Journal of Structural Biology 214 (2022) 107913

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Smart data collection for CryoEM

Tristan Bepler[a], Andrew J. Borst[b], Jonathan Bouvette[c], Giuseppe Cannone[d], Songye Chen[e], Anchi Cheng[a], Ao Cheng[f], Quanfu Fan[g], Fanis Grollios[h], Harshit Gupta[i], Meghna Gupta[j], Theo Humphreys[k], Paul T. Kim[a], Huihui Kuang[a], Yilai Li[l], Alex J. Noble[a], Ali Punjani[m], William J. Rice[n], Carlos Oscar S. Sorzano[o], Scott M. Stagg[p], Joshua Strauss[q], Lingbo Yu[h], Bridget Carragher[a], Clinton S. Potter[a,*]

[a] New York Structural Biology Center, New York, NY, USA
[b] University of Washington, Institute for Protein Design, Seattle, WA, USA
[c] National Institute of Environmental Health Sciences, NIH, Durham, NC, USA
[d] Laboratory for Molecular Biology, Medical Research Council, Cambridge, England
[e] California Institute of Technology, Pasadena, CA, USA
[f] Northwestern University, Evanston, IL, USA
[g] MIT-IBM Watson AI Lab, Cambridge, MA, USA
[h] ThermoFisher Scientific, Eindhoven, The Netherlands
[i] SLAC National Accelerator Laboratory, Menlo Park, CA, USA
[j] University of California at San Francisco, San Francisco, CA, USA
[k] Pacific Northwest CryoEM Center, Portland, OR, USA
[l] University of Michigan, Ann Arbor, MI, USA
[m] Structura Biotechnology, Toronto, Canada
[n] New York University School of Medicine, New York, NY, USA
[o] Biocomputing Unit, Natl. Center of Biotechnology, CSIC, Madrid, Spain
[p] Florida State University, Tallahassee, FL, USA
[q] University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

ABSTRACT

This report provides an overview of the discussions, presentations, and consensus thinking from the Workshop on Smart Data Collection for CryoEM held at the New York Structural Biology Center on April 6–7, 2022. The goal of the workshop was to address next generation data collection strategies that integrate machine learning and real-time processing into the workflow to reduce or eliminate the need for operator intervention.

1. Introduction

A Workshop on Smart Data Collection for CryoEM was organized by the National Resource for Automated Molecular Microscopy (NRAMM) and held at the Simons Electron Microscopy Center, New York Structural Biology Center, New York, NY on April 6–7, 2022. Twenty-three participants from 17 institutions attended the meeting and contributed to this paper. Most of the participants were actively engaged in developing or improving methods for data collection in cryoEM, with a focus primarily on single particle cryoEM. The slides and video from the presentations and discussions are available at: https://nramm.nysbc.org/workshop-on-smart-data-collection-for-cryoem/.

Single particle cryoEM data collection typically requires acquisition of thousands of images from a transmission electron microscopy (TEM) grid. TEM grids are composed of a metal disk (e.g., copper, gold, molybdenum) perforated by a mesh of squares (10′s μm in dimensions) and,

Peer review under responsibility of For special issues, the foot notes must be placed. For the SI “VSI: Coiled coils’, Please insert the below footnote This Special Issue is edited by Andrei N. Lupas and Derek N. Woolfson, and is published as a companion to the meeting on CoiledCoil, Fibrous & Repeat Proteins at Alpbach, Austria, September 38, 2017For the SI ”SI: Protein Tandem Repeats“, Please insert the below footnote This Special Issue is edited by Andrey V. Kajava and Silvio C.E Tosatto, and is published as a companion to the COST Action BM1405 meeting on nonglobular proteins in molecular physiopathology at Belgrade, Serbia, September 1517, 2016.

* Corresponding author.
E-mail address: cpotter@nysbc.org (C.S. Potter).

https://doi.org/10.1016/j.jsb.2022.107913
Received 2 August 2022; Received in revised form 29 September 2022; Accepted 23 October 2022
Available online 28 October 2022
1047-8477/© 2022 Elsevier Inc. All rights reserved.

for cryoEM, the mesh is usually covered with a thin (10′s nm) support film (e.g. carbon, gold) perforated by holes (from 0.2 to 2 μm in diameter). The specimen of interest is immobilized in vitreous ice (10′s nm thick) supported by the holey substrate and images are acquired over regions of the specimen suspended across the holes so that the support film does not contribute additional signal to the images. The goal is to identify areas across the holes where the vitrified ice is of ideal thickness and particles are adequately distributed (not too sparse and not too crowded). Most data collection strategies involve examining images at a sequentially increasing series of magnifications: first obtaining an overview of the entire grid, then imaging squares, followed by imaging regions of interest inside holes, then finally acquiring high-magnification images of the specimen. In a single day, an accomplished electron microscopist can manually acquire hundreds of high-magnification exposures, but automated data acquisition software enables collection of tens of thousands of images per day.

Automated data collection attempts to emulate the performance of an experienced microscopist. The first step is assessing the overall grid quality, so most data collection strategies begin with collecting an overview atlas of the grid by automatically stitching together multiple low magnification images such that a majority of the grid is covered (see Fig. 1a). The atlas reveals overall features of the sample including regions where the ice is absent, very thick regions, dried out or cracked squares, as well as areas that may be appropriate for data collection. The user will typically make judgements by eye, selecting squares from the potentially good areas and avoiding the bad ones. The automated software will then move to the user-targeted squares and acquire a medium magnification image such that the holes can be seen (see Fig. 1b). All automated software systems have procedures for identifying holes and a variety of options for assessing the likelihood that the hole will yield ice of suitable thickness (see Fig. 1c). The operator usually adjusts various parameters of the hole finders to optimize the outcome for a specific grid. This is generally done by collecting a handful of high magnification images for a variety of ice thicknesses and assessing which holes have the highest probability of yielding intact particles in a good distribution (see Fig. 1d). Once that has been determined, the software can be set up for full automation and generally be left unattended for hours to several days, depending on how much data needs to be acquired. The data collection software moves sequentially to each selected square and acquires high magnification images at targets within the holes selected from that square. A variety of housekeeping tasks (focusing, checking drift, aligning the energy filter, pausing to allow for cryogen fills) are taken care of automatically as needed.

