U-Net 3+ for anomalous diffusion analysis enhanced with mixture estimates (U-AnD-ME) in particle-tracking data

Journal of Physics: Photonics

OPEN ACCESS

RECEIVED
25 February 2025

REVISED
2 August 2025

ACCEPTED FOR PUBLICATION
8 August 2025

PUBLISHED
21 August 2025

PAPER

U-Net 3+ for anomalous diffusion analysis enhanced with mixture estimates (U-AnD-ME) in particle-tracking data

Solomon Asghar[1,2], **Ran Ni**[2] and **Giorgio Volpe**[1,*]

[1] Department of Chemistry, University College London, 20 Gordon Street, WC1H 0AJ London, United Kingdom
[2] School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University, Singapore 639798, Singapore
* Author to whom any correspondence should be addressed.

E-mail: g.volpe@ucl.ac.uk

Keywords: particle tracking, machine learning, microscopy data analysis, anomalous diffusion

Abstract

Biophysical processes within living systems rely on encounters and interactions between molecules in complex environments such as cells. They are often described by anomalous diffusion transport. Recent advances in single-molecule microscopy and particle-tracking techniques have yielded an abundance of data in the form of videos and trajectories that contain critical information about these biologically significant processes. However, standard approaches for characterizing anomalous diffusion from these measurements often struggle in cases of practical interest, e.g. due to short, noisy trajectories. Fully exploiting this data therefore requires the development of advanced analysis methods—a core goal at the heart of the recent international Anomalous Diffusion (AnDi) Challenges. Here, we introduce a novel machine-learning framework, U-net 3+ for anomalous diffusion analysis enhanced with mixture estimates (U-AnD-ME), that applies a U-Net 3+ based neural network alongside Gaussian mixture models to enable highly accurate characterisation of single-particle tracking data. In the 2024 AnDi challenge, U-AnD-ME outperformed all other participating methods for the analysis of two-dimensional anomalous diffusion trajectories at both single-trajectory and ensemble levels. Using a large dataset inspired by the Challenge and experimental trajectories, we further characterize the performance of U-AnD-ME in segmenting trajectories and inferring anomalous diffusion properties.

1. Introduction

Due to its ubiquity across a broad range of fields spanning the natural sciences and beyond, diffusion has been widely studied since its first observations by Robert Brown [1]. According to Einstein's relation, the mean squared displacement (MSD) of Brownian motion grows linear with time t: $\mathrm{MSD}(t) \sim Dt$, where D is the diffusion coefficient [2]. Many natural and human processes show deviations from Brownian motion known as *anomalous* diffusion [3, 4], which exhibit non-linear relationships between MSD and time: $\mathrm{MSD}(t) \sim Kt^{\alpha}$, where K is the generalised diffusion coefficient and $\alpha \neq 1$ is the anomalous diffusion exponent [5]. A process is subdiffusive when $\alpha < 1$, and superdiffusive when $\alpha > 1$ [5]. Subdiffusion, which can occur due to crowding or interactions with boundaries, has been repeatedly observed in living cells including within cytoplasms [6], nuclei [7], and cell membranes [8]. Superdiffusion appears in active and directed systems [9–11], such as molecular motors on DNA [12]. Various approaches have been recently put forward for the characterization of these processes [13, 14], also with machine-learning-based methods [15–24].

Within the life sciences, advances in live-cell single-molecule imaging and particle-tracking techniques offer new insights into crucial cellular processes [25, 26]. However, fully leveraging these technical advances requires further development of methods for data analysis. Often, experimental data are extracted in the form of particles' trajectories, and standard analysis methods struggle when these trajectories are, e.g. short, noisy and irregularly sampled [4, 27]. Additionally, there is a need for reliable methods to identify switches between different diffusion behaviours in these trajectories, as these changes are valuable indicators of

biophysical interactions within a system [27]. Examples include variations in the generalized diffusion coefficients K due to conformational changes [28], dimerization (DI) events [29] and ligand binding [30]. Dynamics can also change due to transient immobilization [31] and confinement effects [32].

With live-cell single-molecule experiments in mind, three particularly informative properties for characterizing anomalous diffusion in trajectories are α, K, and the phenomenological behaviour of the diffusing particles (diffusion type, DT), which can be classified as immobilized, confined, freely diffusing, or directed [27]. The international Anomalous Diffusion (AnDi) Challenges aimed to quantitatively assess the quality of existing methods for the difficult, yet important, task of identifying these properties and to spur the creation of new methods [4, 27]. The last AnDi Challenge took place in 2024 and was designed specifically with biological applications in mind, focusing on two-dimensional heterogeneous diffusion in the cellular environment [27]. Specifically, the Challenge aimed to evaluate methods for detecting and quantifying changes in single-particle motion, focusing on trajectory segmentation and the inference of diffusion properties at both single-trajectory and ensemble levels [27]. The Challenge was divided into four tasks (two ensemble-level and two single-trajectory tasks) across two tracks based either on the analysis of trajectories or of videos directly.

Here we introduce our novel machine-learning framework for the highly accurate characterisation of anomalous diffusion properties in single-particle trajectories. We developed our framework, called U-net 3+ for anomalous diffusion analysis enhanced with mixture estimates (U-AnD-ME), to compete in the 2024 AnDi Challenge. U-AnD-ME obtained 1st place for both tasks in the Challenge's Trajectory Track. After briefly introducing the anomalous diffusion models and metrics used for training and performance evaluation, we describe U-AnD-ME's architecture and training, followed by an ablation study (section 2). Next, we benchmark our framework's performance in terms of analysing anomalous diffusion trajectories on both synthetic and experimental data (section 3). Finally, we conclude, discussing possible future improvements and applications for U-AnD-ME (section 4).

2. Methods

2.1. Anomalous diffusion data

To benchmark our method against data of a known ground-truth, we simulated two-dimensional fractional Brownian motion trajectories [33], similar to those of the 2024 AnDi Challenge [27], with the `andi-datasets` Python package [34]. Simulations used generalized units (i.e. pixels and frames). The Challenge considered five different physical models of particles' motion and interaction with the environment (figure 1(a)): single-state diffusion (SS) - particles have a single diffusion state [35]; multi-state diffusion (MS) - particles spontaneously switch between two or more diffusion states with different K and/or α [28, 36, 37]; dimerization (DI) - particles diffuse according to a two-state model, with switching induced by random encounters with other particles [29, 30, 38]; transient confinement (TC) - particles diffuse according to a space-dependant two-state model, being in one state when outside confined regions and the other while inside them [32, 39]; and quenched trap (QT) - particles diffuse according to a space-dependant two-state model, switching between motion and immobilization by traps [31, 40]. Trajectories are at most 200 frames with the minimum segment length being 3 (minimum number of time steps before a change of state or end of trajectory).

Our dataset uses a balanced composition of the same nine numerical experiments of the 2024 AnDi Challenge [27], where the values of the diffusion properties (table 1) were selected to assess the participating methods while representing biologically relevant scenarios [27]. Experiment 1 mimics the multi-state diffusion found in membrane proteins, with simulation parameters reproducing the three fastest states reported for the diffusion of the α2A-adrenergic receptor [41]. Experiment 2 reproduces changes in diffusion due to protein dimerization, as has been reported for the epidermal growth factor receptor ErbB-1 [30]. Experiments 3, 4, and 5 were designed to evaluate the methods' ability to detect changes from a free diffusion state to subdiffusion caused by traps, confinement regions, and dimerization, respectively. Experiments 6 and 7 model dimerization and multi-state diffusion respectively, with both experiments using the same diffusion parameters. Experiment 8 serves as negative control and contains only single-state diffusion trajectories with incredibly broad distributions of α and K, allowing us to test U-AnD-ME performance when diffusion properties vary significantly from its initial training distribution. Experiment 9 is a quenched-trap simulation with very short trapping times and superdiffusion in the free state. In the Challenge, for a given diffusion state, the values of the anomalous-diffusion exponent α and the generalised diffusion coefficient K were randomly drawn from state-specific Gaussian distributions with bounds $\alpha \in (0, 2)$ and $K \in [10^{-12}, 10^{6}]$ pixel^2/frame, parametrized by their means (μ_α and μ_K) and standard deviations (σ_α and σ_K) (table 1).

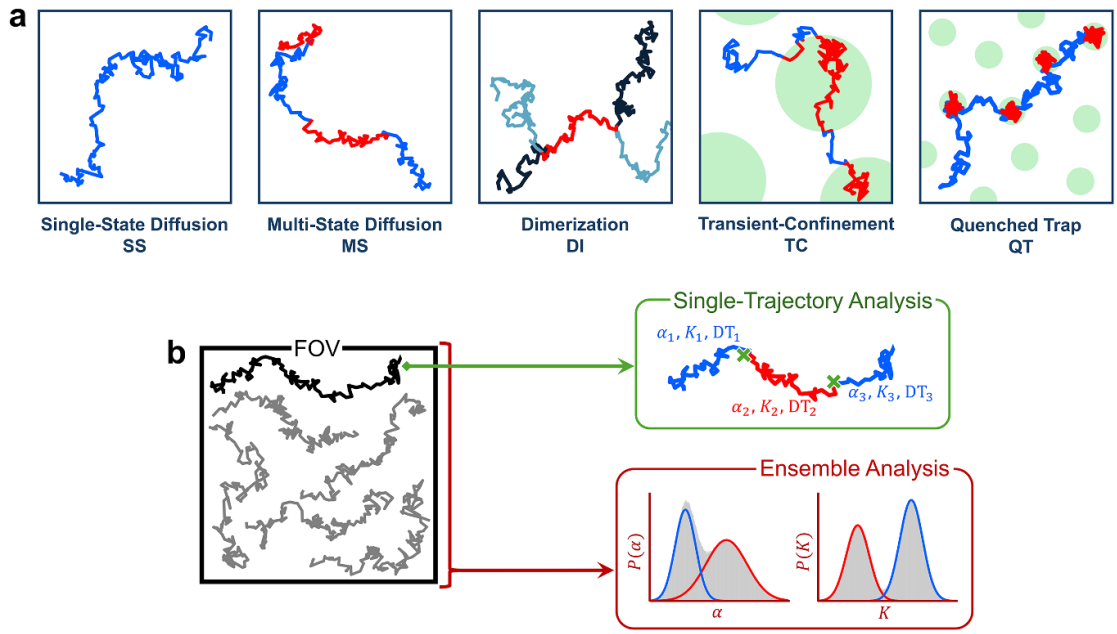

Figure 1. Overview of the 2024 AnDi Challenge. (a) The 2024 AnDi Challenge considered five physical models of diffusion [27] (left to right): single-state diffusion (SS) without change in properties; multi-state diffusion (MS) spontaneously alternating between two states (red and blue); dimerization (DI) of two particles (light and dark blue) interacting and transiently co-diffusing (red); the transient-confinement model (TC) with particles diffusing differently outside (blue) and inside (red) compartments (green) with osmotic boundaries; and the quenched-trap model (QT) with particles (blue) transiently immobilised (red) by traps (green). (b) In the Challenge [27], a field of view (FOV) is composed of several trajectories (left). In the Trajectory Track, these trajectories can be analysed individually (Single-Trajectory Task) or as an ensemble (Ensemble Task). In the Single-Trajectory Task (top right), each trajectory is analysed by detecting change points (green crosses), and, for each segment they demarcate, by inferring the anomalous-diffusion exponent α, the generalised diffusion coefficient K, and the diffusion type (DT). In the Ensemble Task (bottom right), analysis of an ensemble of trajectories returns the distributions for α and K, $P(\alpha)$ and $P(K)$.

Table 1. Simulated experimental properties. Columns show the diffusion model of each numerical experiment, along with the diffusion properties (mean μ and standard deviation σ of the anomalous-diffusion exponent α and the generalized diffusion coefficient K) and weights of each diffusion state.

Exp.	Model	μ_α	σ_α	μ_K	σ_K	Weight
1	MS	1.00	0.0001	0.15	0.01	0.30
		1.00	0.01	0.33	0.001	0.49
		1.00	0.01	0.95	0.01	0.21
2	DI	1.00	0.1	0.28	0.001	0.76
		1.10	0.01	0.0035	0.0001	0.24
3	QT	0.00	0.0	0.0	0.0	0.15
		1.00	0.005	1.0	0.1	0.85
4	TC	0.20	0.001	0.01	0.001	0.56
		1.00	0.005	1.0	0.1	0.44
5	DI	0.20	0.001	0.01	0.001	0.31
		1.00	0.005	1.0	0.1	0.69
6	DI	0.70	0.1	0.1	0.1	0.86
		1.20	0.001	1.0	0.01	0.14
7	MS	0.70	0.1	0.1	0.1	0.74
		1.20	0.01	1.0	0.01	0.26
8	SS	1.00	10	1.0	100	1.00
9	QT	0.00	0.0	0.0	0.0	0.58
		1.99	0.01	1.0	0.01	0.42

As in the Challenge, the structure of our dataset mirrors that of typical experimental data [27]. Each simulated experiment is composed of three hundred fields of views (FOVs), ten times more than in the Challenge dataset. Each FOV represents a 128×128 pixel^2 region where trajectory recording takes place, and encompasses approximately eighty trajectories on average. To better represent measurements from real tracking experiments, the trajectories are corrupted using Gaussian noise with zero mean and a standard deviation $\sigma = 0.12$ pixels. Particles within the same FOV can interact with one another and/or with the FOV's environment.

