

Supplemental Information for Hidden Markov model analysis of fluorescence blinking in fluorescently labeled DNA

Tatsuhiro Furuta¹, Shuya Fan², Tadao Takada³, Yohei Kondo⁴, Mamoru Fujitsuka²,
Atsushi Maruyama⁵, Kiyohiko Kawai^{4,+}, and Kazuma Nakamura^{1,6,*}

¹*Graduate School of Engineering, Kyushu Institute of Technology, 1-1 Sensui-cho, Tobata-ku, Kitakyushu, Fukuoka 804-8550, Japan.*

²*SANKEN (The Institute of Scientific and Industrial Research), The University of Osaka, Mihogaoka 8-1, Ibaraki, Osaka 567-0047, Japan.*

³*Department of Applied Chemistry, Graduate School of Engineering, University of Hyogo, 2167 Shosha, Himeji, Hyogo 671-2280, Japan.*

⁴*Department of Life Science and Technology, Institute of Science Tokyo, Nagatsuta, Midori-ku, Yokohama, Kanagawa 226-8501, Japan.*

⁵*Department of Pharmacy, Tokyo University of Pharmacy and Life Sciences, Horinouchi, Hachioji, 192-0392, Japan.*

⁶*Integrated Research Center for Energy and Environmental Technologies, Kyushu Institute of Technology, 1-1 Sensui-cho, Tobata-ku, Kitakyushu, Fukuoka 804-8550, Japan.*

*kazuma@mns.kyutech.ac.jp

+kawai.k@life.isct.ac.jp

Benchmark for Gaussian emission hidden Markov model simulation

Here we present a benchmark test for the Gaussian emission hidden Markov model (HMM) used in this study. The benchmark validation is based on the method presented in Ref.[1]. Computational procedure and analysis are summarized as follows:

- (1) Construction of the ground-truth synthetic blinking traces:

The true ON and OFF dwell-time distributions were drawn as exponential distributions with time constants $\tau_{\text{ON}} = 17.6$ ms, $\tau_{\text{OFF}} = 7.8$ ms, and these timescales are based on the analysis results for the experimental time series addressed in the main text. One hundred ON and one hundred OFF dwell times were independently sampled from the exponential distributions and concatenated to generate a ground-truth hidden sequence (“synthetic trace 1”), where the mean intensities of the ON and OFF states were set to $1 + \Delta$ and 1, respectively. The synthetic trace 1 is binned in the time-bin 0.5 ms.

- (2) Addition of noise:

Gaussian noise with standard deviation $\sigma = 0.5$ was added to generate realistic noisy observations (“synthetic trace 2”). The signal-to-noise ratio of each synthetic dataset is thus characterized by $S/N = \Delta/\sigma$. We generated 50 independent realizations of synthetic trace 2.

- (3) HMM inference and validation:

We applied our HMM analysis to the standardized synthetic trace 2 and obtained the inferred hidden sequence. Using this inferred sequence, we reconstructed the ON and OFF dwell-time histograms from the 50 samples and compared them with the ground-truth exponential distributions set in the step (1). We performed this procedure for a range of S/N values.

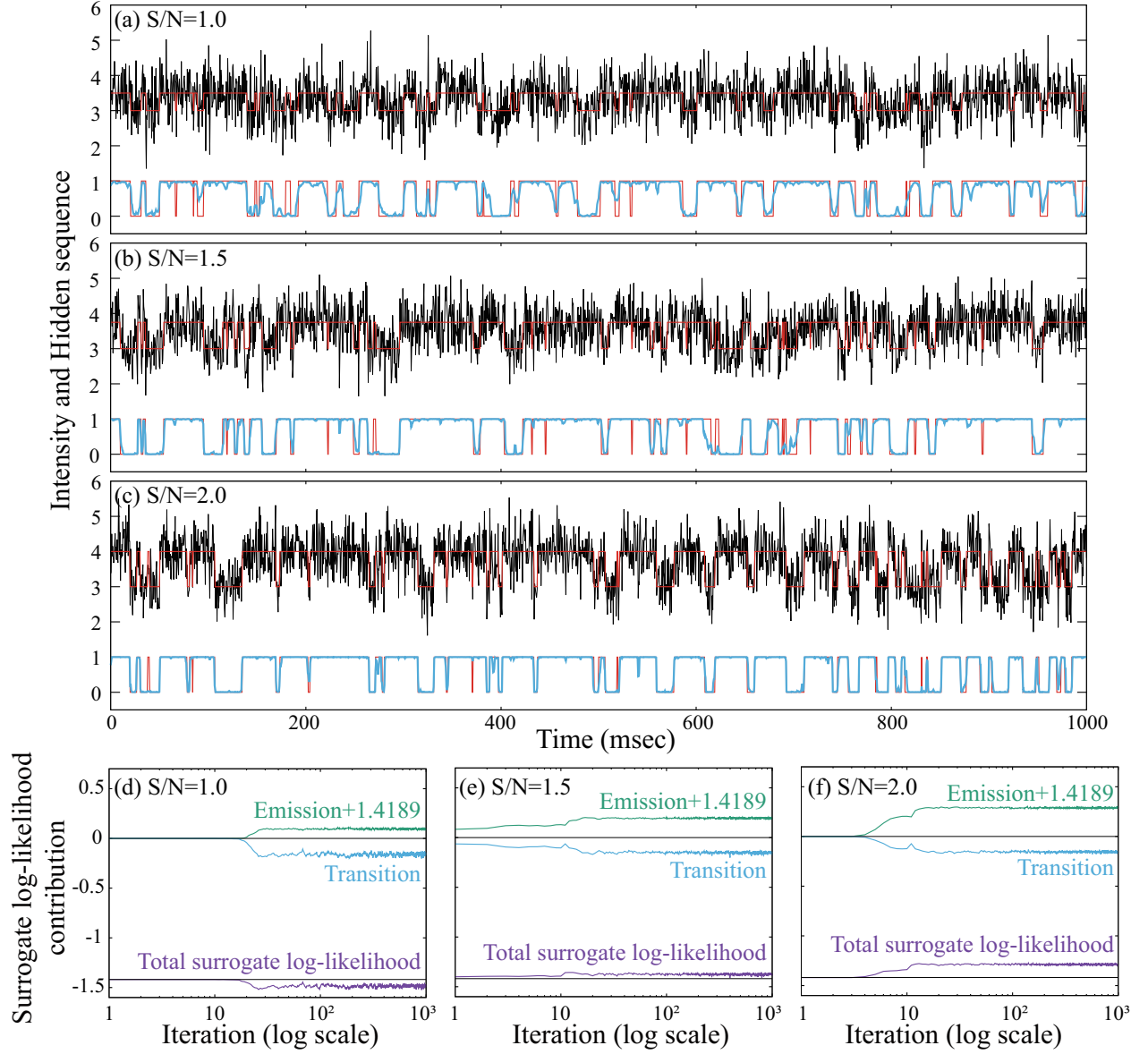
(4) Results: S/N threshold for reliable recovery:

Figure S1 (a–c) compares three analysis results for benchmark time series: (a) $S/N = 1.0$, (b) $S/N = 1.5$, (c) $S/N = 2.0$. The upper part of each figure shows resulting “synthetic trace 1” (ground truth, red line) and “synthetic trace 2” (noisy trace, black line), and the lower part compares the ground-truth synthetic trace 1 and the inferred hidden sequence (cyan line). Figure S1(d–f) represent iteration-step dependence of the surrogate objective of the log-likelihood $\mathcal{L}$ defined in Eq. (22) of the main text, used as a convergence and reliability indicator. The $\mathcal{L}$ is shown together with its decomposition into emission and transition contributions. Panels (d), (e), and (f) correspond to the data shown in panels (a), (b), and (c), respectively. Figure S2 displays our calculated blinking plots (purple bars), compared with the ground-truth dwell-time distribution (green curves). Panels (a,b), (c,d), and (e,f) correspond to $S/N = 1.0, 1.5$, and 2.0 , respectively.

From both qualitative inspection of the ON/OFF blinking plots and quantitative evaluation using the surrogate log-likelihood $\mathcal{L}$, we found that a reliable recovery is achieved around $\Delta/\sigma \sim 1.5$; for $\Delta/\sigma > 1.5$, the HMM accurately recovers both the hidden sequence and the underlying exponential dwell-time distributions. We also see that $\mathcal{L}$ is larger than the baseline value of -1.4189 , which corresponds to a trivial reference case where both the emission and transition contributions to $\mathcal{L}$ vanish. For $\Delta/\sigma < 1.5$, the HMM tends to produce spurious transitions and the reconstructed dwell-time distributions deviate significantly from the ground truth, particularly in the short duration region. We also see that $\mathcal{L}$ is lower than the baseline of -1.4189 .

(5) Implication for the experimental data:

Most of the experimental traces analyzed in this study satisfy S/N above this threshold, because $\mathcal{L}$ is basically larger than the baseline (see “Analysis results for all 40 experimental datasets” in Supporting information), except for the specific trace #40 shown in Figs. 5(c) of the main text, which indeed fell below the baseline and was therefore excluded from the final analysis. We believe that this synthetic-data benchmark provides a quantitative and objective criterion for assessing the reliability of our HMM-based reconstruction of the hidden state sequence $\mathbf{S}$.



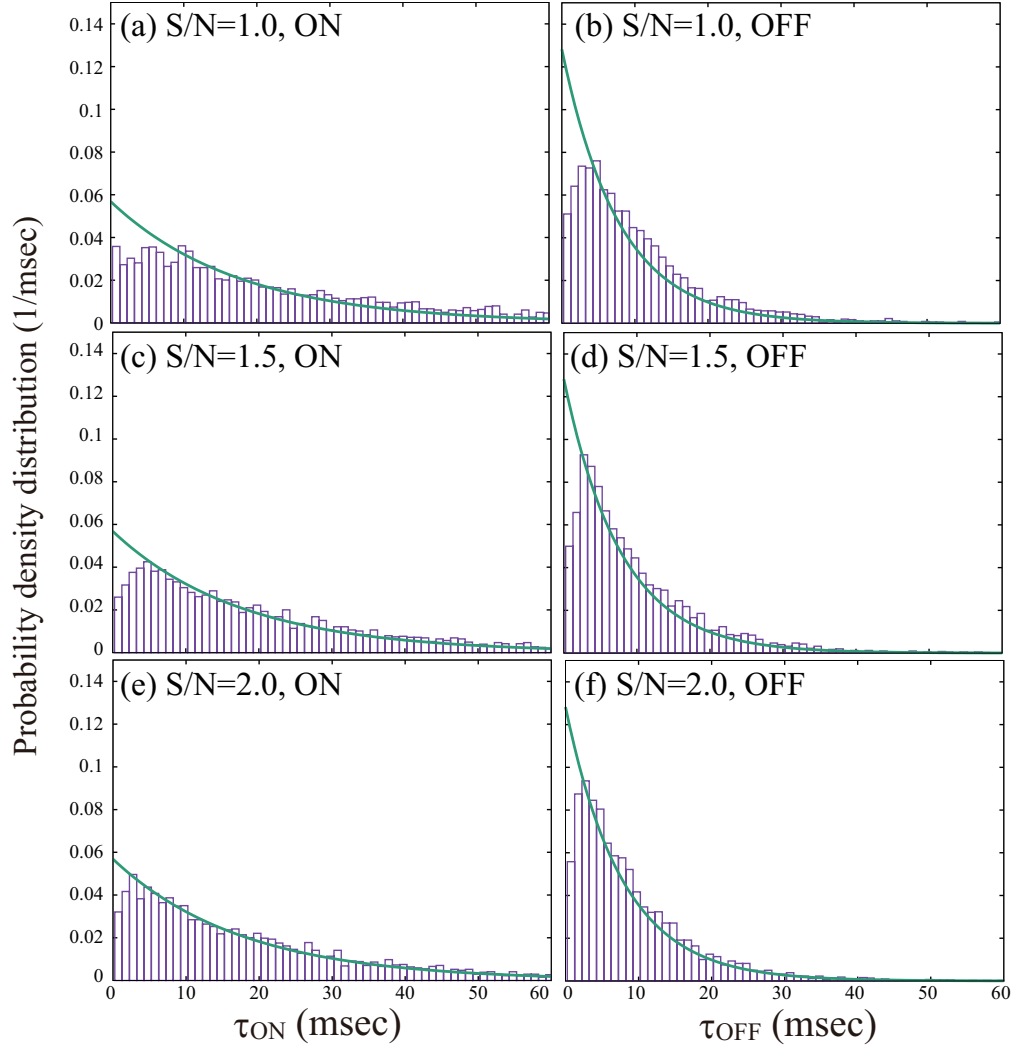


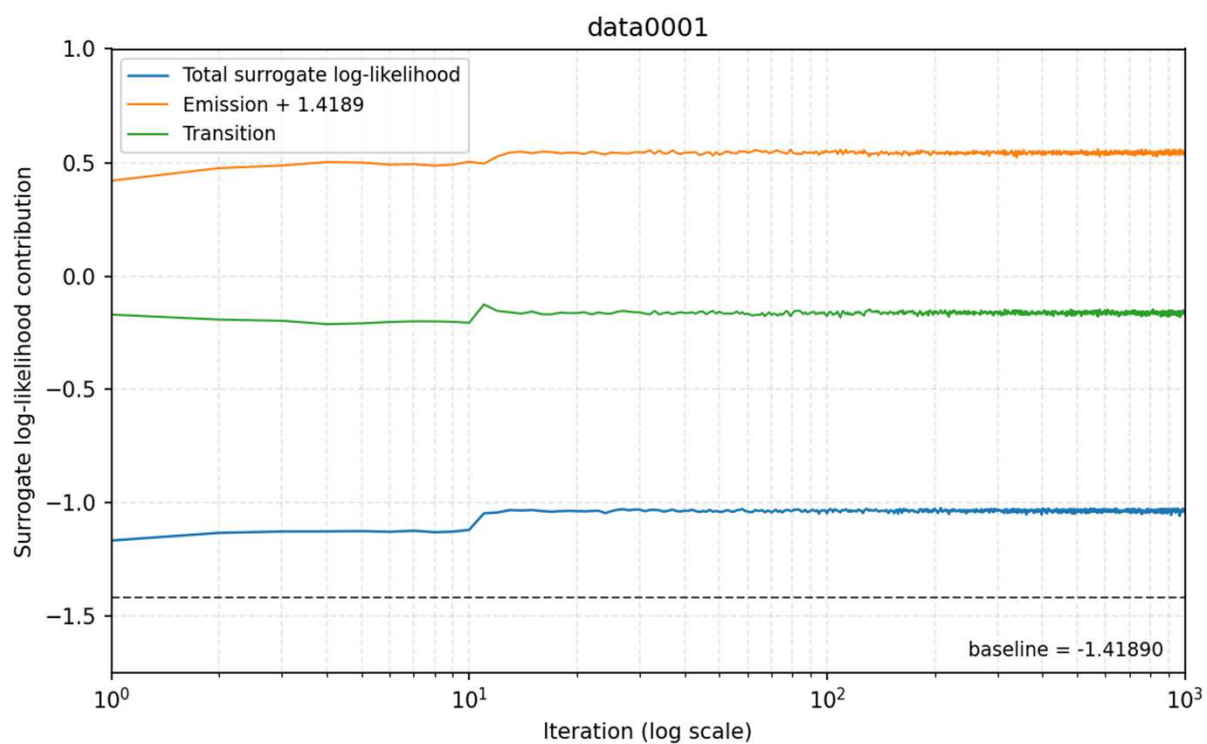
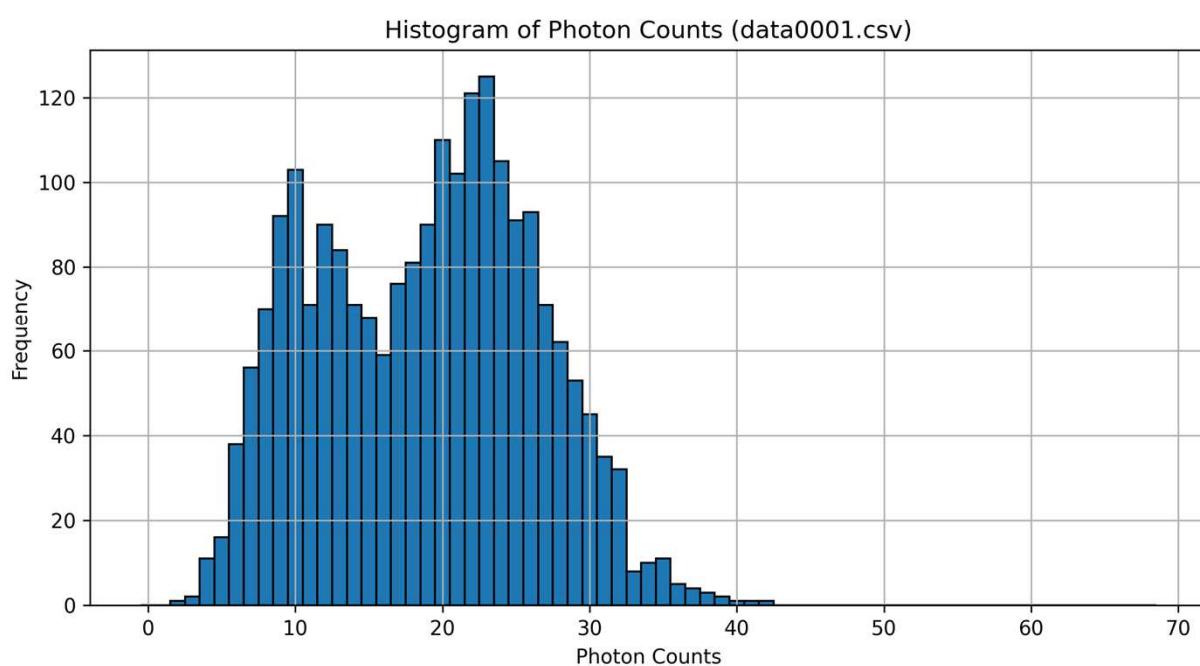
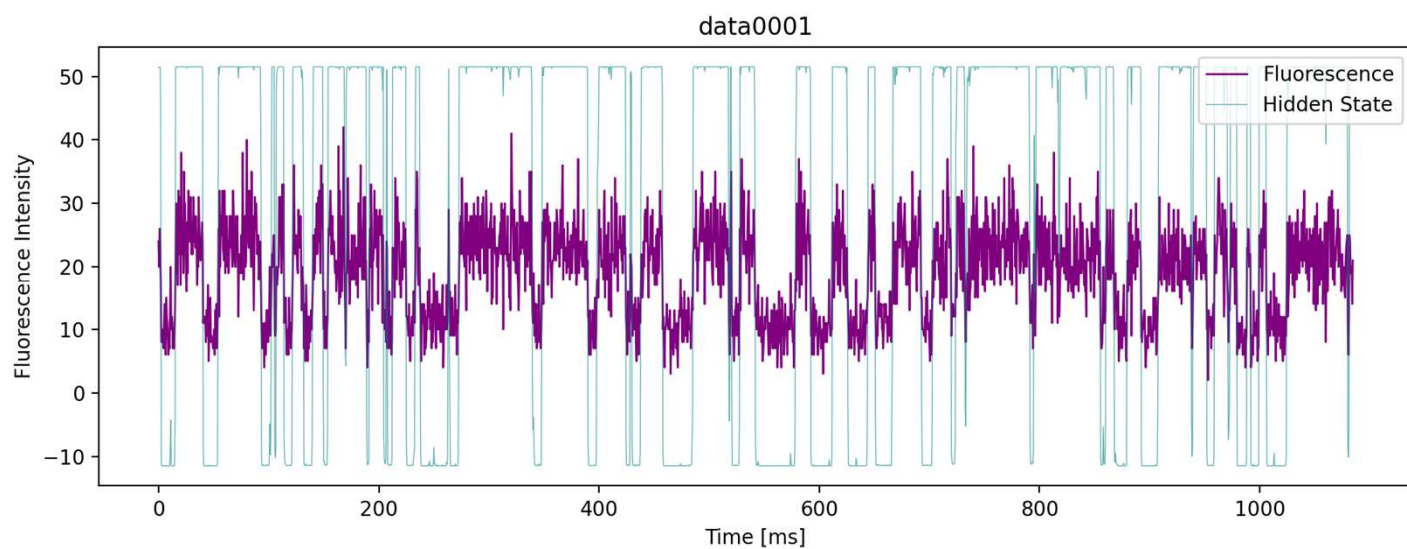
Figure S2: Blinking plots calculated for the benchmark time series. Panels (a,b), (c,d), and (e,f) correspond to signal-to-noise ratios $S/N = 1.0$, 1.5 , and 2.0 , respectively. Histograms show the distributions of the inferred ON- and OFF-state dwell times, averaged over 50 independent realizations. Green curves indicate the ground-truth exponential distributions of the form $(1/s) \exp(-\tau/s)$, with $s = 17.56$ ms for the ON state and $s = 7.82$ ms for the OFF state.

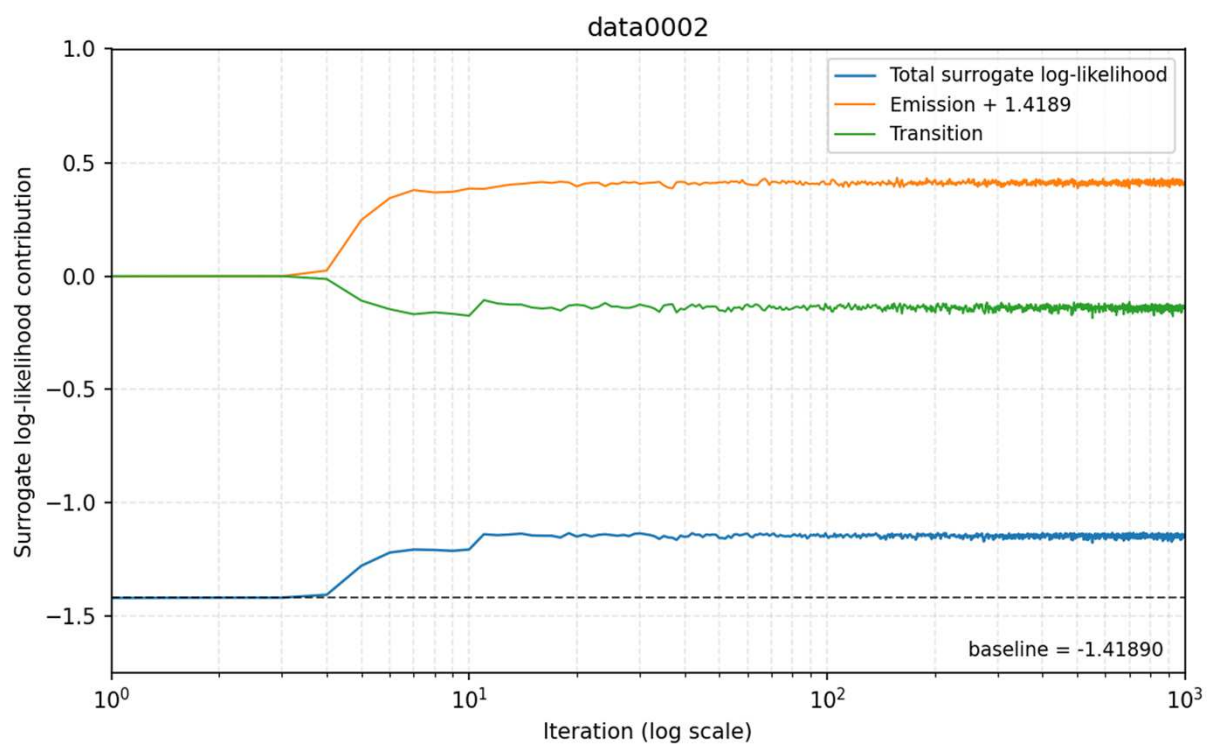
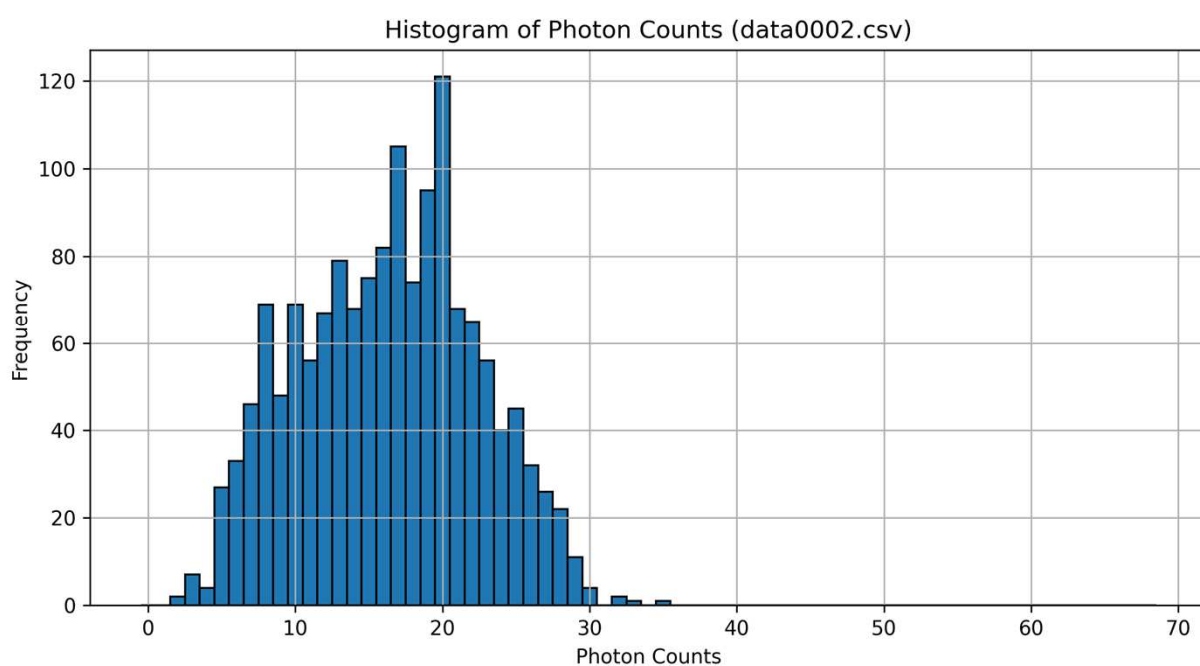
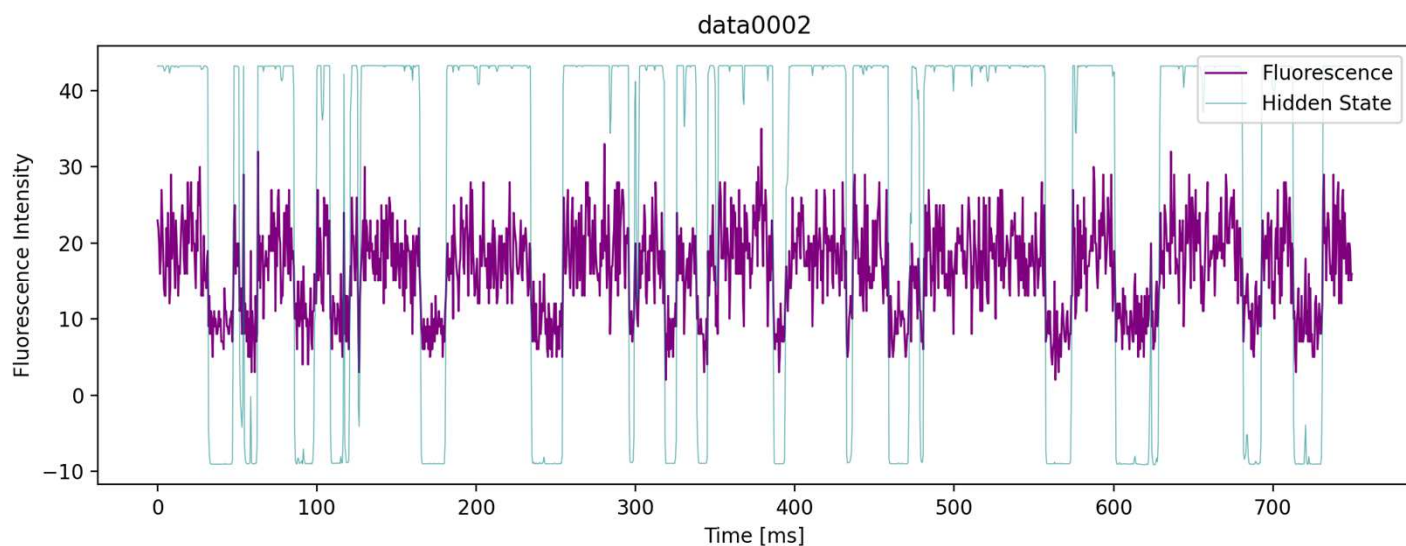
Analysis results for all 40 experimental datasets

In this Supplemental Information, we provide the analysis results for all 40 experimental datasets, including the photon-count time-series analysis, the corresponding histograms, and the iteration-step dependence of the surrogate objective $\mathcal{L}$. Data were collected for a total of 40 single-molecule trajectories (**data0001–data0040**). On each page, the upper panel shows the photon-count time series, where the solid purple line represents the number of photons detected within each 0.5 ms time bin, and the thin green line indicates the hidden-state sequence inferred from the hidden Markov model simulation. The middle panel displays the histogram of photon counts, while the lower panel presents the surrogate objective $\mathcal{L}$ as a function of the iteration step, decomposed into the emission and transition contributions. Among the 40 datasets, **data0040** was excluded from the calculation of the blinking plot because the inferred hidden-state sequence exhibited substantial noise. Notably, the photon-count histogram of **data0040** shows a unimodal pattern, and its total surrogate objective $\mathcal{L}$ is appreciably lower than those of the other datasets.

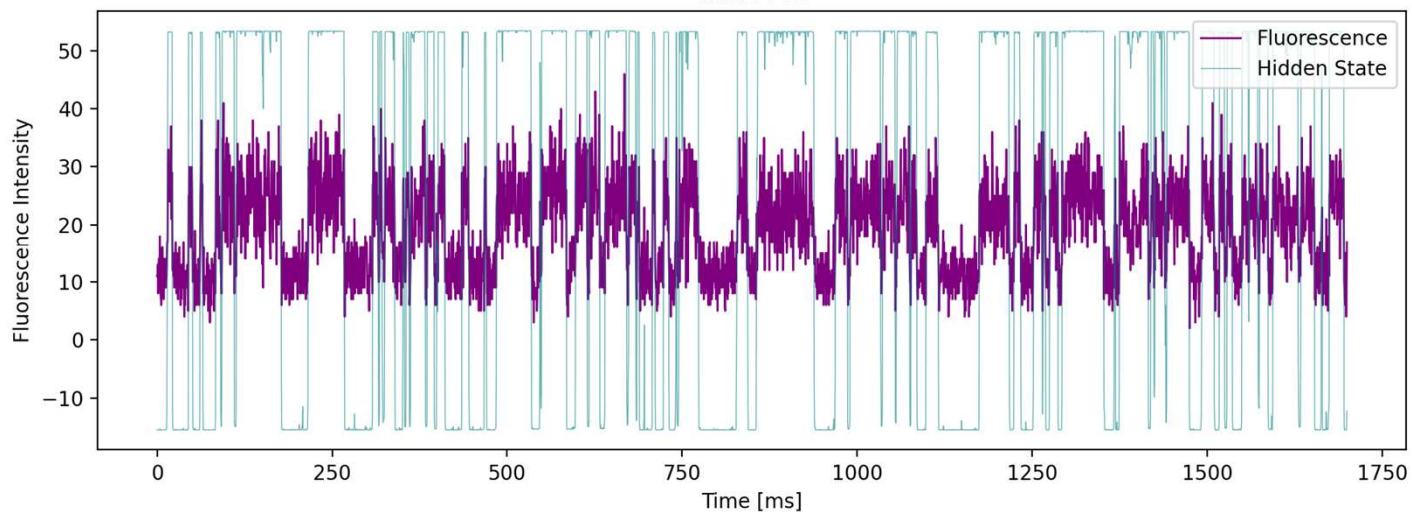
References

- [1] Tatsuhiro Furuta, Keisuke Hamada, Masaru Oda, and Kazuma Nakamura. Hidden markov model analysis for fluorescent time series of quantum dots. *Phys. Rev. B*, 106:104305, Sep 2022.

