Trends in
Biochemical Sciences

CellPress

Special Issue: Pushing boundaries of cryo-EM

Forum

Cloud computing platforms to support cryo-EM structure determination

Yilai Li[1] and
Michael A. Cianfrocco[1,*]

Leveraging the power of single-particle cryo-electron microscopy (cryo-EM) requires robust and accessible computational infrastructure. Here, we summarize the cloud computing landscape and picture the outlook of a hybrid cryo-EM computing workflow, and make suggestions to the community to facilitate a future for cryo-EM that integrates into cloud computing infrastructure.

Cryo-EM is a critical tool for the future of structural biology

Single-particle cryo-EM continues to grow as a mainstream structural biology technique. Key to the growth of cryo-EM has been continued advancements in instrumentation, algorithms capable of processing low signal-to-noise ratio particles, and the growth of cryo-EM-specific atomic modeling tools. The power of cryo-EM is highlighted in response to the SARS-CoV-2 pandemic, where cryo-EM was used to determine the structure of the spike protein alone [1], in complex with Fabs [2], or within intact virions [3].

In parallel to the rapid rise of cryo-EM, there has been a concomitant growth in cloud-computing-based platforms (Box 1). The ever-expanding presence of both cryo-EM and cloud computing will lead to their intersection as the future of structural biology moves to remote, cloud-based high-performance computing (HPC) resources. In this Forum, we highlight areas and pain points related to the spread of cryo-EM into cloud environments.

Computational challenges faced by cryo-EM practitioners

Despite the continued advancements in microscopes, detectors, and algorithms, cryo-EM remains a big data, life sciences research endeavor. For instance, cryo-EM practitioners require routine access to tens to hundreds of terabytes of storage per research group and petabytes at institutional levels. This large data footprint is separate from the computational infrastructure needed to embark on a nondeterministic analysis of cryo-EM data, requiring upwards of hundreds of jobs per dataset. Compared to other fields that demand HPC, such as molecular dynamics simulations, each individual cryo-EM job is light-weighted. While it might have taken a few days to run a single job on a central processing unit (CPU) cluster a few years ago, with the development of new algorithms and graphics processing unit (GPU) acceleration, most of the jobs can be completed within several hours. The rise of shorter job times is due to continued advances in algorithms and compute acceleration using CPU and GPU-specific compilers.

The challenge of cryo-EM computing is due to a variety of job types and the nondeterministic nature of data processing. First, a typical cryo-EM workflow consists of over ten different job types, such as motion correction, CTF-estimation, 3D refinement, and particle polishing, whose optimal computing infrastructure differs. This situation leads to computing infrastructure hurdles for many researchers, as the optimal infrastructure for some job types may not be available to them. Second, although it seems that the total time of a complete workflow (from the data outputted by the microscope to a final structure) is not long, in practice, it is common that a scientist will run hundreds of jobs on a single dataset to achieve the best structure. In other words, each individual cryo-EM job may not take long, but hundreds of such light-weighted jobs with different optimal computing infrastructures are typically required for one promising dataset.

Cloud computing resources for cryo-EM

We believe that the flexibility of cloud computing makes it a valuable tool for cryo-EM computing [4]. Since cryo-EM computing consists of many light-weighted jobs that need to be run separately from either a CPU cluster or a multi-GPU workstation, the idle time is usually significant. Especially because not all job types are GPU accelerated, the usage of GPU is usually relatively low. Cloud computing can address these hurdles precisely. On the one hand, the users would only pay for the 'on' computing time. On the other hand, cloud computing can offer flexible computing services for each different job type. For example, job types that only use CPU can use a cheaper CPU infrastructure, increasing efficiency and reducing the cost.

Box 1. What is cloud computing?

Cloud computing refers to the utilization of remote cyberinfrastructure resources. This broad definition encompasses the public cloud, private cloud, government supercomputers, and any other remote computing resources. The public cloud refers to services offered by companies such as Amazon Web Services and Google Cloud Platform that are available to any paying customer. Private cloud refers to any restricted access remote computing service. Government supercomputers offer free or subsidized computing services for the academic community such as the National Science Foundation funding the Extreme Science and Engineering Discovery Environment (XSEDE)[vii] to support supercomputers across the United States.

To further reduce idle time and cloud computing costs, we believe that intelligent pipelines for streamlining different job types will enable flexible computing. For example, by combining preprocessing steps with deep learning-based data assessments, we designed a pipeline that outputs expert-level curation of particle stacks for subsequent cryo-EM analysis [5]. Such pipelines can significantly reduce idle time by seamlessly knitting the different types of jobs together. Decision-making in such pipelines can either be heuristic or captured by a pretrained machine learning model. Besides, some initial steps in the data processing can be done 'on the fly' locally during data acquisition for real-time feedback before transitioning to cloud-based infrastructure.

