Seeing Living Atoms, One Electron at a Time:

the expectation-maximization algorithm with Poisson statistics for analyzing counting frames of direct electron detectors in electron cryomicroscopy

Geoffrey Woollard
Department of Medical Biophysics, University of Toronto
Princess Margaret Cancer Centre, University Health Network
Toronto, Ontario, Canada

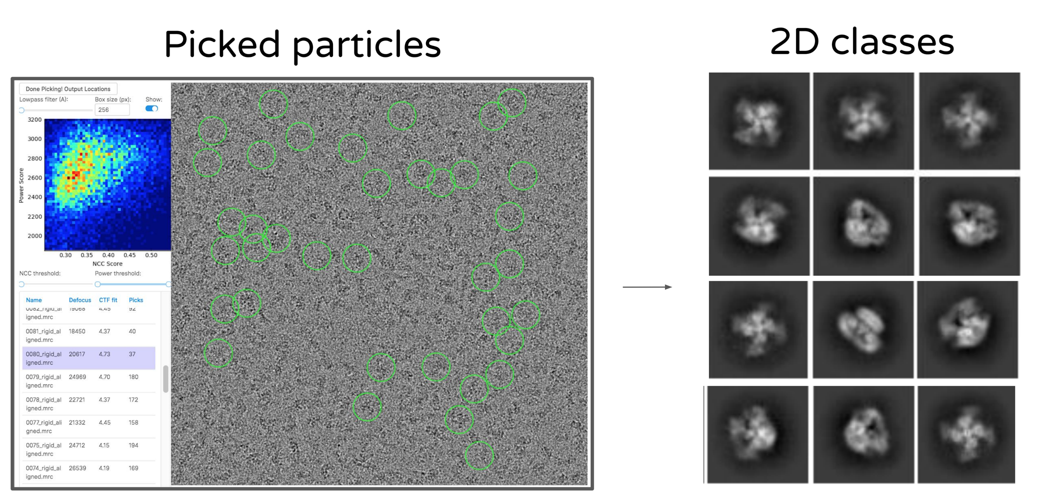
Abstract

→ **Electron cryomicroscopy (cryoEM)**
- Direct electron detectors are now fast and accurate enough to develop algorithms with the raw "counting frames", not the "movie frames" used in standard workflows

→ **Expectation-maximization (EM) algorithm with Poisson statistics**
- The EM algorithm on low dose (electron counts / pixel) counting frames breaks down with Gaussian statistics, while Poisson statistics are robust.

→ **A new wave of algorithms** can preserve information provided by multiple measurements collected by direct electron detectors, and let us see living atoms at higher resolution.