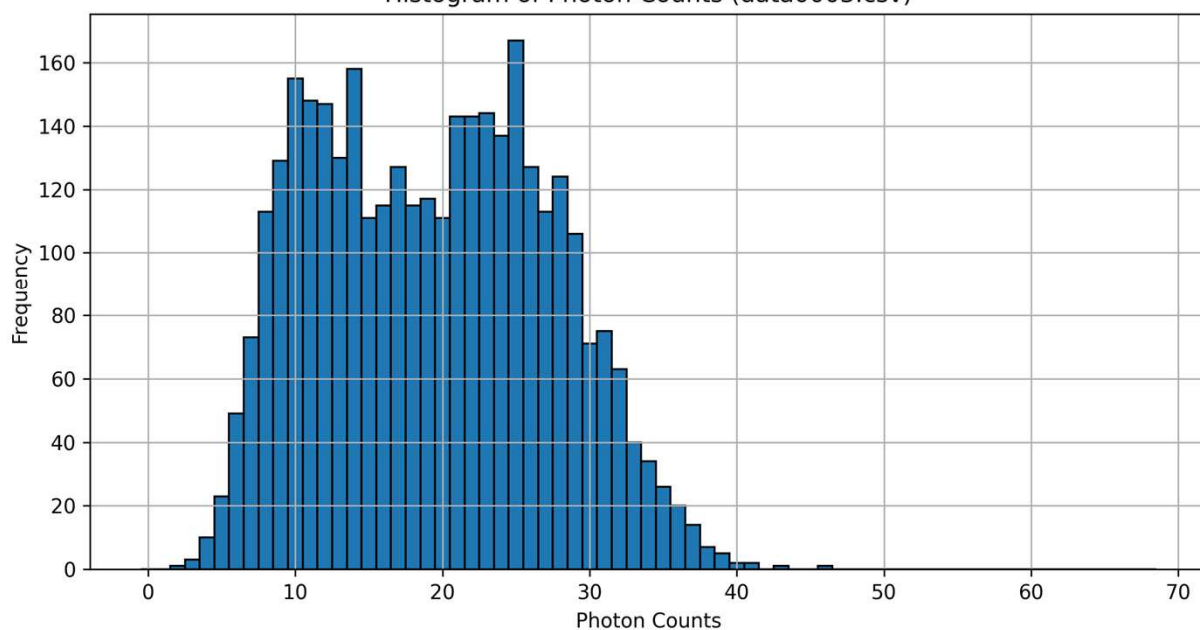




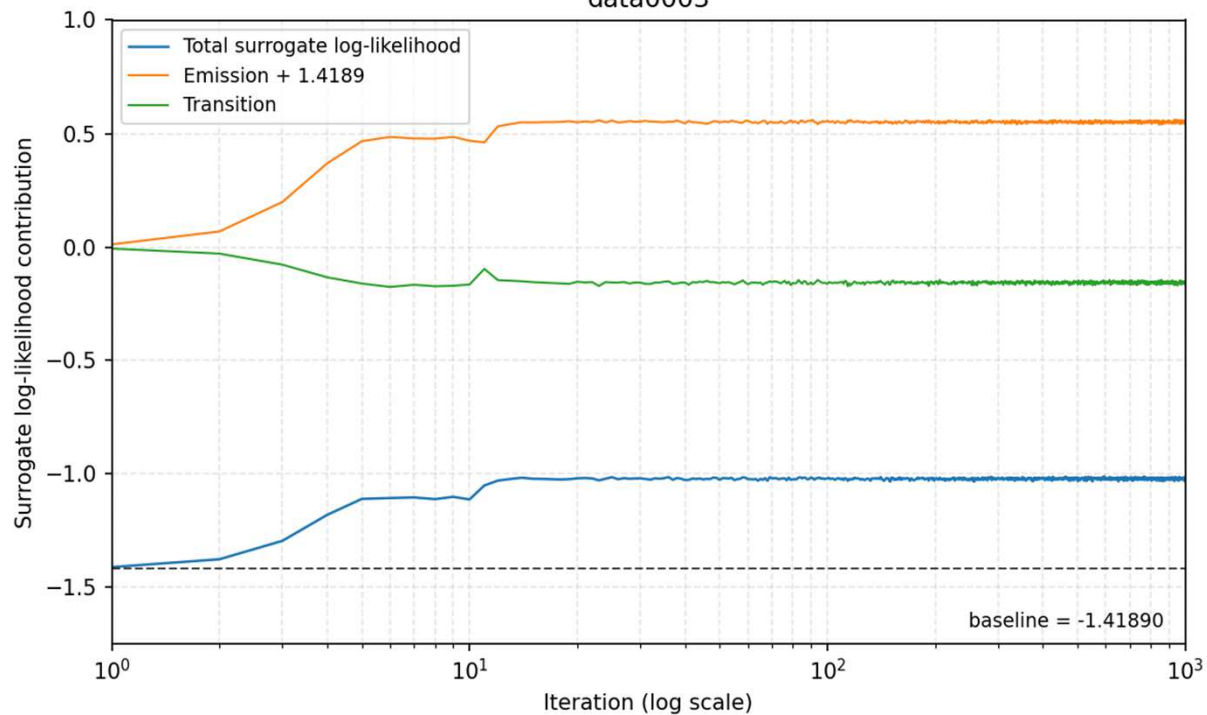
data0003



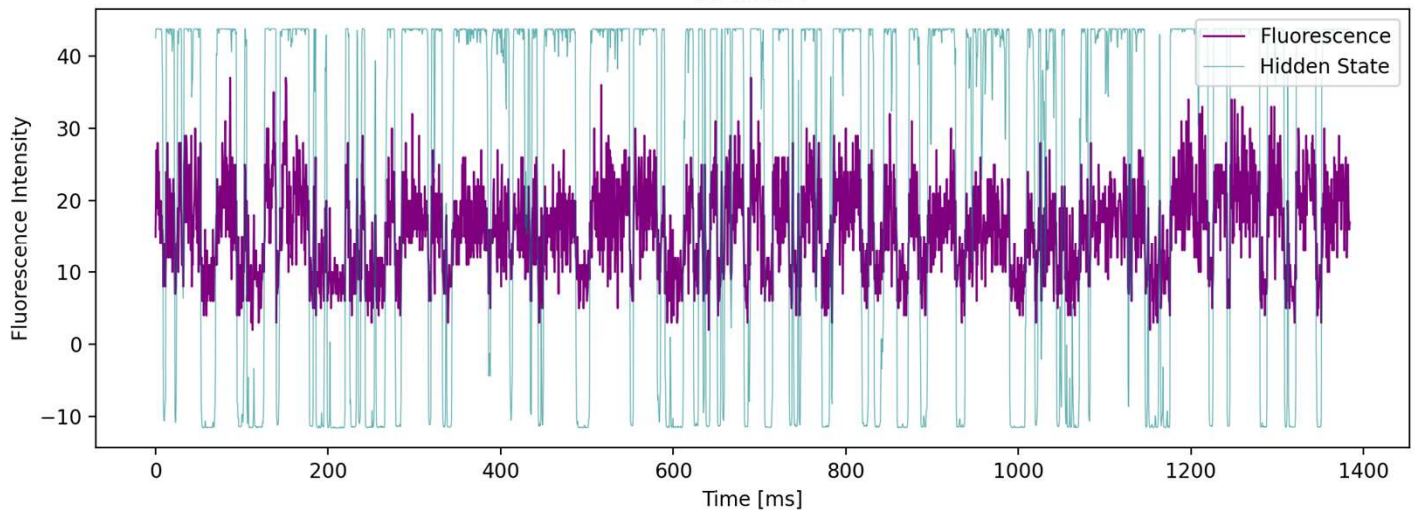
Histogram of Photon Counts (data0003.csv)



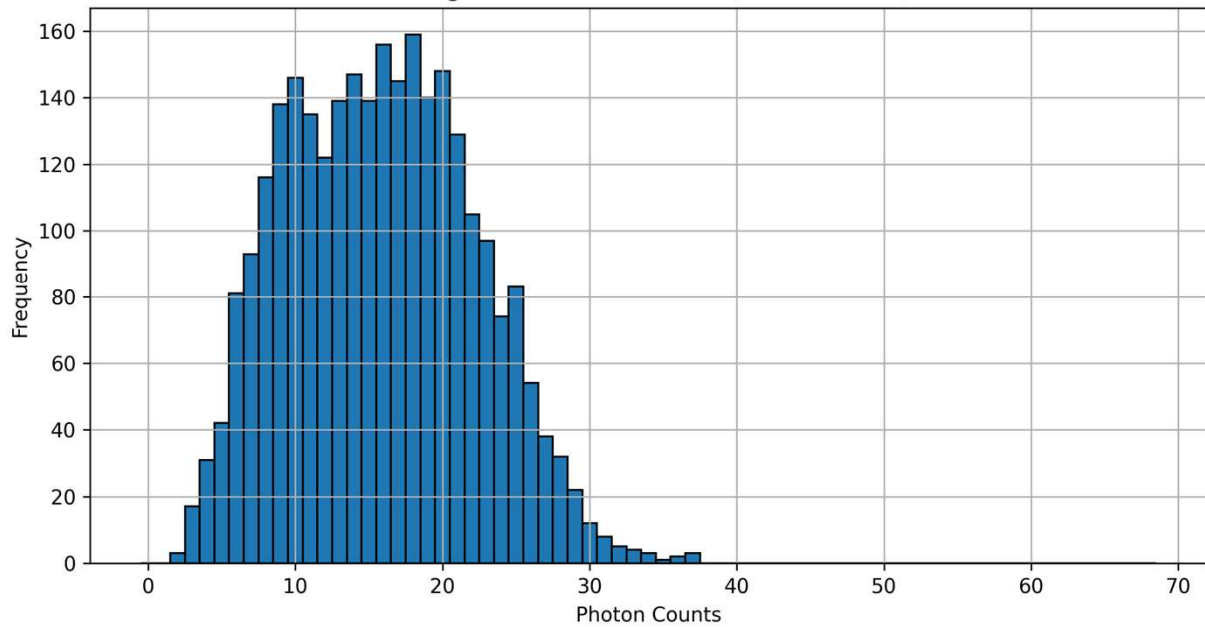
data0003



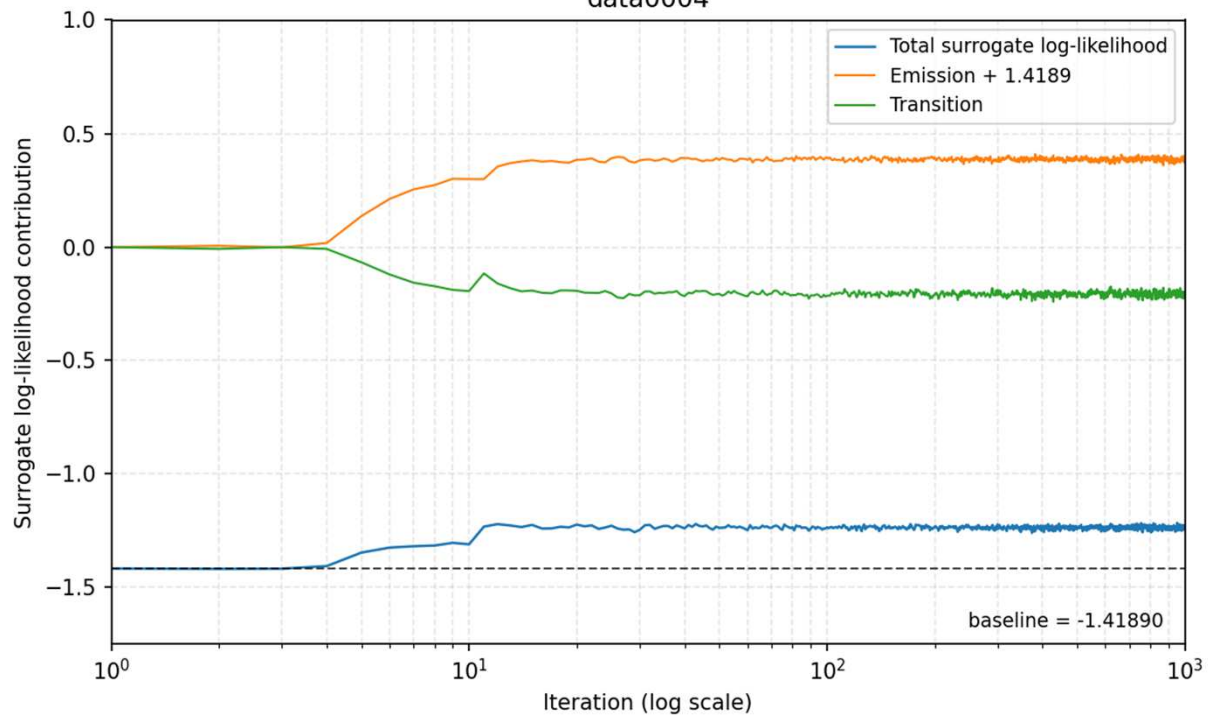
data0004

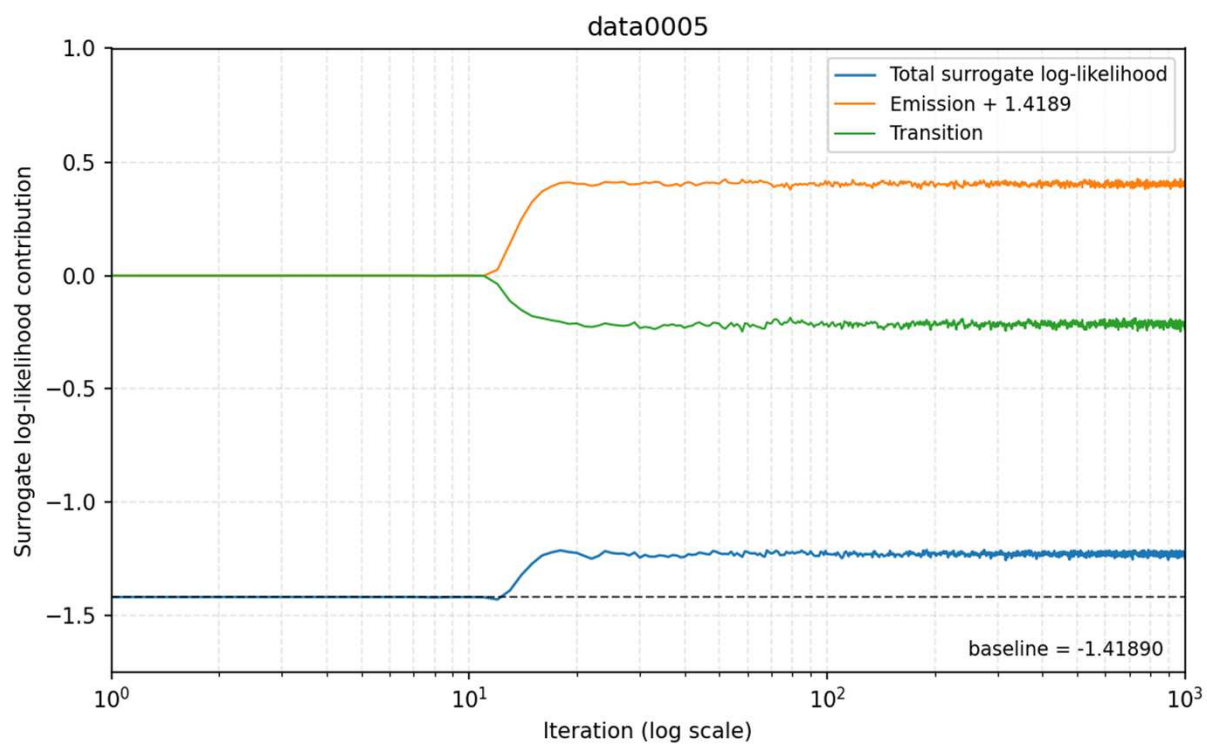
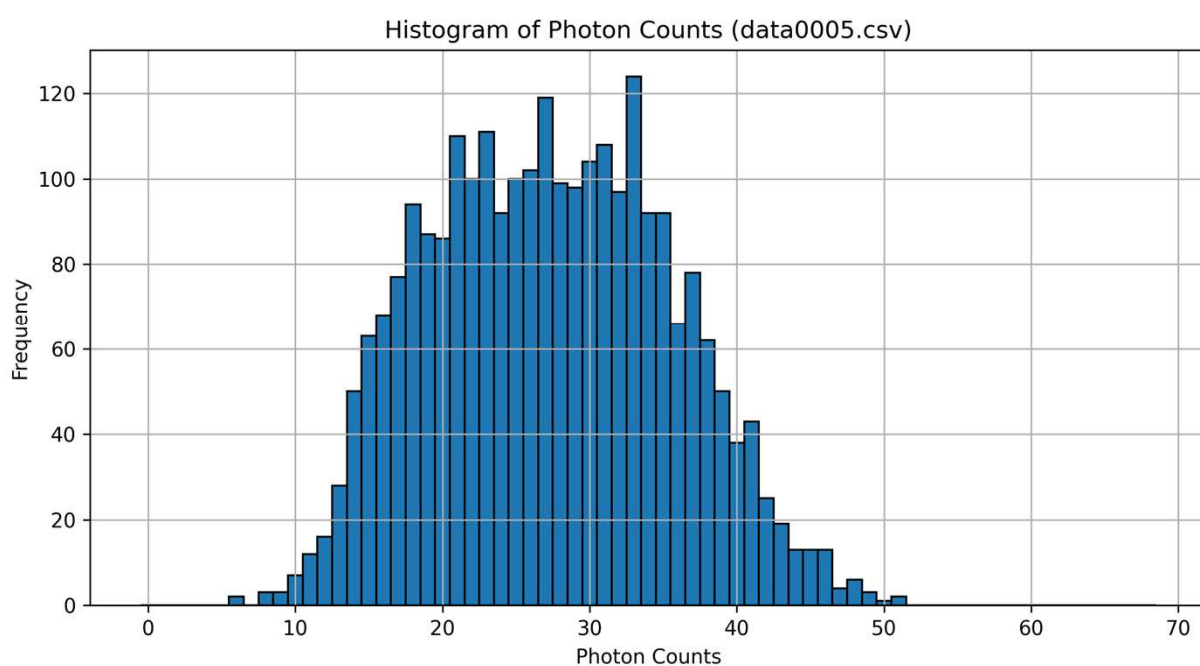
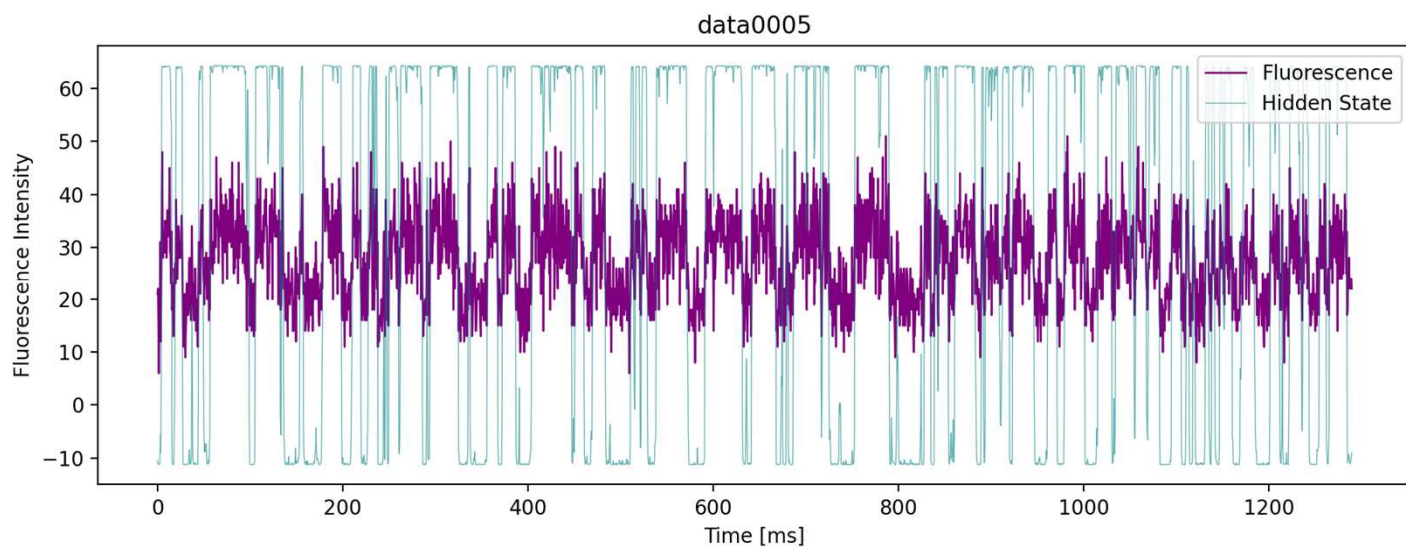


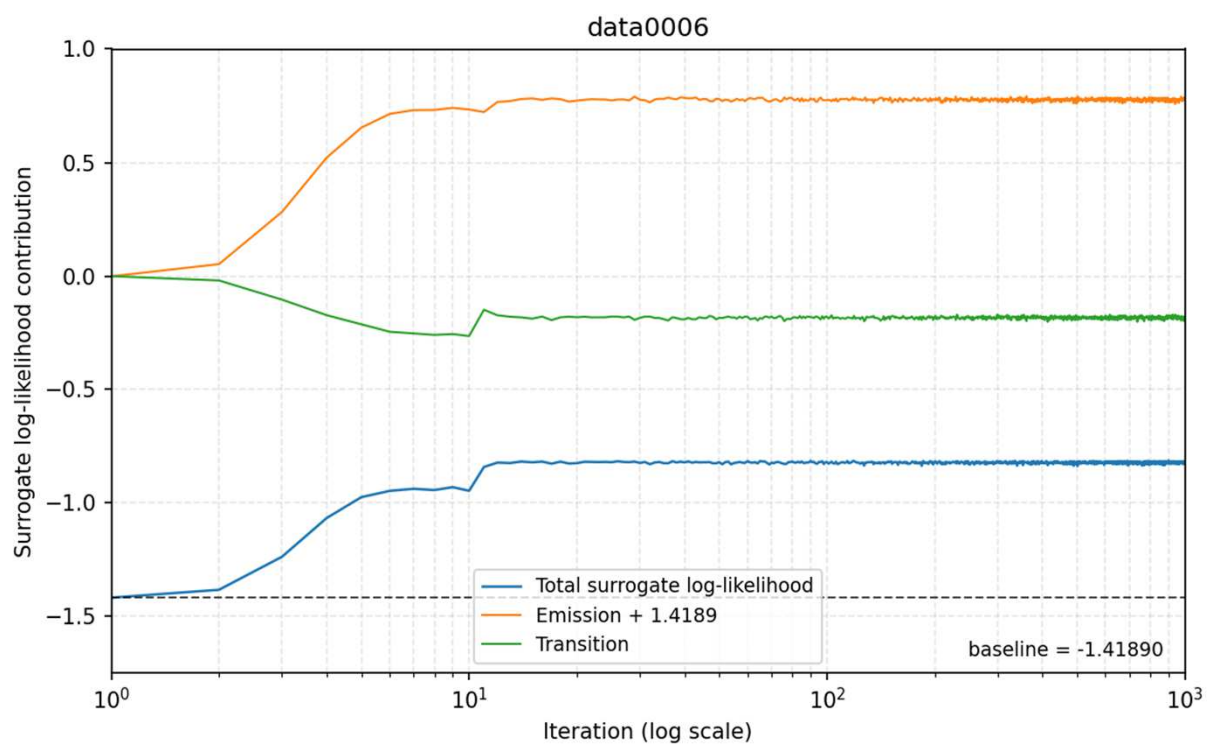
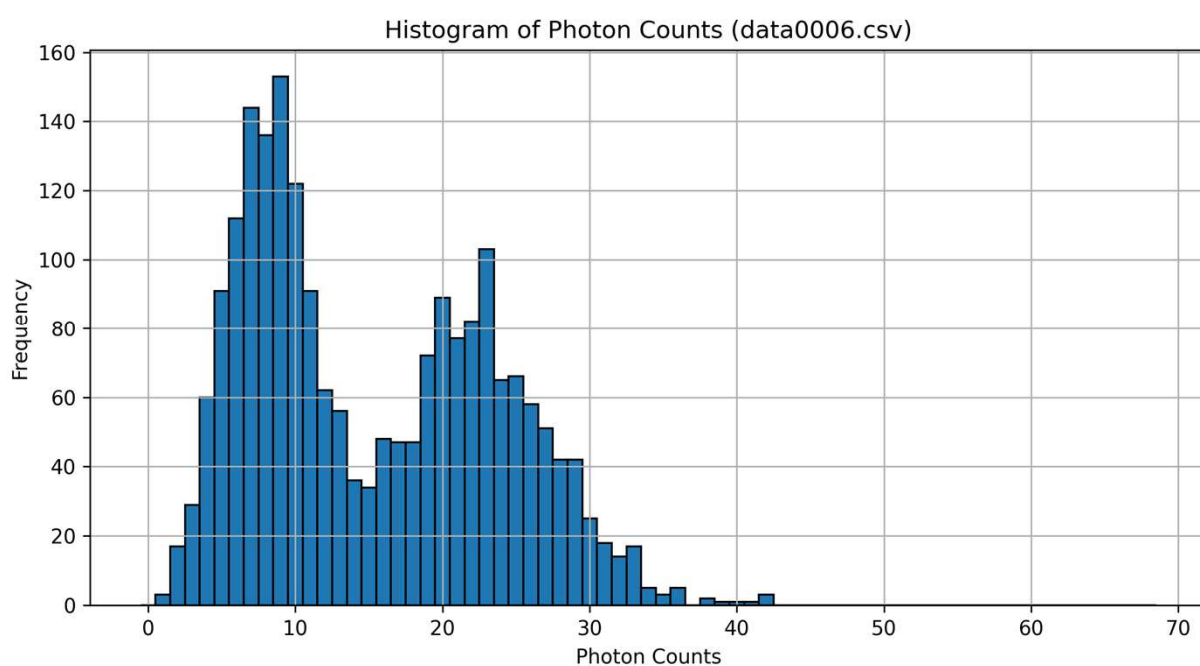
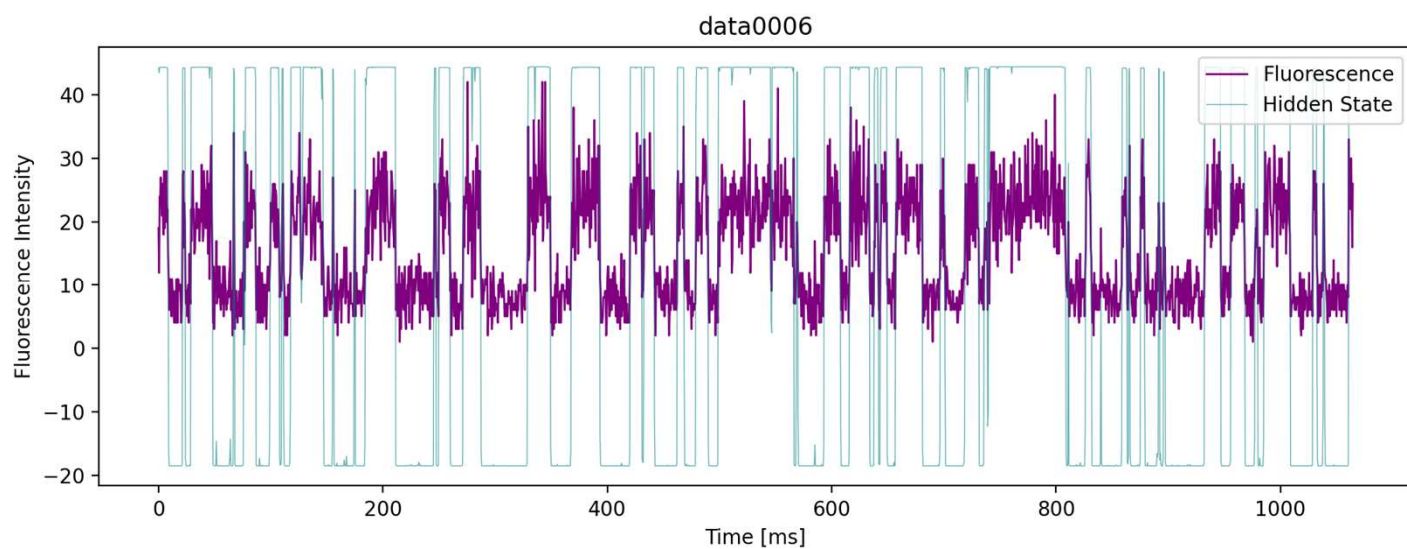
Histogram of Photon Counts (data0004.csv)

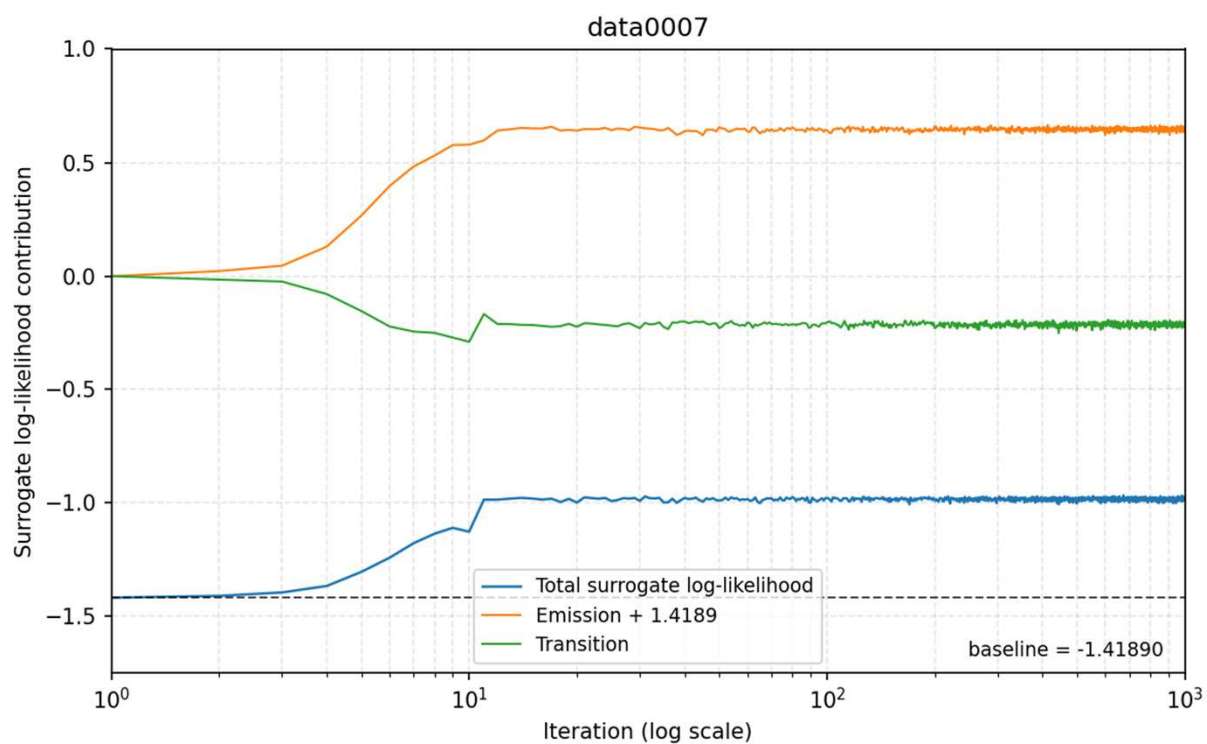
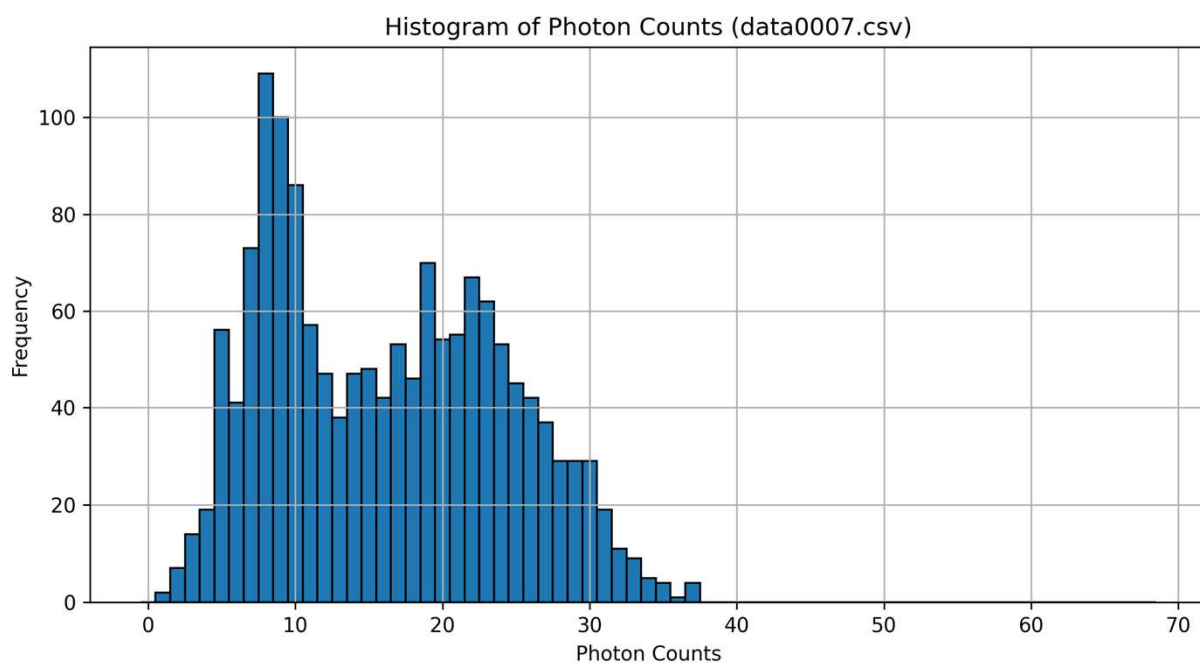
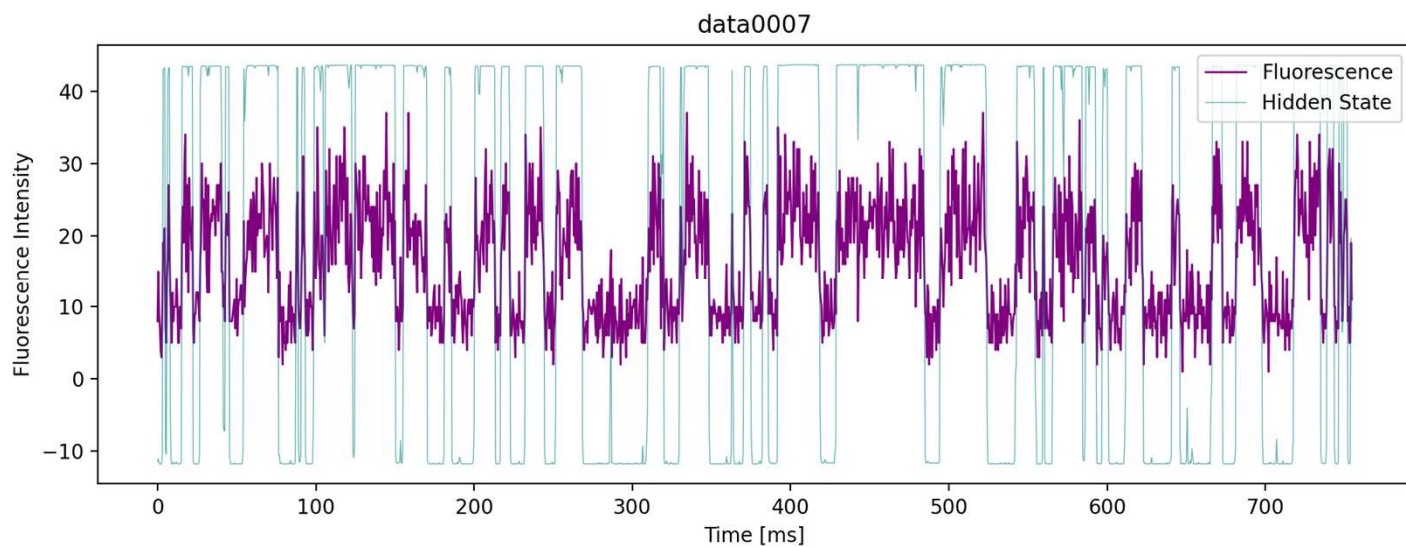