The large data storage footprint of cryo-EM data makes cloud storage a more complicated decision than computing. For example, typical cloud storage ensures a high degree of data durability (i.e., chance of data loss or corruption) as compared to local storage on machines or external hard drives. With high data durability, the cost per gigabyte of cloud storage is larger than the price of local disks. Another feature to consider is how often the data will be accessed; immediate access will cost more than infrequent access. For cryo-EM practitioners, it is important to consider: (i) funding available to pay for data storage; (ii) regulatory compliance with funding agencies (i.e., do funding agencies require the data to be preserved?); and (iii) dataset value (i.e., how hard would it be to regenerate the dataset?). Taken together, cloud storage typically makes the most economic sense for small amounts of data short-term and larger amounts of data for long-term archival storage.

A final consideration with cloud computing involves data transfer rates between local and cloud storage. Because cryo-EM data are usually in the scale of terabytes (for raw movies) or gigabytes (for aligned micrographs), they require fast transfer between on-premises storage and the cloud, such as 200–300 Mb/s, which can be achieved on 10 Gb networking connections. Even though institutions, such as university campuses, may have fast networking out of campus, moving data within campus from the cryo-EM instrument to the cloud may be a limiting step in data movement workflows. Given this, as structural biologists set up cryo-EM instrumentation and computing infrastructure, it is critical to include network upgrades to ensure fast data transfer.

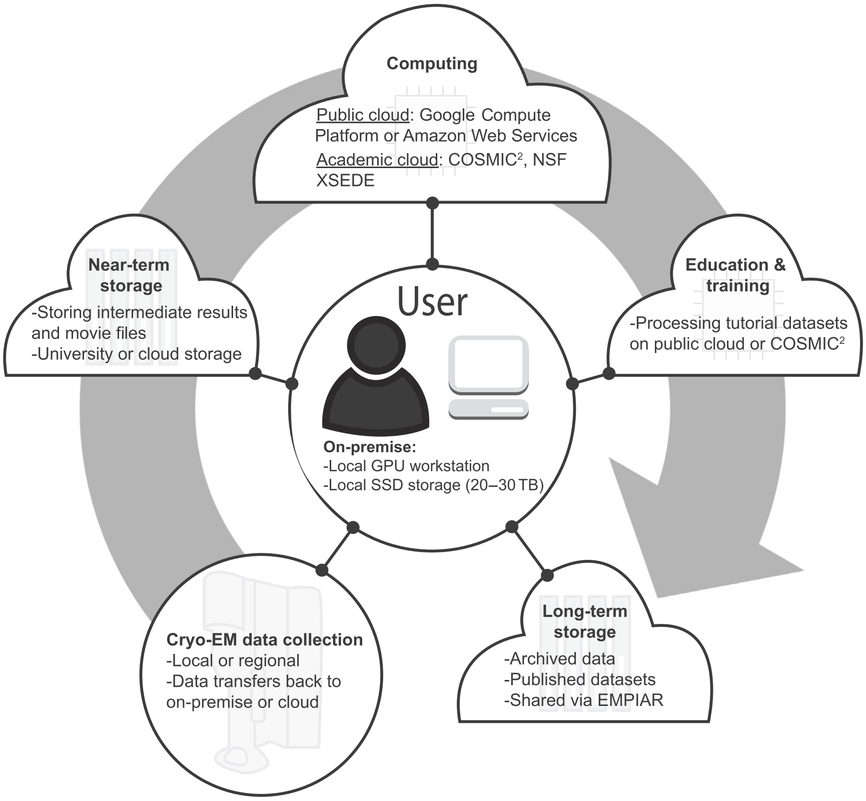
A hybrid future for cryo-EM computing

We believe that the future of cryo-EM computing will be the utilization of hybrid computing resources (Figure 1). In this world, we expect that on-premise machines will serve as the hub around which a hybrid computing ecosystem will assemble to orchestrate data movement, data storage (near term and cold), and accessing flexible software tools. With a modest on-premise workstation, scientists will leverage local storage options at universities or cloud providers to store files that do not need constant access but are not ready for archival (e.g., movies, micrographs, and datasets). This will relieve a local storage burden for the on-premise workstation. In parallel with near-term storage, scientists will also access flexible computing options on public cloud resources or academic platforms such as COSMIC^2 [6], a science gateway for cryo-EM. This will include 'burstable' computing, where jobs can be pushed up to cloud computing platforms when local resources are at 100% utilization, in addition to using specific hardware setups for utilizing software with particular requirements. Finally, these machines will allow access to educational cloud resources to process tutorial and real datasets to learn best practices in cryo-EM[i,ii].

Data security is critical for both on-premises and cloud-environment processing. Utilizing end-to-end encryption will ensure security for data movement, such as that implemented in the Globus data transfer service[iii]. Beyond data movement, utilizing virtual private networks and other security measures will limit the ability of unauthorized access to data and computing. There are academic-oriented cyber security organizations available to the community, such as TrustedCI: NSF Cybersecurity Center of Excellence[iv], which provides consulting and services for data security.

The cryo-EM community needs platforms that allow seamless movement of jobs and data from on-premise to cloud resources to enable this future. We have designed and implemented two different prototypes for this future: cryoem-cloud-tools [7] and COSMIC^2 [6]. cryoem-cloud-tools is a Software as a Service (SaaS) package for moving jobs natively from RELION straight into Amazon Web Services, providing real-time data syncing for users to launch jobs locally, but it ran on the cloud. COSMIC^2 is a Platform as a Service (PaaS) to allow academic users free computing at the San Diego Supercomputer Center.