There are several automated data collection packages currently available both from academic groups and commercial entities. On the academic side, there is Leginon (Suloway, Shi, 2009) and SerialEM (Schorb, Haberbosch, 2019), while the commercially available software includes EPU (Thompson, Iadanza, 2019) for Thermo Fisher Scientific microscopes, JADAS (Zhang, Nakamura, 2009) for JEOL microscopes, and Latitude for microscopes with Gatan cameras. While these packages differ somewhat in how their workflows are set up and the degree of automation, all function similarly and are capable of supporting the acquisition of thousands of images per day.

There are many bottlenecks that slow down the single particle data collection workflow. Several of these concern the need for operator supervision and intervention in screening grid conditions or in setting up and managing a long data collection session. It was the general opinion of the group that met together for this workshop that “smarter” software could further reduce the need for hands-on operator time just as the current automated software packages essentially replaced manual data collection about 10 years ago. In the scenarios discussed at the workshop, we all understood “smart” to mean “capable of some independent action” rather than “quick-witted intelligence,” as the current state of the various software options use a combination of traditional algorithmic approaches and machine learning. We anticipate that true “machine intelligence” may ultimately obviate the need for human operation and supervision of single particle data collection. However, this has yet to be achieved.

Automation of data collection is challenging, because the conditions that provide optimal data differ between samples, preparations, and project goals. There is heterogeneity at the low- and medium-magnification levels that complicates the process of identifying regions-of-interest (ROIs) even prior to assessing data quality. This requires operators to manually tune parameters for the ROI detection algorithms currently in use and then to examine high magnification micrographs to correlate lower magnification ROIs to ideal ice thickness and other sample-specific conditions. Machine learning approaches are thus required at two levels to automate screening and data collection: (i) identifying viable collection locations in low- and medium-magnification images quickly and robustly across different grid types, preparation methods, and microscopes, and (ii) optimally ordering those targets and altering collection strategies to acquire the best data as quickly as possible, learning about which regions contain the best sample as collection proceeds.

2. CryoEM grid screening

While the ultimate goal of automation is to achieve full independence from the need for operator input, one of the most urgent goals identified at the workshop was for reducing operator time required for grid screening. Regardless of the software used for data collection, one of the main limiting factors for throughput is finding “good” grids and finding “good” areas on them for data collection. There can be substantial variability in sample quality from grid to grid even when the same sample is prepared under nearly identical conditions (Sgro and Costa, 2018). What this often means - from a practical perspective - is that many grids must be screened before identifying one where the sample is well-preserved and particles are well distributed in thin ice over sufficiently large areas that will yield enough images to reconstruct a high-resolution 3D map.

Screening can be quite a laborious process, as the grid must be assessed at both low magnifications and high magnification. Low

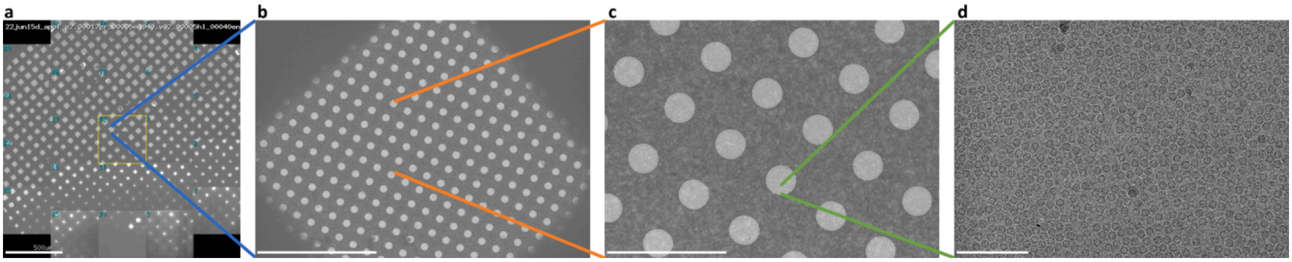


Fig. 1. a. Overview atlas of an EM grid (nominal magnification of 1,550x for each of the 23 atlas grid tile images with a pixel size of 4160 Å). Scale bar is 500 μm. b. Square level image showing an individual square (nominal magnification of 940x with a pixel size of 379 Å). Scale bar is 20 μm. c. Hole level image showing holes in the support film (nominal magnification of 3,600x with a pixel size of 99 Å). Scale bar is 5 μm. d. Final high magnification image showing ice of ideal thickness and good particle distribution (nominal magnification of 105,000x with a pixel size of 0.844 Å). Scale bar is 100 nm.

magnification images can reveal which areas are “bad” for data collection because the ice is too thick, is dried out, or is cracked, or if there is a sufficiently large area with potential for high-resolution data collection. However, high magnification images must also be acquired to assess the quality of the sample itself. During the process of vitrification, the sample is affected by contact with the air–water interface which can lead to denaturation, aggregation, or preferred particle orientation (Noble et al., 2018; D’Imprima, Floris, 2019). While the thinnest possible ice is preferred to maximize contrast, many particles are not stable in very thin ice. As a result, it is often necessary to try several different preparation conditions in order to produce an optimal grid. Conditions typically varied to achieve this goal often include alterations to sample concentration, changes in grid type (gold vs carbon, hole size and geometry), different buffer conditions (pH, salts, sugars, cryo-protectants), alternative instruments used for vitrification (e.g. Vitrobot, Leica EM GP2, chameleon, manual plunger), and/or a suite of detergents or substrates which may help alleviate issues associated with particles interacting at the air–water interface. It is difficult to predict *a priori* how a new sample will behave for a given set of grid preparation conditions. In fact, even for known samples, grid quality can vary even when using the same sample and conditions.

The ultimate result is that for a new sample it is common that many grids need to be “screened” to determine the optimal conditions for high resolution data collection. Practically speaking, many grids will need to be rapidly screened, and an autoloader system makes this process efficient. A 200 keV system is typically used for this purpose with the principal goal being to ensure that particles are embedded and intact within vitreous ice. Projects may spend a significant percentage of time at this stage.