In the Trajectory Track of the Challenge, experiments could be analysed in two distinct ways [27], based on predictions at either single-trajectory (Single-Trajectory Task) or ensemble level (Ensemble Task) (figure 1(b)). Our framework allows for both types of predictions. Single-trajectory predictions involve the detection of all the change points (CPs) within a trajectory and, for each segment these CPs demarcate, the inference of K, α and DT – an identifier of what kind of constraint is imposed by the environment: immobile = 0, confined = 1, free = 2 (unconstrained, $0.05 \leqslant \alpha < 1.9$), directed = 3 ($1.9 \leqslant \alpha < 2.0$). Ensemble predictions describe each experiment collectively, capturing the distributions of α and K across all of its trajectories.

2.2. Evaluation metrics

Single-trajectory predictions used two different metrics to assess the detection of CPs [27]: The Jaccard similarity coefficient ($\mathrm{JSC_{CP}}$, equation (1)) and the root mean squared error ($\mathrm{RMSE_{CP}}$, equation (2)). Given a ground-truth CP at location $t_{\mathrm{GT},i}$ (with i an integer) and a detection at location $t_{\mathrm{P},j}$ (with j an integer), a gated absolute distance is defined as $d_{i,j} = \min(|t_{\mathrm{GT},i} - t_{\mathrm{P},j}|, \varepsilon_{\mathrm{CP}})$, where $\varepsilon_{\mathrm{CP}} = 10$ is a fixed penalty for CPs more than $\varepsilon_{\mathrm{CP}}$ apart. The number of detected CPs may not always match the number of true ones. In these cases, CPs were assigned as a rectangular assignment problem using the Hungarian algorithm [42] by minimising the sum of distances between paired CPs, $d_{\mathrm{CP}} = \min_{\text{paired CP}} \left(\sum d_{i,j}\right)$. After this assignment, we calculated the number of true positive (TP), false positive (FP) and false negative (FN) detections. A detection was considered a TP if it was within $\varepsilon_{\mathrm{CP}}$ of its paired ground-truth value. Predictions not associated with any ground-truth values or more than $\varepsilon_{\mathrm{CP}}$ away from their assigned value were considered FP. Ground-truth CPs with no assigned detection within $\varepsilon_{\mathrm{CP}}$ were considered FN. The overall number of TP, FP and FN was used to calculate the Jaccard similarity coefficient for the change-point detections over an experiment:

$$\mathrm{JSC_{CP}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{1}$$

$\mathrm{JSC_{CP}}$ takes values between 0 and 1, with 1 being a perfect score. The root mean squared error of the TP detections was also calculated:

$$\mathrm{RMSE_{CP}} = \sqrt{\frac{1}{N}\sum_{\mathrm{TP}} \left(t_{\mathrm{GT},i} - t_{\mathrm{P},j}\right)^2} \tag{2}$$

where N is the number of TPs. Lower $\mathrm{RMSE_{CP}}$ values indicate detection with lower localization error. Together, the two metrics quantified the quality of CP predictions in terms of both accuracy and resolution [27].

After identifying CPs, inference of α, K and DT can be made for each segment they delineate. For the N paired segments, inference of the anomalous-diffusion exponent α was evaluated via a mean absolute error (MAE):

$$\mathrm{MAE}_\alpha = \frac{1}{N}\sum_{\mathrm{seg}} |\alpha_{\mathrm{GT},i} - \alpha_{\mathrm{P},j}| \tag{3}$$

where $\alpha_{\mathrm{GT},i}$ and $\alpha_{\mathrm{P},j}$ are the ground-truth and predicted values of α, respectively. Evaluation of the generalised diffusion coefficient K used the mean squared logarithmic error (MSLE):

$$\mathrm{MSLE}_K = \frac{1}{N}\sum_{\mathrm{seg}} \left[\log\left(K_{\mathrm{GT},i} + 1\right) - \log\left(K_{\mathrm{P},j} + 1\right)\right]^2 \tag{4}$$

where $K_{\mathrm{GT},i}$ and $K_{\mathrm{P},j}$ are the ground-truth and predicted values of K, respectively.
Lower values of MAE_α and MSLE_K indicate better predictions. The DT was evaluated using the F_1-score:

$$F_1 = \frac{2\mathrm{TP_c}}{2\mathrm{TP_c} + \mathrm{FP_c} + \mathrm{FN_c}} \tag{5}$$

where $\mathrm{TP_c}$, $\mathrm{FP_c}$ and $\mathrm{FN_c}$ are the TPs, FPs, and FNs with respect to segment classification. Due to the presence of class imbalance, this metric is calculated as a micro-average which aggregates the contributions of all classes [27]. F_1-score takes values between 0 and 1, with 1 being the best possible score.

Finally, ensemble predictions were evaluated using the estimated mean, standard deviation, and relative weight of each state's α and K to define the multimodal distributions $P(\alpha)$ and $P(K)$ (figure 1(b)). The

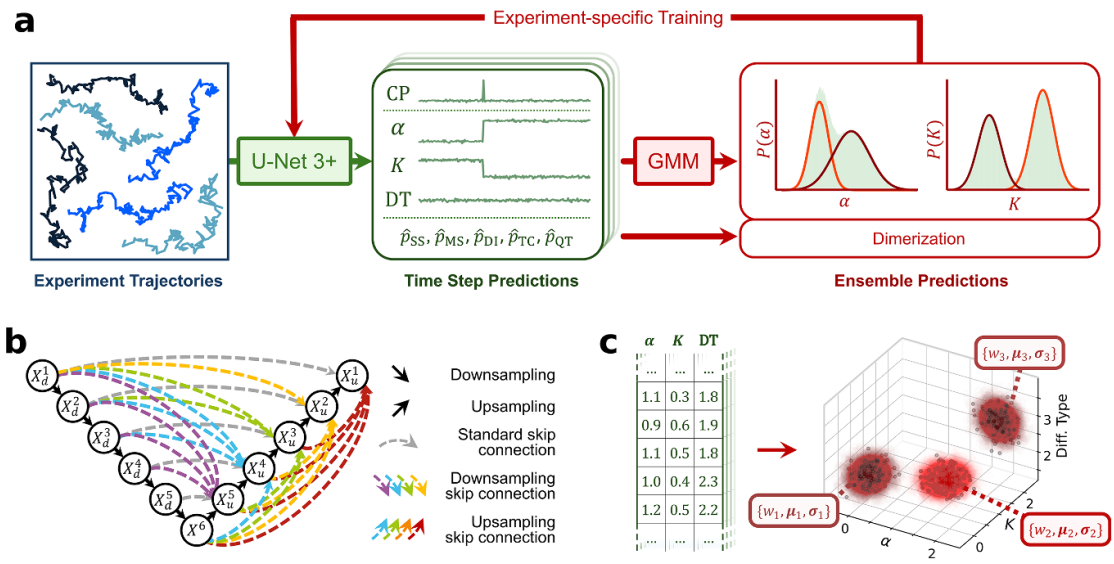

Figure 2. U-AnD-ME workflow. (a) A network inspired by U-Net 3+ processes each trajectory of an experiment. For each time step, it predicts the probability of it being a change point (CP), α, K, diffusion type (DT), and the likelihood of belonging to each of the five diffusion models ($\hat{p}_{SS}$, $\hat{p}_{MS}$, $\hat{p}_{DI}$, $\hat{p}_{TC}$, $\hat{p}_{QT}$). These time-step predictions are processed to produce trajectory-level predictions (not pictured). All diffusion model predictions in an experiment are averaged to predict its most likely model, here dimerization. Additionally, all the α, K and DT predictions in an experiment are used to create a Gaussian mixture model (GMM) and estimate the probability distributions $P(\alpha)$ and $P(K)$. These ensemble predictions inform the training of a second U-Net 3+ inspired network, making it experiment-specific and thereby more accurate. (b) Schematic of a U-Net 3+ architecture. Each X_d (X_u) is a node in the downsampling (upsampling) branch. X^6 is the bridge between the two branches. Solid downwards (upwards) black arrows represent downsampling (upsampling). Dashed grey arrows show standard skip connections. Dashed downwards (upwards) coloured arrows represent downsampling (upsampling) skip connections. The colour of these arrows indicates their output shape based on their end-point node: for example, each green arrow reshapes its input to match the size of X_u^3. (c) The predictions of α, K and DT for each time step of every trajectory in an experiment inform the creation of a three dimensional GMM (using diagonal covariance only). The parameters of each component of the GMM (weight w, mean μ and standard deviations σ) represent a different diffusion state, and, collectively, capture that experiment's ensemble properties. In the example, three components are identified.

similarity of these distributions to the ground-truth distributions $Q(\alpha)$ and $Q(K)$ (table 1) was assessed using the first Wasserstein distance [27]:

$$W_1(P,Q) = \int_{\mathrm{supp}(Q)} |\mathrm{CDF}_P(x) - \mathrm{CDF}_Q(x)|\,dx \tag{6}$$

where CDF refers to a distribution's cumulative distribution function and supp(Q) is the support, i.e. $\alpha \in (0,2)$ or $K \in [10^{-12}, 10^6]$ pixel^2/frame. $W_1(P,Q)$ approaches 0 as the accuracy of the predictions improves. Henceforth, we refer to the first Wasserstein distance of α as W_α and of K as W_K.

2.3. U-AnD-ME framework

U-AnD-ME (figure 2(a)) processing begins by using a neural network based on U-Net 3+ [43] (figure 2(b)) to make predictions for each time step in a trajectory (section 2.3.1), similar to other pointwise inference methods, such as STEP [44]. Before being passed to this network, trajectories must be preprocessed (section 2.3.2). The initial neural network is trained to handle a broad range of experimental conditions, and so we refer to it as the *generalist* network (section 2.3.3). For each time step, this network predicts the likelihood of it being a CP, estimates its α, K, and DT, and its likelihood of belonging to each of the five possible diffusion models. These time-step predictions are used to segment each trajectory and label the properties of each segment (section 2.3.4), solving the single trajectory task of the 2024 AnDi Challenge. Additionally, to make ensemble predictions (section 2.3.5), the combined time-step predictions from all the trajectories in an experiment can be used to infer the most likely diffusion model behind that experiment and for the creation of a Gaussian mixture model (GMM) capturing the experiment's α- and K-distributions (figure 2(c)). This ensemble information can be leveraged to further refine U-AnD-ME predictions (section 2.3.6). Ensemble predictions are used to generate trajectories representative of each experiment, enabling training of a new, more accurate *experiment-specific* network, also based on U-Net 3+ [43] (figure 2(b)). We implemented our framework in Python 3 using TensorFlow 2 and NumPy. All codes pertaining to U-AnD-ME are freely available under an MIT licence [45]. We chose all parameters of our architecture by optimizing the metrics of the 2024 AnDi Challenge. All computations used the same node on

the Gekko cluster of Nanyang Technology University with an Intel Skylake Xeon Gold 6150 processor and Nvidia V100 GPU for network training.

2.3.1. Architecture

U-Net 3+ (figure 2(b)) is the inspiration behind U-AnD-ME's central neural network (for both generalist and experiment-specific cases). It is a convolutional architecture, originally developed for biomedical image segmentation, consisting of an encoding/downsampling branch (figure 2(b), left) followed by a decoding/upsampling branch (figure 2(b), right), with complex skip connections interlinking the two [43].

Although U-Net architectures were originally formulated for image-segmentation tasks, their hierarchical approach to feature extraction has since been applied to time series analysis [46, 47], including to the analysis of anomalous diffusion data [48]. To adapt the architecture to time series, we replaced the two-dimensional convolutions of the original architecture with one-dimensional convolutions. The downsampling branch compresses the input and extracts coarse-grained semantic information, while the upsampling branch combines this semantic information with fine-grained details from the skip connections, enabling context-aware processing of the input. Both branches are roughly symmetric leading to its eponymous 'U' shape. As the shape of the data changes at each step, the nodes X of both branches are sometimes referred to as *scales* [43, 49]. In figure 2(b), the node X_{d}^{1} (X_{u}^{1}) is the largest scale of the downsampling (upsampling) branch, while X_{d}^{5} (X_{u}^{5}) is the smallest scale. X^{6} is the bridge between the two branches. We implement six scales in total, as this is the maximum possible with our input length. In fact, our network accepts 224×2 (time steps $\times$ dimensions) matrices as input, with any shorter trajectories padded to this length as it allows up to $N_{\mathrm{d}} = 5$ downsampling operations, thus enabling deeper and more expressive networks.