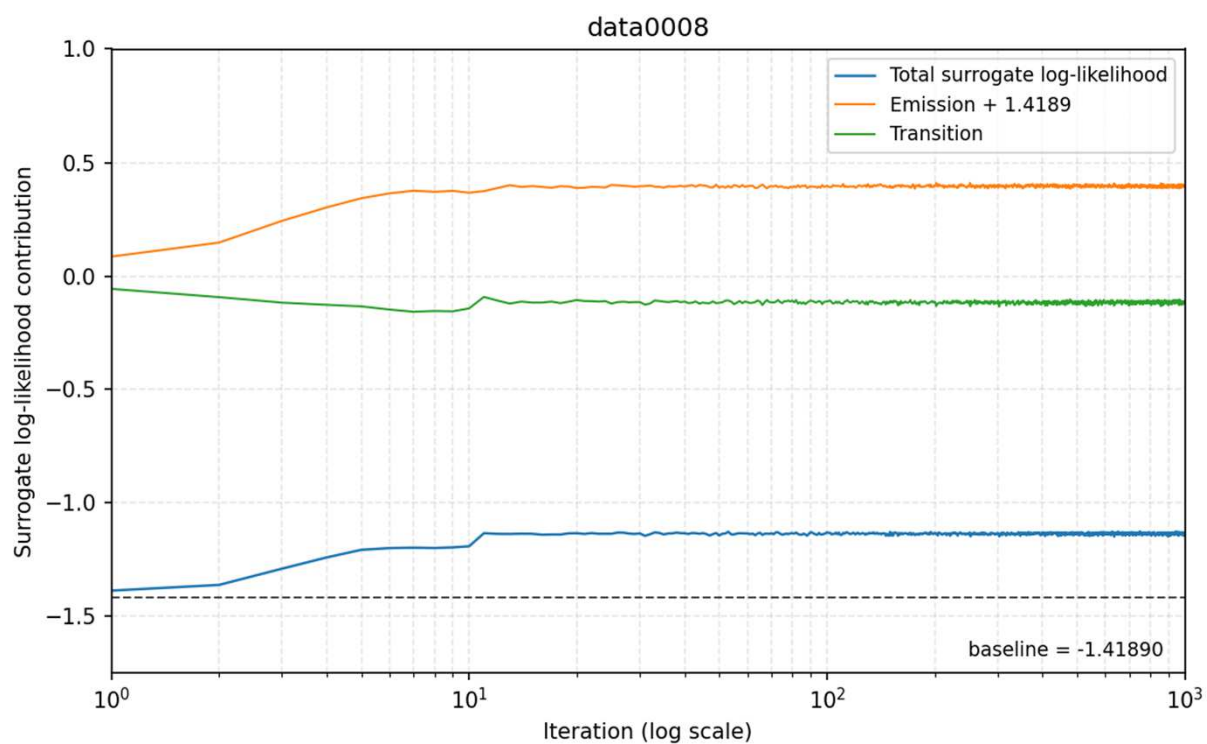
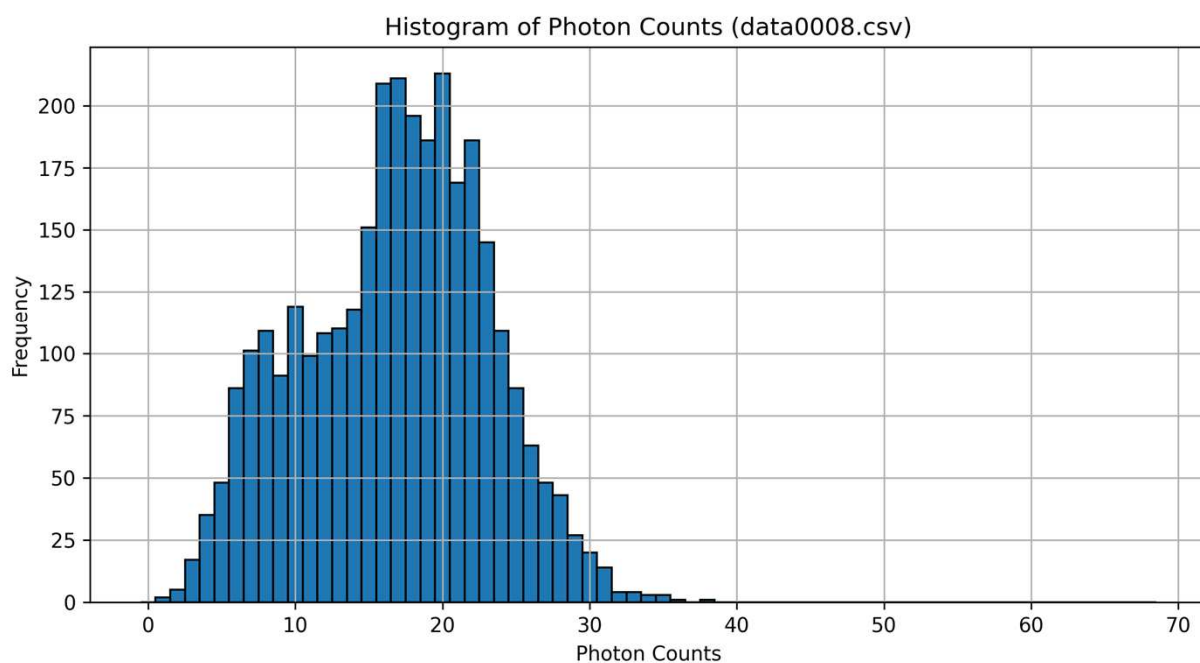
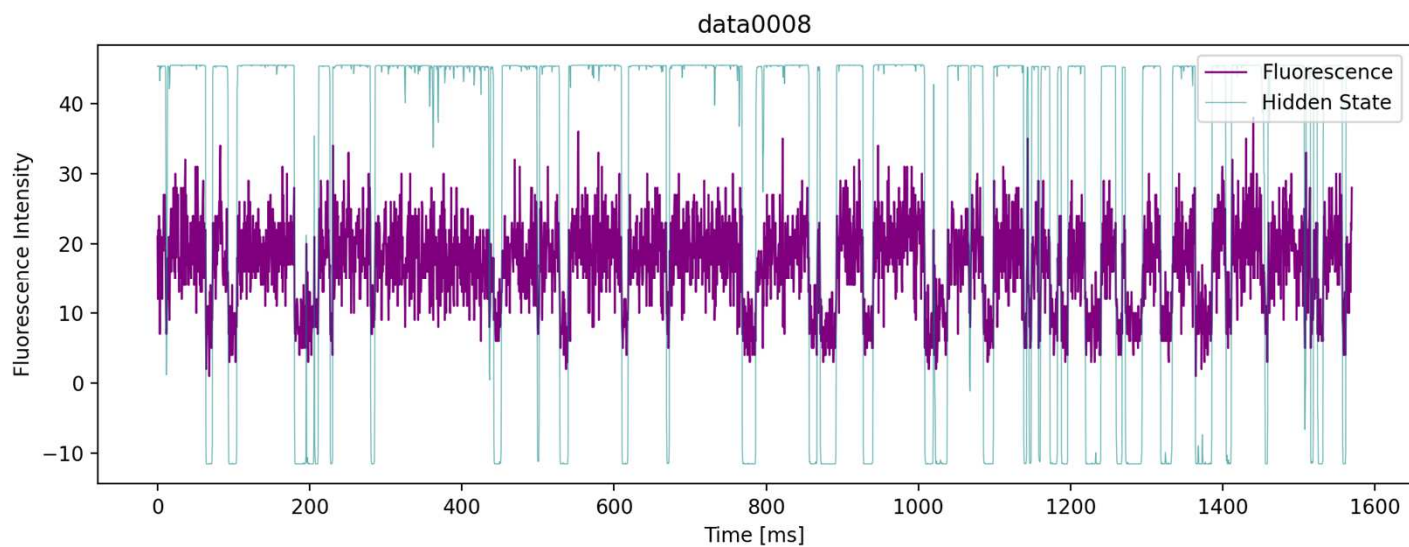
data0004

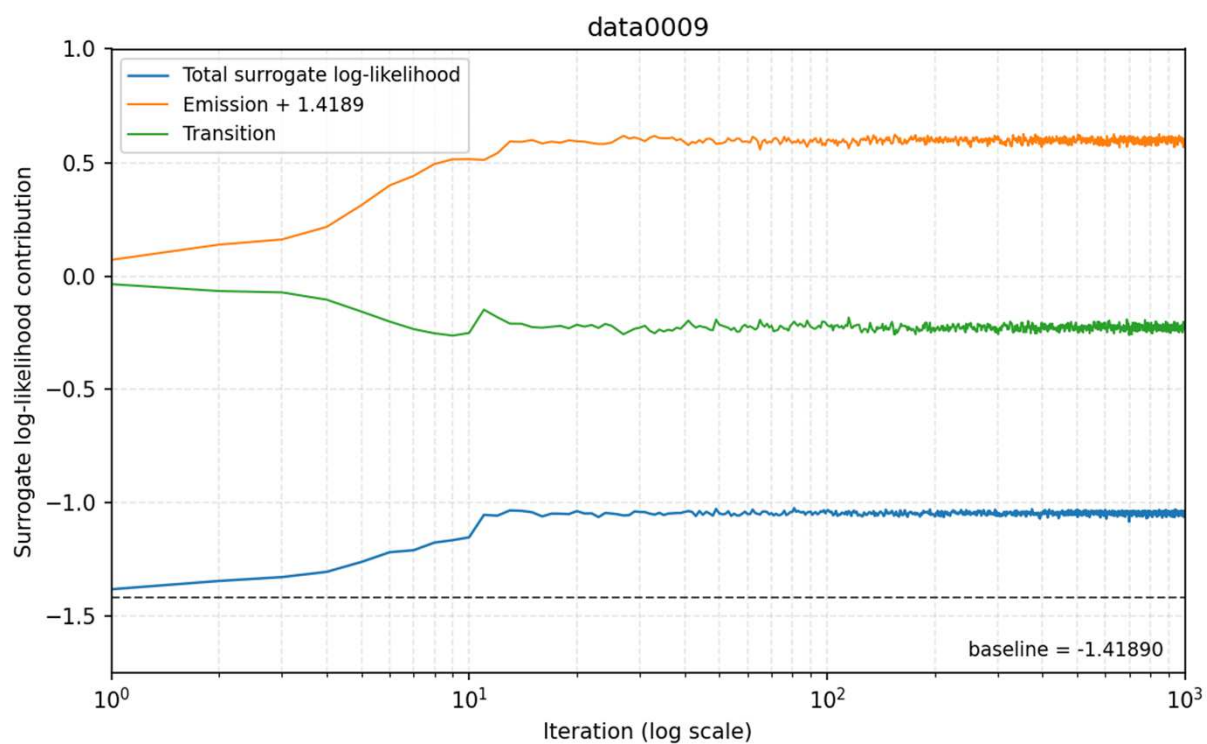
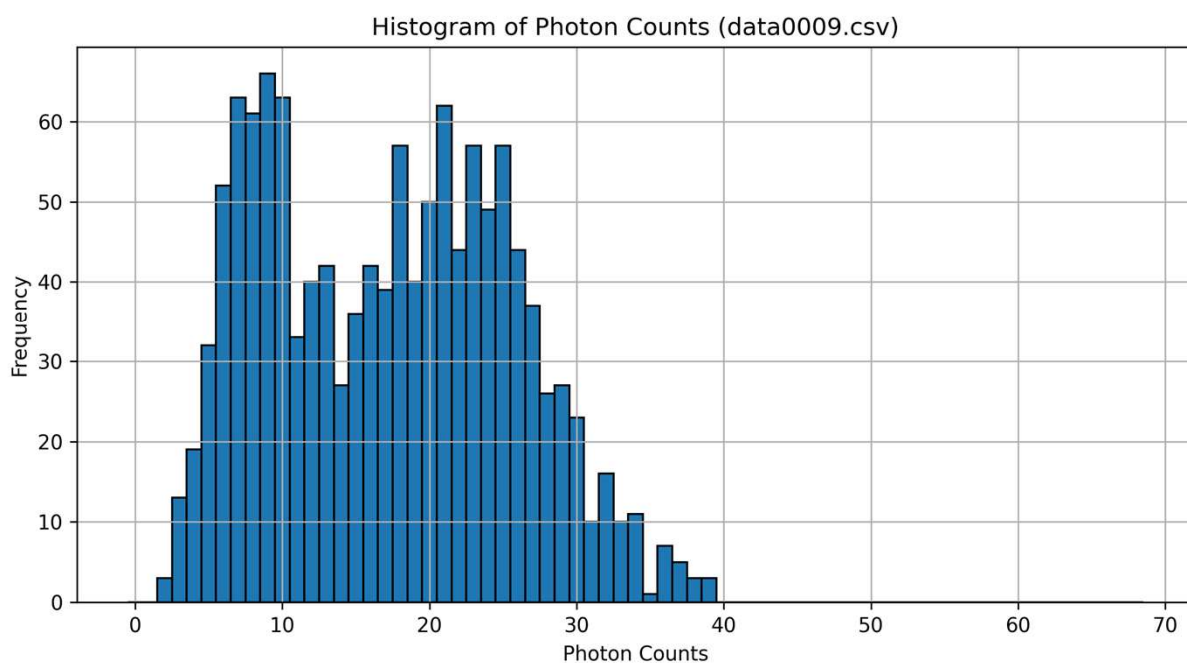
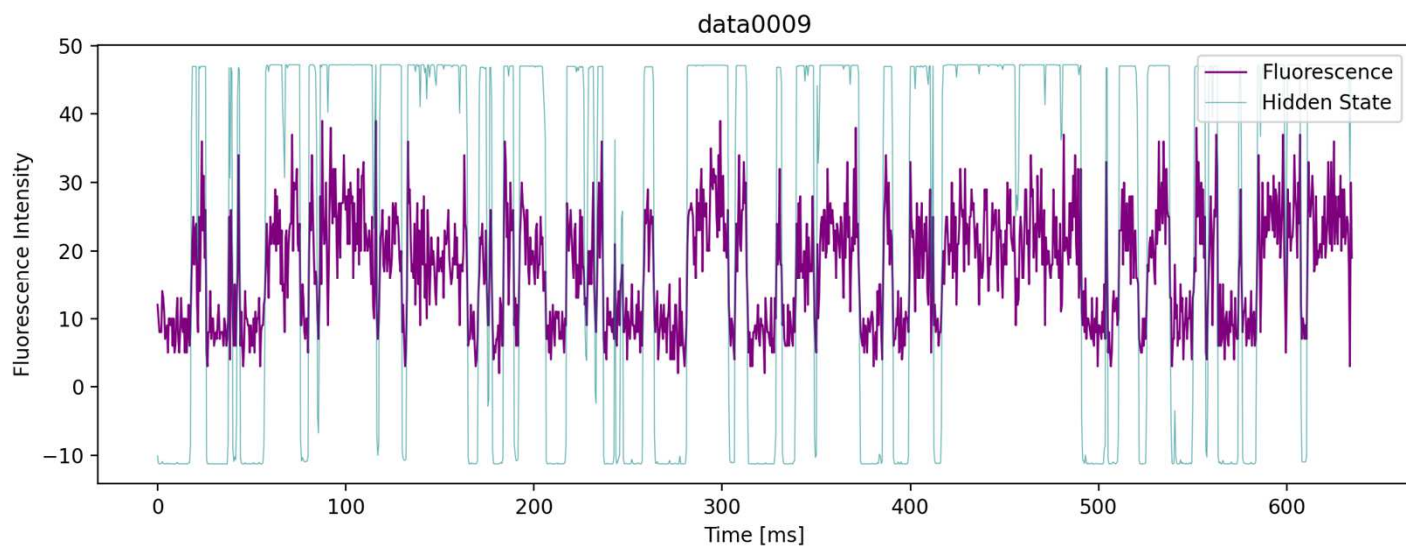


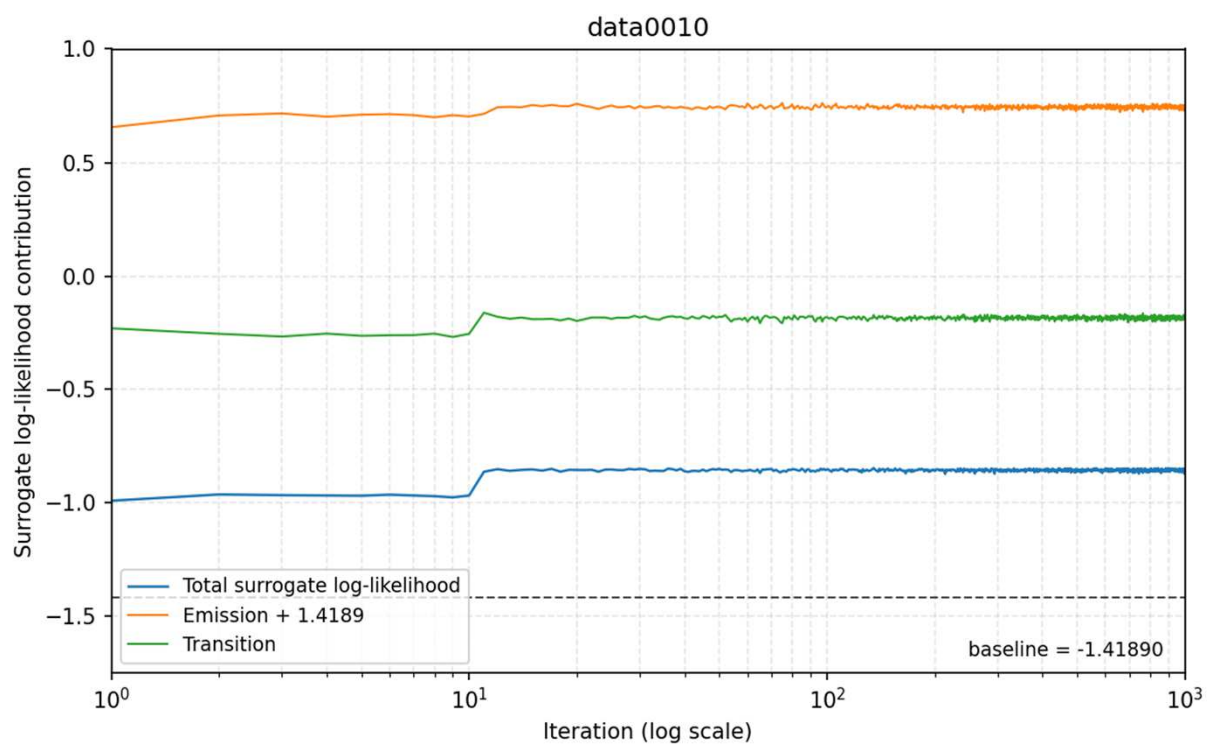
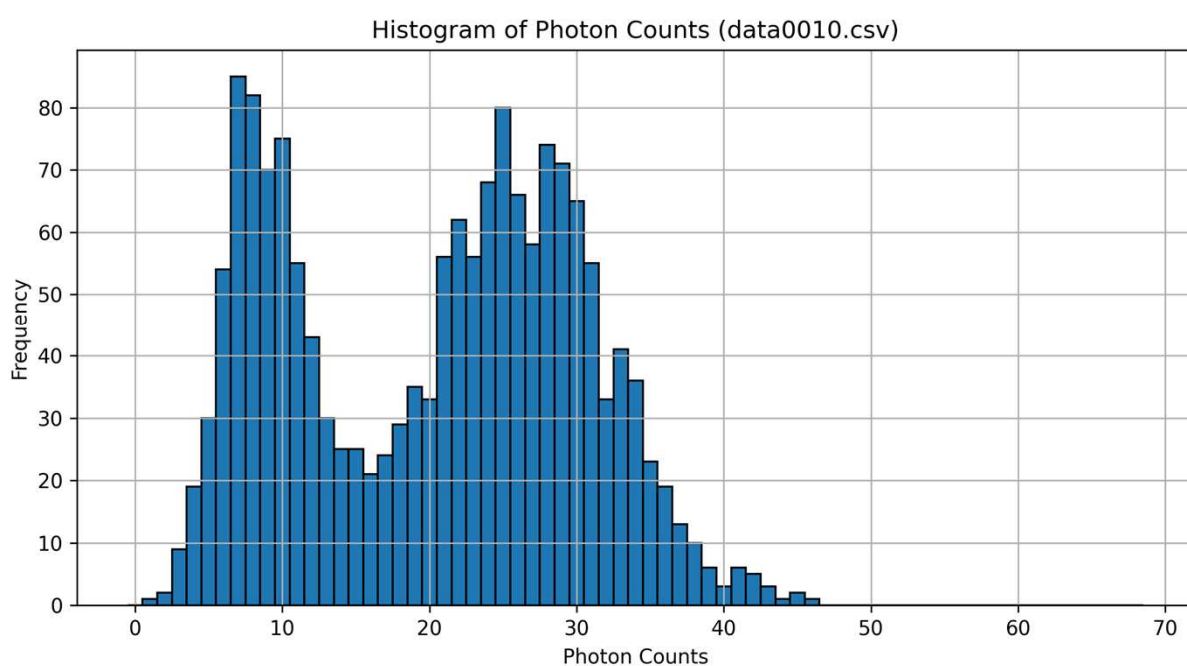
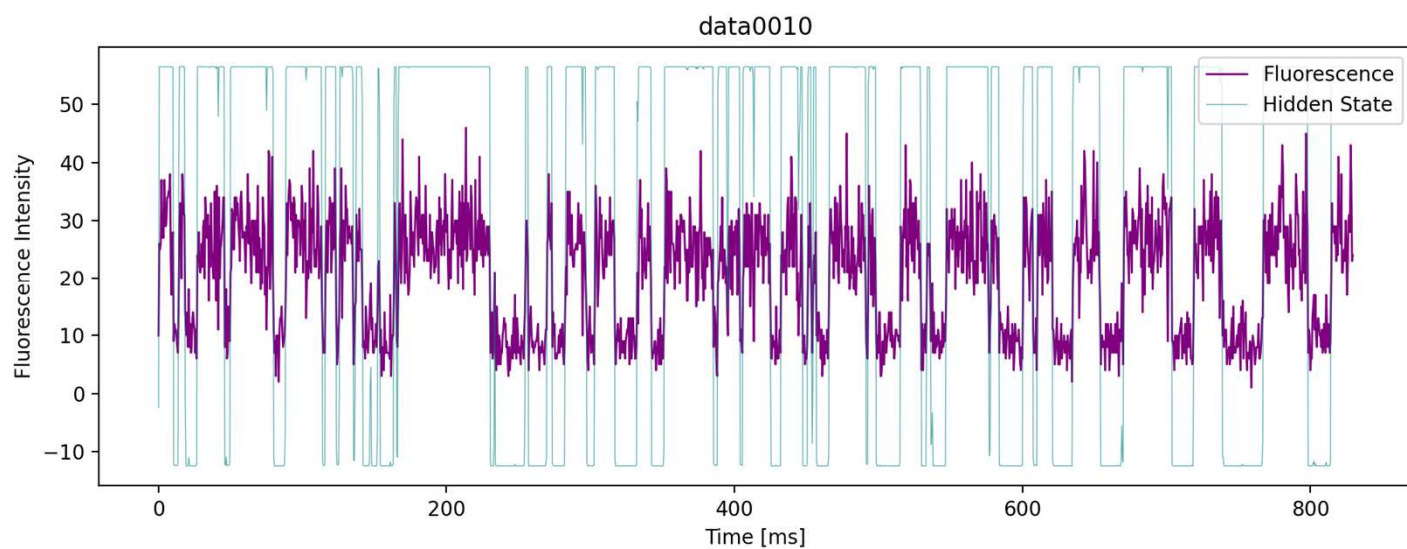


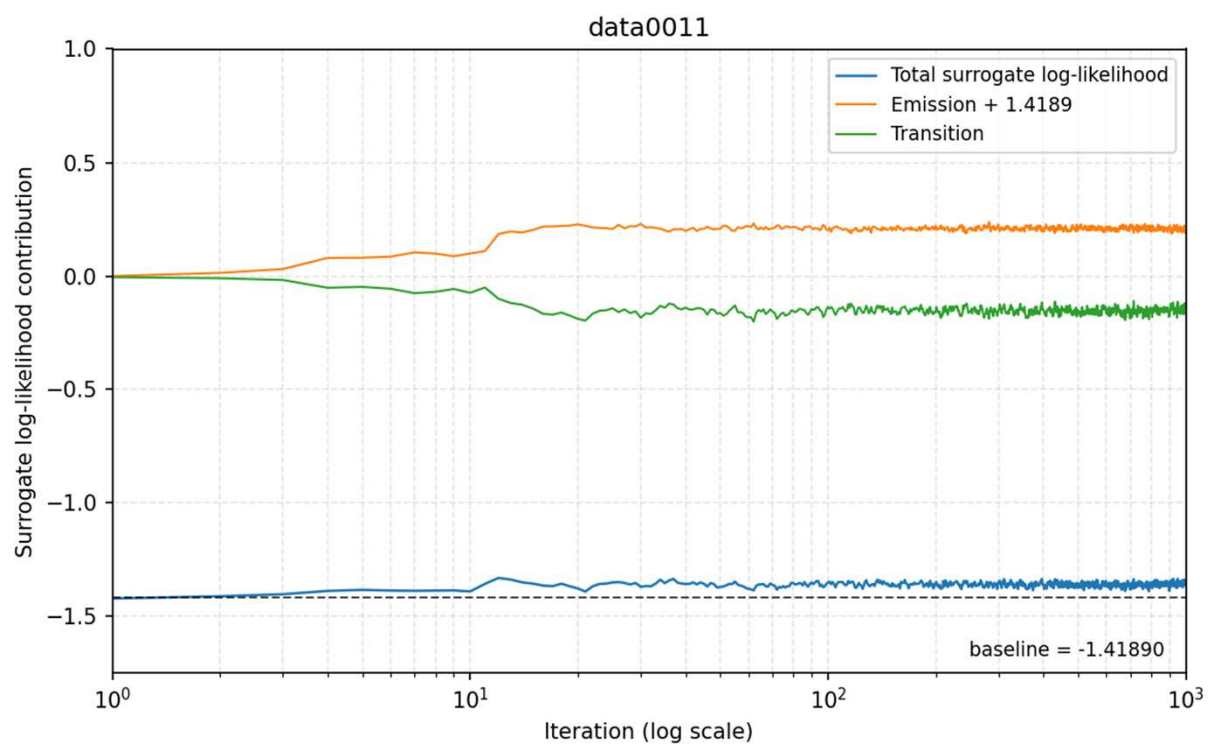
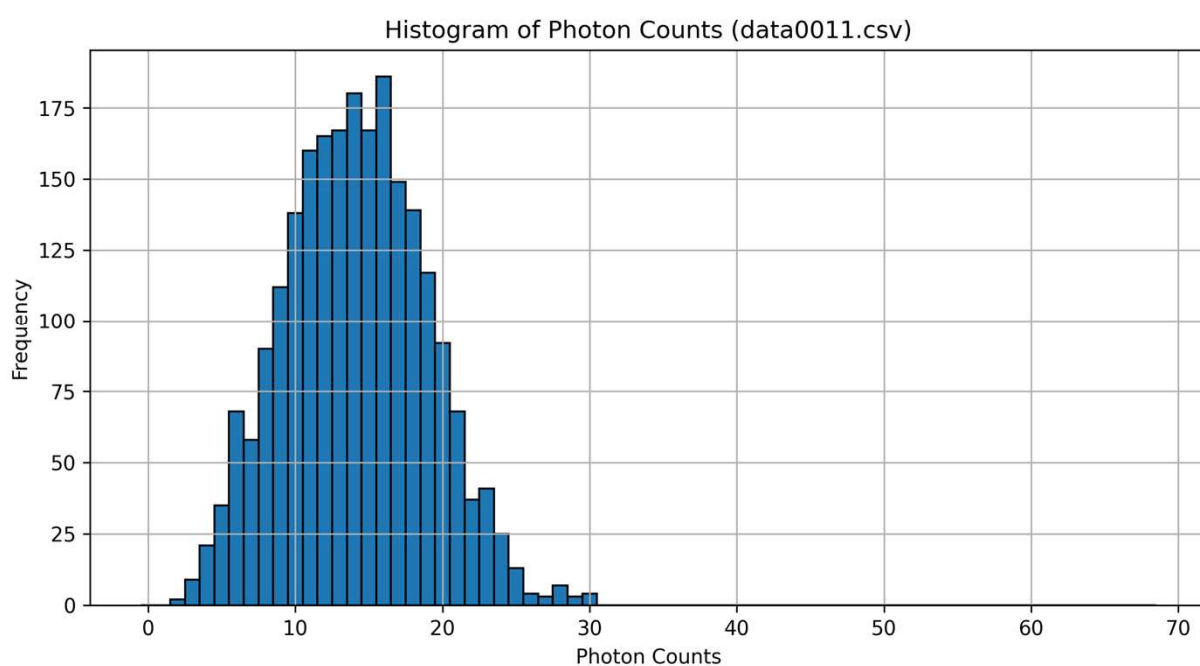
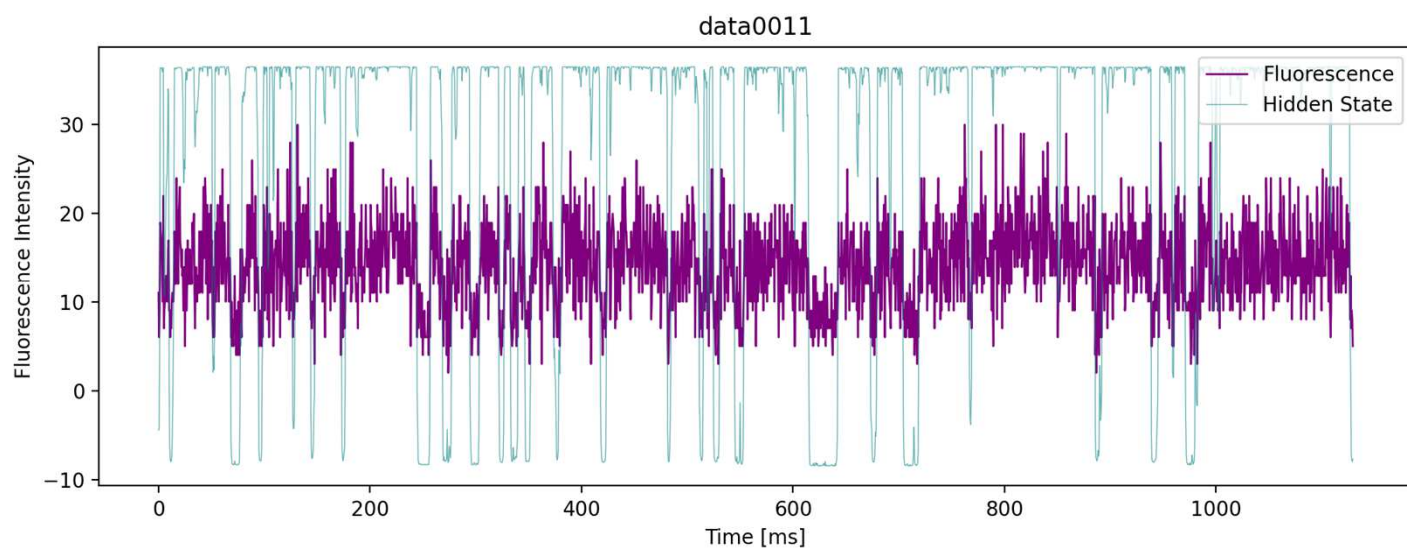