Once a new sample passes these initial assessments, a second screening process may ensue as the grids are optimized for an efficient high resolution data collection session. Small datasets (hundreds of images) may be collected to sample the grids to determine which are most suitable for high resolution collection. Typically, the data will be assessed by calculating 2D class averages and reconstructing preliminary 3D maps.

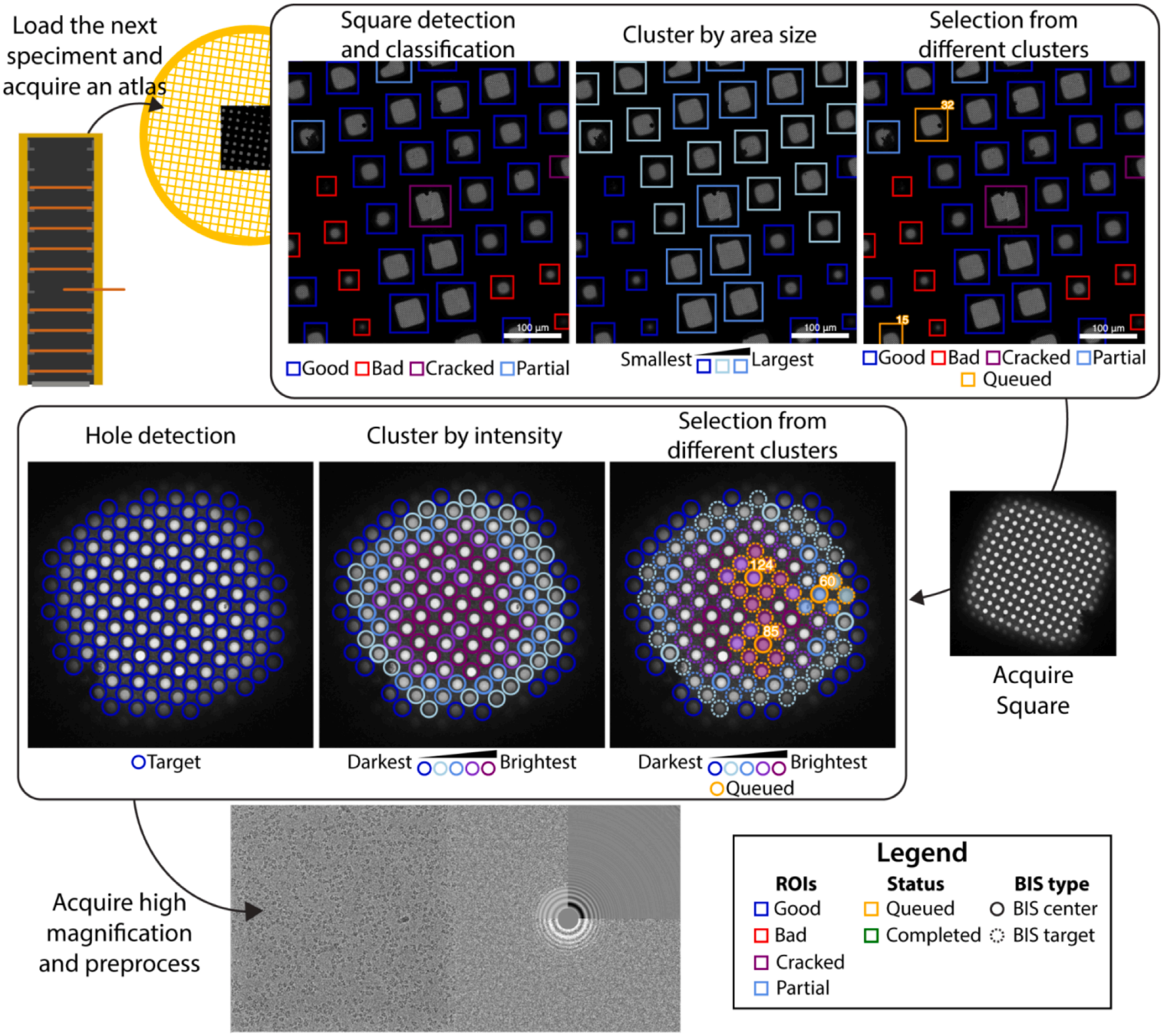

Load the next speciment and acquire an atlas
Square detection and classification
□Good □Bad □Cracked □Partial
100 µm
Cluster by area size
Smallest □□□ Largest
Selection from different clusters
32
15
□Good □Bad □Cracked □Partial □ Queued
Acquire Square
Hole detection
○Target
Cluster by intensity
Darkest ○○○○○ Brightest
Selection from different clusters
124
60
65
Darkest ○○○○○ Brightest ○Queued
Acquire high magnification and preprocess
Legend
ROIs
□ Good
□ Bad
□ Cracked
□ Partial
Status
□ Queued
□ Completed
BIS type
○ BIS center
◌ BIS target

Fig. 2. SmartScope automated workflow for grid screening. The main functions essential for imaging such as specimen exchange, montaging, eucentricity and autofocus are automated. The algorithms for area selection (boxes) include a combination of deep learning and conventional image analysis approaches for feature identification, classification and clustering of the targets against different metrics. The resulting layered approach provides a sampling of different targets during screening.