The downsampling branch follows a standard architecture for convolution networks, being composed of scales implementing repeated (valid) one-dimensional convolutions with a kernel size of 3 and stride of 1, each followed by a rectified linear unit activation function and a one-dimensional max-pooling operation. The latter operates along the time axis with a pooling size of 2 and a stride of 2, thus halving the data dimensionality at each downsampling step. After each of these downsampling steps the number of feature channels (i.e. vectors abstractly encoding extracted information) increases. We set the number of channels for X_{d}^{1}, X_{d}^{2}, X_{d}^{3}, X_{d}^{4}, X_{d}^{5} and X^{6} to 16, 32, 64, 64, 128 and 128, respectively (figure 2(b)). Every scale in the upsampling branch consists of a 1D transposed convolution along the time axis with a kernel size of 2 and a stride of 2; this operation doubles the data dimensionality, inverting the shape changes caused by the downsampling operations. Each transposed convolution in our upsampling branch uses 512 channels. Skip connections allow the output of each up-convolution to be combined with features from the downsampling branch. Every node of the upsampling branch incorporates information from its same-scale counterpart from the downsampling branch and, additionally, from all larger-scale downsampling nodes, and from all smaller-scale upsampling nodes including the bridge (figure 2(b)). This means that the skip connections must also include downsampling and upsampling operations as appropriate, reshaping their input to match the shape of the upsampling branch they connect to. Incorporating skip connections that combine scales in this way enhances the integration of coarse-grained semantic information with finely detailed information, allowing the network to better understand the context of the input.

Finally, a 1×1 convolution operates on the output X_{u}^{1} to ensure that the number of features matches the desired number of outputs: in our case, this convolution reduces the number of channels from 512 to 9, making the shape of the final output 224×9 and encoding the nine predicted feature channels for each of the 224 time steps. The first channel undergoes a sigmoid activation and represents the presence of CPs. The next three have no activation function and represent K, α and DT. The remaining five channels undergo a five-way softmax activation, and represent the probability of a time step to belong to each of the five possible phenomenological diffusion models. Experiment-specific networks can also be created once experimental ensemble properties have been predicted once (figure 2(a)). As the diffusion model is fixed for each experiment, these networks do not need to make any further model prediction, so that the 1×1 convolution after X_{u}^{1} reduces the number of channels to just 4 instead of 9 (presence of CP, α, K and DT).

2.3.2. Trajectory pre-processing

Before being fed into a network, trajectories (of maximum 200 frames) were preprocessed to simplify learning and ensure appropriate tensor sizes. Raw trajectories consist of explicit time-step labels t, with position values (x and y) for each time step (figure 3(a)). They may not span from 1 to 200, as particles may enter a FOV after $t = 1$ and leave before $t = 200$. These trajectories are first differenced in time to yield increments $(\Delta x_t, \Delta y_t) = (x_{t+1}, y_{t+1}) - (x_t, y_t)$. Processed trajectories do not contain explicit time labels t, but their index i implicitly captures this information, i.e. $(\Delta x_i, \Delta y_i) \equiv (\Delta x_t, \Delta y_t)$. Missing values are padded with zeros to a fixed length of 224 time steps (section 2.3.1).

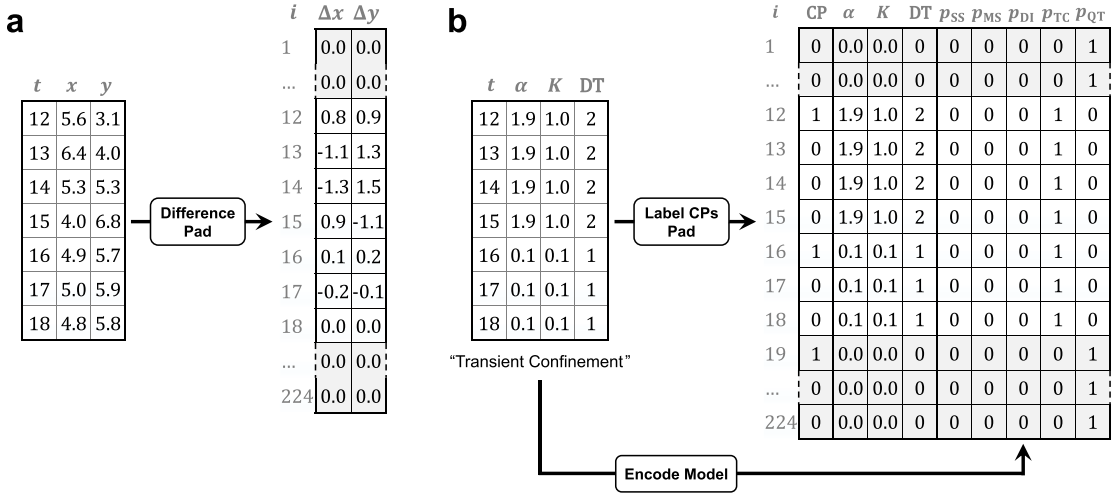

Figure 3. Trajectories pre-processing. (a) Raw trajectories (left) consist of time-step labels t and positions, x and y, for each time step. They are differenced in time to yield the increments $(\Delta x_t, \Delta y_t) = (x_{t+1}, y_{t+1}) - (x_t, y_t)$. Processed trajectories (right) do not contain explicit time labels t, but their index i implicitly captures this information, i.e. $(\Delta x_i, \Delta y_i) \equiv (\Delta x_t, \Delta y_t)$. Missing values are padded with zeros (grey) to a fixed length of 224 time steps. In the example, the original trajectory spans $t = 12$ to $t = 18$ (time when it was within the FOV). The increments $(\Delta x_t, \Delta y_t)$ are therefore defined for $i = 12$ to $i = 17$, while zeros are used to fill the otherwise undefined values, from $i = 1$ to $i = 11$ and from $i = 18$ to $i = 224$. (b) Corresponding raw labels (left) consist of the same time-step labels t, with the respective values of α, K, and DT. Additionally, a label describes the diffusion model of the trajectory, here 'transient confinement'. For both generalist and experiment-specific networks, raw labels are processed (right) by adding explicit CP labels (set to one for the start of a new segment and zero otherwise) for each time step. As for trajectories, the indices i of the new matrix captures time information. Moreover, for generalist networks (as in the depicted example), the diffusion model is one-hot encoded and also forms part of the label through the probabilities p_{SS}, p_{MS}, p_{DI}, p_{TC} and p_{QT}. In the example, for each original time step ($i = 12$ to $i = 18$), the probability of belonging to the TC mode, p_{TC}, is set to one, while all other probabilities are set to zero. CP, α, K, and DT values are padded with zeros (grey). In generalist networks, model labels are padded with zeros too bar the probability of being in a quenched trap, p_{QT}, which is set to one. This padding mimics an immobilizing trap at the FOV edge.

Corresponding raw labels (figure 3(b)) consist of the same time steps t, along with values of α, K, and DT. Additionally, a label describes the diffusion model behind the trajectory. For both generalist and experiment-specific networks, the labels for α, K and DT are used without any additional processing. As DT is an ordinal category, we simply treat it as a float variable. As for trajectories, the label index i implicitly encodes the original time information t. Padding sets the missing values of α, K and DT to zero up to the fixed length of 224 time steps of the processed trajectories. Processed labels include explicit CP information too, which is set to one for the start of any segment and for the first time step after the end of the raw trajectory; it is set to zero otherwise. Unlike experiment-specific networks, generalist networks also require information about the diffusion model. As this information is not ordinal, we apply one-hot-encoding: each time step has five labels encoding whether it belongs to each of the five diffusion models; the label for the true model is set to 1 while all others are 0. For padded time steps, the diffusion model is set as a QT, i.e. $p_{QT} = 1$ while $p_{SS} = p_{MS} = p_{DI} = p_{TC} = 0$. This padding scheme essentially treats the boundary of every FOV as an immobilizing trap, preventing the padding from adding any physically unrealistic behaviour to the training data.

2.3.3. Network training

We simulated fractional Brownian motion trajectories for training using the `andi-datasets` Python package [34]. For generalist networks, we simulated trajectories corresponding to all five diffusion models, with all their parameters randomly selected from a predefined range informed by the 2024 AnDi Challenge pilot dataset [27]. Due to the negative control nature of Experiment 8, K-values can purposely span a very broad range. Training a network over such a large range would be computationally expensive and result in generally poor performance. We therefore limited the range for training the generalist networks to the values more represented in the Challenge dataset (table 1) [27]. Trajectories were therefore simulated for each diffusion state in an experiment with values of α and K sampled from Gaussian distributions with means, $\mu_\alpha \sim \mathrm{U}(0, 1.999)$ and $\mu_K \sim \mathrm{U}(10^{-12}, 15)$, and standard deviations, $\sigma_\alpha = 0.01\mu_\alpha$ and $\sigma_K = 0.01\mu_K$. SS diffusion requires only one state. For MS diffusion, we simulated a maximum of five states, as this was comfortably larger than the maximum of three found in the pilot dataset [27]. MS diffusion also requires the definition of a transition matrix M [27]. At any time step, the transition probability from a state i to a state j is given by M_{ij}. The probability of remaining in the same state i is given by M_{ii}, which we set to a single value

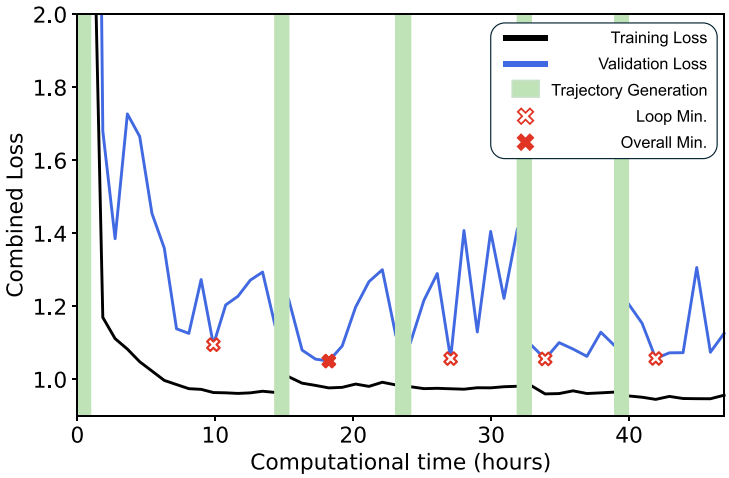

Figure 4. Training and validation loss curves. Exemplary training (black line) and validation (blue line) loss curves for U-AnD-ME. At the start of each training and validation loop (green vertical lines), 50 000 new trajectories are generated and used with a 4:1 random split between training (40 000 trajectories) and validation (10 000 trajectories). In each loop, training proceeds until validation loss (blue line) stagnates for five consecutive epochs. Following this, the parameters from the minimum loss over this loop (red crosses) are restored, new trajectories are generated, and another training loop commences. Training comes to a final stop when the validation minimum loss per loop stops improving between loops. The final network parameters are those from the overall validation loss minimum (filled red cross). The performance of the final networks was evaluated against the unseen dataset of the AnDi Challenge composed of 30 000 trajectories, with ground-truth unknown to the participants.

$M_{ii} \sim \mathrm{U}(0.9{,}0.999)$ for all states. The values for all other transition probabilities are then $M_{ij} = (1 - M_{ii})/(s-1)$, where s is the number of states. The DI model requires two states and the definition of the number of diffusing particles N, their radius r, the probability P_b that two particles bind when at a distance $d < 2r$, and the probability P_u of a dimer unbinding. We generated experiments for this model using $N \sim \mathrm{U}(50, 150)$, $r \sim \mathrm{U}(0.1, 2.0)$, $P_\mathrm{b} = 1$, and $P_\mathrm{u} \sim \mathrm{U}(0, 0.1)$. The TC model also requires two states and the definition of the number of confinement regions N_c, their radius r_c, and the transmittance probability of the boundary T. We used $N_\mathrm{c} \sim \mathrm{U}(10, 100)$, $r_\mathrm{c} \sim \mathrm{U}(1, 15)$, and $T \sim \mathrm{U}(0, 0.5)$. Finally, the QT model requires only the definition of one state, as the other is complete immobilization ($\mu_\alpha = \mu_K = \sigma_\alpha = \sigma_K = 0$). It also requires the number of traps N_t, their radius r_t, the trap binding probability P_t when at a distance $d < r_\mathrm{t}$, and the unbinding probability P_u. We used $N_\mathrm{t} \sim \mathrm{U}(100, 500)$, $r_\mathrm{t} \sim \mathrm{U}(0.1, 2.0)$, $P_\mathrm{b} = 1$, and $P_\mathrm{u} \sim \mathrm{U}(0, 0.1)$.