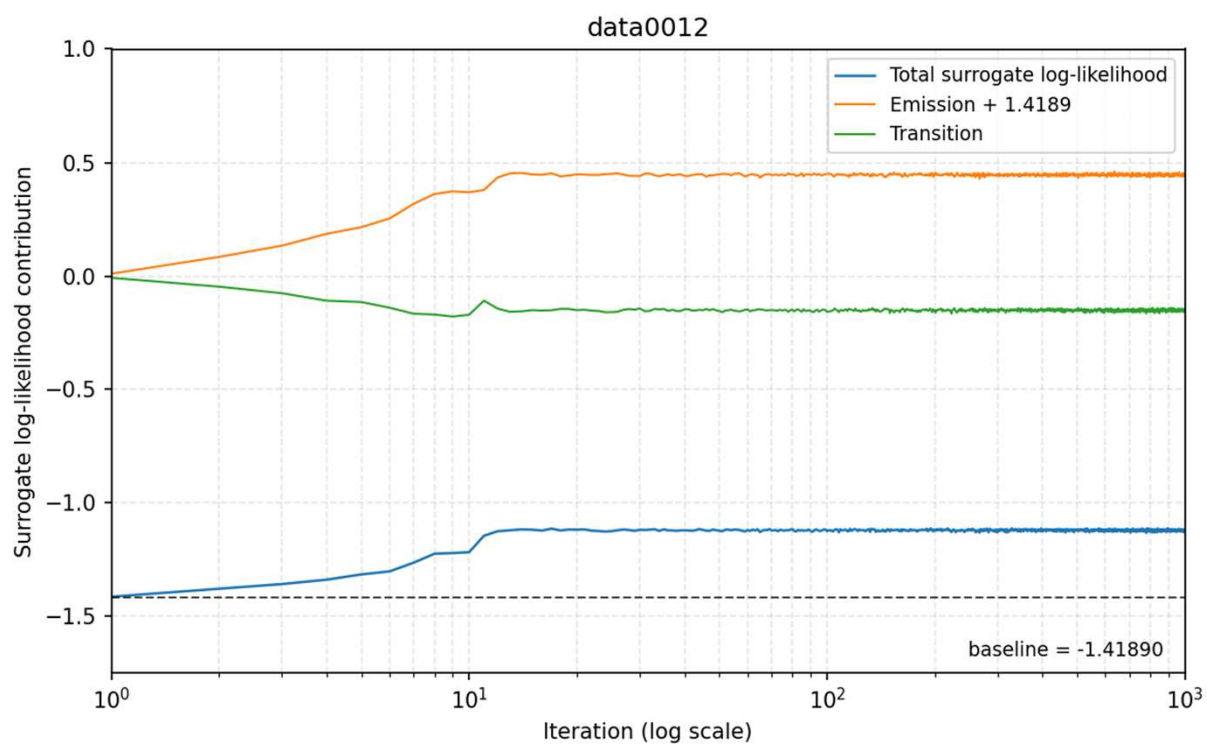
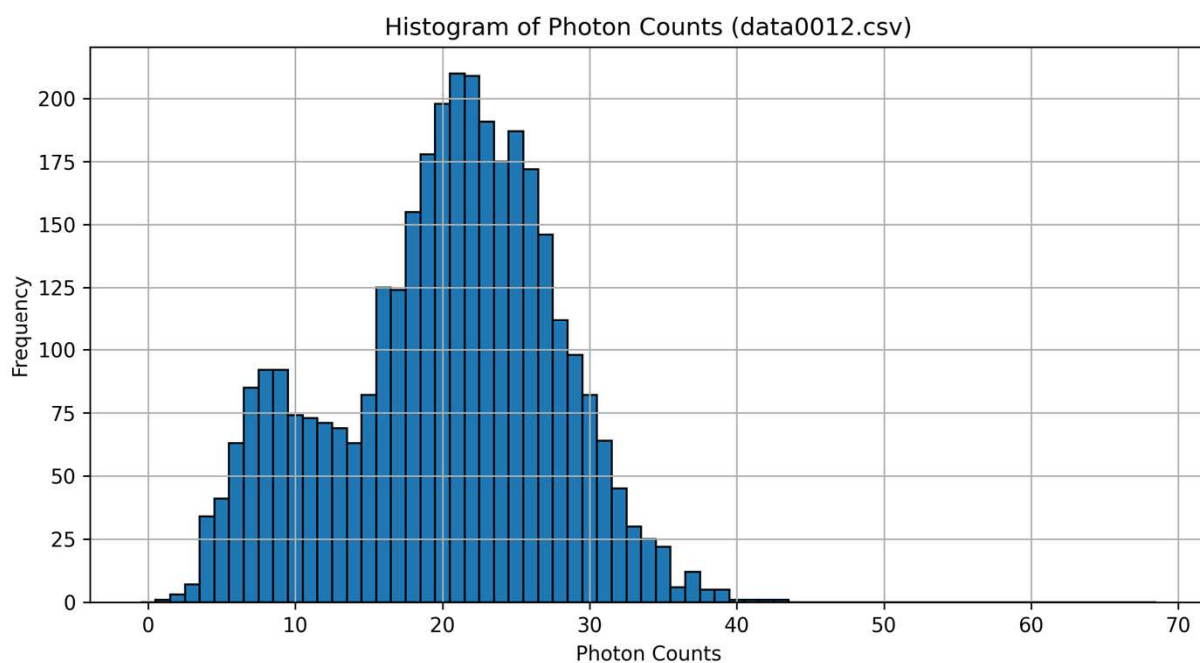
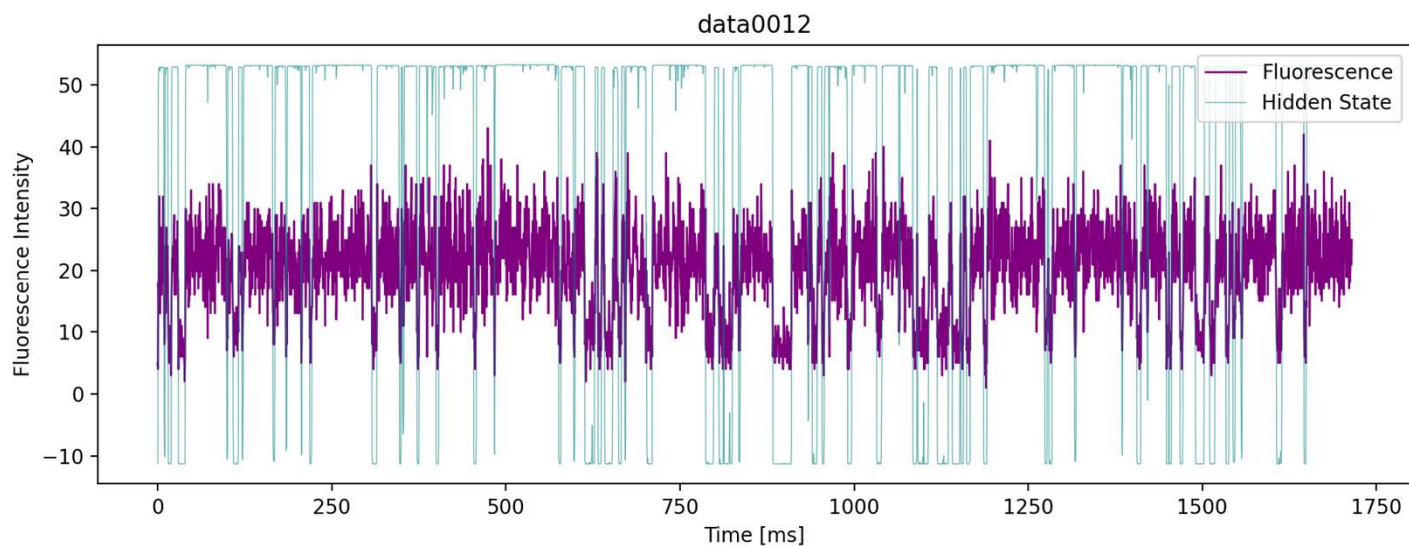


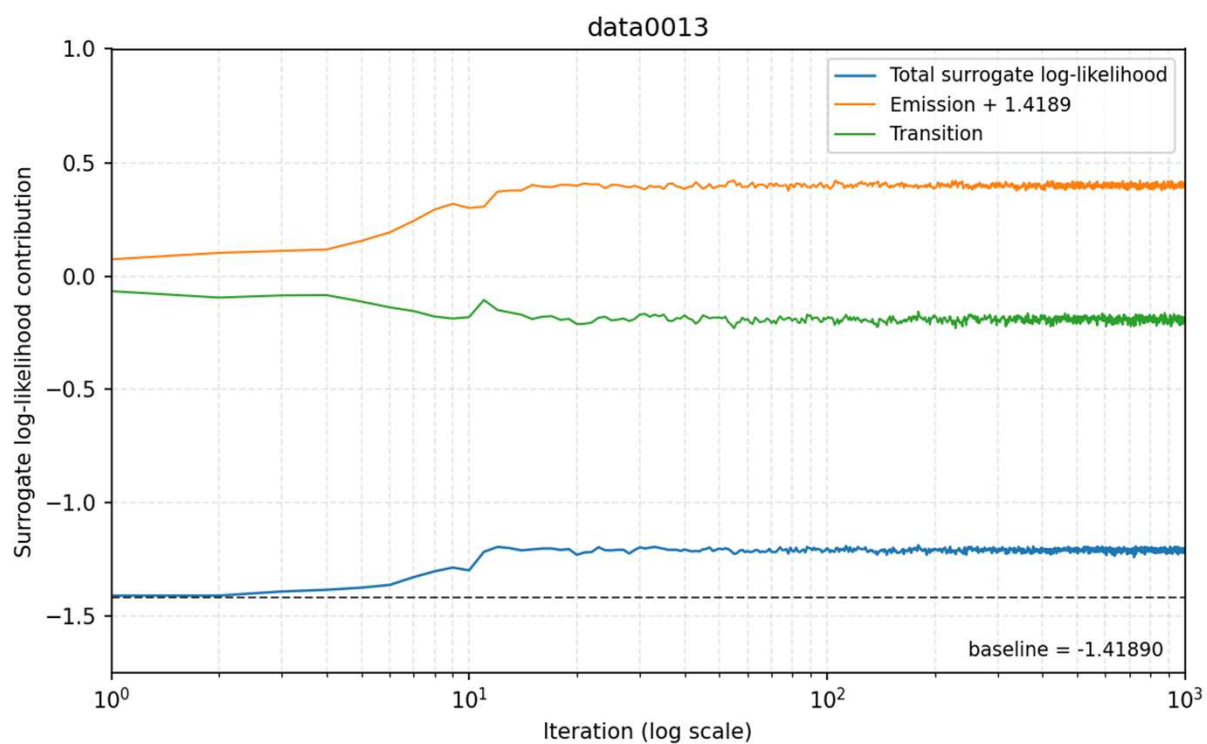
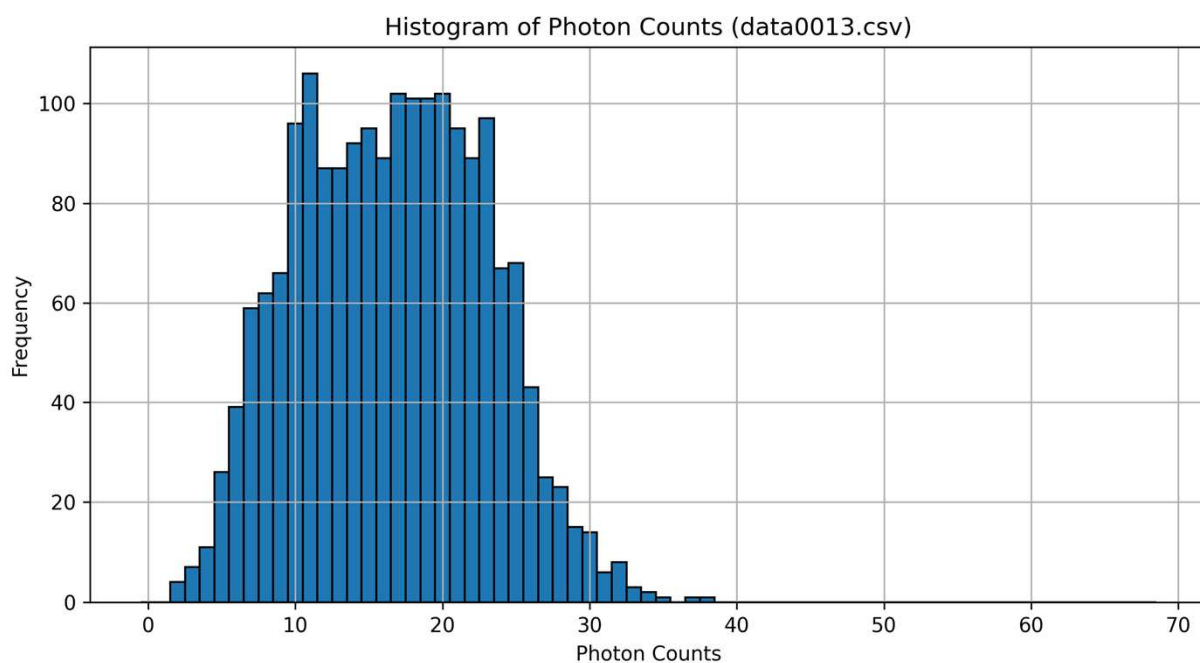
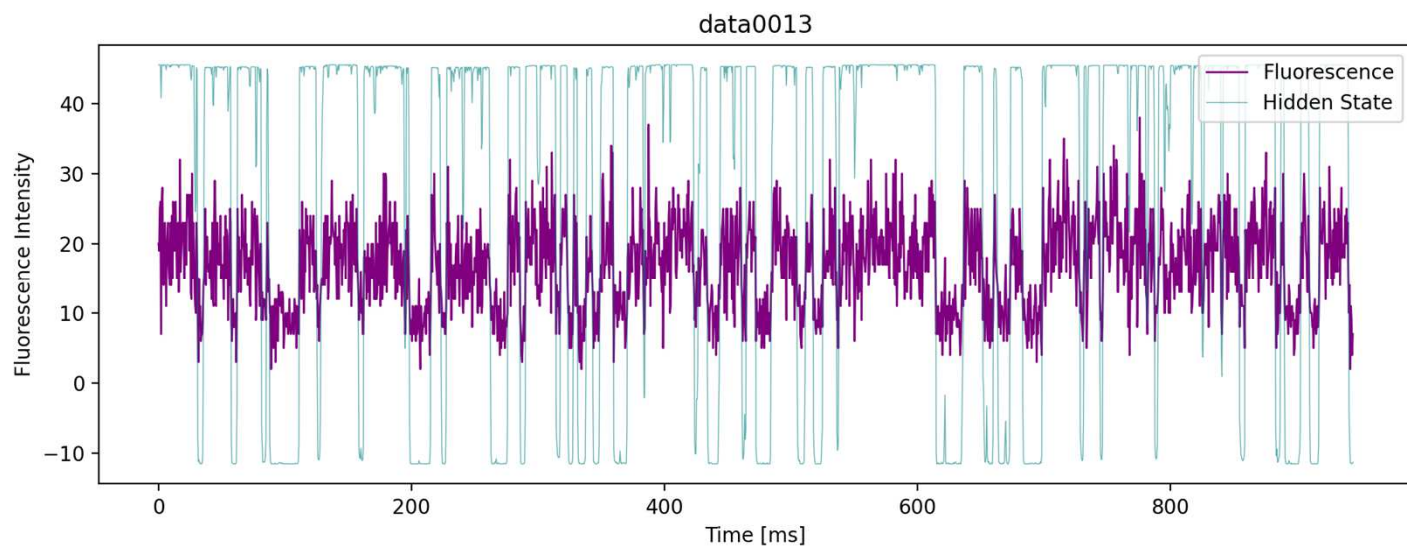




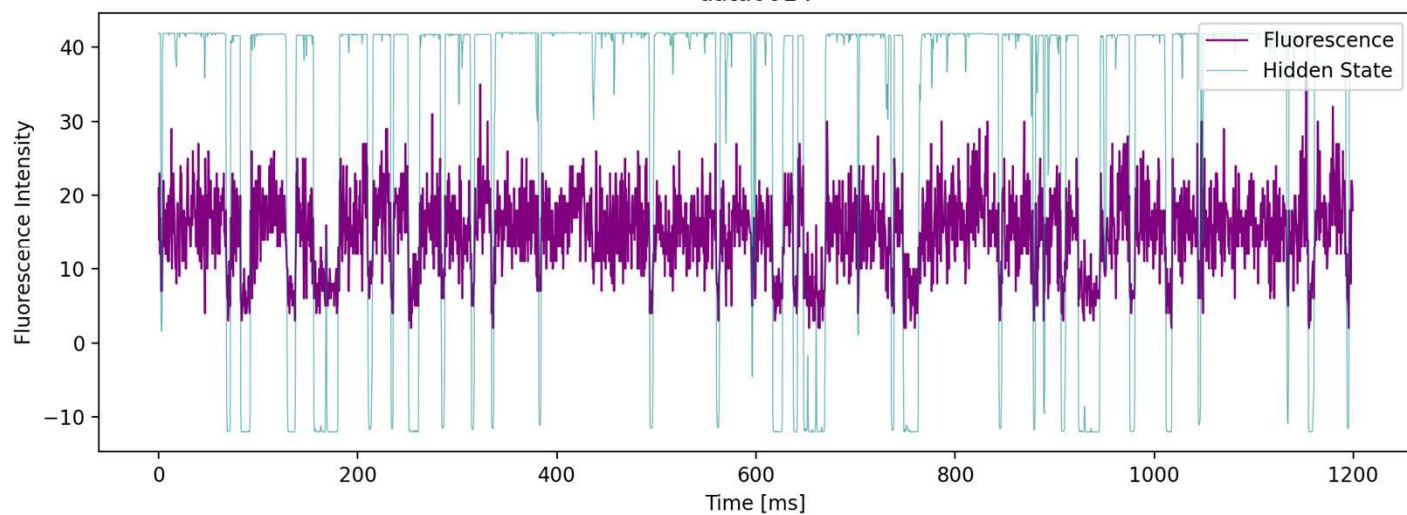




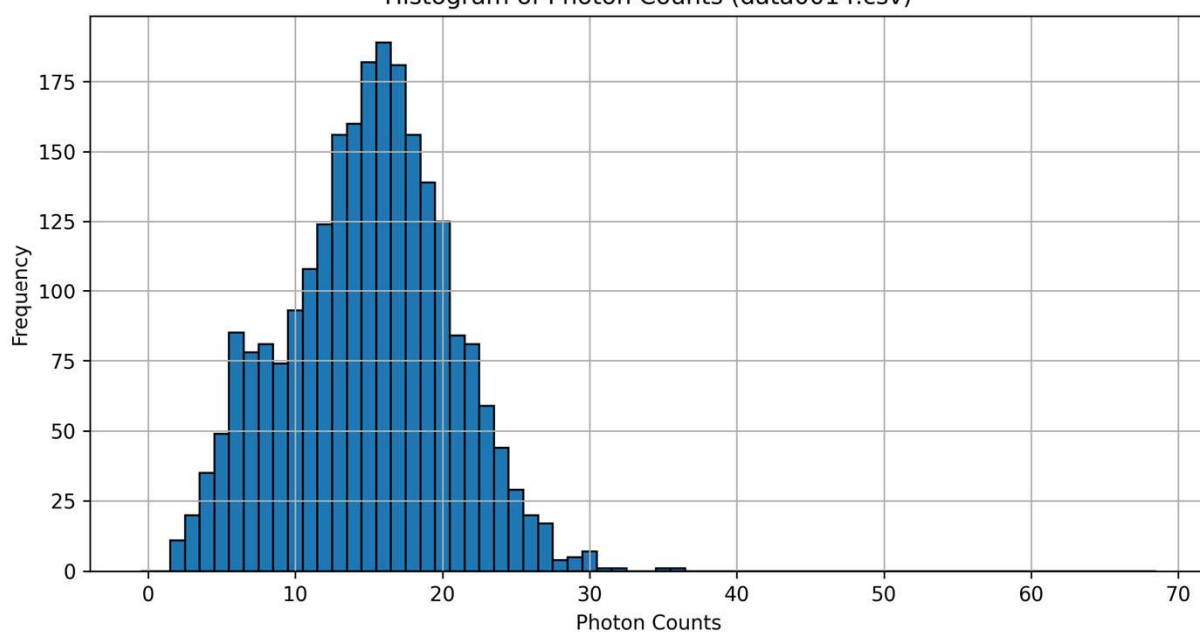




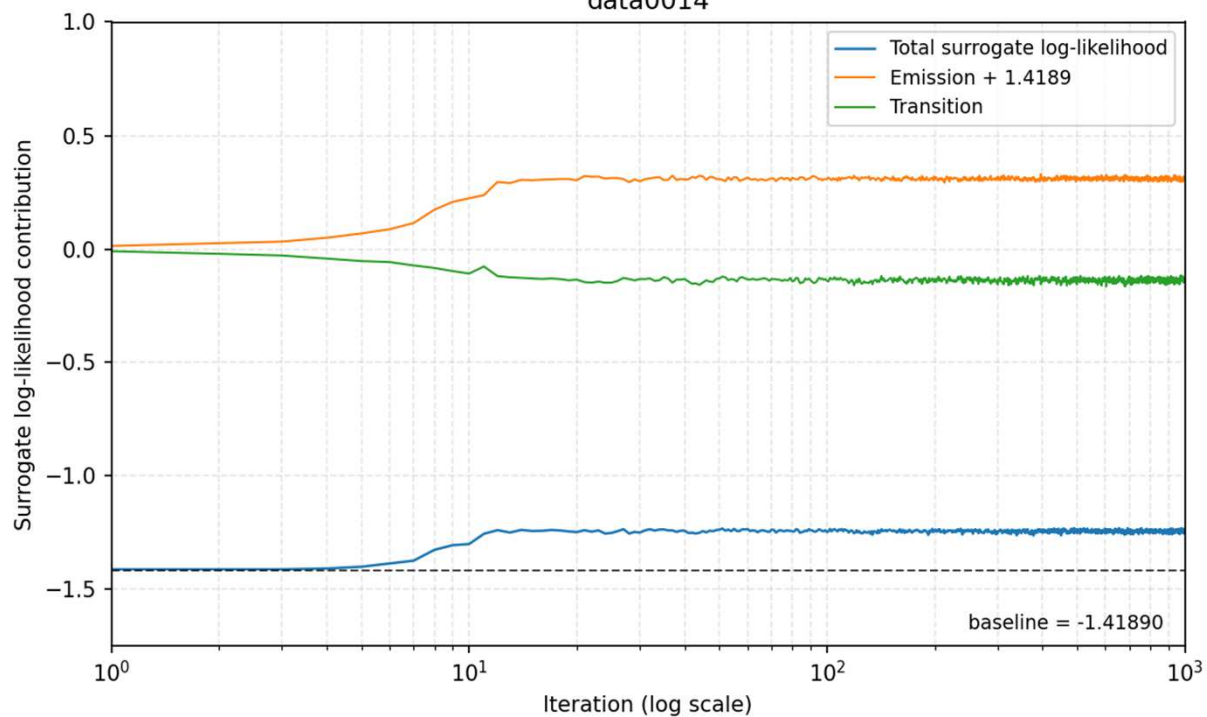
data0014

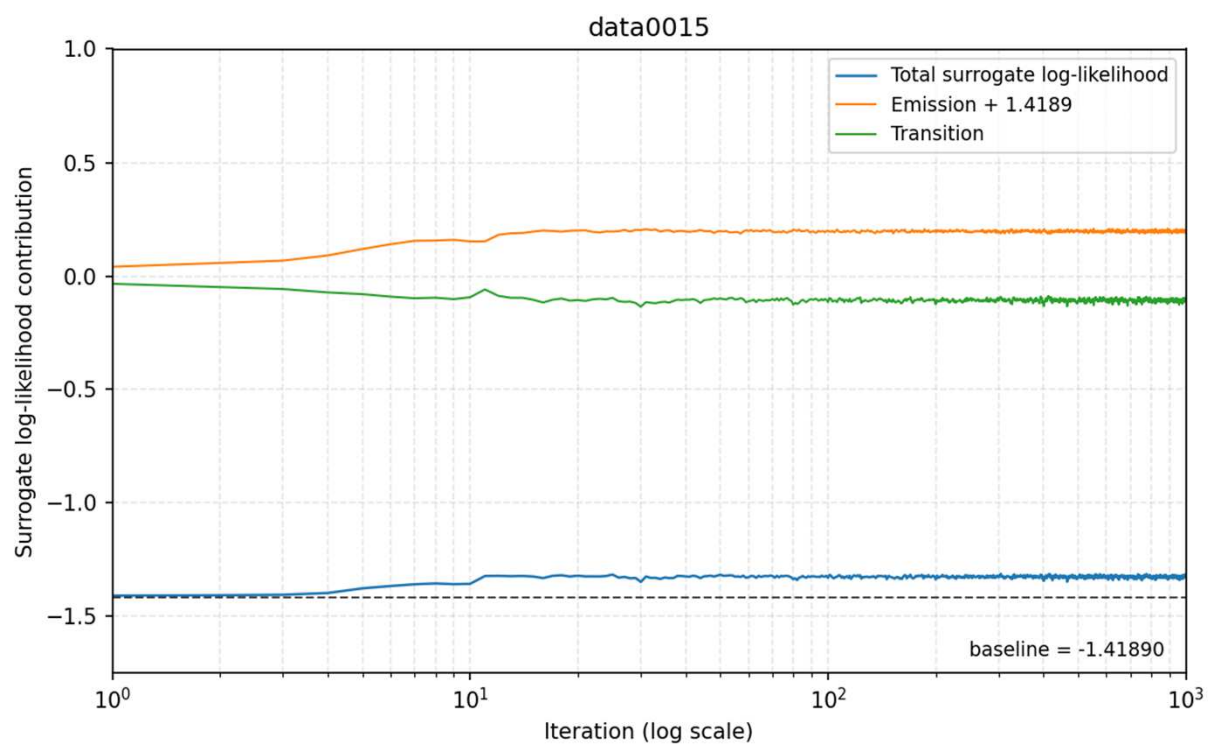
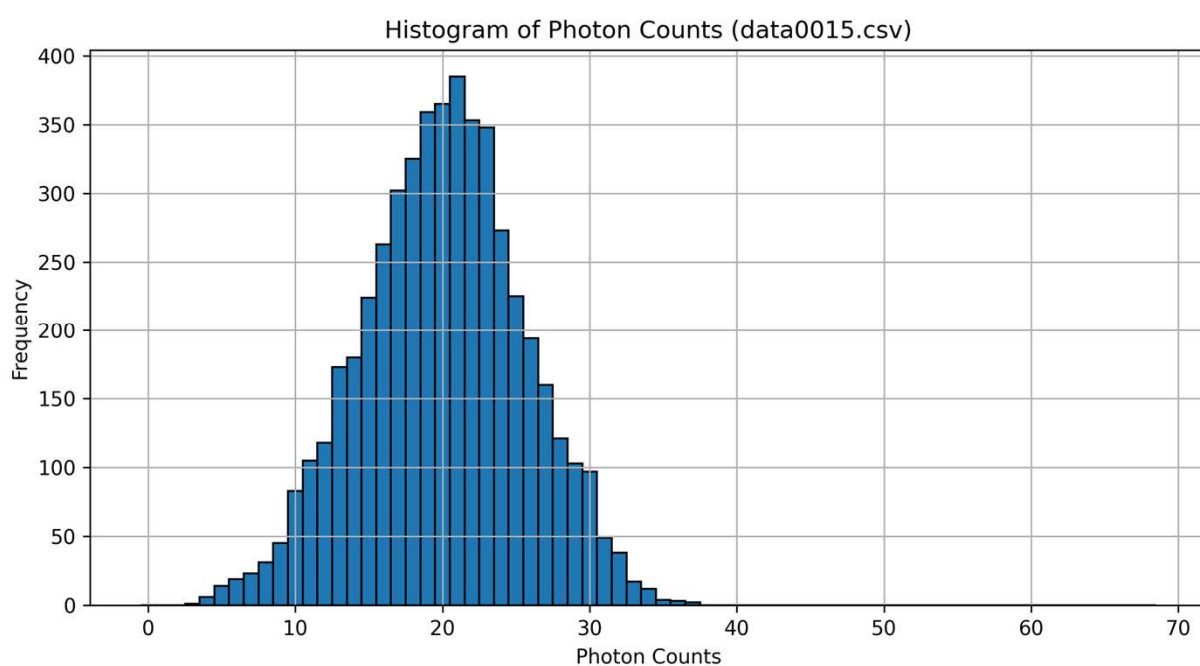
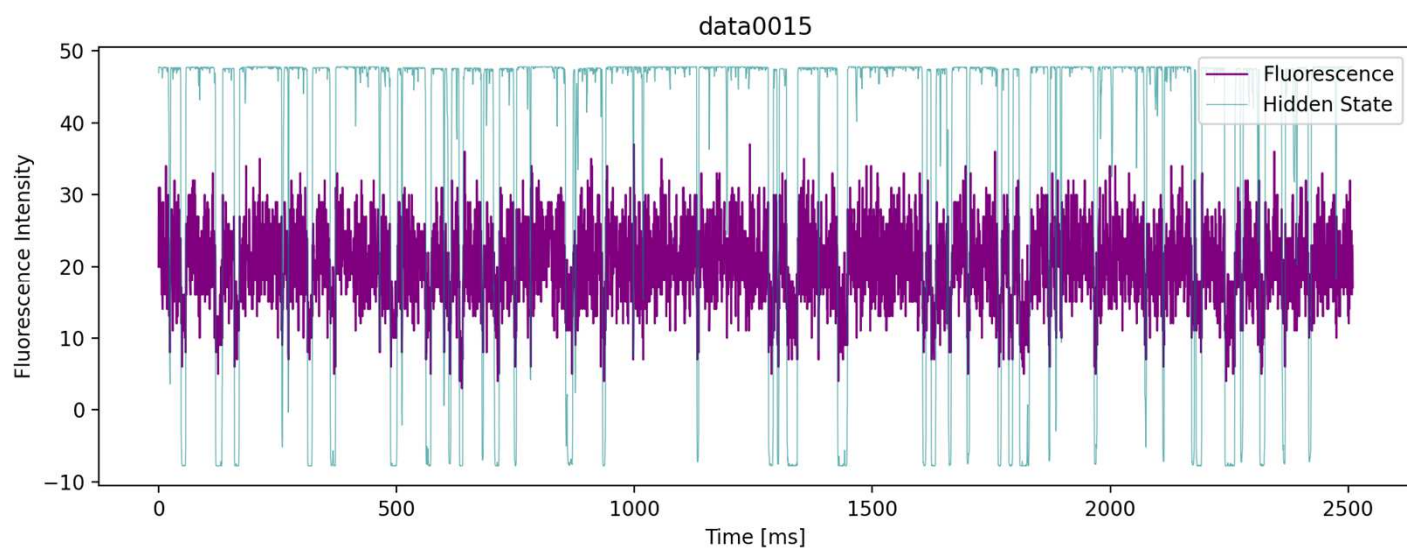


Histogram of Photon Counts (data0014.csv)

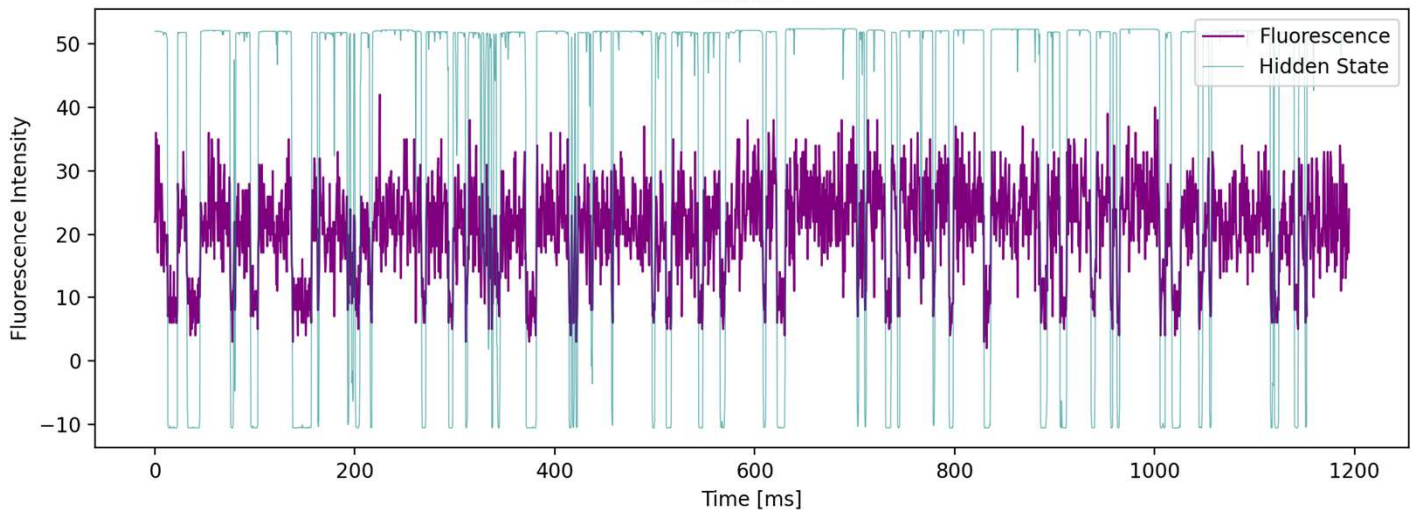


data0014

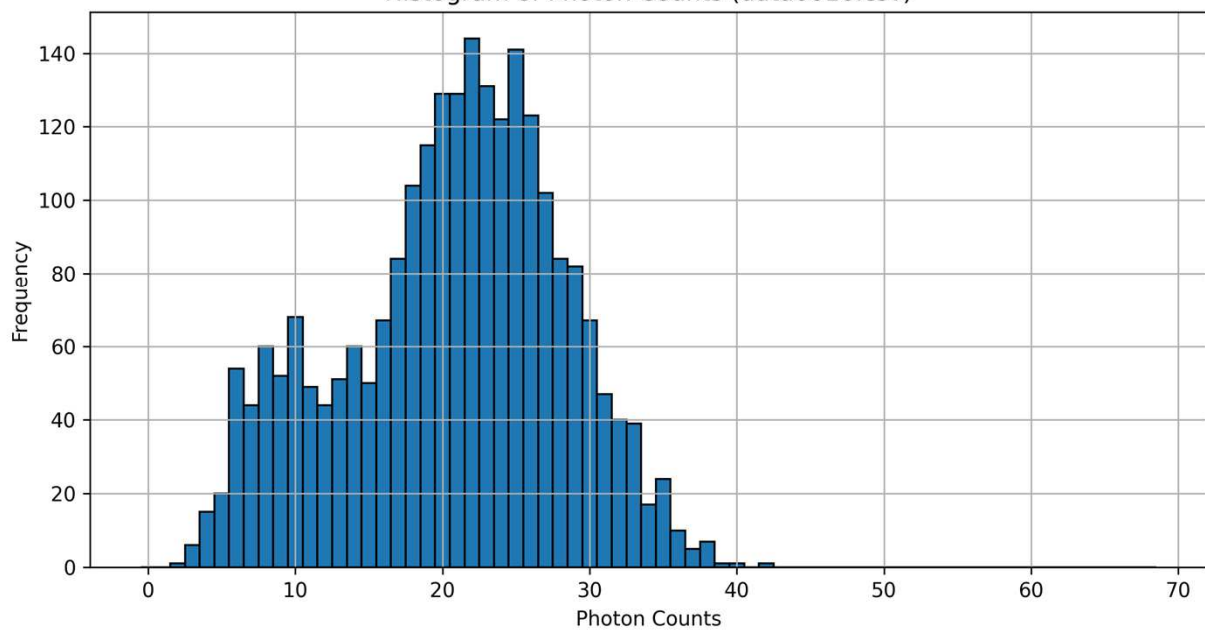




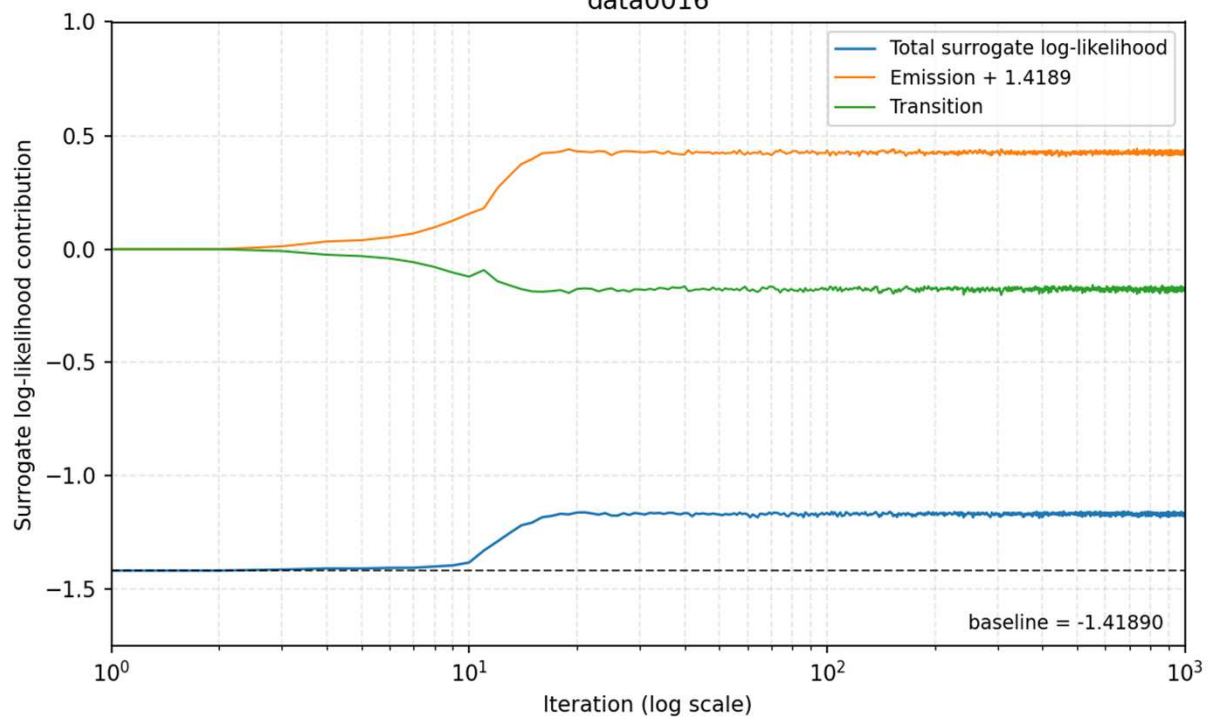
data0016



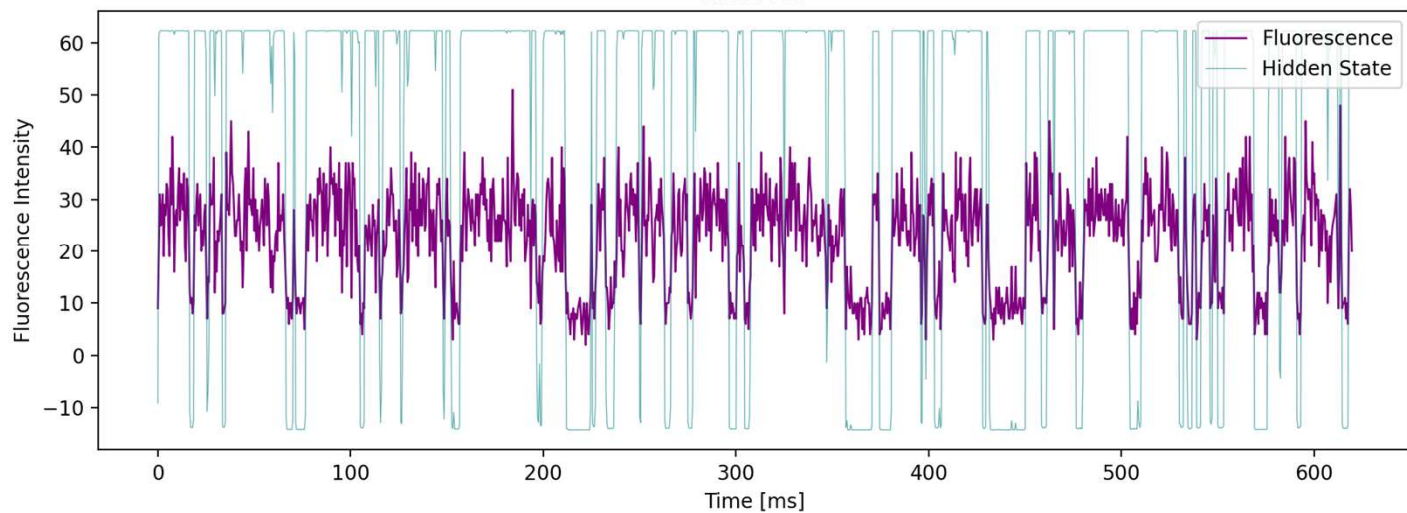
Histogram of Photon Counts (data0016.csv)



