WHITE PAPER

Cryo-EM processing at the pace of medicinal chemistry on AWS

Ieva Drulyte, Adrian Koh, Brian Skjerven, Natalie White, Stephen Litster, Mazdak Radjainia

Introduction

Pharmaceutical companies that discover small molecule drugs use an iterative process known as design-make-test-analyze (DMTA) cycles to generate and optimize lead compounds[1]. In a typical small molecule discovery program, several thousand new compounds are synthesized and evaluated before a drug candidate is identified as suitable for human trials.

The value of Structure-Based Drug Design (SBDD) for rapid and successful progression from initial compounds to a mature drug candidate is well established[2]. With SBDD, a drug discovery program is three times more likely to achieve the desired potency than without[2]. SBDD involves 3D structure determination of compounds in complex with their target protein pinpointing for medicinal chemistry towards optimal compound potency, selectivity and solubility[3]. Structure enablement is most impactful when structures of compound-bound complexes are obtained early on in lead generation when many critical decisions need to be made[1]. Early structures drive detailed mechanistic understanding of how modifying the compound alters its biological activity, thereby contributing to a highly predictive structure-activity relationship (SAR) model after several DMTA cycles. A good SAR model brings the potential of a chemical series to light and allows to design new compounds with high precision. In other words, if structures of compound-bound target complexes can be determined rapidly in the early phases of a program, better drugs can be discovered in less time.

Before the advent of cryo-EM, the benefits of SBDD were largely limited to protein targets for which crystal structures could be determined. Cryo-Electron Microscopy (cryo-EM) is a structural biology technique that does not depend on the crystallization of proteins and is amenable to many previously intractable targets. This includes membrane proteins accounting for approximately 60% of drug targets, and large soluble protein machineries.

The challenge for cryo-EM is the delivery of timely structures, and ultimately, timely repeat structures. For integration of structures into DMTA cycles, turning around multiple compound-bound structures in a matter of days and at resolutions better than 3 Å is critical. Until recently, such timeline and performance were not realistic; however, developments in detector technology and cryo-EM data collection strategies now allow the collection of most datasets in a day or less[4]. The bottleneck has moved to processing terabyte-sized datasets and the question of how to significantly compress data processing timelines.

Intrigued by preliminary benchmarks, we wanted to explore how quickly we could process the larger datasets (**Figure 1**), if the cryo-EM software platform cryoSPARC™ [5] is deployed on Amazon Web Services (AWS). Here, we present an AWS

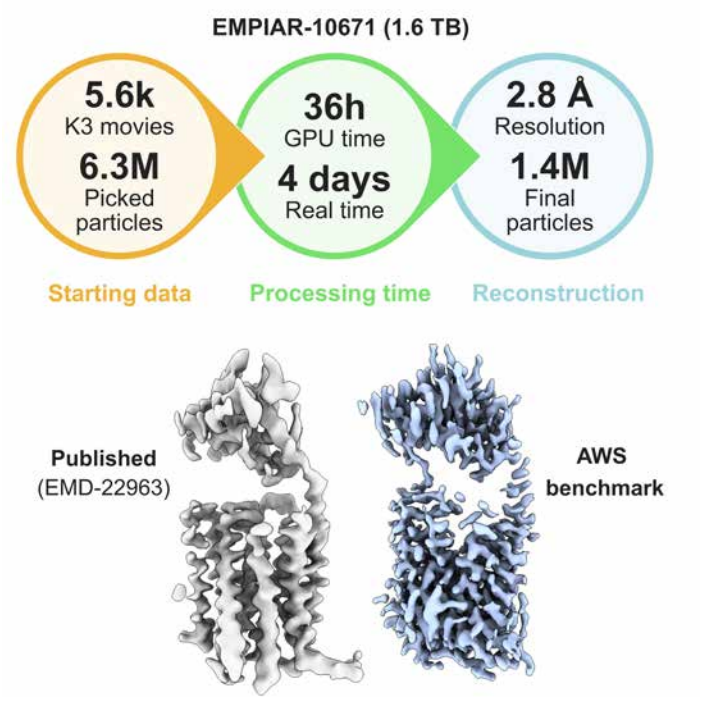

Figure 1. Benchmark of the 74kDa agonist-bound CGRP receptor complex using public data.

ThermoFisher SCIENTIFIC

reference architecture (Figure 2) that we applied to two public G protein-coupled receptor (GPCR) and one internally collected SARS-CoV-2 spike protein datasets showing that <3 Å resolution structures can be obtained in four days or six hours, respectively.