The training of both generalist and experiment-specific networks followed the same procedure (depicted in figure 4), but the simulation parameters used for experiment-specific training came directly from the ensemble predictions of the generalist network for that experiment (see also section 2.3.6). As the networks' output is multimodal, we used several different loss functions in unison for training: a binary cross-entropy loss for the binary classification task of detecting CPs; mean squared error losses for the regression tasks of inferring α, K and DT; and, only for generalist networks, categorical cross-entropy for the multi-class classification task of predicting the diffusion model.

We trained each network of U-AnD-ME through training loops (figure 4). For each loop, we generated and used a total of 50 000 new trajectories of known ground-truth with a 4:1 split between training (40 000 new trajectories per network per loop) and validation (10 000 new trajectories per network per loop). Each loop continued until the validation loss stagnated for five consecutive epochs. Training came to a final stop when the minimum validation loss per loop stopped improving between loops. We finally selected the network parameters from the overall validation minimum (figure 4). Both generalist and experiment-specific networks reached their validation minimum after ca. 18 hours on our computational resources (typically corresponding to two training loops per network). The final networks were tested against the unseen dataset (30 000 trajectories) of the AnDi Challenge, with ground-truth unknown to the participants.

2.3.4. Single-trajectory predictions

The single-trajectory prediction procedure was identical for generalist and experiment-specific networks (figure 5). For each trajectory, after removing padded time steps, CPs were detected first. We considered a time step to be a CP if its CP label was at least 0.25 and a local maximum compared to its immediate neighbours. CPs within two time steps of the start or end of the unpadded output were ignored, as the minimum possible segment length was three. The output tensor was split into segments according to these identified CPs (figure 5). The values of α, K and DT for each time step in a segment were averaged to generate a singular prediction for that segment. This average used a parabolic weighting, where time steps

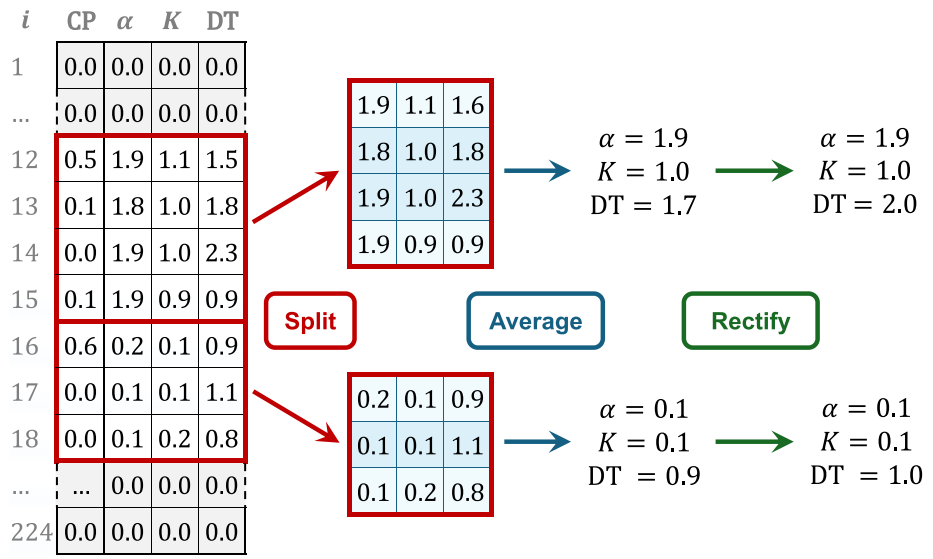

Figure 5. U-AnD-ME single-trajectory prediction procedure. The network output consists of predictions of change point (CP) probability, α, K, and diffusion type (DT) for each of the 224 time steps of the processed trajectories. This output includes padded values (grey). After padding is reverted, the output is split into segments using the change point predictions (red boxes). Each identified segment then undergoes a weighted average (blue shades), with its central time steps assigned higher weights, which yields estimates for α, K, and DT. Finally, these values are rectified, being constrained and rounded as appropriate.

near the centre of the segment contributed more than those at its extremities as errors in the CP localisation make them less reliable (figure 5). For a segment spanning i_{start} to i_{end}, each time step's weighting w_i is given by $w_i = \frac{3}{10}(2 - \tilde{i}^2)$, where $\tilde{i} = 2\frac{i - i_{\text{start}}}{i_{\text{end}} - i_{\text{start}}} - 1$ is a mapping of i from $[i_{\text{start}}, i_{\text{end}}]$ to $[-1, 1]$ and the prefactor $\frac{3}{10} = (\int_{-1}^{1} 2 - \tilde{i}^2 \, d\tilde{i})^{-1}$ is a normalisation factor ensuring that $\sum_{i=i_{\text{start}}}^{i_{\text{end}}} w_i = 1$. Finally, prediction of segment properties were rectified (figure 5) by constraining all values to within physically possible/realistic ranges, and ensuring that they were of an appropriate data type. Predictions for α and K were constrained to $(0, 2)$ and $[10^{-12}, 10^{6}]$, while those for the DT were rounded to the nearest integer and then constrained to $[0, 3]$ (section 2.1).

2.3.5. Ensemble predictions

Ensemble predictions aim at approximating each experiment's multimodal distributions of α and K, P_α and P_K (table 1, figure 2(a)). Using standard expectation maximisation [50], we fitted a GMM to the joint distribution of all values of α, K and DT predicted by our generalist network for all time steps of the trajectories in an experiment (figure 2(c)). In our case, we used this GMM to approximate the multimodal distributions of α, K, and DT as a sum of Gaussian components, where each component i is characterized by means $\mu_{\alpha,i}$, $\mu_{K,i}$, and $\mu_{\text{DT},i}$, standard deviations $\sigma_{\alpha,i}$, $\sigma_{K,i}$, and $\sigma_{\text{DT},i}$, and a weight w_i. While the properties of the DT distribution are not strictly necessary, considering them led to better separation between different diffusion states, and, thus to a better accuracy for the captured distributions P_α and P_K. Our GMM used strictly diagonal covariance matrices, meaning the three variables (α, K and DT) are independent and have different standard deviations. In the ideal case, the number of Gaussian components will match that of diffusion states in an experiment exactly. In practice, this is rarely the case, yet we found that the overall GMM distributions over α and K still closely approximate the experimental distributions. We created GMMs with between one and ten components, and finally selected that with the lowest Bayesian information criterion [51]. We set ten as the maximum number of components as this led to strong performance with good efficiency. Finally, we also used the outputs of the generalist network to predict the probability of an experiment to belong to each of the five possible diffusion models. We used this information to train our experiment-specific networks together with the probabilities P_α and P_K defined by the GMM (section 2.3.6). To avoid error propagation from poor diffusion model predictions, we never used this information to decide the number of GMM components.

2.3.6. Experiment-specific networks

Once approximate experimental properties were available from the predictions of generalist networks, we trained and used experiment-specific networks to improve accuracy for both single-trajectory and ensemble predictions. Training proceeded as outlined in section 2.3.3 but with new trajectories simulated in

Table 2. Ablation study summary. Normalized metric scores ($\tilde{\text{JSC}}_{\text{CP}}$, $\tilde{\text{RMSE}}_{\text{CP}}$, $\tilde{\text{MAE}}_{\alpha}$, $\tilde{\text{MSLE}}_{K}$, $\tilde{F}_1$, $\tilde{W}_{\alpha}$ and $\tilde{W}_{K}$), mean inference time per trajectory ($\tilde{t}_{\text{inf}}$)), and normalized mean reciprocal rank (MRR) for the different variants of U-AnD-ME architecture tested in the ablation study. We tested the following changes to U-AnD-ME basic architecture: not differencing the inputs in the preprocessing ('No differencing'), using a constant number of channels across scales ('No taper') as opposed to the reverse-tapered shape used in the original network (number of channels for X_{d}^1, X_{d}^2, X_{d}^3, X_{d}^4, X_{d}^5 and X^6 all set to 90, as opposed to 16, 32, 64, 64, 128 and 128 of the original network), using a wider network ('No taper + wider', using only 5 scales, all with 640 channels), using a network with a larger convolutional kernel ('Larger kernel', kernel size of 5 instead of 3), as well as removing the parabolic weighting from the post-processing ('No weighting'). The metrics scores are normalized to the respective scores for the basic U-AnD-ME architecture.

	$\tilde{\text{JSC}}_{\text{CP}}$	$\tilde{\text{RMSE}}_{\text{CP}}$	$\tilde{\text{MAE}}_{\alpha}$	$\tilde{\text{MSLE}}_{K}$	$\tilde{F}_1$	$\tilde{W}_{\alpha}$	$\tilde{W}_{K}$	$\tilde{t}_{\text{inf}}$	MRR
U-AnD-ME	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.45
No differencing	0.65	0.53	1.15	1.20	0.24	0.50	1.08	1.00	0.37
No taper	0.66	1.73	0.76	1.12	1.04	0.55	0.74	1.00	0.44
No taper + wider	0.61	2.65	1.35	1.59	0.04	0.41	1.35	1.00	0.31
Larger kernel	0.91	0.83	0.55	0.73	0.91	0.81	0.50	1.36	0.67
No weighting	1.00	1.00	0.99	1.08	0.99	1.00	1.00	1.00	0.42

accordance with the predicted diffusion model using ranges of α and K defined by the distributions predicted by the generalist GMM (section 2.3.5).

Ideally, a single model will be predicted for the experiment with high confidence. However, often multiple models show comparable probabilities, so it is important not to overcommit to just the most probable one, as this risks training the network on the wrong model. Therefore, we trained experiment-specific networks using trajectories from multiple models selected from the softmax output $\{\hat{p}_{\text{SS}}, \hat{p}_{\text{MS}}, \hat{p}_{\text{DI}}, \hat{p}_{\text{TC}}, \hat{p}_{\text{QT}}\}$ of the generalist network encoding the probability of the experiment being SS, MS, DI, TC or QT, respectively. We sorted this set of probabilities into descending order and calculated the difference between any two consecutive probabilities in the sorted list. Training used all models up to the first where this difference exceeded the predefined threshold value of 0.1. We used this value because it stroke an acceptable balance between accuracy (the generated set of diffusion models generally included the true model) and specificity (the generated set of diffusion models was small) on a trial dataset.

Simulating trajectories of an experiment for training requires the definition of μ_{α}, μ_K, σ_{α} and σ_K for each diffusion state in it (section 2.3.3). While the training of the generalist network uses random values for these parameters, when training experiment-specific networks, these values come directly from the components of the generalist GMM capturing the experiment's α- and K-distributions (section 2.3.5). When there were more predicted GMM components than experimental states in the model being considered, the trajectory ensemble generated for training used random subsets of all the predicted components. For example, in a DI experiment with exactly two states and a GMM with five predicted components, each FOV generated for training would use a random set of two components from the five predicted.

2.4. Ablation study

We conducted ablation tests to isolate how individual architectural choices impact U-AnD-ME performance. We tested the following changes to U-AnD-ME basic architecture: not differencing the inputs in the preprocessing ('No differencing' in table 2), using a constant number of channels across scales ('No taper' in table 2) as opposed to the reverse-tapered shape used in the original network (number of channels for X_{d}^1, X_{d}^2, X_{d}^3, X_{d}^4, X_{d}^5 and X^6 all set to 90 for 'No taper', as opposed to 16, 32, 64, 64, 128 and 128 of the original network), using a wider network (using only 5 scales, all with 640 channels, 'No taper + wider' in table 2), using a network with a larger convolutional kernel (kernel size of 5 instead of 3, 'Larger kernel' in table 2), as well as removing the parabolic weighting in the post-processing ('No weighting' in table 2). For better comparison between the variants, we maintained the number of trainable parameters approximately constant apart from the 'larger kernel' network which had 45% more trainable parameters due to the larger kernel size. To reduce the overall computational costs associated with running these tests, all variants were trained using only one iteration of 50 000 trajectories (40 000 training, 10 000 validation), as opposed to the iterative training procedure detailed in section 2.3.3. Evaluation used 30 simulated FOVs per experiment, directly matching the AnDi Challenge evaluation, whereas our other results use 300 simulated FOVs per experiment. Table 2 reports the scores of each scenario relative to the basic U-AnD-ME architecture, evaluating the same metrics of the AnDi Challenge as well as the mean inference time per trajectory. The different scenarios are then compared using the mean reciprocal rank (MRR, table 2). While individual tests can outperform the basic U-AnD-ME architecture for specific metrics, this architecture ranks 2$^{\text{nd}}$ in terms of MRR, striking a good balance between the overall metrics performance and the computational time needed to train. The 'no taper' version ranks close to U-AnD-ME with better performance in inferring diffusion properties (both at the single-trajectory and ensemble level) but significantly worse CP predictions. This is

because giving every encoder scale the same number of channels shifts most of the network's capacity into the early, high-resolution layers. This allows for better learning of the statistics determining diffusion parameters, but leaves fewer parameters for deeper layers that aim to aggregate long-range context, thus making noticing changes in diffusion regime more difficult. It should be noted that even the highest-ranking network (i.e. the one with a larger convolutional kernel) did not outperform U-AnD-ME on all metrics. It performed better in predicting diffusion properties (α and K) and in the ensemble task (W_α and W_K) due to the wider kernel enabling better integration of information across time steps, yet it performed worse in determining the DT (F_1) and in detecting CPs (JSC_{CP}). Interestingly, time resolution ($RMSE_{CP}$) is better, but this metric alone can be misleading in cases with lower JSC_{CP} as it is only calculated on TP detections (section 2.2). Importantly, U-AnD-ME keeps inference time relatively low compared to this variant (approximately 25% faster), thus becoming our architecture of choice in the time-sensitive context of the Challenge.

