P57. Self-Supervised Learning of Molecular Diffusion using Motion-Informed Vision Transformer - MiViT

Emilien Silly[1], José Requejo-Isidro[2], Daniel Sage[1]

1 Center for Imaging and Biomedical Imaging Group, EPFL, Lausanne, Switzerland
2 Centro Nacional de Biotecnología (CNB), CSIC, Madrid, Spain

daniel.sage@epfl.ch

Estimating diffusion of molecule from image-based single particle tracking (SPT) is essential for probing subcellular states. The diffusion coefficient (D) is typically derived from the mean square displacement (MSD) of sub-pixel localizations; however, motion during exposure produces blurry, blob-like shapes that degrade localization precision and diffusion accuracy. Indeed, previous work [Park, 2023] has shown that convolutional neural networks (CNN) can infer D directly from small image patches centered on the localization; however, the lack of temporal context limits the performance. We propose a Motion-Informed Vision Transformer (MiViT), to directly regress the diffusion coefficient (D) from time-series image patches, capturing spatial and temporal features. Trajectory features [Kæstel-Hansen, 2024] are computed to form a temporal token, which is concatenated with CNN-encoded shape tokens. The resulting spatiotemporal tokens are then processed through self-attention layers within a transformer architecture. To train without labeled data, we use self-supervised learning on simulated sequences of Brownian diffusing particles generated under imaging conditions, aligned with the ANDI challenge [Muñoz-Gil, 2021]. MiViT reduces the error on D estimation, (mean squared error: 1.41 for MSD, 0.76 for CNN, and 0.57 for our method) on 10,000 synthetic samples. The transformer architecture captures long-range dependencies and temporal structure more effectively, especially under noise. Our approach could generalize across various experimental conditions, demonstrating the benefit of spatiotemporal self-attention in MiViT models. Although currently validated only on synthetic data, further work is needed to evaluate robustness under real acquisition variability. Our pilot study work suggests that high frame rates are not strictly necessary; improved image quality at lower frame rates may yield more informative diffusion estimates.