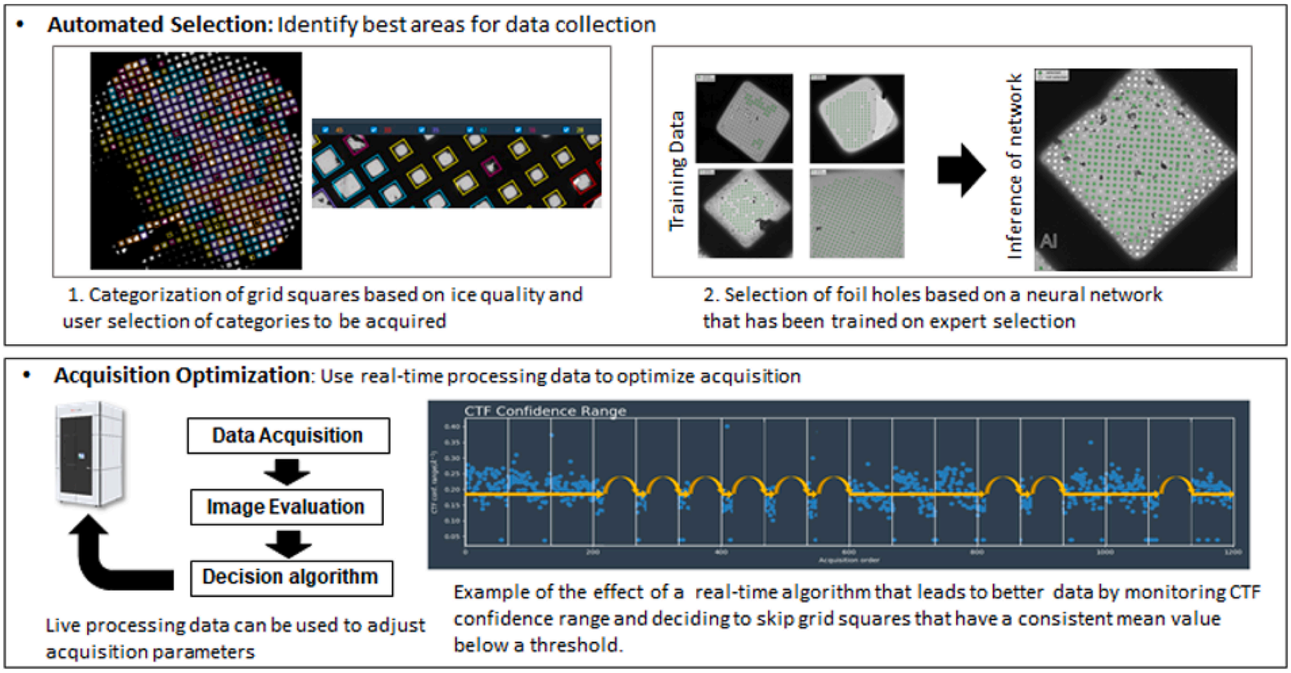
3. Software systems for Smart grid screening and data collection

At the workshop, four different “smart” software packages were presented and discussed. These are laid out in detail in other papers as referenced but a brief overview is provided here.

SmartScope is a software for automated specimen screening in cryoEM based around the data acquisition software SerialEM (Schorb, Haberbosch, 2019). It provides an automated workflow for imaging at multiple magnifications and a web user interface to access results (Bouvette, Huang, 2022). It leverages deep learning algorithms for feature recognition and classification along with clustering methods to provide adequate sampling of a variety of areas across the grid. The workflow proceeds as follows (Fig. 2): 1) A microscopy session is created by filling out basic information about the imaging settings and specimens; 2) The session is started, which triggers SmartScope to connect to the microscope via SerialEM’s python API and start the workflow; 3) For each specimen, the grid is automatically loaded into the column of the microscope; 4) A low magnification atlas is acquired and the squares are identified, classified, and clustered according to their size, from which a subset is selected; 5) A square is acquired and holes are automatically identified and clustered based on their signal intensities and a subset is selected; 6) Finally, high-magnification images of the selected areas are recorded and pre-processed to provide basic quality metrics such as CTF fit. As a session is progressing, the webpage updates in real-time and allows remote interaction with the microscope where area selection can be modified and data annotated. SmartScope was built with a modular design which will allow new and existing algorithms to be integrated as plugins and protocols to be created for new applications and sample types. The interface facilitates access to the instruments and results without granting full access to the instrument to every user. The goal of SmartScope is to assist microscopists by automating the specimen screening process in cryoEM and to lower the barrier of adoption for cryoEM.

Smart EPU is built on Thermo Fisher’s EPU, a single particle data acquisition application that focuses on automation, guidance and user experience. EPU aims to enable more users to benefit from cryoEM by making the workflow more efficient and lowering the entry barrier for new users. Smart EPU is a system of software programs created around EPU that allows for further automation by using machine learning and on-the-fly feedback loops. Smart EPU provides an open interface to enable the development of algorithms that influence the set-up of an experiment or connect to an ongoing acquisition and optimize it in terms of efficiency and quality (Fig. 3). EPU already contains a set of built-in classical algorithms that help users select grid squares and holes to be acquired. For example, a classification method that categorizes and suggests similar-looking squares and hole selections is assisted by automating routine tasks such as finding all holes, removing holes that are located close to grid bars and selecting holes using an intensity-based ice thickness filter. On top of these routines, a new machine learning algorithm has been added to automatically recognize and discard sub-optimal areas that would lead to inferior micrographs. The neural network that powers this smart filtering makes the selection based on the encoded knowledge of experts as it has been trained with numerous selections from previous experiments. These algorithms can be combined with the workflow of EPU for fully automated acquisitions on multiple grids (EPU Multigrid) in order to set-up screening or high-resolution experiments with limited user interaction. Once data acquisition starts, algorithms can leverage, in real-time, the data and metadata produced by the microscope and use the API to fine tune parameters. Smart EPU includes algorithms that digest the results of the EPU Quality Monitor (EQM) routines of motion correction and CTF determination in order to adjust on-the-fly and optimize parameters such as microscope focus, stage stabilization time, or skip an area that consistently produces micrographs with inadequate CTF resolution estimations. For example, Fig. 3 shows the CTF confidence range calculation for 1,200 images out of 18 grid squares on an Apoferritin sample acquired on a Tundra microscope. Even with such a standard sample, not all grid squares are equal in quality, and without any intervention, a lot of data might be thrown away at the end of the session. When Smart EPU algorithms are applied on a similar dataset, the grid squares noted with an arrow on Fig. 3 will be skipped. This leads to better data quality and an increase in throughput. The prediction of suitable areas to acquire together with automatic on-the-fly adjustments should improve the acquisition of good quality images while lowering the time investment of the operator.