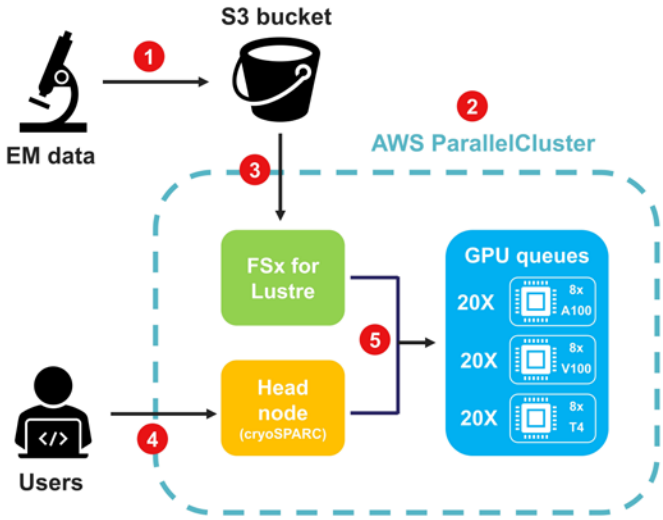

Figure 2. Infrastructure-as-code provides on-demand HPCs for cryo-EM.
1 – EM data flows automatically to an Amazon S3 bucket using Amazon's DataSync tool.
2 – A tailored cluster is deployed from a single configuration file.
3 – Auto-import (not copying) of cryo-EM data to FSx for Lustre as high-speed caching storage for GPU nodes.
4 – Users interact with a low-cost head node where cryo-EM programs like cryoSPARC are installed.
5 – Job schedulers submit to GPU nodes of choice, which are automatically provisioned and shutdown after jobs finish.

AWS architecture for cryo-EM

AWS ParallelCluster for on-demand compute

AWS developed a cryo-EM solution that makes it easy to create, configure and manage high performance compute (HPC) clusters on-demand. Users execute a single command line option to spin up a cluster from a single configuration file. Setting up a cluster is handled by ParallelCluster. ParallelCluster is an AWS service that automatically and securely provisions resources needed for HPC applications in an automated and secure manner. It also supports multiple instance types and job submission queues, as well as job schedulers like AWS Batch, SGE, Torque and Slurm. Using these and other configurable parameters, scientists can design clusters that are tailored to the needs and priorities of cryo-EM users.

The HPC needs of cryo-EM come with several complexities that are difficult to solve on-premises. Cryo-EM analysis requires flexible access to state-of-the-art GPUs. Depending on the stage of the data processing pipeline, speed may or may not scale with the number of GPUs per node. Rather than relying on aging and expensive hardware in on-premises data centers, cryo-EM workloads on AWS can take advantage of the flexibility of state-of-the-art GPUs. Scientists can experiment with and find the optimal balance between cost and performance benchmarking against the number of GPUs per node. AWS ParallelCluster provisions Elastic Compute Cloud (EC2) instances that are perfectly matched with the needs of a given step in the processing workflow. Rather than on-premises resource contention during peak processing times and idle resources during processing valleys, on AWS we only have to pay for what we use and can scale the processing resource up and down to match the workload needs. This eliminates the problem and inefficiency of planning for peak capacity. With ParallelCluster, the required compute is always available, and users pay only for what was used.

High Performance and Cost-Efficient File Storage with FSx for Lustre

There are two reasons why increasing the speed of cryo-EM processing is important. Primarily, high-speed serves to obtain structures faster to meet the timelines of DMTA cycles, as discussed above. Secondly, shorter processing times reduce the spend on GPU nodes. To make GPU nodes perform as fast as possible, it is critical to have a fast-caching file storage solution to keep the GPUs fed. For cryo-EM disk read or write (I/O) performance is particularly relevant as large particle stacks need to pass in and out of the GPU's memory. Disk I/O performance is indeed rate-limiting for cryo-EM processing because it is very I/O intense.

FSx for Lustre is a fully-managed AWS service that delivers a high-performance parallel filesystem, that is optimized for workloads such as HPC, machine learning, analytics, electronic design automation, and media processing. FSx for Lustre integrates well with ParallelCluster and Amazon S3 buckets, the latter presenting a low-cost file system for so-called object storage. Using a seamless combination of S3 and FSx for Lustre gives users the best of both worlds, in terms of costs and performance. Since FSx for Lustre is shared between all compute nodes, it further reduces costs by not requiring compute nodes that come with local storage. In short, FSx for Lustre is a cost-effective and ultra-performant shared drive for rapid read/write access to temporary cryo-EM files in order to speed up cryo-EM reconstructions.