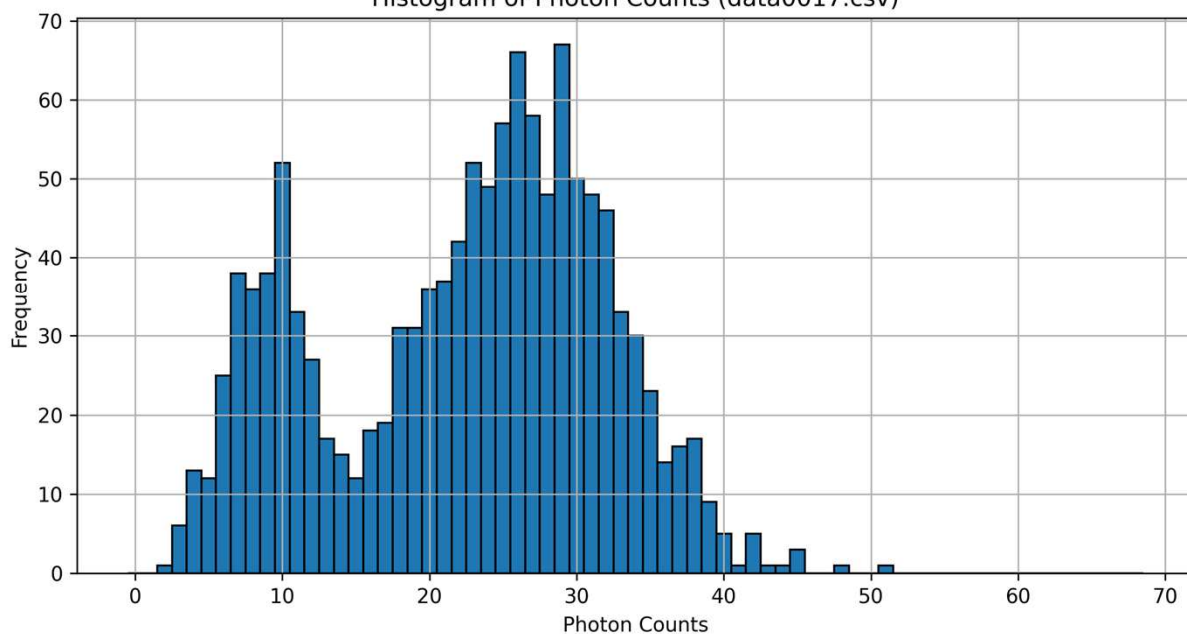
data0016



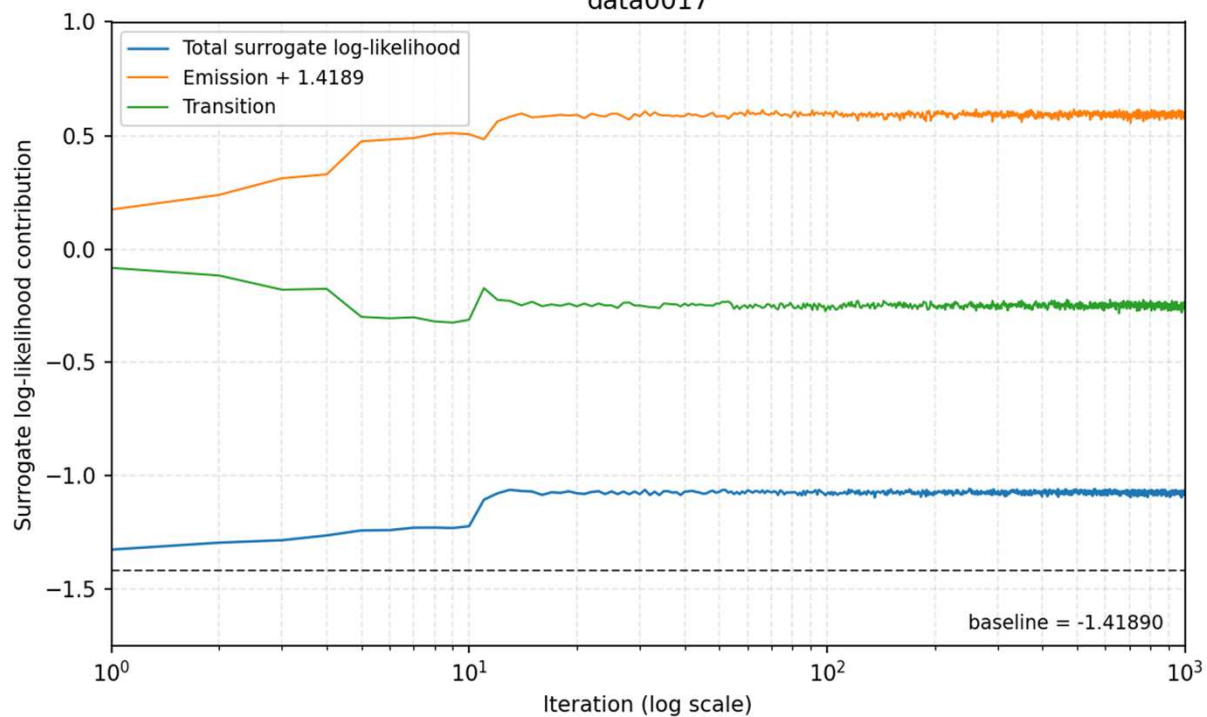
data0017

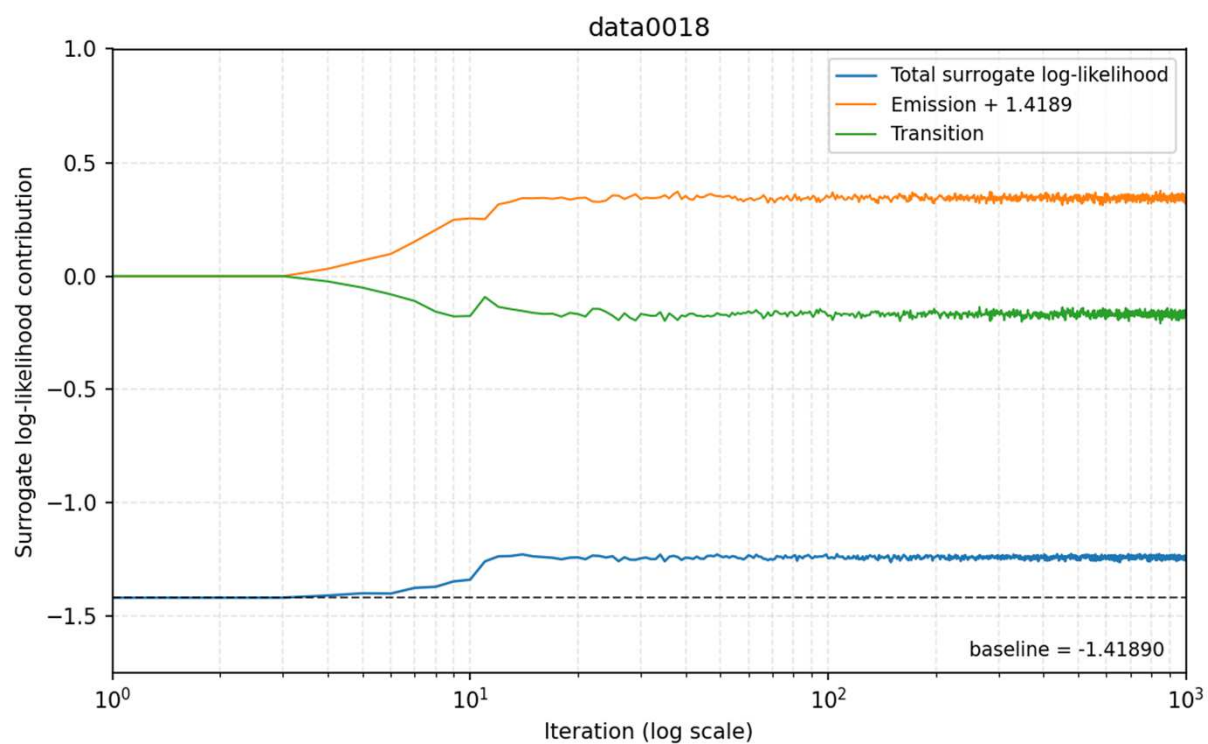
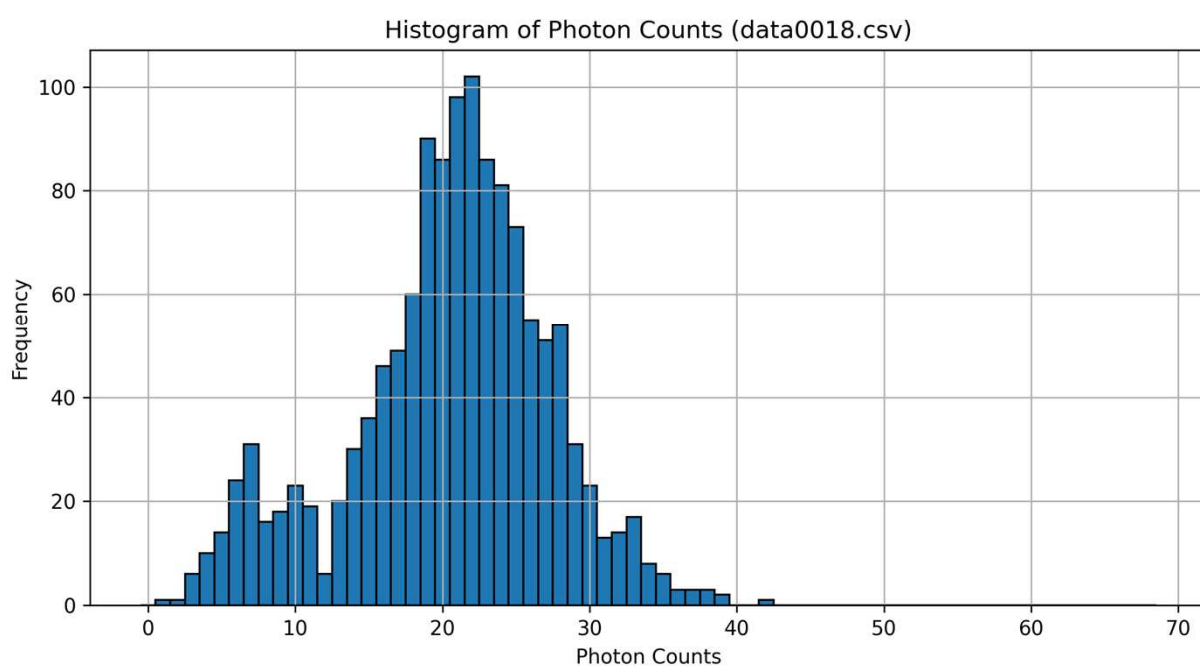
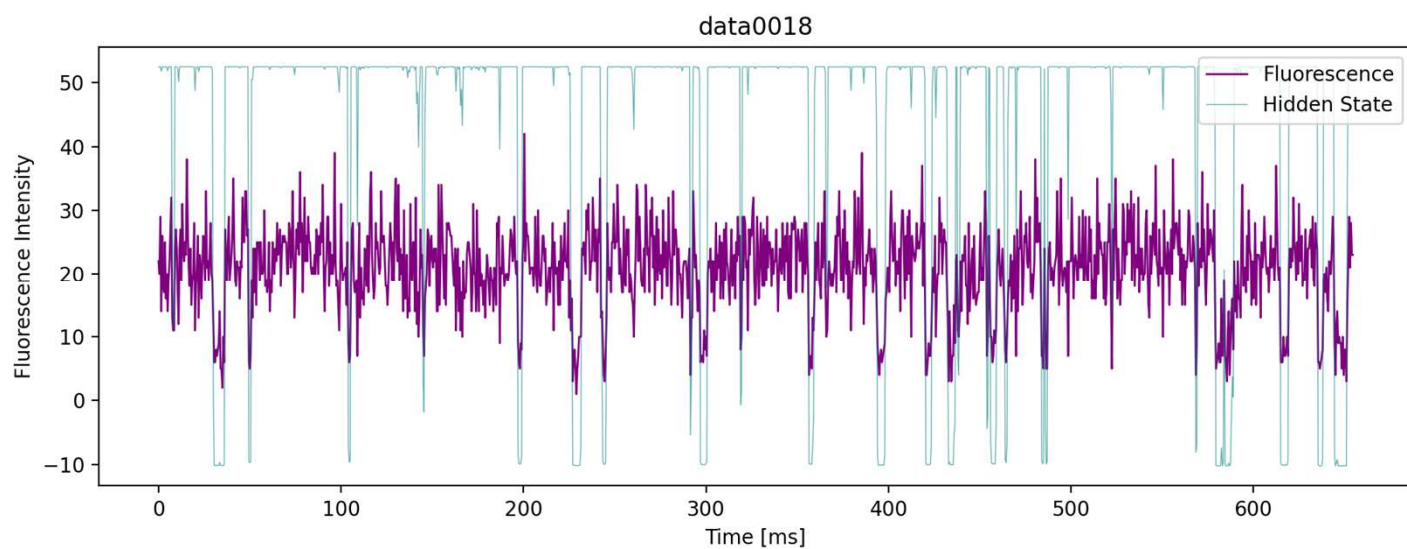


Histogram of Photon Counts (data0017.csv)

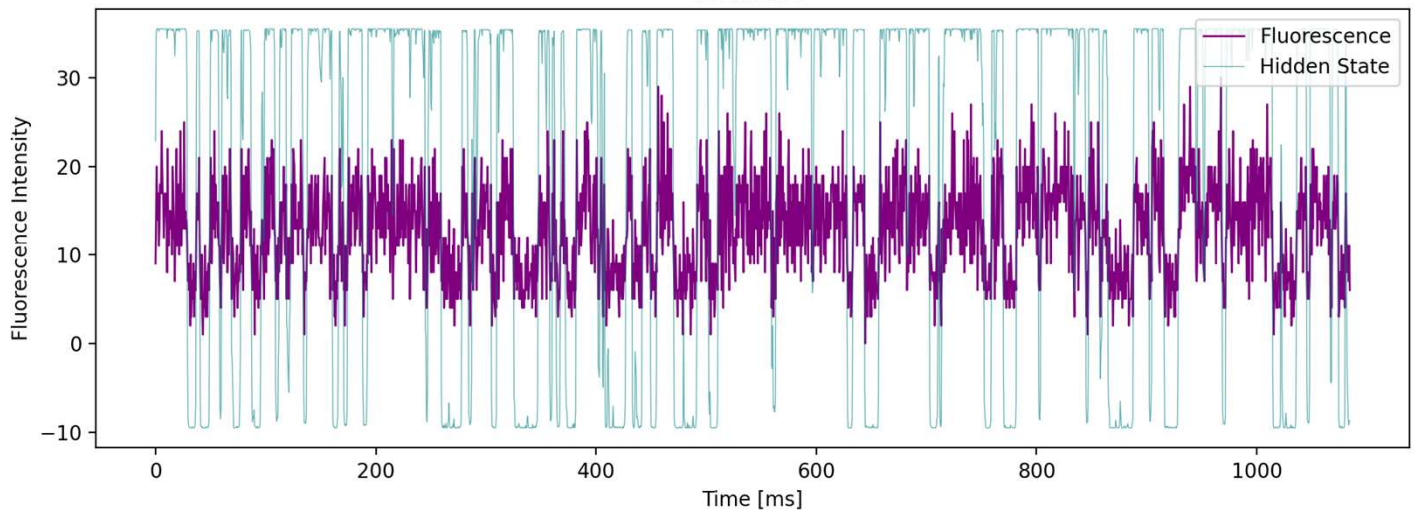


data0017

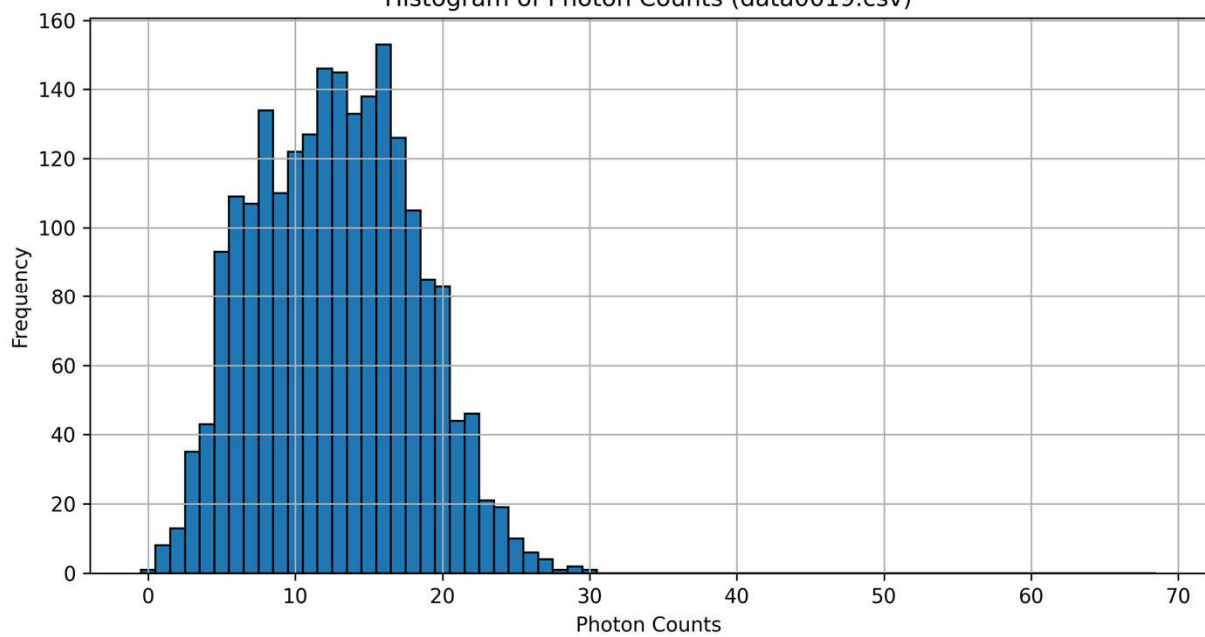




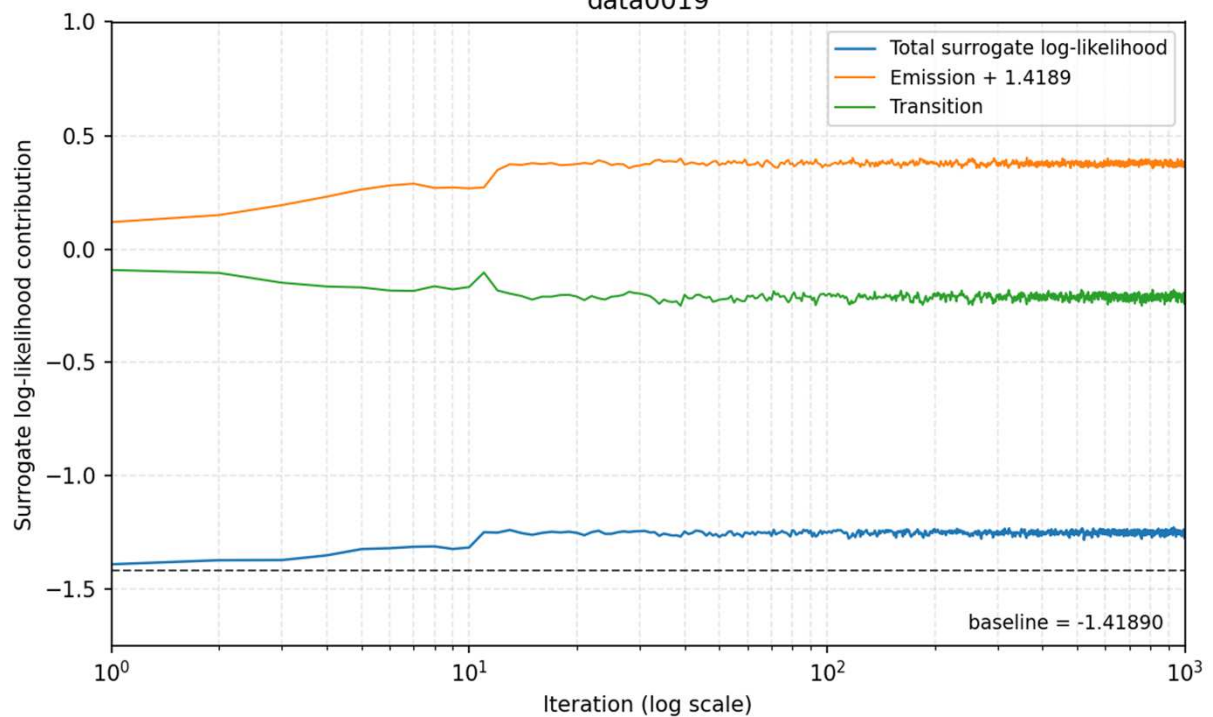
data0019

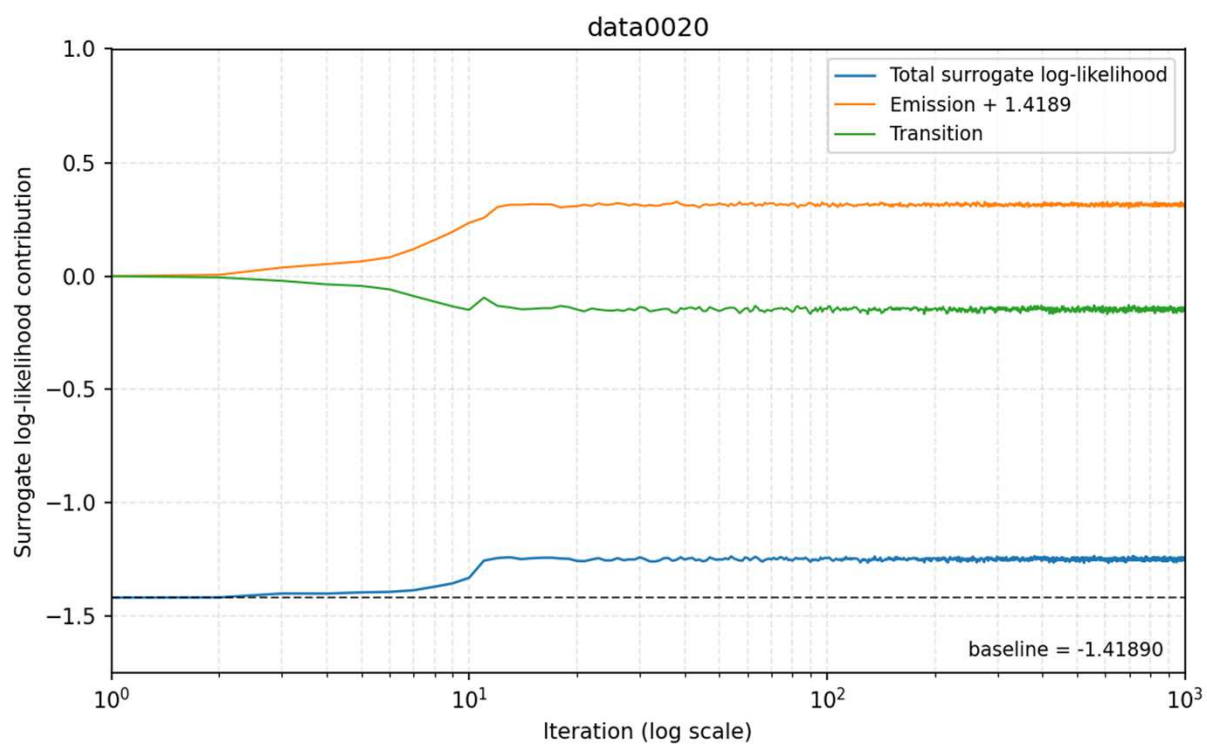
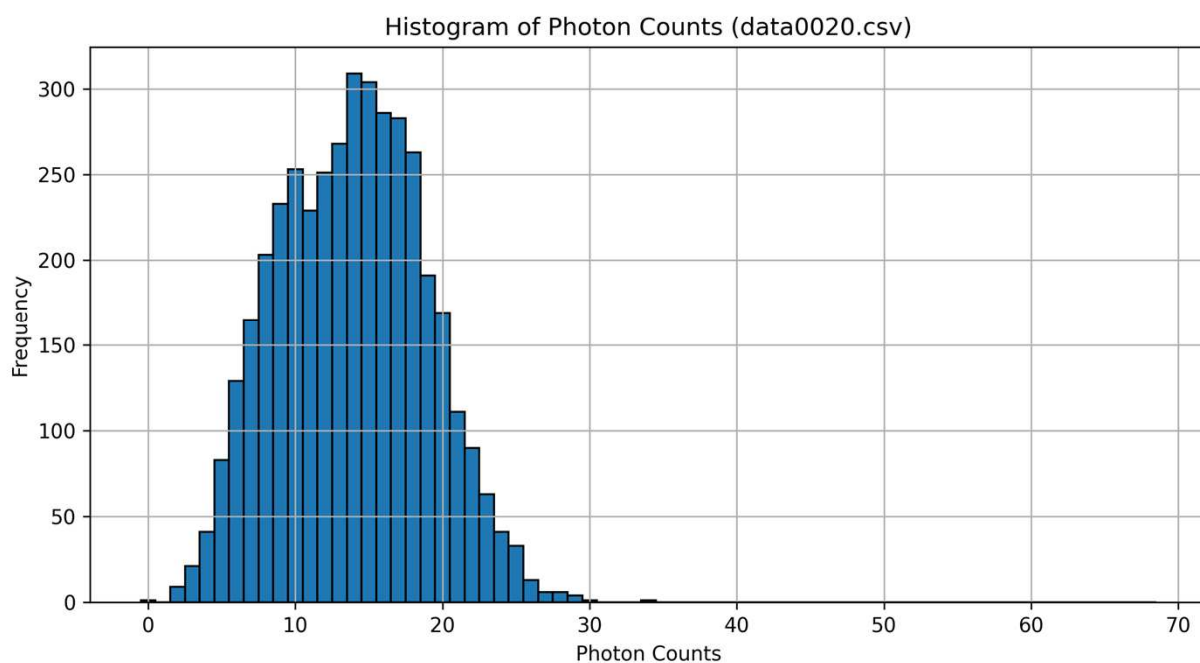
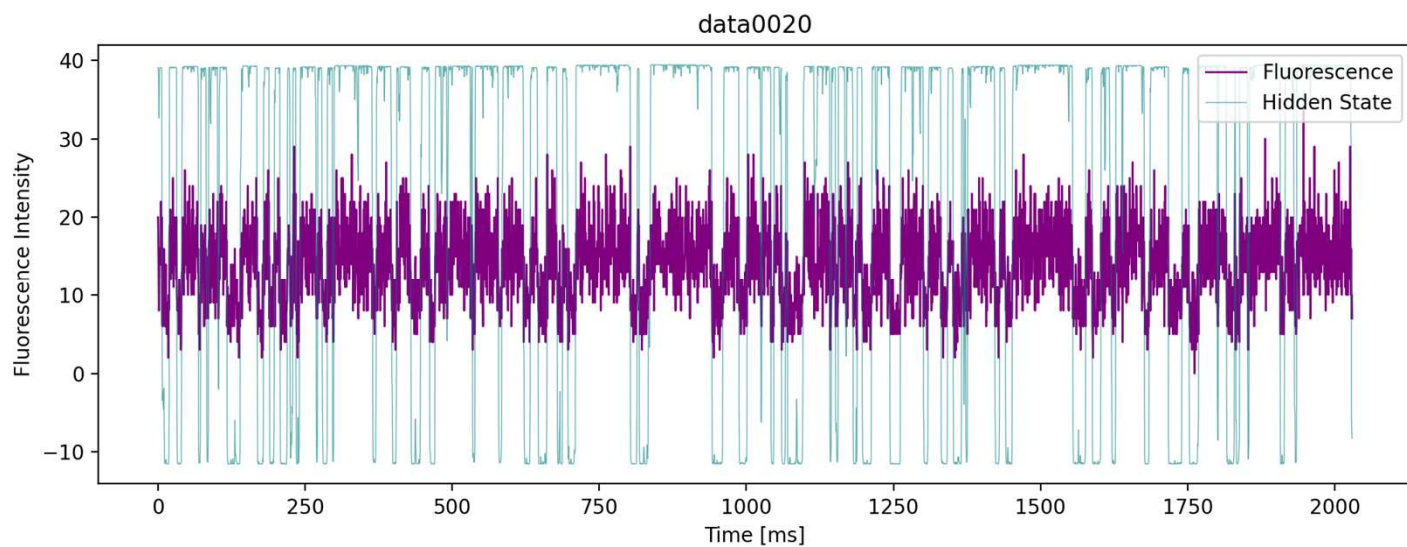