Smart Leginon integrates the Leginon data collection workflow with a machine learning program, Ptolemy (Kim et al., 2021), to provide a solution for fully automated, high-throughput grid screening (Cheng,



- **Automated Selection:** Identify best areas for data collection

1. Categorization of grid squares based on ice quality and user selection of categories to be acquired

2. Selection of foil holes based on a neural network that has been trained on expert selection

- **Acquisition Optimization**: Use real-time processing data to optimize acquisition

Live processing data can be used to adjust acquisition parameters

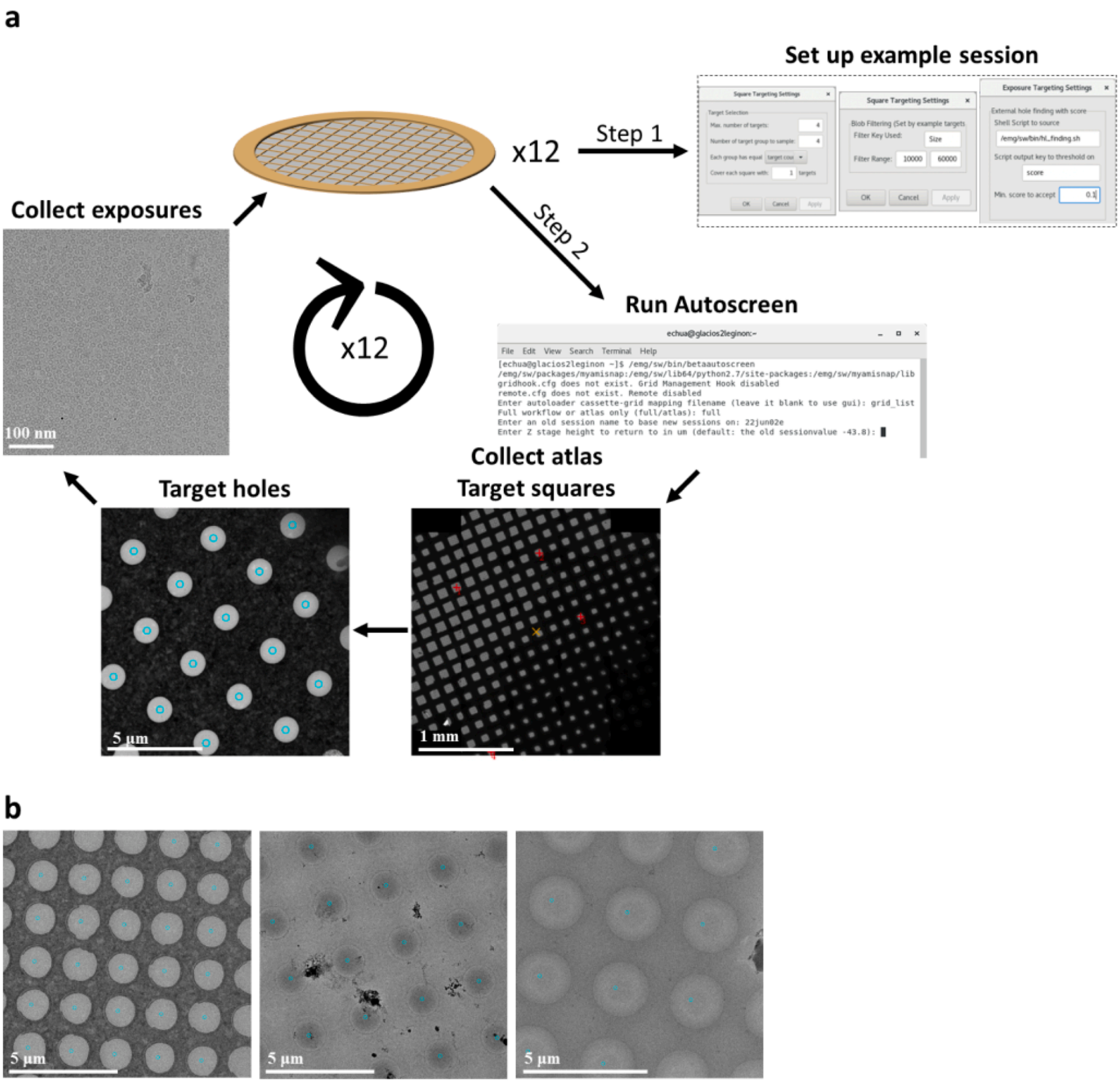
Example of the effect of a real-time algorithm that leads to better data by monitoring CTF confidence range and deciding to skip grid squares that have a consistent mean value below a threshold.

Fig. 3. Automation routines of Smart EPU Software for algorithm assisted target selection and live experiment optimization.

Kim, 2022). Ptolemy provides robust automated square and hole finding and scoring. For example, given a pre-defined square area range, the highest scoring class provides reliable sampling of grid squares on unknown grids in multi-grid screenings. Similarly, Ptolemy's parameterless hole localization is not affected by differences in grid support material, such as gold versus carbon, or grid geometry (hole size and spacing) (Fig. 4). At the start of screening, the only inputs required are the number of square groups to sample, the number of squares to image in each group, a filter "blob" size used by the Ptolemy square finder, and a threshold for the Ptolemy hole finder. The session setup only takes a few minutes, after which it runs unattended for a set of grids that can be automatically exchanged. Initial parameter settings can be re-used so that operator setup can be eliminated in future sessions. As an example, Smart Leginon automatically screened an 11 grid cassette, reducing required operator time from ~ 6 h to < 10 min. This significantly reduces the burden on microscope operators and makes more efficient use of microscope time.

CryoRL has not yet been integrated into a data acquisition program. Its goal is to enable increased data collection efficiency, while eliminating the need for subjective decisions and user intervention, by maximizing the number of micrographs with high-resolution information that are acquired over a given time duration (Fan, Li, 2022, Li, Fan, 2022). By combining supervised classification or regression and deep reinforcement learning, cryoRL provides a new framework for cryoEM data collection (Fig. 5). It aims to return the quality predictions for lower magnification hole level images and also plan the trajectory for higher magnification data acquisition, by balancing the trade offs in the time for stage movement and Z-height adjustment with data quality. The RL network combined with an offline hole classifier or regressor has shown improved performance compared to human users, including cryoEM experts (Li, Fan, 2022). The reward function of the reinforcement learning model makes cryoRL flexible enough to be extended to both data screening and collection over a range of microscopes and cameras.