Automated cryoSPARC deployment on AWS

CryoSPARC is a complete solution for cryo-EM data processing that is developed by Structura Biotechnology Inc, for use in research and drug discovery. It stands out in its speed from raw data to 3D reconstruction. It features unique algorithms[6, 7] that address issues of flexibility and perform particularly well for therapeutically relevant targets, such as membrane proteins. With its inherent speed, cryoSPARC is ideal for integration of cryo-EM with DMTA cycles, while keeping the compute footprint small. We automated our deployment of cryoSPARC on AWS using a configuration file that includes execution of post-install scripts that set up a cryo-EM processing environment together with the infrastructure. With a single command line option, hardware and software stand ready for cryo-EM processing.

CryoSPARC on AWS benchmarks

To understand and showcase the performance of cryoSPARC on AWS, we processed three cryo-EM datasets that are representative of drug targets and traditionally difficult to process. Since the goal was to minimize the time-to-structure, we used AWS p4d.24xlarge instances throughout, which feature eight latest generation NVIDIA A100 GPUs.

Agonist-bound CGRP receptor (74 kDa)

We performed the first benchmark on a public EMPIAR dataset with the accession code 10671, corresponding to agonist-bound CGRP receptor[8]. CGRP receptor is a GPCR and established drug target for the treatment of migraine. Rapid access to structures can help develop small molecule CGRP inhibitors that improve on existing therapies by increased potency and being more brain-penetrant.

EMPIAR-10671 is challenging to process due to the small size of the protein lacking prominent features. The images are also very crowded leading to a very large number of particles. The raw dataset is 1.6 TB in size and comprises approximately 5600 Gatan K3 movies. We reconstructed agonist-bound CGRP to a resolution of 2.8 Å in 4 working days from 6.3 million initial

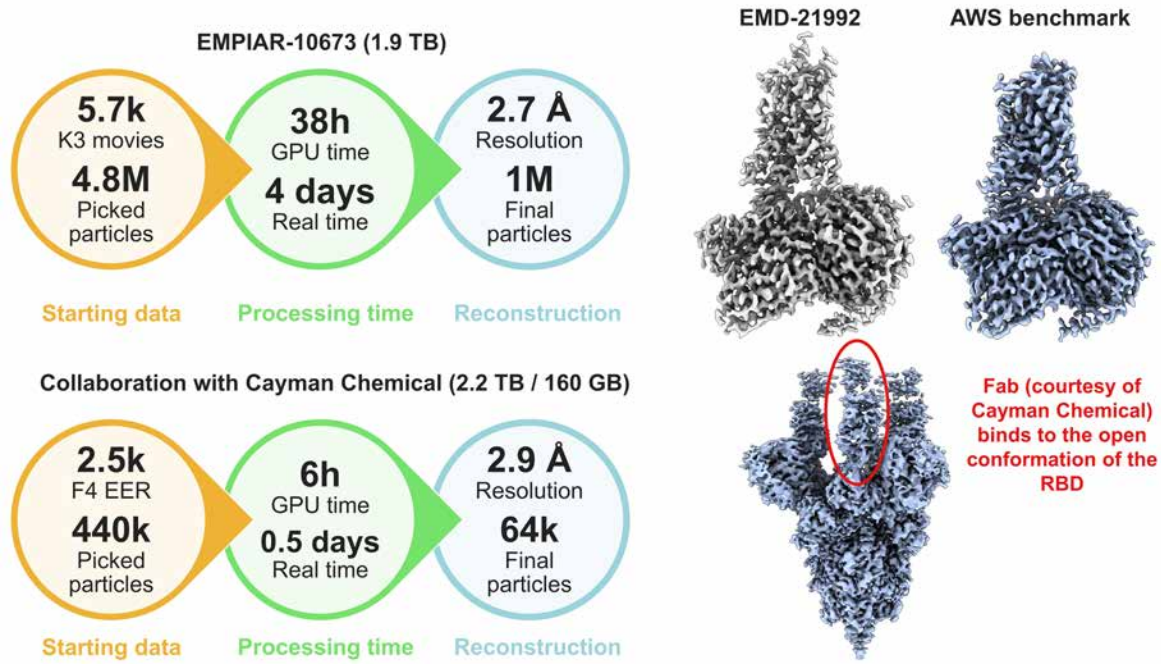

Figure 3. Benchmarks of public GLP1-R/GLP-1 complex dataset (top) and in-house dataset for SARS-CoV-2 spike protein in complex with a Fab (bottom). With cryoSPARC on AWS, repeat structures of average to hard cryo-EM SBDD projects can be obtained within 1-2 days.

particles (Figure 1). The actual processing time was only 36 hours, which is more indicative for repeat structures when the most direct path to reconstruction is established. Our reconstruction presents a marked 0.7 Å resolution improvement on the published CGRP structures, which is likely attributable to so-called non-isoform refinement in cryo-SPARC.

