Sorting of Single–Molecule Trajectories by means of Machine Learning- a status update on the annotation procedure

Lisa Krenkel[1], Tobias Schlosser[2], Danny Kowerko[2], Richard Börner[1]

[1] Laserinstitut Hochschule Mittweida, University of Applied Sciences Mittweida, 09648 Mittweida, Germany, [2] Department of Informatics, Technical University Chemnitz, 09111 Chemnitz, Germany

We use machine learning for the selection and classification of single–molecule trajectories to replace commonly used user–dependent sorting algorithms. Measured fluorescence time series of labelled single molecules need to be sorted into 'good molecules' and 'bad' molecules before further kinetic and thermodynamic analysis. Currently, processing, sorting and analysis of the data is mainly done with the help of laboratory specific programs. Although there are freely available programs for processing smFRET data, they do not offer 'molecular sorting' or it is purely empirical. Only recently, new approaches came up to solve this problem by means of machine learning. Here, we describe a sound terminology for molecular sorting of smFRET data and present an efficient workflow for manual annotation followed by the training of the ML algorithm. Descriptive statistics of our generated dataset are provided and will serve as the basis for supervised ML-based molecular sorting algorithms yet to be developed.

Introduction

In recent years single–molecule fluorescence resonance energy transfer (smFRET) has been established as a mature and adaptable method regarding the study of biomolecular structures as well as their structural dynamics[1]. Förster resonance energy transfer describes the weak dipol–dipol coupling of a donor and an acceptor fluorophore with overlapping emission/excitation spectra. The efficiency of this energy transfer is strongly distance dependent and used to calculate the interdye distance of a specific donor–acceptor–pair. The experimental measure FRET is calculated according to:

$$FRET = \frac{F_A}{F_A + F_D} \quad (1)$$

where F_A refers to the fluorescence emission intensity of the acceptor and F_D to the emission of the donor fluorophore.

In total internal reflection fluorescence (TIRF) microscopy experiments, the (bio–)molecules of interest are labelled with fluorophores and immobilised on a surface before their fluorescence signal is recorded via EMCCD or sCMOS camera resulting in a single–molecule video (SVM) [6]. The analysis of SMV's usually requires two steps: starting with video processing, including the separation of the field of view (FOV) into donor and acceptor channel and the detection of single–molecule spots; consisting of the calculation of biomolecule coordinates and their fluorescence intensities [2]. The subsequent second step is the trace processing including fluorescence intensity correction and reliable state detection [7]. However, the fluorescence trajectories which are extracted from the SMV contain not only useful fluorescence intensity fluctuations e.g., from structural changes in the biomolecule, but photophysical artefacts, such as photobleaching, blinking etc. Because of this, the intensity trajectories are sorted into 'good' and 'bad' molecules, i.e., fluorescence intensity trajectories fulfilling a number of criteria, before further analysis. This molecular sorting (or classification) is currently done mostly by hand or through semi–automated algorithms based on threshold criteria still requiring profound user interference. Admittedly, this is not only problematic in terms of the required expertise and time, but also inevitably introduces user bias. Most of the freely available tools for sorting of fluorescence trajectories rely on purely heuristically approaches like intensity thresholds etc. which requires a rather advanced understanding of FRET [8]. New tools have emerged implementing machine learning approaches to classify smFRET trajectories such as FRETboard or deepFRET while minimising user bias and necessary time [4, 9].

At its simplest, machine learning describes the design of algorithms with the ability to optimize their performance based on example data or prior experience [10]. The chosen approach in case of the sorting algorithm is based on supervised learning, meaning that the algorithm is trained with annotated training data to later sort similar, un–annotated data [11]. Since the annotation is done per hand, based on user experience and/or prior knowledge, the intensity–time traces of a biomolecule are generally depicted in form of graphs containing both the donor and acceptor intensities as well as the FRET

[1] Also known as Förster resonance energy transfer. The effect was discovered and described first theoretically by Theodor Förster in 1948[5].