SmartScope, Smart Leginon, and Smart EPU can all be used to perform triage screening on a cassette of grids. All three packages are able to return an atlas of each grid but Smart Leginon and SmartScope



Max. number of targets: 4
Number of target group to sample: 4
Each group has equal: target cou
Cover each square with: 1 targets

Filter Key Used: Size
Filter Range: 10000 60000

Shell Script to source: /emg/sw/bin/hl_finding.sh
Script output key to threshold on: score
Min. score to accept: 0.1

```
echua@glacios2leginon:~
[echua@glacios2leginon ~]$ /emg/sw/bin/betaautoscreen
/emg/sw/packages/myamisnap:/emg/sw/lib64/python2.7/site-packages:/emg/sw/myamisnap/lib
gridhook.cfg does not exist. Grid Management Hook disabled
remote.cfg does not exist. Remote disabled
Enter autoloader cassette-grid mapping filename (leave it blank to use gui): grid_list
Full workflow or atlas only (full/atlas): full
Enter an old session name to base new sessions on: 22jun02e
Enter Z stage height to return to in um (default: the old sessionvalue -43.8): 
```

Fig. 4. Smart Leginon workflow and functionalities. **(a)** Smart Leginon Autoscreen can screen a multi-grid session fully unattended after setting up several parameters in an example session as shown in Step 1 and then executing a short command line script as shown in Step 2. Subsequently, all grids will be screened according to the example session parameters. **(b)** Smart Leginon hole targeting with Ptolemy provides robust, parameter-less hole targeting, allowing for the Autoscreen process to work across many grid types and characteristics.

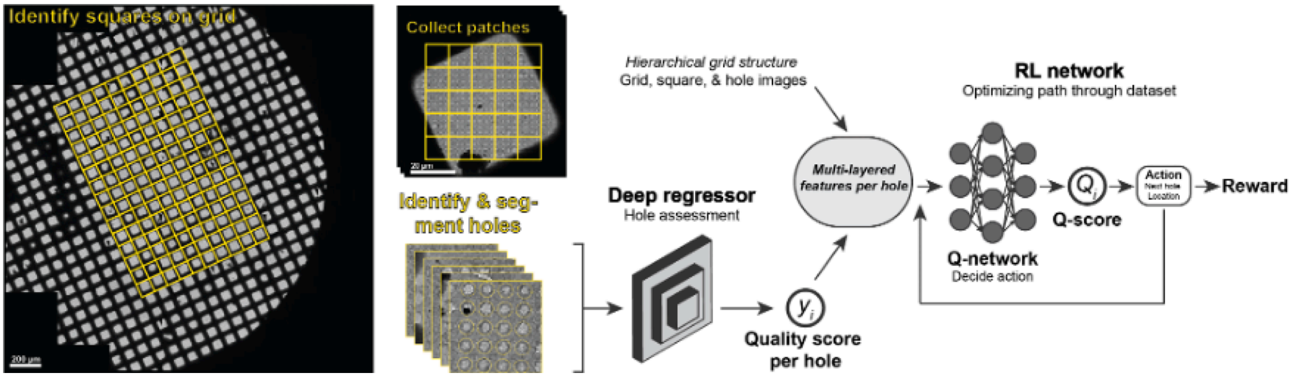


Fig. 5. CryoRL uses medium magnification hole images as candidates. A pre-trained regressor is used to predict the quality scores (for example, CTFMaxRes) for the candidate holes. The output quality scores are combined with the hierarchical grid structure to engineer features which are used as the input for the reinforcement learning (RL) network. The RL network will output an optimized trajectory for data collection, balancing the tradeoff between the time cost of stage movement and the predicted data quality.

also acquire intermediate and high magnification images of each grid while, currently, Smart EPU requires a second pass to select areas either for continuing screening or for automated high-resolution data collection. Ultimately all three packages (and others that might also be developed) are likely to converge to similar functionality.

4. Discussion of current limitations and needs for smarter data collection

Below we summarize some of the topics that were discussed at the workshop.

5. Active learning and fine-tuning

The ideal high-magnification exposure collection locations for a given grid are difficult to know *a priori*, as they depend on particle quality, grid type, and grid preparation method. Indeed, one of the goals of screening is to explore squares and holes so that a human operator can determine the best collection locations for a particular grid. Existing methods for automated targeting use pre-trained, fixed machine learning models trained on static, pre-curated datasets. These models are then applied in a “one-size-fits-all” fashion to all cryoEM grids. To fully automate data collection in a manner that works robustly for any session, a method of refining a model’s selection to be specific to the current session is required. This can be accomplished via fine-tuning of global models, or by using an additional model that is trained only on incoming data for the session. This problem is amenable to an active learning approach, whereby the model attempts to quickly learn by selecting the most informative data points for maximizing a reward function. Successful implementation of an active learning and fine-tuning approach to data collection could make data collection significantly more time-efficient and eliminate the need for human input during the data collection process.

6. Assessment Metrics: How can “good” regions with “good” particles be recognized?

Although it is difficult to determine whether a given acquisition of images will lead to high resolution structures, it is more straightforward to identify images in which the data quality is compromised. The following are some of the indicators of potentially problematic datasets that can be identified during an acquisition session:

- Large global shifts between movie frames. These can occur due to instability in the grid (cracks, tears) or the microscope.
- Significant amount of astigmatism in the power spectrum of the whole micrograph or in local patches. This can arise from a poorly aligned microscope or contamination on the grid.
- The CTF is not fitted well to sufficiently high resolution. A CTF fit that does not extend to high resolution can be due to microscope aberrations, ice that is too thick, or too few particles in the ice contributing to the signal.
- Particle density is too high. If particles are too densely packed they may be overlapping in 2D projection which might compromise alignment and averaging.
- Particle density is too low. Frame alignment and CTF fitting may be compromised and data collection will be inefficient.
- Highly preferred orientation. Assessing this aspect requires processing the data either in 2D or 3D. The preferred orientation may be correlated to specific ice thicknesses.

The quality of a data acquisition is not determined by the quality of only a few micrographs and the indicators discussed need to be collectively evaluated to estimate the proportion of potentially problematic micrographs. These evaluations may then drive decisions to skip regions of the grid or skip to a new grid.