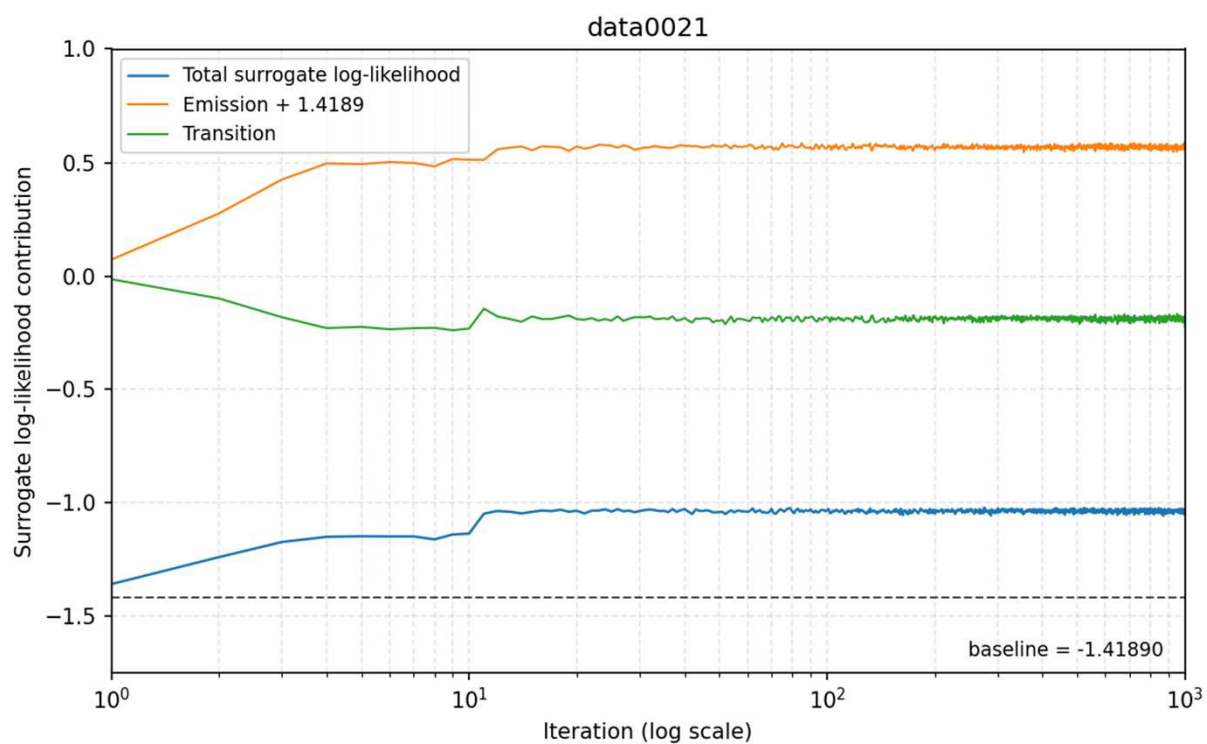
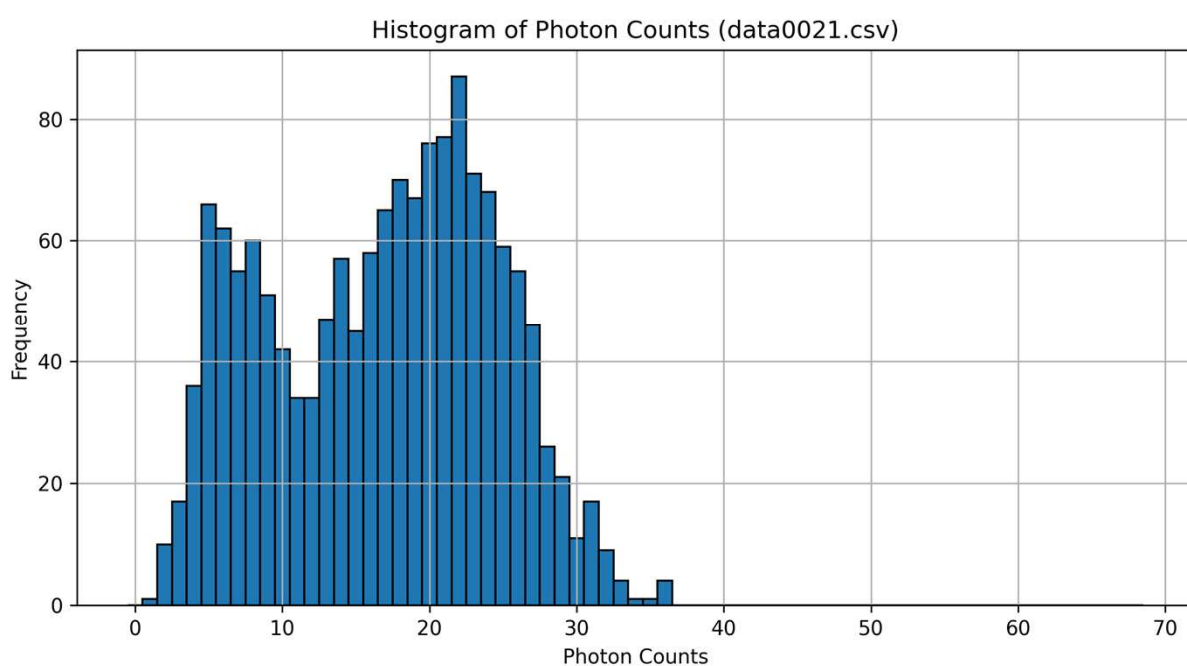
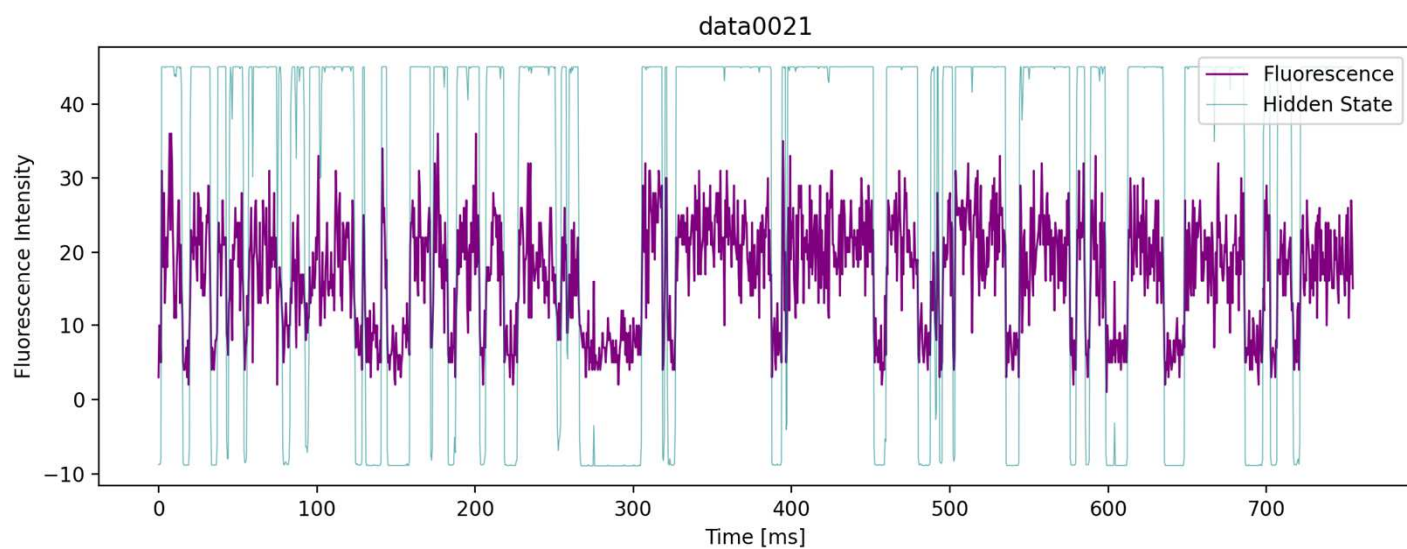
Histogram of Photon Counts (data0019.csv)

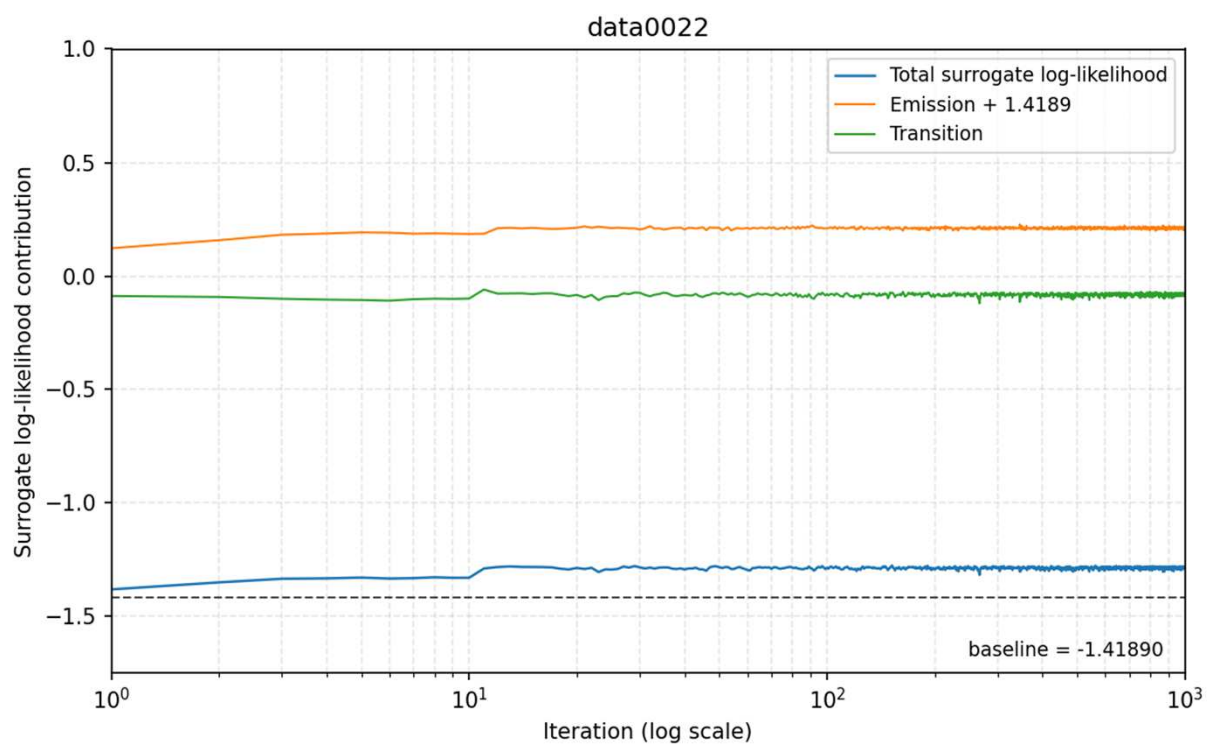
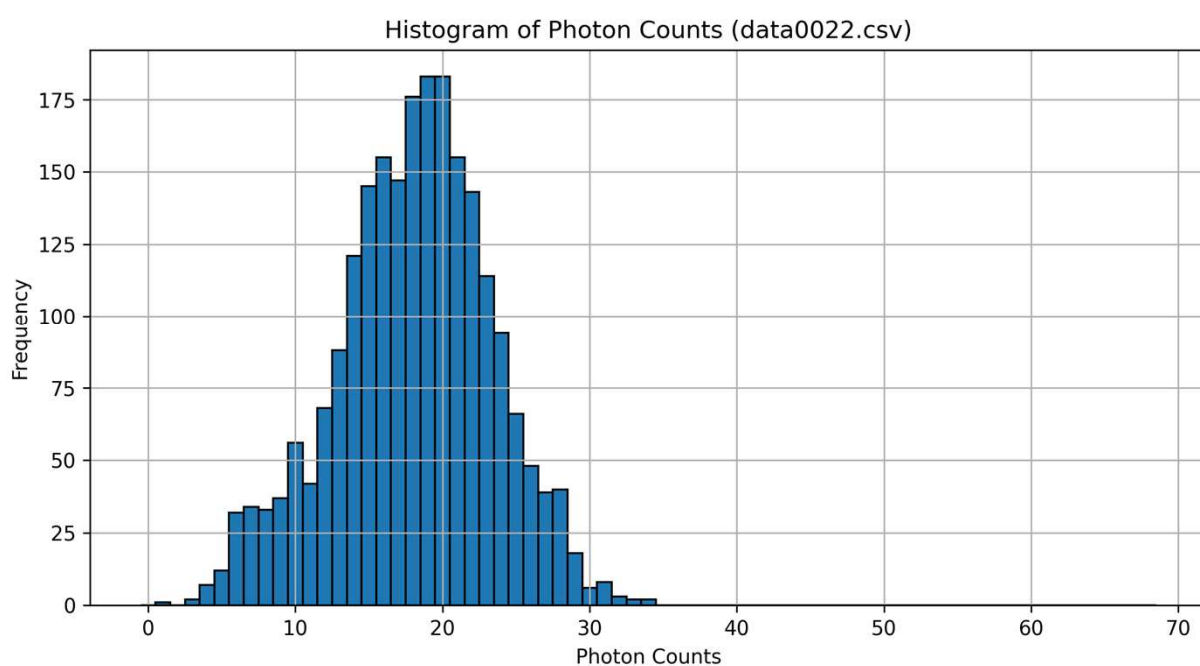
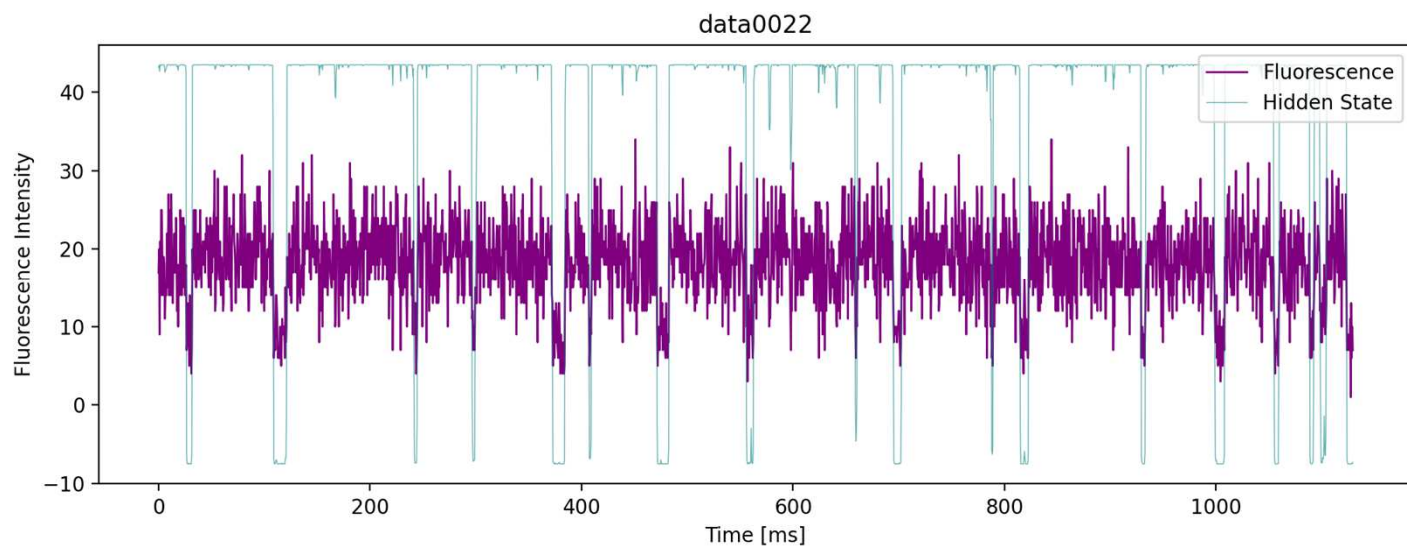


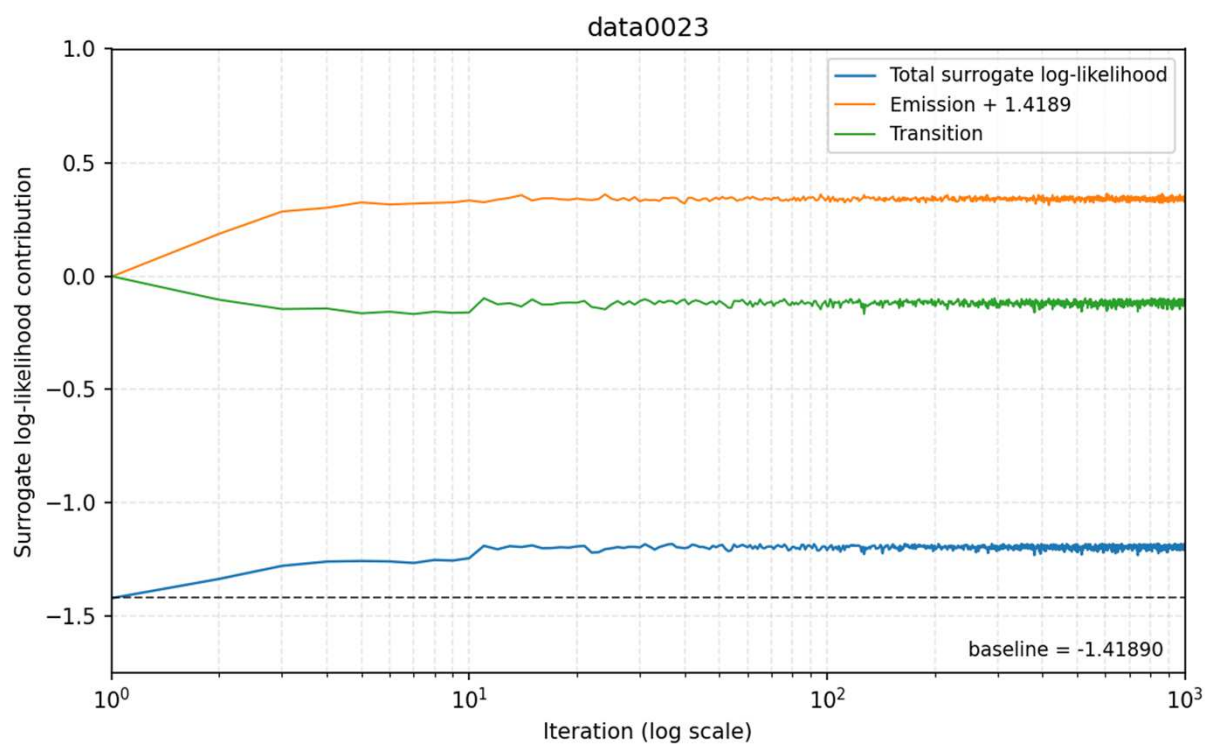
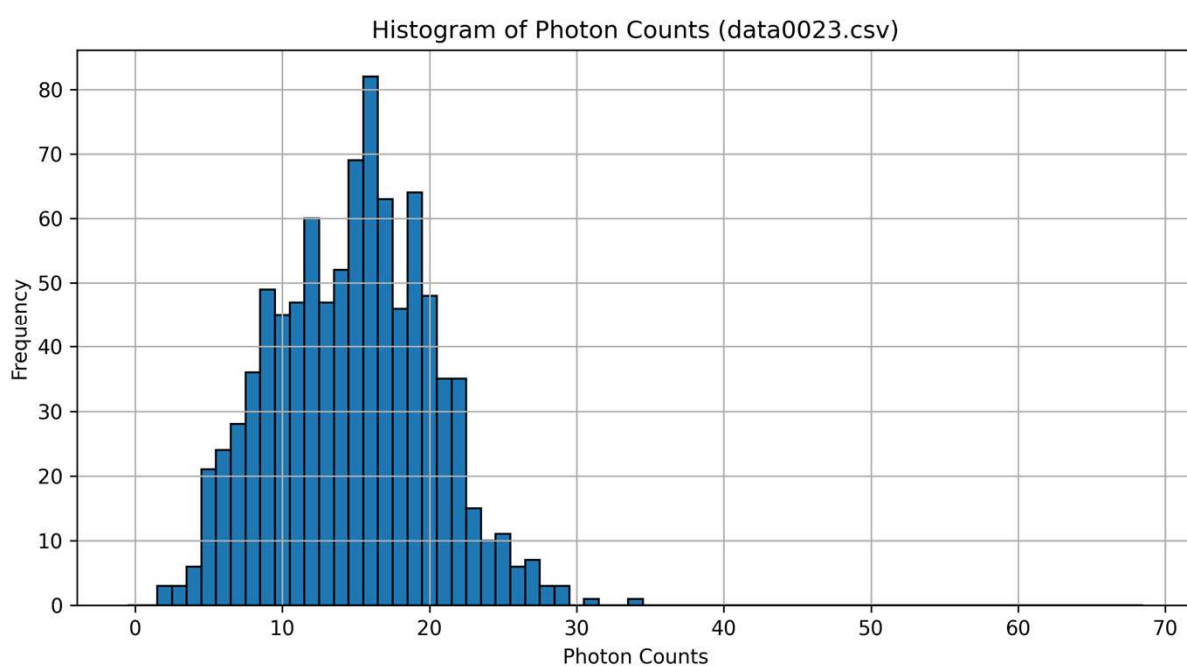
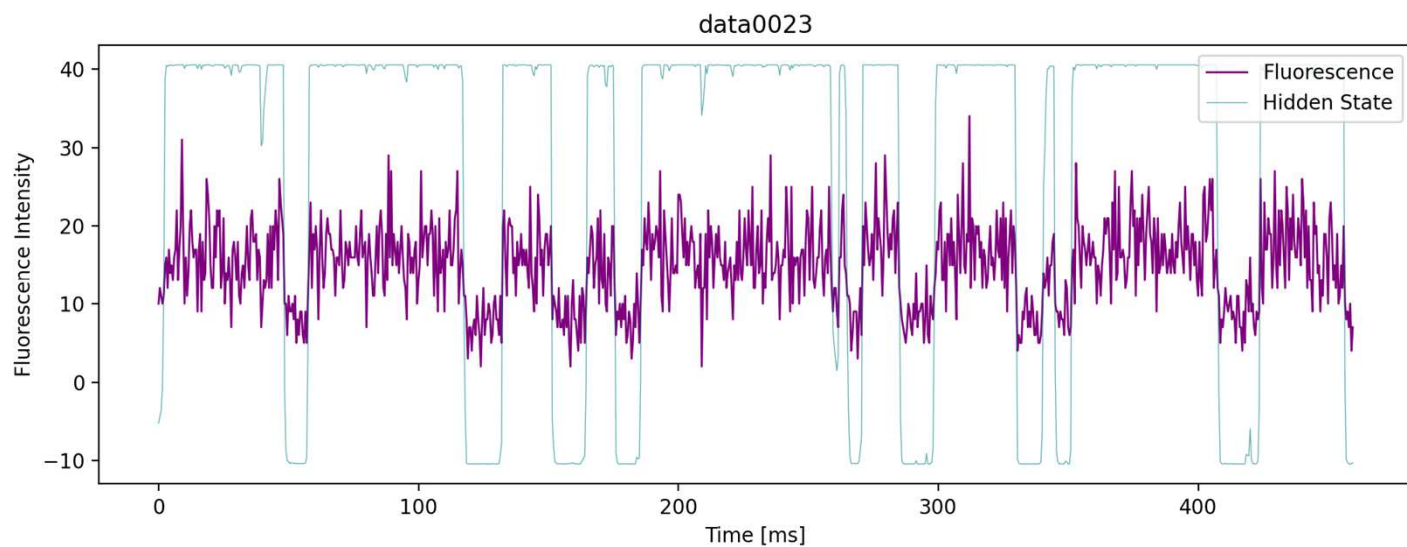
data0019



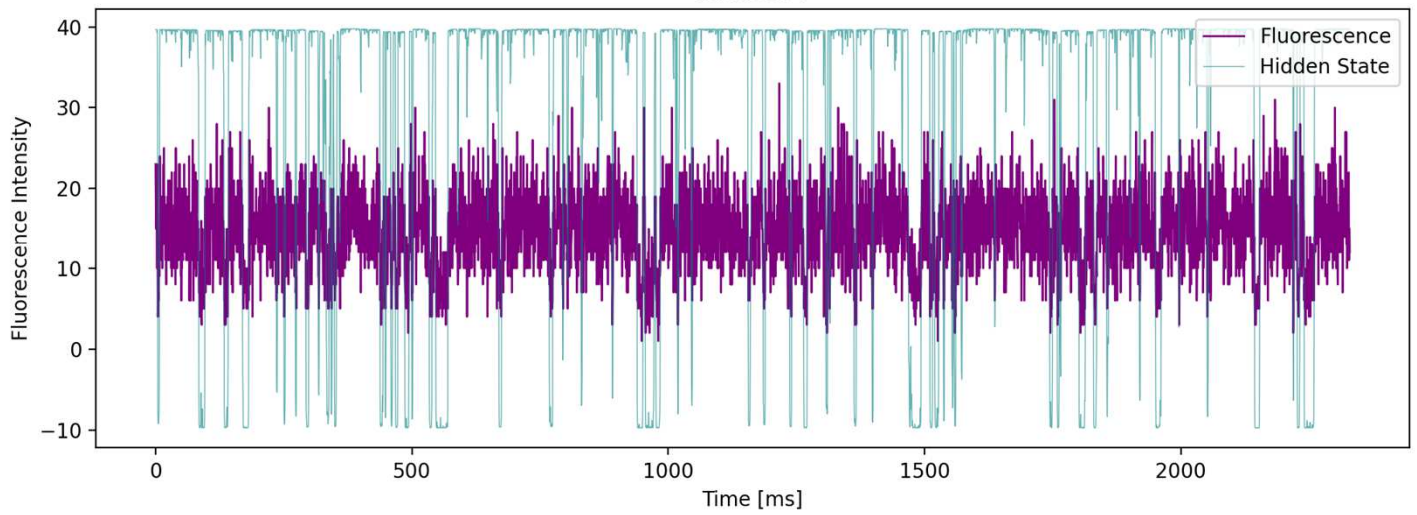




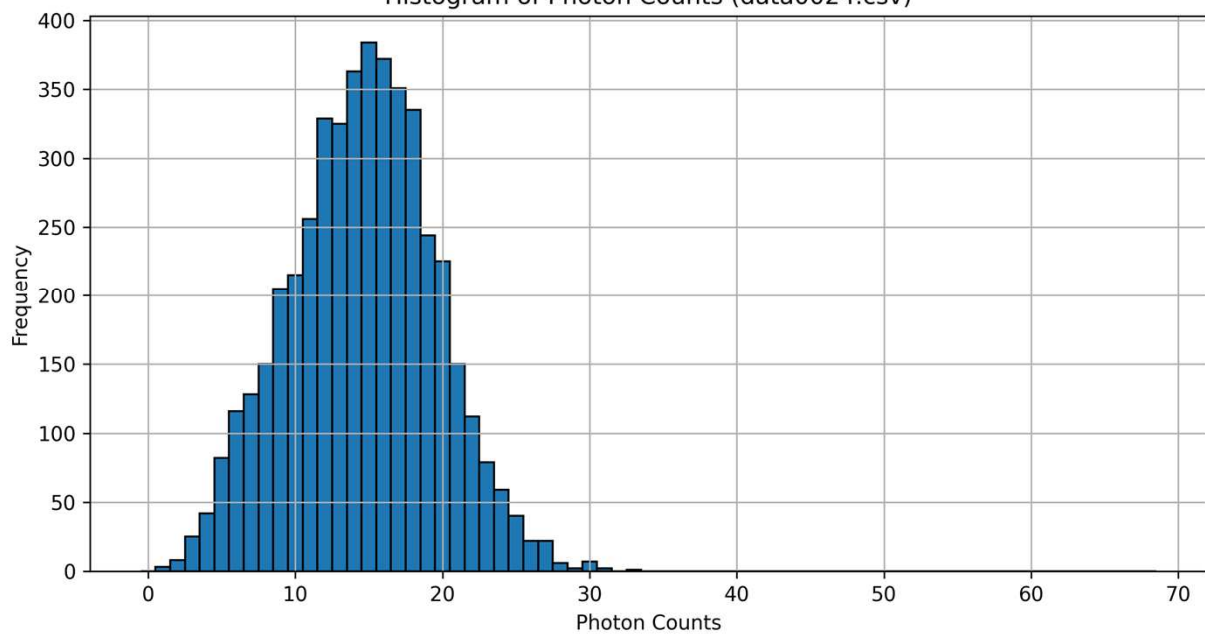




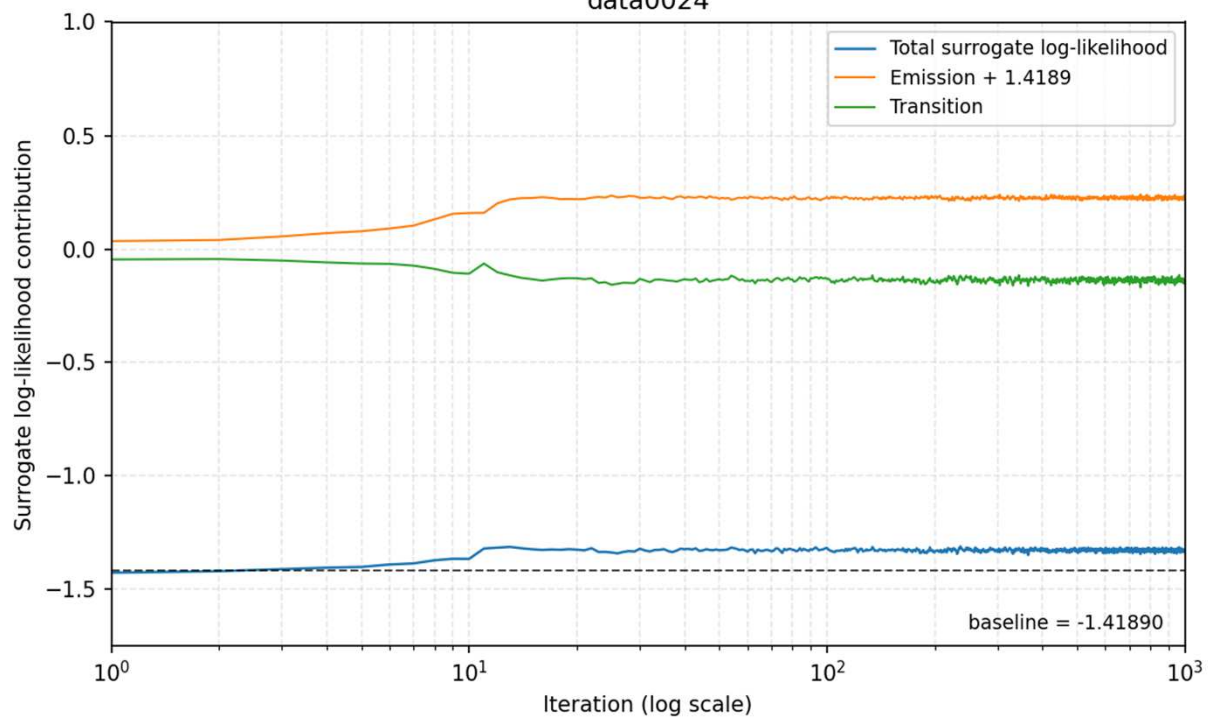
data0024



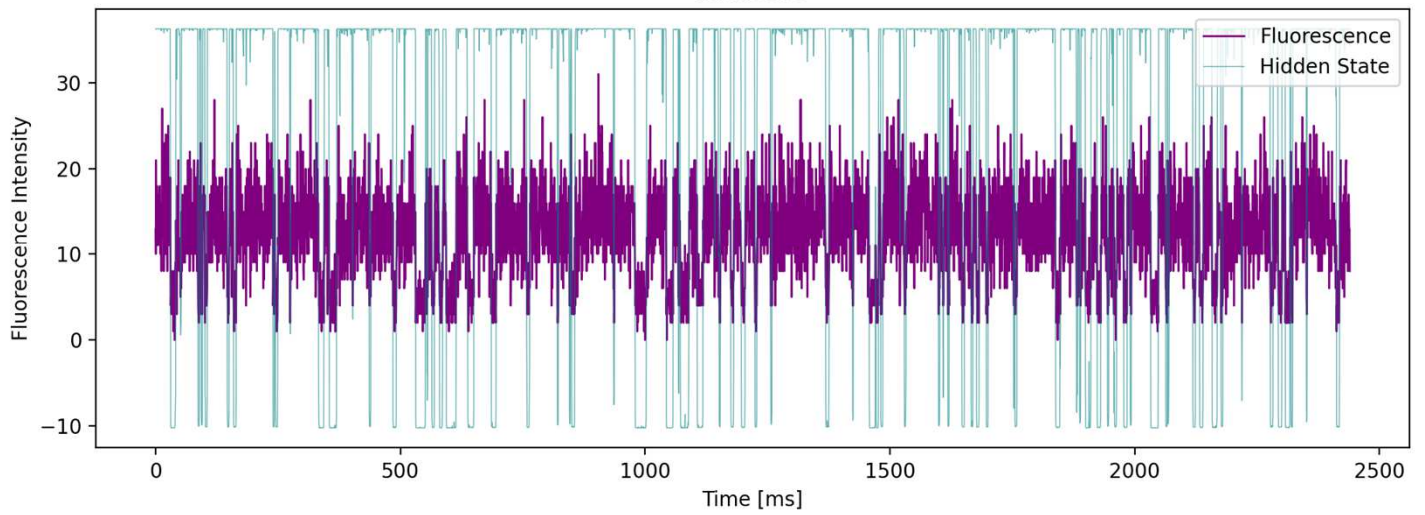
Histogram of Photon Counts (data0024.csv)