In summary, the continued growth of cryo-EM across the life sciences will be enabled by interfacing cryo-EM analysis with cloud computing infrastructure. As a community, we need to: (i) standardize software environments for easy deployment via containers; (ii) remove administrative hurdles for accessing cloud resources; and (iii) develop tools for predictable costs for grant budgets. To date, software distribution, installation, and execution rely on individual practitioners and research groups to maintain cyberinfrastructure. While possible for academic research and development, deploying software for widespread community use would benefit from standardization so that the community can overcome software installation hurdles to more easily determine cryo-EM structures. Software containerization – putting software into a virtual machine housed


Computing
Public cloud: Google Compute Platform or Amazon Web Services
Academic cloud: COSMIC^2, NSF XSEDE
Near-term storage
-Storing intermediate results and movie files
-University or cloud storage
User
On-premise:
-Local GPU workstation
-Local SSD storage (20–30 TB)
Education & training
-Processing tutorial datasets on public cloud or COSMIC^2
Cryo-EM data collection
-Local or regional
-Data transfers back to on-premise or cloud
Long-term storage
-Archived data
-Published datasets
-Shared via EMPIAR

Trends in Biochemical Sciences

Figure 1. A hybrid computing workflow for cryo-electron microscopy (cryo-EM). After data collection on local or national facilities, data are shared across networks to local storage for immediate pre-processing and initial analysis on local solid-state disks (SSDs). Considering the large data footprint for raw cryo-EM data, all raw data should be moved to cheaper near-term storage located on university or cloud storage. To process cryo-EM data, cryo-EM users should have an on-premise graphics processing unit (GPU) machine for running local jobs. When on-premise does not provide enough computational resources, users can move their analysis routines to public or academic cloud resources. At the end of a project, cryo-EM data will be stored long-term on cold storage solutions offered by cloud providers. Finally, to train new users, the cloud offers suitable computing to enable hands-on training to educate new entrants into the field.

within another computer – will enable this future, as containers help to standardize software environments. Continued development of container software and container management such as Docker[v] and Singularity[vi] will facilitate this future.

After creating standardized software environments, users will need to access cloud computing resources without administrative burdens. Typically, universities can partner with cloud platforms to facilitate access to public cloud resources such as Amazon Web Services or Google Cloud Compute. While this is helpful, it still leaves researchers without the tools needed to run and manage jobs on the cloud. Developing easy-to-use platforms to manage data and computing instances will facilitate new users to access cloud resources.

Finally, since academic research laboratories operate on predefined grant budgets, cloud computing costs need to be predictable for cost and performance. When developing grant budgets, it would help researchers to have a fixed yearly cost for accessing cloud resources. Importantly, this yearly cost would come with an expectation of a storage footprint and compute performance to ensure that the budgeted funds are being used appropriately. A predictable budget will fall onto universities and cloud providers to develop financial tools to ensure access and cost, without overcharging or underdelivering cloud infrastructure.

We remain optimistic that these hurdles will be addressed to enable continued discoveries in life sciences research. The power of cryo-EM alongside cloud computing will enable efficient data analysis workflows, data management solutions, and educational opportunities that will impact academic and commercial cryo-EM practitioners.

Resources

[i]https://aws.amazon.com/blogs/publicsector/structural-biologists-learning-cryo-electron-microscopy-have-new-educational-resources-powered-by-aws/

[ii]https://aws.amazon.com/blogs/hpc/stion-a-saas-for-cryo-em-data-processing-on-aws/

[iii]www.globus.org/data-transfer

[iv]www.trustedci.org/

[v]https://docs.docker.com/get-started/overview/

[vi]https://sylabs.io/guides/3.5/user-guide/introduction.html

[vii]www.xsede.org/

[1]Life Sciences Institute & Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA

*Correspondence:
mcianfro@umich.edu (M.A. Cianfrocco).

https://doi.org/10.1016/j.tibs.2021.11.005

References

1. Wrapp, D. *et al.* (2020) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 367, 1260–1263
2. Barnes, C.O. *et al.* (2020) SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* 588, 682–687
3. Ke, Z. *et al.* (2020) Structures and distributions of SARS-CoV-2 spike proteins on intact virions. *Nature* 588, 498–502
4. Cianfrocco, M.A. and Leschziner, A.E. (2015) Low cost, high performance processing of single particle cryo-electron microscopy data in the cloud. *Elife* 4, e06664
5. Li, Y. *et al.* (2020) High-throughput cryo-EM enabled by user-free preprocessing routines. *Structure* 28, 858–869.e3
6. Cianfrocco, M.A. *et al.* (2017) COSMIC^2: a science gateway for cryo-electron microscopy structure determination. In *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, pp. 1–5, Association for Computing Machinery
7. Cianfrocco, M.A. *et al.* (2018) cryoem-cloud-tools: a software platform to deploy and manage cryo-EM jobs in the cloud. *J. Struct. Biol.* 203, 230–235