. Some outputs even fell outside the model's training distribution, including unphysical negative diffusion values. This strongly suggests that the real data lies far outside the domain captured by our simulator, exposing its failure to reproduce the complexity of actual microscopy acquisitions.

Several key factors contribute to this mismatch. First, the intensity of real particles is not stable across frames. While our simulator accounts for motion blur and photon spreading, it does not model more complex phenomena such as inter-acquisition variability or fluorophore blinking and bleaching—effects that have substantial impact but are difficult to simulate accurately. Second, noise in real data is more complex than our assumptions. We modeled background noise as Gaussian and signal noise as Poisson, but real background distributions exhibit clear non-Gaussian features, as shown in Figure 14. The histogram reveals a strong right-skew, suggesting additional noise components that were not captured in our simulation model. Finally, in our simulated data, the particle was always perfectly centered and starting it's frame trajectory in the center of the image, which is not always the case for localized particles. To account for this we should slightly move the particle's start point, although it would render the simulator more noisy.

These discrepancies mean our synthetic training data does not sufficiently match real data to support generalization. Consequently, a model trained purely on simulated data fails to perform reliably on real experiments. Other researchers have encountered similar issues and developed more sophisticated simulators for self-supervised learning (e.g., Spotiflow [5] and DeepTrack[10]). Leveraging such simulators as a foundation for future work could substantially improve realism and enable effective model transfer to real-world applications.

6 Discussion

This project demonstrates that transformer-based architectures offer clear advantages over traditional methods, particularly in their ability to model long-range temporal dependencies in particle trajectories. Our results show that combining raw image data with hand-crafted trajectory features can significantly improve prediction accuracy and robustness.

However, the pipeline still suffers from serious limi-

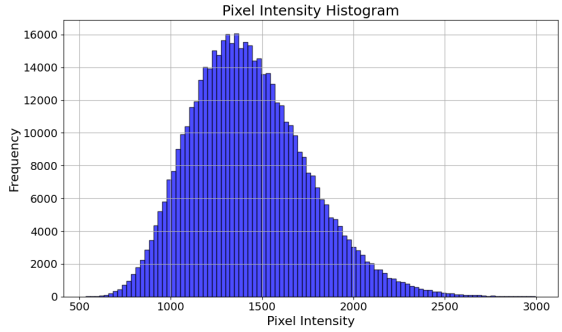

Figure 14: Histogram of background intensity values in real microscopy data.

tations. Most notably, the mismatch between synthetic and real data severely impacts model generalization. In some cases, predictions on real data are completely unreliable, falling outside the training distribution and even returning nonphysical values. In other cases, the lack of ground-truth in real datasets makes it impossible to assess performance with any confidence. This renders current results on real data largely speculative. This is probably the case for other visual methods that suffer from the same problem when applied on real data, and depend greatly on the quality of the acquired data.

To enable effective application of synthetically trained models to real data, two critical improvements are needed. First, the simulator must be enhanced to more accurately capture the physical nuances of fluorescence microscopy that influence final image appearance. Second, advances in experimental acquisition, led by microbiologists—are necessary to reduce noise and imaging artifacts. These complementary efforts would improve both training realism and test-time reliability.

As demonstrated in Section 4.4, our models perform well on clean, synthetic data, setting an upper bound for expected performance. However, further work is needed to quantify how noise and inter-frame variability degrade accuracy on real datasets. Importantly, our findings suggest that high acquisition frame-rates may not be strictly necessary: lower frame-rates with higher signal quality could yield more informative images. By leveraging motion blur as a signal rather than a flaw, our approach can extract positional information even under such conditions, potentially enabling more robust analysis from less noisy inputs.

Another issue we encountered is the inconsistent handling of physical units across simulation components. The AnDi simulator outputs diffusion coefficients in pixel^2/s and trajectories in pixels, while image generation requires real-world scaling parameters (e.g., nanometers per pixel). Additionally, PSF specifications often come in terms of FWHM, requiring further conversions. Without consistent unit handling, results are difficult to interpret and the simulator remains hard to adapt for real-world use. A proper harmonization of all units across the pipeline is essential for any serious application to experimental data. We did not have time to completely implement this unit harmonization across all steps of the project, and it would require thorough investigation with a good physical and/or biological background, which are out of scope of this project.