1. Introduction: A Standard CryoEM Algorithmic Workflow

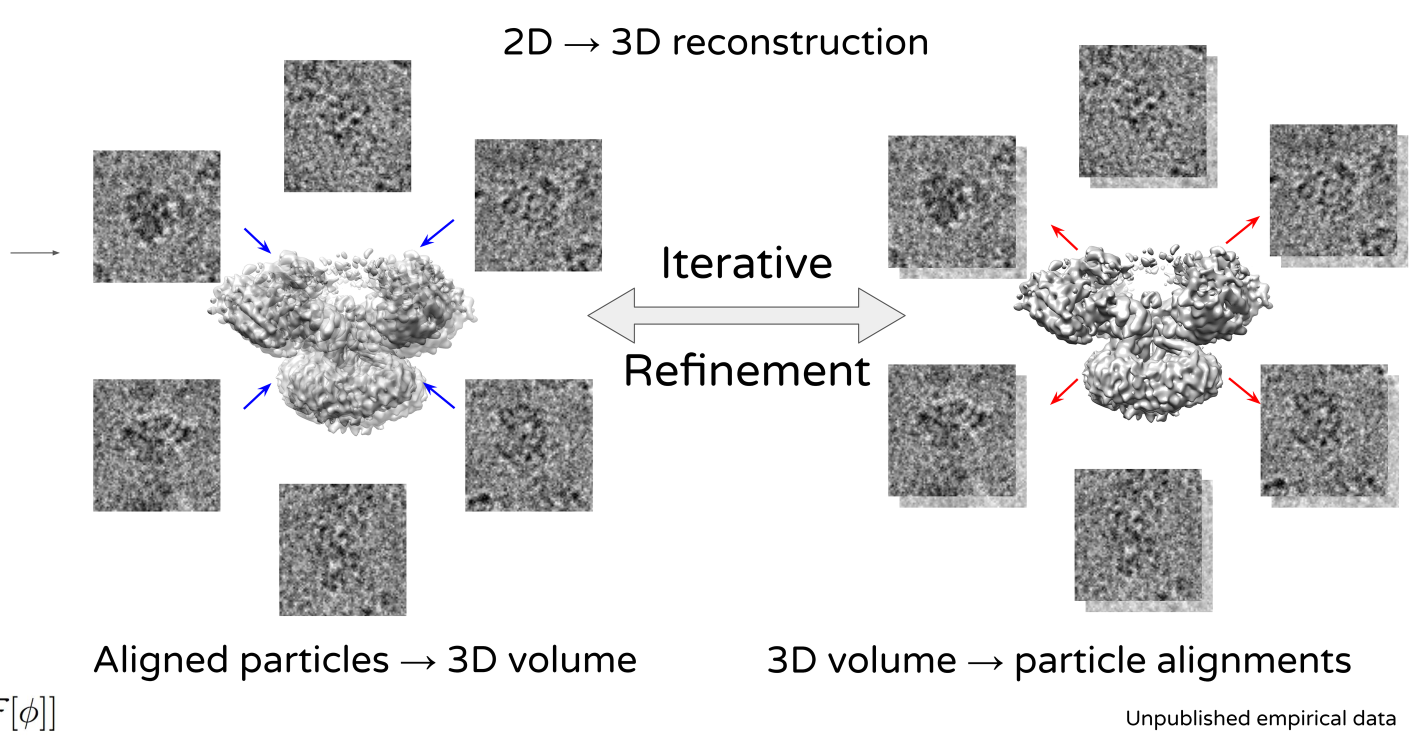


$$I_i(x,y) = psf_i \star \int_{-\infty}^{\infty} \phi(R_i r)\, dz + \sigma, \quad r = (x,y,z)$$

$$\mathcal{F}[I_i] = CTF_i * Pj_i[\mathcal{F}[\phi]] + \sigma$$

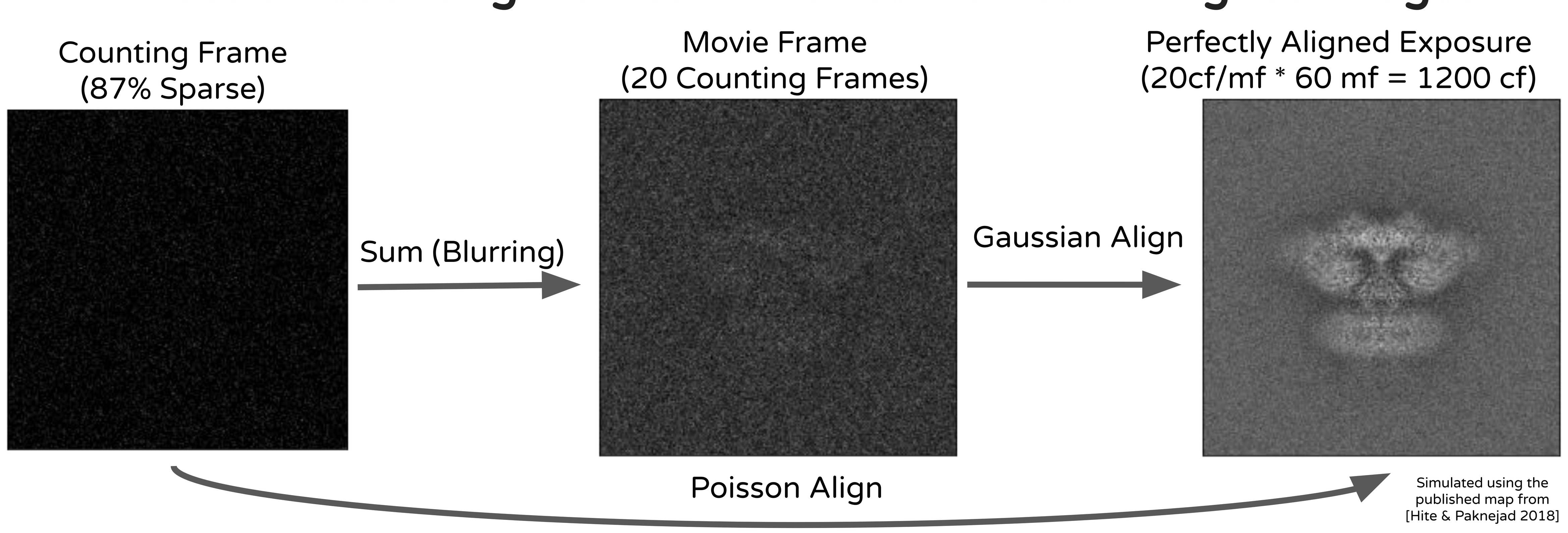
$$p_i = p_i(I_i | Pj_i, \phi_i, CTF_i)$$

$$\arg \min_{Pj_i, \phi_i, CTF_i} \log p_i = \arg \min_{Pj_i, \phi_i, CTF_i} \log \sum_{k \in \mathcal{F}} \mathcal{F}[I_i] - CTF_i * Pj_i[\mathcal{F}[\phi]]$$