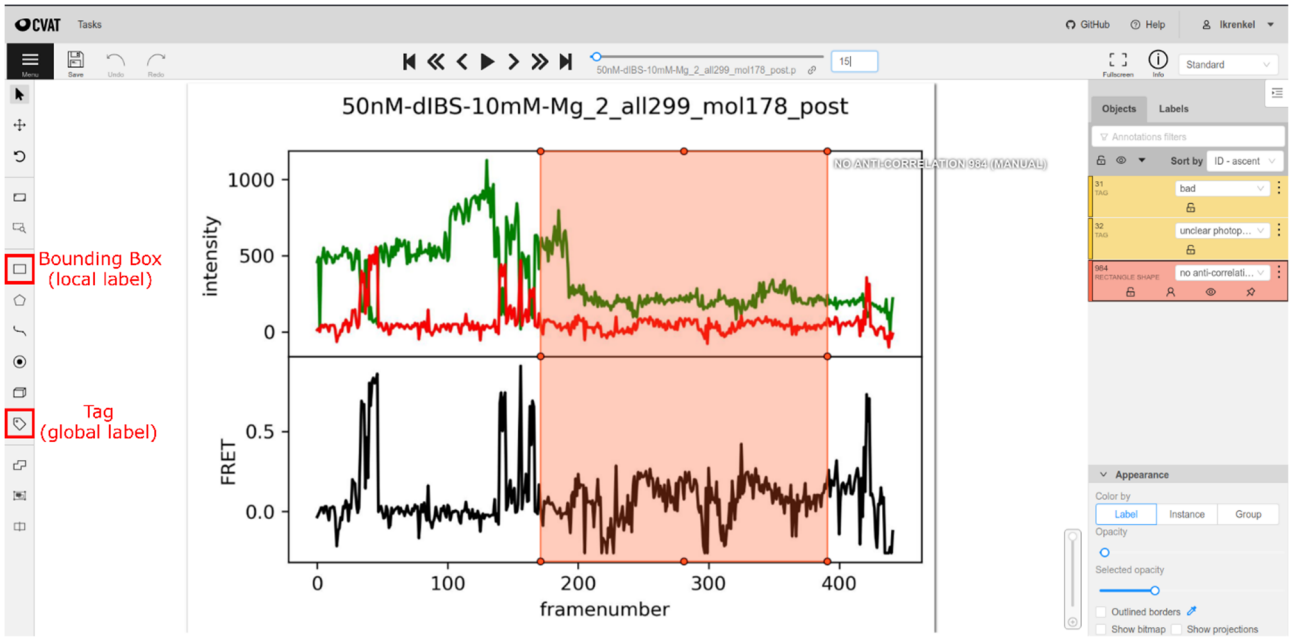

Fig. 1: CVAT Annotation tool: Both, the global and local labels, were assigned via their respective tabs in the CVAT annotation tool as marked on the (left). On the (right) all of the assigned labels are displayed. The annotation progress can be viewed at the top.

trajectory (Fig. 1). Thus, the annotated data is used as training data set.

The field of automated and semi–automated visual inspection utilises classification approaches which can be distinguished according to their functionality: filter–based, projection–based, and hybrid approaches. These are usually accompanied by clustering approaches, mostly as additional classification steps, but also to generate interpretable rule sets for the classification [17]. The problem with these approaches lies within their limited ability to classify novel signatures and patterns as well as the likely need for manual adaptation and parameterisation through the user. Because of this more recent machine learning approaches to the problem of molecular sorting have included support vector and multilayer perceptron–based classifiers. Following the contributions of *de Lannoy et al.* [3], *Thomsen et al.* [4], and *Li et al.* [9] even the latest deep learning motivated research utilized classical single deep and convolutional neural networks. These approaches are assumed to result in improved training and test rates, which results in the definition of an optimum which allows for both, a timely classification as well as a solution independent classification of regions of interest.