7. Metrics for quality

Several metrics for measuring the quality of high-resolution micrographs have been discussed. A major consideration for any metric used for automation is its ability to be computed efficiently and without user input. This has driven interest in methods like CTF resolution (Rohou and Grigorieff, 2015) and machine learning algorithms that can reproduce human qualitative judgements (Li, Cash, 2020). However, these approaches do not quantify the end goal of data collection, which is to reconstruct a high resolution structure. They also can be unreliable as in the case where high resolution from the CTF may be attributed to the edge of the carbon film being included into an image.

Low throughput methods of evaluating data quality usually require calculating the resolution of 2D class averages or 3D reconstructions, assessing the angular coverage, calculating a Guinier plot falloff B factor of the reconstructed map, etc. Evaluating these metrics on-the-fly requires building a fully reliable, fully automated data analysis pipeline. This requires high fidelity particle picking, rapid algorithms for 2D classification and 3D reconstruction, and the ability to automatically assess 2D class averages or 3D structure quality.

Some progress has been made towards evaluating 2D class averages (Li et al., 2020), but the ultimate value of these methods in guiding data collection strategies has not yet been determined. The resolution of 3D structures is susceptible to overfitting and inclusion of particles in 3D reconstructions does not necessarily reflect particle quality, as different reconstruction methods may select different subsets of particles contributing to the final map and yet achieve equivalent resolutions (Sanchez-Garcia, Segura, 2018). An interim solution to a full analysis is to simply count the number of particles in each image, but this still requires reliable user-free particle picking methods and does not provide any feedback on the quality of the picks.

In general, good protein structures should provide 2D classes that exhibit high-resolution details (e.g. evidence of secondary structural features like alpha helices) during early-to-mid stages of analysis. If the images are hard to align and fail to produce high-quality 2D classes, they may also be hard to align in order to produce high-quality and interpretable 3D maps. For particles lacking high symmetry there should also be a range of classes of different appearances if preferred orientation is not present.

Most discussions involving automated assessment of particle quality typically begin with the following question: *"How can we design software to emulate, or improve on, the decisions made by humans?"* Human assessment of particle quality typically relies on prior knowledge of expected particle size, shape, and morphology (which might be deduced from the literature), as well as from prior biochemical, biophysical, and/or structural analyses like homologous structures or negative stain data. Thus, future advancements in computational algorithms for assessment of 2D class averages and/or 3D reconstructions may need to incorporate various levels of operator-supplied input parameters defining the features that are *expected* to be observed for a given sample. Some of the possible operator provided metrics might include (note that all will require an efficient on-the-fly processing suite):

- Prior structures (PDB/EMDB) that might be used to generate 2D projections to compare to acquired data to inform automated assessment of orientational distribution and/or expected particle morphology.
- Negative stain 2D class averages that might help assess if the particles are of the expected size, shape, and morphology in vitreous ice. Comparisons would need to take into account differences in resolution and preferred orientation.
- Pre-calculated cryoEM 2D classes could be used in a similar way to the negative stain classes. This might be more reliable than negative stain 2D averages and could be used during iterative assessment of cryoEM sample quality throughout later stages of screening.
- Basic expectations of particle geometry including particle shape, size, symmetry could be used to assess 2D class averages or initial 3D reconstructions.

Given that each unique sample which is imaged during the screening phase may have different amounts of preexisting information readily available, we propose that all of the above options could be included as optional input parameters, and the operator can decide which to use. Downstream processing and assessment would then need to adapt assessment metric algorithms to account for the different type(s) of inputs potentially provided by the operator.

8. What is the lowest magnification image that can be used to determine if holes are "good"?

Good holes are currently determined by human operators who look for particles and assess image quality in the high-magnification exposures. However, it is possible that a computer vision algorithm could directly detect the presence or absence of particles in holes at lower magnifications, without needing to take a high magnification exposure. The lowest magnification required for a computer-vision algorithm to reliably detect particles has thus far not been well characterized - it almost certainly depends on the size and density of the particles, and is likely upper-bounded by a magnification where the particle dimensions are more than two pixels. Experimental characterization of the lowest magnification required for a particle could potentially allow for the use of more efficient magnification levels during both screening and data collection. For example, images could be taken at a currently unutilized magnification level, where multiple holes are simultaneously visible, with the computer vision algorithm explicitly detecting particle presence or absence at this magnification and directing high magnification exposure collection accordingly. Optimization of the magnification levels used could increase microscope throughput, as well as the amount of high-quality data collected within a given session.

9. Publicly available labeled or non-labeled datasets for ML development and foundational learning

A key requirement for any machine learning algorithm is the availability of representative data of the kind of images that will be encountered in a production environment. ImageNet (Krizhevsky, Sutskever, 2012) with 14 M images is one of the best known publicly available datasets, and there are a plethora of datasets addressing specific tasks (handwriting recognition, human actions, hand gesture recognition, natural language processing, etc.). The public availability of these datasets has been of paramount importance for the development of new methodological ideas, especially those using deep learning that are very data intensive. The variability of kinds of images expected in cryoEM, especially if focused on single particle data acquisition, should not be as large as with natural images. There are several varying factors such as the grid support (copper or gold), mesh size, distribution and size of the holes, presence or not of a carbon or graphene layer, presence of contaminants, ice crystals, aggregation, and varying ice thickness. Although the nature of contaminants can be highly varied, still their complexity and variety should be orders of magnitude below those of natural images. This means that, possibly, a few hundreds or thousands of examples of images at a variety of magnifications may be enough to delineate the variability of the images obtained at a cryoEM facility.

To support a useful cryoEM data depository, various metadata including pixel size, grid mesh size, and hole size should accompany each of the deposited images. Image examples should be available at a variety of magnification levels (low, medium, high magnification). It is not necessary that magnification levels be standardized, variation will help to cover the whole resolution range, but accurate pixel size is critical. It is not a hard requirement that all images in the depository have counterparts at different magnification levels but it will be helpful if many of them have such correspondences. Ideally, not only the correspondence but also the locations of the high magnification images within the lower magnification images will be helpful.

Regarding the labeling of such publicly available images, again it is not necessary that all the images have all the labels. This is a situation known in machine learning as partial or semi-supervised labeling and a typical approach is to use the labels available to infer the missing labels. The underlying algorithm can give different weights to the labels depending on whether they have been provided by the user or by an algorithm. Some of the labels that would be useful might include: local ice thickness, local quality of the image (note that this quality may be different when evaluating squares, holes, or particles), defocus, location of contaminants, ice crystals or carbon edges, location of particles in high magnification images, CTF resolution estimation of high magnification images, etc. In any case, what is important is that the definition of these labels is unambiguous and consistent across datasets deposited by different laboratories. The format of the metadata thus needs to be agreed upon and ideally built on a well-established community standard as for example the EMDB and EMPIAR databases.