6.1 Future Work

Our current approach relies on hand-engineered trajectory features, which are computed once per sequence and provided to the model. A more elegant and flexible alternative would be to input the full localized positions of the particle across time directly into the transformer, alongside the image data. This would allow the model to learn relevant spatiotemporal features on its own. Similar ideas have been explored using RNNs and LSTMs [4], especially in the AnDi Challenge 2024. Adopting a similar architecture might further boost performance and eliminate manual feature design altogether.

Given the central role of noise in limiting model performance—especially for sub-pixel motion blur estimation, a dedicated denoising approach could be beneficial. Existing denoising algorithms struggle under high-noise conditions, often introducing artifacts (as shown in Section 4.3). Since we have access to noise-free ground truth images and sub-frame particle positions during training, a supervised denoising model could be trained to either clean the images or directly estimate sub-positions, in the spirit of Spotiflow [5]. Integrating such a module could significantly improve the accuracy of diffusion coefficient predictions, either by preprocessing the input images or by bypassing the image analysis entirely via learned sub-position extraction.

7 Conclusion

Our results demonstrate that transformer-based models outperform previous approaches in both image-only and hybrid image-plus-trajectory diffusion coefficient prediction tasks. Their ability to model long-range temporal dependencies makes them particularly well-suited for single-particle tracking (SPT), and they may hold untapped potential for trajectory-only prediction, potentially exceeding LSTM-based models.

However, we were unable to achieve satisfactory results on real data, primarily due to the challenge of accurately replicating experimental conditions in our simulations. Improving the realism of the simulator is a necessary step for future work, along with access to labeled real-world datasets that could allow for proper validation and fine-tuning of simulation-trained models.

Our framework is highly modular and can be extended to multitask scenarios. For instance, by adapting the output layer and loss function, the model could be used for the AnDi Challenge 2, which requires simultaneous prediction of the diffusion coefficient, diffusion

regime (e.g., free, confined, directed), and anomalous exponent α.

8 Personal conclusion

This project was a highly engaging and valuable experience. I gained a deeper understanding of Vision Transformers and their application to complex computer vision tasks, as well as the challenges involved in modeling noisy biological data. Applying theoretical knowledge from coursework to a real-world problem—especially one as complex as fluorescence-based particle tracking—was both difficult and rewarding. Simulating realistic conditions and understanding the underlying biophysics pushed me to explore areas outside my comfort zone.

I would like to sincerely thank Daniel Sage for his consistent support and guidance throughout the project. His feedback during our weekly meetings helped steer the work in the right direction, and his domain expertise proved invaluable when I encountered technical or conceptual roadblocks.

I also wish to thank José Requejo-Isidro for his early contributions to the project and for providing access to real experimental data. While we were unable to complete the mitochondria simulation within the available timeframe, it remains a promising direction for future work.

Throughout the project, I had the opportunity to engage with several experts in the field. I am grateful to Giovanni Volpe (creator of DeepTrack), Gorka Muñoz-Gil (organizer of the AnDi Challenge), and Jacob Kæstel-Hansen (creator of DeepSPT) for their valuable insights and for taking the time to thoroughly address my conceptual questions. Their input was instrumental in shaping my understanding of key challenges in the domain. The field of fluorescence microscopy and molecular diffusion is vast, and I hope this project contributes, even modestly, to ongoing research. I am happy to assist anyone interested in using or building upon this work. The complete codebase is available at: `https://github.com/mimsilly/SPT_VisionModel`.

References

[1] Emanuel Di Nardo and Angelo Ciaramella. Tracking vision transformer with class and regression tokens. *Information Sciences*, 619:276–287, January 2023. ISSN 0020-0255. doi:10.1016/j.ins.2022.11.055. URL `http://dx.doi.org/10.1016/j.ins.2022.11.055`.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL `https://arxiv.org/abs/1512.03385`.