Following the beforehand highlighted and emphasised principles of automated and semi-automated visual inspection, this contribution proposes an evaluation within a range different approaches in comparison to contrast and accentuate their respective advantages and differences. For this purpose, said approaches are separated into the following three categories: baseline ML models, DL-based models, as well as more novel, temporal information and region-based convolutional (R-CNN) and deep neural networks in general. Furthermore, this also includes commonly deployed DL-based principles such as pre-trained and optimised building blocks for DNNs (e.g., DenseNets) [18], building blocks with different characteristics, i.e. kernel sizes and connectivities (inception modules) [19], approaches with residual learning (ResNets) [20], as well as combinations of these approaches, i.e. Inception-ResNets [21].

In the last decade, numerous annotation tools for images and videos have emerged in the computer vision community. We use CVAT[2] (Computer Vision Annotation Tool), which supports many easily machine readable annotation formats, developed or originally used in computer vision challenges [12]. As such CVAT fulfils the annotation requirements for this project which are mainly bounding box and global labels, the rich text–based export format support (json, csv), collaborative labelling via web (application) and preserving data privacy via hosting it as own web service/web application. Considering the chosen approach of supervised learning, the first step in developing a neural network for molecular sorting is the annotation of fluorescence trajectories to build a data set for training, validation, and testing.

Herein, we give an overview of the different label categories used, the percentage of the labels within the test data set, as well as a short outlook how this work will continue.

Methods

The video processing and trace generation to obtain fluorescence intensity and thus FRET trajectories for a