Imaging **fragile biological samples** with **transmission electron microscopy** has traditionally been accomplished by embedding them in a glassy state of aqueous solution (vitreous water), and operating the microscope at **low doses**: so called electron cryomicroscopy (**cryoEM**). Low doses give rise to **shot noise**, where the value at a certain pixel varies from its expected value because of low electron count. Furthermore, the sample moves and is damaged during imaging (seconds), **blurring** features. This can be overcome by taking a series of image frames, and **aligning** them computationally. A **first wave of fast detectors and algorithms** was an important driver of the technological advances in cryoEM during the 2010s underpinning the 2017 Nobel Prize in Chemistry.

2. Electron Counting Frames → Movies Frames → Aligned Images



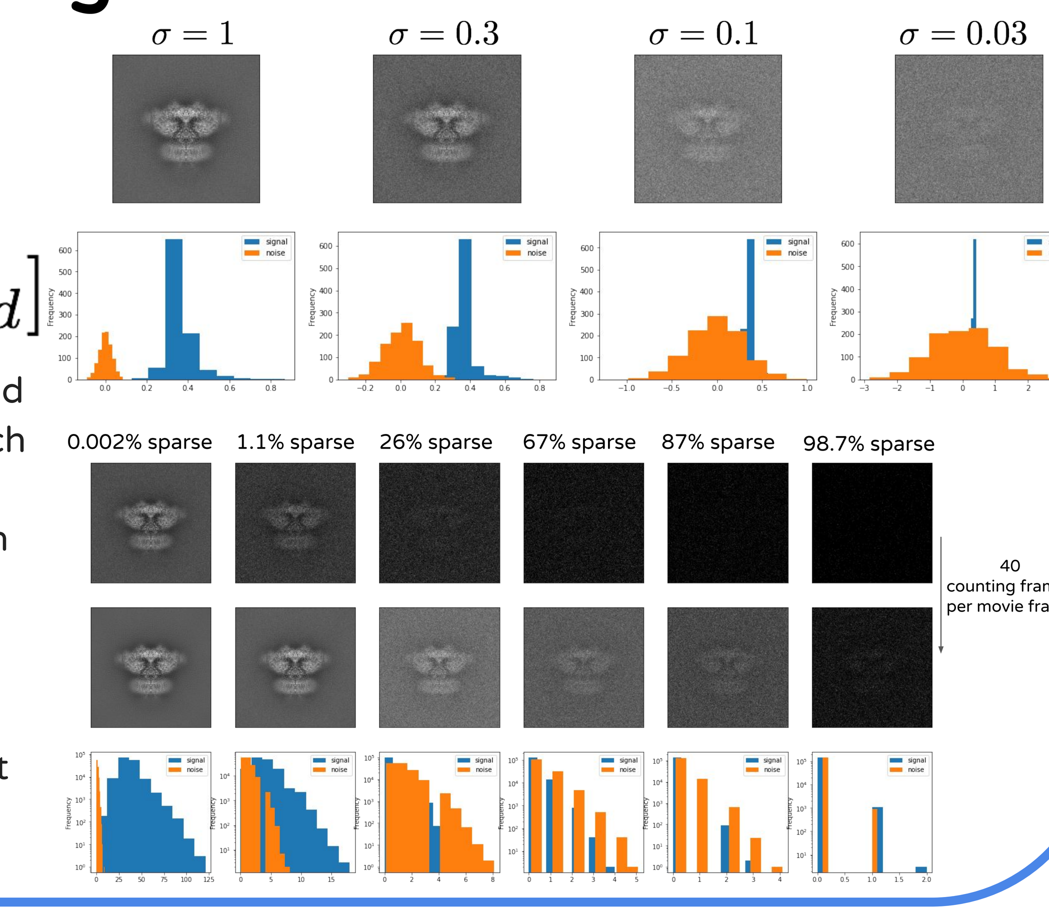
The detector technology has continued to improve in **speed** (frames per second) and **accuracy** (low miscounts of single events, i.e. coincidence loss). Researchers have already developed strategies to encode sparse frames in the range of 0.01 electron counts / pixel (1.6e5 counts in a 4K image) [Datta et al, 2019].