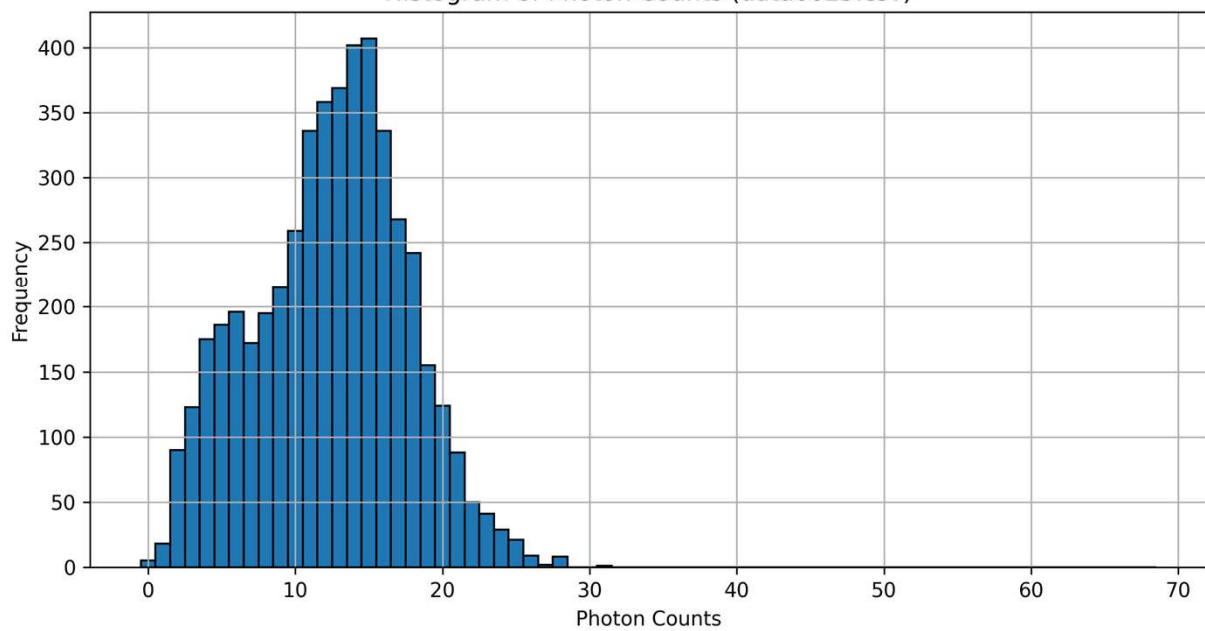
data0024



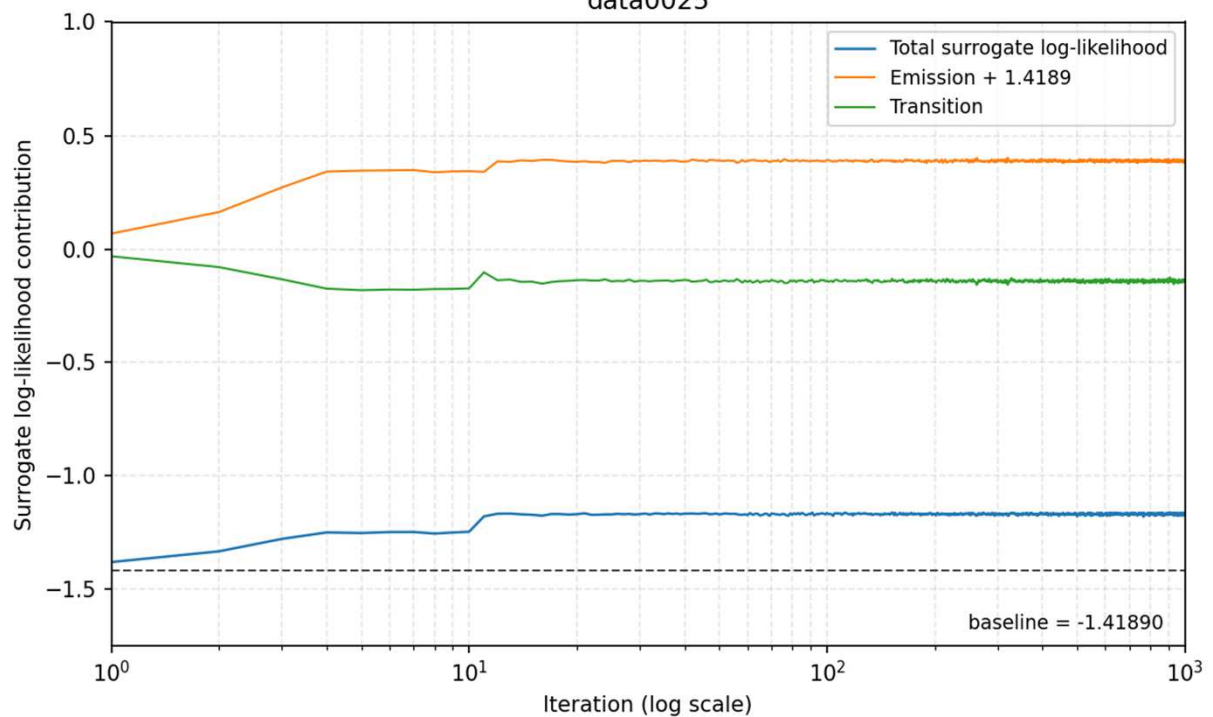
data0025

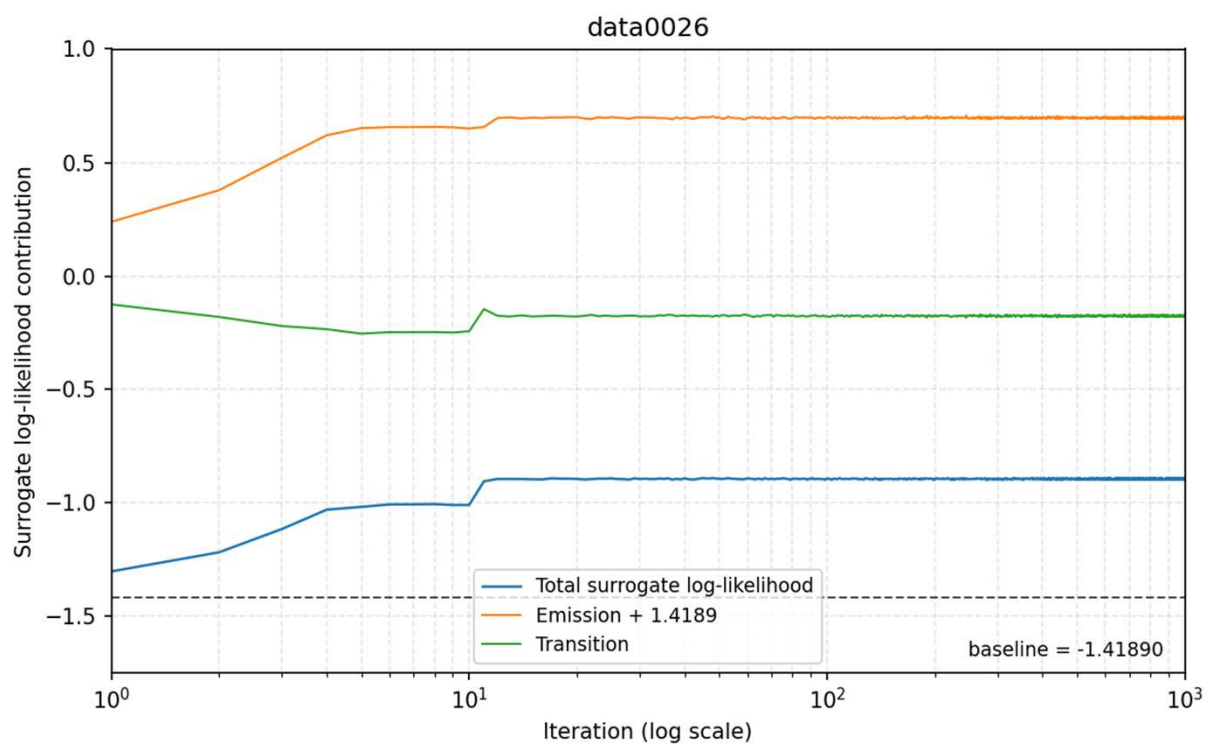
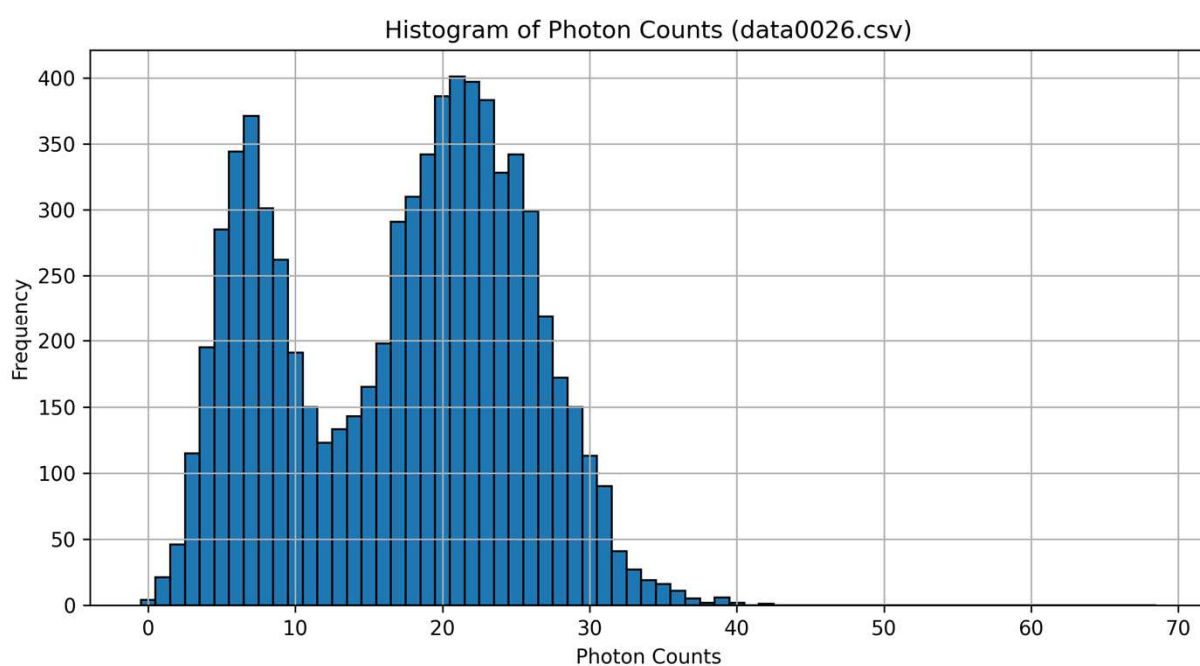
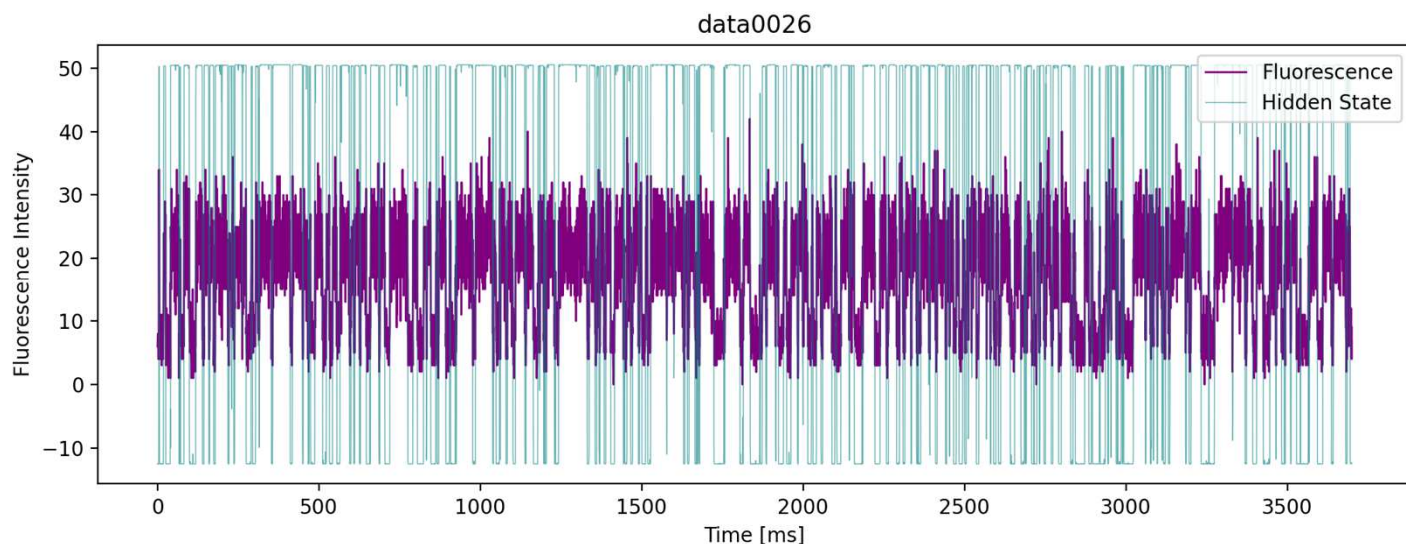


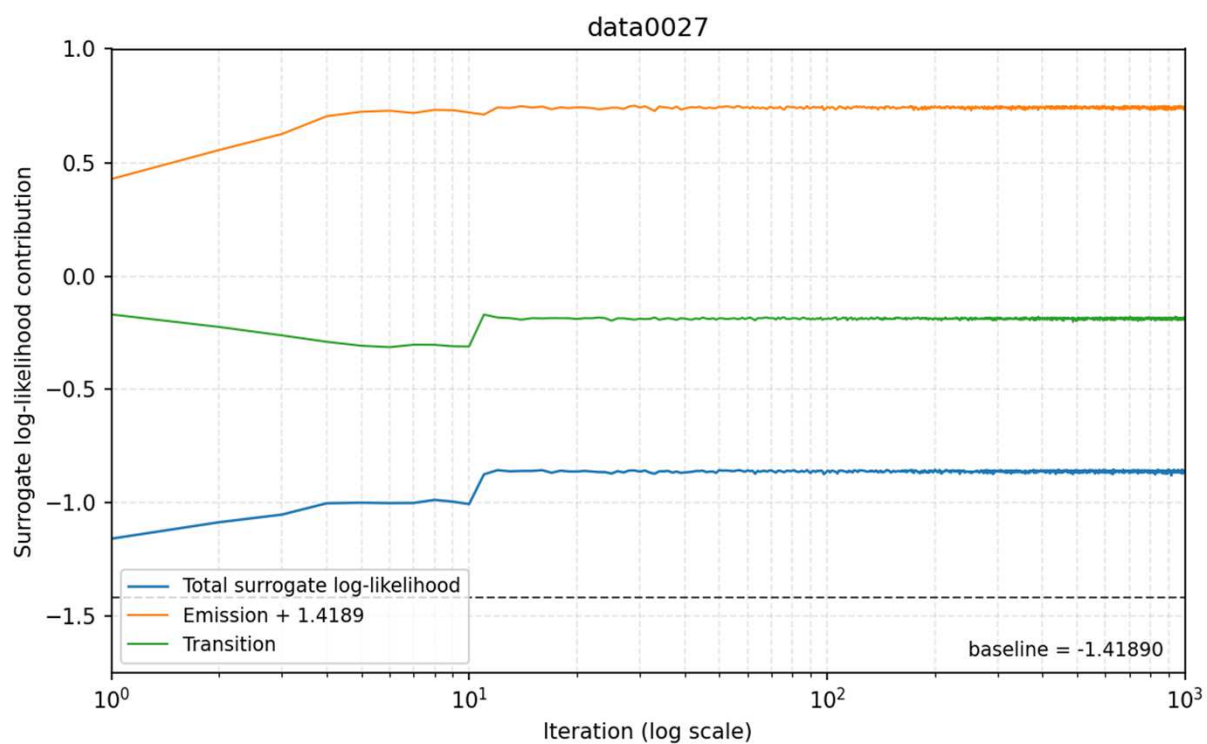
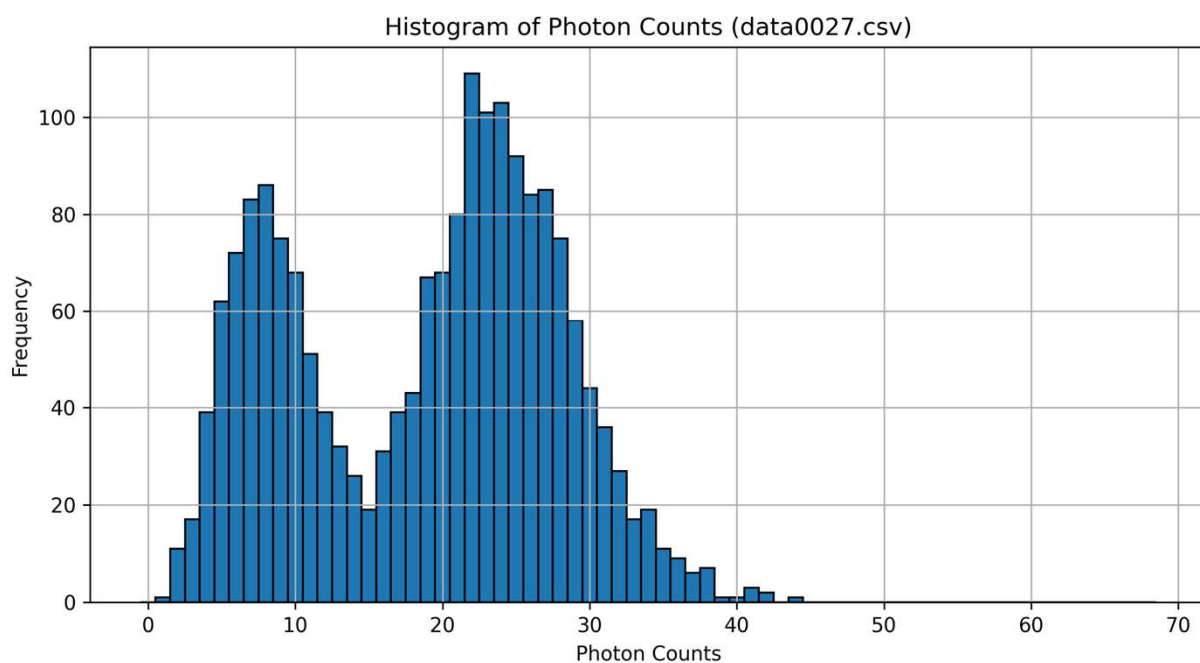
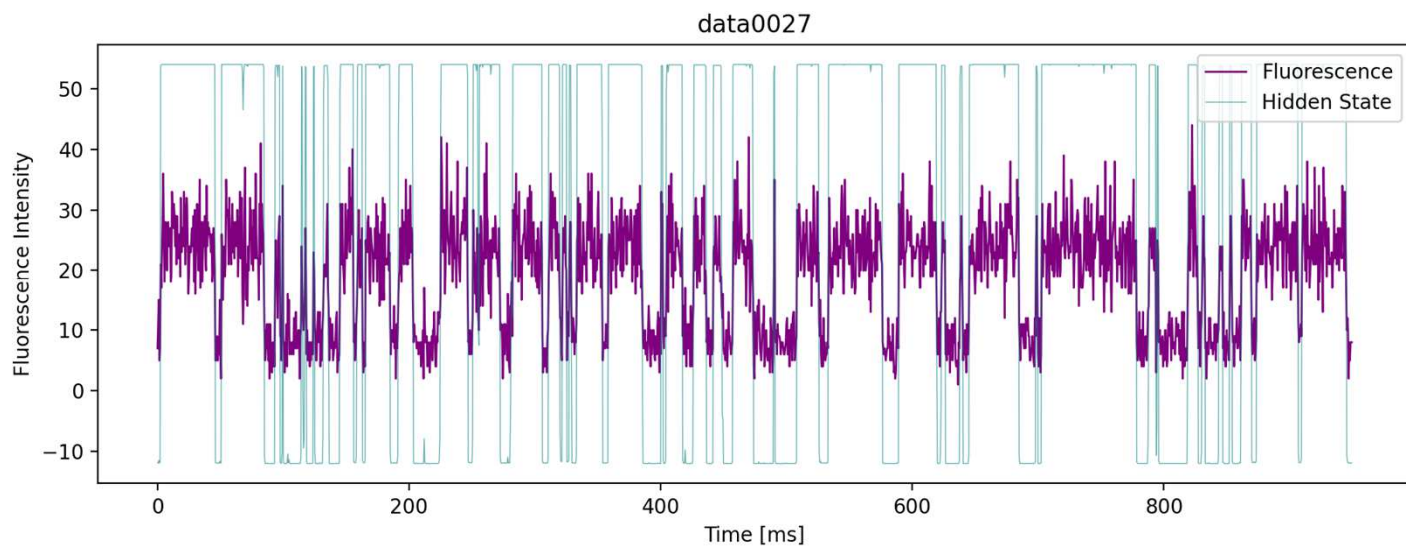
Histogram of Photon Counts (data0025.csv)



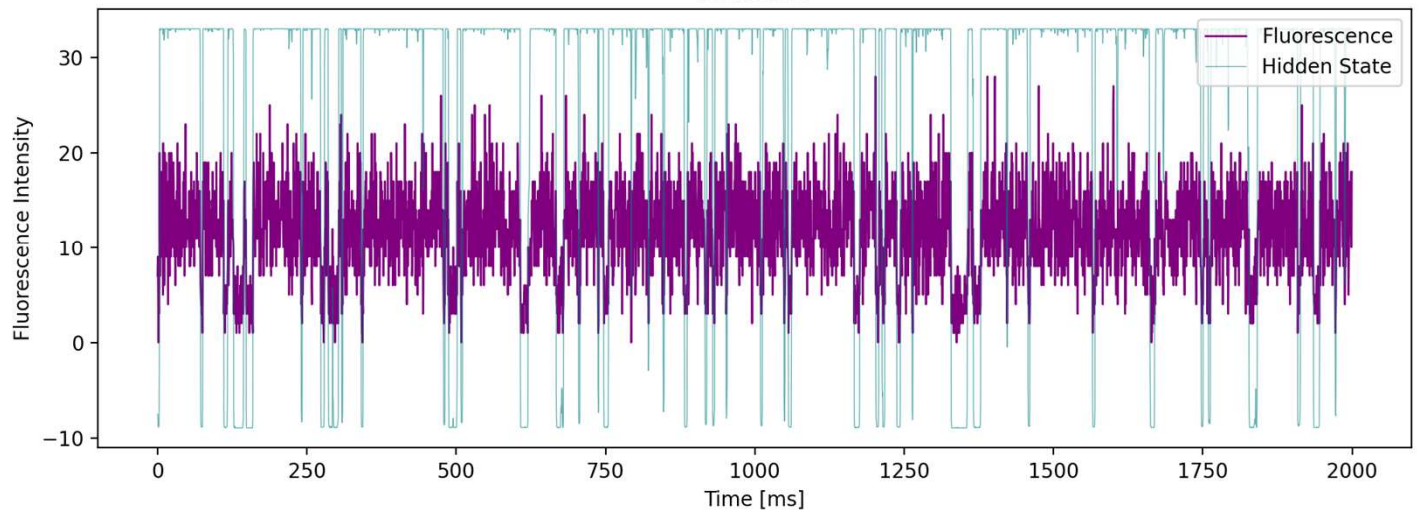
data0025



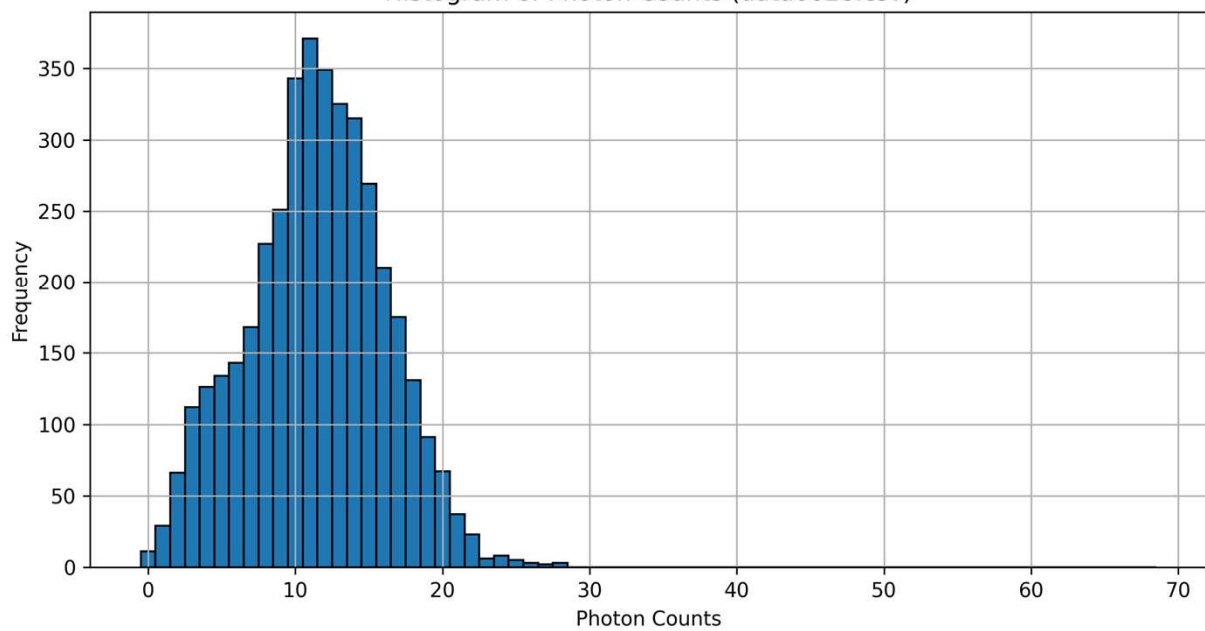




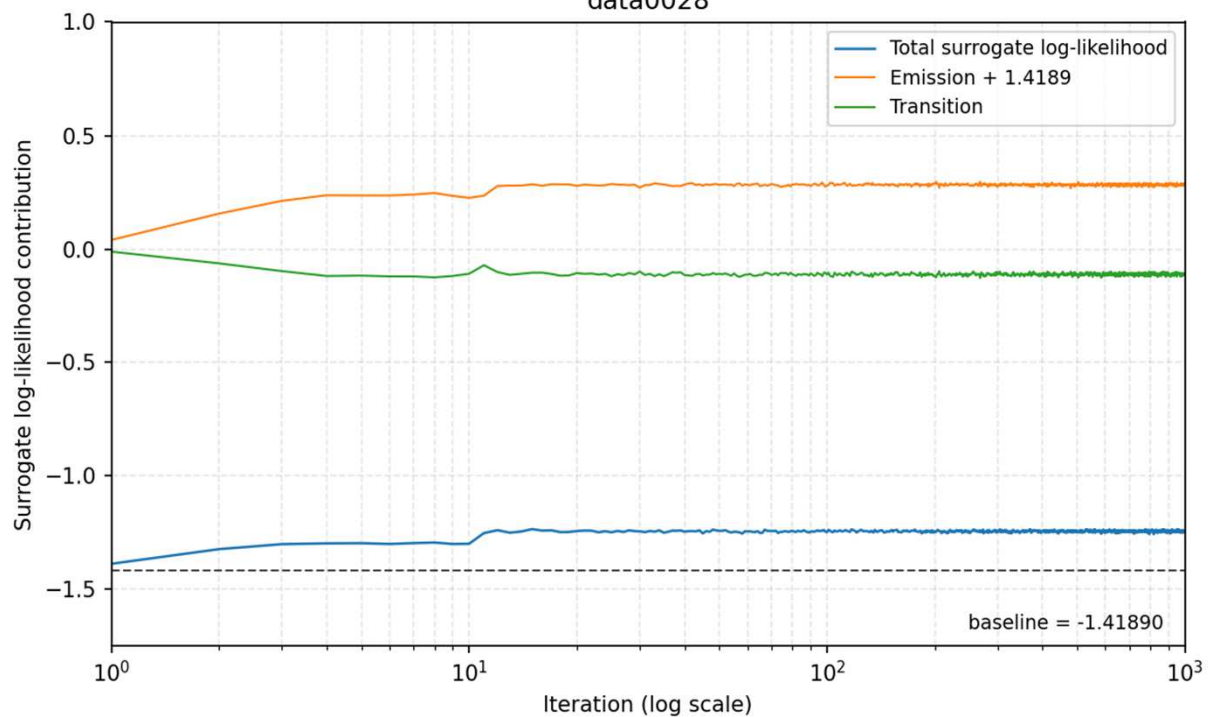
data0028



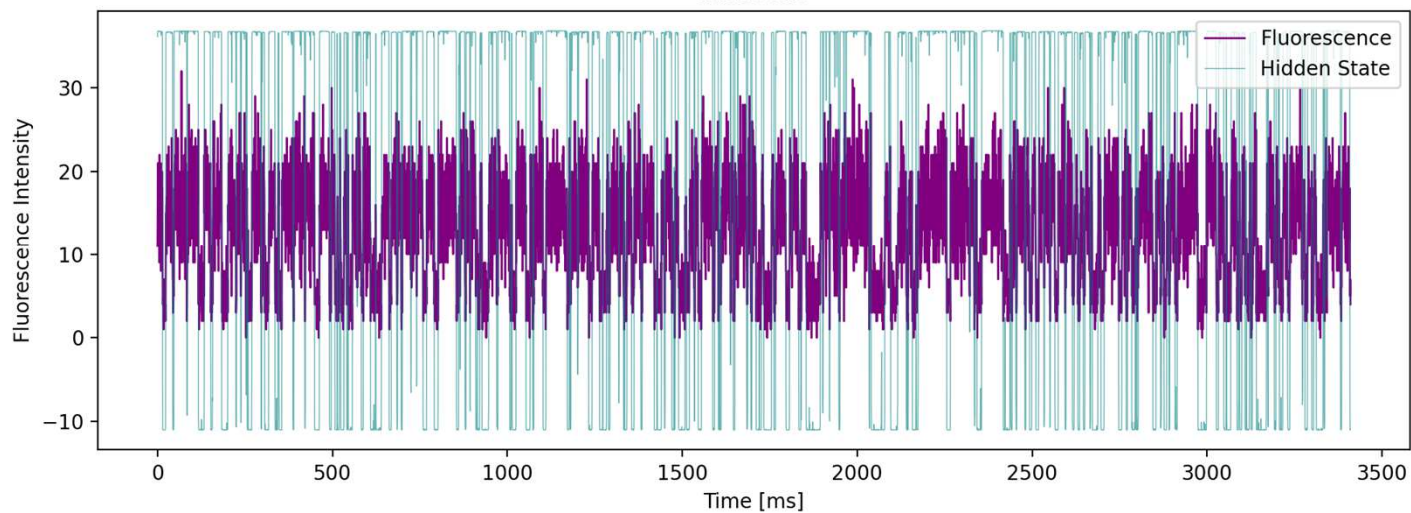
Histogram of Photon Counts (data0028.csv)



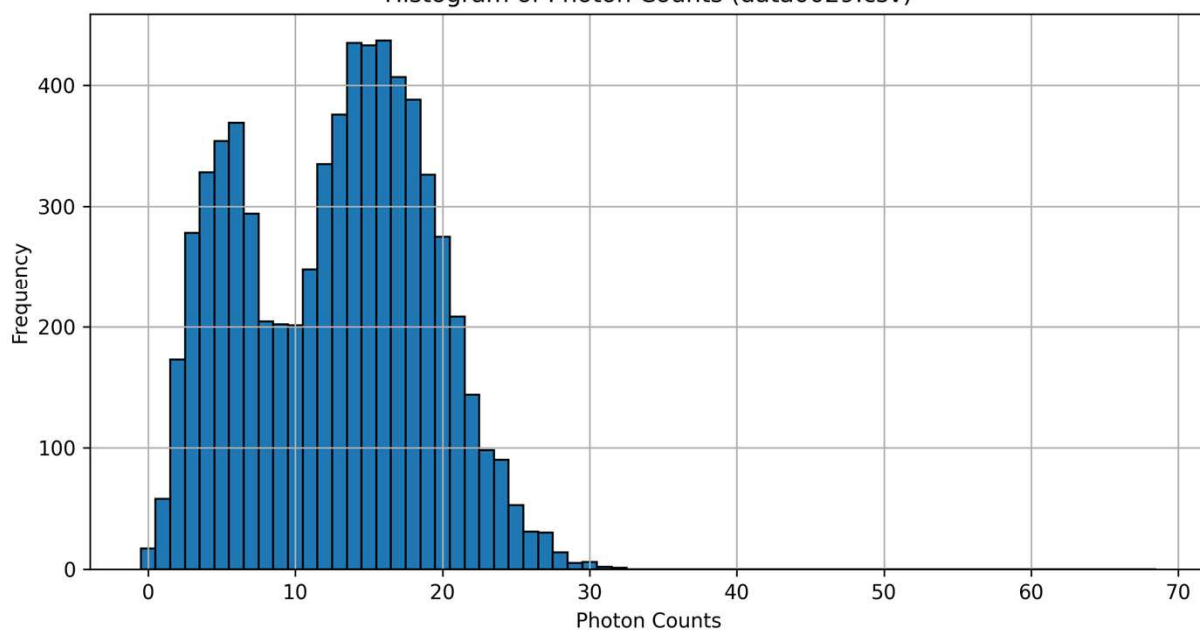
data0028



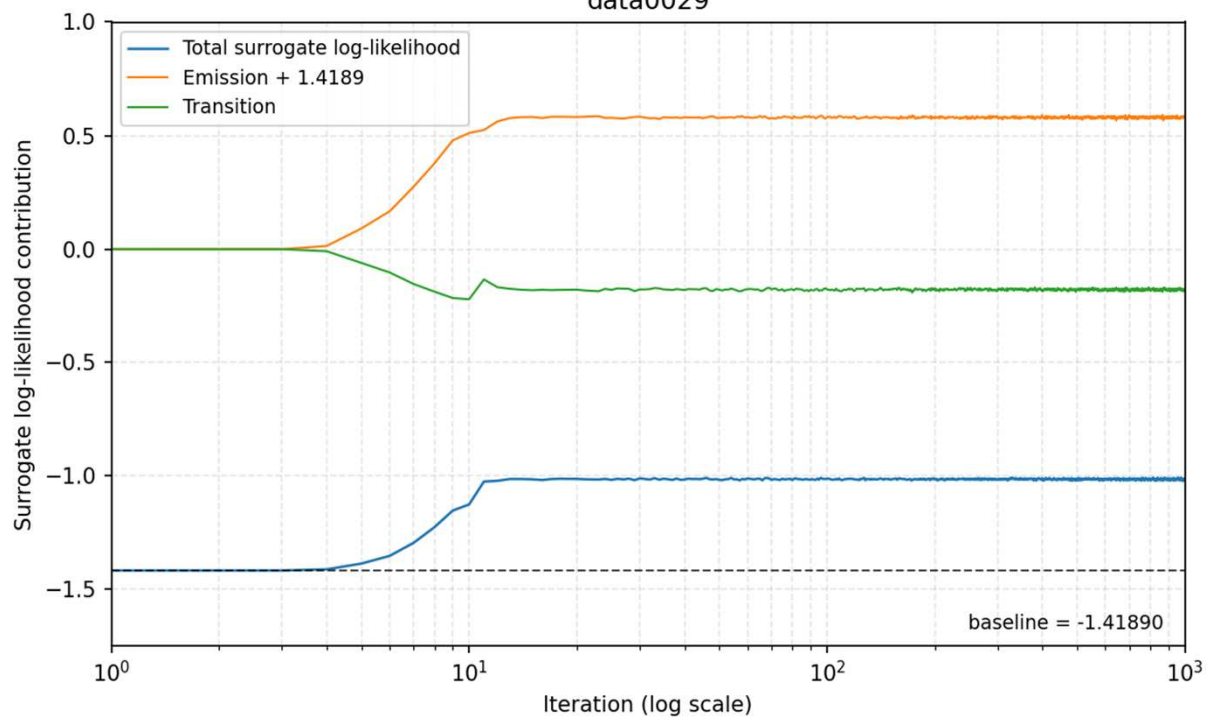
data0029



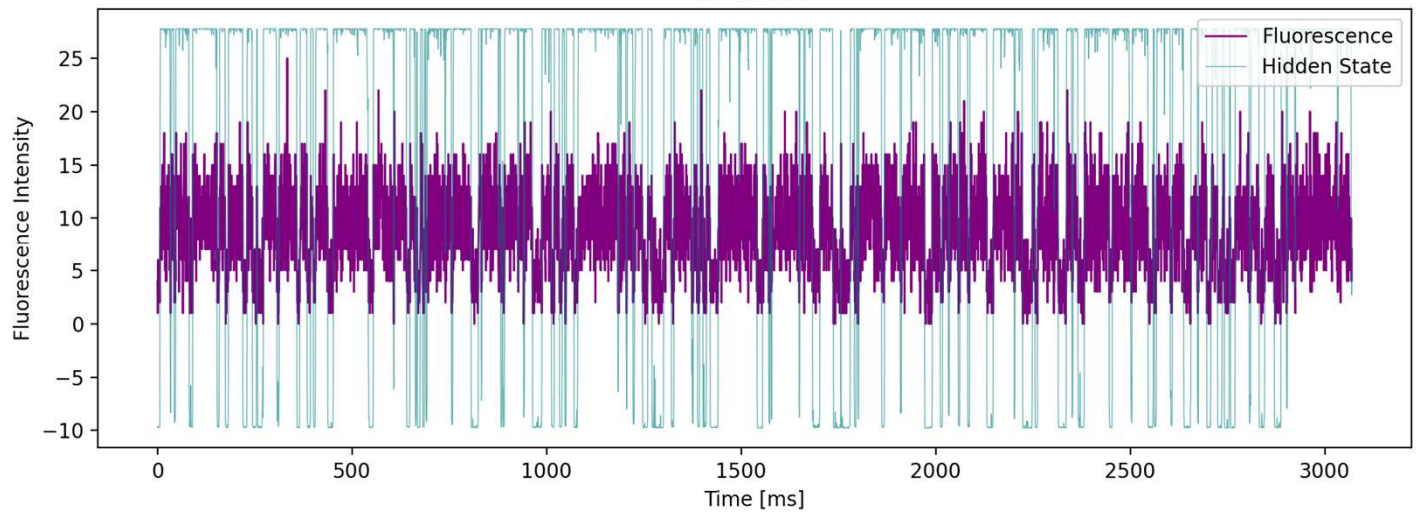
Histogram of Photon Counts (data0029.csv)