Active GLP-1R/GLP-1 complex

We benchmarked our second dataset using public dataset EMPIAR-10673, which is a cryo-EM dataset of the GLP1-R/GLP-1 complex[9]. GLP1-R is a validated target for diabetes and weight loss with peptide GLP-1 being its native ligand. The market for GLP1-R agonists is highly lucrative and fiercely contested. Cryo-EM SBDD can provide edge in arriving at a best-in-class drug, particularly if it can be integrated with DMTA cycles.

Active GPCR datasets are notorious for preferred orientation, which requires excess data collection to enrich rare views. This inevitably increases processing times. In 4 working days, we obtained a 2.7 Å reconstruction of EMPIAR-10673 that is qualitatively comparable to published dataset (**Figure 3**). The actual compute time to structure was only 38 hours starting from approximately 5700 Gatan K3 movies (1.9 TB) and 4.8 million picked particles.

SARS-CoV-2 Spike protein in complex with a Fab

For the third dataset, we chose an internal dataset representing the spike protein of SARS-Cov-2 in complex with a commercially available Fab courtesy of Cayman Chemical. The spike protein of SARS-CoV-2[10] is the primary target of COVID-19 vaccines and therapeutic antibodies.

In this case, processing on AWS started from motion-corrected images, rather than raw movies. This is because the approximately 2500 Thermo Scientific Falcon 4 Direct Electron Detector camera EER movies were already motion-corrected on-the-fly, which reduced upload size from 2.2TB to 160GB. Picking 440 thousand particles, a 2.9 Å structure was obtained in less than a day, involving 6h of actual compute (see **Figure 3**). This result highlights that for many real-life targets, particularly complexes that are larger or have symmetry, high-resolution structures can be obtained in hours (**Table 1** and **Figure 4**).

	GPUs	CGRP	GLP1R	Spike
Patch motion (M)	8	1 h 40 min	2 h 29 min	-
Patch CTF (M)	8	35 min	38 min	13 min
Blob picker	1	33 min	30 min	25 min
Particle extraction	8	15 min	12 min	3 min
2D classification	4	5 h 1 min	4 h 17 min	1 h 28 min
Heterogenous refinement	1	12 h 14 min (5 rounds)	10 h 38 min (4 rounds)	20 min
Ab Initio Reconstruction	1	6 h 34 min	13 h 26 min	-
Particle re-extraction	8	13 min	11 min	5 min
Non-uniform refinement	1	8 h 57 min	5 h 39 min	3 h 7 min
Total runtime		36 h 2 min	38 h	5 h 41 min

Table 1. Breakdown of cryoSPARC processing time by benchmark.

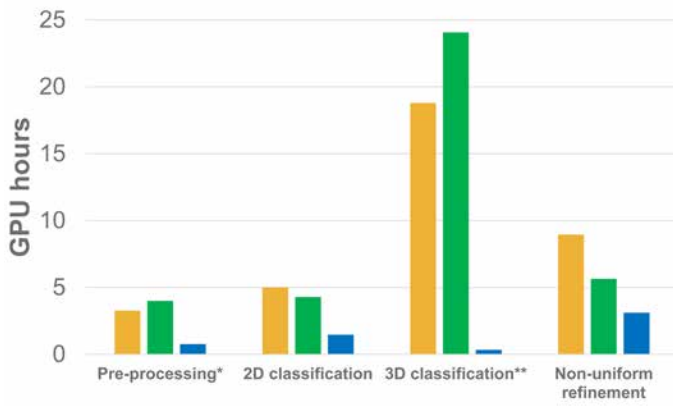

Figure 4. Breakdown of cryoSPARC processing times by processing stage. *Pre-processing includes motion and CTF correction, picking and extraction; ** 3D classification includes heterogenous refinement and Ab Initio Reconstruction.

thermo scientific

Conclusions

With up-to-date hardware and state-of-the-art processing software, it is now possible to obtain repeat structures fast enough for informing DMTA cycles yielding exquisite SAR models. Ultra-fast processing workflows are also valuable for reducing the time to a target's first structure or training novices in cryo-EM workflows. For the former, fast processing can provide immediate feedback on protein quality or grid optimization thereby accelerating project progress. In regard to the latter, new users will gain experience much quicker, the more often they can iterate over the cryo-EM workflow in a short amount of time.