10. Standard ML interfaces to data collection packages

At present there are several programs interfacing the microscope including SerialEM, Leginon, EPU, JADAS, and Latitude. They all support the same kind of operations (changing magnification, setting focus, moving the stage to specific locations, exposing the sample for a given amount of time, etc.). These are the basic operations required to acquire images in an electron microscope. It would be very useful if they all offered a common programmatic, driver interface which would unify basic microscope operation. This would foster the development of sophisticated, machine learning based computer programs that could accelerate the development of cryoEM as an instrumental technique in structural biology.

11. Extending smart data collection beyond single particle data acquisition

Optimizing targeting locations for single particle data collection does not automatically translate to other sample and collection modalities, such as cell specimens intended for cryoelectron tomography (cryoET), microED samples, negative stain grids, liposome characterization, etc. Typically, these use cases will require additional prior knowledge for targeting parameter optimization and the areas of interest may have undesirable characteristics relative to SPA grids, e.g. cryoET collection on a cell specimen is often performed in darker areas on the grid rather than brighter areas. To enable the extension of smart SPA collection software to other sample and collection modalities, these algorithms should be designed to agnostically localize potential areas of interest and to not impose built-in strong priors for SPA collection. We foresee that subsequent iterations of smart collection software will extend beyond SPA, thereby increasing efficiency, reducing operator burden, and decreasing user-bias across all cryoEM collection modalities.

12. Conclusion

The Smart Data Collection for CryoEM Workshop provided an opportunity to discuss many of the challenges of developing the next generation of cryoEM data collection software that incorporates machine learning. These discussions do not often take place at scientific meetings and are not typically the subject of journal publications. The workshop spurred discussion between several development groups and likely accelerated detailed publications and release of the software for each of the packages. This review documents these discussions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

We are grateful for financial support for this workshop from NIH NIGMS GM103310, Simons Foundation (SF349247), and Thermo Fisher Scientific.

References

Bouvette J, Huang Q, Riccio AA, Copeland WC, Bartesaghi A, Borgnia MJ. Automated systematic evaluation of cryo-EM specimens with SmartScope. Elife. 2022 Aug 23; 11:e80047. doi: 10.7554/eLife.80047. PMID: 35997703; PMCID: PMC9398423.

Cheng A, Kim P, Kuang H, Mendez JH, Chua EYD, Maruthi K, Wei H, Sawh A, Aragon MF, Serbynovskyi V, Neselu K, Eng ET, Potter CS, Carragher B, Bepler T, Noble AJ. Fully Automated Multi-Grid Cryo-EM Screening using Smart Leginon. bioRxiv. 2022: 2022.07.23.501225.

D'Imprima, E., Floris, D., Joppe, M., Sanchez, R., Grininger, M., Kuhlbrandt, W., 2019. Protein denaturation at the air-water interface and how to prevent it. Elife. 8. PMCID: PMC6443348.

Fan Q, Li Y, Yao Y, Cohn J, Liu S, Vos SM, Cianfrocco MA. CryoRL: Reinforcement Learning Enables Efficient Cryo-EM Data Collection. arXiv preprint arXiv: 220407543. 2022.

Kim PT, Noble AJ, Cheng A, Bepler T. Learning to automate cryo-electron microscopy data collection with Ptolemy. arXiv preprint arXiv:211201534. 2021.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 25.

Li, Y., Cash, J.N., Tesmer, J.J., Cianfrocco, M.A., 2020. High-throughput cryo-EM enabled by user-free preprocessing routines. Structure. 28 (7), 858–869 e3.

Noble AJ, Dandey VP, Wei H, Brasch J, Chase J, Acharya P, Tan YZ, Zhang Z, Kim LY, Scapin G, Rapp M, Eng ET, Rice WJ, Cheng A, Negro CJ, Shapiro L, Kwong PD, Jeruzalmi D, des Georges A, Potter CS, Carragher B. Routine single particle CryoEM sample and grid characterization by tomography. Elife. 2018;7. PMCID: PMC5999397.

Rohou, A., Grigorieff, N., 2015. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. J Struct Biol. 192 (2), 216–221. PMCID: PMC6760662.

Sanchez-Garcia, R., Segura, J., Maluenda, D., Carazo, J.M., Sorzano, C.O.S., 2018. Deep Consensus, a deep learning-based approach for particle pruning in cryo-electron microscopy. IUCrJ. 5 (6), 854–865.

Schorb, M., Haberbosch, I., Hagen, W.J.H., Schwab, Y., Mastronarde, D.N., 2019. Software tools for automated transmission electron microscopy. Nat Methods. 16 (6), 471–477. PMCID: PMC7000238.

Sgro, G.G., Costa, T.R.D., 2018. Cryo-EM Grid Preparation of Membrane Protein Samples for Single Particle Analysis. Front Mol Biosci. 5, 74. PMCID: PMC6090150.

Suloway, C., Shi, J., Cheng, A., Pulokas, J., Carragher, B., Potter, C.S., Zheng, S.Q., Agard, D.A., Jensen, G.J., 2009. Fully automated, sequential tilt-series acquisition with Leginon. J Struct Biol. 167 (1), 11–18. PMCID: PMC2724967.

Thompson, R.F., Iadanza, M.G., Hesketh, E.L., Rawson, S., Ranson, N.A., 2019. Collection, pre-processing and on-the-fly analysis of data for high-resolution, single-particle cryo-electron microscopy. Nat Protoc. 14 (1), 100–118.

Zhang, J., Nakamura, N., Shimizu, Y., Liang, N., Liu, X., Jakana, J., Marsh, M.P., Booth, C.R., Shinkawa, T., Nakata, M., Chiu, W., 2009. JADAS: a customizable automated data acquisition system and its application to ice-embedded single particles. J Struct Biol. 165 (1), 1–9. PMCID: PMC2634810.