3. Results

Inspired by the 2024 AnDi Challenge, we evaluate U-AnD-ME using a well-balanced dataset with approximately 216 000 trajectories representative of a wide range of biologically relevant phenomena (section 2.1, table 1). We first evaluate how impactful the framework's experiment-specific training is (section 3.1). We then discuss the quality of the single-trajectory analysis, including the CP detection, the inference of the diffusion properties (the anomalous-diffusion exponent α and the generalized diffusion coefficient K, section 3.3), and the classification of the DT (section 3.4). Finally, after discussing the ensemble predictions (section 3.5), we apply U-AnD-ME to infer anomalous diffusion properties in experimental trajectories (section 3.6).

3.1. Generalist vs. experiment-specific networks

Figure 6 compares the performance of the generalist network to that of experiment-specific ones which leverage ensemble predictions. For all metrics, experiment-specific networks led to improvements in average performance (figures 6(a)–(g)), highlighting their benefit over the generalist because of a more relevant selection of diffusion properties for training. In most cases, there is also a strong correlation between the improvement of paired metrics going from generalist to experiment-specific networks (figures 6(h)–(j)): for CP detection, an improved JSC_{CP} typically comes with a lower $RMSE_{CP}$ (figure 6(h)); a similar trend emerges for the joint predictions of the trajectories' diffusion properties (figure 6(i)) and of their ensemble distributions (figure 6(j)). Only the $RMSE_{CP}$ for Experiments 1, 6 and 7, W_K for Experiment 8 (due to its broad range of K values) and the F_1-score for Experiment 6 deviate from this trend, with the decrease seen for this last metric being negligible.

For $RMSE_{CP}$, experiment specificity led to worse performance for Experiments 1, 6 and 7, as they contain more complex CPs to detect due to MS diffusion or DI. As $RMSE_{CP}$ is calculated only on TP CPs, the better $RMSE_{CP}$ less specific networks exhibit is not caused by superior CP detection but by them simply failing to detect more difficult CPs entirely. Generalist networks detect only more distinct CPs and make good predictions on these with relatively low localization error (hence the relatively higher values of $RMSE_{CP}$), while experiment-specific networks additionally pick up more subtle CPs that are harder to localize well (hence the relatively lower values of $RMSE_{CP}$). This can lead to an anti-correlation between JSC_{CP} and $RMSE_{CP}$ where favourable (high) JSC_{CP} can be accompanied by poor (high) $RMSE_{CP}$, as in figure 6(h) for Experiments 1, 6 and 7. This counter-intuitive behaviour is why both $RMSE_{CP}$ and JSC_{CP} were used in tandem to evaluate CP detections in the 2024 AnDi Challenge [27].

Interestingly, for Experiment 8 (serving as negative control to the other experiments [27]), the values of most metrics changed little between generalist and experiment-specific networks. In particular, the latter led only to negligible improvements for JSC_{CP}, $RMSE_{CP}$, and F_1-score. The relatively simpler dynamics of the SS model (single diffusion state with no CPs) meant that even our generalist network could get close to optimal values for these metrics, thus reducing the need for experiment-specificity. Experiment 9 instead stands out due to the improvement that experiment-specific networks introduced over the generalist one on all metrics of the single-trajectory task (figures 6(a)–(e) and (h)–(i)). This experiment in fact possesses two diffusion states that have very different properties (trapped and near ballistic) with narrow distributions (table 1), thus justifying the greatest benefits over other experiments introduced by a better selection of parameters due to experiment-specific training.

For benchmarking, we also compare U-AnD-ME performance in inferring α and K against logarithmic (for α) and linear (for K) fits of the time-averaged MSD (MSD, figures 6(c) and (d)). To this end, focusing on Experiment 8's single-state trajectories enables us to assess U-AnD-ME ability to infer these diffusion parameters avoiding segmentation errors [27]. Under these conditions of absence of CPs, U-AnD-ME experiment-specific networks perform better than the MSD fit, particularly for the estimation of K: while the

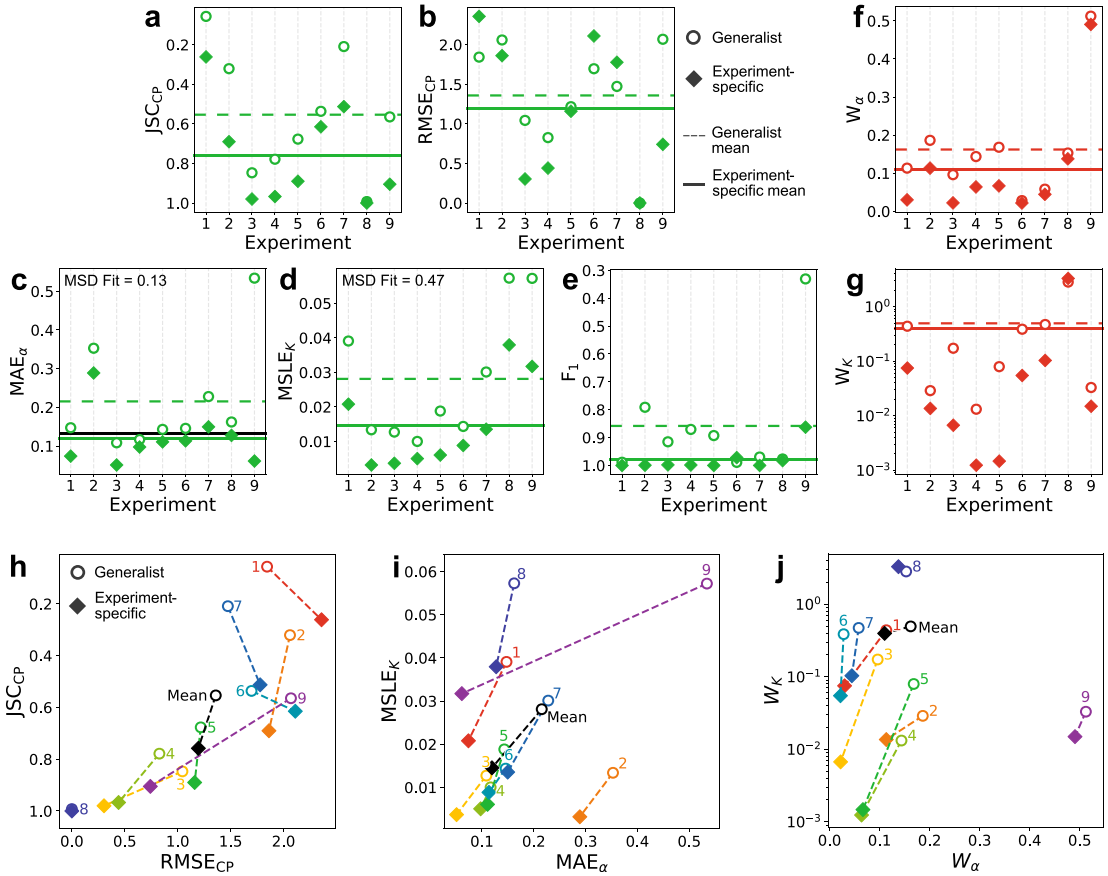

Figure 6. Impact of experiment specificity. (a)–(e) Performance by experiment (numbers as in section 2.1) for each metric of the single-trajectory analysis: (a) JSC_{CP}, (b) $RMSE_{CP}$, (c) MAE_α, (d) $MSLE_K$, and (e) F_1 score. In c and d, the scores obtained by using the fit (MSD Fit) of the time-averaged mean-squared displacement of the single-state trajectories of Exp. 8 are provided as a benchmark; the corresponding line is shown in black for c but not for d as this value lies well beyond the plot limits. (f) and (g) Performance by experiment for each metric of the ensemble analysis: (f) W_α, and (j) W_K. The y-axes show better metric values at the bottom of each plot. Dashed and solid horizontal lines show average performance for generalist and experiment-specific networks, respectively. (h)–(j) Correlation between subtask metrics for (h) change-point detection, (i) inference of diffusion properties and (j) ensemble predictions. The numbers indicate individual experiments. Dashed lines connect each experiment's generalist and experiment-specific metrics' values. Mean values are shown in black. Axes show better metric values at the bottom left of each plot. (a)–(j) Hollow circles and filled diamonds represent metrics values of generalist and experiment-specific networks, respectively.

value for MAE_α from the MSD fit ($MAE_\alpha = 0.133$) is only slightly worse than that from U-AnD-ME ($MAE_\alpha = 0.130$), $MSLE_K$ from the MSD fit ($MSLE_K = 0.478$) is an order of magnitude worse than that from U-AnD-ME ($MAE_K = 0.038$). The results of the AnDi Challenge further benchmark our method against those of the other 17 participating teams in Track 1 [27, 52].

3.2. Change-point detection

U-AnD-ME outperformed all other methods in the 2024 AnDi Challenge in terms of CP detection, scoring better in terms of both JSC_{CP} (accuracy) and $RMSE_{CP}$ (localization) [27]. The JSC_{CP} (figure 7(a)) and the corresponding $RMSE_{CP}$ (figure 7(b)) show that CP detection was highly performant, being both accurate (JSC_{CP}) and precise ($RMSE_{CP}$), for two models (TC and QT) with $JSC_{CP} > 0.91$ and a $RSME_{CP} < 0.72$ time steps. As a reference, all submitted methods in the segmentation task of the previous 2020 AnDi Challenge obtained $RSME_{CP}$ values of at best 10–20 time steps (an order of magnitude larger) when only considering the subset of trajectories with CPs at least 20 time steps away from their start/end [34]. U-AnD-ME's particularly high performance for TC and QT is due to these two models showing two very clearly different states. Unlike the other models with CPs, TC and QT transitions always involve one segment with very low mobility—near zero for TC and precisely zero for QT. This leads to more distinct CPs, reducing mislabelling from the network, as can be seen from the relatively low number of FP (FP, figure 7(c)) and FN (FN, figure 7(d)) detected CPs compared to the other two models (MS and DI). This common segment feature in the CPs may also facilitate a more effective training generalization among different trajectories.

The MS model shows the poorest performance with $JSC_{CP} > 0.32$ and a $RSME_{CP} < 2.15$ time steps as, unlike other models, the CPs in MS trajectories can involve a variety of states, which naturally adds

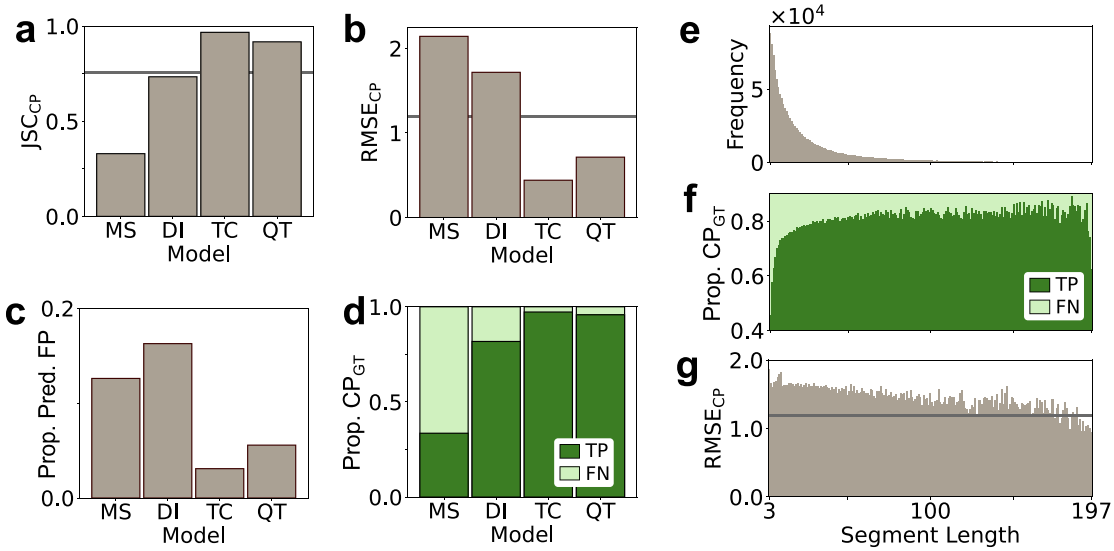

Figure 7. Change-point detection. (a) and (b) Metrics for change-point detection from experiment-specific networks for the four models that exhibit change points (MS, DI, TC and QT) in terms of (a) JSC_{CP} and (b) $RMSE_{CP}$. Horizontal lines represent the average value of each metric. (c) The proportion of false positives (FP) over all detected change points. (d) The relative proportion of true positives (TP, dark green) and false negatives (FN, light green) over all ground-truth change points (CP_{GT}). (e) The frequency of different segment lengths across our dataset. Shorter segments are more represented as predictions are more challenging due to the lower information content per segment. (f) The relative proportion of true positives (TP, dark green) and false negatives (FN, light green) across all ground-truth change points (CP_{GT} as in d) by segment length. (g) $RMSE_{CP}$ of the change-point detections by segment length. Every change point is assigned to two segments, the ones immediately preceding and following it. The horizontal line shows the metric's average value.

complexity to their identification. Also, unlike TC or QT, these transitions can be between high mobility states. The relatively poorer JSC_{CP} for MS is caused both by CPs being missed (low TP and high FN, figure 7(d)) and by a relatively high proportion of false CPs being identified (FP, 7(c)).