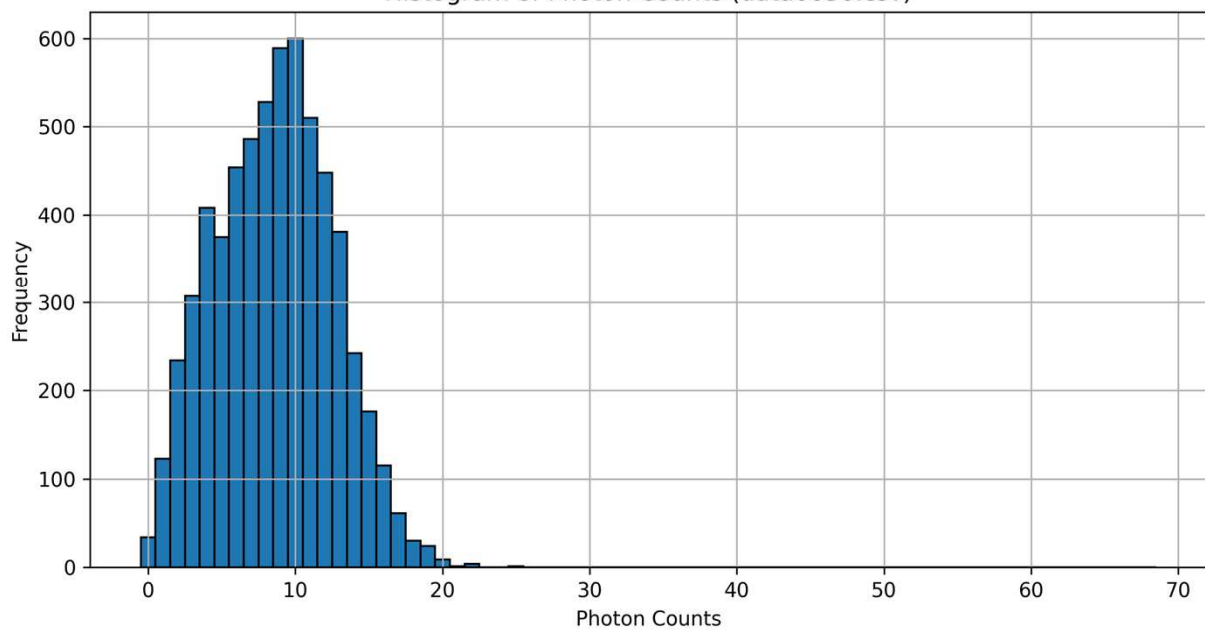
data0029



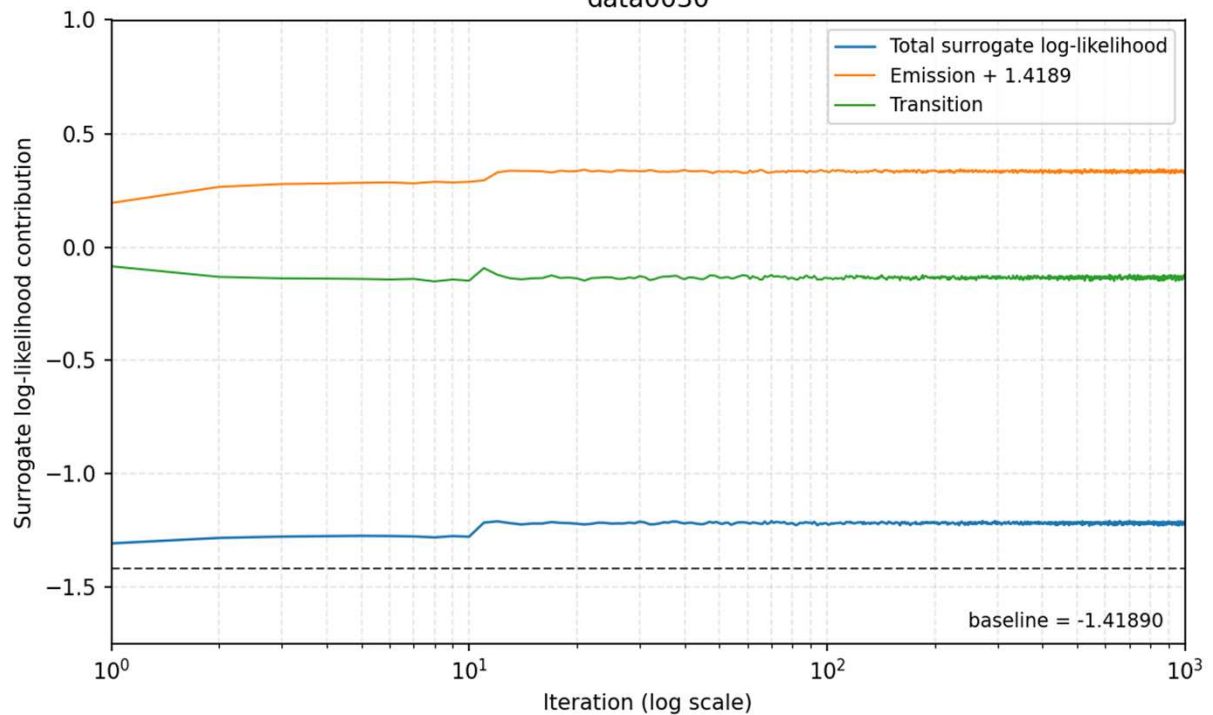
data0030



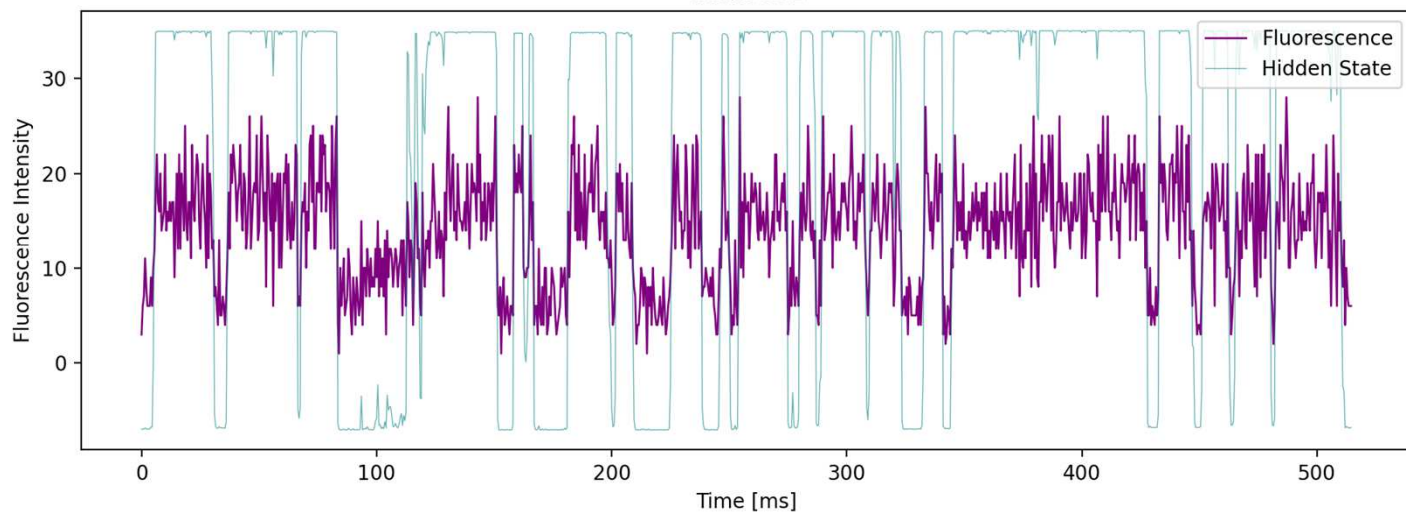
Histogram of Photon Counts (data0030.csv)



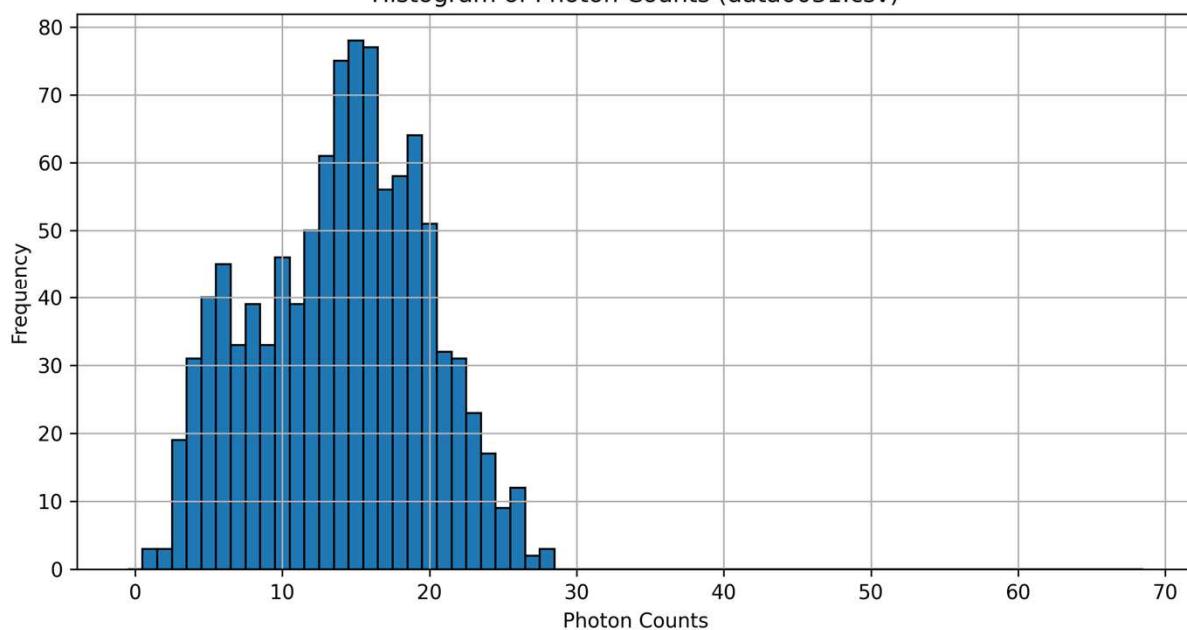
data0030



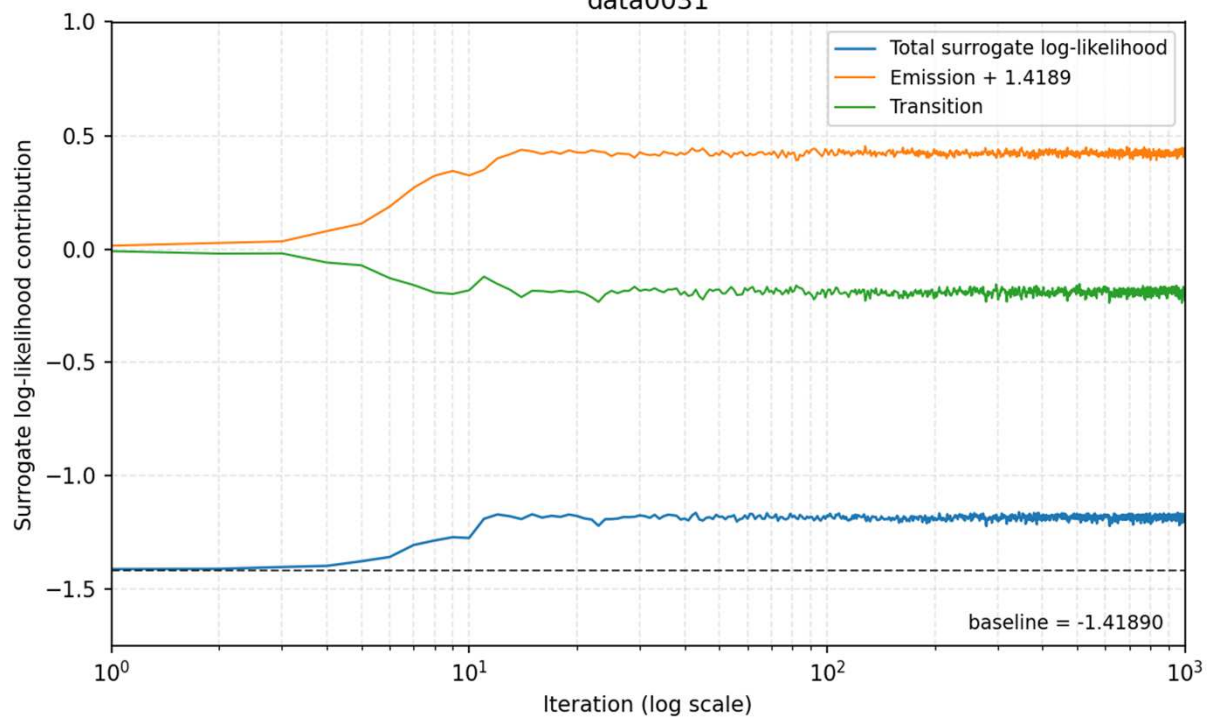
data0031



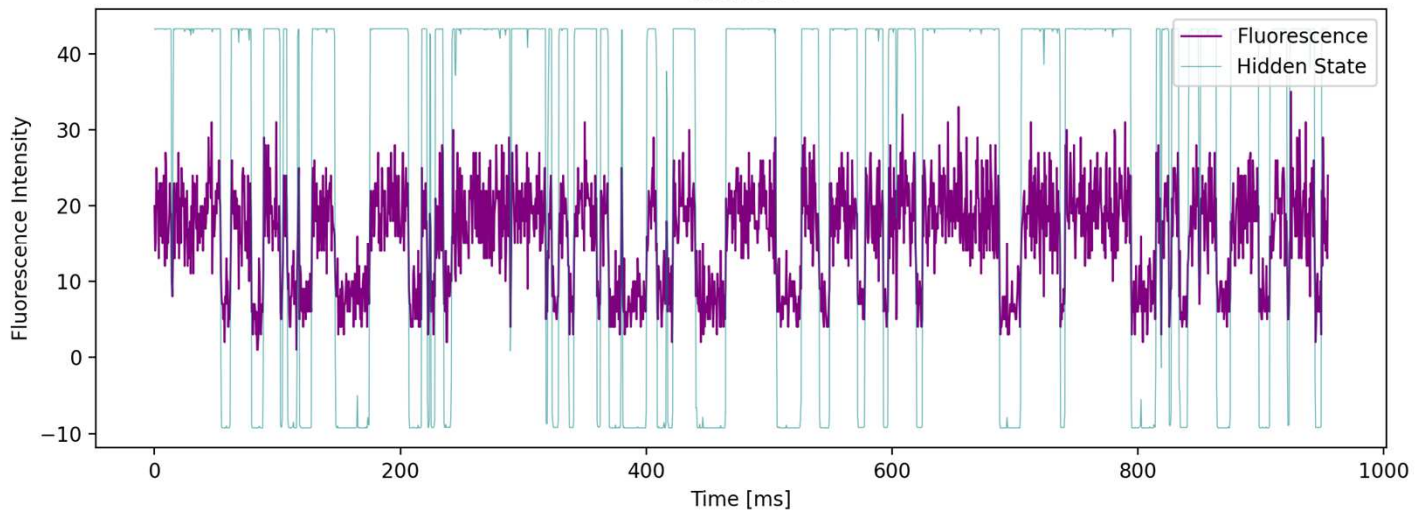
Histogram of Photon Counts (data0031.csv)



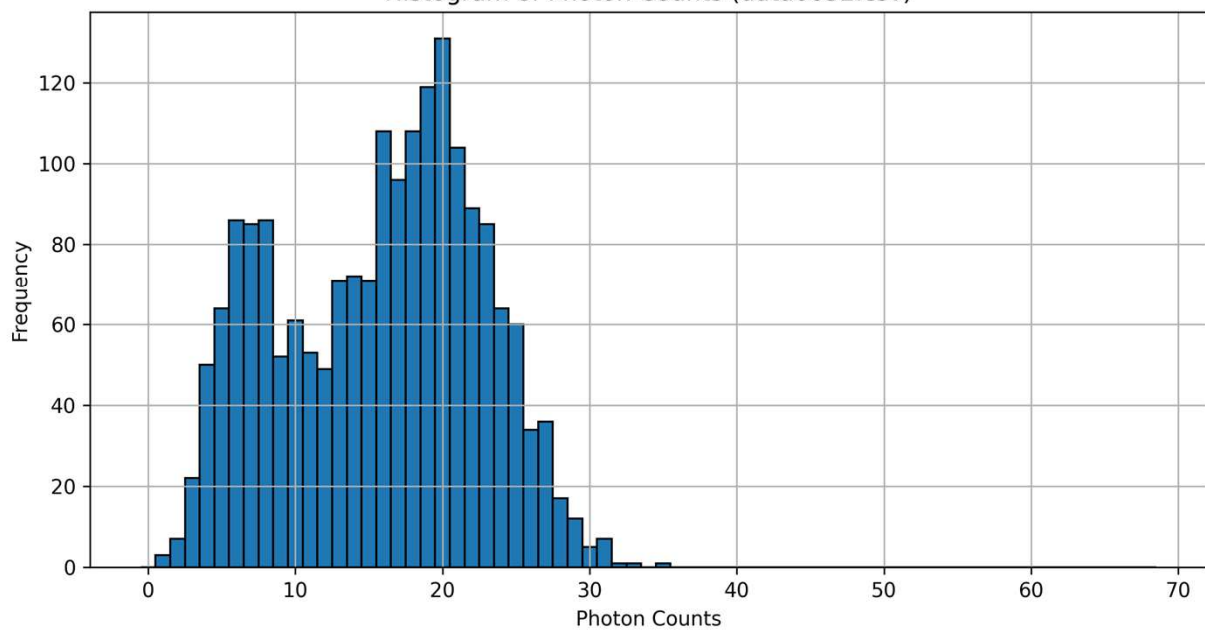
data0031



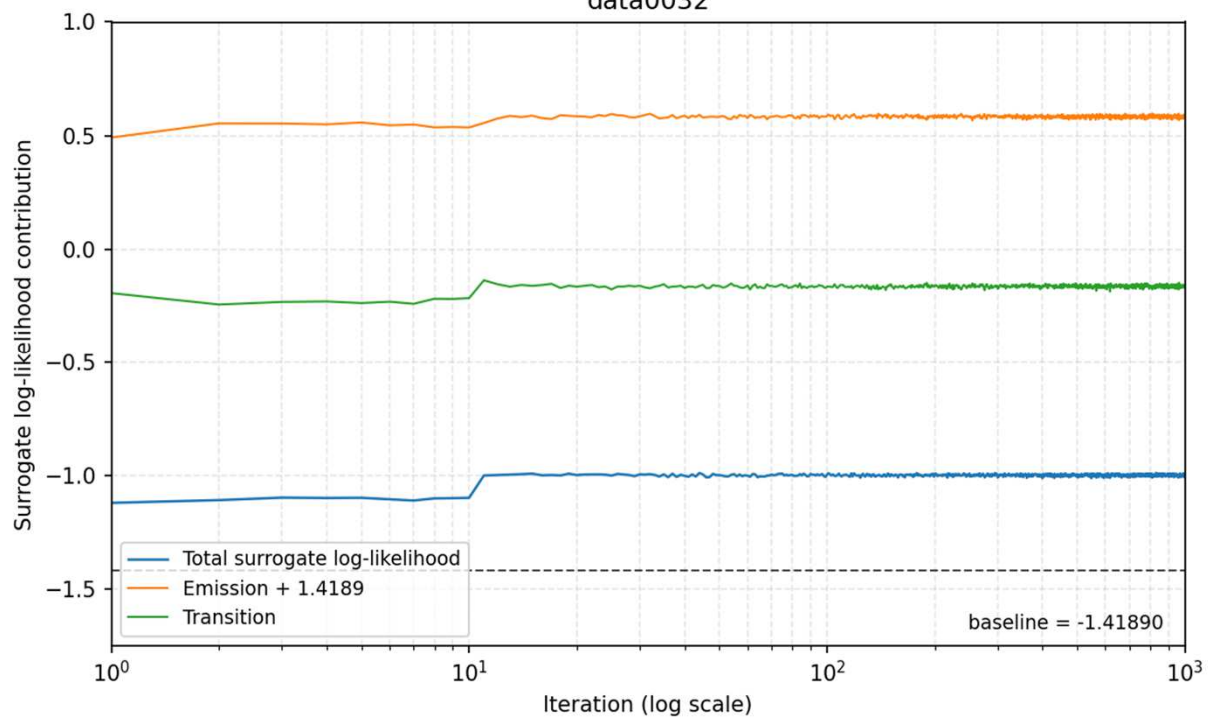
data0032



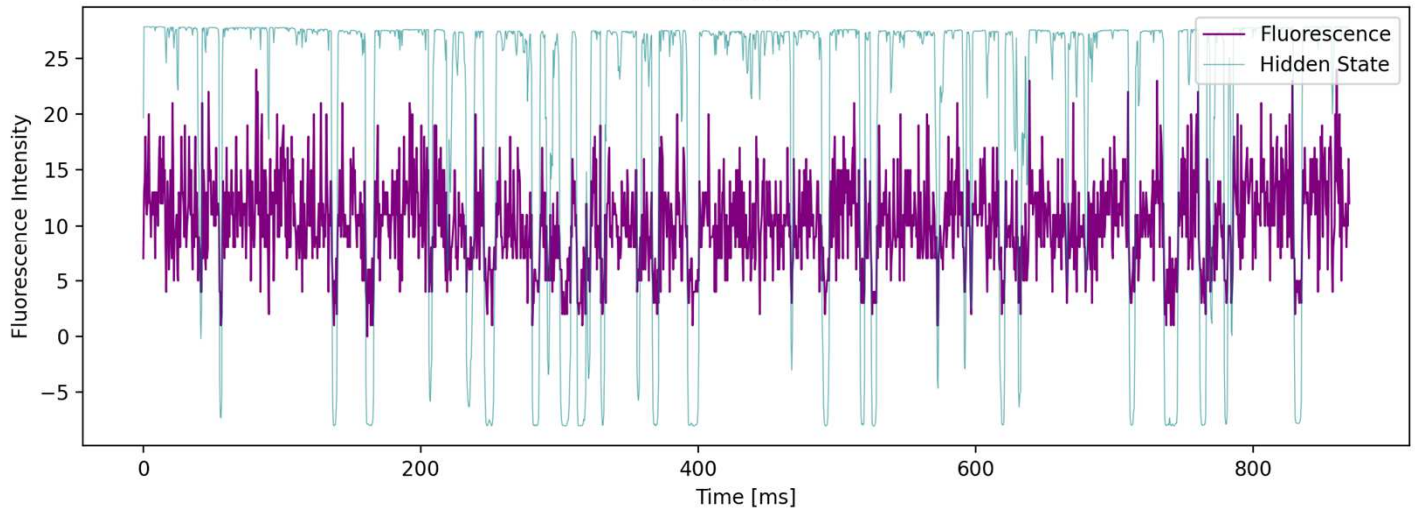
Histogram of Photon Counts (data0032.csv)



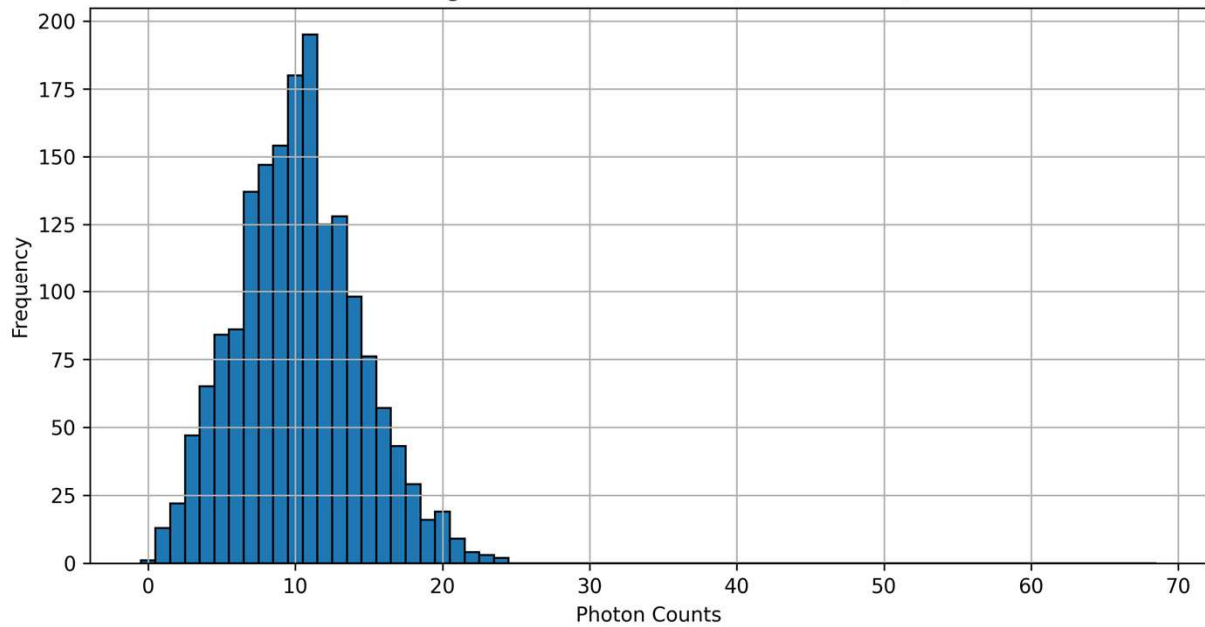
data0032



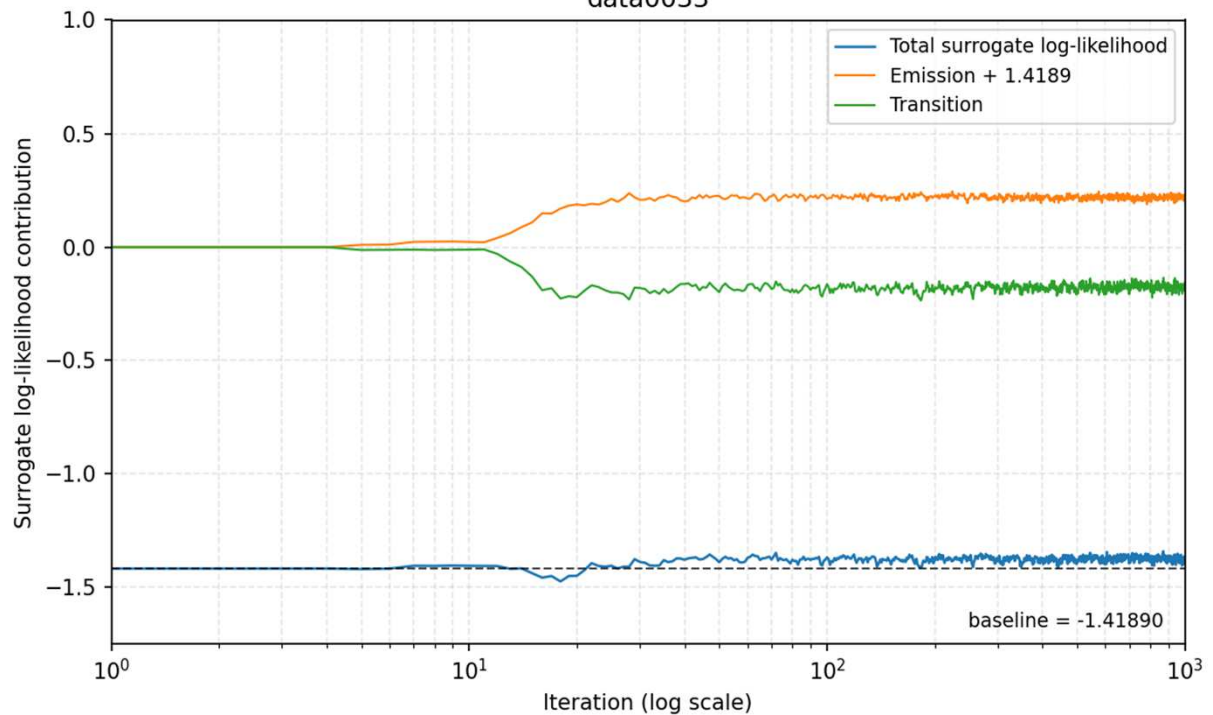
data0033



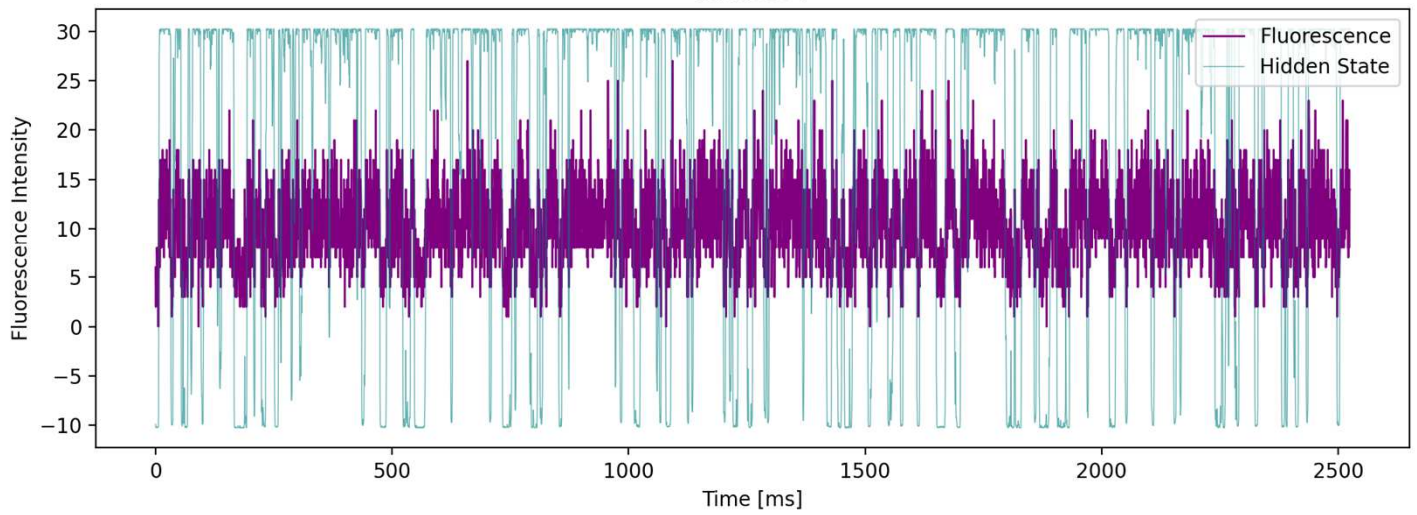
Histogram of Photon Counts (data0033.csv)



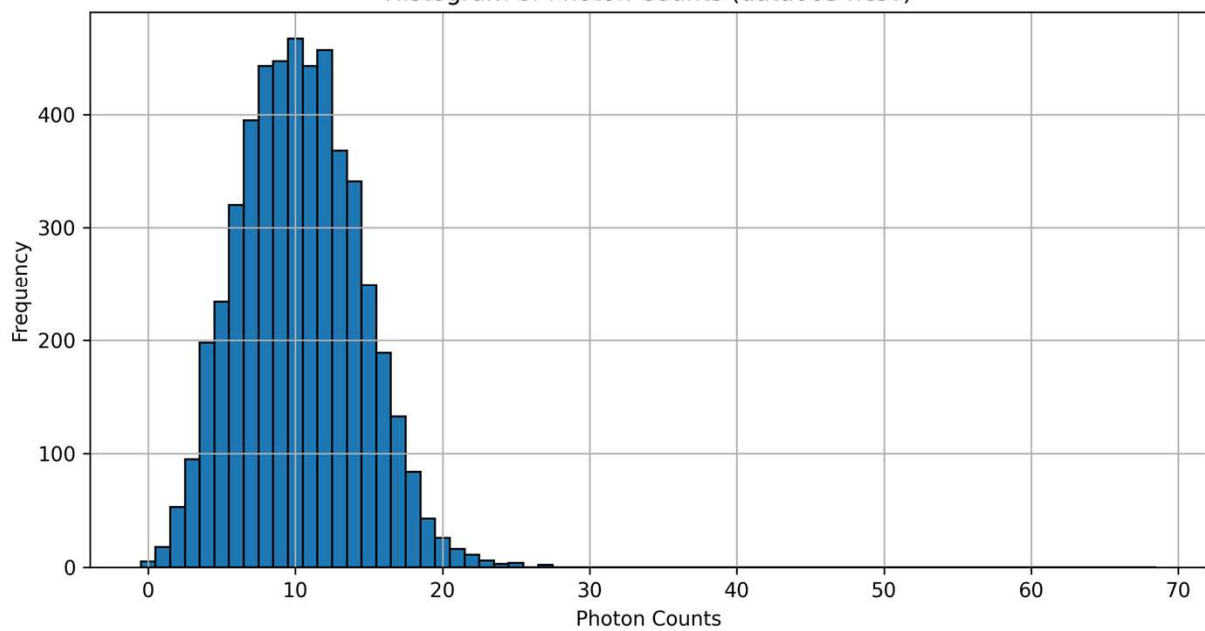
data0033



data0034



Histogram of Photon Counts (data0034.csv)



data0034