Historically the data used in cryoEM inverse problems safely assumed **Gaussian** statistics [Sigworth, 1998]. Although the underlying statistics are known to be modeled better as **Poisson**, using Gaussian statistics was justified through the central limit theorem. A faint outline of the image could be seen, at least if a few **"movie" frames** were averaged together and low pass filtered. But how sparse does the data need to be in order for Gaussian statistics to break down? Can we use **"counting" frames** with the existing approaches, or do we need to develop **new algorithms**?

3. Forward Model of Image Formation

$$\vec{X}_i = S(\vec{q}_i) R(\phi_i) \vec{A} + \sigma \vec{N}[\mu = 0, 1]$$

$$\vec{X}_i = P_{oiss}[S(\vec{q}_i) R(\phi_i) \vec{A} + \lambda_{background}]$$



An experimental observable (**Xi**) of a 2D projection (**A**) that is rotated (**R**) and translated (**S**) is corrupted by **Gaussian** additive noise, with the noise in each pixel iid (top equation)). Alternatively, both signal and noise are **Poisson** random variables (bottom equation), with each pixel having its own Poisson parameter from the unaligned signal, and a shared parameter from background noise.

This **simple model** could be extended to include the effects of the detector (detective quantum efficiency, **DQE**) and electron lens aberrations (contrast transfer function, **CTF**) [Vulović et al, 2013].

4. Expectation Maximization

$$\begin{aligned} p(\vec{A}|\text{data}) &= \int d^N\vec{q}\,d^N\phi\; p(\vec{A}, \{\vec{q}_i, \phi_i\}|\text{data}) \\ &= \int d^N\vec{q}\,d^N\phi\; \frac{p(\text{data}|\vec{A}, \{\vec{q}_i, \phi_i\}) p(\vec{A}, \{\vec{q}_i, \phi_i\})}{p(\text{data})} \\ &= C \prod_i \int d\vec{q}_i d\phi_i\; p(\vec{X}_i|\vec{A}, \vec{q}_i, \phi_i) p(\vec{A}, \vec{q}_i, \phi_i) \end{aligned}$$

$$\boxed{\gamma_i(\vec{q}_i, \phi_i) = p(\vec{X}_i|\vec{A}, \vec{q}_i, \phi_i) p(\vec{A}, \vec{q}_i, \phi_i)}$$

$$\begin{aligned} L(\vec{A}) &= \ln p(\vec{A}|data) \\ &= \sum_i \ln \int d\vec{q}_i d\phi_i \gamma_i \end{aligned}$$

$$\begin{aligned} & p(\vec{X}_i|\vec{A}, \vec{q}_i, \phi_i) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big[-\sum_\alpha \frac{(S(\vec{q}_i)R(\phi_i)\vec{A} - X_{i\alpha})^2}{2\sigma^2}\Big] \quad \text{Gaussian} \\ &= \prod_\alpha \frac{\lambda_{i\alpha}^{X_{i\alpha}} e^{-\lambda_{i\alpha}}}{X_{i\alpha}!} \quad \text{Poisson} \end{aligned}$$