DI, with only two diffusion states as TC and QT, fares in-between as the change in diffusion properties is not always as marked as for these other two models (table 1). Like MS, DI can also include transitions between high-mobility states, but, differently from this model, penalization in detection comes from a higher proportions of FP values rather than disproportionally mislabelling true CPs (figures 7(c) and (d)).

Figures 7(e)–(g) explore the influence of segment length on CP detection. As in the 2024 Challenge, the evaluation dataset is richer in shorter segments (figure 7(e)); this mirrors the fact that single-molecule live-imaging data often contain short trajectories. Additionally, this segment length distribution allows training to focus on identifying and characterising shorter segments, which are known to be more challenging and easier to miss due to their lower information content [4, 48]. As can be expected, prediction quality increases with segment length (figures 7(f) and (g)), confirming that the improved feature information provided by longer trajectories has a notable impact on CP identification [4]. U-AnD-ME struggles most for very short segments (< 10 time steps, figures 7(f) and (g)). The proportion of TP CPs over FNs across all ground-truth values increases steeply for increasing lengths, until it plateaus at approximately 83% for segments longer than 65 time steps (figure 7(f)). Similarly, the $RMSE_{CP}$ shows a roughly linear halving from 1.86 to 0.94 time steps with increasing segment lengths (figure 7(g)). Interestingly, for very long segments (> 194 time steps), while the localization error is low ($RMSE_{CP} < 1$ time steps (figure 7(g)), there is a higher proportion of FN points compared to slightly shorter segments (figure 7(g)). This is an artefact coming from the high proportion of FN points identified on very short segments. As the maximum trajectory length is 200, any detection shortcoming in trajectories with a single CP associated with a very short segment is also bound to propagate to its longer counterpart [4].

3.3. Inference of the diffusion parameters

U-AnD-ME also outperformed all other participating methods in the 2024 AnDi Challenge in terms of inferring both α and K, achieving the lowest MAE_α and $MSLE_K$ [27]. On our dataset, we achieved averages of $MAE_\alpha = 0.12$ and $MSLE_K = 0.015$. Inference strongly improves with segment length, due to the aforementioned increased feature information that longer segments contain (figures 8(a) and (b)) [4]. After an initial steep improvement with segment length, MAE_α plateaus at ≈ 0.7 for segments longer than 150 time steps, while $MSLE_K$ keeps slowly improving (figure 8(b)).

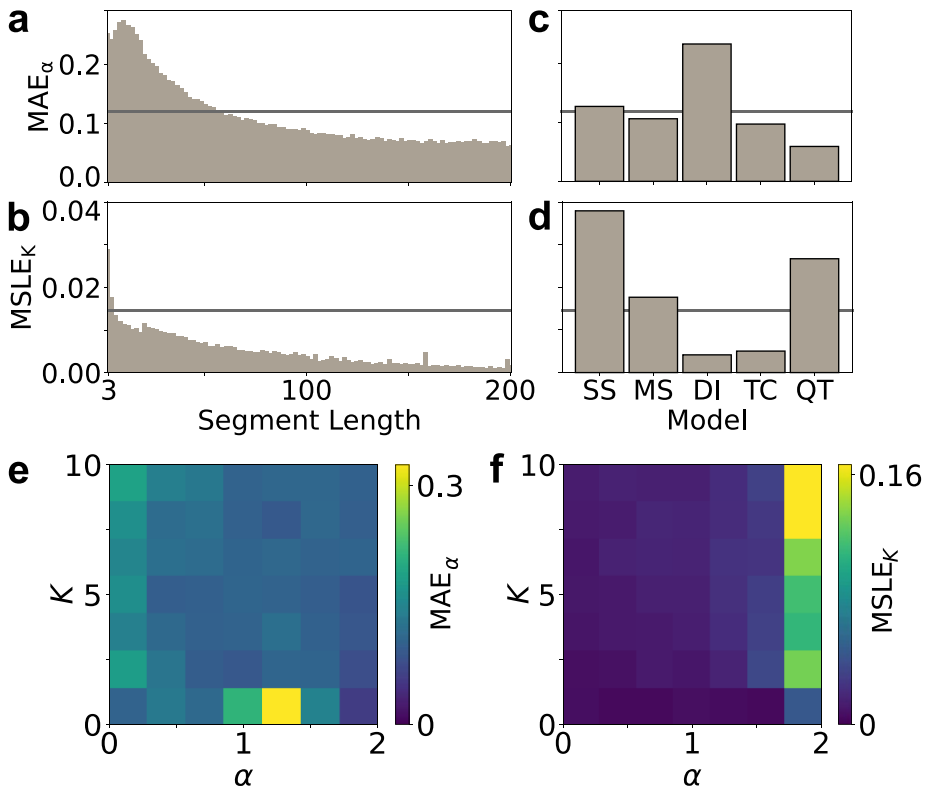

Figure 8. Inference of the diffusion properties. (a) MAE_α and (b) $MSLE_K$ as a function of ground-truth segment length. (c) MAE_α and (d) $MSLE_K$ for each of the five possible diffusion models. Horizontal lines show average values. (e) and (f) Heat-maps showing (e) MAE_α and (f) $MSLE_K$ as a function of α and K.

When evaluating the inference of α by model (figure 8(c)), MAE_α for SS, MS, and TC are all comparable and close to the average value. QT shows the best performance ($MAE_\alpha = 0.06$), likely because inferring α is straightforward for the immobilized state ($\alpha = 0$). The inference of segment properties can also be expected to be most effective when CPs are accurately detected (figure 7), which slightly benefited TC too (section 3.2). At the opposite end, DI was the worst performing model ($MAE_\alpha = 0.23$). This is attributed largely to Experiment 2 which, unlike the two other DI experiments (Experiments 5 and 6) has two overlapping states (diffusion and directed) with very similar values of α, leading to mislabelling the less frequent directed state (table 1). In fact, this experiment has the worst MAE_α of all even after using experiment-specific networks (figure 6(c)). Interestingly, the inference of K by model (figure 8(d)) shows much more variability than that of α. MS performance is the only one close to the $MSLE_K$ average value, with the other models performing much better or worse than average. TC and DI performed best, both with $MSLE_K$ scores of ca. 0.005, as all experiments of these models have well-separated bimodal distributions of K (table 1, section 3.5), which U-AnD-ME could discern well. On the opposite end, SS performed the worst, likely due to the range of K values in its only experiment (the negative-control Experiment 8) being significantly broader than any other experiment. Although better than SS, QT performance was also poorer than the average, hampered by Experiment 9 with an extremely superdiffusive state ($\mu_\alpha = 1.99$). We believe that its very directed trajectories have indeed influenced our network's capability to finely resolve the correct width of the distribution of K-values (see also section 3.5).

Finally, figures 8(e) and (f) show how the values of MAE_α and $MSLE_K$ vary with α and K. Performance is generally quite consistent across different values with two exceptions: MAE_α at $\alpha \approx 1.2$ for low values of K ($K < 1.5$) (figure 8(e)) and $MSLE_K$ for strongly superdiffusive trajectories as K grows (figure 8(f)). Both deviations are however not due to the actual values of α and K, but rather to their representation in the training dataset and the shape of the experimental distributions to be resolved by U-AnD-ME (section 3.5). In fact, while the latter is due to Experiment 8 (the negative-control experiment) and its atypically broad range of K-values compared to the other experiments (table 1), the former is due to U-AnD-ME struggling to resolve the multimodal distributions of Experiments 6 and 7, both featuring a secondary narrow peak at $\alpha = 1.2$ on a much broader underlying distribution centred at $\alpha = 1$ (table 1, section 3.5).

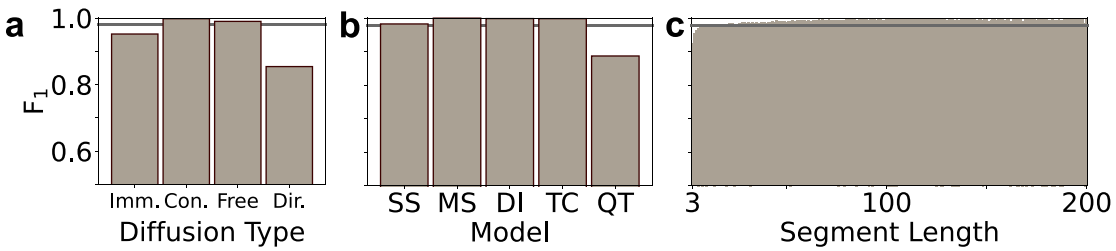

Figure 9. Classification of diffusion type. F_1-score of the classification of diffusion type (DT) by (a) diffusion type (immobilized, confined, freely diffusing, or directed), (b) diffusion model, and (c) ground-truth segment length. The solid lines show the average $F1$ score across all experiments.

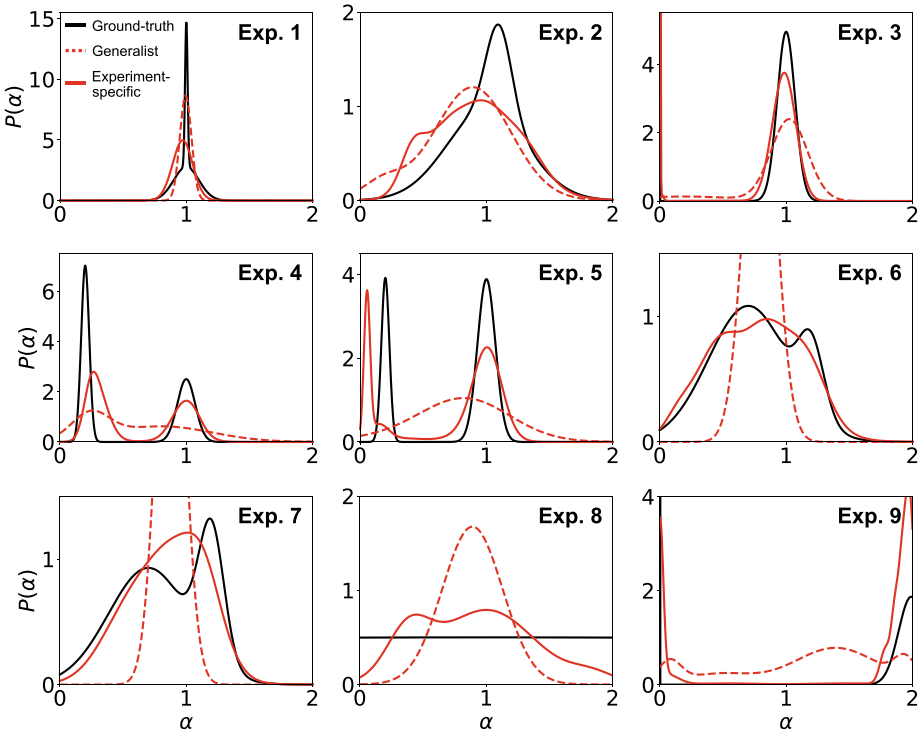

Figure 10. Predicted α-distributions from ensemble analysis. Ground-truth α-distributions $P(\alpha)$ (black lines), distributions predicted by generalist networks (dashed red lines), and distributions predicted by experiment-specific networks (solid red lines) for each experiment.

3.4. Classification of diffusion type

When it came to the classification of DT (among immobilized, confined, freely diffusing and directed), U-AnD-ME achieved a superior F_1-score (with an average score of $F_1 = 0.98$) compared to all other participating methods in the 2024 AnDi Challenge [27]. Figure 9(a) shows that U-AnD-ME was indeed highly accurate in classifying all DTs, being the lowest F_1 score 0.85 for directed diffusion. Looking at the classification by model (figure 9(b)) shows how this relatively poorer performance for directed trajectories is primarily due to the QT model. As noted earlier (section 3.3), this model was hampered by Experiment 9 with an extremely superdiffusive state that was harder to resolve for U-AnD-ME (see also figure 6(e)). Finally, classification as a function of segment length is unsurprising, with relatively poorer predictions for shorter segments due to their lower information content (figure 9(c)). Segments of length 3 achieved $F_1 = 0.92$, with values plateauing to $F_1 \approx 1$ for segments longer than 100 time steps.