[2] https://github.com/openvinotoolkit/cvat

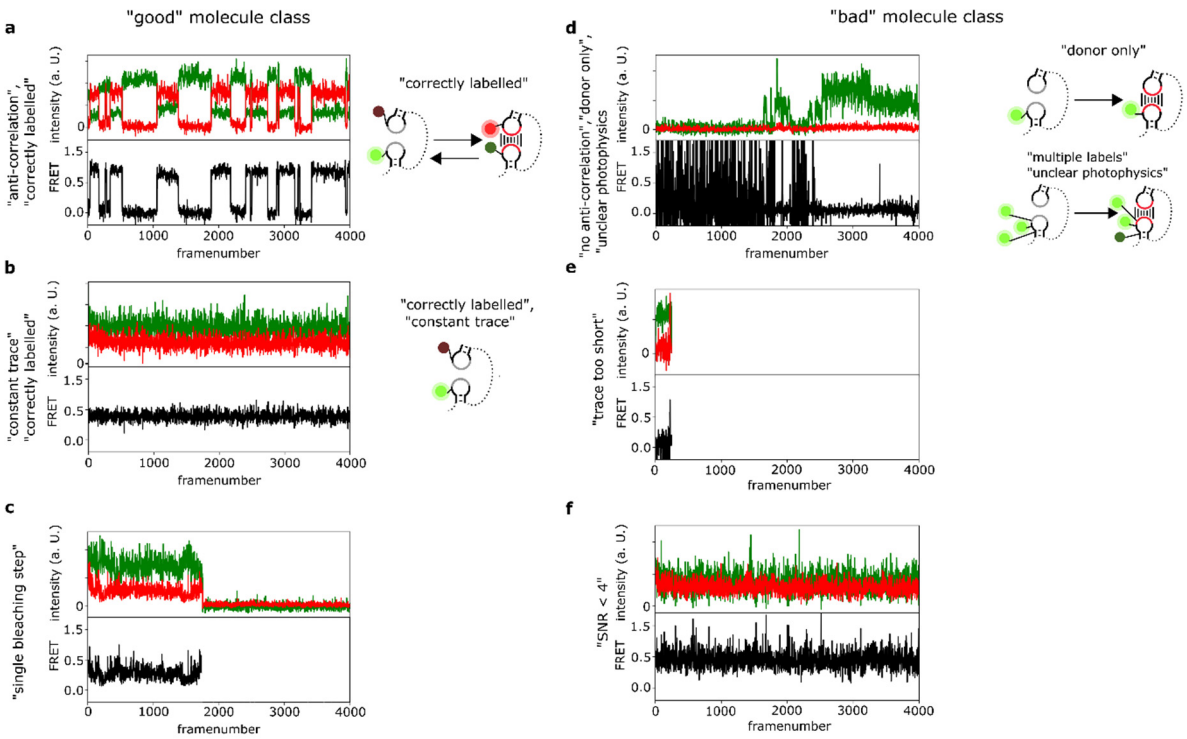

Fig. 2: Label classifiers: (left column) Good molecules (right column) Bad molecules. (a) Correctly labelled, anti-correlated signal of single-molecule fluorescence intensity trajectory. (b) Correctly labelled, static single-molecule intensity trajectory. (c) Correctly labelled, anti-correlated signal of a single-molecule intensity trajectory - both donor and acceptor signal display single bleaching steps. (d) Incorrectly labelled (multiple donor labels) signal for a single-molecule intensity signal with no anti-correlation but unclear photo physics. (e) Too short trace of a single-molecule intensity trajectory. (f) Single-molecule intensity trajectory with a too low signal to noise ratio, resulting in a very 'noisy' signal.

given experimental or simulated SMV is described elsewhere [13]. An annotation for the training dataset was achieved by converting the fluorescence intensity and FRET traces into graphs, which were labelled using the online image/video annotation tool CVAT (Fig. 1). An important distinction was made between local and global labels. Global labels are defined by tags, e.g. bad or good, and were selected based on the combination of assigned local labels (compare Tab. 1). The local labels in turn are assigned using bounding boxes to mark their occurrence in the trace, i.e. by assigning a certain range of frame numbers, as such all traces require a minimum number of two labels – one global and one local. Examples for "good" and "bad" molecules are given in Fig. 2.

The declaration as a "good" or "bad" molecule was made after the local labels were assigned. An important factor for the declaration of a "good" or "bad" molecule in smFRET experiments is the anti–correlation of donor and acceptor signals, inherently linked to the FRET process, where a high donor signal corresponds to a low acceptor signal and *vice versa*. Thus, "anti–correlated" signal traces are labelled accordingly as "good" molecules (Fig. 2a). If the intensity trace of a given molecule does not display anticorrelation, but a constant signal well above the background, the trace is labelled as constant or static (Fig. 2b). This is the case if no structural changes take place during the measurement time. In some instances, a fluorescence signal can be observed to decrease instantaneously to a minimal level, i.e. the background signal. This is referred to as single–bleaching step, an event in which a fluorescent dye is (photo)chemically destroyed. A single bleaching step is usually not considered to be an indicator for a bad molecule, rather these traces can be used to calculate correction factors (Fig. 2c) [14]. In contrast, multiple bleaching steps confirm multiple dyes in the spot (Fig. 2d) and are labelled as "bad" molecules. Usually, it is not possible to determine whether these multiple bleaching steps are a result of the labelling method itself, the immobilization procedure, or an experimental artefact. As such, they were annotated as "unclear photophysics" (Fig. 2d). If the fluorescence signal of a donor–dye remains at the background signal level for the whole trace, the molecule was labelled as "donor only" (Fig. 2d) and consequently as "bad" molecule. In contrast, a signal for both, donor and acceptor well above the background signal, is an indicator for a correctly labelled molecule, and designated as "good" (Fig. 2a, b, and c). Further, smFRET traces vary in their length due to the statistical nature of photobleaching or potential molecular desorption. Heuristically, traces are categorized as "trace too short" to give any significant information (< 10 frames, Fig. 2e). Lastly, some smFRET traces contain important information regarding the different states of the biomolecule but are obscured by experimental noise. Here, observed fluctuations have near to no correlation to actual state changes of the molecules – the noise simply masks the fluctuations due to state changes. Because of this, traces with a low signal

to noise ratio SNR < 4 can generally considered as "bad" molecules (Fig 2f) [7].

Results

The training data set, at the current stage, contains 2409 molecules, of which 74.89 % are designated as "good" and 26.24 % as "bad". The percentages of all defined labels are listed in Tab. 1. It is important to note that some labels like "anti–correlation" were used multiple times on a single trace but count only once. Further, traces which exhibit multiple features were labelled multiple times (compare Tab. 1, right column).