[3] Jacob Kæstel-Hansen, Marilina de Sautu, Anand Saminathan, Gustavo Scanavachi, Ricardo F. Bango Da Cunha Correia, Annette Juma Nielsen, Sara Vogt Bleshøy, Wouter Boomsma, Tom Kirchhausen, and Nikos S. Hatzakis. Deep learning assisted single particle tracking for automated correlation between diffusion and function. November 2023. doi:10.1101/2023.11.16.567393. URL `http://dx.doi.org/10.1101/2023.11.16.567393`.

[4] Alvaro Lanza, Xiang Qu, and Stefano Bo. Recurrent neural network analysis of single trajectories switching between anomalous diffusion states, 2025. URL `https://arxiv.org/abs/2503.09422`.

[5] Albert Dominguez Mantes, Antonio Herrera, Irina Khven, Anjalie Schlaeppi, Eftychia Kyriacou, Georgios Tsissios, Evangelia Skoufa, Luca Santangeli, Elena Buglakova, Emine Berna Durmus, Suliana Manley, Anna Kreshuk, Detlev Arendt, Can Aztekin, Joachim Lingner, Gioele La Manno, and Martin Weigert. Spotiflow: accurate and efficient spot detection for fluorescence microscopy with deep stereographic flow regression. February 2024. doi:10.1101/2024.02.01.578426. URL `http://dx.doi.org/10.1101/2024.02.01.578426`.

[6] Xavier Michalet. Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. *Phys. Rev. E*, 82:041914, Oct 2010. doi:10.1103/PhysRevE.82.041914. URL `https://link.aps.org/doi/10.1103/PhysRevE.82.041914`.

[7] Xavier Michalet and Andrew J. Berglund. Optimal diffusion coefficient estimation in single-particle tracking. *Physical Review E*, 85(6), June 2012. ISSN 1550-2376. doi:10.1103/physreve.85.061916. URL `http://dx.doi.org/10.1103/PhysRevE.85.061916`.

[8] Gorka Muñoz-Gil, Giovanni Volpe, Miguel Angel Garcia-March, Erez Aghion, Aykut Argun, Chang Beom Hong, Tom Bland, Stefano Bo, J. Alberto Conejero, Nicolás Firbas, Òscar Garibo i Orts, Alessia Gentili, Zihan Huang, Jae-Hyung Jeon, Hélène Kabbech, Yeongjin Kim, Patrycja Kowalek, Diego Krapf, Hanna Loch-Olszewska, Michael A. Lomholt, Jean-Baptiste Masson, Philipp G. Meyer, Seongyu Park, Borja Requena, Ihor Smal, Taegeun Song, Janusz Szwabiński, Samudrajit Thapa, Hippolyte Verdier,

Giorgio Volpe, Artur Widera, Maciej Lewenstein, Ralf Metzler, and Carlo Manzo. Objective comparison of methods to decode anomalous diffusion. *Nature Communications*, 12(1), October 2021. ISSN 2041-1723. doi:10.1038/s41467-021-26320-w. URL http://dx.doi.org/10.1038/s41467-021-26320-w.

[9] Ha H. Park, Bowen Wang, Suhong Moon, Tyler Jepson, and Ke Xu. Machine-learning-powered extraction of molecular diffusivity from single-molecule images for super-resolution mapping. *Communications Biology*, 6(1), March 2023. ISSN 2399-3642. doi:10.1038/s42003-023-04729-x. URL http://dx.doi.org/10.1038/s42003-023-04729-x.

[10] Jesús Pineda, Benjamin Midtvedt, Harshith Bachimanchi, Sergio Noé, Daniel Midtvedt, Giovanni Volpe, and Carlo Manzo. Geometric deep learning reveals the spatiotemporal features of microscopic motion. *Nature Machine Intelligence*, 5(1):71–82, January 2023. ISSN 2522-5839. doi:10.1038/s42256-022-00595-0. URL http://dx.doi.org/10.1038/s42256-022-00595-0.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.