3.5. Ensemble distributions

Finally, U-AnD-ME outperformed all other participating methods in the 2024 AnDi Challenge in terms of capturing the ensemble distribution of α. W_α values in figure 6(f) show an average of 0.16 and 0.11 for generalist and experiment-specific networks, respectively. Our method was not as successful in terms of predicting the distribution of K, placing 4$^{\text{th}}$; this was the only metric across the Trajectory Track of the

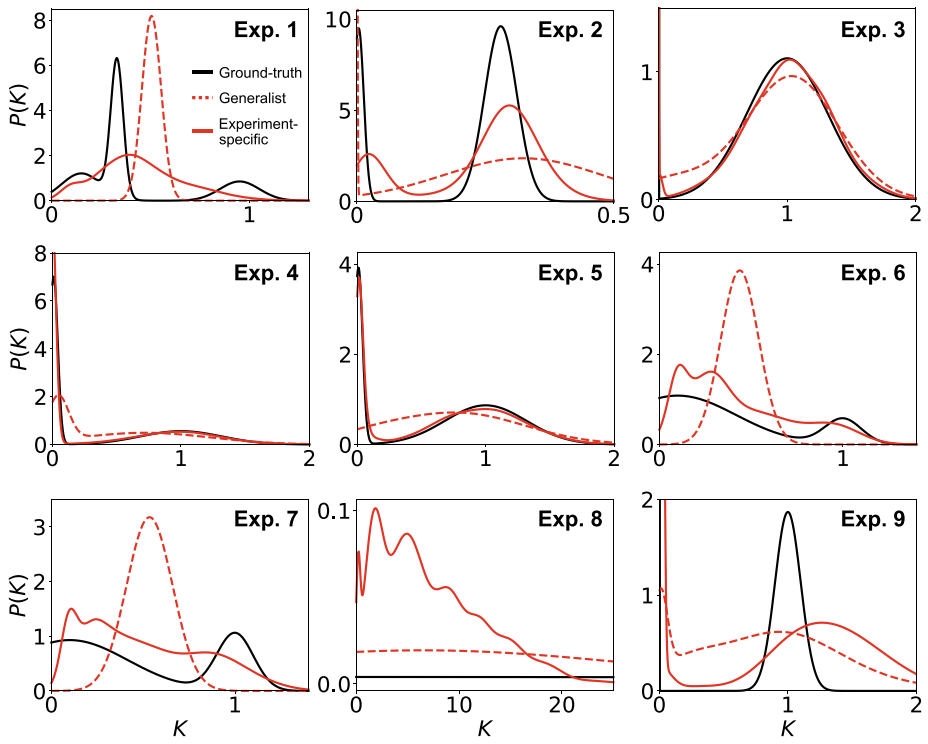

Figure 11. Predicted K-distributions from ensemble analysis. Ground-truth K-distributions $P(K)$ (black lines), distributions predicted by generalist networks (dashed red lines), and distributions predicted by experiment-specific networks (solid red lines) for each experiment.

Challenge for which U-AnD-ME did not outperform all other participating methods. We primarily attribute this decreased performance to the negative-control Experiment 8 (section 3.3), which shows the worst W_K in figure 6(g) with scores of 2.83 and 3.29 against averages of 0.49 and 0.40 for generalist and experiment-specific networks, respectively.

From the reconstructed ensemble probabilities of α and K in figures 10 and 11, we can see how experiment-specific networks tend to perform better than generalist ones as already observed in figures 6(f), (g) and (j). For both properties, generalist networks tend to find unimodal distributions near the average parameter value, while experiment-specific networks capture richer distributions better reflecting those of the ground truth. When two peaks in the ground-truth distributions are close together (e.g. in the α-distributions for Experiments 2, 6 and 7, and in the K-distribution for Experiment 1), experiment-specific networks tend to approximate neighbouring ground-truth modes with a single broad coarse-grained peak, merging the information from the individual modes. Unimodal distributions (e.g. in the α-distributions for Experiments 1 and 3, and in the K-distribution for Experiments 3) and multimodal distributions formed by well-separated modes (e.g. in the distributions of α for Experiments 4, 5 and 9, and of K for Experiments 2, 4, 5, 6, 7 and 8) tend instead to be captured better by U-AnD-ME.

3.6. Application to experimental data

To verify that U-AnD-ME generalises beyond synthetic trajectories, we applied it to extract the diffusion properties of experimental single-particle-tracking data from Granik *et al* [15] as an example (figure 12). These data are available online at [54] and capture the diffusion of fluorescent beads in entangled F-actin network gels with various mesh sizes. These trajectories demonstrate fractional Brownian motion [15], and hence can be directly compared with the dataset used for the Challenge [27]. As no ground truth is available, our estimates are compared to fits of the ensemble average of the time-averaged MSDs (eTAMSD) from the trajectories at different gel mesh sizes. In their work, Granik *et al* [15] introduce a deep-learning framework based on convolutional neural networks (CNNs) that well recovers the dependence of the anomalous diffusion exponent α on the gel mesh size against the MSD fits (mean square error $\epsilon = 7.90 \times 10^{-4}$, figure 12(a)). Using the generalist network trained on simulated data, U-AnD-ME produces consistent predictions of the exponent α ($\epsilon = 3.35 \times 10^{-4}$, figure 12(a)) and, differently from [15], is also able to extract reliable predictions of the generalized diffusion coefficient K ($\epsilon = 0.142$, figure 12(b)) based on the analysis of single trajectories. Ensemble analysis offers comparable results for predictions of both α

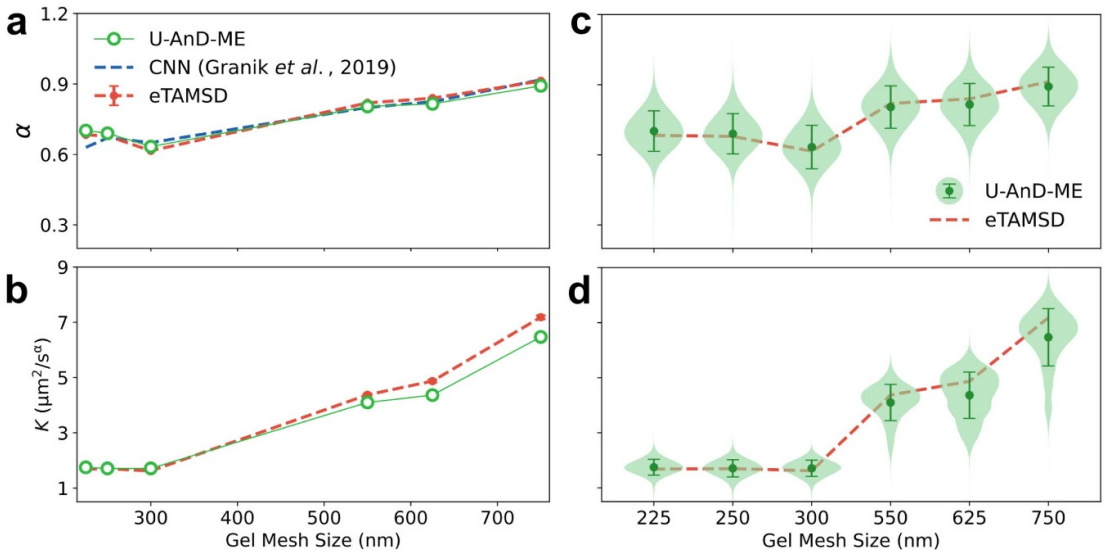

Figure 12. Inference of anomalous diffusion properties in experimental data. Predicted anomalous diffusion properties from experimental trajectories of fluorescent beads diffusing in entangled F-actin network gels with various mesh sizes [15] based on the data available online at [54]. Both U-AnD-ME (a) and (b) single-trajectory and (c) and (d) ensemble predictions are shown for (a, c) the anomalous diffusion exponent α and (b, d) the generalized diffusion coefficient K. In addition to the pre-processing steps detailed in section 2.3.2, input trajectories were also processed to have differences of zero mean in x and y. U-AnD-ME predictions are compared against fits of the ensemble average of the time-averaged mean squared displacements (eTAMSD) calculated from trajectories for each gel mesh size. Bootstrapped standard deviations are also shown for eTAMSD (1000 iterations with 1000 samples, values on the order of line width). In a, U-AnD-ME predictions are also compared against those from the CNN-based approach from [15] (code available online at [53]). In c and d, U-AnD-ME ensemble predictions are shown as violin plots to represent the full spread of the distributions of α and K for each gel mesh size. Dots and error bars respectively show the mean and standard deviation of the distributions.

($\epsilon = 3.35 \times 10^{-4}$, figure 12(c)) and K ($\epsilon = 0.142$, figure 12(d)). Although the true diffusion properties are unknown, U-AnD-ME's reliable performance on these experimental data strengthens confidence in our model and corroborates the results on synthetic trajectories presented earlier.

4. Discussion

The success of U-AnD-ME in the 2024 AnDi Challenge demonstrates its significance and potential for the analysis of data from live-cell single-molecule imaging at both single-trajectory and ensemble levels. Our method came 1st for the Single-Trajectory Task of the Trajectory Track of the Challenge, which was the most subscribed task of the competition. For this highly competitive task, U-AnD-ME was awarded 1st place for every possible subtask, accurately predicting the locations of CPs, the anomalous-diffusion exponent α, the generalized diffusion coefficient K, and the DT. Our method was also awarded 1st place for the Ensemble Task of the Trajectory Track, coming 1st in the subtask dedicated to predicting the distribution of α and 4th in the subtask for predicting the distribution of K. The suboptimal performance for this subtask is due to our framework struggling in cases where the training distribution for the generalist network differs significantly from the experimental distribution. In the Challenge, the generalised diffusion coefficient K could take values in the range $K \in [10^{-12}, 10^{6}]$. As most experiments but the negative-control Experiment 8 generally take smaller values, the training procedure we used for generalist networks focussed on a smaller range ($K \in [10^{-12}, 15]$), as training a network over the full range would be computationally expensive and result in poor performance for the range of interest. As a consequence, experiments with significant density for $K > 15$ may have poor generalist network predictions for the ensemble K-distribution, and thereby poor experiment-specific network predictions for this distribution. No such issue affects α as these values are confined to a much smaller interval ($\alpha \in [0, 2)$), making training over all possible values straightforward. In practice, prior knowledge of the system being analysed could eliminate this issue by adapting the generalist network's training range to better suit the problem at hand.

The inference of segment diffusion properties tends to improve with the quality of the CP detections [4]. Currently, U-AnD-ME uses a standard binary cross-entropy loss for CPs. However, in typical trajectories, fewer time steps are CPs than not, leading to significant imbalances between these two possible classes. *Focal loss* is a modified cross-entropy designed to perform better with class imbalance [55]. Its use for training instead of binary cross-entropy could improve the performance of CP detection and thereby improve the

inference of segment properties too. Moreover, as shown here, experiment-specific training significantly improves the accuracy of all network predictions. Using experiment-specific architectures informed by the physics of each underlying model could further improve our approach. Ensemble predictions could also benefit from increased *a priori* knowledge about experimental data: knowing an experiment's model could inform us about the number of distinct states it has and this could be used to directly enforce the number of GMM components. Finally, while the original U-Net 3+ architecture was designed for image analysis [43], U-AnD-ME extended it to time series. Amongst other changes, this involved using 1D convolutions as opposed to 2D convolutions. Presently, U-AnD-ME analyses trajectories extracted from microscopy videos. Future developments could explore the use of 3D convolutions to enable U-AnD-ME to directly extract information from these videos.

Our ablation study shows that variations to the basic U-AnD-ME architecture can be introduced to enhance specific performance metrics (e.g. inference of diffusion properties versus CP detection) while keeping computational resources approximately constant. Alternatively, architectural modifications can improve overall performance, albeit at the cost of increased computational demand. While these variants can be considered based on specific experimental needs, we find that the U-AnD-ME basic architecture offers a good balance between overall performance and computational efficiency, particularly in tasks requiring the detection of CPs as in the 2024 AnDi Challenge.

In conclusion, built using convolutional operations, our machine learning framework ensures high efficiency and stable training while delivering results that allowed U-AnD-ME to be among the top-performing teams in the 2024 Andi Challenge. Even against comparable frameworks for the analysis of anomalous diffusion data proposed during the 2020 AnDi Challenge [4], U-AnD-ME offers several advantages as it does not require complex feature engineering [19], uses a single network for a wide range of segment lengths [20, 21], and extracts all diffusion parameters with a single architecture [20]. The relative simplicity of our framework is particularly noteworthy given the increased complexity of the 2024 AnDi Challenge tasks compared to those of the previous version. While we applied it to the analysis of both synthetic and experimental 2D trajectories, our architecture could be easily adapted to handle even higher dimensional trajectories. Besides its proven effectiveness in particle-tracking analysis, U-AnD-ME's central architecture could also be relevant for any other task requiring analysis of time series exhibiting complex diffusion behaviour, such as animal migration records [56], search strategy development in microrobotics [57] or financial market data [58]. Additionally, it holds potential for any problem that requires the segmentation of time series. For example, U-AnD-ME framework could be applied to ECG analysis [59], fault detection in manufacturing [60], or detecting ecosystem regime shifts [61].