$$\vec{A}_* = \arg \max_{\vec{A}} L$$

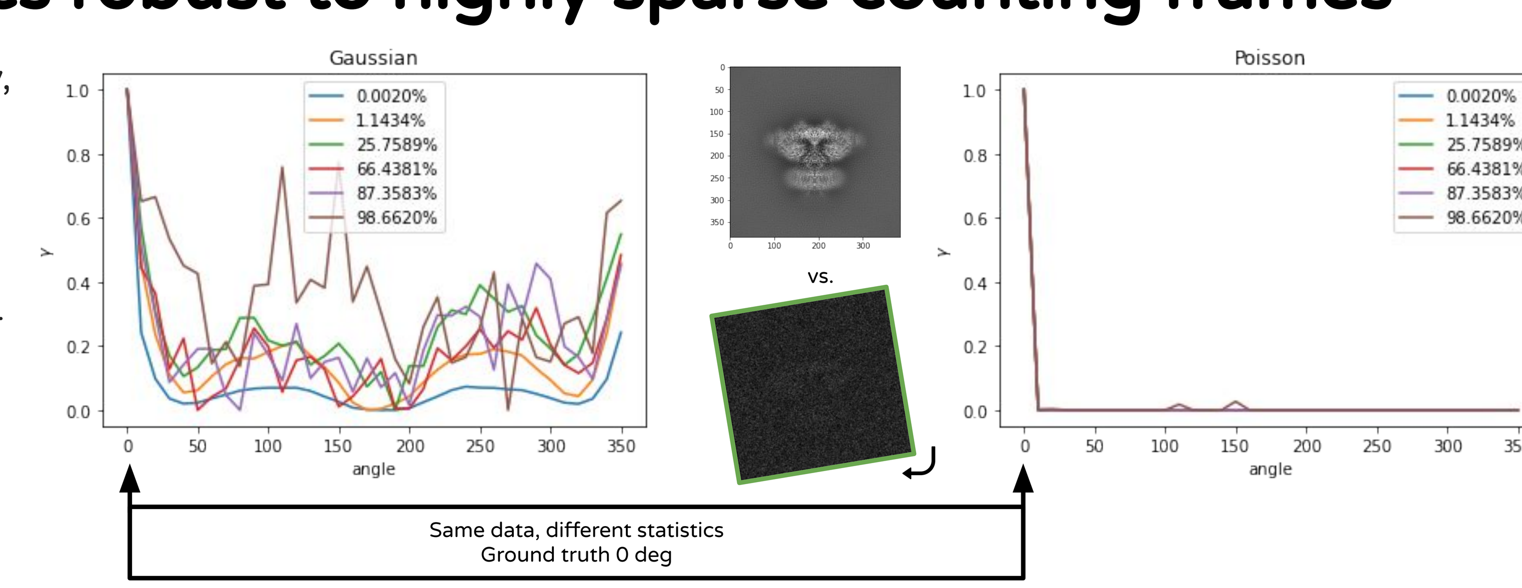
$$\Longrightarrow \vec{A}^{\text{next}} = \frac{1}{N} \sum_i \frac{\int d\vec{q}_i d\phi_i \gamma_i^{\text{prev}}(\vec{q}_i, \phi_i) \vec{X}_i}{\int d\vec{q}_i' d\phi_i' \gamma_i^{\text{prev}}(\vec{q}_i', \phi_i')}$$

The EM algorithm is an iterative maximum a posteriori estimate [Do & Batzoglou 2008, Nelson 2019]. It it a **Bayesian average** of the experimental observables over the nuisance parameters in the model (here rotation and translation alignment). Starting from the Poisson pmf, we can **analytically derive the EM update equations** using Poisson statistics [Loh 2014] following Nelson't treatment of the EM algorithm in [Nelson 2019]. It is common to perform EM in Fourier space with Gaussian statistics, because the Fourier Transform of a Gaussian is Gaussian. However, this is not so with low dose Poisson statistics that include signal and noise [Grob et al 2013, Loh 2014], so the updates done in real space. Effects in Fourier space (DQE, CTF) can be applied in that space before returning to real space. The background noise parameter in the Poisson model can be learned from the data.

5. Poisson statistics robust to highly sparse counting frames

Gaussian statistics break down at high sparsity, while Poisson statistics are robust to sparsity.

For low dose counting frames with Poisson statistics (see "3. Forward Model of Image Formation") the **Bayesian weighting factors** for a noisy simulated experimental observable correctly signal out the ground truth (0 deg rotation) when compared with the noiseless signal (A).



References

1. Datta, A., Ban, Y., Ding, M., Chee, S. W., Shi, J., & Loh, N. D. (2019). ReCoDe: A Data Reduction and Compression Description for High Throughput Time-Resolved Electron Microscopy.
2. Nelson, P. C. (2019). Chapter 12 : Single Particle Reconstruction in Cryo-electron Microscopy. In *Physical Models of Living Systems* (pp. 305–325).
3. Loh, N. D. (2014). A minimal view of single-particle imaging with X-ray lasers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1647), 20130328.
4. Grob, P., Bean, D., Typke, D., Li, X., Nogales, E., & Glaeser, R. M. (2013). Ranking TEM cameras by their response to electron shot noise. *Ultramicroscopy*, *133*, 1–7.
5. Vulović, M., Ravelli, R. B. G., van Vliet, L. J., Koster, A. J., Lazić, I., Lücken, U., … Rieger, B. (2013). Image formation modeling in cryo-electron microscopy. *Journal of Structural Biology*, *183*(1), 19–32.
6. Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology*, *26*(8), 897–899.
7. Sigworth, F. J. (1998). A Maximum-Likelihood Approach to Single-Particle Image Refinement. *Journal of Structural Biology*, *122*(3), 328–339.
8. Hite, R. K. & Paknejad, N. (2018). https://www.emdataresource.org/EMD-7981