Tab. 1: Label distributions across trainings data set

label	fraction of molecules	labels per trace	fraction of molecules
good	74.89 %	1	0.00 %
bad	26.24 %	2	11.50 %
single bleaching step	2.95 %	3	84,85 %
SNR < 4	13.08 %	4	3.28 %
donor only	7.85 %	5	0.30 %
trace too short	0.00 %		
correctly labelled	71.19 %		
no anti–correlation	8.09 %		
anti–correlation	78.41 %		
unclear photophysics	12.41 %		

Discussion and Outlook

A known issue of training data sets for ML approaches is the existence of class imbalance. Here, we observe a higher number of “good” in comparison to “bad” molecules. If the distribution of classes, i.e. labels, is highly imbalanced, classification learning algorithms usually display low predictive accuracy for the infrequent classes [15]. The effect on the classification performance of the resulting neural network is generally detrimental and solved e.g., by under/oversampling which might lead to overfitting of the algorithm [16]. As raw smFRET date sets are rare and if published, do not contain “bad” molecules, raw SMV data needs to be processed to minimise the class imbalance in the training data set. Therefore, we will annotate more smFRET traces especially for the classification of bad molecules.

MASH–FRET as a tool for the processing of smFRET data also includes the option to simulate traces and even the option to simulate whole SMVs – including “bad” molecules [2]. The annotation of simulated data will help to overcome class imbalance in our training data set. Further, simulated data will be of particular interest during the evaluation process of our algorithm - since all relevant parameters are defined in this simulated data sets they are considered as ground-truth [7].

Only recently, deepFRET [4] and AutoSiM [9] introduced AI–assisted automated sorting/classification of smFRET trajectories. But despite the very positive results of both studies, they show a lack of cross–sample and cross–laboratory variability. As such not only data sets have to be continuously updated but algorithms need to be trained with community driven annotated data sets. Therefore, we will pay special attention to use single–molecule data sets of different labs to introduce lab specific variances in the training data set.

Acknowledgements

The authors thank the organizers of the *Interdisziplinäre Wissenschaftliche Konferenz an der Hochschule Mittweida* (IWKM) for giving them the opportunity to present their research. We thank Susann Zelger–Paulus for sharing SMV. We further thank Victoria Birkedal and Thomas Villmann for insightful discussions; we are looking forward to further collaborate with them.

References

[1] Juette, M. et al.: Nature Methods, 13 (2016), 341–344
[2] Börner, R. et al.: PLOS ONE, 13 (2018), e0195277
[3] de Lannoy, C. et al.: bioRxiv, 15 (2020), 2020.08.28.272195
[4] Thomsen, J. et al.: eLife, 9 (2020), e60404
[5] Lerner, E. et al.: eLife, just accepted (2021)
[6] Hellenkamp, B. et al: Nature Methods, 15 (2018), 669–676
[7] Hadzic, M. C. A. S. et al.: The Journal of Physical Chemistry B, 122 (2018), 6134–6147
[8] White, D. S. et al.: eLife, 9 (2020), e53357
[9] Li, J. et al: Nature Communications, 11 (2020), 5833
[10] Alpaydin, E.: MIT Press, 1 (2010)
[11] Bonaccorso, B.: Packt Publishing, 1 (2017)
[12] Liu, L. et al.: International Journal of Computer Vision, 128 (2020), 261–318
[13] Hadzic, M. C. A. S. et al.: International Society for Optics and Photonics, 9711 (2016), 971119
[14] Zelger–Paulus, S. et al.: Methods in Molecular Biology (Clifton, N.J.), 2113 (2020), 1–16
[15] Ling, C. X. et al.: Springer US, 1 (2010)
[16] Buda, M. et al.: Neural Networks, 106 (2018), 249–259
[17] Lakkaraju, H. et al.: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 1675–1684
[18] Huang, G. et al.: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 2261-2269
[19] Szegedy, C. et al.: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, 1-9
[20] He, K. et al.: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 770-778
[21] Szegedy, C. et al.: AAAI Conference on Artificial Intelligence (AAAI-17), 2016, 4278- 4284