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgment

S A and G V are grateful for the studentship funded by the A*STAR-UCL Research Attachment Programme through the EPSRC M3S CDT (EP/L015862/1). R N acknowledges support from the Academic Research Fund from the Singapore Ministry of Education (RG59/21 and MOE2019-T2-2-010) and the National Research Foundation, Singapore, under its 29th Competitive Research Program (CRP) Call (Grant No. NRF-CRP29-2022-0002). G V also acknowledges support for this work by The Chan Zuckerberg Initiative 'Multi-color single molecule tracking with lifetime imaging' (2023-321188).

ORCID iDs

Ran Ni 0000-0001-9478-0674
Giorgio Volpe 0000-0001-9993-5348

References

[1] Oliveira F A, Ferreira R M S, Lapas L C and Vainstein M H 2019 Anomalous diffusion: a basic mechanism for the evolution of inhomogeneous systems *Front. Phys.* **7** 18
[2] Einstein A 1905 On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat *Ann. Phys., Lpz.* **322** 549–60
[3] Klafter J and Sokolov I M 2005 Anomalous diffusion spreads its wings *Phys. World* **18** 29
[4] Muñoz-Gil G *et al* 2021 Objective comparison of methods to decode anomalous diffusion *Nat. Commun.* **12** 6253
[5] Metzler R, Jeon J-H, Cherstvy A G and Barkai E 2014 Anomalous diffusion models and their properties: non-stationarity, non-ergodicity and ageing at the centenary of single particle tracking *Phys. Chem. Chem. Phys.* **16** 24128–64

[6] Tolić-Nørrelykke I M, Munteanu E-L, Thon G, Oddershede L and Berg-Sørensen K 2004 Anomalous diffusion in living yeast cells *Phys. Rev. Lett.* **93** 078102
[7] Wachsmuth M, Waldeck W and Langowski J 2000 Anomalous diffusion of fluorescent probes inside living cell nuclei investigated by spatially-resolved fluorescence correlation spectroscopy *J. Mol. Biol.* **298** 677–89
[8] Kusumi A, Sako Y and Yamamoto M 1993 Confined lateral diffusion of membrane receptors as studied by single particle tracking (nanovid microscopy). Effects of calcium-induced differentiation in cultured epithelial cells *Biophys. J.* **65** 2021–40
[9] Höfling F and Franosch T 2013 Anomalous transport in the crowded world of biological cells *Rep. Prog. Phys.* **76** 046602
[10] Chen K, Wang B and Granick S 2015 Memoryless self-reinforcing directionality in endosomal active transport within living cells *Nat. Mater.* **14** 589–93
[11] Volpe G and Volpe G 2017 The topography of the environment alters the optimal search strategy for active particles *Proc. Natl Acad. Sci.* **114** 11350–5
[12] Lomholt M A, Ambjörnsson T and Metzler R 2005 Optimal target search on a fast-folding polymer chain with volume exchange *Phys. Rev. Lett.* **95** 260603
[13] Krapf D, Lukat N, Marinari E, Metzler R, Oshanin G, Selhuber-Unkel C, Squarcini A, Stadler L, Weiss M and Xu X 2019 Spectral content of a single non-brownian trajectory *Phys. Rev. X* **9** 011019
[14] Sposini V *et al* 2022 Towards a robust criterion of anomalous diffusion *Commun. Phys.* **5** 305
[15] Granik N, Weiss L E, Nehme E, Levin M, Chein M, Perlson E, Roichman Y and Shechtman Y 2019 Single-particle diffusion characterization by deep learning *Biophys. J.* **117** 185–92
[16] Kowalek P, Loch-Olszewska H and Szwabiński J 2019 Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach *Phys. Rev. E* **100** 032410
[17] Bo S, Schmidt F, Eichhorn R and Volpe G 2019 Measurement of anomalous diffusion using recurrent neural networks *Phys. Rev. E* **100** 010102
[18] Muñoz-Gil G, Garcia-March M A, Manzo C, Martín-Guerrero J D and Lewenstein M 2020 Single trajectory characterization via machine learning *New J. Phys.* **22** 013010
[19] Gentili A and Volpe G 2021 Characterization of anomalous diffusion classical statistics powered by deep learning (CONDOR) *J. Phys. A: Math. Theor.* **54** 314003
[20] Argun A, Volpe G and Bo S 2021 Classification, inference and segmentation of anomalous diffusion with recurrent neural networks *J. Phys. A: Math. Theor.* **54** 294003
[21] Li D, Yao Q and Huang Z 2021 Wavenet-based deep neural networks for the characterization of anomalous diffusion (WADNet) *J. Phys. A: Math. Theor.* **54** 404003
[22] Seckler H and Metzler R 2022 Bayesian deep learning for error estimation in the analysis of anomalous diffusion *Nat. Commun.* **13** 6717
[23] Pineda J, Midtvedt B, Bachimanchi H, Noé S, Midtvedt D, Volpe G and Manzo C 2023 Geometric deep learning reveals the spatiotemporal features of microscopic motion *Nat. Mach. Intell.* **5** 71–82
[24] Seckler H, Szwabinski J and Metzler R 2023 Machine-learning solutions for the analysis of single-particle diffusion trajectories *J. Phys. Chem. Lett.* **14** 7910–23
[25] Manzo C and Garcia-Parajo M F 2015 A review of progress in single particle tracking: from methods to biophysical insights *Rep. Prog. Phys.* **78** 124601
[26] Shen H, Tauzin L J, Baiyasi R, Wang W, Moringo N, Shuang B and Landes C F 2017 Single particle tracking: from theory to biophysical applications *Chem. Rev.* **117** 7331–76
[27] Muñoz-Gil G *et al* 2025 Quantitative evaluation of methods to analyze motion changes in single-particle experiments *Nat. Commun.* **16** 6749
[28] Yanagawa M, Hiroshima M, Togashi Y, Abe M, Yamashita T, Shichida Y, Murata M, Ueda M and Sako Y 2018 Single-molecule diffusion-based estimation of ligand effects on G protein—coupled receptors *Sci. Signal.* **11** eaao1917
[29] Tabor A *et al* 2016 Visualization and ligand-induced modulation of dopamine receptor dimerization at the single molecule level *Sci. Rep.* **6** 33233
[30] Low-Nam S T, Lidke K A, Cutler P J, Roovers R C, van Bergen en Henegouwen M P, Wilson B S and Lidke D S 2011 ErbB1 dimerization is promoted by domain co-confinement and stabilized by ligand binding *Nat. Struct. Mol. Biol.* **18** 1244–9
[31] Spillane K M, Ortega-Arroyo J, de Wit G, Eggeling C, Ewers H, Wallace M I and Kukura P 2014 High-speed single-particle tracking of GM1 in model membranes reveals anomalous diffusion due to interleaflet coupling and molecular pinning *Nano Lett.* **14** 5390–7
[32] Ritchie K, Iino R, Fujiwara T, Murase K and Kusumi A 2003 The fence and picket structure of the plasma membrane of live cells as revealed by single molecule techniques (review) *Mol. Membr. Biol.* **20** 13–18
[33] Mandelbrot B B and Van Ness J W 1968 Fractional Brownian motions, fractional noises and applications *SIAM Rev.* **10** 422–37
[34] Muñoz-Gil G, Requena B, Fernández G F, Bachimanchi H, Pineda J and Manzo C 2023 AnDiChallenge/andi_datasets: AnDi Challenge 2 (https://doi.org/10.5281/zenodo.10259556)
[35] Eggeling C *et al* 2008 Direct observation of the nanoscale dynamics of membrane lipids in a living cell *Nature* **457** 1159–62
[36] da Rocha-Azevedo B, Lee S, Dasgupta A, Vega A R, de Oliveira L R, Kim T, Kittisopikul M, Malik Z A and Jaqaman K 2020 Heterogeneity in VEGF receptor-2 mobility and organization on the endothelial cell surface leads to diverse models of activation by VEGF *Cell Rep.* **32** 108187
[37] Achimovich A M, Yan T and Gahlmann A 2023 Dimerization of iLID optogenetic proteins observed using 3D single-molecule tracking in live *E. coli Biophys. J.* **122** 3254–67
[38] Valley C C, Arndt-Jovin D J, Karedla N, Steinkamp M P, Chizhik A I, Hlavacek W S, Wilson B S, Lidke K A and Lidke D S 2015 Enhanced dimerization drives ligand-independent activity of mutant epidermal growth factor receptor in lung cancer *Mol. Biol. Cell* **26** 4087–99
[39] Weigel A V, Tamkun M M and Krapf D 2013 Quantifying the dynamic interactions between a clathrin-coated pit and cargo molecules *Proc. Natl Acad. Sci.* **110** E4591–600
[40] Rossier O *et al* 2012 Integrins β1 and β3 exhibit distinct dynamic nanoscale organizations inside focal adhesions *Nat. Cell Biol.* **14** 1057–67
[41] Sungkaworn T, Jobin M-L, Burnecki K, Weron A, Lohse M J and Calebiro D 2017 Single-molecule imaging reveals receptor—G protein interactions at cell surface hot spots *Nature* **550** 543–7
[42] Crouse D F 2016 On implementing 2D rectangular assignment algorithms *IEEE Trans. Aerosp. Electron. Syst.* **52** 1679–96
[43] Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen Y-W and Wu J 2020 UNet 3+: a full-scale connected UNet for medical image segmentation (arXiv:2004.08790)

[44] Requena B, Masó-Orriols S, Bertran J, Lewenstein M, Manzo C and Muñoz-Gil G 2023 Inferring pointwise diffusion properties of single trajectories with deep learning *Biophys. J.* **122** 4360–9
[45] Asghar S 2024 SolomonAsghar/U-AnD-ME (available at: https://github.com/SolomonAsghar/U-AnD-ME)
[46] Perslev M, Jensen M H, Darkner S, Jennum P J and Igel C 2019 *U-Time: a Fully Convolutional Network for Time Series Segmentation Applied to Sleep Staging* (Red Hook, NY, USA: Curran Associates Inc)
[47] Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum P J and Igel C 2021 U-Sleep: resilient high-frequency sleep staging *npj Digit. Med.* **4** 72
[48] Qu X, Hu Y, Cai W, Xu Y, Ke H, Zhu G and Huang Z 2024 Semantic segmentation of anomalous diffusion using deep convolutional networks *Phys. Rev. Res.* **6** 013054
[49] Zhou Z, Siddiquee M M R, Tajbakhsh N and Liang J 2018 *Unet++: A Nested U-Net Architecture for Medical Image Segmentation* (Springer) pp 3–11
[50] Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the em algorithm *J. R. Stat. Soc. B* **39** 1–22
[51] Hand D J, McLachlan G J and Basford K E 1989 Mixture models: inference and applications to clustering *Appl. Stat.* **38** 384
[52] ANDI Challenge Team 2024 Andi challenge 2024—leaderboard (available at: http://andi-challenge.org/challenge-2024/#andi2leaderboard)
[53] Nano-bio-optics lab 2019 AnomdiffDB/DB (available at: https://github.com/AnomDiffDB/DB)
[54] Nano-bio-optics lab 2019 Software (available at: https://nanobiooptics.net.technion.ac.il/software/)
[55] Lin T-Y, Goyal P, Girshick R, He K and Dollár P 2017 Focal loss for dense object detection *Proc. IEEE Int. Conf. on Computer Vision* pp 2980–8
[56] Revell C and Somveille M 2017 A physics-inspired mechanistic model of migratory movement patterns in birds *Sci. Rep.* **7** 9870
[57] Gentili A, Klages R and Volpe G 2024 Anomalous diffusion of superparamagnetic walkers with tailored statistics (arXiv:2412.13960)
[58] Plerou V, Gopikrishnan P, Amaral L A N, Gabaix X and Stanley H E 2000 Economic fluctuations and anomalous diffusion *Phys. Rev. E* **62** R3023–6
[59] Vocaturo E and Zumpano E 2021 ECG analysis via machine learning techniques: news and perspectives *2021 IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM)* vol 24 (IEEE) pp 3106–12
[60] Netzer M, Palenga Y and Fleischer J 2022 Machine tool process monitoring by segmented timeseries anomaly detection using subprocess-specific thresholds *Prod. Eng.* **16** 597–606
[61] Ahmad W, Shadaydeh M and Denzler J 2024 Regime identification for improving causal analysis in non-stationary timeseries (arXiv:2405.02315)