But deploying and maintaining up-to-date IT for cryo-EM HPC comes with many complexities, lengthy delays as well as high up-front costs. In this white paper, we present an infrastructure-as-a-code approach that is easy to implement and highly geared to obtain structures as fast as possible. We show that for prototypical drug discovery projects, structures can be obtained in hours to a day on AWS. The automated deployment makes it easy to manage one or even multiple HPCs without upfront costs. HPCs can be brought up or down as they're needed, only paying for the storage and compute that was used. ParallelCluster's elasticity is useful for companies of all sizes. Large companies with many programs can easily scale up their resources, when cryo-EM needs to drive rapid progression of multiple programs in parallel. Small companies, particularly small biotech companies with very few cryo-EM programs, may only have intermittent compute needs. The high flexibility of on-demand clusters is therefore broadly applicable, greatly reduces IT barriers to cryo-EM and it enables structure at the pace of medicinal chemistry for previously intractable drug targets.

About the authors

Ieva Drulyte, Adrian Koh, and Mazdak Radjainia are employees of Thermo Fisher Scientific.
Brian Skjerven, Natalie White, and Stephen Litster are affiliated with Amazon Web Services.

References

1. Wesolowski, S.S. and Brown, D.G. (2016). *The Strategies and Politics of Successful Design, Make, Test, and Analyze (DMTA) Cycles*, in Lead Generation, J. Holenz (Ed.).
2. Hajduk PJ, Greer J.; *A decade of fragment-based drug design: strategic advances and lessons learned*, Nat Rev Drug Discov., 2007 Mar 6(3):211-9.
3. Collie GW, Michaelides IN, Embrey K, Stubbs CJ, Börjesson U, Dale IL, Snijder A, Barlind L, Song K, Khurana P, Phillips C, Storer RI. *Structural Basis for Targeting the Folded P-Loop Conformation of c-MET*, ACS Med Chem Lett. 2020 Dec 8 12(1):162-167.
4. Christoph Wigge, Aleksandar Stefanovic, Mazdak Radjainia. *The rapidly evolving role of cryo-EM in drug design*, Drug Discovery Today: Technologies, 2021.
5. Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA. *cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination*. Nat Methods. 2017 Mar;14(3):290-296.
6. Punjani A, Zhang H, Fleet DJ. *Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction*, Nat Methods. 2020 Dec;17(12):1214-1221.
7. Punjani A, Fleet DJ. *3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM*, J Struct Biol. 2021 Jun;213(2):107702.
8. Josephs TM, Belousoff MJ, Liang YL, Piper SJ, Cao J, Garama DJ, Leach K, Gregory KJ, Christopoulos A, Hay DL, Danev R, Wootten D, Sexton PM. *Structure and dynamics of the CGRP receptor in apo and peptide-bound forms*, Science. 2021 Apr 9;372(6538):eabf7258.
9. Zhang X, Belousoff MJ, Zhao P, Kooistra AJ, Truong TT, Ang SY, Underwood CR, Egebjerg T, Šenel P, Stewart GD, Liang YL, Glukhova A, Venugopal H, Christopoulos A, Furness SGB, Miller LJ, Reedtz-Runge S, Langmead CJ, Gloriam DE, Danev R, Sexton PM, Wootten D. *Differential GLP-1R Binding and Activation by Peptide and Non-peptide Agonists*, Mol Cell. 2020 Nov 5;80(3):485-500.e7.
10. Hsieh CL, Goldsmith JA, Schaub JM, DiVenere AM, Kuo HC, Javanmardi K, Le KC, Wrapp D, Lee AG, Liu Y, Chou CW, Byrne PO, Hjorth CK, Johnson NV, Ludes-Meyers J, Nguyen AW, Park J, Wang N, Amengor D, Lavinder JJ, Ippolito GC, Maynard JA, Finkelstein IJ, McLellan JS. *Structure-based design of prefusion-stabilized SARS-CoV-2 spikes*. Science. 2020 Sep 18;369(6510):1501-1505.

Find out more at
thermofisher.com/PharmaceuticalResearchusingCryoEM

ThermoFisher SCIENTIFIC